Chapter 1
Introduction to Precision Machine Design

Why do you like to design machines? "It is the pitting of one's brain against bits of iron, metals, and crystals and making them do what you want them to do. When you are successful that is all the reward you want."

Albert A. Michelson

1.1 INTRODUCTION

Companies can remain competitive in world markets only if they develop new technologies and methods to keep one step ahead of the competition; maintaining the status quo is not acceptable. Hence new machines need to be designed with increased speed, accuracy, and reliability. This leads to the need for designers who have a deep understanding and love of the art and science of design.[1]

In a broad sense, the art and science of design is a potent vitamin that must be taken in balance with other mental nutrients, such as mathematics, physics, manufacturing, hands-on experience, and business skills. Most people exercise to keep physically fit so they can enhance their enjoyment of day-to-day living. Analogous to physical exercise, analysis is a form of mental pushup that trains the mind to be strong and swift. Indeed, many designs would never have even been conceived of if the design engineer did not understand the basic physics behind the process that prompted the need for a new design. Similarly, knowing how to build things can enable the design engineer to develop easily manufacturable products that are a pleasure to use. As illustrated in Figure 1.1.1, the need to integrate various disciplines means today's design engineer must be a Renaissance person. Design engineers must be more creative than their competition and more observant of the world around them. In today's tough international competitive world, if you want something, you can only obtain it with blood, sweat, tears, and design.

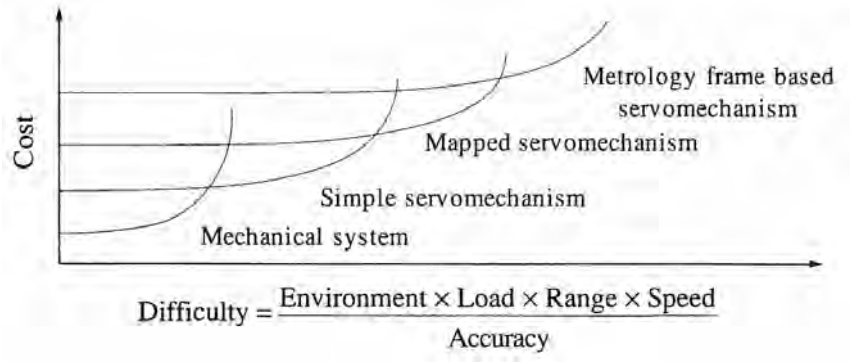


Figure 1.1.1 Increasing difficulty often leads to the integration of engineering disciplines.

In the future, expert systems may evolve to replace mundane engineering tasks. However, it is unlikely that computers will ever be able to do creative design. If a computer program can be designed to do creative design, a computer program can also be designed to design new computer programs. Thus there will always be jobs for creative design engineers. However, high-paying engineering jobs will soon no longer be available for students who lack good creative and/or analytic skills.[2] On the other hand, the future for bright, creative, hard-working design engineers is very promising. How can new design engineers be taught to think and be creative? Integrating theory and application with real-world considerations seems to be a good method and is stressed in following chapters. The remainder of this chapter addresses broad issues including:

---

[1] *"Enthusiasm is one of the most powerful engines of success. When you do a thing, do it with all your might. Put your whole soul into it. Stamp it with your own personality. Be active, be energetic, be enthusiastic and faithful and you will accomplish your object. Nothing great was ever achieved without enthusiasm."* Ralph Waldo Emerson

[2] *"The hero of my tale, whom I love with all the power of my soul, whom I have tried to portray in all his beauty, who has been, is, and will be beautiful, is Truth."* Leo Tolstoi

- Basic economics
- Basic project management skills
- Design philosophies
- The design process in the real world

1.2 FUNDAMENTALS OF ECONOMIC ANALYSIS[3]

The initial specifications for a machine are often generated by a company's salespeople, who are responding to requests from customers; however, it often seems as if the customer wants infinite performance for zero cost. When presented with customer requirements, it is the duty of the design engineer to sketch out realistic options and cost estimates for a family of possible designs that could meet the customer's specifications. This initial step is usually done by senior design and manufacturing engineers with experience in determining just how long and how much it will cost to design and build a new product.

The underlying principle of economic analysis is that money has different values depending upon when it is actually received or spent. The simplest example of this is a traditional passbook savings account; $1000 deposited today at 6% interest will be worth approximately $1349 five years from now. The amount of $1000 today or $1349 in 5 years is equivalent, assuming that 6% is the highest interest rate you could obtain. Because money received or paid out at some future time has a different value when considered at "what it is worth today," economic analysis is vital to decisions on machinery purchases and hence can greatly influence a machine's design. It should be stressed that there are many ways to evaluate an investment decision; the necessary brevity of this section precludes a more detailed discussion, and it is recommended that all engineers complete an engineering economics course at some point in their career.

1.2.1 Cashflow Timelines

The first step in evaluating an investment is to identify the applicable cashflows, as well as when they are expected to occur. One convenient method for visualizing cashflows is to mark them on a timeline.

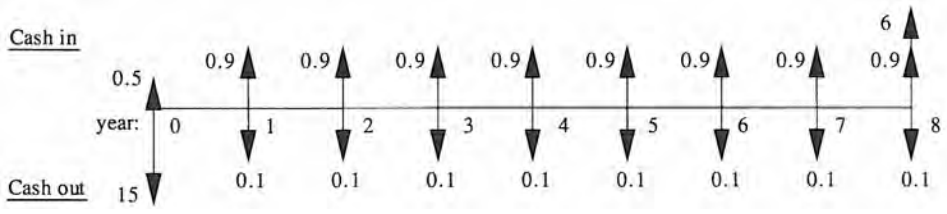


Figure 1.2.1 Cashflow (thousands of dollars) timeline for 1.

Example 1

An automated drill press is under consideration for purchase for $15,000. It is estimated that by replacing the old drill press with a new one, the company will save $900 per year in labor costs, but there will be an added maintenance cost of about $100 per year. The old press can be sold for $500 today, and the estimated salvage value of the new press in 8 years is $6000. Figureq 1.2.1 shows the applicable cashflows as a function of time. Subtracting the "cash out" amounts from the "cash in" amounts for each given year, the result is the net cashflow timeline shown in Figure 1.2.2. The timeline demonstrates that in the beginning of the first year $(t = 0)$, the company has a net negative cashflow of $14,500. In years 1 to 7 there is a net positive cashflow of $800, and in year 8 there is a $6800 positive cashflow from the sale of the machine and that year's profits. To answer the question "Should the company buy the new drill press?", it is necessary to evaluate the predicted cashflows; doing so requires an understanding of compound interest factors.

[3] Section 1.2 was written by Richard W. Slocum III.

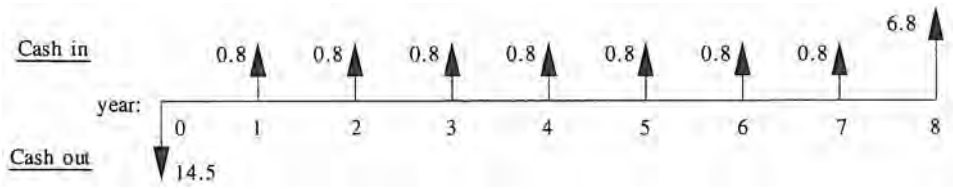


Figure 1.2.2 Net cashflow (thousands of dollars) timeline for Example 1.

1.2.2 Compound Interest Factors

The time value of money concept is also referred to as *interest* and *interest compounding*. The value of various types of cashflows at different times along the timeline can be calculated, using the following mathematical factors/formulas, in which i is the interest rate per period (e.g., $i = 0.1$ for 10% interest) and n is the number of periods.

- *Single-Payment Future Worth Factor*: This gives the value of a cashflow now $(t = 0)$ at some time n periods in the future:

$$(F/P, i, n) = (1 + i)^n \tag{1.2.1}$$

- *Single-Payment Present Worth Factor*: This gives the present value $(t = 0)$ of a cashflow that will occur n periods in the future:

$$(P/F, i, n) = \frac{1}{(1 + i)^n} \tag{1.2.2}$$

- *Uniform Series Future Worth Factor*: This gives the value (at some time n periods in the future) of a uniform series of cashflows occurring once per period:

$$(F/A, i, n) = \frac{(1 + i)^n - 1}{i} \tag{1.2.3}$$

- *Uniform Series Present Worth Factor*: This gives the present value $(t = 0)$ of a uniform series of cashflows occurring once per period:

$$(P/A, i, n) = \frac{(1 + i)^n - 1}{i(1 + i)^n} \tag{1.2.4}$$

- *Capital Recovery Factor*: This gives the uniform series of cashflows over n future periods that are equivalent to one cashflow at the present $(t = 0)$:

$$(A/P, i, n) = \frac{i(1 + i)^n}{(1 + i)^n - 1} \tag{1.2.5}$$

  The product of the principal and the CRF is the payment due each period. This is the factor used to calculate the payment required to repay a car loan or a home mortgage.

- *Sinking Fund Factor*: This gives the uniform series of cashflows over n future periods that are equivalent to one large cashflow in period n:

$$(A/F, i, n) = \frac{i}{(1 + i)^n - 1} \tag{1.2.6}$$

  It used to be that these formulas were too cumbersome to use, so they were tabulated for various interest rates and time periods; however, modern programmable and financial calculators and PC-based spreadsheet programs eliminate this problem.

Example 2

Fred wishes to borrow $\$12,000$ to apply toward purchase of a new car. If the bank is offering 5-year loans at 12%, what will Fred's monthly payment be? The *Capital Recovery Factor* is used here. The period is 1 month, there are 60 periods in 5 years, and the interest rate per period is

12%/12 months per year, or 1%: Monthly car payment = (A/P ,1%,60) $\times \$12,000$. The A/P factor is 0.02224, so Fred's payment is \$266.88 per month.

Example 3

Mary is presently a sophomore in college and she wishes to begin a monthly savings program that will accumulate enough money so that by graduation she will be able to make a down payment on a home. She has 29 months until graduation and estimates that $\$11,000$ will be required for a down payment. If interest rates on deposits remain constant at about 6%, how much should she save each month? The *Sinking Fund Factor* is used here. There are 29 periods and the interest rate per period is 6%/12 months per year, or 0.5%/period: Required monthly savings = (A/F,0.5%,29) $\times \$11,000$. The A/F factor is 0.0321, so Mary must save \$353.10 per month.

1.2.3 Economic Analysis of Projects

Economic evaluation of projects involves a determination of whether a series of cashflows over time, as shown in the timeline of Example 1, meet a company's investment criteria. Usually, a company's management establishes a minimum interest rate that investments must provide in order to be considered worthwhile.[4] It is generally safe to say that this minimum rate is somewhere near the rate for long-term government Treasury bonds (which are virtually risk-free), plus a factor for estimated risk. For example, if Treasury bonds yield a risk-free 9%, why should the company invest in a high-risk project unless that project will yield at least the T-bill rate plus, say, 3% to cover the increased risk? In this instance, management would thus set the minimum interest rate at 12%. The minimum interest rate is referred to by a number of terms including:

- Desired (minimum) rate of return
- Discount rate
- Hurdle rate
- Desired (minimum) yield

All refer to the same interest i as discussed above. There are two principal methods for determining if a series of cashflows meet an established minimum rate of return:

1. Calculate the rate of return that results from the cashflows, and compare it to the minimum rate of return. Calculation of rate of return is a matter of setting up the applicable cash flows with the appropriate factors (P/A, A/F, and so on), and then solving for the interest i in the expression for the sum total of all cashflows (in and out): $PW_i = 0 =$ sum of cash flows discounted at some interest rate i.
2. Calculate the present worth of the cashflows using the minimum rate of return to discount (or bring back) to the present cashflows that occur in future time periods. This analysis calculates the present worth of the investment, expressed as PW_x where x is the percent minimum rate of return, or discount rate. If the PW_x is positive, it is known that the proposed investment's rate of return is greater than the established minimum. If the PW_x is negative, it is known that the proposed investment's rate of return is below the established minimum.

Because calculation of the rate of return is an iterative process, it is time consuming for all but the simplest cashflows. Calculation of the PW_x is relatively straightforward and returns a quick go or no-go decision on the proposed investment.

Example 4

The management of the company in Example 1 has established a minimum rate of return of 12% for all projects. Should the new drill press be purchased? First, set up the equation as shown in Figure 1.2.3. Note that cashflows out of the company are, by convention, expressed as negative numbers, and cashflows into the company are expressed as positive numbers. Since the PW_{12} of $-\$8102$ is negative, the proposed new drill press does not meet the company's required investment criteria, and the press should not be purchased. Many investment decisions are analogous to the preceding example; the key to accurate investment analysis is to correctly identify the applicable cashflows that will result from a proposed investment.

[4] The danger in this method is that it is sometimes applied without including such effects as quality and employee and customer satisfaction.

$$\underset{}{PW_{12}} = \underset{\substack{\text{No multiplier}\\ \text{since t = 0}}}{-\$14,500} + \underset{\substack{\text{P/A gives present worth}\\ \text{of the cashflow for years 1-7}}}{\$800(P/A, 12\%, 7)} + \underset{\substack{\text{P/F gives present worth}\\ \text{of the cashflow in year 8.}}}{\$6,800(P/F, 12\%, 8)}$$

Figure 1.2.3 Present worth calculation for Example 4.

1.2.4 Machine Cost Determination

There have been countless machines designed to perform tasks in a technically elegant manner. However, only those machines that can operate in a cost-effective manner will become commercial successes.[5] There are two components to the cost of any machine: the fixed cost component and the variable cost component. Fixed costs are defined as those costs that are continually present because the machine is there; thus they occur irrespective of whether the machine is producing products. Examples of fixed costs include: cost of required spare parts inventories, certain maintenance costs, and cost of floor space. Variable costs are costs incurred that are directly related to the amounts of product produced. Examples of variable costs include: material, labor, most maintenance costs, and utilities (i.e., power) used by the machine. The task of comparing various alternative machines is mathematically simple, consisting generally of addition, subtraction, and multiplication of the various cost components. The difficult part of the problem is identifying which cost components are relevant to a given situation and developing realistic estimates of production rates, wages, interest rates, taxation rules, and so on.

Initial Capital Expenditure

Initial capital expenditures are the up-front costs associated with placing a given piece of machinery into service. There are numerous expenses to consider, many of which are easily overlooked by engineers in the early stages of design. Nonetheless, they are real costs that will be included in evaluations by potential purchasers. Examples of initial capital expenditures include:

- Cost of the machine itself.
- Cost of freight and rigging; machines that are oversized to the point of not being truckable on a standard-size truck may incur significant extra shipping costs.
- Cost of spare parts inventory.
- Cost of training employees to operate and maintain the machine.
- Cost of physical plant modifications, such as power sources, structural changes to accommodate weight, size, vibration, noise, and so on.

Awareness of these factors can help a design engineer to minimize a machine's associated costs. In many cases this can be done without changing the machine's basic design or increasing its cost. For example, if one is designing a large electric-powered press that utilizes an innovative method of high-speed operation, the design engineer can:

- Make the machine so that it comes apart into major subassemblies for ease of shipping and installation.
- Utilize gearboxes, motors, and other general components that are also used on other machines of this type. This reduces the purchaser's need to develop another entire supply of spare parts for the machine, and may simplify the purchaser's training expenses for repair technicians. For instance, if research shows that 75% of presses use "Brand X" motors with "Brand D" speed controllers, there is some merit in this machine also doing so, unless there is a sound technical or economic reason to do otherwise.
- Utilize control logic systems similar
to other machines of this type, unless there is a sound technical or economic reason to do otherwise. This will reduce operator training expenses associated with the machine.
- Design the component parts' layout for serviceability and operator comfort and safety.

[5] "A hen is only an egg's way of making another egg." Samuel Butler

For this example, the cost of implementing these suggestions would probably be small, yet they may result in a more economical machine design than, for instance, a press that requires double-wide trucks for transport, has many hard-to-find motors and gearboxes, and has an operator control system unlike most other machines of its type.

Fixed Maintenance Costs

Fixed maintenance costs are costs to maintain the machine that are not dependent upon the volume of production. For example, if the lubrication oil must be changed "every month or 250 hours of operation," the cost to do so is a fixed maintenance cost, assuming that the machine is used on a 5-day-per-week, single-shift production schedule. In general, fixed maintenance costs are not as significant as variable maintenance costs for machines that are in relatively constant use. For machines intended primarily for standby use, such as emergency generators and other backup systems, the fixed maintenance costs are usually greater than variable costs.

Variable Maintenance Costs

Variable maintenance costs are directly dependent on the amount of usage a machine receives. Items such as belt, cutting tool and electric motor brush changes are examples of variable maintenance costs. In the example of the preceding paragraph, if it is assumed that the machine will be used on a two-shift-per-day basis (double shift), the cost of lube oil changes can become a variable maintenance cost.

Capital Asset Depreciation

As a machine ages, it generally wears out and loses value.

Depreciation is an accounting term for the cost of the wearing out per year. It is important to note that from the standpoint of cashflows, a company will generally spend the cash at the time of machine purchase. On the other hand, the depreciation charge occurs over a set number of years known as the *recovery period*.[6] There is, in effect, a cash inflow in each of those years because the depreciation charge is not an actual cash outflow. On the contrary, the depreciation charge reduces a company's taxes by the amount of the depreciation charge multiplied by the company's marginal tax rate.[7] There are many methods of calculating exactly how much depreciation to charge in a given year. For simplicity, only the commonly used *straight line method* will be considered here. This method assumes that the machine "wears out" (i.e., depreciates) an equal amount during each year of the recovery period. Thus the annual depreciation charge for an investment is equal to

$$\text{depreciation} = \frac{\text{initial capital expenditure}}{\text{investment recovery period}} \tag{1.2.7}$$

Product Reject Rate

Inasmuch as reject rates directly affect costs, they must be considered when determining a machine's economic feasibility. There are several potential areas of cost impact:

- Material wasted by the reject.
- Damage to other parts caused by the reject. For example, in the case of a robotic microchip installer, an error in gripper position may damage the circuit board that it is working on. If the circuit board is nearing completion, the damage may cost several thousand dollars. Some complex parts, such as turbine rotors, have millions of dollars of value-added worth.
- Downtime caused by the reject. In the case of a stand-alone lathe, the downtime costs may be minimal. In the case of an assembly line robot, downtime costs can be substantial if the entire line must be stopped to correct the error.

Machine design engineers must thus be aware of the operating characteristics of the market for which their machine is intended.

Ripple Effects

The maintenance period and operating characteristics of the machine should be considered with respect to other machines in a plant, and the effect of any resultant interdependency. If a

[6] The 1987 federal tax law specifies various recovery periods for different types of investments. For example, the recovery period for most machinery is 5 years; most commercial buildings have a recovery period of 32.5 years.

[7] Under 1987 federal tax law, the marginal tax rate for corporations with profits over $\$100,000$ is 36%.

shop buys a numerically controlled (NC) lathe to manufacture 1000 gadgets per day, but the shop's stockroom can only supply stock cut to length for 500 parts, additional machinery and equipment will have to be purchased and figured into the cost of using the new lathe. Similarly, if a high price is paid for a new machine that requires little maintenance, but the machine is used to finish parts made on an old machine that often breaks down, the new machine could be idle as a result of another machine's faults. On the other hand, if a new machine is made too cheaply and often needs repairs, it may quickly obtain a bad reputation as a plant stopper and will be shunned by buyers.

Other Tax Considerations

There are very few investment decisions that are made without consideration of tax consequences. The machine design engineer should consider these issues when comparing design alternatives. In general, the tax-related factors that most affect an investment decision are:

- Capital asset depreciation schedules (recovery periods)
- Investment tax credits
- Overall tax rates

Federal and State taxing authorities frequently attempt to institute social change through adjustments to the tax code, and it is thus constantly changing. A machine design engineer does not need to be an expert in all areas of the code; however, a basic familiarity with the sections of the tax applicable to capital investments is very important.

1.2.5 Machine Operator Costs

The task of determining "How much does a machine operator cost?" can be very complex, particularly when factors such as payroll burdens, support personnel, and enforcement of work rules are taken into account. A few of the more common costs that should be included in addition to the basic hourly wage are:

Payroll Burdens: These commonly fall into two groups: taxes (such as social security tax and unemployment tax) and employee benefits (such as insurance, paid vacation, retirement benefits, and savings plans). These costs can be significant and it is not uncommon for payroll burdens to equal the worker's basic hourly wage rate.

Worker Efficiency: The cost of workers' break times and other nonproductive periods must be reflected into calculations of either "cost per hour" or "productivity per hour."

Support Personnel: In many instances the direct efforts of a worker at a machine must be supported by others. For example, a particularly sophisticated machine may require that a special mechanic be on-site to provide service and make adjustments. One mechanic may support, for example, six machines. Thus one-sixth of the mechanic's hourly cost must be included in the cost of the machine's operation. Machines designed with better serviceability and reliability can reduce these types of costs and sometimes make the machine more attractive than an alternative production method.

Work Rules: In many instances the tasks that a worker is allowed to do are restricted by work rules, which is particularly true in unionized plants. For example, a machine operator may not be allowed to adjust the belt tension at his machine; he would be required to call a mechanic from the maintenance department. This results not only in the extra cost for a mechanic, but also in the cost of the machine operator waiting for the mechanic to come and do the work. In order to minimize this problem, the machine design engineer can try to make routine maintenance as infrequent as possible.

Thus the engineer attempting to determine the cost of human labor must include a multitude of factors; the problem is rarely as simple as "What is the operator's hourly wage, and how many widgets can she make per hour?" When precise data on a worker's production rate are unavailable, approximate production rates can often be determined by utilizing various standardized estimating manuals.

1.2.6 Examples

The task of designing a machine can in effect be equated to the task of defining an end use, determining the costs associated with existing machines and/or humans, and then designing a machine which does the same job for a lower cost. The following example will illustrate how various cost factors can affect a purchase decision for a machine, as well as how omitting certain cost factors can lead to incorrect decisions.

Example 1[8]

Widget Metalworking Company currently has a number of standard, manually operated lathes. A contract has been awarded to Widget for production of 3000 stainless steel screw-top acid flask covers per month. The contract guarantees that the purchaser will order 3000 tops per month for the next 5 years. The plant currently works single shifts, 7 days per week, 50 weeks per year. The plant engineer has solicited a proposal from Nifty Machine Tools, a maker of numerically controlled lathes, for equipment to help Widget meet these production requirements. Widget's management has established that investment decisions are to be made on the basis of a 12% discount rate. Working with Widget's engineer, the Nifty sales engineer has developed the cost summary shown in Table 1.2.1. Ignoring tax considerations (for the moment) and Widget's other work load, should Widget purchase a new Nifty lathe?

	Existing lathe	Nifty ×100 lathe	Zipmaster lathe
Operator hourly wage	$25.00	$25.00	$25.00
Annual overhaul cost	$1,500	$7,500	$12,000
Routine overhaul cost	$1,000	$2,000	$3,900
Routine overhaul frequency (h)	2,000	5,000	3,750
Capacity tops per hour	8	35	35
Cost per top	$3.69	$1.19	$1.88
Number of tops per year	36,000	36,000	36,000
Machine cost	$0	($490,000)	($300,000)
Operator training	($500)	($12,900)	($10,000)
Spare parts	$0	($40,000)	($22,000)
Salvage value	$5,000	$240,000	$190,000

Table 1.2.1 Widget's cost summary.

First, the difference between the variable cost of making a top using a manual lathe and a Nifty lathe must be made. Let Wo = operator's hourly wage, Oa = annual overhaul cost, Or = routine overhaul cost, Of = routine overhaul frequency, Q = capacity, in tops per hour, C = cost per top. Then

$$C = \frac{W_o}{Q} + \frac{O_a}{Q \times 8 \text{ hours/day} \times 7 \text{ days/week} \times 50 \text{ weeks/year}} + \frac{O_r}{O_f}$$

Thus, for the old lathe

$$C_{old} = \frac{\$25}{8 \text{ tops}} + \frac{\$1500}{8 \text{ tops} \times 8 \times 7 \times 50} + \frac{\$1000}{2000} = \$3.6920/\text{ top}$$

For the new lathe

$$C_{new} = \frac{\$25}{35 \text{ tops}} + \frac{\$7500}{35 \text{ tops} \times 8 \times 7 \times 50} + \frac{\$1000}{5000} = \$1.1908/\text{ top}$$

Thus the new Nifty lathe could save $2.5012 per top. Note that it is assumed that material costs and reject rates are identical for both machines. The additional profit to Widget Metalworking Company per year, then, is $2.5012 per top times the 36,000 tops that are to be made each year, or about $90,000. In order to determine if this additional profit is worth the required initial capital expenditure, a discounted cashflow analysis is utilized. Let CF_x be the cashflow in year x, with

[8] It is assumed here that Widget, Nifty, and Zipmaster are imaginary companies.

x = 0 being the present and PW_{12} be the present worth using Widget's 12% discount rate. Then, for the new lathe

$$\begin{aligned} CF_0 &= \text{new machine cost} + \text{operator training cost} + \text{spare parts cost} \\ &= (-\$490,000) + (-\$12,900) + (-\$40,000) = -\$542,900 \\ CF_1 &= CF_2 = CF_3 = CF_4 = \$90,041 \\ CF_5 &= \$90,000 + \text{salvage value} = \$90,041 + \$240,000 = \$330,041 \end{aligned}$$

Utilizing the formulas from Section 1.2.2 gives

$$\begin{aligned} PW_{12} &= CF_0 + CF_{1-4}(P/A, 4, 12\%) + CF_5(P/F, 5, 12\%) \\ &= (-542,900) + \$90,041(3.037) + \$330,041(0.5674) = -\$82,139 \end{aligned}$$

The negative PW_{12} indicates that Widget's investment criteria are not met, and purchase of the new lathe cannot be justified.

Example 2

Unhappily for the Nifty sales engineer, the up-front costs of his company's machine were too high for Widget to justify the expenditure. The Widget plant engineer then contacted Zipmaster Machine Tool Company; their machine, he was informed, had the same characteristics as Nifty's except those shown in Table 1.2.1. Should Widget purchase the Zipmaster lathe? Analyzing the Zipmaster machine purchase in the same manner as that of Example 1, Widget's engineer found the PW_{12} to be $11,377. The number is positive, indicating that Widget's investment criteria are met, and the machine could be purchased.

The lesson to be learned from Examples 1 and 2 is that a machine design engineer faces constant economic tradeoffs. Interestingly, in many cases, making the machine more reliable and durable (which is reflected in a higher salvage value and lower maintenance costs) can result in too high an initial cost to allow the product to sell. In essence, the *time value of money* dictates that repair costs and revenues incurred in the future are worth less (when considered at the present time) than their absolute dollar amount at the time of their projected receipt or expenditure in the future. The farther in the future the projected date of expenditure, the less the impact on investment decisions. One must be strongly cautioned, however, to consider systems of machines where downtime on one machine can cripple an entire line. In these cases, it is often worth the cost to pay for reliability.

A machine design engineer must also review the designs of the existing machinery that performs tasks similar to those his machine is intended to replace or compete with. In cases such as the Nifty lathe, the productivity increase of a new machine is simply not enough to overcome its high initial cost. One solution is to change the basic design to reduce the initial costs, even at the expense of higher maintenance costs during the life of the machine, as Zipmaster presumably did.

Example 3

Let us assume that Nifty's machine is produced in the United States, while Zipmaster's is manufactured in Europe. Let us further assume that Widget is a profitable company, paying a marginal tax rate of 42.5% (federal, state, and local). For this example, assume that the tax laws specified that the recovery period and investment tax credit for domestic machines are 4 years and 10%, respectively, and for imported machines 8 years and 0%. How would this affect Widget's decision to purchase a new lathe? For the Nifty machine, the investment tax credit[9] is effectively an outright "rebate" of 10% of the machine's purchase price, or $490,000 10%, or $49,000. By establishing a recovery period of 4 years, the government is, in effect, allowing Widget to "write off" the value of the Nifty lathe in just 4 years; this is a reduction in taxable annual income of (490,000 - 49,000)/4 = $110,250/year. *Because Widget pays a 42.5% tax on its profits, the "reduction" in taxable income results in an annual tax bill that is lower by 42.5% of $110,250, or $46,856. Because Widget is not actually paying out the $110,250 per year depreciation charge, the net effect is that a positive cashflow of $46,856 occurs.*

Recalculating the PW_{12} for the Nifty lathe utilizing the tax rules yields $109,180. The number is positive, indicating that Widget's investment criteria are met by the Nifty lathe as well. Now the

[9] An investment tax credit is the percentage of an investment's cost that is credited to the purchasing company's tax bill by the government offering the tax credit. The amount of the tax credit must be subtracted from the initial capital investment amount before dividing by the number of years over which the asset is to be depreciated.

Widget plant engineer is faced with the choice of which machine to purchase. A similar calculation must be made for the Zipmaster machine. Inasmuch as the tax rules mandate a different treatment for imported machinery, the reduction in Widget's taxable income is less and results in a recalculated PW_{12} of \$68,828. Now the economics show that the Nifty lathe has a higher after-tax present worth, and Widget's decision should be to purchase the Nifty lathe. Tax consequences[10] can play an important role in an investment decision. This is particularly true when comparing machinery to human labor, since tax laws frequently give economic benefits to machinery investments.

1.3 PROJECT MANAGEMENT: THEORY AND IMPLEMENTATION [11]

Once the conceptual ideas for a new machine are approved by management and the customer, the next step is to develop a detailed plan for completing the design.[12] This involves blending heuristic design methodologies with project management tools to ensure that the final design meets the customer's specifications.

The practice of project management is essential to virtually any undertaking, regardless of size or complexity. In many instances the projects at hand are not complicated and their "management" can easily be done in one's head. Formalized project management involves a set of techniques to plan and control projects[13] that are too complex to be handled in one's head. In general, the techniques define the individual activities that compose a project and their various interrelationships. The result is a "road map" of the activities that will occur during the course of a project and the time and order at which they are expected to occur. There are several established methods of project management; the better known ones are *CPM*, (Critical Path Method) and *PERT* (Program Evaluation and Review Technique). All methods involve relatively simple underlying concepts. Until recently, the principal obstacle to more widespread use of these methods has generally been that all but the smallest projects involve great quantities of calculations. Today there are numerous project management software packages for use on personal computers. This section will introduce the reader to some of the concepts and techniques of effective project management. Their use in conjunction with one of the many readily available project management software packages and a personal computer will enable a design engineer or manager to plan and control a project of virtually any size.

1.3.1 General Concepts

In developing a road map for a project, it is necessary to know where the project starts, what it is intended to accomplish, and what steps are required to reach the final goal. Furthermore, attributes of each step, such as cost, required time, and personnel requirements, must also be known. After determining each of these, the project manager can lay them out in block form, examine their various interrelationships, and attempt to eliminate situations which might slow the project. By comparing the actual progress of the project to the initial plan, the manager can quickly identify parts of the project that are, for example, falling behind schedule or are over budget. The various "key points" of a project are known as *events*. Examples of these are:

- Project start
- Engineering completion
- All materials on site
- Project completion

Events are defined as milestones and thus do not involve time, effort, or cost. *Activities* are tasks that must be completed during the course of a project:

- Define project work plan
- Review preliminary designs
- Check final design
- Procure materials
- Construct prototype

[10] Certain tax practices, such as depreciation, recapture, capitalization of spare parts/training, and so on, have been neglected here in the interest of simplicity. These are all factors which marketing generally takes into account when deciding a pricing strategy.

[11] This section was also written by Richard W. Slocum III.

[12] *"Be always sure you're right–then go ahead."* David Crockett

[13] *"Never tell people how to do things. Tell them what to do and they will surprise you with their ingenuity".* General G.S. Patton

Activities generally involve effort by various individuals, groups, or machines (resources), with associated time (activity duration) and cost (activity cost) characteristics. In practice, the breakdown of a project into a large number of discrete activities results in better planning and control than would result from a few large activity descriptions. *Dependencies* are the various relationships between activities, such as:

- Which activities must be finished before other activities can begin
- Which activities may be performed simultaneously

In some cases the dependencies result from process considerations, and in other cases dependencies result from limitations on available resources. Consider the following activities that relate to building a brick wall upon an existing foundation:

Activity A: Stage bricks at the job site.
Activity B: Prepare the cement mixer.
Activity C: Mix the cement.
Activity D: Build the wall.

Because of the fact that the wall cannot be built until the bricks are at the job site, the cement mixer is prepared, and the cement mixed, activity D can only begin after activities A, B, and C are complete. Thus activity D is a *process-induced dependency*. Activities A and B are not dependent on each other due to process considerations because one person could haul bricks to the job site while another could prepare the mixer. However, if there is only one person available to build the wall, that person obviously cannot haul bricks and prepare the mixer at the same time. In the latter case, activities A and B are subject to a *resource-induced dependency*. It is incumbent on the project manager to identify the resources that are available and any resource-induced dependencies that may result. Similarly, the project manager must be familiar with the processes required to complete the project so that process-induced dependencies may be correctly reflected in the layout of activities.

1.3.2 CPM Charts: Basic Concepts

After identifying the various events and dividing the project into as many discrete activities as possible, a project manager may begin to determine how the project will "fit together." To accomplish this, it is generally helpful to make a graphical layout of the various events and activities. Such a layout (known as an *activity network*) is the first step in the creation of a CPM chart[14]. By one popular convention, events are depicted as boxes with rounded corners and activities are depicted as boxes with squared corners, as shown in Figure 1.3.1.

Consider the CPM chart for building a brick wall as shown in Figure 1.3.2. The chart is arranged roughly according to how it is anticipated that the work flow will go, with the left side of the chart representing the beginning of the project. The actual placement of the boxes is not critical and can be adjusted later if needed for pictorial clarity. Each of the activities are still quite general at this stage, and could be broken down further if desired. Next, each activity is assigned a duration time and a resource. In this case, assume that two people (Jane and Mike) are available to do the work. The assignment of these resources is the job of the project manager, and the activities' boxes are annotated with the appropriate names (lower right) and estimated activity duration (upper right). The activity boxes are then connected with lines to indicate their order of performance. Note that Jane and Mike can work simultaneously hauling bricks and preparing the mixer, but the placement of bricks and concrete cannot be started until the bricks have been hauled in, the mixer prepared, and the concrete mixed.

The *critical path* is the sequence of events upon which the total duration of the project is dependent. In this example, the critical path is clear: hauling the bricks to the site (three hours), mixing the cement (1 hour), and placing them and the cement (4 hours) determines how long the project will take (total of 8 hours). "Preparing the mixer" is not on the critical path as long as its duration is less than 3 hours. The critical path is indicated on the CPM chart by using thicker lines.

Each element of the CPM chart also has two other times annotated. *Early start* (upper left corner of the symbol) is the earliest time at which an event or activity can occur. The total time

[14] Both PERT and CPM are effective tools for controlling projects; to a large degree the determination of which to use is a matter of personal preference. In this book we utilize the CPM method. CPM charts can be presented in either the "i – j node" or "precedence network" format; we utilize the latter due to its simpler appearance. The fundamental concepts are the same among all methods.

Figure 1.3.1 CPM chart symbols.

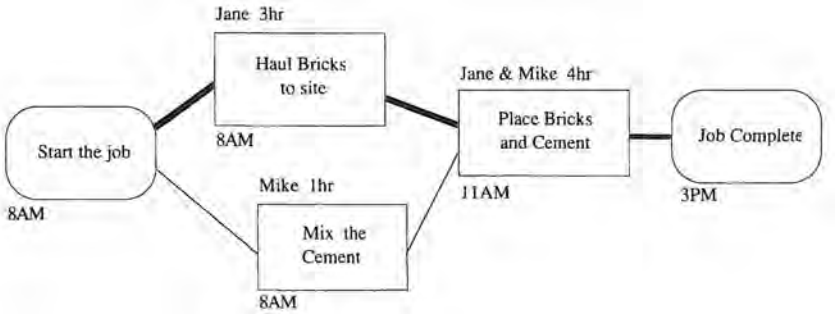


Figure 1.3.2 CPM chart for building a brick wall.

for the project is obtained by beginning at the planned start time of 8 AM, and adding up the time durations of the activities on the critical path. Thus this project should be complete at 3 PM. Late start (lower left corner of the symbol) is the latest time that an activity can begin without causing the completion time to slip. In other words, if the activity does not begin by the time shown as "late start," the critical path will change, and the project will take longer. It should be clear that activities on the critical path have identical early start and late start times, since by definition any delay in activities on the critical path lengthens the entire project. The underscoring of early start times indicates that these are fixed; in this example, it is a "given" that the project will in fact start at 8 AM.

Float is defined to be the extra time available during the project which will not cause a delay in the final completion time if it is used. Float for any activity is calculated as the difference between early start and late start times. In the example, the activity of preparing the mixer has a float of 2 hours, since the activity of preparing the mixer can increase by 2 hours without affecting the overall completion time. Activities that are on the critical path never have float, since by the definition of critical path any lengthening of the time required by those activities will delay the completion of the project.

The CPM chart shows the project manager when the job should be completed and that his initial work plan will result in Mike sitting idle for 2 hours while Jane completes the hauling of the bricks. To minimize costs, an astute project manager will most likely assign Mike to help Jane haul bricks, or find other work for him to do. In a more complex project, extensive "what if" scenarios would be examined. For example, what if a cement truck was called in at the appropriate time and Mike and Jane both hauled bricks?

This example is obviously very simplistic, and it is easy to calculate various early/late start and float times. Visualize, however, a project of developing a new machine. This would involve perhaps dozens of people and hundreds of activities. In practice, calculation of the critical path and experimenting with various "what if" options for such a project is possible only with a computer. As mentioned previously, there are several excellent project management software packages available for use on personal computers. Most of these packages will also calculate project costs, track costs by time and activity (yielding cash flow forecasts), level resources (i.e., assist with fitting the work-load to the available resources), and adjust schedules for long workdays, holidays, and weekends. The astute design engineer in charge of a complex design will become familiar with and make use of a project management software package to the fullest extent possible.

In many instances, the critical path method provides an effective tool for a project manager in justifying time duration estimates, costs, and so on, which may not be clear to those unfamiliar

with the mechanics of how a project is performed. Since the job of most project managers involves coordinating various groups, this type of qualitative supporting information can be very valuable.

Example 1

Prudhoe Bay oil field is located on the North Slope of Alaska. It is the largest U.S. oil field and produces approximately 1.5 million barrels of oil daily. Large quantities of natural gas also come out of the ground with the oil and because there is no pipeline to carry this gas to market, it must be reinjected into the ground. This is performed by 13 General Electric Frame V turbines, located at the central compressor plant (CCP). These turbines, which until recently together comprised the largest assembled collection of operating turbine horsepower in the world, generate tremendous amounts of heat. Through a variety of compressor operating parameter changes and initial design underestimates of heat output, the buildings in which the compressors are housed become extremely hot, 49 °C (120 °F), in the ceiling areas, even when the outside temperature is −17 °C (−30 °F). The assignment given to the project manager was to add ventilation to lower the inside temperature to a reasonable level. As with many large companies, there were a variety of departments and personnel that had areas of authority which included the CCP. As part of his job, the project manager was required to coordinate the interaction of the various groups, drawing on their various areas of expertise so the project could be accomplished. The managers of the CCP felt that the work could be completed within 4 months. After all, they said, "All we really need is to add a few big fans, and how hard can that be?"

As a first step, the project manager had to establish a rough estimate of the schedule; his "gut feel" was that given the normal lead times and organizational review cycles of the company, the project would take at least a year. Proceeding in the manner of the preceding section, he did a rough layout of the activities that were required to accomplish the project, and assigned time durations which reflected what he had experienced to be the company's "normal way of doing things." The CPM chart that resulted from this effort (too large to reproduce here) validated his original "guestimate" of about a year. Recognizing that the other interested parties would not be satisfied with the long time to get the compressor building cooling equipment installed, the Project Manager devised another CPM, denoted as the *best-case* approach. In general, the best-case assumed: 1) that the individuals involved would be able to dedicate themselves to acting faster on this project's activities than they were generally accustomed to doing, and 2) that the appropriate authority within the company would allow the *fast-track* design/construct method[15] to be used for portions of the job.

The best-case method allowed the job to be completed in just 8 months. Many people have a tendency to make commitments that they may not live up to later, and the project manager would be held accountable if he began a project that later failed to meet its schedule. Hence the project manager met with all of the involved groups and individuals. His message was that everyone would have to agree to perform their activities in the time indicated on the best-case CPM chart in order to have the project completed in an accelerated time frame. After reviewing the best-case schedule, if anyone felt that they could not meet the time allowed for their activities, they should speak up. If there were no objections to the times allotted, the individuals responsible for a particular activity were to initial the activity on the CPM chart. A few individuals requested time extensions for some activities, and these were accommodated. The CPM chart became the road map of the project. As the project progressed various individuals had to be reminded occasionally that their activity was falling behind schedule, and that if they did not deliver as promised the project's completion date would slip. The exact amount of slippage could readily be demonstrated by entering the latest projected activity duration into the CPM chart and determining a new finish date.

At various stages of the project the project manager could also input the current date as a given event and the CPM chart would then give a revised completion date. For example, some materials were lost in transit for about 2 weeks. When they finally arrived, that fact was entered as an event with a "given" (i.e., underlined) date, and within seconds, the impact on the activities which were to follow was determined. The project manager then informed the parties whose activities were affected that the early start time for their activity had slipped. The CPM chart thus acts as a plan and update tool to help manage a project.

[15] *Fast track* is a method whereby procurement/construction is begun prior to completion of the detailed engineering. While there is a risk that rework of already constructed areas may be needed due to design changes occurring in the late stages of engineering, the fast-track method provides for rapid execution of projects in an organized manner.

On other occasions, individuals were faced with other last-minute projects that also required their attention and they would ask the project manager how much extra time they could have to perform their activities. If the activity in question was not on the critical path, the CPM chart would quickly identify the float time for a particular activity. This allowed the project manager to allocate some additional time to the performance of an activity without affecting the project's overall completion date.

This example illustrates many of the practical applications of utilizing CPM techniques, specifically:

- A CPM chart provides a means to demonstrate that the initial perceptions of the time required to accomplish a project are often incorrect. A CPM chart can help to focus attention on potential problem areas and allow for the generation of a time estimate agreed upon by team members.
- A CPM chart provides a method of ensuring that individuals or departments keep their commitments. Since they agree to a proposed schedule at the beginning of a project, they later have little choice but to perform as promised.
- A CPM chart provides a means to identify where activities could be rescheduled using available float time. This allows for the timely performance of activities related to other projects and is to the overall benefit of the company.
- A CPM chart provides a means to rapidly adjust the project schedule as inevitable scope changes, problems, and delays occur. This allows the project manager to keep others informed of the project's overall status, as well as the schedule status of their particular activities. It also acts as a common reference tool to all involved parties; this facilitates better communication and minimizes misunderstandings which could otherwise result.

It is important to remember that a CPM chart, like any plan, should be viewed as a dynamic system that is constantly subject to updating and revision. The chart resulting at a project's end can be compared to previous versions to determine "What went right/wrong?", "Why did costs go up/down?," and other pertinent questions. These techniques will aid with improving the overall project management process as well as the execution of the specific project at hand. Those who intend to become involved with virtually any aspect of practical engineering will at some point be faced with the responsibility for seeing that a project is accomplished on time and on budget. While project management is an art and science unto itself, some basic techniques such as those presented here can be used by virtually anyone to effectively control and manage their project, as opposed to having their project manage them.

1.3.3 Estimating: Methodology and Resources[16]

The validity of any project's schedule and budget depends heavily on the accuracy of the estimates upon which they were created. Determining correct values to be used in an estimate is generally an imprecise science that largely relies on the practical knowledge of those performing the estimate. Virtually all estimating manuals provide information on the man-hours and/or costs associated with performing a given activity in a given location with a given labor situation. For example, a popular construction industry estimating manual divides construction projects into many components such as erecting the frame, hanging the Sheetrock, taping the Sheetrock, texturing the wall, and so on. For each activity a required crew size is assumed. The manual lists the pay for each crew member and cost of materials based on nationwide surveys by the manual's editors. The manual also gives estimated installation time and costs for literally thousands of construction activities. When utilizing manuals such as these, an estimating engineer must become familiar with the assumed bases for the estimates that they present. For example, if the manual's data is based on utilizing union crews in a northern city, and the project being estimated is to be built with nonunion labor in a southern city, each estimated component of the project must be adjusted by an appropriate factor to produce an accurate estimate. Most estimating manuals describe their assumptions in an introductory section and give some practical examples demonstrating application of the data they contain. It is imperative

[16] This section is not intended to emphasize any single type of estimating method, but rather to illustrate some of the factors and concerns which must be taken into account when preparing an estimate for a project.

that a practicing engineer does not simply look up data and use it in the same fashion that data from other sources has been used without first determining the facts upon which the data is based. As with development of a project schedule, determination of cost and schedule estimates are assisted by first breaking the project into as many activities as possible. The cost and time duration of each activity is then calculated, and the resulting figures are added to provide the total project cost and schedule. Alternatively, the individual costs and times can simply be entered into a CPM chart.

Example

Consider again the previous project of building a brick wall. Assume that the wall to be built is straight, 30.5 m (100 ft) long, 1.2 m (4 ft) feet high, and 0.2 m (8 in.) (i.e., two bricks) thick, and is to be placed on an existing concrete foundation. The project is to be built in Birmingham, Alabama. What will its direct cost (i.e., cost of labor and materials, excluding allowances for profit and overhead) be? Utilizing a Means Estimating Manual one finds that the basis for estimating brick masonry can be either the surface area of the wall to be constructed or the number of bricks to be used. Because the wall's dimensions are known, making the estimate based on the former is more straightforward. The surface area of the wall is 36.6 m^2 (394 ft^2). Although the wall is two bricks thick, the manual indicates that the backup bricks can be applied faster than the face brick. Thus the "backup" and "face" portions of the wall must be estimated separately, and combined only at the end.

As shown in Table 1.3.1, the total direct cost is approximately $2941. The contractor would have to add to this amount profit and overhead. A person contemplating purchasing the installation of such a wall would also have to factor in allowances for these items. The contractor would most likely have precise data from other projects on which to base the overhead and profit margins. A prospective purchaser would need to rely on published averages such as those presented elsewhere in the manual.

	Daily Output	Time	Material		Labor	
	(ft^2)	(days)	Per ft^2	Total	Per ft^2	Total
Backup wall	240	1.67	$1.54	$616	$2.97	$1,188
Face wall	215	1.86	$1.54	$616	$3.32	$1,328
Subtotal		3.07		$1,232		$2,516
For Birmingham costs:				× 0.841		×0.757
Total				$1,036		$1,905

Table 1.3.1 Brick wall construction cost estimate.

In addition to being of value to project managers for the purposes of scheduling projects and planning budgets, estimating techniques are used widely in the economic evaluation of machines and processes. Thus an engineer working to develop a machine to, for instance, automate the preparation of pipe ends for welding must be familiar with the use of estimating techniques in order to determine the proposed machine's economic value.

Example

Consider the development of a machine to automate the in-field preparation of pipe ends for welding. Steel pipe that is to be joined together is commonly welded using a butt weld. To facilitate a proper weld, the end of each piece to be joined must be beveled. New pipe customarily is delivered to a job site in 7 to 14 m (20 to 40 ft) sticks, with each end properly beveled. As short pieces are cut from the stick, their ends must be prepared. This requires the welder to make an angle cut (the bevel) with a torch, and then use a grinder to form the flat end (the land). The problem faced by the engineer is to develop a machine which can do the activity faster and cheaper than it can be done by the welder. To calculate the cost of the human performing the operation, the engineer may consult a piping estimator's manual.

For example, what is the time required for a welder to field-prepare a 0.25 m (10 in.) diameter Schedule 40 pipe for a butt weld? Referring to the section in the estimating manual on "Flame Beveling Pipe for Welding," one finds that the beveling requires 0.51 hours. Experience shows that grinding the land requires about 5 minutes, or 0.08 hours. If the average cost of a welder is $28.50 per hour, including an allowance for payroll burdens, then

$$\text{cost per end prepared} = \$28.50/\text{hour} \times (0.51 + 0.08)\text{hours} = \$16.82$$

The proposed machine must be able to perform the work for less cost than the welder in order to be an economically feasible alternative. Keep in mind that the welder will still be required to set up and operate the machine; cost savings must come from a reduction in the overall time required to prepare each end. Other factors that may affect the cost of human labor are:

- *Crew size:* In most cases a welder is assisted by a helper; since this estimating manual does not allow for this, the $16.90 per end cost figure must be modified by the helper's average wage rate multiplied by 0.59 hours.
- *Working conditions:* The estimating manual's introduction states that the figures presented assume that the weather is "half good and half bad." If the machine being designed is targeted toward areas where the weather is 100% bad, the man-hour figures must be adjusted accordingly. For example, estimates involving winter fieldwork in Minnesota require the baseline hours to be multiplied by 2.5.

Machines are often not affected by adverse weather to the extent that humans are. A machine which reduces the time to perform the pipe end preparation by 0.25 hours may save a contractor utilizing a single welder on warm California days the following:

$$\text{savings} = (0.59 \text{ hours} - 0.25 \text{ hours}) \times \$28.50/\text{hour} = \$9.69$$

The same machine may save a contractor using a welder with a helper (wage: $21.55/hour) in the Minnesota winter:

$$\text{savings} = (0.59\text{h} - 0.25\text{h}) \times 2.5 \times \$28.50/\text{h} + (0.59\text{h} - 0.25\text{h}) \times 2.5 \times \$21.55/\text{h} = \$42.55$$

Thus an expensive machine may not have economic value to a company operating in southern California, but it may be an excellent tool for a company operating in Minnesota.

Precision and Project Scale

Most estimating manuals present information to several significant figures (i.e., 0.51 hour per bevel). The estimator must realize that estimates are, by definition, approximations. As such, one must use the information from estimating manuals as a guideline which must be adapted to the size of the project. For example, it may be reasonable to say that any given component of the foregoing small brick wall example can be estimated to the nearest dollar or so, but on an aggregate basis, it just is not possible to predict the project cost to that level of precision. The estimator must use judgment and say that for the brick wall example, the estimated cost is $2950 (or better yet, $3000) rather than the $2941 figure obtained by totaling the cost of each individual project component. The surest indication that a given estimate was prepared by a novice is extreme precision; the cost of adding an assembly line to a plant simply cannot be estimated to be $3,234,766.67! However, as with all types of numerical analysis, numbers should not be rounded off until the end of the calculation.

Types of Estimates

Estimates are usually made at various stages of a project, and their accuracy is affected by the amount of information known to the estimator at the given time. This allows managers to make decisions on an ongoing basis, without having to wait until a great deal of effort has been expended to identify all factors required to yield an accurate estimate. An estimate made at the very beginning of a project is usually known as the *conceptual* or *VROM (Very Rough Order of Magnitude) estimate*. It is used to decide whether a given project is even worth pursuing. Assuming that it is, further research and engineering efforts provide more data, which can then be used for a *preliminary* or *ROM (Rough Order of Magnitude) estimate.* This figure would be used in the same fashion, enabling managers to determine whether the project (or machine) is economically feasible. Again assuming that it is, further information allows greater accuracy, resulting in a *control estimate.* Finally, when virtually all factors have been pinned down, a *definitive estimate* is prepared. Some guidelines for the accuracy level of these estimates are given in Table 1.3.2.

It is very important that all parties understand which type of estimate is being discussed at any given time. Moreover, decision makers must combine the estimate's accuracy level with factors such as the project size and risk of major problems that could result should the estimate be at the extreme limits of its range. For example, a small machine design firm may be able to easily proceed with

Conceptual (VROM)	± 70%
Preliminary (ROM)	± 30%
Control	± 15%
Definitive	± 5%

Table 1.3.2 Types and accuracies of estimates.

development of a new drive belt system for a lathe if they determine that the company could easily absorb a loss resulting from the conceptual estimate of $25,000 being 60% too low. The situation (i.e., management decision whether to proceed to the next stage) may be different if the same firm was faced with a $3 million project where the same 60% error may result in the company being wiped out!

Estimating is a key element in virtually every project, since sooner or later somebody will always ask the questions "How long will it take?" and "How much will it cost?" It is generally not technically difficult to prepare an estimate, but a great deal of judgment and skill is involved. While engineers and scientists are sometimes uncomfortable with the imprecision involved, particularly in conceptual and preliminary estimates, they must accept that in estimating one is compelled to produce an answer using the best available information at hand. Users of the estimates should always be aware of "What is behind the estimate" and not take inappropriate risks.

1.4 THE DESIGN OF A DESIGN ENGINEER[17]

Most people take design for granted. For example, consider an automobile and think for a moment how you would go about designing it from the ground up. The number of branching decision paths and engineering skills required to design and manufacture an automobile may seem overwhelming. When the author first decided to pursue a career in design, he was immediately humbled when he really took a close look at things as complex as automobiles, CNC machines, spaceships, and tall buildings. Trying to comprehend how these complex things were designed and built gave him an appreciation for the capability of the human mind to break down even the most complex projects into workable tasks. He then realized that he was also a human and therefore he must also have this same ability. Once he came to this realization, it became simple for him to reduce design problems to summations of numerous conceptual designs. Each task could be broken up into smaller and smaller tasks. Economic feasibility could then be determined by constantly feeding information about the design as it evolves into economic and project management databases. The design process is a dynamic one where design options are generated and discarded until a working design is finally converged upon. Ultimately, the design must meet specifications for function, safety, reliability, cost, manufacturability, and marketability.

In order to design yourself into a design engineer, you must consider several aspects of the task including:

- What makes a good design engineer? How do you make the design help itself?
- How do you keep informed? How do you choose from all the options?
- How do you form a design methodology? How do you make the design safe?
- How do you develop conceptual designs?
- These questions are addressed in following parts of this section. Then as an example, a typical design plan for a machine tool is presented.

1.4.1 What Makes a Good Design Engineer?[18]

Is design an art form only to be practiced by those gifted with its talents, or is it a regimented discipline that can be learned? Virtually everything that humans do involves altering the environment around us, which is essentially what design is all about; thus every individual possesses the ability to

[17] *"I believe that our Heavenly Father invented man because he was disappointed in the monkey."* Mark Twain
[18] *"Hit hard, hit fast, hit often."* Admiral William "Bull" Halsey

design to some extent. While there are few Mozarts in the history of the world, there are numerous musicians who play his music and enable us all to enjoy it. Each person must identify the area in which he feels promise, and must do the best job he can. It is true there is only one queen bee in a hive, but without the workers even the queen cannot survive. In other words, be careful of your ego and always strive to improve your abilities.

It is very difficult to teach people how to become creative design engineers because everyone thinks differently. There are often no clear solutions to a problem. Historical knowledge can also often be a powerful tool to help demonstrate how creative ideas are formed; unfortunately, a discussion of the history of machine tool development is beyond the scope of this book.[19] Systematic methods of analysis and synthesis can be formulated to aid in development of ideas; however, these methods have often been blamed for stifling creativity. A good design engineer often uses systematic methods of analysis and synthesis in order to help evaluate wild and crazy conceptual ideas generated during the initial creative phase of problem solving.

How can creativity be stimulated and enhanced? Perhaps if this question could be answered with an equation, a computer program could be written that could design anything. Good design engineers usually think in terms of pictures instead of in terms of equations or IF THEN ELSE logic. Often, it seems as if daydreams are an inner manifestation of the creative urge within all individuals. The task of the design engineer, therefore, is to install enough reality into his or her memory to enable daydreams to produce useful solutions to real problems. One must also be able to keep a mental catalog of available building blocks and methods in which they can be manufactured and put together. The database must be kept open, so as to not preclude the development of new building blocks, while taking care to keep abreast of new technologies.

A design engineer must also become good at identifying problems. Once a problem is identified, it will usually yield to an unending barrage of creative thought and analysis. High-priced consultants do not necessarily solve detailed problems; they identify the problems for others to solve. Identifying a problem requires careful detective work. In addition to solving and identifying problems, the design engineer must also learn to identify what the customer really needs, which is not necessarily what the customer thinks that he or she needs. This requires interaction with marketing research groups, customers, and manufacturing personnel on a continuing personal basis.

To keep his or her mind tuned, a good design engineer must always ask: "How does that work?" and "Why does that catch my eye?" regarding everything he or she sees in daily life. This will help to develop a feel for the needs and wants of people and the ability to make a realistic assessment of what is technologically feasible. It will also help the design engineer to develop a feel for color, form, texture, and proportion. By being observant, patient, and optimistic, a design engineer will become aware of what people buy and use. If the design engineer notices fault with something, chances are that others do, too, and thus money could be made by correcting the fault.[20] *Opportunity only knocks for those who listen, and it is hard to hear knocking when the radio is turned up too loud.*[21]

In addition, although each engineer must understand the physics of operation of the machine, he or she must realize that the design process is itself a precision dynamic system. If each engineer understands the structure of the design process and what other members of the team have to do, he or she will be less likely to cause problems that adversely affect the project. Once attitudes such as "why should I bother with this detail because someone will catch it" become established, competitiveness is the next thing to be lost. *Design engineers must feel a personal love for their work and the work of the team.*

1.4.2 How to Keep Informed

Of critical importance to the design engineer are catalogs and trade magazines, and the "bingo cards" that are used to order manufacturers' product literature. When one considers that engineers spend

[19] See, for example, Chris Evans, Precision Engineering: An Evolutionary View, Cranfield Press, Cranfield, Bedford MK43 0AL, England [available from the American Society for Precision Engineering, Raleigh, NC (919) 839-8444]; D. Hawke, Nuts and Bolts of the Past: A History of American Technology, 1776-1860, Harper & Row, New York; and Wayne Moore, Foundations of Mechanical Accuracy, Moore Special Tool Co., 1800 Union Ave., Bridgeport, CT 06607, 1970.

[20] See, for example, François Burkhardt and Inez Franksen (eds.), Design: Dieter Rams, Gerhardt Verlag, Berlin, 1981.

[21] Take those mind-numbing headphones off and observe the world around you!

thousands of dollars on college educations, a few hundred dollars per year for subscriptions to trade magazines and catalogs is a justifiable expense. Manufacturing catalogs are also wonderful picture books for kids to look through and they also help mom or dad from reinventing wheels as well as stimulating a flow of new ideas. However, one must keep things in perspective and not develop *catalogitis*, which means "design is the gospel according to catalogs, and thou shalt not deviate from standard parts." The use of standard versus nonstandard parts must be considered carefully. Before specifying a part for production, a good design engineer will also check a listing of companies such as the Thomas Register for many different sources of a part and then carefully evaluate each manufacturer's products.

In addition to building up a catalog file, the design engineer should maintain a small library of reference books. The author has found that the following books contain many useful heuristic engineering rules:

1. T. Busch, Fundamentals of Dimensional Metrology, Delmar Publishers.
2. Machine Tool Specs, Huebner Publishing Co.
3. J. Lienhard, A Heat Transfer Textbook, Prentice Hall.
4. W. Moore, Foundations of Mechanical Accuracy, Moore Special Tool Co.
5. E. Oberg et al., Machinery's Handbook, Industrial Press.
6. E. P. Popov, Engineering Mechanics of Solids, Prentice Hall.
7. J. Shigley and C. Mischke, Standard Handbook of Machine Design, McGraw-Hill Book Co.
8. M. F. Spotts, Design of Machine Elements, Prentice Hall.
9. R. Steidel, Jr., An Introduction to Mechanical Vibrations, John Wiley & Sons.
10. M. Weck, Handbook of Machine Tools,John Wiley & Sons.
11. W. Woodson, Human Factors Design Handbook, McGraw-Hill Book Co..
12. The Principles and Techniques of Mechanical Guarding, Bulletin 197, U.S. Dept. of Labor Occupational Safety and Health Administration.
13. F. H. Rolt, Gauges and Fine Measurements, University Microfilms.
14. H. J. J. Braddick, The Physics of Experimental Methods, Chapman & Hall.
15. T. N. Whitehead, The Design and Use of Instruments and Accurate Mechanisms, Dover Publications.
16. W. Mendenhall, Statistics for the Engineering and Computer Sciences, MacMillan.

Societies exist for precision engineers.[22] Also, it is wise to keep a copy of the standards used that pertain to the products being designed.[23] One of several biographies on Clarence "Kelly" Johnson, who was instrumental in the design of many of the commercial and military aircraft produced by Lockheed Corp., should also be read. It is also suggested that the book Industrial Design, by Raymond Loewy, also be read by every aspiring design engineer. Loewy was perhaps the greatest product designer of the twentieth century. Many corporate logos that are familiar to us today (e.g., Exxon, Shell, Formica, BP, TWA, etc.) were designed by him. The aesthetic aspect of design is stressed because aesthetics are a necessary (but not sufficient) part of most designs.

Last, but not in any way the least, is the importance of networking. You must build a network of friends and associates who can help you to keep informed, and vice versa.

1.4.3 Formulating a Personal Design Methodology

One of the founding fathers of machine tools, Henry Maudslay who was born in 1771, is credited with the development of the compound slide, whose design principle is used on virtually every lathe in the world today. Although it is said that many of his inventions were described before by others in principle, it was Maudslay who reduced many ideas to practice. One of Maudslay's fundamental contributions was to note that the extra cost of making a machine from metal, as opposed to wood, was recouped many times over in terms of the machine's accuracy and life. Maudslay had several maxims which still serve as a set of basic guidelines for all types of designs:

[22] For example, the American Society for Precision Engineering in Raleigh, NC (919) 839-8444, and the Japan Society of Precision Engineering.

[23] For example, a designer of rotary tables or spindles should keep a copy of Axis of Rotation: Methods for Specifying and Testing, ANSI Standard B89.3.4M-1985, and Temperature and Humidity Environment for Dimensional Measurement, ANSI Standard B89.6.2-1973. A catalog of standards is available from the American Society of Mechanical Engineers, United Engineering Center, 345 East 47th St., New York, NY 10017.

l. Get a clear notion of what you desire to accomplish, then you will probably get it.

2. Keep a sharp look-out upon your materials: get rid of every pound of material you can do without. Put yourself to the question, "what business has it there?" Avoid complexities and make everything as simple as possible.

3. Remember the get-ability of parts.

Maudslay's maxims can be used as a good foundation for just about any personal design methodology.

Because many modern systems are often so complex, expertise is required in many different disciplines; hence often it is nearly impossible for one person alone to design an entire system. However, it is possible to be aware of the capabilities of other disciplines. This allows an individual or small group to develop a design plan for a complex system. Although it is often said that *"no matter what you design, somebody has already thought of it before, at least in principle,"* this should not be accepted as a defeatist attitude but rather one of awareness. Continual updating of one's mental memory banks with new knowledge about advances in all fields of engineering and science is a must if a design engineer is to remain competitive. This updating must be the cornerstone of every design engineer's personal design methodology. Other than that, every design engineer has his or her own way of doing things, many aspects of which are borrowed from established methods. Some of these methods are discussed below.

Designs can be categorized as being *original*, *adaptive*, or *scaled*. Original design means developing a new way of doing something (e.g., cutting with waterjets, as opposed to using a saw blade). Adaptive design means using technology developed for another task and adapting it to perform the task at hand (e.g., using lasers to sculpt wood). Scaled design means changing the size or arrangement of a design in order to accommodate a similar change in an existing process (e.g., design a bigger version of an existing machine). Each of these types of design can be equally challenging and all require five basic steps:

1. Task definition
2. Conceptual design
3. Layout design
4. Detail design
5. Design follow-up

Task definition often starts with the customer or sales representative requesting the design department to provide a study regarding the feasibility, cost, and potential availability of a design to perform a specific function. In response to this request, the company's best design engineers get together to sketch out concepts. It is in the *conceptual design* phase that the functional relationships of components and the physical structure are usually defined. Once a few select conceptual designs are chosen, they are expanded in detail through *layout design* where preliminary sizing of components and calculations are made in order to produce rough assembly drawings of the conceptual designs. This enables more accurate feasibility and cost estimates to be developed. After modifying the required specifications and conceptual designs, the project's feasibility can be determined, usually resulting in one design being chosen for detailing. The *detail design* phase is everything that follows in order to bring the design to life. *Design follow-up* involves activities such as the development of a maintenance plan and documentation, which often causes many design engineers to run and hide. However, if the design is not maintained, or if nobody can figure out how to use it, the design will not be used and design effort will have been wasted.

Along each step of the design path, design engineers have to apply their own personal design methodology, which they must develop themselves. Whatever form the methodology takes, it should realize that no design engineer is an island.[24] In general, the method should:

1. *Foster creativity*. The design engineer should always start with wild, crazy, "what if" designs and if necessary scale back to more rational conventional solutions. The design engineer, however, must know when to turn off the wild, crazy, dreaming aspect often associated with generating conceptual designs, and proceed with a systematic consideration of one or two concepts that will lead to the detail design.

[24] *"So long as the mother, Ignorance, lives, it is not safe for Science, the offspring, to divulge the hidden causes of things."* Johannes Kepler

2. *Acknowledge the creativeness of others*. A *not invented here* (NIH) syndrome is unacceptable and has been the downfall of many a firm that was unwilling to adopt an outsider's superior concept.[25] There is no room for prejudice in design. One must take what exists, use it to its fullest potential, and then improve upon it.
3. *Do not depend on luck or ignore a problem in the hope that it will go away*. A wishful attitude has killed many people and generated huge legal fees. Every detail from where to run electric lines and hydraulic hoses, to placement of the warning labels and nameplates, must be carefully considered.
4. *Be disciplined and well organized so the design can be passed onto others for detailing or completion*. This requires knowing how to delegate authority to optimize utilization of an organization's resources.
5. *Respect simplicity and the fundamental knowledge of how and why things work*. This will hasten the convergence process for a design and will help prevent oversights such as placing a measuring element far away from the process to be measured (Abbe error). It will also lead to the minimizing of design cost, manufacturing cost, functional errors, and embarrassment.
6. *Continually subject designs to value analysis in an effort to reduce cost with an equal or increased level of quality*. Not only must the design be subject to value analysis, but the manufacturing and sales must be considered as integral parts of a successful design process. Thus the design engineer must also be knowledgeable in production and marketing skills.

When developing your own personal design methodology, also consider some of the following common methods used by design engineers to develop solutions to design problems:

1. *Persistent questioning*. By always asking *"Why?"* and *"Can it be made simpler and better?"* you will be less likely to settle for less than best, or to overlook a possible improvement.
2. *Known solutions*. By analyzing known solutions to existing similar problems, one can often find a wheel that exists without having to reinvent it. This method also includes systematic analysis of variations of known solutions.
3. *Forward chaining*. Start with a sketch of the problem depicting what you hope to accomplish and then form an expanding tree of ideas.
4. *Backward chaining*. Start with what you know the complete design must look like and then trace back through all the elements that lead to the final design and scale and modify them accordingly. This procedure is usually used to develop manufacturing or process plans.

The array of products that are designed in the world is so varied and complex that it is nearly impossible to list a comprehensive generic design plan that incorporates these concepts for a generic product.

1.4.4 Developing Conceptual Designs

A really good design engineer is differentiated from part detailers by the former's ability to generate conceptual designs from which final designs evolve. A good design engineer also has the patience to consider that nothing is of secondary importance because sooner or later a user will be exposed to each part of the machine. An invaluable exercise to help you become better at generating conceptual ideas and developing an eye for detail is thus to adapt a philosophy "How does that work?" or "Why does that catch my eye," and apply it to everything you see in daily life. This mental exercise will help to develop your design abilities and build up a database of idea building blocks and new technologies.

In addition to observing the world around you, experiment with ideas by having plenty of cardboard, modeling clay, Styrofoam, foam rubber, paper clips, and Popsicle sticks on hand. Cardboard models can do wonders to help visualize a project. In the age of computers, solid modeling

[25] "*Self-conceit may lead to self-destruction*." Aesop

software programs are supposed to replace these crude tools, but a computer is often not available when a brainstorm strikes. Solid modelers and other computer tools, however, can be invaluable once the more detailed phase of developing a conceptual design begins. In addition, there are several methods which can be followed in order to develop good conceptual designs.

Adaptation of Existing Systems

One should always look to see if something proven already exists which can be adapted to solve the problem at hand. This includes suggestions for solutions to standard problems often given by vendors in their product catalogs.

Systematic Analysis of the Process

If an equation can be written to describe the process, then mechanical components representing variables in the equation can often be used to produce the desired effect. This represents what is often known as a "cookbook solution." It is also one of the most desirable solutions because in the cases where it can be applied, it almost always works.[26] Back-of-the-envelope calculations are often used to estimate the feasibility of an idea. The best designers are often those with good back-of-the-envelope skills.

Systematic analysis of the process is also extremely valuable to provide a starting point for more detailed numerical calculations. A common error made by novice design engineers is the belief that they do not need to know analysis because computers will do it all for them. This type of thinking is totally unacceptable and has no basis in reality. Computers are good for checking assumptions and highlighting problem areas, but you must have a good idea about where to start the analysis process. The better your initial estimate, the less computer-assisted iterations you will have to make.

Consider estimates of machine tool accuracy. Assume someone tells you that they can consistently provide you with 0.2 m (8 in.) long aluminum parts accurate to 10 μm (400 μin.), and you notice that the sun is shining through a window in his drafty shop with no heat in the winter. You know that your part will expand at the rate of 20 μm/C^{o} and thus over a temperature change of 10 C^{o} in the guy's shop, a minimum thermal error of 40 μm will be present in the parts. You should think twice about buying parts from this person.

Nature as a Model

Using solutions found in nature is part of the process of being aware of your surroundings, as opposed to being aware only of loud music in your ears. Consider the structure of bamboo, bird wings, human bone, oak trees, marsh reeds, and burrs. Bamboo and bird wings are essentially honeycombed structures that have served as the inspiration for the structure of all modern aircraft frames. Human bone is a unique structure with compliance at the ends to cushion shock at the point of relative motion, yet bones have great strength and stiffness along their length. Oak trees have great strength and grow to enormous size, yet they are often uprooted in violent storms because of their lack of compliance. On the other hand, marsh reeds, which often grow side by side with oak trees at the edge of a forest, bend with the wind and are not blown over.[27] Similarly, the safest cars are those designed to yield on impact. Velcro was invented by an outdoorsman who noticed, upon close examination of a burr, tiny barbed hooks that cling again and again to fuzzy surfaces. These examples illustrate why it is better to stay tuned to the world than to a set of portable headphones.

Model Tests

One of the world's foremost aircraft design engineers, Kelly Johnson, was once discussing the recent development of a prototype forward-swept-wing fighter aircraft. He said that the problem with the design is that it took 12 years of computer modeling to design it. He compared this to the 2 years it took him and his crew at Lockheed Corp.'s "Skunkworks" to design the SR-71 Blackbird reconnaissance aircraft in the early 1960s. The SR-71 is still one of the fastest, highest flying, most advanced planes in the world. The SR-71 design task was similar to that of the forward-swept-wing fighter in that the SR-71 had to have all new engine, frame, skin, and control technology developed to withstand the high altitude, speed, and skin temperatures that it would be subjected to. The SR-71 was designed in minimal time using a balanced blend of analysis and model testing. One should always keep an open mind and gather data in the quickest, most efficient, and most accurate method

[26] This method is illustrated many times in this book.

[27] "The Oak Tree and the Reed" by Aesop. Everyone should read Aesop's fables.

possible. Often the best results are obtained from a balanced mix of analysis and experimentation. Do not let yourself be bullied into not using modeling clay because it is not modern.

Brainstorming

Brainstorming is the method most often thought of for generating conceptual designs. It essentially has no rules other than to keep a free and open mind while generating as many ideas as possible without anyone criticizing them until all ideas are heard. An additional benefit of brainstorming is persons involved become exposed to other design ideas from individuals skilled in other areas of technology. As proposed by Pahl and Beitz,[28] the brainstorming group should have the following composition:

1. The group should have a minimum of 5 and a maximum of 15 members. Too few members produce too few new views and opinions. Too many members form a mob that segments itself into small groups anyway.
2. The group should have experts from a variety of fields to ensure that ideas from all areas of technology will be represented. It should also have a few laypeople involved. Often, an inexperienced nontechnical person will ask a question which causes experts to stop and think.
3. If possible, the boss should not be included in the group. This will help avoid wasting time by some members who seek to speak merely to gain favor with the boss. On the other hand, in many situations the boss is the top expert in the field; thus the personality of the members needs to be considered.
4. A leader should be appointed whose primary function is to organize and keep order. The leader should take notes to ensure that subtle technical comments are not accidentally missed by a nontechnical secretary. In addition, the leader must act to discipline the group if it wanders too far afield (e.g., discussing baseball scores).

Pahl and Beitz also propose that the session have the following format:

1. All participants must speak with free and open minds without fear of being subject to harassment, or fear that someone will think their idea is silly. All thoughts must be complete enough to make sure that everyone understands them.
2. Criticism of ideas must be polite, kept to a minimum, and be constructive. This will allow promising ideas to evolve. Remember, ideas which may seem irrelevant may stimulate another group member to generate a plausible idea.
3. All ideas and discussions must be recorded, to and allow each member to reconstruct the session at a later date.
4. The session should not last more than 30 to 45 minutes. People tend to become less productive after longer periods.

No matter what format is used, it is vital to cast off inhibitions and let the creative juices flow.

Rohrbach's 635 Method

This method requires six people to sketch (on paper) three possible solutions to the problem. Each person's suggestions are read by the other five, who must provide three comments about each suggestion. This method provides a more systematic development of an idea, and it is easier to tell who solved the problem. Also, personality conflicts arising in an open and verbal atmosphere are avoided. This method, however, does not allow for the creative cross-pollination of ideas that often occurs in a group setting. Thus it is often wise to implement Method 635 before a brainstorming session.

1.4.5 Self Principles

Self principles utilize the phenomena the machine is trying to control to help control the phenomena. There are four basic types of self principles: self-help, self-balancing, self-protecting, and self-checking.

[28] G. Pahl and W. Beitz, Konstruktionslehre, Springer-Verlag, Berlin, 1977. Translated and published as Engineering Design, Design Council, London, 1984, pp. 87–88.

Self-help systems use the phenomena to be controlled as a means for controlling the phenomena. For example, some airplane door mechanisms swing the door out when opened, yet the door acts as a self-sealing tapered plug when closed. The modern tubeless tire is a result of careful structural design that utilizes the elastic nature of the tire to make a seal that allows the tire to be inflated. Once inflated, the internal pressure keeps the tire's lips pressed up against the tire rim and a seal is maintained.

Self-balancing systems utilize geometry to reduce an undesirable effect such as high stress. An example is the leaning of turbine blades to offset centrifugal forces. In order to utilize this principle, the design engineer would find the forces and stresses on the blade and calculate the ideal angle for the blade. As the turbine spins, the shape of the blade pumps the air through the turbine and creates large bending stresses on the blade. If the blade is tilted so that its length axis is not along a radius of the turbine, centrifugal forces can be used to create a bending moment opposite to that caused by the airflow. In this manner, cyclic tensile stresses can be minimized.

Self-protecting systems utilize passive elements to ensure that the system is not overloaded to the point of permanent deformation. They do this by providing additional force transmission paths after a predetermined elastic deformation occurs. However, in protecting one component the design engineer must make sure that the effect does not ripple through the system and damage other components. Usually, self-protected systems can return to their normal operating range unaffected by overloads. The most common example of self-protecting systems are found in spring-actuated mechanisms. For extension springs, a chain or cable can be used to prevent the spring from being overextended. When the spring is compressed, the chain or cable is slack. When springs are compressed they can be made so that the coils or disks collapse against each other, or so the mechanism contacts a hard stop. If the helix angle of a linear spring varies with the length, then the coils near the top will make contact first and a hardening effect will gradually occur and increase after the normal linear operating range has been exceeded.

Self-checking systems utilize the concept of symmetry to ensure geometric accuracy. This involves turning the device through 180^{o} and comparing the measurements to the first set. For example, if a coordinate measuring machine (CMM) is very repeatable but not accurate, it can still be used to measure the straightness of a straightedge. The edge of the straightedge is placed vertical so that gravity does not create an error. The straightedge is measured with the CMM and then the straightedge is rotated 180^{o} and measured again. The measurements are subtracted from each other, which cancels out the straightness error in the CMM. Related to self-checking, run-in of mechanical systems and burn-in of electronic systems for a predetermined period can be used to verify that the system will meet specifications for its intended life. However, these processes are time consuming and often require the full prototype product to be tested using actual load cycling conditions. On the other hand, system components such as spindles can be checked for their integrity, not necessarily that of the entire design, by running them on test stands prior to being installed on the machine.

1.4.6 How to Choose from Among All the Design Possibilities

When faced with a myriad of design alternatives, how can the design engineer be assured that he or she is making the best choice? Often one overriding factor, such as a requirement for zero friction in a bearing, will eliminate other choices, but in many instances the choice is not so clear. There are many systematic methods available for evaluating design alternatives, and most rely on some sort of weighting of attributes to arrive at a *desirability score*.

The simplest method is a linear weighting scheme that applies a desirability value to each parameter that affects the performance of a component in a design. When there are only a few design alternatives to consider, this method is the easiest to use provided user bias can be minimized. Consider the selection of a linear bearing for a linear machine tool axis. Assume that there are three bearing options, sliding, rolling, and hydrostatic. Factors to consider in their selection include:

- Accuracy of motion: straightness, smoothness (high-frequency straightness errors)
- Friction characteristics: static friction, dynamic friction
- Cost: purchase, install, maintain

If these parameters are evaluated and given a desirability score, then upon summing the scores, we would in effect be comparing dissimilar values. Even for this simple example, a design engineer who

lacks experience can have difficulty correctly assigning desirability weights, because when assigning weights to one variable, he or she must do so in the context of all other system variables. When there are more than several variables, it is often difficult to evaluate their relative importance. Thus a method is needed to allow the design engineer to decompose the decision problem into discrete manageable components.

Instead of trying to compare all factors at once, it would make more sense first to determine the relative importance (priority) of each characteristic (e.g., accuracy versus friction, accuracy versus cost) at each level in the outline of design attributes and then evaluate the relative characteristics of each linear bearing with respect to the most explicit characteristic (e.g., straightness, smoothness, static friction, and so on). This type of decision analysis is called the *Analytic Hierarchy Process* (AHP) and was developed by Thomas L. Saaty.[29] The AHP combines the two fundamental approaches that humans have developed for analysis of complex systems, the deductive approach[30] and the systems approach.[31] The AHP method enables a person to structure a system and its environment into mutually interacting parts and then to evaluate their relative importance by measuring and ranking the impact of these parts on the entire system. This structured approach to decision making eliminates much of the guesswork and confusion of the common method of synthesizing an overall explanation for choices made in a system from piecemeal explanations arrived at through deduction.

The AHP method is particularly valuable when dealing with complex, unstructured problems whose priorities need to be ordered, and where compromises need to be made to serve the greatest common interest. It is often difficult to agree on which objective outweighs another, particularly in complex issues where a wide margin of error is possible when making design tradeoffs. Intuitive thought processes that work well in the familiar routine of daily life can be misleading when applied to complicated matters where sources of information and opinions are varied. The AHP method structures complex problems in an organized manner that allows for interaction and interdependence among factors and still enables one to think about the factors in a simple way. It also allows the sensitivity of any one parameter on the final answer to be tested merely by varying that one parameter (holding others constant) using a computer simulation.

Applying the AHP

The AHP method is a method of breaking down a complex, unstructured situation into its component parts, arranging these parts in a hierarchical order; assigning numerical values to subjective judgments on the relative importance of each part, and synthesizing the judgment to determine which parts have the highest priority and should be acted upon to influence the outcome of the situation. The subjective judgment on the relative importance of AHP components can be provided by a single design engineer, or better yet, by a group of design engineers.

The assessment of component importance with the AHP method involves three distinct steps: *Step 1*: Setting up the AHP model and determining the relative importance or priority (weight) of its components on a level-by-level basis. *Step 2*: Evaluating the relative importance of design choices (i.e., component types) with respect to the lowest (most elemental) entries in the model. *Step 3*: Computing the desirability of design choices using the results of steps 1 and 2. To illustrate the AHP method, a brief outline of the procedures involved in steps 1 through 3 is presented for a hypothetical model. A detailed example is then presented which shows how the calculations are made.

Step 1: Setting Up the AHP Model

Perhaps the easiest way to follow the construction of a hypothetical AHP for bearing selection is to compare its construction to a system of water tanks as shown in Figure 1.4.1. The distribution of a given volume of water in the reservoir system is analogous to the distribution of priorities (or weights) in our model.[32] Consider the three-level reservoir system of Figure 1.4.1 with the main reservoir located on level 1 and auxiliary reservoirs on levels 2 and 3. The tanks are connected by

[29] T. L. Saaty and J. M. Alexander, Thinking with Models, Pergamon Press, Elmsford, NY,1981.; and T. L. Saaty The Analytic Hierarchy Process, McGraw-Hill, New York, 1980.

[30] Deductive approach: A complex system is broken down into components and structured into a network. Explanations are then obtained for the network components (inputs). These inputs are synthesized to provide an explanation for the whole network. There are, however, no rules of logic for combining (synthesizing) these inputs.

[31] Systems approach: Analysis of a system is done by examining it from a general perspective that does not give much attention to the functioning of its parts.

[32] Dr. Peaslee of the Los Alamos National Laboratories originated the idea of comparing the distribution of weights to the distribution of water.

various-size pipes which determine the distribution of water as it flows from the highest to the lowest level. The desirability or weight for each entry on any particular level with respect to other entries on that level can be equated to the distribution (by percent) of the initial volume of water among the reservoirs on the same level. Note that at any level the total percentage of water distributed among the reservoirs at that level is equal to 100%.

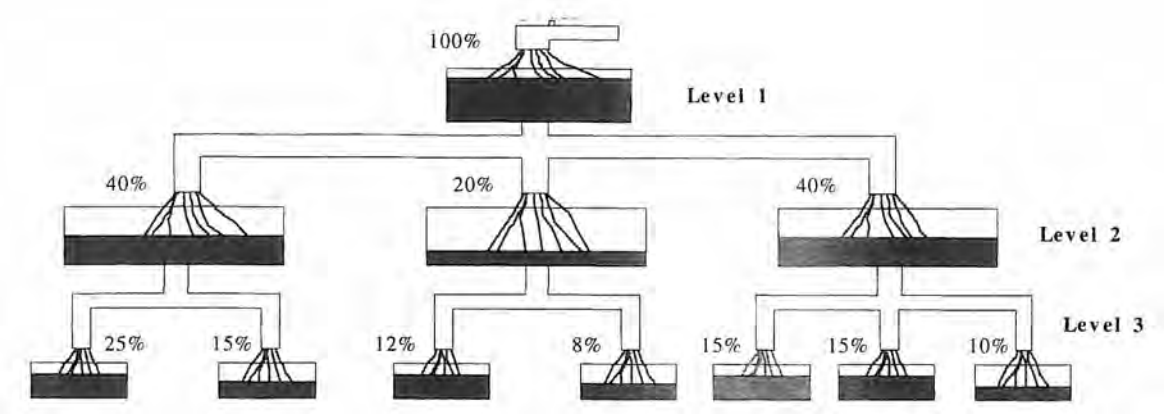


Figure 1.4.1 Visualizing the Analytic Hierarchy Process as the flow of water in a reservoir system.

Step 2: Evaluating Importance of Design Choices

The model set up in step 1 is now ready for use. The evaluation of a given design choice is done by determining its "importance" to each entry in level 3 (the lowest level). The "importance" of a perspective design, component, or characteristic with respect to each entry in Level 3 can be expressed in terms of actual physical values, or on a numerical scale, such as very important = 9, important = 6, marginal = 3, or equal = 1. Any scale or intermediate values can be used.

Step 3: Results

The desirability is calculated by combining the weights obtained in step 1 with the evaluations obtained in step 2.

Example

The model shown in Figure 1.4.2 will be used to determine the relative merits of three linear bearings LB1, LB2, and LB3 for a diamond turning machine design. The first step is to determine the relative importance between the elements in level 2 with respect to the element in level 1. To do this, assign a number to this importance (from 1 to 9) as shown in Figure 1.4.3, and give a rational for each determination (e.g., the importance of accuracy versus friction with regard to controlling motion of the tool point, the importance of repeatability versus cost, etc.)

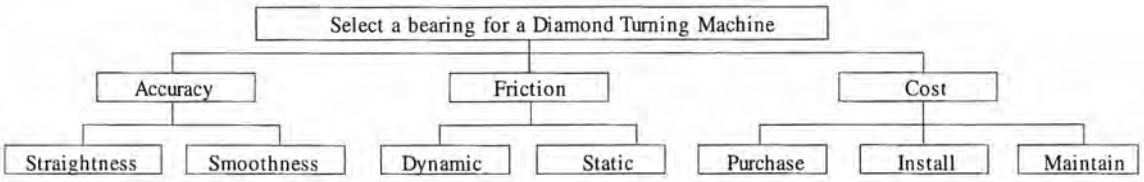


Figure 1.4.2 Applying the AHP to the selection of a precision bearing.

The relative importances of elements in level 2 with regard to the level 1 elements are then arranged in matrix form. In the first row, accuracy is considered slightly more important than friction as noted by the value 2. The relative priority between the elements is determined from the eigenvalue problem $A\omega = \lambda_{max}\omega$. For larger matrices this becomes cumbersome and a sufficient approximation is obtained from the geometric mean of the row, which gives the exact solution when the matrix is consistent. This value is then normalized by the sum of the priorities.

Importance	Definition
1	Equal importance
3	One a little more important than the other
5	One is more important than the other
7	One is much more important than the other
9	Absolute importance

Figure 1.4.3 Definition of importance ratings.

	Accuracy	Friction	Cost	Priority	N. Prior.				Level 2
Accuracy	1.000	2.000	4.000	2.000	0.571				
Friction	0.500	1.000	2.000	1.000	0.286				
Cost	0.250	0.500	1.000	0.500	0.143				
Sum	1.750	3.500	7.000	3.500	1.000				
Accuracy	Straight.	Smooth.	Priority	N. Prior.	N. Wght				Level 3
Straightness	1.000	0.500	0.707	0.333	0.190				
Smoothness	2.000	1.000	1.414	0.667	0.381				
Sum	3.000	1.500	2.121	1.000	0.571				
Friction	Static	Dynamic	Priority	N. Prior.	N. Wght				Level 3
Static	1.000	4.000	2.000	0.800	0.229				
Dynamic	0.250	1.000	0.500	0.200	0.057				
Sum	1.250	5.000	2.500	1.000	0.286				
Cost	Purchase	Install	Maintain	Priority	N. Prior.	N. Wght			Level 3
Purchase	1.000	0.500	0.500	0.630	0.200	0.029			
Install	2.000	1.000	1.000	1.260	0.400	0.057			
Maintain	2.000	1.000	1.000	1.260	0.400	0.057			
Sum	5.000	2.500	2.500	3.150	1.000	0.143			
Bearing comparison:	Straight.	Smooth.	Static mu	Dyn. mu	Purchase	Install	Maintain	Overall Goodness	Norm. Overall Goodness
Sliding teflon	7.000	7.000	1.000	2.000	5.000	3.000	5.000	4.943	0.29
Rolling balls	6.000	5.000	3.000	5.000	3.000	5.000	5.000	4.676	0.27
Hydrostatic	9.000	9.000	9.000	5.000	1.000	2.000	1.000	7.686	0.44

Figure 1.4.4 Spreadsheet output for bearing selection using the AHP.

The level 3 components' normalized relative priorities are determined in a similar manner. These normalized priorities are then multiplied by their respective level 2 category's normalized priorities. The resultant normalized weights represent the relative importance of all of the items being considered. These normalized weights are used to weight the relative comparison that is then done between items being considered to solve the level 1 problem.

For the example given here, the next step is to determine the relative desirability of the three linear bearing systems versus each element in level 3. The bearing comparison values in the matrix can be the inverse of actual values that have been normalized with respect to the column sum, or the scale presented in Figure 1.4.3 can be used. For this example, the latter was chosen.

The relative desirability of the three linear bearings LB1, LB2, and LB3 is the sum of the products of the level 3 weights:

$$\text{Normalized desirability} = \frac{\sum_{j=1}^{7} (\text{level 3 normalized weights}) \times (\text{bearing desirability values})}{\sum_{i=1}^{3} \text{bearing desirability values}} \tag{1.4.1}$$

As shown in Figure 1.4.4, the normalized desirabilities are 4.9, 4.7, and 7.7, respectively, so the best choice for this application is a hydrostatic bearing.

Note that as the size of the nxn matrix increases, the chance of inconsistency also increases. Consistency checks can be made on any of the matrices at any step in the process. First, the *consistency index* (C.I.) is determined. The sum of the values of the first column is multiplied by its respective normalized priority value. This is repeated for all of the columns. These values are then summed and they represent the largest eigenvalue of the matrix λ_{max}. The consistency index is given by:

$$\text{C.I.} = \frac{\lambda_{max} - n}{n - 1} \tag{1.4.2}$$

The consistency ratio compares the consistency of the comparisons made to the value that would have been obtained if the comparisons were random. Table 1.4.1 gives values of C.I._{random}. The

ratio $C.I._{comparison}/C.I._{random}$ should be less than 0.1 for the comparison to be acceptable. As shown in Figure 1.4.5, perfect consistency is obtained when the logical mathematical relations (e.g., $a_{jk} = a_{ik}/a_{ij}$) are imposed on the importance matrix.

n	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C.I.	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51	1.48	1.56	1.57	1.59	1.60

Table 1.4.1 Consistency Index for random matrices.

	B	C	D	E	F	G
6		Accuracy	Friction	Cost	Priority	N. Prior.
7	Accuracy	1	**2**	**4**	=(C7*D7*E7)^0.3333	=F7/$F$10
8	Friction	=1/D7	1	=E7/D7	=(C8*D8*E8)^0.3333	=F8/$F$10
9	Cost	=1/E7	=1/E8	1	=(C9*D9*E9)^0.3333	=F9/$F$10
10	Sum	=Sum(C7:C9)	=Sum(D7:D9)	=Sum(E7:E9)	=Sum(F7:F9)	

Figure 1.4.5 Spreadsheet cell formulas to ensure consistency.

It would have been possible for an experienced design engineer to make this decision, but the AHP allows an inexperienced design engineer to incrementally develop his experience. Furthermore, it allows for the systematic collection and analysis of evidence that can be presented to management or other members of the design team as justification for the decision.

1.4.7 Safety Considerations[33]

Designing a product to be safe is of prime importance. Unfortunately, some consider safety as an afterthought; however, if a product is not safe, it will quickly gain a bad reputation, sales will cease and people will be hurt. There are three principal methods that a design engineer can use to increase the safety of a design: *direct methods*, *indirect methods*, and *warnings*. Direct methods require elimination of the hazard. This usually means changing the process and is sometimes not feasible. Indirect methods involve the use of guards and shields to prevent operator injury or fouling of other machine components. Warning labels should only be considered as a method for drawing attention to the guards and the hazards that may result if the machine is misused.

When properly designed, safety systems protect the operator, the environment, other components in the system, and the integrity of the manufacturing process. Safety hazards which affect humans include mechanical systems (e.g., flying chips, sparks, high-pressure leaks, and rotating cutters), electrical systems (e.g., high-voltage shock hazards and EMI of pacemakers), acoustical sources (e.g. loud-noise-producing systems), optical sources (e.g., light from lasers, welding, and some photo-optical processes), and chemical and radioactive sources.

In addition to safety hazards resulting from the manufacturing process being performed by the machine, hazards can be created by failure of machine components (e.g., rotating parts). These can be prevented by *safe life*, *fail-safe*, and *redundant* design. Safe life design requires the part to be designed with infinite life even when subject to overloads and abuse. Fail-safe design is also known as the leak-before-break criterion, which means that the part gives ample warning via decreased performance before it breaks. Redundant design implies use of more parts than are necessary for operation of the system. With the loss of one part, decreased performance will be noticed (e.g., loss of power in one of several engines on an airplane), but the machine should be able to operate long enough to bring the system to a controlled stop. Parts whose failure could cause damage must often be designed to one of these criteria.

Designing for safety also requires the design engineer to think carefully through the operation of the machine. While searching for pinch points, uncovered processes, and the like, the design engineer is often able to spot other potential problems with the machine. As a final safety check, the design engineer must ask: "Would I be willing to operate the machine on a daily basis?"

[33] *"Always do right. This will gratify some people, and astonish the rest."* Mark Twain

1.4.8 Design Plan for a Machine Tool[34]

Assume that the head of marketing walks in and hands you a customer's specifications for a new machine tool and asks you to produce a design that will meet them. What should you do? The following tasks address this question:

1. *Define machine function and specifications.* The design engineer must determine what the customer really needs and then define a reasonable set of functional specifications that describe what the machine must do. This is accomplished by analyzing the parts and processes the machine is expected to execute. If the customer says that they require axis accelerations of 2g's but a review of their production requirements indicates that an acceleration of 0.2g's is sufficient, the discrepancy should be addressed. The design engineer should also look for ways of modifying the customer's specifications in order to simplify the design of the machine while checking for consistency and realism of the customer's specifications. It is the responsibility of the design engineer to take nothing for granted. The functional requirements for the machine must address the following considerations, which will eventually have to be considered in detail:

    a. *Geometry:* What are the approximate overall size and footprint?
    b. *Kinematics*: What type of mechanism is required, and what are its required repeatability, accuracy, and resolution?
    c. *Dynamics*: What forces are generated, and what are their potential effects on the system and its components? How rigid must the machine be to resist processing forces while maintaining surface finish and part accuracy?
    d. *Power requirements*: What types of actuators can be used, and what systems are needed to control the operating environment (e.g., air conditioning)?
    e. *Materials*: What types of materials can maximize performance of the machine? What are the properties of the materials the machine will be processing?
    f. *Sensors and control*: What types of sensor and control systems are needed? How can they be used to decrease the cost of required mechanical systems and increase reliability?
    g. *Safety*: What are the requirements for the protection of the operator, the environment, and the machine?
    h. *Ergonomics*: How can all design factors be combined to produce a machine that is a pleasure to operate, maintain, and repair?
    i. *Production*: Can the machine's components be manufactured economically?
    j. *Assembly*: Can the machine be assembled economically? For example, it is expensive to specify hand-finished surfaces because few people still have the skill and experience required to hand scrape or lap parts.
    k. *Quality control*: Can the device be manufactured with consistent quality in the quantities required?
    l. *Transport*: Can the device be transported to the customer's plant? What are the implications of a take-apart-put-back-together design?
    m. *Maintenance*: What service intervals will be required, and how will they affect other machines in the customer's plant?
    n. *Costs*: What are the allowable costs for completion of the project?
    o. *Schedules*: What are the time deadlines on the project?

Often, these initial specifications can be developed by considering a block diagram of the system and its inputs and its outputs. Decomposing the design into its component blocks while considering each of these factors can help the design engineer to assemble a detailed list of specifications.

[34] Most of this outline is from a major machine tool manufacturer's in-house literature for design engineers. Used with permission, but the manufacturer wished to remain anonymous.

2. *Perform a state-of-the-art technology assessment.* The purpose of this task is to evaluate existing technology and how it could be used in the context of providing answers to the issues raised above. For example, it is unlikely that you will want to develop metallurgical processes to optimize bearing steels for a ball bearing in a door hinge. Often the best place to start this task is with an analysis of your competition. Be careful, however, for if you merely copy designs without improving upon them, you may always be a generation or two behind. A result of this search might also be the generation of a wish list of technologies that need to be developed. If warranted, they may be developed for this project, or your list may stimulate management to provide resources for development of other products. Also, vendors may develop new products in response to your need if they perceive the market to be large enough.

3. *Preliminary specifications.* At this stage, enough information should be available to enable the design engineer to justify modified customer specifications. The result of this step is a list of design engineer generated preliminary specifications for the machine.

4. *Specification iterations.* The customer will usually be glad to review modifications to their specifications in the hope that it will save them money; however, they will probably be wary that you are trying to sell them something from your existing inventory just to unload it. As a result, several iterations are likely before a final set of specifications is agreed upon.

5. *Develop conceptual designs.* Up until this step, the design engineer has been gathering the seeds of thought, which must now be planted, nurtured, and harvested. This step is thus the most crucial, for it lays the foundation for the whole machine. It is vitally important to the success of each design engineer to develop his or her abilities at this art/science, as discussed in Section 1.4.4.

6. *Formulate a development plan.* Once a plausible conceptual design(s) has been developed, a development plan can be formulated that will enable a design team to proceed with making the dream a reality. This stage is most often done by a senior manager with input from senior design and production engineers. In developing a plan for the execution of the detailed design, the following factors must be considered:

   a. *Corporate objectives*. The product must satisfy corporate objectives. For example, a machine tool company may not want to dilute its resources by designing consumer products.

   b. *Projected return on investment.* Any design entails risking a percentage of the company's financial resources, and most companies like to spread risk among many different projects.

   c. *Required human resources*. A company may not want to devote a substantial amount of its human resources to a single risky project. Note that human resources include personnel required to service and maintain machines once they leave the factory.

   d. *Time plan.* A time plan must be devised to ensure that the design, manufacture, and installation of a machine can proceed within the period required by the customer. This includes allocations for support by the departments of engineering, manufacturing, marketing, service, documentation, training, and maintenance.

7. *Complete the detailed design.* To complete the detailed design, the design team will be assembled by the senior design engineer and engineering management. The team's tasks include the following:

   a. Double check the conceptual designs to ensure that the best are chosen.

   b. Perform preliminary design analysis, including a preliminary formulation of the machine's error budget.

   c. Review existing technology, especially patents, to avoid infringement.

   d. Select major mechanical, electrical, and sensor components along with making provisions for adequate space for seals, bellows, cable carriers, and safety shields. Also, make sure the machine can physically be shipped to and installed at the customer's plant.

   e. Perform a value analysis of major machine components.

f. Evaluate environmental requirements necessary to meet performance specifications (e.g., temperature, humidity, and vibration control).

g. Perform a safety review of the machine.

h. Conduct a design review to check for code compliance and design quality.

i. Perform an ergonomic design review with respect to operator and maintenance personnel acceptance.

j. Perform a design review of systems for tool changing, part transfer, coolant supply and containment, and chip removal.

k. Review required alignment procedures and requirements.

l. Generate initial layout drawings.

m. Generate a preliminary parts list.

n. Obtain a preliminary manufacturing review and cost estimate.

o. Perform an electrical systems review that closely scrutinizes choices of sensors, components, connections, cable carriers, and commutators.

p. Perform a hydraulic and pneumatic systems review that checks timing charts, line size, flow rates, pressure, component selection, thermal control, and filtering requirements.

q. Perform a value analysis of all the hardware.

r. Develop flowcharts and programming methods.

s. Document shipping requirements.

t. Develop final layout drawings.

u. Perform a final manufacturing review that includes costs, tool notes, methods, value analysis, and a review of commercial parts and vendor choices.

v. Perform a design analysis review and complete a final value analysis.

w. Develop a test program.

x. Perform a final design check and a final error budget evaluation.

y. Generate production drawings for the prototype.

z. Manufacture, assemble, and test the prototype.

8. *Complete the design follow-up.* After the design is complete, there is still much work that needs to be done in order to make the design successful, including the following steps:

a. Develop a test and user support program.

b. Award contracts for manufacturing engineering, manufacturing, production control, and assembly.

c. Check the quality control plan for critical parts and assemblies, color coding schemes, spare parts inventories, service parts inventories, and guarding requirements.

d. Review unique tooling requirements and associated field service.

e. Perform a safety audit update.

f. Obtain high-speed safety certification (if applicable).

g. Update design and documentation.

h. Prepare a final bill of materials.

9. *Prepare documentation.* No one will purchase a machine unless its certification, maintenance, and operation procedures are clearly documented. Often the best person to write an outline for these documents is the engineer who designed the machine in the first place. Documentation must include:

a. Records of tests, including part programs and description of tooling and fixtures used.

b. Owner's manuals.

c. Training courses and manuals.

Design is not just producing a stack of drawings of parts to be made. Design is part of the process of bringing a product to market. The best designs marketed are those in which the design engineer has been with the machine from its conception to decommission. No one knows a machine better than the person who designed it, and this is the reason for involving the design engineer in developing the implementation details which are crucial to success.

1.4.9 The Moral Responsibility of the Design Engineer[35]

Farmers grow food to keep people alive biologically. Clergyman keep people alive spiritually. Design engineers make things which make life physically more pleasant. Managers help design engineers to work together, and lawyers and politicians protect us from our human selves. We must all remember that history often repeats itself: power corrupts, and absolute power corrupts absolutely. While a dishonest few can ruin people's lives by cheating people of their hard-earned savings, people can always earn the money back. On the other hand, a bad design engineer's work can kill and maim. Because it cannot often be replaced, living human tissue is often worth more than all the stocks on Wall Street to the individual to which it is attached. *Thus design engineers must always make sure that quality and safety are the primary goals of their efforts lest they be forced to use their own products as punishment.* Always ask yourself, "Would I let my children use this design?"

Also consider the moral responsibility of design engineers to themselves and society to maximize their potential. In the 1800s, the average workweek was 70 hours and there was not a whole lot to do other than work. Today, technology has reduced the average workweek to 40 hours. That same technology, however, has also given us the means to waste time that might otherwise be spent on creative thought. *Rather than spend time glued to the TV set or hypnotized by the radio, consider the alternative: look, think, and analyze the way things work around you. Play mind games with yourself.* For example, while driving down the highway, try to calculate stresses in telephone pole wires, or ask yourself, how was that bridge designed? What is the total thermal expansion of the world from day to night? How many seconds have you lived? The list of possible entertaining questions is limitless. Compare these thoughts to the concept of muscle tone and success in sports. If you are to win, you must be able to respond instantly. Too much radio and TV clogs the mind and slows its response time, just as lack of exercise slows the muscles and decreases coordination.

The author admits needing time to just sit down and relax while watching nature shows on public TV, or listening to classical music. It is during these times that TV, music, and all the other wonderful things nature and technology has brought us all are useful and beneficial to the person as a whole. Unfortunately, many people tend to abuse relaxants. Alcoholics, drug addicts, and bad design engineers are all burdens that society can do without. The best "high" a design engineer can get is that which comes with success. The most relaxing thing in the world is knowing that you can do something and do it well. *Chemical highs that act on the central nervous system come and go and require a flow of capital to maintain them. A design engineer's high, on the other hand, lasts as long as the design is used, which is a function of how well it was designed in the first place.*

1.5 Design Case Study: A High Speed Machining Center[36]

This case study describes the evolution of the design of Cincinnati Milacron's high-speed three-axis (X,Y,Z) computer-controlled machining center. A discussion of the historical background of the project is followed by a discussion of the initiation of the design process for this machine at the company. The conceptual and detailed development of the High-Speed Machining Center (HSMC)

[35] *"Concern for man himself and his fate must always form the chief interest of all technical endeavors, concern for the great unsolved problems of the organization of labor and the distribution of goods - in order that the creations of our mind shall be a blessing and not a curse to mankind. Never forget this in the midst of your diagrams and equations."* Albert Einstein

[36] *"Few things are harder to put up with than the annoyance of a good example"* Mark Twain. The author would like to thank Dick Kegg and Bob Johoski of Cincinnati Milacron for arranging the interviews with their engineers that made this case study possible. The author would also like to extend special thanks to Tad Pietrowski, who was the chief design engineer for the HSMC, Richard Curless, Carl Padgett, Myron Schmenk, Jim Suer, and all the other people at CM who helped make this case study possible.

are then discussed.[37] Many terms used will be unfamiliar to the novice design engineer, but they are all discussed in detail in this book. Just like the design process itself, understanding these sections will require the use of the book's index and a little bit of effort.

1.5.1 Initiation of the Design Process

A major automobile manufacturer inquired as to whether a machine could be designed to drill holes in panels and machine aluminum engine blocks and other components at almost triple current rates. The result of this inquiry was the generation of the Required Specific Objectives (RSO) for a new machine. After the RSO was initially developed, a management team made a quick review to determine market potential. This review was positive, so the head of engineering delegated the job of performing a design feasibility study to a conceptual design manager. The conceptual design manager assembled a team of senior design engineers and analysts. In addition to the senior design team, some less experienced design engineers were included to further their training and allow for fresh thoughts and ideas to be presented.

The conceptual design team centered around a very creative design engineer who was also capable of handling the seemingly infinite number of issues associated with the machine's RSO. The chief design engineer had to be aware of technological limitations of all the machine's component systems, although he usually only considered first-order effects at this point. Using this knowledge, his first responsibility was to guide the design team for the generation of as many plausible designs as possible. All through the early design stage, whenever in doubt as to the feasibility of a process, manufacturing specialists were consulted. The result was the formulation of an acceptable working conceptual design. Providing input and support were specialists in actuators, sensors, control systems, and manufacturing. As the team developed and refined the design, they also updated the RSO. RSO modifications were also continuously reviewed by upper management and marketing to ensure that the customer's needs were met.

Once the conceptual design was developed, analyzed, and refined, an estimate of its cost and time to manufacture was made. At this point, management and marketing determined that the machine could be sold in sufficient quantities to justify the R&D investment needed to build a production prototype. Since management and sales always want to cut costs, the success of the project often depends on engineers with a strong appreciation for manufacturing engineering to make recommendations as to how parts and assemblies can be modified in order to reduce costs and increase quality and performance.

1.5.2 Definition of Machine Function and Specifications

The automobile manufacturer asked the machine tool manufacturer to design a machine that would be compatible with transfer line equipment, be able to quickly remove large amounts of metal, and be able to quickly drill numerous holes in a single part. This sequence was often used to machine engine blocks and other components that used a large number of bolts during assembly. High spindle and axis speeds, on the order of 20,000 rpm and 25 - 50 m/min (1000 - 2000 ipm), were specified. Most machining centers, however, had previously been designed to manufacture iron parts, and their spindle and axis speeds were limited to about 6000 rpm and 10 m/min (400 ipm), respectively. The aircraft industry had long been running machines with 20,000 rpm spindles for machining aluminum, but feed rates were not too high because of chip removal problems. On the other end of the manufacturing spectra, electronics manufacturing companies were using small precision drilling machines, with axis speeds on the order of 25 m/min (1000 ipm), to drill printed circuit boards. It was first determined that to move a tool through aluminum at the required speeds and feeds would require a twenty horsepower motor with a special high-speed spindle. Motor and spindle technology was available to meet the 20,000 rpm goal. However, the highest linear axis speed that could be economically attained was judged at the time to be 30 m/min.

With respect to the spindle speed, at high axis speeds, high spindle speeds are required to maintain minimum cutting rates. It was known that spindle speeds on the order of 20,000 - 40,000

[37] Many terms are used with which the reader may not be familiar. The reader is urged to look in the index to find where these terms are defined in the book.

rpm reduce and in some cases almost eliminate tool wear while cutting aluminum. The decision to limit the spindle speed to 20,000 rpm was based on limitations on current spindle design technology for general-purpose applications and customer surveys which indicated that 1) many customers would not want to pay a significantly higher price for a 40,000 rpm spindle, and 2) many machine operators would never use the 40,000 rpm capability because they were unfamiliar with it.

High acceleration rates to get the axes up to speed also presented difficult challenges. Initially, the RSO specified acceleration rates on the order of 2g to minimize move time between cutting operations. Back-of-the-envelope calculations showed that with present actuator technology, maximum reliable acceleration rates for a large machine tool structure were about $1/4$g. However, a careful analysis of the proposed machining scenario found that at $1/4$, the time spent during the acceleration period was only 5 to 10% of the total typical cycle time. At $1/4$g, the time to attain the maximum speed of 30 m/min would only take $1/5$ of a second or a distance of travel of 5 cm (2 in.).

In the course of their analysis of cycle times, the engineers discovered that a better way to decrease machine cycle time would be to develop a new method for changing tools. Traditionally, most automatic tool changers used an arm to grab the old tool out of the spindle, move it to the tool carousel, deposit the tool, pick up a new one, and move it to the spindle for insertion. As a result, much time is often spent changing tools, even with double-arm tool changers. As an alternative, the engineers proposed using the machine's high-speed axes to move the spindle itself to the tool storage carousel and exchange tools. This strategy would require only minor modifications to a standard tool carousel. However, it would result in the elimination of a complex mechanism from the design and reduce tool changing time by a factor of 2. The RSO was modified to incorporate the changes (30 m/min and $1/4$g), while still maintaining the overall required throughput performance of almost two to three times that of other existing machines.

Perhaps the most critical requirement for the HSMC was low cost. The potential customer anticipated large volumes of these machines being needed to upgrade and outfit existing and new transfer lines. Thus the machines had to be designed to be built quickly and inexpensively while maintaining high quality. Factors considered when analyzing old designs and conceptualizing the new design included:

a. *Geometry.* The previous corporate design strategy for machining centers called for sizes and configurations to be grouped into semimodular categories. To meet one of the principal RSO goals, the ability to be duplexed, only horizontal spindle machines could be considered. Historically, the size of the machine was also linked to the horsepower rating, which was a function of the amount of metal per unit time that could be removed, and the concept of cubic fixturing. The former is a function of the physics of metal cutting, while the latter is a function of how parts are handled and fixtured in production environments. It had been determined that most parts could fall into the shape of a cube, and if they did not, they could be fixtured onto large "tombstone" fixturing blocks to allow many parts to be simultaneously moved into a machine. The machine could then produce a set of parts with a minimum number of tool changes. This also helps to reduce downtime while waiting for a new part to be moved into the machining area. These factors led to the following rough guide ratio of cube size to power rating:

$500 \times 500 \times 500$ mm: 3.7 kW (5 hp)
$600 \times 600 \times 600$ mm: 7.4 - 11.1 kW (10 - 15 hp)
$800 \times 800 \times 800$ mm: 11.1 - 18.5 kW (15 - 25 hp)
$1000 \times 1000 \times 1000$ mm: 18.5 - 29.6 kW (25 - 40 hp)

The actual size of the HSMC's work volume specified in the RSO was 500 mm (20 in.) wide, 380 mm (15 in.) deep, and 500 mm high. A large 14.8-kW (20-hp) spindle motor was required for the high spindle speeds and spindle angular acceleration rates also specified in the RSO.

b. *Kinematics.* To be duplexable, the spindle had to be horizontal and have three linear degrees of freedom. In order to enable the machine to meet the throughput requirements stated in the RSO, the machine had to be designed to enable the spindle to change tools instead of a special tool changing arm. The machine also had to allow for the necessary tool path motions while allowing the part to remain stationary. Although these general capabilities had been achieved before, previous designs could not readily be adapted for high speeds. Among the factors that limited the applicability of previous designs was the distribution of mass relative to the bearing and drive systems and the incompatibility of sealing systems with high-speed axes and high-volume chip formation.

c. *Dynamics.* In order to move quickly and accurately, all bearings needed to have low friction and be preloaded. This would require a fully constrained recirculating rolling element or fluidstatic bearing design. Also, the drive system would be required to actuate the system with minimal creation of reaction moments. By minimizing moments, angular deflections and resultant Abbe errors would also be minimized. With respect to vibrations and mounting the machine to the factory floor, previous designs were large enough to require a many-footed mounting system, which in turn required the user to provide a thick stable concrete slab. The HSMC would use a three-point mount to reduce foundation requirements and installation costs. Passive vibration isolation supports would be provided along with reasonable specifications for vibration levels in the floor.

d. *Power requirements.* Machining aluminum at high-speed requires relatively low axial cutting forces and very high speeds. In fact, almost 80% of the actuator force would be expended accelerating the machine's axes.

e. *Materials.* First and foremost, stability of materials was considered of the utmost importance to maintain accuracy of the machine. Second, since most of the power expended into the machine would go into accelerating the mass of the machine (according to preliminary calculations), candidate materials' stiffness-to-weight ratio and coefficients of thermal expansion became important. Also of concern was materials' ability to damp out vibrations caused by the cutting process. Ceramics, aluminum, and cast iron were considered as possible structural materials. Although ceramics have an exceptional stiffness-to-weight ratio and often a low coefficient of thermal expansion, their cost precluded their use at this time. Cast polymer concrete was also considered for use in the base, but since the rest of the structure was made from iron, no cost advantage over cast iron could be realized at the time. Aluminum and cast iron had roughly the same stiffness-to-weight ratio; aluminum, however, has almost twice the coefficient of thermal expansion and half the elastic modulus of cast iron; thus cast iron was chosen for the primary structural material.

f. *Sensors and control.* Sensors in machining centers typically include linear scales attached to the axes as close as possible to the work region, or angular measuring devices (e.g., resolvers or encoders) attached to leadscrew actuators. Each system has its good and bad points, as discussed in following chapters; however, since customer surveys indicated that a significant number of customers preferred one over the other, the design would have to be able to use either type of sensing system.

With respect to using lasers to map and correct geometric machine tool errors for the accuracies specified in the RSO, mapping techniques would not be needed. Mapping of thermal errors would be useful; however, at the time, no significant industrial experience existed in the field of mapping thermal errors in a machine tool. Developing production methods to harden thermal mapping procedures developed in laboratories was judged to be too new a technology to use on a design that would already introduce so many new features. However, this would not preclude introduction of this technology in the future.

One issue stressed by the electronic systems design engineers was that the mechanical design engineers should be required to write operation flowcharts for the machine. This would greatly simplify the task of designing the machine controller and control software. In addition, this would help to ensure that the design engineers left space for all the required sensors and cables.

g. *Safety.* Any machine tool manufacturer still in business today is aware of the standards and codes governing the requirements for guards and shields on machines. Every design is scrutinized by review panels to make sure that a user would really have to think hard to find a way to hurt himself while using the machine. Chances would then be that if the user was smart enough to figure out a way of hurting himself, then he would be smart enough to know not to misuse the machine. Thus the machine had to be designed so that it would not run unless all guards and shields are in place.

h. *Ergonomics.* As with safety systems, numerous codes and company policies existed to make sure that the machine was serviceable and user friendly. The continual stressing of the importance of communication between design and production engineers and sales personnel helped to maintain the ability to design user-friendly machines.

i. *Production.* Traditionally, the highest-accuracy machines were manufactured using hand finishing techniques (e.g., scraping and lapping). Because of the anticipated production requirements, and the decreasing availability of skilled craftspeople capable of hand-finishing machines, the use of components with only precision ground accuracy was desired. This does not imply a

degradation of accuracy compared to old machines; it implies a recognition that modern grinders can finish parts to tolerances that in the past often required hand finishing to achieve.

j. *Assembly.* The same shortage of skilled craftspeople to hand scrape and lap bearing surfaces also results in a lack of skilled craftspeople to hand finish machine components during assembly. To help ensure quality, one method would be to use precision jigs to hold components in place while the components where fixed together using the technique of replication (filling gaps with zero shrink epoxy).

k. *Quality control.* Using a single precision jig to hold components in place while they are being glued together also helps to ensure uniformity. Because of the tremendous value added to this type of machine during manufacturing, quick self-checking methods also must be employed throughout the manufacturing process.

l. *Transport.* The machine size specified in the RSO historically never had any problem being transported fully assembled to customer plants.

m. *Maintenance.* Since the machine was to be sold as a unit that could easily be integrated into a transfer line, it was imperative to reduce maintenance to a minimum. The use of modular components would help minimize repair costs and downtime.

n. *Costs.* The initial request for a design study by the client emphasized the need for a low-cost, high-volume product. Often when a custom design was done for a company, the engineering cost was on the order of the manufacturing cost, because only a few units were made and it was cheaper to overdesign the structure. Not so on this type of high-volume item; cost was thus an integral factor in many of the design decisions made.

o. *Schedules.* If one company had thought of using a high-speed machining center to speed up their manufacturing capabilities, others probably would too. Thus it was important to complete the design as quickly as possible. Altogether, design-to-prototype time was halved from the traditional 2 year period to 1 year. Increased use of computers, streamlining of the design and review process, and the urge to be more competitive helped to achieve this schedule. However, ideally it would be nice if the time it took to complete this type of design could be reduced even further, to 6 months or less.

1.5.3 Conceptual Designs of the HSMC

A horizontal spindle machine is generally more accurate than a vertical spindle machine because the spindle is not cantilevered off a large C-shaped support structure, which is subject to greater deformations. On vertical machining centers, loads from the spindle create a bending moment that acts on the column. On a horizontal machine, loads from the spindle act as a point force on the column; thus it is more rigid than a vertical machine. Chip removal is also easier on a horizontal machine. Fixturing of parts on horizontal machines, however, is sometimes more difficult because gravity does not always work to help hold the part while it is being clamped in place. When modular palletized fixturing is used, the problem often no longer becomes apparent on the production line.

Three basic designs for the HSMC evolved from numerous brainstorming sessions. The first contained only revolute joints and thus had the characteristic geometry of a human arm. The articulated arm, shown conceptually in Figure 1.5.1, had several good points: (1) it was dexterous, (2) it would be easy to seal out contaminants, and (3) it was sexy. However, it had several fatal flaws: (1) in order to position the cutting tool in three-dimensional Cartesian space, it required five axes of motion; and (2) it had a very low natural frequency, due to its extended cantilever design, which would require the servocontrolled actuators not only to resist cutting forces and accelerate inertias but also to support the weight of the structure when extended. A similar design was the articulating arm on a linear slide, shown conceptually in Figure 1.5.2. This design required only four degrees of freedom to position the cutting tool in three-dimensional Cartesian space; however, it still had the cantilever stiffness problems associated with revolute joints.

Another hybrid design was the scissors jack design shown in Figure 1.5.3. This design required only three actuators to position the tool, but required many revolute joints to form the necessary linkage. The mechanical advantage provided by the linkage was nonlinear and when compared to a traditional linear bearing slide configuration, it cost significantly more and had a greater potential to suffer component failure. Thus this design was also discarded.

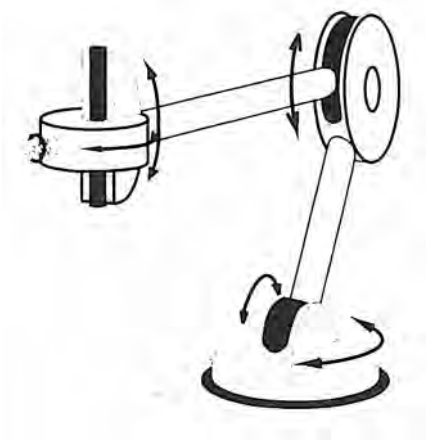


Figure 1.5.1 Fully articulated design for a high-speed machining center.

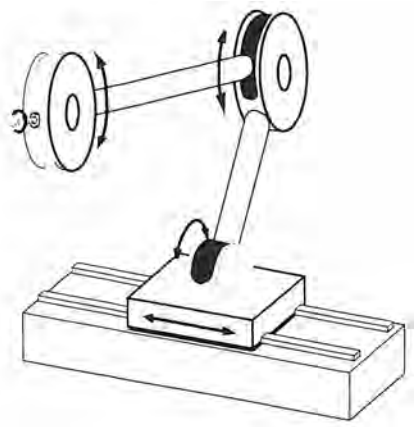


Figure 1.5.2 Traveling fully articulated design for a high-speed machining center.

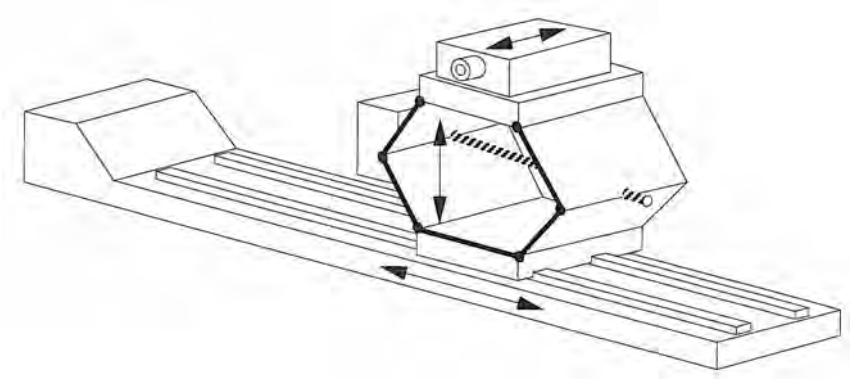


Figure 1.5.3 Conceptual scissor jack design for a high-speed machining center.

Free-spirited thinking with possible combinations of revolute and Cartesian joints did not lead to any plausible designs. In addition, a senior student design contest at a local university also failed to produce any plausible unconventional designs. Hence the design team went back to thinking along the lines of using only linear motion bearings to achieve the desired XYZ motion. It was, however, important to have spent the time investigating all possibilities.

The team's next thought was: *"Why not make the column only move side to side in the X direction?"* Z direction motion would be achieved by moving the axis in and out on a slide which would contain the spindle itself. The Y direction would function as it always had; it would move the spindle up and down. This concept is shown in Figure 1.5.4. This design seemed well suited to the task. It had three axes of motion and the spindle nose had the mobility needed to exchange its own tools at the carousel. The question now remained of how to seal the system against the flood of coolant and chips that would be generated during high-speed machining operations. In fact, when designing a machine of this type, almost 30 to 50% of the design effort may actually be spent incorporating guards, seals, and cable carriers into the design. Usually, these components are available "off the shelf," but their geometry or design properties often influence the grouping of other elements.

The issue of sealing the axes against the immense flood of coolant and chip contaminants thus became an issue of prime importance. Bellows-type way covers could handle the speed, but the corrugations could become clogged given the huge volumes of chips that are often generated in high-speed machining of aluminum. Sliding or telescoping way covers on all three axes could be used, but the issue of drag forces from the sliding seals drew concern, as did buildup of chips on the covers. As mentioned earlier, the design had the image of a spindle moving in and out of a wall.

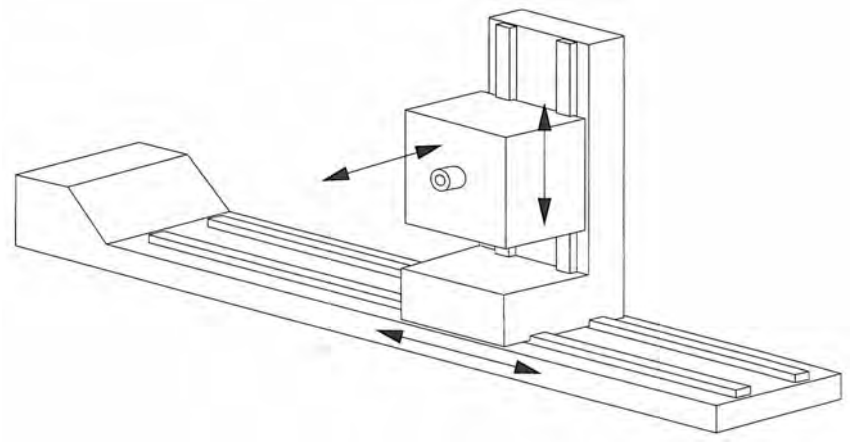


Figure 1.5.4 Conceptual rectilinear design for a high-speed machining center.

As shown in Figure 1.5.5, by pressing the wall against a plate with a cutout large enough to move the spindle through its planer range of motion, a single, square-shaped seal could be used to keep out coolant and chips. This eliminated the need for bellows or way covers on the axes. In turn, this reduced the cost of seals and reduced the effort needed to move the axes by reducing seal friction. The moving wall design also had an impact on the manner in which heat from the cutting process and chips were removed from the machine. By presenting only a smooth wall to the cutting process, chips could not build up and act to create local hot spots on the structure. Furthermore, because the chips would not be able to accumulate, they would be less likely to jam the machine's part transfer mechanism. By providing a totally enclosed sealed housing around the cutting area, a virtual flood of coolant could be used to carry heat and chips away from the work zone. Routing wires and air lines to the various axes became almost straightforward now that the moving wall design had been established. Since the entire structure was to be enclosed inside a sheet metal box, nothing could get in to interfere with cable carriers.

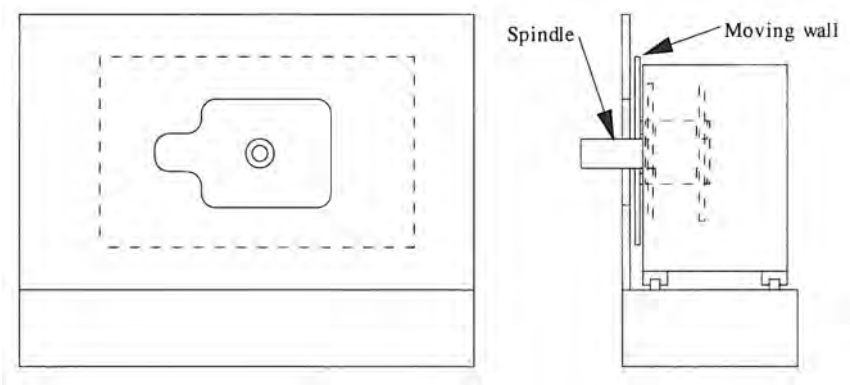


Figure 1.5.5 Concept of the moving wall sealing system.

In summary, with all three axes stacked, the spindle could project and move in and out literally from a moving wall. This would allow for a smooth interface to the world, which would be very efficient at keeping out chips and cutting fluid. Thus the wall design helped significantly to simplify the design of the rest of the machine. This seemingly trivial issue of how to seal all three axes simultaneously turned out to be one of the principal advantages of the new design.

1.5.4 Detailed Design

Once the conceptual design was developed, preliminary component sizing was done to verify the workability of the design. Next, manufacturing personnel were brought in to review the design before any more detail was done. With a consensus that the machine would not be difficult to build, the entire design package was presented to upper management for approval along with a preliminary cost estimate. Engineering, tooling, and inventory costs were evaluated and compared to expected revenues from anticipated orders. Finally, the decision was made to proceed with the detailed design. The detailed design team was then assembled and tasks delegated based on areas identified during the conceptual design phase. In describing the detailed design of the HSMC, the design methodology will first be outlined by considering the dominant physical effects (sources of error) governing the accuracy of the machine. Mechanical and electrical systems will then be discussed.

1.5.4.1 Sources of Error

Before a design engineer can rush off to a terminal and start drafting machine components, he or she must consider how various inputs to the system will affect the machine outlined in the conceptual design phase. Although the subject of causes of errors in machine tools is discussed in detail in Chapter 2, specific sources that should be kept in mind while reviewing the choice of mechanical and electronic system components for the HSMC will be discussed here. Note that machining center performance is now evaluated with respect to ANSI/ASME B5.54 1992 standard for machining centers.

Geometric Errors

Errors in machine geometry occur in all axes and are composed of three translational and three rotational components per axis. There are also orthogonality and parallelism errors between axes. The worst type of geometric errors are angular errors. Angular errors are amplified by the distance between the source and the point of interest and they are called *Abbe errors*. Abbe errors are the most often overlooked source of error. Static deflection errors are caused by the weight of the machine itself and the weight of the workpiece. With careful modeling of the structure, joints, and bearings, static errors can usually be accurately predicted and kept below a desired threshold.

Dynamic Errors

Dynamic errors in the machine are due primarily to structural resonances excited by the cutting action. This causes chatter in the tool and can dramatically degrade the surface finish of a part. This type of vibration is very difficult to analyze and predict because when the tool touches the workpiece, it changes the stiffness properties of the machine. A conservative approach to modeling this phenomenon is to use finite element methods and modal analysis to try and predict structural resonances. The machine controller can then be programmed not to run the cutting tool at a speed which might excite one of the machine's resonant frequencies. The changing stiffness of the machine caused by the tool touching the workpiece can be dealt with by putting a larger safety band around each machine's catalog of resonant frequencies. The controller then can be instructed to avoid these operating speeds. Although this does not address the issue of cutting tool stiffness and vibration modes, identifying resonances also helps design engineers develop specifications for allowable floor vibration levels, and to specify servoloop times and operating speeds.

Dynamic errors are often traceable and characterizable once the machine is built, but they are the most difficult to predict and compensate for during the design phase. In general, making the machine rigid, well damped, and lightweight helps to make it easier for the machine to achieve high-speed operation.

Thermal Errors

Thermal expansion of machine tools is rarely uniform, so measurements cannot necessarily be scaled solely by a linear compensation factor. The cause of these thermal deformations is heat generated by motors, linear and rotary bearings, and the cutting process. Environmental changes in temperature usually affect the machine in a more uniform manner but must also be considered. The ability of the machine to track the temperature of the environment depends on the thermal cycle of the environment, the thermal time constant of the machine, and the heat transfer characteristics between the two. If the environmental temperature is steady, then the machine should be strongly

coupled (thermally) to the environment. If the machine is to be located in a drafty warehouse, it should be shielded from the environment.

To allow the machine to reach thermal equilibrium, most machines require a warm-up time. Traditionally, operators have warmed up machines by "cutting air" for extended periods. Some machine tool builders hand finish machines so that after the machine warms up, it deforms into the proper configuration. Attempts have been made at reducing warm-up time by using heating elements. However, actively using temperature-controlled fluids or resistance heaters for heating and cooling parts of the machine tool works only if the entire machine is treated in this manner. Otherwise, temperature gradients between portions of the structure invariably form, causing the machine to warp (elastically). In the long run, careful selection of components and geometries to minimize thermal distortion and actively cooling key heat-generating regions provides the most economical solution to the problem of thermal errors.

As a result of observations about past attempts at thermal control, for the final design it was decided to cool the spindle bearings with an oil mist, isolate motors from the structure as much as possible, and flood the work zone with coolant. In addition, the structure was made to be blocky in shape to help minimize angular deformations. Finite element analysis of the machine's thermal performance was also done to verify that the worst case assumptions of machine performance would fall within the accuracy specifications for the machine.

Workpiece Effects

Traditional machining centers are often adversely affected by the mass of the part to be machined because the mass distorts the structure of the machine and changes the machine's dynamic characteristics. With the high-speed machining center, the part arrives in front of the machine and is locked into one place until the machining is complete. Probes on the machine can be used to exactly where the part is before machining commences, thus compensating for how the part deforms the stand that supports it. In this respect the HSMC is immune to the effects of part weight.

1.5.4.2 Linear Bearing Selection

In order for the machine to move at maximum speed with minimum power input, low friction preloaded bearings were required. Low friction also helps to minimize heat generated during high-speed moves. Preloaded bearings were required to achieve high stiffness to resist inertial forces. Three types of existing linear bearings were considered for this application: fluidstatic, sliding contact, and rolling element bearings. Fluidstatic bearings, which include hydrostatic and aerostatic bearings, shown schematically in Figure 9.2.1, are the only types of bearings for machine tools that are truly frictionless and preloadable. The former use a cushion of high-pressure oil to float one structure above another. Aerostatic bearings use low-pressure air and are sensitive to shock loads. There has been extensive use of hydrostatic bearings in machine tools, particularly grinders, and they have proven to be among the most accurate bearings available. One of the design goals of the HSMC was to have an all electric machine in order to minimize maintenance and maximize uptime of the machine. Thus hydrostatic bearings would be considered only if no other suitable bearing was found.

Sliding contact bearings for machine tools, discussed in Section 8.2, utilize a thin layer of low-friction material (e.g., a plastic from the PTFE family) bonded to the surface of the moving axis. The pads generally require lubrication and slide on hardened steel or cast iron rails. This type of bearing had been the mainstay of the machine tool industry for many years. It gained an excellent reputation as a high-stiffness medium friction (static coefficient of friction on the order of 0.1) bearing with excellent damping characteristics. The large surface contact areas that could be attained with this type of bearing allowed machines to resist very high cutting and shock loads. Sliding friction bearings could be easily machined, ground, or scraped, and were still a very popular bearing material. However, their finite friction properties meant that power input to the high-speed axes would be more than double that required for a system with a very low friction bearing such as a rolling element type. Also, finite friction sometimes leads to a condition known as *stick-slip*, which can limit the accuracy and resolution of the system. Stick-slip is best characterized by trying to push a book to a desired position on a table. The initial force to get the book going impedes the accuracy to which it can be moved to a desired location. This effect is especially apparent when you try to move the book to a desired position with higher and higher approach speeds. Many sliding contact

bearings have minimal stick-slip (static μ almost equal to dynamic μ). In this case it was not the stick-slip but the heat generation which prevented sliding contact bearings from being used.

The only other bearing alternative was rolling element bearings. Interestingly enough, the machine tool industry traditionally used highly skilled craftspeople to hand finish machines, which in turn were used to build other machines. From the first rock a caveman pulled out of the ground, handmade tools have been used to make it easier to build better next-generation tools. Similarly, the machine tool industry has developed more and more accurate grinding machines which have allowed for the manufacture (by machine) of long travel, high accuracy, recirculating rolling element bearings suitable in many cases for the replacement of sliding contact bearings. There are many types of rolling element linear bearings, as discussed in Section 8.5, which have very low friction characteristics; however, they cannot carry as much load (per area) and do not have as good damping characteristics as sliding contact bearings. Furthermore, once worn out they cannot be refinished or adjusted with a gib. Thus for a long time, the machine tool industry avoided their use, except on lower-power (<7 kW) machines. However, rolling element bearings' modularity, low cost, and low-friction properties were irresistible.

Cylindrical rolling element bearings had been used before on large machines, but they often require hardened steel ways to be custom designed for the structure, and typically required 12 bearing cartridges to fully support an axis. As an alternative, the engineering team considered a self-contained unit, commonly referred to as a linear guide, as shown in Figure 8.5.30. Linear guides used round balls which, although they cannot handle the high shock loads that cylindrical bearings can, were sufficient for the HSMC. Linear guide rails could be bolted to the structure and have parts bolted to the carriage with minimal alignment difficulties. Load capacity and stiffness could be attained simply by adding more bearing units to the linear guide rail. The more units the greater the cost, but a point was reached on a test jig where an economical prototype axis was successfully built. The prototype carriage assembly was used on a conventional machine, and it yielded well-finished parts at a lower cost than for an antifriction pad type of bearing system.

The conclusion of the bearing search was that linear guide type rolling element linear bearings would be well suited to this application if the number of bearing blocks and their placement were carefully chosen. From a manufacturing point of view, it enabled a major precision component to be bolted into place onto precision ground surfaces. Ultimately, it was desired to use the technique of replication to attach the bearings to the structure. Replication would require master craftspeople to hand finish precision jigs in the shape of the bearing rail. The jig would then be sprayed with mold release and suspended in troughs in the machine's rough castings. Special epoxy would then be poured around the jig, where it would take the shape of the bed and of the jig, so a nearly perfect seat for the bearing rail could be obtained. Thus precision machining and finishing operations for linear bearing rail mounting would only have to be performed on a few jigs.

1.5.4.3 Actuators

The axes of the high-speed machining center required high-force, high-speed, high-accuracy linear actuators. The choices available included linear electric motors, rack-and-pinions, and ballscrews. Linear electric motors are capable of moving the axes without the use of an intermediate transmission, as shown schematically in Figures 10.3.10 to 10.3.13. The lack of a transmission to reflect a higher motor inertia and stiffness, however, results in this type of motor being impractical for accurate motion control of high-speed, high-mass systems subject to large cutting force disturbances. However, it seems probable that as motor technology advances (e.g., development of superconducting materials and magnets with ever higher flux densities), linear electric motors may one day be a practical choice for this type of application.

Rack-and-pinion drives, discussed in Section 10.8.1, are commonly used in machine tools where large ranges of motion (>3 m) are required. The gear forms can be ground very accurately, and multiple pinions preloaded against each other can be used to prevent backlash. However, a precision rack-and-pinion system does not have an inherent reduction ratio, so it requires a large motor or costly high-precision speed reducer. Thus as a system a rack-and-pinion drive is more expensive than a ballscrew for moderate lengths of travel. Rack-and-pinion systems also cannot be outfitted with self-cleaning wipers to wipe the gear teeth prior to contact with the pinion.

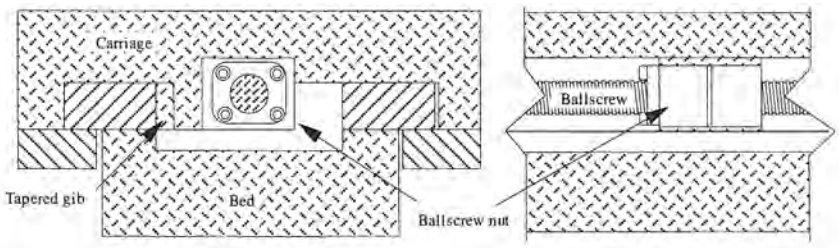


Figure 1.5.6 Schematic of a ballscrew-driven machine tool carriage with sliding contact bearings.

Leadscrews such as ballscrews or rollerscrews provide economic, high-efficiency, accurate rotary-to-linear motion conversion. In general they are the actuator of choice for machine tool actuators requiring moderate range of motion and speed. A leadscrew drive system, shown schematically in Figure 1.5.6, converts rotary power to linear power. For high-force applications, multiple start threads are used, which linearly multiplies the load-carrying capacity of the threads by providing an extra load-carrying path, but does not affect the force-torque relation. With ballscrews, balls are introduced between the threadforms of the screw and the nut to provide a rolling contact interface, which reduces friction to a minimum. Rollerscrews use a planetary system of rollers in lieu of balls and they can support much higher loads than ballscrews, but they generally have smaller leads. For the HSMC, ballscrews were chosen.

Choosing a ballscrew requires knowledge of the motor to be used and the loads anticipated. In order to keep motor rpms at a reasonable level (typically less than 4000 rpm), a screw with a large lead is required if the axis is to move at high speeds. If the lead is too small, then the motor and ballscrew have to turn at a high rate, which can result in a dynamic instability called shaft whip. Ballscrews with high leads, on the other hand, require higher-torque motors to attain the same force as would be attained with a small lead and low-torque, high-speed motor. The former also requires the motor to have higher torque, which means higher current, which can lead to greater heat dissipation. These factors end up imposing physical limitations on the practical speed at which a linear motion system can be driven using ballscrews. With the technology at hand, the engineers found themselves limited to a linear speed range of 25 to 50 m/min (1000 to 2000 ipm), which was within the requirements of the RSO. In order to choose the actual ballscrew/motor combination from all the possible combinations, a spreadsheet program was written that listed available motors and screw leads. The program was used to pick a combination that minimized cost and heat generated.

1.5.4.4 Axis Drive Motors

On a machine tool, motor frame size is directly related to the cost of the motor and the supporting structure. Often the design engineer is faced with increasing the size of the rest of the structure if he or she is uncertain about the size of the motor. Whenever possible, it is better to err on the conservative side and have the option of being able to install a bigger motor if necessary. This allows a smaller motor to be chosen for the prototype to test performance, while allowing for the option of using a larger motor to accommodate changing user requirements. It is far easier to buy and install a larger motor on a machine designed to handle small or large motors than to substantially modify the machine's structure in order to accommodate a larger motor. The design engineer must also allow enough room for installation of motors with integral encoders and tachometers.

Choosing the right motor for a machine tool is often done by considering the temperature rise of the motor. Because the motor windings have a finite electrical resistance, they become very warm during use. The motor housing is often bolted directly to the machine tool's structure; however, this also provides a direct path for the heat conduction. As a result, axis and spindle drive motors are one of the major contributors to the total amount of heat generated in a machine tool (bearings are the other). As discussed and illustrated by experiments in Chapter 6, heat transfer from motors to a machine structure can seriously degrade accuracy. The best method for preventing motor heat from being transferred to the structure is to try to isolate the motors from the structure using insulating high-stiffness ceramic interfaces and radiation shields. Cooling the motors with fans and ducting the

cooling air away from the machine are also effective methods. It is also wise to design the machine to dissipate heat and deform uniformly.

1.5.4.5 Spindle Design

Since the spindle would have to be massive to house a 15-kW (20-hp) motor, it was important to identify the spindle design for the HSMC at the earliest possible time so that detailed work on the rest of the structure could proceed. For ball bearing spindles, speeds higher than 20,000 rpm were generally attainable only in small diameters where centrifugal forces would not cause high radial forces on the balls to brinell (indent) the race and promote rapid fatigue; however, standard tooling for machining centers would require a spindle diameter that was too large for ball bearing technology to perform safely and accurately at speeds greater than 20,000 rpm.

Aerostatic bearings were considered as an alternative to ball bearings to allow the spindle speed to be increased. Although more accurate and able to achieve higher speeds than ball bearing spindles, most air bearing spindles are more susceptible to damage if the spindle were to feed a tool into a part incorrectly (i.e., a crash). Even in installations which would be completely computer controlled, there would still be the chance of a programing error causing a crash. Experience with a typical industrial operator showed that one crash per day was not uncommon. This would not be acceptable with most air bearing spindles. A porous graphite air bearing spindle developed and perfected at Oak Ridge[38] in the 1960s had proven itself very accurate and reliable even when "crashed"; however, the cost of the spindle (on the order of $25,000 from manufacturers that were producing it) made its use impractical. Thus air bearings were ruled out as potential bearings for the spindle.

Hydrodynamic bearings had been used very successfully in many types of machine spindles and were less susceptible to crashes since the parts of the spindle do not have to be assembled with as close tolerances as are required for air bearings. Hydrodynamic bearings operate by dragging a viscous layer of oil between parts in relative motion; thus the position of the spindle axis becomes a function of spindle rotation speed. At high speeds, the viscous shear of the oil also creates large amounts of heat, although the oil can be cooled. The expense and reliability issues associated with adding sensors to monitor spindle location and use the signals as an error correction signal were studied, but the option was too expensive and unreliable to implement on a production machine. In addition, the large number of rapid start/stop cycles required for high-speed tool changing would not allow the hydrodynamic lubricating layer to form properly, and a pressurized fluid assist (hydrostatic) stage would be required. Hydrodynamic bearings are most often used on machines where the spindles run for extended periods of time at constant rpm (e.g., some grinding machines). These factors led to the conclusion that hydrodynamic spindles would not be appropriate for the HSMC.

Magnetic bearings were also considered but were found to be too expensive for the application. Magnetic bearings use servocontrolled electromagnets placed around the periphery of the spindle to keep it centered. Their primary limitation is the cost and complexity of the control system required to keep the spindle centered amid disturbances from large cutting forces.

Once an appropriate off-the-shelf ball bearing spindle was found, the next step was to identify a power and transmission system to drive the spindle. The HSMC was intended for rapid starts, stops, and tool changes. This drastically affected the motor selection criteria. The current technological trend had been toward brushless dc motors because of their low maintenance costs. For the HSMC's spindle, the motor would have to rapidly bring the spindle up to speed, machine a part (e.g., drill a hole), and then rapidly decelerate it for tool changing. Since the energy stored in a system is proportional to the velocity squared, the spindle motor ends up being the largest motor, and source of heat, in the system.

A brushless motor has no dynamic breaking capability and thus requires electrical current to be supplied in order to slow the motor down. Brushed dc motors, on the other hand, act as generators (the power is actually dissipated by an external resistor bank) when the polarity of their leads are switched and thus actually generate current as they quickly slow down. Since almost 25% of the power input to the HSMC spindle would be expended starting and stopping for tool changes, significant power savings could be realized by using a brushed motor. According to the motor supplier, and confirmed by a bench test, there would not be a problem with the motor brushes arcing at the high motor speeds required. Thus choosing a motor that had the ability to brake itself without

[38] See W.H. Rasnick et al., "Porous Graphite Air-Bearing Components as Applied to Machine Tools," SME Technical Report MRR74-02.

large external power inputs (i.e., a dc brushed motor) helped to reduce the overall heat input into the machine.

Another spindle design criteria was the ability to tap holes at high speed and then stop the spindle before the tap reached the bottom of the hole. Unfortunately, to meet this criterion an unrealistic motor would be required to generate the torque levels for this almost instantaneous stop of the entire spindle. Fortunately, tapping mechanisms exist that have a special clutch designed to allow for motion between the tap and the spindle when the tap reached the bottom of the hole. Just knowing that the right tool exists somewhere can often save considerable design time and effort.

Once the spindle motor type was chosen, a motor/transmission configuration had to be specified. Since the spindle would be required to run from 200 to 20,000 rpm, the first thought was to drive the spindle directly using an unhoused motor. The motor would use the spindle bearings to support the rotor with the stator held by the same structure that held the spindle bearings' outer races, as shown in Figure 8.7.2. This option would make for a very clean design; however, the cost of a motor that could operate with this large speed range was prohibitively expensive at the time (in the future it may not be). Also, sealing the motor brushes from the bearing's oil mist lubrication would be difficult. It is possible for a brushed motor to operate at low speeds in an oil bath as long as a reasonable oil flow rate through the motor is maintained. This allows for good temperature control of the motor. However, in this application high spindle speeds would cause too much viscous shear if the motor was flooded with oil.

As a second option, a more economical motor with a two-speed transmission system was considered. Gear transmissions were not available for the speeds required, so a proprietary timing belt transmission was developed. A prototype test stand was built, tested, and abused to determine the reliability of the design. It worked well and no major modifications were required. The design of the spindle and its power transmission elements was thus completed and readied for integration with the rest of the design.

When choosing the best spindle design, great attention also had to be paid to thermal effects. Previously, various manufacturers had tried to control spindle temperature by either cooling or preheating the spindle to maintain constant temperature. Cooling a high-speed spindle is often done with an oil mist from the inside. If the outer race of the spindle receives the cooling action first, it can contract and overload the spindle bearings causing them to seize. Unfortunately, the use of oil to cool a rapidly rotating inner diameter results in the cooling oil being agitated and sheared, which itself generates heat. Thus the cooling oil temperature had to be carefully controlled and pumped to and from the machine via insulated hoses. Preheating the spindle would cause problems with the rest of the structure as discussed earlier, so it was not considered to be a feasible option.

1.5.4.6 Structural Design

The chief structural design criterion was to minimize the weight and maximize the stiffness while using steel plate as the primary structural material. General rules of thumb were followed to accomplish this goal. The first was to have as much mass as possible placed as far away as possible from the center of mass. This creates an I-beam effect, which maximizes the stiffness-to-weight ratio. The second rule was to balance stiffnesses of the structure to make sure that no single member was overstressed while others were understressed. The wisdom of applying the latter principle became apparent when the first round of finite element analysis showed almost zero deflection in a stiffening plate and only a small deflection in the plate it was supporting. Upon closer examination of the strain output tables, it was found that if the position and orientation of the stiffening plate were changed to place more load on the stiffening plate, the deformation in the other main member decreased substantially. The lesson is the design engineer should not be happy with a design that merely works, but should strive for the best performance possible. The third rule of thumb was to decompose the structure into tractable analyzable modules. Each module had a budgeted stiffness, such that the total combined stiffness met the requirements for the machine.

1.5.4.7 Hydraulic and Pneumatic Support Systems

One of the goals set for this machine was to utilize only electric motors or actuators powered by compressed air, which is available to all machines in a machine shop. This would require changing

the design of the company's tool storage carousel motor system. Historically, the company's tool-holding carousels used a hydraulic motor and hard stops to index the carousel. For the HSMC, an electric-powered system was designed that used a servocontrolled electric motor to position the carousel. Although the electric system's manufacturing cost was $300 greater than a hydraulic system, maintenance engineers estimated that it would cost six times less annually to maintain the electric system.

1.5.4.8 Sensor Systems

The sensors and control system are the central nervous system of the machine tool. Without sensors to monitor position and other variables (e.g., temperature and pressure warning sensors), the machine would shut down. Unfortunately, sensors are also the most delicate part of the machine. In general, it was desired to minimize the number of sensors in order to increase potential uptime of the machine. On the HSMC, position sensors are needed to measure the linear position of the axes along the full length of travel and to measure discrete points of travel (i.e., home positions). To measure linear motion, there were a number of economical alternatives available: inductive scales, linear encoders, magnetic scales, and rotary sensors on precision ballscrews.

For this application, the degree of resolution required (1 μm) limited cost-effective choices to linear scales or resolvers (or encoders) on ballscrews. Linear scales ideally would be located as close to the toolpoint as possible to minimize Abbe errors. Unfortunately, there would be chips flying, coolant splashing, and lubricating oil dripping. The next best location for linear sensors would be close to the ballscrews. Resolvers are desirable rotary sensors because they are easy to seal and are tolerant of dirt. The principal problem with resolvers (or encoders) is that any error in the ballscrew caused by expansion due to heat generated as the ball nut rotates at high speed would not be measured. Linear scales, on the other hand, measure the motion of the carriage directly. As long as the ballscrew could provide smooth backlash free motion, the controller could use the information from the linear scales to accurately control the motion of the carriage without limit cycling. Thus although linear scales are somewhat more expensive and harder to install than resolvers, they can increase accuracy and reliability. Still, due to customer demand as determined by surveys, both types of sensors were made available.

Sensors to measure discrete position were used frequently in the machine to help define axes home positions and positions of tools in the tool storage carousel. For high-cycle applications, noncontact switches are the sensors of choice. To prevent motion beyond the intended range of travel, three barriers are provided. The first is a software axis travel limit that interrupts the main control routine if a linear sensor reading exceeds a maximum value. The next barrier is a noncontact proximity sensor that signals a relay to turn off power to the motor. If that fails, a hard mechanical stop prevents the axis from falling on the floor.

1.5.5 Summary

This case study highlighted the design process for a typical high-performance, computer controlled horizontal machining center. The description of the design process may lead one to believe that the design decisions were straightforward; on the contrary, extensive amounts of engineering experience, benchtop testing, and calculations were required to complete the design and prepare it for production. The finished product is shown in Figure 1.5.7 and has been very successful commercially. Try to imagine all the components discussed earlier as they would appear with the sheet metal covers removed. The ability to mentally disassemble a machine in this way is a skill that can help you find the strengths and weaknesses of your competition.

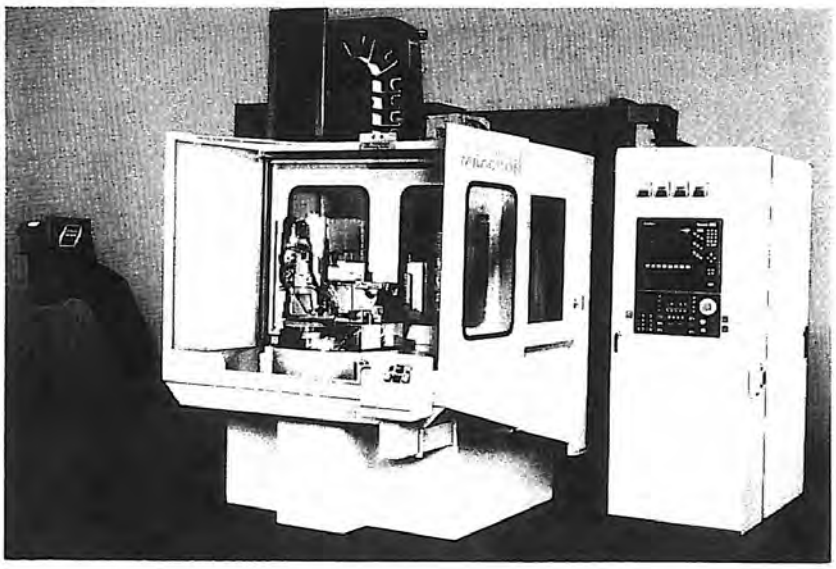


Figure 1.5.7 High-speed machining center. (Courtesy of Cincinnati Milacron.)

1.6 Design Case Study: A Coordinate Measuring Machine[39]

Coordinate measuring machines (CMMs) help form the cornerstone of the concept of part interchangeability by measuring the accuracy of parts to a higher degree of accuracy than that to which the parts were machined. As a result, manufacturers can use CMMs to help control the quality of their manufacturing processes. An example of the evolutionary design process for a CMM is provided in this section by studying the design of Sheffield's Apollo Series Cordax® coordinate measuring machine.

CMMs cannot necessarily be used to determine *why* a part is out of tolerance. For example, one might ask is a hole too big, or is it elliptical in shape with the major axis greater than the allowable hole diameter? For the former case, the boring tool needs to be adjusted. For the latter case, the accuracy of the spindle may not be great enough to allow the part to be machined to the proper tolerance. A CMM is usually used to probe three or more points along the hole contour which are used to determine if the mean radius of the hole is within tolerance. In order to determine the source of error, many more points along the contour would have to be measured, which would slow throughput. In most applications, the former assessment of part accuracy is sufficient because machine spindle accuracy has been certified previously and the cause of out-of-tolerance parts is usually due to improper setups.

Typical high-performance coordinate measuring machines are characterized not only by accuracy, but also by high-speed axes which move quickly from point to point while making "measurements on the fly." They accomplish this with the use of a device called a *touch trigger probe* which is shown in Figure 1.6.1 being used on an Apollo CMM. The touch trigger probe is rigid until it touches the part surface. As the tip of the probe touches an object's surface, it triggers a signal to the CMM's controller to measure and record the CMM's axes' position. A special linkage then allows the probe tip to be subjected to several millimeters (often up to a centimeter) of overtravel. The controller signals the CMM to decelerate quickly when the probe triggers. This operating mode allows a series of rapid high-speed measurements to be made. Greater accuracy can be attained at the price of speed, depending on the design of the particular machine and probe.

[39] The author would like to thank Dr. Stephen Fix at Sheffield for arranging interviews with the Apollo's designers. Special thanks to Fred Bell, Bob Brandstetter, Don Greier, and Tom Hemmelgarn for their help in making this section possible.

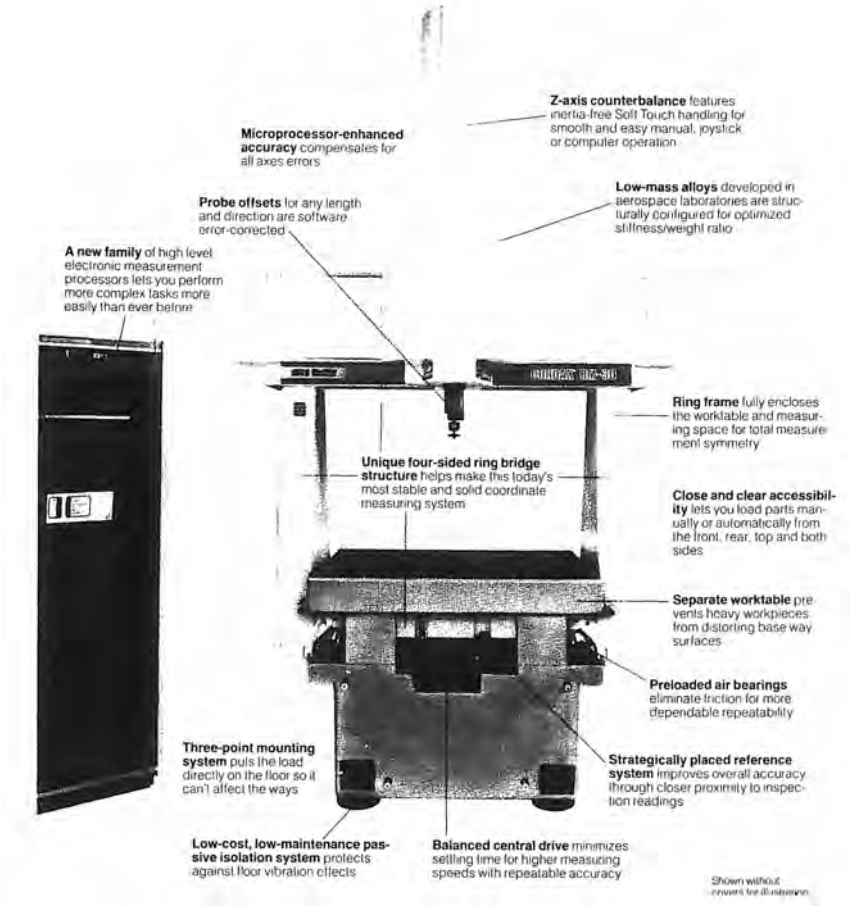


Figure 1.6.1 Touch trigger probe being used with an Apollo CMM to measure a part. (Courtesy of Sheffield Measurement, a Cross & Trecker Company.)

1.6.1 Foundation of Functional Specifications

Historically, there have been many types of coordinate measuring machine configurations developed for specific purposes. The principal goal of Sheffield's engineers was to increase the number of parts per hour that a CMM could accurately measure, and to make the accuracy of the machine independent of the weight of the part. In order to learn from previous CMM designs produced by many different manufacturers, market surveys were done which led to the formulation of a competitive analysis (CA) factor:

$$\text{CA} = \frac{\text{volume} \times \text{throughput}}{\text{accuracy} \times \text{price}} \tag{1.6.1}$$

The CA factor provided the design engineers with a qualitative tool with which to judge potential new designs.[40]

At the time the need for a faster, more accurate machine was identified, roughly three market brackets existed for coordinate measurement machines: ultraprecision, high throughput, and small manual machines. Comparatively few machines from the first group were being sold worldwide because of their high cost and the great amount of hand finishing, which required long lead times, necessary to achieve submicron accuracy. These machines, however, are essential for use in tool and die making and precision instrument fabrication. As a result, their manufacturers enjoyed a special niche in the marketplace that was difficult to penetrate. Thus it would not have been wise to attempt to compete in this bracket. The second bracket was responsible for the sale of a great number of high-throughput machines worldwide to all types of manufacturing industries. These machines

[40] Also see Methods for Performance Evaluation of Coordinate Measuring Machines, ANSI/ASME Standard B89.1.12M-1985.

typically had accuracies on the order of 10 to 15 μm (0.0004 to 0.0006 in.) and were capable of making rapid computer controlled measurements. This was the segment Sheffield was already competing in, and they enjoyed a significant share of the market. The last segment of the market centered around small, benchtop-type machines, and the market was judged to be fairly limited and highly competitive. Thus from a product and market share point of view, if Sheffield could achieve a factor of 2 or more increase in the competitive analysis factor, a larger portion of the high-throughput market could be won.

Having reaffirmed the wisdom of competing and designing machines for high-throughput markets, a design study for a new machine with a high CA was launched. Factors to be considered when analyzing old designs, conceptualizing the new design, and comparing the two included:

a. *Geometry.* The CMM should be of a modular design that could be scaled to small to midsize machines. Three basic machine sizes were identified by marketing to meet customers' needs: the machines were to have maximum part weight capabilities ranging from 10 to 18 kN (2200-4000 lbf). Work volumes were to enable handling of parts from 1000 to 1200 mm (40 to 48 in.) wide, 750 to 2000 mm (30 to 80 in.) long, and 620 to 1000 mm (25 to 40 in.) high. The footprint of the machine was, of course, to be as small as possible and allow large parts to be loaded using an overhead crane.

b. *Kinematics.* In order to overcome many problems inherent with other existing designs (as discussed later), such as accuracy dependence on part weight, a new type of machine configuration would have to be developed. In general, it would be desirable to have a symmetrical machine driven through its center of gravity, with the measurement structure fully decoupled and independent from the part support structure; hence, elastic deformation of the part support structure would not affect measurement accuracy, regardless of part load. The entire machine should also be mounted on a three-point mount to minimize the chance of machine accuracy being affected by deformations of the floor.

c. *Dynamics.* In order to move quickly and accurately, all bearings should be frictionless and preloaded. In general, this would require the use of opposed air bearing pads. Also, the drive system would be required to actuate the system with minimal creation of reaction moments. By minimizing dynamic moments, angular deflections and resultant Abbe errors would also be minimized. Passive vibration isolation supports should be provided along with reasonable specifications for vibration levels in the floor.

d. *Power requirements.* Since coordinate measuring machines do not remove material from the part, axis power requirements are much less than are required for most machine tools. Also, since CMMs are often used in a manual or teach mode where an operator guides the probe, the mass of the structure should be minimized to prevent operator fatigue. Minimizing the mass of the CMM also minimizes the actuator size, which would in turn minimize heat input to the structure. Despite operation in temperature-controlled environments, heat generation within the machine could still lead to thermal gradients, which could cause significant distortions of the machine. The use of frictionless air bearings would also help to reduce overall power requirements.

e. *Materials.* Stability of materials used for the structure was of the utmost importance. However, it would be nice to know that geometrical errors in the machine could be measured with a laser interferometer and corrected for in software[41]; thus in case the machine was ever abused and its accuracy degraded, it could be recalibrated. Other than stability, the stiffness-to-weight ratio was the chief factor affecting materials choice. Maximizing this ratio would help to minimize power required to move the machine's axes. This in turn would reduce heat input to the machine and minimize thermal deformations. Numerous types of structural materials had become available since the days of cast iron and a thorough review was conducted.

Another advantage of aluminum is its ability to rapidly achieve thermal equilibrium with its environment. This is due to aluminum's high thermal diffusitivity, which allows the thermal response of aluminum to be an order of magnitude faster than that of most ceramics or granite. Thus aluminum structures are less likely to develop thermal gradients, which can cause bending moments and Abbe errors.

The aerospace industry had considerable success in developing high stiffness-to-weight ratio, stable alloys. Thus in the search for lightweight alternatives to traditional materials such as cast iron

[41] Not all CMM manufacturers do this, but Sheffield had previously determined that this capability was strategically important.

and granite, it seemed logical to consider aerospace industry experiences. In addition to the requirement for high stiffness-to-weight ratios, the aerospace industry also required highly dimensionally stable materials. Stability was required in the aerospace industry because often 90% of the material from a billet was removed during machining, and if the alloy was not stable, massive deformations of the part could occur. Thus despite the fact that many aerospace alloys (e.g., aluminum) had higher coefficients of thermal expansion than other traditional machine tool materials, the benefits of decreased weight and ease of fabrication were considered important for this application, and an aluminum alloy was thus chosen for the primary structural material.

f. *Sensors and control.* Sensors for measuring linear motion of coordinate measuring machines typically consist of linear scales attached to the structure near the linear bearings. The use of linear scales is one of the most accurate methods for measuring linear distance. Greater accuracy can only be attained with the use of laser interferometers. Accordingly, in the 1970s, Sheffield chose to develop its own linear scale technology in order to gain a competitive cost advantage over their competitors who purchased scales from other companies. This R&D effort was successful to production of linear scales, which proved to be as accurate as other "off the shelf" scales, but at a greatly reduced cost.

In addition to the use of high-accuracy linear scales, other methods for decreasing the cost of the machine while maintaining accuracy were sought. This led to the decision to map geometric errors in the machine with a laser interferometer and incorporate the error maps into software error correction routines. This new approach to increasing machine accuracy had been successfully implemented in a laboratory situation on a precision coordinate measuring machine at the National Institute of Standards and Technology (formally the National Bureau of Standards), and was deemed feasible for adaptation to industrial use. The use of error maps also allowed for easier field maintenance and repair of the machine.

g. *Safety.* Safety is always of prime importance in any design. Since the operator sometimes works in the measuring volume to guide the measurement probe, all potential pinch points had to be identified and then shielded or eliminated. To prevent the operator from being struck by a CMM moving under computer control, electronic curtains of light could be installed around the machine. If the operator passed through the curtain of light, the machine would shut down or slow down. When the operator was in the work volume during manual operation, the same factors which allowed the machine to be moved easily by hand would also allow the operator to stop the machine from striking him.

h. *Ergonomics.* Plastic covers would prevent the operator from radiating body heat to the machine, protect the operator from the machine, keep airborne dirt off the bearing ways, and provide a more futuristic-looking machine for the operator to use. If the machine looked elegant, then some of the pride spent making the machine might wear off on the operator. The overall design of the structure was also required to be as open as possible to allow easy access to the work volume. Similarly, all control panel operations had to be user friendly. Much experience in all these areas was available as a result of user feedback on previous designs. For the repair and service person, the ability to correct errors in mechanical motion with laser measurement and software-based error corrections would also prove invaluable.

i. *Production.* Traditionally, machines with the highest accuracy were manufactured using hand finishing techniques. However, because of the anticipated great number of machines that would be sold, and the decreasing availability of skilled craftspeople capable of hand finishing machines, it was decided that only tolerances that grinding could achieve would be specified. Final accuracy would be achieved using software-based error mapping and compensation techniques.

j. *Transport.* Most machines could be transported and installed in one piece. However, if disassembly of a large model was required, error mapping techniques could be implemented on-site in a customer's plant after the machine was reassembled.

k. *Maintenance.* The air bearings proposed to support the linear axes' structures were somewhat self-cleaning in that they would blow dust off the ways. However, airborne oils and moisture could cause a heavy film to build up on the ways. This layer could in turn cause the bearings to drag. Past experience showed that air bearings have an averaging effect over surface irregularities; thus once an oil film had been removed, bearing performance could usually be restored.

l. *Costs.* If the competitive analysis showed that a significant gain could be made over existing technology, then any reasonable cost would justify the development. This type of cost analysis is proprietary to the company, and thus is not presented in this case study.

m. *Schedules.* There is an old saying: "No matter what you think of, someone else will also think of it at the same time"; thus it was important to complete the design as quickly as possible.

1.6.2 State of the Art Technology Assessment (1984)

The fixed table cantilever coordinate arm measuring machine shown in Figure 1.6.2 has the following characteristics:

- Lightweight moving mass.
- Fast measuring speed.
- Limited Y axis travel.
- Limited Z axis travel.
- Easy access to workpiece from three sides (high throughput).
- Good accessibility for part loading.
- Restricted to relatively light parts because the table deflects under the part's weight.
- Cantilevered design can have a low natural frequency and large Abbe errors can occur.

Moving bridge CMMs, such as shown in Figure 1.6.3, could be built larger than cantilevered machines and had higher natural frequencies. Their design, however, presented other problems unique to the structure. Characteristics of this type of CMM are:

- Operator access open on three sides with bridge to the rear.
- Large Y axis range available.
- Part's width cannot exceed opening of bridge.
- Part's weight limited because the structure deforms under the weight of the part.
- Problems with the outer leg "walking" because the bridge cannot be driven through the center of gravity without the addition of a costly overhead drive system suspended above the Y axis, or a second (slave) actuator on the outer leg.
- Bridge bearings are typically not preloaded, which decreases the speed at which the machine can operate.

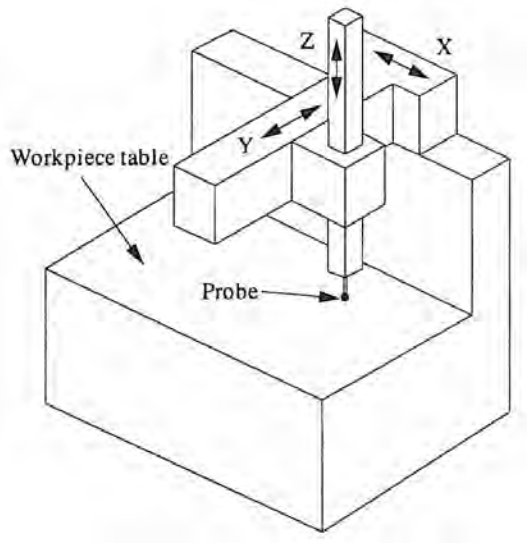


Figure 1.6.2 Fixed table cantilever arm CMM.

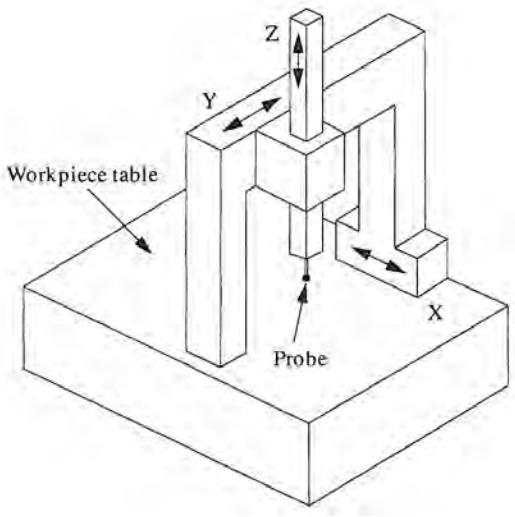


Figure 1.6.3 Moving bridge CMM.

To alleviate the problem of the walking bridge, the column-type fixed bridge moving table design evolved. It has the following characteristics:

- Very accurate system, provided that the part's weight is limited.
- Moving table axis can be made very massive, to support heavy parts.
- Measurement time is slow because of massiveness of the table.
- Part's width cannot exceed opening of the bridge.

In an effort to increase the work volume, the column CMM evolved as shown in Figure 1.6.4. The vertical axis was isolated from the planer axes to reduce the effect of Abbe errors, and the open table design allowed for larger parts to be put on the table.[42] Although table deformations still affected accuracy, this type of CMM has the following characteristics:

- Very accurate system, although the open C section is prone to thermal gradients, causing the column to arch back and induce large Abbe errors.
- Massive table can support large parts, but slows the measuring time. Accuracy is still a function of the weight of the part and how it deforms the machine.
- Part height is limited because of the open C column.
- Requires a highly skilled operator to coordinate motion of table and probe.
- Expensive machine to manufacture.

Designs took a semifull circle with the development of the moving ram horizontal arm CMM shown in Figure 1.6.5. The horizontal arm can probe deep into horizontal recesses in parts. The machine has the following characteristics:

- Medium accuracies for large part sizes.
- Excellent probe penetration into the side of a part.
- Limited vertical accessibility for most parts.
- Accuracy still a function of part weight.
- Cantilevered design leads to lower natural frequencies and is susceptible to large Abbe errors.

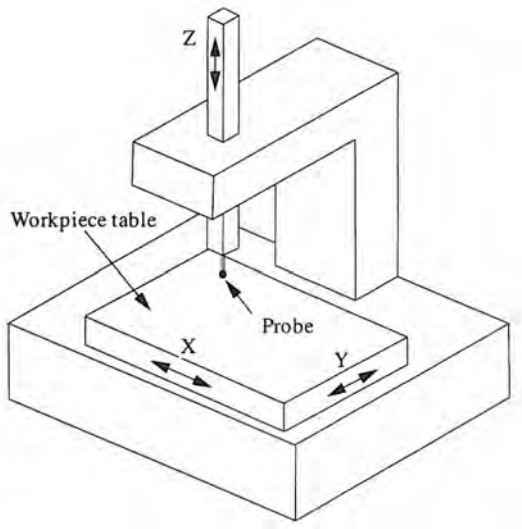


Figure 1.6.4 Column-type CMM.

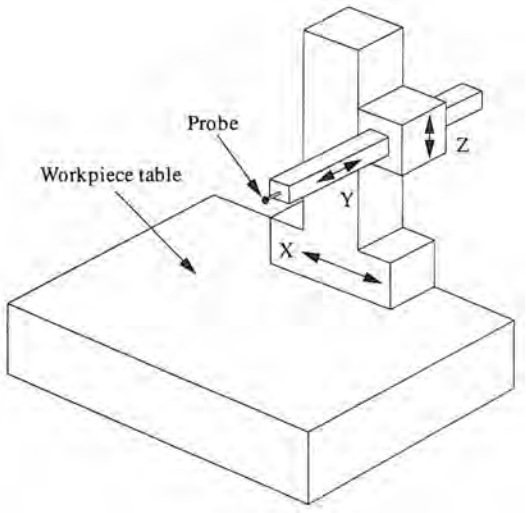


Figure 1.6.5 Moving ram horizontal arm CMM.

A variation of the moving ram horizontal arm CMM evolved as the moving table horizontal arm coordinate measuring machine shown in Figure 1.6.6. It has the following characteristics:

- Measures large parts on a relatively small machine, but accuracy is still a function of part weight.
- Excellent penetration of probe into the side of a part.
- Limited accessibility from top and bottom of part.
- Cantilevered arm subject to horizontal beam vibrations (common form of excitation from the floor).

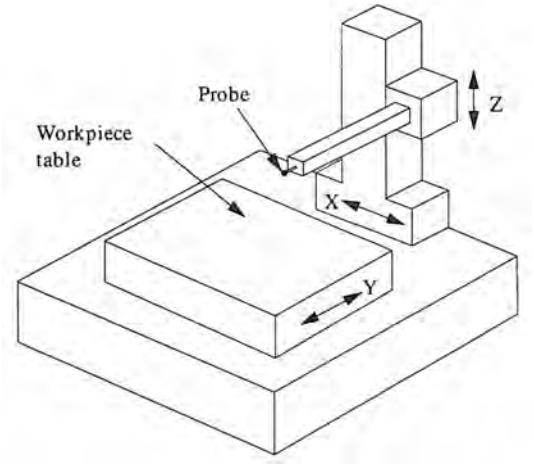


Figure 1.6.6 Moving table horizontal arm CMM.

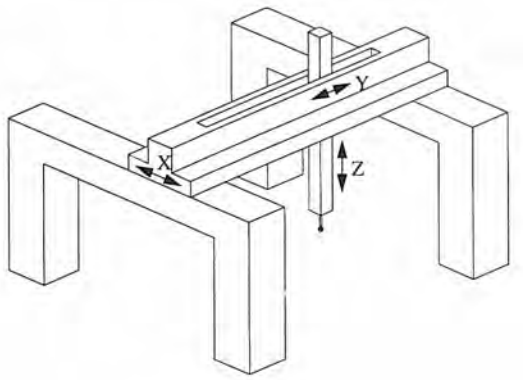


Figure 1.6.7 Gantry-type CMM.

All designs discussed so far were part-weight limited because the surfaces the parts rested on were in close physical contact with the measuring frame. The weight of the part caused static deflections in the structure, which decreased accuracy. This led to the evolution of the gantry-type CMM shown in Figure 1.6.7. Gantry-type CMMs have the following characteristics:

- Excellent for large axis travels.
- Physical machine is large relative to part size and requires massive construction to minimize bending deformations in beams supporting the axes.
- Operator has free access to all parts of larger machines.
- On a practical basis, the design is limited to large machines.
- As long as the foundation supporting the part is isolated from the foundation that supports the measuring system's legs, the weight of the part will not affect the accuracy.
- Ideal application for the use of a three-dimensional tracking laser interferometer system.

After careful analysis of other CMM designs, it was concluded that to maximize usefulness, the measuring system should be isolated from the structural system that supports the part. Furthermore, a ring-type structure appeared to be the most stable, and it should be driven through its center of percussion to avoid walking problems. These two requirements seemed to be a natural combination that ultimately led to the design of the ring bridge Apollo CMM. The design evaluation will be discussed in more detail later on, but is shown here for completeness of the survey of different types of CMMs. The ring bridge CMM concept is shown in Figure 1.6.8 and has the following characteristics:

- Optimized stiffness-to-weight ratio, resulting in high static and dynamic stiffness.
- Ring bridge is driven through the vertical plane of symmetry.
- All air bearings are preloaded, and therefore well suited for high-speed use.
- Worktable is isolated from the measuring system and base way surfaces; thus accuracy is insensitive to part weight.
- Part width cannot exceed width of bridge.
- Large Y axis range is practical.
- Operator has open access on three sides with ring to the rear.
- Bridge can be moved to the extreme end of table so that a part can be loaded using an overhead crane.

As a result of this survey (and the development of the ring bridge concept), the types of coordinate measuring machines described previously were categorized and evaluated with respect

[42] For an informative design case study of this type of CMM, see J. B. Bryan and D. L. Carter, "Design of a New Error-Corrected Coordinate Measuring Machine," Precis. Eng., Vol. 1, No. 3, 1979, pp. 125–128.

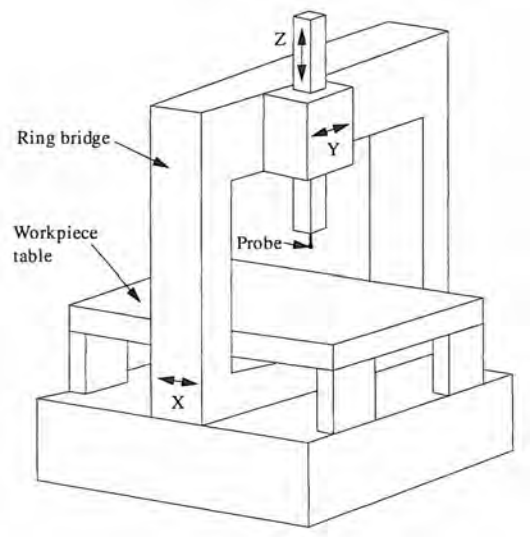


Figure 1.6.8 Ring bridge CMM.

to their competitive analysis quotient. From the CA values determined,[43] it became apparent that an all-out design effort on the ring bridge design had the potential to be a very profitable venture. Based on this survey of existing machines, it was also determined that the earlier specifications for size were reasonable and correct. The accuracy goal of the machine was specified at 0.012 mm (0.0005 in.) for a 400 mm ball bar[44] anywhere in the work volume. This would suit the needs of most manufacturers, and would be attainable at a reasonable cost if software-based error correction techniques were used.

Technical Evaluation of the Apollo's Predecessor

The predecessor to the Apollo CMM was the 1810/1820 series fixed table cantilever arm CMM manufactured by Sheffield. The smaller 1800 series would still be offered after the introduction of the Apollo, which was meant to measure much larger parts. An 1810/20 series upgrade feasibility study was thus intended to study the feasibility (performance versus cost) of improving the throughput and accuracy of Sheffield's 1810/20 series vertical arm machines and to help guide development of the Apollo. The 1810/20 series upgrade included the evaluation of mechanical and electrical system positioning hardware, the controller hardware and software, and a system level analysis of the servo position control loops.

The mechanical linkages for the X and Y axes consisted of a traction-type leadscrew shown, for example, in Figure 10.8.18. The traction drive rollers were arranged to ride on a 25.4 mm (1 in.) drive shaft at a 9.04° pitch, which resulted in 12.7 mm (1/2 in.) travel for each revolution of the drive shaft. The traction drive was engaged by air pressure, which controlled the axial force (and axis acceleration) possible through the drive. This allowed the nut to be uncoupled from the drive axis for ease of movement manually. The X and Y axis axial acceleration was thus limited by slippage of the nut friction drive. The axial speed of the X and Y axes was limited by the maximum critical speed of drive shaft (i.e., the shaft "whip" frequency).

The traction nut drive used for the X and Y axis actuators had the following advantages: (1) minimal backlash; (2) easy mechanical coupling ratio variability (via roller pitch adjustment); (3) easy preload adjustability (by air pressure) to compensate for wear; (4) low stiction; (5) medium rigidity; (6) common parts (as presently configured) for X and Y axes; and (7) releasable drive for manual operation. However, it had the following disadvantages: (1) poor motor coupling ratio, resulting in low motor/load response time; (2) low-velocity limits for long-axis travels due to critical shaft speed limits; and (3) varying friction levels meant variability in maximum acceleration.

[43] These values cannot be provided here due to proprietary concerns.

[44] A ball bar is a bar with round balls on the ends. The bar is clamped in the middle and the ends are probed to determine the length of the bar. See ANSI B-89.1.12. Also see J. B. Bryan, "A Simple Method for Testing Measuring Machines and Machine Tools," Precis. Eng., Vol. 4, No. 2, 1982, pp. 61–69.

The Z axis drive used a flat steel band to couple the Z axis shaft to a motor-driven pulley mechanism. An air piston counterbalance mechanism reduced the amount of torque on the motor. No practical limits for velocity or acceleration existed in the Z axis linkage for the range of acceleration and velocities considered. The Z axis drive/counterbalance had the following advantages: (1) integral pneumatic counterbalance mechanism, (2) simplicity, (3) low stiction, (4) no backlash, and (5) releaseable drive for manual operation. It had the following disadvantages: (1) Potentially low rigidity, due to steel tape drive, and (2) poor motor coupling ratio. Note that end-of-travel stops for all three axes were provided for by rubber bumpers. The maximum velocities for safe deceleration travel distance were designed to be 130 cm/s (5 in./s) for all three axes.

Due to the high mismatch of reflected load inertia and the motor inertia on all axes, different transmission mechanisms for all axes were sought. Matching inertias would help to reduce servo settling times. It was also recommended that the cost aspect of both drive mechanisms be reviewed. In addition, the low critical shaft speed of the traction nut drive could be addressed by a larger diameter drive shaft, a parallel carriage for damping, a higher pitch screw, or a middle support point.

1.6.3 1.6.3 Design Evolution of the Apollo Ring Bridge CMM

The Apollo CMM project began with the development of conceptual designs based on both traveling and fixed bridge configuration designs. The fixed bridge configuration was the best conceptual concept for meeting the high accuracy requirements and high dynamics for fast servo operation. A conventional traveling bridge configuration would have been the best choice for the lower-accuracy, small manual machines. After much reflection and group discussion, an individual brainstorming session in the shower one morning then led to the development of the ring bridge concept shown in Figure 1.6.8. It had the following general characteristics:

- Optimized stiffness-to-weight ratio resulting in high static and dynamic stiffness.
- Preloaded bearings.
- Very repeatable and stable, which is required for effective software-based error correction.
- Worktable is isolated from the measuring system, thus part weight does not affect accuracy.
- Ring bridge can be driven close to its center of gravity to prevent walking.
- Part width cannot exceed the bridge opening.
- Large Y axis range is possible.
- Operator access open on two sides, and limited on the other sides.

The initial ring bridge concept shown in Figure 1.6.9 had the following drawbacks:

- “V” ways would be difficult to manufacture and inspect.
- Ring’s support bearings were cantilevered and reduced ring stiffness.
- Workpiece loads could affect the bearing ways because of the weak unsupported legs.
- Vertical uprights of the ring were tall and had an effect on machine stiffness.
- Difficult to install the ring structure around the base legs.
- Holes for forklift access had to be incorporated in the base, which would weaken the base.
- Due to packaging problems, the bridge drive actuator could not be centered.
- Ring structure close to the floor created a potential pinch point.
- Machine cables could not be efficiently packaged in the base and thus would require exterior cable loop carriers.

The second ring bridge concept was the final design and is shown in Figure 1.6.1. It has the following characteristics:

- Square ways are easy to manufacture and inspect.
- Bearing ways directly support the lower member of the ring, which made for a stiffer structure.
- Workpiece loads directly transferred to the floor through rigid members that cause no deformation in the bearing ways or measurement structure geometry.
- Vertical ring uprights are shorter than in the first ring bridge design, which helped to stiffen the ring.
- Ring could easily be assembled onto the completed base.

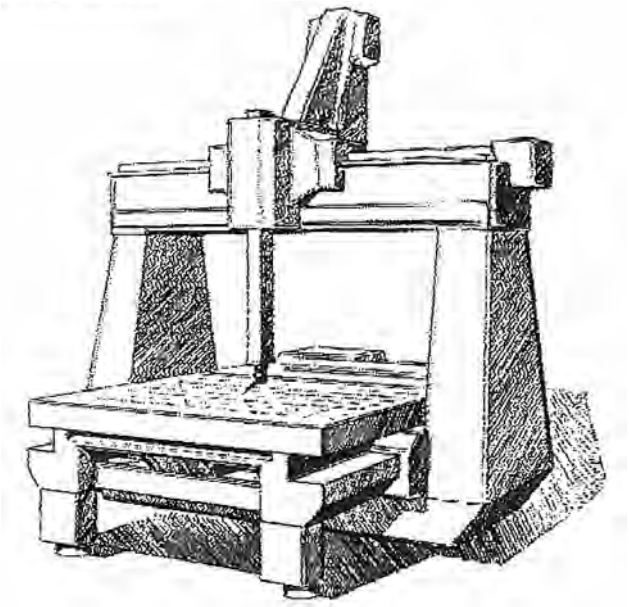


Figure 1.6.9 Initial ring bridge concept. (Courtesy of Sheffield Measurement, a Cross & Trecker Company.)

- Forklift access was provided under the base, negating the requirement for forklift holes, thus resulting in a stiffer base structure.
- Bridge drive was near its center.
- Pinch point between ring structure and the floor was eliminated.
- Machine cables could be packaged on top of the base using simple sliding loops.
- Center of gravity of the entire machine was low due to the low base position.

The ring bridge was thus selected over the fixed and conventional traveling bridge designs because it offered a technological "leapfrog" concept: good dynamics for fast servo operation, high accuracies with software error correction, and the same configuration for both manual and servo machines, resulting in commonality of parts and cost benefits. In addition, releasable drives for manual operation could be added to servocontrolled machines in order to increase market competitiveness. The ring bridge's main feature was the ability to decouple the structural and measuring systems so that accuracy would not be a function of part weight. In addition, software-based error correction techniques and aerospace alloys would help provide a distinctive competitive edge.

The ring bridge concept presented a great opportunity to make a significant advance in the state of the art of CMM design. The seemingly radical departures from traditional machine tool manufacturing techniques (e.g., use of aluminum, and error mapping) yielded a simpler, more accurate, easier-to-manufacture product that was also more difficult for other manufacturers to copy.

1.6.4 Detailed Design of the Apollo CMM

In developing the detail design of a machine such as the Apollo coordinate measuring machine, it was important to consider three critical areas: sources of mechanical error, mechanical systems, and sensor and electronic systems.

1.6.4.1 Sources of Error

Sources of mechanical error include machine geometry, machine dynamics, thermal effects, and workpiece effects. Each of these areas is discussed below.

Machine Geometry

Errors in machine geometry occurred in all three axes and were composed of three translation and three rotational components per axis. In addition, there was an error associated with the orthogonality of the axes with respect to each other. Altogether, this causes 21 sources of geometric error in the machine. Fortunately, all these errors could be measured as a function of position of occurrence in the structure. Using a laser interferometer and other measuring techniques to make maps of the geometric errors within the structure, software based error correction algorithms could use these maps to compensate for geometric errors. This mapping, however, did not include temperature effects, and therefore the machine would be required to come to equilibrium in its environment before being error mapped. In addition to providing corrections for errors in the machine's geometry, the software would be required to compensate for probe tip offset errors caused by angular errors (i.e., pitch, yaw, and roll) in the machine's axes. Thus as long as the user input the probe geometry into the machine controller, effects of probe offsets could also be compensated for.

Machine Dynamics

The primary dynamic errors in the machine were due to (1) the position of the moving probe tip relative to the machine's linear sensors as the probe tip made contact, and (2) vibration from within the machine, or transmitted to the machine from the floor. The first effect could be predicted with some certainty by electronically measuring the time delay between the time when the probe actually made contact and the time the machine received the signal that the probe has contacted the part. This could only be done using a test part that was placed a precise distance away from a known starting point. After the machine moved the known distance, the time delay could be evaluated as the time the probe triggered. This would give a measure of the accuracy as a function of speed of the machine. In practice, this error was not great because the probe could signal the controller as quickly as 1 μs after contact was made. The mechanical error was thus the product of the delay time conversion and the speed at which the CMM was moving. Note that it was important to maintain the capability of the machine to probe while moving rapidly. If measurements could only be made after the machine had slowed way down, throughput would have been severely degraded.

The second effect, vibration from the machine and the floor, was difficult to predict and correct for. This is where careful finite element modeling of the structure helped to identify resonances in the structure. This allowed structural deficiencies to be corrected before the machine was built, and allowed specifications to be made for allowable floor vibration levels. It also enabled the design engineers to specify servoloop times and operating speeds. The controller could also be set to prevent electric motors in the CMM from operating at speeds which may have excited structural resonances.

Although they are often traceable and characterizable, dynamic errors are the most difficult to predict and compensate for. Making the machine as stiff and lightweight as possible helped to make it easier for the machine to achieve high-speed operation.

Thermal Effects

Most structural metals expand at a rate in the range 11 to 22 μm/m/C^{o} (steel and aluminum, respectively), so typical shop temperature variations ($\pm$5 C^{o} and up) can have a dramatic effect on the accuracy of a large machine, particularly if the rate of change is significant (greater than 3 C^{o}/hour). Hence inspection tasks are supposed to take place in a temperature controlled environment. The degree of accuracy of temperature control depends on the required accuracy to which the part is to be inspected. Note that if the part and the axis position measurement scales are made of materials with the same coefficient of thermal expansion, then as long as both are at the same temperature, an accurate inspection can take place. This, however, is not a good condition to hope for, as materials used in manufacturing are quite varied.

Software corrections for thermal errors are difficult to make because there are almost an infinite number of temperature gradient profiles that can exist in the machine from site to site. Errors due to temperature gradients are the worst kind, since they often cause warping of the structure, which introduces angular errors. Thus it was decided to specify allowable room temperatures in which the machine should be operated.

Position along each of the axes was to be measured using stainless steel measurement scales pinned at one end to the aluminum frame and secured along their length with adhesive tape. The adhesive tape had a low shear modulus, so differential expansion between the steel scales and alu-

minum frame would be allowed to occur. Thus the advantages of steel scales, which have a low coefficient of thermal expansion, and the advantage of aluminum, which has an order-of-magnitude better thermal conductivity than steel, could be combined in one machine.

The issue of different materials having different transient thermal response times was and is very difficult to address. It thus becomes very important to allow the to-be-measured part to come to thermal equilibrium in the measurement room before it is placed on the coordinate measuring machine. Allowing the part and machine to come to the same temperature is often referred to as *soaking*.

Workpiece Effects

The accuracy of most previous coordinate measuring machine designs was affected by the weight of the part to be measured, due to the fact that the weight of the part distorted the structure of the measuring system. The ring bridge design solved this problem by isolating the measuring system from the table that supported the part. Since all measurements on the part are relative to some datum on the part, once the part was placed on the measuring table, initial deflections of the table due to the weight of the part would not be important. The part, however, must be able to support its own weight without deforming significantly. The part must also be mounted so that it will not wobble. Thus the table only had to be sized not to deflect when loaded by parts that were incapable of supporting their own weight. Such parts are inherently lightweight; hence Apollo's table ended up being much lighter than one would expect on a "traditional" coordinate measuring machine. Naturally, this could become a sensitive point when the salesman is trying to convince an old-time inspector that the Apollo is actually more accurate than machines of a much more massive construction. Thus, educating the consumer about the structure and working principles of the Apollo would end up being a very important task.

1.6.4.2 Mechanical Systems Design

Besides the ring bridge structure, the probes, bearings, and actuators are vital components of the coordinate measuring machine's mechanical systems. The probes must be able to measure accurately on the fly, or far more stringent conditions would have to be placed on the rest of the machine. Proven reliable accurate probes were commercially available and were therefore not considered a problem area. The bearings had to be frictionless, repeatable, and preloaded. The actuators had to provide quick acceleration and deceleration and smooth motion; accuracy was not crucial because the probe could measure while the axes were moving.

The bearings had to be absolutely frictionless to eliminate distortion of the structure due to drag forces. In addition, the bearings had to be preloaded to allow the lightweight axes to be accelerated rapidly. The only type of bearings that simultaneously are truly frictionless and preloaded are hydrostatic and aerostatic bearings. The former use a cushion of high-pressure oil to float one structure above another. Using oil in a clean inspection environment, however, was not a desirable option. Aerostatic bearings use air instead of oil, but the lower viscosity of air requires closer mechanical tolerances between moving parts in order to keep airflow rates low and provide high bearing stiffness.

The technology required to manufacture high-accuracy aerostatic bearings had been known for some time; however, most bearings used orifices and had rather high flow rates. Orifice technology had the advantage of being well understood analytically. However, stiffness of orifice compensated bearings was usually less than that attainable with porous bearings. To decrease air consumption and increase stiffness, Sheffield made the decision to develop porous air bearing technology. Porous air bearing technology had been in use for decades but it was considered a difficult technology to master because it was thought to be difficult in production to control flow resistance accurately through a porous medium. However, if the effort proved successful, it would give Sheffield a good competitive advantage. They took the risk and it paid off well. The air bearings for the Apollo coordinate measuring machine are designed using in-house porous graphite pad technology that borrowed from technology developed at Oak Ridge National Labs. The porous graphite also had the advantage of not harming the ways if air pressure was lost. The bearings are preloaded by having one pad on one side of the way push against a pad on the other side of the way. This configuration proved to be stable to the submicron (microinch) level.

One of the key criteria for selection of the actuators was the ability to uncouple the drive train from the moving structure. This would allow the user to move the structure freely without having to backdrive the power train. Actuators considered for the Apollo drive train included rack and pinion, ballscrew, cogged belt, linear motor, traction drive, and chain drive, all of which are discussed in detail in Chapter 10. Their applicability to this problem is discussed briefly below.

In the past, rack-and-pinion and ballscrew drives had proven difficult to disengage and unforgiving when run into a rigid stop. On machine tools, the extremely heavy structure is built to withstand cutting forces and can take this kind of abuse, but CMMs cannot. In addition, these types of drives can be very accurate, but require periodic backlash adjustment. Finally, and perhaps most important, they require extremely careful alignment procedures to prevent the actuator from causing error motions in the bearings.

Cogged belts (i.e., timing belts) and sprockets are usually used to transmit torque from one shaft to another. However, a clamp can be used to attach the moving structure to the belt. As the drive sprocket rotates the belt, the structure is pulled along. When the operator wants to move an axis in a manual mode, he activates a switch which uncouples the belt from the structure. Should the belt try and move the structure past its allowable limit of travel, end-of-travel sensing would prevent further movement. The principal problem with cogged belts in this type of application, however, is their low stiffness in a stretch mode. In CMMs, however, there are no cutting forces to resist and measurements are made while the machine is still moving. Thus servo accuracy problems associated with low-stiffness actuators would probably not be a significant problem. To verify this, a test jig was built and it proved successful. Since the position feedback element directly reads the position of the structure, no loss of accuracy would occur on future moves. Thus low belt stiffness was judged not to be a fatal concern. Its main limiting effect was to limit the attainable servoloop bandwidth.

Dc linear electric motors were also potential candidates for this application. However, they were very expensive compared to cogged belt drives. Their positioning accuracy was potentially higher, but since the machine would be capable of measuring on the fly anyway, higher actuator positioning accuracy did not outweigh their higher cost. The major disadvantage of linear motors is they are major heat sources within the structure and they do not provide any means for gear reduction, which conventional rotary motors can achieve easily and inexpensively. This precludes matching the motor and load inertias. Various friction (traction) drives were also considered. Previous coordinate measuring machines built by Sheffield used these types of drives, but the mechanism to engage and disengage the drive from the flat drive rail or shaft was deemed too costly.

A chain drive would operate on the same principle that the cogged belt system would. However, it is much more difficult to clamp and unclamp an axis to the chain. Chain drives also had a history of jerky (cogging) motion because the links are differential elements, whereas the cogged belt's fibers cause it to bend around drive sprockets in a continuous manner. Lubrication requirements also made chain drives a poor choice for this application.

The best choice for this application was deemed to be the cogged belt system. Note that other manufacturers have decided that friction drives are the best actuator to use. Because one manufacturer makes a design decision does not mean that it is the best one for all cases.

1.6.4.3 Sensor and Electronic Systems

As discussed above, stainless steel measurement scales manufactured in-house would be used on the Apollo design because of their proven accuracy, reliability, and low cost. The sensors would be placed near the bearings and Abbe errors at the probe minimized by mapping the structure. Coordinate measuring machines are not subject to flying chips, splashing coolant, or dripping lubricating oil; thus the task of sensor location and protection would be much easier than on most machine tools.

The electronic control systems would all have to be custom designed if the machine were to be able to register precisely the time at which the probe made contact with the part and then read the positions of the axes. The architecture would be based on a 16-bit microprocessor with a math coprocessor and standard modular memory and communication boards. All software control routines would have to be written using double precision; hence the ultimate desire is to use a controller with 32-bit architecture. On future models of the Apollo CMM, it would be a straightforward task to provide a 32-bit controller.

1.6.5 Design Follow-Up

The design follow-up for this project has shown the Apollo CMM to be a very successful venture. Since the introduction of the Apollo CMM in 1986, the Apollo has sold very well and has set a new standard for CMM design which emphasizes throughput as well as accuracy. The use of aluminum as the primary structural material in the design required some adjustment in the manufacturing division, but this can be considered a normal process of adjustment to technology. Now that they have adjusted to machining aluminum, there is no desire to switch back to cast iron.

Chapter 2
Principles of Accuracy, Repeatability, and Resolution

If in other sciences we should arrive at certainty without doubt and truth without error, it behooves us to place the foundations of knowledge in mathematics.

Roger Bacon

2.1 INTRODUCTION[1]

An engineer who has taken a couple of design courses, is familiar with a few machine designs, and has an entire filing cabinet of catalogs is not necessarily a good machine design engineer.[2] A good machine design engineer understands and has developed an intuitive feeling for the factors that affect machine performance. Equally important, the design engineer must also understand the basic physics that characterize a machine component or system. This knowledge is essential for the development of good designs and the proper selection of components.

The design of quality precision machines depends primarily on the ability of the design and the manufacturing engineers to predict how the machine will perform before it is built. Kinematics of a machine are easily tested for gross functionality using mechanism synthesis and analysis software. Wear rates, fatigue, and corrosion are often difficult to predict and control, but for the most part are understood problems in the context of machine tool design. Hence perhaps the most important factors affecting the quality of a machine are the accuracy, repeatability, and resolution of its components and the manner in which they are combined. These factors are critical because they affect every one of the parts that will be manufactured using the machine. Accordingly, minimizing machine cost and maximizing machine quality mandate predictability of accuracy, repeatability, and resolution. This allows the design engineer to optimize his choice of components and to specify manufacturing tolerances. Similarly, if during the manufacturing process a variance in procedure occurs, say due to a change in suppliers, it must be possible to quickly determine the effect of the substitute component on the machine's performance. These concepts also form the roots of statistical quality control.

Designing a machine that has good accuracy, repeatability, and resolution has often been considered a black art helped along by scientific principles.[3] As noted by Donaldson[4]: "A basic finding from our experience in dealing with machining accuracy is that machine tools are deterministic. By this we mean that machine tool errors obey cause-and-effect relationships, and do not vary randomly for no reason." Herein lies the key, for an aspect of design or manufacture will appear to be a black art only when the observer lacks the time or resources to use scientific principles to discover the true nature of the phenomena.

2.1.1 Accuracy, Repeatability, and Resolution

In the context of taking a deterministic approach to design, it is very important to understand the rules of the game. There are three basic definitions to remember with respect to how well a machine tool can position its axes: *accuracy, repeatability (precision),* and *resolution*. These terms can be represented diagrammatically by a marksman's target as shown in Figure 2.1.1.

Accuracy is the ability to tell the truth. Accuracy is the maximum translational or rotational error between any two points in the machine's work volume. As shown in Figure 2.1.1, accuracy can be represented as the difference between the root mean square radius of all the bullet holes in a

[1] A good general reference that discusses many of the topics in this chapter and more is Technology of Machine Tools, Vol. 5, Machine Tool Accuracy, Robert J. Hocken (ed.), Machine Tool Task Force. Available from U.S. Dept. of Commerce National Technical Information Service as Report UCRL-52960-5. There are numerous references on this subject, far too many to list here, but the Technology of Machine Tools provides a very good summary of the art and science of machine tool errors, as it was written by a task force of leading researchers.

[2] *"Nothing in education is so astonishing as the amount of ignorance it accumulates in the form of inert facts."* Henry Adams

[3] *"To most people, nothing is more troublesome than the effort of thinking."* James Bryce

[4] R. Donaldson, "The Deterministic Approach to Machining Accuracy," SME Fabricat. Technol. Symp., Golden, CO, Nov. 1972 (UCRL Preprint 74243).

target and the radius of the bull's eye. Linear, planar, and volumetric accuracy can all be similarly defined for a machine.

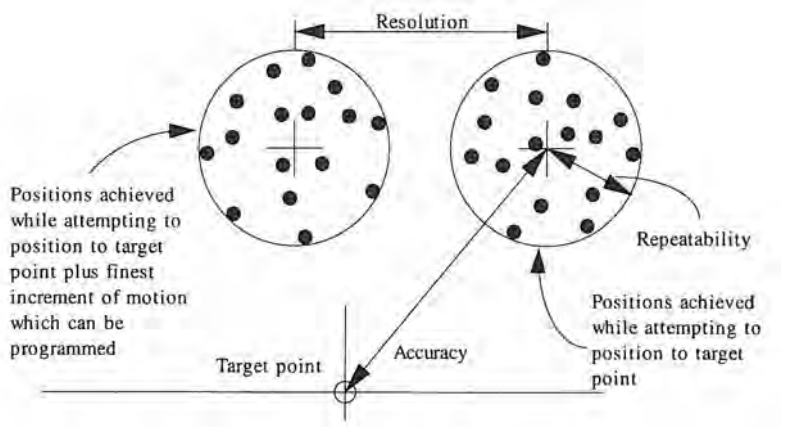


Figure 2.1.1 Defining accuracy, repeatability, and resolution.

Repeatability (precision) is the ability to tell the same story over and over again. Repeatability is the error between a number of successive attempts to move the machine to the same position. As shown in Figure 2.1.1, repeatability can be represented by the diameter of the circle, which contains N% of the bullet holes in a target. N is typically defined based on what appear to be random occurrences, as shown in Table 2.1.1. Bidirectional repeatability is the repeatability achieved when the point is approached from two different directions. This includes the effect of backlash in a leadscrew. For a set of N data points with a normal (Gaussian) distribution, the mean x_{mean} and the standard deviation σ are defined as

$$x_{mean} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.1.1}$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - x_{mean})^2} \tag{2.1.2}$$

The standard deviation is used in the determination of the probability of occurrence of an event in a system that has a normal distribution. Table 2.1.1 gives the percent chance of a value occurring within a number of standard deviations of its expected value.[5] One must be very careful not to confuse an offset of the mean with the allowed random variation in a part dimension. For example, precision molded plastic lenses for cameras have their molds hand finished to make the mean size of lenses produced equal to the nominal size required.

It is interesting to note, however, Bryan's[6] observation of the issue of using probabilistic methods to characterize repeatability: "The probabilistic approach to a problem is only a tool to allow us to deal with variables that are too numerous, or expensive to sort out properly by common sense and good metrology." One must not belittle probability, however, for it is a mathematical tool like any other that is available to the design engineer. Required use of this tool, however, might be an indication that it is time to take a closer look at the system and see if the system can be changed to make it deterministic and therefore more controllable. Often the key to repeatability is not within the machine itself, but in isolating the machine from variations in the environment.[7]

Resolution is how detailed your story is. Resolution is the larger of the smallest programmable step or the smallest mechanical step the machine can make during point-to-point motion. Resolution

[5] The equation for the computation of the percent chance was obtained from A. Drake, Fundamentals of Applied Probability Theory, McGraw-Hill Book Co., New York, 1967, p. 211. Values for $\Phi(k)$ were computed using Mathematica®. Another valuable reference to have is M. Natrella, Experimental Statistics, NBS Handbook 91.

[6] J. Bryan, "The Power of Deterministic Thinking in Machine Tool Accuracy," 1st Int. Mach. Tool Eng. Conf., Tokyo, Nov. 1984 (UCRL Preprint 91531).

[7] *"In designing an experiment the agents and phenomena to be studied are marked off from all others and regarded as the field of investigation. All others are called disturbing agents. The experiment must be so arranged that the effects of disturbing agents on the phenomena to be investigated are as small as possible."* James C. Maxwell

k	N % chance of occurrence
1.0	68.2689
2.0	95.4500
3.0	99.7300
4.0	99.9937
5.0	99.9999
6.0	100.0000

Table 2.1.1 Chance of a value falling within kσ of its expected value for random processes.

is important because it gives a lower bound on the repeatability that one could obtain if one really tried.

Although these definitions seem straightforward enough, how measurements are best made to determine them is sometimes a source of great debate. Of a primary concern is: What is the *certainty* of the measurements themselves used to characterize the accuracy, repeatability, and resolution of a machine, and what parameters (e.g., machine warm-up time) are these measurements themselves a function of? These issues will be addressed throughout Chapters 3 through 6.[8]

2.1.2 Amplification of Angular Errors

Perhaps the greatest sin in precision machine design is to allow an angular error to manifest itself in a linear form via amplification by a lever arm. Mathematically this error has a magnitude equal to the product of the lever arm's length and the sine of the angle. Hence this type of error can be referred to as a *sine error*. A *cosine error*, on the other hand, represents the difference between the distance between a point and a line, and the distance along the measurement path between a point and a line. Note that by definition, the former is along a line orthogonal to the line.

With respect to dimensional measurements, in the late 1800s, Dr. Ernst Abbe noted *"If errors in parallax are to be avoided, the measuring system must be placed coaxially with the axis along which displacement is to be measured on the workpiece."* The Abbe principle can be visualized by comparing measurements made with a dial caliper and a micrometer as shown in Figure 2.1.2. The dial caliper is often used around the shop because it is easy to use; the head slides back and forth to facilitate quick measurement. However, note that the measurement scale is located at the base of the jaws. When a part is measured near the tip of the jaws, the jaws can rock back slightly, owing to their elasticity and imperfections in the sliding jaw's bearing. This causes the caliper to yield a slightly undersize measurement. The micrometer, on the other hand, uses a precision measuring device located in line with the part dimension, so there is no Abbe error. Both instruments are sensitive to how hard the jaws are closed on the part. A micrometer often has a torque limiting adjustment to provide a very repeatable measuring force. It is impossible to overstress the importance of Abbe errors.

This principle extends to locating bearing surfaces far from the workpiece area of the machine tool.[9] Errors in the bearing's motion can be amplified by the distance between the bearing and the workpiece, and transmitted to the workpiece. This can result in horizontal and vertical straightness errors and axial position errors. The same is true for the effects of all other types of errors on machine components.

2.2 Formulating the System Error Budget

To define the relative position of one rigid body with respect to another, six degrees of freedom must be specified. To further complicate matters, error in each of the six degrees of freedom can have numerous contributing components. In fact, considering all the interacting elements in a typical

[8] Also see A. T. J. Hayward, Repeatability and Accuracy, An Introduction to the Subject and a Proposed Standard Procedure for Measuring the Repeatability and Estimating the Accuracy of Industrial Measuring Instruments, Mechanical Engineering Publications, New York, 1977.

[9] See J. B. Bryan, "The Abbe Principle Revisited-An Updated Interpretation," Precis. Eng., Vol. 1, No. 3, 1989, pp. 129–132. An extension of the Abbe principle to this type of situation is referred to as the *Bryan principle*, which is discussed in Section 5.2.1 along with a discussion on where to mount sensors for various types of measurements.

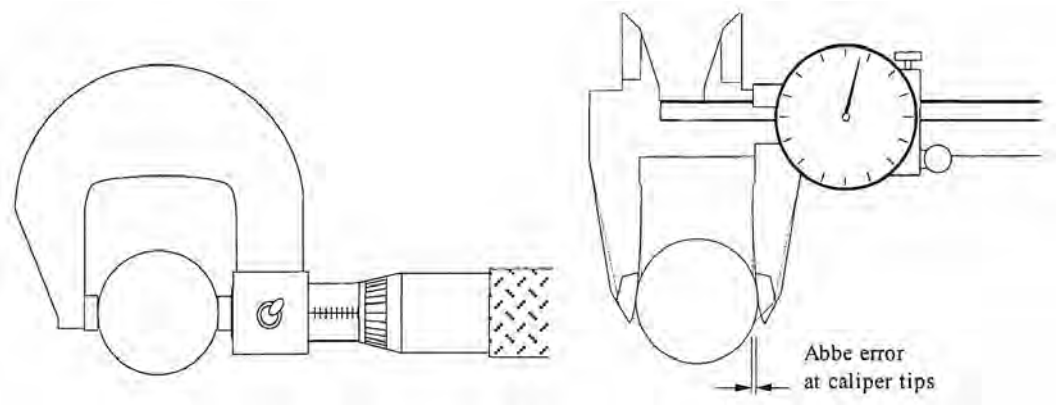


Figure 2.1.2 Abbe error illustrated through the use of a dial caliper and a micrometer.

machine tool, the number of errors that must be kept track of can be mindboggling. Therefore, the best way to keep track of and allocate allowable values for these errors is to use an *error budget*. An error budget, like any other budget, allocates resources (allowable amounts of error) among a machine's different components.[10] The goal is to allocate errors such that the ability of any particular component to meet its error allocation is not exceeded. Minimizing the amount and complexity of developmental work needed to meet the requirements of the error budget then becomes the main goal of distributing and redistributing allowable errors among components. An error budget can be viewed as an engineering management tool to help control and guide the design process. It is also a tool to help predict how the final design will behave. Like a CPM chart, an error budget is a dynamic tool that must be continually updated during the design process.

An error budget is formulated based on connectivity rules that define the behavior of a machine's components and their interfaces, and combinational rules that describe how errors of different types are to be combined. The first step in developing an error budget is to develop a kinematic model of the proposed system in the form of a series of homogeneous transformation matrices (HTM). The next step is to analyze systematically each type of error that can occur in the system and use the HTM model to help determine the effect of the errors on the toolpoint position accuracy with respect to the workpiece. The result is a list of all endpoint error components, their sources, and amplification-at-the-toolpoint factors (called the *error gains* or *sensitivities*). Different combinational rules can then be applied to yield upper and lower bound estimates of the total error in the machine.

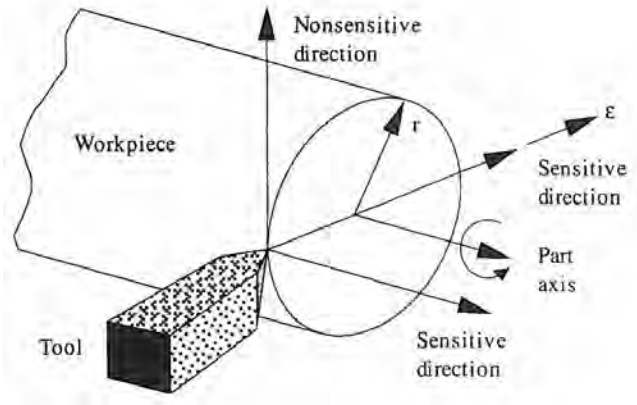


Figure 2.2.1 Illustration of sensitive directions.

When using an error budget to check the designed versus specified accuracy of the toolpoint or workpiece location with respect to the workpiece, it is important to note which are the *sensitive*

[10] See R. Donaldson, "Error Budgets," in Technology of Machine Tools, Vol. 5, Machine Tool Accuracy, Robert J. Hocken (ed.), Machine Tool Task Force.

directions of the machine. For example, as illustrated in Figure 2.2.1, an error motion ε of a tool tangent to the surface of a round part of radius r in a lathe results in a radial error in the workpiece of magnitude ε^2/r, which is much smaller than ε. Sensitive directions can be *fixed*, such as when the tool is stationary and the part is moving (e.g., a lathe), or *rotating*, such as when the tool is rotating and the part is fixed (e.g., a jig borer). Effort should not be expended to reduce errors that are already inconsequential. However, in general an error budget should be formulated that includes *all errors,* to avoid accidentally discarding a sensitive error. When the final list is tabulated, errors in nonsensitive directions can then be ignored.

2.2.1 Homogeneous Transformation Matrix Model of a Machine[11]

In order to determine the effects of a component's error on the position of the toolpoint or the workpiece, the spatial relationship between the two must be defined. To represent the relative position of a rigid body in three-dimensional space with respect to a given coordinate system, a 4x4 matrix is needed. This matrix represents the coordinate transformation to the reference coordinate system (X_R Y_R Z_R) from that of the rigid body frame (X_n Y_n Z_n), and it is called the *homogeneous transformation matrix* (HTM). The first three columns of the HTM are direction cosines (unit vectors i, j, k) representing the orientation of the rigid body's X_n, Y_n, and Z_n axes with respect to the reference coordinate frame, and their scale factors are zero. The last column represents the position of the rigid body's coordinate system's origin with respect to the reference coordinate frame. P_s is a scale factor, which is usually set to unity to help avoid confusion. The pre-superscript represents the reference frame you want the result to be represented in, and the post-subscript represents the reference frame you are transferring from:

$$ {}^{R}\mathbf{T}_{n} = \begin{bmatrix} 0_{ix} & 0_{iy} & 0_{iz} & P_x \\ 0_{jx} & 0_{jy} & 0_{jz} & P_y \\ 0_{kx} & 0_{ky} & 0_{kz} & P_z \\ 0 & 0 & 0 & P_s \end{bmatrix} \tag{2.2.1} $$

Thus the equivalent coordinates of a point in a coordinate frame n, with respect to a reference frame R, are

$$ \begin{bmatrix} X_R \\ Y_R \\ Z_R \\ 1 \end{bmatrix} = {}^{R}\mathbf{T}_{n} \begin{bmatrix} X_n \\ Y_n \\ Z_n \\ 1 \end{bmatrix} \tag{2.2.2} $$

For example, if the $X_1Y_1Z_1$ coordinate system is translated by an amount x along the X axis, the HTM that transforms the coordinates of a point in the $X_1Y_1Z_1$ coordinate frame into the XYZ reference frame is

$$ {}^{XYZ}\mathbf{T}_{X_1Y_1Z_1} = \begin{bmatrix} 1 & 0 & 0 & x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.3} $$

If the $X_1Y_1Z_1$ coordinate system is translated by an amount y along the Y axis, the HTM that transforms the coordinates of a point in the $X_1Y_1Z_1$ coordinate frame into the XYZ frame is

$$ {}^{XYZ}\mathbf{T}_{X_1Y_1Z_1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.4} $$

[11] The HTM representation of structures has existed for many decades. See, for example, J. Denavit and R. Hartenberg, "A Kinematic Notation for Lower-Pair Mechanisms Based on Matrices," J. Appl. Mech., June 1955. Perhaps the most often referenced work with respect to its application to manufacturing tools is R. Paul, Robot Manipulators: Mathematics, Programming, and Control, MIT Press, Cambridge, MA 1981.

If the $X_1Y_1Z_1$ coordinate system is translated by an amount z along the Z axis, the HTM that transforms the coordinates of a point in the $X_1Y_1Z_1$ coordinate frame into the XYZ frame is

$$XYZ\mathbf{T}_{X_1Y_1Z_1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.5}$$

If the $X_1Y_1Z_1$ coordinate system is rotated by an amount θ_x about the X axis, the HTM that transforms the coordinates of a point in the $X_1Y_1Z_1$ coordinate frame into the XYZ frame is

$$XYZ\mathbf{T}_{X_1Y_1Z_1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x & 0 \\ 0 & \sin\theta_x & \cos\theta_x & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.6}$$

If the $X_1Y_1Z_1$ coordinate system is rotated by an amount θ_y about the Y axis, the HTM that transforms the coordinates of a point in the $X_1Y_1Z_1$ coordinate frame into the XYZ frame is

$$XYZ\mathbf{T}_{X_1Y_1Z_1} = \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.7}$$

If the $X_1Y_1Z_1$ coordinate system is rotated by an amount θ_z about the Z axis, the HTM that transforms the coordinates of a point in the $X_1Y_1Z_1$ coordinate frame into the XYZ frame is

$$XYZ\mathbf{T}_{X_1Y_1Z_1} = \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 & 0 \\ \sin\theta_z & \cos\theta_z & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.8}$$

For axes with simultaneous combinations of these motions, these HTMs can be multiplied in series to obtain a single HTM for the axis, as discussed below. Extreme care must be taken, however, with axes subject to large multiple rotations because the order of rotation becomes very important. You can see this for yourself by experimenting with a book. Rotate a book 90° about various axes in different orders and observe which side faces up after an equal number of rotations.

Machine structures can be decomposed into a series of coordinate transformation matrices describing the relative position of each axis and any intermediate coordinate frames that may assist in the modeling process, starting at the tip and working all the way down to the base reference coordinate system (n = 0). If N rigid bodies are connected in series and the relative HTMs between connecting axes are known, the position of the tip (Nth axis) in terms of the reference coordinate system will be the sequential product of all the HTMs:

$$R\mathbf{T}_N = \prod_{m=1}^{N} {}^{m-1}\mathbf{T}_m = {}^0\mathbf{T}_1\ {}^1\mathbf{T}_2\ {}^2\mathbf{T}_3 \ldots . \tag{2.2.9}$$

Often, however, it can be difficult to determine how a part modeled as a rigid body actually moves; thus care must be taken when evaluating the error terms in the HTMs of systems with multiple contact points. Nonserial link machines (e.g., a four bar linkage robot) require a customized formulation to account for interaction of the links.

In a similar combinational method described by Reshetov and Portman,[12] the elements of the matrices may be mathematical functions and the entire representation is called the *form-shaping function* for the machine. This allows a closed-form mathematical representation of the machine's performance to be made. Although more elegant and conducive to some optimization studies, because of increasing complexity with the increase in the number of axes, a closed-form solution is often not practical for a complex machine and can lead to too simplified a model. The HTM method

[12] See D. Reshetov and V. Portman, Accuracy of Machine Tools, translated from the Russian by J. Ghojel, ASME Press, New York, 1988.

described here can readily accommodate any number of coordinate frames. Using a digital computer, mathematical functions describing errors and paths of motions are then readily made to correspond to elements in the HTM model.

Linear Motion Errors

Consider the case of an ideal linear motion carriage shown in Figure 2.2.2, with x, y, and z offsets of a, b, and c, respectively. Its HTM with respect to a parallel Cartesian reference frame is obtained by using Equations 2.2.3–2.2.5 in Equation 2.2.9. This is easily done in one's head, or by recognizing that pure translations affect only the last column of the HTM:

$$ {}^{R}\mathbf{T}_{n} = \begin{bmatrix} 1 & 0 & 0 & a \\ 0 & 1 & 0 & b \\ 0 & 0 & 1 & c \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.10} $$

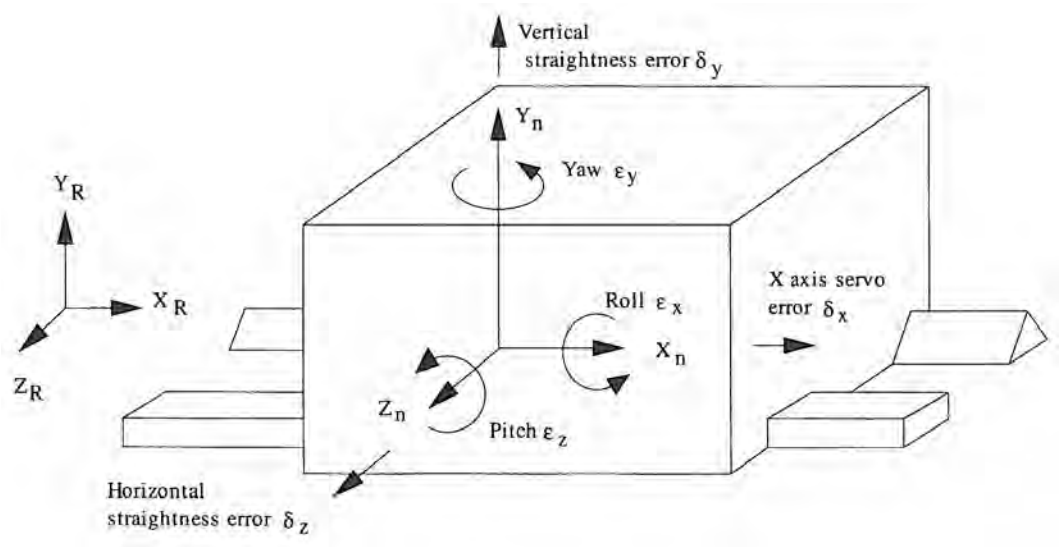


Figure 2.2.2 Motion and errors in a single-axis linear motion carriage (prismatic joint).

All rigid bodies have three rotational (ε_X, ε_Y, ε_Z) and three translational (δ_X, δ_Y, δ_Z) error components associated with their motion, as shown, for example, for a simple linear carriage in Figure 2.2.2. These errors can be defined as occurring about and along the reference coordinate system's axes, respectively. Often the errors will be a function of the position of the body in the reference frame. Remember, the elements of the matrix are the positions, given in the reference frame, of the unit vectors' tips.

For the linear carriage, the HTM that describes the effects of errors on carriage motion can be found by multiplying Equations 2.2.3 - 2.2.8 in series with error terms δ_X, δ_Y, δ_Z, ε_X, ε_Y, and ε_Z for x, y, z, θ_x, θ_y, and θ_z, respectively. The same result can be obtained in the following manner. Observing the right-hand rule, a rotation ε_X about the X axis (*roll*, as in roll over in bed) causes the tip of the Y-axis vector to move in the positive Z direction by an amount proportional to sin ε_X, and in the negative Y direction by an amount proportional to 1 - cos ε_X. Since the error terms are very small, at most on the order of minutes of arc, small-angle approximations are valid and will be used from now on in this context. Hence the element $o_{ky} = \varepsilon_X$. The roll error ε_X also causes the tip of the Z-axis vector to move in the negative Y direction, so that the element $o_{jz} = -\varepsilon_X$. A rotation ε_y about the Y axis (*yaw*, as in turn your head away from others when you yawn) causes the tip of the X-axis vector to move in the negative Z direction so that $o_{kx} = -\varepsilon_Y$, and causes the tip of the Z-axis vector to move in the positive X direction so $o_{iz} = \varepsilon_Y$. Similarly, a rotation ε_Z about the Z axis (*pitch*, as in pitch forward when you trip) causes the tip of the X-axis vector to move in the positive Y direction so $o_{jx} = \varepsilon_Z$, and causes the tip of the Y axis vector to move in the negative X direction so that $o_{iy} = -\varepsilon_Z$. Translational errors δ_X, δ_Y, and δ_Z directly affect their respective axes, but care must be taken

in defining them.[13] Having neglected second-order terms, the resultant HTM describing the error in position of the carriage with respect to its ideal position is

$$\mathbf{E}_n = \begin{bmatrix} 1 & -\varepsilon_Z & \varepsilon_Y & \delta_X \\ \varepsilon_Z & 1 & -\varepsilon_X & \delta_Y \\ -\varepsilon_Y & \varepsilon_X & 1 & \delta_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.11}$$

The first column describes the orientation of the carriage's X axis by defining the position of the tip of a unit vector, parallel to the carriage's X axis. Assuming that the base coordinates of the unit vector are in the reference frame at 0, 0, 0, the X, Y, Z coordinates of the X unit vector's tip are 1, ε_Z, -ε_Y. Remember, the pitch ε_Z caused the X axis to move in the positive Y direction, while the yaw ε_Y caused the X axis to move in the negative Z direction. The same logic applies to the formulation of the other columns. The actual HTM for the linear motion carriage with errors is thus ${}^R\mathbf{T}_{nerr} = {}^R\mathbf{T}_n\mathbf{E}_n$:

$${}^R\mathbf{T}_{nerr} = \begin{bmatrix} 1 & -\varepsilon_Z & \varepsilon_Y & a+\delta_X \\ \varepsilon_Z & 1 & -\varepsilon_X & b+\delta_Y \\ -\varepsilon_Y & \varepsilon_X & 1 & c+\delta_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.12}$$

Axis of Rotation Errors[14]

Consider the rotating body shown in Figure 2.2.3. Ideally, the body rotates about its axis of rotation without any errors; however, in reality the axis of rotation revolves around an axis of the reference coordinate frame with radial errors δ_X and δ_Y, an axial error δ_Z, and tilt errors ε_X and ε_Y. All of these errors may be a function of the rotation angle θ_Z. For a point in the spindle coordinate frame $X_nY_nZ_n$, one would first use the rotation angle θ_Z to transform the point into roughly the reference frame. Then, since the other error motions are small, the order of multiplication of their HTMs would then not be critical. The sequential HTM multiplication would thus have the form

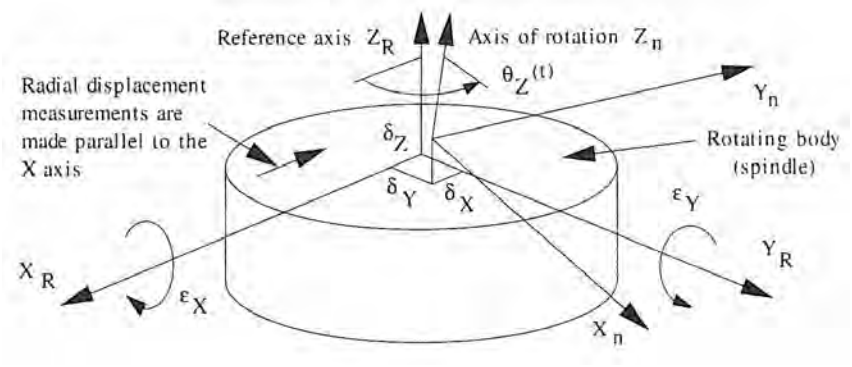


Figure 2.2.3 Motion and errors about an axis of rotation (revolute joint).

$${}^R\mathbf{T}_{nerr} = \prod_{i=3}^{8} \text{Eq. 2.2.i} = (\text{Eq. 2.2.3})(\text{Eq. 2.2.4}) \ldots (\text{Eq. 2.2.8}) \tag{2.2.13}$$

[13] See, for example, Section 5.7.2.

[14] Definitions used in this section were condensed from those provided in Axis of Rotation: Methods for Specifying and Testing, ANSI Standard B89.3.4M-1985, American Society of Mechanical Engineers, United Engineering Center, 345 East 47th Street, New York, NY 10017. This document also contains appendices which describe measurement techniques and other useful topics pertaining to axes of rotation.

With the operators S = sine and C = cosine, the general result is

$$ {}^{R}\mathbf{T}_{nerr} = \begin{bmatrix} C\epsilon_Y C\theta_Z & -C\epsilon_Y S\theta_Z & S\epsilon_Y & \delta_X \\ S\epsilon_X S\epsilon_Y C\theta_Z + C\epsilon_X S\theta_Z & C\epsilon_X C\theta_Z - S\epsilon_X S\epsilon_Y S\theta_Z & -S\epsilon_X C\epsilon_Y & \delta_Y \\ -C\epsilon_X S\epsilon_Y C\theta_Z + S\epsilon_X S\theta_Z & S\epsilon_X C\theta_Z + C\epsilon_X S\epsilon_Y S\theta_Z & C\epsilon_X & C\epsilon_Y \\ \delta_Z 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.14a} $$

Note that this general result may also be used for the general case of a linear motion carriage if ε_Z is substituted for θ_Z. Most often, second-order terms such as $\varepsilon_X \varepsilon_Y$ are negligible and small-angle approximations (i.e., $\cos\varepsilon = 1$, $\sin\varepsilon = \varepsilon$) can be used, which leads to

$$ {}^{R}\mathbf{T}_{nerr} = \begin{bmatrix} \cos\theta_Z & -\sin\theta_Z & \epsilon_Y & \delta_X \\ \sin\theta_Z & \cos\theta_Z & -\epsilon_X & \delta_Y \\ \epsilon_X \sin\theta_Z - \epsilon_Y \cos\theta_Z & \epsilon_X \cos\theta_Z + \epsilon_Y \sin\theta_Z & 1 & \delta_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.14b} $$

However, since this matrix is usually evaluated using a spreadsheet, it does not hurt to use the exact values. When nanometer performance levels are sought, second order effects can start to become important.

In the context of evaluating errors of rotating bodies, particularly when making measurements and discussing the results, it is necessary to consider the following terms which are defined in ANSI B89.3.4M.

"*Axis of rotation* - a line about which rotation occurs." As shown in Figure 2.2.3, the axis of rotation revolves, with translational and angular errors, around the Z reference axis. These errors are defined with reference to the reference coordinate system as δ_X (radial motion), δ_Y (radial motion), δ_Z (axial motion), ε_X (tilt motion, generally causing sine errors in a nonsensitive direction), and ε_Y (tilt motion, generally causing sine errors in a sensitive direction). ε_Z is the error in angular position error and it is assumed that it is incorporated into the rotation angle θ_Z. Second-order effects [i.e., $\varepsilon_Z(1-\cos\varepsilon_X)$] can generally be ignored when compared to ε_Z.

"*Spindle* - a device which provides an axis of rotation."

"Perfect spindle - a spindle having no motion of its axis of rotation relative to the reference coordinate axes." Thus for a perfect spindle, all the error terms, $\delta_X \ldots \varepsilon_Y$, are equal to zero.

"Perfect workpiece - a rigid body having a perfect surface of revolution about a center line."

"Error motion - changes in position, relative to the reference coordinate axes, of the surface of a perfect workpiece with its center line coincident with the axis of rotation." This definition excludes thermal drift errors.

"Sensitive and nonsensitive directions - the sensitive direction is perpendicular to the ideal generated workpiece surface through the instantaneous point of machining or gaging." The *fixed sensitive direction* is where the workpiece is rotated by the spindle and the point of machining or gaging is fixed (e.g., in a lathe). The *rotating sensitive direction* is where the workpiece is fixed and the point of machining or gaging rotates with the spindle (e.g., in a jig borer).

"Radial motion - error motion in a direction normal to the Z reference axis and at a specified axial location". Note that the term *error motion* refers to the distance between the axis of rotation and the reference axis. The radial motion will be a function of the position along the Z axis, and the rotation angle. The term *radial runout* includes errors due to radial motion, workpiece out-of-roundness, and workpiece centering errors; thus runout is not equivalent to radial motion.

"Runout - the total displacement measured by an instrument sensing against a moving surface or moved with respect to a fixed surface." The term total indicator reading (TIR) is equivalent to runout. Unfortunately, all too often an imperfect workpiece is eccentrically mounted to a spindle and used to evaluate the performance of a spindle; hence the runout can be a misleading measurement.

"Axial motion - error motion colinear with the Z reference axis." "Axial slip," "end camming," and "drunkenness" are nonpreferred terms which have been used in the past.

"Face motion - error motion parallel to the Z reference axis at a specified radial location." Face motion includes sine errors caused by tilt motions. The term *face runout* includes errors in the workpiece in a manner similar to radial runout; thus face runout is not equivalent to face motion.

"Tilt motion - error motion in an angular direction relative to the Z reference axis." Tilt motion creates sine errors on the spindle which is why the radial error motion is a function of Z position and face motion is a function of radius. In Figure 2.2.3, tilt motion about the Y axis is in the sensitive direction because it causes an error in X direction (along the assumed measurement axis). Note that "coning," "wobble," and "swash" are sometimes used to describe tilt motion, but they are nonpreferred terms.

"Pure radial motion - the concept of radial motion in the absence of tilt motion."

"Squareness - a plane surface is square to an axis of rotation if coincident polar profile centers are obtained for an axial and a face motion polar plot or for two face motion polar plots at different radii." Squareness is equivalent to orthogonality.

Various measurements must be made in order to evaluate these error motions. The data that is collected is typically analyzed with the aid of different types of polar plots because linear plots of the data are often difficult to interpret.

"Error motion polar plot - a polar plot of error motion made in synchronization with the rotation of the spindle." Error motion polar plots are often decomposed into plots of various error components. Some of the various types of error motion polar plots are shown in Figure 2.2.4. Note that it is also very important to consider the frequency spectrum of the errors (See Section 2.4).

"Total error motion polar plot - the complete error motion polar plot as recorded."

"Average error motion polar plot - the mean contour of the total error motion polar plot averaged over the number of revolutions." The average error motion has components that include fundamental and residual error motion components. Note that asynchronous error motion components do not always average out to zero, so the average error motion polar plot may still contain asynchronous components.

"Fundamental error motion polar plot - the best-fit reference circle fitted to the average error motion polar plot." The fundamental error motion polar plot of an eccentric workpiece would actually be a limacon (the polar plot of a pure sinusoid). As the eccentricity decreases, the limacon approaches being a circle.

"Residual error motion polar plot - the deviation of the average error motion polar plot from the fundamental error motion polar plot. For radial error motion measurements, this represents the sum of the error motion and the workpiece (e.g., ball) out-of-roundness. The workpiece out-of-roundness can be removed using a reversal technique as described below.

"Asynchronous error motion polar plot - the deviations of the total error motion polar plot from the average error motion polar plot." Asynchronous in this context means that the deviations are not repetitive from revolution to revolution. Asynchronous error motions are not necessarily random (in the statistical sense).

"Inner error motion polar plot - the contour of the inner boundary of the total error motion polar plot."

"Outer error motion polar plot - the contour of the outer boundary of the total error motion polar plot."

It is often useful to consider the total error motion as the sum of the asynchronous error motion and the repetitive average error motion. The repetitive error is composed of only errors with frequencies that are integer multiples of the axis of rotation frequencies. Asynchronous error motion is not necessarily random in a statistical sense. It is often due to non-random sources (e.g., motors) with frequencies which are not integer multiples of the axis of rotation frequencies.

For each of the different types of polar plots, different centers can be defined. These polar chart centers are used to help define the values of the measured displacements.

"Polar chart (PC) center - the center of the polar chart."

"Polar profile center - a center derived from the polar profile."

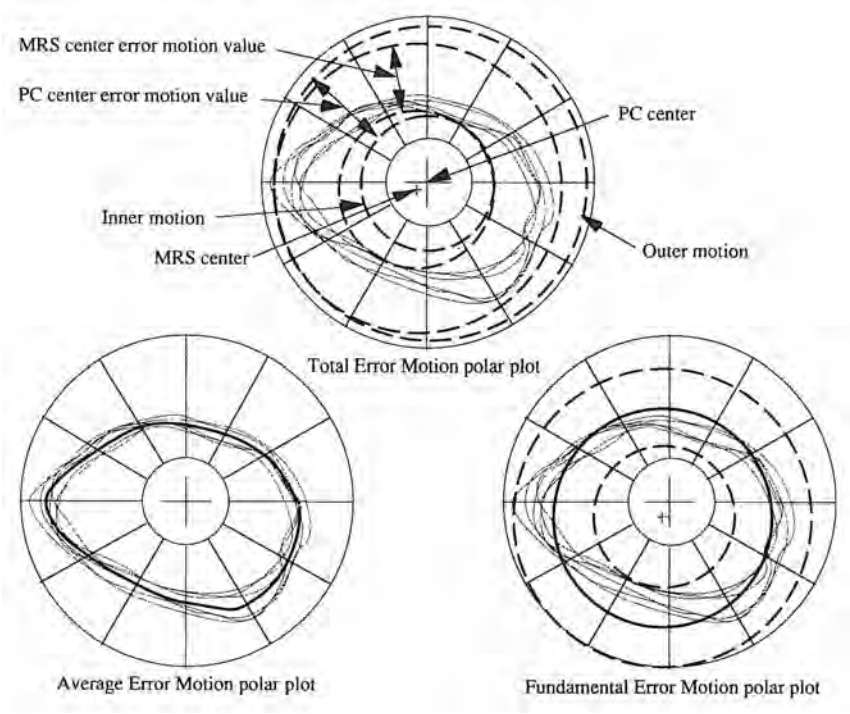


Figure 2.2.4 Examples of error motion and error motion component polar plots.

"Minimum radial separation (MRS) center - the center which minimizes the radial difference required to contain the error motion polar plot between two concentric circles."

"Least squares center - the center of a circle which minimizes the sum of the squares of a sufficient number of equally spaced radial deviations measured from it to the error motion polar plot."[15]

"Axis average line - a line passing through two axially separated radial motion polar plot centers." Note that the default is the MRS center.

Preferred centers: Unless otherwise specified, the following motions are assumed to be measured with respect to:

Radial motion:	MRS center	Face motion:	PC center
Axial motion:	PC center	Residual face motion:	MRS center
Residual axial motion:	MRS center	Tilt motion:	MRS center

Often one wants present just a few numbers to describe the performance of a spindle, instead of an entire set of polar plots. Different types of error motion values are defined for the different types of error motions.

"Total error motion value - the scaled difference in radii of two concentric circles from a specified error motion center just sufficient to contain the total error motion polar plot."

"Average error motion value - the scaled difference in radii of two concentric circles from a specified error motion center just sufficient to contain the average error motion polar plot." The average error motion value is a measure of the best roundness that can be obtained for a part machined while being held in the spindle (or the roundness of a hole the spindle is used to bore).

"Fundamental error motion value - twice the scaled distance between the PC center and a specified polar profile center of the average error motion polar plot." This value represents the once-per-revolution sinusoidal component of an error motion polar plot. Thus when a perfect

[15] See J. I. McCool, "Systematic and Random Errors in Least Squares Estimation for Circular Contours," Precis. Eng., Vol. 1, No. 4, 1979, pp. 215–220.

workpiece is perfectly centered, the fundamental radial error motion value will be zero. In the same manner, fundamental tilt motion does not exist.

"Residual error motion value - the average error motion value measured from a specified polar profile center." This represents the difference between the average and fundamental error motions.

"Asynchronous error motion value - the maximum scaled width of the total error motion polar plot, measured along a radial line through the PC center." For a lathe spindle, given the tool profile, cutting angle, and feed rate, the theoretical surface finish can be predicted (e.g., like the grooves in a phonograph record). The asynchronous error motion values can be used to help predict the deviation from the theoretical surface finish that will be produced on the machine when a "perfect" tool (e.g., a diamond) is used to cut a material that machines well (e.g., some copper and aluminum alloys). Note that a built-up edge of the material being removed tends to form more easily on cutting tools not made from diamond, and a built-up edge can degrade the surface finish.

"Inner error motion value - the scaled difference in radii of two concentric circles from a specified error motion center just sufficient to contain the inner error motion polar plot."

"Outer error motion value - the scaled difference in radii of two concentric circles from a specified error motion center just sufficient to contain the outer error motion polar plot." This plot can be used to help predict the potential out-of-roundness of a part machined while being held in the spindle.

When measuring error motions, in practice, it is very difficult to obtain a perfect workpiece (e.g., a ball) and to mount it perfectly. Thus methods are needed, for example, when measuring radial error motion, to separate the three error components: ball eccentricity, ball out-of-roundness, and spindle radial error motion. The ball eccentricity is represented by the fundamental error motion value, and should be minimized in order to minimize distortion of the polar plots. In order to determine the average error motion value of the spindle, it is necessary to separate the ball out-of-roundness and spindle radial error motion. This can be accomplished using the *Donaldson reversal principle* or the *multistep method*.[16]

The Donaldson reversal technique is used for radial error motion measurements where $P(\theta)$ represents the part out-of-roundness and $S(\theta)$ represents the spindle radial error motion. It must be assumed that the spindle has no significant asynchronous radial error motion (i.e., the error motions are highly repeatable which is usually the case with fluid-film bearings). The first technique, referred to as *procedure P*, yields the roundness error of the part. The second technique, referred to as *Procedure S*, yields the spindle radial error motion.

Initially, 0° positions are marked on the spindle housing, spindle, and ball (workpiece). All the 0° marks are aligned and the sensor is mounted at the 0° position. A set of measurements (profiles) is then taken, and the measured value $T_1(\theta)$ will be the sum of the part out-of-roundness and the spindle radial error motion: $T_1(\theta) = P(\theta) + S(\theta)$. Note that it is assumed that the hills and valleys on the polar plot correspond to hills and valleys on the ball.

Next, the spindle's 0° mark is aligned with the housing's 0° mark and the ball is rotated so its 0° mark is oriented 180° with respect to the housing's 0° mark. The sensor is also moved so it is aligned with the ball's 0° mark. For procedure P, a set of measurements (profiles) is then taken using the same sign convention as for the first set of measurements. The measured value $T_{2P}(\theta)$ will now be the difference between the part out-of-roundness and the spindle radial error motion: $T_1(\theta)$ = $P(\theta)$ - $S(\theta)$. The ball's out-of-roundness can thus be found by averaging $T_1(\theta)$ and $T_{2P}(\theta)$. On a polar plot, $P(\theta)$ is just the curve drawn equidistant between the $T_1(\theta)$ and $T_{2P}(\theta)$ curves. The deviation of the resultant curve from a best-fit circle represents the ball's out-of-roundness.

For procedure S, the sign convention is reversed for the second set of measurements; therefore, $T_{2S}(\theta)$ = -$T_{2P}(\theta)$ = -$P(\theta)$ + $S(\theta)$, and the spindle's radial error motion will be the average of $T_1(\theta)$ and $T_{2S}(\theta)$. Once again, the average can be obtained graphically by drawing a curve equidistant between the $T_1(\theta)$ and $T_{2S}(\theta)$ curves. The deviation of the resultant curve from a best-fit circle represents the spindle's radial error motion. The average error motion value for the spindle can be determined as defined above.

[16] The former was developed by Bob Donaldson at LLNL, and the latter by Spragg and Whitehouse. See Appendix B of ANSI B89.3.4M-1985 and the various cited references. Also see the ANSI standard Measurement of Out-of-Roundness, ANSI B89.3.1-1972(R1979).

Note that with polar graph paper and a dial indicator, one can easily make the measurements and calculations described above. However, if the asynchronous radial error motions are significant, the spindle error has to be represented by the average radial error motion polar plot. The accuracy will depend upon being able to obtain a repeatable average radial error motion for each of the two measurement sets obtained when using the reversal method. In this instance, digital data acquisition systems are often required in order to average many measurement profiles which helps to reduce the effects of asynchronous error motions.[17] Note that repeatability can be improved, for example, when measuring rolling element bearing spindles, by turning the spindle backward to the same starting point.

If one is using a digital data acquisition system, then the multistep method can alternatively be used. For this method, the sensor position remains fixed, and the ball is incrementally rotated through N 360°/N increments with respect to the spindle for each set of measurements. The part error will rotate with each step while the spindle error will not; hence it is possible to separate the two types of errors. The part error is obtained by choosing one angle of the spindle's rotation and then recording the sensor reading at this position for all the different orientations in a sequence. To obtain spindle errors, a fixed angle on the part is chosen instead. Each data set must be normalized so that the profile radius and eccentricity are the same.

Coordinate Frame Location

Perhaps the most important step in assembling the error budget for a machine is the placement of the coordinate frames and the assignment of linear and angular errors corresponding to the axes. Angular motion errors suffer no ambiguity in their definition since they are unaffected by other errors and therefore can be defined with respect to any set of axes. Linear motion errors, on the other hand, must be carefully defined in terms of linear motion caused directly and linear motion that is the result of an Abbe error. Unless the coordinate systems are located at the origin of the angular errors,[18] where pitch, yaw, and roll errors do not cause Abbe errors, then the design engineer must carefully incorporate the Abbe effects into his model. Note that often it is desirable to locate the coordinate system on the surface of the part, so measurements can be made more easily; hence the Abbe errors must often be incorporated in the model. To avoid confusion in the design stage, one should place a coordinate frame at the roll center and one on the surface of the machine, where errors are readily measurable. This is illustrated by an example in Section 5.7.2. All the effort expended in carefully building an error budget can be well rewarded if the machine actually makes a part to tolerance; however, if the part is out of tolerance, the engineer who designed the machine may be required to find out what went wrong. Placing coordinate frames at measurable points on the machine greatly facilitates forensic engineering.

Determination of the Relative Errors Between the Toolpoint and Workpiece

Several steps are necessary to use HTMs to determine the effects of errors in the machine on the relative position and orientation errors of the toolpoint with respect to the workpiece:

1. Formulate the kinematic model of the machine using the sequential process of determining all the machine axes' HTMs that describe an axis's position relative to the axis upon which it is stacked. This is done by starting at the *tool tip* and continuing until the last axis is described with respect to a convenient fixed reference frame. The same procedure is done starting at the ideal tool contact point on the *workpiece* and ending at the same fixed reference frame.

2. For all but the simplest problems, the HTM matrices must be multiplied numerically to obtain the HTMs for the tool and workpiece. To model errors as the machine moves along a specific path, the elements of the HTMs can be numerically assigned values at each point along the machine's path. These values (e.g., position of the machine's axes) can even be numerically generated by the same means that the machine's controller will use to determine where the axes should be relative to each other. Error maps from other machines for which data exists may also be usable in the model.

Regardless of the method used to step the machine's axes numerically through their ideal paths, ideally the HTM products (Equation 2.2.9) for the position of the point on the workpiece the

[17] Typically, an encoder is used to measure spindle rotation for the q coordinate on the polar plot. It is not always feasible to hook up an encoder, so autocorrelation techniques can be used as described in Section 8.8.

[18] This point is often called the *roll center* and its position is a function of the arrangement of an axis's bearings. The roll center is usually at the center of stiffness as illustrated by Equation 2.2.29 (pivot point) which is the point at which when a force is applied, no angular motion occurs. It can be found easily using force-moment balance equations and data on the location and stiffness of bearing components.

tool contacts and the toolpoint with respect to the reference frame will be identical. The relative error HTM $\mathbf{E}_{\text{rel}}$ representing position and orientation errors between the tool and workpiece is determined from ${}^{\text{R}}\mathbf{T}_{\text{work}} = {}^{\text{R}}\mathbf{T}_{\text{tool}}\mathbf{E}_{\text{rel}}$:

$$\mathbf{E}_{\text{rel}} = {}^{\text{R}}\mathbf{T}_{\text{tool}}{}^{-1}{}^{\text{R}}\mathbf{T}_{\text{work}} \tag{2.2.15}$$

The relative error HTM is the transformation in the toolpoint coordinate system that must be done to the toolpoint in order to be at the proper position on the workpiece. The position vector $\mathbf{P}$ (see Equation 2.2.1) component of $\mathbf{E}_{\text{rel}}$ represents the translations in the toolpoint's coordinate frame that must be made to the toolpoint in order to be at the proper location on the workpiece.

For implementation of error correction algorithms on numerically controlled machines, one must consider how the axes will be required to move in order to create the desired motion, specified by Equation 2.2.15, in the tool reference frame. For the general case of a machine with revolute and translational axes (e.g., a five-axis machining center), one would have to use inverse kinematic solutions such as those developed for robot motion path planning. Most machine tools and CMMs have only translational axes, and thus the error correction vector ${}^{\text{R}}\mathbf{P}_{\text{correction}}$ with respect to the reference coordinate frame can be obtained from

$$\text{R}\begin{bmatrix} \mathbf{P}_x \\ \mathbf{P}_y \\ \mathbf{P}_z \end{bmatrix}_{\text{correction}} = \text{R}\begin{bmatrix} \mathbf{P}_x \\ \mathbf{P}_y \\ \mathbf{P}_z \end{bmatrix}_{\text{work}} - \text{R}\begin{bmatrix} \mathbf{P}_x \\ \mathbf{P}_y \\ \mathbf{P}_z \end{bmatrix}_{\text{tool}} \tag{2.2.16}$$

Because of Abbe offsets and angular orientation errors of the axes, ${}^{\text{R}}\mathbf{P}_{\text{correction}}$ will not necessarily be equal to the position vector $\mathbf{P}$ component of $\mathbf{E}_{\text{rel}}$. ${}^{\text{R}}\mathbf{P}_{\text{correction}}$ does represent the incremental motions the X, Y, and Z axes must make on a Cartesian machine in order to compensate for toolpoint location errors.

3. For purposes of minimizing computational energy and numerical error, the *error gain* (or *sensitivity*) of each of the six error types (δ_X, δ_Y, δ_Z, ε_X, ε_Y, and ε_Z) of a particular axis can be calculated and used as the multiplier for all the sources of error that contribute to each component of error at a particular axis. For example, the Y direction straightness of carriage n can have numerous components, such as geometric, thermal, and load-induced errors. The product of the *gain* for the Y-direction straightness of carriage n and each of these errors will yield their respective contributions to the total toolpoint error. The *error gain*[19] for a particular error type in a particular axis is determined numerically by setting all errors in the system equal to zero and then setting the value of the error type of interest equal to a percentage of the machine size (e.g., 0.1%). Equation 2.2.9 is then evaluated for the tool and the work to yield ${}^{\text{R}}\mathbf{T}_{\text{work}}$ and ${}^{\text{R}}\mathbf{T}_{\text{tool}}$. Equation 2.2.16 is then evaluated and the result is divided by the assumed error value. The result is the gain of the particular error for the toolpoint X, Y, and Z errors in the reference coordinate frame. The gain for an angular error on an angular error is always 1. Similarly, the gain for a translational error on a translational error is also always 1. The gain for a translational error on an angular error is always 0. In the determination of the error gains, setting the value of the error type equal to a percentage (e.g., 0.1%) of the machine size instead of a typical error value (ppm) helps to minimize numerical roundoff errors.

4. Once all the error gains are computed and stored, the error between the toolpoint and workpiece can be determined using Equation 2.2.15 or 2.2.16. For each axis, evaluate (and list) all sources of error for each of the six types of error which can occur. The total error is obtained from a combinational rule (discussed below) for the product of all the errors and their respective gains.

5. The HTM model is also useful for systems where the errors are measured and values stored so that the errors can be compensated for in real time. How the machine axes must move in order to produce the desired translation and rotation of the toolpoint depends on the machine's configuration but is a readily solvable geometry problem.

The result of this analysis is a list of all the errors in the system and their corresponding gains (amplification factors) which cause a resultant error at the endpoint. In addition to being well suited for compilation by computer, this method serves to flag errors with large gains (amplification) at the toolpoint. Using the systematic HTM method to aid in formulating a machine's error budget not only removes some of the guesswork from the machine design process, but also provides the quickest way to learn many of the heuristic rules of design for accuracy.

[19] The term *sensitivity* of the machine to a particular type of error is also used. Since a multiplying factor is what is being determined, the author, however, believes that the term *gain* is more appropriate.

One way of minimizing error is maximizing the efficiency of the machine's structural loop. The structural loop is defined as the structure that joins the tool to the fixture to which the workpiece is attached. During cutting operations, the contact between the tool and workpiece can change the structural loop's characteristics. Maximizing the efficiency of the structural loop generally requires minimizing the path length of the mechanism. As can be seen from Abbe's principle or the HTM analysis method, the shorter the path length, the less the error amplification and the total endpoint error. Structural loop design considerations are discussed in Section 7.4.

2.2.2 Combinational Rules for Errors[20]

The error gain matrix provides the amplification factors for each of the error components ε_i in the machine (e.g., $\varepsilon_i = \delta_X, \delta_Y, \delta_Z, \varepsilon_X, \varepsilon_Y, \varepsilon_Z$) at each of the points where homogeneous transformation matrices were defined for the machine. The next step is to define the types of error components that can exist. Like acid and water, different types of errors do not mix unless handled properly. Once all error components are multiplied by their respective error gains, a final combination of errors can then be made to yield an educated guess as to the machine's expected performance. There are three common types of errors, which are defined as:[21]

1. *"Random* - which, under apparently equal conditions at a given position, does not always have the same value, and can only be expressed statistically."
2. *"Systematic* - which always have the same value and sign at a given position and under given circumstances. (Wherever systematic errors have been established, they may be used for correcting the value measured.)" Systematic errors can generally be correlated with position along an axis and can be corrected for if the relative accompanying random error is small enough.
3. *"Hysteresis - i*s a systematic error (which in this instance is separated out for convenience). It is usually highly reproducible, has a sign depending on the direction of approach, and a value partly dependent on the travel. (Hysteresis errors may be used for correcting the measured value if the direction of approach is known and an adequate pre-travel is made.)" Backlash is a type of hysteresis error that can be compensated for to the extent that it is repeatable.

Systematic and hysteresis errors can often be compensated for to a certain degree using calibration techniques. Random error cannot be compensated for without real-time measurement and feedback into a correcting servoloop. Thus when evaluating the error budget for a machine, three distinct *sub-budgets* based on systematic, hysteresis, and random errors should be kept. Often inputs for the error budget are obtained from manufacturers' catalogs (e.g., straightness of linear bearings), and they represent the peak-to-valley amplitude errors e_{PV}. A peak-to-valley error's equivalent random error with uniform probability of occurrence is given by $\varepsilon_{equiv.random} = \varepsilon_{PV}/K_{PV}$. If the error is Gaussian, then K_{PVrms}= 4 and there is a 99.9937% probability that the peak-to-valley error will not exceed four times the equivalent random error.

In the systematic sub-budget, errors are added together and sign is preserved so cancellation may sometimes occur. The same is true for the hysteresis sub-budget. In the random sub-budget, both the sum and the root-mean-square error should be considered, where the latter is given by

$$\varepsilon_{irms} = \left(\frac{1}{N} \sum_{i=1}^{N} \varepsilon_{irandom}^2 \right)^{1/2} \tag{2.2.17}$$

Note that in the random sub-budget, all the random errors are taken as the 1σ values. For the final combination of errors, the 4σ value is typically used, which means that there is a 99.9937% chance that the random error component will not exceed the 4σ value. In this case the total worst-case error for the machine will be

$$\varepsilon_{iworst\ case} = \Sigma\varepsilon_{isystematic} + \Sigma\varepsilon_{ihysteresis} + 4\Sigma\varepsilon_{irandom} \tag{2.2.18}$$

The best-case error for the machine will probably be

$$\varepsilon_{ibest\ case} = \Sigma\varepsilon_{isystematic} + \Sigma\varepsilon_{ihysteresis} + 4(\Sigma\varepsilon_{irandom}^2)^{1/2} \tag{2.2.19}$$

[20] Engineers should periodically review a good statistics book. See, for example, W. Mendenhall, Statistics for Engineering and Computer Science, Macmillan Publishing Co., New York, 1990.

[21] These definitions are from the CIRP Scientific Committee for Metrology and Interchangeability's "A Proposal for Defining and Specifying the Dimensional Uncertainty of Multiaxis Measuring Machines," Ann. CIRP, Vol. 27, No. 2, 1978, pp. 623–630.

In practice, the average of these two values is often used as an estimate of the accuracy the design is likely to achieve. In Section 2.3, factors that contribute to each of the different types of errors are discussed in detail.

After the machine is built and tested, errors will always exist in its performance; thus it is instructive to glimpse at how these errors are evaluated. Uncorrected errors constitute the *measuring uncertainty* in the machine, and can only be fairly evaluated given uniform and consistent testing methods (e.g., number of hours the machine has been allowed to warm up and the mechanical sequence in which it is tested). The uncertainty U is defined as

$$\Delta\ell_{\text{desired}} = \Delta\ell_{\text{observed}} \pm \text{U} \tag{2.2.20}$$

The manner in which these errors combine is demonstrated for a single-axis linear system whose *error graph* is shown in Figure 2.2.5, which utilizes the following terms:[22]

- $e_j$is the uncorrected systematic error, at the point j, with respect to the starting point of the calibration.
- H_j is the uncorrected hysteresis error at the point j.
- P_j is the unidirectional repeatability at the point j.
- R_j is the bidirectional repeatability, at the point j, which includes hysteresis effects, thus $R_j = P_j + H_j$.

Each of these values is taken as four times the standard deviation of the mean with at least five measurements taken from each direction. The total uncertainty between the measurements of length ℓ_1 and length ℓ_2 is $U_{1,2}$

$$U_{1,2} = |e_1 - e_2| + \frac{H_1 + H_2 + P_1 + P_2}{2} \tag{2.2.21}$$

By summing the components H_i and P_i, as opposed to taking the root mean square value, a certainty of 99.5% is obtained. Note that when n individual measurements are made and used to form a mean, the uncertainty of the mean will approximately be the product of $n^{-1/2}$ and the uncertainty of the measurements themselves.

This represents the case for the uncertainty for a one-dimensional machine, but what should be done for a multiaxis machine? For two- or three-axis machines, a point must be approached using each of the different axes from first the positive and then the negative directions while holding all other axes fixed. In this manner, the systematic, hysteresis, and random errors for each of the axes can be found, and then combined as for the single-axis case to give the uncertainty $\pm U_X, U_Y, U_Z$ for the machine. If the machine was well modeled and the error budget correctly formulated, then there should be close agreement between the measured uncertainty and the average of Equations 2.2.18 and 2.2.19.

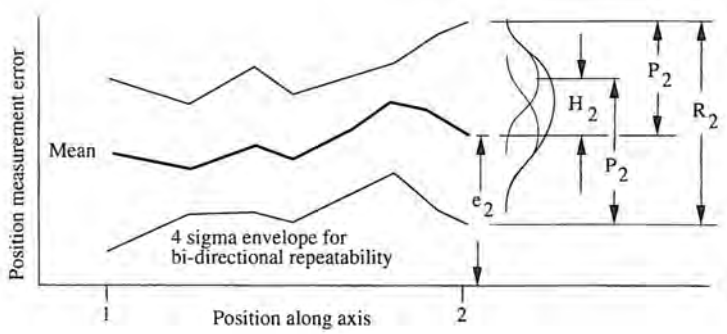


Figure 2.2.5 Error graph for single-axis errors.

[22] Ibid.

2.2.3 Formulation of a Two-Axis Cartesian Machine's Error Gain Matrix

This section will detail the formulation of the error gain matrix for a two-axis Cartesian machine shown schematically in Figure 2.2.6. Note that all the coordinate systems assigned, one per moving axis and a fixed reference coordinate system, have their X, Y, and Z axes parallel. This greatly simplifies formulation of the error budget.

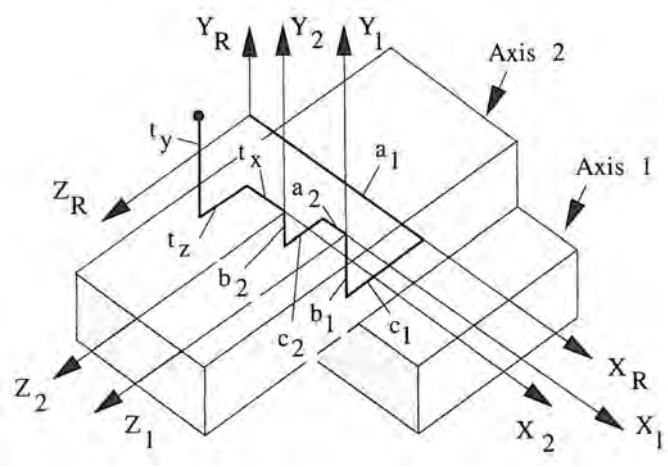


Figure 2.2.6 Coordinate frame definition for a two-axis Cartesian machine.

For the two-axis machine, the errors in each carriage are identical to those of the carriage shown in Figure 2.2.2. By keeping all the coordinate system's axes parallel, the HTM given by Equation 2.2.12 can be used for each of the carriage's axes:

$$ {}^{R}\mathbf{T}_1 = \begin{bmatrix} 1 & -\varepsilon_{z_1} & \varepsilon_{y_1} & a_1 + \delta_{x_1} \\ \varepsilon_{z_1} & 1 & -\varepsilon_{x_1} & b_1 + \delta_{y_1} \\ -\varepsilon_{y_1} & \varepsilon_{x_1} & 1 & c_1 + \delta_{z_1} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.22} $$

$$ {}^{1}\mathbf{T}_2 = \begin{bmatrix} 1 & -\varepsilon_{z_2} & \varepsilon_{y_2} & a_2 + \delta_{x_2} \\ \varepsilon_{z_2} & 1 & -\varepsilon_{x_2} & b_2 + \delta_{y_2} \\ -\varepsilon_{y_2} & \varepsilon_{x_2} & 1 & c_2 + \delta_{z_2} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.2.23} $$

Proceeding with the logic used to obtain Equation 2.2.16, for a tool located at point t_x, t_y, t_z,the actual coordinates of the toolpoint in the reference coordinate system are given by

$$ \begin{bmatrix} X_t \\ Y_t \\ Z_t \\ 1 \end{bmatrix}_{\text{actual}} = {}^{R}\mathbf{T}_1 \quad {}^{1}\mathbf{T}_2 \begin{bmatrix} t_x \\ t_y \\ t_z \\ 1 \end{bmatrix} \tag{2.2.24} $$

The ideal coordinates of the toolpoint would be the sum of all the individual components along their respective axes:

$$ \begin{bmatrix} X_t \\ Y_t \\ Z_t \end{bmatrix}_{\text{ideal}} = \begin{bmatrix} a_1 + a_2 + t_x \\ b_1 + b_2 + t_y \\ c_1 + c_2 + t_z \end{bmatrix} \tag{2.2.25} $$

The translational errors in the toolpoint position are thus given by

$$ \begin{bmatrix} \delta_{x_t} \\ \delta_{y_t} \\ \delta_{z_t} \end{bmatrix} = \begin{bmatrix} X_t \\ Y_t \\ Z_t \end{bmatrix}_{\text{actual}} - \begin{bmatrix} X_t \\ Y_t \\ Z_t \end{bmatrix}_{\text{ideal}} \tag{2.2.26} $$

Evaluating Equation 2.2.26 while neglecting second-order terms gives

$$\begin{bmatrix} \delta_{x_t} \\ \delta_{y_t} \\ \delta_{z_t} \end{bmatrix}_{\text{actual}} = \begin{bmatrix} -t_y(\varepsilon_{z_1}+\varepsilon_{z_2}) + t_z(\varepsilon_{y_1}+\varepsilon_{y_2}) - b_2\varepsilon_{z_1} + c_2\varepsilon y_1 + \delta_{x_1} + \delta_{x_2} \\ t_x(\varepsilon_{z_1}+\varepsilon_{z_2}) - t_z(\varepsilon_{x_1}+\varepsilon_{x_2}) + a_2\varepsilon_{z_1} - c_2\varepsilon x_1 + \delta_{y_1} + \delta_{y_2} \\ -t_x(\varepsilon_{y_1}+\varepsilon_{y_2}) + t_y(\varepsilon_{x_1}+\varepsilon_{x_2}) - a_2\varepsilon_{y_1} + b_2\varepsilon x_1 + \delta_{z_1} + \delta_{z_2} \end{bmatrix} \tag{2.2.27}$$

The error gains are the coefficients of the errors δ_{x1}, δ_{y1}, δ_{z1}, ε_{x1}, ε_{y1}, ε_{z1}, δ_{x2}, δ_{y2}, δ_{z2}, ε_{x2}, ε_{y2}, ε_{z2}. Note that in a machine with a rotary axis, an angular error about one axis may have components about other axes; however, in both of these cases, the gain associated with these errors will not represent any amplification by distance. Thus they would not typically be reported in a table of error gains. Table 2.2.1 shows the error gain matrix for the two-degree-of-freedom carriage discussed above.

	δ_{xt}	δ_{yt}	δ_{zt}
Axis 1 errors			
ε_{x1}	0	$-t_z - c_2$	$t_y + b_2$
ε_{y1}	$t_z + c_2$	0	$-t_x - a_2$
ε_{z1}	$-t_y - b_2$	$t_x + a_2$	0
Axis 2 errors			
ε_{x2}	0	$-t_z$	t_y
ε_{y2}	t_z	0	$-t_x$
ε_{z2}	$-t_y$	t_x	0

Table 2.2.1 Error gain matrix for a two-axis Cartesian carriage system.

Remember, the error gain matrix serves to call attention to errors which may be dominant. Also, even though signs are given with the gains, when the total error is evaluated by the root-mean-square method, the signs have no effect. Furthermore, each type of error, such as δ_{x1}, can have numerous components associated with different physical phenomena, causing error motions in the X direction. Finally, note how the translational errors are static in nature (i.e., their gains are unity.) It is the angular errors that can have potentially large gains associated with their effect on translational errors. These Abbe errors are often the dominant errors in a machine.

2.2.4 Error Motions Caused by Bearing Point Deflections[23]

Many high-precision machines are used primarily to scan the surface of a sample. Under these operating conditions, the machine axes are moving with a quasi-steady velocity. This section will describe a procedure that can be used to assess the error motions associated with a machine's steady-state operating conditions. First kinematically supported carriages will be considered, and then the solution method will be described for a non kinematic configuration of bearings.

First assume that a kinematically supported carriage and bearing way coordinate systems' axes are coincident. Next assume that the carriage is free to move along the X axis and that the position vectors of the five bearing contact points in the carriage coordinate system are given by $[P_{bi}]^T = [X_{bi}\ Y_{bi}\ Z_{bi}]_{i=1,5}$. The bearing stiffnesses at the five bearing points are assumed to be $K_{bi\ i=1\ to\ 5}$. The direction cosines of the five bearing reaction forces are defined[24] in the carriage coordinate system are given by $[\Theta_{bi}]^T = [\alpha_{bi} \quad \beta_{bi} \quad \gamma_{bi}]_{i=1,5}$. Note that is assumed that the direction of motion is always along the X axis, so $\alpha_{bi} = 0$ always. Three types of kinematically supported carriages are shown in Figure 2.2.7. For small (micron level) bearing deflections, the effect of the resultant angular motions of the carriage on the bearings' direction cosines will be second order and thus can be ignored.

At each of the five bearing points, there are static and dynamic coefficients of friction, μ_{Fi} and μ_{vi}. From a standstill when the loads are first applied, the motions of the bearing contact points have components in the YZ plane as the carriage deflects and the system geometry changes accordingly.

[23] The study of how geometric constraints affect the motion of bodies is called *Screw Theory*. See, for example, J. Phillips, Freedom in Machinery: Introducing Screw Theory, Cambridge University Press, London, 1982.

[24] Given that the orientation of the bearing reaction force is parallel to the vector from 0,0,0 to a,b,c, the direction cosines are defined as, $\alpha = a/(a^2+b^2+c^2)^{1/2}$, $\beta = b/(a^2+b^2+c^2)^{1/2}$, and $\gamma = c/(a^2+b^2+c^2)^{1/2}$.

As motion progresses and the bearing contact points reach their equilibrium YZ positions, there will be no tendency toward motion in the YZ plane and the direction of the friction force vectors will lie principally along the X axis.

The direction of the friction forces is dependent on the direction of the applied X direction force, the location of the pivot point, and the net moment caused by all external forces. The center of friction and the center of stiffness govern the location of the pivot point of the system (point about which rotations occur when forces are applied). Note the similarity to the center of mass calculation (for this case, M = 5)

$$\xi_{\text{center of friction}} = \frac{\sum_{i=1}^{M} F_{\text{normal i}}\mu_i\xi_i}{\sum_{i=1}^{M} F_{\text{normal i}}\mu_i} \text{ for } \xi = X, Y, Z \tag{2.2.28}$$

$$\xi_{\text{center of stiffness}} = \frac{\sum_{i=1}^{M} K_i\xi_i}{\sum_{i=1}^{M} K_i} \text{ for } \xi = X, Y, Z \tag{2.2.29}$$

When a moment is applied to the carriage, friction forces on one side of the pivot point (e.g., Y coordinate > 0) will be oriented in one direction, while friction forces on the other side of the pivot point will have the opposite direction. However, determination of the bearing reaction forces also depends on the friction forces' magnitudes and directions; and friction force directions for resisting moments will for moments about the Y and Z axes.

A conditional iterative model for these effects would be unduly complex considering that there must be an actuator or clamping point to maintain position of the system in the X direction. Thus it is conservative (larger than actual errors will be predicted) to assume that the static friction force is zero, and that moments are resisted by the bearing geometry and bearing preload. Thus the only frictional force accounted for is the dynamic friction force which acts opposite to the velocity vector:

$$F_{X_{\mu vi}} = \mu_{vi} v_x \tag{2.2.30}$$

If one was interested in the transient solution, the pivot point information and the inertia and acceleration of the carriage in the YZ plane would have to be incorporated into the model and an iterative solution process used. The friction force vectors would then be oriented along the respective lines that connect the past t - 2Δt time step XYZ points and the past t - Δt time step XYZ points for each of the bearing reaction force points.

The carriage is loaded by a set of N force vectors $[F_{fj}]^T = [F_{fXj}F_{fYj}F_{fZj}]_{j=1,N}$, which are applied at points defined in the carriage coordinate system $[P_{fj}]^T = [X_{fj}Y_{fj}Z_{fj}]_{j=1,N}$. In addition, a set of torques (moments) Γ_X, Γ_Y, Γ_Z are assumed to be applied to the carriage. Note that for a moment, the point of application does not need to be known in order to determine the resultant forces at the bearing points.

In summary, the known parameters of the problem are:

- P_{bi} bearing point coordinate vectors
- Θ_{bi} direction cosine vectors of the bearing contact points
- K_{bi} bearing stiffnesses at the bearing points
- μ_{vi} dynamic velocity coefficients of friction
- $F_{f\xi j}$ N generic applied force vectors
- $P_{f\xi j}$ coordinate vectors of the generic forces
- Γ_ξ generic applied torques about the X, Y, and Z axes

The 16 unknowns of the problem are the bearing reaction forces, the steady-state velocity of the carriage in the X direction, the gap change at each of the five bearing points, and the error terms in the carriage's homogeneous transformation matrix:[25]

[25] This assumes that the carriage, bearing rails, and machine structures are rigid bodies. The actual deformation of the structure would change the coordinates of the bearing reaction force points, but the effect on determination of bearing deflection errors would be second order.

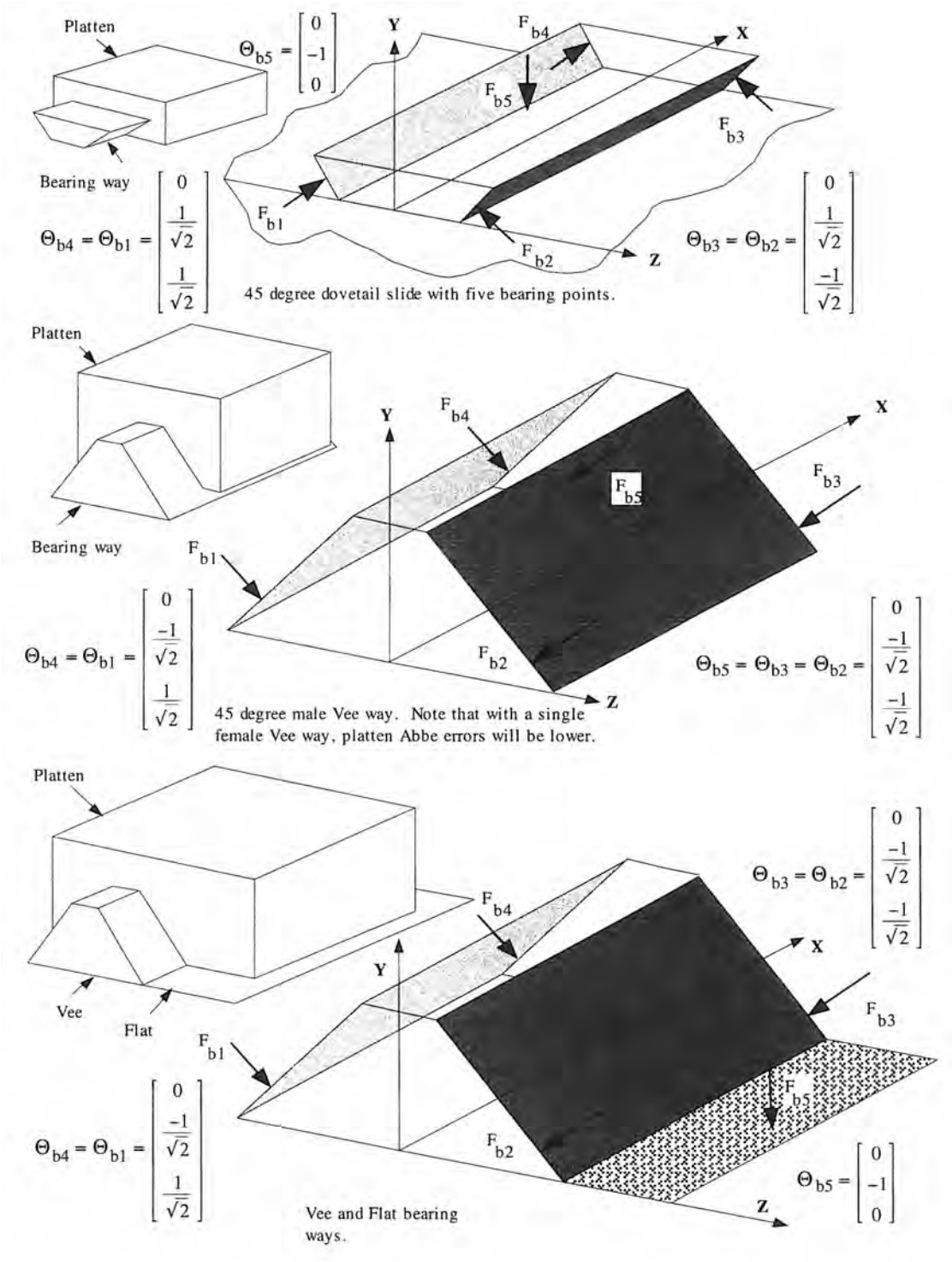


Figure 2.2.7 Common kinematic bearing way and carriage configurations.

- F_{bi} magnitudes of the five bearing reaction forces
- v_x steady-state velocity of the carriage in the X direction
- δ_{bi} gap change at each of the five bearing points
- ε_X rotation about the X axis (roll)
- ε_Y rotation about the Y axis (yaw)
- ε_Z rotation about the Z axis (pitch)
- δ_Y translation along the Y axis (Y-direction straightness)
- δ_Z translation along the Z axis (Z-direction straightness)

Note that any error motions along the X axis are assumed to be measured and compensated for with the X-axis servo. In order to solve for the 16 unknowns, force and geometric equilibrium equations are used.

The first step is to determine the five bearing reaction forces (F_{bi}) and the X-direction velocity (v_x). The effect of the bearing deflections on the equations used to determine the forces is second order. If the deflections were to be incorporated into the model, then the equations would be nonlinear and an iterative solution would be required, so second-order effects are ignored here.

$$\Sigma F_X = 0 = -\sum_{i=1}^{5} \mu_{vi} v_x + \sum_{j=1}^{N} F_{fx_j} \tag{2.2.31}$$

$$\Sigma F_Y = 0 = \sum_{i=1}^{5} F_{bi}\beta_{bi} + \sum_{j=1}^{N} F_{fy_j} \tag{2.2.32}$$

$$\Sigma F_Z = 0 = \sum_{i=1}^{5} F_{bi}\gamma_{bi} + \sum_{j=1}^{N} F_{fZ_j} \tag{2.2.33}$$

$$\Sigma M_x = 0 = \Gamma_X + \sum_{i=1}^{5} F_{bi}(-Z_{bi}\beta_{bi} + Y_{bi}\gamma_{bi}) + \sum_{j=1}^{N} (-Z_{fj}F_{fYj} + Y_{fj}F_{fZj}) \tag{2.2.34}$$

$$\Sigma M_Y = 0 = \Gamma_Y - \sum_{i=1}^{5} \mu_{vi} v_x Z_{bi} + \sum_{i=1}^{5} F_{bi}(Z_{bi}\alpha_{bi} - X_{bi}\gamma_{bi}) + \sum_{j=1}^{N} (Z_{fj}F_{fXj} - X_{fj}F_{fZj}) \tag{2.2.35}$$

$$\Sigma M_Z = 0 = \Gamma_Z + \sum_{i=1}^{5} \mu_{vi} v_X Y_{bi} + \sum_{i=1}^{5} F_{bi}(-Y_{bi}\alpha_{bi} + X_{bi}\beta_{bi}) + \sum_{j=1}^{N} (-Y_{fj}F_{fXj} + \Sigma X_{fj}F_{fYj}) \tag{2.2.36}$$

Once these equations are expanded and represented in matrix form, numerical values for the unknowns F_{b1}, F_{b2}, F_{b3}, F_{b4}, F_{b5}, and v_x can then be found using a spreadsheet. Thus 6 of the 16 unknowns can be found.

The next step is to find the changes δ_{bi} in the bearing gaps caused by the bearing reaction forces acting on the finite stiffness bearings. Five of the remaining 10 unknowns can now be found:

$$\delta_{bi} = F_{bi}/K_{bi} \tag{2.2.37}$$

In order to determine the remaining five unknowns, the error motions, geometric compatibility relations must be used. The new coordinates $[P_{bi}]_{new}$ of the bearing points in the carriage's coordinate frame are equal to the old coordinates $[P_{bi}]$ minus the deflections along the direction cosines of the bearing reaction forces:

$$[P_{bi}]_{new} = [P_{bi}] - [\Theta_{fbi}]\delta_{bi} \tag{2.2.38}$$

In order to determine the values of the five remaining unknowns (the five error motions), geometric constraints need to be considered. There are five geometric constraint equations relating

the Y and Z positions of the bearing points in the bearing way's coordinate system. These constraints represent the planes the bearing points are restricted to move in as the bearing gaps change due to changing forces on the bearing contact points as various forces and moments are applied. The new coordinates of the bearing contact points in the bearing way coordinate system are found using a homogeneous transformation matrix[26] with no translational offsets since the original coordinate systems of the carriage and bearing way were considered to be coincident:

$$[P_{bi}] = \begin{bmatrix} 1 & -\varepsilon_Z & \varepsilon_Y & 0 \\ \varepsilon_Z & 1 & -\varepsilon_X & \delta_Y \\ -\varepsilon_Y & \varepsilon_X & 1 & \delta_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} [P_{bi}]_{new} \tag{2.2.39}$$

where the elements of $[P_{bi}]$ are $[X_{\text{bearing way}}\ Y_{\text{bearing way}}\ Z_{\text{bearing way}}\ 1]^T$. Assuming that the bearing direction cosines given by Equation 2.2.29 are normal to the bearing ways, then the constraint equations between the bearing coordinates are

$$Y_{\text{bearing way i}} = -(Z_{\text{bearing way i}} - Z_{bi})\gamma_{bi}/\beta_{bi} + Y_{bi} \tag{2.2.40}$$

Expanding equations 2.2.39 and 2.2.40 and putting the results in matrix form allows for the determination of the unknowns ε_X, ε_Y, ε_Z, δ_Y, and δ_Z: Twenty simultaneous equations are generated in the variables ε_X, ε_Y, ε_Z, δ_Y, δ_Z and $(X_{inew}, Y_{inew}, Z_{inew})_{i=1,5}$. One can also add to the error terms of the HTM the effects of deformations of the carriage itself. In addition, if it was desired to determine how error motions in the bearing way surfaces affected the motion of the carriage, one could use Equations 2.2.38–2.2.40, where the deflections of the bearing points are dependent on the deviations of the bearing ways beneath them.

This solution can be generalized for any number of contact points if it is assumed that all points are in contact and the change in gap at a point is proportional to a change in force (e.g. no gap opening). If the force-deflection relation cannot be linearized over the expected deflection interval, then the same general approach can be used but the solution method may have to be iterative. The capability to determine the error motions of a carriage as a function of its geometry and input forces is a powerful tool that can be used for studies of carriage design. For example, as the bearing spacing increases, the angular error terms generally decrease, but the increased weight of the carriage itself causes greater errors.

2.3 QUASI-STATIC MECHANICAL ERRORS

Quasi-static mechanical errors are *errors in the machine, fixturing, tooling, and workpiece* that occur relatively *slowly*. This means that the errors occur at a frequency much lower than the bandwidth of axes on the machine that could be used to correct the errors. For example, if a linear axis travels at 2.5 mm/min (0.1 in./min) during a finish cut and the axis has a straightness error with a wavelength on the order of 0.1 m (4 in.), an orthogonal axis would only have to be able to move back and forth at 0.00043 Hz with a range of motion equal to the straightness error. In general quasi-static errors can be compensated for as illustrated by the case study in Chapter 6. Sources of these types of errors include:

- Geometric errors
- Kinematic errors
- External load-induced errors
  - Errors caused by gravity loads
  - Errors caused by accelerating axes
  - Errors caused by cutting forces
- Machine assembly load-induced errors
- Thermal expansion errors
- Material instability errors
- Instrumentation errors

Some errors have a period of hours or even years. These types of errors include errors caused by thermal growth and material instability (e.g., austenitic steel transforming to the more stable

[26] Remember that it was assumed that any error motion in the X direction would be measured and corrected for by the carriage's servo; thus $\delta_X = 0$. Note that even if δ_X were finite, it could not affect the values of δ_Y and δ_Z.

ferritic state), respectively. Because the periods may be lengthy, it is often difficult to measure and correct for this type of error, as the stability of the measuring system itself often comes into question. Hence the longer the anticipated time constant of the error, the more effort should be expended to eliminate it. The exceptions to this generalization are: (1) the error soaks through the machine and achieves a uniform steady-state value, and (2) the machine can readily be recalibrated (or offsets measured) so that during the time the machine is working on a single part the error is insignificant.

In order to visualize quasi-static errors in a machine, first assume that every machine component is infinitely rigid. Then, for each single-degree-of-freedom machine component (e.g., a linear carriage), superimpose on the component five off-axis errors (vertical and horizontal straightness, and yaw, pitch, and roll errors) and one on-axis error. The sixth error will include errors in the sensor and any servo error. The magnitude of the linear errors will be highly dependent on where the HTM coordinate axes are assigned. Next, repeat the process while imagining that every part is made of soft rubber: How will the machine deform in each case?

2.3.1 Geometric Errors

Geometric errors are defined here as *errors in form of individual machine components* (e.g. straightness of motion of a linear bearing). Geometric errors are concerned with the quasi-static accuracy of surfaces which move relative to each other, such as components of linear and rotary axes as shown in Figures 2.2.2 and 2.2.3. Geometric errors can be smooth and continuous (systematic) or they can exhibit hysteresis (e.g., backlash) or random behavior. Many factors affect geometric errors including

- Surface straightness
- Surface roughness
- Bearing preload
- Kinematic versus elastic design principles
- Structural design philosophies

How can the design engineer estimate geometric errors? In addition to part catalogs, one must consider the accuracy of the manufacturing process and how it is characterized.

Surface Straightness

Straightness is the deviation from true straight-line motion. One generally thinks of the straightness error as primarily dependent on the overall geometry of the machine and applied loads. As shown in Figure 2.3.1, straightness error can be considered the deviation from a straight line of the motion of a linear axis. The unofficial term *smoothness* can be used to describe straightness errors that are dependent on the surface finish of the parts in contact, the type of bearing used, and the bearing preload. In other words, the smoothness of motion would be the deviation from the best-fit polynomial describing the straightness of motion. Smoothness is not intended to be a descriptor of surface finish as described below, but rather a descriptor of high-frequency straightness errors whose wavelength is typically on the order of the magnitude of the error, normal to the surface of two moving bodies in contact.

Surface Roughness

Surface roughness is a characterization of the profile of the surface and often has an effect (although difficult if not impossible to characterize) on the smoothness of a bearing's motion. In terms of the manufacturing process, smoothness of motion of a bearing can only be quantified in terms of the surface roughness and bearing design:

1. Sliding contact bearings tend to average out surface finish errors and wear less when the skewness is negative. A positive skewness (defined below) can lead to continued wear of the bearing.

2. For rolling element bearings, if the contact area is larger than the typical peak-to-valley spacing, an elastic averaging effect will occur and a kinematic arrangement of rollers will produce smooth motion. If this condition is not met, the effect will be like driving on a cobblestone street. If numerous rolling elements are used, the effects of elastic averaging can help to smooth out the motion. If the elements are recirculating, however, noise may be introduced into the system as the rolling elements leave and enter the load-bearing region.

3. Hydrostatic and aerostatic bearings are insensitive to surface finish effects when they are considerably less than the bearing clearance.

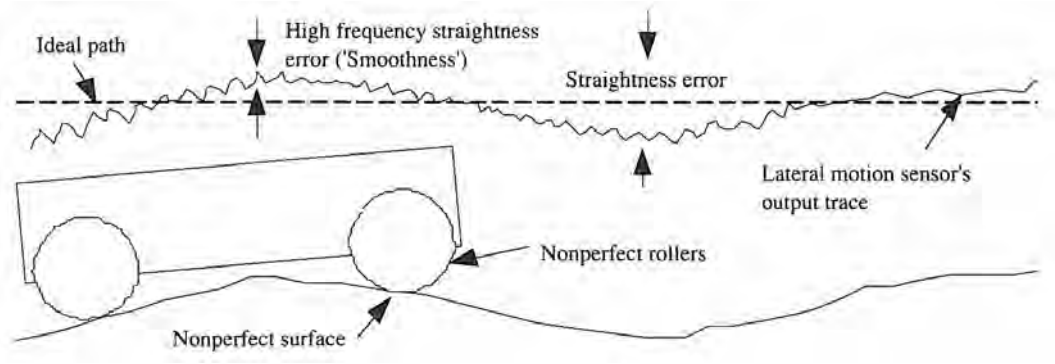


Figure 2.3.1 Straightness errors caused by surface form and finish errors.

There are three common parameters for specifying the surface finish or roughness[27], the *root mean square* (rms or R_q), the *centerline average* (R_a), and the International Standards Organization (ISO) 10-point height parameter (R_z). The latter is with respect to the five highest peaks and five lowest valleys on a sample. A surface profile measurement yields a jagged trace (like that under the wheel in Figure 2.3.1). If a best fit straight line is drawn through a section of the trace of length L, then the R_q, R_a, and R_z surface finish are defined, respectively, from deviations y from the line as a function of distance x along the sample:

$$R_q = \sqrt{\frac{1}{L}\int_0^L y^2(x)\,dx} \tag{2.3.1a}$$

$$R_a = \frac{1}{L}\int_0^L y(x)\,dx \tag{2.3.1b}$$

$$R_z = \frac{\sum_{i=1}^{5} y_{peak}(i) - y_{valley}(i)}{5} \tag{2.3.1c}$$

The former two are defined in terms of continuous functions of distance along the sample (X axis). Realistically, these integrals would be evaluated numerically (usually by the software in the measurement system).

Unfortunately, these values do not provide any information as to the topographical characteristics of the surface. As shown in Figure 2.3.2, the surface topography can be characterized by the *skewness*. The skewness is the ratio of the third moment of the amplitude distribution and the standard deviation σ from the mean line drawn through the surface roughness measurements. Hence the skewness provides a measure of the shape of the amplitude distribution curve. Skewness is often used to quantify the acceptability of a surface for a particular application. For a bearing, sharp deep valleys separated by wide flat planes may be acceptable. This form would have a negative skewness value and is typically in the range -1.6 to -2.0 for bearing surfaces. Sharp spikes would soon grind off, creating wear debris and more damage, and hence positive skewness values are unacceptable for contact-type bearing surfaces. The skewness is defined mathematically as:

$$\text{skew} = \frac{\mu_3}{\sqrt{\mu_2}} = \frac{\mu_3}{\sigma} \tag{2.3.2a}$$

where the nth moment μ_n of the amplitude distribution is defined as

$$\mu_n = \int_{-\infty}^{\infty} (y-\mu)^n f(y)\,dy \tag{2.3.2b}$$

[27] See, for example, K. Stout, "How Smooth is Smooth," Prod. Eng. May 1980. The following discussion on surface roughness is derived from this article. For a detailed discussion of this subject, see Surface Texture (Surface Roughness, Waviness, and Lay), ANSI Standard B46.1-1985, American Society of Mechanical Engineers, 345 East 47th St., New York, NY 10017. Also see T. Vorburger and J. Raja, Surface Finish Metrology Tutorial, National Institute of Standards and Technology Report NISTIR 89-4088 (301-975-2000).

and the mean μ is defined as

$$\mu = \int_{-\infty}^{\infty} yf(y)\,dy \tag{2.3.2c}$$

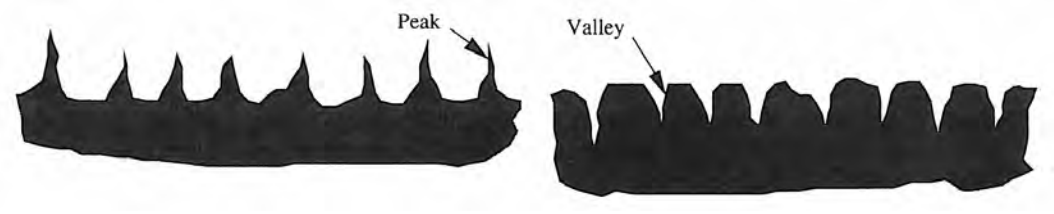


Figure 2.3.2 Surfaces with positive (left) and negative (right) skewness.

The probability density function (PDF) f(y) simply defines the chance of a value occurring. For the surface roughness measurement data, given a value of y, the probability of a reading having a value in the range y - ε to y + ε (where ε is infinitesimally small) is equal to the projected length of the surface roughness curve that lies in the region y - ε to y + ε onto the X axis divided by the length of the trace on the X axis. Once again, these calculations are usually computed in the measuring instrument's software.

Numerous other methods exist for defining the shape and intensity of surface roughness features. For example, the autocorrelation function is used to check for the degree of randomness in a surface. This can be used to help track down periodic components (e.g., those caused by tool chatter), which can then sometimes be reduced in subsequently made parts. A frequency spectrum analysis can also be used to accomplish this. The science of surface metrology is constantly evolving as surface finish requirements increase, and the interested reader should consult the literature.[28]

Bearing Preload

For bearing systems where there are more contact points than degrees of freedom restrained (nonkinematic bearing systems), a high preload can decrease the chance of an isolated rough spot affecting the smoothness. However, high preloads accelerate wear and can lead to stiction, which decreases controllability. The ratio of preload to applied load is also an important consideration where nonlinear deflection relations (Hertzian deflection of two curved bodies in contact as discussed in Section 5.6) exist. Insufficient preload is also synonymous with at least some of the bearing points periodically losing contact with the bearing surface. This results in difficult-to-map error motions and decreased stiffness, possibly leading to chatter of the tool. Tool chatter, in turn, degrades the surface finish of the part.

Unless the axes of the machine are designed to utilize the weight of the machine itself as a preload, in many cases the preload and accuracy of the machine will change with time. This is particularly true if contact-type bearing surfaces wear and if the structure relaxes due to internal stress needed to maintain the preload. In an effort to counter these effects, the classic method has been to use a device known as a gib[29] to generate the preload. In general, a gib can be defined as a mechanical device that can apply preload to a bearing. In the classical sense, a gib was a wedge-shaped part that was advanced by the action of a screw. To assess the accuracy of a properly constructed gibbed surface, one treats the gib as another degree of freedom. That is, when the gib is adjusted, its finite motion introduces another set of geometric errors that must be included in that particular axis's errors.

When a gibbed machine wears, the gibs, bearings, and sometimes the ways must be re-hand-finished and the gibs adjusted to provide the proper preload. In many machines it is common to use modular rolling bearing components whose preload is set by using oversized balls or rollers, or by tightening bolts that push on a plate contacting the rollers. When the latter type of system wears, it is often discarded and replaced with a new unit. Hence large economies of scale can be obtained by some bearing component manufacturers.

---

[28] See, for example, Journal of Surface Metrology edited by K. Stout and published by Kogan Page, London.

[29] Gib design is discussed in Section 8.2.2.

Kinematic Versus Elastic Averaging Design Principles

With a kinematic (nonoverconstraint) mounting, there is no forced geometric congruence and the total error of a structure is more predictable. Elastically averaged systems force geometric congruence (massive overconstraint) and thus evaluation of errors is more difficult. These design philosophies greatly affect structural, bearing, and actuator design philosophies. Hence these design philosophies will be encountered throughout this book and the precision machine design engineer's career.

A geometric error affected by these design philosophies is the error resulting from an indexing motion caused by breaking mechanical contact, repositioning, and then reestablishing mechanical contact. Analysis of this type of error requires discussion of two principles: the principle of *kinematic design* and the principle of *elastic averaging.* The former is a deterministic approach (e.g., a three-legged stool) and the latter is a probabilistic approach (e.g., a five-legged chair with a flexible frame). These concepts are discussed in detail in Section 7.4.3.

Structural Design Philosophies[30]

Regardless of the manner in which a system is designed, geometric errors will always be present. Pure translational errors are not as serious as Abbe errors because the former are not amplified at the tool tip. Compensating curvatures and error-mapping techniques are often used in an attempt to compensate for errors. In the former, the bearing ways are manufactured to have a straightness profile to compensate for deflection of the machine as a heavy machine axis traverses the length of the bearing. The method of compensating curvatures is the old school of thought method for dealing with errors before computer control was widely available. Unfortunately, when the load on the machine changes, the amount of correction also changes. Thus, even without software error corrections, if possible it is better to make the machine stiff enough to prevent the error in the first place or use counterbalancing weights suspended from noninfluencing load frames to help support the load.

The errors in a machine can sometimes be mapped and used as part of the servo-feedback signal for an orthogonal axis to correct for the error. If error-mapping techniques are to be effective, then all axis' straightness, yaw, pitch, and roll errors must be mapped as a function of geometry and temperature. The error correction algorithm must also have access to the tool and workpiece geometry so that Abbe errors at the tool tip can be compensated for. Note the complexity that results from interaction between errors, such as X errors, which may be a function of both Y and Z errors in some machines. Thus as a potential purchaser of machines for a factory you may one day manage, remember *caveat emptor.*[31]

2.3.2 Kinematic Errors

Kinematic errors are defined here as *errors in an axis's trajectory that are caused by misaligned or improperly sized components*. For example, kinematic errors include orthogonality (squareness or perpendicularity) and parallelism of axes with respect to their ideal locations and each other. Translational errors in the spatial position of axes are also a form of kinematic error. The dimensions of an axis's components can also cause the tool or workpiece to be offset from where it is supposed to be and is also classified as a kinematic error. Translational errors, however, are usually easily compensated for in Cartesian machines by using a tool offset. Errors incurred as a result of loads induced during assembly also are kinematic in nature, but they are considered in a class by themselves and are discussed in Section 2.3.4.

As shown in Figure 2.3.3, given two machine axes of motion in the XZ plane with one machine axis aligned with the X reference axis, the orthogonality error is defined as the deviation ε_y from 90° between the machine's other axis and the reference X axis. This is a straightforward definition which is simple to specify on a drawing, but it is not necessarily simple to measure or control in production. As is also shown in Figure 2.3.3, parallelism between two axes has horizontal and vertical forms that define the relative taper and twist between the two axes, respectively. An example of horizontal parallelism error (taper) would be an axis on a lathe that was not parallel to the spindle's axis of rotation. Using this axis to move a tool along the outer surface of a part would result

[30] See Section 7.4.

[31] Let the buyer beware.

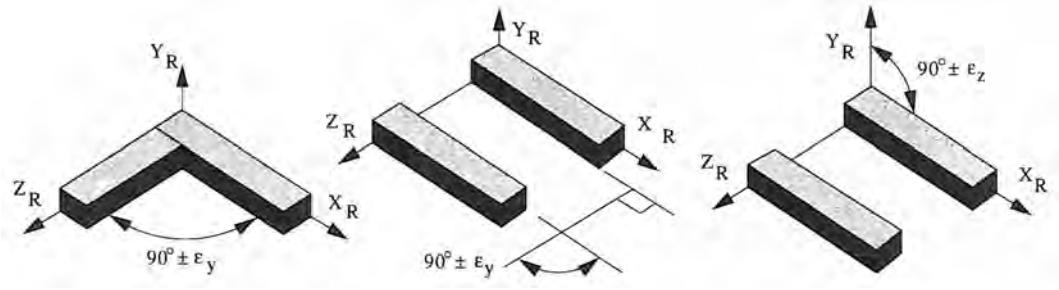


Figure 2.3.3 Orthogonality and horizontal and vertical parallelism errors.

in the part becoming tapered along its length. An example of vertical parallelism error is two axes used to support a milling machine bed. If one end of one of the axes is high, the bed itself will be warped when bolted to the axes, and parts machined while bolted to the bed will later wobble like a four-legged chair with one short leg. Note that these errors are sometimes in nonsensitive directions. For the lathe, a horizontal parallelism error moves the toolpoint in a direction normal to the surface, while vertical parallelism moves the toolpoint tangent to the surface in a nonsensitive direction.

The accuracy to which orthogonality and parallelism can be specified is entirely dependent on machining and grinding processes if no hand finishing work (e.g., scraping and/or lapping) is to be done. Best-case estimates for how good orthogonality and parallelism can be specified can thus be made by considering the accuracies of the machines in the local shop. Of course the skill of the machinist, type of fixturing used, and other factors also contribute to these errors. Bearing surfaces are often bolted in place, and thus some amount of hand alignment in the form of shimming, scraping, or lapping is possible to correct for angular errors.

Kinematic errors in a well-designed and manufactured machine should be very repeatable and can be compensated for if the controller is well designed. However, the fundamental principle of modern precision machine design is still to *maximize mechanical performance for a reasonable cost before using special controllers, algorithms, and sensors to correct for mechanical errors.*

2.3.3 External Load-induced Errors[32]

External loads that cause errors in a machine include gravity loads, cutting loads, and axis acceleration loads. They do not include errors caused by stresses generated internal to the machine caused by assembly errors (e.g., loads caused by tightening of bolts and forced geometric congruence between parts), which are discussed in Section 2.3.4.

The difficulty in modeling load-induced errors lies in their often distributed and/or varying effects. Other types of errors discussed thus far have been geometrically induced and were a function of position. Thus they could be relatively easily included in the HTM model of a machine. Load-induced errors, on the other hand, are often distributed throughout the structure, and thus in order to incorporate them into the HTM model, a method for lumping them at discrete points must be devised. Depending on the structure, the bearing interface is often the most compliant part of the structure and thus it can make sense to lump load-induced errors at the bearing interfaces. For more complex structures, it may be necessary to introduce additional coordinate frames into the HTM model.

Lumped Parameter Modeling Considerations

To illustrate conservative modeling of machine stiffness, consider the simple cantilever beam shown in Figure 2.3.4. It is desired to model the elastic beam as a rigid body connected to a wall with a spring at its base. If equivalent endpoint deflections were found for the cantilever beam and the spring model, the lateral system stiffnesses would match, and the resultant translational error would ripple through to the endpoint of other axes that may be attached to the end of the beam. However, the angular stiffness would be in error, and effects of Abbe (or sine) errors on other components that

[32] Also see Section 5.6 for a discussion of deformations caused by high contact stresses typically incurred at the interface between a contact-type sensor and a part.

might be mounted to the end of the beam would be seriously underestimated. If the lateral deflections were matched, then the lateral and angular stiffnesses and deflections (lateral and angular) would be:

	Cantilever beam	Angular spring	Deflection error
$K_{lateral}$	$3EI/L^3$	$3EI/L^3$	0.0
$K_{angular}$	$2EI/L^2$	$3EI/L^2$	33% low

On the other hand, if the slopes of the beams at their endpoints were matched, then Abbe errors would be correctly modeled and the translational error would be overestimated, which is conservative:

	Cantilever beam	Angular spring	Deflection error
$K_{lateral}$	$3EI/L^3$	$2EI/L^3$	33% high
$K_{angular}$	$2EI/L^2$	$2EI/L^2$	0.0

This example can be repeated for beams and plates held by almost any type of boundary conditions and loaded by almost any type of load. Remember that most machine components are bulky and require that shear strains also be accounted for. If the length-to-depth ratio is less than 3, shear strain deflections can account for an increasingly larger proportion of the total deflection. In many cases finite element analysis should be used to evaluate structural stiffnesses.

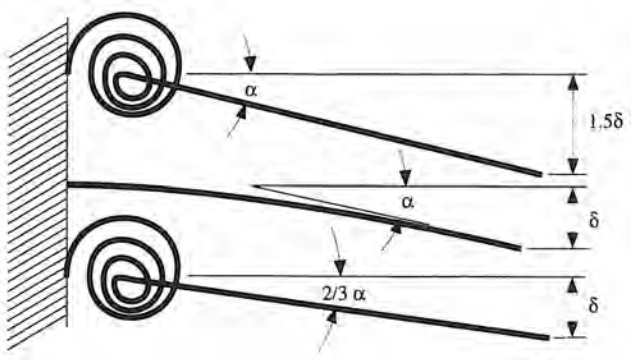


Figure 2.3.4 Two possible torsional spring-rigid beam models for a cantilever beam.

It should always be remembered that Abbe errors are some of the most often overlooked errors affecting the performance of an instrument or machine tool. Thus it is best to be conservative when estimating stiffnesses that give rise to angular motion, since you never know what will be bolted to the worktable and how it will amplify an angular error. However, you do know that translational errors are never amplified and are thus less likely to cause problems even if they are overestimated.

In order to become good at modeling machine structures, remember that most components become more monolithic and blocky in nature as the tool tip is approached. This is partly due to the fact that trimming material off these areas often does not save much in the way of material and labor costs unless the machine is meant for high-speed motion. As the base of the machine is approached, the structure often opens up into a truss type of configuration that can be modeled as a series of plates and spars. Regardless of the component, however, always consider local and global deformations and try to identify dominant stiffnesses (i.e., the most compliant member).

An example is the bolting of a linear bearing rail to a box section, as shown in Figure 2.3.5. Before rushing off to a handbook to try and find an equivalent load case, take a careful look at the structure and try to imagine what is happening. Draw several deformed shapes, and if you still cannot picture in your mind how it will deform, make a model from clay, foam rubber, or cardboard and load it accordingly. Remember St. Venant's principal.[33] If the beam were infinitely stiff, the bolts that held the rail to the beam and the carriage would probably be the most compliant members

[33] St. Venant's principal states: *The effect of forces or stresses applied over a small area may be treated as a statically equivalent system which, at a distance approximately equal to the width or thickness of a body, causes a stress distribution which follows a simple law.*

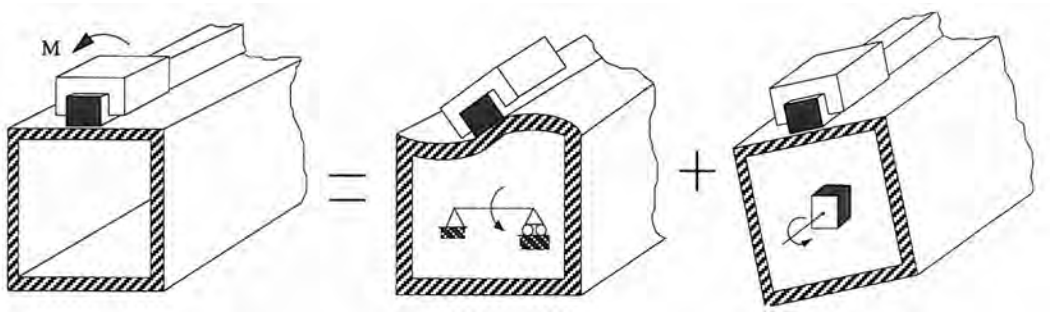


Figure 2.3.5 Modeling the torsional stiffness of a linear bearing mounted to a box beam by decomposition and superposition.

and the beam could be ignored. If the beam were made of thin-walled aluminum tube, then it would be fairly compliant and would have to be modeled as a flexing diaphragm and as a beam in torsion.

Once you have thought out how to model a structure, then pick up a handbook to get the equations for the deflection. If the equations for deflection are not there (and in most cases they will not be since shear deformations will be dominant), then derive them using energy methods.[34] After you make an attempt at modeling the structure, no matter how intractable it seems to be, do a simple finite element model to compare with your prediction. This will give insight as to why structures deform the way they do, thus enabling you to design more efficient structures in the first place. At any rate, the more "exact" the structure is to begin with, the fewer iterations you will have to go through to finish the job.

Once an equivalent spring for a part of the structure has been found, it can be lumped together in mechanical series with all other modeled springs at a common node (HTM reference frame). The total stiffness will be given by

$$K_{total} = \frac{1}{\sum_{i=1}^{N} \frac{1}{K_i}} \tag{2.3.3}$$

Estimations of the structure's lateral natural frequency using this stiffness will be conservative if springs modeled by the slope-matching method are used. However, caution must be exercised in evaluating equivalent stiffnesses of springs at discontinuous interfaces, such as bearings. In most cases, bearing manufacturers provide reasonably good stiffness data. Stiffness of bolted assemblies is also of concern and is discussed in Section 7.5. In many cases, finite element methods or model tests will be required to validate lumped parameter models of the structure's expected dynamic and static performance. One should realize that as many equivalent springs as needed can be used to model the six load-induced error components that can occur at every HTM reference frame. The error gains derived from the HTM method analysis can be applied to deflections caused by the springs as well as pure geometric errors. Thus the output from the error budget will also help to flag areas with insufficient stiffness.

2.3.3.1 Errors Caused by Gravity

Gravity loads every molecule in the machine and in the machine that was used to manufacture the machine's components. The latter effect is often the initial source of geometric errors in components.[35] Hence extreme care must be taken when evaluating the straightness of off-the-shelf components. One must ask, for example, how is the straightness specified a function of the method used to support the component? Similarly, one must evaluate how the weight of the workpiece and of other parts of the machine cause the structure as a whole to deform.

Gravity can be used to help provide a repeatable preload force to a machine's bearings; hence even as some bearings wear, dynamic performance can be maintained since the preload will remain

[34] This procedure is not as painful as many perceive it to be, and it is an excellent form of mental pushup to keep the mind lean, mean, and in competitive condition. Several examples are given throughout this book.

[35] "*A hen is only an egg's way of making another egg.*" Samuel Butler

constant. Thus a machine could periodically be remapped and new error corrections used to compensate for wear. This, in effect, would allow a machine to be rebuilt without ever having to take it apart. In general, however, machines that rely only on gravity to preload their bearings are not used for high-speed or high-force operation. High-speed requires low mass which, in turn, would provide insufficient preload to prevent the structure from lifting.

When designing a critical part for a machine, such as the frame which supports the bearings, the highest accuracy will be obtained if the part is designed so that it can be machined while being supported kinematically. This requires the part to be fixtured in a manner that is free of forced geometric congruence stresses (i.e., stresses imposed when bolting a straight piece to a curved machine bed, or vice versa). In some cases, this can lead to a vicious circle, for in order to support its own weight and resist machining forces, the part becomes so heavy that it will deform other parts, including the machine used to manufacture it. On the other hand, if this goal cannot be attained, then the part may not designed well enough to resist forces imposed on it by the machine of which it is to be a part. For this reason, each part must be carefully designed with consideration given to the type of manufacturing process that will be used to make it and the loads it is expected to bear.

Gravity loads are best modeled and evaluated using the HTM method by choosing equivalent torsional and extensional springs at coordinate frame interfaces. The equivalent springs should be chosen to match uniformly loaded models in a manner equivalent to that described in the previous example. The amount of load applied to the springs must include the weight of the structure, and the varying weight associated with the lightest and heaviest parts the machine will be required to manufacture.

Many methods have evolved over the years to compensate for deflections caused by the weight of a machine component. In addition to compensating curvatures, auxiliary structural frames have been built around machines to support counterweight systems. The former method made manufacturing difficult to control. The latter method increased the cost of the machine and essentially doubled the mass that most axis motors are required to move. The era of the laser interferometer and the microprocessor are changing the manner in which machines are designed. For example, if software corrections for mapped errors are to be incorporated into the design of the machine, then it is usually not necessary to worry too much about how the structure will deform under its own weight as long as the stiffness is high enough to allow for controllability and to prevent tool chatter. Often it is only necessary to make sure that extreme smoothness of motion exists in all axes and to make the structure rigid enough to resist dynamic loads.

Example: Deflection of a Simply Supported Beam Loaded by Its Own Weight

Consider the case of a generalized, symmetrical, isotropic beam, often used as a straightedge, loaded by its own weight and symmetrically supported as shown in Figure 2.3.6. Deflection of the beam will be caused by bending moments and shear stresses generated by the beam's own weight. When the beam is symmetrically supported and the deflection at the ends equals that in the middle, the minimum overall deflection will be achieved.[36] The exact solution for this problem requires use of the theory of elasticity which is impractical for everyday use; however, a good estimate can be made of the ideal support point location if bending and shear deformations are considered. Most engineers know how to find deflections due to bending or can look them up in a handbook, but equations for shear deformations are not found in most handbooks.

The first step is to write the loading function for the beam which, uses singularity functions[37] to activate the load at appropriate times:

$$q(x) = -w < x >^0 + \frac{w\ell}{2} < x - \ell_1 >_{-1} + \frac{w\ell}{2} < x - \ell_2 >_{-1} \tag{2.3.4}$$

The shear is

$$V(x) = -\int q(x)\,dx = wx - \frac{w\ell}{2} < x - \ell_1 >^0 - \frac{w\ell}{2} < x - \ell_2 >^0 + C_1 \tag{2.3.5}$$

[36] Another support condition that is desired for end standards (e.g., a standard meter bar) is locating the supports so that the slope of the end faces of the beams are vertical. This is known as supporting the beam at its *Airy* points. See F. H. Rolt, Gauges and Fine Measurements, University Microfilms, Ann Arbor, MI.

[37] An expression in $< \; >$ equals zero when the "exponent" is negative and the subscript or the expression inside the brackets is less than zero.

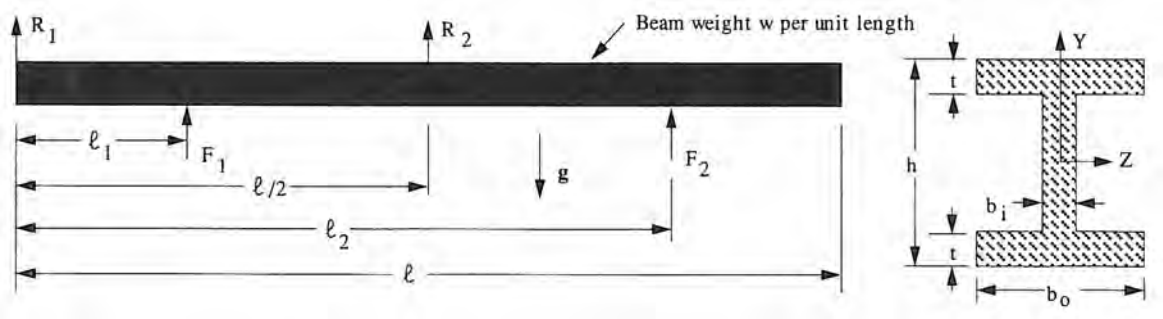


Figure 2.3.6 Generalized beam loaded by its own weight and applied fictitious forces R_i.

The shear is zero at x = 0, so $C_1 = 0$. The moment is

$$M(x) = -\int V(x)\,dx = \frac{-wx^2}{2} + \frac{w\ell}{2} < x - \ell_1 > + \frac{w\ell}{2} < x - \ell_2 > + C_2 \tag{2.3.6}$$

The moment is zero at x = 0, so $C_2 = 0$. The slope is found from

$$EI\alpha(x) = \int M(x)\,dx = \frac{-wx^3}{6} + \frac{w\ell}{4} < x - \ell_1 >^2 + \frac{w\ell}{4} < x - \ell_2 >^2 + C_3 \tag{2.3.7}$$

Generalizations cannot be made about the slope, so going on to the deflection:

$$EI\delta(x) = EI \int \alpha(x)\,dx = \frac{-wx^4}{24} + \frac{w\ell}{12} < x - \ell_1 >^3 + \frac{w\ell}{12} < x - \ell_2 >^3 + C_3 x + C_4 \tag{2.3.8}$$

With the boundary conditions $\delta 0$ at $x = \ell_1, \ell_2$, after considerable algebra (good mental pushups to quicken the mind) one finds the deflection at the end (1) and in the middle (2) to be, respectively[38]

$$\delta_{end} = \frac{w\ell_1}{24EI}\left[\ell_1^3(\ell_2 + \ell_1)(\ell_2^2 + \ell_1^2) + 2\ell(\ell_2 - \ell_1)^2\right] \tag{2.3.9}$$

$$\delta_{middle} = \frac{w}{24EI}\left[\frac{-\ell^4}{16} + \ell_1^4 + \frac{(\ell_2 - \ell_1}{2}\left[(\ell_2 + \ell_1)(\ell_2^2 + \ell_1^2) - \frac{3\ell}{2}(\ell_2 - \ell_1)^2\right]\right] \tag{2.3.10}$$

where w is the weight per unit length of the beam. For the generalized beam with material density ρ, w is

$$w = \rho\{2tb_o = b_i(h - 2t)\} \tag{2.3.11}$$

The slope deflections caused by shear stresses are of significance only when the length of the beam is on the order of three to five times the height or less. Most introductory mechanics of solids courses skip over this important topic because most engineers will never use it; however, as a machine design engineer, one will need this knowledge quite often.

To find the shear deflections at the points of interest, it is extremely useful to make use of energy methods (Castigliano's theorem), which requires fictitious forces, R_1 and R_2, respectively, to be applied at the points where the deflections are to be found. By symmetry and force-moment balance, the reaction forces can be found to be

$$F_1 = \frac{-R_1\ell_2}{\ell_2 - \ell_1} - \frac{R_2}{2} + \frac{w\ell}{2} \tag{2.3.12}$$

$$F_2 = \frac{R_1\ell_1}{\ell_2 - \ell_1} - \frac{R_2}{2} + \frac{w\ell}{2} \tag{2.3.13}$$

The shear force in the beam, which causes the shear deformation, is

$$V = -R_1 + wx - F_1 < x - \ell_1 >^0 - R_2 < x - \ell/2 >^0 - F_2 < x - \ell_2 >^0 \tag{2.3.14}$$

[38] Note that when $\ell_1 = 0$, $\ell_2 = \ell$, $\delta_{end} = 0$ and $\delta_{middle} = 5\omega\ell^4/384EI$. Similarly, when $\ell_1 = \ell_2 = l/2$, $\delta_{end} = \omega(l/2)^4/8EI$ and $\delta_{middle} = 0$.

The shear stress in a beam is

$$\tau = \frac{VQ}{bI} \tag{2.3.15}$$

where V is the shear force, Q is the first moment of the area of the cross-section from the outermost fiber to the level where the shear stress exists, b is the beam width at the point in the cross section where the shear stress is evaluated, and I is the second moment of the cross-sectional area of the beam about its neutral axis. Note that this solution assumes that the shear force is parabolically distributed across the width of the beam.[39] In addition, near the application of the shear force itself, the shear stress profile given by the engineering approximation (looks like a U on its side) is significantly in error compared to the distribution obtained from a theory of elasticity solution (looks like a W on its side).[40] However, at a distance equal to about one-fourth the beam height, the engineering solution is nearly correct, and at a distance of one-half the beam height, the engineering solution is essentially correct. The theory of elasticity solution is complicated for a rectangular cross-section beam and nearly intractable for the general case considered here or often found in practice. Hence the engineering solution is used throughout this book with the stipulation that for critical calculations, a detailed finite element study should be made.

For the generalized rectilinear cross-section beam the second moment of the area about the neutral axis is

$$I = \frac{b_0 h^3 - (b_0 - b_i)(h - 2t)^3}{12} \tag{2.3.16}$$

From the top of the flange to the bottom of the flange

$$\tau\Big|_{h/2-1}^{h/2} = \frac{V}{b_0 I}\int_y^{h/2} y b_0\, dy = \frac{v}{2I}\left[\left(\frac{h}{2}\right)^2 - y^2\right] \tag{2.3.17}$$

For the inner section (web)

$$\tau\Big|_0^{h/2-t} = \frac{Vt(h-t)}{2I} + \frac{V}{b_i I}\int_y^{h/2-t} y b_i\, dy = \frac{V}{2I}\left[\left(\frac{h}{2}\right)^2 - y^2\right] \tag{2.3.18}$$

With the shear stress known as a function of position for the various portions of the beam, the elastic strain energy can be found:

$$U_{shear} = \int_{Volume} \frac{\tau^2}{2G} dVolume \tag{2.3.19}$$

The volume increment will be b dy dx, where b is a function of y. Assuming that the beam's cross section is constant along the length of the beam, the integral can be expressed as

$$U_{shear} = \frac{1}{8GI^2}\int_0^{\ell} V^2 \left\{ b_i \int_{-\frac{h}{2}-t}^{\frac{h}{2}-t} \left[\left(\frac{h}{2}\right)^2 - y^2\right]^2 dy + 2b_0 \int_{\frac{h}{2}-t}^{\frac{h}{2}} \left[\left(\frac{h}{2}\right)^2 - y^2\right]^2 dy \right\} dx \tag{2.3.20}$$

Beginning with the innermost terms, the first yields

$$2b_i\left[\frac{h^5}{60} - \frac{h^2t^3}{3} + \frac{ht^4}{2} - \frac{t^5}{5}\right] \tag{2.3.21}$$

The second inner term yields

$$2b_0\left[\frac{h^2t^3}{3} - \frac{ht^4}{2} + \frac{t^5}{5}\right] \tag{2.3.22}$$

Note that if $b_o = b_i$, or t = 0, then the sum of the two terms is $bh^5/30$ which is what is obtained from evaluating the left inner integral from -h/2 to h/2 (e.g., a rectangular beam). To simplify the remainder of the problem, let K be equal to:

$$K = \frac{\frac{b_i h^5}{60} + (b_i - b_0)\left(-\frac{h^2t^3}{3} + \frac{ht^4}{2} - \frac{t^5}{5}\right)}{4GI^2} \tag{2.3.23}$$

[39] For the derivation of this simple engineering approximation, see any introductory mechanics of solids book (e.g., S. Timoshenko, Strength of Materials, Part I, 3rd ed. Robert Krieger Publishing Co. Melbourne, FL, pp. 113–118.)

[40] See S. Timoshenko and J. N. Goodier, Theory of Elasticity, McGraw-Hill Book Co. New York, 1951, pp. 57–59.

Next evaluate the integral:

$$U_{shear} = K \int V^2 \, dx \tag{2.3.24}$$

From Castigliano's theorem, the deflection at a point i, where the fictitious force R_i is applied, will be given by

$$\delta_i = \frac{\partial U}{\partial R_i} = 2K \int_0^{\ell} V \frac{\partial V}{\partial R_i} \, dx \tag{2.3.25}$$

Since the R's are fictitious forces they are set equal to zero after the differentiation and before evaluation of the integral. For i = 1, deflection at the end, the integral becomes

$$\begin{aligned} \delta_1 &= 2K \textstyle\int_0^{\ell} \left(wx - \frac{w\ell}{2} < x - \ell_1 >^0 - \frac{w\ell}{2} < x - \ell_2 >^0 \right) \\ &\times \left(-1 + \frac{\ell_2}{\ell_2 - \ell_1} < x - \ell_1 >^0 - \frac{\ell_1}{\ell_2 - \ell_1} < x - \ell_2 >^0 \right) dx \end{aligned} \tag{2.3.26}$$

This integral must be evaluated in three parts, from x = 0 to ℓ_1, x = ℓ_1 to ℓ_2, and x = ℓ_2 to ℓ, due to the presence of the singularity functions. The first segment is[41]

$$\lim_{\varepsilon \to 0} \int_0^{\ell_1} -wx \, dx = \frac{-w\ell_1^2}{2} \tag{2.3.27}$$

The second segment is found to be

$$\lim_{\varepsilon \to 0} \int_{\ell_1}^{\ell_2} \left(wx - \frac{w\ell}{2} \right) \left(-1 + \frac{\ell_2}{\ell_2 - \ell_1} \right) dx = 0 \tag{2.3.28}$$

The third segment is

$$\lim_{\varepsilon \to 0} \int_{\ell_2}^{\ell} \left(wx - \frac{w\ell}{2} - \frac{w\ell}{2} \right) \left(-1 + \frac{\ell_2}{\ell_2 - \ell_1} - \frac{\ell_1}{\ell_2 - \ell_1} \right) dx = 0 \tag{2.3.29}$$

Thus the deflection of the left-hand part of the beam due to shear forces caused by the weight of the beam is:

$$\delta_{\text{shear end}} = -Kw\ell_1^2 \tag{2.3.30}$$

For the middle section of the beam, the deflection integral becomes:

$$\begin{aligned} \delta_2 &= 2K \textstyle\int_0^{\ell} \left(wx - \frac{w\ell}{2} < x - \ell_1 >^0 - \frac{w\ell}{2} < x - \ell_2 >^0 \right) \\ &\times \left(- < x - \frac{\ell}{2} >^0 + \frac{< x - \ell_1 >^0}{2} + \frac{< x - \ell_2 >^0}{2} \right) dx \end{aligned} \tag{2.3.31}$$

This integral is evaluated in three parts, from x = ℓ_1 to l/2, x = l/2 to ℓ_2 and x = ℓ_2 to l. The first segment is

$$\lim_{\varepsilon \to 0} \int_{\ell_1}^{\ell_2} \left(wx - \frac{w\ell}{2} \right) \left(\frac{1}{2} \right) dx = \frac{w}{16} (4\ell_1 \ell_2 - \ell^2) \tag{2.3.32}$$

The second segment is

$$\lim_{\varepsilon \to 0} \int_{\ell_2}^{\ell_2} \left(wx - \frac{w\ell}{2} \right) \left(-1 + \frac{1}{2} \right) dx = \frac{w}{16} (4\ell_1 \ell_2 - \ell^2) \tag{2.3.33}$$

The third segment is

$$\lim_{\varepsilon \to 0} \int_{\ell_2}^{\ell} \left(wx - \frac{w\ell}{2} - \frac{w\ell}{2} \right) \left(-1 + \frac{1}{2} + \frac{1}{2} \right) = 0 \tag{2.3.34}$$

The deflection of the middle part of the beam due to shear forces caused by the weight of the beam is

$$\delta_{\text{shear middle}} = \frac{wK}{4} (4\ell_1 \ell_2 - \ell^2) \tag{2.3.35}$$

[41] Note that when x = ℓ_1, < x - ℓ_1 >^{0} equals <0>^{0}, which equals 1; however, this would mess up the integration, so we actually evaluate the integral from 0 to $\ell_1 - \varepsilon$, where ε is one-billionth of the width of an iron atom, which is too small to affect the accuracy of our calculations, so $< -\varepsilon >^0 = 0$.

In order to find the optimum values for ℓ_1 and ℓ_2, the following equation must be minimized

$$\Sigma = \{\delta_{\text{end bending}} + \delta_{\text{end shear}}\} - \{\delta_{\text{middle bending}} + \delta_{\text{middle shear}}\} \tag{2.3.36}$$

This equation is best solved numerically in an iterative manner. If the shear deformations are ignored, the solution is independent of the beam cross section and $\ell_1 = 0.2232\text{l}$ and $\ell_2 = 0.7768\ell$. Note that the slope above the support points will have a finite value not necessarily equal to zero. If the bending deformations are ignored, then the solution is also independent of beam cross section and $\ell_1 = 0.2500\ell$ and $\ell_2 = 0.7500\ell$. Since the magnitude of the bending and shear deflections each change in their own unique nonlinear way, for the combined case the point of minimum deflection will depend on the beam cross section, but will usually lie somewhere between the two values.

Most beams are designed with the idea that as much mass as possible should be located as far away from the beam's neutral axis as is possible. This concentrates the beam's mass in regions of maximum stress; this is why for maximum stiffness-to-weight ratios, I beams with holes in their center web are used. The maximum shear stress in a beam, however, is located along the neutral axis, and the deflection is a more complicated function of geometry and loading, as shown by the preceding calculations and the results in the tables. The method used for this example can be used to obtain a good estimate of deformation of virtually any type of beam under virtually any type of loading.[42] All that is needed is a little time to carefully reevaluate the integrals.

2.3.3.2 Errors Caused by Accelerating Axes

In addition to increasing accuracy for enhancement of quality, machines are being required to move at greater speeds in order to increase productivity. Machine tools are usually thought of as big, bulky, slow-moving structures. The next generation of machine tools, however, will probably require axes to have acceleration capabilities in excess of 1 g. As an example, consider the situation shown in Figure 2.3.7. A high-speed machining center has an X-axis carriage that holds two other axes and the spindle. For the simple two dimensional case shown, it is assumed that the dominant compliance in the system is due to the X-axis bearings, whose deflection also causes the greatest Abbe errors. Assume that the Z axis, which holds the spindle and 20-hp motor, has a mass M of 682 kg, is supported by bearings with stiffness of 350 MN/m (2×10^6 lb/in.), and is required to accelerate and decelerate at 0.5g while drilling or boring a hole. The acceleration a produces deflections in the Y and Z directions and about the X axis. For geometry of the case shown, the equations describing the Y-axis deflection of the rear X axis bearing and the roll of the structure about the X axis are (units are kg, m, s)

$$\delta_{\text{y-rear bearing}} = \frac{-1.016\text{aM}}{0.726\text{K}_{\text{x axis bearing}}} \tag{2.3.37}$$

$$\theta_{\text{x}} = \frac{\delta_{\text{y-rear bearing}} - \delta_{\text{y-front bearing}}}{0.762} \tag{2.3.38}$$

When the spindle starts accelerating into a part at 0.5g, the resultant bearing deflections are found to be $\delta_{\text{y-front bearing}} = 12.7\mu\text{m}$ (500 μin.), and $\delta_{\text{y-rear bearing}}$= -12.7 μm (-500 μin). The pitch $\varepsilon_{\text{x}}$of the structure is thus -33.3 μrad. Assuming that the tool tip's coordinate reference frame origin is nominally located at the origin of the machine's reference XYZ coordinate system, the tool tip's HTM is

$${}^{\text{R}}\text{T}_{\text{tool tip}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 3.33 \times 10^{-5} & 1.27 \times 10^{-5} \\ 0 & -3.33 \times 10^{-5} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.3.39}$$

When the tool is at its forward position, Z = 0.508 m (20 in.) from the front bearing, the HTM is postmultiplied by the toolpoint vector $\mathbf{TP}^{\text{T}} = [01.0160.5081]$, which describes the position of

[42] Note that an exact solution for this type of problem would require the application of the general theory of elasticity, which is needed to account for the strain field in the region of the supports. For simple shapes and loads, such as a beam with rectangular cross section supported on knife edges, this can be done, although the mathematics become very complex with respect to those used above. Also for more complex problems, such as an I beam with other added loads, the procedure becomes almost unmanageably complex. R. Reed solved the optimal support location problem for a beam with rectangular cross section ("A Glass Reference Surface for Quality Control Measurements," Int. J. Mech. Sci., Vol. 9, 1966) and showed that when bending theory was used, the positions of the knife edges obtained resulted in deflections that were about twice those that were obtained if the optimal knife-edge positions were found using a solution based on the theory of elasticity.

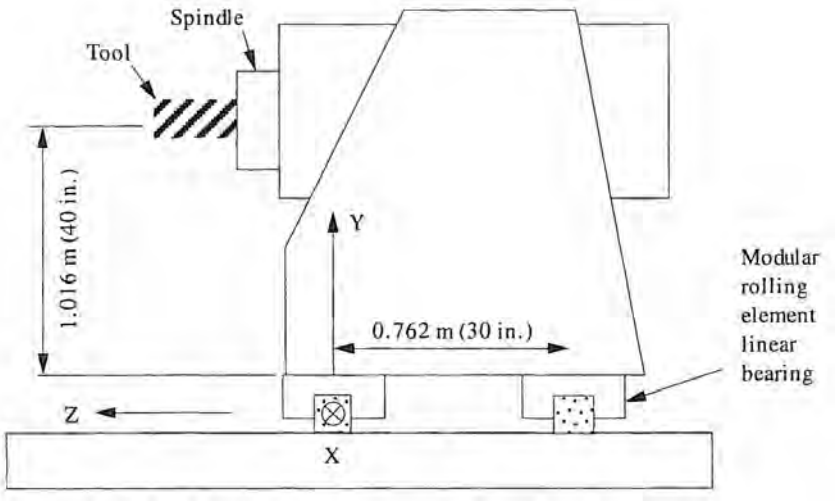


Figure 2.3.7 Machine geometry for inertial load example.

the tool in its own coordinate system. The resultant position vector $\mathbf{TR}^{\mathrm{T}}$ locating the tool in the machine's XYZ reference coordinate system is [0 1.0160296 0.5079662 1]. The resultant tool tip error is thus $\delta_y = 29.6\mu$ m (1165 μin.) and $\delta_z = -33.8\mu$m (-1332 g μin.). Note that these values would scale linearly with the stiffness of the X axis bearings.

This deflection caused by accelerating the mass is acceptable for many drilling operations which would allow high spindle speeds and high axis feed rates to be used to increase productivity. Ramp and dec (acceleration and deceleration) rates of high-speed manufacturing equipment's axes may one day routinely approach several g. In designing this type of equipment, accuracy along the path of motion may or may not be critical. However, final placement and settling time, where the maximum accelerations and inertial forces are present, will be important. Note that in the design of this type of machinery, cutting forces are often insignificant compared to inertial forces.

2.3.3.3 Errors Caused by Cutting Forces

Another major contributor to load-induced errors in machine tools and some robots are cutting forces. Fortunately, high-speed cutting processes often generate only low cutting forces, so at least the problems of high acceleration and high cutting force usually do not occur simultaneously. However, cutting forces are applied at the tool tip and act on every element in the machine (spring in the HTM model). In order to estimate forces generated by the cutting process, actual cutting forces on machines with similar tools should be measured or appropriate handbooks consulted. In many cases the rapid advance of new types of cutting tool materials and tool shapes will require the design engineer to consult with a tooling manufacturer or make experimental tests. Remember that when the tool contacts the workpiece, it can weakly close the machine's structural loop and thus change stiffness characteristics. This effect and the effect of chatter are discussed in Section 2.4.

When cutting or assembly (arising in automatic assembly equipment) forces are used as inputs to the HTM to evaluate machine deflection under load, it must not necessarily be assumed that errors can be compensated for electronically, because the process may take place too quickly for the servosystem to compensate for errors. An exception to this rule of thumb may be some slow-speed, high-precision diamond turning operations using single-axis piezoelectric-actuated fast tool servos to correct for radial error at the tool tip.[43] A device of this type was designed and built at Lawrence Livermore Laboratory and had a resolution of 0.01 μm, a range of 5–6 μm, and a bandwidth of 50 Hz. In most cases, however, attempts at compensating for the magnitude of cutting load errors will be out of phase with the occurrence of the error. The brittle nature and limited force characteristics of most piezoelectric actuators may also prevent their use in many machining applications.

[43] See E. Kouno and P. McKeown, "A Fast Response Piezoelectric Actuator for Servo Correction of Systematic Errors in Precision Engineering," Ann. CIRP, Vol. 33, 1984, pp. 369–373. Also see S. Patterson and E. Magrab, "Design and Testing of a Fast Tool Servo for Diamond Turning," Precis. Eng., Vol. 7, No. 3, 1985, pp. 123–128. Also see Figure 10.5.1.

2.3.4 Load-induced Errors from Machine Assembly

Even if all machine components are within required tolerance prior to assembly, additional load-induced errors can be introduced during assembly. The first type of error is forced geometric congruence between moving parts. The effects of this type of error are illustrated by the design case study presented in Section 2.5.

A common example is mounting a mirror at its four corners, which often creates a visible distortion (see Section 7.4.6). The second type of error is the effect of the assembly process on the stiffness of the structure itself, and how the stiffness can be evaluated and incorporated into the HTM model. A third type of error, one that can be predicted, is the deformation of the machine when forces are applied to preload bearings and bolts.[44] In addition, errors may also be caused by clamping or locking mechanisms.

No part is made to exact specified dimensions, and when two such inexact parts are bolted together, they attain an equilibrium shape. The resultant internal stresses can also lead to apparent material instabilities. Traditionally designed machines minimized this problem by hand finishing components to fit before final assembly. Today, on the 1- to 2- μm level, modern grinders and careful fixturing procedures can also yield closely matched surfaces. Below 1 - 2 μm, however, gravity errors become dominant and exceptional care is needed to achieve higher accuracy. Often the only way to match two surfaces is to use the technique of scraping described in Section 7.2.4. The alternative is to mount parts kinematically. The former requires highly skilled personnel, which are harder and harder to find these days. The latter asserts that alignment between two surfaces should be exactly kinematic; however, this requires a change in design methodology that can increase structural costs but also increases reliability of design performance prediction.

As discussed earlier, elastic averaging is based upon the principle that the mean surfaces of two parts can be finished to match exactly, and then only small errors in surface finish prevent the atomic planes from bonding together (i.e., a cold weld). By hand finishing or precision grinding two flat surfaces, or by providing hundreds of meshing "gear teeth,"[45] the two parts can be made to overcome their surface finish discrepancies. Applying a high clamping force causes the asperities, the peaks and valleys that populate the part surface on the submicron level, to mesh together. This has been a most successful method for building indexing tables, as repeated clamping and unclamping will eventually wear-in the two surfaces. The problem with wearing-in two surfaces in this manner, however, is that they will not yield the same accuracy if a new un-worn-in component is added to the group (e.g., a pallet holding system that has a new pallet added). The latter condition depends, of course, on the level of accuracy being aimed for and whether or not the parts are to be joined together permanently. In the extreme case, parts that are flat enough can be permanently "cold welded" together by carefully finishing the mating surfaces, cleaning the parts, putting them together with dry alcohol between the surfaces, placing the assembly in a vacuum chamber, and pumping out the air. When done properly, the surfaces fuse and the assembly behaves like a single piece. This process is also sometimes referred to as *optical contacting*.

Kinematic design represents implementation of the fact that only one coordinate is required to define each degree of freedom in a body. This results in a holonomic system that is uniquely defined by a set of closed-form mathematical equations. This is the principle that makes a three-legged chair stable as opposed to some rigid, four-legged chairs that wobble because of one short leg. Kinematically designed systems, however, usually have much heavier structures to support themselves between the points of contact. Analysis of their characteristics, however, is easier because of the closed-form nature of their behavior (see Section 7.7 for detailed design equations).

If the surfaces are to be mated for life, then the stiffness of the joint should be considered along with the effects of forced congruence on the accuracy of the overall structure. First consider the effect of forcing two structures together. The mean of their errors can be computed and entered into the error budget. However, with large surface areas to be mated, what are the chances of a dirt particle becoming embedded between the surfaces? Can this variable in the manufacturing process cause problems? To design around this problem, a surface can have channels cut in it to act as a trap for particles. A particle moving between two surfaces only will travel half as far, so manufacturing

[44] For analysis methods, see Section 7.5 and Chapter 8.

[45] Curvic and Hirth couplings are often used on indexing devices to achieve high rotational indexing accuracy and high stiffness. These types of couplings are essentially two face gears that are clamped together (see Figure 6.1.5).

requirements for how much the surfaces are to be *wrung* together in order to force contaminants into the channels need to be specified. Second, the stiffness per unit area of bolted joints decreases rapidly with increasing surface finish amplitude.[46]

The effect of interface surface finish on the stiffness of a bolted joint can be characterized as a collection of randomly sized prismatic rods. The rods are under simultaneous axial compression, bending, and shear. Models describing this phenomenon provide good qualitative agreement with experiments, but it is difficult to arrive at quantitative conclusions. In general, it is desirable to make the surface as smooth as possible in order to maximize stiffness. Rougher surfaces exhibit greater damping characteristics through greater elastic strain energy expenditures, but also have correspondingly lower stiffnesses.[47] However, there is a limit to how fine a surface finish should be specified. Finishes finer than about 0.1 μm give rise to the problem of the absence of room to accommodate contaminants in the air which tend to settle on surfaces. Unless the components are to be assembled in clean rooms, it may not be advantageous to specify higher finishes. In the extreme case, finer surfaces can also lead to cold welding (optical contacting) of parts before they are properly aligned. In addition to the much-studied aspect of stiffness in a direction normal to a surface, the shear stiffness of a joint is of prime importance. Shear stiffness is more sensitive to contact pressure because it is primarily friction dependent. In general, however, joint friction provides more then adequate shear stiffness (see Figure 7.5.7). Dowel pins can help increase shear stiffness, but their primary purpose is often to allow parts to be relocated after disassembly.

The number of bolts specified along with their torque levels can have a profound effect on the final shape of an assembled part. For example, consider a round flange held by two bolts. The tightening of the bolts locally compresses the upper surface of the flange, which causes it to curve up. As more and more bolts are added, the magnitude of the bending deformation is reduced as the surface approaches a state of uniform compression. On a bolt circle for assemblies that have critical dimensional stability requirements, a good rule of thumb is to make sure that the bolts are spaced on the circumference not more than three or four bolt diameters apart and they are tightened so that the stress in the threads is only 10-20% of yield. On a linear arrangement of bolts, too few bolts, or overtightened bolts, can warp a bearing way and give the impression of increased straightness and angular (pitch) errors.

Tightening of bolts should be preceded by assuring that the threads and interface between bolt head and part surfaces are clean and lubricated. This is done by initially screwing the bolts in the mating piece deeper than they will go when the parts are assembled. During assembly, a lubricant should be applied to the threads and underneath the head. This will minimize friction and help to ensure uniformity of preload. Where vibration is present, a hardening thread lubricant can be used. The hardening lubricants also act to better distribute loads among the threads, which has the effect of increasing the apparent stiffness of the bolt.[48]

A third type of mounting is a compromise between kinematic and elastic averaging methods. This third option, *replication*, is an offshoot of the practice of grouting. Grouting is the process of locating two surfaces with respect to each other and then injecting a bonding agent (e.g., cement or epoxy) between them. This does not deform the structures since geometric congruence is not forced, yet it yields a high stiffness interface since voids between the surfaces are filled. To be effective, however, grouting requires a finite gap between the surfaces. Replication involves pouring a low-shrinkage epoxy around the form of a master jig. The jig is removed and the part is then bolted in the impression left by the jig. Grouting and replication are discussed in greater detail in Section 7.5.2.

The final act of assembly is installation of the machine at the customer's plant. For small-to-medium-sized machines with footprints less than about 2m × 2m, it is feasible to use a three-point kinematic mount. This ensures that the machine is not deformed by an uneven or unstable floor. Very large machines, such as planing mills, require many points of support under their bases. Installation of such machines often requires them to be adjusted on site to compensate for varying properties in the foundation.

---

[46] See D. M. Abrams and L. Kops, "Effect of Waviness on Normal Contact Stiffness of Machine Tool Joints," Ann. CIRP, Vol. 34, 1985, pp. 327–330.

[47] See M. Burdekin et al., "An Elastic Mechanism for the Microsliding Characteristics between Contacting Machined Surfaces," J. Mech. Eng., Vol. 20, No. 3, 1978.

[48] See Section 7.5.1 for a more detailed discussion of bolted joint design and bolting procedures.

2.3.5 Errors Caused by Thermal Expansion[49]

The need for ever-increasing accuracy and greater machine speeds makes thermal errors ever the more important to control. Errors caused by thermal expansion are among the largest, most overlooked, and misunderstood form of error in the world of machine design. Thermal errors affect the machine, the part, and the tool. Even the warmth of a machinist's body can disrupt the accuracy of an ultraprecision machine. Figure 2.3.8 shows thermal effects that must be accounted for in the design of a precision machine.[50] Thermal errors are particularly bothersome because they often cause angular errors that lead to Abbe errors.

Temperature changes induce thermal elastic strains, ε_T, that are proportional to the product of the coefficient of expansion, α, of the material and the temperature change, ΔT, experienced by the material:

$$\varepsilon_T = \alpha \Delta T \tag{2.3.40}$$

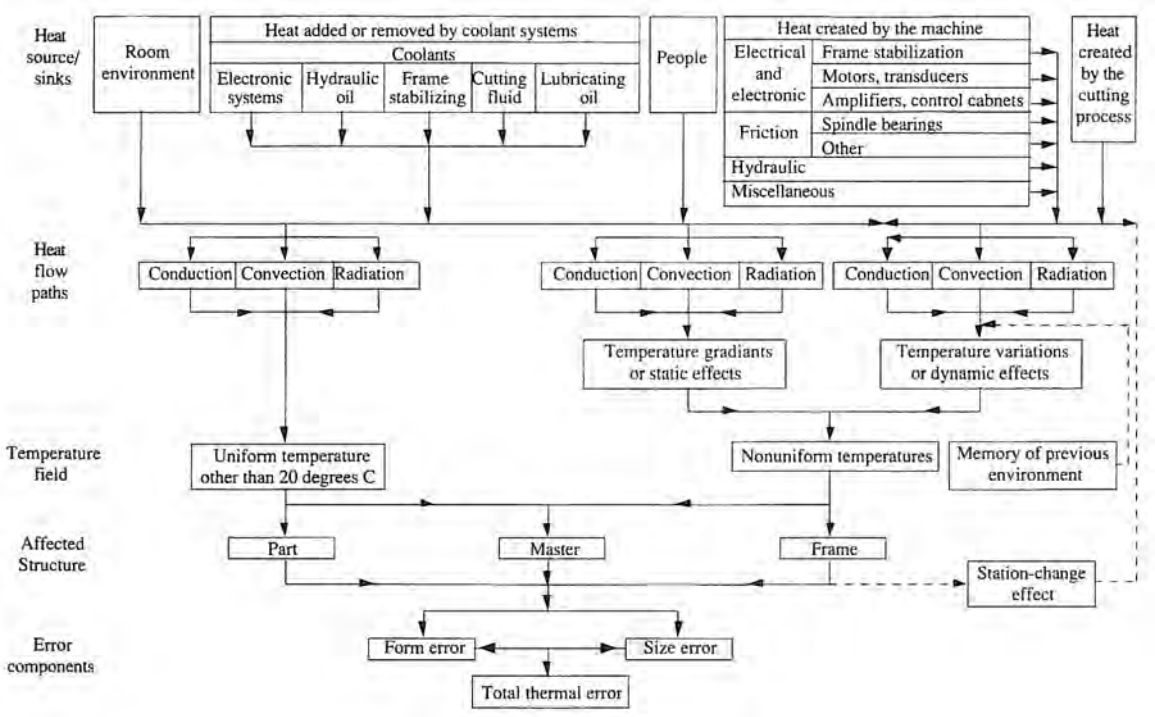


Figure 2.3.8 Thermal effects in manufacturing and metrology. (After Bryan.)

In addition, temperature gradients often cause angular errors, which lead to Abbe errors. If a machine, a tool, and a workpiece all expanded the same amount and could all be kept at the same temperature, then the system might expand uniformly with respect to the standard (always measured at 20°C) and everything would be within tolerance when brought back to standard temperature. However, different metals manufactured at different temperatures can experience serious dimensional metrology problems. Fortunately, standards have been developed (e.g., ANSI B89.6.2[51]) that define in great detail the effects of temperature and humidity on dimensional measurement and how measurements of these effects should be made. In this section, examples of analysis methods for estimating thermal deformations will be discussed. Although thermal strains can be minimized by using materials that do not expand very much, and by accurately controlling the environment; in practice, however, these can be difficult solutions to implement for economic reasons.

---

[49] C° denotes temperature difference whereas °C denotes an absolute temperature. This helps to avoid confusion, particularly when reference is made to temperature increases above ambient.

[50] J. Bryan, figure presented in the keynote address to the International Status of Thermal Error Research, Ann CIRP, Vol. 16, 1968. Also see R. McClure et al., "3.0 Quasistatic Machine Tool Errors," Technology of Machine Tools, Vol. 5, Machine Tool Accuracy, NTIS UCRL-52960-5, Oct. 1980.

[51] Available from ASME, 22 Law Drive, Box 2350, Fairfield, NJ 07007-2350, (201) 882-1167.

Metals can be designed to have low coefficients of expansion and acceptable dimensional stability properties; however, generally their coefficient of expansion will vary with temperature (e.g., Invar with $\alpha < 1\mu m/m/C^\circ$). These materials are often used in instruments and sometimes in machine hot spots such as spindles, which also often utilize active cooling systems. Other materials which have low coefficients of expansion include some types of ceramics (e.g., Zerodur® with $\alpha = 0.0 - 0.1\mu m/m/C^\circ$), which are also very stable; however, many of these materials also have low thermal conductivity, which can lead to hot spots and greater local deformations.

Most machine structures are still made from cast iron, which is inexpensive, stable, well damped, and machines easily. On the other hand, as an example of the problems that can occur in a machine made from cast iron, consider that gray cast iron can have an outer shell of white cast iron on it as it comes from the mold, although in most cases the thickness is inconsequential. The difference in thermal expansion between the two is $\alpha_{\text{whiteiron}} = 9.45\mu m/m/C^\circ$ and $\alpha_{\text{gray iron}} = 11.0\mu m/m/C^\circ$; thus when specifying machining of cast parts, bending deformations caused by differential thermal strains must be considered.[52] In many cases, annealing at very high temperatures can cause variations in material composition (e.g., white iron) to diffuse out, yielding a uniform structure. If structures are not properly heat treated and a white iron shell is retained, stability and thermal growth problems may occur.

2.3.5.1 Examples of Thermal Errors

In any room, warm air rises and cold air sinks, which gives rise to temperature gradients. Errors imposed by a temperature gradient on a large machine structure can be significant. If the height of the machine is significantly greater than the characteristic dimension of the footprint, the machine will probably grow taller without warping. For a linear temperature gradient, the amount of growth is equal to

$$\delta = 0.5\alpha h(T_{top} - T_{base}) \tag{2.3.41}$$

In a typical industrial inspection room, machines where the vertical distance from the table to the tool tip is on the order of 1 m may be subjected to a 1 C° gradient from the environment. For a one meter tall cast iron structure ($\alpha = 11\mu m/m/C^\circ$), this can lead to a 5.5 μm (220 μin.) error. Note that in many machine tools, the gradient caused by heat from motors and bearings may be an order of magnitude greater. For precision coordinate measuring machines, however, environmental gradients usually dominate.

For a big machine where the characteristic footprint dimension is larger than the height of the machine (e.g., a large thick surface plate), substantial thermally induced bending strains can exist. The thermal strain ε_T at an elevation y from the neutral axis in a horizontal beam of height h due to a temperature gradient $\Delta T = T_{top} - T_{bottom}$ is $\varepsilon_T = \alpha y \Delta T/h$, which is equivalent to a bending strain, which would also be dependent on the y coordinate. From elastic beam theory, the radius of curvature of the thermally deformed beam is related to the thermal strain by

$$\varepsilon_T = \frac{y}{\rho} = \frac{\alpha y \Delta T}{h} \tag{2.3.42}$$

The bending moment is related to the curvature by

$$M = \frac{EI}{\rho} \tag{2.3.43}$$

The resulting thermally induced error δ_T in the beam's straightness is equal to the tip deflection of a cantilever beam, of half the length of the beam being deformed, with a moment applied at its end:

$$\delta_T = \frac{M(\ell/2)^2}{2EI} = \frac{\ell^2 \alpha \Delta T}{8h} \tag{2.3.44}$$

The thermally induced slope error θ_T at the ends of the plate, which cause Abbe errors in parts mounted to the plate, is given by

$$\theta_T = \frac{\alpha \Delta T \ell}{2h} \tag{2.3.45}$$

[52] R. Wolfbauer, "The Effect of Heat on High Precision Machines Using a Coordinate Controlled Machine as an Example," Maschin Markt, Vol. 1, Oct. 1957 (MTIRA Translation T.1).

For a cast iron surface plate (a large flat plate) that is 1×0.3m, and a temperature difference of $\Delta T = 1/3C^\circ$, perhaps caused by radiant coupling between the top of a surface plate and strong overhead lights, the errors are $\delta = 1.5\mu$m (60 μin.) and $\theta_T = 6.1\mu$rad. For precision applications, these errors can be very significant, and thus temperature gradients must be carefully controlled and materials carefully selected as discussed in Section 7.3.2.

A bimaterial structure can be subject to deformations even in the absence of a gradient. For example, consider a cast iron surface plate that has not been properly heat treated. A layer of white iron may exist on one side, while the other side has been machined flat so that the surface is of gray iron. This structure will bow if it is used at any temperature other than that at which it was manufactured, as is illustrated in the following example and modeled in Figure 2.3.9. Another example would be a polymer concrete base with steel bearing rails bonded to its surface.

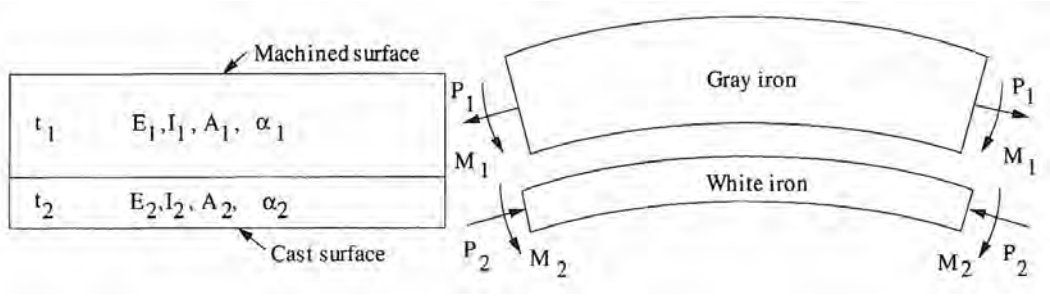


Figure 2.3.9 Bending of a bimaterial structure caused by differential thermal expansion.

To determine the amount of deformation that can occur in these situations, note that shear forces and strains are not present in this case, even for short beams. Forces between the two elements are assumed to act through the neutral axis of each respective element, thus acting as a couple which balances the respective bending moments that are generated in each beam. With the assumption that the white iron layer is much thinner than the gray iron layer, $t_2 << t_1$, the force-moment balance is

$$\frac{Pt_1}{2} = M_1 + M_2 = \frac{E_1 I_1}{\rho} + \frac{E_2 I_2}{\rho} \tag{2.3.46}$$

Geometric compatibility at the interface is visualized by taking two independent beams and subjecting them to thermal axial and bending loads so that the bottom side of one beam fits into the top side of the other. The gray iron layer is compressed by the white iron, which has a lower coefficient of expansion ($\alpha = 9.5\mu$m/m/C$^\circ$). The white iron is pulled by the gray iron and so its upper surface is in tension due to bending:

$$\alpha_1 \Delta T - \frac{P}{E_1 A_1} - \frac{t_1}{2\rho} = \alpha_2 \Delta T + \frac{P}{E_2 A_2} + \frac{t_2}{2\rho} \tag{2.3.47}$$

Substituting from Equation 2.3.46 for the load P into Equation 2.3.47 and solving for the radius of curvature ρ yields

$$\frac{1}{\rho} = \frac{(\alpha_1 - \alpha_2)\Delta T}{\frac{t_1 - t_2}{2} + \frac{2}{t_1}(E_1 I_1 + E_2 I_2)\left(\frac{1}{E_1 A_1}\frac{1}{E_2 A_2}\right)} \tag{2.3.48}$$

From elastic beam theory $1/\rho = M/EI$; hence the bow of the plate will be $\delta = M(l/2)^2/2EI$, and the slope $\theta = M(l/2)/EI$. Therefore, the deflection and slope of the beam are

$$\delta = \frac{(\alpha_1 - \alpha_2)\Delta T(\ell/2)^2}{t_1 - t_2 + \frac{4}{t_1}(E_1 I_1 + E_2 I_2)\left(\frac{1}{E_1 A_1} + \frac{1}{E_2 A_2}\right)} \tag{2.3.49}$$

$$\theta = \frac{(\alpha_1 - \alpha_2)\Delta T(\ell/2)}{\frac{t_1 - t_2}{2} + \frac{2}{t_1}(E_1 I_1 + E_2 I_2)\left(\frac{1}{E_1 A_1} + \frac{1}{E_2 A_2}\right)} \tag{2.3.50}$$

Consider the case of a 1 m^2 surface plate 0.3 m deep with 3 cm wall thickness. If the top is ground flat while the bottom retains a 0.5 cm layer of white iron, then assuming that the elastic moduli are

about equal, the deflection will be on the order of 0.10 μm for a 1 C° temperature gradient across the plate. The angular error will be on the order of $\theta = 0.41\mu$rad. These errors are not significant for many applications, but could be if the surface plate is used as a flatness reference for building precision diamond turning machine components.

As another example, consider using materials with different coefficients of expansion to cancel out thermal strains. When steel bolts are used to connect components with a higher or lower coefficient of thermal expansion than steel, an aluminum or Invar sleeve can be used to maintain constant tension in the bolt, as shown in Figure 2.3.10. The relative length of the sleeve and the bolt can be tuned to have the same amount of expansion as occurs in the length of the column of metal from the center of the threads to the base of the sleeve. To prevent ambiguity concerning the contact region, the engaged threaded length should not be longer than one bolt diameter. Assuming a temperature gradient T(x) in the direction of the bolt length, the differential increment of expansion $d\delta$ is $\alpha T(x)\,dx$. The required length of the sleeve is found by equating the deflections of the assembly to the deflection of the bolt head:

$$\int_0^{h_1} \alpha_1 T(x)\,dx + \int_{h_1}^{h_2} \alpha_2 T(x)\,dx + \int_{h_2}^{h_3} \alpha_s T(x)\,dx = \int_0^{h_3} \alpha_b T(x)\,dx \tag{2.3.51}$$

If the assumption of a uniform temperature is made, then the required sleeve length is

$$h_3 = \frac{h_1(\alpha_1 - \alpha_2) + h_2(\alpha_2 - \alpha_s)}{\alpha_b - \alpha_s} \tag{2.3.52}$$

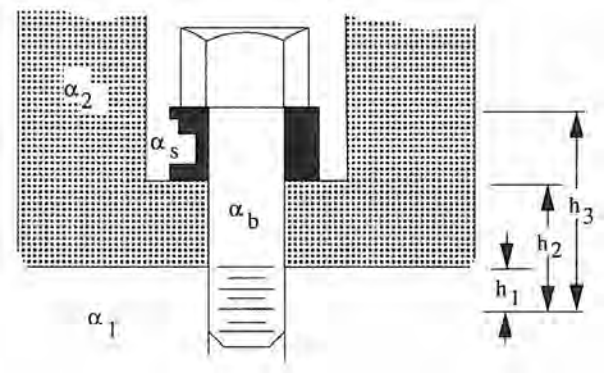


Figure 2.3.10 Matching thermal expansions to maintain constant preload.

Since joint stiffness is a strong function of interface pressure, and changing of the pressure also changes the deflection of the joint itself, maintaining constant joint pressure in the face of changing temperature is very important.[53] In general, when designing structures to resist thermal loads, try to use closed sections, such as a square tube as opposed to an I beam, because the former tend to dissipate gradients more readily. The best defense, however, is a good offense, which requires elimination of heat sources and gradients from the machine and its environment.

Temperature variations also affect readings from laser interferometers, autocollimators, and capacitance probes and other sensors. For example, atmospheric temperature changes cause errors on the order of 1μm/m/C° in laser interferometers. Corrections can be made through the use of refractive index measuring devices (refractometers); however, in some advanced machines the laser beam is buried deep within the machine as a feedback element, and it is difficult to accurately measure the local atmospheric conditions, including content of hydrocarbon products in the air. As a further example, consider that when using an autocollimator to measure straightness of a part, a lateral temperature gradient of only 1 C°/m causes an error on the order of 1.0 μrad/m. Capacitance probes can be subject to changes of 400 μm/m/C°; however they are usually only used over a range on the order of 10 μm, so total thermal error would be 40 Å/C°. These values are listed here to

[53] See for example A. Slocum, "Design to Limit Thermal Effects on Linear Motion Bearing Performance," Int. J. Mach. Tools Manuf., Vol. 27, No. 2, 1987, pp. 239–245.

give a general feeling for the sensitivity of measurements. Temperature-induced errors in sensors are discussed in greater detail in Chapters 3–5.

As is shown by the examples above, not only are the average temperature of the room and the machine important, but the gradient and changes in gradients are important as well. Horizontal gradients can be controlled with good cross flow and by the fact that air likes to settle in layers of even temperature (stratify). Vertical gradients are more difficult to prevent.

2.3.5.2 Identification, Control, and Isolation of Heat Sources[54]

Heat that causes thermal errors is introduced into the machine from a number of sources including moving parts (e.g., high-speed spindles, leadscrew nuts, linear bearings, transmissions), motors, the material removal process, and the external environment (e.g., sunshine through a window, direct incandescent lighting, heating ducts, the floor, operator's body heat, etc.). Heat transfer mechanisms in a machine include conduction, convection, evaporation,[55] and radiation. They are executed by recirculating lubricating oil and cutting fluid, chips from the cutting process, conduction through the frame of the machine, convection of air in and around the machine, and internal and external sources of radiation.

Obviously, some heat sources are more important than others; the sensitivity of the machine to different heat sources can be judged by modeling it as a series of thermal expansion elements that are included in the HTM method of formulating the system error budget. Only then can it be determined what the allowable range of temperature variations are for a particular component. Once this preliminary estimate is made and a structure is designed, a more detailed finite element model can be used to verify and correct the model used in the HTM model.

There are three schools of thought regarding minimization of thermally induced errors. The first is to prevent thermal expansion in the first place. The second is to minimize the time it takes for the machine to reach its equilibrium temperature, and in some cases, bring the entire machine to a uniform temperature (i.e., warm up the machine), which can help to minimize differential expansion. The third is to disregard the effects of thermal errors and simply map them, which is not easy. Regardless of the methodology followed, a thorough understanding of the effects of heat sources and transfer mechanisms is required.

Motors and spindles generate the greatest amounts of heat. One way to minimize heat transmission into the rest of the machine is by the use of radiation shields and high-thermal-resistance mounting materials (e.g., some ceramics) to form a *thermal break,* as shown in Figure 2.3.11. In this example, the temperature of the polymer concrete section could be monitored and used to control the flow of a cooling media. Mounting a motor externally and using a piece of reflective sheet metal as a radiation shield in an open space between a hot motor casing and a cast iron base can also significantly reduce heat transfer of the motor into the machine.

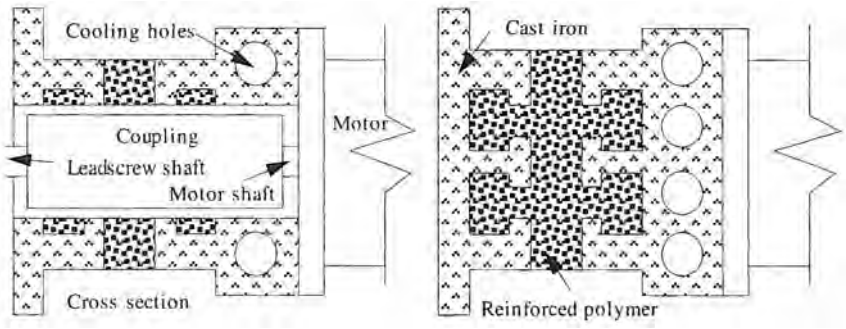


Figure 2.3.11 Conceptual drawing of a motor thermal break.

Radiation from external sources varies in intensity from sunshine to human operators and lighting in the room. Therefore, indirect fluorescent lighting, protective clothing for operators, and

[54] A vital reference to have that discusses many of the issues discussed here in greater detail is Temperature and Humidity Environment for Dimensional Measurement, ANSI Standard B89.6.2-1973.

[55] Evaporative cooling occurs when aqueous fluids are used. Evaporative cooling is usually uneven and represents one of the biggest temperature control problems facing the precision machine design engineer.

machine guards are needed for operating some high-precision machines (e.g., some diamond turning machines). Plastic PVC curtains have been found to be particularly effective at blocking infrared radiation. Often it is difficult to remove the human operator because of the need to rely on operator senses and judgment to control a manufacturing process. As an example, consider a CMM with a 1-m^3 work volume. With fluorescent lights overhead turned on in the morning, the thermal drift over a 2-h period can be on the order of 5 μm. If the CMM is turned on and the overhead lights are left off, the thermal drift may only be about 1 μm.

In machines with high-speed components (e.g., spindles and high reduction transmissions), it is important to carry heat away from the component without decreasing accuracy. For example, cooling the outer spindle structure typically causes the inner bearing race to expand more than the outer race.[56] If the initial (cool) preload on the spindle bearings was too high, the added load from thermal expansion can lead to a catastrophic failure known as "freezing the spindle," which is not recommended. The best way to cool a high-speed spindle supported by rolling element bearings is with an oil mist applied to the inner bearing race. However, this means that the oil is also swished around at high speed, generating viscous friction and heat; thus the oil's temperature and flow rate will also have to be carefully controlled. In order to control the cooling oil's temperature, it must either be circulated through the machine to help bring the machine to uniform temperature quickly or else it must be carried to an external temperature-controlled tank via insulated hoses. Alternatively, the spindle housing can be cooled if the bearings have a low initial preload; however, then the machine cannot be used for heavy-duty work until it has warmed up. Leadscrews are also now available with hollow centers in which temperature-controlled oil can be passed through to control temperature. In many cases, it may be desirable to have a number of independent coolant paths through the machine to avoid formation of gradients within the machine. Note that a spindle can also be mounted kinematically to a housing with sliding bearings for an interface as shown in Figure 2.3.12.[57] The structure and bearing interfaces must be very carefully designed to avoid introducing more error into the spindle than would have existed had the spindle been attached directly to the housing. As shown in Figure 2.3.13, a simpler, yet effective design is to support the front of the spindle with horizontal beams rigidly mounted to the machine and spindle housing, and the rear of the spindle by flexures which allow for axial growth.

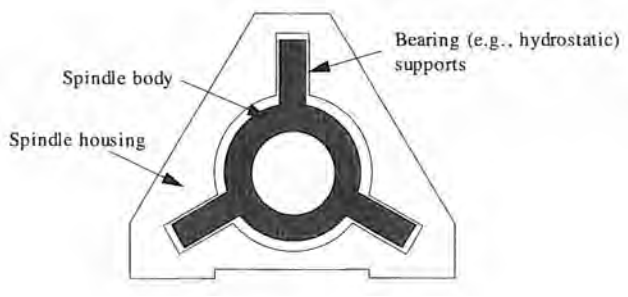


Figure 2.3.12 Kinematically supported spindle. (After Reshetov and Portman.)

The process of metal removal generates heat, much of which is carried away by cutting fluid. The cutting fluid splashes on the machine and warms it. A flood of temperature-controlled cutting fluid is thus needed to reduce heat transferred to the machine by this mechanism. A flood of cutting fluid also helps to carry the heat away from the part and tool. As with the recirculating oil, the cutting fluid should be collected and led away from the machine to be temperature controlled as soon as possible. Note that some large surface grinders have the opposite problem; evaporative cooling requires that the coolant be warmed prior to being sprayed onto the machine. The chips from the cutting process also carry away a considerable amount of heat and should not be allowed to accumulate. A flood of coolant will help to wash them into a conveyor system for removal. The conveyor system should also be thermally isolated from the rest of the machine.

Conduction of heat through the machine from all these sources and cooling methods depends on the geometry, material, and attachment method of various components. The greater the mass and

[56] For analysis methods, see Section 8.9.

[57] From D. Reshetov and V. Portman, Accuracy of Machine Tools, translated from the Russian by J. Ghojel, ASME Press, New York, 1988.

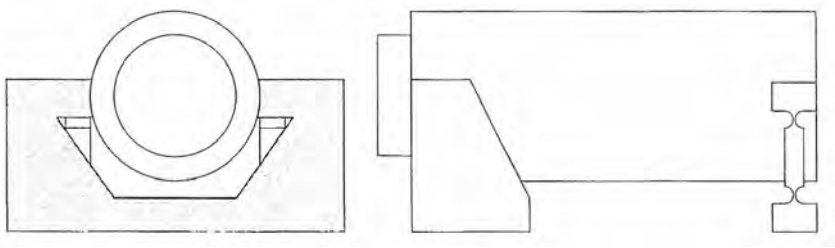


Figure 2.3.13 Quasi-kinematically supported spindle with rigid front mounts, and rear support flexures to allow for thermal growth. (Courtesy of Rank Taylor Hobson Inc.)

path length, the longer it takes for the heat to penetrate through a part. The heat transfer rate also varies widely across bolted joints for the same reason that the stiffness does: bolted joints actually make contact only at the high points of the surfaces. If it is desired to decrease heat transfer across a joint, a rigid insulating material layer (e.g., some ceramics) can be placed between metal components at a joint. Similarly, if it is desired to increase heat transfer at a joint, in addition to specifying finer surface finishes, a heat-conducting epoxy can be used to grout the joint. Thermal deformations of the machine can also be greatly affected by welded joints.[58] Welded joints' metallurgical properties are usually different from those of the parent metal because of the different alloy content created by the welding process. Only a low total input energy, single-pass welding process such as laser or electron beam welding can generate a weld with properties nearly identical to that of the parent metal. In most instances, however, the most important thing to specify is a full penetration weld along the entire joint interface.

Some assembly operations utilize the ability of the machinist to hand-scrape surfaces for final configurations. Examples of this include scraping the rear mount of a spindle so that when the main front bearing heats up more than the rear (the front bearing is usually bigger), the spindle would come into alignment. However, it is difficult and sometimes unreliable to specify this type of fine tolerancing for a new type of machine. Thus it is best to rely on active methods of temperature control or mapping and software-based error correction algorithms to minimize thermal errors.

Many temperature control schemes include air showers to carry away heat from the room. The optimal velocity to accomplish this is on the order of 1 m/s (200 fpm); however, workers feel cold and uncomfortable in rooms with flow greater than 1/8 m/s (25 fpm).[59] It is possible to have personnel leave the room and then turn up the flow, but the effects of noise and air turbulence on the machine itself must then be considered. In addition, any changes in the environment must be followed by a thorough soaking period during which the entire system is allowed to come to thermal equilibrium. In some cases, the machine itself can be temperature controlled with a liquid flowing through internal passages or over its surface. This can result in order-of-magnitude increases in the machine's thermal stability.[60] Always remember the effectiveness of fins, fans, and radiation shields when planning temperature control methods.

Efforts to model thermal errors should result in equivalent thermal expansion elements that can be used in the HTM model to help predict machine errors. By careful design, environmental temperature control, and use of error maps and software error correction techniques, a machine that meets performance goals can usually be designed. In the end, it is the sensor and servo systems that together have the greatest potential to measure temperatures and control the compensation of temperature induced errors.

---

[58] See L. Kops and D. M. Abrams, "Effect of Shear Stiffness of Fixed Joints on Thermal Deformation of Machine Tools," Ann. CIRP, Vol. 33, 1984, pp. 233–238.

[59] See Temperature and Humidity Environment for Dimensional Measurement, ANSI Standard B89-6-2-1973. This document also contains useful discussions of various thermal properties and design issues.

[60] See J. Bryan et al., "An Order of Magnitude Improvement in Thermal Stability with Use of Liquid Shower on a General Purpose Measuring Machine," SME Precis. Eng. Workshop Technical Paper IQ82-936, St. Paul, MN, June 1982; and D. B. DeBra, "Shower and High Pressure Oil Temperature Control," Ann. CIRP, Vol. 35, No. 1, 1986. The design of an effective low-cost temperature-controlled room is described in the report "A Cost Effective Realization of Environmental Temperature Control" by T. McKnight and C. Mossman of Martin Marietta Energy Systems, Oak Ridge National Laboratory, Oak Ridge, TN 37927.

2.3.6 Errors Caused by Material Instability

No matter how well a machine is mechanically designed and manufactured, if the materials used are not stable, with time the machine will lose its accuracy. Growth and warpage in materials is dependent on the internal alloying structure as well as the state of stress in the materials. For example, steels which contain martensite and austenite [face-centered cubic (FCC)] crystals are in a higher-energy state than steels with a ferritic structure [Body-centered cubic (BCC)]. Hence an austenitic structure will always try to revert to the BCC state, where the transition rate is affected by the internal state of stress and temperature in the material. Not all hardened steels behave in this manner; however, to be conservative, it is best to specify case hardening of a surface only for wear resistance, unless high contact stresses are expected. Alloying elements and heat treatment processes also play an important role in a material's stability. The manner in which alloying elements form hard particles to resist crack growth can create a high state of internal stress, which can also warp a structure over a period of time. The manufacturing process used can also impart residual stresses into the material which can lead to long-term stability problems. In addition, for large machines requiring support from foundations, it is imperative that a well-drained (consistent moisture) subgrade and stable concrete slab be used. The machine structure should be supported at numerous points using adjustable height shoes that allow for differential elongation between the machine and the floor. Mold release should be used if the machine is to be grouted in place. This will help the foundation to accommodate differential longitudinal growth between itself and the machine.

One way to avoid instability caused by internal stress relaxation and long-term chemical or metallurgical processes is to use a ceramic material. Many ceramic materials are brittle, so the manner in which they are manufactured does not create plastic strains, which can relax with time. In addition, the inert nature of many ceramic materials helps to ensure dimensional stability. Dimensional stability of 0.01-0.1 ppm/year can be obtained with ceramic materials.

To help ensure stability of the structure once it is put together, tangential stresses at component interfaces should be minimized. As discussed in Section 7.2.5.2, vibrating two components at the structure's natural frequency can often help to seat components properly. Extreme care, however, must be taken not to transmit "tapping" forces through rolling element bearings, or else Brinelling and premature failure may result.

Since it is difficult to determine in what direction material instability errors will occur, it is almost impossible to include their effects in the HTM, other than to assume that they act equally in all directions. As a general precaution, therefore, care should be taken when specifying untested alloys for parts which require stability.

2.3.7 Instrumentation Errors

Sensors are truly at the heart of a machine's accuracy, and all the causes of error in a machine structure can also affect the accuracy and repeatability of sensors. From being mounted in the wrong position, to being heated up, to being deformed by mounting bolts, sensors are even more sensitive to these sources of error than are tons of cast iron. So extreme care must be taken when specifying types of sensors and how they are mounted and used in a machine. Chapters 3–5 discuss types of sensors and mounting methods.

2.4 Errors Caused by Dynamic Forces

Dynamic forces can cause the machine to vibrate in a manner which creates an undesirable surface finish on the part, or which prevents the machine from servoing to the desired position. These effects are caused primarily by structural vibrations and friction, respectively. Increasing structural stiffness decreases the magnitude of structural vibrations but can create a penalty in terms of increased manufacturing cost. Generally, the higher the friction in the bearings, the better the damping and the less chance of generating chatter in the tool. However, it then becomes more difficult to move the tool at high speed and with high resolution. Furthermore, high friction causes heat to be more rapidly generated in the bearings.

Structural Vibration

For large metal-cutting machines, the greatest source of vibration is often from the cutting process itself. Most parts are machined from blanks which have a rough surface to start with. As the tool moves across the surface of the part, the varying depth of cut and associated varying force acts to excite the machine structure and the part itself. This forced excitation of the structure causes deflection of the tool tip, which in turn imparts roughness to the freshly machined surface. After several passes, the amplitude of the imparted roughness may be small, but it is always present. Usually a heavy roughing cut is made, followed by a light finish cut.

Elaborate models of the cutting process exist to help predict this effect for certain specific cutting tools, which aids in their design,[61] but with the ever-changing technology of cutting tools, the machine should be able to adapt to virtually any cutting situation. On the other hand, characterizing the dynamic response of a machine tool during the design stage can help the design engineer to achieve increases in stiffness and savings in weight. There is still a great deal of ambiguity arising from the issue of what are the true stiffness and damping factors associated with many types of joints in a machine (e.g., bolted and welded joints);[62] however, as discussed in Section 7.4, there are ways to help increase the machines damping far beyond what is provided for by the joints. This allows the design engineer to make assumptions about joint stiffness using the data that exist in the literature, as discussed in Section 7.5. In addition, although it has been shown that when the structural loop is closed during the machining process the dynamic response of the machine changes considerably,[63] dynamic modeling of the machine's open structural loop can help to show trends and spotlight obvious deficiencies.

After the design engineer does the best he can to model, design, and check the machine's performance using finite element methods or simple models, a prototype machine will usually be built. The prototype can reveal much information about the design engineer's assumptions about the seemingly innumerable variables affecting the dynamic performance. By testing the prototype using the method of modal analysis, data is obtained about the actual damping and stiffness of the structure. This allows the design engineer to correct his machine models. The performance of the machine can then be tested numerically for a wide range of applications, and modifications can be made accordingly. In practice, this has been a most effective iterative design method for machine tools.[64]

Since a machine will have an infinite number of modes, it is likely that at least one will be excited by some particular cutting frequency. Since no machine can be designed to be totally insensitive to vibration, the best thing to do is adapt and avoid it. Before fast microcomputers were widely available, this was easier said than done. With the present level of microprocessor technology, however, it is possible to monitor vibration levels and change the speed or feed accordingly to move away from one of the modal frequencies of the machine. Accurate placement of accelerometers to sense cutting vibrations requires careful finite element analysis of the structure to allow the sensors to be placed away from nodes and near points of maximum vibrational amplitude.[65]

Vibration in other types of precision machines not used for heavy cutting is often caused by more subtle sources. Precision assembly machines, electron microscopes, diamond-turning machines, and wafer steppers are all examples of machines whose performance relies on isolation from the environment. Any one of the following sources could potentially degrade performance of this type of machine:

- Vibration transmission through the floor.
- Rotating mechanical components (e.g., motors, transmissions).

---

[61] See, for example, J. Tlusty, "Criteria for Static and Dynamic Stiffness of Structures," Technology of Machine Tools, Vol. 3 USDOC NTIS UCRL-52960-3, Oct. 1980, Sect. 8.5.

[62] M. Yoshimura, "Evaluation of Forced and Self Excited Vibration at the Design State of Machine Tool Structures," ASME J. Mech. Trans. Automat. Des., Vol. 108, Sept. 1986, pp. 323–329.

[63] G. Staufert, "Estimation of Dynamic Behavior of Working Machine Tools by Cepstral Averaging," Ann. CIRP, Vol. 28, 1979, pp. 229–234.

[64] See N. Okubo and Y. Yoshida, "Application of Modal Analysis to Machine Tool Structure," Ann. CIRP, Vol. 31, 1982, pp. 243–246; S. Haronath et al., "Dynamic Analysis of Machine Tool Structures with Applied Damping Treatment," Int. J. Mach. Tools Manuf., Vol. 27, No. 1, 1987; I. Inamuru and T. Sata, "Stiffness and Damping Identification of the Elements of a Machine Tool Structure," Ann. CIRP, Vol. 28, 1979, pp. 2235–2240; and M. Yoshimura and K. Okushima, "Computer Aided Design Improvement of Machine Tool Structures Incorporating Joint Dynamic Data," Ann. CIRP, Vol. 28, 1979, pp. 241–246.

[65] S. A. Tobias et al., "On Line Determination of Dynamic Behavior of Machine Tool Structures During Stable Cutting," Ann CIRP, Vol. 32, 1983, pp. 315–318.

Harmonic	Waveform f limit	Amplitude f limit	For a frequency of 70 Hz: Fundamental frequencies to avoid:	
2	0.667	2.000	46.67	140.00
3	0.500	1.000	35.00	70.00
4	0.400	0.667	28.00	46.67
5	0.333	0.500	23.33	35.00
6	0.286	0.400	20.00	28.00
7	0.250	0.333	17.50	23.33
8	0.222	0.286	15.56	20.00
9	0.200	0.250	14.00	17.50
10	0.182	0.222	12.73	15.56

Figure 2.4.1 Limitations imposed on spindle speeds obtained from using Equation 2.4.2 with the assumption that sufficient energy exists at higher order harmonics.

- Rolling bearings (e.g., microstructure making noise as asperities grind together).
- Limit cycling in servo loops.
- Turbulence in fluid supply lines.
- Sound pressure.
- Aerostatic instability in air bearings (pneumatic hammer).

Vibration in a machine can be an insidious source of error because combinations of various sources of energy at various frequencies can lead to resultant frequencies that excite the machine structure. In order to investigate this effect, first consider the ideal case of two waveforms of equal amplitude that interact. As illustrated by Figure 4.5.2 and Equations 4.5.11 - 4.5.14, two waves having nearly identical frequencies f_1 and f_2 and amplitudes Y_1 and Y_2 can interact to produce a waveform with a frequency equal to the average of the two frequencies, and the frequency of the amplitude of the resultant waveform varies as a function of the difference of the two frequencies:

$$f_{waveform} = \frac{f_2 + f_1}{2} \qquad f_{amplitude} = \frac{f_2 - f_1}{2} \tag{2.4.1}$$

In order to avoid vibration problems in the machine, the following relations should be observed:

$$f_{source} \neq \left(\frac{2}{N+1}\right)\omega_n \qquad f_{source} \neq \left(\frac{2}{N-1}\right)\omega_n \tag{2.4.2}$$

Figure 2.4.1 illustrates the restrictions on the spindle speed which are based on the assumption that there is sufficient energy at various higher order harmonics to develop the detrimental waveforms. Of course one can use a spreadsheet to graphically examine the frequencies of waveforms obtained from the combination of different amplitude harmonics. For example, during a consulting job, the author found that the Fourier series sum of a spindle's radial error motion frequency spectrum yielded a waveform that closely matched the surface finish of a part ground using the spindle!

In addition, one must consider that there are numerous other frequencies that are functions of fundamental frequencies. For example, as modeled by Equations 8.3.6-8.3.10 in Section 8.3, the various elements in a rolling element rotary motion bearing generate 5 different error motion frequencies. Often, in a spindle that has many sets of bearings, one is not even sure of the relative phase between the bearings.

For machines that make light cuts, the machine's structural loop is unlikely to change due to contact between the tool and the workpiece; however, the machine's natural frequencies will be a function of the position of the machine elements. Thus one can see that the number of taboo frequencies can be quite large and the best way to avoid dynamic problems is to build in as much damping as possible into the system. Preferably, damping or isolation should be built in at the source of the vibration.

Accordingly, in all cases, 62.5 grams of prevention is worth a kilogram of cure. Prevention includes eliminating sources of vibration from within the machine and isolating the machine from the outside world. Many different commercially available vibration isolation systems exist. One must recognize that any connection to the outside world can act as a conduit to carry vibrational energy into the structure. Connections to the outside world, therefore, should be made through low stiffness interfaces while ensuring that the machine does not shake itself off its mounts.

Frictional Errors

Friction between surfaces helps to damp out vibrations, but it often does more harm than good by causing wear in bearings, limit cycling in servos, and forces that deform components. For example, consider a leadscrew actuator used to apply a force to a linear carriage. Since it is difficult to align exactly the axis of motion of the leadscrew and the axis of motion of the carriage, and error motions between the two also exist, a laterally compliant flexural coupling is often used as an interface between the two.[66] Laterally compliant couplings, however, make the system less stiff axially than if the actuator were to be coupled directly to the carriage.

When an electrical signal from the controller is sent to the motor that controls the position of an axially compliant system, some of the torque generated by the motor is spent accelerating the inertia of the screw, some of the torque goes into overcoming friction in the nut, and some of the torque is converted into linear force, which compresses the coupling. The force spent compressing the coupling is also transferred to the carriage; however, the magnitude of this force is not known. Thus when the controller thinks the carriage is at the proper position and tells the motor to stop turning the leadscrew, the energy stored in the compressed spring of the coupling can continue to push the carriage forward. The controller may try to correct for this action by reversing motor current, but this just causes the problem to occur in the opposite direction. This is known as limit cycling and is one of the most difficult servo errors to correct for. A good servocontroller can often reduce limit cycling to near the resolution of the feedback sensor, and hence this is one of the things to evaluate when shopping for a controller. You can demonstrate this effect to yourself by using a springy strip of metal (an old saw blade) to try and push a book across a table and then to stop on a line. If you do not look to see how much the blade is bending, it is difficult to stop the book accurately. Some mechanisms, such as servovalves, use a constant high-frequency dither to help overcome static friction problems.

As a result of this type of problem, many manufacturers do not use flexural couplings between an actuator and a carriage; instead, they spend their efforts carefully aligning the two. There will always be some misalignment, however, and its effects can be evaluated as illustrated in Section 2.5. Design alternatives range from four-degree-of-freedom couplings, to adding a second feedback element to the carriage to measure the compression of an auxiliary coupling as discussed in Section 10.9. Note that in the last case, merely measuring the spring force with a load cell may not be sufficient, due to the limited resolution of many force-sensing devices compared to the resolution of position-sensing devices. Friction can also act to cause erroneous sensor readings. For example, optical encoders or resolvers are often used to measure rotation of a screw and convert the number of turns times the lead of the screw into linear position. However, friction in the leadscrew support bearings and nut can cause shaft wind-up. The sensor ends up measuring an error equal to the twist in the shaft that is not converted into linear motion by the screw. Proper shaft sizing can reduce this type of error as discussed in Section 5.2.2. In addition, there are various software algorithms to correct for this effect; however, they are all based on an initial calculation of the wind-up torque. As the machine wears, the error in the algorithms increases.

Friction also leads to wear, which changes the dimensions of parts. There are many elaborate general-form equations available for predicting wear rates of most types of surfaces[67]; however, since almost all are empirical in nature, the experience of the design engineer or the bearing manufacturer is most valuable. Many empirical equations also carry a precautionary footnote stating that any number of factors can invalidate the analysis. As long as bearing ways are hardened and the bearing materials are softer, wear will occur in the bearings, which are generally easier to replace than the long bearings ways.

---

[66] See Chapter 10 for a detailed discussion of power transmission elements.

[67] For example, see A. G. Suslov, "Possibility of Ensuring the Wear Resistance of Machine Parts in the Preproduction Planning Engineering Stage," Sov. J. Frict. Wear (translated), Vol. 17, No. 4, 1986, pp. 19–24.

2.5 Design Case Study: Carriage Straightness Errors Caused by Leadscrew Misalignment

A common method for building a precision carriage is to couple a leadscrew to the carriage of a linear bearing as shown in Figure 1.5.6. Most systems constructed in this manner specify precision manufacturing and assembly tolerances to minimize misalignment-induced errors. Some systems even sacrifice axial stiffness of the drivetrain in order to use a flexible coupling to prevent component misalignment from further degrading accuracy. This case study develops a generalized methodology for using information regarding component compliance and assembly tolerances to predict the performance of assembled linear carriages. The analysis is based on the compliances of the components and their relative lateral and angular misalignment before being forced into geometric congruence by assembly. The results can be incorporated into a closed-form analysis, or more appropriately, evaluated numerically.

Forced Geometric Congruence

Manufacturing and assembly errors in linear system bearing and actuator components create the situation depicted in Figure 2.5.1. Note that, in reality, errors in a plane normal to the paper would also exist. For this two-dimensional model, there is a difference in the vertical position and angular alignment (slope) of the components, which is a function of position along their length. When bolted together, the system will reach an equilibrium position that will also be a function of the leadscrew nut's position along the bearing. This forced geometric congruence results in increased straightness and pitch errors of the bearing carriage.

To visualize a model of the system, consider two springs which are of different lengths. When arranged in a linear fashion with opposite ends tied to a wall and the adjacent ends tied together, each spring undergoes a displacement until forces are balanced and an equilibrium position is established. If the springs are anchored to a common plane and a bar attached across their tops, in addition to an equilibrium height position, an equilibrium slope of the bar will also be achieved.

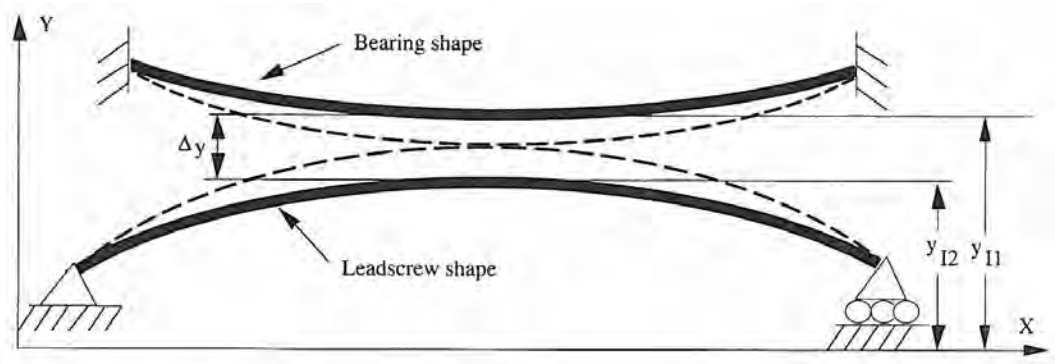


Figure 2.5.1 Example of coupling geometry between a linear bearing and a leadscrew.

A similar situation exists for a linear carriage assembly. The forces caused by imposing geometric congruence of displacements orthogonal to the length axis[68] causes lateral and angular displacement of the components at the point. Similarly, imposing geometric congruence of angular displacements of the components at the same point causes angular and lateral displacements. Regardless of system configuration, lateral displacements caused by forces and moments can be related to vertical displacement compliances $C_{\delta F}$ and $C_{\delta M}$, respectively. Similarly, angular displacements caused by forces and moments can be related to angular displacement compliances $C_{\alpha F}$ and $C_{\alpha M}$, respectively.

Once a linear bearing carriage and leadscrew nut are fastened together, forces and moments between them will be equal and opposite. The displacements of the components from their respective equilibrium positions, however, may not be equal. The vertical and angular coordinates of the linear bearing carriage (subscript 1) and the leadscrew nut (subscript 2) can be written as functions of the

[68] Referred to generically herein as *lateral displacements* .

initial position of the component plus the displacements caused by forces and moments. For the linear bearing

$$y_1 = y_{I1} + F_1 C_{\delta F1} + M_1 C_{\delta M1} \tag{2.5.1}$$

$$\alpha_1 = \alpha_{I1} + F_1 C_{\alpha F1} + M_1 C_{\alpha M1} \tag{2.5.2}$$

For the leadscrew

$$y_2 = y_{I2} + F_2 C_{\delta F2} + M_2 C_{\delta M2} \tag{2.5.3}$$

$$\alpha_2 = \alpha_{I2} + F_2 C_{\alpha F2} + M_2 C_{\alpha M2} \tag{2.5.4}$$

After the bearing carriage and leadscrew nut are forced together and clamped, the following equilibrium conditions must exist:

$$F_1 F_2 = 0 \tag{2.5.5}$$

$$M_1 M_2 = 0 \tag{2.5.6}$$

$$y_1 = y_2 \tag{2.5.7}$$

$$\alpha_1 = \alpha_2 \tag{2.5.8}$$

For purposes of simplifying terms in the analysis, assume the following notation:

$$\Delta_y = y_{I1} - y_{I2} \tag{2.5.9}$$

$$\Delta_\alpha = \alpha_{I1} - \alpha_{I2} \tag{2.5.10}$$

$$C_{\delta F} = C_{\delta F1} + C_{\delta F2} \tag{2.5.11}$$

$$C_{\alpha F} = C_{\alpha F1} + C_{\alpha F2} \tag{2.5.12}$$

$$C_{\delta M} = C_{\delta M1} + C_{\delta M2} \tag{2.5.13}$$

$$C_{\alpha M} = C_{\alpha M1} + C_{\alpha M2} \tag{2.5.14}$$

After substituting these values into the difference of Equations 2.5.1 and 2.5.3, and 2.5.2 and 2.5.4, the following expressions for the equilibrium forces and moments are found:

$$M_1 = \frac{\Delta y C_{\alpha F} - \Delta_\alpha C_{\delta F}}{C_{\alpha M} C_{\delta F} - C_{\delta M} C_{\alpha F}} \tag{2.5.15}$$

$$F_1 = \frac{\Delta_y C_{\alpha M} - \Delta_\alpha C_{\delta M}}{C_{\alpha F} C_{\delta M} - C_{\delta F} C_{\alpha M}} \tag{2.5.16}$$

The equilibrium lateral and angular positions of the carriage are then found, respectively, from Equations 2.5.1 and 2.5.2 with the forces and moments given by Equations 2.5.15 and 2.5.16. When the initial shape of the linear bearing and leadscrew are represented as a polynomial function of position along the bearing, a quantitative assessment of the situation is best determined numerically. For analysis at a point, a spreadsheet program is useful.

Effect of Constant Axial and Torsional Leadscrew Stiffness on System Compliance

If the lateral compliance of the leadscrew and its relative position with respect to the linear bearing have an effect on system accuracy, the question arises: Can the lateral compliance of the leadscrew be increased to increase carriage straightness without increasing axial compliance and thus decreasing the system's dynamic performance? The lateral compliance of the leadscrew can be increased by reducing the diameter or increasing the length. The effect of these two options, however, also increases the axial and torsional compliance of the system, which can decrease the controllability of the system. The question then arises: Can the lateral compliance be increased while maintaining constant axial and torsional compliance?

Consider the case where a leadscrew's length ℓ and radius R are changed from the initial (i) to final (f) state such that the axial compliance remains constant; then the following must be true:

$$\frac{F\ell_i}{\pi R_i^2 E} = \frac{F\ell_f}{\pi R_f^2 E} \tag{2.5.17}$$

For constant axial compliance of a beam whose length is changed, the before and after radii must have the following proportion:

$$R_f = R_i \sqrt{\frac{\ell_f}{\ell_i}} \tag{2.5.18}$$

The lateral compliances for several different mounting cases are shown in Figure 2.5.2, and all vary with ℓ^2, ℓ^3, ℓ, ℓ^2 for the force slope and displacement and moment slope and displacement compliances, respectively. As a result, the initial-to-final compliance ratios for each of these cases when the diameter is increased while maintaining constant axial stiffness are:

$$\frac{C_{\alpha Fi}}{C_{\alpha Ff}} = 1 \tag{2.5.19}$$

$$\frac{C_{\delta Fi}}{C_{\delta Ff}} = \frac{\ell_i}{\ell_f} \tag{2.5.20}$$

$$\frac{C_{\alpha Mi}}{C_{\alpha Mf}} = \frac{\ell_f}{\ell_i} \tag{2.5.21}$$

$$\frac{C_{\delta Mi}}{C_{\delta Mf}} = 1 \tag{2.5.22}$$

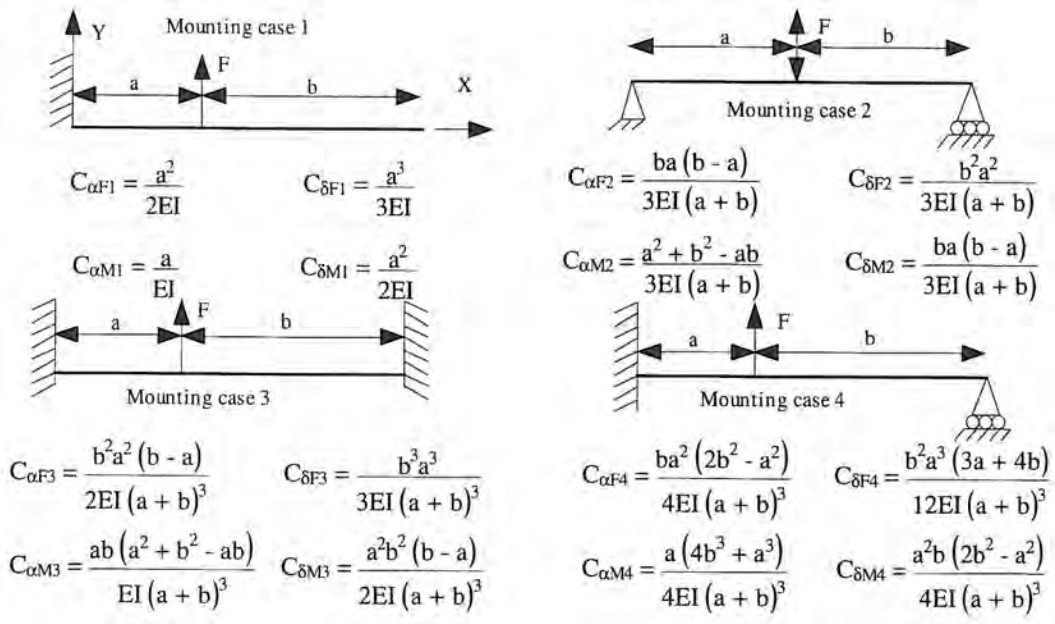


Figure 2.5.2 Compliances associated with different mountings.

If constant axial compliance is imposed as a condition for increasing the length of the leadscrew, the force-angular and force-lateral compliance of the leadscrew remain constant and increase, respectively, which is good. On the other hand, the moment angular and moment lateral compliances decrease and stay the same, respectively. The former is bad. As is shown in the following example, the effect of the moment-angular compliance decreasing (stiffness increasing) is to increase errors in the system as the leadscrew length is increased while maintaining constant axial stiffness.

For constant torsional compliance, the relation between initial and final lengths and radii is found from the relation between angular deflection and the applied torque Γ:

$$\frac{2\Gamma\ell_i}{G\pi R_i^4} = \frac{2\Gamma\ell_f}{G\pi R_f^4} \tag{2.5.23}$$

The before and after radii must therefore be in the following proportion:

$$R_f = R_i(\ell_f/\ell_i)^{1/4} \tag{2.5.24}$$

The effect of this relation on the lateral compliance is less stringent than the one imposed by constant axial compliance. None of the compliances actually decrease with increasing leadscrew length while the diameter is also increased to maintain constant torsional stiffness.

Example: An Air Bearing Carriage Actuated by a Leadscrew

When determining mounting configurations and manufacturing tolerance specifications for linear bearings and leadscrews, four common cases exist as shown in Figures 2.5.2 and 2.5.3.[69] The deflection compliances all assume that the beam length is at least three times the height of the beam, so the error from neglecting shear deformations would be on the order of 10%. The cantilevered beam can be used for short ranges of travel where an effectively kinematic mount is desired for minimum cost. For applications involving heavier loads, a beam rigidly mounted at one end and rigidly mounted or simply supported at the other end is often used. As will be shown by example, in order to minimize coupling errors, a leadscrew should be simply support it at each end.

If errors orthogonal to the axis of motion will not be corrected by servocontrol, then accuracy is of primary concern. To create a workable design, it is desirable to know the relative effects of various parameters on the accuracy of the carriage. Eight factors of primary importance are considered here:

1. Angular alignment between the bearing carriage and the leadscrew nut.
2. Lateral misalignment between the bearing carriage and the leadscrew nut.
3. The weight of the bearing rail.
4. Carriage loads: weight of the carriage, parts, and fixtures and cutting forces.
5. The weight of the leadscrew.
6. The ratio of the length of the bearing to the leadscrew.
7. Maintaining constant axial leadscrew stiffness if its length were to be increased.
8. The manner in which the bearing and leadscrew are mounted.

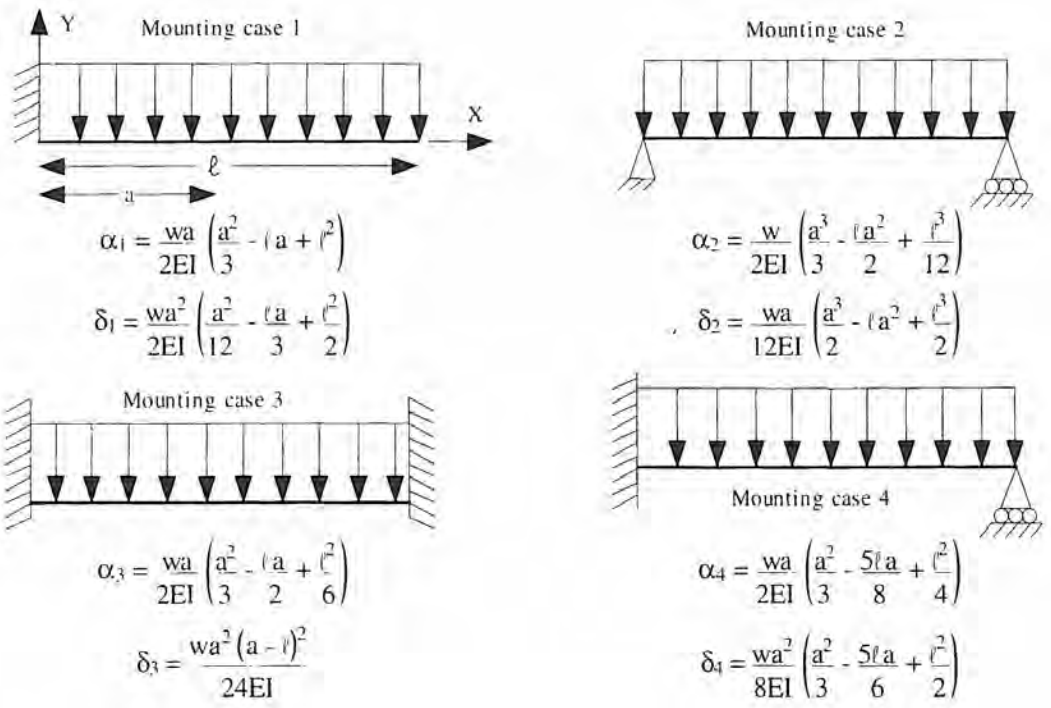


Figure 2.5.3 Deflections caused by a beam's own weight.

These eight factors are a function of the user's system design and must be carefully understood. The straightness of the bearing itself is a first-order effect that is directly controlled by purchase specifications, so it will not be included here, although it can be modeled as a lateral mis-

[69] When the beams are mounted so that they are supported along their length, typically their lateral compliance can be modeled as a constant.

alignment. To demonstrate the effects of these factors on the total error in the system, the following conditions were assumed for a rectangular cross-section air bearing:

- Bearing length 0.508 m, width 0.140 m, height 0.076 m.
- Carriage length 0.216 m, width 0.178 m, height 0.113 m.
- Bearing elastic modulus 69 GPa, density 25.8 kN/m^3.
- Cutting force 45 N.
- Air bearing lateral stiffness 6.30 x 10^7 N/m, pitch stiffness 90 kN-m/rad.
- Initial leadscrew length 0.508 m, diameter 0.041 m.
- Leadscrew elastic modulus 210 GPa, density 78.8 kN/m^3.
- Angular misalignment 200 μrad, lateral misalignment 5 μm.

For the numerical analysis of the performance of various bearing/leadscrew configurations, when the leadscrew was cantilevered the free end was always assumed to be in line with one end of the bearing. For other configurations where the leadscrew was supported at both ends, the bearing was assumed to be centered over the length of the leadscrew. Effects of the various factors on carriage lateral and angular errors are shown in Figures 2.5.4 and 2.5.5 and are summarized for all the various cases in Tables 2.5.1 and 2.5.2.

The complexity of the coupling interaction precludes any catch-all statement about the performance of the system other than errors due to angular misalignment and the weight of the components dominate. Specific observations summarized in the tables include:

1. As the leadscrew length increases and the diameter increases to maintain constant axial stiffness, the compliance $C_{\delta F}$ increases while the compliance $C_{\alpha M}$ decreases. The decreasing $C_{\alpha M}$ is bad because the system is very sensitive to angular misalignment in the first place. This effect is coupled to the lateral deflection as shown by Equations 2.5.15, 2.5.16, and 2.5.1.

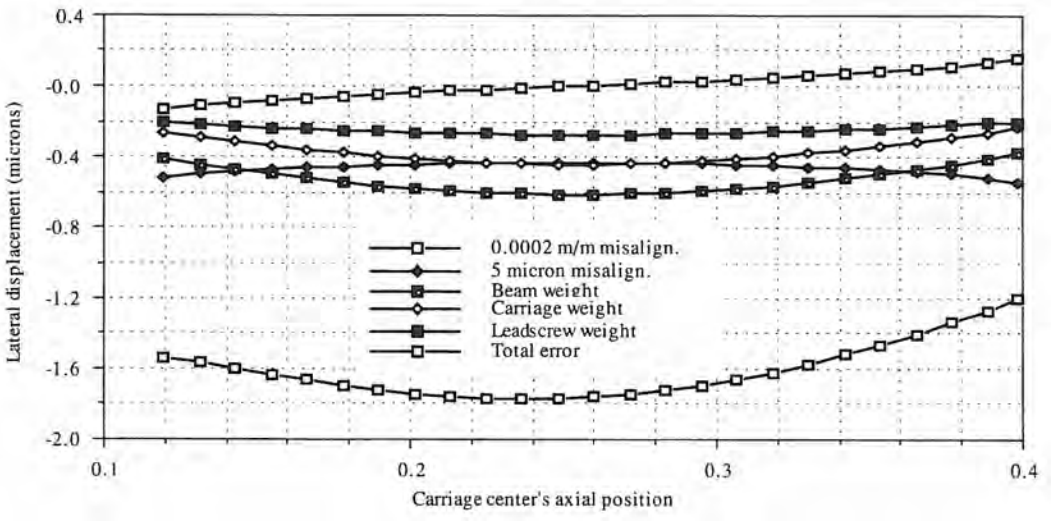


Figure 2.5.4 Carriage lateral errors from a simply supported 0.5-m-long bearing rail clamped to a simply supported 0.5-m-long 4.1-cm-diameter leadscrew's nut.

2. When only an angular misalignment is present in the system, the lateral error increases as the leadscrew length increases with constant axial stiffness. This is due to the fact that it takes a larger moment to achieve angular congruence between the leadscrew and the bearing.

3. When only a lateral misalignment is present, as the leadscrew length is increased along with the diameter increasing to maintain axial stiffness, the lateral compliance increases. This reduces the lateral force needed to achieve lateral alignment. Hence lateral and angular errors decrease accordingly.

4. With respect to the beam weight, as the leadscrew length increases along with its diameter to maintain constant axial stiffness, the leadscrew's decreasing angular compliance tends to decrease

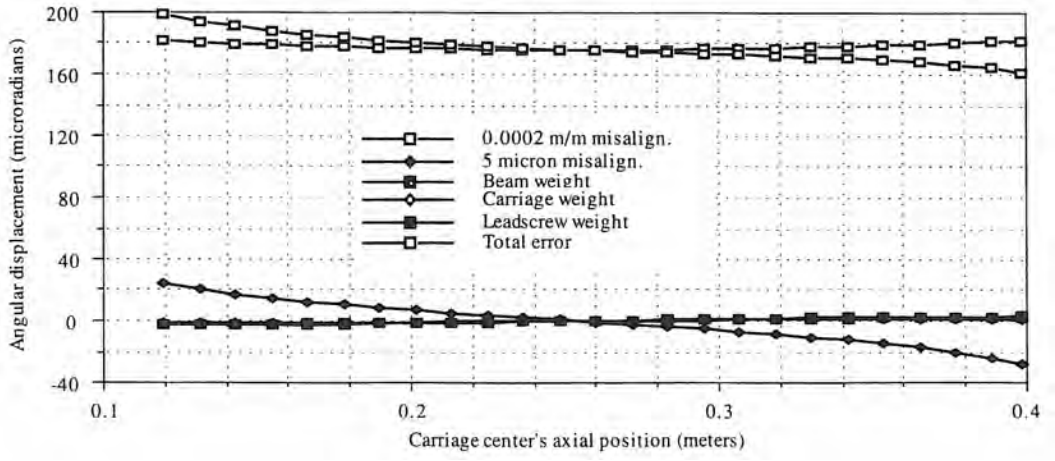


Figure 2.5.5 Carriage angular errors from a simply supported 0.5-m-long bearing rail clamped to a simply supported 0.5-m-long 4.1-cm-diameter leadscrew's nut.

	Leadscrew		Maximum total carriage error	
	Length (m)	Diameter (cm)	Lateral (μm)	Angular (μrad)
Constant0.51	4.06	-1.77	200.5	stiffness:
	0.64	4.57	-1.88	193.4
	0.76	5.08	-2.08	191.0
Constant	0.51	4.06	-1.77	200.5
diameter:	0.64	4.06	-1.70	184.3
	0.76	4.06	-1.72	172.4

Table 2.5.1 Maximum total carriage errors for the system of Figure 2.5.4.

Mounting case number		Maximum total carriage errors	
Bearing	Leadscrew	Lateral (μm)	Angular (μrad)
1	1	9.35	220
1	2	16.67	180
1	3	23.75	189
1	4	19.00	195
2	1	2.95	219
2	2	1.78	200
2	3	3.40	203
2	4	3.35	208
3	1	1.19	230
3	2	0.56	206
3	3	1.52	217
3	4	1.40	220
4	1	1.30	228
4	2	0.76	204
4	3	2.08	214
4	4	1.52	218

Table 2.5.2 Maximum total carriage errors for a variety of mounting cases for the system of Figure 2.5.4.

the angular error in the beam. The leadscrew acts to support the beam, and the same effect is true for the carriage loads.

5. As the leadscrew length increases, so does its weight. The weight must be supported by the bearing since the lateral compliance of the leadscrew increases. The angular error also increases because the bearing has to apply a larger moment to correct the increasing slope in the sagging leadscrew whose angular compliance is decreasing.

6. When the leadscrew length is increased along with the diameter to maintain constant axial stiffness, the lateral error increases slightly while the angular error decreases slightly. When the leadscrew length is increased and the diameter is held constant, both the lateral and angular errors decrease, until the leadscrew stiffness-to-weight ratio decays to the point where the amount the leadscrew must rely on the carriage to support it outweighs the benefits of increased leadscrew lateral compliance.

7. The mounting configuration plays the most important role in determining the effect that different factors have on system performance. As shown in Table 2.5.2, in general the best mounting configuration is to have the bearing fully constrained at both ends, and have the leadscrew simply supported at each end. Note that if the leadscrew is driven by a timing belt, lateral loads on the leadscrew and any bending errors they may introduce must be carefully evaluated.

Chapter 3
Analog Sensors

Weights and measures may be ranked among the necessaries of life, to every individual of human society. They enter into the economical arrangements and daily concerns of every family. They are necessary to every occupation of human industry; to every transaction of trade and commerce; to the labors of the husbandman; to the ingenuity of the artificer; to the studies of the philosopher; to the researches of the antiquarian; to the navigation of the mariner, and the marches of the soldier; to all the exchanges of peace, and all of the operations of war. The knowledge of them, as in established use, is among the first elements of education, and is often learnt by those who learn nothing else, not even to read and write. This knowledge is riveted in memory by the habitual application of it to the employments of men throughout life.

John Quincy Adams

3.1 INTRODUCTION

In order to be able to design an accurate machine, one must have an appreciation for how to measure errors in order to design to prevent them. A good machine design engineer must know what sensors exist, their characteristics, and how and when to use them. *Quite often a sensor limitation will change the overall configuration of a system*. An example of this situation is a linear axis that requires more accuracy than can be obtained by sensing the angular position of a leadscrew. Room in the design must be allotted for use of a scale or laser interferometer. Most precision machines operate under closed-loop control, which requires high-performance sensors. To illustrate the importance of closed-loop control, try marking the position of the hot and cold water knobs when you take a shower. The next time you get into the shower, turn the knobs to the same positions, and jump in the shower without first testing the water temperature (open-loop control). Similarly, to avoid damaging the machine or part, critical processes must be accurately measured and the signal fed back to the controller, which then makes adjustments.

It is beyond the scope of this book to discuss all the detailed methods of measurement that exist, so primarily methods used for position measurement of computer-controlled machine components will be discussed. For a broader discussion of classical and modern dimensional metrology, the reader is referred to other books.[1] On a similar note, although thermal and fluid flow sensors are vital to the operation of many machines, they usually do not require substantial machine modifications in order to incorporate them into the design; hence thermal and fluid flow sensors are not discussed in this book.

Section 3.2 discusses basic definitions and properties regarding sensor performance. Detailed descriptions of nonoptical sensors are provided in the remainder of this chapter. Chapter 4 will discuss optical sensors. Chapters 3-6 are devoted to measurement issues to help the reader develop an understanding for how the sensors work and what conditions in a machine tool can generate errors in the sensors. Methods for electronically interfacing the sensors are not discussed because they do not have an appreciable effect on the way a machine tool is designed.[2]

3.1.1 Sensor Definitions

Continuous control of physical systems generally requires the use of a device to measure the process (i.e., the measurand) to be controlled. A *sensor* is a device that responds to or detects a physical quantity and transmits the resulting signal to a controller. For example, a cam with a cam follower converts rotary mechanical motion to linear mechanical motion and could be called a sensor of rotary

[1] See, for example, T. Busch, Fundamentals of Dimensional Metrology, Delmar Publishers, Albany, NY, 1964; F. Farago, Handbook of Dimensional Measurement, Industrial Press, New York, 1968; G. Thomas, Engineering Metrology, John Wiley & Sons, New York, 1974; and A. T. J. Hayward, Repeatability and Accuracy, Mechanical Engineering Publications, New York, 1977.

[2] For a broad treatment of the subject, see, for example, D. Wobschall, Circuit Design for Electronic Instrumentation, McGraw-Hill Book Co., New York, 1987; and as an excellent general-purpose electronics book, The Art of Electronics, by P. Horowitz and W. Hill, Cambridge University Press, London, 1980.

motion. A *transducer*, on the other hand, is a sensor that converts (transduces) one form of energy to another form, although not necessarily with a direct exchange of energy (e.g., optical encoders). Most often, a transducer will represent the value of the process being sensed in terms of an analog voltage or a digital signal. The following definitions pertain to basic properties of sensors:

Absolute: The output is always relative to a fixed reference, regardless of the initial conditions (e.g., where the sensing element was when the power was first turned on). An example is a common ruler with numbers on the markings representing the distance of each marking from the endpoint.

Analog: The output is continuous and proportional to the physical quantity being measured. An example of an analog sensor is a mercury bulb thermometer.

Digital: The output can only change by an incremental value given a change in the measured physical quantity. An example of a digital sensor is a digital thermometer.

Incremental: The output is a series of binary pulses, where each pulse represents a change in the physical quantity by one resolution unit of the sensor. The measured quantity is only known by counting pulses with respect to an initial state that must be defined when the measurement first begins (e.g., when the power is turned on). An example would be a ruler without numbers on the markings.

In many cases there are numerous types of sensors available for measuring similar physical quantities. When choosing a sensor, obviously the performance, cost, and availability must all be considered. Quite often it is the environmental operating conditions that dictate the type of sensor that should be used. Very rarely will a sensing system be found that is not affected by some environmental factor (e.g., temperature); thus all that one can do to the first order is recognize that errors will occur and try to make the measurement system robust enough to prevent expected changes in environmental parameters from causing significant errors. In critical applications, a test jig should be used to test the sensor's performance under the intended operating conditions. Accordingly, in order to be able to compare sensor technologies, appropriate terms regarding sensor performance characteristics must be defined:

Accuracy. As shown in Figure 2.1.1, all sensors are accurate in that an input causes an output. It is a lack of knowledge of how real-world parameters (e.g., temperature) affect a sensor's output that leads to an error caused by the difference between what the sensor output says and what one thinks it says.

Averaged output. By collecting many data points and using their average value as the output of a sensor, random errors can be reduced by an amount approximately equal to the square root of the number of averages taken. By reducing the noise level in the system, the resolution and sometimes the repeatability can be increased. If the repeatability can be increased, then with the use of a mapped response, the apparent accuracy can also be increased. Averaging will not, however, reduce systematic errors such as truncation of an analog measurement signal by an analog-to-digital converter or hysteresis errors. In order to determine what sampling rate to use, assume that the resolution δ of the sensor is limited by random noise components. If the maximum slew rate is v (e.g., velocity) and it is desired to increase the resolution by a factor of N, then the total required sampling time is

$$t_{\text{total sample}} = \frac{\delta}{Nv} \tag{3.1.1}$$

During this time, the measurand will not have changed by more than 1/N times the resolution of the sensor. In order to increase the resolution by averaging out random noise, about N^2 data points have to be taken in this time; therefore, the minimum required sampling period is

$$t_{\text{sample}} = \frac{\delta}{N^3 v} \tag{3.1.2}$$

Note that the resolution corresponding to the least significant bit of the analog-to-digital converter used to take the samples must be at least equal to the resolution that one hopes to achieve by averaging.

Frequency response is the effect on the output of the sensor of the physical quantity being measured as it varies in time.

Hysteresis is the maximum difference in sensor output between measurements made from 0 to 100% full-scale output (FSO) and 100 to 0% FSO. Although hysteresis is easily measured, its mechanism is not completely understood. Note that it is possible to map hysteresis effects.

Linearity is the variation in the constant of proportionality between the output signal and the measured physical quantity. It is often expressed in terms of a percentage of the full-scale output. Before the days of microprocessors, the accuracy of the system was often defined in terms of the linearity. There are three different ways of fitting a straight line to the sensor's output verses input graph as shown in Figure 3.1.1: *endpoint line*, *best straight line*, and *least squares line*. The *endpoint line* merely connects the endpoints of the sensor's response curve. The *best straight line* is the line midway between the two parallel lines that completely envelop the sensor's response curve. The *least squares line* is the line drawn through the sensor's response curve such that the sum of the squares of the deviations from the straight line is minimized.

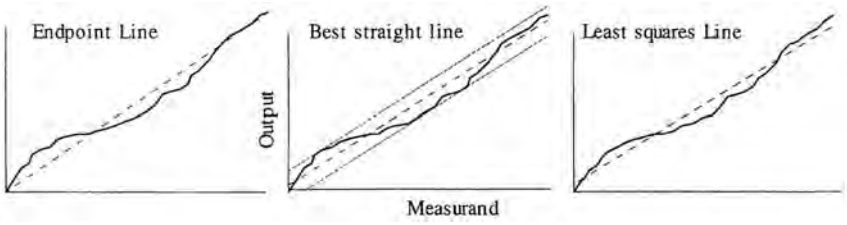


Figure 3.1.1 Different types of linear response lines for a sensor.

Mapping involves measuring the response of a sensor to a known input under known conditions, and then storing the results in a look-up table or fitting a mathematical expression to the data. Effects such as nonlinear response, hysteresis, and temperature effects can be compensated for with the use of mapping. In this manner, the apparent accuracy of the sensor can be greatly increased over accuracy as defined by linearity.[3]

Noise is the magnitude of any part of the sensor's output that is not directly related to the physical quantity being measured.

Noise input margin is the maximum input noise level (e.g., deviation in the supply voltage of a sensor) that can be tolerated before it affects the desired performance of the sensor.

Repeatability is how consistently the sensor gives the same output when subject to the same input.

Resolution (see Figure 2.1.1) of a sensor is the smallest detectable change in the measured physical quantity that can be detected.

Sensitivity is the variation in sensor output caused by a variation of the physical quantity being measured (e.g., volts/mm).

Slew rate error is how the accuracy of the sensor changes with the rate of change of the measured physical quantity. This can be obtained from the frequency response. If the resolution is considered the period of a cycle, then the equivalent excitation frequency for determining the slew rate from a frequency response graph (Bode plot) is the rate of change of the measured physical quantity (slew rate) divided by the resolution.

Standoff distance is how far the sensor is nominally placed from the target. Many noncontact sensors (e.g., capacitance probes) are placed so that their ends are a fixed distance from the nominal position of the surface to which they are sensing the distance. Their sensing range is then a fraction of the standoff distance.

Step response is the time-varying change in sensor output given a step change in the measured physical quantity. Since for position-measuring sensors a step change would require infinite acceleration, the frequency response and slew rate error are better descriptive terms. When describing the step response, the following time increments are typically used, as shown in Figure 3.1.2:

- *Delay time:* The time it takes for the sensor output to rise to 10% of the nominal peak value once the step input is applied.
- *Rise time:* The time it takes for the sensor output to rise from 10% to 90% of the nominal peak value once the step input is applied.
- *Storage time:* The time it takes the sensor output to fall to 90% of the nominal peak value once the step is removed.

[3] See, for example, J. Moskaitis and D. Blomquist, "A Microprocessor Based Technique for Transducer Linearization," Precis. Eng., Vol. 5, No. 1, 1983, pp. 5–8. Note that the mapped response is sometimes referred to as the *linearized response.*

- *Fall time:* The time it takes for the sensor output to fall from 90% to 10% of the nominal peak value once the step is removed.

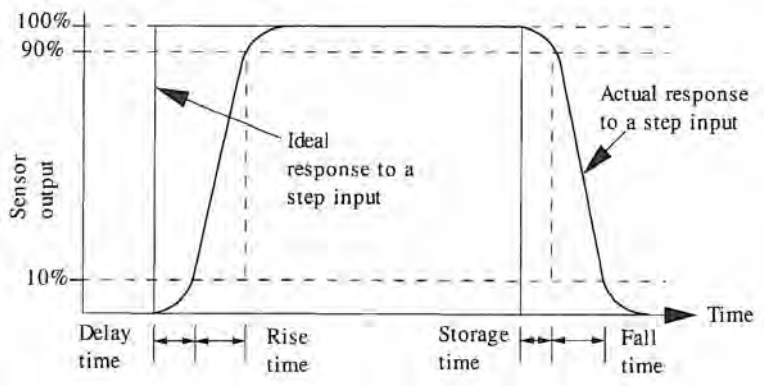


Figure 3.1.2 Characteristics of a sensor's response to a step input.

3.1.2 System Dynamics[4]

The frequency response of a sensor system is the system's ability to respond to changes in the measurand. Generally, the faster the process being measured, the less accurate the measurement. For example, the human eye cannot see a good video terminal flicker because the screen is updated at 60 times per second; on the other hand, the human eye has a greater overall range of perception (e.g., color, shape, texture, size, depth of field,) than perhaps any artificial vision system. The frequency response of a sensor is usually defined in terms of the frequency a sensor's output tends to decrease because it can no longer accurately detect changes in a rapidly changing measurand. This frequency can be determined experimentally or sometimes analytically. In order to demonstrate how quickly system accuracy can be degraded with increasing frequency, a review of basic dynamic system properties will be made.

Most dynamic systems can be modeled mathematically as a continuous time-dependent function $\wp(t)$ of an input to the system. For convenience, time-based mathematical models of dynamic systems are mapped into the Laplacean domain where the linear operator d/dt is essentially represented by the variable s. The transformation is accomplished using the Laplace transform, which is defined as

$$\mathcal{L}\left[\wp(t)\right] = F(s) = \int_0^\infty \wp(t)e^{-st}\,dt \tag{3.1.3}$$

For example, consider the force balance equation for a mass, spring, damper system driven by a forcing function u(t):

$$\frac{md^2x}{dt^2} + \frac{bdx}{dt} + kx = u(t) \tag{3.1.4}$$

When substituted into Equation 3.1.3 it yields

$$ms^2x + bsx + kx = U(s) \tag{3.1.5}$$

where U(s) = 1 for a unit impulse and U(s) = 1/s for a unit step. The response of the system can thus be represented by

$$G(s) = x = \left\{\frac{1}{ms^2 + bs + k}\right\} U(s) \tag{3.1.6}$$

The term on the right in braces is called the *system transfer function* G(s) and is formulated in terms of a numerator, N(s), and denominator, D(s). In the same manner that a machine tool is built up from

[4] For a more detailed discussion of the properties of dynamic systems, consult a text such as Modern Control Engineering by K. Ogata, Prentice Hall, Englewood Cliffs, NJ, 1970; or Introduction to System Dynamics by J. Shearer et al., Addison-Wesley Publishing Co., Reading, MA, 1967.

bearings and actuators and other modular components, a complete system's transfer function can be factored into a series of first- and second-order equations in the numerator and denominator, where the response of each individual factored component is easily computed:

$$G(s) = \frac{\prod_{i=1}^{L} N_i(s)}{\prod_{j=1}^{K} D_j(s)} \tag{3.1.7}$$

If the logarithm of the transfer function is taken, it can be represented as

$$Log_{10}G = \sum_{i=1}^{L} \log_{10}N_1 - \sum_{j=1}^{K} \log_{10}D_j \tag{3.1.8}$$

This allows the response of a complex system to be obtained easily by graphical addition of its components' responses. Because most numerical analysis of dynamic systems is now done on a computer, the merits of this method may not be appreciated, but before personal computers became commonplace, most engineers did the calculations manually. An added benefit from this method results from the fact that dynamic systems often respond over a large frequency range, and log-log graphs allow for better resolution readings from the graph over a wide range of frequencies.

The magnitude of the response is usually represented as 20 times the logarithm of the magnitude, which defines the *decibel* (dB):

$$\text{response(dB)} = 20\log_{10}(G) \tag{3.1.9}$$

The logic behind using the factor of 20 resides in the fact that the magnitude of the response is equal to the square root of the sum of the squares of the real and imaginary (sine and cosine) parts of the system's dynamic response. These parts are obtained by substituting $s = j\omega$ into the transfer function, where ω is frequency and $j^2 = -1$. Thus Equation 3.1.9 yields the magnitude of the response as a function of the frequency. The logarithm of the square root of a function is just one-half of the logarithm of the function; thus the factor of 20 reduces to 10. The factor of 10 is needed so that people do not have to talk in terms of a fraction of a decibel.[5]

A nice result of this analysis method is that as the frequency ω approaches infinity, all first-order (d/dt) and second-order (d^2/dt^2) terms of a dynamic system's response decrease, respectively, according to

$$\begin{aligned} dB_{\text{1st order}} &= -20\log_{10}\left(1+\omega^2\tau^2\right)^{0.5} \\ &= 0 \quad \text{for} \quad \omega << \frac{1}{\tau}; \\ &= -20\log_{10}\omega\tau \quad \text{for} \quad \omega >> \frac{1}{\tau} \end{aligned} \tag{3.1.10a}$$

$$\begin{aligned} dB_{\text{2nd order}} &= -20\log_{10}\left[\left(1-\frac{\omega^2}{\omega_n^2}\right)^2 + \left(2\xi - \frac{\omega}{\omega_n}\right)^2\right]^{0.5} \\ &= 0 \quad \text{for} \quad \omega << \omega_n; \\ &= -40\log_{10}\frac{\omega}{\omega_n} \quad \text{for} \quad >> \omega_n \end{aligned} \tag{3.1.10b}$$

For every doubling of the frequency, called an *octave*, the first- and second- order responses change by -6.02 dB and -12.04 dB respectively. For every order of magnitude increase in frequency, the first- and second- order responses change by -20 dB and -40 dB respectively. The -3 dB response point of a first order system is where the excitation frequency equals the natural frequency of the system and the response is 0.707 of the response at zero frequency (dc) input. It is also the point where the response lags the input by 180°. These concepts are illustrated in Figure 3.1.3. *Most sensors' frequency responses are given in terms of the -3 dB point.*

[5] Although this explanation may seem frivolous to an experienced system dynamicist, most machine design students and practicing designers who do not use these relations every day do not have an intuitive feel for the logic behind a decibel and thus soon forget what it means.

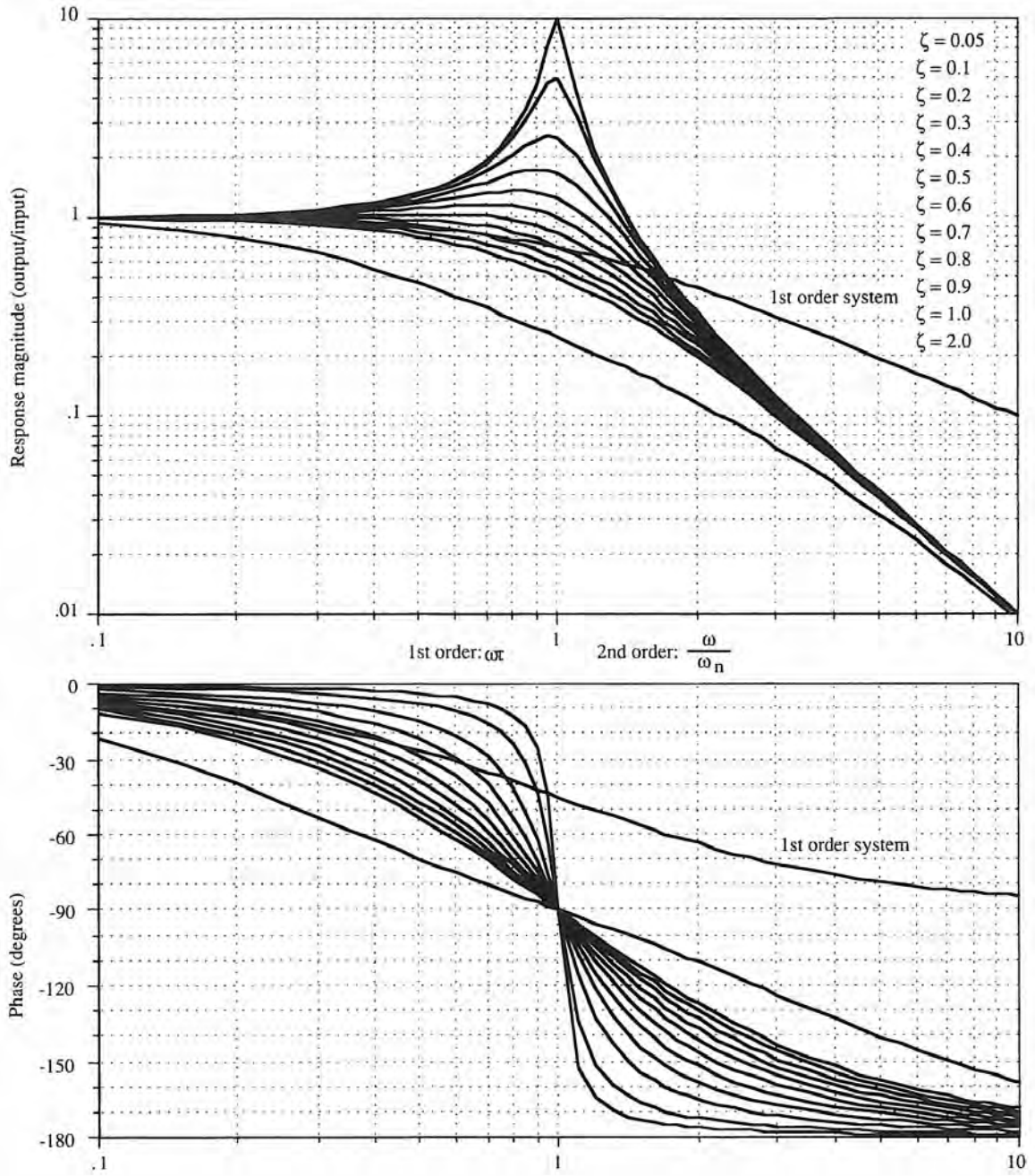


Figure 3.1.3 Magnitude and phase of simple first-order and second-order systems computed using a spreadsheet.

If a sensor is used to detect motion of a part and the output from the sensor used to control an axis to correct for the error, the sensor would probably be operated well before its -3 dB frequency response point. The justification for this can be seen in Table 3.1.1. For high accuracy applications, most sensors must be operated in the quasi-steady state unless the sensor's frequency response is mapped. Note that the phase angle portion of the dynamic response also affects whether a sensor can be effectively used in a control system for a machine. If the response of the sensor lags behind the actual physical process too much, then it may not be possible for the mechanism to correct for errors sensed because the error may have already irreversibly affected the process.

Decibels (dB)	Error
-0.0000087	1 ppm
-0.000087	10 ppm
-0.000869	100 ppm
-0.008690	1000 ppm
-0.087	1 %
-0.915	10 %
-3.0	30 %

Table 3.1.1 Measurement error corresponding to gains on a sensor's Bode plot.

3.2 NONOPTICAL SENSOR SYSTEMS[6]

Nonoptical sensors are defined here as sensors that generate analog signals or digital pulses in response to a physical process by other than optical means. Nonoptical sensors discussed here include:

- Capacitance sensors
- Hall effect sensors
- Inclinometers
- Inductive digital on/off proximity sensors
- Inductive distance measuring sensors
- Inductosyns®
- Linear and rotary variable differential transformers
- Magnetic scales
- Magnetostrictive sensors
- Mechanical switches
- Piezoelectric material-based sensors
- Potentiometers
- Syncros and resolvers
- Velocity sensors

Each of these types of sensors is discussed in detail in following sections. In most cases, representative examples of size, cost, accuracy and so forth are given to provide the reader with a "feel" for what is available.

3.2.1 Capacitance Sensors[7]

A capacitive sensor, as shown in Figure 3.2.1, is a device that determines the distance (gap) between a probe and a target surface. The sensor measures the capacitance that is formed by two parallel plates, one being the face of the probe and the other being the target. The measurement is termed noncontact because the probe does not physically touch the target. Thus capacitance sensors often find use in applications where motion of the target cannot be measured by a contacting probe due to the speed of the target or where the target cannot be touched without deforming the target surface.

Most performance parameters of a capacitance sensor, such as accuracy and linearity, are determined primarily by the value of the probe capacitance, which can range from as high as 10-100 pF for large probes to as low as 0.01-0.1 pF for small probes. For the smaller values of probe capacitance, careful circuit design is necessary in order to achieve acceptable system performance.

Capacitance sensors are able to sense a wide variety of materials, including metals, dielectrics, and semiconductors. The sensor output will be affected by the type of material. All

[6] A detailed theoretical discussion of the operating principles of many of these sensors is provided in the book Control Sensors and Actuators by C. Silva, Prentice Hall, Englewood Cliffs, NJ, 1989.

[7] This section was written principally by Wayne Haase, formally of Pioneer Technology Corporation (a manufacturer of precision capacitance sensors) in Palo Alto, CA.

conductive materials affect a capacitance sensor's output equally; consequently, a capacitance sensor calibrated over a stainless steel target will also measure correctly over brass and aluminum. The sensor is therefore not affected by changes in alloy constituencies or grain size, a problem for eddy current (impedance probes) gages. Changes in calibration, however, occur for different dielectric materials such as Zerodur® or quartz.

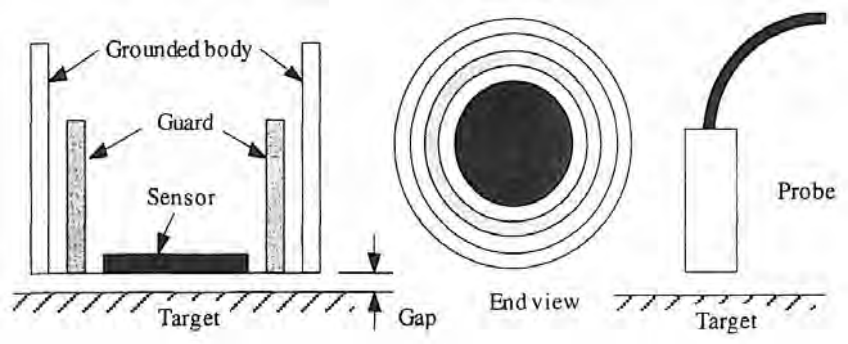


Figure 3.2.1 Noncontact capacitive distance sensor.

Uses of Capacitance Sensors

Capacitance sensors are often used to measure motion of rotating parts such as spindles and bearings. They can also be used to map the flatness of delicate objects, such as lenses and silicon wafers. For example, they are used to measure the surface characteristics of hard disk blanks during manufacture. Parameters such as displacement, velocity, acceleration, thickness, and repetitive and nonrepetitive runout of the disk as it turns are obtained from the analog output of the sensor. Because capacitance sensors are capable of providing greater resolution than virtually any other type of analog sensor, they are often used as feedback devices for short-range precision micropositioners.

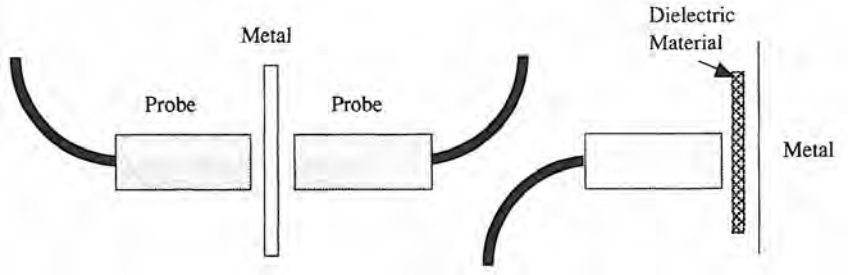


Figure 3.2.2 Thickness measurement using capacitive sensors.

Capacitance sensors can be used to measure thickness of a part, as shown in Figure 3.2.2. For metals, two probes are required, one on each side of the material. The thickness of the material is determined by subtracting the sum of the two probe-to-target distances from the probe-to-probe distance. For measuring the thickness of dielectric materials, one probe is used in conjunction with a grounded metal surface. The dielectric material is introduced into the gap between the probe and the surface and the resultant change in probe-to-ground capacitance is a measure of the thickness of the material.

Capacitance sensors are also used extensively to measure pressure. This type of gage incorporates a rigid frame with a thin flexible diaphragm. A pressure differential across the diaphragm created by fluid or acoustic wave pressure causes the diaphragm to deflect, creating a change in capacitance between the diaphragm and the sensor. Many microphones operate on this principle. Capacitance sensors can also be used in a presence/no-presence mode to detect proximity of a nonmetallic object or liquid level. However, for sensing proximity of metallic objects, less expensive inductive proximity sensors are most often used.

Probe Design for Capacitance Sensors

The manner in which a capacitance sensor is designed depends on the particular measurement application and on the electronic circuit configuration that will be used to amplify sensor output. Improved performance, particularly in accuracy and linearity, can be achieved by matching the shape of the probe's face to that of the target. It will be assumed throughout this discussion that the surfaces to be measured are flat, so the face of the probe should also be flat.

As discussed above, many capacitive gages use probes that have a very small value of capacitance, often in the range 0.01-1.0 pF. In this situation, stray capacitance, which might occur between the probe sensor and the outer body of the probe, must be eliminated, or at least reduced to 10^{-14}-10^{-16} F. This can be done by employing a guard electrode around the sensing electrode, as shown in Figure 3.2.1. The equivalent circuit is shown in Figure 3.2.3. The guard electrode and the sensing electrode are driven by identical voltage waveforms, such as a constant-frequency sine wave. The guard electrode serves two purposes: First, its charge is maintained by the source current, so any nearby stray capacitance affects the guard with only second-order effects on the sensor. The second purpose of the guard electrode is to collimate the electric field lines between the sensor and the target, as illustrated in Figure 3.2.4. As a result the capacitor C_{st} acts more like a classical parallel-plate capacitor:

$$C = \frac{\varepsilon A}{d} \tag{3.2.1}$$

where C is the capacitance in farads, ε is the dielectric constant of the material in the gap, A is the area of the gage plate, and d is the distance between the plates. The extent of how environmental changes affect these parameters in a particular application must be assessed and if necessary a reference probe used to compensate.

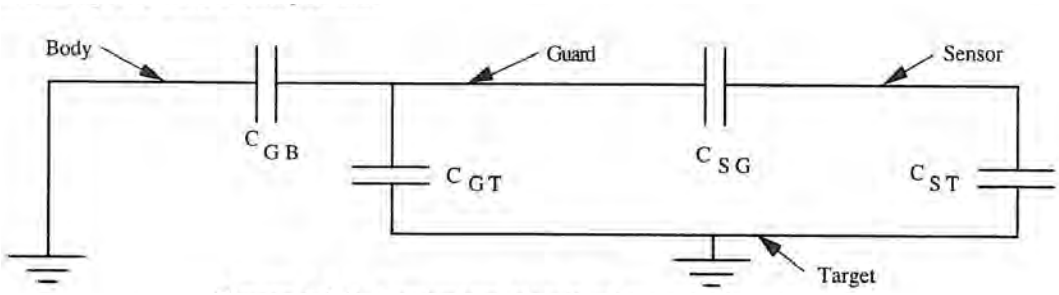


Figure 3.2.3 Equivalent circuit of a capacitive sensor.

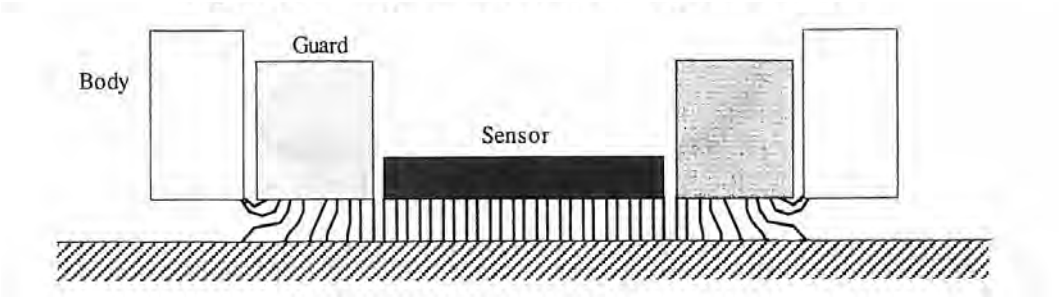


Figure 3.2.4 Electric field lines for a capacitive sensor.

The dielectric constant ε is a measure of how easily electromagnetic waves can travel through a medium. The dielectric constant of any media other than a vacuum[8] will be affected by changes in temperature, pressure, humidity, and media type:

$$\varepsilon = \{(\text{Temperature, Pressure, Humidity, Media Type}) \tag{3.2.2}$$

[8] The ideal medium is a vacuum whose dielectric constant is used as a standard reference and is given a value of unity.

Humidity and temperature can be controlled, but barometric pressure cannot be regulated easily; thus in ultraprecision applications (less than 0.l μm) a second reference gage which monitors a fixed gap is often used to compensate for changes in barometric pressure. Pressure and humidity usually do not affect the dimensional size of a precision machine, but temperature will cause the machine to change shape. Hence it is desirable to make the effect of temperature change on the sensor much less than the temperature's effect on the machine. Motion measured by the probe will then be actual motion and not temperature-induced drift of the probe output. Changes in content of the atmosphere caused by cutting fluid or oil showers for temperature control must also be prevented, often by the use of positive pressure guards.

The area of the capacitance sensor tip also greatly affects the accuracy of the system. The greater the ratio of the area of the sensor to the distance from the target, the greater the accuracy and resolution of the probe. This is due to the fact that the relative influence of electromagnetic waves, which are warped by the edge geometry of the probe, on the capacitance decreases with the increase in the area-to-standoff distance ratio. Also, the ratio of the probe area to the characteristic surface finish dimension of the part should be as great as possible to provide an averaging effect. If the probe is too small, the effect of changing surface topography may be registered by the probe as an apparent motion of the target.

Typical Characteristics of Capacitance Sensors

Most applications involve either high-speed motion or sensitive target materials which dictate noncontact measurements. The following summary of capacitance sensor characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as gospel.

Size: Round sensors from 5 mm diameter × 50 mm long to 10 mm diameter × 50 mm long. Flat sensors from 115 mm long × 15 mm wide × 4 mm thick to 200 mm long × 20 mm wide × 6 mm thick.

Cost: $900 to $1500 for the probes, $2300 for a printed circuit card that conditions the signal and provides output suitable for an ADC, $4600 for a self-contained black box with analog output meter and differential operating mode.

Measuring Range: On the order of ±0.13 mm (0.005 in.).

Accuracy (Linearity): On the order of 0.10-0.20% of full-scale range.

Repeatability: Highly dependent on environmental conditions, but can be on the order of two to five times the resolution.

Resolution: On the order of 1 part in 10^5 of full-scale range. Typically can be as small as 25 Å (0.1 μin), and with special probes and electronics the resolution can be on the order of angstroms. The resolution depends on the quality of the power supply, the full-scale voltage range output from the probe, and the resolution of the analog-to-digital conversion device. For off-the-shelf probes, noise levels are typically 0.02 μin. (5 Å) at 1 Hz, 0.04 μin. (10 Å) at 1 kHz, and 0.05 μin. (13 Å) at 40 kHz.

Environmental Effects on Accuracy: 0.04%/C^o of full-scale range (400 μm/m/C^o). Thus a sensor with a sensing range of 10 μm will have an error of 40 Å/C^o. Pressure and humidity also affect performance, but to a much lesser degree than temperature. Impurities in the air such as oil may also act to change the dielectric constant and thus must be guarded against. The sensor can be operated submerged in nonair environments (e.g., oil) as long as it is calibrated in the same medium.

Life: The sensor is noncontact, so life can be infinite.

Frequency Response (-3dB): Up to 20-40 kHz.

Starting Force: Noncontact and no electromagnetic force coupling, so there is no force between the sensor and the sensing surface.

Allowable Operating Environment: Accumulation of debris will change the dielectric constant in the gap and measurement error will result. The sensors can be operated in fluid environments as long as the black box electronics are tuned accordingly. Operating temperatures range from 0^{o}C to 50^{o}C, 0-90% noncondensing humidity. Precision systems should be operated at 20^{o}C (68^{o}F).

Shock Resistance: On the order of 10g.

Misalignment Tolerance: Misalignment causes errors that are proportional to the cosine of the angle of misalignment.

Support Electronics: Requires a special black box which usually represents 50-75% of the cost of the system. Most black boxes only require 115 V ac at 0.30 A. Ultraprecision systems require stable dc input.

3.2.2 Hall Effect Sensors[9]

If a charged particle, such as an electron, moves in a magnetic field, a force is exerted on the particle, which results in the particle being deflected from its original path. This is known as Lorentz's law. A Lorentz force acts at right angles to the direction of particle motion and the magnetic field. The Hall effect is a result of a Lorentz force acting on electrons flowing through a semiconductor. It is exhibited when certain materials conducting a current are placed in a magnetic field whose orientation is orthogonal to the direction of current flow. The result is a potential produced in a direction orthogonal to the excitation current and the magnetic field, as shown in Figure 3.2.5.

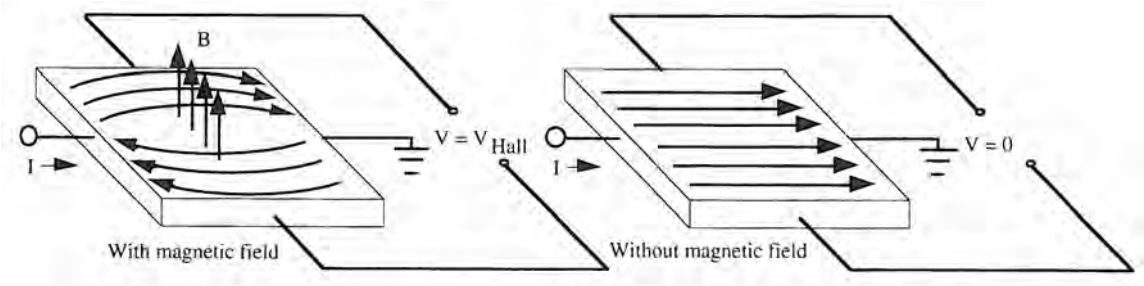


Figure 3.2.5 Operating principle of a Hall effect sensor. (Courtesy of Micro Switch, a Honeywell Division.)

The output voltage from a Hall effect device is only on the order of millivolts and thus requires signal conditioning electronics to boost its output. The combination of a Hall element and signal conditioning electronics onto a single element constitutes a modern solid-state Hall effect transducer. Output from this type of transducer is typically on the order of 0.5 V dc and is proportional to the product of the reference current, which is typically 4-20 mA at 4-24 V dc, the strength of the magnetic field, and the cosine of the phase angle between the direction of reference current flow and the magnetic flux lines.

Hall effect sensors can also be designed to be sensitive to the type of magnetic poles used to trigger them. For example, a bipolar digital on/off Hall effect sensor has a plus (south pole) maximum trigger point and a minus (north pole) minimum release point. Bipolar Hall effect sensors are thus often used with pairs of bar magnets or with magnets which are magnetized on the outside diameter with alternating north and south poles. The south pole "turns the sensor on" and the north pole "turns the sensor off." Unipolar Hall effect sensors are triggered by the south pole of a magnet and are released when the magnet is removed. Figures 3.2.6a-g show some of the ways that Hall effect sensors can be used to sense position.

Once in the presence of a magnet, a Hall effect sensor can output an analog voltage proportional to the magnetic field strength, or it can be designed integral with a transistor that makes it operate in a current sinking or current sourcing mode. In the analog mode, a Hall effect sensor acts as a distance measuring sensor where voltage is proportional to magnetic field strength, which varies with distance. The primary disadvantage of using an analog Hall effect sensor as opposed to another noncontact distance sensor, such as a capacitance or impedance sensor, is that it usually requires a magnet to be mounted to the object being sensed, and its accuracy depends on the stability of the

[9] The Hall effect was discovered by Dr. Edwin Hall in 1879 while he was a doctoral candidate at Johns Hopkins University. A useful reference for users of Hall effect transducers is Hall Effect Transducers, available from Micro Switch, a Honeywell Division, Freeport, IL 61032.

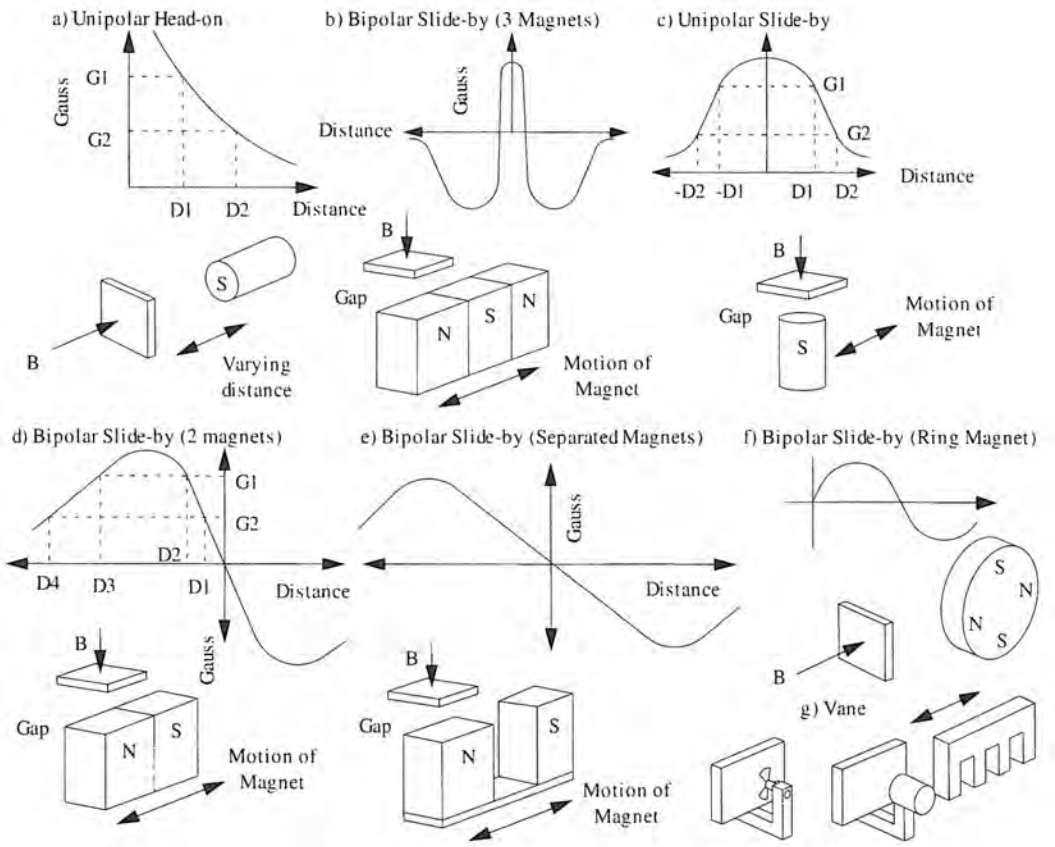


Figure 3.2.6 Hall effect sensor mounting methods. (Courtesy of Micro Switch, a Honeywell Division.)

magnetic field. Many magnets lose their strength with time (demagnetization) and therefore Hall effect sensors are generally not suitable for ultraprecision applications where resolutions greater than about 5 μm are required. However, for less stringent applications they can be almost two orders of magnitude less expensive than other electronic distance sensing means. An inexpensive Hall effect sensor does require a stable power supply, typically 5 +- 0.001 V dc to ensure maximum accuracy. In fact, the accuracy will be proportional to the accuracy of the power supply used. Calibrated Hall effect sensors are available to check magnetic fields to ensure that they are sufficient to operate "off the shelf" sensors effectively. Where extreme accuracy is required, these calibrated Hall elements can themselves be used as a sensor, although the price is an order of magnitude greater (e.g., $60 vs. $6).

In the current sinking mode, the switch is normally closed and an output voltage is registered across the sensor. When the sensor is triggered by a magnet, the switch is opened and current no longer flows. In the current sourcing mode, the opposite case is true and the switch is normally open. The type of sensor used depends on what the output is used to control. To avoid bounce in a Hall effect sensor, a certain amount of electrical hysteresis is built in. Thus the operate point where the Hall effect is triggered by a certain magnetic field strength is different from the release point, where the Hall effect is "turned off" by lack of sufficient magnetic field strength.

Magnets for Hall Effect Sensors

The magnet is every bit as important to the operation of a Hall effect sensor as is the Hall effect sensor itself. As shown in Figure 3.2.7, lines of magnetic flux move from the north pole to the south pole of a magnet. The strength of the field is described in terms of the *flux density* and is measured in terms of *gauss*.[10] A typical refrigerator magnet may have a field strength on the order of

[10] Metric units for magnetic field strength are in gauss. One tesla = 10^4 gauss.

100 G near its surface. Although the design of magnetic systems is beyond the scope of this book,[11] the main point to keep in mind is that a magnet's field can be permanently changed by interaction with other magnets or strong electric fields. For precision sensing applications, therefore, external magnetic fields should be minimized and periodic recalibration of the sensing system scheduled.

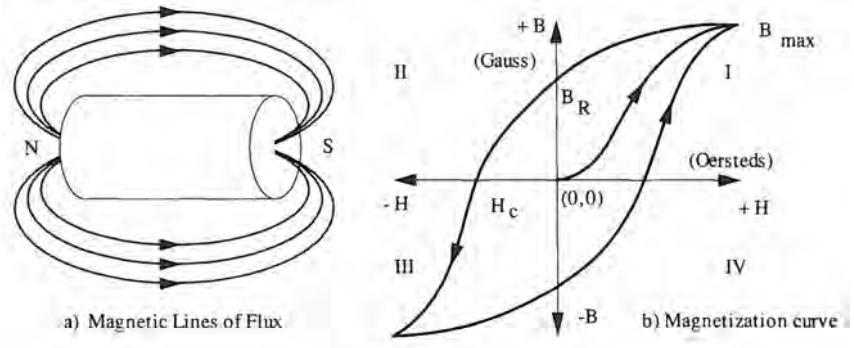


Figure 3.2.7 Properties of a magnet.

With the use of a pole piece, a magnetic field can be provided with a lower-resistance path for the lines of flux. As shown in Figure 3.2.8, the effect of the pole piece on the performance of a Hall effect sensor is to increase the distance at which the sensor is triggered, or to allow the use of a less sensitive sensor for the same trigger distance, or to allow the use of a less expensive magnet. Often a magnet can be placed inside an iron cavity that acts to focus the magnetic flux in a manner similar to that caused by a pole piece, thereby increasing the flux gradient (gauss per units of length). This increases the resolution of Hall effect sensors. Various magnets are available for use with Hall effect sensors, as shown for example in Table 3.2.1. Figure 3.2.9 shows a typical magnet's characteristics. Cylindrical bar magnets for use with Hall effect sensors are typically 5-10 mm in diameter and 10-30 mm long. Ring magnets may have 20 poles and be 45 mm in diameter. Of course miniature magnets can be used for short range applications.

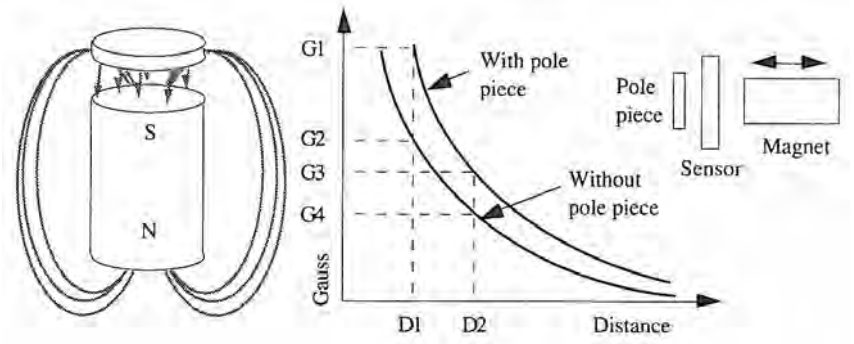


Figure 3.2.8 Effect of a pole piece on a magnetic field and a Hall effect sensor's output. (Courtesy of Microswitch, a Honeywell Division.)

Typical Applications

Figure 3.2.6 showed several methods of mounting Hall effect sensors and magnets. The *unipolar head-on mode* shown in Figure 3.2.6a uses a single magnet to trigger a Hall effect sensor that approaches the south pole of the magnet along a line that is parallel to the north/south pole axis of the magnet. A problem with this type of setup is that there is not much room for mechanical overtravel.

The *unipolar slide-by mode* shown in Figure 3.2.6c uses a single magnet whose axis is perpendicular to the plane of motion of the Hall effect sensor. This setup can be designed to have equally

[11] See, for example, Handbook of Magnetic Phenomena by H. E. Burke, Van Nostrand Reinhold, New York, 1986.

Material	Mechanical Process	Temp. strength	Mag. shock range (C)	Demag. resistance	Gap distance (mm) from sensor: 0.25 resistance	0.76	1.27	2.54	3.81	5.06 (Gauss level at 25 C)	Catalog No.
Alnico V Cast	Good	-40 to 300	Poor	Fair	1460	1320	1170	810	575	420	101MG3
											101MG4
Alnico VIII Sintered	Good	-40 to 250	Good	Excellent	1050	900	755	470	295	195	101MG7
											102MG11
											102MG15
					730	550	410	205	115	75	101MG2L1
Indox 1 Pressed	Good	0 to 100	Good	Excellent	700	520	375	175	85	45	105MG5R2
											105MG5R4
Rare Earth Pressed	Poor	-40 to 250	Good	Excellent	1110	630	365	120	55	25	103MG5
					2620	2100	1600	940	550	350	106MG10

Table 3.2.1 Properties of some magnets available for use with Hall effect sensors. (Courtesy of Micro Switch, a Honeywell Division.)

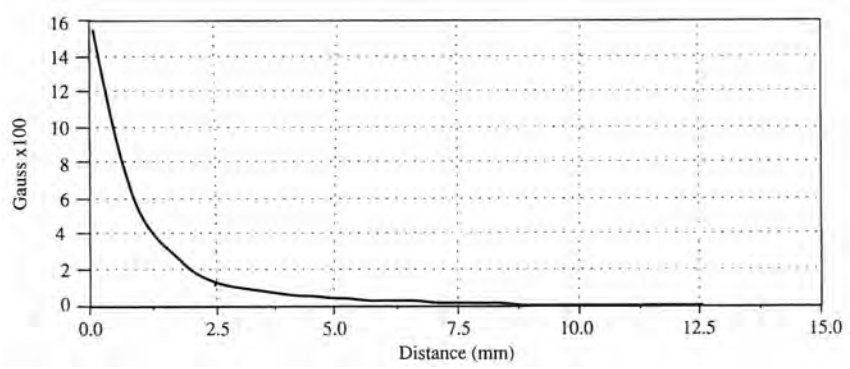


Figure 3.2.9 Typical properties of a magnet used with Hall effect sensors. Micro Switch magnet: 103MG5 measured head-on with a 732SS21-1 calibrated Hall element. (Courtesy of Micro Switch, a Honeywell Division.)

sensitive triggering capabilities as the unipolar head on, but it can allow for overtravel. Since a machine cannot stop instantaneously, this type of setup is often used on machines to sense the home position of the machine. To maximize repeatability, the gap between the sensor and the magnet should be made as small as possible. This will increase the effective flux density the sensor sees and, as a result, decrease the distance the sensor travels from the point of not being triggered to the point of being triggered.

In order to obtain directionality, the *bipolar slide-by mode* is used as shown in Figure 3.2.6d and e. The steepness of the Gauss vs. distance curve is a function of the spacing of the magnets. For maximum resolution, the magnets should be placed close to each other but without causing demagnetization. If an analog sensor is used, then the voltage will be a measure of position. If a digital on/off proximity sensor is used, then it will trigger at one point on the curve and release at another point on the curve. Note that the design is unidirectional, as a bipolar sensor requires a high positive magnetic field to trigger, but it will release only if exposed to a low negative magnetic field.

Figure 3.2.6b shows another variation of the *bipolar slide-by mode* with three magnets. This allows the sensor to trigger and release from either direction. The steepness of the curve indicates that the sensor will have a very high repeatability, but will also be more subject to bounce triggering in the presence of mechanical vibration. By spacing the magnets farther apart, resolution will decrease, but so will the chance of bounce.

Another variation of the bipolar slide-by mode uses a round magnet that is alternately poled in a N-S-N-S... manner around its circumference. This allows the sensor to be triggered and released a number of times per revolution equal to the number of pole pairs on the magnet. This makes this type of sensor useful for counting. In the analog version, a pair producing sine and cosine output becomes equivalent to a resolver for measuring rotational position of a shaft.

If a whole row of magnets were used with a bipolar slide-by-mode sensing system, the effect would be a sawtooth pattern of on/off signals. Use of this many magnets would be expensive, so

instead, the vane-type sensor is used, as shown in Figure 3.2.6g. In a vane sensor, a ferromagnetic target passes between a magnet and Hall effect sensor, which channels away the magnetic field and causes the sensor to release. A typical application is for a fully electronic ignition to replace mechanical contact points on an internal combustion engine's distributor. Similar arrangements of magnets on an engine's flywheel can be used to trigger fuel injection systems.

Consider a typical miniature Hall effect analog position sensor about 5 mm square. This sensor may have a useful operating range of about 1000 G and produce an output over a range of about 5 V. The magnet shown in Figure 3.2.9 has a maximum standoff distance from the sensor of about 1.27 mm (0.050 in.), and it has a field strength of about 1200 G per 1.27 mm. The miniature sensor has a sensitivity of 3 mV/G $\times$ 1200 G/1.27 mm = 2.835 V/mm (72 V/in.). With a 10 mV accuracy power supply and voltmeter, which are relatively common, the mechanical resolution of the sensor will be about 3.5 μm (140 μin.).

Hall effect sensors can also be used as digital on/off proximity sensors. To evaluate the repeatability of this type of sensor conservatively, assume that repeatability is a function of the flux density (G/mm) of the magnetic field and the differential field strength required for triggering of the sensor. If a magnet has a field strength of 1000 G/mm (25,400 G/in.) a typical sensor with a trigger point difference of 40 G would have a position-sensing repeatability of about 40 G/1000 G/mm = 0.40 mm (0.0016 in.). Hall effect sensors also allow the monitoring of the position of objects that are hidden from view as long as the barrier does not block magnetic field lines. This allows Hall effect sensors to sense through nonconductive materials and nonferrous metals such as aluminum and austenitic stainless steel. In general, a Hall effect sensor has the same operating advantages as inductive sensors, although the latter cannot sense through metals.

Hall effect sensors are becoming the sensor of choice for monitoring discrete position of pneumatic and hydraulic pistons with austenitic stainless steel or aluminum walls.[12] In one configuration, the piston itself is manufactured with an integral magnet, and sensors are placed on the outside of the cylinder at discrete locations. In another configuration, the piston is made from a ferrous material and the Hall effect sensor is placed between a magnet and the piston. When the piston moves by, it acts like a pole piece which focuses the magnetic field and triggers the sensor. Some manufacturers offer this type of piston configuration with rails located on the cylinder body for mounting sensors at any desired location. With the use of a three-position four-way valve, a feedback signal from the sensor can be used to position the cylinder midstroke with a repeatability on the order of 6 mm (0.24 in.). Hall effect sensors are also useful for controlling sequencing operations. Hall effect sensors are also used on dc brushless motors to sense the location of the rotor and thus allow for the controlled switching of the current to the motor windings.

Because they sense the presence and strength of magnetic fields, Hall effect sensors can also be used to measure the current in a wire or the temperature of a material whose resistance to current flow changes with temperature. To sense current, a wrap of wire around, or a cable running through a ferrous toroid induces a magnetic field whose strength can be measured by a Hall effect sensor. With this method, it is possible to sense currents ranging from 1/4 to 1000 A. Although this type of sensing system must be carefully calibrated, it has the ability to withstand huge overloads without damage.

Typical Characteristics of Hall Effect Sensors

The following summary of Hall effect sensor characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as gospel.

Size: Analog sensors: 5 $\times$ 5 $\times$ 2.5 mm to 12 mm diameter by 25 mm long with maximum standoff distance on the order of 1.27 mm (0.050 in.). Proximity sensors: 4 $\times$ 3 $\times$ 1.5 mm to 12 mm diameter by 25 mm long. k

Cost (One piece): Analog sensors: $5 for the small sensor, $15 for the large sensor and $5 for a precision stable magnet. Proximity sensors: $2 for a small sensor and $15 for a large sensor. In quantity, prices can be substantially lower.

Measuring Range: On the order of 0.25-2.5 mm (0.01-0.1 in.).

[12] The traditional method for sensing discrete positions of a piston uses a magnet on the piston to activate a reed switch attached to the outside of the cylinder. Reed switches use a thin cantilevered piece of metal that when deflected by a magnet, close a circuit; however, they are subject to mechanical "bounce" caused by machine vibration, which can give false readings.

Accuracy (Linearity): On the order of 1.0 to 0.1% of full-scale range, depending on the range.

Repeatability: Depends on the magnetic field, supply voltage, and the mechanical device. Repeatability is also affected by temperature. If the temperature is held to within 2 C^o, and the electrical system has millivolt stability, sensor repeatability can typically be 10-20 μm (0.0004-0.0008 in.) for analog position sensors, and 25-50 μm (0.001-0.002 in.) for digital sensors.

Resolution: Analog position sensors: as high as 0.5 μm (20 μin.), but more typically with a one millivolt system resolution is on the order of 5 μm (200 μin.). Digital sensors: with custom matched components as good as 1 μm (40 μin.), with typical off-the-shelf components 10 μm (0.0004 in.).

Environmental Effects on Accuracy: Temperature effect on full-scale output is on the order of 1%/C^o, but in some conditions can be as low 0.01%/C^o.

Life: Noncontact and therefore theoretically infinite.

Frequency Response: (-3 dB): Up to 100 kHz.

Starting Force: For some applications, the sensors should be mounted to a nonferrous structure so that when it is brought near the triggering magnet, appreciable forces will not be generated by the magnet.

Allowable Operating Environment: Hall effect sensors are solid-state devices which can be made to withstand virtually any environment as long as it does not interfere with the magnetic properties of the sensor or physically abrade it. Hall effect sensors operate from -40^{o}C to 150^{o}C.

Shock Resistance: Off-the-shelf models are typically rated to 10g. Special models can withstand 40g.

Misalignment Tolerance: The output is a function of the cosine of the misalignment angle.

Support Electronics: The resolution and repeatability of analog and digital sensors is directly related to the performance of the power supply used. Analog output Hall effect sensors require an analog-to-digital converter to digitize their output. Typically, the maximum current drawn by a Hall effect switch is only 20 mA. The cost for the power supply may be $1.00 for a low-cost battery, to $500 for a stable precision power supply.

3.2.3 Inclinometers

Inclinometers are electromechanical levels that are used in many applications where the angular position of a body with respect to a horizontal or vertical reference must be determined with great accuracy, typically on the order of microradians or less. Electrical output allows for data recording or for use as an error signal in a closed-loop control system. Application examples include:

- Collection of movement data during construction and use of all types of structures.
- Measuring motion of local geologic formations.
- Measuring machine platform stability.
- As feedback devices for control of a machine's orientation.
- Slope measurement for rapid determination of straightness.

Precision inclinometers are typically constructed as shown in Figure 3.2.10. They are essentially precision pendulums that use a flexural mount or a precision ball, fluid, or magnetic bearing to support a pendulum. Another type of inclinometer uses a mercury bubble which wets a linear resistor. The more the device tilts, the more the resistor is wet and the greater the change in the output voltage. This type of inclinometer has relatively low resolution because of surface tension effects, but is significantly less expensive than a pendulum-type inclinometer. Since inclinometers are often used in rugged outdoor environments, they are usually enclosed in a sturdy hermetically sealed housing. They require a supply voltage of ±12-18 V dc at 0.015 A and they output a voltage (up to ±5 V dc) proportional to the sine of the angle of tilt.

As an inclinometer is tilted through an angle θ, a position sensor generates an electrical signal which is amplified and fed back to a galvanometer. The galvanometer produces a torque that attempts

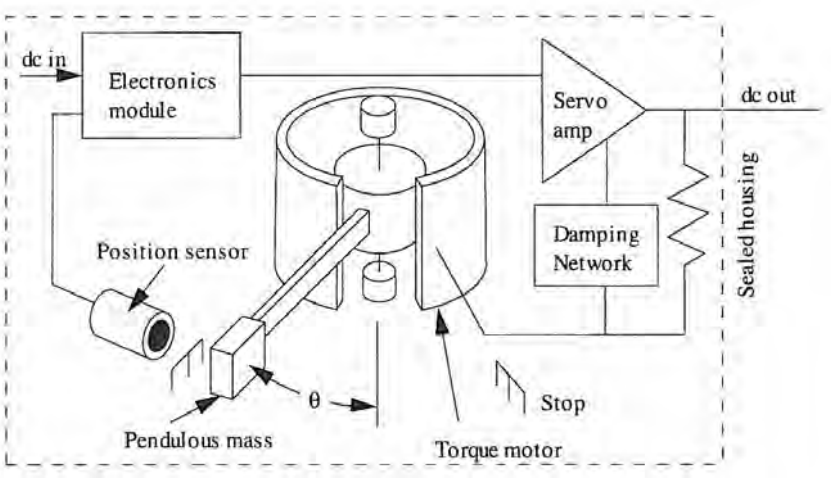


Figure 3.2.10 Construction of a pendulum-type inclinometer. (Courtesy of Lucas Schaevitz.)

to keep the pendulous mass in its original position with respect to the position sensor. When the mass is at this home position, the current applied to the galvanometer to generate the balancing torque is proportional to the sine of the angle of tilt. Passing this current through a large resistance in parallel with the galvanometer generates a voltage also proportional to the sine of the angle of tilt that can be read by external instruments. Because the instrument is designed around a pendulum, its natural frequency is dominated by a $(g/\ell)1/2$ term. The longer the pendulum, the higher the resolution of the device because the tilt angle is amplified by the pendulum length and then measured by a sensor. As would be expected, the higher the resolution, the lower the frequency response.

Typical Characteristics of Inclinometers

The following summary of inclinometer characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as gospel.

Size: 40 mm cube to a cylinder 40 mm in diameter and 50 mm long.

Cost: On the order of $1000-$1400 for 0.1 arcsecond resolution models, $180 for 0.01 degree resolution models. Cost of an analog-to-digital converter or chart recorder is extra.

Measuring Range: From $\pm 1^{o}$ to $\pm 90^{o}$.

Accuracy (Linearity): 0.02% deviation from theoretical sine value for the high-resolution models, 2% for the inexpensive models.

Repeatability: Typically, 0.01 to 0.001% of full-scale output for precision sensors, 1% for low cost ones.

Resolution: 0.1 arcsecond to 0.01 degree (typically 12-16 bits).

Environmental Effects on Accuracy: 0.05-0.005% full-scale/C^{o}.

Life: Internal mechanical contacts are elastically designed for very low stress levels so that life is virtually infinite.

Frequency Response (-3 dB): 0.5 Hz on 1^{o} range-of-motion models to 40 Hz on 90^{o} range-of-motion models.

Starting Torque: Inclinometers are bolted to a structure and thus increase the total mass, which may affect the force or torque required to actuate small structures.

Allowable Operating Environment: Typical inclinometers can operate from 250^{o}K to 70^{o}C. They are hermetically sealed, so fluids and dirt cannot hurt them. They do require a power supply, which must also be protected if the system is to be mounted outdoors.

Shock Resistance: 35-50g constant, 1500g shock survival.

Misalignment Tolerance: A cosine error results from the inclinometer not being properly aligned with the tilt axis.

Support Electronics: A 12 V dc power supply with 0.1% accuracy and 0.01% stability, and a system to read the 5 V dc full-scale output.

3.2.4 Inductive Digital on/off Proximity Sensors

An inductive proximity sensor essentially consists of a wire-wound ferrite (iron) core, an oscillator, a detector, and a solid-state switch, as shown in Figure 3.2.11. The oscillator generates a high-frequency electromagnetic field centered around the axis of the ferrite core, which focuses the field in front of the sensor. When a metal object moves into the electromagnetic field, eddy currents are induced in the object, which results in energy being drawn from the field. This causes a decrease in the amplitude of the oscillations, which switches a transistor in the sensor on or off, depending on the type of sensor used (normally closed or open). When the metal object is removed, the transistor switches back to its original state. The response time of the sensor depends on the effective inductance and resistance of the circuit.

Usually, some amount of hysteresis is built into the sensor to give it different trigger (on) and release (off) points. If the sensor had the same on and off point, then even a small amount of mechanical vibration, electrical noise, or temperature drift could cause readings to flutter (bounce). The difference between the trigger and release points is typically 3-15% of the full-scale range and is set at the factory. Once the release point is set, it is generally as repeatable as the triggering point itself.

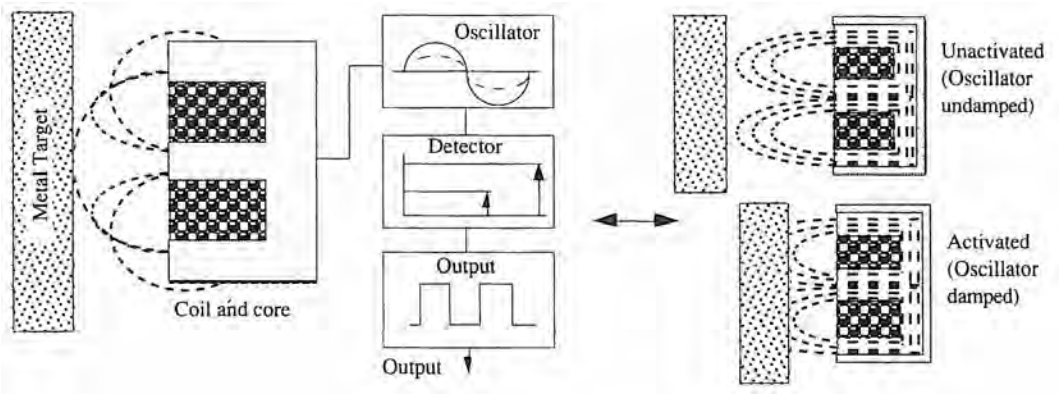


Figure 3.2.11 Operation of an inductive digital on/off proximity switch. (Courtesy of Turck Inc.)

The diameter of the sensor is essentially proportional to the allowable standoff distance of the sensor from the surface. The position of an object can be sensed as it approaches the sensor axially or laterally (head-on or from the side), as shown in Figure 3.2.12. For head-on applications, a shielded sensor is used to focus the sensing field into a sharper point in front of the sensor. For edge (slide-by) sensing, an unshielded unit is used, which results in a broad field. Note that if a head-on sensor is used, it must be protected by a sturdy metal collar to prevent it from being smashed in the event of machine overtravel. In the slide-by mode, if the sensor is used to sense (count) the passing of a line of objects, it is important to space the objects at a distance equal to twice their maximum width. This will ensure that the sensor's oscillating field can be restored once an object moves past, which enables the sensor to release before the next object triggers it.

Like Hall effect sensors, inductive digital on/off proximity sensors provide an inexpensive way to determine the presence of a metal object without touching it. Their advantage over Hall effect sensors is they do not require the use of a magnet, which can attract ferritic dirt. Inductive proximity sensors also can easily be designed to have a large standoff distance on the order of 25 mm. Only a conductive target needs to be supplied, which is often the object itself and the sensor can sense through nonconductive materials. The sensor itself very rarely fails because there are no moving parts in its design; thus inductive proximity sensors are rapidly replacing old-style mechanical limit switches in applications requiring direct interface to an electronic controller.

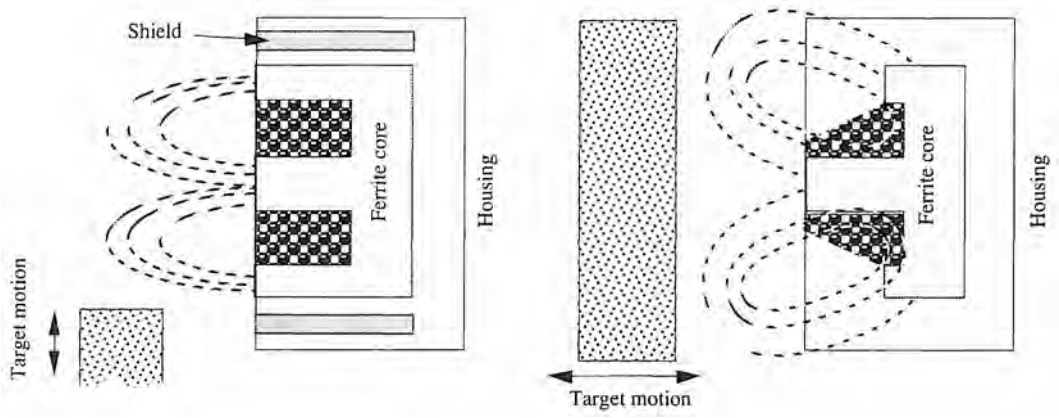


Figure 3.2.12 Shielded and unshielded inductive digital on/off proximity switches for head-on and slide-by sensing, respectively. (Courtesy of Turck Inc.)

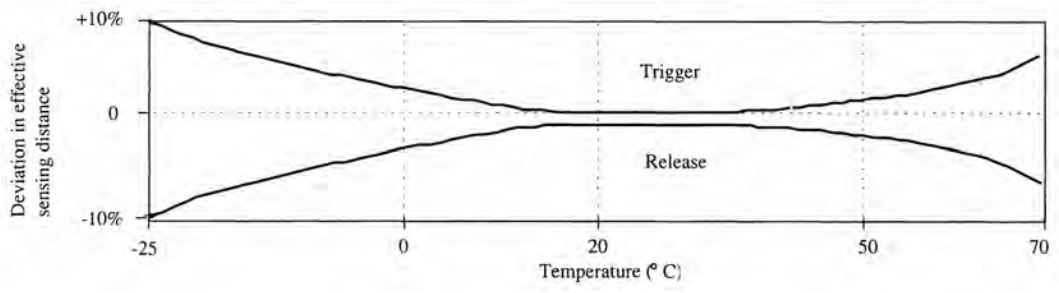


Figure 3.2.13 Effect of temperature on inductive digital proximity sensors. (Courtesy of Turck Inc.)

Regardless of the application, inductive proximity sensors should preferably be mounted with face and target in a vertical plane so that metal filings or chips (dirt) cannot accumulate on the surface, thereby preventing a possible change in the trigger point. Temperature affects their electronics' performance as shown in Figure 3.2.13. A typical family of inductive proximity sensors is shown in Figure 3.2.14. The sensing ranges assume that a ferrous target is used. Other metal targets generally have correction factors associated with their use that decrease the sensing range, as shown in Table 3.2.2.[13]

Inductive proximity sensors only require a power supply and simple user interface circuit to operate. Manufacturers' catalogs usually provide simple direct instructions on how to hook up their sensors. In many cases, the sensors can be ordered with a wide range of outputs, including TTL, which allows them to be plugged into most controllers and personal computers.

[13] From Turck Inc., Minneapolis, MN, (612) 553-9224, Inductive Proximity Switch Catalog, p. 7.

Flush mount	Non-Flush mount	Switch dist. (mm)	Drawing #	Diameter (mm)	L1 (mm)	L2 (mm)	L3 (mm)	Switch freq. (Hz)
•		2	1	12	50	75		1500
•		5	1	18	50	76		1000
•		10	1	30	50	76		500
	•	4	2	12	50	75	10	1500
	•	8	2	18	50	76	10	1000
	•	15	2	30	50	76	15	500

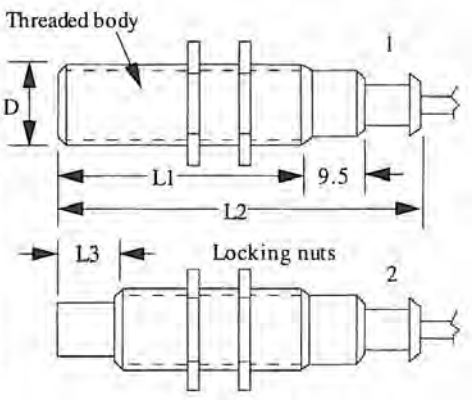


Figure 3.2.14 Typical sizes of inductive digital proximity sensors. (Courtesy of Turck Inc.)

Material	Shielded units	Unshielded units
Aluminum (chunk)	0.30	0.55
Aluminum (foil)	1.00	1.00
Brass	0.40	0.55
Copper	0.25	0.45
Lead	0.50	0.75
Mercury	0.60	0.85
Stainless Steel	0.35-0.65	0.50-0.90

Table 3.2.2 Inductive proximity sensor sensing distance correction factors.

Typical Characteristics of Inductive Proximity Sensors

The following stated characteristics of inductive proximity sensors are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as gospel.

Size: Body size: 8 mm diameter $\times$ 30 mm long to 45 mm diameter $\times$ 60 mm long. Switching distance 1-25 mm.

Cost: \$20-\$40 for the small ones to \$100+ for the large ones.

Measuring Range: From 0.8 to 60 mm standoff distance.

Accuracy (Linearity): Not applicable for this type of sensor.

Repeatability: Typically, 1.0-0.1% of standoff distance.

Resolution: Not applicable for this type of sensor.

Environmental effects on repeatability: From a nominal of 20°C, about 0.2%/C°, as shown in Figure 3.2.13.

Life: If the cable is not flexing, life can be infinite. For applications requiring cable flexing, virtually infinite life can be achieved through proper cable carrier design.

Frequency response: Up to 5000 Hz.

Starting Force: The sensors are noncontact and produce no appreciable electromagnetic forces on the object being sensed.

Allowable Operating Environment: Epoxy potted to meet almost any requirement. The primary consideration is whether contaminants in the area contain metallic components that could accumulate on the sensor of the target and cause the trigger point to change. Allowable operating temperatures are on the order of -25-70°C.

Shock Resistance: On the order of 30g impulse. Continuous vibration: 1 mm amplitude at 55 Hz and 12g.

Misalignment Tolerance: Since the sensors do not contact the object, there is no effect other than error in the actual distance being sensed which can be compensated for by calibration during installation.

Support Electronics: Nominally only a dc (5-65 V) or ac (20-250 V) power supply is required. Output from the sensors can usually be sent directly to another binary device.

3.2.5 Inductive Distance Measuring Sensors[14]

Inductive distance measuring sensors operate as shown schematically in Figure 3.2.15. An ac current in the reference coil creates an electromagnetic field, which combines with the field produced in the active coil. The resultant electromagnetic field will interact with a conductor producing a current flow on the surface and within the target.[15] The induced current produces a magnetic field which opposes and reduces the intensity of the original field. This changes the effective impedance (dynamic characteristics) of the active coil, which is detected by the signal conditioning electronics. The result is an analog voltage output from the black box that is proportional to the distance of the probe from the target.

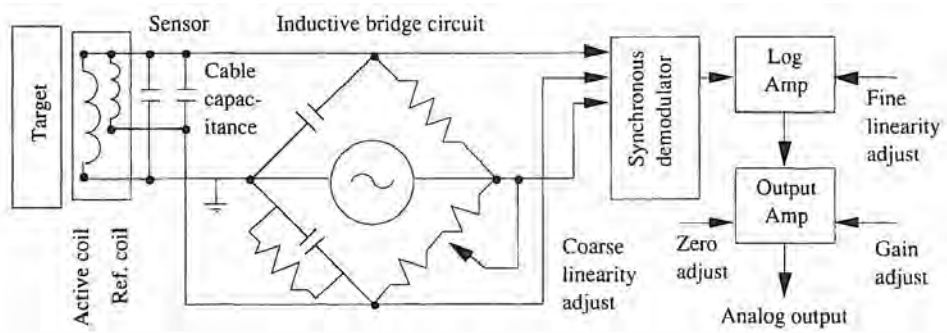


Figure 3.2.15 Schematic operating diagram of an inductive distance measuring system. (Courtesy of Kaman Instrumentation Corp.)

Since the amplitude of the output from the sensor itself decays exponentially with distance from the target, a log-amplifier is often used which provides more amplification the weaker the signal gets. The result is an apparent linear output from the black box. However, with this type of log amplifier, thermally induced electrical noise is also amplified. Thus for maximum stability and linearity, inductive sensors should be used at only a fraction of their measuring range, as shown in Figure 3.2.16. Figure 3.2.17 shows possible methods for achieving increased performance characteristics of inductive distance measuring sensors.

Unlike capacitance sensors, an inductive distance measuring sensors' performance is dependent on the properties of the target material. To achieve a high level of performance, the target should have uniform electrical properties, be a very good conductor, and have low magnetic permeability (should not support a magnetic field). The best target materials are aluminum, copper, and brass. Ferrous objects make poor targets and ideally would have a thin piece of a good target material epoxied or plated on them. The grain size and grain boundaries of ferrous materials' microstructure locally affect the permeability, which causes inductive sensors to see apparent motion when they scan over the surface of a ferrous target. The thickness of the target should be on the order of 0.5 mm (0.02 in.) for gold, silver, copper, and aluminum targets. The thickness of the target should be on the order of 1.3 mm (0.05 in.) for magnesium, brass, bronze, and lead targets.

Since inductive probes' output are affected only by conductive metals, the presence of dirt will not affect their accuracy unless the dirt contained substantial amounts of metallic particles. Temperature changes will affect the output of the electronics, but this can be compensated for by using a reference probe. Sensors are available for operation from 650°C to absolute zero. Figure

[14] The source for this subject material, including figures and tables, is Measurement Solutions Handbook, Kaman Instrumentation Corp., 1500 Garden of the Gods Road, P.O. Box 7463, Colorado Springs, CO 80933, (303) 599-1825. Used with permission.

[15] The current flow is in a circular pattern, which leads to the term *induced eddy currents*.

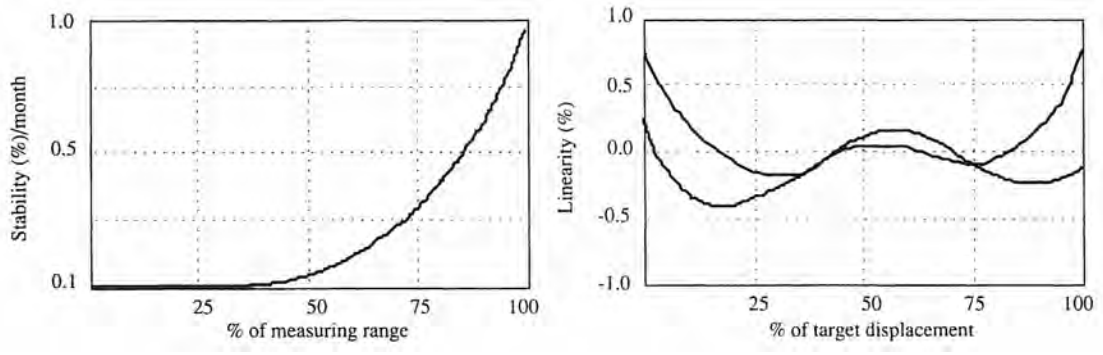


Figure 3.2.16 Typical stability and user adjustable linearity of an inductive distance measuring system. (Courtesy of Kaman Instrumentation Corp.)

	Use first half of full scale measuring range	Use larger diameter sensor	Decrease measuring range	Order temperature compensation option	Increase target thickness to 3 skin depths	Order linearity calibration option	Cover mag. target with non-mag. material	Strap cables down	Add band pass filter on output	Order increased frequency response option	Adjust gain and linearity potentiometers	Decrease offset slightly
Increase sensitivity and output voltage											●	
Increase frequency response										●		
Reduce thermal sensitivity	●	●		●	●		●					●
Increase stability	●	●	●				●	●				●
Improve resolution	●		●				●	●	●			●
Increase linearity			●			●					●	
Increase range		●										

Figure 3.2.17 Methods for improving performance of inductive distance measuring systems. (Courtesy of Kaman Instrumentation Corp.)

3.2.18 shows some of the many types of applications that inductive distance measuring sensors are used for. Note that the larger the range, the lower the resolution, because analog voltages can only be measured with about 12-bit resolution (1 part in 4096). For ultraprecision systems, differential mode sensors must be used.

A differential measuring system allows for the output of one sensor to be subtracted from the output of another, which cancels errors. Assuming that one sensor is looking at a fixed target or the opposite side of the other sensor's target, motion will be recorded and environmental errors will still be canceled. This requires the systems to be matched and hybrid electronic circuits to be used. Matching ensures that both sensors have similar characteristics and they will be affected equally by changes in the environment. Hybrid circuits ensure that all signal processing is done in a small enclosed area to ward off other environmentally induced errors.

Typical Characteristics of Inductive Distance Measuring Sensors

The following summary of impedance probe characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as gospel.

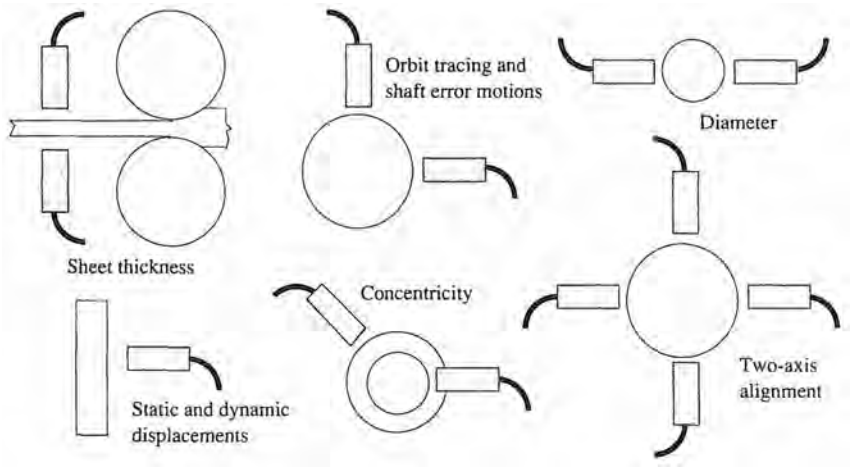


Figure 3.2.18 Some of the many uses for noncontact displacement sensors.

Size: Standard sensors, 5 mm diameter, 20 mm long with 0.5 mm range to 75 mm diameter, 110 mm long, with 50 mm range. Differential sensors, 5 mm diameter, 25 mm long, with 0.23 mm range.

Cost: $750-$1000 for the sensor and black box. $250-$400 for the sensor. $4500 for a dual-axis differential measurement system and $3700 for a single-axis differential measurement system.

Measuring Range: 0.5-50 mm (0.02-2 in.) to ±0.23 mm (0.009 in.) for a differential system.

Accuracy (Linearity): Typically, 0.1% at 25% of full-scale range, 0.5% at 75% of full-scale range and 1.0% at full-scale range.

Repeatability: Typically twice the resolution. With a differential system, stability may be 0.13 μm (5 μinup per month.

Resolution: Varies with sensor size from 0.1 μm (4 μin.) to 3 μm (120 μin.). When used in differential mode with hybrid electronics, resolution can be on the order of angstroms.

Environmental Effects on Accuracy: 0.1% full-scale range/C^o, 0.02% full-scale range/C^o with temperature compensation.

Life: Noncontact solid-state device, so life is virtually infinite.

Frequency Response (-3 dB): Up to 50 kHz.

Starting Force: The sensors are noncontact and produce no measurable or electromagnetic forces on the object being sensed.

Allowable Operating Environment: Environment must not contain loose particles of conductive metal. The sensors are tolerant of fungi, moisture, vacuum, pressure, and immersion (noncorrosive, nonconductive, nonabrasive). Operating temperatures of the sensor and cable range from -55°C to 105°C. Operating temperatures of the electronics range from 0°C to 55°C. Precision systems should be operated at 20°C (68°F).

Shock Resistance: On the order of 20g for the sensor.

Misalignment Tolerance: Since the sensors do not contact the object, no effect other than a cosine error in the actual distance being sensed results from misalignment.

Support Electronics: Nominally only ±12 or ±15 V dc power supply and the factory supplied black box. Output from the black box is an analog signal, so an analog-to-digital converter is required to interface to a computer or controller. Stability of the power supply will affect the output.

3.2.6 Inductosyns®[16]

Inductosyns® utilize inductive coupling between two coils with many overlapping windings to help average out errors. They are available in both linear and rotary form. A linear Inductosyn® is a linear motion transducer consisting of an electromagnetically coupled scale and slider as shown schematically in Figure 3.2.19. The scale, which can be almost any length, is fixed to a machine axis parallel to the direction of linear motion. In general, the scale is constructed from a strip of metal, typically stainless steel, that is covered by an insulator. Bonded to the surface of the insulator via printed circuit technology is a strip of wire that forms a continuous rectangular waveform with cyclic pitch typically 0.l in., 0.2 in. or 2 mm. The slider is fixed to the carriage so that it travels about a tenth of a millimeter above the surface of the scale, and it is constructed in a manner similar to the scale. The required straightness of motion of the mechanical slide depends on the scale manufacturer, but typically is on the order of 10-20 μm.

An Inductosyn® acts like an electrical transformer, such that if the scale is excited by a 5-10 kHz signal (A sin ωt), the outputs from the slider will be

$$S_{13} = B \sin \omega t \sin\left(\frac{2\pi X}{S}\right) \tag{3.2.3a}$$

$$S_{24} = B \sin \omega t \cos\left(\frac{2\pi X}{S}\right) \tag{3.2.3b}$$

where B is the magnitude, X is linear displacement, and S is the spacing of the printed circuit waveform. From these two outputs, the two unknowns B and X can be found. Because the amplitude B of the output waveforms is also found, small-amplitude variations caused by a slightly varying gap between the scale and slider or the presence of foreign material in the gap will not affect accuracy. Accuracy and resolution of motion depend on the number of waveforms per inch, and the resolution of an Inductosyn®-to-digital converter (IDC) used to tell a microprocessor-based machine tool controller where the slide is. Typically, the spacing S between the waveforms is 0.5 mm (0.02 in.) and the outputs are resolved to 12 bits, which gives a resolution of 0.12 μm (5 μinup. Harmonic errors produced by the sine and cosine waveforms are a great nuisance and may require mapping to minimize them. Noise rejection and high accuracy are obtained by the averaging effect of having many coils on the slider and scale overlapping at one time. Inductosyns® are therefore essentially coarse/fine position sensing systems, where the waveform provides coarse position information and the sine wave interpolation provides fine position resolution.

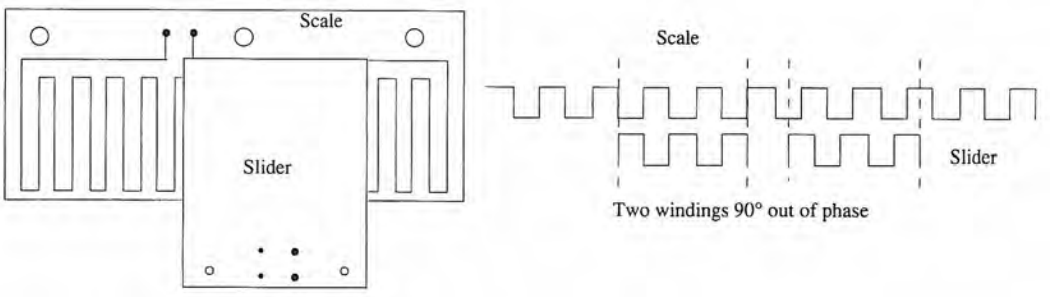


Figure 3.2.19 Operating principle of a linear Inductosyn®. (After Boyes).

There are many other scale-type devices for measuring linear motion, including linear optical encoders and magnetic scales. The linear Inductosyn® (and magnetically encoded scales), however, is one of the most rugged types of linear scales. Its performance is seriously affected only when its parts are abraded by dirt until an electrical short is generated. Note that dirt can cause a linear optical encoder system to miss a count.

[16] For an detailed discussion of the operation, application, and data conversion techniques of Inductosyns®, resolvers, and synchros, see the book Synchro and Resolver Conversion Handbook, G. S. Boyes (ed.), Analog Devices Inc., Norwood, MA. Published by Memory Devices Ltd., Central Ave., East Molesey, Surrey KT8 0SN, England.

There are four types of commercially available linear Inductosyns®. The first two are standard or narrow-width 10 in. segments. Each segment is connected to the oscillator, so for long lengths there is not a drop in signal level. The segments also conform to the shape of the surface they are attached to. Tape-type Inductosyns® use a tape that is anchored at one end of the machine component and at the other end via a tension adjusting mount. The scale on the tape is printed while the tape is under tension. By adjusting the tension during mounting, absolute accuracy over the length of the tape scale can be fine tuned. Once the tension is adjusted, the tape is locked in place until the machine needs to be recalibrated. If the tape is bonded or laid onto a flat surface so that it follows the contour, mapping and calibration with a laser interferometer may be required. An adjustable Inductosyn® has clamps and tension-adjusting screws every 3 in. along its length. For systems where software-based error correction techniques cannot be used, the clamps and tension adjusters allow the length of the scale to be adjusted discretely along the entire length.

The stator of a rotary Inductosyn® has two separate rectangular printed track waveforms arranged radially around the disk. The sine track is made up of sections which alternate with the cosine track. Both tracks cover the entire outer region of the stator. Similarly, the rotor corresponds to a linear Inductosyn's® scale, with the entire outer region also covered by a nearly rectangular waveform. Because the entire waveform of the rotor covers both of the waveforms on the stator, exceptionally good random noise reduction is obtained by means of an averaging effect, but periodic errors are not reduced. The latter once again must be mapped and compensated for in the sensor's black box. To help minimize eccentricity errors, the rotor is usually mounted directly to the rotary axis structure and uses its bearings for support.

Rotary Inductosyns® can have a high number of poles, which creates an averaging effect, and the ability to resolve a clean analog signal to 16 bits gives a potential resolution of 0.1 μrad (0.02 arcsecond). Accuracy can be 1.0 μrad and repeatability on the order of 0.5 μrad. Only the best optical encoders can match this performance; however, optical encoders cannot tolerate contamination the way an Inductosyn® can. Hence Inductosyns® are often used on precision rotary tables.

It is important to note that as the number of poles increases, the sine and cosine waveforms become so closely spaced that coupling between the two can occur. Therefore, although rotary Inductosyns® with a large number of poles may have high resolutions, they may be less accurate than Inductosyns® with fewer poles. If mapping and software-based compensation techniques are used, this effect can be negated and the advantages of ultrahigh resolution and insensitivity to dirt and contamination can still be realized.

Typical Linear and Rotary Inductosyn® Characteristics

The following summary of linear and rotary Inductosyn® characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as gospel.

Size: Linear Inductosyns®, (1) standard-type 1000-mm and 250-mm segments are 60 mm wide and 10 mm tall, while the slider is 75 mm wide, 100 mm long, and 80 mm tall; (2) narrow-type 250-mm segments are 25 mm wide and 10 mm tall, while the slider is 30 mm wide, 100 mm long, and 40 mm tall; (3) tape scale is 0.30 mm thick and 20 mm wide and can be 30 m long. 300-m-long scales have been proposed for shipyard use. The slider is similar to that used for the narrow segment scale. Rotary models come in 300, 150, and 75 mm (12, 7, and 3 in.) diameters. Self-contained units are on the order of 50-75 mm (2-3 in.) thick. Modular units' disks are only a few millimeters thick.

Cost: (1,2) Standard and narrow 250-mm segments cost about $140 for the scale and $300 for the slider. (3) Tape scales cost about $0.23/mm ($70/ft) and $175 for the slider. They also come prepackaged (scale and slider) as a bolt-on unit for about $1400 + $0.33/mm travel. A black box to convert sine rotary waveforms to digital counts which can be sent to a CNC controller or personal computer costs about $1000.

Measuring Range: Up to 30 m (100 ft.) linear, continuous rotary.

Accuracy (Linearity): For linear systems, as good as 1 μm (40 μin.). For rotary systems, accuracy depends on the number of poles and if compensation methods are used to prevent sin/cos wave coupling in units with a large number of poles. Typically, a 305-mm (12 in.) diameter unit will have 5-10 μrad accuracy.

Repeatability: Typically, 10 times the accuracy.

Resolution: Depends on the Inductosyn®-to-digital converter but can be as high as 0.025 μm (1 μinup for linear Inductosyns® and 0.05 arcsecond for rotary models. More typical values are 0.5 μm (20 μinup and 0.5 arcsecond respectively.

Environmental Effects on Accuracy: 0.1% full-scale range/C°, 0.02% full-scale range/C° with temperature compensation.

Life: Noncontact device, so life is virtually infinite if the sensor is not abraded by dirt.

Frequency Response: Limited by discretization and count rate of electronics. Maximum rate of linear travel is typically on the order of 1.5 m/s (60 in./s). The electronics can typically handle 16 bits of information in 10 μs, which corresponds to 15 μm (0.0006 in.) of motion in this interval at maximum speed.

Starting Force: Inductosyns® are noncontact and thus they produce no measurable relative electromagnetic forces between their components.

Allowable Operating Environment: Inductosyns® have been designed for use in vacuum and high-pressure oil environments from 10°K to 180°C. They can operate effectively as long as the medium they are in does not contain an abrasive conductive slurry that could wear off the protective insulating layer and short out the device. Some highly conductive media may lower the output signal, which can affect the resolution of the device. The environment should not contain loose particles of conductive metal. Note that the black box electronics must be kept in a protected environment.

Shock Resistance: The scale and slider are essentially printed circuit devices mounted to metal blocks, so as long as machine vibration does not cause contact between the two, they are insensitive to vibration. Note that wires must be properly held by strain reliefs to prevent their connections from fatiguing.

Misalignment Tolerance: Since the sensors do not contact the object, no effect other than a cosine error in the actual distance being sensed results from misalignment. If the gap between the slider and scale increase by more than about a tenth of a millimeter, the signal level will also decrease, thereby decreasing resolution. For rotary systems, cyclic angular error $\Delta\theta$ is a function of radial motion ε_r and Inductosyn® diameter D: $\Delta\theta \approx 0.1\varepsilon_r/D$.

Support Electronics: Inductosyns® are analog devices that require a precision power supply, oscillator, demodulator, and analog-to-digital conversion device. Normally, the latter two are combined in an Inductosyn®-to-digital converter (IDC) that can have up to 16 bits resolution per cycle of the Inductosyn®. Cost of the black box electronics, assembled and ready to plug in and supply digital position data to a CNC controller, may be on the order of $1000 per axis. Because they are analog devices, resolution is mostly dependent on the electronics. Cost of a 16-bit IDC by itself is on the order of $400.

3.2.7 Linear and Rotary Variable Differential Transformers

LVDTs and RVDTs employ the principle of electromagnetic induction to sense linear and rotary motion with a range of motion typically less than 10-20 cm or one revolution, respectively. An LVDT consists of three elements as shown in Figure 3.2.20: (1) the armature or core, which is made from a ferritic (magnetic) alloy; (2) the stem, which is typically made of a nonmagnetic alloy and anchors the core to an object; and (3) the transformer, which consists of a primary ac excited coil and two secondary coils enclosed in a protective insulated magnetic shield. The armature moves within the hollow core of the coil without making physical contact. When the primary is excited by an ac current, the armature induces a voltage in the secondary coils. The position of the armature affects the output voltage from the two secondaries, one being plus and the other being minus, to achieve directionality.

To measure the relative position of two bodies, LVDTs can be used with the core attached to one body and the transformer attached to a second body, or they can be purchased as complete assembled units with spring-loaded sensing tips. LVDTs have the following practical operating characteristics:

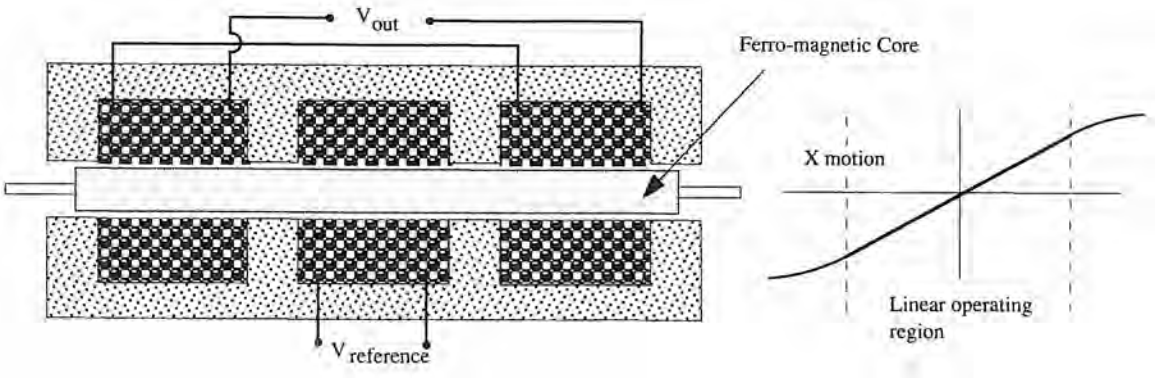


Figure 3.2.20 Schematic construction of a linear variable differential transformer (LVDT). (Courtesy of Lucas Schaevitz.)

1. There is no mechanical contact between the windings and the armature. As a result there is no friction, mechanical hysteresis, or wear. Therefore, life is essentially infinite and reliability is extremely high.
2. LVDTs require a stable ac excitation source and a black box signal conditioner to convert the ac output from the secondary coils into a dc voltage. Their symmetrical construction leads to an extremely stable null point (zero voltage output at the home position), which allows them to be used in very high gain closed-loop control systems.
3. When the core is properly supported, there is no stick/slip; thus LVDTs virtually have infinite resolution and have been used to measure displacements in the nanometer range. Accuracy and resolution are limited only by the signal conditioning electronics and the analog-to-digital converters.
4. Their high-level output simplifies black box circuitry.
5. They can over-range without harm and they are relatively insensitive to lateral motion of the core.
6. They are rugged and shock resistant and require virtually no maintenance. They can operate from cryogenic temperatures to very high radiation levels in nuclear reactors. They can also be made to operate inside a high-pressure hydraulic cylinder.
7. LVDTs are most often used to measure relative motion between objects whose surfaces only move a little bit with respect to each other. They are not practical, for instance, for use in measuring the radial error motion of a high-speed spindle.

Conventional LVDTs require a stable ac excitation source and a black box to amplify and convert the output to a dc signal. Old black box technology required a relatively large amount of printed circuit board real estate, on the order of 3×5 in.; however, single monolithic circuits are now available. If digital output from an LVDT system is desired, then it is best to use a tracking LVDT to digital converter. This is a monolithic device that continually tracks the output from the LVDT. Input to the device is ± 15 V dc, 5 V dc, the excitation signal to the LVDT, and the output signals from the LVDT. Internally the device uses an up-down counter and a comparator to update continually (on the order of microseconds) the digital word representing position. Cost of this device is on the order of $110 for the new monolithic versions and $300 for the older hybrid versions.

A dc LVDT uses a special hybrid circuit to convert a dc supply voltage to an ac signal to drive the LVDT and to amplify and demodulate (convert from ac to dc) the output from the LVDT. These are more convenient to use than conventional LVDTs because the electronics fit into a smaller package, and only dc instrument-level voltage is required to power them. They are significantly less expensive (50%) than an ac LVDT with its black box; however, they are not as stable or accurate and are much more environmentally limited in their application. Still, dc LVDTs have valuable utility because they can greatly simplify measuring system design.

RVDTs look very much like small electric motors. They produce an output voltage whose magnitude varies linearly with the angular position of its shaft. In place of the core of the LVDT, the RVDT has by a specially shaped ferromagnetic rotor. The angular motion of the rotor emulates the

displacement of the core of the LVDT. The coupling between the rotor and the stationary windings is electromagnetic only.

Typical Characteristics of LVDTs and RVDTs

The following summary of LVDT characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as gospel.

Size: Linear models: 5 mm diameter by 11 mm long with 0.13 mm range to 21 mm diameter by 1.6 m long with 0.6 m range. Rotary models: 25 mm diameter by 25 mm long with 5 mm diameter shaft 10 mm long.

Cost: For the LVDTs above (one-piece price): $150, $1100, and $230, respectively. A "black box signal conditioner" to provide a dc output voltage costs about $450. For a dc LVDT unit with 0.13 mm (0.005 in.) range total measuring system, cost is $250. Rotary LVDTs cost on the order of $250 plus the cost of a black box. A monolithic LVDT to digital converter costs about $110.

Measuring Range: Linear 0.1 mm to 1 m (0.004 in. - 40 ft.). Rotary $\pm 60^{o}$.

Accuracy (Linearity): 0.01-0.05%, depending on range of motion. Rotary transducers typically are not as accurate (one-half or less) as linear transducers.

Repeatability: The device itself is noncontact and repeatability depends on the signal conditioning electronics and electrical noise. Typically, repeatability is 10 times the accuracy.

Resolution: Theoretically infinite, and typically as small as 0.1 μm on linear models and 25 μrad on rotary models. Primary limit of resolution depends on the resolution of analog-to-digital conversion device or gage used to convert the output to a readable value.

Environmental Effects on Accuracy: On the order of $0.1\%/C^{o}$ of full-scale range.

Life: Linear models are noncontact and therefore can have infinite life. Rotary models' life is governed by load on bearings and number of cycles. Typical life of RVDT's can be 10-100 x 10^{6} cycles.

Frequency Response (-3 dB): 400 Hz to 10 kHz.

Starting Force and Torque: Most linear models are non contact, so starting force is zero. Rotary models' starting torque is on the order of 0.14 N-mm (0.02 oz-in.).

Allowable Operating Environment: Noncontact LVDTS can operate virtually anywhere as long as gunk does not build up and cause contact and wear between the core and coils. Operating temperatures range from 218^{o}K to 150^{o}C for regular models to 3^{o}K to 175^{o}C for special models.

Shock Resistance: 1000g for 11 ms, 20g up to 2 kHz continuous.

Misalignment Tolerance: Lateral misalignment of the core and coils will not affect the measurement as long as the two do not make physical contact. Angular misalignment causes a cosine error. RVDTs require the use of flexible couplings.

Support Electronics: Require a black box to provide an excitation frequency (l kHz) and to convert the output to a dc level. The black box typically uses 115 V ac power. An alternative is to use a separate 1 kHz oscillator, ± 15 V dc and 5 V dc power supply, and an LVDT to digital converter. Dc LVDTs require only ± 15 V dc at 0.02 A.

3.2.8 Magnetic Scales

Magnetic scales use a sliding sensing head to detect sine and cosine waves from a magnetically recorded scale. The scale is typically a wire imprinted with thousands of north/south pole pairs. As shown in Figure 3.2.21, this type of sensor exists in linear form, such as that sold under the trade name Magnescale® by Sony Magnescale, Inc. A Magnescale's® principal advantage over optical scales is that it is relatively immune to dirt and fluid contamination. Only when the dirt builds up to the point where the unit jambs or parts become abraded will the unit fail. As with all sensors, however, it is best to protect them as well as is possible. For short stroke systems, the scale is often

mounted on the moving member, and the read head is mounted on the fixed member in order to simplify cable handling requirements.

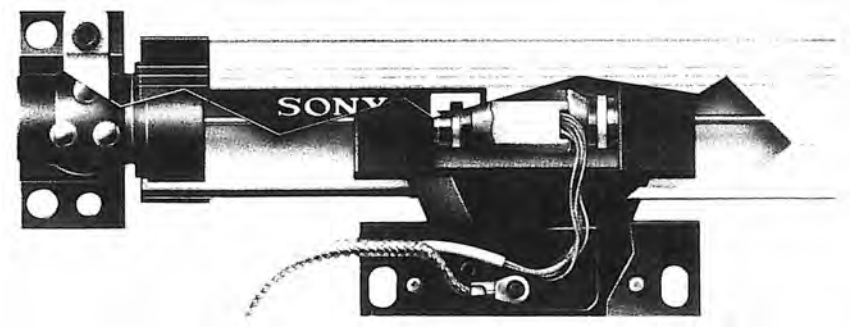


Figure 3.2.21 Magnetically encoded linear scale and sliding read head. (Courtesy of Sony Magnescale Inc.)

Magnescales® are incremental position measuring devices. Operation of the scale generates two square waves which the electronics decode to produce a digital word representing change in position from the initial startup point. The magnetic scale is recorded onto a thin wire which passes through the sensor head; thus minor misalignment (0.1 mm overall) of the scale with the axis of motion being measured results only in a measurement error, not in increased wear of the sensor or loss of signal. When numerical control is used, the performance of the entire machine, including axis mechanical and sensor errors can be mapped with a laser interferometer. In general, magnetic scales are fairly fault tolerant of the mechanical system.

Typical Characteristics of Magnetic Scales

The following summary of linear magnetic scale characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as gospel.

Size: Three basic sizes (of Magnescales®), head and scale combined envelope: (1) 40 mm wide, 20 mm tall, 140 mm + stroke = length (to 1.3 m). Read head 40 mm long. (2) 55 mm wide, 25 mm tall, 150 mm + stroke = length (to 2.1 m). Read head 60 mm long. (3) 75 mm wide, 55 mm tall, 350 mm + stroke = length (to 3 m) long. Read head 60 mm long. Ribbon scales up to 30 m long are available.

Cost: Scale and head: (1) \$240 + \$0.95/mm (\$460 minimum cost); (2) \$230 + \$0.95/mm; (3) consult vendor. Interface to machine tool controller or personal computer runs from \$325 to \$1200, depending on features desired.

Measuring Range: To 30 m.

Accuracy (Linearity): 2.5 μm + 2.5 μm/m scale length.

Repeatability: Included in the accuracy specification and thus not quoted separately.

Resolution: The resolution is a function of the spacing of the magnetically encoded scales and the phase resolving electronics. Resolution is on the order of 0.5 μm independent of scale length.

Environmental Effects on Accuracy: Unless abraded, temperature affects accuracy via expansion of the scale by about 11 μm/m/C°. The sensor is shielded, so external magnetic fields do not affect this type of sensor under normal operating conditions found in most machine tools.

Life: Seals can wear out, and if not replaced the scale can be abraded by dirt and be ruined. In a properly designed and sealed system, life can be infinite. In an improperly designed system where the cutting fluid is allowed to become caustic and full of dirt, life can be reduced to a period of only a few years.

Frequency Response: Limited by count rate of electronics. Maximum rate of linear travel is on the order of 1.5 m/s (60 in./sec). The sampling period is 20 μs, which corresponds to 20 μm (0.0008 in.) displacement at the maximum rate of linear travel.

Starting Force and Torque: On the order of a half a newton to overcome seal friction. For precision applications in clean areas, the seals can be removed to reduce this value to near zero.

Allowable Operating Environment: Magnetic scales are designed to be placed on a machine tool without any special sealing or protection requirements. However, there is no use taking added risks or chances with such a critical part of the machine. Like all sensors, they should be treated with as much respect as bearings. Operating temperatures range from 0 to 50°C (limited by the electronics).

Shock Resistance: On the order of 30g.

Misalignment Tolerance: 0.1-mm overall allowable misalignment. For very long spans, systems will inevitably sag, so calibration may be required. Unless the scale is damaged, misalignment results in cosine errors but not in loss of scale performance.

Support Electronics: 115 V ac and factory-supplied black box to provide digital output.

3.2.9 Magnetostrictive Sensors

"Magnetostriction refers both to the dimensional changes that occur in ferromagnetic materials in the presence of imposed magnetic fields and to the magnetization changes that occur in ferromagnetic materials exposed to mechanical stress."[17] This is a profound physical effect that lends itself to the design of some interesting sensors. Among the most useful for the machine design engineer are noncontact torsion sensors and linear position sensors.

Noncontact Torsion Sensors

It is often necessary to monitor the torsional stress in a rotating shaft in order to make a direct measurement of the power being transmitted, or to prevent overstressing of the shaft. The traditional method for monitoring shaft stress is to use strain gages mounted to the shaft and supplied with commutation means to supply voltage and transfer the output from the gages. This is an expensive, electrically noisy, and potentially unreliable method.

The Villari effect is a change in magnetization that occurs in the direction of mechanical strain. When a shaft is in torsion, shear strains are produced that are oriented at 45° to the axis of shaft rotation. By placing several Villari differential torque transformers around the circumference of a shaft, the torsional stress can be accurately measured, as shown in Figure 3.2.22. Adding the output from all the sensors helps to negate the effects of small shaft eccentricity (up to 0.1 mm or so) and varying magnetic properties of the shaft. Note that this arrangement of sensors must be initially calibrated because ferromagnetic materials will not necessarily behave in the same manner even under similar conditions. The closer the sensors are to the shaft, the higher the output signal, which is typically on the order of millivolts. Thus this type of sensor is accurate to about 1 part in 1024 (10 bits) when properly calibrated.

Linear Position Sensors

The Guillemen effect is exhibited when a magnetic material is immersed in a magnetic field, causing it to change its dimensions. For a long slender rod, the diameter of the rod changes locally. This local change in diameter can serve as a point of reflection for a longitudinal stress wave. By placing an ultrasonic transducer at one end of the rod, a stress wave can be sent down the rod. The time it takes for the wave to travel to the point of diameter change and return can be measured. The elapsed time can then be used as a measure of the distance between the magnetic field and ultrasonic transducer.

The Wiedemann effect is the twist produced in a wire, located in a longitudinal field, when a current flows through the wire. The longitudinal field and the circular field of the wire interact to form a helical resultant. The magnetic material expands (contracts) parallel to the helical lines of force, which causes torsional strain in the wire. In a Temposonics® transducer, a torsional strain pulse is induced in a specially designed magnetostrictive tube by the momentary interaction of two magnetic fields. One of these fields emanates from a permanent magnet which passes along the outside of the tube. The other is produced by a current pulse which travels at ultrasonic speed down

[17] H. E. Burke, Handbook of Magnetic Phenomena, Van Nostrand Reinhold, New York, 1986. There are many variations of magnetostrictive sensors which are discussed in this reference.

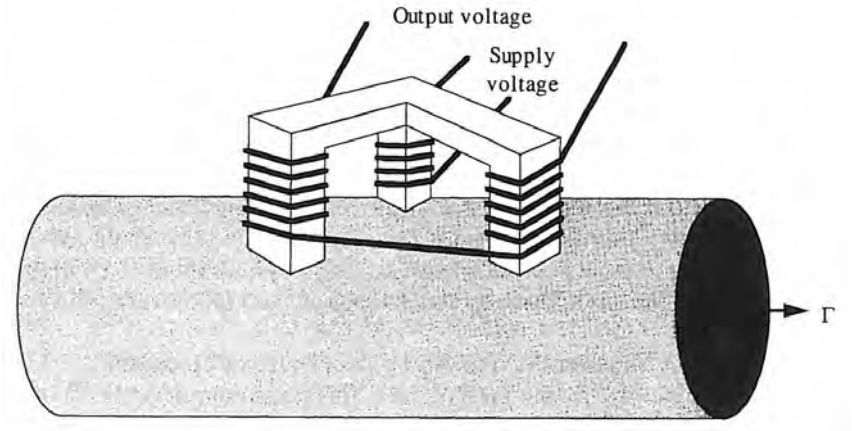


Figure 3.2.22 Villari noncontact torque transducer for rotating shafts. (After Burke.)

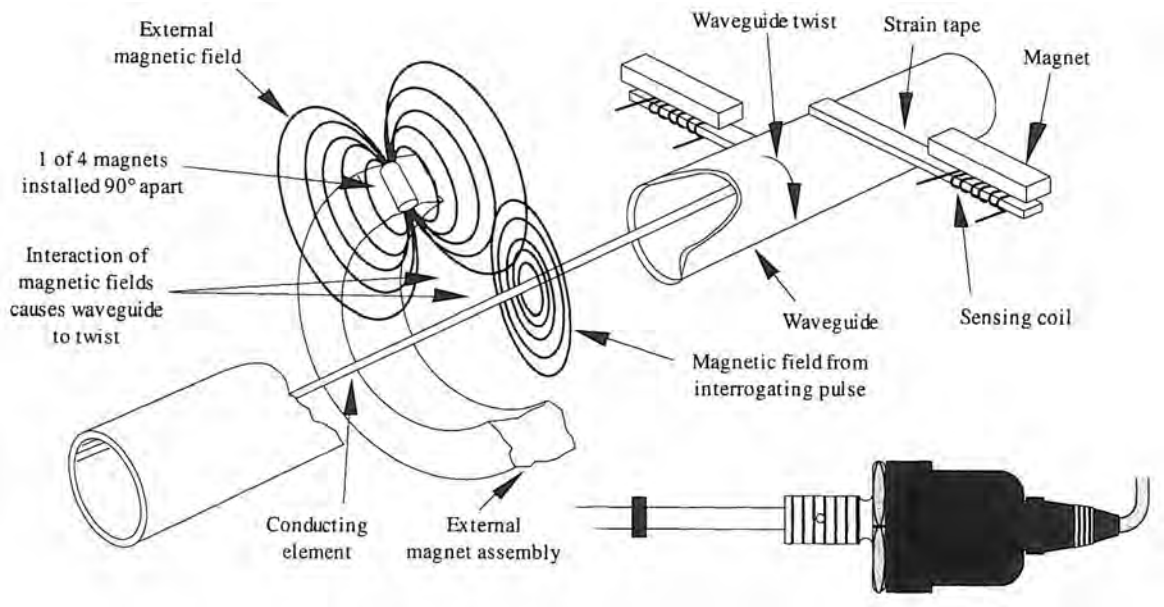


Figure 3.2.23 Operating principle of a Temposonics® linear position sensor and a typical housing configuration. (Courtesy of MTS Systems Corp.)

the tube that serves as a waveguide, and is detected by a coil arrangement at the end of the device.[18] The interaction between the fields and a typical transducer's construction are shown in Figure 3.2.23. Both rigid and flexible tubes can be obtained.

These types of sensors are ideal for use *inside* hydraulic cylinders, particularly those with very long strokes, and on other machines where extreme robustness and moderate resolution are required. They have also found application on large Cartesian systems such as robots and overhead cranes. Perhaps the nicest thing about them is that the traveling head consists only of a permanent magnet and therefore requires no wires. The scale and transducer are mounted to the fixed part of the machine.

[18] For example, a sensor of this type is sold under the trade name Temposonics™ by MTS Systems Corp., Box 13218, Research Triangle Park, NC 27709. Lucas Schaevitz Inc. also makes a sensor of this type which it sells under the tradename MagnaRule®: Lucas Schaevitz Inc. 7905 North Route 130, Pennsauken, NJ 08110.

Typical Properties of Linear Motion Magnetostrictive Effect Sensors

The following summary of linear motion magnetostrictive sensors are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as gospel.

Size: Diameter of magnetostrictive rod housing: 10 mm, diameter of magnet: OD = 35 mm, ID = 14 mm, Coil housing: Diameter, 40 mm, length, 100 mm.

Cost: $700, including electronics.

Measuring Range: Up to 3 m (9 ft.).

Accuracy (Linearity): $\pm 0.05\%$ of full-scale range.

Repeatability: Typically, repeatability is two to five times the accuracy.

Resolution: Practical resolution in on the order of 12 bits, with 16 bits attainable, of the full-scale range.

Environmental Effects on Accuracy: Temperature affects accuracy via expansion of the scale by about 11 μm/m/C°. Also, temperature may affect the speed of pulse propagation, but for small temperature excursions this generally should not be a problem.

Life: Noncontact and therefore infinite unless physically damaged.

Frequency Response (-3 dB): Typically, 200-50 Hz for lengths of 0.6-2.5 m, respectively.

Starting Force and Torque: Noncontact and balanced magnetically; therefore, no force or torque is associated with their operation.

Allowable Operating Environment: They are designed to be placed on a machine tool without any special sealing or protection requirements. Operating temperatures range from 2°C to 66°C. Temperature affects performance by causing the rod to expand linearly. They can operate inside a high-pressure hydraulic cylinder.

Shock Resistance: Not specified.

Misalignment Tolerance: As a rule of thumb, every 0.5 mm (0.020 in.) displacement of the magnet perpendicular to the axis will produce a change of 25 μm (0.001 in.) in the axial measurement.

Support Electronics: 115 V ac and factory-supplied black box. Optional digital or analog output signal.

3.2.10 Mechanical Switches

Mechanical switches can trace their history back to the discovery of electricity and are available in an almost infinite array of sizes and shapes. The two ends of the available spectrum are represented by small, unsealed switches for use in office-type environments, up to rugged, first-sized, sealed switches for use in sawmills and coal mines. They are commonly used to detect open doors and as overtravel limit switches for machine tool axes. Mechanical switches generally are not as accurate as a noncontact sensor such as an inductive proximity sensor. However, they are much easier to mount and the standoff distance from the object is not very critical. Repeatability is limited to about 10 μm (0.0004 in.), and they are available to withstand almost any operating condition. In general, however, they are not meant for high-cycle or high-speed applications. Inexpensive models can cost less than $1, while rugged environmentally sealed models can cost on the order of $100.

3.2.11 Piezoelectric Material Based Sensors

Materials which exhibit the piezoelectric effect have a crystalline structure which begins to oscillate when a force is applied along a certain axis. This releases higher-energy electrons in the material, inducing a flow of current. Measuring the induced piezoelectric current allows for a quantization of the physical process that induced the state of stress in the piezoelectric material. The most common piezoelectric transducers are accelerometers, precision load cells, thin plastic film pressure transducers, and ultrasonic transducers.

Accelerometers

Accelerometers can be used to help control velocity, sense vibration, or determine an object's position. Acceleration feedback is required for velocity control of machine systems in the same manner that velocity feedback acts in position control systems. Measuring vibration allows rotating machinery to be dynamically balanced and aligned. In some applications, measuring acceleration can also be used to detect tool wear.[19] When used as part of inertial guidance systems, position is obtained by twice integrating their signals, where drift is compensated for by periodic update of true position.

For inertial applications, a force balance system constructed much like that of the tilt sensor shown in Figure 3.2.10 is usually used instead of a piezoelectric element. A pendulum-type accelerometer's frequency response is limited to about 200 Hz, which is much lower than that of a piezoelectric accelerometer, but their resolution may be 1 μg which is often much higher. Piezoelectric accelerometers are monolithic devices consisting of a mass attached to a piezoelectric element and preloaded with a spring. Since they have no moving parts supported by bearings, they can have frequency responses up to 100 kHz and can be made as small as a thumbnail. Monolithic integrated circuit accelerometers have even been made. When the mass anchored to the end of a piezoelectric element is accelerated, it induces a force on the piezoelectric element. The resultant current flow can be calibrated to an acceleration level using a standard reference accelerometer or a precision laser interferometer-based accelerometer calibrator. Piezoelectric accelerometers are the most common form of accelerometer used for analyzing vibrating machinery. Their cost can range from $50 to over $1000, depending on the required performance level.

An anecdote well worth mentioning regarding the use of accelerometers relates to a case of improper mounting. In this instance, a student was trying to measure the vibration levels in a machine using an accelerometer. He *glued* the accelerometer to the machine using silicon rubber. What he got was a mass (the accelerometer) attached to a vibrating platform via a soft spring. As a result, his system had a natural frequency well below that he was trying to measure, and the system behaved like a low pass filter. The moral of the story is "think about the whole system!"

Typical Characteristics of Accelerometers

The following summary of accelerometer characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as gospel.

Size: From thumbnail to fist size.

Cost: From $100-$5000. A typical instrumentation quality accelerometer costs $1000.

Measuring Range: From tenths to thousands of gs.

Accuracy (Linearity): Typically, 1-0.05% of full-scale (full ±range).

Repeatability: Typically, two to five times the accuracy.

Resolution: On the order of 0.001% of full scale.

Environmental Effects on Repeatability: On the order of 0.02%/C°.

Life: Accelerometers must be periodically recalibrated, but physical life can be infinite.

Frequency Response (-3 dB): From 200 to 100 kHz for pendulum and piezoelectric types, respectively.

Starting Force: The effect of the mass of the accelerometer on the object may have to be considered in some cases.

Allowable Operating Environment: Hermetically sealed to meet almost any requirement from 200°K to 90°C. Piezoelectric accelerometers are sensitive to higher temperatures, which can "anneal" the piezoelectric material.

Shock Resistance: Accelerometers are designed to measure shock and vibration.

Misalignment Tolerance: A cosine error results from misalignment. Note that they are, however, susceptible to errors caused by cross axis motion by an amount on the order of 0.002g/g.

[19] See, for example, K. Yee and D. Blomquist, "An On-Line Method of Determining Tool Wear by Time-Domain Analysis," SME Tech. Paper MR 82-901, 1982.

Support Electronics: Pendulum type requires ± 15 V dc power supply. Piezoelectric types require amplification of the output signal, which may be in the milli- or microvolt range.

Precision Load Cells

Precision load cells use a thin film of crystalline piezoelectric material to measure nanostrains using a gage only centimeters across. This gives a sensitivity on the order of a nanostrain, which is two to three orders of magnitude more sensitive than available with metal strain gages. The measurable load range, however, is limited because the maximum strain level may only be a few microstrain. This type of load cell is usually custom designed.

Tactile Sensors

Plastic piezoelectric thin films[20] can be made into large thin sheets, in series as large as square meters, costing as little as \$300/m^2. Plating desired regions with a thin metal film produces discrete pressure-sensitive locations. Compared to ceramic piezoelectric materials, they produce 10 to 20 times the voltage for a given state of stress; however, they can only withstand pressures of about 10^6 N/m^2 (150 psi). The piezoelectric constant for a typical piezoelectric film is 216×10^{-3} (V/m)/(N/m^2). This means that if a 0.25-mm (0.010 in.)-thick piece of film had 100,000 N/m^2 pressure applied to it, about 5 V would be generated. Sustainable current levels are almost zero, however, so a very high impedance voltage measuring device must be used.

Because the plastic film is flexible, unlike brittle ceramic piezoelectric materials, it has a growing number of applications in consumer products and robotics. It can be used to make lightweight, miniature microphones and touch-sensitive keypads. For robotic grippers, a sheet of the material can have sensing regions deposited onto it in printed circuit board fashion. This creates a layer of "skin" with "nerve" endings that enable a gripper to determine the pressure distribution on the surface of a robot hand as it picks up an object. How the robot controller would use this information to better grasp and manipulate objects remains a very difficult and complex problem.

Ultrasonic Piezoelectric Sensors

Ultrasonic sensors use a stress wave generating/receiving element to create a pressure wave in an object and record the return of echo pulses. Measuring the time it takes the echoes to return gives a measure of the distance of inclusions or features within the object. There are three types of ultrasonic transducers: piezoelectric, magnetoresistive, and electrostatic. They are discussed in Section 3.2.14.

3.2.12 Potentiometers

Any device in which a change in a physical process changes the device's effective electrical resistance could be classified as a potentiometer. However, as the term has evolved in position measurement applications, a potentiometer has become synonymous with a transducer that measures linear or rotary mechanical position by means of a variable resistor. A potentiometer basically consists of a coil of wire or a high-resistance film and a wiper, whose position along the coil or film is established by the motion of the process being measured. A dc voltage is applied across the entire length of the coil or film, and the wiper acts as a pickoff point for an intermediate voltage; thus the device acts like a continually variable resistor. Measurement of the voltage between one end of the potentiometer and the wiper provides a measure of the position of the wiper's position along the wire coil or film. A typical linear potentiometer's construction is shown in Figure 3.2.24. Rotary potentiometers are made using similar technology. For sensing more than 360^o rotation, rotary potentiometers typically use a screw shaft to create motion of a nut, which is connected to a linear potentiometer's wiper. Rotary potentiometers are available with 10 turn capability. All types of potentiometers are available with virtually any type of mounting bracket.

Potentiometers are generally the least expensive type of linear or rotary motion sensor available, and they have high output voltages, which do not need to be amplified. Their principal drawback is that they rely on mechanical contact in order to function. This leaves them susceptible to dirt and oil contaminating the wiper/film interface, which can change the resistance and cause errors. Various forms of seals can be used to minimize the chance of contamination, but there is still

[20] For a more detailed discussion of the characteristics and applications of thin plastic film piezoelectric materials, see "Piezoelectric Plastics Promise New Sensors," Mach. Des., Oct. 23, 1986, pp. 105–110.

the issue of friction and wear between the wiper and film. To allow for wear of the film, the wiper typically contacts a region that is a small percentage of the total cross section; thus the wiper acts to read location along the film, and any wear of the film by the wiper minimally affects the output.

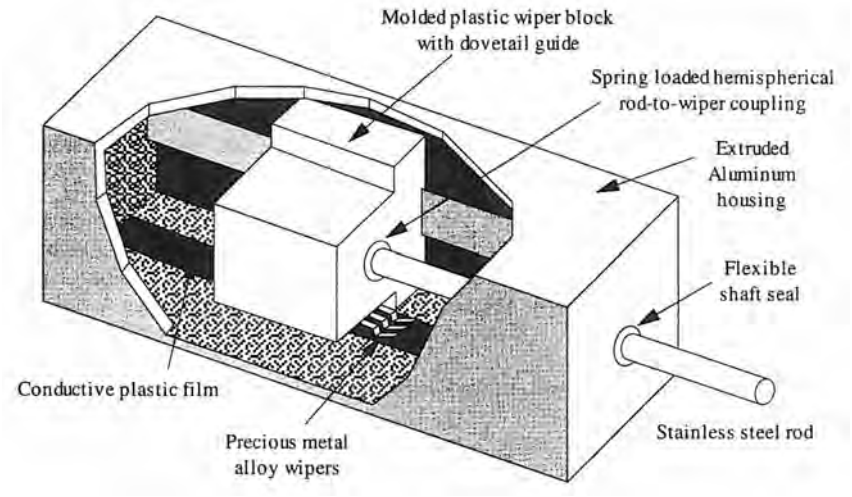


Figure 3.2.24 Construction of a precision linear potentiometer. (Courtesy of Vernitron Corp.)

Modern conductive plastic films are rapidly replacing old-style wire-wound designs. Use of films also allows tailoring of deliberate nonlinearities into the output. Since the film is continuous and potentiometers are analog devices, they can have submicron resolution, although they are still subject to electrical noise and manufacturing nonuniformity. Functionally, their resolution is dependent on a dc power supply and a digital-to-analog converter used to digitize the output signal. Typically, a reliable reading on the order of 1 part in 4000 can be attained. With care, levels of 1 part in 10,000 can be achieved. In addition, systematic nonlinearities can be mapped and compensated for.

Potentiometers have very little overhead associated with their size to sensing range. In fact, one popular type of potentiometer looks like a small, fist-sized box from which a wire protrudes.[21] The wire is connected to a rotary potentiometer and constant torque spring. By routing the wire over pulleys, it is possible to place the body of the sensor away from the mechanism whose motion is being monitored. In many cramped applications, this can be extremely useful, although stretch of the wire can introduce errors and should be considered when choosing this type of sensor. This type of potentiometer can be ordered sealed for underwater use with strokes up to 19 m (750 in.). Cost is on the order of $400 to $50/m stroke. Integral tachometers can also be ordered so that the sensor outputs a position and velocity signal.

Typical Characteristics of Potentiometers

The following summary of potentiometer characteristics are generalizations only. Some manufacturers specialize in making custom units in large quantities, so single-piece prices will be high. Other manufacturers make only a standard line, so prices are low, but the model cannot be modified in any way (e.g., change the size of the thread on the shaft). This general summary should by no means be taken as gospel.

Size: Linear models: Small, diameter = 12 mm, length = stroke plus 38 mm up to to 100 mm stroke. Large, diameter = 64 mm, length = stroke plus 150 mm up to 1.5 m stroke. Rotary models: Small single-turn 12 mm diameter and 25 mm long. 10-turn models may be 25 mm diameter and 64 mm long.

Cost: Small linear model with 0.1% linearity, $490. Large long-stroke-length model with 1/2% linearity, $1000 + $0.78/mm ($20/in.) stroke. Small rotary model $150. Multiturn rotary model $340. Note that inexpensive $10 potentiometers can always be found, but one must be careful to evaluate how long and well they will perform.

[21] Celesco Transducer Products, Canoga Park, CA.

Measuring Range: Linear plunger models to 1 m. String-type models to 19 m. Rotary models to 10 turns.

Accuracy (Linearity): 0.01-0.05% of full-scale output at best depending on range of motion. 1/2–1/4% linearities are more often quoted. With software linearization techniques it does not make sense to spend a lot of money for a highly linear pot. Rotary potentiometers typically are not as accurate as linear motion potentiometers.

Repeatability: Typically, 10 times the accuracy.

Resolution: Ideally, as small as 0.1 μin. on linear models, and 5 arcsecond on rotary models. Actual resolution depends on full-scale voltage and range and resolution of analog-to-digital conversion device. Typically, a potentiometer with a 100-mm stroke and a 12-bit ADC will have a resolution of 25 μm (0.001 in.).

Environmental Effects on Accuracy: Typically, 0.05%/C^{o} of full-scale output.

Life: Linear models 50×10^6 cycles; rotary models to 200×10^6 cycles for 50% degradation of properties.

Frequency Response: Potentiometers act as pure resistors and their response is thus limited by their mechanical construction. Linear models, 1.0-2.5 m/s maximum velocity; rotary models, 100 rpm maximum.

Starting Force and Torque: Linear; 1-4 N. Rotary; 1-10 N-mm.

Allowable Operating Environment: The slower the speed, the longer the life. Potentiometers can be ordered sealed to withstand very adverse conditions (e.g., military specifications MIL-R-39023); however, without protection, wear can rapidly occur. They can be tolerant to fungi, moisture, vacuum, pressure, and immersion (noncorrosive, nonconductive, nonabrasive). Operating temperatures range from -200°C to 125°C.

Shock Resistance: On the order of 50g, depending on the sensor.

Misalignment Tolerance: With couplings or mounting trunions, significant misalignment can be tolerated if planned and designed for. Cosine errors result from misalignment.

Support Electronics: Resolution, accuracy, and repeatability are all directly dependent on the power supply and analog-to-digital converter.

3.2.13 Syncros and Resolvers

Synchros and resolvers have been used since the 1930s as elements of electromechanical servo and shaft angle positioning systems. They are essentially rotary transformers, analogous to RVDTs, but with infinite multiturn capacity. Their principal attribute is good accuracy, low cost, and insensitivity to contaminants. Their principal limitation is that they are analog devices in a world of primarily digital control systems. However, specialized synchro/resolver to digital integrated circuits have evolved to overcome this problem.[22] Note that a resolver is just a special form of a synchro. A synchro is basically a variable transformer in which the magnitude of the electromagnetic coupling between primary and secondary coils, which determines the magnitude of the output voltage, varies with the relative angular position of the coils. The synchro can operate in one of two modes: (1) as a generator of electrical signals in response to shaft rotation, or (2) as a generator of shaft rotation in response to electrical input signals.

These characteristics define two overlapping groups: *control synchros* and *torque synchros,* respectively. Control synchros include transmitters, differentials, control transformers, resolver transmitters, resolver differentials, resolver control transformers, and two hybrid units called transolvers and differential resolvers. Torque synchros include transmitters, differentials, and receivers. Control synchros are designed for use with analog or digital servocontrolled axes, while torque synchros are generally used to transmit rotary position signals that turn dial gages.[23]

Transmitters and Receivers

[22] See, for example, Synchro and Resolver Conversion Handbook, G. S. Boyes (ed.), Analog Devices Inc., Norwood MA. Published by Memory Devices Ltd., Central Ave., East Molesey, Surrey KT8 0SN, England.

[23] See, for example, Analog Components Catalog, Clifton Precision, Litton Systems Inc., Clifton, PA, (215) 622-1000.

Transmitters and receivers are constructed from a single-phase dogbone-shaped rotor that is excited via two slip rings and is electromagnetically coupled to a three-phase stator, as shown schematically in Figure 3.2.25. In order to transmit angular position information, the rotor winding is excited by an ac voltage (60 or 400 Hz) which induces a voltage in the stator windings proportional to the cosine of the angle between the rotor coil axis and the stator coil axis. Accuracy, repeatability, and linearity are all dependent on the quality of the windings. The voltages at each of the stator windings, which are 120° apart, are

$$V_{\text{stator1}-3} = kV_{\text{rotor2}-1} \sin\theta \tag{3.2.4a}$$

$$V_{\text{stator3}-2} = kV_{\text{rotor2}-1} \sin(\theta + \pi/3) \tag{3.2.4b}$$

$$V_{\text{stator2}-1} = kV_{\text{rotor2}-1} \sin(\theta + 2\pi/3) \tag{3.2.4c}$$

k is the coupling transformation ratio characteristic of the synchro, which is also defined as $k = V_{\text{max out}}/V_{\text{in}}$, θ is the rotor position angle, and the subscripts on the voltages refer to respective terminals. In order to use the information from a transmitter, a device must be able to decode the information contained in these equations. Prior to the development of fast, economical, integrated circuit technology, an analog device, the receiver, was used to accept information from the transmitter. With today's technology, the analog signals can be converted to digital format using a synchro-to-digital converter (SDC).

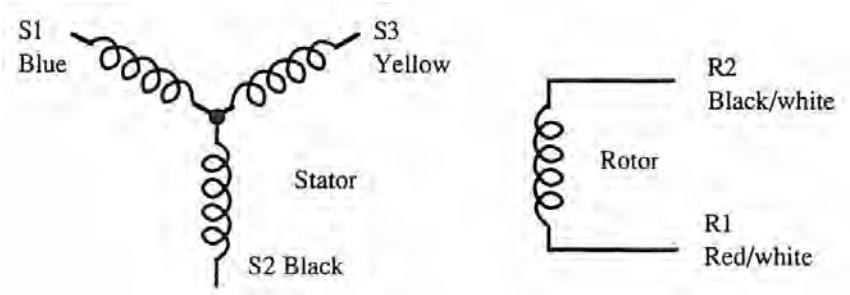


Figure 3.2.25 Synchro transmitter winding configuration.

Torque transmitters and receivers are constructed identically. They are used to form a pure analog electrical system, which is used to transmit angular information from one point to another without mechanical linkages. A torque transmitter is connected electrically to a torque receiver, as shown in Figure 3.2.26. Often, a torque differential transmitter is connected between the latter two, as shown in Figure 3.2.27, such that when positions θ_{CG} and θ_{CD} are dialed in on the transmitter and differential, the receiver shaft assumes a position $\theta_{CR} = \theta_{CG} \pm \theta_{CD}$. Whether addition or subtraction is done depends on the exact configuration of the wiring. It is also possible to replace the receiver with another transmitter such that a torque differential receiver acts as a receiver and its shaft position becomes $\theta_{CD} = \theta_{CD1} \pm \theta_{CD2}$. In a typical application, the torque transmitter shaft is driven by a high helix screw whose nut is a float in an oil or gas tank. The torque transmitter sends a signal to a torque receiver at the control panel, which turns a gage dial. Torque receivers have high internal damping to prevent oscillation of the output shaft.

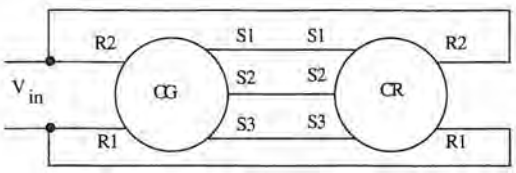


Figure 3.2.26 Coupling a synchro transmitter (CG) to a receiver (CR).

Control transmitters are also mechanically driven and are most often used in conjunction with control transformers and control differential transmitters to electrically add or subtract signals

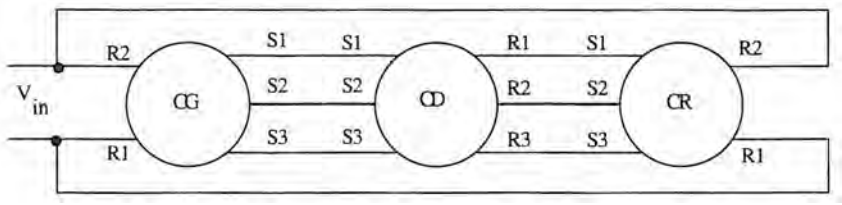


Figure 3.2.27 Coupling a synchro transmitter (CG) and a differential (CD) to a receiver (CR).

before sending them off to be amplified for controlling the position of a motor. Because their signals are to be amplified, control transmitter windings do not have to be as large and powerful as torque transmitters.

Standard analog positioning accuracy for transmitters and receivers is on the order of ± 10 arcminutes. Maximum torque levels are only on the order of 3 g-mm per degree of receiver displacement; the larger the resistance torque, the larger the angular error. This is enough torque to turn a dial; thus amplification and the use of a more powerful motor coupled to the transmitter via a control transformer are required if large torque level is to be generated (e.g., to turn a radar antenna).

Differential

The differential is an element illustrated in Figure 3.2.28 and is used in conjunction with transmitters and receivers as discussed above. It has a three-phase stator and rotor and uses slip rings (brushes) to transmit current to the rotor windings.

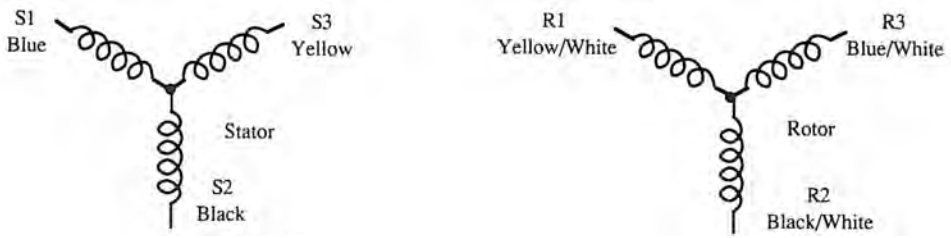


Figure 3.2.28 Synchro differential winding configuration.

Control Transformer

A control transformer has the same stator and rotor coil configuration as a transmitter. With the transmitter the rotor is the primary coil and the stator is the secondary coil, but the opposite is true for the transformer. As shown in Figure 3.2.29, this allows a transmitter to send position information from the angularly positioned rotor over long distances via three wires (noise insensitive) and then have the information converted back to a voltage across two wires to be amplified and used to power a motor.

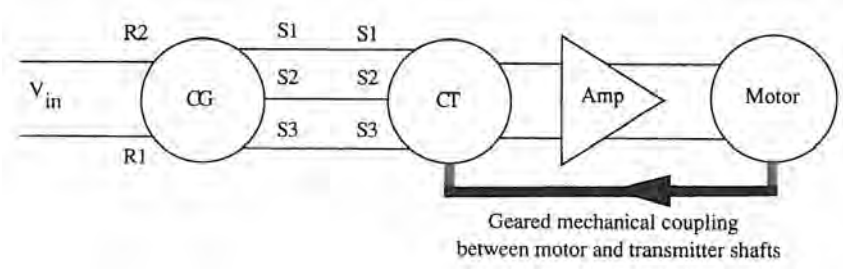


Figure 3.2.29 Coupling a synchro transmitter (CG) to a transformer (CT) and an electric motor.

In order to make the angular motion of the motor shaft equal to the angular motion of the transformer shaft, a mechanical feedback loop is required. As shown in Figure 3.2.29, the motor

shaft will rotate and will, through gearing, also cause the transformer shaft to rotate. The motor will continue to see an amplified voltage until the shaft angle of the transformer equals the shaft angle of the transmitter. As long as the motor sees a voltage, its shaft will continue to turn, also causing the transmitter's shaft to turn. This closed-loop process continues until all the shafts are at their respective desired angular positions.

Transolver and Its Inverse, the Differential Resolver

A transolver is essentially a control transformer with an additional winding in quadrature (displaced 90°) to the main rotor winding, as shown in Figure 3.2.30. It can be used as a transmitter or transformer by shorting the unused rotor winding to ground or dummy loading it with the other winding, respectively. Transolvers are not often used. The inverse of a transolver, however, the differential resolver, is quite frequently used to convert three-wire data from a synchro system into four-wire data of a resolver system. As shown in Figure 3.2.31, the stator has the two windings in quadrature, while the rotor has three windings. This allows the differential resolver to operate with only three slip rings as opposed to four required for a resolver. Hence reliability is increased and cost decreased.

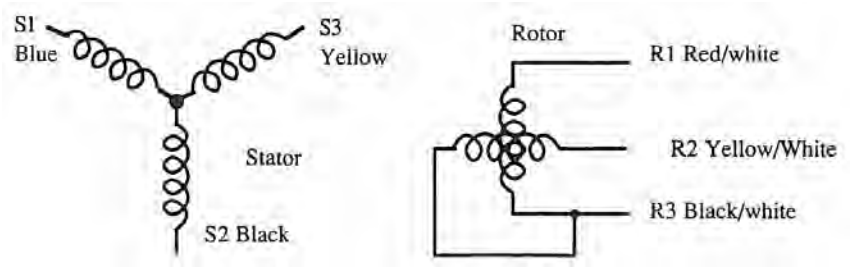


Figure 3.2.30 Transolver winding configuration.

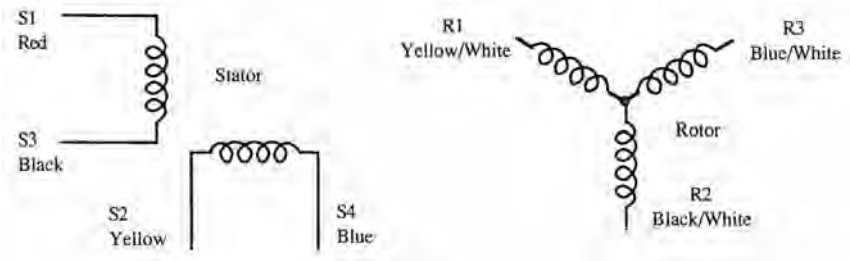


Figure 3.2.31 Differential resolver winding configuration.

Resolvers: Transmitters, Differentials, and Control Transformers

A resolver is a special form of a synchro in which the windings on the stator and rotor are displaced at 90° instead of 120°. They originally evolved to be used as analog devices for computing coordinate transformations in conjunction with inertial guidance systems in missiles and spacecraft. For ultrahigh-reliability space systems they are still used in this way. For applications to machine tools, resolvers generally provide better accuracy and resolution than synchros.

The family of resolvers include resolver transmitters, resolver differentials, and resolver control transformers. Each of these types of resolvers is shown schematically in Figures 3.2.32–3.2.34 and functions in a fashion similar to its synchro counterpart. On most resolvers, two ends of the rotor windings are tied together internally, so only one frequency excitation to the rotor needs to be supplied. A special high-accuracy winding-compensated resolver is available which has integrated temperature and phase compensation built into its windings.

For a two-pole resolver excited by V = A sin ωt, the voltages on the stator windings are

$$S_{13} = kA \sin \omega t \sin \theta \tag{3.2.5a}$$

$$S_{42} = kA \sin \omega t \cos \theta \tag{3.2.5b}$$

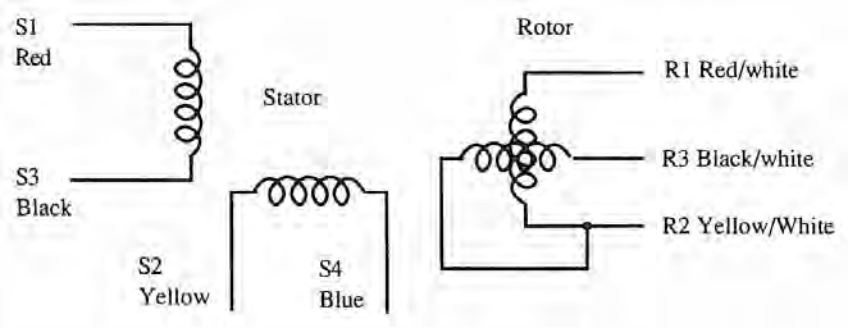


Figure 3.2.32 Resolver transmitter winding configuration.

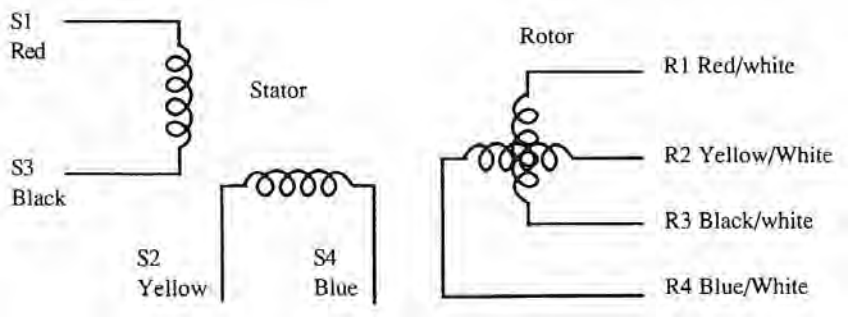


Figure 3.2.33 Resolver differential winding configuration.

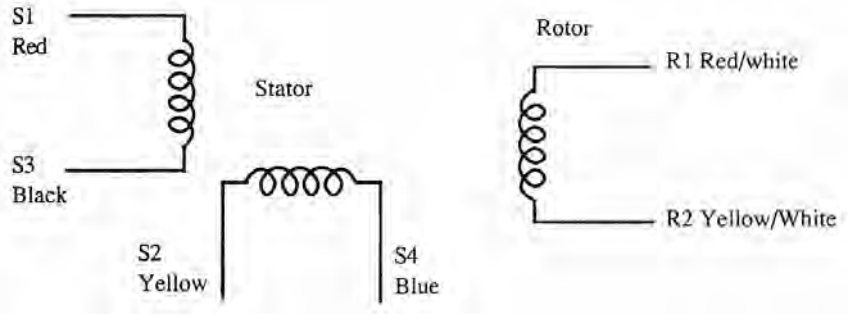


Figure 3.2.34 Resolver control transformer winding configuration.

The value of k is the coupling coefficient, which does not necessarily have to be constant for the resolver since its effect cancels out. The angle of rotation is the inverse tangent of the ratio of the stator windings' voltages. By using additional windings, multispeed resolvers can be electrically geared to give course and fine signals that effectively increase the resolution of the system by a factor equal to the electrical gear ratio. This requires the use of two resolver-to-digital converters and some combinational logic, as discussed below.

Synchro and Resolver-to-Digital Converters

With respect to providing accurate position feedback, synchro-to-digital and resolver-to-digital converters (SDCs and RDCs) are commonly available with 16-bit resolution. This corresponds to about 96 μrad (20 arcseconds resolution). Allowable resolver shaft speeds up to 5000 rpm are also common. With course/fine systems, 21-bit resolution is not uncommon (0.6 arcsecond or 3 μrad).

The simplest and least expensive type of resolver-to-digital converter is the direct dc conversion device shown in Figure 3.2.35. The signals from the resolvers are sampled and held until they can be converted into digital form by analog-to-digital converters. As a result of the inevitable time delay in sampling the two signals, any error in the excitation source or noise spike in the wires will cause the sine or cosine waveform to be distorted. This distortion will be converted to digital format, and thus large errors can be induced when the arctangent function is used to calculate the shaft angle. Since resolvers are often chosen for their ability to survive in harsh environments, noise can be a problem if a direct dc conversion device is used. However, 10-bit devices cost only about $30 and therefore are very useful in systems where a large number of resolvers are used and high accuracy is not required.

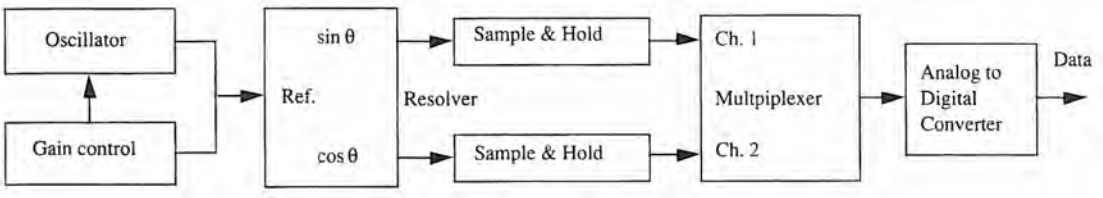


Figure 3.2.35 Direct dc conversion of resolver-to-digital signals.

Phase analog conversion requires a stable oscillator and the resolver to be run backwards with the stator as the primary coils, as shown schematically in Figure 3.2.36. The oscillator generates precise sine and cosine signals which are fed into the stator windings. When the input angle of the signal equals the input angle of the resolver rotor, the output from the resolver goes to zero. At this instant, the zero-crossing detector signals the high-speed counter to note the angle of the oscillator. This system is more prone to errors than the direct dc system, due to errors in accuracy of the oscillator and phase coordination between components, and is not often used.

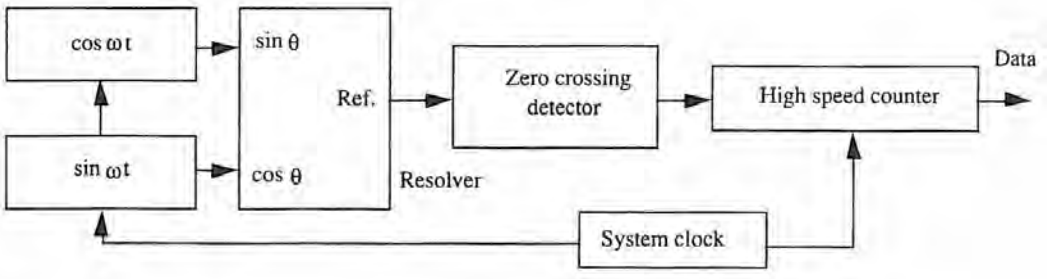


Figure 3.2.36 Phase analog conversion of resolver-to-digital signals.

Tracking resolver-to-digital converters use a simultaneous ratiometric calculation of the resolver outputs to determine the shaft angle while being insensitive to electrical noise. Resolutions to 16 bits are common. Resolutions to 18 bits are available, and 20 bits is on the horizon. As shown

in Figure 3.2.37, tracking converters use special digital-to-analog (D/A) converters which convert a digital angle to sine and cosine voltage waveforms. The sine and cosine waveforms of the digital angle are multiplied by the cosine and sine outputs from the resolver, respectively. The difference between them provides an error signal whose amplitude varies at the same frequency as the resolver excitation source. The resolver excitation source is used as an input to the circuit to demodulate the error signal and provide a dc voltage proportional to $\sin(\theta_{\text{resolver}} - \theta_{\text{digital}})$. The error signal is fed to an integrator and then to a voltage-controlled oscillator which drives an up/down counter. The integrator ensures that the steady-state error will always be zero. The tracking R/D converter thus functions in a differential analog mode to maintain very fast conversion times while canceling out analog noise sources. These devices are available as single-chip devices. Note that a wonderful by-product of the tracking R/D converter is the output from the integrator itself. By reading this signal, rotational velocity can be determined to within 1% accuracy without the danger of noise spikes that often occur when other rotation sensing elements' outputs are differentiated to obtain velocity feedback.

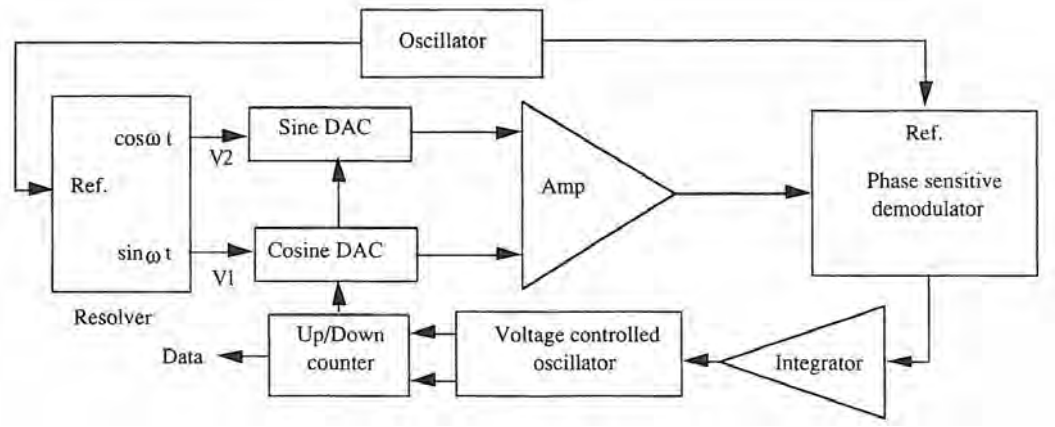


Figure 3.2.37 Tracking resolver-to-digital converter.

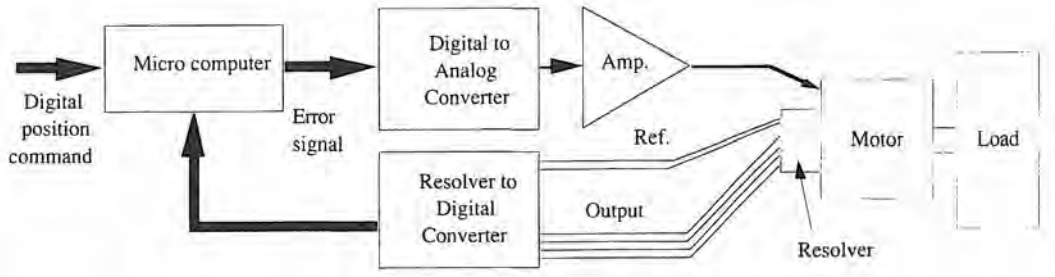


Figure 3.2.38 A typical servo-controlled system with resolver feedback.

A typical closed-loop digital servo system that uses resolver feedback is shown schematically in Figure 3.2.38. Resolver-based sensing systems can operate up to speeds of 100,000 rpm, but the design engineer must realize that resolver-to-digital converter systems have finite settling times. For example, a typical high-accuracy RDC has a gain of essentially 0 dB at 10 rev/s;[24] however, at 20 rev/s, a typical response is about 1 dB. Since $1\text{ dB} = 20\log_{10}(\theta_{\text{inp}}/\theta_{\text{out}})$, this means that there will be a 12% error between the shaft angle and the digital angle. When a machine tool is programmed to move from point A to point B, the motor that drives the axis must also speed up and slow down. For motion parallel to a machine axis this is generally not a problem, for the machine slows way down before finishing a cut. This allows the resolver to settle and the final point to be reached with almost no error. On the other hand, if two axes' motions are being coordinated, the speed may have to be

[24] The speed at which a leadscrew with a 5-mm lead must be driven to move a linear motion carriage at 3 m/min.

curtailed to ensure that the toolpoint moves along the proper path. Usually, tracking errors are due to fundamental physical limitations of the mechanical components and the control algorithms used. Only a small part of the total error is usually due to dynamic limitations of the resolver-to-digital converter. Most RDC manufacturers provide dynamic modeling software that enable the design engineer to determine how the dynamics of the RDC will affect the position error of the system as a function of speed.

Single-stage resolver-to-digital converters can have a resolution of about 16 bits. If greater resolutions and accuracies are required, a multispeed resolver system is required, as shown in in Figure 3.2.39. A mechanically geared resolver is generally impractical to use because of gear backlash and accuracy. An electrically geared resolver has multiple poles which are internally arranged and coupled to provide coarse and fine outputs. They also require the use of two

RDCs and appropriate digital logic to combine the coarse and fine digital words. Electrically geared resolvers are available with up to a 64: 1 "gear ratio" (6 bits). The total resolution in bits of a coarse/fine system will therefore be 6 bits + $\#\text{bits}_{\text{fine RDC}}$, where # is typically 16. Using a 16-bit fine RDC, a resolution of 22 bits or 1.5 μrad is possible, which allows a precision leadscrew system to potentially realize submicroinch resolution.[25] With a coarse/fine system, the course RDC usually also provides a tap for obtaining an analog voltage out that is proportional to the shaft speed; thus it also doubles as a tachometer with a linearity of about 1%. Multispeed resolvers have a host of problems associated with their operation, including a decreased magnetic coupling factor, which requires the windings to have more turns in order to achieve acceptable transformation ratios; and increased phase shift of dynamic response (not angular shaft degrees), on the order of 10° for 16 speed units, 25° for 36 speed units, and 40° for 64 speed units. The design engineers of the control algorithms and the machine must consider this to make sure that any system which uses them will remain controllable.

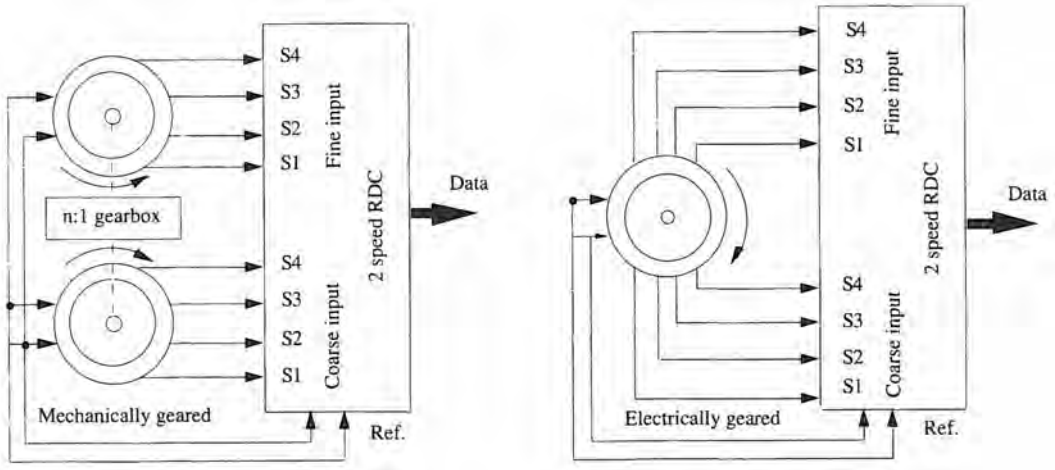


Figure 3.2.39 Mechanically and electrically geared coarse/fine resolver systems. (After Boyes.)

Applications

One problem that may be anticipated with synchros or resolvers is wear of the brushes used to transmit the excitation voltages to the rotor. In order to increase life and operating speed and decrease susceptibility to contamination, brushless resolvers evolved which use a circular transformer to transmit the excitation voltage to the rotor. This makes them essentially insensitive to virtually any form of contamination; hence many machine tools use brushless resolvers to measure the rotation of motor shafts that drive leadscrews. Brushless resolvers are often favored over optical rotary encoders because of the former's greater resistance to hostile environments.

[25] There are numerous problems associated with mechanically achieving submicroinch accuracy, as is the subject of a good deal of this book. With the availability of modern laser interferometers one would probably never use an electrically geared resolver on a leadscrew.

Typical Characteristics of Resolvers

The following summary of resolver characteristics are generalizations only. There are numerous manufacturers of resolvers and synchros, so the user should be very careful to make sure to find exactly the unit that is desired. This general summary should by no means be taken as gospel.

Size: Sized in tenths of an inch: for example, a synchro with a 1.5-in.-diameter would be a size 15, with common sizes being 08, 11, 15, 18, and 23. Length is on the order of 1.5 times the diameter. Large-diameter *slab* or *pancake* units may be 6 in. or more in diameter and 2 in. thick. A large diameter allows numerous windings to be used to increase their electrical gear ratio.

Cost: $50-$10,000.

Measuring Range: Unlimited rotation.

Accuracy (Linearity): 7 arcminutes is standard. Precision electrically geared pancake resolvers can have arcsecond accuracies.

Repeatability: Typically, two to five times the accuracy.

Resolution: From sub arcseconds to several arcminutes. These are analog devices, so resolution is dependent on the performance of the electronic components used. Highest resolution would be from a 6-in.-diameter 64-speed resolver (which may cost $10,000).

Environmental Effects on Repeatability: If a tracking RDC is used, most effects on the resolver or synchro will cancel each other out. However, whenever high precision is required, winding compensated units should be specified.

Frequency Response: The limiting frequency response factor is usually the resolver-to-digital converter. For 14-bit resolution, a hybrid RDC's response is 0 dB to about 10 Hz (600 rpm). For a 16-bit hybrid RDC, response is 0 dB to about 5 Hz (300 rpm). A new monolithic 16-bit RDC's 0 dB response is also about 20 Hz.

Starting Torque: On the order of 5 g-mm (0.007 oz-in.) to overcome bearing and seal friction.

Allowable Operating Environment: Brushless resolvers and synchros can operate almost anywhere. They were originally designed for use in military combat equipment, so they have a long history of being able to resist harsh conditions.

Shock Resistance: On the order of 10-50g.

Misalignment Tolerance: Shaft misalignment can introduce cyclic errors into the system. For ultraprecision systems, the resolver can be purchased to mount directly on an extension of the motor shaft or leadscrew. The latter is often desirable in order to thermally isolate the resolver from the motor.

Support Electronics: Requires a resolver-to-digital converter, 400-Hz or 2.6-kHz 26-V excitation source, $\pm$15-V dc and 5-V dc precision power supplies.

3.2.14 Ultrasonic Sensors

Ultrasonic sensors use a stress wave generating/receiving element to create a pressure wave in a medium and measure the amplitude and return time of echoes. Measuring the time it takes the echoes to return gives a measure of the distance of inclusions or features within the object. There are two popular types of ultrasonic transducers used in manufacturing environments: piezoelectric and electrostatic. Piezoelectric transducers are widely used in nondestructive evaluation of materials. They must be mechanically or fluidically coupled to the part since the amplitude of their waves is very small. They can, however, generate frequencies in the 100-kHz range. Piezoelectric transducers are finding increased use in manufacturing applications as sensors for determining the thickness and surface finish of parts (e.g., shells and thin-walled parts held in vacuum chucks) while the part is still held in its fixture on the machine. Resolution on the order of 10 to 12 bits[26] is possible. They are also widely used to measure thickness and uniformity of layers in built-up composite structures, and to search for voids. Cost of these types of systems is on the order of $5000.

[26] From conversations with Dr. Gerry Blessing at the National Institute of Standards and Technology.

Electrostatic or capacitive transducers apply a voltage to a metallized film to generate ultrasound. Actually, they use two conductive materials, one the metallized film and the other a metallic backing dish. The film itself is the dielectric barrier. An oscillating voltage is applied, causing the dish and metallized film to be alternately attracted and repelled by the electric field. The dish and film have very low inertias and can respond quickly to a driving signal. They couple very well to air and have the large displacement amplitude needed to couple well with low-density media such as gases. With various signal processing techniques, electronic transducers can be used to provide accurate dimensional measurements of metal, wood, rubber, glass, plastic, and so on. An example of this type of sensor is Polaroid Corporation's acoustic range-finding sensor. It is accurate to about 1 part per 100 to a range of about 10 m and is finding great popularity among mobile robot design engineers. This sensor costs about $150 in small quantities, and is about 40 mm in diameter and 15 mm thick. For use in hostile environments, an additional ruggedized housing is required.

Greater resolution and accuracy at close range can be realized using more complex systems that operate at higher frequencies. A single-channel version of this type of black box costing about $5000 uses a thumb-sized sensor which costs about $250. With a 10-cm (4 in.) focal length and 5-cm (2-in.) range of sensing motion, resolution is about $\pm 0.1\%$ (50 μm). The speed of sound changes in air by about $0.15\%/C^{o}$. To compensate for changes in air temperature, pressure, humidity, and so on, which can greatly affect accuracy, the black box has a special calibration port to which a reference sensor can be connected. In regions of high air turbulence, however, accuracy degrades to about $\pm 1\%$. Measurement update time can be up to about 500 times per second, making the black box suitable for use in closed-loop control systems. Typical applications include sensing liquid levels in bottles on a production line and sensing the distance to a surface that cannot be touched to prevent contamination.

3.2.15 Velocity Sensors

Closed-loop feedback position and velocity control systems often require velocity feedback in order to obtain a clean low-noise velocity signal. Differentiating a position signal to obtain a velocity signal often leads to noise spikes, which can be detrimental to a control system's performance unless the position sensor's resolution is about 10 times greater than the required mechanical resolution. Hence a sensor to measure velocity is often required for a high-performance servoloop. Analog linear and rotary velocity sensors operate on the principle that a moving coil (or magnet) in the presence of a stationary permanent magnet (or coil) will induce a voltage in the coil proportional to the relative speed of the coil and magnet: Faraday's law of induction says that voltage induced in a coil is proportional to the rate of change of magnetic flux caused by a moving magnet. Because the flux in the magnet is constant, the rate of change of flux with respect to the coil is a function of the relative velocity between the magnet and the coil.

Linear Velocity Transducers

A linear velocity transducer (LVT) is constructed in a manner similar to that of an LVDT except that an excitation voltage is not used. If both poles of the magnetic core entered a single coil, opposing magnetic poles would cancel any voltage induced in the coil. Using two coils whose voltages are summed differentially eliminates this problem. The polarity of the total output voltage depends on the direction of motion of the magnet core. An LVT can be modeled as a pure inductance and a resistor, and assuming that it is connected to a very high impedance meter or ADC, it will behave as a first-order system. Thus the time constant of its response is just $\tau = 2\pi L/R$. After 7 time constants, $e^{-t/\tau} = e^{-7}$ or 0.1% of the final value has been reached. A typical time constant is on the order of 0.001 s. At higher velocities and accelerations, the response may lag behind the input, thereby decreasing the linearity. Take note that a coil of wire acts like an AM antenna and thus a great introducer of noise into the system. LVDTs do not have this problem because they use an ac excitation source.

Tachometers

The rotary form of a velocity transducer is known as a *tachometer* and includes permanent magnet stator dc tachometers, drag-torque tachometers, capacitor tachometers, digital tachometers and dc brushless tachometers. The most common form is the dc tachometer. Note that for accurate

speed and position control of electric motor driven systems, the use of angular velocity feedback is often essential to obtain proper damping.

Permanent-magnet stator dc tachometers have a wound rotor (armature) and commutator assembly. Rotation of the armature causes the windings to pass through the stator's magnetic field, thereby inducing a voltage in the windings proportional to the rotation rate. The voltage is passed through the commutator to the output terminals. Dc tachometers are thus essentially the analog opposites of dc motors. The rotor contains a wound coil, and the unit spins inside a magnetic housing. Brushes and a slip ring are used to transmit the generated voltage (which is proportional to the velocity) to the outside world. Brushes are typically rated for 100,000 h of continuous use, and models are available to handle up to 12,000 rpm. Ideally, a dc tachometer should provide a pure dc voltage output proportional to rotation rate; however, the discrete number of armature coils (poles) leads to a pulsating output, known as ripple, that is a percentage of the voltage being output. To reduce ripple, filtering techniques and greater numbers of coils can be specified. The former can have detrimental time-delay effects on precision servoloops, while the latter increase cost.

Other sources of error include contamination of the commutator interface, temperature changes, and magnetic fields. To avoid effects of the former, it is important to specify a tach that is designed for use in the type of environment it will be operating in. Trying to save a few dollars on an unsealed model can cause problems in the long run. Stray magnetic fields and nearby iron can also affect the output of a dc tach, so the tach should be isolated from such sources by an air gap of a centimeter or two, or provisions should be made for recalibration.

Drag-torque tachometers consist of a magnet that rotates inside a cup. A torque on the cup proportional to velocity is produced by eddy currents induced by the rotating magnet. A spring attached to the cup and a potentiometer shaft provide means for limiting and measuring the rotation of the cup. This type of tachometer is not very common because it is prone to large thermally induced errors which often occur in the neighborhood of an electric motor. This design, however, was the basis for many dial-type automobile speedometers and tachometers.

Digital tachometers essentially differentiate the output from an optical encoder and are prone to noise spikes if the differentiation is not executed with care. However, they are not affected by electrical noise.

Brushless tachometers are essentially inside-out versions of brushed tachometers. They have a magnetic rotor and wound stator. In order to generate a uniform dc signal proportional to shaft speed, however, a means for switching the different windings on the coil is required. This adds cost and increases complexity, but minimizes noise associated with running brushed models at low rpm and also eliminates brush maintenance. However, just as motors have torque ripple, dc or brushless tachometers will have velocity ripple that is prominent at low speeds.

Typical Characteristics of dc Tachometers

The following summary of dc tachometer characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as gospel.

Size: Commonly used industrial tachometers are from 19 mm in diameter and 40 mm long to a 65-mm^2 mounting plate 75 mm long. Unhoused versions are also available that mount on an existing shaft. High-performance tachometers for sensing almost quasi-steady motion may have much larger diameters and shorter lengths.

Cost: Brushless models on the order of $100 plus $100 to convert sine-wave output to dc signal. Brushed models range from $150 to $250 and provide a dc output. High-performance tachometers may cost $1000 or more.

Accuracy (Linearity): Above a few hundred rpm, linearity is typically on the order of 0.1%.

Repeatability: About five times the accuracy.

Resolution: About 0.01% of full-scale voltage. At low speeds, brushed models tend to generate more ripple in their output voltage unless they are specially constructed.

Environmental Effects on Accuracy: Most quality units have temperature-compensated windings, so the effect of temperature on accuracy is on the order of 0.01%/C^o change from 20^{o}C.

Life: Brushed models, on the order of 100,000 h at 3600 rpm. Brushless models' life is limited by bearing life.

Frequency Response (3 dB): On the order of 200 Hz at 0.1% accuracy and 450 Hz at 1% accuracy.

Driving Torque: On the order of 3.5-7 N-mm (0.50-1.0 oz-in.).

Allowable Operating Environment: They can be designed for use in almost any environment, but performance will be affected by temperature.

Shock and Vibration Resistance: 10-20g.

Misalignment Tolerance: Use of a flexible coupling to compensate for misalignment decreases dynamic performance. Best performance obtained through use of hollow-center models that can be mounted directly over the motor shaft.

Support Electronics: None required for brushed models other than an analog-to-digital converter. Brushless models require a $100 electronic black box.

Chapter 4
Optical Sensor Systems

For the sake of persons of different types, scientific truth should be presented in different forms, and should be regarded as equally scientific, whether it appears in the robust form and the vivid coloring of a physical illustration, or in the tenuity and paleness of a symbolic expression.

James Clerk Maxwell

4.1 INTRODUCTION

All sensor systems have to be traceable to a standard, and perhaps no reference can be made more stable than the wavelength of light. Hence it is of prime importance that machine designers and manufacturing engineers be aware of the operating principles of optical sensor systems[1].

Most optical sensor systems operate using intensity, interference, or time-of-flight measurements. Sensors based on the intensity of reflected light are principally analog devices that, for example, either generate a signal proportional to distance, or an on/off voltage as in the case of an optical limit switch. Issues of accuracy, repeatability, and resolution are similar to those for nonoptical sensors. Sensors whose operation is based on the phenomena of optical interference generally have very high accuracy (parts per million) and very high bandwidth. Output linearity is based on stability of the speed of the light and is therefore affected primarily by environmental factors. Sensors whose operation is based on time of flight (e.g., radar and some surveying instruments) can have exceptional range. In general, optical sensors are far more accurate and have greater frequency response than nonoptical sensors.

There are many variations on optical sensor types, too many to discuss here; hence only those most commonly used in conjunction with precision manufacturing processes will be discussed:

- Autocollimators
- Optical encoders
- Fiberoptic sensors
- Interferometric sensors
- Laser triangulation sensors
- Photoelectric transducers
- Time-of-flight sensors
- Vision systems

It is extremely important to note that accuracy and often repeatability are highly dependent on the manner in which the sensor is mounted. Therefore, Chapter 5 should be consulted for a detailed discussion of sensor mounting methods.

4.2 AUTOCOLLIMATORS

Autocollimators are devices for precise measurement of small rotations around axes orthogonal to an optical sighting axis. Autocollimators are more often used as inspection devices than as integral parts of a sensor system for servo-controlled machines. Both manual and electronic autocollimators are available, although the latter are most widely used today. An autocollimator is actually the marriage of a collimator to a telescope. A collimator takes diverging light (e.g., light from a bulb) and focuses it into a nondiverging column of light (i.e., focus the light at infinity). A telescope, on the other hand, takes light from a source at infinity and focuses it onto a point; thus when the angle of incidence of the light from infinity on the telescope changes, the position of the focused image on the focal plane of the telescope also changes. Because one focus of the telescope is at infinity, the axial position of the target mirror does not affect the position of the focused image. An autocollimating telescope is an instrument that combines a collimator and telescope into a single unit.

The components of a typical electronic autocollimator system are shown in Figure 4.2.1. Light from the the illumination source is focused by the condenser optics and projected onto the collimator reticle. The collimator reticle projects a desired pattern (e.g., a crosshair) via a beamsplitter. The rays from the collimating reticle are turned 90^{o} toward the objective lens, which takes the image

[1] A useful book for engineers to read is: W. Welford, Useful Optics, The University of Chicago Press, Chicago, 1991.

and projects it in a nondiverging manner (focus at infinity) at the target mirror. The target mirror reflects the image back through the objective lens to the beamsplitter.

The beamsplitter is actually a semireflecting surface which allows 50% of the light to pass through to an eyepiece reticle. The eyepiece reticle is simply a scaling device which allows for rapid manual assessment of the motion of the image related to angular motion of the target mirror. The remaining 50% of the light is projected onto a photodiode. Output from the photodiode can then be recorded electronically for analysis or used as a feedback signal for some servo systems that control small angular motions.

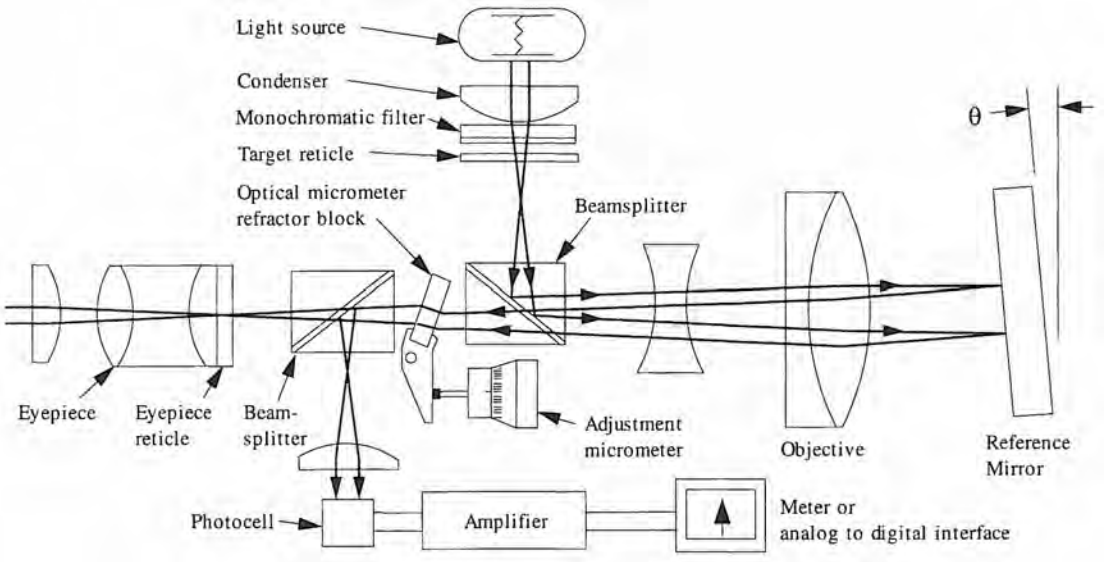


Figure 4.2.1 Schematic drawing of an autocollimator and target. (Courtesy of Rank Taylor Hobson.)

Autocollimators provide a fast, simple method to measure straightness or flatness of a surface. In order to measure the straightness of a surface, such as a bearing way or surface plate, a mirror mounted to a sled is incrementally moved along a straight path (linear or crisscross). At each incremental stop, the autocollimator is used to measure the slope from which elevations can then be derived. Care must be taken to use uniform step sizes of the appropriate length; and as discussed by Bryan,[2] the selection of step size can be affected by the frequency of the straightness error. Properly used, an autocollimator can check the straightness or flatness of a surface to the 1/2–1/4 μm level in an order of magnitude less time than would be required to set up a laser interferometer and a straightedge. The relation between the lateral displacement d, the step length L, and the change in angle θ is just $\delta = L \tan \theta$. However, this does not help to plan the measurement experiment.

Virtually any surface can be represented using a Fourier series; thus assume that a surface has an elevation $x = A \sin(2\pi x/\lambda)$. It can be shown that the elevation (straightness) at any step number N determined by measurement with an autocollimator is

$$\delta_N = L_{step} A \left\{ \frac{\sin\left[2\pi\left(\frac{NL_{step}L_{sled}}{\lambda}\right)\right] - \sin\left[\frac{2\pi NL_{step}}{\lambda}\right]}{L_{sled}} \right\} + \delta_{N-1} \tag{4.2.1}$$

A short step length is desirable if a high-resolution map of the surface is to be obtained. A long sled is desired to minimize the effect of variations in contact height between the sled feet and the surface. The step length must be less than the sled length, and if the ratio of λ/L_{sled} is less than about 20, then the sled length must equal the step length (which is not a bad rule of thumb to follow anyway) in order to have theoretically zero error. In addition, the ratio of λ/L_{sled} should be greater than about 5. Assuming that the proper step size and sled length are chosen, the error in the calculation of straightness is a function of the knowledge of the step size L_{step} between measurements. Since θ is

[2] J. B. Bryan, "The Abbe Principle Revisited: An Updated Interpretation," Precis. Eng., Vol. 1, No. 3, 1979, pp. 129–132.

small, small-angle approximations apply, and the straightness measurement error is $\Delta\delta = \theta\,\Delta L_{step}$. If θ were, for example, 0.01 rad (0.5730°) and ΔL_{step} = 1/4 mm (0.010 in.), then $\Delta\delta$ = 2.5 μm (100 μin.). More typically, however, ΔL_{step} = 25 μm (0.001 in.), θ = 100 μradians, and $\Delta\delta$ = 0.0025 μm (0.01μ in.). Note that the surface and the sled feet must be very clean, and the area of the sled feet must be much greater than any local depressions (e.g., scrape marks). Even though elevation has to be obtained by a process of slope integration, it is often simpler to implement than direct measurement of straightness by interferometry.

With respect to the measurement of a moving carriage, one should note that the autocollimator will also measure the angular motion caused by the moving carriage's bearings. In addition, the spacing of the bearings on the carriage and its relation to the frequency of the straightness error also confuses the measurement. Hence although some autocollimator manufacturers advertise that their autocollimators can be used to measure straightness and orthogonality of moving axes, the measurements obtained will often not be reliable to the highest order. On the other hand, autocollimators can provide a means for evaluating the angular errors in a linear axis, and then this information could be used by the machine tool controller to calculate Abbe errors and compensate for them. Autocollimators are widely used to aid in testing the accuracy of rotary indexing tables and to determine the curvature of precision optics.

Operating Considerations

The electronic autocollimator relies on the use of an analog device, a photodiode, and precision optics to determine change in angular position of a target mirror. Accordingly, there are several factors which affect the resolution or sensitivity of an autocollimator, including photodiode type, electrical noise, bandwidth, measuring range, mirror distance, and objective lens focal length.

An autocollimator's resolution is affected by defocusing resulting from beam divergence as the target mirror moves farther and farther away. In order to alleviate electrical noise problems and achieve very high resolutions, more electrical filtering is required. This decreases the bandwidth, which corresponds to the allowable rate of change of the angle being measured. For most inspection tasks, this is not a problem; however, when using an autocollimator to measure vibration, bandwidth limitations can become important. Typical values for resolution versus bandwidth are shown in Table 4.2.1.[3] The product of the desired sensitivity and the electronic resolution (i.e., the resolution of an analog-to-digital converter and the amount of noise in the system) governs the allowable measuring range. Typical electronic resolution is 1 part per 1000 (10 bits), although some models are capable of 12- to 16-bit resolution. Thermal growth errors within the instrument itself can affect angular accuracy and electrical noise generation in the photodiode. These errors can be minimized and resolution and accuracy increased with the use of an external light source.

Bandwidth (Hz)	Microradians	Arcseconds	Hz/mrad
0.1 or less	0.005	0.001	–
1	0.05	0.01	20
5	0.1	0.02	50
30	0.5	0.1	60
100	1.2	0.25	83
300	3	0.6	100
1000	5	1	200

Table 4.2.1 Typical effect of bandwidth on Autocollimator resolution. (After Thurston.)

The distance of the mirror from the autocollimator is of concern because it affects the behavior of the beam of light in several ways. First, some autocollimators use a photodiode whose output is sensitive to intensity as well as position of a light beam. However, as discussed in Section 4.7, it is possible to make the output of the diode insensitive to variations in light intensity by using a ratiometric comparison of outputs from a lateral effect diode. The farther the target mirror is from the objective lens, the greater the distortion (caused by atmospheric turbulence) of the ideally circular beam of light. As a result, the center of intensity of the beam can shift from its ideal position within the beam shape; consequently, the lateral effect diode, whose output is proportional to the location of the center of intensity, registers an apparent angular change in position of the target mirror.

[3] From T. Thurston, "Specifying Electronic Autocollimators," Proc. SPIE Opt. Test. and Metrol., 1986, pp. 399–401.

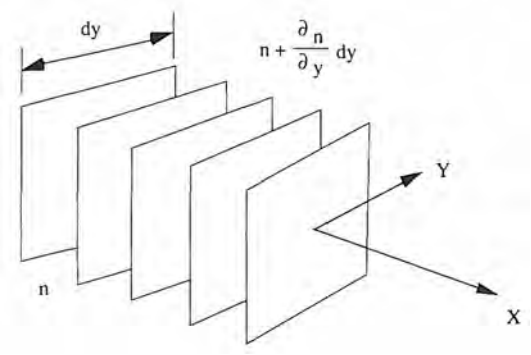


Figure 4.2.2 Effect of an index of refraction gradient (e.g., caused by a temperature gradient) on the propagation of a plane of light.

Temperature gradients cause changes in the refractive index of air, bending the light and causing an apparent angular motion of the target mirror. For example, as shown in Figure 4.2.2, if light is envisioned as a series of planes orthogonal to the optical path with one edge of the plane being dragged through molasses (air of different temperature), the planes will twist and veer off to one side. The index of refraction, n, is defined as the ratio of the speed of light in a vacuum, c, to the speed of light, v, in a medium:

$$n = \frac{c}{v} \tag{4.2.2}$$

The index of refraction is proportional to the composition and density of a medium[4]:

$$(n-1) \times 10^7 = (n_{nominal} - 1) \times 10^7 \times \frac{\text{Pressure}}{760\ \text{mm}} \times \frac{293^\circ\text{K}}{T(^\circ\text{K})} \tag{4.2.3}$$

where $(n - 1)_{nom} \times 10^7$ is about 2808 for a helium neon laser in air at standard temperature and pressure.[5] During a time interval dt, the left edge of the plane of light travels dx = c dt/n while the right side of the plane travels a distance dx':

$$dx' = \frac{c\,dt}{n + \frac{\partial n}{\partial y}\,dy} \tag{4.2.4}$$

The change in angle dθ of the plane of light is

$$d\theta = \frac{dx' - dx}{dy} \tag{4.2.5}$$

Substituting for dx, dx', and dt = n dx/c into Equation 4.2.5, the change in angle with respect to the length of the optical path is found:

$$\frac{d\theta}{dx} = \frac{-\frac{\partial n}{\partial y}}{n + \frac{\partial n}{\partial y}\,dy} \tag{4.2.6}$$

Differentiating Equation 4.2.3 with respect to the gradient direction y yields

$$\frac{dn}{dy} = \frac{-K\,dT}{T^2\,dy} \tag{4.2.7}$$

where:

$$K = (n_{nominal} - 1) \times \frac{\text{Pressure (mm)Hg}}{760\ \text{mm}} \times 293^\circ\text{K} \tag{4.2.8}$$

[4] The term (n - 1)*10^7 is used because air's refractive index n would equal 1.0002808, which is more cumbersome to write than 2808. For a more detailed expression, see Equation 4.5.28.

[5] See Handbook of Chemistry and Physics for tables of the refractive index of various wavelengths of light in different media.

Noting that n = 1 + K/T, the change in angle as a function of optical path distance caused by a transverse temperature gradient dT/dy (orthogonal to the beam path) is

$$\frac{d\theta}{dx} = \frac{\frac{K}{T^2}\frac{dT}{dy}}{1 + \frac{K}{T} - \frac{K}{T^2}\frac{dT}{dy}dy} \tag{4.2.9}$$

For small gradients dT/dy in air

$$\frac{d\theta}{dx} \approx \frac{K}{nT^2}\left(\frac{dT}{dy}\right) \tag{4.2.10}$$

This error can occur in horizontal or vertical angle measurement. For a typical gradient of 1 C°/m, dθ/dx = 1.0 μrad/m. When measuring straightness of a bed of length L, the resultant Abbe error for the straightness measurement caused by this error will be $\Delta\delta = L d\theta/dx$. For a l m bed, the error may be l μm, which is significant. Fortunately, air tends to stratify so 1 C°/m gradients orthogonal to the beam path are rare. More often, turbulent, nonuniform temperature air causes fluttering of the autocollimator's output. This type of error, however, can be averaged out if data are taken for a period that is several times as long as the characteristic period of turbulent motion. In precision applications these assumptions need to be verified by monitoring the measurements. Furthermore, as in the case of using an autocollimator to align large, operating machinery in a factory environment, one must be extremely careful of temperature gradients.

Two other factors which affect the measuring range and resolution of an autocollimator are the diameter and distance of the target mirror from the objective lens, and the focal length and diameter of the objective lens. The former affects the measuring range but does not directly affect the value of the angle measured by the autocollimator. It is a property of the objective lens that light incident at an angle on the lens, regardless of its source, is displaced from the principal axis along the focal plane; so it is only important to prevent the target mirror from moving so far away from the objective lens that the reflected light misses the objective lens. Of course, atmospheric effects on the beam between the target mirror and objective lens must still be considered. As the focal length of the objective lens increases, the "optical lever arm" also increases and a given angular motion of the target mirror will cause a larger lateral displacement on the photodiode which increases the resolution but decreases the range of detectable motion.

Typical Characteristics of Autocollimators

The following summary of autocollimators' characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as "gospel." Furthermore, it is extremely important to note that accuracy is highly dependent on the manner in which the optics are mounted and how the environment is controlled.

Size: From units that can be held in the palm of your hand, to units as big as your forearm.

Cost: In the range of $2000-$5000 for the autocollimator. Signal conditioning electronics can run from $500 to $4000.

Measuring range: From arcseconds to a few degrees.

Accuracy (Linearity): On the order of 0.1-0.05% of full-scale range, depending on environmental conditions. Higher accuracies can be achieved by mapping.

Repeatability: Dependent on environmental conditions, but can be on the order of two to five times better than the accuracy.

Resolution: Typically as small as 0.1 arcsecond but models with 0.001 arcsecond resolution are available.

Environmental Effects on Accuracy: On the order of 1 μrad/m for a 1 C° gradient and l μrad/C° for the instrument itself.

Life: The sensor is noncontact, so life is limited only by the associated electronics.

Frequency Response: See Table 4.2.1 for typical dynamic responses.

Starting Force: The effect of the inertia of the target mirror on system dynamic performance needs to be considered.

Allowable Operating Environment: To maintain accuracy, the system should ideally be used at 20°C with no gradients.

Shock Resistance: Autocollimaters for military applications typically can withstand 80g shock. Precision models for aligning machinery should not be bumped around.

Misalignment Tolerance: Misalignment results in a cosine error. For measuring straightness it is important that the step distance be kept constant and matched to the pitch of the contact areas.

Support Electronics: Typically a black box signal conditioner for the photodiode, and a display or computer interface.

4.3 OPTICAL ENCODERS

Optical encoders typically operate on the principle of counting scale lines (slits) with the use of a light source and a photodiode. They can be configured to measure angular rotation or linear motion. Most optical encoders produce a digital output based on counts of scale lines and they are usually immune to electrical noise encountered by synchros, resolvers, and RVDTS. Some optical encoders also produce analog sine and cosine wave output and they may be susceptible to electrical noise. However, unlike synchros and resolvers, optical encoders are extremely sensitive to dust, dirt, and fluid contamination and therefore must be very carefully sealed or used in a clean environment. Because of their sensitivity to contamination, optical encoders were once considered delicate devices with questionable reliability; however, years of design evolution have produced designs that will withstand virtually any type of operating environment. Figure 4.3.1 schematically shows how a typical incremental rotary optical encoder is constructed. Output from an encoder can be absolute or incremental. Absolute encoders typically provide 10–12 bits of resolution, but 16 bit units are available. Incremental encoders typically provide 10- to 16-bit resolution, and ultra-high-resolution encoders can provide 21 bits of resolution (3 μrad).

High resolution can also be obtained by gearing together two low-resolution optical encoders to obtain a coarse-fine system in a manner similar to that used with some resolver systems. Although gear backlash and nonlinearities are often worse than the resolution of a single high-resolution system, multispeed systems were often the only alternative when high-resolution encoders were not available. Also, in extreme environments where a single high-resolution encoder may be damaged by vibration or shock, precision geared systems with rugged low resolution encoders can still survive.

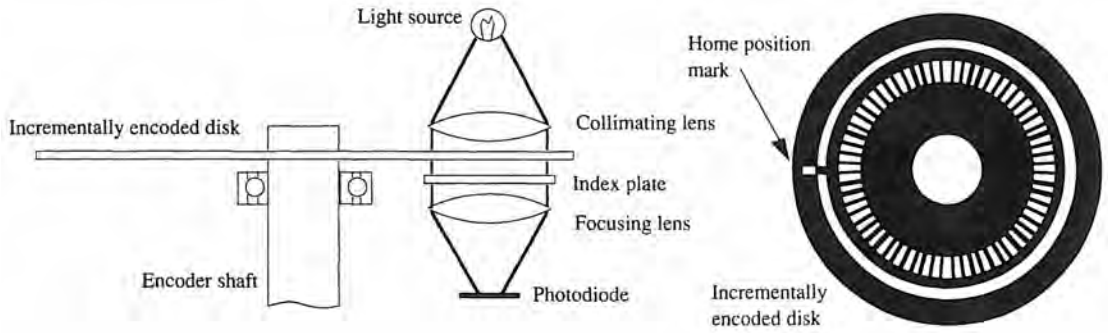


Figure 4.3.1 Construction of an incremental optical encoder.

4.3.1 Incremental Position Encoders

A typical incremental encoder is constructed as shown schematically in Figure 4.3.1. The input shaft can be made integral with the device whose rotation is being measured, or the encoder shaft can be connected to the device by means of a coupling. The encoder disk is divided into alternately optically opaque and transparent sectors. A mask between the light source and the disk allows for many windows of light to be turned on and off simultaneously; thus local errors in line spacing are averaged out and average light intensity is increased. Light from an LED on one side of the disk is collimated. After the light passes through the disk and the mask, it is focused by a lens onto a

photodiode. As the disk rotates, a photodiode on the other side of the disk senses pulses of light passing through the transparent slits in the disk and the mask. The mask serves to average individual line spacing errors. The photodiode's output is electronically conditioned to produce square-wave pulses whose sum at any time indicates angular position of the disk and shaft. A second diode out of phase with the first is sometimes used to cancel dc noise. Position information (the number of counts) is stored in an external counter. If power to the system is lost, the position of the shaft is lost as well. This is one drawback of incremental encoders which led to the development of absolute encoders. When using an incremental encoder, to determine shaft position upon startup, a single slit (home bit) is needed on the encoder, or an external limit switch must be used.

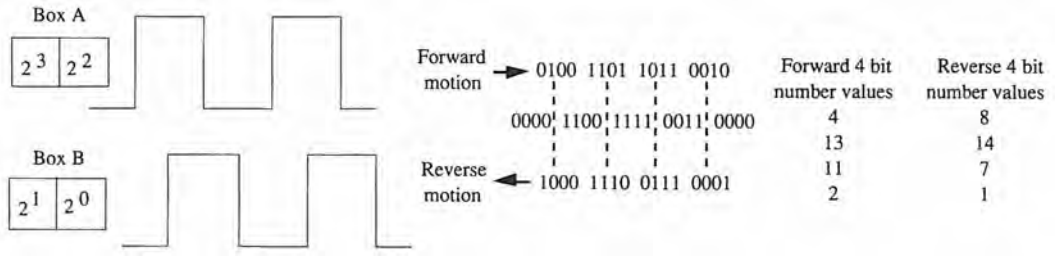


Figure 4.3.2 Using 90° out-of-phase square waves to detect direction and increase resolution by a factor of 4 using quadrature logic.

Output resolution of an encoder is therefore primarily a function of the number of opaque and transparent sectors and typically is between 100 and 100,000 counts for l complete revolution. Most incremental encoders have a viewing window, light source, and detector(s) whose output is 90° out of phase with respect to the first set; this type of encoder is called a *quadrature encoder*. This construction allows the encoder to generate two square-wave signals that are 90° out of phase, as shown in Figure 4.3.2. These two square waves allow for determination of shaft rotation and an increase in resolution by a factor of 4 using quadrature logic.

Assume that there are two square waves that are 90° out of phase. Next, assign a two-element box to each square wave and assume that the boxes move in phase with each other to the right while the square waves remain stationary. When a box encounters a change of state in its respective square wave (i.e., goes from high to low or from low to high), then the old left element (a one or a zero) gets tossed out, the old right element of the box moves to the left element, and the new right element assumes the value of the state of the square wave (i.e., high = 1, low = 0). This is accomplished electronically using a device known as a shift register. If the A box becomes the first two elements of a 4-bit word (most significant bits), and the B box becomes the last two elements, then during forward motion as successive states are crossed, the decimal values of the 4 bit word are 4, 13, 11, and 2. As the box moves to the left during reverse motion, the values of the 4-bit word become 8, 14, 7, and 1. Note two facts: (1) During one period of the square waves, four changes of state are recorded in the 4-bit word; and (2) if the encoder sits on the edge of a wave, the value of the 4-bit word will oscillate back and forth between two numbers, for example, 4 and 8, or 13 and 14, or 11 and 7, or 2 and 1. Thus it is impossible to generate false counts. If only one square wave were used and counts were triggered by high-low readings, then vibration could cause the reading to rapidly oscillate back and forth over one point, creating an illusion of high-speed motion. For instance, it is unwise to assume that a single sensor can be used to count gear teeth and to use the high-low signal that is generated as a means for determining angular position of the gear. Two sensors 90° out of phase must be used. Single chips are now available[6] to decode the information (square waves) from the A and B channels and store counts in memory for servocontroller access.

[6] For example, Hewlett-Packard's HCTL-2000 Quadrature Decoder/Counter Interface IC. See HCTL-2000 Technical Data and Specifications, Hewlett-Packard, P.O. Box 10301, Palo Alto, CA 94303-0980.

Moiré Fringes and Interpolation Encoders

When two grids of comparable spacing are laid over one another with one at a slight incline, as shown in Figure 4.3.3, an interference pattern is generated that is called a Moiré fringe.[7] If the grids are formed from fine light transmitting lines and the system is illuminated from behind, then light will only be transmitted through to the observer in regions where the lines intersect. Consequently, as one grid moves with respect to the other, the transmitted points of light will appear to move orthogonally with respect to the direction of motion of the grids. If the grids are coarse with respect to the wavelength of light used, then diffraction effects can be ignored. For small angles θ between grids of equal line spacing ℓ, the vertical range of motion of the transmitted points of light will be ℓ/θ. The resolution of the device is thereby ideally increased from ℓ to $\ell\theta$ although the width of the band of light also places a limit on the attainable resolution, as can be seen from the figure.

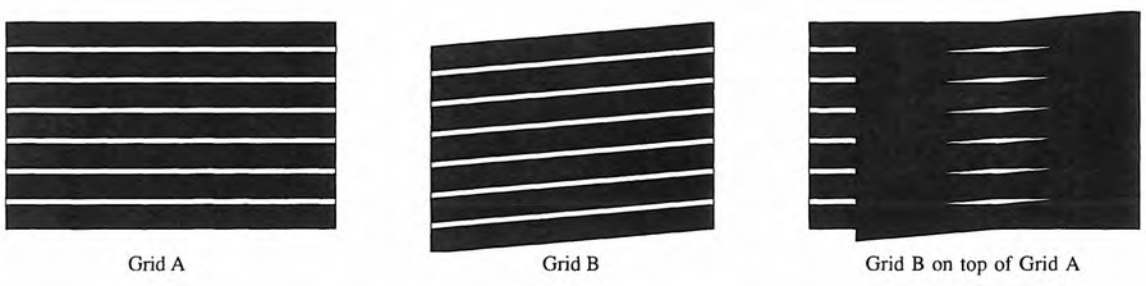


Figure 4.3.3 Moiré fringes produced by superimposing two grids.

Perhaps a better technique, based on the same type of design logic, uses reference grid windows that are not tilted but have sections that are spaced 1/4 pitch. This increases the signal as well as compactness of the scale and read head for a given desired increase in resolution. This is the technique commonly used in linear and rotary optical encoders to increase resolution by generating sine and cosine wave output from the photodiodes. Where the sine wave has poor resolution ($0, \pi, 2\pi, \ldots$) the cosine wave provides high resolution, and vice versa; hence the position between the grid lines can be accurately interpolated. Note, however, that it is assumed that the intensity is constant. For rotary encoders, this method can yield an increase in the encoder's resolution typically by a factor of about 25, although a factor of 80 is possible; this type of encoder is called an *interpolation encoder*. After interpolation, two 90^{o} out-of-phase square waves are generated. The square waves can then be used in quadrature to give a further $4\times$ increase in resolution. Thus it is possible to increase the resolution of an interpolation encoder electronically by a factor of 100, although linearity begins to degrade at these high magnifications.

4.3.2 Absolute Position Encoders

As shown in Figure 4.3.4, an absolute encoder's disk is divided into N sectors, where N is the number of bits resolution of the encoder. On one side of the disk is a light source that uses a parabolic reflector to produce a plane of light, and on the other side of the disk is a set of N photodiodes arranged radially. The sectors are arranged so that a digital word is formed by each set of opaque and transparent sections in a sector. Each sector increments one power of 2 from the previous sector; hence the detectors produce a parallel word representing the shaft angle, and position of the shaft from 0 to 360^{o} is known even when the power is first turned on. The digital word produced by an absolute encoder is not a straight binary word. Instead, a gray scale is used to help avoid errors in reading the word. A straight N-bit binary word is based on the formula

$$\text{Number}_{\text{base}}\ 10 = \sum_{n=1}^{N} 2^{n-1} \tag{4.3.1}$$

[7] See J. Meyer-Arendt Introduction to Classical and Modern Optics, Prentice Hall, Englewood Cliffs, NJ, 1972; M. Stecher. "The Moiré Phenomenon," Am. J. Phys., Vol. 32, 1964, pp. 247–257; and A. T. Shepard, "25 Years of Moiré Fringe Measurement," Precis. Eng., Vol. 1, No. 2, 1979, pp. 61–69. Also see O. Kafri and I. Glatt, The Physics of Moiré Metrology, John Wiley & Sons, New York, 1990.

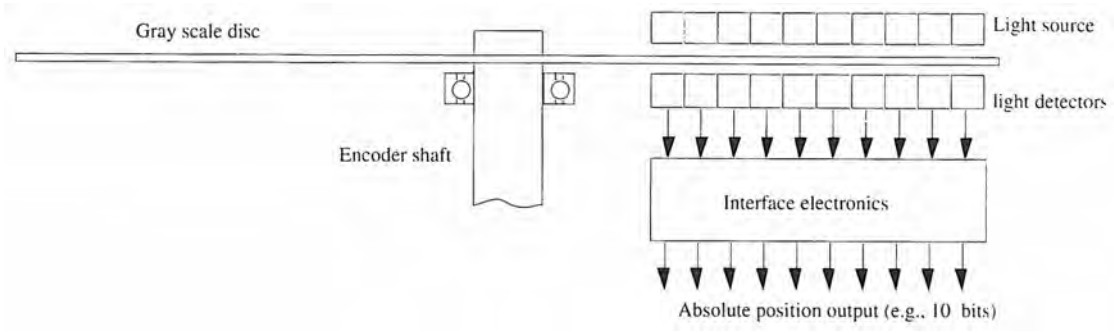


Figure 4.3.4 Operating principle of an absolute optical encoder.

This can lead to significant measurement errors. For example, consider the two 3-bit encoders shown in Figure 4.3.5.[8] The shaded regions represent areas where the light cannot shine through to the detectors. One disk is encoded in straight binary and one is encoded in gray scale, and both have a resolution of $360^o/2^3 = 45^o$. The problem arises when vibration or foreign contamination causes a photodiode to trigger at the transition before the other diode releases. For example, assume that when moving from 45^o (001) to 90^o (010) on a straight binary-encoded disk, a forklift runs into the machine and causes the middle track to trigger before the outer track releases. This jiggles the encoder and causes 011 to appear for an instant. Thus the control computer would see 001 then 011 and then 010, which corresponds to 45^o, then 135^o, then 90^o. In this instant, the servo controlling the machine would try to move from 45^o to get to 135^o and then back to 90^o. The gray scale, on the other hand, never has more than one state changing at a time, as shown in Table 4.3.1; thus it is far more forgiving and safe.

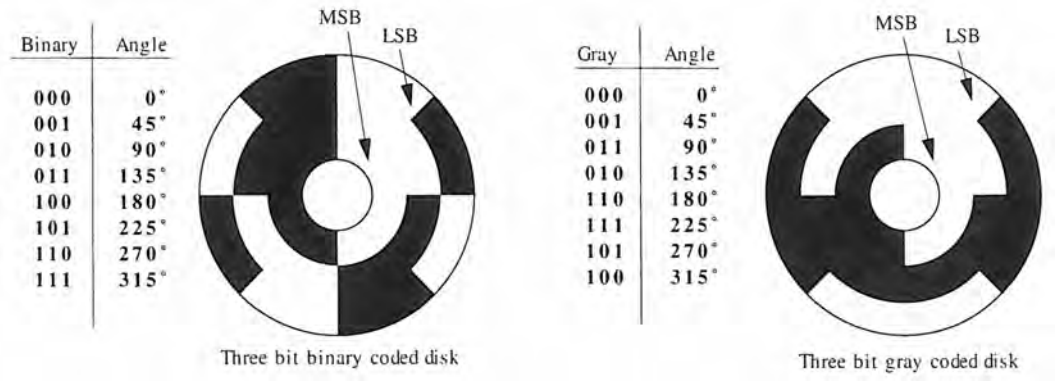


Figure 4.3.5 Comparison between binary and gray scale absolute encoder disks. (After Fletcher.)

As shown in Figure 4.3.6, an N-bit gray code number is computed from an N-bit binary word by using the following logic:

1. Place a zero to the left of the binary word's most significant bit.
2. Beginning at the left of the number, perform an EX-OR operation[9] on all pairs of bits.

To go from gray code to straight binary, use the following logic:

1. Copy the first "1" as it stands.
2. Write 1's until the next "1" is met.
3. Write a "0."
4. Write 0's until the next "1" is met.
5. Write a "1."
6. Go to step 2.

[8] For more on codes and digital logic, see, for example, W. Fletcher, An Engineering Approach to Digital Design, Prentice Hall, Englewood Cliffs, NJ, 1980.

[9] The EX-OR (EXclusive-OR) function definition is: If two bits are identical, the result is "0". If the two bits are different, the result is "1".

Decimal	Straight Binary	Gray Binary
0	000000	000000
1	000001	000001
2	000010	000011
3	000011	000010
4	000100	000110
5	000101	000111
6	000110	000101
7	000111	000100
8	001000	001100
9	001001	001101

Table 4.3.1 Comparison of decimal, straight binary, and gray binary numbers

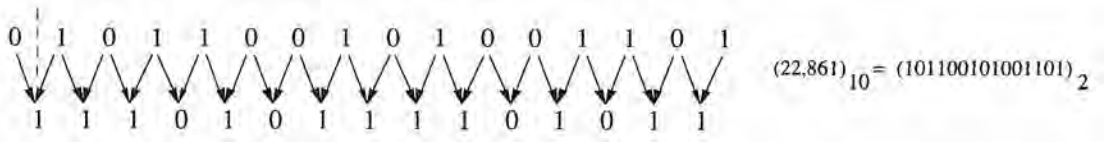


Figure 4.3.6 Conversion of a straight binary word to a gray binary word by the EXclusive-OR rule.

Despite the seemingly complex output from an absolute encoder, its principal advantage is that its angular position (from 0 to 360°) is always known, even from initial power-up, and no external interpolating electronics are needed. However, for more than one turn, as needed to determine the rotation of a leadscrew, an external counter that keeps track of full revolutions is still needed. Their primary disadvantage is that interpolation and quadrature techniques cannot be used to increase their accuracy.

4.3.3 Diffraction Gratings and Encoders[10]

Conventional optical encoders are limited in resolution because as the scale lines (slits) are made narrower in order to increase resolution, increasing diffraction effects reduce the signal-to-noise ratio of the photodiode. Scale line pitch is limited by diffraction effects to about 125 lines/mm. Since the slit width is limited, conventional encoders have to increase their diameter in order to increase resolution. This increases the encoder inertia and also makes mounting more difficult; however, a larger encoder diameter does decrease cyclic errors caused by radial error motions of the shaft. If the pitch of the disk and grating becomes very fine, then the light will diffract as it passes through. The different orders of diffracted light will have phase variations between them. Hence if they can be made to interfere, the resultant fringes can then be focused onto photodiodes that produce square or sinusoidal wave output.[11] This phenomenon can be used in the design of rotary and linear encoders. The cost to produce the gratings is substantially more, but encoder resolution can be increased by one to two orders of magnitude. The remainder of this section will discuss the design of a very compact encoder, called a laser rotary encoder, that uses this method. The laser rotary encoder is shown in Figures 4.3.7 and 4.3.8. There are two types of signal outputs: One type of output is pure sine and cosine waves which allow for 4× interpolation and 4× quadrature multiplication for a total resolution of 1,296,000 counts per revolution (1 arcsecond). The other type of output is square-wave output, which yields 324,000 counts per revolution (19.4 μrad) with quadrature logic.

Detection Mechanism of Rotation Angle[12]

The phase of a light wave diffracted by a diffraction grating is changed (modulated) by movement of the grating.[13] As shown schematically in Figure 4.3.9, a laser beam illuminates a grating

[10] This section was written principally by Dr. Katsuji Takasu of Canon USA Inc.

[11] See J. Burch, "The Metrological Applications of Diffraction Gratings," in Progress in Optics, E. Wolf (ed.), John Wiley & Sons, New York, 1963.

[12] For a more detailed discussion, see T. Nishimura and K. Ishizuka, "Laser Rotary Encoders," Motion, July/Aug. and Sept./Oct. 1986.

[13] See, for example, K. Matsumoto, "Method for Optical Detection and/or Measurement of Movement of a Diffraction Grating", U.S. Patent 3,726,595, 1973; and F. A. Jenkins and H. E. White, Fundamentals of Optics, McGraw-Hill Book Co., New York, 1981, pp. 355–377.

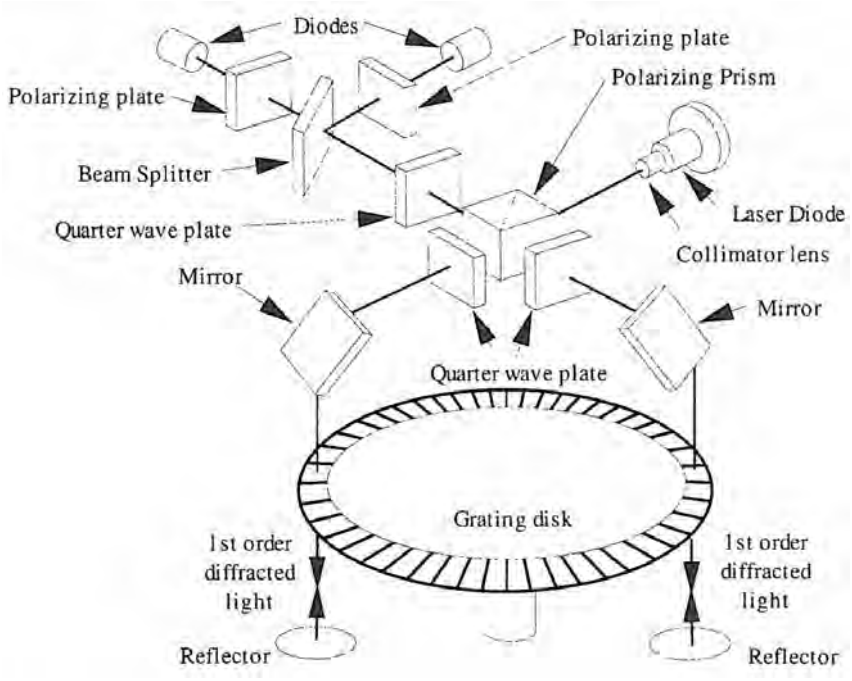


Figure 4.3.7 Operating principle of Canon's laser rotary encoder. (Courtesy of Canon USA, Inc.)

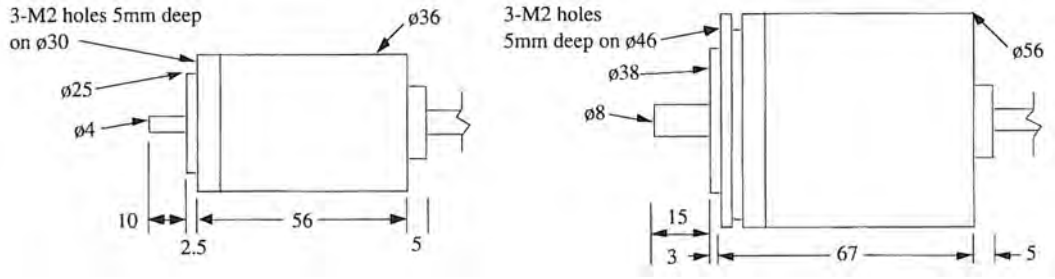


Figure 4.3.8 Laser encoders with (left) 81,000 pulses/rev and 4 N radial, 9 N axial allowable shaft loads, and (right) 50,000 pulses/rev and 15 N radial, 19 N axial allowable shaft loads. (Courtesy of Canon USA, Inc.).

having many slits periodically placed around its circumference. When the grating is moved by an amount x, the optical path of a diffracted light with angle α is changed by

$$\Delta\ell(x) = x \sin\alpha \tag{4.3.2}$$

Substituting the diffraction equation gives

$$P \sin\alpha = m\lambda, (m: \text{ the order of diffraction}) \tag{4.3.3}$$

where P is the distance between the leading edges of the slits and λ is the wavelength.

Substituting Equation 4.3.3 into Equation 4.3.2 gives

$$\Delta\ell(x) = m\lambda x/P \tag{4.3.4}$$

Movement of the grating causes a phase shift of the mth-order diffracted light wave by

$$\Delta\phi(x) = \Delta\ell(x)2\pi/\lambda = 2m\pi x/P \tag{4.3.5}$$

If the grating moves by P, the phase of first-order diffracted light wave is shifted by 2π.

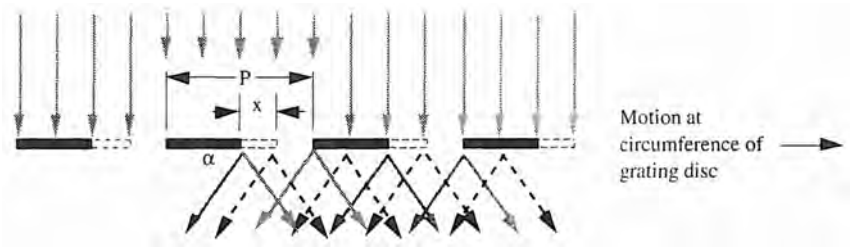


Figure 4.3.9 Light path through a laser encoder's grating disk. (Courtesy of Canon USA, Inc.)

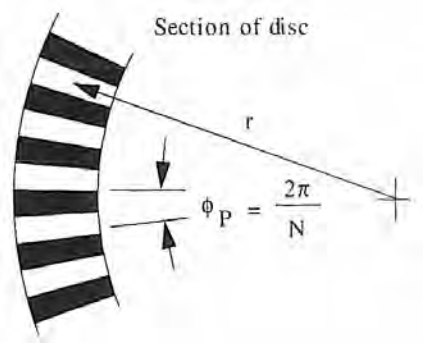


Figure 4.3.10 Laser rotary encoder disk geometry. (Courtesy of Canon USA, Inc.)

Next consider the phase change of a light wave diffracted by a grating disk having N slits on its circumference, as shown in Figure 4.3.10. An angle ϕ_p in radians which corresponds to one period of the slit pattern on the disk is

$$\phi_p = 2\pi/N \tag{4.3.6}$$

A distance of one period P_r of the slit pattern at a radius r, is

$$P_r = r\phi_p = 2\pi r/N \tag{4.3.7}$$

When the grating disk rotates by an angle θ, the movement of the slit pattern at a radius r becomes P_r. The corresponding phase change, $\Delta\phi(\theta)$, of the mth-order diffracted light wave is obtained by substituting both $x = r\theta$ and $P = P_r$ into Equation 4.3.5:

$$\Delta\phi(\theta) = 2m\pi r\theta/P_r = mN\theta \tag{4.3.8}$$

For example, the phase of the first-order diffracted light wave (m = 1) is changed by $2N\pi$ radians when the grating disk rotates by 2π radians.

Equation 4.3.8 shows that the rotation angle of the grating disk can be calculated from a measurement of phase change of diffracted light waves. In general, the relative phase of two light waves can be measured by measuring the intensity of the interference pattern generated by two light waves. Figure 4.3.11 illustrates this method. The phase of the two beams and therefore the intensity of the interference pattern is modulated by rotation of the grating disk. The intensity of the interference pattern from the two first-order diffracted waves is

$$I(\theta) = \left|e^{i(\omega t+\phi_0-N\theta)} + e^{i(\omega t+\phi_0+N\theta)}\right|^2 = 2\{1+\cos(2N\theta)\} \tag{4.3.9}$$

where ω is the angular frequency of the laser light, and ϕ_o is a constant phase angle that depends only on the optical path between the laser diode and the photosensor. Equation 4.3.9 shows that the intensity of the mixed light wave is expressed by a cosine wave with a period of $2\pi/2N$ radians. Thus one rotation of the grating disk generates 2N pulses of the signal.

If mirrors are used to reflect the diffracted laser light back through the grating disk as shown in Figure 4.3.12, the phase change of the mixed light wave is doubled:

$$I(\theta) = \left|e^{i(\omega t+\phi_0-2N\theta} + e^{i(\omega t+\phi_0+2N\theta)}\right|^2 = 2\{1+\cos(4N\theta)\} \tag{4.3.10}$$

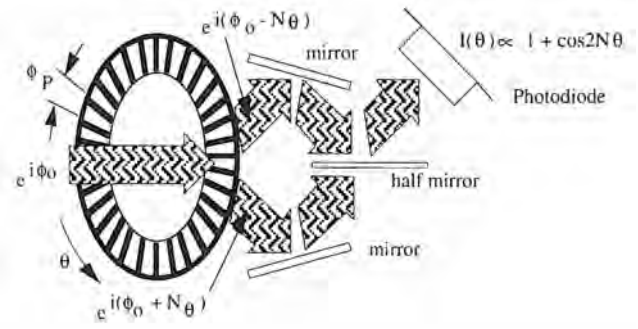


Figure 4.3.11 First-order diffracted light combining to form a cosine wave. (Courtesy of Canon USA, Inc.)

Thus the number of signal pulses detected is four times the number of slits on the disk.

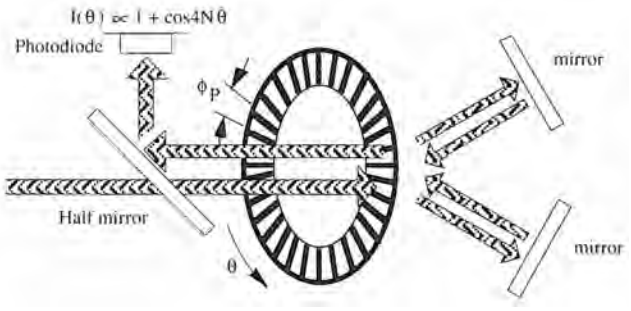


Figure 4.3.12 Mirrors used to make diffracted waves pass back through grating disk to increase resolution. (Courtesy of Canon USA, Inc.)

A rotary encoder must have two signals, the phases of which are 90° apart, in order to allow for determination of the direction of rotation. Thus the rotary encoder must have both sine and cosine wave outputs. To meet this requirement, the laser rotary encoder utilizes a special optical cavity design to split a single coherent laser beam into two beams: one for generation of the sine wave and one for the generation of the cosine wave. Sine and cosine waves can be used to increase the resolution by interpolation.

Figure 4.3.13 shows a detailed view of the construction of the laser encoder, which operates as follows:

1. Plane-polarized light from the diode laser is collimated and passed through a polarizing beamsplitter which breaks the light into orthogonal components ($\mathbf{E_X}$ and $\mathbf{E_Y}$ components). One of the components proceeds straight on through, and the other one is diverted to the right; half of the light goes off to location M1 and half of the light off to location M2.
2. The two beams of light are then each directed through the grating on the disk, where the light is diffracted and reflected back through the disk along the line of its incoming path.
3. The $\mathbf{E_X}$(M1) and $\mathbf{E_Y}$(M2) beams of light each make two passes through a quarter-wave plate on their way to and from the disk, which rotates their respective polarization angles by 90°. The beam which was previously internally reflected by the prism now passes straight through, and the beam which passed straight through is now internally reflected. As a result, both beams emerge from the prism in line. However, the two beams are 180° out of phase, which is the same as having a $\cos(\theta)$ and $-\cos(\theta)$. To obtain $\cos(\theta)$ and $\sin(\theta)$ waves, the beams are passed through a quarter-wave plate which changes the relative phase of the $\mathbf{E_Y}$ $(\cos\theta)$ and $\mathbf{E_X}$ $(-\cos\theta)$ beams by 90°.
4. The beams now pass through a beamsplitter which sends 50% of the light from both beams in two orthogonal directions. The light from these two beams passes through polarizing plates, which only let $\mathbf{E_Y}$ $(\cos\theta)$ and $\mathbf{E_X}$ $(\sin\theta)$ beams through to the respective diodes.

The two diodes then generate sine and cosine analog waveforms to be interpolated and quadratured. The sine and cosine signals are used to determine the position of the disk between graduations. Where the sine wave has poor resolution (0, π, 2π, ...) the cosine wave provides high resolution, and vice versa. Note, however, that it is assumed that the intensity is constant.

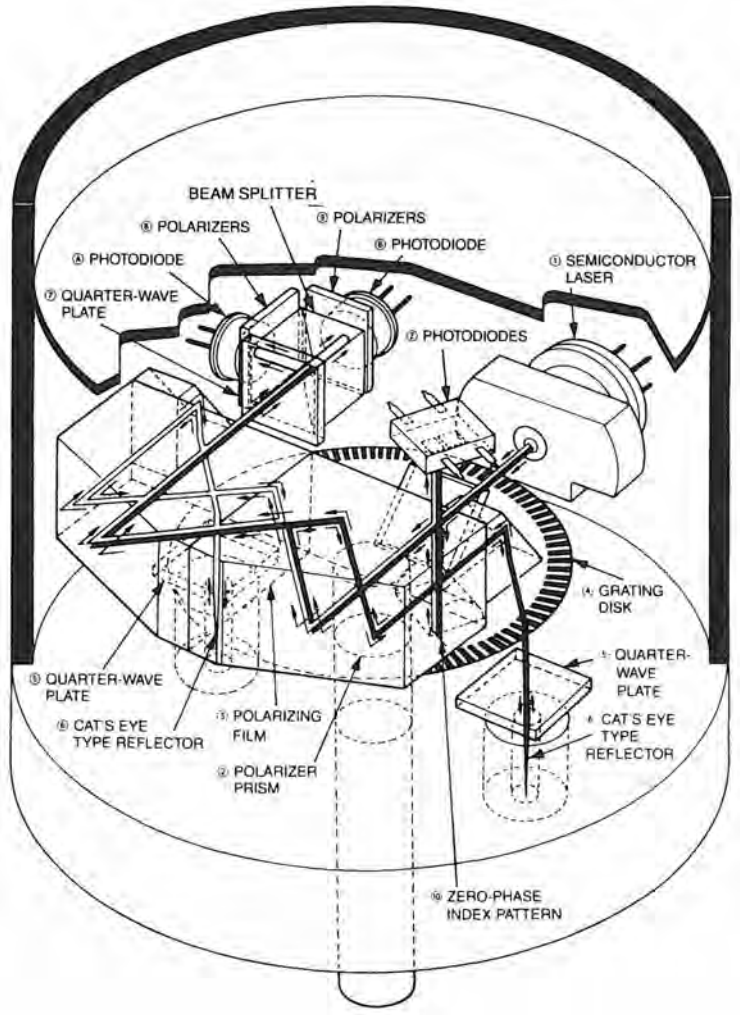


Figure 4.3.13 Internal construction of Canon's laser rotary encoder. (Courtesy of Canon USA, Inc.)

Many high-resolution encoders now make use of diffracted light to increase resolution. A new technique has also been developed to make the encoder less sensitive to contamination and relative alignment of the head and disk. Three signals are formed (R, S, T) that are phased $2\pi/3$ radians apart. This also allows the intensity to be determined, which greatly increases the accuracy of the interpolation by sine and cosine waves. As a result, submicron (microinch) measurements can be made with conventional linear encoder scales and a retrofitted read head.[14]

4.3.4 Linear Encoders

Linear encoders are constructed using the same opto-electronic logic as rotary incremental encoders; only the mechanical construction is different. Alignment of the scale, light source, and photodiodes

[14] M. Hercher and G. Wyntjes, "Fine Measurements with Coarse Scales," 1989 Joint ASPE Annu. Meet. & Int. Precis. Eng. Symp., Monterey, CA, preprint pp. 86-92. Patents pending by Optra Inc., Cherry Hill Park, 66 Cherry Hill Drive, Beverly, MA 01915-1065.

is very critical because most electronic interpolation assumes the intensity of light is constant; thus many models have integral bearings, which make them look like a box with a movable plunger rod protruding. Since this is impractical for long travels on precision machines, some manufacturers offer a linear scale and traveling read head that for external appearance and mounting considerations are similar to magnetic scales (e.g., see Figure 3.2.21). Linear optical encoders are often used on CNC machine tools and are also often used with digital readouts on manual machines. Unhoused versions are also available which consist of a glass scale and a read head which straddles it.

Traditional linear encoders' scale line pitch is limited by diffraction effects to about 125 lines/mm. This limits the encoder's resolution to about 0.2 μm (after interpolation and quadrature logic), although 0.1 μm resolution can be obtained with more elaborate electronics. Linear encoders based on the diffraction effect are now made by several different manufacturers. Linear diffraction encoders can easily achieve 0.01 μm resolution after interpolation and quadrature multiplication.[15] Accuracy is on the order of 3 ppm at 20°C, and they cost two to three times more than a conventional linear encoder and about one-half to one-third as much as a laser interferometer. With any scale, one needs to be concerned with the stability of the glass itself which may be only 1 part in 10^6 or 10^7 per year due to phase transformations in the material; however, the scale can periodically be calibrated with a laser interferometer.

There are not many types of linear optical encoders that are available because of the difficulty in encoding the scale. A new type of absolute linear encoding system has been developed by Parker Hannifin Corp.[16] Called the Phototrak®, the encoder uses pairs of slits in a stainless steel scale. The distance between the centroid of the slits is constant, but the distance between the slits varies with position along the scale. The maximum range is about 2 m, the resolution is 1 micron, and the accuracy is about 5 microns. This type of encoder costs about twice as much as an incremental encoder.

4.3.5 Selecting an Optical Encoder

There are many manufacturers of different types of optical encoders for different applications in different environments.[17] Units are available with virtually any desired shaft type, mounting flange type, integral shaft coupling type, and so on. Thus when selecting an optical encoder, one must consider the type of environment and loading the encoder will be subject to and the accuracy or resolution required. If the device operates slowly or intermittently, an absolute encoder should be chosen. In extreme environments, a shock-proof case should be specified. An encoder in a shock-proof case is mounted inside a heavy cast iron or aluminum housing by means of vibration absorbing elastomers. The encoder shaft is coupled with a flexible coupling to the separate shaft housing so that high external loads cannot be transmitted to the delicate encoder shaft. For clean environments, a simple inexpensive incremental model costing less than $100 can often be used which utilizes the bearings and shaft of a machine to support the disk (i.e., a modular encoder mounted to a motor with a shaft protruding from its rear). Note that modular encoder designs have far less cyclic error due to radial error motions of the shaft, than do housed encoders, which must be coupled to the device shaft. This is discussed in detail in Section 5.3.

Typical Characteristics of Optical Encoders

The following summary of optical encoders' characteristics are generalizations only. Manufacturers are always advancing the state of the art, so this general summary should by no means be taken as gospel. Furthermore, it is extremely important to note that accuracy and often repeatability are highly dependent on the manner in which the sensor is mounted. Chapter 5 should be consulted for a detailed discussion of sensor mounting considerations.

Size: Rotary models: Incremental: From 12 mm diameter and 12 mm long for a miniature encoder with 8 bits resolution, to 65 mm diameter and 65 mm long for an average 12-bit resolution model, to 150 mm diameter and 150 mm long or 170 mm (6.7 in.) diameter and 50 mm long for arcsecond-resolution models. Absolute: 65 mm diameter and 75 mm long for an average

[15] See, for example, A. Teimel, "Technology and Application of Grating Interferometers in High Precision Measurement, Progress in Precision Engineering, P. Seyfried, et al. (Eds.), Springer-Verlag, New York, 1991, pp. 15–30.

[16] B. Hassler "A New Absolute Linear Position Encoding Technology," Motion, Nov./Dec. 1990, pp. 3–8.

[17] See G. Avolio, "Encoders, Resolvers, Digitizers," Meas. Control, Sept. 1986, pp. 232–245.

12-bit model, to 110 mm diameter by 130 mm long for a 16-bit model. Laser encoder 36 mm diameter and 50 mm long. Linear Models comparably sized to Magnescales®.

Cost: Rotary models: From less than $100 for modular encoders that are assembled on the shaft of a motor, to $250 for an average 12-bit incremental encoder, to $500 for a 12-bit absolute encoder, to $6300 for a 20-bit (1 arcsecond) resolution encoder, to $10,000 for a 23-bit resolution encoder. Laser rotary encoder costs on the order of $2000 and linear models about $5000. Conventional linear encoders are priced comparable to Magnescales®.

Measurement Range: Rotary incremental encoders are not rotation limited. Absolute encoders can have a range to 10 turns. Linear encoders' range to 3 m on glass scales. With stainless steel tape reflecting scales, virtually any length scale can be produced.

Accuracy: Highly dependent on method used to couple the encoder to a shaft. Rotary models: For conventional encoders, accuracy can be half (worse) the repeatability. Large 22-bit incremental resolution encoders can have 5 μrad accuracy. Laser encoder cumulative error in one rotation is on the order of 73 μrad (15 arcseconds). Linear models: Typically l μm with some units having 1/4 μm.

Repeatability: Rotary models: On the order of half (worse) the resolution for most encoders. Large 22-bit resolution incremental encoders can have 2.5 μrad repeatability. Laser encoder repeatability is around 10 μrad (2 arcseconds). Linear models: Conventional encoders typically 1/2 μm with some models having 1/4 μm. Linear diffraction encoders can have 0.05 μm repeatability.

Resolution: Rotary models: The cost of almost any single model encoder up to 10 bits is relatively constant. 12-bit incremental and absolute encoders are very common. 16 bits is the highest resolution for an absolute encoder. Some incremental encoders can achieve over 10^7 counts per revolution (24 bits) with 25 × interpolation and 4 × multiplication. A laser encoder can have 324,000 counts per revolution with quadrature logic (4 × multiplication) and 1,296,000 counts per revolution with 4 × interpolation and 4 × multiplication. Conventional linear encoders typically have 1/2 μm resolution, with some units having 0.1 μm resolution. Linear diffraction encoders can achieve 0.01 μm resolution after interpolation and multiplication. Expect nanometer and then angstrom resolution in the near future.

Environmental Effects on Accuracy: Thermal gradients can cause differential expansion of the encoder structure and affect alignment of internal components. Special construction can help to balance thermal expansion and reduce effects on accuracy. In general, thermal effects on the machine will be much more significant.

Life: Depends on the environment and the mechanical design of the housing, bearings, and seals. 10^7 cycles of the shaft are not uncommon; however, as the mechanical system wears, accuracy is likely to degrade.

Frequency Response: Rotary models: Typical allowable mechanical shaft speeds are 4000-10,000 rpm. Allowable shaft speed to prevent missing counts depends on the electronics used, but a rule of thumb is to assume the count rate is one-tenth the rate of the system clock. Thus a system with an 8 MHz clock can accept 800,000 counts per second. For a 12 bit encoder, this corresponds to a shaft speed of about 11,700 rpm. For linear encoders the maximum velocity is typically 0.5-1 m/s.

Starting Torque and Force: Rotary models: Unsealed models for clean room operation do not use seals and with instrument bearings have no appreciable starting torque (<0.1 N-cm). Sealed models for operation in a splashed environment can have starting torques on the order of 1–5 N-cm. Linear models without seals will have no starting force, while sealed models may have 1/2 - 1 N starting force.

Allowable Operating Environment: Operating temperatures range from 0°C to 50°C for most models, with special designs allowing for operation from 250°K to 80°C. Precision systems should be operated at 20°C (68°F).

Shock and Vibration Resistance: Typically, 50g shock for 10 ms. Vibration at 10g up to 2000 Hz.

Misalignment Tolerance: Rotary models: Misalignment will not affect full revolution counts, but will affect linearity of the output. Linear models: Misalignment causes errors that are proportional to the cosine of the angle of misalignment.

Support Electronics: Requires power supply from ± 5 V dc ±5% to 24 V dc ±5%. An A quad B interface may cost on the order of $100. Interpolation electronics $400 for 5 ×, $900 for 25 ×.

4.4 FIBER OPTIC SENSORS[18]

Optical fibers transmit light using the property of total internal reflection: Light which is incident on a media's interface will be totally reflected if the incident angle is greater than a critical angle known as Brewster's angle. As is illustrated in Figure 4.4.1, this condition is satisfied when the ratio of the index of refraction of the fiber and the cladding is in proper proportion. This ratio also governs the efficiency at which light from a source will be captured by the fiber: The more collimated the light from the source, the more light that is transmitted by the fiber. The acceptance cone angle θ is known as the numerical aperture and has the value

$$\theta = \sin^{-1}(n_1^2 - n_2^2) \tag{4.4.1}$$

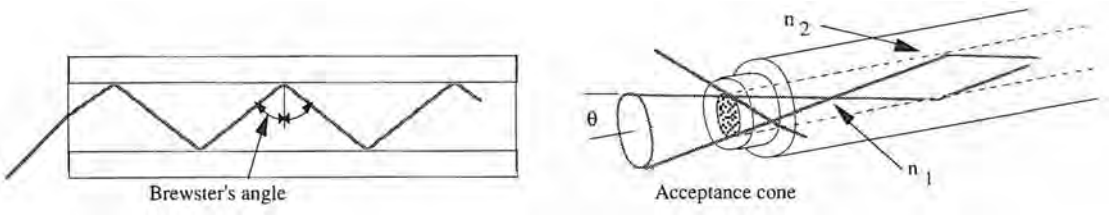


Figure 4.4.1 Propagation of light through a fiber. (Courtesy of 3M.)

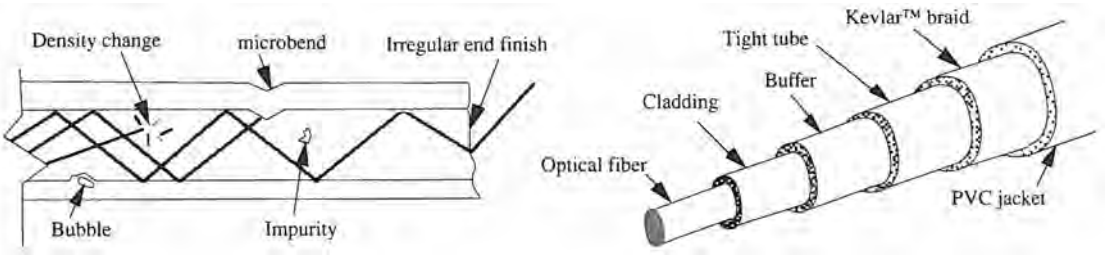


Figure 4.4.2 Construction of a fiber optic cable and typical defects. (Courtesy of 3M.)

A multimode fiber optic cable of the type of interest here actually has a multilayered structure, as shown in Figure 4.4.2, that provides protection and support for the cable. Also shown are common defects in individual fibers, although the grade of the fiber is measured simply by its light transmission efficiency. For telecommunications work, ultrahigh efficiency is needed to minimize the number of expensive amplification stations that are needed to transmit signals over long distances.

There are basically three kinds of reflective fiber optic bifuricated probe configurations: hemispherical, random, and fiber pair, as shown in Figure 4.4.3. The front slope of the reflection curve is due to the fiber being held extremely close to the surface. The back slope follows a $1/R^2$ curve as the distance increases. Bifurcated bundles have one common end and the other end is split evenly

[18] For a detailed discussion of the physics of operation of optical fibers, see H. Haus, Waves and Fields in Optoelectronics, Prentice Hall, Englewood Cliffs, NJ, 1984. Fiberoptic displacement sensors are easily made using an LED, fiber optic cable, and CCD. For information on availability of parts and assemblies, contact Marketing Coordinator, 3M Corp., 420 Frontage Road, West Haven, CT 06516. Also see information from MTI Instruments, 968 Albany-Shaker Rd., Latham, NY 12110.

into two, as shown in Figure 4.4.4. Bifurcated bundles are used to sense the intensity of reflected light from an object; one branch of the divided end emits light and the other branch receives the light. This is a commonly used arrangement to obtain distance sensing performance as indicated by the curves in Figure 4.4.3. All three types of probes shown in Figure 4.4.3 can be used to sense distance, and the type of probe used depends on the accuracy required. Randomized bundles have no order of fiber arrangement at either end of the cable. Many photoelectric fiber optic sensors use such bundles.

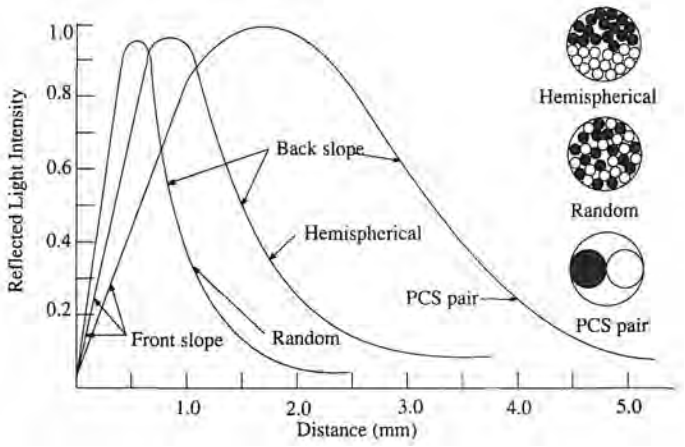


Figure 4.4.3 Generalized performance characteristics for three types of reflective fiber optic probes. (Courtesy of 3M.)

The distance of an object from the face of a fiber optic bundle consisting of source and receiving fibers can be determined based on the intensity of reflected light that is sensed.[19] When the sensor is very close to the surface, the light cannot be reflected into the receiving fibers; hence the occurrence of the front slope phenomenon shown in Figure 4.4.3. The receiving fibers transmit the reflected light back to a photodiode, which measures the intensity. As the sensor continues to move away from the surface, the cross-sectional area of the reflected light beam in the plane of the receiving fibers increases and the resultant intensity decreases. Ideally, the performance of the sensor is a function of the cross-sectional geometry of the bundle, the illumination exit angle, and the distance from the surface. Tilting or contamination of the surface can quickly lower resolution. For a seven-fiber bundle, measurement ranges of 1 mm are possible with 12-bit resolution.

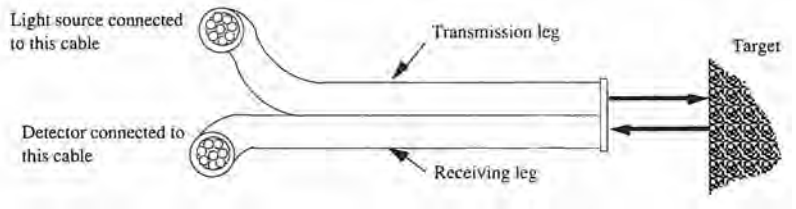


Figure 4.4.4 Bifurcate probe used in a reflective scanning mode. (Courtesy of 3M.)

Optical fibers are an ideal medium for transmitting signals because they are immune to electromagnetic interference. In addition, fiber optic sensors typically have small mass and can withstand shock loads of 100g or more. Their temperature range of operation is limited only by the temperature limits of the sheathing and epoxy used in their fabrication. Some fiber optic sensors can operate at temperatures of 200–400°C. By contrast, most photoelectric sensors are limited to 100°C as a maximum operating temperature. Because the principle of total internal reflection is unaffected by moderate bending of the fiber, fiber optics can greatly extend the reach of light sources and sensors into difficult to reach places such as inside the human body.

[19] See Giallorenzi et al., "Optical Fiber Sensor Technology," IEEE J. Quantum Electron., Vol. QE-18, No. 4, 1982, pp. 626–665.

Furthermore, they can be used to measure physical quantities that change the properties of light transmission through the fiber itself (e.g., as interferometers).

Typical Characteristics of Fiber Optic Displacement Sensors

The following summary of fiber optic sensors' characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as "gospel."

Size: Cable diameter can be as small as 1.0 mm but typically is 3-5 mm.

Cost: $100 to $1000 or more, depending on the type of sensor.

Measurement Range: From several meters for proximity sensors to a few millimeters for measurement of small displacements.

Accuracy (Linearity): Can be as good as 0.1% of full-scale range for distance sensors kept normal to a clean reflecting surface.

Repeatability: Highly dependent on environmental conditions, but can be on the order of 0.05% of full-scale range.

Resolution: If the fiber is very small and held very close to the surface of an object, resolutions as small as 0.1 μm are achievable. More typically, a fiber optic sensor will have a resolution of 10 μm and a range of 1–5 mm.

Environmental Effects on Accuracy: Almost entirely dependent on the effect of the environment on the surface at which the sensor looks. Dirt on the surface of a target can degrade the performance of a fiber optic sensor; hence air wipes can be helpful. The photodiode used to monitor the strength of the reflected signal is usually kept back in a protected environment.

Life: The probes are noncontact, so life can be infinite if the cable is not fatigued.

Frequency Response (-3 dB): Up to 10 kHz, depending on the photodiode used to monitor the return signal.

Starting Force: Fiber optic sensors are noncontacting, and therefore no force exists between the probe and sensing surface. As is true with all sensors, the flexing force of the cable may need to be considered in some situations.

Allowable Operating Environment: To maintain accuracy the surface of the probe and the sensing surface must be kept clean. Operating temperatures for the region of the probe may range from 200°C to 400°C. However, the light source and photodiode must often be protected from these extremes. The individual fibers must not be exposed to moisture or they will eventually erode to the point of failure.

Shock Resistance: On the order of 100 g.

Misalignment Tolerance: A varying angle between probe and target will be directly measured by the probe. Thus multiple calibrated sensors are needed if the target translates and rotates with respect to a probe.

Support Electronics: Light source and detector.

4.5 INTERFEROMETRIC SENSORS

Interference results from superimposing two or more waveforms, and an interferometer is a means for measuring the otherwise invisible effect of interference. There are numerous types of interferometers in use today. In order to understand and appreciate the power of an interferometer as a measurement tool, one must first understand the fundamental physics of the properties of waves and light.[20]

Interference of Waves with Equal Frequency, Amplitude, and Velocity

Consider two waves traveling in the same direction with the same amplitude, frequency, and velocity, but one lags behind the other by a phase angle ϕ:

$$Y_1 = A\sin(kx - \omega t - \phi) \tag{4.5.1}$$

[20] By now, much of this knowledge has been forgotten by most engineers. If it were put in an appendix, no one would read it and readers would miss what they should know. Thus this review is presented here.

$$Y_2 = A\sin(kx - \omega t) \tag{4.5.2}$$

The physical significance of the phase angle is that at any instant in time t, if the two waves are compared, one will be displaced along the X axis by a constant distance ϕ/k. The constant k is the inverse of the wavelength λ; $k = 2\pi/\lambda$. At any point x, one wave will appear to lag the other in time by an amount ϕ/ω. If the two waves are superimposed, the resultant waveform has the equation:[21]

$$Y = 2A\cos\left(\frac{\theta}{2}\right)\sin\left(kx - \omega t - \frac{\phi}{2}\right) \tag{4.5.3}$$

The resultant wave has the same frequency of the original wave. When the phase ϕ is zero, the resultant amplitude is twice that of the original waves; the waves are constructively interfering. When the phase ϕ is 180^o, the resultant amplitude is equal to zero everywhere; the waves are destructively interfering.

When two waves travel different paths to get to the same point (e.g., one beam of light used as a reference and the second is reflected from a target), they will constructively interfere if the path length difference to and from the mirror is 0, λ, 2λ, 3λ,..., corresponding to a phase change of 0, 2π, 4π, 6π,... On the other hand, the waves will destructively interfere when their path length differences to and from the mirror are $\lambda/2$, $3\lambda/2$, $5\lambda/2$, $5\lambda/2$,... corresponding to a phase change of π, 3π, 5π, 7π,...

The occurrence of this phenomenon for all waves, even waves created by vibrating strings and pebbles dropped into a pond, leads to this thought: *If I could generate a very stable, high-frequency, observable waveform, and somehow think of ways to make a process to be measured to cause the waves to interfere by changing the phase angle ϕ, I should have a very accurate and stable means of measuring the process. Furthermore, if I observe the interference at a fixed position, as time progresses and the process changes the phase angle, I should see light and dark bands move across my field of view.* Did early instrument designers ponder this same thought while observing interference patterns in nature?

The Doppler Effect

The Doppler effect is well known to anyone who has walked down a highway and heard the apparent pitch of a horn increase or decrease as the vehicle moves toward or away from you while blowing its horn. Doppler radar (radio detection and ranging) works on the same principle. Conventional radar emits a radio-frequency pulse and measures the time it takes for the echo to return. Thus it senses any object in the path of the broadcast beam, which can be directionalized. Doppler radar, on the other hand, depends on frequency shift to detect a moving object and thus only detects moving objects.

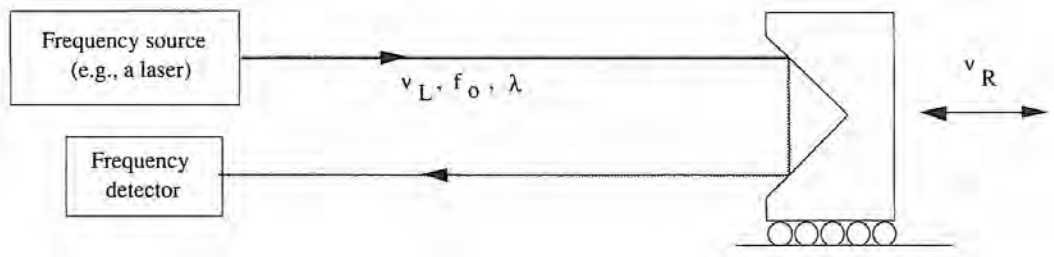


Figure 4.5.1 A frequency source and a detector used to determine the velocity of a moving target by the Doppler shift of the light beam.

As illustrated in Figure 4.5.1, a laser can be used as a stable directionalized frequency source whose beam of light travels to a retroreflector which reflects the incoming beam back toward the source parallel and laterally shifted to itself. Motion of the retroreflector at a velocity v_R will *stretch* the beam of light (of wavelength λ_o and frequency f_o) as it is being reflected. The detectors can use the effects of this phenomenon to determine the distance the retroreflector has traveled.

If a source of light (e.g., a laser) was fixed in space, then in a time interval t, a target (e.g., the retroreflector) would receive $v_L t/\lambda_o$ waves of wavelength λ_o (m/cycle) that were moving at speed v_L.

[21] Recall the trigonometric identity: $\sin\alpha + \sin\beta = 2\sin(\alpha/2 + \beta/2)\cos(\beta/2 - \alpha/2)$.

If the detector were moving toward the laser (via the retroreflector) at a velocity v_R, it would receive $v_R t/\lambda_o$ additional waves in the same time t. The frequency f (cycle/s) seen by the detector is the total number of waves received in time t:

$$f = \frac{v_L \frac{t}{\lambda_0} + v_R \frac{t}{\lambda_0}}{t} = \frac{v_L + v_R}{\lambda_0} = \frac{v_L + v_R}{v_L/f_0} \tag{4.5.4}$$

The change in frequency is f - f_o. For motion of the detector toward the laser, the change in frequency is positive, and for motion away from the source the change in frequency is negative:

$$\Delta f = \frac{v_R f_0}{v_L} \tag{4.5.5}$$

If the detector was stationary and the laser was moving toward it, the effect would be an apparent shortening of the wavelength. Assuming that the laser is emitting light at a frequency f_o, then if the laser is traveling at a velocity v_R, it moves a distance v_R/f_o in one cycle. The wavelength λ_o appears to be shortened by this amount, and thus the wavelength λ seen by the detector is

$$\lambda = \frac{v_L}{f_0} - \frac{v_R}{f_0} \tag{4.5.6}$$

The apparent frequency f as the laser moves toward the detector is v_L/λ and it appears to increase:

$$f = \frac{v_L f_0}{(v_L - v_R)} \tag{4.5.7}$$

For motion of the laser toward the detector the change in frequency is positive, and for motion away from the detector the change in frequency is negative:

$$\Delta f = \frac{v_R f_0}{v_L - v_R} \tag{4.5.8}$$

If the velocity of the mechanical components with respect to the velocity of the light is small (e.g., 1 m/s compared to 3×10^8 m/s), then the change in frequency is given by Equation 4.5.5. Thus the total change in frequency caused by the retroreflector moving toward (+) or away (-) from the detector, as seen by the detector in Figure 4.5.1, is

$$\Delta f = \frac{2 f_0 v_R}{v_L} \tag{4.5.9}$$

For a given time interval Δt, the retroreflector will move an amount $\Delta x = v_R \Delta t$, and the resultant phase shift $\Delta\phi$ in radians is just $2\pi \Delta f\, \Delta t$. The distance traveled by the retroreflector is thereby found by measuring the phase and using it in the relation

$$\Delta x = \frac{v_L \Delta\phi}{4\pi f_0} = \frac{\Delta\phi\lambda}{4\pi} \tag{4.5.10}$$

An instrument called a phase detector can be used to measure the phase difference between the laser beam as it leaves the laser head and the reflected beam. Since the value of $\Delta\phi$ will vary between 0 and 2π, a counter must be used to keep track of the number of crossings of these boundaries. Note that red light from a He:Ne laser has a wavelength of 632.8 nm, $f_o \approx 4.8 \times 10^{14}$ Hz, and a velocity $v_L \approx 3 \times 10^8$ m/s.

Another variation on the Doppler technique is to use two pairs of two frequency beams incident on a diffuse reflecting surface to measure motion orthogonal to the beams. This technique can be very useful for detecting small changes in length of a surface caused by thermal growth or mechanical loads.[22]

Beat Phenomena and Heterodyne Detection

Have you ever noticed that when two adjacent notes on the piano are struck simultaneously, their sound rises and falls periodically? Have you ever wondered why the sound of a twin engine

[22] M. Hercher et al., "Non-contact Laser Extensometer," paper presented at OE Lase'87 Conf. SPIE, Los Angeles, CA, Jan. 1987. This device is manufactured by OPTRA Inc., 66 Cherry Hill Dr., Beverly, MA 01915 (617) 921-2100.

propeller-driven aircraft changes periodically? These sounds are characteristics of the *beat phenomenon*[23] which provides a simple, accurate way to detect phase differences between two waves with nearly identical frequencies. As shown in Figure 4.5.2, two waves having nearly identical frequencies f_1, and f_2 and amplitudes Y_1 and Y_2 can be represented by

$$Y_1 = A\cos(2\pi f_1 t) \tag{4.5.11}$$

$$Y_2 = A\cos(2\pi f_2 t) \tag{4.5.12}$$

When added together, $Y_1 + Y_2$ yields

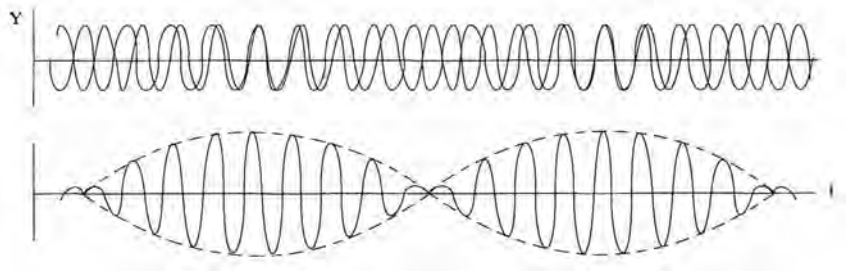


Figure 4.5.2 Two waves of nearly the same frequency superimposed to form a beat.

$$Y_{12} = 2A\cos\left(\frac{2\pi(f_2 - f_1)t}{2}\right)\cos\left(\frac{2\pi(f_1 + f_2)t}{2}\right) \tag{4.5.13}$$

The waveform has a frequency equal to the average of the two frequencies, and the frequency of the amplitude of the resultant waveform varies as a function of the difference of the two frequencies:

$$f_{\text{waveform}} = \frac{f_2 + f_1}{2} \qquad f_{\text{amplitude}} = \frac{f_2 - f_1}{2} \tag{4.5.14}$$

A beat (maximum amplitude) occurs whenever Equation 4.5.14 equals 1 or -1. Each of these values occurs twice in each cycle; thus the beat frequency is one-half the difference between the two frequencies: $(f_2 - f_1)/2$. This fundamental phenomenon allows the difference in two very fast, otherwise nearly immeasurable frequencies (e.g., interfering light waves) to be measured fairly easily by superimposing the waves and measuring the occurrence of intensity peaks. The use of the beat phenomenon to assist in the measurement of the frequency difference in two electromagnetic waves is known as *heterodyne detection.*[24] *dynamis*), meaning dynamic. Hence *heterodyne detection* is the detection of changes (a dynamic process) in frequency.

The Importance of Polarized Light

Previous sections have discussed methods for comparing the two frequencies of light in order to detect motion of a target. Practically, this requires the use of a reference beam and a measurement beam, both of which typically start out in phase and with their waveforms aligned in the same plane. For the beams to be handled easily, however, a method is needed to treat them as separate entities, even when they follow a common path. As will be discussed later, in order to maintain accuracy and reduce environmental errors, various optics used in interferometric measurement rely on the measurement and reference beams following a common path from the source to the region near where the measurement is made, and then back to the detector. The phenomenon of polarization makes the common path design goals possible.

All electromagnetic radiation (e.g., radio waves and light) is described by electromagnetic theory as being composed of transverse waves as shown in Figure 4.5.3: The *direction* of the vibrating electric **E** and magnetic **B** fields is at right angles to the direction of propagation. Natural light (e.g., from the sun or other luminous source) has **E** and **B** fields that are randomly oriented around

[23] The beat phenomenon can also be observed between two pendulums connected by a spring. See Section 3.7 of L. Meirovitch, Elements of Vibration Analysis, McGraw-Hill Book Co., New York, 1975.

[24] From *hetero,* meaning different, and *dyne,* (from Greek

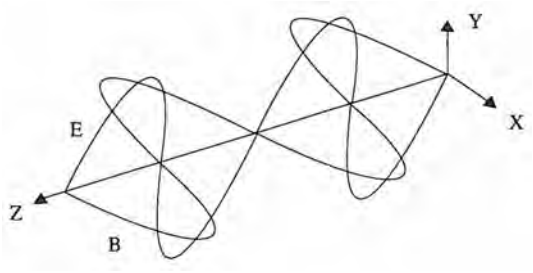


Figure 4.5.3 Plane-polarized light composed of electric (**E**) and magnetic (**B**) field vector components.

the direction of propagation, but always at right angles to each other. If natural light is incident on a special type of filter that has the characteristic of a fettuccini pasta maker, only the components of the light whose electric *resultant* **E** vectors are parallel to the plane of the noodles will be allowed through. The resultant waveform is shown in Figure 4.5.4. This type of filter is called a *polarizing filter.*[25] Instead of actual slits like a pasta maker, a polarizing filter relies on the atomic structure of a crystal lattice to transmit light oriented in only one plane. Note that if two Polaroids are placed on top of each other with their polarizing directions orthogonal, no light can get through. The intensity I of light transmitted through two sheets of a Polaroid is governed by *Malus's Law*[26]:

$$I = I_0 \cos^2 \theta \tag{4.5.15}$$

Malus's law can be used to determine if an object is a polarizer. If the two sheets of material are placed between a light source and a detector, the intensity of light transmitted through the sheets must eventually vanish as the sheets are rotated. Note that if the angle θ between two Polaroids is 45°, the intensity is cut in half. If you wear a pair of polarized glasses and rock your head from side to side while looking at reflected sunlight, you will notice that the intensity of light changes.[27]

As shown in Figure 4.5.4, when two light beams' electric field components $\mathbf{E_X}$ and $\mathbf{E_Y}$ are in phase, the resultant vector component **E** of light has a fixed orientation at a 45° angle to both the X and Y axes. Any plane polarized wave **E** can always be decomposed into orthogonal electric vectors $\mathbf{E_X}$ and $\mathbf{E_Y}$. As one moves along the Z axis, the orientation of the resultant **E** vector with respect to the X and Y axes does not change, but its magnitude goes to zero and then increases in the other direction. Plane-polarized light is thus referred to as light in the P-state.

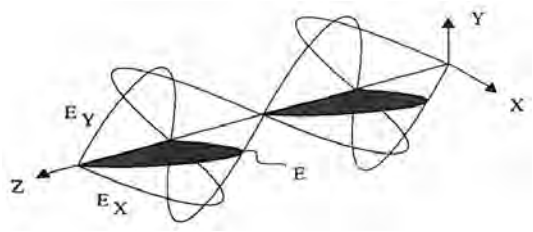


Figure 4.5.4 Plane-polarized light formed by two orthogonal in phase **E** vectors.

As the relative phase of the $\mathbf{E_X}$ and $\mathbf{E_Y}$ beams changes, their relative position along the Z axis changes and the *magnitude and direction* of the resultant vector **E** change. If their amplitudes are equal and the phase angle between the $\mathbf{E_X}$ and $\mathbf{E_Y}$ beams is $\phi = \pm\pi/2 + 2m\pi$, where m is any

[25] *Polaroid* is the commercial name of a sheet of polarizing material. For the curious of heart, see J. Walker's "The American Scientist" in Sci.Am., Dec. 1977.

[26] Malus's Law was discovered experimentally by Étienne Louis Malus (1775-1812) in 1809 while observing light reflected off the windows of the Luxembourg Palace in Paris. See how important it is to pay attention to the world around you instead of frying your brain with portable headphones!

[27] In 1812, Sir David Brewster (1781-1868) found that at a critical angle of incidence with a glass or other dielectric material (e.g. water, ice, car polish), all the components of a light beam except one polarized component are refracted (they travel into the material and do not reflect). A portion of the polarized beam is also refracted, but some is reflected. Brewster found that the angle between the refracted light and the reflected polarized component was equal to 90°. From Snell's law, $n_1 \sin\theta_p = n_2 \sin\theta_r$, he deduced that the critical angle of incidence was $\theta_p = \tan^{-1}(n_2/n_1)$.

real integer, then the magnitude of the **E** vector is constant and its direction continuously changes, thereby sweeping out a circle. This form is called *circularly polarized* light. Light can be right circularly polarized (i.e., the R-state corresponding to clockwise rotation of the **E** vector) or left circularly polarized (i.e., the L-state) depending on the sign of the $\pi/2$ term. As a matter of fact, plane and circularly polarized light are actually degenerate forms of *elliptically polarized* (i.e., the E-state) light.

In order to fully utilize the phenomena of interference and heterodyne detection in a modern interferometer, it is necessary to generate and separate two orthogonal P-states. One will be the measurement beam, and one will be the reference beam. By having two orthogonal P-states of different frequencies, the light can travel to a point near where the measurement is to be made, split, and then be recombined after the measurement beam returns from the target. However, before a phase difference measurement can be made using heterodyne detection techniques, the polarization angles of the P-states must then again be made equal (nonorthogonal), so they can add to form a beat.

These three tasks, generation, separation, and recombination, can all fundamentally be accomplished using crystals that have the property of *birefringence*. A birefringent crystal's[28] index of refraction depends on the polarization angle and frequency of the incident light with respect to the material's *optic axis*[29]; thus the material is optically anisotropic. P-state light, which is composed of $\mathbf{E_X}$ and $\mathbf{E_Y}$ components, will be refracted by the crystal and either be *split* into two single orthogonal P-state waves, or emerge as a single beam with a *potentially different polarization* character such as the E-state. Birefringent materials are difficult to use in the mass production of optical components, and thus whenever possible, special coatings on conventional substrates (e.g., glass) are used.

How can a birefringent crystal or a special coating discriminate between differently polarized light rays? The answer lies in the manner in which light is transmitted through a transparent media. Light propagates through a transparent substance by striking and exciting electrons in the substance. The light's **E** field causes the substance's electrons to vibrate, which in turn generates light waves that travel through the substance and excite other electrons. These waves are called *secondary wavelets*. The superposition of billions of little wavelets gives rise to the effect of the light wave passing through the crystal. The electrons act like little masses held to the nucleus by little springs. When the substance's atomic structure is arranged so that the stiffness of the springs varies with direction, then the amplitude of the vibration that a plane of light causes depends on the stiffness of the springs in its plane of vibration. Hence the incident plane of light can be bent (refracted) by the material if the wavelets formed by the $\mathbf{E_X}$ and $\mathbf{E_Y}$ components are different. Figure 4.5.5 shows a mechanical model of the electron cloud of an atom bound to the nucleus with soft and hard springs. The $\mathbf{E_X}$ electric field is shown vibrating in the plane of the hard springs. The natural frequency of the vibrating electrons in the XZ plane will thus be higher than those with the soft springs[30] in the XY and YZ planes. As a result, the secondary wavelets generated by the $\mathbf{E_X}$ component will travel through the media faster than those generated by the $\mathbf{E_Y}$ component. The X direction is hence defined as the orientation of the *optic axis* of the material.

The manner in which P-state light composed of $\mathbf{E_X}$ and $\mathbf{E_Y}$ components is refracted by a crystal depends on: (1) the shape of the crystal, (2) the orientation of the optic axis with respect to the crystal, (3) the angle of incidence of the light, and (4) the type of crystal. The components of P-state light incident to a birefringent crystal will appear to diverge when: (1) the crystal's front and back surfaces are not parallel, (2) the optic axis is not parallel to these surfaces, and (3) the light is not incident perpendicular to the front surface. Indeed, two components of light, which are orthogonally polarized P-states, will emerge from the crystal; however, remember that a light wave can always be broken down into two orthogonal vector components (i.e., $\mathbf{E_X}$ and $\mathbf{E_Y}$ components), including the "$\mathbf{E_X}$" or "$\mathbf{E_Y}$" components that have emerged from a birefringent crystal. Hence the emerging components are referred to as the *ordinary* (o-ray) and the *extraordinary*[31] (e-ray) rays. Remember, the magnitude of the two indices of refraction of a birefringent crystal depend on the polarization angle of the light; thus the two indices of refraction are referred to as n_o and n_e. By considering the

[28] Calcite (calcium carbonate, $CaCO_3$) ice, mica, quartz, and many other crystals are birefringent.

[29] Not to be confused with the *optical axis* of a lens.

[30] Remember that the natural frequency of a system is proportional to the square root of the stiffness divided by the mass.

[31] *Extra* as in "additional"; there is nothing particularly special about the second wave other than the fact that it is orthogonally polarized with respect to the first.

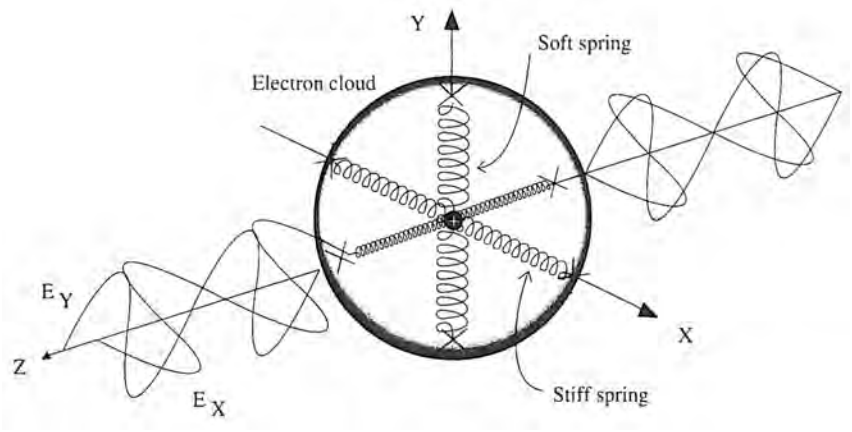


Figure 4.5.5 Mechanical vibration model of how a birefringent material's electron cloud changes the phase between two orthogonal **E** vectors. (After Hecht.)

three criteria for the behavior of P-state light as it hits a birefringent crystal, different shape crystals can be designed to make P-state light do almost anything. Prisms, alone or in combination, and slabs that are properly inclined to a beam of light can be made to polarize the light into orthogonal components that travel along the same path.

Light can also be made to travel through a birefringent material without being separated but with its polarization state changed. For example, assume that one has a crystal with two parallel surfaces and the optic axis is parallel to these surfaces. Also assume that P-state light is incident *orthogonal* to the optic axis of a crystal with the plane of the $\mathbf{E_X}$ component aligned with the hard springs, and the plane of the $\mathbf{E_Y}$ component aligned with the soft springs. The $\mathbf{E_X}$ component will travel through the media faster than the $\mathbf{E_Y}$ component. Since the light was orthogonal to the optic axis, the $\mathbf{E_X}$ and $\mathbf{E_Y}$ components will travel in the same path straight through the material without being refracted; however, the motion of the $\mathbf{E_Y}$ components will be retarded with respect to the motion of the $\mathbf{E_X}$ component. This causes a relative phase shift between the $\mathbf{E_X}$ and $\mathbf{E_Y}$ components, thus changing the polarization angle.

Birefringent materials can therefore be made to selectively change the polarization angle of light. When the thickness d of the material is just right, it causes a net phase shift of 90^o in the $\mathbf{E_X}$ and $\mathbf{E_Y}$ components (one quarter of a wave):

$$d = \frac{(4m+1)\lambda_0}{4|n_0 - n_e|} \qquad m = 0, 1, 2, 3 \ldots \tag{4.5.16}$$

This causes the light to change from the P-state to an R or L-state: Plane-polarized light is now circularly polarized. Hence the optical component that causes this effect is called a *quarter-wave plate*.

If the light is then passed through a second quarter-wave plate, another 90^o phase shift between the two components occurs and the R or L-state converts to a P-state that is orthogonal to the light's original P-state. This provides a way to change the polarization angle of a measuring beam to match that of a reference beam so they can then interfere. A measuring beam can be made to make two passes through a quarter-wave plate or one pass through a *half-wave plate*. The thickness of a half-wave plate must be

$$d = \frac{(2m+1)\lambda_0}{2(n_0 - n_e)} \qquad m = 0, 1, 2, 3 \ldots \tag{4.5.17}$$

Manufacturers of optics can implement strict quality control procedures; however, if the user is not careful to maintain required angles of incidence, then slight changes in the polarization state may occur. This can cause the relative magnitudes of a measurement beam and a reference beam to change slightly as well as allow polarization leakage and mixing between reference and measurement beams. These effects can give rise to a distortion of the overall combined waveform used

in a heterodyne detection process. Therefore, attention to detail is most important when setting up and aligning the optics of a laser interferometer measurement system. Remember, interferometric measurement systems are chosen for their ability to measure to the submicron level, a level at which most engineer's intuition does not work; hence procedure must be strictly followed.

Construction of Stable Light Sources

Previous discussions assumed that light was available in nice, neat, well-behaved waveforms; unfortunately, it is virtually impossible to have a light source of a single wavelength. There is no known true source of pure sine waves of a fixed frequency and amplitude. The real world does not allow for such things in the same manner that no measurement is exact. Both are a function of how much the user wants to pay in order to achieve a more accurate approximation.

The term *coherence* is used to describe how closely light from a source approaches being truly monochromatic (of a single frequency). The *coherence time*, Δt_c, is a measure of the time interval during which the phase of a lightwave can be predicted at a fixed point in space (i.e., the time it resembles a sinusoid). The coherence time is merely the inverse of the frequency bandwidth which is the total range of frequencies that comprise a light wave. *Temporal coherence* is a term used to describe light that has a large coherence time. Even though the coherence time is equal to the inverse of the frequency bandwidth of light, broadband white light can still be made to interfere over short distances or if it is focused. In fact, focused broadband sources were used with early interferometers, as discussed later. In general, the more coherent the light, the better the attainable fringe resolution from an interference pattern.

The *coherence length* represents the spatial interval over which a disturbance at the beginning of the spatial interval can be correlated to a disturbance at the end of the interval. For example, consider the surface of a pond during a calm day when the surface is still and smooth like a sheet of glass. If you drop a rock in on one side of the pond, the waves will travel across the pond to the other side. From there one could deduce that a rock was dropped in on the opposite side. On a windy day, however, the surface is a random collection of waves, and if you drop a rock in on one side it would not be evident from the other side of the pond. Similarly, how well a light source can be used to generate interference patterns is a function of its coherence. *Spatial coherence* is the degree of coherence perpendicular to the direction of travel.

As the coherence of a light source decreases, fringes formed by an interference pattern become less sharp and the resolution of the measurement system decreases. Recognition of this fact is all the machine design engineer needs at this point; sensor manufacturers are well aware of coherence design issues. The machine design engineer therefore needs only to make sure that the sensor is not used beyond its intended range without consulting the manufacturer.

A good way to obtain monochromatic light is to excite a substance whose characteristic emission spectrum is closely monochromatic. Every element has a characteristic emission spectra that is made up of different wavelengths of light and by filtering, selective wavelengths can be obtained. However, the best way to obtain monochromatic light is with a laser. All common substances are composed of atoms constructed from a nucleus and surrounding electron cloud that likes to maintain a minimum energy configuration, known as the ground state. When energy is supplied or pumped into the substance, the electrons are raised to an excited state and then drop back down, like the circulation of popcorn inside a hot air popper. When the electrons fall from the excited state to the ground state, they emit a photon. Unfortunately, light from a source like a regular incandescent bulb is composed of photons emitted at random from random energy levels; thus the light is incoherent.

In 1917, Einstein showed that an excited electron can fall from an excited state to a lower state by the mechanism described above. The electron can be dragged down to another specific intermediate state by electromagnetic radiation of a particular frequency. This is known as *stimulated emission*. The photon emitted by the falling electron will be *in phase, have the same polarization angle, and propagate in the same direction* as the stimulating radiation. In order to have a continuing chain reaction that produces a steady stream of light, however, all the electrons must first be raised to their excited state, creating a *population inversion*, and then be triggered by photons at the proper frequency. When triggered, a virtual avalanche of in-phase photons is created. As long as sufficient energy is pumped into the substance, the avalanche continues.

In the 1950s the detailed physics of this process was worked out and tested using microwaves, resulting in the *maser*.[32] It was but a matter of time before someone (Theodore Maiman, in 1960) applied the same principles to light, and thus was born the *laser*. In practice, the laser cavity can be made from a crystal, such as a ruby, or from a glass tube filled with a gas such as carbon dioxide or a mixture such as helium and neon. One end of the cavity is mirrored and the other end is only partially silvered. When the cavity is *pumped* by a source, typically by conventional light or electrical discharge, the lasing action builds upon itself to a high-intensity level. Since only the ends of the cavity are silvered, nonaxial direction light leaks out the sides and the remaining coherent light becomes focused along the length of the cavity. As the intensity builds, it eventually allows some of the light to leak out the partially silvered end of the cavity. The quality of the light is measured by the coherence and degree of collimation. As one would expect, the longer the optical cavity, the more coherent the light.

Note that in order to have very coherent light, the lightwaves both inside and outside the cavity must appear to be a continuous wave. Thus if one were to look at the entire waveform at its frequency of oscillation, one should see a stationary sinusoidal wave. In order to achieve this effect, the distance between the mirrors of the cavity must be closely controlled so that the waves bouncing back and forth between the mirrors constructively reinforce and stay in phase. During steady-state operation, the wave in the laser cavity and its reflected image can be considered to have the same amplitude, frequency, and velocity. Increases in energy (amplitude) of the wave as it traverses the tube and picks up energy from the lasing action are regulated by allowing part of the wave to leak it out the partially silvered end of the laser tube. The equations for wave and its reflection are

$$Y_1 = A\sin(kx - \omega t) \tag{4.5.18}$$

$$Y_2 = A\sin(kx + \omega t) \tag{4.5.19}$$

Their resultant in the lasing cavity is the linear superposition of the two:

$$Y = 2A\sin(kx)\cos(\omega t) \tag{4.5.20}$$

Note that there are points where the amplitude is always zero: kx = 0, π, 2π, 3π and so on. Since these points are always zero, only the region between them oscillates up and down as a function of time. These are called *standing waves*. Their maximum amplitude is a function of position along the wave. Once the wave leaves the laser cavity, however, there is no opposing wave, and the wave looks like a regular sine wave whose amplitude at a point is a function of time only.

For a cavity filled with a substance of refractive index n, in order to maintain standing waves, the length L of the cavity must be equal to an integer number (m) of half wavelengths:

$$L = \frac{m\lambda_0}{2n} \tag{4.5.21}$$

Note that there are an infinite number of standing waves that can be supported in the cavity, and since the frequency equals the velocity divided by the wavelength, the frequencies are

$$f_m = \frac{mc}{2Ln} \tag{4.5.22}$$

The frequency difference between any two modes m and m + 1 is

$$\Delta f = \frac{c}{2Ln} \tag{4.5.23}$$

Thus, in order for the laser to approach monochromaticity, the cavity length must be carefully controlled, the media carefully chosen, the energy state of the population inversion carefully controlled, and the frequency of the initial triggering wave must be monochromatic.

It would seem that the cards are stacked against the laser designer. Fortunately, however, there are some media whose electrons like to exist in specific excited energy levels and fall to specific lower, but not ground state, energy levels. The most widely used lasing medium for dimensional

[32] The 1964 Nobel Prize in Physics was awarded to Charles Townes (United States), Alexander Prokhorov (USSR), and Nikolai Basov (USSR) for their development of the technique of microwave amplification by stimulated emission of radiation (Maser).

metrology applications is a mixture of helium and neon. The Helium-Neon laser has a visible wavelength of 632.8 nm, which is bright red. The cavity length is often controlled using an electric heater. Two frequencies are generated in the lasing cavity and their relative intensity is used as a feedback parameter to control the cavity length. Alternatively, piezoelectric actuators can be used to move one of the mirrors at the end of the tube.

4.5.1 Optical Flats and Fizeau Interferometry

An optical flat is a plate of transparent material whose front and back surfaces deviate from a true plane by not more than λ/4. Typical commercially available optical flats are usually plane to within λ/10. Optical flats are used in conjunction with a quasimonochromatic light source as a quick accurate manual method to check the flatness of a surface visually. The only constraint is that the surface must be reflective, as in a finely finished piece of metal or another optical component.

Since the goal of surface flatness measurement is to determine deviation from a true surface, a method is needed to project an even wavefront of light to the surface from a stable reference. Then one can observe the interference patterns that result from parts of the wave traveling farther than others before they are reflected back. The optical flat itself acts as the reference plane from which a quasi-monochromatic light wavefront emanates. Any change in distance between the observer and the optical flat does not affect the light or its interference pattern. The same phenomenon that is responsible for this effect also produces the colored lines in soap bubbles; each fringe or color is representative of a line of constant thickness.

An optical flat is a piece of glass, with flat parallel surfaces, that is used to measure the thickness of a layer of air between the bottom surface and a part. The thickness of the air film trapped between an optical flat and the surface equals the height deviation of the surface. As quasimonochromatic light passes through the air gap and is reflected back across the air gap, it interferes with itself. If the optical path length through the air gap differs by an integer multiple of the half-wavelength of light (0, λ/2, λ, 3λ/2, 2λ, 5λ/2,...), then the reflected light's phase will invert when it is reflected. The reflected light will thus be phase shifted 180^{o} from the incident light and it will destructively interfere with the incident light, causing a dark band to appear. In other words, the distance between light and dark bands represents a change in gap between the optical flat and the surface by one-quarter wavelength. Note that when light reflects off a surface, it can be phase shifted by an amount that depends on the type of surface it is reflected from. Thus optical flats are useful as a tool for measuring the relative change in height across a surface.

Typically, one edge of the optical flat is made to contact the part and the other edge is lifted up so that a wedge of air is formed. Nonuniformity of the fringe spacing indicates peaks or valleys. Curved fringes also indicate deviation from flatness. Together they provide a contour map of the surface. By changing the location of the reference edge, one can determine the exact topography of the surface by reading the dark bands like lines on a topographic map. This takes skill and experience, but because of its overall simplicity, optical flats remain one of the most useful manual measuring tools available in the shop. Typical prices for λ/10 optical flats 80, 160, and 200 mm in diameter (3, 6, and 8 in.) are on the order of $450, $1100, and $1600, respectively.

The diffuse light source can be replaced by a coherent collimated source; and then the reference surface provided by the optical flat can be moved away from the surface. This is called a Fizeau interferometer after the types of fringes formed, and it is a very powerful metrological tool. The two-dimensional fringe pattern can then be projected onto a TV monitor or interpreted digitally. If the surface being analyzed is not flat, the fringes will appear warped like the lines on a topographical map. If the surface is flat but tilted with respect to the reference optics, the fringe spacing will be proportional to the angle of tilt. If the object is rotating, the fringes will rotate.[33]

4.5.2 Michelson and Optical Heterodyne Interferometers

When machine tool design engineers think of an interferometer for use in the measurement of machine tool parameters, they are generally referring to the type of interferometer developed by the

[33] See A. Gee et al., "Interferometric Monitoring of Spindle and Workpiece on an Ultraprecision Single-Point Diamond Facing Machine," Vol. 1015, SPIE Micromachining Optical Components and Precision Engineering, 1988, pp. 74-80.

American physicist Albert A. Michelson (1852-1931).[34] The knowledge of optics and how light waves could interfere with each other to form fringes was known before Michelson developed his interferometer. However, Michelson was the one that first pulled this knowledge together to design a measuring tool that used the concept of interference to measure displacement.[35] All length measurement standards are traceable to a standard which is now defined as the speed of light in a vacuum. A de facto standard is the wavelength of red light from an iodine stabilized Helium-Neon laser.[36] For most applications, a relatively inexpensive stabilized Helium-Neon laser is sufficient. Transferring this standard to mechanical measure is possible only through the use of interferometric methods. Thus at least from a design point of view, the historical development of Michelson's interferometer is well worth studying.

In the 1860s, the Scottish physicist James Clerk Maxwell developed his now famous equations describing the behavior of electromagnetic waves. Since all other waves required a medium for transmission (e.g., water or air), it was postulated that the transmission of light depended on a *light-carrying ether* that existed everywhere, including space; however, no one had been able to measure its physical properties (e.g., mass or even its presence). Michelson had been very active in performing experiments to determine the speed of light using the rotating toothed wheel technique developed by Fizeau in France in 1849. Michelson replaced the toothed wheel with a mirrored wheel and in 1880 was able to determine the speed of light to be about 299,910 km/s (accurate to 0.039%).[37] Because light moves so fast, Michelson reasoned that in order to measure the presence of the ether, light itself would have to be used to determine the properties of the ether.

Michelson proposed to measure the presence of the ether by measuring the effect that the speed of the Earth had on the ether as the Earth moved through it. To accomplish this task, Michelson reasoned that if he passed light across the path of the ether and compared it to light flowing with the ether, he would be able to detect the presence of the ether by then interchanging the two paths (rotating the experiment 90^{o}). Accomplishing this task would require devising a method to make the two beams interfere with each other and then looking for changes in the fringe's position as the experiment was rotated. Michelson's experimental apparatus is shown schematically in Figure 4.5.6. A ground-glass plate diffuses the light from the source and makes any light that reaches the experiment appear to be a uniform wavefront as opposed to a bright spot. The source light travels to a half-silvered mirror which lets half of the light through (the measurement beam) and sends half of the light off at a right angle (the reference beam). Because Michelson did not have a laser, he also had to contend with the very broad bandwidth of light that was generated. In order for broadband light to interfere, a wavefront must be split and then recombined such that the optical path length[38] differs by an odd multiple of $\lambda/2$. The allowable value of the odd multiple depends on the coherence of the light. Since Michelson did not have a laser, he had to make sure that the distances through air and glass that the measurement beam and reference beam traveled were the same. To equalize the optical path lengths, a compensator plate was placed in the measurement beam's path.

The diffuse light strikes the mirrors at all angles; hence the optical path difference of the different rays that are incident on the mirrors separated by a distance d over the angle θ is

$$\Delta\ell = 2\mathrm{d}\cos\theta \tag{4.5.24}$$

Sir George Stokes (1819-1903) first showed that light internally reflected from a beamsplitter is phase shifted by 180^{o} from light externally reflected by the beamsplitter[39] (i.e., the reference and measurement beams, respectively). Therefore, when the relation given by Equation 4.5.24 equals an integer multiple of the wavelength, interference fringes occur. For a fixed distance between the plates, this means that the occurrence of fringes is a function of the angle θ. The lens acts to focus parallel rays at a fixed point; hence when all the rays from the diffuse source are drawn, a pattern of circular fringes will be seen. As the plates move relative to each other, the fringes will move toward

[34] For a more detailed historical account of Albert Michelson's achievements, see L. Swenson, Jr., "Measuring The Immeasurable," Am. Heritage Invent. Tech., Fall 1987, pp. 42–49.

[35] "*But in Science the credit goes to the man who convinces the world, not to the man to whom the idea first occurs.*" Sir Francis Darwin.

[36] Approximately 6328 Å, which can be stable to several parts per billion.

[37] Michelson later repeated the experiment in 1926 and determined the speed of light to be 299,796 km/s (0.0012% error). See D. Halliday and R. Resnick, Physics, Parts I and II, John Wiley & Sons, New York, 1978, pp. 922–927.

[38] When determining the spatial relation between two light beams, one must consider that light travels at different speeds through different media.

[39] This effect is seen from equations known as *Stokes relations*. For derivations of this reflection phenomenon, see E. Hecht, Optics, 2nd ed., Addison-Wesley Publishing Co., Reading, MA, 1987.

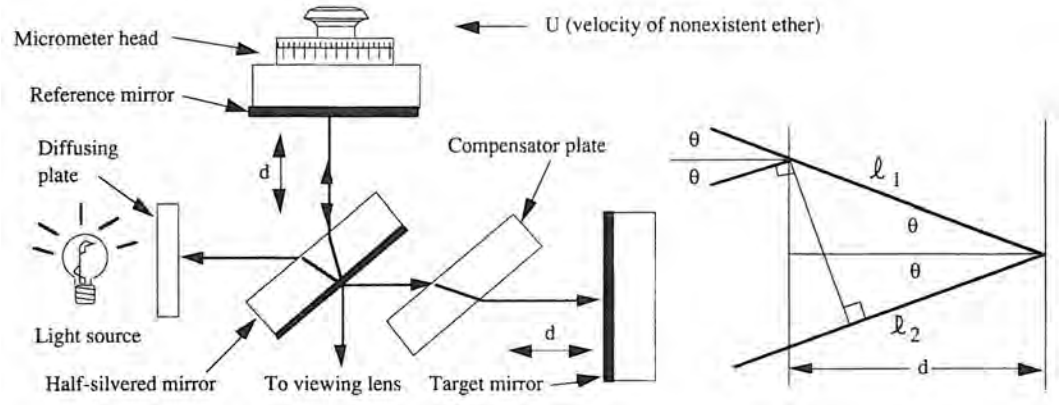


Figure 4.5.6 Illustration of Michelson's Interferometer for the purpose of determining if a light-carrying "ether" existed, and the interferometer's mathematical equivalent.

the center. The fringes are only one-half wavelength apart, making it difficult to see more than a few at a time, and these circular segments appear as parallel fringes.

The distance between successive dark bands corresponds to an optical path difference of λ. The motion of the target by $\lambda/2$ changes the optical path by λ; thus motion of the target can be resolved, without fringe interpolation, to $\lambda/2$. With a He:Ne light source, this is equivalent to about 0.32 μm (12.5 μin.). As the target starts to move, the number of light-dark fringe interferences that go by equals twice the velocity of the target divided by the wavelength of light. At a snail's pace of 1 mm/s, with He:Ne light, 3,161 counts per second must be made.

Once the relative optical path lengths in Michelson's apparatus were adjusted using the micrometer screw to fine tune the position of the reference mirror, circular fringes would appear at the detector (e.g., lens and eye combination). If the ether existed, and the earth was moving through it at a velocity u, the time it took for the light to travel across the stream of ether and back would be the distance (2d) divided by the velocity $[(c^2 - u^2)^{1/2}]$. Hence the time it took for the light to travel to the target mirror would be $d/(c + u) + d/(c + u)$. Expanding these expressions using the binomial theorem, it can be shown that the difference in time would have to be approximately $\Delta t = du^2/c^3$. So if the experiment were rotated 90^o while the fringe pattern was being observed, a phase shift would occur between the two beams.

Much to his surprise, Michelson could not detect any change in the fringe pattern, no matter how many times he repeated the experiment. This led him to collaborate with a friend and fellow scientist Edward Morley to construct an interferometer that bounced the light back and forth several dozen times between mirrors to increase the optical path length (distance d) to tens of meters. This acted to increase any effective optical path length change by a factor equal to the number of passes. Even with this enhanced resolution, they saw no indication of the presence of the ether. This seemed to indicate that there was no ether, and eventually led Einstein to figure out the problem. Although he failed to gain fame as the first person to prove or disprove the existence of the ether, Michelson did gain fame as the designer of the forerunner of modern interferometers. He went on to design interferometers for a wide range of applications, including interferometers for the determination of the distances of stellar objects from Earth. Michelson also won the 1907 Nobel Prize for Physics by showing that the standard meter was equal to 1,553,163.5 wavelengths of red cadmium light.

A modern *optical heterodyne interferometer*, shown schematically in Figure 4.5.7, uses Michelson's idea of splitting a beam into measuring and reference beam components; however, it is different from Michelson's interferometer in several very important ways, most of which owe their origin to development of modern electronics. Hewlett-Packard first commercialized optical heterodyne interferometers for general machine tool metrology use in the early 1970s. They made the interferometer system components available in modular form so that the user could easily build up a measuring system to meet his needs. Today, many other major suppliers of Optical Heterodyne Interferometers exist, each providing systems with their own special attributes. Thus the reader is

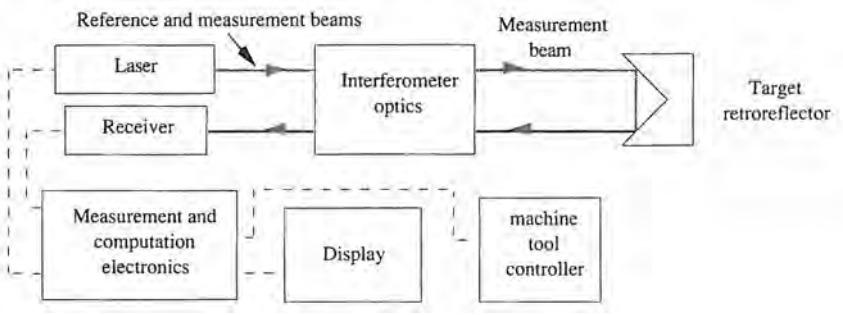


Figure 4.5.7 Principal components of an optical heterodyne interferometer.

cautioned to look carefully at all available systems and think about the design and operation of the system components as well as customer service before specifying a particular system.

The principal advances over Michelson's interferometer that are incorporated in modern optical heterodyne interferometers include:

1. Coherent laser light is used, so the optical path length of the measuring (or measurement) and reference beams does not have to be equal. Furthermore, the fringes are far less "fuzzy" than those obtained with noncoherent laser light sources.
2. The crisp, clear fringe patterns generated with laser light allow for the use of optical heterodyne detection techniques.
3. Beam distortion is minimized with modern optical components, so resolution and accuracy are increased. They also allow quantities such as linear displacement, angular rotation, straightness, squareness, parallelism, and relative refractive indices of gases to be accurately measured.

Two Frequency Laser Light Sources

In order for heterodyne detection techniques to be used and to provide ultrahigh resolution, the laser light source must emit two extremely stable frequency-shifted orthogonally polarized beams whose frequencies are known with great precision. A He:Ne laser head capable of meeting these requirements is shown in Figure 4.5.8. The lasing cavity emits two orthogonal linearly polarized beams that are frequency shifted by about 640 MHz. A partially silvered mirror is used to sample the beam, and the sample is separated by a birefringent prism. Detectors sense the intensities of two components of the sampled beam, and their output is used to control the temperature of the laser tube. By adjusting the laser tube temperature, the distance between the ends can be controlled and the frequencies stabilized to about 1 part in 10^8. With this method, frequency stability is in the range of 1 part in 10^7.

The 640 MHz frequency-shifted beam is used to help control the stability of the laser; however, the frequency difference is too large for practical use with present optical heterodyne detection techniques. A more acceptable frequency difference would be on the order of 20 MHz[40]; however, in order to obtain a 20 MHz frequency difference, a laser cavity about 7.5 m long would be required (Equation 4.5.23). A laser with this length of cavity would provide a very collimated coherent beam, but is impractical for most purposes. Thus, in order to generate the 20 MHz frequency-shifted beam, the output from the laser head is first put through a polarizing plate which blocks one of the two orthogonally polarized beams. The single remaining beam, which can still be resolved into two orthogonal $\mathbf{E}_X$ and $\mathbf{E}_Y$ components, is passed through an *acousto-optic frequency shifter.*[41]

An acousto-optic frequency shifter, shown schematically in Figure 4.5.9, uses a glass block onto which a piezoelectric transducer is cemented. The transducer is driven by a 20 MHz crystal oscillator, causing a travelling acoustic wave in the glass block which alters its refractive index. As the beam passes through the glass block, half of the light is transmitted unaffected while the other half is diffracted at a small angle and upshifted in frequency by 20 MHz. After the beams exit the acousto-optic frequency shifter, they immediately enter the birefringent combining prism before they

[40] This is the frequency used by Zygo Corp.'s laser.
[41] There are other methods that can be used, such as Zeeman splitting.

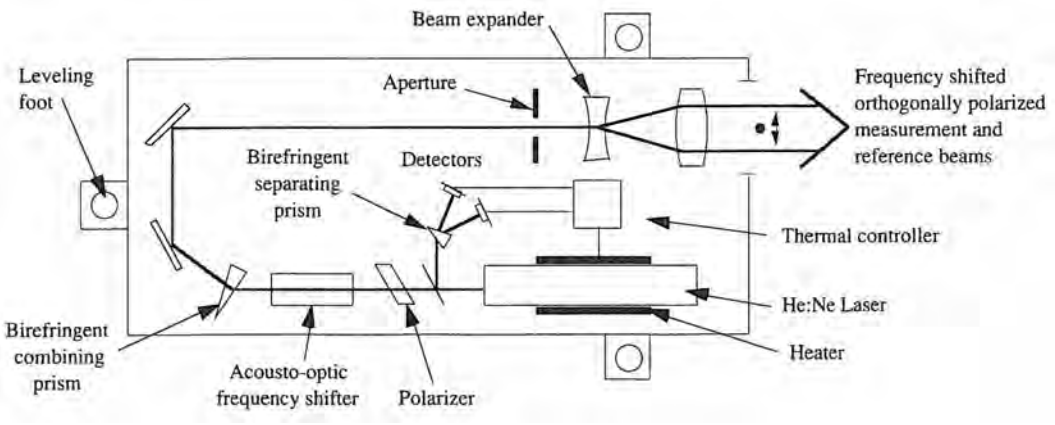


Figure 4.5.8 Schematic drawing of a laser head used with an optical heterodyne interferometer. (Courtesy of Zygo Corp.)

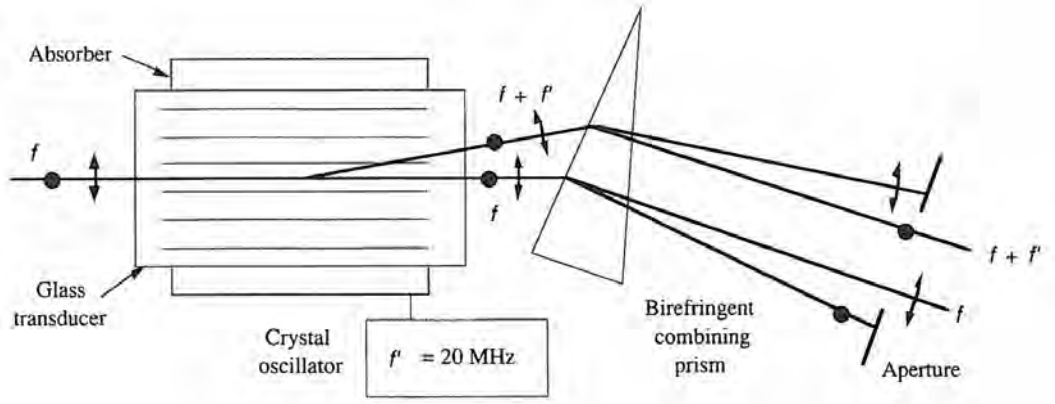


Figure 4.5.9 Acousto-optic frequency shifter used to phase shift a pair of orthogonally polarized laser beams. (Courtesy of Zygo Corp.)

diverge appreciably. The prism refracts one polarization component of the undiffracted beam and the orthogonal polarization component of the diffracted beam so they leave the prism collinearly. Although the beams are actually separated by a fixed distance, for all practical applications they can be considered as one beam. The beam is then expanded and collimated before leaving the laser head. The oscillator's 20 MHz signal is also used as a reference with which to compare the phase of the beat frequency signal generated by the interference between the Doppler-shifted measurement beam and the reference beam.

Electronic and Optical Heterodyne Detection[42]

If modern high-speed electronics were used to observe the fringes from Michelson's original interferometer design, higher resolutions than $\lambda/8$ could be obtained. Instead of using just one photodiode to sense the interference fringes, the interfering reference and measurement beams can be split into two equal components. Half of the interfering beams are directed to a photodetector and the other half is retarded 90^{o} by passing it through a quarter-wave plate before it is sensed by a second photodiode. As a result, sine and cosine signals are generated by the photodiodes, respectively. As with optical encoders, this allows for interpolation and quadrature multiplication. Resolutions of $\lambda/100$ and better are therefore possible. As the speed of the target is increased and the fringes flashed by at an ever-increasing rate, the resolution of this type of system decreases. In addition, as the velocity of the target increases, the measuring beam becomes Doppler shifted and the fringe pattern changes. This problem can be overcome with optical heterodyne detection.

[42] For a detailed mathematical treatment of heterodyne detection, see Section 14.6 of H. Haus, Waves and Fields in Optoelectronics, Prentice Hall, Englewood Cliffs, NJ, 1984.

Initially, the reference and measuring beams from the previously described laser head recombine in the interferometer optic to form a resultant beat frequency wave with a nominal frequency of 20 MHz. As the velocity of the target changes over a range of ± 1.8 m/s (70 in./s), the phase changes by ± 6 MHz. Current electronics technology allows for extremely accurate edge detection of waves within a restricted bandwidth, in the range of 10-15 MHz. Thus if the position of the edges of the wave produced by the beat phenomena could be compared to the position of the edges of a wave produced by a stable oscillator, an accurate assessment of the frequency of the beat wave can be made. Using Equation 4.5.10, the distance the target mirror traveled can thus be accurately determined independent of the velocity or acceleration of the target mirror. In fact, the frequency response of the electronics used is flat (0 dB) out to the maximum velocity, and then drops off very rapidly. If one were to try merely to detect the edge of the measurement beam wave as it was Doppler shifted by the target mirror, a detector with frequency response of 12 MHz centered about a nominal frequency of 4.8×10^{14} Hz would be required. The waveform produced by the beat phenomenon has all the wonderful characteristics of high-frequency laser light (i.e., it is visible, and collimated); however, it has an effective "wavelength" for detection purposes that is $4.8 \times 10^{14}/20 \times 10^{6} = 24 \times 10^{6}$ times "longer" than the original light beam.

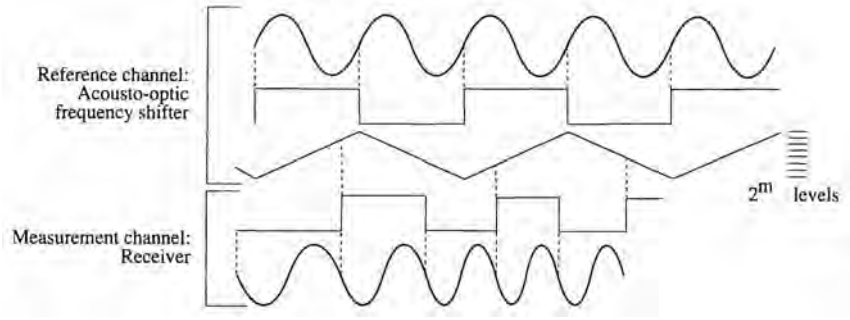


Figure 4.5.10 A method for detecting changes in optical path length using phase measurement. (Courtesy of Zygo Corp.)

As shown in Figure 4.5.10, for Zygo Corp.'s interferometer the 20 MHz stable reference signal from the crystal oscillator and the beat wave are both converted into square waves. The reference square wave from the oscillator is then integrated to yield a triangular wave. This is done for two reasons: (1) it is relatively easy to generate a very accurate square wave from a sinusoidal wave, and (2) it is very easy to discretize the intensity level of a triangular wave so that resolution along the wavelength will be constant, whereas the sensitivity of intensity detection of a sinusoidal wave varies with position along the wave. The rising edge of the beat phenomenon beam's square wave is then used to trigger an analog-to-digital converter that measures the amplitude of the reference triangular wave. If the target mirror does not move, there will be no Doppler shift of the measuring beam and the amplitude will be constant. As the target mirror moves, the rising edge position will change, which is a measure of the beat frequency or phase shift between the two beams. Furthermore, since one is using only the midpoint (zero crossing) between the peak and valley of the beat phenomenon beam's intensity as a trigger, intensity variations in the measuring beam will not affect the resolution or accuracy of the measurement. Note that there are many other ways to measure phase shift between two waves.

If the frequency shift between the reference and measurement beams were made smaller than 20 MHz, higher resolution could possibly be obtained; however, the allowable velocity of the target mirror would decrease in direct proportion. This is due to the fact that the Doppler frequency shift caused by motion of the target mirror must be kept below the amount of intentional frequency shift between the measuring and reference beam. This is an observance of the Nyquist sampling theorem, which prevents aliasing. Furthermore, as far as the analog-to-digital converter used to discretize the triangular waveform is concerned, a point is reached where giving it more time to discretize a process will not always appreciably increase its accuracy.

4.5.2.1 Beam Handling Components

At the time this book was written, two major suppliers of off-the-shelf Michelson-type optical heterodyne interferometers for use in servo-controlled machines were Hewlett-Packard Corp.[43] and Zygo Corp.[44] Most of their optics for interferometric measurement are interchangeable; the user must consider what is already at hand and what type of measurement is to be made before specifying a particular manufacturer's product. Keep in mind that optics such as plane mirrors, beam benders, beamsplitters, and retroreflectors are also available from a host of other sources; one must merely check the specified accuracy.

Beambenders

The function of the beambender is to bend the light through 90° as shown in Figure 4.5.11. It is comprised of a plane mirror mounted inside a hollow square cube, which is sometimes made from Invar for temperature stability. A typical high-quality mounted beambender costs about $450.

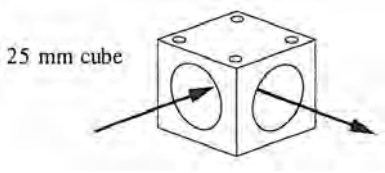


Figure 4.5.11 Cube type beambender (fold mirror).

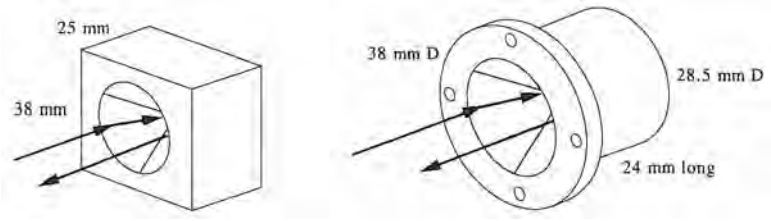


Figure 4.5.12 Block and flange mount retroreflectors (corner cubes).

Retroreflectors

A retroreflector is an optically ground and polished tetrahedral prism that reflects an incoming laser back parallel to itself at a separation distance twice that of the incoming beams' distance from the corner apex. The prism is typically mounted inside a metal (e.g., Invar or stainless steel) cube as shown in Figure 4.5.12. Lateral motion of the retroreflector causes no net change on the optical path length. If the retroreflector moves up by δ, the distance from the laser to the retroreflector and back decreases by an amount 0.707δ, but the distance between the beams also increases by 0.707δ. A similar argument can be made with regard to rotations of the retroreflector. Laser light is always returned parallel to its incoming path, and only axial motion of the retroreflector changes the optical path length. Errors in manufacture of the component can affect the path length; however, as long as the ingoing and outgoing beams are parallel to one another, these errors are generally negligible. A typical high-quality linear retroreflector mounted in a metal cube costs about $600.

Beamsplitters

A single laser head usually has enough power, on the order of 0.5 mW, for several measurement axes; thus a method is needed for dividing up the beam. The beamsplitter accomplishes this task with the use of two prisms cemented together, as shown schematically in Figure 4.5.13. The amount of light that is split off from the main beam depends on the type of

[43] For more detailed discussions and methods for setting up Hewlett-Packard optics, see Hewlett-Packard Corp.'s 5528A Laser Measurement System User's Guide, available from your local HP representative.

[44] For more detailed discussions and methods for setting up Zygo optics, see Zygo Corp.'s Axiom 2/20 Laser Measurement System User's Guide, available from Zygo Corp., Middlefield, CT.

interface between the two prisms. The phenomenon is called *frustrated total internal reflection* (FTIR). Both 33% and 50% beamsplitters are available, as well as more complex models that can break up a single incoming beam into numerous components. Note that beamsplitters can be made to separate (or to not separate) light of different polarities. When used as a device to separate the orthogonally polarized reference and measurement beams of an interferometer, a *polarizing beamsplitter* is used. A typical high-quality 50% beamsplitter mounted in a metal cube costs about $500.

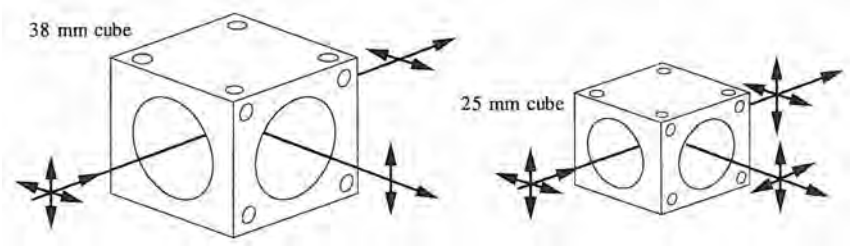


Figure 4.5.13 Polarizing and nonpolarizing beamsplitters.

Linear Displacement Interferometers

An optical heterodyne interferometer is very sensitive to any phenomena (e.g., temperature gradients) that can cause a relative phase shift between the measurement beam and the reference beam. Hence a method is needed whereby two essentially collinear beams[45] can be brought near the point of measurement, split into measurement and reference beams, and then recombined.

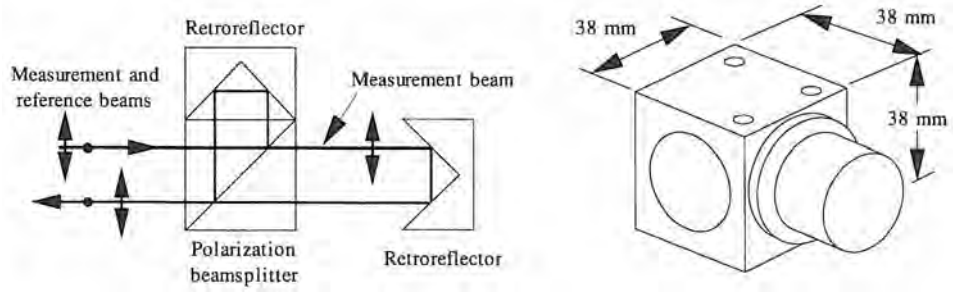


Figure 4.5.14 Linear displacement interferometer.

The linear displacement interferometer is made up of a polarizing beamsplitter and a retroreflector as shown in Figure 4.5.14. The reference beam, which is polarized perpendicular to the plane of incidence of the beamsplitter interface, is reflected at the interface. The reference beam exits the beamsplitter and is incident upon the retroreflector. The retroreflector sends the beam back into the beamsplitter, where it is again reflected from the interface, and it then emerges in an opposite but parallel direction to the incoming beam. The measuring beam, which is polarized 90° with respect to the reference beam and is in the plane of incidence, is transmitted straight through the beamsplitter to the retroreflector. The retroreflector sends the measuring beam back through the beamsplitter, where it recombines with the reference beam. Both beams then proceed to the detector. A typical high quality linear displacement interferometer mounted in a metal cube costs about $2000.

Plane Mirror Interferometer

Plane mirror interferometers are commonly used with XY motion plattens such as those found on wafer steppers. Two plane mirrors are placed at right angles to each other and an interferometer is used with each mirror. Figure 4.5.15 shows the beam path through the interferometer. The measurement and reference beams have a common path, so thermal expansion of the interferometer

[45] The beams must be collinear up to near the point a measurement is to be made in order to prevent environmental effects from causing a relative phase shift between the beams; if the beams were not collinear, then even small turbulent gradients in the air could cause a time-varying phase shift between them.

optics causes minimal errors. The optical path length changes by twice the amount the mirror moves (see Equation 4.5.25). Earlier designs did not always incorporate common path optics and thus they suffered from significant thermal growth errors. Note that there is still the potential for a significant deadpath error. A typical high-quality plane mirror interferometer (not including the target mirror) costs about $4000.

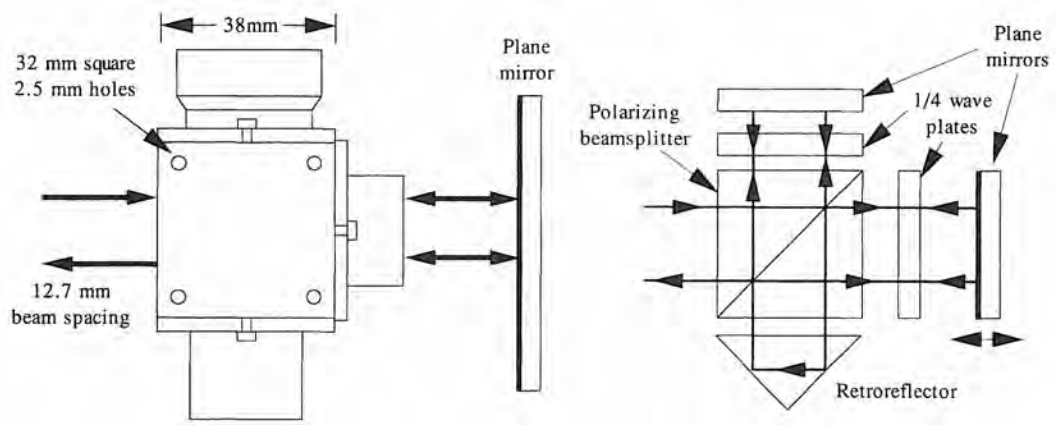


Figure 4.5.15 Plane mirror interferometer with common beam paths for temperature stability. (Courtesy of Zygo Corp.)

Differential Plane Mirror Interferometer (DPMI)

There are different versions of the DPMI,[46] but the overall goal is to minimize thermal growth errors experienced by other interferometer designs by providing for a common path of the measurement and reference beams through the interferometer. Because the beams travel along a common path, any thermal distortion of the device will affect the reference and measurement beams equally and the effects will cancel. When used in conjunction with different reflecting optics, a differential plane mirror interferometer can serve as a platform for the measurement of displacement, angular motion, and straightness.

As shown in Figure 4.5.16, the incoming beam, which is composed of two orthogonally linearly polarized frequency-shifted components, is split by the polarization shear plate. The half-wave plate rotates one component, so they both have the same linear polarization angle. This allows both beam components to pass through the polarization beamsplitter. The measuring beam passes through a quarter-wave plate, out to a moving-target mirror, and back through the quarter-wave plate. Its polarization angle is thus rotated by 90^{o}, and the polarization beamsplitter reflects it down to the retroreflector. The retroreflector sends the measurement beam back into the polarization beamsplitter, but displaced from the incoming beam. The measuring beam is reflected off to the target mirror, making one pass through the quarter-wave plate. On the return trip it makes one more pass through the quarter-wave plate. Its polarization angle is once again rotated 90^{o}, allowing the beam to pass through the polarization beam splitter toward the polarization shear plate. The reference beam makes a similar trip, although it reflects off of a reference mirror instead of the target mirror. The beams are recombined by the polarization shearplate and proceed back to the detector. If one traces the path of the measurement and reference beams in the interferometer, one will find that the same distance is traveled. Therefore, if the interferometer experiences uniform thermal growth, no optical path difference will exist between the measuring and reference beams in the interferometer.

Measurement of linear or angular displacement depends on how the paths of the beams are directed from the interferometer to the target and reference mirrors, as shown in Figures 4.5.16 and 4.5.17, respectively. For the linear displacement arrangement, the optical path change l_{opc} caused by a target mirror displacement δ is 4δ. Note that this is twice the optical path length sensitivity of a regular linear displacement interferometer. Accordingly, the maximum allowable velocity of the target mirror is reduced by half. The number of counts recorded by the system electronics is a

[46] See G. Siddall and R. Baldwin, "Some Recent Developments in Laser Interferometry," Proc. NATO Adv. Study Inst. Opt. Metrol., Viana do Castelo, Portugal, July 16–27, 1984, pp. 69–83.

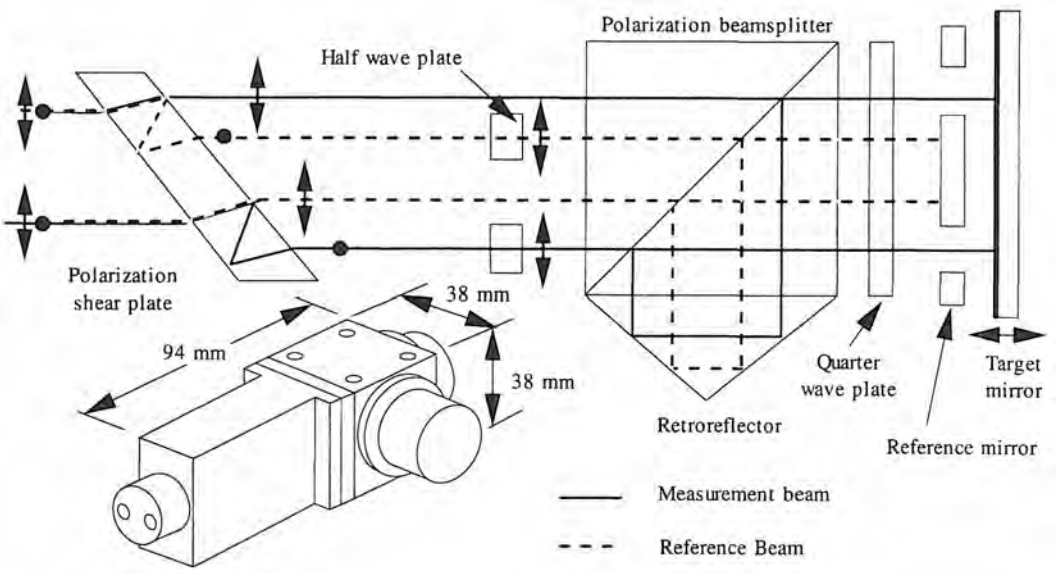


Figure 4.5.16 Differential plane mirror interferometer for linear displacement measurements, with common beam paths for temperature stability. (Courtesy of Zygo Corp.)

function of the optical path length change and the discretization level M. So for a motion δ of the target mirror, the number of counts recorded by the electronics will be

$$N = \frac{4\delta}{(2^M - 1)\lambda} \tag{4.5.25}$$

Typically M = 8 bits, and the last bit is noise, so the resolution is $\lambda/508 \approx 12.5$. For measurement

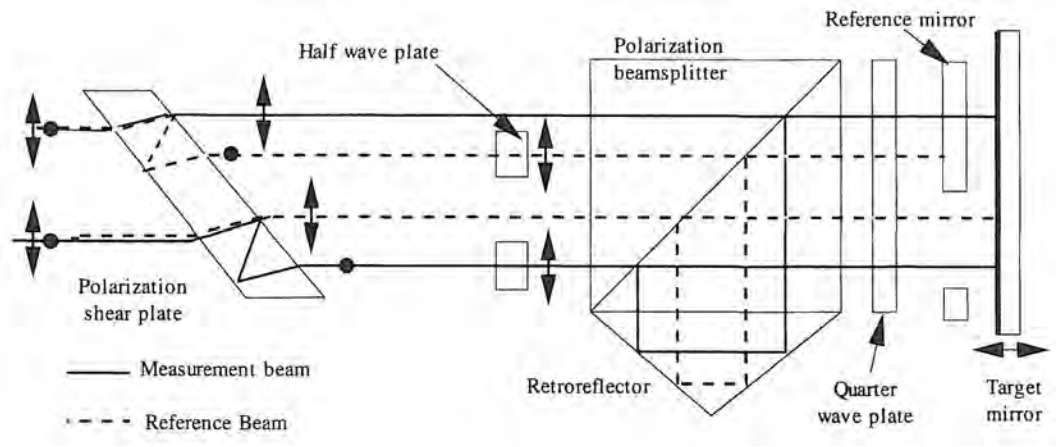


Figure 4.5.17 Differential plane mirror interferometer for angular displacement measurements, with common beam paths for temperature stability. (Courtesy of Zygo Corp.)

of angular displacement, as shown in Figure 4.5.17, any rotation of the target mirror about the X axis changes the angle of incidence of the two beams on the polarization shear plate, effectively changing the relative optical path length when the two beams are recombined. Rotations about the Y axis also cause a change in optical path length; however, this error is only on the order of 0.057 arcsecond (0.28 μrad) for θ_Y rotations of up to 2 arcminutes (582 μrad). In order to measure rotation about the Y axis, the entire configuration would have to be rotated about the Z axis by 90°. The number of counts generated by the electronics is approximately equal to $N = 3.65\theta_X$ where θ_X is in microradians. The error associated with this approximation is about one count per

2300 μrad. If greater range or accuracy is required, then an exact equation can be used.[47] Using the angular interferometer, straightness measurements can also be made by integrating a series of angular measurements as is done with autocollimators. This requires that the target mirror be moved an equal distance each time.

In addition to its versatility, the body of the differential plane mirror interferometer can easily be separated from the reference mirror. This allows the interferometer to be located away from the process, and a small reference mirror to be located near the process, thus minimizing space requirements near the point of action. Typical cost of a high-quality DPMI with reference mirrors for linear or angular displacement measurement is on the order of $6000.

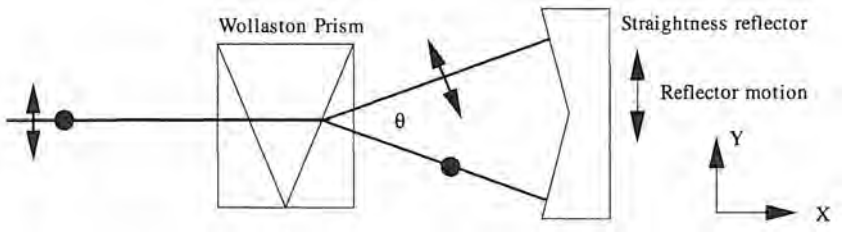


Figure 4.5.18 Hewlett-Packard's straightness interferometer measurement system. (Courtesy of Hewlett-Packard Co.)

Straightness Interferometer and Reflector

One manufacturer's straightness optics are shown schematically in Figure 4.5.18, and they must be used as a matched pair so that the reflector will return the two frequency components directly back to the interferometer. The interferometer contains a Wollaston prism, which has a different index of refraction for each of the two perpendicular polarity components of the laser beam. The Wollaston prism splits the beam from the laser head into its two components which travel to the reflector along precisely controlled paths. The orientation of the plane of the two exit paths is adjusted by turning the interferometer so that vertical or horizontal straightness can be measured. The reflector contains two plane mirrors which reflect the beam components back along their respective paths to the interferometer. Either component can be fixed or attached to the moving target. Initially, the two beam paths have the same relative length, but Y direction motion ΔY will cause the path lengths to differ by an amount $2\Delta Y \sin(\theta/2)$.

Another type of straightness interferometer is based on a differential plane mirror interferometer, as shown in Figure 4.5.19. The DPMI is modified by replacing the linear retroreflector with a split retroreflector, the reference mirror is replaced with a straightness prism assembly, and the plane target mirror is replaced with a straightness mirror assembly similar to that used by the HP system. The straightness prism is mounted to the moving body. The straightness ΔY is related to the number of counts from the electronics by

$$\Delta Y = \frac{N\lambda}{1016 \sin(\alpha/2)} \tag{4.5.26}$$

where $\alpha/2$ is the half angle engraved on the straightness mirror assembly. $\alpha/2$ has a nominal value of 1.8°, so the minimum detectable straightness error is about 0.02 μm (0.8 μin.). Limitations on the mirror flatness and knowledge of the angle $\alpha/2$ give a practical limit of the accuracy to about 0.3-0.5 μm (12-20 μin.) over a range of linear travel of about 3 m. To compensate for mirror flatness errors, the straightness mirror assembly can be rotated 180° and the measurements can be retaken and averaged with corresponding values from the previous pass. Pitch, yaw, and roll of the straightness prism assembly do not appreciably affect the straightness measurement with respect to the nodal point of the prism assembly. From this one can see why the preferred method for measuring straightness with an interferometer is to use a plane mirror interferometer and a long straightedge that has been lapped to a high degree of straightness.

With either system, an initial error in alignment will seem to cause a steadily increasing ΔY error as the optics are moved along the X axis, but this is easily subtracted off using a first order

[47] For a detailed analysis and the "exact" equations, see G. Sommargren, "A New Laser Measurement System for Precision Metrology," Precis. Eng., Vol. 9, No. 4, 1987, pp. 179–184.

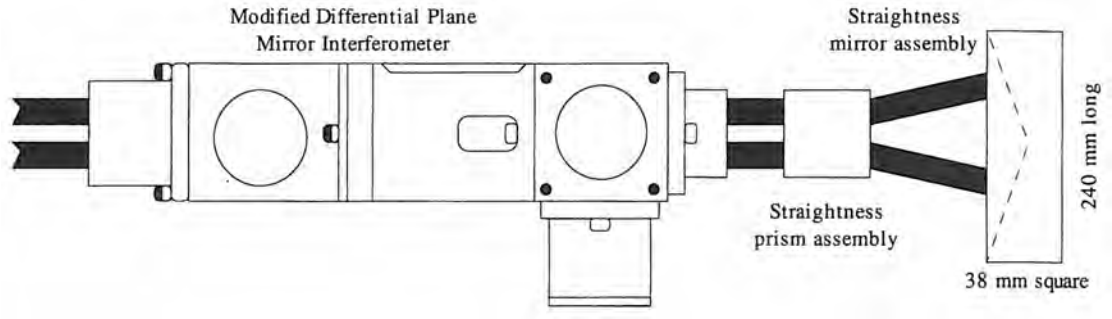


Figure 4.5.19 Zygo's straightness interferometer measurement system. (Courtesy of Zygo Corp.)

curve fit routine. Variations in the straightness optics setup can be used to measure squareness and parallelism. Typical cost of a straightness measurement system is on the order of $9000.

Linear/Angular Displacement Interferometer

A two-axis stage's XY position and yaw are typically measured as shown schematically in Figure 4.5.20. Yaw is determined by the ratio of the difference in X and X' measurements to their separation distance. The principal problems associated with the yaw measurement include: large amount of real estate occupied by the X' measuring system, coordinating information from the receivers to yield a low noise measurement, and stability and knowledge of the separation distance.

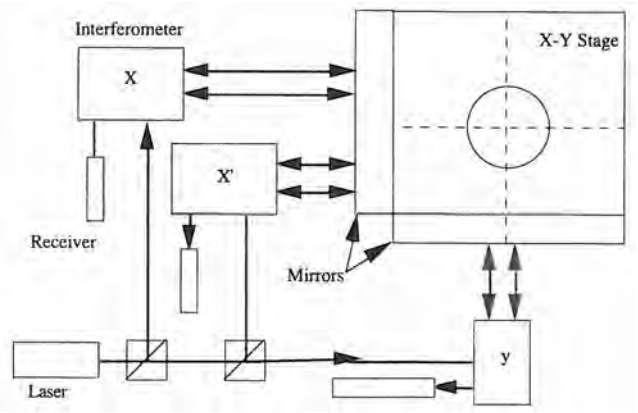


Figure 4.5.20 Typical method for wafer stage metrology using a laser measurement system (Courtesy of Zygo Corp.).

These problems are overcome by the linear/angular displacement interferometer[48] shown in Figure 4.5.21. The errors described above are eliminated by this interferometer's monolithic athermal design and the direct measurement of yaw by differential plane mirror interferometry. Fiber optic links also allow for the remote location of the receivers, which minimizes required space and thermal input to the stage system. This interferometer is essentially the combination into a single unit of linear and angular differential plane mirror interferometers described earlier, so resolution and accuracy are governed by the same equations. The approximate cost of this device, not including the receivers or stage mirror, is $9,000.

Measurement Receivers

The measurement receiver is shown in Figure 4.5.22. It converts the beat frequency wave of the measurement and reference beams into the series of square-wave pulses. These pulses are sent to the electronics, where they are used in conjunction with the reference signal from the laser

[48] U.S. Patent 4,733,967 by Zygo Corp. See G. Sommargren, "Linear/Angular Displacement Interferometer for Wafer Stage Metrology," SPIE Symp. Microlithog., San Jose, CA, Feb. 1989.

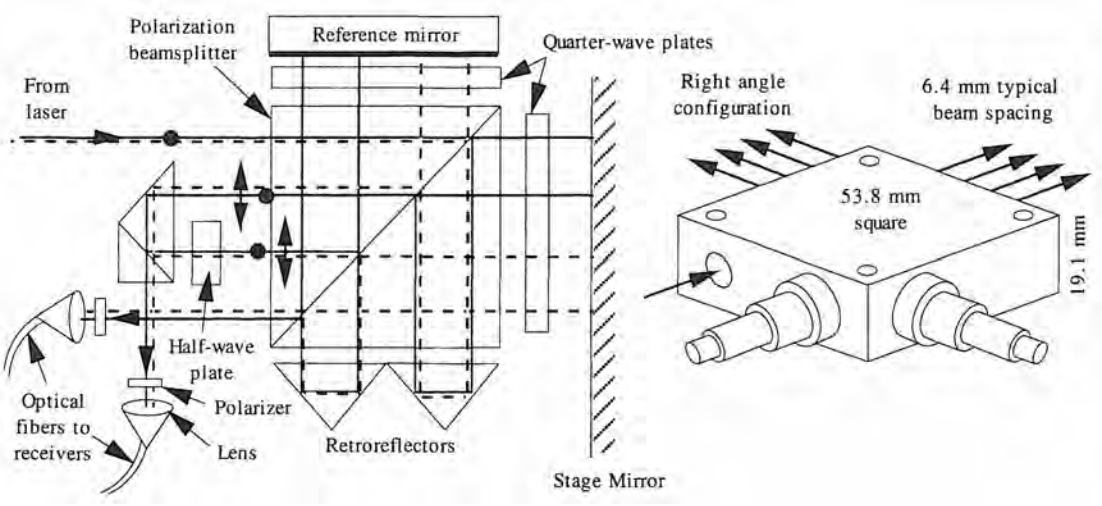


Figure 4.5.21 Zygo's linear/angular displacement interferometer. (Courtesy of Zygo Corp.)

head to determine the phase shift of the measurement beam. One receiver is needed for each axis of measurement, and it is possible to bring the beams back to the receiver from the interferometer optics using a fiber optic cable. A single receiver costs about $700.

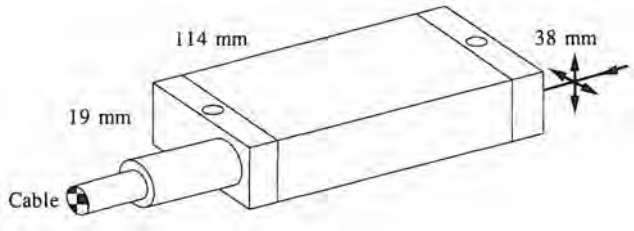


Figure 4.5.22 Laser interferometer's measurement receiver. (Courtesy of Zygo Corp.)

Refractometers

Laser interferometric measurements rely on a good knowledge of the atmosphere's refractive index (see the discussion below). Relative changes in the refractive index can be measured using the device shown in Figure 4.5.23. The device uses a differential plane mirror interferometer where the reference mirror and target mirrors are one and the same, but the reference beams are enclosed in evacuated quartz tubes. Only the physical path length of the measurement and reference beams are equal. A change in the atmosphere's refractive index will cause a change in the optical path length. By interferometrically comparing the reference beam (in vacuum) to the measuring beam, the change in optical path length can thus be determined. A refractometer of this type costs about $10,000. As discussed below, an Edlen's box must still be used to establish the initial conditions.

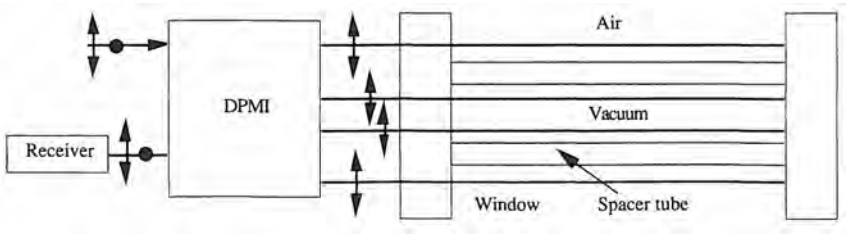


Figure 4.5.23 Refractometer for measuring relative changes in the refractive index of the operating environment. (Courtesy of Zygo Corp.)

4.5.2.2 Sources of Error

The error resulting from distance measurement using light as the waveform is a function of the accuracy of the knowledge of the media's refractive index n, light wavelength λ, detector's rms electronic noise $< \Phi_n^2 >^{1/2}$, and the misalignment angle θ[49]:

$$\frac{\Delta x}{x} = +\frac{\Delta n}{n} + \frac{\Delta \lambda}{\lambda} + \frac{\left\langle \Phi_n^2 \right\rangle^{1/2} \lambda}{4\pi x} + \frac{\theta^2}{2} \tag{4.5.27}$$

In addition, errors are introduced into the system by imperfections in the optics and their alignment which contribute to the types of errors above.

Refractive Index Errors

Since the accuracy of an interferometric measurement depends on the stability of the wavelength of light, any phenomenon that affects the wavelength of light will affect the accuracy. Temperature, pressure, humidity, and gas composition affect the refractive index of light in the measurement area, with the environmental error being about 1 μm per meter of measuring beam path length per degree Celcius temperature change. The relation between the refractive index, temperature, pressure, and humidity is given by Edlen's equation.[50] A modified version of this equation is given by Schellekens[51] as

$$n - 1 = \frac{2.879294 \times 10^{-9}(1 + 0.54 \times 10^{-6}(C - 300))P}{1 + 0.003671 \times T} - 0.42063 \times 10^{-9} \times F \tag{4.5.28}$$

where C is the CO_2 content in ppm, F is the water vapor pressure in Pa, P is the air pressure in Pa, and T is the air temperature in °C. The effect of errors in C, F, P, and T on the refractive index are found from the respective partial derivatives of Equation 4.5.28:

$$\frac{\partial n}{\partial C} = \frac{1.55482 \times 10^{-15} \times P}{1 + 0.003671 \times T} \text{ppm}^{-1} \approx 1.45 \times 10^{-10} \text{ppm}^{-1} \tag{4.5.29a}$$

$$\frac{\partial n}{\partial F} = -4.2063 \times 10^{-10} \times \text{Pa}^{-1} \tag{4.5.29b}$$

$$\frac{\partial n}{\partial P} = \frac{2.87929 \times 10^{-10}(1 + 5.4 \times 10^{-7}(C - 300))}{1 + 0.003671 \times T} \text{Pa}^{-1} \approx 2.67 \times 10^{-9} \times \text{Pa}^{-1} \tag{4.5.29c}$$

$$\frac{\partial n}{\partial T} = \frac{-1.05699 \times 10^{-11}(1 + 5.4 \times 10^{-7}(C - 300))P}{(1 + 0.003671 \times T)^2} \text{K}^{-1} \approx -9.20 \times 10^{-7} \times \text{K}^{-1} \tag{4.5.29d}$$

From these equations, it can be seen that the temperature is the most critical parameter.

The best way to measure the effects of relative changes in these quantities on the refractive index from an initial startup time is to use a refractive index measuring device (*refractometer*). There are numerous ways in which such devices can be constructed. A common device for measuring parameters used to determine the absolute index of refraction is an *Edlen's box*, which is a set of instruments that measures air pressure, temperature, and humidity. An Edlen's box, however, does not take into account atmospheric composition and local changes in the refractive index along the measurement beam path. Environmental factors and changes in gas composition can be determined by using an absolute refractometer that is constructed essentially like the refractometer shown in Figure 4.5.23 with the addition of means to allow the atmosphere to circulate through the evacuated tube, and then to pump out the air once again. With this method, the refractive index can be determined with an uncertainty of about 5×10^{-8}.

However, it is impractical to place a refractometer along every beampath. Some machines use oil showers for cutting lubrication and temperature control, and hydrocarbon vapors greatly change the refractive index of air. Even air turbulence caused by air showers, which are often used for temperature control of a machine, can cause errors if a precision machine's interferometric measurement system is not enclosed. Table 4.5.1 shows experimental results for a 175 mm measurement beam path.[52] In these experiments performed by Bobroff, the nozzle flows simulated the effect of turbu-

[49] C. Wang, "Laser Doppler Displacement Measurement," Lasers Optron., Sept. 1987, pp. 69–71.

[50] B. Edlen, "The Refractive Index of Air," Metrologia, Vol. 2, No. 2, 1965, pp. 71–80.

[51] P. Schellekens et al., "Design and Results of a New Interference Refractometer Based on a Commercially Available Laser Interferometer," Ann. CIRP, Vol. 35, No. 1, 1986, pp. 387–391.

[52] From N. Bobroff, "Residual Errors in Laser Interferometry from Air Turbulence and Non-linearity," Appl. Opt., Vol. 26, No. 13, 1987, pp. 2676–2681.

Beampath condition	Rms optical path fluctuation (Å)
Enclosed	4
Unenclosed	15
0.5 m/s	24
1.0 m/s	45
0.5 m/s nozzle	24
1.0 m/s nozzle	45

Table 4.5.1 Observed rms optical path fluctuations in a 175 mm measurement beam path. (After Bobroff.)

lent flow caused by air shower flow as it flows around optical mounts. Thus it is sometimes wise to enclose the beam paths in evacuated, or more practically, in helium-filled pathways.[53]

The index of refraction is also dependent on the wavelength of light. An experimental two wavelength (244 nm and 488 nm) argon laser interferometer was shown to be insensitive to changes in the index of refraction caused by atmospheric effects.[54] Because the beams shared a common beam path, there was not a question of local effects, such as turbulence, not being accounted for. At present, this type of system is too expensive and delicate to compete with He:Ne-based systems; however, as higher-accuracy machines become needed and the impracticality of enclosing all beam paths in a vacuum continues to manifest itself, dual-wavelength interferometers may become more commonplace.

Changes in the refractive index cause a cumulative error over the path the measurement beam makes. The *deadpath* error is the error incurred in the region traversed by the measuring beam that is not traversed by the object being measured. For submicron systems, the deadpath error is usually due to local gradients, so measuring the temperature or barometric pressure in the room does not provide sufficient information to compensate for this type of error. By placing the interferometer or reference mirror as close to the end of travel of the target as possible, environmental effects on the untraveled region of the measuring beam's path can be minimized. If there is a large deadpath, then in addition to increasing the effects of environmental changes on the optical path length of the measuring beam, dimensional changes of the machine can occur in this region. Depending on the position of the optics, the measured motion will be that of the target plus the motion caused by thermal growth of the material between the target and the reference mirror. Once again, there are many advantages to using enclosed beam paths.

Light Wavelength Errors

The accuracy to which the wavelength is known directly affects the accuracy of the determination of distance. The wavelength in a vacuum of He:Ne laser light may be known to 1 part in 10^9-10^{10} for iodine-stabilized lasers and 1 part in 10^7-10^8 for commercially available He:Ne lasers. Frequency stabilities of 1 part in 10^{12} are obtainable with Helium-Neon lasers that are stabilized with respect to the emission spectra of a stable iodine isotope; however, this type of laser is expensive and delicate. On the other hand, stabilities over a period of a few years of 1 part in 10^7-10^{10} are possible with a He:Ne laser stabilized, for example, using the Zeeman effect[55] and are readily commercially available. Short-term (months) stability of a stabilized He:Ne laser can be 1 part in 10^8-10^{10}. The laser can also be calibrated periodically with respect to an iodine-stabilized laser. Unstabilized or free-running lasers have been found by the National Institute of Standards and Technology to be stable to about 1 part in 10^5-10^6, which is sufficient for some applications.[56]

The generation of stable coherent laser light is a self-renewing process that requires some of the "old light" to be used to generate "new light." It is the property of light generated by stimulated emission that the new light generated is in phase with and has the same polarization angle of the old

[53] See J. B. Bryan, "Design and Construction of an Ultraprecision 84 Inch Diamond Turning Machine," Precis. Eng., Vol. 1, No. 1, 1979, pp. 13–17, and J. B. Bryan and D. L. Carter, "Design of a New Error-Corrected Coordinate Measuring Machine," Precis. Eng., Vol. 1, No. 3, 1979, pp. 125–138.

[54] See A. Ishida, "Elimination of Air Turbulence Induced Errors by Nanometer Laser Interferometry Measurements," 1989 Joint ASPE Annu. Meet. & Int. Precis. Eng. Symp., Monterey, CA, pp. 13–16.

[55] When the atoms of a light source are subjected to a magnetic field, the energy levels the atoms fall from to emit photons are divided into two levels. Thus the emitted light contains two frequency components f_1 and f_2, whose frequency is proportional to the strength of the magnetic field. The two frequencies form a beat frequency (see Section 5.3.5.3) which is detected and used to stabilize the laser frequency.

[56] This discussion on laser stability is from conversations with Prof. R. Hocken of the University of North Carolina at Charlotte.

light. However, just like a plumbing system where a little bit of fluid can leak back into the system even with a check valve in place, some of an interferometer's Doppler-shifted measuring beam works its way back into the laser cavity by means of partial reflection and transmission. This results in the generation of noise within the laser. Fortunately, other real-world effects, such as partial absorption, keep the relative amounts of this noise low. As discussed in Section 5.7.1, there are methods for filtering the laser light to prevent feedback of frequency shifted light into the laser cavity.[57]

Electrical Noise Errors

The amount of electrical noise in the detector that is used to detect phase typically corresponds to an rms phase noise $< \Phi^2{}_n >^{1/2}$ of about 1 part in 10^3. Using Equation 4.5.27, a He:Ne interferometer system will thus typically have an electrical noise error component of about 0.5 Å.

Alignment Errors

If "exact" linear displacement is to be determined, it must be measured at the toolpoint or the angular displacement must also be measured to account for Abbe errors. In addition, if the measurement beam is not parallel with the axis of motion by an amount θ, then an error equal to $\ell(1 - \cos\theta)$ will occur where l is the length of travel. In most cases θ will be less than a tenth of a degree, so the error can be accurately approximated by $\ell\theta^2/2$. Similarly, measurement of angular displacements of magnitude γ must be done about the correct axis or an error of $\gamma\theta^2/2$ will occur. These types of errors are called *cosine* errors. For a large precision diamond turning machine where ℓ may be equal to 1 m of travel, in order to keep the misalignment error below 100 Å (0.4 μin.), the misalignment angle must be less than 141 μrad (0.141 mm/m). This is achievable, but requires great care when setting up the optics. If the misalignment error is to be less than 0.1 Å over a range of 0.1 m, say for an atomic resolution measuring machine, the misalignment error must be less than 4.5 μrad, which is not easy to attain.

Optical Component Errors

Surface finish effects and coating property variations on the polarization beamsplitter can allow some of the reference beam to travel to the target and be Doppler shifted. Similarly, some of the measuring beams may leak to the retroreflector in the interferometer. This creates noise in the optical signal seen by the receivers and limits the resolution at which the zero crossing of the signal can be detected. In practice, the signal quality errors limit displacement resolution with a plane mirror interferometer or DPMI to about 10 - 20 Å. Methods do exist for keeping the beams separate in order to avoid this problem. Until recently there has been no commercial justification for development of systems with higher resolutions.

With the increasing popularity of scanning tunneling microscopes, there has been a realization that soon calibrations with Å resolution over a range of millimeters will be required.[58] To make these measurements with an interferometer will require the design of new optics. This need has led to the development of a crystal DPMI by Zygo Corp.[59] This interferometer uses birefringent materials to achieve the beam separation and bending requirements; hence coatings are not required. Stable resolutions of $\lambda/1000$ were obtained with the prototype and $\lambda/4000$ is deemed possible without too much difficulty. The cost of the crystal interferometer should also be within reason. In addition, high resolution is required so that digital servos' recursive algorithms can be executed without generating excessive differentiation noise. Typically, for a digital servo to be effective, the sensor feedback resolution must be at least 10 times greater than the desired position resolution of the machine. Ultimately, as resolution requirements increase, X-ray interferometers[60] may have to be developed for commercial applications.

Alignment errors in the beam handling optics can also cause polarization mixing of the reference and measuring beams. When the angle of incidence between the laser and an optical component made from a birefringent material (e.g., a polarization beamsplitter) is not exact, some leakage of the two frequencies occurs. This is referred to as polarization leakage and it contributes to the overall

[57] See A. Rosenbluth and N. Bobroff, "Optical Sources of Nonlinearity in Heterodyne Interferometers," 1989 Joint ASPE Annu. Meet. & Int. Precis. Eng. Symp., Monterey, CA, pp. 57–61.

[58] See Section 8.8.2.

[59] G. Sommargren, "Linear Displacement Crystal Interferometer," 1989 Joint ASPE Annu. Meet. & Int. Precis. Eng. Symp., Monterey, CA., paper unpublished. For information, contact Zygo Corp.

[60] D. Bowen "Sub-nanometer Transducer Characterization by X-ray Interferometry," 1989 Joint ASPE Annu. Meet. & Int. Precis. Eng. Symp., Monterey, CA.

noise seen by the system. In addition, as the light beams are directed around corners using beam benders, small changes in the polarization angle can result if the beams are not bent by exactly 90°. This effect occurs because the light beam is not a single line, but it is a rod of finite cross section. So if the light is not incident on the reflecting surface at exactly 45°, the inner part of the beam will be slightly retarded with respect to the outer part by a fractional portion of a wavelength. Just like a wave plate changes the polarization angle by retarding one of the beam's components, bending a beam through an angle other than exactly 90° can cause a small change in the polarization angle. As long as the deviation from 90° is less than about a tenth of a degree (1.7 mm/m) this effect can be ignored for most applications. Overall, optical nonlinearities may contribute a fixed error into the system on the order of 5 nm.[61]

The optics and their mounts will also change in size with temperature. This can result in a change in optical path length between the reference and measurement beams. Recall that the DPMI discussed earlier was symmetrical and had equal path lengths for the reference and measurement beams. The index of refraction of the glass used in interferometer optics will also vary with temperature and can cause measurement errors with a period of hours or days. Since some of the laser energy is continually dissipated in the optics, it is important to provide a means for controlling their temperature. Air showers would introduce more error than they would help, so in many cases one can only mount the optics to a large temperature controlled thermal mass and let the system come to equilibrium.

Although they are not encountered as often as thermal effects, the effects of strong magnetic fields on interferometric measurements may have to be considered if interferometers are to be used in close proximity to magnetic bearings or electric motors. A magnetic field will not affect the behavior of light itself, but it can ever so slightly change the refractive index of dielectric materials (e.g., quartz) which are often used in the beam handling optics. As long as the measurement and reference beams share a common path through the optics, however, even a strong varying magnetic field should not affect the accuracy of the system unless it causes increased polarization mixing.

Error Summary

Measurement errors must be incorporated into the system's error budget as discussed in Section 4.2. Recall that the purpose of the error budget was to allow the design engineer to pinpoint the dominant errors in a system and thus allow the dominant errors to be targeted for reduction. Interferometers, like all other systems, are subject to real-world limitations. Once identified, many of the nonlinearities can be corrected for using appropriate optical and electrical filtering techniques. Laser interferometers have been used in harsh machine tool environments with great success, provided that they are protected from cutting fluids, harsh temperatures and so on. In general, if conditions are so bad that a way cannot be devised to protect the optics and beam path, then the rest of the machine will probably be distorting so badly that the resolution and accuracy of an interferometer is not needed in the first place.

4.5.2.3 Typical Characteristics of Michelson-Type Optical Heterodyne Interferometers

The following summary of Michelson-type optical heterodyne interferometer characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as "gospel." Furthermore, it is extremely important to note that accuracy is highly dependent on the manner in which the optics are mounted and how the environment is controlled.

Size: Laser head, 125 mm high × 140 mm wide × 450 mm long. See Figures 4.5.11-4.5.22 for sizes of beam handling components. The main electronics box comes in two parts, each about 25 × 50 × 50 cm.

Cost: About $10,000 for the laser head and support electronics for up to four axes of measurement. For each axis of measurement, the cost of the optics, mounts, and additional electronics is about $6000.

Measuring Range: Up to 30 m.

[61] See, for example, C. Steinmetz, "Accuracy in Laser Interferometer Measurement Systems," Lasers Optron., June 1988. Also see W. Hou and G. Wilkening "Investigation and Compensation of the Nonlinearity of Heterodyne Interferometers," Progress in Precision Engineering, P. Seyfried, et al. (Eds.), Springer-Verlag, New York, 1991, pp. 1–14.

Accuracy: In a vacuum, the accuracy is usually a function of the alignment. If perfectly aligned and used in a vacuum, the accuracy can be on the order of half (worse) the resolution. In nonvacuum conditions, the environment greatly affects accuracy, as discussed previously.

Repeatability: Depends on the stability of the environment and the laser head, but can be as repeatable as the resolution.

Resolution: Depends on optics used, but can be as high as λ/508 (12 Å = 0.05 μin.). Look for resolutions of λ/1024 and λ/4096 in the future as better optics (e.g., crystal interferometers) and phase measurement techniques evolve. Of course, this resolution does not mean anything unless the mechanical system is designed to be able to respond accordingly.

Environmental Effects on Accuracy: About 1 μm/m/C^{o}. The user should also consider the effects of air turbulence and thermal expansion of the optics, mounts, and the machine itself.

Life: Essentially infinite, although for ultraprecision applications where nanometer accuracy is sought, one must consider the design of the optics being used, for optics can fog ever so slightly over a period of many years of continuous use and cause the center of beam intensity to shift slightly. For some measurements, this may cause small cosine errors. In addition, the laser head must be checked periodically against an iodine-stabilized laser.

Frequency Response: Frequency response is typically flat (0 dB) out up to the maximum velocity of the system (1.8 m/s for the Zygo system). At higher speeds, an error signal is generated. Note that resolution is independent of speed, and for some systems, any acceleration is allowable as long as the maximum velocity is not exceeded.

Starting Force: Essentially zero.

Allowable Operating Environment: To maintain accuracy, the system should ideally be used in a vacuum, or in air at 20^{o}C with no gradients. Consult with the manufacturer for operation in other environments.

Shock Resistance: For precision measurements at high frequencies, one must consider that inertial effects will cause deformation of the optics. In general, it is best not to mount the electronics or laser head in a high-g environment unless arrangements have been made with the manufacturer to harden the system.

Misalignment Tolerance: Misalignment causes cosine errors and loss of beam power. Misalignment also increases polarization leakage which degrades accuracy.

Support Electronics: Extensive support electronics are required. Systems are available for interface to machine tool controllers which do not allow for manual reading. Systems are also available for manual reading of the data (LED display).

4.5.3 Laser Interferometer-Based Tracking Systems

Laser interferometer-based tracking systems, such as developed by Hocken and Lau,[62] have the greatest potential resolution of any externally based endpoint measurement system (1 part in 10^5). Accuracy is situation dependent and is limited in a large part by temperature gradients causing the laser beam to bend (see Equation 4.2.10). These systems can have a bandwidth on the order of 10-100 Hz, which is fast enough to measure the endpoint position of a robot or coordinate measuring machine once it reaches its final destination, but it is only marginally fast enough to use as a position feedback signal for continuous path motion; however, one might expect that soon servo-level frequencies will be available. Hocken and Lau's device consists of a laser interferometer used to measure the distance from a two-axis gimbaled platform to a two-axis gimbaled retroreflector mounted near the toolpoint of a machine or robot. The laser interferometer measures the distance to the retroreflector and precision encoders measure the inclination angles. A beamsplitter placed in front of the retroreflector diverts part of the beam to a lateral effect diode. Output from the diode is used to keep the servo-controlled gimbals positioned so that the laser stays centered on the retroreflector. Another type of interferometric-based tracking system uses several lasers to measure distances to a collection of cat's eye retroreflectors on the target. The distances are used to determine

[62] This system was developed at NIST in 1985 by K. Lau, R. Hocken, and W. Haight. The results are presented in "An Automatic Laser Tracking Interferometer System for Robot Metrology," Precis. Eng., Vol. 8, No. 1, 1986, pp. 3–8.

the position and orientation of the target.[63] Both of these types of systems are finding applicability with large (room-size) coordinate measuring machines used to inspect large complex parts such as submarine propellers.

4.5.4 Mach-Zehnder Interferometers

A Mach-Zehnder interferometer operates on the same principle as a Michelson interferometer. Like the Michelson interferometer, there is the original classic version that uses broadband light from a source that shines through a diffusing plate, and there is a modern optical heterodyne version. As shown in Figure 4.5.24, a beamsplitter divides light from a source into a measuring beam and a reference beam. The reference beam is guided along an environmentally stable path. Meanwhile, the measuring beam is directed along a path that shines the beam through a process, such as a chemical reaction or fluid flow. The two beams are brought back together with the use of another beamsplitter, where they interfere to form a fringe pattern. As the process changes, the optical path length of the measurement beam changes, and the fringe pattern shows this change. In addition to straight fringes merely moving across the eyepiece, patterns reflecting the nature of the process occurring within the cross section of the beam can be seen.

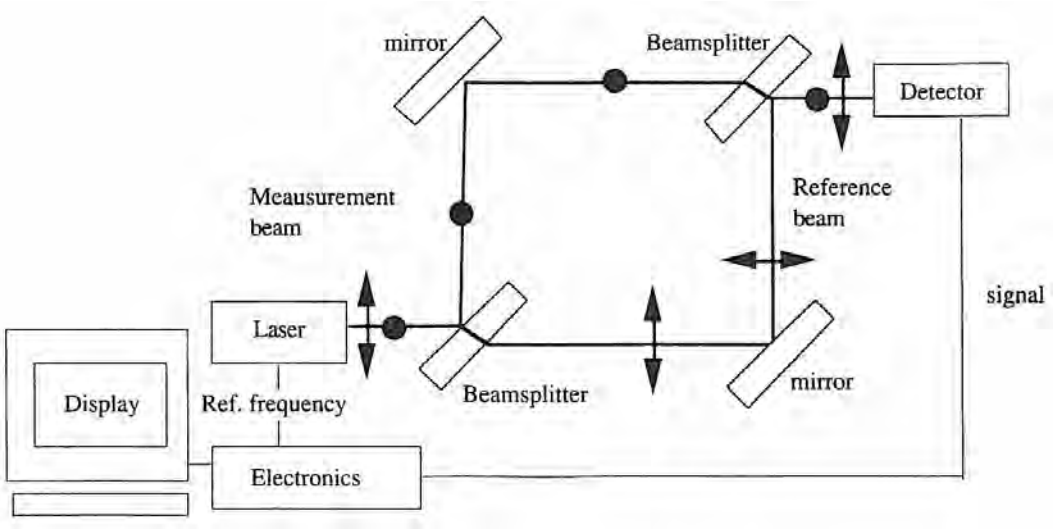


Figure 4.5.24 Operating principle of an optical heterodyne-type Mach-Zehnder interferometer.

If the process takes place on a scale much larger than the beam, then optical heterodyne techniques can be used to increase the resolution and speed of the measurements; however, only one process variable will be extractable from the system. For the optical heterodyne version, a measuring beam and a frequency-shifted reference beam are generated by the laser head. The beams are split by polarization beamsplitters and then recombined. The recombined beams form a beat frequency wave which is then compared to the output from the laser head (e.g., the output from an acousto-optic frequency shifter). A measure of how the process is affecting the measurement beam is then attained with the use of phase measurement techniques, as described previously. In fact, with the use of appropriate optics, a Michelson interferometer can be easily converted into a Mach-Zehnder Interferometer.

4.5.5 Sagnac Interferometers

A Sagnac interferometer in the form of a *ring laser gyroscope*[64] is illustrated in Figure 4.5.25. It has an optical cavity design that allows two beams of light to travel the same path in a ring, with one

[63] This laser interferometer triangulation system is available from Chesapeake Laser Systems Inc., Lanham, MD.
[64] Fiber optic gyroscopes also operate on the Sagnac principal; they differ from ring laser gyroscopes in that they use an external laser source.

beam going clockwise and the other beam going counterclockwise. A triangle is the simplest shape for the cavity, although any number of mirrors that cause the light to follow a closed path will do. The Sagnac effect is exhibited when the device is rotated and the fringe pattern formed by the beams appears to move with respect to the body of the interferometer. In actuality, the standing waves produced in the laser cavity remain fixed in space, and the cavity rotates around them. A device such as a photodiode can then be used to observe the motion of the wave nodes relative to the cavity.

The cavity is often filled with a mixture of helium and neon gas. Electrical discharge between the cathode and the anodes provides the energy input necessary for the lasing action to occur. Although not shown here, like the He:Ne laser head used for a Michelson interferometer, some of the beam can be bled off and used to control thermal heaters or piezoelectric actuators on the mirrors to keep the cavity tuned to the proper frequency. When the mirrors are spaced just right, standing waves are produced and noncoherent portions of light radiate out of the laser. The standing waves have nodes that are fixed in space; when the mechanical assembly rotates, the photodetector can see the nodes pass by beneath it.

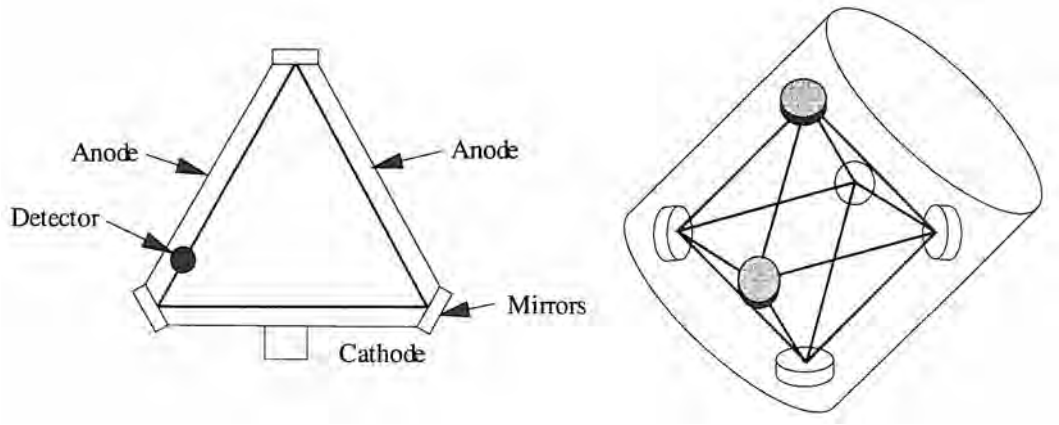


Figure 4.5.25 Sagnac interferometers used as ring laser gyroscopes. (After Koper.)

Because the node spacing is equal to half of the wavelength of the light generated, as the diameter of the ring laser gyro is increased, the angular resolution also increases. Current limits on the minimum rate of detection are also a function of the *lock-in effect*. Rotation rates below this level cannot be detected. Current state-of-the-art ring laser gyros can detect motion as slow as 0.001°/h (1 rev/41 years or 4.8×10^{-9} rad/s) to as fast[65] as 500°/s. However, laser gyro guidance systems with this accuracy are expensive (on the order of $100,000 for three axes).

A medium-size ring laser gyro typically has three orthogonal laser rings set up in a common cavity. This type of gyro can be made as small as a grapefruit. As a navigational tool for aircraft and space vehicles, the ring laser gyro is rapidly replacing mechanical gyroscopes. Perhaps one day when precision engineers are as common as lawyers, new processes and machines will be invented that will allow every car, robot, and bulldozer to be equipped with a ring laser gyro.

4.5.6 Fabry-Perot Interferometers[66]

Similar to the way that averaging data acts to increase the accuracy of an analog sensor, a Fabry-Perot interferometer (developed by Charles Fabry and Alfred Perot in 1897) uses multiple beams of light to make the resultant interference ring pattern very defined and sharp.[67] Michelson's interferometer essentially interfered imperfect sine waves to yield fuzzy bands of light and dark fringes, although the use of laser and optical heterodyne techniques has allowed resolution to be increased by two orders of magnitude. On the other hand, increasing the narrowness and clarity of the fringes

[65] J. Koper, "A Three Axis Ring Laser Gyroscope," Sensors, March 1987, pp. 8–16.

[66] For an in-depth reference, see J. M. Vaughan, The Fabry Perot Interferometer: History, Theory, Practice, and Application, IOP Publishing, Philadelphia PA, 1989.

[67] See H. Polster, "Multiple Beam Interferometry," Appl. Opt., Vol. 8, No. 3, 1969, pp. 522–525.

can allow for even greater resolutions, although a price is paid in terms of range of measurement. A Fabry-Perot interferometer is a multiple-beam interference device. As shown schematically in Figure 4.5.26, a single light ray reflects numerous times between two mirrors, with each reflection leaking off a component of light which contributes to the interference pattern. As a result, imperfections in the mirror surfaces and nonmonochromaticity of the light are essentially averaged out. As a result, very sharp narrow bright fringes are formed as opposed to the equal-width light and dark fringes seen with a Michelson interferometer.

When the distance between the two partially silvered mirrors of a Fabry-Perot interferometer is essentially fixed, the device becomes known as an *etalon*. This device forms the basis for laser cavities and for spectroscopic instruments. Recall that all substances have an electromagnetic signature (e.g., copper produces green light in an oxidizing flame). Also recall that a diffraction grating has the ability to break broadband light up in to its component colors (like a prism). Consequently, a substance can be identified by analyzing the light components in its characteristic spectrum. When materials are closely related atomically, however, resolving the spectral lines becomes difficult with most diffraction gratings, although some gratings can have resolutions on the order of 1 part in 10^6. Fabry-Perot spectroscopy, on the other hand, produces superimposed circular fringe patterns, where each fringe pattern is associated with a certain frequency of light.

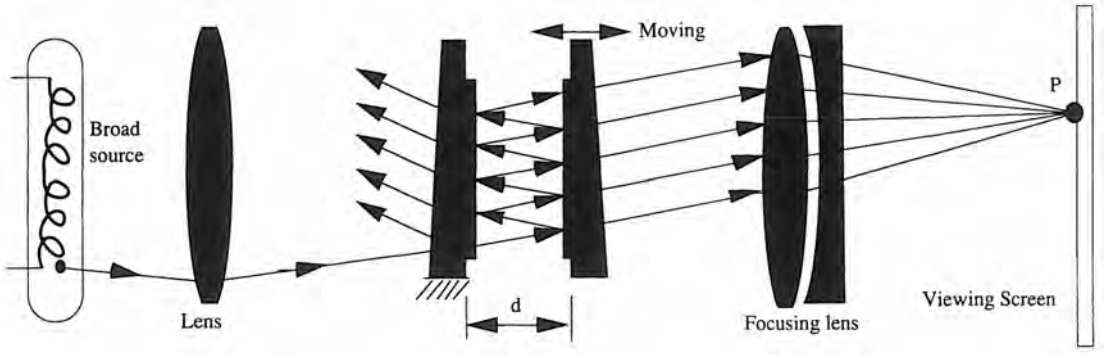


Figure 4.5.26 Operating principle of a Fabry-Perot interferometer. (After Hecht.)

The ratio of the wavelength λ to the least resolvable wavelength difference $\Delta\lambda$ is called the chromatic resolving power $\mathcal{R}$ of a spectroscope. The *finesse* $\mathcal{F}$ of an optical system is a measure of how perfect the optics are and the number of passes the light makes in order to form the interference fringes: The finesse $\mathcal{F}$ is thus equal to the ratio of the separation of the fringes to the half-width of the fringe. For a Fabry-Perot spectroscope with the mirrors separated by a distance d and refractive index n_f, the resolving power $\mathcal{R}$ is

$$\mathcal{R} = \lambda/\Delta\lambda = 2\mathcal{F}n_f d/\lambda \tag{4.5.30}$$

and thus

$$\Delta\lambda = \lambda^2/2\mathcal{F}n_f d \tag{4.5.31}$$

If the relation between frequency, speed, and wavelength $f = c/\lambda$ is differentiated with respect to λ, ($|\Delta f| = |c\ \Delta\lambda/\lambda^2|$) and substituted into Equation 4.5.31, an expression for the bandwidth is obtained:

$$\Delta f = \frac{c}{2\mathcal{F}n_f d} \tag{4.5.32}$$

The more sensitive the interferometer, the smaller the bandwidth it can be used to detect, as can be deduced from the definition of the finesse. When used as a spectroscopic device, increased resolution means decreased bandwidth, which can be a limiting factor in its ability to distinguish between various compounds. When used as a displacement measuring device, one must consider that as the distance d between the optics changes, the fringe position will move due to the changing position of the mirrors and the phase shift associated with the velocity of the mirrors.

If we assume, for example, that the finesse of the system is about 30 (typical) and $n_f d$ = 0.010 m, then the minimum resolvable change in wavelength and frequency will be about $\lambda/10^6$ and

5×10^8 Hz, respectively. As a spectrometer, the Fabry-Perot interferometer is an excellent bandpass filter. If it were used as a displacement measuring interferometer, one of the mirrors would have to be moving at a velocity of about 316 m/s in order to cause a Doppler phase shift that could be resolved. At a small fraction of this speed, the resultant phase shift should not affect the displacement accuracy measured in terms of motion of the fringes.

With a Fabry-Perot interferometer, the peaks (light bands) are separated by $\lambda/2$, but the width of the fringe γ is only $\lambda/2F$. If we assume that an N-bit analog-to-digital converter (where the least significant bit is noise) will be used to discretize the intensity of the fringe, then the resolution of the device can be $\lambda/2^N F$. For N = 12 bits, F = 30, and λ = 6328 Å, the resolution is λ/122,880 = 0.05 Å! Unfortunately, allowing for edge effects, the actual useful range of measurement is only on the order of $\lambda/4F \approx 53$ Å at this resolution level. The sharpness of the fringe corresponds to a low level of optical noise. Hence resolution is principally a function of the noise level in the photodiode, the resolution of the ADC used to discretize the output from the diode, and how well the cavity mirrors are kept parallel. Greater range of motion might be attained by focusing the images of two neighboring fringes on the surfaces of staggered photodiodes, so at least one photodiode is always illuminated by a fringe. Fabry-Perot interferometers for measuring small displacements are generally not available as off-the-shelf items because they are rather awkward for this purpose and simpler-to-use heterodyne interferometers are beginning to break the angstrom barrier, so before rushing out and designing a system that incorporates a Fabry-Perot interferometer, a machine design engineer must consult with a team of good optical and electronic system designers. Remember, the greater the resolution, the greater the problems associated with designing, manufacturing, and assembling the system, and Fabry-Perot interferometers are not known for their ease of use.

4.6 Laser Triangulation Sensors

Measurement of distance by triangulation is as old as the pyramids. When used in conjunction with lasers and photodiodes, triangulation can be a powerful tool for noncontact distance measurement at a standoff distance that would not allow the use of capacitive or inductive probes; however, the resolution is not nearly as high. Optical triangulation systems can also be designed to reflect off the surface at a very focused point, thus avoiding averaging effects, which can hide detection of sharp changes in surface contours. A typical laser triangulation system is shown schematically in Figure 4.6.1. Light from the laser makes a spot on the surface of the target and a lens is used to focus scattered light onto the surface of a photodetector. The output of the photodetector is proportional to the position of the center of intensity of the focused image. Thus uniform ambient light has no effect on the reading, but a bright spot, caused by the laser incident on the surface of an object, will create a reading. Based on the relative geometry of the laser, lens, and photodetector, and the position of the spot on the photodetector, the distance from the probe to the surface of the object can be determined. Note that this is an absolute, not incremental, distance measuring device. Typical applications for this type of device include continuous inspection of surface topographies of objects such as complex castings and propellers. Many other geometries are available for this type of probe, including pencil-thin forms that can be inserted into tight spaces.[68] Since the photodetector finds the center of intensity of the reflected beam, the surface need only be randomly diffuse in order to make resolution largely independent of surface finish.

Typical Characteristics of Laser Triangulation Sensors

The following summary of laser triangulation sensors' characteristics are generalizations only. Note that manufacturers are always advancing the state of the art, so this general summary should by no means be taken as "gospel."

Size: As small as a pencil or a pack of cigarettes. Typically $30 \times 80 \times 70$ mm.

Cost: Probes cost on the order of $5000; black box signal conditioning and digitizing electronics may also cost on the order of $5000.

Measuring Range: From about 5-50 mm (0.2-2 in.).

Accuracy (Linearity): On the order of $\pm 0.2\%$ of full-scale range.

[68] Laser triangulation gages are made, for example, by Chesapeake Laser Systems, Lantham, MD.

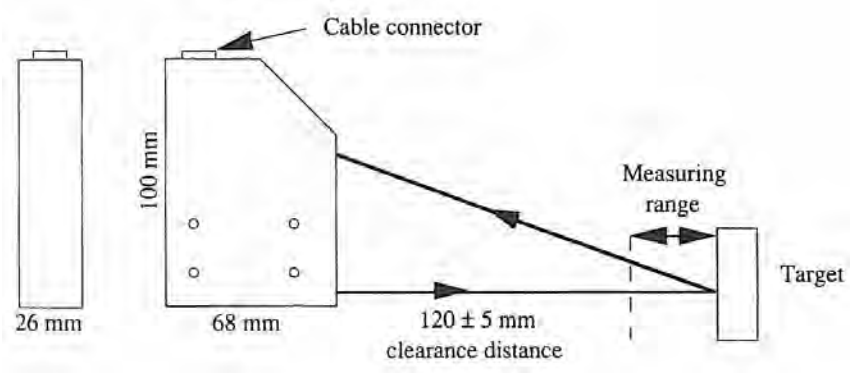


Figure 4.6.1 A laser triangulation distance sensor. (Courtesy of Candid Logic Inc.)

Repeatability: Depends on the repeatability of the surface finish, but can be on the order of 0.05% of full-scale range.

Resolution: On the order of 0.02% of full-scale range.

Environmental Effects on Accuracy: On the order of 0.01%/C^{o} of full-scale range from the nominal 20^{o}C operating temperature.

Life: The probe is noncontact, so life can be infinite if the optical surfaces are not degraded by the environment.

Frequency Response (-3 dB): 2-5 kHz with off-the-shelf black box. Higher rates are possible depending on the manufacturer, the speed of the analog-to-digital converter, and the filtering performed.

Starting Force: No force between probe and sensing surface.

Allowable Operating Environment: To maintain accuracy the probe optics and the sensing surface must be kept clean. This can be accomplished with the use of air wipes in a dirty environment. Systems typically can operate from 0 to 40^{o}C.

Shock Resistance: On the order of 10g.

Misalignment Tolerance: If the incident beam is not normal to the surface, small cosine errors will occur with respect to sensor accuracy. Typically, the maximum allowable misalignment tolerance is $\pm 15^{o}$.

Support Electronics: Analog output of the probe can be filtered and digitized by user-supplied electronics, or an optional black box can be bought with the sensor to digitize, filter, and average the output signal as desired by the user. Most systems operate off line voltage. The probe itself requires a precision power supply which is included as part of the black box.

4.7 PHOTOELECTRIC TRANSDUCERS

Photoelectric materials absorb incident light and use the energy to move electrons to higher energy levels (valence band) or to a state where the electron is freed, thereby increasing the conductivity of the material. Conversely, current applied to some semiconductors causes photons to be emitted. Hence there are three different ways in which photoelectric transducers can utilize the photoelectric effect: photoconduction, photoemission, and photovoltaic action.

In photoconduction, light incident on the photosensitive material causes the material to change its conductance (resistance). Many photodiodes' construction are based on photoconductive principles. When an excitation (bias) voltage is applied across a photoconductive diode, the output voltage can be proportional to the intensity of light striking the diode, or a ratio of the voltage outputs from the different leads can be used to determine the position of the center of intensity of the light beam on the diode surface.

In photoemission, an applied current frees photons from the photosensitive material's surface. This type of transducer is known more commonly as a light-emitting diode (LED). Not a sensor itself, LEDs are often used as light sources for transducers such as optical encoders. LEDs are

also used to transmit information from remote probes to a main controller (e.g., from a probe on a rotating component). A common application of phototransducers in machine tool applications is for wireless transmission of sensor output where radio frequencies may be distorted by other radio sources or strong magnetic fields created by pulse-width-modulation-driven electric motors. The position sensing probes are held in the tool carousel along with the cutting tools. After the part has been cut, the probe is put into the spindle and the machine tool acts as its own coordinate measuring machine to verify part geometry. Although this will not detect errors in part due to the accuracy limits of the machine tool, it will allow the operator to check the part program for accuracy or allow for the detection of out-of-tolerance surfaces caused by worn or broken tools. Most of these probes use an array of battery-powered LEDs on the probe and a strategically mounted photodiode as a receiver.

In photovoltaic action, a voltage is generated when light is incident on the photosensitive material. Since photovoltaic cells are only about 10-20% efficient, the designer must be careful to ensure that the dissipated thermal power does not cause unacceptable thermal errors in ultraprecision systems. In addition to acting as power sources for devices such as calculators, satellites, and remote instrumentation sites, photovoltaic devices in combination with LEDs are often used as *opto-isolators*. Unlike fuses and buffers through which power surges may leak, these device pairs will transmit signal-level voltages, while the physical gap between the components will cause them to saturate before destructive power surges can be transmitted. Because of their fast response time, photovoltaic devices are most often used to construct one of the most fundamental optical transducers, the *photodiode*.

Photodiodes

Photodiodes are principally used as sensors to measure the intensity of an incident light beam or the position of the center of intensity of a light beam on the surface of the diode. The types of photodiodes available include discrete array and monolithic one- and two-dimensional devices which can provide discrete or analog output. Their light-wavelength sensitivity can range from infrared to ultraviolet. Monolithic analog output photodiodes are commonly used in conjunction with LEDs in optical encoders and proximity sensors, and in conjunction with lasers in autocollimators, interferometers, and triangulation sensors. Two-dimensional discrete arrays of individual photodiodes form the imaging device of most video cameras. Photodiodes are the fundamental building blocks of many optical sensor systems because they allow optical signals to be converted into a voltage which can then be digitized for use by computers.

In its simplest form a photodiode is a solid-state device that uses the energy from photons (light) to switch on a transistor, thus enabling current to flow. This indicates that light is incident on the surface of the diode. To detect size and shape of the incident image, many individual photodiodes can be placed in one- or two-dimensional arrays in order to discretize the image. Arrays with 1024×1024 elements on a 0.5×0.5 in. surface are available. Arrays with 4096×4096 elements have been constructed; however, one must realize that the number of bits of information to be processed is equal to 1024^2 versus 4096^2, respectively. So even though diodes with greater resolution can be constructed, one must consider the ability of the supporting hardware to process all the information.

Discrete photodiode arrays that measure the intensity of incident light are also available. The voltage generated at the point is read with a fast analog-to-digital converter. The more levels of intensity required, the slower the response of the diode. Note that if only one very fast (10 MHz) 6-bit analog-to-digital converter were used to give 64 intensity levels, a 1024×1024 array could only be scanned at 10 Hz. Obviously, more ADCs can be used and only changing areas can be scanned. There are basic physical limitations, therefore, that vision systems must face in their quest for resolution and speed.

For use in optical position sensors, monolithic diodes are most often used, which provide an analog voltage output. Since the diode is made from a continuous photovoltaic or photoconductive media, it could ideally provide infinite resolution. They can be used to measure the intensity of an incident light beam or the position of the center of intensity of an incident beam. When used to measure the intensity of incident light, they are sensitive to varying ambient light unless a narrow bandpass filter is used (e.g., a 630 to 635 nm bandpass filter would be used with systems that use He:Ne laser light). Applications include sensing in optical encoders and in optical heterodyne detection. When used in a differential voltage mode, they can be used to detect the center of intensity

of the incident light signal while being insensitive to diffuse ambient light and variations in intensity of the light signal. Applications include light spot position sensing elements in autocollimators and laser triangulation systems.

Position sensing photodiodes have two common forms: quadrature and lateral effect. The output from a quadrature diode drops off very rapidly as the spot moves away from the center. Near the center, the output rises rapidly. This makes a quadrature diode very useful as a centering (nulling) device. In this type of application, when a large error exists, the spot is near the edge of the diode and high-resolution is not needed. When the spot is near the center of the diode, highly sensitive and linear output is desired to maximize controllability of a servo system designed to keep the spot at the center. This type of application is typically required for pointing and tracking systems. In the lateral effect mode, the voltages from the four quadrants are proportional to the position of the light spot on the surface of the diode. The ratio of these voltages is proportional to the location of the center of intensity of the incident light on the diode. When operated in the photovoltaic (PV) mode, no bias voltage is required, and minimum noise levels (i.e., dark current) are present in the diode; however, the rise time and fall time are longer. In the photoconductive mode, an applied bias voltage does not change the level of the output, but it decreases the rise and fall times, thereby making the diode more responsive; on the other hand, the dark current in the diode increases. The uniformity of response (X%) is a measure of the symmetry of the output of the diode. If the spot is incident on a point 1,1 in the first quadrant, then the magnitudes of the voltages must not vary by more than X% when the light spot is incident on a point -1,1, or -1,-1, or 1,-1.

A two-dimensional lateral effect diode is typically built with a ground contact at the center and four leads spaced 90° apart around the outer circumference. The position of the center of intensity of an incident light beam is determined by monitoring the photogenerated currents from each lateral contact. When light is incident on the diode, the current generated from each photon must travel to a lead. The resistance along the path to the lead determines the net contribution of each photons' energy to the current at each lead. In this manner, the lateral effect diode acts as a light-controlled variable resister for measuring the position of the light spot on the X and Y axes of the detector. Linearity of the response is dependent on the uniformity of the resistance of the diode surface. Since no manufacturing process is perfect, mapping of the diode's response is sometimes required. If leads A and C lie on the X axis, and leads B and D lie on the Y axis, the X and Y positions of the center of intensity of an incident light beam are given by the ratios of the difference and sum of the voltages measured at the leads:

$$X = \frac{A - C}{A + C} \qquad Y = \frac{B - D}{B + D} \tag{4.7.1}$$

The responsivity of the diode is the product of the accuracy and the scan frequency required. O'Kelly[69] gives the following relation for determining the responsivity of the diode DL in distance resolution per square root of the desired bandwidth with a signal-to-noise ratio of 1:

$$\Delta L = \frac{(4KT/R_s + E_n^2/R_s^2 + 2R_\lambda P_d q)^{1/2} \times L}{2P_d R_\lambda} = \frac{\text{length}}{\sqrt{\text{Hz}}} \tag{4.7.2}$$

The first term in the numerator is diode Johnson noise, the second term is the amplifier noise and the third term is the diode shot noise. For example, consider a typical 30 × 30 mm diode:

K: Boltzmann's constant = 1.381×10^{-23} J/K°
T: temperature = 293°K
R_s: resistance between contacts = 1000 Ω
E_n: amplifier noise = 10^{-8} V/(rad/s)$^{1/2}$
P_d: monochromatic incident power = 0.001 W
R_λ: detector responsivity = 0.25 A/W
q: electron charge = 1.60x10^{-19} C
L: distance between contacts = 0.03 m

For a signal-to-noise ratio of 10, the sensitivity is ΔL = 84 Å/Hz$^{1/2}$. The typical rise time for this diode is 5 μs. At 1 kHz, a resolution of 0.27 μm (10 μin.) could be achieved. Often, the principal limiting performance factor in lateral effect diodes use is the cleanliness of the system and the speed and resolution of the analog-to-digital converters used.

[69] B. O'Kelly, "Lateral-Effect Photodiodes," Laser Focus/Electro-Opt., March 1976, pp. 38–40.

Photoelectric Proximity and Distance Sensors

Photoelectric proximity sensors have two major elements: an emitter and a receiver. The emitter can be a light-emitting diode, collimated light source, or diffuse light source operating in the infrared or visible light range. The receiver is usually a photodiode. Light is transmitted from the emitter to the receiver in one of five different ways: opposed, retroreflective, diffuse, convergent, or specular.

In the opposed mode, shown in Figure 4.7.1, the emitter and receiver are placed opposite one another so that the source light is incident normal to the receiver. This mode is often referred to as the *through beam, interrupted beam*, or *beam-break* mode because when an object moves through the light beam, the light beam is broken and the output from the photodiode goes to zero. Opposed mode sensing usually provides the greatest optical contrast, especially if the object is opaque like a bottle filled with a liquid.

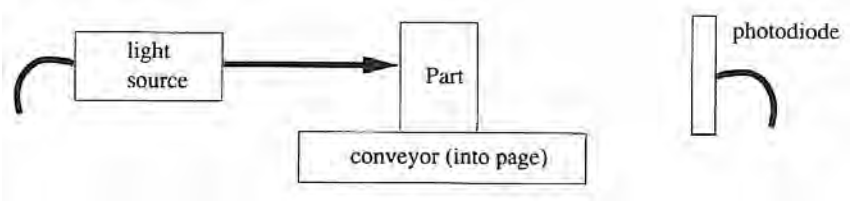


Figure 4.7.1 Opposed mode (interrupted beam) operation of a photoelectric proximity sensor.

In the retroreflective mode, shown in Figure 4.7.2, the emitter and receiver are in the same unit. A retroreflector is placed behind the object being sensed. In many instances, an ordinary bicycle reflector can be used. The retroreflective mode is not as reliable as the opposed mode when conditions are dirty, when the scanning distance is great, or when the object being sensed is itself reflective. The retroreflective device has the advantage of only having to mount one active component and associated signal cable.

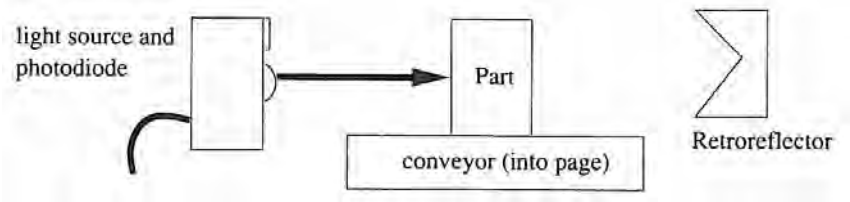


Figure 4.7.2 Retroreflective mode operation of a photoelectric proximity sensor.

A diffuse mode sensor, shown in Figure 4.7.3, also contains the emitter and reflector in the same unit. Light from the emitter hits the target and diffuses in all directions. The emitter-receiver unit can thus be placed at any reasonable angle with respect to object being sensed, because some portion of the reflected light will reach the receiver. In order to prevent stray reflections from triggering the device, the beam power must be tuned with respect to the distance to the surface to be sensed. The diffuse mode is most often used for general proximity sensing.

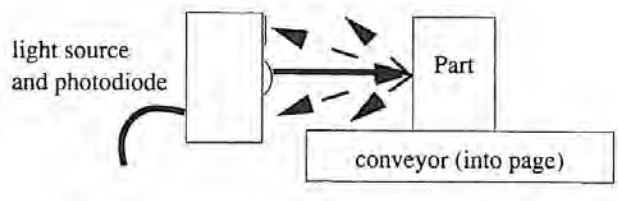


Figure 4.7.3 Diffuse reflection mode operation of a photoelectric proximity sensor.

Convergent mode sensing is a special form of diffuse sensing employing additional optics similar to those used with laser triangulation sensors. The added focusing optics create a small,

well-defined image at a fixed distance from the sensor. The receiver collects reflected light from the object being sensed similarly to the diffuse mode. This mode is often used when it is necessary to detect objects which are very close to a reflective surface. Specular mode sensing operates in a similar manner as shown in Figure 4.7.4. The emitter and receiver are mounted at equal angles from the perpendicular to a reflector. The distance from the sensors and the objects must remain constant.

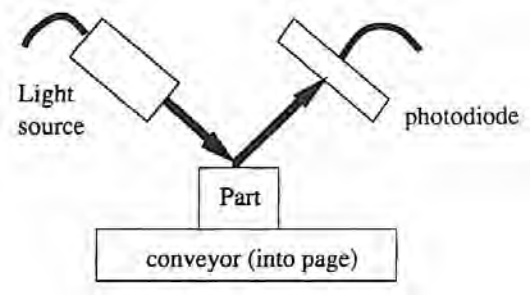


Figure 4.7.4 Specular reflection mode operation of a photoelectric proximity sensor.

Photoelectric sensors are usually used in the proximity (on/off) mode, although models are available with analog outputs which provide a dc voltage or current proportional to the distance from the object. The outputs may be linear, parabolic, or sometimes logarithmic. Output voltage ranges are usually 0-30 V dc depending on the amplifier supplied with the sensor.

Typical Applications

Photoelectric sensors are widely used as counting and indexing control devices in production line machinery. Since this type of machinery often utilizes pneumatic components such as pistons, air is usually available for use as a wipe to keep the surface of the optics clean. The principal advantage of photoelectric proximity sensors over other types of proximity sensors (e.g., capacitive, inductive, and mechanical) is that they are much less sensitive to object position and orientation, although this also means that their resolution is more limited.

An example of the use of photoelectric proximity sensors in machinery for other than counting purposes is shown schematically in Figure 4.7.5, where a pair of photoelectric sensors are used to control the position of a sanding belt on a large industrial sander. Manufacturing and economic limitations prevent the belt and rollers from being so perfectly made that the belt will always track perfectly on the rollers. Thus the upper roller is made to be able to rotate about an axis normal to its length so that as the belt starts moving off to the right, it breaks the beam of an interrupted-mode photoelectric proximity sensor. The output from the sensor triggers a relay signaling a valve to move a piston that rotates the length of the roller in a counterclockwise manner, thereby causing the belt to move to the left. When the belt moves off too far to the left, it breaks the beam of another opposed-mode photoelectric sensor. The output triggers a relay signaling the valve to move the piston, which rotates the length of the roller in the clockwise direction, thereby sending the belt off to the right. An added benefit of this automatic tracking mode is that the sanding belt ends up oscillating back and forth, giving a better finish to the part.

Typical Characteristics of Photoelectric Sensors

There are many modular photoelectric sensors that are meant to bolt into standard limit switch mounts. These sensors may be fist-sized and are usually designed for harsh environments. The following summary of photoelectric sensors' characteristics are generalizations only. Manufacturers are always advancing the state of the art, so this general summary should by no means be taken as "gospel."

Cost: From $2 for a small Radio Shack® version with 1 mm range to $100+ for industrially hardened models with a range of many meters.

Size: As small as an aspirin capsule, to heavy-duty fist-sized models for extended range industrial conditions.

Range (Standoff Distance): Opposed mode ranges vary from 0.3-30 m (1-90 ft) to 2.5-25 cm (1-10 in.). Retroreflective mode ranges are on the order 5 cm to 3 m (2 in. to 9 ft). Diffuse

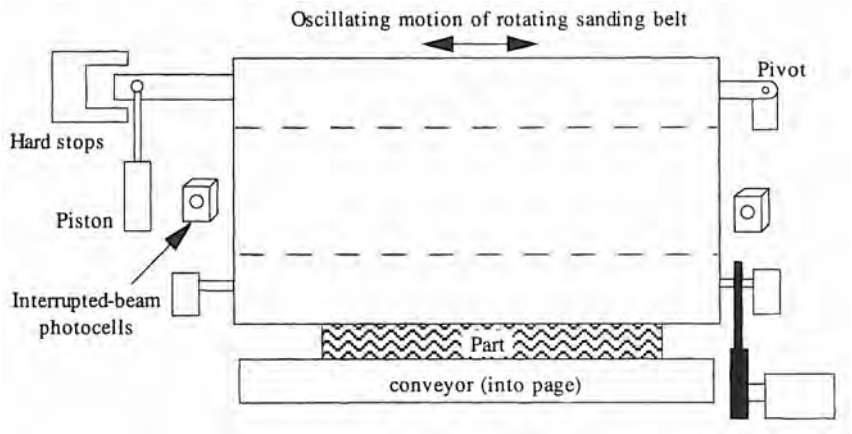


Figure 4.7.5 Photoelectric proximity sensor used to control oscillating motion of a sanding machine's belt.

reflective mode sensors have ranges of a few millimeters to 50 cm (fractions of an inch to 20 in.).

Accuracy (Linearity): Not applicable for this type of sensor.

Repeatability: From 0.1 mm (0.004 in.) to 10 mm (0.4 in.) depending on the beam size, uniformity of the target, and cleanliness conditions.

Environmental Effects on Repeatability: Performance is very dependent on keeping the optics clean. An air wipe can be used for this purpose in a dusty environment.

Life: If the cable is not flexing, life can be infinite. For applications requiring cable flexing, virtually infinite life can be achieved through proper cable carrier design.

Frequency Response: To generate an acceptable signal, illumination times vary from 100 μs to 0.1 s.

Starting Force: None.

Allowable Operating Environment: Most photoelectric sensors are hermetically sealed to meet almost any requirement, including underwater operation. The primary consideration is whether an air wipe can be provided for in environments where accumulation of gunk can act to block the lens. Operating temperature range can be from 240°K to 70°C.

Shock Resistance: On the order of 10g.

Emitter/Receiver Misalignment Tolerance: Highly dependent on the sensing application and type of object being detected but can be as high as 10-15° between the emitter and receiver for some models.

Support Electronics: Nominally, only a dc (6-24V at 0.010-0.050 A) power supply is required. TTL-compatible output voltages are usually output from the sensors. Note that for transmission of the signal over long distances, units without an internal amplifier may need to have an external amplifier added. Control units may also be purchased for sensors. Control units provide various logic outputs for input into machine controllers, and so on. For example, if a sensor is to be used as a switch, a control unit can be used to provide SPDT (single-pole double-throw) logic. Other logic units may include delays, one-shots, counters, repeaters, and so on. Controllers can cost from $10 to several hundred dollars.

4.8 TIME OF FLIGHT SENSORS

Time of flight sensors measure the distance to an object by emitting a pulse of energy and measuring the time it takes for the reflected wave to return. The higher the frequency of the wave, the less it is affected by atmospheric conditions. Radar (radio detection and ranging), EDMI (electronic distance measuring instrument), and GPS (global positioning system) are usually used for applications ranging from locating aircraft to measuring long distances on construction sites, but are also finding

increasing applications in mobile robot position sensor systems. Thus these sensors warrant at least a brief discussion here.

Unlike an interferometer, which is an incremental position sensing device, a radar system attains absolute position measurements by emitting and timing a continuous series of short electromagnetic wave pulses, with each pulse lasting for perhaps only a microsecond. Since electromagnetic waves travel at the speed of light, detecting the time of flight to within 1 μs allows for 150 m resolution. Numerous forms of radar have evolved, including focused plane systems that allow range, altitude, and bearing to be detected. Doppler radar measures the Doppler shift of longer pulses and thus only detects moving objects. Doppler radar units are used as ground speed detectors for airplanes and heavy machinery such as bulldozers. Doppler radar is more expensive than conventional radar, but it is less likely to think that a hill is a speeding automobile. Frequency-modulated (FM) radar generates a continuous beam of varying frequency radiation. The receiver sees energy reflected from the target as well as energy directly from the source. As the frequency varies over a range Δf, a varying beat frequency is generated where the frequency of the beats is proportional to the distance of the object. Like optical heterodyne detection, utilizing beat frequencies dramatically increases the apparent length of the electromagnetic wave while maintaining the advantages of short wavelengths; hence it allows for more accurate measurement of the occurrence of peaks and valleys of the wave. The beat frequency

itself will also be a function of the known reference distance from the source to the detector and the distance to the target. The longer the reference distance, the greater the range and resolution of the system. Scanning laser versions of this type of radar have been developed and are sometimes referred to as *laser radar vision systems.*[70]

EDMIs operate on a similar principle except that they use a focused beam of infrared light that is reflected off a retroreflective target. With pulse-shaping techniques and very fast detection electronics, accuracies of 1-5 mm $\pm$1–5 ppm are available with measurement update times on the order of 1–5 s. The parts per million error is due to variations in the refractive index of light along the beam path affecting the speed of light value used in determining distance. Costs vary from about $3500 to $15,000 depending on the accuracy, speed, and accessories desired. Size of an EDMI is about that of two coffee mugs[71]. Since these instruments are designed for making quasistatic measurements (0.1 Hz), they would most likely be used in mobile robotic systems as a position update device when the robot came to rest. For control of a moving robot, encoders on the wheels or Doppler radar could be used as a means of determining instantaneous velocity and position.

A relatively new position measuring system is the *global positioning system* (GPS), which with the rapid advance of microelectronics has the potential to revolutionize travel and construction. Parts of the GPS system are already in place and by the early 2000s, 20 to 30-GPS satellites (depending on how many spares are put up) will be in geosynchronous orbit around the Earth. Ideally, at least three satellites will always be visible to an observer, with a maximum of six being visible at any one time. Each satellite emits a characteristic signal which has a timing signal and position information for each satellite. Using this information and time-of-flight and triangulation measurement techniques, current systems can provide global positioning accuracies for civilian purposes to 30 m absolute and 1-10 ppm differential with 15-20 min of scan time. Although the receiver and required black box electronics can cost from $50K to $100K, the rate at which electronic systems technology advances may one day allow for a wristwatch version to be developed with centimeter global accuracy and seconds response time. At this point in time, however, the GPS system is strictly a surveyor's tool as far as civilian use is concerned.

4.9 VISION SYSTEMS

A vision system can be classified in a broad sense as a noncontact optical sensor system that obtains geometric information about a part such as position, orientation, size, shape, or surface contour. When considering the use of a vision system as part of a machine system, one must be careful to think

[70] Commercially available from Digital Optronics Corp. of Herndon, VA. For a more detailed description, see F. Goodwin, "Coherent Laser Radar 3-D Vision System," SME Tech. Paper MS85-1005.

[71] See the annual EDMI survey presented in the surveyor's trade journal, Point of Beginning, published by POB Publishing Co., P.O. Box 810, Wayne, MI 48184.

about the real applications and limitations of these devices. When designed around a specialized application, vision systems can be one of the most useful, reliable, and efficient tools available (e.g., using a scanning laser to perform 100% inspection of every part to verify, for example, that all pins on electronic connectors are present and are not bent); on the other hand, when vision systems are expected to have capabilities portrayed in science fiction movies, unfulfilled expectations and disappointment are usually the result.

For precision engineering applications, vision systems are useful for mapping the shape of a tool. This is especially important if a contoured surface is to be machined using a Cartesian machine without a rotary table. Everyday applications are found in the form of optical comparators.

The machine vision process can be divided into four basic steps: image formation, image preprocessing, image analysis, and image interpretation.[72] Regardless of their sophistication level, all vision systems utilize these four steps. The components of a typical vision system include the environment, the illumination source, the light receiver, and the signal processing system. The former two are perhaps the most critical part of the system because they determine the degree of complexity required in the latter two in order to extract useful information from the system.

The Environment

A classic problem that vision sensors are thought of as being ideal to solve is the *random part orientation in the bin* problem. In this problem, a robot is presented with a bin of random parts randomly piled on top of each other in a bin. The robot's job is to pick out the parts in an orderly fashion and put them where they belong. Now ask yourself the following question: *Who dumped all the parts in the bin in the first place? Why not use a vibratory bowl feeder to feed the parts in an orderly manner? If the parts are too complex or delicate to be bowl fed, are they not also too delicate to be tossed into a bin, where they can get nicked and gouged?* On the other hand, there are applications such as casting deburring where a large number of unfinished parts are dumped out of a vibratory deburring station and they need to be sorted. Hence there is the need to analyze the entire process and not just try to solve a difficult problem with the application of expensive high technology.

In general, as a problem becomes more defined and ordered, vision system technology becomes more applicable, economical, and reliable. On the other hand, a vision system is often the only type of sensor system that can be used in conjunction with problems that are undefined and disordered. The designer's goal must thus be to minimize cost and maximize productivity and quality of the *entire machine-sensor system.*

Illumination

A vision system recognizes different features on the object and the background either by *reflected* or *through illumination* of the object. Reflected illumination is the most versatile form of illumination because it allows complex surface topographies to be measured; on the other hand, it requires a commensurate amount of signal processing software to extract this information. There are many types of reflected illumination schemes. With diffuse front lighting, the light source is next to the light receiver and in front of the object. Front-polarized lighting uses polarized light to help minimize effects of ambient light. With front-directed bright-field lighting, the light source is set an angle so that the light reflected off the object is reflected in a specific direction. Structured light uses light with a special geometric pattern (e.g., a scanning dot, line, or grid) to reduce the complexity of the signal processing required.

Through illumination or *back lighting* has the light source on one side of the object and the receiver on the other, thereby producing a silhouette of the object. It produces the greatest contrast and is used, for example, in optical comparators, which compare the silhouette of an object to a reference pattern. This type of system can also be used to map the shape of a tool for turning of complex shapes without a rotary axis to keep the tool normal to the work surface. Through-illuminated vision systems are among the most easy to implement and the most accurate.

The type of light that can be used includes diffuse white light, quasimonochromatic diffuse light, and structured light. Normal diffuse light is the most convenient but the most difficult to implement, because reflective surfaces and shadows create images that can trick the system into thinking there is an edge where there is only a shadow line. Quasimonochromatic diffuse light and

[72] For a more complete glossary of image processing terminology, see R. Clouthier, "Glossary of Image Processing Terminology," Lasers Optron., Aug. 1987, pp. 60–61.

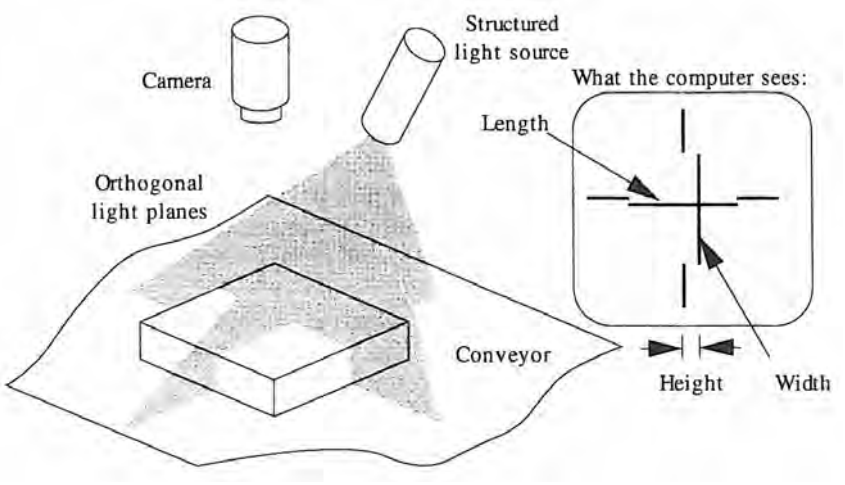


Figure 4.9.1 Two orthogonal structured light planes used to measure three dimensional information. (After Landman.)

narrow bandpass filters can help to eliminate ambient light problems, but shadows can still cause errors. Structured light (e.g., a plane of light), on the other hand, illuminates the object with a pattern of light (e.g., parallel stripes or an array of dots). This allows one to observe the deformations of the pattern caused by the object. When structured light is incident at an angle on an object, the object distorts the light in a manner unique to the shape of the object. The effect is akin to that of lines on a contour map.

If structured light is applied in a scanning mode (e.g., a single line or dot moving across the surface of an object), then shadows, reflections, and other phenomena that produce nonunique images can usually be avoided. For example, as shown in Figure 4.9.1, two planes of light can be used to determine the height and width of objects on a conveyor. If the conveyor speed is accurately known, length dimensions can also be determined. Types of structured light used for similar purposes include multiple planes, grids, and dots.

Image Acquisition and Data Processing

The receiver is responsible for determining the intensity of the reflected light as a function of position (one- or two-dimensional), and converting the information into numerical array elements that the image processing system can use to evaluate what the system is "seeing." Receivers use vidicon tubes or arrays of photodiodes as their sensing elements. A vidicon camera forms an image by passing light through a series of lenses which focus it on photoconductive plate. An electron beam within the vidicon tube scans the photoconductive surface. An output is produced which is proportional to the changes in light intensity along the scan line. Vidicon cameras have a tendency to "burn in" images on the photoconductive surfaces and are thus being replaced by solid-state photodiode arrays. Solid-state cameras use a lens to focus the image onto a one- or two-dimensional array of photodiodes called pixels. Typically, 256 $\times$ 256 pixel arrays are used. Solid-state cameras are small (palm sized), rugged, and their photosensitive elements do not wear out with use.

Two popular formats used to evaluate light intensity are *binary* and *gray* scale. In the binary format an analog signal is converted into one of two possible values, a high (1) value or a low (0) value. A pixel is assigned the low value unless its analog signal is above a certain threshold. This is the simplest form of conversion. It has a very low level of resolution, yet it is usually adequate for simple inspection tasks such as silhouette matching. For gray scale format, analog-to-digital converters are used to convert the analog signal from a photodiode, which is proportional to the intensity of the incident light, into a digital value varying from 0 to as high as 255 (8 bits). Resolving the intensity of the incident light permits surface characteristics to be inspected and compared; however, the processing requirements are significant. A 256 $\times$ 256 array requires over 65,000 8-bit storage locations, and the time needed to process this large amount of data can be on the order of tenths of seconds. To reduce processing time, if a certain feature on an object is to be detected, then only the area containing the feature need be processed.

Various aspects of the image can be deduced from the array of numbers generated by the image acquisition system, including position, orientation, geometric configuration, and surface topography. Since the computer hardware and software to accomplish this task change and evolve so rapidly, it is beyond the scope of this book to discuss them in detail, so only basic concepts will be discussed.

The distance of the object with respect to the camera can be determined using stadimetry, triangulation, or stereo vision techniques. Stadimetry or direct imaging techniques compare the apparent image size to at least two known image size/distance ratios stored in memory. Triangulation was discussed in Section 4.6. Stereo or binocular vision employs parallax to determine the distance to an object; an object's relative motion with respect to a reference object will appear to be different when viewed by two receivers. By combining the two receivers' images, the distance of the object from the reference can be calculated.

If the relative position of three noncollinear points on the object is known, then orientation of the surface can be determined if the distance to the points can be determined. If the usual intensity distribution is known for an object (e.g., a standard part on a conveyor), its orientation can be determined by comparing the light intensity distribution of the image to the known intensity distribution of the part. Areas which are farther away will appear darker than usual. Areas which are closer will appear lighter than normal. The most robust method for determining the orientation of an object, however, is probably with the use of structured light.

In inspection and robotic applications it is often necessary to identify a part. An object's shape can be identified using basic features (e.g., lines and arcs). The shapes and contours within a region are determined by first outlining the shape. Often the outline of the object is sufficient to identify the object. If more detailed information is necessary, then regional components are structured along with their relationships to determine the overall shape. The processed image can also be compared to information stored in order to interpret what the system "sees." Two methods of image interpretation are *feature weighting* and *template matching*. Feature weighting assigns a relative weight value to each feature to interpret the image. Template matching overlays a template of the expected image onto the processed image and compares the two.

The evaluation of a vision system for a particular application should thus be made by considering such factors as resolution, processing speed, discrimination, and accuracy. The resolution of a vision system is the field of view divided by the number of pixels in a particular direction of the array. Cost usually goes up as the square of the resolution. The processing speed of a vision system is a function of the complexity of the processing scheme and the number of pixels. The ability of the vision system to discriminate between features and variations of light intensity is a function of the complexity and resolution of the system. More discrimination usually requires more resolution and complexity and is usually more expensive. The accuracy of the vision system is a function of the percentage of correct decisions made about a group of objects being examined.

Application Examples

The most basic of vision systems utilizes interrupted beam photosensors, scanning laser triangulation sensors, and scanning laser telemetric systems as their system building blocks. It is much easier to construct the geometry by scanning the surface of a part because two of the degrees of freedom, obtained from the scanner, are already known. Examples of these types of systems are shown schematically in Figure 4.9.2.

A scanning laser triangulation system is often used to check electrical connector pin height, row-to-row spacing, and pin position. With this type of system, inspection rates on the order of 0.2–0.4 m/s (8–16 in./s) are possible. It then becomes feasible to provide for 100% inspection of some types of parts on a conveyor. This can enable a manufacturer to meet requirements for zero-defect shipment of parts, which is often required for parts used in high-speed, automated electronics manufacture. Similar systems can be combined with a vibratory bowl feeder to sort and/or check the quality of small parts. The system can "learn" by passing known quality parts through the feeder. Air wipes can be used to make sure that the optics do not become contaminated. Another example of a scanning system uses a laser radar vision system to determine the distance to an array of points on an object. This type of system can have a range on the order of 3–10 m and depth of field of about 1–3 m, which is one to two orders of magnitude greater than most laser triangulation systems. Resolution with 1 or 0.1 second dwell times can be on the order of 25 or 100 μm, respectively.

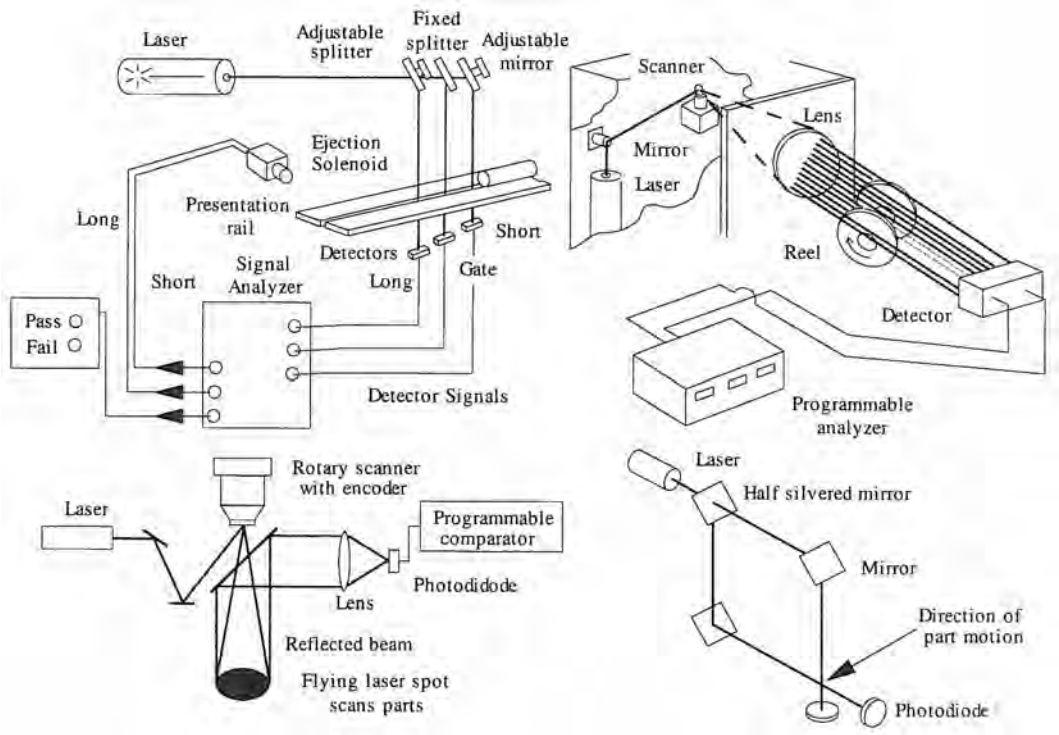


Figure 4.9.2 Various types of simple effective vision systems for high-speed 100% part inspection. Clockwise from upper left: sequence interruption, shadowed signals, transmitted signals, and circular scanning using reflected signals. (Courtesy of Sperry Rail Inc.)

Chapter 5
Sensor Mounting and Calibration

Finally we shall place the Sun himself at the center of the Universe. All this is suggested by the systematic procession of events and the harmony of the whole Universe, if only we face the facts, as they say 'with both eyes open.'

Nicholas Copernicus

5.1 INTRODUCTION

If a sensor is improperly mounted and/or calibrated, then mounting and/or calibration induced errors may dominate in the machine. Factors to consider when designing a sensor mounting system include:

- Sensor location
- Sensor alignment
- Mounting structure design
- Mounting environment
- Contact between curved surfaces
- Metrology frames
- Sensor calibration
- Performance verification

Ideally, the system should be mounted and calibrated on the machine it is going to be used on. Since this is often not practical, the sensor should at least be mounted in the same manner as that in which it was calibrated.

5.2 SENSOR LOCATION

The most important consideration in mounting a sensor is where to mount it in order to ensure that the desired quantity is accurately measured. A common error caused by improper mounting is an *Abbe error,* which occurs when the measurement axis is not collinear with the axis of the quantity being measured (see Section 2.1.2). Exposure to the environment and danger of being physically damaged, however, usually dictate that the sensor be mounted inside the structure of the machine where it can be protected.

Another concern is whether the sensor should be mounted on the input or output ends of a transmission. If the sensor is mounted on the input end of a transmission along with the motor, then the resolution of the sensor will be enhanced by a factor equal to the transmission ratio; however, any nonlinearities in the transmission, such as a nonlinear transmission ratio or backlash, will also affect the output of the sensor. On the other hand, if the sensor is mounted on the output end of the transmission, although it may more accurately measure the process, the controllability of the servo system may decrease if there is backlash in the transmission. This problem can often be overcome if an antibacklash device is used, such as a constant force preload spring, on the transmission.

5.2.1 Where to Mount Linear Displacement Sensors

Of the mounting locations for a linear displacement sensor shown in Figure 5.2.1, which is the best? There are two principal options: (1) fixed scale and moving read head, or vice versa, and (2) inboard or outboard mounting. With respect to the former, if the carriage size is large with respect to the range of motion, then it may be advantageous to mount the scale on the carriage and the read head on the base. In this manner the wiring for the read head does not have to be run through flexible conduit. On the other hand, if the carriage is small with respect to the range of motion, then if the scale is attached to the carriage, it will overhang the length of the carriage and it is sure to get knocked out of alignment. With respect to the inboard or outboard mounting, the principal comparative operating characteristics of the inboard and outboard system are:

1. When mounted in the outboard position, the sensor is farther away from heat generated by friction in the leadscrew nut and linear bearings. Even when these elements incorporate low friction recirculating ball bearings, considerable heat can be generated during rapid motion. If the

sensor were located in the inboard position, the internal environment's composition (e.g., humidity and hydrocarbon content from lubricating fluids) may affect some sensor's performance, such as laser interferometers. In some cases, internal bellows or tubes can be provided to protect the sensor.

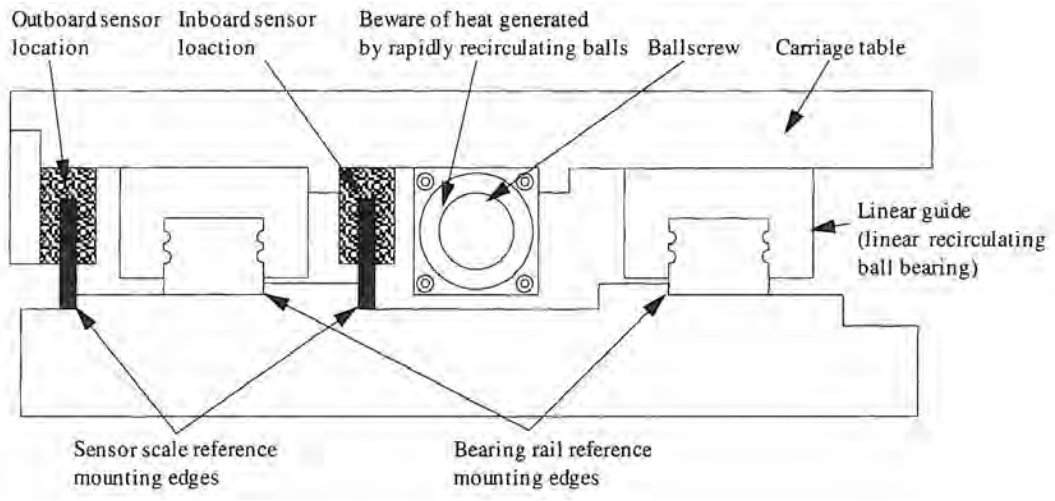


Figure 5.2.1 Possible mounting locations for a linear scale sensor.

2. When mounted in the outboard position, the sensor may not be protected by the machine's structure and way covers, and thus it might be exposed to a flood of chips and coolant. Most linear displacement sensors designed for this type of application are made to resist this type of abuse, but the design engineer should verify the manufacturer's claims by talking to machinists who have operated machines with similar sensors and ask them if there has been any trouble with them. Also, part loading and unloading methods should be considered. If heavy parts are loaded with an overhead crane, what is the likelihood that a part will swing and crush the sensor? Can the sensor be placed on the other side of the axis where this possibility could be minimized? Can a protective metal overhang be designed into the axis, which could also serve as a reference edge for mounting of the sensor?

3. An outboard mounted sensor is easier to install and maintain. Therefore, crucial elements of the sensor system design process should include observing existing designs and having discussions with the manufacturing personnel in charge of fabricating the machine and maintenance personnel who have to service the machine. These types of discussions often result in useful suggestions such as: *If you grind a reference edge on the inside of the base to put the sensor up against, will it alleviate alignment problems?*

4. An outboard-sensor will measure the position of the toolpoint more accurately when the tool is over the sensor but more poorly when the tool is on the other side of the carriage, as shown in Figure 5.2.2. The position errors of the toolpoint over the left side of the carriage would be 0 and $\ell\theta/2$, respectively, if the sensor were mounted at the left outboard or yaw center positions. For the toolpoint over the right side of the carriage, the errors would be $-\ell\theta$ and $-\ell\theta/2$, respectively, if the sensor were mounted at the left outboard or yaw center positions. For the toolpoint over the center of the carriage, the errors would be $-\ell\theta/2$ and 0, respectively, if the sensor were mounted at the left outboard or yaw center positions. In most cases, it is desirable to mount the sensor closest to where the tool will be most of the time. The center of a machine table is most often used, thus it is advantageous for the sensor Abbe error to be minimized in this region.

The design engineer must also consider whether or not the machine's performance will be mapped to allow software-based error correction algorithms to be used. If this is the case, then the mounting location becomes less critical. In the end, however, only when this information is used in the machine's error budget can the design engineer quantitatively determine which mounting locations are acceptable from an accuracy standpoint.

As another example of how sensor placement can affect the measurement, consider the measurement of straightness of motion of a machine tool carriage. As shown in Figure 5.2.3, there

are two methods commonly used for measuring straightness of motion with respect to a reference straightedge, type M and type F. In both cases the X position of the straightedge in the reference frame must be known if one were to incorporate a mapped compensation curve for the straightedge in the calculation of straightness.

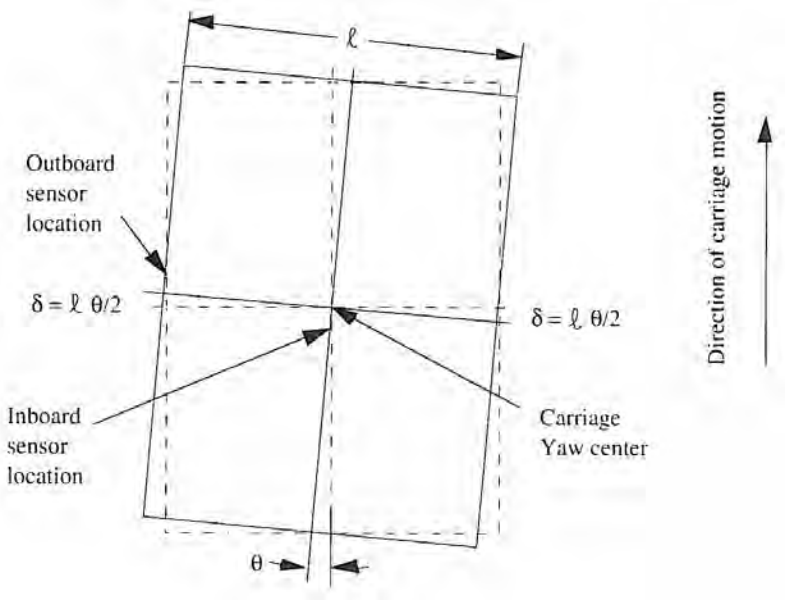


Figure 5.2.2 Abbe errors resulting from mounting a sensor at inboard and outboard locations.

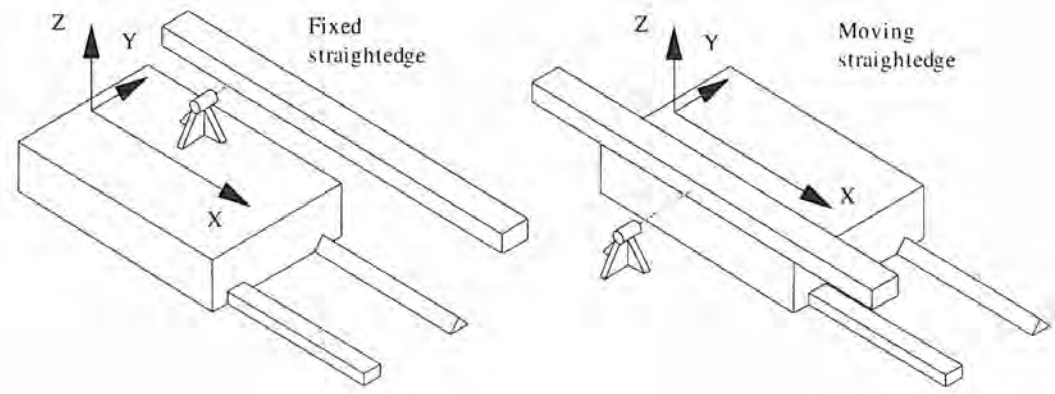


Figure 5.2.3 Type M (moving) and type F (fixed) straightness measurement.

For type M (moving) straightness, the straightedge is attached to the carriage and the Y and Z coordinates are needed to define the location of the straightedge in the carriage coordinate system. This is due to the fact that in addition to the actual straightness as a function of carriage position along the X axis, there is an Abbe error present in the measurement caused by yaw and roll motions that are amplified by Y and Z offsets, respectively. If the Y and Z coordinate are known along with the yaw and roll errors, then the errors can be combined with the straightness reading in the HTM model. Mounting of the sensor is easy (with type M straightness) because one does not have to worry about the sensor cable flexing. On the other hand, if the carriage is small and the length of travel large, mounting the straightedge on the carriage can be difficult.

For type F (fixed) straightness, the straightedge is fixed to a reference coordinate system and the sensor is mounted to the carriage. The Y and Z coordinates of the sensor in the carriage coordinate system must also be known. With type F straightness, it is relatively easy to mount the straightedge, but then one must contend with flexing of the sensor cable. It is possible to use a plane

mirror interferometer or DPMI that uses the straightedge as the target mirror. Ideally, the carriage will behave like a rigid body and straightness measurements made by either method will be equal as long as yaw and roll errors are also measured.

The *Bryan principle*, which is an extension of the Abbe principle, helps make the decision as to which type of straightness measurement to make by stating that the straightedge should be placed to emulate the part and the sensor placed to emulate the tool motion:[1] "A straightness measuring system should be in line with the functional point whose straightness is to be measured."

5.2.2 Where to Mount Angular Displacement Sensors

Angular motion sensors often are attached to power transmission elements (e.g., ballscrews and transmissions) and thus often end up measuring the angular deflection (twist) of a shaft in addition to its angular displacement. The most common example of this situation occurs when a resolver or encoder is mounted to an electric motor which drives a ballscrew. The question is: *Does the shaft windup (torsional deflection) contribute significantly to the overall system error?*

To answer this question, consider that the goal is to find the axial position of the carriage being driven by the leadscrew. There are thus two load-induced errors in the measurement: (1) the error due to windup of the screw, and (2) the error due to axial compression of the shaft. These two errors are intrinsically coupled by the geometry and material properties of the leadscrew shaft.

The twist $\Delta\phi$ of a circular shaft is a function of the applied torque Γ, the length L, the polar moment of inertia I_p, and the shear modulus G:

$$\Delta\phi = \frac{\Gamma L}{I_p G} \tag{5.2.1}$$

The axial displacement of a carriage driven by a leadscrew with a lead l (distance/revolution) is

$$\Delta X = \frac{\ell \Delta\phi}{2\pi} \tag{5.2.2}$$

Thus when a torque is applied to the leadscrew, the equivalent axial displacement error caused by the sensor measuring the twist is

$$\Delta X = \frac{\ell L \Gamma}{I_p G 2\pi} \tag{5.2.3}$$

In order to find the equivalent axial stiffness of this effect, the relation between force and torque in the leadscrew must be considered. Equating the work in to the work out (no friction) gives

$$F = \frac{2\pi\Gamma}{\ell} \tag{5.2.4}$$

The equivalent displacement can be written as

$$\Delta X = \left(\frac{\ell^2 L}{4\pi^2 I_p G}\right)\left(\frac{2\pi\Gamma}{\ell}\right) \tag{5.2.5}$$

The equivalent axial stiffness of the torsional system is thus

$$K_{\Gamma \text{axial eq.}} = \frac{G I_p 4\pi^2}{\ell^2 L} = \frac{\pi^3 r^4 E}{(1+\eta)\ell^2 L} \tag{5.2.6}$$

where r is the leadscrew shaft radius, E is Young's modulus and η is Poisson's ratio. Assuming a fixed-simple supported condition, the axial stiffness of the leadscrew is

$$K_{\text{axial}} = \frac{\pi r^2 E}{L} \tag{5.2.7}$$

[1] For a detailed case study of straightness metrology, see J. B. Bryan and D. Carter, "Straightness Metrology Applied to a 100 Inch Travel Creep Feed Grinder," 5th Int. Precis. Eng. Sem., Monterey, CA, Sept. 1989. Also see J. B. Bryan, "The Abbe Principle Revisited-An Updated Interpretation," Precis. Eng., July, 1989, pp. 129–132.

The ratio of the torsional-axial equivalent stiffness and the axial stiffness is

$$\frac{K_{\Gamma \text{axial eq.}}}{K_{\text{axial}}} = \frac{\pi^2 r^2}{(1+\eta)\ell^2} \tag{5.2.8}$$

Since the lead is usually less than the radius, the torsional-axial equivalent stiffness is at least an order of magnitude greater than the axial stiffness of the leadscrew. Thus any equivalent displacement errors due to torque windup of the shaft are often insignificant compared to the axial compression of the shaft. A linear position sensor would measure the shaft compression, whereas a rotary sensor on the shaft would not.

Note that this is an idealized calculation; the effect of the efficiency and breakaway torque on the *system*, which is caused by seal friction and bearing and leadscrew nut preload, must also be considered. Typically, as the breakaway torque is approached, the shaft is turning at the motor end, but it is not turning at the nut and the carriage does not move. Thus there will be an equivalent displacement error caused by the angular displacement sensor measuring the twist of the shaft as the breakaway torque is approached. Even if the sensor were placed on the other end of the shaft, there would still be the issue of decreased controllability caused by high breakaway torque. Thus reducing the breakaway torque is the best way to increase the resolution and accuracy of a precision machine.

Another important concern is backlash resulting from imperfect mating between transmission parts. If the sensor is mounted integral with the motor, then backlash error will always be present. If the sensor is mounted at the transmission output (e.g., use a linear displacement sensor instead of an angular displacement sensor), then although the accurate position of the carriage may be known, it may be difficult to control the position of the carriage to better than the amount of backlash due to limit cycling problems. Backlash can be reduced with higher-quality components or by increasing the preload on the system, which also increases the breakaway torque. In order to maximize controllability, backlash must be minimized.

5.3 Sensor Alignment

In addition to choosing the best position for the sensor, a method for aligning the sensor to the axis of motion must be provided for. For example, if the machine uses ground linear bearing rails, a reference surface for one of the rails to rest against is usually ground into the machine bed. The design engineer may also want to consider specifying a similar reference surface for a linear displacement sensor. In this manner, alignment is not difficult because the sensor can be placed against the reference surface and fastened in place. As long as the bearing and sensor reference surfaces where machined at the same time without refixturing the part, alignment of the axis of motion and the sensor will be ensured.

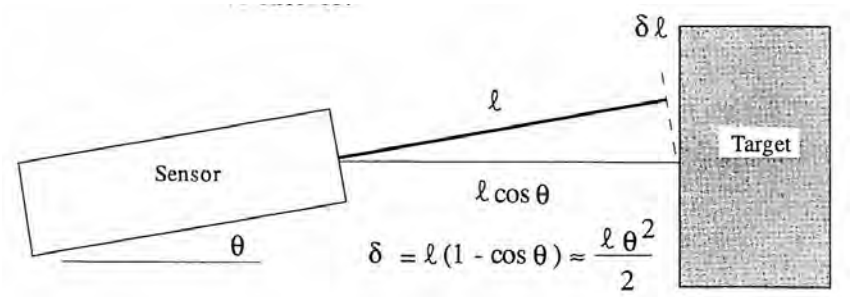


Figure 5.3.1 Cosine error resulting from sensor misalignment.

If a sensor is misaligned with respect to the axis of motion that it is measuring, then the sensor will only be measuring a component of the motion, as shown in Figure 5.3.1. For linear displacement measuring sensors, the error as a function of the measured displacement l and the misalignment θ will be

$$\varepsilon = \ell(1-\cos\theta) = \frac{\ell\theta^2}{2} \tag{5.3.1}$$

hence the term *cosine error*.

There will always be some finite misalignment between the sensor and the axis of motion, but even if the error is acceptable, a method for coupling the sensor to the axis is required to ensure that forced geometric congruence between the two does not overload the sensor's structure and possibly cause additional errors. The sensor manufacturer may recommend that if the alignment error is below a certain amount, the sensor can be rigidly attached to the axis of motion. This type of arrangement, or the use of some other type of flexural coupling, is usually best because it eliminates the possibility of backlash in the sensor's coupling. However, the design engineer must make sure that the natural frequency of the system comprised of the sensor and the flexural coupling does not lie within the operating frequency of the machine. Also, the dynamic deflection of the sensor/coupling system must be considered. Although couplings with sliding interface elements may be more rigid than flexural couplings, the sliding interfaces are themselves sources of backlash and wear. Figure 5.3.2 shows an encoder coupled to a shaft by a flexible coupling. Note the use of reference surfaces.

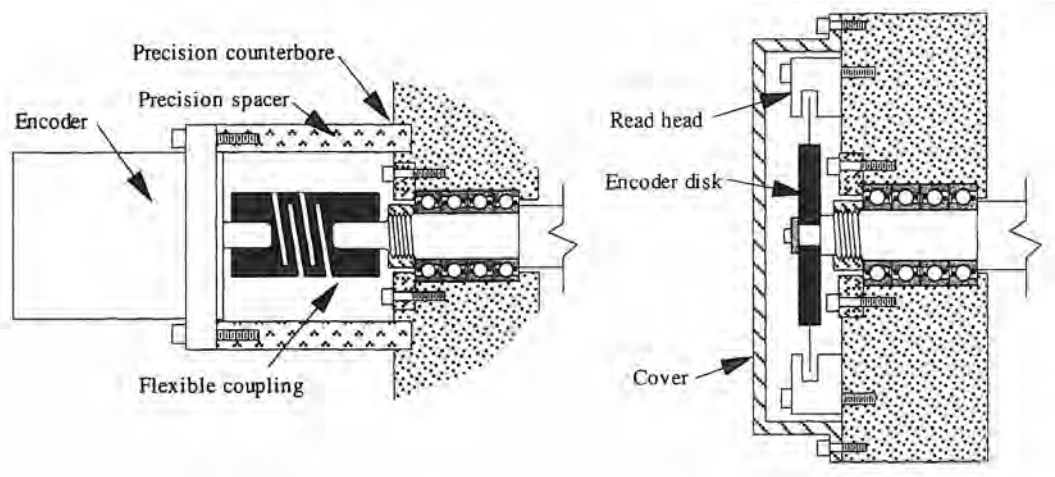


Figure 5.3.2 Mounting methods for rotary encoders.

For angular displacement measuring sensors, as long as there is no slippage in the coupling between the sensor and axis of motion, then there will be minimal error in measuring a complete (360^o) rotation,[2] but there may be a cyclic error in the measurement for angles between 0 and 360^o. The magnitude of the error and its function of position depends on the type of coupling used and the relative shaft eccentricity e.

A conservative generic method for visualizing the error is provided by the pin-in-slot model shown in Figure 5.3.3. Imagine a sensor's shaft of radius r_c has a pin sticking in its outer circumference. The pin extends into a slot in the shaft whose angle of rotation is to be measured. As the output shaft moves through an angle θ, during the first 180^o of motion, the sensor shaft will lead the output shaft by an amount ε_θ. From the law of cosines, the length l_e from the eccentric circle radius to the center of the pin is

$$\ell_e = \sqrt{e^2 + r_c^2 - 2e\, r_c \cos\theta} \tag{5.3.2}$$

From the law of sines

$$\frac{\sin \varepsilon\theta}{e} = \frac{\sin\theta}{\ell_e} \tag{5.3.3}$$

Hence the angular error ε_θ in the sensor output will be

$$\varepsilon_\theta = \sin^{-1}\left[\frac{e\,\sin\theta}{\sqrt{e^2 + r_c^2 - 2e\, r_c \cos\theta}}\right] \tag{5.3.4}$$

[2] See Section 10.7.5 for a discussion of flexural couplings and the error they may themselves cause.

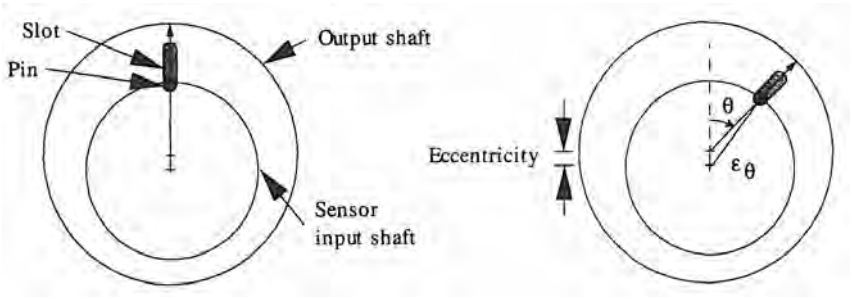


Figure 5.3.3 Pin-in-slot model of coupling rotational error caused by shaft eccentricity.

Fortunately, $e << r_c$, so the error can be approximated as

$$\varepsilon_\theta \approx \frac{e \sin\theta}{r_c} \tag{5.3.5}$$

The maximum error is thus

$$\varepsilon_\theta \approx \frac{e}{r_c} \tag{5.3.6}$$

In order to minimize the error and maximize the accuracy of the system, the coupling radius should be as large as possible. If the coupling is made from a flexural element, the error will be very repeatable and can be mapped. For example, the periodic axial positioning error in a system driven by a leadscrew is a function of periodic errors in the leadscrew and periodic error due to leadscrew-to-sensor shaft eccentricity.

As shown in Figure 5.3.2, the most accurate systems mount the sensor's measuring disk (e.g., the code disk on an optical encoder) onto the output shaft so that no coupling error due to shaft eccentricity is present. There can still be an eccentricity between the sensor disk and the output shaft, but now the error will be a function of the sensor disk's radius. If the sensor has two measuring elements placed 180^o apart, then averaging the element's signals will cause their periodic errors to cancel each other; however, this method has no effect on the error in a system that uses a shaft coupling between the sensor and the output shaft because the coupling introduces the error.

For both linear and angular sensor coupling designs, loads transferred across the interface are usually small, and repeatability and accuracy are of the utmost concern. Thus a flexural coupling element is usually preferred over sliding element couplings even though the former can also have some small hysteresis errors, which must be included in the error budget for the system.

5.4 Sensor Mounting Structure Design

When designing the mounting structure for a sensor, one must consider all the same design issues that went into design of the rest of the machine. For example, when the sensor is accelerated by the machine, will the acceleration-induced loads cause the mounting structure to deform? Too often sensors are mounted with *wimpy* brackets designed with incorrect assumptions, as shown figuratively in Figure 5.4.1, because some design engineers incorrectly assume that sensors only measure and are not subject to any loads. In noncritical, quasi-static, isothermal, vibration-free nonexistent applications this may be true, but these conditions are very rare. The sensor will usually experience thermal, static, dynamic, and vibration-induced loads; thus when designing the structural mounting system for a sensor, all the error sources discussed in Chapter 2 must be considered.

As with the rest of the machine, thermally induced errors in the sensor system are the most often overlooked. Problems usually occur when the sensor is mounted near a heat source, such as a motor or bearing housing. The design engineer must be very careful to consider the effects of a temperature change on the output of the sensor, and on local thermal deformations of the structure that may be measured by the sensor. Another important factor to consider is the thermal time constant of the sensor and structural systems. If a cast iron machine is used to machine cast iron, then as long as the entire system expands at the same rate, accuracy can be maintained. However, if the sensor system is not thermally coupled to the structural system, then chances are that the temperature

of the sensor system with its lower thermal mass will vary more quickly and often than that of the overall structure. If the machine is to be error mapped, this effect acts like an additional degree of freedom which complicates interpretation of the calibration measurements and development of the compensation algorithm.

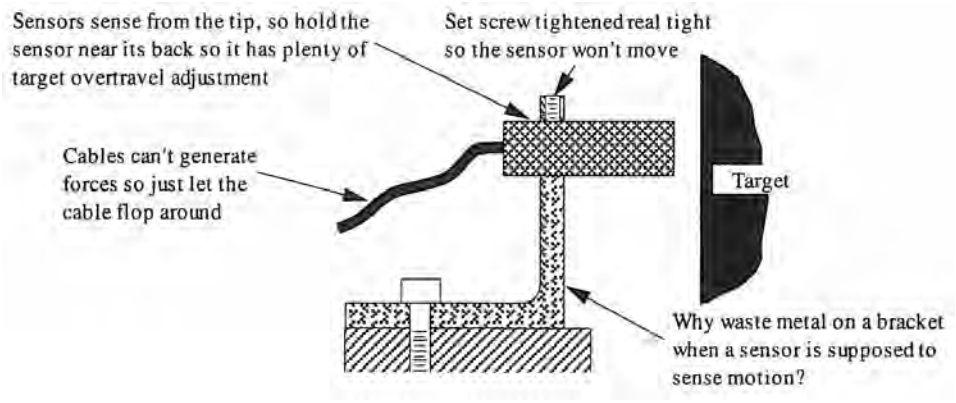


Figure 5.4.1 How not to mount a sensor, and some false assumptions.

Often the design of the sensor itself plays an important role in designing a suitable mount. For example, consider sensors such as LVDTs which have cores that move inside an outer housing. Ideally, the core fits inside the sensor body without contact between them. However, when the stroke becomes long and the sensors are used in a horizontal mode, the cantilevered core can sag and make contact with the housing. This problem can be remedied, as shown in Figure 5.4.2, with the use of a long rod held at both ends where the core is attached to the middle of the rod. Lubrication is to be avoided because it can attract dirt, which can lead to an abrasive situation. Some manufacturers[3] provide LVDTs where the core has a hard round tip which contacts the object to be measured, and the body of the core is supported either with an instrument-grade ball bushing or an air bearing. The latter type uses air pressure that leaks out around the shaft to act as a constant force spring.

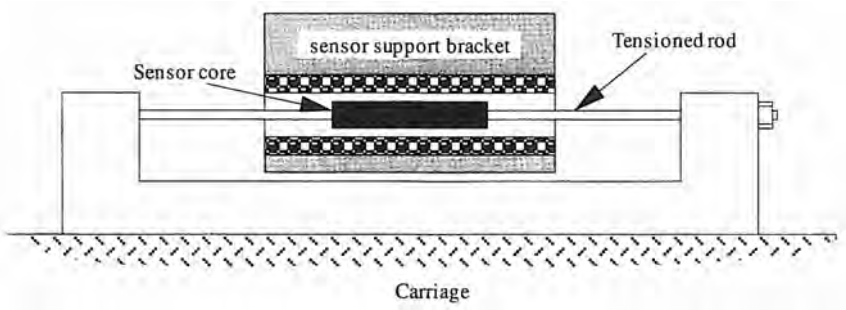


Figure 5.4.2 Method for preventing the core of a sensor from contacting sensor housing for long stroke sensors.

If the sensor is improperly secured to the mounting structure, seemingly insignificant distortions of the sensor's housing can cause a change in the sensor's output characteristics. Improper mounting implies the use of threaded body sensors with locking nuts which are "really torqued down good to make sure the sensor doesn't move." Threaded fasteners used with sensors should only be gently hand-tightened with plenty of lubrication applied to the threads. The assembly should then be potted in epoxy as shown in Figure 5.4.3. Preferably collet or split-ring clamping devices that apply uniform pressure to the sensor housing should be used, as described in Figure 5.4.4. The primary function of the fastener should be to hold the sensor gently in place while it is being aligned. Once the sensor is aligned, the system should be stress relieved (e.g., with a vibrator attached to the

[3] Air bearing LVDTs are available, for example, from Rank Taylor Hobson in Keene, NH; and Cranfield Precision Engineering Ltd. in Cranfield, Bedford, England.

structure), rechecked for alignment, and then fixed in a position with a low-shrinkage epoxy. This procedure will prevent residual stresses from causing the system to creep and offset the calibration. Furthermore, mounting a sensor in this stress-free manner minimizes the chances of damaging the sensor or altering its properties by distorting its structure.

If it is desired to measure small displacements in a harsh environment, such as spindle error motions or straightness of linear slides, consider using a reference surface that acts as a nonwearing component upon which an error map for the machine can be made and spindle error motions or straightness can be continuously monitored. This surface should be physically connected to the machine so it is subject to many of the same loads that cause error in other components.

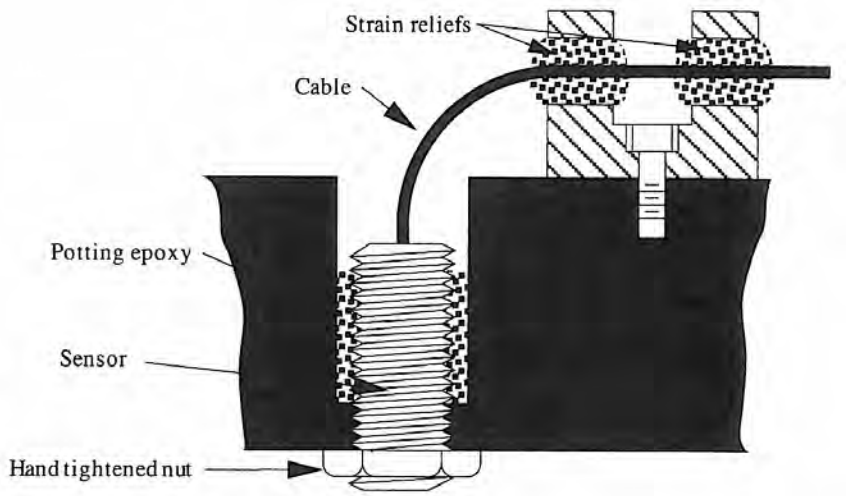


Figure 5.4.3 Method for mounting sensors with threaded bodies.

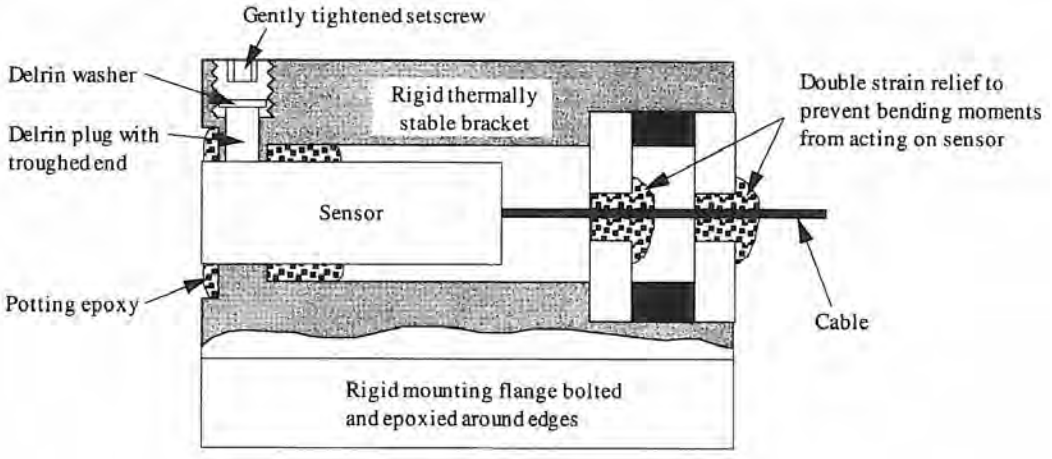


Figure 5.4.4 Method for mounting a smooth body sensor for ultraprecision applications. Alternatively, a large split housing can be used which applies uniform circumferential pressure to the sensor body.

In order to prevent contamination of the reference surface which may affect the sensor readings, a steady laminar stream of lubricant or cooling fluid or cutting oil can be used to keep the region clean, as shown in Figure 5.4.5. Of course the sensor will have to be calibrated with the fluid flowing in the gap between the sensor and the surface. To compensate for composition and temperature of the fluid, a reference sensor that continuously measures a fixed distance can be placed in the same fluid. Note that the dielectric constant of a fluid is much more sensitive to temperature changes than that of a gas; hence this method may not be appropriate for use with capacitance sensors.

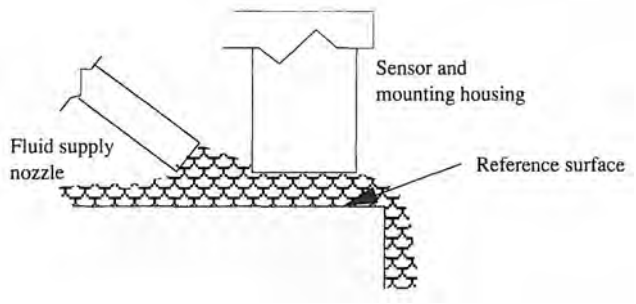


Figure 5.4.5 Use of a reference surface and cleansing fluid flow for continuous measurement of machine tool axis straightness in an adverse environment.

5.5 SENSOR MOUNTING ENVIRONMENT

Other than thermal effects, the primary environmental factor affecting sensor performance is the presence of dirt, coolant, and other debris often found in a machining environment. Many types of sealed unit sensors have been designed with these contaminants in mind. Still, the design engineer should always try and minimize the amount of punishment a sensor must endure. A contradiction of goals that design engineers must deal with, however, is that in order to decrease the Abbe error, the sensor should be mounted as close to the process as possible; yet the closer the sensor is mounted to the process, the harsher the environment.

In cases where extreme accuracy and repeatability are required, environmental conditions can be assessed by using a second sensor to measure the apparent distance to a stationary surface. If this "weather station" is properly designed so that its geometry is unaffected by environmental effects,[4] then the system will provide a measure of environmental effects on the sensor and medium through which it is sensing. To be effective, however, groups of sensors and their electronics must be matched to have similar characteristics, and the weather station must be located adjacent to the actual measurement being made, or else local atmospheric conditions[5] (e.g., an oil mist near a spindle) may degrade its effectiveness.

Electromagnetic interference (EMI) can also cause problems. The U.S. Navy learned about EMI the hard way in 1967 when a fighter plane taxiing on the deck of the *U.S.S. Forrestal* crossed a radio beam which caused one of the plane's missiles to launch. The ensuing explosion killed 134 sailors and caused $72 million in damage. Similarly, with the proliferation of electronic systems in the manufacturing environment, EMI can cause a computer-controlled machine to make an unexpected move which may ruin a part or the operator. Although the machine design engineer may not explicitly be required to design the EMI shielding for a machine,[6] he or she must provide space in the design to accommodate it.

One of the most common EMI problems results from crosstalk between power and signal lines that act as antennas. For example, because the current in the power lines to the servomotors varies with the type of operation the machine is performing, a sensor's power and signal lines should be shielded and run through conduit separate from power lines to the servomotors. As an example, consider the fact that motors powered by *pulse width modulation* (PWM) essentially act as radio transmitters, while sensors with wire-wound cores (e.g., LVDTs and LVTs) act like receivers. Sensors that require ac excitation should use a common oscillator or have individual oscillators that differ in frequency (but not even multiples). Individual wires and cables must also be shielded from each other to prevent crosstalk.

Another factor to consider is the effect of flexing of the signal cable on the output of a sensor. Flexing a cable changes properties, such as its capacitance, which can affect the output of some high-resolution sensors. There are three ways to avoid this problem: (1) do not flex the cables, (2)

[4] This can be done by making the weather station from a material with a zero coefficient of thermal expansion such as Invar or Zerodur.

[5] See for example N. Bobroff, "Residual Errors in Laser Interferometry from Air Turbulence and Nonlinearity," Appl. Opt., Vol. 26, No. 13, 1987, pp. 2676–2681.

[6] A good overview of methods for combating EMI is given by D. Bahniuk, "New Weapons to Combat EMI," Mach. Des., May 21, 1987, pp. 99–103.

choose a sensor system which is not affected by flexing cables, or (3) convert the sensor output to a digital signal before sending it through a flexing cable.

Sensors which make contact with the target surface (e.g., spring-loaded LVDT probes) must contend with the fact that all surfaces in contact deform even under the slightest loads. In the elastic region, the amount of deformation is a function of the load, component geometry, and material type. The deformation occurs because wherever curved surfaces contact other surfaces, point or line contact exists and infinite stresses cannot exist because all materials have a finite amount of elasticity; thus the surfaces at these contact points deform until an equilibrium contact area is attained, as discussed in the next section.

5.6 CONTACT BETWEEN CURVED SURFACES

Many reference books give graphs and equations for determining the contact stress and elastic deformation that exist between two bodies in contact. Often equations are not given for the case at hand, or it requires graphical interpretation. This is not suitable for numerical analysis; thus in this section equations suitable for back-of-the-envelope calculation of stress and deflection at the contact interface between a noncylindrical object and any other object are presented first. The "exact" equations are then discussed which can be implemented on a spreadsheet. Regardless of which formulas are used, consider that the magnitudes of deflections of two bodies in contact are often in the submicron range; thus surface finish characteristics can play an important role.

Approximate Solution for Point Contact Between Objects

The gap bending hypothesis states that *the effect of geometry on the system in the contact region is a function of the algebraic sum of the curvatures of the two surfaces in contact. Thus the contact between two curved surfaces can be approximated by an equivalent contact problem between a sphere and a plane for which the solution is known.*[7] This is useful for quick back-of-the-envelope engineering estimates. The first step is to determine the equivalent modulus of elasticity of the system based on the elastic modulii and Poisson ratios of the two materials in contact:

$$E_e = \frac{1}{\frac{1-\eta_1^2}{E_1} + \frac{1-\eta_2^2}{E_2}} \tag{5.6.1}$$

Next, the equivalent radius of the system is found from the gap bending hypothesis. Note that convex surfaces' (e.g., a ball) radii are positive, concave surfaces' (e.g., a groove) radii are negative, and flat plane radii are infinite.

$$R_e = \frac{1}{\frac{1}{R_{1_{major}}} + \frac{1}{R_{1_{minor}}} + \frac{1}{R_{2_{major}}} + \frac{1}{R_{2_{minor}}}} \tag{5.6.2}$$

The radius of an equivalent circular contact area between the two bodies is given by

$$a = \left(\frac{3FR_e}{2E_e}\right)^{1/3} \tag{5.6.3}$$

Note that for many cases the contact zone will be elliptical, and hence if one is interested in the size and shape of the contact zone (e.g., for calculating the amount of slip in a rolling body), the general contact-between-bodies solution must be used. The deflection of the system due to elastic deformation of the bodies at the contact interface is

$$\delta = \frac{1}{2}\left(\frac{1}{R_e}\right)^{1/3}\left(\frac{3F}{2E_e}\right)^{2/3} \tag{5.6.4}$$

Often the amount of deformation will be small compared to the desired accuracy of the system, or it can be calculated to a reasonable degree so it can be accounted for. In other instances, repeatability is all that is being measured, so as long as the force variation over the range of travel of the probe is small, the changing amount of deformation will be negligible.

[7] J. Tripp, "Hertzian Contact in Two and Three Dimensions," NASA Tech. Paper 2473, July 1985.

As an example of the magnitude of contact deformations, consider the case of a 3 mm (0.118 in.) diameter steel ball probe used to measure the location of a steel surface with the applied force on the order of 1 N (0.22 lbf). The equivalent modulus of elasticity for the system is about 110 GPa (16×10^6 psi). The equivalent radius is 0.75 mm (0.030 in.). The diameter of the equivalent circular contact region (which in this case actually is circular) is 21.7 μm (854 μin.). The deflection, or error in measurement, is about 0.31 μm (12.2 μin.). Note that a variation in the force by 10% causes a change in the deflection by only about 6% or 0.021 μm (0.83 μin.). If the characteristic surface finish dimension of the part is much larger than the radius of the contact area, then measurement errors will be dominated by surface finish effects. Often when making precision measurements in the submicron range, a region of the part must be polished, or a polished block must be attached to the surface.

Allowable Contact Stress

In order to avoid damaging the surfaces, the contact stress must be kept below the elastic limit. The maximum contact pressure is at the center of the interface. For a sphere on a flat plate, or a system equated to a sphere on a flat plate, the Hertz contact stress (contact pressure) is

$$q = \frac{1}{\pi}\left(\frac{1}{R_e}\right)^{2/3}\left(\frac{3E_e^2F}{2}\right)^{1/3} = \frac{aE_e}{\pi R_e} \tag{5.6.5}$$

The stress at the center of the contact circle of radius a and as a function of depth z (z from 0 $\rightarrow +\infty$), and contact pressure q is

$$\sigma_z(z) = q\left\{-1 + \frac{z^3}{(a^2+z^2)^{3/2}}\right\} \tag{5.6.6a}$$

$$\sigma_r(z) = \sigma_\theta(z) = \frac{q}{2}\left\{-(1+2\eta) + \frac{2(1+\eta)z}{\sqrt{a^2+z^2}} - \frac{z^3}{(a^2+z^2)^{3/2}}\right\} \tag{5.6.6b}$$

To predict failure in metals, the shear stress is often used

$$\tau = \frac{\sigma_\theta - \sigma_z}{2} = \frac{q}{2}\left\{\frac{1-2\eta}{2} + \frac{(1+\eta)z}{\sqrt{a^2+z^2}} - \frac{3z^3}{2(a^2+z^2)^{3/2}}\right\} \tag{5.6.7}$$

The maximum shear stress τ_{max} occurs at a depth z:

$$z = a\sqrt{\frac{2(1+\eta)}{7-2\eta}} \tag{5.6.8a}$$

$$\tau_{max} = \frac{q}{2}\left\{\frac{1+2\eta}{2} + \frac{2}{9}(1+\eta)\sqrt{2(1+\eta)}\right\} \tag{5.6.8b}$$

Note that when η = 0.3, z = 0.637a, and τ_{max} = 0.333q. σ_z, σ_r, and τ are plotted in Figure 5.6.1, where they are normalized in terms of the contact pressure q and contact circle radius a. The severe stress state exists in the body with the larger radius of curvature and hence only values below the contact surface are plotted. For metals, one can assume $\tau_{max\ allowable} = 1/2\sigma_{max\ tensile}$. Therefore, for metals, one can conservatively assume that the allowable Hertz contact stress[8] (contact pressure) is

$$q_{Hertz\ metals\ max} = \frac{3\sigma_{allowable\ tensile\ stress}}{2} \tag{5.6.9}$$

For the example above, the contact pressure is 1012 MPa (147 ksi) and the maximum shear stress is 337 MPa (49 ksi). For hardened steel, this is an acceptable level, but for mild steel or aluminum, one must be very careful to keep the contact force low, or high stresses and permanent microscopic dents in the surface will result. This is known as Brinelling.

For brittle materials (e.g., ceramics), the flexural strength is the limiting design factor. For brittle materials, one can assume that the allowable Hertz contact stress[9] is:

$$q_{Hertz\ brittle\ materials\ max} = \frac{2\sigma_{allowable\ flexural\ stress}}{1-2\eta} \tag{5.6.10}$$

[8] See R. J. Roark and W. C. Young, Formulas for Stress and Strain 5th ed., McGraw-Hill Book Co., New York, 1975.

[9] From conversations with John Lucek of Cerbec Bearing Company, 10 Airport Park Road, East Grandby, CT 06026.

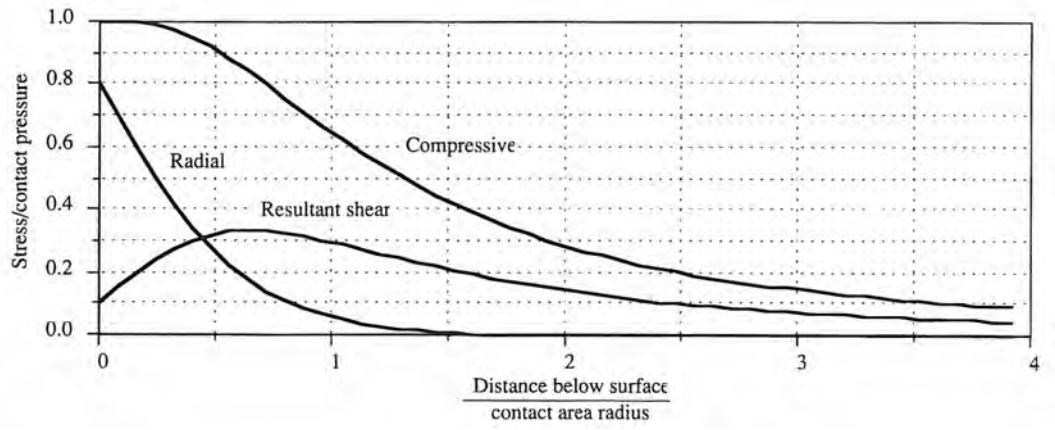


Figure 5.6.1 State of stress in a material below the contact region.

Note that there is a volume effect for failure in materials. As the size of the specimen decreases, the likelihood of a defect initiating failure decreases. Thus in practice, the allowable Hertz stresses may be 50% greater than Equations 5.6.9 and 5.6.10 allow for. For example, hot pressed silicon nitride bearing components can withstand millions of cycles at a Hertz stress level of 6.9 GPa (10^6 psi), and the single cycle failure load is often in the neighborhood of 10 GPa (1.5×10^6 psi). Note that it has been observed that the life of a rolling ball or of a rolling cylinder is inversely proportional to the cube of the load or the fourth power of the load, respectively (the eighth and ninth power of the Hertz stress). For example, 12.5 mm (0.5 in.) RC 63-64 steel balls withstood 17.5 million cycles at a stress of 1.20 GPa (174 ksi) before 10% of the balls failed, and 700 million cycles before 90% of the balls failed).[10]

Rolling Contact Stress Considerations

When the bodies are in rolling contact (e.g., a capstan drive), the tangential stress component due to the tractive force may have to be considered. The theory and results are complex[11] but they show how the addition of a tangential stress can lead to the formation of tensile stresses in the surface. When one purchases precision modular bearings, one can usually be confident that the manufacturer has taken the complete stress state into account in the design of the bearings. For new designs, as an initial engineering estimate of the effect of the tangential stresses on the maximum resultant shearing stress, assume the following:

1. The maximum tangential load (tractive effort) that can be supported is equal to the product of the coefficient of friction μ and the normal load.
2. The tangential stress is equal to the tangential load divided by the contact area.
3. For metals, the maximum shear stress is equal to the sum of the shear stresses due to the normal and tangential loads.
4. For brittle materials (e.g., ceramics), the allowable flexural stress, which is used with Equation 5.6.10 to determine the allowable Hertz stress, is decreased by an amount equal to the tangential stress.

At high contact pressures, which occur when high preloads are applied to achieve stiffness or when high normal loads occur, it is best to assume that the asperities in the contact region are in intimate contact, so μ is ≈ 0.30. This generally corresponds to the coefficient of friction for an unlubricated material on a similar material. If by this conservative estimate an unrealizable design is obtained, then the "exact" solution obtained by Liu should be used.

Fretting Corrosion

In addition to high stress at the contact interface, fretting corrosion must be guarded against. Fretting can occur when submicron-size surface features (asperities) of similar materials are forced

[10] H. Styri, "Fatigure Strength of Ball Bearing Races and Heat Treated Steel Specimens," Proc. ASTM, Vol. 51, 1951.

[11] See J. Smith and C. Liu, "Stress Due to Tangential and Normal Loads on an Elastic Solid with Application to Some Contact Stress Problems," J. Appl. Mech., June 1953, pp. 157–166. Also see F. Seely and J. Smith, Advanced Mechanics of Materials, John Wiley & Sons, New York, 1952.

into repeated contact without a lubricant between them. Note that vibration can cause microscopic motion of the surfaces, which is enough to make bearings wear through a lubrication layer. Each time the surfaces come in contact, the asperities can bond to each other via interatomic forces. When the surfaces are pulled apart, the asperities are ripped apart which exposes fresh metal that allows the process to be repeated. Thus care must be taken to use inert materials if possible (e.g., ceramics, a synthetic ruby, or hardened stainless steel) for sensor probe heads. For bearings, it should be noted that if an axis is locked in position for an extended period of time while the machine is subject to vibration, the bearings in rotary and linear axes and in ballscrews can fret the surfaces they contact. Upon cleaning and inspection, little dimples can be seen that may be mistaken for Brinelling. Fretting can only be avoided with the use of dissimilar materials.

"Exact" Solution for Point Contact between Objects

The derivation of the "exact" form of the equations for point contact between two bodies was made by Hertz in 1881.[12] Assumptions in the theory include:

- The bodies are linear elastic.
- The contact area is small with respect to the smallest radius of curvature (ratio is < 0.1).

A detailed discussion of the derivation is beyond the scope of this book and the results require the evaluation of complete elliptic integrals.[13] Evaluation of the elliptic integrals is done with an infinite series and thus can be quite time consuming. However, these integrals have been tabulated for a range of cases and thus one can use the "exact" solution without too much difficulty.[14]

The equivalent modulus of elasticity E_e and radius R_e are defined as in Equations 5.6.1 and 5.6.2, respectively. A function $\cos\theta$ is defined as

$$\begin{aligned}\cos\theta &= R_e\left[\left(\frac{1}{R_{1_{major}}}-\frac{1}{R_{1_{minor}}}\right)^2+\left(\frac{1}{R_{2_{major}}}-\frac{1}{R_{2_{minor}}}\right)^2\right.\\ &+ \left. 2\left(\frac{1}{R_{1_{major}}}-\frac{1}{R_{1_{minor}}}\right)\left(\frac{1}{R_{2_{major}}}-\frac{1}{R_{2_{minor}}}\right)\cos 2\phi\right]^{1/2}\end{aligned} \tag{5.6.11}$$

where ϕ is the angle between the planes of principal curvature of the two bodies as shown in Figure 5.6.2. For example, for a friction drive roller on a round drive bar, $\phi = 90^o$. The function $\cos\theta$ is then used to find the factors α, β, and λ whose values are given in Table 5.6.1. Fifth-order polynomials can be used to represent α, β, and λ which makes the "exact" solution to the Hertzian contact problem more amenable to spreadsheet analysis. Table 5.6.2 gives the coefficients for the polynomials. Alternatively, using the arc cosine function, one can work with much simpler expressions.

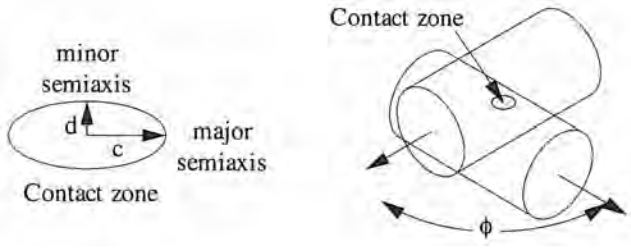


Figure 5.6.2 Two bodies in contact.

$$\alpha = A0_\alpha + A1_\alpha\cos\theta + A2_\alpha\cos^2\theta + A3_\alpha\cos^3\theta + A4_\alpha\cos^4\theta + A5_\alpha\cos^5\theta \tag{5.6.12a}$$

$$\beta = A0_\beta + A1_\beta\cos\theta + A2_\beta\cos^2\theta + A3_\beta\cos^3\theta + A4_\beta\cos^4\theta + A5_\beta\cos^5\theta \tag{5.6.12b}$$

[12] When the problem approaches that of a sphere indenting an elastic cavity ($R_1/R_2 < 1.1$), the Hertz theory assumptions start to break down. In this region, Hertz theory overestimates the deflection; thus for precision machine designers, Hertz theory provides a conservative estimate of point contact deflections. See, for example, L. E. Goodman and L. M. Keer, "The Contact Stress Problem for an Elastic Sphere Indenting an Elastic Cavity," Int. J. Solids Structures, Vol. 1, 1965, pp. 407–15.

[13] See F. Seely and J. Smith, Advanced Mechanics of Materials, John Wiley & Sons, New York, 1952.

[14] See R. J. Roark and W. C. Young, Formulas for Stress and Strain 5th ed., McGraw-Hill Book Co., New York, 1975.

Cosθ	0.000	0.100	0.200	0.300	0.400	0.500	0.600	0.700	0.750	0.800	0.850	0.900	0.920	0.940	0.960	0.980	0.990
α	1.000	1.070	1.150	1.242	1.351	1.486	1.661	1.905	2.072	2.292	2.600	3.093	3.396	3.824	4.508	5.937	7.774
β	1.000	0.936	0.878	0.822	0.769	0.717	0.664	0.608	0.578	0.544	0.507	0.461	0.438	0.412	0.378	0.328	0.287
λ	0.750	0.748	0.743	0.734	0.721	0.703	0.678	0.644	0.622	0.594	0.559	0.510	0.484	0.452	0.410	0.345	0.288

Table 5.6.1 Values of α, β, and λ as a function of cosθ

	$\cos\theta < 0.9$	$\cos\theta \leq 0.9$		$\cos\theta < 0.9$	$\cos\theta \leq 0.9$		$\cos\theta < 0.9$	$\cos\theta \leq 0.9$
$A0_\alpha$=	0.99672	−4522789.91	$A0_\beta$=	1.00000	51254.01	$A0_\lambda$=	0.75018	70770.70
$A1_\alpha$ =	1.27860	24146274.74	$A1_\beta$ =	−0.68865	−273306.28	$A1_\lambda$ =	−0.04213	−378446.23
$A2_\alpha$ =	−6.72010	−51557740.00	$A2_\beta$ =	0.58909	582926.00	$A2_\lambda$ =	0.29526	809436.14
$A3_\alpha$ =	27.37900	55036391.00	$A3_\beta$ =	−1.32770	−621625.00	$A3_\lambda$ =	−1.75670	−865562.50
$A4_\alpha$ =	−41.82700	−29371139.00	$A4_\beta$ =	1.77060	331436.01	$A4_\lambda$ =	2.67810	462760.42
$A5_\alpha$ =	23.47200	6269014.53	$A5_\beta$ =	−0.99887	−70684.52	$A5_\lambda$ =	−1.55330	−98958.33

Table 5.6.2 Values of polynomial coefficients for fifth order curve fits to α, β, and λ.

$$\lambda = A0_\lambda + A1_\lambda \cos\theta + A2_\lambda \cos^2\theta + A3_\lambda \cos^3\theta + A4_\lambda \cos^4\theta + A5_\lambda \cos^5\theta \tag{5.6.12c}$$

$$\alpha = 1.939e^{-5.26\theta} + 1.78e^{-1.09\theta} + 0.723/\theta + 0.221 \tag{5.6.12d}$$

$$\beta = 35.228e^{-0.98\theta} - 32.424e^{-1.0475\theta} + 1.486\theta - 2.634 \tag{5.6.12e}$$

$$\lambda = -0.214e^{-4.95\theta} - 0.179\theta^2 + 0.555\theta + 0.319 \tag{5.6.12f}$$

The dimensions of the major (c) and minor (d) semiaxis of the elliptical contact area are given by

$$c = \alpha \left(\frac{3FR_e}{2E_e}\right)^{1/3} \quad d = \beta \left(\frac{3FR_e}{2E_e}\right)^{1/3} \tag{5.6.13}$$

The contact pressure q is given by

$$q = \frac{3F}{2\pi cd} \tag{5.6.14}$$

Note that once the contact pressure is evaluated, it can be used with Equations 5.6.6 - 5.6.10 to evaluate the stress state below the surface. For contact at a single interface (e.g., an elastic hemisphere on an elastic flat plate), the distance of approach of two far field points in the bodies is

$$\delta = \lambda \left(\frac{2F^2}{3R_e E_e^2}\right)^{1/3} \tag{5.6.15}$$

As the bodies approach sphericity, the approximate (gap bending hypothesis) and "exact" solutions given in Equations 5.6.13-5.6.15 converge. When the bodies are far from being spherical (e.g., the case of a friction drive roller in contact with a round bar), the gap bending hypothesis yields conservative results: The gap bending hypothesis gives estimates for the stress and deflection that are much higher (by up to 30%) than they really are; hence they yield a more conservative design.

Line Contact Between Objects

A similar set of equations exists for the contact stresses between surfaces with only one radius of curvature each. If the cylinder's axes are not parallel, then the analysis is quite complex,[15] but for most applications the axes are parallel. For a cylinder of length L and diameter d_1 loaded by a

[15] See earlier references and J. Lubkin "Contact Problems," in Handbook of Engineering Mechanics, W. Flugge (ed.), McGraw-Hill Book Co., New York, 1962. Also see T. Harris, Rolling Bearing Analysis, John Wiley & Sons, Inc., New York, 1991.

force F/L along its length and in contact with a cylinder[16] of diameter d_2, the contact area between cylinders is a rectangle of width 2b

$$b = \left(\frac{2Fd_1d_2}{\pi LE_e(d_1 + d_2)} \right)^{1/2} \tag{5.6.16}$$

The expression for the deflection of a cylinder is more complicated than for the case of a sphere because of end effects. For a cylinder of diameter d_1 compressed between two flat rigid surfaces (E_2 is infinite), the displacement of one of the contact surfaces relative to the center of the cylinder is

$$\delta_{cylinder} = \frac{2F}{\pi LE_e} \left[Log_e \left(\frac{2d_1}{b} \right) - \frac{1}{2} \right] \tag{5.6.17}$$

The diametrical shortening would be twice this value. Note that d_2 and E_2 are assumed infinite in Equation 5.6.16, which is used in the evaluation of Equation 5.6.17. For the case of two rollers in contact, determination of the total relative motion between their centers would require Equation 5.6.17 to be evaluated once for each cylinder and then the results summed.

For an elastic cylinder on an elastic flat plate, Equation 5.6.17 blows up, which is a peculiar result of the plane stress theory used to obtain the equation. To determine the displacement of the center of a cylinder with respect to a point at distance d_o below the surface, a two-part solution is required. The first part is the displacement due to the deformation of the cylinder as given by Equation 5.6.17. The second part is for the deformation of the elastic flat plate as a rigid cylinder is pressed into it:

$$\delta_{flat} = \frac{2F}{\pi LE_e} \left[Log_e \left(\frac{2d_o}{b} \right) - \frac{\eta}{2(1-\eta)} \right] \tag{5.6.18}$$

Typically, d_o is set equal to the cylinder diameter. Note that in this case E_1 is infinite in Equation 5.6.1. The total deflection of a system composed of an elastic cylinder compressed between *two* elastic flat plates is thus

$$\delta_{system} = 2(\delta_{cylinder} + \delta_{flat}) \tag{5.6.19}$$

The maximum contact pressure q in any of the cylindrical contact cases is

$$q = \frac{2F}{\pi bL} \tag{5.6.20}$$

With the x axis aligned along the cylinder's axes, the stresses as a function of position on the Z axis (z from 0 to + ∞) are

$$\sigma_X = -2q\eta \left\{ \left(1 + \frac{z^2}{b^2} \right)^{1/2} - \frac{z}{b} \right\} \tag{5.6.21}$$

$$\sigma_Y = -q \left\{ \left(2 - \frac{b^2}{b^2 + z^2} \right) \left(1 + \frac{z^2}{b^2} \right)^{1/2} - \frac{2z}{b} \right\} \tag{5.6.22}$$

$$\sigma_Z = -q \left(\frac{b^2}{b^2 + z^2} \right)^{1/2} \tag{5.6.23}$$

In order to determine the yield condition, each of the shear stresses would have to be determined as a function of the position on the Z axis:

$$\tau_{YX} = \frac{\sigma_Y - \sigma_X}{2}, \quad \tau_{ZX} = \frac{\sigma_Z - \sigma_X}{2}, \quad \tau_{ZY} = \frac{\sigma_Z - \sigma_Y}{2} \tag{5.6.24}$$

The maximum shear stress value is $\tau_{ZY} \approx 0.3q$ at $z/b \approx 0.786$. When plotted as a function of depth below the surface, similar curves are obtained as for the case of spherical bodies. Once again, the most severe stress state is in the body with the larger radius of curvature.

[16] For cylinders, the second radius of curvature is infinite. For a cylinder in contact with a flat plane, $d_2$is infinite; d_2 is negative for a concave surface.

Tangential Stiffness of the Contact Interface

Studies have also been made to determine the tangential displacement of bodies in point contact that are also subject to tangential loads. In these situations, it is assumed that the tangential force is aligned with respect to the contact ellipse along a semiaxis of dimension a. This axis may be parallel to the major (c) or minor (d) semiaxis of the contact ellispe.

For the case of a circular contact region and equal modulii of elasticity, Mindlin gives the following solution for when there is no slip between the bodies:[17]

$$\delta_{tan} = \frac{F_{tan}(2-\eta)(1+\eta)}{4aE} \tag{5.6.25}$$

When there is slip between the bodies, the shear traction must be limited to a value equal to the product of the coefficient of friction and the contact pressure. For the general case of elliptical contact between two curved bodies made from the same material, Deresiewicz provides a solution.[18] The semiaxes (a' and b') of the inner ellipse (inside of which there is no slip) are given by

$$\frac{a'}{a} = \frac{b'}{b} = \left(1 - \frac{F_{tan}}{\mu F}\right)^{1/3} \tag{5.6.26}$$

where F is the contact (e.g., preload) force. The tangential displacement of a point on either body with respect to a far-field point on the body is given by

$$\delta_{tan} = \frac{3\mu F(2-\eta)(1+\eta)}{8aE}\left[1 - \left(1 - \frac{F_{tan}}{\mu F}\right)^{2/3}\right]\Phi \tag{5.6.27a}$$

$$\delta_{tan\,unload} = \frac{3\mu F(2-\eta)(1+\eta)}{8aE}\left[2\left(1 - \frac{T^* - F_{tan}}{2\mu F}\right)^{2/3} - \left(1 - \frac{T^*}{\mu F}\right)^{2/3} - 1\right]\Phi \tag{5.6.27b}$$

where T^* is the initial tangential force, F_{tan} is the new lower force, and Φ is given by

$$\Phi = \left[\frac{4a}{\pi b(2-\eta)}\right]\left[\left(1 - \frac{\eta}{k^2}\right)\mathbf{K} + \frac{\eta \mathbf{E}}{k^2}\right] \quad a < b \tag{5.6.28a}$$

$$\Phi = 1 \quad (\text{spherical contact}) \quad a = b \tag{5.6.28b}$$

$$\Phi = \left[\frac{4}{\pi(2-\eta)}\right]\left[\left(1 - \eta + \frac{\eta}{k_1^2}\right)\mathbf{K}_1 + \frac{\eta \mathbf{E}_1}{k_1^2}\right] \quad a > b \tag{5.6.28c}$$

Once again, take care to note that the value of a is the dimension of the semiaxis *parallel* to the direction of the applied tangential force. Also note that the material constants η and E are for the body (e.g., roller or drive bar in a friction drive) of interest. The constants k and k_1 are the arguments of the complete elliptic integrals $\mathbf{K}$, $\mathbf{E}$, and $\mathbf{K}_1$, $\mathbf{E}_1$, respectively. Figure 5.6.3 shows the value of Φ for various values of a/b computed using exact values of the elliptic integrals for a/b obtained from Mathematica®. The nature of the elliptic functions causes an asymptotic "kink" near the transition point of $a/b = 1$. To smooth the curve, the first 150 terms of the series representing the elliptic integrals were used in their calculation and the resultant value of Φ was scaled by the endpoint of the exact value. This yields polynomials of reasonable accuracy that can be used for spreadsheet analysis:

$$\begin{aligned}\Phi \;=\; & 0.13263 + 1.4325(a/b) - 0.54754(a/b)^2 + 0.12303(a/b)^3 \\ & -0.013591(a/b)^4 + 0.0005729(a/b)^5 \text{ for } 0.1 \le a/b \le 8\end{aligned} \tag{5.6.29a}$$

$$\begin{aligned}\Phi \;=\; & 1.9237 + 0.11029(a/b) - 2.8323 \times 10^{-3}(a/b)^2 + 4.3109 \times 10^{-5}(a/b)^3 \\ & -3.3497 \times 10^{-7}(a/b)^4 + 1.0257 \times 10^{-9}(a/b)^5 \text{ for } 8 < a/b < 90\end{aligned} \tag{5.6.29b}$$

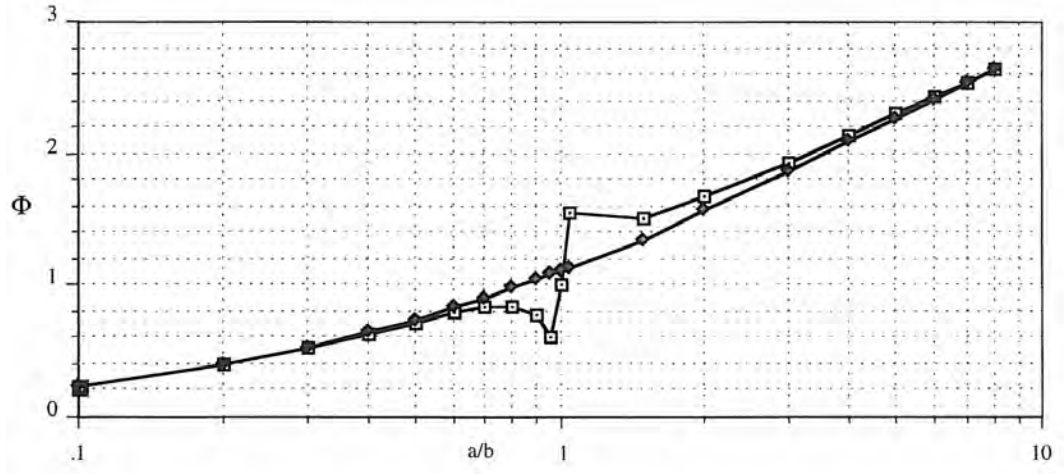


Figure 5.6.3 Φ from Equation 5.6.29 as a function of the ratio of semiaxis dimensions.

If the tangential force applied varies from a negative to a positive value, but always with a magnitude less than μF, then a hysteresis loop will be formed. The area enclosed in the loop is a measure of the energy dissipated by friction in one cycle of loading.

From Equation 5.6.27a the stiffness of each of the two contacting bodies is found as the inverse of the compliance:

$$K_{tan} = \frac{1}{\frac{\partial \delta_{tan}}{\partial F_{tan}}} = \frac{4Ea}{(2-\eta)(1+\eta)\Phi}\left(1 - \frac{F_{tan}}{\mu F}\right)^{1/3} \tag{5.6.30}$$

This equation reflects the condition where the slip zone size increases with the tangential force. For example, for a friction drive, Equation 5.6.30 can be used to determine the approximate stiffness of the contact zone on the roller and then the drive bar. The net stiffness of the system will be the stiffness of two springs in series. The use of these equations for friction drive design is discussed in Section 10.8.2.

5.7 METROLOGY FRAMES

It is very difficult to attain submicron accuracy because even the slightest change in the applied force on a large structure can often cause a significant deformation to occur. When axes are stacked on top of each other, as well as the respective sensors on the axes, the sensors only measure the degree of freedom of the axis they are attached to. The sensors do not necessarily measure the motions of other axes, even though they may be in line with a sensor's measuring range. In order to uncouple the sensors from the structure, a separate stationary reference frame is required whose position the machine tool's axes are measured with respect to. This type of stationary reference structure is known as a *metrology frame*.

The first documented use of a metrology frame was in the Rogers-Bond Universal Comparator developed in 1883.[19] One of the original motivating factors for its design was to allow for an accurate comparison of the yard and meter standards in use at the time. Microscopes used to view the objects were mounted on a separate structure as per Rogers' requirement for "the complete separation of the standards to be compared from the framework to which the microscopes are attached." Interestingly enough, this concept remained relatively unused until recently when the high accuracy requirements of some numerically controlled lathes that used diamond cutting tools (diamond turning machines) began to exceed the capabilities of conventional designs.

[17] R. Mindlin, "Compliance of Elastic Bodies in Contact," J. Appl. Mech., Vol. 16, 1949, pp. 259–28.

[18] H. Deresiewicz, "Oblique Contact of Nonspherical Bodies," J. Appl. Mech., Vol. 24, 1957, pp. 623–64.

[19] See Chris Evans, Precision Engineering: An Evolutionary View, Cranfield Press, Cranfield, Bedford, England.

The design considerations for a metrology frame are similar to those for the machine tool itself. The primary difference being that ideally the metrology frame is a static structure that acts as a fixed reference frame for sensors to be mounted to, or as a reference surface for moving sensors to measure against. A properly designed metrology frame will be unaffected by dynamic or static loads within the machine. As the machine's axes move with respect to the metrology frame, linear and angular displacement measurements of the structural system are made with respect to the metrology frame. If direct measurements of the position of the tool tip cannot be made, which is often the case, then straightness and angular error displacements are measured so that the actual position of the tool tip can then be calculated using the methods described in Section 2.2. It should also be noted that it is desirable to make the metrology frame as small as possible to minimize environmental effects. This also helps to minimize the structural loop discussed in Section 7.4.

5.7.1 Design of the Large Optics Diamond Turning Machine[20]

As shown in Figure 5.7.1 the large optics diamond turning machine (LODTM), built at the Lawrence Livermore National Laboratory (LLNL), is a vertical spindle bridge-type (portal) machine. It was designed to machine large optical components (e.g., mirrors for telescopes) using a diamond tool, to an accuracy of 0.028 μm rms (1.1 μin.) in the work volume with a surface finish on the order of 42 Å R_a (0.17 μin.). This would allow infrared optics to be machined without the need for subsequent polishing. The machine was designed with a vertical spindle to minimize nonaxisymmetric deformations of the part being machined. The part can have maximum weight of 13.4 kN (3000 lbf) and dimensions of 1.63 m diameter by 0.51 m thick (64 × 20 in.). Since LODTM is primarily a finishing machine, the rate of chip production during the machining process is low enough to enable chips to be easily removed from the nominally horizontal work surface.

LODTM's structural system is made from welded mild steel plates. The metrology frame is made from Super Invar and it is kinematically mounted to the structural frame with three flexures, so that thermal and mechanical loads cannot be transmitted between the two systems. Variable forces, such as the stretching of the bellows enclosing the interferometers, are supported by attachment of the fixed ends to the structural frame. Internal passages allow 100 gpm of water, temperature controlled to 0.001 F^o, to be circulated through the structure to control its temperature. The external air temperature can be controlled to 0.01 F^o when the machine is operated remotely. The X axis, referred to as the *carriage*, moves horizontally (radially along the part) and carries the Z axis. The Z axis, referred to as the *toolbar*, moves the tool vertically along the axial direction of the part. As shown in Figure 5.7.2, seven measurements are made by the interferometric measuring system. The resolution of the interferometric system is about 6 Å (0.025 μin.). The position of the metrology frame with respect to the faceplate of the spindle is continually measured using four capacitance probes. With the set of 11 measurements taken, the metrology frame can undergo small rigid-body motions in the XZ plane without changing the measured relative position of the tool and workpiece. This system allows the position of the toolpoint to be determined accurately with respect to the workpiece even though the structural frame may deform under load by as much as $6^1/_4$ μm (250 μin.).

The toolbar's X direction straightness and pitch and the carriage's axial position and pitch are measured using two straightedges vertically mounted to the toolbar and four interferometers mounted on the metrology frame. The straightedges move vertically with the toolbar and are kinematically mounted so that deformations and accelerations of the toolbar do not cause appreciable deformation of the straightedges. The difference between upper and lower measurements is used to determine the combined pitch of the carriage and toolbar, which allows the Abbe error at the tool tip to be determined. Their average is a measure of the combined X direction straightness of the toolbar and the X position of the carriage. Averaging the measurements from each side allows symmetric expansion errors of the metrology frame to be rejected. Note that Y direction straightness and yaw of the carriage cause motion of the tool tip in nonsensitive directions and therefore do not need to be measured.

[20] See R. Donaldson and S. Patterson, "Design and Construction of a Large Vertical-Axis Diamond Turning Machine," SPIE's 27th Ann. Int. Tech. Symp. Instrum. Display, Aug. 21-26, 1983 (also available as a technical report from NTIS, UCRL-89738). Also see J. B. Bryan, "Design and Construction of an Ultra Precision 84 Inch Diamond Turning Machine," Precis. Eng., Vol. 1, No. 1, 1979, pp. 13–17. The author is very grateful to Dr. Robert Donaldson of LLNL for reviewing this section.

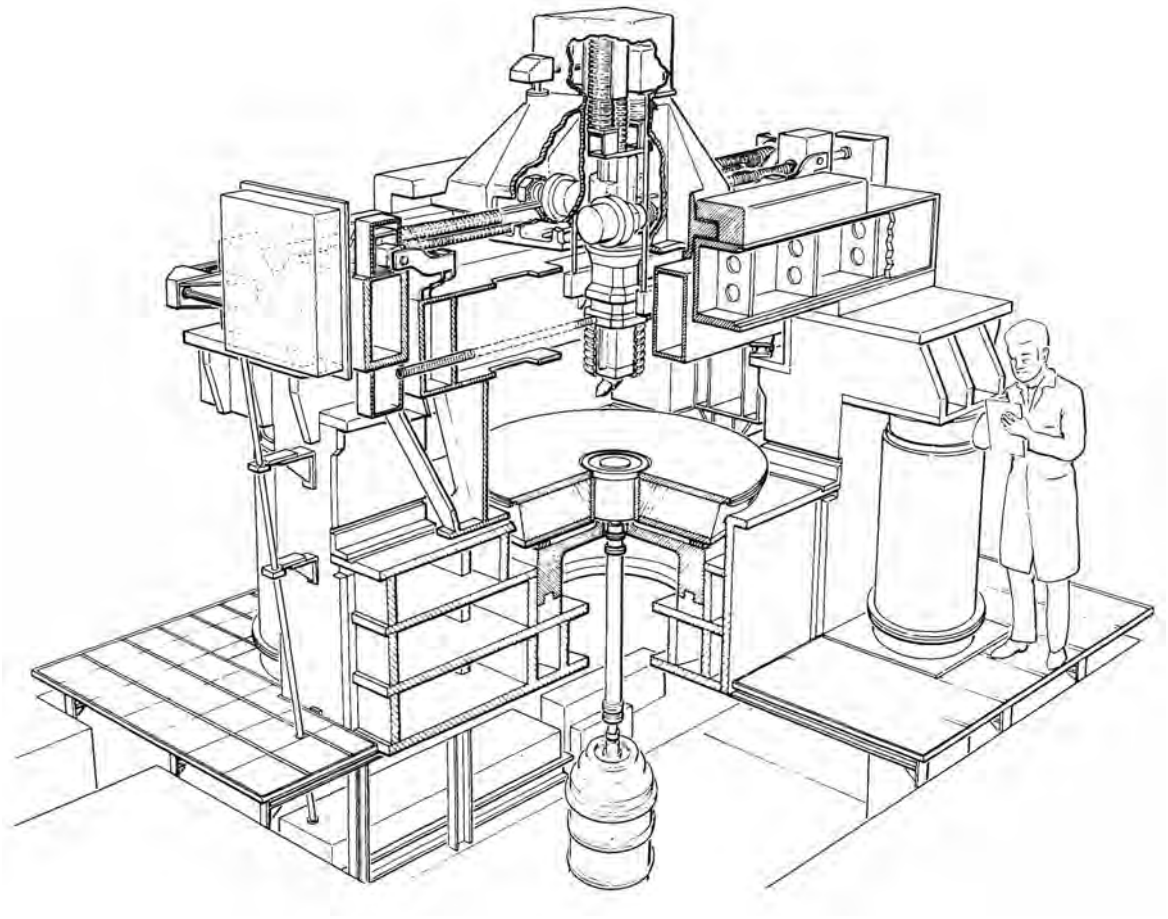


Figure 5.7.1 Large Optics Diamond Turning Machine (LODTM) designed and built by Lawrence Livermore National Laboratory. (Courtesy of LLNL.)

Because the toolbar is attached to the carriage, it is not feasible to measure the Z displacement of the tool tip directly with respect to the metrology frame. Instead, the Z displacement of the tool tip with respect to a reference plane on the carriage is measured and combined with Z direction straightness measurements of the carriage to yield the position of the tool tip in the Z direction with respect to the metrology frame. The Z direction straightness of the carriage is measured by the two outer interferometers on the top optics box. The optics box is mounted to a mini-metrology frame that is made from Super Invar and attached kinematically to the carriage with a vee, tetrahedron, flat kinematic coupling. The interferometers measure the differential motion of the top optics box with respect to the carriage's two straightedges which are mounted to the main metrology frame. By using two straightedges placed equidistant from the centerline of the carriage, roll errors, which could effect the Z direction straightness measurement of the carriage, can be eliminated by subtracting the average of the two outer laser readings from the measurement made through the center of the toolbar. The straightedges are optically finished so they function as long plane mirrors. Recall from Section 4.5 that the accuracy of a straightness interferometer is on the order of 0.3-0.5 μm (12-20 μin.). LODTM's straightedges are themselves accurate to about 150 nm and greater accuracy is achieved after installation by measuring artifacts placed in the work zone to generate error maps.

The entire machine is housed in a special environmentally controlled room with a high-velocity air shower temperature controlled to 0.01 F^o when operated remotely. In order to minimize refractive index errors caused by turbulence, temperature changes, pressure changes, humidity changes, and tramp gases, evacuated steel bellows with optical windows at their ends are used to encase the interferometer beam paths that measure large motions. Thus the distance that the beam has to travel in air can be kept very small, on the order of 0.8 mm (0.03 in.). Rigid tubes are used where

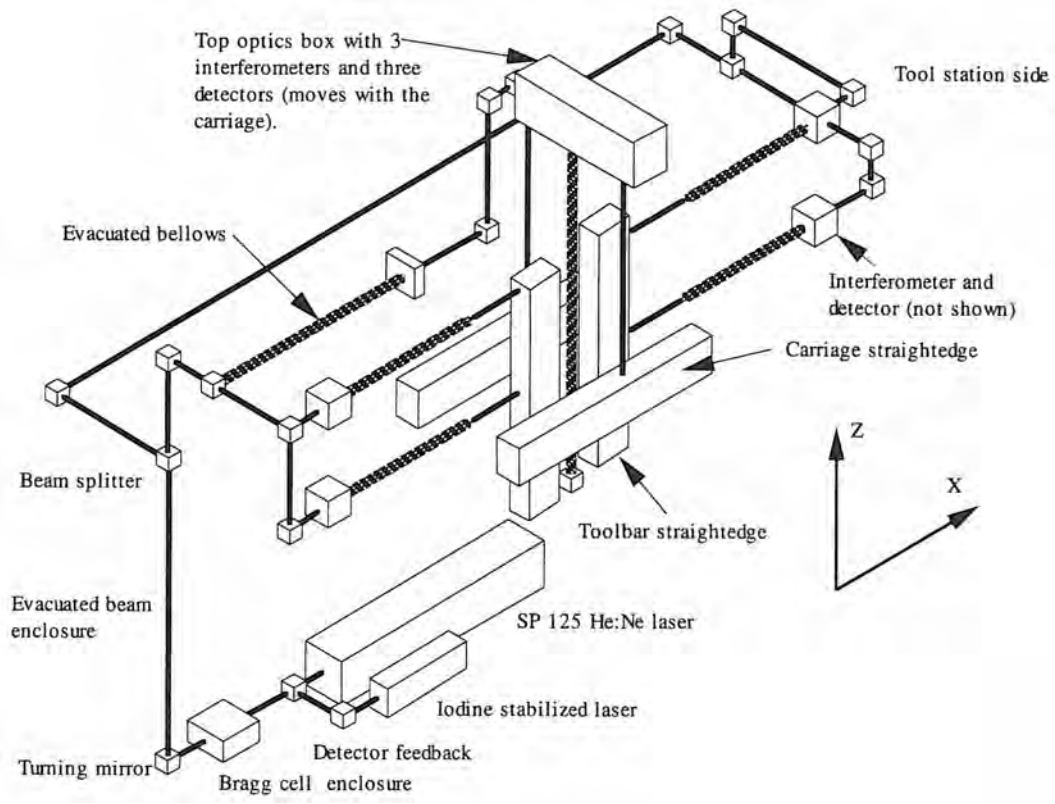


Figure 5.7.2 Components of LODTM's interferometric measurement system. (Courtesy of LLNL.)

only small motions are measured. The use of a bellows, as opposed to a tube with a sliding seal, eliminates the need for a sliding seal and the associated stick-slip characteristics which would affect the dynamic performance of the axes' servosystems. There are two types of bellows used, those for enclosing the interferometers and those for force counterbalance. To minimize heat generation in the axes' drive motors, their average motor current is used as an error signal to drive a variable-vacuum control loop for the counterbalance bellows. Two large bellows support the 2200 N weight of the toolbar and the controller adjusts for the spring rate of these bellows and the measurement bellows. For the carriage, the evacuated beampath bellows provides a constant bias force and an opposed bellows with controlled vacuum controls the bias and spring rate effects.

Sources of error in the Michelson-type heterodyne interferometric displacement measurement system include thermally induced changes in the index of refraction of the optical components, and polarization mixing of the measuring and reference beams. Polarization mixing errors are minimized through the use of very high quality optical components. Keeping the interferometer's measurement and reference beams collinear as much as possible helps to minimize the effects of index or refraction errors. Where only the measurement beam travels (e.g., quarter-wave plates and vacuum windows), index of refraction errors are minimized through the use of two measurement systems such as the two outer interferometers used for detecting X direction straightness of the Z axis and the two sets of two interferometers used to measure the pitch and Z position of the X axis. The difference between pairs of measurements helps to cancel out the thermally induced index of refraction errors in the optics to the extent that the temperature of the optics are equal.

The effect of retroreflective leakage of the measurement beam back into the laser is minimized by using the system shown in Figure 5.7.3. The large 15 mW He:Ne laser is stabilized with respect to a smaller iodine-stabilized laser. The stabilized single frequency output passes through an arrangement of optics similar to that used for a Mach-Zehnder interferometer where the beam's two orthogonally polarized components are separated. Each component passes through an acousto-optic frequency shifter, which increases the frequencies of the two beams by 60 and 61.75 MHz,

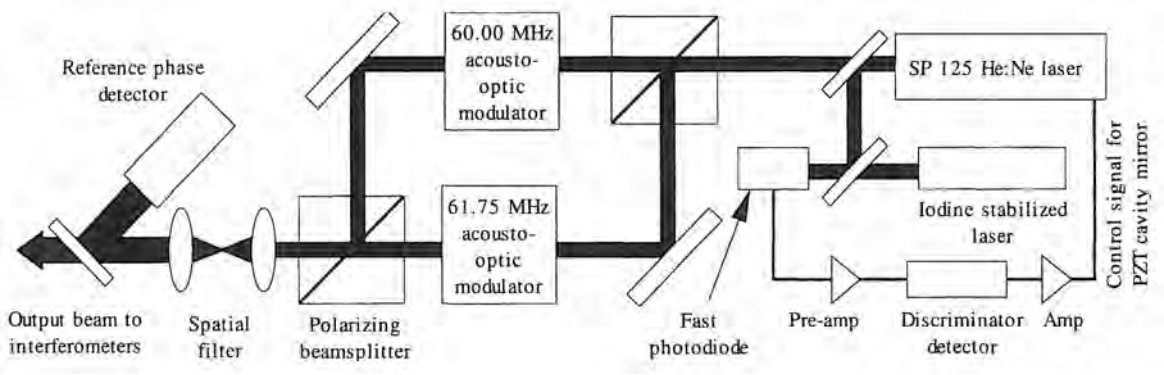


Figure 5.7.3 LODTM interferometer laser source. (Courtesy of LLNL.)

respectively. The resulting 1.75 MHz frequency difference is used for heterodyne detection of phase change as discussed in Section 4.5.7. Any retroreflected portion of the beam will thus enter the laser cavity at a frequency of minimum gain, so pollution of the laser cavity from exterior light sources is minimized.

Even with the use of the metrology frame, however, great care must still go into the design of the structural system in order to ensure stability and controllability of the servoed axes to the desired level of mechanical resolution.

5.7.2 Metrology Frame Design Concept for a T-Base Lathe[21]

Consider the T-base two-axis lathe shown schematically in Figure 5.7.4. The tool is mounted on the carriage whose radial position in the horizontal plane with respect to the spindle axis is numerically controlled. The axial position between the tool and the spindle is numerically controlled by motion of the spindle. The two axes intersect to form a "T" shape. A nice feature about this design is that the axes are not stacked on top of each other so the geometry and associated nesting of mechanical components is greatly simplified. Axes used to define the spindle system are X_{1S} Y_{1S} Z_{1S} and X_{2S} Y_{2S} Z_{2S}. The former is located at the centroid of the spindle's linear bearings and the latter is located at the spindle faceplate. Axes used to define the carriage are X_{1C} Y_{1C} Z_{1C} and X_{2C} Y_{2C} Z_{2C}. The former are located at the carriage bearing's centroid and the latter is located at the toolpoint.

Figure 5.7.5 shows the error gain matrix for the tool. One can see that the X and Z axes errors are most sensitive to Abbe effects. Unfortunately, they are both sensitive directions. Figure 5.7.6 shows the total errors for two positions of the carriage and spindle. Note that all the errors are assumed to be random at this early design stage, so when the errors are summed, the absolute values are taken first. The errors are an order of magnitude smaller than would ever be required of a typical machining operation (e.g., for an automobile engine); however, for precision optics and computer hard disks, they are still an order of magnitude too large.

There are three possible methods for increasing the accuracy of the lathe: (1) use specialized hand-finishing processes; (2) map the performance of the lathe and use software-based error correction algorithms; or (3) build a metrology frame around the lathe and measure and compensate for the errors in real time. The first method would be expensive and the machine would have to be periodically refinished unless frictionless aerostatic or hydrostatic bearings were used throughout. If fluidstatic bearings were not used, in the period between rebuilds, inferior parts might be produced. The second method may be more economical but would still require periodic remapping of the machine unless frictionless aerostatic or hydrostatic bearings were used throughout. The third method would provide the greatest increase in accuracy but also requires the largest initial capital investment. Note that the metrology frame may itself also require periodic recalibration to compensate for long term creep of the structural materials, and even with the use of a metrology frame, an initial mapping of the system may be required.

[21] Ibid.

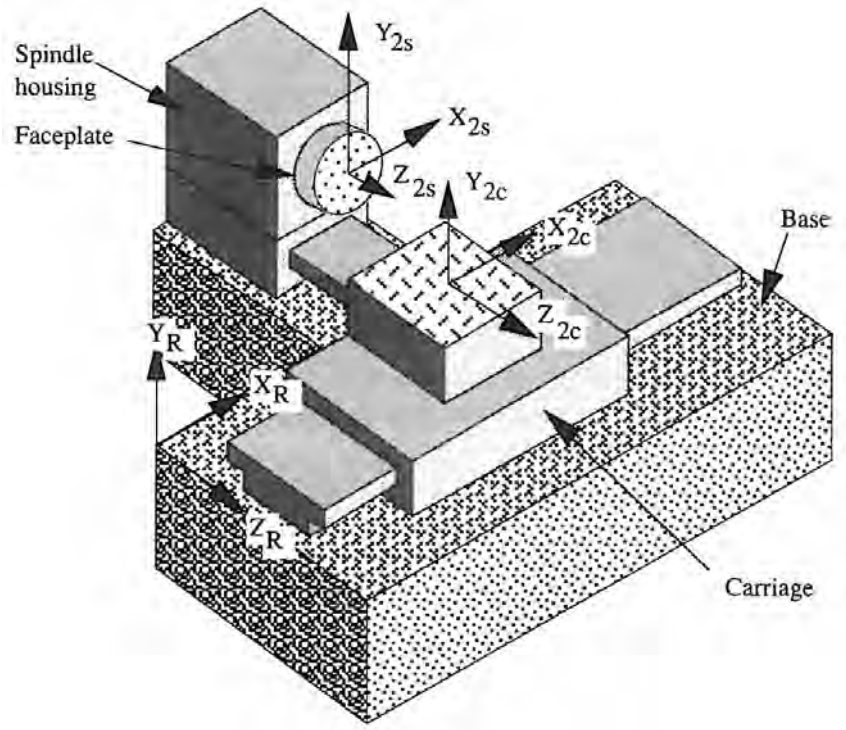


Figure 5.7.4 Illustrative example of T-base lathe.

In order to attain submicron accuracies, it appears that one of the latter two options is most desirable. Mapping techniques are discussed in detail in Chapter 6, so assume that a metrology frame is to be designed for this lathe. How does one go about designing a metrology frame? The first step is to identify the error motions that must be measured. These can be seen from the error gain matrix. The degrees of freedom that need to be measured for the spindle are: Z position, δ_x straightness, ε_x, and ε_y. The ε_z error of the spindle carriage does not have to be measured if the spindle carriage δ_x error is measured at an elevation equal to that of the center of the faceplate. For the carriage: X position, δ_z straightness, ε_x, ε_y, and ε_z all need to be measured. Note that with a crossed axis design, as was used for LODTM, some of these degrees of freedom could be measured simultaneously. The T-base design, on the other hand, generally requires a separate measurement system for the spindle and carriage assemblies.

In deciding how to make the required measurements, one must consider the sources of the errors in greater detail. Beginning with the spindle assembly, consider that the goal is to determine

Axis 1 error gains						
	DeltaX	**DeltaY**	**DeltaZ**	**EpsilonX**	**EpsilonY**	**EpsilonZ**
DelX(1)	-1.0	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
DelY(1)	0.00E+00	-1.0	0.00E+00	0.00E+00	0.00E+00	0.00E+00
DelZ(1)	0.00E+00	0.00E+00	-1.0	0.00E+00	0.00E+00	0.00E+00
EpsX(1)	0.14E-12	-0.25	-0.40	-1.0	0.00E+00	0.00E+00
EpsY(1)	0.25	0.690E-13	0.12E-04	0.00E+00	-1.0	0.00E+00
EpsZ(1)	0.40	-0.20E-04	0.00E+00	0.00E+00	0.00E+00	-1.0
Axis 2 error gains						
	DeltaX	**DeltaY**	**DeltaZ**	**EpsilonX**	**EpsilonY**	**EpsilonZ**
DelX(2)	-1.0	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
DelY(2)	0.00E+00	-1.0	0.00E+00	0.00E+00	0.00E+00	0.00E+00
DelZ(2)	0.00E+00	0.00E+00	-1.0	0.00E+00	0.00E+00	0.00E+00
EpsX(2)	0.14E-12	0.69E-13	0.00E+00	-1.0	0.00E+00	0.00E+00
EpsY(2)	0.00E+00	0.69E-13	0.00E+00	0.00E+00	-1.0	0.00E+0-0
EpsZ(2)	0.00E+00	0.69E-13	0.00E+00	0.00E+00	0.00E+00	-1.0

Figure 5.7.5 Toolpoint error gain matrix.

For axes positions:

Spindle frame l:	0.75, 0.00, -0.25	Spindle frame 2:	0.00, 0.40 0.25
Carriage frame 1:	0.75, 0.00, 0.25	Carriage frame 2:	0.00, 0.40, -0.25

Sum of the axes summed random errors (meters, radians)

DeltaX	DeltaY	DeltaZ	EpsilonX	EpsilonY	EpsilonZ
0.701E-05	0.350E-05	0.451E-05	1.00E-04	0.100E-04	0.100E-04

RMS of the axes RMSed random errors (meters, radians)

DeltaX	DeltaY	DeltaZ	EpsilonX	EpsilonY	EpsilonZ
0.337E-05	0.190E-05	0.287E-05	0.707E-05	0.707E-05	0.707E-05

Average of random sum and rms errors (meters, radians)

DeltaX	DeltaY	DeltaZ	EpsilonX	EpsilonY	EpsilonZ
0.519E-05	0.270E-05	0.3690E-05	0.854E-05	0.854E-05	0.854E-05

For axes positions:

Spindle frame 1:	0.75, 0.00, -0.50	Spindle frame 2:	-0.25, 0.40, 0.50
Carriage frame 1:	0.50, 0.00, 0.25	Carriage frame 2:	0.00, 0.40, -0.25

Sum of the axes summed random errors (meters, radians)

DeltaX	DeltaY	DeltaZ	EpsilonX	EpsilonY	EpsilonZ
0.826E-05	0.600E-05	0.576E-05	0.100E-04	0.100E-04	0.100E04

RMS of the axes RMSed random errors (meters, radians)

DeltaX	DeltaY	DeltaZ	EpsilonX	EpsilonY	EpsilonZ
0.401E-05	0.314E-05	0.313E-05	0.707E-05	0.707E-05	0.707E-05

Average of random sum and rms errors (meters, radians)

DeltaX	DeltaY	DeltaZ	EpsilonX	EpsilonY	EpsilonZ
0.614E-05	0.457E-05	0.445E-05	0.854E-05	0.854E-05	0.854E-05

Figure 5.7.6 Error combinations for the axes: What the tool must do to be at the proper place on the work.

where the *part* is. The part is attached to the rotating faceplate, so the first thought should be: *The faceplate rotates so it is bound to have many error sources, so what are they and how can position and orientation of the faceplate be measured?* In order to answer this question, place an imaginary part on the spindle and see what happens.

When a part is placed on the faceplate the spindle is displaced in the negative Y direction and rolls forward about the X axis. The former error occurs in the nonsensitive direction, while the latter causes an Abbe error in the Z direction (along the axis of the part). If an aerostatic bearing spindle is used, then variations in spindle speed will not cause an appreciable change in the radial position of the faceplate; on the other hand, if a hydrostatic bearing spindle is used, then the faceplate's radial position will vary with the spindle speed. Thus faceplate motion in the Z direction and X direction straightness of the spindle assembly must both be measured. For this design example, it is assumed that a general-purpose crash-resistant lathe is required; thus hydrostatic bearings are preferred over aerostatic bearings and effects of varying hydrodynamic lift, causing spindle displacements in the X direction need to be considered. If they are measured on a plane on the same elevation as the center of the faceplate, then they will pick up Abbe errors caused by pitch (ε_z). Roll errors result from a difference between the front and rear spindle Y direction bearing displacements. These displacements vary due to the different loads they see and a varying bearing stiffness caused by hydrodynamic lift that varies with spindle speed. Hence the Z position, δ_x straightness, and ε_x and ε_y angular errors of the faceplate need to be measured. There are many metrology frame designs that could be developed to accomplish this task. One such system and the logic behind its development will be presented here.

The faceplate moves along the Z axis with a range of $1/4$ m and it rotates about the Z axis. Thus any sensors chosen to measure the error motions must do so in the presence of these large motions. Reviewing the types of sensors available for measuring motion with submicron accuracy, there are three possible choices: laser interferometers for measuring linear and small angular displacement, autocollimators for measuring small angular displacement, and capacitance and differential impedance probes for measuring small linear displacements.

In choosing sensors and a mounting configuration, the first thought that comes to mind is that three points determine a plane which could define the Z position and ε_y and ε_x of the back of the

faceplate. Thus if the Z position of three points of the back of the faceplate can be measured, then Z, ε_y, and ε_x for the faceplate can be uniquely determined. In order to measure these three positions that experience large linear displacements (while rotating) with respect to a fixed reference frame, three plane mirror or differential plane mirror interferometers (DPMIs) can be used. In order to use the interferometers to measure the position of three points on the back of the *rotating* faceplate, the back of the faceplate could be polished to optical quality to function as a plane mirror. This design is shown conceptually in Figure 5.7.7.

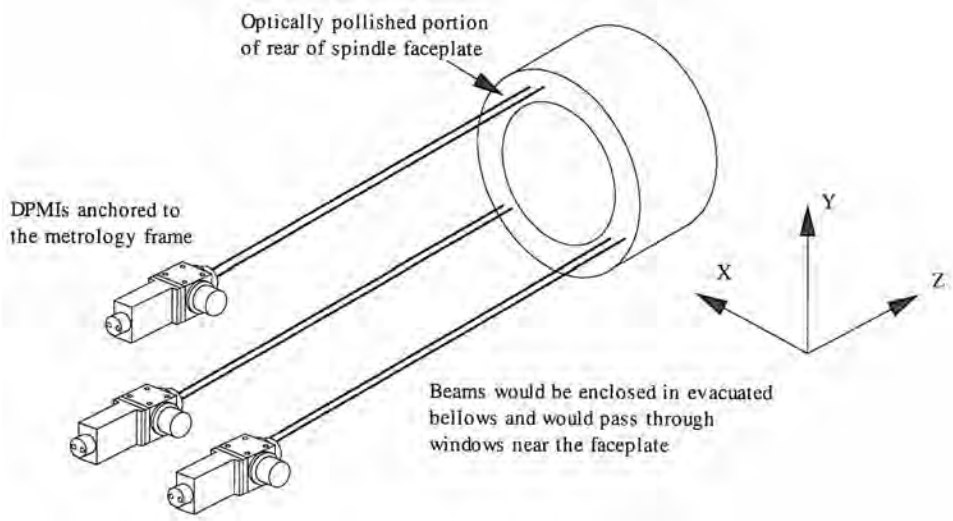


Figure 5.7.7 System used to measure axial position and yaw and pitch of a spindle faceplate.

Since the faceplate is a structural element that is subject to the weight of the part and thermal loads from the spindle bearings, at first glance it does not seem desirable to use its back as a reference surface from which measurements are to be made. On the other hand, the measurements would be made close to the position of the object that directly affects the accuracy of the part. Hence all that needs to be done is make the faceplate massive enough to ensure that the back surface remains flat to the tolerance that is desired. Back-of-the-envelope calculations necessary to determine the feasibility of this idea are described below.

The spindle faceplate is loaded by the weight of a part held to its surface by vacuum or some other means (e.g., bolts). Gravity pulls down on the center of mass of the part, which is cantilevered out from the faceplate, and a varying (probably linear) force distribution is generated with the top of the faceplate being pulled forward toward the carriage and the bottom being pushed back toward the spindle. To facilitate a quick back-of-the-envelope feasibility calculation for using the faceplate as the measuring reference surface, assume the following:

The moment produced by the part is equivalent to a force couple applied at the top and bottom of the faceplate.

The top half of the faceplate behaves like a semicircular plate of radius R that is cantilevered from a wall with a force F from the couple applied at the tip.

With these assumptions and the use of energy methods, an upper bound estimate can be found for the deflection of the tip of the faceplate. The result will be an upper bound because a concentrated load gives a greater deflection than a distributed load of equal total magnitude. Furthermore, the faceplate will in actuality be attached to the spindle shaft of diameter on the order of half that of the faceplate, and the spindle shaft will greatly stiffen the faceplate. Thus if the results indicate that for a reasonable thickness of the faceplate that it can act as a reference measuring surface, the metrology frame design can proceed. Of course, a more detailed finite element study of the best thickness to use should be made, but this analysis will give the design engineer a starting point.

The upper bound model of the loading for the spindle faceplate is shown in Figure 5.7.8. The sectional moment of inertia for the disk can be expressed as

$$I = \frac{bh^3}{12} = \frac{h^3 R \sin\theta}{6} \tag{5.7.1}$$

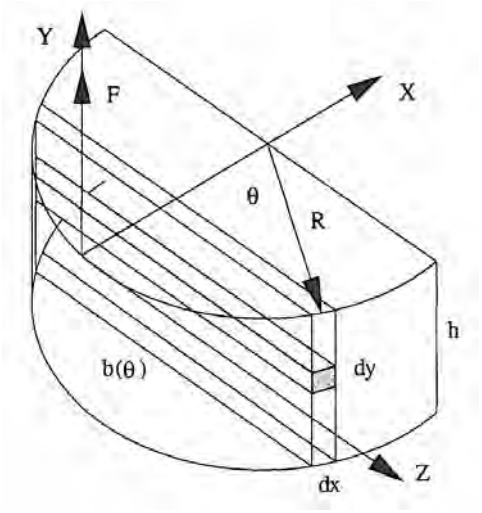


Figure 5.7.8 Upper bound model of loading condition for spindle faceplate.

The shear and bending deflection of the plate, based on the assumptions above, are found using energy methods. The respective strain energies in general form are

$$U_{bend} = \int \frac{M^2\,dx}{2EI} \tag{5.7.2}$$

$$U_{shear} = \int_V \frac{\tau^2\,dV}{2G} \tag{5.7.3}$$

The deflection is then given by

$$\delta = \frac{2U}{F} \tag{5.7.4}$$

For the half disk, the bending strain energy is

$$U_{bend} = \frac{3F^2R^2}{EH^3}\int_0^{\pi/2}(1-\cos\theta)^2\,d\theta = \frac{3(3\pi-8)F^2R^2}{4Eh^3} \tag{5.7.5}$$

The shear stress in a cantilevered beam with rectangular cross section is

$$\tau = \frac{F}{2I}\left(\frac{h^2}{4}-y^2\right) \tag{5.7.6}$$

The shear strain energy is thus

$$U_{shear} = \frac{F^2}{8G}\int_V \frac{1}{I^2}\left(\frac{h^2}{4}-y^2\right)^2 dV = \frac{3\pi F^2}{20Gh} \tag{5.7.7}$$

The total deflection is thus

$$\delta = \frac{3F}{Eh}\left(\frac{(3\pi-8)R^2}{2h^2}+\frac{\pi(1+\eta)}{5}\right) \tag{5.7.8}$$

This expression conservatively assumes that the part held onto the faceplate has no rigidity.

Table 5.7.1 summarizes the results of this analysis made with the assumption that the faceplate is 400 mm (15.75 in.) in diameter and the part is of equal diameter but may be from 20 to 200 mm thick. If a faceplate thickness is chosen that seems reasonable, on the order of 1/4 of the diameter, then the deflection with a 50 mm thick steel part that weighs about 500 N (112 lbf) is 0.0164 μm (0.64 μin.). With the largest possible part,[22] which is 200 mm thick, the deflection is 0.263

[22] Note that the deflection is a function of the square of the part thickness because the force couple is a function of the product of the weight and the thickness of the part.

μm (10.35 μin.). These results are most likely very conservative (and hence good for preliminary design feasibility studies), especially when you consider the fact that a 400 mm diameter, 200 mm thick part is itself going to be fairly rigid. In addition, one should note that the manner in which the faceplate is held to the spindle and the mass distribution of the faceplate and parts bolted to it will have an effect on the faceplate shape as the spindle speed increases.

Faceplate thickness	δ500N	δ1000N	δ1500N	δ2000N
(mm)	(μm)	(μm)	(μm)	(μm)
20	1.615	6.461	14.538	25.845
40	0.209	0.835	l.879	3.340
60	0.065	0.261	0.587	1.044
80	0.030	0.118	0.266	0.472
100	0.016	0.066	0.148	0.263
120	0.010	0.042	0.094	0.167
140	0.007	0.029	0.065	0.116
160	0.005	0.022	0.049	0.087
180	0.004	0.017	0.038	0.068
200	0.003	0.014	0.031	0.055

Table 5.7.1 First-order approximate faceplate deflections when subject to loading by steel parts 400 mm in diameter that are 50, 100, 150, and 200 mm thick (rigidity of part ignored).

The next calculation considers the angular rigid-body motion of the faceplate. This will determine if the plane mirror surface of the rear of the faceplate tilts so far as to cause the measurement beam to become critically misaligned. First the contribution to the angular misalignment from the worst-case deflection is found to be 0.263 μm/0.200 m = 1.32 μrad. Next, assume that the spindle bearings are 400 mm apart and the distance from the front bearing to the center of mass of the 200 mm thick steel part (weight 2000 N) is 400 mm, then the force components in the front and rear bearings due to the weight of the part will be on the order of 4000 N (900 lbf) and 2000 N (450 lbf), respectively. A reasonable assumption for the spindle's radial stiffness is 4×10^8 N/m (2.3×10^6 lb/in.). Thus the angular deflection component from the deflection of the spindle bearings is about 37.5 μrad. Assuming that the bend in the spindle shaft causes an angular deformation on the order of that caused by the spindle bearings, the total pitch of the back of the faceplate will be about 75 μrad. Amplified over the distance the light must travel from the plane mirror surface to the DPMI, which on the order of 1 m gives plenty of room to insulate the interferometers from the spindle motor, the lateral displacement of the measuring beam is about 0.075 mm. The cosine error resulting from this misalignment will be 28 Å (0.11 μin.).

This first-order upper bound analysis has shown that it is structurally feasible to use the back surface of the spindle faceplate as a target mirror for a set of three differential plane mirror interferometers. The reference surface would have to be kept clean with an air jet and/or wiper system. Note that if an air bearing spindle was used, a cleaning system would not be needed. The accuracy to which the angular measurement can be determined will depend on the accuracy of the knowledge of the spacing of the three DPMIs.[23] The spacing of the DPMIs can be determined by direct measurement of the location of the center of intensity of the light beams. Assuming that the laser interferometric system has 0.013 μm (0.5 μin.) resolution, then the resolution of the faceplate angular measurements can be on the order of 0.087 μrad. The other measurement system alternative would be to use one laser interferometer and two autocollimators. However, laser interferometers have much greater bandwidths than autocollimators and this arrangement should provide more than adequate resolution.

The next step is to determine a way to measure the radial motion of the faceplate in the Z direction. Since the faceplate is turning, it must be the measurement target; however, in order for a sensor to observe continually the same point (axially) on the faceplate, it must move with the spindle assembly in the Z direction. Hence X direction straightness errors of the spindle assembly must also be measured and added to the X radial motion of the faceplate in order to determine the X position of the faceplate with respect to the reference frame. The best way to accomplish this is

[23] Section 5.9 presents a detailed error analysis for a system with three sensors used to measure axial position and yaw and pitch.

to use a straightedge kinematically mounted to the spindle assembly such that temperature changes in the straightedge do not affect its straightness. Two sensors mounted to the metrology frame (e.g., capacitance probes) would measure the distance to the straightedge which would be used to determine the X direction straightness of the spindle assembly. Combined with a measurement of the X distance from the straightedge to the faceplate, the X motion of the faceplate with respect to the metrology frame could be determined. By using two straightedges, one on each side of the spindle, thermal effects could be minimized by differential measurements. The spindle housing is about 500 mm long and two sensors must be able to measure the straightedge's Z position with respect to the metrology frame. If the angular resolution is to be 0.1 μrad, and capacitance probes with 0.025 μm (1 μin.) accuracy are used, then they must be placed 500 mm apart. This requires the straightedge to be at least 750 mm long, which will require an overhang out the back of the spindle by 250 mm. Since the back of the spindle is easily protected, this type of arrangement would be acceptable. The other alternative would be to increase the length of the spindle.

The carriage serves as a place to mount cutting tools and as a place to mount parts that are to be flycut if a tool is held in the spindle. Thus the top surface of the carriage must remain free of obstructions including sensor reference surfaces. The error analysis indicated that in addition to the axial position of the carriage along the X axis, Z direction straightness and yaw, pitch, and roll motions must be measured. Laser interferometers and capacitance probes would also be good choices for measuring the motions of the carriage.

Beginning with the X motion of the carriage, if two X measurements are made in a horizontal plane, then ε_y can also be determined. Note that if a reference straightedge where mounted in the center of the carriage parallel to the Z axis and appropriate pathways allotted for, two sets of two measurements along the X axis could be made from either side of the carriage. In this manner, thermal effects on the index of refraction of the optics could be accounted for in the same manner as was done for the X axis of LODTM. Evacuated bellows still need to be used as was done on LODTM.

The Z direction straightness must be measured using a straightedge arrangement because a straightness interferometer cannot provide the requisite accuracy. Recall that on LODTM two straightedges were required to measure the X direction straightness and roll of the Z axis even though the roll angle caused an Abbe error in the nonsensitive direction at the tool tip. However, for the T-base lathe, an ε_x error could cause an Abbe error in the measurement of the Z straightness and thus it too has to be measured. In this case, an ε_y error can also cause Abbe errors in the measurement of the Z direction straightness, but since ε_y is measured by the X axis laser interferometers, it can be calculated. The ε_x angular error and Z direction straightness errors on the T-base's carriage can thus be measured using a similar method that was used with the Z axis on LODTM.

Two sets of two sensors mounted on the carriage could be used to measure the Z displacements of two points each on the straightedges with respect to the carriage. These measurements can be used to determine the Z direction straightness and roll and pitch of the carriage. Depending on how the straightedges are protected, a continuous oil shower, as shown in Figure 5.4.5, might be used to keep them clean. If the carriage is long with respect to the travel, then it may be advantageous to mount the straightedges on the carriage for two reasons: (1) so the surfaces of the straightedges face away from the cutting process, and (2) so the sensors' signal cables (if capacitance probes are used) do not have to flex. Once again, the straightedges must be kept at a uniform temperature to prevent them from bowing along their length. Thus careful analysis would have to be made to determine whether a low thermal expansion, low thermal conductivity material like Zerodur or a high thermal expansion, high thermal conductivity material like aluminum should be used. Whichever material is used, however, the straightedges must be attached kinematically to the carriage. Note that vertical deflection of the straightedges due to their own weight would not cause an error in the direction being measured unless the deflection in some way introduced a twist into the beam.

LODTM was designed to cut parts without an accompanying oil shower to help maintain constant temperature of the part (an air shower was used). For some precision machines, temperature control is achieved by showering the machine with temperature-controlled cutting oil.[24] For the measurement system developed for the T-base lathe, the interferometers and their target surfaces can be sealed with bellows, or in the case of the spindle, environmentally protected using a labyrinth

[24] See J. Bryan et al., "An Order of Magnitude Improvement in Thermal Stability with Use of Liquid Shower on a General Purpose Measuring Machine," SME Tech. Paper IQ82-936, June 1982.

seal and a positive internal pressure, temperature-controlled gas environment. The straightedges and capacitance probes, on the other hand, might require a complex sliding seal whose friction properties would affect the dynamic performance of the axes. An alternative to trying to keep the gap between the capacitance probes and the straightedges clean is to bath the gap continually with a temperature-controlled stream of oil, as was shown in Figure 5.4.5. Note that an air wipe might not be appropriate because the expanding air could cool the probe and straightedge which could cause thermal errors.

Design of the the metrology frame structure itself must consider issues such as stability, dynamic stiffness (resistance to vibrations from the subgrade), and thermal growth. These and other structural design considerations are discussed in Chapter 7. Assuming that proper environmental control is maintained and the metrology frame's structure is properly designed, use of the metrology frame could allow for the machine to achieve a volumetric accuracy on the order of 0.025 μm (1 μin.). This is two orders of magnitude better than the accuracy (computed) for the machine without a metrology frame.

5.7.3 Goniometers

Metrology frames are stationary structures that cannot easily be used to measure the endpoint position of devices such as articulated robots or five-axis milling machines. A *goniometer*, on the other hand, ideally consists of nondeforming (kinematically mounted) measuring beams that are mounted to moving structural members. The relative position of measuring beams attached to an articulated structure can be measured by clusters of sensors attached to the ends of the measuring beams. Goniometers have been used successfully to measure motion of human limbs to study gait for the design of artificial limbs; however, goniometers have not been widely used in the industrial sector as devices for measuring the position of robots and machine tools because of the complexity associated with making all the required measurements of the relative position of each measuring beam. For example, goniometers for robots have been designed,[25] but often the cost and complexity is greater than would be required to redefine the task and the mechanical design so the system specifications can be met with simpler systems.

5.8 SENSOR CALIBRATION

The reader should note that in many instances specific standards for calibration and use of various sensors and measuring machines have been developed by the American National Standards Institute (ANSI) and similar institutions abroad. These standards describe how objects should be measured, and how to utilize the information to characterize part geometry. For example, to describe the roundness of a hole, one method is to measure the contour of the hole and calculate a mean radius for all the points. Another method would be to define the size of the hole as bounded by the largest circle that contains all the measured points. Variations on this example exist for many other types of geometries, and the reader is cautioned to check the standards for different countries when writing specifications for machine performance for a machine that may be sold internationally.

It is important for the machine design engineer to appreciate how sensors can be calibrated so that manufacturers' claims can be better assessed. After all, if the machine does not work because the sensor is no good, the machine design engineer, not the sensor manufacturer, will have to bear the brunt of the blame. Furthermore, often a precision machine design engineer is called upon to design calibration equipment. Before small, fast laboratory computers became widely available, sensors were usually calibrated by manually reading the measurement of the sensor's response to a known input such as the thickness of a gage block. This process was slow and tedious, and as a result, the output of a sensor was most often fit to a straight line. Hence product literature usually defines accuracy as the linearity of the sensor's response. With the advent of small, fast computers, a whole new world has opened up for the dimensional metrologist and the design engineer of his equipment.

[25] See, for example, A. Slocum, "Development of a Six Degree-of-Freedom Position and Orientation Sensing Device: Design Theory and Testing," Int. J. Mach. Tool Des., Vol. 28, No. 4, 1988, pp. 325–340. Also see U.S. patents 4,606,696 and 4,676,002.

Accuracy is highly dependent on the care taken in performing the calibration.[26] The primary goal in a calibration experiment is to measure the response of a sensor to a known quantity while holding all other variables constant (e.g., temperature, supply voltage). Ideally, all sensors should be calibrated with respect to a traceable National Standard (e.g., the wavelength of light from a stabilized He:Ne laser) while mounted in the fixture in which they are to be used. When this is not plausible, care must be taken to duplicate the operating and mounting conditions for the sensor.

As accuracy requirements are pushed further and further, it often becomes desirable to determine how the sensor itself is affected by other variables (e.g., temperature). With the aid of a small, fast, computerized data acquisition system and a controlled environment, this can be accomplished by repeating the calibration measurements over their full range each time the other variable(s) is incremented. The effect is to create a series of nested loops, with the innermost loop being the response of the sensor over its intended range of motion. This allows for the multidimensional mapping of the sensor's response to a primary input (e.g., position) while simultaneously being subjected to varying secondary variables (e.g., temperature, supply voltage, frequency, etc.).

Regardless of the method used to calibrate a sensor, the certainty of the calibration can be obtained from a careful evaluation of the calibration system's error budget. In general, there are five dominant error categories to be considered:

- R_I the repeatability of the sensor.
- R_C the repeatability of the calibrator.
- R_L the repeatability of the link between the instrument and the calibrator.
- S_L the systematic uncertainty of the link.
- S_C the systematic uncertainty of the calibrator.

The repeatability of the sensor is defined as the standard deviation from the mean of the sensor's response map. The repeatability of the calibration device is essentially a function of how stable the device is. The repeatability and systematic uncertainty of the link between the sensor and the calibrator are obtained by measurement or estimation from the error budget for the calibration system. The error budget must include errors caused by such as temperature changes and signal digitization truncation (least significant bit errors). The systematic uncertainty of the calibrator is a function of how well it relates to the recognized national standard, and is usually supplied with the calibrator.

Designing the Calibration Experiment

When designing a calibration system for static or dynamic calibrations, the same error budgeting techniques need to be applied to the calibration system design as are used for the design of the machine the sensor is used in. The ideal goal is to make the calibration system 5-10 times more accurate than the application requires. In many cases it will not be possible to calibrate the sensor in situ. In such cases the sensor should be physically mounted in the same manner in which it will be used. It is often easier to design the calibration experiment when it is not done in situ because one does not have to contend with other structures on the machine. For example, it is usually easier to set up an interferometer that is in line with the sensor on a calibration stand than in the machine, hence allowing for almost complete elimination of Abbe errors in the calibration.

Not everyone has access to a laser interferometer, but often a commercially available computer-controlled stage with servo accuracy greater than is required for the calibration of the sensor will be available. One must be careful when reading manufacturers' claims for component accuracy and how they define accuracy. For example, linear positioning accuracy at the pitch center of a stage may be 1 μm, but if the mounting surface of the stage is a centimeter above the pitch center and the pitch error of the stage is 10 μrad, then a 0.1 μm Abbe error results. The engineer should always ask: *Does the manufacturer base his accuracy claims on the supposed accuracy of the components used in the system, or does the manufacturer measure each stage with a laser interferometer to certify the accuracy of each system?* Not many stages are mass produced with greater than 1μm repeatability at a reasonable distance (within a few centimeters) from the top surface of the stage. Thus for critical application, if a laser interferometer is not available, the user should arrange for an independent lab to verify the performance of a system.

[26] For an enlightening discussion of calibration issues, the reader is referred to a 50-page book Repeatability and Accuracy: An Introduction to the Subject, and a Proposed Standard Procedure for Measuring the Accuracy of Industrial Measuring Instruments, by A. T. J. Hayward, Mechanical Engineering Publications, New York, 1977.

In summary, whether performing a static or dynamic calibration, the mechanical design of the calibration system should at least be representative of the manner in which the sensor is held in the machine it will be used. The properties of the structure around the sensor (e.g., stiffness and mass) should be as representative of the operating system as possible. In addition, if a target is needed, it should be of the same material and shape as the target in the intended application. An error budget for the system must also be made that includes static and dynamic mechanical, electrical, and environmental factors. This will enable the user to make an estimate of the certainty of the calibration.

Stability Determination

In order to determine the stability of a sensor's output, the sensor must be mounted in a stable environment where the sensor is measuring a fixed quantity (e.g., distance). Typically, for probe-type sensors, a cap is made from the same material as the probe body, which fits over the end of the sensor and is secured with a light squeeze type of collar, or by its own weight if the sensor is oriented vertically.[27] The system is allowed to sit and measure the distance from the sensor to the cap for an extended period of time. Any drift, whether it be inherent in the electronics or the sensor itself, is characteristic of the stability of the sensing system, and thus will be measured.

Static Response Calibration

Determining the static response of a sensor is the most often performed calibration. A static calibration is one where the measurand is incremented and the sensor's response is not measured until the system, which often includes mechanical and electrical components, has stabilized. Typically, this means waiting a few seconds before taking the readings. Once stabilized, numerous readings can be made and averaged to yield an accurate assessment of the sensor's response. With the widespread availability of personal computers and computer-controlled linear motion stages, it is becoming more commonplace to map the static performance of a sensor's response by taking numerous data points and fitting an nth-order polynomial to the data.[28] The computer can analyze the output from the sensor and only begin taking readings once it determines that the output has stabilized. Without a computer-controlled calibration system, calibration is a tedious, boring task where the risk of human error is great.

In order to increase mapping accuracy, a set of maps of the sensor response may be made and used to find an averaged set of coefficients. However, it is unlikely that a computer-controlled calibration stage can stop at the exact place each time. Still for each run a set of nth-order polynomial coefficients can be determined and a program can then be written to digitally increment the position variable of each polynomial. All the corresponding responses generated from the polynomials can then be averaged and used to generate an averaged polynomial for the sensor. An alternative would be to attain a common set of incremented points from initial multiple pass readings by interpolating between points.

Regardless of the type of position calibration performed, ideally a stabilized He:Ne laser interferometer should be used as the traceable incremental position measurement reference. Since the laser is only an incremental position measuring device, a home position sensing device is still required, such as an LVDT or a sensitive limit switch. If the axis of measurement of the laser interferometer is coincident with the axis of measurement of the sensor, then Abbe errors can be virtually eliminated and the stage is only required to move a fixed increment, stop, and allow the laser and the sensor to measure their position simultaneously. Using an interferometer and a stage with a finite amount of friction also allows the servo control system to be turned off when it is time to make a measurement. This eliminates possible limit cycling errors caused by discretization limits present in most digital servos.

Dynamic Response Calibration

Determination of the dynamic response of a sensor is difficult but can be accomplished using one of two following methods: The first method requires the use of a shaker table whose oscillation frequency and amplitude can be measured. White noise or a swept sine wave is used as the input

[27] This type of test is thus often known as a *cap test*. See Temperature and Humidity Environment for Dimensional Measurement, ANSI Standard B89.6.2-1973, p. 24, for details.

[28] See, for example, J. V. Moskaitis and D. S. Blomquist, "A Microprocessor Based Technique for Transducer Linearization," Precis. Eng., Vol. 5, No. 1, 1983, pp. 5–8.

command to the table and the sensor output is measured and stored. If the transfer function of the table is known, digital signal processing techniques can then be used to determine the sensor's transfer function.[29] The transfer function for the sensor obtained by this technique assumes that the response of the sensor is linear with respect to changing distance for a constant frequency; thus it must be combined with the static response map generated for the sensor to obtain a frequency-amplitude calibration for the sensor. The second method provides greater accuracy with commensurate increase in complexity of the calibration. A range of displacements and frequencies are input to the sensor to produce a map of the sensor's output as a function of these two variables. Note that if accuracy of this magnitude is required from a position sensor, it is often easier to use a laser interferometer or high-accuracy encoder as the sensor in the first place. The high cost of a laser interferometer can often offset the high cost of elaborate initial and periodic calibrations.

Sampling Rate[30]

When making discretized measurements of a dynamic system, the measurements must be made fast enough in order to mathematically reconstruct the response of the system. If the sampling rate is not fast enough, then a higher-frequency waveform (e.g., noise) may take on the identity of a lower-frequency component and cause a condition known as *aliasing*. The *minimal* sampling rate to avoid this problem, referred to as the *Nyquist* frequency, is equal to twice the frequency of the process being measured. Since there is always noise at all frequencies (i.e., white noise) in an analog system, before an analog signal from the sensor is used to reconstruct the response of the system or used in a digital servo, it often must be low-pass filtered with an analog or digital filter. Since all filters introduce their own dynamic characteristics into the system, the faster the sampling rate, the higher the filter's cutoff frequency can be with respect to the desired frequency of the system. The greater the ratio of the cutoff frequency to the desired system frequency response, the less the effect of edge characteristics of the filter on system performance. Thus although the Nyquist criteria says the sampling frequency should be at least twice the cutoff frequency of the filter, ideally it should be 5-10 times greater.

Environmental Effects on Calibration Accuracy

If a sensor is calibrated at standard temperature and pressure, 20°C and 760 mm Hg, and then used in a different environment, the calibration may be in error. It is vital that thermal effects be carefully considered when evaluating the calibration system error budget. It is equally vital that the person performing the calibration observes the assumptions made in the error budget and does not introduce new sources of error. The most common mistake made is for the person to pollute the calibration experiment thermally with their own body heat. Staring at an experiment at close range can subject the apparatus to temperature changes caused by radiation of body heat and a person's warm breath. Handling components and then not letting the temperature of the system come to thermal equilibrium[31] is another common error.

Other considerations, such as temperature and velocity gradients may also have to be considered. For example, if a capacitance probe is used to measure the error motions of a spindle, what is the effect of the density and temperature of the air as it is sheared between the probe tip and the spindle? On the micron level, these effects may be ignorable, but what about on the submicron level? Often the best way to handle this type of situation is to use an arrangement of probes whose outputs are combined differentially to minimize these errors. For example, when measuring the error motions of a spindle, the differences in the output from the sensors is a measure of the change in lateral position of the spindle. Any environmental factor, including expansion of the spindle due to thermal and internal effects, will affect both sensors equally and thus will mutually cancel each other out. Geometric symmetry can be used in many other ways to increase and check the accuracy of measurements. Other examples of the use of symmetry in measurement are described in Chapter 6.

Related to environmental effects is the cleanliness of the measurement surfaces. A stream of clean, dry, pressurized air can often be used to keep dirt from building up on the sensor and target surfaces, although then one has to worry about refrigeration effects. As discussed in the metrology frame example and shown in Figure 5.4.5, instead of an air wipe, in some cases a steady laminar flow of temperature-controlled oil can be used. In any case, care must be taken to ensure that the

[29] See, for example, Chapter 11 of A. Oppenheim and R. Schafer, Digital Signal Processing, Prentice Hall, Englewood Cliffs, NJ, 1975.

[30] See also Section 3.1.

[31] Allowing components to reach thermal equilibrium in their environment is called *letting the system soak*.

calibration is performed under the same conditions that the sensor will be used in. In many instances, particularly in a measurement laboratory or clean room, dirt is not a problem. On the other hand, in a machine tool environment dirt can affect sensor's performance in many ways. For example, dirt can decrease the intensity of a light spot and change the location of its true center of intensity. Dirt can also change the dielectric constant in the gap between a capacitance probe and a target. Dirt containing metallic particles can affect the output of an impedance probe. Dirt can get between the tip of a spring-loaded LVDT and the target, causing the thickness of the dirt particle to be measured also.

5.9 Effects of Sensor Output and Location Errors on Accuracy

In this section we investigate the effects of accuracy of a sensor and its location on the accuracy of a device that uses three sensors to determine the distance and pitch and roll angles between two nominally parallel plates. An example of an application for this type of device includes measuring the position and orientation of a plane with respect to a metrology frame (e.g., a spindle faceplate). A generic arrangement for this type of sensor system is shown in Figure 5.9.1.[32] The general system equations describing the degrees of freedom l_{XY} (distance between the planes at any point X, Y), α, and β (yaw and pitch) are formulated with the assumption that rotations of the target plane (X'Y' plane) occur about the X and Y axes, respectively, in the sensor coordinate plane as shown in Figure 5.9.1. The error associated with this nonEulerian selection of the angles is on the order of $\alpha\sin\beta$. With the assumption that α and β are at most on the order of 200 μrad, the angular error in αgr β due to this assumption will be about 0.04 μrad, which is acceptable. The relations for the system are

$$\ell_{XY} = \ell_3 + (b + Y)\sin\alpha - X\sin\beta \tag{5.9.1}$$

$$\alpha = \tan^{-1}\left(\frac{\ell_2 - \ell_3}{a + b}\right) \tag{5.9.2}$$

$$\beta = \tan^{-1}\left(\frac{\ell_1 - (\ell_2 b + \ell_3 a)/(a + b)}{c}\right) \tag{5.9.3}$$

Note that a, b, and c are dimensions as shown in Figure 5.9.1, they are not coordinates. If small-angle approximations are assumed, then the error in computing ℓ_{XY} over the expected range of α and β will be on the order of 4 parts per million. Thus small-angle approximations should only be used to evaluate the error caused by parameter variation. For the purposes of evaluating the sensitivity of ℓ_{XY} to variations in ℓ_i, a, b, and c, it can be assumed that $\tan^{-1}(\,) = (\,)$ and upon substituting Equations 5.9.3 and 5.9.2 into Equation 5.9.1:

$$\ell_{XY} = \frac{-X\ell_1}{c} + \frac{\ell_2}{a + b}\left(b + Y + \frac{Xb}{c}\right) + \ell_3\left(1 - \frac{b + Y - Xa/c}{a + b}\right) \tag{5.9.4}$$

Effect of Sensor Output Errors

To determine the error $\delta\ell_{XY\ell i}$ the calculation of the distance between plates due to an error $\delta\ell_i$ in sensor #i's output, the partial derivative $\partial\ell_{XY}/\partial\ell_i$ of Equation 5.9.4 is evaluated:

$$\delta\ell_{XY\ell 1} = -\frac{X}{c}\delta\ell_1 \tag{5.9.5}$$

$$\delta\ell_{XY\ell 2} = \frac{b + Y + Xb/c}{a + b}\delta\ell_2 \tag{5.9.6}$$

$$\delta\ell_{XY\ell 3} = \left(1 - \frac{b + Y - Xa/c}{a + b}\right)\delta\ell_3 \tag{5.9.7}$$

[32] Note that it is also possible to monitor the tilt and position of a plane with respect to another through the use of a Fizeau interferometer. See A. Gee et al., "Interferometric Monitoring of Spindle and Workpiece on an Ultraprecision Single-Point Diamond Facing Machine," SPIE Vol. 1015, Micromachining Optical Components and Precision Engineering,1988, pp. 74–80.

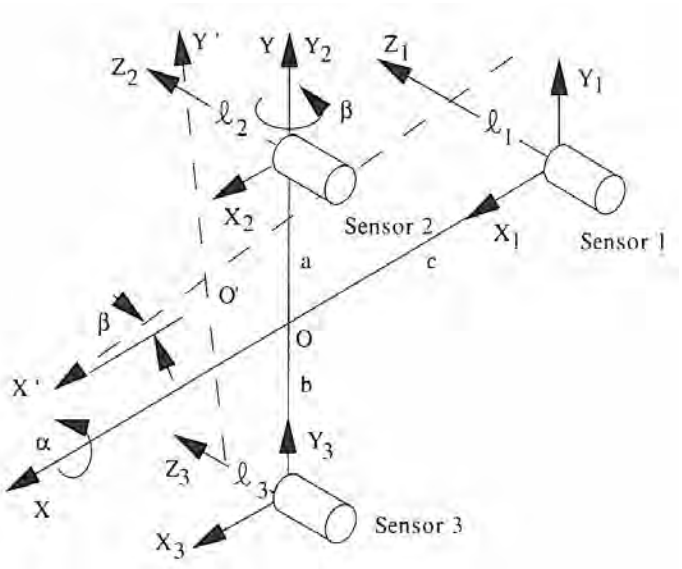


Figure 5.9.1 Triad of distance measuring sensors used to determine distance and yaw and pitch between two plates.

The angular errors $\delta\alpha_{\ell i}$ due to errors in the sensor reading $\delta\ell_i$ are determined in a similar manner:

$$\delta\alpha_{\ell 2} = \frac{1}{a+b}\delta\ell_2 \tag{5.9.8}$$

$$\delta\alpha_{\ell 3} = \frac{-1}{a+b}\delta\ell_3 \tag{5.9.9}$$

Similarly, the angular errors $\delta\beta_{\ell i}$ are

$$\delta\beta_{\ell 1} = \frac{1}{c}\delta\ell_1 \tag{5.9.10}$$

$$\delta\beta_{\ell 2} = \frac{-b}{c(a+b)}\delta\ell_2 \tag{5.9.11}$$

$$\delta\beta_{\ell 3} = \frac{-a}{c(a+b)}\delta\ell_3 \tag{5.9.12}$$

Effect of Errors in Probe Spacing

The effect of an error in the knowledge of the spacing between the sensors on the determination of the distance and angles between the two plates is found by taking the partial differential of l_{XY} with respect to a, b, and c:

$$\delta\ell_{XYa} = \frac{(b + Y + Xb/c)(\ell_3 - \ell_2)}{(a+b)^2}\delta a \tag{5.9.13}$$

$$\delta\ell_{XYb} = \frac{(Y - a - Xa/c)(\ell_3 - \ell_2)}{(a+b)^2}\delta b \tag{5.9.14}$$

$$\delta\ell_{XYc} = \frac{X}{c^2}\left(\ell_1 - \frac{\ell_2 b + \ell_3 a}{a+b}\right)\delta c \tag{5.9.15}$$

The effect of errors in orthogonality between the coordinate axes used to define the location of the sensors in the plane on the determination of l_{XY} can be equated to errors in a, b, and c.

Next determine the effect on the calculation of the angles α and β from errors in the sensor spacing. Proceeding as before, the following relations are found:

$$\delta\alpha_a = \frac{(\ell_3 - \ell_2)}{(a+b)^2}\delta a \tag{5.9.16}$$

$$\delta\alpha_b = \frac{(\ell_3 - \ell_2)}{(a+b)^2}\delta b \tag{5.9.17}$$

$$\delta\beta_a = \frac{(\ell_2 - \ell_3)b}{c(a+b)^2}\delta a \tag{5.9.18}$$

$$\delta\beta_b = \frac{(\ell_3 - \ell_2)a}{c(a+b)^2}\delta b \tag{5.9.19}$$

$$\delta\beta_c = \left(\frac{\ell_1 - (\ell_2 b + \ell_3 a)/(a+b)}{c^2}\right)\delta c \tag{5.9.20}$$

Effect of Probe Misalignment

The target plane's rotation is defined by rotations α and β about the X and Y axes, respectively. Ignoring cross coupling effects,[33] the equivalent errors in the distance measurement ℓ_i caused by mounting orthogonality errors ε_{xi} and ε_{yi} can be found from the law of sines and small angle approximations to be

$$\delta\ell_{i\varepsilon x} \approx \frac{\ell_i \varepsilon_{xi}(2\alpha - \varepsilon_{xi})}{2} \tag{5.9.21}$$

$$\delta\ell_{i\varepsilon y} \approx \frac{\ell_i \varepsilon_{Yi}(2\beta - \varepsilon_{Yi})}{2} \tag{5.9.22}$$

where ℓ_i is the actual (erred) distance measured by the sensor.

Even a relatively simple device can require extensive calculations in order to fully evaluate errors in the system. With Equations 5.9.4–5.9.22, the effect of any system perturbation on the calculated physical quantities ℓ_{XY}, α, and β can be determined and incorporated into the system error budget. Regardless of the measurement system design, a similar procedure can be followed and the resulting equations used to help formulate the error gain matrix for the system. Once the error gain matrix is found, specific contributors to the error (e.g., thermal and long-term stability effects) can be applied to each of the variables in the system (e.g., the dimensions a, b, c and the sensor readings ℓ_i). Note that some of the expressions for errors contain values of the sensor output, and in order to incorporate the results of the expressions into the error budget, worst-case values for the sensor output should be used.

5.10 Design Case Study: Design of a Laser Telemetric System[34]

A large number of manufactured components and raw materials are cylindrical in shape and are manufactured in continuous processes. The materials of these components include metal, rubber, plastic and glass. In addition to the various types of materials, the motion and temperature of the components during manufacturing can vary over a large range. The preferred method to measure soft, delicate, hot, or moving objects is with a noncontacting sensor such as capacitive, eddy-current, air, or optical type. Optical sensors have a number of advantages over the other types including: They can measure parts of any material, and they allow the distance from the sensor to the object being measured to be large. There are a number of various techniques used in optical dimensional gaging. These include shadow projection, diffraction phenomena, linear arrays, and scanning laser beams.

In 1972 two manufacturers asked Zygo to develop a noncontact optical sensor to measure extruded and rolled material for in process quality control. Sensor requirements were 50 mm (2 in.) measurement range, 0.005 mm (0.0002 in.) accuracy, and a 100 mm (4 in.) measurement throat which could tolerate a large lateral motion of the part in the throat without loss of accuracy. At

[33] For example, the angle β does not affect the error in length measurement associated with the orthogonality error ε_{xi}, whereas α does.

[34] The work described here benefited from the contributions of many people at Zygo Corp.. Besides James Soobitsky, major technical contributors were Frank C. Demarest, George C. Hunter, and Carl A. Zanoni. The text by James Soobitsky was edited by A. Slocum. This system is now sold by Z-Mike, a Division of Laser Mike, 430 Smith Street, Middletown, CT, 06455 (203) 635-2100.

that time, other manufacturers of scanning laser beam sensors did not offer an instrument with these capabilities. Also, other types of noncontacting sensors (e.g., capacitance probes) could not perform under these conditions. Zygo felt that the development of such a sensor was within its capabilities and that a sufficient market existed for the sensor to justify its design and development. Zygo Corp. chose a scanning laser beam for their laser telemetric system (LTS) because of its proven accuracy, reliability, and versatility. Other advantageous characteristics of the LTS design concept include:

1. Noncontact measurement allows the LTS to measure moving, delicate, soft, moving, hot, radioactive, and (with special software) cylindrical objects made of transparent materials.
2. The position of the object being measured can be anywhere in the measurement volume of the instrument. This means that the object can move laterally and axially as its dimension is being measured, and the distance between the sensor and the object can be anywhere from several inches to several feet.
3. More than one object can be measured simultaneously.
4. A typical system will cost on the order of $5000.

5.10.1 Design Concepts

The basic function of the LTS instrument is to provide high-speed, noncontact precision measurements of various types of parts in many different environments. As shown in Figure 5.10.1, the laser beam is directed through the collimating lens, which acts as a beam shaping optic and is reflected off a rotating mirror known as a scanner to produce an angularly scanned laser beam. The surface of the scanner is located at the focus of the collimating lens, which converts the angularly scanned laser beam to a parallel scanned laser beam. The parallel scanned laser beam is partially blocked by the object being measured, placed at what is known as the *passline* of the system, which creates a shadow of the object. The shadow is actually a time-varying loss of light. What seems to the eye to be a static shadow projection is in reality a time-dependent on off on signal when the shadow is viewed as a function of time.

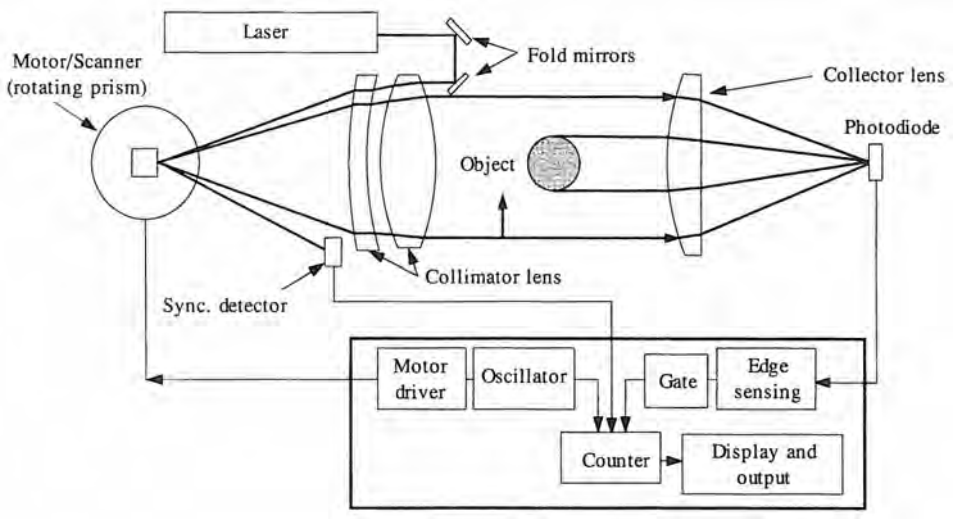


Figure 5.10.1 Laser scanner for dimensional measurement; simplified schematic of models 110 and 120. (Courtesy of Zygo Corp.)

The parallel scanned laser beam is then focused by a collector lens onto a photodetector which converts the light signal to an electrical signal. The signal is equivalent to the Gaussian energy distribution of the laser beam convolved with the attenuation of the scanned beam caused by the object. The electrical signal will be a nearly constant value during the time when the light is not blocked by the part, and nearly zero when the light is blocked by the part. Since the intensity of the laser beam has a Gaussian profile, the electrical signal generated by the photodiode during the transition from light to dark, or vice versa, will be the integral of this function, commonly referred

Error source	Type
Alignment of object being measured	Systematic
Motion of object being measured	Systematic and random
Atmospheric effects	Random
Dirt, dust, oil	
In measurement region	Random
On object	Systematic and random
Temperature	Systematic
Surface finish of object	Systematic and random
Edge-sensing errors	Systematic
Stray light	Systematic and random
Process effects	Systematic and random

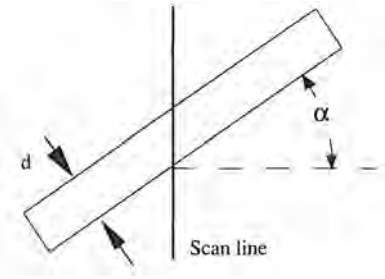


Figure 5.10.2 Error caused by object inclined with respect to scan line. (Courtesy of Zygo Corp.)

to as the cumulative-distribution function. Amplification and electronically differentiating (with respect to time) the signal gives the Gaussian shape again. Electronically differentiating the signal a second time gives an S-shaped signal with a zero crossing at the center of the Gaussian curve. Where the signal is zero (the peak of the Gaussian curve), represents the exact edge of the part (U.S. Patent 3,907,439). This method of edge sensing also eliminates any sensitivity to power fluctuations of the laser. The time between these edge signals is measured and corrected, thus yielding dimensional information about the object. The dimensional output is calculated in only a fraction of a second.

The internal components of an LTS are normally grouped into three distinct subsystems: the transmitter, the receiver, and the controller. The transmitter houses the laser, the scanner/motor, the optical bar, the collimating optic, the synchronous (sync.) detector or autocalibration mask (autocal.), and the beam shaping optics. The receiver houses the collector optic, the photodiode, and the preamplifier or digitizing electronics. The controller contains the remaining electronics for the measurement calculations, such as the reference clock, correction table, operating software, and power supply.

The LTS is used in a variety of environments which range from an inspection room to the rolling line of a steel mill. Hence the instrument needed to have separable transmitter and receiver and allow the object being measured to be placed anywhere in the measurement space with no degradation of performance. The sensor also had to be insensitive to mounting orientation and provide for a simple mounting technique. Although the concept of an LTS is quite simple, these environmental requirements make implementing it into an industrial product with the required performance a significant technical challenge. Limitations on LTS performance are caused by errors introduced either by the instrument used to make the measurement (internal sources) or by the geometrical or environmental factors related to the particular measurement situation (external sources). These errors, whether internal or external, are either systematic or random in nature.[35]

5.10.2 External Error Sources

For the laser telemetric system, the primary external error sources and their classification include:

[35] Recall that systematic errors can be mapped but not reduced by averaging data, while random errors cannot be mapped but can be reduced by averaging.

Alignment of Object Being Measured

If the object being measured is tilted with respect to the scan line of the measurement beam, the geometrically introduced error shown in Figure 5.10.2 appears. The geometrical error is

$$\varepsilon_a = d\left(\frac{1}{\cos\alpha} - 1\right) \tag{5.10.1}$$

where d is the dimension of interest and α is the tilt angle. Equation 5.10.1 shows that for a constant error, the allowable tilt decreases as the dimension being measured increases.

Motion of Object Being Measured

The dimension of an object is directly related to the time the beam takes to scan that object. If the object is moving along the same axis as the beam is scanning, there will be an error induced which is given by

$$\varepsilon_m = \frac{dV_{tp}}{V_b - V_{tp}} \tag{5.10.2}$$

where d is the dimension being measured, V_{tp} and V_b are the transverse velocity of the object being measured and the measurement beam, respectively. It is important to realize that only the velocity component of the object parallel to the measurement beam motion is a source of error. If the part motion is unidirectional, this error is systematic. If the part motion is due to vibration, then the error will be random in nature and can be substantially reduced by averaging.

Atmospheric Effects

A 0.0025 mm (0.0001 in.) accuracy requirement over a 250 mm (10 in.) throat corresponds to 2 arcseconds (10 μrad) of collimation. Note that variations in the refractive index of the air due to temperature differences are capable of deflecting a light beam about 5 arcsecond/C ^{o}. Thus individual measurements may fluctuate by approximately 0.0075 mm (0.0003 in) due to atmospheric turbulence and temperature inhomogeneities. Averaging can substantially reduce this source of error; however, since turbulence has a 1/f noise spectrum, averaging will only reduce the error a limited amount.[36]

Dirt, Dust, and Oil

Particulate matter, oil droplets, or other light-beam-interrupting substances present in the measurement region can shift the energy profile of the measurement beam and therefore introduce uncertainty in the measurements. As with other random errors, averaging will reduce these effects. On the object being measured, an oil film or other particulate matter can introduce systematic errors. The magnitude of the errors depends on the size and nature of the particular contaminant. A high-pressure air wipe on the part prior to its entering the measurement region can help to eliminate this error.

Temperature

The temperature of the part being measured directly effects its size. Since all engineering materials have some level of sensitivity to temperature, there will be a systematic error given by

$$\varepsilon_T = d_{To}(\alpha_p - \alpha_i)(T - T_o) \tag{5.10.3}$$

where d_{To} is the dimension of the object at standard temperature T_o, α_p is the coefficient of thermal expansion of the part's material, α_i is the coefficient of thermal expansion of the instrument, and T is the temperature at the time the measurement is taken. If both coefficients of thermal expansion and the temperature are well known, this error can be corrected for in software.

Surface Finish of Object

If the object being measured has a surface finish which has a roughness as large as the measurement beam diameter, the dimension of the part will have an error associated with the microfinish of that part. The exact value of the error depends on the type of surface finish on the object. The exact amount of error the LTS senses is a function of the beam diameter and the part roughness profile, as shown in Figure 5.10.3. The LTS has the inherent capability to sense a measurement near the

[36] A complete explanation of 1/f noise is beyond the scope of this section. In brief, 1/f noise has a spectral density proportional to a power of 1/f. When averaging time is increased, the noise in the bandwidth 1/t is also greater so that errors are only reduced to a limited extent. For an in-depth discussion, see C. D. Motchenbacher, Low Noise Electronic Design, Wiley-Interscience, New York, 1973.

rms surface of the object being measured. This is important to consider when comparing LTS data with contact gage data since contact gages tend to measure peak-to-peak dimensions. One type of surface finish is that produced by turning a part on a lathe. The pattern can be similar to the threads of a screw. In this case the error is given by

$$\varepsilon_L = p\frac{\tan\left(\frac{\delta}{2}\right)}{2} \tag{5.10.4}$$

where p is the feed per revolution of the cutting tool and δ is the angle of the cutting toolpoint. Other more random surface finishes present different errors.[37]

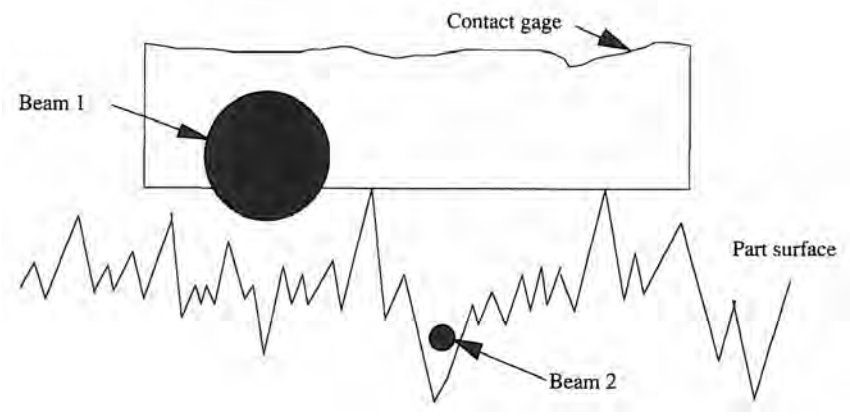


Figure 5.10.3 Errors caused by relative beam and part dimensions. (Courtesy of Zygo Corp.)

Edge Sensing Errors

Edge-sensing errors are known to occur on cylindrical objects made of transparent materials, with the greatest effect seen on hollow tubes. The amount of error is affected by part diameter, laser beam diameter, and wall thickness (if the part is tubular). Typically, a transparent object will produce a signal with well-defined outer edges, poorly defined inner edges (if the part is tubular), and many other phantom edges caused by reflection, refraction, dirt, or interference as the light passes through the part. The extra edges may be ignored through the use of special software or hardware which inhibits the signal during the noisy interval. However, under some conditions the outer edges of the part do not produce good enough edge transitions to achieve accurate measurements.

There are a number of theories which explain the source of this error. One is that the wall thickness (in tubular objects) is less than the diameter of the beam, causing the signal from the outer edges to be distorted by the signal from the inner edges. Another theory is that a portion of the beam is refracted and/or reflected through the part with a small enough angle that it can still enter the collector lens and recombine with the actual edge signal, thus causing an error. Generally, an instrument with a smaller beam diameter will produce the most accurate results, but each application should be reviewed to determine measurement limitations.

Stray Light

Stray light from external sources does not usually cause errors to occur unless it is great enough to saturate the preamp or detector. Stray light will also cause a problem if the intensity variation is at a high-enough frequency to pass through the differentiators. An effective method for controlling stray light is to filter it out using an optical bandpass filter which transmits only the frequency of the laser light. For example, this is the method used for applications involving measurement of incandescent hot objects.

Process Errors

Process errors are "wild card" errors that are dependent on the particular process and thus cannot be generalized. A particular process may produce conditions in which any or all of the error

[37] For example, hot-rolled steel bars fresh out of the mill.

sources mentioned are present. Other error sources could be produced by the introduction of additional optics in the laser beam path or by unique conditions of the process. Special considerations must be taken if a process error is believed to exist.

5.10.3 Optical Error Sources

Optical error sources in the LTS are due to material and manufacturing process inhomogeniety, and they can be controlled by proper specification of components and fabrication processes. The precise reasons for these errors are not fully understood, but enough empirical information has been gathered to make intelligent predictions and recommendations for the components in question.

Laser Related Errors

The laser has some inherent characteristics which cause errors in the LTS. These are laser beam amplitude noise and beam pointing changes, spot shape changes, and sensitivity to optical feedback. Amplitude noise on the laser beam at low frequencies (less than 1 kHz) will not affect the measurement because of the high-frequency response of the differentiation electronics. Amplitude noise at higher frequencies, however, may affect the measurement. Differentiators are usually combined with low-pass filters to eliminate noise at frequencies above the range of interest.[38] This noise may be random or systematic. Random noise, as stated before, can be reduced by averaging. An analog differentiator is essentially a single-pole high-pass filter used below the passband. This results in a gain proportional to frequency, or with two differentiators, gain will be proportional to frequency squared. For example, if the differentiators are designed for signals with frequency components in the 2 MHz range, noise from the laser (20 kHz) power supply would be attenuated by a factor of 10,000.

Systematic noise in the range 1 to 2 MHz may be caused by electrical oscillation of the laser tube due to too small a ballast resistance, or in the 20 kHz range by insufficient filtering on the laser power supply output. Systematic noise may affect accuracy by biasing a measurement slightly to be a multiple of the noise period. Systematic noise may also cause dimensions that are exact multiples of the noise period to appear more stable than other dimensions. The amount of possible measurement error can be determined by

$$\varepsilon_m = \frac{kNV_b}{s_r} \tag{5.10.5}$$

where ε_m is the measurement error in microinches per millivolt, k is a scaling constant (to match units), N is the laser noise in millivolts, V_b is the beam velocity in inches per second, and s_r is the slew rate of the second derivative in volts per microsecond.

When the laser tube changes length due to temperature variations, small changes in the beam direction occur as the laser oscillation finds its most stable axis in the laser cavity. This effect occurs when the laser *modes* and causes the beam to strike the optics at a different position. The output beam of a typical laser is usually a Gaussian energy distribution known as TEM_{00}. If the laser end mirrors are misaligned or some other nonuniformity exists in the laser resonator cavity, the energy distribution could be different and also cause errors in the measurement.

To minimize the effect of the first three error sources, a polarized laser was selected with good stability and acceptable noise levels that are verified during incoming inspection. The solution to the last error was to mount all optics near the laser with sufficient tilt to prevent back reflection from reentering the laser and changing the energy profile characteristics of the beam. Antireflection coatings on the optics were tried but met with limited success.

Collimating Lens Errors

The error sources related to the collimating lens include ray slope errors, internal interference effects, high-frequency slope errors, cosmetic surface quality errors, scan speed linearity errors, and thermal expansion errors. Ray slope errors, thermal expansion errors, and scan speed linearity errors can be minimized in the optical design of the lens. The ability of the lens design to correct for these errors is limited by the type of lens or mirror chosen.

[38] See, for example, G. E. Tobey et al., Operational Amplifiers, Design and Applications, McGraw-Hill Book Co., New York, 1971.

Ray slope errors are the errors in the design of the lens that can be seen in the collimation accuracy of the lens. The total amount of error is affected by manufacturing tolerances and the physical limitations of the lens design. The amount of ray slope error which can be tolerated in a scan lens design is given by

$$R_s = \tan^{-1}\left(\frac{\varepsilon}{2M_T}\right) \tag{5.10.6}$$

where R_s is the maximum ray slope error in the lens, ε is the desired measurement error of the system, and M_T is the measurement throat length of the system.

As an example, the lens used in the first LTS was an air-spaced doublet which was designed to give less than 2 arcsecond ray slope errors and constant scan speed (Ktheta) performance.[39] In a common lens the output ray height for a given input ray angle is approximately equal to the focal length of the lens times the sine of the input angle. The Ktheta design produces an output ray height equal to a constant (K) times the input ray angle (θ). This lens was used with a simple electronics package that had no stored correction tables to maintain accuracy. The lens used in the later LTSs had nearly perfect collimation (parallelism of outgoing rays) but required a stored correction table to correct for the scan speed linearity errors inherent in the design.

Internal interference effects are also design related and can be avoided by following a few guidelines; all lenses should be used with the optical axis tilted perpendicular to the scan direction with an angle sufficient to displace the beam one beam diameter. If tilting is not possible, a high efficiency antireflection coating should be used. In any cemented lens design the index of refraction of the cement should match the index of refraction of one of the elements.

High-frequency slope errors and cosmetic surface quality are fabrication process related. Optics used in the LTS should have the final polishing done on low-speed polishing machines since high-speed polishers produce high-frequency slope errors which will not be filtered out well by the differentiators. Cosmetic surface quality errors are the result of defects in the surface of the optic in the system.

The operation of the laser telemetric system is based on the intensity variation of the laser as it encounters an object. For this reason any defect on the optics or dirt particle which causes an intensity variation of the light signal as the beam encounters the object can cause an error in the measurement. Since the laser beam is small in diameter, ranging from 0.05 to 5 mm (0.002 to 0.2 in.), on the optics in the system, a defect of 0.03 mm (0.0012 in.) in diameter can cause a significant error in the measurement. The allowable defect size for a given instrument can be approximated by

$$S_L = 2\varepsilon\left(\frac{d_o}{d_p}\right) \quad \text{For lenses} \tag{5.10.7a}$$

$$S_M = \frac{\varepsilon}{2}\left(\frac{d_o}{d_p}\right) \quad \text{For mirrors} \tag{5.10.7b}$$

$$D_L = d_o\sqrt{\frac{\varepsilon}{d_p}} \quad \text{For lenses} \tag{5.10.7c}$$

$$D_M = \frac{d_o}{2}\sqrt{\frac{\varepsilon}{d_p}} \quad \text{For mirrors} \tag{5.10.7d}$$

where S_L, S_M, D_L, and D_M are scratch and dig[40] sizes for lenses and mirror surfaces respectively, d_o is the laser beam diameter on the optic in question, d_p is the laser beam diameter at the passline, and ε is the accuracy error desired for the instrument. One can see that mirror surfaces are approximately four times more sensitive than lens surfaces to defects. Also, these equations show that a large beam on the optic (d_o) and a small beam at the passline (d_p) increases the allowable scratch or dig size for a constant error. The only way to control the cosmetic surface quality is through tight quality control and in-process inspection.

---

[39] U.S. Patent 3,973,833.

[40] A *scratch* is a line defect of width in microns and length defined in terms of a geometric parameter of the optics, such as lens thickness. A *dig* is a pit in the optics defined in terms of diameter in tens of microns. A size 10 dig is 100 μm in diameter.

Window Related Errors

The windows are necessary to protect the internal components of the LTS from environmental contaminants and air turbulence. The windows on an LTS are subjected to the same design considerations as the other optical components of the system. The windows cause the same type of errors as caused by the lens with one important addition. Any internal wedge (nonparallelism of the front and back surfaces) of the windows causes an interference phenomenon that modulates the output beam of the instrument. By tilting the window and fabricating the window with the wedge perpendicular to the scanned beam, this error can be minimized. The collection of dirt, dust, and oil on the windows will also lead to the degradation of accuracy in the instrument. Periodic cleaning, improved environment, or special air wipes can reduce or eliminate this error.

Collector Lens Errors

The function of the collector lens, as the name implies, is to collect the light and focus it onto the photodetector. Because it is so close to the end of the beam path, the collector lens is the least critical optical element in the measurement beam of the LTS. The errors in the collector lens which can cause errors in the accuracy of the LTS include cosmetic surface quality, internal interference, and lens aberrations. Lens aberrations, primarily spherical aberration, in the collector lens can cause two distinct errors in the LTS.

Since the photodetector active area is usually quite large, the collector lens should only need to produce a focused spot smaller than this area in order for the photodiode to "see" the entire scanned beam. However, as the diameter of the part increases, the amount of reflected light from the edge of the part which can enter the receiver increases. The error due to reflection off the part is given by

$$\varepsilon_R = d\left[1 - \cos\left(\frac{\gamma}{2}\right)\right] \tag{5.10.8}$$

where ε_R is the error due to reflection, d is the diameter of the part, and γ is the field angle of the receiver.

The error introduced by this effect can be enhanced by the aberrations in the collector lens. A good example of this occurs when a simple plano-convex lens is used as the collector. The positive spherical aberration in this lens at the point of minimum blur will cause the edge rays to have an effective field angle greater than the rays at the inner portions of the lens shown in Figure 5.10.1. This will increase the error given by Equation 5.10.8. By focusing this system slightly ahead of the "best" focus position[41] normally chosen, the errors can be reduced. The error can also be reduced by improving the lens design performance to reduce the aberrations present.

The light beam is also diffracted as it passes over the edge of the part. However, this diffraction pattern is small and more than 99% of the beam energy is collected and focused by the collector lens so that no error in the measurement from diffraction is detected.

Beam Shaping Optics

The beam shaping optics are used to expand or focus the laser beam to the desired size for a particular instrument or application. The Gaussian nature of the power distribution of the laser beam causes the beam to form soft, focused zones known as waists. As discussed in the section on surface roughness, the beam size plays an important role in determining the sensitivity of an LTS system to part roughness. Also, the beam size affects the ability of the electronics to sense an edge accurately as well as the measurement throat range of the instrument and the sensitivity of the instrument to cosmetic defects on the optics. A small beam size increases the slew rate of the second derivative and decreases the zero-crossing detection error. A small beam also diverges at a faster rate as given by

$$w(z) = w_0\sqrt{1 + \left(\frac{\lambda z}{\pi w_0^2}\right)^2} \tag{5.10.9}$$

where $w(z)$ is the beam radius at distance z, w_0 is the beam waist radius, and λ is the wavelength of the laser light. Since the gain of the electronics is optimized at the passline of the instrument (where the beam waist is located), the signal degrades as the beam size increases. The limit of the beam size is usually given by $w(z) = 1.414\ w_0$. This criterion forces instruments with long measurement throat

[41] The *best focus position* is located where the beam has a minimum spot diameter.

requirements to have large-diameter beams and high-accuracy instruments to have small-diameter beams.

The types of beam shaping optics vary from single-element lenses to two-element telescopes which can expand or contract the beam and can be set to various effective focal lengths. The beam waist size and position in the measurement region can be calculated by consecutive iterations of the following equations:

$$\frac{1}{w^2} = \frac{1}{w_1^2}\left(1 - \frac{d_1}{f}\right)^2 + \frac{1}{f^2}\left(\frac{\pi w_1}{\lambda}\right)^2 \tag{5.10.10}$$

$$d_2 = (d_1 - f)\left[\frac{f^2}{(d_1 - f)^2} + \left(\frac{\pi w_1}{\lambda}\right)^2\right] \tag{5.10.11}$$

where d_1 and d_2 are the distances from the optic to the first and second beam waists, respectively, f is the focal length of the optic, λ is the wavelength of the laser light, and w_1 and w_2 are the first and second beam waist radii, respectively. Beam shaping (control of spot size) can be as simple as using a single lens or a two-element telescope. Also, it is common to pass the laser beam through the collimating lens off axis from the scanned beam to act as a beam expander. Beam expanders can be made using cylindrical lenses to expand the beam in only one direction. The instruments with these types of beam expanders have been used to measure extremely rough surfaces.

Air Turbulence Errors

The sources of air turbulence internal to the LTS are somewhat different from the external sources; however, the ray pointing errors are essentially the same. External air turbulence is due primarily to the environment and is sometimes impossible to correct for. There are two sources of internal air turbulence. One source of the internal air turbulence errors are temperature gradients which occur between the internal components of the LTS. The second source is due to the pumping action of the rotating scanner. Since the nature and location of the heat-producing components is well known, steps can be taken to minimize their effects. The main heat sources in an LTS are the laser, the laser power supply, the scanner motor, and the internal electronics. The heat generated by these components can be isolated from the main optical path by careful design of the instrument. This removes the direct heat gain to the beam path. Indirect heating remaining can best be controlled by baffling the optical path to reduce convective heat transfer. The circulation-reducing baffle[42] accomplished this in the following manner. In order to reduce the heat transfer to purely conduction and radiation, air motion needs to be prevented. By designing a baffle system with the proper plate spacing, the air motion can be significantly reduced and system noise can be reduced by a factor of 5.

The second source of air turbulence is the pumping action caused by the scanner as it rotates. The scanners used in the LTS have evolved over the years in a variety of forms. Monolithic prisms with four or five sides, cemented glass/metal assemblies, and two-sided glass plates have all been utilized with good results. With all these designs there is a certain amount of air motion created by the rotating prism, which causes errors in the measurement. The exact amount of air turbulence created by the scanner is a function of the diameter, the rotational speed, the number of facets, the type of construction, and the geometry of the surrounding structure. Effective scanner shrouds to minimize turbulence have been fabricated by placing an enclosure with a small opening for the laser beam to enter and exit around the scanner.

5.10.4 Mechanical Error Sources

Mechanical error sources can be grouped into three categories: temperature effects, stress effects, and vibration effects. The transmitter optical bar is the area of major concern when considering the mechanical error sources and methods of reducing them. The optical bar consists of the components which hold the optics and electronics in precise alignment and allow the user of the system to mount the sensor in the particular application without affecting its accuracy. The optical bar is usually made from combinations of Invar, aluminum, and stainless steel.

[42] U.S. Patent 4,427,296.

The dimensional stability of the LTS is derived from the ability of the system to accurately locate the edges of a part in time and reference them to a known standard. Thus the optical path must remain constant from the time of manufacture to the time the measurement is taken. Also, since the operating temperature of the instrument can vary, the temperature sensitivity of the instrument must be controlled by design. Hence stability is a direct measure of the successful design of the optical bar.

Temperature-Induced Errors

Temperature-induced effects are by far the most difficult errors to correct or prevent by design. As mentioned earlier, all engineering materials have some measurable coefficient of thermal expansion. There are a few which are less sensitive, but they can be expensive and difficult to work with. Some examples are Invar, Zerodur®, U.L.E., and fused silica. Invar is an iron-nickel alloy containing 36 wt% nickel. Zerodur® and U.L.E. are glass ceramic alloys with proprietary compositions, and fused silica is synthetic quartz. There are also composite materials which are insensitive to temperature but are sensitive to humidity. Also all optical materials used for visible light have a thermal coefficient of refractive index.

There are two methods for designing temperature insensitivity into an LTS optical bar. The first method used is to match various materials, usually a low-expansion material with a higher-expansion material. This produces an optical bar that has a temperature coefficient equal to that of the collimating optic:

$$\frac{\mathrm{d(O.P.L.)}}{\mathrm{dT}} = \frac{\mathrm{d(O.B.L.)}}{\mathrm{dT}} \tag{5.10.12}$$

where d(O.P.L.)/dT is the change in optical path length of the collimating optic due to temperature, and d(O.B.L.)/dT is the change in optical bar length due to temperature. These two quantities are given by

$$\frac{\mathrm{d(O.P.L.)}}{\mathrm{dT}} = \mathrm{f}\alpha_{\mathrm{eff}} \tag{5.10.13a}$$

$$\frac{\mathrm{d(O.B.L.)}}{\mathrm{dT}} = \{\mathrm{L}_1\alpha_1 + \mathrm{L}_2\alpha_2 + \cdots + \mathrm{L}_\mathrm{n}\alpha_\mathrm{n}\} \tag{5.10.13b}$$

where f is the focal length of the collimating optic, α_{eff} is the effective coefficient of thermal expansion for that optic which includes dimensional and refractive index changes due to temperature, L_1, L_2, . . ., L_n are the lengths of the optical bar components, and α_1, α_2, . . ., α_n are the coefficients of thermal expansion for the respective optical bar elements. This approach athermalizes the optical bar and is used for instruments which require a large depth of measurement range or high accuracy.[43]

The second method utilizes the fact that a ray entering a lens at a given angle maintains the same ray height at the front focus of the lens for small changes of the origin of that ray along the optical axis. This limits the position where the instrument can accurately measure to requiring the user to place the center of the object at the front focal point. On the other hand, it enables the design engineer to simplify the optical bar design considerably by allowing the use of high expansion easy to manufacture materials such as aluminum without any adverse effects.

Other thermally induced errors are related to differential expansion of the optical mounts and optics. Metals exist such as Kovar, which are intended to be thermally matched to some optical materials, but the selection is limited and cost of fabrication may present problems. This forces the use of common materials with special mounting techniques. One widely used method of optical mounting is to use an elastic adhesive such as RTV silicone rubber. RTV allows differential expansion to occur but holds the optic securely to its locating surface during use. Another technique clamps the optic in pure compression with resilient pads. There is some risk since the mounting surfaces must match exactly in order to guarantee that no bending moments are induced. A third method athermalizes the optical mount by calculating the amount of adhesive thickness necessary to compensate for the differential growth of the mount and optic. In most cases kinematic design practices are a must to prevent distortion of the optics.

[43] This is a similar method to that shown by example in Section 2.3.5 for maintaining bolt preload in temperature-varying systems.

Stress-Induced Errors

Stress-induced error sources refer to any built-up stress in the LTS which is not due to differential thermal expansion. They include assembly and mounting stresses induced by dimensional errors, and weight-induced stresses. Assembly- and mounting-induced stresses can best be controlled by kinematic design and designs which allow for component inaccuracies. Weight-induced stresses can be compensated for with careful analysis and choice of sections to maximize stiffness-to-weight ratios.

Vibrational Errors

Vibrational error sources can be induced by the motor scanner assembly and external sources. These usually occur if the scanner is not balanced and another component in the system has a natural frequency near that of the rotational frequency or drive frequency of the motor. The entire LTS should be mounted on vibration isolation mounts.

5.10.5 Electrical System Error Sources

The pure electrical error sources include preamp electrical noise, insufficient preamp bandwidth, signal quality degradation, signal distortion, group delay distortion, time interval measurement errors, offset voltage errors, response time errors, and photo detector response nonuniformity.

Preamplifier Noise

Preamplifier noise sources include shot noise from detector dark current and op-amp bias current, Johnson noise from detector source resistance, feedback resistor or detector load resistor, amplifier short-circuit noise voltage, and external noise. Shot noise is caused by current flowing through a semiconductor junction. Amplifier short-circuit noise refers to the noise that an amplifier would have if its input was connected to a noiseless short circuit and is usually the largest contributor since the detector capacitance and feedback resistance combination causes it to be amplified at higher frequencies. External noise may be capacitively or inductively coupled into the circuit, from within the equipment, or RF energy could be coupled in to affect operation directly. Noise from the power supply is easily reduced with proper regulation and filtering. Capacitively coupled noise may be reduced by connecting a capacitor from circuit ground to chassis ground near the preamp. The preamplifier design should be optimized for low noise over the entire bandwidth of interest. Usually, the signal from the detector is great enough so that its noise is not a problem.

Preamplifier Bandwidth

The bandwidth required for the preamp and signal processing path depends on the laser beam spot size on the part and the beam scanning velocity. The electrical signal of the second derivative of the edge signal has its spectral peak at a frequency given by

$$f_{d''} = \frac{4V_b}{d_p} \tag{5.10.14}$$

where $f_{d''}$ is the spectral peak frequency, and V_b and d_p are the beam velocity and beam diameter, respectively. The electronics should ideally have a bandwidth about three times this frequency; however, since the noise also increases with bandwidth, the actual bandwidth is usually only twice this frequency.

Signal Quality Degradation

If the electronics is partitioned with the preamp in a separate enclosure from the other electronics, proper cable driving and terminating practices should be observed so that the signal quality does not depend on cable length. Still it may be difficult to prevent external noise from adding to the signal on the cable; thus if the preamp, differentiators, and zero-crossing detector are in a separate enclosure from the other electronics, digital signals may be sent through the cabling with less worry about noise and signal degradation. The associated logic should be designed so that a single pulse on a single wire is produced for any edge transition. If this is not done, variations in logic threshold voltage or device propagation delay may affect the measurement.

Signal and Group Delay Distortion

Signal distortion usually falls into two categories, slew rate limiting and group delay distortion. Slew rate limiting occurs when the signal changes more rapidly than the amplifier can follow. This is distinctly different from bandwidth limitations, and may be recognized by the way rising and falling edges of the signal have constant slope rather than smooth curves. Group delay is defined as the derivative of phase of the transfer function. Group delay distortion occurs when some frequency components of the signal are delayed more than others as they pass through the electronics. When the scanned laser beam is focused to a smaller spot in the measurement region, the line of sharpest focus is slightly curved. Therefore, the frequency spectrum produced by one edge of the object being measured may be slightly different from the spectrum produced by the other edge. If some frequencies are delayed more than others in the electronics, an apparent measurement error may occur. By using a technique of group delay equalization described in U.S. Patent 4,427,296, this error can be reduced.[44]

Time Interval Resolution

The resolution of the time interval measurement must be finer than random errors from other sources. If the random error sources are greater than the quantization uncertainty, both may be reduced together by averaging. If the quantization errors are larger than the random error sources, the result of averaging will still contain the quantization errors. This results in a paradox, that reducing the random errors from other sources below this level might actually result in less accurate measurements after averaging unless the quantization errors are also reduced. Equal effort is required on all fronts. This may seem counter-intuitive, but will be clearer with an example:

- Assume that the exact measurement = 499.60. Gaussian 3σ noise = ± 3 quanta, and we take 1000 readings: We have two readings at 503, 27 readings at 502, 155 readings at 501, 356 readings at 500, 324 readings at 499, 118 readings at 498, and 18 readings at 497, for an average of 499.601, or an error of 0.001 quanta.
- Assume the same exact measurement, no noise, and we take 1000 readings: We have 1000 at 500, for an average of 500. The error is now 0.4 quanta.

The result is the noise acts like a dither signal that keeps the error from locking in on one value.

The time interval error depends on the counting resolution of the reference clock. As mentioned earlier, the LTS determines the dimension of an object by timing the separation of two edge signals. The time between these two edge signals is a function of the focal length of the collimating optic and the rotational velocity of the scanner motor. Thus the accuracy of an LTS is partially determined by the number of clock cycles per scan. Direct counting of time intervals by counting clock cycles becomes difficult above 40 MHz (25 ns). A typical high-accuracy LTS system with 0.25 μm (10 μin.) repeatability requires 1.1 ns counting ability. To obtain finer resolution, two methods are most attractive: delay line interpolation and analog interpolation.[45]

A delay line is an electrical device which delays a signal by a predictable amount of time. The most obvious example of this is a coaxial cable, which will delay a signal about 1 ns per foot of cable. A more practical implementation uses discrete or distributed inductors and capacitors in a small package resembling an integrated circuit. The delay line passes the edge signal through a tapped delay line with N taps and a total delay equal to the reference clock period. Each cycle of the reference clock takes a snapshot of the N delayed edge signals to determine a fractional time interval. This approach has a practical resolution limit of about 5 ns. Analog interpolation uses a pulse which is generated with a width of one clock period plus the unknown fraction. A constant current is integrated during the width of this pulse, and the resulting voltage is measured. This approach is more complicated, but has a practical resolution limit of 0.5-1.0 ns.

Offset Voltage Error

Offset voltage errors occur if the zero-crossing detector for the second derivative does not switch at exactly zero volts. This effect can be seen when the two edge transitions are in opposite directions, as for a hole or a diameter measurement. The dimension will measure either smaller

[44] See H. J. Blinchikoff and A. I. Zverev, Filtering in the Time and Frequency Domains, Wiley-Interscience, New York, 1976.

[45] U.S. Patents 4,332,475 and 4,427,296, respectively.

or larger than it actually is, depending on the polarity of the error. The amount of error may be calculated in a manner similar to the calculation shown for laser amplitude noise (Equation 5.10.5). This error will cancel out if the direction of edge transition is the same on both edges.

Response Time Error

The zero crossing of the second derivative is usually sensed by a comparator: An integrated circuit device which compares two analog signals (in this case, one is zero) and gives a digital output, depending on which is greater. Errors due to response time may be present since most comparators have a response time that is affected by the slope of the input signal, the direction of the slope of the input signal, temperature, and load on the output. To reduce this error a comparator must be chosen which has fast and predictable response.

Detector Response

As the laser beam scans, the point where it is focused on the receiver diode moves slightly. If the detector has nonuniform response over its active area, motion of the spot on the detector surface will appear to be a change in signal. This can produce confusing symptoms similar to those caused by optical interference in the windows.

Software Errors

Discretization and roundoff errors occur in all digital computations. The more complicated the software program, the higher the likelihood of bugs and the more difficult it is to find them. By careful study of the software by more than one person and by statistical analysis of the measurement results (looking for unexplainable error distributions), these bugs can usually be found and eliminated.

5.10.6 Electro-optomechanical Errors

Electro-optomechanical errors occur when a component is subject to a combination of optical, mechanical, and electrical effects and the resulting error cannot be recreated or identified under different conditions. These errors are present in components including the scanner motor, the scanner and sync. detector, and autocal.

Scanner Motor

Motor speed errors can be caused by any or all of the following: low-frequency (average) motor speed changes, high-frequency (instantaneous) motor speed changes, on/off effect, and electromechanical vibrational coupling. Since the beam velocity is directly proportional to the motor velocity, motor speed errors directly affect the accuracy of the instrument. There are two types of motor speed errors present in any electric motor. They are low-frequency or average speed fluctuations and high-frequency or instantaneous speed variations. Low frequency motor speed changes cause the measured size of a part to fluctuate inversely with the speed of the motor. This can be corrected by ratiometrically comparing a stored reference edge dimension with the apparent reference edge dimension and correcting the part size accordingly. In a system without a reference aperture, the motor speed must be derived from the same clock used to measure the time intervals. A synchronous motor operated on power derived from the measurement clock is one possibility. This has the disadvantage of requiring a power amplifier to provide typically 10 W to power the motor. The ideal motor for this system may be a brushless dc motor with a very accurate, high-resolution encoder and phase-locked speed control system; however, this alternative tends to be rather expensive.

High-frequency speed variations are caused by the motor poling effects or feedback loop effects in the case of a dc motor. High-frequency speed changes cannot be removed; instead, the motor needs to have enough inertia to reduce this effect to an acceptable level. In fact, a large rotational inertia improves both low- and high-frequency motor speed fluctuations. Another method of reducing the effects of high-frequency motor speed errors is to average the measurement data in multiples of the number of scanner facets. This guarantees that each measurement is the average of one or more complete revolutions of the motor. A related error is caused by turning the instrument off and on. This error is very unusual and was discovered during testing of early systems. In the first prototype instruments, the scanners were four-sided glass cubes. When the units were being tested, a large repeatable error was observed when the unit was turned off and back on. After much further

testing the problem was identified to be related to the high-frequency speed variations of the motor and the number of facets on the scanner. The synchronous motor being used would synchronize on a different pole each time the motor was turned on. This caused the speed profile over each scan to be different each time the unit was powered up. The effect was also found to occur only when the number of poles in the motor was an even multiple of the number of facets on the scanner. The solution to the problem was simply to increase the number of scanner facets to five. The five-sided scanner produced scan speed fluctuations which were random for any pole position of the motor. Thus when the measurement data was averaged, the on/off error was eliminated.

Electromechanical vibrational coupling occurs most often in systems in which the motor is driven off the ac line voltage. This error is caused by resonant vibrations in the rotor, scanner, and electrical windings, as the motor is subjected to slightly varying line frequencies. The resonant vibrations of the rotor scanner assembly beat against the motor poles and cause high-frequency motor errors to be a function of drive frequency. This error can be reduced by reducing the scanner inertia or increasing the motor shaft diameter to raise the mechanical natural frequency well above the motor drive frequency. It is important to note that a 0.25 μm error can correspond to a motor angle change of as little as 0.2 μrad.

The Scanner

The scanner evolved over the years in various forms. Two-, four-, and five-faceted scanners have been used with success. Both single-piece glass scanners and composite scanners of metal and glass are in use today. The reasons for the various scanner types are many. Two reasons already mentioned were electro-mechanical vibration and on/off poling effects. Two other reasons for the different types of scanners are the duty cycle and number of scans required for a given time interval. For a given system the percentage of time actually used for an active scan during one revolution of the motor can be doubled simply by doubling the number of facets on the scanner. Also, if the system is electronically limited to a maximum beam velocity, the only method to increase the number of scans per second is to increase the number of facets on the scanner.

The other properties important to the scanner are angular pointing error between facets, distance of the scanner facet from the best-fit cylinder, and the figure and cosmetic surface quality of the facet surfaces. Angular pointing errors between facets on a scanner refer to tilt of a scanner facet perpendicular to the scan direction. This error can be seen if the scanned beam is viewed at a great distance from the scanner and separate scan lines are present. In the LTS, this causes the scans from different scanners to hit the other optics and auto-cal at different places, which may produce errors. If the distance of the scanner facets to the center of rotation of the scanner is not equal, there will be a slight focus error from facet to facet as the system operates. The errors caused by this effect will vary for each system and may cause an increase in the random noise of the measurement. The scanner is also as sensitive to cosmetic defects and surface features as the other components in the transmitter. In some cases the scanner is near an intermediate beam waist and cosmetic defects become even more critical.

These stringent requirements for the scanner made it necessary to use only optical polishing to produce the required surface for the LTS. Because of process control and inspection limitations available in the late 1970s, it was more economical to build composite scanners from glass mirrors bonded to metal hubs. Even in 1987, the technology for fabricating monolithic scanners with the desired performance was still being developed.

Sync. Detector and Autocal.

The synchronous (sync.) detector is used to trigger the electronics to start looking for measurement edge data. This in effect tells the electronics where the beam is at one point in time and space. It is the method used on the early systems and relies on a constant motor speed for accuracy. The autocalibration (autocal.) replaced the sync. detector on later systems and provided a reference aperture to be used to correct for low-frequency motor speed fluctuations in order to increase system accuracy. Also, the autocal. can correct for the inaccuracies in the athermalization of the optical system.

The autocal. can be placed in a number of different places with little effect to its function. The autocal. consists of a mask, or reference aperture, with two reference edges or slots with a set of photodiodes behind each edge or slot. Whether an edge or a slot is used is determined by the electronics of the system, but the function remains the same. The time between reference pulses is

measured on each scan and ratiometrically compared with a stored value. The object measurement is then multiplied by this ratio to obtain the correct output dimension.

The material used for the reference mask depends on the system configuration. A system which is athermalized will require a mask with low thermal expansion, while a system working at one focal length will require a mask matched to the thermal expansion of the optical bar material. The most critical feature of the autocal mask is the parallelism and surface finish of the reference edges or slots. Any nonparallelism or surface roughness may be seen by the instrument since temperature or scanner errors may move the beam along the autocal. The apparent change in the reference aperture will cause the instrument to output an error in the measurement.

5.10.7 Summary and Closing Remarks

Despite its simple concept, design of the LTS was a significant technological challenge and it evolved by considering all the error sources and making calculations and judgments based on them. It also must be realized that this information was compiled over a period of 13 years of experience and testing. The design engineer had to consider many aspects of electrical, optical, and mechanical engineering, plus the characteristics of the components, in order to produce the required accuracy and stability. The engineering design criteria were equally based on theoretical and empirical data determined through testing. Empirical data must often be considered when designing new instruments so further improvements can be made.

The first commercial LTS from Zygo was introduced in 1975. The LTS had a 50 mm (2 in.) measurement range, 250 mm (10 in.) measurement throat, 0.005 mm (0.0002 in.) accuracy, separable transmitter and receiver, and did not require frequent recalibration. This system relied on its opto-mechanical design to achieve the required system accuracy. The edge of the part was precisely sensed using a technique of electronically produced derivatives of the light signal.[46] With the advent of microprocessors, a new electronics system design for the LTS evolved which allowed the use of internal error correction tables and operating software. This allowed one system to function with different optical systems and increased the accuracy of the instrument. Also, the new software provided special process control functions and internal computation capability, making it possible for the LTS to supply the user with corrected data directly. A current state-of-the-art LTS can provide the user with measurement ranges from 25 mm (1 in.) to 460 mm (18 in.) and corresponding accuracies of 0.0008 mm (0.00003 in.) to 0.013 mm (0.0005 in.). The controller supports software for statistical process control, data trend histograms, multiple measurements, user-definable internal calculations, transparent object measurement, and expressions which use data from two or more sensors. Also, the controller can monitor up to four other types of sensors (i.e., temperature probes, linear encoders, etc.) through a digital interface.

[46] U.S. Patent 3,907,439.

Chapter 6
Mapping Geometric and Thermal Errors in a Turning Center

Quidquid agis, prudenter agas et respice finem - if you measure, do it with the greatest care and remember the measuring error.

Geoffrey G. Thomas

6.1 INTRODUCTION

This chapter[1] is a condensation of a Ph.D. dissertation, "A General Methodology for Machine Tool Accuracy Enhancement: Theory, Application and Implementation"[2] by Dr. Alkan Donmez of the National Institute of Standards and Technology. This work represents one of the first times that error maps and feedback from strategically placed thermocouples have successfully been used with software-based error correction algorithms to correct for dominant geometric *and* thermal errors in a machine tool. Accordingly, the primary intent of this chapter is to demonstrate the method whereby a machine's errors can be measured, mapped, and then partially compensated for.

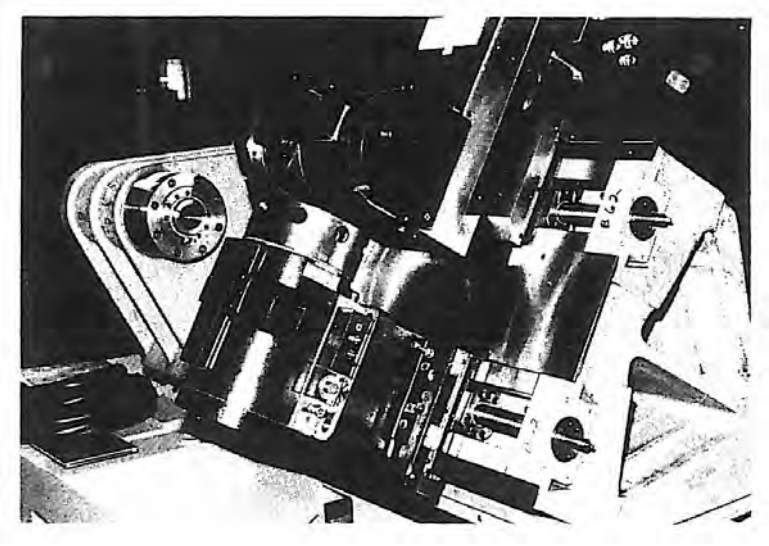


Figure 6.1.1 Two-axis slantbed turning center after assembly of principal components. The servomotor drives will be installed next followed by the protective covers. (Courtesy of Hardinge Brothers, Inc.)

The machine tool used in the experiments was a Hardinge Superslant® two-axis turning center shown partially assembled in Figure 6.1.1. For the purpose of generating an HTM representation as discussed in Section 2.2.1, the machine's coordinate frame assignment is shown in Figure 6.1.2. The structure of the machine consists of a spindle connected to the bed of the machine, modeled as a revolute (rotating) joint; a carriage, which is connected to the bed with a prismatic (sliding) joint; a cross slide, which is connected to the carriage with a prismatic joint; and a tool turret, which is connected to the cross slide with a revolute joint. The turret is locked in place for the cross slide when the machine is cutting. Finally, there is a cutting tool, which is rigidly connected to the tool turret, and a workpiece, which is rigidly held in the spindle.

[1] Section 6.1 was written by Alex Slocum. Sections 6.2–6.7 were written by Alkan Donmez and edited by A. Slocum.

[2] Alkan Donmez received the Ph.D. degree from Purdue University in 1985. The experimental portion of his thesis was conducted at the National Institute of Standards and Technology (formerly, NBS).

Before proceeding with the analysis of the kinematics of the machine, which includes defining the homogeneous transformation matrix representation of the structure, each of the major components will be discussed in detail.

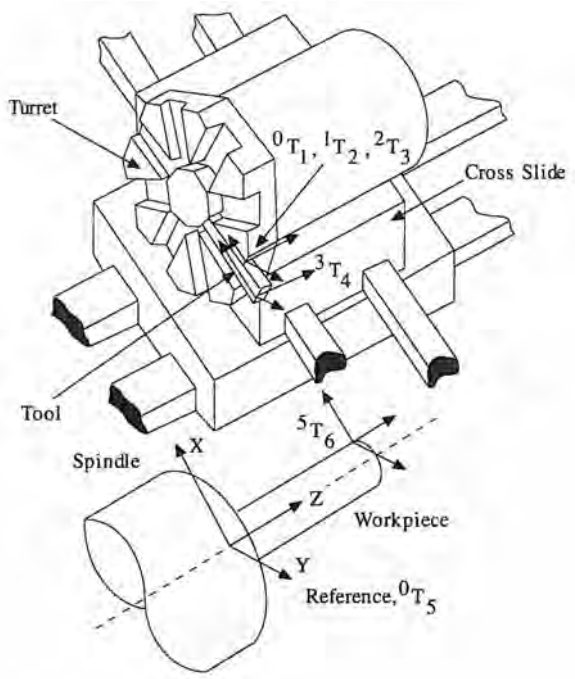


Figure 6.1.2 Coordinate frame assignments on a two-axis slant-bed turning center. (Courtesy of NIST.)

Construction of the Superslant®

In order to better understand the way a typical machine tool is actually put together, Figures 6.1.3 – 6.1.14 show the key components and assemblies of the Superslant®. The Superslant® was given its name for its slant-bed design. In high-volume manufacturing operations, chips from the cutting process can build up at an amazing rate. If the hot chips are not removed, they can act as a heat source and may interfere with the cutting process and degrade surface finish. A slanted or vertical bed on a lathe allows gravity to cause the chips generated during the cutting process to fall onto a moving conveyor which removes them from the machine.

Figure 6.1.3 shows the upper turret, which is octagonally shaped and designed to hold eight tools. The upper turret mount is shown in Figure 6.1.4. The tools can project radially, for turning the outside diameter of a part, or axially for boring the inside diameter of a part. The design of the system for holding the tools to the turret was developed through close collaboration with various tooling manufacturers, so virtually any standard lathe tooling can be mounted to the turret. The lower turret is round and has eight stations for tools such as drill bits and live centers. The final holes for seating these tools are machined into the turret using the assembled spindle of the machine itself; hence axial alignment is assured. In order to index the rotary position of the turret, a highly repeatable and accurate system with extreme rigidity is needed. Since only discrete motions are needed, a gear-type (Hirth or Curvic) coupling is best suited for this application.

As shown schematically in Figure 6.1.5, belleville washers apply a high constant force to a shaft which pulls the turret into the turret mount and forces the two face gears to mesh tightly. In order to index the turret, a pneumatic piston compresses the washers and then an electric or pneumatic motor rotates the turret. Rotation of the turret by the motor only has to be accurate to a degree or so, because the face gears will automatically align the turret when they are forced together. This is a common method used to achieve indexing motion. The accuracy of the coupling depends on a large number of engaging teeth (360 teeth are common), which act to help average out manufacturing errors in the tooth form. When kept clean and lubricated, they can achieve submicron repeatability.

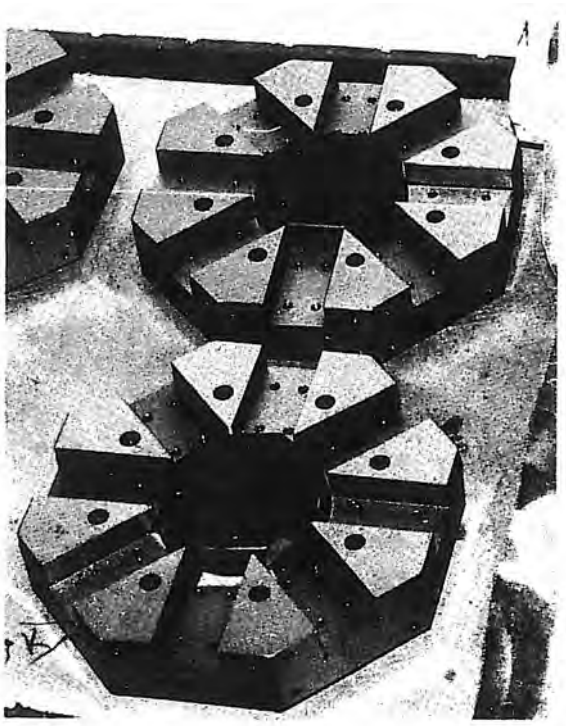


Figure 6.1.3 Upper tool turret. (Courtesy of Hardinge Brothers, Inc.)

On some machines, the lower turret is positioned by a cross slide which allows it to hold tools to cut a profile in the part. This allows the machine to machine the OD and ID of the part simultaneously, which enhances productivity. The machine studied herein has a lower turret which does not move in the X direction and thus represents the most accurate configuration. Figure 6.1.6 shows the lower turret mount riser mounted to the lower turret carriage. A turret mount similar to the one shown in Figure 6.1.4 is bolted and pinned to this riser. Figure 6.1.7 shows the components of the riser. A massive tee-shaped block with sliding contact bearing pads bolts to the bottom of the riser and sandwiches the bearing rails between them. Only one rail's edges are used for establishing straightness of motion along the length of travel, and its edges are sandwiched between an integrally cast edge on the riser and the neck of the tee so that horizontal parallelism of the rails is not critical. Precision hand-finished gibs are used to preload the sliding contact (Turcite® was used here) linear bearings. Figure 6.1.8 shows a bottom view of the riser and the widely spaced linear bearings. This configuration yields a design with high stiffness, damping, and repeatability.

The upper carriage subassembly is shown in Figure 6.1.9. Its design is similar to the lower carriage, except that the upper portion of the riser is replaced by the cross slide mechanism. The ballscrew actuator for the cross slide is mounted to the carriage assembly. The ball nut is anchored to a cast iron coupling block which is bolted and pinned to the cross slide. The craftsman who assembles the cross slide will hand scrape these surfaces to match. Thus when the cross slide is put in place over the bearings and bolted to the coupling block, minimal lateral loads will be imposed on the ball screw from forced geometric congruence. With this type of design, resolutions on the order of 1 to $1/2\,\mu$m can be achieved. Figure 6.1.10 shows the bottom of the cross slide and the tee-shaped blocks that act as bearing rails for the cross slide.

In order to finish scrape the carriages' bearing surfaces and gibs, which guide motion along the Z axis, ground-hardened steel bearing ways are mounted, aligned, and pinned to the cast iron bed. The first step is to scrape the surfaces of the bed flat so that the bearing rails will rest in the same plane (vertical parallelism). The next step is to establish one of the rails as a reference and then indicate in all other rails from the one reference rail to maintain horizontal parallelism between the spindle axis and the upper and lower turrets.

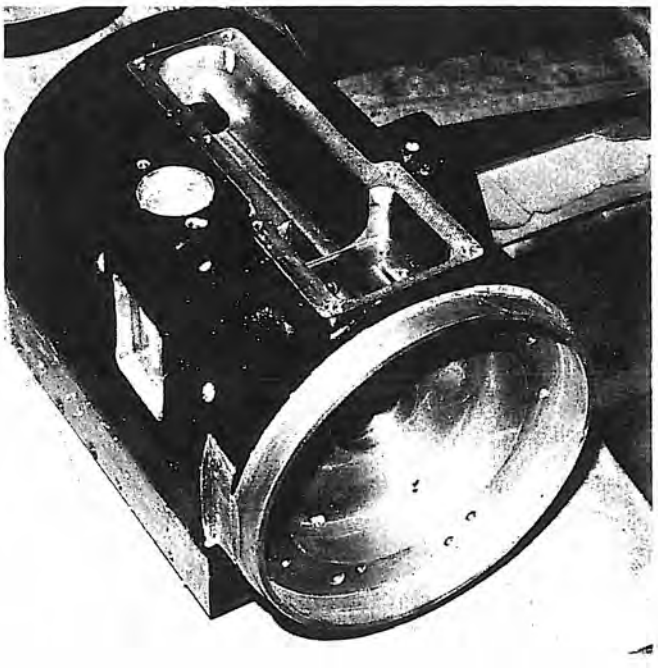


Figure 6.1.4 Upper turret mount to cross slide. (Courtesy of Hardinge Brothers, Inc.)

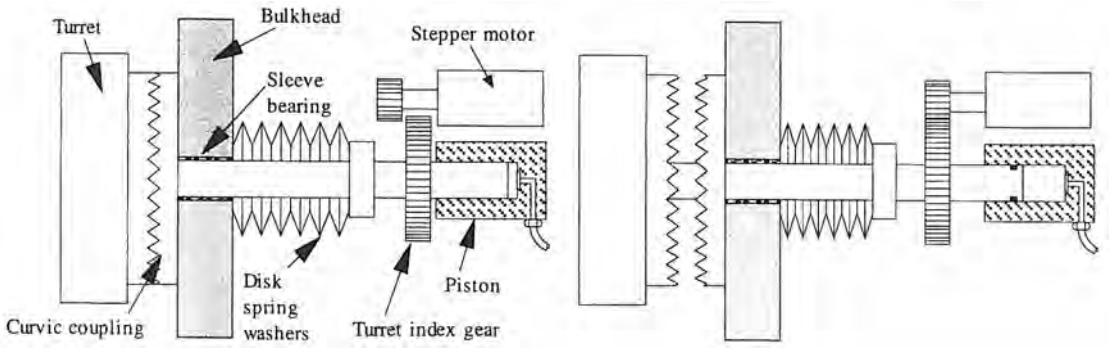


Figure 6.1.5 Typical turret indexing and locking mechanism in engaged and disengaged positions. (Courtesy of Hardinge Brothers, Inc.)

After the bearing rails are bolted and pinned in place, the end supports for the ballscrews can be installed, as shown in Figure 6.1.11. Next come the carriages, ballscrews, and the headstock, as shown in Figure 6.1.1. As shown in Figures 6.1.12 and 6.1.13, the headstock is a very stiff structure that is also hand finished prior to assembly. The casting for the bed is a rigid triangular structure that rests on three points, as shown in Figure 6.1.14.

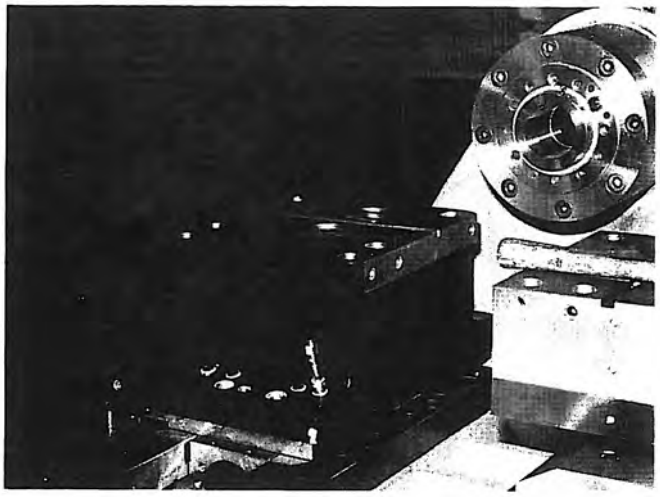


Figure 6.1.6 Lower turret's riser mount on lower carriage. (Courtesy of Hardinge Brothers, Inc.)

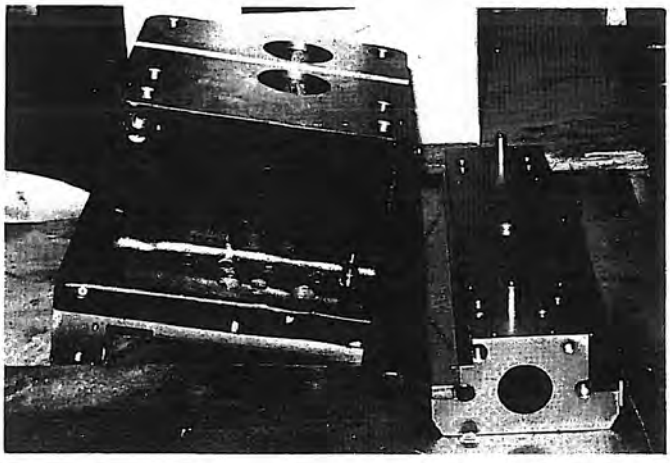


Figure 6.1.7 Lower carriage components. (Courtesy of Hardinge Brothers, Inc.)

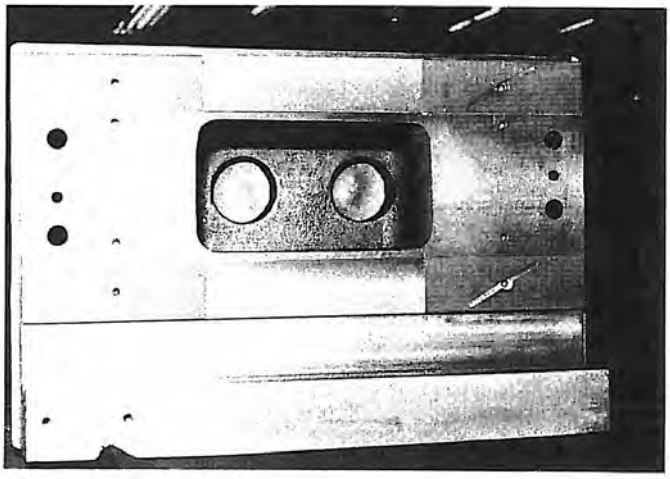


Figure 6.1.8 Lower carriage's antifriction sliding contact bearing pads. Note the oil distribution grooves. (Courtesy of Hardinge Brothers, Inc.)

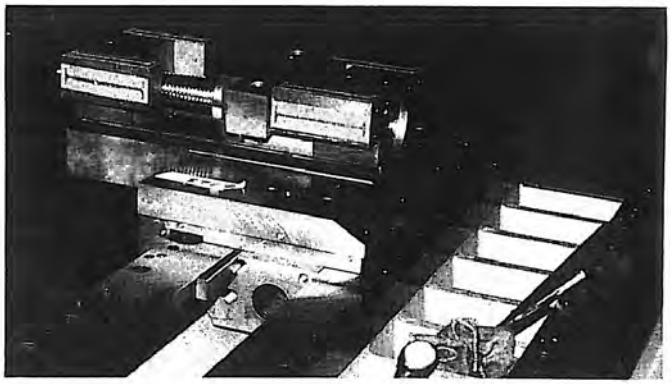


Figure 6.1.9 Upper carriage mounted on hardened bearing rails which are bolted and pinned to the cast iron bed. The cross slide's bearing blocks and leadscrew are mounted to the carriage ready to receive the cross slide. (Courtesy of Hardinge Brothers, Inc.)

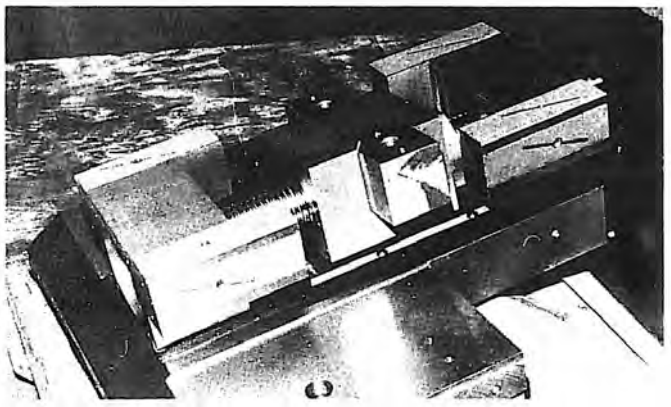


Figure 6.1.10 Upper carriage subassembly. Note the antifriction sliding bearing pads. (Courtesy of Hardinge Brothers, Inc.)

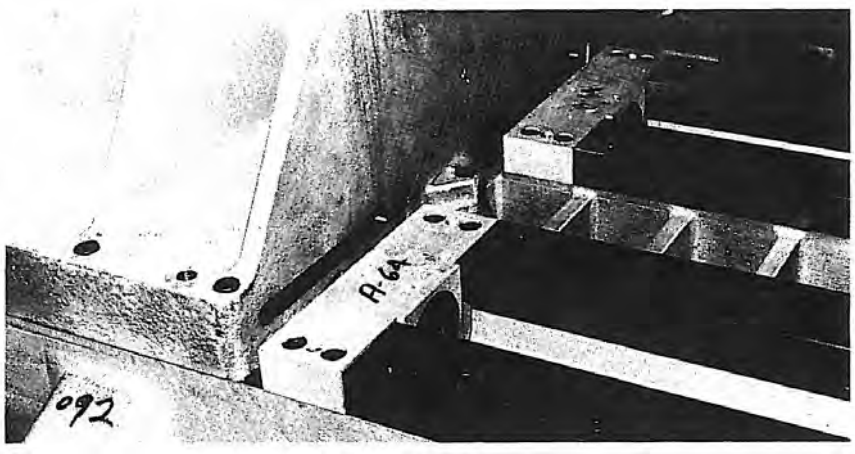


Figure 6.1.11 Headstock bolted to cast iron base. Also shown are upper and lower carriage bearing rails, and bearing supports for carriage ballscrews. (Courtesy of Hardinge Brothers, Inc.)

6.2 ASSEMBLING THE HTM MODEL OF THE SUPERSLANT®

As shown in Figure 6.1.2, the global reference frame is chosen to coincide with the spindle frame at the spindle nose. When the machine is cold, ${}^{0}T_{1}$ is the upper carriage coordinate frame with respect to the reference frame, ${}^{1}T_{2}$ is the upper cross-slide coordinate frame with respect to the carriage, and ${}^{2}T_{3}$ is the tool turret coordinate frame with respect to the cross-slide frame, all of which are located at a point on the turret where measurements of error motions can easily be made. ${}^{3}T_{4}$ is the cutting tool coordinate frame with respect to the tool turret frame, ${}^{0}T_{5}$ is the spindle coordinate frame with respect to the reference frame (they coincide), and ${}^{5}T_{6}$ is the workpiece frame with respect to the spindle frame.

For the carriage, the only degree of freedom is in the Z direction, so the ideal HTM for this axis is

$$ {}^{0}T_{1} = \begin{bmatrix} 1 & 0 & 0 & X_1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2.1} $$

where z is the servo-controlled motion along the Z axis and X_1 is a constant offset that is a function of the cross-slide X position. For the cross-slide the servo-controlled degree of freedom is in the X direction:

$$ {}^{1}T_{2} = \begin{bmatrix} 1 & 0 & 0 & x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2.2} $$

For the tool turret, motions are restricted during cutting; thus the ideal HTM is given by

$$ {}^{2}T_{3} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2.3} $$

The cutting tool, which is rigidly attached to the turret, has tip coordinates X_4, Z_4, so the ideal HTM is

$$ {}^{3}T_{4} = \begin{bmatrix} 1 & 0 & 0 & X_4 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & Z_4 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2.4} $$

For the spindle, although there is a rotational motion about the Z axis during cutting, there is no significance to the angular position about Z axis; hence this motion is assumed to be zero and the ideal HTM is

$$ {}^{0}T_{5} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2.5} $$

For the workpiece which is rigidly attached to the spindle, the ideal HTM is

$$ {}^{5}T_{6} = \begin{bmatrix} 1 & 0 & 0 & x(w) \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & z(w) \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2.6} $$

where x(w) and z(w) are the coordinates of the ideal cutting point on the workpiece.

The homogeneous transformation matrices representing the actual position and orientation of each element, given all possible errors, can be obtained using the error matrices discussed in Chapter 2. For the carriage the errors are

$$ {}^{0}T_{1} = \begin{bmatrix} 1 & -\varepsilon_z(z) & \varepsilon_y(z) & \delta_{(x)}z + X_1 \\ \varepsilon_z(z) & 1 & -\varepsilon_x(z) & \delta_y(z) \\ -\varepsilon_y(z) & \varepsilon_x(z) & 1 & \delta_z(z) + z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2.7} $$

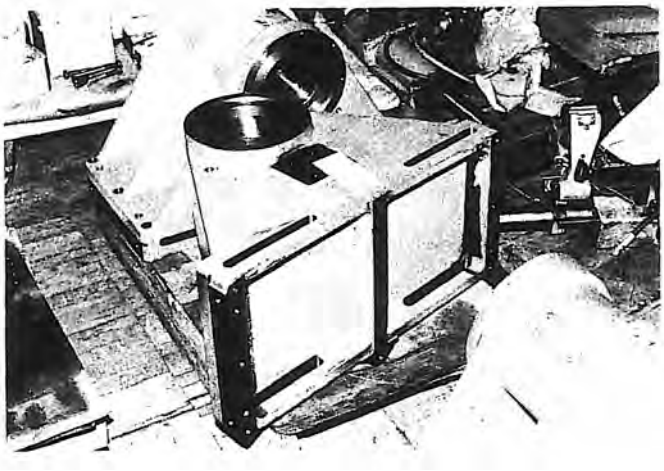


Figure 6.1.12 Headstock casting after machining. Note horizontal slots in the base, which allow sliding way covers for the carriage axes to pass through. (Courtesy of Hardinge Brothers, Inc.)

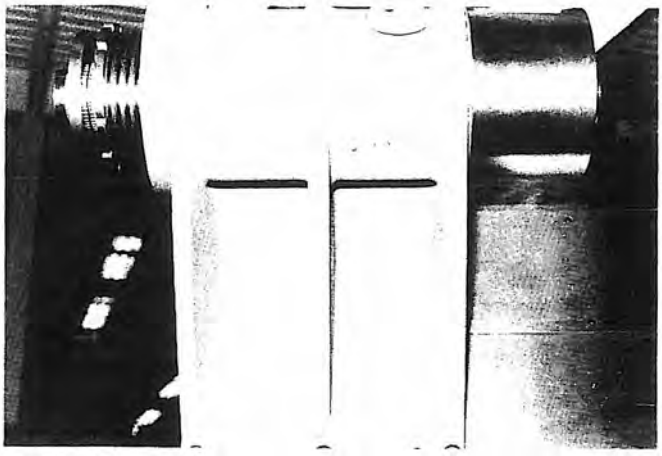


Figure 6.1.13 Headstock with spindle assembly in place, mounted on bed. All mounting surfaces are hand scraped prior to final assembly. (Courtesy of Hardinge Brothers, Inc.)

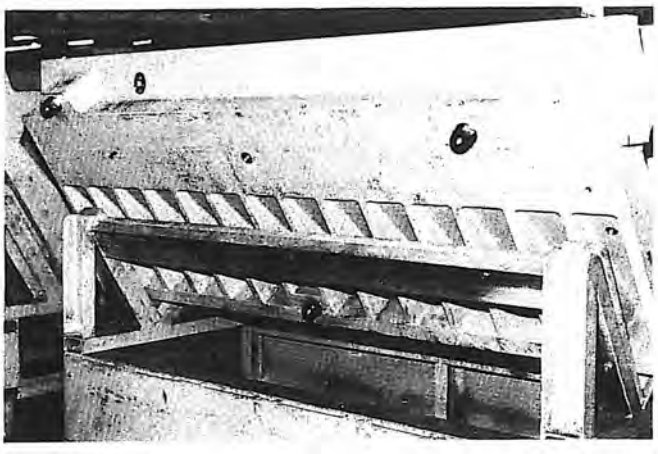


Figure 6.1.14 Three point kinematic mount on bed's bottom plane, shown prior to bed being placed on steel stand. (Courtesy of Hardinge Brothers, Inc.)

where the errors (z) are a function of the carriage's position along the Z axis. Note that:

- $\varepsilon_z(z)$ is the roll error of the carriage.
- $\varepsilon_x(z)$ is the pitch error of the carriage.
- $\varepsilon_y(z)$ is the yaw error of the carriage.
- $\delta_y(z)$ is the Y straightness of the carriage.
- $\delta_z(z)$ is the displacement error of the carriage.
- $\delta_x(z) = \delta'_x(z) + \alpha_p \Delta z$:

where

- $\delta'_x(z)$ is the X direction straightness of the carriage as it moves in the Z direction.
- α_p is the horizontal parallelism error (rotation about the Y axis) between Z motion and the mean axis of rotation of the spindle.
- Δz is the incremental Z motion that amplifies α_p to cause an Abbe error in the X direction.

For the cross-slide the errors are

$$ {}^1T_2 = \begin{bmatrix} 1 & -\varepsilon_z(t) & \varepsilon_y(t) & \delta_{(x)}x + x \\ \varepsilon_z(x) & 1 & \varepsilon_x(x) & \delta_{(y)}(x) \\ -\varepsilon_y(x) & \varepsilon_x(x) & 1 & \delta_{(z)}x \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2.8} $$

where the errors (x) are a function of the cross-slide's X position:

- $\varepsilon_x(x)$ is the roll error of the cross-slide.
- $\varepsilon_z(x)$ is the pitch error of the cross-slide.
- $\varepsilon_y(x)$ is the yaw error of the cross-slide.
- $\delta_y(x)$ is the Y straightness of the cross-slide.
- $\delta_x(x)$ is the displacement error of the cross-slide.
- $\delta_z(x) = \delta'_z(x) + \alpha_o \Delta x$:

where

- $\delta'_z(x)$ is the Z direction straightness of the cross-slide as it moves in the X direction.
- α_o is the orthogonality error between X motion of the cross-slide and the Z axis average line of the spindle rotation.
- Δx is the incremental X motion that amplifies α_o to yield an Abbe error in the Z direction.

For the tool turret which is similar to the spindle shown in Figure 2.2.3 the errors are

$$ {}^2T_3 = \begin{bmatrix} 1 & -\varepsilon_z(t) & \varepsilon_y(t) & \delta_x(t) \\ \varepsilon_z(t) & 1 & -\varepsilon_x(t) & \delta_y(t) \\ -\varepsilon_y(t) & \varepsilon_x(t) & 1 & \delta_z(t) \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2.9} $$

where the errors (t) are functions of the tool turret's rotational position:

- $\delta_x(t)$, $\delta_y(t)$, and $\delta_z(t)$ are translational error motions of the turret.
- $\varepsilon_x(t)$, $\varepsilon_y(t)$, and $\varepsilon_z(t)$ are rotational error motions of the turret.

For the cutting tool, since the coordinate frame is assigned at the tip of the tool,[3] rotational errors have no effect on its position,[4] but the dimensions X_4 and Z_4 will amplify angular errors in other axes. The errors for the cutting tool are thus

$$ {}^3T_4 = \begin{bmatrix} 1 & 0 & 0 & X_4 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & Z_4 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \delta_x(c) \\ 0 & 1 & 0 & \delta_y(c) \\ 0 & 0 & 1 & \delta_z(c) \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & X_4 + \delta_x(c) \\ 0 & 1 & 0 & \delta_y(c) \\ 0 & 0 & 1 & Z_4 + \delta_z(c) \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2.10} $$

where

[3] The XYZ location of the tip of the tool must be determined by direct measurement on the lathe by using a device called a *tool setting station.*

[4] The notation (c) designates an error in the assumed geometry of the cutting tool which is not a function of position.

- X_4 and Z_4 are the ideal tool dimensions in the X and Z directions, respectively.
- $\delta_x(c)$, $\delta_y(c)$, and $\delta_z(c)$ are the change in the tool length in the X, Y, and Z directions, respectively, caused by wear or mounting errors and measured by the tool-setting station.

For the spindle, due to unrestricted rotational motion about the Z axis, assume that $\varepsilon_z(s) = 0$ and the errors[5] are

$$ {}^0T_5 = \begin{bmatrix} 1 & 0 & \varepsilon_y(s) & \delta_x(s) \\ 0 & 1 & -\varepsilon_x(s) & \delta_y(s) \\ -\varepsilon_y(s) & \varepsilon_x(s) & 1 & \delta_z(s) \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2.11} $$

where

- $\delta_x(s)$ is the radial motion of the spindle in a sensitive direction.
- $\delta_y(s)$ is the radial motion of the spindle in a nonsensitive direction.
- $\delta_z(s)$ is the axial motion of the spindle.
- $\varepsilon_x(s)$ is the tilt of the spindle in a nonsensitive direction.
- $\varepsilon_y(s)$ is the tilt of the spindle in a sensitive direction.

Note that only thermal drift components of these error motions will be considered. The dynamic spindle error motions as a function of rotational position cannot be corrected due to the fact that machine tool controller does not have sufficient bandwidth. For example, a 1000 rpm (16.7 Hz) spindle speed requires a controller bandwidth of 167 Hz (10 times), but the machine controller only has a bandwidth of 50 Hz. In addition, the cross-slide only has a bandwidth on the order of 10 Hz and could not respond fast enough to correct for dynamic radial errors.

Assuming no rotational errors, the workpiece error matrix is

$$ {}^5T_6 = \begin{bmatrix} 1 & 0 & 0 & x(w) \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & z(w) \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \delta_x(w) \\ 0 & 1 & 0 & \delta_y(w) \\ 0 & 0 & 1 & \delta_z(w) \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & x(w) + \delta_x(w) \\ 0 & 1 & 0 & \delta_y(w) \\ 0 & 0 & 1 & z(w) + \delta_z(w) \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.2.12} $$

where

- x(w) and z(w) are the coordinates of the point on the workpiece which ideally interacts with the cutting tool to obtain desired workpiece geometry.
- $\delta_x(w)$, $\delta_y(w)$, and $\delta_z(w)$ are the errors in the location of this point in the X, Y, Z directions, respectively.

Using these transformation matrices, the position of the cutting tool with respect to the reference frame is represented by the following matrix multiplication:

$$ \begin{aligned} {}^{Ref}T_{tool} &= {}^{Ref}T_{carriage}\,{}^{carriage}T_{cross\ slide}\,{}^{cross\ slide}T_{turret}\,{}^{turret}T_{tool} \\ &= {}^0T_1\,{}^1T_2\,{}^2T_3\,{}^3T_4 \end{aligned} \tag{6.2.13} $$

Similarly, the ideal workpiece-cutting tool interface point on the workpiece with respect to the reference frame is given by

$$ \begin{aligned} {}^{Ref}T_{work} &= {}^{Ref}T_{spindle}\,{}^{spindle}T_{workpiece} \\ &= {}^0T_5\,{}^5T_6 \end{aligned} \tag{6.2.14} $$

The error matrix **E** could be calculated using Equation 2.2.15; however, as noted in the discussion of Equation 2.2.16, for a Cartesian machine it is desired to obtain the error correction vector in the reference frame. Thus the error correction vector is $\mathbf{p}_E = {}^{Ref}\mathbf{p}_{work} - {}^{Ref}\mathbf{p}_{tool}$ where the vector **p** is defined in Equation 2.2.1:

$$ \begin{aligned} \mathbf{P}_{Ex} &= x(w) + \delta_x(w) + \varepsilon_y(s) * z(w) + \delta_x(s) - X_4 - \delta_x(c) - [\varepsilon_y(z) + \varepsilon_y(x) + \varepsilon_y(t)] * Z_4 \\ &- \delta_x(t) - x - X_1 - \delta_x(x) - \delta_x(z) \end{aligned} \tag{6.2.15} $$

[5] The notation (s) designates spindle error, which is not a function of position.

$$\begin{aligned} \mathbf{P}_{\mathbf{E}y} &= \delta_y(w) - \varepsilon_x(s) * z(w) + \delta_y(s) - [\varepsilon_z(z) + \varepsilon_z(x) + \varepsilon_z(t)] * X_4 - \delta_y(c) \\ &+ [\varepsilon_x(z) + \varepsilon_x(x) + \varepsilon_x(t)] * Z_4 - \delta_y(t) - \varepsilon_z(z) * x - \delta_y(x) - \delta_y(z) \end{aligned} \tag{6.2.16}$$

$$\begin{aligned} \mathbf{P}_{\mathbf{E}z} &= -\varepsilon_y(s) * x(w) + z(w) + \delta_z(w) + \delta_z(s) + [\varepsilon_y(z) + \varepsilon_y(x) + \varepsilon_y(t)] * X_4 - Z_4 \\ &- \delta_z(c) - \delta_z(t) + \varepsilon_y(z) * x - \delta_z(x) - \delta_z(z) - z \end{aligned} \tag{6.2.17}$$

In these equations, the position coordinates are:

- X_4 and Z_4 are the cutting tool offsets.
- x and z are the nominal machine positions.
- x(w) and z(w) are obtained from workpiece geometry.

The error components ε and δ in these equations are primarily functions of the following:

- $\delta_x(w)$ and $\delta_z(w)$ are workpiece position errors caused by thermal and static load deformations.
- $\delta_x(s)$, $\delta_z(s)$, and $\varepsilon_y(s)$ are spindle thermal drift.
- $\delta_x(c)$ and $\delta_z(c)$ are functions of thermal and static load deformations and wear of the cutting tool.
- $\delta_x(t)$, $\delta_z(t)$, and $\varepsilon_y(t)$ are functions of angular position of the turret.
- $\delta_x(x)$, $\delta_z(x)$, $\varepsilon_y(x)$, $\delta_x(z)$, $\delta_z(z)$, and $\varepsilon_y(z)$ are functions of machine tool geometry and thermal and static load deformation of the cross-slide and carriage, respectively.

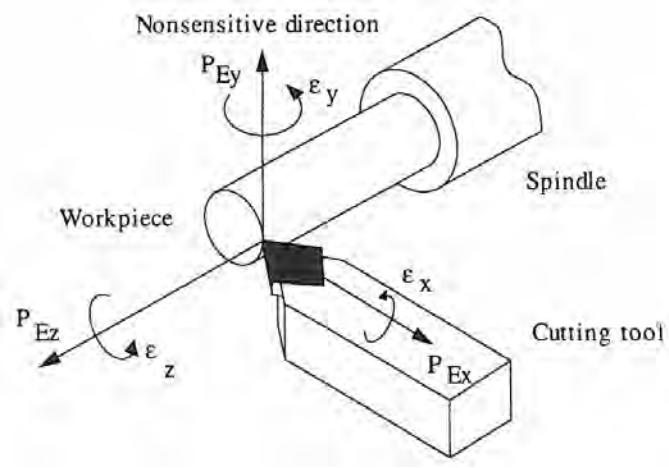


Figure 6.2.1 Resultant error components at the tip of the cutting tool. (Courtesy of NIST.)

Figure 6.2.1 illustrates the resultant error components at the tip of the cutting tool. Since the machine tool's servoed axis can only correct for displacement errors, angular errors at the tool tip are not of concern for the error correction algorithm, but they may be of concern for the machine's design engineers and users. Fortunately, the angular errors were not of significance.

This model, which was applied to the two-axis turning center, can be modified for other types of machine tools, such as multiaxis machining centers. The major difference would be additional matrices corresponding to the additional elements of the machine structure, in Equations 6.2.13 and 6.2.14. For example, to include other types of errors, such as fixturing error, a matrix representing the fixture should be premultiplied with the workpiece matrix. *An important criterion in the selection of the locations of the coordinate frames is that the relative errors should easily be defined in order to facilitate measurement.* Once the model is generated, the next step is to measure or predict the individual error components appearing in the resultant error equations. This will be demonstrated in following sections.

6.3 MACHINE TOOL METROLOGY

In Section 6.2 the resultant error at the tip of the cutting tool was modeled in terms of the combination of individual error components corresponding to different structural elements of the machine tool. In order to determine the resultant error at any location and time in the machine work zone, all of these error components must be measured or an intelligent guess made as to their magnitude. To complicate matters, geometric error components of the structural elements of the machine tool also change as a result of thermal effects and loading conditions. The Superslant® was designed to be rigid enough that nonstatic loading errors were insignificant; thus only geometric and thermally induced changes of the geometric errors had to be evaluated.

Since two independent variables, the nominal position and the thermal state of the machine, are considered to affect geometric errors, measurements must be made in order to determine the relationships between these variables over their possible ranges. Error maps can be obtained by using the machine's own scales to measure position, while the machine uses a probe to measure the shape of a known artifact[6]; however, the degree of accuracy obtained will not match that obtained by the external mapping process described here.

For each machine axis, measurements were made when the slide under study was at one end of its travel range. Then the slide was moved toward the other end of its travel range while a reading was taken at every measuring interval. The motion was reversed at the end of the travel, and the slide was sent back to its starting position again, with a reading taken at every measuring interval. With this procedure, it was possible to measure the reversal error (hysteresis) on each axis. This measurement must be made because the leadscrew backlash compensation algorithms built into the machine tool controller cannot fully account for nonlinearities of the leadscrew or thermally induced errors.

When selecting the measuring intervals, periodic error components, such as those caused by leadscrew misalignment,[7] also had to be considered. In order to eliminate periodic lead errors from the measurements, the measuring interval was selected at even multiples of the leadscrew lead. In order to characterize the periodic lead error, without thermal contamination of the data, measurements of the periodic error were taken separately over a very short travel range (one or two lead lengths) with the assumption of uniformity over the whole travel range.

Figures 6.3.1 and 6.3.2 show typical temperature profiles of several locations on a machine structure obtained during the preliminary temperature measurements. These measurements indicated that approximately 8 to 10 h was required for a typical machine to reach thermal equilibrium under continuous running conditions in an average machine shop environment. In order to determine the effect of temperature changes on machine errors, these errors must be measured over the whole temperature spectrum of the machine. Hence a typical measurement period often lasted for 2 or 3 days.

The error measurement started when the machine was cold. After five cycles of measurements, each of which consisted of movement over the whole travel range of the axis being measured, the machine was then allowed to warm up a little bit. Two types of warm-up procedures were used. In the first one, the machine was on and the axis drive motors were being powered by the servo loops holding the slides at a given position, but there was no movement. This caused a slow warm-up. In the second one, warm-up was induced by moving the machine slide back and forth. By using these procedures, the machine was gradually warmed up over a 2-day period, and the measurements were taken after each warm-up exercise. During each measurement cycle, in addition to the error movements of the slide under study, the temperatures of the locations around the slide and machine were monitored. The critical locations at which the temperatures were monitored included bearing housings at both ends of the leadscrew,[8] the slideway, the drive motor, the bed on which the slide is moving, the fixture for the measurement sensor, and the ambient air. This measurement procedure allowed for the investigation of changes in geometric errors corresponding to changes in the thermal state of the machine.

---

[6] See F. Jouy, "Theoretical Modelization and Experimental Identification of the Geometrical Parameters of Coordinate-Machines by Measuring a Multi-directed Bar," Ann. CIRP, Vol. 35, No. 1, 1986, pp. 393–396.

[7] If the leadscrew has a bow in it, every time it goes through one revolution the relative forced lateral geometric congruence will go from plus to minus and back to plus, thereby appearing periodic.

[8] Often, the temperature of the leadscrew or nut is also monitored.

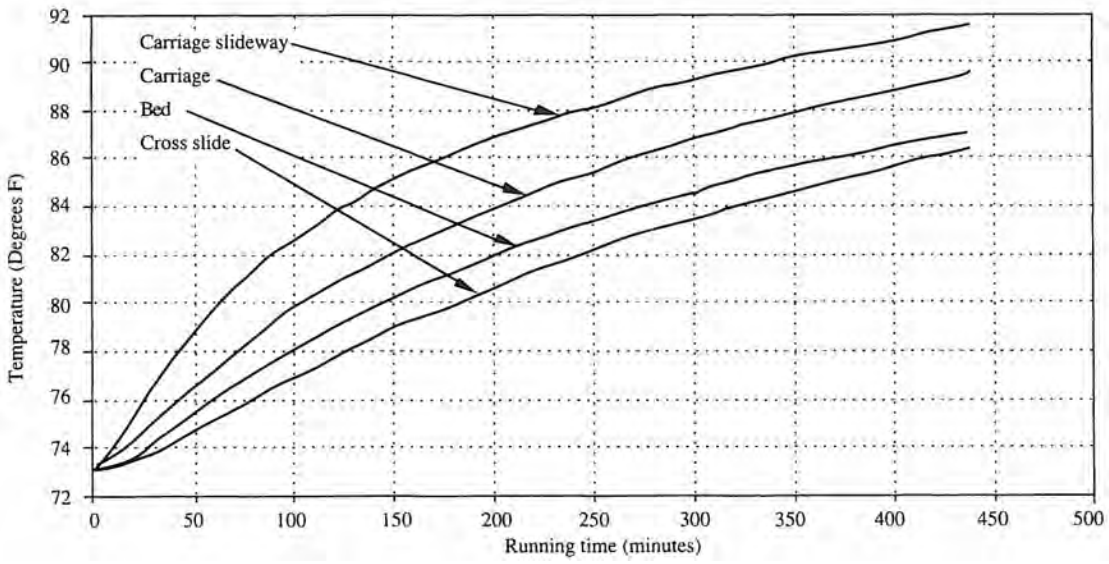


Figure 6.3.1 Temperature profile of the machine elements under continuous running conditions (2000 rpm, 100 ipm). (Courtesy of NIST.)

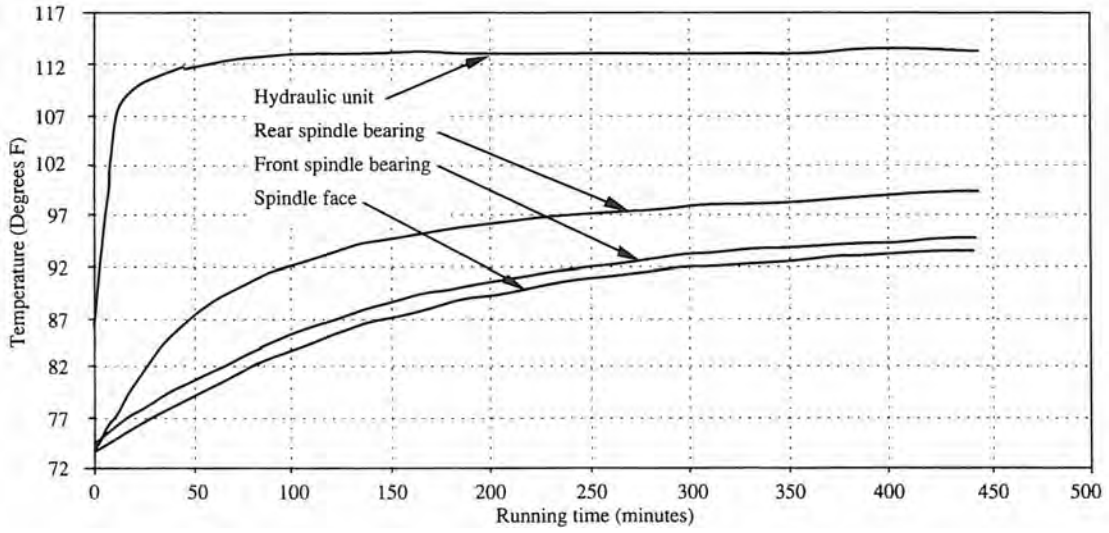


Figure 6.3.2 Temperature profile of the spindle under continuous running conditions (2000 rpm, 100 ipm). (Courtesy of NIST.)

The error components in Equations 6.2.15–6.2.17 were divided into four groups with respect to characteristic similarities, measurement procedures required, and the sensors used:

1. Linear displacement errors
2. Angular errors
3. Straightness, parallellism, orthogonality
4. Spindle thermal drift errors

6.3.1 Linear Displacement Errors

Linear displacement error is defined as the translational error motion of a machine element along its axis of motion. In general, this type of error is caused by the geometric inaccuracies of the drive mechanism and the feedback unit. In the case of ballscrew-actuated slides, lead errors of the ballscrew, misalignment between its axis of rotation and its centerline, irregularities in its geometry, and coupling errors between the feedback unit and the ballscrew cause linear displacement errors.

The best available device for displacement measurements over the travel ranges of machine slides is the laser interferometer. A Hewlett-Packard Model 5528 laser interferometer system was used for linear displacement error measurements in this study. Figure 6.3.3 shows the setup of the laser interferometer optics for linear displacement measurements along the X axis of the turning center. A similar setup was used for the Z axis. To maintain high accuracy levels, the pressure and temperature of the ambient air in the close proximity of the laser beam were monitored during the measurements, and the readings from the interferometer were compensated for accordingly.

6.3.2 Angular Errors

Angular errors are rotational errors caused by geometric inaccuracies of the slideways and the misalignment in the assemblies of structural elements of the machine tool. Yaw error is the rotational error of the slide around the axis perpendicular to the plane in which axis of motion lies. Roll error is the rotational error of the slide around the axis of motion, and pitch error is the rotational error of the slide around the third orthogonal axis of the slide. Although the contributions of all three rotational errors to the resultant error are significant in three or more axis machining centers, in turning centers, the primary components of roll and pitch errors are in the nonsensitive direction,[9] which is the direction perpendicular to the plane in which the two machine slides are moving. Therefore, for most turning centers, only the yaw error needs to be measured and mapped.

To measure the yaw error measurements, a laser interferometer was also used, although an autocollimator could have been used. Except for the angular measurement optics, the data acquisition system for the yaw measurements is the same as the one used for the linear displacement error measurements. Since the principle of the measurement is based on the comparison of the path lengths of the laser beam reflected from two retroreflectors, the environment does not significantly affect accuracy of the measurements. The setups of the laser optics for the cross-slide yaw error measurements as a function of position along the X axis is shown in Figure 6.3.4. A similar setup, rotated 90^{o}, was used to measure the yaw error of the carriage as a function of position along the Z axis.

6.3.3 Straightness, Parallelism, and Orthogonality Measurements

Straightness is the translational error of the machine element that can occur in either of the two directions orthogonal to a slide's axis of motion. Although the laser interferometer system can be used for straightness measurements, the size of the optical attachments required and the minimum allowable distance between them limits its use in turning centers, which usually have short cross-slide travel ranges that are located in cramped quarters. Therefore, high-precision, noncontact capacitance probes and precision lapped test arbors were used for this group of measurements.

The setup for the measurement of X straightness of Z motion is shown in Figure 6.3.5. The principle of this measurement is to measure the change in the gap between the probes which are attached to the carriage while it is moving along the Z axis, and the test arbor mounted on the

[9] Pitch can cause a radial tool position error, but in this machine the error was inconsequential.

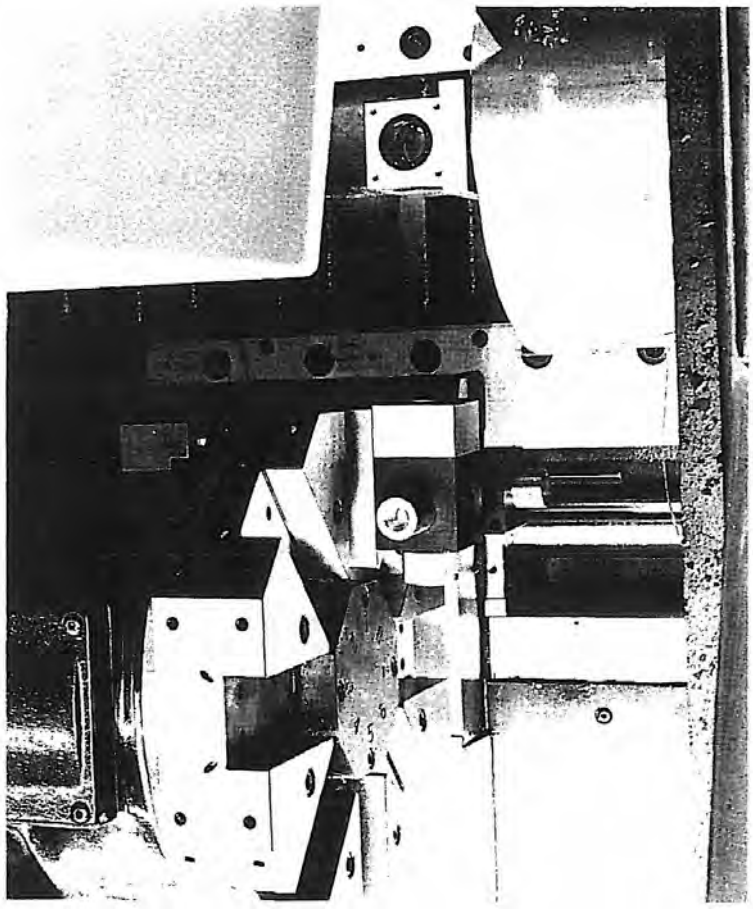


Figure 6.3.3 Laser interferometer setup for X axis displacement error measurements. (Courtesy of NIST.)

spindle. For this measurement, in addition to the axis straightness error, the probe outputs include nonstraightness of the test arbor profile and the misalignment between the arbor and the spindle. To eliminate the arbor profile error, the reversal technique[10] is used. The best-fit slope subtraction on the resultant data eliminates the misalignment errors.

As shown in Figure 6.3.6, in order to apply the reversal technique, two sets of measurements are required along the Z axis. In the first set, the carriage is moved along the Z axis and the probe outputs are recorded at every measuring interval. After the first set, the test arbor is rotated 180° and the test repeated. At any point along Z axis, the first measurement is represented by the following equation:

$$m_1(z) = a(z) - s(z) \tag{6.3.1}$$

where

- $m_1(z)$ is the output of probe l at position z.
- a(z) is the profile nonstraightness of the test arbor.
- s(z) is the straightness of Z motion.

Similarly, the second measurement is given as

$$m_2(z) = a(z) + s(z) \tag{6.3.2}$$

[10] Axis of Rotation: Methods for Specifying and Testing, ANSI Standard B89.3.4M-1985, American Society of Mechanical Engineers, United Engineering Center, 345 East 47th St., New York, NY 10017.

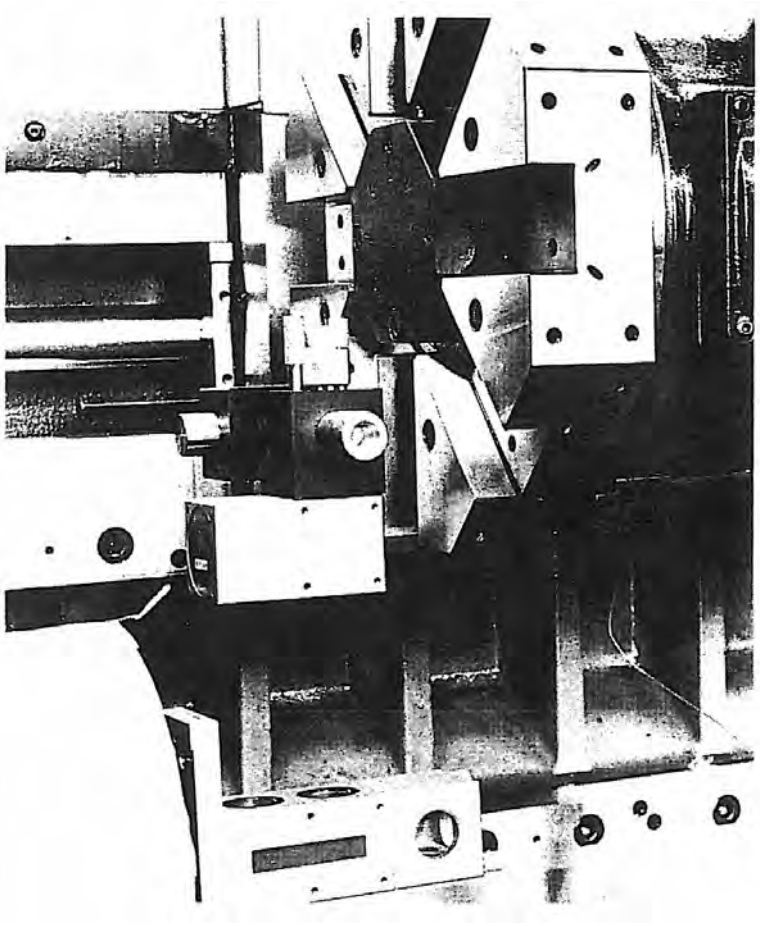


Figure 6.3.4 Laser interferometer setup for X axis yaw error measurements. (Courtesy of NIST.)

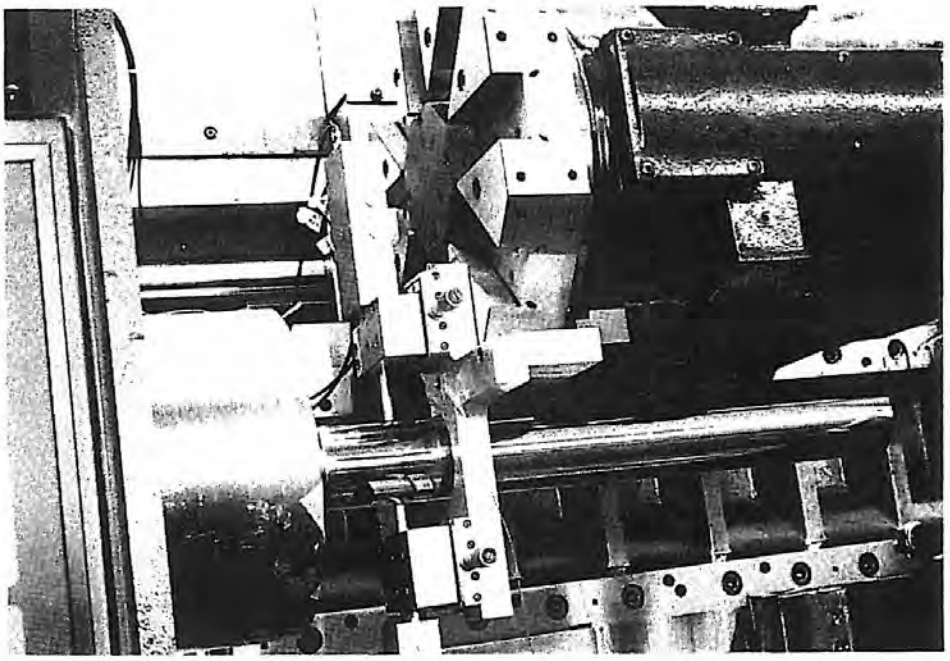


Figure 6.3.5 Setup for the X straightness of the Z motion and the parallelism measurements (probes are at the end of the Z travel). (Courtesy of NIST.)

where $m_2(z)$ is the output of probe 2 at position z. The straightness of the Z motion is thus

$$s(z) = \frac{-m_1(z) + m_2(z)}{2} \tag{6.3.3}$$

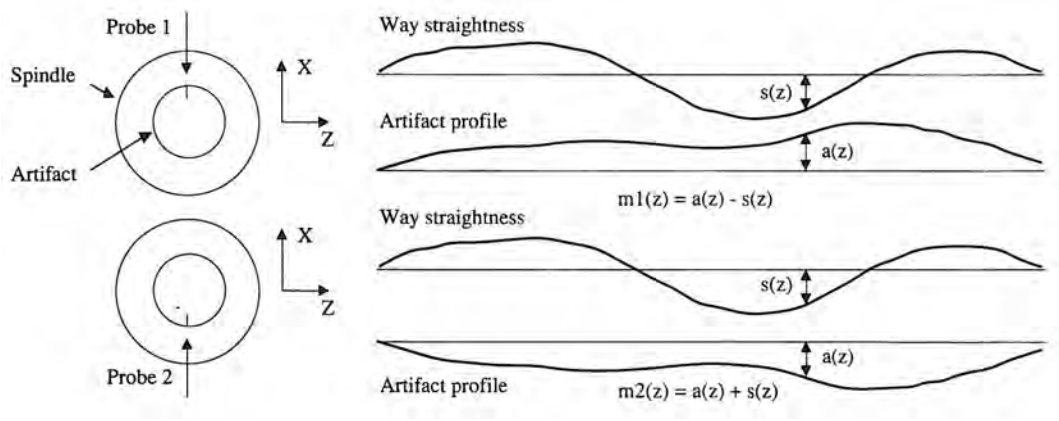


Figure 6.3.6 Reversal technique (artifact rotated 180° in spindle for second set of measurements) for calculating the X straightness of the Z motion. (Courtesy of NIST.)

Ideally, the parallelism between the axis of Z motion and the axis average line of the spindle can be determined by taking two measurements on the test arbor spaced a distance l apart. The parallelism error would be the difference divided by the length between these points. But as mentioned in the preceding section, there are several errors involved in such a measurement. Z axis straightness error, test arbor profile error, and the misalignment between the test arbor and the spindle are the major contributors, although a modified procedure can also be used here where the arbor is rotated 180°.

To measure parallelism without having other errors affect the measurements, two probes spaced 180° apart (either side of the artifact) were used. The spindle was rotated at a very low speed in order to minimize the spindle error motions, and measurements at 512 points in one revolution were recorded for each of the two probes at two locations, 12 in. apart, along the Z axis. It is known that the misalignment between a perfectly round artifact and a spindle with no error motion creates a limaçon as an output of such a measurement.[11] The artifact nonroundness and the spindle error distort this limaçon. In order to eliminate the effects of these errors on the data, a best-fit circle out of 512 data points for one revolution of the spindle must be calculated and used for the analysis. The calculation of best-fit circle radius R from the data is performed using the following formula.

$$R = \frac{\sum r_i}{n} \tag{6.3.4}$$

where r_i is the probe output at the angular position i and n is the number of data points.

To eliminate the axis straightness errors, these best-fit circles are constructed by the outputs of the two probes 180° apart (either side of the artifact). The parallelism error α_p is then calculated using the following formula:

$$\alpha_p = \frac{(R_{21} - R_{11}) - (R_{22} - R_{12})}{2\Delta z} \tag{6.3.5}$$

where

- R_{11} is the least square's radius from probe 1 at location 1.
- R_{21} is the least square's radius from probe 1 at location 2.
- R_{12} is the least square's radius from probe 2 at location 1.
- R_{22} is the least square's radius from probe 2 at location 2.
- Δz is the distance between location 1 and location 2.

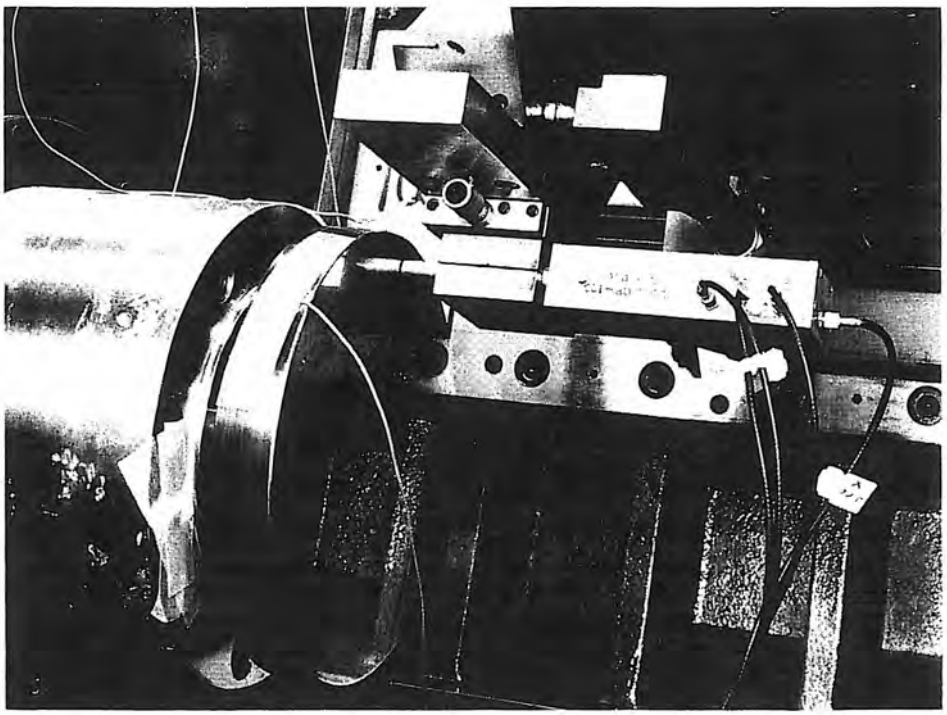


Figure 6.3.7 Setup for the Z straightness of the X motion and the orthogonality measurements. (Courtesy of NIST.)

For the Z direction straightness of the X motion and the orthogonality of the X axis to the axis average line of the spindle, another test arbor was used, as shown in Figure 6.3.7. The arbor was 7 in. in diameter and had a ground and lapped flat face which could be scanned. It was not possible to look at the back side of the artifact while it was mounted on the spindle. Hence it was impossible to apply the reversal technique to eliminate arbor profile errors from the straightness measurements; thus the flat face of this arbor had to be calibrated before its use. This was done on a Moore-3 coordinate measuring machine using the reversal technique, this time to eliminate machine straightness errors. It was found that the surface was flat to within 10 μin, which was within the required accuracy limits of the straightness measurements. Accordingly, there was no need for surface calibration. This arbor was then mounted on the spindle and a capacitance probe was mounted on the turret to make the measurements.

In order to eliminate misalignment and arbor squareness errors, the following method was used: While the cross-slide was moving along the face of the test arbor, the capacitance probe readings were taken at every measuring interval. Then, the spindle was rotated 180° and the measurement repeated with the same probe. With this procedure, it was possible to find the orthogonality as well as the Z straightness of the X motion. The following relationships[12] are obtained from the geometry shown in Figure 6.3.8.

$$m_1(x) = -\delta_z(x) + e(x) \tag{6.3.6}$$

$$m_2(x) = -\delta_z(x) - e(x) \tag{6.3.7}$$

where

- $m_1(x)$ is the probe output for the first set of measurements.
- $m_2(x)$ is the probe output after the spindle rotated 180°.
- $\delta_z(x)$ is the Z straightness of X motion.
- $e(x)$ is the combination of arbor squareness and misalignment errors.

[11] Axes of Rotation, ANSI Standard B89.3.4-1985.
[12] Assumes perfect artifact profile.

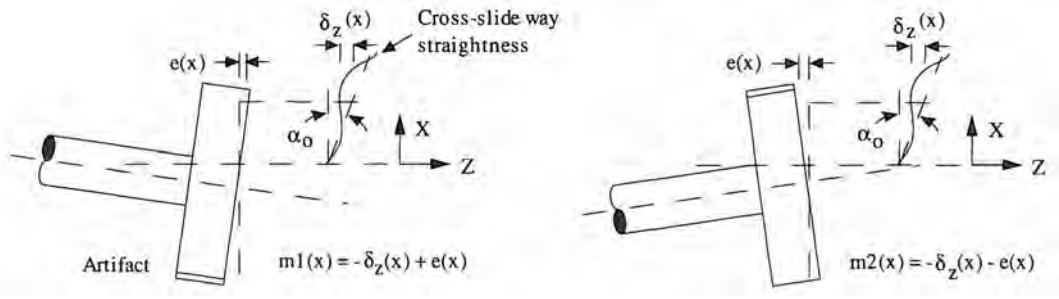


Figure 6.3.8 Reversal technique (spindle rotated 180° for second set of measurements) for calculating the Z straightness of the X motion. (Courtesy of NIST.)

From Equations 6.3.6 and 6.3.7 the straightness is calculated:

$$\delta_z(x) = \frac{-m_1(x) - m_2(x)}{2} \tag{6.3.8}$$

The orthogonality can then be calculated as the best-fit line slope from the data obtained by averaging respective sets of measurements M_1 and M_2.

6.3.4 Spindle Thermal Drift

Thermal drift is defined in the ANSI Standard B89.6.2 as "a changing distance between two objects, associated with a changing temperature distribution within the structural loop due to internal and external sources." As in the case of spindle error motions, three components of spindle thermal drift are critical to the overall performance of the machine tool: (1) axial thermal drift, which is the displacement of the spindle along the Z axis; (2) radial thermal drift, which is the displacement perpendicular to the Z axis in the sensitive direction; and (3) tilt thermal drift, which is the rotation of the spindle in the X-Z plane of the machine.

The equipment used for the measurements of these errors consisted of a data acquisition system, capacitance probes, and precision ground test arbors. In order to eliminate possible error sources in the measurements such as the carriage and cross-slide displacements due to thermal effects and to minimize thermal errors in the setup, the fixture which held the probes was mounted on the machine bed as close to the spindle as was possible and measurements made without other axes turned on.

The conventional way to investigate thermal drift characteristics is to find the changes in the position and orientation of the spindle with time. Using time as an independent variable in the relations complicates the modeling for the prediction of the drift, due to the fact that it requires knowledge of the time history of the operations. In this investigation, a special effort was made to avoid using a time factor in the relationships; therefore, correlations found depend only on physical parameters measured during the operation, such as spindle speed and temperatures of various locations around the spindle.

In order to find the locations where changes in temperature correspond to spindle drift, temperatures were simultaneously monitored at numerous locations around the headstock during the spindle drift measurements, including: (1) the spindle bearings on both ends; (2) the spindle face, the four corners of the headstock base; (3) a location on the bed between the headstock and the fixture; (4) the ambient air; and (5) the probe fixture. During a typical measurement cycle, the spindle was run at a constant speed for 8 hours, and every 10 minutes, temperature and radial and axial measurements were sampled. In order to eliminate test arbor roundness error and the fundamental spindle error motion, the probe reading was always sampled at the same angular position of the spindle. After 8 hours of running time, the machine was turned off and allowed to cool down overnight, while the temperature and the probe sampling continued. This procedure was repeated for different spindle speeds. The amount of tilt drift at any particular time was obtained by dividing the difference between the two radial displacements measured by the two probes by the distance between the two probes.

6.4 CALIBRATION MEASUREMENT RESULTS

The measurement techniques discussed were used with consideration of the geometry and construction of the machine. For example, the lead of the ballscrews used on the Superslant® was 0.2 in, and the usable travel range of the carriage, after the laser optics were attached to the spindle and the carriage, was 13 in. The usable travel range of the cross-slide was 3.4 in. Since it was necessary to obtain enough points along each axis for statistical analyses, the measuring interval was selected to be 1 in. on the carriage motion (Z axis), and 0.2 in. on the cross-slide motion (X axis). Since the measuring intervals selected were even multiples of the ballscrew lead, periodic errors generated by the lead of the ballscrew were measured separately for overtravel ranges of 0.4 in. along each axis, with measuring intervals of 0.002 in. For reference, the machine zero point was at the spindle nose. The points farthest from the zero point on the X and Z axes, x = 4.75 in. and z = 16.5 in., were called "home" positions.

Results of calibration measurements of the following components are discussed in this section:

- Carriage (Z) linear displacement error
- Carriage yaw error
- X straightness of the Z motion
- Spindle radial and tilt thermal drift
- Spindle axial thermal drift

Carriage (Z) Linear Displacement Error

Typical displacement error data taken along the Z axis during the measurement period are shown in Figure 6.4.1. The initial observation from this figure indicates that there is about 200 μin of backlash[13] in the leadscrew despite the use of a preloaded ball nut/leadscrew assembly. This error varies along the length of the leadscrew due to the nonlinearities of the leadscrew and is typical for a large ballscrew used in machine tools. Therefore, different error calibrations for forward and reverse directions would be required to compensate for the backlash. Initially, measurements showed that the home position also drifted randomly, not as a function of temperature, with time, due to instability of the home position limit switch. To alleviate this situation, it was decided that the next point in the data set (which corresponds to the error at the nominal position of 15.5 in.), would be taken as a reference point, and the complete data set would be normalized with respect to this point. This procedure would require the determination of the absolute position of this reference point during machining using a tool-setting station, as discussed in Section 6.5.

The next observation was the influence of the thermal state of the machine on the geometric errors. As the machine warmed up, the slope of the error curve changed significantly, as shown in Figure 6.4.1. The temperature profiles of various points around the carriage during the period of these measurements are shown in Figures 6.4.2 and 6.4.3 and appear to have similar profiles. The best locations (most sensitive) for monitoring temperature to determine its effect on carriage (Z) linear displacement error were the bearing housings at each end of the ballscrew, and the ballnut assembly on the carriage.

Initially a two-variable nonlinear least square's regression analysis was used to obtain the best curve fits for the data. The first independent variable was the nominal carriage position. As second variables, nut temperature, bearing housing temperature, and slideway temperature were tried; however, when incorporated into the regression analysis, none gave acceptable standard deviations for the estimated displacement error.

Finally, error behavior at every individual measurement point (1 in. apart on the Z axis and 0.2 in. apart on the X axis) was analyzed with respect to temperatures of the previously selected locations. For this purpose, a single-variable nonlinear least squares curve-fitting technique was used, taking the temperature as the only variable. This type of curve fitting was done for each nominal position and for both travel directions. Also, to select the best representative point for the thermal effects, the same type of analysis was carried out on the temperatures of varying locations. The curve fit of displacement error with respect to temperature at the right-end bearing housing of

[13] This backlash, or lost motion, is typical for machines with sliding contact bearings because the breakaway friction force may be on the order of 1000 N. As the force is reversed in the ballscrew, the loading of the balls changes drastically and the resulting deflection is very nonlinear as predicted by the Hertz equations discussed in Section 5.6.

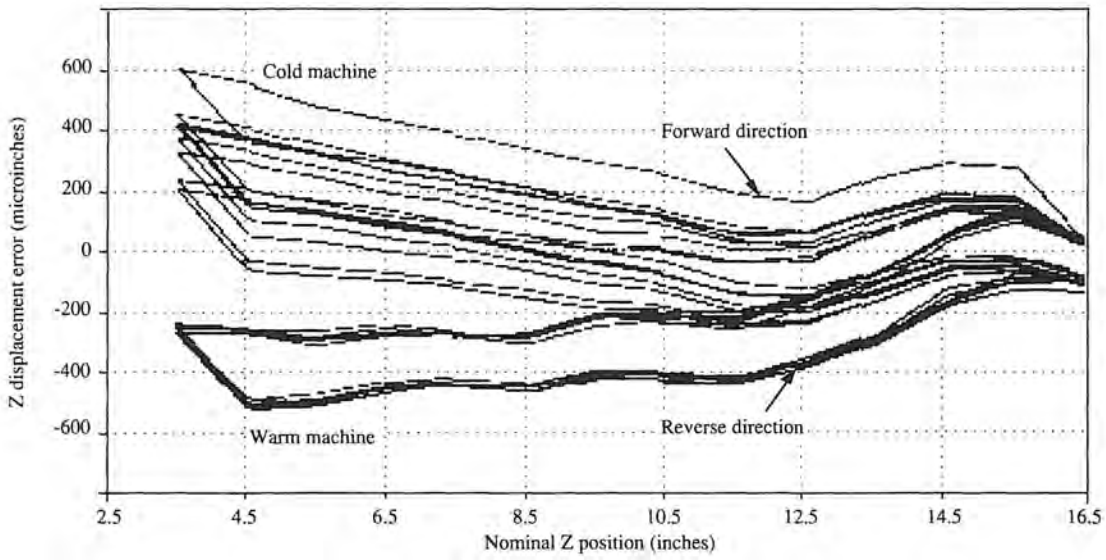


Figure 6.4.1 Z displacement error measurement raw data taken as the machine warms up. (Courtesy of NIST.)

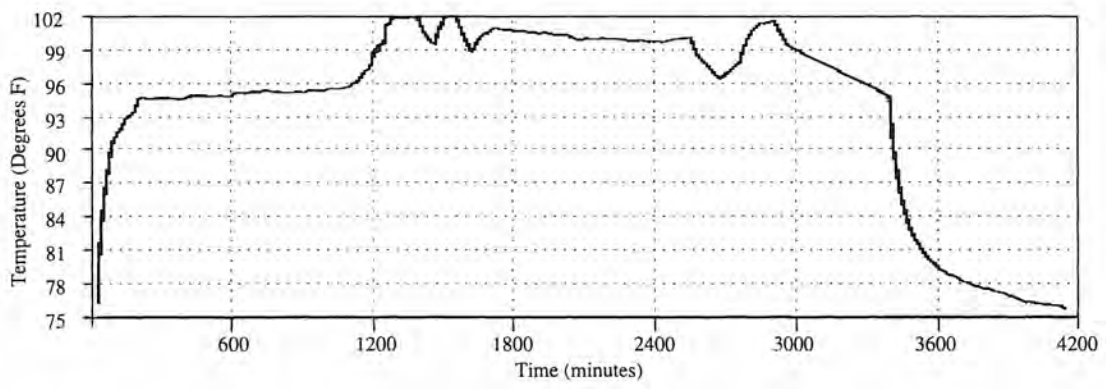


Figure 6.4.2 Z-axis drive motor temperature profile during the Z displacement error measurement period. (Courtesy of NIST.)

the ballscrew gave the smallest standard deviations. Figure 6.4.4 shows an example of the curve-fitting results. For the reverse direction, a very different-shaped curve results. The relationships between the carriage Z displacement error $\delta_z(z)$ and the temperature T of the rear bearing housing of the carriage ball screw for both forward and reverse directions of the carriage motion were obtained from a least square's curve fitting:

$$\delta_z(z) = a_0 + a_1 T + a_2 T^2 + a_3 T^3 + a_4 T^4 \tag{6.4.1}$$

Twelve sets of five coefficients each were thus required in order to map the linear displacement performance of the carriage. However, in order to predict errors between these nominal positions, an interpolation scheme had to be developed. In order to obtain an accurate Z position for use in this equation, a prediction of the periodic displacement error is required.[14] This expectation was confirmed by measurements at 0.02-in. intervals, as shown in Figure 6.4.5.

In order to determine the periodic error without thermal contamination of the data, the measurements were taken over a very short (0.4 in.) travel range with the assumption of uniformity over the whole travel range. The difference between the incremental motion as measured by the laser

[14] The use of a ballscrew-nut drive system and resolver feedback unit creates a periodic type of displacement error.

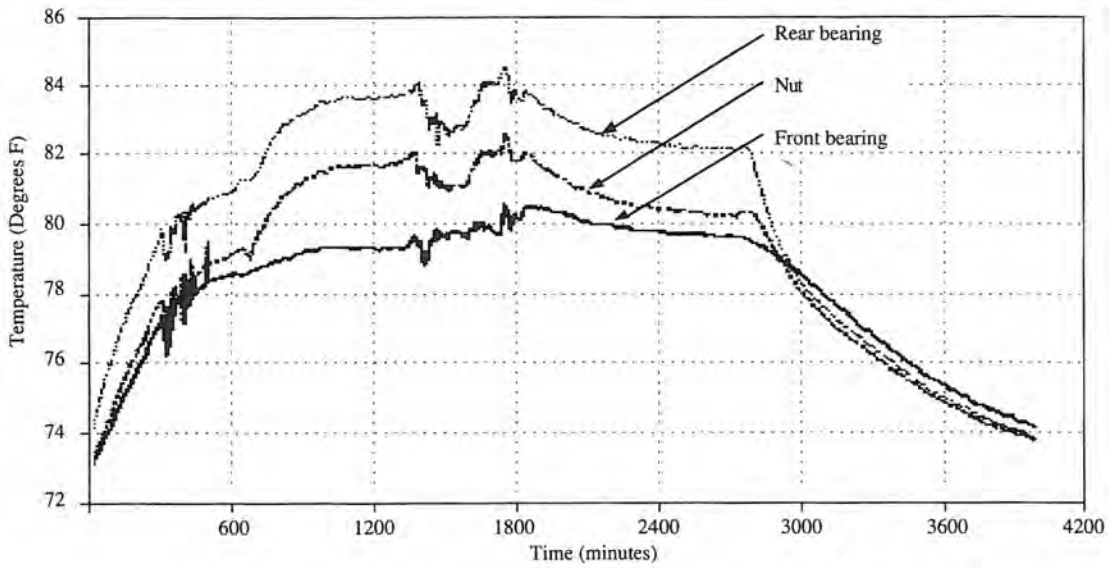


Figure 6.4.3 Temperature profiles of the carriage leadscrew nut and the bearing during the Z displacement error measurements. (Courtesy of NIST.)

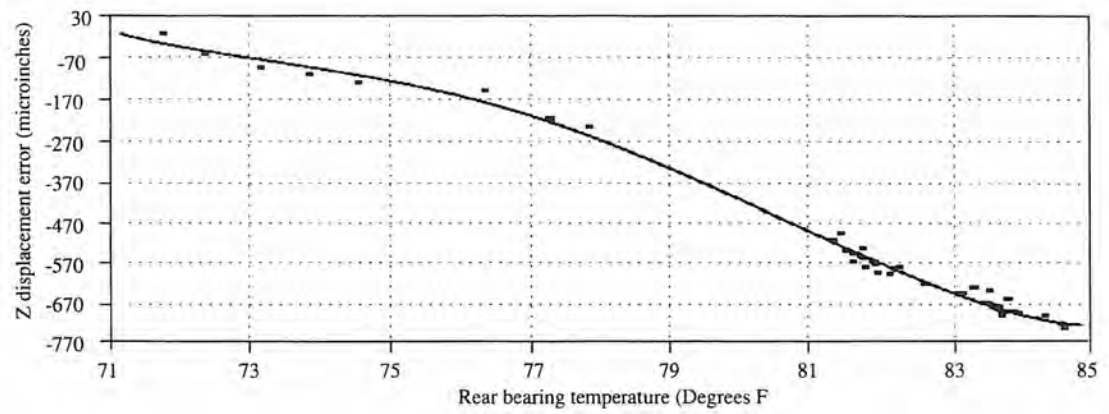


Figure 6.4.4 Typical Z displacement error data (forward direction) at a nominal fixed Z position and varying temperature. (Courtesy of NIST.)

interferometer and given by the machine tool controller is the net periodic error motion. The net error motion data for one ballscrew revolution and the fitted curve corresponding to this data are shown in Figure 6.4.6. As is seen from this figure, there are two components of the periodic error. The first one has a period of one revolution of the ballscrew, which equals 0.20 in.. The other component is a one-tenth of a revolution periodic error caused by the resolver unit, which is attached to the ballscrew through a gear drive with a ratio of 1:10. Because of the timing limitations of the compensation system[15] and the fact that the 10th harmonic of the periodic error only has an order of magnitude of 20 μin, the 10th harmonic error was disregarded. The resulting statistical analysis gave the following equations. For the forward direction of the Z periodic error:

$$\delta'_z(z) = 3.194 + 0.164\cos(31.4159z) + 3.542\sin(31.4159z) - 301.632z + 1693.67z^2 \quad (6.4.2)$$

For the reverse direction of the Z periodic error:

$$\delta'_z(z) = 15.136 - 6.453\cos(31.4159z) + 4.423\sin(31.4159z) + 1209.1z + 5953.20z^2 \quad (6.4.3)$$

where

[15] The compensation system was two machine servo cycles lag between the nominal position and the corresponding error correction.

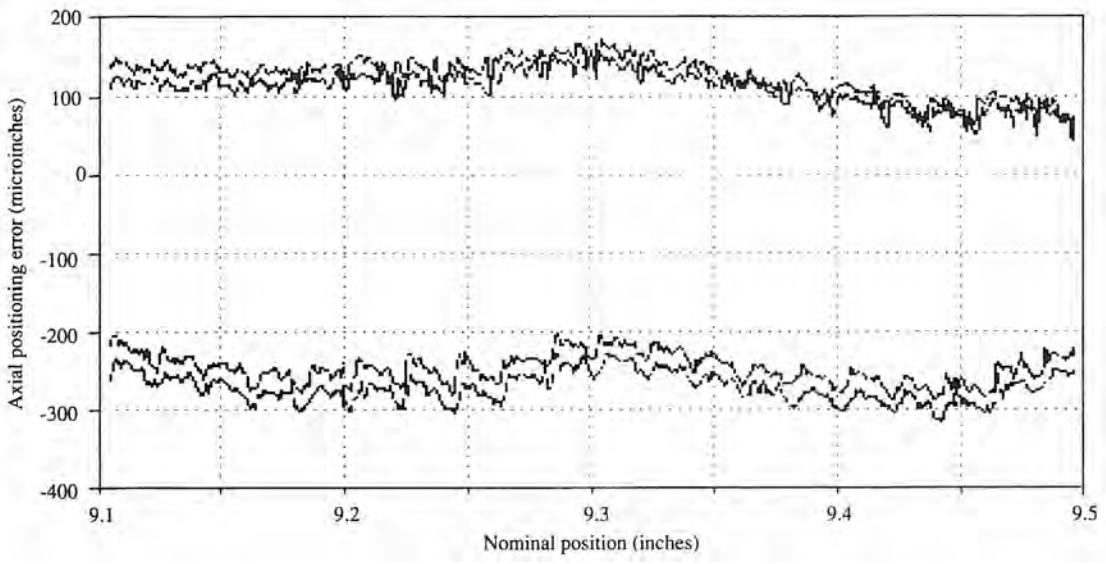


Figure 6.4.5 Typical periodic Z displacement error raw data. (Courtesy of NIST.)

- δ_z' is the periodic ball screw error for Z motion, and
- z is the incremental nominal displacement in Z direction.

Using Equations 6.4.2 and 6.4.3, a sinusoidal interpolation procedure based on superposition can be applied to find the periodic error at any point. When combined with the thermal error, the total linear displacement error for the carriage is obtained. A similar procedure was performed for the cross-slide (X direction) linear displacement error, and similar results were found. Only different coefficients resulted from the curve fitting of the data.

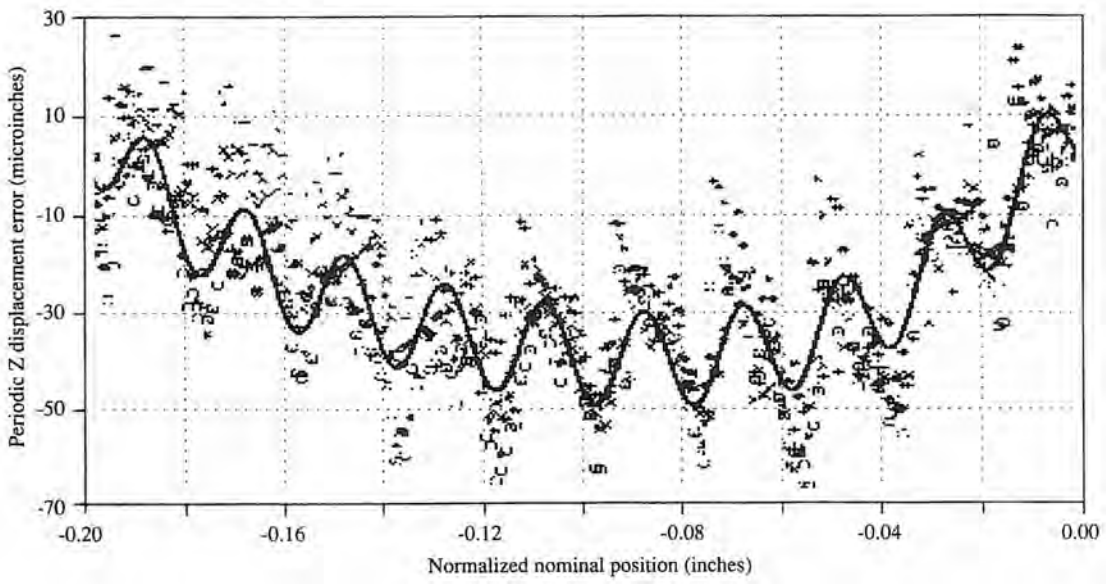


Figure 6.4.6 Net periodic Z displacement error (reverse direction) for one ballscrew revolution (including the 10th harmonic). Four sets of data were taken and then the curve was fitted. (Courtesy of NIST.)

Carriage Yaw Error

The measurement of yaw error was done in two stages: In the first stage, the yaw error of carriage-cross-slide assembly was measured when the machine was at its home position, over a period in which the machine gradually warmed up from the cold state. The relationship between this error and the temperature T on the carriage body is shown in Figure 6.4.7, and is described by the following equation:

$$\begin{aligned}\varepsilon_y(x+z) &= 24054.7 - 1136.62T + 18.5346T^2 - 0.100344T^3 \\ &\quad -0.235989 \times 10^{-3}T^4 + 0.283955 \times 10^{-5}T^5\end{aligned} \tag{6.4.4}$$

In the second stage, the yaw error was measured as a function of position as the cross-slide and the carriage moved away from their corresponding home positions.

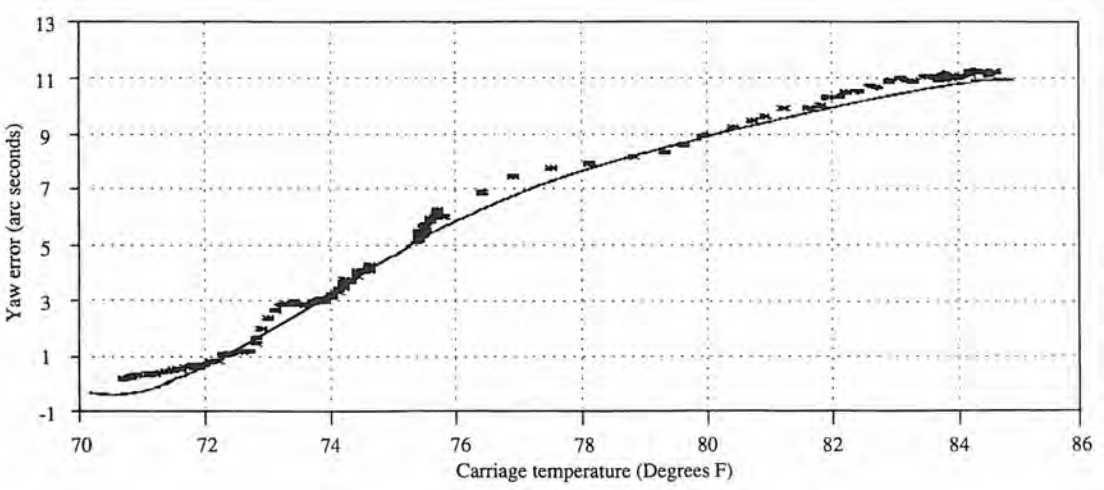


Figure 6.4.7 Yaw error of the carriage-cross-slide assembly as the machine warms up. (Courtesy of NIST.)

Figure 6.4.8 shows a sample of the yaw data taken over a period of 2 days. Since there is a difference in forward and reverse directions, they must be analyzed separately. The change in the yaw error of any point along the axis, relative to the home position, is not significant (total spread is about 1 arcsecond and the distance from the yaw axis is about 1 in.) as the machine warms up. In other words, the effect of temperature on yaw error is constant along the length of travel of the carriage. So a regression analysis with a single variable (nominal position z) is sufficient to find the yaw error model. For the forward direction:

$$\varepsilon_y(z) = 15.0864 - 1.63607z + 0.109098z^2 - 0.003966z^3 \tag{6.4.5}$$

For the reverse direction:

$$\varepsilon_y(z) = 16.2995 - 1.81443z + 0.127474z^2 - 0.004663z^3 \tag{6.4.6}$$

A similar procedure was used for the cross-slide.

X Straightness of the Z Motion

A sample of the raw data from two probes is shown in Figure 6.4.9. From this data, the X direction straightness of the carriage as a function of the Z position is calculated using Equation 6.3.3, and then linear regression analysis is carried out to find the best fit line. Figure 6.4.10 shows the straightness error results after the best-fit line slope is subtracted for the forward direction. It can be seen from these figures that the machine's thermal state does not significantly alter the straightness characteristics; thus, the average values for each direction are used as the straightness curve. Due to the irregularities in the shape of the curve, least square's curve-fitting attempts do not give satisfactory correlations. Therefore, a look-up table with linear interpolation between data points should be used. A similar procedure was used for the Z straightness of the X motion of the cross-slide.

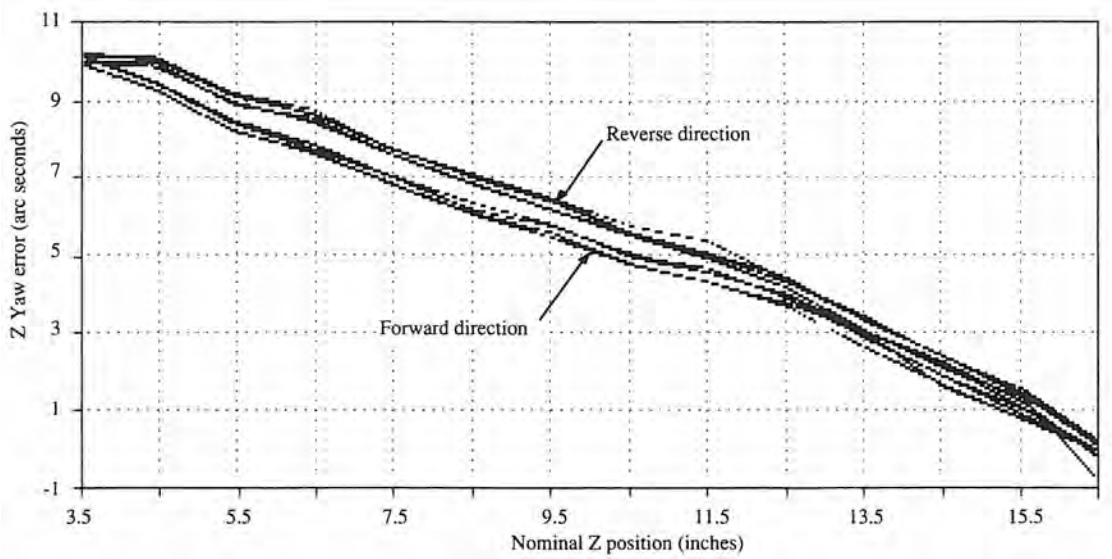


Figure 6.4.8 Carriage yaw error raw data sample. (Courtesy of NIST.)

In order to determine the parallelism error between the carriage motion and the axis average line of the spindle, the best-fit circles were calculated using Equation 6.3.4 on the data measured by two probes 180° apart from each other. The results of the calculations based on the data taken over a 2-day machine warm-up period gave a spread of less than 1 arcsecond. With such a small amount of spread, thermal effects on this error are considered to be insignificant, and a constant average value of -14 μrad parallelism error was used in the error compensation scheme.

The orthogonality between the cross-slide and spindle axis was calculated as the best-fit line slope from the data obtained by Equation 6.3.8, as shown in Figure 6.4.11. The relationship between orthogonality α_o and the cross-slide body temperature T was obtained from this analysis:

$$\alpha_o = 345.098 - 7.33739T + 0.051235T^2 \tag{6.4.7}$$

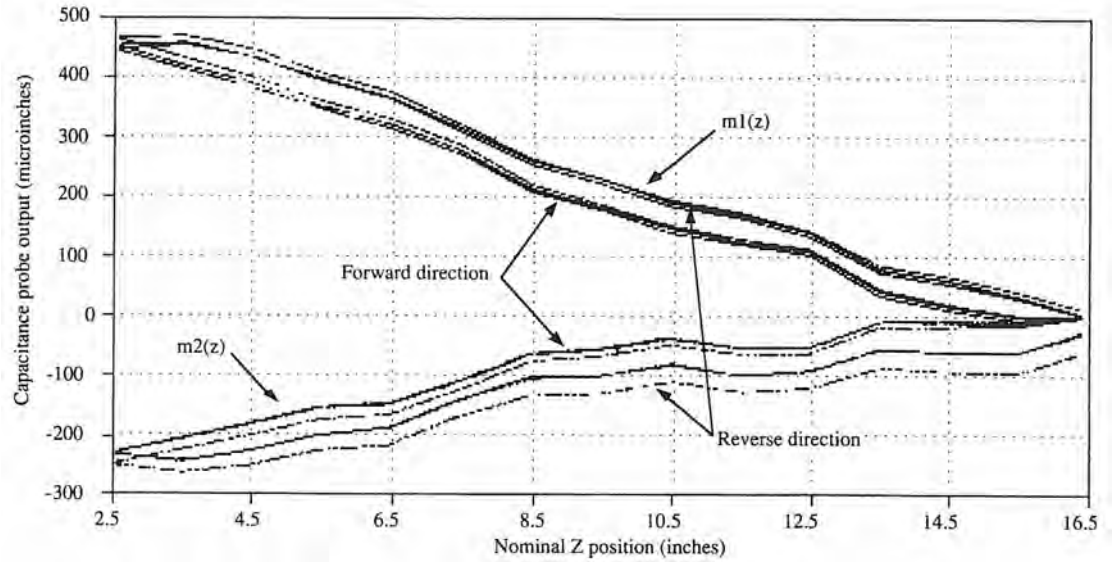


Figure 6.4.9 X straightness of the Z motion sample raw data. (Courtesy of NIST.)

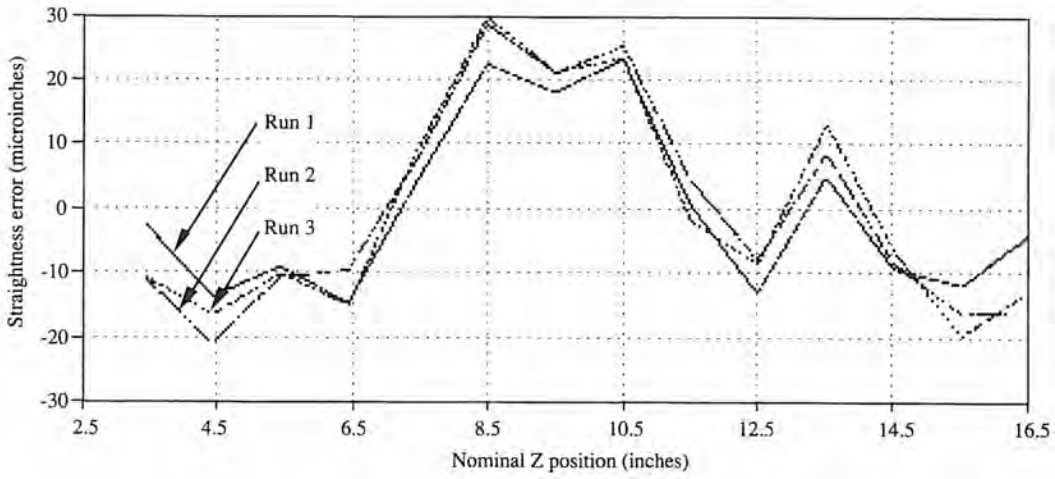


Figure 6.4.10 Calculated X straightness of the Z motion data for the reverse direction of the Z motion. (Courtesy of NIST.)

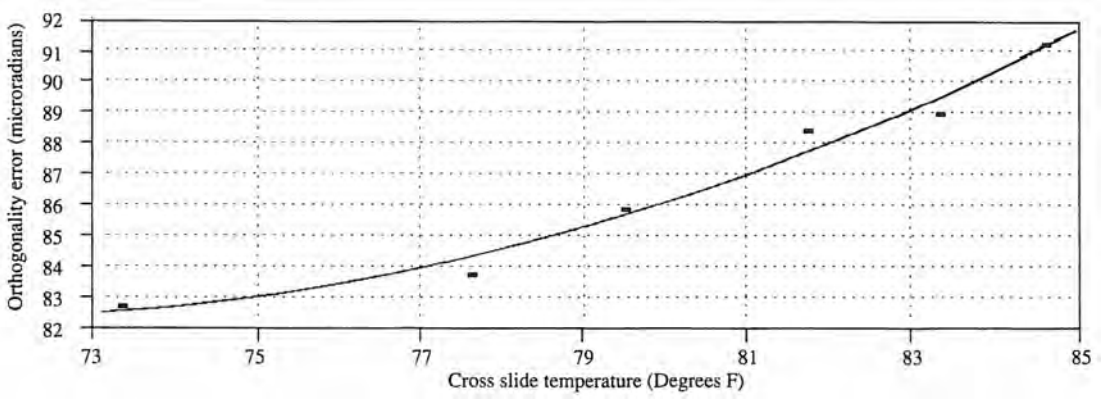


Figure 6.4.11 Orthogonality error between the X motion and the axis average line of spindle rotation as the machine warms up. (Courtesy of NIST.)

Spindle Radial and Tilt Thermal Drift

The data used for calculating the radial and tilt drift of the spindle was obtained using two probes mounted 8 in. apart along the test arbor. The difference between the radial displacements measured by the two probes divided by the distance between them give the amount of tilt at any time. Using this value, the pure radial displacement at the spindle nose was also calculated. Figure 6.4.12 shows sample data obtained during these measurements. It should be noted that the noise in the data is caused by the random component of the spindle error motion. Since this motion is amplified by the tilt motion, the data from the probe which is at the unsupported end of the arbor is noisier. Spindle tilt data is shown in Figure 6.4.13. The maximum tilt error is about 5 arcseconds, and the time to reach the equilibrium is about 3 h. The seven-point running averages technique was used[16] to make the data smoother for analysis. With this technique, each data point is replaced by the average of the data corresponding to three points before and after it.

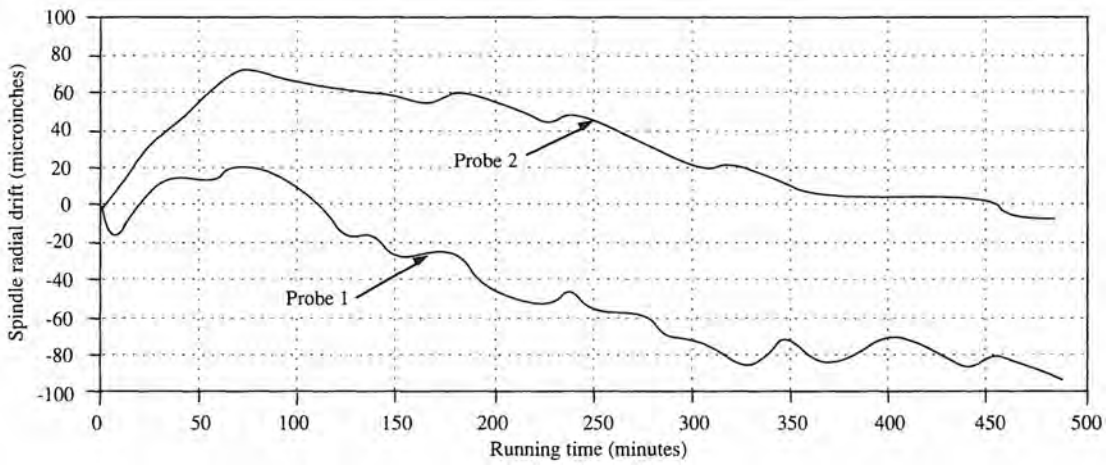


Figure 6.4.12 Spindle radial drift when the spindle is turning at 2000 rpm. (Courtesy of NIST.)

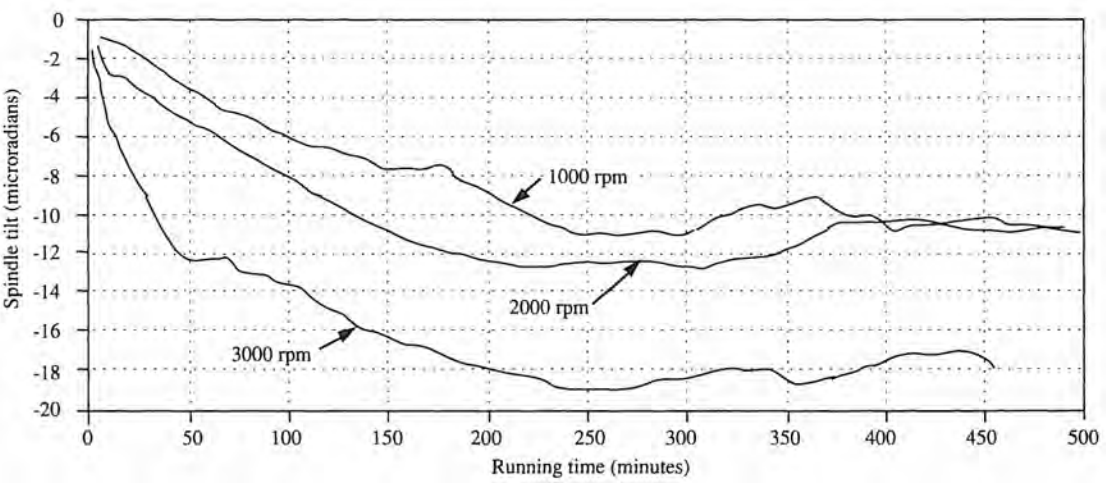


Figure 6.4.13 Spindle tilt drift calculated from radial drift data. (Courtesy of NIST.)

The relationships between the tilt motion and the various combinations of temperatures were then investigated. The temperature profiles of typical locations around the headstock during this period are shown in Figure 6.4.14. The spindle tilt thermal drift behavior can be represented by an equation that is valid for all spindle speeds, and only the temperature rise/drop in the rear spindle bearing (ΔT) is necessary for prediction of the tilt value $\varepsilon_y(s)$:

[16] A. Savitsky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least-Square's Procedures," Anal. Chem., July 1964.

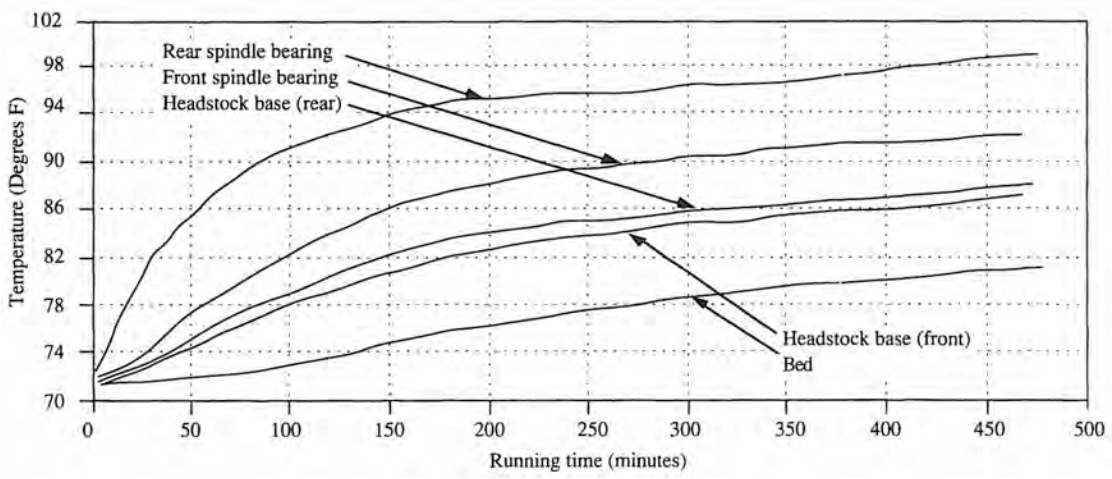


Figure 6.4.14 Temperature profiles during spindle drift measurements at 2000 rpm. At 4000 rpm, the upper bound approaches 108°F. (Courtesy of NIST.)

$$\varepsilon_y(s) = 1.67831 - 0.168393\Delta T - 0.0206067\Delta T^2 + 0.259273 \times 10^{-3}\Delta T^3 \tag{6.4.8}$$

The radial spindle thermal drift has a more complicated behavior. In addition to temperature, spindle speed also significantly affects radial drift. It was also discovered that the radial errors were different during warm-up and cooling cycles. The following equations show the radial spindle thermal drift during warm-up and cooling periods. For the warm-up period:

$$\begin{aligned}\delta_x(s) \quad = \quad & -16.1931 + 3.00839(\Delta T_4 - \Delta T_1 + \Delta T_5) + 24750.7/\omega \\ & +1.55985(\Delta T_4 - \Delta T_1)\Delta T_5 + 10015.2(\Delta T_4 - \Delta T_1)/\omega \\ & -0.291304(\Delta T_4 - \Delta T_1)\Delta T_5^2 + 0.130897 \times 10^{-3}(\Delta T_4 - \Delta T_1)\Delta T_5\omega \\ & +0.538512 \times 10^{-4}(\Delta T_4 - \Delta T_1)^2\Delta T_5\omega\end{aligned} \tag{6.4.9}$$

For the cooling period:

$$\begin{aligned}\delta_x(s) \quad = \quad & 39.959 - 21.2883(\Delta T_4 - \Delta T_1 + \Delta T_5) - 0.012886\omega \\ & -7.29782(\Delta T_4 - \Delta T_1)\Delta T_5 + 59019.9(\Delta T_4 - \Delta T_1)/\omega \\ & -0.293672(\Delta T_4 - \Delta T_1)\Delta T_5^2 + 224.213(\Delta T_4 - \Delta T_1)\Delta T_5^2\omega \\ & -51018.8(\Delta T_4 - \Delta T_1)\Delta T_5/\omega - 9164.44(\Delta T_4 - \Delta T_1)^2\Delta T_5/\omega\end{aligned} \tag{6.4.10}$$

where

- $\delta_x(s)$ is the radial spindle drift.
- ΔT_4 is the temperature change in the rear bearing of spindle.
- ΔT_1 is the temperature change in the front bearing of the spindle.
- ΔT_5 is the temperature change in the bed.
- ω is the spindle speed (rpm).

These are not simple equations, and although each machine tool will be different, the methods used to measure and analyze the data will be the same.

Spindle Axial Thermal Drift

The axial drift of the spindle was measured for different speeds during warm-up and cooling cycles and typical data is shown in Figure 6.4.15. The initial investigation suggested that there was a strong relationship between the axial spindle displacement, the average temperature along the spindle, and the temperature at the location where the sensing fixture was mounted. The difference between the measured and calculated drift at a particular spindle speed was found to be only about 20 μin based on the following equation:

$$\delta_z(s) = \alpha[L_1(\Delta T_1 + \Delta T_4)/2 - (L_1 + L_2)\Delta T_5] \tag{6.4.11}$$

where

- $\delta_z(s)$ is the axial spindle drift.
- α is the coefficient of thermal expansion of cast iron.
- T_1 is the rear spindle bearing temperature.
- T_4 is the front spindle bearing temperature.
- T_5 is the temperature on the bed where the fixture is mounted.
- L_1 is the length of the headstock.
- L_2 is the distance between the headstock and the fixture.

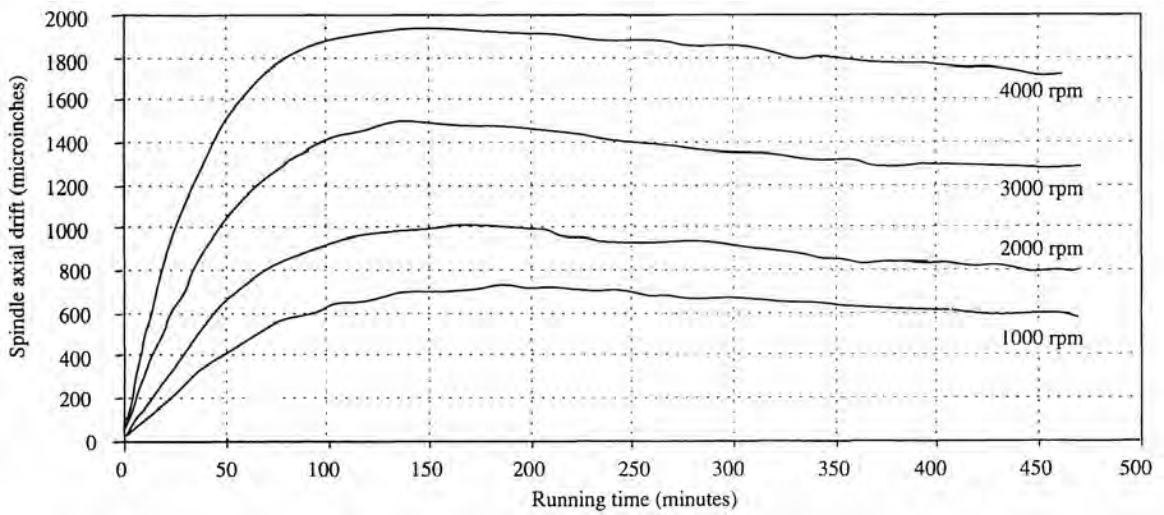


Figure 6.4.15 Spindle axial drift data for different spindle speeds. (Courtesy of NIST).

Based on this investigation, a multivariable polynomial regression analysis was carried out, and the following relationships were obtained. For the warm-up period:

$$\begin{aligned}\delta_z(s) &= 32.5964(\Delta T_1 + \Delta T_4) - 89.4055\Delta T_5 + 0.0022638\omega(\Delta T_1 + \Delta T_4) \\ &\quad -0.153 \times 10^{-7}\omega^2(\Delta T_1 + \Delta T_4)\Delta T_5\end{aligned} \tag{6.4.12}$$

For the cooling period:

$$\begin{aligned}\delta_z(s) &= 22.6971(\Delta T_1 + \Delta T_4) - 56.3001\Delta T_5 + 0.0038758\omega(\Delta T_1 + \Delta T_4) \\ &\quad +0.3 \times 10^{-7}\omega^2(\Delta T_1 + \Delta T_4)\Delta T_5\end{aligned} \tag{6.4.13}$$

These equations show that the spindle drift characteristics are more complex than any other error components measured. In addition, in order to predict axial and radial spindle drift, any compensation system has to keep track of recent thermal history to determine whether the machine is in a warm-up or cooling cycle. Fortunately, there is a way to avoid these complicated decision-making and calculation procedures by locating a tool-setting station in a proper location on the spindle. This approach will be explained in the next section.

6.5 Compensating for the Measured Errors

The regression models of spindle radial and axial thermal drift are difficult to evaluate and incorporate into the resultant error expression given by Equations 6.2.15 and 6.2.17. Also, the cutting tool error terms $\delta_x(c)$ and $\delta_z(c)$ have to be determined. These terms include thermal expansion and the static force deflection of the cutting tool, fixturing errors of the tool holder and the insert, and tool wear. Among these, thermal expansion and static deflections can be found analytically if sufficient information is known about the temperature profile and the static force. However, there is no satisfactory model available for determination of the amount of the tool wear at any given time. In addition, due to the nature of the errors involved in fixturing and replacing of tooling and cutting

inserts, it is not possible to predict the terms $\delta_x(c)$ and $\delta_z(c)$ with sufficient accuracy. Therefore, the method of using a *tool-setting station* was adopted in this study.

The Tool-Setting Station

The tool-setting station is illustrated in Figure 6.5.1. Its principal measuring element is a linear variable differential transformer (LVDT). The tool-setting station measures the position of tooling or a reference bar with respect to the machine reference point along both the X and the Z axes of the turning center. The tool must touch the tool-setting station for measurement, but must be retracted during cutting.[17] The location of the tool-setting station is selected such that it can simplify the resultant error equations, 6.2.15 and 6.2.17. For example, the measurements of a tool-setting station mounted on the spindle include the radial and the axial spindle thermal drift errors $\delta_x(s)$ and $\delta_z(s)$. In addition, any measurement, regardless of the location of the tool-setting station, includes the combined effects of the tool and turret errors. Therefore, the tool-setting station output in the X direction replaces the expression $-x_4 - \delta_x(c) - \varepsilon_y(t)z_4 - \delta_x(t)$ in Equation 6.2.15, and the tool-setting station output in the Z direction replaces the expression $\varepsilon_y(t)x_4 - z_4 - \delta_z(c) - \delta_z(t)$ in Equation 6.2.17.

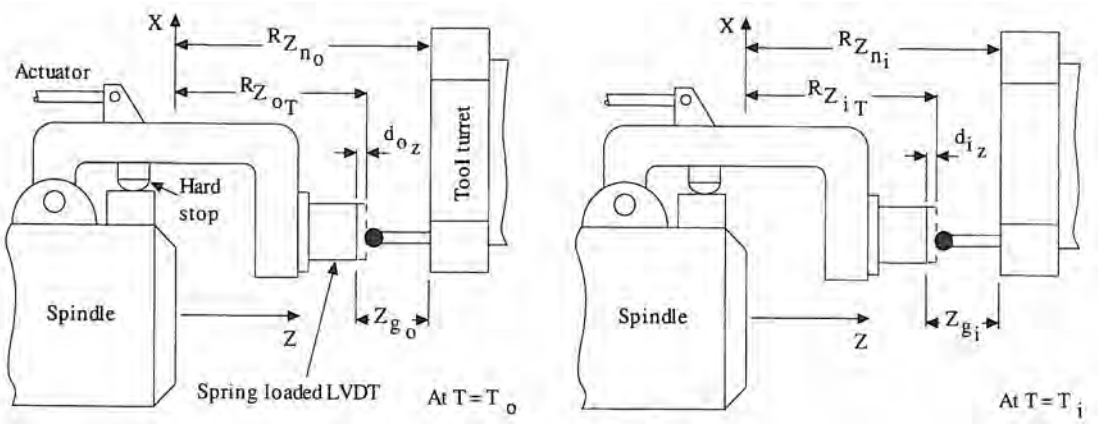


Figure 6.5.1 Tool-setting station geometry. (Courtesy of NIST.)

In the preceding section, the models for linear displacement errors along both axes of motion were established based on the data normalized with respect to the error at the reference points. This normalization was necessary because of the unpredictable variations of the reference locations. Therefore, in order to predict the linear displacement error at any nominal position in addition to the error value obtained from the model, the current location of the reference point must be known. The reference point can be determined by measurements from the tool-setting station with a gage bar permanently mounted on one of the tool stations on the turret. These measurements are used to compensate for the reference point migration. The details of the machine reference offset and the tool dimension offset calculations are given in the following section.

Obtaining the Machine Reference and Tool Dimension Offsets

Since there are errors involved in moving the axes to make contact with the tool-setting station, the information obtained from the measurements cannot be used directly as machine reference offsets. Instead, this measurement data has to be treated to account for the linear displacement errors of the axes and the thermal expansion of the tool-setting station itself. The procedure for the Z axis is given as an example.

Referring to Figure 6.5.1, assume that the initial tool-setting station Z location in the reference coordinate frame at temperature T_0 is given by $^{R}Z_{0T}$. At this thermal condition, a gage bar of length Z_{g0} mounted on the tool turret touches the tool-setting station, causing a reading d_{0z}. Also, let the machine nominal position at this instant be Z_{n0}. Later, when the machine is warmed up to a temperature of T_i, the same type of measurement is performed. The data corresponding to the tool-setting station position, the nominal axis position, and the tool-setting station output are $^{R}Z_{it}$, Z_{ni}, and

[17] The repeatability of the tool-setting station must be better than the machine itself, and thus should be on the order of 10 μin.

d_{iz}, respectively. Assuming that the current length of the gage bar is Z_{gi}, the following expressions can be written about both cases:

$$\text{At } T = T_0, \quad {}^{R}Z_{0_T} - d_{0_z} = {}^{R}Z_{n_0} - Z_{g_0} \tag{6.5.1}$$

$$\text{At } T = T_1, \quad {}^{R}Z_{i_T} - d_{i_z} = {}^{R}Z_{n_i} - Z_{g_i} \tag{6.5.2}$$

where ${}^{R}Z_{n0}$ and ${}^{R}Z_{ni}$ are the machine slide positions in reference coordinates at temperature fields of T_0 and T_i, respectively. For any temperature T, the machine axes' position in reference coordinates ${}^{R}Z_n$, can be expressed as

$${}^{R}Z_n = Z_n + {}^{R}\delta_z(Z_n, T) \tag{6.5.3}$$

where Z_n is the nominal machine axis position and ${}^{R}\delta_z(Z_n, T)$ is the error corresponding to this location and temperature T, with respect to the reference coordinate frame.

The error term ${}^{R}\delta_z(Z_n, T)$ is also expressed as the vectorial sum of the error at any nominal position along the Z axis, Z_n, with respect to a reference position along the same axis, Z_{nr}, and the error of this reference position with respect to the reference coordinates:

$${}^{R}\delta_z(Z_n, T) = {}^{R}\delta_z(Z_r, T) + {}^{r}\delta_z(Z_n, T) \tag{6.5.4}$$

where ${}^{R}\delta_z(Z_r, T)$ is the error of the Z axis reference position with respect to the reference coordinates, and ${}^{r}\delta_z(Z_n, T)$ is the error of any nominal position along the Z axis with respect to the Z axis reference position. Assuming that at $T = T_0$, ${}^{R}\delta_z(Z_r, T_0) = 0$,

$${}^{R}\delta_z(Z_n, T_0) = {}^{r}\delta_z(Z_n, T_0) \tag{6.5.5}$$

But, at $T = T_i$,

$${}^{R}\delta_z(Z_n, T_i) = {}^{R}\delta_z(Z_r, T_i) + {}^{r}\delta_z(Z_n, T_i) \tag{6.5.6}$$

Substituting Equations 6.5.5 and 6.5.6 into Equations 6.5.1 and 6.5.2, respectively, we find that

$${}^{R}Z_{0_T} - d_{0_z} = Z_{n_0} + {}^{r}\delta_z(Z_{n_0}, T_0) - Z_{g_0} \tag{6.5.7}$$

$${}^{R}Z_{i_T} - d_{i_z} = Z_{n_i} + {}^{R}\delta_z(Z_r, T_i) + {}^{r}\delta_z(Z_{n_i}, T_i) - Z_{g_i} \tag{6.5.8}$$

Subtracting Equation 6.5.7 from Equation 6.5.8, the machine Z axis reference offset, ${}^{R}\delta_z(Z_r, T_i)$, is found:

$$\begin{aligned} {}^{R}\delta_z(Z_r, T_i) \quad = \quad & (Z_{g_i} - Z_{g_0}) + ({}^{R}Z_{i_T} - {}^{R}Z_{0_T}) - (d_{i_z} - d_{0_z}) \\ & -(Z_{n_i} - Z_{n_0}) + {}^{r}\delta_z(Z_{n_0}, T_0) - {}^{r}\delta_z(Z_{n_i}, T_i) \end{aligned} \tag{6.5.9}$$

In the equation above, the second term on the right can be approximated by measuring the tool-setting station body temperature and correcting for thermal expansion. Similarly, the first term on the right can be calculated by measuring the gage bar temperature.

With similar reasoning, the machine X axis reference offset is calculated as

$$\begin{aligned} {}^{R}\delta_x(X_r, T_i) \quad = \quad & (X_{g_i} - X_{g_0}) + ({}^{R}X_{i_T} - {}^{R}X_{0_T}) - (d_{i_X} - d_{0_X}) \\ & -(X_{n_i} - X_{n_0}) + {}^{r}\delta_x(X_{n_0}, T_0) - {}^{r}\delta_X(X_{n_i}, T_i) \end{aligned} \tag{6.5.10}$$

where

- ${}^{R}\delta_x(X_r, T_i)$ is the machine X axis reference offset at $T = T_i$.
- X_{gi} and X_{g0} are the gage lengths in the X-direction at $T = T_i$ and $T = T_0$ respectively.
- $({}^{R}X_{iT} - {}^{R}X_{0T})$ is the change in the tool-setting station location along the X direction.
- d_{ix} and d_{0x} are the tool-setting station outputs in the X direction at $T = T_i$ and $T = T_0$, respectively.
- X_{ni} and X_{n0} are the nominal X positions when the tool-setting station is touched off to read d_{ix} and d_{0x}, respectively.
- ${}^{r}\delta_x(X_{n0}, T_0)$ and ${}^{r}\delta_x(X_{ni}, T_i)$ are the linear displacement errors corresponding to these nominal positions at $T = T_0$ and $T = T_i$.

Once the machine axis reference offsets are determined, the tool dimension offsets are found. At any temperature T_i,

$$ {}^{R}X_{i_T} - d_{i_x} = {}^{R}X_{n_i} - X_{tool} \tag{6.5.11} $$

$$ {}^{R}Z_{i_T} - d_{i_z} = {}^{R}Z_{n_i} - Z_{tool} \tag{6.5.12} $$

Using Equations 6.5.3 and 6.5.4, Equations 6.5.11 and 6.5.12 are rewritten as

$$ X_{tool} = X_n + {}^{R}\delta_x(X_r, T_i) + {}^{r}\delta_x(X_n, T_i) + d_{i_x} - {}^{R}X_{0_T} + ({}^{R}X_{i_T} - {}^{R}X_{0_T}) \tag{6.5.13} $$

$$ Z_{tool} = Z_n + {}^{R}\delta_z(Z_r, T_i) + {}^{r}\delta_z(Z_n, T_i) + d_{i_z} - {}^{R}Z_{0_T} + ({}^{R}Z_{i_T} - {}^{R}Z_{0_T}) \tag{6.5.14} $$

where

- x_n and z_n are the nominal axis positions when a tool touches off the tool-setting station.
- d_{ix} and d_{iz} are the tool-setting station readings in the X and Z directions, respectively, at T = T_i.
- ${}^{R}Z_{0T}$ and ${}^{R}X_{0T}$ are the initial tool-setting station locations measured when T = T_0.
- ${}^{R}\delta_x$ (X_r, T_i) and ${}^{R}\delta_z$ (Z_r, T_i) are the machine axis reference offsets.

Although this method may seem complex, it is less computationally intense than the spindle thermal drift calculations and thus actually makes implementation easier and more reliable, because it is not so dependent on using past records of thermal history to predict dimensional changes. Unlike the thermal mapping of the machines linear axes, the spindle thermal drift was too dependent on the past thermal history, as opposed to what is the temperature of the moment.

6.6 Real-time Implementation of the Error Compensation System

Once the resultant error vector is determined at the cutting tool, some means of compensating for the errors must be developed in order to improve the accuracy of the machined workpiece. The common and easy way to compensate for errors is by numerical control (NC) tape modification (change the part program). In addition to the obvious long time lag, tape modification has many other shortcomings. Usually, part programs contain endpoints of the movement and the type of the interpolation required between these points. Based on this information, the computer numerical control (CNC) controller determines the intermediate points through which the cutting tool has to move; hence except for correcting errors at the endpoints (and very few errors are linearly distributed), modifying the NC tape is not helpful. Also, NC tape modification does not compensate for thermally induced errors and is impractical for small batch sizes. Therefore, it was decided to implement the error compensation algorithm for this machine by using a dedicated, low-cost, single-board microcomputer. This section describes the principles of the compensation system, the interface requirements, and the hardware and software characteristics of the system.

In a typical machine tool, an axis servomotor drives the leadscrew based on the error signal derived from the position command and the position and velocity feedback signals. A typical CNC calculates the position command signals, compares the command value to the digitized position feedback signal, and performs the speed control either by monitoring velocity feedback or by deriving the speed feedback signal from the position feedback signal. Figure 6.6.1 illustrates a block diagram for such a controller (with the error compensation system added). A *real-time error compensation system* is an addition to the controller which injects the error compensation signals into the position servo loop. The device can be designed into the CNC unit or added on. In this case the controller hardware and software were fixed and the latter option was required.

There are two methods of injecting error compensation signals in real time. One method is to inject these signals into the servo position feedback signal in the form of an analog voltage. A compensation system using this technique had been implemented before at NIST. In some machine tool controllers, it is not possible to break into the servo control loop electronics to inject an error compensation signal, due to the fact that the axis servo control is implemented in a software algorithm. In this case, the compensation signal must be injected into the controller, in digital form, through input/output ports to be manipulated by the control software. By either method, the injection can be made without interfering with the normal operation of the machine tool controller and requires

no extensive modifications to the controller's electronic hardware. Figure 6.6.1 shows the block diagram of a control system modified by the injection of the compensation signal into the feedback signal. The compensation system developed in this case study injects the error compensation signals into the axis servo control loop software.

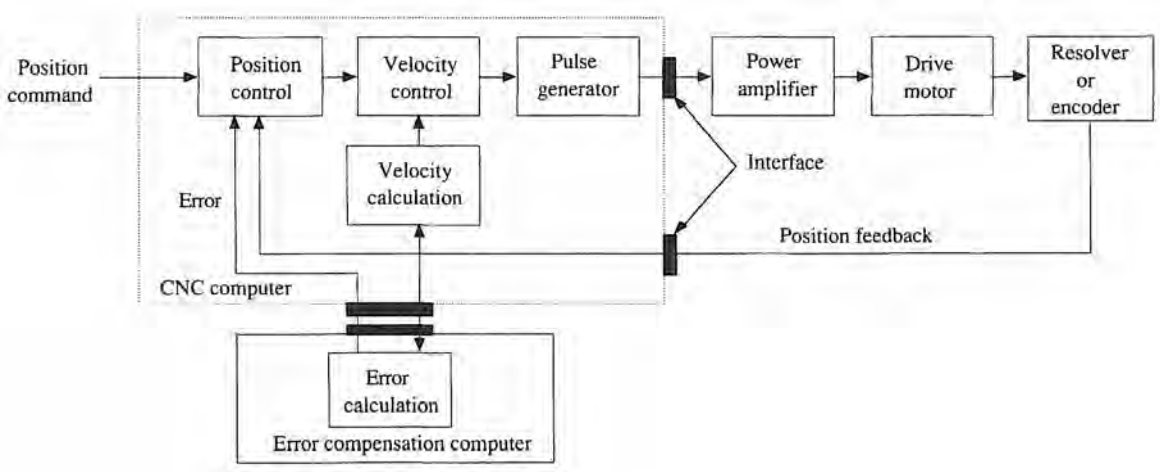


Figure 6.6.1 Error-compensated CNC axis drive. (Courtesy of NIST.)

In order to calculate the errors, the system used three types of independent parameters: (1) nominal axis position; (2) direction of motion; and (3) temperature measurements.[18] In addition, the system required knowledge of the initial conditions at which the location of the tool-setting station or the tool dimension offsets were to be determined. All the changes in the machine were referenced to these initial conditions during operation. The tool-setting station measurements of the reference bar location and the tool dimension offsets were also used by the system in the error calculations.

Based on the current nominal position, the direction of motion, the temperature data, the tool-setting station data and the initial conditions, the errors were calculated for both machine axes and sent to the machine tool controller. The controller received these values via parallel input/output (IO) ports and added them to the registers containing the *following* errors.[19] The position command signal, summed with the following error was the error signal used to drive the axis servo in the next servo cycle. Using this technique, the servo control timing of the machine tool controller was unchanged and the basic servo control algorithm was undisturbed. The schematic of the software servo control loop and the interaction with the error compensation system is shown in Figure 6.6.2.

System Hardware

The overall compensation system, shown in Figure 6.6.3, consisted of the following components:

1. Remotely controlled temperature measurement system.
2. Turning center keyboard interface module.
3. Tool-setting station.
4. Machine tool controller interface board.

The temperature measurement system was used to monitor the temperature at the previously described points on the machine. T-type copper-constantan thermocouples were used as the temperature sensors. The system made temperature measurements with 0.1 F^o resolution. A 10-channel scanner made it possible to monitor 10 channels simultaneously. Communication with the compensation controller (ISBC 86/30) was done through an RS-232 serial interface module, which was also part of the temperature measurement system. Upon receiving a command from the compensation

[18] The locations selected for temperature measurements were (1) the spindle rear bearing housing, (2) the leadscrew bearing housings for the cross-slide and the carriage, (3) the cross-slide, and (4) the carriage body. In addition, the temperatures on a tool-setting station base plate and reference gage bar were monitored for machine reference offset calculations.

[19] A *following* error is the lag of actual position from the commanded position by an amount proportional to the velocity of the axis motion.

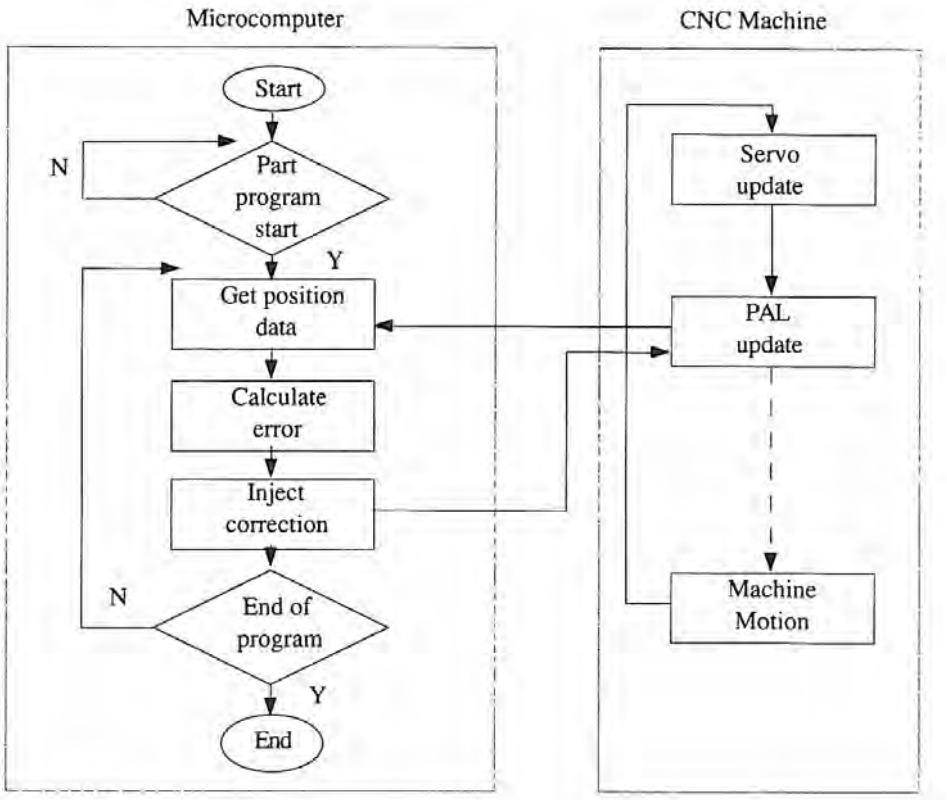


Figure 6.6.2 Schematic of the interaction between the machine tool software servo control loop and the error compensation software. (Courtesy of NIST.)

controller, the temperature measurement system digitized the appropriate channels and returned the temperature information.

The custom-designed keyboard interface module was based on an Intel 8048 single-component microcomputer. The function of this module was to translate the commands from the compensation controller and enter them into the machine tool controller by emulating the keyboard operation. This module was necessary for moving the machine tool axes to the proper positions so that the tool-setting station can determine the machine reference position offsets. The tool-setting station itself essentially consisted of a linear variable differential transformer (LVDT) displacement measuring device. It contained an Intel 8088 microprocessor to linearize the output of the LVDT's and increase the linearity from $\pm$ 0.25% to $\pm$ 0.025% based on mapping the LVDT's output with a laser interferometer.

A multibus single-board microcomputer, Intel ISBC 86/30, with 128k of RAM and 64k of EPROM memory, was the main controller of the overall compensation system. This microcomputer board contained a 16-bit Intel 8086 microprocessor as the CPU and a high-speed version of the Intel 8087A numeric coprocessor for floating point arithmetic operations. The architecture of this board was optimized for high-speed numeric computations which were required for real-time error correction. The combination of 8086 and 8087A made it possible to run the computer at an 8-MHz clock rate to further increase the computational speed, thus meeting the requirement of high servo bandwidth necessary for contouring cuts. This microcomputer used a multibus communications expansion board with four RS-232 serial I/O ports to communicate with the other components of the system, such as the temperature measurement system, the tool-setting station, the keyboard interface module, and a CRT terminal which was used for data entry and manual control of operations. Each I/O port on this board was controlled by an Intel 8251 USART (universal synchronous/asynchronous receiver/transmitter) chip, which could be programmed according to the communication requirements of the components.

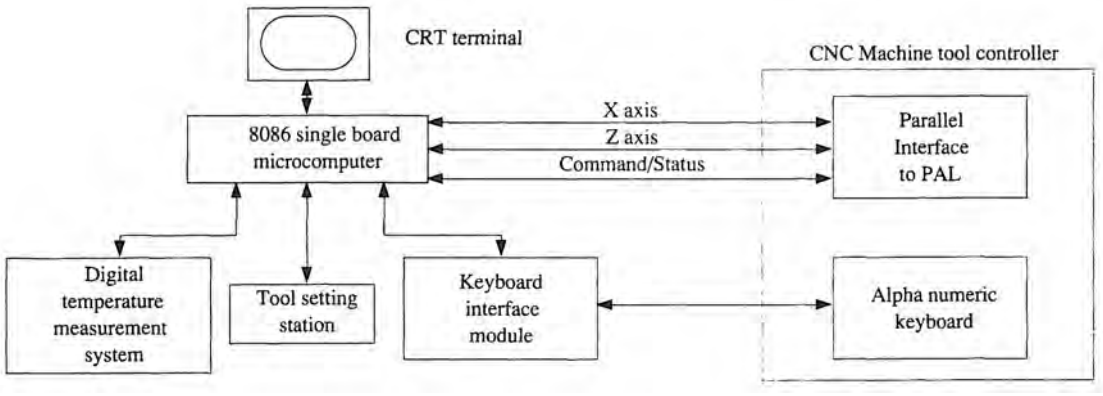


Figure 6.6.3 Error compensation system overall block diagram. (Courtesy of NIST.)

Compensated (μin.)	Uncompensated (μin.)	Improvement (ratio)
650	1300	2.00530
	1050	1.98
270	1030	3.81
-130	1470	11.31
-150	2230	14.87

Figure 6.6.4 Error in diameter (nominal diameter: 1.605 in.). (Courtesy of NIST.)

Cutting Tests

After the real-time error compensation system was implemented on the turning center, cutting tests were carried out under transient thermal conditions to determine the effectiveness of the error compensation system. One group of test parts were machined with no error compensation. The second group of test parts were machined under similar operating conditions but *with* the error compensation system working in real time. Later on, these two groups of parts were measured on a coordinate measuring machine.

The results of these cutting tests are summarized in Figures 6.6.4–6.6.7 in the temporal order in which the parts were machined. The first set of data in each table corresponds to the pair of parts machined when the machine was cold, while the last set of data corresponds to the pair of parts machined when the machine was warmed up after approximately 7 hours of running. As seen from these tables, the ratio of the dimensional accuracy improvement of the error compensated parts over the uncompensated parts increases as the machine warms up. In the last machined pairs, the accuracy of the diameter improved from 2230 μin (56.6 μm) oversize to 150 μin (3.7 μm) undersize, and the length improved from 6390 μin (162 μm) oversize to 450 μin (11.4 μm) oversize.

The increasing improvement of the dimensional accuracy over time was the result of the degrading accuracy of the machine tool itself as it warmed up. Alternatively, even though there seems to be an average factor of 3 improvement on the taper of the parts machined with the error compensation, there is not strong evidence of improvement. This lack of improvement is due to the fact that all the measurement data are close to the measurement uncertainty. Also, taper decreases as the machine warms up. Therefore, once steady-state temperature is reached, error compensation

Compensated (μin.)	Uncompensated (μin.)	Improvement (ratio)
-150	570	3.80390
	4410	11.31
-250	5240	20.96
-90	-480	5.33
450	6390	14.20

Figure 6.6.5 Error in length (nominal length: 3.44 in.). (Courtesy of NIST.)

Compensated (μin.)	Uncompensated (μin.)	Improvement (ratio)
26	85	3.27
18	88	4.89
-44	95	2.16
-8	-7	0.88
4	17	4.25

Figure 6.6.6 Error in taper (1.605 in. diameter, along 2.7 in. length). (Courtesy of NIST.)

Compensated (μin.)	Uncompensated (μin.)	Improvement (ratio)
33.5	44.1	1.3239.5
	8.7	0.22
25.7	36.8	1.43
54.3	78.6	1.45
32.0	58.2	1.82

Figure 6.6.7 Error in squareness (on 1.99 in. radius). (Courtesy of NIST.)

for tapers is not as essential as it is for length and diameter accuracy on this particular machine. Perhaps the greatest improvement can be seen in the diameter and length of parts. On the other hand, there is no significant improvement observed on squareness of the parts machined with the error compensation system. This result suggests that some additional squareness measurement on the machine and machined parts are needed. Based on these measurements, the coefficients for the squareness terms in the software could then be adjusted to further improve the squareness. On the other hand, perhaps on some lathes squareness errors are not major errors that need to be corrected.

6.7 Summary and Conclusions

Errors due to temperature changes in a machine tool were shown to be of the same order of magnitude as the geometric errors. Furthermore, the existence of a large number of error components required a systematic approach to the problem of measuring, mapping, and compensating for them.

To demonstrate the procedure, geometric and thermally induced errors were measured as functions of nominal axis positions and the temperatures of selected locations on a two-axis turning center. Since the major heat sources on a machine tool are fixed and coupled by the structure, the temperature distribution at any given time can be predicted by monitoring the temperatures of a few locations. Consequently, the error behavior is unique with respect to temperature distribution, unless a drastic change in thermal conditions occurs, such as an attachment of a high-intensity external heat source to a part of the machine tool. If there is a permanent drastic change in the thermal conditions, the machine tool has to be recalibrated for the new thermal conditions. The errors were measured using a laser interferometer and high-precision capacitance probes while measuring the temperatures of the selected locations as the machine was gradually warmed up. A least-square's analysis of the data was used to characterize the errors. In predicting the thermal effects, the changes in the machine reference positions were monitored on-line using a tool-setting station. With this approach, it was possible to predict the errors when the machine was in a transient thermal state.

The software compensation system was implemented on a single-board microcomputer. This microcomputer, which was interfaced with the CNC machine tool controller, determined the errors, which were found to be functions of the nominal axis positions, the temperatures of the selected locations, and the last measurement taken from the tool-setting station. The error was then injected into the machine tool controller software servo loop. The error injections were updated at the cycle rate of the machine tool controller, which was 20 ms.

Cutting tests were carried out under transient thermal conditions. Along with the elimination of the nonproductive warm-up period, frequently lasting as long as 10 hours, accuracy enhancements of up to 20 times over the uncompensated workpieces were achieved.

It is interesting to note that although the results were good, substantially more effort was required to map the thermal errors than was required to map the geometric errors. In general, many manufacturers support the idea of geometric error correction, but they are skeptical about the effectiveness of thermal error corrections. This is due to the fact that thermal error mapping is difficult and time consuming to do and it often depends on how a machine is designed, built, and used. Perhaps a better overall method for reducing errors, short of using a metrology frame, is to map the geometric errors and temperature control the machine. The latter can be achieved by better design of components and interfaces and by active cooling of the structure. Active cooling (e.g., with oil showers) has been proven to be a most effective means of controlling thermal errors.[20]

[20] See Section 2.3.5.2 and J. Bryan et al., "An Order of Magnitude Improvement in Thermal Stability with Use of Liquid Shower on a General Purpose Measuring Machine," SME Precis. Eng. Workshop.technical paper IQ82-936, June 1982, St. Paul, Minn.; and D. B. DeBra, "Shower and High Pressure Oil Temperature Control," Ann CIRP, Vol 35, No. 1, 1986.

Chapter 7
System Design Considerations

Things should be made as simple as possible, but not too simple.

Albert Einstein

7.1 INTRODUCTION

The first six chapters focused on the nature of errors and how to measure them; now the time has come for the study of the methods and components that are used to make machines. This chapter will be concerned primarily with manufacturing, structural, and system design considerations. Section 7.2 discusses manufacturing considerations, including various processes used in the manufacture of machine tools. This subject in itself could fill a library with information, so the material presented here is presented only as a brief review. Section 7.3 discusses materials commonly used in machine tools and new materials on the horizon (e.g., ceramics). Once again, the material in this section is intended to be a brief review only. Section 7.4 discusses issues in the design of the overall machine structure. Section 7.5 discusses issues in joint design including bolted, bonded, and interference fit joints. Section 7.6 discusses the often neglected but always critical support systems such as safety systems. Section 7.7 presents an in-depth study on the design of kinematic couplings. Section 7.8 is a case study about the design of a large precision grinding machine.

Additional Thoughts on Systems Design Philosophy[1]

Once the nature of errors in a machine and how to measure them are understood, the overall structure of a machine can be more easily configured. In addition, in order to be able to effectively formulate the overall design, the design engineer must simultaneously envision in his head the functions the machine must perform (e.g., milling, turning or grinding) alongside a mental pictorial library of component technologies (e.g., bearings, actuators, and sensors), generic machine configurations (e.g., cast or welded articulated and/or prismatic structures), analysis techniques (e.g., back-of-the-envelope and finite element methods), and manufacturing methods (e.g., machine, hand, or replication finished). In addition, the machine design engineer must be aware of the basic issues faced by the sensor and electronics engineer, the manufacturing engineer, the analyst, and the controls engineer. Only by a simultaneous consideration of all design factors can an optimal design be rapidly converged upon. Awareness of current technological limitations *in all fields* can also help a design engineer to develop new processes, machines, and/or components.[2]

Just as machines and companies that do not adapt to take advantage of new technologies fade away, so do design engineers. Thus at this point, it is worthwhile once again to note that the way you view and live life has a profound effect on your ability to adapt to rapidly changing technologies.[3] *Design is not just a job; it is a love, a passion, an artform at which only the dedicated can succeed.* If you have this inner drive to design, one of the best ways to train yourself to be able to effectively and creatively perform a simultaneous viewing of all design issues is by taking note of the "editorial advice" presented in Chapter 1. Sugar is sweet and makes life more pleasurable, which in turn increases productivity and creativeness; however, if you eat too much sugar and don't brush your teeth, your teeth will decay, fall out, and make life difficult.

---

[1] *"...the man who aspires to fortune or to fame by new discoveries must be content to examine with care the knowledge of his contemporaries, or to exhaust his efforts inventing again what he will probably find has been better executed before."* Charles Babbage

[2] "*The higher type of man may not know how to deal with small matters, but he can undertake great ones. The lower type of man cannot undertake great things, but he may know how to deal with small ones*." Confucius

[3] The serious machine design engineer should regularly read/scan the following publications (in addition to magazines like Machine Design or Design News) which often have good articles, but more important, have lots of informative advertisements that keep you informed as to what components and machine tool systems exist: American Machinist and Automated Manufacturing, published monthly by McGraw-Hill, Inc., 1221 Avenue of the Americas, New York, NY 10020; Modern Machine Shop, published by Gardner Publications, 6600 Clough Pike, Cincinnati, OH 45244-4090; and Assembly Engineering, published by Hitchcock Publishing Company, 25W550 Geneva Road, Wheaton, IL 60188-2292. Serious machine designers should also regularly scan the various machine design journals that are available including, for example, Precision Engineering, available from ASPE, Box 7918, Raleigh NC 27695-7918; International Journal of Machine Tools and Manufacture and Mechanism and Machine Theory, both published quarterly by Pergamon Press, Maxwell House, Fairview Park, Elmsford, NY 10523; Journal of Dynamic Systems, Measurement, and Control, published quarterly by ASME, 345 East 47th St., New York, NY 10017; and International Journal of Mechanical Sciences, published monthly by A. Wheaton & Co., Exeter, England.

Just as those who train for sports must continually practice and observe strict patterns of diet, exercise, and sleep, so too must a design engineer maintain his mental and physical health. As drugs, alcohol, and junk food are bad for a marathon runner, so too are they bad for a design engineer, for what slows the body often also slows the mind. Furthermore, just as physical training is required prior to athletic competition, mental training is constantly required for a design engineer. This training can take the form of observing the world around you and continually asking yourself why is that? and how can I make it better? *Only through constant training can design engineers face their daily competition at work where they try to design products better than the competition. You cannot become a good athlete or a good design engineer if you spend all your spare time sitting in front of the TV or sitting in a bar.* Furthermore, if you do not plan on becoming a good design engineer, you might as well quit now and not waste your time. There are many other hungry people (including design engineers of expert systems) in the world who do have the will to make you obsolete and expendable.

7.2 MANUFACTURING CONSIDERATIONS

One of the most important design considerations is: Can the design be economically manufactured with sufficient quality control to ensure that performance specifications are met? Furthermore, it must be ensured that any failures that might occur are few and far between, and they do not give the product and/or the manufacturer a bad name. Often a design can earn a bad reputation based on simple ergonomic considerations. For example, there is a general rule for industrial equipment that says do not use any bolts smaller than about 6 mm ($^1/_4$ in.) diameter, even for mundane items such as filter cover plates, because small bolts are too easy to over-torque and break during assembly. Similarly, small drills and taps are more likely to break than large ones. Not that everything should be overdesigned, but the design engineer should always subject the design and its components to continual value analysis with respect to fabrication, use, and maintenance issues. *A good design engineer considers all aspects of a design simultaneously, including how it will be manufactured even when just sketching out the conceptual idea. A product can function smoothly as an integrated system only if design decisions are made in an integrated fashion.*

Manufacturing processes can be divided up into several major groups, including assembly, casting and forming, and material removal (among many others). Within each of these major subgroups there can be hundreds of different types of specific processes. Numerous books have already been written about these processes, so the goal of this section is to briefly review various manufacturing processes,[4] to give the reader an opportunity to reformulate in his mind what are their true characteristics in the context of the material discussed in Chapters 1-6. The manufacturing technologies that will be reviewed in this section include:

- Assembly
- Casting and powder materials technology
- Material removal
- Treatment operations

In order to help develop a "feel" for these various manufacturing process, one should play a mental game, "how was it manufactured," with everything you see[5] and to spend lots of time with the people who actually fabricate the parts you design. The best design engineers know that fabrication personnel are often very creative and can help you to learn what can be done[6] and suggest ways to improve your design.

---

[4] For more detailed discussions of production processes with lots of informative pictures, see, for example, R. Bolz, Production Processes: The Productivity Handbook, Industrial Press, New York; and E. P. DeGarmo, Materials and Processes in Manufacturing, 5th ed., Macmillan Publishing Co., New York, 1979. Also see the Metals Handbook Volumes 4–7, published by the American Society for Metals, Metals Park, OH.

[5] Once again, the author would like to stress the importance of continually observing and analyzing the world around you whether you're at work, rest, or play.

[6] Of course, when told that something cannot be done, you should always ask "why?" because it may lead to a new design opportunity.

7.2.1 Assembly

Most products are assembled from parts, and thus careful consideration of the performance of the system, as well as of individual parts, must always be kept in mind. This includes consideration of how the design's parts will be manufactured and assembled. When sophisticated tools were not available for the manufacture of parts, and when design engineers often had to build the first prototype themselves, the art of design emphasized the need to minimize manufacturing and assembly effort while maintaining product quality and function. As more advanced manufacturing and design tools and materials became available, the complexity of many designs increased dramatically. As complexity increased, many assumed that specialization was the only way to cope with the expanding array of technologies. Unfortunately, many also regarded specialization as an excuse for ignorance of manufacturing concerns. In some industries manufacturing became known as a dirty job unbefitting of a real engineer. As a result, although many products met their performance goals, many were difficult to manufacture, assemble, and were thus expensive. Fortunately, that attitude seems to be dying out. Although specialists will always be needed, all personnel associated with the design process should be able to regard their work objectively and answer the question: Is there a simpler way to manufacture and assemble this product's components? It is critically important that the design engineer keep well informed of the latest technological advancements. Education must be a never-ending process.

Design for manufacture and assembly *must* consider the product as a system in order to be effective. For example, a metrology frame may be expensive and difficult to design and assemble, but a machine tool structure that can resist moving loads and still achieve submicron accuracy over a range of a meter is even more difficult to design. As a simpler example, consider the axle assemblies shown in Figure 7.2.1. It might be argued that any of the individual steel parts are far less complex then the one-piece molded plastic part; however, when one considers how much it will cost to fabricate the parts (including tooling) and assemble them, for large production runs the molded nylon design is an order of magnitude less expensive than the formed steel design.[7]

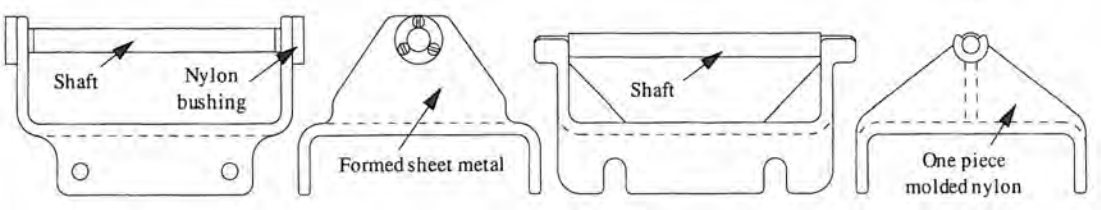


Figure 7.2.1 An axle assembly made from many bolted steel parts, and one made from a few snap together parts. (After Boothroyd and Dewhurst.)

The ability to visualize simultaneously different ways of accomplishing a task and how to fabricate and assemble the associated parts is vital to the success of a good design engineer. There has been a recent surge of research in this area, particularly research concerned with developing expert systems to estimate manufacturing costs. But one should consider an expert system as a tool like a calculator that is used to help the design engineer converge upon a solution. The design engineer must still develop the experience that will enable him from the start to pick a near-best solution which will facilitate rapid convergence to the final solution.

There have been attempts to develop sequential rules (design axioms) for maximizing the manufacturability of parts and products. Axioms are useful tools in the design process and in the future they may evolve into expert systems that will help the design engineer to explore more design options.[8] A few broad design guidelines are:

- Subject all decisions to an "is there a better way" value analysis based on system considerations.

[7] For a more detailed discussion, see the source of this example, G. Boothroyd and P. Dewhurst, Product Design for Assembly Handbook, Boothroyd and Dewhurst, Inc., Wakefield RI, 1987. Also see S. Miyakawa and T. Ohashi, "The Hitachi Assemblability Evaluation Method (AEM)," Proc. 1st Int. Conf. Prod. Des. Assem., April 1986.

[8] *Cunning words may confound the principle of virtue. Impatience in little things may confound mighty plans.* Confucius

- Always picture in your mind how the system will be manufactured, assembled, used, and maintained.
- Minimize the number of parts in an assembly and minimize their complexity.
- Maximize the number of instances where reference surfaces and self-locating "snap together" parts can be used.
- Whenever possible, take advantage of kinematic design principles.
- Try to avoid an abundance of small pieces and threaded fasteners.
- Try to make assembly occur from one direction (i.e., everything fits in from the top).
- Utilize new materials and technologies to their fullest potential.
- Read, read, read, and familiarize yourself with everything.
- Observe, observe, observe, and familiarize yourself with everything.
- Remember Maudslay's Maxims.

These points are illustrated in principle by the examples shown in Figures 7.2.2 – 7.2.20.

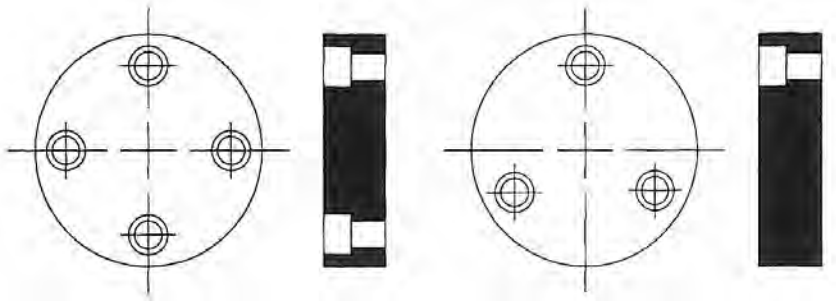


Figure 7.2.2 Subject all bolt patterns to value analysis: How many bolts are really needed?

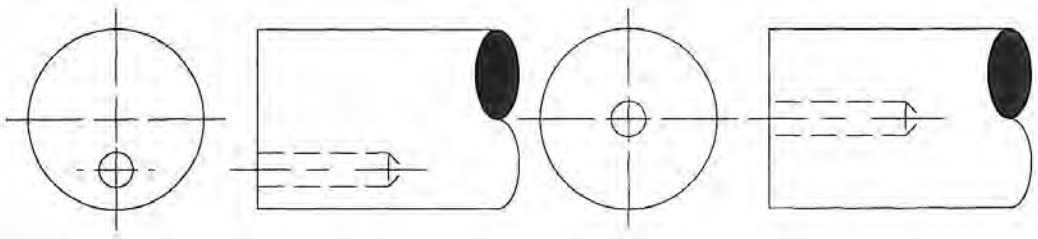


Figure 7.2.3 If the hole is in the center, it can be made while the shaft is still in the lathe.

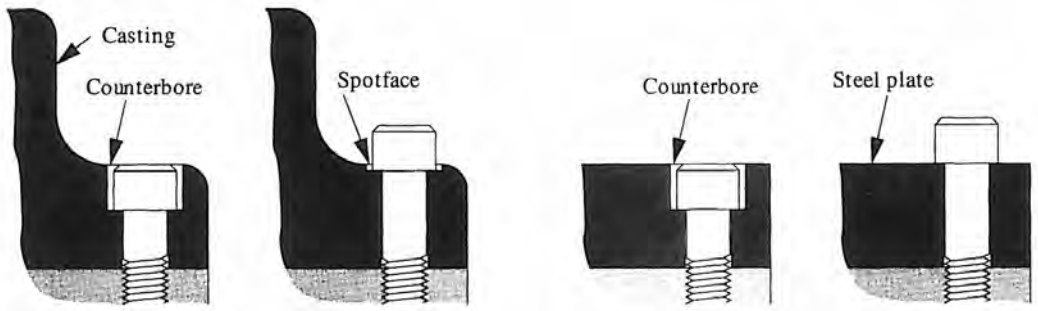


Figure 7.2.4 Do not specify a counterbore unless it is needed.

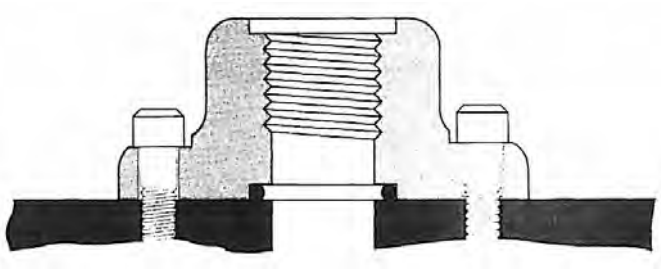


Figure 7.2.5 Put added features, such as an O-ring counterbore, in the easier to handle, less costly part.

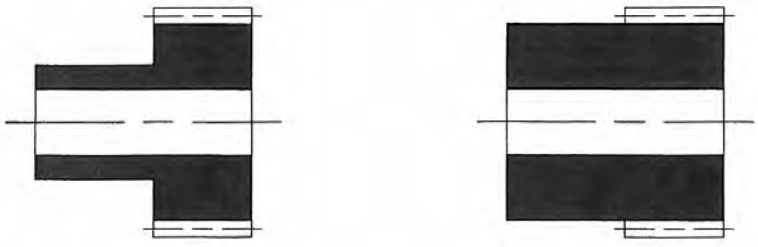


Figure 7.2.6 Do not specify that excess material is to be removed unless required for weight reduction or aesthetic purposes. Subject all decisions to value analysis.

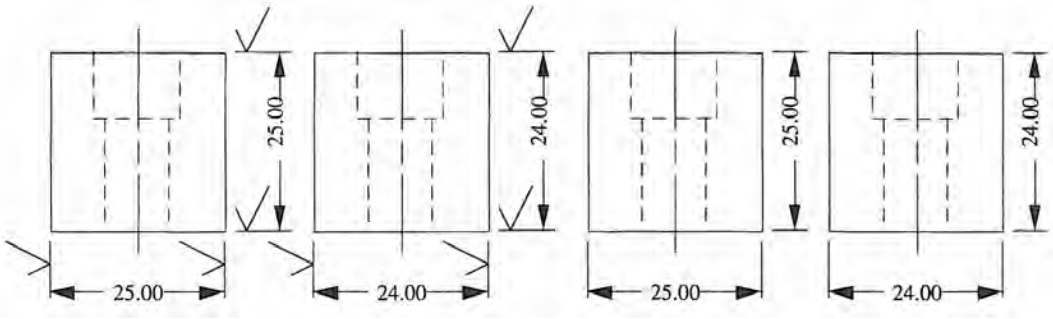


Figure 7.2.7 When possible, size and dimension parts so they can be made from stock or from stock that requires minimal removal. Avoid unnecessary finishing requirements if possible.

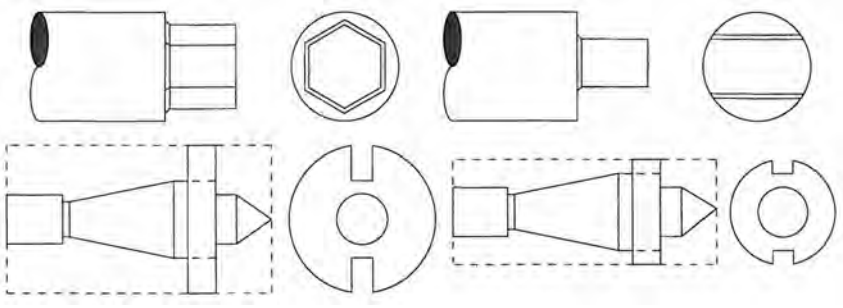


Figure 7.2.8 Avoid specifying unnecessary features.

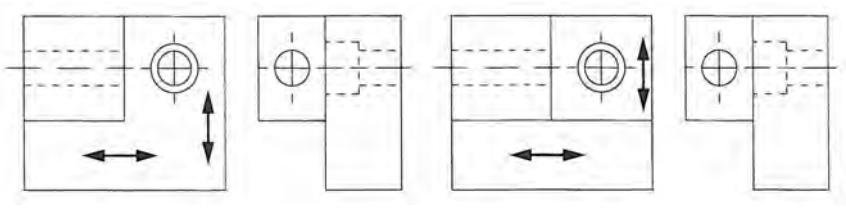


Figure 7.2.9 Avoid implied finishing requirements (e.g., implied tolerance on matching surfaces) unless they are required.

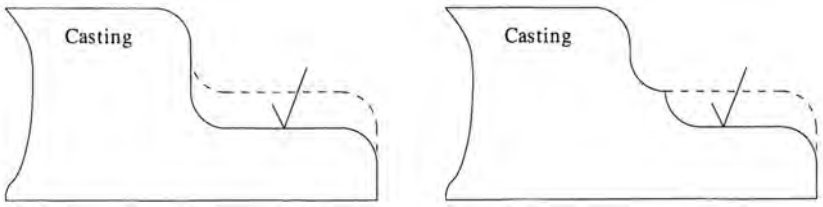


Figure 7.2.10 Allow for "mismatch" region between surfaces produced by different processes. Be wary of stress concentration producing corners.

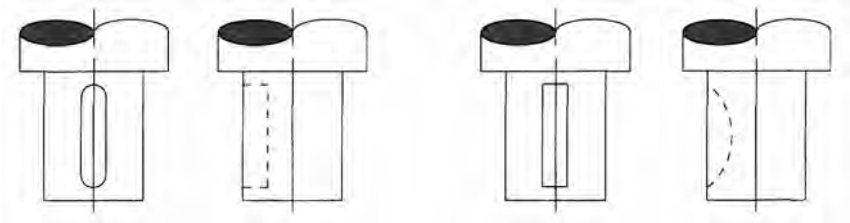


Figure 7.2.11 Consider the most economical machining method and any requirements it may have on part dimensioning and assembly.

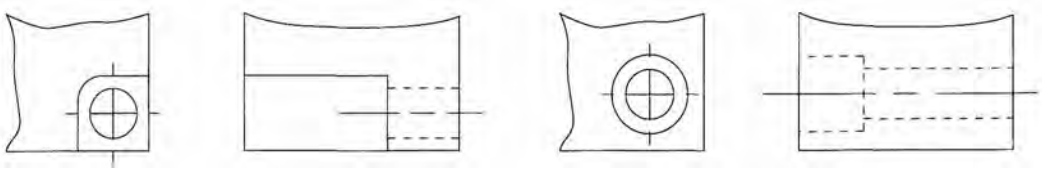


Figure 7.2.12 Consider the machining operation and its effect on part design.

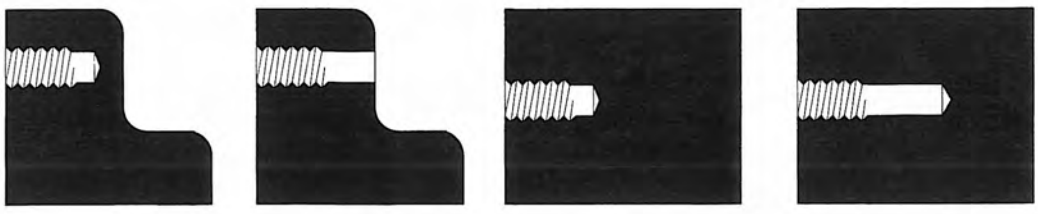


Figure 7.2.13 Think about where the chips from the cutting process go.

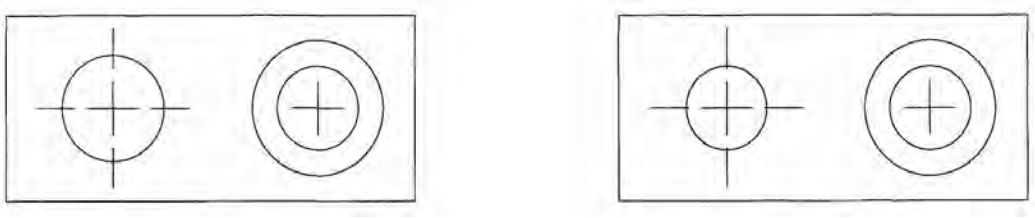


Figure 7.2.14 Specify same-size holes whenever possible.

Figure 7.2.15 Use reliefs and bosses to aid drilling and other machining and assembly operations.

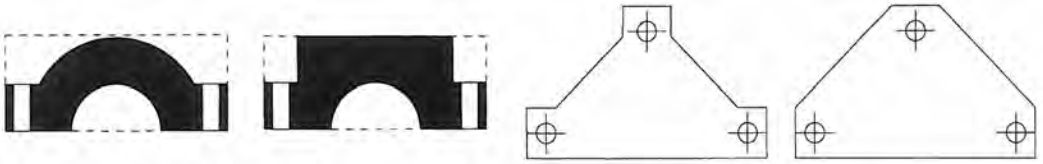


Figure 7.2.16 Try to minimize time and effort required to machine a part.

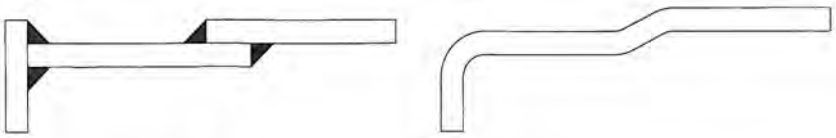


Figure 7.2.17 Consider forming or machining a part from a solid rather than welding small parts together.

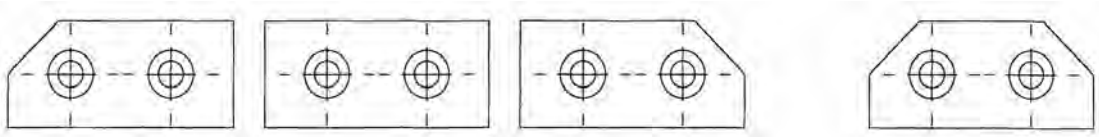


Figure 7.2.18 Minimize the number of part numbers required and maximize symmetry.

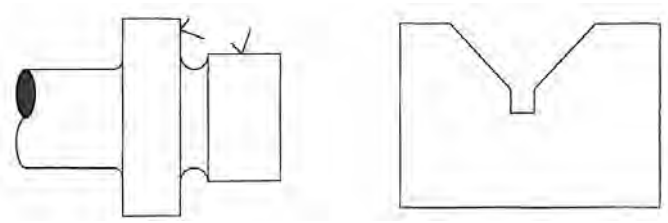


Figure 7.2.19 Provide undercuts to avoid impracticality of having to machine in sharp corners.

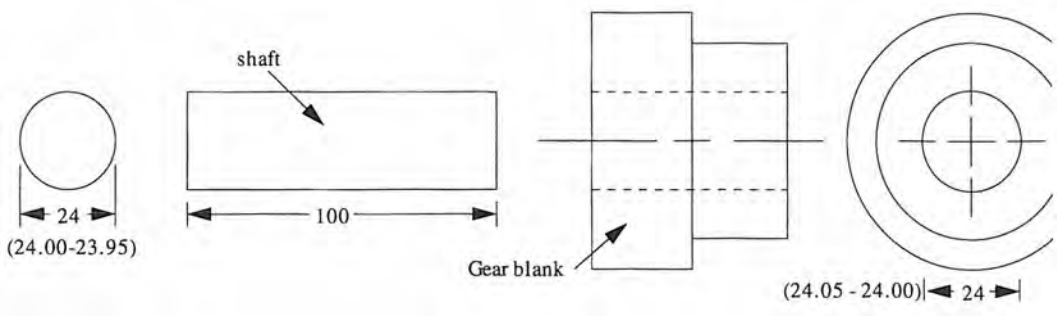


Figure 7.2.20 Variations on part sizes must be anticipated to allow for assembly even in the worst case. Dimensions and tolerances in parentheses would have been required on the part drawings to ensure this.

Details and Tolerances[9]

Success as a design engineer depends on one's ability to accurately convey the details of a design to draftspeople and manufacturing engineers. In fact, it is the often overlooked minute detail (e.g., cable routing) or tolerance error (e.g., unwanted interference) that causes a design to fail. Perhaps the best way for an engineering student to learn about details and tolerancing is thus to continually be involved in "hands-on" research projects. If a picture is worth a thousand words, experience is worth a million words. It is vital that the design engineer maintain a close professional relationship with manufacturing engineers and fabrication personnel who will be responsible for producing the design and who probably have seen countless good and bad designs.

When a product is made up of many different parts, one must always consider how part dimension variations will affect the product's performance, manufacturability, and assembleability. Many engineers are unfamiliar with part tolerancing, in that they dimension all parts to have exactly fitting dimensions as shown figuratively in Figure 7.2.20, while others specify unneeded and unrealistic tolerances. However, after reading and comprehending Chapters 1-6 of this book, the reader should understand the nature of errors and therefore easily be able to figure out the effect of dimensional variations on the assembleability and performance of a system. In effect, tolerancing is just another form of error budgeting. In addition, when tolerancing a part, one needs to keep in mind the ability of the intended manufacturing process to achieve the tolerance. Figures 7.2.21 and 7.2.22 show commonly achievable tolerance ranges and surface finishes, respectively, for various manufacturing processes.

Quite often a drawing will have an implied tolerance between details as drawn, and in this type of situation the design engineer should explicitly note when tolerances are not critical, as shown in Figure 7.2.23. When dimensioning a drawing, if you are unsure as to the appropriate symbol, or cannot find one, it is better to write in plain language what you want and suffer verbal abuse on the order of "can't you draw?!" rather than "can't you design?!" Even simple notes such as "lightening hole" to denote noncritical features can help to reduce manufacturing costs.

[9] See, for example, F. E. Giesecke et al., Technical Drawing, 7th ed., Macmillan Co., New York, 1980, and Dimensioning and Tolerancing Y14.5M-1982, available from the American National Standards Institute, 1430 Broadway, New York, NY 10018.

Size range (inches)										
From	Through	Tolerances (inches) ±								
0.000	0.599	0.00015	0.00020	0.0003	0.0005	0.0008	0.0012	0.0020	0.003	0.005
0.600	0.999	0.00015	0.00025	0.0004	0.0006	0.0010	0.0015	0.0025	0.004	0.006
1.000	1.499	0.00020	0.00030	0.0005	0.0008	0.0012	0.0020	0.003	0.005	0.008
1.500	2.799	0.00025	0.0004	0.0006	0.0010	0.0015	0.0025	0.004	0.006	0.010
2.800	4.499	0.0003	0.0005	0.0008	0.0012	0.0020	0.003	0.005	0.008	0.012
4.500	7.799	0.0004	0.0006	0.0010	0.0015	0.0025	0.004	0.006	0.010	0.015
7.800	13.599	0.0005	0.0008	0.0012	0.0020	0.003	0.005	0.008	0.012	0.020
13.600	20.999	0.0006	0.001	0.0015	0.0025	0.004	0.006	0.010	0.015	0.025
Lapping & honing		XXXXXXXXXXXXXX								
Grinding, diamond turning, boring			XXXXXXXXXXXXX							
Broaching			XXXXXXXXXXXXXX							
Reaming				XXXXXXXXXXXXXXXXXXXXXX						
Turning, boring, slotting, planing, shaping						XXXXXXXXXXXXXXXXXXXXX				
Milling							XXXXXXXXXXXXXXXXX			
Drilling								XXXXXXXXXXXXX		

Figure 7.2.21 Tolerance ranges for various machining processes. (Courtesy of Cincinnati Milacron.)

Error Mapping

The commercial importance of error mapping was illustrated in Section 1.6, where it was shown how it greatly simplified the design and manufacturing of a coordinate measuring machine. Chapter 6 was devoted to how error mapping can be accomplished by using external sensors to map a machine's errors. Accordingly, it only needs to be reiterated here that error mapping is one of the most significant developments in the use of computers in machine tools since the introduction of numerical control. The serious machine design engineer would be well advised to realize this.

7.2.2 Casting and Powder Materials Technology

Casting and powder materials technology are among the most useful processes available. They are inherently efficient because they yield near net shape parts that require minimal machining. Casting, however, requires a mold whose design often influences the shape of the part and casting also requires the part to be annealed to relieve stress and attain stability. Powder materials technology requires several steps, including molding, pressing, and baking. Still, when one considers the most common alternative, machining from a solid, one often finds that casting or powder materials technology are often the best processes to use.[10]

7.2.2.1 Metal Castings[11]

Casting processes are typed according to the type of mold used. Virtually any type of metal can be cast into virtually any type of shape with virtually any required quality level.[12] However, regardless of the casting process used, precision surfaces often still have to be machined and/or ground. Table 7.2.1 lists common casting methods and their characteristics. When one thinks of a machine tool, one often thinks of a large cast iron structure, and sand casting is the process most often used to manufacture the large cast iron parts of a machine tool. After the cast components are received from the foundry, they are stress relieved and mating surfaces are finish machined, ground, and scraped.

[10] Note that there are many other bulk processes, such as extrusion, forming, forging, and molding that are also used to make parts used in precision machines; however, casting and powder metallurgy are the ones that have the most impact on the overall design of a precision machine.

[11] See, for example, the magazine Casting Design and Application, published by Penton Publishing Inc., 1100 Superior Ave., Cleveland, OH 44114.

[12] For example, aluminum alloy casting methods have advanced to the state where they can be used as blanks for diamond machined optical surfaces. See R. Dahlgren and M. Gerchman, "The Use of Aluminum Alloy Castings as Diamond Machining Substrates for Optical Surfaces," Proc. SPIE OE LASE 1987.

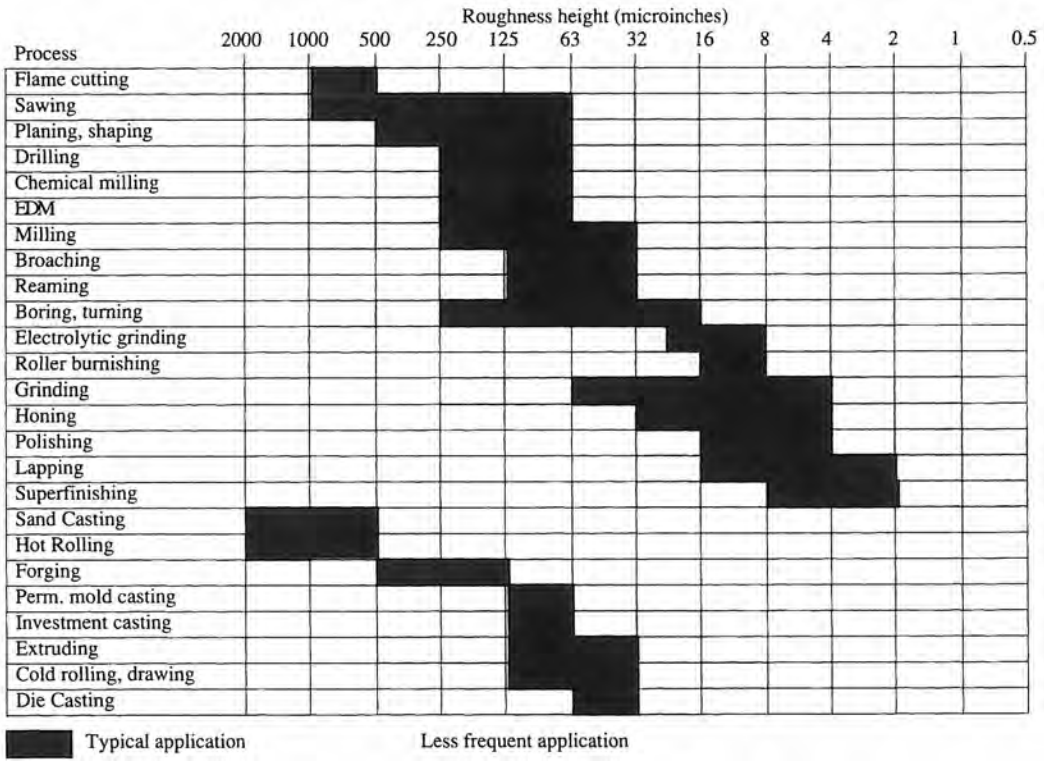


Figure 7.2.22 Typical surface roughness (average of deviations from true surface) for various manufacturing processes. (After Spotts.)

Note that the unmachined surfaces of a casting should *not* show unsightly marks from the casting process. Just as one buys a car for transportation purposes but still wants the body to look nice, a junky-looking casting is often perceived as a sign of a junky machine. Hence one should consult with the foundry to make sure that the design of the casting will also result in a good-looking part.

Sand casting can be used to make virtually any size part, and it basically involves making a pattern out of a suitable material (e.g., foam, wood, or metal) and packing sand around it. Parting lines are used to allow the sand mold components to be disassembled from around the pattern and then reassembled after the part is removed. Sand cores are often inserted into the mold to form cavities inside the mold (e.g., the cylinders of an engine casting). Regardless of the design of the part, one must also consider that as the metal cools it shrinks (on the order of 5-10% for most metals), and that in order to remove the part from the mold without breaking the mold, a taper (draft) of about 1:10 is required. In addition, extra metal should be added to surfaces that will have to be machined (a machining allowance), and locating surfaces should be added so that the part can be fixtured to facilitate machining. Thus in order to specify a casting, there are a few basic guidelines one needs

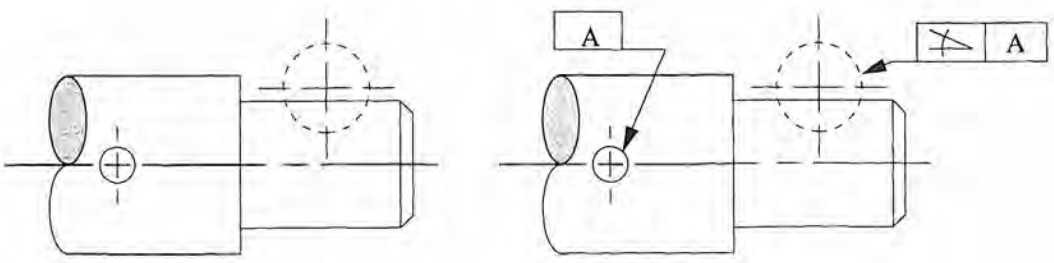


Figure 7.2.23 Avoid implied accuracy; make notes on the drawing designating noncritical dimensions and features.

Casting Process	Min-Max weight(g, kg)	Min. section thickness (mm)	Machining allowance (mm)	Tolerance (±mm)	Surface finish (μm)	Minimum lot size
Sand	100-800k	3	2-10	1/2–6	6–24	1
Perm. mold	100-25	3	0.8–3	1/4–1.5	2.5–6	1000
Die	<1-30	3/4	0.8–1.6	0.025–0.125	0.8–21/4	3000
Plaster mold	100-100	1	3/4	1/8–1/4	0.8–1.3	1
Investment	<1-50	1/2	1/4–3/4	0.05–1.5	0.5–2.2	20

Table 7.2.1 Various casting processes and their characteristics. (After Kutz.)

to know in order to minimize the work that a professional mold design engineer has to do to clean up your design. These guidelines are discussed below.[13]

When molten metal solidifies, it wants to expel impurities. Hence when the outside of a casting solidifies, the inside tends to contain the impurities. This can lead to loss of strength of the casting. Thus the first rule is to design the casting so that it cools evenly to minimize porosity and maximize homogeneity. In order to help identify potential problem areas (hot spots) where sections meet, as shown in Figure 7.2.24, one can draw little arrows normal to the surface of the casting. The arrows should not overlap. Figure 7.2.25 shows relative cooling rates for various sections and how the addition of fillets can facilitate heat transfer out of the casting and into the mold. Figure 7.2.26 shows how all sharp angles should be replaced with generous fillets. As shown in Figure 7.2.27, directional solidification and multiple feed gates can also be used to control the metallurgical qualities of the casting. Section components should be made to have even thickness, and when necessary to deviate from constant thickness it should be done gradually as shown in Figure 7.2.28. Figures 7.2.29 and 7.2.30 show how these guidelines can be applied to more complex patterns.

The casting is often the structural framework for the machine and thus usually has numerous components attached to it. Thickened regions for bolts or mounting surfaces are called *bosses*. Figures 7.2.31 and 7.2.32 show methods for designing bosses. Since bosses are often thickened regions, one must take care to make the transition from a section to a boss gradual and generously filleted.

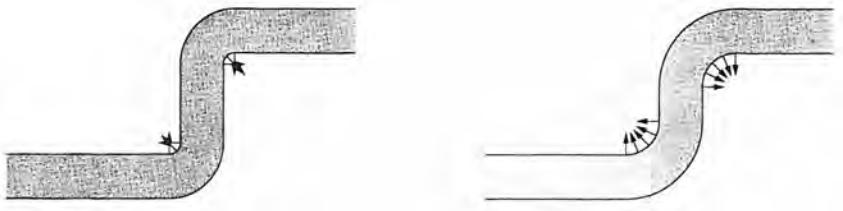


Figure 7.2.24 Identifying hot spots in castings by using outward projecting arrows of length half the casting thickness. Where arrows overlap, hot spots may develop. (Courtesy of Meehanite Metal Corp.)

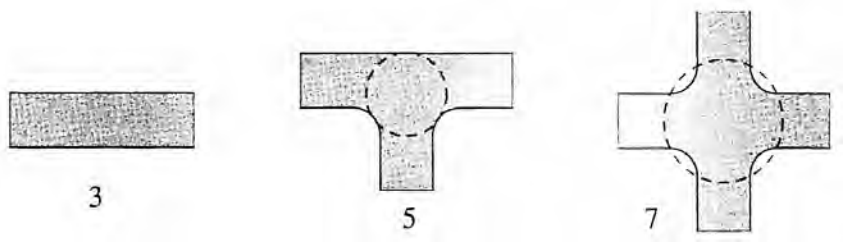


Figure 7.2.25 Examples of relative cooling times. (Courtesy of Meehanite Metal Corp.)

[13] See, for example, the handbook Casting Design as Influenced by Foundry Practice, Meehanite Metal Corp., Marietta, Georgia.

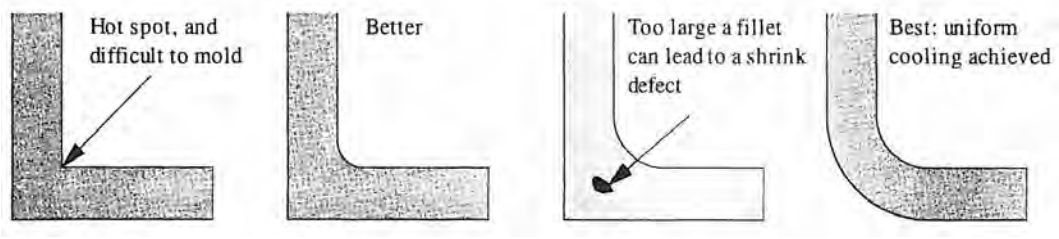


Figure 7.2.26 Fillet all sharp angles. (Courtesy of Meehanite Metal Corp.)

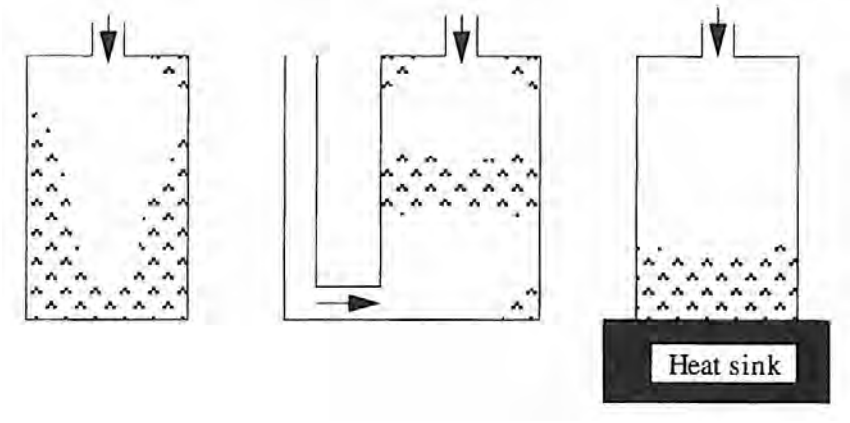


Figure 7.2.27 Directional solidification (risers not shown). (Courtesy of Meehanite Metal Corp.)

One of the principal advantages of castings is that they allow the design engineer to maximize a structure's stiffness-to-weight ratio. This is accomplished by placing material far away from the neutral axis and ribbing it to prevent buckling, or diaphragm-like vibrational modes. Ribs welded to a plate or machined from a solid are comparatively more expensive than cast ribs. Figures 7.2.33–7.2.35 illustrate some basic rules of rib design. Of course, these are just guidelines, and all designs should first be developed using back-of-the-envelope calculations and then checked and modified with the aid of finite element methods.

Permanent Mold Casting

When a part is to be manufactured in lots of hundreds or thousands, it is often advantageous to make a metal mold. A metal mold will yield a part that has better surface finish, grain structure, and dimensional accuracy than those of a sand-cast part. Like sand castings, permanent mold castings are often poured from a ladle.

Die Casting

Die casting is a form of permanent mold casting where the liquid metal is injected into the mold cavity under high pressure using a plunger mechanism. Because the liquid metal is forced into the die, it will fill all parts of even very complex molds. When the part cools in the mold, it is ejected with a plunger rod, and metal for a new part is then injected into the mold. In a sense,

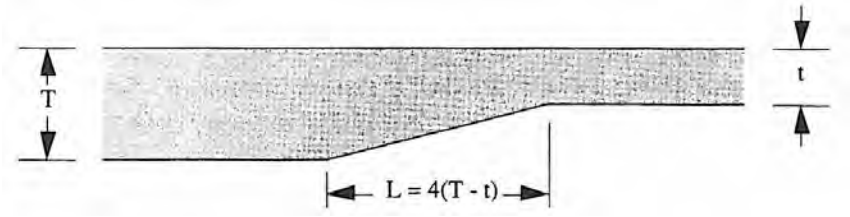


Figure 7.2.28 Avoid abrupt section changes. (Courtesy of Meehanite Metal Corp.)

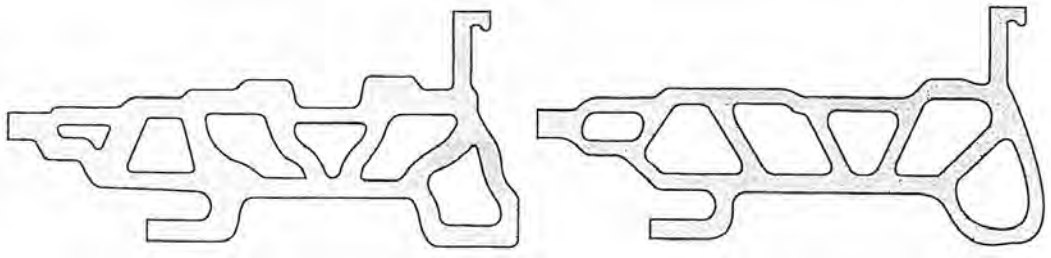


Figure 7.2.29 Design for uniform thickness in sections. (Courtesy of Meehanite Metal Corp.)

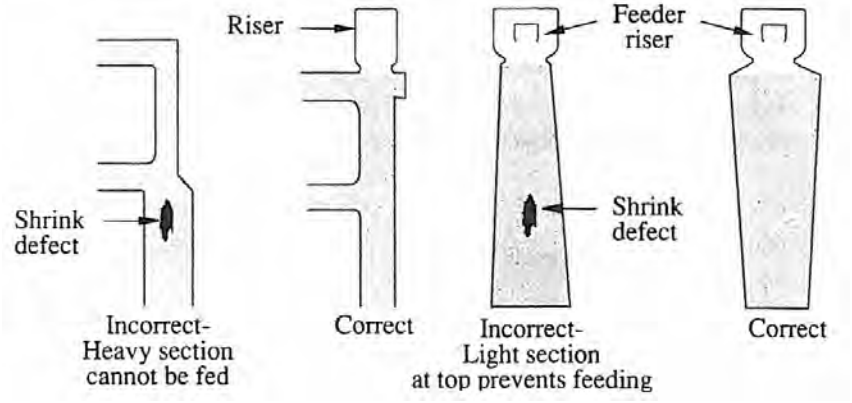


Figure 7.2.30 More intersection details. (Courtesy of Meehanite Metal Corp.)

Figure 7.2.31 Design for bolting or bearing bosses. (Courtesy of Meehanite Metal Corp.)

die casting is the metals analogy of injection molded plastics. The most common die cast parts are made from aluminum or magnesium and include household items (e.g., two-cycle gasoline engine housings with many thin, closely spaced ribs for air cooling).

Plaster Mold Casting

Plaster mold casting is similar to permanent mold casting, but it involves the use of a plaster mold whose thermal characteristics can be varied to suit the part shape and alloy. As a result it is possible to cast parts with very fine detail and little or no warpage or residual stresses. For example, pump impellers are routinely plaster cast with tolerances of 0.13 mm (0.005 in.) held over a range of 15 cm (6 in.). Even some types of gears can be cast that do not require any machining of their teeth. Because the plaster can be dissolved, far more complex shapes can be made than with permanent mold or die castings. However, plaster mold casting is generally limited to nonferrous metals that have a melting temperature below 1000 ^{o}C (1900^{o}F). Beryllium-copper alloys are often plaster mold cast, and after appropriate heat treating can attain yield strengths on the order of 1.17 GPa (170 ksi).

Investment Casting

Investment casting involves making a wax mold of the part and then coating it with a refractory ceramic. The wax is then melted out of the mold and the cavity can be filled with molten metal. Because a ceramic forms the mold, virtually any metal can be investment cast. Properties of investment castings are similar to those obtained with die or plaster mold casting.

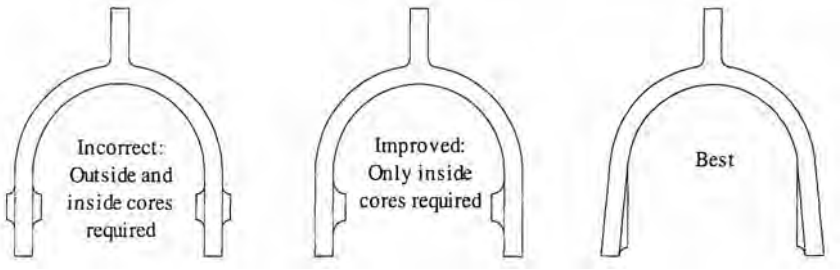


Figure 7.2.32 Omit outside bosses and the need for cores. (Courtesy of Meehanite Metal Corp.)

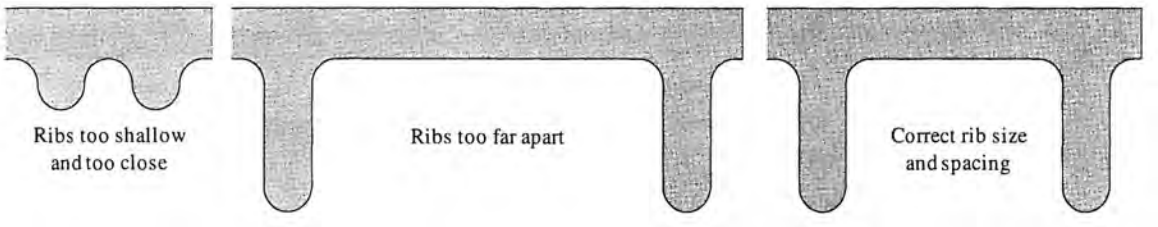


Figure 7.2.33 Proper rib size and distribution. (Courtesy of Meehanite Metal Corp.)

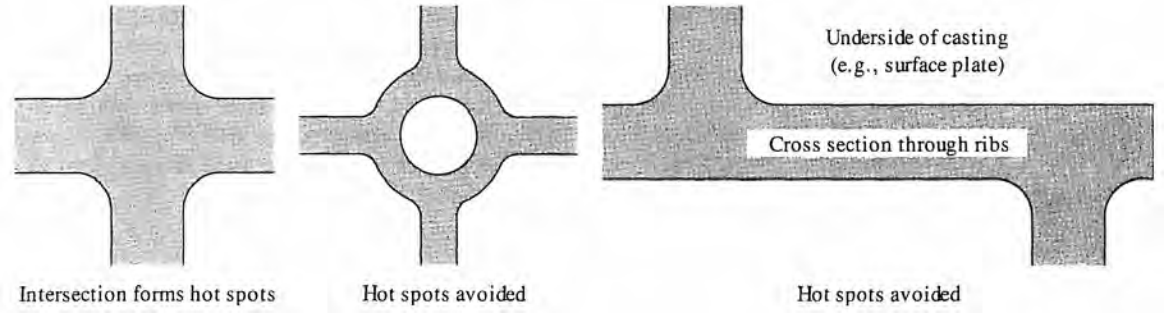


Figure 7.2.34 Avoid cross ribbing, and if necessary, make ribs for Tees or meet at a roundabout. (Courtesy of Meehanite Metal Corp.)

Centrifugal Casting

Centrifugal casting involves mounting molds on a fixture and then spinning the fixture at high speed to cause molten metal to be forced into the recesses of the molds; hence light impurities such as gas bubbles or dirt migrate to the center of the fixture. Thus a casting with superior metallurgical and strength properties can be produced compared to one made from a sand casting.

7.2.2.2 Polymer Concrete Castings[14]

Portland cement-based concrete is not dimensionally stable enough, due to its own internal structural variations with time and its hygroscopic nature, to allow it to be used for the main structure of a precision machine tool. Although in many applications, properly cured reinforced concrete on a stable, dry subgrade can provide a reasonably stable foundation for very large machines that are not self-supporting, unreinforced Portland cement concrete itself is not dimensionally stable due to: (1) reaction shrinkage from cement hydration, (2) shrinkage due to loss of excess nonstoichiometric water, which leaves conduits for humidity-induced expansion or contraction, and (3) nonelastic di-

[14] See, for example, T. Capuano, "Polymer Concrete," Mach. Des., Sept. 10, 1987, pp. 133–135; J. Jablonowski, "New Ways to Build Machine Structures," Am. Mach. Automat. Manuf., Aug. 1987, pp. 88–94; and P. A. McKeown and G. H. Morgan, "Epoxy Granite: A Structural Material for Precision Machines," Precis. Eng., Vol. 1, No. 4, 1979, pp. 227–229.

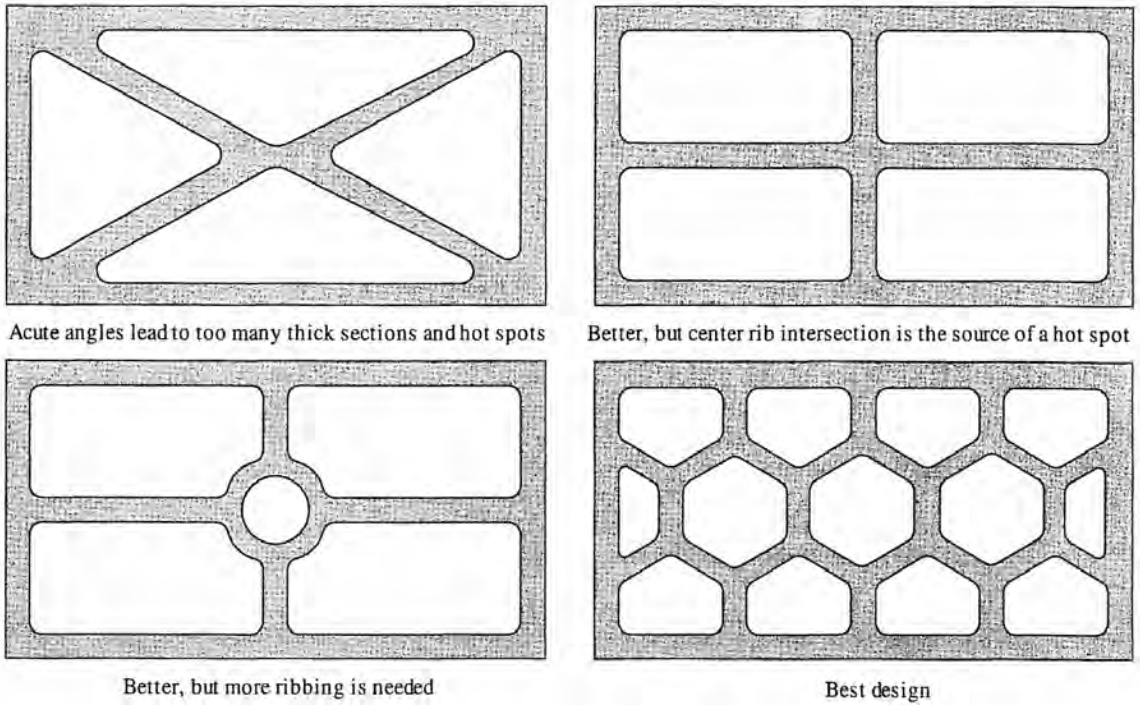


Figure 7.2.35 Avoid using ribs which meet at acute angles. (Courtesy of Meehanite Metal Corp.)

mensional changes (e.g., creep and microcracking in the inherent brittle/porous structure). Overall, strain variations with time may be as high as 1000 μm/m.[15]

Fortunately, a number of different types of polymer-based concretes have been developed which can be used to cast machine tool quality structures. For example, Fritz Studer AG, a prominent manufacturer of precision grinders in Switzerland, discovered that special polymers can be used to bind together specially prepared and sized aggregate to yield a stable, strong, high-rigidity material with a damping coefficient much higher than that of cast iron[16] as discussed in Section 7.3.3. By carefully controlling the manufacturing process and selection of binder and aggregate, properties can be varied somewhat to suit the user. The polymer concrete material and process developed by Studer is known as Granitan® and its composition and manufacturing process was patented. Numerous companies have licensed the process and will make castings from Granitan® to order. Other companies have developed their own proprietary polymer concretes with similar high performance properties.

For polymer concrete castings, the same rules for draft allowance apply as for metal castings if the mold is to be removed. Instead of ribs, polymer concrete structures usually use internal foam cores to maximize their stiffness-to-weight ratio. Unlike metal castings, a polymer concrete casting will not develop hot spots while curing even in thick, uneven sections. Polymer concrete castings can readily accommodate cast in place components such as bolt inserts, conduit, bearing rails, hydraulic lines, etc. as shown in Figure 7.2.36. Figure 7.2.37 shows the cross section through a threaded insert. It should be noted that the bolt will fail before the insert. Because the material is made from an aggregate, one must pay careful attention to the thickness of cross sections to ensure that they are one to two orders of magnitude thicker than the typical dimension of the aggregate used. This will help the cast part behave as if it were made of an isotropic material.

[15] From correspondence with Jack Kane, Gandalf Inc., 206 San Jose Drive, Dunedin, FL 34698. The author would like to thank Jack for his help in preparing this section. See S. H. Kosmatka and W. C. Panarese, Design and Control of Concrete Mixtures, 13th Edition, Portland Cement Assoc., 5420 Old Orchard Road, Skokie, IL 60077-1083, pp. 151–160.

[16] R. Kreienbühl, "Experience with Synthetic Granite for High Precision Machines," Proc. Symp. Mineralguss im Maschinenbau, FH Darmstadt, Sept. 19-20, 1990; and H.J. Renker "Stone Based Structural Materials," Precis. Eng., Vol. 7, No. 3, 1985, pp. 161–164.

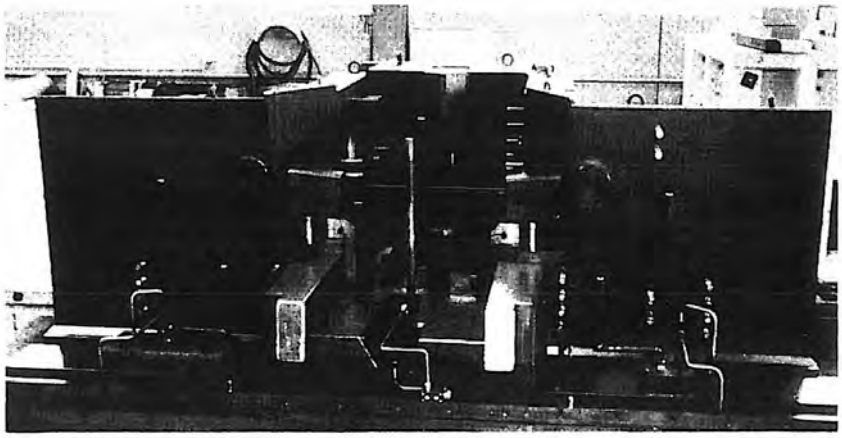


Figure 7.2.36 Cast-in inserts in polymer concrete casting. (Courtesy of Fritz Studer AG.)

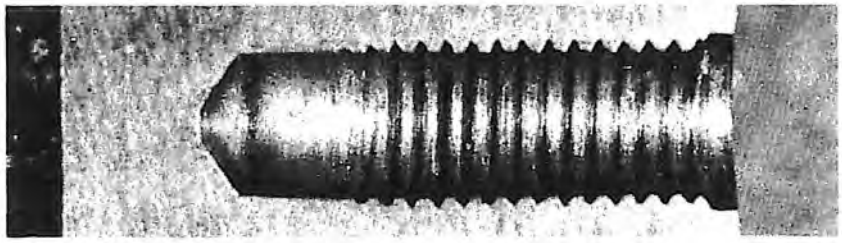


Figure 7.2.37 A threaded insert in a polymer concrete casting. (Courtesy of Gandalf Inc.)

For some very large machines, especially one-of-a-kind machines, it makes sense to make a mold from thin welded steel plate that remains an integral part of the machine after the polymer concrete is poured in. However, one should be careful to consider the effects of differential thermal expansion when designing the steel shell, and the steel shell should be fully annealed after it is welded together. One should note that a continuous steel skin serves as a vibration conduit between members attached to the structure. If metal components are bolted or cast into a polymer concrete bed, maximum vibration isolation will be achieved only if they are separated by the polymer concrete.

One might think that the polymer binder cannot have the long-term dimensional stability of for example cast iron. Fortunately, properly formulated polymers can have good stability, and even though polymers are known to creep, the strength and stability of polymer concrete seem to be achieved primarily by the physical contact between the stable aggregate particles with the polymer acting primarily as a binder. It is recommended here, however, that before a polymer concrete is specified for use in a machine with greater accuracy requirements than for which the polymer concrete had been used previously, a sample structure should be tested for long-term dimensional stability.

For any nonmetallic structural material intended for precision applications, the issue always arises concerning the absorption of water, which can cause swelling and dimensional instability in some materials; however, epoxy aggregate castings can made so they are immune to water and humidity by carefully selecting the epoxy, by making sure that no hydrophyllic components are used in the mix, and by coating the casting with a hydrophobic material. A time and cost saving technique is the in-mold-coating (IMC) technique. With this technique, a combined release agent and coating material are sprayed into the mold prior to casting. When the casting is removed from the mold, it has the desired surface finish and coating and the casting is ready to use.

With appropriate section design, polymer concrete structures can have the stiffness of cast iron structures and much greater damping than cast iron structures.[17] However, due to polymer concretes' lower strength, heavily loaded machine substructures (e.g., carriages) are still best made from cast iron. As can be seen in Table 7.3.1, polymer concretes do not diffuse heat as well as cast iron structures and thus attention must be paid to the isolation of heat sources to prevent the formation of hot spots in a polymer concrete structure. Also, when bolting or grouting nonepoxy granite components to an epoxy granite bed, one must consider the bimaterial effect described in Section 2.3.5; and note that grout can have a coefficient of thermal expansion (on the order of 40 $\mu m/m/C^o$) many times that of other structural materials.

Pound for pound, on average a polymer concrete casting will consume one-third less energy in its preparation and will cost 30-40% less than an iron casting.[18] In addition, due to decreased vibration, tool life is extended and part quality increases. Thus in the future, as energy prices and accuracy requirements rise, it is likely that more and more machines will utilize polymer concrete components, particularly for large sections such as the machine base.

7.2.2.3 Powder Materials Technology[19]

Shapes as simple as washers to those as complex as ceramic turbine compressor rotors can be produced by powder materials technologies. For cold isostatic pressing, a die or mold is filled (charged) with the powdered material and an organic binder. The material is then compacted under pressures ranging from 0.1 to 1.0 GPa (15 - 150 ksi). The resulting green part can easily be machined using carbide, ceramic, or diamond tools. The part is then fired in a kiln. Final dimensional accuracy must be obtained by grinding or lapping. Ceramics are often processed in this manner. For hot isostatic pressing, heat and pressure are applied and the little particles are forced into intimate contact and then fused together.

Unlike casting technology, CIP or HIP parts' grain size can be controlled more easily because the material does not actually melt and then solidify. In addition, hot spots do not form and the porosity can be better controlled. Thus parts can be made so they will absorb liquids via capillary action. In this manner self-lubricating oil-impregnated bearings or metal matrix composites can easily be produced. Subsequently, the parts can be machined or ground.

As ceramic materials technology advances, powder materials technology will become even more commonplace. Consequently, a whole new market is likely to form for ultraprecision machine tools for processing abrasive ceramic materials.

7.2.3 Material Removal Processes

Often, factors such as lot size, accuracy, surface finish, and size will lead to the conclusion that the best way to make the part is to remove excess metal from a blank or to finish machine a near net shape part. The following subsections briefly discuss the basic properties of metal removal processes used in the production of precision machines.

Broaching

Broaching is an inexpensive process that imparts the mirror image of the tool form into the part by making successive small cuts with a gradually changing tool form. For example, a square hole can be made from a round one with the use of a broach whose shape gradually changes from round to square, or a keyway can be cut on the inside of a round hole. The shape change is discrete, with each level being a sharp cutting surface. Unlike turning or milling, in which a single cutting edge continually removes metal from a surface, each tooth on the broach cuts, accumulates, and

[17] See, for example, M. Weck and R. Hartel "Design, Manufacture, and Testing of Precision Machines with Essential Polymer Concrete Components," Precis. Eng., Vol. 7, No. 3, 1985, pp. 165–170; and I. Salje et al., "Comparison of Machine Tool Elements Made of Polymer Concrete and Cast Iron," Ann. of the CIRP, Vol. 37, No. 1, 1988, pp. 381–384.

[18] From conversations with Jack Kane, President, Gandalf Inc., 206 San Jose Drive, Dunedin, Florida 34698. The author would like to thank Jack for his help in reviewing this section. Also see R. Kreienbühl, "Granitan - 10 years of Experiences in Machine Tool Application," Proc. Symp. Reaktionsbeton im Maschinenbau, Techn. Hochschule Darmstadt, March 10–11, 1988.

[19] For an extensive discussion of powder materials technology, see, for example, the Metals Handbook, 9th ed., Vol 7, American Society for Metals, Metals Park, OH. Continuous up-to-date information can also be obtained from the Metal Powder Industries Federation, 105 College Road East, Princeton, NJ 08540. MPIF also publishes a useful quick reference entitled P/M Design Guidebook.

carries out the material from the operation. Once the operation is complete, the material is then removed from the broach teeth (e.g., with a high-pressure air jet). Broaches can be designed to cut shapes with linear or circular forms (e.g., high-helix-ID threads). In the latter case, the broach is rotated as it is fed linearly into the part. Because material is removed by cutting, very high accuracies (1-10 μm) and surface finishes (1 μm R_a or better) can be achieved. Automated broaching machines can produce hundreds of parts per hour, and NC broaches can produce contoured shapes by rotating the tool as it moves axially (e.g., for the rifling on a gun barrel). Broaching is a process that is widely employed to cut splines (blind and through) and other intricate shapes as part of a final finishing operation on parts formed by a variety of other processes.

Drilling, Reaming, Boring, and Tapping

Holes must be formed in parts for shafts, pins, bolts, fluid flow passages, and so on. Since holes are often used to support power transmission components, a machine's dynamic performance is often directly related to the accuracy at which the holes were placed and the accuracy of their size.

Drilling is an inexpensive process that is used to make a hole when accuracies on the order of 1 part in 10^2 (depending on the depth) are adequate.[20] The longer the hole, the more the drill bit wanders. For deep holes, a constant force feed is desired, as opposed to constant feed that most NC machines provide. There are innumerable types of drills from the well-known spiral form to the single straight fluted gun drill used for boring holes with very large depth-to-diameter ratios (e.g., for gun barrels). A design engineer should never be shy about wanting to specify a hole in a part; he must merely consult with the manufacturing department about what tooling is available. If drilling itself is not accurate enough, it is often used as a means of roughing out a hole that will then be cleaned up by reaming, single-point boring, honing, or internal grinding.

A reamer is a tool that is used to slightly (on the order of 1-10%) enlarge a drilled hole. Any time an interference fit part is to be used, the drilled hole in which the part fits should be reamed. Reaming can attain accuracies on the order of 1 part in 10^3 to 10^4 and surface finishes on the order of 1/2 μm R_a. Skewness is typically -0.5 to -1.0.

A drill or reamer's accuracy is limited because many cutting edges are used to form the hole. Single-point boring, on the other hand, utilizes only one cutting edge whose radial position is carefully controlled (either fixed or NC controlled). The boring tool can be fixed and the part rotated, or vice versa. With boring, the accuracy of the hole that is formed can be on the order of the accuracy of the machine used in the boring process. Holes from a few millimeters to many meters in diameter can readily be bored. Boring can provide the high location and size accuracy and surface finish required for an internal surface of revolution.

Because boring is such an accurate hole-forming process, it should be specified for precision bearing bores used to support spindles, leadscrews, motor shafts, and other important powertrain components. When the distance between two in-line bores is not too great (less than 5-10 hole diameters), it is important that they be formed by a process known as *line boring*. Line boring ensures that the holes are of the same diameter and that their centerlines are parallel and coincident. In line boring, the boring tool bores the first hole and is fed into the part until it bores the second hole also. Because the length of the boring tool projecting from the spindle is constant (the part or the spindle itself moves), the deflection of the boring bar will be constant and the holes will have nearly identical shapes and centerlines (assuming that the part and fixturing is reasonably rigid). Hence when the bearings are pressed into the bores, the shaft they support can be supported with minimal forced geometric congruence. When parts are line bored, however, one must ensure that there are appropriate reference surfaces with respect to which to locate the bores (e.g., the bores would have to be parallel to the axis of motion of a slide). In some cases line boring is used to bore a single hole through a stack of flat plates to ensure that when the plates are spaced apart, the holes all line up (e.g., boring three holes into two plates that are later separated by a spacer and then the assembly rides on three rails). When the boring bar might be too long to allow for line boring, one hole can be bored, the part rotated 180° on a precision index table, and then the other hole bored. Once again, reference surfaces must be provided for on the part to ensure that the centerline of the hole is located accurately with respect to the part.

Tapping is the process used to cut or form threads into a hole. The former utilizes a tool that cuts the metal away (a fluted tap), while the latter uses a tool that cold forms the metal (a

[20] See Machinery's Handbook, published by the Industrial Press, New York.

fluteless tap). Cutting threads requires less torque on the tool and thus lessens the chance of breakage. Forming the threads requires twice the torque but leaves the threads with a grain structure that is much stronger. Since stiffness is most often required for machine tools, fluted tapped holes (which are generally the assumed type when one specifies a hole) are most often used.

Although a seemingly mundane detail, tapped holes are among the most often improperly designed details. A common mistake is to assume that the threads can be tapped to the bottom of the hole or that a bolt can be threaded to the bottom of a hole. In the former case, a bottom tap can be used to tap threads to within a quarter or so of the bolt diameter; however, a bottom tap cannot be used to start a hole, and thus bottom tapping requires a tool change. In the latter case, if the bolt extends to near the bottom of the hole, then higher precision is imposed on the bolt, washer, and bolted part. If a hole can be drilled through the section of the part so that chips can be blown through, or drilled deep, to allow for plenty of excess thread and a reservoir for chips to accumulate during the tapping process, then the design engineer should detail the part accordingly.

Another critical issue in the manufacture of threads is the stiffness of the threadform. A conventional bolt threadform (e.g., Acme) is symmetric and may require the use of a thread locking compound or a lock washer to keep the fastener from coming loose under vibratory conditions. Inaccuracies in the threadform make the effective threaded length of a bolt typically less than one diameter. To achieve this effective length, a significant amount of torque must often be applied. The use of a hardening thread lubricant can help to ensure that all threads transfer loads. The hardening thread lubricant (e.g., epoxy) acts to replicate the threads of the bolt in the threaded bore.

A Spiralock® threadform has a land at the root of the threadform. The angle of the land is small with respect to the hole's longitudinal axis, so axial motion of the bolt creates very high normal forces between the contact regions of the bolt's conventional thread and the Spiralock® thread. As a result, local yielding of the bolt threads at the contact points occurs, which causes the metal to flow, creating line contact along the thread helix. This type of thread has greater stiffness and resistance to loosening[21] than does a plain Acme thread; however, the taps cost more than conventional taps and many shops do not stock Spiralock® taps.

Chemical Machining

There are many different forms of chemical machining, including photoforming, chemical milling, and electrochemical machining and grinding. Photoforming is used to remove small amounts of material from parts that cannot be machined, either because the material is too hard or the pattern too complex. It is used to make parts too small or intricate for stamping processes, to etch lines on printed circuits and some encoder disks, and to etch the patterns for integrated circuits on silicon wafers. Chemical machining works by covering the regions not to be machined with a mask and then placing the part in the etching solution. Typically, a photosensitive material is used for the mask. The material is coated with the photoresist and a pattern is optically projected onto the part. The photoresist is developed and the unexposed areas are dissolved, along with the metal beneath them by dipping in a special acid or alkali bath. The process is fairly automated, so small lot sizes can be processed economically.

Chemical milling is used to remove metal from unmasked areas up to a depth of about 1-2 cm. It is not as nearly as accurate as conventional milling, but is useful for making noncritical dimensioned holes, pockets, and ribs in large heavy parts that would be difficult to fixture on a conventional machine.

Electrochemical machining is the chemical analog of broaching. Instead of the tool cutting material as it is advanced into the workpiece, the tool acts as a cathode to eat away at the metal. The part is submerged in an electrolytic bath and the negative terminal of a high-current low-voltage dc power supply is connected to the tool while the positive terminal is connected to the part. As the tool is fed into the part, the metal dissolves via a reverse plating process (electrolysis) and the mirror image of the tool is formed in the part. Virtually any shape can be imparted into virtually any metal with this process and there is little or no tool wear. However, this is not a precision process on the order of boring or grinding.

Electrochemical grinding uses a special grinding wheel where the electrochemical action removes 90% of the metal and the grinding wheel removes the remaining 10%. It is relatively easy to convert a conventional grinder, but the process is not commonly used in small shops. This process

[21] See H. Holmes, "A Spiral Lock for Threaded Fasteners," Mech. Eng., May 1988, pp. 36–39.

is used to grind parts that are too delicate to be subjected to the heat and forces of conventional grinding. In the quest to ductility grind glass at an economical removal rate to obtain a finish that does not require subsequent polishing, it may be found that chemical grinding is the best method.[22]

Electrodischarge Machining

In electrodischarge machining (EDM) the part is placed in a pool of nonconducting dielectric fluid (e.g., oil) and a tool of the desired shape is pushed into the part. When the tool comes in close proximity to the workpiece, numerous tiny electric arcs span the small gap thousands of times per second. Each arc vaporizes a small amount of metal and the resulting thermal shock often dislodges additional small particles. When magnified the surface appears to be covered with many microscopic craters. The rapid quenching of the surface can leave a thin hard brittle layer on hardenable materials, which may be undesirable for some parts. Under these circumstances the surface may be acid etched to remove the hard layer.

Like chemical machining processes, the EDM process is insensitive to the hardness of the metal; thus it is a very useful process for cutting shapes in hardened steel parts without having to worry about loss of temper or the introduction of deep residual stresses, although the cratered surface finish is generally not acceptable as a precision bearing surface. The EDM tool can be of virtually any shape, including a taut wire that can be used to cut out intricate shapes (wire EDM). The EDM process allows for virtually any part geometry to be achieved easily while easily maintaining accuracy on the order of 1 part in 10^3 to 10^4, depending on the type of electrode used. Very high relative accuracies can be obtained by using the mating part of an assembly as an electrode for final finishing of the rough-machined mating part (e.g., in dies for automobile body parts). EDM surface finishes may vary from 1 to 10 μm R_a (or more), depending on the electrode used. Skewness is typically 0.0-1.2. The EDM process is often used to make elaborate die sets that are used to stamp complex metal parts (e.g., motor laminations and razor blades). However, one must still contend with manufacture of the electrode, but a least the electrode can be assembled from many discrete parts (e.g., a collection of simple-to-make pins used to form hundreds of holes in a part simultaneously).

Grinding

Grinding is perhaps the most accurate of the common manufacturing processes and is one of the most used processes in the manufacture of precision machine components. Most people think of grinding as a process used to create flat surfaces or round objects, but grinding is often used to create precision contours in many different types of parts, including bearing raceways and ballscrews.

Most other metal cutting machines use a discrete number of cutting surfaces, but in grinding thousands of tiny cutting surfaces on the grinding wheel produce an averaging effect as they pass across the surface of the metal. Furthermore, the materials used to make grinding wheels can be orders of magnitude harder than conventional cutting tools, so the effect of tool wear on accuracy can be greatly reduced. Also, it is possible to dress (sharpen and shape) a grinding wheel continuously (or frequently) so that machine downtime for tool changing can be minimized. Virtually any part can be ground to precision tolerances,[23] and often in high volumes at a very economical rate.

There are two broad categories of grinding, conventional grinding and creep-feed grinding. The former is the best known and is used as a cleanup operation to remove only a little metal (tens of microns to a few millimeters at most). Most people who have been in a machine shop have seen a surface grinder moving rapidly back and forth across a part under a flood of coolant. With each pass, the wheel is indexed into the work a few microns or so. Other types of grinding processes that also feed into the work a few microns at a time can be used to grind virtually any shape into any surface (e.g., cylindrical ID and OD grinding of bearing raceways, and jig grinding complex forms into dies). Figure 7.2.38 shows some common wheel forms. It is important to realize that with the use of diamond tools to dress the grinding wheel, virtually any profile can be imparted to the grinding wheel. In creep-feed grinding, continuous wheel dressing techniques allow complete forms to be cut into a blank piece of material with depth of cuts that may be on the order of centimeters. Continuous wheel dressing, extremely high coolant flood rates, and high axis actuation forces are required for creep-feed grinding.

[22] From discussions with Kevin Lindsey of NPL. See Lindsey's U.K. Patent 8,928,299, Methods of Preparation of Surfaces and Applications Thereof, Dec. 14, 1989. Also see O. Podzimek, "Residual Stress and Deformation Energy Under Ground Surfaces of Brittle Solids," Ann. CIRP, Vol. 35, No. 1, 1986, pp. 397–400.

[23] See R. Moore and F. Victory, Holes, Contours, and Surfaces, Moore Special Tool Co., Bridgeport, CT.

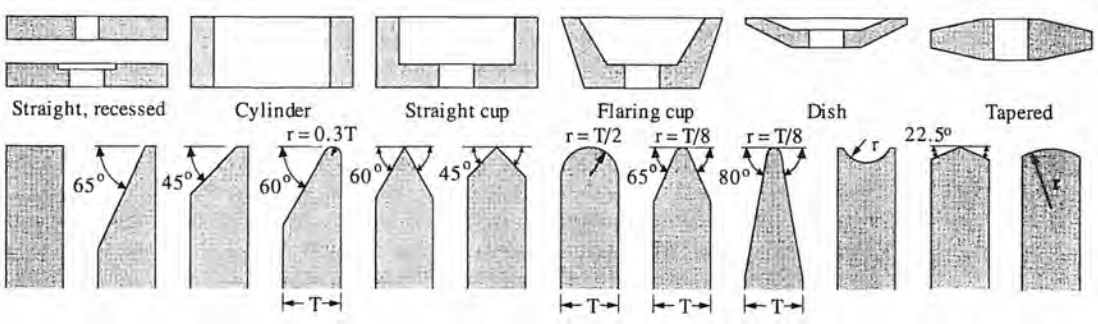


Figure 7.2.38 Common standard grinding wheel shapes and contours.

Accuracy of the grinding process is a function of many factors, including the accuracy of the machine, the dressing of the wheel, the temperature control of the coolant, the fixturing method, and the feed rate. When these factors are properly controlled, accuracies on the order of 1 part in 10^4 to 10^5 are possible. Surface finishes produced by precision grinding are typically 1/4 μm R_a and can be as good as 1/20 μm R_a. Skewness is typically 0 to 0.8. Specifically, for the following materials,[24] obtainable surface finishes are:

Hardened steel	0.05 μm R_a (2 μin.)	Brass	0.30 μm R_a (12 μin.)
Soft steel	0.13 μm R_a (5 μin.)	Aluminum	0.38 μm R_a (15 μin.)
Cast iron	0.13 μm R_a (5 μin.)		

The best surface finish is produced when the wheel retraces its path over the part without being fed into the part, until sparks are no longer produced. This is referred to as *spark-out*.

The surface produced by grinding has many sharp microscopic peaks and valleys, caused by the sharp abrasive particles on the wheel. Steel parts will quickly oxidize unless they are protected, and will be most susceptible to fretting during contact with other steel surfaces unless they are well lubricated.[25] In addition, one must realize that when designing an ultraprecision machine, surface peaks and valleys formed by grinding will emulate the ride produced by driving on a cobblestone street; thus any bearings that make contact with the surface must do so in the presence of these features. Either an elastic averaging effect is needed, or another process such as honing, lapping, or superfinishing (see sections below), is needed to smooth out the peaks and valleys.

There also will always be situations where it will be difficult to achieve the accuracy one desires for large parts without spending one's entire budget. Hence one should always be on the watch for ways to make symmetry work for you. Examine each part in the assembly and ask yourself, for example, "can a long part be from two shorter parts made side by side?" Or "can two parts that are supposed to mate be ground together to ensure their fit?" This process, called *match grinding*, does not necessarily allow for interchangeability of parts, but does allow for interchangeability of sets of parts.

On a final note, the design engineer should be sure to minimize the cost of the grinding operation by using generous undercuts in corners whenever possible to give the grinding wheel plenty of room, as was shown in Figure 7.2.19. Grinding sharp corners is difficult and leads to stress concentrations in the part.

Honing

Honing is an even more accurate process than grinding because it involves area contact of the abrasive surface rather than line contact; hence an even greater averaging effect is obtained. As illustrated in Figure 7.2.39, the honing tool rotates and reciprocates with respect to the workpiece, so the peaks and valleys formed by grinding are eliminated. The resulting surface finish is crosshatched, and can be made to optimize lubrication system performance in systems such as reciprocating engines. Although honing can do wonders for surface finish and shape (i.e., roundness or flatness), it

[24] From R. Bolz, Production Processes: The Productivity Handbook, 5th ed., Industrial Press, New York, pp. 20–26 to 20–27. Values were originally quoted in terms of the root-mean-square values. Approximate equivalent R_a values are given here. See Equation 7.3.2.

[25] See the discussion on fretting in Sections 5.6 and 7.7.

does not correct for location errors (e.g., the position of a bore in a part). Hence the machining or grinding process used prior to honing must adequately establish the position of the hole. Honing is most often specified for round surfaces, but can be used for flat surfaces also. Size accuracy on the order of 1 part in 10^4 can be obtained with surface finishes on the order of 0.5-0.01 μm R_a. Skewness is typically -0.5 to -1.0. The total amount of material that may be removed during the honing process is typically less than 0.01-0.1 mm, but can be as much as 1 mm.

A variation on honing is superfinishing, which utilizes lighter pressures, more rapid reciprocating motion, and a flood of coolant. The result is a surface finish on the order of 0.1-0.01 μm, which increases fatigue and bearing surface life (e.g., the wear surfaces of camshafts and crankshafts are often superfinished).

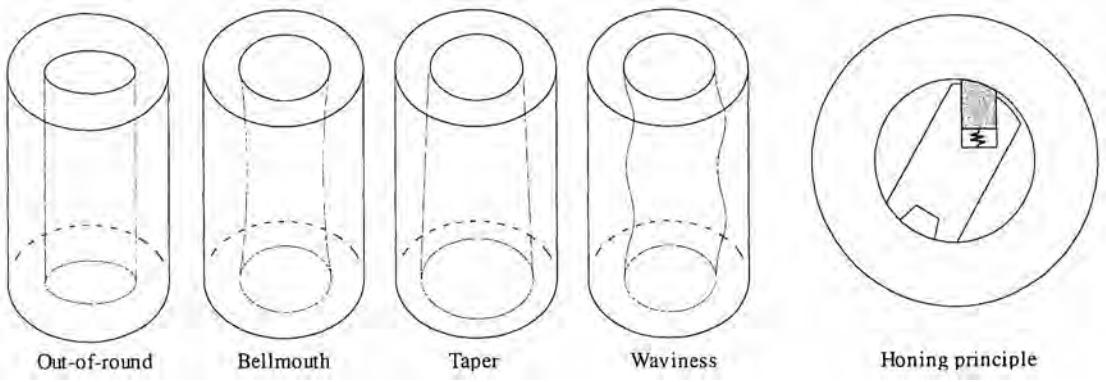


Figure 7.2.39 The honing process generates true roundness through the rotating and reciprocating motion of a tool that is composed of one or more abrasive stones and unequally spaced guide shoes. This process can rapidly and automatically correct for out-of-roundness bellmouth, taper, and waviness.

Lapping

Lapping is an even more accurate process than honing, because lapping uses a random motion pattern along with a much finer abrasive that is "charged" into a material (the *lap*) that is softer than the part being lapped. The abrasive particles cut the surface being lapped, but only embed themselves in the softer surface. With this process, only about 10 μm of material is typically removed, but the best possible accuracy is obtainable. Hence lapping is usually used only to remove fine scratches left by grinding or honing and/or to make fine corrections to increase dimensional accuracy (e.g., straightness, flatness, roundness). Brittle materials (e.g., ceramics) tend to lap better than ductile materials (e.g., steel) because the lapping compound seems to cut the material without causing any plastic deformation (smearing).

The lap can be hand finished (e.g., scraped cast iron) to the desired shape (e.g., a flat plane) and then used to correct the geometry of another surface (e.g., hardened steel bearing rails). In addition, small hand laps can be used to correct local irregularities. Given time, a skilled person with a hand lap and accurate measuring means can achieve virtually any desired accuracy and smoothness. Note that even a perfectly straight hardened and ground bearing rail bolted to a perfectly straight scraped cast iron bed can be deformed by the mounting bolts. The bearing rail can be lapped flat or the error compensated for on NC machines with error mapping; however, note that in either case, the construction of the machine must be such that mounting stresses will not relax with time and cause the geometry to change.

Various forms of automated lapping machines also exist for lapping optical components and round parts such as ball bearings. Because lapping is a slow process, lapping should only be specified in critical applications requiring exceptional accuracy (better than 1-5 μm) where grinding and honing are no longer effective. Before specifying lapping, however, one should check with the shop to see what can and cannot be done.

Milling

Milling is a process whereby metal is removed from the part by a rotating cutter. Most engineers are familiar with this process and all that one must realize is that virtually any shape can be milled given the right shape cutters and the required controlled *relative* degrees of motion between the part and the cutter (e.g., two-, three-, four-, and five-axis machining centers). Accuracy of the milling process is on the order of 1 part in 10^3 to 10^4, and surface finishes on the order of 1 μm R_a can be obtained, depending on the type of tool, material, and feed rate. Skewness is typically 0.2 to -1.6. Because milling is such a common process of extreme usefulness, no more will be said about it here, and those that wish to know more should go visit their local machine shop.

Scraping

Scraping is perhaps the most historically significant manufacturing process because it is the basic manufacturing process of all machine tools. Before any machine tools existed, there had to be a way to form flat surfaces to act as reference planes (i.e., surface plates). From a flat surface, one could then generate straightedges and right-angled surfaces.[26] Once a set of master references is made, other machines and references can be made from it and the process cascades to the point where we are now; large precision machines are available to automate the production of other precision machines. Note that cast iron is the easiest ferrous metal that can be scraped (it is possible to scrape steel weldments), which coincides well with the development of metallurgical and mechanical technologies. Brass, polymers, and many other soft materials can also be scraped.

Cast iron has numerous graphite flakes in its structure which allow small quantities of metal to be removed from the surface, just like scraping paint off of a window. To identify areas to be scraped, a thin layer of colored compound (e.g., red rouge) is applied to the part and it is pressed and rubbed on the reference surface until the high spots on the part wear off the colored compound (this is called *giving up a bearing*). Through progressive scrapings and bearings, a skilled scraper can correct for form (e.g., flatness) and surface irregularities as shown in Figure 7.2.40. Scraping can be used, for example, to match two surfaces that are to be bolted together. Scraping is also used to impart surface texture to bearing surfaces so that they will retain a lubricant better.

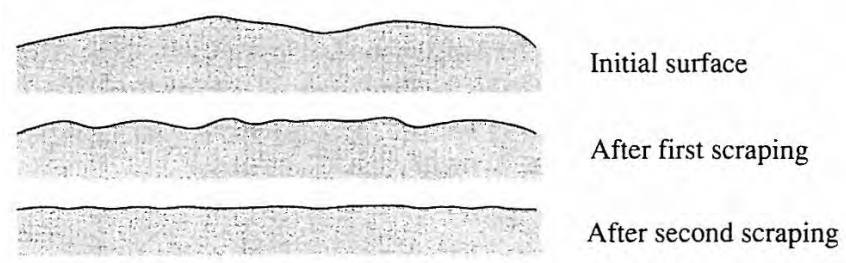


Figure 7.2.40 Progressive scraping can correct for form and minimize the distance between high spots (bearing points).

A flat plane can be created by a multicomparative process that starts with three surfaces, A, B, and C. To remove twist, one surface must be rotated 90° and a new bearing made. For flatness greater than about 1/2 μm, other more localized measuring techniques (e.g., measurement of a grid of straightness lines with an autocollimator) are often used. Figure 7.2.41 shows the process for using a surface plate to make a straightedge. One side of the straightedge blank is first scraped flat using the surface plate as a reference. Side 2 is scraped flat using the surface plate as a reference. Parallelism to side 1 is achieved by using a height gage. Side 3 is also scraped flat using the surface plate as a reference, and perpendicularity to sides 1 and 2 is checked by flipping the straightedge end for end (so that side 1 is on top) and checking it against a lateral position-measuring tool referenced of the bottom edge of side 3. Side 4 can then be made parallel to side 3 in the same manner that side 2 was made parallel to side 1. With continual iteration, patience, and skill, virtually any desired straightness and squareness can be achieved within boundaries set by gravitational deformations.

[26] Scraping is a process that should be used only in the most critical applications. A good reference describing the "how to" of the process is T. Busch, Fundamentals of Dimensional Metrology, Delmar Publishers, Albany, NY, 1964. A pictorial essay of how the components of a precision machine are scraped and lapped to fit is given by W. Moore, Foundations of Mechanical Accuracy, Moore Special Tool Co., Bridgeport, CT.

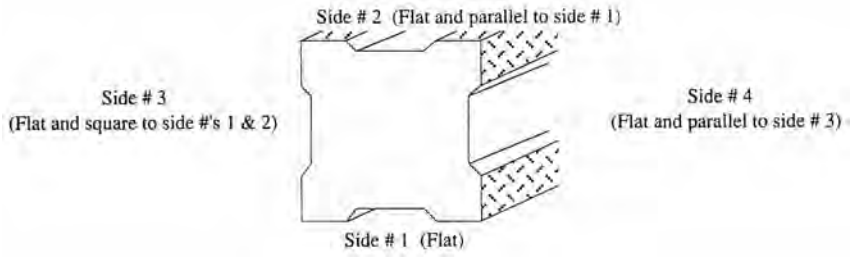


Figure 7.2.41 Process for scraping a straightedge. (After Moore.)

The means to make a flat plane and a straightedge can be combined with methods for graduating length scales[27] to provide the basis for most types of precision machines. Thus it is no wonder that the demise of precision scraping is self-induced: As better and better machines were built by scraping, they were able to make other machines that previously had to be scraped. More recently, the introduction of zero-shrink epoxies for the process of replication is reducing the need for surfaces to be scraped. It is also becoming more common to locate surfaces kinematically and then fill the regions between them with special epoxy. Scraping will never die, but it may fade away asymptotically. However, note that skilled scrapers can compensate for future deflections under load, and this skill is hard to program into a machine.

Single Point Diamond Machining[28]

On a microscopic scale most cutting tools have jagged edges that tear and rip the surface of a part. On the other hand, by comparison a diamond tool can have a smooth, sharp edge that allows it to cut a surface and leave a surface finish on the order of nanometers. Accuracy depends on the characteristics of the machine, but generally, diamond turning centers are the most accurate machines available, and some can attain shape accuracies on the order of 1 part in 10^5 to 10^6. However, one should note that ferrous alloys cannot be machined (except in some cases in a pure methane atmosphere) and thus must be plated (e.g., with electroless nickel) and only the plating itself machined. Diamond machinable materials include aluminum, brass, copper, beryllium, copper, gold, silver, lead, platinum, cadmium teluride, electroless plated nickel, alkali halides (salts), zinc sulfide, zinc selenide, germanium, most plastics, and machinable ceramics.[29] Diamond machining has been instrumental in the economical manufacture of precision components for many common items, such as contact lenses, photocopy machines, laser printers, computer hard disks, and video recorders.

Turning

Turning is a process where metal is removed by rotating the part with respect to a tool whose position is controlled along the radial and axial directions of the part. Like milling, it is one of the most useful and common processes and can attain accuracies on the order of 1 part in 10^3 to 10^4 and surface finishes to 1/2–14 μm R_a. Typical skewness is 0.2 to 1.0. One of the greatest advances in turning centers has been the addition of live tooling. Live tooling is a rotating cutting tool (e.g., an end mill or drill) that can be used to cut slots, holes, and so on, in a part while it is still held in the lathe's spindle. In this manner, other operations are done on the part while it is still fixtured, thereby maximizing accuracy of the operations with respect to the turned dimensions of the part. In some cases live tooling can be used while the part is turning, to make very complex forms such as deep screw threads.

---

[27] One can always divide a length in two using a compass. Do you remember from high school geometry how to do this? Isn't it interesting that these simple tools have allowed manufacturing science to evolve to the point where it is today!

[28] For a detailed discussion of optical component machining techniques, including single point diamond machining, see R. R. Shandin and J. C. Wyan (eds.), Applied Optics and Optical Engineering, Vol. X, Academic Press, New York, pp. 251–387.

[29] G. Sanger, "Optical Fabrication Technology, the Present and Future," Proc. SPIE, Vol. 433, Aug. 1983.

7.2.4 Treating Operations[30]

In the process of obtaining the desired shape of a part, sharp edges can form, grime can accumulate, and internal stresses can develop. Sharp edges are a safety hazard, increase assembly problems, and are sites for raised dents to occur when banged. Grime can interfere with assembly or pollute clean systems, and internal stresses can compromise long-term stability and structural integrity. Hence it is vitally important that appropriate processes be specified to correct these conditions. In addition, after making the part to the desired shape in the soft state of the material, the part must sometimes be treated to increase its strength and make its surface harder and more wear- and corrosion-resistant.

7.2.4.1 Cleaning and Deburring

The need for cleaning chips, coolant, and grime off a part after completion of a manufacturing process is usually self-evident. However, one must also consider what happens to the part after it is cleaned. For example, a steel part that has been cleaned for bonding should not sit around too long or the surface will rust. The design engineer must often incorporate these manufacturing considerations into the design when specifying materials and processes. For example, a special cleaning process known as passivation uses an acid bath to remove ferrous particles from the surface of stainless steel. These particles can become embedded during manufacturing, and if they are not removed they may oxidize and discolor the surface.

There are almost as many methods for deburring as there are methods for manufacturing. The design engineer need not worry too much about which method to use (the manufacturing engineer can determine the best manner to debur a part). However, the design engineer should be aware of the reasons for deburring a part: (1) a sharp edge can cut a person; (2) if a sharp edge is banged by another object, a dent will form which will prevent another part from resting flat on the surface; (3) burrs can break loose and clog lubrication systems; and (4) sharp edges make assembly of mating components more difficult. Hence, whenever possible, one should specify generous chamfers on part edges.

Burrs and surface irregularities can be removed by a process that cuts them off, or they can be smeared by a process called *burnishing*. Burnishing involves rubbing a smooth hard tool across the surface of the part. For example, roll burnishing uses a set of rollers forced radially outward as a means of burnishing the inside diameter of a part. All burnishing operations have the advantage of imparting a net compressive stress into the surface of the part which if it does not distort the part, improves fatigue and wear resistance, and decreases flow resistance in the case of parts that act as fluid guides.

7.2.4.2 Relieving Internal Stresses

Casting, welding, machining, or any other process that radically changes the internal or external form of the part can induce stresses, which with time can act to warp the part.[31] Internal or residual stresses are thought to be one of the major causes of dimensional instability, and thus in critical applications it may be necessary to anneal or otherwise stress relieve the part (e.g., a casting from the foundry) before and after rough machining just prior to final machining (i.e. grinding, scraping, lapping, or honing). Also, if the stresses are tensile in nature and the part is subjected to a corrosive environment, cracking and degradation of the surface will be accelerated.

On the other hand, in some cases it is desirable to induce a residual compressive stress on the surface of the part in order to increase the fatigue life of the part. This is particularly true in some heavily loaded bearing applications, where failure by fatigue is of greater concern than failure by dimensional change.[32] Residual compressive stress can be imposed on a precision steel surface by carburizing or nitriding. In the latter case, the dimensional stability of the part can be retained

[30] For more detail, see, for example, Metals Handbook, Vols. 4 and 5, published by the American Society for Metals, Metals Park, OH.

[31] See, for example, "Surface Integrity," Manuf. Eng. July 1989 (adopted from the Tool and Manufacturing Engineers Handbook series from SME).

[32] C. Stickels and A. Janotik, "Controlling Residual Stresses in 52100 Bearing Steel by Heat Treatment," Residual Stress for Designers and Metallurgists, American Society for Metals, Metals Park, OH, 1981.

indefinitely if the proper alloy is chosen (e.g., Nitralloy), as discussed later. For nonprecision parts, residual stresses can be imposed by shot peening.[33]

Thermal Processes

Historically, the most common method for relieving internal stresses was to subject the part to a thermal process whereby the part is typically raised to a temperature that causes a phase change in the material, and then to cool the material slowly giving time for another phase change to occur without forming any internal stresses.[34] This process removes dislocations in the material and hence eliminates internal stresses. As shown in a generic equilibrium diagram for a generic steel in Figure 7.2.42, there are several processes that can be used to stabilize a part, including *normalizing, full annealing, subcritical (process) annealing,* and *spheroidizing annealing*.

Ferrous parts which have been hot worked (e.g., cast, forged, or rolled) typically have nonuniform structures, grain sizes, and hardness. *Normalizing* is a process that makes these properties uniform throughout the material and removes internal stresses in order to facilitate further processing or to attain dimensional stability. Normalizing involves heating to about 50-60 C^o above the A_3 or A_{CM} lines in Figure 7.2.42. Since normalizing causes a phase change in the material from its previous state to a uniform austenite structure (face-centered cubic structure with 74% packing efficiency between the atoms), the cooling rate will determine the ratio of hard cementite (Fe_3C) to soft ferrite (body-centered cubic with 68% packing efficiency) and the resulting characteristics of the microstructure (pearlite) that is formed. Hence the normalizing process can also become a hardening process if the material is rapidly quenched. After normalizing, tempering is often required, which modifies the ferrite-cementite ratio and structure.

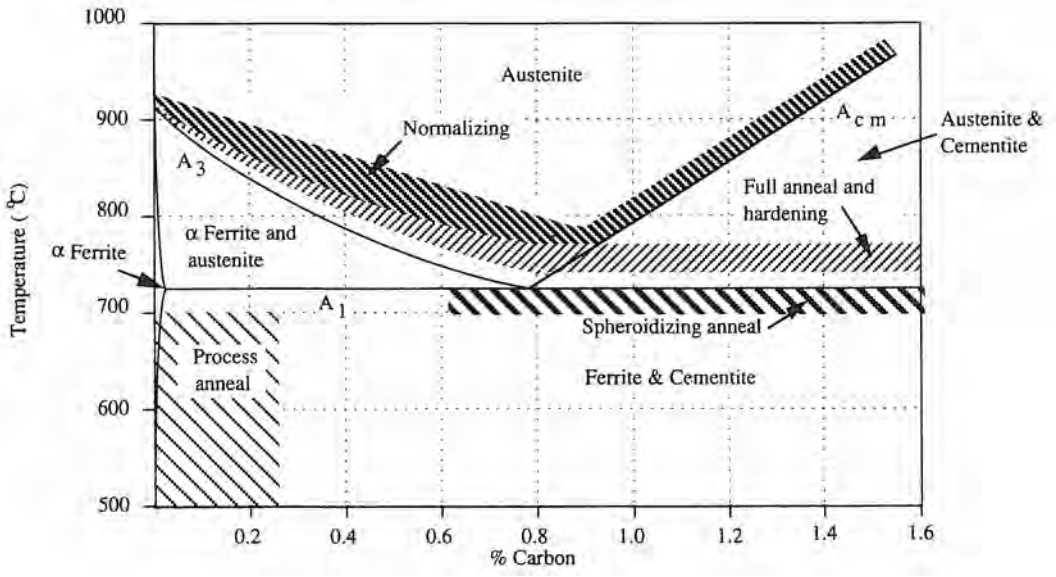


Figure 7.2.42 Typical relationships between phases and heat treatments for steel.

Annealing

is a generic term used to describe the process of heating a part and maintaining it at an appropriate temperature, followed by a controlled cooling cycle to achieve desired properties (generally speaking, softening). A *full anneal* involves heating the material until it attains a uniform (e.g., austenite) structure, and then slowly cooling it to room temperature. This results in maximum softening, uniformity, and stability in the structure. For example, an iron casting right out of the mold has a brittle white iron shell which can cause bimetallic thermal distortion problems. A full anneal can be used to change the white iron layer to gray iron. For a proper full anneal, the temperature

[33] For a discussion of various other methods for imparting beneficial residual stresses and measuring them, see W. Young (ed.), Residual Stress in Design, Process, and Materials Selection, (Proceedings of the ASM Conference on Residual Stress, April 1987).

[34] To describe the annealing process more accurately requires consideration of equilibrium diagrams for the particular material under consideration.

specified, the time at temperature, and the cooling rate all must be carefully chosen according to the type of alloy and part that is being annealed. If the proper combination of these variables is not chosen, then dimensional stability may not be achieved. In general, the higher the temperature, the longer the time at temperature, and the slower the cooling rate, the better the dimensional stability of the part; however, this results in maximum cost.

A *subcritical* or *process anneal* is often done between stages of the manufacturing process of the part in order to relieve stresses generated by cold working. For example, after a part is cast, it will probably be full annealed and then rough machined to remove scale and attain the approximate final dimensions. The rough machining operation creates nonuniform residual stresses in the surface of the part. To relieve these stresses does not, however, require heating the entire part until a phase change occurs. The part may only have to be heated to the point where internal stresses cause the stressed material to creep and relax at the process anneal temperature. This is a substantially cheaper (factor of 2-3) process than a full anneal.

When a ferrous part is cooled from the austenite state, the resulting structure is composed of intertwined layers of ferrite and cementite, where the ratio and structure depend on the cooling cycle. This structure is often not the most dimensionally stable state because the crystals were usually not given enough time to form a stress-free configuration before the temperature dropped to the point where they were locked in position. In order to allow the crystals to be totally relaxed, and hence the material to be totally softened, the iron carbide (cementite) particles must be allowed to collect and form globs (spheres) that are then surrounded by the soft iron matrix. This is called a *spheroidizing anneal*. It requires the material to be heated to just below the A_1 line in Figure 7.2.42 and then held there for a period that may last days, or furnace cooling from the normalization temperature, which can also last for days. A fully spheroidized steel is very soft, which makes it difficult to machine but easy to radically cold work (e.g., cold form tubing). In this soft state the material has maximum metallurgical stability but minimum strength.

Mechanical Stress Relief

The biggest problem with thermal stress relieving is that it is energy intensive and causes the surface of the part to oxidize and discolor unless the part is annealed in an inert environment, which is a very expensive process. On the other hand, if a material has high residual stresses and is allowed to sit for thousands of years, it will eventually reach a relaxed state. *Mechanical stress relief* is a process that speeds up the aging process. It consists of stretching or compressing the part or blank stock sufficiently to cause typically 1-3% plastic deformation. This induced plastic flow relieves internal stresses and helps achieve dimensional stability. This process is most often used to impart better dimensional stability to heat-treated aluminum alloys so that massive quantities of material can be machined away to reduce weight for aircraft components. Aluminum alloys that are mechanically stress relieved have the digits "51" or "52" following the basic temper, T#, designation (e.g., 6061-T651) to indicate tensile or compressive stress relief, respectively. When a part made from mechanically stress relieved heat-treated aluminum stock is rough machined, residual stresses will be induced in the surface of the material. These can be relieved by heating the part to within 175-205°C and holding it there for 1-2 hours and then air cooling the part. This will result in some loss of strength, but will help to ensure that dimensional stability is maintained.[35]

Mechanical stress relief can also be applied to complex parts (e.g., aluminum or iron castings) by *vibration annealing* once the parts have been normalized (or heat treated) and rough machined. Vibration annealing is performed by placing accelerometers and pneumatic or electric vibrators at various points on the structure and then adjusting the frequency until the accelerometers' outputs are maximized. At this point, one is exciting the structure at one or more of its natural frequencies. Hence the energy applied to the structure builds upon itself and eventually "shakes" the residual stresses out of the structure. One must be careful not to impart new stresses into the structure; hence this process is somewhat of a black art. In order to avoid imparting new stresses in large parts, it is advisable to use several small vibrators applied locally instead of just one big vibrator. This is by far the most energy-efficient method for relieving stresses in large structures that would otherwise be very expensive to stress relieve. This process has been adopted by some machine tool manufacturers

[35] See Vol. 4 of the Metals Handbook for more detailed discussions of treating aluminum and other alloy metals. An advisable recreational pastime is to spend a couple of hours a week reading through the Metals Handbook.

and it has been used for many years in the shipbuilding industry, where it is impractical to anneal a ship.

7.2.4.3 Hardening

The process of thermal stress relieving generally leaves the material in a soft state. For wear surfaces or where strength is critical, the metallurgical properties of the part can be altered by heat or strain hardening to give the desired wear and strength properties. When applied loads are low and wear resistance is the primary concern, surface hardening is preferred over a through hardening process because it is often less expensive and leaves the material in a more dimensionally stable state. Contact stresses must be determined as a function of surface depth and it must be ensured that they are below the yield strength of the hardened layer. If the stresses are too high, deep in the material, then a through hardening may be required.

Various handbooks on surface treatment give the hardness as a function of depth, and this information can be used in conjunction with calculation of the contact stresses to determine the suitability of the surface hardening process. The Brinell hardness B can be approximately related to the tensile strength of plain carbon and low-alloy steels by[36]

$$\sigma(\text{MPa}) \approx 3.45\,\text{B} \tag{7.2.1}$$

$$\sigma(\text{psi}) \approx 500\,\text{B} \tag{7.2.2}$$

The Rockwell C hardness (denoted by HRC and then the number) can be related to the tensile strength of the material by[37]

$$\sigma(\text{MPa}) \approx 3.45 \left\{ \frac{1590}{122 - \text{HRC}} \right\}^2 \tag{7.2.3}$$

For purposes of determining failure in metals, the maximum allowable shear stress is sometimes assumed to be

$$\tau_{\text{max metals}} \approx 0.5\sigma_{\text{tensile strength}} \tag{7.2.4}$$

The stresses as a function of depth in the part beneath the contact region can be found from the discussion of Hertz contact stresses in Section 5.6. When the elastic moduli of the base and a plated material are about the same, the Hertz equations can be used. When the moduli differ (e.g., chrome on aluminum) a more elaborate solution is needed.[38]

Selective Heating

When a steel part to be surface hardened contains more than about 0.3% carbon, the surface can be hardened by heating it so rapidly that the inner structure does not have time to heat up also. As the heat source moves along the part, the inner regions act as a heat sink to quench the outer surface and harden it. In this manner the bulk of the material remains cool and dimensional accuracy of the part is maintained. This process is often used for hardening of ferrous gear teeth and bearing ways that are an integral part of a cast iron structure. Numerous methods exist for selective hardening, including flame, laser, electron beam, and induction hardening. When the iron alloy contains insufficient carbon to allow for selective heating, a process must be used where the metallurgical properties of the surface are changed by diffusing new elements into the microstructure. These processes most commonly include carburizing and nitriding.

Carburizing

In order to diffuse sufficient amounts of carbon into a ferrous material's surface microstructure, it is necessary to surround the part with a carbon-rich environment and heat the system to about 900 ^{o}C for 6-72 h, depending on the thickness of the hard layer desired. After soaking at the proper temperature, the parts must then be quenched; hence final grinding may be necessary to remove any thermal warpage that may be induced by uneven quenching. The amount of dimensional change can

[36] 1 MPa = 10^6N/m^2. 1 psi = 1 lb/in^2= 6890 N/m^2.

[37] F. McClintock and A. Argon, Mechanical Behavior of Materials, Addison-Wesley Publishing Co., Reading, MA, 1966, p. 448.

[38] See, for example, Y. Chiu and M. Hartnett, "A Numerical Solution for Layered Solid Contact Problems with Applications to Bearings," J. Lubri. Technol., Vol. 105, Oct. 1983.

be on the order of 1/4 -1%. Case depths from a few microns to a few millimeters can be attained in this manner.

Nitriding

Surfaces can be hardened by formatting nitrides of alloying elements in the steel (e.g., aluminum, chromium, molybdenum, tungsten, and vanadium).[39] Parts must first be heat treated and tempered at a temperature at least 30 C^o higher than the nitriding temperature. The parts are then finish ground to final dimensions and to remove any decarburized surface material. The parts are then heated in an ammonia-rich environment (gaseous or liquid) for 10–120 h at 495–565°C. The nitrogen diffuses into the steel and forms hard (in excess of Rockwell 65C) wear-resistant nitrides to a depth of $1/2 - 1$ mm. Note that typical quench and tempered bearing surfaces may have a hardness on the order of Rockwell 55–65C.

Nitriding will also help improve the fatigue life and corrosion resistance of a part. Furthermore, because the temperatures are relatively low and the parts can be allowed to furnace cool slowly, dimensional stability of nitrided parts is better maintained than with a quench and temper process. The expanded volume of the outer nitride case causes it to be in compression while the core is in tension. This can warp the part if it is not symmetrical. Sharp corners and edges should be avoided because built-up brittle projections can form on them during nitriding. These projections are brittle and susceptible to chipping. Similarly, thin or sharp sections become nitrided through the section and thus are very brittle. Dimensional changes will be identical for identical parts nitrided in different batches with the same nitriding cycle; thus a test part can be nitrided, measured, and subsequent parts machined with an allowance that compensates for dimensional changes. Since the nitrided case and core are in static equilibrium, only final lapping of the surface is allowable. Removal of more than a few microns of the outer core can cause an imbalance in the stress equilibrium that leads to dimensional changes. If the part must be ground, then afterwards it must be stabilized by heating to 565°C for 1 h and then finish ground or lapped. Precision parts, such as gage blocks made from Nitralloy, which are only finish lapped after nitriding, are usually highly dimensionally stable.

Although nitriding may appear to be more expensive than other surface hardening treatments, nitriding allows many types of parts to be finished to size while still in the unhardened state. Other processes, such as carburizing, often require parts to be finish ground afterwards. The savings in machining costs often offset the increased expense of the process itself. Nitriding is a very attractive process to use for precision components that need to be surface hardened.

Shot Peening

The surface of a material can be stabilized by imparting a net compressive stress into it, which in turn makes it difficult for cracks to start at the surface and then propagate into the material. One way to form this surface layer is to direct a stream of small, high-velocity spheres at a surface, which is called *shot peening*. Peening also helps to prepare the surface prior to bonding, as it forms thousands of tiny dish-shaped craters that increase the surface area and hence increase the shear strength of the bonded joint.

Through Hardening

When a part is subject to high loads that act over its cross section, a through hardening process may be required. In general, if a precision machine's part is so highly stressed that it needs to be through hardened, then the part is not a high-accuracy part or it has been poorly designed. Exceptions to this rule are, of course, bearing and tooling components. Some materials (e.g., ferrous alloys) become so hard that they can only be ground or lapped to achieve close tolerance final dimensions. Other materials, such as aluminum, can readily be machined after hardening.

In order to be hardenable, the material the part is made from must contain the proper amounts of alloying elements. Choosing the proper alloy and hardening process can be a formidable task. Thus it is wise to list all the process variables and operating conditions the part will be subjected to, and then consult with a materials engineer or major materials supplier for assistance. Section 7.3 also discusses some of the basic issues involved in materials selection for machine tool components.

Regardless of the hardening process used, there are some design considerations for parts that are to be through hardened:

[39] Note that ion implantation can also be used to introduce nitrogen into the surface. See J. Destefani, "Ion Implantation Update," Met. Prog., Oct. 1988.

1. Consider the quenching process and how it will affect the shape of the part. What sort of differential cooling could occur that may warp the part?
2. How will the shape of the part affect the hardness as a function of depth, and does it matter? Very thick section pieces will quench rapidly near the outside edges and more slowly near the center.
3. Do not harden parts with tapped holes or other sharp concave features, as they have high stress concentration factors and become prime sites for initiation of fracture. Make all edges and corners well rounded. If a hardened part is to be held in place with bolts, it is better to drill and counterbore the to-be-hardened part and tap the softer metal to which the hardened part is to be attached.
4. In general, hardened steels can be finished only by abrasive, electrodischarge, or electrochemical means. One should consult with manufacturing personnel to make sure that they can finish a part once it is hardened.

7.2.4.4 Plating

The surface of a part does not necessarily have to be of the same material as the core. Plating is an effective means for allowing the core to be made of an inexpensive or easy-to-machine material (e.g., steel), while the outside has the characteristics of an expensive or difficult-to-machine material (e.g., chrome or nickel). There are three common plated finishes: electrolytic, electroless, and vacuum deposited. The most commonly electroplated metals are tin, cadmium, chromium, copper, platinum, titanium, and zinc. The part(s) to be plated is placed in a vat containing dissolved salts of the metal to be deposited. The part is made the cathode and a block of the to-be-deposited metal is made the anode. When a dc voltage is applied, metallic ions of the to-be-deposited metal migrate to the part, contact it, give up their charge, and adhere to it. For precision surfaces, hard chrome plating is most often specified. The resulting surface can be up to 3/4 mm thick and have a hardness on the order of HRC 65-70. Electroplating does not smooth over surface defects, and thus in many instances the surface must be ground, lapped, or honed after plating with hard metals. Unfortunately, the resulting plating thickness is not always uniform and is also highly dependent on the shape of the part.

Electroless plating is a chemical process that provides nearly uniform plating thickness on the entire surface of the part virtually independent of part complexity. Nickel is the most common electroless plated metal, and it can provide an order-of-magnitude better corrosion protection than even hard chromium can. As plated, electroless nickel plate has a hardness only on the order of HRC 45-50. By heat treating it (400°C for 1 h), a hardness on the order of HRC 60-65 can be obtained. Note that hard chromium wears less by a factor of 2-5 depending on the application.[40] Electroless nickel also diamond machines very well when the phosphorus content is on the order of 12%.[41] It is also possible to impregnate an electroless nickel plated surface with Teflon®, thereby decreasing the sliding coefficient of friction to about 0.06. The plated layer's thickness is also self-limiting, so it gives an even buildup over all regions of the part. In general, electroless nickel is the plating material of choice to use for precision machine components where stability, diamond machinability, wear resistance, and hardness are required.

Vacuum coating is an economical process for depositing thin films (e.g., 1/2 μm) on materials. It can be used to deposit virtually any material onto virtually any other material. Hence it is most often used to deposit gold or silver to form reflective surfaces. Large telescope mirrors are made by vapor depositing a thin layer of aluminum onto polished glass mirror blanks.

For components used in severe environments, coatings can be applied using thermal spray techniques. Used extensively in aircraft engines, thermal spray processes are used to lay down a wide variety of different materials onto surfaces for different purposes. The material to be applied is heated to a molten or semimolten state and then propelled at sufficient velocity against the substrate to produce the required bond strength.[42]

---

[40] From product literature of Enthone Inc., West Haven, CT.

[41] C. Syn, J. Taylor, and R. Donaldson, "Diamond Tool Wear Versus Cutting Distance on Electroless Nickel Mirrors," Proc. SPIE, Vol. 676, 1986.

[42] See G. Kutner, "Thermal Spray by Design," Met. Prog., Oct. 1988.

7.2.4.5 Anodizing

Aluminum is a wonderful material to machine and it has excellent heat transfer characteristics, but it has poor wear characteristics. Aluminum oxide, on the other hand, wears so well that it is used as an abrasive. To increase the usefulness of aluminum in machine tool components (e.g., for air bearing rails), an oxide layer on aluminum parts can be formed by anodizing. Anodizing is a reverse plating process whereby the part is made the anode in an electrolytic circuit and the reaction progresses inward, increasing the thin layer of aluminum oxide that normally exists on the surface of an aluminum part. Unlike rust on steel, which is not hard or dense and tends to flake off, aluminum oxide is harder and denser than the base metal, and adheres well to the aluminum surface. With anodizing, no new material is added to the part, so there is essentially no net growth of the part and the oxide layer is typically 1-100 μm thick. Hence, only finish grinding, honing, or lapping may be required. Depending on the process, the anodized layer can be microscopically porous, which allows it to be dyed to yield virtually any color and can also be impregnated with PTFE (e.g., Teflon®) or other lubricants to make it self-lubricating.

7.3 MATERIALS

There are hundreds of thousands of different types of metals, plastics, composites, and ceramics from which the design engineer can choose. Even within a specific subgroup (e.g., cast iron) there may be hundreds of different alloys. A machine design engineer cannot be expected to be familiar with all different types of materials, but he should be aware of the basic properties of different groups of materials.[43] For example, Table 7.3.1 lists the basic types of materials a machine tool design engineer typically has to choose from. Before choosing a material for an unfamiliar application, the design engineer should always discuss the application with an experienced materials engineer: Too many seemingly well-designed machine components have ended up on the scrap heap because the wrong material was chosen for the wrong application.

Numerous books and articles exist on how the alloy and structure of various materials affect strength and endurance (fatigue and corrosion resistance) properties and how one should choose a material based on these parameters.[44] Thus in this section only a brief review of strength and corrosion parameters is made, and emphasis is placed on material damping and thermal properties.

7.3.1 Strength Properties of Materials

There are many different failure modes for different types of materials, and hence there is no single quantifier of strength for the design engineer to use.[45] Figure 7.3.1 shows a generic stress-strain curve for a material. Up to a stress level of σ_P, the material is assumed to be *linearly elastic* (stress is linearly proportional to strain) as governed by Hooke's law. From σ_P to σ_E the material is still elastic: The unload stress-strain curve will trace the loading curve back to the origin. After the elastic limit, once the load is removed, then the unload stress-strain curve will follow the dashed line, which is parallel to the proportional elastic loading line, back down to the strain axis. The amount of plastic strain will be where the line intersects strain axis. The *yield point* is usually defined in terms of the stress that causes 0.2% permanent plastic deformation. The *ultimate tensile stress* is where the part starts necking and the force begins to decrease with increasing strain until fracture occurs. Of course, this is a simplified diagram. Many materials such as plastics behave in a very different nonlinear manner. However, as far as most machine design engineers are concerned, critical machine elements will behave in a manner similar to that shown in Figure 7.3.1, with the exception that hard materials (e.g., hardened steel or ceramic components) will fail without necking soon after the yield point is reached.

The state of stress in a part is often quite complex, and the most common way to evaluate when failure will occur is to find an equivalent maximum shear or tensile stress in the part. Hard

[43] It is not a bad idea to become a member of the American Society for Metals (Metals Park, OH 44073), which entitles you to receive their magazine Advanced Materials and Processes. This will help to keep you up to date on all types of new materials. The annual Advanced Materials and Processes Guide to Selecting Engineered Materials is particularly useful.

[44] For example, the Metals Handbook (Volumes 1-7), published by the American Society for Metals, Metals Park, OH.

[45] See, for example, J. Shigley and L. Mitchell, Mechanical Engineering Design, McGraw-Hill Book Co., New York, 1983.

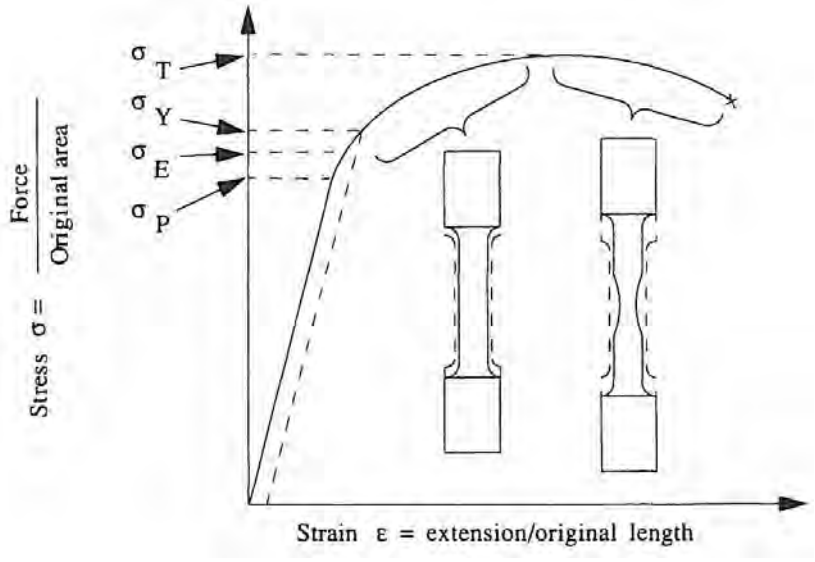


Figure 7.3.1 Definition of stress levels in a material.

materials with low ductility (e.g., ceramics) usually fail when their maximum tensile or compressive stress is exceeded. Materials with higher ductility (e.g., most metals including bearing steels) usually fail when their maximum shear stress is exceeded. In either case, the complex state of stress can be reduced to a set of equivalent principal stresses using Mohr's circle, or an overall equivalent stress found from the general state of stress using the *Mises yield criterion:*

$$\sigma_{yield} = \sqrt{\frac{(\sigma_x - \sigma_y)^2 + (\sigma_y - \sigma_z)^2 + (\sigma_z - \sigma_x)^2}{2} + 3\tau_{xy}^2 + 3\tau_{xz}^2 + 3\tau_{yz}^2} \qquad (7.3.1)$$

Mathematically, this represents a cylinder of radius $\sigma_{yield}\ (2/3)^{1/2}$ whose axis makes equal angles with each of the σ_1, σ_2, and σ_3 principal stress axes. Recall that the principal stress axis system is found from the general state of stress using Mohr's circle, and it is where shear stresses are absent. The maximum allowable shear stress criterion says that yielding will occur when

$$\tau_{max} = \frac{\sigma_{max} - \sigma_{min}}{2} = \frac{\sigma_{yield}}{2} \qquad (7.3.2)$$

Of course, allowances must be made for safety factors, stress concentrations, and cyclic loading.

A measure of the *ductility* of a material, which is a measure of the resistance to fracture, is the *elongation,* which is the amount (%) that a material deforms before it fails. The *fracture toughness* is a measure of how easily cracks can grow in the material. The *brittle-to-ductile transition temperature* is the temperature where a marked increase in toughness occurs. In general, body-centered-cubic materials (e.g., ferritic steels) are susceptible to cold-temperature cracking. Face-centered-cubic materials (e.g., aluminum and austenitic stainless steels) can usually be cooled to near absolute zero temperatures without suffering a loss of ductility.[46]

A material's strength is determined by a complex interaction of metallurgical properties, whereas wear resistance is a function of the type and distribution of particles on the surface of the part. For example, 7075-T6 aluminum has a yield strength of 462 MPa (67 ksi), almost twice that of A36 structural steel, yet in a wear test with the two rubbing against each other, the steel would win. Generally, wear-resistant materials have a hard carbide, nitride, or oxide layer on their surfaces. Hardened steels, ceramics, and thickly anodized aluminum all have good wear properties. Other materials, such as brass, cast iron, and most plastics, wear well when used in conjunction with a hard material because of the former's inherent lubricity.

[46] When the possibility of fatigue or fracture exists, the designer should investigate the issues thoroughly and possibly consult an expert. A nice reference to consult initially is S. Rolfe and J. Barsom, Fracture and Fatigue Control in Structures: Applications of Fracture Mechanics, Prentice Hall, Englewood Cliffs, NJ, 1977.

7.3.2 Bulk Properties of Materials[47]

Bulk properties of isotropic materials are those which depend on the composition and internal structure of the material, and generally not the process used. For example, heat treating an isotropic steel part, one without grain orientation due to rolling stresses, will not appreciably affect the modulus of elasticity. On the other hand, some materials, such as those prepared using powdered materials processes such as hot isostatic pressing (HIP), have bulk properties that are very much dependent on the preparation of the material. With respect to dimensional properties of materials, the primary bulk properties are E, η, and α. The modulus of elasticity E is the material type's effect on the spring constant of the system. Poisson's ratio η is a measure of how much the material contracts (or expands) in directions orthogonal to the direction in which the stress is applied. The coefficient of thermal expansion is α and defines the change in length per unit length per degree change in temperature. For an isotropic material,[48] these constants relate stress and strain by Hooke's law:

$$\begin{aligned} \varepsilon_x &= \frac{\sigma_x - \eta(\sigma_y + \sigma_z)}{E} + \alpha\Delta T \\ \varepsilon_y &= \frac{\sigma_y - \eta(\sigma_x + \sigma_z)}{E} + \alpha\Delta T \\ \varepsilon_z &= \frac{\sigma_z - \eta(\sigma_x + \sigma_y)}{E} + \alpha\Delta T \end{aligned} \tag{7.3.3}$$

$$\gamma_{xy} = \frac{\tau_{xy}}{G} \qquad \gamma_{yz} = \frac{\tau_{yz}}{G} \qquad \gamma_{xz} = \frac{\tau_{xz}}{G}$$

Using Mohr's circle, one can show that for isotropic materials, the shear and elastic moduli have the following relation:

$$G = \frac{E}{2(1+\eta)} \tag{7.3.4}$$

Unfortunately, for many materials, there is a different modulus for tension and bending. When one calculates the deflection of a member due to bending, the bending modulus can be directly substituted for the isotropic elastic modulus E that normally appears in the expression. For simple tension or compression where there are no transverse stress components (plane stress), one can use the tensile or compressive moduli, as required.

Nonisotropic materials are usually thought of as being members of the composites family (e.g., fiber-reinforced materials), but some cast metals and porous materials are also nonisotropic. However, even the manufacturing process used to make everyday materials such as cold-rolled steel can directionally orient a material's grain structure, thereby yielding an elastic modulus that changes with orientation in the part by as much as 20%.[49] For most machine design purposes, one should use the lowest value, which is usually equal to the Young's modulus for the material when there is no anisotropy. In addition, note that for all polycrystalline metals with a given structure (e.g., BCC and FCC iron and FCC aluminum alloys), the modulus of elasticity, Poisson's ratio, and density are not affected by small additions of alloying elements or the heat treatment process unless they preferentially align the grain structure along a particular axis. Don't ever let anyone tell you, for example, that a steel bracket should be heat-treated in order to increase its stiffness.

Dimensional Stability

How well a part made from a given material holds its shape with time and stress is referred to as the *dimensional stability* of the part and the material. Most materials undergo plastic deformation to some degree when subject to stress (e.g., cutting forces), and this imparts residual stresses in the material; thus the goal is to make the net stress level low enough so that the creep rate is not noticeable for the life of the machine or in-between calibrations. In order to maximize dimensional stability, the machine design engineer should try to minimize the ratios of applied and residual stress

[47] See H. Boyer and T. Gall (eds), Metals Handbook: Desk Edition, American Society for Metals, Metals Park, OH, 1985.

[48] An isotropic material has the same properties in all directions. For anisotropic materials (e.g., composite materials) a generalized form of Hooke's law also exists. See, for example, S. Tsai and H. Hahn, Introduction to Composite Materials, Technomic Publishing Co., Westport, CT.

[49] See S. Crandall et al., An Introduction to the Mechanics of Solids, McGraw-Hill Book Co., New York, 1972, pp. 311. Also see F. McClintock and A. Argon, Mechanical Behavior of Materials, Addison-Wesley Publishing Co., Reading, MA, 1966.

to yield strength of the material.[50] When metals are hardened by heat treatment, they are more likely to contain high levels of residual stress; thus when possible it is best to use unhardened or surface-hardened metals. Note that ceramic materials, such as aluminum oxide or silicon nitride, are usually far less sensitive to chemical attack than are metals; and after being properly fired in a kiln and then cooled, ceramic materials can be virtually free from internal stresses. Also, processes such as grinding cause minimal dislocation growth and hence plastic deformation in the surface structure of brittle ceramics, which minimizes residual stress levels. Some materials that can be dimensionally stable are listed in Table 7.3.1.[51] In addition, assemblies of parts must be made in a low stress manner to ensure stability of the part interfaces.

Material[a]	ν	E (GPa)	ρ (Mg/m^3)	K (W/m/C^o)	Cp (J/kg/C^o)	K/ρCp (10^{-6}m^2/s)	α (μm/m/C^o)	Hard[b]	σ_{com} (MPa)	σ^c_{ten} (MPa)	σ_{flex} (MPa)
Aluminum (6061-T651)	0.33	68	2.70	167	896	69	23.6	No		$\approx$255	
Aluminum (cast 201)	0.33	71	2.77	121	921	47	19.3	No		$\approx$100	
Aluminum oxide (99.9%)	0.22	386	3.96	38.9	880	11.2	8.0	Yes	3792	310	552
Aluminum oxide (99.5%)	0.22	372	3.89	35.6	880	10.4	8.0	Yes	2620	262	379
Aluminum oxide (96%)	0.21	303	3.72	27.4	880	8.4	8.2	Yes	2068	193	358
Beryillium (pure)	0.05	290	1.85	140	190	398	11.6	No		$\approx$345	
Copper (OFC)	0.34	117	8.94	391	385	114	17.0	No		$\approx$90	
Copper (free machining)	0.34	115	8.94	355	415	96	17.1	No		$\approx$125	
Copper (bery.-copper)	0.29	125	8.25	118	420	34	16.7	No		300-900	
Copper (brass)	0.34	110	8.53	120	375	38	19.9	No		$\approx$125	
Granite	0.1	19	2.6	1.6	820	0.8	6	No	$\approx$300	-	$\approx$20
Iron (Class 40 cast)	0.25	120	7.3	52	420	17	11	No		$\approx$270	
Iron (Invar)	0.3	150	8.0	11	515	2.7	0.8	No		$\approx$210	
Iron (Super Nilvar)	0.3	150	8.0	11	515	2.7	0	No		$\approx$210	
Iron (Nitralloy 135M)	0.29	200	8.0	4.2	481	1.1	11.7	Yes		310-2760	
Iron (1018 steel)	0.29	200	7.9	60	465	16	11.7	No		$\approx$270	
Iron (303 stainless)	0.3	193	8.0	16.2	500	4.1	17.2	No		$\approx$310	
Iron (440C stainless)	0.3	200	7.8	24.2	460	6.7	10.2	Yes		310-2760	
Polymer concrete	0.23-0.3	45	2.45	0.83-1.94	1250	0.27-0.63	14	No	150	35	45
Zerodur	0.24	91	2.53	1.64	821	0.8	0.05	Yes	-	-	$\approx$60
Silicon carbide	0.19	393	3.10	125	-	-	4.3	Yes	2500	307	462
Silicon nitride (hip)	-	350	3.31	15	700	13	3.1	Yes	-	-	906
Tungsten carbide	-	550	14.5	108	-	-	5.1	Yes	5000	-	2200
Zirconia	0.28	173	5.60	2.2	-	-	10.5	Yes	-	-	207

[a]All properties at 20^{o}C. [b]"Hard" refers to whether or not the material can support dynamic point bearing loads (e.g., rolling elements). For wear resistance, virtually any material can be plated with a wear-resistant material such as electroless nickel or hard chrome. [c]For metals, the approximate yield strength is given.

Table 7.3.1 Properties of various materials for use in precision machines.

Thermal properties

Because thermal errors are often the dominant type of error in a precision machine, the thermal characteristics of structural materials deserve special consideration. Often one must consider not only the coefficient of expansion of a material, but also the thermal conductivity and the thermal diffusivity of the material. In addition, the manner in which a part is made (e.g., a solid section from granite compared to a hollow section from steel), will affect the temperature gradient across the part and hence the thermal deformations of the part. For example, when subject to a given heat flux across its height, similar shaped beams designed for use in a coordinate measuring machine experienced the following bending deformations (normalized to granite): $\delta_{\text{Granite}} = 1.00$, $\delta_{\text{96\% Alumina (cored)}} = 0.60$, $\delta_{\text{Solid aluminum}} < 0.10$, $\delta_{\text{Hollow aluminum}} = 0.25$, $\delta_{\text{Hollow steel}} = 1.80$.[52]

As discussed in Section 2.3.5, often the best strategy for control of thermal deformations is to isolate heat sources and actively cool them, insulate the structure, maximize thermal conductivity of structural elements, and actively attempt to control the temperature of the machine and the environment.

Manufacturability

A design engineer must integrate part configuration with material choice and manufacturing methods, particularly when the part will be produced in high volumes. Ideally, a design engineer

[50] A good rule of thumb is to keep the static stress level below 10-20% of yield.

[51] Note that zirconia can undergo phase transformations which makes its stability not wonderful; however, it is a tough wear-resistant material that has the same coefficient of expansion as steel, and thus it has useful applications in the paper and food processing industries.

[52] K. H. Breyer and H. G. Pressel, "Paving the Way to Thermally Stable Coordinate Measuring Machines," Progress in Precision Engineering, P. Seyfried, et al. (Eds.), Springer-Verlag, New York, 1991, pp. 56–76.

would be familiar with all materials and manufacturing processes, so the situation would never arise where a part could not be manufactured, or a better configuration bypassed because the design engineer thought the part could not be manufactured when in reality it could. Fortunately, material manufacturers are usually well aware of different manufacturing methods and are usually happy to help with the selection of materials and manufacturing methods.

7.3.3 Material Damping[53]

The effect of material damping can be readily observed by placing your ear against a desk and then hitting it (i.e., the desk) and listening to the sound as it decays. In a machine tool, vibrations can be induced by cutting action or by some other excitation mechanism (e.g., a rotating component that is slightly out of balance) that causes the toolpoint to move as it passes by; hence it is very important to build a structure that has high damping to minimize this effect. Vibrations in a structure are dampened by energy losses in the material and in the interfaces between components.[54]

Although it has been extensively studied, the mechanism of damping in a material is difficult to quantify and one must generally rely on empirical results.[55] In fact, damping is highly dependent on alloy composition, frequency, stress level and type, and temperature. Structural damping levels are often quite low, and frequently the dominant source of damping is the joints in an assembly. In fact, one must be extremely wary of damping data that is presented in the literature, because often it is presented without a discussion of the design of the test setup.

There are several damping quantifiers that are used to describe energy dissipation in a structure. The quantifiers include:

- η Loss factor of material
- η_s Loss factor of material (geometry and load dependent)
- A_r Resonance amplification factor
- ϕ Phase angle ϕ between stress and strain (hysteresis factor)
- δ_{Ld} Logarithmic decrement[56] $_{Ld}$.
- ΔU The energy dissipated during one cycle
- ζ The damping factor associated with second order systems

The various damping terms are related (approximately) in the following manner:

$$\eta = \frac{1}{A_r} = \frac{\delta}{\pi} = \phi = \frac{\Delta U}{2\pi U} \tag{7.3.5}$$

Condition	Damping energy integral α	Strain energy integral β
Tension/compression	1	1
Rectangular beam (uniform bending)	$\frac{2}{n+2}$	0.5
Cylindrical beam (uniform bending)	$\frac{1}{n+1}$	0.33

Figure 7.3.2 Stress distribution and damping functions. Note that $\beta v \alpha = 1$ for all cases if n = 2. (After Lazan.)

[53] *"The danger of more or less perpetual vibration of significant magnitude is one of the bugbears of designers of accurate instruments, and research leading to some practical data on this subject for various types of members is urgently required."* T. N. Whitehead.

[54] Mechanical dampers (e.g, shear and tuned mass dampers) are discussed in Section 7.4.1. Joint damping is discussed in Section 7.5 (e.g., see Figure 7.5.8).

[55] A discussion of the many different microstructural mechanisms that generate damping in materials is beyond the scope of this book. For a detailed discussion see B. J. Lazan, Damping of Materials and Members in Structural Mechanics, Pergamon Press, London, 1968.

[56] Most texts on vibration refer to the log decrement as δ; however, to avoid confusion with discussions on displacement termed δ, the log decrement will be referred here to as δ

The loss factor $\eta_s$an be determined experimentally by subjecting a specimen to various frequencies and stresses while measuring the amplification. This allows for the damping to be determined as a function of frequency and stress. The loss factors are equated by the following relation:

$$\eta = \eta_s \frac{\beta}{\alpha} \tag{7.3.6}$$

where α and β are functions of load and geometry as shown in Figure 7.3.2. The factor n is a measure of the stress in the material. When n = 2.0, the material is subject to a low stress.

The phase angle is the ratio of the apparent modulus of elasticity E_2 at low frequencies and the apparent modulus of elasticity E_1 at high frequencies:

$$\phi = \frac{E_2}{E_1} \tag{7.3.7}$$

The logarithmic decrement δ_{Ld}s a measure of the relative amplitude between N successive oscillations of a freely vibrating system (one excited by an impulse):

$$\delta_{Ld} = \frac{-1}{N} \log_e \left(\frac{a_N}{a_1} \right) \tag{7.3.8}$$

The logarithmic decrement can also be related to the damping factor ζ, velocity damping factor b, mass m, and natural frequency ω_n of a second order system model:

$$\zeta = \frac{\delta_{Ld}}{\sqrt{4\pi^2 + \delta_{Ld}^2}} \tag{7.3.9}$$

$$b = 2m\zeta\omega_n \tag{7.3.10}$$

Note that the amplification at resonance ($\omega = \omega_{dpeak}$ by Eq. 7.4.8b) of a second order system is given by

$$A_r = \frac{1}{2\zeta\sqrt{1-\zeta^2}} (\zeta \leq 0.707) \tag{7.3.11}$$

Damping values for various materials are given in Table 7.3.2. The amount of damping one obtains from a material is very low compared to the amount of damping that one can obtain with the addition of a damping mechanism. Damping mechanisms can range from simple sand piles to more complex shear dampers or tuned mass dampers as discussed in Section 7.4.1.

7.3.4 Environmental Properties

In addition to being able to withstand mechanical loads, a part must be configured and a material chosen to ensure that performance in adverse environments will be satisfactory. A precision machine may operate in a clean temperature-controlled room, but thermal performance and corrosion resistance must still be considered.

Thermal Properties[58]

Table 7.3.1 listed common precision engineering materials and their thermal properties. The thermal conductivity is a measure of how well heat is conducted through the material. Materials with low thermal conductivities tend to develop hot spots or large temperature gradients. As illustrated in Section 2.3.5, gradients cause bending moments, which lead to Abbe errors. So even if a material has a very low coefficient of thermal expansion (e.g., Invar or Zerodur), if it is subjected to large thermal gradients or local heat sources, a part made from it may deform more than if the part where made from a material which diffuses heat well (e.g., aluminum). In practice, however, precision instruments rarely encounter such large gradients, so a minimal coefficient of thermal expansion is often the dominant property that affects material choice. The specific heat of a material is a measure of how much thermal energy can be stored in the material. Along with the thermal conductivity, this

[58] Although this section on thermal properties may seem short, one should remember the emphasis placed on thermal errors in Chapter 2. Ideally, every machine designer should have taken a good course in heat transfer.

Material	Load	T_1 (°K)	T_2 (°K)	σ_1 (ksi)	σ_2 (ksi)	f_1 (Hz)	f_2 (Hz)	ζ_1	ζ_2	A_{r1}	A_{r2}
Alumina								5.00E-06	1.50E-05	100000	33300
Aluminum (6063-T6)	bending			1	6			2.50E-04	2.50E-03	2000	200
Aluminum (pure annealed)	axial	50	300					3.50E-06	1.00E-05	143000	50000
Beryillium (18.6%Be)	unspec.			2	50			7.50E-03	4.10E-01	66.7	1.3
Copper (brass)	bending					50	600	1.50E-03	3.00E-03	333	167
Copper (pure annealed)	bending					20	550	3.50E-03	1.00E-03	143	500
Glass	bending					10	100	1.00E-03	3.00E-03	500	167
Granite (Quincy)	bending					140	1600	2.50E-03	5.00E-03	200	100
Iron (cast, annealed)	bending					100	2000	6.00E-04	1.50E-03	833	333
Iron (mild steel)	bending			2.5	5.5			4.50E-04	7.00E-04	1110	714
Lead	bending					20	160	4.00E-03	7.00E-03	125	71.4
Polymer concrete	bending							3.50E0-03		143	
Portland cement concrete	bending							1.20E-02		41.7	
Quartz (ground, piezo)	unspec.					65k		5.00E-06		100000	
Sand (loose on an Al beam)											
beam alone	bending					1000	4000	1.00E-03		500	
50% wt. layer of sand	bending					1000	4000	4.00E-02	9.95E-02	12.5	5.1
100% wt. layer of sand	bending					1000	4000	9.95E-02	4.10E-01	5.1	1.3
Silica (fused, annealed)	axial	73	1073					5.00E-07	5.00E-05	1000000	10000
Silicon nitride (n)	unspec.							1.25E-05		40000	
Soil (misc.)	unspec.					6	30	4.99E-02		10.0	

Table 7.3.2 Damping factors for a various materials for $\beta v \alpha = 1$.[57]

allows the design engineer to determine how long it may take for the part to reach thermal and hence dimensional equilibrium.[59] Radiation coupling is governed by the surface geometry, temperature difference, and thermal emissivity. The latter is a strong function of the surface finish and chemistry. The decision on which material to use based on thermal growth considerations may require careful finite element modeling or a simple experiment to simulate the operating environment and machine configuration.

Corrosion Resistance

Corrosion is a generic term that refers to many different types of material[60] performance degradation. The types of corrosion a design engineer must be aware of include:

- *Cavitation.* High flow rates over curved surfaces can lower the pressure (Bernoulli effect) and cause a vapor bubble to form. When the vapor bubble travels to a higher-pressure region, it collapses and acts like a mini hammer which slowly erodes the surface. This is a common failure mode of pump components.
- *Corrosion fatigue*. When a part is subject to cyclic stresses, it is often more susceptible to corrosion.
- *Dezincification*. The zinc in a zinc alloy (e.g., yellow brass) can corrode preferentially, leaving a porous skeleton of copper and corrosion products.
- *Fretting corrosion*. When two similar materials are in contact, slight relative motion (i.e., submicron or greater) between the two pushes aside any lubricating material and allows the materials' asperites to metallically bond, be ripped apart, and then oxidize. Continual motion pushes aside the corrosion debris and allows the process to continue.[61]
- *High temperature corrosion*. Temperature can speed many different types of corrosion processes.
- *Intergranular corrosion*. Localized attack of the material's grain boundaries, which often act as anodes, can occur.
- *Parting*. A material in an alloy can corrode preferentially (e.g., copper-base alloys containing zinc or aluminum).
- *Pitting*. Localized corrosion often occurs on the surface of an untreated part.

[59] See the conceptual design case study in Section 8.8.2.

[60] For a more detailed discussion corrosion, see, for example, H. Uhlig and R. Revie, Corrosion and Corrosion Control, John Wiley & Sons, New York, 1985.

[61] Fretting corrosion is discussed in more detail in Sections 5.6 and 7.7.

- *Stress corrosion cracking*. When a part is subject to a high constant (tensile) stress, it is often more susceptible to corrosion.
- *Transition zone corrosion*. A change of state (i.e., boiling liquid) enhances corrosion at the interface.
- *Uniform attack*. The entire surface of the part often corrodes (e.g., common rusting of iron). A design engineer must consider the effect of oxidation on the performance of the part and the environment the part will be used in. For example, a ground oiled cast iron surface will not oxidize as quickly as a ground steel surface because the former has tiny pores which hold the oil. On the other hand, if a process such as nitriding is used to harden the surface, then the surface will resist corrosion. Resistance to uniform attack is often obtained by forming a stable oxide on the surface (e.g., an oxide of aluminum or chrome).

Each of these types of corrosion can be brought about or enhanced by a variety of specific environments. For example, a machine tool with a poor coolant management system can be subject to an increasingly caustic wash of old coolant. It is the job of the materials engineer to query the design engineer about the operating conditions (and then double-check them) and choose the proper material and protection mechanism (e.g., nitriding or plating) for the application.

Remember, ignorance is no excuse in materials selection. For example, an architectural engineering firm designed a swimming pool to be installed on the top floor of a parking garage. To help carry the weight, they specified stainless steel tie rods, thinking that stainless steel is obviously impervious to corrosion. However, the stainless steel they chose was susceptible to stress corrosion cracking in the presence of chlorine ions. After the swimming pool crashed through the garage a couple years after it was installed, the jury found that the firm was at fault for not consulting with a materials engineer. The worst aspect, of course, was that several people were killed.

Related to corrosion resistance is the compatibility of a material with other materials, particularly in the presence of an electrolyte or corrosive media. For example, when a copper water pipe is joined directly to a galvanized water pipe, the copper has a higher potential than the zinc, and the zinc dissolves away from the steel pipe and plates itself inside the copper pipe. If the steel pipe does not then corrode and burst, the copper pipe will soon close up with a zinc plug. For machine tools, one must consider the presence of cutting fluid and the fact that it often deviates from neutral pH when the user does not change it or add the appropriate rust inhibitors. If steel and aluminum components are used in the machine, in an aqueous environment the aluminum may slowly dissolve.

Be careful and conservative when choosing materials. Do not be shy about calling a materials supplier and asking for assistance. If you do not get the answer you want (particularly ask why they made the decision they did so you can learn and cross-check recommendations from different suppliers), then call a supplier who will tell you what you want to know. Remember that if the design fails because the material fails, you are first in line to get blamed.

7.4 Structural Design

Like the bones in your body, the structure of a machine provides the mechanical support for all of the machine's components. When considered in the context of the design of the machine as a system, some of the major design issues include:

- Stiffness and damping
- Structural configuration
- Structural connectivity

Each of these issues is intrinsically related to the strategy the design engineer intends to use to obtain accuracy. As discussed in Section 2.1.1, accuracy is the relation of a measurement to a standard, and perhaps the primary goal of a system is to attain repeatability. Without repeatability, one can never hope to achieve accuracy. Repeatability, however, may be necessary, but it is not a sufficient purveyor of accuracy. Accuracy must be obtained by one or more of the following means:

- Accuracy obtained from component accuracy
- Accuracy obtained by error mapping

- Accuracy obtained from a metrology frame

Which strategy the design engineer chooses can have an enormous impact on the rest of the design process. One common denominator, however, is the design axiom that one should always strive to obtain as good a level of accuracy from the components as is economically possible before more advanced design strategies are applied. This is due to the fact that the more advanced design strategies seem to be able to provide only a certain level of improvement (e.g., a tenfold improvement when error mapping is used). It is not unusual to find that the controllability of a mechanical system can sometimes be related to the quality (potential accuracy) of its components.

7.4.1 Stiffness and Damping

As shown in Figure 3.1.3, the response of a structure to a time-varying input depends on the stiffness, damping, and mass of a structure. Hence good stiffness and damping are each necessary but not individually sufficient requirements for a precision machine. One of the most commonly asked questions is: How stiff does the structure need to be? Section 10.2.1 addresses this question in the context of actuators, and the discussion applies here as well. Furthermore, methods for analyzing the stiffness (deflection) of a structure have been discussed in detail in this book, and most engineers are not lacking in their exposure to stiffness design issues. Sadly, it is often assumed that very little can be done to increase the damping of a structure; and as shown in Figure 3.1.3, the damping can have a significant effect on performance as stiffness.

Machine tools traditionally have been built of cast iron, which has moderately good damping properties. When more damping was required for machines that made heavy cuts or were subject to high-frequency vibration (e.g., grinders), the structure's cavities were sometimes filled with lead shot and oil for viscous and mass damping or concrete for mass damping. Polymer concretes were then developed, and they had twice the damping capacity of cast iron and were easier to cast. Polymer concretes are now used extensively for machines subject to high levels of vibration (e.g., grinders). However, as higher speeds and greater accuracies are sought, even better means to damp vibration will be required. Mechanisms to damp structures include shear plate and tuned mass dampers.

Constrained Layer Damper Design

The structural joints in a machine tool have long been known to be a source of damping by the mechanisms of friction and microslip. A study of structural joint damping has shown that numerous theories are available for predicting damping by these mechanisms[62]; however, the amount of damping obtained is still orders of magnitude less than what is required for critical damping, and controlling the surface interface parameters at the joint to achieve uniform results from machine to machine is difficult. In addition, as far as the accuracy of the machine is required, it would be best if the joints behaved as a rigid interface which as discussed below can be obtained by bolting and grouting (or bonding) a joint. A better method is to add damping applying alternate layers of viscous and structural materials to the surface of a machine tool structure,[63] as shown in Figure 7.4.1.

In order to design this type of damping mechanism into a structure, consider that damping can be achieved by friction caused by relative motion, and motion of a structure is generally greatest far from the neutral axis. The greatest degree of damping can therefore be attained by relative motion between two structures whose outer interface surfaces move in opposite relative directions. To accommodate this type of large motion and to dissipate large amounts of energy, a viscous or viscoelastic material is needed. For two parallel flat plates a distance h apart moving relative to each other by a velocity v_{rel}, the shear stress in the viscous material will be

$$\tau = \frac{\mu v_{rel}}{h} \tag{7.4.1}$$

As an example of how to calculate the amount of damping (power dissipation) that can be obtained, consider the simple model of Figure 7.4.1 where a time-varying force F excites the struc-

[62] See, for example, M. Tsutsumi and Y. Ito, "Damping Mechanism of a Bolted Joint in Machine Tools," Proc. 20th Int. Mach. Tool Des. Res. Conf., Sept. 1979, pp. 443–448; and A. S. R. Murty and K. K. Padmanabhan, "Effect of Surface Topography on Damping in Machine Joints," Precis. Eng., Vol. 4, No. 4, 1982, pp. 185–190.

[63] See, for example, S. Haranath, N. Ganesan, and B. Rao, "Dynamic Analysis of Machine Tool Structures with Applied Damping Treatment," Int. J. Mach. Tools Manuf.Vol. 27, No. 1, 1987, pp. 43–55.

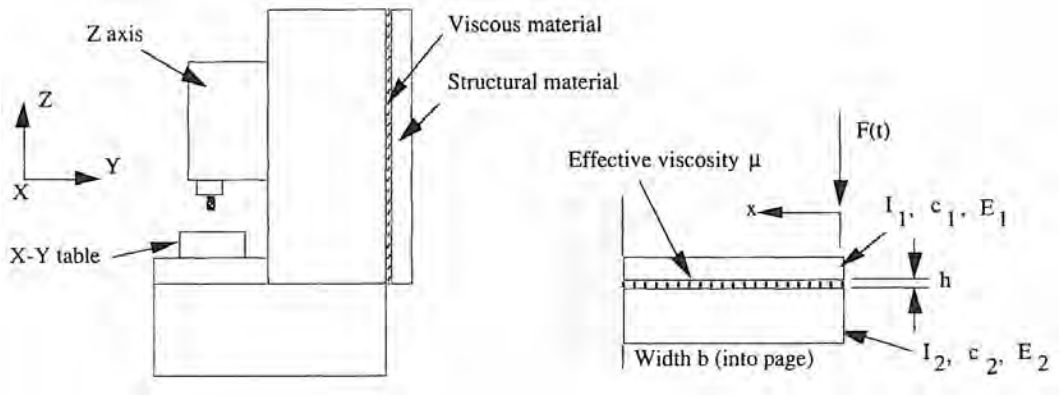


Figure 7.4.1 Increasing structural damping by adding alternate layers of viscous and structural materials.

ture. Shear deformations will not cause relative axial motion between the two structural beams, so only bending needs to be considered. The two beams must have the same displacement, therefore

$$\delta = \frac{F_1 L^3}{3E_1 I_1} = \frac{F_2 L^3}{3E_2 I_2} \tag{7.4.2}$$

$$F_1 + F_2 = F \tag{7.4.3}$$

Each beam thus supports a force of

$$F_1 = \frac{FE_1 I_1}{E_1 I_1 + E_2 I_2} \qquad F_2 = \frac{FE_2 I_2}{E_1 I_1 + E_2 I_2} \tag{7.4.4}$$

The stress as a function of position along the beams is given by $\sigma = Fxc/I$, where c is the distance from the beams' respective neutral axes to the outer surfaces. The resultant strain in the outer surface is just $\varepsilon = \sigma/E$. Over a distance dx on the surface, the change in axial length is εdx. At a point x along the beam, therefore, the axial displacement will be

$$\Delta_{axial} = \int \frac{cFx}{EI}\, dx = \frac{cF(L^2 - x^2)}{2EI} \tag{7.4.5}$$

The velocity v_{rel} is just $d\Delta_{axial}/dt$, and note that F is the only function of time. The power dissipated along the length of the beam is the product of the force and the velocity, where the force is the product of the shear stress and the beam width b and length dx along the beam:

$$\begin{aligned} \Delta\mathcal{P} &= \int d\Delta\mathcal{P} = \int \frac{\mu b v_{rel}^2}{h}\, dx = \left(\frac{dF}{dt}\right)^2 \frac{\mu b c^2}{4E^2 I^2 h} \int_0^L (L^2 - x^2)^2\, dx \\ &= \left(\frac{dF}{dt}\right)^2 \frac{2\mu b L^5 c^2}{15 h E^2 I^2} \end{aligned} \tag{7.4.6}$$

For the two-beam system, one beam's surface is in compression while the other beam's adjacent surface is in tension. With Equations 7.4.4 and 7.4.6, the power dissipation is found to be

$$\Delta\mathcal{P} = \left(\frac{dF}{dt}\right)^2 \frac{2\mu b L^5 (c_1^2 + c_2^2)}{15h(E_1 I_1 + E_2 I_2)^2} \tag{7.4.7}$$

This basic equation can be used in many ways to optimize a structure and its damping mechanism before it is designed, or to optimize the design of a damping mechanism for an existing structure.

Consider Equations 7.3.5 and 7.3.9 and Figure 3.1.3. For the natural frequency of the structure not to be a potentially limiting performance factor, there must be no amplification at resonance. For a damped structure modeled as a second order system, the results of the magnitude of Equation

3.1.6 (let s = jω) can be used to show that damped natural frequency and the frequency at which maximum amplification occurs are

$$\omega_d = \omega\sqrt{1-\zeta^2} \tag{7.4.8a}$$

$$\omega_{dpeak} = \omega\sqrt{1-2\zeta^2} \tag{7.4.8b}$$

The amplification at the damped natural frequency and the peak frequency can thus be shown to be

$$\frac{\text{Output}}{\text{Input}} = \frac{1}{\sqrt{4\zeta^2-3\zeta^4}} \tag{7.4.9a}$$

$$\frac{\text{Output}_{peak}}{\text{Input}} = \frac{1}{2\zeta\sqrt{1-\zeta^2}} \tag{7.4.9b}$$

For unity gain or less, the damping factor ζ must be greater than 0.707. As shown in Table 7.3.2, most structural materials yield damping factors that are orders of magnitude lower. Structures designed from classical materials have been used successfully used for years in a wide variety of machines. However, as nanometer accuracies or very high operating speeds are sought, and difficult machining operations are undertaken (e.g., ductile machining of brittle materials[64]), the limitations of existing designs will begin to manifest themselves more often. Fortunately, with simple methods, such as modeled above, critical damping can be achieved as described below.

In order to make use of Equations 7.3.5 and 7.3.9, the total energy lost per cycle and the energy input per cycle must be found. Assuming that the force input is sinusoidal, F(t) = Asinωt:

$$\begin{aligned}\Delta U_{cycle} &= \frac{2\mu b L^5(c_1^2+c_2^2)}{15h(E_1I_1+E_2I_2)^2}4A^2\omega^2\int_0^{\frac{\pi}{2\omega}}\cos^2\omega t dt \\ &= \frac{2\mu b L^5(c_1^2+c_2^2)A^2\pi\omega}{15h(E_1I_1+E_2I_2)^2}\end{aligned} \tag{7.4.10}$$

Conservatively, the energy input to the cantilever beam during the cycle is the time integral of the product of the force and the velocity:

$$\begin{aligned}U_{cycle} &= \int_0^{\frac{2\pi}{\omega}} F\frac{dF}{dt}\frac{L^3}{3(E_1I_1+E_2I_2)}dt \\ &= \frac{L^3}{3(E_1I_1+E_2I_2)}4A^2\omega\int_0^{\frac{\pi}{2\omega}}\sin\omega t\cos\omega t dt = \frac{2L^3A^2}{3(E_1I_1+E_2I_2)}\end{aligned} \tag{7.4.11}$$

With Equations 7.3.5, 7.4.10 and 7.4.11, this results in the following:

$$\delta_{Ll} = \frac{\mu\pi b L^2(c_1^2+c_2^2)\omega}{10h(E_1I_1+E_2I_2)} \tag{7.4.12}$$

We are interested in the damping at the damped natural frequency, so for the cantilever beam model, the first eigenfrequency can be used: $\omega_{1d} = 3.52\{(1-\xi^2)(E_1I_1+E_2I_2)/[(A_1\rho_1+A_2\rho_2)L^4]\}^{1/2}$, where A_i and ρ_i are the cross-sectional areas and densities, respectively. The mass of other axes and internal ribbing must also be considered. As a conservative estimate, the lumped mass M_{lump} will be added to the distributed mass. The log decrement is found from

$$C = \frac{3.52\,\mu b\,(c_1^2+c_2^2)}{h\sqrt{(E_1I_1+E_2I_2)(A_1\rho_1+A_2\rho_2+M_{lump}/L)(1+h/L)}} \tag{7.4.13a}$$

$$\delta_{Ld} = \pi\sqrt{-2+\frac{\sqrt{100+C^2}}{5}} \tag{7.4.13b}$$

Most structures are short and stocky; thus the factor (1 + h/L), where h is the height of the beam, is added to make an estimated account for the decreased stiffness caused by shear deformations, which the classical eigenfrequency solution does not account for. I, A, and c will all be set by the stiffness

64 See, for example, K. Puttick et al., "Single-Point Diamond Machining of Glasses," Proc. R. Soc. Lond. A, Vol. 426, 1989, pp. 19–30.

requirement of the machine. For many types of structures a large increase in structural damping (by a factor of 5 or more) can be achieved if:

Forming a small gap over a large area may seem like an expensive proposition; however, using the process of replication, a smooth surface can be manufactured economically on both the structure and the damping mechanism. The viscous film is maintained by smearing the high-viscosity fluid on the plates and then pushing them together using a spring preloading mechanism. Bolts alone may have the tendency to cause metal-to-metal contact if they are too tight. On the other hand, if they are too loose, the added structural layer may peel away from the primary structure. Sometimes achieving the critical damping value of $\zeta = 0.707$ requires the use of several layers of damping mechanisms. In other cases, an order-of-magnitude increase in the structural damping coefficient will be deemed sufficient. Figure 7.4.2 shows a design example. Although it may be possible to model the overall structure as a beam, the structure will often be made of plate elements (e.g., a welded structure or a very large casting). In addition to a large overall damping mechanism, individual plate elements within the structure should have damping mechanisms attached to prevent them from resonating.[65]

- A very viscous fluid is used (e.g., $\mu = 2\ 10^4$ centipoise = 20 N-s/m^2).
- A small gap is used (e.g., h = 1 to 5 μm).

Viscosity (N-s/m^2)	20	Added lumped mass (kg)	500
Gap h (μm)	2	h_{inner}= factor x h_{outer}	0.9
Width (m)	1	Beam thickness (m)	1.107
Length (m)	2	Beam wall thickness (cm)	5.5
Elastic modulus: $E_1 = E_2$(Pa)	1.2E+11	Damping plate thickness (cm)	1.0
Density: $\rho_1 = \rho_2$(kg/m^3)	7300	δ_{Ld}	0.539
$K_{desired}$(N/m)	1.75E+09	ζ	0.0855
Amplification at ω_d:	5.82		

Figure 7.4.2 Example of achievable structural damping with a single damping plate.

In addition, it should be noted that the dynamic response of the machine will depend on the inertia, stiffness, and damping. Often, for a particular application it is desirable to tune the machine by careful control of all three parameters.[66]

The amplification at a particular frequency can also be minimized with the use of a tuned mass damper. A tuned mass damper is simply a mass, spring, and damper attached to a structure at the point where vibration motion is to be decreased. The sizes of the mass, spring, and damper are chosen so that they oscillate out of phase with the structure and thus dissipate energy. The design of tuned mass dampers is relatively straightforward,[67] and they have been used with great success in many different types of structures (e.g., the John Hancock Building in Boston and numerous offshore oil platforms). Since a structure has an infinite number of modes of vibration, tuned mass dampers are used primarily to prevent vibration when a machine has a vibration mode which is often excited and is performance degrading.

Tuned Mass Damper Design

In a machine with a rotating component (e.g., a grinding wheel), there is often enough energy at multiples of the rotational frequency (harmonics) to cause resonant vibrations in some of the machine's components. This often occurs in cantilevered components such as boring bars and some grinding wheel dressers. The amplification at a particular frequency can be minimized with the use of a tuned mass damper. A tuned mass damper is simply a mass, spring, and damper attached to a structure at the point where vibration motion is to be decreased. The size of the mass, spring,

[65] A commercially available product exists which consists of a thin metal sheet with a viscous glue on one side [available from Soundcoat Inc., 1 Burt Drive, Deer Park, NY 11729 (526) 242-2200]. The analysis above can be used to determine if the properties of this product are "optimal" for the intended application; if not, a custom damping mechanism can be designed as described above.

[66] See, for example, E. Riven and Hongling Kang, "Improvement of Machining Conditions for Slender Parts by Tuned Dynamic Stiffness of Tool," Int. J. Mach. Tools Manuf., Vol 29, No. 3, 1989, pp. 361–376.

[67] See, for example, L. Meirovitch, Elements of Vibration Analysis, McGraw-Hill Book Co., New York, 1975, pp. 93–100, 117.

and damper are chosen so they oscillate out of phase with the structure and thus help to reduce the structure's vibration amplitude.

Consider the single-degree-of-freedom system shown in Figure 7.4.3. The system contains a spring, mass, and damper. For a cantilevered steel beam, the spring would represent the beam stiffness, the mass would be that which combined with the spring yielded the natural frequency of the cantilevered beam, and the damper would be that which caused a 2% energy loss per cycle. As also shown in Figure 7.4.3, a second spring-mass-damper system can be added to the first to decrease the cantilevered beam's amplitude at resonance. The equations of motion of the system are

$$m\ddot{x}_1(t) + (c_1 + c_2)\dot{x}_1(t) - c_2\dot{x}_2(t) + (k_1 + k_2)x_1(t) - k_2x_2(t) = F(t) \tag{7.4.14a}$$

$$m_2\ddot{x}_2(t) - c_2\dot{x}_1(t) + c_2\dot{x}_2(t) - k_2x_1(t) + k_2x_2(t) = 0 \tag{7.4.14b}$$

These equations can be represented in matrix form as

$$\begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \ddot{x}(t) + \begin{bmatrix} c_1 + c_2 & -c_2 \\ -c_2 & c_2 \end{bmatrix} \dot{x}(t) + \begin{bmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 \end{bmatrix} x(t) \tag{7.4.15}$$

In the frequency domain, in order to present a solution for the motion of the system, the following notation is introduced:

$$Z_{ij}(\omega) = -\omega^2 m_{ij} + i\omega c_{ij} + k_{ij} \qquad i, j = 1, 2 \tag{7.4.16}$$

The amplitudes of the motions of the component and the damper as a function of frequency are given by[68]

$$X_1(\omega) = \frac{Z_{22}(\omega)F_1}{Z_{11}(\omega)Z_{22}(\omega) - Z_{12}^2(\omega)} \tag{7.4.17}$$

$$X_2(\omega) = \frac{-Z_2(\omega)F_1}{Z_{11}(\omega)Z_{22}(\omega) - Z_{12}^2(\omega)} \tag{7.4.18}$$

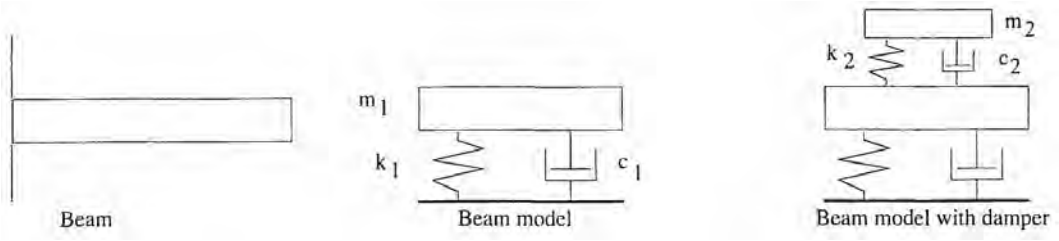


Figure 7.4.3 Cantilever beam, beam model, and beam model with tuned mass damper.

The design of a tuned mass damper system for a machine component may involve the following steps:

- Determine the space available for the damper and calculate the mass (m_2) that can fit into this space.
- Determine the spring size (k_2) that makes the natural frequency of the damper equal to the natural frequency of the component.
- Use a spreadsheet to generate plots of component amplitude as a function of frequency and damper damping magnitude (c_2).

As an example, consider the design of a tuned mass damper for an 80 mm diameter, 400 mm long steel cantilever beam. Figure 7.4.4 shows a portion of the spreadsheet for design of a damper for this system, and Figure 7.4.5 shows the dynamic response of the system.

When the damper consists only of a spring and a mass (c_2 = 0), the response of the beam at resonance can essentially be eliminated; however, the resonance response is then flanked by large-amplitude peaks. As the damper's damping coefficient increases, the response flattens out. As shown in Figure 7.4.6, viscous damping can be obtained by placing a cylinder in an oil-filled bore. Disk spring washers can be used to obtain the desired spring constant.

[68] Ibid., p 115.

Inputs			
Modulus (N/m^2)	2.07E+11	Cylinder diameter (m)	0.025
Density (kg/m^3)	7800	Cylinder length (m)	0.05
Length (m)	0.400	Damper density (kg/m^3)	7800
Diameter (m)	0.080	Damper fluid viscosity (N-s/m^2)	0.04
Bore radial clearance (μm)	1	Number of damping cylinders	2
Calculated beam properties			
Area A (m^2)	5.03E-03		
Inertia I	2.01E-06	**Calculated damper properties**	
First natural frequency (rad/s)	1207	Damper mass (kg)	0.53
Stiffness (N/μm)	19.5	Damper damping [N/(m/s)]	314
Equivalent mass (kg)	13.38	Damper stiffness (N/m)	767477
Damping value (c) [N/(m/s)]	51.43	Unit spring stiffness (N/m)	191869

Figure 7.4.4 Portion of a spreadsheet for the design of a tuned mass damper design for an 80-mm-diameter, 400-mm-long steel cantilever beam.

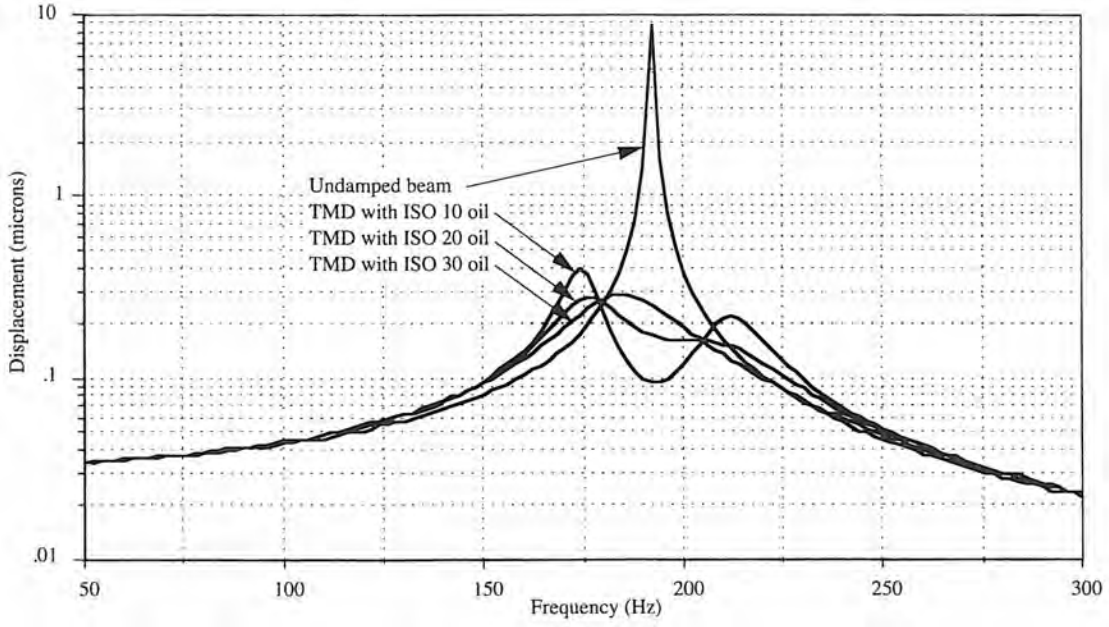


Figure 7.4.5 Numerical simulation of the dynamic response of an 80-mm diameter, 400-mm-long steel cantilever beam equipped with a tuned mass damper.

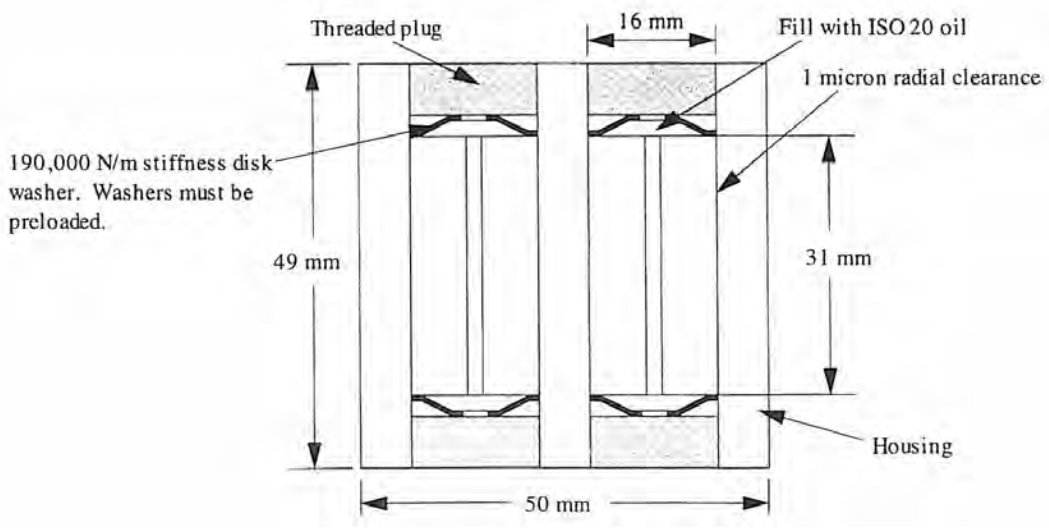


Figure 7.4.6 Cross section of a tuned mass damper design for an 80-mm diameter, 400-mm-long steel cantilever beam.

7.4.2 Structural Configurations

Perhaps the best way to approach the task of defining what the overall configuration of a machine will be is to ask yourself the following questions:

- What do the raw and finished parts look like?
- What sort of tools are needed to make the parts?
- What kinds of debris will be generated, and at what rate?
- How will the machine systems interact?

In the context of answering these questions, try to envision surfaces and how they can be configured around the part and tool. If the problem of structural configuration is approached as a problem of topology, creativity is likely to be enhanced. One of the key questions in topological problems is: What are the symmetries of the problem? Nature has shown that symmetry is beautiful and functional. One must also consider what types of structures already exist and the reasoning behind their development.[69] Now would be a good time to reread Sections 1.5 and 1.6.

For proper functioning of moving axes and operational stability, often it is important to minimize thermal and elastic structural loops. By this it is meant that the path length from the toolpoint to the workpiece through the structure should be minimal. The less material that separates the part of the structure that holds the tool and the part of the structure that holds the part, the more likely the entire system will quickly reach and maintain a stable equilibrium.

Note that in some cases, it may be better to design an axis's stiffness to be constant as opposed to designing it to also have maximum stiffness. As an example, consider the axis that supports the spindle of a jig borer or surface grinder. As the tool is moved into the workpiece, with a constant stiffness design, the error will be constant (if the load is constant which it should be for a finish cut). With a maximum stiffness design, the error will start out much smaller than for a constant stiffness design, but will increase to that of the constant stiffness design when the axis is fully extended. This can lead to the generation of taper in the part. Remember, on the other hand, maximum stiffness designs are still often needed because some machines need to have a maximum stiffness point for severe machining requirements.

Furthermore, taller machines are more likely to be affected by thermal gradients. One must be careful, however, because when designs are made too compact, room for support systems is sometimes inadequate. Furthermore, as the structure deforms under load, it changes the geometry in which the bearings are mounted, which can increase geometric errors. If the design is not kinematic, bearings can become overloaded by the forced geometric compliance between the structure to which the bearing rails or races are mounted and the smaller structure of the platten or spindle. Metrology frames or mapping techniques can be used to compensate for all or some of the errors due to structural deformation, but at a relatively high cost; thus cooling systems are usually preferred for control of thermal deformations and to help maintain bearing performance.

Open Section (C or G) Frames

Most small machines are designed with an open frame, as shown in Figure 7.4.7, which greatly facilitates workzone access for fixturing and part handling. Note that the machine could be designed with the spindle oriented in the horizontal or the vertical direction. Unfortunately, open section frames are the least structurally and thermally stable. The lack of symmetry leads to undesirable thermal gradients and bending moments. The fact that a critical part of the structure is cantilevered means that Abbe errors abound; hence great care must be used when designing a precision machine with an open frame. Note that there are many different variations on this design for different types of machines (e.g., milling machines and lathes). The common feature of all is the fact that the structural loop is open.

Closed Section (Bridge or Portal) Frames

Most large machines are designed with a closed frame as shown in Figure 7.4.8. When the Z motion is built into the bridge, a second actuator must often be slaved to the primary actuator

[69] A good overall reference is M. Weck, Handbook of Machine Tools, Vols. 1 and 2, John Wiley & Sons, New York, 1984. This reference supplies many design examples and shows how actual deformed shapes compare to those predicted by finite element methods. Also see R. R. Shandin and J. C. Wyan (eds.), Applied Optics and Optical Engineering, Vol. X, Academic Press, New York, pp. 251–387, for a detailed discussion of the design of diamond turning machines.

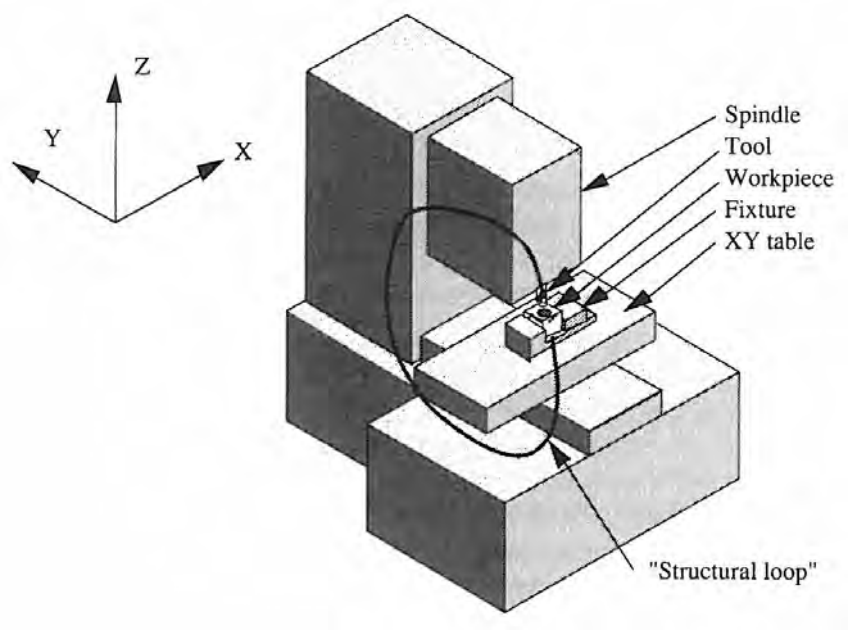


Figure 7.4.7 Structural loop in an open-frame machine tool.

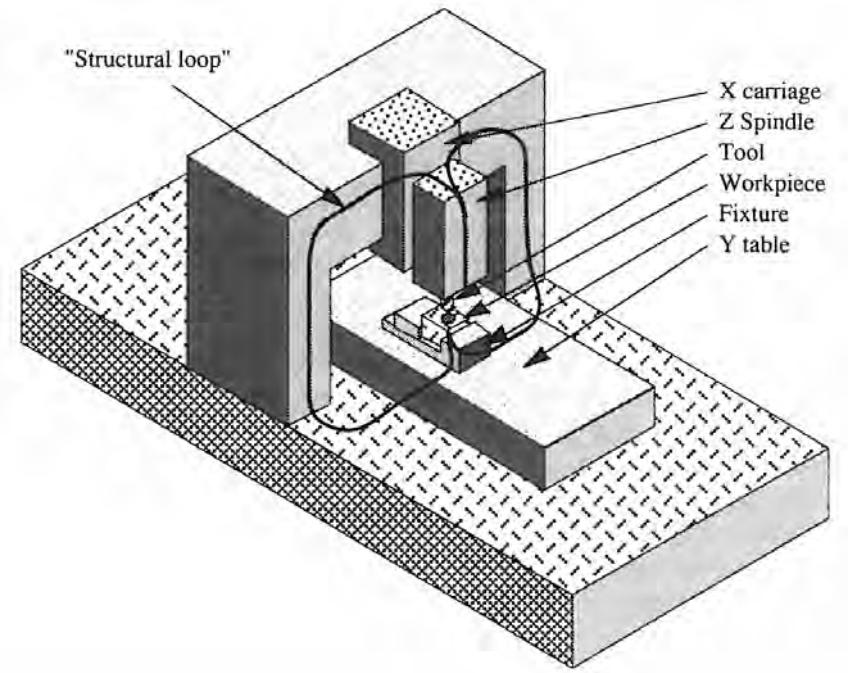


Figure 7.4.8 Structural loop in a closed-frame (bridge) machine tool.

that moves the bridge. This prevents the bridge from yawing (walking). Note that there are many different variations on this design for different types of machines (e.g., milling machines and lathes). The common feature of all is the fact that the structural loop is closed.

Tetrahedral Frames[70]

Nature invented the tetrahedron and found it to be an immensely stable and powerful form (e.g., diamonds). In engineering and architecture, the tetrahedron represents the three-dimensional application of the age-old structure of stability, the triangle. Lindsey of NPL in England took these basic building blocks of nature and added well-engineered damping mechanisms to yield a major advancement in the structure of machine tools. The *Tetraform* machine tool concept is shown in Figure 7.4.9. The structure owes its exceptional dynamic performance to the following:[71]

- Damping in the legs is achieved through the use of inner cylinders which dissipate energy through viscous shear. Energy is dissipated via squeeze film damping and relative sliding motion damping.
- Damping at the joints is achieved by application of the sliding bearing technology developed for the Nanosurf 2 (see Section 8.2.3). Briefly, Lindsey and Smith found that when PTFE is applied to a surface in layers less than 2 μm thick and the bearing slides on a polished surface, the resulting coefficient of friction is two to three times less than a thick PTFE layer and wear is almost eliminated. When the tension studs that hold the legs to the joint nodes are properly tightened, the legs of the tetrahedron are structurally decoupled (they behave independent of each other) due to the high degree of damping provided by the sliding bearing interface, yet the bearing interface's finite friction provides enough support so that the legs have a stiffness somewhere between that of a simply supported beam and a beam clamped at both ends.
- Microslip at the joints does not affect the dimensional stability of the machine because the minimum energy form of the tetrahedron wants to be preserved. Unlike a plane joint which can continue to slip and lead to dimensional instability, the tetrahedron's legs' spherical ends want to stay on the spherical joint nodes.

The latter point has even more profound consequences, in that it makes the use of composite materials in the structure an attractive alternative to metals or ceramics. Wound carbon fiber tubes can be designed to have a zero coefficient of thermal expansion along their length and they can have stiffness-to-weight ratios two times higher than are obtainable with metals. It would be difficult to design a conventional machine tool that made economical use of the desirable properties of carbon fibers.

The damping mechanism of the cylinder inside the tubes is more complex than that described previously in that the following must be considered:

- The neutral axes of the tubes are coincident; hence the inner tube should be mounted so that it does not bend within the outer tube or else there will be no relative motion between the two tubes. Note in the figure that the inner tube is shown held in tension by bolts and thus satisfies this condition.
- If the inner tube does not bend, the gap between the two tubes will change. This will result in the viscous fluid between the tubes being alternately squeezed out and sucked in as the outer beam vibrates. This results in *squeeze film damping*. Squeeze film damping will place normal forces on the inner tube, which will cause it to bend. The resultant bending will cause the amount of sliding motion damping to be reduced, so the relative stiffness of the components must be chosen carefully.
- To accurately model the combined process of squeeze film and sliding motion damping would require straightforward but tedious analysis. For purposes of illustrating how the relative dimensions of the tubes can be optimized, it will be assumed that the plus and minus effects described above cancel and the power dissipation is derived from Equation 7.4.6, where $c = R_I \sin\theta$ and $b = R_I d\theta$:

[70] The author would like to thank Kevin Lindsey of the National Physical Laboratory in England for his kind hospitality and gracious supply of information, which enabled this section to be written.

[71] This concept is protected by worldwide patents. See, for example, UK patent 8,719,169, or contact the British Technology Group, 101 Newington Causeway, London, SE1 6BU, England.

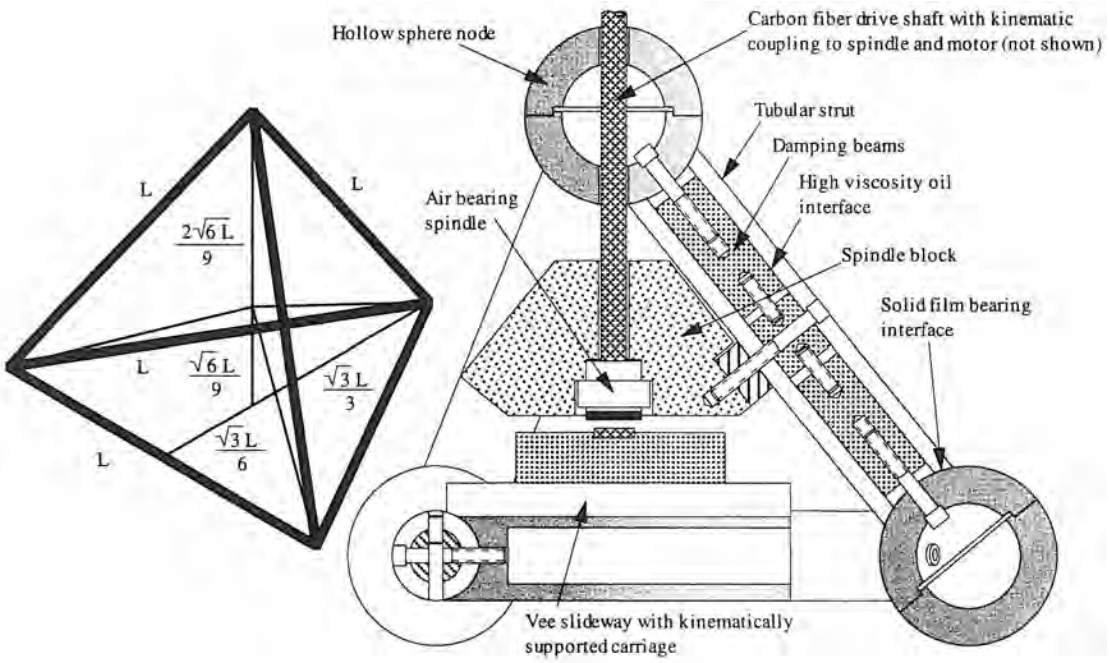


Figure 7.4.9 The Tetraform structural concept for machine tools and instruments developed by Lindsey. (Courtesy of the National Physical Laboratory.)

$$\Delta\mathcal{P} = \left(\frac{dF}{dt}\right)^2 \frac{2\mu L^5}{15E^2I^2}\int_0^{2\pi} R_1^3 \sin^2\theta d\theta = \left(\frac{dF}{dt}\right)^2 \frac{2\mu\pi L^5 R_1^3}{15hE^2I^2} \tag{7.4.19}$$

The structure can be designed for many different applications. For an instrument, damping may be of primary concern. For a machine tool. damping is important, but stiffness may be much more important. For the latter case, Figure 7.4.10 shows how the stiffness, power dissipated, and natural frequency can be used to "optimize" the inside diameter of the legs.

Counterweights and Counterbalances

Counterweights, shown in Figure 7.4.11, can be used to minimize gravity loads. This helps minimize the holding torque required of servomotors, which in turn minimizes motor size and heat input to the machine. For a very high precision machine, however, the bearings used to guide the counterweights and support the pulley bearings should have negligible static friction. Counterweights also increase the mass of the system, which can decrease dynamic performance. In the case of quasistatic axes (e.g., large gantry-type surface grinders), the counterweight can increase the load on the structure that supports the axes, and if the structure is cantilevered, then as the counterbalanced axis moves, the deflection errors caused by the weight of the counterbalanced axis will change. The latter effect can be decreased with the use of compensating curvatures, but they too have their problems. Alternatively, a separate frame and low-accuracy slave carriage can be used to support the counterweights. Thus counterweights are most often used on axes that are not supported by other moving axes (e.g., the vertical axis of a typical three-axis machining center) or in large bridge-type machines.

A counterbalance can be any passive means used to support the weight of an axis that moves in a vertical direction. Pistons have been used successfully but they can impart frictional and misalignment forces. A very effective counterbalance method is to use a float. This also provides viscous damping if the fluid used is a viscous oil. Some early ruling engines (machines used to make diffraction gratings) used carriages guided by kinematically arranged sliding contact bearings where most of the carriage weight was supported by pontoons in oil-filled troughs, which also help to provide damping.

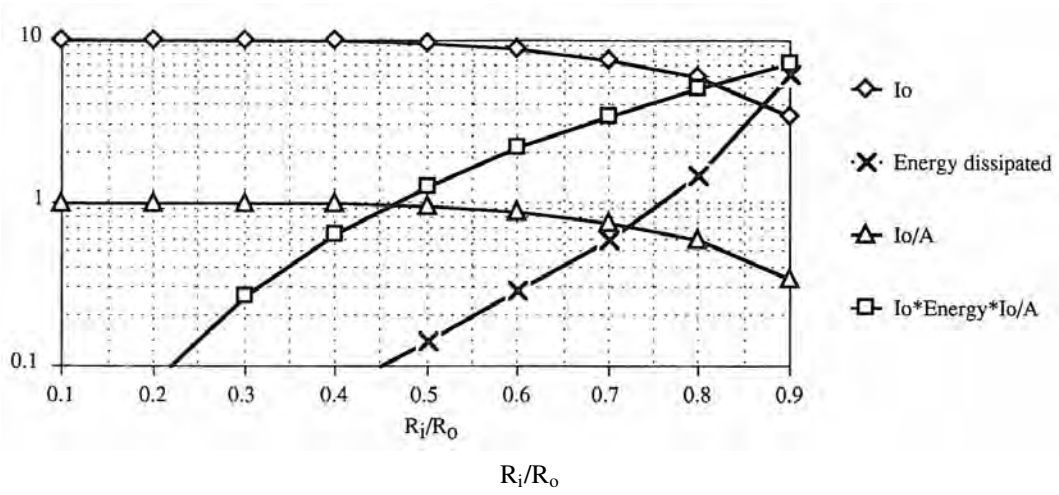


Figure 7.4.10 Bending stiffness ($\propto_o$), natural frequency ($\propto_o$/A), and damping ($\propto_i^3/I_o^2$) optimization of the tubular Tetraform members with a solid tensile rod for a 1-m-long tube with an outside diameter of 10 cm.

Compensating Curvatures

All structures have finite stiffness, and when loads are applied, lateral and angular displacements will result. To compensate for these effects, it is possible to shape the otherwise straight ways of an axis so that the sum of the deflections and the intentional nonstraightness results in minimal net straightness error. This type of correction is known as a *compensating curvature*. Usually, it is very difficult to also compensate for angular errors. If the error budget is properly assembled, it can be used to plot the total error as a function of the position of an axis. Once the plot is found, it can be used to help design a shape (ideally, the inverse of the plot) that will cancel the lateral and hopefully also angular errors for the given bearing design. Sometimes the shape of the compensating curvature is determined by measuring errors on a prototype and then refinishing the bearing ways accordingly. Complex compensating curvatures, however, can be expensive to manufacture because they may require a large surface grinder with two-axis contouring capabilities (rare) or hand finishing of components. Simple (e.g., a bow) compensating curvatures can be made on a surface grinder by loading the to-be-ground structure at strategic points. Compensating curvatures can easily be measured using an autocollimator.

When the load does not greatly change on parts moving along axes that are curvature compensated (e.g., axes that carry a measuring head), the method can be effective. If the loads do change greatly (e.g., a table carrying different weight parts and fixtures), the compensating curvature can sometimes decrease performance. One of its main attractions, however, is that once the correct compensated curvature and manufacturing process is found for a particular machine, it can be the least expensive way to correct for straightness errors.

With modern mapping techniques, compensating curvatures are used primarily on very large structures or when angular errors caused by deflection are too large and cannot be corrected for by another axis. For example, they might be used in a situation where otherwise the toolpoint would enter the workpiece at an angle rather than being perpendicular. A rotary axis would be required to compensate for this type of angular error, as opposed to just another linear axis, which can only be used to compensate for an Abbe error.

Temperature Controllability

The need for careful consideration of thermal errors and temperature control, discussed in Section 2.3.5, is reemphasized here. This is due to the fact that how the temperature is to be controlled can have a large impact on the machine's design. For example, consider the following temperature control design strategies in the context of the representative machine configurations discussed above:

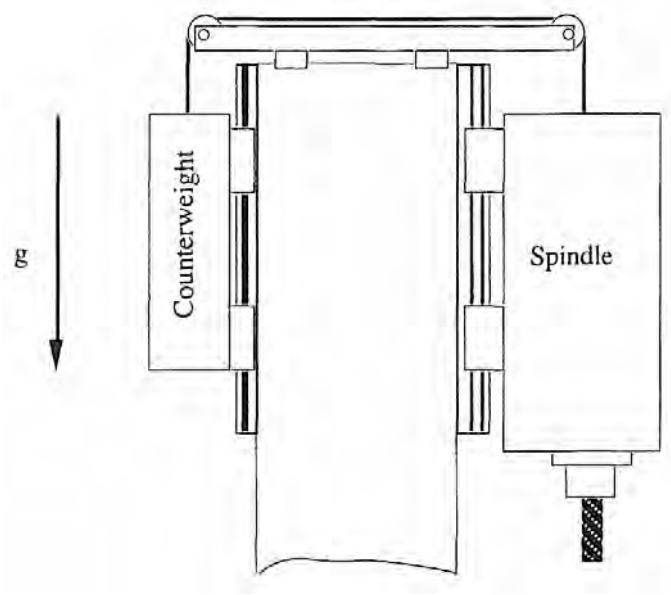


Figure 7.4.11 Use of a counterweight to balance the deadweight of an axis.

- Passive temperature control:
  - Minimize and isolate heat sources.
  - Minimize coefficient of thermal expansion.
  - Maximize thermal conductivity to minimize thermal gradients.
  - Maximize thermal diffusivity to quickly equilibrate transient thermal effects.
  - Minimize thermal emissivity of the structure to minimize radiant coupling to the environment, or maximize the emissivity to couple the structure to an environmental control enclosure.
- Active temperature control:
  - Air showers can be used to control the temperature and minimize thermal gradients in the environment around the machine.
  - Circulating temperature-controlled fluid within the machine can control temperature and minimize gradients.
  - Oil showers can be used to control the machine's temperature and minimize the effects of the external environment on the machine.
  - Thermoelectric coolers can be used for precise, fast, dynamic temperature control of hot spots.

Each conceptual design option must consider how the temperature within the machine will vary if each of these different temperature control strategies were applied.

7.4.3 Structural Connectivity

Most people have had the displeasure of having an end table or chair wobble because either one leg was too short or the floor was not level. On the other hand, many people have also had the displeasure of tipping over backwards in an office chair with four legs (good-quality swivel office chairs have five legs). These two examples illustrate the principle of *kinematic design* and the principle of *elastic averaging*.

The principle of kinematic design, states that point contact should be established at the minimum number of points required to constrain a body in the desired position and orientation (i.e.

six minus the number of desired degrees of freedom). This prevents overconstraint, and thus an "exact" mathematically continuous model of the system can be made. Kinematic locating mechanisms range from simple pins to Gothic arch-grooved three-ball couplings shown in Figure 7.4.12. Kinematic designs, however, are subject to high-contact stresses, which often may require the use of ceramic components (e.g., silicon nitride balls and grooves) if the highest level of performance is to be achieved. If stress and corrosion fatigue are controlled, repeatability of a kinematic system can be on the order of the surface finish of the points in contact if the loads are repeatable or the preload high enough (see Section 7.7 for a detailed discussion). Finite contact areas do exist, and they effectively elastically average out the local errors due to surface roughness. In addition, note that friction and microindentation will limit the accuracy of the kinematic model.

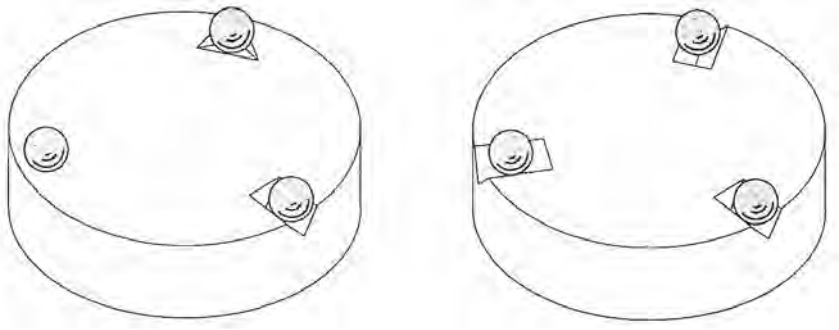


Figure 7.4.12 Flat-groove-tetrahedron (Kelvin clamp) and three-vee-groove kinematic couplings. In both cases, the balls are mounted to an upper plate (not shown) which is held in position with respect to the lower plate by the coupling.

Kinematic support of a structure is often desired to ensure that the structure is not deformed by inaccuracies or instabilities of the mounting surface. For a small instrument, only one of the kinematic mounting points may be rigidly connected and the other two may consist of flexures that will allow for differential thermal growth between the instrument and the mounting surface. Remember that friction does exist in kinematic couplings, and thus forces can be transmitted between a mounting surface and an instrument.

Any machine as small as a desk can usually be made to support its own weight, so a kinematic three-point footing can be used to make the machine's accuracy insensitive to the levelness of the floor. For some structures, such as surface plates or large mirrors, in order to minimize deformations, more than three points are required to support the structure, yet one cannot afford to take the chance that the support itself may be uneven. As shown in Figure 7.4.13, a stacked arrangement of two- or three-point supports on balance beams can allow for multiple support points on the structure while requiring only a three-point mount on the floor. If you examine the windshield wiper blade holders on various automobiles, you will notice that most use a similar cascaded system to ensure that the wiper maintains uniform pressure on the window, even as the curvature changes beneath the wiper blade as it wipes the window.[72]

The principle of *elastic averaging* states that to accurately locate two surfaces and support a large load, there should be a very large number of contact points spread out over a broad region. Examples include curvic or Hirth couplings, which use meshed gear teeth (of different forms respectively) to form a coupling. The teeth are clamped together with a very large preload. This mechanism is commonly used for indexing tables and indexing tool turrets. Figure 7.4.14 shows an indexing and clamping mechanism for a lathe tool turret. Note that many different types of clamping preload systems exist. However, this type of mechanism causes the system to be overconstrained; on the other hand, if an elastically averaged system is properly designed, fabricated, and preloaded, the average contact stress will be low, high points will wear themselves in with use, and errors will be averaged out by elastic deformation. The system itself will then have very high load capacity and stiffness. For a worn-in elastically averaged system, the repeatability is on the order of the accuracy of the manufacturing process used to make the parts divided by the square root of the number of

[72] Once again, you as a designer are advised to remove your headphones and continually observe and analyze everything around you.

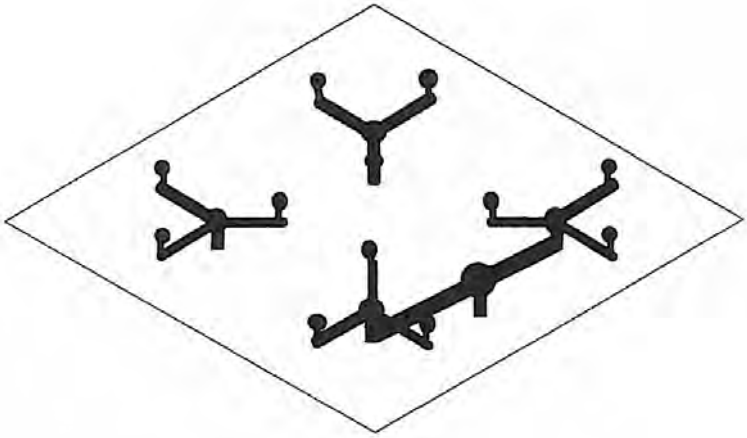


Figure 7.4.13 Use of three-point mounts with ball joints to form a *wiffle tree* to achieve greater support while preventing geometric overconstraint.

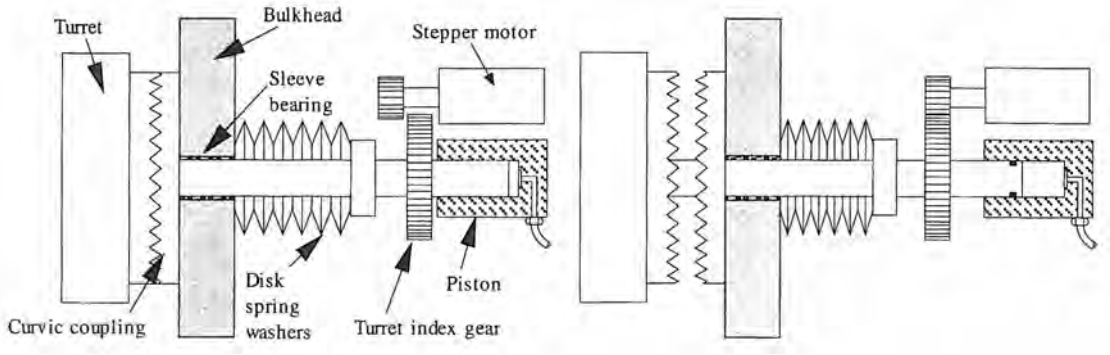


Figure 7.4.14 Typical turret indexing and locking mechanism in engaged and disengaged positions. (Courtesy of Hardinge Brothers, Inc.)

contact points (i.e., teeth in a toothed coupling). Still, because of the large number of contact points, the chance of dirt contaminating the interface increases.

A machine used to manufacture the wings of an airplane is so huge that for support it must rely on many contact points with a level, aged concrete floor of the proper thickness that has been poured on an appropriate subgrade.[73] The point where one switches from kinematic to multipoint mounting, and how many mounting points should be used, is somewhat vague. Based on observations of various machine tools, typically a machine tool with a footprint larger than 4 m^2 will usually require a multipoint mounting system, and mounts should be located about every 500 to 1000 mm. Accordingly, this often requires the adjustment of the mounts to be made in conjunction with measurement of the machine's performance on site. The use of resilient rubber pads under the mounting feet can lessen the sensitivity of the structure to variations in floor flatness and can increase the vibration isolation between the machine and the floor. Resiliant mounts also decrease the likelihood of deformation of the machine due to dimensional changes in the floor. When a large machine is supported by a foundation, the engineer should also consider that the bottom of the foundation will often remain at a constant (deep ground) temperature. The top of the foundation, however, will be exposed to the shop environment. This can result in warping of the foundation as the shop temperature changes with the seasons unless the shop temperature is carefully controlled. It may be possible for the foundation to be instrumented with thermocouples and the errors mapped and used to make corrections in the machine or as a guide in readjusting the machine's mounting pads' height as the

[73] See, for example, B. S. Baghshahi and P. F. McGoldrick, "Machine Tool Foundations - A Dynamic Design Method," Proc. 20th Int. Mach. Tool Des. and Res. Conf., Sept. 1979, pp. 421–427.

seasons change. The foundation must also be kept dry because concrete absorbs moisture and can swell.

With respect to the connectivity between structural elements (e.g., a headstock and a bed), elements of a structure can either be connected together via:

- Kinematic design:
  - Deterministic
  - Less reliance on manufacturing
  - Stiffness and load capacity limited
- Elastically averaged design:
  - Nondeterministic
  - More reliance on manufacturing
  - Stiffness and load capacity not limited

It should be noted that overconstrained systems cannot accommodate differential thermal growth and hence are more prone to warping. Furthermore, finite deformation of the contact surface leads to micromechanical constraints that limit the true kinematicity of the structure. The use of hard materials (e.g., ceramics) helps to minimize the latter problem. For permanently connected components, one can align them using a kinematic location system and then inject epoxy or grout to create a bond between all the surfaces in close proximity. Although with this method, one must make sure that the shrinkage of the bond material does not warp the components.

Which is better, the principle of kinematic design or the principle of elastic averaging? Consider Evans' comments in his assessment of the history of precision engineering: "Contrast, for example, Pollard, in the introduction to his monograph on instrument design, bewailing nonkinematic design practices he attributes to machine tool design practice being brought to instrument design, and Rosenhain who bewails the flimsy designs of instrument makers and calls for an approach more akin to machine tool design."[74]

7.4.4 Materials Considerations

For the conceptual design phase, one should design the structure using differently available materials, and then also design a multimaterial hybrid. For example, cast iron can be made into virtually any shape, so the design engineer has greater freedom, but large sections are expensive to thermally anneal, which must be done to achieve material homogeneity and stability. Granite is usually used in the form of simple rectangular, circular, or planar shapes. Ships are made from welded steel plate and thus conceivably any size of machine tool could also be welded together. Polymer concrete can be cast into virtually any shape and requires no stress relief or prolonged aging cycle. With new ceramics and composite materials, the choices become even more varied, so one really must be alert and consider all options.

Cast Iron Structures

The good general properties of cast iron and the ease with which parts can be cast have made cast iron the foundation of the machine tool industry. Generally, when a machine tool component is smaller than a compact car, it is a candidate for being made of cast iron, although castings weighing hundreds of tons have been made. For larger parts or where economy is of prime importance, one should consider welding together plates and standard structural shapes (e.g., boxes, angles, I-beams, and channels), as discussed below.

Granite Structures[75]

Granite is used as a structural material in machines that are generally used in dry environments because granite can absorb moisture and swell. Thus it may not be appropriate to use granite

[74] C. Evans, Precision Engineering: An Evolutionary View, Cranfield Press, Cranfield, Bedford, England, 1989.

[75] A good source of design information about what shapes and sizes can be made is available from Rock of Ages Corp., Industrial Products Group, P.O. Box 482, Barre, VT 05641.

in a machine where cutting fluid splashes all over it, although there is some debate as to how susceptible to swelling granite actually is. Granite can be sawed into a part of virtually any shape or size (deviations from round slabs or rectilinear shapes can be expensive). Since it is a brittle material, sharp corners are not allowed, and most structures are built from pieces that are bolted (using inserts) and grouted or bonded together. Since granite is brittle, threaded holes cannot be formed, and thus threaded inserts must be glued or press-fit into round holes in the granite. Common applications of granite components in machine tools include structural members and air-bearing ways in coordinate measuring machines and other inspection machines. Note that the porosity of some granite makes it unsuitable for air bearings even after it has been polished.

The low thermal conductivity of granite makes it slow to absorb heat. This makes granite, particularly black granite, susceptible to thermal distortions caused by the top surface absorbing heat from overhead lights. Granite's coefficient of thermal expansion is less than that of most metals, so in the process of manufacture, shipping, and use, one must consider how differential thermal expansion will affect a machine's performance if metal components have been bolted to it. Although granite has been sitting in the ground for millions of years and thus may seem to possess the ultimate stability, one must consider the effects of relieving the stress upon the granite's dimensional stability after it has been quarried. There are suppliers of very stable granite components, so the design engineer must carefully shop around. A very desirable property of granite (or any other brittle material) is that it will chip when banged instead of forming a crater with a raised lip (Brinell). Granite is also relatively inexpensive to quarry, cut, and lap. Hence granite is often the material of choice for coordinate measuring machine tables.

Welded Structures[76]

Welded structures can be made from any weldable alloy (e.g., iron alloys such as 1018 structural steel or Invar). Welded structures (weldments) are commonly used, for example, where (1) the cost of a large structure is to be minimized, (2) a high degree of material damping is not needed or the structure will be filled with a damping material, and/or (3) the part is too large to be cast and thermally stress relieved economically. If welded properly, so that the weld material is in a metallurgically stable form, residual stresses can sometimes be removed using vibratory stress-relief methods.

A welded structure is similar to a cast structure, in that strength and stiffness are achieved through the use of sections that are reinforced with ribs; consequently, structural analysis can be difficult and beyond the conceptual design phase often requires the use of finite element methods. Damping and heat transfer characteristics across welded joints are also difficult to model because they are dependent on depth of weld penetration, composition of the weld material, and contact pressure between member surfaces at the joints where the weld does not penetrate. One obvious solution is to specify full penetration welds. To minimize the cost of welded structures, the number of parts and linear meters of weld must be minimized. Furthermore, the more welds that are made, the greater the thermal distortion that is likely to occur as a result of the manufacturing process. However, if too few ribs are used, large plate sections can vibrate like drums. Damping and thermal performance of a welded structure can improve greatly if a viscous temperature-cooled fluid is recirculated throughout the structure (e.g., as was done with the LODTM described in Section 5.7), or if damping mechanisms are used as described above. As described in Section 7.8, a welded structure can also be used as a mold for a polymer concrete casting which creates a heavy but well-damped and stiff structure, but care must be taken to avoid bimaterial thermal deformation problems.

Concrete Structures[77]

Concrete is defined here as a material composed of an aggregate held together with a binder. Aggregates can range from regular crushed rock (e.g., granite) to ceramics (e.g., aluminum oxide or quartz). Binders can range from Portland cement to special epoxies. Concrete made from aggregate and Portland cement has been used for the principal structural element in some machine tools,[78] and these machines had much higher damping and thermal inertias than those of cast iron machines;

[76] A good reference to have is O. Blodgett, Design of Weldments, James F. Lincoln Arc Welding Foundation, P.O. Box 3035, Cleveland, OH 44104, 1963.

[77] See J. Jablonowski, "New Ways to Build Machine Structures", Am. Mach. Automat. Manuf., Aug. 1987. Also see Section 7.2.3.2 for a discussion of polymer concrete structures.

[78] See, for example, H. Sugishita et al., "Development of Concrete Machining Center and Identification of Dynamic and Thermal Structural Behavior," Ann. CIRP, Vol. 37, No. 1, 1988, pp. 377–380.

however, dimensional stability and the susceptibility of the cured concrete to water absorption and expansion make cement-based concretes unsuitable for precision machines as structural elements and in some cases marginal even as a surface on which to mount a large precision machine.[79]

Traditionally, when large cast iron machine tool parts are made, sand cores are used to form the internal ribbing. These cores are then broken up and removed during the cleaning process. Cast concrete structures, on the other hand, can be removed from a permanent mold after a few hours and thus there is motivation to make the mold as simple as possible, which means minimizing the amount of ribbing. Often instead of ribbing, foam cores are used to enable light-weight closed-form sections to be formed. Since the concrete casting processes are generally not sensitive to differential solidification and variation of material properties caused by hot spots at thick sections, the design engineer can more readily lean toward the conservative design side (uniform thick sections). Of course, one must consider the effects of added weight on the deformation of the machine structure and the costs associated with an increased amount of material.

Ceramic Structures[80]

The first introduction of a machine that was made almost entirely of ceramics was in 1984 at the Tokyo machine tool show. Although all-ceramic machines generally perform admirably, they have yet to compete economically with machines made from cast iron or polymer concrete. However, in the future as more and more ceramic components are made for use in consumer products (e.g., automobiles), it is likely that precision machines will contain more and more ceramic components.[81]

Hard materials (e.g., ceramics) offer advantages over conventional materials in terms of dimensional stability, strength, and stiffness over a wide range of temperatures. In applications ranging from adiabatic internal combustion engines for maximum efficiency to X-ray mirrors for X-ray photolithography, the ability to manufacture components from hard materials is clearly of vital importance to the future of the manufacturing industry. Unfortunately, most ceramic components are finish machined on machines built from cast iron and the abrasive nature of ceramics limits the life of these machines. Thus a new family of machine tools and machine tool components will have to be developed specifically for the manufacture of ceramic components. Consider several key properties of some ceramic materials that can help guide the design process for new machines:

- Most ceramic materials (e.g., aluminum oxide and silicon nitride) will not corrode in any fluid environment that might be used in the manufacture of ceramic components (i.e. fluids from oil to water).[82] Water might be an ideal fluid for hydrostatic spindle bearings because of its low viscosity and high heat capacity. However, the use of water as a lubricant requires very close tolerances to avoid inordinately high flow rates and Reynolds numbers. In spindles, the water can be collected and drained with minimal evaporation; however, for linear bearings, it would be more difficult to prevent the water from spreading over the length of the rails and so evaporative cooling effects would have to be carefully controlled.
- The more brittle a material is, the less plastic deformation that is generated during finish grinding or lapping; hence the surface is more likely to have a negative skewness. A surface with a negative skewness minimizes the need for a lubricant that has good wetting properties that allows for water to be considered as a lubricant. A surface with a negative skewness will also suffer less damage in the event that the lubricant is lost.
- The more brittle a material is, the less plastic deformation that is generated during finish grinding; hence the surface is less likely to contain high residual stress levels that can lead to dimensional instability. Dislocations are generated in any material during machining, which leads to some residual stress, and thus a free abrasive process (i.e., lapping) is most desired for final finishing regardless of the material the part is made from. In addition, unlike some metals, elements do not precipitate out of a ceramic material's microstructure with time (e.g., carbides do not form like in some iron alloys) so dimensional stability is enhanced.

[79] J. B. Bryan and D. Carter, "Straightness Metrology Applied to a 100 Inch Travel Creep Feed Grinder," 5th Int. Precis. Eng. Semin., Sept. 1989.

[80] See, for example, T. Ormiston, "Advanced Ceramics and Machine Design," SME Tech. Paper EM90-353, Sept. 1990.

[81] Y. Furukawa et al., "Development of Ultra Precision Machine Tool Made of Ceramics," Ann. CIRP, Vol. 35, No. 1, 1986, pp. 279–282.

[82] Aluminum oxide components are subject to stress corrosion cracking in aqueous environments and thus care must be taken to minimize tensile stresses. Note that silicon-based ceramics, which have predominantly covalent molecules, are far less reactive with water. Only at high temperature and pressure do silicon-based ceramics become appreciably affected by aqueous environments.

- Most ceramics are pure and thus the achievable surface finish is limited only by the size of the grain structure formed during the sintering process; however, most advanced ceramic materials have submicron sizes and this effect is usually not a problem the way it can be with metals. Also note that metals contain descrete hard particles that are dragged across the softer surface during machining which degrades surface finish.
- Ceramic materials have a high modulus that is good for machine stiffness, but some have poor thermal properties (e.g., alumina) that can lead to increased thermal deformations of the machine. Note that silicon carbide has very good thermal properties but it is considerably more expensive than alumina.
- Ceramic materials in intimate contact will not gall or fret the way many metals often do. However, when ceramic materials are in intimate sliding contact, traction forces can cause the local tensile strength to be exceeded, which produces surface cracks that lead to spalling. In such situations, it may be desirable to use a ceramic material with a high fracture toughness and tensile strength (e.g., silicon nitride). Thermal stresses can also initiate local or gross failure.
- Ceramic materials have a higher modulus of elasticity and lower density than do bearing steels; hence ceramic rolling elements have a smaller contact zone that leads to less heat generation. Hybrid bearings (e.g., metal rings and ceramic rolling elements) generate 30-50% less heat than do steel bearings. This means that grease can be used to lubricate the bearings at much higher speeds.[83] In general, for smaller diameter ultraprecision bearings, hybrid bearings can have up to three times the DN value of steel bearings (i.e., 4.5 million versus 1.5 million).
- Ceramic materials can have an allowable Hertz stress that is many times greater than steel. For example, silicon nitride has an allowable Hertz stress on the order of 6.9 GPa (10^6 psi); thus ceramic components are less likely to suffer from brinelling, which is a common source of error in a well-used machine tool.

Ceramics are brittle, which makes them easier to damage when subjected to shock loads; however, a precision machine should not be subjected to shock loads in the first place. For metrology masters (e.g., squares and straightedges), ceramics are so much lighter than conventional materials that they are less likely to be damaged (dropped) in the first place, and they are less likely to become scratched or worn in everyday use. Ceramic components can also have virtually unmatched dimensional stability. For general machine tool applications, such as bearing rails, the element geometry generally provides more than enough strength to withstand impact loads generated when a machine crashes. Polished ceramic parts also will not gall when they slide on each other the way a metal-on-metal system can. In the future, as tougher ceramics are developed for the aerospace and automotive industries, more and more machine components will be made from ceramics.

The brittle nature of ceramics is actually more of a blessing than a curse for precision machines and instruments. Because ceramics are brittle, minimal residual stresses are not imparted into the component during finish grinding. When finished with a free-abrasive process (i.e., lapping), residual stresses can be virtually eliminated. Thus ceramic bearing rails can be made straighter and with a higher degree of surface finish than can hardened steel-bearing rails. Ceramic chips generated during the grinding or lapping process are also brittle so the surface tends to have a negative skewness.

Because ceramics are brittle, generous radii must be used in all corners, and threaded steel inserts must be used if parts are to be bolted to ceramic components. To bond two ceramic components together, conventional adhesives can be used, or for high-performance applications, ceramic parts can be frit bonded. In frit bonding, a glass powder is applied to the two mating surfaces and then fused by heating the parts above the melting point of the glass. The bond will not be as strong as the ceramic, but it will be almost as stiff.

In general, since ceramics have a fully dense fine-grained structure, they have exceptional corrosion resistance. Ceramic rolling elements also will not fret a steel raceway and thus could increase the applicability of modular rolling element linear bearings for use in production CNC

[83] Steel bearings require oil mist lubrication at higher speeds. Introducing an oil mist (oil dripped into a high-pressure air stream) into a bearing increases the chance that water and dirt particles might be introduced into the bearing, which can lead to premature failure. Actually, a precision bearing cooled by an oil mist should have its air cleaned and dried to a level usually associated with air bearings (e.g., 3 μm filter and $H_2O < 50$-100 ppm).

machines that sometimes run for extended periods with one axis locked in place. Properties of common structural ceramic materials were given in Table 7.3.1.

There are many different types of ceramics for many different types of applications. Aluminum oxide has good overall properties for structural applications such as fluidstatic bearing rails and CMM structures. Zirconia is very tough, and its coefficient of thermal expansion matches that of steel, so it can be used as a bearing rail liner for rolling element bearings without worrying about bimaterial expansion problems. However, it should be noted that zirconia is a multiphase material and thus it is not suited for applications where high dimensional stability is required (e.g., in precision bearings). Silicon nitride has the best overall properties including very high toughness, which makes it ideal for rolling element bearings, but it is too expensive to use for large structural components. Silicon carbide and tungsten carbide are extremely hard and wear-resistant and they are often used in cutting tool applications.

Aluminum oxide components can be made by cold pressing followed by machining, firing and grinding. Note that there is significant volume shrinkage during the firing process. Hence designing ceramic structural parts often requires assistance from the manufacturer. In general, the same shape design rules apply as for metal castings, and the wall thickness should not be greater than about 25 mm. Ceramic components (e.g., those made from silicon nitride) can also be made by hot pressing followed by grinding and lapping.

Filamentary Composite Materials

Carbon fibers can have twice the strength and elastic modulus of steel and less than half the weight. The primary limiting factor in their application in the machine tool industry is cost and dimensional stability of the epoxy resin matrix that binds the fibers together. The former is due primarily not to the cost of raw materials, but to the cost of labor required to lay down sheets of the material (called "prepreg"), and then vacuum cure the assembly at a specific temperature. Note that winding a shape on a rotating mandrel is a fairly inexpensive process, so making tubes of graphite epoxy composites is fairly inexpensive. By controlling the wrap angle, various properties can be tailored, such as axial, lateral, or torsional stiffness, and axial coefficient of thermal expansion. The latter can actually be made equal to zero. Tubes are commonly found as shapes used in spindles,[84] and with the Tetraform machine tool concept, the use of tubes in machine tools made from filamentary composites may become commonplace.

For machine tools, which usually operate in humid environments created by the cutting fluids, water absorption might cause unacceptable dimensional instability problems. The composites industry is working on this problem for the aircraft industry; hence it should not be a problem for too long.

7.4.5 Design Example: A Cast Iron Surface Plate

As an example of a preliminary structural design optimization problem,[85] consider the design of a surface plate. There are many different surface plate designs, but the most common (and least expensive) have a thick top with heavy ribbing. As a result, they essentially behave like a T-shaped beam. Look at any bridge and you will see that I-beams are used for good reason: The more mass the farther away from the neutral axis, the stronger and stiffer the beam is. Remember that for machine tool structures which are blocky, shear stiffness is also important, so all the mass cannot be removed from the neutral axis area. Hence the ideal surface plate will have a solid top and bottom with a fair amount of ribbing joining the top and bottom. Holes through the ribbing could allow for placement of sand cores inside. In order to obtain a back-of-the-envelope estimation for the ribbing design of this type of surface plate, the following assumptions are made:

1. The casting thickness (of the ribs and top and bottom plates) is uniform.
2. The overall stiffness of the surface plate should be about equal to the stiffness of the plate that spans each cell.

---

[84] D. G. Lee et al., "Manufacturing of a Graphite Epoxy Composite Spindle for a Machine Tool," Ann. CIRP, Vol. 34, No. 1, 1985, pp. 365–369.

[85] There are numerous texts on design optimization methods such as G. Reklaitis, Engineering Optimization, John Wiley & Sons, New York, 1983. Some finite element codes are even available (e.g., ANSYS®) that adjust the mesh to allow the program to converge automatically on an "optimal" solution to a fairly well defined problem. M. Weck studies this type of problem in great detail in his book, Handbook of Machine Tools, Vol. 2, John Wiley & Sons, New York, 1984.

3. For maximum accuracy during scraping, the plate should be square so that it can be rotated 90 degrees and rechecked against other plates for twist. Thus it is also assumed that the cells are square. The cells could be hexagonal, but the design equations become too complex for presentation here.
4. The plate is about 1 m × 1 m and its thickness is on the order of 0.25 m, so bending deflections dominate and shear can be ignored. At worst, the shear deflection would equal the bending deflection, so the design stiffness should be assumed to be twice the actual desired stiffness.

The following analysis uses these assumptions to maximize the stiffness-to-weight ratio of the design using the parameters shown in Figure 7.4.15. First, the overall plate stiffness is assumed to be equal to that of a simply supported beam:

$$K_{plate} = \frac{48EI}{L^3} \tag{7.4.20}$$

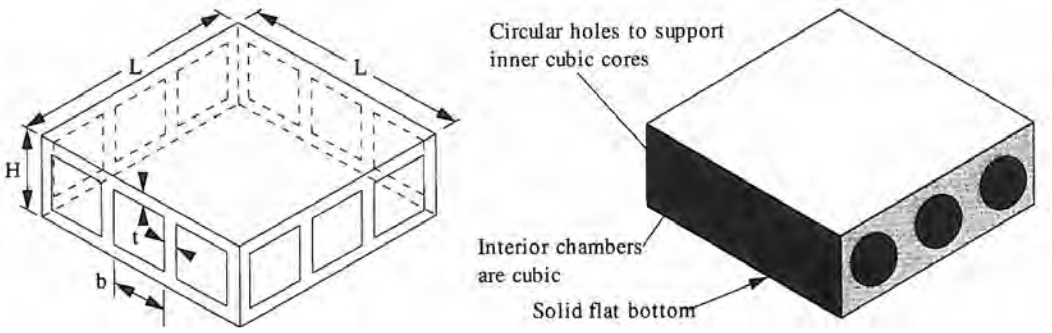


Figure 7.4.15 Cast iron surface plate with solid top and bottom surfaces for maximum stiffness and dimensions used for optimization analysis.

In reality the plate would be supported at three points or on 3 three-point mounts, as shown in Figure 7.4.13; however, the closed-form solution of this mounting is complex, so a conservative, simply supported beam model is used. From the parallel axis theorem, the second moment of the plate's cross-sectional area about the neutral axis (moment of inertia) is

$$I = 2\left[\left(\frac{H}{2} - \frac{t}{2}\right)^2 Lt + \frac{Lt^3}{12}\right] \tag{7.4.21}$$

Hence the plate stiffness is approximately

$$K_{plate} = \frac{24Et\left[H^2 - 2Ht + \frac{4t^2}{3}\right]}{L^2} \tag{7.4.22}$$

The length L is defined by the problem. H is specified as a proportion of L, typically H/L = 1/4 to 1/3. Here H is assumed to be equal to γL. For a balanced design, the cell stiffness must equal the plate stiffness, so each by itself equals twice $K_{desired}$ (referred to as just 2K):

$$\frac{-KL^2}{12Et} + H^2 - 2Ht + \frac{4t^2}{3} = 0 \tag{7.4.23}$$

For the cells, their stiffness is somewhere between that of plates simply supported and plates clamped around the edges. Assume that the cell stiffness is the average of these two cases:[86]

$$K_{cell\ avg.} \cong \frac{12Et^3}{(b-t)^2} = 2K \tag{7.4.24}$$

[86] The formulas for these two cases are found in R. J. Roark and W. C. Young, Formulas for Stress and Strain, 5th Edition, McGraw-Hill Book Company, New York, 1975, pp. 386 and 393. The plate width is the cell width b minus the wall thickness t.

To support the casting cores, all cell walls need access holes which are assumed to have an area

$$A_{\text{core hole}} = \frac{b-t}{2}\left(\frac{H-2t}{2}\right) = \frac{A_{\text{core}}}{4} \tag{7.4.25}$$

In reality the hole would probably be round or oval. The number of rib sections (walls), including the outside edges of the plate, is

$$N_{\text{rib}} = 2N + 2\sqrt{N} \tag{7.4.26}$$

Note that the number of cells N is assumed to be β^2, where β = L/b. The approximate volume of each cell wall is simply b(H - 2t)t, which assumes that the extra iron where the walls meet goes to the intersection detail such as shown in Figure 7.2.34. The total plate volume is thus

$$V + \frac{2Lt}{b}\left(\frac{L}{b}+1\right)\left[b(H-2t) - \frac{(b-t)(H-2t)}{4}\right] + 2L^2t \tag{7.4.27}$$

The mass M of the surface plate is the product of the density and the volume, which simplifies to

$$M = \rho Lt\left[\frac{1}{2b}\left(\frac{L}{b}+1\right)(H-2t)(3b+t) + 2L\right] \tag{7.4.28}$$

The goal is to minimize the mass while meeting the stiffness criteria:

$$\frac{L^2}{24ET\left[H^2 - 2Ht + \frac{4t^2}{3}\right]} + \frac{(L/\beta - t)^2}{12Et^3} \le \frac{1}{K} \tag{7.4.29}$$

Fortunately, neither the cell nor plate stiffness can drop below K_{desired} or the stiffness criteria will be unobtainable. Figures 7.4.16 and 7.4.17, respectively, show that as the number of cells increase, the mass and wall thickness decrease, although with diminishing returns. One would need to consult with the foundry as to the increase in cost associated with more cells and see how it affects the total price of the casting. As far as the mass of the casting that must be dealt with during manufacturing and shipping, the differences are not that large.

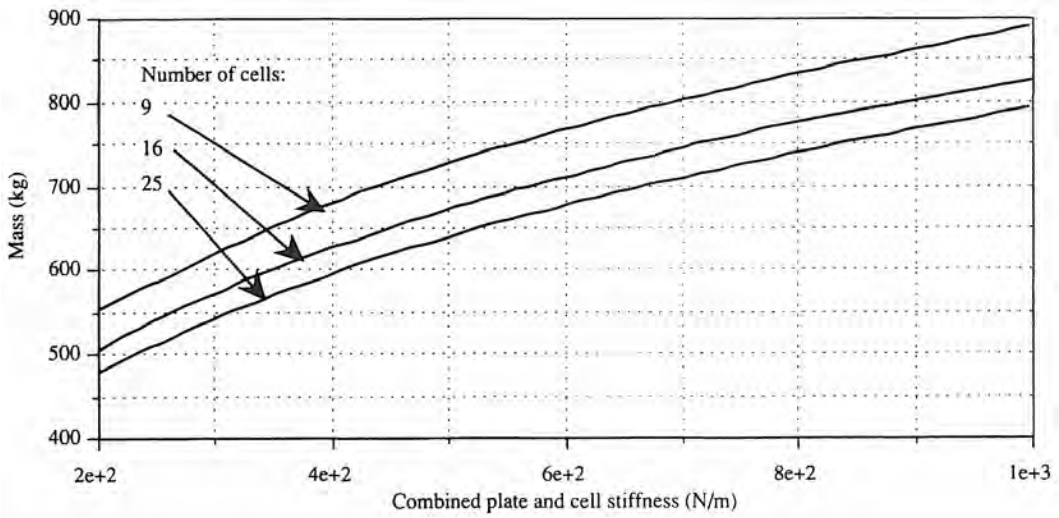


Figure 7.4.16 Effect of number of cells on a cast iron surface plate's mass for a required stiffness (L = 1 m, H = 0.25 m).

This example shows the method and types of calculations that are required for a preliminary optimization of a structural design problem. The next step would be to conduct a finite element analysis on an accurate (e.g., three-point support) model of the surface plate designed using the methods above. The design can then be modified as needed. In some instances for small parts, it may be cheaper to build and test models than to go through an extensive finite element analysis. This is particularly true when one does not have in-house finite element analysis capabilities.

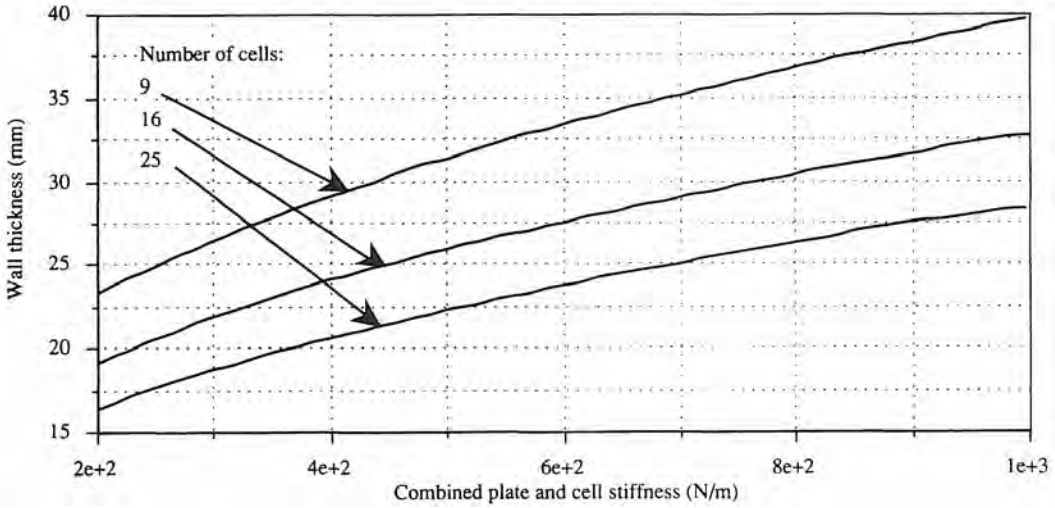


Figure 7.4.17 Effect of number of cells on a cast iron surface plate's wall thickness for a required stiffness (L = 1 m, H = 0.25 m).

7.4.6 Design Case Study: Effects of Mounting Methods on Mirror Distortion[87]

The advantages and disadvantages of kinematic and nonkinematic systems have been discussed in detail. To provide a comparative example, consider how one might mount a square mirror (e.g., a mirror in a camera), as shown in Figure 7.4.18.

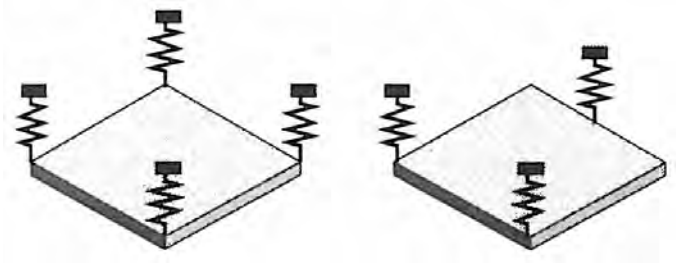


Figure 7.4.18 Two options for mounting a square mirror.

Figure 7.4.19 shows the deflection of a flat glass plate supported at four corners and loaded by its own weight. Note that the deformation is symmetrical about two axes, but it is assumed that the support points are all located in the same plane. Table 7.4.1 shows how the deflection varies with changing plate size. The values can be used to construct a spreadsheet for evaluating different mirror sizes.

[87] The finite element analysis results for this study were generated by Babu Anisetti and Tom Doherty of the Polaroid Corporation.

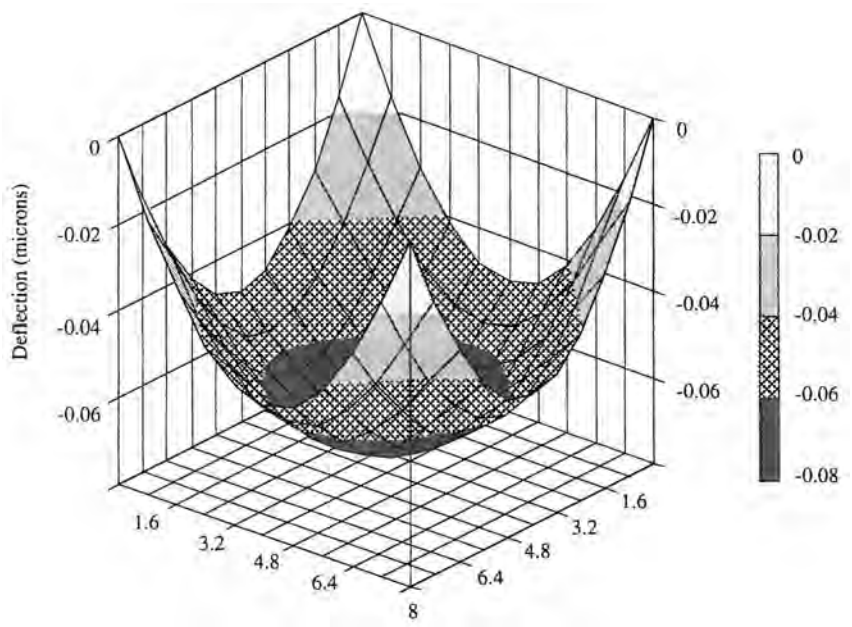


Figure 7.4.19 Deflection of an 8-mm-thick flat glass plate supported at four corners and loaded by its own weight. The plate is 8 cm wide and 8 cm long. (Courtesy of Polaroid Corp.)

L/w	Deflection (microns) for various w/t						
	5	10	15	20	25	30	40
1.00	0.020904	0.073406	0.160782	0.283032	0.440233	0.632358	1.121004
1.25	0.034544	0.125222	0.276123	0.487350	0.758927	1.090854	1.935048
1.50	0.058979	0.219786	0.485521	0.862330	1.344193	1.933143	3.431032
1.75	0.099797	0.378257	0.841731	1.490447	2.324481	3.343910	5.936488
2.00	0.163373	0.628396	1.402664	2.486457	3.879850	5.582920	9.914382

Table 7.4.1 Maximum deflection of a flat glass plate loaded by its own weight and supported at each of its four corners. The plate is 8 cm wide. (Courtesy of Polaroid Corp.)

Figure 7.4.20 shows the deflection of a flat glass plate loaded by its own weight and supported at two corners and in the middle of the opposite edge. This is a kinematic mounting that produces a deformed shape that is symmetric about only one axis. Since there are three support points, they will always be coplanar, and hence the shape you see is the shape you get. Table 7.4.2 shows how the deflection varies with changing plate size.

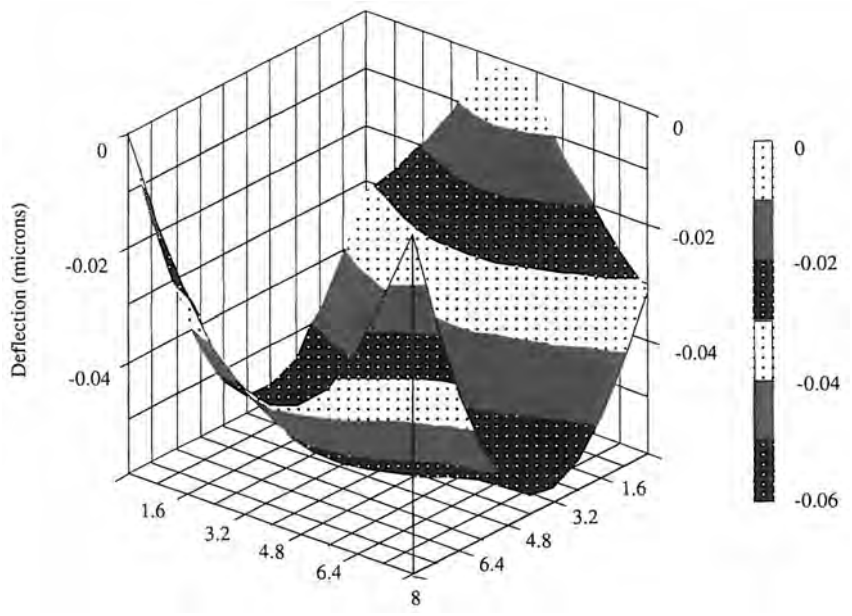


Figure 7.4.20 Deflection of an 8-mm-thick flat glass plate loaded by its own weight and supported at two corners and in the middle of the opposite edge. The plate is 8 cm wide and 8 cm long. (Courtesy of Polaroid Corp.)

L/w	Deflection (microns) for various w/t						
	5	10	15	20	25	30	40
1.00	0.017069	0.057709	0.125171	0.219583	0.340919	0.489204	0.866369
1.25	0.032791	0.117780	0.259105	0.456921	0.711200	1.021994	1.812468
1.50	0.059665	0.222250	0.492862	0.871677	1.358671	1.953870	3.467862
1.75	0.102133	0.388823	0.866267	1.534592	2.393823	3.443986	6.115050
2.00	0.165735	0.639760	1.429309	2.534564	3.955542	5.692394	10.109962

Table 7.4.2 Maximum deflection of a flat glass plate loaded by its own weight and supported at two corners and in the middle of the opposite edge. The plate is 8 cm wide. (Courtesy of Polaroid Corp.)

Figure 7.4.21 shows the deflection of a flat glass plate supported at four corners by springs, each of whose stiffness is equal to a fraction of the central stiffness of the plate if it were simply supported along all its edges and loaded by one of the springs being displaced by 1 μm. This represents a real situation that results from manufacturing not being able to provide a four-point mount that will always be coplanar. Table 7.4.3 shows how the deflection varies with changing plate size.

These were generated using finite element analysis, and experimental results would confirm the numerical results. By studying the figures above, one can come to the following conclusions:

- A symmetrical body where mounts have the same axes of symmetry ideally has symmetric distortions.
- When mountings are overconstrained, significant nonsymmetrical distortions can occur due to forced geometric congruence.
- A kinematic mounting's deterministic nature makes it possible to compute the amount of distortion, which may not have the same axes of symmetry as the body, and then design the system so that the distortion is below a threshold. Hence the system will be less sensitive to manufacturing considerations.

One can always design an error to be below an allowable threshold if the system is deterministic (kinematic). One cannot always rely on manufacturing to hold a tight tolerance. Hence, whenever possible, one should use kinematic design principles in the design and manufacture of precision products, be they machine tools, instruments, or consumer products.

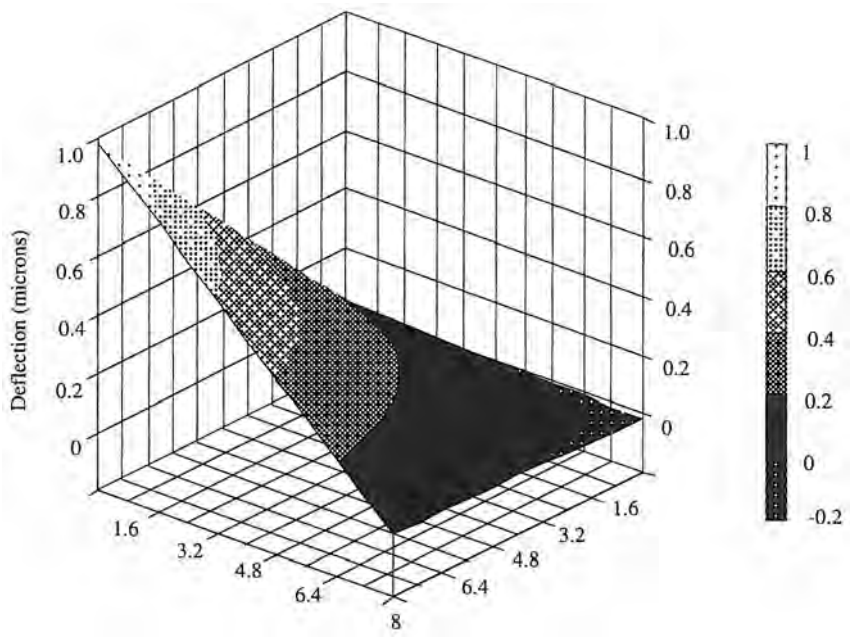


Figure 7.4.21 Deflection of an 8-mm-thick flat glass plate supported at four corners by springs, one of which is displaced by 1 μm. The spring stiffness is equal to the central stiffness of a simply supported plate. The plate is 8 cm wide and 8 cm long. (Courtesy of Polaroid Corp.)

L/w	Deflection (microns) for various w/t			
	1	5	10	20
1.00	0.980491	0.929157	0.890041	0.847293
1.25	0.981177	0.931291	0.892683	0.849869
1.50	0.982548	0.935584	0.899093	0.855294
1.75	0.984047	0.940486	0.904443	0.861898
2.00	0.985444	0.945236	0.910819	0.868756

Table 7.4.3 Maximum deflection of an 8-mm-thick flat glass plate supported at four corners by springs, each whose stiffness is equal to a fraction of the central stiffness of the plate if it was simply supported along all its edges and loaded by one of the springs being displaced by one micron. The plate is 8 cm wide. (Courtesy of Polaroid Corp.)

7.4.7 Structural Design Summary

After a configuration is chosen from among all the conceptual designs, a few simple rules of thumb to follow include:

1. When sizing overall components, keep the proportions of the golden rectangle in mind. These proportions usually yield structurally stiff and aesthetically pleasing designs.
2. Utilize symmetry whenever possible. Asymmetric structures often have internal gradients which are an indicator of potential problems.
3. Remember the principle behind the strength and stiffness of an I-beam, but also remember that shear strains are greatest near the neutral axis.
4. Minimize the structural loop and use closed sections whenever possible, including in the design of the overall design.
5. Start at the tooltip or workpiece with estimates of cutting forces[88] and acceleration requirements and then work backward through the structural system. Use guesstimates for sensor, bearing, and actuator limitations to help size structural components.

[88] See, for example, K. Ing et al., Specific Cutting Force Data for Metal Cutting, Verlag Stahleisen, Dusseldorf, Germany, 1982.

6. Large plate sections should be stiffened with ribs or other means to keep them from vibrating like drumheads. When needed, use active damping systems.
7. Maximize the thermal diffusivity of the machine and minimize heat input.
8. Locate the work volume at the center of mass and in the plane of support. Also try to make the natural frequencies of the various vibration modes (e.g., translational and rotational) close together. These steps will help to minimize cross coupling between modes.[89]

In addition to these general guidelines, one of the best ways to become proficient at structural design is to observe the world around you. For example, continually examine the configurations of existing machinery.[90][91] Developing an aptitude for laying out the general configuration of a machine (i.e., where to place axes with respect to one another, what range of travel is needed, etc.), can also be achieved by dreaming and taking courses in kinematics and dynamics of machinery.[92] It is also very important that the design engineer always question conventional wisdom and try to think of new and different ways of doing things, before a conventional approach is taken. Similar comments pertain to analysis methods. Proficiency at back-of-the-envelope calculations is a must in order to arrive at a suitable point from which finite element analysis can be used effectively. Finite element analysis itself is no panacea; it takes experience to know what types of elements and boundary conditions to use and how many elements are required to obtain a convergent solution.[93] In many cases, back-of-the-envelope calculations can be used to check parts of the finite element model to determine if the mesh is appropriate.

It should be apparent to the design engineer that the structure is not a stand-alone entity. The best machine design engineers integrate the designs of all components so that they work together, and the structure is the means by which all the components are brought together. For example, the job of the structure is not just to support the part and axes but to provide protection from (for) the surrounding environment, and provide regions in which seals and power, sensor, and coolant lines can be easily installed and maintained. More than one brilliant structural design has been thrown out because there was inadequate room for support systems. In many cases, especially for prototypes, it is also advisable to leave extra room in the structure to accommodate next-larger-size motors or bearings that may be used. This makes it easy to modify the prototype or the production machine to suit a customer's special requirements.

7.5 JOINT DESIGN[94]

Joint design is one of the most difficult aspects of machine design because there are so many variables that can affect the performance of a joint. In addition, complex geometries make modeling of joints extremely difficult, and finite element methods often can provide the only models (which still in many instances leave much to be desired). It is costly and time consuming to do finite element analysis of joints; hence, most joints in machine tools are designed using back-of-the-envelope calculations, a conservative nature, and rules of thumb. Once the joint is designed by these methods, then finite element methods can be used to help check the design.

This section will discuss issues in the design of permanent joints between parts and attempt to provide at least some rules of thumb to use when designing joints to achieve desired stiffness for

[89] See, for example, H. Braddick, The Physics of Experimental Method, Chapman and Hall Ltd. London, 1963.

[90] For example, the biannual International Machine Tool Show (IMTS) in Chicago (held the second week of September in even-numbered years) is one of the biggest and best places to see what's available.

[91] See, for example, Huebner's Machine Tool Specs, Huebner Publishing Co., Solon, OH.

[92] References include V. Faires, Kinematics, McGraw-Hill Book Co, New York, 1959; B. Paul, Kinematics and Dynamics of Planer Machinery, Prentice Hall, Englewood Cliffs, NJ, 1979; J. Phillips, Freedom in Machinery, Vol. 1, Introducing Screw Theory, Cambridge University Press, New York, 1984; and R. Paul, Robot Manipulators: Mathematics, Programming, and Control, MIT Press, Cambridge, MA 1981. Many computerized kinematic design packages are currently available.

[93] See, for example, M. Weck and A. Heimann, "Analysis of Variants of Machine Tool Structures by Means of the Finite Element Method," Proc. 18th Int. Mach, Tool Des, and Res, Conf., Sept. 1977, pp. 553–559.

[94] Joints are usually associated with assembly operations, and a good trade magazine to read is thus Assembly Engineering, Hitchcock Publishing Co., 25W550 Geneva Road, Wheaton, IL 60188. Also see Machine Design's annual Fastening, Joining, and Assembly reference issue. Other references include A. Blake, Design of Mechanical Joints, Marcel Dekker, New York, 1985, and the section on joint design in M. Kutz (ed.), Mechanical Engineer's Handbook, John Wiley & Sons, New York, 1986. Also see R. Connolly and R.H. Thornley, "The Significance of Joints on the Overall Deflection of Machine Tool Structures," 6th Int. Mach. Tool Des. and Res. Conf., Sept. 1965, pp. 139–156.

machine tool applications.[95] Joints are defined here as being permanent in the sense that significant effort is needed to take them apart, as opposed to sliding joints with bearing interfaces (Chapters 8 and 9) and coupled joints for periodic mating of parts (Section 7.7). The most common types of permanent joints in machine tools include[96] bolted joints, pinned joints, bonded joints, and interference fit joints. Welded joints were discussed briefly in the preceding section and enough design text is available so they will not be discussed further.

7.5.1 Bolted Joints[97]

Bolts can be used to prevent two parts from separating or sliding relative to one another. For the former, the tensile forces across the joint are transferred through the bolt shaft. For the latter, sliding motion is resisted by frictional forces generated from the normal load on the joint produced by tightening of the bolt and the coefficient of friction between the joint's parts. Because more than one bolt is usually used at a joint, it would be virtually impossible to ensure a tight fit of the bolt shafts in the holes, so it is not even worth trying. Sufficient lateral stiffness is usually provided by bolt preload and joint friction. For better resistance to shock loads, parts can be bolted in place and then holes drilled, reamed, and pinned with hardened steel dowels or roll pins. In situ drilling and reaming of the holes through both parts while they are bolted together maintains hole alignment, so multiple pins can be used.

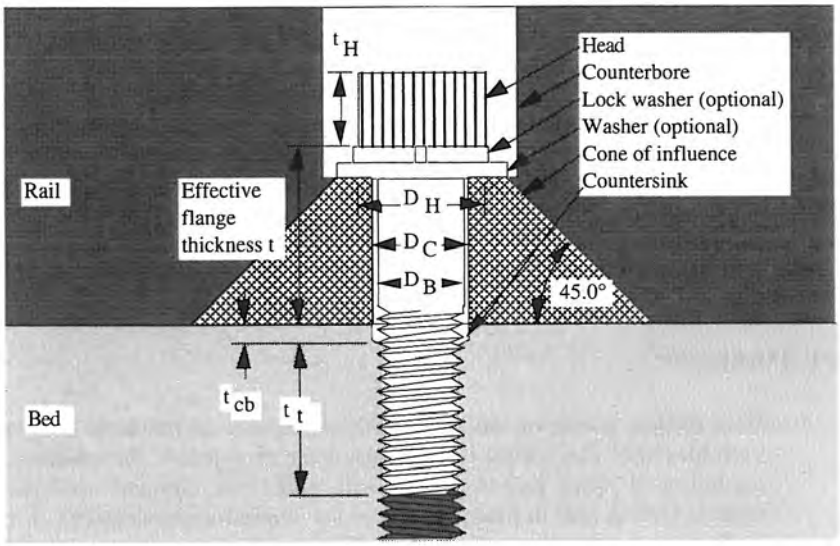


Figure 7.5.1 Typical components of a bolted assembly.

Figure 7.5.1 illustrates the cross section of a typical portion of a bolted joint. A common bolt configuration used to bolt bearing rails for T-slides is shown in Figure 7.5.2. Many rails have a double row of bolts. In general, the cantilevered length should not exceed the bedded length. Ideally, the bedded length should be about 1.5 times the cantilvered length, but sadly this often takes up too much room.

Some issues to consider in the design of bolted joints include:

---

[95] Note that there is a virtual avalanche of literature associated with bolted joints. It is one of the oldest joining methods still commonly used. Most literature focuses on discussion of joint strength and gasket sealing ability. This section will focus on designing for stiffness because that is the primary concern of a machine tool designer.

[96] See, for example, A. Blake, Design of Mechanical Joints, Marcel Deckker, New York, 1986, and M. Kutz (ed.), Mechanical Engineers' Handbook, John Wiley & Sons, New York, 1986.

[97] For a more detailed discussion, see, for example, J. H. Bickford, An Introduction to the Design and Behavior of Bolted Joints, Marcel Deckker, New York, 1981, as well as A. Blake's book.

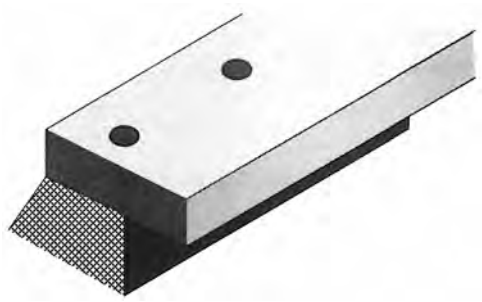


Figure 7.5.2 Counterbored and bolted bearing rail.

- Tensile stiffness
- Compressive stiffness
- Lateral stiffness
- Preload
- Pinning
- Stability
- Bolting hardware

Tensile Stiffness of Bolted Joints

The region of material under the bolt head that effectively acts to resist compression by the bolt is equal to that of a 45° cone.[98] Hence the stiffness of the material under the bolt head is approximately

$$K_{\text{flange comp}} = \frac{1}{\displaystyle\int_0^t \frac{dy}{\pi E_f \left\{ \left(\frac{D_H}{2} + y \right)^2 - \frac{D_C^2}{4} \right\}}} = \frac{\pi E_f D_C}{\log_e \dfrac{(D_C - D_H - 2t)(D_C + D_H)}{(D_C + D_H + 2t)(D_C - D_H)}} \tag{7.5.1}$$

The 1/logarithmic term approaches zero as D_C approaches D_H, hence the value of using a large-diameter head or a washer. Note that as the thickness goes to zero, the stiffness goes to infinity; however, one must bear in mind that the bending stiffness of the flange goes to zero along with the thickness. If the bolt is counterbored and rests on a massive flange (one or more bolt diameters thick), then locally around the bolt the stiffness of the flange will be due primarily to the ring of material in the flange shearing. Beyond two bolt radii, the shear stress is well diffused out into the material around the counterbore. For the bed which is more massive, the threads are assumed to diffuse the stress out so that inclusion of a compression term would only result in a second-order effect.

To find the deflection of the flange of thickness t due to the bolt being in tension, energy methods are used. The shear stress is

$$\tau = \frac{F}{2\pi R t} \tag{7.5.2}$$

The strain energy is given by Equation 2.3.19, where the differential volume element in the flange is $dV = 2\pi R t dR$. The deflection is given by

$$\delta = \frac{\delta U}{\delta F} = \frac{F}{2\pi t G} \int_{R_B}^{2R_B} \frac{dR}{R} = \frac{F \log_e 2}{2\pi t G} \tag{7.5.3}$$

The shear stiffness of the counterbored flange region is thus

$$K_{\text{flange shear}} = \frac{\pi t E_f}{(1 + \eta) \log_e 2} \tag{7.5.4}$$

From geometric compatibility with the bolt head, this includes the effect of shear strain in the bolt head.

[98] M. Spotts, Design of Machine Elements, Prentice-Hall, Englewood Cliffs, NJ, 1985, and J. Shigley and L. Mitchell, Mechanical Engineering Design, McGraw-Hill Book Co., New York, 1983. Both say that the region in compression is a cone with angle of 45° to the centerline of the bolt. Other references say the angle can be as small as 25°. In either case, one finds that the bolt itself is the most compliant spring in the system.

If it is assumed that the effective length of thread engagement is equal to one bolt diameter, then the shear stiffness of the threaded region in the bed (neglecting the countersunk region) is

$$K_{\text{bed shear}} = \frac{\pi D_B E_t}{(1+\eta)\log_e 2} \tag{7.5.5}$$

Assuming that the bolt is threaded in one bolt diameter and the threads start a countersink (or counterbore) distance t_{cb} below the surface to avoid forming a crater lip upon tightening, then the bolt stiffness is approximately (effective length = $D_B/2 + t + t_{cb}$)

$$K_{\text{Bolt}} = \frac{\pi E_B D_B^2}{4(D_B/2 + t + t_{cb})} \tag{7.5.6}$$

For steel bolts in cast iron and steel, the stiffness of the flange structure (when the thickness is greater than at least one-half bolt diameter) is always stiffer than the bolt. A rule of thumb is to design to the flange to be at least one to two bolt diameters thick. It is also desirable to make the stress zone cones beneath the bolt heads overlap. This will help minimize straightness errors caused by bolt tightening but unfortunately, can lead to the use of a plethora of bolts.

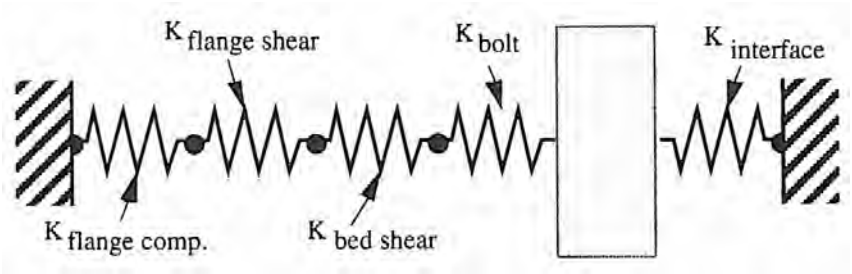


Figure 7.5.3 Spring model of a bolted joint.

As long as the applied load is less then the preload, then the total joint stiffness can be represented by the model shown in Figure 7.5.3. As long as the applied load does not exceed the preload of the joint, a simple force balance[99] shows that the effective stiffness of the bolted joint will be

$$K = K_{\text{interface}} + \frac{1}{\dfrac{1}{K_{\text{flange comp.}}} + \dfrac{1}{K_{\text{flange shear}}} + \dfrac{1}{K_{\text{bed shear}}} + \dfrac{1}{K_{\text{bolt}}}} \tag{7.5.7}$$

The joint stiffness decreases with the difference in preload and applied load (as the preload is unloaded) as discussed later. For a balanced design, the joint interface stiffness should equal at least twice the equivalent stiffness of the bolt system. Thus when fully loaded in tension, the preload will be sufficient to maintain the desired joint stiffness. When the flange is not really massive, one can approximate the stiffness of the flange using beam or plate theory (do not forget shear deformations in thick stubby sections) and then use the same type of solution method.[100] In order to determine which are the dominant (least stiff) areas, imagine that the components are made of soft rubber and then apply loads and see how they deform.

Not all joints need this type of analysis, or in some cases the modeling becomes so complex that an experienced bolted joint design engineer is consulted to design the joint. The best way for a novice design engineer to gain experience about bolted joints is to do these types of calculations, observe all the machinery he/she can, and exploit every chance possible to fix or assemble machinery (e.g., his/her car).

Example

Consider a bearing rail shown in Figure 7.5.4. How does one decide how large and how many bolts to use? Ideally, the bearing rail would behave as if it were built into a wall, which would require an infinite number of bolts (too many). The more bolts that are used, the more costly the

[99] See Equations 8.2.1 and 8.2.2.

[100] A. Blake, Design of Mechanical Joints, Marcel Deckker, New York, 1986, provides a detailed discussion on modeling of flanged joints.

manufacturing operation, and the weaker the bearing rail gets because it becomes perforated with holes. In order to maintain a balanced design, the stiffness of the bearing rail and the bolted joint system should be the same. With respect to the number of bolts used, in order for the following analysis to be reasonably accurate, it is assumed that the bolt spacing L/γ is on the order of $^1/_2$ to $^1/_3$ of the bearing rail width. With this rule of thumb, there will usually end up being at least two bolts near each bearing pad that rides on the rail. The model of the cross section of the bearing rail is shown in Figure 7.5.4. For the purposes of a back-of-the-envelope determination of what size bolts to use, the following assumptions are made:

- The rail is not as stiff as if it were built into a wall, but on the other hand, it is stiffer than a simply supported section if the bolts behaved like a knife edge pushing down along the entire length of the rail. Thus a simply supported model should be conservative.
- The bolts do not provide knife-edge support; thus the stiffness will be less than modeled. This should be offset by the assumption above.
- When a bearing pad exerts a force on the bearing rail, neighboring sections provide support also, thus lending credence to the knife-edge assumption.
- The surface to which the rail is bolted supports the rail only at the rear point; however, because it actually provides some support due to the bowing of the rail, it will be assumed that the rear support point is infinitely stiff and the surface the rail is bolted to does not provide resistance to bowing.
- Bending and shear deformations must be considered.
- Previous analyses for the stiffness of bolts and flanges are valid.
- The bolts are counterbored, and bolt sizes are to be found for the flange thicknesses equal to 1.0, 1.5, and 2 bolt diameters, respectively. Other geometry assumptions are $D_C = D_B$ and $D_H = 1.5D_B$, $t_{cb} = D_B/4$, and the effective width of the rail between bolt centers is ℓ.
- The contact stiffness of the rail/bed interface is generally much greater than any of these terms; hence it is not included here but just as well should be when using a spreadsheet program.
- The rail and bolt are steel, and the base is cast iron, so $E_f = E$, $E_B = E$, and $E_t = E/2$.

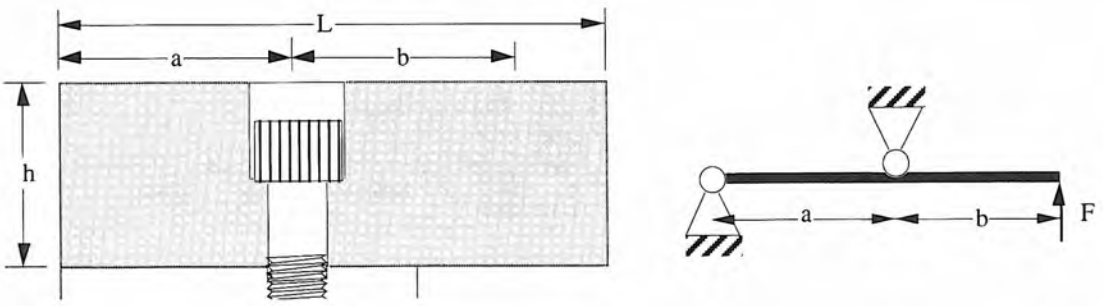


Figure 7.5.4 Model of a bolted rail.

Equations 7.5.1 - 7.5.7 are first evaluated. The results are shown in Figure 7.5.5. The next step is to find the equivalent stiffness at the end of the rail. The system behaves like a spring attached to a fulcrum so the equivalent stiffness is just

$$K_{\text{eq bolt system}} = \frac{Ka^2}{(a+b)^2} \tag{7.5.8}$$

For the remainder of this example, it is assumed that

$$K_{\text{eq bolt system}} = CD_B \tag{7.5.9}$$

The loading function for the bearing rail is as shown in Figure 7.5.4 is

$$q = F < x >_{-1} - F\left(\frac{a+b}{a}\right) < x - b >_{-1} \tag{7.5.10}$$

The shear is

$$V = -F < x >^0 + F\left(\frac{a+b}{a}\right) < x-b >^0 + C_1 \tag{7.5.11}$$

The shear is zero at x = 0, hence C_1 is zero. The moment is

$$M = F < x >^1 - F\left(\frac{a+b}{a}\right) < x-b >^1 + C_2 \tag{7.5.12}$$

The moment is zero at x = 0, hence C_2 is zero. The slope due to bending is thus

$$\alpha = \frac{F}{EI}\left\{\frac{x^2}{2} - \frac{a+b}{a}\frac{< x-b >^2}{2} + C_3\right\} \tag{7.5.13}$$

The deflection due to bending is thus

$$\delta = \frac{f}{EI}\left\{\frac{x^2}{6} - \frac{a+b}{a}\frac{< x-b >^3}{6} + C_3 x + C_4\right\} \tag{7.5.14}$$

The deflection is zero at b and a + b; the bending deflection is thus found to be

$$\delta_{bend} = F\frac{(a+b)b^2}{3EI} \tag{7.5.15}$$

Energy methods are used to find the shear deflection. The shear stress for a rectangular beam is given by Equation 2.3.17 (τ = V[(h/2)^2- y^2]/2I). The strain energy is

$$\begin{aligned} U &= \int \frac{\tau^2}{2G}\,dV = \ell \int \frac{V^2}{8GI^2}\left[\left(\frac{h}{2}\right)^2 - y^2\right]^2 dydx \\ &= \frac{1}{8GI^2}\int_0^{a+b} V^2 \int_{-h/2}^{h/2}\left[\left(\frac{h}{2}\right)^2 - y^2\right]^2 dydx = \frac{\ell h^5}{240GI^2}\int_0^{a+b} V^2\,dx \end{aligned} \tag{7.5.16}$$

Because of the singularity functions, the integral must be done in parts from 0 to b and from b to (a + b). Before integrating, recall that the deflection is given by δ_{shear} = dU/dF and the shear was a function of F; hence

$$\begin{aligned} \delta_{shear} &= \frac{\ell h^5}{120GI^2}\int_0^{a+b} V\frac{dV}{dF}\,dx \\ &= \frac{F\ell h^5}{120GI^2}\left[\int_0^b (-1)(-1)\,dx + \int_b^{a+b}\left(-1+\frac{a+b}{a}\right)\left(-1+\frac{a+b}{a}\right)dx\right] \\ &= \frac{F\ell h^5(a+b)}{120GI^2a} \end{aligned} \tag{7.5.17}$$

Recall from Equation 7.3.4 that G=0.5E/(1 + η), and with the beam's moment of inertia I = l h^3/12, the shear deflection is

$$\delta_{shear} = \frac{Fbh^2(a+b)(1+\eta)}{5aEI} \tag{7.5.18}$$

The total deflection is thus

$$\delta_{beam} = \frac{Fb}{EI}\left[b(a+b)over3 + \frac{h^2(I+\eta)(a+b)}{5a}\right] \tag{7.5.19}$$

Substituting I = lh^3/12 and l = L/γ D_B, the stiffness of the beam is found to be

$$K_{beam} = \frac{5Eah^3[L/\gamma - D_B]h^3}{4b(a+b)[5ab + 3h^2(1+\eta)]} \tag{7.5.20}$$

Assume the following notation:

$$A = \frac{5Eah^3}{4b\,(a+b)(5ab+3h^2(1+\eta))}\quad B = \frac{L}{\gamma} \tag{7.5.21}$$

The inverse sum of Equations 7.5.20 and 7.5.9 must equal a portion χ of $K_{desired}$ (remember Equation 7.5.7). The factor N indicates how many bolts there are per section of rail that the desired stiffness is required. This results in a quadratic in D_B:

$$\frac{K_{desired}}{\chi} = \frac{1}{\frac{1}{NK_{beam}} + \frac{1}{K_{eq.boltsystem}}} \tag{7.5.22a}$$

$$D_B^2 \frac{\chi^{NCA}}{K_{desired}} + D_B \left[\frac{\chi^{NCAB}}{K_{desired}} - NA + C \right] + NAB = 0 \tag{7.5.22b}$$

where the coefficients of the powers 2, 1, and 0 of D_B are a_{term}, b_{term}, and c_{term}, respectively. The bolt diameter is then found using the quadratic formula. For a balanced design, the interface stiffness should be equal to the bolt and rail stiffness. By making the interface stiffness much larger, the bolt diameter will decrease. However, if the interface stiffness is too high, the displacement caused by the preload will be so small that manufacturing defects may cause loss of preload around a bolt. This will result in a soft spot on the rail. To ensure stiffness is maintained in the presence of maximum tension, the bolts should be torqued to make the interface stiffness equal to $2K_{desired}(1 - 1/\chi)$. Typically, χ should be 2 but this may yield unrealistic bolt diameters, so a value as high as 4 is not unreasonable. If unreasonable bolt diameters are still obtained, perhaps a longer section of rail should be considered (e.g., N = 2).

	Meters and Newtons			Inches and pounds		
$K_{desired}$	5.26E+08	5.26E+08	5.26E+08	3.00E+06	3.00E+06	3.00E+06
χ	3.00	3.00	3.00	3.00	3.00	3.00
E	2.07E+11	2.07E+11	2.07E+11	3.00E+07	3.00E+07	3.00E+07
N	1	1	1	1	1	1
Rail width/Bolt spacing	1.0	1.0	1.0	1.0	1.0	1.0
Flange thickness	t = 2Db	t = 1.5Db	t = Db	t = 2Db	t = 1.5Db	t = Db
K flange comp/E*D_B	2.530	2.714	3.075	2.530	2.714	3.075
K flange shear/E*D_B	6.973	5.230	3.486	6.973	5.230	3.486
K thread shear/E*D_B	3.486	3.486	3.486	3.486	3.486	3.486
K bolt/E*D_B	0.286	0.349	0.449	0.286	0.349	0.449
C*(a/(a+b))^2	1.29E+10	1.51E+10	1.79E+10	1.87E+06	2.19E+06	2.59E+06
Rail width L	0.150	0.150	0.150	5.91	5.91	5.91
a	0.065	0.065	0.065	2.56	2.56	2.56
b	0.060	0.060	0.060	2.36	2.36	2.36
h	0.050	0.050	0.050	1.97	1.97	1.97
Bolt spacing	0.150	0.150	0.150	5.905	5.905	5.905
D_B(20% thread allowance)	0.019	0.016	0.014	0.741	0.634	0.533
Max. allowable bolt stress	6.90E+08	6.90E+08	6.90E+08	1.00E+05	1.00E+05	1.00E+05
Max. allowable bolt force	9.91E+04	7.28E+04	5.15E+04	2.23E+04	1.64E+04	1.16E+04
3Dbolt max. contact pressure	7.67E+07	7.67E+07	7.67E+07	1.11E+04	1.11E+04	1.11E+04

Figure 7.5.5 Spreadsheet results for bolt sizing example.

Typically, L and h are set by a standard rail size. As mentioned before, γ is typically 2 or 3. The bolt diameter D_B is shown for three different flange thicknesses in the spreadsheet output of Figure 7.5.5. An experienced machine design engineer will tell you that these bolt diameters look about right; thus it is possible for a novice to make reasonable back-of-the-envelope calculations to determine bolt sizes. One still needs to make sure that the bolts can be tightened by an amount that will create the necessary joint preload and associated lateral stiffness without causing the rail to deform. Lateral deformations due to the Poisson effect should also be prevented, if the rail is also a guiderail, by maximizing the ratio of rail width to bolt diameter, minimizing the bolt spacing, and minimizing the bolt torque. Also note that if $K_{desired}$ was very high, more than the system could reasonably achieve, an unreal bolt diameter would be obtained. This is why a spreadsheet is nice because it allows you to play with the numbers.

What about the rest of the system? Figure 7.5.6 shows bearing rails that can be subject to balanced or unbalanced preload forces. Which design is best? Would making the bearing rails integral with the casting make for a better design? What must be done to accommodate integral bearing rails? How would the model above have to modified if the keeper rail was used for a hydrostatic bearing application? (See the end of Section 8.7.)

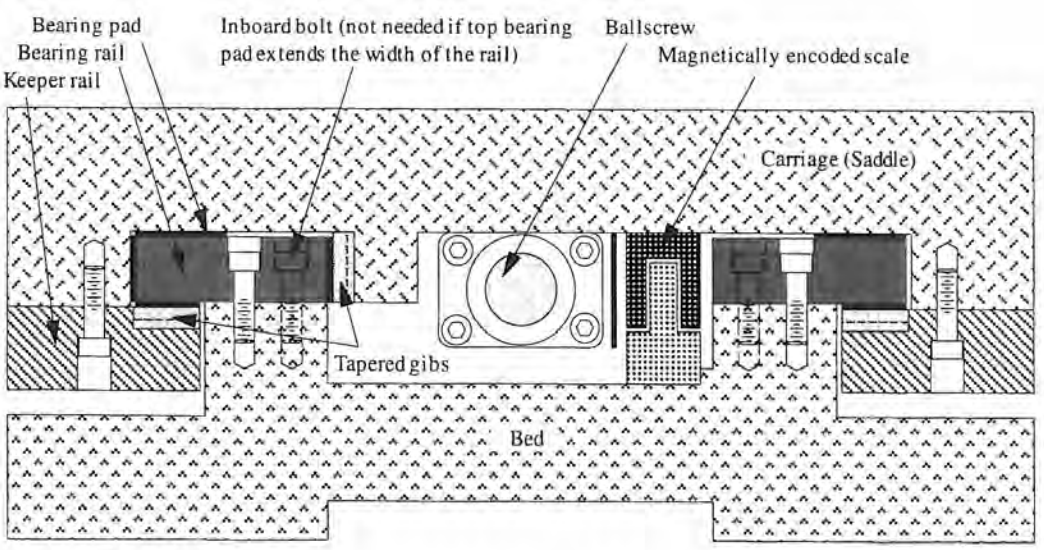


Figure 7.5.6 T-slide bearing configuration.

Compressive Stiffness of Bolted Joints

Ideally, in compression a bolted joint would behave as if the assembly were made from a solid piece of material and calculation of the stiffness would be straightforward; however, due to surface finish variability and imperfections, this is not the case. The finer the surface finish and the tighter the preload, the fewer and smaller the gaps between surface asperities on the surfaces. Hence where compressive stiffness is critical, surfaces should be ground, scraped, or lapped. Vibration annealing can help to seat the asperities. Grouting the joint with an adhesive can fill both large and micro voids and greatly increase joint stiffness as well as increase the damping. The joint may even be designed so it has only three height adjustable contact points (e.g., tapered gibs, tapered horizontal screw threads, etc.). After the entire structure is aligned, the joints can then be grouted with an appropriate material (e.g., cement or epoxy).

In order to estimate joint stiffness, one can consult experimental data as shown in Figures 7.5.7 through 7.5.9. Figure 7.5.7 shows joint stiffness as function of the contact pressure between ground steel specimens[101] up to relatively high contact pressures. Figure 7.5.8 shows the effect of contact pressure on the damping properties of the joint. Figure 7.5.9 shows a joint's compressive stiffness as a function of contact pressure[102] for lower contact pressures. Often lower contact pressures are typically found in components that must be assembled with minimal preload to prevent deformation. One reason to space bolts not too far apart from each other or the boundaries of the contact surface (e.g., not farther than the rail thickness) is so that the joint pressure can be estimated as the total force provided by all bolts divided by the total joint area. The joint stiffness would thus be the product of the stiffness per unit area value obtained from either Figure 7.5.7 or 7.5.9 and the joint area.

Lateral Stiffness of Bolted Joints

The lateral stiffness of a bolted joint is also a function of the surface finish and the preload on the joint. The higher the preload, the greater the lateral force that can be resisted by the coefficient of friction. Lateral stiffness seems to be a function of how the surface asperites interlock when the joint is preloaded, and how they subsequently bend and shear when lateral forces are applied to the joint. Attempts at modeling the asperites as a random distribution of peaks and valleys have yet to prove satisfactory, and empirical data is best used, such as shown in Figures 7.5.7 or 7.5.9. If very high lateral stiffness is required from a bolted joint, it may be appropriate to key or pin the joint after the structure has been aligned. For the former, a loose-fitting key is grouted or epoxied in place as

[101] From M. Yoshimura, "Computer Aided Design Improvement of Machine Tool Structure Incorporating Joint Dynamics Data," Ann. CIRP, Vol. 28, 1979, pp. 241–246. The test specimens were made of 0.55% carbon steel and the surfaces were ground and coated with a light machine oil.

[102] From M. Dolbey and R. Bell, "The Contact Stiffness of Joints at Low Apparent Interface Pressures," Ann. CIRP, Vol. 19, pp. 67–79.

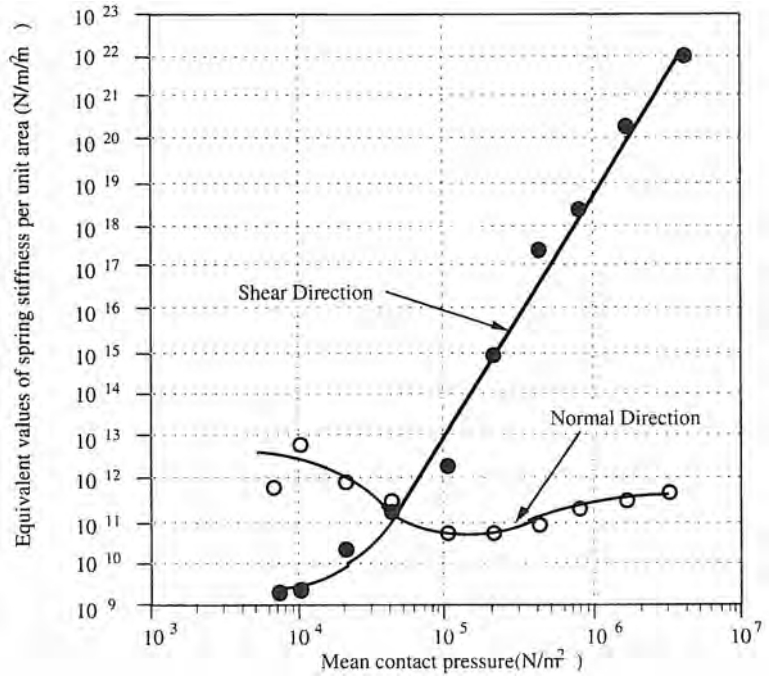


Figure 7.5.7 Joint stiffness of 0.55%C ground steel parts. (After Yoshimura.)

a final manufacturing procedure. For the latter, the dowel pin holes are drilled and reamed into the bed using the finally positioned rail as a template.

Preload[103]

When a torque Γ is applied to a bolt with lead l (1/l threads per unit length), the axial force generated is

$$F = \frac{2\pi\Gamma e}{\ell} \tag{7.5.23}$$

For lubricated threads, the efficiency e can vary from 0.2 to 0.9, depending on the surface finish and accuracy of thread mating.[104] In addition to the efficiency of the threads, one must consider the friction μ under the bolt head. Hex socket cap screws are most often used in machine tools, and the bolt head diameter is typically 1.5 times the bolt diameter. Equating the work in to the work out, the axial bolt force as a function of torque is

$$F = \frac{4\pi\Gamma}{\frac{2\ell}{e} + 3\pi D_B \mu} \tag{7.5.24}$$

One should assume that the coefficient of friction will be 0.3 normally, and 0.1 if the bolt is lubricated and vibration stress relieved during final tightening.

The tensile stress in the bolt shaft is $\sigma = 4F/\pi D_B^2$, and the shear stress in the threads is $\tau = F/\pi D_B L$ where L is the thread engagement length. Whether the bolt or threads fail depends on the relative strength of the bolt and thread material, and the thread engagement length L. Note that if L is greater than one or two thread diameters, then due to imperfections in thread geometry all the threads cannot engage anyway. Thus it does not make sense to specify tapping holes for more than two bolt diameters of usable thread. If a bolt is to be removed from the part periodically, and the bolt must carry a high load and/or preload, then one should consider specifying hardened-steel coiled-thread inserts. In addition, female threads should never be specified in a material that is harder than about RC30, because the sharp thread corners are prime stress concentrators, which can lead to failure.

When many bolts are used on a joint, uniform bolt preload is necessary to minimize distortion of the part and maintain consistency of performance from machine to machine. Special bolts have

[103] See J. Bickford, "Preload: A Partially Solved Mystery," Mach. Des., May 21, 1987.

[104] See Section 10.8.3.1 for a detailed discussion of the calculation of the efficiency based on thread geometry and the coefficient of friction.

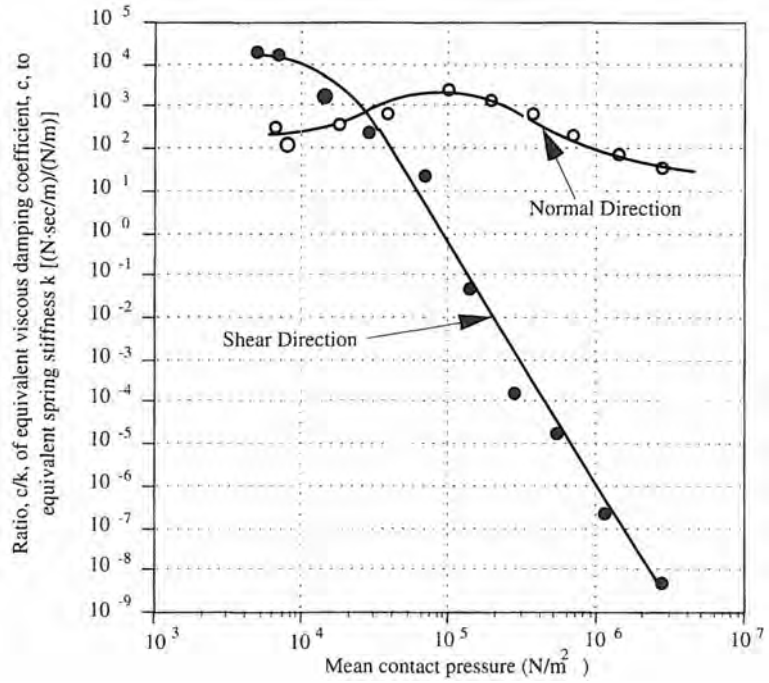


Figure 7.5.8 Joint damping of 0.55%C ground steel parts. (After Yoshimura.)

been developed that have an indicator on the head that lets the bolt tightener know when the proper bolt tension has been achieved. For inexpensive common bolts, it is often desirable to minimize the coefficient of friction and its variability. This can be done in the following manner:

- The tapped threads should start 1/4 bolt diameter below the surface to avoid raising the surface and causing a crater lip to form should the bolts be overtightened.
- Inspect bolts to make sure that the thread forms have a good surface finish and are well formed. If necessary, buff the threads. Avoid the use of cheap bargain basement bolts with poorly formed threads.
- Thoroughly clean and degrease the male and female threads, counterbore seat, washers, and the underside of the bolt head. Thoroughly lubricate the parts and make sure that the bolt threads easily into the hole.
- Make sure that the surface finish of the counterbore seat and the underside of the bolt head are smooth and flat. For critical assemblies where all bolt preload forces must be the same,

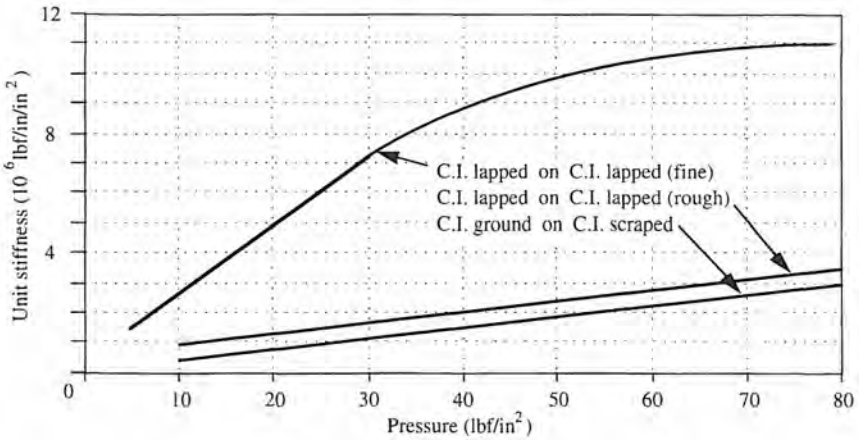


Figure 7.5.9 Compressive joint stiffness at lower contact pressures (After Dolby and Bell).

place a superfinished steel washer in the counterbore (finished side up) and a lubricated superfinished brass washer (finished side facing the superfinished steel washer) on top of it, then insert the bolt. The washers will act as a thrust bearing to minimize friction at the head.
- For permanent joints or joints subject to vibration, use a thread lubricant that hardens after several hours. Under shock and vibration loads, bolt preload may be reduced by a factor of 2 after only a few hours of service if a locking mechanism is not used.
- Tighten bolts in an incremental sequence. Vibration anneal the assembly, and then retighten the bolts. Repeat this process until the bolts do not require further retightening.
- If possible, pin the assembly.
- Perform one final retightening of the bolts while the structure is being vibration stress relieved. The dither action provided by the vibration anneal process can help to overcome stick-slip friction in the bolt threads and head-to-counterbore interface.

The preload force, area, and surface finish determine the interface pressure between the parts of a joint, which affects the stiffness across the joint as discussed above. High bolt preloads are often necessary for sealing pressurized joints with gaskets, or where high alternating stresses must be made a small proportion of the total stress state in the bolt. For machine tool joints where the bolts are not subject to high alternating stresses, the preload should be low and more bolts used to achieve the total desired force and joint interface pressure. This will also help to distribute the force over the joint, thereby minimizing deformation of the joint. To minimize warping, compression, or Poisson expansion of the parts that make up the joint, a rule of thumb is to make the bolt spacing on the order of the thickness of the part or flange.

Remember, the tighter the bolt, the larger the local deformation (dishing) around the bolt, which can deform critical components such as bearing rails; however, the looser the bolt, the lower the stiffness of the joint. Note that bolt diameter can be increased to keep stresses low ($<$10-25% of yield) to help maintain long-term dimensional stability of the system,[105] and larger bolt diameters increase the stiffness of the bolts. The following guidelines can be used to determine the required preload and bolt size for a joint.

- From experimental data (e.g., Figures 7.5.7-7.5.9) determine the required joint interface pressure to achieve the required stiffness, and the required force on the joint area to achieve the interface pressure. The stiffness of the joint should be greater than the stiffness of the parts that make up the joint, as discussed in the context of Equation 7.5.22. Assume that the bolt force acts over an area of four to five bolt diameters (the area of the cone of influence).
- Determine the maximum tensile force exerted on the joint. This force would lower the preload on the joint and thus decrease joint stiffness. The minimum total preload force that must be exerted by the bolts will be the larger of the sum of the maximum tensile force and the required joint preload force or four times[106] the maximum tensile force.
- Make sure that the product of the minimum force and the coefficient of friction for the joint are at least a factor of 5-10 greater than the maximum expected shear force on the joint. If the joint is to also be pinned, then this may not be necessary.
- Space the bolts according to one of the rules of thumb given above (i.e., bolt spacing equal to the part thickness or a portion of the width of the bearing rail).
- Size the bolts so that the stresses in the bolts are below 25% of yield, and compare this value to the bolt size found from a minimum stiffness criteria (as shown in the example). Check the bending, compression, and Poisson expansion of the rail (see next section) to make sure that they are within acceptable limits.

As an example, consider the bolted rail designed in Figure 7.5.5. Figure 7.5.10 summarizes the selection of the torque levels to be used. As shown, the primary design consideration is the

[105] To maximize fatigue strength and consistence of the preload force applied, it has been proposed that bolts should be tightened to the point where they begin to yield. For joints that seal or only require micron accuracy, this method is useful; however, data on long-term dimensional stability of yield-tightened joints is not yet available. For more information on the yield tightening process, see J. Monaghan and B. Duff, "The Effects of External Loading on a Yield Tightened Joint," Int. J. Mach. Tools Manuf., Vol. 27, No. 4, 1987.

[106] This will help to maximize the fatigue life of the bolt. Note that one should check company policy regarding design of fatigue-loaded bolted joints, as a value different than 4 may be used.

	Meters and Newtons		Inches and pounds	
D_B	0.016	0.016	0.625	0.625
Area	0.0002	0.0002	0.307	0.307
$4D_B$ area	0.0030	0.0030	4.602	4.602
Lead	0.002	0.002	0.08	0.08
Thread friction	0.1	0.3	0.1	0.3
Efficiency	0.28	0.11	0.29	0.12
$K_{desired}$	5.26E+08	5.26E+08	3.00E+06	3.00E+06
Joint pressure (from Figure 7.5.7)	4.00E+04	4.00E+04	6	6
Joint pressure (from Figure 7.5.9)	1.38E+05	1.38E+05	20	20
Force required (Ground on scraped cast iron)	416	416	92	92
4 × Alternating force on carriage	40000	40000	4494	4494
Number of bolts under carriage	8	8	8	8
Total force per bolt	2916	2916	654	654
Torque required	6.9	18.8	60.6	164.9
Shear stress	8.52E+06	2.33E+07	1263	3440
Tensile stress	1.45E+07	1.45E+07	2131	2131
Mises equivalent stress	2.07E+07	4.29E+07	3054	6328

Figure 7.5.10 Bolt loads for the example of Figure 7.5.5 (χ = 3.00).

alternating force level, as only a very low bolt torque is required to attain the desired stiffness. Another reason for using a minimum bolt torque would be to ensure that the bolts do not loosen during the life of the machine.

Rail Deformation Due to Preload

Even when a "perfect" rail is bolted to a "perfect" bed, significant deformations can result in the form of bending, compression, and Poisson expansion. These deformations can be seen in the way that light is reflected off a precision-bolted bearing rail. Unfortunately, it is not always possible to finish the rails after they are bolted in place, so it is necessary to estimate the amount of deformation so that it can be minimized beforehand. Also, if the rails were finished while bolted and then transferred to another machine, it would be desirable to be able to determine the amount of deformation that could result from a variation in the bolt force.

Two cases will be considered here. The first case is where the counterbore diameters are small with respect to the bearing rail width b_{rail}, and thus the bearing rail can be modeled as a beam on an elastic foundation, as shown in Figure 7.5.11. The second case is where the bolt holes take up a significant part of the beam cross section and thus the other model shown in Figure 7.5.11 is used. Both cases assume that the beams (rails) are preloaded, so in effect the foundation (bed) can exert downward as well as upward forces on the beams. For the first case, the relative lateral deflection is found from Roark:[107]

$$\delta = y_A \left\{ \text{Cosh}\frac{\beta \ell}{2} \cos \frac{\beta \ell}{2} - 1 \right\} + \frac{M_A \, \text{Sinh}\frac{\beta \ell}{2} \sin \frac{\beta \ell}{2}}{2EI\beta^2} \tag{7.5.25}$$

where

$$y_A = \frac{-F}{4EI\beta^3} \left[\frac{C_2 C_{a1} + C_4 C_{a3}}{C_{14}} \right] \qquad M_A = \frac{F}{2\beta} \left[\frac{C_2 C_{a3} + C_4 C_{a1}}{C_{14}} \right] \tag{7.5.26}$$

and

$$\begin{aligned} \beta &= \left(\frac{b_{bed} k}{4EI} \right)^{1/4} & I &= \frac{b_{rail} \, h_{rail}^3}{12} \\ C_2 &= \cosh\beta e \, \sinh\beta e + \sinh\beta e \, \cosh\beta e & C_4 &= \cosh\beta e \, \sin\beta e - \sinh\beta e \, \cos\beta e \\ C_{a1} &= \cosh\frac{\beta e}{2} \cos\frac{\beta e}{2} & C_{a3} &= \sinh\frac{\beta e}{2} \sin\frac{\beta e}{2} \\ C_{14} &= \sinh^2\beta e + \sin^2\beta e \end{aligned} \tag{7.5.27}$$

In order to estimate the foundation modulus k (stiffness per unit area, hence it has units of N/m^3), consider the following. The bolt pulls down on the rail and up on the bed. If the bed's

[107] R. J. Roark and W. C. Young, Formulas for Stress and Strain, 5th edition, McGraw-Hill Book Co., New York, 1975, p. 134.

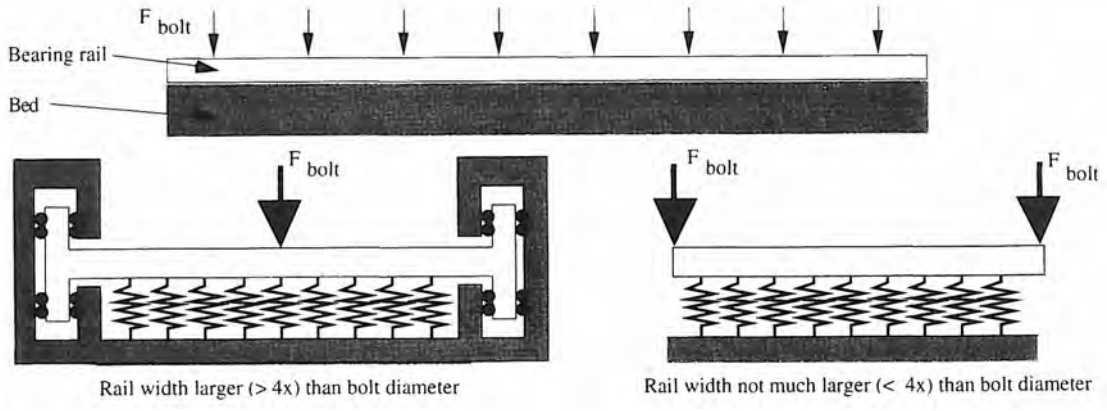


Figure 7.5.11 Model of rail deformation due to bolt preload. When the counterbore diameter is much less than the rail width, the model of a guided beam on an elastic foundation of modulus k is used. When the counterbore diameter is a significant portion of rail width, the simply supported beam model on the right is used.

moment of inertia is much greater (e.g., >10) than that of the rail, then bending effects of the bed can be ignored. For a precision machine this should be an accurate assumption. The deeper the section of the bed, the stiffer it is in bending, but the more compliant it is in compression. If the bolt extended to the bottom of the bed, then it would be easier to make an assessment of k. Assume that the depth of the section is that which makes the bed moment of inertia 10 times greater than that of the rail. Assuming that the bed width is b_{bed} and the rail width is b_{rail}, the effective depth h_{bed} of the bed is

$$h_{bed} = h_{rail}\left(\frac{10\,b_{rail}}{b_{bed}}\right)^{1/3} \tag{7.5.28}$$

The foundation modulus k for the bed is thus

$$k\frac{E}{h_{rail}\left(\frac{10\,b_{rail}}{b_{bed}}\right)^{1/3}} \tag{7.5.29}$$

In many instances where a "standard" type of bed is used, simple experiments could be performed to evaluate k. This value could then be used with the rail deflection equations to scale new designs. For the second case, modeled in Figure 7.5.11, Timoshenko[108] gives the relative deflection between the ends and the middle to be

$$\delta = \frac{2F\beta\left[\cosh\beta e + \cos\beta e - 2\cosh\frac{\beta e}{2}\cos\frac{\beta e}{2}\right]}{b_{bed}k(\sinh\beta e + \sin\beta e)} \tag{7.5.30}$$

This equation would be used, for example, when bolting down modular recirculating rolling element linear bearing rails (e.g., linear guides). Note that these solutions do not consider shear deformations. An analysis that took into account shear deformations would be too complex for general application; however, a conservative engineering estimate is to assume that the shear deformations are on the order of the bending deformations, because in general this type of rail is fairly stocky and the bolt spacing is usually equal to an amount less than the rail width. For the previous example, the foundation modulus (k) is 2.7529E + 11N/m^3, and 2 × Equation 7.5.25 (to account for shear deformations) is 0.152 μm (6 μin.). For other geometries and loads, deflections obtained from the theory presented above correlates reasonably well with experimental data provided by Levina.[109]

[108] S. Timoshenko Strength of Materials, Part II, 3rd ed., Robert E. Krieger Publishing Co., Melbourne, FL, p. 17. Note that "Timoshenko's k" is the stiffness per unit width whereas Roark's k (and k given by Equation 7.5.29) is the stiffness per unit area, and hence Equation 7.5.30 has the term "$b_{bed}k$" where k is defined by Equation 7.5.29.

[109] See Z. M. Levina, "Research on the Static Stiffness of Joints in Machine Tools", Proc. of the 8th Int. Mach. Tool Des. and Res. Conf., Sept. 1967, pp. 737–758.

Even if the subgrade were infinitely rigid, the bolts would still compress the metal in the rails. The result would be low regions near the bolt heads and high regions between them. This is referred to here as compressive deformation of the rail. To estimate the compressive deformation, use the area of the rail encompassed by three to five bolt diameters.[110] In addition, recall the Poisson effect given by Equations 7.3.3. Hence the compressive deformation is accompanied by a lateral expansion that is on the order of 0.3 times the compressive deformation (for most metals), depending on the rail geometry. For large rails such as shown in Figure 7.5.1, the Poisson expansion is usually negligible because it is diffused into the surrounding metal. For thin rails, such as those used by many types of modular linear bearings, the effect can be more prominent, which is why manufacturers' bolt-tightening recommendations should be carefully followed.

Since bolt torques and thread efficiencies are never exactly the same, varying vertical and horizontal straightness errors in the rail will often be present with a period equal to that of the bolt spacing. Bolt preload, thread friction, and tightening methods should all be chosen with care when designing and manufacturing a precision machine.

Pinned Joints

Bolts cannot transmit shear loads because their clearance holes are too inaccurate to allow for a tight fit in a joint that may have many bolts. If a bolted joint were subject to a high shock load or sustained vibration, it is possible that the joint may slip. A tight-fitting solid pin of length L_{pin} in a hole bored through the joint can prevent this from occurring, as well as greatly increase the lateral stiffness of a bolted joint. Far away from the joint interface at the pin tip, the area A_{part} of the part that bears the shear load is equal to the cross section of the part. At the joint interface the pin's cross-sectional area A_{pin} supports all the shear load. Assuming a symmetric joint with linearly varying area, the shear stress as a function of position in the joint is

$$t = \frac{F}{A_{part} - \left(\frac{A_{part} - A_{pin}}{L_{pin}/2}\right) y} \tag{7.5.31}$$

Steel pins are almost always used, regardless of whether the parts are made from steel, cast iron, or aluminum. A conservative assumption is thus to assume that the shear modulus of the pins and parts are equal. Thus the stiffness of the region from the top of the pin to the bottom of the pin is

$$K + \frac{F}{\delta} = \frac{F}{2\int_0^{l_{pin}/2} = \frac{\tau}{G}\, dy} = \frac{G(A_{part} - A_{pin})}{L_{pin} Log_e(A_{part}/A_{pin})} \tag{7.5.32}$$

Note that in the limit as $A_{part} \Rightarrow_{pin}$, $K \Rightarrow GA_{part}/L_{pin}$.

Because the desired amount of lateral stiffness can often be obtained with bolt preload, often only two pins are used to maintain alignment should the joint be subject to vibration. If the joint is subject to shock loads and it is not desired for the joint to give, more pins may be needed to prevent deformation around the pins. After drilling and reaming a hole through both parts (after they are bolted and aligned), the pins can be pressed in. If desired, the first part can be reamed slightly larger, so the pin is a tight press fit in one part and a light press fit in the other part.[111] In this manner, if the parts are separated, the pins always reside in the same part. Care must be taken to make the pin hole 25% deeper than the pin to allow room to compress air in the bottom of the blind holes, or to grind a small flat on one side of the pin to allow the air to escape. Ideally, the hole should be a through hole, so the pin can be knocked out if required at a later date.

Solid steel pins (dowels) provide the greatest shear stiffness; however, they require the holes to be reamed, which is an extra manufacturing step. A roll pin is a round, hollow, hardened steel pin with an axial slit in one side. When pressed into a drilled hole that is a few percent smaller than the pin, the slit width is forced to decrease, thereby allowing the roll pin to fit the hole perfectly. Roll pins are well suited to general manufacturing applications or where very high shear stiffness is not needed (one just wants to locate and hold the parts for assembly or to resist light service loads). It

[110] It would be nice to see detailed graphs of compressive and Poisson deformations caused by bolts in various-size rails. Such graphs would probably have to be generated using finite element methods.

[111] One can calculate the desired press fit, or use standard tables provided in Machinery's Handbook, Industrial Press, New York. For large sections where there is obviously no danger of splitting the section, it is easier to do the latter.

is also less expensive, for example, to roll pin a part (e.g., a handle) to a shaft, then cut a keyway. Also, a roll pin will never loosen the way a setscrew can.

Stability of Bolted Joints

Most common joints that rely on mechanical contact between parts (i.e., not bonded or welded joints) may suffer a loss of initial preload with applied cyclic stress. This loss of preload seems to be due to the applied stress, causing microslip between the surfaces. For maximum dimensional stability, the design engineer may wish to specify vibrational stress relief and retightening of bolts, with the final tightening done while the vibration stress relief process is being done. In addition, if the joint is to be subject to vibrational loads or extreme stability is required, a hardening thread lubricant should be used.

For example, when a bearing rail is bolted in place, the bolts and threads should be prepared as discussed above, and then assembly should proceed in the following manner:

- The bearing rail should be aligned and the bolts incrementally tightened according to a specific pattern (e.g., from the inside out to prevent clamping in a bow) until the desired bolt torques are reached. The alignment should be checked after each tightening increment.
- The assembly should be vibration stress relieved.
- The alignment should be checked and the bolts retightened.
- The alignment should be checked, and then the assembly vibration stress relieved while the bolts are retightened.
- A final checking of the alignment should be performed.
- Ideally, the rail would be pinned to the structure with dowels and also potted in place with an epoxy, which helps increase joint stability and damping. As an alternative to pinning, a loose-fitting key that runs the length of the rail can be grouted or epoxied in place as the final assembly procedure.

The vibratory procedure accelerates the wear-in period of the permanent joints and helps prevent the machine from changing after the customer gets it. When properly implemented along with control of the maximum stress levels caused by the bolting process, the stability of a joint bolted by this process will probably be as good as one can get it. Most machine tools do not require the use of such an elaborate bolting procedure; however, as nanometer performance is sought, this type of procedure may become more common.

Bolting hardware[112]

For the machine design engineer, there are two basic types of bolts used to hold together structures: hex head bolts and socket head cap screws. The former are tightened with an open-end wrench, and the latter are tightened with a hex key (Allen) wrench. Handbooks give all necessary geometrical dimensions, such as head, body, and thread dimensions and required wrench clearances. There are also a number of other head configurations available, including decorative and tamperproof types. One need only consult a catalog of fastener companies to be overwhelmed with all the various types of fasteners that are available.

From an ergonomic viewpoint, it is wise never to specify the use of a bolt with a diameter smaller than 4-6 mm (No. 8-$^1/_4$ in.) and on heavy equipment 10 mm ($^3/_8$ in.) to avoid having the bolt head twisted off by overzealous bolt tighteners that may be tightening other larger bolts in the neighborhood. The best way for a design engineer to develop an intuition for "what bolt is right" is to take apart old machinery and fix old cars (e.g., rebuild an engine or change the rear end in an old truck). Remember, education teaches best those who teach themselves.[113]

Most machine tool applications use threads cut into one of the parts, but some applications require the use of nuts. As with bolts, there are many different types of nuts that are available to provide many different functions through special features such as:

- Special shapes to minimize the stress state in the threads to maximize fatigue life.

[112] The next three sections discuss bolts, nuts, and washers. For illustrative examples of different types, consult Machinery's Handbook, Industrial Press, New York, or a Thomas Register of American Manufacturers, Thomas Publishing Co., New York, which often shows informative pictures in the company listings.

[113] Translated for students in contemporary educational institutions, take off those damn headphones and observe the world around you. Signs, bridge girders, construction machinery, all have dozens of fascinating bolted joints to be looked at and analyzed in your head.

- Nonround threads and inserts to provide a locking effect.
- Special geometries (e.g., convex, serrated) on the seating face to enhance clamping action of the bolt on the parts.

For machine tool applications, a plain nut or a nut with a nylon insert to prevent loosening is most often used. The specific choice is usually a matter of company standards.

There are two basic types of washers used with bolts: flat washers, which prevent the surface from being marred by the bolt head and enhance load distribution on thin parts, and lock washers, which help maintain constant preload and prevent loosening of a bolt subject to vibration. In counterbored holes, flat washers are not used because the counterbore is only slightly larger than the bolt head. Split lockwashers are sometimes used with counterbored bolts; however, often a hardening thread lubricant is used instead. A split lockwasher is compressed by the bolt and thus can help maintain preload, even in the presence of bolt length changing, due to creep of highly stressed thread.

Summary

Summarizing recommendations for bolt torque selection:

- The theory will often yield a recommended torque level which is far below the allowable torque for a particular bolt:
  - The theory represents the minimum required torque.
  - The higher the torque, the better as far as machine stability is concerned.
  - Too high torques can deform components (e.g., cause rail straightness errors).
- It is best to experiment:
  - See what the component deformations are at different torque levels.
  - Use as high a torque level as possible.

7.5.2 Adhesive Joints[114]

Adhesive joints are fast becoming the most preferred type of joint in assemblies ranging from consumer products to airplanes. Bonded joints can be strong and fatigue resistant, and the process by which they are made is far more automatable than that used for almost any other joint. Virtually anything can be bonded to anything to perform in any type of condition desired. As shown in Figure 7.5.12, adhesives fill small voids between mating parts and thus can dramatically increase strength, stiffness, damping, and heat transfer characteristics of the joint. Even when two ground surfaces are to be bolted together, one can specify the use of a few drops of low viscosity adhesive that will flow and fill the small voids.[115] Some manufacturers have found that a sliding fit joint held by an adhesive (e.g., Loctite® Retaining Compound) makes a more dimensionally stable, accurate, and long-lasting joint than does a press-fit joint. This is particularly true for precision mechanisms (e.g., index tables and computer disk drives). Bearing races can also be fixed to bores in this manner to avoid altering the preload by press fitting.

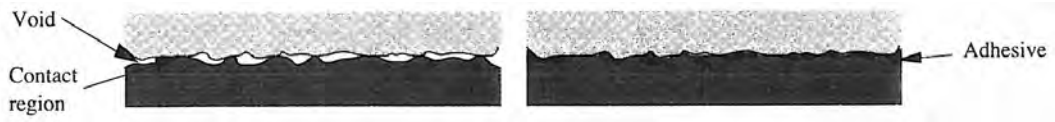


Figure 7.5.12 Joints typically make contact over 30% of the surface area. An adhesive can fill gaps, thereby increasing strength, stiffness, and damping. [From Guide to Adhesive Selection and Use (LT-1063), courtesy of Loctite Corp.]

[114] For a more detailed discussion of adhesive joints, see, for example, A. Blake, Design of Mechanical Joints, Marcel Dekker, New York, 1985, and M. Kutz (ed.), Mechanical Engineers' Handbook, John Wiley & Sons, New York, 1986.

[115] A quantitative study of how much adhesives help stiffness and damping as a function of surface finish and joint preload was not available at the time of this writing.

As with bolted joints, there is a plethora of information available on adhesive joints; thus only adhesive processes for machine tools will be discussed here, such as bonding, potting, and replication. In all cases, joint preparation is of prime importance, and the user should be careful to follow the manufacturer's recommendations.

Bonding

Bonding quite simply entails sticking one piece of material to another using an adhesive. The goal is to design the best geometry for the joint and choose the correct adhesive. Volumes of text exist on this subject, and adhesive manufacturers are usually happy to provide design assistance. As with other joints on machine tools, bonded joints are usually designed for stiffness, not strength. If the layer is thin enough, then the effective stiffness of the adhesive layer will often be much greater than that of the part itself. Because they act as continuous media for forming a joint, adhesives used by themselves or in conjunction with bolts are a must in every design engineer's tool kit.[116] A typical adhesive polymer (of the epoxy family) for machine tool use will have properties similar to the ones shown in Tables 8.2.2 and 8.2.3.

Bonding does not always mean that an isotropically rigid joint must be formed. For example, in Section 1.6 the coordinate measuring machine described had an aluminum frame and steel measuring scales. In this case it was necessary to prevent the aluminum frame from stretching the scales when the temperature the machine was used at was different than that at which it was assembled, or to prevent damage during shipping. To do this, one end of the scales was pinned to the aluminum frame, and the rest of the scale was bonded along its length using an adhesive that had excellent peel resistance, but low shear resistance. This type of joint can be used when thermal property mismatch between materials or temperature gradients can be expected to cause differential thermal expansion between components, and one wants to hold the two components together with constant spacing while still allowing the components to slide relative to each other.[117] A good application for this type of adhesive might be to hold a damping plate to a structure such as the one designed in Section 7.4.1.

Potting

A potted (sometimes referred to as grouted) joint uses an adhesive to lock in place a joint that uses mechanical means (e.g., bolted or gravity-held kinematic mount) to hold parts in alignment and to support the static weight of the structure. Dams are then built around the joint to prevent the adhesive from leaking out, and then the adhesive is poured to fill the region around the joint. The mechanical joint supports the static load and the potting compound (e.g., epoxy or cement grout) resists microslip of the mechanical joint and greatly increases joint stiffness and damping. Examples include grouting machine tool beds to concrete floors, potting major machine tool components to the machine tool structure (e.g., the headstock to the bed), and potting sensors in sensor mounts.

Note that in addition to external potting, internal potting can be used to fill keyways and keyholes that have loose-fitting keys or pins. This allows two parts to be locked together without pinning with press-fit dowels. An adhesive can be poured into a cavity (e.g., a cavity with an access hole) in the region between the parts of a joint to make the parts appear to mate perfectly. In this way, the adhesive forms a castable pin or key. Often the adhesive is injected around a loose-fitting pin or key in order to minimize the thickness of the adhesive layer.

Replication[118]

Replication is a process whereby a very low shrinkage polymer is poured or injected around a master shape that has been coated with mold release. When the polymer cures, it has the shape of the master, which can then be removed and used again. Because hardening of many polymer resins is an exothermic reaction (it gives off heat) one of the important aspects of this process is to minimize the amount of polymer used, and to maximize the stiffness and thermal diffusivity of the master and the part. This is required to prevent the heat of polymerization from heating the structure, deforming it, and then the polymer hardening to the thermally deformed shape. Unlike bolted assemblies, once the polymer cures to form the desired shape (e.g., a trough in which a linear bearing rail is placed),

[116] See, for example, M. Chowdhary, M. Sadek, and S. Tobias, "The Dynamic Characteristics of Epoxy Resin Bonded Machine Tool Structures," Proc. 15th Int. Mach. Tool Des. Res. Conf., 1975.

[117] For example, 3M Corp.'s EC-801 has a high tensile but low shear strength.

[118] See, for example, A. Devitt, "Replication Techniques for Machine Tool Assembly," East. Manuf. Tech. Conf., Springfield MA, Oct. 1989. Available from Devitt Machinery Co., Twin Oaks Center, Suite G, 4009 Market Street, Aston, PA 19014.

alignment is no longer possible. Note that for machines which are to be error mapped or used with a metrology frame, small misalignments are not usually of major consequence. Figures 7.5.13-7.5.15 show a replication process being used to replicate hydrostatic bearing pads. Note that to prevent porosity, the replicant must be thoroughly vacuum degassed.

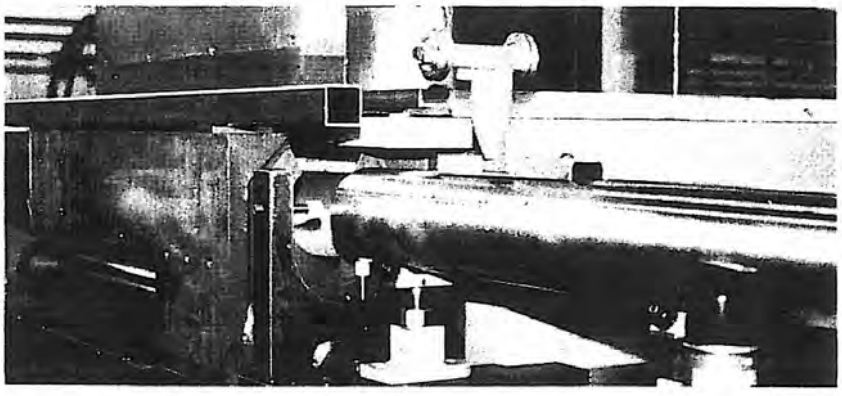


Figure 7.5.13 Beginning of replication process for hydrostatic bearing pads: The round bearing rail is being inserted into the bearing pad region, where it will be positioned using precision bored alignment fixtures. (Courtesy of Devitt Machinery Co. and Bryant Grinder Corp.)

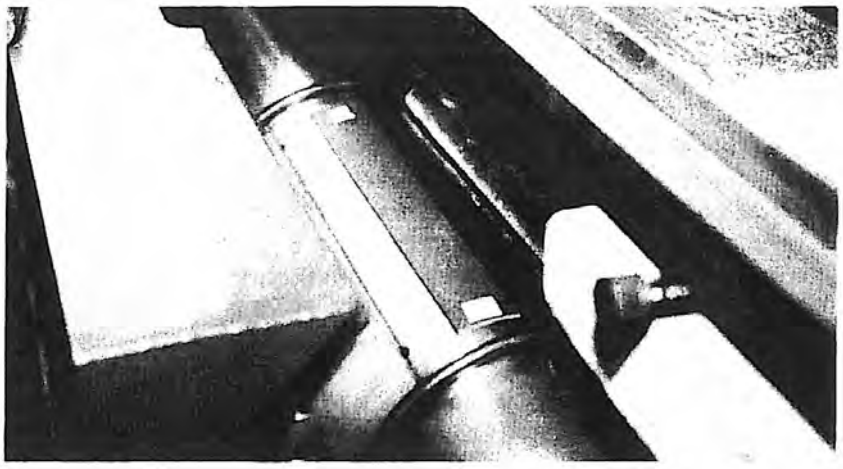


Figure 7.5.14 Replication process for hydrostatic bearing pads: O rings seal in the casting polymer and magnetic strips form the hydrostatic bearing pockets and drain grooves. (Courtesy of Devitt Machinery Co. and Bryant Grinder Corp.)

Figure 7.5.16 shows schematic details of surface preparation required to ensure that debonding does not occur. This sawtooth pattern can be obtained in a number of different ways. In addition to the rough surface finish required, the surface must also be thoroughly cleaned and a mold release apply to the master. Characteristics of replication polymers for bearing applications are discussed in Section 8.2.

Replicated parts and assemblies can be made to be as dimensionally stable as the rest of the machine. For example, as discussed in Chapter 1, diffraction gratings are among the most precise and scientifically important devices ever made, yet most modern gratings are actually replicated from a master grating.[119] In machine tools, mounting surfaces for linear bearing rails can be replicated, or in some cases for sliding or hydrostatic bearings the replicated surface functions as the way surface.

[119] See, for example, E. Loewen, Diffraction Gratings: Ruled and Holographic, Academic Press, New York, 1983.

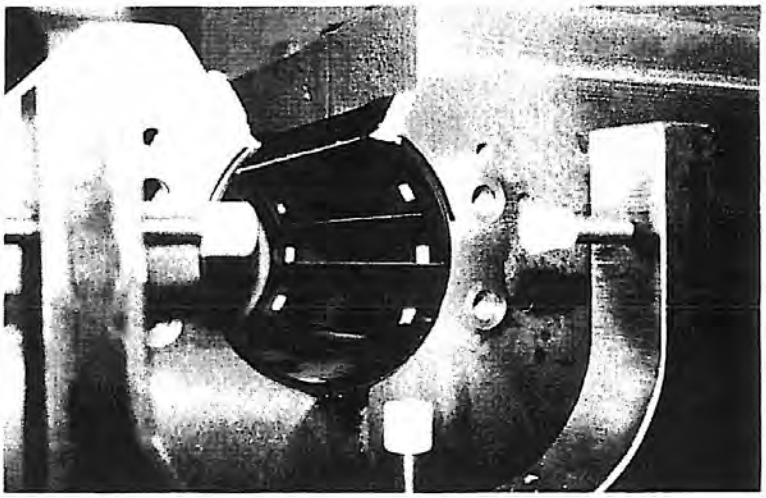


Figure 7.5.15 Replication process for hydrostatic bearing pads. After the replication material is mixed and then degassed in a vacuum, it is injected. After curing, the rail is removed and the bearing is complete. (Courtesy of Devitt Machinery Co. and Bryant Grinder Corp.)

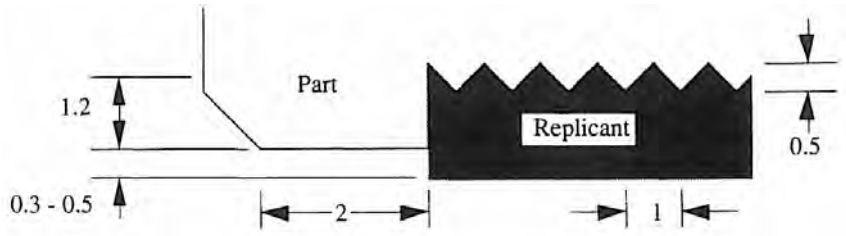


Figure 7.5.16 Details of surface and region for replication. Dimensions are in mm. (Courtesy of "DIAMANT" Metallplastic GmbH.)

Replicated surfaces are also used as historical records of components' surfaces, and as samples of large surfaces that need to be analyzed on a remote instrument.[120]

Remarks

Adhesives technology is advancing so rapidly that one must be careful to consult with the proper experts before choosing a specific adhesive, or ruling out adhesives in the first place. For example, "epoxy" is a term known to most all people, yet it is as generic as steel or wood: There are many different types of epoxies available for many different types of applications. Other types of adhesives will outperform epoxies in many applications (e.g., elevated temperatures). One should always check with various manufacturers to see what they think best suits your application. Tell them what you want to bond to what and what the operating conditions will be. Have them recommend an adhesive and explain why they chose that particular adhesive. Check with several manufacturers and make sure that their stories corroborate one another.

[120] See, for example, P. J. James and A. S. Collinge, "Assessment of Certain Epoxy Resins for Replicating Metal Surfaces," Precis. Eng., Vol. 1, No. 2, 1979, pp. 70–74, and P. J. James and W. Thum, "The Replication of Metal Surfaces by Filled Epoxy Resins," Precis. Eng., Vol. 4, No. 4, 1982, pp. 201–204.

7.5.3 Interference Fit Joints[121]

Shrinking a part over another part, or pressing a part into another part, has traditionally been a common, effective, economical, and reliable way to hold two parts together (e.g., to shrink fit a gear to a shaft); yet one of the most misunderstood forms of tolerancing is that associated with an interference fit. Care must be taken to ensure that the tolerances on the parts are such that when the dimensions differ minimally, sufficient contact pressure at the interface exists to transmit the desired load. Furthermore, if the parts are machined so that the bore is at its smallest allowable size and the shaft is at its greatest allowable size, the stresses around the bore must not be so great to cause the material to yield, even in the presence of other stresses in the system, including axial, torsional, bending, shear, pressure, thermal, and inertial stresses. The first two types of stresses often rely on the interference fit itself to transfer them across the joint; thus they directly affect the minimum required interference fit pressure. Bending stresses can be equated to axial stresses at the joint interface and generally are only of significance if a radial support is located far from the joint. Shear force stresses act to decrease the radial dimensions of the parts and can thus cause loosening of the joint. Thermal stresses can cause a part to come loose or split apart. Inertial stresses can also cause the part to come loose or split.

Axial Loads

The product of the minimum interface pressure, the coefficient of friction, and the interface area must be greater than the axial force. Assuming that the axial force has the safety factor built into it, the minimum interface pressure to allow transmission of the axial force without slipping is

$$P_{min} = \frac{F}{\mu\pi D\ell} \tag{7.5.33}$$

An axial force applied to a round member will cause the diameter of the member to change. From Hooke's law (Equations 7.3.3) one can show that the change in diameter is approximately

$$\Delta D = \frac{-4\eta F}{\pi D E} \tag{7.5.34}$$

After the required interference fit to support the axial load is found, the absolute value of the change in diameter must be added to the diametrical interference fit required.

Torsional Loads

The product of the minimum interface pressure, the interface area, the coefficient of friction, and the radius of the interface must be greater than the applied torque. Assuming that the torque has the safety factor built into it, the minimum pressure to allow transmission of the torque without slipping is

$$P_{min} = \frac{2\Gamma}{\mu\pi D^2\ell} \tag{7.5.35}$$

Since the torsional and axial motions are orthogonal, the interface pressure calculated must be able to resist the resultant of the axial and tangential force vectors.

Bending Loads

When a beam bends, one surface is in tension and the other surface is in compression. In order to prevent the joint from working itself loose, the product of the interface pressure and the coefficient of friction must be greater than the maximum tensile or compressive stress in the beam. Furthermore, in order to transfer the bending moment effectively across the joint, the second moment of the area (commonly called the moment of inertia) of the outer part's cross section must be greater than the second moment of the area of the inner part's cross section. In general, the Poisson contraction caused by the tensile forces will cancel the Poisson expansion caused by the compressive forces, so the application of a static moment should not cause loosening of the joint if the interface pressure criterion is met.

[121] For further information on interference fits, see O. J. Horger (ed.) the ASME Handbook, American Society of Mechanical Engineers, New York, 1953, p. 178. For a rigorous discussion of this subject, see S. Timoshenko and J. Goodier, Theory of Elasticity, McGraw-Hill Book Co., New York, 1951, p. 388. A good condensation of this information is also provided by M. Kutz (ed.), Mechanical Engineers' Handbook, John Wiley & Sons, New York, 1986.

Shear Loads

Shear loads are often present in shafts but are of little consequence to interference fit joints unless they are applied within one or two diameters of the joint (e.g., a bearing support placed near a gear). The shear load compresses the shaft, making it into an oval shape, and because of the Poisson effect, the expansion will only be a fraction of the contraction. Two shear loading situations commonly exist: (1) the shear load is applied through the part that was shrunk fit onto the shaft, and (2) the shear load is applied by another part that fits on the shaft. In the former case, the applied radial load will increase the pressure on one side of the bore, and will decrease it on the other side of the bore. The net result is that the axial holding capability of the joint will be maintained; however, on the loosened side, the interference pressure minus the shear load's stress should still be higher than that calculated to support all other types of loads; otherwise, microslip and fretting of the interface region may result.

The latter case typically occurs when a pulley on a shaft turns a large flywheel. The weight of the flywheel compresses the shaft in the bearing housing. The pulley is usually located next to the bearing housing to minimize bending stresses caused by belt tension, so the interference fit is subject to the resulting decrease in shaft diameter. An upper bound estimate on the amount of decrease in diameter can be found by assuming that the shear load is a line load applied to the shaft (cylinder), and then calculating the compression of the shaft using Hertz contact stress theory as discussed in Section 5.6. The diameters of the shaft would be D and infinity, and the diameters of the bearing bore would be assumed to be 1.1D and infinity. A lower bound estimate of the decrease in interference pressure would be to assume that it equals the shear load V divided by the product of the diameter and length of the bearing support: $P_{decrease} = V/DL$.

Pressure Stresses

If the assembly is subject to internal or external hydrostatic pressure, then the components may expand or contract, causing tightening or loosening of the joint. Tightening could cause bursting, while loosening could cause the joint to also fail. Thus one must evaluate the effect of hydrostatic pressure on the interface pressure between the components of an interference fit assembly.[122]

Thermal Loads

Temperature changes can cause the diameters of the assemblies to vary, which directly affects the allowable minimum and maximum interference fits. For the inner and outer cylinders, the change in outside diameter relative to inside diameter at the diameter D of the interference fit for a uniform temperature change ΔT from the assembly temperature is

$$\Delta D = D\Delta T(\alpha_1 - \alpha_2) \tag{7.5.36}$$

If the coefficient of thermal expansion α_O is greater than α_I, ΔD will be negative and the fit will "loosen up" as the assembly warms up. Even for shafts and gears bathed in oil, this temperature change can be significant (10 C^o or more) and thus so can the change in diameter.

Assuming that the calculated ideal interference fit was in the range of δ_{min} to δ_{max}, if ΔD is negative, then in order to maintain the holding strength of the joint when the assembly heats up, the interference fit must lie in the range $\delta_{min} - \Delta D$ to δ_{max}. If ΔD is positive, then in order to prevent yielding in the materials when the assembly heats up, the interference fit must lie in the range δ_{min} to $\delta_{max} - \Delta D$. Thermal deformations may not seem significant, but for cases where the temperature change can be in the hundreds degrees (e.g., in engines) they can be significant, and matching of material expansion coefficients can be very important. For cases where temperature gradients are to be encountered, first design the assembly assuming that a uniform temperature difference exists between the two parts (e.g., outer part at T_o and inner part at T_i), then check the performance of the design in a gradient using closed-form solutions if possible, or finite element analysis.

Inertial Stresses

Many interference fits transmit large forces or torques at little or no speed; however, in many applications (e.g., gearboxes) the assembly may be turning at many thousands of RPM. When a body spins, centrifugal forces tend to expand the body and cause internal radial and circumferential stresses. The stresses should not cause the gear to fly apart and the gear must not become loose on

[122] See Sections 7.5.3.2 and 7.5.3.3.

the shaft. A loose gear will be inaccurate and will rapidly wear. Determination of the stresses and deformation produced by these forces is a well-researched subject that is briefly discussed below.

7.5.3.1 Finding the Required Amount of Interference

In order to determine the required interference between two parts, one must determine what pressure is required to transmit the incident loads across the interface in the presence of a finite coefficient of friction between the materials at the interface. One can, for example, perform the following steps:

1. Size the components of the assembly to enable them to carry the applied loads (axial, torsional, bending, shear, thermal, and inertial loads) in whatever combination is dictated by the application. In addition to the usual multipliers used to account for stress concentrations and fatigue life, include a multiplying factor of 2 to account for the interference fit stresses. As a by-product of this step, one should obtain the resultant stresses due to the loads and the displacements at the interface due to each of the loads on the system. It is wise to incorporate safety factors into the loads during this step.
2. Find the amount of interference between the two parts that will result in an interface pressure acting over the interface area that will enable the system to support the applied loads. Note that the most conservative approach to this step would be to assume that all maximum loads act simultaneously; however, this is not always the case. For example, an electric motor driving a shaft has a torque-speed curve, which says that the motor outputs maximum torque at minimum speed and maximum speed at minimum torque. To the value of the required interference, add the value of the worst combination of high and low tolerances for the part and its mate, which manufacturing says they can readily and continuously achieve. For example, a tolerance of ± 0.002 mm on a shaft and a tolerance of ± 0.001 mm on the bore means that the interference may be 0.003 mm larger or smaller than what the design engineer anticipated. If manufacturing errs on the minus side of the tolerance, the part assembly should still have enough holding power. If the error is made on the plus side, the assembly should not break due to increased interference pressure.
3. Calculate the interference stress in the assembly assuming that the interference equals each of the high and low values found in step 2, plus the absolute value of the total displacement (caused by the loads) at the interface found in step 1. This will ensure that the interface stress calculated includes the effects of the applied loads. With this method, if the applied loads act to decrease the diameter of the inner part (e.g., a shaft in tension decreases in diameter due to the Poisson effect), the interference will have to be that found in step 2 plus the amount the shaft shrinks. This will ensure that when the loads are applied, the interference fit will still have enough holding capacity. If the applied loads act to increase the diameter of the shaft, the calculation represents an upper bound for the stresses at the joint.
4. Calculate the resultant state of stress in the body due to the applied loads and the interference fit, using Mohr's circle or an equivalent yield criteria, such as Von Mises. If the product of the equivalent stress and the safety and fatigue loading factors exceeds the allowable yield stress, you must increase the size of the parts and go back to step 2. If the product is less, you may want to decrease the size of the components and return to step 2.

Often implementation of steps 1-4 will be complex enough to warrant the use of a computer and a recursive algorithm to home in on the "optimal" values. Steps 1-4 themselves outline the structure of the algorithm. In addition, in order to determine what the tolerances on an interference fit must be, one must understand how to determine the pressure at the interface and determine the resulting stresses in the parts. Thin- and thick-walled part assemblies are discussed below. Note that the analysis methods themselves are synonymous with the design process and thus they should not be skipped over just to find the final answer at the end of the sections.

Implementation Details

The holding power of an interference fit joint depends on the coefficient of friction and the amount to which the surface asperites of the two parts dig into each other, forming a mechanical link. If friction is the assumed dominant source of holding power, it should be maximized, which means assembling the parts after they have been cleaned and degreased. For similar metals in contact (e.g.,

steel on steel), the coefficient of friction can be on the order of 0.2-0.3. For cast iron, in which the graphite acts as a lubricant, the coefficient of friction may be about 0.1. For dissimilar materials (e.g., steel on brass), one should assume that the coefficient of friction is only about 0.1. Stress concentration at the joint can be minimized through the use of radii, as shown in Figure 7.5.17.

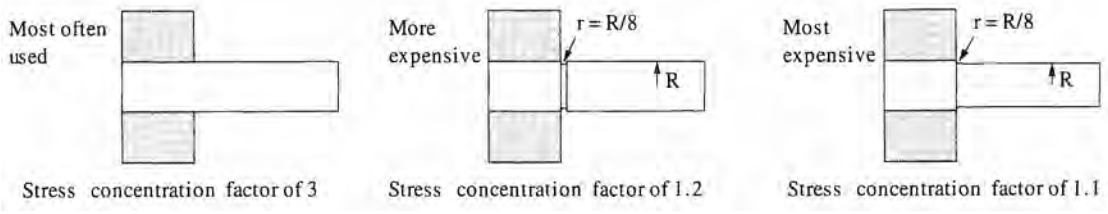


Figure 7.5.17 Shrink fit configurations for a shaft and part.

Microslip occurs at even small tangential stress levels. One might tend to think that the rougher the mating surfaces, the better the chance that the peaks and valleys will interlock, thereby decreasing microslip and increasing the holding power of the joint. However, as the surface roughness increases, the stiffness and dimensional location ability decreases. In general, the finer the surface finish (on the order of 0.5 μm R_a) the more the joint appears to be solid, and perhaps clean, high surface finish parts actually cold weld together after they are press fit. Design engineers interested in designing interference fit joints to achieve a desired stiffness, as opposed to the usual design for load-holding capability, must consult empirical data.[123]

To increase the holding power, stiffness, and corrosion resistance of an interference fit joint while removing some of the quality problems associated with having to carefully control interference fits and surface finish, the joint can also be brazed or have an adhesive used in the interference fit assembly process. Brazing is used when the assembly is to be heat treated, and an adhesive is used when the assembly does not have to be heat treated or used at elevated temperatures. When the assembly is heat treated, the brazing metal melts and by capillary action is drawn into the spaces between the asperites in the parts. The heat treatment process may relieve high interference stress, but the metallic bond formed by brazing maintains joint strength and the joint's creep resistance at high temperatures. Thus for brazing, the purpose of the interference fit is primarily to act as a centering device. An adhesive applied to the parts before the joint is assembled would fill the gaps between the asperites and increases the holding strength of the joint, but interference stresses would still exist. Thus with an adhesive, the holding power of the joint is the sum of that provided by the interference pressure and the adhesive, but the parts are more highly stressed.

With the use of a brazing process or adhesive, the surface finish should be rough (1 or 2 μm R_a) so as to provide space for the brazing metal or adhesive to form a bond between the parts. To assist in the flow of metal or adhesive, a crosshatch pattern surface finish can be specified (such as obtained with a flex-hone).

Interference Fits Between Thin Walled Tubes

In order to determine the allowable interference between a thin[124] outer tube that is to be shrunk over a thin inner tube, consider that change in diameters is simply a function of to-be-determined compliances and desired interface pressure P:

$$\delta_0 = C_0 P \tag{7.5.37a}$$

$$\delta_I = C_I P \tag{7.5.37b}$$

Geometric compatibility implies that the outside diameter of the inner tube will be equal to the inside diameter of the outer tube after they are assembled. Since the outer tube is smaller than the inner tube, its diameter must increase, while the inner tube's diameter must decrease:

$$D_0 + C_0 P = D_I - C_I P \tag{7.5.38}$$

[123] R. H. Thornley and I. Elewa, "The Static and Dynamic Stiffness of Interference Shrink-Fitted Joints," Int. J. Mach. Tools Manuf., Vol. 28, No. 2, 1988.

[124] A "thin-walled tube" is one where D >> t (e.g., D > 10t).

The pressure can only cause a circumferential stress in both the inner and outer tubes. By symmetry, the pressure must be constant around the circumference. The circumferential stress creates a strain ε_θ in the circumference given by Hooke's law to be $\varepsilon_\theta = \sigma_\theta/E$. The change in circumference $\delta_\theta$orresponding to a circumferential strain ε_θ is

$$\delta_\theta = \pi D \varepsilon_\theta \tag{7.5.39}$$

The change in diameter ΔD corresponding to a change in circumference δ_θ is

$$\Delta D = \frac{\delta_\theta}{\pi} \tag{7.5.40}$$

Thus the change in diameter caused by a circumferential strain is

$$\Delta D = D \varepsilon_\theta \tag{7.5.41}$$

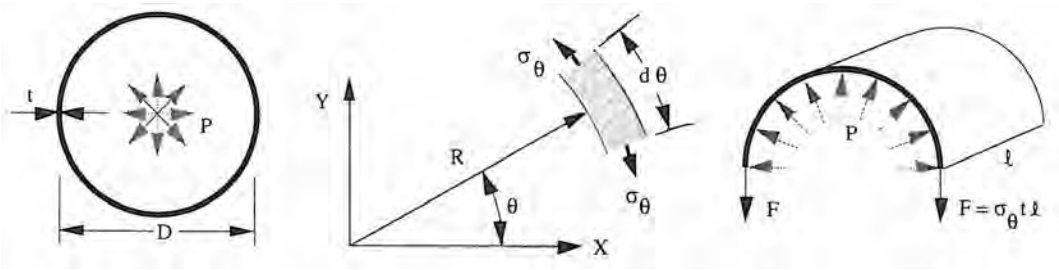


Figure 7.5.18 Thin-walled tube subject to internal pressure P.

The next step is to find σ_θ in terms of P, D, and t (the wall thickness). As shown in Figure 7.5.18, if the hoop is cut along a diameter, the circumferential stress σ_θ must balance the pressure forces. A radially varying stress exists and varies from the internal pressure to the exterior pressure (e.g., 0). This stress, however, is insignificant compared to the circumferential stress, as will be shown. The differential force in the Y direction created by the pressure acting on a differential element of unit length $d\theta$ is $0.5D\theta P\sin\theta$, so the σ_θ stress is

$$\sigma_\theta = \int_0^\pi \frac{D\sin\theta}{2} d\theta = \frac{PD}{2t} \tag{7.5.42}$$

By definition, for a thin-walled tube $D >> t$ and thus $\sigma_\theta >> \sigma_r$ so σ_r can be neglected. The circumferential strain is thus

$$\varepsilon_\theta = \frac{PD}{2tE} \tag{7.5.43}$$

Thus the change in diameter ΔD of a thin-walled tube caused by a pressure[125] acting on the tube is

$$\Delta D = \frac{PD^2}{2tE} \tag{7.5.44}$$

The compliance C is thus $D^2/2tE$, and upon substitution into Equation 7.5.38,

$$D_0 + \frac{P D_0^2}{2t_0 E_0} = D_I - \frac{P D_I^2}{2t_I E_I} \tag{7.5.45}$$

Substituting $D_O = D_I - \Delta D$ into Equation 7.5.45 yields

$$D_I - \Delta D = + \frac{P\,(D_I^2 - 2D_I \Delta D + \Delta D^2)}{2t_0\, E_0} = D_I - \frac{P\, D_I^2}{2t_I\, E_I} \tag{7.5.46}$$

[125] P is positive when applied internally and P is negative when applied externally.

Terms like ΔD^2 and $2D_I\Delta D$ are small with respect to ${D_I}^2$. The required diametrical interference ΔD in terms of the required interface pressure P is thus

$$\Delta D \approx \frac{P\,D_I^2\,(t_0E_0 + t_I\,E_I)}{2t_0\,t_I\,E_0\,E_I} \tag{7.5.47}$$

The circumferential stresses in the outer and inner tubes are found from Equation 7.5.42:

$$\sigma_{\theta 0} = \frac{D_0 t_I E_I E_0 \Delta D}{D_I^2(t_I E_I + t_0 E_0)} \qquad \sigma_{\theta I} = \frac{t_0 E_I E_0 \Delta D}{D_I(t_I E_I + t_0 E_0)} \tag{7.5.48}$$

Consider tubes made of the same material with equal wall thicknesses, Equations 7.5.48 are on the order of

$$\sigma_{\theta max} = \frac{E(\Delta D/2)}{D_I} \tag{7.5.49}$$

The wall thickness term cancels out and we are left with an expression that says the product of the diametrical strain in either of the tubes (half of the total diametrical interference, since each tube expands, divided by the diameter) and Young's modulus is equal to the stress. As the wall thickness increases, a proportionately larger pressure is required to expand the tube to accommodate the diametrical interference, and hence the thickness term drops out.

The axial extension of the inner part beyond the axial dimensions of the other part will create added resistance to compression by the outer part at its ends and thus the interface pressure will be higher in this region. By ignoring this effect, a conservative estimate for the force or torque that can be transmitted across the interface will be obtained; however, one must then be more generous with the safety factor for the allowable stress since the interface pressure will be higher. In addition, in order to minimize stress concentrations, the entrances to the outer parts inside diameter should be beveled, or the inner part radiused, so the transmission is gradual and does not lead to a stress concentration.

Interference Fits Between Thick-Walled Bodies

For thick-walled bodies, the circumferential stress varies with the radius, and there is a stress component in the radial direction which can be on the order of the circumferential stress and thus cannot be neglected. Figure 7.5.19 shows a section of unit thickness cut from a thick-walled cylinder. In order to determine the minimum and maximum interference required, we proceed in a similar manner as was done for the thin-walled cylinder. A major assumption is that the problem is two-dimensional, so cross sections of the cylinder remain plane after deformation.

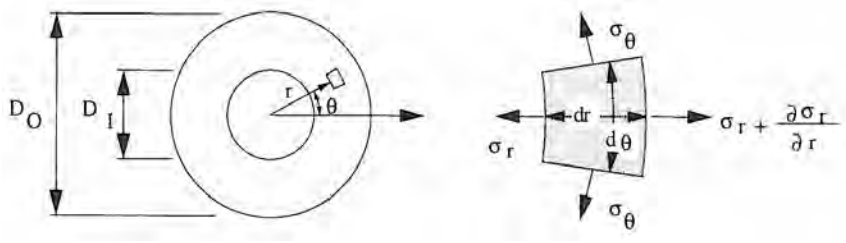


Figure 7.5.19 Section of a thick-walled cylinder subject to internal and external pressures.

The radial stress σ_r varies across a dr thick circumferential element by an amount $(\partial\sigma_r/\partial r)$ dr. Because of symmetry, the circumferential stress σ_θ must be constant for this element, but σ_θ will vary with the radius r. Summing the forces in the radial direction

$$-\sigma_r r d\theta - \sigma_\theta dr d\theta + \left(\sigma_r + \frac{\partial \sigma_r}{\partial r}\,dr\right)(r + dr)d\theta = 0 \tag{7.5.50}$$

Neglecting higher terms such as dr^2, Equation 7.5.50 reduces to

$$\sigma_\theta - \sigma r - r\frac{\partial \sigma_r}{\partial r} = 0 \tag{7.5.51}$$

There are two unknowns, σ_θ and σ_r, so geometric compatibility must be considered.

The deformation of a thick-walled cylinder subject to uniform pressure is a symmetrical problem and thus must consist only of the radial displacement of all points in the cylinder. The radial displacement cannot vary with circumferential position because of symmetry, but it may vary radially. Thus the other physical parameter of the problem is the radial displacement u of material at a given radius. If u and r can be related to the strains ε_θ and ε_r produced by the stresses σ_θ and σ_r, then Equation 7.5.51 can be reduced to one equation in one unknown. The known relation for the displacement u can then be used to find ε_θ and ε_r and then σ_θ and σ_r.

If the displacement at any radius r is u, then at a radius r + dr the displacement will be u + (?u/?r)dr, so the change in radial displacement is just (∂u/∂r)dr. The total elongation of a differential element in the radial direction is equal to the product of the strain and the length dr:

$$\Delta u = \varepsilon_r \, dr = \frac{\partial u}{\partial r} \, dr \tag{7.5.52}$$

The radial strain ε_r is thus

$$\varepsilon_r = \frac{\partial u}{\partial r} \tag{7.5.53}$$

In the circumferential direction, the change in circumference for a given displacement of a point that used to be at radius r but due to the applied pressure has moved to r + u, equals the product of the circumference and the circumferential strain:

$$2\pi r \varepsilon_\theta = 2\pi(r + u) - 2\pi r \tag{7.5.54}$$

Thus the circumferential strain is

$$\varepsilon_\theta = \frac{u}{r} \tag{7.5.55}$$

From Hooke's law, the stresses σ_r and σ_θ are found in terms of the displacement u:

$$\sigma_r = \frac{E}{1-\eta^2}\left(\frac{d\,u}{dr} + \eta\frac{u}{R}\right) \qquad \sigma_\theta = \frac{E}{1-\eta^2}\left(\frac{u}{r} + \eta\frac{du}{dr}\right) \tag{7.5.56}$$

Substituting these expressions into Equation 7.5.51 yields

$$\frac{d^2u}{dr^2} + \frac{1}{r}\frac{d\,u}{dr} - \frac{u}{r^2} = 0 \tag{7.5.57}$$

From an introductory differential equations textbook, one finds that the general solution to this problem is

$$u = C_1 r + \frac{C_2}{r} \tag{7.5.58}$$

where the constants C_1 and C_2 depend on the boundary conditions. The boundary conditions are functions of the pressures applied at the outside and inside surfaces of the cylinder: $\sigma_{r=b}$ = -P_O and $\sigma_{r=a}$ = -P_I. The radial displacement at any radius r in a cylinder is

$$u = \frac{1-\eta}{E}\left(\frac{D_I^2\,P_I - D_0^2 P_0}{D_0^2 - D_I^2}\right) r + \frac{1+\eta}{E}\left(\frac{D_I^2 D_0^2 (P_I - P_0)}{4\left(D_0^2 - D_I^2\right) r}\right) \tag{7.5.59}$$

For a cylinder subject to internal pressure, the radial displacement of the inner surface is

$$u_{\text{inner surface}} = \frac{D_I P}{2E}\left(\frac{D_I^2 + D_O^2}{D_O^2 - D_I^2} + \eta\right) \tag{7.5.60}$$

For a cylinder (or shaft when D_I = 0) subject to external pressure, the radial displacement at the outer surface is

$$u_{\text{inner surface}} = \frac{-D_O P}{2E}\left(\frac{D_I^2 + D_O^2}{D_O^2 - D_I^2} - \eta\right) \tag{7.5.61}$$

For a shrink fit, the displacements will be equal and opposite. Their difference will thus be one-half of the diametrical interference Δ. The interface pressure will thus be

$$P_{interface} = \frac{\Delta}{D_{interface}\left\{\frac{1}{E_O}\left(\frac{D_{interface}^2 + D_{outside}^2}{D_{outside}^2 - D_{interface}^2} + \eta\right) + \frac{1}{E_I}\left(\frac{D_{inside}^2 + D_{interface}^2}{D_{interface}^2 - D_{Inside}^2} - \eta_I\right)\right\}} \tag{7.5.62}$$

For the case of a solid shaft press fit into a very large ($D_{outside} >> D_{interface}$) structure (e.g., a press fit plug), the interface pressure is

$$P = \frac{A}{D_{interface}\left\{\frac{1+\eta}{E_O} + \frac{1-\eta}{E_I}\right\}} \tag{7.5.63}$$

The next step is to determine the stresses produced by the interface pressure. By substituting Equation 7.5.59 into Equations 7.5.56, and applying the boundary conditions, expressions for the stresses in a cylinder subject to internal and external pressures:

$$\sigma_r = \frac{D_I^2 P_I - D_O^2 P_O}{D_O^2 - D_I^2} - \frac{(P_I - P_O)D_O^2 D_I^2}{4r^2(D_O^2 - D_I^2)} \tag{7.5.64a}$$

$$\sigma_\theta = \frac{D_I^2 P_I - D_O^2 P_O}{D_O^2 - D_I^2} - \frac{(P_I - P_O) D_0^2 D_I^2}{4 r^2 \left(D_O^2 - D_I^2\right)} \tag{7.5.64b}$$

When a cylinder is only subjected to an internal pressure P_I, the stresses are

$$\sigma_r = \frac{D_I^2 P_I}{D_0^2 - D_I^2}\left(1 - \frac{D_0^2}{4r^2}\right) \qquad \sigma_\theta = \frac{D_I^2 P_I}{D_0^2 - D_I^2}\left(1 + \frac{D_0^2}{4r^2}\right) \tag{7.5.65}$$

The maximum stresses occur at the inner surface of the cylinder where r = D_I/2. The radial stress is compressive and equal to -P_I and the circumferential stress is tensile and its magnitude is always larger than the radial stress. For the case where a cylinder is subject to an external pressure P_O, the stresses are

$$\sigma_r = \frac{-D_0^2 P_0}{D_0^2 - D_I^2}\left(1 - \frac{D_I^2}{4r^2}\right) \qquad \sigma_\theta = \frac{-D_0^2 P_0}{D_0^2 - D_I^2}\left(1 + \frac{D_I^2}{4r^2}\right) \tag{7.5.66}$$

Note that σ_r and σ_θ are both compressive. These equations give the equivalent stress state for the inner and outer surfaces of inner and outer cylinders of a shrinkfit assembly. In order to determine for the maximum allowable interface pressure, one must combine these stresses with the stresses formed at the inner and outer surfaces of the parts by the other applied loads using, for example, the Mises criteria.[126]

Rotating Assemblies

When the assembly is rotating, the radial force acting on the differential element of Figure 7.5.19 is (ρ is the *mass* per unit volume)

$$dF_r = \rho\omega^2 r^2\, dr\, d\theta \tag{7.5.67}$$

Incorporating this term into Equation 7.5.50, the equilibrium equation becomes

$$\sigma_\theta - \sigma_r - \frac{r\partial\sigma_r}{\partial r} - \rho\omega^2 r^2 = 0 \tag{7.5.68}$$

Proceeding as before, the equilibrium equation in terms of the displacement becomes[127]

$$\frac{d^2u}{dr^2} + \frac{1}{r}\frac{du}{dr} - \frac{u}{r^2} + \frac{\rho\omega^2 r}{E^*} = 0 \tag{7.5.69}$$

[126] For metals, failure occurs when the maximum shear stress $\tau_{max}$is exceeded. The maximum shear stress is equal to one-half of the difference between the maximum and minimum principal stresses. With the tresca relation, $\tau_{max} = 0.5\sigma_{yield}$. With the Mises relation (Equation 7.3.1) $\tau_{max} = 3^{1/2}\sigma_{yield}$. Since interference fit calculations require other stresses to be incorporated into the analysis, the Mises criteria is easier to apply.

[127] Recall that $E^* = E/(1 - \eta^2)$.

The solution to this equation is

$$u = \frac{-\rho\omega^2 r^3}{8E^*} + c_1 r + \frac{c_2}{r} \tag{7.5.70}$$

The constants are first evaluated for a disk with a hole at the center, which is the case where the shaft of the interference assembly is hollow, then the boundary conditions are $\sigma_r = 0$ at $r = D_I/2$, $D_O/2$. After much algebra[128] the radial displacement and stresses are found:

$$u = \frac{\rho\omega^2}{8E^*}\left[-r^3 + (3+\eta)\left(\frac{(D_0^2 + D_I^2)r}{4(1+\eta)} + \frac{D_0^2 D_I^2}{16(1-\eta)r}\right)\right] \tag{7.5.71}$$

$$\sigma_r = \frac{\rho\omega^2(3+\eta)}{8}\left(\frac{D_I^2}{4} + \frac{D_0^2}{4} - r^2 - \frac{D_I^2 D_0^2}{16r^2}\right) \tag{7.5.72}$$

$$\sigma_\theta \frac{\rho\omega^2(3+\eta)}{8}\left[\frac{D_I^2}{4} + \frac{D_0^2}{4} - \frac{1+3\eta}{3+\eta} r + \frac{D_I^2 D_0^2}{16r^2}\right] \tag{7.5.73}$$

The maximum stress occurs at $r = 0.5(D_O D_I)^{1/2}$. Thus for the assembly, the equivalent stress state must be found at the outside surface of the outer part, $r = 0.5(D_O D_I)^{1/2}$, the interference surface, and the inner surface of the inner part. This gets complicated, and the best way to keep track of all the terms is to evaluate the stress state numerically and plot it as a function of interface pressure. Note that in such a program, one must keep track of the displacements of the interface caused by the loads and determine what loosening or tightening effect they may have on the joint.

When the inner cylinder is a solid shaft, the constant $C_2 = 0$ and the boundary condition $u = 0$ at $r = 0$ exists. The stresses are thus

$$\sigma_r = \frac{\rho D_0^2 \omega^2(3+\eta)}{32}\left(1 - \frac{4r^2}{D_0^2}\right) \tag{7.5.74a}$$

$$\sigma_\theta = \frac{\rho D_0^2 \omega^2(3+\eta)}{32}\left(1 - \frac{(1+3\eta)4r^2}{(3+\eta)D_0^2}\right) \tag{7.5.74b}$$

When designing high-speed rotating machinery, extreme care must be used. More than one reference should be consulted and design equations and calculations should be checked and rechecked. Also, one must ensure that the design conforms to applicable codes and standards.

7.6 SUPPORT SYSTEMS

In addition to the overall geometry and selection of principal components, the design engineer must carefully consider the selection of *support systems,* including:

- Electrical systems
- Chip removal systems
- Fixturing systems
- Fluid systems
- Safety systems
- Sealing systems
- Tooling systems
- Vibration control systems

Each of these systems can be as critical to the successful implementation of a design as the obvious critical systems. However, it is the support systems that are most often left until last, which can result in poor integration. Hence basic characteristics of each of these systems will be discussed along with how they can affect the performance of a machine tool's design.

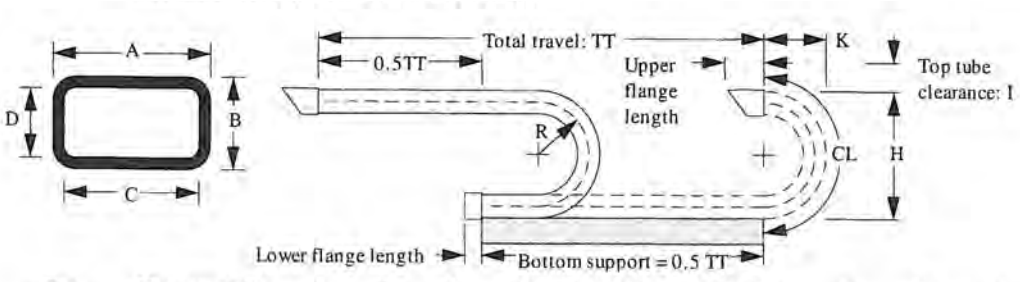
[128] For a more detailed discussion, see S. Timoshenko, Strength of Materials, Part II, Theory and Problems, Robert Krieger Publishing Co., Melbourne, FL.

7.6.1 Electrical Systems

Electrical systems should be treated with the same respect due your own nervous system. The mechanical design engineer may not design the electrical systems, but he must work closely with the electrical system design engineer to provide adequate space and wire mounting locations for neatly running wiring in a manner that is conducive to manufacturing the machine, and for tracing wires in the event repair is needed. Anyone who has ever spent time trying to trace a short in an automobile wiring system (particularly behind the dashboard) will appreciate this comment.[129]

Electrical systems are most often responsible for failures in moving axes. This is because it is difficult to make connections to moving axes. The design engineer must pay careful attention to the choice of cable carriers and critical bend radii for wires in the carrier. A typical cable carrier design data sheet is shown in Figure 7.6.1. These carriers can also be used for fluid lines. Also, power and sensor cables should be shielded and separated from each other to prevent electromagnetic interference (EMI). When selecting a cable carrier, there are several important issues:

- Assume that the cables are 10% larger when determining the required interior carrier size.
- The bending radius should be eight times the diameter of the largest cable, or larger if recommended by the cable manufacturer.



Type	Max span* (ft)	A	B	C	D	R (±10%)	H 9±10%)	K	CL	I
C0-4	3	1.18	0.79	0.90	0.62	1.8	4.4	3.2	6.5	1.0
C1-6	5	1.97	1.18	1.71	0.90	2.4	5.9	4.0	9.0	1.5
C1-9	5	1.97	1.18	1.71	0.90	4.0	9.1	5.6	15.5	1.5
C1-13	5	1.97	1.18	1.71	0.90	5.9	13.0	7.5	23.5	1.5
C2-10	6.5	3.15	1.77	2.87	1.50	4.0	9.7	5.9	14.5	2.0
C2-175	6.5	3.15	1.77	2.87	1.50	7.9	17.5	9.8	30.0	2.0
C2-22	6.5	3.15	1.77	2.87	1.50	9.9	21.5	11.8	33.5	2.0
C2A-12	7.5	3.74	1.97	3.46	1.70	4.9	11.9	7.0	17.5	2.5
C3-135	8	4.33	2.36	4.01	2.05	5.5	13.5	7.7	19.0	2.5
C3-20	8	4.33	2.36	4.01	2.05	8.9	20.1	11.1	32.5	3.0
C3-26	8	4.33	2.36	4.01	2.05	11.8	26.0	14.0	44.5	3.0
C4-18	9.5	6.69	3.15	6.38	2.83	7.3	17.8	9.9	25.0	3.0
C4-23	9.5	6.69	3.15	6.38	2.83	9.9	22.8	12.5	35.5	3.0
C4-31	9.5	6.69	3.15	6.38	2.83	13.8	30.7	16.4	51.0	3.0
C5-22	10.5	6.69	3.54	6.38	3.23	9.1	21.7	11.9	36.5	3.0
C6-23	11	7.87	3.93	7.56	3.62	9.5	22.8	12.5	38.0	3.0
C7-24	11.5	8.66	4.33	8.35	4.02	9.9	24.0	13.0	39.5	3.0

* Addition of one roller increases maximum span by 50%, 2 rollers by 100%.
All dimensions are in inches unless noted.

Figure 7.6.1 Typical cable carrier design data sheet. (Courtesy of A&A Manufacturing Co.)

[129] One of the greatest assets a design engineer can have is an old car that frequently breaks down. Hours spent scraping knuckles while fixing it can be very enlightening as to how to design something and how not to design something with respect to maintainability. One of the best gifts you can give your kid is a beat-up old car and the resources and help to fix it, as opposed to a new car.

7.6.2 Chip Removal Systems

If the machine is designed to remove material, regardless of the process, careful attention must be paid to how the waste material will be removed from the work zone. The less material that is removed, the easier the cleanup may appear; however, the less material that is removed, the more sensitive subsequent operations are likely to be to the presence of waste material from previous operations.

For example, in turning, if the proper tooling and feed and speed rates (all selected by the user) are not chosen, a terrible tangle of chips can form on the part. The average design engineer might say "hey, that's not my fault," whereas the good design engineer might say, "how can I design the system to prevent that type of failure mode from happening?" The answer in some cases is to use high-pressure coolant applied through a tiny orifice near the toolpoint to break off the chip as it forms. This type of system is known as a *hydraulic chipbreaker*. Hence the extreme value of asking users what the problems with existing *systems* are, and asking venders for system components what types of solutions they may have.

In production milling and turning, a common problem is modern tools and machines can remove metal so fast that the process can easily get buried in chips if active chip removal is not provided. Active chip removal includes the use of vacuums, coolant floods, and chip conveyors. These systems are often integrated with the machine geometry to enhance chip removal (e.g., a slanted or vertical bed, so the chips fall down into the conveyor). In modern integrated plants, a central coolant system is often used, where a flood of coolant washes chips (and heat) from the machining process into a central trough to carry them to a central point for separation and collection. In systems with self-contained coolant systems, care must be taken to filter out the chips and monitor the condition the coolant.

7.6.3 Fixturing Systems

All machines must perform some function on parts, and thus some method must be provided to hold the parts. Often the fixturing system is provided by a third-party vendor or the user himself. However, the design engineer must take care to ensure that the machine can interface to different systems,[130] and in the case of fixturing systems that use pneumatic or hydraulic power, investigate the possibility of providing fluid power lines to the machine's axis through cable carriers in the machine. At the entrance to the cable carrier and the exit on the axis, manifolds can be provided for the user to tie into his fluid power and fixturing systems accordingly. Not everyone will want this option, but it is easy to design the machine for the next size larger cable carrier to accommodate extra lines.

Other than providing standard tee slots or pallet-changing capability, sometimes there is little else the design engineer can do to prepare himself for fixturing challenges other than continually observing what systems are in use and what future systems are being proposed. Whenever possible, try to design insensitivity into the system. For example, the design case study of the coordinate measuring machine in Section 1.6 discussed how the machine's accuracy was made insensitive to how much the part weighed.

7.6.4 Fluid Systems

Fluid systems include those used to deliver power, lubrication, and coolant. They provide the lifeblood for the machine, and are as important as your own body's blood and blood vessels. Accordingly, they should be designed with the same attention that you would give your own body. This includes attention to layout (you would not want your aorta running across your face) and attention to specifying a maintenance schedule (you drink fluids and go to the bathroom on a regular basis).

Hydraulic and Pneumatic Power Systems

Fluid power systems can deliver higher power densities than any other type of system. Accordingly, extreme care must be taken to route their lines, as a broken line can cause damage to man

[130] See, for example, F. Wilson and J. Holt Jr. (eds.), Handbook of Fixture Design, McGraw-Hill Book Co., New York, 1962.

and machine. Neatness is a key to performance, and the design engineer should carefully consider how the lines will be run through the machine, how they will be anchored, and how they will be routed to moving axes. Careful use of swivel joints and cable carriers is a must. Hoses that must be interfaced to a moving linear axis should usually be constrained inside a flexible cable carrier. This helps prevent the hoses from becoming abraded, tangled, or damaged. The exception is when the forces imposed on the carriage by the cable carrier are too great such as might be the case for a machine with submicron accuracy.

Lubrication Systems

Bearings must often be lubricated either continuously or intermittently. Hydrostatic bearings require routing of high-pressure fluid lines for continuous lubrication, and sliding bearing materials (e.g., PTFE or cast iron on hardened steel) require a periodic (once every several minutes to once a day) squirt of lubricant. It is up to the design engineer to determine what the bearings need, and then design an automatic or manual system accordingly. In some instances, the lubrication system also serves as the coolant system. For example, an oil mist spray is often used to lubricate and cool spindle bearings.

Coolant Systems

Coolant systems almost always must be provided for material removal machines, and many precision machines require a cooling system to actively control the machine's temperature. In the former case, the coolant is in an open system exposed to the air and is thus subject to degradation. In the latter case, temperature control systems can often be sealed and can thus be protected through the use of chemicals. Care must be taken to provide corrosion protection for all exposed metal parts. Where water-based systems are used, the design engineer should realize that the humidity levels around the machine are likely to be higher and thus all unprotected metal surfaces are subject to corrosion.

Coolants for cutting process are often water based and are thus prone to becoming *stinky*. There is nothing worse than stinky coolant to make a machine operator's job unpleasant. How can the design engineer help? Even when coolant is changed, there are nooks and crannies where stinky coolant hides, and unless all the stinky coolant is removed, the bacteria which cause bad odors to form will quickly multiply. Hence regions that are awash with coolant should be smooth with rounded contours. For example, a spherical tank is far less susceptible than a rectangular tank to the buildup of stinky sludge. Coolant should enter the machining area and leave as soon as possible. In addition to helping to quickly remove chips and prevent stinky buildup, this will also help to quickly remove heat from the machine. Accordingly, for a precision machine, the design engineer may want to pay attention to how coolant lines can act as heat sources and insulate them accordingly. In the owner's manual for the machine, the design engineer should also stress the importance of a coolant management plan, and recommend that stinky-resistant coolants be used. The less often the coolant needs to be changed, the less the strain on the user's wallet and on the environment. In many areas, coolant from machining processes is now classified as a hazardous waste.

The design engineer must also consider how direct (splashing) and indirect (increased humidity or hydrocarbon vapors) exposure to coolants will affect the machine components. Many precision systems, such as grinders and diamond turning machines, operate under a flood of coolant. Hydrocarbon vapors and humidity can change the speed of light; hence if machines with coolant systems also use laser interferometers, it may be wise to enclose the interferometers in evacuated or helium-filled passageways. Special passageways require more room in the machine, however, and often increase the length to travel ratio in a slideway design. Support systems can indeed have a big effect on the design of the entire machine system.

7.6.5 Safety Systems

One problem with safety systems is that it is difficult to write anything about them because of liability issues. Hence in this section, only some basic philosophies will be discussed, and the reader

is referred to government and company policy literature.[131] This statement in itself should make the design engineer aware of their importance. There are three overall design-for-safety rules:

- Design safety into the machine itself. Do not design the machine and then figure out how to make it safe.
- Design the machine as if your life depended on it, lest as punishment for poor design you might be forced to work on the machine for the rest of your life.
- Remember that people are generally careless with regard to safety and they will often push a machine to see how far they can get away with something.

As an anecdote worth remembering regarding safety, recall the case in the early 1980s where two drunk men decided to pick up a lawn mower and use it to trim some bushes. The men lost their fingers and were awarded several million dollars because it was said the lawn mower manufacturer should have made the mower so that this could not happen. Now lawn mowers are supposed to have a clutch installed that stops power from being transmitted to the blade if the user takes his hands off the mower handle. But wait, what if the user uses wraps tape around the handle to hold the clutch lever? It seems that the manufacturer should install a vision system to watch the operator. Who cares if the mower that used to cost $200 may then costs tens of thousands of dollars! Does this seem ridiculous and stupid? Well it is not half as ridiculous or stupid as the average user often acts with regard to safety. On the other hand, the design engineer should not use this anecdote as a means to ridicule safety systems.[132]

Up until the mid-1900s, safety systems were ignored on many machines, and many careful, intelligent operators lost life and limb. The design engineer must realize that people are human, and like all living things, even people deserve to be treated with compassion and respect. So if nothing else, design your machine so that you would feel comfortable with someone you love operating it.

7.6.6 Sealing Systems[133]

One of the deadliest assassins of machines is dirt. Dirt can be generated by people working in the environment, by chips from the cutting process, and by wear of tools and components on the machine. The extent to which a dirt source can damage a machine depends on the sensitivity of the machine, how well it is sealed, and the amount and type of dirt to which it is exposed. Unless dirt is kept away from mechanical contact bearings and out of fluid and pneumatic systems, enhanced wear will result that can lead to premature loss of accuracy and machine failure. Similarly, if a machine generates dirt (e.g., wear particles or oil dripping), it could contaminate the manufacturing process (e.g., from food processing equipment to machines for fabricating integrated circuits)

Fortunately, sealing systems exist for virtually any type of application; all that is required is careful evaluation of the machine and working environment. There are thousands of seal options ranging from common O-ring and wiper-type seals for high-pressure sealing, to labyrinth seals, bellows and wipers to keep chips and dirt out of bearings. Because there are so many types of seals and seal materials, the design engineer should always consult with a seal manufacturer for their advice on the details of a seal design. This does not, however, preclude the design engineer from laying out the design for a machine using assumed space requirements for sealing systems.

A typical machine tool with a several-micron resolution of motion is generally not concerned with seal friction. This is due to the fact that even a ballscrew-driven slide supported by sliding contact bearings can easily overcome stiction effects encountered with wiper-type seals. Machines with submicron resolution or high speed/acceleration capabilities, on the other hand, require that friction be minimized, and thus these systems are sealed with bellows or labyrinth seals that are under slight internal pressure to keep dirt out. Other types of sealing are also often required. For example, laser interferometer beam paths often need to be enclosed in evacuated tubes or bellows to prevent the refractive index along the beam path from varying.

---

131 Recall that safety issues were discussed in Chapter 1. A design engineer must also consult the Occupational Safety and Health Administration (OSHA) to obtain the latest regulations regarding the type of equipment he is designing.

132 "*As an arbiter in litigation I am no better than other men. But what I do strive for is that there shall be no litigation at all.*" Confucius

133 A good reference to consult is L. Martini (ed.), Practical Seal Design, Marcel Deckker, New York, 1986. Also see Machine Design's annual Mechanical Drives and Fluid Power reference issues. An indispensable comprehensive industrial reference is Parker Packing Engineering Handbook available from Parker Packing, P.O. Box 30505, Salt Lake City, UT 84125.

Dynamic Seals for Round Parts

Round parts can undergo rotary motion (e.g., bearings and shafts) or linear motion (e.g., piston shafts or round shafts that act as rails for linear bearings). Round parts are easy to seal because the seals themselves are so easily made to fit the part exactly; in addition, it is easy to design wipers to scrape dirt off the round part and prevent it from damaging the seal (e.g., dirt off a hydraulic cylinder rod on a piece of construction equipment).

Typical rotary motion seals for keeping dirt out are shown in Figure 7.6.2. Labyrinth seals are used where the environment is relatively clean and friction must be minimized. Contact-type seals can keep bearings operating even in environments as harsh as the front axle of your car. When the seal is required to hold high pressure, a lip or a squeeze seal is commonly used, as shown in Figure 7.6.3. Lip-type seals work well at high differential pressures, but can leak at low differential pressures, because they rely on the pressure to force the seal against the surfaces; on the other hand, lip-type seals have the lowest coefficient of friction of any type of pressure seal. A squeeze-type seal has an initial preload that seals at low as well as high differential pressures, but as a result also has a higher coefficient of sealing friction. Both of these types of seals, however, can usually work on pistons that undergo axial and rotary motion.

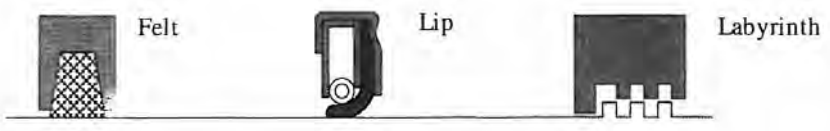


Figure 7.6.2 Common rotary seals.

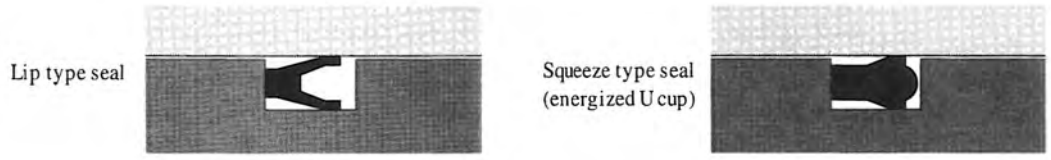


Figure 7.6.3 Lip and squeeze type seals for rotating and/or reciprocating round shafts.

High-pressure sealing depends on precise amount of contact pressure between the seal lip and the shaft. In a properly designed and installed seal, the lip rides on a film of lubricant. The film does the actual sealing. Therefore, the film thickness (approximately 25 μm) must be precisely controlled by the mechanical pressure of seal element and shaft surface finish. If the film is too thick, fluid leaks. If film is too thin, the lip wears, which increases friction and may result in slip-stick oscillations. Stick-slip oscillations in turn cause surface waves to form in the seal, which permits leakage. The main factors affecting film thickness are mechanical pressure, sealed pressure, and temperature. In general, as sealed pressure increases, slip contact pressure increases, and the film thickness decreases. Increasing temperature, which may be caused by increasing shaft speed, reduces fluid viscosity, which also causes the film thickness to decrease.

Shaft conditions have a significant effect on sealing. Shafts should be hardened to at least Rockwell C 30 and have an R_a surface finish on the order of $^1/_4$-$^1/_2$ μm (10-20 μin.). Finishes finer than $^1/_4$ μm generally show no improvement in seal life, and too fine a finish can shorten seal life because the shaft surface cannot support a thin lubricating oil film. In order to prevent the film of oil becoming exposed to the outside environment, a wiper seal is also often used. In order to prevent dirt on the outside from being drawn back into the seals on a reciprocating shaft, a scraper is often used.

Dynamic Seals for Nonround Parts

Unfortunately, all parts are not round, and sealing nonround parts can be difficult. The difficulty arises in the corners that invariably accompany nonround parts. These corners are usually where geometric mismatch between seal and part occur, and these regions act as points for dirt

intrusion and enhanced seal wear. In order to prevent dirt from getting into linear bearings or leadscrews, one should use a wiper seal on the bearing or leadscrew, and then encase the entire system in a bellows or sliding way cover. These devices are especially important when recirculating balls are used in the bearings or leadscrew. Note that sliding contact bearings or acme threads are less susceptible to the grinding action of dirt particles because the lubrication systems used with these components helps to flush them out; thus, for example, sliding bearings often just use a wiper seal.

Way Covers

Bellows and way covers are surprisingly inexpensive, even for custom applications; thus there is no excuse for not using them. Bellows-type seals are used where friction is to be minimized, but chips can collect in the folds and can cause the bellows to tear. Bellows are typically made of an elastomer, although metal bellows are commonly used for vacuum applications. The former are generally custom made without mold charges. Generally, the convolutions are 3-5 mm thick when folded (slope is 90 ^{o}). As shown in Figure 7.6.4, the slope between the ID and OD should be at most 45^{o}; hence a convolution with a 50 mm ID and a 75 mm OD will be typically 4 mm thick compressed and 54 mm when extended. Virtually any cross section can be specified to match the shape of round, square, triangular, or any combination. For long span bellows, plates with bearing surfaces can be incorporated into the convolutions at regular intervals to keep them from dragging and snagging.

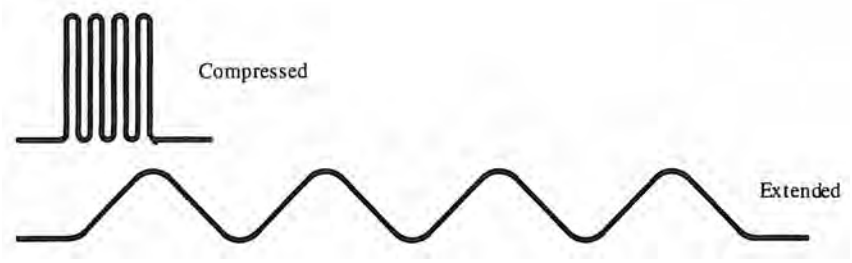


Figure 7.6.4 Compressed and extended bellows convolutions.

Related to bellows are diaphragms. Flat and convoluted diaphragms are used where there is relatively little motion. Rolling diaphragms are used for relatively long motions. Rolling diaphragms should not be used where subject to pressure reversals because reversal may force the diaphragm out of rolling convolution and shorten its life. Special diaphragms that accept pressure reversals without sidewall distortion are available.

Sliding way covers are made of metal or hard plastic. Sliding way covers are very rugged, but there is friction between the segments caused by sliding seals (of little consequence on most machine tools but they can take up considerable space on large machines). To allow for the thickness of the metal and the wiper seals, allow about 6-8 mm per section. Recall from the design case study in Section 1.5 that sealing considerations can often make or break a design.

Windowshade-type seals are wound on a spring-loaded drum, and although they take up little room, they do not seal as well as the other types. Every design engineer should familiarize himself with sealing manufacturers' products.

Static Seals

Static seals are used commonly to prevent flow from a high-pressure region to a lower pressure region across a joint between components. Static seals are often encountered in fittings for pressurized hose (e.g., pneumatic or hydraulic lines), or between manifolds (e.g., between a valve and a pressure distribution manifold).

Hose fittings often use tapered pipe threads, which seal by virtue of the fact that as the fitting is tightened, contact between all surfaces of the threads ideally form, hence forming a seal. There are always small micropassages, however, so a moldable medium (i.e., pipe dope or Teflon tape wrapped on the threads) is required to fill the voids. The principal problem with pipe threads is that they often leak, and assembly personnel in their quest to eliminate leaks sometimes tighten them so tight that it becomes impossible to remove the fitting. A better alternative to pipe threads is the JIC fitting, which uses a straight thread with an O ring around the flange of the fitting that seats on the

surface of the part the fitting screws into. The fitting thus only has to be tightened enough to seat the O ring. Tapered pipe threads are cheaper than JIC fittings, but tapered pipe threads should only be used for pneumatic lines. JIC fittings should be used for high-pressure hydraulic lines, or where ease of assembly and assurance of quality is desired. Note that JIC fittings require a counterbore for the O ring to seat in. An alternative sealing method is to use a straight thread fitting and a bonded seal. A bonded seal is essentially a washer with a rubber seal bonded to the inside diameter. The rubber makes a seal between the underside of the fitting and the surface of the part. The metal washer keeps the pressure from blowing the seal out.

In order to form a seal between two parts that are bolted together, one can cut an O-ring groove into one of the parts, or one can use a gasket. When high-pressure fluid encounters an O-ring seal, it causes the seal to flatten out and push even harder on the bolted faces of the assembly. Thus an O ring operates on the principle of self-help. The depth of the O-ring groove is thus less than the cross-sectional diameter of the O ring, and the width larger. This gives room for the tightened joint to compress the O ring, and for fluid to push against the entire cross section of the O ring. There is always some space between some areas of mating parts, and the O ring can be extruded into this region and fail if the gap or the pressure are too great. Thus for high pressures, defined by seal manufacturers for their products, a T seal with backup rings can be used. Regardless of the seal used, it is very important to follow manufacturers' guidelines for dimensioning of seal grooves.

Gaskets seal by compressing a deformable material between the two surfaces. The material takes the shape of the surfaces and thus compensates for any flatness mismatch between the surfaces. However, the pressure of the fluid tries to pry apart the surfaces, and thus gasketed joints require very high preloads to prevent leakage.

An alternative to O rings and gaskets is to use a sealant (e.g., silicone rubber) between mating surfaces. Components are clamped together when the sealant is still fluid. Metal-to-metal contact very nearly occurs at high spots, and sealant fills voids in other areas. The result is a very thin gasket that is almost perfectly matched to surface irregularities on flanges.

7.6.7 Tooling Systems

The design engineer must consider the types of tooling that will be used and what are the required force, power, and feed rates needed for proper operation of the tooling. Requirements for rapid material removal (e.g., hogging) are quite different than for precision finishing (e.g., diamond turning). For example, forces from hogging operations can exceed hundreds and even thousands of newtons while forces from diamond turning may be less than 1 N.[134] Often, tooling standards will dictate the design of tool holding turrets and tool changers, and location of coolant jets. For example, hydraulic chipbreakers require high-pressure coolant (5-10 MPa). The fluid travels through the tool and is sprayed at the toolpoint, thereby "shooting off" the chip as it forms.

7.6.8 Vibration Control

Internal and external sources of vibration can degrade machine performance as discussed in Section 2.4.1. In Sections 7.3 and 7.4 materials and methods were discussed for substantially increasing the damping factor of a machine. In this section, sound control, vibration isolation, and shock control are discussed. In order to appreciate the effects that these vibration sources have on the machine, one must carefully generate a dynamic model of the machine tool structure, which is unfortunately beyond the scope of this book.[135]

Sound Control[136]

For precision machines where nanometer resolution is desired, sound energy on the order of a person's voice can be detrimental. For a typical machine tool, it is desirable to control sound so

[134] J. Drescher and T. Dow, "Measurement of Tool Forces in Diamond Turning," 5th Int. Precis. Eng. Symp. (IPES), Monterey CA, Sept. 1989.

[135] See, for example, S. Timoshenko et al., Vibration Problems in Engineering, John Wiley & Sons, New York, 1974, and J. Tuzicka and T. Derby, Influence of Damping in Vibration Isolation, Technical Information Division, Naval Research Laboratory 1971, Library of Congress No. 71-611802. Also see the annual International Conference on Vibration Control in Optics and Metrology, sponsored by the SPIE.

[136] For a more in-depth discussion of sound control with respect to machine tools, see M. Weck, Handbook of Machine Tools, Vol. 2, John Wiley & Sons, New York, 1984, pp 62-73.

as to make the machine more pleasant to operate. One unpleasant characteristic of most machine tools is the operation of dc brushless motors using a pulse-width-modulated (PWM) power signal. By sending the motor a high-frequency input as opposed to a dc signal, much greater electrical efficiency is obtained; unfortunately, most PWM drivers operate at 3 kHz, which creates a high-pitched whine that is most unpleasurable. Advances in switching power transistors will hopefully soon push this frequency up over 20 kHz, so it will be in the inaudible range. In either case, the noise is generated by vibration, which can in turn show up on the surface of a part that is being machined to the nanometer level if the machine is not properly designed (i.e., it does not have a very rigid, well-damped, tight structural loop). In addition, one should note that an electric motor, which contains many coils of wire, essentially acts like a transmitter. Thus when high-frequency, high-current waveforms are passed through the coil of wire, it generates substantial electromagnetic interference.

In order to control sound from internal or external sources, surfaces that reflect sound must be covered with sound-absorbing material. There are a number of commercially available adhesive-backed sheets of material that one merely has to stick onto surfaces to make a room or machine sound dead.

Vibration Isolation

The world around us is filled with vibration. Vibration is generated in a structure by people walking in the hallway, big trucks driving by, and by fans and compressors in a building's environmental control system. Vibration is also generated by rotating components within a machine or by PWM-driven motors as discussed above. The best way to isolate a precision machine from vibration is first to try to reduce the level of vibration at the source. For example, all rotating components such as motors should be dynamically balanced, and equipment in mechanical equipment rooms should be isolated from the building.[137] This is the responsibility of the building design engineer, but the machine or facilities design engineer should be careful to ascertain the properties of the systems in a building.

The next step, ideally, is to sit the machine on its own concrete slab that is isolated from the rest of the world by a series of springs and dampers (e.g., a viscoelastic material). Each mass/spring/damper set acts as a second-order system with 40 dB/decade attenuation capability. Two sets, one on top of the other, provide 80 dB/decade attenuation. It is possible to obtain virtually any degree of isolation at any cutoff frequency that one desires. However, one must be careful of distributed mass effects in mechanical springs introducing resonances at higher frequencies. Hence air bags or viscoelastic materials are often used as springs. With the former, damping is provided by connecting the air bag to a reservoir via an orifice (typically, the reservoir should have eight times the volume of the air bag).[138]

Next the design engineer should make sure that the machine itself is in contact with other structures only through viscoelastic materials, the one allowable exception being wires from sensors and to motors. One should make sure that this guideline is strictly adhered to. Even a mundane metal wire conduit can serve as a vibration path into a machine from the outside world. There are a number of viscoelastic materials that one can use to absorb vibration energy; however, because they are viscoelastic, one should not rely on them for dimensional stability without specifically requesting appropriate data from the manufacturer. Since these materials are polymer based, it seems like a new one is being developed every month. Hence one can only keep abreast of latest advancements through continual reading of design journals.

Shock Control

Machines with reciprocating components can generate shock loads, or a machine that is being shipped to customer's plant can be subject to shock loads. In either case, it is important to ascertain what the probable load will be and then design an appropriate mount for the component or the machine in its shipping container.[139] Shock control devices can range from using foam rubber, Styrofoam, or viscoelastic material pads to piston-type shock absorbers.

---

[137] See, for example, A. Campanella, "Vibration Isolation Criteria for Elevated Mechanical Equipment Rooms," Sound Vibrat., Oct. 1987.

[138] D. Debra, "Overview of Vibration Isolation Techniques", 5th Int. Precis. Eng. Symp. (IPES), Monterey CA, Sept. 1989. Also see T. Takagami and Y. Jimbo, "Study of an Active Vibration Isolation System," Precis. Eng., Vol. 10, No. 1, 1988, pp. 3–7.

[139] See, for example, J. Arimond, "Shock Control", Mach. Des., May 21, 1987.

7.7 KINEMATIC COUPLING DESIGN[140]

Kinematic couplings have long been known to provide an economical and dependable method for attaining high repeatability in fixtures.[141] Properly designed kinematic couplings are deterministic: They only make contact at a number of points equal to the number of degrees of freedom that are to be restrained. Being deterministic makes performance predictable and also helps to reduce design and manufacturing costs.[142] On the other hand, contact stresses in kinematic couplings are often very high and no elastohydrodynamic lubrication layer exists between the elements that are in point contact; thus for high-cycle applications, it is advantageous to have the contact surfaces made from corrosion-resistant materials (e.g., ceramics). When non-stainless steel components are used, one must be wary of fretting at the contact interfaces, so steel couplings should only be used for low-cycle applications.

Tests on a heavily loaded (80% of allowable contact stress) steel ball/steel groove system have shown that microinch repeatability could be attained;[143] however, with every cycle of use, the repeatability worsened until an overall repeatability on the order of ten μm was reached after several hundred cycles. At this point, fret marks were observed at the contact points. Tests on a heavily loaded (80% of allowable contact stress) silicon nitride/steel groove system have shown that 50 nm repeatability could be attained over a range of a few dozen cycles, and that with continued use the overall repeatability asymptotically approached the surface finish of the grooves (on the order of $^1/_3$ μm R_a). An examination of the contact points showed an effect akin to burnishing, but once the coupling had worn in, submicron and better repeatability was ultimately obtained. Unfortunately, references were not found in the literature to make an extensive comparison of the effects of load and surface finish on kinematic coupling repeatability.

The tests that were reported also showed that with the use of polished corrosion-resistant (preferably ceramic) surfaces, a heavily loaded kinematic coupling can easily achieve submicron repeatability with little or no wear-in required. Regretfully, too many designers still consider kinematic couplings to be useful only for instrument or metrology applications.

7.7.1 Coupling Configuration and Stability

Symmetry aids in reducing manufacturing costs, and for practical fixturing applications, in general, the use of grooves for all contact regions minimizes the overall stress state in the coupling. Thus it is assumed here that the kinematic coupling to be designed is a three-groove type.

Two forms of three-groove couplings are illustrated in Figure 7.7.1. Planar couplings are often found in metrology applications. They can also be used in the manufacture of precision parts. For example, a planar three-groove coupling can be used to hold a grinding fixture on a profile grinder. A matching three-groove plate on a CMM allows the grinding fixture to be transferred to the CMM with the part. The part can be measured and then placed back onto the grinder so the errors can be corrected. To minimize Abbe errors in some applications, vertically oriented couplings can be designed where the preload is obtained with a clamping mechanism or by gravity acting on a mass held by a cantilevered arm. An example would be a three-groove kinematic coupling used to hold photolithographic masks in a wafer stepper whose projection axis must be horizontal because of its size.

With three grooves, the question naturally arises as to what is the best orientation for the grooves. Mathematically, to guarantee that the coupling will be stable, James Clerk Maxwell stated the following:[144]

---

[140] Also see A. Slocum, "Kinematic Couplings for Precision Fixturing - Part 1: Formulation of Design Parameters," Precis. Eng., Vol. 10, No. 2, 1988, pp. 85–91; and A. Slocum and A. Donmez, "Kinematic Couplings for Precision Fixturing - Part 2: Experimental Determination of Repeatability and Stiffness," Precis. Eng., Vol. 10, No. 3, 1988, pp. 115–122.

[141] C. Evans, Precision Engineering: an Evolutionary View, Cranfield University Press, Cranfield England, 1989, pp 21-29.

[142] A. Slocum, "Kinematic Couplings for Precision Fixturing - Part I - Formulation of Design Parameters," Precis. Eng., Vol. 10, No. 2, April 1988, pp. 85–91.

[143] A. Slocum and A. Donmez, "Kinematic Couplings for Precision Fixturing - Part II - Experimental Determination of Repeatability and Stiffness," Precis. Eng., Vol. 10, No. 3, July 1988, pp. 115–121.

[144] J. C. Maxwell, "General Considerations Concerning Scientific Apparatus," in The Scientific Papers of J. C. Maxwell, Vol. II, W.D. Niven (ed), Cambridge University Press, London, 1890, pp. 507–508.

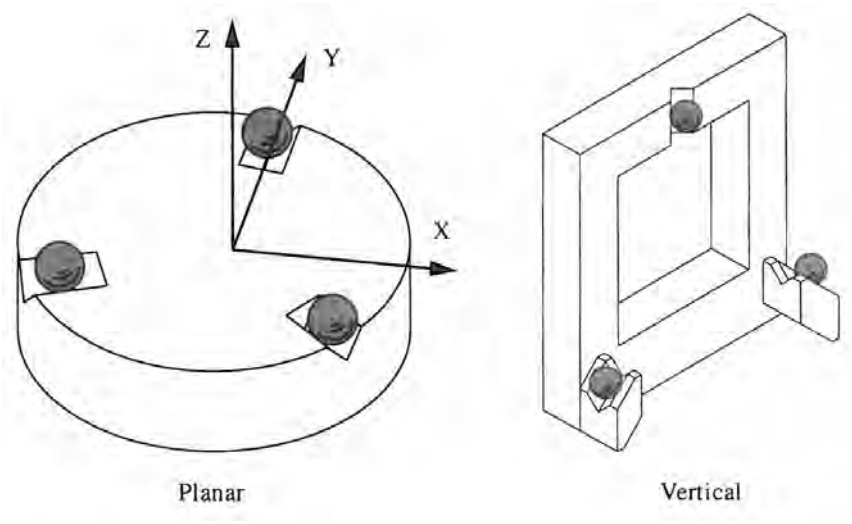


Figure 7.7.1 Examples of three-groove kinematic couplings for horizontal and vertical fixturing applications. For clarity, the coupling components to which the balls are permanently affixed are not shown.

> "When an instrument is intended to stand in a definite position on a fixed base it must have six bearings, so arranged that if one of the bearings were removed the direction in which the corresponding point of the instrument that would be left free to move by the other bearings must be as nearly as possible normal to the tangent plane at the bearing.
>
> (This condition implies that, of the normals to the tangent planes at the bearings, no two coincide; no three are in one plane, and either meet in a point or are parallel; no four are in one plane, or meet in a point, or are parallel, or, more generally, belong to the same system of generators of an hyperboloid of one sheet. The conditions for five normals and for six are more complicated.)"

In a footnote to this discussion, Maxwell references Sir Robert Ball's pioneering work in *screw theory*. Screw theory asserts that the motion of any system can be represented by a combination of a finite number of screws of varying pitch that are connected in a particular manner. This concept is well illustrated for a plethora of mechanisms by Phillips.[145] Ball's work on screws spanned the latter half of the 19th century and a detailed summary of his work on screw theory was published in 1900.[146] Ball's treatise describes the theory of screws in elegant, yet easily comprehensible linguistic and mathematical terms. Currently, research in automation is attempting to use screw theory to determine what is the best way to grasp an object (e.g., with a robotic hand) or to fixture a part (e.g., for automated fixture design for manufacturing).[147]

Screw theory is an elegant and powerful tool for analyzing the motion of rigid bodies in contact, but it is not always easy to apply. With respect to practical implementation of the theoretical requirement for stability, for precision three-groove kinematic couplings, stability, and good overall stiffness will be obtained if the normals to the plane of the contact force vectors bisect the angles of the triangle formed by the hemispheres (e.g., balls) that lie in the grooves, as shown in Figure 7.7.2.[148]

Furthermore, for balanced stiffness in all directions, the contact force vectors should intersect the plane of coupling action at an angle of 45 degrees. Note that the angle bisectors intersect at a

[145] J. Phillips, Freedom in Machinery, Vol. I, Cambridge University Press, London, 1982, p 90.

[146] R. S. Ball, A Treatise on the Theory of Screws, Cambridge University Press, London, 1900.

[147] J. J. Bausch and K. Youcef-Toumi, "Kinematic Methods for Automated Fixture Reconfiguration Planning," IEEE Conference on Robotics and Automation, 1990, pp. 1396–1491. This reference summarizes work done by John Bausch for his Ph.D. thesis in the Mechanical Engineering Department at MIT. Dr. Slocum, who was a member of John's thesis committee, first suggested to John that screw theory would provide a good theoretical method for studying the problem of automated fixture design.

[148] From conversations and observations with Dr. William Plummer, Director of Optical Engineering, Polaroid Corp, 38 Henry Street, Cambridge, MA 02139.

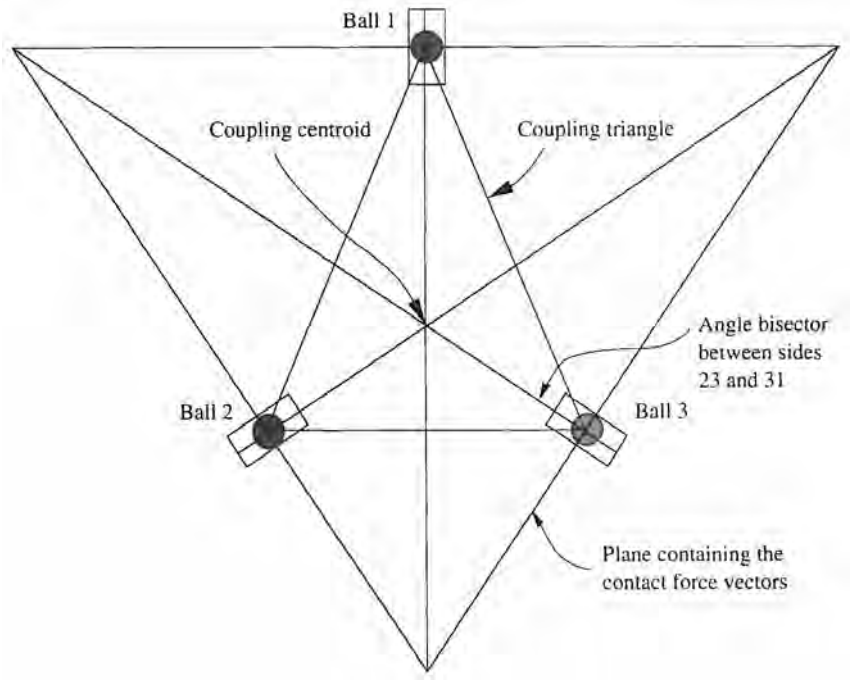


Figure 7.7.2 For good stability in a three-groove kinematic coupling, the normals to the planes containing the contact force vector pairs should bisect the angles between the balls.

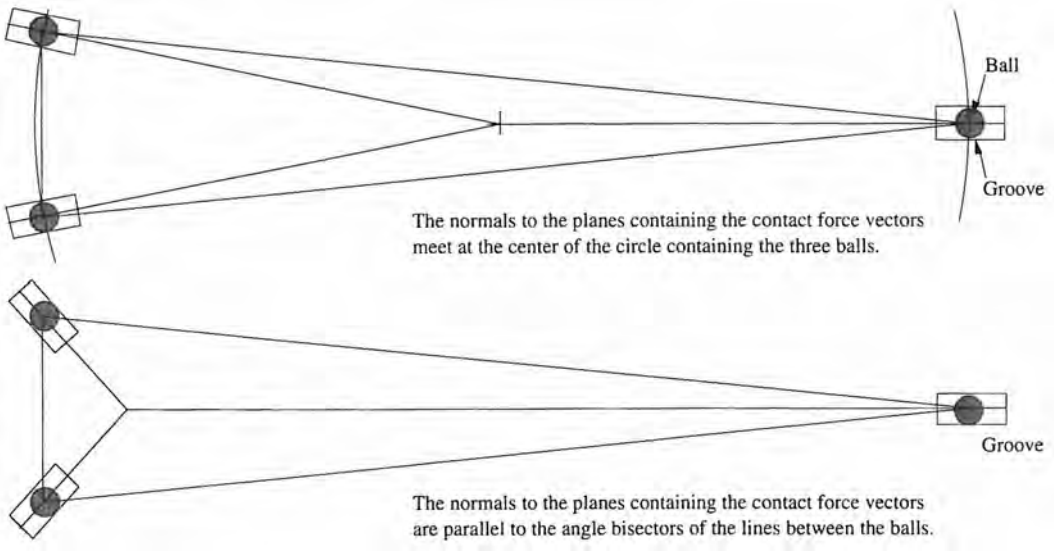


Figure 7.7.3 Consider the design of a long coupling to locate a laser head on an instrument. Compare the stability of couplings designed by two methods that give the same solution for a coupling where the balls lie on the vertices of an equilateral triangle.

point that is also the center of the circle that can be inscribed in the coupling triangle. This point is referred to as the *coupling centroid* and it is only coincident with the coupling triangle's centroid when the coupling triangle is an equilateral triangle. Fortunately designers of precision kinematic couplings are not faced with the generic grasp-a-potato problem faced by researchers in robotics.

For a coupling where the balls lie on the vertices of an equilateral triangle, the angle bisectors also intersect at the triangle's centroid. If the normals to the planes containing the contact forces' vectors were to always point towards the coupling triangle's centroid instead of along its angle bisectors, then the coupling's stiffness will decrease as the coupling triangle's aspect ratio increases. This concept is illustrated in Figure 7.7.3. Most coupling designs seek to obtain good stiffness in all directions; however, in some cases it may be desirable to maximize the stiffness in a particular direction.

Note that any three-groove kinematic coupling's stability can be quickly assessed by examining the intersections of the planes that contain the contact force vectors. For stability, the planes must form a triangle as illustrated in Figure 7.7.4.

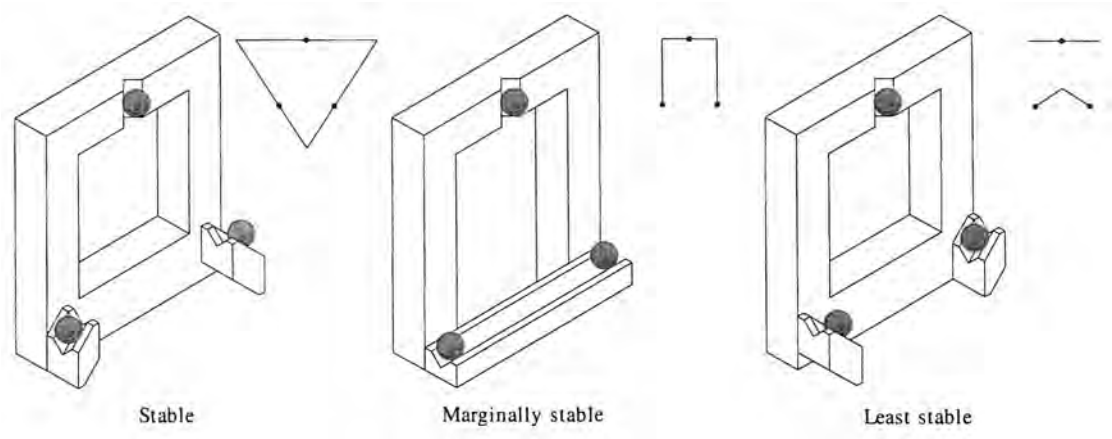


Figure 7.7.4 Different configurations for a kinematic coupling that illustrate how the intersections of the planes containing the contact force vectors can be used to make an assessment of the coupling's stability.

7.7.2 Analysis of Three-Groove Couplings

Figure 7.7.5 illustrates the information needed to characterize a three-groove kinematic coupling. To design a three-groove kinematic coupling, the designer must provide the following information:

- The balls' diameters and the grooves' radii of curvature.
- The coordinates x_{Bi}, y_{Bi}, and z_{Bi} of the contact points of the balls in the grooves.
- The contact forces' direction cosines α_{Bi}, β_{Bi}, and γ_{Bi}.
- The coordinates x_{Pi}, y_{Pi}, and z_{Pi} of up to three preload forces.
- The x-, y-, and z-direction preload forces' magnitudes $F_{P\xi i}$ at each of the three points.
- The coordinates x_L, y_L, and z_L of an external applied load (the effect of more loads can be evaluated using superposition).
- The x-, y-, and z-direction magnitudes $F_{L\xi}$f the externally applied load.
- The moduli of elasticity and Poisson ratios of the ball and groove materials.

The output from the analysis will be:

- The contact forces (F_{Bi})
- The contact stresses
- The deflections at the contact points
- The six error motion terms (δ_x, δ_y, δ_z, ε_x, ε_y, ε_z) that exist at the coupling's centroid

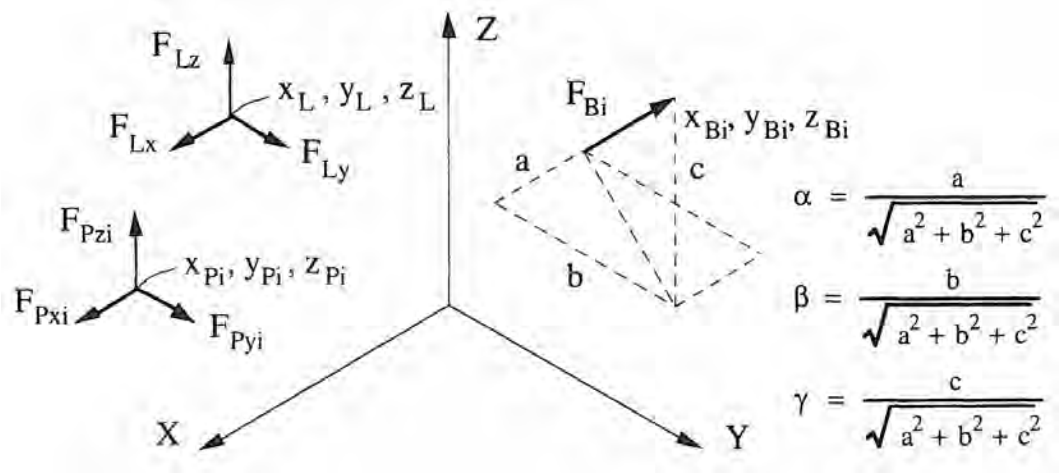


Figure 7.7.5 Information required to define a three-groove kinematic coupling.

Force and Moment Equilibrium

The force and moment balance equations for the system are

$$\sum_{i=1}^{6} F_{Bi}\alpha_{Bi} + \sum_{i=1}^{3} F_{Pxi} + F_{Lx} = 0 \tag{7.7.1}$$

$$\sum_{i=1}^{6} F_{Bi}\beta_{Bi} + \sum_{i=1}^{3} F_{Pyi} + F_{Ly} = 0 \tag{7.7.2}$$

$$\sum_{i=1}^{6} F_{Bi}\gamma_{Bi} + \sum_{i=1}^{3} F_{Pzi} + F_{Lz} = 0 \tag{7.7.3}$$

$$\sum_{i=1}^{6} F_{Bi}(-\beta_{Bi}z_{Bi} + \gamma_{Bi}y_{Bi}) + \sum_{i=1}^{3} -F_{Pyi}z_{Pi} + F_{Pzi}y_{Pi} - F_{Ly}z_L + F_{Lz}y_L = 0 \tag{7.7.4}$$

$$\sum_{i=1}^{6} F_{Bi}(\alpha_{Bi}z_{Bi} - \gamma_{Bi}x_{Bi}) + \sum_{i=1}^{3} F_{Pxi}z_{Pi} - F_{Pzi}x_{Pi} + F_{Lx}z_L - F_{Lz}z_L = 0 \tag{7.7.5}$$

$$\sum_{i=1}^{6} F_{Bi}(-\alpha_{Bi}y_{Bi} + \beta_{Bi}x_{Bi}) + \sum_{i=1}^{3} -F_{Pxi}y_{Pi} + F_{Pyi}x_{Pi} - F_{Lx}y_L + F_{Ly}x_L = 0 \tag{7.7.6}$$

The magnitudes of the six contact point forces are easily calculated using these equations and a spreadsheet. Once the magnitudes of the forces are known, they can be used to determine the stress and deflection at the contact points using Hertz theory as discussed in Section 5.6.

Elastic Characteristics of the Ball-Groove Interface

Hertz theory, as discussed in Section 5.6, can be used to evaluate the stress and stiffness characteristics of the contact interfaces in a kinematic coupling. The interface geometry is defined by the ball radius R_{ball}, and the major and minor radii of the groove R_{groove} = -(R_b[1 + γ]) and -∞s respectively. The radius of an equivalent ball on a flat plate is

$$R_e = \frac{R_{ball}(1+\gamma)}{(1+2\gamma)} \tag{7.7.7}$$

The effects of the ball/groove radius ratio γ on the contact stress and deflection are shown in Figure 7.7.6 based on calculations made using the gap-bending hypothesis (for an actual design analysis, one would use the exact form of Hertz theory). To avoid problems with contamination at the interface, γ should be as large as possible. To minimize the effect on contact stress and deflection, γ

should be as small as possible. If the effect on stress caused by increasing γ is kept at a 10% level, a good compromise is to let $\gamma = 0.20$.

The ratio of the preload to the applied load also has an effect on the stiffness of the system. If the cutting force F_c is a percentage ζ of the preload F_p ($F_C = \zeta F_P$), then the deflection varies as

$$\delta = C_{constant} F_P^{2/3} \left((1 + \xi)^{2/3} - 1 \right) \tag{7.7.8}$$

Figure 7.7.7 shows an effectively linear relation for this function over a range of expected values for ζ. As expected, the ratio of applied load to preload should be kept as small as possible. If the coupling is used in a fixture for precision machining operations, note that during finish cuts the cutting forces may only be on the order of Newtons.

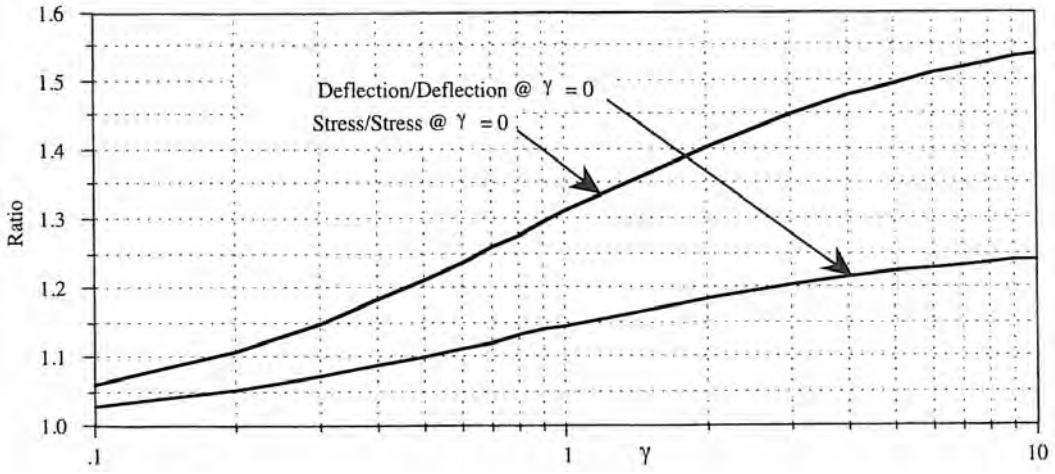


Figure 7.7.6 Effect of ball/groove radius ratio γ on stress and deflection.

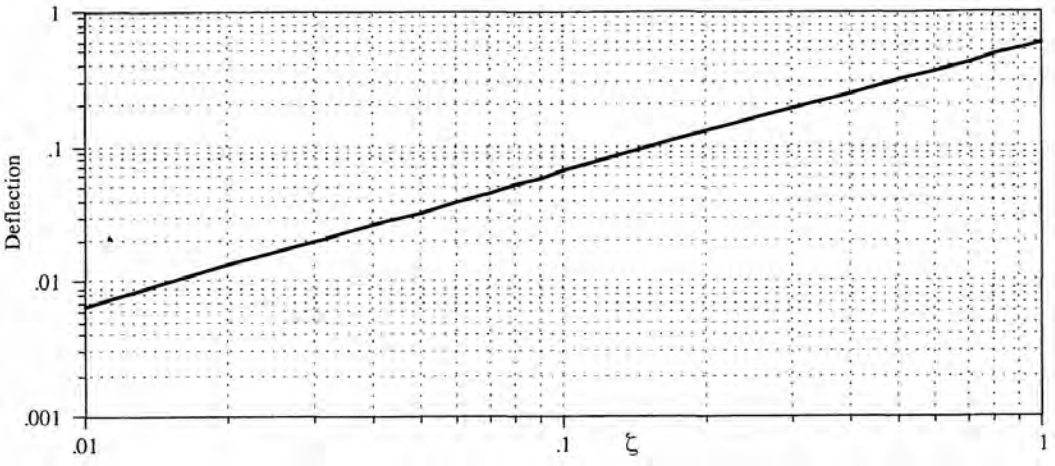


Figure 7.7.7 Effect on deflection of ratio ζ of applied force to preload ($F_c = \zeta F_p$).

Kinematics of the Coupling's Error Motions

Contact between the ball and the groove actually results in an elastic indentation of the region. Combined with a finite coefficient of friction, it is reasonable to assume that there is no relative motion between the ball and the groove at the contact interface. If one makes this assumption and then calculates the new position of the balls' centers using the contact displacements and contact forces' direction cosines, one finds that there is not a unique homogeneous transformation matrix that relates the old and new ball positions. These factors make the calculation of a kinematic coupling's error motions a non-deterministic problem.

Fortunately, if the distances between the balls, determined using their new coordinates, do not change greatly, then reasonable estimates can be made of the coupling's error motions. Using the design theory presented herein, a spreadsheet can be used to show that the change in distance between the balls is typically five to ten times less than the deflection at the contact points. Furthermore, the ratio of the change in the distance between the balls to the distance between the balls is typically an order of magnitude less than the ratio of the deflection of the ball to the ball diameter. Thus estimates of the coupling's error motions can be made in the following manner:

The product of the deflection of the balls with the contact forces' direction cosines are used to calculate the ball's deflections. The displacements of the coupling triangle's centroid, which are $\delta_{\xi c}$ (ξ = x, y, z), are assumed to be the equal to the weighted average (by the distances between the balls and the coupling centroid) of the ball's deflections:

$$\delta_{\xi c} = \left(\frac{\delta_{1\xi}}{L_{1c}} + \frac{\delta_{2\xi}}{L_{2c}} + \frac{\delta_{3\xi}}{L_{3c}} \right) \frac{L_{1c} + L_{2c} + L_{3c}}{3} \tag{7.7.9}$$

The rotations of the coupling about the x and y axes are conveniently determined for the case of a coupling whose grooves lie in the xy plane (other orientations confuse the angle definition in the spreadsheet analysis). To determine the rotations, the altitudes of the coupling triangle and its sides' orientation angles must be determined as shown in Figure 7.7.8. With these geometric calculations, the rotations about the x and y axes can be determined:

$$\varepsilon_x = \frac{\delta_{z1}}{L_{1,23}} \cos\theta_{23} + \frac{\delta_{z2}}{L_{2,31}} \cos\theta_{31} - \frac{\delta_{z3}}{L_{3,12}} \cos\theta_{12} \tag{7.7.10}$$

$$\varepsilon_y = \frac{\delta_{z1}}{L_{1,23}} \sin\theta_{23} + \frac{\delta_{z2}}{L_{2,31}} \sin\theta_{31} - \frac{\delta_{z3}}{L_{3,12}} \sin\theta_{12} \tag{7.7.11}$$

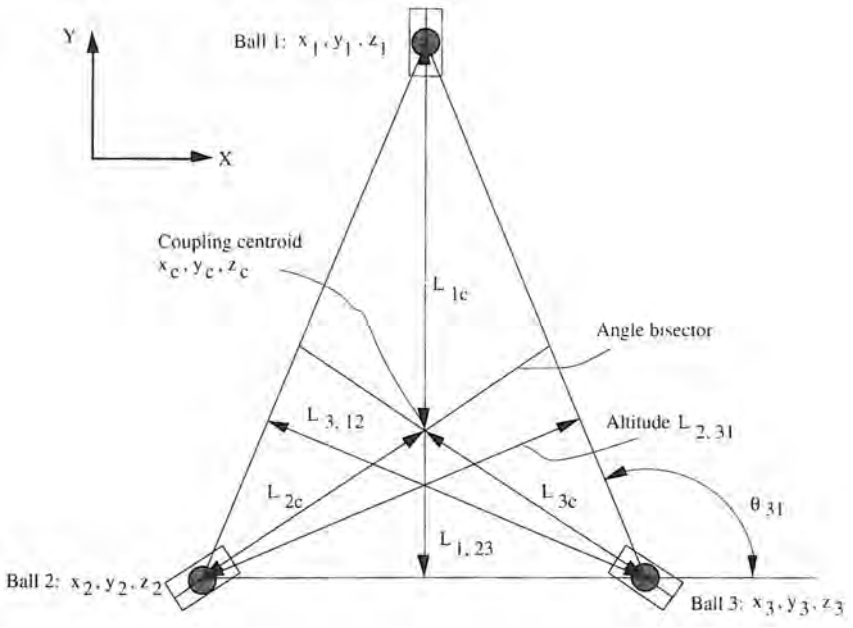


Figure 7.7.8 Geometry of a planar kinematic coupling.

The coupling's rotation about the z direction at the coupling centroid is assumed to be the average of the rotations calculated for each ball. For example, the rotation about the z direction at the coupling centroid caused by ball 1 is

$$\varepsilon_{z1} = \frac{\sqrt{(\alpha_{B1}\delta_1 + \alpha_{B2}\delta_2)^2 + (\beta_{B1}\delta_1 + \beta_{B2}\delta_2)^2}}{\sqrt{(x_1 - x_c)^2 + (y_1 - y_c)^2}} \text{SIGN}(\alpha_{B1}\delta_1 + \alpha_{B2}\delta_2) \tag{7.7.12}$$

The rotation error about the z axis of the coupling is assumed to be

$$\varepsilon_z = \frac{\varepsilon_{z1} + \varepsilon_{z2} + \varepsilon_{z3}}{3} \tag{7.7.13}$$

The errors can then be assembled into a homogeneous transformation matrix (HTM) for the coupling that allows for the determination of the translational errors δ_x, δ_y, and δ_z at any point x, y, z in space around the coupling:

$$\begin{bmatrix} \delta_x \\ \delta_y \\ \delta_z \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & -\varepsilon_z & \varepsilon_y & \delta_x \\ \varepsilon_z & 1 & -\varepsilon_x & \delta_y \\ -\varepsilon_y & \varepsilon_x & 1 & \delta_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x - x_c \\ y - y_c \\ z - z_c \\ 1 \end{bmatrix} - \begin{bmatrix} x - x_c \\ y - y_c \\ z - z_c \\ 0 \end{bmatrix} \tag{7.7.14}$$

In the HTM it has been assumed that the rotations are small, so small angle trigonometric approximations are valid. Also, the error motions had been calculated about the coupling triangle's centroid which may not be coincident with the coordinate system's origin; hence the centroid coordinates are subtracted from the location at which the errors are to be determined.

7.7.3 Practical Design Considerations

With a spreadsheet, the design engineer can easily play "what-if" design games to arrive at a theoretically workable kinematic coupling for virtually any application. However, the problem remains how best to manufacture the coupling.

Silicon nitride or silicon carbide are the best materials for the spherical parts of the coupling.[149] Either balls or cylinders with a hemispherical end can be used in the coupling. A cylinder with a spherical end can be pressed or epoxied into a hole to obtain near monolithic properties. Mounting of a ball takes more care to insure that the compliance of the mounting method is very low compared to the compliance of the coupling. Ball mounting methods include:

- A shaped seat can be machined, ground, or electrodischarge machined into the mounting surface for the ball. Shapes for the seat, in order of increasing compliance, include:
  - Hemisphere
  - Cone
  - Tetrahedron

  For the hemisphere, the bottom of the hole should be counterbored to prevent contact of the ball near its pole which would increase lateral compliance. For any of these seats, an extra ball of the same size should be burnished in place or pressed in until the surface is brinelled, which will help to ensure that the ball does not make contact at only two points in the case of a spherical or conical seat. A ball can then be brazed or epoxied into the seat to make the ball act as an integral part of the structure.
- A surface can be ground flat and then annular grooves ground around the ball locations. Sleeves can then be pressed into the grooves. The balls can then be pressed into the sleeves until they contact the flat surfaces. The balls should deeply brinell the flat surface in order to increase the bearing area and decrease the compliance.

With any of these methods, the difficulty in accurately locating the balls from fixture to fixture may suggest that the balls should be affixed to a rough machined fixture. The fixture would then be clamped to the grooved portion of the coupling and finish machined.

The ideal material for the grooves would also be a hard ceramic because it would not corrode and the coefficient of friction between the ball and the groove would be minimized, which would maximize the repeatability of the coupling. The grooves can be profile ground in a monolithic plate using a profile grinder and an index table, or the grooves can be made in modular inserts which are bolted or bonded into place on the coupling.

[149] Standard size silicon nitride balls are available from Cerbec Bearing Company, 10 Airport Road, East Granby, CT 06026 (203-653-8071). Cylinders with spherical ends can also be manufactured.

Effects of Contamination on the Coupling Interfaces

Minimizing contamination of the coupling interface may require cleaning with a solvent such as Freon. However, removing all the dirt and oil can lead to fretting corrosion of metal surfaces. Fretting corrosion is caused by repeated alternating sliding contact of two adjacent surfaces. As noted by O'Connor,[150] damage can occur with slip amplitude as small as 1 μm (40 μin.) and Tomlinson[151] puts the figure at values as low as 0.03 μm (1 μin.). This value is also indicated by Waterhouse.[152] The action of fretting tends to increase the coefficient of friction until stresses high enough to initiate a fatigue crack form. For a kinematic coupling with its high contact stresses, this could be a problem. Also, metal-to-metal contact results in higher coefficients of friction. This occurs when metal-to-metal contact causes local cold welds between asperities, which are then torn apart. The newly exposed fresh metal surface quickly oxidizes and the process repeats.

Fretting can be prevented by using ceramics for the balls and/or grooves. A ceramic such as silicon carbide has a Hertz strength on the order of 6.9 GPa (1000 ksi), a Young's modulus of 415 GPa (60×10^6 psi), and a fatigue life approximately one to two orders of magnitude higher than for bearing steels.[153] However, fracture toughness of silicon carbide is about one-third that of SAE 52100 bearing steel, so care is required during handling. Another alternative is silicon nitride, which has a working Hertz stress on the order of 6.9 GPa (1000 ksi) and a Young's modulus of 311 GPa (45×10^6 psi). Note that silicon nitride has been made tough enough to allow it to be certified for gas turbine engine ball bearings.

7.7.4 Experimental Determination of Repeatability

There are several factors which theoretical considerations cannot readily address, including friction, surface finish, and surface contamination. The best way to determine the effects of these factors on kinematic coupling repeatability is to test a full-scale model. Before embarking on an elaborate experimental investigation of the effect of various real-world parameters on kinematic coupling repeatability, it was decided to see how good the coupling could be when fabricated with a modest budget. A steel load frame was constructed and a pneumatic piston was used to raise, lower, and provide the preload between two 356 mm (14 in.) diameter, 102 mm (4 in.) thick, kinematically coupled cast iron disks. A 5800 N (1300 lbf) preload with 1% repeatability was used in all experiments. The lower disk was itself mounted on a pseudo kinematic mount of tapered bolt heads threaded through the steel frame and seated into grooves in the lower disk. The piston would raise the upper disk about 13 mm (0.5 in.) above the lower disk, then lower it onto the lower disk and apply the preload.

The piston ram transferred its force through a shear beam type of load cell to a 38 mm (1.5 in.)-thick aluminum plate. The plate was loosely connected to the upper cast iron disk with three 10 mm (3/8 in.) bolts in 13 mm (1/2 in.) clearance holes located over the disks' contact points. Neoprene rubber pads 13 mm (0.50 in.) thick and 5 cm (2 in.) in diameter were placed between the aluminum plate and the upper cast iron disk. The upper disk was raised only about 13 mm (0.50 in.) between cycles. This design formed a very tight structural loop which minimized mechanical and thermal noise. The massiveness of the system coupled with room temperature stability of ± 1 C^o made thermal growth errors in the measurements negligible. The sensitivity of the system was about 6.5 mV/μm (0.17 mV/μin.). Stability of the system output over a 24-hour period was on the order of 0.05 μm (2 μin.).

Error motions of the upper disk in axial and radial directions as well as tilt motions about two orthogonal axes are of interest. To measure these error motions, six LVDTs were used. Three of these LVDTs were mounted vertically in the steel brackets around the circumference of the disk and 120 degrees apart from each other. These LVDTs were used to measure axial and two tilt

[150] J. J. O'Connor, "The Role of Elastic Stress Analysis in the Interpretation of Fretting Fatigue Failures," in Fretting Fatigue, R.B. Waterhouse (ed.), Applied Science Publishers, London, 1981, p. 23.

[151] G. A. Tomlinson et. al., Proc. Inst. Mech. Eng., Vol. 141, 1939, p. 223.

[152] R. B. Waterhouse, "Avoidance of Fretting Failures," in Fretting Fatigue, R.B. Waterhouse (ed.), Applied Science Publishers, London, 1981, p. 238.

[153] J. R. Walker, "Properties and Applications for Silicon Nitride in Bearings and other Related Components," Workshop Conserv. Subst. Technol. Crit. Met. Bearings Relat. Components Ind. Equip. Opportun. Improv. Perform., Vanderbilt University, Nashville, Tenn, March 12-14, 1984.

Axial:	0.90 μm (35 μin.)
Radial:	1.40 μm (55 μin.)
Tilt-X:	~ 5 μrad
Tilt-Y:	~ 5 μrad

Figure 7.7.9 Unlubricated kinematic coupling's 3σ repeatability.

error motions. The other three LVDTs were mounted radially on the same brackets that held the vertical LVDTs. They were used to determine the radial error motion of the upper disk with respect to the lower disk. One of the radial LVDTs was redundant, and used to determine closure of the measurements.

The data acquisition system consisted of a desktop computer, a relay-activated transducer amplifier for the LVDTs, a load cell to measure the preload, and several digital voltmeters to monitor the load cell and LVDT outputs. The computer controlled and monitored the entire experiment including raising and lowering of the disk, reading the LVDT outputs, and calculating the error motions. After each application of preload, the data acquisition software waited for about 20 seconds to allow for the system to settle. After this delay, the computer began scanning the LVDT outputs. For each LVDT, it took 10 readings and stored the average. Finally, after taking the load cell reading, the computer raised the upper disk. This cycle was repeated 600 times with 15-second intervals between each cycle. In this manner, automated testing took place over a period of days to obtain several sets of data for each condition tested.

Results

The first round of experiments were performed using steel balls epoxied into hemispherical seats in the lower disk, and hardened steel Gothic arches in the upper disk. With time, repeatability decayed to only about 2.5 μm (100 μin.). To check the reliability of the measurements, radial LVDT outputs were used to back-calculate the radius of the coupling. Agreement between sets of measurements at the start of the tests and the end was about 0.0025 μm; thus accuracy of the measurement was confirmed. Upon examining the ball-groove interface, it was found to have brown rust marks. This occurrence of fretting corrosion meant that steel balls could not be used with steel grooves for extended periods.

The steel balls were then replaced with silicon nitride balls, but the repeatability was found to be only on the order of 2.5 to 12.5 μm (100 to 500 μin.). It was thought that the ball was rocking in the hemispherical seat despite the epoxy. At the ball groove interface, no fretting was observed, but the interface did contain smear marks visible with the naked eye. No brinelling was observed. The hemispherical seat was then ground out into a Gothic arch shape to match the other groove. The logic was that in production, it would be cheaper to make all hardened steel inserts the same. Even when epoxied in place, the balls migrated along the grooves with changing load eccentricities, and repeatability was found to be only on the order of 12.5 to 25 μm (500 to 1000 μin.).

The next step was to EDM a tetrahedral seat into three of the vee blocks for holding the silicon nitride balls. This tetrahedral seat ensured that the balls themselves were kinematically mounted. The balls were placed in the tetrahedral seats, and the preload mechanism was cycled. The upper disk was then raised and liberal amounts of epoxy poured around the balls in the tetrahedral seats. The preload was then applied and the epoxy allowed to cure.

The balls and grooves were then carefully cleaned with Freon to ensure that the test would not have an error component caused by foreign matter contamination. The results are shown in Figure 7.7.9. The radial repeatability was only 1.40 μm (55 μin.), and the system never stabilized. A careful analysis was made of the situation, and first-order calculations indicated that the error could be caused by friction in the system. When stability tests were made, the system would take hours to stabilize. This indicated that high interface stresses were present that were slowly trying to relax. The solution therefore seemed to be to lubricate the ball-groove interface. It would be interesting to determine how the surface finish affects the need for a lubricant.

The next series of experiments were preceded by a generous application of grease to the balls and the Gothic arch grooves. As the upper disk was raised and lowered by the piston, it wobbled considerably, and would thus wipe a new smear of grease onto the contact points before the preload was applied. Settling time for this system was on the order of 10 to 20 seconds. The results obtained

		Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Avg (2-6)
Axial:	(μm)	0.76	0.28	0.38	0.23	0.28	0.35	0.30
Radial:	(μm)	0.68	0.35	0.33	0.43	0.30	0.25	0.33
Tilt-X:	(μrad)	1	2	1	2	2	1	1.6
Tilt-Y:	(μrad)	3	4	3	2	3	2	2.8

Figure 7.7.10 Lubricated kinematic coupling's 3σ repeatability.

from three different sets of 600 readings each are shown in Figure 7.7.10. Figures 7.7.11 and 7.7.12 show typical data from these tests.

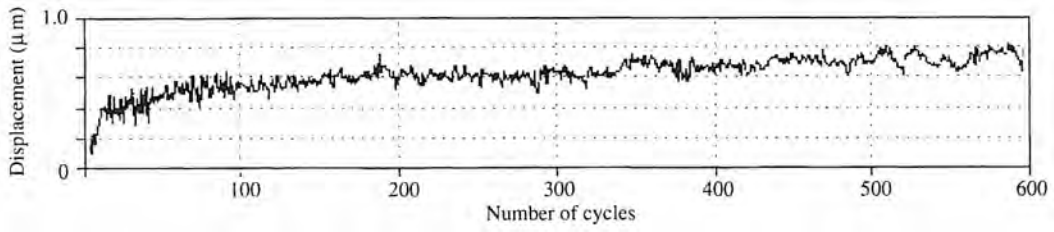


Figure 7.7.11 Radial repeatability of the lubricated kinematic coupling.

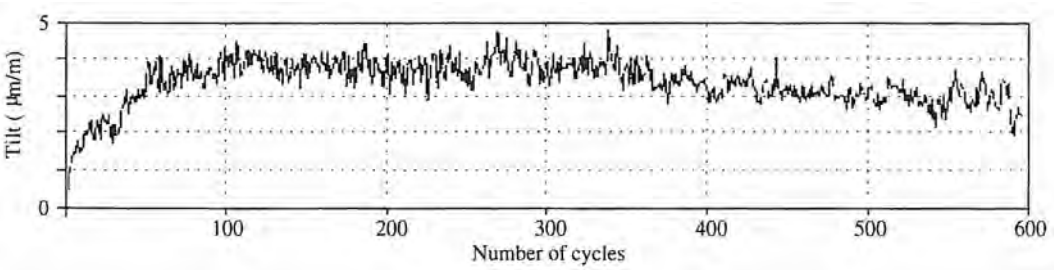


Figure 7.7.12 Y axis tilt repeatability of the lubricated kinematic coupling.

With respect to the first cycle, when lubricated, the coupling repeats radially to about 0.68 μm (27 μin.) and axially to about 0.76 μm (30 μin.). After a wear-in period of about 50 cycles, the system quickly stabilized. The average repeatability after wear-in was 0.30 μm (12 μin.) radially and 0.33 μm (13 μin.) axially as shown in Figure 7.7.10. The wear-in period is attributed to the "rough" surface finish of the ground arches (0.5 μm). If polished arches were used to begin with, the wear-in period should not be necessary. When subjected to a 90 N (20 lbf) radial force applied 5 cm above the line of coupling action, the coupling was still very repeatable as shown in Figure 7.7.13. Figure 7.7.13 also shows the average stiffness of the coupling along and about its axes. The three-groove kinematic coupling is definitely stiff enough to withstand light machining forces, such as encountered in diamond turning operations, and still maintain submicron accuracy.

In general, with respect to the scatter in the data, consider that the surface finish of the ground Gothic arches was only about 0.5 to 0.8 μm (20 to 30 μin.) R_a. Also consider that the coefficient of friction of silicon nitride on silicon nitride without lubrication would be lower than that of silicon nitride on lubricated steel. Furthermore, the footprint of silicon nitride on silicon nitride would be about half that of silicon nitride on steel. Thus a silicon nitride on silicon nitride kinematic coupling would more closely achieve a true kinematic interface which requires point contact between surfaces.

Since these tests yielded acceptable repeatability levels, it was decided to construct a base and pot chuck that used the arches and balls from the repeatability tests, to hold and cut a hemispherical part on a CNC lathe. A vacuum system was not available for the lathe, so three L-shaped brackets were bolted to the lower disk, and setscrews, positioned over the location of the balls in the upper disk, were used to preload the coupling. A 400-mm (8-in.) diameter 304 stainless steel blank was held by the pot chuck, and both roughing and finishing cuts were made. It is interesting to note that

		Set 7	Set 8	Set 9	Avg (7-9)	Stiffness
Axial:	(μm)	0.65	1.25	0.58	0.83	1.09×10^8 N/m
Radial:	(μm)	0.23	0.78	0.70	0.57	1.58×10^8N/m
Tilt-X:	(μrad)	11	12	5	9.3	4.83×10^5N-m/rad
Tilt-Y:	(μrad)	25	3	7	11.7	3.85×10^5 N-m/rad

Figure 7.7.13 Lubricated kinematic coupling's repeatability with 90 N applied radial force.

even though the set screws were tightened only to a few newton meters the setscrews never loosened during the roughing or finishing operations. The depth of the finish cut on the contoured section of the part was 0.13 mm (0.005 in.) and the feed rate was 55 surface meters per minute (180 SFM). With a new ceramic insert tool, the surface finish was 0.3 μm R_a (12 μin.). On a different section, a 0.52 mm (0.020 in.) depth of cut was used and the surface finish was 0.4 μm R_a (15 μin.). Thus when subjected to the ultimate test, roughing and finishing machining tests in a difficult to machine material, the coupling performed as well as the machine itself.

7.8 Design Case Study: A Unique Machine for Grinding Large, Off-axis Optical Components[154]

By incorporating large off-axis aspheric elements, the optical system designer has more options at his disposal. The lack of symmetry of these elements renders them extremely difficult or unsuitable for manufacture by classical methods. Current approaches place the optical center of the element at the center of rotation of a rotary table and all operations that are required to produce the form and surface finish proceed in a radial manner. In the segmented system of production, where many different mirrors are assembled together to produce one large parent optic, this approach cannot be used effectively. Each off-axis element can be generated more efficiently by the use of advanced three-axis CNC control. Figure 7.8.1 shows a typical off-axis optical system configuration.

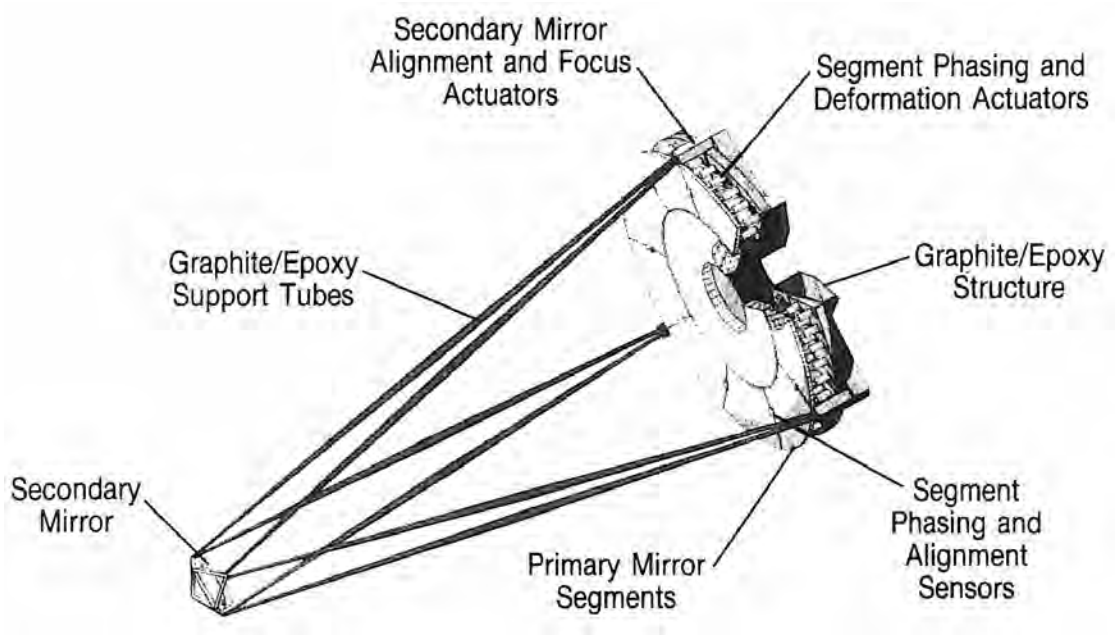


Figure 7.8.1 Kodak's active optical control system. (Courtesy of Eastman Kodak Company.)

[154] This case study was written by P. B. Leadbeater, M. Clarke, W. J. Wills-Moren, and T. J. Wilson. It originally appeared in Precision Engineering as "A Unique Machine for Grinding Large, Off Axis Optical Components: the OAGM 2500," (Oct. 1989, Vol. 11, No. 4). Used in the edited form with permission of Butterworth & Co. (Publishers) Ltd.

In 1980, Eastman Kodak Company developed a 1-m-capacity, computer-controlled grinding system capable of grinding glass and glass/ceramic axisymmetric optical components to within 6 μm of the required aspheric surface. This accuracy was generally independent of the amount of aspheric departure from the best-fit sphere. Using this system, a number of 0.8-m diameter aspheric lenses with approximately 900 μm of aspheric departure were generated. By producing a surface figure with less than 6 μm of departure from final form, an optic can readily be tested interferometrically after a small amount of polishing.

In order to expand this capability to larger, off-axis elements, Eastman Kodak Company contracted Cranfield Precision Engineering to design, manufacture, and commission a machine with a working envelope of 2.5 m × 2.5 m × 0.6 m. A basic assumption in the design requirements was that the machine coordinate system would be based on the segment rather than the parent optic. This approach also reduces the working envelope requirements and allows a wide degree of flexibility in the off-axis component configuration. A second basic assumption was that the machine must incorporate in situ metrology for rapid evaluation of the optical figure during the grinding process. This would allow analysis and compensation of such items as tool wear as well as the ability to pass surface figure data to the next polishing process. In this way, the surface figure can continue to be addressed while the surface finish is being improved.

In summary, the functions of the machine, known as the OAGM, are:

- To generate within the working volume these dimensional surfaces by use of a grinding wheel of partial spherical form.
- To measure the geometry of the workpiece using a contact stylus in X, Y, Z coordinates.[155]

7.8.1 Design Concept

During the early stages of the design, many configurations for the machine were considered. The decision to mount the workpiece horizontally was made primarily to ease the handling of very large and expensive components. Other factors considered were machine symmetry, overall stiffness, ease of manufacture and the results of error budget analysis. As stated earlier, the strategy of workpiece mounting is to tilt the segment to minimize surface slope, reducing the requirements for accuracy in the X and Y directions. However, the requirement for high accuracy in the Z direction remains and the error budget analysis showed that this must be better than 1 μm over the full 2.5 m × 2.5 m area. It was concluded that it would not be practical to expect this accuracy to be achieved directly from the bearings of the machine with the large moving masses involved and therefore a separate frame of reference for measuring Z motion was adopted. Following from this and the practical considerations stated earlier the configuration shown in Figure 7.8.2 was selected.

The workpiece together with its fixture is mounted on a stationary table surrounded by the base unit. The bearing rails for the traveling gantry are located at either side of the base and additional X axis travel of the gantry is incorporated to allow for ease of component loading. Surrounding the worktable is the metrology frame carrying precision optical reference straight edges on either side of the worktables. The bridge forms the Y-axis bearing rail system and on this is carried the Y, Z assembly. Also forming part of this assembly is the Y reference beam, inside which is mounted a third reference straightedge. The precision grinding spindle together with the retractable air bearing measuring probe is mounted at the lower end of the Z-axis slide.

Bearings

The X and Y axes utilize hydrostatic bearings fed with temperature-controlled oil. The vertical Z axis is aerostatic. Throughout the design stage careful consideration was given to the collection and return of the exhaust oil from the bearings, which is then filtered and temperature controlled to ±0.1 C° before being pumped back through the bearings. Total oil flow for the X and Y axes is only 1.6 liters/min with nominal gaps of 25 μm. Vertical stiffness of each X bearing is 9500 N/μm and 8000 N/μm horizontally. The Y axis stiffness values are approximately half those of X.

The X-axis bearings take the form of a support bearing together with guidance bearings on either side of the right-hand guide rail and a support bearing only on the left-hand side of the machine as shown in Figure 7.8.3. The configuration of the Y axis hydrostatic bearings is shown in Figure

[155] J. B. Bryan, "Design of a New Error Corrected Coordinate Measuring Machine," Precis. Eng., Vol. 1, No. 3, 1979, pp. 125–128.

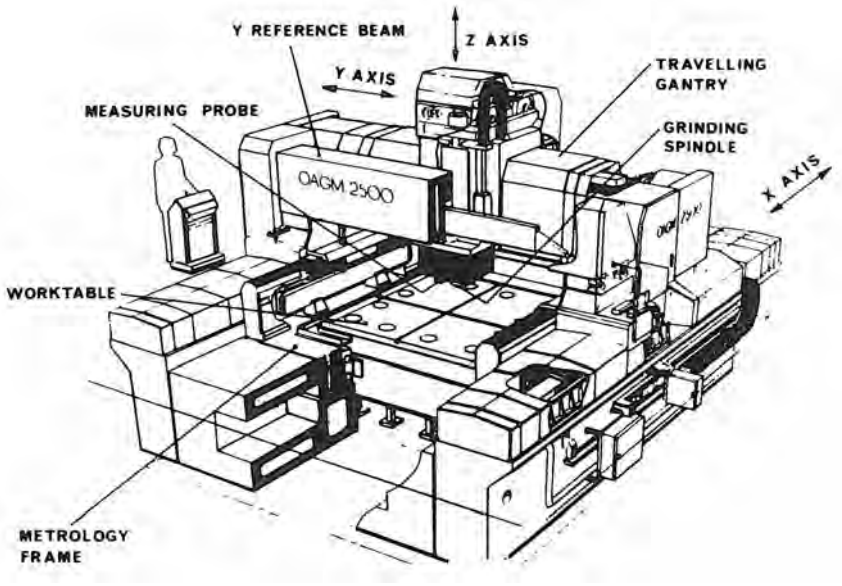


Figure 7.8.2 OAGM 2500 machine configuration. (Courtesy of Cranfield Precision Engineering Ltd.)

7.8.4 and consists of an upper bearing having vertical and horizontal constraint and a lower bearing having only horizontal constraint. A fully constrained prismatic aerostatic bearing carriage system is used for the vertical Z axis.

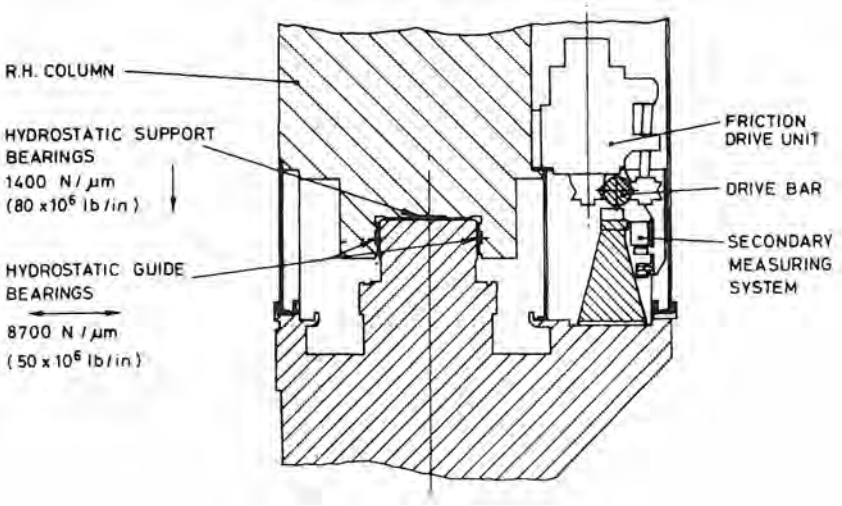


Figure 7.8.3 X-axis guide and drive assembly. (Courtesy of Cranfield Precision Engineering Ltd.)

Machine Structure

The main structure of this machine consists of lightweight steel weldments which are filled with Granitan® S100 cored with polystyrene blocks. The Granitan dominates producing a structure with high stability and excellent internal damping.[156] The weldments themselves form the molds and provide a suitable means of attachment for other components.

The base is made of four major components. Each section weighs approximately 17 tons. Outer bearing members mount on either side of the two center spacing sections. This four-part

[156] T. J. Philips, "A Specific Polymer Concrete for Machine Structures," 5th Int. Congr. Polym. Concr., 1987, p. 127.

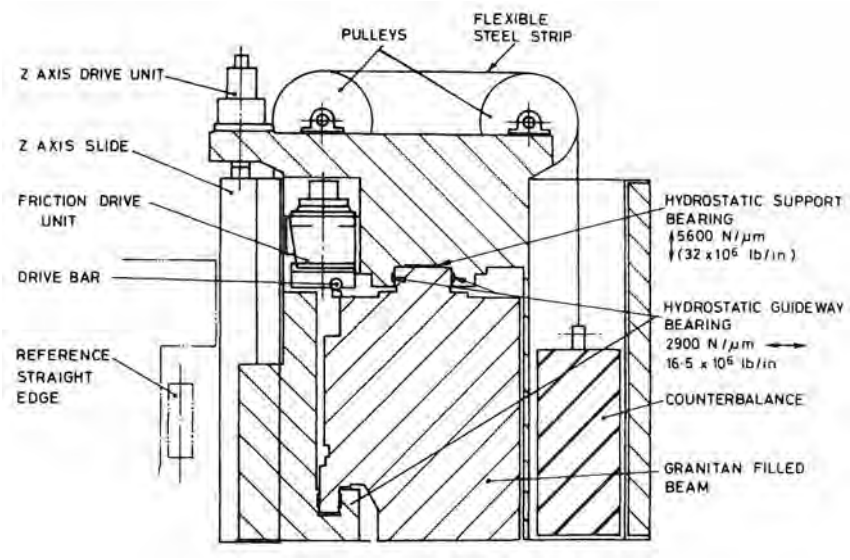


Figure 7.8.4 Y and Z axes. (Courtesy of Cranfield Precision Engineering Ltd.)

construction facilitates manufacture, shipping, and installation. To ensure precise reassembly of the base unit at the final site, the joints of the base unit were scraped flat prior to assembly.

The 2.5 m $\times$ 2.5 m cast iron worktable is mounted kinematically to the base structure in a manner such that any deflections of the base caused by motions of the gantry will not cause distortions in the table. In order to minimize the loading on the kinematic locations, a mechanical weight-relieving system is employed. The surrounding metrology frame is connected kinematically to the worktable, and again a weight-relieving system is built in. The table and metrology frame were designed to compensate for thermal changes in the structure and to link tightly the optical component being fabricated with the optical reference flats in the metrology frame. Corning ULE 7971 titanium silicate material is used for each of the three 2.75 m $\times$ 0.3 m $\times$ 0.1 m reference flats in the metrology frame. The traveling gantry is more conventional in its construction, but here again all major structural members are formed from Granitan®-filled weldments.

Actuators

The principal drive system for the horizontal X and Y axes is by means of traction (friction) drives.[157] A synchronized twin friction drive is provided in the X axis with a drive system on either side of the machine, while a single drive system is employed in the Y direction. Each drive consists of a unit as shown in Figure 7.8.5 and employs a dc torque motor directly driving a vee roller, which in turn operates against a circular section drive bar. This drive bar is rigidly coupled at each end of the stationary member. A series of spring-loaded devices spaced at intervals along the traction bar are provided to prevent sagging of the bar under its own weight. A further preload roller provides a force to maintain the drive roller against the traction bar with sufficient force to prevent any slippage due to accelerating and grinding forces.

To provide motion to the shorter Z axis, a ballscrew system operating through a backlash-free and noninfluencing nut is employed. The drive to the screw is again a directly coupled dc torque motor and the total Z axis assembly is mechanically counterbalanced. As a precaution against any malfunction, a fail-safe braking system is fitted to each axis drive.

Metrology Frame and Measuring System[158]

The reference measurement system for this machine is based on the metrology frame concept, an approach which has been incorporated in the past on a number of specialized high-precision

[157] M. Douglas, "Friction Drive Matches Encoder Resolution," Drives Controls, July/Aug, 1988.

[158] For a more detailed discussion, see W. J. Wills-Moren and T. Wilson "The Design and Manufacture of a Large CNC Grinding Machine," Ann. CIRP, Vol. 38, No. 1, 1989, pp. 529–532.

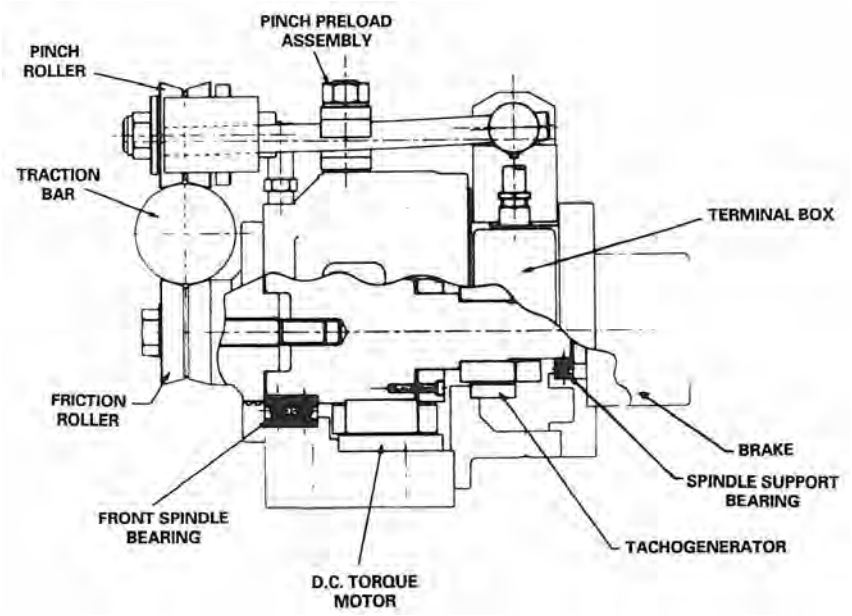


Figure 7.8.5 54-mm friction drive assembly. (Courtesy of Cranfield Precision Engineering Ltd.)

machines.[159] In the design of the OAGM, the application of this concept has been limited to the Z axis, which is the primary sensitive direction. The basic reference frame as set up by the three precision glass reference bars is shown in Figure 7.8.6. The principle of operation applies to both the grinding and measuring modes.

In practice, two reference bars are mounted on either side of the worktable in the X direction and are nominally parallel to the table and coplanar. The third reference bar is mounted above and at right angles to the X bars, forming the reference for movements in the Y direction. It can be seen that as the distances A and B are precisely known and any variations to these are taken into account, then dimension C can also be accurately established, having referred in the first instance to a suitable datum. The OAGM 2500 metrology frame is shown in more detail in Figures 7.8.7 and 7.8.8.

In order to control the machine to the required accuracy, a multipath laser interferometer system is built in. Three Zygo Axiom 2/20® lasers and associated optics are incorporated with the output beam of each laser being split three ways. These form a comprehensive monitoring system for each axis. Additionally, any change in the environmental conditions which occurs during operation of the machine is detected, allowing compensation to be made automatically. Figure 7.8.9 shows an illustration of the laser metrology system.

[159] See, for example, J. B. Bryan, "Design and Construction of an 84 Inch Diamond Turning Machine," Precis. Eng., Vol. 1, No. 1, 1979, pp. 13–17; R. R. Donaldson and S. R. Patterson, "Design and Construction of a Large Vertical Axis Diamond Turning Machine," Proc. SPIE Ann. Int. Technol. Symp., 1983, pp. 433–438; J. B. Arnold, R. R. Burleson, and R. M. Pardue, "Design of a Positional Reference System for Ultra-Precision Machining," Oak Ridge Y12 Plant Report Ref. Y2202, 1979.

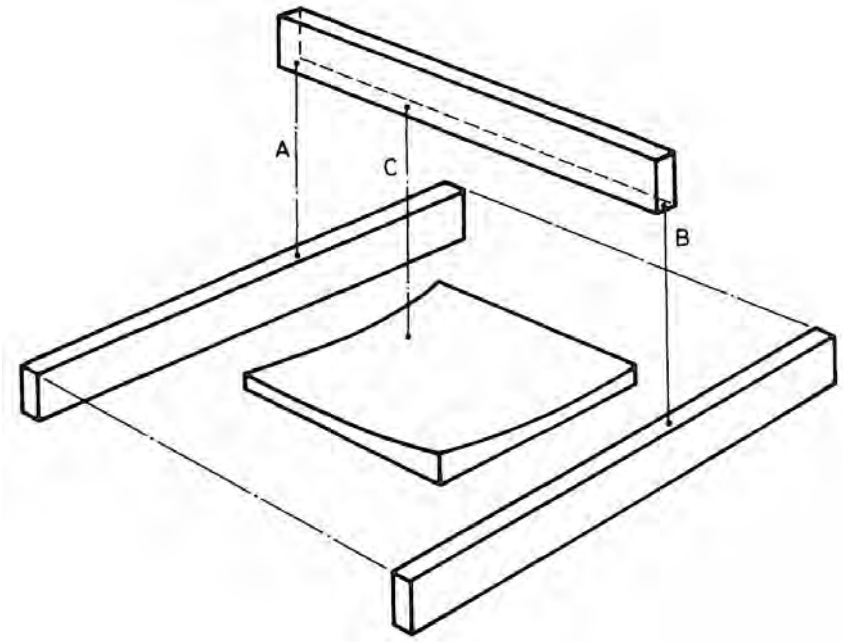


Figure 7.8.6 Basic metrology reference frame. (Courtesy of Cranfield Precision Engineering Ltd.)

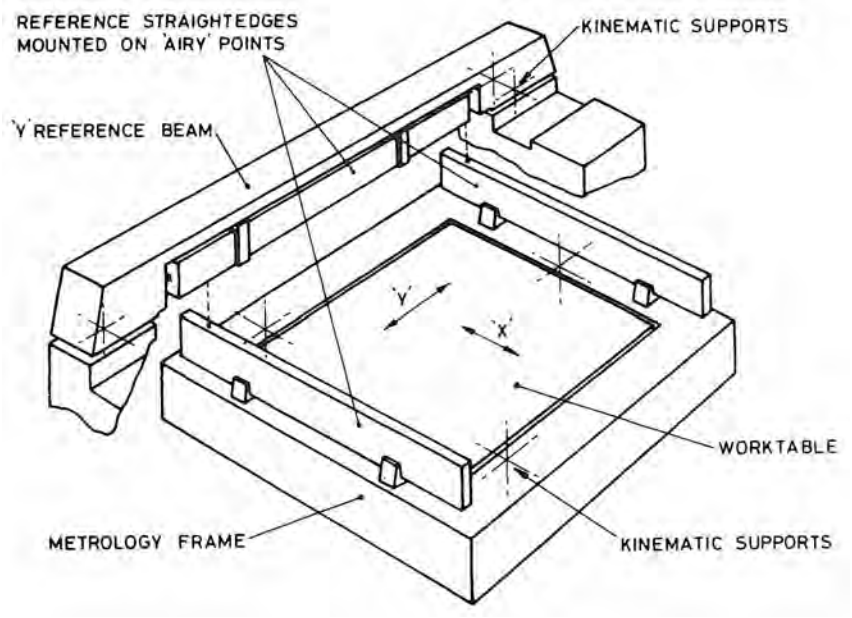


Figure 7.8.7 OAGM 2500 metrology frame. (Courtesy of Cranfield Precision Engineering Ltd.)

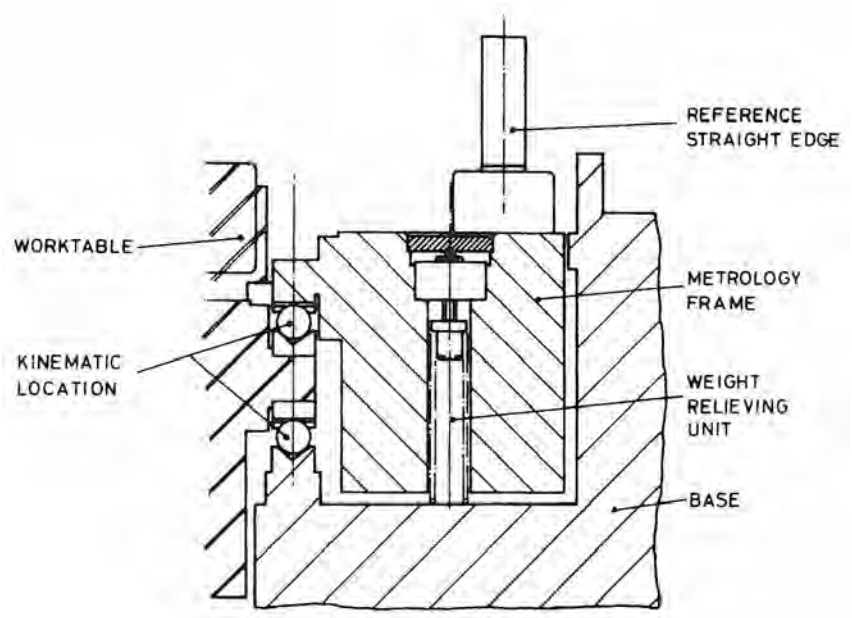


Figure 7.8.8 OAGM 2500 metrology frame interface. (Courtesy of Cranfield Precision Engineering Ltd.)

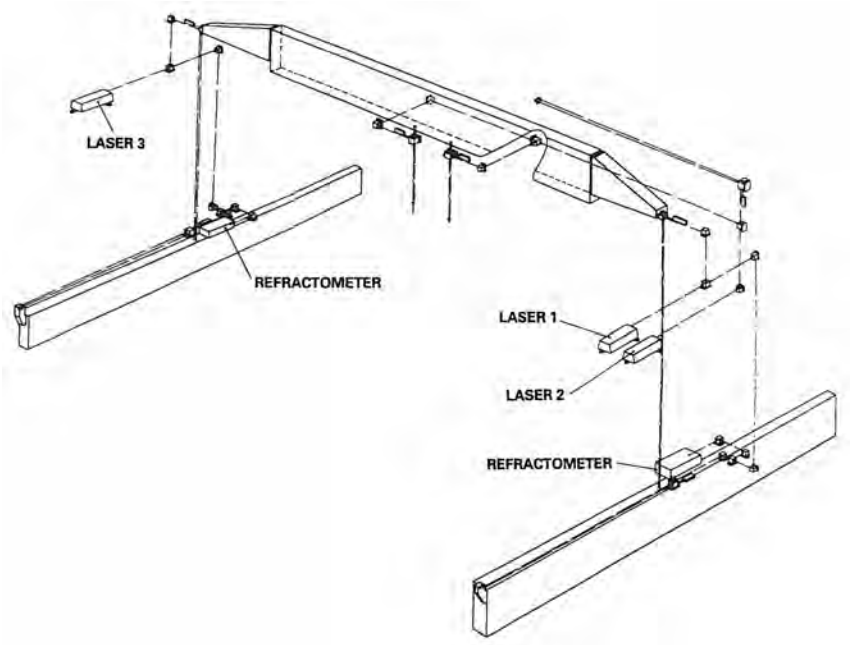


Figure 7.8.9 Measurement system for off-axis grinding machine. (Courtesy of Cranfield Precision Engineering Ltd.)

Grinding Spindle

An externally pressurized air bearing spindle inclined at 10° to the horizontal is provided. This spindle has variable speed to a maximum of 3000 rpm and is electrically powered. Temperature-controlled cooling water is circulated in the spindle housing to maintain the temperature to within $\pm 0.1C^{o}$. The rotor is accurately dynamically balanced to ensure the highest possible surface quality on the glass workpiece. Figure 7.8.10 shows the spindle arrangement, which is designed to take a 200-mm diameter wheel of partial spherical form. This geometry ensures effective contact with the component surface up to a maximum of 7° slope. Full flood coolant temperature controlled to within $\pm 0.1C^{o}$ is supplied to the grinding wheel. Complete segregation of the coolant from the hydrostatic bearing oil prevents contamination of the bearing oil with glass swarf. Multiple covers provide protection to both the bearings and the metrology frame optical elements.

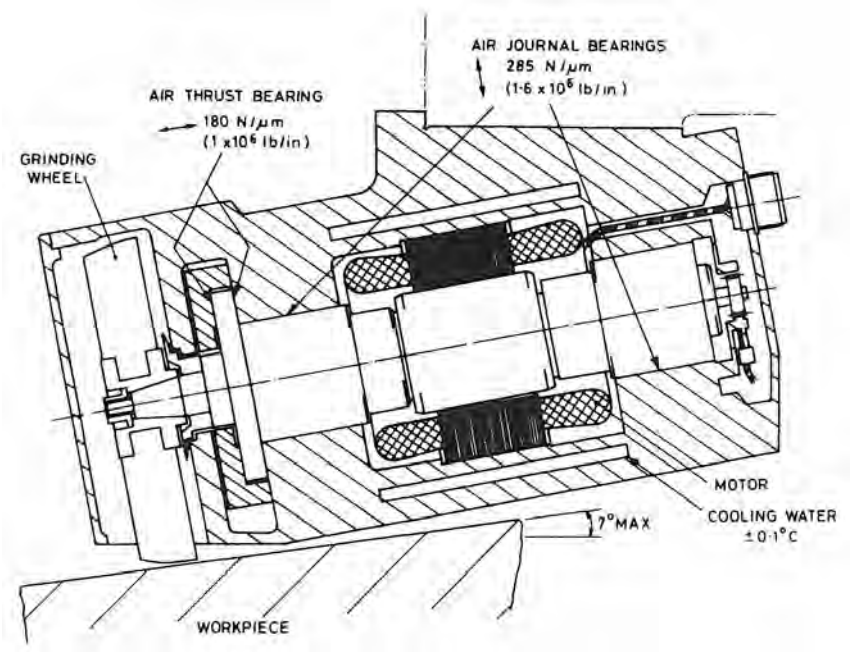


Figure 7.8.10 OAGM 2500 grinding spindle. (Courtesy of Cranfield Precision Engineering Ltd.)

In situ Metrology System[160]

The machine's positional system is based around the reference metrology frame which effectively sees no variation in loading. By use of a vertically acting probe mounted adjacent to the grinding spindle, the form of the workpiece can be determined. The workpiece optical surface can be completely scanned to provide a global evaluation of the surface. Surface data from the profilometry probe can be analyzed using Kodak's proprietary interferometry evaluation software.

This probe is retractable, as shown in Figure 7.8.11, into a sealed housing for its protection during grinding operations. The spindle of the probe which carries the contact stylus is produced in Zerodur to minimize errors caused by any thermal changes. This spindle is carried in an externally pressurized air bearing, allowing a counterbalance system to operate and hence provide a low and adjustable contact force. A retroreflector referencing to the Y reference straightedge together with suitable optics performs the vertical (Z) displacement measurement function.

In order to stay within the small stroke of the probe, the Z-axis slideway is under servo control, and is caused to operate such that a null-seeking feedback device built into the probe is kept centralized. Additionally, small movements of the stylus (up to 4 mm) can be performed by means of electrically operated bias coils built into the probe body.

[160] See P. A. McKeown, W. J. Wills-Moren, and R. F. Read, "In-Situ Metrology and Machine Based Interferometry for Shape Determination," Proc. SPIE, Vol. 802, 1987.

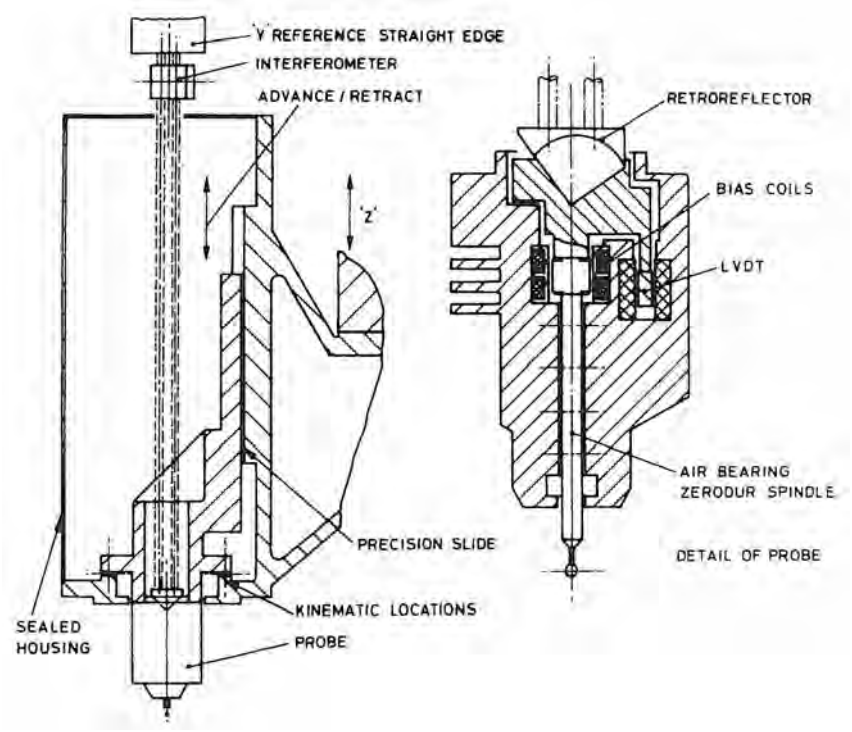


Figure 7.8.11 OAGM 2500 in-situ probing system. (Courtesy of Cranfield Precision Engineering Ltd.)

7.8.2 Summary

The machine was installed at Eastman Kodak Company, Rochester, NY, in 1990, and was found to meet spec. For primarily ergonomic reasons, the worktable and floor are at approximately the same level, with the base of the machine being installed in a pit. The foundations onto which the machine rests go down to bedrock and are isolated from the building foundations.

Temperature control on such a large machine also plays an important part in achieving the overall accuracy of component and therefore the environment is controlled to $\pm$ 0.5 C^o. All of the main heat sources, such as electrical control cabinets, hydraulic pumps, and temperature control units, are situated outside the working environment with the relevant services passing through ducts into the machine pit.

Chapter 8
Bearings with Mechanical Contact between Elements

The man who doesn't make up his mind to cultivate the habit of thinking misses the greatest pleasure in life. He not only misses the greatest pleasure, but he cannot make the most of himself. All progress, all success, springs from thinking.

Thomas Edison

8.1 INTRODUCTION

Since bearings are such a critical element in precision machines, one must *think* about the seemingly innumerable number of bearing design considerations that affect the performance of a machine, including:

- Speed and acceleration limits
- Range of motion
- Applied loads
- Accuracy
- Repeatability
- Resolution
- Preload
- Stiffness
- Vibration and shock resistance
- Damping capability
- Friction
- Thermal performance
- Environmental sensitivity
- Sealability
- Size and configuration
- Weight
- Support equipment
- Maintenance
- Material compatibility
- Mounting requirements
- Required life
- Availability
- Designability
- Manufacturability
- Cost

Because there are always design tradeoffs in choosing a bearing for a precision machine, all of these factors must be considered simultaneously by the design engineer. Each of these design considerations are discussed in general below, and also later with respect to the different types of bearings that make mechanical contact between their elements, including: sliding contact bearings, rolling element bearings, and flexural element bearings. Bearings without mechanical contact between elements (e.g., hydrostatic bearings) are discussed in Chapter 9.

Speed and Acceleration Limits

Speed and acceleration limits affect bearing performance in many different ways. For example, sliding bearings (e.g., cast iron on oiled cast iron) have a higher coefficient of friction than rolling bearings, and thus if they undergo high-speed motion through repeated cycles, they generate larger amounts of heat. On the other hand, rolling bearings are usually not as stiff or well damped as sliding bearings, and hydrodynamic bearings can suffer whirl instability if not properly designed. High accelerations require minimal friction, so more energy can be spent accelerating the mass of the machine, but if the acceleration levels are too high, rolling motion components can slip or jamb in recirculating tubes if they are not properly designed. One can always find an appropriate bearing for the job, but one must be careful not to blindly use the first bearing that comes to mind.

Range of Motion

The range of motion a bearing allows can greatly affect applicability. For example, a flexural bearing can provide perhaps the smoothest, most accurate motion attainable, yet its range of motion is severely limited. Also, bearings have a certain amount of overhead associated with them. For example, the carriage on a recirculating roller bearing may be 10 cm long, which makes it appropriate for a machine with a stroke of a meter, but may be inappropriate for an instrument with a stroke of 1 cm. As another example, non-recirculating rolling element bearings do not have the inaccuracies introduced by recirculating components, but the carriage length is coupled to the stroke length.

Applied Loads

Loads are an important consideration for choosing a bearing. For example, a roller bearing can support higher loads than a ball bearing because it has line contact instead of point contact, or a hydrostatic bearing can support a much larger load than an aerostatic bearing because the operating pressures are an order of magnitude larger. One must consider not only how large a load the bearing can support, but how the bearing reacts when it is overloaded and what happens if the bearing fails.

Accuracy

Accuracy has a dual meaning for bearings. The first is: accuracy of the motion of the components supported (e.g., total error motion). The second is: assuming that the rest of the system is perfect, the ability to allow the components supported to be servoed to a desired position. In general, the lower the friction and the higher the accuracy of bearing components, the greater the accuracy (both kinds) of the bearing.

Given that a bearing is supposed to support a device as it moves linearly from point A to point B, accuracy of the first kind is usually quoted by manufacturers as the *lateral deviation* from the ideal motion path or the *straight-line* accuracy or running parallelism. Since all components deflect under their own weight, it is sometimes better to not worry about accuracy and instead, worry about repeatability. With the development of systematic methods for measuring bearing motion and storing the results, in many cases repeatability is becoming a more important factor in the selection of a bearing for a precision machine.

Accuracy of the second kind is difficult to quantify because of the dependence on the rest of the machine; hence bearing manufacturers usually only make claims like "low friction and high accuracy of bearing components allow for excellent attainable accuracy of servo-controlled axes."

Repeatability

Like accuracy, repeatability has a dual meaning for bearings. First is repeatability of the motion of the components supported (e.g., asynchronous error motion and running parallelism repeatability). Second is assuming that the rest of the system is perfect, the ability to allow the components supported to be servoed to the same position. In general, the lower the friction and the higher the accuracy of bearing components, the greater the repeatability (both kinds) of the bearing.

Repeatability of the first kind is the ability of the bearing to repeat its motion. Once mapped, it can be used to allow other axes to compensate for errors in a particular bearing's accuracy. For linear motion bearings, this is often referred to as the *straight-line* repeatability or running parallelism repeatability. Repeatability is usually affected by the surface finish and accuracy of the components in the bearing.

Repeatability of the second kind is difficult to quantify because of the dependence on the rest of the machine; hence bearing manufacturers usually only make claims like "low friction and high accuracy of bearing components allow for excellent repeatability of servocontrolled axes."

Resolution

Resolution is the ability of the bearing to allow for a small increment of motion. This is affected by the bearing's friction level and smoothness of motion, which is related to the surface finish and shape accuracy of the bearing components for contact-type bearings. On the submicron level, distortions of bearing components and surface finish limitations can create a situation where, in order to move, the bearing element must move over a hump, hence giving the bearing a resolution on the order of the width of the hump. The obvious way to overcome this problem is to use components with great accuracy, high surface finish, and low friction or to use noncontact bearings such as fluid, air, or magnetic. For short travel ranges, flexural bearings should be used.

Preload

Many geometries deform nonlinearly; thus when subjected to a high initial load which causes a great deal of deformation, a small additional load causes a proportionately smaller amount of deflection.[1] Similarly, the change in interface stress will be less per given change in applied stress. Hence in order to achieve greater stiffness and fatigue life, a bearing may be preloaded. Care should be taken when selecting bearing spacing and preload to avoid a teeter-totter effect. Note that the

[1] For example, see the discussion on Hertzian contact stresses and deflections in Section 5.6.

higher the preload, the greater the deformation of the components and the generation of friction; therefore, the greater the chances for lower repeatability and resolution.

Stiffness

High stiffness is almost always desirable because a stiff machine can respond more quickly and accurately than can a floppy machine. On the other hand, for mechanical contact bearings, one often pays a price for stiffness in terms of cost and, when preload is involved, repeatability and resolution. With respect to bearing location and how forces are applied to prevent moments, see the discussion on Equation 2.2.29 regarding the *center of stiffness* of a system.

Vibration and Shock Resistance

In order to provide vibration and shock resistance, there must be no mechanical contact between bearing components, or the preload must be high enough so that the resultant varying stress level is a small percentage (typically 10%) of the stress level due to the preload. Sliding contact and fluid bearings (hydrostatic or hydrodynamic bearings) typically provide the best shock and vibration resistance.

Damping Capability

The damping capability of a bearing can have a significant impact on the ability of the machine as a whole to damp out vibrations and withstand shock and vibration. In many cases, it is the bearing interface which provides the only damping mechanism other than material damping; hence since the latter is often low, a well-damped bearing can be a significant asset. Sliding contact and fluid bearings (hydrostatic or hydrodynamic bearings) typically provide the best damping capability because of the viscous nature of the lubricating film between the surfaces. Unfortunately, by generalizing like this one can come to the incorrect conclusion that sliding bearings are always better than rolling bearings for general machine tool use. One must consider all the other factors which go into selection of the bearing.

Friction

High static friction with lower dynamic friction results in *stick-slip* or *stiction,* which leads to limit cycling in servos and generally should be avoided. Any bearing that relies on sliding or rolling mechanical contact between components has stick-slip to some degree, although manufacturers are constantly decreasing the difference between static and dynamic coefficients of friction. The difference between the static and dynamic coefficients of friction in rolling bearings is usually low enough to ignore. Since the static friction force always opposes the direction of force, when axes are contouring, if an axis's velocity changes direction, there will be a force discontinuity as it passes through the zero point and the friction force creates a force discontinuity. The effect on position and velocity performance will of course depend on the design of the rest of the system and the servos.[2] Sliding or dynamic friction is usually desirable because it helps damp out motion of the part supported by the bearing along the axis of motion provided by the bearing. Dynamic friction does cause heat to be generated.

Thermal Performance

There are several aspects regarding the thermal performance of a bearing:

1. How do the friction properties change with temperature, and how do these changing properties affect the dynamic (e.g., servo controllability) performance of the machine? If bearing performance changes significantly during machine warm-up, then the machine may only be usable after the machine is warmed up, or an adaptive controller or thermal performance map may be needed.
2. How are the bearing's accuracy, repeatability, and resolution affected by temperature changes? If the bearing components expand differentially, will it cause failure of the bearing? The latter point can be crucial in the design of temperature control systems for bearings.
3. What are the heat transfer characteristics across the bearing? If the bearing acts as an insulator, will it increase the warm-up time of the machine too much? If it acts as a conductor, will it transmit unwanted heat from one part of the machine to another?

[2] Digital control algorithms can be designed to change their servo constants when the velocity is zero. Feed-forward algorithms can also be effectively used to overcome the effects of stick-slip.

Environmental Sensitivity

In addition to thermal issues, one must consider moisture and dirt, and how a bearing's performance will be affected by their presence. For example, moisture content greater than 100 ppm in an oil mist lubrication system (oil dripped into a pressurized air stream) causes rolling element bearing life to decrease exponentially. One must also consider how a bearing's support systems (e.g., air or oil pump) are affected by dirt and moisture.

Seal-ability

If the bearing must operate in an unpleasant environment that could do harm to the bearing, one must consider how easily the bearing configuration lends itself to means for sealing out the environment.

Size and Configuration

For small precision mechanisms (e.g., a watch) some bearings and their attendant support systems may not be feasible (e.g., a magnetic or hydrostatic bearing is not easily worn about one's wrist). Note that the use of high-performance ceramic materials can increase the structural efficiency of a bearing and hence decrease its size for a given application. One should always be on the lookout for new materials and how they can be used to make better bearings.

Weight

Similar to the size issue is the weight issue. Once again there are fundamental limitations to the performance of different types of bearings, and materials often play a very key role. For example, sliding bearings made of ruby or sapphire (e.g., as in a jeweled watch) outperform virtually every other type of precision rotary bearing in a precision application where strict size and weight constraints exist.

Support Equipment

The ideal bearing requires no attention once installed; unfortunately, most high-performance bearings require continual upkeep or mechanisms to perform that upkeep: for example, automatic lubricators for sliding or rolling bearings, or a source of clean pressurized air or oil for aerostatic or hydrostatic bearings, respectively.

Maintenance Requirements

Is it better to choose a less expensive bearing that requires some maintenance than a more expensive bearing that requires no maintenance? One must be aware of the time value of money; and one must also consider ripple effects such as if the machine goes down, the rest of the line may also go down. More than one contract has been lost primarily on the basis of a manufacturer's perceived less than admirable service record. Also, one must consider that if a bearing requires support equipment, then the support equipment will require maintenance.

Material Compatibility

One must consider how the bearing materials will interact thermally and chemically with nearby materials. On the thermal side, a steel bearing press fit into an aluminum housing may come loose when the machine warms up. On the chemical side, consider the problem of fretting corrosion discussed in Sections 5.6 and 7.7. Using different materials (e.g., ceramics or stainless steel) can often easily lead to a solution without painful design headaches.

Required Life

Bearing life is difficult to define because for any bearing there are factors which can cause its performance to degrade. Therefore bearing life is defined as the allowable degradation of operating performance as described by other parameters pertinent to the bearing's operation. Most manufacturers are very good about providing life data for their bearings, although very few define life in terms of dimensional performance.

Availability

A modular bearing from a catalog can be easy to specify and will often more readily satisfy other criteria (e.g., cost and availability), but the design engineer must ask himself how far one goes in changing the machine to fit the bearing in order to allow for a modular off-the-shelf bearing to be used.

Designability

Designability also implies how much one changes the machine to fit the bearing. If a bearing is to be custom designed for the application, how does one's experience affect the chances of success? For example, air bearings are generally considered to be moderately difficult to design (and manufacture), whereas hydrostatic bearings are not that difficult to design (or manufacture).

Manufacturability

As an example of manufacturability, consider that it is easy to specify 45^o angles for bearing rails, but given a choice, production personnel would generally rather only have to deal with 90^o angles. Ease of alignment and preload adjustment are also manufacturability issues.

Cost

The biggest headache for any design engineer is considering cost. Cost manifests itself in terms of procurement, maintenance, and design integration costs. For example, a magnetic bearing may provide the ultimate accuracy, but are the costs associated with the need for increased space and sensor and microelectronic support hardware worth the effort? An air bearing provides exceptional dimensional performance, but it is very sensitive to machine crashes, which can easily overload and destroy the bearing and void the warranty.

Summary

Evaluating these parameters for different types of bearings represents an accumulation of a large amount of information. Sometimes manufacturers' catalogs will not contain the necessary information, and then it is up to the engineer to obtain it through persistent questioning. If a suitable bearing is not available from a reputable manufacturer, the design engineer can then consider the use of a custom-designed bearing. Always remember, *caveat emptor!*

8.2 SLIDING CONTACT BEARINGS

Sliding contact bearings are the oldest, simplest, least expensive bearing technology, and they still have a wide range of applications, from construction machinery to machines with atomic resolution. Sliding bearings are thus a very important element in the machine design engineer's tool kit. Sliding contact bearings utilize a variety of different types of lubricants between various interface materials. Lubricants range from light oil to grease to a solid lubricant such as graphite or a PTFE polymer. Because they often distribute loads over a large area, contact stresses and space requirements are often low while stiffness and damping are usually high. In this section, general properties of sliding contact bearings are discussed followed by a discussion of design considerations.

8.2.1 General Properties[3]

There are numerous types of sliding contact bearings available and in the context of discussing their general properties, some specific categories will be discussed. In general, one should note that all sliding contact bearings have a static coefficient of friction that is greater (to some degree, no matter how small) than the dynamic coefficient of friction (static μ $>$ dynamic μ). The difference between the static and dynamic friction coefficients will depend on the materials, surface finish, and lubricant.

Material Combinations: Cast Iron on Cast Iron

For a long time, cast iron on cast iron was the dominant type of linear bearing used in machine tools because of the inherent lubricity provided by the graphite in cast iron, and the fact that cast iron can be hand scraped to a very high degree of accuracy. However, it is good to have one surface harder than the other in order to decrease wear.

Material Combinations: Cast Iron on Steel

Once precision grinding machines and high-strength steel became commonplace, it was found that by making one surface in a sliding contact bearing harder than the other, better wear

[3] Which came first, the chicken or the egg? So it is with the issue of which to discuss first, general properties or design considerations. Ideally, Section 8.2.1 will be read once again after 8.2.1 and 8.2.2 are read in order.

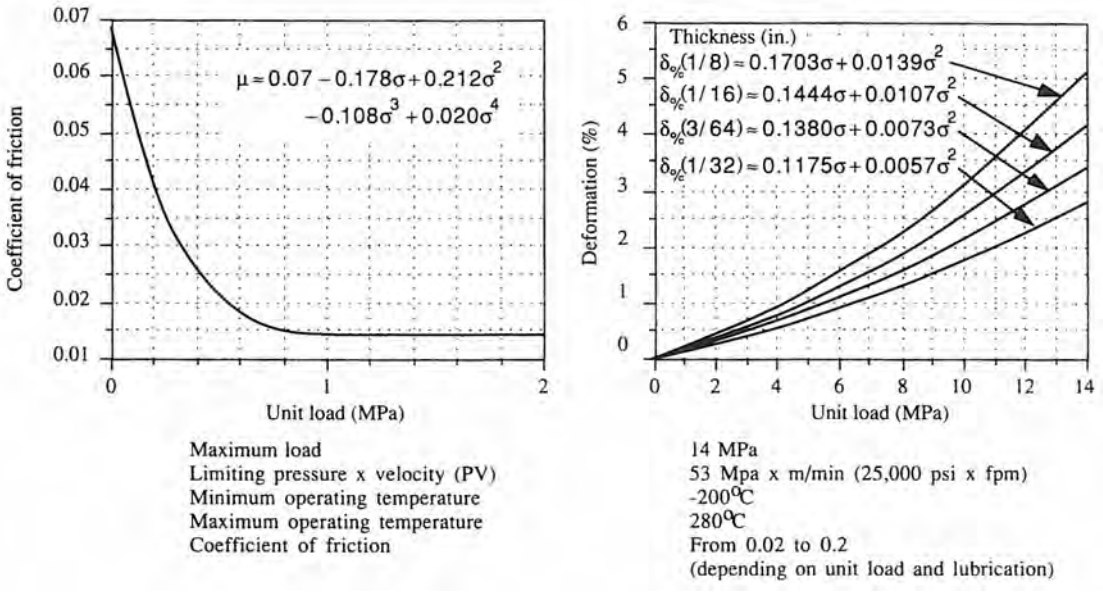
characteristics were generally obtained. When the time came to rebuild the machine, only one side of the bearing interface generally had to be refinished. Hence cast iron on steel has largely replaced cast iron on cast iron as a sliding contact bearing. Many machine tool carriages use this combination, particularly where high stiffness and load-carrying capability are required in a small space such as in screw machines.

Material Combinations: Brass on Steel

Brass also has an inherent lubricity when in contact with steel. The powder metallurgy process can also be used to make the brass part porous, thus forming a reservoir for a lubricant. As the bearing heats up, the lubricant is brought to the surface, where it reduces the coefficient of friction and the heat generated; hence the system behaves in a closed-loop manner to regulate friction and heat. This type of bearing is usually used where high-speed reciprocating motion is encountered. It is also very dirt resistant and is often used in articulated joints of nonprecision machinery (e.g., a backhoe).

Material Combinations: Polymers on Most Anything

Polymer based [e.g., PTFE (polytetrafluoroethylene)] bearings have been engineered to minimize stick-slip (typically static μ within 10-20% of dynamic μ) and thus have virtually replaced metal-on-metal bearings in most machine tool applications. Figure 8.2.1 shows how load and speed affect various properties of a commonly used sliding bearing material (i.e., Turcite®[4]) frequently used in machine tools. Tables 8.2.1 - 8.2.3 give the properties of commonly used polymer-based bearing materials. This material commonly is used in the form of a thin sheet a few millimeters thick that is bonded to the machine carriage. After bonding, the bearing surfaces are often scraped to achieve maximum accuracy and impart good oil retention characteristics into the surface. In order to maintain good quality control, careful surface preparation is required, although once the process is perfected, the bonds usually last for the life of the machine. It is very important to keep bonded sliding bearings free of dirt, however, because particles can cause the bonded layer to tear and then fail catastrophically.



Maximum load	14 MPa
Limiting pressure x velocity (PV)	53 Mpa x m/min (25,000 psi x fpm)
Minimum operating temperature	-200°C
Maximum operating temperature	280°C
Coefficient of friction	From 0.02 to 0.2 (depending on unit load and lubrication)

Figure 8.2.1 Properties of Turcite® sliding bearing material. (Courtesy of W. S. Shamban & Co.)

[4] See Turcite® design information from W. S. Shamban & Company, Newbury Park, CA.

Maximum load	
Normal	140 N/mm^2
Special circumstances	250 N/mm^2
Compressive yield strength	310 N/mm^2
Maximum rubbing velocity	2.5 m/s
Specific load x rubbing velocity (PV factor)	
Continuous	1.75 N/mm^{2}x m/s
Short periods	3.5 N/mm^{2}x m/s
Minimum operating temperature	-200^{o}C
Maximum operating temperature	280^{o}C
Coefficient of friction (unlubricated)	From 0.02 to 0.2, depending on load
Electrical resistance	1 to 10 ohms/cm^2
Nuclear radiation resistance	Unaffected by gamma-ray dose of 10^8rad

Table 8.2.1 Properties of Glacier DU® bearing material available as bushings or in sheet form. (Courtesy of The Glacier Metal Company Ltd.)

Figure 8.2.2 shows the frictional characteristics of a castable bearing material that is applied by the process of replication (see Section 7.5.2). One surface must be very rough (as shown in Figure 7.5.16), while the other surface should be smooth and covered with mold release. To prevent porosity, the replicant must be thoroughly vacuum degassed before application. Replication greatly decreases the cost of bearing surfaces for the following reasons:

- Only the replication master needs to be accurately finished.
- An exact fit is obtained, so gibs are not needed.
- The machine is easily rebuilt.
- Replicated bearings are less sensitive to dirt and they generally do not have to be scraped.

Material Combinations: Almost Anything on a Ceramic

Ceramic materials can be made harder than any plastic or metal and finished smoother without worry of surface degradation due to oxidation. In many ways, ceramics seem to make ideal bearing surfaces. Typical ceramic bearing materials include Zirconia, silicon nitride, silicon carbide, Zerodur®, and sometimes aluminum oxide. The material that slides on the ceramic surface should generally have inherent lubricity like a polymer bearing. Note that because ceramics are brittle, after finishing their surfaces generally have a negative skewness, so a ceramic can potentially slide on a ceramic for a very long time with negligible wear and low friction. In addition, since ceramics are brittle, residual stresses are not easily built up in their surfaces during finishing and hence it is easier to grind or lap a ceramic bearing rail flat than it is a steel one. The obvious drawbacks of ceramics are their difficulty to manufacture and their cost relative to cast iron or steel. Nevertheless, some applications warrant the use of a ceramic such as coordinate measuring machine bearing rails or precision hydrostatic bearing rails. An example is the National Physical Laboratory's Nanosurf 2 machine, which used PTFE pads on an inverted polished Zerodur® vee to achieve nanometer smoothness of motion (see Section 8.2.3).

A common type of ceramic bearing application is the use of a jewel (e.g., sapphire or ruby) as a bearing surface. When forces are low, jewel bearings are a type of sliding contact rotary bearing that have been used in precision instruments for centuries. Although the coefficient of friction is

Compressive strength	96.5 MPa
Shear strength	31.7 MPa
Compressive modulus	4.0 GPa
Shear modulus	0.8 GPa
Shrinkage	400 micron/m
Coefficient of thermal expansion	42.5 micron/m/C^o
Lubricated static coefficient of friction	Approx. 0.11
Lubricated dynamic coefficient of friction	Approx. 0.09
Density	2.1 g/cm^3
Pot life	45 minutes
Cure time	40 hours

Table 8.2.2 Properties of a typical castable high lubricity polymer. (Courtesy of ITW-Philadelphia Resins.)

Specific weight	1.6 g/cm^3
Dynamic strength	1450 N/cm^2
Static strength	14,000 N/cm^2
Minimum operating temperature	-40°C
Maximum operating temperature	125°C
Shrinkage	About $^1/_4$%
Moisture absorption	Very good resistance

Table 8.2.3 Properties of Moglice® high lubricity castable bearing replication material. (Courtesy of "DIAMANT" Metallplastic GmbH.)

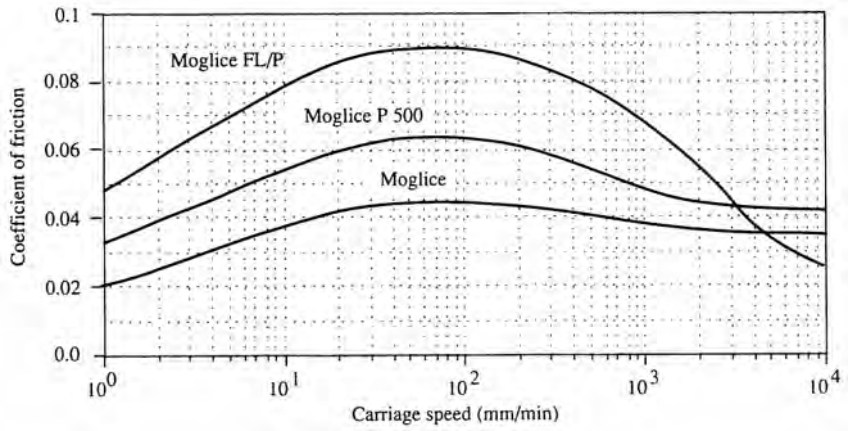


Figure 8.2.2 Frictional properties of Moglice® castable bearing material: contact pressure 5 daN/cm^2 with mineral oil having a viscosity of 25 centistokes at 50°C. (Courtesy of "DIAMANT" Metallplastic GmbH.)

high compared to ball bearings, the bearing region is only a point; hence the radius the bearing friction force acts over is orders of magnitude less than with a ball bearing and thus so is the friction torque. Jewel bearings can be finished to a very high degree and have a high compressive strength and modulus, which enable them to be used in this manner.

Speed and Acceleration Limits

Because the coefficient of friction of a sliding contact bearing is generally higher than that for a rolling contact or fluidstatic bearing, for machine tool applications, sliding contact bearings are usually used only where the maximum speed and acceleration levels are less than on the order of 0.25 m/s (600 in./min) and 0.1g, respectively. Note that a sliding contact bearing can be used as a backup bearing for hydrostatic or aerostatic bearings (e.g., plastic or graphite pads) in case fluid or air pressure should fail.

Range of Motion

Linear motion sliding contact bearings can have as long a range of motion as it is possible to machine the ways they ride on. Bearing ways can be made in sections and then spliced together to make rails tens of meters long. Angular rotary motion sliding contact bearings are not rotation limited.

Applied Loads

Because sliding contact bearings typically distribute the load over a large area, huge loads can typically be supported. The exception of course is when a kinematic arrangement of bearings with rounded contact surfaces are used to support an instrument platten. In general, surface contact pressures are typically less than about 1 MPa (150 psi), but can be as high as 10 MPa.

Accuracy

The accuracy of a sliding contact bearing depends greatly on the type of bearing, the accuracy of the parts, and on the contact pressure and lubricant type. Typically, in a machine tool after wear-in, a linear motion sliding contact bearing can yield straight-line accuracies on the order of 5-10 μm

when the surfaces are ground, and submicron accuracy is possible if the surfaces are hand finished. When the surfaces are prepared as discussed in Section 8.2.3, accuracy of motion (normal to the direction of travel) can be on the order of nanometers. Since sliding contact bearings typically have a high coefficient of friction (on the order of 0.02-0.1), the servo-controllable accuracy of the supported axis depends greatly on the axial stiffness and controllability of the system. Assuming that everything else in the system is perfect, the typical servo-controlled accuracy of a heavily preloaded system can be as high as 5 μm. Lightly preloaded systems (e.g., wafer steppers and instruments) can achieve submicron start and stop accuracy.

Repeatability

One nice aspect of most sliding contact bearings is that they wear-in with use, but in order to prevent abrasion, the surface finish is still as critical as it is with rolling contact bearings. As the preload is maintained, repeatability often increases with time and can typically be in the range 0.1-1.0 μm if the bearing design is not overconstrained. Assuming that everything else in the system is perfect, the typical servo-controlled repeatability of heavily preloaded systems can be as high as 2 μm. For example, ruling engine carriages for diffraction gratings were typically supported by five sliding contacts on vee and flat ways and achieved submicron repeatability of line widths. As discussed in Section 8.2.3, angstrom repeatability is possible for specially designed and manufacturer sliding contact bearings.

Resolution

As long as the platten of a machine is kept moving (e.g., during a contouring cut), resolution of relative motion between axes or steadiness of motion (constant velocity) can be as good as for many rolling bearings. For start and stop moves, the inherent friction in most sliding contact bearings often limits their resolution to about 2-10 μm for carefully designed well-worn-in bearings. Many PTFE-based bearings have very nearly equal static and dynamic coefficients of friction (within 10% at low velocity); hence systems supported by these bearings can typically readily achieve submicron and better resolution of motion.

Preload

Sliding contact bearings generally have a large contact surface area, so the load deflection curve is essentially linear once the bearing is worn in. Thus for precision applications the preload is typically only on the order of 5-10% of the allowable load. Where precision is not required, a light sliding fit is all that is needed. High preloads may be required to achieve desired stiffness levels, and for low-speed motion the resulting friction force may be acceptable. If the same machine must make high-speed positioning (i.e., rapid traverse) moves, an air or oil assist may be used to relieve the load on the bearing and hence reduce the drag force. This entails providing pockets in the bearings: One pocket is required for pressurized fluid to take the weight of the machine and preload off the bearing, and often a second, smaller pocket is required on the opposing pad to relieve the opposing preload force.

Stiffness

As shown in Figure 8.2.3,[5] sliding contact bearings can be exceptionally stiff because the surface area can be so large. The highest stiffness will occur when the bearing is worn-in. If the design is not kinematic, the preload must be sufficient so that all parts of all bearing pads are in contact with their respective mating surfaces.

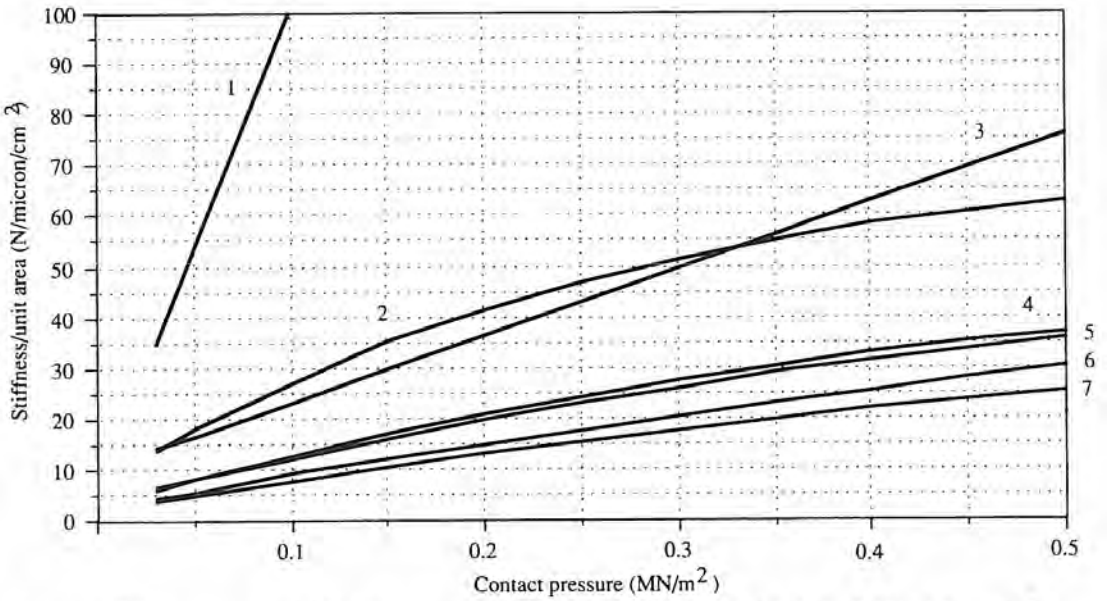
A captive preloaded bearing is one where one bearing pad pushes against another, such as in the bearing pads of a rectangular (e.g., T shaped) or dovetail carriage design. As the force on a pair of opposing pads preloaded against each other is applied, one pad pushes harder on the rail by an amount equal to the product of the pad stiffness and the carriage deflection, and one pushes less by the same amount. A simple free-body diagram of the system shows the sum of the forces to be

$$F_{load} - (F_{preload} + K_{upper\ pad}\,\delta) + (F_{preload} - K_{lower\ pad}\,\delta) = 0 \tag{8.2.1}$$

From Equation 8.2.1 and the relation $F_{load} = K_{total}\,\delta$, the total stiffness of the bearing pad set is

$$K_{total} = K_{upper\ pad} + K_{lower\ pad} \tag{8.2.2}$$

[5] From M. Dolbey and R. Bell, "The Contact Stiffness of Joints at Low Apparent Interface Pressures," Ann. CIRP, Vol. 19, pp. 67–79.



#	Way	Bearing	Stiffness/unit area (k/a) as a function of contact pressure P
1	Lapped cast iron	Ground cast iron	k/A = 6.2105 + 9.5583e-4P
2	Ground cast iron	Ground cast iron	k/A = 7.2247 + 2.3291e-4P - 3.4428e-10P^2 + 2.0422e-16P^3
3	Ground cast iron	Scraped cast iron	k/A = 10.129 + 1.3275e-4P
4	Ground cast iron	Ground ferobestos	k/A = 3.4536 + 9.7855e-5P - 5.5217e-11P^2 - 1.0544e-17P^3
5	Ground cast iron	Ground plain DX	k/A = 3.8624 + 8.7409e-5P - 3.7974e-11P^2 - 1.8380e-17P^3
6	Ground cast iron	Ground dimpled DX	k/A = 2.1344 + 7.5302e-5P - 5.8759e-11P^2 + 4.2540e-17P^3
7	Ground cast iron	DU as received	k/A = 2.1064 + 6.1535e-5P - 3.0397e-11P^2 + 2.2446e-18P^3

Figure 8.2.3 Constant stiffness of various sliding contact bearings lubricated with light oil and after wear-in. (After Dolbey and Bell.)

As the load increases, the stiffness of one pad increases while the other one decreases, due to changing contact pressure. When the applied force equals the preload force, one pad loses contact and the total pad set stiffness is just that of the single pad in contact. Since the preload is a small fraction of the total maximum load the bearing can expect, a conservative assumption is to assume that the stiffness of a pad set is just the stiffness of a single pad.

Vibration and Shock Resistance

When sliding contact bearings have applied loads distributed over a large surface area, their shock and vibration resistance is virtually unmatched by any other type of bearing except hydrostatic or hydrodynamic bearings.

Damping Capability

Sliding contact bearings' damping capability is equaled only by hydrostatic or hydrodynamic bearings. This is due to the viscoelastic nature of many types of the sliding bearing interfaces and the large contact surface area, which often is coated with a thin layer of oil or grease which provides squeeze film and viscous damping.

Friction[6]

Sliding contact bearings usually have an appreciable amount of static friction associated with their use. Once motion begins, the dynamic friction level is usually slightly lower than the static friction, which results in a step rise in acceleration. This condition is often referred to as *stick-slip* or *stiction*. As the speed increases, the friction coefficient continues to decrease as a hydrodynamic (for lubricated bearings) fluid layer builds up. Soon a transition point is reached where viscous drag begins to increase the net sliding coefficient of friction. This overall behavior is represented by the *Stribeck curve*. Stiction is very pronounced in metal-on-metal bearings but it is less pronounced

[6] The mechanism of friction is complex and not fully understood. Numerous volumes of research describe studies of dynamic and static friction coefficients, but no clear understanding exists, and to present all opposing viewpoints might confuse the reader even more than the author.

in modern PTFE-based bearings that have been worn-in. Some manufacturers claim that stiction is nonexistent and they prove it by supplying data which shows that the friction level at very low speeds (e.g., 10^{-6} m/s) equals the friction coefficient at moderate speeds (10^{-1} m/s); however, what is important is the friction at 0 m/s.

It is very easy to show that a sliding bearing's static friction coefficient is higher than the dynamic friction coefficient. All one needs is a sample of the bearing material mounted on the bottom of a carriage, and a bearing rail whose angle of inclination can be finely adjusted. When the inclination angle is zero, the carriage will not slide off the rail. As the angle is increased to a few tenths of a degree, a dial indicator will show whether or not the carriage is sliding (even over a period of hours). If the carriage does not move, then the static coefficient of friction is higher than the dynamic coefficient of friction. As the rail's inclination angle is slowly increased, there will come a point where the carriage begins to slide. This is the transition between the static and dynamic regimes of friction. The static friction will be equal to the tangent of the angle of inclination. The dynamic coefficient of friction can be accurately determined only if the velocity of the carriage is also measured.

For PTFE-based bearings, with a wear-in period of a few hundred to a few thousand cycles, the static coefficient of friction may decrease from a high of about 0.3 to about 0.03-0.1 The dynamic coefficient of friction of worn-in sliding contact bearings may be on the order of 0.02-0.1. The coefficients of friction can also depend on the applied load, surface finish, and bearing design (see Section 8.2.3). For example, as shown in Figure 8.2.1, one manufacturer shows that the coefficient of friction decreases with applied load. In this case in order to minimize the dynamic coefficient of friction for maximum contouring accuracy (as long as reversal does not take place where the velocity goes from plus to minus), it makes sense for the preload to be at least 10% of the rated load. On the other hand, when minimal starting force is required, the preload should be minimized. If high speeds are to be encountered (>0.5 m/s), one may want to consider the use of a rolling element or fluidstatic bearing.

Because of the presence of some stiction in all sliding contact bearings, one must be careful when designing a servo *system*. For example, imagine a typical machine tool carriage driven by a leadscrew. If a linear scale is used as a feedback device, when the axis is first given a move command the static friction force prevents the axis from moving for the first few servo cycle times. As a result, the integrator in the servo algorithm rapidly starts increasing the force being applied.[7] The leadscrew will actually turn, but the displacement generated goes into compressing and twisting the leadscrew shaft. When the carriage does start moving and the friction level suddenly drops as it switches from static to dynamic friction, the elastic energy stored in the leadscrew shaft and the high rapidly increasing force level can cause the carriage to lurch forward. On the other hand, if a resolver or encoder connected to the leadscrew was used, as soon as the move command was given, the encoder would be reading the rotational motion in the leadscrew that was causing the leadscrew to be compressed as it tried to overcome the static friction; hence the integrator would not cause the force to rise as rapidly and the leadscrew would not have stored as much energy when the transition from static to dynamic friction occurred and there would be less of a lurch forward. The leadscrew and encoder act to filter out the dynamic effects of the controller, which would otherwise make control of the high-friction sliding bearing system more difficult. Overall, some design engineers like linear scales because it makes the accuracy of the leadscrew less critical as long as there is no backlash. Other design engineers like encoders or resolvers because of the problem described above. Fortunately, as controller and bearing technology advances, stiction is becoming less and less of a problem.

Thermal Performance

A worn-in sliding bearing's frictional properties will generally change with temperature if the viscosity of the lubricant used changes with temperature. Self-lubricating bearings (e.g., oil-impregnated bronze) release lubricants to the interface when the bearing temperature rises, thereby lowering the coefficient of friction and the heat generated by motion at the bearing interface. If a lubricant is chosen whose viscosity does not change with the bearing operating temperature range, then after the warm-up period, a worn-in sliding contact bearing's dynamic performance (servoabil-

[7] The integrator is used to eliminate steady-state error. Servos with tight velocity loops and feedforward loops have essentially zero steady-state error, and they are less susceptible to the phenomenon of windup.

ity) will be consistent. As noted in Chapter 6, where a machine tool's geometric and thermal errors were mapped, heat generated by the relatively slow moving linear bearings was not as significant as that from motors and high-speed rotary joints (e.g., the spindle and ballscrew). It was also shown that the warm-up time for a small machine tool (e.g., 4000 kg) could be as long as 12 h.

Because the surface area of sliding contact bearings is so large, they have the potential to transmit heat better than do rolling contact bearings. If forced lubrication or a plastic or graphite bearing pad is used, however, then the bearing can act as an insulator. Because of all the variables (material types, surface finish, contact pressure, lubricant type, etc.), it is not possible to generalize further.

Environmental Sensitivity

One interface surface of a sliding contact bearing should always be made harder than the other, so if dirt gets into the bearing, it will become embedded in the softer part. This prevents the dirt from wandering around in the bearing and causing continual wear. For this reason sliding contact bearings can be far more dirt resistant than rolling element bearings. Since the bearing rail (way) geometries required for a sliding contact bearing are usually simple geometric shapes (e.g., a rectangle as opposed to arced grooves for some rolling contact ball bearings), they are simple to seal with wiper-type seals and often do not require bellows or way covers. Since contact stresses are low at the bearing interface and dissimilar materials are used (e.g., cast iron or Teflon on steel), a simple surface treatment on the bearing rails (e.g., carburizing or nitriding) can be used to enhance the corrosion resistance of tool steel ways should moisture find its way into the bearing or the lubricant supply.

Sliding contact bearings can generate particulate matter which can be detrimental to the performance of a clean room environment. Care must be taken in designing sliding contact bearings for clean room applications. Even if the bearing is not expected to generate particles, the high cost of clean rooms may warrant the use of a fluid lubrication system to trap particles and a sealing method to contain the fluid. In some instances a ferrofluid with a magnetic seal may be used.

Seal-ability

As mentioned above, even though sliding contact bearings are generally environmentally tough, it makes no sense to tempt fate. In all but some instrument applications, at least wipers should be used on linear bearings and wiper-type seals on rotary bearings. Since the bearing geometries are usually so simple, sealing sliding contact bearings is generally not a problem.

Size and Configuration

By their nature, sliding contact bearings take up less space than do almost any other type of bearing. For machine tool applications, lubrication lines often need to be provided. Most machine tool applications of sliding contact bearings are linear. For large revolute joints, such as those that support the spindle assembly on five-axis machining centers, curved rails and sliding contact bearings can be used.

Weight

Because of their simplicity, sliding contact bearings have the highest performance-to-weight ratios.

Support Equipment

Of all the different types of bearings, sliding contact bearings are the least sensitive to abuse; however, for a precision machine, tender loving care is always in order, and thus sliding contact bearings should usually be installed with an automatic lubrication system.

Maintenance Requirements

Sliding contact bearings for precision applications generally require an automatic lubrication system, and because they rely more on a wear-in period to achieve good steady-state performance, they should be designed with a gib to allow the preload to be adjusted easily. Note there are some high-precision low-force applications where a PTFE material slides on polished surfaces and no liquid lubricant is used.

Material Compatibility

Sliding contact bearings have high stiffness to resist cutting forces but use low preloads to minimize starting forces; hence a small amount of differential thermal growth between bearing components can lead to loss of preload and in some cases opening of a bearing gap. Fortunately, sliding contact bearings also generally have good heat transfer characteristics, so as long as the design maintains preload in the presence of uniform temperature changes,[8] it should perform well as the machine warms up. Of course for the final design, a more detailed finite element analysis or a model test should be completed to see how the preload varies in the presence of differential temperatures.

The surface finish of the interface components is very crucial to proper wear-in and long-term life. In general, one wants a surface finish on the order of 0.1-0.5 μm with a random pattern on the hard surface the bearing rides on. This provides fine grooves in which a lubricant can be trapped. Rougher finishes tend to cause the asperities to drag like grappling hooks, and finer surfaces cannot support a lubricating layer.[9] The better the surface finish of the bearing material, the less time that is needed to wear-in the bearing and the less adjustment that is needed after wear-in.

Required Life

A review of various manufacturers' catalogs shows that when properly lubricated and not overloaded, after wear-in sliding bearings may wear at a rate of about 10^{-11} m/meter of travel. Some PTFE bearings that ride on surfaces with nanometer-level surface finishes (so they are nonabrading) are thought to build up a film on the bearing way which transfers back and forth between the bearing pad, leading to essentially zero wear (see Section 8.2.3). The Glacier Metal Company has found that performance of PTFE-based bearings is increased substantially when the PTFE is impregnated with lead and bonded to a porous bronze layer. The porous bronze is bonded to a steel layer, which can be bonded to the machine. This design is referred to as a DU® bearing, and it makes for a very tough bearing which virtually never experiences scoring, tearing, and then delamination even in the most abrasive of environments. Glacier DX®bearings consist of an acetal polymer bonded to a porous bronze layer that is bonded to a steel backing. DX® bearings are designed to operate with marginal lubrication, and often they have a dimpled surface which holds a lubricant in place. UHMW plastics (e.g., ultrahigh molecular weight polyethylene) are also commonly used in abrasive environments (e.g., coal bin liners) and bearings are available.

It used to be that machines were designed for a life of 5-10 years, with the intent that they could then be rebuilt over and over. Ideally this should still be the case, but for some markets where machine cost is a primary selling point, it pays to use the least expensive components that are available and then plan on rebuilding the machine more often. It also used to be that sliding contact bearings often had to be hand scraped, which made them expensive, but modern accurate surface grinders make it possible to, in many instances, just assemble ground components and attain reasonable repeatability (on the order of 5-10 μm) after wear-in and readjustment of the gibs.

Availability

Modular sliding contact bearings or materials for custom manufacture are among the easiest items to procure.

Designability

Sliding contact bearings influence the design of a machine by often requiring the use of a gib and an automatic lubrication system. In general, sliding contact bearings are very easy to design.

Manufacturability

The comments for designability apply here as well.

Cost

The cost of sliding contact bearing materials themselves are negligible, and the cost of an automatic lubrication system is moderate. The principal cost associated with precision sliding contact bearings is that of attaining the desired accuracy and surface finish of the bearings and the surface they slide on. For example, it also takes considerable skill to make a tapered gib.

---

[8] See Section 8.9.

[9] See, for example, H. Moalic et al. "The Correlation of the Characteristics of Rough Surfaces with Their Friction Coefficients," Proc. Inst. Mech. Eng., Vol. 201, No. C5, and A. Hamouda, "An Investigation of Some Factors Affecting Stick-Slip Mechanism in Relation to Precision of Sliding Components," IMEKO-Symp. Meas. Estimat., Bressanone, 1984.

8.2.2 Design Considerations

The best way to assimilate the thoughts presented above is through design considerations. Naturally, they are biased toward precision machines, and as noted, Section 8.2.1 should be reread after reading this section.

Bearing Spacing

For rotary bearing systems, it is simple task to calculate bearing forces and thus choose bearing size and spacing. For linear bearing systems, similar calculations can be made to size bearings and their spacing to yield desired load and stiffness capabilities. In order to prevent a linear bearing system from walking (yawing or pitching during low forward velocities), assuming that the actuation force is in line with the center of mass, the ratio of the spacing of bearing pads should be on the order of 2:1 length to width. A rule of thumb for a lower bound ratio is to make the bearing pad centers lie on the perimeter of a golden rectangle,[10] with the long side along the direction of motion. The absolute minimum length-to-width ratio is 1:1. For large machines where this is not possible (e.g., some moving bridge designs) two drive systems are required, with one slaved to the other. In general, the higher the speed, the higher the coefficient of friction, or the higher the moment loads on the system, the higher the length-to-width ratio should be. With respect to the kinematics of motion of a carriage assembly kinematically supported by sliding bearings, refer to the generalized analysis presented in Section 2.2.4 and illustrated in Figure 2.2.7. This analysis can be extended to nonkinematic systems as long as the appropriate spring constants and geometries are used.

Because sliding contact bearings are often so stiff and their coefficient of friction is relatively high compared to rolling bearings, preload forces are usually kept low; however, if a slight amount of wear or material relaxation takes place, the preload can be lost. Hence the configuration of the bearings is most important. Often the amount of preload required for a sliding contact bearing corresponds to a deflection of only a few microns in the structure; therefore, rather than trying to make all parts fit with the precise interference, a gib is used, as discussed below.

Automatic Lubrication Systems

For precision machine applications where loads greater than about 100 N are encountered, sliding contact bearings should be lubricated. As shown in Figure 6.1.8, this is usually accomplished by cutting oil distribution grooves in the bearing pad. Lubrication can be supplied continuously or periodically through a timer-controlled pump. One should not rely on the maintenance man to grease the bearings periodically or to push a lever to give them a squirt of oil.

Gib Design

In order to allow the preload to be adjusted easily, particularly after the wear-in period has been completed, a machine element called a *gib* should be used with linear sliding contact bearings. Four common types of gibs are shown in Figure 8.2.4. Straight gibs are preloaded using setscrews along their length or with a roller. Tapered gibs are preloaded by a single screw that pushes them along a matching taper, thereby creating uniform lateral displacement along the length of the gibs. The former is easy to manufacture, while the latter requires more complicated machine setups and considerable skill if it is to be hand scraped. Still, tapered gibs provide the highest degree of stiffness. Note that gibs for dovetails that are locked in place with retaining bolts can provide twice the stiffness to overturning moments than do tapered gibs.[11]

For the T carriage design shown in Figure 8.2.5, a straight gib is used to control the lateral clearance between the saddle and rail. Note that lateral motion is restricted by one rail only. To use both rails for lateral restraint would overconstrain the system, increase the likelihood that Poisson expansion of the bolted rails would create straightness errors, and increase manufacturing costs considerably. In order to design the gib, an estimate can be made of the deformations on the gib resulting from the setscrew pressure. As illustrated in Figure 8.2.6, assume that a single setscrew

[10] The golden rectangle was the basis for most Greek temple architecture. The ratio of the width minus the height and the height equaled the ratio of the height and the width, or $2/(5^{0.5}- 1) = 1.618$. Physically, this meant that if you begin with a rectangle and cut a square from it whose side length equals the short side of the rectangle, the remaining rectangle will be in proportion to the original rectangle. As far as bearing spacing is concerned, this is a pure rule of thumb and the author is not aware of a mathematical proof of why it works for bearings.

[11] See Z. Levina, "Research on the Static Stiffness of Joints in Machine Tools," Proc. 8th Int. Mach. Tool Des. Res. Conf., Sept. 1967, pp. 737–758.

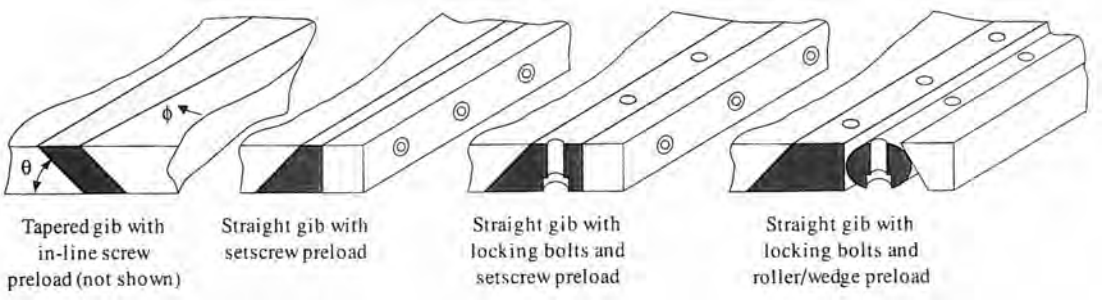


Figure 8.2.4 Some of the many types of gibs that can be used to preload bearings.

acts at the center of a plate simply supported around its edges, where the size of the plate is the size of the bearing pad, and the thickness of the gib (plate) is to be determined.[12]

Ideally, only one setscrew per bearing pad region is used, or else uneven tightening of the setscrews can cause uneven wear of the pad and render the analysis too inaccurate and too difficult. The thickness t_{gib} of the gib should be chosen so that the deflection is one-half the desired repeatability δ of the system:

$$t_{gib} = \left(\frac{2\alpha\eta a b^3 P_{max}}{\delta E} \right)^{1/3} \tag{8.2.3}$$

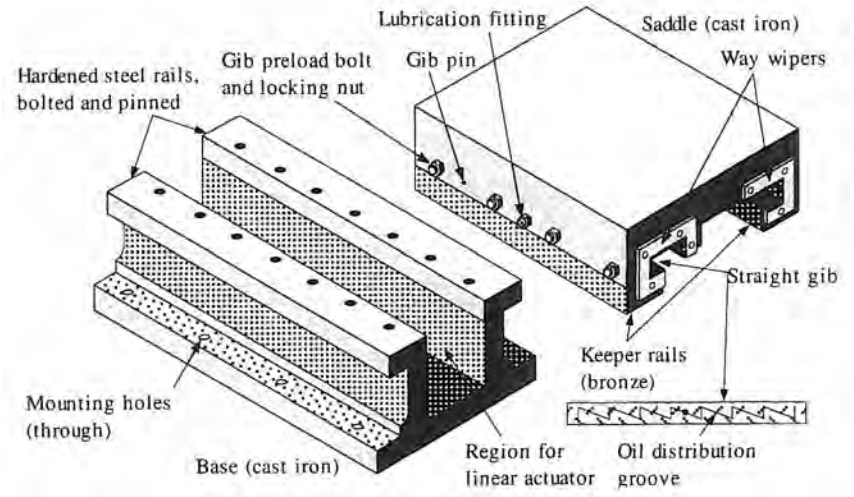


Figure 8.2.5 Components of a tee-shaped slide. Note the straight gib and way wipers. (Courtesy of Setco Industries Inc.)

For example, assume that a brass gib is used and E = 110 GPa (16×10^6 psi), $\delta = 10^{-6}$ m, a = 0.1 m, b = 0.05 m, P_{max} = 0.5 MPa (75 psi), and η = 0.3; then t = 0.0183 m (0.72 in). Note that the force on the gib is applied entirely through the setscrew, so the stiffness of the bearing pad can be no greater than the stiffness of the setscrew and its thread and contact interfaces, and thus the motivation for using more setscrews. If the space between the gib and carriage were potted with epoxy after the machine was worn-in, then the stiffness would be greatly increased. Whenever the gibs need to be adjusted for wear, they could be removed, the epoxy cleaned out, and then the gibs reinstalled, adjusted, and epoxy potted.

A tapered gib acts like a wedge; hence it is supported along its full length and can be much thinner than a setscrew preloaded straight gib and provides a much greater degree of stiffness. This minimizes the amount of overhang of the keeper rail and helps to minimize machine size. Tapered

[12] The formula and values for a as a function of a/b are from R. J. Roark and W. C. Young, Formulas for Stress and Strain, 5th ed., McGraw-Hill Book Co., New York, 1975. This is one of the books that you are supposed to have as a reference.

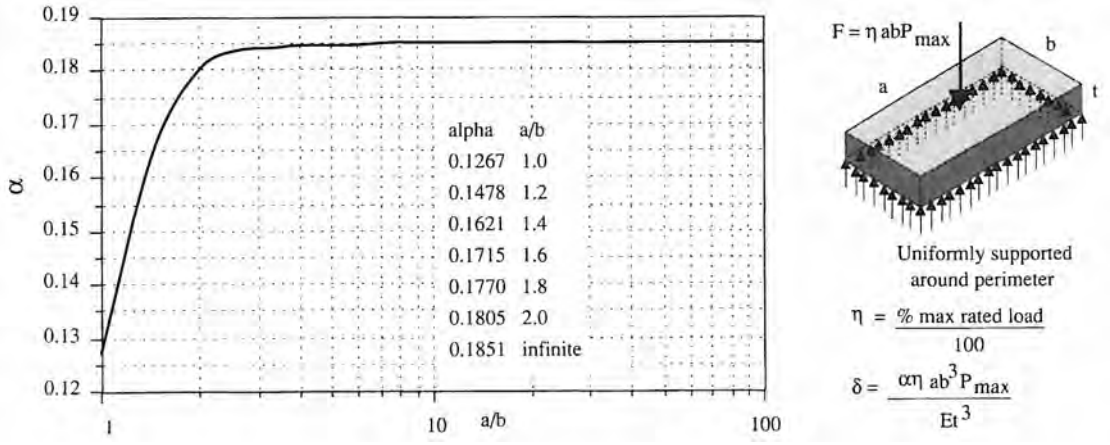


Figure 8.2.6 Model for calculating straight gib thicknesses.

gibs usually require them to be hand scraped to achieve performance commensurate with their design. If manufacturing personnel are not adept at scraping, then tapered gibs should be avoided. An alternative is to use a roller wedge and a straight gib, but this takes up more room than a tapered gib.

Rotary Bearing Configurations

Sliding contact rotary bearings are not often used in precision machines because ball bearings are so inexpensive and easy to use. However, sliding contact bearings may be needed where high loads may be encountered, or extreme accuracy requirements warrant lapping a shaft and sliding contact bearing to fit each other perfectly. Figure 8.2.7 shows various configurations for sliding contact rotary bearings. Note that various manufacturers make similar shaped bearings from a large variety of materials. Linear sliding contact bearings are also made by a number of manufacturers and are commonly used in applications not hostile to high-performance polymers (e.g., consumer products and industrial machinery, including earth moving equipment). At very high or low operating temperatures where lubrication cannot be provided, metal-impregnated graphite modular sliding contact rotary bearings are often used.[13]

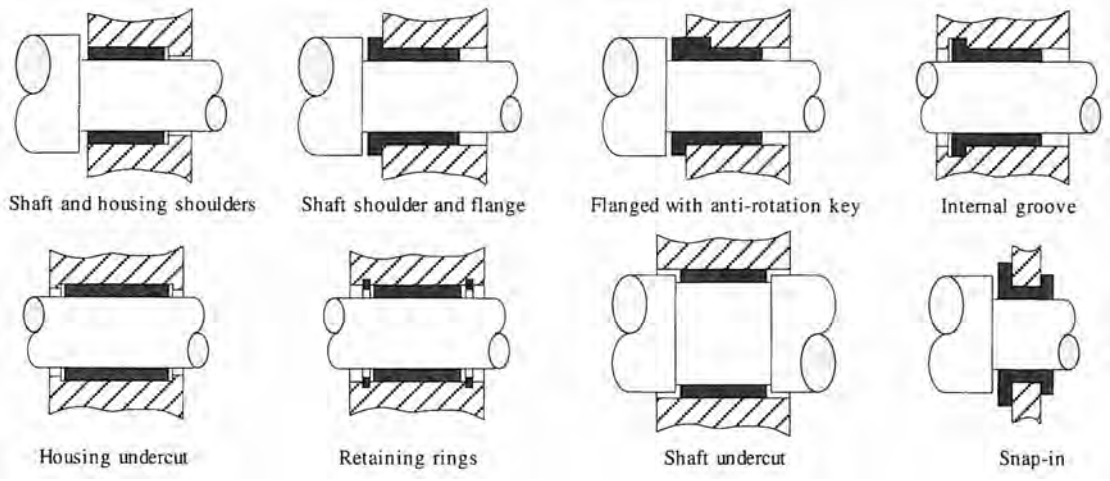


Figure 8.2.7 Various configurations for rotary sliding contact modular bearings. (Courtesy of Thomson Industries, Inc.)

For very high precision applications, such as instruments, or consumer product designs, such as a mouse trackball, where the friction torque is to be minimized, one should consider using a jewel bearing (synthetic sapphire or ruby). Figure 8.2.8 shows various available jewel shapes. The high

[13] Graphite Metallizing Corp., Yonkers, NY.

surface finish, high compressive strength, high modulus, and chemical inertness of a jewel bearing usually ensures dimensional stability. These properties also allow jewel bearings to be used with a small-diameter or pointed shaft, thereby minimizing tare torque. Detailed design considerations for jewel bearings are provided by various manufacturers.[14]

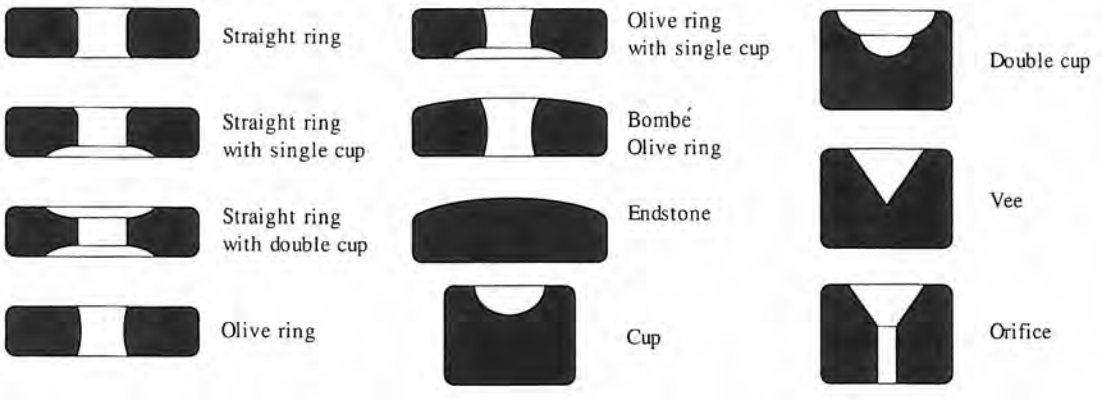


Figure 8.2.8 Standard jewel bearing designs. (Courtesy of Swiss Jewel Co.)

Closed Rectangular Linear Bearing Configuration

A rectangular linear sliding contact bearing configuration is shown in Figure 8.2.9 with four pads on the top, four pads on the bottom, and two on each side. It can also be made where the entire interface region acts like a bearing pad. Either way, it is an overconstrained design, but if well made and preloaded, it can provide exceptional stiffness and damping capability. Note that the full wraparound design of the carriage means that the bearing rail is subject to bending loads. Hence this design is intended primarily for long strokes where accuracy is not required and weight is to be minimized, or for short strokes where the bearing rail is stiff enough to prevent large deflections. The rails can even be curved to allow for travel along a curved path.[15] Since the bearing is preloadable, stiffness can be very high in all directions. Wear among the bearing surfaces is usually symmetric, so alignment of the carriage does not change much with wear.

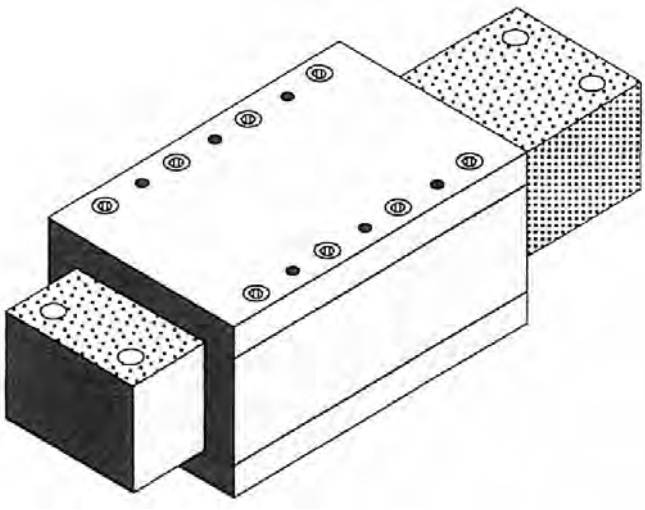


Figure 8.2.9 General configuration of a rectangular linear motion sliding contact bearing. Bearing surface may be composed of pads or be the entire interface (use of gibs not shown).

Open Rectangular (T) Linear Bearing Configuration

[14] See, for example, product literature by Swiss Jewel Co., Philadelphia, PA, and Bird Precision, Waltham, MA.
[15] See, for example, literature from Precision Laminations Inc., Rockford, IL.

A typical open rectangular (T-shaped) sliding bearing configuration was shown in Figure 8.2.5. Some other variations are shown in Figure 8.2.10. T configurations are the most common type of sliding bearing configuration seen in machine tools. It is overconstrained but provides very high stiffness. Wear among the bearing surfaces is usually symmetric, so alignment of the carriage does not change much with wear. With careful design and manufacturing quality control, machines with 5 to 10 μm repeatability or better can be built using this bearing design. Figures 8.2.11 and 8.2.12 show typical off-the-shelf modular T-shaped bearing components and load capacities, respectively. It is very easy to add a leadscrew or piston actuator to one of these modular assemblies along with a position sensor and control system to yield a servo-controlled single-axis system. Hence these modular carriages, which are available from a number of manufacturers, are very popular with design engineers, especially those who design one-of-a-kind machines or machines that are only produced in small lots.

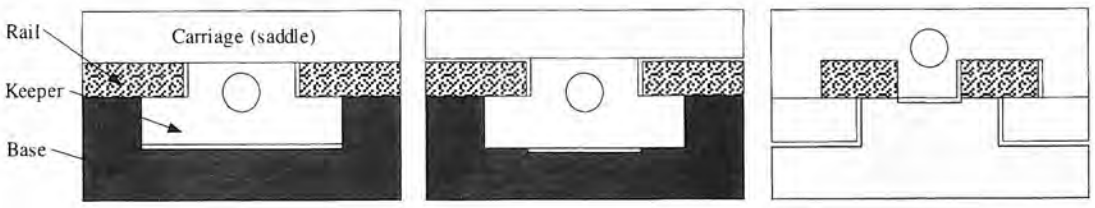


Figure 8.2.10 Some rectangular slideway configurations. How many other versions are there? Where would bearings of different types be put? Where could gibs be included?

Dovetail Linear Bearing Configuration

The dovetail shape can be preloaded and it is a better approximation to a kinematic configuration than is a rectangular or T shape. A typical dovetail bearing configuration is illustrated in Figure 8.2.13. Typically, there are four bearing pads on the top surface and two on each of the angled surfaces. If one bearing pad is on the top surface and two bearing pads are on each side, then the design would be kinematic. If the gib adjustment is provided at the one pad, then repeatability and accuracy would be a primary function of the parallelism of the angled sides.[16] This type of bearing is also available in off-the-shelf modular form, as shown in Figure 8.2.14. Since the dovetail is preloadable, stiffness can be very high in all directions. Wear among the bearing surfaces is usually symmetric, so alignment of the carriage does not change much with wear.

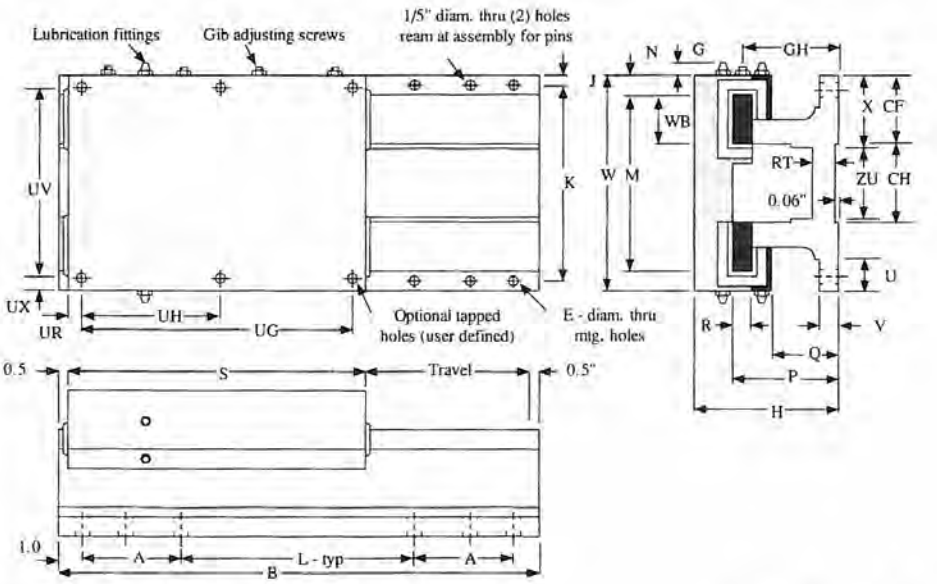
Vee and Flat Linear Bearing Configuration

The vee and flat is a true kinematic bearing configuration, as illustrated and discussed in detail in Section 2.2.4. A vee and flat supported carriage is generally preloaded by the weight of the machine, so it is often acceleration and speed limited. It is possible to preload a vee and flat supported carriage with a friction drive roller, but then the preload force on the bearings becomes a function of carriage position. Vee and flat supported carriages are also prone to walking problems should the center of mass of the system change dramatically, as when a heavy part is put on one side of the carriage. When the center of mass changes, the distance from the axial motion guiding the vee to the center of mass changes; hence the distribution of frictional forces changes, which causes a yaw motion on the carriage. This causes the carriage to yaw differently when different weights are applied at different positions.

Nevertheless, vee and flat designs are often used in machines where the loads are low and 1 μm or better accuracy and repeatability are required, such as in some precision surface grinders. A generalized analysis for vee and flats and other kinematic bearing configurations was presented in Section 2.2.4 and illustrated in Figure 2.2.7. Vee and flat bearing arrangements are very simple to build and are generally not available as off-the-shelf modular components.

Double-Vee Linear Bearing Configuration

[16] Imagine the slide with the angled side directions not parallel to each other. Then as the carriage moves axially, the three side bearings squeeze the rails, so as the rails converge, the carriage is lifted up and the three bearing pads on top lose their preload.



	SHL9	SHL12	SHL15	SHL18	SHL24	SHL32
CF	2.62	3.62	3.62	4.75	6.00	8.00
CH	3.75	4.75	7.75	8.50	12.00	16.00
E	9/16	9/16	9/16	11/16	11/16	11/16
G	0.38	0.50	0.50	0.50	0.62	0.62
GH	3.38	4.38	5.38	6.50	7.50	8.00
H	5.5	7.5	8.5	10.0	12.0	12.0
J	0.75	0.75	0.75	1.00	1.00	1.00
K	7.5	10.5	13.5	16.0	22.0	30.0
M	6.50	9.00	12.00	14.25	19.00	26.00
N	1.25	1.50	1.50	1.88	2.50	3.00
P	3.75	5.00	6.00	7.50	8.50	9.25
Q	2.25	2.75	3.75	4.50	5.50	5.25
R	0.75	1.25	1.25	2.00	2.00	2.50
RT	0.94	1.00	1.50	1.75	2.00	2.75
U	1.50	2.00	2.00	2.50	3.00	3.75
V	0.75	1.00	1.00	1.50	1.75	2.00
W	9	12	15	18	24	32
WB	1.50	2.50	2.50	3.50	4.50	5.00
X	2.88	4.19	4.19	5.62	7.25	8.25
ZU	3.25	3.62	6.62	6.75	9.50	15.50

B	A	L	F
18	8	-	6
24	5	12	8
30	8	12	8
36	8	12	8
42	11	12	10
48	11	12	10
60	11	12	12
72	11	12	14
84	11	12	16
96	11	12	18
108	11	12	20
120	11	12	22

Figure 8.2.11 Dimensions (inches) of a family of modular T slides. (Courtesy of Setco Industries, Inc.)

The double vee is quasi-kinematic: It is more kinematic than the dovetail but not as kinematic as the vee and flat; however, it has been found to be easier to manufacture than a dovetail, and a carriage supported by twin vees is less susceptible to walking than is the vee and flat when heavy off-center loads are applied. A double-vee arrangement is shown in Figure 8.2.15. There is always a vee to guide axial motion near the applied load, so yaw errors are not greatly influenced by changing loads or load position. Double-vee sliding bearing ways are generally not available as off-the-shelf modular components.[17] Note that the carriage and ways of a double-vee way system can both be made as females and the same master tools used to check them. Then hardened steel square rails can be bolted into the female ways to form the male ways for the carriage. Note that preload is provided by gravity, so axis accelerations are limited and the stiffness may not be as great as for preloadable systems. Wear between the two rails is usually symmetric, so alignment of the carriage does not change much with wear.

[17] A very thorough discussion of manufacturing considerations for twin-vee rails is given by W. Moore, Foundations of Mechanical Accuracy, published by Moore Special Tool Co., Bridgeport, CT.

Model	Width	Standard saddle lengths			Base lengths (1 " increments)		Load rating (lbf/in saddle length)			Approx. weight (lbf/in length)	
	(in)	(in)	(in)	(in)	Min.	Max.	Horiz.	Vert.	Sidewall	Saddle	Base
SHL9	9	9	13.5	18	12	120	50	25	15	6	4
SHL12	12	12	18	24	15	120	75	38	25	12	7
SHL15	15	15	22.5	30	18	120	80	40	27	15	10
SHL18	18	18	27	36	21	120	105	53	35	19	17
SHL24	24	24	36	48	27	120	140	70	45	32	24
SHL32	32	32	48	-	35	120	150	75	50	38	33

Figure 8.2.12 Load capabilities for a family of modular T slides. (Courtesy of Setco Industries, Inc.)

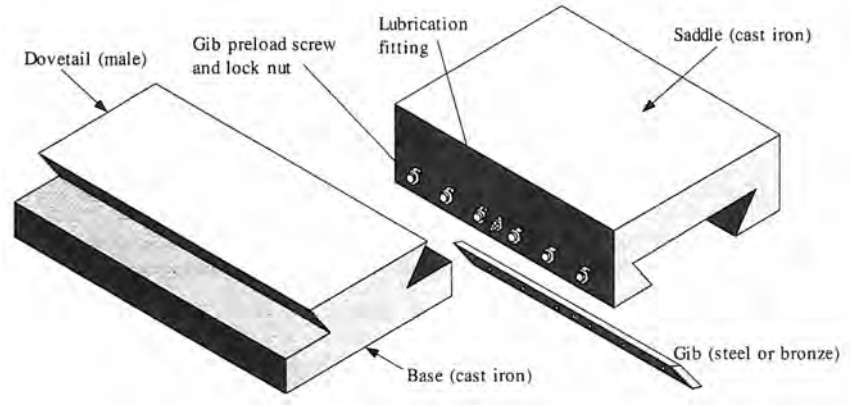


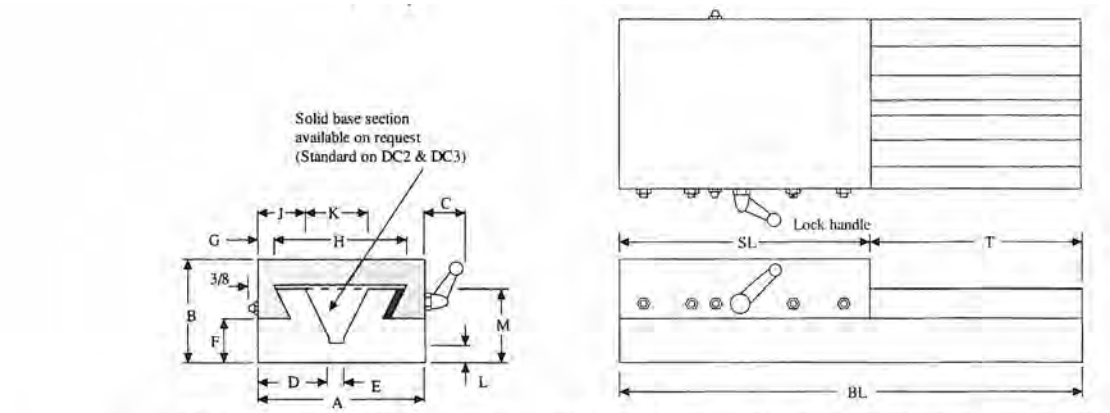
Figure 8.2.13 Construction of a modular dovetail slide assembly. (Courtesy of Russell T. Gilman, Inc.)

Thoughts on Design of Sliding Bearing Machine Ways

When designing linear bearings for a machine, one must also consider how the ways will be manufactured. It is during the design stage that the largest savings in cost and machine reliability can be realized. For example, the double vee seems like a simple design, but seven errors must be controlled in the manufacturing process: form, lean, center distance, vertical straightness and parallelism, and horizontal straightness and parallelism. Other bearing configurations are subject to similar errors. Thus one should always consider the following points when choosing a bearing design:

1. How does the design perform kinematically, assuming that it can be manufactured perfectly?
2. How do errors change when loads and their applied position change?
3. Is the design preloadable so that stiffness is high along and about all axes (except along the direction of motion)?
4. What happens to accuracy as the components wear? Does the position of the carriage merely shift the Cartesian offsets, or is angular alignment of the axes affected?
5. Can simple tests such as reversal of a checking master be used to check the accuracy of the manufacture and assembly process?
6. How difficult is the design to manufacture, and how easy is it to correct manufacturing mistakes?
7. Does the design require a gib?

Note that any linear bearing is usually used in the center of its range of motion; hence with use the carriage will slide easier in the center than at the ends; hence gibs cannot merely be tightened to take up slack because this will cause the carriage to bind near the ends of travel. When slack due to bearing rail wear appears in the most used range of travel, the entire bearing surface must be refinished or replaced.



Model	DC 2	DC 3	DC 4	DC 6	DC 8	DC 10	DC 12	DC 16	DC 20
SL: Min.	2	3	4	6	8	10	12	16	20
SL: Max. (1 in increm.)	12	18	36	36	36	36	36	36	36
BL: Min.	3	4	5	7	9	11	13	17	21
BL: Max. (1 in increm.)	36	48	60	96	96	96	96	96	96
A	2	3	4	6	8	10	12	16	20
B	1 3/8	1 5/8	2 1/4	3	4	5	6	7 1/2	9
C	3/8	3/8	1 1/4	1 1/4	1 1/4	1 11/16	1 11/16	2 3/8	2 3/8
D	-	-	1 11/16	2 3/8	3 1/4	4 1/4	5	6 1/2	8
E	-	-	3/4	1 1/4	1 1/2	1 1/2	2	3	4
F	5/8	3/4	1 1/8	1 1/2	2	2 1/2	3	3 3/4	4 3/4
G	5/16	13/32	31/64	3/4	1 3/32	1 13/32	1 45/64	2 27/32	3 19/32
H	1 3/8	2 3/16	3 1/32	4 1/2	5 13/16	7 3/16	8 19/32	11 1/16	13 9/16
J	-	-	1 15/64	1 41/64	2 13/32	3 3/32	3 45/64	5 1/8	6 5/16
K	-	-	1 17/32	2 23/32	3 3/16	3 3/4	4 19/32	5 3/4	7 3/8
L	-	-	11/16	23/32	31/32	1 3/8	1 1/2	2 1/4	2 3/4
M	1 1/64	1 9/64	1 5/8	2 1/8	2 5/8	3 1/2	4	5 1/4	6 1/4
Saddle weight (lbf/in)3/8	1/2	7/8	1 3/4	3 3/8	4 7/8	7 3/8	11 7/8	17 3/8	
Base weight (lbf/in)	1/2	7/8	1 1/4	2 3/8	4 1/8	6 3/4	9 3/8	16 1/8	24 1/2
Maximum load (lbf)	115	200	350	800	1500	2500	3500	6000	10000
Pitch moment (lbf-in) 44	119	265	900	2250	4688	7875	17625	35625	

Figure 8.2.14 Dimensions (inches) of modular dovetail slides. (Courtesy of Russell T. Gilman, Inc.)

Example: Sizing Sliding Bearing Pads

For the T carriage shown in Figure 8.2.16, assume that the rails are ground cast iron and the bearings are DU®as received (curve 7 in Figure 8.2.3). For a 1000 N (225 lbf) preload force it is desired to size the bearing pads to achieve 1.75×10^9 N/m (10^7 lb/in.) bearing stiffness in the vertical direction. From Equations 8.2.1 and 8.2.2, for a nominal operating point about $F_{applied} = 0$, the stiffness will be essentially twice what either individual set of pads will be; hence

$$K_{pad} = \frac{K_{desired}}{2 \times 4 \text{ pad sets}} = \frac{K_{desired}}{8} \tag{8.2.4}$$

$$F_{pad} = \frac{F_{preload}}{4 \text{ pad sets}} \frac{F_{preload}}{4} \tag{8.2.5}$$

From Figure 8.2.3, for low contact pressures the stiffness per unit area as a function of contact pressure can reasonably be approximated by

$$\frac{K_{pad}}{A_{pad}} = a + \frac{bF_{pad}}{A_{pad}} + \frac{cF_{pad}^2}{A_{pad}^2} \tag{8.2.6}$$

where the coefficients a, b, and c are the product of those given in Figure 8.2.3 and the unit conversion factor $(10^4 \text{cm}^2/\text{m}^2)(10^6 \mu\text{m/m})$. Solving for the bearing area yields

$$A_{pad} = \frac{-(bF_{pad} - K_{pad}) + \sqrt{(bF_{pad} - K_{pad})^2 - 4acF_{pad}^2}}{2a} \tag{8.2.7}$$

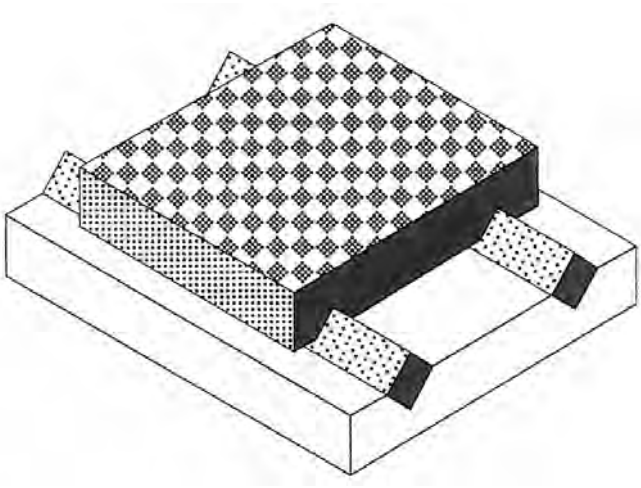


Figure 8.2.15 General configuration for a double-vee linear motion sliding bearing. Bearing surfaces may be composed of two bearing pads on each side of the vees (eight pads total), three pads per vee (six pads total), or pads covering the entire interface with the vees.

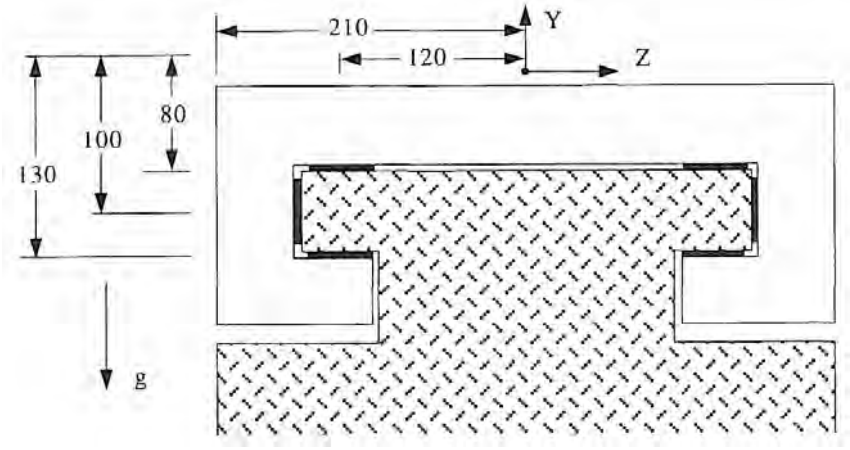


Figure 8.2.16 T slide for sliding bearing design example. The carriage rides on sliding contact bearings as shown, along with another identical set located a distance 480 mm into the page (all dimensions in mm).

For the 1000 N total system preload, A_{pad} = 0.00308 m^2. The pads should be narrow, to minimize keeper rail overhang. For a length-to-width proportion of 2:1, each pad should be about 8 cm long by 4 cm wide.

8.2.3 Design Case Study: The Nanosurf 2's Bearing Design[18]

Surface roughness measurements are often made over a small portion of the surface, while form measurements are made over the entire surface of a part. The former typically uses a transducer with 1 nm or better resolution (e.g., an LVDT) and the measurement is made over a small region of the part. The latter usually uses a transducer with less resolution and the measurement is made over a small region of the part. However, some optical components require both form and surface finish to be inspected with high-resolution across their width. For example, some laser gyroscope optical components were found to have 0.1 nm surface profile deviations over a 0.05 mm range, but a 20

[18] See K. Lindsey, S. Smith, and C. Robbie "Subnanometre Surface Texture and Profile Measurement with Nanosurf 2," Ann. CIRP, Vol. 37, No. 1, 1988, pp. 519–522. The author would like to thank Kevin Lindsey and Stuart Smith for their hospitality and help in preparing this section. Note that Rank Taylor Hobson Inc. is now manufacturing surface inspection machines that utilize this technology.

nm deviation over a 4 mm range. Hence the need for a long range of motion high resolution surface profilometer. The Nanosurf 2, shown in Figure 8.2.17, was designed at NPL by Lindsey and Smith to meet these goals.

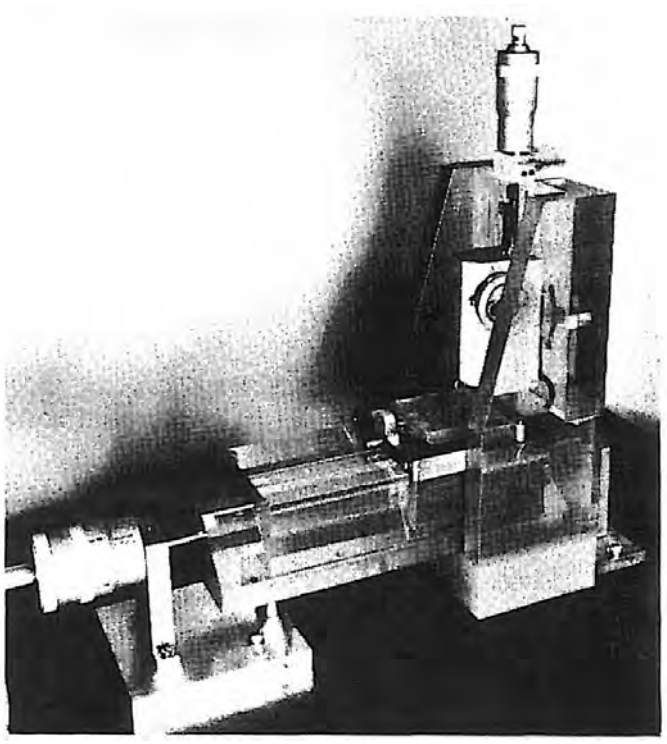


Figure 8.2.17 NPL's Nanosurf 2. (Courtesy of the National Physical Laboratory.)

Ideally, a surface profilometer must be mass producible, inexpensive, and easy to use. These characteristics are ideally suited to sliding bearings; however, nanometer motion resolution and repeatability (normal to the axis of motion) had never been documented before with sliding bearings. The primary problem was that the wear rate of sliding contact bearings prevented their use in applications requiring nanometer accuracy. Extensive detective work by Smith and Lindsey lead to the following conclusions:

- Most sliding contact bearings had surface features on the $^1/_4 - \mu m$ range, which led to the formation of a cobblestone street effect unless a massive carriage supported by very large area bearing pads was used to average out these features. For instruments, this type of design was impractical.
- Most sliding contact bearings have a bearing material thickness that may be from 30 to 300 μm thick. When manufactured, the plastic's structure is isotropic, but after use, a thin oriented (anisotropic) layer forms on the bearing surface. The isotropic layer beneath the oriented layer does not conduct heat well, and as a result, elliptical bubbles form in the thin oriented layer due to a large difference in axial and lateral coefficients of thermal expansion in the thin layer.[19] As these bubbles pop and the material wears away, the bearing's geometry changes (albeit on the nanometer level). For machine tool applications, the wear rates are low and acceptable. The wear rates are not acceptable for precision instruments. A very thin PTFE film, on the other hand, transfers heat very quickly to the backing material and thus will not experience this problem.
- When the bearing rail surface is polished, a thin film PTFE bearing will transfer a thin film to the rail and an equilibrium situation is established; hence wear can actually be

[19] See K. Tanaka, Y. Uchiyam, and S. Toyooka, "The Mechanism of Wear of Polytetrafluoroethylene," Wear, No. 23, 1973, pp. 153–172.

essentially eliminated.[20] Generally, an optical quality surface finish is required to achieve this effect.

As a result of these observations, the bearing design shown in Figure 8.2.18 was developed.[21] A thin film, 2-3 μm thick, of polymeric bearing material such as PTFE is deposited onto the convex bearing pads[22] and the bearing rail is lapped to have a surface finish in the nanometer range. The carriage must utilize five of these pads in a kinematic configuration such as those shown in Figure 2.2.7. Five radiused pads touch down on the bearing rail to provide a kinematic support condition; and the pads deform by the mechanism of Hertzian contact to actually contact the rails in circular regions. The circular contact region of diameter $D_{contact}$ may be on the order of 100 μm in diameter, but it depends on the resultant contact stresses and possible loss of kinematic condition as the radius increases to infinity.

Figure 8.2.18 Cross-section of the sliding bearing pad design for the Nanosurf 2. (Courtesy of the National Physical Laboratory.)

During the wear-in period of the carriage and bearing rail, a thin polymer film is transferred to the bearing rail. Soon an equilibrium transfer rate between the bearing pads and the bearing rail is attained. In this steady-state condition, the carriage motion will be repeatable over the length of travel even if it is used over only a portion of the travel distance. If the rail surface finish is too rough, then the polymer will abrade, albeit slowly, and the equilibrium layer will never form. Also, if the polymer layer is too thick, then wear will also occur as described above.

The accuracy of the system depends on the shape and surface finish of the bearing rail, while the repeatability and resolution depend on the surface finish. If the surface finish of the rails is R_a, then due to averaging effects, the smoothness of motion δ_{normal} of the carriage will be on the order of

$$\delta_{normal} = \sqrt{\frac{R_a^3}{D_{contact}}} \tag{8.2.8}$$

With a surface finish on the order of 5 nm R_a, which is readily achievable in glass by lapping, smoothness of motion normal to the direction of motion on the order of 0.05 nm is possible.

For machine tool applications a thin film bearing would probably not withstand the daily punishment; thus this type of bearing technology is probably restricted to machines that are pampered (e.g., instruments and wafer steppers).

8.3 Rolling Element Bearings

For rotary motion bearings, Figures 8.3.1 and 8.3.2 show representative types of ball and roller bearings which will be discussed in greater detail in Section 8.4. Typical linear rolling element bearing configurations are shown in Figure 8.3.3 and will be discussed in greater detail in Section 8.5. Note that in general it is easier to make a ball spherical than a roller cylindrical; hence ball bearings are more commonly used in precision machines than are roller bearings, the exception being when very high loads must be withstood. A good roller bearing may be better than a moderate ball bearing, and hence in the end the most important thing is to compare manufacturers' specifications.

For linear motion applications, it is more difficult to maintain quality control of a curved surface a ball rides on than a planar surface a roller rides on because the latter can be self-checking.

[20] B. Mortimer and J. Lancaster, "Extending the Life of Aerospace Dry Bearings by the Use of Hard Smooth Surfaces," Wear, No. 121, 1988, pp. 289–305.

[21] See K. Lindsey and S. Smith, U.K. Patent 8,709,290, Precision Motion Slideways, April 1988, or U.S. patent 4,944,606.

[22] See, for example, H. Boenig, Plasma Science and Technology, Cornell University Press, Ithaca, NY, 1982.

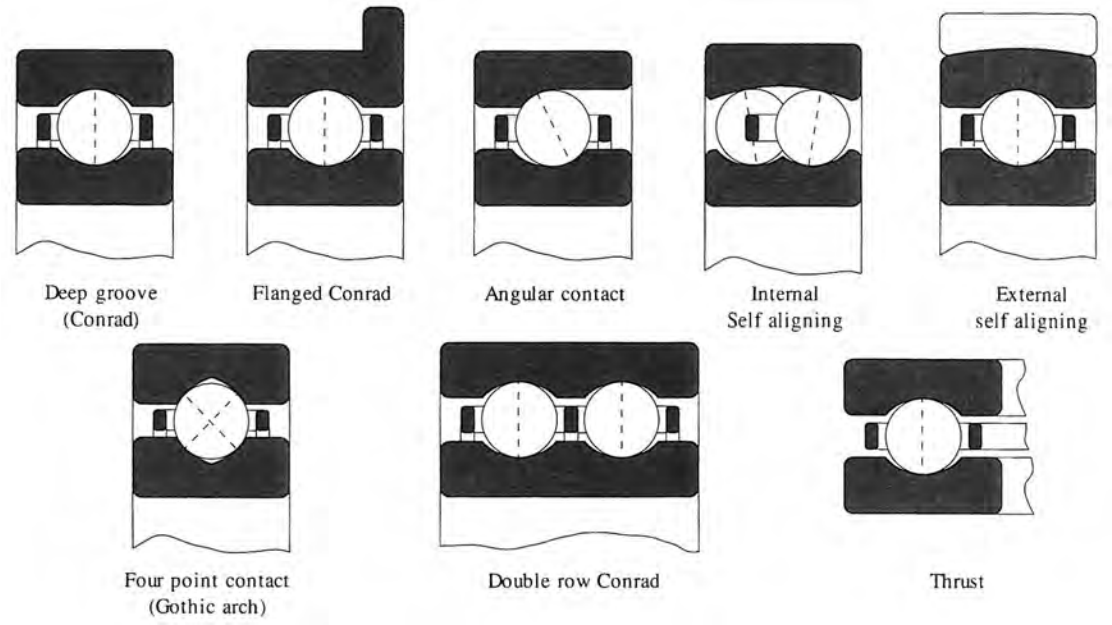


Figure 8.3.1 Typical ball bearing configurations for rotary motion bearings.

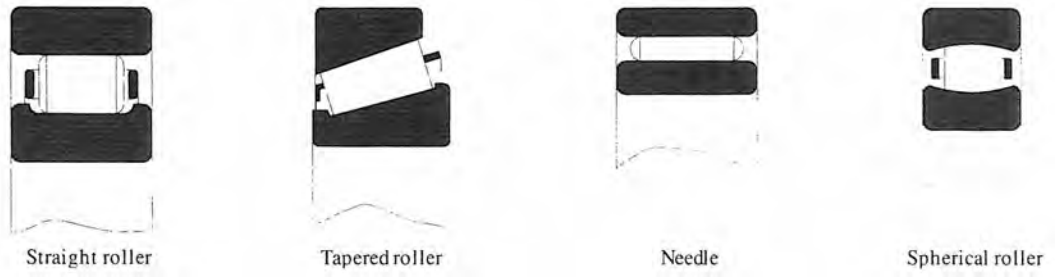


Figure 8.3.2 Typical roller bearing configurations for rotary motion bearings.

Note that machine-made linear bearing rails will have the same errors as the linear bearings on the machine that was used to make them. For rotary axes, the raceway and the grinding tool can both be rotating at different speeds, which combined with the random nature of precision spindle radial error motions means that the form of the raceway can be uniform around its circumference. Thus rotary motion rolling element bearings can be made more accurate than linear motion rolling element bearings. Typical total radial error motion on a precision spindle is on the order of $^1/_4$-1 μm. For higher-performance linear or rotary motion bearings, one would typically move into the realm of aerostatic, hydrostatic, or magnetic bearings.

The design and production of a rolling element bearing requires careful analysis, materials selection, manufacturing quality control, and testing. Few companies other than bearing manufacturers have the resources for this type of effort. Whenever possible, one should use off-the-shelf bearing components. In addition, whenever possible, one should use modular components such as spindles and linear axes, particularly for non-submicron machines made in small lots (less than about 10-20 machines). The savings in design time, prototype testing, spare parts inventory, and repair and replacement costs often far outweigh the potential of saving a few dollars in manufacturing costs.

Designing with rolling element bearings can be intimidating because there are so many subtle details that can ruin a design if they are not considered. The best way to learn about how to handle these details is to work with others with experience, and if possible to experiment.[23] In addition, manufacturers of precision bearing components are also usually willing to work with design engi-

[23] *"When there's no experimenting there's no progress. Stop experimenting and you go backward. If anything goes wrong, experiment until you get to the very bottom of the trouble."* Thomas Edison

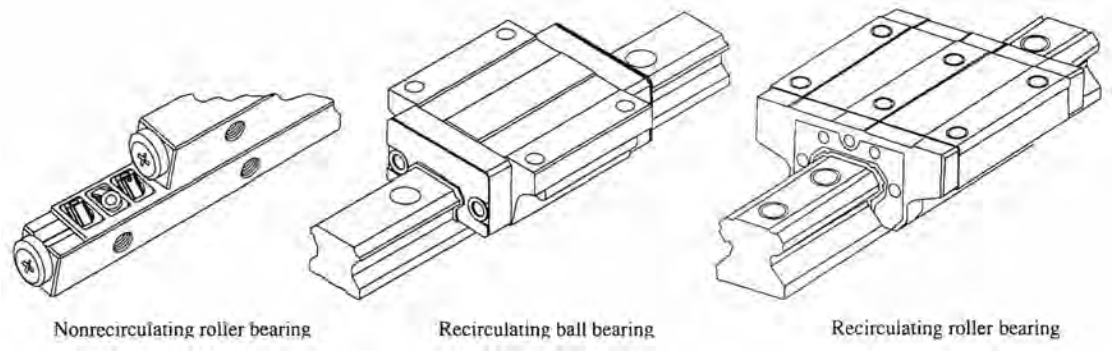


Figure 8.3.3 Typical linear rolling element bearing configurations.

neers to integrate their bearing components into a design. Furthermore, as discussed in Section 7.2.2 and illustrated in Chapter 6, mapping and metrology frames can be used to increase accuracy given adequate repeatability, resolution, and controllability of the machine *if* the machine is designed with these error-reducing methods in mind.

Speed and Acceleration Limits

The speed limits of a rotary motion rolling element bearing are a function of the dynamic balance of the assembly, the centrifugal forces on the assembly, the centrifugal force of the balls or rollers on the outer race of the bearing, and the type of lubrication and operating environment. The centrifugal force acts to expand the outer race (due to its own mass and the mass of the rolling elements), which can cause the bearing to lose its preload. At the same time, heat generated at high speeds from viscous shear in the lubricant and imperfect rolling contact causes the entire bearing to expand, but the outer race usually is attached to a bigger heat sink, so it tends to expand less. Furthermore, the centrifugal force acting on the balls causes the contact stress between the balls and the outer race to increase with the square of the speed. In addition, when the line of contact is not aligned with the angular acceleration vector (e.g., as is the case in an angular contact bearing), gyroscopic forces are generated which can cause the ball to spin as it rolls at high speed.[24] All these factors govern the maximum speed the bearing can be used at. Table 8.3.1 shows representative speed limits for various bearing configurations. As discussed below under accuracy, the lubrication in a bearing is dependent on the contact pressure between bearing elements. As the pressure increases from centrifugal force, the lubricant viscosity increases, which also increases the viscous drag on the bearing. In addition, the type of cage (e.g., pressed steel welded, molded plastic, or bronze) used to maintain spacing between the rolling elements can also have a significant effect on bearing performance.

All these factors combine to make high-speed spindle design a very difficult task not suited for novice design engineers. Most rotary motion systems with rolling element bearings have a maximum speed on the order of 5000 rpm, although 10,000-rpm machine tool spindles are becoming more common and 20,000-rpm machine tool spindles exist. Note that there are many manufacturers of spindles, and in most cases unless in-house design and manufacturing ability is readily available, it is wise to consider buying a modular spindle assembly from a reputable, experienced manufacturer. Because of the high speed at which spindles are often run, it is virtually impossible to correct for dynamic spindle error motions by moving the cutting tool. One can only account for long-term changes such as thermal growth or deflection caused by static loads.

Linear motion rolling element bearing speed limits are typically on the order of 1-2 m/s and are a function of the design of the bearing and the applied loads. Since linear bearings usually are not subject to the continuous high speeds of rotary bearings, active servo control of error motions using orthogonal axes can often be easily achieved if the machine and controller are designed properly.

Acceleration limits for rotary and linear bearings are usually not quoted in manufacturers' catalogs because associated with high accelerations are high speeds, which usually dominate the selection criteria. High accelerations can harm rolling element bearings if they cause the bearing elements to slip, which enhances wear. Fortunately, machines that undergo high accelerations often

[24] See Section 3.4 of B. J. Hamrock and D. Dowson Ball Bearing Lubrication, John Wiley & Sons, New York, 1981.

Bearing Type	Type of Cage	ABEC-1 Grease[b]	ABEC-1 Oil[c]	ABEC-3 Grease[b]	ABEC-3 Oil[c]	ABEC-7 Grease	ABEC-7 Circulating oil	ABEC-7 Oil mist
Single-row, nonfilling slot type	Molded nylon PRB	200,000	250,000	200,000	250,000	250,000	250,000	250,000
	pressed steel	250,000	300,000	250,000	300,000	300,000	350,000	400,000
Single-row, filling slot type	Molded nylon PRB	200,000	200,000	-	-	-	-	-
	pressed steel	200,000	250,000	-	-	-	-	-
Single row, radial and angular contact	Molded nylon PRC composite CR (ring piloted)	300,000	350,000	300,000	400,000	400,000	600,000	750,000
Angular-contact Single and double row	Molded nylon PRB	200,000	250,000	-	-	-	-	-
	pressed steel	200,000	250,000	-	-	-	-	-
Single-row, angular contact	Metallic (ring-piloted)	250,000	300,000	-	-	-	-	-

[b] Grease filled to 30-50% of capacity. Type of grease must be carefully chosen to achieve the speed values shown. Consult Torrington for complete recommendations.

[c] For oil bath lubrication, oil level should be maintained between one-third and one-half from the bottom of the lowest ball.

Table 8.3.1 DN speed values [Bore [mm] rpm] for ball bearings. Note that single- or double-sealed bearings should not exceed 250,000 DN. (Courtesy of The Torrington Company.)

have fairly large bearing preloads to increase stiffness to prevent large deflections due to inertial loads. A high bearing preload lessens the likelihood of slip. A tractive fluid[25] can also be used as a lubricant to prevent slip in machines with high accelerations.

Range of Motion

Rotary motion bearings usually provide unlimited rotary motion. One-piece rail linear motion bearings are generally available in up to 3 m lengths. Longer lengths usually require rails to be spliced together. When spliced together, there is no limit to the total length. Accuracy and repeatability will of course be affected by the total length. The longer the length, the greater the chance that thermal gradients and material and foundation stability will affect accuracy and repeatability.

Applied Loads

Many rolling element bearings can support radial and axial loads. In order to size a rotary motion bearing, an equivalent radial load F_e must be found:

$$F_e = K_\omega K_r + K_A F_A \tag{8.3.1}$$

where

K_ω = rotation factor = 1 for rotating inner ring and 2 for a rotating outer ring.

K_r = radial load factor = 1 (almost always).

K_A = axial load factor.

For radial contact ball bearings, $K_A = 1.4$. For shallow- and steep-angle angular contact ball bearings, $K_A = 1.25$ and 0.75, respectively. Similar equations are used for linear motion bearings to combine loads and moments. Most manufacturers also have load correction factors to accommodate the operating environment for the bearing.

Empirical load-life equations given by manufacturers for their bearings are fairly accurate; hence for a given load it is easy to select a desired bearing size. For example, a typical load life equation for rolling element rotary motion bearings has the form

$$L_a = a_1 a_2 a_3 (C/F_e)^\gamma \tag{8.3.2}$$

[25] For example, Monsanto Corp.'s Santotrac® lubricant, which solidifies under high pressure (>15,000 psi) and liquefies upon release of the pressure.

where

L_a = millions of revolutions.
a_1 = 1.0 for a 10% probability of failure.[26]
a_2 = a materials factor which is typically 1.0 for steel bearings and 3.0 for bearings with plated races.
a_3 = lubrication factor, which is typically 1.0 for oil mist.
C = basic dynamic load rating from a table of available bearings.
F_e = applied equivalent radial load, determined by bearing type.
γ = 3 for balls and $^{10}/_{3}$ for rollers.

With this type of equation it is easy to solve for the basic dynamic load and then choose an appropriate bearing from a catalog. Of course, for high-speed applications, as defined by the particular manufacturer, the operating speed must be taken into account when calculating the equivalent radial load. Cleanliness during assembly and use is critical to attaining long life.

For rolling element linear motion bearings, the basic dynamic load, C_N, typically refers to the load under which 90% of a group of bearings will support while traveling a distance of 50 km. For an applied load F_C, the load-life relation for rolling balls is typically[27]

$$L\,(\mathrm{km}) = 50\left(\frac{C_n}{\oint_w F_C}\right)^3 \tag{8.3.3}$$

For rolling cylinders the load-life relation is typically

$$L\,(\mathrm{km}) = 50\left(\frac{C_n}{\oint_w F_C}\right)^{10/3} \tag{8.3.4}$$

The service factor $\oint_w$ depends on the type of operating conditions:

$\oint_w$ = 1.0-1.5 for smooth operation with no impact or vibration loads (e.g., semiconductor equipment).
$\oint_w$ = 1.5-2.0 for normal operation (e.g., CMMs).
$\oint_w$ = 2.0-3.5+ for operation with impact or vibration loads (e.g., machine tools).

For very severe load and vibration situations, such as creep feed grinders, $\oint_w$may be as high as 10. For varying loads F_i each acting over a distance L_i, the equivalent load is given by

$$F_C = \left(\frac{\sum_{i=1}^{M} F_i^3\, L_i}{\sum_{i=1}^{M} L_i}\right)^{1/3} \quad \text{or} \quad F_C = \left(\frac{\int F(L)^3\, dL}{L}\right)^{1/3} \tag{8.3.5}$$

When properly selected, loaded, and maintained (cleanliness is critical), rolling element linear bearings can provide hundreds to thousands of kilometers of accurate motion. Rolling element rotary bearings have been built that are 10 m in diameter and support millions of newtons; hence the average machine design engineer is not likely to encounter a situation where a standard bearing is not available to handle the required loads.

Also of great importance is the ratio of preload to applied load. As the rolling elements roll by, the raceways and balls see an alternating stress on their surfaces. Hence it seems that the lower the preload, the longer the bearing life. In general, this is true, as long as there is sufficient preload to remove any bearing gap, so the elements are always in contact. Vibration loads, on the other hand, can generate an effective number of cycles orders of magnitude greater than the number of rolling cycles; hence a heavier preload is often desired to minimize the alternating stress levels caused by vibration. For shock or vibration applications, one should check with the bearing manufacturer.

[27] From NSK Corp. catalog, Precision Machine Parts and Linear Motion Products.

Note that if one axis is locked in position for an extended period while the machine is operating and generating vibration, the bearings in the immobile axis may be subject to fretting corrosion (as discussed in Sections 5.6 and 7.7). If stainless steel components or ceramic rolling elements are used, fretting would not be a problem.

Accuracy

Accuracy of bearings is usually defined as the deviation from the path along which the bearing is supposed to guide the machine; for example, the radial and axial error motions in a rotary motion bearing and straightness errors in a linear motion bearing (see Chapters 2 and 6). With bearings, accuracy also has a second meaning, that of the position accuracy that the bearing will allow the component it supports to be servo controlled. Of course the latter is really a function of the entire system design. Unfortunately, almost everything affects the accuracy of bearings.

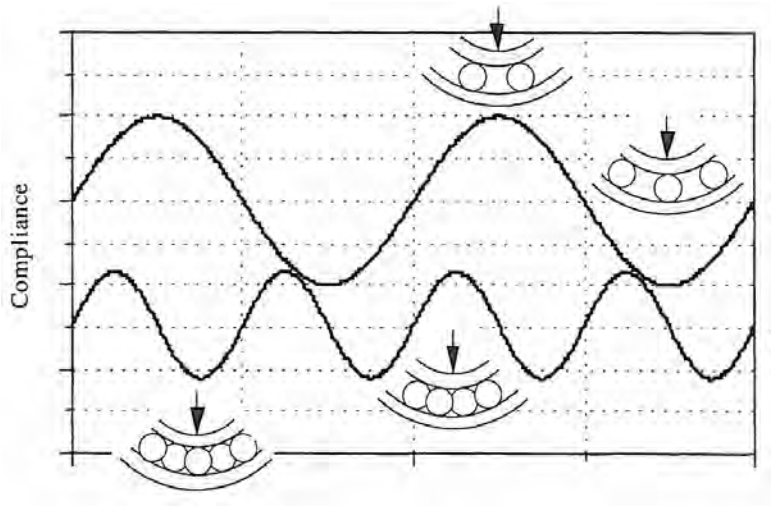


Figure 8.3.4 Illustrative comparison of the compliances of single- and double-row bearings.

The accuracy of a rolling element bearing depends primarily on the accuracy of the machine component onto which it is mounted, the accuracy of the bearing components, and the number of rolling elements. A perfectly round bearing pressed into an out-of-round hole will have large errors. A perfectly straight linear bearing rail bolted to a curved surface will take on the shape of the curve. Also, the form accuracy of a linear or rotary bearing raceway will directly affect the accuracy of the bearing. The number and accuracy of the rolling elements also affect the accuracy of the bearing, because the more rolling elements present, the greater the averaging effect of errors and the less sinusoidal variation in stiffness as illustrated in Figure 8.3.4. In fact, where it is important to minimize vibration, a double row bearing may be used where the cage holds the two rows out of phase by one-half pitch. Hence in order to determine the accuracy of a bearing, one cannot merely consider the numbers given in a manufacturer's catalog; one must consider the accuracy of every component in the system, their arrangement, and how other factors, such as preload and speed of operation, affect the system.

Figure 8.3.4 merely illustrates the change in compliance as a function of position of the rolling elements. As discussed in Section 2.4, there are many contributing factors to dynamic radial error motion. The elements in a rolling element rotary motion bearing generate error motions whose frequency can be predicted.[28] For inner and outer ring speeds ω_i and ω_o, respectively in rpm, the *rotation frequency* is:

$$f_i = \frac{\omega_i}{60} \tag{8.3.6a}$$

$$f_0 = \frac{\omega_0}{60} \tag{8.3.6b}$$

[28] See Section 2.3 of B. J. Hamrock and D. Dowson Ball Bearing Lubrication, John Wiley & Sons, New York, 1981. Also see M. Weck, et al., "Konstruktion Von Spindel-Lager-System für die Hochgeschwindigkeits," Material-bearbeitung; Expert-Verlag, 1990.

Note that most frequently either ω_i or ω_o is zero. A bearing ring will never be perfectly round, and the frequency given by Equation 8.3.6 must be multiplied by the number of lobes (e.g., 2 for an egg-shaped ring) to obtain the primary error motion frequency.

As the rolling elements roll, they move the cage along with them. This assembly revolves about the bearing's rotation axis with a speed of ω_c, which generates errors at the *cage frequency*:

$$f_c = \frac{\omega_c}{60} = \frac{1}{60\,D_{pitch}}\left(\frac{D_{pitch} - D_{ball}\cos\beta}{2}\omega_i + \frac{D_{pitch} + D_{ball}\cos\beta}{2}\omega_0\right) \tag{8.3.7}$$

where β is the ball-groove contact angle and the pitch diameter is usually the average of the bore and outside diameter. The frequency of the error motion caused by the rolling elements themselves is the *rolling element frequency*:

$$f_b = \frac{\omega_b}{60}\frac{1}{60\,D_{ball}\cos\beta}\left(\frac{D_{pitch} - D_{ball}\cos\beta}{2}\omega_i - \frac{D_{pitch} + D_{ball}\cos\beta}{2}\omega_0\right) \tag{8.3.8}$$

Note that if a rolling element has a high spot, then it will create a straightness error with a period of πD. If the rolling element is oval in shape, then the error period will be $\pi d/2$.

The *inner race frequency* and *outer race frequency* are, respectively:

$$f_{ir} = |f_i - f_c|\,N \tag{8.3.9}$$

$$f_{or} = |f_o - f_c|\,N \tag{8.3.10}$$

where N is the number of rolling elements. Often, in a spindle that has many sets of bearings, one is not even sure of the relative phase between the bearings, so the above frequencies may beat in very unusual ways. Fortunately, using frequency analysis methods, one can take the frequency spectrum of a spindle's error motion and determine what are the sources of the error motion frequency components. Figure 8.3.5 shows the frequency spectrum, averaged over 256 revolutions, of a ball bearing spindle's error motion.

As was illustrated in Figure 2.3.1, the surface of a raceway and rolling elements are not perfectly smooth or perfectly shaped. Recall the Hertz contact stress equations from Section 5.6 that a little bit of applied force initially causes a lot of deflection, so without much effort, the mean radius and roundness of all the rolling elements can be made equal so that error motions can be minimized. In addition, because of the nonlinear behavior of the contact stress equations, stiffness increases with preload. Increased stiffness can lead to increased accuracy in the presence of high loads. With continued preload, however, the rolling elements' ovality increases and so does the static and dynamic friction of the bearing.[29] Increased friction means that the servomechanism will have more difficulty positioning an object that the bearing is supporting and also leads to increased heat generation, which degrades accuracy.

As the contact pressure between elastic surfaces (i.e., steel bearings and raceways) increases it also increases the viscosity of the lubricating media used in the bearing. This helps maintain a film of finite thickness between the bearing components, which decreases friction and wear. If the bearing is moving, the asperities on the surface of the rolling elements and the raceways will be separated by the thin (at times only a few microinches thick) layer of the lubricant. This is referred to as *elastohydrodynamic lubrication*[30] because as the elements roll, they drag lubricant into the contact region, where its viscosity increases with contact pressure, thereby helping to maintain a film even in the presence of high contact pressures. Bearing design engineers must make extensive, but straightforward, calculations to ensure that under the rated load the lubricant film thickness is at least three times the surface roughness of the bearing. Fortunately, when choosing a bearing from a catalog, if one follows the operating condition recommendations, one usually need not be concerned with these calculations.[31]

[29] To appreciate this effect, take a tennis ball and roll it back and forth between your hands. Some of the energy expended in continually elastically deforming the ball is dissipated as heat due to hysteresis effects within the deforming ball, and as slip between portions of the contact region and your hands.

[30] See B. J. Hamrock and D. Dowson, Ball Bearing Lubrication, John Wiley & Sons, New York, 1981.

[31] A nice overview of bearing design considerations for load and life capabilities of many different types of bearings along with references for more detailed investigations is provided by B. J. Hamrock, "Lubrication of Machine Elements," Mechanical Engineers' Handbook (edited by M. Kutz), John Wiley & Sons, New York, 1986. Also see J. Brahney, "Film Thickness: The Key to Bearing Performance," Aerosp. Eng., June 1987.

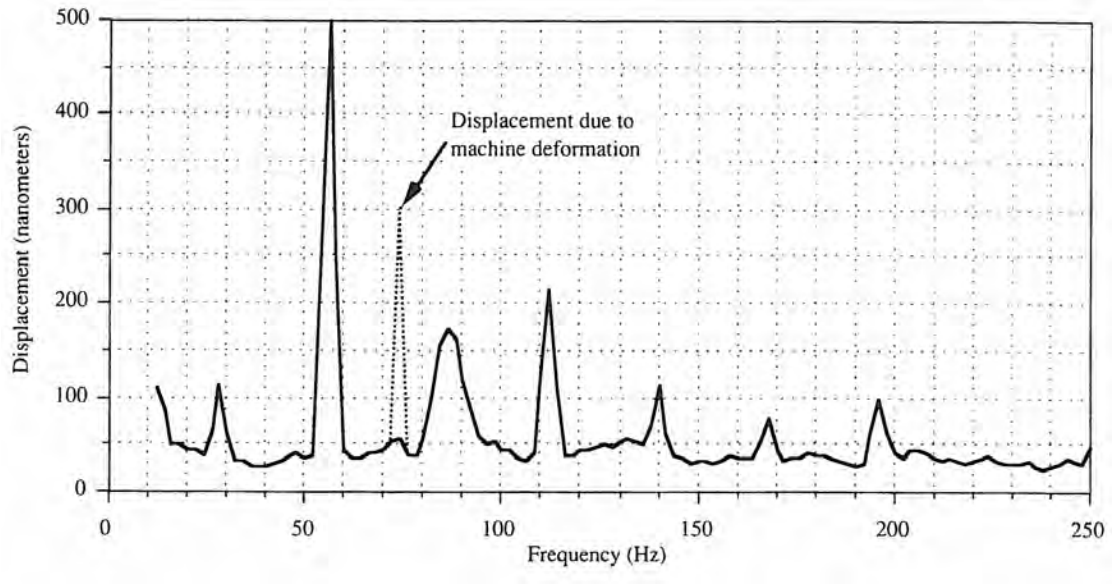


Figure 8.3.5 Frequency spectrum of a rolling element bearing spindle's radial error motion. The spindle speed was 1680 rpm (28 Hz), the bearing inner diameter was 75 mm, the outer diameter was 105 mm, the number of balls was 20, the ball diameter was 10 mm, and the contact angle was 15°.

The speed at which the bearing operates also affects accuracy. Increased speed leads to increased heat generation and thermal deformation. For this reason, many systems are designed so they are properly preloaded when operating temperature is reached. Increased speed also leads to increased out-of-balance forces and moments that cause deflection of bearing and machine components. Furthermore, increased speed leads to increased centrifugal forces, which cause the contact angle of a ball between the inner and outer race of an angular contact bearing to differ, which affects the elastohydrodynamic lubrication layer thickness and the amount of spin and possible slippage of the ball.[32] This also leads to increased gyroscopic effects. Rollers have line contact, so they are principally affected by a changing elastohydrodynamic lubrication layer thickness.

All these factors combine to make high-speed (greater than a few thousand rpm) or high-accuracy (less than 1 μm or so total radial motion error) rotary motion bearing design a task that requires care and experience. Precision rolling element spindle design is considered an art form by many design engineers, although in reality spindle design just requires extra careful attention to details in analytical modeling and design.

Linear motion rolling element bearings generally do not have to contend with centrifugal forces in the raceways but do need to be concerned with the dynamics associated with rolling elements in the raceway. With recirculating types, the rolling elements move from the raceway to the return tube, and then move from the return tube back into the raceway. This causes noise, which limits straightness and smoothness of motion to about the $^1/_2$ μm level. When high accuracy (better than about $^1/_2$ μm) and high speed (greater than a few thousand rpm or 1-2 m/s) are required for rotary or linear motion bearings, it is often desirable to use an aerostatic or hydrostatic bearing.

When a rotary motion bearing size is chosen, one must specify the accuracy to which the bearing components are made, which is represented by the ABEC number. Tables 8.3.2 and 8.3.3 show the tolerances associated with various ABEC numbers and bearing sizes. The higher the ABEC number, the higher the accuracy of the bearing components. For example, bearings used in appliances are usually ABEC 1, while bearings used in common machine tools are ABEC 5. Precision spindles use ABEC 7 or 9 bearings. Remember, the accuracy of the installed bearing will in large part be a function of the roundness of the bore and shaft diameters.

For linear bearings, accuracy of the first kind is given in catalogs as the deviation in straightness per meter travel when the bearing is mounted on a perfect surface. This is sometimes referred to

[32] Ibid.

Bearing bore (mm) over	incl.	Bore diameter +0 to:					Width variation				Total radial error motion					Side runout with bore				Raceway runout with side				Width inner and outer rings +0 to:	
ABEC #:		1	3	5	7	9	3	5	7	9	1	3	5	7	9	3	5	7	9	3	5	7	9	1,3,5,7	9
0	10	-8	-5	-5	-4	-3	8	5	3	1	8	5	4	3	1	8	8	3	1	15	8	3	1	-127	-25
10	18	-8	-5	-5	-4	-3	10	5	3	1	10	8	4	3	1	10	8	3	1	15	8	3	1	-127	-51
18	30	-10	-5	-5	-4	-3	10	5	3	1	13	8	4	3	3	10	8	4	1	15	8	4	3	-127	-51
30	50	-13	-8	-5	-5	-3	10	5	3	1	15	10	5	4	3	10	8	4	1	15	8	4	3	-127	-51
50	80	-15	-10	-8	-5	-4	13	5	4	1	20	10	5	4	3	13	8	5	1	18	8	4	3	-127	-51
80	120	-20	-13	-8	-6	-5	13	8	4	3	25	13	6	5	3	13	8	5	3	18	10	5	3	-127	-51
120	150	-25	-15	-10	-8	-6	15	8	5	3	30	15	8	8	3	15	10	8	3	23	10	8	3	-127	-51
150	180	-25	-15	-10	-8	-6	15	8	5	4	30	15	8	8	5	15	10	8	4	23	10	8	5	-127	-51
180	250	-30	-18	-13	-10		15	10	5		41	20	10	8	5	15	10	8		23	13	8	5	-254	

d_{MIN} (the smallest diameter of a bore) and d_{MAX} (the largest diameter of a bore) may fall outside the limits shown. $(d_{MIN} + d_{MAX})/2$ will be within the bore diameters tabulated. For further details, see AFBMA Standard 20.

Table 8.3.2 Standard ABEC tolerances (microns) for bearing inner races.

as the *running parallelism* of two surfaces on the bearing. One must be careful to determine whether the values given are cumulative or absolute. In the former case, a bearing with 1 μm/m straightness error will have a maximum straightness error of 3 μm for a 3-m-long bearing. For the latter case, the maximum straightness error will be 1 μm. Typical ranges of running parallelism (cumulative) for linear recirculating rolling element modular bearings are from 1-μm/m to 30 μm/m. When components are lapped, an order-of-magnitude increase can be realized. Remember that straightness errors exist in two directions orthogonal to the axis of motion. One must also ask the manufacturer to quote pitch, yaw, and roll accuracies, although in many cases the manufacturer will not have the data because it is assumed that the bearings are used in pairs that are widely spaced so that the angular errors will not manifest themselves. Note that for all types of bearings, cost increases substantially with accuracy.

Repeatability

Ideally, rolling contact bearings do not require a wear-in period. In practice, an assembly is usually run for several days to ensure that everything seats properly and that failure does not occur. During this wear-in period, tall asperities are worn down (if the bearings were not run-in at the factory) and the residual stress state due to rolling contact stresses hopefully reaches a steady state.[33] It is a good idea to have the assembly thoroughly flushed with clean lubricant after the wear-in period. The more accurate the bearing components, the closer to true rolling contact, the sooner a steady-state repeatability is reached, and the longer the bearing will last.

For recirculating element linear motion rolling bearing, the elements are not perfectly round (partially due to the preload) and a finite surface finish causes a small amount of slippage; thus when the bearing rolls back to its starting position, each element may not be in the place it was when it started, and since no two rolling elements are exactly the same, the bearing will be in a slightly different position. For a nonrecirculating element bearing, one might assume that the elements are forced into contact with the surface and thus act like miniature gears on racks, so they should be infinitely repeatable because the elements would always come back to the same position. This can be true for kinematic arrangements of rolling elements, where there is no substantial elastohydrodynamic lubrication layer. For lubricated nonkinematic designs, slippage can and does occur, which

[33] See K. Johnson and J. Jeffries, "Plastic Flow and Residual Stresses in Rolling and Sliding Contact," Proc. Symp. Fatigue Roll. Contact, Institute of Mechanical Engineers, London, March 1963.

Bearing OD (mm) over	incl.	Outside diameter +0 to:					Width variation				Total radial error motion					Raceway runout with side				OD runout with side			
ABEC #:		1	3	5	7	9	3	5	7	9	1	3	5	7	9	3	5	7	9	3	5	7	9
0	18	-10	-8	-5	-5	-3	10	5	3	1	15	10	5	4	1	20	8	5	1	10	8	4	1
18	30	-10	-8	-5	-5	-4	10	5	3	1	15	10	5	4	3	20	8	5	3	10	8	4	1
30	50	-13	-8	-5	-5	-4	10	5	3	1	20	10	5	5	3	20	8	5	3	10	8	4	1
50	80	-13	-10	-8	-5	-4	15	5	3	1	25	13	8	5	3	20	10	5	4	13	8	4	1
80	120	-15	-10	-8	-8	-5	15	8	5	3	36	18	10	5	3	23	13	5	5	13	8	5	3
120	150	-20	-13	-10	-10	-5	20	8	5	3	41	20	10	8	3	25	13	8	5	15	10	5	3
150	180	-25	-15	-13	-10	-6	20	8	5	3	46	23	13	8	5	30	15	8	5	15	10	5	3
180	250	-30	-18	-13	-10	-8	20	10	8	4	51	25	13	10	5	36	15	10	6	15	10	8	4
250	315	-36	-20	-13	-13	-8	25	13	8	4	61	30	15	10	6	41	18	10	6	18	13	8	4

D_{MIN} (the smallest diameter of an OD) and D_{MAX} (the largest diameter of an OD) may fall outside the limits shown. $(D_{MIN} + D_{MAX})/2$ will be within the outside diameters shown. For further details, see AFBMA Standard 20.

Table 8.3.3 Standard ABEC tolerances (microns) for bearing outer races.

causes the geometric constraints to change ever so slightly, which affects the repeatability of the bearing. If the design is not kinematic, then the greater the number of rolling elements, the greater the amount of elastic averaging.

For rotary motion bearings, there are many components that are deduced from the total error motion, including average, fundamental, residual, asynchronous, inner, and outer error motions. These are discussed in Section 2.2.1. Quite often bearing manufacturers simply refer to the total radial motion as *total nonrepetitive runout* (TIR), which is now not a standard term. Most manufacturers test their bearings using artifacts that are more accurate than the bearings themselves; hence TIR for be used as an estimate of the total radial error motion. A precision ball bearing spindle may have a total radial error motion on the order of $^1/_4$-1 μm, in contrast to an appliance bearing, in which the total error motion can be an order of magnitude higher. For linear motion bearings, one must ask for the straightness and pitch, yaw, and roll repeatabilities and whether or not they are cumulative.[34] Most manufacturers do not quote repeatability because most design engineers have historically been more concerned with accuracy. In most cases, one will find that the repeatability is 2-10 times better than the accuracy. To be sure, however, one needs to test a manufacturer's bearing before it is specified for a precision design.

Resolution

Resolution of a bearing refers only to the ability of the bearing to roll and allow small motions of the components it supports. Assuming that the rest of the machine system is perfect, as long as the carriage of the machine is kept moving (e.g., during a contouring cut), resolution of motion between rolling element bearing supported components can be very high, but ultimately depends on preload, lubrication, accuracy of components, and alignment of the bearings. For rotary motion, precision ball bearing resolution can be from microradians to nanoradians. For linear motion, precision ball or roller bearings' dynamic resolution can typically be from nanometers to microns. The minimum resolvable motion increment the bearings will allow from a standstill may only be 10 times worse, and depends mainly on the preload and accuracy of the bearing components (e.g., roundness of the rollers). The higher the preload, the poorer the resolution is likely to be. Most manufacturers will not attempt to quote figures for bearing resolution because of the great dependency on the rest of the

[34] See Section 2.3.

machine; however, assuming that the rest of the machine is perfect, a good indicator of the resolution is the starting torque or force required to begin motion, both of which are often provided by the manufacturer. The lower the starting torque, the greater the chance of achieving high resolution from the bearing. With careful nonlinear numerical simulation of the system's servo-control system using the starting torque or force as a deadband value, an estimate of the resolution can be made.

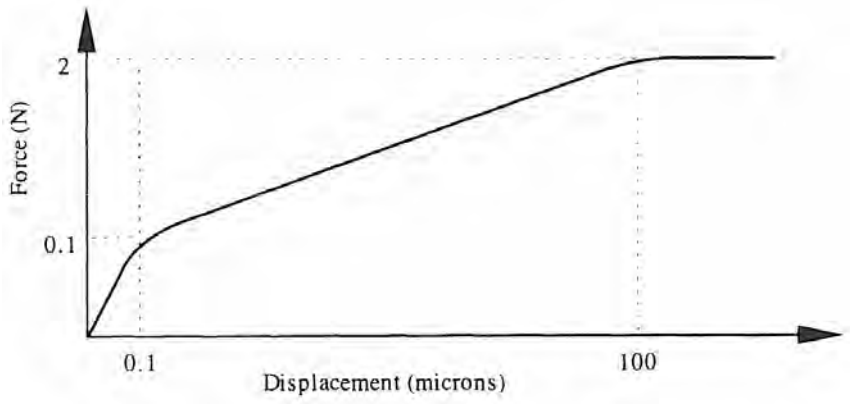


Figure 8.3.6 Behavior of the force-to-displacement relationship of a lightly preloaded, circular arc groove, recirculating rolling ball linear motion bearing. (After Futami et al.)

As illustrated in Figure 8.3.6, for a range of motion on the submicron level, the rolling element may not actually roll, but may just deform. For very small motions the rolling element acts as a flexural bearing and angstrom resolution may be possible; however, as soon as the element begins to roll, a large decrease in performance will be seen (the cobblestone street effect) with respect to nanoscale performance.[35] On the millimeter range-of-motion scale, the resolution will appear to be smooth once again. The exact transition from flexing to rolling is probably very difficult to predict and itself may not be repeatable. It would be interesting to see what the effect of an increasing thickness lubrication layer had on the resolution and whether or not a tractive fluid would help.

Preload

The preload greatly affects all aspects of bearing performance, particularly stiffness, in often a very nonlinear way. For example, stiffness increases continually with preload, while accuracy and repeatability increase as the bearing gap is closed and all the elements come into contact and then decrease as the friction levels begin to increase with preload. Resolution almost always decreases with increasing preload. For rotary motion systems, preload is achieved by displacing one race axially with respect to the other, expanding or contracting the inner or outer race on a taper, or by loading the bearing with oversize balls or rollers. Details of appropriate mounting conditions for various types of rotary bearings are discussed in Section 8.4. For linear motion systems, preload is achieved with the use of gravity, gibs, or loading with oversize balls or rollers. Table 8.3.4 shows examples of preload, operating conditions, and application examples from a manufacturer of linear motion recirculating ball modular bearings. *Heavy, medium, light,* and *very light* preload typically corresponds to about 5%, 3%, 2%, and 1%, respectively, of the static load capacity of the bearing unit, but these values are very approximate and manufacturers' catalogs must be consulted for details and application recommendations.

Stiffness

Stiffness of rolling element bearings is dependent on the preload and the size and number of rolling elements. Preload affects how many rolling elements are actually in contact, as shown in Figure 8.3.7 for a rotary bearing. As noted by a major bearing manufacturer:

A deflection analysis of the total system to include all of the parameters of the bearing, housing, and spindle requires a very complex computer program. The Timken Company has developed a sophisticated program which handles all of these parameters using a transfer matrix method of calculating the spindle stiffness. The spindle is evaluated using the four "state vectors" (deflection,

[35] S. Futami, A. Furutani, and S. Yoshida, "Nanometer Positioning and Its Microdynamics," Nanotechnology, Vol. 1, No. 1, 1990, pp. 31-37.

Class of preload	Operating conditions	Application examples
Heavy Medium	Vibration and shock loads Overhanging or offset loads Heavy cutting forces	Machining center, turning center
Medium Light	Light vibration Light overhanging or offset loads Light and medium cutting loads	Surface grinder, jig grinder, robots, laser processing machine, PCB drilling machine
Light	Slight vibration	XY table for semiconductor manufacturing
Very light	No overhanging or offset loads Light and precise operation	CMM tables, high-speed machines, EDM machines
Very light Clearance	Machines with large amounts of thermal growth High accuracy is not required	Welding machines, automatic tool changers, material handling equipment

Table 8.3.4 One manufacturer's recommendations for preloads for different applications of recirculating ball modular linear bearings. (Courtesy of NSK Corp.)

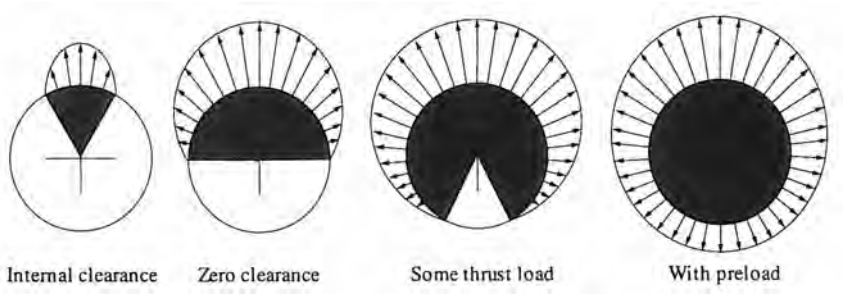


Figure 8.3.7 Effect of bearing preload on the region of the bearing that effectively acts to support the load. (Courtesy of The Timken Co.)

slope, moment, and shear). The housing and bearings are then evaluated based on the criteria established by the program for the bare spindle deflection. The bearing "spring rate" is calculated based on the bearing load zone. Since the load zone depends on the bearing setting, shaft stiffness, housing stiffness, load, internal bearing geometry, and the effects of thermal expansion, a complex iteration process is required for the calculation. This program can perform deflection analysis for any number of bearing supports or for any bearing configuration.

One of the unique features of the deflection analysis program is that it can calculate the effect of loose bearing fits on either the cup or the cone. Studying the effects of loose cone fits or loose cup fits on stiffness is very revealing. Often, a designer will specify a loose fit at the adjustable position because he feels looseness will aid in assembling and adjusting the spindle bearings. Recent studies on a typical spindle, however, have shown that a 0.010 mm (0.0004 in.) loose cone fit reduces spindle stiffness by 48%.

Some manufacturers have similar analysis programs that are generally available to customers who have a modem and a terminal.

With respect to selection of the number of rolling elements, consider a linear bearing with recirculating balls. For a given length of ball contact zone, is it better to have many small balls or a few large balls? The ball radius and force per ball are inversely proportional to the number of balls:

$$R_b \propto N^{-1} \tag{8.3.11a}$$

$$F_b \propto N^{-1} \tag{8.3.11b}$$

From the Hertz contact equations, the deflection and stress have the following proportional relationships to the ball radius and force, respectively:

$$\delta \propto R_b^{-1/3} F^{2/3} \tag{8.3.12a}$$

$$\delta \propto R_b^{-2/3} F^{1/3} \tag{8.3.12b}$$

Hence the deflection and stress are affected in an opposite manner by the number of balls used.

$$\delta \propto N^{-1/3} \tag{8.3.13a}$$

$$\delta \propto N^{1/3} \tag{8.3.13b}$$

Hence to increase stiffness more balls should be used, while to increase load-carrying ability fewer balls should be used. For rotary bearings one must be careful not to jump to the same conclusion because as the ball diameter increases, the angular relation of the ball with respect to the load line changes and hence so does the ability to carry loads. In addition, for rolling element bearings, the radial stiffness changes as the balls move circumferentially with respect to the load line; hence more balls or two rows of balls one half-pitch apart can greatly decrease vibration as discussed above. Usually, one just looks at the load and stiffness ratings for modular bearings, so one need not be concerned with these relationships; however, should a custom bearing need to be designed, with careful modeling and consideration of performance specifications one can usually find the proper-size rolling elements.

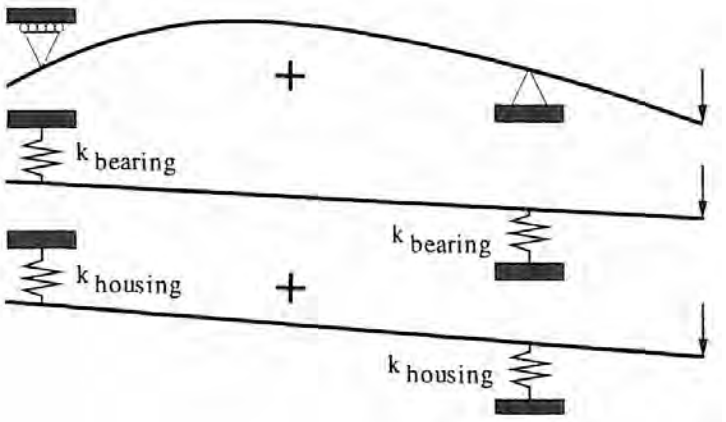


Figure 8.3.8 Sources of deflection in a spindle include contributions from the spindle shaft, bearings, and housing.

It is also very important to consider how much stiffness is really needed. For example, in the design of linear motion systems, often the bearing stiffness specified is many times greater than the carriage that the bearing supports. For spindles, one must carefully consider the stiffness of the bearings, shaft, and housing as shown in Figure 8.3.8. In general, one should strive to achieve a balanced design.

Vibration and Shock Resistance

With rolling element bearings, the contact area is small, so one must be careful to maintain adequate preload when vibration or shock loads are present. Bearing manufacturers have long considered this type of application and can readily suggest a suitable bearing and preload to prevent premature failure of the bearing. Take care, however, to ensure that the bearing is not subject to vibration during long periods when it is not moving. During these periods, the vibration causes the rolling elements to jiggle back and forth, which causes microslip. This causes the bearing to work its way through the lubricating layer and make physical contact with the race. With steel bearing components, this can lead to fretting corrosion[36] and premature failure. If rolling element bearings must be used in this type of application, stainless steel components or ceramic rolling elements should be used. A less effective alternative would to be to use hard chrome-plated steel components.

Damping Capability

For a long time, many linear motion carriage design engineers refused to consider the use of modular rolling contact bearing units because of the fear that they would not provide adequate

[36] See Section 7.7.4 for a discussion of fretting corrosion.

damping capability. However for general machine tool use, consider the fact that long ago rolling element bearings became the dominant type of bearing used in spindles, and ballscrews replaced brass nut leadscrews. Since a machine is only as good as its spindle or ballscrews, there must be a way to choose rolling element bearings to support linear motion carriages also. It is true that sliding contact bearings, which have a large contact area, are better damped than rolling element bearings; however, rolling bearings' modularity and low friction make them more attractive in many computer-controlled machine tool applications. With proper bearing selection, preload choice, and prototype testing, a rolling element bearing can often be used in place of sliding contact bearings. Exceptions may be where space is at a premium (e.g., a screw machine) or where very high stiffness, damping, and shock and fretting resistance are preferred over resolution and where hydrostatic bearings are inappropriate.

It is possible to support the primary load with rolling element bearings and then have an added sliding bearing unit that helps to damp vibration. Some linear rolling element bearing manufacturers now sell modular sliding bearing carriages that fit onto the same modular rails on which rides their modular rolling element bearing carriages.[37]

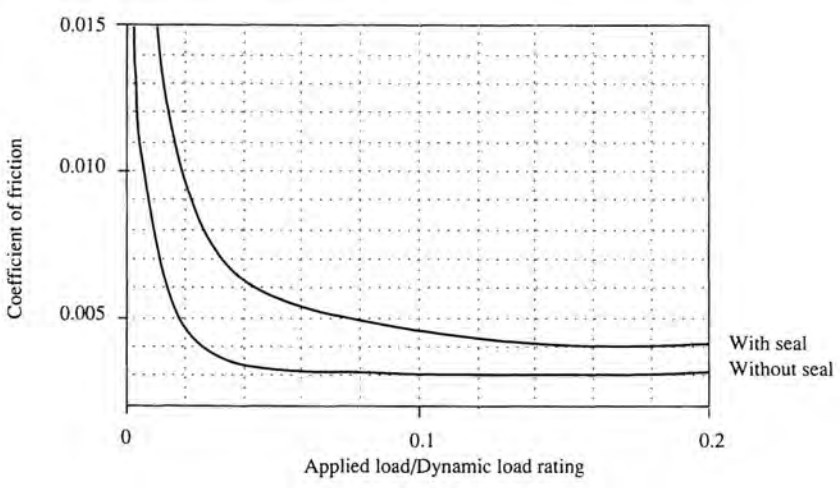


Figure 8.3.9 Effect of applied load on friction coefficient μ of a recirculating ball, circular arc groove linear motion guide, where there is minimal differential slip at the rolling interface. (Courtesy of THK Co., LTD.)

Friction

Due to elastic deformation of the surfaces, rolling element bearings' coefficient of friction (static and dynamic) depend to a large extent on the preload applied. Usually, the static and dynamic coefficients of friction are almost equal. With increasing preload, the coefficient of friction for bearings that are separated from each other by a cage or spacer balls may increase from a low of about 0.001 to about 0.01, depending on the bearing design. For bearings with rolling elements that are not separated (e.g., in some recirculating rolling element linear bearings), load is required to make the bearings roll efficiently without rubbing. Figure 8.3.9 shows the friction coefficient as a function of load for a high-quality bearing. It is interesting to note that with ceramic bearings, the contact zone is smaller, which leads to a purer rolling contact condition. As a result, ceramic bearings typically generate 30-40% less heat which allows them to use grease lubrication at higher speeds. This can more than pay for the additional cost of ceramic bearings.

Note that the dynamic coefficient of friction can change if the lubricant's viscosity changes with temperature. If a manufacturer cannot provide the data you need, then go to another manufacturer or perform tests on sample bearings. Because the static and dynamic coefficients of rolling friction are so low, the stick-slip problems associated with many sliding contact bearing systems are generally not encountered. As a result, on rolling element bearing supported carriages, one is likely to find linear scales used as feedback elements.

[37] Also see, for example, U.S. Patent 4,529,255.

Thermal Performance

Despite their low coefficients of friction, rolling element bearings can generate significant amounts of heat when the components they support move at an appreciable velocity. A rolling element bearing's frictional properties will also generally change with load and speed, which also affects the rate of heat generation. The interdependence of this relation makes modeling bearing thermal performance difficult but not impossible. As discussed in Section 8.4.9, there are a number of computer aided design packages which help evaluate thermal performance of bearings. These programs can help answer such questions as:

- How much oil (or grease) should be used?
- What oil (or grease) viscosity should be used?
- What will be the operating temperature?
- Must the oil be cooled?
- How much cooling is required?

Heat generated by relatively slow moving linear bearings is not as significant as that from motors and high-speed rotary joints (e.g., spindle and leadscrew journal bearings), but for a precision machine, all heat sources must be considered. An estimate of the power generated can be obtained from the friction, load, and speed and this value used to design the local region so that it can transfer the heat to a flow of temperature-controlled lubricant. As discussed in Section 2.3.5, in most cases one must be careful to cool bearings from the inside or else differential thermal expansion can cause the outer bearing parts to shrink more than the inner parts. This can cause an increase in the preload and in some cases can lead to catastrophic failure. The exception would be if when cold the bearing was unpreloaded, but then the machine should not be used until it is warmed up.

Bearings may be cooled with flowing oil if the speed is not so high that viscous shear in the oil will generate more heat than the oil can remove. For high-speed bearings, an oil mist is often used where oil is dripped into a high-pressure air stream, which blows over the bearings. In either case, it is vital that water be kept out of the lubrication system (e.g. < 100 ppm) or else premature bearing failure will result.

In order to determine how heat diffuses through the machine, finite element analysis or model testing may be required. Even with finite element analysis, one has the problem of the uncertainty of the heat conduction coefficient of joints. In addition, because the surface area of rolling contact bearings is small, the rolling elements can act as thermal insulators, which increases the machine's thermal time constant. Numerous methods have been tried to remedy this situation, including providing heaters for the machine or circulating temperature-controlled fluid through the machine. In general, one should isolate the heat sources, cool them, and use thermally insensitive designs as discussed in Sections 2.3.5 and 8.7. A further example of a thermally insensitive design is to neutrally support components as shown in Figure 2.3.12, and control the temperature of the supports so that thermal growth does not greatly affect the centerline location of the component.

In addition to radial dimension and preload changes caused by thermal effects, the designer must be concerned with axial thermal growth which typically is of a larger magnitude. For example, when supporting a shaft, one set of bearings usually is used to withstand radial and axial loads and a second set of bearings usually is used to withstand only radial loads. The second set must not be axially overconstrained to allow for thermal growth. There are four methods for dealing with the problem of axial thermal growth:

1. Let one set of bearings be free to slide in the bore. This is the most common, least expensive, and least desirable method because it can lead to decreased radial stiffness and accuracy if the fit is too loose or early bearing failure if the fit is too tight.
2. Preload sets of bearings to yield a thermocentric design. As discussed in Section 8.7, this is a very difficult configuration to design and thus is not often used for a multi-purpose spindle.
3. Use a hydraulic device to maintain a fixed preload on the bearings (see Figure 8.4.19). This is an effective, but more expensive, method.
4. Use a diaphragm, as shown in Figure 8.3.10, to support the bearing set which must have freedom to move axially. This is a moderate cost method that does require extra room and thus is not often used.

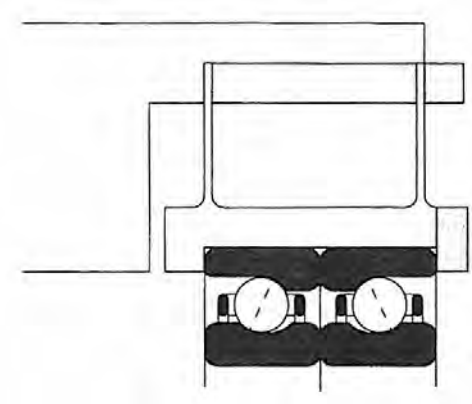


Figure 8.3.10 Method for bearing mounting that provides reasonable radial stiffness and low axial stiffness to accomodate axial thermal growth of the supported shaft.

Environmental Sensitivity

Precision rolling element bearings are extremely sensitive to dirt and will quickly grind themselves into oblivion (i.e., lose their accuracy and repeatability) unless they are carefully protected. Rolling element bearings can themselves generate micro wear particles, which can be very detrimental to a clean room environment. No matter what the application, rolling element bearings require some sort of seal, examples of which are discussed in following sections with mounting recommendations.

Few bearings for precision machines can run without some sort of lubricant,[38] so provisions must be made to permanently seal a bearing lubricated with grease or oil, or provide a steady flow of lubricant. When it is absolutely necessary to seal the lubricant in, a fluid with microfine ferrous particles (a ferrofluid) can be used along with a magnetic seal to keep the lubricant in the bearing. An example of this sort of application is in hard disk drives for computers. Remember that rolling element bearings must not be used for small-motion applications where the elements will not roll enough to drag lubricant between them and the raceways, or else fretting can occur (unless stainless steel or ceramic components are used). For limited motion applications one should consider using a flexural bearing.

Seal-ability[39]

Rotary motion rolling element bearings are relatively easy to seal using labyrinth or wiper seals. Linear motion bearings with flat or round way geometries are also relatively easy to seal with wipers. On the other hand, multigroove linear bearings (e.g., in a linear guide) are more difficult to seal because of the problems with accurately matching a wiper to the cross-sectional geometry of the rail. For clean room applications where dirt, cutting fluid, and chips are not generated, it is usually okay just to use a simple wiper for all types of rolling element bearings. When the bearings must operate in an adverse environment, way covers or bellows should be used. In severe environments, the inside of the way covers should be slightly pressurized to keep dirt out. In a clean room, the inside of the way covers should be subject to a slight negative pressure to keep micro wear particles in.

Size and Configuration

The number of types and sizes of rolling element bearings are almost innumerable. One has only to walk the aisles at a machine tool trade show to get a feeling for the many different types and sizes available. If you can imagine a bearing, chances are that a bearing manufacturer can make it. Do not be shy or too hastily think that you have invented the wheel.

[38] Note that there are innumerable material combinations for special applications. For example, the coefficient of friction of silicon nitride on silicon nitride is about the same as steel on lubricated steel; thus the former can be used in unlubricated high temperature bearings. See, for example, P. Dvorak, "Specialty Bearings Fill Nearly Every Niche," Mach. Des., June 23, 1988.

[39] See Section 7.6.6 for a discussion of different types of seals for rotary and linear sealing systems.

Weight

Weight is generally not a concern with rolling element bearings used in precision machines that do not have to go into outer space.

Support Equipment

Sealed lubricated-for-life ball bearings are available that do not need any support equipment. Most recirculating ball linear motion bearings must periodically be charged with grease through a grease fitting. Some spindle bearings and other bearings that must be cooled need a temperature-controlled oil supply/return system. It is vital that the cooling system fluid (oil or oil mist) be kept clean and dry.

Maintenance Requirements

Sealed-for-life bearings theoretically need no maintenance (until they fail). Some bearings need a shot of grease now and then in their grease fitting. Oil-cooled bearings require scheduled maintenance to change filters and check the quantity and quality of the oil. It is essential that oil-mist systems, which use high pressure air to form the mist, utilize air that is clean and dry (e.g., less than 100 ppm water).

Material Compatibility

As mentioned earlier, one can always find suitable materials from which to manufacture rolling element bearings for use in virtually any environment: examples being various types of stainless steels, ceramics (e.g., glass, silicon nitride, sapphire), and plastics. Again, note the possibility of fretting corrosion and that it can be prevented through the use of stainless steel components or dissimilar materials or by making sure that all axes are regularly moved enough to rebuild the elastohydrodynamic lubrication layer.

Rolling element bearings generally have poor thermal conductivity across the elements. It is very important to make sure that the manner in which the bearings are mounted and cooled, either by forced cooling or conduction to the machine and natural convection, does not result in loss of preload or too high an increase in preload. Thermal growth usually only causes failures in high-speed relatively constant duty devices, such as spindles. All rolling element bearing systems, however, generate some heat, and one must make sure that the resulting thermal expansion does not degrade the machine's performance too much.

Required Life

Properly designed, manufactured, installed, and maintained, rolling element bearings can provide millions of cycles of trouble-free accurate motion. The key is in the care taken during all design steps, from concept to maintenance documentation of a precision machine. Note that it is also important to realize that some wear will occur with every cycle, so it might be advantageous to allow for a periodic remapping of the machine's geometry so that wear can be compensated for, or to use easily replaceable modular bearings.

Availability

Modular rolling element bearings are usually available off the shelf, although unstocked standard product line items can take 6-12 weeks to obtain. Some manufacturers may take as long as 4-6 months to fill orders for items that are not stocked. Even longer lead times may be required for custom designs.

Designability

It is fun to design with modular rolling element bearings. However, because they can be sensitive to thermal growth and they do not wear-in to compensate for misalignment, one must be extremely careful designing the mounting method, as discussed in Sections 8.4 and 8.5. Manufacturers' catalogs usually have numerous design examples that show in detail how to incorporate their bearings into typical machines. One should never be shy about asking the manufacturer's representative for application assistance.

It is very important to note that care must be taken when designing the structure in which the bearings are mounted. A substantial structure is usually required to provide adequate stiffness. A substantial structure is also better able to dissipate the heat generated by the bearings.

Manufacturability

Modular rolling element bearings are highly dependent on the accuracy of the surface they are mounted on. For example, a perfectly straight linear bearing will take the shape of the bed it is bolted to; hence one must be careful with tolerance and assembly details. A mistake sometimes made is to assume that if a bearing is perfect by itself, then it will be perfect after it is mounted in the machine. Mounting bolts can easily be over tightened which can result in bearing distortion.

Cost

Rolling element bearings can cost from a few dollars for a small ABEC 1 rotary element bearing to several thousand dollars for a large 3-m-long high-accuracy linear bearing. When one goes from a normal-accuracy class to a high-accuracy class, the cost can increase by an order of magnitude, depending on the type of the bearing. Thus one must be careful not to specify more accuracy than is required.

8.4 ROLLING ELEMENT ROTARY MOTION BEARINGS

In this section design considerations will be discussed for the following types of rotary motion rolling element bearings:

- Radial contact
- Angular contact
- Four point contact
- Self aligning
- Thrust
- Straight roller and needle
- Tapered roller
- Spherical roller
- Thin section bearings

Typical ball and roller bearing construction were shown in Figures 8.3.1 and 8.3.2, respectively. Ball bearings' load capacity and stiffness depend on the radius of curvature of the groove, ball-groove contact angle, and number of balls. A ball bearing is typically assembled by radially displacing the inner race and loading the balls from one side. The balls are then distributed around the race and their circumferential spacing is then maintained by the retainer (cage). The deeper the groove, the fewer balls that can be loaded unless a filling hole is used. A filling hole increases the number of balls and hence load capacity and stiffness, but the plug may act like a micro bump if the bearing takes bidirectional thrust loads.[40] Roller bearings, on the other hand, have the desirable characteristic of allowing for the rollers to be loaded with almost 100% packing density. The cage is required primarily to keep the rollers from rubbing against each other as they roll. Since the rollers are in line contact, roller bearings generally have very high load capacity and stiffness.

Regardless of the type of bearing is used, the design engineer must be certain to follow manufacturers' guidelines for choosing the proper size of bearing and preload for the application. Furthermore, it is imperative that the manufacturer's recommendations for bore and shaft size and bore alignment be carefully followed. This includes determining the slope angle of the shaft caused by radial shaft loads and making sure that the slope angle does not exceed the misalignment tolerance. Chances are that if the slope angle is too large, then the shaft is not properly designed anyway.

Other factors to consider when designing a bearing mount that most affect the overall layout of the machine include[41]:

- Radial and axial error motions[42]
- Radial, thrust, and moment load capacity
- Allowance for thermal growth
- Alignment
- Preload adjustment

A useful tool for designing systems with rotary motion bearings is the ability to visualize forces as "fluids" and see how they flow from the shaft to the bearing to the housing. There should be only

[40] Nice reference on the mechanisms of ball bearings are B. J. Hamrock and D. Dowson, Ball Bearing Lubrication, John Wiley & Sons, New York, 1981; and T. Harris, Rolling Bearing Analysis, John Wiley & Sons, Inc., New York, 1991.

[41] Recall that sealing systems are also a crucial part of many designs. They are discussed in Section 7.6.6.

[42] Because the accuracy of a bearing is so highly dependent on the mounting surface, it is usually not quoted by bearing manufacturers. Instead, the repeatability is usually quoted and it is called nonrepetitive runout for a rotary motion bearing. Note that nonrepetitive runout is no longer a preferred term (see Figure 2.2.3) but the term persists.

one axial flow path in each direction or else the system is overconstrained axially. In general, there should be at least two radial force flow lines, to allow the system to withstand a moment. With this visualization method, one can easily determine if a system is properly constrained.

8.4.1 Radial Contact Bearings

There are two basic types of radial contact bearings: large-radius-of-curvature shallow groove and small-radius-of-curvature deep groove. Deep-groove bearings are often available with flanges to simplify mounting requirements. A large-radius-of-curvature shallow-groove bearing is meant to be preloaded through the use of oversize balls and have essentially no axial load-carrying capability. Thus when used as a radial support bearing in conjunction with another bearing that gives radial and thrust support to a shaft, the shallow-groove bearing allows for thermal growth of the shaft without having to allow the inner or outer race slide on the shaft or in the bore, respectively. To accommodate axial thermal growth, the balls roll axially a very small amount. The shallow groove prevents the balls from migrating out of the bearing. This capability maximizes attainable radial accuracy.

Typical applications of shallow groove bearings are in high-precision instruments, where high radial accuracy is required, so a loose fit of the inner or outer race, needed in other types of bearings to accommodate axial thermal growth, is not tolerable. Note that because the radius of curvature of the groove is much greater than that of the balls, the radial load capacity and stiffness are much lower than compared to those of a deep groove bearing.

Deep-groove bearings, also called *Conrad* bearings, have a groove radius of curvature close to that of the balls, and the groove has an arc length on the order of 90°; hence radial load capacity and stiffness are very high. The balls can contact the groove at a large angle from the radial position and thus accommodate fairly large axial loads. Preload is controlled by the manufacturer, who uses oversize balls to preload the bearing. Conrad bearings are not meant to have to be preloaded against each other. Where two Conrad bearings are used to support a shaft, both bearings' inner and outer rings can be fixed in place to the shaft and bore, respectively. This allows both bearings to support axial loads in addition to radial loads. However, unless the shaft and bore spacers are precisely matched, too large a preload may be incurred during assembly. Also, the preload may increase with temperature. Hence one bearing's non-rotating ring is usually allowed to float axially unrestrained. Mounting details are shown in Figures 8.4.1 and 8.4.2.[43]

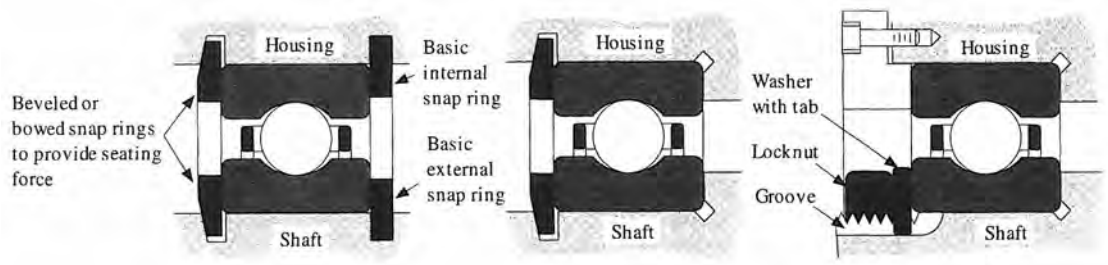


Figure 8.4.1 Methods for mounting Conrad bearings with full axial restraint in a bore and on a shaft.

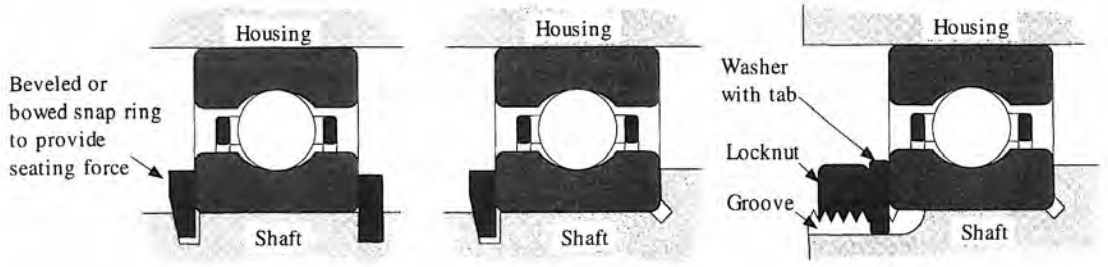
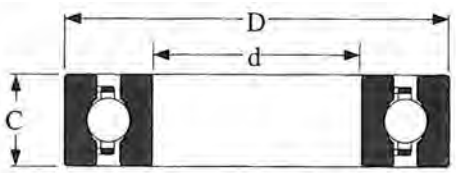


Figure 8.4.2 Methods for mounting Conrad bearings with full axial restraint on a shaft and axial freedom in a bore, for use with mountings shown in Figure 8.4.2.

[43] Shields and seals have been left out of the drawings for clarity.

Deep-groove bearings are widely used in consumer products and industrial equipment. Deep-groove bearings do not require a mechanism to preload them and thus can be held easily with inexpensive snap rings. Typically available deep-groove bearings are shown in Figure 8.4.3.



Bearing Number	Bore d	Tolerance +0 to -	OD D	Tolerance +0 to -	Width C	Fillet Radius	Mass	Static load Rating CO	Dynamic load Rating CE
	mm	μm	mm	μm	mm	mm	kg	N	N
300K	10	8	35	11	11	0.6	0.054	3750	9000
301K	12	8	37	11	12	1.0	0.064	3750	9150
302K	15	8	42	11	13	1.0	0.082	5600	13200
303K	17	8	47	11	14	1.0	0.109	6550	15000
304K	20	10	52	13	15	1.0	0.141	7800	17600
305K	25	10	62	13	17	1.0	0.236	12200	26000
306K	30	10	72	13	19	1.0	0.354	15600	33500
307K	35	12	80	13	21	1.5	0.458	20000	40500
308K	40	12	90	15	23	1.5	0.644	24500	49000
309K	45	12	100	15	25	1.5	0.862	30000	58500
310K	50	12	110	15	27	2.0	1.125	35500	68000
311K	55	15	120	15	29	2.0	1.424	41500	80000
312K	60	15	130	18	31	2.0	1.765	48000	90000
313K	65	15	140	18	33	2.0	2.168	56000	102000
314K	70	15	150	18	35	2.0	2.617	63000	116000
315K	75	15	160	25	37	2.0	3.175	71000	125000
316K	80	15	170	25	39	2.0	3.756	80000	137000
317K	85	20	180	25	41	2.5	5.008	90000	146000
318K	90	20	190	30	43	2.5	5.121	98000	156000
320K	100	20	215	30	47	2.5	7.085	127000	186000

Figure 8.4.3 Typically available radial contact bearings which can also generally be supplied with shields or seals. Bearings with bores from 12-105 mm are also available in extra- and super-precision series. (Courtesy of The Torrington Company.)

Radial and Axial Error Motion

Total radial error motion of radial contact bearings can be on the order of $^1/_2$-1 μm or better for high-precision grades, while axial error motion may be large (a few microns) because of the ideally pure radial contact. Because of radial contact, ideally no preload (provided by a mechanism) other than that provided for with the use of slightly oversize balls is required. High-performance radial contact bearings can be used in instruments where low total radial error motion and minimal coefficient of friction and starting torque are required but axial repeatability is not very important and speeds and cost requirements are low.

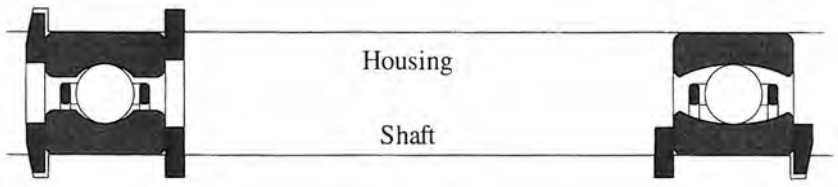


Figure 8.4.4 High-radial-accuracy, low-speed, low-starting torque assembly that utilizes a deep groove bearing and a shallow large-radius-of-curvature bearing.

A precision, low-speed, low-starting torque mounting is shown in Figure 8.4.4. Note that the bearings' outer races are mounted in a single straight-through bore, so concentricity and alignment are ensured. The shaft OD is ground in one pass until spark-out is reached, to maximize shaft accuracy and concentricity of the ends. Snap rings (shown) or a spacer can be used to axially

separate the bearings. The bearings themselves should be purchased as a matched pair. The deep-groove bearing's outer race is held in position by snap rings. If a step were machined in the bore and a single tapered snap ring used to seat the outer race, then a straight bore could not be used and some precision would be lost. For the inner race on the shaft, the snap rings are efficient, but one needs to make sure that the assembly is statically and dynamically balanced. To minimize friction, a labyrinth seal would be used instead of a contact type seal. However, this configuration is not suitable for high-speed operation because of the presence of the the snap rings on the shaft and the two-piece retaining ring in the deep-groove bearing. For high-speed, high-accuracy applications, one would probably use angular contact bearings with a single component retaining ring.

Radial, Thrust, and Moment Load Support Capability

As discussed above, large-radius-of-curvature shallow-groove bearings have low to moderate radial load capacity and negligible axial load capability. Deep groove radial contact bearings, on the other hand, have very high radial load capacity and moderate axial load capability. Moment loads are resisted by using pairs of radial contact bearings that are spaced suitably far apart. If radial load-carrying capability is to be increased by using more than one bearing at one end of a shaft, match ground bearings must be used because of the likelihood that slight differences in unmatched bearings will lead to unequal load sharing with failure of one bearing and then the next. Another alternative if high radial load capability is needed and limits on the shaft and housing sizes are stringent is to use a double-row bearing. A double-row bearing is essentially two deep-groove bearings that share common inner and outer races and a retainer. A big advantage of using a double-row bearing is that the balls can be one-half pitch apart, which can reduce sinusoidal variation in radial stiffness as the balls roll by up to 70%.[44] Note that the allowable shaft misalignment for a double row bearing is much less than that for a single-groove bearing. If possible, one should design the system so that only one bearing is needed, at each end of the shaft.

Allowance for Thermal Growth

Bearings generate heat when rotating and the housing usually has much better heat transfer to the outside world than the shaft; thus the bearings must be mounted in a manner that will allow the shaft to expand axially without placing axial loads on the bearings. Hence one needs to axially and radially restrain one bearing, while only radially restraining the other and allowing it to axially float on the shaft or in the bore while being held by the bore or shaft respectively. Figures 8.4.1 and 8.4.2 illustrated some of the many ways in which radial contact bearings can be mounted to achieve either of these two effects, respectively. Note the use of undercuts to ensure that the bearing races seat properly and to minimize stress concentration in the shaft. Chamfers also help to ensure that the bearing races seat properly. Note also that the stationary race should always be the one that is free to move axially. If the shaft were stationary and the housing rotating, then the inner race should be allowed to move axially and the outer race restrained. To lock a shaft nut in place, a backup washer with a tab in an axial groove in the shaft prevents torque from the inner race from being transferred to the nut. To prevent a nut from being loosened by vibration, a cotter pin could be used, although one could also use a backup washer with a tab that folds over onto one of the nut flats. Snap ring assemblies can work as well as more complex bolted assemblies with a shaft nut, but one must take care to observe the recommendations of the snap ring manufacturer.

Alignment

Radial contact bearings, particularly the large radius of curvature shallow-groove bearings, are more forgiving of misalignment than are other non-self-aligning ball bearings. Typical allowance for nonparallelism of the shaft to bore centerline may be as high as 1 mm/m.

Preload Adjustment

The preload in a radial contact bearing is generally set with the use of oversized balls; hence it is controlled by the manufacturer and the user need only make sure that the inner and outer races are properly secured to their respective assemblies. If one wants higher axial stiffness and load-carrying capabilities, one should use angular contact bearings. Note that deep-groove bearings can be purchased with a slight radial clearance and then preloaded against each other in the same manner that angular contact bearings are. Although deep groove bearings have fewer balls than do angular con-

[44] M. Weck, Handbook of Machine Tools, Vol. 2, John Wiley & Sons, New York, 1984, p. 187.

tact bearings, which means they will have lower load capacity and accuracy, they can have integral shields or seals which decrease the complexity of an assembly.

8.4.2 Angular Contact Bearings

An angular contact bearing's balls contact the races along a line inclined to a plane orthogonal to the axis of rotation. One side of the inner or outer race is open, so a large number of balls can be loaded into the bearing; hence radial load capacity and thrust load capacity in one direction are high. In order to support bidirectional thrust loads, a second bearing facing the opposite way is needed. To attain high stiffness and load capacity in a small space, two, three, or four (or more) bearings can be used forming duplex, triplex, or quadruplex sets. Angular contact bearings can be used in sets when their races are match ground, thus ensuring that even load distribution among the balls will occur. To achieve preferential load and stiffness capability, one could, for example, have two bearings facing the direction of maximum thrust and a third bearing preloading the other two and providing resistance to moderate thrust reversal. Figure 8.4.5 illustrates some of the many configurations that are used with angular contact bearings. Most angular contact bearings are universal; they can be preloaded by appropriately clamping the inner races (back-to-back) or the outer races (front-to-front).

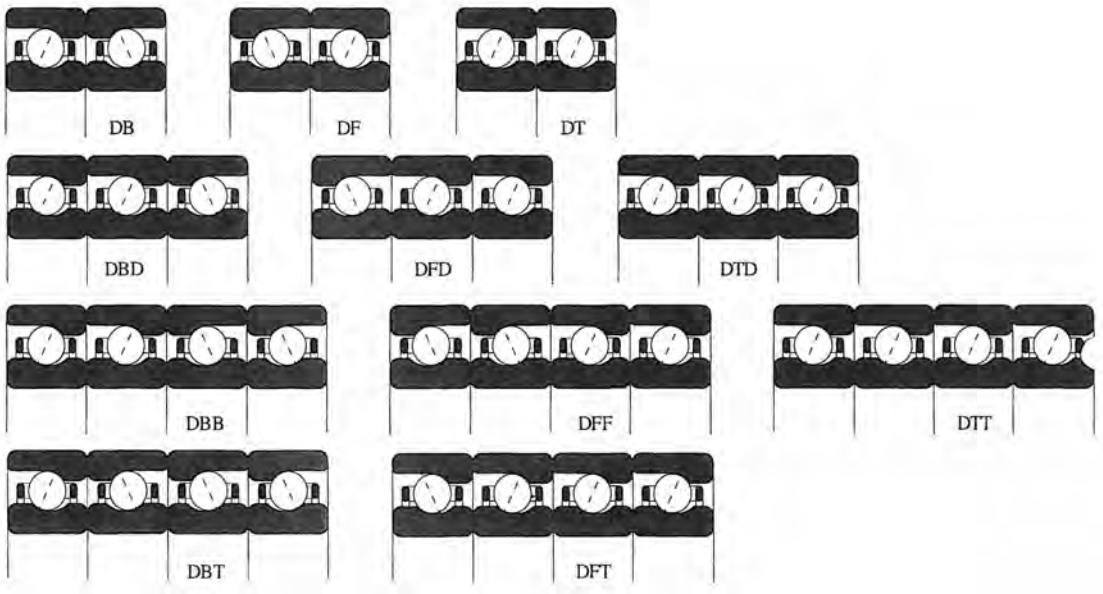


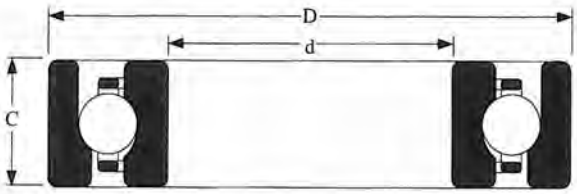
Figure 8.4.5 Some of the many possible configurations for angular contact bearings.

Angular contact bearings come in many sizes and grades and all are intended for precision machine tool use. Four types of angular contact bearings offered by one major manufacturer include:

- 2M-WI and 2MM-WI types with a 15°initial contact angle are for applications where the bearings must generate minimal heat and support very high radial loads.
- 3M-WI and 3MM-WI types with a 25°initial contact angle are for applications where the bearings must support very high axial loads.
- 2MM-WO types with a 15°-18° initial contact angle are for very high speed applications where centrifugal force is the primary load source. The outer race has full shoulders on both sides and the inner ring has a low shoulder on the nonthrust side to permit loading the bearing with a maximum complement of balls and a one-piece cage which pilots against the precision ground lands of the outer ring.
- 2MM-WN types with a 15° initial contact angle are for applications where maximum oil flow through the bearing is required for cooling and lubrication. The nonthrust sides of both the inner and outer rings have low shoulders, which also allows the bearing to be

loaded with a maximum complement of balls and a one-piece cage which pilots against the precision-ground lands of the outer ring.

Representative superprecision angular contact bearings for precision machine tool use are shown in Figure 8.4.6. Superprecision angular contact bearings with very steep contact angles are also available for supporting large thrust loads generated by ball screws, as shown in Figure 8.4.7.



Bearing Number Contact angle	Bore	Tol.	OD	Tol.	Width	Fillet	Mass	Static load Rating CO (kN)		Dynamic load Rating CE(kN)	
(25^o) (15^o)	d	+0 to -	D	+0 to -	C	Radius					
	mm	µm	mm	µm	mm	mm	kg	(2MM)	(3MM)	(2M)	(3MM)
(3MM) 2MM301WI-CR	12	3.8	37	5	12	1	0.068	4.05	3.90	11.0	10.8
(3MM) 2MM302WI-CR	15	3.8	42	5	13	1	0.091	5.00	4.80	12.0	11.6
(3MM) 2MM303WI-CR	17	3.8	47	5	14	1	0.113	6.20	6.00	17.0	16.3
(3MM) 2MM304WI-CR	20	3.8	52	5	15	1	0.159	8.65	8.30	21.6	20.8
(3MM) 2MM305WI-CR	25	3.8	62	5	17	1	0.236	13.20	12.70	30.5	30.0
(3MM) 2MM306WI-CR	30	3.8	72	5	19	1	0.354	17.30	16.60	38.0	37.5
(3MM) 2MM307WI-CR	35	5.1	80	5	21	1.5	0.458	22.00	21.20	46.5	45.0
(3MM) 2MM308WI-CR	40	5.1	90	8	23	1.5	0.644	27.50	26.50	56.0	54.0
(3MM) 2MM309WI-CR	45	5.1	100	8	25	1.5	0.862	33.50	32.00	67.0	64.0
(3MM) 2MM310WI-CR	50	5.1	110	8	27	1.5	1.125	40.00	38.00	78.0	75.0
(3MM) 2MM311WI-CR	55	5.1	120	8	29	2.0	1.424	47.50	45.00	90.0	86.5
(3MM) 2MM312WI-CR	60	5.1	130	10	31	2.0	1.765	55.00	52.00	104	99.0
(3MM) 2MM313WI-CR	65	5.1	140	10	33	2.0	2.168	69.50	67.00	125	120
(3MM) 2MM314WI-CR	70	5.1	150	10	35	2.0	2.617	80.00	76.50	140	134

Figure 8.4.6 Typically available superprecision angular contact bearings. (Courtesy of The Torrington Company.)

Bearing Number d	Bore +0 to -	Tol. D	OD +0 to -	Tol. C	Width Radius	Fillet	Mass	Preload	K_{axial} torque	Drag thrust	Dynamic thrust	Max.
	mm	µm	mm	µm	mm	mm	kg	(N)	(N/µm)	(N-m)	(kN)	(kN)
Duplex (add suffix DU)												
MM9306WI-2H	20.000	3.8	47.0	5	31.75	0.8	0.272	3340	750	0.34	25.0	25.0
MM9308WI-2H	23.838	3.8	62.0	5	31.75	0.8	0.527	4450	1100	0.45	29.0	35.5
MM9310WI-2H	38.100	5	72.0	5	31.75	0.8	0.590	6670	1300	0.45	36.0	45.5
MM9311WI-3H	44.475	5	76.2	5	31.75	0.8	0.590	6670	1390	0.56	38.0	51.0
MM9313WI-5H	57.150	5	90.0	8	31.75	0.8	0.859	7780	1655	0.79	40.5	61.0
MM9316WI-3H	76.200	5	110	8	31.75	0.8	0.980	10000	2100	1.02	44.0	76.5
MM9321WI-3H	101.600	6.4	145	8	44.45	1.0	2.16	13340	2455	1.36	85.0	150
MM9326WI-6H	127.000	7.5	180	10	44.45	1.0	3.86	17790	3150	2.27	91.6	186
Quadruplex (add suffix QU)												
MM9306WI-2H	20.000	3.8	47.0	5	63.50	0.8	0.545	6670	1500	0.68	40.3	50.0
MM9308WI-2H	23.838	3.8	62.0	5	63.50	0.8	1.053	8900	2200	0.90	47.5	71.0
MM9310WI-2H	38.100	5	72.0	5	63.50	0.8	1.180	13340	2600	0.90	58.5	91.0
MM9311WI-3H	44.475	5	76.2	5	63.50	0.8	1.180	13340	2780	1.13	61.0	102
MM9313WI-5H	57.150	5	90.0	8	63.50	0.8	1.707	15570	3300	1.58	65.5	122
MM9316WI-3H	76.200	5	110	8	63.50	0.8	1.961	21000	4200	2.03	72.0	153
MM9321WI-3H	101.600	6.4	145	10	88.90	1.0	4.32	26700	4900	2.71	127.0	300
MM9326WI-6H	127.000	7.5	180	10	88.90	1.0	7.72	35580	6300	4.50	147.8	375

Figure 8.4.7 Typically available superprecision angular contact bearings with 60^o contact angle for ballscrew journals. (Courtesy of The Torrington Company.)

To preload angular contact bearings against each other, they can be mounted back to back or face to face as shown in Figure 8.4.8. In the back-to-back mode the lines of force action project away from each other and give the pair a high moment-resisting capability. For a fixed outer race, note that as the shaft expands axially and radially more than the housing, the preload remains relatively

constant[45]: The axial expansion decreases the preload and the radial expansion increases the preload, with the result that the two effects cancel. For a fixed outer race, a face-to-face mounting has the opposite effect: The moment resistance is low and the system is thermally unstable for a rotating inner race because differential radial and axial growth both tend to increase the preload. For sets of bearings on the far ends of a rotating shaft, it would seem that the face-to-face mounts would allow for greater misalignment; however, the thermal instability factor is usually deemed more critical and thus the face-to-face mounting method is not often used. Note that when the outer race is rotating the inner race is fixed, the reverse is true. The amount of relative overhang in the races determines the amount of preload when a pair of angular contact ball bearings are preloaded against each other. Bearings are available flush ground (no preload after assembly) or with a moderate amount of overhang (preload obtained after assembly).

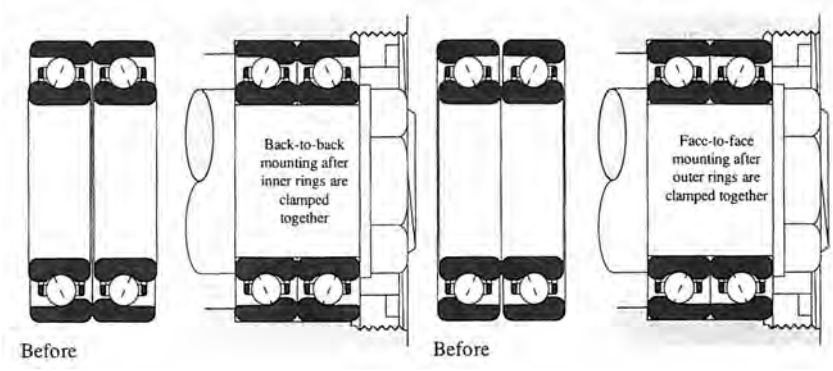


Figure 8.4.8 Back-to-back and fact-to-face mounting of angular contact bearings.

Radial and Axial Error Motion

Because they can have many balls loaded into them, total axial and radial error motions of angular contact bearings can be as low as $^1/_4$ μm but typically are on the order of $^1/_2$-1 μm. Commercially available spindles that use angular contact bearings are available with total error motions of $^1/_4$-1 μm. Angular contact bearings typically have three times the speed capability of radial contact bearings. Angular contact bearings are generally the most often used type of ball bearing in precision machinery.

Radial, Thrust, and Moment Load Support Capability

When properly mounted, angular contact bearings can support very high radial, thrust, and moment loads. Angular contact bearings' radial load capability is about the same as for a similar size Conrad bearing, but the former can support up to three times the thrust load of a Conrad bearing. Note that a radial force loads only one side of the ring, whereas an axial force loads the entire ring; hence the resultant thrust force from a radial force on one of the bearings often does not appreciably change the equivalent radial load on the other bearing. In many cases when radial loads are applied to a back-to-back or face-to-face mounting, the axial force components cancel. When a thrust load is applied, it only "flows" through the bearing whose contact angle faces it, and acts to unload the other bearing. Unfortunately, it is more difficult to envision how moment loads are supported by angular contact bearings.

A free-body diagram of a generic back-to-back mounting situation is shown in Figure 8.4.9. The following assumptions are made:

- The shaft, inner races, and balls are assumed to constitute a rigid free body in space.
- Reaction forces are applied through the ball-to-outer-race contact points.
- The reaction forces are assumed to act through the top and bottom balls in each of the two bearings.
- Once the maximum reaction force is determined, it is assumed to be distributed among the balls in that bearing in the same manner that a radial load would be distributed (as defined by the manufacturer for a particular bearing).

[45] Section 8.7 discusses design methods for achieving thermal growth equilibrium.

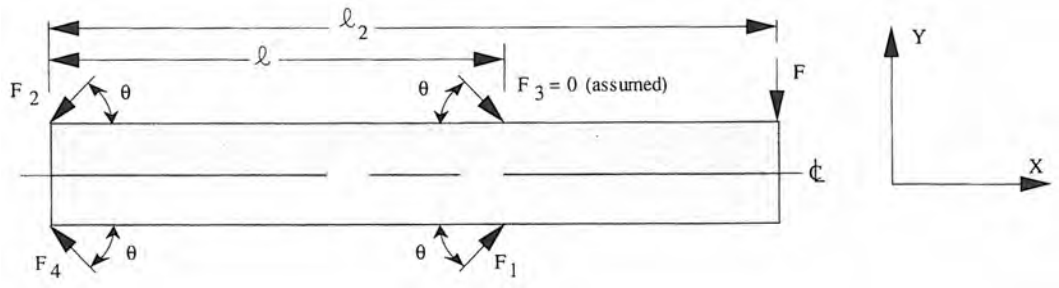


Figure 8.4.9 Free-body-diagram of back-to-back angular contact bearing mounting.

Note that there are four reaction forces, yet for this two-dimensional rigid-body model we have only three equations to work with. Recall that the balls can only transmit compressive loads. Hence we assume that one of the forces is zero and proceed with the calculations. If one of the reaction forces then turns up to be negative, then we made the wrong assumption. For the case shown, F_1 will be much higher than F_2, so F_4 will have to exist in order to act with F_2 to balance the high X-direction component of F_1; hence F_3 is assumed to be zero. Force and moment equilibrium give

$$\Sigma F_x = 0 = F_1 \cos\theta - F_2 \cos\theta - F_4 \cos\theta \tag{8.4.1}$$

$$\Sigma F_y = 0 = F_1 \sin\theta - F_2 \sin\theta + F_4 \sin\theta - F \tag{8.4.2}$$

$$\Sigma M_2 = 0 = F_1(R\cos\theta + \ell\sin\theta) + F_2 R\cos\theta - F_4 F\cos\theta - F\ell_2 \tag{8.4.3}$$

For the present example of a back-to-back mounting, after a little algebra one finds that

$$F_1 = \frac{F(\ell_2 \sin\theta + R\cos\theta)}{R\sin 2\theta + \ell\sin^2\theta} \tag{8.4.4}$$

$$F_2 = \frac{F(\ell_2 - 0.5\ell)}{2R\cos\theta + \ell\sin\theta} \tag{8.4.5}$$

$$F_4 = F/2\sin\theta \tag{8.4.6}$$

The free-body diagram and equation formulation for back-to-back mounting can also be used for a face-to-face mounting. For the former, θ may be on the order of 75^o and for the latter, θ may be on the order of 115^o. It is left up to the reader to find the reaction forces for the case when off-center axial loads are present.

When a shaft is supported by sets of angular contact bearings at each end (e.g., a spindle), the bearings offer moment resistance, so the shaft cannot really be modeled as simply supported. To do otherwise, however, requires modeling the bearings as springs, to provide consideration of geometric compatibility in order to solve for all the forces in the resulting overconstrained system. Procedurally, this is not difficult, but it is tedious to implement. As a result, one is probably best off doing a first-order conservative analysis that assumes that the bearing reaction forces act as simple supports, as shown in Figure 8.4.9. After the bearings have initially been selected to support the loads and have the required stiffness, to check the choice one could build a detailed finite element model that even includes each ball (modeled as a nonlinear spring using Hertz theory to avoid the debate over how to model contact interfaces accurately). One could then possibly use this same model for heat transfer and thermal growth calculations. As shown in Figure 8.3.8, one must also consider the behavior of the spindle shaft and housing.

In some cases a vertically oriented shaft must be held in position precisely, so the bearing mounting should be as kinematic as possible. To accomplish this, one could use a single angular contact bearing to support the axial load of the shaft and radially locate one end, and then use a large-radius-of-curvature shallow-groove bearing to radially locate the other end of the shaft. Note that this arrangement will result in the lightest-possible preload and hence starting torque, but it will only support downward thrust forces. Radial forces' resultant axial components in the angular contact bearing must be kept very low to avoid loss of preload.

Allowance for Thermal Growth

As with deep-groove bearings, one must axially and radially restrain bearing sets on one end of the shaft and generally only radially restrain bearing sets on the other end of the shaft. Near the front of the machine (e.g., spindle faceplate), typically several bearings (e.g., a quadruplex set) are preloaded against each other and axially and radially restrained on the shaft and in the housing. On the rear end of the shaft, several other bearings (e.g., a duplex pair) are preloaded against each other and are radially and axially restrained on the shaft but only radially restrained in the housing; hence the shaft is allowed to thermally expand axially while minimally affecting the bearing preload or the axial position of the front end (e.g., spindle faceplate). Only when speeds are low and relative shaft and bore thermal expansions negligible should a front bearing be preloaded against the rear bearing as discussed in Section 8.7.

Alignment

A single angular contact bearing can tolerate a misalignment of about ± 2 minutes-of-arc (0.6 mm/m). A duplex pair acts to resist moment loads and thus has an alignment tolerance of essentially zero degrees with another pair used to support a long shaft. Hence when using sets of angular contact bearings at each end of a shaft (e.g., spindle), extreme care must be taken to ensure that the bore centerlines are coincident and that the surfaces the races mate with at each end of the shaft are concentric. Typically, spindle bores are line-bored in one pass, or the spindle is mounted on a precision rotary table, then bored from one end, rotated 180^{o}, and then bored from the other end.

Preload Adjustment

Angular contact bearings are usually used in pairs and thus are preloaded by clamping the races until the overhung surfaces are forced together, as shown in Figure 8.4.8. Hence the amount of preload is set by the manufacturer and there is little danger of over-preloading the bearings during assembly when instructions are followed. Note that press fitting the races into a bore or onto a shaft causes the width of the races to change because of the Poisson effect; hence one must be careful to follow the manufacturer's recommendations. One should avoid using angular contact bearings, which require the user to measure axial spacing and then grind and insert shims to control preload. This method is inaccurate and a source of quality control nightmares.

Because of the nature of Hertzian contact, preload greatly affects stiffness in a nonlinear manner. To maintain machine performance over a wide range of operating conditions, it is often desirable to ensure that when the system is loaded, no bearing's preload is lost. Bearing manufacturers have a wide range of application experience, and if they cannot recommend a proper preload, then usually they will be willing to work with you to find a suitable preload.

8.4.3 Four-Point Contact Bearings

Four-point contact bearings have a Gothic arch shape groove in the inner and outer races. The balls make contact at two points on the inner race and two points on the outer race, which allows a single four-point contact bearing to support radial, axial, and moment loads. Hence this type of bearing is often used where space constraints exist, such as in robots and rotary turntables.

Radial and Axial Error Motion

Four-point contact bearings are essentially a duplex pair that have been cut in half along their radial centerlines and then fused into one bearing that is preloaded through the use of oversize balls. However, due to the fact that there is a finite contact area at each of the contact points, a four-point contact bearing is actually overconstrained and there is significant slippage.[46] Hence four-point contact bearings are not meant for high-speed, high-precision operation. Where high precision is required (better than about 3-5 μm total axial or radial error motion), one should use angular contact bearings.

Radial, Thrust, and Moment Load Support Capability

A four-point contact bearing essentially has the load capabilities of a duplex pair. Because a four-point contact bearing is itself fully constrained, only one bearing is needed to support a shaft (unless the shaft is very long). Bearing manufacturers give allowable radial, thrust, and moment

[46] See Figure 8.5.2 and Section 8.5 for a more detailed discussion of this effect.

loads for their four-point contact bearings and formulas for combining the loads into a single equivalent load. Designing with four-point contact bearings is thus usually straightforward. On a large industrial scale, this type of bearing can be manufactured with integral gear teeth and used as a turntable bearing for a crane.

Allowance for Thermal Growth

Since only one bearing is used per axis, one need only be concerned that the inside-diameter assembly will expand radially more than the outside-diameter assembly, which could cause the effective radial load to become too high. Note that sometimes a long shaft may be radially supported at one end by a four-point contact bearing and at the other end by a Conrad bearing that is not restrained axially in the bore; however, this is an odd arrangement.

Alignment

Again, since only one bearing is usually used, alignment is not a problem. One need only make sure that the mounting surfaces are ground to the manufacturer's specifications.

Preload Adjustment

The preload in a four-point contact bearing is set with the use of oversize balls; hence it is controlled by the manufacturer and the user need only make sure that the inner and outer races are properly secured to their respective assemblies.

8.4.4 Self-Aligning Bearings

The cross sections of external and internal self-aligning bearings were shown in Figure 8.3.1. The former is essentially a Conrad bearing whose outer race rests in a spherical seat. This design is meant to allow the bearing to be statically aligned to the shaft and not meant to readjust itself continuously; small motions of this type would invariably lead to fretting. Internal self-aligning bearings, on the other hand, utilize a spherical-shaped outer race groove which allows the balls to roll in the groove and perform the self-aligning function.

Radial and Axial Error Motion

Self-aligning bearings are typically suited for industrial machinery power transmission shaft supports and are generally not used in precision machines in the commercially available form. Note that spindles with self-aligning, hemispherical journal, porous air bearings are commercially available that have submicron error motions.[47]

Radial, Thrust, and Moment Load Support Capability

External self-aligning bearings generally have the same load-supporting capabilities as Conrad bearings. For internal self-aligning bearings, because the radius of curvature of the outer race groove is large, they can only support roughly 70% and 30% of the radial and axial loads, respectively, of a similar-size Conrad bearing. Note that if two bores that are meant for bearings to support a shaft are not concentric, then a self-aligning bearing must be used at each end.

Allowance for Thermal Growth

As with a Conrad bearing, one must be sure to provide only axial restraint for the shaft with one bearing. The other bearing should be free to move axially in the bore and provide only radial support to the shaft.

Alignment

An external type of self-aligning internal bearing can typically accommodate 5^{o} of misalignment. An internal self-aligning internal bearing can typically accommodate 2.5^{o} of misalignment.

Preload Adjustment

Self-aligning bearings are radially preloaded through the use of oversized balls, or with a tapered race forced onto a tapered shaft by an adjustment nut.

[47] For example, spindles built by Rank Taylor Hobson Inc., Keene, NH, and Cranfield Precision Engineering Ltd., Cranfield, Bedford, England.

8.4.5 Ball Thrust Bearings

Ball thrust bearings, shown in Figure 8.3.1, are designed to carry large thrust loads but require the use of an additional bearing to maintain radial alignment. For high-precision machines, ball thrust bearings are not often used because of the good job angular that contact bearings do in supporting both radial and thrust loads.

8.4.6 Straight Roller and Needle Bearings

As shown in Figure 8.3.2, straight roller bearings are cylinders that roll between cylindrical races, so they can only support radial loads. Needle bearings are long thin relatives that are used in applications where space is at a premium and high load-carrying capability is required (e.g., automotive driveshafts). Due to the line contact that exists at the roller race interface, straight roller bearings can support very high loads and have very high stiffness. They are often used to support heavily loaded shafts (e.g., a roll used in steelmaking or printing, or a heavy-duty spindle) that are axially constrained by tapered roller or angular contact bearings. Note that if the load is not evenly distributed along the roller, it can cause the roller to migrate axially and cause loaded contact with the bearing shoulders. Hence two rows of rollers (needles) with rounded ends may be needed along with hardened ground shoulders on the rings.

Radial and Axial Error Motion

Straight rollers can easily be individually inspected and the cylindrical races easily ground and inspected, so very small radial and axial total error motions can be obtained. Some large machine tools use straight roller bearings in their spindles. However, in a high-precision machine one rarely needs the load-carrying capabilities of straight roller bearings and thus submicron performance bearings of this type are not generally available on an off-the-shelf basis. On the other hand, when needed, cylindrical rollers can be lapped and inspected to achieve virtually any performance level desired.

Radial, Thrust, and Moment Load Support Capability

Straight roller bearings are ideally used only to support radial loads. In some types, however, the races have lips which allow the ends of the rollers to ride against, thereby allowing them to support light axial loads. Note that the ends of the rollers are slipping against the race lips as opposed to rolling. When used in conjunction with other bearings to form a couple (e.g., a pair of tapered roller bearings) straight roller bearings can support very large moment loads. Since the rollers do not establish the relative axial position of the races, the inner and outer races must both be axially restrained on the shaft and in the bore, respectively, using one of the clamping methods shown previously.

Allowance for Thermal Growth

Since straight roller bearings are not meant to resist axial loads, one need only make sure that differential radial expansion between the bore and the shaft does not cause the preload of the bearing assembly to change beyond an acceptable amount.

Alignment

Any shaft misalignment causes the rollers to be compressed into tapered shapes, which increases wear and friction and the tendency to migrate axially as noted above. Hence alignment must be within about 10-100 μm/m.

Preload Adjustment

Straight roller bearings and needle bearings are preloaded by using oversize rolling elements or by using a bearing with a slight taper on the inner race inside diameter or outer race outside diameter. This taper rides on a mating taper on the shaft or in the housing, respectively. Preload is controlled during assembly by setting the axial location of the bearing on the mating taper.

8.4.7 Tapered Roller Bearings

Tapered rolling bearings are one of the most often used types of bearing. They have a wide range of applications from supporting the axles in a car to supporting heavy-duty machine tool spindles. The components of a basic tapered roller bearing assembly are shown in Figure 8.4.10, and like angular contact bearings, they are meant to be used in pairs with one bearing preloaded against the other. An example application is shown in Figure 8.4.11.

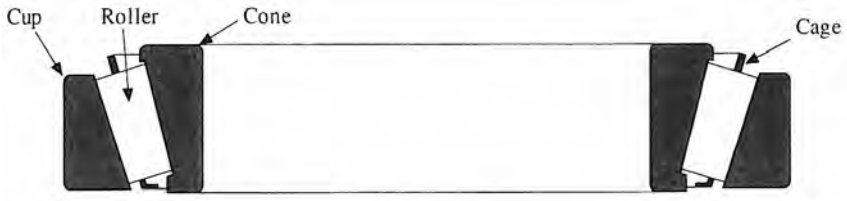


Figure 8.4.10 Cross section of a typical tapered roller bearing.

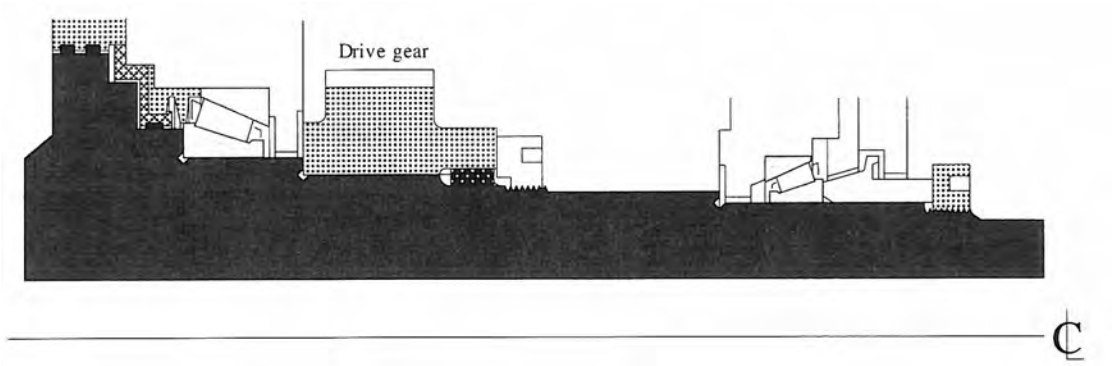


Figure 8.4.11 A high power precision spindle configuration (>20 kW) that uses tapered roller bearings.

Tapered roller bearings have line contact between the rollers and races and thus they have very high stiffness and can support very high axial and thrust loads. The angle of the taper is such that the lines tangent to the raceway and roller surfaces meet at a common point on the axis of rotation; hence true rolling motion of the tapered roller is attained. However, the contact forces between the roller and raceway are not parallel, so the effect is to squeeze the roller out, which causes the flat end of the roller, which is orthogonal to the cone axis, to seat on a lip on the inner race, as shown in Figure 8.4.12. This helps keep the rollers from skewing but adds to the friction coefficient of the bearing. Like angular contact bearings, tapered roller bearings' rolling elements are also subject to gyroscopic forces. For high-precision machine applications, if load and stiffness requirements can be met with angular contact ball bearings (without having to use too many), they should probably be used instead of tapered roller bearings.

Radial and Axial Error Motion

It is more difficult to finish a tapered roller than a spherical ball, but in the context of a bearing system, some tapered roller bearing manufacturers can deliver tapered roller bearings with specifications equal to an ABEC 9 ball bearing (most ball bearing manufacturers will only provide up to an ABEC 7 bearing). Tapered roller bearings are often used for general machine tool applications where heavy spindle load capability is required. In these applications, radial and axial total error motions on the order of 1 μm can be obtained. Recently, class AA tapered roller bearings were introduced by The Timken Company which have greatly enhanced performance, as shown in Figure 8.4.13.

Radial, Thrust, and Moment Load Support Capability

Tapered roller bearings can be mounted back-to-back or face-to-face just like angular contact ball bearings, and the same first-order force analysis methods apply as described above (even the

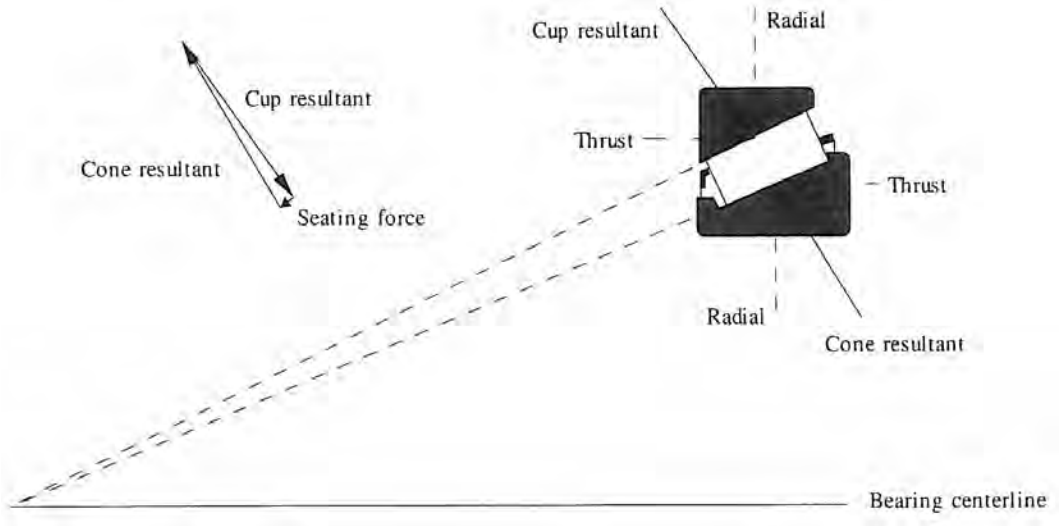


Figure 8.4.12 Force components and seating force resultant for a tapered roller bearing. True rolling motion is obtained because the extensions of lines along the rollers' profile intersect at an apex on the bearing centerline.

same equations can be used). Typically, available sizes for precision machine tool applications are shown in Figure 8.4.14. Note that sets of tapered roller bearings are also often grouped at each end of a shaft to give increased load capacity without a bigger bore, just like sets of angular contact ball bearings are used in duplex, triplex, and so on, mountings. Some manufacturers provide extensive examples of reaction force calculations, particularly where gearing is involved, and the reader is encouraged to obtain manufacturer's literature.[48] Manufacturers also have complex empirical formulas for calculating life based on the applied loads and operating condition. Since a radial load on a tapered roller bearing induces a thrust load, and vice versa, one must take this into account when calculating equivalent radial loads, as shown in Figures 8.4.15 and 8.4.16. Similar figures are provided by manufacturers for cases where multiple bearings are used.

Allowance for Thermal Growth

Tapered roller bearings are mounted using the same philosophy as used for angular contact ball bearings. The back-to-back mounting method is generally thermally stable for rotating inner ring assemblies, whereas the face-to-face method generally is not. When back-to-back sets of bearings are used at each end of a long shaft, the stationary races of one set must be free to move axially as differential thermal expansion between the shaft and housing occur.

Some manufacturers are kind enough to provide information to allow design engineers to make preliminary estimates of heat generation. The Timken Company provides the following method for finding the heat generated in their tapered roller bearings:

1. Find the equivalent axial load using the bearing's load factor K and axial load factor f_t shown in Figure 8.4.16. If $KF_{axial}/F_{radial} > 2.5$, then $F_{eq\ axial} = F_a$, else $F_{eq\ axial} = f_t F_r/K$.
2. The heat generated (Watts) is a function of the equivalent axial load, heat factor G_1 (given with other bearing properties in Figure 8.4.14), viscosity μ (centipoise), and speed S (rpm):

$$Q = 2.7 \times 10^{-7}\, G_1 S^{1.62} \mu^{0.62} F_{\text{axial equiv.}}^{0.30} \tag{8.4.7}$$

3. The viscosity of the oil depends on the oil temperature and hence the flow rate of the oil. The flow rate (liters/min) required to remove all of the heat generated by the bearing is

$$F = \frac{9.6 \times 10^{-9}\, G_1 S^{1.62} F_{\text{axial equiv.}}^{0.30}}{T_{\text{oil out}} - T_{\text{oil in}}} \tag{8.4.8}$$

[48] Just look in the Thomas Register under *Bearings, Roller* for a list of roller bearing manufacturers.

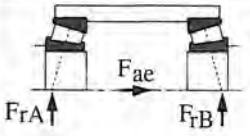
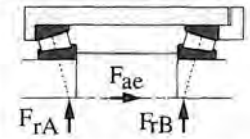
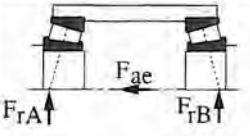
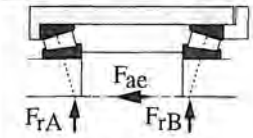
Cup OD		Class			
Over (mm)	Including (mm)	C (μm)	B (μm)	A (μm)	AA (μm)
30	50	6	2.5	2	1
50	120	6	3.5	2	1
120	150	7	3.5	2	1
150	180	8	4	2	1
180	250	10	5	2	1
250	315	11	5	2	1

Figure 8.4.13 Typical tapered roller bearing average radial error motion. The asynchronous radial motion value is about one-half this amount. (Courtesy of The Timken Company.)

Bearing #	Bearing size (mm)	F_{radial} (kN)	Load factor K	Heat factor G	K_{radial}* N/μm	K_{axial} N/μm
JP6049-JP6010	60 x 100 x 21	21.0	1.24	39.5	770	310
JP7049-JP7010	70 x 110 x 21	22.0	1.27	51.1	842	314
JP8049-JP8010	80 x 125 x 24	27.2	1.29	69.7	962	345
JP10044-JP10010	95 x 145 x 24	30.1	1.24	104	1123	441
JP10049-JP10010	100 x 145 x 24	30.1	1.24	104	1123	441

* Stiffnesses based on 180^{o} load zone

Figure 8.4.14 Tapered roller bearings with low heat generation for precision machine tool applications. (Courtesy of The Timken Company.)

Thrust condition	Thrust load	Dynamic Equivalent Radial Load
 $\frac{0.47F_{rA}}{K_A} \le \left(\frac{0.47F_{rB}}{K_B} + F_{ae}\right)$	$F_{aA} = \frac{0.47F_{rB}}{K_B} + F_{ae}$ $F_{aB} = \frac{0.47F_{rB}}{K_B}$	$*P_A = 0.4F_{rA} + K_A F_{aA}$ $P_B = F_{rB}$
 $\frac{0.47F_{rB}}{K_B} > \left(\frac{0.47F_{rA}}{K_A} + F_{ae}\right)$	$F_{aA} = \frac{0.47F_{rB}}{K_B} - F_{ae}$ $F_{aB} = \frac{0.47F_{rB}}{K_B}$	$*P_A = 0.4F_{rA} + K_A F_{aA}$ $P_B = F_{rB}$
 $\frac{0.47F_{rB}}{K_B} \le \left(\frac{0.47F_{rA}}{K_A} + F_{ae}\right)$	$F_{aA} = \frac{0.47F_{rA}}{K_A}$ $F_{aB} = \frac{0.47F_{rA}}{K_A} + F_{ae}$	$P_A = F_{rA}$ $*P_B = 0.4F_{rB} + K_B F_{aB}$
 $\frac{0.47F_{rA}}{K_A} > \left(\frac{0.47F_{rB}}{K_B} + F_{ae}\right)$	$F_{aA} = \frac{0.47F_{rA}}{K_A}$ $F_{aB} = \frac{0.47F_{rA}}{K_A} - F_{ae}$	$P_A = F_{rA}$ $*P_B = 0.4F_{rB} + K_B F_{aB}$

*If $P_A < F_{rA}$, use $P_A = F_{rA}$ and if $P_B < F_{rB}$, use $P_B = F_{rB}$

*If $P_A < F_{rA}$, use $P_A = F_{rA}$ and if $P_B < F_{rB}$, use $P_B = F_{rB}$

Figure 8.4.15 Dynamic equivalent radial load equations for single-row mounting of tapered roller bearings. (Courtesy of The Timken Company.)

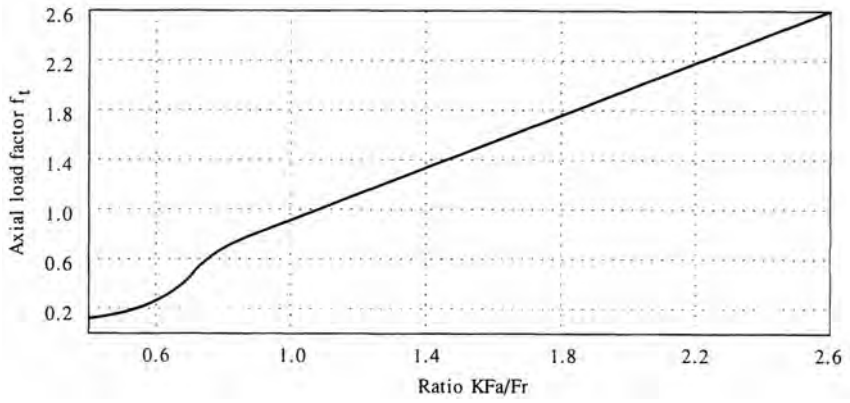


Figure 8.4.16 Equivalent axial load factor. If KFa/Fr > 2.5, use axial load directly. (Courtesy of The Timken Company.)

To accommodate various operating conditions, the flow rate should be chosen for the worst case and then a closed-loop control system used to measure the oil temperature and adjust the flow rate accordingly.

Alignment

A single tapered roller bearing can tolerate a misalignment of about 0.001 radians without seriously degrading the life of the bearing. However, for precision applications in order to obtain accurate motion, better alignment should be maintained: The cup and cone seat roundness should be at least as good as the radial error motion of the bearing, and the backing squareness to the bearing centerline should also be at most equal to the bearing's radial error motion.

A pair of bearings at one end of a bore acts to resist moment loads and thus has an alignment tolerance of essentially zero degrees with another pair used to support a long shaft. When using sets of tapered roller bearings at each end of a shaft (e.g., spindle), extreme care must be taken to ensure that the bore centerlines are coincident and that the surfaces the races mate with at each end of the shaft are concentric.

It is also very important to ensure that the surfaces the cup and cone seat against be square with respect to the bearing centerline. Seating surface squareness can be enhanced by grinding the seating surfaces and the shaft all at once, as shown in Figure 8.4.17. Since adjusting nut threads are never perfect, an alternative is to use a ground spacer with an adjusting nut whose inner surface is ground flat.

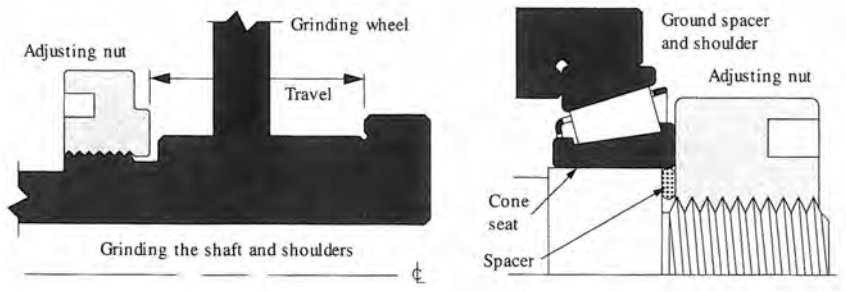


Figure 8.4.17 Methods for improving seating surface squareness. (Courtesy of The Timken Company.)

Preload adjustment

A tapered roller bearing has a higher stiffness than an angular contact bearing, and thus it is more difficult to grind an offset in the races or make a spacer that will allow the preload to be established via meeting a geometric displacement constraint; hence when economy is of prime importance the preload is often set using bolts that are tightened until the proper torque is reached.

Furthermore, because the bearing is stiffer, the Poisson effect (axial expansion of the outer race when it is pressed into a bore or axial contraction of the inner race when it is pressed over a shaft) is more significant for tapered roller bearings than for angular contact bearings; hence when spacers are used to control the preload, extreme care must be taken and the manufacturer's guidelines followed to the letter.

The preload has a dramatic effect on the number of rollers loaded and hence on the stiffness as was shown in Figure 8.3.7.[49] As shown in Figure 8.4.18, the difference in stiffness as one goes from 10% of the radial load that gives a 10% chance of failure at 500 rpm after 3000 h to 90% is small. This is due to the fact that the increase in load does not necessarily lead to more rollers being loaded, the way an axial preload does. However, when axial preload is supplied to increase the load zone, a huge increase in stiffness is realized. Figure 8.4.18 shows other effects of preload on physical parameters of tapered roller bearing systems. For most other types of bearings, the same trends will be exhibited. Note that if too much clearance between the cup and bore and cone and shaft exist, then the beneficial affects of preload can be greatly diminished; hence one must be careful to follow tolerance specifications provided by the manufacturer.

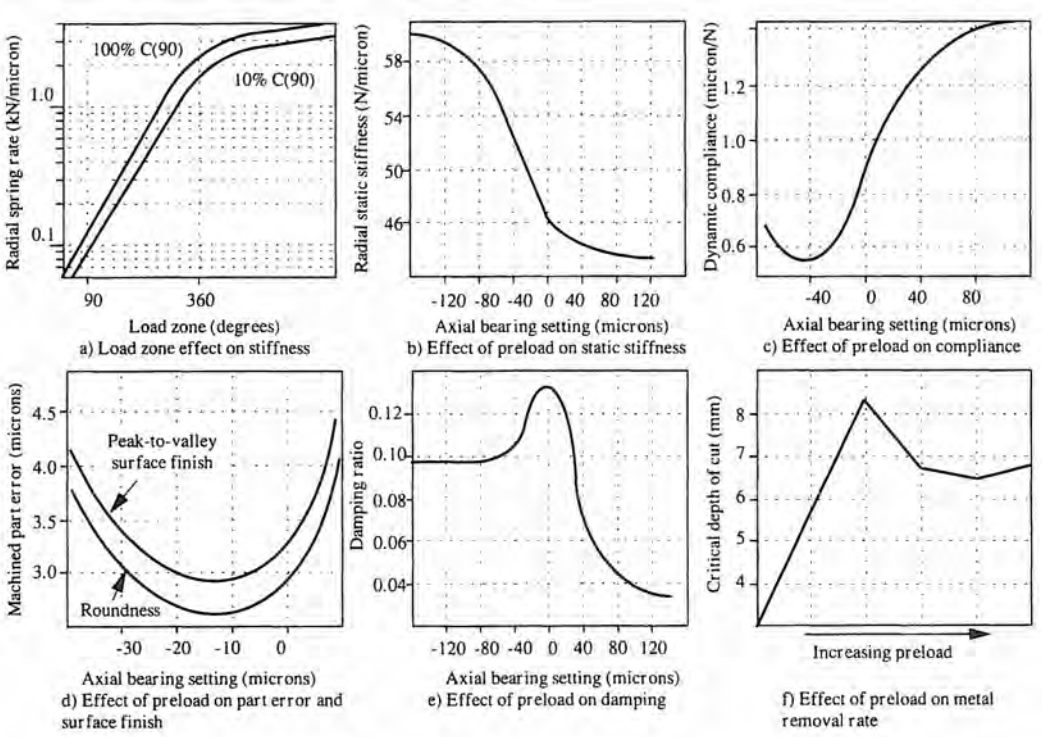
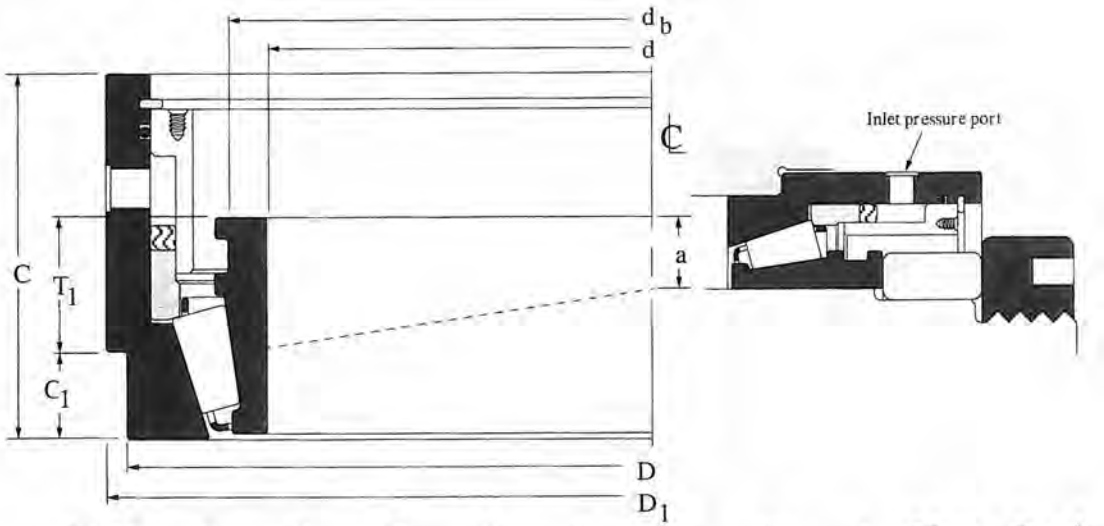


Figure 8.4.18 Effects of tapered roller bearing preload on system performance. (Courtesy of The Timken Company.)

Thermal growth and applied loads can change the preload on bearings, which in turn changes performance. To help maintain constant preload, one could use a large spring to seat the cup or cone, but then stiffness would decrease. An alternative is to force the tapered rollers farther into the space between the cup and cone. Figure 8.4.19 shows a Hydra-Rib® bearing designed to do this. The Hydra-Rib® bearing is usually used as the rear bearing in a spindle assembly. With the Hydra-Rib® design and a servo-controlled pressure system, the dynamic stiffness of the assembly can be changed in real time to suit the process. This capability provides the design engineer with a whole new way to approach spindle design.

[49] See the discussion on stiffness in Section 8.3.1.



Series #	d	D	D_1	T_1	C	C1	d_b	a	G factor	F_{radial}*
JP5000	50	104	103.5	31	66	15	61	-13.0	32.5	20.6
JP7500	75	122	130	22	65	10	85	-10.2	60.5	23.1
JP8500	85	140	148	23.5	66	10	96	-9.1	86.6	30.1
JP11000	100	170	178	27	70	12	114	-10.4	140	34.1
JP11000	110	170	178	27	70	12	122	-10.4	140	34.1
JP12000	115	180	190	26	70	12	128	-7.1	173	42.0
JP12000	120	180	190	26	70	12	132	-7.1	173	42.0
JP13000	125	190	200	29	72	12	138	-7.6	212	49.9
JP13000	130	190	200	29	72	12	142	-7.6	212	49.9
JP14000	135	205	213	27	72	13	148	-3.8	235	50.6
JP14000	140	205	213	27	72	13	152	-3.8	235	50.6
JP16000	155	227	235	30	76	15	169	-5.1	318	60.9
JP16000	160	227	235	30	76	15	172	-5.1	318	60.9
JP17000	170	240	248	30	79	15	182	-4.8	365	62.9
JP18000	180	260	268	30	84	15	193	-1.8	398	63.6
JP20000	200	282	290	32	83	17	213	-0.8	480	74.6
JP22000	220	308	316	32	83	17	233	+5.3	588	80.8

Units are mm and kiloNewtons
* Radial load rating (kN) @ 500 rpm for 3000 hours L10.

Figure 8.4.19 Characteristics of Timken's Hydra-Rib® bearing. (Courtesy of The Timken Company.)

8.4.8 Spherical Roller Bearings

Spherical roller bearings, illustrated in Figure 8.3.2, are the roller bearing equivalent of self-aligning ball bearings. The spherical rollers rest in a spherical race and allow for shaft misalignment on the order of 2°. Single-row spherical roller bearings, where the centers of the rollers are parallel to the rotation axis, have very high radial load capacity but low axial load capacity. Double-row spherical roller bearings with inclined roller axes have high radial and thrust capacity. These bearings are even more difficult than tapered roller bearings to manufacture as far as obtaining high precision, and thus are used mainly in large industrial machinery, where it is difficult to align large bores with the precision required by tapered roller bearings. Spherical roller bearings are typically preloaded by axially displacing the inner ring with a tapered inside diameter over a tapered shaft. For reversing thrust loads, care must be taken to provide an oppositely configured bearing or an opposed axial restraint for the inner ring so that the thrust load does not pull the bearing further onto the tapered shaft, thereby causing overloading of the bearing.

8.4.9 Thin Section Bearings

The problem with many bearings is as the bore diameter increases, so does the bearing cross section.[50] In many cases where space is at a premium (e.g., robotic applications) a large bore is required, so power transmission shafts and wires can pass through a joint, yet to use a traditional bearing would require too large a bearing for the application. Hence thin section bearings evolved which are bearings that have a constant cross section regardless of diameter. Figure 8.4.20 shows typical size ranges of available thin section ball bearings. Thin section bearings are available in Conrad, angular contact, four-point contact, crossed roller, and tapered roller types, and have essentially the same mounting and performance characteristics as their thicker cousins.

		Bore diameter (inches)																										
Bearing series:		1	1.5	2	2.5	3	3.5	4	4.25	4.5	4.75	5	5.5	6	6.5	7	7.5	8	9	10	11	12	14	16	18	20	25	30	35	40
KAA series	A	•	•																										
3/16" radial	C	•	•																										
cross section	X	•	•																										
KA series	A			•	•	•	•	•	•	•	•		•	•		•			X										
1/4" radial	C			•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•								
cross section	X			•	•	•	•	•	X	•	•	•	•	•	•						X								
KB series	A			•	•	•	•	•	X				X	•					•										
5/16" radial	C			•	•	•	•	•	•	•		•	•	•				•											
cross section	X			•	•	•	•	•	•	•			•	•	•			•	•										
KC series	A							•		•	X	•	X	•		•	X				X								
3/8" radial	C							•	•	•	•	•	•	•	•	•	•	•	•	•	•	•							
cross section	X							•		•		•	•	•	•	•		•	•	•	•	•		•					
JU series	A																												
3/8" radial	C							•		•		•	•	•	•		•	•	•	•	•								
cross section	X							•		•		•	X		•		•	•	•	•	•								
KD series	A							•	•	•	•	•	•	•	•	•	X	•	•										
1/2" radial	C							•	•	•	•	•	•	•	•	•	•	•	•	•	•	•							
cross section	X							•	•	•		•	•	•	•	•	•	•	•	•	•	•	•		•		•		
KF series	A										•		•	X	•		X	X	X	X		•							
3/4" radial	C							•	•	•	•	•	•	•	•		•	•	•	•	•	•	X						
cross section	X							•		•	•	•	•	•	•		•	•	•	•	•	•	•						
KG series	A								X					•	X		•	•	•			•	•	X	X	•	•	X	
1" radial	C								X		X	•	•	•	•	•	•	•	•	•	•	•	•	•	•	X	X	X	
cross section	X									•		•		•		•	•	•	•	•	•	•	•	•	•	•		•	

• Available from stock X Limited availability

Figure 8.4.20 Typical size ranges of angular contact (A), Conrad (C), and four-point contact (X) thin section ball bearings. (Courtesy of Kaydon Corp.)

Cross roller bearings are a type of thin section bearing that can support high radial, axial, and moment loads because of line contact between the rollers and the races, and they can do so with less slip than can four-point contact Gothic arch ball bearings. This allows cross roller bearings to achieve high accuracy and have low friction. Figure 8.4.21 compares two designs for guiding the rollers to prevent them from skewing. The spacers allow a larger portion of the roller surface to be in contact with the race. As also shown in the figure, if the cage thickness and grinding relief width are not carefully chosen, then a greater force couple (moment) will exist on the roller than for the design that uses a spacer. Cross roller bearings are used where space for only a single bearing exists and the bearing must support radial, axial, and moment loads. When mounting cross roller bearings, it is important to make sure that the split outer ring is securely clamped by the structure. Figure 8.4.22 shows typically available sizes of cross roller bearings. Most cross roller bearings are available with a lubrication hole (1-3 mm diameter) along the line of the split ring.

[50] The bearing cross section is defined as follows: The width is the projected axial dimension of the inner and outer races. The height is the difference between the outer-race outer radius and the inner-race inner radius (i.e., one half the difference between the housing and shaft bores).

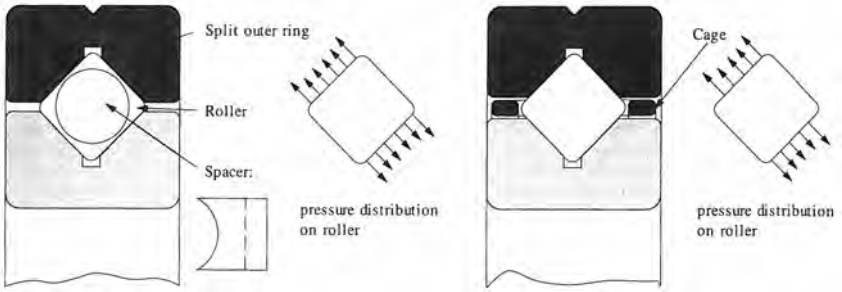


Figure 8.4.21 Comparison of cross roller bearings with spacers and cages. (Courtesy of THK Co., LTD.)

Model	Weight N	ID mm	OD mm	Height mm	Inner ring height tolerance (mm)	Outer ring height tolerance (mm)	C kgf	Co kgf
RB3010	1.2	30	55	10	+0 -0.120	+0 -0.150	660	520
RB4010	1.6	40	65	10	+0 -0.120	+0 -0.150	750	650
RB5013	2.7	50	80	13	+0 -0.120	+0 -0.150	1510	1290
RB6013	3.0	60	90	13	+0 -0.150	+0 -0.200	1630	1500
RB7013	3.5	70	100	13	+0 -0.150	+0 -0.200	1750	1700
RB8016	7.0	80	120	16	+0 -0.150	+0 -0.200	2720	2590
RB9016	7.5	90	130	16	+0 -0.200	+0 -0.250	2830	2780
RB10020	14.5	100	150	20	+0 -0.200	+0 -0.250	3020	3160
RB11020	15.6	110	160	20	+0 -0.200	+0 -0.250	3110	3350
RB12025	26.2	120	180	25	+0 -0.200	+0 -0.250	6350	6570
RB13025	28.2	130	190	25	+0 -0.250	+0 -0.300	6590	7020
RB14025	29.6	140	200	25	+0 -0.250	+0 -0.300	7090	7910
RB15013	6.8	150	180	13	+0 -0.250	+0 -0.300	2440	3240
RB15025	31.66	150	210	25	+0 -0.250	+0 -0.300	7280	8360
RB15030	53	150	230	30	+0 -0.250	+0 -0.300	9210	9840

Figure 8.4.22 Typically available crossed roller bearings. Sizes up to 1250 mm ID are available. (Courtesy of THK Co., LTD.)

8.4.10 Summary

When sketching the concept of a machine, the discussion presented here combined with information in manufacturers' catalogs should be sufficient for initial selection of bearings and developing layout drawings of a machine. However, detailed design of precision rotary bearings systems can be difficult because of all the subtle parameters that one must consider and the complex manner in which they interact. Practically, in order to quickly and efficiently apply all the necessary equations to predict load capacity, stiffness, life, temperature rise, and so on, one must use a digital computer. This also will allow one to conduct parametric studies such as what will be the effect of increasing preload or bearing diameter on system stiffness and heat generation? Fortunately, rotating machinery design software systems are available.[51]

8.5 Rolling Element Linear Motion Bearings

Linear motion rolling element bearings are among the most important elements in precision machine tools; hence basic operating characteristics as well as different available types will be discussed here in detail. The following types of linear motion rolling element bearings are considered:

- Nonrecirculating balls or rollers

[51] For example, the ROMAX® software design package for rotating machinery which is available from Engineering Technology Associates of Southfield, MI. In addition, some bearing manufacturers have time-sharing programs that allow the designer to access them with a computer and a modem. This can allow the designer to rapidly play "what if" games when selecting a bearing and predicting its performance in the assembly, for example, Timken Company has a SELECT-A-NALYSIS® bearing selection and analysis program. Most bearing manufacturers also have in-house programs that can predict performance as a function of operating parameters (e.g., bearing temperature as a function of load, speed, and type of lubrication).

- Recirculating balls
- Recirculating rollers

Before choosing a rolling element linear motion bearing, there are several fundamental issues to consider, including:

- Balls or rollers, which to use?
- To retain or not to retain?
- Shape of the contact surface
- To recirculate or not to recirculate?
- Bearing spacing
- Selection criteria

There are as many manufacturers of linear rolling element bearings as there are manufacturers of rotary motion rolling element bearings; hence *caveat emptor!*

Balls or Rollers, Which to Use?

Balls can easily be made round to the submicron level by an automated lapping process. Generally, the variation of ball sizes in one lot is very small, so each ball does not have to be measured. When it is desired to size each ball, finished balls are rolled down two slowly diverging rails and same-size balls will drop through the rails at specific locations. Once installed in a bearing, balls do not skew sideways like rollers can and thus are always aligned in the grooves they ride in; however, point contact with the groove limits load capacity and stiffness. Rollers are typically produced to the $^1/_2$-1-μm level of roundness and size by centerless grinding. Rollers can (with greater difficulty than balls) be lapped to submicron accuracy. Line contact with the rolling surface enables rollers to support large loads and have high stiffness. It is also more difficult to grind a groove round than lap a plane flat, but it is easier to make a ball spherical than a roller cylindrical.

In order to quantify the differences between load capacity and stiffness of ball or roller systems, one can review the equations for stiffness and contact stress given in Section 5.6, but since the forms of the equations for deflection of spheres and cylinders are very different (cube roots as opposed to natural logarithms) simple ratiometric calculations and comparisons are not plausible. Hence in order to choose between balls or rollers, one needs to do a detailed numerical comparison and/or compare stiffness values given by bearing manufacturers. In general, roller bearings are used for large machines (spindle power $>$ 20 kW) and ball bearings are used in smaller machines.

In either case, when submicron accuracy or repeatability is required, inspection of individual balls or rollers may be required. If a rolling element has a high spot, then it will create a straightness error with a period of πD. If the rolling element is oval in shape, then the error period will be πd/2.

To Retain or Not to Retain?

A retainer can have three functions: (1) to keep the balls or rollers from falling onto the floor when the bearing unit is removed from the rail, (2) to keep rollers moving in a straight line and prevent them from skewing sideways, and (3) to keep the balls or rollers from rubbing against each other. No balls or rollers are perfectly round, and thus during rolling some slip can occur, which changes the spacing between rolling elements. The concept of nonround and different-size elements rolling in a bearing is shown in Figure 8.5.1. It is possible to statistically consider the out-of-roundness and size variation and then mathematically predict what the load, stiffness, and lateral force variation characteristics will be.[52] However, this type of analysis is not used when choosing bearings from a catalog and is used only by engineers who design their own rolling element bearings.

Rolling elements will also invariably touch unless a method is used to separate them. Unfortunately, when rolling elements touch, their surface velocity vectors are in opposition and rubbing occurs. Many types of linear bearings and ballscrews cannot use a mechanical retainer to keep the balls or rollers separated, and hence there will be some intermittent rubbing unless smaller diameter spacer balls are used to separate all the load-carrying balls. At low speeds the amount of rubbing is usually insignificant; but at high speeds rubbing balls act like a noise source which can affect the position repeatability of the servodrive and can create nonrepeatable angular and straightness errors on the microradian and submicron level, respectively. Note that the use of spacer balls or rollers decreases the number of load-carrying elements by two, so load capacity and stiffness are decreased.

[52] See Z. M. Levina, "Research on the Static Stiffness of Joints in Machine Tools," Proc. 8th Int. Mach. Tool Des. Res. Conf., Sept. 1967, pp. 737-758.

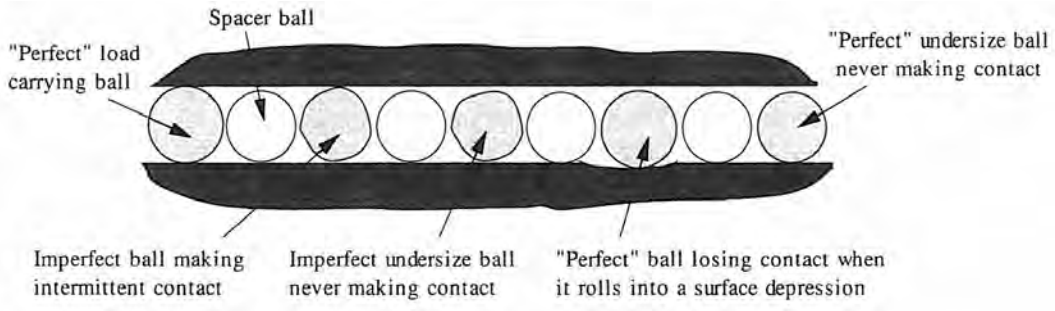


Figure 8.5.1 Rolling elements are not necessarily round and of the same size.

Shape of the Contact Surface

For rolling balls, the shape of groove the balls roll in affects the load capacity, stiffness, and dynamic and static coefficients of friction. Figure 8.5.2 illustrates the two most common groove designs, the circular arc and the Gothic arch. A ball in a circular arc groove makes only two-point contact with the groove, and thus at least four grooves are required to support bidirectional loads. The advantage is that almost pure rolling motion exists because there is very little differential slip caused by the contact surface's latitudinal range. The balls can be arranged in a back-to-back or a face-to-face configuration; the former provides higher moment resistance, while the latter provides more self-aligning capability. Thermal stability associated with such arrangements is generally not an issue since the velocities are usually low compared to rolling element rotary motion bearings.

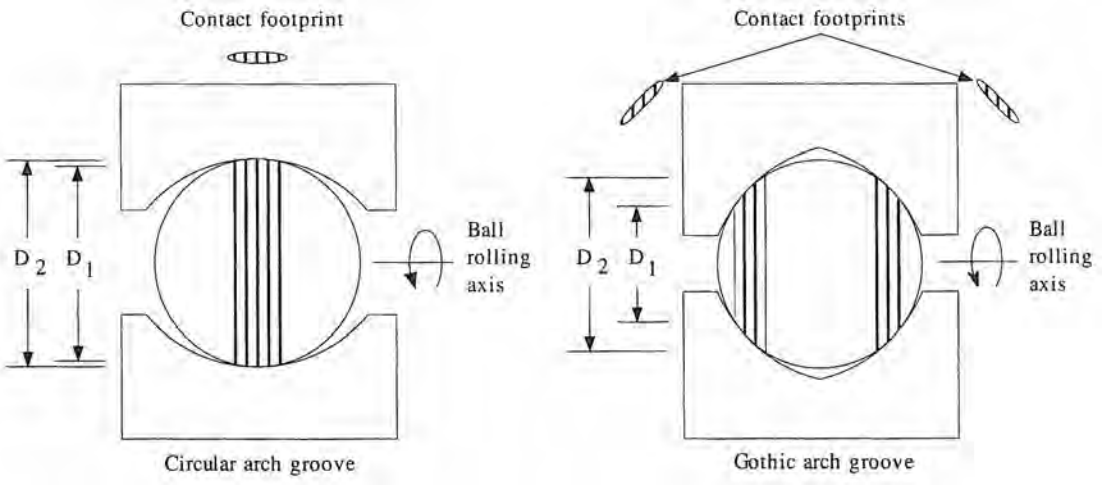


Figure 8.5.2 A circular arch groove typically experiences 3% slip during rolling compared to about 40% for a Gothic arch groove.

A ball in a Gothic arch can make four-point contact with the groove, and thus only two grooves are required to support bidirectional loads. This can, however, result in substantial differential slip, which greatly increases friction, so sometimes four Gothic arch grooves are used, where an offset exists between the grooves on the bearing rail and the bearing carriage; thus the balls make two-point contact with the grooves only under normal load conditions. When overloaded, one set of balls in a groove makes contact with the other side of the arch, thereby helping to support the overload.

Figures 8.5.3 and 8.5.4 illustrate the design dilemma faced by a Gothic arch groove linear motion bearing that has full four-point contact between each set of balls and their respective grooves (i.e., no groove offset). In order to reduce the differential slip, a much larger groove radius is required, which decreases the load capacity. In fact, if the differential slip were to be reduced to that of a circular arch, then the load capacity would be on the order of the circular arch groove bearing. Some design engineers want the extra load capacity a Gothic arch has the potential to provide;

Gothic arch groove bearings are thus available with an offset between the grooves, so the balls nominally have two major points of contact and two minor points of contact. This results in load capacity and differential slip somewhere between a two-point circular arc design and a four-point Gothic arch design.

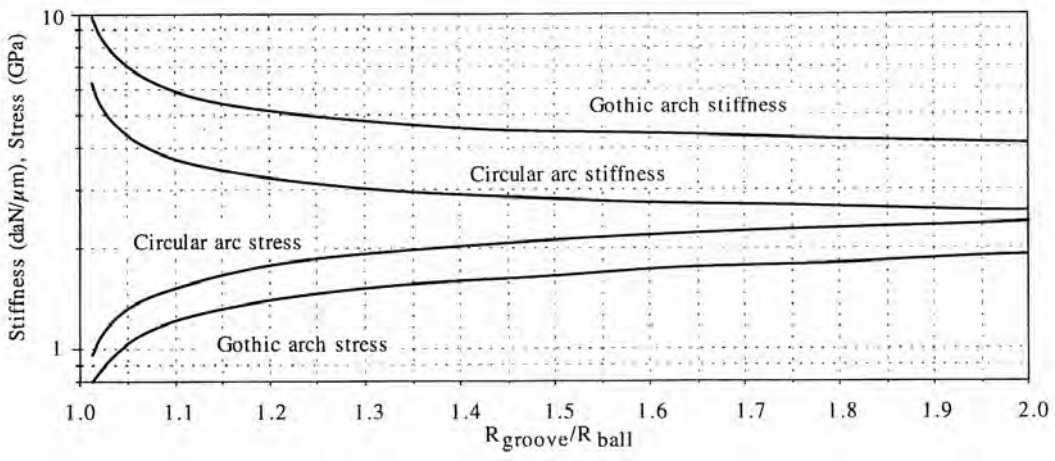


Figure 8.5.3 Hertzian stress and stiffness between balls and different groove shapes: Numerical results are for 5 mm diameter balls, each supporting 100 N of force.

It is difficult to generalize further because other issues (e.g., component accuracy and surface finish, return path design, metallurgical properties, etc.) also affect bearing performance. In addition, one can always find a bearing, of either groove design, that meets one's load capacity and stiffness requirements. Therefore, in order to evaluate a manufacturer's products, one should ask for data on their bearings' load capacity, stiffness, and static and dynamic coefficients of friction. It is perhaps most important to obtain data on accuracy as a function of travel and applied load. Do not be shy about asking the manufacturer how his product compares with those of his competitors. How well a company knows other products and technologies is an indication of the worth of their product recommendations.

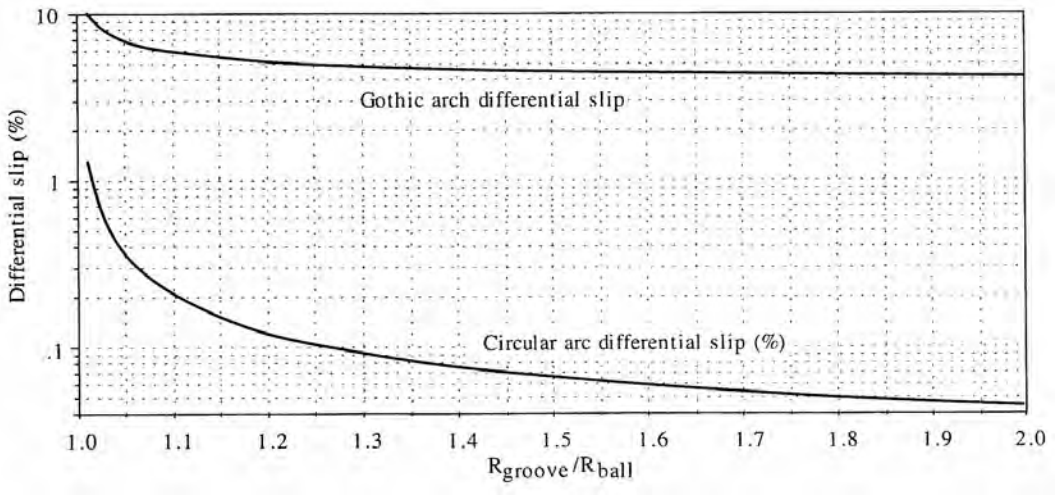


Figure 8.5.4 Differential slip between rolling balls and different groove shapes: Numerical results are for 5 mm diameter balls, each supporting 100 N of force.

To Recirculate or Not to Recirculate? That Is the Question

The preload on a bearing is needed to increase stiffness. This means that when a ball leaves the load path on its way to the recirculating tube, it expands. Similarly, when it leaves the recirculat-

ing tube and reenters the load path, it is compressed. The result on the submicron level is tiny force inputs orthogonal to the axis of motion similar to what one feels when one uses a serrated knife. The more balls or rollers in the path, the greater the stiffness of the bearing and the less the effect this orthogonal force noise has on the bearing's high-frequency straightness error (smoothness). The bumpiness can be reduced when the manufacturer provides gradual entrance/exit regions; however, some noise will always be present. For general machine tool applications where 1 μm performance is sufficient, recirculating bearings perform admirably; however, for submicron (1-10 μin.) applications the design engineer should try to use nonrecirculating bearings if possible or at least carefully measure the performance in a prototype machine of a statistically significant group of bearings intended for use.[53] Other bearing alternatives, such as hydrostatic or aerostatic bearings, should also be considered for submicron applications.

Bearing Spacing

When using multiple rolling elements to support a load, typically the system is overconstrained anyway, so one should not be shy about supporting a carriage at all four corners. When using modular bearing elements, in order to minimize the length of the bearing rail, one usually tries to minimize the axial spacing between the bearing elements. However, because rolling element bearings still have a finite amount of static and dynamic friction, the greater the ratio of the longitudinal-to-latitudinal (length-to-width) spacing, the smoother the linear motion will be and the less the chance of walking.

One should first do calculations to size bearings and their spacing to yield desired load and stiffness capabilities with the notion that in order to prevent walking (rapid yawing or pitching during low forward velocities) the ratio of the longitudinal to latitudinal spacing of bearing elements should ideally be on the order of 2:1. For the space conscious, the bearing elements can lie on the perimeter of a golden rectangle (ratio about 1.6:1). The absolute minimum length to width ratio is 1:1. The higher the speed, the higher the length-to-width ratio should be. For large moving bridge machines it is often necessary to provide actuators and position sensors on both sides of the bridge with one servo system slaved to the other.

Sadly, when systems are overconstrained, it is difficult to perform neat closed-form analysis like that presented in Section 2.2.4. However, this type of analysis can be extended to nonkinematic systems as long as the appropriate spring constants and geometric compatibility equations are used. Section 8.5.2.4 discusses analysis methods for determining forces and deflections of linear guide bearing systems which may be extrapolated to other types of bearing systems.

Selection Criteria

Selection criteria to consider when designing a system supported by linear bearings should include:

- Running parallelism, repeatability, and resolution
- Lateral and moment load support capability
- Allowance for thermal growth
- Alignment requirements
- Preload and frictional properties
- Life

As with rotary motion systems, one must be able to visualize forces and moments as "fluids" and see how they flow from the carriage to the bearing to the machine. For machine tool applications where high cutting forces and moments must be resisted, one is virtually required to use an overconstrained bearing arrangement; however, one still needs to visualize how forces and moments are distributed so that one can determine the number, size, and spacing of bearing elements to be used.

There are many different types of linear rolling element bearings suitable for use in precision machines, and hence one must diligently consider all types and all manufacturers before selecting a bearing.[54] Commercially available systems assembled from precision-ground components can typically attain 1-10 μm/m running parallelism (moving bearing surface with respect to rail surface), while hand lapped and inspected systems can sometimes realize a one- to two-order-of-magnitude

[53] See Figure 8.3.7 and the accompanying discussion.

[54] Once again, the Thomas Register can prove to be an invaluable reference source. Also, one must continually read design magazines and update one's catalog files with company literature.

improvement. Of course these rough performance guidelines are subject to many factors, as discussed in Section 8.3.1.

From a production point of view, one wants high quality with good price and delivery. From a prototype point of view, if you choose a manufacturer without comparison shopping and without subjecting each manufacturer's product to a value analysis with respect to the design it will be used in, the prototype is likely to fail and the project to be canceled if the bearing's performance is bad. It is the wise design engineer that selects at least two possible options and does preliminary tests to determine which bearing is best for the intended application. One should also ask questions like:

- What is the manufacturer's reputation?
- Has the manufacturer supplied bearings for a similar application before?
- Can the manufacturer provide data on the accuracy as a function of load and travel?
- Is friendly intelligent design assistance offered?
- Does the manufacturer offer a usable bearing from stock?
- How long would a custom order take to be delivered?
- How does the use of an available stock bearing affect the rest of the design?
- What are the prototype and production quantity costs?

The reader, his management, and others are likely to think of many other points to consider, but these should at least help stimulate the discussion.

One should always consult with the bearing manufacturer for advice on selecting a bearing. For selection of linear bearings, however, one should note that some salesmen are not as conservative as they should be because they realize that the higher material cost associated with rolling element linear bearings makes some design engineers want to use sliding bearings even though the cost of the system is often roughly the same; hence the salesmen are sometimes inclined to recommend the cheapest (smallest) bearing they honestly believe will work for the application.

However, nobody knows the machine as well as the engineer designing it, and thus it is wise to construct a test jig using the recommended bearing and the next-larger-size bearing. The test jig can be mounted on the carriage of a larger machine with known good damping characteristics so that the dynamics of the test jig will be dominant. The test carriage can be loaded with weights to simulate maximum part or cutting loads. To simulate vibration loads from cutting, a pneumatic or electric vibrator can be attached. The carriage can even be repeatedly "crashed" to check its shock resistance. One may even wish to put a nonmoving linear bearing-supported axis on top of the test carriage to evaluate the ability of the design to resist fretting for situations where one axis may not be used while making a production run of parts. Some test cuts may also be made to evaluate the machineability performance of the design.[55]

8.5.1 Bearings with Nonrecirculating Balls or Rollers

Nonrecirculating elements travel half as far as the carriage they support, as shown in Figure 8.5.5. Their primary application is where short travel and compact design are needed. They are easy to seal with a lip seal or bellows enclosure, and they are relatively inexpensive and easy to install and repair. They can be lubricated for life or lubricated with a wick or some other automatic device. Various types of nonrecirculating bearing element linear bearings are discussed below.

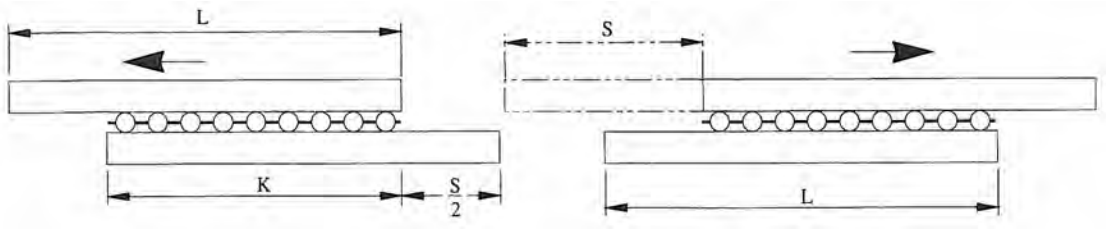


Figure 8.5.5 Rolling motion of nonrecirculating rolling elements in a linear bearing.

[55] *"It takes less time to do a thing right than to explain why you did it wrong."* Henry Wadsworth Longfellow.

8.5.1.1 Balls on Cylindrical Surfaces

Balls can be held by a cage and then allowed to act as an interface between a cylindrical housing and a shaft. This configuration enables the shaft to undergo limited linear and unlimited rotary motion. If the shaft and bore are carefully lapped to size and precision balls are used, then it is possible, for example, to design a quill for a spindle that has submicron accuracy and nanometer resolution. For more common applications, such as in punch presses, commercially available shafts, bores, and balls-in-cages are available. Note that contact stress will be higher and stiffness lower for a ball on a round shaft than for any other rail geometry.

Running Parallelism, Repeatability, and Resolution

Running parallelism is a measure of straightness accuracy, repeatability is a measure of how well the straightness errors repeat, and resolution is a measure of how friction and rolling quality affect how small a motion increment a servo system (e.g., sensor, actuator, and controller) can make the bearing-supported component move. Nonrecirculating balls on a ground round shaft can achieve rotary motion with about 2-5 μm of total radial error motion, running parallelism of about 5-10 μm/m, and resolution (depending on the servo system) on the order of 0.1-1.0 μm. If the bore and shaft are lapped and the balls carefully inspected, a one- to two-order-of-magnitude increase in performance can be obtained, but at a substantial increase in price. Note that craftsmen who can perform this type of lapping are not common, so one should not specify it unless one has the in-house capability.

Lateral and Moment Load Support Capability

This type of arrangement is meant primarily to provide two degrees of freedom. In order to support large moment loads about other axes, one should use two sets of bearings, one at each end of the shaft. Load capacities are similar to those of recirculating balls on round rails (see Section 8.5.2.2). For custom-designed systems, one would have to perform a detailed analysis similar to that used to determine loads on individual balls in a rotary motion ball bearing. Note that this type of bearing is not generally meant to be used in parallel to support a linear carriage.

Because the balls are resting on a cylindrical surface, this type of bearing arrangement is less stiff than a bearing where the ball rests in a circular arc or Gothic arch groove. As a first-order calculation, compare the equivalent radius of curvature of two cases. The first is where the ball rests in a circular arch groove where the groove radius of curvature is, for example, 1.2 that of the ball. In this case the equivalent radius of curvature is

$$R_e = \frac{1}{1/R_b + 1/R_b - 1/1.2R_b + 1/\infty} \approx 0.857\, R_b \tag{8.5.1}$$

Next assume that the ball rests on a shaft whose diameter is 10 times that of the ball; hence the equivalent radius of curvature is

$$R_e = \frac{1}{1/R_b + 1/R_b + 1/10R_b + 1/\infty} \tag{8.5.2}$$

For a ball the stiffness is proportional to the cube root of the equivalent radius of curvature; hence for this example the ball in the circular arch is about 22% more stiff than the ball on the cylinder. The contact stress is proportional to the $^{-2}/_{3}$ power of the equivalent radius, so the contact stress of the ball on the cylinder is 48% higher. Fortunately, there is usually plenty of room to load this type of bearing with lots of balls so that the stress and stiffness levels can be made acceptable.

Allowance for Thermal Growth

Since the bearings do not restrain axial motion, one only needs to make sure that differential radial expansion between the shaft and the housing is not too great. As with rotary motion spindles, this can be achieved by cooling the bearings with an oil mist.

Alignment Requirements

Each ball-cage assembly can resist moment loads so that the alignment requirements for the bore and shaft are the same as for a spindle supported by sets of angular contact bearings at each end.

Preload and Frictional Properties

Preload is obtained by loading the cage with oversized balls. This implies that the housing must be sufficiently rigid to prevent the inner pressure from the oversize balls from causing the housing bore to deform to an out-of-round shape, which will affect the error motion of the shaft. For this complicated type of loading, finite element analysis may have to be used. Start with the wall thickness of the housing at least equal to the radius of the shaft. Because the balls are in point contact at two points, rolling friction will be very low. Static and dynamic coefficients of friction on the order of 0.005-0.01 can be expected.

8.5.1.2 Balls or Rollers in Grooved Rails

Various types of nonrecirculating ball or roller bearings in grooved rails are shown in Figure 8.5.6. When preloaded against each other as shown in Figure 8.5.7, vertical, horizontal, and moment loads can be supported. Lower contact stress (and stiffness) and rolling friction can be obtained when the bearings are mounted with a vertical orientation so only the weight of the carriage preloads them. In applications where precise axial position control is needed, such as in an inspection machine, the latter mounting method is often preferred. One of the more common forms of this type of bearing is the crossed roller bearing, illustrated in Figure 8.5.8. The rollers are held by a cage, the length of the rollers is slightly less than the diameter, and the rollers are oriented 90°with respect to their adjacent neighbors. This prevents the ends of the rollers from rubbing on the side of the vee they are not primarily in contact with while allowing the assembly to support bidirectional loads.

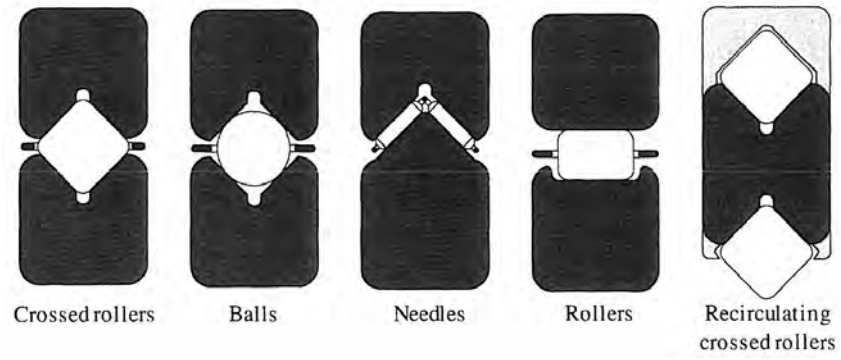


Figure 8.5.6 Variations on the ball or roller in groove theme.

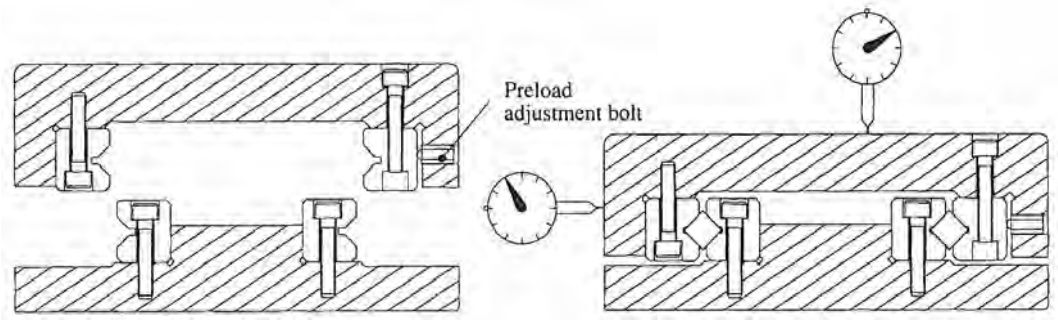


Figure 8.5.7 Typical assembly of crossed roller supported slide. (Courtesy of NSK Corp.)

Figure 8.5.9 shows a kinematic three-ball configuration. With this system, repeatability can be on the order of the surface finish of the components. Note that the balls in the vee groove would be in four-point contact and would thus be subject to slippage. One of the female vees could be replaced with a male vee and four balls could be used instead of the two while still maintaining a kinematic configuration. Preload for this type of system is attained by the weight of the carriage or by a capstan (friction) drive roller (not shown), which would also provide the axial drive force.

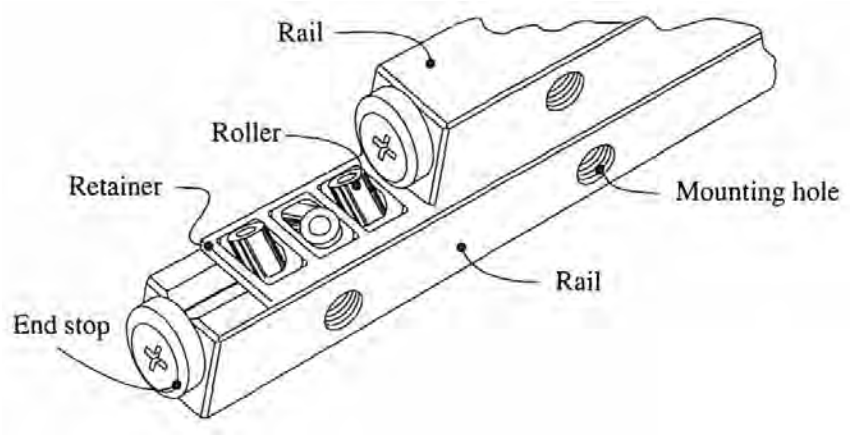


Figure 8.5.8 Crossed roller bearing construction. (Courtesy of NSK Corp.)

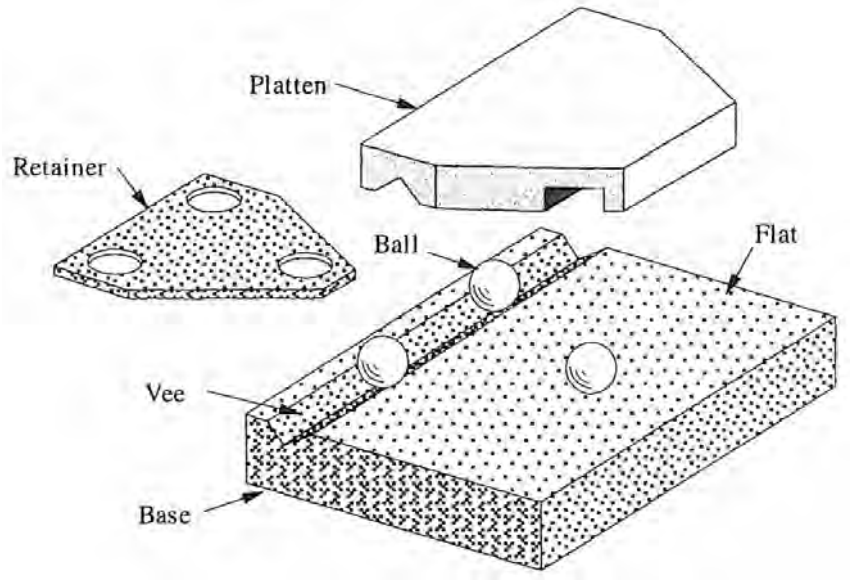


Figure 8.5.9 Vee and flat rails for nonrecirculating rolling ball support of an instrument platten.

Running Parallelism, Repeatability, and Resolution

Machine ground components in this configuration can achieve running parallelism of about 5-10 μm/m, repeatability of about $^1/_4$-$^1/_2$ μm, and resolution (depending on the servosystem) on the order of 0.1-1.0 μm. Figure 8.5.10 shows the typical effect of rail length on running parallelism. If the components are lapped and the rolling elements carefully inspected, a one- to two-order-of-magnitude increase in performance can be obtained, but at a substantial increase in price. An alternative is to use lapped balls, which are much less expensive than lapped rollers. However, note that the balls would be in four-point contact, so preload and loads must be kept low. Alternatively, one of the female vees could be replaced with a male and five balls and a special retainer used.

Lateral and Moment Load Support Capability

For crossed roller bearings, because adjacent rolling elements' roll axes are at 90° with respect to each other, the number of load-carrying elements depends on the way the bearings are loaded. When forces act to compress the two rails together, the number of loaded rollers equals the number of rollers. When forces act to shear the two rails with respect to each other, the number of loaded rollers is one-half the total number of rollers. When balls are used, each ball contacts each rail at two points, so ideally all balls are loaded irrespective of the way the rails are loaded.

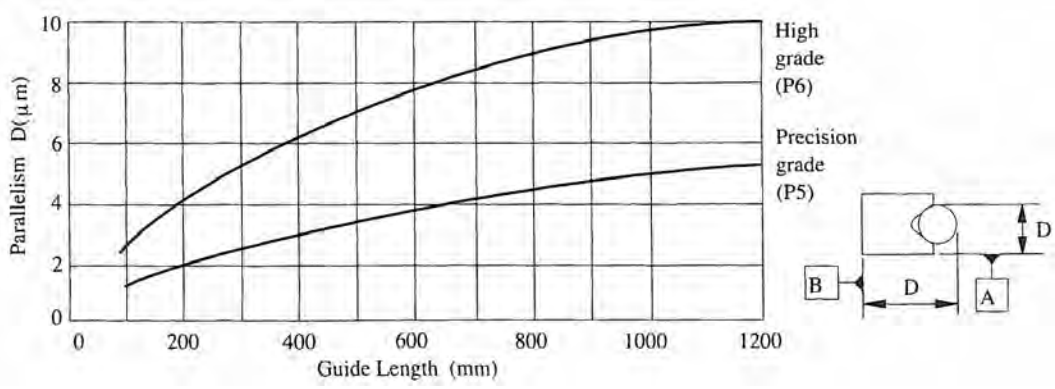
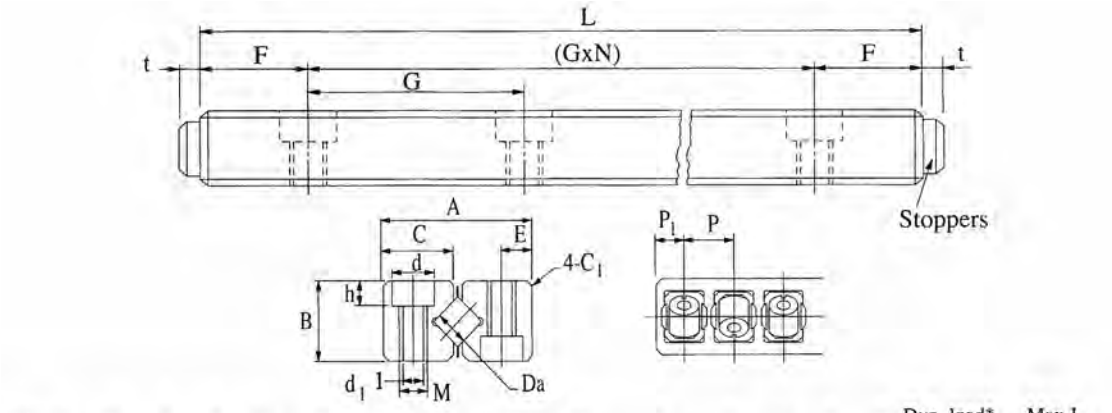


Figure 8.5.10 Accuracy grades of crossed roller bearings. (Courtesy of NSK Corp.)

Figure 8.5.11 shows one manufacturer's standard line of crossed roller bearings, and Figure 8.5.12 shows the load correction factor for the number of rollers used. To obtain the dynamic load capacity, one multiplies the rated capacity by the load correction factor corresponding to the actual number of loaded rollers. The static load capacity is just the rating for one roller times the number of loaded rollers. These relations are true for compressive or shear loads on the rails. One only needs to account for the actual number of loaded rollers and make sure the mounting is rigid so that the rails remain aligned. If the rails are not aligned properly, the cylindrical rollers may be loaded to form a tapered roller. The stiffness of crossed roller bearings is also often provided by the manufacturer. For example, Figure 8.5.13 shows the stiffness and stiffness correction factors for the bearings shown in Figure 8.5.11. If the rolling elements may be subject to fretting corrosion conditions (i.e., very slow speed and high vibration levels) they could be made from a ceramic material (e.g., silicon nitride), or stainless steel components could be used. To give an idea of cost, from the smallest to the largest bearing in a high-accuracy grade, the price varies from about $100 to $5000.



Model	Da	A	B	C	C1	E	d	h	d1	M	F	G	t	P	P1	Dyn. load* (kgf)	Max L P5	Max L P6
CRG04	4	24	12	11.3	0.3	5	8.0	4.2	4.3	M10x1.5	20	40	2.3	6.5	3.8	1000	200	300
CGR04	4	26	10	12.3	0.5	5	8.0	4.2	5.0	-	12/15	38/40	2.3	6.5	3.8	1000	200	300
CRG06	6	31	15	14.5	0.8	6	9.5	5.2	5.2	M6x1	25	50	3.2	9.5	5.8	2720	400	600
CRG09	9	44	22	20.8	1.0	9	11	6.2	6.8	M8x1.25	50	100	4.0	14	8.0	7420	600	900
CRG12	12	58	28	27.6	1.5	12	14	8.2	8.5	M10x1.5	50	100	5.0	20	12	13300	900	1200

* For twenty rollers. Dimensions are in mm.

Figure 8.5.11 Typical dimensional specifications for crossed roller bearings. (Courtesy of NSK Corp.)

Allowance for Thermal Growth

Since two parallel rails represent an overconstrained mounting, there is no way to accommodate thermal growth other than through (hopefully) elastic deformation of the components. If the bearings are used in quasi-steady applications, then thermal growth should not be a problem as

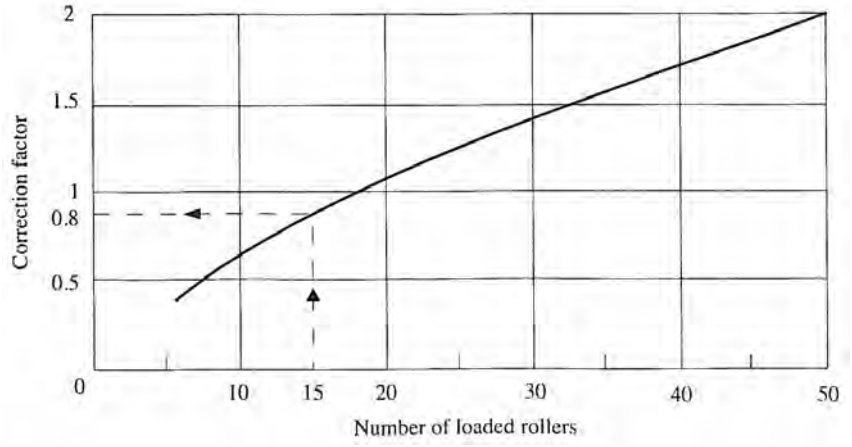


Figure 8.5.12 Load correction factor for crossed roller bearings. (Courtesy of NSK Corp.)

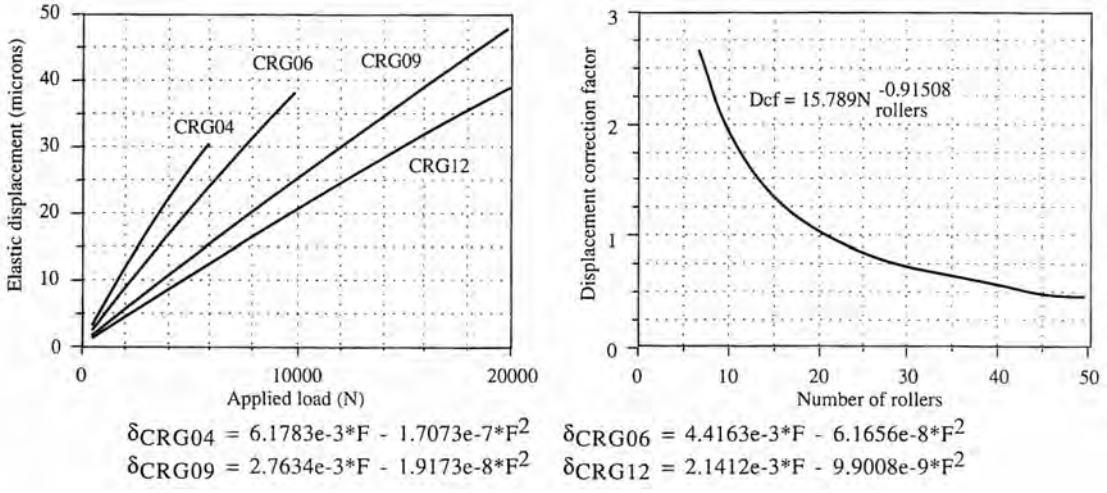


Figure 8.5.13 Stiffness of crossed roller bearings and approximate curve fits. (Courtesy of NSK Corp.)

long as the carriage and base to which the rails are mounted are made of the same material. For a high-speed reciprocating system, an active cooling system such as an oil mist should be considered if very high precision is to be maintained.

Alignment Requirements

As shown in Figure 8.5.7, one rail is typically mounted against a reference edge and the the other is preloaded against it through the use of preload bolts. This requires, however, that the two rails fixed to the base be absolutely parallel. Any misalignment between the two fixed rails will result in the worst case of an equivalent error being imparted on the carriage and nonuniform coefficient of friction over the carriage's range of travel. This in turn could make the servo-system response a function of position. Also note that as shown, the preload bolts will have a tendency to bend the rail so that a large number are needed. If a backing plate is used, one could only require that two preload bolts be used, which would make assembly easier but the system large. Another alternative would be to use one of the gib designs shown in Figure 8.2.4 or Figure 8.5.14. It is up to the design engineer to make the appropriate beam deformation calculations to determine how many bolts are needed, as discussed below.

Preload and Frictional Properties

Figure 8.5.13 shows the classical hardening effect that results from line contact. With a light preload on the order of 5-10% of the rated load, the static and dynamic coefficients of friction will

be between 0.01 and 0.005. For the purposes of dynamic modeling, one should make sure that the system will function anywhere in this range. There are many ways to preload crossed roller bearings, as shown in Figure 8.5.14. The longitudinal wedge (tapered gib) should be recognizable from sliding bearings. It is not easy to machine and usually requires that it be hand finished (scraped). The double longitudinal wedge can be machine finished (ground) but requires more space, and the extra surface decreases stiffness. The easiest and most common method is to use preload screws. Unfortunately, this leads to the most problems in terms of deforming the rails, as discussed below.

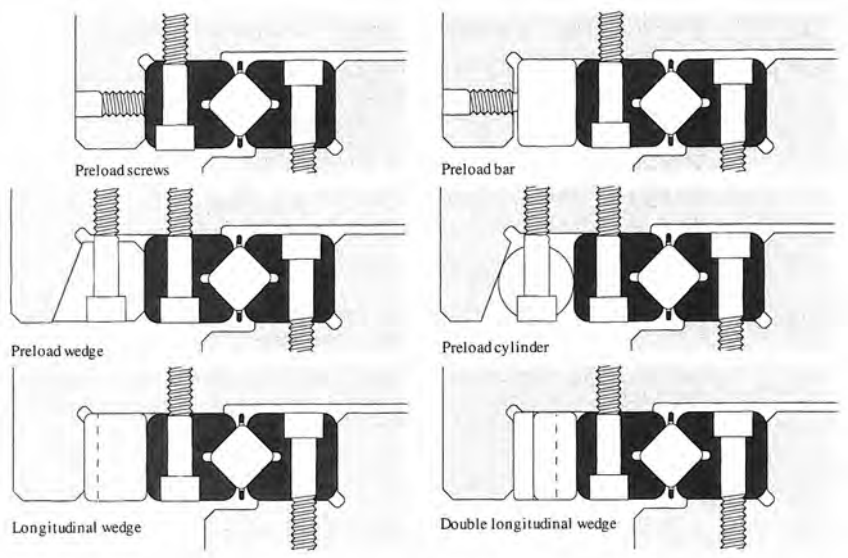


Figure 8.5.14 Methods for preloading crossed roller bearings. (Courtesy of Schneeberger Inc.)

The beam length l_b is modeled as being equal to the spacing between the preload bolts, and the beam width is b_o. A conservative first-order calculation for the case when there are many rollers (>5) is to assume that the bearing rail, or combination of the bearing rail and backing plate with their individual stiffnesses added, rests on an elastic foundation and that the ends of the beam are not guided (zero slope is not imposed). A positive preload force W per bolt is that which forces the beam into the foundation. The stiffness of the elastic foundation k is assumed to be the total stiffness k_b of the bearing rail being preloaded divided by the total length l_b of the bearing region (the product of the number of rollers and the pitch) and the width of the rail:

$$k = \frac{k_b}{b_o \ell_b} \tag{8.5.3}$$

Equation 7.5.30 gives the expression for the deformation of such a beam on an elastic subgrade. This expression does not include shear deformation effects, which can be on the order of the bending deformations for this type of stubby beam problem; hence one should double the result obtained from Equation 7.5.30 if l_b is less than twice b_o. An alternative to this expression would be to model each roller as a spring and use a somewhat complex closed-form solution or finite element model to determine the shape of the beam when it is supported by the discrete springs.

Example

Assume that we want to design a carriage like that shown in Figures 8.5.7 and 8.5.5 with 64 mm of travel. As a rule of thumb, the length of the roller cage should be at least twice this value (128 mm), and thus as shown in Figure 8.5.5, the rail length should be 160 mm. The table is used for positioning small assemblies for inspection so that any size of bearing can be used. To simplify the assembly of the table, it is desired to use only two preload bolts to preload the bearings used to at most 10% of the rated dynamic load in the vertical direction before they are locked into place with the mounting bolts. Is this goal feasible?

First try the largest bearing available which would only require two preload bolts, thereby simplifying assembly. The CRG12 series has a roller pitch length of 20 mm, so 8 rollers would be

required and l_b is 160 mm. From Figure 8.5.10 with the precision class bearings, one can expect a running parallelism of about 1.8 μm (72 μin.). The rated dynamic load in the vertical direction is

8220 kgf × 2 rails × 1/2 of bearings in contact × 0.5 load factor ≈ 4000 kgf

The rails are way oversized for the application, so only a 2% preload is required. Thus the total preload should be about 80 kgf or about 400 N per preload screw. The stiffness of the foundation beneath the rail being preloaded is found from Figure 8.5.13 and Equation 8.5.3. From Equation 7.5.30 the deflection is found to be 1.07 μm (42 μin). The ends will overhang 30 mm, so the deflection of the ends of the rail will then be in the middle.

Next try the smallest bearing available, the CRG04 series, which has a roller pitch length of 6.5 mm, so 25 rollers would be required. The preload bolt spacing is 40 mm, so four preload bolts would be required. The rated dynamic load in the vertical direction is

617 kgf × 2 rails × 1/2 of bearings in contact × 1.2 load factor ≈ 740 kgf

Thus the total preload should be about 74 kgf or about 185 N per bolt. Proceeding as before, the deflection is found to be 1.09 μm (43 μin.). For multiple preload bolts, one would have to model the effects of varying bolt friction and tightening torque on the deflection of the rail and design an appropriate quality control procedure. The cost of a more complex assembly procedure would have to be taken into account along with the cost of the bearings and structure required to hold the bearings. Which design would you choose?

8.5.1.3 Rollers on a Flat Rail

Wherever a sliding bearing is used to provide linear motion, rollers held by a cage and rolling on a flat rail can often be used instead. This includes T, dovetail, double vee, and vee and flat configurations such as those shown in Figures 8.2.5, 8.2.13, 8.2.15, and 2.2.7 respectively. Modular bearings of this type are shown in Figure 8.5.15, and examples of available sizes are shown in Figure 8.5.16. Note that modular rolling surfaces (rails) are also available. When high load capacity and stiffness are needed, rollers held by a retainer and rolling on hardened ground (or lapped) flat rails can be used. "Flat" does not necessarily mean a horizontal plane; the flat surfaces can be inclined.

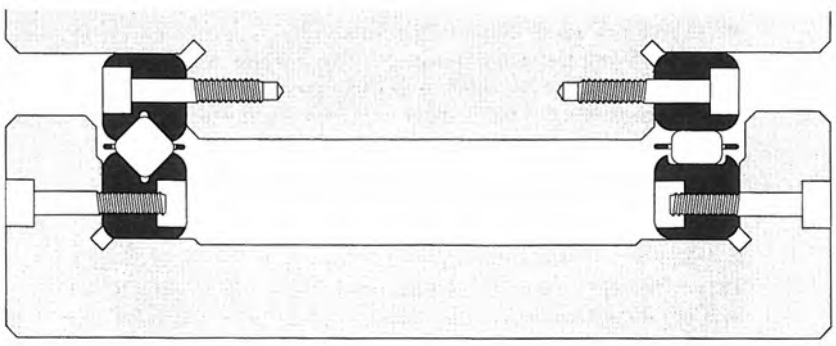


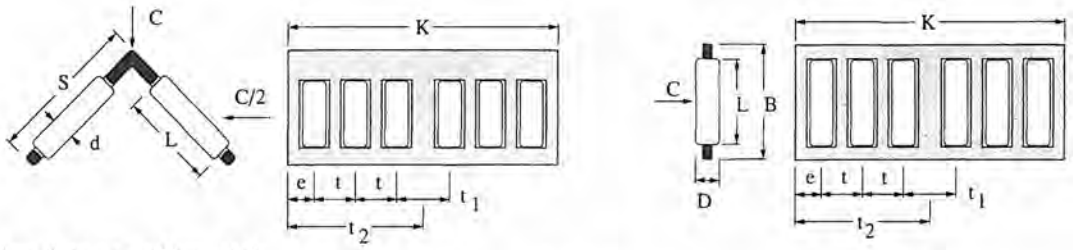
Figure 8.5.15 Quasi-kinematic arrangement of cross roller bearings and rollers on flat rails.

Running Parallelism, Repeatability, and Resolution

Machine ground nonrecirculating rollers on a flat rail can achieve running parallelism of about 5-10 μm/m, repeatability of about $^1/_2$-1 μm, and resolution (depending on the servosystem) on the order of 0.1-1.0 μm. If the rails are hand finished and the rollers individually inspected and sorted, then submicron accuracy and better repeatability can be obtained. Once again, however, one should not rely on hand-finishing operations unless the proven capability exists in-house.

Lateral and Moment Load Support Capability

Each assembly of rollers in a cage can support compressive loads. If the assembly is long enough and the preload sufficient, it can also support moment loads about an axis parallel to the



Rollers (3 to a set of length t_1):

Vee cages:

Size	D	L	S	e	t	t_2	t_1	$C/(2t_1)$
VSF 412	4	6	15	6	6	8	20	11.4
VSF 612	6	5.8	18.5	8	9	11	29	17.2
VSF 624	6	18	29.5	8	9	11	29	42
VSF 1024	10	14	32	13	15	18	48	69.2
VSF 1435	14	20	47	17	21	24	66	137.2

Flat cages:

Size	D	L	B	e	t	t_2	t_1	$C/(2t_1)$
SF 422	4	15.8	22	6	6	8	20	22.8
SF 624	6	18	24	8	9	11	29	29.6
SF 636	6	30	36	8	9	11	29	66.4
SF 1024	10	14	24	13	15	18	48	48.8
SF 1050	10	40	50	13	15	18	48	146.8
SF 1435	14	20	35	17	21	24	66	97.0
SF 1471	14	56	71	17	21	24	66	286.8

Needles (More than 3 to a set of length t_1):

Size	D	L	S	e	t	t_2	t_1	$C/(2t_1)$
HW 15	2	6.8	15	3.5	4.5	4.5	31.5	8.7
HW 20	2.5	9.8	20	4	5.5	5.5	44	17.9
HW 25	3	13.8	25	4.5	6	6	54	33.8
HW 30	3.5	17.8	30	5	7	7	70	55

Size	D	L	B	e	t	t_2	t_1	$C/(2t_1)$
H 10	2	6.8	10	3.5	4.5	4.5	31.5	7.2
H 15	2.5	9.8	15	4	5.5	5.5	44	14.8
H 20	3	13.8	20	4.5	6	6	54	27.9
H 25	3.5	17.8	25	5	7	7	70	45.5

Figure 8.5.16 Typically available modular nonrecirculating roller linear bearings. Units are mm and kN. (Courtesy of Schneeberger Inc.)

rollers' axes. Manufacturers usually supply compressive load data for the bearing assembly or allowable load per roller. If one wishes to determine the moment capability and stiffness of the assembly, a first-order model is to assume that the loads on the rollers vary linearly from a maximum value to zero. This assumes that the stiffness of each roller is the same, which is not true because Hertz theory shows it to be dependent on load, but suffices for a first-order estimate of stiffness about a nominal operating point. Figure 8.5.17 shows this first-order model, which assumes that the carriage is rigid with respect to the rollers and initially all the rollers are perfectly round and the rolling surfaces perfectly flat. As the moment causes the carriage to pitch, it moves as a rigid body through the pitch angle, so the displacement from roller to roller is assumed to vary linearly. The equations representing the system are thus

$$\Sigma F_y = 0 = -F_p - F + \sum_{i=1}^{N} F_i \tag{8.5.4}$$

$$\Sigma M_z = 0 = -F_p \ell_p - F\ell_f + \sum_{i=1}^{N} F_i \, \ell_i \tag{8.5.5}$$

$$\delta_i = \delta_1 - \frac{(\delta_1 - \delta_N)(\ell_i - \ell_1)}{\ell_N - \ell_1} \tag{8.5.6}$$

$$F_i = K_i \delta_i \tag{8.5.7}$$

where for the linear model the roller stiffnesses K_i are constant. If one wanted to obtain a better linear model, one could assume that the stiffnesses were constant, but their values varied from a condition with no preload, to a condition with full preload. Equations 8.5.4-8.5.7 would still be linear and solvable with matrix methods. Figure 8.5.18 shows a similar model for a system where one one set of rollers is preloaded against another.

In order to better estimate the actual pitch stiffness and load capacity of the system, five alternatives can be considered: (1) assume that the K_i's are constant but vary in value from K_1 with no preload to K_N with full preload, then use Equations 8.5.4-8.5.7 to evaluate the forces and then iterate; (2) measure a test system; (3) assume that each roller acts like a spring whose stiffness can be obtained from Hertz theory for rollers where the distance d_o is equal to one roller diameter; (4) build

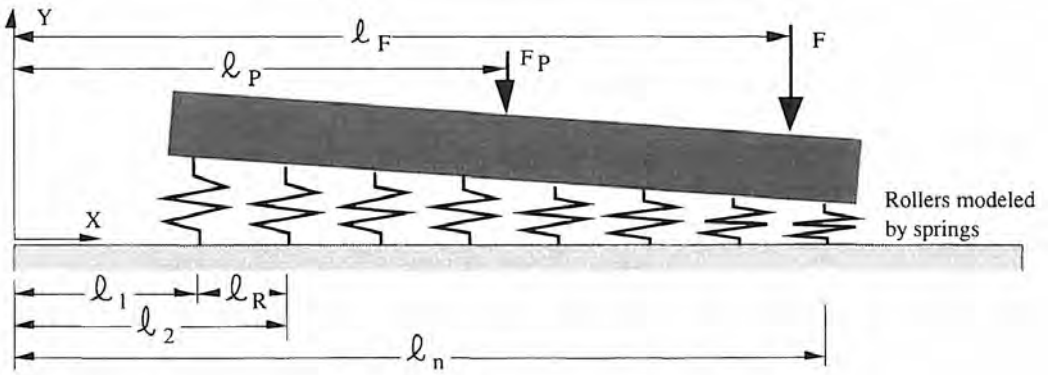


Figure 8.5.17 First-order model of response to an offset load of a rolling element linear bearing-supported slide preloaded by a weight.

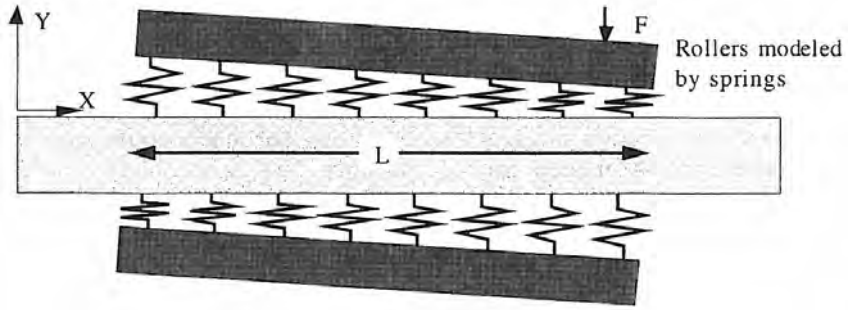


Figure 8.5.18 First-order model of response to an offset load of a rolling element linear-bearing supported slide preloaded by opposed bearings.

a finite element model of the system using the nonlinear spring constants defined by Hertz theory; or (5) build a finite element model that also models each rolling element and the contact zone.

Typically a system is preloaded and the number of rollers is large (i.e., L/roller diameter $>$ 10) so it is fair to assume that the system has a uniform stiffness over its length of $K_{lateral}$, which is defined at the bearing's geometric center. The rotational stiffness of the system about the centroid will thus be

$$K_{rotational} = \frac{K_{lateral} L^2}{12} \tag{8.5.8}$$

Thus when a force F is applied a distance λ from the bearing centroid, the bearing sees a force F and a moment $F\lambda$ which act on the springs $K_{lateral}$ and $K_{rotational}$ to produce lateral and rotational motions. This same reasoning can be used with sliding contact bearings.

Whichever method is used, one still needs to contend with the assumption that the geometry of the rollers and rolling surface was perfect to start. To address this assumption, one needs to consider the equations that describe the Hertz equations for the rollers in greater detail. First of all, one must choose a value for the reference distance d_o. The deflection is actually rather insensitive to this parameter as it increases only by 10% as d_o goes from one to three roller diameters; hence we will assume that d_o equals one roller diameter. For a steel roller of diameter d_r and length[56] l_r between two steel plates and loaded by a force F_r, the displacement of a far-field point in one steel plate with respect to a far-field point in the other steel plate is thus[57]

$$\delta = \frac{4F_{roller}}{\pi \ell_{roller} E_e} \left(2 \log_e \left(\frac{\pi \ell_{roller} E_e}{F_{roller}} \right) - 0.7143 \right) \tag{8.5.9}$$

[56] Note that rollers' ends usually are gradually tapered to minimize the stress concentration at the ends; hence one should use the untapered length in the calculations.

[57] From Equations 5.6.16-5.6.18. It is assumed that $d_o = d_{roller}$.

The compliance C is the partial derivative of Equation 8.5.9 with respect to the force, and the stiffness is the inverse of the compliance:

$$K = \frac{\pi \ell_{roller} E_e}{4\left(2\log_e\left(\frac{\pi \ell_{roller} E_e}{F_{roller}}\right) - 2.7143\right)} \tag{8.5.10}$$

The design engineer must be careful to consider real-world effects when using Equations 8.5.9 and 8.5.10. For example, Tables 8.5.1, 8.5.2, and 8.5.3 show the force, stiffness, and deflection of a single roller between flat steel plates a contact pressure of 200 MPa. This represents about 15% of the maximum contact pressure one would impose on a steel bearing. Note that these tables were easily generated with a spreadsheet. Ask yourself the question: If 100 of these rollers were used, would the preload force be distributed evenly over the rollers? Is the accuracy of the manufacturing process good enough to ensure that some rollers are not effectively preloaded because they are sitting in a trough created by an error in the manufacturing process? Is the level of deflection on the order of the roundness of the rollers? The more rollers that are used, the lower the stress levels and the greater averaging effect of manufacturing errors; however, the greater the chance that all rollers will not be loaded evenly, the more likely the total stiffness of the system will *not* be simply equal to the product of the total number of rollers and the stiffness of a single roller predicted by Hertz theory.

	Roller length (diameters)			
	1	1.5	2	2.5
Roller diameter (mm)	Force (N)			
5	13.8	20.7	27.6	34.5
10	55.3	82.9	110.5	138.2
15	124.3	186.5	248.7	310.8
20	221.0	331.6	442.1	552.6

Table 8.5.1 Force to cause contact pressure of 200 MPa in a single steel roller compressed between two flat steel plates.

	Roller length (diameters)			
	1	1.5	2	2.5
Roller diameter (mm)	Stiffness (N/m)			
5	1.29E+07	1.93E+07	2.58E+07	3.22E+07
10	2.69E+07	4.03E+07	5.37E+07	6.71E+07
15	4.13E+07	6.19E+07	8.26E+07	1.03E+08
20	5.60E+07	8.41E+07	1.12E+08	1.40E+08

Table 8.5.2 Stiffness of a single steel roller compressed between two flat steel plates when the contact pressure is 200 MPa (do/d = 1).

	Roller length (diameters)			
	1	1.5	2	2.5
Roller diameter (mm)	Deflection (μm)			
5	1.13	1.13	1.13	1.13
10	2.18	2.18	2.18	2.18
15	3.20	3.20	3.20	3.20
20	4.19	4.19	4.19	4.19

Table 8.5.3 Deflection of a single steel roller compressed between two flat steel plates when the contact pressure is 200 MPa (do/d = 1).

Once again the real world is at odds with the ideal world. Ideally, one would like to keep the preload force low and use many rollers in order to attain high stiffness yet still have low rolling friction. However, the lower the preload, the more likely that the amount of preload deflection will be less than the accuracy to which the rollers and rolling surfaces are manufactured.

The higher the preload, the better the chances of all rollers being forced into contact with the rolling surfaces and hence the better the accuracy of the theory. One must consider these effects to carefully balance the choice of preload, manufacturing tolerances, and sensor, control, and actuator system design that must move the axis in the presence of rolling friction caused by the preload.

Until an effective mathematical model[58] of this situation is built, or unless one is able to build a physical model, the precision machine design engineer may want to assume that the stiffness is determined based on only a kinematic arrangement of rollers supporting the load. With respect to the maximum load-carrying capability, one can probably assume that all the rollers are in contact. The validity of this assumption is based on the fact that precision machines are usually designed for stiffness as opposed to load-carrying capability, and for a precision machine the first few microinches are the most critical and they also happen to be at the transition from where only a few rollers are properly loaded to where all the rollers are properly loaded. For Equations 8.5.4 and 8.5.7 this means that only the first and the last roller should be used (1 and N); all the rollers in between are there for load capacity insurance purposes. Even with this conservative assumption, a system can easily have a stiffness of several hundred newtons per micron.

Allowance for Thermal Growth

Pseudokinematic arrangements of rollers are more tolerant of thermal growth than are nonkinematic systems. For the latter, there is no way to accommodate thermal growth other than elastic (hopefully) deformation of the components. If the bearings are used in quasi-steady applications, then thermal growth should not be a problem as long as the components are made of the same material. For precision high-speed reciprocating systems, the use of an active cooling system should be considered.

Alignment Requirements

Alignment is a function of the assumptions made in modeling the system, for both error budgeting and stiffness and load calculations. If bearing rails are not parallel, then rollers may not be preloaded properly, and in extreme cases a rocking motion will exist unless the preload is high enough to ensure that roller contact always exists.

Preload and Frictional Properties

Preload can be applied by the weight of the carriage being supported or by an opposed bearing configuration. In the latter, tapered gibs or preload bolts are required, design principles for which were discussed earlier. In most cases, static and dynamic coefficients of friction on the order of 0.01-0.001 can be expected. For system design purposes, one can assume a static coefficient of friction of 0.01 and a dynamic coefficient of friction of 0.005.

8.5.2 Recirculating Balls

Balls can establish their own spacing in recirculating linear motion bearings, so for low speeds spacer balls are not needed unless submicron performance is required. In either case, most bearings have a fixed retainer to prevent the balls from dropping onto the floor should the bearing unit be removed from the rail. There are numerous types of linear bearings that use recirculating balls, and those considered here include wheels on rails, recirculating balls on round or grooved rails, and linear motion guides.

8.5.2.1 Rotary Motion Bearings as Wheels on Rails

Since the accuracy of rotary motion rolling element bearings is so high and they are so inexpensive, one of the simplest ways to get accurate unlimited linear travel is to use a rotary motion bearing as a wheel riding on a rail. The load on the rolling elements behaves like a smooth sine function and thus there is no cogging effect. Balls are often used, but in heavy load applications rollers are used. Various types of bearing wheels are shown in Figure 8.5.19, and they are also often used as cam followers. The chamfered units are often used as a long series to transfer moving loads.

[58] In the manner one shows that random error is reduced by a factor about equal to the square root of the number of samples taken (minus one), an interesting research project would be to devise a model for the stiffness of a system given the required preload deflection, the number of rolling elements, and the accuracy of the system elements.

The chamfer acts to keep the load centered so that an extra set of rollers is not required. The plain rollers are often used as wheels on a carriage that runs on fixed rails. The vee series has slippage associated rolling contact over a range of diameters but provides the simplest geometry for attaining linear motion as only three wheels can be used. Units are shown with integral threaded shafts but are also available with a through hole or the shaft mounted eccentrically in the inner race. As an adjustment nut is turned, the center of the inner race is displaced radially with respect to the shaft, thereby preloading the system. Of course, if a style or accuracy grade is not available in the modular form, this type of roller is easily designed and manufactured.

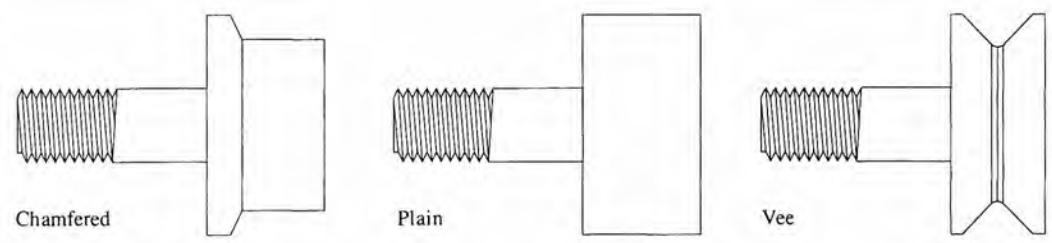


Figure 8.5.19 Various types of cam followers. Often the plain type has a crowned surface.

The most accurate configuration using bearing wheels is a kinematic one shown in Figure 8.5.20, for which the analysis done in Section 2.2.4 can be applied. Other configurations use multiple rollers attached to a moving carriage, a carriage assembly rolling along rails, or a long series of stationary rollers used to transfer loads from station to station. Figure 8.5.21 shows typical off-the-shelf plain-stud-style bearing wheels. Note that if pairs of rollers are anchored to a table via a swivel joint, the table can be made to move following a curved rail.

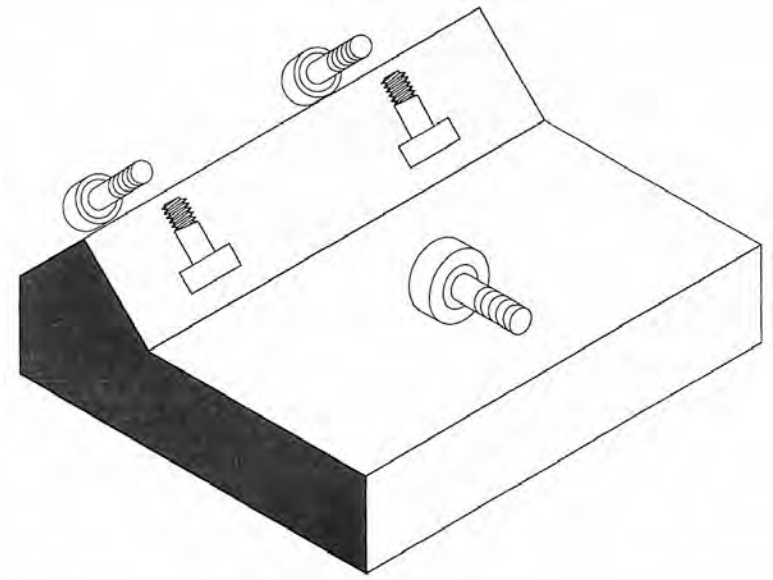


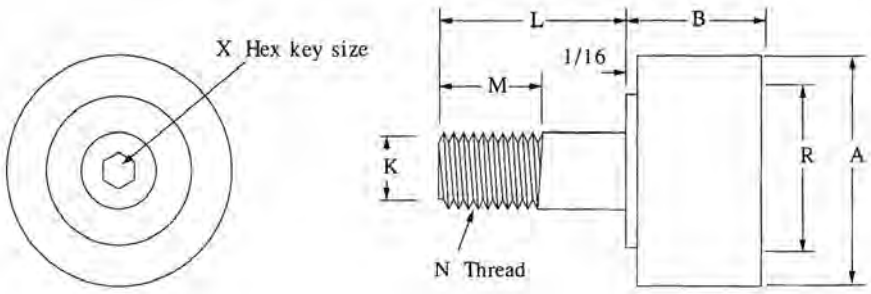
Figure 8.5.20 Kinematic configuration of rollers on a vee and flat. Crowned rollers must be used if slip-noise is to be avoided.

Running Parallelism, Repeatability, and Resolution

For a kinematic arrangement of wheels on rails, running parallelism can be as good as the rails can be made parallel. For properly supported ground systems this will be about 3-5 μm/m. If the assembly is hand scraped or lapped, order-of-magnitude increases can be attained. Note, however, that a single piece of dirt can act like a bump as a roller rolls over it, so cleanliness and wipers are essential for high performance. For fully constrained two-rail systems, one needs to consider alignment effects, which may decrease parallelism to 5-100 μm/m depending on the care

taken during machining and assembly. With great care, nonkinematic configurations can also have submicron repeatability if the preload is constant and all rollers are always in contact.

Kinematic systems are generally preloaded only by the weight of the carriage they support, and the bearings in the wheels themselves usually are only lightly preloaded; hence the rolling coefficient of friction is on the order of 0.01-0.005, and with the proper actuator/sensor/control system, submicron and better resolution of motion is possible. Nonkinematic configurations tend to have higher coefficients of friction and are more difficult to control.



Part #	A	B	K	L	M	N	R	X	L_{10} radial load
PLR-1 1/2	1.500	1.1875	0.625	1.500	0.750	5/8-18	0.750	5/16	1050
PLR-1 3/4	1.750	1.1875	0.750	1.750	0.875	3/4-16	1.000	5/16	1050
PLR-2	2.000	1.6875	0.875	2.000	1.125	7/8-14	1.000	5/16	1450
PLR-2 1/4	2.250	1.6875	0.875	2.000	1.125	7/8-14	1.000	5/16	1450
PLR-2 1/2	2.500	1.6875	1.000	2.250	1.500	1-14	1.250	1/2	1980
PLR-2 3/4	2.750	1.6875	1.000	2.250	1.500	1-14	1.250	1/2	1980
PLR-3	3.000	2.000	1.250	2.500	1.750	1 1/4-12	1.750	1/2	6000
PLR-3 1/4	3.250	2.000	1.250	2.500	1.750	1 1/4-12	1.750	1/2	6000
PLR-3 1/2	3.500	2.000	1.250	2.750	1.750	1 1/4-12	1.750	1/2	6000
PLR-4	4.000	2.000	1.250	2.750	1.750	1 1/4-12	1.750	1/2	6000
PLR-4 1/2	4.500	2.000	1.250	2.750	1.750	1 1/4-12	1.750	1/2	6000
PLR-5	5.000	3.000	2.000	4.500	2.500	2-12	3.250	5/8	15100

All dimensions are in inches. L10 life is in lbf for 3000 hours at 100 rpm. Roller diam. +0.000 - 0.001

Figure 8.5.21 Typical sizes of plain-stud-style cam followers. Sizes up to 10 in. roller diameter are available. (Courtesy of Osborn Manufacturing/Jason Incorporated)

Lateral and Moment Load Support Capability

Rotary motion bearings are usually designed so their races receive support from the housing and shaft; thus one cannot simply use the load ratings for rotary motion bearings in this type of application. Cam followers are intended for this type of application and load ratings are given by the manufacturers, which provides incentive for using them as opposed to a regular rotary motion bearing.

Allowance for Thermal Growth

Kinematic systems' accuracy and repeatability are affected by thermal growth, but they do not tend to lock up or become loose the way nonkinematic systems can; therefore, kinematic systems do not have to worry about lack of allowances for thermal growth affecting dynamic response. Nonkinematic systems must rely on elastic deformation of the components or careful design to balance thermal growth.[59] If the bearings are used in quasi-steady applications, then thermal growth should not be a problem as long as the carriage and base the rails are mounted to are made of the same material. For high-speed precision reciprocating systems, an active cooling system should be considered.

[59] See Section 8.9.

Alignment Requirements

Kinematic configurations have no alignment requirements other than the outer race of the cam followers must ride flat on the rails. If the wheels do not ride flat, the system will still be kinematic, but enhanced wear on the rolling edge will result. To ensure proper contact, the outer race can be ground so as to have a spherical profile. Nonkinematic systems must be aligned so that all rollers' preloads stay within desirable limits, and desired accuracy of motion is achieved.

Preload and Frictional Properties

Kinematic systems are preloaded by the weight of the carriage being supported, and nonkinematic systems are typically preloaded by an opposed bearing configuration. If a friction drive is used, the capstan roller can be used to preload the kinematic arrangement of rollers. As mentioned above, eccentric shafts are often used in the latter. In most cases, static and dynamic coefficients of friction on the order of 0.01-0.005 can be expected. For system design purposes, one should consider that the static and dynamic coefficients of friction will have the worst possible values and thus the dynamic model should try different permutations to find out which is worst. There are times when pessimism pays off.

8.5.2.2 Recirculating Balls on Round Shafts

Recirculating balls that ride on round shafts individually have only radial stiffness and allow for rotational alignment but not continuous rotational motion. Figure 8.5.22 shows the basic elements of a linear bearing which incorporates recirculating balls on a round shaft. This type of bearing is typically available in the form of single units that are pressed into a bore in the machine structure or they are available already mounted in aluminum or steel pillow blocks or flange blocks. Self-aligning types must be used in pairs on a shaft in order to support moment loads. Twin pillow blocks can support radial and moment loads.

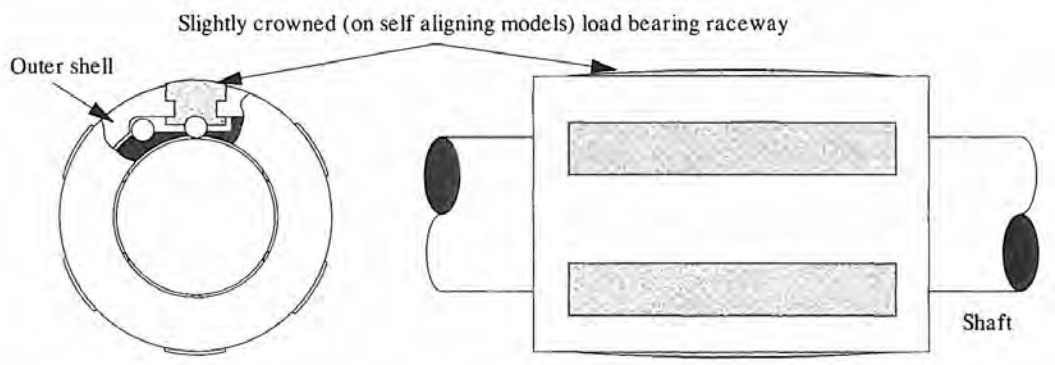


Figure 8.5.22 The basic element of a linear bearing which incorporates recirculating balls on a round shaft (e.g., a Ball Bushing®. Both open and closed styles are typically available. (Courtesy of Thomson Industries.)

Shafts can be simply supported and used with closed Ball Bushings® or shafts can be supported along their lengths and used with open-type Ball Bushings®, as shown in Figure 8.5.23. The former are used primarily where short strokes are needed or where the carriage serves as a guide for reciprocating motion and straightness of travel is not of primary concern. The latter allows heavy loads to be supported along the entire range of travel, but stiffness in a direction outward from the shaft is somewhat less than inward toward the shaft. Off-the-shelf shafts are available in many different diameters (metric and inch) and are cut to lengths up to 5 m long. Shaft support rails can be purchased from stock or machined into the structure the shafts are attached to.

Running Parallelism, Repeatability, and Resolution

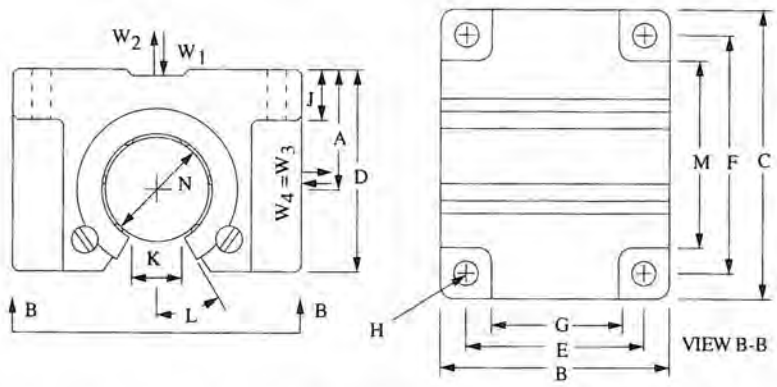
This type of Ball Bushings® is typically used on pairs of shafts because a single unit is not meant to support a torsional load; hence the running parallelism depends primarily on the quality of the shafting used and how well it is aligned. Typical machine-finished shaft straightness of off-the-shelf shafting is on the order of 100 μm/m and can easily be aligned to this value. Order-of-

magnitude increases in shaft straightness are available, although at an increased cost. Preloaded Ball Bushings® for general use can have micron repeatability, while instrument grades (available for up to 6 mm shafts) running on lapped shafts can have microinch repeatability. With the proper servosystem, standard units are used in machines with $^1/_2$ to 1 μm resolution while instrument grades can allow for microinch resolution of motion.

Lateral and Moment load Support Capability

Single or self-aligning Ball Bushings® can only support radial loads. A machine tool carriage will typically be supported by three or four self-aligning Ball Bushings® at two corners of a carriage and one as an outrigger, or at four corners of a carriage respectively. The use of three bearing units would make the assembly quasi-kinematic, so vertical parallelism of the shafts is not as critical, although lack of horizontal parallelism can still cause the carriage motion to bind. The self-aligning feature refers only to angular misalignment; if the shafts are divergent, then the open-type bushings will be spread apart as the carriage moves along the diverging rails.

When heavy loads and moments are applied to a machine tool table from many directions and locations, the carriage must be supported at each corner. If very heavy loads are encountered, then more than four bearing units may be required. Once one crosses over from quasi-kinematic to overconstrained, it is generally acceptable to add more bearing units to build up load-carrying capability. Beware that static and dynamic friction will generally increase with an increase in overconstraint, as more bearing mounts must elastically deform to accommodate relative error motions. Ball Bushing® pillow blocks and shaft support rails for heavy-duty machine tool applications are shown in Figures 8.5.23 and 8.5.24, respectively. Load deflection curves for these bearings are shown in Figure 8.5.25. Ball Bushings® pillow blocks for light-duty applications (e.g., some materials handling equipment) are shown in Figure 8.5.26.



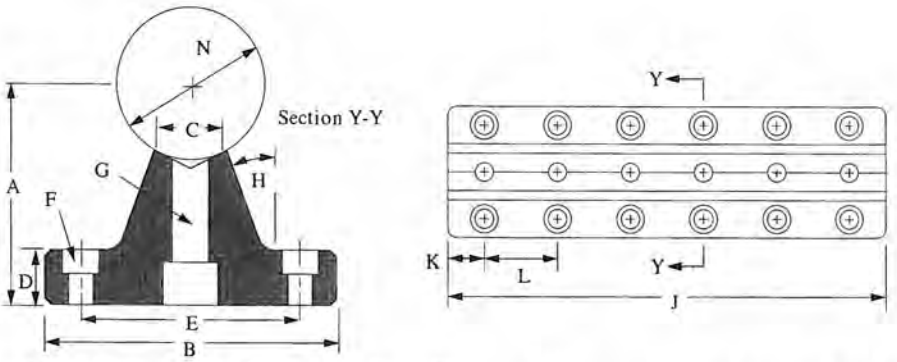
#*	A**	B	C	D	E	F	G	H_{bolt}	H_{hole}	J	K	L	M	N	Wt. (lbf)	load (lbf)+ W_1	W_2	W_3
32	2.375	6	$4^7/_8$	$3^7/_8$	5.000	3.750	$3^3/_4$	$^1/_2$	$^{17}/_{32}$	$^7/_8$	1	27^o	$2^5/_8$	2	27	4500	2100	2500
48	3.500	$8^3/_8$	$7^1/_4$	$5^7/_8$	7.000	5.875	$5^1/_2$	$^5/_8$	$^{21}/_{32}$	$1^1/_4$	$1^1/_2$	30^o	$4^{51}/_8$	3	55	10000	4500	5500

* Pillow block number given by XPBO- # -OPN. All dimensions are in inches.
**+0.000-0.001. Centerline of shaft must be parallel to base to within 0.0005.
+ Rolling load capacity based on a life of 2 million inches travel. and shaft hardness of Rockwell 60C.

Figure 8.5.23 Ball Bushing® pillow blocks for heavy duty machine tool applications. (Courtesy of Thomson Industries.)

Allowance for Thermal Growth

As with all overconstrained systems, allowance for thermal growth must be designed into the system. Fortunately, rolling element bearings have low coefficients of friction, so heat buildup is gradual and usually gives ample time for the structure to grow uniformly. Note that several manufacturers sell preassembled units that have aluminum bases and carriages with steel rails and pillow blocks. As the temperature changes, the steel rails on the aluminum base will act as a bimetallic strip and one can expect bowing to occur. Be careful to evaluate the possible magnitude of this error before specifying this type of preassembled component. In many applications the design will be adequate and the light weight will be desirable.



Rail #	N	A*	B	C	D	E	F**	G***	H	J	K	L	Wt. (lbf)
XSR-32	2	2.375	$4\frac{1}{2}$	$\frac{7}{8}$	1	$3\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{2}$-13x2	27^{o}	$23\frac{13}{16}$	$1\frac{31}{32}$	4	16
XSR-48	3	4.000	6	$1\frac{1}{4}$	$\frac{15}{16}$	$4\frac{1}{4}$	$\frac{5}{8}$	$\frac{3}{4}$-10x2$\frac{3}{4}$	30^{o}	$23\frac{13}{16}$	$1\frac{31}{32}$	6	31

*+0.000-0.001. Centerline of shaft must be parallel to base to within 0.0005.
**Drilled and counterbored for cap screw. ***Recommended cap screw size.

Figure 8.5.24 Cast ductile iron shaft supports for large rails used in heavy-duty machine tool applications. (Courtesy of Thomson Industries.)

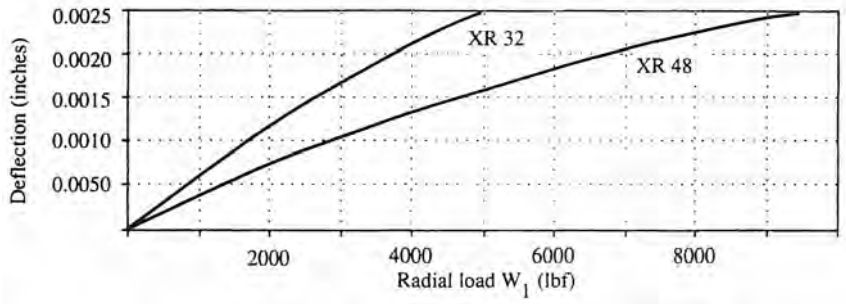


Figure 8.5.25 Vertical load-deflection curves for Ball Bushings®. (Courtesy of Thomson Industries.)

Alignment Requirements

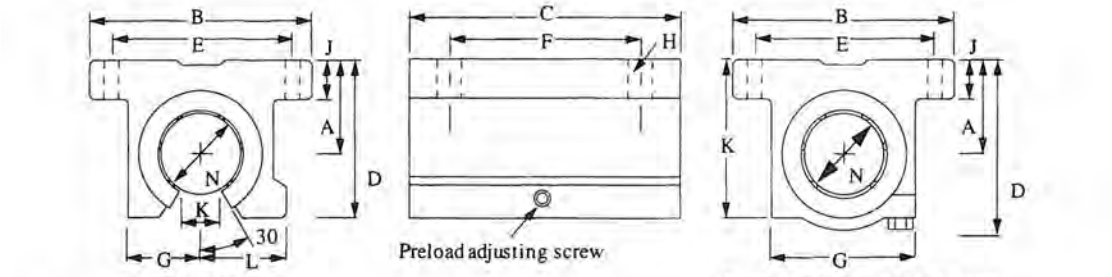
One rail is located with respect to a machine reference and then bolted down securely according to a procedure provided by the manufacturer. The second rail is made to be parallel to the first through the use of spacer blocks (e.g., gage blocks) and then the second rail is bolted in place. Measurements are made along the lengths of the two rails to make sure that alignment is better than the desired running parallelism. The pillow blocks are then placed on the rails and the carriage bolted to them. The assembly is tested by measuring the force required to sustain slow steady motion and noting any deviations in this force along the length of travel.

Preload and Frictional Properties

Ball Bushings® can be preloaded through the use of oversize balls in both open and closed types, or through the use of a clamp with some open types. As with most rolling element bearings, static and dynamic coefficients of friction can be anywhere in the range of 0.01-0.001 depending on load, preload, alignment accuracy, and lubrication.

8.5.2.3 Recirculating Balls on Grooved Shafts

To increase load capacity and allow for torque transmission, recirculating balls can ride on a grooved shaft. This type of design is typically referred to as a *ball spline*. The construction of a ball spline is shown in Figure 8.5.27. Ball splines are used in applications where only one bearing rail is desired, such as in shuttle tables within a machine, or where torque must be translated to a rotating and translating shaft, such as in a quill. As illustrated by Equations 8.5.1 and 8.5.2, a ball in a groove



Open type

Closed (adjustable preload) type

Open type:

#*	N	A**	B	C	D	E***	F***	G	H	J	K	L	F_{radial} (lbf)
8	1/2	0.687	2	1 1/2	1 1/8	1.688	1.000	11/16	5/32	1/4	5/16	3/4	180
10	5/8	0.875	2 1/2	1 3/4	1 7/16	2.125	1.125	7/8	3/16	9/32	3/8	15/16	320
12	3/4	0.937	2 3/4	1 7/8	1 9/16	2.375	1.250	15/16	3/16	5/16	7/16	1	470
16	1	1.187	3 1/4	2 5/8	2	2.875	1.750	1 3/16	7/32	3/8	9/16	1 1/4	780
20	1 1/4	1.500	4	3 3/8	2 9/16	3.500	2.000	1 1/2	7/32	7/16	5/8	1 5/8	1170
24	1 1/2	1.750	4 3/4	3 3/4	2 15/16	4.125	2.500	1 3/4	9/32	1/2	3/4	1 7/8	1560
32	2	2.125	6	4 3/4	3 5/8	5.250	3.250	2 1/4	11/32	5/8	1	2 7/16	2350

* Pillow block number given by SPB- # -OPN. **±0.003. ***±0.010. All dimensions are in inches.

Closed (adjustable preload) type:

#*	N	A**	B	C	D	E***	F***	G	H	J	K	F_{radial} (lbf)+
4	1/4	0.437	1 5/8	1 3/16	1 3/16	1.312	0.750	1	5/32	3/16	3/4	42
6	3/8	0.500	1 3/4	1 5/16	1 5/16	1.437	0.875	1 1/8	5/32	3/16	7/8	70
8	1/2	0.687	2	1 11/16	1 1/4	1.688	1.000	1 3/8	5/32	1/4	1 1/8	180
10	5/8	0.875	2 1/2	1 15/16	1 5/8	2.125	1.125	1 3/4	3/16	9/32	1 7/16	320
12	3/4	0.937	2 3/4	2 1/16	1 3/4	2.375	1.250	1 7/8	3/16	5/16	1 9/16	470
16	1	1.187	3 1/4	2 13/16	2 3/16	2.875	1.750	2 3/8	7/32	3/8	1 15/16	780
20	1 1/4	1.500	4	3 5/8	2 13/16	3.500	2.000	3	7/32	7/16	2 1/2	1170
24	1 1/2	1.750	4 3/4	4	3 1/4	4.125	2.500	3 1/2	9/32	1/2	2 7/8	1560
32	2	2.125	6	5	4 1/16	5.250	3.250	4 1/2	11/32	5/8	3 5/8	2350

* Pillow block number given by SPB- # -ADJ. **±0.003. ***±0.010. All dimensions are in inches.

+ Rolling load capacity based on a life of 2 million inches travel. and shaft hardness of Rockwell 60C.

Figure 8.5.26 Typically available self-aligning single-pillow blocks. Twin-pillow blocks are also available. (Courtesy of Thomson Industries.)

can support considerably more load than a ball on a shaft, and hence the load capacity of a single ball spline is much higher than that of a single bearing where the balls ride on a round shaft.

Running Parallelism, Repeatability, and Resolution

The running parallelism of a ball spline carriage with respect to its rail is shown in Figure 8.5.28. With the proper servo system, standard units can readily allow for $^1/_2$ to 1-μm motion resolution.

Lateral and Moment Load Support Capability

Ball spline nuts transmit torque and only permit axial motion. Sometimes two nuts are used on a rail to support a small table. Rarely are four nuts on two rails used to support a large table because it would be more efficient to use a linear motion guide. Ball spline nuts come in a large variety of exterior configurations, including: (1) plain cylindrical nut with a centrally located keyway and circumferential lubrication groove, (2) cylindrical nut with a flange at one end or in the middle and a centrally located keyway and circumferential lubrication groove, (3) cylindrical nut with an integral gear in the middle, and (4) rectangular nut which a flat surface can be bolted to. Figure 8.5.29 illustrates typical sizes and load capacities of ball splines.

Allowance for Thermal Growth

For single bearing rail systems, although the bushings are actually overconstrained, their small baseline makes them somewhat insensitive to thermal effects. Care must be taken, however,

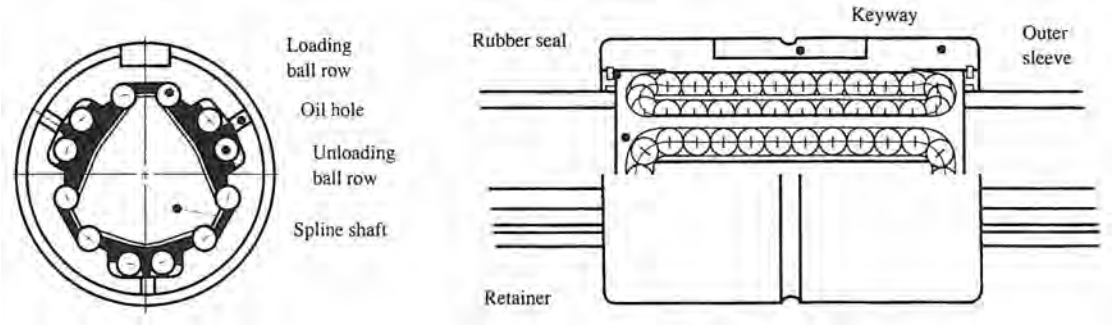


Figure 8.5.27 Construction of a ball spline for supporting radial and torsional loads. (Courtesy of THK Co., LTD.)

Shaft (mm)		Nominal shaft diameter (mm) 15, 20			25, 30			40, 50			60, 70			85, 100, 120			150		
Over	Up to	N	H	P	N	H	P	N	H	P	N	H	P	N	H	P	N	H	P
-	200	56	34	18	53	32	18	53	32	16	51	30	16	51	30	16	-	-	-
200	315	71	45	25	58	39	21	58	36	19	55	34	17	53	32	17	-	-	-
315	400	83	53	31	70	44	25	63	39	21	58	36	19	55	34	17	-	-	-
400	500	95	62	38	78	50	29	68	43	24	61	38	21	57	35	19	46	36	19
500	630	112	-	-	88	57	34	74	47	27	65	41	23	60	37	20	49	39	21
630	800	-	-	-	103	68	42	84	54	32	71	45	26	64	40	22	53	43	24
800	1000	-	-	-	124	83	-	97	63	38	79	51	30	69	43	24	58	48	27
1000	1250	-	-	-	-	-	-	114	76	47	90	59	35	76	48	28	63	55	32
1250	1600	-	-	-	-	-	-	139	93	-	106	70	43	86	55	33	80	65	40
1600	2000	-	-	-	-	-	-	-	-	-	128	86	54	99	65	40	100	80	50
2000	2500	-	-	-	-	-	-	-	-	-	156	-	-	117	78	49	125	100	68
2500	3000	-	-	-	-	-	-	-	-	-	-	-	-	143	96	61	150	129	84

Figure 8.5.28 Running parallelism (microns) of a ball spline nut with respect to its rail. (Courtesy of THK Co., LTD.)

to allow one end of the bearing rail to axially float, or else it may bow as it heats up, due to environmental changes and rolling friction.

Alignment Requirements

Since only one rail is generally used with this type of bushing, one must only align the rail to the rest of the machine. When mounting two bushings to a carriage, one must take care to ensure that the mounting surfaces are aligned and properly located.

Preload and Frictional Properties

Ball spines are preloaded by using oversize balls. As with most rolling element bearings, static and dynamic coefficients of friction can be anywhere in the range of 0.01-0.005, depending on the load, preload, accuracy grade, and lubrication used.

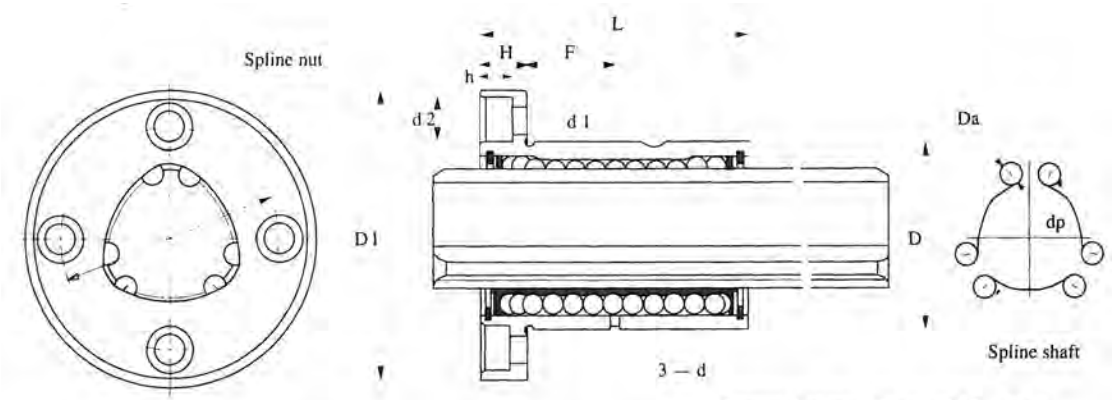
8.5.2.4 Linear Motion Guides

Linear motion guides are characterized by an essentially rectangular cross section rail and rectangular box-shaped carriages that contain passages for recirculating balls.[60] This type of design was originally patented in France in 1932,[61] and since then there have been many improvement patents, and many manufacturers of linear motion guides currently exist.

Typically, two rails and four carriages (blocks) are used to support an axis, as illustrated in Figure 8.5.30. Linear guides are replacing sliding contact linear bearings in larger and larger machines as confidence in their performance under heavy loads increases and manufacturers of linear motion guides provide larger and larger bearings. Rolling element bearings are used in large spindles and ballscrews and thus it also makes sense to use them to support moving axes. When

[60] Various manufacturers now also makes linear motion guides that uses short rollers instead of balls to yield increased load capacity and stiffness; however, cylindrical rolling element linear motion guides generally do not have as good accuracy as ball-type linear motion guides.

[61] French Patent 730,922.



Size*	dp	Da	D	L	D1	H	F	d	P.C.D	d1	d2	h	Torque (kgf-m) Stat. C_t	Torque (kgf-m) Dyn. C_{ot}	F_{radial} (kgf) Stat. C	F_{radial} (kgf) Dyn. C_o	M_A* (N-m)
15	15	2.381	23	40	43	7	13	2	32	4.5	8	4.4	2.7	4.5	3.9	5.1	22.4
20	20	3.175	30	50	49	7	18	3	38	4.5	8	4.4	6.6	9.6	7.0	9.0	26.6
25	25	3.969	37	60	60	9	21	3	47	5.5	9.5	5.4	13.7	18.8	11.6	14.4	47.4
30	30	4.763	45	70	70	10	25	3	54	6.6	11	6.5	24.3	32.4	17.2	20.4	78.0
40	40	6.350	57	90	90	14	31	4	70	9	14	8.6	53.4	68.4	28.4	32.2	177.5
50	50	7.938	70	100	108	16	34	4	86	11	17.5	11.0	97.6	117.0	41.5	44.0	226.1
60	60	9.525	85	127	124	18	45.5	4	102	11	17.5	11.0	166.5	230.9	59.0	72.6	534.0
70	70	11.112	95	110	142	20	35	4	117	14	20	13.0	194.4	229.3	59.1	61.6	446.7
85	85	11.906	115	140	168	22	48	5	138	16	23	15.2	322.3	383.4	80.6	85.0	767.8
100	100	14.288	135	160	195	25	55	5	162	18	26	17.5	526.8	760.0	112.0	143.0	1035.0

*Model number without seal LBF- #. Model number with seal LBF- #UU
**Moment about a radius of the shaft, for two spline nuts in close contact. A single nut has only 10%M_A.

Figure 8.5.29 Typical sizes of flanged ball splines with grooved rails. (Courtesy of THK Co., LTD.)

linear guides cannot provide the desired load and stiffness capacity in the allotted space, recirculating roller bearings can be used.

There are a number of different manufacturers of linear motion guides, but there are basically two different types of linear motion guides (linear guides): those with circular arc grooves and those with Gothic arch grooves. These two types of grooves were illustrated side by side in Figure 8.5.2, and their relative performance characteristics were compared in Figures 8.5.3 and 8.5.4. There is also a design variation on the Gothic arch groove whereby the grooves are offset so that the balls nominally contact only one surface of each arch, and contact with the other arch surface is made only when the bearing is overloaded. Hence the system is somewhat self-protecting.

Each type of linear motion guide has its advantages and disadvantages. In order to be objective, they will be discussed in alphabetical order by type, and representative manufacturers' catalog information will also be presented in alphabetical order. As with other such examples of available products, this does not represent a product endorsement. A design engineer must carefully consider many criteria, and often tests must be made with different products before a selection decision can be made. Remember, if a manufacturers' catalog does not have the information you need, do not hesitate to call the manufacturer and ask for the information. How well a sales engineer responds to your questions is often a good indication of the type of service you will receive at a later date.

Running Parallelism, Repeatability, and Resolution

Running parallelism of the different types of linear guides made by most different manufacturers are generally competitively equivalent as far as published specifications are concerned. Typical running parallelism values are shown in Figure 8.5.31. Of course when selecting a line of bearings for a major production run, one would want to spot check various manufacturers' bearings. When selecting a linear motion guide for a particular type of application (e.g., for a measuring machine or a heavy metal cutting machining center), it is helpful to ask the bearing manufacturer to provide data on running parallelism as a function of life.

Linear motion guide rails are often ground when bolted to a grinding jig. Therefore, in order to ensure that the rails deform to the same shape at which they were when the grooves were ground, it is imperative to specify that the tapped mounting holes be cleaned out and the manufacturers'

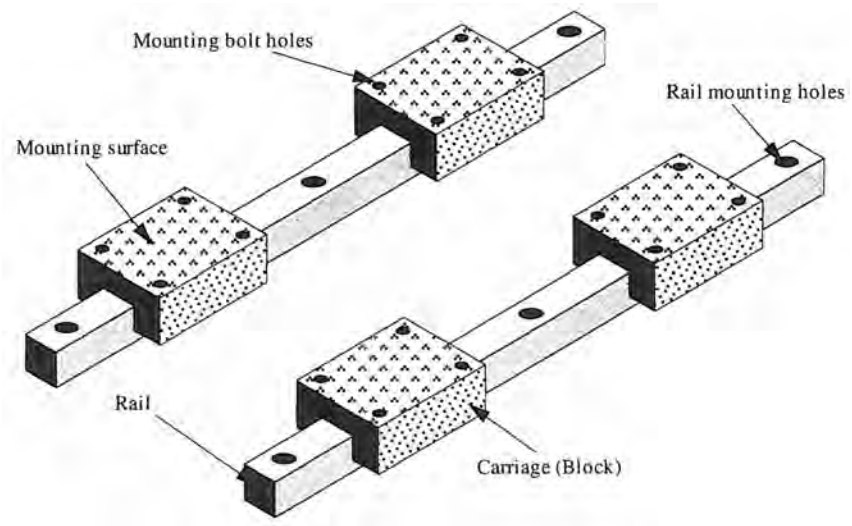


Figure 8.5.30 Basic components of a linear motion guide bearing system.

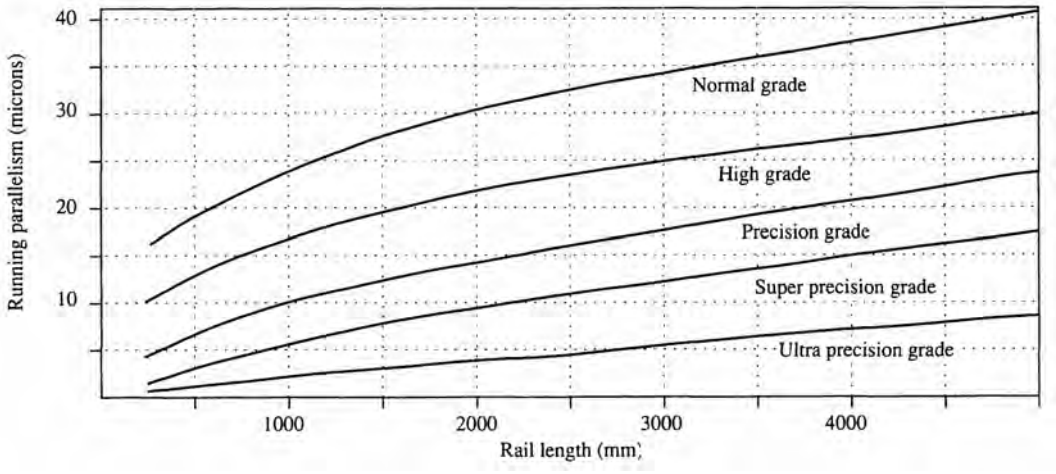


Figure 8.5.31 Typical running parallelism of upper and side surfaces of a linear guide bearing carriage with respect to upper and side surfaces of the bearing rail. (Courtesy of THK Co., LTD.)

suggested torque levels and tightening procedure be carefully followed. After tightening the bolts, the rail straightness should be checked. If necessary, the bolt torquing procedure may have to be modified.

Repeatability depends on the accuracy grade and can be anywhere from 0.l to 10 μm. A light preload is usually desired for high repeatability. Similarly, allowable resolution given the proper servo system can be from 0.1 to 10 μm. As was shown in Figure 8.3.6, in order to maximize attainable resolution of motion, friction should be minimized.

Lateral and Moment Load Support Capability

Circular arc grooves contact balls at two points; hence in order to have bidirectional load capacity, four grooves per rail are required. The ball-groove contact vectors can be arranged in face-to-face or back-to-back configurations. Examples of these two types are illustrated in Figures 8.5.32-8.5.35. Back-to-back designs have a higher moment capacity than do face-to-face designs. High moment capacity is useful for single-rail applications. For machine tools, moment capacity is obtained by spacing rails far apart and relying on the bearings' normal load capacity; thus high

individual bearing carriage moment capacity can actually make a multiple bearing system more difficult to align and assemble without providing a much greater moment capacity for the system.

Gothic arch grooves that contact balls at four points can have a large amount of differential slip, as was shown in Figure 8.5.4; hence often the arches are offset so that the balls nominally make contact at two points, and thus achieve rolling motion with lower friction. In the event of an overload, the deflection of the balls in one of the grooves causes the balls to make four-point contact with that groove. This helps to prevent damage in case of overload. Note that when subjected primarily to side loads, four point contact is made. This increases friction due to differential slip, but also provides a measure of overload protection. Typical sizes and properties of one type of Gothic arch linear guide are shown in Figures 8.5.36 and 8.5.37.

It is not possible to generalize and say that one design type can carry more load or is stiffer than the other. One can always select a unit from a design type that has the load capacity and stiffness that are desired. Design engineers are therefore strongly recommended to consider more important factors such as accuracy as a function of life, resolution of motion, and many other factors that are discussed in Section 8.5 in the subsection *Selection Criteria*. Regardless of the type of linear bearing selected, it is also important to be careful if it will be subject to very slow motion and extended periods of vibration, which could lead to fretting. Manufacturer's catalogs usually scale down the allowable loads under these circumstances. Note that some manufacturers will provide stainless steel bearing rails or ceramic rolling elements for cases were fretting may be a problem.

A typical machine tool carriage has a rectangular footprint and has one linear guide bearing carriage mounted near each corner. For purposes of accurately analyzing the load-carrying capacity and stiffness, a finite element model is needed that includes characteristics of the bed and carriage; however, before one can build a finite element model, one must use back-of-the-envelope calculations to preliminarily size members and choose bearings. The structure of the carriage must be designed so that its deformations are within those allotted by the error budget; thus for the purposes of finding approximate bearing reaction forces, it will be assumed that the carriage and structure the rails are mounted to behave like a rigid body.

Consider the general case in Figure 8.5.38. The bearings are symmetrically located about a Cartesian coordinate system[62]. There are three types of forces, F_X, F_Y, and F_Z, which can act anywhere in space with respect to the carriage coordinate system. Forces of the F_X type can include cutting forces, actuator forces, center-of-mass acceleration forces, and F_X forces from axes stacked on top of the carriage. Forces of the F_Y and F_Z type can include cutting forces and forces from other axes stacked on top of the carriage. Gravity can act in any direction, depending on the machine configuration. In order to estimate the magnitudes of the resultant forces on the bearings, two assumptions must be made:

Moment stiffness of the bearings is insignificant.

Forces are distributed in relation to the bearings' proximity to them.

With these assumptions, careful scrutiny of the geometry represented by Figure 8.5.38, and some algebra one can find the effect of the generic forces on each of the four bearing carriages. With superposition, the net forces on the bearing carriages caused by a number of generic forces applied at different points can be determined.

X direction forces cause Y direction forces in the bearing carriages in the form of couples whose magnitude is assumed to be independent of the Z or X position of the X direction force; hence the total force couple is evenly distributed between bearings at respective ends:

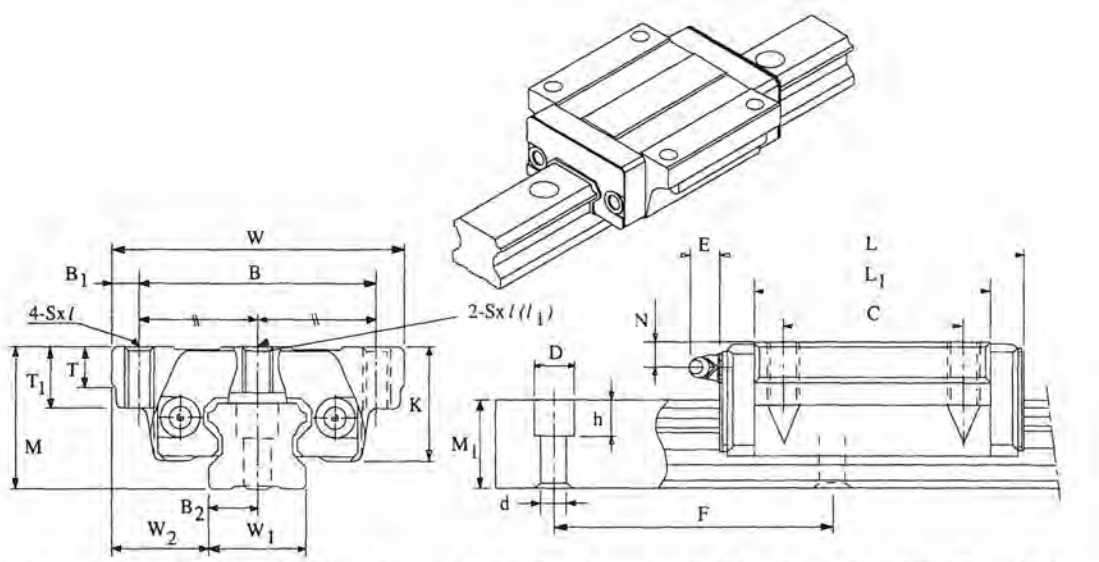
$$F_{1Y,FX} = F_{2Y,FX} = \frac{F_X\, y_{FX}}{2\,(x_1 - x_4)} \tag{8.5.11}$$

$$F_{3Y,FX} = F_{4Y,FX} = \frac{-F_X\, y_{FX}}{2\,(x_1 - x_4)} \tag{8.5.12}$$

X direction forces also cause Z direction forces in the form of a couple which is assumed to be unaffected by the Y or X position of the X direction force and therefore is evenly distributed between bearings at respective ends.

$$F_{1Z,FX} = F_{2Z,FX} = \frac{F_X\, z_{FX}}{2\,(x_1 - x_4)} \tag{8.5.13}$$

[62] See Equation 2.2.29 for a discussion on the center of stiffness of a bearing system.



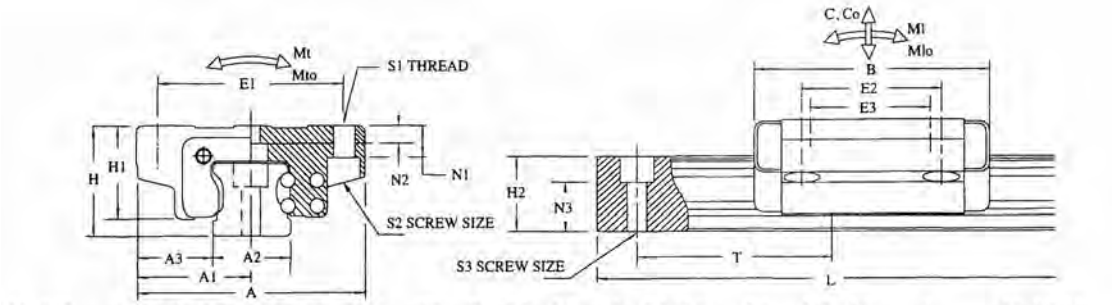
Model	W	B	B_1	L	C	M	Sxl	A	T	T_1	K	L_1	E	W_1	W_2	B_2	M_1	F	d,D,H
15TA	47	38	4.5	53.5	30	24	M5x11	12.2	7	11	19.4	40.5	5.5	15	16	7.5	15	60	4.5,7.5,5.3
20TA	63	53	5	70	40	30	M6x10	14.5	10	10	25	50	12	20	21.5	10	18	60	6,9.5,8.5
20HTA	63	53	5	86	40	30	M6x10	14.5	10	10	25	66	12	20	21.5	10	18	60	6,9.5,8.5
25TA	70	57	6.5	79	45	36	M8x16	18	10	16	29.5	59	12	23	23.5	11.5	22	60	7,11,9
25HTA	70	57	6.5	103	45	36	M8x16	18	10	16	29.5	83	12	23	23.5	11.5	22	60	7,11,9
30TA	90	72	9	94	52	42	M10x18	21	10	18	35	72	12	28	31	14	26	80	9,14,12
30HTA	90	72	9	116	52	42	M10x18	21	10	18	35	94	12	28	31	14	26	80	9,14,12
35TA	100	82	9	105	62	48	M10x21	24	13	21	40	81.3	12	34	33	17	29	80	9,14,12
35HTA	100	82	9	134	62	48	M10x21	24	13	21	40	110	12	34	33	17	29	80	9,14,12
45TAX	120	100	10	139	80	60	M12x15	30	14	25	50	98	16	45	37.5	22.5	38	105	14,20,17
45HTA	120	100	10	171	80	60	M12x15	30	14	25	50	130	16	45	37.5	22.5	38	105	14,20,17
55TAX	140	116	12	163	95	70	M14x17	36	15	29	57	118	16	53	43.5	26.5	44	120	16,23,20
55HTA	140	116	12	201	95	70	M14x17	36	15	29	57	156	16	53	43.5	26.5	44	120	16,23,20
65TAX	170	142	14	186	110	90	M16x23	43	23	37	76	147	16	63	53.5	31.5	53	150	18,26,22
65HTA	170	142	14	246	110	90	M16x23	43	23	37	76	207	16	63	53.5	31.5	53	150	18,26,22
85TA	215	185	15	247	140	110	M20x30	51	30	55	94	179	16	85	65	42.5	65	180	24,35,28
85HTA	215	185	15	303	140	110	M20x30	51	30	55	94	236	16	85	65	42.5	65	180	24,35,28

*Prefix by HSR. All dimensions are in mm.

Figure 8.5.32 Face-to-face circular arch linear guides. (Courtesy of THK Co., LTD.)

	Stiffness (K_Y, K_Z) (N/μm)	Load capacity (kgf)($F_Y = K_Z$)		Static moment capacity (kgf-m)		
Model	Medium preload	Dyn. C	Static C	Static M_X	Static M_X	Static M_X
HSR 15TA		760	1150	6.0	6.0	8.4
HSR 20TA	490	1230	1790	11.7	11.7	17.4
HSR 20HTA	686	1900	2380	20.2	20.2	23.2
HSR 25TA	647	1770	2580	20.2	20.2	29.4
HSR 25HTA	872	2420	3440	34.4	34.4	38.1
HSR 30TA	833	2500	3510	32.2	32.2	48.4
HSR 30HTA	1117	3320	4680	54.7	54.7	64.5
HSR 35TA	960	3320	4580	48.1	48.1	77.0
HSR 35HTA	1284	4470	6110	81.7	81.7	102.9
HSR 45TAX	1215	5350	7170	93.8	93.8	156.8
HSR 45HTA	1627	7170	9550	159.6	159.6	208.9
HSR 55TAX	1470	7890	10300	162.2	162.2	272.3
HSR 55 HTA	1960	10600	13800	275.5	275.5	363.7
HSR 65TAX	1842	12600	16100	316.7	316.7	497.6
HSR 65HTA	2479	17100	21500	538.4	538.4	664.5
HSR 85TA	2244	18700	23200	762.1	762.1	942.8
HSR 85HTA	2999	25200	30900	930.6	930.6	1255.0

Figure 8.5.33 Stiffness and load capacity of linear guides of Figure 8.5.32. (Courtesy of THK Co., LTD.)



Model*	A	A1	A2	A3	H	H1	H2	B	E1	E2	E3	E3 (Alt)	N1	N2	N3	T	S1, S2, S3
ACG-25	70	35	23	23.5	36	29.5	24.5	81	57	45	40	45	9.0	4.5	16.0	60.0	M8,M6,M6
ACG-35	100	50	34	33.0	48	40.0	32.0	106	82	62	52	62	12.0	7.0	21.0	80.0	M10,M8,M8
ACG-45	120	60	45	37.5	60	50.0	40.0	135	100	80	60	80	15.0	8.0	24.0	105.0	M12,M10,M12

* Maximum rail length is 3000 mm.

Figure 8.5.34 Back-to-back circular arch linear guides. (Courtesy of Thomson Industries.)

Model	Load Capicity (N)		Moment Capacity			
	C	Co	Mt	Mto	Ml	Mlo
ACG-25	13500	25000	190	350	95	175
ACG-35	25500	44000	530	910	265	455
ACG-45	42500	70500	1160	1910	580	905

Figure 8.5.35 Load capacity of linear guides of Figure 8.5.34. (Courtesy of Thomson Industries.)

$$F_{3Z,FX} = F_{4Z,FX} = \frac{-F_X z_{FX}}{2(x_1 - x_4)} \tag{8.5.14}$$

Y direction forces cause Y direction forces in the bearing carriages that are assumed to be proportional to the bearing carriages' relative XZ location with respect to the point-of-force application. First consider how the relative X position affects the distribution of forces between bearing carriages 1 and 2 and 3 and 4. Lumping carriages 1 and 2 together and 3 and 4 together, one finds that the relative X positions of the carriages and applied force cause the following allocation of forces:

$$F_{1+2Y} = -F_Y \left(\frac{x_4 - x_{FY}}{x_4 - x_1} \right) \tag{8.5.15}$$

$$F_{3+4Y} = F_Y \left(\frac{x_1 - x_{FY}}{x_4 - x_1} \right) \tag{8.5.16}$$

The distribution of the force allocated to carriages 1 and 2 between carriages 1 and 2 then depends on their Z location relative to the force:

$$F_{1Y,FY} = -F_Y \left(\frac{x_4 - x_{FY}}{x_4 - x_1} \right) \left(\frac{z_2 - z_{FY}}{z_2 - z_1} \right) \tag{8.5.17}$$

$$F_{2Y,FY} = F_Y \left(\frac{x_4 - x_{FY}}{x_4 - x_1} \right) \left(\frac{z_1 - z_{FY}}{z_2 - z_1} \right) \tag{8.5.18}$$

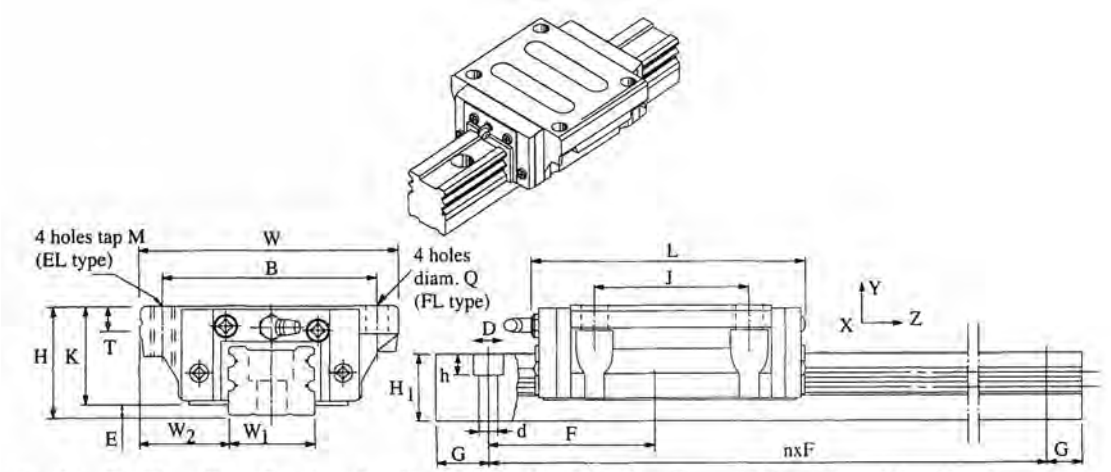
Similarly for carriages 3 and 4:

$$F_{3Y,FY} = -F_Y \left(\frac{x_1 - X_{FY}}{x_4 - x_1} \right) \left(\frac{z_4 - z_{FY}}{z_3 - z_4} \right) \tag{8.5.19}$$

$$F_{4Y,FY} = F_Y \left(\frac{x_1 - X_{FY}}{x_4 - x_1} \right) \left(\frac{z_3 - z_{FY}}{z_3 - z_4} \right) \tag{8.5.20}$$

A Z direction force with an X axis offset causes Z direction forces in the bearing carriages, which are assumed to be evenly distributed between pairs acting as couples:

$$F_{1Z,FZ} = F_{2Z,FZ} = \frac{-F_Z}{2} \left(\frac{x_4 - x_{FZ}}{x_4 - x_1} \right) \tag{8.5.21}$$



#*	H	E	W_2	W	B	L	J	K	T	Q	M	W_1	H_1	F	d,D,h	G	Std. rail lengths
15	24	4.5	16	47	38	55	30	19.5	8	4.5	M5	15	14	60	4.5,7.5,5.3	20	60,280,400,520
20	30	7	21.5	63	53	69.4	40	23	10	6	M6	20	19	60	6,9.5,8.5	20	280,400,520,700
25	36	5.5	23.5	70	57	80.8	45	30.5	11	7	M8	23	22.5	60	7,11,9	20	340,460,640,820,1000
30	42	7.5	31	90	72	95.2	52	34.5	11	9	M8	28	27.5	80	9,14,12	20	440,660,760,1000,1240
35	48	7.5	33	100	82	111	62	40.5	12	9	M10	34	31	80	9,14,12	20	600,760,1000,1240,1560
45	60	10	37.5	120	100	137	80	50	13	11	M12	45	37.5	105	14,20,17	22.5	780,990,1200,1410,1620
55	70	13	43.5	140	116	160	95	57	14	14	M14	53	45	120	16,23,20	30	1020,1260,1500,1980
65	90	14	53.5	170	142	186	110	76	23	16	M16	63	53	150	18,26,22	35	1270,1570,2170,2620

*Prefix by LY and suffix by EL or FL. Note that LY15&20 have a single row on each side.

Figure 8.5.36 Gothic arch groove linear guides. (Courtesy of NSK Corp.)

	Stiffness (K_Y, K_Z) (N/micron) Preload				Load capacity (kN) $F_Y = F_Z$		Moment capacity (N-m)		
Model	Heavy	Medium	Light	Very light	Dyn. C	Static C	Static M_X	Static M_X	Static M_X
LY15		167	137	98	6.05	7.45	70	50	50
LY20		196	167	127	9.8	11.3	140	90	90
LY25	461	392	284	167	17.4	26.5	310	210	210
LY30	578	480	323	196	25.7	38.4	540	360	360
LY35	657	578	363	245	35.9	52.2	900	590	590
LY45	862	735	500	314	52.6	78.8	1800	1180	1180
LY55	1019	882	598	372	80.9	115.0	3130	2060	2060
LY65	1558	1343	911	559	171.0	230.0	8530	5440	5440

Figure 8.5.37 Stiffness and load capacity of linear guides of Figure 8.5.36. (Courtesy of NSK Corp.)

$$F_{3Z,FZ} = F_{4Z,FZ} = \frac{F_Z}{2}\left(\frac{x_1 - x_{FZ}}{x_4 - x_1}\right) \tag{8.5.22}$$

A Z direction force with a Y axis offset causes Y direction forces in the bearing carriages with bearing carriages 1 and 4 acting as a couple with carriages 2 and 3:

$$F_{1Y,FZ} = F_{4Y,FZ} = \frac{F_z y_{FZ}}{2(z_1 - z_2)} \tag{8.5.23}$$

$$F_{2Y,FZ} = F_{3Y,FZ} = \frac{-F_z y_{FZ}}{2(z_1 - z_2)} \tag{8.5.24}$$

Various manufacturers' catalogs give specialized examples of this general case, but the general form is more amenable to inclusion in a spreadsheet program, where one can enter all types of forces and locations and then click the mouse and see their effect on the bearing reaction forces. With these reaction forces and bearing stiffnesses, estimates of the carriage error motions can easily be determined for inclusion in the machine's error budget.

Four bearing carriages at the corners of a structure work well to support boxy-type structures such as machine columns that support other axes; however, in some cases when supporting a boxy structure it may be desirable to distribute the load along the rails to decrease local deformations. In

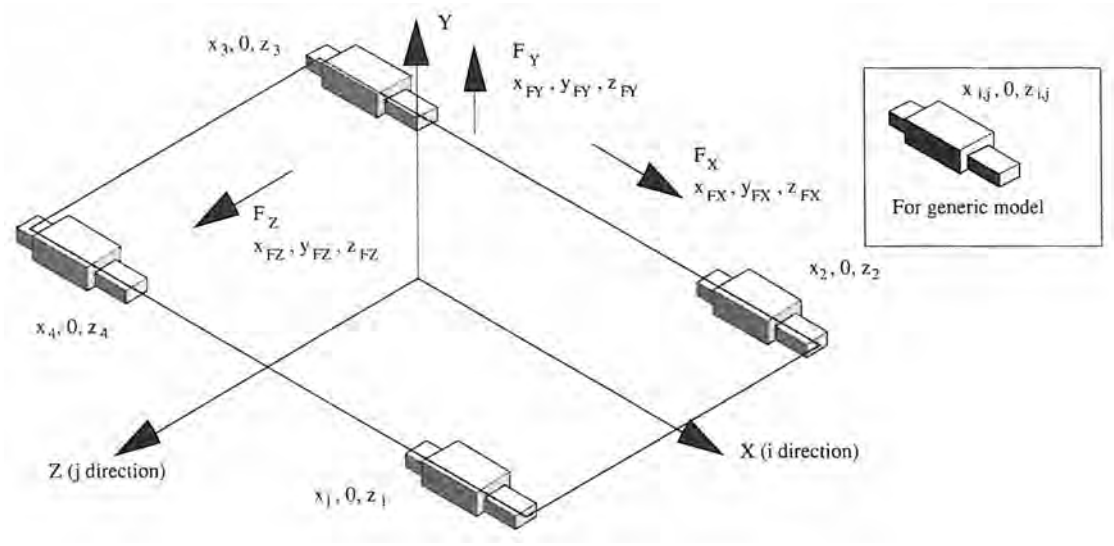


Figure 8.5.38 Generalized model of a four-bearing block system.

this case, the structure can still be considered to be rigid and loads on the carriages can be found in a manner similar to that described for finding the loads on individual rollers. In this case the spring constants are those associated with the stiffness of the bearing carriages. The general case is also shown in Figure 8.5.38, and superposition is again used to combine the effects of different force sources at different locations.

An X direction force F_X can cause Y and Z direction forces on the bearing carriages. The Y direction forces are caused by the Y offset. There are M × N bearing carriage Y direction forces and deflections or 2 × M × N unknowns. The equilibrium equations are

$$\Sigma F_Y = 0 = \sum_{i=1}^{M}\sum_{j=1}^{N} F_{i,jY} \tag{8.5.25}$$

$$\Sigma M_x = 0 = \sum_{i=1}^{M}\sum_{j=1}^{N} F_{i,jY} z_{ij} \tag{8.5.26}$$

$$\Sigma M_z = 0 = -F_X y_{FX} + \sum_{i=1}^{M}\sum_{j=1}^{N} F_{i,jY} x_{ij} \tag{8.5.27}$$

These provide three equations. The force displacement relation provides M × N equations:

$$F_{ijY} = K_{ijY}\delta_{ijY} \tag{8.5.28}$$

The deflections are constrained such that the respective ends of the "springs" remain in planes which are not necessarily parallel:

$$\delta_{ijY} = \delta_{1,1Y} - \frac{(\delta_{1,1Y} - \delta_{M,1Y})(x_{ij} - x1, 1)}{x_{M,1} - x_{1,1}} - \frac{(\delta_{M,1Y} - \delta_{M,NY})(z_{ij} - z_{M,1})}{z_{M,N} - z_{M.1}} \tag{8.5.29}$$

The second term is the contribution from the bearing carriages' relative X position and the third term is the contribution from the relative Z position. This provides M × N - 3 equations; hence there are a total of 2 × M × N equations and 2 × M × N unknowns, so the system is determinate.

The M × N Z direction forces and the M × N deflections of the bearing carriages are caused by a Z offset of the X direction force F_X and they are found in a similar manner:

$$\Sigma F_Z = 0 = \sum_{i=1}^{M}\sum_{j=1}^{N} F_{i,jZ} \tag{8.5.30}$$

There is no moment about the X axis since the YZ plane is in the plane of the bearing carriages. The moment about the Y axis is

$$\Sigma M_Y = 0 = F_X z_{FX} - \sum_{i=1}^{M} \sum_{j=1}^{N} F_{i,jZ} x_{i,j} \tag{8.5.31}$$

These provide two equations and the force displacement relation provides M × N more equations:

$$F_{i,jZ} = K_{i,jZ} \delta_{i,jZ} \tag{8.5.32}$$

The motions are constrained to remain in a plane and where each bearing carriage is anchored to a rigid plane (the carriage) and rides on a rigid rail; thus its deflection is a function only of its X position:

$$\delta_{i,jZ} = \delta_{1,1Z} - \frac{(\delta_{1,1Z} - \delta_{M,1Z})(x_{ij} - x_{1,1})}{x_{M,1} - x_{1,1}} \tag{8.5.33}$$

This provides M × N - 2 equations, so once again the system is determinate.

To find the Y forces and deflections on bearing carriages due to a force F_Y with an X and Z offset:

$$\Sigma F_y = 0 = F_Y + \sum_{i=1}^{M} \sum_{j=1}^{N} F_{i,jY} \tag{8.5.34}$$

$$\Sigma M_z = 0 = F_Y x_{FY} + \sum_{i=1}^{M} \sum_{j=1}^{N} F_{i,jY} x_{ij} \tag{8.5.35}$$

$$\Sigma M_x = 0 = -F_Y z_{FY} - \sum_{i=1}^{M} \sum_{j=1}^{N} F_{i,jY} z_{ij} \tag{8.5.36}$$

The Y direction deflections and geometric constraints are given by Equations 8.5.28 and 8.5.29, respectively.

A Y offset for a Z direction force F_Z causes Y direction forces in the bearing carriages, which sum to zero as in Equation 8.5.25. The moments about the X and Z axes are given by

$$\Sigma M_x = 0 = -F_z y_{FY} - \sum_{i=1}^{M} \sum_{j=1}^{N} F_{i,jY} z_{ij} \tag{8.5.37}$$

$$\Sigma M_x = 0 = \sum_{i=1}^{M} \sum_{j=1}^{N} F_{i,jY} x_{ij} \tag{8.5.38}$$

The deflections and geometric constraints are given by Equations 8.5.28 and 8.5.29, respectively.

An X offset for a Z direction force F_Z causes Z direction forces in the bearing carriages, where

$$\Sigma F_Z = 0 = F_Z + \sum_{i=1}^{M} \sum_{j=1}^{N} F_{i,jZ} \tag{8.5.39}$$

$$\Sigma M_Y = 0 = -F_Z x_{FZ} - \sum_{i=1}^{M} \sum_{j=1}^{N} F_{i,jZ} x_{ij} \tag{8.5.40}$$

and the deflections and geometric constraints are given by Equations 8.5.32 and 8.5.33, respectively.

These equations for forces and deflections of bearing carriages based on generic forces applied at generic locations can practically be solved only using numerical methods. Once the analysis is programmed it will run much faster than a finite element program; hence this method is very useful for initial sizing of bearings and error budgeting prior to the construction of finite element models.

For large machines where a big flat table is in motion (e.g., a very large machining center), it is often desired to minimize the weight of the moving table. In this case, one would make the

base of the machine rigid and use multiple bearing carriages on two or more rails to support a light machine tool table and make it perform as well as a heavier more rigid table supported by only four bearing carriages. To make a first-order design of this type of table, one merely makes sure that any portion of the table bounded by four bearing carriages is rigid with respect to those four bearing pads. However, because the bearing carriages on a linear guide are each fully constrained and no machined surface is perfect, when multiple bearing units are used next to each other on a single rail, each unit's allowable load should be multiplied by a load factor f_C: For 1, 2, 3, 4, and 5 units the load factors are 1.00, 0.81, 0.72, 0.66, and 0.61, respectively.[63] Note that this factor will vary somewhat between manufacturers. It also depends on how the bearing carriages are fastened to the machine tool carriage. For example, if they are just bolted in place, then the weighting factor above should be applied. If three or four bearing carriages are bolted to the machine tool carriage and then the other bearing carriages are shimmed and grouted in place before their mounting bolts are installed, then better alignment is ensured and the values above might be too conservative.

The analysis above also yields the horizontal and vertical loads that each bearing carriage will be subjected to. These loads are added linearly to obtain an equivalent load to be used in linear motion guide life calculations. When the load varies with travel distance, Equation 8.3.5 is used to obtain the equivalent load. Temperatures above 100°C also affect the equivalent load. A temperature load factor f_T is defined as

$$f_T = 1.2335 - 4.1138 \times 10^{-3} - T + 2.7506 \times 10^{-5}\, T^2 - 9.7125 \times 10^{-8} T^3 \tag{8.5.41}$$

The life equation given by Equation 8.3.3 thus becomes

$$L(\text{km}) = 50 \left(\frac{f_T f_C C_N}{f_w F_C} \right)^3 \tag{8.5.42}$$

With a spreadsheet program, these equations make it fun and easy to include lots of different types of forces from every source you can imagine. It gives the design engineer an efficient real-time tool to play the game of: What if I change this: how will it affect the sizing of my bearings? Spreadsheets also allow a design engineer to easily perform parametric studies (graphs) of different design parameters, something which is difficult to do with finite elements. When the general model that includes bearing carriage stiffnesses is used along with a model for the overconstrained system of how the "rigid" carriage moves in space when one of the bearings deflects, this spreadsheet can also generate values of the error motions for the carriage. This program might be linked to an error budget program so that one can instantly see how a change in an applied force causes a change in the tool tip position.

Allowance for Thermal Growth

Like a spindle supported by rotary motion angular contact bearings, the moving component of a bearing assembly generally becomes warmer faster than the fixed part because the former generally has less thermal mass. Hence when the linear guide rails are stationary and the carriages are moving, face-to-face linear guides are more thermally stable[64] with respect to maintaining constant preload, although in general this type of unit does not see high enough speeds for thermal growth to cause a significant change in preload.

Like all other overconstrained systems, one must try to control heat generation or allow it to dissipate uniformly through the structure so that the entire machine grows evenly, thereby minimizing thermal Abbe errors. Unfortunately, overconstraint is the price one pays for high load-carrying capability; hence careful first-order estimates of thermal errors followed by detailed finite element models and tests on similar systems are often a necessary part of the development of a new precision machine.

Alignment Requirements

One must be sure to consider mounting details such as those shown in Figure 8.5.39 (this is true of course for all types of bearing carriages and rails). For applications where the machine

[63] From THK Corp. bearing literature.

[64] For linear motion bearings, the carriage (analogous to the outer race) of the face-to-face type moves, whereas in rotary bearings the inner race (analogous to the rail) moves.

is subjected to shock and vibration loads, both rails should be fully constrained. In this case it is imperative that the reference edges, which the rails are pushed up against, are made parallel to the accuracy desired for the machine. For general applications, one rail, which is referred to as the master rail, is fully constrained. The secondary, or subsidiary, guide rail is made parallel to the master using gage blocks or a dial indicator. The secondary rail is then just bolted to the machine without using fixed mechanical means for lateral restraint. In either case, after the rails are parallel and bolted in place,[65] the bearing carriages are attached to the machine tool carriage in positions as determined from the analysis above.

There will always be a parallelism error between the rails. Using the methods described in Section 2.5, it is possible to estimate the effect of this error on the accuracy of the carriage supported by the linear guides. Other methods have also been developed to estimate this type of error.[66]

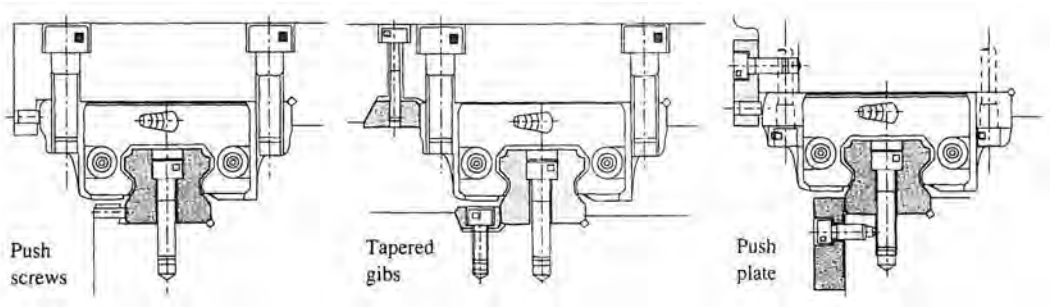


Figure 8.5.39 Linear guide mounting methods. (Courtesy of THK Co., LTD.)

The type of groove also affects the ability of the linear guide to accommodate misalignment. Figure 8.5.40 compares the effects of alignment errors on system performance for a circular arc groove bearing and a Gothic arch bearing with its balls in full four-point contact with the grooves. Alignment insensitivity implies less stiffness in the accompanying direction, and alignment sensitivity implies greater stresses. As always, one rarely gets something for nothing. Note that in addition to rolling resistance caused by preload or misalignment, one needs to consider the applied load and coefficient of static and dynamic friction, which can be in the range of 0.01-0.005.

Preload and Frictional Properties

There are essentially five preload classes for linear guides, the same five as are used for rotary motion bearings, as shown in Table 8.3.4. As was shown earlier, stiffness is greatly affected by preload and is often the dominant factor in determining what preload to specify. Preload and the shape of the groove also affect rolling friction, as shown in Figures 8.3.6, 8.5.2, and 8.5.4.

8.5.3 Recirculating Rollers

Maximum load capacity and stiffness in a linear motion bearing are obtained with rollers (e.g., cylinders). Also, like rotary motion bearings with rollers, linear motion bearings with rollers need a method to keep the rollers from skewing sideways as they roll. This can be accomplished in a number of ways, including placing each roller in a plastic carrier, having each roller's turned-down ends attached to a chain so that all rollers are connected together, or through the use of a belt that encircles the roller recirculation path and passes across a groove in the middle of each roller. Each of these methods will be illustrated when specific types of bearing units are discussed below.

8.5.3.1 Cylindrical Rollers on Flat Rails

There are two main types of recirculating roller linear bearings where the rollers contact flat ways that are referred to here as crawler track and merry-go-round types. Crawler track bearings have many variations as shown in Figure 8.5.41. These bearing blocks are essentially meant to replace

[65] Mounting procedures for bolted joints to ensure stability were discussed in Section 7.5.1.

[66] S. Shimizu and N. Furuya, "Accuracy Average Effect of Linear Motion Ball Guides System for NC Machines-Theoretical Analysis," Progress in Precision Engineering, P. Seyried et al. (eds.), Springer-Verlag, New York, 1991, p. 324.

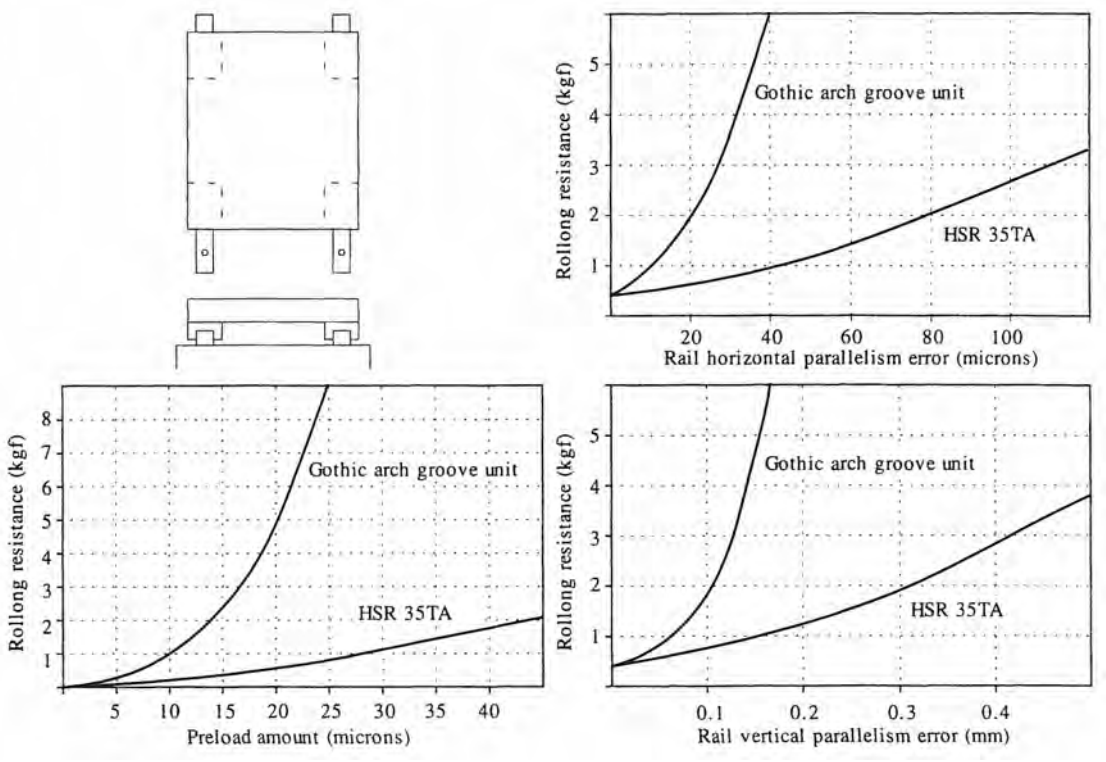


Figure 8.5.40 Comparison of Gothic and circular arch groove rolling resistance versus preload and alignment. Table weight was 30 kgf (Courtesy of THK Co., LTD.)

sliding element bearings in virtually any type of sliding bearing application (where space allows). Other types of crawler track linear bearings are configured in bearing blocks with multiple tracks which can be fitted to rectangular or angled rails as shown in Figures 8.5.42 and 8.5.43, respectively. Sizes of single-track units are shown in Figure 8.5.44. The merry-go-round recirculating roller bearing shown in Figure 8.5.45 uses shorter rollers than the crawler track type, but it has a much lower profile.

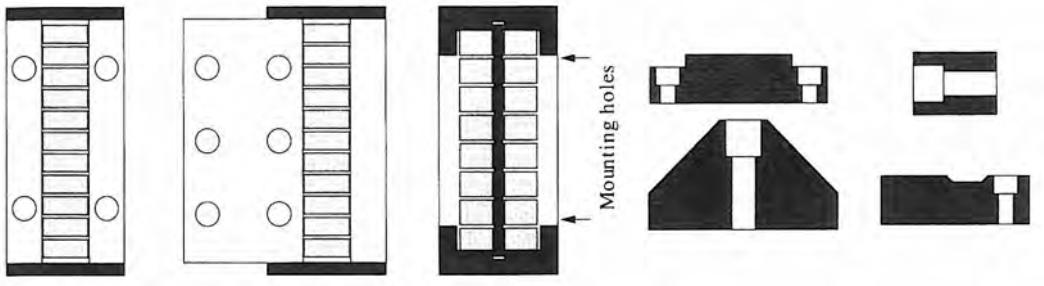


Figure 8.5.41 Basic crawler track type recirculating roller bearings for linear motion and some typically available rail types.

Running Parallelism, Repeatability, and Resolution

For a quasi-kinematic arrangement of five bearing blocks on a vee and flat rail arrangement, running parallelism can be as good as the rails can be made parallel. For properly supported and ground systems this can be about 5-10 μm/m. This type of bearing is not often used on machines that are hand finished, but if it were, it could potentially achieve an order-of-magnitude and better increase in performance. For fully constrained two-rail systems, one needs to consider alignment

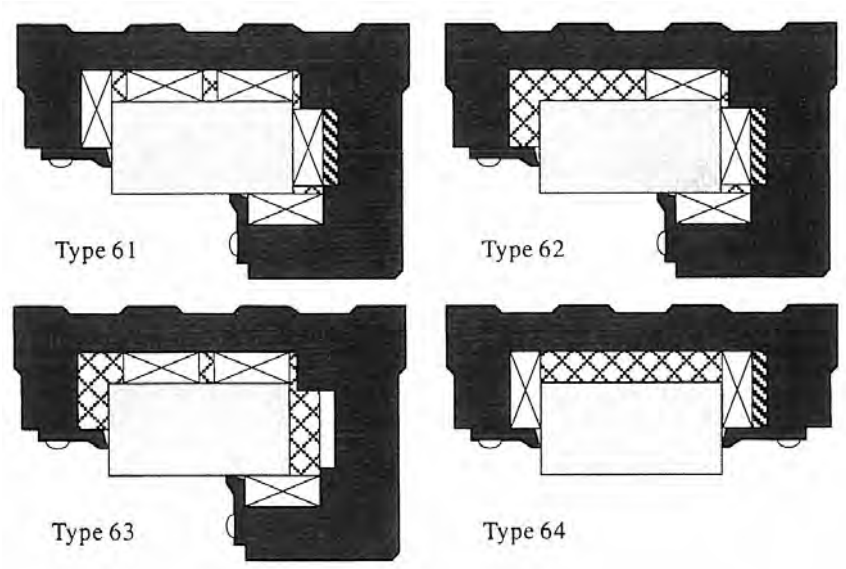


Figure 8.5.42 Multiple-crawler-track-tread recirculating roller linear bearings for use on rectangular rails. (Courtesy of Schneeberger Inc.)

effects, which may decrease parallelism by an order of magnitude depending on the care taken during machining and assembly.

Repeatability of these types of bearings when they are machine ground can be as good as 1 μm when a medium preload (5% of rated load) is used. Even with rolling elements, after extensive use they wear in and repeatability can increase. With a properly designed servo-control system, $^1/_2$-1 μm resolution should be readily attainable.

Lateral and Moment Load Support Capability

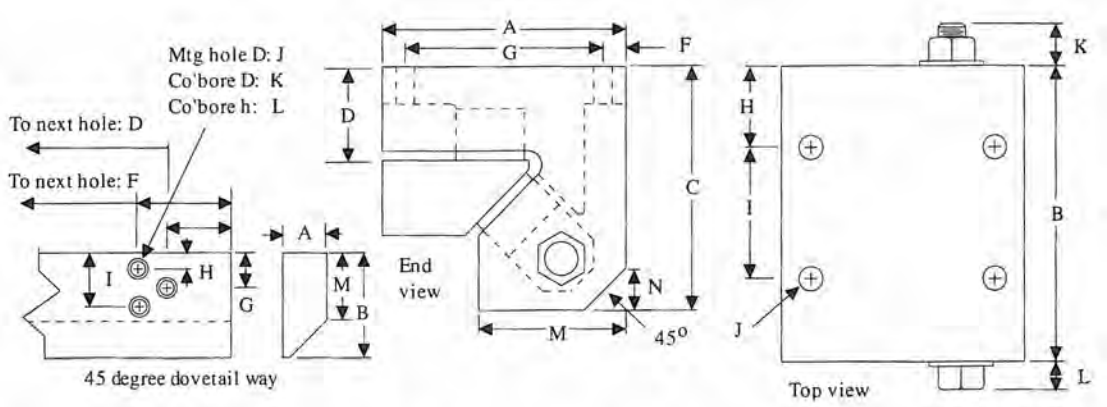
Recirculating rolling element linear bearings are meant primarily to support loads normal to the surfaces the rollers contact and thus must be used in pairs to support moment loads. For their size, linear bearings with recirculating rollers can support very large loads compared to bearing blocks with recirculating balls, and the former generally have greater stiffness; however, recirculating roller units are more sensitive to misalignment. In order to determine the loads on individual bearing blocks, one would use a method similar to that described above for linear guides.

Figure 8.5.44 shows sizes for a typical family of single-crawler-track recirculating linear bearings. This type of bearing is very versatile and gives the design engineer quite a bit of freedom to accommodate many different types of rail geometries. Figures 8.5.42 and 8.5.43 show multiple-track versions that are meant to ride on rectangular and dovetailed rails, respectively. In order to minimize alignment and assembly headaches, if possible one may want to use one of these types of multiple-track units. Figure 8.5.45 shows available sizes of merry-go-round recirculating roller bearings and Figure 8.5.46 shows typical mounting configurations. Figure 8.5.47 shows load/deflection curves for the merry-go-round recirculating roller bearings.

Allowance for Thermal Growth

It is possible to design quasi-kinematic arrangements for recirculating roller bearings for linear motion as long as one considers five bearing blocks on a vee and flat rail arrangement kinematic; however, each bearing block in addition to resisting normal loads to a large degree also resists pitching loads to some degree. Thus if the surfaces the blocks are bolted to are not aligned, only one or two rollers on each block can end up supporting most of the load. This type of misalignment problem can occur for most types of modular bearings.

As always, nonkinematic systems must rely on elastic deformation of the components and/or careful design to balance thermal growth. If the bearings are used in quasi-steady applications, then thermal growth should not be a problem as long as the temperature is uniform and the carriage and



Model	A	B	C	D	F	G	H	I	J	K	L	M	N	Fstatic (lbf)	Fdyn. (lbf)
150-2D	3.125	3.500	2.814	1.1250	0.281	2.500	1.125	1.125	0.344	0.535	0.320	1.72	0.44	4330	2200
200-2D	4.375	4.375	3.813	1.5000	0.375	3.625	1.125	2.125	0.406	0.785	0.415	2.50	0.50	7400	4030
300-2D	6.625	6.000	5.688	2.3125	0.625	5.375	1.625	2.750	0.656	1.035	0.575	3.75	1.13	18310	10010
400-2D	9.875	8.000	7.750	3.0625	0.750	8.375	1.875	4.250	0.781	1.285	0.720	5.50	2.00	33830	18780

Dovetail way dimensions All dimensions are in inches.

MODEL	A	B	C	D	E	F	G	H	I	J	K	L	M
150-DW	1.000	2.000	1.000	6.000	-	-	0.625	-	-	0.344	0.500	0.344	1.250
200-DW	1.375	2.750	1.000	6.000	-	-	1.125	-	-	0.406	0.594	0.406	1.625
300-DW	1.750	4.000	-	-	1.000	8.00	-	0.750	1.875	0.656	0.969	0.656	2.625
400-DW	2.500	6.000	-	-	1.000	10.00	-	0.875	2.375	0.781	1.156	0.781	4.125

Figure 8.5.43 Recirculating roller bearings for dovetail ways. (Courtesy of Detroit Edge Tool Co.)

base the rails are mounted to are made of the same material. For high-speed reciprocating systems, an active cooling system should be considered.

Alignment Requirements

Although these types of bearings were meant to replace sliding contact bearings in many applications, they cannot be hand finished during assembly; hence they must rely on the preload force to make sure that all the rollers in each bearing block contact the rail. Since the rollers rely on line contact, it is very important that the rail surfaces be well aligned. For some large gantry machines, rails are spliced end to end to form a rail that may be tens of meters long.

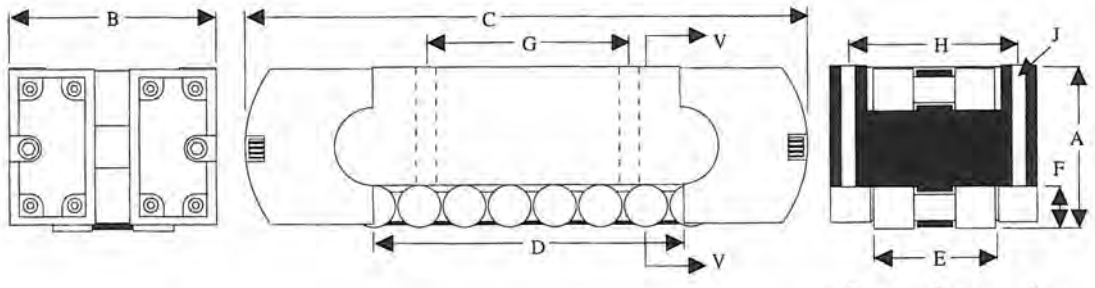
For applications where the bearing rails lie in a horizontal plane, when two rectangular rails are used, the two vertical sides of *one rail only* should be used to guide horizontal straightness of travel. The second rail serves only to withstand vertical forces. This makes the design's performance much less sensitive to manufacturing or environmental effects. Similar axioms can be thought of for other axis orientations.

Preload and Frictional Properties

Preload is provided by the weight of the carriage for quasi-kinematic systems and by gibs (typically, setscrew or tapered gib) or eccentric mounting pins in other cases. In most cases, static and dynamic coefficients of friction on the order of 0.01-0.005 can be expected. For system design purposes, one should consider that the static and dynamic coefficients of friction will have the worst possible values and the dynamic model should try all different permutations.

8.5.3.2 Cylindrical Rollers on Crowned Races

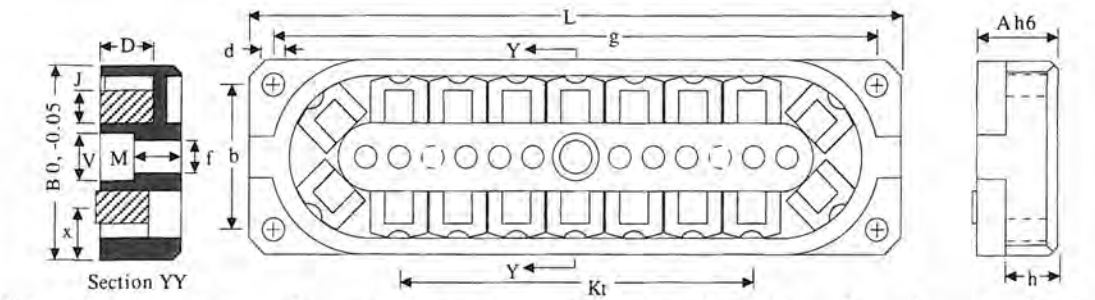
The obvious load carrying advantages of rollers over balls is somewhat tempered by their sensitivity to misalignment, skewing, and edge loading. Due to load dependent deformation, manufacturing tolerances, and misalignment of roller bearings, the relative race orientation is subject to translation. As the parallelism of the races degrades, the onset of edge loading and roller taper occurs. The edge loading condition creates a very high stress in the roller and race that leads to shortened life and decreased accuracy. The taper roller condition prohibits straight line rolling, causing the roller to roll toward the more heavily loaded end. While the replacement of cylindrical rollers with convex



Model	A	B	C	D	E	F	G	H	J	F_{static} (lbf)	$F_{dyn.}$ (lbf)	K (lbf/µin)
21050	0.5625	0.875	2.00	1.12	0.45	0.15	0.75	0.672	#4	4330	2200	3.2
21100	0.7500	1.000	2.88	1.50	0.600	0.200	1.000	0.812	#4	7920	4240	5.1
21150	1.1250	1.500	4.00	2.25	0.900	0.300	1.500	1.219	#8	17260	9830	7.0
21200	1.5000	2.000	5.50	3.00	1.200	0.400	2.000	1.625	#10	29460	17750	9.5
21250	2.2500	3.000	8.12	4.50	1.800	0.600	3.000	2.437	0.25	68200	43300	15.3
21300	3.0000	4.000	11.00	6.00	2.400	0.800	4.000	3.250	0.31	122000	81000	19.1

Dimensions are in inches.

Figure 8.5.44 Single-crawler-track recirculating roller bearing. For speeds in excess of 1200 ipm, special models are available. (Courtesy of Tychoway Bearings Co.)



Size	L	kt	A	B	h	g	b	d	f	V	M	D	J	x	C_{100}*
3RE	70	33	7	16	5	64	10	3	2.5	4	5.5	3	2.8	3.5	4500
6RE	90	52	10	22	6.7	83	15	3.8	3	5.2	7.8	6	4	5	12000
9RE	135	78	15	33	10.5	125	23	5	5	8.2	11.8	9	5.6	7.5	27000

* Loading (N) with a service life of 100 km and lubricated at intervals with grease or oil. All dimensions are in mm

Figure 8.5.45 Typically available sizes of merry-go-round recirculating roller bearings for low-profile requirements. (Courtesy of Schneeberger Inc.)

rollers alleviates the edge loading condition, the taper roller condition still exists. Edge crowning of the races lessens both undesirable conditions, but does not eliminate them.

These problems can be overcome with the use of cylindrical rollers on a continuously radiused race, such as the Accumax® roller bearing sold by Thomson Industries and shown in Figure 8.5.48. In this design, the cylindrical rollers are loaded between two arcuate races, that when translated or rotated relative to each other still present the roller with large radii to roll upon. The design is a four circuit, face-to-face design that allows for equal loading in all directions. The result is that the Accumax® roller bearing has characteristics similar to those of linear guide type bearings discussed in Section 8.5.2.4, but with substantially higher capacities and stiffnesses.

Running Parallelism, Repeatability and Resolution

The arcuate race roller bearing's running parallelism conforms to the typical running parallelism for linear guides, shown in Figure 8.5.31. The repeatability of the roller system depends upon the accuracy class and may vary from between 0.1 to 10 microns. Likewise, the resolution of the roller bearing system may vary from 0.1 to 10 microns, dependent upon the servo system. The preload of the roller bearing will effect the resolution of the system.

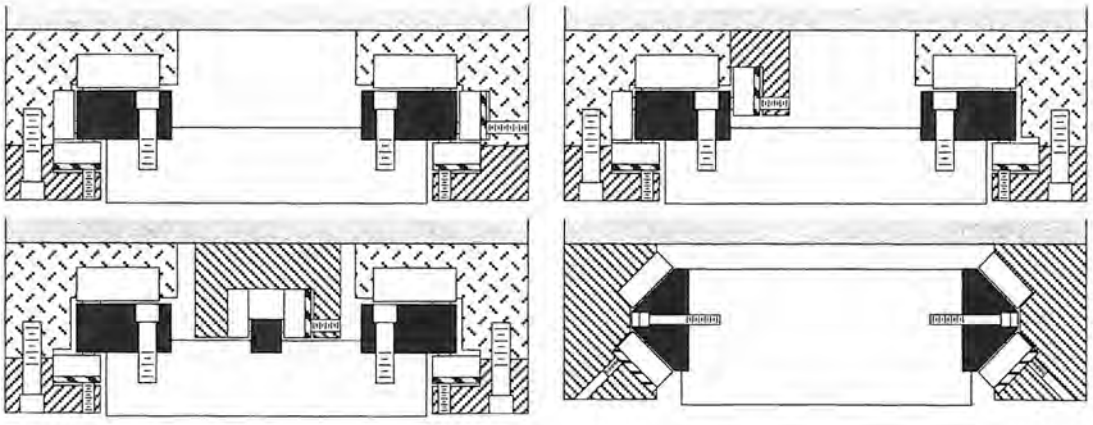


Figure 8.5.46 Typical mounting arrangements for merry-go-round recirculating roller bearings. (Courtesy of Schneeberger Inc.)

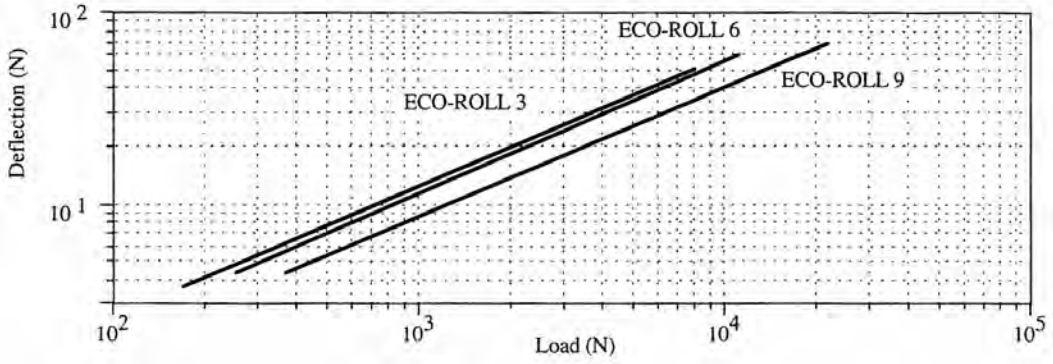


Figure 8.5.47 Load-deflection curves of merry-go-round recirculating roller bearings listed in Figure 8.5.43. (Courtesy of Schneeberger Inc.)

Lateral and Moment Load Support Capability

The Accumax® roller bearing system utilizes a face-to-face rolling element orientation, as shown in Figure 8.5.48. Characteristic load capacities and dimensions are shown in Figure 8.5.49. Because of this rolling element orientation, sensitivity to roll misalignment is low, but moment capacity is reduced; however, this lends itself to machine tool applications, where the rails are well separated, as discussed in section 8.5.2.4.

Allowance for Thermal Growth

As the relative velocity of any bearing system increases, the system tends to generate more heat, with a potential for the occurrence of increased friction and variation of the effective preload. Because typical linear bearing applications in machine tools do not require very high speeds, compared to the surface speeds on spindles, and because of the face-to-face configuration of the rolling elements, thermally dependent deviation of performance is not a significant factor. Again, as discussed in section 8.5.2.4, careful assessment of the thermal loads should be made.

Alignment Requirements

Roller bearings are inherently very rigid, so care must be taken in the design of multi-rail systems. Generally, this is accomplished through the use of one master rail and one floating rail (prior to final bolting) to achieve parallelism between the rails. By utilizing an arcuate rail technology and face-to-face geometry, sensitivity to variation in rail and carriage height is minimized. Because of the arcuate races, this design can be less sensitive to height variation than some face-to-face ball systems, as the latter has a very large contact angle rotation associated with axial misalignment.

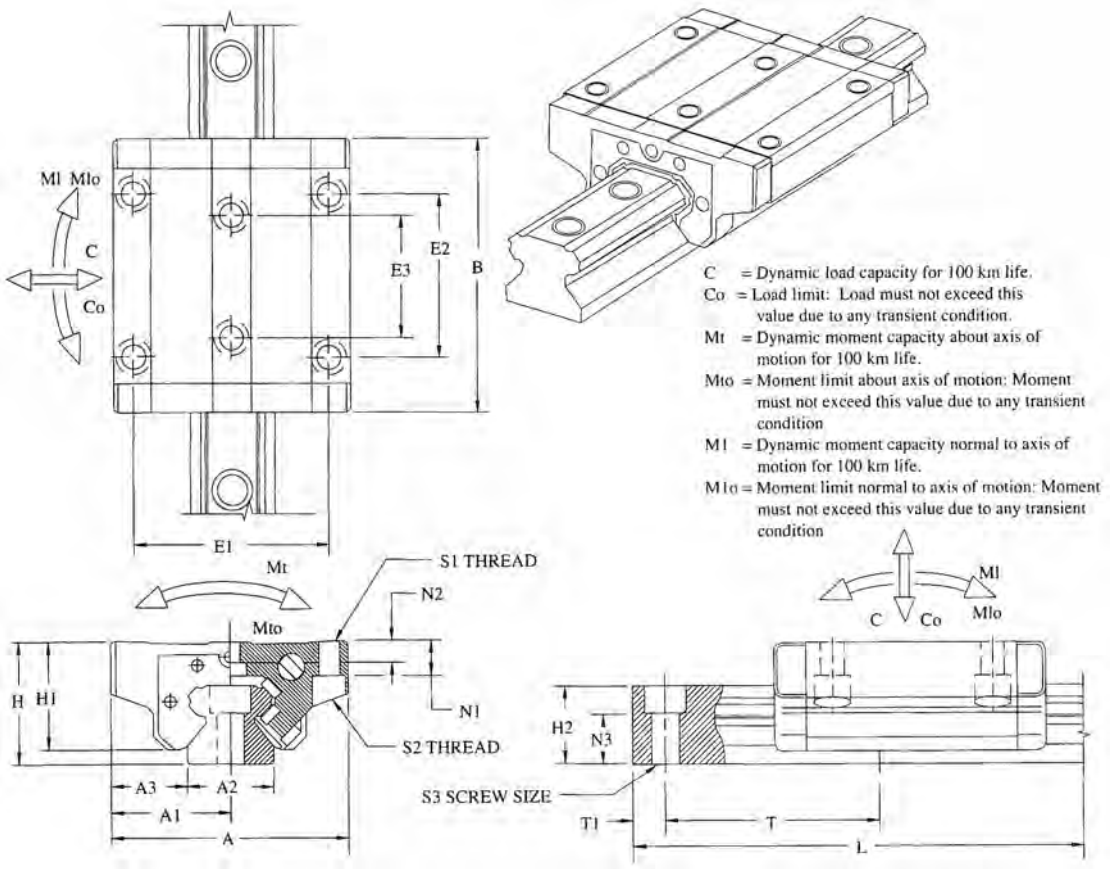


Figure 8.5.48 Accumax® roller bearing. (Courtesy of Thomson Industries.)

Preload and Frictional Properties

The bearing is available in various preload classes to achieve the required stiffness. As the roller reaction is significantly stiffer than the ball reaction, preload effectiveness is reduced at higher preload values. For system design considerations, one may assume a static coefficient of friction of 0.01 and a dynamic coefficient of 0.005.

8.5.3.3 Hourglass-Shaped Rollers on Round Ways

One of the earliest types of unrestricted linear motion bearings designed to support very large loads and have a very low coefficient of friction was the Roundway® bearing, manufactured by Thomson Industries. This design uses hourglass-shaped rollers that are connected together in a crawler track arrangement that rides on round ways (rails) as shown in Figure 8.5.50. These units can be sealed with molded plastic seals that enclose the rolling elements and interface with the round way with rubber wipers.

Running Parallelism, Repeatability, and Resolution

Roundway® bearings are designed for heavy-duty applications where many kilonewtons must be supported. The hourglass-shaped rollers require that the rails be parallel or else one side of the hourglass will bear a larger load, unless a quasi-kinematic configuration of bearings is used as shown in Figure 5.5.51. Running parallelism is typically 10-100 μm/m. Repeatability can be a factor of 10 better. Resolution depends on the servo system used, but because of the low rolling coefficient of friction, repeatability can be 1 μm or better.

Lateral and Moment Load Support Capability

A single block can only support a load normal to the shaft surface. A double block can support loads from two directions. When used in combination, there are a large number of ways that

	ACM-35	ACM-45	ACM-55
A (mm)	100	120	140
A1 (mm)	50	60	70
A2 (mm)	34	45	51
A3 (mm)	33.0	37.5	43.5
H (mm)	48	60	70
H1 (mm)	42.4	53	61.6
H2 (mm)	31.0	38.5	46.5
B (mm)	106	135	160
E1 (mm)	82	100	116
E2 (mm)	62	80	95
E3 (mm)	52	60	70
E3 (alt) (mm)	62	80	95
S1	M10x1.5	M12x1.75	M14x2.0
S2	M8x1.25	M10x1.5	M12x1.75
S3	M8x1.25	M12x1.75	M14x2.0
N1 (mm)	12.0	16.2	19.5
N2 (mm)	7.0	9.5	11.5
N3 (mm)	19.7	24.6	19.5
T (mm)	40.0	52.5	60.0
Tmin (mm)	12	16	18
Lmax (mm)	3000	3000	3000
C (kN)	46.6	80.0	115.7
Co (kN)	79.5	132.4	191.1
Mt (Nm)	590	990	1470
Mto (Nm)	1010	1690	2510
M1 (Nm)	790	1310	1950
M1o (Nm)	1140	2210	3320
Stiffness* (no preload) (N/μm)			
Vertical (down)	1,538	2,000	2,857
Vertical (up)	599	826	1,333
Horizontal	909	1,124	1,471
Stiffness (light preload, 5%) (N/μm)			
Vertical (down)	1,695	2,174	3,030
Vertical (up)	617	847	1,429
Horizontal	1,020	1,266	1,667
Stiffness (medium preload, 10%) (N/μm)			
Vertical (down)	1,786	2,326	3,333
Vertical (up)	633	862	1,471
Horizontal	1,111	1,351	1,724

* Linearized value.

Figure 8.5.49 Characteristics of the Accumax® roller bearing shown in Figure 8.5.48. (Courtesy of Thomson Industries.)

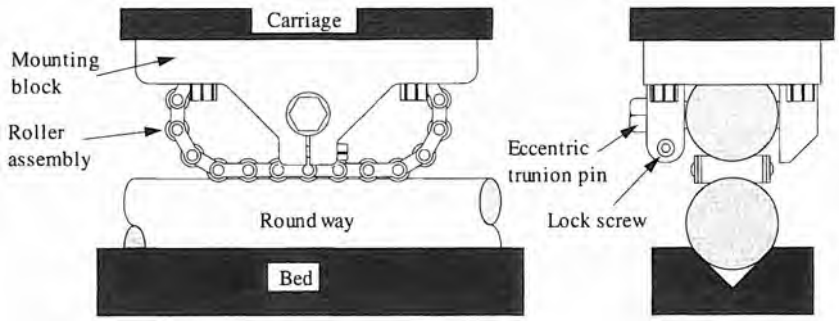


Figure 8.5.50 Roundway® bearings. (Courtesy of Thomson Industries.)

the two basic blocks can be used, two of which are shown in Figure 8.5.51. Sizes and load-carrying capabilities are shown in Figures 8.5.52 and 8.5.53. To compute bearing loads for complex loadings, one can use a method similar to that described for linear motion guides as long as force direction is considered.

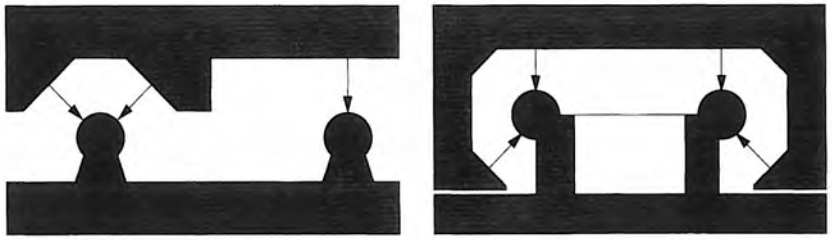


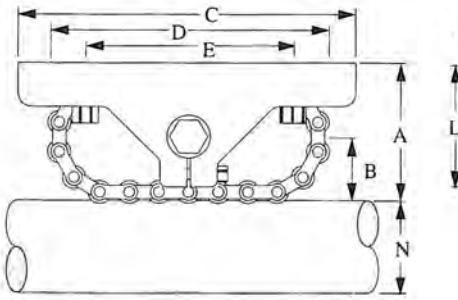
Figure 8.5.51 Two of the many possible mounting configurations for Roundway® bearings. (Courtesy of Thomson Industries.)

Allowance for Thermal Growth and Alignment Requirements

As always, nonkinematic systems must rely on elastic deformation of the components and/or careful design to balance thermal growth. The trunion mount allows the bearing blocks to accommodate pitch misalignment easily and the hourglass shape allows roll misalignment to be accommodated easily. Horizontal parallelism must be strictly controlled to prevent uneven loading of the rollers. An eccentric trunion pin allows for adjustment of carriage parallelism with respect to the plane of the round ways. Typical rail alignment requirements are 0.1 mm/m noncumulative.

Preload and Frictional Properties

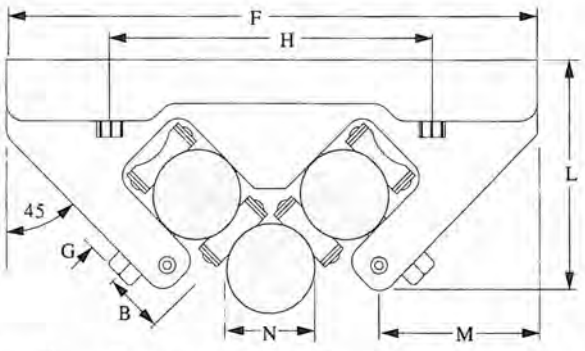
Preload can be provided by the weight of the carriage or by use of an eccentric trunion pin in opposed unit designs. The rollers' varying contact diameter leads to some slippage, but the diameter difference is not too large and the manufacturer claims coefficient of rolling friction as low as 0.005 for the single units and 0.007 for the double units.



Size*	N	A	B	C	D	E	F	G	H	J	L	F_{stat}(lbf)	F_{roll}(lbf)**
8	0.500	1.000	0.450	3	$2\frac{3}{8}$	$1\frac{1}{2}$	$1\frac{1}{4}$	$\frac{3}{16}$	$\frac{15}{16}$	$\frac{5}{16}$	$\frac{7}{8}$	1130	970
16	1.000	1.750	0.8000	5	$3\frac{3}{4}$	$2\frac{1}{2}$	$2\frac{1}{8}$	$\frac{1}{4}$	$1\frac{5}{8}$	$\frac{1}{2}$	$1\frac{1}{2}$	3280	3020
24	1.500	2.500	1.150	$6\frac{1}{2}$	$5\frac{3}{8}$	$3\frac{1}{2}$	$2\frac{7}{8}$	$\frac{5}{16}$	$2\frac{1}{8}$	$\frac{5}{8}$	$2\frac{1}{8}$	6260	6020
32	2.000	3.250	1.500	$8\frac{1}{2}$	$7\frac{3}{8}$	$4\frac{1}{2}$	$3\frac{5}{8}$	$\frac{3}{8}$	$2\frac{3}{4}$	$\frac{3}{4}$	$2\frac{7}{8}$	12500	12360
48	3.000	5.000	2.300	13	11	7	6	$\frac{1}{2}$	$4\frac{1}{4}$	$1\frac{1}{4}$	$4\frac{1}{4}$	25000	24000
64	4.000	6.500	3.000	17	$14\frac{7}{8}$	9	$7\frac{3}{4}$	$\frac{1}{2}$	$5\frac{1}{2}$	$1\frac{1}{2}$	$5\frac{7}{8}$	50000	48000

*Model number is RW-*Size*-S. **For a travel life of 10^7 inches, round way hardness of R-60C and lubrication.

Figure 8.5.52 Sizes of Roundway® bearings and mounting blocks. (Courtesy of Thomson Industries.)



Size*	N	A	B	C	D	E	F	G	H	J	L	M	F_{stat}(lbf)	F_{roll}(lbf)**
8	0.500	1.000	0.450	$1\frac{3}{8}$	$2\frac{3}{8}$	1	3	$\frac{3}{16}$	$2\frac{1}{4}$	$\frac{5}{16}$	$1\frac{3}{8}$	$\frac{11}{16}$	1600	1370
16	1.000	1.750	0.8000	$2\frac{1}{4}$	$3\frac{3}{4}$	$1\frac{5}{8}$	$5\frac{3}{4}$	$\frac{1}{4}$	4	$\frac{1}{2}$	$2\frac{3}{8}$	$1\frac{9}{16}$	4600	4300
24	1.500	2.500	1.150	$2\frac{3}{4}$	$5\frac{3}{8}$	2	$7\frac{7}{8}$	$\frac{5}{16}$	6	$\frac{5}{8}$	$3\frac{3}{8}$	$2\frac{1}{8}$	8900	8600
32	2.000	3.250	1.500	$3\frac{1}{2}$	$7\frac{3}{8}$	$2\frac{1}{2}$	$9\frac{3}{4}$	$\frac{3}{8}$	$7\frac{1}{2}$	$\frac{3}{4}$	$4\frac{3}{8}$	$2\frac{1}{2}$	17700	17500
48	3.000	5.000	2.300	$5\frac{1}{2}$	11	4	$15\frac{1}{2}$	$\frac{5}{8}$	12	$1\frac{1}{4}$	7	$4\frac{1}{4}$	36000	35000
64	4.000	6.500	3.000	7	$14\frac{7}{8}$	5	$19\frac{1}{4}$	$\frac{3}{4}$	15	$1\frac{1}{2}$	$8\frac{5}{8}$	5	72000	70000

*Model number is RW-*Size*-V **For a travel life of 10^7 inches, round way hardness of R-60C and lubrication.

Figure 8.5.53 Sizes of Roundway® bearings and mounting blocks. (Courtesy of Thomson Industries.)

8.6 Flexural Bearings[67]

Sliding, rolling, and fluid film bearings all rely on some form of mechanical or fluid contact to maintain the distance between two objects while allowing for relative motion between them. Since no surface is perfect and no fluid system is free from dynamic or thermal effects, all these bearings have an inherent fundamental limit to their performance. *Flexural bearings* (also called *flexure pivots*), on the other hand, rely on the the stretching of atomic bonds during elastic motion to attain smooth motion. Since there are millions of planes of atoms in a typical flexural bearing, an averaging effect is produced that allows flexural bearings to achieve atomically smooth motion. For example, flexural bearings allow the tip of a scanning tunneling microscope to scan the surface of a sample with subatomic resolution.[68] There are two categories of flexural bearings, monolithic and clamped-flat-spring, as shown in Figures 8.6.1 and 8.6.2.

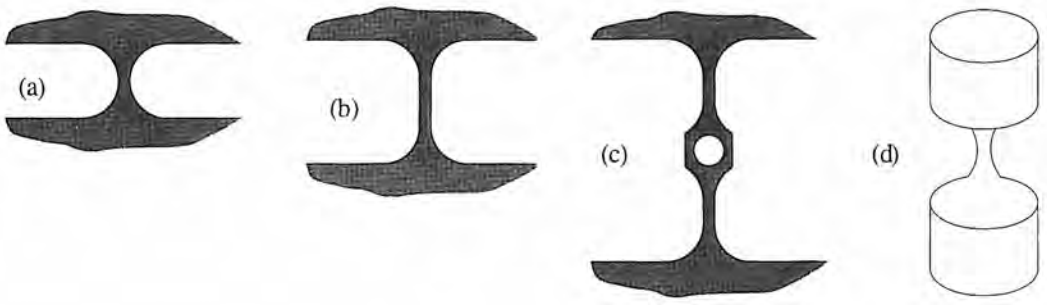


Figure 8.6.1 Types of monolithic flexures.

[67] The author would like to thank Roger Reiss and Graham Siddall for their contributing material which helped in writing this section. The reader may wish to consult the following references: F. S. Eastman, "Flexure Pivots to Replace Knife Edges and Ball Bearings," Univ. of Wash. Eng. Exp. Sta. Bull., No. 86, Nov., 1935; F. S. Eastman, "The Design of Flexure Pivots," J. Aerosp. Sci., Vol. 5, Nov. 1937, pp. 16-21; R. V. Jones, "Parallel and Rectilinear Spring Movements," J. Sci. Instrum., Vol. 28, 1951, p. 38; R. V. Jones and I. R. Young, "Some Parasitic Deflections in Parallel Spring Movements," J. Sci. Instrum., Vol. 33, 1956, p. 11; G. J. Siddall, "The Design and Performance of Flexure Pivots for Instruments," M.Sc. thesis, University of Aberdeen, Scotland, Department of Natural Philosophy, Sept. 1970.

[68] See, for example, G. Binnig and H. Rohrer, "Scanning Electron Microscopy," Helv. Phys. Acta, Vol. 55, 1982, pp. 726-735.

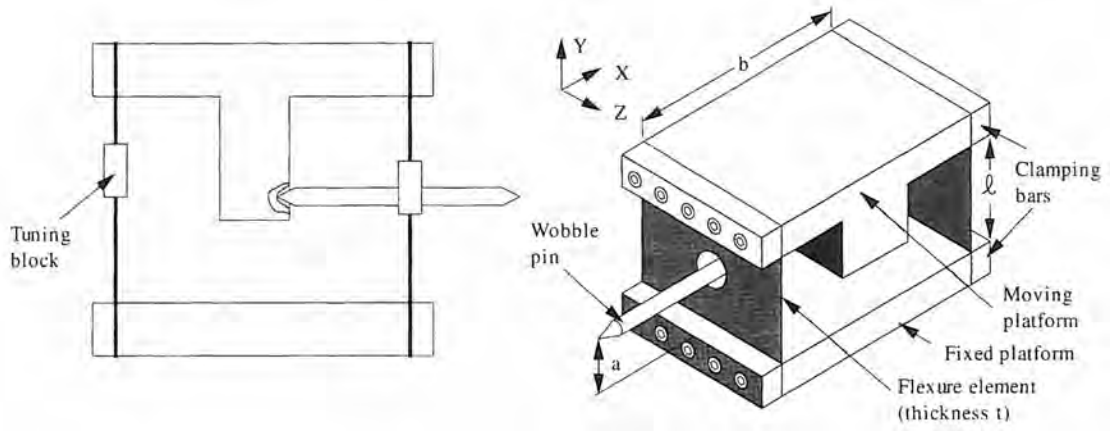


Figure 8.6.2 Four bar linkage flexures. Note that if tuning blocks are to be used, then the end flexural element must be composed of two separate elements spaced a finite distance apart. Also, the wobble pin is shown used with cup jewel bearings. Ideally, the end of the wobble pin would be spherical and it would rest in a seat formed by three balls; thus it would be kinematic.

8.6.1 General Properties

Speed and Acceleration Limits

The speed and acceleration limits of a flexural bearing are limited only by the system natural frequency and the stress levels in the bearing. A common example of a device that uses flexural bearings is an audio loudspeaker where the cone acts as a diaphragm (a type of flexure) which holds the coil and allows it to move with respect to the magnet.

Range of Motion

Flexural bearings are rather limited in their range of motion. For monolithic bearings, the ratio of range of motion to bearing size is on the order of 1/100. For clamped flat spring flexural bearings, the ratio of range of motion to bearing size is on the order of 1/10; thus they are used primarily in small range of motion fine-positioning devices.

Applied Loads

Flexural bearings have been designed to support large precision machine components weighing hundreds of pounds. Flexural bearings are easily custom made using spring steel plates, so one should not be scared away by large applications. Leaf springs on large trucks are flexural bearings: They guide the motion of the truck body with respect to the axle, while also providing a restoring force, although often in flexural bearing design it is desirable to minimize the restoring force.

Accuracy

Accuracy of a flexural bearing depends on how well the bearing was assembled or machined. Even if there is a small off-axis error motion associated with the primary motion, the error motion is usually very predictable and highly repeatable. Flexural bearings cannot attain perfect motion because of:

- Variation in spring strength.
- Variation in spring geometry.
- Overall inaccuracies of manufacture.
- Bending of the bearing in an unintended manner.
- Bending of structure.
- External applied loads (e.g., gravity and the manner in which the actuation force is applied)

These effects can be minimized with careful machining of matched components; furthermore, these effects can often be negated with the use of metal blocks clamped to the flat springs as shown in

Figure 8.6.2. By adjusting the position of the blocks along the length of the flexure, the flexure's performance can be tuned to yield order of magnitude increases in accuracy.

The most common errors in flexures are the pitch angle and vertical motion that accompany linear motion in a four bar linkage flexure shown in Figure 8.6.2. For small displacements, the errors are a function of the distance moved x, the length of the springs l, the spring thickness t, the platform length b, the distance of force application, a, above the fixed end of the springs:

$$\theta_{\text{pitch}} = \left(\frac{6(\ell - 2a)t^2}{3b^2\ell - 2t^2\ell + 6at^2} \right) \left(\frac{x}{\ell} \right) \tag{8.6.1}$$

$$\delta_{\text{vertical}} \approx \frac{x^2}{2\ell} \tag{8.6.2}$$

Note the importance of the placement of the applied force. If the force is applied at a point other than halfway between the platforms, a bending moment is generated which causes the pitch angle to occur.

As an example of the effect of manufacturing inaccuracies on flexure accuracy, consider a difference δ_{spring} in the length of a clamped spring flexure's spring and a difference δ_{platform} in the length of platforms. The pitch angles caused by the difference in spring length and difference in platform length are respectively:

$$\theta_{\text{spring}} = \frac{\delta_{\text{spring}} x^2}{2\ell^2 b} \tag{8.6.3}$$

$$\theta_{\text{platform}} = \frac{\delta_{\text{platform}} x}{\ell b} \tag{8.6.4}$$

These equations have been experimentally verified by Jones[69] who gives typical tolerances on the spring and platform length to be 25 μm and 1-3 μm respectively.

Repeatability

Monolithic flexural bearings are probably repeatable to the subangstrom level, but the repeatability obtained will depend on the stress level and any bending hysteresis associated with the particular material used. If the bearing elements are clamped to the structure, then very high repeatability, on the order of angstroms to nanometers, can also be obtained if there is no slip between components.

Resolution

Flexural bearings provide a reaction force proportional to displacement. There is no stiction so resolution is essentially entirely a function of the servo control system.

Preload

Flexural bearings are inherently preloaded.

Stiffness

Unfortunately, one does not get something for nothing, and the more range of motion a flexural bearing provides, the lower the stiffness. However, the stiffness is fairly easy to calculate because the elements are considered continuous without sliding or rolling interfaces.

Vibration and Shock Resistance

Flexural bearings are often used where a high-frequency motion is to be generated. As long as stress levels do not cause failure in a monolithic flexure or microslip in a clamped flexure, they are vibration and shock insensitive.

Damping Capability

Flexural bearings only provide as much damping as the material they are constructed from, so in general it is negligible. However, one can easily bond a vibration-absorbing elastomer to the flexural elements to absorb vibrational energy (see Section 7.4.1). In the case of a clamped flexure, the elastomer should be applied only to the exposed area of the flexure, not between the spring steel and the clamp.

[69] Ibid.

Friction

There is no static or dynamic coefficient of friction associated with flexural bearings. There is a restoring force proportional to the size and geometry of the flexure and the displacement.

Thermal Performance

Flexural bearings are sensitive to thermal growth because the ratio of surface area to volume is very large. A bonded layer of elastomer will not only help increase damping, it will make the flexure less likely to suffer from undesirable thermal effects.

Environmental Sensitivity

Other than thermal effects, flexural bearings are insensitive to environmental effects, such as dirt and slime buildup, as long as corrosion is not induced in the flexural elements.

Seal-ability

No seals are generally required except maybe to help maintain thermal equilibrium of the bearing.

Size and Configuration

There are an unlimited number of sizes and configurations of flexural bearings.

Weight

Because of their simplicity, flexural bearings can be made very lightweight.

Support Equipment

None required.

Maintenance Requirements

None required.

Material Compatibility

For clamped flexural bearings, it is very important to make sure that the material used as the flexural element is compatible with the material it is clamped to.

Required Life

As long as generous radii are used and the fatigue stress is not exceeded, life can be infinite.

Availability

Flexural bearings are easily designed and built. Modular versions are also available.

Designability

It is very easy to design a flexural bearing, but one must be aware of the fine points such as fretting at clamped interfaces, stress concentration, overconstraint, and thermal sensitivity.

Manufacturability

The comments for designability apply here as well.

Cost

Flexural bearings are generally inexpensive.

8.6.2 Design Considerations

Flexural bearings provide maintenance-free, dirt-insensitive limited motion with very high stability and repeatability. In order to design a flexural bearing, the following points need to be considered:

- Stress ratio
- Monolithic or clamped design?
- Thermal effects
- Flexure kinematics

Stress Ratio

When flexing, the maximum stress in the flexure should be as low as possible, and should be at most 10-15% of the yield strength of the material. If the maximum stress is kept low, then the flexure will remain dimensionally stable for longer periods of time. For monolithic designs, the

flexure may be heat treated locally with a laser after it is machined. For clamped designs, hardened tempered spring steel is clamped to the structure and used for the flexures.

Monolithic or Clamped Design?

There are two basic types of flexures, monolithic and clamped. The former entails starting with a solid block of metal and machining away all the material that is not supposed to be there. This leaves a rigid frame that is connected to a rigid platten by very thin sections. With this type of design, there can be little question as to the integrity of the connection between the flexural elements and the frame and platten; hence instability and micro-noise associated with microslip are avoided. Wire EDM machines are quite often used to manufacture monolithic flexural bearings.

Various types of geometries for the regions near monolithic flexures exist, as shown in Figure 8.6.1. The simplest involves cutting a notch equal to the diameter of a milling cutter from each side of the flexure, as shown in Figure 8.6.1a. When the milling cutter is used to cut a notch wider than the mill's width, as shown in Figure 8.6.1b, care must be taken by the machinist so that the cutting forces do not cause the remaining thin membrane to break. If a large membrane is needed, one can machine the slot from one direction, fill the slot with castable material (e.g., epoxy after coating the sides with mold release), and then machine the other side to form the flexure. As shown in Figure 8.6.1c, an island can also be left in the middle and a bolt used to clamp the island during machining. One can also consider other types of machining operations, such as wire EDM followed by chemical deburring. Two degrees of freedom can be obtained with a single flexure when it is shaped like an hourglass, as shown in Figure 8.6.1d.

Clamped flexures offer the most economical way to design flexural systems with large (mm range) motion. However, if the joint between the flexural element and the part is not designed properly, then microslip will occur and dimensional stability will be not be maintained. In order to clamp the flexure properly, the product of the clamping pressure and the coefficient of friction (μ nominally being 0.1) must be greater than the tensile stress that exists at the outermost fibers of the flexural element. By greater is meant by a factor of 3 or so; however, the clamping pressure should be less than one-half to one-third of the yield strength of the materials at the joint. Also, the deeper the hinge element extends into the structure the better. If there is any slippage, then fretting can occur, which creates the potential for eventual structural failure or increased mechanical noise from the joint.

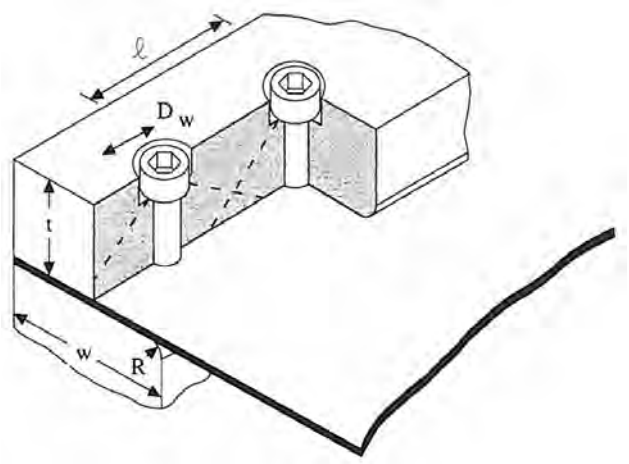


Figure 8.6.3 Parameters for a clamped flexural bearing.

A typical clamped flexural assembly is shown in Figure 8.6.3. A clamping bar squeezes the flexure between it and the base. The clamping bar should be several times thicker than the bolts that are used to clamp it, and the bolt spacing should be such that there is about 30% overlap of the 45° pressure cones emanating from under the bolt heads. With a diameter under the bolt head (with lock washer) of D_w and a clamping bar thickness of t, the bolt spacing should be

$$\ell_{\text{bolt spacing}} = D_w + 5t/3 \tag{8.6.5}$$

The width of the clamping region should allow for the the pressure cone under the bolt head to end about one-third of the way after the edge of the bar to minimize the chance of fretting by there being an unclamped region that may undergo microslip:

$$w_{clamp} = D_w + 4t/3 \tag{8.6.6}$$

Note that the leading edges of both the base and the clamping bar should have generously *radiused*, not chamfered, corners to prevent a stress concentration at this point. Hence the clamping bar should have a width on the order of

$$w_{bar} = D_w + 2t \tag{8.6.7}$$

The bearing tries to pry the bolts up as it flexes; thus one must make sure that in the presence of this stretching of the bolts, sufficient preload exists to keep the clamping pressure on the underside of the flexure a factor of 2 or 3 greater than the bending stress in the flexure. To lessen the prying force on the bolts and at the same time decrease the stress concentration where the bolts extend through the flexural element, there should be two to three bolt diameters between the center of the bolt holes and the edge of the clamping bar. Where space is at a premium, one can use a thin film adhesive to help prevent microslip.

Thermal Effects[70]

Have you ever noticed that in colder climates the animals tend to grow bigger? Have you ever noticed that a spoonful of hot oatmeal cools more quickly than a bowl of hot oatmeal? The larger the body, the better it stores heat and the less it is affected by its surroundings. Unfortunately, flexural elements usually have a very large surface area-to-weight ratio and thus they are especially prone to thermal distortion, so one must be careful to isolate them from heat sources. In some cases, one may wish to insulate the flexure by bonding an elastomer to it, which would also help to increase damping. A viscoelastic layer bonded to the flexure can also help to increase damping, as discussed in Section 7.4.

Flexure Kinematics

The number of different types of flexures is limited only by the imagination. In fact, one fun game to play while you are waiting in the airport for your flight is to try and think up new types of flexural bearing designs.

For simple linear motion, it was shown that for the four bar linkage design of Figure 8.6.2 that there can be parasitic motions. This design is also sensitive to external applied loads. By using symmetry, one can have a more accurate and more robust design as shown in Figure 8.6.4. Note the use of a wobble pin, which is also sometimes referred to as a pivot arm, that preloads the actuator (e.g., leadscrew) and minimizes friction and coupling forces. The moving platform effectively doubles the resolution of the actuation system.

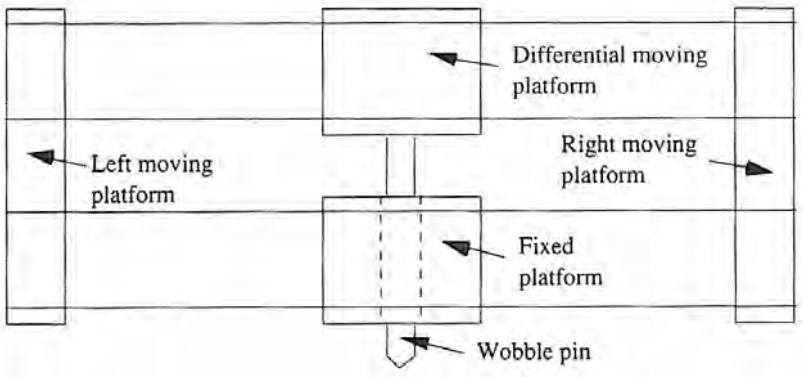


Figure 8.6.4 Symmetrical dual four bar linkage with differential actuation system. Left and right platforms would be connected to a common rigid top-plate.

[70] See Section 8.6.3.

Note that one can also use a differential screw to move a intermediate stage with very fine resolution of motion. The torque from the screw, however, will cause roll motion in the stage. Thus this stage drives the final precision stage via a wobble pin as shown in Figure 8.6.5. The wobble pin can seat in a jewel vee bearing or a cluster of three balls. The former will have better isolation from roll moments, while the latter will form a true kinematic interface.

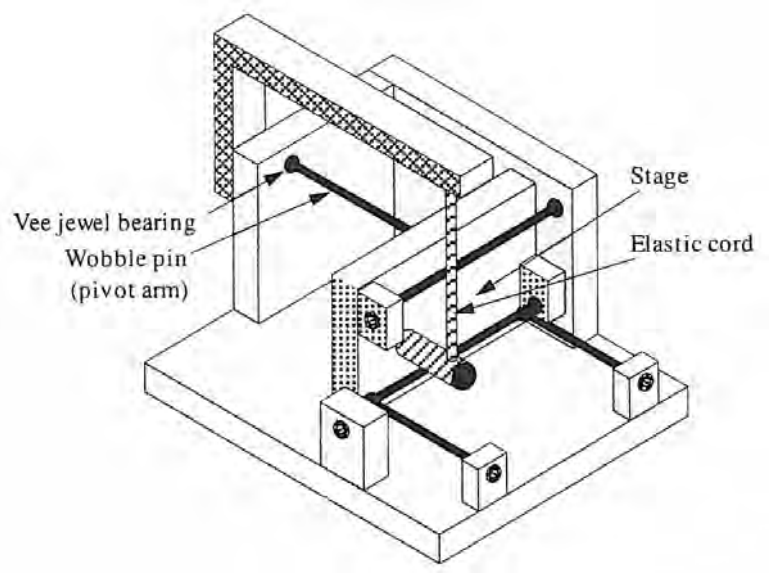


Figure 8.6.5 Use a differential actuation system to increase resolution, and a wobble pin to prevent parasitic force actuation errors.

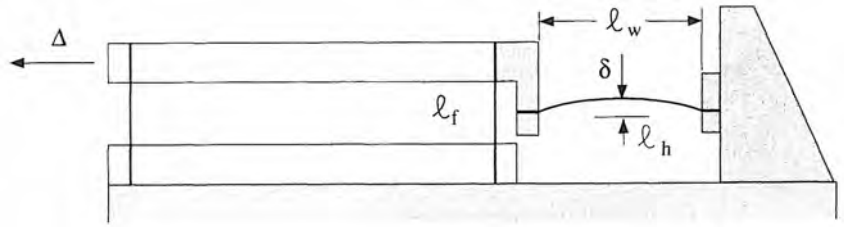


Figure 8.6.6 Bowed flexure used as a high-reduction transmission.

Flexures are often used where very fine motions are required. A flexure can also be used to help generate very fine motion, as shown in Figure 8.6.6. Based on a triangular shaped beam, the downward motion δ causes a lateral motion Δ:

(8.6.8)

For a curved beam, the deflection may be one-half as much, and it is likely that the relation varys with position. Nevertheless, it is possible to chain together a series of these bowed flexures, each at right angles to each other, and obtain a very high transmission ratio. For example, assume that l_h = 2 mm and l_w = 20 mm, then the transmission ratio is 5. In series, the ratios become 25, 125, 625,... for 2, 3, 4,..., units respectively. A variation to this bowed beam approach is to use two beams laid on top of each other and tied together at one end. When the beams bend there will be slip along the interface between them. At the interface, one beam's side will be in tension and the other side will be in compression (see Equation 7.4.5). Using beams in a differential mode is a common flexure design technique. Figure 8.6.7 illustrates a device which provides two angular degrees-of-freedom adjustment for a lens assembly.[71] Note that the center of the lens is placed so there is no net axial motion.

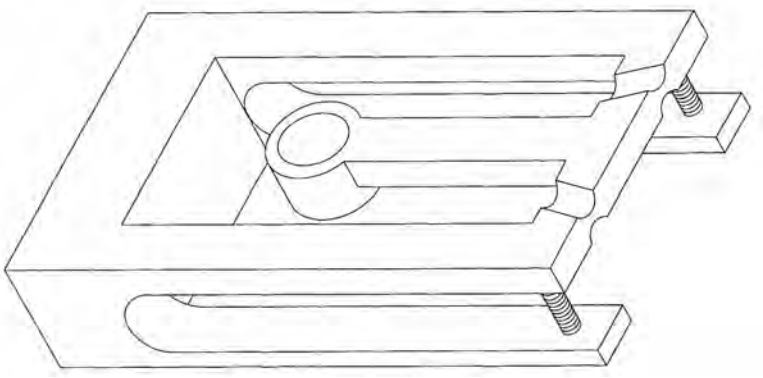


Figure 8.6.7 Device for adjusting pitch and yaw. (Courtesy of Polaroid Corp.)

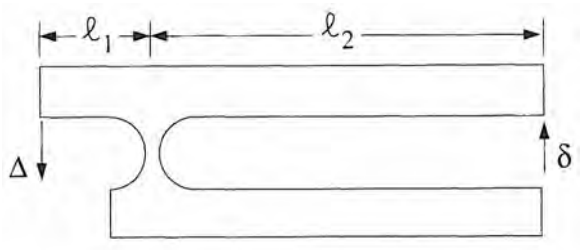


Figure 8.6.8 Lever and fulcrum transmission.

A mechanical advantage can also be obtained with a lever and fulcrum, as shown in Figure 8.6.8. Here the transmission ratio is simply

(8.6.9)

It is not difficult to imagine a monolithic design that uses several such devices in series to attain very high transmission ratios. When combined with a simple high-resolution actuator such as a voice coil or piezoelectric actuator, atomic-level motions are possible for a very low cost. Note that the range of motion will not be very great, so this type of system is used most often as a fine motion stage mounted on top of a coarse motion stage.

Flexures are often used in the design of mechanisms that provide fine adjustment capability for optical components. There are innumerable such devices, many of which are proprietary and buried within complex assemblies, so they will never see the light of day.

To obtain angular motion, several types of flexures are available. A cross-strip flexure, shown in Figure 8.6.9 with equal leg lengths, makes a very effective hinge with a well-defined axis of rotation. Note that the earlier monolithic hinges' axes of rotation are less deterministic. Still, even with a cross-strip flexure, there is a parasitic motion (lateral motion accompanies the rotation) when a force is used to create the torque. A popular modular off-the-shelf-design cross-strip flexure pivot is available from Lucas Aerospace[72] as shown in Figure 8.6.10. Angular motion can also be achieved using a three-spoke monolithic wheel arrangement as shown in Figure 8.6.11. To achieve multiple degrees of freedom, one can stack systems on top of each other or think of other designs, such as that shown in Figure 8.6.12. The spring clip allows the centerpiece to undergo two translational degrees of freedom, but beware of fretting corrosion.

[71] This device was designed by Tony Gonsalves and Bill Plummer of Polaroid Corp.
[72] Lucas Aerospace Power Transmission Corporation, 211 Seward Avenue, P.O. Box 457, Utica, NY 13503-0457.

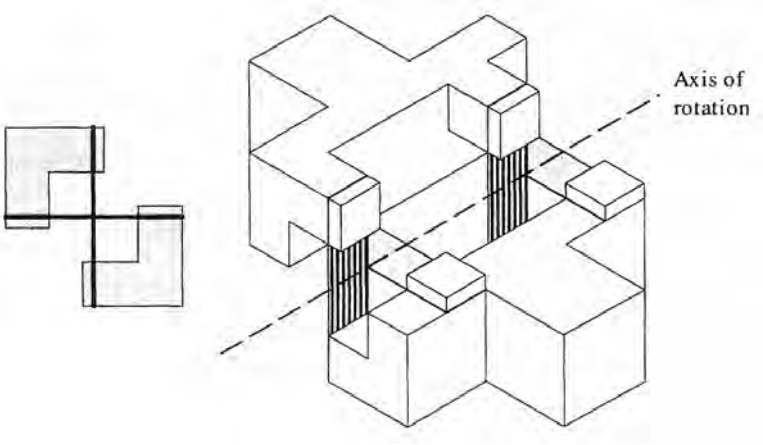
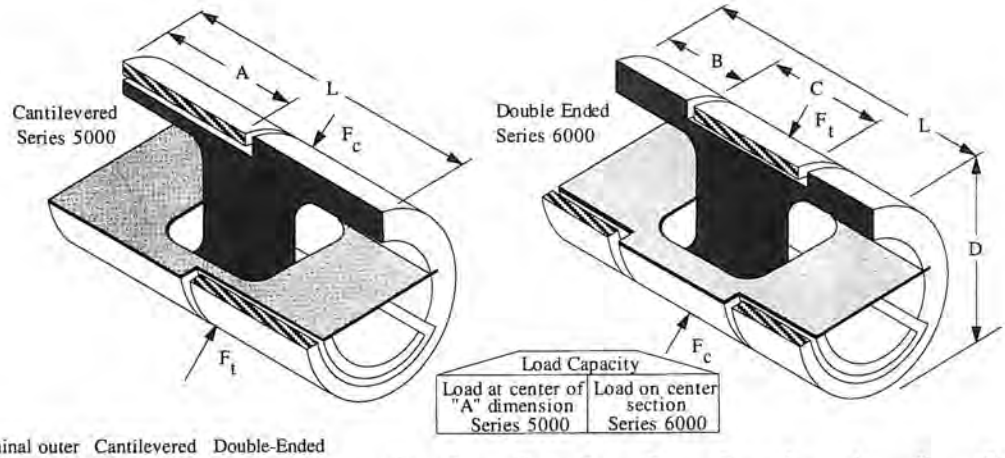


Figure 8.6.9 Cross-strip flexure for angular motion.



Nominal outer diameter (inch)	Cantilevered Series 5000 (Size-Type)	Double-Ended Series 6000 (Size-Type)	K_θ (lbf-in/rad)	Fc (lbf)	Ft (lbf)	Fc (lbf)	Ft (lbf)	D (in.)	L (in.)	A (in.)	B (in.)	C (in.)
1/8	5004-400	6004-400	0.800	25.00	25.0	28.0	28.0	0.1250	0.200	0.095	0.045	0.085
	5004-600	6004-600	0.100	8.80	12.5	17.7	25.0					
	5004-800	6004-800	0.011	0.88	3.5	2.2	4.7					
5/32	5005-400	6005-400	1.600	39.00	39.0	44.0	44.0	0.1562	0.250	0.120	0.057	0.110
	5005-600	6005-600	0.200	13.80	19.5	27.6	39.0					
	5005-800	6005-800	0.025	1.39	5.5	3.5	7.4					
3/16	5006-400	6006-400	2.710	56.0	56.0	63.0	63.0	0.1875	0.300	0.142	0.067	0.130
	5006-600	6006-600	0.326	19.8	28.0	39.6	56.0					
	5006-800	6006-800	0.041	2.1	6.8	4.9	9.0					
1/4	5008-400	6008-400	6.540	100.0	100.0	113.0	113.0	0.2500	0.400	0.190	0.090	0.175
	5008-600	6008-600	0.817	35.4	50.0	70.7	100.0					
	5008-800	6008-800	0.102	3.4	14.0	8.5	19.0					
5/16	5010-400	6010-400	12.800	156.0	156.0	176.0	176.0	0.3125	0.500	0.238	0.112	0.220
	5010-600	6010-600	1.640	55.0	78.0	110.0	156.0					
	5010-800	6010-800	0.204	5.7	21.9	14.0	29.0					
3/8	5012-400	6012-400	22.000	225.0	225.0	253.0	253.0	0.3750	0.600	0.285	0.135	0.265
	5012-600	6012-600	2.750	80.0	113.0	159.0	225.0					
	5012-800	6012-800	0.331	7.9	31.5	19.8	42.0					
1/2	5016-400	6016-400	52.000	400.0	400.0	450.0	450.0	0.5000	0.800	0.380	0.180	0.355
	5016-600	6016-800	6.500	141.0	200.0	283.0	400.0					
	5016-800	6016-800	0.813	14.2	56.3	35.4	75.0					
5/8	5020-400	6020-400	106.00	625.0	625.0	703.0	703.0	0.6250	1.000	0.475	0.225	0.445
	5020-600	6020-600	13.30	221.0	312.0	442.0	625.0					
	5020-800	6020-800	1.69	22.1	87.8	55.0	117.0					
3/4	5024-400	6024-400	182.00	900.0	900.0	1013	1013	0.7500	1.200	0.570	0.270	0.535
	5024-600	6024-600	22.80	318.0	450.0	636	900					
	5024-800	6024-800	2.85	31.1	127.0	78	169					
1	5032-400	6032-400	431.00	1600	1600	1800	1800	1.0000	1.600	0.770	0.370	0.735
	5032-600	6032-600	53.80	566.0	800	1131	1600					
	5032-800	6032-800	6.73	56.6	225	141	300					

Figure 8.6.10 A commercially available flexural pivot. (Courtesy of Lucas Aerospace Power Transmission Corporation.)

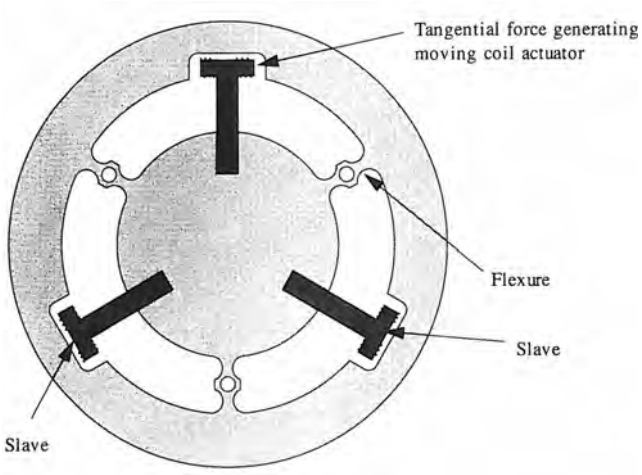


Figure 8.6.11 Three-spoke monolithic wheel for angular motion.

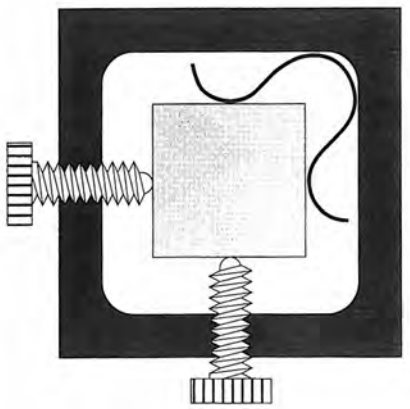


Figure 8.6.12 Two-dimensional flexural bearing system.

8.6.3 Design Case Study: Flexure Thermal Sensitivity[73]

Optical wafer steppers project and align patterns onto photoresist-coated semiconductor wafers to define the processing of each layer during the production of integrated circuits. This process is known as *optical lithography*. Wafer steppers are truly production-oriented precision instruments, used in three-shift operation around the world. All optical wafer steppers utilize three essential systems: the optical system, the alignment system, and a step-and-repeat stage. Reduction optical systems reduce the pattern formed by a chrome-on-glass reticle and project the reduced image onto the wafer. A key element of the optical system of any stepper is resolution. Today's systems can routinely produce submicron images over a 22-mm-diameter field. The resolution is an important factor since it defines the minimum size of the features of the integrated circuit. Smaller features allow for smaller, denser, faster chips.

Resolution alone does not constitute successful IC production. Unless the image aligns with the previous layer, workable circuits cannot be produced. This function is performed by the alignment system. By detecting the location of alignment marks placed on the wafer during previous processing, the stepper can then properly place the new images over the underlying levels. This is performed using an alignment microscope mounted on the system. When the alignment microscope is not mounted on the optical axis of the reduction lens, it is known as an off-axis alignment system. The better the alignment system performs, the smaller the chips can be made, and the yield of the chips improves.

The third element of the optical wafer stepper is the X, Y stage system. Since the lens field is much smaller than the diameter of the wafer, the wafer must be moved from site to site. This step-and-repeat cycle is carried out until the wafer is completely exposed. Stages must be highly accurate and are typically metered by laser interferometers, with resolution as high as 0.01 μm with a total range of 200 mm by 200 mm. The speed of the stages is also critical since it is a major contributor to the overall throughput of the stepper.

Flexures and Their Use in Steppers

Several adjustments in the stepper require extremely precise (submicron) motion over limited ranges. These include the alignment of the reticle, motions to focus the image onto the wafer, and even submicron motions of the stage itself. All of these can be accomplished using flexures. The flexure used for focus motions on the DSW® Wafer Stepper[74] system is a parallel spring flexure. To

[73] This section was written by Stanley W. Stone of GCA, a Unit of General Signal, and edited by A. Slocum. For a more detailed discussion, see S. Stone, "Flexure Thermal Sensitivity and Wafer Stepper Baseline Drift," paper presented at SPIE OPTCON 1988, Precision Instrument Design Section, Nov., 1988.

[74] DSW Wafer Stepper is a trademark of GCA Corp., Andover, MA.

move the column (which contains the lens) up and down to focus the image, a parallel spring flexure is placed between the optical column and the bridge, as shown in Figure 8.6.13. The linear motion of the column must be straight; that is, it must not have any tipping, tilting, or twisting motion. Tipping or tilting causes the image plane of the lens to deviate from that of the wafer, thus causing the image to go out of focus at the edges of the field. Twisting motions would cause the image to rotate and therefore not overlay the previous level. Slight translations in X or Y are allowable in this design, since the mirror positions on the column are monitored by the laser interferometer. The interferometer system measures the differential motion between the stage and the column, thus translations of the column due to foreshortening of the flexure arrangement are corrected.

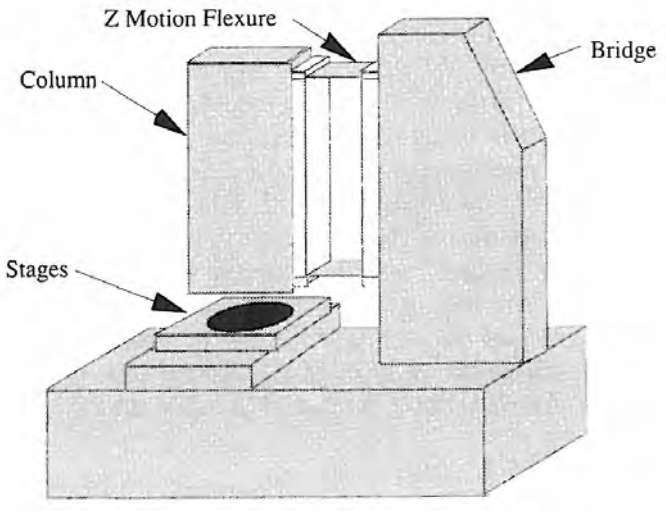


Figure 8.6.13 A parallel spring flexure mounted between the column and the bridge of a wafer stepper. (Courtesy of GCA, a Unit of General Signal.)

To move the column, a voice coil linear motor is used. The voice coil is an excellent actuator for high-precision applications, since it provides a frictionless drive with infinite resolution. In this case, the voice coil drives a lever mechanism that reduces its motion by a factor of 10. To accommodate the weight of the column (so the motor does not carry all of the weight), a counterbalance spring is used. This spring has a low spring rate so it does not load the motor excessively.

Stepper Alignment Systems

In an off-axis alignment system, the microscope and alignment system are mounted away from the optical axis of the reduction lens. For proper system operation, the distance between the alignment microscope and the optical axis must be known precisely. Test procedures determine this distance and record it into the stepper's control system. Once the alignment marks have been located by the alignment system, this distance is subtracted from the current stage position to determine the correct stage position to expose the sites on the wafer.

Two marks on the wafer are used to align the wafer in X, Y, and θ. Two microscope objectives are used, one for each mark. The right objective is used to align X and Y, and the left to align the rotation theta. The distance between the optical centerline of the reduction lens and the X, Y alignment microscope objective is the alignment baseline. The baseline must remain stable during the exposure of a batch of wafers, since drift in the baseline causes misalignment on the wafers, requiring wafers to be reworked and/or lowering the final yield of devices.

Baseline Drift

Baseline drift was discovered on batches of wafers of 100 or more, when a customer was attempting to attain better performance from the system than was specified. After tracking several lots, the drift appeared to be consistent on a particular machine, but variable from machine to machine. At worst, the drift was 0.35 μm. One factor was common to all machines: the drift was more prevalent in the X axis than the Y axis and was accompanied by microscope rotation. The drift

was in the negative X direction and demonstrated a classic exponential response curve. Microscope rotation causes the entire array of exposures to be rotated on the wafer.

There have been several theories about baseline drift. Most were centered around the voice coil motor used to position the X fine stage. The stage steps more in X than Y, and therefore the X motor's temperature is higher as a result. The thermal time constant of the X voice coil motor was measured and found to be about 20 minutes, close to the time constant of the baseline drift. Measurement of the stage expansion as a function of the coil temperature was also measured, but found to have a much longer time constant (approximately 7 hours). Speculation that heat from the voice coil was causing a scale change in the X axis interferometer was also discussed, although measurements did not support this theory. Ultimately, the X voice coil motor-induced drift theories were false since the alignment process aligns the wafer with the system at each wafer, correcting for any mechanical drift in the stage. The alignment system compensates for this at each wafer. This concluded that the source of the baseline drift was definitely in the alignment system.

The Interaction Between Thermal Sensitivity of Flexures and Baseline Drift

Close examination of the Z motion assembly, illustrated in Figure 8.6.14, shows an important clue to the cause of the drift. The voice coil motor is located very close to one edge of the bottom flexure. As the column moves to focus the image, the temperature of this voice coil increases. Heat from this motor is transferred to the edge of the bottom flexure. This particular flexure element is made of a thin spring steel sheet. Because of its geometry, the heat transfer across the element is very poor and a large thermal gradient is generated. This thermal gradient causes a differential expansion across the flexure. This in turn results in an angle between the front and rear plates of the Z motion actuator, which induces a rotation in the optical column. How does this rotation cause the baseline of the system to drift? The key to this question is in the way the alignment microscope is mounted to the column. Since the microscope is mounted to the front of the optical column, it is not on the optical axis of the machine. Because the microscope measures at the point away from the optical axis of the reduction lens (where the measurement *should* take place), it has a great deal of Abbe offset, as shown in Figure 8.6.15.

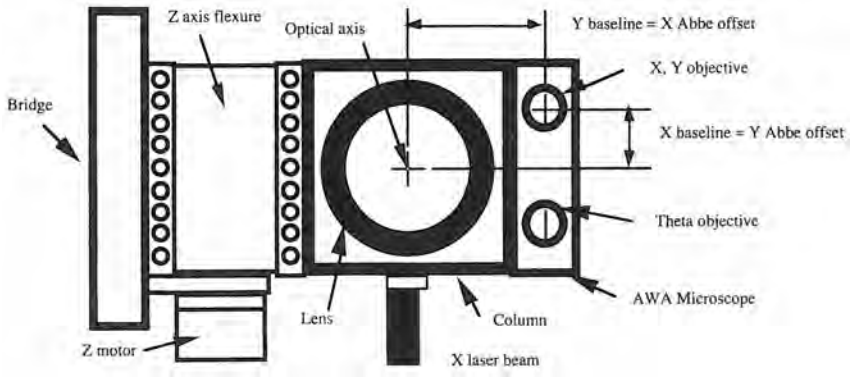


Figure 8.6.14 Top view of optical column and bridge assembly. The automatic wafer alignment microscope is mounted to the front of the column, creating a long baseline in the system, resulting in a large Abbe error. (Courtesy of GCA, a Unit of General Signal.)

Abbe Errors

Abbe error is a common measurement problem. Often it is impossible to move the scale of measurement tool as close to the object as necessary. Any angle introduced in the "parallels" between the scale and the object to be measured results in an error. This Abbe error is a function of the length of the parallels and the angle between them. A simple example is a set of vernier calipers with a loose jaw as shown in Figure 2.1.2. The Abbe offset in the alignment microscope combined with column rotation caused by heating of the lower flexure results in a drift of the alignment microscope position relative to the optical axis. This represents a drift in the baseline.

The rotational sensitivity of the column as a function of the thermal gradient across the flexure is easy to calculate. A simple model shows that when one edge of the flexure is a different

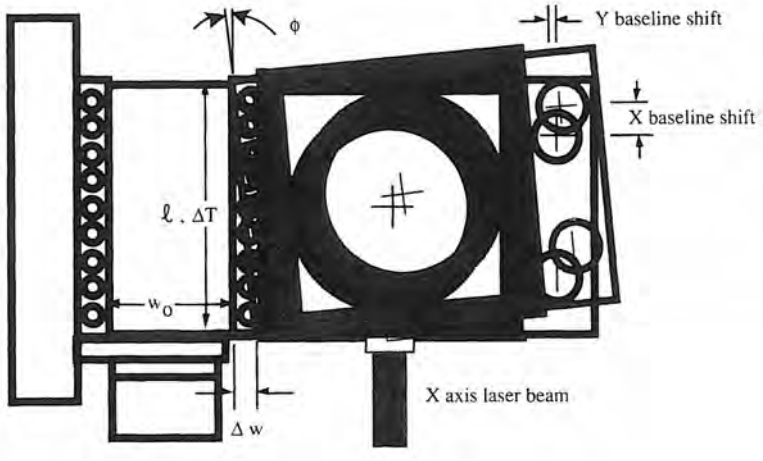


Figure 8.6.15 Flexure thermal distortion leading to rotation and translation of the optical column. (Courtesy of GCA, a Unit of General Signal.)

temperature than the other, the gap between the front and rear plates attached to the flexure will be different at the two edges. The plates will therefore no longer be parallel and some small angle will exist between them:

$$\Phi = \frac{\Delta w}{\ell} \tag{8.6.10}$$

where ϕ is the angle between the plates, Δw is the difference in length between the free edges of the flexure, and l is the length of the flexure as shown in Figure 8.6.14. The change in length of the flexure as a function of temperature difference is

$$\Delta w = w_0 \Delta T \alpha \tag{8.6.11}$$

where w_o is the nominal width of the flexure, ΔT is the temperature difference between the free edges of the flexure, and α is the coefficient of thermal expansion of steel. Substituting Equation 8.6.11 into 8.6.10 yields an expression for the thermal sensitivity:

$$\frac{\Phi}{\Delta T} = \frac{w_0 \alpha}{\ell} \tag{8.6.12}$$

To determine the magnitude of this sensitivity, the values found on the system are substituted into Equation 8.6.12:

$$\frac{\Phi}{\Delta T} = \frac{3.25\,\text{in.}\,(11.7 \times 10^{-6}/\text{C}^\circ)}{6.375\,\text{in.}} = 5.97\,\mu\text{rad/C}^\circ = 1.2/\text{arcsec/C}^\circ$$

To determine the effect this thermal sensitivity has on the baseline stability, this value is multiplied by the Abbe offset. Since the Abbe offset for X is the baseline (and vice versa), the overall sensitivity for the X baseline is

$$\frac{\Delta x}{\Delta T} = \frac{\Phi}{\Delta T}\, y = (5.97\,\mu\text{rad/C}^\circ)(0.111\,\text{m}) = 0.66\,\mu\text{m/C}^\circ \tag{8.6.13}$$

The Y baseline sensitivity is less because the Abbe offset is less. The sensitivity of 0.66 μm/C° is very high since the maximum registration error is 0.35 μm. This means that a temperature change greater than 0.5 C° across the width of the flexure would cause the system to move out of specification. This drift is caused by heat from the voice coil motor differentially expanding the flexure. This rotates the column, causing the microscope to drift due to the Abbe offset. This drift is strictly a function of the temperature difference between the free edges of the flexure, and therefore it exhibits a classic exponential response (being a function of the time to heat the coil and the time to transfer that heat to the flexure). The column appears to rotate around the optical axis, as shown in Figure 8.6.15, because the X, Y interferometer system measures at the optical axis (i.e., it has zero Abbe offset). Translational errors that occurs are therefore compensated.

Solutions to the Problem

Once the problem is understood, the solutions are fairly simple. The first solution attempted was to change some of the materials from aluminum to Delrin®. Delrin® has approximately 1/1000 of the thermal conductivity of aluminum. This solution worked well but could not be installed in the field. A second solution was to actively control the temperature of the motor. This was done by adding a heater to the motor and sensing the motor temperature. By holding the temperature of the motor at a fixed temperature above ambient, with either the coil or the heater providing the necessary power to maintain the temperature, the flexure maintained a fixed temperature differential after warm-up. This solution worked well, except for the complexity and expense of heaters, sensors, and controllers added to the system. Another solution was to make the system more symmetrical. When the motor is placed near the center of the flexure instead of the edge, motor heat does not cause the rotation of the assembly. This solution was used on new instruments, but again it could not be installed in the field.

The last solution was to design a more efficient voice coil motor. This was accomplished using neodenium iron magnets, which have extremely high energy products, and a redesigned coil. By reducing the coil resistance from approximately 12 ohms to less than 2 ohms while maintaining the same force constant, the power dissipated is reduced by a factor of 6. This solution was the best since it was inexpensive and easy to install on existing machines.

Conclusions

There are several lessons in practical machine design that can be learned from this experience. The first is to understand the Abbe principle. This does not mean that all systems must be designed with zero Abbe offset, but one must understand the errors that can be introduced when this is ignored. It *is* possible to design a microscope system with zero Abbe offset that does not pass through the lens. A second lesson is symmetry. The voice coil placed to one side of the flexure assembly caused problems because it heated the flexure unevenly. The problem was not the heat, but its uneven distribution or thermal gradient across the flexure's width. Another lesson is to watch for high thermal sensitivities in the design. Although it was not obvious in this design, it was the high thermal sensitivity of the flexure coupled with the Abbe offset of the alignment microscope that caused the baseline drift.

8.6.4 A Flexural Bearing-Based Adjustable Optical Mount[75]

The widespread use of lasers as integral parts of many engineering projects has created an entire industry devoted to the production of laser peripherals. These peripherals include beam-handling optics anchored to various types of mounts which are designed to ease the installation of a laser-based system. Unfortunately, not all devices are designed according to kinematic principles, with the result that substantial cross coupling exists between adjustment axes. This makes many designs very difficult to use.

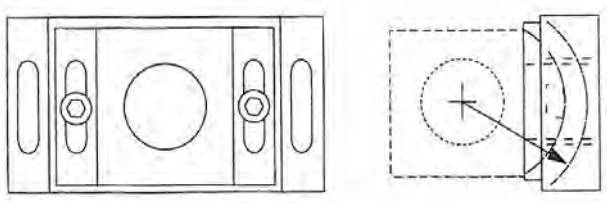


Figure 8.6.16 Typical optic mount with sliding bearings and axis of rotation centered in the mounted component.

A typical mount shown in Figure 8.6.16, consists of a mounting plate whose curved underside rides in the trough of an accompanying baseplate, thus providing several degrees of roll. Yaw and lateral motion are achieved by loosening two screws running from the baseplate to the table and manually swiveling or displacing the entire device. Roll is adjusted through the use of shims. Due

[75] This section was written with the help of one of the author's research assistants, Miles Arnone.

to friction, and roughness of the surfaces of the device, as well as the screws used for adjustment, it is difficult to achieve fine-resolution adjustment and then lock the system in place. In addition, the design used is not very stable and it must be potted in epoxy once it has been adjusted.

As a result of experiencing much dissatisfaction with this type of mount, it was decided to try to design a better one. In so doing, several design parameters were defined:

- The number of pieces should be minimized; the goal was to make a monolithic device which would lead to reduced friction and thereby improve resolution and stability.
- In both size and range it should be similar to current models in order to allow existing components to be mounted on it.
- Pitch rotation was specified as $\pm$ 2.5°.
- The mount should be easy to manufacture so that it can be produced for about $10 to $20. Many existing kinematic five-degree-of-freedom systems cost several hundred dollars.

Once the design parameters were defined, various design ideas were qualitatively explored. The ideas centered around the concept of a monolithic device which would act as a very stiff spring, deflecting to provide the necessary pitch and then returning to its initial state when required. Roll, yaw, and translational motion were not of primary concern because their effect on beam guidance is of secondary importance in many situations (e.g., fold mirrors) compared to roll control. Roll or yaw could be achieved by fastening a mount orthogonally on another mount.

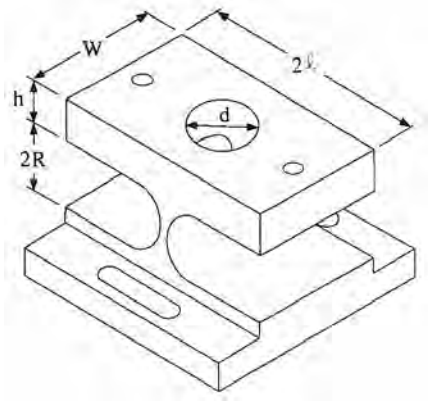


Figure 8.6.17 Proposed flexural bearing-based optical component mount.

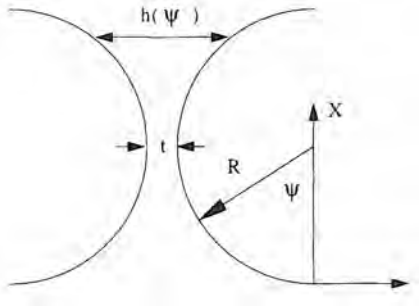


Figure 8.6.18 Flexural bearing parameters.

To meet the low-cost criteria and provide high resolution over a small range of motion, it was obvious that a flexural bearing had to be used. An initial concept is shown in Figure 8.6.17. It was chosen in large part because of the ease with which its behavior can be modeled. It can be modeled as two cantilever beams, one vertical and the other a horizontal platform attached to the vertical beam. A central hole of radius r allows the optical mount to be used in various configurations. When a setscrew is turned on either side of the platform, the horizontal beam bends and the horizontal surface pitches, with the vertical beam acting as a hinge. A conservative first-order model assumed that the device behaves like a part without the hole of width w:

$$w = W - d \tag{8.6.14}$$

The question then remains: how do we find the other dimensions?

Determination of Platform Beam Thickness

The platform beam can be modeled as two cantilever structures of length l, width w, and thickness h. A conservative estimate of the deflection in the top plate caused by the force from the screw is

$$\delta = \frac{4F\ell^3}{Ewh^3} \tag{8.6.15}$$

The allowable deflection is assumed to be some fraction of the required deflection:

$$\delta \approx \ell \sin \gamma\theta \tag{8.6.16}$$

δ gives the deflection of the horizontal beam that corresponds to γ percent of the ideal deflection caused solely by the rotation of the vertical beam through an angle θ.

Determination of R

From a manufacturing point of view it is desirable to make the center web as shown in Figure 8.6.18. The value of R arises from a determination of the characteristics of the support beam, which can be modeled as a series of stacked rectangles with base w and height varying as a function of R and x. The angle of the platform is θ, which is the slope of the vertical support beam at x = 2R and moment M = Fl:

$$\theta = \int_0^L \frac{M\,dx}{EI} \tag{8.6.17}$$

The most convenient way to define the moment of inertia I is in terms of the angle Ψ, not x, and thus the integral will require conversion to polar coordinates. From Figure 8.6.17

$$\int_0^L \frac{F\ell\,dx}{EI\,(x)} = \int_0^\pi \frac{F\ell R\sin\psi d\psi}{EI(\psi)} \tag{8.6.18}$$

$$I(\psi) = \frac{w(t+2R-2R\sin\psi)^3}{12} \tag{8.6.19}$$

Equation 8.6.17 is thus

$$\theta = \frac{3F\ell R}{2Ew}\int_0^\pi \frac{\sin\psi d\psi}{\left(\frac{t}{2}+R-R\sin\psi\right)^3} \tag{8.6.20}$$

This integral is evaluated by parts[76] yielding

$$\theta = \frac{3F\ell R}{2Ew\left[(0.5t+R)^2-R^2\right]}\left\{\frac{1}{(0.5t+R)}+\left[\frac{1}{(0.5t+R)^2-R^2}\right]\times \left[\frac{2R^2+(0.5t+R)^2}{0.5t+R}+\frac{3R(0.5t+R)\left(\frac{\pi}{2}-\tan^{-1}\left(\frac{-R}{\sqrt{(0.5t-R)^2-R^2}}\right)\right)}{\sqrt{(0.5t+R)^2-R^2}}\right]\right\} \tag{8.6.21}$$

Equation 8.6.21 can be solved numerically for R given E, w, l, θ and t, for example by using the Newton-Raphson method or by stepping through possible ranges of values using micron increments. Note that shear deformations were not considered because there is no shear, only a moment, applied to the flexure. The thickness t is found from a maximum stress criteria for bending in the center of the web:

$$t = \sqrt{\frac{6F\ell}{w\sigma_{max}}} \tag{8.6.22}$$

With this value of t, R can be determined.

Unfortunately, in order to obtain an angular rotation of $\pm$ 2.5°, while keeping σ on the order of one-half the yield stress, the device height would have to be over 1.5 in. This is not acceptable in light of the original goal to retain the physical dimensions of existing mount designs as closely as possible. Reflecting on the initial design idea, the initial curvature of the support beam had been chosen for two reasons: (1) stress concentrations would be lower at the support beam to platform beam interface with the semicircular shape than if the support beam had been rectangular, and (2) the part could be made using casting and milling operations, which would make manufacture of the part easier than it would have been with orthogonal surfaces.

As is often the case, the design modification was a compromise between the qualities of manufacturability and physical dimension. The support beam was redefined with a rectangular section of length ℓ_w separating the two regions of curvature. This allowed the overall height of the device to be reduced to less than 1 in. while keeping stress concentrations within acceptable levels. This

[76] See for example, I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series and Products, Academic Press, New York, 1980.

design would probably need to be wire electric discharge machined (EDM). Since the center section is loaded by a moment, determination of the final design's characteristics was achieved by superimposing beam bending analysis for a rectangular section and the analysis of a beam section like that of the original design, although with much smaller R. The device height is thus $2(R + h) + \ell_w$.

Prototype Tests

A prototype aluminum mount was manufactured and then tested by mounting a mirror on it and then reflecting a laser off of the mirror onto a target mirror 10 m away and then back to a viewing target near the mount. For a varying pitch angle the position of the laser beam was then plotted. Initial tests showed that the path that the mount traced out for the beam of light was a highly repeatable jagged spiral rather than a straight line.

Figure 8.6.19 shows the desired versus observed response for vertical displacement as a result of tightening one of the two adjustment screws. This represented a very undesirable cross-coupling characteristic between pitch and yaw. It seems that the device, not being torsionally stiff relative to yaw, was susceptible to twisting when the adjusting screw was torqued. In order to remedy this, the threads for the adjusting screws were moved from the upper platform to the base and cap screws, rather than setscrews, were specified. In addition, Teflon® washers are placed between the screwhead and the platform. This reduced the torque applied to the platform due to the turning of the screw. Torque is applied to the base, which is rigidly affixed to its operating area. This simple design change would have been overlooked but for the simple tests that were performed. Alternatively, a wedge-type actuation device could greatly increase the angular adjustment sensitivity.

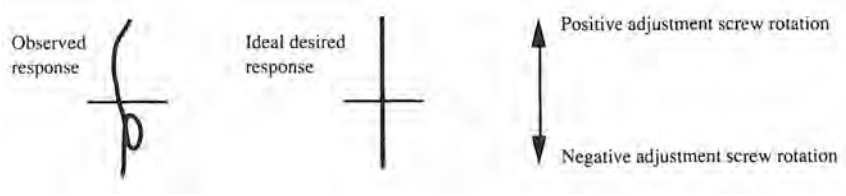


Figure 8.6.19 Path of light beam reflected off the first prototype mount as the mount was moved through its range of motion.

8.6.5 Flexures for Compliant Support of Drive Components[77]

Drive components (e.g., a friction drive) are often mounted on flexures to minimize the effects of forced geometric congruence.[78] In the case of a friction drive and the friction drive bar, the two main components would be mounted on flexures so that the axes of the flexural supports are orthogonal to each other. Figure 8.6.20 shows the element that is often used in place of a plain leaf spring in a parallel linkage flexure. This element allows the stiffness in the X and Z directions to be maximized while minimizing the stiffness in the Y direction. The axial stiffness is given simply by

$$K_x = \frac{wtE}{2\ell} \tag{8.6.23}$$

To analyze this structure, one can imagine that if a moment existed on the middle clamping plate set, then its angle could be set arbitrarily. This is not the case, and thus the system can be modeled as two cantilevered beams whose ends are tied together with the middle clamping plate. The deflection and slope of a cantilevered beam are, respectively,

$$\delta_{bend} = \frac{F\ell^3}{3EI} \quad \delta_{shear} = \frac{F\ell h^2(1+\eta)}{5EI} \quad \alpha = \frac{F\ell^2}{2EI} \tag{8.6.24}$$

The geometric constraint on the system is that the total deflection Δ is the sum of the individual deflections of the cantilevered beams plus the Abbe error caused by the product of the slope and the

[77] The motivation to write this section came from Keith Carlisle of Cranfield Precision Engineering Ltd.
[78] See Section 2.5.

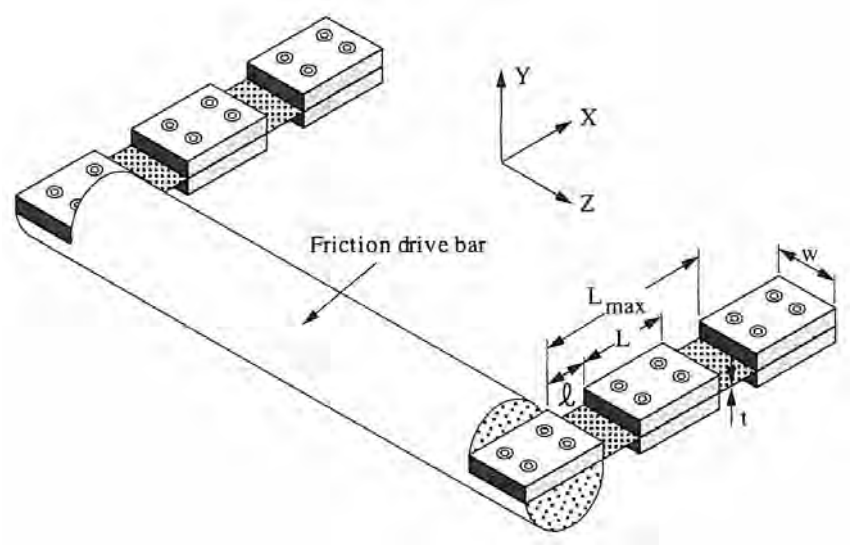


Figure 8.6.20 Parallel flexures with a clamping bar in the middle to increase the ratio of transverse/lateral (K_Z/K_Y) stiffness.

length of the middle clamping plates. The deflection Δ of a pair of these flexures in parallel is

$$\Delta = \frac{F\ell^2}{2EI_{ZZ}}\left[\frac{2\ell}{3}+\frac{L}{2}\right]+\frac{F\ell h^2(1+\eta)}{5EI_{ZZ}} \tag{8.6.25}$$

The stiffness in the Y direction is thus

$$K_Y = \frac{2EI_{ZZ}}{\ell^2\left[\frac{2\ell}{3}+\frac{L}{2}\right]+\frac{2\ell t^2(1+\eta)}{5}} \tag{8.6.26}$$

where $I_{ZZ} = wt^3/12$. For the Z direction the stiffness is

$$K_Z = \frac{2EI_{YY}}{\ell^2\left[\frac{2\ell}{3}+\frac{L}{2}\right]+\frac{2\ell w^2(1+\eta)}{5}} \tag{8.6.27}$$

Note that for the ratio $l/L_{max} = 0.5$, no clamping bar in the middle, it is difficult to achieve high lateral compliance while maintaining high axial stiffness. Also note that the thinner the flexure, the more "efficient" it is. However, the flexure must not be made so thin as to make it susceptible to buckling.

8.7 DESIGN TO LIMIT THERMAL EFFECTS ON BEARING PERFORMANCE[79]

Changes in bearing preload due to thermal growth can affect the kinematic and dynamic performance of a machine. Kinematic errors can be compensated for in control software; however, dynamic errors are not so easily corrected, and depending on the structure, a moderate uniform temperature increase may alter the preload of the bearing and the dynamic performance of the machine. This section will illustrate how to balance thermal expansion of components in an attempt to maintain constant preload on linear or rotary bearings. Note that the analysis technique used for angled contact linear bearing arrangements shown in Figure 8.7.1 can also be applied to most roller bearing configurations (including rotary ones).

8.7.1 Linear Bearing Example[80]

As shown in Figure 8.7.1, it is assumed that various structural materials are used and that the temperatures in the structure and carriage are uniformly T_s and T_c, respectively. They are assumed uniform in their respective structures but may be different from each other. As the system's temperature increases, the bearing gap will close/open due to X direction growth but will open/close due to Y direction growth. For crossed rollers the dimensions e and f would be zero. The transformations between the bearing coordinate systems (i,j) and the XY coordinate system are

$$\Delta_i = \Delta X \cos\theta + \Delta Y \sin\theta \tag{8.7.1}$$

$$\Delta_j = -\Delta X \sin\theta + \Delta Y \cos\theta \tag{8.7.2}$$

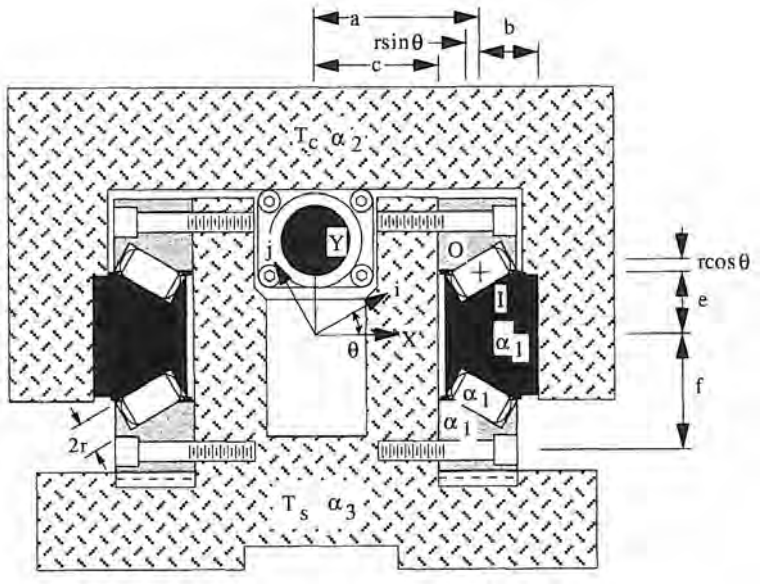


Figure 8.7.1 Linear motion carriage geometry.

The X and Y direction thermally induced displacements of point 0 with respect to the origin of the XY coordinate system are

$$\Delta X_0 = T_s[\alpha_1 (a - c - r\sin\theta) + \alpha_3 c] \tag{8.7.3}$$

$$\Delta Y_0 = T_s[-\alpha_1 (f - e - r\cos\theta) + \alpha_3 f] \tag{8.7.4}$$

The X and Y direction thermally induced displacements of point I with respect to the origin of the XY coordinate system are

$$\Delta X_1 = T_c[\alpha_2 (a + b) - \alpha_1 (b - r\sin\theta)] \tag{8.7.5}$$

[79] *"The gods see what is to come, wise men see what is coming, ordinary men see what is come."* Appolonius

[80] See A. Slocum, "Design to Limit Thermal Effects on Linear Motion Bearing Components," Int. Jou. Mach. Tool Manuf., Vol. 28, No. 2, 1988.

$$\Delta Y_1 = T_c \alpha_1 (e - r\cos\theta) \tag{8.7.6}$$

If the distance between points O and I is to remain constant with respect to uniform temperature changes in the structure, then the differences between the two surfaces' displacements must equal the increase in bearing element thickness (e.g., roller diameter). It is assumed that the rollers' temperature is $(T_s + T_c)/2$

$$\Delta j_0 - \Delta j_1 = r\alpha_1 (T_s + T_c) \tag{8.7.7}$$

With Equations 8.7.2-8.7.7, the critical angle θ is found, at which the bearing surfaces should be oriented with respect to the XY coordinate frame:

$$\theta = \tan^{-1}\left(\frac{T_s[\alpha_1(e-f)+\alpha_3 f] - T_c(\alpha_1 e)}{T_c[\alpha_1 b - \alpha_2(a+b)] - T_s[\alpha_1(c-a)-\alpha_3 c]}\right) \tag{8.7.8}$$

It can also be shown that the angle θ is not a function of the distance j along the length of the roller. In addition, the relative sliding displacement between the two planes, which is found using Equation 8.7.1, is typically only on the order of a few microns/C^o. This small amount of sliding is of little consequence since rollers would establish an equilibrium axial position during motion.

With respect to tapered roller bearings on steel shafts in a steel housing, axial and radial thermal growth will cancel each other out when the bearings are spaced such that the lines of the rollers' center meet at a point between the bearings. This is referred to as a *thermocentric* design.

8.7.2 Thermocentric Angular Contact Bearing Mounting

As discussed in Section 8.4.2, angular contact ball bearings are often used in a back-to-back mounting configuration when the inner race is rotating because this configuration is thermally stable compared to a face-to-face mount. In order to avoid problems with axial spindle growth affecting bearing preload, many spindles are designed with the front set of bearings a duplex, triplex,... set and a duplex pair in the back, which is allowed to float in the bore as shown in Figure 8.7.2.

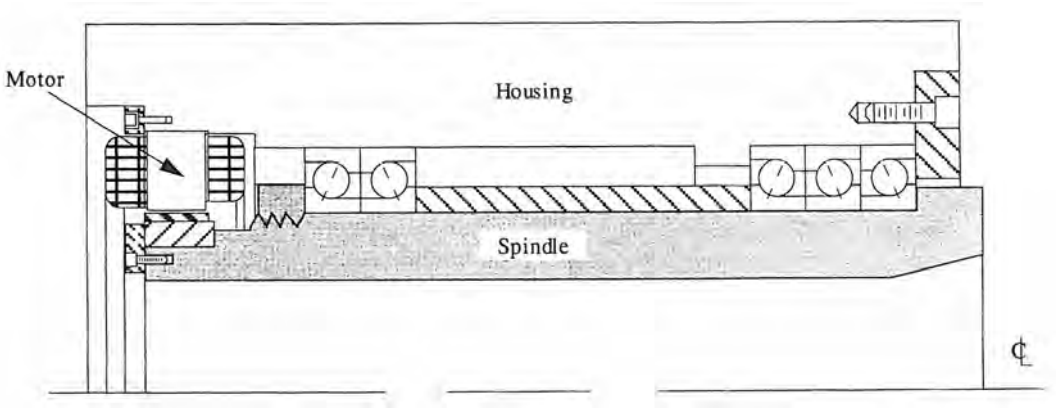


Figure 8.7.2 Back-to-back angular contact bearing spindle design with requirement for rear bearing's outer ring to slide in housing bore (seals and oil mist lubrication system are not shown).

The design shown in Figure 8.7.2 is relatively insensitive to differences in spindle and housing temperature; however, this design has the following disadvantages:

- More bearings are needed.
- Tighter alignment tolerances between the front and rear bores are needed.
- Expansion of the rear bearings' outer rings could prevent them from being allowed to float in the rear bore.
- The rear bearings do not have as good a heat transfer coefficient to the spindle housing as the front bearings; thus the rear bearings can run hotter.

The last point is perhaps the most difficult to deal with because it requires small sliding motions between two surfaces in a constantly changing thermal environment. The spindle is often warmer than the housing because the spindle has a smaller thermal mass than the housing and the bearings have poor thermal conductivity across the rolling elements.

If one were to use a back-to-back arrangement where the front bearings all face one way and the rear bearings all face the other way, then all of the above-mentioned problems would go away; however, could the bearing spacing be designed such that the preload would be constant regardless of the change in spindle and housing temperature?

Consider the angular contact ball bearing design shown in Figure 8.7.3. The centers of curvature of the two circular arc grooves are displaced from each other along X and Y axes by amounts a and b, respectively. The balls ride in circular arc grooves and in order for the balls (spheres) to be tangent to both circular arcs at the same time, the balls' diameters must lie along radii of the circular arcs. Hence the contact angle θ must satisfy b = acotθ. Assume that the groove radii are in proportion to the ball radii: $r_{groove} = \gamma r_b$, then the ball diameter must be

$$r_b = \gamma r_b - \frac{\sqrt{a^2 + b^2}}{2} = \gamma r_b - \frac{a\sqrt{1 + \cot^2\theta}}{2} = \gamma r_b - \frac{a}{2\sin\theta} \tag{8.7.9}$$

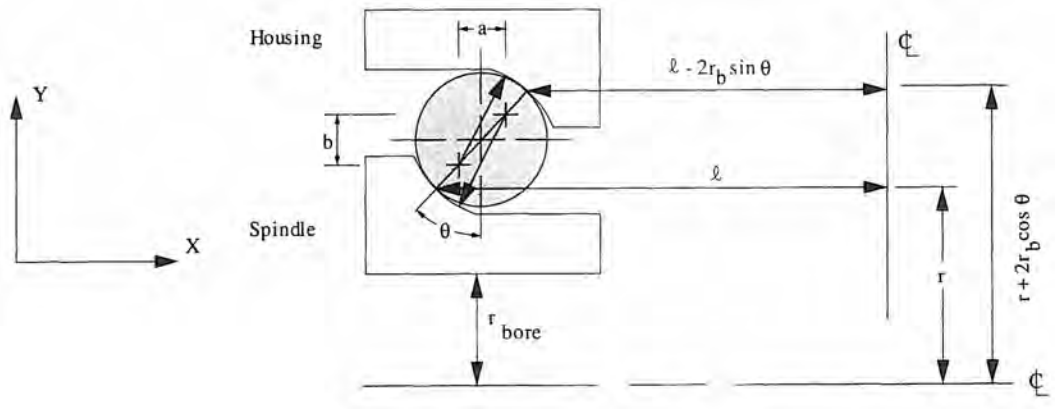


Figure 8.7.3 Geometry definition for back-to-back angular contact bearing spindle design with front and rear bearing sets preloaded against each other.

The offsets a and b can thus be expressed in the following terms:

$$a = 2r_b(\gamma - 1)\sin\theta \tag{8.7.10}$$

$$b = 2r_b(\gamma - 1)\cos\theta \tag{8.7.11}$$

The thermal expansions of the housing and spindle in the X and Y directions are[81]

$$\delta_{ah} = \alpha_h \Delta T_h(\ell - 2r_b \sin\theta) \tag{8.7.12}$$

$$\delta_{bh} = \alpha_h \Delta T_h(r + 2r_b \cos\theta) \tag{8.7.13}$$

$$\delta_{as} = \alpha_s \Delta T_s \ell \tag{8.7.14}$$

$$\delta_{bs} = \alpha_s \Delta T_s r \tag{8.7.15}$$

Note that for a face-to-face mounting, the second term in parentheses in Equation 8.7.12 would be positive. This would also cause the solution for ℓ to be negative, thus indicating that the bearing should be turned around to achieve thermal stability. The change in the dimensions a and b are given by

$$\Delta a = \ell(\alpha_h \Delta T_h - \alpha_s \Delta T_s) - 2r_b \alpha_h \Delta T_h \sin\theta \tag{8.7.16}$$

[81] Note that it is assumed that the bearing rings' coefficient of thermal expansion is nearly equal to that of the housing and spindle, respectively (e.g., steel versus cast iron). The relative dimensions of the rings and spindle make errors resulting from this assumption small. When designing a "real" spindle, one might wish to take these differences into account, based on the relative dimensions of the bearings' inside and outside diameters. Note that it is also assumed that $r_{groove} = \gamma r_b$ always.

$$\Delta b = r(\alpha_h \Delta T_h - \alpha_s \Delta T_s) + 2r_b \alpha_h \Delta T_h \cos\theta \tag{8.7.17}$$

The expressions for the change in the distances a and b (Δa and Δb) added to a and b, respectively, must satisfy the relation given by Equation 8.7.9 with the inclusion of a term for the thermal expansion of the ball due to a change in ball temperature of T_b:

$$2r_b(1 + \alpha_b \Delta T_b - \gamma) = -\sqrt{(a + \Delta a)^2 + (b + \Delta b)^2} \tag{8.7.18}$$

The expansion becomes lengthy, but straightforward. Assume the following:

$$A_1 = 4r_b^2(1 + \alpha_b \Delta T_b - \gamma)^2 \tag{8.7.19a}$$

$$A_2 = 2r_b \sin\theta(\gamma - 1 - \alpha_h \Delta T_h) \tag{8.7.19b}$$

$$A_3 = \alpha_h \Delta T_h - \alpha_s \Delta T_s \tag{8.7.19c}$$

$$A_4 = [2r_b \cos\theta(\gamma - 1 - \alpha_h \Delta T_h) + rA_3]^2 \tag{8.7.19d}$$

The equation for the bearing spacing 2l between the front and rear bearings is thus

$$2\ell = 2\left(\frac{-A_2 + \sqrt{A_1 - A_4}}{A_3}\right) \tag{8.7.20}$$

In the implementation of this analysis, the relation b = a cotθ should be checked with the values of a + Δa and b + Δb. One may have to iterate with the new value of θ until convergence is reached. In doing the analysis, one finds that the ideal spacing varies with temperature by only a few percent but depends on the relative temperature ratio between the bearing components. Unfortunately, without previous spindle experience, prototype tests, or detailed finite element models the relative temperatures cannot be predicted easily. Figure 8.7.4 shows how the spindle spacing 2l varies with temperature ratios K_T for a spindle bearing with a 70-mm ID (r = 45 mm, r_b = 10 mm, γ = 1.1, θ_i= 15° and 25°). The value of 2l given above is for "exact" equilibrium.

It may be acceptable to solve for l given a maximum allowable change in ball diameter (change in preload). Figure 8.7.5 shows representative axial displacement values to achieve various preload and axial stiffness levels for an ABEC 7 angular contact ball bearing with a 70-mm bore. If this allows for a reduction in manufacturing costs and an increase in reliability, then a small change in preload with temperature may be acceptable. Once the spindle was designed using these back-of-the-envelope calculations, a finite element model could be used to see how the net axial displacement changes as a function of different temperature distributions in the spindle. In order to determine what the actual temperature difference between the spindle and housing is likely to be, one could measure existing spindles or attempt to use bearing design programs[82] and/or thermal finite element models. Ultimately, before one goes into production with an untested design, it should be bench-tested. If the spindle is run below about 4000 rpm, in many cases it is possible just to use grease-lubricated bearings and thereby avoid the cost of a temperature controlled oil mist system that would be applied to the spindle bearings.

To help minimize the temperature difference between the spindle and the housing:

- Rough-shot-peen the inside of the housing and the outside of the spindle and make the distance between them large enough to guarantee turbulent airflow. This will increase surface area and turbulence in the air gap between the two bodies; hence the heat transfer coefficient will be increased. One can also buy special coatings which cause the bodies to transmit heat better (so they behave like blackbodies).
- Introduce clean filtered pressurized air at different points in the spindle to housing gap. The expanding air will cool the spindle and also increase flow out of the spindle; however, if one is not careful, the air may blow out the grease in the bearings.

Based on the temperature effects on preload problems discussed here, one can understand the motivation behind the development of the Hydra-Rib® bearing shown in Figure 8.4.19.

[82] For example, ROMAX® software design package for rotating machinery which is available from Engineering Technology Associates of Southfield, MI. In addition, some bearing manufacturers have time-sharing programs that aid the designer in selecting a bearing and predicting its performance in the assembly, for example, The Timken Company's SELECT-A-NALYSIS® bearing selection and analysis program.

Contact angle	$T_{spindle}/T_{housing}$	$T_{bearing}/T_{housing}$	2l(m)
15°	1.5	1.25	0.293
15°	1.25	1.5	1.076
15°	1.25	1.125	0.855
15°	1.125	1.25	2.204
25°	1.5	1.25	0.149
25°	1.25	1.5	0.580
25°	1.25	1.125	0.447
25°	1.125	1.25	1.178

Figure 8.7.4 Minimum spindle bearing spacing 2ℓ as a function of temperature distribution ratios, for 70-mm-ID bearings to prevent spindle preload from increasing as spindle temperature increases. Cast iron spindle αg 11.0 μm/m/C°, bearing steel αg 10.2 μm/m/C°, cast iron housing αg 11.0 μm/m/C°.

Preload N (lbf)	Single bearing preload displacement μm (μin)	Approx stiffness N/μm (lb/μin)
780 (175)	20 (800)	39.0 (0.22)
1890 (425)	33 (1300)	85.4 (0.49)
3450 (775)	43 (1700)	156 (0.89)

Figure 8.7.5 Stiffness and inner ring axial displacement resulting from various preloads of a 70 mm ID ABEC 7 angular contact ball bearing with 25° contact angle and 76,500 N (17,000 lbf) static load capacity.

8.8 Case Study: Measurement of a Spindle's Error Motions[83]

A precision multi-axis grinder workhead spindle was fitted with tapered roller bearings and was driven by a variable-speed electric motor through a 41-inch, 130-tooth timing belt. The motor pulley had 40 teeth, the spindle 80. After the spindle was built, it was desired to determine the characteristics of the error motions. In order to study the error motions of the spindle, capacitance probes are normally used, but in this instance capacitance probes were not available. Since the spindle was to be tested at relatively low speeds, sensitive-direction measurements were made with an air-bearing LVDT equipped with a flat-tipped carbide probe tip indicating a 0.5 in. diameter, grade 5 steel metrology ball attached to the spindle face. The ball was centered to within 50 μin. for each test. The probe/ball interface was lubricated with light mineral oil. Figure 8.8.1 shows the frequency spectra of the background noise measured in the radial and axial directions at zero shaft speed. The peak-to-valley magnitudes were 5-6 μin. The large components at multiplesand fractions of about 1100 Hz associated with the spindle drive motor power supply mask smaller 30-Hz components from fluorescent shop lights.

The spindle speed was measured with an optical tachometer (strobe). The data acquisition rate was adjusted for each test to provide power-of-two data points per shaft revolution, which aided in performing the frequency-domain analysis. The average number of points per revolution was determined as part of the data analysis by counting the number of points between the first and last autocorrelation function maxima and dividing by the total number of maxima minus one.

The radial error motion was measured at three rotational speeds (20, 370, and 700 rpm) and two distances from the faceplate (42 and 172 mm). Axial error motion was measured at all three speeds at the 42 mm position. 256 data points per shaft revolution were collected for 64 revolutions at each test condition. The error motion measurements were averaged over 32 shaft revolutions. The average error motion determined was subtracted, revolution by revolution, from the total error motion to obtain the apparent asynchronous error motion. The steady-state and once-per-revolution components of average error motion were removed with a high-pass FFT filter. Figures 8.8.2 - 8.8.10 show the following types of error plots:

- Average waveform polar plots
- MRS (minimum radial separation) error plots

[83] This Section was written by J. David Cogdell of the The Timken Company, Timken Research, 1835 Dueber Avenue, S.W., Canton, Ohio 44706-2798.

- Asynchronous error motion plots (for 20 revolutions)

Both the plots and the numerical data include the effects of the background mechanical andelectrical noise. The lower half of each figure consists of a frequency spectrum of the average error motion. The analysis technique used to generate average error motion plots suppresses non-integer cycle/revolution frequency components. These sub- and fractional-frequency components, typically associated with a bearing'sfundamental defect frequencies, are available by calculating the power-spectrum of multiple revolutions of either the total or the asynchronous error motion. An example is included with Figure 8.8.10.

Figures 8.8.2, 8.8.3, and 8.8.4 show the axial error motion of the spindle at 20, 370, and 700 rpm, respectively. At the two lower speeds the motion is dominated by a 2-cycles-per-revolution component. The 2-cycles-per-revolution component persists at 700 rpm, but is supplemented by a major component at 15 cycles-per-revolution and minor components at 6 and 12 cycles-per-revolution. The average axial error motion is less than 10 μin. at 20 and 370 rpm, and 25 μin. at 700. The asynchronous axial error motion is 23-35 μin.

The radial error motion at the 42-mm position is shown in Figures 8.8.5 - 8.8.7 for spindle speeds of 20, 370, and 700 rpm, respectively. In addition to several high-frequency, low-amplitude components, there are substantial 2, 3, 4, 5, and 6-cycles-per-revolution components at 20 rpm. At both 370 and 700 rpm, the higher-frequency components have largely disappeared. At the higher speeds, the average error motion is dominated by a 6-cycles-per-revolution component (with indications of separate 6-cycles-per-revolution components from both bearings). The average radial error motion ranges from 6 to 15 μin., and the asynchronous radial error motion ranges from 15 to 28 μin.

Figures 8.8.8, 8.8.9, and 8.8.10 show the radial error motion at the 172-mm position. As shown in Figure 8.8.8, at 20 rpm the motion is dominated by a 2-cycles-per-revolution component, with as before, several small higher-frequency components. As shown in Figures 8.8.9 and 8.8.10, at both 370 and 700 rpm, the average error motion consists primarily of 2- and 6-cycles-per-revolution components. The average radial error motion varies from 10 to 13 μin., and the asynchronous error motion varies from 28 to 74 μin.

Figure 8.8.10 includes a frequency spectrum of sixteen shaft revolutions of asynchronous radial error motion (yielding an angular resolution of one-sixteenth revolution). The fundamental defect frequencies for the two bearings are:

	Rear bearing	Nose bearing
f_{roller} (cycles/rev)	4.95	5.21
f_{cup} (cycles/rev)	9.94	11.82
f_{cone} (cycles/rev)	12.06	14.18

These values do not match the measured major frequencies of 0.06, 0.88, 1.00, 1.44, 1.81,2.19, 5.63, and 7.44 cycle/rev particularly well. Since the drive belt frequency is 0.61 Hz and the motorpulley frequency is 2.00 Hz, there are several resonances possible between the fundamental frequencies. Figure 8.8.11 summarizes the spindle error motions measured.

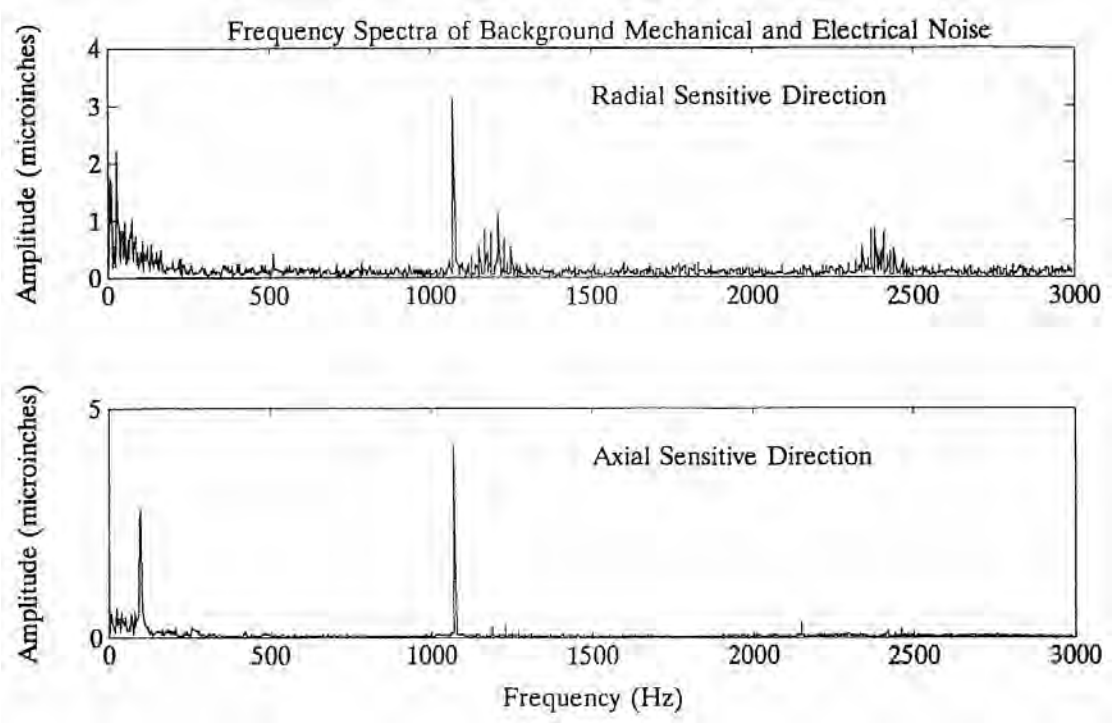


Figure 8.8.1 Background mechanical and electrical noise present during spindle error motion measurements. (Courtesy of the Timken Company.)

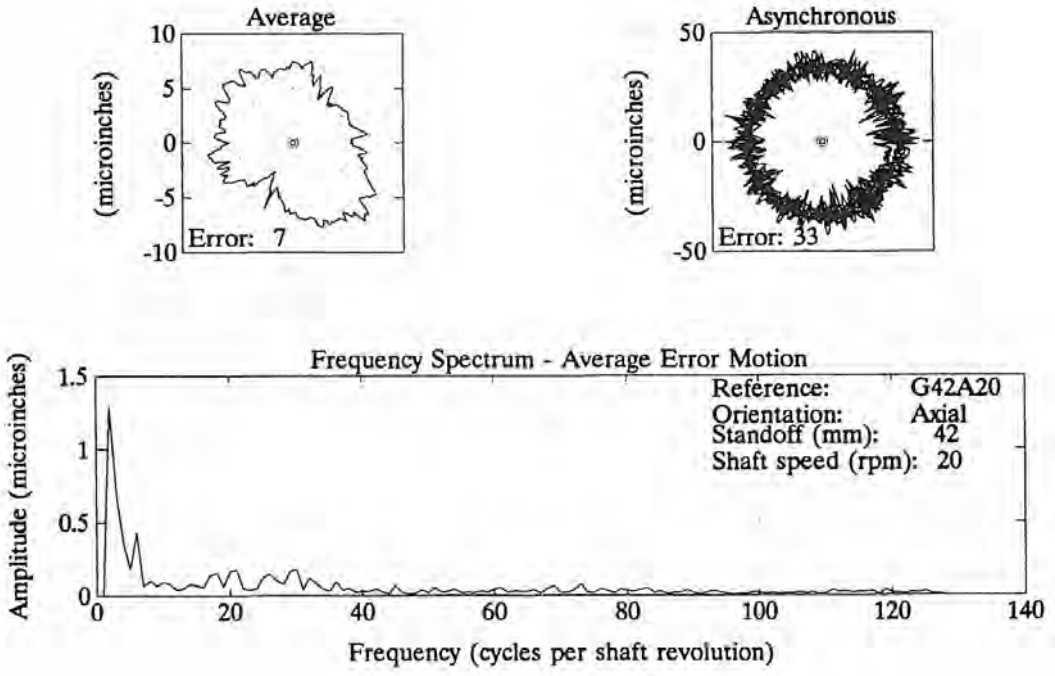


Figure 8.8.2 Axial error motion at 20 rpm. (Courtesy of the Timken Company.)

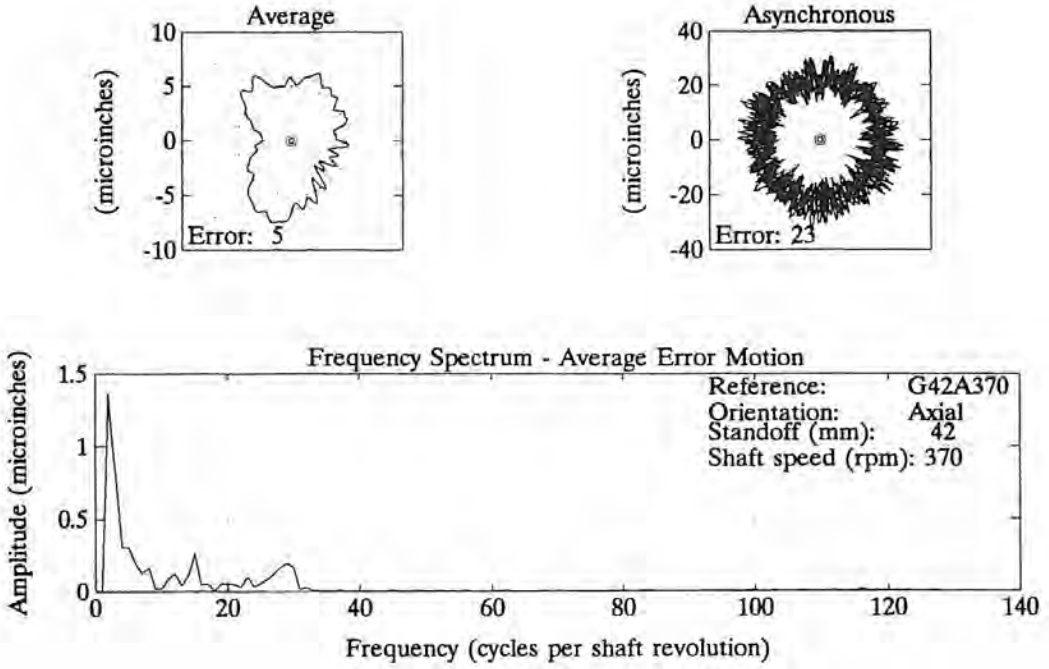


Figure 8.8.3 Axial error motion at 370 rpm. (Courtesy of the Timken Company.)

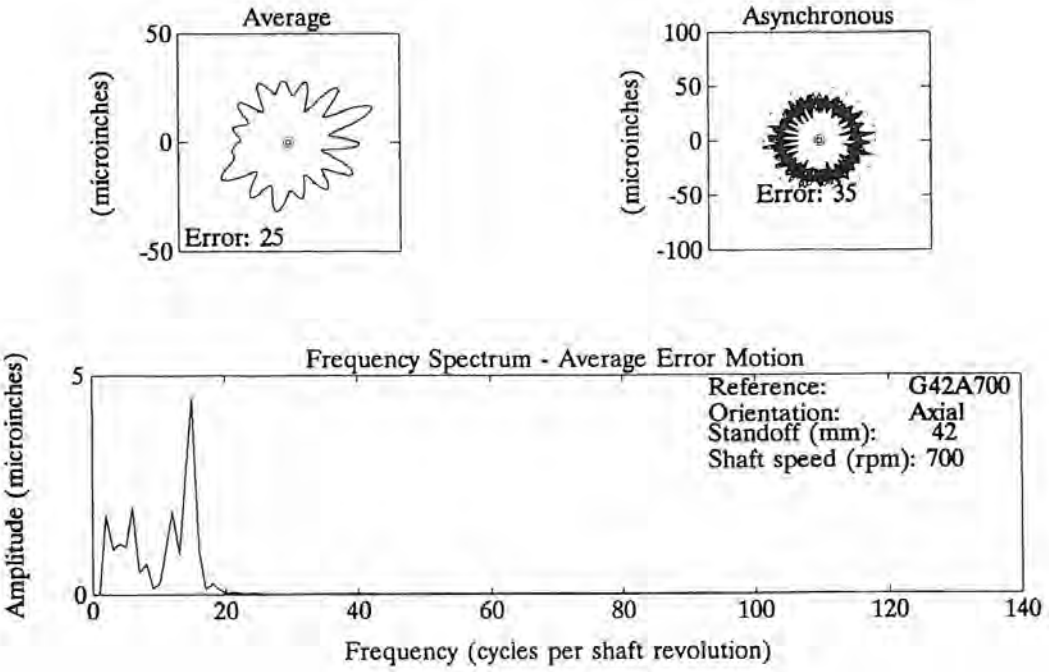


Figure 8.8.4 Axial error motion at 700 rpm. (Courtesy of the Timken Company.)

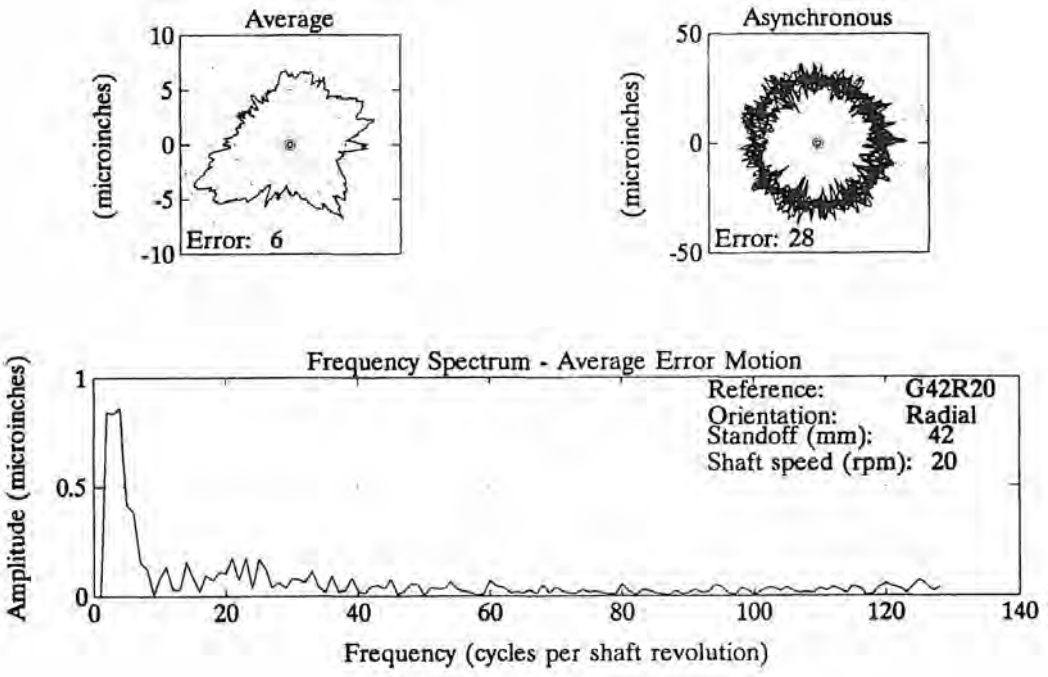


Figure 8.8.5 Radial error motion 42 mm from the spindle faceplate at 20 rpm. (Courtesy of the Timken Company.)

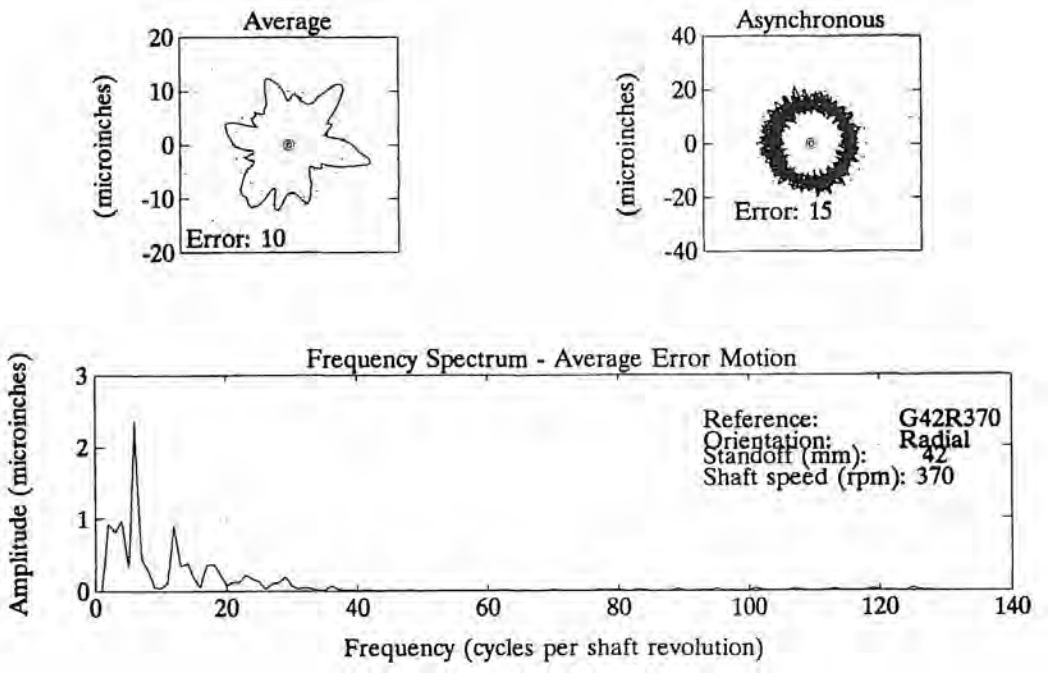


Figure 8.8.6 Radial error motion 42 mm from the spindle faceplate at 370 rpm. (Courtesy of the Timken Company.)

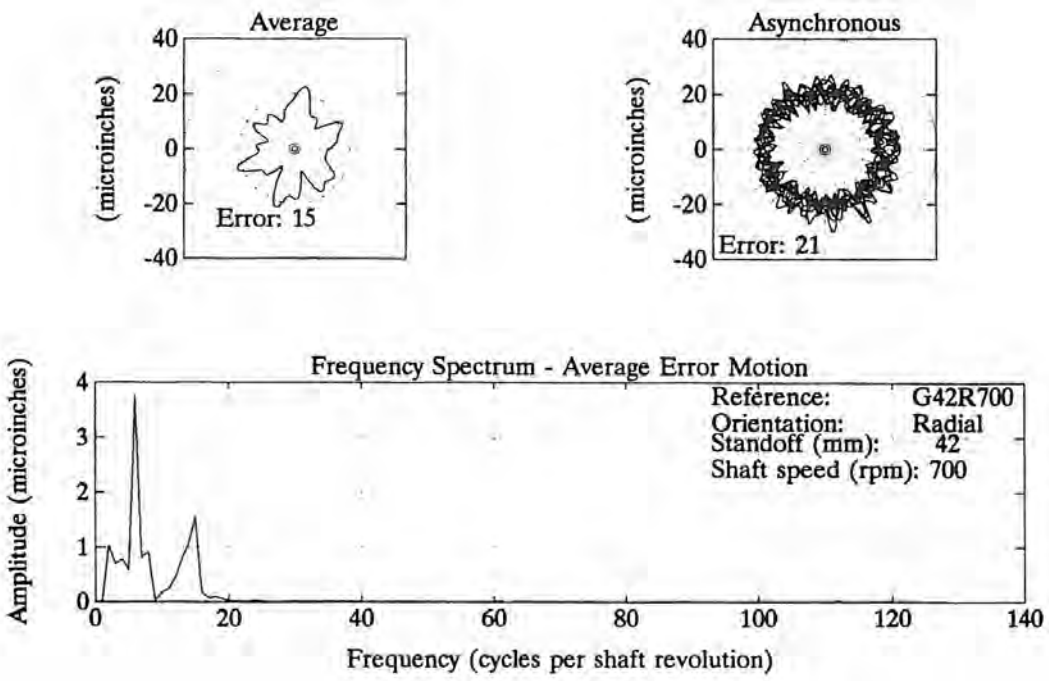


Figure 8.8.7 Radial error motion 42 mm from the spindle faceplate at 700 rpm. (Courtesy of the Timken Company.)

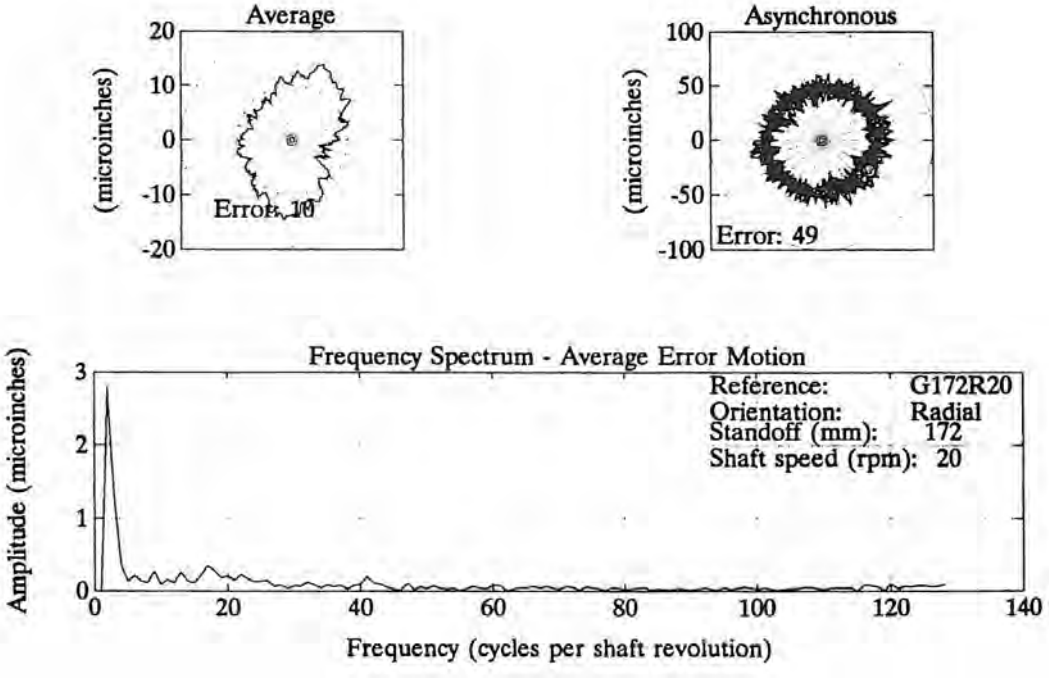


Figure 8.8.8 Radial error motion 172 mm from the spindle faceplate at 20 rpm. (Courtesy of the Timken Company.)

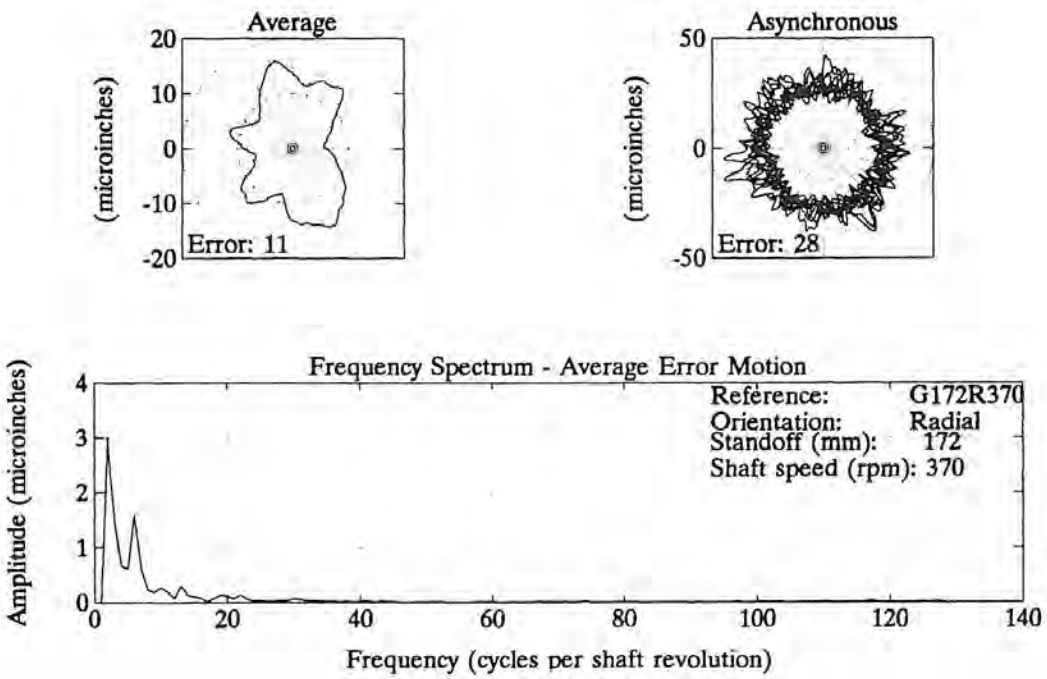


Figure 8.8.9 Radial error motion; 172 mm from the spindle faceplate at 370 rpm. (Courtesy of the Timken Company.)

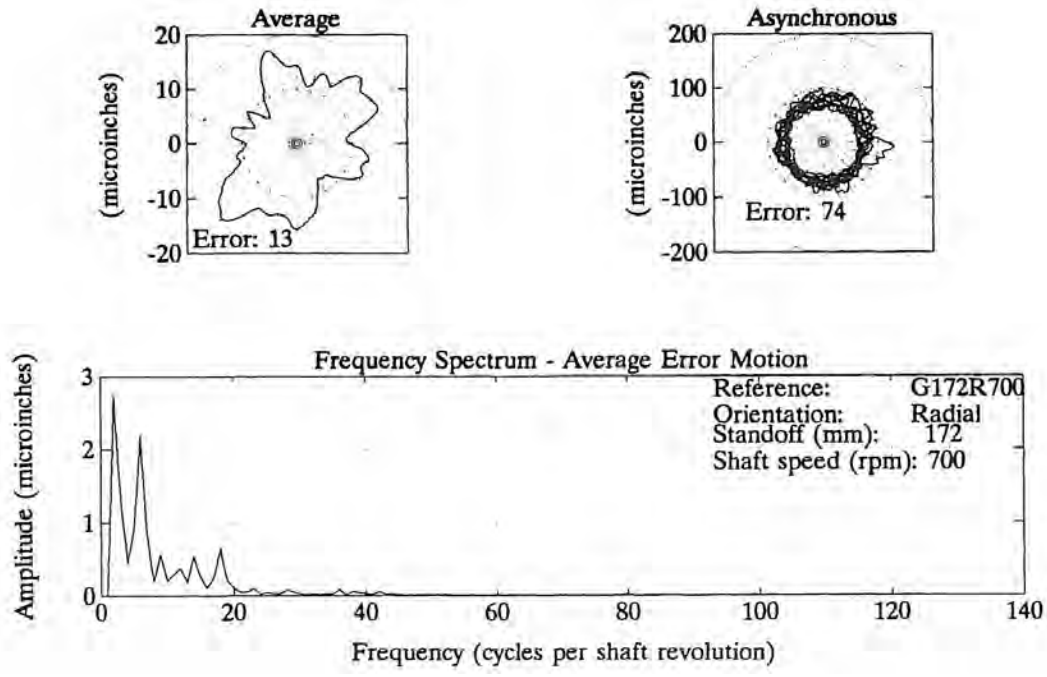


Figure 8.8.10 Radial error motion; 172 mm from the spindle faceplate at 700 rpm. (Courtesy of the Timken Company.)

Spindle speed	Average error motion			Asynchronous error motion		
	Axial error motion (μin.)	Radial error motion (μin.) Distance from faceplate 42 mm	172 mm	Axial error motion (μin.)	Radial error motion (μin.) Distance from faceplate 42 mm	172 mm
20 rpm	7	6	10	33	28	49
370 rpm	5	10	11	23	15	28
700 rpm	25	15	13	35	21	74

Figure 8.8.11 Summary of spindle error motions. (Courtesy of the Timken Company.)

Chapter 9
Bearings without Mechanical Contact between Elements

Without this playing with fantasy no creative work has ever yet come to birth. The debt we owe to the play of imagination is incalcuable.

Carl Gustav Jung

9.1 INTRODUCTION

One of the primary problems associated with bearings with mechanical contact between elements, with the exception of flexural bearings and properly designed sliding element bearings, is that the errors in form and surface finish of the elements can have a significant effect on error motions of the bearing. This is particularly true of higher frequency errors which can be very difficult to eliminate. Bearings without mechanical contact between elements often have error motion frequency spectrums that have well-defined peaks at intervals of the rotation frequency. When error motions are not caused by the motor or coupling, these peaks often decay with increasing speed. This is illustrated in Figure 9.1.1, which should be compared to Figure 8.3.6.

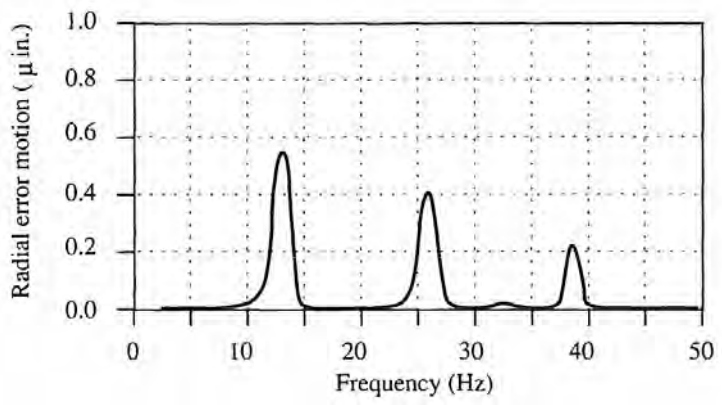


Figure 9.1.1 Radial error motion frequency spectrum of an air bearing spindle built in the 1970s and still being used in production. (Courtesy of Polaroid Corp. and Professional Instruments Co.)

For this reason, when accurate and quiet motion must be attained, a bearing without mechanical contact between its elements is often chosen. Bearings without mechanical contact between elements that are discussed in this chapter include hydrostatic, aerostatic, and magnetic bearings.

9.2 HYDROSTATIC BEARINGS[1]

Hydrostatic bearings utilize a thin film of high-pressure oil to support a load. In general, bearings gaps can be rather large, on the order of 1-100 μm. There are five basic types of hydrostatic bearings: single pad, opposed pad, journal, rotary thrust, and conical journal/thrust bearings, as shown in Figure 9.2.1. All operate on the principle of supporting a load on a thin film of high-pressure oil that flows continuously out of the bearing; hence a method is needed for supplying the pressurized oil and collecting and recirculating the oil that flows out of the bearing.

[1] Four good references for hydrostatic bearing design are W. Rowe and J. O'Donoghue, "A Review of Hydrostatic Bearing Design," Institution of Mechanical Engineers, London, 1972; W. Rowe, Hydrostatic and Hybrid Bearing Design, Butterworth, London, 1983; F. Stansfield, Hydrostatic Bearings for Machine Tools, Machinery Publishing Co., Ltd. London, 1970; and D. Fuller, Theory and Practice of Lubrication for Engineers, 2nd ed., John Wiley & Sons, New York, 1984.

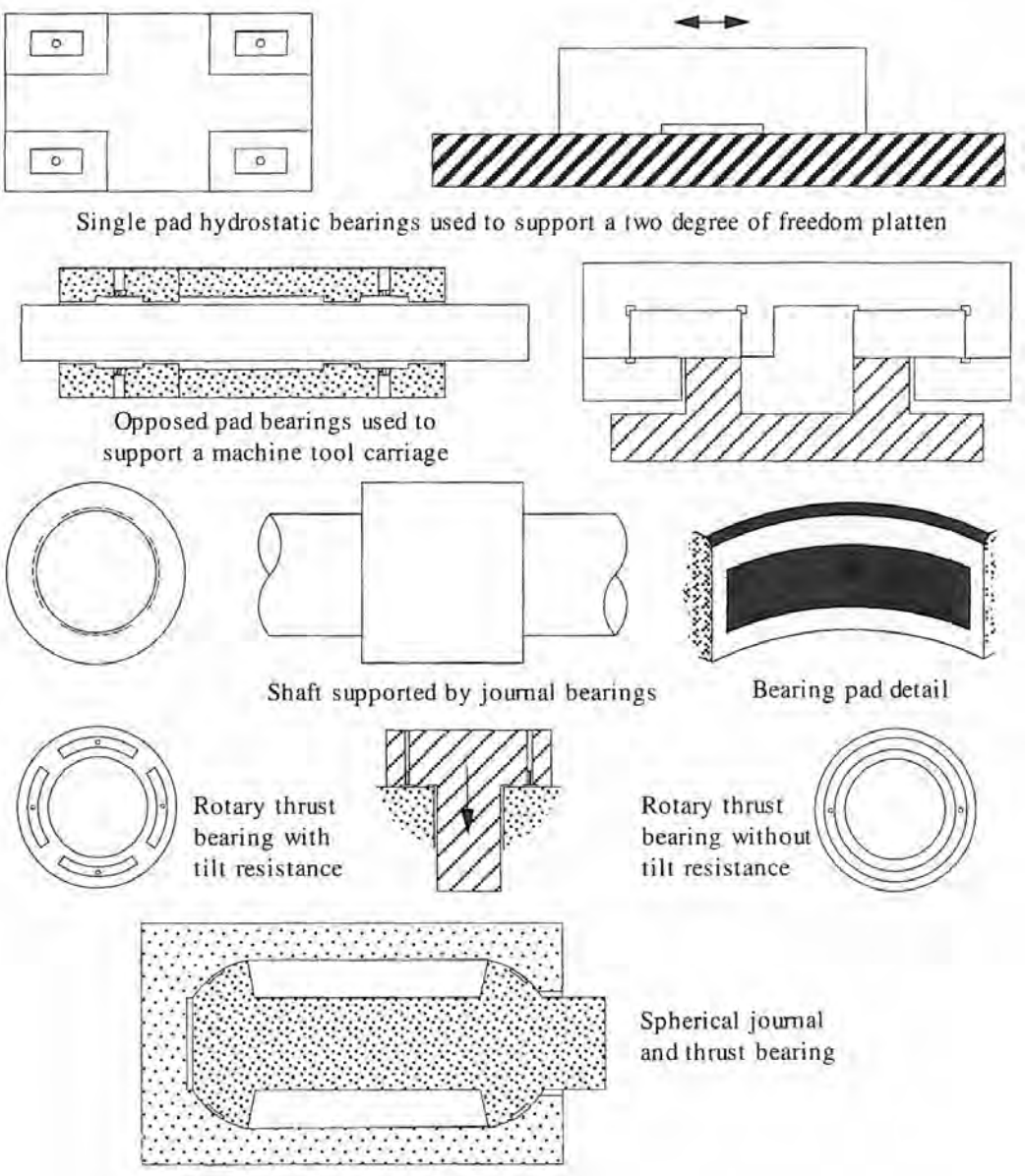


Figure 9.2.1 Various hydrostatic bearing configurations.

The inner region or pocket serves three main functions. First, when the bearing is at rest with no pressure applied, the gap is zero. When the bearing is pressurized, the pocket provides an area for the pressurized fluid to act over to provide the initial lift force. Second, the pocket acts as a region of constant high pressure that allows the bearing to carry more load. Third, the pocket helps to distribute the fluid to all parts of the bearing in noncircular or nonsquare pads (e.g., quadrants of a circle).

9.2.1 General Properties

Speed and Acceleration Limits

Hydrostatic bearings have only viscous friction associated with a fluid film layer being sheared during motion of the bearing. At high speeds (e.g., in spindles) the heat generated can be considerable and may require closed-loop temperature control of the oil. In addition, the flow rate out of the bearing should be high enough to ensure that the flow rate of oil dragged into the bearing by viscous shear is not greater than the flow rate that would normally be exiting the bearing. This could leave one side of the bearing "dry" and result in a loss of stiffness. For rotary bearings, the oil that exits one pad still clings to the rotating member and is pulled into the next pad. Hydrostatic rotary bearings operated at high speeds can experience hydrodynamic effects if the lands are too wide, and they can generate considerable heat. Hydrodynamic bearings suffer from oil whirl instability. This causes vibration at a frequency equal to half the rotating speed. Hydrostatic bearings are much stiffer than plain hydrodynamic bearings, and thus the former generally do not have whirl instability problems if the land widths and pocket depth are properly chosen.[2]

Acceleration is not limited by the bearing design; however, if the speed and acceleration are high and the oil very viscous, a hydrodynamic wedge will form which will increase the pitch error in a linear motion bearing.

Range of Motion

Linear motion hydrostatic bearings can have as long a range of motion as it is possible to machine the slideways they rest on. Slideways tens of meters long have been built. Angular motion hydrostatic bearings are not rotation limited.

Applied Loads

Because hydrostatic bearings distribute the load over a large area, huge loads can be supported. For example, machine tools with multiton carriages often use hydrostatic bearings, and offshore oil platform decks, which may weigh 20,000 tons, are transferred from the fabrication yard to a barge using firehoses to supply water to hydrostatic bearings on the deck's feet. When hydrostatic bearing pads are used in an opposed mode to preload each other, very high bidirectional stiffness can be obtained.

Accuracy

Overall accuracy of motion (e.g., straightness) of a hydrostatic bearing depends on the accuracy of the components. Hydrostatic bearings are insensitive to small random irregularities in the ways and pads which make them perhaps the smoothest running of all bearings. Although some averaging occurs if the bearings are long with respect to the bearing rail straightness errors, if the rails are warped by improper bolt tightening as discussed in Section 7.5, the hydrostatic bearing supported carriage's straightness will correlate very repeatably to the warped way shape. Peak-to-valley surface finish of hydrostatic bearing components should be less than one-fourth of the bearing clearance. There is no wear-in period associated with hydrostatic bearings, and accuracy therefore depends on keeping the fluid flow restrictors clean and the pressure source free from pulsations.[3] Hydrostatic linear motion bearings have been built with submicron/meter accuracy.

With hydrostatic bearings, the accuracy to which a carriage can be axially servoed depends entirely on the actuator, sensor, and controller. Because they provide high normal and tangential

[2] See, for example, P. Allail, "Design of Journal Bearings for High Speed Machines," in Fundamentals of the Design of Fluid Film Bearings, ASME, New York.

[3] See, for example, T. Viersma, Analysis, Synthesis and Design of Hydraulic Servosystems and Pipelines, Elsevier SciencePublishers, Amsterdam, 1980.

damping, hydrostatic bearings are the ideal bearing to use to support a precision linear carriage or a rotary axis that is not moving at high speeds.

Repeatability

Repeatability depends on the stability of the fluid supply system, including the pump and the devices that regulate the flow of oil into the bearings (i.e., flow restrictors). As long as the bearing design is symmetrical, then a flow surge in the line should affect all bearing pads equally and the carriage motion should not be affected. Hence it is important to keep fluid flow supply lines of equal length and to try and filter out as much noise from the pump as possible. With these precautions, low-speed hydrostatic bearings have been built with submicron repeatability. There is no fundamental reason why nanometer repeatability could not be achieved if thermal effects were also controlled.

One must be careful, however, about the viscous behavior of the fluid film in the bearing gap. During a high-speed move, a viscous oil will more quickly build up a hydrodynamic layer than a light spindle oil. This layer or wedge can cause the bearing gap to increase. When the machine comes to its desired position or slows appreciably, the hydrodynamic layer subsides and the bearing gap changes. This problem can be avoided by using a low-viscosity fluid (e.g., a light spindle oil such as ISO 5 or ISO 10) and a small bearing gap (e.g., on the order of 5-10 μm).

Resolution

Since hydrostatic bearings have zero static friction, motion resolution of an object supported by them is virtually unlimited. This makes the design of the actuator and control systems comparatively easy, especially since hydrostatic bearings are so well damped. For example, hydrostatic bearings are used to support large telescopes that often must make very slow sweeps of the night sky and thus require very smooth high-resolution motion.

Preload

Hydrostatic bearings need to be preloaded in order to give them bidirectional stiffness. If a single hydrostatic bearing is used to support a carriage, then as the carriage accelerates or forces act to increase the bearing gap, the bearing will have very little stiffness unless a hardening flow restrictor (e.g., a diaphragm type) is used. Therefore, an opposed pad configuration is the most common one used in precision machine tools. *Only opposed pad systems will be considered here*.

Stiffness

Hydrostatic bearing stiffness can easily be in the newton per nanometer range. Hydrostatic bearing stiffness is not difficult to calculate, and hydrostatic bearings do not have loss of contact problems that sliding or rolling contact bearings that are preloaded against each other can have. Designing with hydrostatic bearings is more deterministic than designing with most other types of bearings.

Vibration and Shock Resistance

Hydrostatic bearings provide better vibration and shock resistance than any other type of bearing. The viscous oil film in the bearing gap makes an excellent energy absorber.

Damping Capability

The thin oil film in the bearing gap gives hydrostatic bearings excellent damping capabilities in both normal, via squeeze film damping, and tangential, via viscous shear, bearing directions. Estimation of the squeeze film damping coefficient for use in dynamic models of the machine can be difficult.[4] Since most dynamic machine models use finite element models, it is recommended that a fluid finite element program be used to determine the damping factor for the pad geometry and machine configuration used. Note, however, that squeeze film damping depends on a change in the bearing gap to dissipate energy by viscous flow, and thus if the bearing is much stiffer than the machine structure, very little effective damping will occur. Hence the need for a balanced design, which is one of the continuing sagas of this book.

Tangential damping is very easy to estimate (see Equation 9.2.34) and is independent of the bearing stiffness, although a stiff bearing often has a small gap which is also indicative of high tangential damping. Because of high tangential damping, controlling the motion of a hydrostatic bearing supported carriage is truly a joy.

[4] See, for example, W. Rowe, Hydrostatic and Hybrid Bearing Design, Butterworth, London, 1983, pp. 199–204.

Friction

Hydrostatic bearings have absolutely zero static friction. The amount of dynamic friction force generated is proportional to the velocity of the carriage, bearing dimensions, and oil viscosity. The dynamic friction force on a hydrostatic bearing is independent of the loads applied insofar as they do not change the bearing gap.

Thermal Performance

Energy is input into a hydrostatic bearing in the form of a flow at a pressure. The oil oozes out of the bearing and into a drip pan. In the pan its flow rate and pressure are essentially zero, so all the power that is represented by the initial flow and pressure is expended in the viscous shear the fluid undergoes as it oozes out of the bearing. This power is dissipated as heat, where power (watts) = flow (m^3/s) $\times$ pressure (N/m^2). The temperature rise of the oil depends on how much heat is conducted by the machine. Also, very viscous oils become hot when pumped, so unless it is cooled, the oil may arrive at the bearing hot. It is possible to precool the oil, so the net effect is that the machine maintains a desirable temperature (e.g., 20^oC). One must be careful to consider this effect, which gives motivation for using as low a pressure and flow rate as possible. In general, hydrostatic bearings are not used where speeds greater than about 2 m/s are encountered because viscous shear of the fluid in the bearing gap also generates too much heat.

Environmental Sensitivity

Because oil is always flowing out of the bearing, hydrostatic bearings are self-cleaning. The oil must be collected, filtered, and reused, so it is important to keep chips and cutting fluid out of the bearing area. This requires the use of bellows or sliding way covers, or grooves connected to a suction pump, which surround the bearing pads.

Seal-ability

Hydrostatic bearings generally ride on square or dovetail rails, so it is not difficult to seal them and collect the oil using a suction pump if so desired. Rotary motion hydrostatic bearings are easily sealed using rotary seals. Provisions must be made for collecting the oil and not allowing it to collect in stagnant pools, such as in a bellow's folds.

Size and Configuration

Hydrostatic bearings take up very little space themselves, but the plumbing requirements may be significant. It is generally desirable to have only one hose coming to the carriage, so the carriage itself may look like a block of Swiss cheese after all the drilling is done in order to get the oil to all the different bearings pads. However, it is imperative that the "block of cheese" be thoroughly internally deburred and cleaned to prevent the fluid flow restrictors from becoming clogged. Generally, hydrostatic bearings are used in opposed pad configurations. Kinematic designs are not always necessary because opposed pad bearings act like constant force springs that fill whatever gap exists. As the equilibrium gap changes, the stiffness changes, but stiffness will always be finite; hence hydrostatic bearings are essentially completely forgiving of rail and carriage misalignments that could cause loss of preload with a sliding or rolling bearing.

Weight

Because of their simplicity, hydrostatic bearings have very high performance-to-weight ratios, but only if one excludes the size and weight of the pump, oil collection and distribution system, and oil temperature control system.

Support Equipment

The biggest drawback of hydrostatic bearings is they require a pump, distribution, collection, and filtering system, and often a means to temperature control the oil. The system must be kept extremely clean to prevent foreign contamination form clogging the oil flow regulation devices which typically are small orifices or slots. The oil is usually filtered using a filter with particle passing size four times smaller than the gap or flow restrictor size.

Maintenance Requirements

Oil level and cleanliness must be monitored and filters on the pump changed according to a fixed maintenance schedule. The system should be inspected periodically for signs of contamination and the bearing rails for signs of wear should a bearing pad's flow restrictor become clogged and

the pad starved for oil. Oil quality should also be monitored to make sure that its pH level remains within desirable limits and that the oil does not become contaminated with bacteria.

Material Compatibility

Hydrostatic bearings are compatible with virtually all materials, and the presence of a small bearing gap usually leaves ample room for differential thermal expansion between components; however, one needs to determine the order of this gap change and make sure that it does not alter the bearing's performance too much. If the gap becomes too small, then the flow restrictors can saturate the bearing pressure from both sides of opposed pads, with the resultant loss of stiffness and load-carrying capability. If the gap opens up too much, then the bearing will be starved for oil and a loss of stiffness will also occur. It is a good idea to choose bearing materials such that if a loss of pressure occurs, the bearing can coast to a stop without being damaged.

Required Life

Hydrostatic bearings whose oil supply systems are properly maintained can have essentially infinite life. There are hydrostatic bearings that have been around for decades without ever having been rebuilt. Often the only reason to take such machines out of service is that other parts of the machine need to be rebuilt.

Availability

Hydrostatic bearing spindles are available as off-the-shelf items. Linear hydrostatic bearings are generally custom designed and manufactured. By the time one considers all the support systems and special plumbing required, the design and manufacture of a hydrostatic bearing is not a trivial task.

Designability

It is relatively easy to design a hydrostatic bearing-supported system; however, there are many design details that must be considered, such as those associated with oil distribution, collection, and temperature control.[5]

Manufacturability

In order to prevent a plethora of hoses from choking a design, it is a good idea to have oil passages machined integral with the bearing structure. In this way only one hose needs to be connected. The structure laced with holes needs to be carefully internally deburred and cleaned. Holes which need to be plugged should be done so with press-fit plugs[6] unless access may be required at a later date for system cleaning. Pipe thread plugs should never be used; instead, use straight thread plugs with O ring seals. An even better option is to use flange-type couplings so that the fluid never flows over a threaded region.

Cost

The primary costs associated with hydrostatic bearings are those of the fluid supply system, the cost of machining all the oil supply holes, and the cost of machining long straight rails or very round bores.

Summary

At low speeds (<2 m/s) hydrostatic bearings provide the best performance characteristics (friction, accuracy, stiffness, and load capacity) of any bearing, with the exception perhaps of aerostatic bearings in some applications. Hydrostatic bearings' principal drawback is the cost associated with the fluid supply system and the mess associated with having to collect the oil. It is for this reason that one does not see them in widespread use in machine tools. In high-precision machines and machines that require a high number of accurate cycles (e.g., grinders), they are more commonplace. Following sections describes how hydrostatic bearings can be designed.

[5] There are software packages available to aid with the design of systems supported by hydrostatic and hydrodynamic bearings. For example, programs are available from Rotor Bearing Technology & Software Inc. (Conshohocken, PA) to analyze rotating machinery dynamics, bearings, and seals.

[6] Tapped plugs can be used but are expensive if one considers the tapping operation. A simple press-fit plug called a *Lee plug* is available from the Lee Company (Westbrook, CT, and various international offices).

9.2.2 Analysis of Opposed Pad Linear Motion Hydrostatic Bearings

Virtually any number of pads can be combined in virtually any configuration; therefore, it is important to develop a feel for the theory so that one will be able to handle one's own special cases. Figure 9.2.2 shows the electrical circuit analogy for an opposed pad hydrostatic bearing. In electrical systems E = IR, and in fluid systems P = QR where P is pressure, Q is flow, and R is fluid resistance. It is fairly simple to provide a pressure source (analogous to a voltage source) for a hydrostatic bearing system; however, a method is needed to regulate the flow, or else there will be no pressure difference across the bearing and it will not be able to support a load. Flow restrictors of resistance R at the entrance to each bearing can provide a means for regulating the flow.[7] As a force is applied to the system, the bearing gaps will change and so will the upper and lower resistances. A simple circuit analysis shows that the difference in pressure between the upper and lower pads of the bearing is

$$\Delta P = P_u - P_\ell = P_s \left(\frac{R_u}{R + R_u} - \frac{R_\ell}{R + R_\ell} \right) \tag{9.2.1}$$

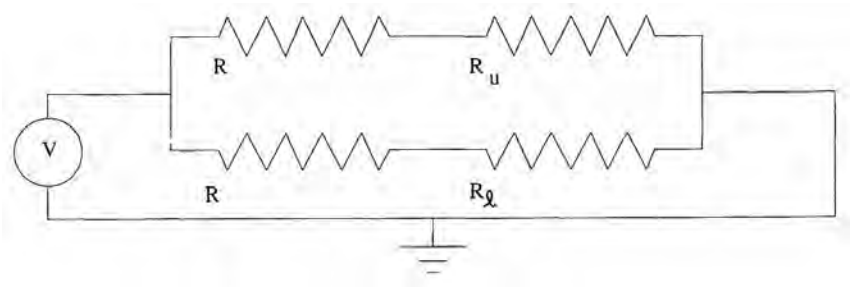


Figure 9.2.2 Electrical circuit analogy for opposed pad hydrostatic bearings.

For a nominal gap h and small excursions δ of the structure supported by the bearings, the upper and lower gap resistances vary as[8]

$$R_u = \frac{\gamma}{(h - \delta)^3} \quad R_\ell = \frac{\gamma}{(h + \delta)^3} \tag{9.2.2}$$

where the constant of proportionality γ is to be determined later. The resulting difference in pressure across the bearing is thus

$$\Delta P = P_s \gamma \left[\frac{1}{R(h - \delta)^3 + \gamma} - \frac{1}{R(h + \delta)^3 + \gamma} \right] \tag{9.2.3}$$

If the inlet flow resistance R was zero, the bearing could support no load. Also, if the inlet flow resistance was infinite, the bearing could support no load. Thus there must be some ideal inlet resistance (compensation) between these two extremes. Taking the partial derivative of the pressure difference with respect to the inlet flow resistance and ignoring all terms with δ^2 and higher terms[9] yields

$$\frac{\partial \Delta P}{\partial R} = P_s \gamma h^2 \left[\frac{-(h - 3\delta)}{\left[Rh^2(h - 3\delta) + \gamma \right]^2} - \frac{(h + 3\delta)}{\left[Rh^2(h + 3\delta) + \gamma \right]^2} \right] \tag{9.2.4}$$

Setting this result to zero and ignoring terms with δ^2 and higher, one finds that the "optimal" inlet flow resistance to maximize load capacity is

$$R = \frac{\gamma}{h^3} \tag{9.2.5}$$

---

[7] Hardening restrictors (e.g., a diaphragm type) can also be used, as discussed later. Fixed resistances are commonly used and provide a means of introducing the analysis methods for hydrostatic bearings.

[8] In the next section it will be shown that the upper and lower resistances $R_u$and R_l,respectively, are inversely proportional to the cube of the gaps.

[9] For a high-precision machine, a large change in gap would result in unacceptable motion of machine members anyway; hence $\delta << h$ and this assumption is justified.

Thus for maximum load support capability, the inlet low restrictor's resistance should be equal to the nominal resistance of the bearing pad. Substituting Equation 9.2.5 into 9.2.3, the pressure difference across the bearing is found:

$$\Delta P = P_s h^3 \left(\frac{1}{(h-\delta)^3 + h^3} - \frac{1}{(h+\delta)^3 + h^3} \right) \tag{9.2.6}$$

If the displacement of the bearing is assumed to be a portion of the nominal gap, $\delta = \alpha h$:

$$\Delta P = P_s \left(\frac{1}{(1-\alpha)^3 + 1} - \frac{1}{(1+\alpha)^3 + 1} \right) \tag{9.2.7}$$

Linearizing Equation 9.2.7 about $\alpha = 0$ gives

$$\Delta P = P_u - P_\ell \approx \frac{P_s}{2-3\alpha} - \frac{P_s}{2+3\alpha} \approx \frac{3P_s}{2}\alpha \tag{9.2.8}$$

Figure 9.2.3 shows the correction factor resulting from this linearization. Equation 9.2.8 must be multiplied by this correction factor to get the same value as given by Equation 9.2.7. At 50% gap closure, the correction factor is 0.88, and one should not load a hydrostatic bearing so that the gap closes more than 50%. Most precision hydrostatic bearings are designed so that $\alpha < 0.1$, and then the factor is 0.996, so the linearization is reasonably accurate. Note that these calculations for the pressure assume that the source is not flow limited. If it is, then the effective supply pressure may decrease and this expression would not be valid. One must always check that the pressure source can provide sufficient flow. For an opposed pad bearing with supply pressure P_s and inlet restrictor resistance R, the total flow is just $Q = P_s/R$.

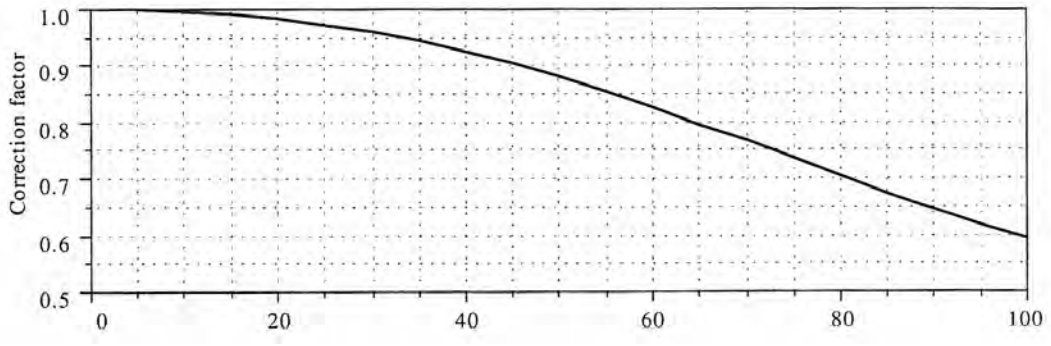


Figure 9.2.3 Effect of linearization on accuracy of Equation 9.2.8.

Bearing stiffness is the change in load for a given change in bearing gap or $A\partial\Delta P/\partial\delta$, where A is the effective bearing area.

$$K = A\frac{\partial \Delta P}{\partial \delta} = 3P_s A h^3 \left[\frac{(h-\delta)^2}{[(h-\delta)^3 + h^3]^2} + \frac{(h+\delta)^2}{[(h+\delta)^3 + h^3]^2} \right] \tag{9.2.9}$$

For initial estimation purposes, it can be assumed[10] that A is a portion of the bearing pad of width a and length b, where A = a(3b - a)/4. This expression cannot be readily simplified to a form where insight can be obtained as to the optimal flow restrictor resistance for maximum stiffness. However, one can deduce that it should be the same as that for maximum load-carrying potential. Linearizing Equation 9.2.9, the bearing stiffness can be approximated by

$$K \approx \frac{3P_s A}{2h} \tag{9.2.10}$$

For example, if P = 2MPa (290 psi), a = b = 0.05 m (2 in.), A = 0.001250 m^2 (1.937 $in.^2$) and h = 10 μm (0.0004 in.), then K = 375 N/μm (2.1 lb/μin.) which is a very stiff bearing.

[10] See Equation 9.2.33.

The load the bearing can support is just $F = K\delta$, where $\delta = \alpha h$:

$$F \approx \frac{3P_s A\alpha}{2} \tag{9.2.11}$$

With $\alpha = 0.5$ and a correction factor from Figure 9.2.3 of 0.88, the bearing of the previous example would thus be able to support a load on the order of 1650 N (371 lbf). This is a good estimate which is based on a bearing with a central pocket and a land width equal to 25% of the bearing width.

Pad Flow Resistance

As shown in Figure 9.2.1, hydrostatic bearing pads often have a rectangular shape with a central pocket. A reasonable model of the flow out of the bearing is to assume that it is composed of parallel rectangular and circular plate regions as shown in Figure 9.2.4. The four corner quadrants together act like a circular annulus.

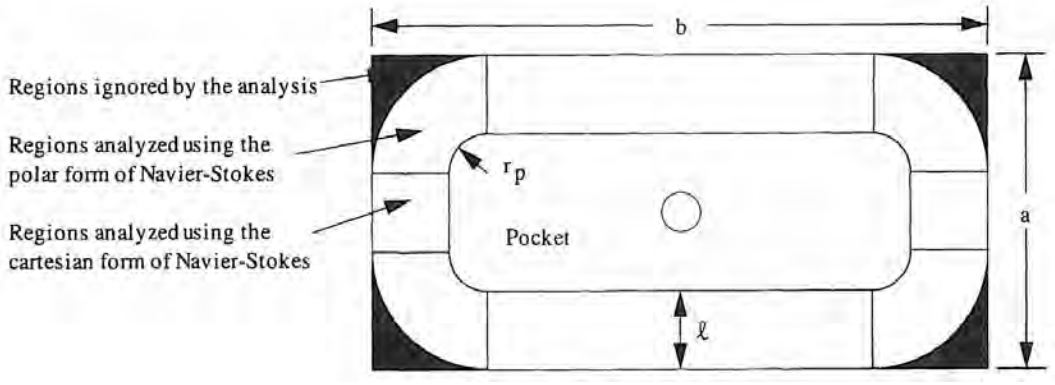


Figure 9.2.4 Flat pocketed rectangular pad geometry.

For the rectangular plate region, steady-state laminar (nonturbulent) flow exists, as shown in Figure 9.2.5. Since the ratio of bearing gap to land length is generally very small, it is often acceptable to assume that the velocity profile is fully developed, and thus the Navier-Stokes equations reduce to

$$\frac{1}{\rho}\frac{\partial P}{\partial X} = \nu\frac{\partial^2 u}{\partial y^2} \tag{9.2.12}$$

where ρ = fluid density (kg/m^3), $\partial P/\partial X$ is the pressure gradient (N/m^3) along the direction of the flow, ν is the kinematic viscosity (m^2/s), and $\partial^2 u/\partial^2 y$ is the second derivative of the fluid velocity u out of the bearing with respect to position in the gap. The boundary conditions for this problem assume that there is no relative motion between the plates.

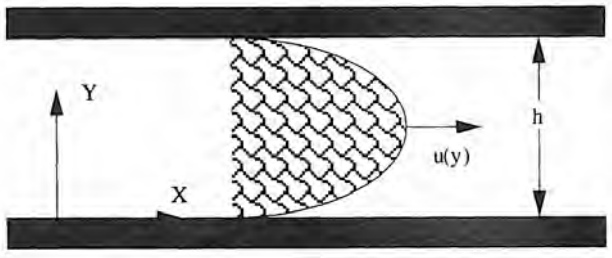


Figure 9.2.5 Fully developed fluid flow between two parallel plates.

Integrating twice and applying the boundary condition, $u = 0$ at $y = 0$ and h, the velocity profile is found as a function of position in the gap and gap height[11]:

$$u = \frac{1}{2\rho\nu}\frac{\partial P}{\partial X}(y^2 - yh) \tag{9.2.13}$$

[11] $\rho\nu = \mu$, which is often referred to as the dynamic viscosity or just the viscosity. It has units of kg/m/s = N-s/m^2.

This flow is parabolic. If the plates were moving with respect to each other at a velocity V, a linearly varying term Vy/h would be superimposed upon the parabolic flow. The flow rate is found by integrating across the gap. It is assumed that the depth of the parallel plates into the page (along the Z direction) is d. The flow rate between the plates due to the pressure gradient ∂P/∂X is

$$Q_{fp} = \int udA = \frac{1}{2\mu}\frac{\partial P}{\partial x}\int_0^d\int_0^h (y^2 - yh)\,dydz = -\frac{dh^3}{12\mu}\frac{\partial P}{\partial x} \tag{9.2.14}$$

The pressure can only be a function of position across the land. Integrating Equation 9.2.14 across the land width l in the X direction:

$$\int_{P_p}^{0} dP = \frac{-12\mu\, Q_{fp}}{dh^3}\int_0^{\ell} dx \qquad P_p = \frac{12\ell\mu}{dh^3}Q_{fp} \tag{9.2.15}$$

With the pad model shown in Figure 9.2.4, the distance d for the rectangular (unshaded) land regions is

$$d = 2(a + b - 4(\ell + r_p)\,) \tag{9.2.16}$$

Hence the fluid flow resistance of the straight sections of the flat pad bearing is

$$R_{ss}\frac{6\ell\mu}{(a + b - 4(\ell + r_p)\,)h^3} \tag{9.2.17}$$

Note that the flow rate due to one flat plate moving at a velocity V with respect to the other would be $Q_f p = Vhd/2$. In general, to be conservative the flow due to the pressure difference should be on the order of twice that due to the flow dragged into the bearing by relative motion. Hence the maximum speed of the structure supported by the bearing, whose inlet flow restrictor's resistance equals the nominal pad resistance, should be

$$V_{max} = \frac{P_p h^2}{12\ell u} \tag{9.2.18}$$

where $P_p = P_u$ or P_l, whichever is smaller. This prevents one side of the bearing pad from becoming starved for oil. Note that the flow out the sides is unaffected and the flow out the other end is increased, so the net pressure across the bearing pad essentially remains unchanged.

What about the corners of the bearing pad? As shown in Figure 9.2.4 the inner corner usually has a radius r_p as a result of milling the pocket. As shown in the figure, it can be assumed that the outer corners have a radius $r_p + l$. Ignoring the outer little bit (shown shaded) lends to the conservativeness of the analysis. The four quadrants together make a circle. To evaluate the resistance of the corners, the polar form of the Navier-Stokes equations are needed. For the steady-state flow of an incompressible fluid in the radial direction only,

$$\frac{1}{\mu}\frac{\partial P}{\partial r} = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{\partial^2 u}{\partial z^2} - \frac{u}{r^2} \tag{9.2.19}$$

The relative order of the terms of the right side are respectively $1/r^2$, $1/h^2$, and $1/r^2$. Since $r >> h$, Equation 9.2.19 reduces to

$$\frac{1}{\mu}\frac{\partial P}{\partial r} = \frac{\partial^2 u}{\partial z^2} \tag{9.2.20}$$

Integrating twice and applying the boundary conditions u = 0 at z = 0,h, the velocity profile is found:

$$u = \frac{1}{2\mu}\frac{\partial P}{\partial r}(z^2 - zh) \tag{9.2.21}$$

The flow out the four rounded corner sections is equal to the flow out of a full 360° annulus:

$$Q_a = \int udA = \frac{1}{2\mu}\frac{\partial P}{\partial r}\int_0^{2\pi}\int_0^h (z^2 - zh)\,dzrd\theta = \frac{-\pi h^3 r}{6\mu}\frac{\partial P}{\partial r} \tag{9.2.22}$$

Again, the pressure can only be a function of position along the land:

$$\int_{P_p}^{0} dP = \frac{-6\mu\, Q_a}{\pi h^3}\int_{r_p}^{rp+\ell}\frac{dr}{r} \qquad P_p = \frac{6\mu \log_e\left(\frac{r_p+\ell}{r_p}\right)}{\pi h^3}q_a \tag{9.2.23}$$

Thus the resistance of the rounded corner regions is

$$R_a = \frac{6\mu \log_e\left(\frac{r_p+\ell}{r_p}\right)}{\pi h^3} \tag{9.2.24}$$

The total resistance of the rectangular flat pad pocketed bearing is thus

$$R = \frac{1}{\frac{1}{R_a} + \frac{1}{R_{ss}}} = \frac{6\mu}{h^3\left[\frac{\pi}{\log_e\left(\frac{r_p+\ell}{r_p}\right)} + \frac{a+b-4(\ell+r_p)}{\ell}\right]} \tag{9.2.25}$$

Total Opposed Pad Flow Rate

Given the supply pressure P_s, the total flow rate at the nominal gap for the opposed bearing pad system will be

$$Q_{total} = \frac{P_s}{R} \tag{9.2.26}$$

Bearing Pad Effective Area

To determine the force exerted by the upper and lower pads, one must individually consider the regions shown in Figure 9.2.4. The pocketed region contributes a force equal to

$$F_{pocket} = Pp\left[(a-2\ell)(b-2\ell) + r_p^2(\pi-4)\right] \tag{9.2.27}$$

For the land regions not at the corners, the pressure decays linearly and they contribute a force of

$$F_{land} = Pp\ell\left[a+b-4(\ell+r_p)\right] \tag{9.2.28}$$

For the rounded corner regions, the pressure decays logarithmically and the four corners together act as a single beveled ring. To find the pressure as a function of radial position in the ring, consider that the portion of the total pad flow Q that goes out the corners is inversely proportional to the total flow resistance of the corners, R_a, and proportional to the pocket pressure, $Q_a = P_p/R_a$. This with Equation 9.2.24 substituted into Equation 9.2.22 yields

$$\frac{dP}{dr} = \frac{-Pp}{\log_e\left(\frac{r_p+\ell}{r_p}\right)}\frac{1}{r} \tag{9.2.29}$$

Integrating and applying the boundary condition P = 0 at $r = r_p + l$, the pressure as a function of radial position is found:

$$P = \frac{-Pp}{\log_e\left(\frac{r_p+\ell}{r_p}\right)}\log_e\left(\frac{r}{r_p+\ell}\right) \tag{9.2.30}$$

The force increment is dF = PdA, so the total force on the four rounded corner sections is:

$$F_{rc} = \frac{-2\pi P_p}{\log_e\left(\frac{r_p+\ell}{r_p}\right)}\int_{r_p}^{r_p+\ell} r\log_e\left(\frac{r}{r_p+\ell}\right)dr = \pi P_p\left[\frac{\ell(2r_p+\ell)}{2\log_e\left(\frac{r_p+\ell}{r_p}\right)}\right] \tag{9.2.31}$$

From Equations 9.2.27, 9.2.28, and 9.2.31, the effective area for the rectangular flat pad pocketed bearing is found:

$$\begin{aligned} A &= (a-2\ell)(b-2\ell) + r_p^2(\pi-4) + \ell\left[a+b-4(\ell+r_p)\right] \\ &+ \pi\left[\frac{\ell(2r_p+\ell)}{2\log_e\left(\frac{r_p+\ell}{r_p}\right)} - r_p^2\right] \end{aligned} \tag{9.2.32}$$

This is the area to be used in evaluating Equations 9.2.10 and 9.2.11 for the stiffness and force. It is difficult to make quick engineering estimates of bearing performance using Equation 9.2.32. It would be nice to find a relation between l and r_p that makes all the r_p terms disappear: the relation is $r_p = 0.4142l$, which makes the effective area

$$A = ab - \ell(b + a) \tag{9.2.33}$$

Often l = 0.25a, so A = a(3b - a)/4.

Power Dissipation and Choice of the Land Width

For precision applications one may wish to optimize the land width l to:

- Minimize pumping power
- Minimize friction power
- Minimize flow
- Maximize allowable speed
- Maximize stiffness
- Maximize load capacity

As discussed earlier, all the energy that enters the bearing in terms of pressure and flow is dissipated as heat, power = QP_s. Since the flow through the bearing is $Q = P_s/R$, the net pumping power required and heat dissipated in the bearing by the flowing oil are each $P_s{}^2/R$. R is given by Equation 9.2.25. Note that although R is the value of each of the resistors in Figure 9.2.2, the total value of the circuit's (opposed pad bearing system) resistance is also R. In general, the pressure should be kept as low as possible in order to minimize pumping power. In order to obtain the desired stiffness and flow, larger pads should be used in place of high pressure. This is not always possible, but should serve as a general guideline.

The friction power is the heat generated by viscous shear as the carriage moves back and forth. For high-speed or high-cycle machine components (e.g., grinder tables) this power can be significant. The shear stress in the fluid between two plates moving relative to each other at a velocity V_{rel} is

$$\tau_{yx} = \mu\frac{du}{dy} = \frac{\mu V_{rel}}{h} \tag{9.2.34}$$

The friction force is the product of the shear stress, and the area and the friction power is the product of the friction force and the velocity. For the area of the pocket the shear stress acts over, four times the actual area is used to account for viscous power generated by recirculatory flow[12]:

$$\mathcal{P}_{\text{Friction}} = \mu V_{rel}^2 \left(\frac{\text{Area}_{\text{lands}}}{h_{\text{lands}}} + \frac{4\text{Area}_{\text{pocket}}}{h_{\text{pocket}}} \right) \tag{9.2.35}$$

When the land width $\ell = \beta a$, the "optimal" value of β will depend on the values of a, b, h, and P_s. In general, as β decreases, force and stiffness increase, but so do pumping power and flow; on the other hand, friction power decreases. The inverse is also true. A good compromise for many applications is $\beta = 0.25$. Many references present bearing optimization parameters that allow the designer to choose "optimal" fluid viscosity and bearing dimensions; however, these models are almost always based on an isoviscous model.[13] In addition, real-world factors such as available oil viscosity and the occurrence of turbulent flow often provide an impediment to the application of handbook equations. As discussed below, iterative methods using spreadsheets allow the designer quickly to chart performance trends over a wide range of design parameters. The capability to play "what if" games can give a designer a powerful competitive advantage.

Example

Consider the design of hydrostatic bearings to replace the sliding bearings used in the example of Figure 8.2.16. Figure 9.2.6 shows the results of a spreadsheet analysis of the problem that uses the design equations presented in this section. To help compare performance, an estimate of power dissipation can be made. For the sliding bearing system, there are eight pads on the rails in the Y direction and each one sees 250 N of preload force, including the weight of the carriage. Thus

[12] The factor of 4 is recommended by J. N. Shinkle and K.G. Hornung, "Frictional Characteristics of Liquid Hydrostatic Journal Bearings," J. Basic Eng. Trans. ASME, Vol. 87, No. 2, March 1965, pp. 163–169.

[13] In an isoviscous model, the viscosity does not change with temperature, which is practical only for quasi-steady speeds. For a nonisoviscous model, see Figure 9.2.21.

% gap closure factor: alpha	0.10	0.10
Supply pressure (Pascals, psi)	3,450,000	500
Oil viscosity (N-sec/m^2)	0.01	
Nominal gap (m, in)	0.000010	0.0004
Desired force at alpha (N, lbf)	4450	1000
Required effective area/pad (m^2, in^2)	0.002151	3.3333
a (b = 2a, Beta = 0.25, and Eq. 9.2.33)	0.0415	1.633
b	0.0830	3.266
Resultant stiffness (N/m, lbf/in)	4,375,000,000	25,000,000
Flow (m^3/sec, gpm)	2.15E-06	0.034
Pumping power (Watts, Hp)	7.41	0.010
Max. velocity (m/s, ipm)	0.14	338
Friction power (Watts, Hp)	0.35	-
Total power (Watts, Hp)	7.8	0.010
Sliding bearing friction coeff.	0.10	0.10
Sliding bearing weight (N, lbf)	2074	466
Sliding bearing preload (each) (N, lbf)	250	56
Sliding bearing power (Watts, Hp)	44.0	0.059

Figure 9.2.6 Portion of spreadsheet for T bearing design example.

the force to get the carriage moving (and keep it moving) due to the preload and a coefficient of friction of 0.1 is 8 × 250 × 0.1 = 200 N. To estimate the carriage weight, assume that with all the components that might be bolted to the carriage, the weight is equal to a block of cast iron 420mm × 480mm × 150mm. The density of cast iron is about 7000 kg/m^3, so the weight of the carriage is about 2074 N. Thus the total sliding force is 407N. The spreadsheet shows the maximum velocity and power for the hydrostatic bearing. When it is moving, the hydrostatic bearing system introduces an order of magnitude less power than the sliding bearing system. When idling, the hydrostatic bearing generates heat, whereas the former does not. What other comparisons can be made between the two systems, and which would you choose for different applications?

9.2.3 Variations on the Opposed Pad Design

The basic methods and theory presented in the preceding section can be applied to many different types of bearing configurations. This section will discuss design parameters for journal bearings for supporting slowly rotating shafts, bidirectional annular thrust bearings, and quasi-kinematic arrangements of flat pad bearings. It is assumed that all of these bearings use laminar inlet flow restrictors.

Hydrostatic Journal Bearings

Because of hydrodynamic effects, journal bearing design can be a very difficult task, especially if one considers the problem of heat buildup caused by the rotating shaft. The discussion presented here allows one to make first order estimates of load and stiffness capacity for quasi-static operation. Note that it is assumed that for maximum load and stiffness capacity, the inlet restrictor resistance equals the pad resistance as was derived in the previous section. See Section 9.2.6 for a discussion of some of the thermal design issues for hydrostatic spindles.

For low-speed applications,[14] the analysis presented in the preceeding section can be used to help determine the performance characteristics of hydrostatic journal bearings. At high speeds, the rotating shaft tends to pull the oil exiting from one pad into the adjacent pad, causing hydrodynamic lift. Hydrodynamic lift greatly increases the bearing capacity but also causes the shaft that is supported to move off center. For precision applications this is not always desirable, and hence hydrodynamic bearings are not considered here. Most engineering libraries have numerous volumes on hydrodynamic bearing design.

As shown in Figure 9.2.7, a journal bearing pad that supports a shaft of diameter D_s is geometrically defined in terms of its pad width a, land width l, and arc angle ϕ. The drainage grooves between the pads help minimize the buildup of a hydrodynamic fluid wedge. This can help to minimize radial motion of the shaft, but it will decrease the maximum allowable speed of the shaft. Four

[14] Low speed is defined here by the condition of Equation 9.2.18.

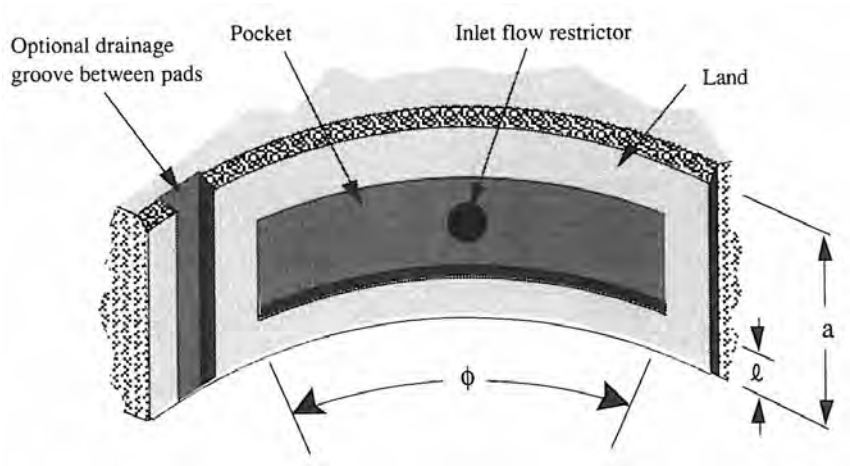


Figure 9.2.7 Hydrostatic journal bearing pad geometry.

pads are commonly used, but additional pads can help to minimize variations of load and stiffness with circumferential position.

For the purposes of determining the journal bearing pad's mean perimeter and fluid resistance, the pad can be considered flat, where the pad length b is $D_s\phi/2 + 2l$. A journal bearing's pockets are usually made using a T-slot cutter, so the corners of the pocket are square and rp is essentially zero, which prevents evaluating the pad resistance with Equation 9.2.25. In order to determine the pad resistance, a mean perimeter d_m is defined where the land area outside the perimeter equals that inside:

$$d_m = 2\sqrt{(a+b)^2 - 4\ell(a+b-2\ell)} \tag{9.2.36}$$

From the parallel-plate model, the fluid resistance for each pad's land regions is thus

$$R = \frac{6\ell\mu}{h^3\sqrt{(a+b)^2 - 4\ell(a+b-2\ell)}} \tag{9.2.37}$$

The restrictor resistance typically has the same value. The flow resistance for the four bearing pads that surround the shaft is formed by four sets of resisters in parallel, where each set of resistances is composed of two resistors (Equation 9.2.37) in series; thus $Q = 2P_s/R$.

The bearing pad cannot be considered flat for the purposes of determining the radial force the bearing can support. For a hydrostatic journal bearing where the pocket pressure is one-half of the supply pressure, the incremental radial force from a trapezoidal force wedge across the lands and the pocket at any angle θ, not including the end lands, is

$$dF_R \approx \left(\frac{3P_s\alpha}{2}\right)(a-\ell)\left(\frac{D_s\,d\theta}{2}\right) \tag{9.2.38}$$

The effective incremental force that supports the shaft is thus

$$dF_s \approx \frac{3P_s\alpha(a-\ell)D_s\cos\theta\,d\theta}{4} \tag{9.2.39}$$

The net force on the shaft, not including a contribution from the end lands, is

$$F_{pocket} \approx 2\int_0^{\phi/2} \frac{3P_s\alpha(a-\ell)D_s\cos\theta\,d\theta}{4} \approx \frac{3P_s\alpha(a-\ell)\,D_s\sin(\phi/2)}{2} \tag{9.2.40}$$

It is assumed that the end lands help support the shaft with a force equal to the product of the volume of the end lands' force pyramid and $\sin(\phi/2)$. Hence the total force on the shaft due to a pair of opposed journal bearing pads is

$$F_S \approx \frac{3P_s\alpha}{2}\left\{(a-\ell)D_s + \frac{4\ell^2}{3} + (a-2\ell)\ell\right\}\sin\left(\frac{\phi}{2}\right) \tag{9.2.41}$$

Note that it is assumed that a drainage groove is cut in the longitudinal direction between pads as shown in Figure 9.2.7. The width of this groove is such that the land width is constant around the bearing perimeter. The effective area A to use when evaluating the stiffness with Equation 9.2.10 is

$$A = \left\{ (a - \ell)\, D_s + \frac{4\ell^2}{3} + (a - 2\ell)\ell \right\} \sin\left(\frac{\phi}{2}\right) \tag{9.2.42}$$

Bidirectional Annular Thrust Bearings

An annular thrust bearing is shown in Figure 9.2.8. The annular thrust bearing described herein relies on journal bearings at each end of the shaft to provide angular rigidity. If the bearing were required to have angular rigidity, the annulus would have to be divided up into four quadrants, each with its own inlet flow restrictor.

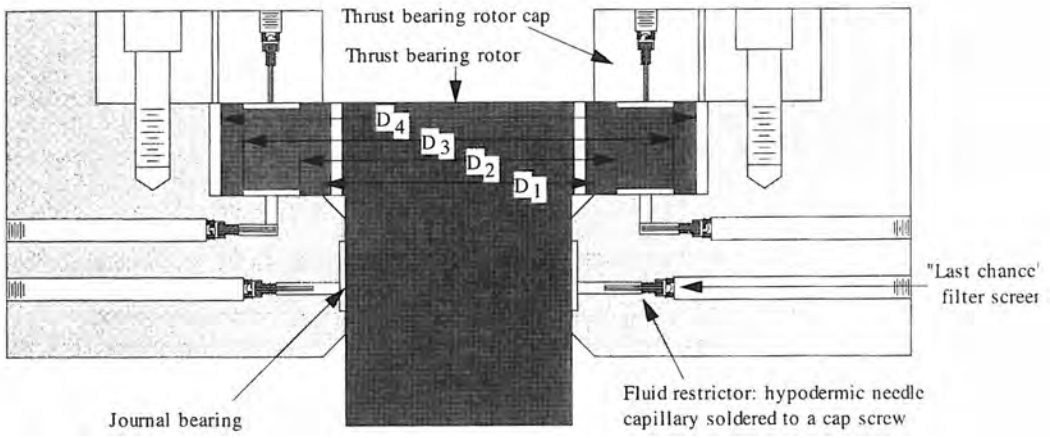


Figure 9.2.8 Annular thrust bearing used to support bidirectional thrust loads on a shaft.

Because the annular thrust bearing is circular, the polar form of the Navier-Stokes equations are used to determine the fluid resistance of the bearing pad. It is still assumed that the inlet flow resistance equals the bearing pad resistance, so Equation 9.2.8 is valid for ΔP. Given the pad and pocket inside and outside diameters D_1, D_4, D_2, D_3, respectively, the total pad resistance can be found using the analysis techniques of Equations 9.2.19–9.2.25, and summing the inner and outer pad resistances in parallel:

$$R = \frac{6\mu \log_e (D_4/D_3) \log_e (D_2/D_1)}{\pi h^3 [\log_e (D_4/D_3) + \log_e (D_2/D_1)]} \tag{9.2.43}$$

The effective area is found using the analysis techniques used for Equations 9.2.29-9.2.31, and adding the effective areas of the inner and outer lands and the pocket:

$$A = \frac{\pi}{4} \left[\frac{D_4^2 - D_3^2}{2\log_2(D_4/D_3)} - \frac{D_2^2/ - D_1^2}{2\log_e(D_2/D_1)} \right] \tag{9.2.44}$$

The force and stiffness are given by Equations 9.2.10 and 11.

Note that since there are no end lands, this type of annular thrust bearing is not speed limited by oil being dragged out of the end land. This helps to maximize the allowable speed of a shaft supported by a pair of journal bearings and a bidirectional thrust bearing.

Quasi-kinematic Arrangements of Flat Pad Bearings

In order to minimize the number of bearing pads (and manufacturing costs) required to support a linearly moving carriage, an arrangement of six bearing pads may be used as shown in Figure 9.2.9. To be ideally kinematic, a single pad in the middle of the top of the carriage would be needed. However, this would require the use of a keeper rail that is cantilevered too far to be practical. Since hydrostatic bearings fill gaps between the bearing pads and rails with high-pressure oil, one does not have to worry about overconstraint; hence the six (or any larger number) pad design

behaves kinematically and is called a quasi-kinematic design. Note that one could use an eight-pad design where each lower pad is preloaded by an upper pad directly above it. This would yield a design that has much greater pitch stiffness and easier flow restrictor maintenance (see Section 9.2.4) but would require slightly more complicated porting and more flow. Whenever an angled pad design is used, one must be careful to make the base way sufficiently massive so that the lateral forces from angled pads do not cause the base way to spread open.

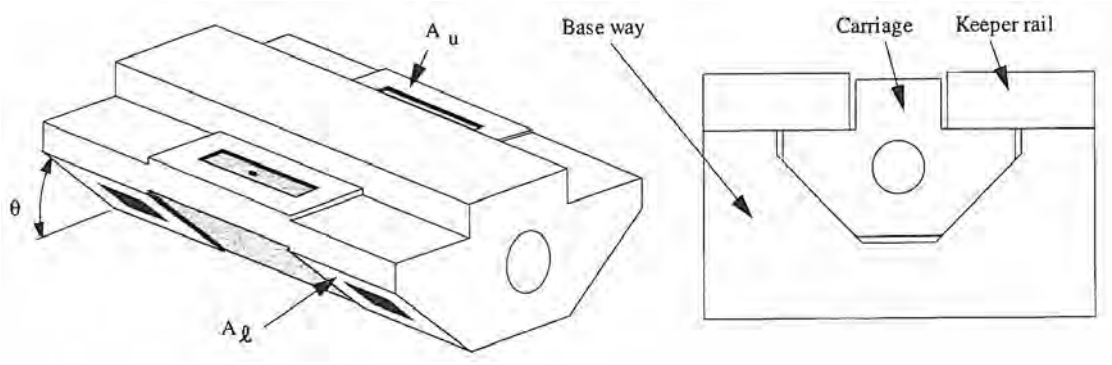


Figure 9.2.9 A quasi-kinematic arrangement of hydrostatic bearings for a linear slide.

One advantage of this design over dovetail or rectangular ways is that it is easy to manufacture and the geometry is self-checking. To manufacture this system, the angled surfaces in the base way and carriage are first finished by grinding, scraping, or lapping. The carriage is then placed in the base way with shim stock of thickness H between the angled surfaces. The top surface of the base way and carriage are then ground flat together as a unit. When the shim stock is removed, the nominal upper and lower bearing gaps will be

$$h = \frac{H}{1 + \cos\theta} \tag{9.2.45}$$

A motion δ_u of the upper gap will correspond to a motion of the lower gap $\delta_l = \delta_u \cos\theta$.

If the inlet flow restrictor resistances equal the bearing pad resistance at the nominal position, then the upper and lower pad pocket pressures will be equal; thus in order to have force equilibrium while maintaining a nominal gap h, the effective upper and lower pad areas must satisfy the following:

$$A_{\text{effective upper pad}} = 2A_{\text{effective lower pad}} \cos\theta \tag{9.2.46}$$

9.2.4 Flow Restrictor Design

The resistance and flow predicted by the models above can be used initially to size the inlet restrictors and the pressure supply unit, and thus help choose the bearing dimensions (e.g., land width). In this section, some of the many methods for providing inlet flow restriction will be discussed.

Orifices

Orifices create turbulent flow, which is undesirable for precision applications. They also yield a bearing with lower stiffness than do laminar flow devices, but in most cases this is easily overcome simply by making the bearing larger. For extremely small bearing gaps, a series of cascaded orifices may be the only practical way to attain a high desired restrictor resistance. Fortunately, one can purchase high-resistance cascaded orifice assemblies complete with last-chance filters.[15] These devices are typically about 5 mm in diameter and 20 mm long.

Flat-Edge Pins

A simple way to manufacture a flow resistance is to take a solid pin (e.g., a dowel pin) and grind a flat edge on one side, as shown in Figure 9.2.10. When pressed into a hole, the flow region looks like a flattened arch. One must be careful, however, to make sure that the pin does not create

[15] For example, from the Lee Company, 2 Pettipaug Road, P.O. Box 424, Westbrook, CT 06498-0424.

shavings when it is pressed into the hole. Consider a cylinder of length L with the top ground off, the region of flow is bounded by the angle θ_{max} that is a function of the amount ε ground off the pin of diameter D_p:

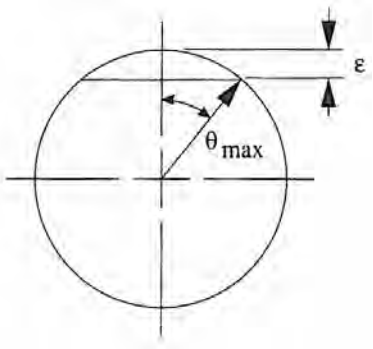


Figure 9.2.10 Laminar flow inlet restrictor made by grinding off the top of a pin and then later pressing it into a hole.

$$\theta_{max} = \cos^{-1}\left(1 - \frac{2\varepsilon}{D_p}\right) \tag{9.2.47}$$

The height of any section the fluid flows through is

$$H = 0.5\, D_p(\cos\theta - \cos\theta_{max}) \tag{9.2.48}$$

The incremental width of a section for purposes of integrating across the chord is

$$dw = 0.5\, D_p \cos\theta\, d\theta \tag{9.2.49}$$

The resistance of each incremental width section is found from Equation 9.2.15. The total resistance is that of all the incremental width resistances in parallel[16]:

$$\begin{aligned} R &= \frac{R}{\Sigma\frac{1}{R}} = \frac{1}{\Sigma\frac{H^3\, dw}{12\mu L}} = \frac{1}{\frac{D_p^4}{192\,\mu L}\int_0^{\theta_{max}} \cos\theta(\cos\theta - \cos\theta_{max})^3\, d\theta} \\ &= \frac{192\,\mu L}{D_p^4\left\{\frac{3}{8}\,\theta_{max} + \frac{1}{4}\sin 2\theta_{max} + \frac{1}{32}\sin 4\theta_{max} \right.} \\ &\qquad -3\cos\theta_{max}\left(\frac{1}{12}\sin 3\theta_{max} + \frac{3}{4}\sin\theta_{max}\right) \\ &\qquad \left. +3\cos^2\theta_{max}\left(\frac{1}{4}\sin 2\theta_{max} + \frac{\theta_{max}}{2}\right) - \cos^3\theta_{max}\sin\theta_{max}\right\} \end{aligned} \tag{9.2.50}$$

where all the terms in $\{.\}$ are in the denominator. The bearing pads and gaps are usually designed first and then the flow restrictors are designed. The values of ε, L, and D_p must all be balanced to yield the proper resistance and a machinable design. If ε is too small, then manufacturing errors can cause the desired resistance to be too great or too small. Typically, $\varepsilon > D_p/20$ and $L > 100\varepsilon$.

One should also check the Reynolds number for the design, and for all restrictor designs and regions in the bearing, to ensure that the flow is laminar. Turbulence generates noise and heat. In addition, to prevent turbulent flow occurring as is the case in an orifice, the length of the device should be about 100 times the gap ε. For flow in pipes, $R_e < 2{,}000$ (2000 - 4000 is a laminar to

[16] For help in evaluating the integral, expand the integrand and see I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series and Products, Academic Press, New York, 1980, p. 132, Article 2.5.13, nos. 11, 12, 13.

turbulent transition zone) and for flow between parallel plates, $R_e < 7700$.[17] To be conservative, one should require that $R_e < 2{,}000$:

$$2,000 > \frac{\text{velocity} \times \text{gap}}{\text{kinematic viscosity}} = \frac{(Q/A)\varepsilon}{\nu} \tag{9.2.51}$$

The cross-sectional area of the fluid flow region of the inlet restrictor is

$$\begin{aligned} A &= 2\int_0^{\theta_{max}} H dw = \frac{D_p^2}{2}\int_0^{\theta_{max}} \cos\theta(\cos\theta - \cos\theta_{max})\, d\theta \\ &= \frac{D_p^2}{4}\{\theta_{max} - \cos\theta_{max}\sin\theta_{max}\} \end{aligned} \tag{9.2.52}$$

In most cases it is simple to design the restrictor (using a computer) while meeting the Reynolds number criteria.

Capillary Tubes[18]

A small-diameter hole (radius r_c) is difficult to drill deep in a metal part, but glass or stainless steel (hypodermic) capillary tubing is readily available. The diameter of the hole (i.e., fluid resistance) is often not specified to the accuracy desired and the diameter is difficult to measure; thus a test jig may have to be made to measure the fluid resistance of a known length of the capillary tubing. The resistance is proportional to length, so the capillary can be cut to yield the desired resistance. The alternative is to measure the pocket pressure when a trial resistor made from a piece of the capillary is used. The latter is the most effective means of tuning a bearing, and the wise bearing designer includes pressure measuring ports in his design. In order to ensure laminar flow through the capillary, the length-to-diameter ratio should be greater than 50. If necessary, a bundle of smaller-diameter tubes can be used to obtain a resistance equivalent to a single large-diameter short length tube. Note that it is impractical, from a manufacturing point of view, to use a capillary tube with a diameter smaller than about 0.4 mm (0.016 in.). If a higher resistance is needed, then a series of cascaded orifices should be used.

The cylindrical form of the Navier-Stokes equations are used to find the fluid resistance of a capillary. By definition, a capillary's length-to-diameter ratio is large, so end effects can be ignored and the flow will be fully developed, laminar, and steady without circular flow:

$$\frac{1}{\mu}\frac{dP}{dx} = \frac{1}{r}\frac{d}{dr}\left(r + \frac{du}{dr}\right) \tag{9.2.53}$$

Holding $\mu dP/dx$ constant, integrating twice, and simplifying yields

$$u = \frac{1}{\mu}\frac{dP}{dx}\left(\frac{r^2}{4} + C_1 \log_e r + C_2\right) \tag{9.2.54}$$

The boundary conditions are u = 0 at r=r_c and u is finite at r = 0; hence

$$u = \frac{1}{4\mu}\frac{dP}{dx}\left(r^2 - r_c^2\right) \tag{9.2.55}$$

The flow Q through the capillary is

$$Q = \int udA = \frac{\pi}{2\mu}\frac{dP}{dx}\int_0^{r_c}(r^2 - r_c^2)\, rdr = \frac{-\pi r_c^4}{8\mu}\frac{dP}{dx} \tag{9.2.56}$$

Integrating along the length yields

$$P_p = \frac{8\ell\mu Q}{\pi r_c^4} \tag{9.2.57}$$

[17] See the textbook by M. Potter and J. Foss, Fluid Mechanics, Published by Potter and Foss, Michigan State University, Lansing, MI, 1982, p. 298.

[18] It is often difficult to find a reliable supplier of tubing that is willing to sell you less than a kilometer. One supplier with a wide in-stock (from 0.004 in. ID on up) selection is Cooper's Needle Works Ltd. 261 Aston Lane, Birmingham, West Midlands B20 3HS, England, phone 011 44 21 356 4719.

The flow resistance is

$$R = \frac{8\ell\mu}{\pi r_c^4} \tag{9.2.58}$$

How does this compare with the flow resistance of the flat-edge pin? Given a nominal pin diameter of 6 mm, Figure 9.2.11 plots the ratio of the capillary to flat-edge pin fluid resistances as a function of characteristic dimension. The characteristic dimension is the largest particle that could pass through the device, which is equal to the amount ground off the edge of the pin or to the capillary diameter. The capillary provides greater flow resistance; however, consider that the resistance is a function of the characteristic dimension to the fourth power. For the range of resistances usually encountered in bearing pad design for precision machine tools, a capillary tube will in general pass a particle that is twice the size of a flat-edge pin.

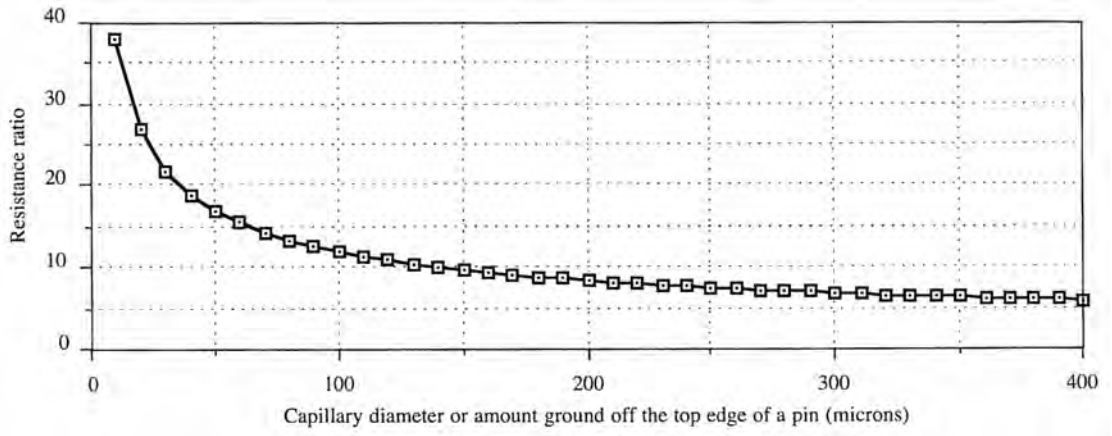


Figure 9.2.11 Flow restrictor resistance ratio for a capillary and a 6 mm pin with the top edge ground off.

Constant-Flow Devices

If the flow to the bearing pads can be precisely controlled regardless of gap resistance, then greater stiffness and flow can be achieved. Constant-flow devices are available that are about 7 mm in diameter and 30 mm long. Internally, they have a spring-loaded tapered plug and seat. The plug responds to changes in flow in order to keep the flow constant regardless of the pressure difference across the plug until they saturate. Given a large constant-flow device and a small gap (high bearing resistance) the pressure in the pocket will saturate and less than the rated flow will occur.

One note of caution regarding the use of constant-flow devices, if the viscosity varies from the assumed value, then the supply flow may easily be too large or too small for the bearing. If the flow supply is marginally too large, the bearing quickly saturates and performance degrades. If the flow becomes too much, then the bearing may be saturated at equilibrium and not be able to support a load at all. Given that all oils have large variations in viscosity with temperature, it is best to avoid using constant-flow devices in hydrostatic bearings unless the oil characteristics have been well established. The pressure difference across a bearing pad that uses an inlet resistance is not a function of viscosity; therefore, in most cases laminar flow inlet resistors should be used.

Proportional Flow Restrictors

It is possible to design an inlet restrictor with a resistance that is proportional to flow using a diaphragm as shown schematically in Figure 9.2.12.[19] This is called a *diaphragm-type restrictor.* As the upper gap increases, a pressure differential is generated across the diaphragm, which causes it to decrease the flow to the upper pad's inlet restrictor. The pressure on the upper pad drops far more quickly than is the case with a single laminar flow restrictor. This creates a hardening spring effect and bearing stiffness is greatly enhanced. In fact, over a finite load range, the bearing can appear to be infinitely stiff. However, with infinite stiffness, one will not obtain any squeeze film damping which requires motion to dissipate energy. Since a laminar flow inlet restrictor is still used to supply

[19] See J. Degast, U.S. Patent 3,442,560, May 6, 1969.

flow to the pad, the viscosity of the fluid will not affect the force and stiffness characteristics of the bearing. Other types of self-regulating inlet restrictors exist, and most operate on the same variable resistance principle. Note that these devices must still be tuned to the particular pad geometry and gearing gap used.

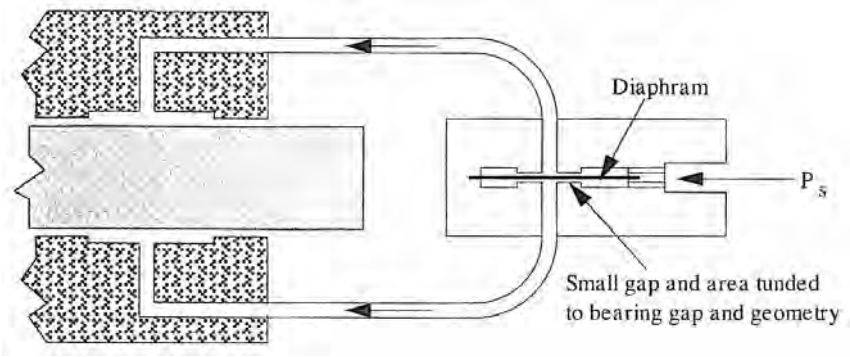


Figure 9.2.12 Operating principle of a diaphragm-type flow restrictor.

One can go a step further and introduce a means to servo-control the pressure differential across the bearing gap.[20] Depending on the bandwidth of the servo, five of an axis's six geometric static and dynamic errors (to a certain degree) could be compensated for. Servovalves can be used to control the pressure across the bearings, but the initial and maintenance costs would probably be prohibitive. A better alternative is to use a diaphragm-type restrictor that uses a piezoelectric actuator to bias the position of the diaphragm. In fact, a bimorph piezoelectric element could be used as the diaphragm itself. Note that should the servo fail, the machine would continue to function, although without error compensation. This would at least allow the machine to be used for some parts while the servo problem was being fixed.

The problem with these devices is their mechanical complexity and potential for introducing dynamic instabilities into the system. The former increases manufacturing costs while the latter introduces more uncertainty into the design of a prototype. If at all possible, it is recommended that this type of device not be used. If requirements for high stiffness, low power, and small space cannot be met with a conventional design, self-regulating inlet restrictors can be used but extensive dynamic modeling is required.

Numerous other alternatives, including adjustable devices, are discussed in the reference by Stansfield cited at the beginning of Section 9.2. Regardless of the method used to obtain fluid resistance, the key is to keep the flow laminar to avoid noise caused by turbulence.

9.2.5 Self-Compensating Bearings

Although designing restrictors may be straightforward, the third- and fourth-power relationships in some of the calculations make most restrictors extremely sensitive to manufacturing tolerances. *Self-compensation* or *gap compensation* is based on the principle that high-pressure fluid can be regulated through passages on the bearing surface. The fluid flows out of a pocket on the surface of the bearing, across special lands, and into a small collection pocket that is connected to a large bearing pocket on the other side of the bearing rail. As the bearing rail is loaded and the gap on one side begins to close, the resistance to flow increases to the pad that generates a force in the direction of the load. Conversely, fluid flows more easily to the pad that is resisting the load. A self-compensating design is insensitive to the manufacturing variations in the bearing gap. In addition, a properly designed self-compensated bearing is minimally affected by deflection of bearing components (e.g., keeper rails). Self compensating designs are also far less sensitive to dirt because there are no small diameter passages. Once again, the principle of self-help illustrates its importance. There are many

[20] See, for example, J. Zeleny, "Servostatic Guideways-A New Kind of Hydraulically Operating Guideways for Machine Tools," Proc. 10th Int. Mach. Tool Des. Res. Conf. Sept. 1969. Also see U.S. Patents 4,630,942 and 4,080,009.

different types of self-compensated bearing designs. Figure 9.2.13 illustrates a novel design[21] for a self compensated modular hydrostatic bearing.

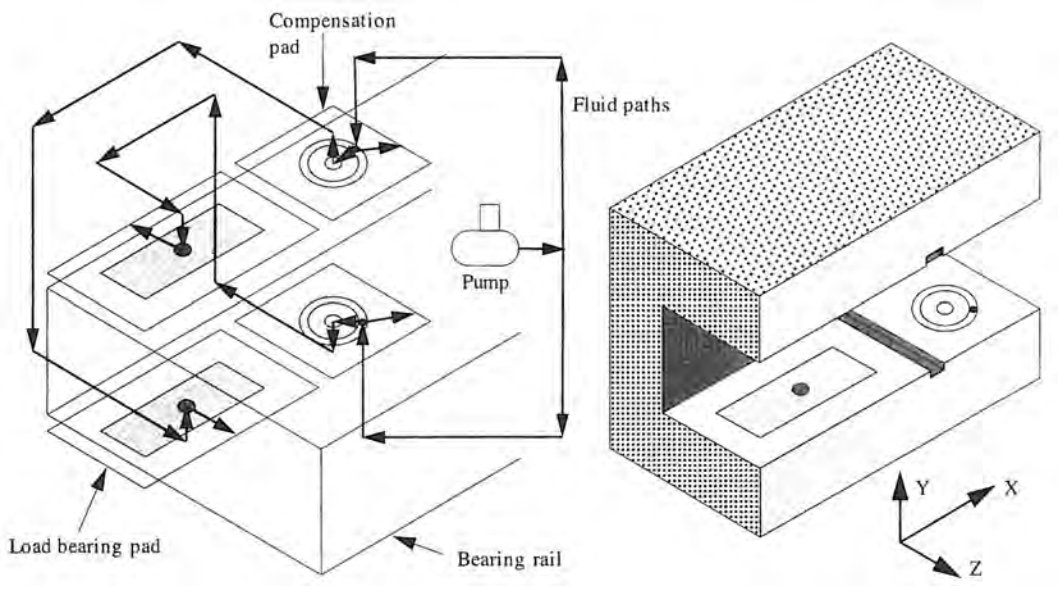


Figure 9.2.13 The Hydroguide®, a self-compensating modular hydrostatic bearing. (Courtesy of Aesop, Inc.)

Assuming that the self-compensation unit is properly designed so that there is no detrimental leakage flow,[22] Equation 9.2.3 can be modified so that the restrictor resistance R is proportional to the opposed bearing gap, and the nominal restrictor resistance is equal to the product of γ and the opposed pad bearing resistance:

$$\Delta P = P_s \left(\frac{\dfrac{1}{(h-\delta)^3}}{\dfrac{\gamma}{(h+\delta)^3} + \dfrac{1}{(h-\delta)^3}} - \frac{\dfrac{1}{(h+\delta)^3}}{\dfrac{\gamma}{(h-\delta)^3} + \dfrac{1}{(h+\delta)^3}} \right) \tag{9.2.59}$$

The load capacity is equal to the product of Equation 9.2.59 and the effective area (Equation 9.2.32). The stiffness is equal to the product of the effective area and $\partial/\partial\delta$ of Equation 9.2.59:

$$K = 3A P_s \left\{ \frac{\dfrac{\gamma}{(h-\delta)^4} - \dfrac{1}{(h+\delta)^4}}{(h+\delta)^3 \left(\dfrac{\gamma}{(h-\delta)^3} + \dfrac{1}{(h+\delta)^3} \right)^2} + \frac{1}{(h+\delta)^4 \left(\dfrac{\gamma}{(h-\delta)^3} + \dfrac{1}{(h+\delta)^3} \right)} \right.$$
$$\left. - \frac{\dfrac{1}{(h-\delta)^4} - \dfrac{\gamma}{(h-\delta)^4}}{(h-\delta)^3 \left(\dfrac{1}{(h-\delta)^3} + \dfrac{\gamma}{(h+\delta)^3} \right)^2} + \frac{1}{(h-\delta)^4 \left(\dfrac{1}{(h-\delta)^3} + \dfrac{\gamma}{(h+\delta)^3} \right)} \right\} \tag{9.2.60}$$

Figures 9.2.14 and 9.2.15 show the stiffness and load capacity of an ideal (e.g., the Hydroguide®) self-compensating bearing. Note that for the same resistance ratio, the stiffness level will be about twice the value for an equivalent capillary compensated bearing. To give a more uniform response, the resistance ratio γ would be about 3 or 4. Note that this still gives a greater stiffness than is obtained with a capillary compensated bearing with a resistance ratio of 1. Figure 9.2.16 shows spreadsheet design results for a self-compensated bearing in which a group of six opposed pad pairs are used to support a machining center's axes.

[21] "Self Compensating Hydrostatic Linear Bearing," #5,104,237, April 14, 1992, Aesop, Inc., 81 North State Street, Concord, NH 03301. Foreign patents pending

[22] As mentioned, there are many self-compensator designs available. Virtually all have leakage flows which make the performance differ from the ideal case described here. However, there is a new design, called the Hydroguide®, that has been developed and patented by the present author.

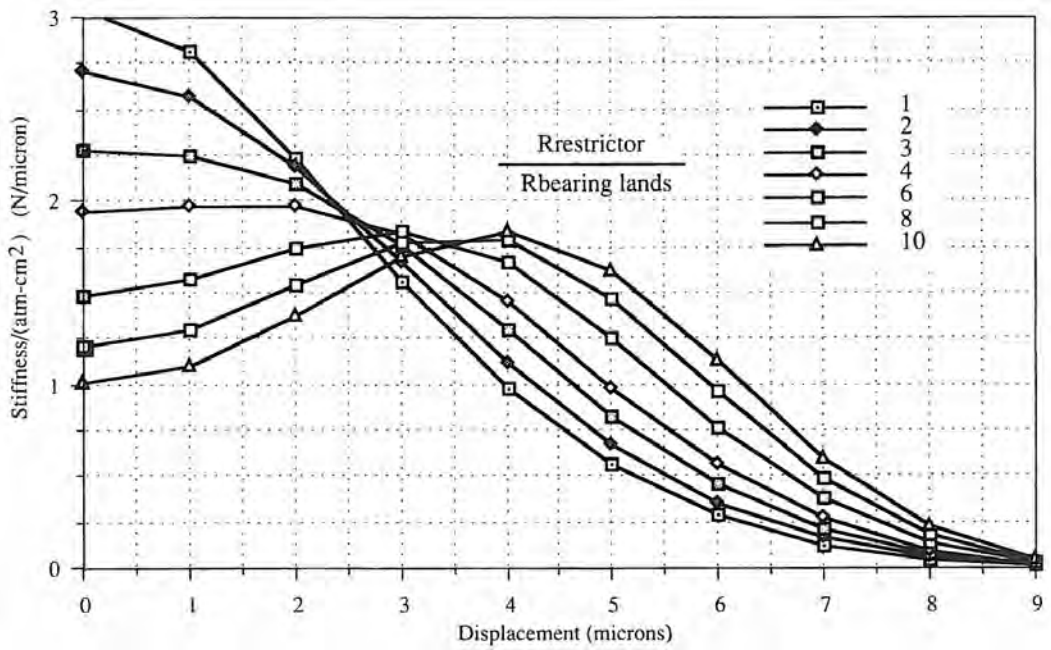


Figure 9.2.14 Stiffness of an ideal self-compensated bearing with a 10 micron nominal gap.

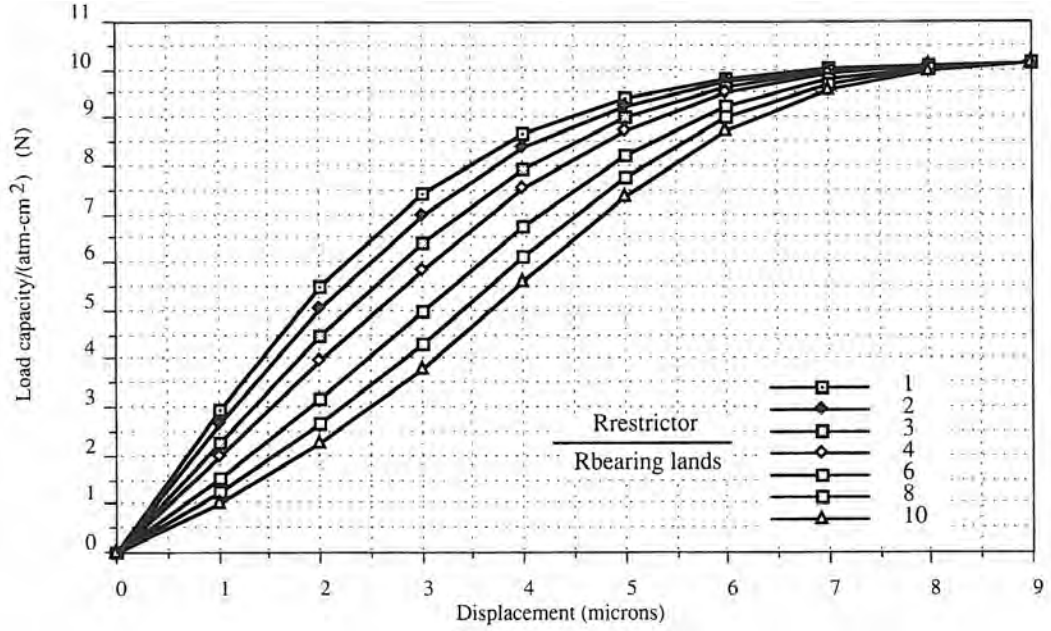


Figure 9.2.15 Load capacity of an ideal self-compensated bearing with a 10 micron nominal gap.

Supply Pressure Ps (N/m^2, psi,atm)	2,028,600	294	*20*	
Viscosity mu (N-s/m^2) (water)	*0.001*			
Nominal bearing gap h (m, in, μm)	**0.000010**	0.000394	*10.0*	
Rectangular bearing characteristics (one pad either side of compensator)				
Width a (m, in, mm)	**0.0300**	1.18	*30.0*	
Length b (m, in, mm)	**0.0600**	2.36	*60.0*	
Land width l (25% of width) (m, in, mm)	0.0075	0.30	*7.5*	
Pocket radius rp (m, in, mm)	**0.0040**	0.16	*4.0*	
Fluid resistance across bearing lands Rbearing (Nsec/m^5)	*6.79E+11*			
Effective pad area (cm^2, in^2, mm^2)	11.14	1.73	*1114*	
Results are for each pad pair:				
	Self compensating:		Capillary:	
gamma=Rrestrictor/Rbearing	***3***		**1**	
Load capacity at 50% gap closure (N, lb)	*2,006*	451	1,492	335
Initial stiffness (N/μm, lb/μin)	*508*	2.91	339	1.94
Stiffness at 25% gap closure (N/μm, lb/μin)	*436*	2.49	310	1.77
Stiffness at 50% gap closure (N/μm, lb/μin)	*184*	1.05	214	1.22
Flow (liters per minute)	*0.24*		0.18	
Pump power (Watts)	*8.05*		6.06	

Figure 9.2.16 Spreadsheet results for the design of an ideal self-compensated bearing.

9.2.6 Step-Compensated Bearings

One of the problems associated with hydrostatic bearings is their design complexity. This is particularly true for hydrostatic journal bearings where it is difficult to bore integral passageways into the spindle head. A step-compensated bearing can be an order of magnitude simpler to manufacture than any other type of hydrostatic bearing and thus it can be well-suited to hydrostatic journal bearing design.

A step-compensated bearing typically utilizes a constant-depth pocket whose depth is a typically 2 to 2.5 times the nominal gap between the bearing lands and the bearing rail or bore. Sometimes a tapered pocket is used to optimize performance for a given application. Figure 9.2.17 illustrates the design principle for a constant-depth pocket. Fluid enters the bearing unrestricted from the pressure source. A pressure gradient exists across the pocket width and across the land width. As the bearing is displaced, the gradient increases on the side the load is applied and hence a restoring force is generated.

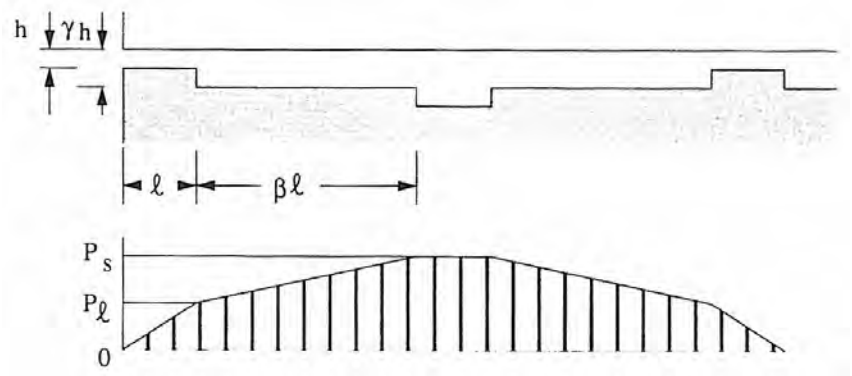


Figure 9.2.17 Geometry of a step-compensated bearing and equivalent resistance diagram.

The fluid resistances of the upper pocket and land are given respectively by:

$$R_{\ell u} = \frac{12\ell\mu}{(h+\delta)^3} \qquad R_{pu} = \frac{12\beta\ell\mu}{(\gamma h-\delta)^3} \tag{9.2.61}$$

The fluid resistances of the lower pocket and land are given respectively by:

$$R_{\ell 1} = \frac{12\ell\mu}{(h-\delta)^3} \qquad R_{p\,1} = \frac{12\beta\ell\mu}{(\gamma h-\delta)^3} \tag{9.2.62}$$

The force, per unit depth into the page, that the bearing can support is:

$$F_{step} = P_s\ell(1+\beta)\left[\frac{R\ell_u}{R_{pu}+R_{\ell u}} - \frac{R_{\ell 1}}{R_{p1}+R_{\ell 1}}\right] \tag{9.2.63}$$

Conversely, the force, per unit depth into the page, that a capillary compensated bearing can support has the scale factor $2P_sl(0.5+\beta)$.

In order to design a step bearing, the ratios γ and β must be chosen. This is best accomplished using a spreadsheet to make a parametric study of the design variables. Typically β is 10 and γ is in the range of 2.0-2.6. Typically a capillary compensated bearing has almost four times the load capacity and stiffness. Note that a self-compensated bearing can typically have twice the load capacity and stiffness of a capillary compensated bearing. Hence although a step-compensated bearing is easier to manufacture, when bearing pad size must be minimized, a capillary or self-compensating bearing should be used.

Where space is not greatly restricted but manufacturing costs and thermal considerations are of prime importance, a step-compensated bearing can be desirable. For example, in spindles, recirculatory flow is established in pocketed bearings which cause increased heat generation. If the pocket depth is increased to lessen the shear power in the pocket, then the Reynolds number increases which can lead to turbulent flow and greatly increased heat generation. Note that there is no recirculatory flow, and at high speeds the hydrodynamic wedge can be substantial which leads to increased load capacity. Should whirl instability become a problem, then the annulus could have a lobed profile ground into it to reduce whirl instability.

Step-Compensated Journal Bearing Design

For a step-compensated spindle, there will be circumferential leakage flow and hence the length to diameter ratio should not be too large. On the other hand, the larger the ratio of the pocket to the land width, the greater the load capacity of the bearing and the lower the friction power. It is assumed here that the decrease in load capacity due to circumferential leakage is offset by the hydrodynamic wedge.

The ideal gap ratio γ can allow for a reasonable amount of variation in the nominal gap, on the order of $\pm 10\%$, and this must be taken into consideration when choosing a bore diameter and bearing gap. Typically, the relative diameters of a large bore and mating shaft (e.g., 100 mm) can be held to 0.005 mm. Hence the bearing gap variation can be expected to be about 0.0025 mm and the nominal bearing gap between the lands and the bore should be 0.025 mm.

The force is calculated using Equation 9.2.63, which is per unit depth, where the depth is $Rd\theta$ and the bearing gap is a function of the angle θ. The radial force is represented by the equation:

$$F_{radial} = \frac{P_sD\ell}{2}\int_0^{2\pi} \frac{(1+\beta)\cos\theta d\theta}{\dfrac{\beta(1-\alpha\cos\theta)^3}{(\gamma-\alpha\cos\theta)^3}+1} \tag{9.2.64}$$

A closed form solution of this integral is very difficult to obtain, but fortunately it can be numerically evaluated on a spreadsheet using a simple macro.

Figure 9.2.18 shows spreadsheet output that gives the load capacity, stiffness, viscous shear power, and pocket Reynolds number of cylindrical step compensated journal bearings. The maximum load capacity is at $\alpha = 0.5$, where the nominal gap has decreased by one-half. For the initial application of a load, α is set at a value of 0.01 which represents the allowable radial error motion of the spindle (in this case where h is 15 microns, error motion = 0.15 microns). For an 80 mm diameter bearing, the initial stiffness is a respectable 362 N/μm (2 lbf/μin), and the temperature rise at 4000 rpm is a moderate 2.1 C^o. The temperature rise is 4-10 times less than a comparable rolling element bearing would experience.

Note that leakage flow can act to reduce the staic load capacity of this type of bearing by 50%. For air bearings of this type, the problem is overcome with the use of axial grooves instead of

a single large circumferential step region. Groove compensation was first developed by Professional Instruments Corp. (See Figure 9.3.49). Note that step compensated bearings do act very efficiently when used in a circular flat-pad configuration.

Enter bold #s					Results are for one journal only! Two are needed to make a spindle.							
Gap (m, microns)	1.5E-5	**15**										
L/D	**1**											
Speed (rpm, rad/sec)	**4000**	419										
Ps (N/m^2, psi, atm)	3.0E+6	441	**30**									
Viscosity	**0.001**	r	**97**	Cp	**4180**							
Gamma	**2**		Force (N)		Stiff (N/μm)		Power			Power	Power	Temp
Beta	**5**		alpha	alpha	alpha	alpha	Shear	Re #	Flow	Pump	Total	rise
D(m)	l	beta*l	**0.01**	**0.5**	**0.01**	**0.5**	Watts		lpm	Watts	Watts	C
0.050	0.0042	0.021	21	1081	141	146	33	314	2.4	121	156	0.9
0.055	0.0046	0.023	26	1308	171	176	49	346	2.4	121	172	1.0
0.060	0.0050	0.025	31	1557	204	210	69	377	2.4	121	193	1.2
0.065	0.0054	0.027	36	1827	239	247	96	408	2.4	121	220	1.3
0.070	0.0058	0.029	42	2119	277	286	129	440	2.4	121	254	1.5
0.075	0.0063	0.031	48	2433	318	328	170	471	2.4	121	296	1.8
0.080	0.0067	0.033	54	2768	362	373	220	503	2.4	121	347	2.1
0.085	0.0071	0.035	61	3125	409	422	280	534	2.4	121	409	2.5
0.090	0.0075	0.038	69	3503	458	473	352	565	2.4	121	482	2.9
0.095	0.0079	0.040	77	3903	511	527	437	597	2.4	121	569	3.4
0.100	0.0083	0.042	85	4325	566	583	536	628	2.4	121	671	4.1
0.105	0.0088	0.044	94	4768	624	643	651	660	2.4	121	788	4.8

Figure 9.2.18 Spreadsheet output for a step-compensated spindle journal bearing.

The Reynolds number indicates that the flow will be laminar, and the flow, total power, and temperature rise are all very modest. Hence a water, hydrostatic, step compensated, journal bearing represents a significant advancement in the state of the art of hydrostatic spindles. A design for this type of spindle is shown in Figure 9.2.19.

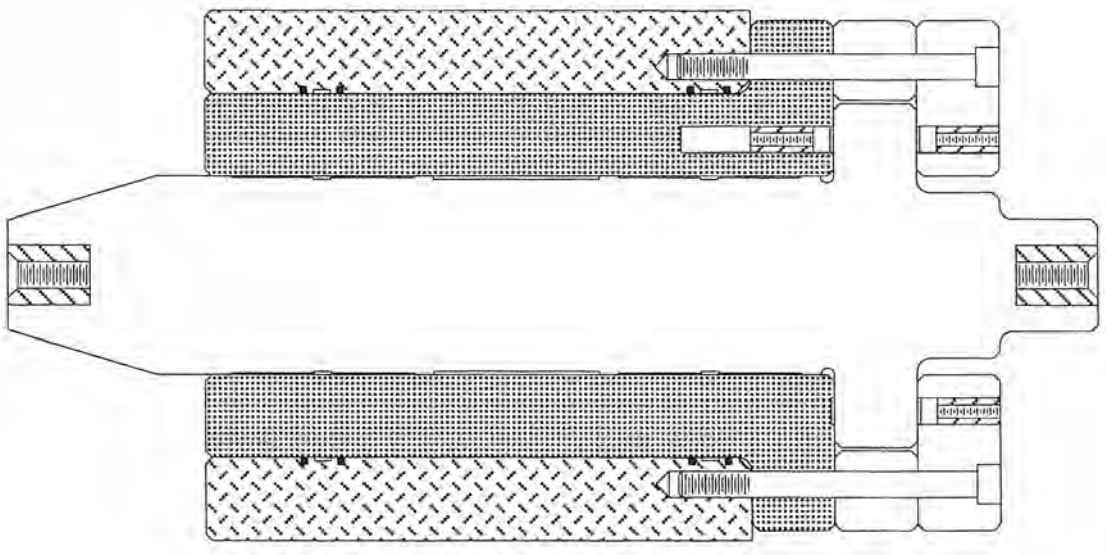


Figure 9.2.19 Cross section of a prototype ceramic step-compensated water hydrostatic bearing spindle. (Courtesy of Aesop, Inc.)

9.2.7 System Design Considerations

In order to become proficient at designing hydrostatic bearings, one should understand the theory so spreadsheets can be written and wisely used to help converge upon a design. The following rules

Temperature (^oC)	Actual viscosity(N-sec/m^2)	Computed viscosity(N-sec/m^2)	% error
20	0.0210	0.0213	-1.8
40	0.0100	0.0087	13.2
100	0.0025	0.0026	-5.5

Figure 9.2.20 Viscosity[26] of a typical "light weight" hydraulic fluid as a function of temperature: ISO 10 oil, μ (N-sec/m^2) = $1.05*T^{-1.3}$.

of thumb can be used to help minimize the number of unknown variables in the design process and to help guide the design of flat opposed pad bearings. Similar conclusions can be drawn for other bearing configurations.

Fluid pressure: Most hydrostatic bearings operate at about 3.5 MPa (500 psi) so low pressure fittings can be used and minimal pumping power will be generated. To obtain greater stiffness, one may want to increase the pressure to 7MPa, and in extreme cases one may want to use the maximum economically available pressure which is 21 MPa (3000 psi). The pump used should generate minimal noise. Gear pumps are best for most applications because they generate less noise than piston pumps and the noise they do generate is high frequency which is more easily attenuated. Various types of line filters can also be added to minimize noise in the hydraulic system.[23]

The fluid pressure will affect the stiffness of a hydraulic hose; in fact hose forces can be one of the biggest force inputs to an ultra precision machine tool carriage. Hose forces can be relieved by using a balanced design (i.e., two hoses mounted so their forces cancel) or a linear hydraulic commutator. There are two main designs of hydraulic commutators: (1) those that use a long tube, with a radial hole in the middle, that passes through the carriage, or (2) a pressurized groove on the bearing surface.[24] For the tube to be effective and not impart any friction forces onto the carriage, it must be supported with respect to the carriage by hydrostatic bearings. PTFE sliding bearings will not work for this application; the friction forces will be greater than the hose forces will be. The pressurized groove design will work well, but often the leakage flow will be very high which leads to too much pumping power and heat generation.

Fluid viscosity: A common hydraulic fluid used in hydrostatic bearings is shown in Figure 9.2.20. Note that the more viscous the oil, the greater the chance that a hydrodynamic wedge will build up during high speed moves causing angular and lateral errors in the motion of the body being supported.

The fluid viscosity also directly affects the amount of heat generated by the bearing. The total power generated in a hydrostatic bearing is a function of the viscous shear power and the pumping power. Figure 9.2.21 shows an algorithm that can be used to model the effect of temperature induced viscosity changes on hydrostatic bearing performance. As the temperature of the oil increases, the viscosity of the oil decreases and hence the flow increases. The increased flow then acts to carry more heat away from the bearing (e.g., a spindle)[27] and as a result thermal equilibrium is reached at a much lower temperature than would be predicted if the viscosity were assumed constant. The fluid viscosity decreases exponentially across the resistances in the bearing, but the simplified model here assumes that the resistance is ultimately based on a weighted average of the input and output viscosities. The weighting is small ($\mu_{avg} = (\mu_{in}+2\mu_{out})/2$) so the results on temperature rise will be conservative but the flow may be underestimated. Hence the pump chosen should have 25% greater capacity than predicted. Once the solution converges and the viscosity is used to find the temperature rise, it must be checked that the iterations' temperature rise limit was greater than the final temperature rise.

This algorithm would be used to help size the fluid restrictors using the weighted average of the viscosity of the oil flowing through them and the weighted average of the viscosity of the oil flowing out of the lands. In this manner, the fluid restrictors would be sized so their resistance would be equal to that of the pads and an "optimal" 50% pressure drop would be achieved.

Bearing gap: For ultra precision machines (e.g., diamond turning machines) where the carriage velocity is usually very low (0.1 m/s), one wants the gap as small as reasonably possible to

[23] See, for example, T. Viersma, Analysis, Synthesis and Design of Hydraulic Servosystems and Pipelines, Elsevier Science Publishers, Amsterdam, 1980.

[24] See U.S. Patent 4,865,465 by Sugita et al. of Citizen Watch Co.

[27] The affect of temperature on the oil's specific heat is considered second order here.

Temperature (^{o}C)	Actual viscosity(N-sec/m^2)	Computed viscosity(N-sec/m^2)	% error
20	0.0210	0.0213	-1.8
40	0.0100	0.0087	13.2
100	0.0025	0.0026	-5.5

Figure 9.2.21 Section (edited) of a macro used to iterate to find the viscosity and oil temperature at different points in the bearing.

manufacture. Typically the gap is on the order of about ten microns (0.0004 in.). A small gap helps to maximize stiffness while minimizing pressure and flow required. This in turn minimizes pumping power. Higher speed systems (1-2 m/s) may have gaps on the order of 20-50 μm (0.0008-0.002 in.) to help prevent one end land from running dry.

Pocket design: The pocket area should be large enough to lift the structure being supported when the bearing is first turned on and one side's pads' bearing gaps are essentially zero. As a rule of thumb, the product of the projected horizontal pocket area and the supply pressure should thus be at least twice the weight of carriage being supported. In order to evenly distribute the fluid to the lands, the pocket depth should be at least 10-20 times the bearing gap. The pocket corner radius should be one-quarter to one-half of the land width. Recall that $r_p = 0.4142l$ simplifies bearing calculations greatly although this is a moot point when one is using a spreadsheet for calculations. The fluid in the pocket and in the line leading to the pocket are compressible; however, in the quasi-steady state, any compression of the fluid that results in a change in gap will also cause an increase in the pressure differential which brings the bearing back to its equilibrium position.

Gap displacement: The gap displacement factor α should never be more than 0.5. It gives the precision machine designer a conservative feeling if one can comfortably obtain desired bearing performance, including acceptable flow rates, with $\alpha < 0.1$ at maximum load.

Land width: As mentioned earlier, a good choice for the land width is 25% of the bearing pad width. For high speed applications, the land width may have to be decreased to decrease the viscous shear forces and the chance of one side of the bearing running dry. The lands must be large enough to comfortably support the weight of the system when the supply pressure is turned off. Section 9.2.6 describes a design where the land width is optimized to minimize heat generation.

Pad length-to-width ratio: For linear bearing systems, it is advantageous to have the pads long so the keeper rail's cantilevered length can be minimized[28]; on the other hand, the longer the bearing pads the farther apart they must be spaced in order to provide pitch resistance. The longer and farther apart the pads are, the wider the carriage must be in order to have an acceptable bearing length to width spacing ratio to provide yaw resistance from moments generated by dynamic friction and cutting loads. The effect of the length-to-width ratio on the design of the keeper rail is of far greater importance than its effect on any other factors such as pumping power.

Number of pads and their configuration: The number of pads used can be minimized to minimize plumbing requirements; however one doesn't have to worry about loss of contact as is the case when designing with sliding or rolling bearings. Thus more pads can be used to support a large carriage which helps to minimizing the carriage weight by decreasing the amount of material needed to span the distance between two pads in order to maintain carriage stiffness. More pads also helps to average out errors in the bearing rails. How the pads are configured (e.g., rectangular, dovetail, etc.) is limited only by the designer's imagination and ability to analyze the design. It is not unusual to see a T-carriage with pads along all the surfaces that otherwise would be in contact.

Fixed flow restrictor design: Commonly used fixed flow resistance devices are the capillary tube or the flat edge pin. Because of the relatively small passageways through these devices, they must be installed in a manner that facilitates easy maintenance as shown in Figure 9.2.22. Note that this design assumes that there is a pressure tap in the pocket so the pocket pressure can be monitored. The pocket pressure must be initially checked in order to trim the restrictor, and during use, the pocket pressure must be monitored so the machine can be shut down if the filter clogs up. In addition, there are typically two bearing pad designs which one must consider, that in which the carriage's bearing pads surround the rail (e.g., a conventional T-carriage design) and that in which

[28] The keeper rails' dimensions and the bolts used to hold it in place must be chosen carefully. It does no good to have a very high stiffness bearing and a low stiffness structure. See the example of the design of a T-slide rail in Section 7.5.1.

the rails surround the carriage's bearing pads (e.g., a quasi-kinematic bearing pad arrangement). The case of journal bearings is equivalent to the former. All flow restrictors must be protected by a last chance screen filter. In all restrictor designs, modularity and ease of maintenance is a must.

For fixed compensation devices, since the fluid resistance of the bearing pad changes with the cube of the gap, it is a good idea to manufacture the bearing components, measure the gaps, and then size the flow restrictors for each pad. When a pin is used with the top edge ground off, typically the flow restrictor's diameter is set at a reasonable value such as 8 mm and the length is set at 20 mm, so all one has to do is solve the equation for the amount ε to be ground off the top which requires the use of a coarse-fine numerical iteration. The pin can be made from steel, brass, or aluminum. Steel dowel pins are available off-the-shelf ground to within 5-10 microns accuracy. In addition, reamers are also available matched to standard dowel pin sizes to ensure the dowel pins fit correctly without any material being shaved off as they are pressed in. The edge of a steel dowel pin is also easily ground off. Brass and aluminum machine well and their lower moduli decrease the required tolerance on the diameter; however, because they are relatively soft metals, there is a danger of shaving of metal occurring as they are pressed into the hole. The shavings can get into and clog the fluid passageway. If possible, it is perhaps best to use a stainless steel hypodermic tube as a capillary and solder or braze it into a threaded insert.

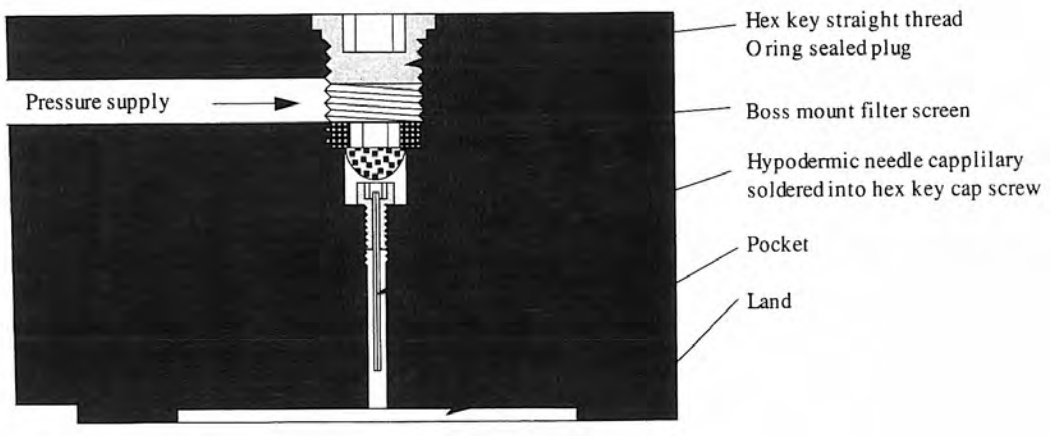


Figure 9.2.22 Restrictor mounting for flat bearing pads that surround the bearing rails.

Bearing rail design: It is very important to note that with most hydrostatic bearing arrangements that have cantilevered members (e.g., keeper rails), the deflection of the keeper rails caused by the pocket pressure can be on the order of the intended bearing gap if the keeper rails are sized solely to provide the desired carriage stiffness.[29] If the gap increases by a factor of 2, which is typical for a keeper rail sized for a sliding bearing application that is then used for a hydrostatic application, the flow out of the bearing will be eight times what was expected. The method discussed in Section 7.5.1 for determining rail deformations can be modified to determine the deflection profile of a cantilevered beam subject to a pressure over a portion of its length. In some cases, it makes sense to replicate the bearing to an exact sliding fit and let the pocket pressure expand the bearing to the desired gap.

9.3 Aerostatic Bearings[30]

Aerostatic bearings utilize a thin film of high-pressure air to support a load. Since air has a very low viscosity, bearings gaps need to be small, on the order of 1-10 μm. There are five basic types of aero-

[29] The principle of self-help (see Section 1.4.5) can be applied to this design problem to develop a bearing configuration whereby moments from bearing pressures in one direction cancel moments from a bearing in the other direction. However, this type of design requires that the bearing rails have excellent horizontal parallelism. See J. DeGast's U.S. Patent 3,583,774, June 8, 1971.

[30] This section was written in collaboration with Prof. Kenneth J. Stout, Lucas Professor and Head of the School of Manufacturing and Mechanical Engineering, Department of Engineering Production, Southwest Campus, Birmingham, B15 2TT England. Prof. Stout has written PC based software for easy application of the theory presented in this section.

static bearing geometries: single pad, opposed pad, journal, rotary thrust, and conical journal/thrust bearings similar to those configurations shown for hydrostatic bearings in Figure 9.2.1. All operate on the principle of supporting a load on a thin film of high-pressure air (typically 690 kPa) which flows continuously out of the bearing and into the atmosphere.

Aerostatic bearings have a relatively recent history. One of the first known publications on the subject was related to an experimental investigation by Willis[31] in 1828. His experimental work on thin air films identified the existence of a lubricating regime. The first experimental work on compressible fluid bearings was conducted by Hirn and published in 1854[32] relating to self-acting bearings. Hirn's work indicated the possibility of reducing friction in machinery by the use of a thin film of high-pressure air. A typical application reported by Girard[33] in 1863 concerned a water hydrostatic journal bearing devised for a railway propulsion system employing a linear impulse turbine. Considerable interest was generated in the subject of fluid film bearings in the middle to late nineteenth century, although most of this work was concerned with liquid bearings because a liquid's higher viscosity allowed larger bearing gaps and tolerances to be used.

Kingsbury[34] in 1897 reported experimental results on a 6 in. diameter gas journal bearing which used both air and nitrogen as the lubricant. These now famous experiments demonstrated that gas bearings were feasible and that the low friction these bearings afforded could be realized in practice. Many patents and designs followed in the early 1900s. Typical examples include those in 1904 by Westinghouse[35], who developed an air thrust bearing to support a vertical steam turbine, and in 1920 Abbott,[36] who patented a design for an externally pressurized air journal bearing. However, because of the close tolerances required to produce the gas bearing components, limited progress was made in their application in these early times. Theoretical methods were limited to approximate solutions of the Reynolds equation[37] for incompressible fluids. These solutions were improved by analytical work by Sommerfeld in 1904 (journal bearings) and by Mitchell in 1905 (thrust bearings). It was not until 1913 that an appropriate form of the Reynolds equation for compressible fluids was developed by Harrison.[38] The nonlinearity of the governing equations and the lack of adequate computing facilities for solving them resulted in very limited progress in providing theoretical predictions to supplement early experimental work.

It was not until the years following World War II that significant progress was made in further developing air bearing technology. This can be attributed to the needs of the nuclear power and defense industries, which expanded rapidly in the 1940s. Bearing systems were required to operate in exacting conditions of high speed, high stiffness, extremes of temperature, and low friction. As these demands could not be met by sliding contact, rolling element, or hydrostatic bearings, alternative types of bearings were considered. Preliminary experimental work on gas bearings indicated that these bearings could meet the operating requirements. The subsequent research that followed in the 1950s and 1960s, which was largely experimental, solved many early design problems and provided general design data. More recently, other industries have taken advantage of this technology. The most notable of these include manufacturing and turbomachinery.

In the late sixties, the vast majority of theoretical design methods for aerostatic bearings were based on approximate mathematical models for flow in the bearing clearances and control devices. Solutions were mainly restricted to steady-state operation, and conventional orifice flow equations were used for the control devices, based on an adiabatic process. Laminar flow in the clearance was described by a suitable form of the Reynolds equation and by assuming that flow into the bearing clearance was a line source to simplify the calculations. At this time it was more realistic to treat the calculation in this way rather than to attempt to solve the true case of discrete entry points. Much of

[31] Willis, Rev. R., "On the Pressure Produced on a Flat Surface when Opposed to a Stream of Air Issuing from an Orifice in a Plane Surface," Trans. Cambridge Philos. Soc. Vol. 3, No. 1, 1828, pp. 121–140.

[32] G. Hirn, "Study of the Principal Phenomena shown by Friction and of Various Methods of Determining the Viscosity of Lubricants," Bull. Soc. Ind., Mulhouse 26, No. 129, 1854, pp. 188-277 (in French).

[33] L. Girard, "Application des Surfaces Glissantes," Bachelier, Paris, 1863 (in French).

[34] A. Kingsbury, "Experiments with an Air Lubricated Bearing," J. Am. Soc. Nav. Eng., No. 9, 1897.

[35] G. Westinghouse, Vertical Fluid Pressure Turbine, U.S. Patent 745,400, 1904.

[36] W. Abbott, Device for Utilizing Fluid under Pressure for Lubricating Relatively Movable Elements, U.S. Patent, 1,185,571, 1920.

[37] O. Reynolds, "On the Theory of Lubricating and Its Application to 'Mr. Beauchamp Towers' Experiments, Including an Experimental Determination of the Viscosity of Olive Oil," Philos. Trans., Vol. 177, 1886, pp. 157–234.

[38] W. Harrison, "The Hydrodynamic Theory of Lubrication with Special Reference to Air as a Lubricant," Trans. Cambridge Philos. Soc., Vol. 22, 1913, pp. 39–54.

the design data which was originally produced has since been shown to be reasonably accurate for bearings operating at low eccentricity ratios where flow is predominantly axial.

Developments in computer power which enabled improved computational techniques to be employed have lead to finite difference and finite element analysis methods being applied to gas bearing analysis and as a result a considerable number of papers have resulted from many workers throughout the world. Much of the analysis conducted during the middle and late 1970s was directed at improving the accuracy of experimentation and supporting theoretical methods, including attempts to model the pressure loss in the orifices. In addition, other workers have analyzed the effects of manufacturing variations on performance, in an attempt to help guide the design engineer in the selection of manufacturing tolerances. The results produced enabled the designer to tolerance the critical dimensions of the bearing.

Knowledge of bearing characteristics has greatly increased in recent years, so the majority of new work will either lie in extending the range of bearing types covered by current design, investigating new applications which are steadily being introduced, or most important, the transfer of the existing knowledge into easily usable data for design engineers. There is a requirement for simple graphical design procedures for gas bearings which significantly reduce tedious calculation and provide the designer with information on the interaction of the effects of one design parameter on another. It is hoped that the design information and the procedures presented in this chapter, which have been extensively developed and used by Prof. Stout for designing systems for industry, will encourage more design engineers to use air bearings in relevant industrial applications.

9.3.1 General Properties of Externally Pressurized Gas Bearings[39]

The choice of liquid or gas lubrication depends upon the type of application for which the bearings are intended. For example, moderate loads and moderate stiffness at high speed will favor gas bearings, while the requirement for high load and high stiffness at moderate speeds favors hydrostatic bearings. Figure 9.3.1 compares liquid and gas film bearings. Hydrostatic bearings exhibit greater film stiffness than aerostatic bearings, usually by a factor of about 5, due mainly to the higher pressures of the lubricating medium. Important differences lie in the performance under dynamic loading. Hydrostatic bearings have superior damping characteristics compared to aerostatic bearings, which may be a critical feature in some applications. The major differences in performance relate to the effects of compressibility of the air film and the viscosity of the two lubricant mediums. To ensure that flow rates of aerostatic bearings are kept to realistic levels, bearing clearances must be smaller than those for hydrostatic bearings; hence the quality of manufacture must be higher and tolerances on size must be rigorously controlled, thereby increasing manufacturing costs. In addition, running costs tend to be higher for gas-lubricated bearings as more power is expended in compressing a gas than in raising an incompressible fluid to the same pressure.

More and more machines are using air bearings. Modular air bearings used on a CMMs are made in the form of a simply detachable shoe, held in place by a screwed ball which enables the air bearing pad to take up an alignment parallel to the location face. The ball enables the shoe to have 3 degrees of rotation, which ensures accurate alignment. The gap between the faces is on the order of 6 to 10 μm to ensure that air flow rates are modest.[40] Air bearings are now commonly used to support LVDT cores to improve reliability and reduce gauging forces. An air bearing live center has been developed recently for improving accuracy for gear checking. Other applications include profile projection equipment employing air bearing slides, rotary measuring tables and machine tool lead screw measuring heads. Modern developments in both manufacture and measurement are placing greater demands on accuracy and quality, particularly in the area of nanotechnology. It is therefore probable that there will be a greater need to utilize high quality spindles incorporating air bearings in the foreseeable future. As demand for spindle and carriage speeds and accuracy increases, greater use of aerostatic bearings in precision machine tools can be anticipated.

Speed and Acceleration Limits

Aerostatic bearings have only viscous friction associated with the air film layer being sheared during motion of the bearing. When using high-speed spindles (surface speeds greater than 10

[39] See, for example, M. Tawfik and K. Stout, "Characteristics of Slot Entry Hybrid Gas Bearings," 8th Int. Gas Bearing Symp., April 1981.

[40] The CMM discussed in Chapter 1 used porous graphite aerostatic bearings.

Liquid	Mixed phase	Gas
Capillary, orifice, slot, or diaphragm restricted	Liquid/steam when slot restricted	Porous, orifice, or slot restricted
Highly suitable for machine tools	If correctly designed, bearing will operate in either media	Highly suitable for textile machines and where contamination is unacceptable
High load capacity		Moderate load capacity: grinding spindles, diamond turning spindles, and instruments
Very high stiffness		High stiffness
Very high damping		Moderate-low damping
Low friction at low speed		Very low friction at all speeds

Figure 9.3.1 Capabilities and applications for externally pressurized bearings.

m/s), the bearing gap should be large enough to ensure that the friction power is less than twice the pumping power. In such situations the temperature rise due to friction within the bearing gap is offset by the refrigeration effects of the gas film as it expands in the gap after leaving the orifice.

Range of Motion

Linear motion aerostatic bearings can have as long a range of motion as it is possible to machine the slideways they rest on, which can be tens of meters. Angular motion aerostatic bearings are not rotation limited.

Applied Loads

Because aerostatic bearings distribute the load over such a large area, large loads can typically be supported. A hydrostatic bearing (3.5 MPa operation pressure) for the same-size pad area can typically support five times the load of an aerostatic bearing. High-pressure hydrostatic bearings (20 MPa) can support even larger loads but may suffer from significant heat generation. As shown in Figure 9.3.2, an estimate of an air bearing's load capacity can be found by multiplying the effective (projected) area by the entry pressure of the gas as it enters the bearing clearance. The effective projected area may be approximated as the area contained between the inlet orifices plus half the area outside the plane of orifices toward the edges of the bearing. The entry pressure is typically one-half the supply pressure. An estimate of the bearing stiffness is obtained by dividing the value above by half the no-load bearing gap. The load capacity of journal bearings is estimated in a similar manner to that described for flat pad bearings. The effective (projected) area of a journal bearing is approximately 0.3(L - a)D, where L is the length of bearing, D is the diameter of bearing, and a is the distance from the row of orifices to the outlet end of the bearing.

Accuracy

Overall accuracy of motion (e.g., straightness) of an aerostatic bearing depends on the accuracy of the components. An aerostatic bearing averages out local irregularities to make them perhaps the smoothest running of all bearings. Maximum peak-to-valley surface roughness of air bearing components, however, should not be greater than one-fourth of the bearing gap.[41] There is no wear-in period associated with aerostatic bearings and accuracy therefore depends on keeping the fluid flow restrictors clean and the pressure source free from pulsations. Aerostatic linear motion bearings have been built with submicron/meter accuracy.

Repeatability

Repeatability depends on the stability of the fluid supply system, including the pump and the devices that regulate the flow of air into the bearings (i.e., flow restrictors). If pneumatic hammer instability, pressure surges, and temperature changes can be avoided, aerostatic bearings can achieve submicron (perhaps soon to be nanometer) repeatability.

[41] For a Gaussian surface, peak to valley roughness R_y= 7 x average roughness R_a ($7R_a = R_y$).

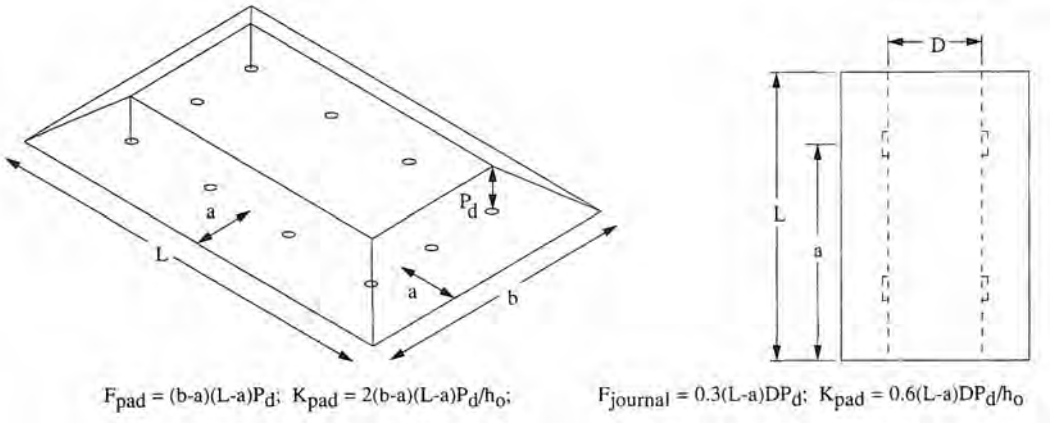


Figure 9.3.2 Formulas for estimating the load capacity (F) and stiffness (K) of aerostatic bearings.

Resolution

Since there is no stiction with aerostatic bearings, motion resolution of an object supported by them is virtually unlimited. However, the design of the actuator and control systems is not trivial since aerostatic bearings are not well damped.

Preload

Aerostatic bearings need to be preloaded in order to give them bidirectional stiffness. If a single aerostatic bearing is used to support a load, then as the load accelerates or forces act to increase the bearing gap, the bearing will have very little stiffness. Hence the opposed pad configuration is the most common one used in machine tools. It is also possible to preload a single pad aerostatic bearing with a vacuum pad which usually surrounds the pressurized pad region. The problem with this technique is that it is difficult to achieve less than a negative pressure of 10.5 atm, so the vacuum pad area has to be an order of magnitude larger than the area of the pressurized pad.

Stiffness

Aerostatic bearing stiffness can easily be in the 100 N/μm range. An estimate of the stiffness can be obtained by dividing the estimated load capacity by one-half of the nominal bearing gap. Aerostatic bearing stiffness is also not difficult to calculate accurately, and aerostatic bearings do not have the problem of loss of contact that sliding or rolling contact bearings can experience. Thus designing with aerostatic bearings is usually more deterministic than designing with contact bearings.

Vibration and Shock Resistance

Aerostatic bearings have very good shock and vibration resistance because there is no mechanical contact between moving parts. The air film collapses very quickly in the event of loss of pressure; hence aerostatic bearings should have a large backup reservoir and a control interlock to shut the machine down in the event of pump failure. In this manner touchdown and catastrophic failure can be avoided. In addition, if the bearing is not properly designed, the bearing itself can resonate as the air film alternately compresses and expands. This condition is known as *pneumatic hammer* and methods to avoid it are discussed later.

Damping Capability

The thin, low-viscosity air film in the bearing gap gives aerostatic bearings moderate to low damping capabilities in the normal and tangential bearing directions respectively.

Friction

Aerostatic bearings have absolutely zero static friction. Dynamic friction forces are negligible at low speeds (less than 2 m/s). The dynamic friction force on a aerostatic bearing is independent of the loads applied insofar as they do not change the bearing gap.

Thermal Performance

The viscosity of air is very low, so aerostatic spindle bearings are very tolerant of small changes in bearing clearance caused by viscous heating. Most hydrostatic bearings, on the other hand, are significantly less tolerant, and great emphasis must be given to optimizing hydrostatic spindle bearings in terms of friction power and pumping power for surface speeds greater than about 2 m/s. It is important to realize that air cools as it expands, and thus for a precision machine it is important to minimize flow and the resultant refrigeration effect.

Environmental Sensitivity

Because air is always flowing out of the bearing, aerostatic bearings are self-cleaning. The escaping air is not generally collected, so it is not necessary to keep chips and cutting fluid out of the bearing area, although it is still preferable to use bellows or sliding way covers to keep the bearings surfaces clean and prevent them from being damaged. In addition, unlike bearings with oil lubrication, there is no mess associated with air bearings.

Seal-Ability

Aerostatic linear bearings generally ride on rectangular or dovetail rails, so it is not difficult to seal them if required. Rotary motion aerostatic bearings usually do not need to be sealed.

Size and Configuration

Aerostatic bearings take up very little space themselves, but the plumbing requirements may be significant. It is generally desirable to have only one hose coming to a bearing, so the bearing itself may look like a block of Swiss cheese after all the drilling is done in order to get the air to all the different bearing pads. Kinematic pad arrangements are not necessary because the bearings act like constant force springs that fill whatever gap exists. As the equilibrium gap changes, the stiffness changes but stiffness will always be finite; thus aerostatic bearings are more forgiving of rail and carriage misalignments, as long as they do not cause the gap to change too much, which could cause loss of preload with sliding or rolling bearings.

Weight

Aerostatic bearings have moderate-to-high performance-to-weight ratios.

Support Equipment

The biggest drawback of aerostatic bearings is that they require a pump or connection to an air supply system. The system must be kept extremely clean to prevent foreign contamination from clogging the flow regulation devices, which typically are small orifices or slots. Typically, the air is filtered to 1 μm and dried with a desiccant to minimize condensation within the bearings as the air expands and cools.

Maintenance Requirements

Air cleanliness must be monitored and filters changed according to a fixed maintenance schedule. The air supply system should be inspected periodically for signs of contamination and the bearing rails for signs of wear that would result if a bearing pad's flow restrictor becomes clogged and the pad starved for air. Properly maintained and serviced, an air bearing should never experience any wear. There are many examples which show no wear after 10 years of continuous operation.

Material Compatibility

Aerostatic bearings are compatible with virtually all materials, and the presence of a small bearing gap usually leaves ample room for differential thermal expansion between components; however, one needs to determine the order of this gap change and make sure that it does not alter the bearing's performance too much. If the gap opens up too much, then the bearing will be starved for air and a loss of stiffness will occur. If the gap decreases, little change in stiffness would be seen, but eventually, further reductions in gap would cause the stiffness to deteriorate because the inlet flow restrictors would be improperly sized for the prevailing gap.

Required Life

Aerostatic bearings, whose air supply systems are maintained, can have essentially infinite life.

Availability, Designability and Manufacturability

Aerostatic bearing spindles and linear bearings are available as off-the-shelf items. It is not difficult to design and build custom aerostatic bearings if basic design rules are followed and one has some experience. The critical parameter in manufacturing aerostatic bearings is maintaining proper orifice and clearance dimensions. For orifices, one can use watchmaker's jeweled bearings. The orifice length should not be greater than four times the diameter, and the edges of the orifice should be as sharp as possible.

There are four basic types of gas bearings: aerodynamic, squeeze film, aerostatic, and hybrid. They are most commonly used in rotary applications, although linear bearing designs exist as well. The aerodynamic bearing is often called self-acting because it generates its pressure within the gas film by the mechanism of velocity-induced viscous shearing in a converging film, a process similar to that found in hydrodynamic bearings. Unfortunately, the film pressures generated are relatively low. The advantage of this type of bearing is that it is entirely self-contained and is independent of any external source for gas supply. Aerodynamic bearings are used, for example, to support the read/write heads on computer disk drives. The squeeze film bearing is also independent of an external supply source, but this type of bearing, due to the poor squeeze properties of gas films, has not been found to be generally practical as a solution to a wide variety of engineering problems, although laboratory tests have demonstrated experimental feasibility. The aerostatic bearing, or the externally pressurized gas bearing as it is often called, has its pressure in the gas film generated from an external source (compressor). A hybrid bearing combines both the aerostatic and the aerodynamic contributions to load. In practice, the combined performance obtained does not greatly enhance the purely aerostatic component of load for speeds used in most applications.

The most widely used bearing types are journal bearings for rotating shafts, and thrust bearings, either rectangular or circular form, often found in machine tool slideways, or of an annular type found in precision spindle assemblies. The journal bearing normally requires both axial and radial location, and this requirement may be met by separate journal and thrust bearings, conical or spherical bearings. Conical bearings offer the advantage that separate thrust faces are unnecessary, although problems may occur because a load in one direction will affect displacement and load capacity in the orthogonal direction. The spherical configuration has the useful property of allowing limited misalignment to occur, but unfortunately the ratio of radial to axial load capacity is poor.

Cost

The principal costs associated with aerostatic bearings are those of machining all the air supply passages and machining long straight rails or very round bores with close tolerances. The cost of maintaining the air supply system should also be considered. An air filter dryer unit, which can provide air for dozens of air bearings, can cost on the order of $650.

9.3.2 General Operating Characteristics[42]

As shown in Figure 9.3.3, gas at supply pressure P_o is admitted into the bearing clearance by a restricting device (often an orifice) which reduces the gas pressure from P_o (supply pressure) to P_d (orifice downstream pressure). Downstream of the restrictor, the gas flows through the bearing clearance, where its pressure reduces further, to atmospheric pressure P_a at the outlet to the bearing. Changes in film clearance modify the restriction over the bearing lands and affect the orifice downstream pressure P_d, which in turn affects load capacity. A smaller clearance leads to a higher P_d for a given restrictor and consequently a higher load capacity. Increasing the film clearance has the opposite effect. Therefore an optimum condition exists at which maximum film stiffness occurs where the rate of change of load when divided by rate of change of clearance is a maximum. One of the major objectives when designing aerostatic bearings is to choose the restrictor dimensions in conjunction with film clearance to achieve the optimum condition for stiffness. Incorrect component sizing will result in low bearing stiffness and inefficient operation, although often increased load capacity may be achieved when the bearing departs from the optimum stiffness condition. Hence, designs which seek maximum stiffness may require some load capacity to be sacrificed.

Flow Through Inlet Orifices

[42] For the remainder of this chapter only aerostatic bearings that use a discrete number of flow restrictors will be considered in detail.

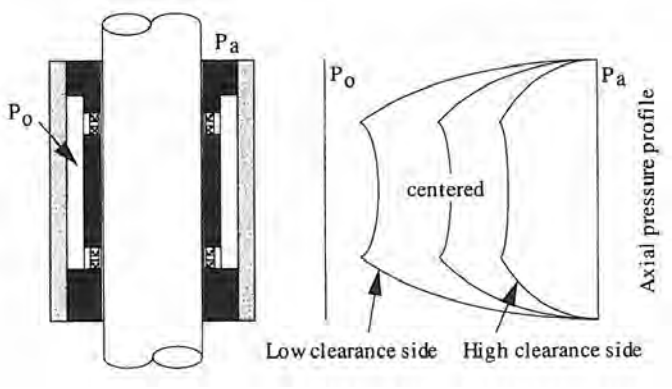


Figure 9.3.3 Principle of aerostatic bearing operation.

Figure 9.3.4 illustrates typical designs of pocketed and annular orifices, both of which are turbulent flow devices. The pressure drop that occurs is due to the acceleration of the gas as it expands. The pocketed orifice design shown can be manufactured by using a pierced insert that is press fit to an appropriate depth below the surface of the bearing bore. The annular (inherently compensated) orifice is made by simply drilling through the walls of the bearing with an appropriately sized drill.

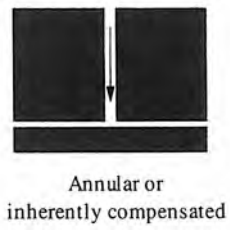


Figure 9.3.4 Typical orifice designs.

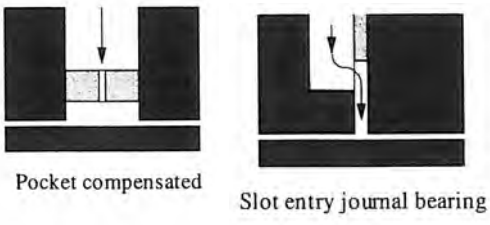


Figure 9.3.5 Slot-entry journal bearing.

Flow Through Inlet Slot

As shown in Figure 9.3.5, an inlet slot can be formed by a thin shim which is fitted between two adjacent sections of the bearing. This type of inlet restriction provides a laminar flow device, and has the effect of marginally increasing stiffness at higher operating eccentricity ratios. Manufacture of the shim and assurance of sealing between the lands can be difficult. An alternative construction which produces laminar inlet flow can be obtained by machining one or more flats on a round plug fitted into a reamed hole.

Flow Along Axial Grooves

Professional Instruments developed a method whereby air at full line pressure is introduced between two bearing surfaces by means of specially formed grooves in one of the bearing surfaces.[43] The grooves control the pressure gradient while maintaining a high degree of stability because squeeze film effects are maximized by the long flow paths and the absence of upstream restrictors (e.g., orifices). Groove-compensated air bearings work extremely well and are often considered the near equals of porous air bearings. The amount of design information available on this type of bearing, however, is limited, and if a company wanted to develop in-house capability for designing other than orifice-compensated air bearings, it would probably be most worthwhile to undertake a porous air bearing development effort.

Flow Through Porous Media

The ideal design would supply air evenly to the entire bearing pad. This can be accomplished with the use of a porous media (e.g., graphite). Porous air bearings can have substantially greater stiffness and load capacity than other types of externally pressurized aerostatic bearings. In addition,

[43] U.S. Patent 3,305,282, granted to Professional Instruments Co., 4601 Highway 7, Minneapolis, MN 55416, (612) 927-4494.

porous air bearings are generally not subject to pneumatic instabilities (hammer) and porous graphite air bearings are very crash resistant. The hemispherical spindle described in Figure 9.3.54, which was developed at Oak Ridge and is now manufactured by several different companies, indeed exhibits these characteristics, and many such spindles there have been in use in an abusive environment for decades. However, accurately characterizing the flow through porous media and controlling the manufacturing process (application of a lacquer) that tunes the viscosity are both challenging tasks. A discussion of these tasks is beyond the scope of this book, but they are well documented.[44]

In some bearings, inlet restriction is provided by a porous plug inserted in the body of the bearing. The porous material has a matrix of pores throughout its structure, each small in size. By machining the plug to the appropriate length, the flow resistance through the porous plug should match the flow resistance through the segment of the bearing being supplied by the plug. Flow through a porous plug usually obeys Darcey's equation, but is difficult to control in practice.

Static Instability

The phenomenon of pneumatic hammer instability is caused by the compressibility of gases and the consequent delay between bearing clearance changes and the response to this change through variations in pressure in the orifice pocket. If the pocket volume is too large, and the time delay is too long, the resulting pressure in the pocket may increase excessively. This causes the clearance between the bearing surfaces to be increased. This increase in clearance reduces the pressure in the pocket, and hence the clearance between the mating surfaces will reduce again. If this occurs too slowly, overcompensation will occur.

The reduction in clearance between the two surfaces increases gas flow resistance, and this increase causes the pressure in the gas to build up. As a consequence of pressure increasing, the clearance between the two mating surfaces gets larger and the cycle described above repeats. The instability which occurs is due to very low damping in a compressible fluid, and is termed *pneumatic hammer instability* or simply *pneumatic hammer*. Pneumatic hammer is most often found in association with thrust faces. This is because the volume of gas within the orifice pockets is often large when compared with the volume of gas within the lands. There are two ways of overcoming this problem: one is to reduce the pocket depth and diameter, which reduces the total pocket volume. The second method is to use inherently compensated orifices and accept reduced load capacity.

Types of Aerostatic Bearings

The geometry of a circular thrust bearing is shown in Figure 9.3.6. The pressurized air enters the bearing through the central feed hole and the compressible fluid disperses radially to the outlet at the outer circumference of the bearing. This form of bearing has virtually no resistance to tilting. Several of these bearings can be used to support a structure and keep it from tilting or other bearings (e.g., radial support bearings) can be used. This type of bearing does not generally suffer from pneumatic hammer since the pocket volume is normally small when compared with the land volume.

The geometry of an annular thrust bearing is shown in Figure 9.3.7 and is more widely used in linear and rotary applications because of its superior tilt resistance. It is also often used as a thrust face in conjunction with a journal bearing. The major problem with annular thrust bearings is they they may suffer from pneumatic hammer if they have not been designed correctly. Inherently compensated annular bearings are often more stable, although they have less load capacity and stiffness than pocketed orifice-compensated bearings of the same size. Inherently compensated annular bearings may be preferred for narrow land bearings where the land area is small and the number of supply orifices is comparatively large.

Rectangular flat-pad bearings are normally found in slideway assemblies, for example those found on CMMs. Given an adequate number of inlet restrictors, these bearings may have any length-to-width ratio to suit the application's load capacity and stiffness requirements. Typical rectangular bearing geometries are shown in Figure 9.3.8. Rectangular bearings are often fitted with two rows of inlet devices to provide tilt stiffness in operation and to increase load efficiency and bearing stiffness. Opposed pad configurations are often convenient when the loading on the bearing may be bidirectional. Opposing pads can be preloaded against each other so that bearing stiffness can typically be increased to approximately double that achieved with single-acting bearings.

[44] See W. H. Rasnick et al., "Porous Graphite Air-Bearing Components as Applied to Machine Tools," SME Tech. Report MRR74-02.

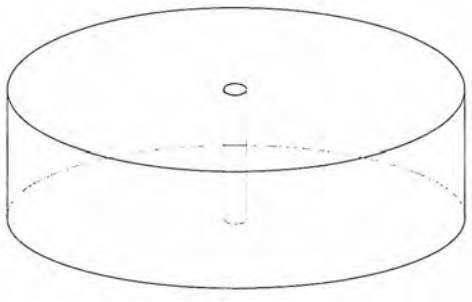


Figure 9.3.6 Circular thrust bearing.

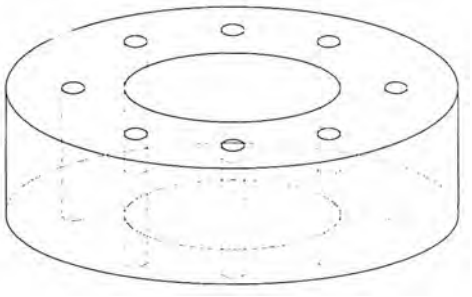


Figure 9.3.7 Annular thrust bearing.

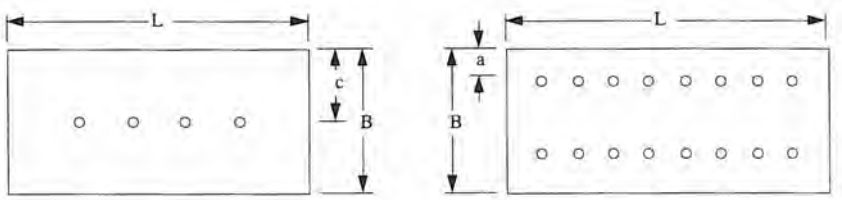


Figure 9.3.8 Double-entry (a/b = 0.25) and single-entry (c/B = 0.5) thrust bearings.

A very common type of externally pressurized gas bearing is the journal bearing, which usually contains two rows of entry sources around the bearing circumference. A typical journal bearing configuration was shown in Figure 9.3.3. Generally, very short journal bearings should be avoided (L/D $<$ 0.5), since they are difficult to design properly. Most journal bearings require axial constraint, and a common method is to employ a thrust flange centered between two annular pads. Normally, both pads are arranged at one end of the bearing, usually at the end which must be most accurately located. This helps to eliminate axial thermal expansion errors. Other forms of combined journal and thrust bearings are possible and include conical, partial spherical, and Yate-type configurations as shown in Figure 9.3.9. Although these latter configurations are possible as alternatives to the journal with thrust flange type, they are more complex to manufacture. It is therefore necessary to decide whether the advantages they may offer in terms of lower gas flow rate is worth the extra cost and manufacturing complexity required to make these bearings.

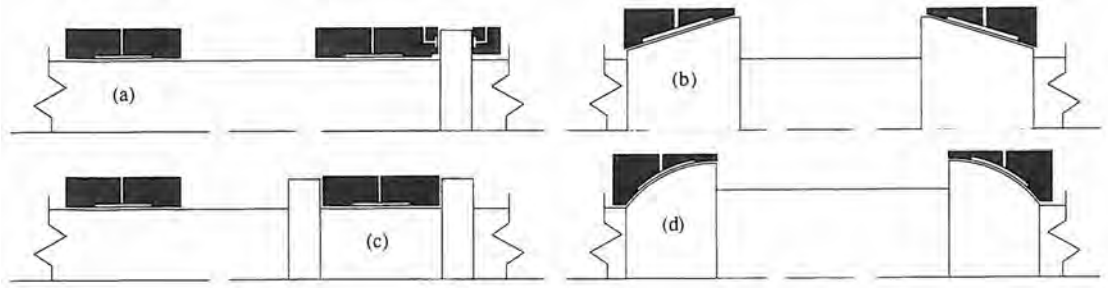


Figure 9.3.9 Spindle designs for externally pressurized (gas or liquid) fluid bearings: (a) cylindrical journals with annular thrust bearings; (b) conical journal bearings; (c) Yates bearing; (d) hemispherical journals.

9.3.3 Analysis of Orifice Compensated Bearings

As discussed earlier, some sort of resistance must exist at each inlet source between the pressure source and the bearing pads to allow a pressure differential between opposed pads to form. This gives the bearing stiffness and is sometimes referred to as *compensation*. Orifice bearings, which

are made by drilling a hole through the surface of the bearing to an air supply reservoir, are inherently compensated. Pocketed compensated bearings are made by recessing the area around the orifice to give a larger area where the pressure is equal to the entry pressure, hence allowing the bearing to support larger loads. However, pocketed orifice bearings are more likely to experience pneumatic hammer. As discussed previously, orifices are not the only means of providing an inlet restrictor, but they have the advantage in that they are compact and normally easy to install. Orifices may be provided by a jewel with a hole through its center, typically used in watches and clocks, or by producing a specially drilled plug which is inserted into a reamed hole in the bearing surface. It can be shown that bearings that have pocketed orifices yield greater stiffness than those having inherently compensated orifices by a factor of up to 1.5.

Inherent Compensation

The gas flow path through a typical inherently compensated restrictor[45] is shown in Figure 9.3.10 along with its resultant experimental pressure profile. The pressure profile has been obtained by measurements conducted in the clearance of a journal bearing by traversing a pressure transducer across an orifice. This enables pressure variations over small distances to be identified, which is particularly significant in the feeding region around the restrictor. Immediately under the orifice feeding area, the supply pressure is recorded. This pressure corresponds to the supply conditions in the inlet device. At the inlet to the bearing film, the gas accelerates through the curtain area $\pi d_f h$ from the supply pressure P_o, and as a consequence, its static pressure P_t is reduced. As the flow continues, the gas recovers some of its dynamic pressure and viscous losses prevail to the atmospheric boundary.

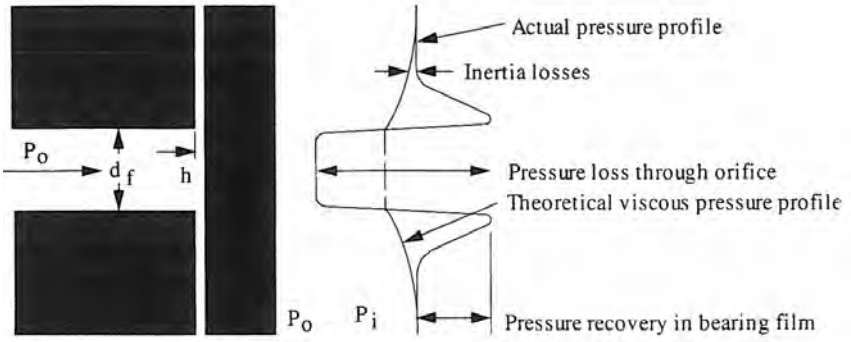


Figure 9.3.10 Pressure losses local to an inherently compensated restrictor.

Assuming that the expansion of the gas is an isentropic process (i.e., reversible adiabatic) the mass flow is given by

$$\dot{m}_o = \pi d_f h\, C_d\, P_o \left[\frac{2\nu}{(\nu-1)\,RT} \left\{ \left(\frac{P_t}{P_o}\right)^{\frac{2}{\nu}} - \left(\frac{P_t}{P_o}\right)^{\nu+\frac{1}{\nu}} \right\} \right]^{1/2} \tag{9.3.1}$$

where for choked flow[46]

$$\frac{P_t}{P_o} = \left[\frac{2}{\nu+1}\right] \frac{\nu}{\nu-1} \tag{9.3.2}$$

This equation relates the mass flow rate to the orifice dimensions, supply pressure P_o, Boltzmann's constant R, and temperature T, and the throat pressure ratio P_t/P_o. The value of the discharge coefficient C_d in the flow equation largely accounts for the vena contracta effect at the entrance to the bearing film. The values used for C_d are dependent on such effects as the sharpness of the corner around which the flow passes. From experimentation, it has been found that typical values of C_d are

[45] Alternatively called an annular compensated restrictor or orifice, which is in practice a small hole drilled into the surface of the bearing.

[46] Choked conditions occur in the flow of air through an orifice when the reduction in diameter causes the flow to accelerate to a maximum (reaches supersonic velocity). Further reduction in the orifice diameter will not allow the air to accelerate any further. Hence the flow becomes choked.

on the order of 0.8. Hence the choked mass flow rate in kg/s per inlet for air at 20°C is given by

$$\dot{m}_o = 7.48 \times 10^{-4} C_d\, d_f h \frac{P_o}{P_a} \tag{9.3.3}$$

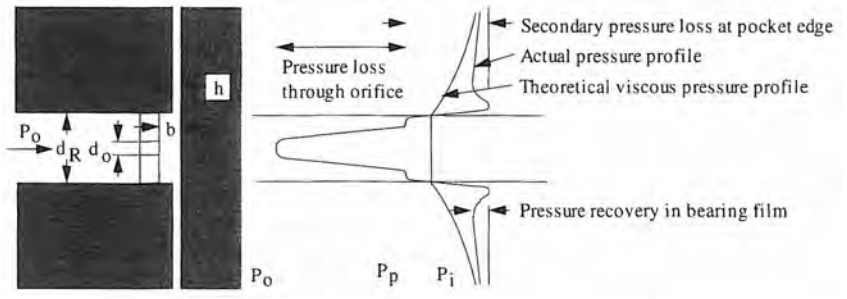


Figure 9.3.11 Pressure losses local to a pocketed orifice.

Pocketed Compensation

The gas flow path through a typical pocket compensated restrictor is shown in Figure 9.3.11. Measured and predicted values generally agree within about 10%.[47] Initially, the gas flows through the orifice flow area $\pi d_o^2/4$ and attains a pocket pressure P_p. At the edge of the pocket the gas further expands through a secondary restrictor given by a curtain are $\pi d_R h$ (where d_R = diameter of curtain area), where a vena contracta occurs as the gas enters the bearing film. The gas subsequently recovers some of its dynamic pressure as the velocity reduces and eventually viscous losses prevail in the bearing clearance. The isentropic flow equation can be used in conjunction with a discharge coefficient C_d to account for the departure from idealized flow:

$$\dot{m} = \frac{C_d\, \pi d_o^2\, P_o}{4\sqrt{1+\delta_L^2}} \left[\frac{2\nu}{(\nu-1)RT} \left\{ \left(\frac{P_p}{P_o}\right)^{\frac{2}{\nu}} - \left(\frac{P_p}{P_o}\right)^{\frac{\nu+1}{\nu}} \right\} \right]^{1/2} \tag{9.3.4}$$

where

$$\frac{P_p}{P_o} = \left[\frac{2}{\nu+1}\right]^{\frac{\nu}{\nu+1}} \tag{9.3.5}$$

and

$$\delta_L = \frac{d_o^2}{4 d_R\, h} \tag{9.3.6}$$

The inherent compensation factor δ_L accounts for the effect of the resistances in series on the mass flow rate. The mass flow rate above is expressed in terms of pocket pressure P_p rather than throat pressure P_t. The throat pressure P_t cannot be easily measured, while P_p is directly obtainable from measured pressure profiles. The effect of expressing pressures in these terms is that the values of C_d, which can be obtained from experimentation, include the effects of pressure recovery within the pocket.

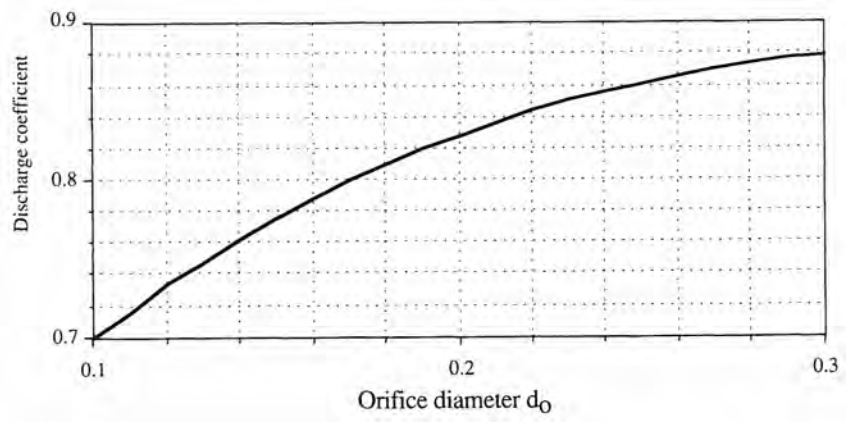
Figure 9.3.12 shows a curve for the choked (free jetting) discharge coefficient $C_d{}^*$. The results relate to orifices produced by the use of watchmaker's ruby jewels free-jetting to atmospheric conditions.[48] The variation of $C_d{}^*$ for a given orifice diameter may vary by 15% and are often due to deviations in orifice dimensions from the nominal values stated by the manufacturers. Figure 9.3.13 shows the dependence of the discharge coefficient C_d with P_p/P_o compared with choked flow. It can be seen that as P_p/P_o increases from the choked flow case, the coefficient of discharge decreases. Similar trends have been reported by Marsh et al.[49] and Markho et al.[50]

[47] E. Pink, "The Application of Complex Potential Theory to Externally Pressurized Gas Lubricated Bearings," Proc. 8th Int. Gas Bearing Symp., Leicester Polytechnic, England, 1981.

[48] E. Pink and K. Stout, "Design Procedures for Orifices Compensated Gas Journal Bearings Based on Experimental Data," Tribol. Int., Feb. 1978, pp. 63–75.

[49] H. Marsh et al., "The Flow Characteristics of Small Orifices Used in Externally Pressurized Bearings," Proc. 7th Int. Gas Bearing Symp., University of Cambridge, England, 1976.

[50] P. Markho et al., Discussion of Ref. 3.3, pp. 41–42, Proc. 7th Int. Gas Bearing Symp., University of Cambridge, England, 1976.



Orifice diameter d_o

Figure 9.3.12 Experimental C_d* at choked conditions for ruby jewel orifices: $Cd^* = 0.4540 + 3.1762d_o - 7.8571d_o^2 + 6.6666d_o^3$.

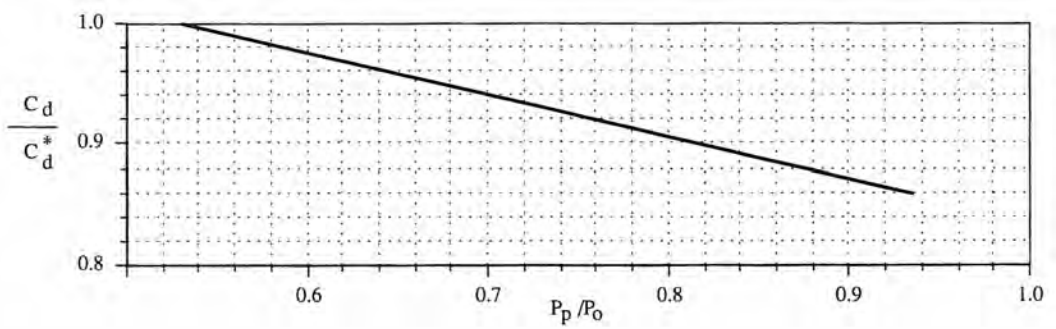


Figure 9.3.13 Correlation of C_d with pressure ratio based on recovered conditions in the pocket.

When brass bushings are used as orifices, severe distortion of the hole diameter can occur as the orifices are pressed into the bearing.[51] Experiments conducted indicate that the coefficient of discharge can vary by a factor of 2 for the same nominal orifice sizes. A possible cause of this variation is attributed to small geometric differences and/or surface roughness effects. The values of C_d^* obtained through experimentation are typically in the range 0.5-0.92, with the trend of increasing C_d^* with increased orifice diameter. It must be noted, however, that the problems encountered due to the deformation of hole geometries cannot occur with watchmaker's jewels. This is because jewels, unlike brass bushings, are extremely hard and thus do not deform appreciably or burr. If excessive stresses are encountered, the jewels crack or shatter. In addition, as jewels are translucent, any cracking that occurs is easily detectable by the use of a microscope and a strong light source. For a pocketed compensation bearing, the choked flow rate in kg/s per inlet for air at 20°C is thus given by

$$\dot{m} = 1.87 \times 10^{-4} C_d d_o^2 \frac{P_o}{P_a} \tag{9.3.7}$$

Flow Between Parallel Plates

After the air enters the bearing through the orifice, it can be modeled as flow between parallel plates because predominantly viscous flow exists in this region. It may also be applied to laminar inlet restrictors which commonly have the form of narrow rectangular passages typically found in slot-entry bearings. In either case, the assumptions used to determine the flow are:

[51] H. Marsh et al., "The Flow Characteristics of Small Orifices used in Externally Pressurized Bearings," Proc. 7th Int. Gas Bearing Symp., University of Cambridge, England, 1976.

1. The flow throughout is assumed to be purely viscous with no slip at the boundaries. This means that gas inertia is neglected, which is a valid assumption for situations involving low Reynolds numbers.
2. The gas flows at constant temperature (i.e., isothermal conditions). This implies that the heat generated by viscous shearing is efficiently dissipated. Since typical bearings used in practice employ both small gas film clearances and metallic materials are used for the bearing faces, this assumption is generally valid.
3. Pressure is constant across the height of the bearing film. The pressure actually varies only marginally across the film for the small clearances used in practice.

Consider the flow element shown in Figure 9.3.14. Equating forces gives an expression for the shear stress:

$$\tau = \frac{dP}{dy} z \tag{9.3.8}$$

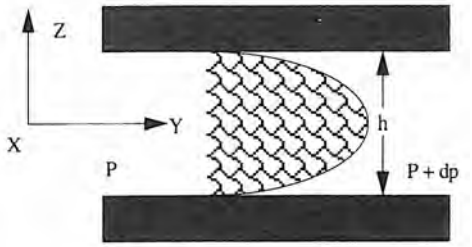


Figure 9.3.14 Pressure induced flow.

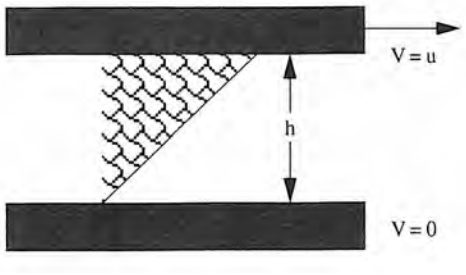


Figure 9.3.15 Velocity induced flow.

With the assumptions above, the Navier Stokes equations reduce to

$$dv = \frac{-dP}{dy} \frac{z}{\eta} dz \tag{9.3.9}$$

Integrating across the field with v = 0 at z ± h/2 gives

$$v = \frac{1}{2\eta} \frac{dP}{dy} \left[\frac{h^2}{4} - z^2 \right] \tag{9.3.10}$$

This gives a parabolic velocity distribution. The mean velocity is obtained by integrating the velocity profile and dividing by the air film thickness h to give

$$v_{mean} = \left(\frac{h^2}{12\eta} \right) \left(\frac{dP}{dy} \right) \tag{9.3.11}$$

The mass flow rate is simply the product of the area, density, and mean velocity:

$$\dot{m} = A\rho\, v_{mean} \tag{9.3.12}$$

Substituting Equation 9.3.11 into 9.3.12, assuming that the flow is isothermal (i.e., ρ = P/RT), and substituting A = hx yields

$$\dot{m} = \left(\frac{Ph^3}{12\eta RT} \right) \left(\frac{dP}{dy} \right) x \tag{9.3.13}$$

The equation above expresses the mass flow rate in terms of the pressure gradient in the direction of flow, the gas properties, and the flow channel dimensions. It can be applied to any viscous flow resistance, such as that provided by the bearing clearance.

Equation 9.3.13 can also be applied to inlet restrictors which have the form of narrow rectangular passages typically found in slot-entry bearings. For these bearings, Equation 9.3.13 can be used directly to account for the pressure drop, to give

$$P_o^2 - P_d^2 = \frac{24\eta\, \dot{m}\, RT\ell}{ah^3} \tag{9.3.14}$$

where a, h, and l are the width, thickness, and length of the slot, respectively.

In the case of one-dimensional flow through a single orificed bearing, the boundary conditions of $P = P_d$ at $y = 0$ and $P = P_a$ at $y = Y$ are introduced to give

$$P_d^2 - P_a^2 = \frac{12\eta\,\dot{m}\,RT}{\pi h^3}\,\xi \tag{9.3.15}$$

The shape factor ξ defines how the flow length dy varies with flow width x

$$\xi = \int_{y_1}^{y_2} \frac{2\pi}{x}\,dy \tag{9.3.16}$$

Note that if the bearing was a circular thrust bearing $x = 2\pi y$. The total mass flow rate through the bearing clearance from the inlet planes to the atmospheric boundaries is given by

$$G = \left[\left(\frac{P_d}{P_o}\right)^2 - \left(\frac{P_a}{P_o}\right)^2\right] \frac{\pi h^3 P_o^2}{12\eta RT\xi} \tag{9.3.17}$$

Velocity Induced Flow

Velocity induced flow occurs where the bearing surfaces are moving relative to each other as shown in Figure 9.3.15. The resulting velocity profile is linear and the velocity induced flow is given by:

$$\dot{m} = A\rho v = \frac{hP_{mean}\,U}{2RT} \tag{9.3.18}$$

Feeding Parameter $\Lambda_s\xi$ and Pressure Ratio K_{go}

Equating mass flow rates through N orifices (Equations 9.3.1 and 9.3.4 for inherently and pocketed compensated bearings, respectively) for flow through the bearing clearance with the assumption of no dispersion (Equation 9.3.7) yields

$$\Lambda_s\,\xi C_d \left\{\frac{2\nu}{\nu-1}\left[\left(\frac{P_t}{P_o}\right)^{\frac{2}{\nu}} - \left(\frac{P_t}{P_o}\right)^{\frac{\nu+1}{\nu}}\right]\right\}^{\frac{1}{2}} = \left(\frac{P_d}{P_o}\right)^2 - \left(\frac{P_a}{P_o}\right)^2 \tag{9.3.19}$$

where

$$\Lambda_s\,\xi = \frac{6\eta\sqrt{RT}\,Nd_0^2}{4P_o h_o^3\sqrt{1+\delta_o^2}}\,\xi \qquad \text{(pocketed orifices)} \tag{9.3.20a}$$

$$\Lambda_s\,\xi = \frac{6\eta\sqrt{RT}\,Nd_f}{P_o h_o^2}\,\xi \qquad \text{(inherently compensated)} \tag{9.3.20b}$$

and

$$\frac{P_t}{P_o} = \left\{\frac{P_d/P_o}{\left[\frac{2}{\nu+1}\right]^{\frac{\nu}{\nu-1}}}\right\} \tag{9.3.21}$$

The term $\Lambda_s\xi$ is commonly known as the feeding parameter and has been defined by MTI.[52] This parameter gives a measure of the ratio of the pressure drop through the orifices against the pressure drop through the bearing clearance at concentric conditions. The value indicates the degree of matching of the flow resistance. An alternative method is to express the pressure drop in terms of the design pressure ratio K_{go}, where

$$K_{go} = \frac{P_d - P_a}{P_o - P_a} \tag{9.3.22}$$

For a given bearing design, the calculation of $\Lambda_s\xi$ is comparatively simpler than K_{go}. Also, depending upon the restrictor loss analysis employed, differing values of K_{go} can be attributed to the same

[52] D. Wilcock, Design of Gas Bearings, Mechanical Technology Inc., Lotham, NY, 1967.

bearing. In addition, the effect of dispersion further complicates the analysis. These conditions make the selection of K_{go} for a particular bearing design an arbitrary matter. By contrast $\Lambda_s\xi$ is defined directly from the bearing geometry and is therefore a purely independent variable. The relationship between $\Lambda_s\xi$ and K_{go} is shown in Figure 9.3.16. It can be seen that for a wide range of P_d/P_o ratios, the value of K_{go} varies only slightly for a given value of $\Lambda_s\xi$.

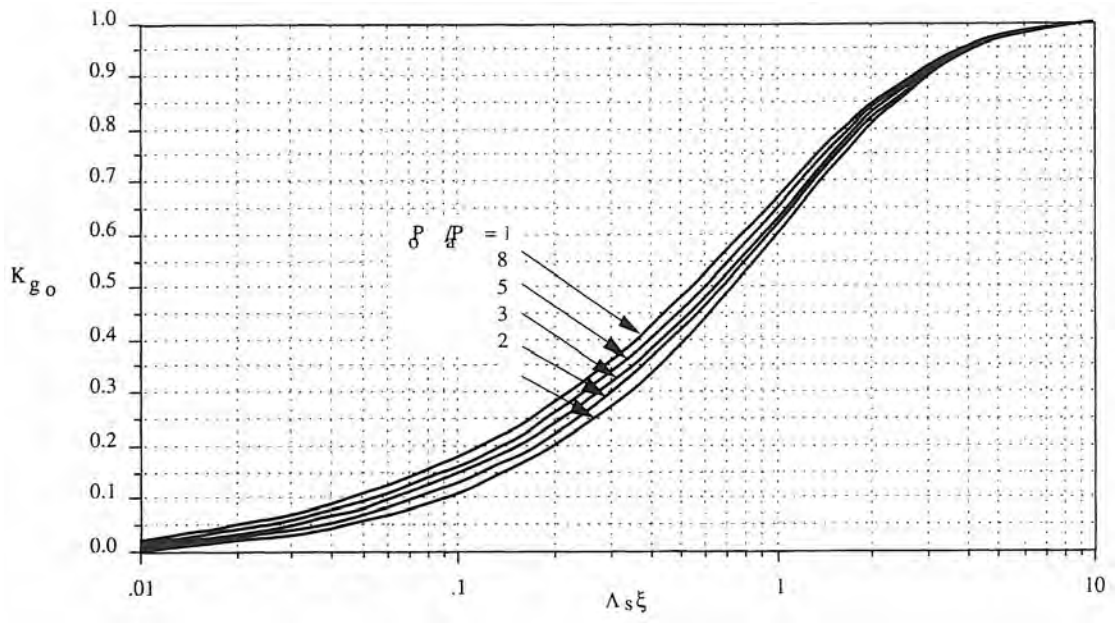


Figure 9.3.16 Relationship of feeding parameters with K_{go} where $C_d = 0.8$ and $\nu = 1.4$.

Orifice Design

Nondimensionalizing mass flow rate given by Equation 9.3.7 to the following form:

$$\overline{G} = \frac{G12\eta RT\xi}{\pi P_o^2 h^3} \tag{9.3.23}$$

gives a measure of the pressure drop through the bearing film which can be expressed as

$$\overline{G} = \left(\frac{P_d}{P_o}\right)^2 - \left(\frac{P_a}{P_o}\right)^2 \tag{9.3.24}$$

Figure 9.3.17 shows that the nondimensional mass flow rate is less sensitive to changes in supply pressure than it is to changes in $\Lambda_s\xi$.

Line Feed Solution Corrected to Account for Dispersion

When a number of entry sources are present in a bearing, dispersion pressure losses occur between inlets. The analysis presented in the preceding section assumed a line feed model which ignored these dispersion losses. To account for dispersion, an infinite number of sources are required around the feeding plane, or alternatively, a small groove is employed to connect the orifices. However, when discrete orifice restrictors are employed without a connecting groove, a pressure loss occurs between the inlet sources, so they do not act like a line source. In this section it is shown how an appropriate line feed correction factor can be deduced to account for dispersion effects. The method used to deduce the line feed correction factor is that which was previously suggested by Lund[53] for journal bearings and from which the design data presented by MTI was based. This method involves relating the assumed line source pressure[54] P_L to the orifice downstream pressure

[53] J. Lund, "The Hydrostatic Gas Journal Bearing with Journal Rotation and Vibration," J. Basic Eng., Trans. ASME, Ser. D , Vol. 86, pp. 328–336, 1964.

[54] This is the pressure in the bearing along a line that connects all the orifices. This is not the pressure in the line from the compressor, which is P_o.

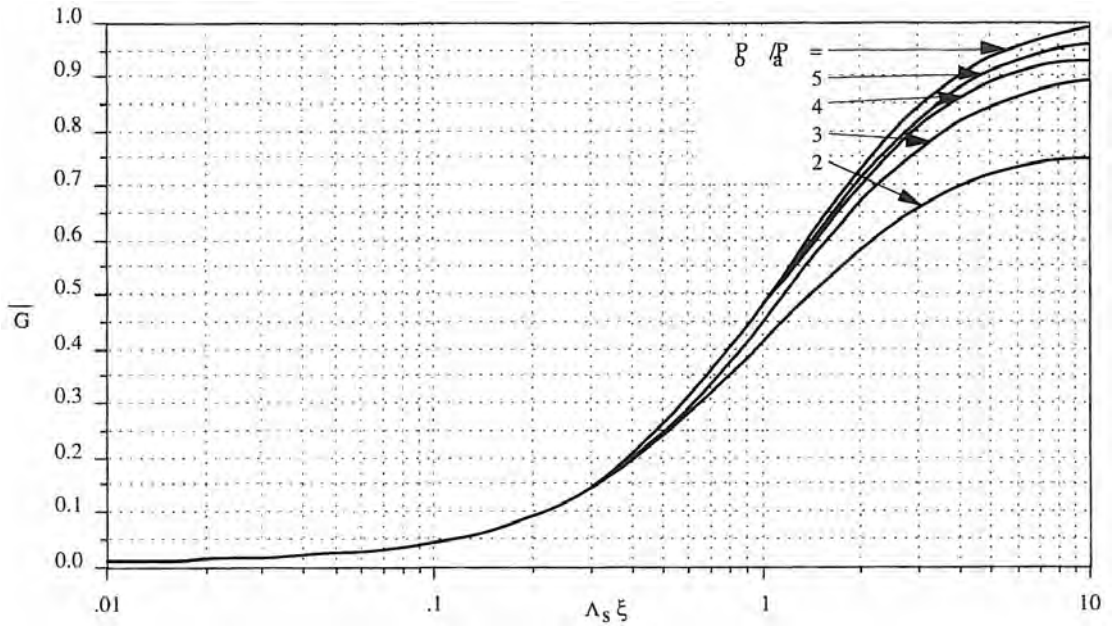


Figure 9.3.17 Nondimensionalized mass flow rate as a function of the feeding parameter $\Lambda_s \xi$, where $C_d = 0.8$, $\nu = 1.4$, and $1/\lambda = 1.0$.

P_d, by the factor $1/\lambda$:

$$\frac{1}{\lambda} = \frac{P_L^2 - P_a^2}{P_d^2 - P_a^2} \tag{9.3.25}$$

Figure 9.3.18 shows l/λ for a given bearing geometry. This diagram was constructed using complex potential theory.[55] A measure of the dispersion losses can be obtained directly from this figure by entering the appropriate bearing design parameters. Notice that as the number of orifices is increased, the spacing between them decreases. Hence the pressure drop which occurs and becomes greatest at the midpoint between the orifices is reduced when the orifices become closer together. N refers to the total number of orifices per pad and n refers to the total number of orifices per row. The clearance is not important in terms of l/λ, since orifice diameter is selected to balance clearance (flow in = flow out). $N\xi$ relates to the performance as a function of pad shape, whereas nd/D is a measure of how the discrete entries approach acting like a line feed source.

Assuming viscous, isothermal flow, but now defining the pressure around the inlet plane as a line pressure P_L, the flow through the bearing clearance is given by

$$G = \left[\left(\frac{P_L}{P_o} \right)^2 - \left(\frac{P_a}{P_o} \right)^2 \right] \frac{\pi h^3 P_o^2}{6\eta RT \xi} \tag{9.3.26}$$

Substituting $P_d{}^2$ for $P_L{}^2$ from Equation 9.3.25:

$$G = \left[\left(\frac{P_d}{P_o} \right)^2 - \left(\frac{P_a}{P_o} \right)^2 \right] \frac{\pi h^3 P_o^2}{6\eta RT \xi \lambda} \tag{9.3.27}$$

The general effect of l/λ upon the flow rate and P_d/P_o is shown in Figure 9.3.19. As l/λ decreases, P_d/P_o increases and flow rate decreases as shown by Equation 9.3.26, P_L/P_o also decreases which in turn leads to a reduced load capacity.

[55] For further details, see E. Pink, "The Application of Complex Potential Theory to Externally Pressurized Gas Lubricated Bearings," Proc. 8th Int. Gas Bearing Symp., Leicester Polytechnic, England, 1981.

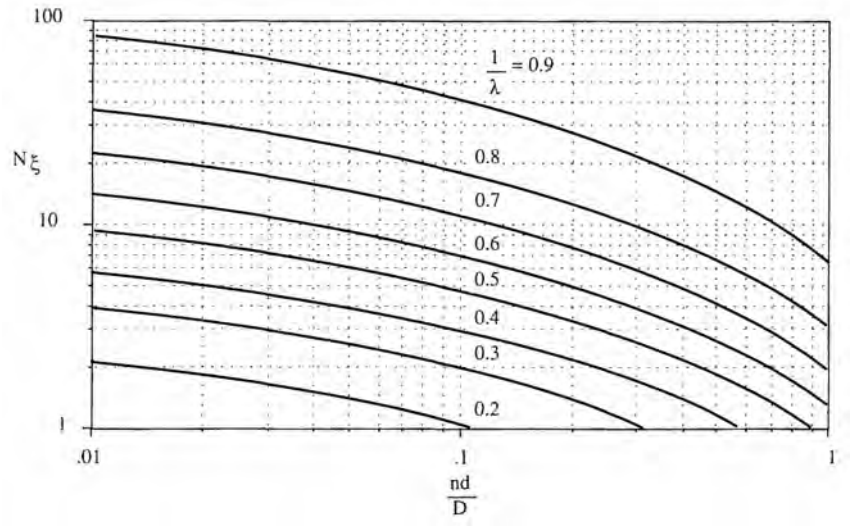


Figure 9.3.18 Determination of l/λu

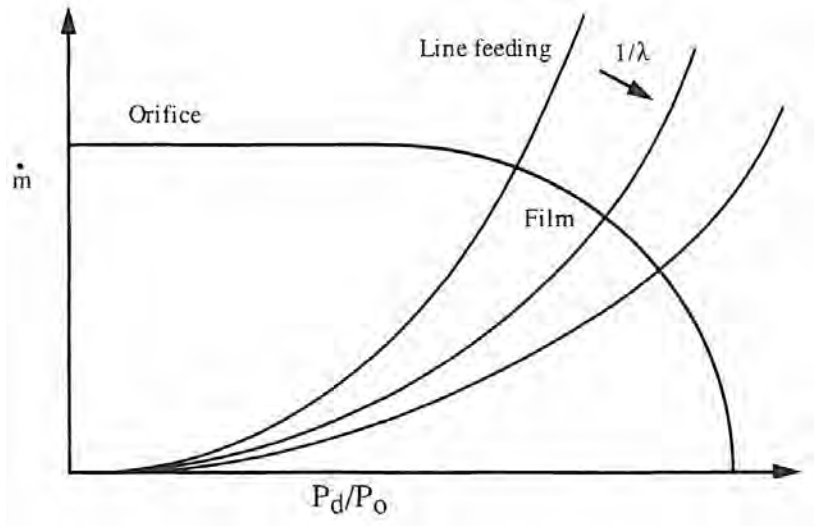


Figure 9.3.19 Effect of l/λ on orifice and film pressures.

Equating the flow rate through the orifices, with the flow through the bearing clearance, as was done to obtain Equation 9.3.19, gives

$$\Lambda_s \xi C_d \left\{ \frac{2\nu}{\nu-1} \left[\left(\frac{P_t}{P_o}\right)^{\frac{2}{\nu}} - \left(\frac{P_t}{P_o}\right)^{\frac{\nu+1}{\nu}} \right] \right\}^{\frac{1}{2}} = \left(\frac{P_L}{P_o}\right)^2 - \left(\frac{P_a}{P_o}\right)^2 = \frac{1}{\lambda}\left[\left(\frac{P_d}{P_o}\right)^2 - \left(\frac{P_a}{P_o}\right)^2 \right] \tag{9.3.28}$$

where

$$\frac{P_t}{P_0} = \left[\frac{2}{\nu+1}\right]^{\frac{\nu}{\nu-1}} \quad \text{(choked orifices)} \tag{9.3.29}$$

and

$$\frac{P_t}{P_o} = \frac{P_d}{P_o} \quad \text{(unchoked orifices)} \tag{9.3.30}$$

Figure 9.3.20 shows how the orifice and film pressures vary. As $\Lambda_s\xi$ increases, greater disparity occurs between P_L/P_o and P_d/P_o. This indicates that a single orifice eventually loses its ability to act as a restrictor for too large a bearing, whereas a line source reaches a stable limit.

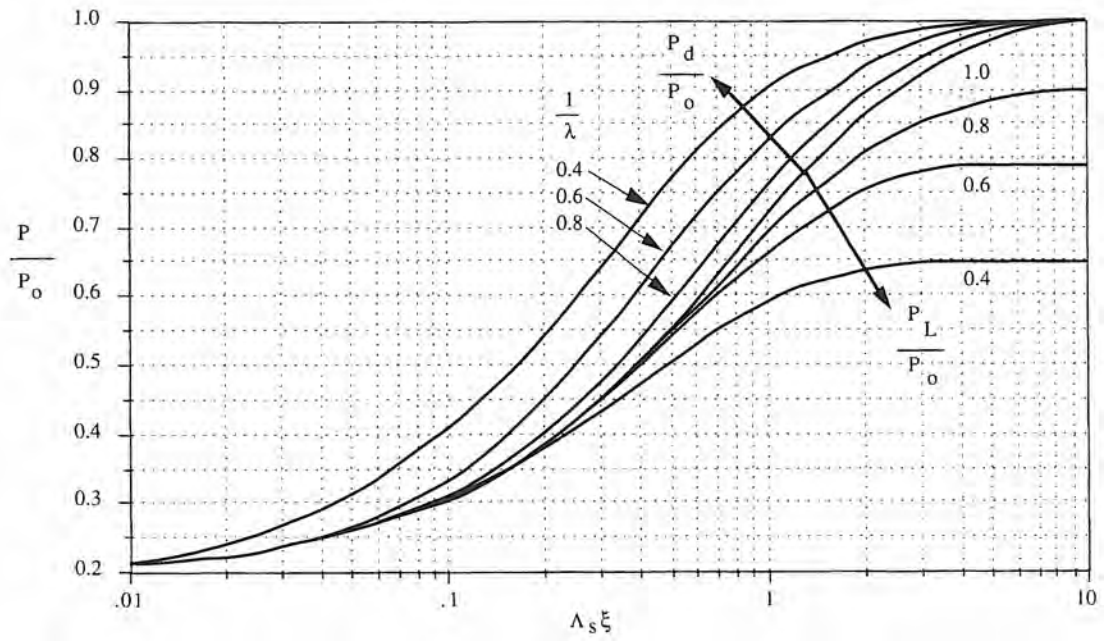


Figure 9.3.20 Orifice and film pressure as a function of $\Lambda_s\xi$, where $C_d = 0.8$, $\nu = 1.4$, and $P_o/P_a = 5.0$.

Effect of λ *on Flow Rate*

Nondimensional flow for the line source may be expressed in a manner similar to that for a single orifice:

$$\overline{G}_L = \left(\frac{P_L}{P_o}\right)^2 - \left(\frac{P_a}{P_o}\right) = 1\lambda\left[\left(\frac{P_d}{P_o}\right)^2 - \left(\frac{P_a}{P_o}\right)^2\right] \tag{9.3.30}$$

At low $\Lambda_s\xi$, there is a larger pressure drop across the bearing clearance than through the orifices so there is plenty of flow to act to connect up the orifices; hence the comparative difference between line source pressure and orifice pressure (l/λ) has little effect on mass flow rates. In Figure 9.3.21 it can be seen that as $\Lambda_s\xi$ increases, the effect of l/λ on flow rate becomes more significant.

Effect of λ *on Load Capacity*

The effect of l/λ on load capacity can be demonstrated by considering a single-acting thrust bearing where parallel film conditions exist. The load capacity would follow the trend shown in Figure 9.3.22. At low values of $\Lambda_s\xi$, little differences exist, but as $\Lambda_s\xi$ increases, the effect of l/λ becomes more pronounced. When $1/\lambda$ decreases, the orifices behave less like a line source and the bearings load capacity drops.

Effect of λ *on Stiffness*

The film stiffness is largely dependent on the change in line pressure as a function of bearing gap (dP_d/dh). To demonstrate the effect of l/λ upon the bearing stiffness, a small eccentricity analysis was made to determine the film pressure response of dP_d/dh. The results are shown in Figure 9.3.23. As $1/\lambda$ decreases, the maximum obtainable dP_d/dh reduces significantly, which would result in lower bearing stiffness. However, as $1/\lambda$ decreases (increasing dispersion) so does the optimum $\Lambda_s\xi$ for maximum dP_d/dh (and stiffness) conditions. The lower the value of $\Lambda_s\xi$, the greater the resistance of the bearing clearance compared to the resistance of the orifices; hence the orifices are more likely to behave as a line source which increases stiffness. For typical values of $1/\lambda$ used in practice, the optimum feeding parameter $\Lambda_s\xi$ or maximum stiffness would be in the range of 0.45 to 0.67.

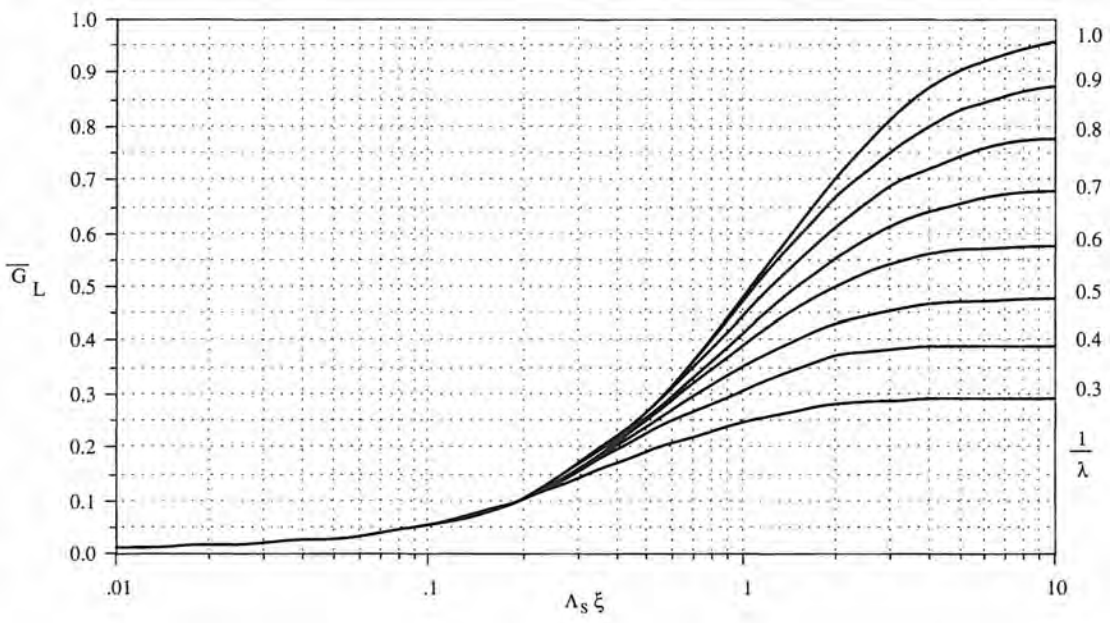


Figure 9.3.21 Nondimensional line source flow as a function of $\Lambda_s \xi$ or various l/λ where $C_d = 0.8$ and $\nu = 1.4$.

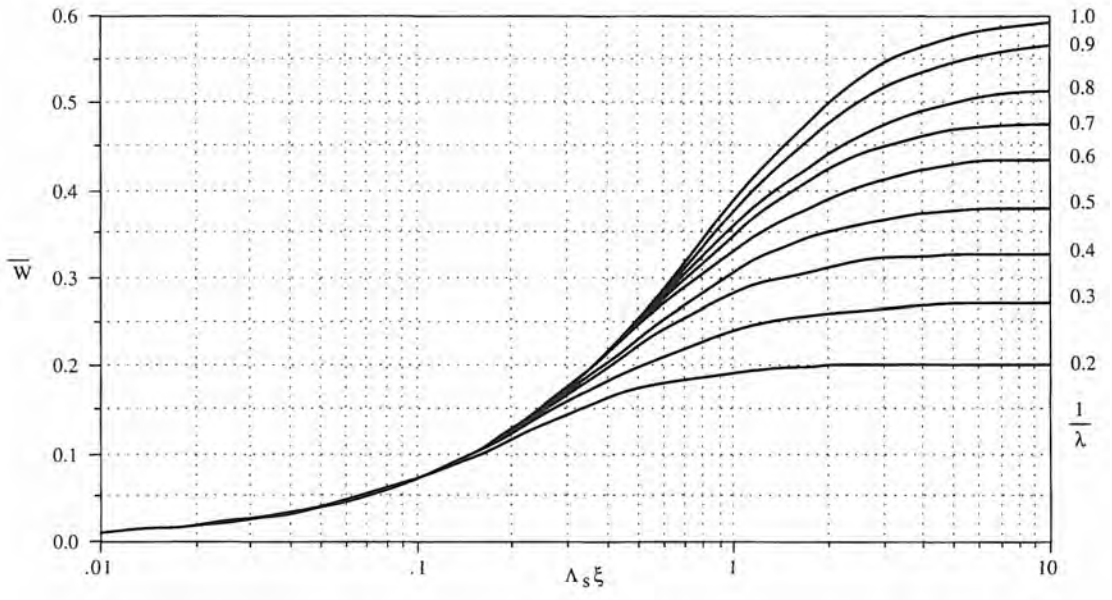


Figure 9.3.22 Effect of l/λ on loading characteristics for an inherently compensated bearing where $C_d = 0.8$, $\nu = 1.4$, $R_o/R_i = 1.5$, and $P_o/P_a = 5.0$.

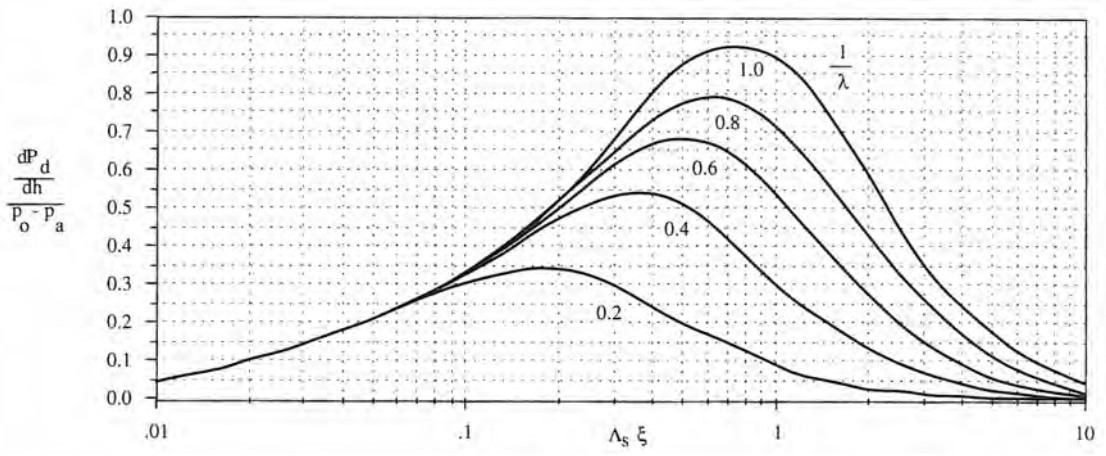


Figure 9.3.23 Sensitivity of film pressures with changes in clearance for various l/λ for a pocketed orifice bearing where $C_d = 0.8$, $\nu = 1.4$, and $P_o/P_a = 5.0$. For inherently compensated bearings, multiply the ordinate by $^2/_3$.

9.3.4 Design of Circular and Annular Thrust Bearings

A circular thrust bearing, as shown in Figure 9.3.24, can support unidirectional loads such as axial shaft loads. These bearings typically can have a relatively low flow rate. However, this configuration cannot be used at the faceplate end of a spindle, and therefore their use is limited to a small range of simple thrust-only applications. Note that the angular stiffness of this type of bearing is essentially zero. Annular thrust bearings are also shown in Figure 9.3.24. The bearing arrangement consists of a circular row of inlet holes at an intermediate radius between the outer and inner bearing radii. This bearing is often used in an opposed pad arrangement in the design of spindle assemblies.

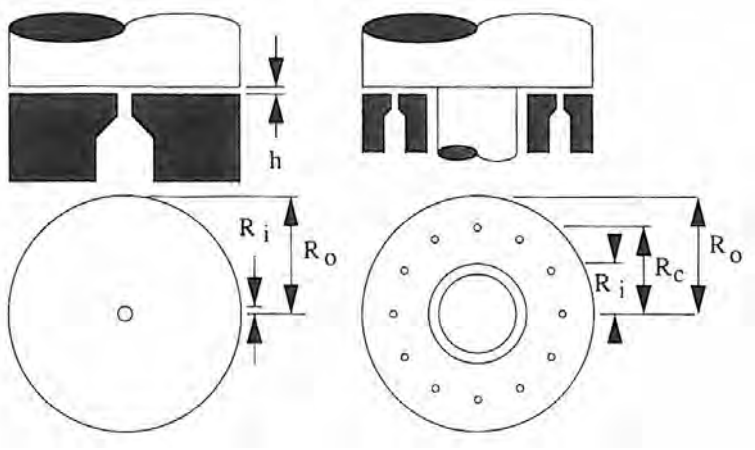


Figure 9.3.24 Circular and annular thrust bearing geometries.

From the point of view of the loading characteristics, it is desirable to choose pocketed orifices, as these give up to 1.5 times greater stiffness than annular orifices. However, if pocketed orifices are used, due consideration has to be paid to the avoidance of pneumatic hammer. To prevent pneumatic hammer, experience has shown that a good criterion is to design the pocket geometry such that the total volume enclosed in the pockets is less than one-twentieth of the bearing land volume. In practice, this means that the circular thrust bearings can normally be designed to incorporate a pocketed orifice. However, for annular and rectangular thrust bearings, the choice of pocketed or inherently compensated bearings is largely dependent upon the land area. Small land areas tend to favor inherently compensated orifices as excessive pocket volume/film volume ratios can occur.

For pocketed orifice bearings the recess depth should be equal to or greater than the orifice diameter. Also, the orifice geometry should be designed such that the curtain flow area $\pi d_R h$ is at least twice the orifice flow area $0.25\pi {d_o}^2$. This ensures that predominantly pocketed compensation is achieved:

$$\frac{\pi d_o^2}{4} \times \frac{1}{\pi d_R h_o} = \frac{d_0^2}{4 d_R h_0} \leq 0.5 \tag{9.3.31}$$

For bearings which have inherently compensated orifices, a good design criterion is to ensure that the inlet flow area is at least twice the curtain flow area:

$$\frac{\pi d_f^2}{4} \times \frac{1}{\pi d_f h_o} = \frac{d_f}{4 h_o} \geq 2 \tag{9.3.32}$$

Attention to these guidelines will ensure that bearings will operate in a predictable manner and their performance will correspond to that calculated using the design data presented in this chapter.

Parameter	Pocketed orifices	Inherently compensated orifice
Max. stiffness (N/μm)	$K = \frac{0.27\pi R_o^2 (P_o - P_a)}{h_o}$	$K = \frac{0.18\pi R_o^2 (P_o - P_a)}{h_o}$
Max. Load (N) @ $\varepsilon = 0$	$W = 0.15\pi R_o^2 (P_o - P_a)$	
Max. Load (N) @ $\varepsilon = -0.25$	$W = 0.21\pi R_o^2 (P_o - P_a)$	
Max. Load (N) @ $\varepsilon = 0.25$	$W = 0.11\pi R_o^2 (P_o - P_a)$	
Air flow rate (m^3/s)	$Q = \frac{0.34\, h_o^3 P_o^2}{3.42 \times 10^6 \times 2\log_e(R_o/R_i)}$	
Orifice diameter (mm)	$d_o = \sqrt{\frac{\Lambda_s \xi P_o h_o^3}{7890 \times 2\log_e(R_o/R_i)}}$	$d_f = \frac{\Lambda_s \xi P_o h_o^2}{31.55 \times 2\log_e(R_o/R_i)}$
Pocket depth (mm)	$b \leq \frac{0.05 h_o}{(R_i/R_o)^2 \times 10^3}$	Not applicable
Pocket compensation when:	$\frac{125 d_o^2}{R_i h_o} < 0.5$	Not applicable

Figure 9.3.25 Design equations for circular thrust bearings where $\Lambda_s \xi = 0.85$. Load capacities are based on $R_o/R_i = 20$. Units of pressure are N/mm^2. Units of dimensions are mm. Units of bearing gap h_o are μm.

For spreadsheet design purposes, typically available performance characteristics of circular and annular thrust bearings are Figures 9.3.25 and 9.3.26, respectively. For the equations given in these figures, any consistent set of variables may be used to determine the approximate bearing performance. The equations relate to "optimum" bearing performance. For multiinlet (annular) bearings, allowance has to be made for realistic dispersion losses between inlets, so it can be assumed that $1/\lambda$ is about 0.7.

9.3.4.1 Design of Circular Thrust Bearings

To ensure that pneumatic hammer is avoided when pocketed orifices are used, the ratio of pocketed volume/clearance volume should be at least 1:20 so the pocket depth b in mm should be

$$b \leq \frac{0.05\, h_o}{(R_i/R_o)^2 \times 10^3} \tag{9.3.33}$$

where h_o is the nominal bearing gap in microns. The pocket depth can be calculated initially assuming that $R_i/R_o = 0.05$ in the equation above where R_i is the pocket radius ($R_i = d_R/2$), and R_o

Parameter	Pocketed orifices	Inherently compensated orifice
Max. stiffness (N/μm)	$K = \frac{0.44\pi(R_o^2 - R_i^2)(P_o - P_a)}{h_o}$	$K = \frac{0.29\pi(R_o^2 - R_i^2)(P_o - P_a)}{h_o}$
Max. angular stiffness (Nm/rad)	$K_A = \frac{0.23\pi(R_o^2 - R_i^2)R_oR_i(P_o - P_a)}{h_o}$	
Max. load (N) @ $\varepsilon = 0$	$W = 0.26\pi(R_o^2 - R_i^2)(P_o - P_a)$	
Max. load (N) @ $\varepsilon = -0.25$	$W = 0.37\pi(R_o^2 - R_i^2)(P_o - P_a)$	$W = 0.35\pi(R_o^2 - R_i^2)(P_o - P_a)$
Max. load (N) @ $\varepsilon = 0.25$	$W = 0.18\pi(R_o^2 - R_i^2)(P_o - P_a)$	$W = 0.20\pi(R_o^2 - R_i^2)(P_o - P_a)$
Air flow rate (m^3/s)	$Q = \frac{0.27\, h_o^3 P_o^2}{3.42 \times 10^6 \times 2\log_e(R_o/R_i)}$	
Orifice diameter (mm)	$d_o = \sqrt{\frac{\Lambda_s \xi P_o h_o^3}{7890 \times n \times 0.5\log_e(R_o/R_i)}}$	$d_f = \frac{\Lambda_s \xi P_o h_o^2}{31.55 \times n \times 0.5\log_e(R_o/R_i)}$
Pocket depth (mm)	$b \le \frac{0.2(R_o^2 - R_i^2)h_o}{nd_R^2 \times 10^3}$	Not applicable
Pocket compensation when:	$\frac{R_o^2 \times 10^3}{d_R h_o} < 0.5$	Not applicable

Figure 9.3.26 Design equations for annular thrust bearings where $\Lambda_s\xi = 0.60$. Units of pressure are N/mm ^{2}. Units of dimensions are mm. Units of bearing gap h_o are μm.

is bearing radius. If the calculated pocket depth is smaller than can conveniently be manufactured, then R_i/R_o has to be reduced. Reducing R_i/R_o reduces load capacity as well as reduces the flow rate. Alternatively, increasing R_i/R_o reduces the pocket depth and increases load capacity and flow rate.

Figures 9.3.27–9.3.29 illustrate the load capacity and film stiffness characteristics for circular thrust bearings. In Figures 9.3.27 and 9.3.28 performance is plotted against feeding parameter $\Lambda_s\xi$, where for pocketed and inherently compensated orifices, respectively:

$$\Lambda_s \xi = \frac{15.8 \times 10^3\, d_o^2}{P_o h_o^2} \log_e\left(\frac{R_o}{R_1}\right) \tag{9.3.34}$$

$$\Lambda_s \xi = \frac{63.1\, d_f}{P_o h_o^2} \log_e\left(\frac{R_o}{R_1}\right) \tag{9.3.35}$$

The data shown applies specifically to $R_i/R_o = 0.05$ and a correction factor is used for the load or stiffness for differing values of R_i/R_o as shown in Figure 9.3.29. The units are N/mm^2 for pressure, microns for bearing gap and mm for radii.

It can be seen from Figure 9.3.28 that the stiffness is a maximum at $\Lambda_s\xi = 0.85$. For this optimum condition, the pocket radius (radius of the orifice plug) can be calculated from Equation 9.3.33.

Example

A circular thrust bearing is required to support a minimum load of 700 N (160 lb). It must also have a stiffness exceeding 70 N/μm (0.4 lb/μin.). The bearing is additionally required to have moderate flow rate, hence the air film thickness h_o is required to be small. Assume that the outside

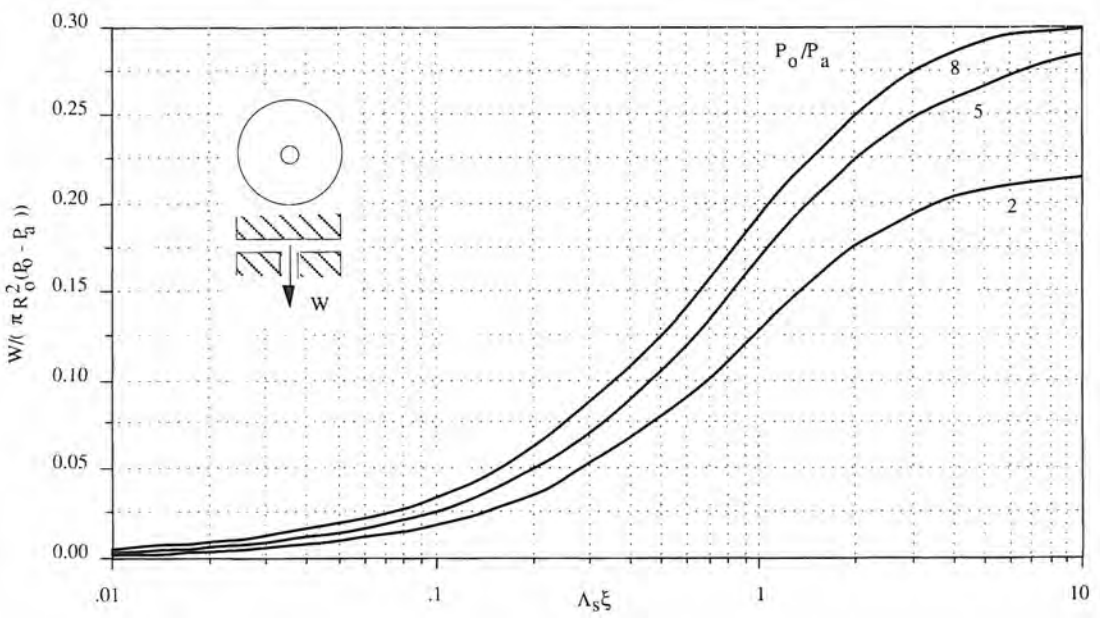


Figure 9.3.27 Pocketed circular thrust bearing load capacity where $R_i/R_o = 0.05$, $C_d = 0.8$, and $\nu = 1.4$. For inherently compensated bearings, multiply the ordinate by $^2/_3$.

radius of the bearing is constrained so that it does not exceed 70 mm (2.75 in). The input values will be: $R_o = 70$ mm, $R_i = 3.5$ mm, $P_o/P_a = 5$ ($P_o = 505$ kN/m^2), $h_o = 20$ μm. From Figure 9.3.25 the equation for stiffness has units R_o = mm, (P_o - P_a) = N/mm^2, and h_o = μm. When $P_o/P_a = 5$, P_o-P_a = 4 atm = 0.405 N/mm^2, the stiffness is

$$K = \frac{0.27 \times \pi \times 70^2 \times 0.405}{20} = 84.17\,\text{N}/\mu\text{m}$$

From Figure 9.3.25 the equation for load capacity yields W = 935.2 N. From Figure 9.3.25 the equation for flow rate yields

$$Q = \frac{0.34 \times 20^3 \times 0.506^2}{3.42 \times 10^6 \times 2 \times \log_e(20)} = 3.4 \times 10^{-5}\,\text{m}^3/\text{sec}$$

and $\Lambda_s\xi = 0.85$ from Figure 9.3.28. From Figure 9.3.25 the equation for the pocketed orifice diameter d_o is

$$d_o = \sqrt{\frac{0.85 \times 0.506 \times 20^3}{7.89 \times 10^3 \times 2 \times \log_e(20)}} = 0.27\,\text{mm}$$

From Equation 9.3.33, the pocket depth is found to be $b \leq 0.4$ mm.

9.3.4.2 Design of Annular Thrust Bearings

It was stated previously that annular thrust bearings are often used in conjunction with journal bearings. A consequence of their use in space is often limited and therefore land widths are often narrow. A consequence of narrow land widths is that the air film volume between the bearing faces is relatively small. If pocketed orifices are to be used, then the possibility exists that pneumatic hammer instability will occur. Hence many annular thrust bearings have inherently compensated orifices. Design procedures are presented below for bearings having either pocketed orifices or inherently compensated orifices.

Position of Inlet Sources

Some designers position the inlet sources at a radius R_c midway between the bearings inside radius R_i and the outside radius R_o:

$$R_c = \frac{R_i + R_o}{2} \tag{9.3.36}$$

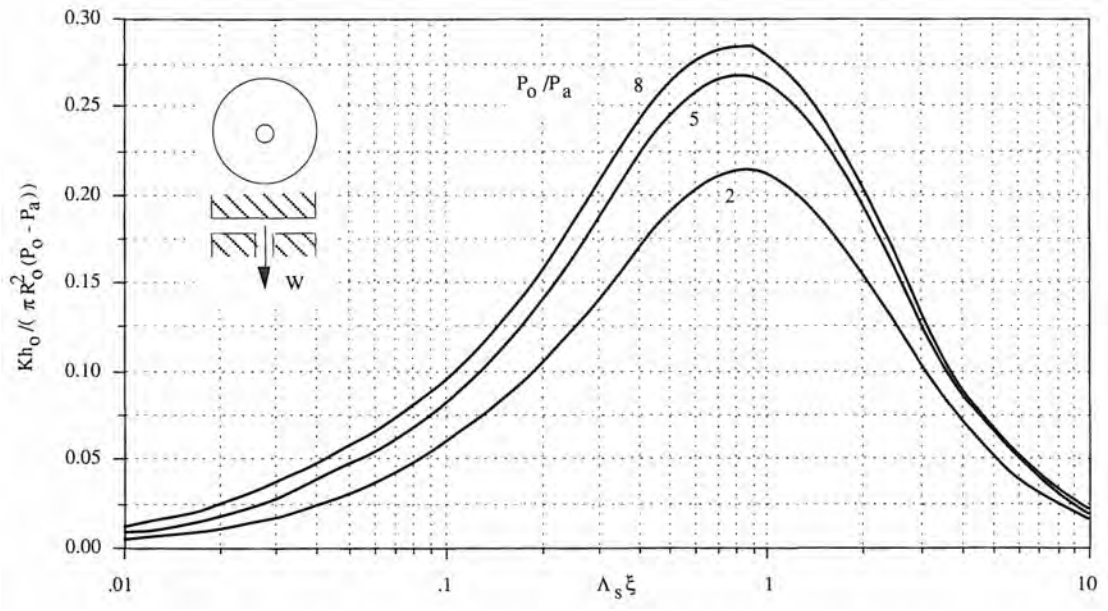


Figure 9.3.28 Pocketed circular thrust bearing stiffness where $R_i/R_o = 0.05$, $C_d = 0.8$, and $\nu = 1.4$. For inherently compensated bearings, multiply the ordinate by $^2/_3$.

This position, although acceptable, does not lead to the minimum flow rate, since there is less resistance to outward flow than there is to inward flow. The absolute minimum flow rate occurs where the inward flow resistance equals the outward flow resistance:

$$R_c = \sqrt{R_o R_i} \tag{9.3.37}$$

This is derived by equating flows inward to the flows outward using the appropriate logarithmic expression for flow in a circular segment.Curves for load capacity and stiffness as a function of the feeding parameter $g\Lambda_s\xi$ are similar to those shown in Figures 9.3.27 and 9.3.28.

The feeding parameter $\Lambda_s\xi$s here the units are the same as for Equation 9.3.34, is given by

$$\Lambda_s\xi = \frac{7.89 \times 10^3\, n d_o^2}{P_o h_o^2} \times \frac{\log_e}{2}\left(\frac{R_o}{R_i}\right) \tag{9.3.38}$$

for pocketed orifices, and for inherently compensated orifices:

$$\Lambda_s\xi = \frac{31.55\, n d_f}{P_o h_o^2} \times \frac{\log_e}{2}\left(\frac{R_o}{R_i}\right) \tag{9.3.39}$$

To account for dispersion losses, the corrected line feed model shown in Figure 9.3.18 has been used to calculate bearing performance, where

n	=	number of inlets per circumferential row
N	=	total number of inlets in the bearing (for more than one circumferential row)
ξ	=	$0.5 \log_e(R_o/R_i)$
d/D	=	$d_f/2R_c$ for inherently compensated bearings
d/D	=	$d_R/2R_c$ for pocketed bearings (where $D = R_o + R_i$)

Calculation of Flow Rate

The nondimensional flow rate was plotted as a function of $\Lambda_s\xi$ in Figure 9.3.17 from which the flow rate in m^3/s can be calculated by using the equation

$$G = \frac{\overline{G} h_o^3 P_o^2}{1.71 \times 10^{12} \log_e (R_o/R_i)} \qquad \text{(free air, } m^3/s) \tag{9.3.40}$$

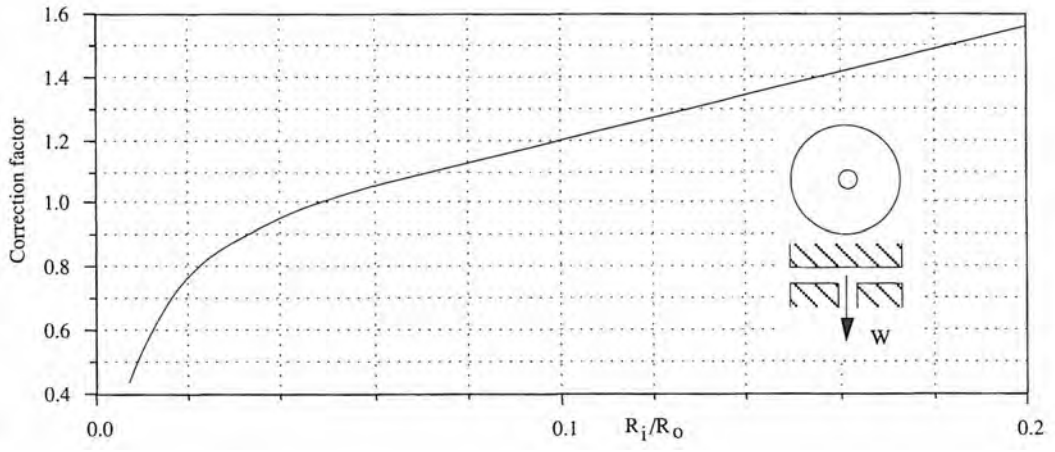


Figure 9.3.29 Load and stiffness correction factor (= 1.0 @ R_i/R_o = 0.05) for a circular thrust bearing.

where P_o = supply pressure (kN/m^2 absolute), h_o = film clearance (μm), d_o, d_f = orifice diameter (mm). In designs where pocketed orifices are used in annular thrust bearings, the following general rule for the pocket depth b in mm will ensure that pneumatic hammer will be avoided:

$$b \leq \frac{0.2\,(R_0^2 - R_i^2)\,h_o}{d_R^2\,N} \tag{9.3.41}$$

Example

An annular thrust bearing having pocketed orifices is required to support the thrust load of a small rotary table. The bearing must be designed so that the stiffness exceeds 600 N/μm (3.4 lb/μin.). The outside radius of the thrust bearing R_o = 200 mm (8 in.). The inside radius is variable but must not be less than R_i = 100 mm (4 in.). (Preferably the inside radius should be larger.) The air film thickness must not be smaller than 15 μm, but if stiffness permits, a large clearance would be preferable. The input values will be: R_o = 200 mm, R_i = 114 mm, P_o/P_a = 5 (506 kN/m^2), h_o = 20 μm, n = 10 orifices, d_R = 3.2 mm. From Figure 9.3.26 the equation for stiffness is obtained:

$$K = \frac{0.44 \times \pi \times (200^2 - 114^2) \times 0.405}{20} = 755\,\text{N}/\mu\text{m}$$

From Figure 9.3.26 the equation for load capacity is obtained:

$$W = 0.26 \times \pi(200^2 - 114^2) \times 0.405 = 8933\,\text{N}$$

From Figure 9.3.26 the equation for flow rate is obtained:

$$Q = \frac{0.27 \times 20^3 \times 0.506^2}{3.42 \times 10^6 \times 0.5 \times \log_e(1.75)} = 5.7 \times 1^{-4}\,\text{m}^3/\text{s}$$

From Figure 9.3.25 the equation for orifice diameter d_o is obtained with $\Lambda_s\xi$ = 0.6:

$$d_o = \sqrt{\frac{0.6 \times 0.506 \times 20^3}{7.89 \times 10^3 \times 10 \times 0.5 \times \log_e(1.75)}} = 0.33\mu\text{m}$$

From Figure 9.3.26 pocket depth is given by

$$b \leq \frac{0.2 \times (200^2 - 114^2) \times 20}{10 \times 3.2^2 \times 10^3} \leq 1.05\,\text{mm}$$

It is necessary to check that the restriction to the air as it enters the bearing clearance is due primarily to the orifice pocket is described by Equation 9.3.31. This is when

$$\frac{d_o^2 \times 10^3}{4 d_R h_o} = \frac{0.33^2 \times 10^3}{4 \times 3.2 \times 20} = 0.425$$

Since the value is less than 0.5, the design is satisfactory. From Figure 9.3.26 the equation for angular stiffness is obtained:

$$K_A = \frac{0.23 \times \pi \times (200^2 - 114^2) \times 200 \times 114 \times 0.405}{20} = 8.98 \times 10^6 \text{ N-M/rad}$$

When checking for the likelihood of pneumatic hammer in a pocketed orifice annular thrust bearing the following condition must apply (from Equation 9.3.33):

$$\frac{4(R_o^2 - R_i^2)h_o}{1000\, N d_R^2\, b} = \frac{4(200^2 - 114^2) \times 20}{1000 \times 10 \times 3.2^2 \times 1.05} = 20.09 > 20$$

The value just exceeds 20; if a further margin of safety is required, reduce the pocket depth b. By reducing b to 0.7 mm, the ratio of (land volume/pocket volume) = 30.1 could be acceptable, but Equation 9.4.18 would have to be rechecked to ensure that the resulting value is less than 0.5.

As with the circular thrust bearing example, the graphical and numerical results are close. The graphs are easier to use, but the equations are more accurate.

9.3.5 Design of Rectangular Flat Pad Bearings

Rectangular flat pad bearings typically may have either a single-row-entry configuration or a double-row-entry configuration as shown in Figure 9.3.30. The double-row configuration has the advantage that load capacity and stiffness are improved, although a penalty is paid in that the gas flow rate is doubled. Rectangular thrust bearings are commonly used on linear motion carriages in grinding machines, diamond turning machines, and measuring machines.

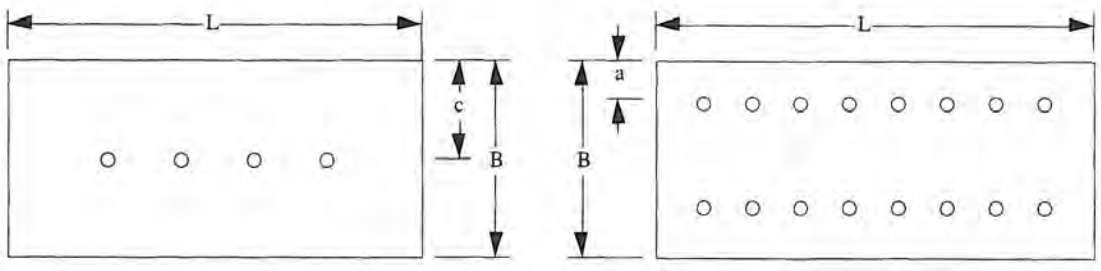


Figure 9.3.30 Single-entry (c/B = 0.5) and double-entry (a/b = 0.25) thrust bearings.

The rectangular flat pad bearing is possibly one of the more difficult configurations to design efficiently because of the complex flow paths from the inlet orifices to the bearing outlet, which occur within the bearing clearance. It is often possible to obtain inefficient pressure profiles within the bearing clearances, and a number of designs have yielded poor or unstable performance. As a consequence of the difficulty of attempting to predict flow characteristics within the bearing clearances in a general design sense, it has been realized by many designers that simple approximate models or simple analytical solutions do not yield satisfactory solutions. To overcome these difficulties it is necessary to use more sophisticated techniques such as finite difference calculations to evaluate the load, flow, stiffness and angular stiffness coefficients. These coefficients are then used in developing design procedures to evaluate other aspects of rectangular thrust pad bearing design. It is possible to design rectangular thrust pad bearings using the charts presented in this section to yield a satisfactory solution for a range of bearing sizes. The design charts presented here are based on considerable computation and have been verified in a number of bearing systems designed for industrial applications.

The major factor which influences the stiffness of the bearing system is the line feed factor l/λ. The theoretically ideal bearing would have the entry sources so close together that there was no drop in pressure between them (i.e., a continuous line of entries). In practice such an arrangement would often be impractical, due to manufacturing considerations and expense,[56] so pressure drops

[56] Unless, of course, one uses a porous medium.

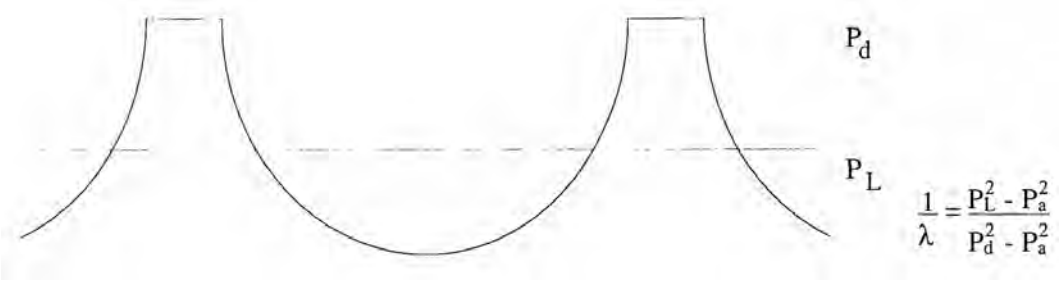


Figure 9.3.31 Line feed correction factor l/λ.

between inlet orifices are normally tolerated and a typical situation is shown in Figure 9.3.31. An important aspect of the design is to ensure that the positioning of the inlet orifices does not seriously detract from the bearing's operational performance. When undertaking the design of rectangular flat pad bearings, attention to the complexities of orifice sizing and positioning is of such importance that the line feed factor has to be calculated first when standard equations are used to evaluate bearing performance.

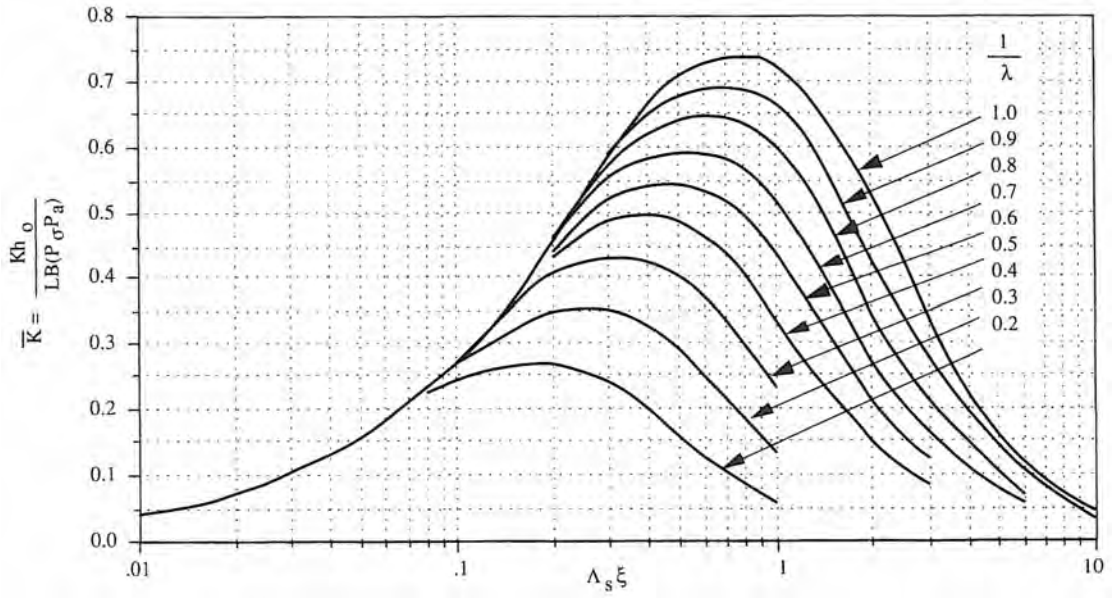


Figure 9.3.32 Stiffness parameter for rectangular double-entry thrust bearings with pocketed orifices where a/B = 0.25, $P_o/P_a = 5$, $C_d = 0.8$, and $\nu = 1.4$. For inherently compensated orifices, multiply the ordinate by 0.67. For single-entry bearings, multiply the ordinate by 0.75.

Rectangular Bearing Pad Design Parameters

$\Lambda_s\xi$ is the feeding parameter, which is a measure of the ratio of pressure drop through the orifices to the pressure drop through the bearing clearance and is an optimum at 0.55 for rectangular thrust bearings. l/λ is a measure of how well the string of orifices approaches a true line source. Variations of 1/λ affects the bearing stiffness, load capacity, and mass flow rate. As orifices become closer together, 1/λ increases as a line source is approached. Figure 9.3.32 shows dimensionless stiffness coefficients for rectangular thrust bearings. Figure 9.3.32 shows that the greater the value of 1/λ, the greater the stiffness. A similar situation is seen in Figure 9.3.33, which shows dimensionless load capacity against $\Lambda_s\xi$ for various 1/λ. It is also shown in Figure 9.3.34 that the flow coefficient increases with 1/λ, which provides more entry sources into the bearing, thereby improving the bearing performance.

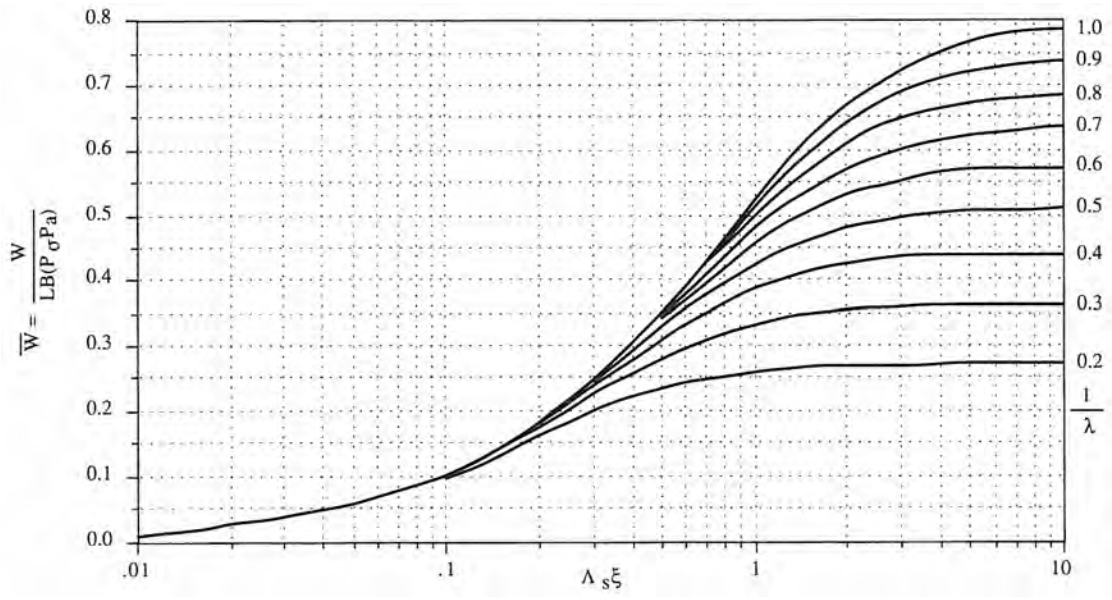


Figure 9.3.33 Load parameter for rectangular double-entry thrust bearings with pocketed orifices where $\varepsilon = 0$, a/B = 0.25, $P_o/P_a = 5$, $C_d = 0.8$, and $\nu = 1.4$. For single-entry bearings, multiply the ordinate by 0.75.

Figures 9.3.32–9.3.34 can be used to provide the coefficients necessary to determine the stiffness, load, and flow parameters of the bearing. To determine these coefficients it is necessary to determine $1/\lambda$. With reference to Figure 9.3.30, the term $1/\lambda$ is evaluated by first calculating $N\xi$, and $nd\pi/L$ (nd/D) and then using Figure 9.3.18:

$$\xi = \frac{\pi B}{L} \qquad \text{(single entry bearings)} \tag{9.3.42a}$$

$$\xi = \frac{2\pi a}{L} \qquad \text{(double entry bearings)} \tag{9.3.42b}$$

n = number of orifices per row.
N = number of orifices per bearing (where there are two rows of inlets N = 2n).
d = d_f for inherently compensated bearings
d = d_R for pocket-compensated bearings

Calculation of Flow Rate

Nondimensional flow rate is plotted against $\Lambda_s\xi$ in Figure 9.3.34, and it can be used to calculate the flow rate G (m^3/s):

$$G = \frac{\overline{G} h_o^3 P_o^2}{3.42 \times 10^6 \xi} \tag{9.3.43}$$

where P_o and h_o have units of N/mm^2 and mm, respectively. Note that the equation above applies to both single-entry and double-entry configurations. The term ξ accounts for the difference in flow rate of the two bearings.

Ensuring Stability in Bearings Having Pocketed Orifices

In designs where pocketed orifices are used, the following general rule for the pocket depth b in mm will help to ensure that pneumatic hammer will be avoided.

$$b \leq \frac{0.2\, LBh_o}{\pi d_R^2 N \times 10^3} \tag{9.3.44}$$

where h_o is in microns and other dimensions are in mm.

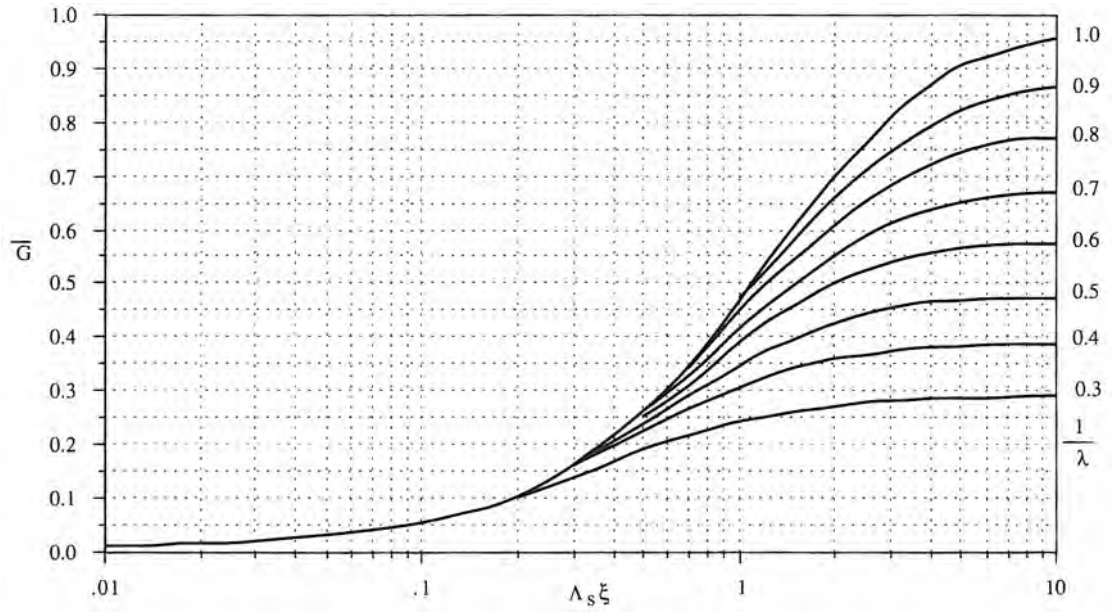


Figure 9.3.34 Nondimensional flow rate for rectangular double-entry thrust bearings with pocketed orifices where a/B = 0.25, P_o/P_a = 5, C_d = 0.8, and ν = 1.4. For single-entry bearings, multiply the ordinate by 0.5.

Calculation of Orifice Dimensions

Orifice dimensions control the pressure drop through the orifices and subsequently through the bearing clearances. This pressure drop is specified in terms of the feeding parameters discussed previously. The feeding parameters may be defined for pocketed orifices as

$$\Lambda_s\xi = \frac{7.89 \times 10^3\, n d_o^2 \pi B}{P_o h_o^3 L} \tag{9.3.45a}$$

and for inherently compensated orifices:

$$\Lambda_s\xi = \frac{31.55\, n d_f \pi B}{P_o h_o^2 L} \tag{9.3.45b}$$

Units of pressure are N/mm^2, units of gap h_o are microns, and all other dimensions are mm. It should be noted that $\Lambda_s\xi$ is the same for both single- and double-row entry configurations. This applies for a/B = 0.5 (single-entry) and a/B = 0.25 (double-entry) bearing configurations. This is because the double-entry bearing incurs twice the flow rate of the single-entry bearing at the ratios of a/B specified; hence twice as many orifices are required at the orifice diameter specified.

Example

A rectangular air bearing is to be designed for use in a machine where the bearing is lightly loaded. The design requirement is for a bearing stiffness of 50 N/μm (0.3 lb/μin.). The bearing is to be operated with a regulated air supply P_o/P_a = 5 (P_o = 0.507 N/mm^2). The length of the bearing must not exceed 100 mm. The design air film clearance h_o is 30 μm. Determine the bearing width, flow rate, and load capacity. Select an appropriate number of orifices and define the orifice and pocket dimensions when the bearing has two rows of orifices at quarter stations (a/B = 0.25). The input values will be: bearing length L = 100, bearing width B = 50, double row entry a/B = 0.25, number of pocketed orifices per row n = 8 (N = 16), supply pressure P_o/P_a = 5.0, and air film thickness h_o = 30 μm. The steps in the design are:

1. Calculate the bearing shape factor ξ:

$$\xi = \frac{\pi 2a}{L} = \frac{\pi \times 2 \times 12.5}{100} = 0.785$$

$$N\xi = 16 \times 0.785 = 12.86$$

2. Calculate the diameter of the pocketed orifices from

$$\Lambda_s\xi = \frac{7.89 \times 10^3\, n d_o^2\, \pi B}{P_o h_o^3 L} \tag{9.3.46}$$

$$d_o = \sqrt{\frac{\Lambda_d\, \xi\, P_o\, h_o^3\, L}{7.89 \times 10^3\, n\pi B}} \tag{9.3.47}$$

$\Lambda_s\xi = 0.55$ for optimum stiffness of rectangular thrust pads, so the orifice diameter is

$$d_o = \sqrt{\frac{0.55 \times 0.506 \times 30^3 \times 100}{7.89 \times 10^3\, 8 \times 50 \times \pi}} = 0.275\,\text{mm}$$

$$\frac{n d_o \pi}{L} = \frac{8 \times 0.275 \times \pi}{100} = 0.69$$

3. Using values of $N\xi = 12.56$ and $nd_o\pi/L = 0.069$ in Figure 9.3.18 (note that d_o replaces d_f), l/λ is found to be 0.65.

4. Determine the stiffness using Figure 9.3.32:

$$K = \frac{\overline{K} LB(P_o - P_a)}{h_o} = \frac{0.57 \times 100 \times 50 \times 0.405}{30} = 38.5\,\text{N}/\mu\text{m}$$

5. Determine the load capacity using Figure 9.3.33:

$$W = LB\,(P_o - P_a)\overline{W} = 100 \times 50 \times 0.405 \times 0.36 = 729\,\text{N}$$

6. Determine the flow rate using Figure 9.3.34:

$$G = \frac{h_o^3 P_o^2 \overline{G}}{3.42 \times 10^6 \xi} = \frac{30^3 \times 0.506^2 \times 0.27}{3.42 \times 10^6 \times 0.785} = 6.95 \times 10^{-4}\text{m}^3/\text{s}$$

7. Determine the pocket diameter:

$$d_R = \frac{d_o^2 \times 10^3}{2h_o} = \frac{0.275^2 \times 10^3}{2 \times 30} = 1.26\,\text{mm}$$

8. Determine the pocket depth:

$$b \leq \frac{0.2\,LBh_o}{\pi d_R^2 N \times 10^3} = \frac{0.2 \times 100 \times 50 \times 30}{\pi \times 1.26^2 \times 16 \times 10^3} = 0.38\,\text{mm}$$

Opposed Pad Thrust Bearings

Thrust bearings are often used in an opposed pad configuration to provide location against motion in two directions. Load capacity, flow rate, and stiffness can be determined from the opposed pad configuration. Load capacity of an opposed pad bearing is 0.77 × single-acting bearing load capacity for pocketed orifice bearings, and 0.62 × single-acting bearing load capacity for inherently compensated bearing. The flow rate is doubled for an opposed pad configuration. The stiffness is also doubled for opposed pad configurations as is summarized in Figure 9.3.35.

9.3.6 Orifice Compensated Journal Bearings

The basic principles of operation of externally pressurized journal bearings were illustrated in Figure 9.3.3. Pressurized gas, normally air, is fed to an annular plenum chamber around the bearing at supply pressure P_o. The gas is admitted into the bearing clearance by the orifices, which reduce the gas pressure to P_d at the inlet. The gas subsequently flows into the bearing clearances, its pressure reducing to atmospheric pressure P_a at the outlet of the bearing. For concentric conditions, $\varepsilon = 0$, the

	Pocketed orifice bearings			Annular orifice bearings		
Bearing type	Load coeff.	Stiffness coeff.	Flow coeff.	Load coeff.	Stiffness coeff.	Flow coeff.
Double entry single acting	0.62	1.0	1.0	0.58	1.0×0.67	1.0
Double entry opposed pad	0.47	2.0	2.0	0.36	2.0×0.67	2.0
Single entry single acting	0.46	1.0×0.75	1.0	0.43	$1.0 \times 0.75 \times 0.67$	1.0
Single entry opposed pad	0.35	2.0×0.75	2.0	0.27	$2.0 \times 0.75 \times 0.67$	2.0

Figure 9.3.35 Correction factors for opposed pad thrust bearings. The values of B/L and B/2L in the equation for flow rate gives the appropriate multiplier for the flow rate for single- and double-entry bearings.

inlet pressures at the restrictors are normally equal and are designated as P_{do}. As with the other types of aerostatic bearings discussed previously, when the bearing is loaded, the shaft is displaced and an eccentricity occurs. The eccentricity leads to increased flow resistance in the low clearance side and hence reduced flow through the restrictors on this side of the bearing. Thus the inlet pressure in the low-clearance P_{d1} increases. Conversely, in the high-clearance side the inlet pressure P_{dh} decreases. The pressure differential across the bearing multiplied over the projected area of the bearing allows the bearing to support a load.

The design data presented in this section has been obtained from extensive analytical and experimental studies conducted over many years. The analysis used has taken into account a line feed model which corrects for both dispersion and circumferential flow losses. The data presented in this section is applicable for a typical range of supply pressures safely achievable from a commercial compressor, P_o/P_a, in the range 3-8.

9.3.6.1 Design Considerations

Figure 9.3.36 compares the load, flow rate, stiffness, and tilt stiffness of different typical journal bearing geometries and arrangements. A good compromise in design which gives high load capacity without incurring excessive flow rate is shown by configuration (a), which has double-row admission at a/L = 0.25. Configuration (b) shows that load capacity stiffness and tilt stiffness are reduced by using a single-row entry bearing, and gas flow rate is halved. Configurations (c) and (d) are double versions of (a) and (b), respectively, that can also support moment loads.

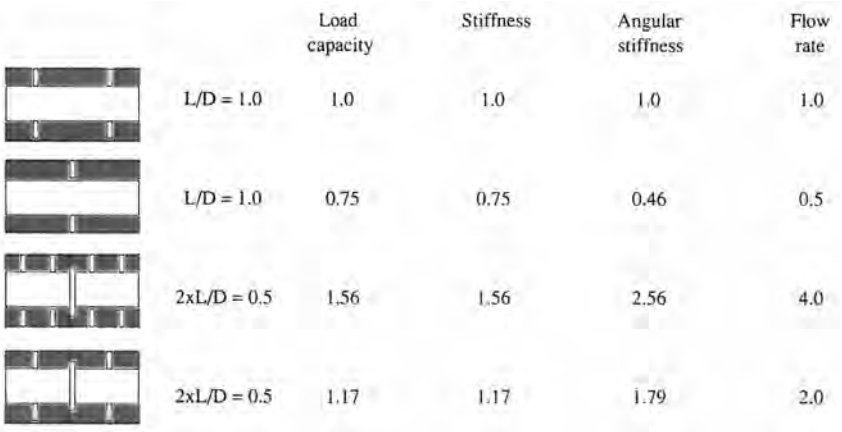


Figure 9.3.36 Performance comparison of load, flow rate stiffness and tilt stiffness for typical journal bearing geometries and arrangements.

Bearing Clearance

Bearing clearance affects both flow rate and stiffness. It is good design practice to select a clearance equal to the smallest value consistent with manufacturing constraints. As an approximate guide to clearance, gas journal bearings are usually manufactured to have an air film thickness in the range

$$\frac{2h_o}{D} = 0.0005 \text{ to } 0.00015 \tag{9.3.48}$$

These small clearances require accurate manufacture in terms of diameter and roundness. It is therefore necessary to employ measuring instruments to check the sizes of bearing elements that have high accuracy and resolution.

Orifice Design

Orifice design can take the form of either pocketed orifices or inherently compensated orifices, as discussed earlier. It is preferable to use pocketed orifices, as these provide greater stiffness by a factor of about 1.5. However, when pocketed orifices are used, to avoid pneumatic hammer the pocket depth in mm should be

$$b \leq \frac{0.2\,DLh_o}{d_R^2\,N \times 10^3} \tag{9.3.49}$$

where h_o is in microns and other dimensions are in mm. When pocketed orifices are used, the major flow restriction is provided by the orifice flow area $0.25\pi {d_o}^2$. However, a secondary flow restriction exists downstream as the flow enters the bearing film at the pocket circumference, which is the curtain flow area $\pi d_R h$. The orifice geometry should be designed such that the curtain flow area $\pi d_R h_o$ is at least twice the orifice flow area $0.25\pi {d_o}^2$. This ensures that predominantly pocketed compensation is achieved:

$$\frac{\pi d_o^2}{4} \times \frac{10^3}{\pi d_R h_o} = \frac{d_o^2 \times 10^3}{4 d_R\, h_o} \leq 0.5 \tag{9.3.50}$$

Furthermore, the pocketed orifice recess depth is equal to, or greater than, the orifice diameter. These measures ensure that the flow area $\pi {d_f}^2/4$ is at least twice the curtain flow area:

$$\frac{\pi d_F^2}{4} \times \frac{10^3}{\pi d_f h_o} = \frac{d_F \times 10^3}{4h_o} \geq 2 \tag{9.3.51}$$

This ensures that the orifice diameter acts as the predominant restriction in the bearing.

Feasibility Determination

For an initial bearing design assessment, performance characteristics for journal bearings are given in Figures 9.3.37–9.3.39. The data allows a feasibility study to be undertaken before detailed calculations are made. For the equations given, any consistent set of variables may be used to determine appropriate bearing characteristics. These equations yield a bearing for general applications, and they have a realistic allowance made for dispersion losses between inlets (l/λ = 0.7). Figure 9.3.40 shows the load capacity nondimensionalized in terms of bearing projected area and supply pressure, and therefore gives a measure of the bearing efficiency. The effect of increasing L/D ratio is to increase circumferential losses, which adversely affect the pressure profile within the bearing clearance and hence reduces the bearing efficiency. However, this is more than offset by a larger bearing area and therefore the total capacity is increased. An undesirable feature of the loading characteristic is the negative stiffness region which occurs at high eccentricity, particularly on large L/D ratio bearings ($L/D > 1$). This phenomenon is known as static instability or lockup and may be seen as a sudden collapse of the air film when the region is approached.

Line Feed Correction Factor $1/\lambda$

The effect of dispersion losses around the bearing is accounted for by the line feed correction parameter $1/\lambda$. The effect of $1/\lambda$ on load capacity is illustrated in Figures 9.3.41. The line feed correction $1/\lambda$ is determined from Figure 9.3.18, where n = number of inlets/row, N = total number of inlets in the bearing, and:

$$\xi = \frac{L}{D} \quad \text{for central admission } (a/L = 0.5) \tag{9.3.52a}$$

Parameter	Pocketed orifices	Inherently compensated orifice
Stiffness (N/μm) at $\varepsilon = 0$	$K = \frac{\overline{K}\,(P_o - P_a)\,D^2}{h_o}$	$K = \frac{0.67\,\overline{K}\,(P_o - P_a)\,D^2}{h_o}$
Angular stiffness (Nm/rad) at $\varepsilon = 0$	$K_A = \frac{(0.043\text{-}0.003\,(L/D)^2)\,L^3\,D(P_o - P_a)}{h_o}$	$K_A = \frac{(0.033\text{-}0.002\,(L/D)^2)L^3\,D(P_o - P_a)}{h_o}$
Max. Load (N) at $\varepsilon = 0.5$	$W = \overline{W}\,(P_o - P_a)D^2$	$W = 0.67\,\overline{W}\,(P_o - P_a)D^2$
Air flow rate (m^3/s) at $\varepsilon = 0$	$G = \frac{0.235\,h_o^3\,P_o 2}{3.42 \times 10^6\,\xi}$	
Orifice diameter (mm)	$d_o = \sqrt{\frac{\Lambda_s\,\xi\,P_o\,h_o^3\,D}{7890\,nL}}$	$d_f = \frac{\Lambda_s\,\xi\,P_o\,h_o^2\,D}{31.55\,nL}$
Orifice pocket diameter	$d_R = \frac{d_o^2 \times 10^3}{2h_o}$	Not applicable
Pocket depth (mm)	$b \le \frac{0.2\,DLh_o}{N\,d_R^2 \times 10^3}$	Not applicable

Figure 9.3.37 Design equations for optimized double entry journal bearings where a/L = 0.25. Figure 9.3.39 gives values for $\Lambda_s\xi$. Units of pressure are N/mm^2. Units of dimensions are mm. Units of bearing gap h_o are μm.

$$\xi = \frac{L}{2D} \quad \text{for double admission } (a/L = 0.25) \tag{9.3.52b}$$

$$\frac{d}{D} = \frac{d_f}{D} \quad \text{(inherently compensated bearings)} \tag{9.3.53a}$$

$$\frac{d}{D} = \frac{d_R}{D} \quad \text{(pocketed bearings)} \tag{9.3.53b}$$

Typical values of l/λ used in practice are between the values of 0.5 and 0.8, which may be considered the optimum range.

Feeding Parameter $\Lambda_s\xi$

Load capacity is plotted against feed parameter $\Lambda_s\xi$ in Figure 9.3.41, where:

$$\Lambda_s\xi = \frac{7.89 \times 10^3\,nd_o^2}{P_o\,h_o^3}\,\frac{L}{D} \quad \text{(pocketed orifices)} \tag{9.3.54a}$$

$$\Lambda_s\xi = \frac{31.55\,nd_f}{P_o\,h_o^2}\,\frac{L}{D} \quad \text{(inherently compensated orifices)} \tag{9.3.54b}$$

For a particular l/λ value, load capacity is maximized at a particular value of $\Lambda_s\xi$. l/λ significantly affects load capacity, with lower values of l/λ yielding lower load capacity.

Bearing Stiffness

In general, as the eccentricity increases, stiffness reduces to values considerably lower than at $\varepsilon = 0$. It is therefore good practice for engineers to design for, and specify concentric stiffness which will necessarily have an affect on the load that bearings can support in the concentric or near-concentric condition. In other words, do not load a bearing to the point where stiffness decreases below an acceptable limit.

Parameter	Pocketed orifices	Inherently compensated orifice
Stiffness (N/μm) at $\varepsilon = 0$	$K = \frac{\overline{K}\,(P_o - P_a)D^2}{h_o}$	$K = \frac{0.67\,\overline{K}\,(P_o - P_a)D^2}{h_o}$
Angular stiffness (Nm/rad) at $\varepsilon = 0$	$K_A = \frac{(0.021 - 0.002\,(L/D)^2)\,L^3\,D\,(P_o - P_a)}{h_o}$	
Max. Load (N) at $\varepsilon = 0.5$	$W = \overline{W}\,(P_o - P_a)D^2$	$W = 0.67\,\overline{W}\,(P_o - P_a)D^2$
Air flow rate (m^3/s) at $\varepsilon = 0$	$G = \frac{0.235\,h_o^3\,P_o 2}{3.42 \times 10^6\,\xi}$	
Orifice diameter (mm)	$d_o = \sqrt{\frac{\Lambda_s\,\xi\,P_o\,h_o^3\,D}{7890\,nL}}$	$d_f = \frac{\Lambda_s\,\xi\,P_o\,h_o^2\,D}{31.55\,nL}$
Orifice pocket diameter	$d_R = \frac{d_o^2 \times 10^3}{2h_o}$	Not applicable
Pocket depth (mm)	$b \le \frac{0.2\,DLh_o}{N\,d_R^2 \times 10^3}$	Not applicable

Figure 9.3.38 Design equations for optimized single entry journal bearings. Figure 9.3.39 gives values for $\Lambda_s\xi$. Units of pressure are N/mm^2. Units of dimensions are mm. Units of bearing gap h_o are μm.

	L/D ratio			
	0.5	1	1.5	2
Pocketed orifices	0.5	0.42	0.63	0.7
Inherently compensated orifices	0.52	0.45	0.67	0.67

Figure 9.3.39 Values of $\Lambda_s\xi_{optimal}$ deduced from experimental data.

Angular Stiffness

In some design situations it may be necessary to determine the angular stiffness of the bearing, which is the resistance to angular movements along the axis of the bearing. A journal bearing subjected to angular torque is shown in Figure 9.3.42. Nondimensional torque at $\varepsilon_T = 0.5$ is plotted against $\Lambda_s\xi$ for various l/λ???n Figures 9.3.42 and 9.3.43 for L/D=1. Angular stiffness is independent of the type of compensation and the values shown apply for bearings with either pocketed or inherently compensated orifices. For single-admission bearings, Figure 9.3.42 shows that increasing l/λ and $\Lambda_s\xi$ both increase angular stiffness. The double-admission bearing torque characteristics are shown in Figure 9.3.43, where tilting affects restrictor pressures at the feeding planes. Unlike single-admission bearings, the value of angular torque maximizes at an optimum value of which fortunately corresponds to the value of $\Lambda_s\xi$, which leads to maximum load capacity.

The angular torque (tilt stiffness) of journal bearings may be calculated from the equations given in Figures 9.3.37 and 9.3.38. Note that for double-entry journal bearing the equations are different for pocketed orifice and inherently compensated orifices. When determining tilt stiffness for single-entry bearings it is seen from Figure 9.3.38 that a single value is given for both pocketed and inherently compensated bearings. This is because the differences between the tilt stiffness of bearings with the two orifice forms is negligible.

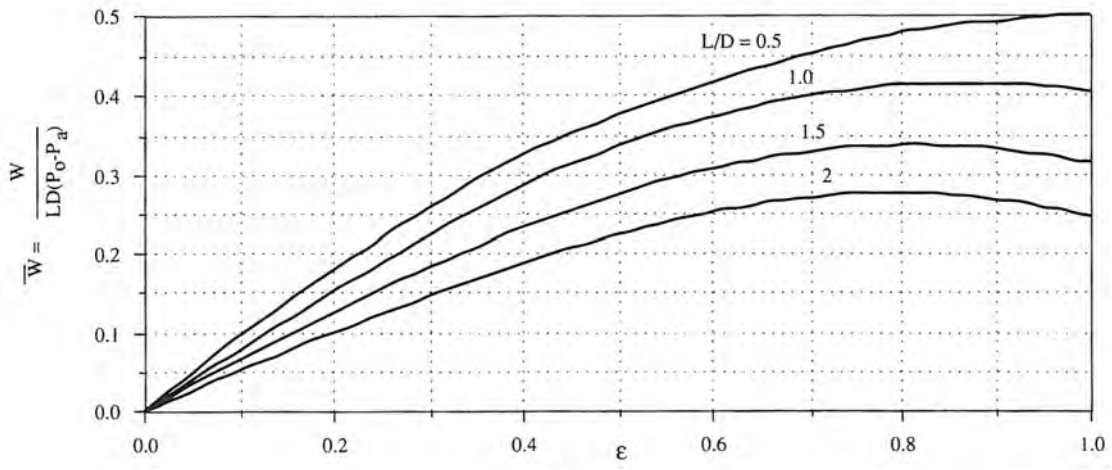


Figure 9.3.40 Load parameter for double-entry journal bearings with pocketed orifices where a/L = 0.25, $P_o/P_a = 5$, $\Lambda_s\xi = 0.5$, $1/\lambda = 0.7$, $C_d = 0.8$, and $\nu = 1.4$.

9.3.7 Material Selection and System Design[57]

The design and fabrication of an air bearing requires more than just determining the proper dimensions. Materials selection and system design considerations must also be studied carefully.

The correct selection of materials for aerostatic and aerodynamic bearings is of considerable importance if high reliability is to be achieved. Bearing materials must be able to tolerate rubbing contact, which may occasionally occur at start, stop, or overload conditions. Rubbing conditions which occur with aerostatic bearings differ from those which occur with aerodynamic bearings. In the former case, rubbing occurs at high loads but relatively low speed, while in the latter case, rubbing may occur with low loads at high speeds. When contact occurs, particularly in aerostatic bearings, few materials can withstand the high kinetic energy dissipated by the rotor. The design engineer should therefore ensure that adequate aerostatic capacity is available to withstand sudden overloads. Materials used in air bearing applications should have closely similar properties to plain bearing materials. In particular:

Corrosion Resistance: The designer is recommended not to use materials which differ in electrochemical potential by more than 0.25 V.

Thermal Expansion: It is wise to match thermal expansion rates as far as is practical. If precise matching is not possible, it is necessary to investigate the consequences and to adjust clearances if necessary.

Machinability: It is important that the materials selected are easily machinable if the close tolerances required are to be achieved.

Bushing Materials

Bronzes are highly recommended. Lead bronze is corrosion resistant and has good antiseizure properties. These materials can be easily repaired when damaged and should be used as thin shells shrink fitted into body bores.

Shaft Materials

The choice of materials for the shaft of an air bearing is very wide. The need for corrosion resistance is limited to the surfaces of the journal and thrust faces. Plating process should be avoided because plating on bearings has been known to peel and lie in the clearances or orifices with catastrophic effects. Often the material of choice is hardened stainless steel. If a ceramic (e.g., aluminum oxide) could be used as the rotor, the bushing could also be made from ceramic and a very robust bearing could be made.

[57] A good discussion of real-world design considerations is provided by J. Powell, Design of Aerostatic Bearings, Machinery Publishing Co., London.

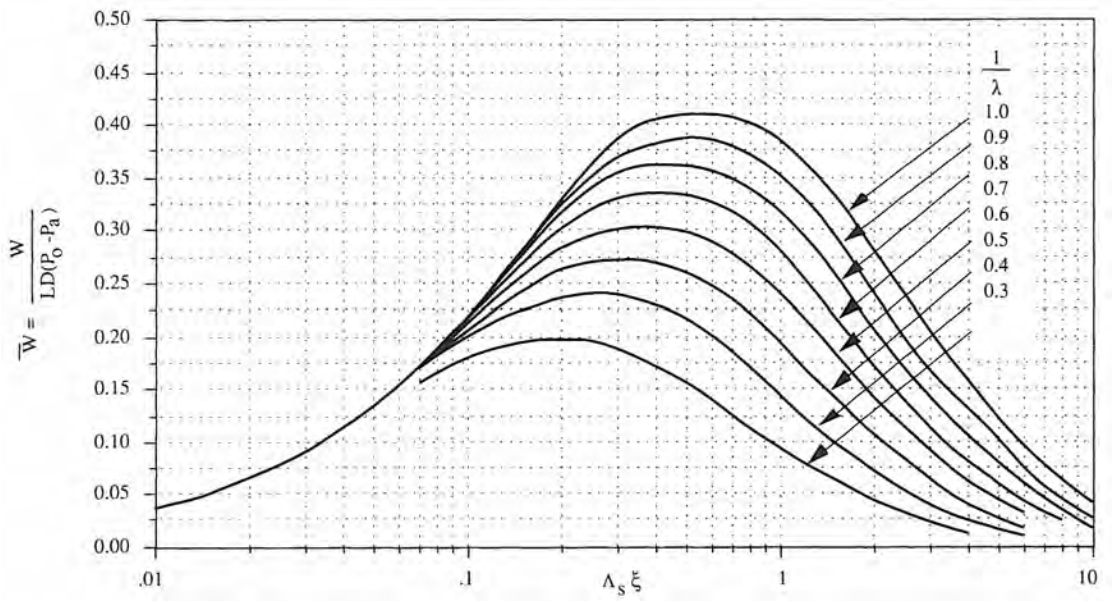


Figure 9.3.41 Load parameter as a function of $\Lambda_s \xi$ for double-entry journal bearings with pocketed orifices for various l/λ values where $\varepsilon = 0.5$, $L/D = 1$, $a/L = 0.25$, $P_o/P_a = 5$, $C_d = 0.8$, and $\nu = 1.4$.

Mechanical Installation

As design clearances are so small, it is important to minimize any stresses which may be imposed during assembly. Radial clamping tends to cause distortion irrespective of the care which is taken in design. Such distortions are likely to lead to seizure. The best method of locating a spindle assembly is by means of a bolted flange at one end of the bearing. This method of assembly has the disadvantage that the spindle body diameter has to be a good fit in the body of the housing. Distortion of the bearing surfaces may also arise because of temperature changes. In applications involving either high or low temperatures, the problem must be minimized to acceptable limits by matching rates of thermal expansion.

Filtration

Filters should be selected which will restrict the maximum particle size to one third of the minimum film thickness (assuming that this particle size is not detrimental to the pump or other components in the systems). An approximate guide to filter size is to assume 3 μm filtration per micron of the minimum film thickness. The filter should be positioned on the high-pressure side of the pump to prevent pump wear debris from contaminating the bearing. In addition, the filter construction should preclude any possibility of the filter material shedding and finding its way into the bearing. When orifice or capillary controlled bearings are employed, an additional safeguard may be achieved by utilizing coarse line filters immediately adjacent to the control device. Small sintered porous plugs may be obtained commercially for the purpose. When slot entry bearings are employed, higher-quality filtration is required to prevent blockage by silting in the feed slots. Regular filter maintenance should be ensured to maintain system reliability. In addition, filters should be fitted with a pressure drop indicator.

Electrical Circuit Breakers

Electrical circuit breakers should be introduced into the system on the high-pressure side of the filter to protect the bearing against sudden pressure failure. If these circuit breakers are introduced in conjunction with the provision of a pressure reservoir to cope with the bearing stopping period, good protection results. This is absolutely necessary for gas bearings. In addition, a supply pressure indication light should be incorporated. It is important to ensure that the bearings are not moved until the pressurized gas film is present. It is highly undesirable to rotate the bearings by hand in the absence of an air film, as this may damage them.

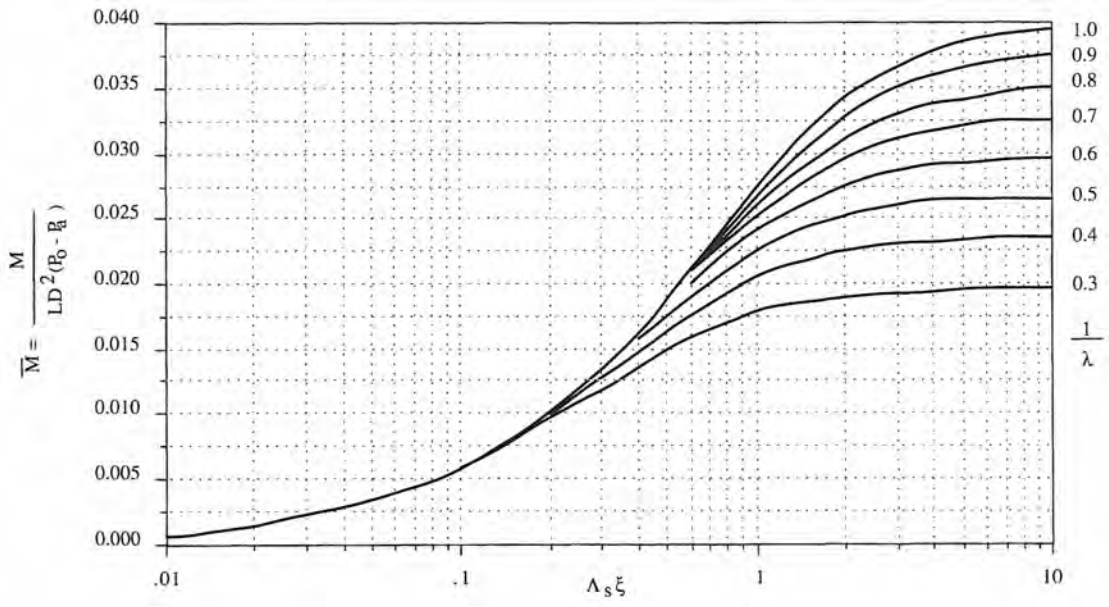


Figure 9.3.42 Torque parameter as a function of $\Lambda_s\xi$ for single-entry journal bearings with pocketed orifices with various l/λ values where $\varepsilon = 0.5$, $L/D = 1$, $a/L = 0.5$, $P_o/P_a = 5$, $C_d = 0.8$, and $\nu = 1.4$.

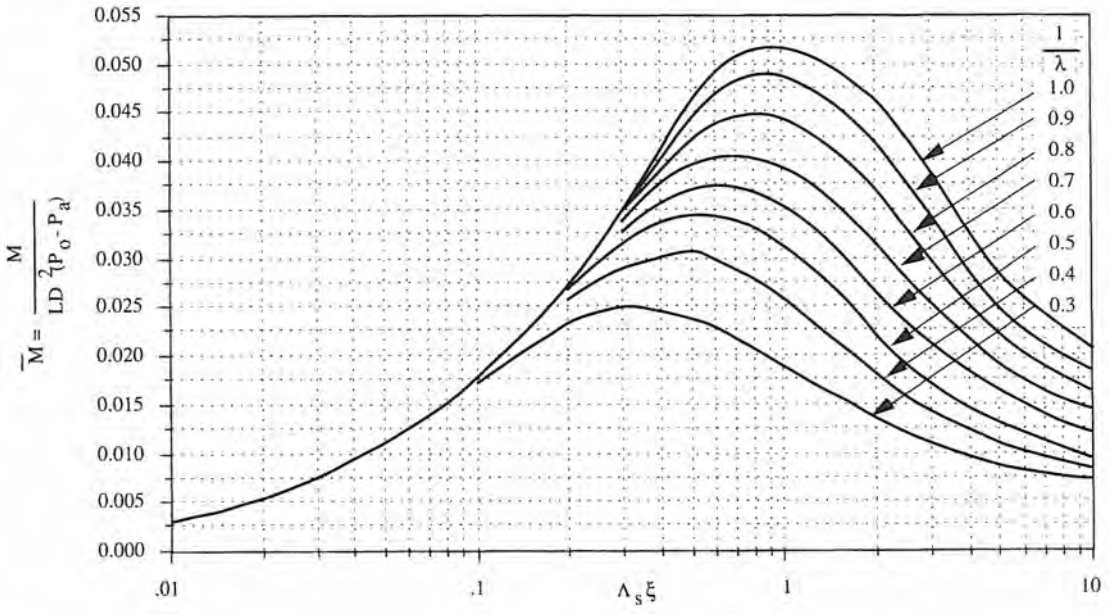


Figure 9.3.43 Torque parameter as a function of $\Lambda_s\xi$ for double-entry journal bearings with pocketed orifices with various l/λ values where $\varepsilon = 0.5$, $L/D = 1$, $a/L = 0.25$, $P_o/P_a = 5$, $C_d = 0.8$, and $\nu = 1.4$.

System Flushing After Assembly

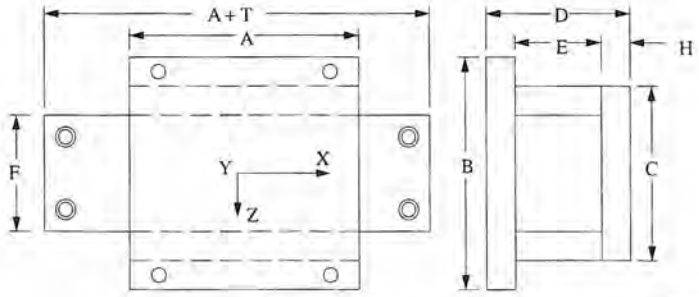
Flushing of the lubricating system during manufacture and after assembly is important and often inadequately performed in practice. Considerable damage may be caused initially to the bearings if insufficient care is taken during flushing.

Oil Vapor in Compressed Air

Oil vapor presents difficulties because of its tendency to condense into a wax-like deposit within the feed holes and bearing clearances. An effective method of preventing oil vapor reaching the bearing is to have an activated charcoal filter element in the supply line. This element must be replaced periodically. If feed holes do become blocked, a solvent, such as industrial alcohol, can be passed into the system with the air supply for a short period to cleanse the bearing. If a porous air bearing is being used, one must make sure that the solvent will not dissolve the lacquer that is often used to tune the porosity.

9.3.8 Modular Aerostatic Bearings[58]

Figures 9.3.44 and 9.4.45 show a family of aerostatic linear carriages and rails that are used in a wide variety of applications.[59] Both the carriage and the rail are made from hard anodized aluminum with optional Teflon®-impregnated surfaces. Note that these rails are typically supported at their ends and the carriages travel between the supports; thus for long travels, the stiffness is dominated by that of the bearing rail.



	Minimum	Maximum
A (in.)	1.0	8.5
B (in.)	1.5	10.0
C (in.)	1.0	10.0
D (in.)	0.83	7.0
E (in.)	0.36	5.5
F (in.)	0.5	8.0
T: Travel (in.)	0.5	48.0

Figure 9.3.44 Geometry range of a typical family of modular linear motion air bearings with bearing rails supported at their ends. Approximately 30 standard air slides are available. (Courtesy of Dover Instrument Corporation.)

For short travel devices that are not subject to cutting forces (e.g., a measuring machine), often the desired accuracy can be achieved with this design for a lower cost than a T-shaped carriage. These units are also often used for long-distance, high-speed shuttle carriages, where straightness over the length of travel is not crucial. To drive the carriages through their center of mass, the rail can be machined through its center and a linear motor or leadscrew installed. Air bearing journals can also be designed to ride on round ways like ball bushings. Modular air bearing bushings are also available with shaft sizes from 0.25-1 in. and prices ranging from $170 to $250. When lateral stiffness and accuracy are to be maximized, a T configuration should be used as shown in Figures 9.3.46 and 9.3.47.

Air bearing pads, shown in Figure 9.3.48, are also available, so designers can create their own carriage configurations. The load capacity and stiffness values are a function of the orifice

[58] There are many manufacturers of modular aerostatic bearings and the wise design engineer will obtain performance specifications and price and delivery quotes from all reputable manufacturers before choosing a particular bearing.

[59] Dover Instrument Corp., 200 Flanders Road, P.O. Box 200, Westboro, MA 01581-0200, (508) 366-1456.

	Minimum	Maximum
Z-Axis stiffness (lb/in.)	15,000	2,500,000
Z-Axis working load (lb)	1	200
Y-Axis stiffnesss (lb/in.)	15,000	1,000,000
Y-Axis working load (lb)	1	150
Pitch stiffness (in.-lb/rad)	1,000	15,000,000
Pitch working load (in.-lb)	0.5	700
Yaw stiffness (in.-lb/rad)	1,000	15,000,000
Yaw working load (in.-lb)	0.5	700
Roll stiffness (in.-lb/rad)	500	15,000,000
Roll working load (in.-lb)	0.5	1,000
Air pressure (psi)	60	100
Air flow (SCFM)	0.1	1.0

Figure 9.3.45 Load characteristics of a typical family of modular linear motion air bearings with bearing rails supported at their ends. These values do not include the properties of the bearing rail. Approximately 30 standard air slides are available. (Courtesy of Dover Instrument Corporation.)

and vacuum pocket layout. The pads are typically designed to optimize a particular application requirement. These pads cost on the order of $225 for small ones, $375 for medium-sized ones, and $975 for the large ones. For nonopposed pad designs, a vacuum preload pad is often used and these pads cost on the order of $715 and $1595 for medium and large pads, respectively. Care must be taken to ensure that the pads remain parallel to the bearing rail surface after the pads are attached to the carriage. This could be accomplished through strict control of the final machining of the pad mounting surfaces or by replicating the mounting surfaces. A circular thrust pad can also be easily designed or purchased off-the-shelf.[60]

For two-dimensional planar applications, a platten can be kinematically supported by three pads and spherical seats used to ensure that the bearing faces are not misaligned with respect to the surface; however, greater stability with a smaller footprint can be obtained with four pads, but care must be taken to ensure that all the pad surfaces are in the same plane. This is readily accomplished by precision grinding, scraping, or replication.

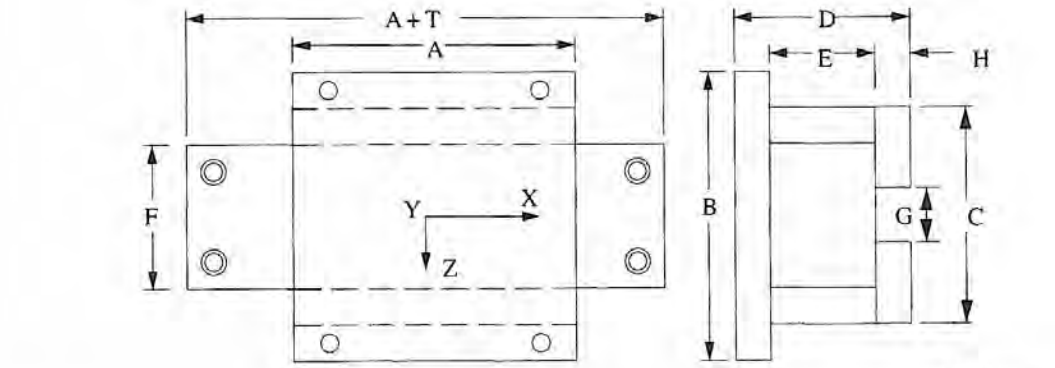
Air bearing spindles can be made to operate so quietly and smoothly that it is difficult to see if they are even rotating. Unlike ball bearing spindles, air bearing spindles give little or no warning when then fail. When an air bearing spindle fails, even at high speed, it can stop in less than a revolution. The resultant inertial forces can rip the spindle apart. It is wise to design blow proof housings around high-speed spindles.

Air bearing spindles can be designed in a manner similar to ball bearing spindles with a large shaft length/shaft radius ratio, which allows several spindles to be ganged together for drilling or milling applications, or for use as a grinding wheel spindle.[61] Alternatively, tilt stiffness can be achieved using two large annular thrust bearings. The latter design minimizes shaft bending and possible misalignment of the radial journal bearings so total radial error motion can be minimized. Figures 9.3.49–9.3.51 show modular spindles of this type[62] which carry the tradename Blockhead®. Figures 9.3.52 and 9.3.53 show the performance characteristics of Blockhead® spindles. These spindles can have total error motions on the order of 1 μin. An unmotorized 4 in. Blockhead® spindle costs on the order of $2700 while an unmotorized 10" Blockhead® costs on the order of

[60] A number of companies sell modular bearings of this type, including Dover Instrument Corp., Westboro, MA, (508) 366-1456; Fox Instrument and Air Bearings Corp., Livermore CA, (415) 373-6444; NSK Corp., Chicago, IL, (312) 530-5777; and Schneeburger Inc., Bedford, MA, (617) 271-0140.

[61] Spindles of this type are commercially available from Federal Mogul Corp., which has worldwide sales offices. Federal Mogul Corp., Westwind Division, 745 Phoenix Drive, Ann Arbor, MI 48108, (313) 761-6826.

[62] Air bearing spindles of this type (over 20 standard models) are also available from Dover Instrument Corp.



	Minimum	Maximum
A (in.)	4.00	15.00
B (in.)	5.00	9.00
C (in.)		
D (in.)	3.00	6.00
E (in.)	1.50	3.00
F (in.)	3.00	6.00
G (in.)		
T: Travel (in.)	2.00	48.0

Figure 9.3.46 Geometry range of a typical family of modular linear motion air bearings with bearing rails supported along their length. Approximately 10 standard air slides are available. (Courtesy of Dover Instrument Corporation.)

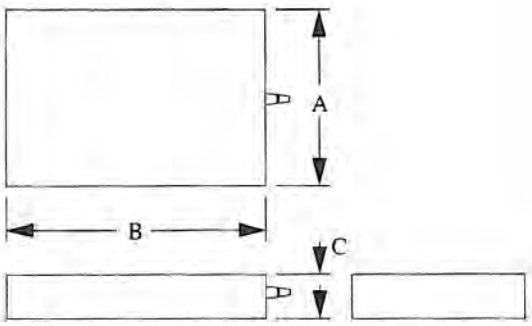
$8900. Note that a brushless dc motor drive with a precision tachometer may add $4000-$8000 to the cost of this type of spindle.

A hemispherical bearing does not experience misalignment problems; however there exists coupling between the two orthogonal radial directions. Oak Ridge National Laboratory developed a large hemispherical journal, porous graphite spindle, and the technology was transferred to many different companies, some of which still produce the spindle.[63] The spindle has a 410 mm diameter, 50 mm thick faceplate and the body is about 700 mm long (from the front of the faceplate to the back of the housing) by about 470 mm high and 420 mm wide. The spindle shaft extends out the back beyond the 700 mm housing. The spindle shaft is 135 mm long and about 76 mm in diameter (3 in.) and it has a 50 mm through-hole to the faceplate. A direct-drive motor housing can be fitted to the spindle or a pulley and belt can be used to drive the spindle. The spindle is mounted to a machine bed by means of six 16 mm bolts through flanges on the base of the spindle. Figure 9.3.54 shows the characteristics of this type of spindle. With an integral drive motor, this type of spindle can cost from $30,000 to $60,000 depending on the level of accuracy and type of drive motor desired. In general, if greater capacity is needed, one should consider the use of a hydrostatic bearing.

[63] This design was developed at Oak Ridge National labs for their own internal use. The design and manufacturing technology was then transferred to a number of companies, including Rank Taylor Hobson in Keene, NH (formally Rank Pneumo), and Cranfield Precision Engineering Ltd. in Cranfield, Bedford MK43 0AL, England. See W. H. Rasnick et al., "Porous Graphite Air-Bearing Components as Applied to Machine Tools," SME Tech. Report MRR74-02.

	Minimum	Maximum
Z-Axis stiffnesss (lb/in.)	300,000	2,000,000
Z-Axis working load (lb)	20	150
Y-Axis stiffnesss (lb/in.)	200,000	1,000,000
Y-Axis working load (lb)	15	125
Pitch stiffness (in.-lb/rad)	300,000	30,000,000
Pitch working load (in.-lb)	30	1,000
Yaw stiffness (in.-lb/rad)	300,000	22,000,000
Yaw working load (in.-lb)	30	800
Roll stiffness (in.-lb/rad)	400,000	9,000,000
Roll working load (in.-lb)	50	1,000
Air pressure (psi)	60	100
Air flow (SCFM)	0.4	2

Figure 9.3.47 Load characteristics of a typical family of modular linear motion air bearings with bearing rails supported along their length. These values do not include the properties of the bearing rail. Approximately 10 standard air slides are available. (Courtesy of Dover Instrument Corporation.)



Reference:	A	B	C
1P	2	4	0.75
2P	4	6	1
3P	6	10	1.5
2V	6	10	1.5
3V	6	10	1.5

Figure 9.3.48 Geometry range of a typical family of modular linear motion air bearing pads. Type P is often used in an opposed pad mode. Type V is vacuum preloaded and is not used in an opposed pad mode. (Courtesy of Dover Instrument Corporation)

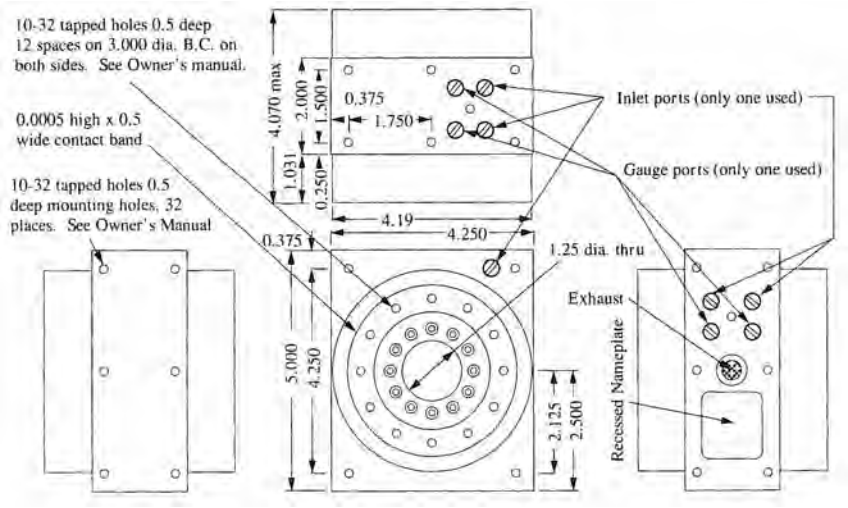


Figure 9.3.49 Blockhead® spindle model 4B. (Courtesy of Professional Instruments Co.)

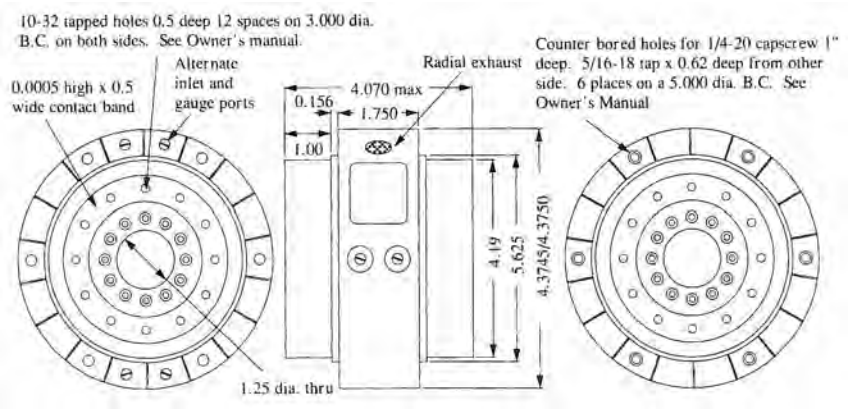


Figure 9.3.50 Blockhead® spindle model 4BR. (Courtesy of Professional Instruments Co.)

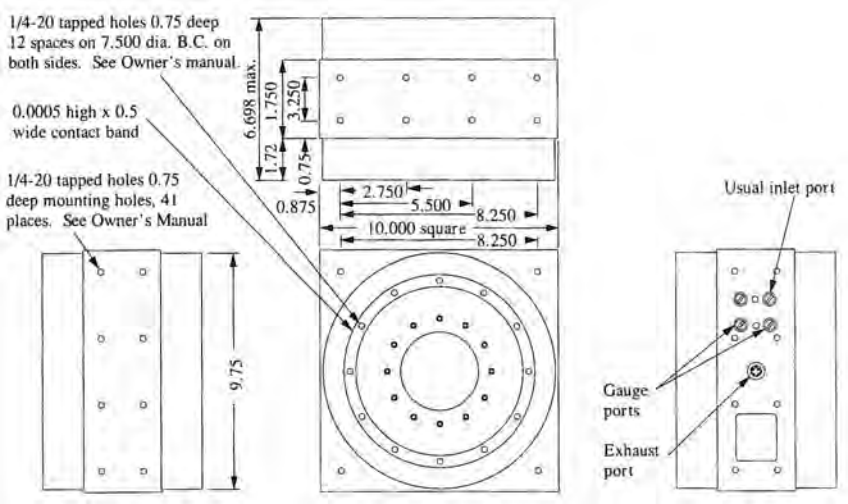


Figure 9.3.51 Blockhead® spindle model 10B. (Courtesy of Professional Instruments Co.)

Model	$F_{rad\ max}$ (N)	$F_{rad\ work}$ (N)	$F_{ax\ max}$ (N)	$F_{ax\ work}$ (N)	M_{max} (N-m)	M_{work} (N-m)	K_{radial} (N/μm)	K_{axial} (N/μm)	$K_{angular}$ (N-m/μrad)
4B & 4R	441	225	1764	892	45	23	118	353	0.45
10B	1764	892	10780	5292	539	274	225	1764	11.76

Figure 9.3.52 Load and stiffness characteristics of Blockhead® spindles operating with a 10 atm supply pressure. Loads and stiffnesses scale linearly with pressure. Since axial loads and moments are supported by the same bearing, the sum of the ratios of applied loads to allowable loads must be less than 1: $F_{axial}/F_{\ axial\ work} + M/M_{work} \leq 1$. Ultimate load capacity occurs when metal-to-metal contact is made. (Courtesy of Professional Instruments Co.)

Model	Axial error (μm)	Radial error (μm)	Angular error (μrad)	Air flow (liter/min)	Rotating weight (N)	Total weight (N)
4B & 4R	<0.05	<0.05	<0.2	60	34.3	323.4
10B	<0.05	<0.075	<0.2	120	80.4	647.8

Figure 9.3.53 Characteristics of Blockhead® spindles operating with a 10 atm supply pressure. Flow varies with the square of the pressure. (Courtesy of Professional Instruments Co.)

Maximum speed	1500 rpm
Working radial load (at faceplate)	1900 N
Working axial load	1900 N
Radial stiffness	525 N/μm
Axial stiffness	525 N/μm
Axial thermal drift	2 μm in 75 minutes from 0 to 1000 rpm.
Supply pressure	6.2 atm.
Flow rate	283 liters/min (maximum)
Rotor mass	206 kg
Shaft inertia	2.1 kg-m^2
Overall mass	650 kg
Power consumption @1000 rpm	<0.11 kW

Figure 9.3.54 Characteristics of the Oak Ridge porous graphite hemispherical journal spindle.

9.4 MAGNETIC BEARINGS[64, 65]

It was first proven mathematically in the late 1800s by Earnshaw that using only a magnet to try and support an object represented an unstable equilibrium; however, it was found in the 1930s that by using an electromagnet and measuring the air gap and using it as a feedback parameter, the system could be stabilized. Although it is beyond the scope of this book to discuss how magnetic bearings are designed, an attempt will be made to introduce the reader to some of the characteristics of magnetic bearing-supported systems. Magnetic bearings will most likely become more commonplace in the machine design engineer's world as the quest begins for machines with nanometer accuracy to manufacture next generation microelectronic and optical components.

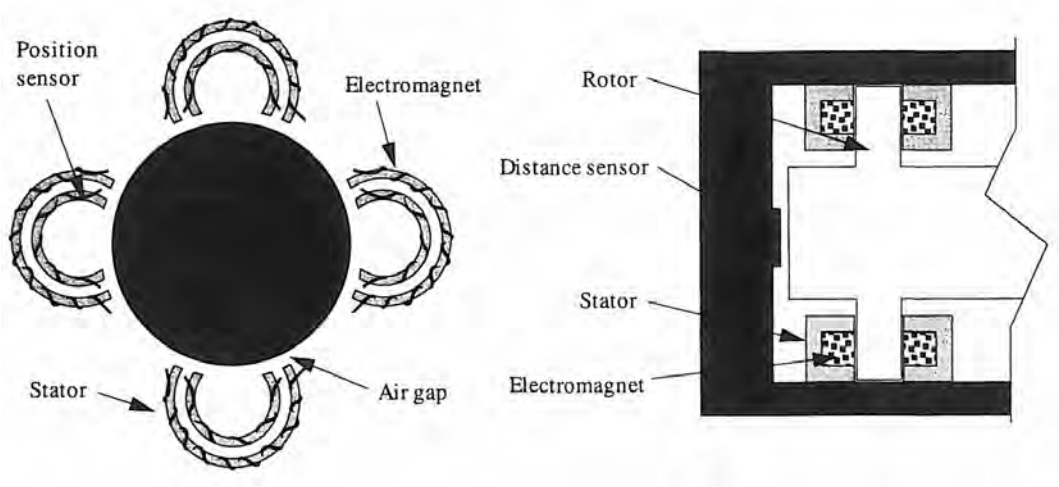


Figure 9.4.1 Magnetic bearing configurations for supporting radial and thrust loads.

Basic Operating Principles

Magnetic bearings are most often used to support radial and thrust loads in rotating machinery. Common design configurations are shown in Figure 9.4.1. The coils in magnetic bearings have virtually infinite life, but the control system can be affected by power outages or component failure; thus auxiliary rolling element bearings must be incorporated into the design as shown in Figure 9.4.2. The rolling element bearings operate at half the magnetic bearing air gap. Should the magnetic bearing control system fail, the auxiliary bearings would catch the spinning rotor before it contacted the magnetic bearing coils and destroyed them.

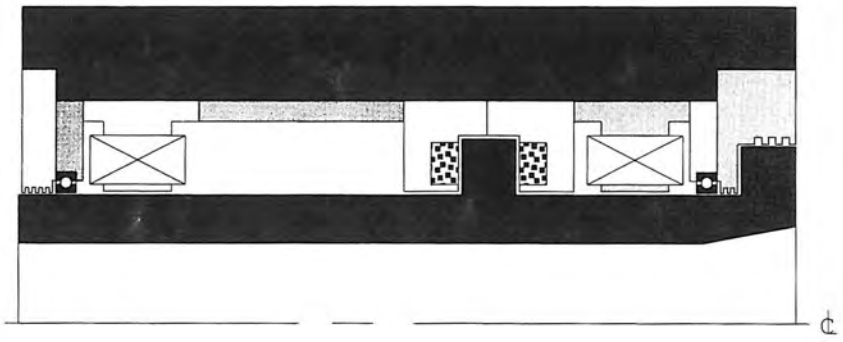


Figure 9.4.2 Conceptual design of a magnetic bearing spindle.

[64] Magnetic bearing actuators are a variation on moving coil and solenoid actuator design discussed in Section 10.4.2. This section was written with considerable help from David Eisenhaure of SatCon Technology Corp., 12 Emily St., Cambridge, MA 02139.

[65] Also see B. V. Jayawant, Electromagnetic Levitation and Suspension Techniques, Edward Arnold Publishers, Ltd., London, 1984.

Magnetic bearings can also be used to support linear motion devices. A planar version of the horseshoe-shaped magnets often used to support rotating shafts can be used, but this causes vertical motion control to be coupled with angular motion control. An alternative is to use a number of round bearings, similar to those used to resist shaft thrust loads and shown in Figure 9.4.3, in a kinematic configuration. This particular bearing has a bias force of about 30 N supplied by the permanent magnet and a control force of about 15 N supplied by the coil. It is discussed further in Section 9.4.2.

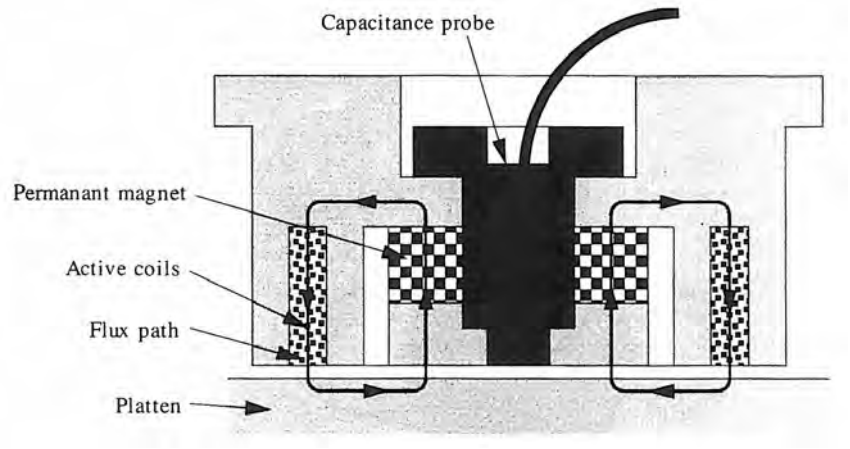


Figure 9.4.3 Magnetic bearing design used in a kinematic arrangement of magnetic bearings for supporting a precision linear motion. (Courtesy of SatCon Technology, Inc.)

Regardless of the type of load supported, magnetic bearings require a closed-loop control system for stability, as shown schematically in Figure 9.4.4. Typically, an analog control loop is used for coarse position control and a digital loop is superimposed on it for fine motion control and compensation for analog component drift. An analog sensor would be used with the analog control system. A laser interferometer, which has digital output, can be used as a high-resolution, high-bandwidth sensor for a fine-motion control digital control system.

Some magnetic bearings use a permanent magnet to provide a bias force. The magnetic flux produced by the permanent magnet acts only to levitate a portion of the object's weight in order to minimize the power expended by the active coil. The control force exerted on the object is thus proportional to the current supplied to the coil.

The attraction force between the bearing and the object is produced when energy stored in the magnetic field and in the air gap is transformed into mechanical work. The force required to do work is given by the spatial derivative of the energy W_m:

$$\overline{F} = \overline{\nabla} W_m \tag{9.4.1}$$

The energy stored is a function of the flux strength and the area of the gap:

$$W_m = \frac{B_g^2 A_g \ell_g}{2\mu_o} \tag{9.4.2}$$

where B_g is the magnetic flux, A_g is the air gap area, l_g is the gap width, and μ_o is the permeability of free space. By differentiating with respect to the gap width, the force as a function of gap width is obtained:

$$F_{gap} = \frac{dW_m}{d\ell_g} = \frac{B_g^2 A_g}{2\mu_o} \tag{9.4.3}$$

The flux is given by

$$B_g = \frac{\mu_o NI}{\ell_g} \tag{9.4.4}$$

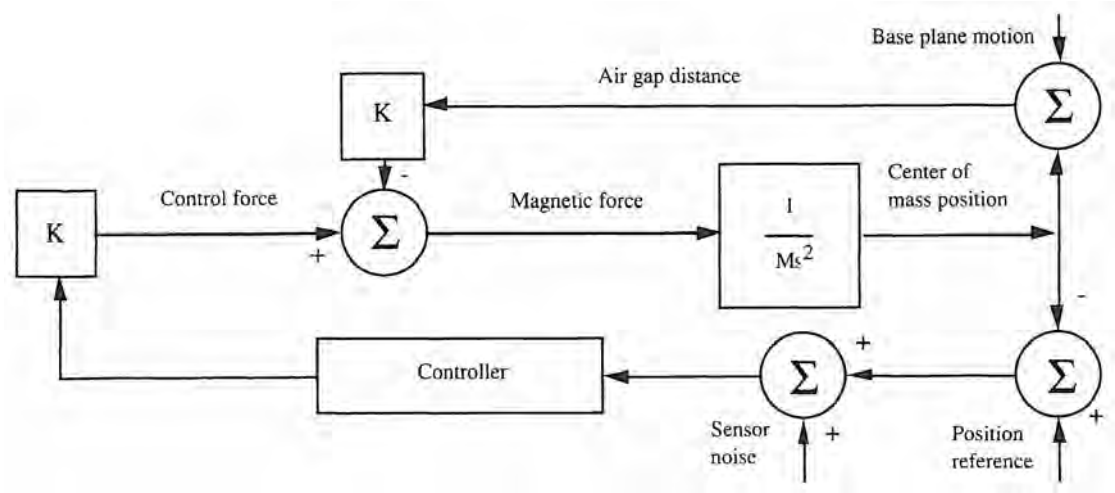


Figure 9.4.4 Basic magnetic suspension block diagram. (Courtesy of SatCon Technology, Inc.)

where N is the number of turns of the coil and I is the current through the coil. Substituting Equation 9.4.4 into 9.4.3 gives

$$F_{gap} = \frac{N^2 I^2 \mu_o A_g}{2\ell_g^2} \tag{9.4.5}$$

Any constant disturbance which causes a perturbation from the nominal gap length produces a proportional perturbance of the nominal force which tends to increase the gap length further if the disturbance is constant; hence closed-loop servo control is required for stability. The magnetic force is a nonlinear function of the input current and the gap length. A servo-controlled magnetic bearing operates by suspending an object at a nominal point. Therefore, it is reasonable to linearize the force equation about the nominal gap length and current. Linearizing the force in terms of the current I and gap length l_g is done by:

$$\delta F_{gap} = \frac{\partial f_p}{\partial I}|_o(\delta I) + \frac{\partial f_p}{\partial \ell_{go}}|_o(\delta \ell_g) \tag{9.4.6}$$

Simplifying, Equation 9.4.6 becomes

$$\delta F_{gap} = \frac{NB_{go}A_{go}}{\ell_{go}}\delta I - \frac{B_{go}^2 A_{go}}{\mu_o \ell_{go}}\delta \ell_g \tag{9.4.7}$$

The force exerted on the mass being levitated is given by

$$f = m\ddot{x} \tag{9.4.8}$$

The change in the mass position δx is equal and opposite to the change in gap length, $\delta x = -\delta l_g$. Substituting Equation 9.4.8 into 9.4.7, and taking the Laplace transform yields

$$\frac{\delta x}{\delta I}(s) = \frac{1}{m}\left[\frac{\frac{NB_{go}A_{go}}{\ell_{go}}}{s^2 - \left(\frac{B_{go}^2 A_{go}}{m\mu_o \ell_g}\right)}\right] \tag{9.4.9}$$

Let C_i be the coefficient associated with the current and C_g be the coefficient associated with the gap width; then

$$\frac{\delta x}{\delta I}(s) = \frac{1}{m}\left[\frac{C_i}{s^2 + C_g}\right] \tag{9.4.10}$$

Note that the poles of the system are symmetric about the jω-axis of the s-plane; thus the system is open-loop unstable. The unstable pole is known as the *minimum bandwidth frequency*, which is the minimum frequency for introduction of a lead filter to ensure good stability in a closed-loop system:

$$\omega_u = \pm\sqrt{\frac{B_{go}^2 A_{go}}{m\mu_o \ell_g}} \tag{9.4.11}$$

Note that the poles are a function of the flux density. A typical Bode plot would show that there is 180 degrees of phase shift for all frequencies and there is finite gain for the open-loop system.

The minimum required system bandwidth for stability is about 10 Hz for many suspensions. Maximum achievable bandwidths range from 100 Hz for a simple attractive-type system to 40 kHz or higher for voice coil systems (see Section 10.4). There is also a limit to the bandwidth of the system due to an additional pole representing the coil. This pole, called the L/R break pole, is usually ignored during the design of the control system as the pole is most often well beyond the bandwidth of the closed-loop system. However, it should be noted that it does limit the practical bandwidth of the system. Thus the required bandwidth of the system is selected to be below its L/R value.[66]

For simple, nonrotating systems a high-order analog lead filter is often used to compensate for the unstable pole. The purpose of the lead is to add phase at the crossover frequency (where the system gain is 0 dB) in order to increase the phase margin so as to guarantee stability. The additional pole due to the coil limits the increase in bandwidth to about a decade improvement.[67] The typical transfer function for such a filter is given by

$$H(s) = \frac{K}{s}\,\frac{(s + a\omega)(s + \omega)}{(s + b\omega)(s + c\omega)} \tag{9.4.12}$$

where a, b, and c may typically have values of 0.1, 10, and 50, respectively. Note that a free integrator (1/s) has been added to increase the apparent stiffness of the system by minimizing the steady-state error. This filter has the advantages of producing essentially infinite dc stiffness and no effect on the crossover frequency. The poles are placed well beyond the pole of the system to get the crossover frequency at around $3\omega_u$. This will produce good phase margin (on the order of 45^o) and increase the bandwidth of the system.

Results from attempts at increasing phase margin with lead compensators have been only marginal.[68] State-feedback implementations using states created by observers have had the best success. In particular, for high stiffness, integral feedback was shown to be highly successful in increasing the stiffness of the bearing. Linear quadratic regulator (LQR) design techniques have also proven useful for magnetic bearing control.[69]

9.4.1 General Properties[70]

Speed and Acceleration Limits

Magnetic bearings do not limit the speed or acceleration of components they support. Systems of 100,000 rpm and higher have been built for applications ranging from special pumps to spindles for ultrahigh-peed machining.

Range of Motion

Linear motion magnetic bearings can be used to support a carriage that moves linearly. In general, if the coils are stationary, then range of motion will be limited by the variation in force

[66] See, for example, T. Hawkey and R. Hockney, "Magnetic Bearings for an Optical Disk Buffer," Phase I, National Science Foundation SBIR Program, Final Report R08-87, Sept. 1987, SatCon Technology Corp.

[67] Ibid.

[68] K. Reistad et al., "Magnetic Suspension for Rotating Equipment; Phase I Project Final Report," Spin and Space Systems, Inc., Phoenix, AZ, for NSF Award DAR-7916703, SBIR Program 1980.

[69] B. Johnson, Active Control of a Flexible Two-Mass Rotor: the Use of Complex Notation, Sc.D. thesis, MIT, Mechanical Engineering Department, Sept. 1986.

[70] See, for example, D. Weise, "Present Industrial Applications of Active Magnetic Bearings," 22nd Intersoc. Energy Convers. Eng. Conf., Aug. 1987, Philadelphia, PA.

produced as the center of gravity moves with respect to the coils. Magnetically levitated trains have been proposed that use a whole series of coils that are energized as the train passes by them; however, the economics of such a design have yet to be proven. Other designs transfer power to coils on the train via a live rail. Perhaps high-temperature superconducting materials would help make magnetically levitated trains an economical reality.[71] Rotary motion magnetic bearing-supported systems are more common in industrial environments and are not motion limited.

Applied Loads

Virtually any magnitude load can be supported by a suitable magnetic bearing, depending on how much one wishes to pay and how much room one has. Increasing the proportion of the load that is supported by permanent magnets decreases the current that must pass through the coils and the resultant heat generated. The magnitude and frequency of applied disturbance forces combined with the controller bandwidth has an effect on achievable resolution as discussed below.

Accuracy

Typically, achievable rotational accuracy is 50 μm and 0.1 μm systems have been built. Since magnetic bearings depend on a closed-loop servosystem to achieve stability, the performance of the position sensor and servocontroller will directly affect the accuracy of the system. With new high-speed digital signal processor technology and better sensors for fine-position sensing, there is no reason why nanometer and better accuracy cannot be obtained if one wanted to pay for it. Note that magnetic bearings generate a relatively large amount of heat, so thermal control plays an enhanced role in achieving accuracy with magnetic bearings.

Repeatability

Repeatability of a magnetic bearing system depends on the sensor and control system. Impedance probes are commonly used as analog position sensors, so typical repeatability is in the micron range unless precision position sensors are used.

Resolution

Motion control resolution of the bearing gap is also limited by the sensor and control system. There is virtually no mechanical damping in a magnetic bearing-supported system unless the suspended object is also in contact with a viscous fluid. Hence disturbance forces acting on the system play an important role in determining motion control resolution of the bearing gap. At low frequencies, performance is determined almost completely by the ability of the controller to cancel disturbances. The primary control system parameter affecting disturbance cancellation is the controller gain, which determines suspension stiffness. The higher the suspension stiffness, the greater the ability to reject force disturbances. Depending on the nature of the source, at high frequencies, the disturbance forces are generally absorbed by the supported object's inertia and internal damping characteristics. Figures 9.4.5 and 9.4.6 show first-order estimates of how the disturbance force affects achievable resolution as a function of bearing stiffness and controller bandwidth for a system designed to support a 10 kg instrument platten. Figure 9.4.7 shows the equivalent carriage disturbance forces generated by support structure motion for the same instrument platten.[72]

A first-order estimate of the minimum resolvable gap length increment that the bearing can be servoed to can also be determined. If η is the proportion of the mass of the object supported by the permanent magnets in a magnetic bearing, then the net vertical force exerted on the object will be given by

$$F_{net} = (1 - \eta)\, mg \tag{9.4.13}$$

This is also the minimum force the coil must exert to levitate the mass. Typically, the coils are designed so that they can exert twice this force. Using an n bit A/D converter, the resolution of the control force is 2^n parts. The minimum resolved force $\varepsilon_{\Delta F}$ is given by

$$\varepsilon_{\Delta F} = \frac{(1 - \eta)\, mg}{2^{n-1}} \tag{9.4.14}$$

[71] See, for example, E. R. Laithwaite, Propulsion Without Wheels, Hart Publishing Co. New York, 1966; E. R. Laithwaite, Transport Without Wheels, Scientific Books, London, 1977; and B. V. Jayawant, Electromagnetic Levitation and Suspension Techniques, Edward Arnold Publishers, London, 1981.

[72] See Section 9.4.3.

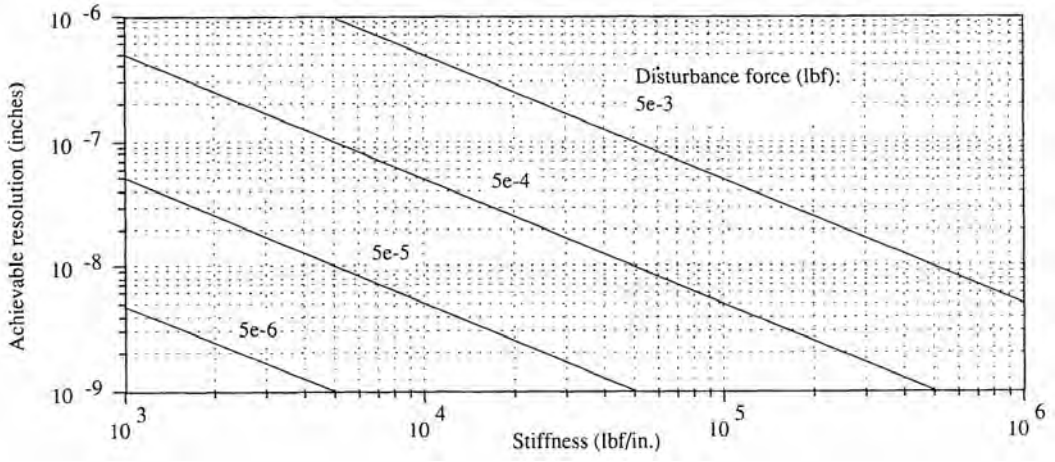


Figure 9.4.5 Achievable suspension resolution for a 10 kg platten and bearings with 150 N force capability. (Courtesy of SatCon Technology, Inc.)

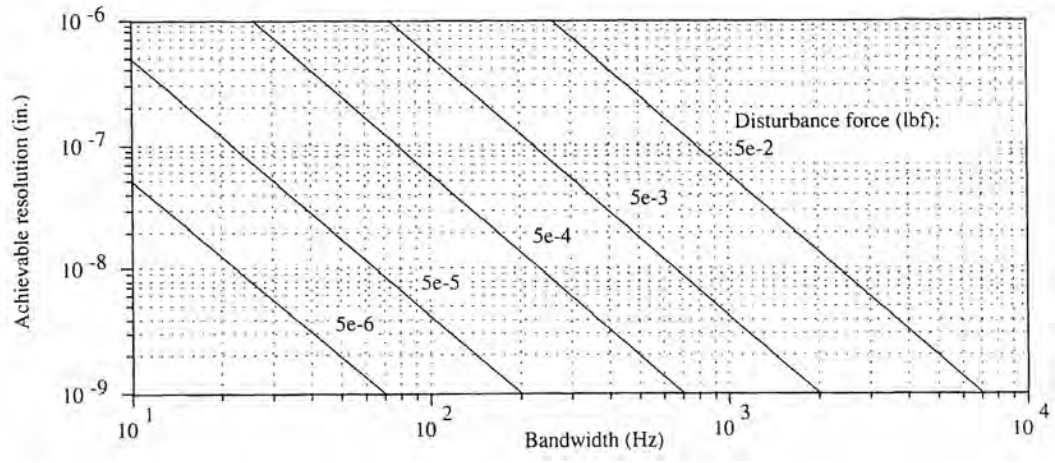


Figure 9.4.6 Achievable suspension resolution for a 10 kg platten and bearings with 150 N force capability. (Courtesy of SatCon Technology, Inc.)

The acceleration seen by the mass is given by

$$a = \frac{(1-\eta)g}{2^{n-1}} \tag{9.4.15}$$

The distance the mass moves toward the magnetic bearing due to the minimum resolved force acting during a time t is

$$\delta = \frac{(1-\eta)gt^2}{2^n} \tag{9.4.16}$$

This formula can be used in the calculation of the maximum allowable servo update time as well.

Motion Resolution of the Supported Load

Since there is no friction with magnetic bearings, motion resolution of an object supported by them is limited only by the actuator, sensor, and control system used.

Preload

Magnetic bearings are efficient only in the attraction mode. In order to obtain high performance in systems with randomly oriented force components, magnetic bearings should be used in an opposed mode design. For precision instrument plattens, it is feasible to use gravity to preload the bearing system.

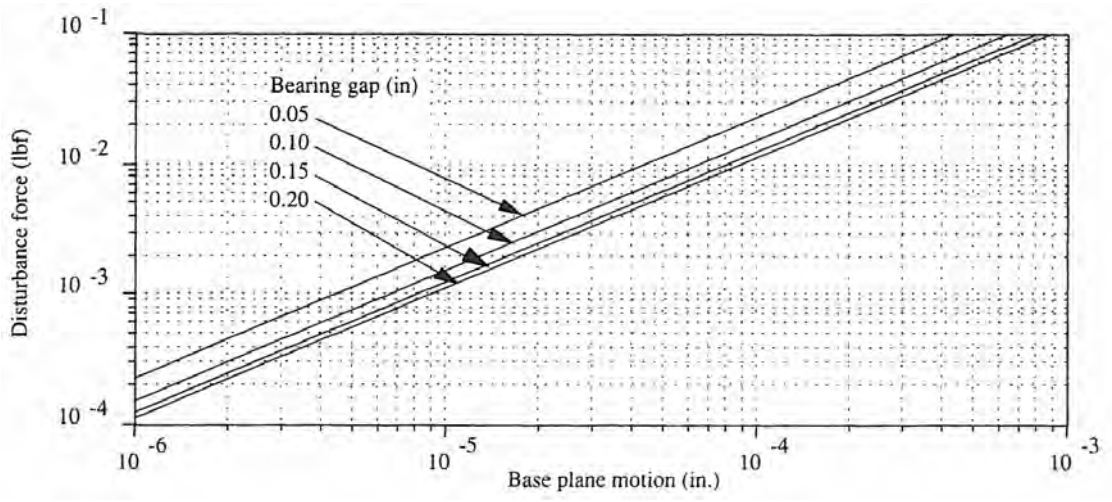


Figure 9.4.7 Slide disturbance force for a 10 kg platten and bearings with 150 N force capability. (Courtesy of SatCon Technology, Inc.)

Stiffness

The steady-state stiffness of magnetic bearings can be essentially infinite, depending on how the closed-loop control system is designed. Magnetic bearing dynamic stiffness depends on the frequency of the applied load and the bandwidth of the control system.

Vibration and Shock Resistance

There are several modes in which a magnetic bearing can be operated in order to actively control vibration[73]:

1. *Inertial axis control.* The frequency of rotation is measured, amplified, and subtracted from the control signal sent to the coils. This creates a zero stiffness condition for the bearing at the frequency of rotation. As a result, quasi-static and disturbance forces are still resisted, but the rotor is then free to spin about its inertial axis. This virtually eliminates dynamic rotor imbalance. Care must be taken, however, to "turn off" this inertial axis control when the rotor passes through critical speeds (natural frequencies).
2. *Peak of gain.* Instead of subtracting the rotational frequency component from the control signal, it can be added to achieve very high stiffness at the rotational frequency. This is in effect a feedforward control system which can greatly minimize total radial error motion for low-speed (<1000 rpm) systems.
3. *Vibration control.* A magnetic bearing's closed-loop control system can be used in conjunction with feedback from accelerometers to produce forces opposite to those created by vibration. The net effect is to cancel the vibration. Vibration can be reduced by 20 dB using this type of system.[74]
4. *Alignment.* Often just achieving proper alignment between rotating components can do wonders to minimize vibration. A magnetic bearing's ability to control the distance between the rotor and stator allows for precise alignment control of large components.
5. *Dynamic balancing.* Magnetic bearings can be used for in-situ dynamic balancing of components. Rotor speed, angular position, and gap displacement measurement information can be collected and used to determine h_o to balance the rotor.

Damping Capability

A magnetic bearing's damping capability is attained from the closed-loop control system. Additional magnetic bearing modules can be added at various points along a shaft and used as

[73] See, for example, W. S. Chung et al., "Ultra Stable Magnetic Suspensions for Rotors in Gravity Experiments," Precis.Eng., Vol. 2, No. 4 , 1980, pp. 183–186.

[74] D. Weise "Present Industrial Applications of Active Magnetic Bearings," 22nd Intersoc. Energy Convers. Eng. Conf., Philadelphia, PA, Aug. 1987.

vibration dampers. In this mode, the gap measurement signal is differentiated and used as a velocity feedback signal.

Friction

There is no friction, static or dynamic, associated with magnetic bearings. However, at high speeds hydrodynamic drag may become a problem if the bearing gap is too small.

Thermal Performance

Magnetic bearings can generate significant amounts of heat and therefore may require external cooling devices, such as recirculating chilled water jackets. For systems where the load does not vary greatly, a large percentage of the load can be supported by permanent magnets which minimize coil size and current required to levitate the load.

Environmental Sensitivity

As long as the coils are protected (e.g., hermetically sealed), magnetic bearings can operate in virtually any environment. They have been used successfully in the following environments: air with temperatures ranging from -235 to 450°C, 10^{-7} torr to 8.5 MPa, water, seawater, steam, helium, hydrogen, methane, and nitrogen. One must ensure that in a corrosive environment the system's materials do not fail to perform their structural or sealing functions.

Seal-ability

In a normal environment there is really no need to seal magnetic bearings; however, it is a good idea to protect the auxiliary bearings from becoming damaged by contamination.

Size and Configuration

Magnetic bearings are typically 2-10 times larger than the rolling element bearings they can replace; however, in many applications, accommodating a magnetic bearing's larger size is not too much of a problem.

Weight

Magnetic bearings are very heavy compared to the rolling element bearings they replace. In some applications, such as precision mechanical gyroscopes, the forces encountered by the bearing are so small that the weight of the required bearing is inconsequential anyway. In large stationary industrial applications, such as pipeline compressors or print roll diamond turning machines, bearing weight is not a primary design consideration.

Support Equipment

Magnetic bearings require a closed-loop servo system to make them stable. The servo system must have precision displacement sensors to measure the bearing gap and amplifiers for the output signal from the controller. To increase reliability, redundant control and power supply systems are often used on critical magnetic bearing applications such as those used in pipeline compressors.

Maintenance Requirements

Magnetic bearings have virtually no maintenance requirements. This makes them especially suitable for equipment that must be kept continually running, such as pipeline compressors.

Material Compatibility

Magnetic bearings require wound coils (typically, copper wire) for the stator, and an iron rotor or iron rotor laminations (to minimize eddy current losses). It is possible to hermetically seal the copper windings in a can to protect them from hostile environments. Similarly, it is possible to plate the iron rotor or laminations with a noncorroding material (e.g., chrome or nickel). Since magnetic bearings are noncontact devices and run dry, problems with material compatibility are usually not encountered.

Required Life

Since magnetic bearings are noncontact devices, they can have essentially infinite life.

Availability, Designability, and Manufacturability

Magnetic bearings are generally custom designed for the application and are thus as yet not available off the shelf except for some preengineered complete spindle assemblies. There are many successful commercial applications of magnetic bearings, such as[75]:

Pipeline compressor:

Speed:	5250 rpm.
Radial load:	14 kN.
Axial load:	50 kN.
Environment:	Rough industrial.
Advantages:	Drive power reduced from 26 kW to 3.4 kW.

Print roller diamond turning machine:

Speed:	0-1500 rpm.
Radial load:	18 kN.
Environment:	Machine shop.
Advantages:	High rotational accuracy (better than 1 μm) independent of load.

Turbomolecular vacuum pump:

Speed:	0-30000 rpm.
Radial load:	75 kN.
Environment:	Room temperature high vacuum.
Advantages:	Reduced pump size by allowing higher speeds.

Magnetic bearings' proven track record means that one should not be afraid to pursue their use if the application warrants.

Cost

Magnetic bearings are probably the most expensive type of bearing one can use; however, for the problems they solve, effective system cost can be low compared to design solutions that use other bearings.

9.4.2 Design Case Study: An Ultraprecision Magnetic Bearing Supported Linear Motion Carriage[76,77]

Ultimately, as quantum effect devices are developed for the microelectronics industry, machines with angstrom motion resolution over an area the size of an integrated circuit chip will be required. Several devices with angstrom motion resolution currently exist. For example, scanning tunneling microscopes (STMs) are piezoelectric-actuated flexural linkage structures with angstrom resolution and range of motion on the order of 1 μm.[78] Because of an STM's small range of motion, it can be made small enough to make it virtually immune to thermal and vibration problems that would plague an angstrom resolution machine with centimeter range of motion. Current designs of other precision machines such as diamond turning machines are principally performance limited by mechanical contact between moving parts, misalignment between actuators and bearings, bearing stability, and attainable temperature control.[79] To help overcome these problems, coarse-fine positioning systems have evolved as shown in Figure 9.4.8. Although coarse-fine systems can be effective, they are mechanically cumbersome and can be difficult to control. The design envisioned for an atomic resolution measuring machine (ARMM) would address some of these issues by way of its kinematic magnetic bearing design.

---

75 From Magnetic Bearings Inc. product literature.

76 This work was supported by the Center for Manufacturing Engineering of the National Institute of Standards and Technology, in conjunction with their Molecular Measuring Machine project. See E. Teague, "The NIST Molecular Measuring Machine Project: Metrology and Precision Engineering Design," J. Vac. Sci. Tech. A, Dec. 1989.

77 This section was derived from a paper by A. Slocum and D. Eisenhaure, "Design Considerations for Ultra-precision Magnetic Bearing Supported Slides," NASA Conf. Magn. Suspens. Technol., Hampton, VA, Feb. 2-5, 1988.

78 G. Binnig and H. Rohrer, "Scanning Electron Microscopy," Helv. Phys. Acta, Vol. 55, 1982, pp. 726-735.

79 See, for example, R. Donaldson and S. Patterson, "Design and Construction of a Large Vertical-Axis Diamond Turning Machine," SPIE's 27th Annu. Int. Tech. Symp. Instrum. Display, Aug. 21-26, 1983, and J. Biesterbos et al., " A Submicron I-Line Wafer Stepper," Solid State Technol., Feb. 1987, pp. 73–76.

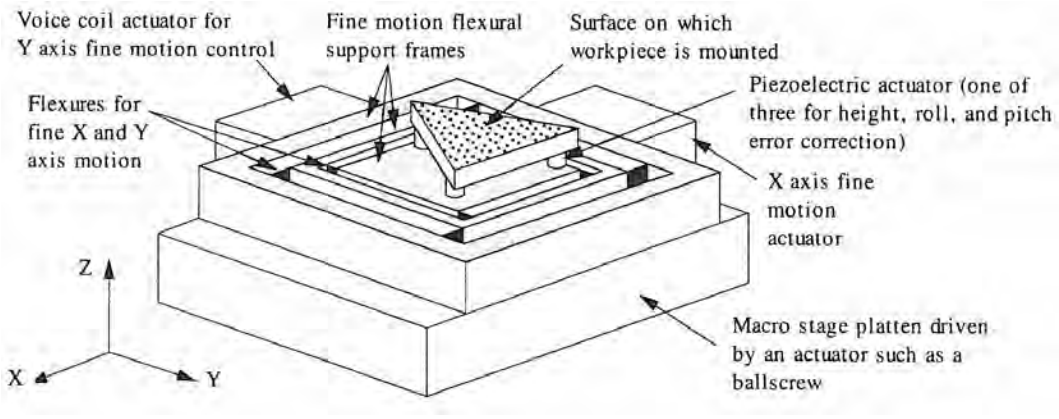


Figure 9.4.8 Example of a coarse-fine positioning system used to correct for slide errors caused by errors in slide geometry and forces caused by a misaligned actuator.

The design principle for the ARMM evolved from a crossed axis concept developed at the National Institute of Technology (NIST) for its Molecular Measuring Machine (M^3) project. A model of M^3 is shown in Figure 9.4.9. M^3 was designed as a sphere which would be radiantly coupled to an external temperature-controlled spherical shell. In this way the operating environment would be isolated from vibration effects encountered with forced fluid temperature control systems.

Early tests with sliding bearings showed that performance was a large function of the manufacturing process used to finish the bearing ways.[80] If the bearings did not perform as expected, then many months and dollars would be lost while the bearings were refinished for another try. What was needed was a bearing whose performance could be adjusted easily. At MIT, under sponsored research from NIST, the author designed a magnetically supported platten, with one-large-degree-of-freedom motion, illustrated in Figure 9.4.10. Ultimately, a two-axis version would have to reside in a sphere like the M^3. SatCon Technology Corp. designed the magnetic bearings, and a Lincoln Laboratory researcher designed the control system for the bearings.[81]

The remainder of this section briefly discusses some of the system design considerations for the ARMM, including bearing, actuator, sensor, motion control, and temperature control design issues. There are numerous other factors to consider, and an in-depth discussion of all the engineering calculations required would fill many volumes; however, even this limited discussion should give the reader a feeling for all the subtle factors that can affect the design.

Bearings

The properties desirable in a bearing for the ÅRMM include (1) ultrahigh resolution, (2) low friction, and (3) high stiffness. Only magnetic bearings have the potential to meet these requirements *and allow for easy adjustment of performance after fabrication*. The magnetic bearings designed for the ARMM prototype were shown in Figure 9.4.3 and consist of a steel disk-shaped shell with three concentric ring units. Each ring is called a pole of the bearing. A permanent magnet is placed in the center ring of the shell and a control coil lies in the gap between the two outer rings. The bearing has a hollow core to provide space for the gap-sensing element. The flux paths for the magnetic circuit for both the permanent magnet and the fully activated coil are also shown in Figure 9.4.3. The flux path design is the major element in magnetic bearing design, as it directly affects the amount of force that can be exerted and the efficiency of the bearing. Unfortunately, flux path design is beyond the scope of this book and will not be discussed further.

It should be noted that the magnetic bearing design for the prototype ARMM is not necessarily "optimal" with respect to the bearing itself. Space limitations on the width of the platten limited the diameter of the bearings; thus the poles have a very narrow cross section. The bearings could not be made wider as this would force the platten to be wider which would make the platten too heavy.

[80] NIST's M^3 will use a coarse-fine system to compensate for this problem.

[81] The author wishes to thank Jim Downer and Tim Hawkey of SatCon Technology Corp. for designing the bearings, and Dave Trumper, formally a Ph.D. student at MIT in the EE Dept. and now a professor of electrical engineering at UNC Charlotte, for designing the control system and getting the hardware up. Also thanks to Van Pham, whose S.B. thesis was to work with Dave on the controller design.

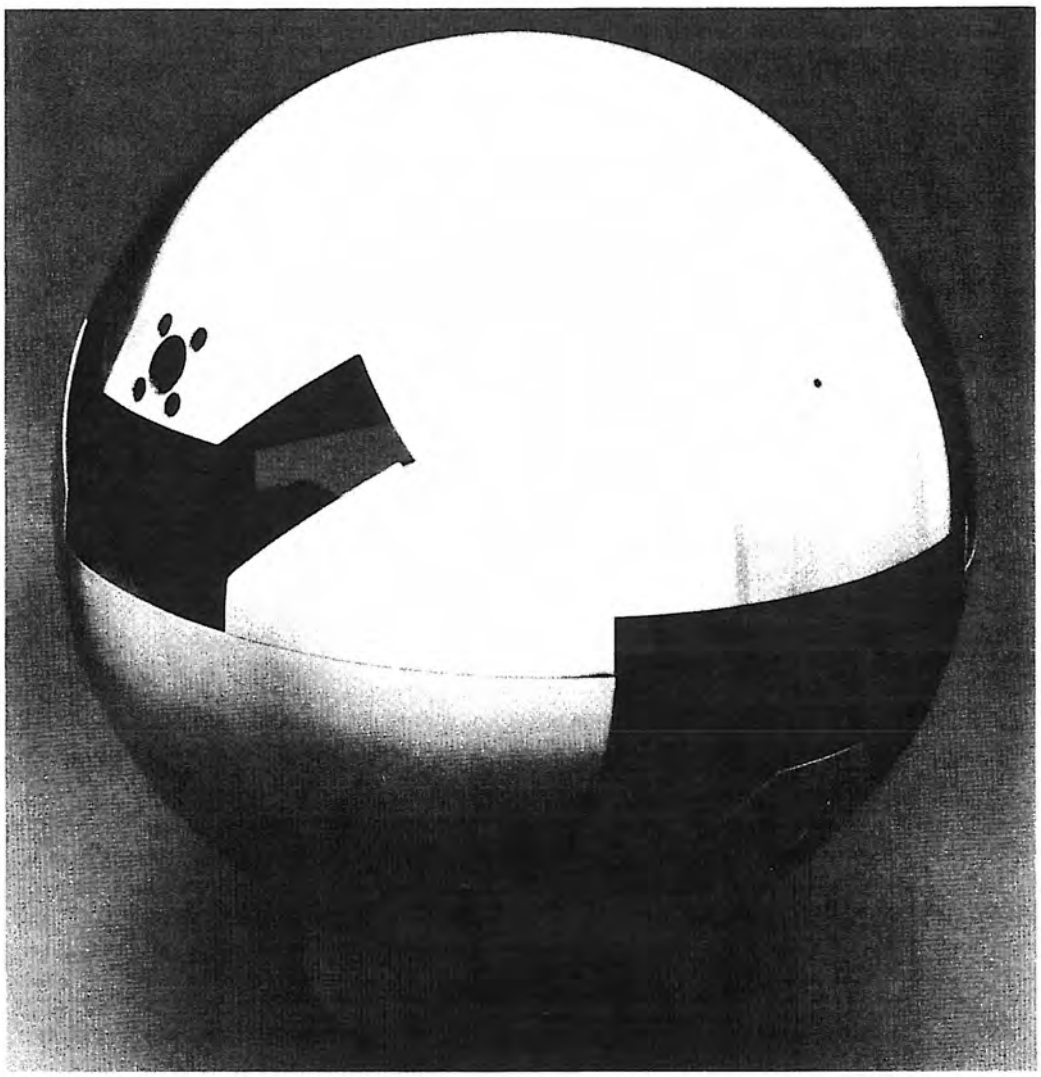


Figure 9.4.9 Model of NIST's molecular measuring machine. (Courtesy of the National Institute of Standards and Technology.)

Ideally, the cylindrical poles should be designed short and thick like a disk. The gap between the bearing and the platten is small to minimize flux leakage, but the gap between the concentric rings of the poles should be as wide as possible to provide a longer path for the flux. The longer the path, the more work the flux does and the more force is available to levitate the object.

Figures 9.4.5 and 9.4.6 showed the achievable bearing gap resolution at various total disturbance force levels as a function of suspension stiffness and bandwidth, respectively. The disturbance force represented was modeled as a broadband disturbance over the entire frequency range of interest. For the simple control system considered, the bandwidth is equal to the natural frequency of the system. The principal disturbance force acting on the carriage in a precision laboratory environment is caused by ground plane motion.[82] If necessary, the sensitivity to base motion could be reduced by one order of magnitude by incorporating magnetic flux feedback in the control loop. For this application, the total disturbance force level should be kept below about 0.01 N (0.005 lb).

[82] Commercially available servo-controlled systems are capable of keeping table motion amplitudes below 100 Å at 10 Hz. For example, see systems available from Barry Control and Newport Corp.

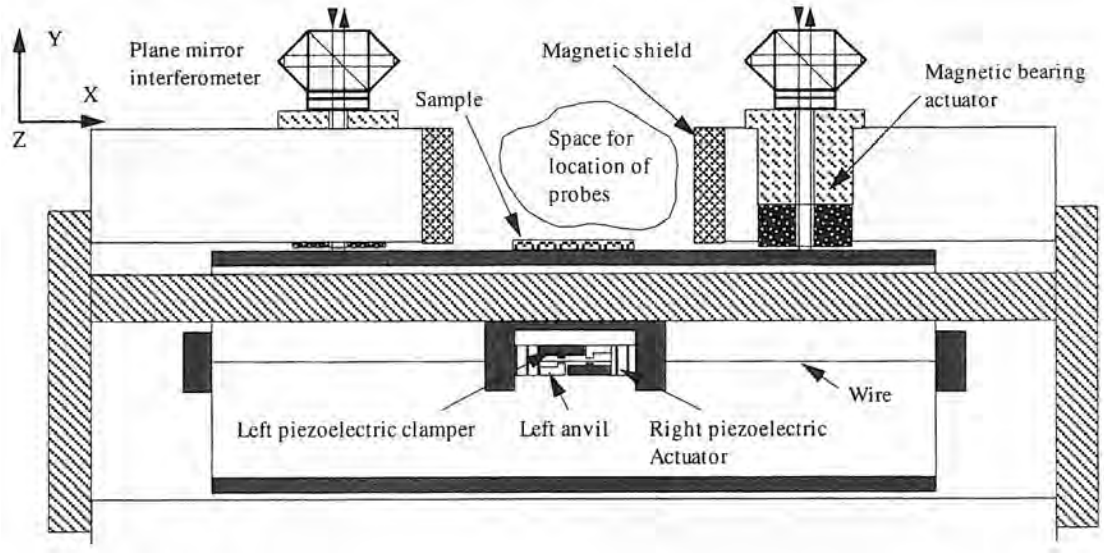


Figure 9.4.10 Cutaway side view of atomic resolution measuring machine (ARMM) with kinematic arrangement of magnetic bearings.

Actuators

For axial motion of the platten, there are numerous actuator possibilities for the ÅRMM, especially if one considers combinations of macro and micro motion systems. However, because of the inherent complexity of a coarse-fine actuator system, only direct-drive systems were considered including linear electric motors and piezoelectric microsteppers.[83] Either actuator used, however, would require a means to couple it to the platten without causing disturbance forces to be transmitted to the platten. In order to maximize the coupling efficiency, a wire-type coupling was used in the design.[84]

Linear electric motors have had a practical resolution limit on the order of 0.1-0.01 μm, based on the fact that mechanical coupling and thermal errors in existing systems began to dominate at this resolution. With the wire-type coupling and frictionless bearings, these problems could be avoided. By mapping the motor's response to current input and using a dc amplifier in place of a PWM drive,[85] it is conceivable that angstrom motion resolution could be obtained.

Piezoelectric microsteppers have high resolution and large ranges of motion. Their resolution is a function of the increment of each step, which can be on the order of angstroms, and how the device clamps between steps; a jerky clamping action may induce unacceptable disturbance forces. As shown schematically in Figure 9.4.10, the piezoelectric actuator designed for the ARMM could ideally deliver 0.1 Å axial resolution with less than 0.1 Å lateral error motion. Design of a control system to achieve smooth jerk-free motion from any microstepper actuator is potentially very difficult. On the other hand, the advantages offered by piezoelectric microsteppers include very low thermal energy generation and virtually infinite resolution.

Current actuator research for the ARMM is focusing on a linear electric motor design integrated with the bearing design to produce a two-axis motor whose bearing gap and planar position can be controlled simultaneously. This will eliminate the need for a coupling device between the motor and the platten which in turn allows one to design a two-dimensional large motion-controlled platten with servo-controlled bearing gap and pitch and roll control.[86]

---

[83] Piezoelectric actuators are discussed in Section 10.5.

[84] See Section 10.8.1 for a discussion of coupling methods, including a detailed discussion of wire-type couplings. Section 10.9 discusses a control system that makes the wire appear axially stiffer than it actually is.

[85] See Section 10.3.

[86] Two-dimensional plattens currently exist that are driven by Sawyer motors as discussed in Section 10.3; however these systems cannot control the bearing gap, pitch, or roll and thus require the use of a fine motion stage for precise (better than submicron) final positioning.

Sensor Systems

There are four principal sensor systems that need to be designed for the ÅRMM:

1. *Environmental.* When the ARMM is operated outside of a vacuum, the temperature, pressure, and humidity all need to be monitored to enable the environment to be controlled to ensure accuracy of the machine. It is anticipated that temperature control in the structure of the machine will have to be on the order of 0.01-0.001 C°/cm.
2. *Position feedback sensors.* In order to keep the actuators decoupled from one another, a through-the-bearing measurement method should be used. This maintains accuracy of the kinematic model and provides for checking measurement closure should direct measurement of the position and orientation of another region of the carriage (e.g., sample area) also be made. Through-the-bearing analog sensors would be used to stabilize the platten coarsely with analog control loops. Interferometers, either making through-the-bearing measurements or measurements in the neighborhood of the sample being tested,[87] would be used for fine motion digital control.

   When a laser interferometer system is used as a through-the-bearing position feedback sensor, the platten must be polished to optical quality. It is envisioned that in the near future the resolution of single-crystal differential plane mirror interferometers will approach the 1 Å level.[88] A similar laser interferometer would also be used for axial position measurement. Precision capacitance probes, on the other hand, can have very high resolution over short ranges of motion and they do not require the surface of the carriage to be optically polished. In addition, the finite area of their measuring tips creates an averaging effect which reduces the effect of surface finish errors on the gap measurement. Capacitance probes can allow for servoing on the angstrom level[89]; however, long-term drift problems could render the probes inadequate for continuous operation of the system over a period of days or weeks.
3. *Sample measurement.* Existing technology developed for STMs could be used. Note, however, that a single probe would take many thousands of years to map with angstrom accuracy an entire 50 mm diameter specimen. Thus multiple probe techniques will probably be needed if such a large area is to be mapped.

Motion Control System

The resolution of the signal sent by the controller to the magnetic bearing affects the total force acting on the platten. In Equation 9.4.16, if $\eta = 0.9$, $\delta = 0.1$ Å, and $N = 12$, then the maximum servo update time is 204 μs. Using 32-bit digital signal processors (DSPs) for control makes this loop time achievable. In the future, in order to achieve even higher resolutions, superconducting coils powered by a ultrahigh-accuracy microwave-superconducting-Josephson junction power supply might be investigated for use. This type of power supply has been shown by NIST researchers to produce voltages with part-per-billion stability. Using available superconductors, which operate at liquid nitrogen temperatures, would require the bearings to be wrapped in an insulating blanket, liquid nitrogen passed around the coils, and an electric heater used to balance the heat flow into the structure. This is probably not a feasible option.

Temperature Control

Controlling system temperature without introducing large gradients could be achieved by using constant or low-power devices, operating the system at a steady thermal state, and configuring the system as a sphere hanging inside another temperature-controlled evacuated sphere. The two spheres could be radiantly coupled with the outer sphere cooled by a high-velocity coolant source.[90]

Temperature control by radiant coupling can be effective and is not as prone to the formation of gradients as convective cooling is. Consider the heat transferred between two bodies by radiation:

$$Q_{net} = F_{1-2}\sigma A_1 (T_1^4 - T_2^4) \tag{9.4.17}$$

[87] NIST personnel are developing a five-axis (X, Y, yaw, pitch, roll) laser interferometer that can be located around the probe used to measure the sample.

[88] Discussions with Carl Zanoni, Zygo Corp., Middlefield, CT.

[89] J. R. Leteurte, "Capacitance Based Sensing and Servo Control to Angstrom Resolution," Proc. 1987 Precis. Eng. Conf., Cranfield, England.

[90] The inner sphere cannot be cooled directly by a high-velocity source because the high velocities needed to eliminate gradients also generate turbulence and vibration. The dual-sphere concept was proposed by NIST researchers.

Material	ρ (kg/m^3)	E (GPa)	$\alpha_{expansion}$ (μm/m/C^o)	K (W/m/C^o)	C_p (J/kg/C^o)	t (s)	ΔT (C^o x 10^{-6})	α (m^2/s x 10^{-6})
Aluminum	2707	69	22.0	231	900	133	1.8	94.8
Beryllium	1848	275	11.6	190	1,886	361	3.4	54.5
Copper	8954	115	17.0	398	384	243	2.3	116
Grey cast iron	7200	80	11.8	52	420	308	3.4	17.2
Invar	8000	150	0.9	11	515	5736	44.4	2.7
Lead	11373	14	26.5	35	130	67	1.5	23.7
Zerodur	2550	90	0.15	6	821	22,960	267	2.9

Table 9.4.1 Properties of various materials and relative time for temperature change in a 1m diameter inner sphere caused by disturbance of 0.001 C^o in the outer temperature control envelope to cause 0.1 Å thermal expansion in a 0.25 m segment.

If 10 W of waste heat is to be removed from 1.0 or 0.5 m diameter spheres, and the temperature is to be maintained at 20°C (293°K), then Equation 9.4.17 can be used to determine that spheres surrounding the 1.0 or 0.5 m inner spheres must be kept at 19.440407°C and 17.742068°C respectively.[91]

Assuming steady-state conditions, small deviations δT (less than 0.1 C^o) in the temperature of the outer sphere will affect the amount of power retained by the inner spheres by

$$\Delta Q_{0.5m,\ T=17.742068} = 4.378\delta T_{outer\ sphere}(W/C^\circ) \tag{9.4.18a}$$

$$\Delta Q_{1.0m,\ T=19.440407} = 17.820\delta T_{outer\ sphere}(W/C^\circ) \tag{9.4.18b}$$

If deviations in the outer sphere temperature occur with a time constant much faster than that of the inner sphere, then the thermal mass of the inner sphere should prevent it from being affected by these variations.

In order to make a first-order evaluation of thermal suitability of various materials for the ÅRMM structure, consider the time t it takes the inner sphere to reach a new equilibrium temperature $T_i + \delta T_i$ given a step change δT_o in the outer-sphere temperature. Assume that the new equilibrium temperature is T + δT_i, where δT_i is the change in temperature that causes 0.1 Å thermal growth in a 0.25 m segment of the structure. Given a change in the outer-sphere temperature of δT_o (e.g., 0.001 C^o) then to the first order[92] the time it takes the inner sphere to change its temperature by δT_i via radiant coupling is

$$t = \frac{\rho c_p R \delta T_i}{3F_{1-2}\sigma\left[T_i^4 - T_o^4 - (T_i + \delta T_i)^4 + (T_o + \delta T_o)^4\right]} \tag{9.4.19}$$

Table 9.4.1 shows the value of t for various candidate materials Based on the time evaluation, Zerodur or Invar should be used. If the thermal diffusivity is considered, a material such as copper might be used in order to minimize gradients.[93]

By tuning power dissipation and the equivalent blackbody view factor with the size of the sphere, it may be possible to utilize an inexpensive accurate temperature control process (i.e., a phase-change process) for controlling the inner sphere's temperature. For example, assume that the outer sphere is contained in an ice water bath and the temperature of the outer sphere is fine tuned with an electric heater. If the inner sphere still needs to dissipate 10 W of power, the diameter of the inner sphere should be about 0.1759 m if the spheres still behave like black bodies. Unfortunately, this may not leave enough room for ARMM. This preliminary thermal analysis gives a good indication of where to start the design process, although it does not include transient affects, nor does it model hot spots within the sphere. Hot spots could probably be prevented if the structure were suitably instrumented and zone temperature control is used.

Summary

The tools for the development of the ARMM currently exist, all that is needed is commercial interest in order to make it an everyday reality. With careful design, resolution would be increased

[91] The view factor is 1, and σ is Boltzmann's constant, σ = 5.6697 x 10^{-8} W/m^2-K^4.

[92] This assumes that thermal conductivity of the body is infinite.

[93] Zerodur®, Invar, copper, aluminum, beryllium and cast iron are all known to have very stable forms. See, for example, J. Berthold et al., "Dimensional Stability of Fused Silica, Invar, and Several Ultra-low Thermal Expansion Materials," Metrologia Vol. 13, 1977, pp. 9–16.

not by requiring a change in the mechanical design (expensive), but by advances in sensor, control system, and control algorithm designs. Perhaps with the introduction of "warm" (20-30^{o}C) superconductor technologies, new sensors and actuators will become available that will push resolution limits even further.

Chapter 10
Power Generation and Transmission

The yearning of man's brain for new knowledge and experience and for pleasanter and more comfortable surroundings never can be completely met. It is an appetite which cannot be appeased.

Thomas Edison

10.1 INTRODUCTION

When designing a servo-controlled machine, one must choose power sources, power conversion devices, and coupling mechanisms while considering how they interact dynamically and statically. In this chapter the following are discussed:

- Dynamic matching of elements
- Linear and rotary electric motors
- Electromagnetic actuators
- Piezoelectric actuators
- Hydraulic and pneumatic actuators
- Rotary power transmission components
- Linear power transmission components

It is hoped that combined with continued real-world exposure, the material presented here will give the machine design engineer a feel for how to size components initially so that the engineer can complete the design of the rest of the machine. The machine design engineer can then interact with dynamics and controls personnel to develop detailed dynamic models and make the final selection of dynamic system components.

10.2 DYNAMIC MATCHING OF COMPONENTS

All the components of a machine must be in proper proportion, both in relation to their physical size and the capabilities of the servocontroller and power systems. If a component is oversized, it may increase the cost of the machine while performance may not be increased. If a component is too small, the rest of the machine's components may never reach their potential and machine performance will suffer. Note that size is a function of static and dynamic qualifiers. Components that behave well statically do not necessarily have good dynamic performance. This was illustrated in Section 7.4 for the case of the design of the machine's structure.

Machine elements are usually first sized assuming a static configuration (e.g., what is the bearing deflection under an applied load?), where the loads applied are the highest of those encountered in static or dynamic cases. The dominant static design factor is stiffness and the dominant dynamic design factor is natural frequency and damping. In order to size the components of a precision servo system, the following factors need to be considered:

1. The static axial (angular) mechanical stiffness of the system should be high enough such that the smallest force (torque) input into the system causes a deflection less than a maximum allowable limit. If the deflection were greater than this limit, then the closed-loop servo system may not be able to compensate for it. The smallest force (torque) increment will be a function of the resolution of the control system's output to the power source and the relative servo loop and mechanical time constants. For a point-to-point measuring machine, the axes need mainly to be positioned quickly and the probe accommodates overshoot; hence the actuator stiffness is not always of principal concern and the actuator may sometimes be sized based on a force criterion.
2. In the period during calculation of the next output value to the power system, the system is essentially running open loop. Assuming that small fast motions are dominated by the dynamics of the inertia of the system, the servo-loop time must be chosen such that when the minimum force (torque) acts on the system inertia for this open-loop period, the inertia does not move further than a maximum allowable limit. As a conservative estimate, one can assume that the system has no damping when making an estimate of this type of error.

3. To maximize the efficiency of the system and thereby decrease thermal errors, the transmission ratio, the inertias of the drivetrain elements and of the load, and the applied external forces must be in proper proportion.

It is possible to arrive at a series of equations that interrelate all the system variables. Section 10.2.1 discusses the minimum stiffness and servo-loop time issues. Section 10.2.2 discusses selection of the optimal transmission ratio.

10.2.1 Minimum Required Actuator Stiffness

The larger the mass that an actuator has to move, the less the actuator will feel high-frequency external force disturbances because the mass acts like a low-pass filter. The higher the stiffness of the actuator, the faster it will be able to move the mass, which means increased productivity and increased ability to resist external forces. However, the actuator must have sufficient stiffness to prevent greater than desired deflections from occurring when the servo system is not able to compensate for them (e.g., at high frequencies).

Sizing an actuator requires attention to stiffness and power requirements, and both depend on the mass of the system. Starting at the terminal axis that moves the tool tip, from the required acceleration and speed profile curves the design engineer should estimate the power required to move the axes. In many cases, frictional and cutting forces far outweigh inertial forces. It is a good idea to overestimate (e.g., by as much as 25%) the amount of power required, as there is nothing worse than a machine which does not have enough power. If the machine has too much power, then chances are the motors will run cooler and thermal errors will be less. In the preliminary design calculations, as the design engineer works back through the machine, about 0.3 N/W (50 lbf/hp) should be added to the weight of each axis to account for the weight of the powertrain components.

In order to determine the required stiffness of the actuator, first an estimate of the time constant (in seconds/cycle) of the system must be made:

$$\tau_{\text{mech}} = 2\pi\sqrt{\frac{M}{K}} \tag{10.2.1}$$

where M is the total system mass[1] and K is the actuator stiffness. The control system loop time τ_{loop} must be at least twice as fast, to avoid aliasing. Faster servo times create an averaging effect, thereby effectively increasing the resolution of the signal sent to the electric motor. Like taking the average of a number of data points with random error components, it is assumed here that a fast servo-loop time helps to increase the force resolution approximately by the factor $(\tau_{\text{mechanical}}/2\tau_{\text{loop}})^{1/2}$. For a controller with N bits of digital-to-analog resolution,[2] the incremental force input is thus assumed to be

$$\Delta F = \frac{F_{\text{max}}}{2^N\sqrt{\frac{\tau_{\text{mech}}}{2\tau_{\text{servo}}}}} \tag{10.2.2}$$

The deflection δ_K that this incremental force input causes becomes part of the servo system error that must be included in the error budget. The minimum axial stiffness[3] for the actuator is thus

$$K \geq \left\{ \frac{F_{\text{max}}\,\tau_{\text{servo}}^{1/2}}{2^N\,\pi^{1/2}\,M^{1/4}\,\delta_K} \right\} \tag{10.2.3}$$

For rotary systems, the force is replaced by torque and the mass by inertia. But how does one determine what the servo-loop time is?

While the controller is calculating the next value to send to the digital-to-analog converter (DAC), which provides the control signal to the power electronics, the power signal remains equal

[1] As discussed in the next section, the system mass is that of the carriage and the reflected inertia of the actuator. For optimal power transmission, the carriage mass and the actuator equivalent mass should be about equal so $M = 2m_{\text{carriage}}$.

[2] One cannot arbitrarily set N very high, as noise limits the attainable resolution. For example, although one may specify the use of a 16 bit ADC, one may actually only get 14 bits if one is lucky. For high-power systems, resolutions of 1 part per 1000 (10 bits) are considered good.

[3] Recall Section 5.2.2, where it was shown that a leadscrew's equivalent axial torsional stiffness is generally much greater than the axial stiffness, and thus can be ignored.

to the last value in the DAC. Hence regardless of the ideal input the system requires to minimize the position error, it is receiving an old signal and is therefore running open loop. It is thus important to make sure that during this time, τ_{servo}, the application of the incremental force only causes the system to move an amount which is insignificant. The most conservative assumption is that there is no damping in the system, so the error δ_M due to the mass being accelerated by the force resolution of the system for a time increment τ_{servo} is

$$\delta_M = \frac{1}{2}\left(\frac{\Delta F}{M}\right)\tau_{servo}^2 \tag{10.2.4}$$

The maximum allowable servo-loop time is thus

$$\tau_{servo} = \sqrt{\frac{2\delta_M M}{\Delta F}} \tag{10.2.5}$$

Combining Equations 10.2.2, 10.2.3, and 10.2.5 yields an expression for the required actuator stiffness to ensure that the axis will be controllable to the resolution desired (not accounting for externally applied forces or nonlinear effects):

$$K \geq \frac{F_{max}\,\delta_M^{1/4}}{2^{N-1/4}\,\pi^{1/2}\,\delta_K^{5/4}} \tag{10.2.6}$$

It is important to note that one must also determine the stiffness required to ensure that the error caused by externally applied forces will be below a desired maximum value.

Note that the total servo error δ_{servo} has been separated into its actuator deflection and mass acceleration motion error terms δ_K and δ_M respectively. This is due to the fact that the servo system cannot always be expected to respond fast enough to eliminate the actuator error term, so it is included separately. Typically, one would set $\delta_K = \delta_M = 1/2\delta_{servo}$. Also note that the maximum force cannot be set arbitrarily low or else the system will not have enough force to dynamically control the system.

When using this value of the actuator stiffness to design the drive train, one must be careful to properly add all the stiffnesses of all the drivetrain components. For example, take note that if the leadscrew's support bearings are axially much stiffer than the leadscrew, then the axial stiffness of the system will be dominated by the leadscrew. This allows the leadscrew diameter and system inertia to be minimized. One must also consider the stiffness of the structure between the actuator and the tool tip. Ultimately, before any hardware is built, a detailed numerical dynamic model of the system should be developed and tested by an experienced dynamic systems analyst.

These equations can also be used to find the maximum allowable servo-loop time:

$$\tau_{servo} \leq \sqrt{\frac{\pi^{1/2}\,2^{(4N+3)/4}\,\delta_M^{3/4}\,M\,\delta_K^{1/4}}{F_{max}}} \tag{10.2.7}$$

Figure 10.2.1 shows a plot of required stiffness and servo-loop time as a function of servo error for a typical high-accuracy machine. This minimum servo-loop time assumes that there is no phase lag in the controller. For a digital control algorithm that saves L past values for use in a recursive algorithm, the actual servo-loop time should be chosen to be τ_{servo}/L or faster. L is typically 2 for a PID algorithm, but may be as high as 10 if a nice filter is used to eliminate noise from the differentiation. For controllers that have an analog velocity loop, L may be 1. Note that these equations serve as initial design layout tools only. It is vital that before design details are made, a detailed dynamic model of the system be built.

The design engineer must also consider the natural frequency of each individual component and make sure that it will not be excited by its own motion or the motion of other components. There have been too many horror stories of simple items such as sensor brackets that were vibrating uncontrollably because their fundamental frequency was equal to that of a fan motor in the machine.

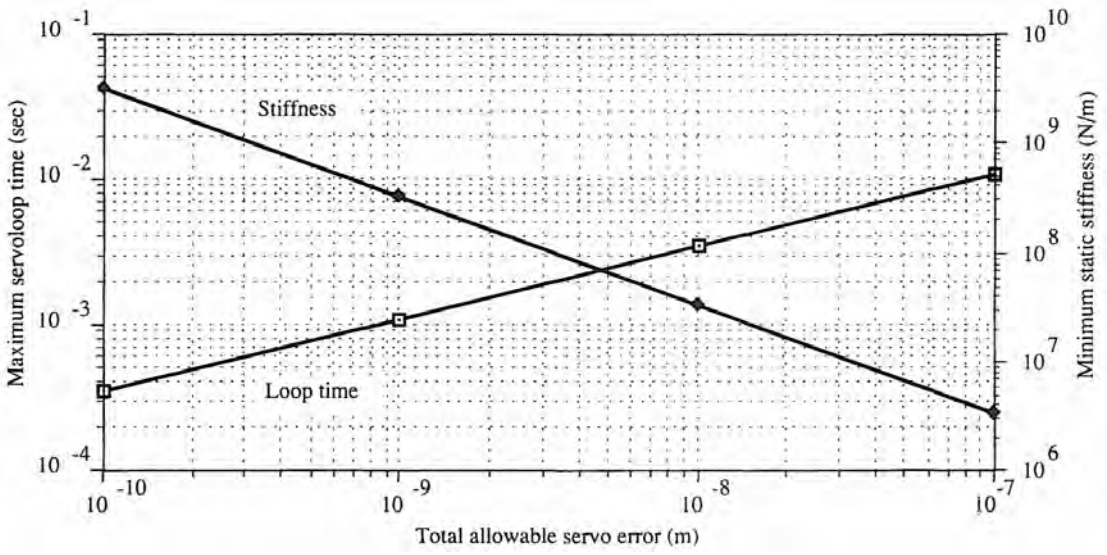


Figure 10.2.1 Required static stiffness and servo-loop time for a diamond turning machine with a 1000 N maximum axial force, 200 kg system mass, and 12 bit DAC.

10.2.2 Transmission Ratio Selection[4]

In this section, a procedure is presented for selecting a servomotor and transmission ratio to drive a specified load. In distilling this procedure from the available literature, the following assumptions were made:

1. Only permanent magnet brushed and brushless commutator-type motors are considered. These are by far the most common types of motors used in performance servo applications.
2. For intermittent applications, the motor's peak torque rating can be used in the calculations. For continuous duty, the motor's continuous torque rating should be used. In either case, in order to determine the motor's steady-state temperature, the manufacturer of the motor should be consulted.
3. It is assumed that the user wishes to choose the smallest, least expensive motor that will perform the operation satisfactorily. It is difficult to give an absolute cost-per-watt figure for a given class of motors, but since it is generally true that the cost of servomotor/amplifier packages rises with power, it will be assumed that motor power can be used as an indicator of approximate system cost. Thus, the following method is implicitly trying to minimize the motor power.
4. If inertia forces dominate, a simplified selection procedure can be used. If external forces dominate, their time history must be used to choose a transmission ratio to minimize motor power.

Based on these assumptions and the specifics of the particular application, the optimal system transmission ratio and the desired output power and power rate of a dc motor can be specified. In general, the accuracy of the final result is expected to be quite good. However, it is generally good practice to multiply the final motor power by a reasonable factor of safety (e.g., 25-50%) to allow for unmodeled effects. If the difference between the selected motor inertia and the load inertia is greater than several orders of magnitude, the optimal transmission ratio becomes unreasonably large to implement. In such cases, it is recommended that a more practical coupling ratio and a larger motor be used.

[4] The material for the inertia matching part of this section was obtained in large part from a paper written as part of an independent studies project by Bruce M. Schena while he was one of the author's graduate students at MIT.

Motor and Load Power Rate for Motor Selection

Before a motor and coupling ratio can be selected, a few parameters of the system must be evaluated. First, regardless of the type of drive system being used, be it a leadscrew, friction drive, belt and pulley, or gear train, the mass (kg) or inertia (kg-m^2) of the load and drive train components must be known. Second, the design engineer must have an estimate for the approximate velocity profile as a function of time. In some instances, this profile can be determined directly by the application, such as in the case where a parameter, such as velocity or maximum acceleration, is set by an outside constraint such as a customer specification. In other cases, the exact velocity profile may be of secondary importance, while the primary goal is obtaining a specified move time. For the purpose of this analysis, it will be assumed that the design engineer is attempting to do the latter, where it is desired to move a specified load a specified distance in a given time.

The optimum velocity profile (minimum power) for moving a load between two points in a given time is parabolic in shape; however, because it is difficult from a control standpoint to achieve such a profile, it is rarely used. Two common alternatives to the parabolic profile are the triangular and trapezoidal profiles. Both of these are relatively easy to implement. It has been shown by Tal[5] that if the parabolic profile is assumed to represent 100% efficiency in a particular application, a trapezoidal profile with equal acceleration, slew (constant speed), and deceleration times is 89% efficient, and a triangular profile is 75% efficient. The triangular profile, although it requires more power per move, is the fastest way of moving a load a given distance.

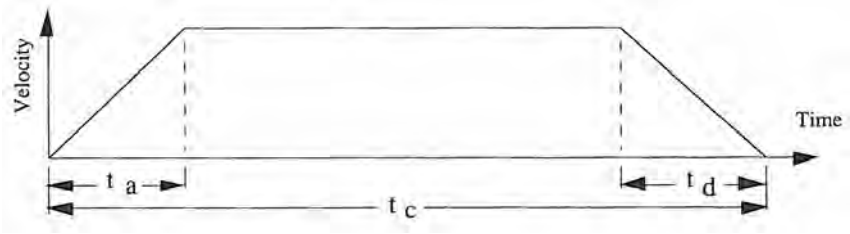


Figure 10.2.2 Trapezoidal velocity profile.

A trapezoidal velocity profile is shown in Figure 10.2.2. Given the distance D that the load is required to move and the total time t_c required, the trapezoidal profile is represented by

$$D = \frac{a_{max}}{2}\left(\frac{t_c}{3}\right)^2 + \left(\frac{a_m t_c}{3}\right)\left(\frac{2t_c}{3}\right) + \frac{a_{max}}{2}\left(\frac{t_c}{3}\right)^2 \tag{10.2.8}$$

The maximum acceleration and velocity during the move are

$$a_{max} = \frac{9D}{2t_c^2}, \quad \text{maximum acceleration (m/s}^2 \text{ or rad/s}^2\text{)} \tag{10.2.9}$$

$$v_{max} = \frac{3D}{2t_c^2}, \quad \text{maximum velocity (m/s or rad/s)} \tag{10.2.10}$$

The next step is to determine the *power rates* for the motor and load. Unlike torque or velocity, power rate is unfortunately a nonintuitive concept. The unit of power rate is W/s and the power rate is basically a measure of the electrical-to-mechanical power conversion (transduction) efficiency of a given actuator. A motor can be thought of as a black box in which electrical power is turned into mechanical power. In an ideal world, this transduction would be instantaneous. In the real world, the motor resistance and inductance combine to create a dynamic system which has a time constant of τ_m. The power rate combines both the power of the actuator and the mechanical time constant into a single figure of merit. Although an intuitive feel for power rate is difficult to develop, its actual value can be calculated with knowledge of the system inertia and information from motor data sheets. For any motor used to accelerate a mass for a period t_a, the power rate is defined to be

$$P_{R\ motor} = \frac{\Gamma_{motor}^2}{J} = \Gamma_{motor}\,\alpha = \frac{I^2 R}{t_a} \tag{10.2.11}$$

[5] J. Tal, "The Optimal Design of Incremental Motion Control Systems," Proc. 14th Symp. Increment. Motion Control Syst. and Dev., May 1985, p. 4.

A relative feel for this equation can be obtained if one realizes that Γ^2_{motor}/J is a measure of how fast the motor can accelerate its own inertia and I^2R is the power into the motor. Thus the left-hand side of the equation is a measure of the torque into the system and of the acceleration of the core and all components directly connected to it. The right side of the equation is a measure of the electrical power into the system and the actual speed of response.

The maximum motor power during the move in the case of the trapezoidal profile is just the power rate multiplied by the acceleration time:

$$P_{motor} = PR\frac{t_c}{3} \tag{10.2.12}$$

The motor selection process can thus be thought of as an attempt to determine the required power rate P_R and motor power P_{motor} and make sure that they are matched to the rest of the system. These values can then be used to evaluate candidate motors.

In order to determine the required motor power rate, it is first necessary to determine the power required to move the load efficiently. Since the load inertia and required acceleration are quantities specified by the machine configuration, the load's power rate can be calculated for a linear motion system by

$$P_{R\ load} = (M_{load}\, a + F_f)\, a \tag{10.2.13a}$$

and for a rotational system by

$$P_{R\ load} = (J_{load}\, \alpha + \Gamma_f)\, \alpha \tag{10.2.13b}$$

The load inertias are of all the components on the load side of the transmission (e.g., they would not include the inertia of the leadscrew). The factors F_f and Γ_f are the constant friction force or friction torque opposing the motion of the load. The maximum static and dynamic coefficients of friction under the maximum load and speed conditions should be used. The friction effects are often of considerable, and surprising, magnitude, so they must not be ignored. Where friction is not explicitly known, but drivetrain efficiency is known (e.g., the ratio of torque to speed multipliers for a gearbox), one can adjust the power rate required by the load accordingly. For example, if the efficiency of a ballscrew is 90%, then the power rate for the carriage driven by the ball screw should be multiplied by 1/0.9 = 1.11. Similarly, the cutting force can be significant, if not greater than the force required to accelerate the load, as discussed in the next section. Hence the *load power rate* for cutting $P_R = F^2_{cut}/M$ or Γ^2_{cut}/J should be used if it is greater than the power rate given by Equation 10.2.13. One must also make sure that the torque-speed curve for the motor is adequate.

From Equation 10.2.12 the load power is simply

$$P_{load} = P_{R\ load}\, t_a \tag{10.2.14}$$

For systems dominated by inertial loads, it can be shown that during acceleration and deceleration times, maximum power is transferred from the motor to the load under what is called the *matched inertia doctrine.*[6] When the matched inertia condition is met, it has been shown by Newton[7] that the power expended in causing motion of the motor's core and directly attached drivetrain components is the same as that expended in causing motion of the load. Thus, the motor must have a total power rating of at least twice the load power requirement (assuming no losses) in order to move itself and the load:

$$P_{motor} \geq 2P_{load} \tag{10.2.15}$$

Similarly, it has also been shown by Newton[8] that when inertias are matched, the required motor power rate must be at least four times the load power rate:

$$P_{R\ motor} \geq 4P_{R\ load} \tag{10.2.16}$$

With Equations 10.2.15 and 10.2.16, one can search manufacturers' catalogs for a motor which has the correct PR_{motor} and P_{motor} characteristics.

---

[6] J. Tal, "Optimal Design Motion Control Systems," Motion, July/Aug. 1986, p. 20.

[7] G. Newton, "Selecting the Optimum Electric Servo-motor for Incremental Positioning Applications," 10th Symp., Increment. Motion Control Syst. and Dev., B. C. Kuo (ed.), p. 5.

[8] Ibid.

Optimal Transmission Ratio for a System with Small External Loads

For the case where the external loads are low (i.e., frictional and cutting forces), the matched inertia doctrine can be used to find the "optimal" transmission ratio. First assume that all the rotational power leaving the motor arrives at the load, and power equals the product of torque and angular speed:

$$\Gamma_{\text{motor}}\,\omega_{\text{motor}} = \Gamma_{\text{load}}\,\omega_{\text{load}} \tag{10.2.17}$$

But torque also equals the product of inertia and angular acceleration; thus

$$J_{\text{motor}}\,\alpha_{\text{motor}}\omega_{\text{motor}} = J_{\text{load}}\,\alpha_{\text{load}}\,\omega_{\text{load}} \tag{10.2.18}$$

Substituting for the angular velocity in terms of constant acceleration and time gives

$$J_{\text{motor}}\,\alpha^2_{\text{motor}} = J_{\text{load}}\,\alpha^2_{\text{load}} \tag{10.2.19}$$

The transmission ratio n relates the motor and load velocities by $\omega_{\text{motor}} = n\omega_{\text{load}}$, so differentiating this relation once with respect to time, and substituting into Equation 10.2.19, one finds that the accelerations are related by the same coupling ratio:

$$J_{\text{motor}}\,n^2\,\alpha^2_{\text{Load}} = J_{\text{Load}}\,\alpha^2_{\text{Load}} \tag{10.2.20}$$

The "optimal" transmission ratio for a pure rotational motion system dominated by inertial loads is thus

$$n_{\text{opt}} = \sqrt{\frac{J_{\text{load}}}{J_{\text{motor}}}} \tag{10.2.21}$$

A similar method of analysis can be used to show that for a friction (capstan) or belt drive system the optimal drive wheel radius r in meters is

$$r_{\text{roller}} = \sqrt{\frac{J_{\text{motor}}}{M_{\text{load}}}} \tag{10.2.22}$$

For a leadscrew-driven carriage, the lead in mm/rev is

$$\ell = 2\pi \times 10^3 \sqrt{\frac{J_{\text{motor}}}{M_{\text{load}}}} \tag{10.2.23}$$

In each of these cases, the load inertia is everything on one side of the transmission, and the motor inertia is everything on the motor side of the transmission (e.g., the motor inertias and the gear or leadscrew inertia that is directly attached to the motor). After selecting the transmission ratio, one must estimate the efficiency of the drivetrain and see if the initial estimate used to find the load power rate was correct. The iterative nature of motor size and transmission ratio calculations should now be apparent. The required peak torque of the motor will be within the motor's capabilities if the motor's power rate is greater than or equal to the required value. However, one should always double check with the motor manufacturer to make sure that this is true.

It is also necessary to check the required velocity and torque of the motor. For rotational systems the motor speed in rpm is

$$\omega_{\text{motor}} = n_{\text{opt}}\,\omega_{\text{load}} \tag{10.2.24}$$

For linear friction or belt drives with a carriage velocity in m/s, the motor speed in rpm is

$$\omega_{\text{motor}} = \frac{30V_{\text{load}}}{\pi r_{\text{roller}}} \tag{10.2.25}$$

For leadscrews with lead l (mm/rev) and a carriage velocity in m/s, the motor speed in rpm is

$$\omega_{\text{motor}} = \frac{6 \times 10^4\,V_{\text{load}}}{\ell} \tag{10.2.26}$$

When the load inertia is very large, one often finds the transmission ratio becoming very large, which in turn can require that the motor have a very high speed. If the required peak velocity of the motor is greater than the actual capability of the motor/amplifier system, then the transmission

ratio n would have to be reduced (r or l increased) and thus the required motor power rate would have to be increased. It is this limiting speed factor which makes inertia matching methods often impractical for choosing motors for large massive systems; however, as shown below, some large massive systems often have large external (i.e., frictional and cutting) loads that increase the required ratio in order to minimize motor heat generation.

Optimal Transmission Ratio for a System with Large External Loads[9]

The inertia matching doctrine is concerned with motor power during acceleration and deceleration, but what about long periods of constant-velocity motion where the system may encounter large frictional loads and cutting forces? The thermal power generated by the motor, with winding resistance R_{motor} and motor constant K_t (N-m/A), during a trapezoidal move defined by Figure 10.2.2, and with a constant applied external load (i.e., frictional and cutting forces) is

$$W_{thermal} = \frac{cR_{motor}\,\omega_{load}^2\,J_{load}^2}{K_t^2\,t_c}\left[n^2\left(\frac{J_{motor}}{J_{load}}+\frac{1}{n^2}\right)^2+\frac{r}{n^2}\right] \tag{10.2.27}$$

where

$$c = \frac{t_c}{t_a}+\frac{t_c}{t_d} \tag{10.2.28a}$$

$$r = \frac{\Gamma_{load}^2\,t_c^2}{c\omega_{load}^2\,J_{load}^2} \tag{10.2.28b}$$

Remember, Γ_{load} is the external load due to friction and cutting forces, and it does not include loads associated with acceleration of the load's inertia. The thermal power can be minimized with respect to the transmission ratio n by $\partial W_{thermal}/\partial(n^2) = 0$, hence giving the optimal transmission ratio:

$$n_{opt} = \sqrt{\frac{J_{load}\sqrt{1+r}}{J_{motor}}} \tag{10.2.29}$$

In these equations, if the load is moving linearly, then J_{load} is the mass (kg), Γ_{load} is the force (Newtons), and ω_{load} is the maximum load speed (m/s) at the maximum force level.

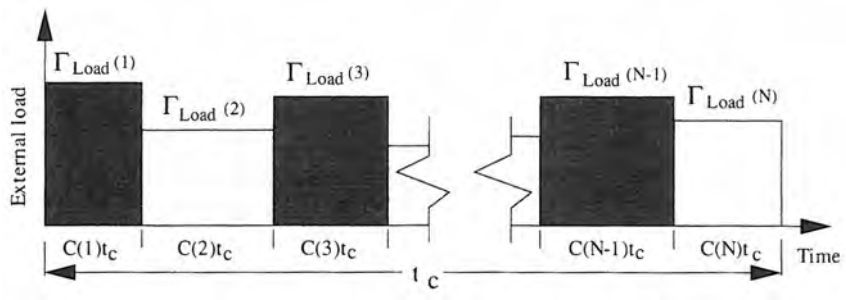


Figure 10.2.3 External load profile during a move.

In many cases, the load torque requirements during the move will vary greatly. If the load application is divided up into N segments each of length $C(i)t_c$, as shown in Figure 10.2.3, then an overall equivalent load can be defined:

$$\Gamma_{Lequ} = \sqrt{\sum_{i=1}^{N} r_{load}^2\,(i)C\,(i)} \tag{10.2.30}$$

where the sum of all C(i) must equal 1.0. The optimum transmission ratio is then determined using $\Gamma_{load} = \Gamma_{Lequ}$ in Equation 10.2.28b. Note that if the external loads are zero, then the optimal ratio is equal to that given by the inertia-matching criterion in Equation 10.2.21.

[9] For a detailed analysis and the original paper that is summarized here, see J. Park and S. Kim, "Optimum speed reduction ratio for d.c. servo drive systems," Int. J. Mach. Tools Manuf., Vol. 29, No. 2, 1989. Also see J. Park and S. Kim, "Computer aided Optimum motor selection for D.C. servo drive systems," Int. J. Mach. Tools Manuf., Vol. 30, No. 2, pp. 227–236, 1990.

During a normal operating period of several hours or days, one must consider all of M various load cycles encountered, each with its own Γ_{Lequ}. In this case, the equivalent external loads $\Gamma_{Lequ}(i)$ are found for each trapezoidal move $t_c(i)$ using Equation 10.2.30, and the total thermal power dissipated by the motor is

$$W_{thermal} = \sum_{i=1}^{M} k(i) \left[n^2 \left(\frac{J_{motor}}{J_{load}} + \frac{1}{n^2} \right)^2 + \frac{r(i)}{n^2} \right] \tag{10.2.31}$$

where

$$k(i) = \frac{c(i) R_{motor}\, \omega_{load}^2(i) J_{load}^2}{K_t^2\, t_c\,(i)} \tag{10.2.32a}$$

$$r(i) = \frac{\Gamma_{Lequ}^2(i) t_c^2(i)}{c(i) \omega_{load}^2(i) J_{load}^2} \tag{10.2.32b}$$

$$c(i) = \frac{t_c(i)}{t_a(i)} + \frac{t_c(i)}{t_d(i)} \tag{10.2.32c}$$

The optimal reduction ratio n for the system is obtained from $\partial W_{thermal}/\partial(n^2) = 0$:

$$n_{opt} \sqrt{\frac{J_{load}\sqrt{1+R}}{J_{motor}}} \tag{10.2.33}$$

where

$$R = \frac{\sum_{i=1}^{M} s(i) r(i)}{\sum_{i=1}^{M} s(i)} \tag{10.2.34a}$$

$$s(i) = \frac{c(i) \omega_{loadmax}^2(i)}{t_c(i)} \tag{10.2.34b}$$

Equation 10.2.34b represents a portion of the ith trapezoidal segment's contribution to the total performance requirement.

In order to minimize peak cooling requirements for a machine, it may be desirable to minimize the thermal output during any one move; hence the "optimal" reduction ratio should be found for each particular move using Equation 10.2.29 and then the heat dissipation calculated using Equation 10.2.27. Note that all of this type of analysis is quite amenable to spreadsheet implementation.

For a capstan (friction) or belt-driven system where J_{motor} has units of kg-m^2 and J_{load} is linear so it has units of kg, n_{opt} has units of 1/m. The optimal drive wheel radius in meters is thus

$$r_{capstan} = \frac{1}{n_{opt}} \tag{10.2.35}$$

For a leadscrew-driven system, the optimal lead l in mm/rev is

$$\ell = \frac{2\pi \times 10^3}{n_{opt}} \tag{10.2.36}$$

In both cases, n_{opt} is obtained from Equation 10.2.29 or 10.2.33.

Optimal Transmission Ratio Determination Summary

For measuring machines the external loads are negligible and hence the simpler-to-implement inertia-matching doctrine yields the same results as the generalized formulas; hence Equations 10.2.21-10.2.23 can be used to determine the optimal transmission ratio. For systems with large external loads, Equation 10.2.33 should be used. In both cases, load and motor power rates need to be determined in order to select a motor. In addition, after selecting a transmission ratio and a motor, it must be determined if the system will be able to handle the maximum speeds and torques that will be imposed on it.

The selection procedure for determining the specifications for a dc servomotor system capable of moving a machine component with specified trapezoidal profiles and external load histories can be summarized as follows:

1. Calculate the inertia of the load (e.g., machine tool table).
2. Estimate the drivetrain component inertias from preliminary component selection made using the minimum stiffness criteria of Equation 10.2.6. Calculate the optimal transmission ratio using the inertia-dominated or external load-dominated formulas, as required.
3. Check the maximum motor and transmission speeds to see if they are acceptable. If unrealistic values for the transmission ratio or motor speed are found, then a larger motor with higher torque and inertia should be used. In some cases, one might just have to live with an inefficient motor situation. In general, the motor peak torque can be called upon for acceleration or deceleration periods, but the operating torque should not be exceeded while moving against external loads.
4. Assuming an ideal trapezoidal velocity profile, calculate the worst-case load power rate (Equations 10.2.13) from all the moves, and be sure to include friction and external loads.
5. Calculate the worst-case load power (Equation 10.2.14).
6. Multiply the load power by 2 and the load power rate by 4 to get the minimum motor power and power rate, respectively.
7. Find a motor which has the proper power and power rate.

Maximum power rate corresponds to minimum move time. This means that if one motor has a higher power rate than another, the higher power rate motor has the potential to move the load faster than the lower power rate motor. The tradeoff is usually in cost.

Despite the attempt at inclusion of real-world effects such as transmission element inertias and external loads in the theory presented in this section, it is still possible that the motor selected through this procedure will be underpowered for the application. Additional linear and nonlinear effects, such as bearing and brush friction, internal damping, motor winding resistance, and thermal limits, will all act to reduce the net power transmitted to the load. Thus, it is strongly advised that for critical applications, the final motor selection should be based on a detailed dynamic model and consultation with the manufacturer. If possible, a prototype system should be bench-tested.

10.3 Linear and Rotary Electric Servomotors [10]

A servomotor is a motor whose output can be proportionately controlled and can therefore be used in a closed-loop system. Linear and rotary electric servomotors are the most common types of servomotors used to control the motion of precision machines' axes. Electric motors' versatility is virtually unlimited and new designs are constantly emerging; hence an attempt is made here not to generalize beyond identification of a few trends. Although a mechanical design engineer does not have to be an expert in motor design and control,[11] because motor manufacturers are in general very helpful, the design engineer should always maintain a "show me" attitude and ask for a clear explanation of why the manufacturer recommends a particular type of motor.

Performance Qualifiers

When design engineers must be responsible for choosing the motor, power supply, and controller[12] for a precision machine's axes, they should choose a company that will work with them to answer questions about how the motor will perform on the machine. The design engineer may want to ask the motor manufacturer about the following issues pertaining to the performance of the motor system:

[10] A good trade journal with informative "how to" articles is Motion, the official journal of the Electronic Motion Control Association, Chicago IL, (312) 372-9800.

[11] See, for example, A. Fitzgerald et al. Electric Machinery, McGraw-Hill Book Co., New York, 1983; P. Ryff, Electrical Machines and Transformers, Prentice Hall, Englewood Cliffs, NJ, 1987; G. Dubey, Power Semiconductor Controlled Drives, Prentice Hall, Englewood Cliffs, NJ, 1989; W. Leonhard, Control of Electrical Drives, Springer-Verlag, New York, 1985; and S. Nasar and I. Bolden, Linear Electric Motors, Prentice Hall, Englewood Cliffs, NJ, 1987.

[12] Some motors require a power supply and circuitry (i.e. a *driver*) to switch the power to different windings, so one must consider the motor, power supply, and driver as an integral system.

Accuracy (linearity): Some people think that with a closed-loop system, they do not need to be too concerned with the linearity of the relation between input current and output torque (or force). However, any nonlinearity in this relation manifests itself as a varying gain on the system which can be detrimental to the stability and accuracy of the overall system. On the micron level this may not be of concern, but on the nanometer level it most certainly is. Users of stepper motors, which are run open loop, need to be particularly concerned with accuracy.

Controllability: The ideal question to be able to ask of the motor manufacturer is "in this application, will I be able to control the position of the system to within X microns?" A good motor manufacturer will ask you about the system specifications (mass, stiffness, friction, servocontroller type, etc.), do some modeling, and then call you back with a yes or no answer. The design engineer should then ask for examples of past and present applications. If a new type of motor has been recommended or if this is a new type of application, one must be prepared to add bench-test time to the project as well as seeking alternative possibilities. If the motor manufacturer cannot answer the question with a "yes" or "no", the designer should be wary and inquire elsewhere. For example, what is the time constant of the motor and its driver (controller)? Also, low total motor and power supply inductance is analogous to low mass, which is indicative of a responsive system; therefore, one should be wary of systems that add large inductors to the driver to help filter out noise. Furthermore, is the motor driver a device that has feedback loops in it to control position or speed, or is it just a power amplifier with phase commutation circuitry? One must be careful to make sure that analog feedback loops are properly buffered. If they are not, a small change in the variable resistor in one section can necessitate the readjustment of all the other variable resistors. In general, dc brushed motors with simple proportional power amplifiers will be the most controllable.

Resolution: Given an ideal feedback device and driver, what is the minimum increment of force or torque that the motor will produce? Are the motor and power electronics made to have minimal electrical noise in order to help increase resolution?

Repeatability: Given an ideal feedback device and driver, how repeatable is the motor's performance as a function of voltage level, time, speed, and temperature.

Service: If a manufacturer does not provide complete detailed easy-to-read documentation and friendly courteous help before the sale, drop them like a hot potato. Ask for references and call the references to check what their experience with the manufacturer was. Some manufacturers say things like "we will make the adjustments for you." However, if the adjustments are made wrong, or the motor manufacturer does not respond quickly, you will get blamed. You or someone who works with you must be able to understand how to install, adjust, and maintain the system.

Thermal Characteristics: It is vital that the motor manufacturer be able to tell you how the motor will perform thermally in the application you envision. If you provide the motor manufacturer with mounting details and probable duty cycles, they should be able to give you a projected history of motor operating temperature and thermal output. Remember, unless a cooling system is designed for the motor, its heat will go into your machine and cause it to expand. As the motor heats up, its resistance increases and the strength of its magnets decreases. Both these factors cause the motor to draw more current. Usually, an equilibrium point is reached, but in some cases the motor may burn up.[13]

Size Constants

Size constants are motor parameters which remain constant independent of winding changes, and they include:

Maximum Stall Torque (N-m): The maximum stall torque (peak rated torque) is defined as that which produces a specified winding temperature rise in a given time with the motor shaft locked and no external cooling or heat sinking.

T_c*, Continuous Stall (N-m)*: The stall torque is defined as that which results in a specified steady-state temperature rise. A motor can operate continuously at this stall torque in a specified maximum ambient temperature.

Maximum Continuous Output Power (watts): The maximum shaft output power that can be obtained from a specific motor without exceeding a specified steady-state temperature rise. To achieve this point, the voltage must be varied or the motor must be wound to operate at a specific speed torque point.

[13] See, for example, W. Fleisher "How to Select DC Motors," Mach. Des., Nov. 10, 1988.

K_M, *Motor Constant (N-m/W$^{1/2}$)*: Ratio of peak torque to the square root of power input at stall and a specified ambient temperature.

$$K_M = \frac{T_p}{\sqrt{P_p}} = \frac{K_T}{\sqrt{R_M}} \tag{10.3.1}$$

K_M indicates the ability of a motor to convert electrical power into torque.

TPR, Temperature Rise Per Watt (C^o/watt): Essentially the worst-case ratio of winding temperature rise to average power continuously dissipated from the armature.

F_o, *Damping Coefficient (N-m/rpm)*: Torque loss due to rotational losses, mostly eddy current, which is proportional to speed. This should be included as a friction effect when calculating the load power rate for inertia matching.

T_F, *Hysteresis Drag Torque (N-m)*: Magnetic friction due to hysteresis in armature laminations, T_F typically includes the cogging torque. T_F may be decreased with the use of high-efficiency laminations and/or large air gaps. However, there will be a resultant performance penalty with an increased air gap. Its effect should also be included as part of the load power rate.

Cogging Torque (N-m): Reluctance torque resulting from the alignment of the magnet (in a brushless motor) and lamination tooth edge. This is the torque you feel when you rotate a motor by hand. Cogging torque is one the primary cause of problems in precision systems:

- It is the major cause of control problems at low speed.
- It is minimized with the use of slanted windings.
- It is minimized with the use of an increased air gap (although efficiency is lost)
- It is minimized with the use of a sine-wave or mapped-wave controller triggered by the Hall effect sensors.
- It is minimized by maximizing the number of poles.
- It is minimized by making sure that the number of slots is not an integer multiple of the number of poles.

Number of Poles: The number of magnetic poles used in the design of the motor's permanent-magnet field.

Torque Ripple (%): The variation in torque produced as the powered motor turns and the orientation of the windings and poles changes.

Winding Constants

Winding constants are motor parameters which are winding dependent and will change if the motor is wound differently or operated at a different voltage level.

T_p, *Peak Torque (N-m)*: The nominal value of developed torque with the peak current I_p applied to the motor (I_p x K_T = T_p).

I_p, *Peak Current (A)*: The rated value of current produced when the design voltage is divided by the motor terminal resistance at a specified temperature (e.g., typically 25 ^{o}C).

K_T, *Torque Sensitivity (N-m/A)*: The ratio of developed torque to armature input current for the designated winding. This torque/current relationship prevails regardless of the armature speed. Thus, any value of current in a winding will develop a corresponding value of torque whether the motor is running or at a standstill. Normal tolerance is $\pm$10%.

No-Load Speed (rpm): The typical operating speed of the motor at a specified voltage with no external load applied. No-load speed is a function of voltage applied and is not necessarily the maximum speed of the motor. In the published data, consideration is often given to iron loss and drive electronics loss. Theoretical no-load speed (with no consideration given to losses) can be calculated by dividing the applied voltage by K_B. Theoretical no-load speed is that speed at which the generated back EMF is equal to the applied voltage.

K_B, *Voltage Constant (volts/krpm)*: Also known as back EMF constant, K_B is the ratio of voltage generated in the armature to the speed of the armature. Since both K_B and K_T are determined by the same factors, K_B is directly proportional to K_T. When torque is in N-m and K_B is in volts/krpm, the relationship is K_T x 0.00522 = K_B. Normal tolerance is $\pm$10%.

R_M, *Terminal Resistance (ohms)*: The resistance measured between any two motor terminals at 25 ^{o}C. The temperature coefficient of copper causes a rapid resistance increase with increases in temperature.

L_M, *Terminal Inductance (mH)*: The series equivalent of armature inductance as measured at the motor terminals. Normal tolerance is $\pm 30\%$.

Peak Efficiency: Each winding has one load point (speed, torque) which gives optimum performance for the specified voltage. Efficiency is the ratio of output to input power (W).

Maximum Continuous Output Power (W): The maximum shaft output power that can be obtained from a specified winding at the stated voltage without exceeding a specified steady-state temperature rise.

These definitions are used by many motor manufacturers. If a catalog does not provide the numbers you need, do not be shy about asking for information.

10.3.1 Servomotor Types

In order to provide power to a precision servo-controlled axis, there are essentially six basic types of servomotors that can be used: dc brushed motors, dc brushless motors, ac induction motors, synchronous reluctance motors, hysteresis motors, and stepper motors. Each type has its own advantages and disadvantages that makes it appropriate for different applications.[14] They are described briefly below in the context of rotary applications. Linear versions are essentially rotary motors that have been cut at one point on the circumference and then pressed flat.[15]

DC Brushed Servomotors

In a dc brushed servomotor[16] a stationary magnetic field is induced in the stator's coils with a dc current, and a rotating field is induced in the rotor's coils with a dc current transferred to the rotor windings via brushes making contact with a segmented slip ring on the rotor. The rotor and stator fields try to align themselves. As the rotor moves so that the magnetic fields come into alignment, the brushes move onto a conductor for a different set of windings, so the magnetic fields are no longer aligned and hence the rotor keeps turning. A similar phenomenon occurs in a brushed linear dc motor.

There are many different ways in which the stator and rotor windings can be configured, depending on the application. Conventional radial pole and axial pole (e.g., ServoDisc®[17]) motors are illustrated in Figure 10.3.1. Performance characteristics obtainable from these two types are contrasted in Figure 10.3.2. In an iron core motor, radial magnetic fields are used. In an axial pole motor, current flows radially in the disk orthogonal to an axial magnetic field, and the result is a force orthogonal to both, which creates a torque on the disk. Since there is no iron needed to focus the magnetic field generated by rotor windings, rotor inertia can be greatly reduced and the number of current carrying paths can be very large. As a result, torque-to-inertia ratios are much higher and cogging can be virtually eliminated; however, compared to a brushless dc servomotor, for a given torque rating, an axial pole motor may be three times the diameter and twice the length. For iron core designs when the field winding voltage is held constant and the armature (rotor winding) voltage is varied, a constant torque versus speed output is obtained. When the armature voltage is held constant and the field winding voltage varied, a constant power output is obtained.

Because there are only a discrete number of windings on the rotor of an iron core motor, there will be a sinusoidal variation in the torque output of a dc servomotor. One distinct advantage of a brushed motor over a brushless motor is that with the former it is easier to add more slots (windings) to reduce cogging torque than it is to add windings and switching circuitry to the latter. The more slots a motor has, the greater the potential for reducing cogging if they are properly positioned relative to each other and the magnet spacing. A technique which can be used with virtually any conventional motor is to slant the armature's teeth (i.e., the windings) in the rotor or stator. This makes the windings form a very gradual helix instead of running purely longitudinally. This adds to manufacturing cost and heat generated but can greatly decrease cogging torque. Cogging torque can also be decreased by increasing the air gap between the rotor and stator, but this also results in a loss of torque and efficiency. It is possible to model the effect of cogging torque on the performance of the system and compensate for it with a suitably robust control system. The cogging torque can

[14] A good general reference about motors and motor control systems is Machine Design's electrical/electronic annual reference issue.

[15] This is not how they are made, but it is a good visualization tool.

[16] These motors are usually just called *dc servomotors*.

[17] ServoDisc® is a trade name used by PMI Motion Technologies.

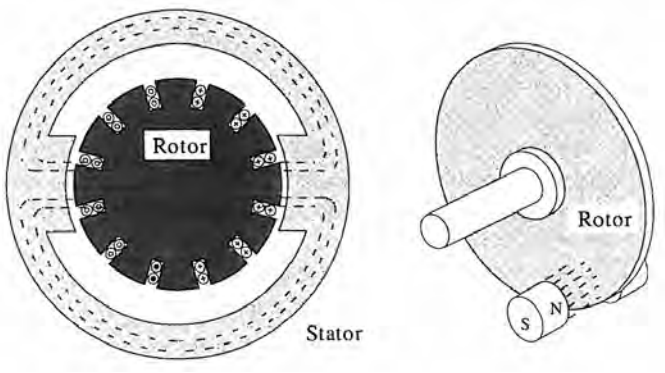


Figure 10.3.1 Radial pole iron core and axial pole (e.g., ServoDisc®) dc brushed motors. (Courtesy of PMI Motion Technologies.)

also be mapped as a function of rotor speed and position and then be accounted for in the control algorithm with a look-up table of torque correction factors; however, the dynamics of the system will limit the accuracy of this approach.

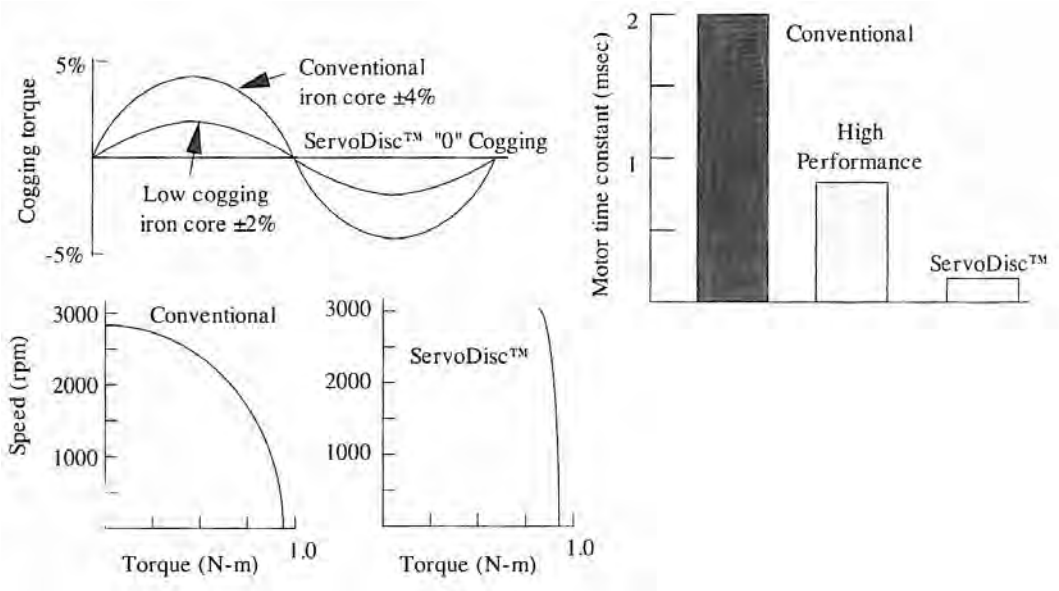


Figure 10.3.2 Comparison of typical radial pole iron core and Servo Disc® motor performances. (Courtesy of PMI Motion Technologies.)

One of the primary advantages of brushed dc servomotors is their simplicity of operation. All that is needed to run a dc servomotor is a power supply of the appropriate voltage and current capability, a suitable power op-amp, and the analog output from a servo-control algorithm. Another advantage of dc servomotors is their ability to dynamically brake themselves without requiring power input that would otherwise add to the heat generated in the system. Braking is achieved by disconnecting the armature from the power supply and then connecting the leads together. Without power applied and the rotor still turning, the motor is essentially a generator, and the current generated in the armature creates a magnetic field that opposes motor rotation. The degree of braking generated depends, among other parameters, on the resistance of the device used to short the two armature leads.

There are three main disadvantages of dc servomotors:

1. The brushes wear and give off small particles which can be detrimental to clean rooms, so one must be careful to seal dc servomotors for clean room applications. Also, regardless of speed, under high-torque conditions, very high current through the brushes can cause them to erode (burn out) very rapidly.
2. Since there are windings on the rotor, heat will be generated by the resistance of the windings; unfortunately, the object the rotor is connected to (e.g., a ballscrew) serves as one of the principal heat conduction paths out of the rotor. At low speeds, it is possible to operate the motor with temperature-controlled oil flowing through the motor, as long as the oil viscosity and motor speed are chosen so that a hydrodynamic wedge does not build up beneath the brushes.
3. Small sparks are generated at the brush interface which makes dc servomotors unsuitable for use in explosive environments.

Note that for all types of dc motors, dc power supplies used with power op-amps dissipate large amounts of heat. For high-power systems, a constant-voltage power supply is often used and the power pulsed. Pulsing the power at high frequency causes power dissipation to be in the form of radio waves, as opposed to heat, so that components do not burn out. The torque output of the motor will be proportional to the width of the pulses. The pulses are generated by a switching power supply that accepts a signal from the servocontroller to tell it how wide to make the pulses. This is called *pulse width modulation* (PWM). When one stands next to a CNC machine and hears a high-pitched ringing (typically at 3 kHz), one is hearing the vibration in the motor caused by the PWM drive. For machines with resolutions up to about $^1/_2\mu$m, these vibrations are usually not mechanically noticeable. For submicron-accuracy machines the vibrations can sometimes cause surface finish problems. Note that in an effort to eliminate the ergonomic problems associated with 3 kHz ringing and associated machine vibration at the submicron level, higher and higher switching frequencies are being used; however, the resulting signal is literally being broadcast into the environment and can disturb other sensitive electronic systems nearby. The higher the frequency, the shorter the wavelength and the easier it is for short lengths of wire to act as antennas.

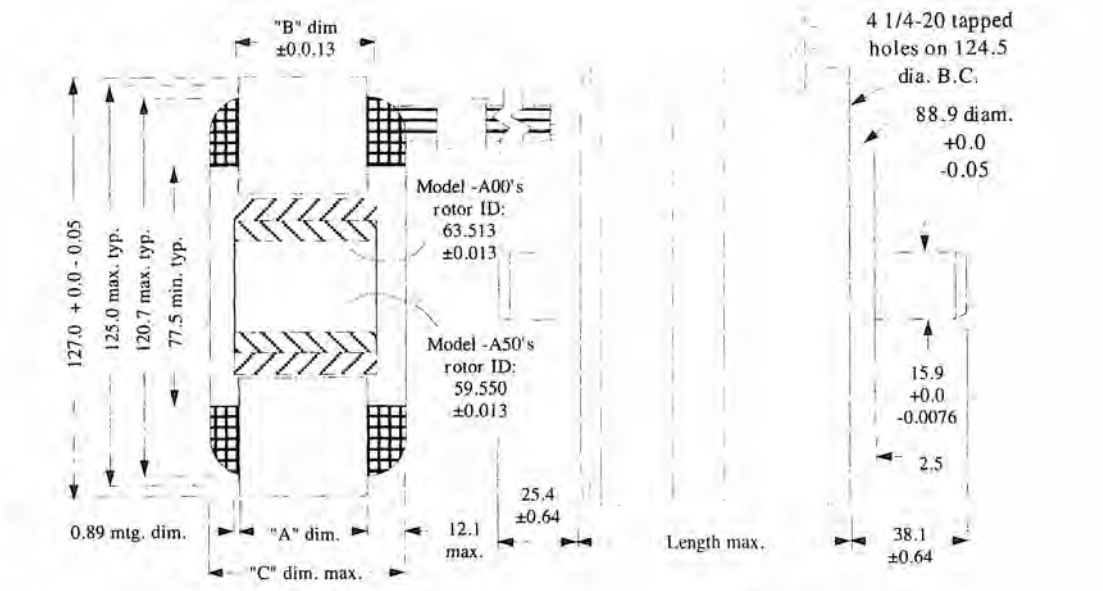
DC Brushless Servomotors

In a dc brushless servomotor, the magnetic field in the rotor is provided by permanent magnets. Hall effect sensors in the stator or resolver output are used to signal a *motor driver* when to switch the current in the stator's windings to create a rotating magnetic field that the rotor tries to follow. The motor driver still depends on the servocontroller to tell it what torque output is desired from the motor. Motor output torque is also directly related to the type of magnets used in the rotor, and no appreciable thermal energy is generated in the rotor. Heat generated in the stator windings is dissipated through the motor housing, which can easily be thermally isolated from the machine and/or cooled.

Figure 10.3.3 shows a family of brushless servomotors available in housed or frameless form. Figure 10.3.4 shows the data sheets for the motors which show that dc brushless motors can have very high power densities. Housed motors are also available with integral resolvers and/or tachometers, but an increase in motor length results. Because there are no brushes, dc brushless motors can be operated at zero rpm and high torque indefinitely as long as the winding temperature does not exceed the limit.

Because there are only a discrete number of magnets on the rotor, there will be a sinusoidal variation (coging) in the torque output of a dc brushless servomotor. Also, the shape of the signal (e.g., sinusoidal or square) sent to the windings by the motor driver affects the magnitude of the cogging torque. Cogging torque can sometimes be reduced by increasing the air gap or skewing the windings, but these methods also reduce efficiency. As discussed earlier, it is possible to model the effect of cogging torque on the performance of the system and compensate for it with a suitably fast control system. The cogging can also be mapped as a function of rotor speed and position, but using the information can be difficult unless the servocontroller and motor driver designs are carefully integrated.[18] Presently, motor drivers are available that vary the current to the windings in

[18] A new type of precision motor driver and servo controller, invented by the author and colleagues, is designed to map the cogging torque and then adjust the voltage to each of the motor's windings so as to eliminate the torque variations. See U.S. Patents 4,878,002 and 5,023,528.



Frameless motor Model	"A" dim.	"B" dim.	"C" dim.
RBE-03000-_50	13.3	17.9	35.2
RBE-03001-_50	26.7	31.2	48.5
RBE-03002-_50	41.3	45.8	63.1
RBE-03003-_50	54.6	59.2	76.5
RBE-03004-_50	66.7	71.2	88.5
RBE-03005-_50	82.6	87.1	104.4
RBE-03006-_50	95.3	99.8	117.1

Housed motor Model	Length
RBEH-03000-_50	58.9
RBEH-03001-_50	72.3
RBEH-03002-_50	86.9
RBEH-03003-_50	100.2
RBEH-03004-_50	112.3
RBEH-03005-_50	128.1
RBEH-03006-_50	140.8

Figure 10.3.3 Typically available frameless and housed dc brushless motors. All dimensions are in mm, and the drawing is not to scale. (Courtesy of Inland Motor Division, Kollmorgen Corp.)

a manner which can reduce cogging torque to the part per thousand range; however, most (not all) use switching power supplies that still cause the motors to "sing" at high frequency. Most machines can tolerate this high-frequency vibration, and those that cannot generally use dc servomotors with linear dc power supplies.

A motor driver's function could be performed by the servo controller, which could output appropriately shaped signals to a power op-amp associated with each winding. The motor's Hall effect sensors' outputs or resolver output would be connected to the servo controller. In this way a dc power supply could be used and the user would be free to experiment with various cogging torque mapping and correction algorithms. This type of system may become common with the advent of a high-processing-power digital signal processor-based (DSP) servo controller, such as referenced above.

Dc brushless motors and their drivers used to be very expensive, but new technologies and applications are making them cost competitive with other types of motors for machine tool axes. Dc brushless servomotors can provide very high torque in a small space with very little heat generation and thus are becoming more commonplace on machine tools, robots, and inspection machines. However, there are three main disadvantages of dc brushless servomotors:

1. The rotor can become demagnetized in the presence of high overload currents, other strong magnetic fields, or high temperatures. These factors must be prevented and in general factory environments where machine tools are used they are not a problem (in general, a brushed motor will be subject to similar environmental limitations). Also extreme care must be used when assembling unhoused motors into machine assemblies as the rotor magnets are extremely strong and brittle. When trying to put the rotor on a shaft, it will invariably move off to the side and stick to the housing unless a guide is used.

Winding	0-A00	0-A50	1-A00	1-A50	2-A00	2-A50	3-A00	3-A50	4-A00	4-A50	5-A00	5-A50	6-A00	6-A50
Stall T N-m*	5.9	14.7	11.3	28.2	16.9	42.2	21.9	55.2	26.5	66.7	32.4	82.0	38.3	97.4
Max T N-m	2.6	2.9	4.6	5.1	6.4	7.2	7.7	8.9	9.3	10.5	11.4	12.9	13.3	15.2
Max power W	208	210	247	249	291	293	318	320	354	356	424	426	465	466
K_MN-m/W	0.34	0.38	0.56	0.63	0.75	0.85	0.88	1.00	1.01	1.14	1.18	1.35	1.34	1.54
TPR C^o/W	1.2	1.2	1.1	1.1	1.0	1.0	0.91	0.91	0.85	0.85	0.78	0.78	0.73	0.73
Cogging N-m	0.08	0.1	0.14	0.18	0.20	0.25	0.26	0.32	0.31	0.39	0.38	0.46	0.43	0.54
Damping N-μm/rpm	38	44	70	82	100	119	130	153	160	184	190	224	240	274
J g-m^2	0.23	0.28	0.40	0.48	0.59	0.70	0.76	0.90	0.91	1.08	1.12	1.33	1.28	1.53
Mass g	1080	1123	1814	1882	2617	2739	3345	3487	4026	4167	4904	5131	5642	5897
# poles	12	12	12	12	12	12	12	12	12	12	12	12	12	12
Voltage	100	100	100	100	100	100	100	100	100	100	100	100	100	100
K_TN-m/amp	0.90	1.00	1.37	1.52	1.64	1.85	1.84	2.09	1.97	2.24	2.05	2.33	2.20	2.51
Max rpm	1040	940	690	620	575	510	510	455	480	425	460	410	430	375
R_{term} Ohms	7.2	7.2	5.8	5.8	4.7	4.7	4.4	4.4	3.8	3.8	3.0	3.0	2.7	2.7
L_{term}mH	19.4	17.7	22.5	20.5	22.3	20.3	21.8	19.8	20.7	18.7	18.2	16.5	17.3	15.7
Max continuous output power:														
Power W	208	210	247	249	291	293	318	320	354	356	424	426	465	466
Torque N-m	2.47	2.76	4.33	4.81	6.06	6.77	7.40	8.33	8.74	9.84	10.7	12.1	12.5	14.2
Speed rpm	800	730	545	495	460	415	410	365	385	345	380	335	355	315

* For 100C^o temp. rise in 15 sec.

Figure 10.3.4 Typical performance characteristics of Model RBE-0300 series dc brushless motors. (Courtesy of Inland Motor Division, Kollmorgen Corp.)

2. In an effort to remain cost competitive with brushed dc servomotors, brushless systems are sometimes supplied with motor drivers that do not attempt to minimize cogging torque because they output square waves to the windings. *Sine drives* output sine waves to the windings, where the amplitude is proportional to the position of the rotor as determined from a feedback device. Unfortunately, sine drives are often expensive. As always, *caveat emptor*!
3. Most motor drivers brake dc brushless motors by reversing the current input to the windings; hence almost as much power is expended in stopping them as was needed to get them going. For servoed axes this is not a problem, but for high-speed spindles the power usage and heat generation can become prohibitive.

AC Induction Servomotor

In an ac induction servomotor's stator, typically two 90 ^{o}-out-of-phase sine waves are input to two stator windings, which causes a rotating magnetic field to be established. Conductors on the rotor are at right angles to the magnetic field, which causes a current to be induced in them. This in turn causes a force which tries to push the conductor out of the field, or in this case, the rotor to turn. Torque is proportional to the amplitude of the sine wave and speed is limited by the frequency of the wave. The rotor always lags behind the sine wave, so its motion is *asynchronous*. Ac induction motors are perhaps the simplest and most rugged type of motor available.

A discrete number of windings in the stator and conductors in the rotor make ac induction servomotors also have cogging torque, but it is not feasible to minimize cogging through mapping. Also, some heat is dissipated in the rotor conductors. However, the simplicity and economy of ac induction servomotors makes them very attractive as high-horsepower motors, and they are often used to power spindles even though they must be braked by reversing the current to the windings.

Synchronous Reluctance Servomotors

Most engineers remember Ohm's law E = IR, and for magnetic circuits a similar law holds: $\mathcal{F} = \phi\mathcal{R}$, where $\mathcal{F}$ is the magnetomotive force, which, like voltage, is the potential that forms the magnetic field; ϕ is the flux of the circuit, which is like current; and R is the reluctance, which is the resistance to the field (note that for both laws, all the constants are functions of temperature). When an iron object such as a rod is placed in a magnetic field, a magnetic field is induced in the object. As a result, the two fields want to align and a torque is generated. When the objects are aligned, the reluctance has been minimized, and hence this is referred to as *reluctance torque*. By creating a rotating field in the stator, a rotor with a radial configuration of iron bars will follow the rotating field. The lag between the two is proportional to the torque exerted on the rotor, and as long as the rotor does not lag too far behind, its speed will be the same as that of the rotating field. If the rotating field were to get too far ahead of the rotor, then as it went past the rotor and came upon the rotor from the other side, the torque would reverse itself. As a result, the rotor would just vibrate back and

forth. Thus a method is also needed to start the rotor turning until its speed is synchronous with that of the rotating field.

In order to start a reluctance motor, one can vary the frequency of the current that creates the rotating field. The frequency can be controlled if the rotor position is measured. Another method is to combine the rotor structure of an induction motor with that of a reluctance motor. The result is called an *Amortisseur* winding. Remember that in order for current to be induced in the rotor, there must be relative motion between the rotor and the rotating magnetic field; hence there comes a point where the "radial bars" take over from the longitudinal bars and the motor moves from asynchronous to synchronous operation.

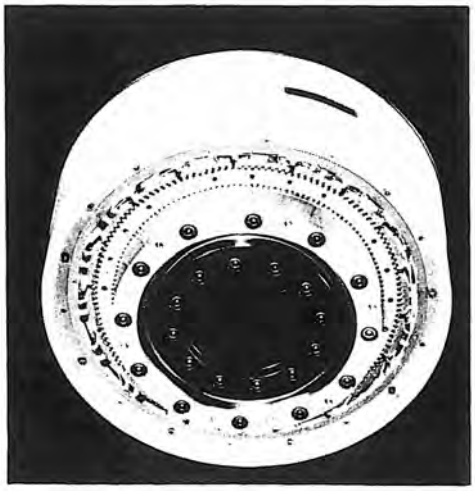


Figure 10.3.5 The Megatorque® high-torque, low-speed motor for direct-drive applications. (Courtesy of NSK Corp.)

The torque output and smoothness depend on the number of radial "teeth" on the rotor, the number of windings on the stator, and the resolution of the rotor position sensor. Often the stator has its own sets of radial teeth for focusing the magnetic field. An example of this type of configuration is shown in Figure 10.3.5, which shows a high-torque, low-speed motor that is designed for directly driving large loads. To generate very high torques, the rotor rides in an annulus between an inner and an outer stator. Figure 10.3.6 shows torque-speed curves for several available models, and Figure 10.3.7 shows the data sheets for the motors. Among their intended applications are in direct-drive robot arms and rotary index tables. Linear versions of these motors are also available.

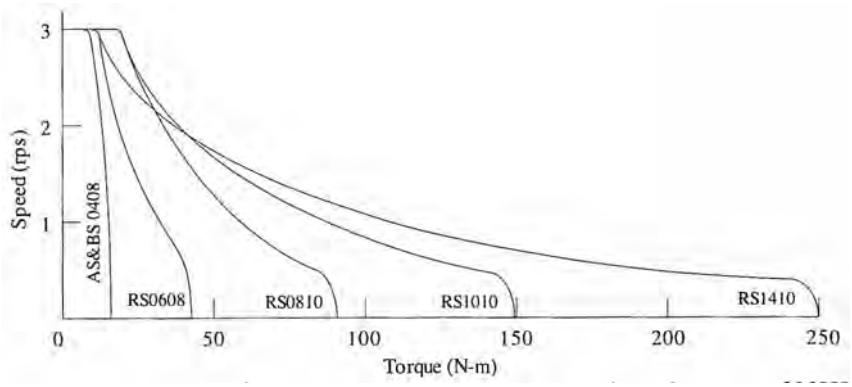


Figure 10.3.6 Megatorque® motor torque-speed characteristics. (Courtesy of NSK Corp.)

Specifications / Motor type		RS1410	RS1010	RS0810	RS0608	AS&BS0408
Max. Speed	(rps)	1.0/3.0				1.5/4.5
Max. Torque	(N-m)	250	150	90	40	10
Max. Amps per phase	(A)	7.5				6
Winding Voltage	(VDC)	330 or 165				
Resolver Resolution	(counts/rev)	614400/153600				409600/102400
Rotor Inertia	(kg-m-m)	0.267	0.076	0.02	0.0075	0.0023
Resolver Accuracy	(arc-sec)	± 30				± 60
Resolver Repeatability	(arc-sec	± 2.1/± 8.4				± 3.2/±12.8
Max. Friction Torque	(N-m)	8.0	5.5	4.5	3.0	1.0
Axial Load	(kN)	20	9.7	4.6	3.8	1.8
Moment Load	(N-m)	400	160	80	60	20
Axial Stiffness	(N/micron)	1000	710	330	250	400
Moment Stiffness	(N-m/microrad)	3.33	0.67	0.4	0.28	0.33
Mass	(kg)	73	40	24	14	6.5
Diameter	(mm)	380	290	220	180	120
Length	(mm)	170	160	140	120	120

Figure 10.3.7 Megatorque® motor characteristics. (Courtesy of NSK Corp.)

Linear versions of synchronous reluctance motors designed for high-accuracy positioning systems are often referred to as *Sawyer motors*[19] and operate as shown in Figure 10.3.8. A two-dimensional version is available where the base the slider moves over looks like a waffle iron, except that the waffle teeth are only a few millimeters across.[20] The slider is constructed from two orthogonal sets of teeth magnetically energized with windings as illustrated in Figure 10.3.9, so an umbilical to the slider is required. Additional forcers can be added to the motor and controlled differentially to obtain yaw angle control. Position of the platten is typically measured using differential plane mirror interferometers with long stick mirrors mounted on the motor as the target motor. Because of the strong attraction between the slider and the base, an air bearing is required to allow the slider to move across the base. For one-dimensional versions of this type of motor, air or roller bearings can be used.

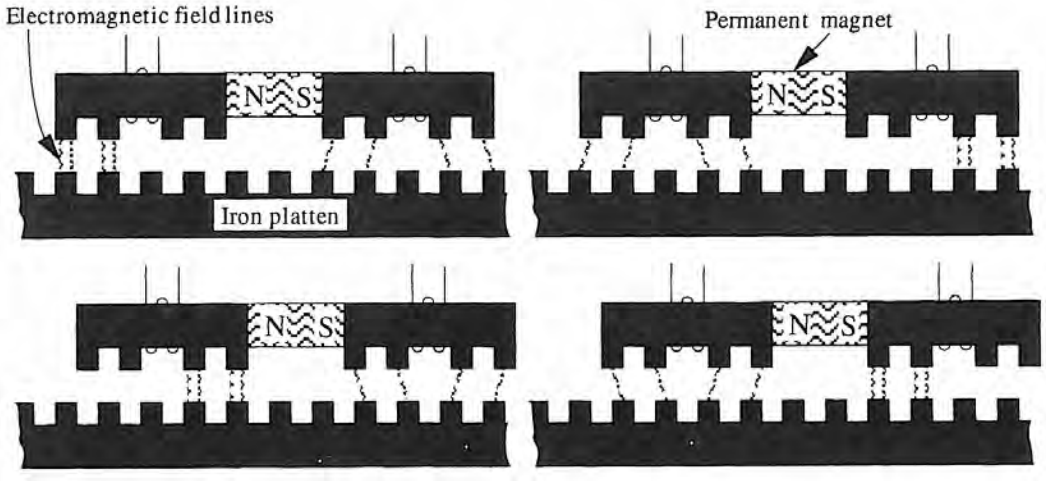


Figure 10.3.8 Current sequence in a Sawyer motor to produce linear motion. (After Pelta.)

[19] See, for example, E. Pelta, "Sawyer Motor Positioning Systems, Theory and Practice," Conf. Appl. Motion Control, University of Minnesota, June 10-12, 1986. Also see E. Pelta, "Precise Positioning without Geartrains," Mach. Des., April 23, 1987.

[20] See, for example, E. Pelta, "Two Axis Sawyer Motor," IECON'86, 12th Annu. IEEE Ind. Electron. Soc. Conf., Milwaukee, WI, Sept. 29-Oct. 3, 1986. Also see Xynetics Products (Santa Clara, CA) company literature.

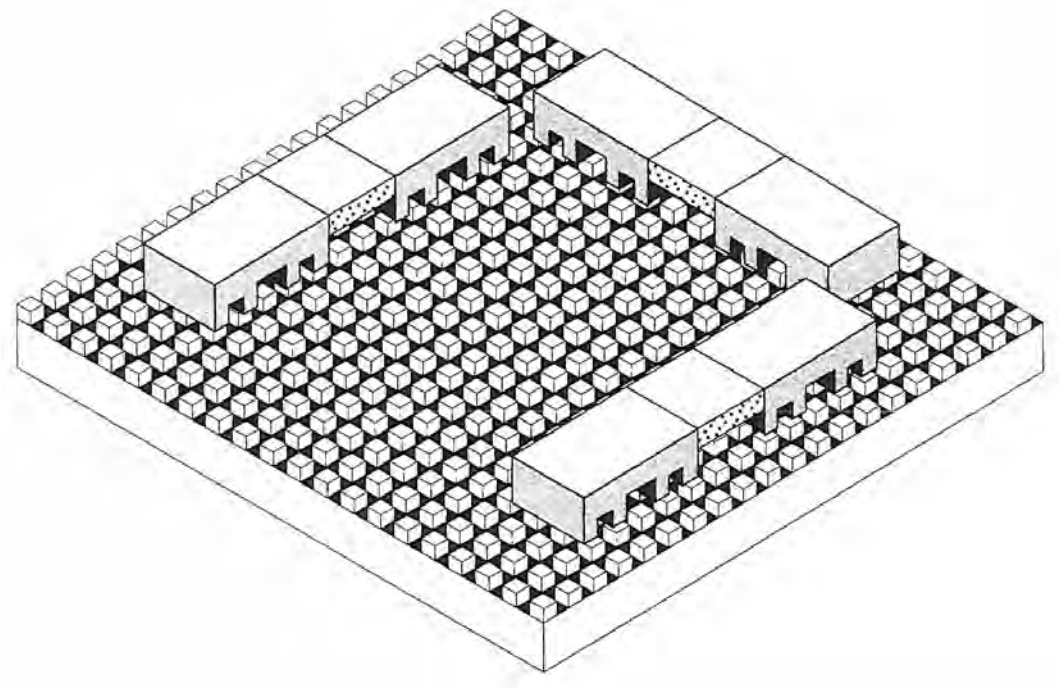


Figure 10.3.9 Concept for a two axis Sawyer motor. Platten and air bearing to support the motor are not shown. Multiple forcers could be used to control yaw. (After Pelta.)

Hysteresis Motors

Hysteresis motors have rotors that are made from a solid smooth piece of a magnetically hard material. As a rotating magnetic field is produced by the stator coils, it induces a magnetic field in the rotor; however, due to hysteresis in the rotor iron, its field lags behind that of the rotating field. The result is the generation of torque on the rotor that tries to minimize the lag. In addition, eddy currents are induced in the rotor which create magnetic fields of their own which also contribute to the torque. In general, torque is roughly constant until the synchronous speed is approached, where the torque drops to zero. Because of the solid rotor, there is virtually zero cogging torque in a hysteresis motor; however, motor torque/inertia is relatively low, so these motors are not generally intended for rapid start and stop operations. They are excellent for systems requiring constant or slowly varying speed capability without velocity variations caused by cogging torque. Since they are asynchronous motors, tachometer feedback is required if constant speed is to be obtained.

Stepper Motors

Stepper motors are synchronous motors that use a special motor driver to control the rate of the rotating magnetic field and keep track of the field's position. Hence the rotor position is known without the use of a feedback sensor, unless the maximum torque is exceeded. The step angle is determined by the design of the motor and the driver. Typically, the step angle is a few degrees, but with advanced drivers it can be a few milliradians. For consumer products, such as printers' paper position control, stepper motors are ideal because they do not require the added expense and complexity of a position sensor. For some types of machinery, such as assembly machines, index tables, and conveyors, they are also preferred for the same reason.[21] However, for axes that must be used for precision contouring, stepper motors generally have unacceptable cogging (detent) torques.

10.3.2 Motor Mounting Methods

Motors can be purchased that have their own housing that protects the windings and supports the motor shaft, or frameless motors can be purchased where the rotor and stator fit into the user's

[21] See, for example, C. Marino, "Selecting Stepmotors for Incremental Motion Applications," Motion, Nov./Dec. 1986.

assembly. The main advantage of housed motors is they are modular, so repair usually involves unbolting the old and bolting on the new. The principal disadvantage is they require the use of a coupling between their output shaft and the input shaft of the device. The coupling is usually flexible to accommodate shaft misalignment and prevent premature bearing failure, but flexible couplings decrease drivetrain stiffness.

Housed motors are usually mounted with a flange on the end of the motor where the output shaft protrudes, although foot mounts are available for larger motors. Often it is desirable to leave room in the design for the next larger or smaller motor size and even have the mounting holes drilled. This allows for an easy retrofit of the machine should it be desired. It is important to follow the manufacturer's recommendations for the heat dissipation capability of the object to which the motor is mounted. If the motor was completely insulated from the rest of the machine, then chances are it would quickly overheat unless auxiliary cooling was provided.

This brings up an interesting design requirement, for as shown in Chapter 6, one of the primary heat sources in a precision machine is the servo-axis drive motors. Most machine tool manufacturers bolt their servo-axis drive motors directly to cast iron machine structures, with the idea that the entire system should come to thermal equilibrium in a short time. For a high-precision machine where all heat sources must be minimized, a thermal break needs to be used. This means using a low-thermal-conductivity material as a spacer between the motor and the machine it is bolted to, and a sheet metal cover over the motor so that its heat does not radiate heat to the machine. An example of a spacer design is the polymer concrete spacer with integral air cooling passages shown in Figure 2.3.11.

When thermally isolated from the rest of the machine, the motor often does not have sufficient thermal mass or surface area to dissipate the heat it may generate during continuous operation. Hence the motor may have to be cooled either with a water jacket or by a fan. In the latter case, the air must be ducted so that it blows away from the machine. If the entire machine lives in a room whose temperature is being carefully controlled, then cooling the motor with temperature-controlled fluid may be desired to minimize the introduction of thermal gradients into the room.

The rotor of a frameless motor is hollow and is made to fit onto a power input shaft (e.g., a ballscrew). As a result, there are no coupling errors or loss of stiffness due to a flexible coupling. In addition, for very fine motion control where the shaft may be supported by aerostatic or hydrostatic bearings, one does not have to worry about the drag torque from a housed motor's bearings and seals.

The bearings that support the power input shaft must be of very high quality to maintain concentricity of the rotor and the stator. The stator is usually housed in a bore that was machined at the same time as the bore for the shaft bearings. The stator is usually held in place with a key and an adhesive, so removal is very difficult. To avoid thermal problems, the housing assembly is often thermally isolated from the rest of the machine and is cooled as described above.

One must also be careful of assuming that a motor is dynamically balanced. Many motor manufacturers seem to assume that because their designs are symmetrical, the design will be balanced. However, variations in material density and component size can lead to dynamic imbalance that can cause varying levels of vibration in the system at different speeds. For a motor driving a leadscrew on a CNC turning center, this may not be a problem. For the drives of diamond turning machine spindles, each motor may have to be dynamically balanced and the magnetic center made to coincide with the shaft's geometric center.

Linear Motor Mounting Methods[22]

Linear electric motors can have moving or fixed coil designs. The rail the coil interacts with can be a permanent magnet design in the case of a dc brushed or brushless design or a toothed design in the case of a linear stepping motor (Sawyer motor). In either case, there is a very high attractive force between the coil and the rail which must be prevented from causing the coil to crash into the rail. This attractive force can be many times the maximum axial force the motor can provide. This attractive force can have a sinusoidally varying component that changes with position, which in turn can translate into sinusoidal straightness errors on a precision slide if the axis bearings are also used to support the motor. To minimize the effect of the attractive forces on the system, a balanced design is often used where the moving coil moves between two rows of magnets (or vice versa). The forces

[22] See, for example, W. E. Barkman, "Linear Motor Slide Drive for Diamond Turning Machine," Precis. Eng., Vol. 3, No. 1, 1981, pp. 44–47.

on this double-rail configuration are therefore balanced and only small lateral forces are generated as a result of the moving coil not being exactly centered, or because of variations in the magnets' field strengths. The amount of the resulting straightness variation may range from microinches to microns depending on the design of the system and the manner in which it was manufactured.

Figures 10.3.10 – 10.3.13 show four different potential configurations for a high-precision air bearing-supported carriage powered by a linear motor. In the first case, a simply supported air bearing carriage is actuated by a linear motor. The carriage is driven off center, so there will be a moment on it which causes Abbe errors on the working surface[23]. In addition, heat rising from the coil goes directly into the system. In the second case, a T-shaped bearing is used to minimize rail deformations and the motor is mounted to one side. This has less of a thermal problem, but a moment is still generated on the carriage. In the third design one motor is used on each side of the carriage. Here the problem of coordinating the motion of the two motors exists, as well as the expense of having two motors. In the fourth design, the motor is within the bearing rail and drives the carriage through its center of mass, but then heat is input directly into the system and a custom bearing design is required. Other configurations are possible and active cooling can be added, but these figures show that using linear motors is not as simple as one might think.

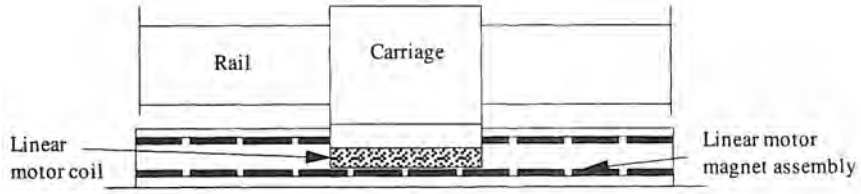


Figure 10.3.10 Linear motor moving coil mounted to underside of an air bearing carriage that rides on a simply supported rail.

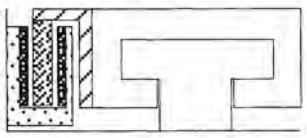


Figure 10.3.11 Linear motor moving coil mounted to the side of an air bearing carriage that rides on a T-shaped rail that is fully supported along its length.

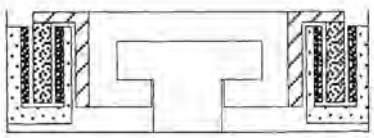


Figure 10.3.12 Linear motor moving coil mounted to the sides of an air bearing carriage that rides on a T-shaped rail that is fully supported along its length.

Linear or Rotary? That is the Question

What is better to use for a given application, a linear or a rotary motor? For applications that require the machine to resist substantial cutting forces (e.g., hundreds of newtons or more), in almost all cases a drive train with a rotary motor powering a leadscrew should be used. This is due to the fact that the cost of a motor and its driver increases only slightly with speed capability but increases dramatically with torque (or force) capability. Hence you get more for your money by using a motor with a transmission element (e.g., a leadscrew). This is also true when one considers that one of the

[23] See Equation 2.2.29 for a discussion on the center of stiffness of a bearing system.

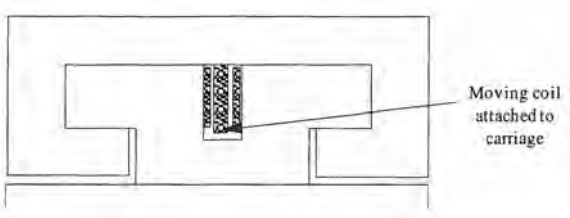


Figure 10.3.13 Linear motor moving coil mounted to the inside of an air bearing carriage that rides on a T-shaped rail that is fully supported along its length.

principal costs of a dc brushless motor is that of the magnet. In addition, the transmission ratio of the leadscrew acts to greatly increase system stiffness.

For a measuring machine, particularly large ones, it is often impractical to use leadscrews or other mechanical transmissions, and thus linear motors would seem ideal.[24] More and more measuring machines are using linear motors; however, one should note that one of the major concerns is it is difficult to thermally isolate the coil from the rest of the machine. The coil usually is dumping its waste heat right into the most sensitive part of the machine. Greater use of linear motors is inevitable (and welcome if systems are properly designed), so design engineers should build up suitable catalog files of manufacturers' literature.

Notes on Permanent Magnets[25]

Permanent magnets are the heart of many types of electromagnetic actuators, and thus various types of available magnets will be discussed briefly here. Up until the mid 1970s, only ceramic and alnico permanent magnets were used in actuators. Ceramic magnets are made from sintered barium ferrite or strontium ferrite, both of which are relatively inexpensive, and have energy products[26] in the range 3-4 MGOe. Alnico magnets contain cobalt, which is rare and expensive, and have energy products in the range 7-9 MGOe. Magnets made from rare earth metals were then developed, generically called samarium cobalt magnets, which have energy products in the range of 20-30 MGOe. Rare earth elements are expensive, but the high-energy products meant that less magnetic material was needed to begin with; hence most high-performance actuators now use rare earth magnets.

In the early 1980s, neodymium-iron-boron magnets were developed[27] by Magnaquench®, a business unit of the Delco-Remy Division of General Motors. These magnets are sold under the trade name magnaquench® and in the sintered form can have energy products in the range of 27-35 MGOe. When the material is powdered, mixed with a polymer, and injection molded, energy products in the range of 7-9 MGOe are obtained. Neodymium is 10-20 times more plentiful then samarium, so as the technology develops for this new magnetic material, its performance is expected to increase to twice that of rare earth magnets while the cost drops to one-half. One of the largest applications of this new material is for automobile starter motors, but servomotor manufacturers will probably also rapidly embrace this new technology.

[24] Recall the CMM case study in Chapter 1.
[25] From an in-house memorandum by Jack Kimble at BEI-Kimco. Thanks Jack.
[26] The *energy product* is a figure of merit for magnetic materials, and it is expressed in millions of gauss oersteds (MGOe).
[27] See U.S. Patent 4,496,395. Also see B. Carlisle, "Neodymium Challenges Ferrite Magnets," Mach. Des., Jan. 9, 1986.

10.4 LIMITED RANGE OF MOTION ELECTROMAGNETIC ACTUATORS[28, 29]

A solenoid uses a coil of wire to generate a magnetic field that attracts an iron component toward the coil. A typical high-efficiency solenoid is constructed as shown in Figure 10.4.1, and there are a seemingly infinite number of other design variations. The approximate force (N) generated by a solenoid of the type formed by plane parallel surfaces is primarily a function of the number of turns N of the coil, the current I (A), the pole area A (m^2), the air gap h (m), and magnetic permeability of air $\mu(4\pi \times 10^{-7}\ N/A^2)$:

$$F = \frac{N^2 I^2 A\mu}{2h^2} \tag{10.4.1}$$

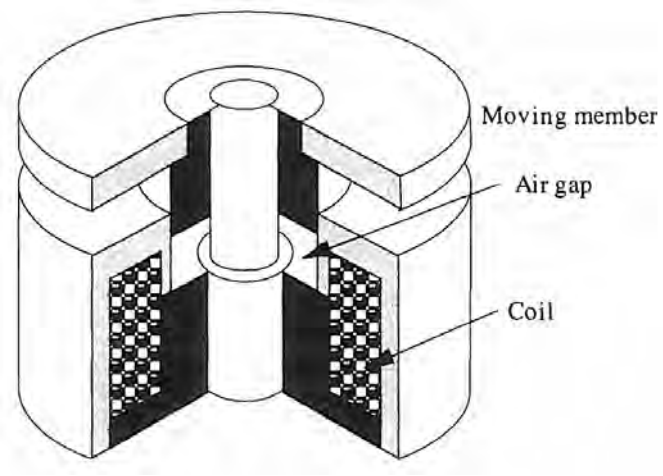


Figure 10.4.1 Construction of a high-efficiency solenoid actuator. (Courtesy of Lucas Ledex, Inc.)

The length of the iron path, the magnetic saturation properties of the structure, and the area and shape of the pole pieces also affect performance. The nonlinearity of the force response is apparent from this relation. Although many other solenoid designs exist, most have similar nonlinearities. Since solenoid actuators are entirely dependent on the coil to establish the magnetic circuit, they also have a slow electromechanical time constant. Hence solenoids are most often used as inexpensive actuators for forcing components against fixed mechanical stops .

Wound coil/permanent magnet actuators, often called *voice coil actuators* because they are used in loudspeakers, use permanent magnets to establish a magnetic circuit; thus any infinitesimal amount of current applied to a coil inside the magnet structure almost instantly produces a commensurate change in the magnetic field and force on the coil. It is as if the permanent magnets act to "preload" the system, thereby increasing its bandwidth and linearizing its force-displacement characteristics; Lorentz's equation shows that the force is proportional to the permanent magnet's magnetic field B, the engagement length l of the moving coil, the current i, and the number of turns N:

$$F = B\,\ell\,iN \tag{10.4.2}$$

For small motions, l is essentially constant. Moving coil actuators provide significantly higher performance characteristics than solenoids, but at a commensurate increase in cost. For example, a large actuator that can supply 500 N of force and operates without external cooling may cost $7500, an order of magnitude greater than the cost of a solenoid. Both linear and rotary motion moving coil actuator designs can be manufactured.

The coil is often made the moving member because it usually has a lower mass than the magnet assembly. Since the stroke is short compared to the size of the actuator, a loop of wire to the coil allows power to be transmitted to the moving coil without applying any appreciable reaction

[28] For a good design reference, see H. Rotors, Electromagnetic Devices, John Wiley & Sons, New York.

[29] The author would like to thank Jack Kimble of BEI Motion Systems Co., Kimco Magnetics Division, San Marcos, CA, for his help in preparing this section.

force. Note that the coil must be supported by a linear bearing, which is usually not provided by the actuator manufacturer. It is up to the designer of the machine system to choose the bearing type and incorporate it into the design. Since strokes are comparatively short for this type of actuator, flexural bearings are often used to support the coil.

Two design variations for linear motion moving coil actuators are shown in Figure 10.4.2. In the conventional moving coil actuator design, a cylindrical coil moves axially inside a cylindrical magnet assembly that is radially magnetized. When the coil's length exceeds the length of the magnet by at least the stroke length, the response is more linear, heat is dissipated better, and efficiency is increased; however, the increased mass and inductance of the coil decreases the system bandwidth. When the coil is shorter than the magnet assembly, the inverse is true. Where maximum force-to-weight ratio is required, the flux focusing actuator design should be used. This design allows the surface area of the magnet to be much larger than the air gap cross section of the conventional design. There are also fewer flux leakage paths with the flux focusing design, so that almost all the magnetic flux passes through the air gap. Flux focusing actuators have much greater force-to-weight ratios, larger bandwidths, and higher efficiency than those of conventional designs.

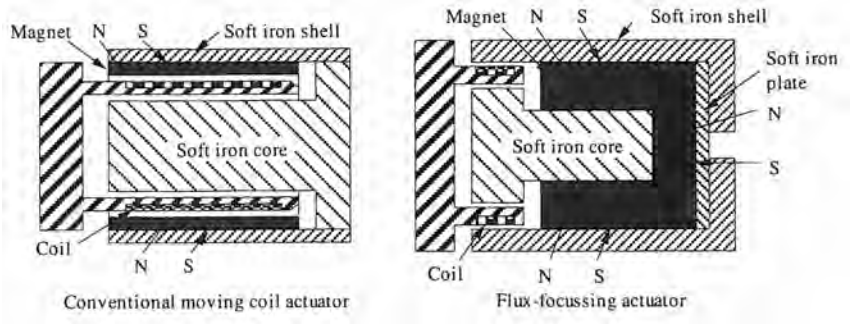


Figure 10.4.2 Two types of moving coil actuators. (Courtesy of BEI Motion Systems Company, Magnetics Division.)

For limited-range-of-motion applications, voice coil actuators are far superior to most other actuators because voice coil actuators have zero mechanical hysteresis (but they do have magnetic hysteresis which affects the control system design), zero force or cogging torque, and zero backlash. The same type of actuator constants are used for evaluation of voice coil actuators as are used for electric motors. When suitable bearings are used to support the core, the positioning capability of a system that uses voice coil actuators is limited only by the sensor and servocontroller. Nanometer and nanoradian resolutions are possible with voice coil actuators, over a range of many millimeters. Their principal drawback is, like electric motors, that they can generate considerable heat. Hence, if possible, piezoelectric or magnetostrictive actuators should be used.

Moving coil actuators are used in applications ranging from computer disk drive head positioners to mirror positioners for high-speed optical scanning equipment and precise positioning of mirrors and lenses used with high-power lasers.[30] They have also been used to provide six-degree-of-freedom control to a robot wrist.[31] As mentioned in Section 8.6.3, moving coil actuators are often used for the precise position control of silicon wafers in photolithography applications. Even a two-axis hemispherical configuration can be designed to sweep out a solid angle. Figure 10.4.3 shows the properties of some available linear and rotary moving coil actuators, and Figures 10.4.4 and 10.4.5 show representative physical mounting details.

For optical scanning applications, it may be desirable to purchase a complete unit that one merely fastens the mirror to, and plugs into a controller. This type of device is usually referred to as a *galvanometer.*[32] In order to economically support the shaft while allowing for very high bandwidths, these devices use precision instrument-grade ball bearings. Because of the extremely low coefficient

[30] See, for example, J. Kimble, "Rare Earth-Cobalt Magnets as applied to Linear Moving Coil Actuators," 3rd Int. Workshop Rare Earth-Cobalt Perm. Magnets Their Appl., University of California, San Diego, CA, June 27-30, 1978.

[31] R. L. Hollis, A. P. Allan, and S. Salcudean, "A Six-Degree-of-Freedom Magnetically Levitated Variable Compliance Fine Motion Wrist," 4th Int. Symp. on Robot. Res.,Santa Cruz, CA, Aug. 9–14, 1987.

[32] Galvanometers are available, for example, from General Scanning, Inc., (617) 924-1010.

Linear actuators

Model	Cont. force	Peak force	Stroke ±	K_f	K_m	R_{elec}	t_{elec}	R_{therm}	Weight coil	Total weight	Height	Width or diam.	Length
	lbf	lbf	in	lbf/amp	lbf/W$^{1/2}$	Ohms	μsec	C°/W			in	in	in
LA10-12	1.20	1.3	0.03	0.53	0.42	1.8	20.0	10.5	0.35oz	3.2oz	-	1.09	1.20
LA13A-30	1.12	1.5	0.69	0.67	0.31	4.6	731	6.58	0.71oz	15.6oz	1.33	1.25	3.00
LA14-15	2.00	3.5	0.09	1.25	0.61	4.20	54.0	8.0	0.89oz	8.0oz	-	1.49	1.50
LA14B-24	2.73	4.0	0.88	1.20	0.54	5.00	740.0	3.4	0.77oz	22.7oz	1.36	2.75	2.38
LA15-15	5.54	5.0	0.04	1.35	1.00	1.90	100.0	2.8	0.80oz	9.6oz	-	1.50	1.53
LA26-29A	3.20	8.0	0.30	1.30	0.89	2.10	-	6.7	0.28lbf	2.4lbf	-	2.62	2.88
LA20B-26	6.00	11.0	0.08	1.30	1.60	0.66	159.0	-	0.13lbf	2.2lbf	-	2.05	2.70
LA22-34A	6.84	15.0	0.06	0.75	1.09	0.47	35.0	2.2	0.08lbf	3.0lbf	-	2.19	3.38
LA30-27	12.25	15.0	0.13	9.40	2.00	21.1	536	2.3	4.00oz	3.9lbf	-	3.00	2.78
LA30-41-2	8.70	30.0	0.30	2.60	1.80	2.20	1800	3.7	0.50lbf	5.3lbf	-	3.00	4.08
LA33-58A	14.20	30.0	0.05	1.60	2.05	0.61	200	1.8	0.21lbf	12.3lbf	-	3.31	5.84
LA40-52	15.00	30.0	0.63	2.00	1.70	1.40	220	1.1	0.40lbf	12.4lbf	-	5.31	4.48
LA40-60	13.10	40.0	0.10	3.70	2.44	2.40	-	3.0	0.81lbf	13.5lbf	-	5.31	5.13
LA42A-44A	30.60	70.0	0.18	4.30	3.30	1.70	81.0	1.0	0.37lbf	20.3lbf	4.44	4.44	4.40
LA90-49A	73.40	95.0	0.57	7.90	5.60	2.00	250.0	0.5	1.70lbf	34.4lbf	-	8.12	4.86
LA78-54	73.40	100.0	0.25	7.90	5.60	2.00	-	0.5	2.30lbf	34.3lbf	-	7.75	5.39

Rotary actuators

Model	Cont. torque	Peak torque	Stroke ±	K_f	K_m	R_{elec}	t_{elec}	R_{therm}	Weight coil	Total weight	Height	Width or diam.	Length
	oz-in	oz-in	degrees	oz-in/amp	oz-in/W$^{1/2}$	Ohms	μsec	C°/W	oz	oz	in	in	in
RA23-06	6.42	6	13.0	5.6	3.0	3.4	560	17.8	0.22	1.6	1.14	0.65	1.79
RA25-11	14.00	25	16.0	18.0	4.7	14.6	180	9.7	0.25	7.0	1.47	1.03	2.20
RA60-10	56.11	80	15.0	19.0	13.8	1.9	730	5.2	0.75	14.7	3.12	1.19	3.75
RA68-12	88.28	110	10.0	45.0	21.5	3.5	800	5.1	1.10	18.0	3.40	1.19	3.75
RA55-22	356.40	2300	15.0	209.3	55.7	13.7	2300	2.1	11.00	147.2	2.75	2.19	-

Figure 10.4.3 Properties of some commercially available moving coil actuators. (Courtesy of BEI Motion Systems Company, Magnetics Division.)

of friction in this type of ball bearing (<0.005), mechanical damping is limited. In order to damp out transient response oscillations within two undamped oscillation periods, the back EMF can be used as a feedback signal. Better response can be obtained from units with integral precision feedback sensors. When a mirror is mounted on the galvanometer shaft, the bandwidth will decrease by a factor equal to the square root of the ratio of the mirror plus galvanometer inertia to the galvanometer inertia. Typically, galvanometers can be sinusoidally driven at 85% of their natural frequency. With discontinuous waveforms (e.g., triangular or square waves), they can be driven at 5% of their natural frequency. Two galvanometers with their axes of rotation placed orthogonal to each other can be used to make a laser scan a two-dimensional surface (e.g., the laser beam path in laser light shows). By feeding various sine and cosine waveform combinations into the galvanometers, virtually any pattern can be generated.[33]

10.5 PIEZOELECTRIC ACTUATORS[34]

A piezoelectric material changes shape when a voltage is applied across it. An angstrom or so of motion typically results from every volt applied to a single piezoelectric crystal used as an actuator. Hence crystals are often stacked upon each other with electrodes in between and high voltages used to obtain microns of motion. Depending on its mass and stiffness, the bandwidth of a piezoelectric actuator can be on the order of several kilohertz. A piezoelectric actuator's resolution and high bandwidth are often orders of magnitude better than those of most other actuators.[35] A typical piezoelectric actuator dissipates only milliwatts of power, which will not disturb the thermal equilibrium

[33] This makes a wonderful exercise in applications of Fourier transforms and spatial geometry. Of course one can also just hook up the system and input various waveforms, such as from various types of music and see what happens. Obviously, nonclassical music will produce extremely distorted images (no kidding, try it).

[34] My wonderful engineer wife wrote the original version of this section as part of her M.Sc. thesis: D. L. Thurston, Design and Control of High Precision Linear Motion Systems, MIT, Electrical Engineering Department, April 1989.

[35] See, for example, T.G. King, "Piezoelectric Ceramic Actuators: A Review of Machinery Applications," Precis. Eng., Vol. 12, No. 3, July 1990, pp. 131–136.

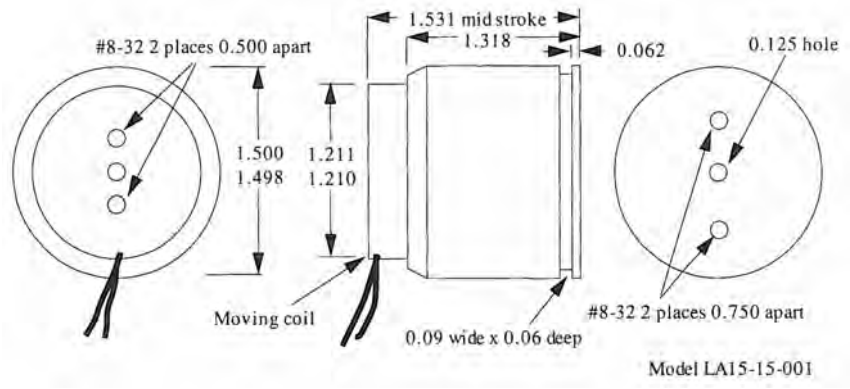


Figure 10.4.4 Typical configuration of a moving coil actuator. (Courtesy of BEI Motion Systems Company, Magnetics Division).

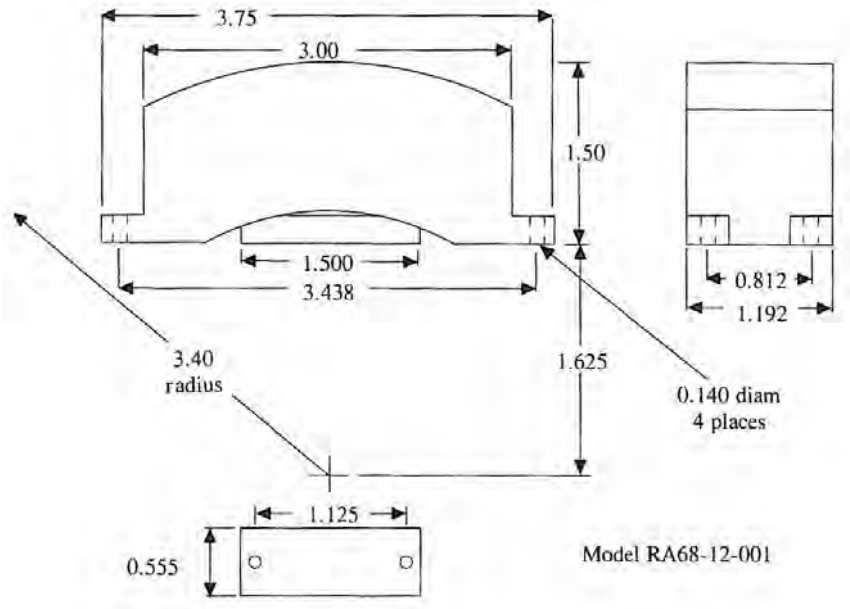


Figure 10.4.5 Physical configuration of a rotary motion moving coil actuator. (Courtesy of BEI Motion Systems Company, Magnetics Division.)

of a precision machine. Equivalent electromagnetic actuators dissipate heat orders of magnitude higher, due to winding resistances and eddy current losses.

Piezoelectric actuators have proven themselves in many precision applications, ranging from fast tool servos (FTSs), used to compensate for spindle error motions in large diamond turning machines,[36] to indexing diffraction grating ruling machines,[37] to scanning tunneling microscope actuators.[38] An example of an FTS developed for a large diamond turning machine (see Section 5.7.1) is shown in Figure 10.5.1.

[36] See, for example, S. Patterson and E. Magrab, "Design and Testing of a Fast Tool Servo for Diamond Turning," Precis. Eng., Vol 7, No. 3, 1985, pp. 123–128, and A. Gee et al. "Interferometric Monitoring of Spindle and Workpiece in an Ultra-Precision Single-Point Diamond Facing Machine," Proc. SPIE, Vol. 1015, 1988, pp. 74–80.

[37] See, for example, A. Gee, "A Piezoelectric Diffraction Grating Ruling Engine with Continuous Grating-Blank Position Control," Proc. ICO Conf. Opt. Methods in Sci. and Ind. Meas., Tokyo, 1974 Japan., J. appl. Phys., Vol. 14, 1975, Suppl. 14-1, pp. 169–174.

[38] See, for example G. Binnig and H. Rohrer, "Scanning Tunneling Microscopy," Helv. Phys. Acta, Vol. 55, 1982. Also see P. Atherton, "Micropositioning using Piezoelectric Translators," Photon. Spectra, Vol. 21, No. 12, 1987, pp. 51–54.

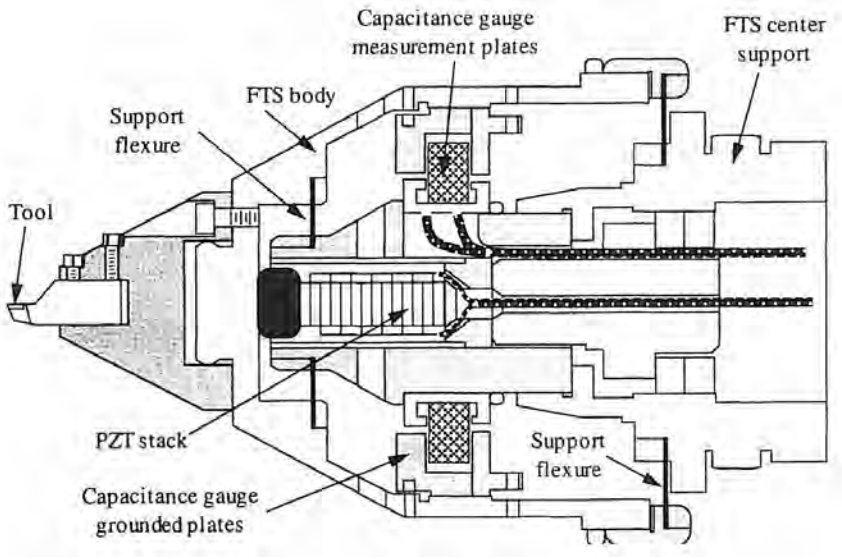


Figure 10.5.1 Piezoelectric actuated fast tool servo (FTS). (After Patterson and Magrab, Courtesy of Lawrence Livermore National Laboratory.)

Spindle designs have also been proposed and tested that use piezoelectric actuators to radially adjust the position of fluidstatic bearing pads supported by flexures as shown in Figure 10.5.2.[39] With this design, dynamic radial motion of a spindle at 1000 rpm can be reduced by a factor of 2.

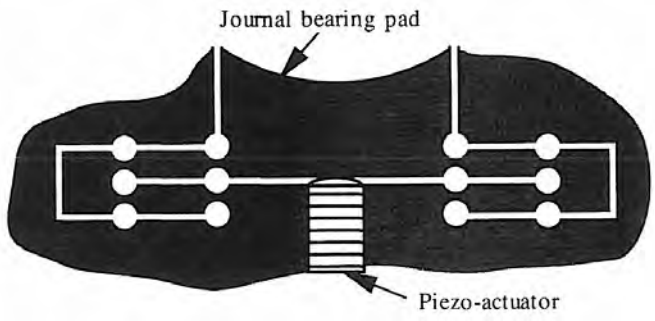


Figure 10.5.2 Method for controlling radial motion of spindle bearing pads. (After Horikawa et al.)

Piezoelectric materials can provide linear motion in three ways:

1. Stacks of piezoelectric disks can be epoxied together with electrodes between disks. When a voltage is applied to the electrodes, the stack expands or contracts, as is the case with the fast tool servo.
2. Piezoelectric materials can be epoxied together to form a composite beam which bends as the two materials expand a different amount when a voltage is applied (i.e., a bimorph).
3. Piezoelectric materials can be assembled to form a microstepper (e.g., an Inchworm®[40]. A rod is placed through the center of three piezoelectric cylinders (each comprised of a stack) with a hole bored through their centers as shown in Figure 10.5.3. The two end stacks operate in a radial expansion/contraction mode and the middle stack operates in an extension/contraction mode. Figure 10.5.3 shows the sequence of element motions needed

[39] See O. Horikawa et al., "Vibration, Position, and Stiffness Control of an Air Journal Bearing," 1989 Int. Precis. Eng. Symp., Monterey CA, preprints pp. 321–332.

[40] *Inchworm*® is a trademark of the Burleigh Corp. The term denotes the motion of the actuator which mimics the caterpillar namesake.

to enable the actuator to move along the rod. This sequence continues until the destination is reached. Other designs of this type of drive also exist.[41]

The resolution of piezoelectric microstepper actuators is a function of the resolution of the actuator between the clamping cycle, which can be on the order of an angstrom, and how the device clamps between steps. Any misalignment is likely to produce error motions orthogonal to the direction of motion. Misaligned clampers, improper phasing, and debris (e.g., formed by circumferential clampers) can also produce a jerky clamping action which will induce errors in the system. Many different high-resolution piezoelectric microstepper actuators have been developed and some are commercially available.[42] The primary problem with microstepper actuators is they do not necessarily provide continuous motion over a large distance. Within steps, the motion is of course continuous. Near continuous motion can be achieved by two microstepper motors operating out of phase with respect to one another: when one pulls, the other rests, and vice versa; however, the control algorithms needed are not trivial.

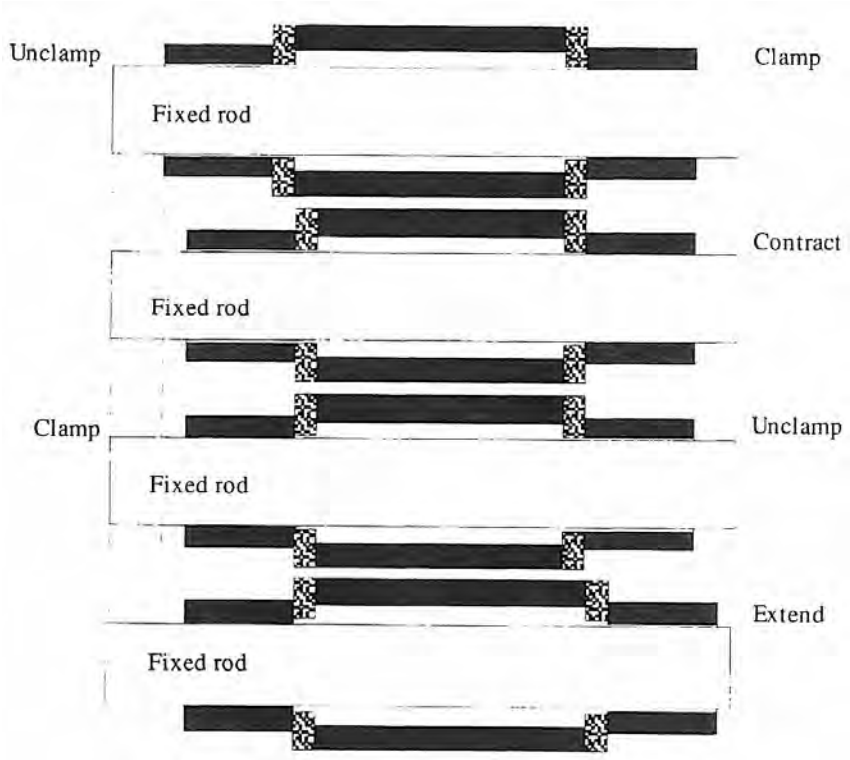


Figure 10.5.3 Control sequence for a piezoelectric microstepper-type actuator.

10.5.1 Properties of Piezoelectric Materials[43]

A piezoelectric material has a crystalline structure that produces a voltage at its surface when stretched or compressed. When pressure is applied to the structure along a certain axis, higher-valence electrons are freed by one or more of several mechanisms which generate a charge in the

[41] See A. Gee, "A 'Micro-incher' Machine Carriage Drive with Automatic Feedback Control of Step-Pitch, Step-Phase, and Inter-Step Positioning," Precis. Eng., Vol. 4, No. 2, 1982, pp. 85–91.

[42] See, for example, the catalog New Micropositioning Products, Burleigh Instruments, Fisher, NY. For other commercially available piezoelectric actuators, also see Physik Instrumente: The PI System Catalog, Physik Instrumente (PI) GmbH & Co., Waldbronn, Germany, and Piezoelectric Ceramics: Catalog and Application Notes, EDO Corporation, Western Division, Salt Lake City, UT. There are numerous other manufacturers of piezoelectric materials themselves.

[43] References to consult for a more detailed discussion include J. Herbert, Ferroelectric Transducers and Sensors, Gordon and Breach, New York, 1982, and W. Beam, Electronics of Solids, McGraw-Hill Book Co., New York, 1965. Also see H. Jaffe, "A Primer of Ferroelectricity and Piezoelectric Ceramics," Vernitron Corp. Tech.Report TP-217, and H. Jaffe, "Piezoelectricity," Vernitron Corp. Tech. Paper TP-238, Bedford, OH, 1961.

material. Piezoelectric materials can be used for pressure, acoustical, and acceleration sensing. Fortunately for machine designers, the opposite mode exists; a voltage applied to a piezoelectric material causes it to expand (or contract), which allows piezoelectric materials to be used as actuators.

For some crystals, each unit cell of the crystal is a dipole. A voltage applied to the crystal causes the dipoles to try to align themselves, resulting in expansion or contraction of the crystal. When the crystal's unit cells are not dipoles, it is the crystal's lack of symmetry which produces the piezoelectric effect. There are 32 types of crystals in nature and 21 are nonsymmetrical. Only one of the 21 nonsymmetric classes does not demonstrate the piezoelectric effect. When a nonsymmetric crystal has unit cells which are dipoles, the crystal is said to be *pyroelectric*. Pyroelectric crystals have the property that thermal expansion of the crystal causes the dipole to contract or expand, which causes a net charge to appear on the faces near the ends of the dipoles. Ten of the 20 nonsymmetric piezoelectric crystals are pyroelectric.

Ferroelectricity is the term used to describe the effect of dipoles changing orientations when an electric field is applied. Ferroelectric crystals therefore experience a hysteresis effect when an electric field is applied and reversed. All ferroelectric crystals are also piezoelectrics, but the inverse is not always true. To determine if a material is ferroelectric, a field is applied to pole the material, and many dipoles align themselves with the field. As the field increases, the amount of charge within the crystal saturates, as there is only a finite amount of charge the crystal can hold. The field is then decreased, passes through zero, and reverses direction. Suddenly the dipoles reverse direction to align themselves with the field. The field continues to increase in magnitude, and eventually the charge within the crystal saturates once again. The field is decreased again, passes through zero, and switches direction once again. The dipoles will again reverse direction when the field is large enough and the process can repeat itself. The charge remaining in the crystal when the field is zero is known as the *remnant charge*.

Ferroelectric crystals have dipole arrangements in *domains*, or areas where all the dipoles are aligned in the same direction. The domains form when the crystal has been heated to near an anneal point and then cooled. When a strong field is applied to opposite faces of an annealed crystal, the dipoles tend to align themselves with the field. This is called *poling* and is how many ceramics initially gain their piezoelectric properties. Not all of the dipoles in the various domains will change their direction, but enough will do so to increase the piezoelectric properties of the crystal; however, a poled material is never as strong as a single-domain crystal. A plot of strain versus voltage applied for a ferroelectric material will yield a Lissajous figure.

Ferroelectric crystals can be used as actuators, but they may be difficult to control because of the large amount of hysteresis encountered upon charge reversal. There are methods that can be used to make ferroelectric (and piezoelectric) actuators more controllable.[44] Take note that many ceramic piezoelectric crystals are also ferroelectric, but this is often not mentioned in the catalogs. Very high **d** constants ($>50 \times 10^{-12}$ m/V) are characteristic of ferroelectric materials. Ferroelectric ceramics were initially developed for acoustic transmission and reception devices where hysteresis effects are not necessarily undesirable but greater range of motion for less voltage was required. Much of the sales volume in custom-manufactured piezoelectric materials appears to be for acoustic applications, and thus the buyer of piezoelectric materials for actuators must be careful not to unknowingly choose a ferroelectric material from a catalog of piezoelectric materials. The supposed bad reputation that piezoelectric materials have gained as hard-to-control actuators has been from the application of ferroelectric materials as actuators without the users taking care to design the control systems to be able to deal with the increased hysteresis.

10.5.2 Piezoelectric Material Constants[45]

Piezoelectric materials usually behave anisotropically with respect to their piezoelectric properties and there are many constants that are used to characterize their behavior, including the $\mathbf{d}_{ij}$, $\mathbf{g}_{ij}$, $\mathbf{k}_{ij}$,

[44] See, for example, H. Kaizukz and B. Siu, "A Simple Way to Reduce Hysteresis and Creep When Using Piezoelectric Actuators," Jpn. J. Appl. Phys., Vol. 27, No. 5, 1988, pp. L773–L776; and H. Kaizukz, "Application of Capacitor Insertion Method to Scanning Tunneling Microscopes," Rev. Sci. Instrum., Vol. 60, No. 10, 1989, pp. 3119–3122. Also see H. S. Tzou, "Design of a Piezoelectric Exciter/Actuator for Micro-displacement Control: Theory and Experiment," Preci. Eng., Vol. 13, No. 2, 1991, pp. 104–110.

[45] See, for example, C. Germano, "On the Meaning of 'g' and 'd' Constant as Applied to Simple Piezoelectric Modes of Vibration," Vernitron Corp. Tech. Report TP-222, and R. Gerson, "On the Meaning of Piezoelectric Coupling," Vernitron Corp. Tech. Report TP-224, Bedford, OH.

and **K** constants. A 1, 2, or 3 subscript corresponds to one of the material's axes, as shown in Figure 10.5.4. A 4 represents the plane parallel to the 2 and 3 axes which is normal to the 1 axis; a 5, the 1-3 plane; and a 6, the 1-2 plane. When used as the second subscript in one of the constants, values greater than 3 indicate that a shear strain or shear stress occurs in that plane. Remember that by using Mohr's circle and changing orientation, shear stresses and shear strains can be resolved into axial components.

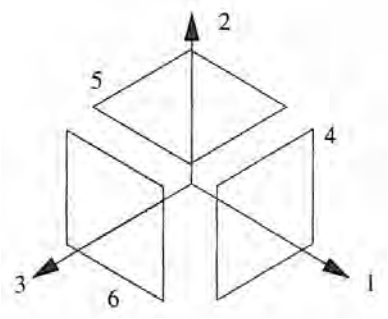


Figure 10.5.4 Axis and shear plan definition for piezoelectric materials.

The shape and symmetry about axes of the crystal affect the number of piezoelectric constants associated with each crystal. Not all constants are possible with all crystals. The more symmetrical the crystal, the fewer the piezoelectric constants it has. Therefore, the easiest piezoelectrics to work with are the crystals which are asymmetric about only one or two axes, which helps prevent unwanted cross-axis coupling. For example, quartz, a member of a trigonal crystal system with three twofold axes of symmetry and a threefold principal axis, has only 2 piezoelectric constants.[46] Figure 10.5.5 shows some typical values of piezoelectric materials' constants.

*The **d** Constant*

The piezoelectric strain constant $\mathbf{d}_{ij}$ quantifies the relation between electric *field*[47] E in the i direction (E = volts/thickness) and mechanical *strain*[48] ε in the j direction ($\varepsilon = \Delta\ell/\ell$):

$$\varepsilon_j = \mathbf{d}_{ij} E_i \tag{10.5.1}$$

The letter h is also sometimes used as a subscript for the **d** constant for ceramics, where $\mathbf{d}_h$ indicates that stress is generated or applied equally in the 1, 2, and 3 directions (i.e., hydrostatic stress). It also indicates that the electrodes, which are used to connect the piezoelectric to the voltage source or meter, are perpendicular to the 3-axis. It should be noted that only certain classes of piezoelectric materials will have a piezoelectric effect when hydrostatic pressure is applied. This occurs in ceramic piezoelectric materials containing only a single polar axis. For example, zinc sulfide has only one piezoelectric constant, $\mathbf{d}_{14} = \mathbf{d}_{25} = \mathbf{d}_{36}$.

*The **g** Constant*

The piezoelectric stress constant $\mathbf{g}_{ij}$ is the relation between the field (volts/thickness) in the i direction and the stress (force/area) in the j direction:

$$E_i = \mathbf{g}_{ij} \boldsymbol{\sigma}_j \tag{10.5.2}$$

The constants **d** and **g** are linearly related by

$$\mathbf{d}_{ij} = K\varepsilon_o \mathbf{g}_{ij} \tag{10.5.3}$$

K is the relative dielectric constant and ε_o is the dielectric constant of free space (8.85×10^{-12} F/m).

*The Piezoelectric Coupling Coefficient **k***

[46] Twofold means that there is symmetry in two directions about an axis. The piezoelectric effect my be perpendicular or parallel to the twofold axis of symmetry.

[47] *Field* is defined as applied voltage per unit distance between electrodes.

[48] *Strain* is defined as change in length per unit length. Shear strain is defined as the tangent of the amount of out of squareness generated.

	T limit (°C)	Field dir.	Force dir.	d x10^{-12}	g	E (GPa)	K	k	Density (10^{3}kg/m^3)	Wave vel. (m/s)
Quartz	550	X	X	2.3	0.0578	80	4.5	0.1	2.65	5400
		X	Y	2.3	0.0578	80	4.5	0.1		5400
Ammonium dihydrogen phosphate	120	Z	45° XY	24	0.1750	19	15.5	0.29	1.8	3250
Rochelle salt	45	X	45° YZ	290	0.0936	18	350	0.68	1.77	3200
		Y	45° ZX	27	0.3316	10	9.2	0.3		2400
Barium titanate ceramic	45	Z	Z	190	0.0126	106	1700	0.52	5.7	4300
		Z	X	78	0.0052	110	1700	0.22		4400
Lead titanate zirconate (45/55)	100	Z	Z	140	0.0352	71	450	0.6	7.6	3100
		Z	X	57	0.0143	87	450	0.26		3400
Lead strontium titanate zirconate	300	Z	Z	250	0.0235	67	1200	0.64	7.5	3000
		Z	X	105	0.0099	81	1200	0.3		3300
Lead metaniobate	300	Z	Z	80	0.0402	60	225	0.42	6	3200
	300	Z	X	11	0.0055	60	225	0.04		3200

Figure 10.5.5 Physical properties of common piezoelectric materials. (After Jaffe.)

Piezoelectric coupling is defined as the ability of the material to change (transduce) one form of energy into another, such as the conversion of mechanical energy, by applied pressure, into electrical energy. It is also defined as the converse, the ability to convert electrical potential (applied voltage) into mechanical force. Much of the applied mechanical or electrical energy is stored in the material by the mechanism of elastic deformation or electrical capacitance, respectively; thus the **k** coefficient is not a measure of the efficiency of the crystal. Efficiency is a term associated only with losses. The efficiency of piezoelectric crystals is typically very high (>90%) because piezoelectric materials have high electrical resistances. Therefore, one should think of the **k** constant as a kind of "transmission ratio." In some cases one wants as much motion as possible per volt (e.g., for moving a tool in a FTS) and in other cases one wants very little motion per volt (e.g., for a scanning tunneling microscope's probe control).

The measure of piezoelectric coupling is defined as the piezoelectric coupling coefficient $\mathbf{k}^2$ but is quoted in catalogs as **k** because the values of $\mathbf{k}^2$ are sometimes very small:

$$\mathbf{k}_1^2 = \frac{\text{electrical energy generated}}{\text{mechanical energy applied}} \tag{10.5.4a}$$

and

$$\mathbf{k}_2^2 = \frac{\text{mechanical energy generated}}{\text{electrical energy applied}} \tag{10.5.4b}$$

$\mathbf{k}_1$ is used when the piezoelectric is used as a pressure- or force-sensing device. $\mathbf{k}_2$ is used when the piezoelectric is used as an actuator (note that $\mathbf{k}_1 = \mathbf{k}_2$). Values of **k** vary from zero to 1. Values for the coupling coefficients will generally be less than those quoted in catalogs, due to frictional and resistive elements in the system, but **k** can be as high as 0.8, with 0.5 not uncommon. The coupling coefficient is usually denoted $\mathbf{k}_{ij}$, where the i subscript denotes the direction of the electric field and the j subscript denotes the direction of the mechanical strain.

It is interesting to note that the coupling coefficient **k** can be determined by measurements of the modulus of elasticity E. The modulus gives different values when the electrodes of the piezoelectric are short (sc) and open (oc) circuited. The equation relating E_{sc}, E_{oc}, and **k** is

$$E_{oc}(1 - \mathbf{k}^2) = E_{sc} \tag{10.5.5}$$

The Dielectric Constant K

The dielectric constant[49] K of a material is a measure of the amount of charge that the material can store when it separates two electrodes. Piezoelectric crystals are dielectrics, and a material that is dielectric is also an insulator. Insulators are materials with no free charges available to conduct electricity. However, the common property of all dielectrics is their ability to store electrical energy.

[49] See W. H. Hayt, Jr., Engineering Electromagnetics, McGraw-Hill Book Co., New York, 1981.

Energy is stored by the shifting of dipoles or positive and negative charges within the material. These dipoles are bound; they can shift position only when an electric field is applied.

There are two kinds of dielectrics, polar and nonpolar. Polar dielectrics have permanent offsets between the center of gravity of positive and negative charges (i.e., each molecule is a dipole). The dipoles are randomly oriented until an electric field is applied. Then the molecules try to align with the electric field. Nonpolar dielectrics have no dipole arrangement until an electric field is applied; these materials need to be poled.

When an insulator is placed between two electrodes, a charge will build up between the electrodes as a potential is applied, thus forming a capacitor. The amount of charge or capacitance increases between the electrodes when a dielectric material is used, as opposed to just having air or free space between the electrodes. This capacitance is linearly related to the capacitance C_o in vacuum by K:

$$C = KC_o \tag{10.5.6}$$

K is always > 1 for dielectrics. In other words, when used in a quasi-static state, a piezoelectric behaves like a capacitor. This is very important for the modeling of piezoelectric actuator dynamics.

It is interesting to note that the coupling coefficient can be calculated using measured values of the dielectric constant. K_{free} is the dielectric constant of the material without mechanical stress applied. $K_{clamped}$ is the dielectric constant when the material is clamped so that it cannot deform any further. The relation for $\mathbf{k}$, K_{free}, and $K_{clamped}$ is

$$K_{free}(1 - \mathbf{k}^2) = \mathbf{K}_{clamped} \tag{10.5.7}$$

10.5.3 Effects of Stress and Temperature[50]

Many ceramic materials become piezoelectric (and ferroelectric) upon poling. After poling, the ceramic constantly tries to unpolarize itself. Aging of a piezoelectric material is the attempt by the material to unpolarize itself. The aging rate is a logarithmic function of time and is given in decades of time. When a stress is applied to the crystal, it becomes repolarized to a degree, and thus a new aging cycle begins. If aging is left unchecked, the dipoles will re-unalign themselves back into domains until the material is no longer piezoelectric. Since piezoelectrics are used to measure force or create forces, aging is not typically a factor in design considerations. Only when the piezoelectric will seldom (approximately once a year or so) be used is aging potentially a factor.

High stresses, typically greater than about 140 MPa (20 ksi), can alter piezoelectric constants. Continued high-stress cycles can cause permanent changes in these characteristics for some soft piezoelectric materials. Under high cyclic stresses, soft donor-doped ceramic piezoelectrics show serious degradation, including depoling and dielectric and mechanical losses. Unlike hard ceramics, which eventually stabilize to new levels, soft ceramics continue to degrade upon application of continuous high stress.

The *Curie* temperature of a ferroelectric crystal is the temperature at which the piezoelectric nonsymmetric structure returns to its nonpiezoelectric symmetric structure. This temperature is constant for a given material. When piezoelectric materials are used as precision actuators, one usually does not have to worry about the Curie temperature because (1) a ferroelectric crystal should generally not be used as an actuator, and (2) piezoelectric actuators do not generate appreciable amounts of heat, so a precision actuator should not get that hot.

10.5.4 Common Piezoelectric Materials and Shapes

There are many types of piezoelectric materials available for commercial use, including quartz, lithium niobate, tourmaline, and many types of ceramics. Quartz is a natural piezoelectric with a very low $\mathbf{d}$ constant ($\mathbf{d}_{11} = 2.3 \times 10^{-12}$ m/v) and is most commonly used for pressure- and acceleration-sensing devices. Remember that in general one should not use a ferroelectric material, which is also piezoelectric, in an actuator; but often in order to obtain the range of motion desired, one will have

[50] See, for example, H. Krueger and D. Berlincourt, "Effects of High Static Stress on the Piezoelectric Properties of Transducer Materials," Clevite Corporation Engineering Memo. 61–12, Cleveland, OH, May 1961, and H. Krueger and H. Helmut, "Stress Sensitivity of Piezoelectric Ceramics: Part 1: Sensitivity to Compressive Stress Parallel to the Polar Axis," J. Acoust. Soc. Am., Vol. 42, No. 3, 1967, pp. 636–645.

to use a ferroelectric material. High relative dielectric constants (>100) are characteristic of ferroelectric materials. Note that many ceramic materials are ferroelectric. Two of the most common ceramic piezoelectric materials are barium titanate and lead zirconate. Other ceramic piezoelectric materials are lead metaniobate and sodium bismuth titanate. There may be many variations of one type of material. For example, one manufacturer of ceramic piezoelectric crystals offers four variations of barium titanate and five variations of lead zirconate-lead titanate. For the four different versions of barium titanate, the $\mathbf{d}_{33}$ constant varies from 51×10^{-12} to 152×10^{-12} m/V. Until recently only high voltage (e.g., 1000 V) piezoelectric materials were available. Low voltage (e.g., 100 V) piezoelectric materials are now available. Low-voltage devices are more desirable, as it is difficult to obtain low-noise, high-voltage power supplies. In addition, high voltages are dangerous to work with. Most of the low voltage materials are also ferroelectric, and thus extra care must be taken when designing control systems for them.

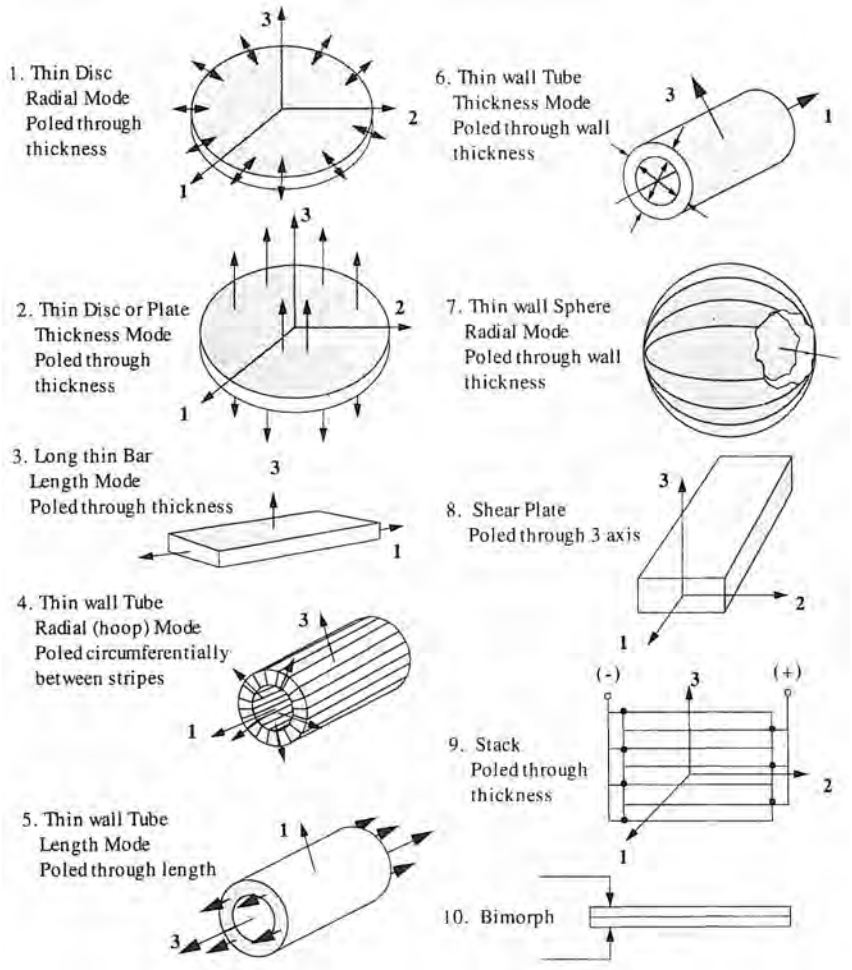


Figure 10.5.6 Common piezoelectric transducer shapes. (Courtesy of EDO Corporation)

Piezoelectric materials can be manufactured in a variety of shapes. Various available shapes of piezoelectric transducers are shown in Figure 10.5.6, and there is usually more than one mode of operation for each shape. Manufacturers usually do not readily have information on constants for unusual shapes.[51] Cylinders are usually manufactured with the polarization axis parallel to the length of the cylinder. Electrodes are placed at the ends of the cylinder and the cylinder expands or contracts in the longitudinal direction. A disk is usually manufactured with the polarization axis perpendicular to the face of the disk. The disk can operate in a thickness mode like a cylinder or in a planar mode where strain or stress occurs in a radial direction.

[51] It is usually best to try to design an actuator with the shapes available.

Plates can be manufactured in a variety of ways, yielding many different modes of operation. When the plate is made with the polarization axis perpendicular to the plate's face, an applied voltage can generate motion parallel to the length or width of the plate. These modes are called *transverse longitudinal and width modes*. It is also possible to generate shearing forces, but this is rarely required for actuators. For the $\mathbf{d}_{11}$ constant of a plate, note from Equation 10.5.1 that the change in thickness is constant regardless of the thickness of the plate. Hence to make an actuator with more than a few microns of motion, many plates must be stacked upon each other, as shown in Figure 10.5.6.

Tubes can move in longitudinal, radial width, and thickness extension modes. The axis of polarization is radial in these cases and the electrodes are on the inner and outer surfaces of the tube. A tube can also be sliced longitudinally. In this case the poling axis is in the θ-direction in cylindrical coordinates. This cut can operate in the parallel width mode, meaning that it twists, and in the transverse longitudinal or length expansion mode. Tubes themselves behave poorly as a radial expander, so if one wants a radial expander, a tube made from longitudinal sections, like a barrel, with electrodes between sections should be used.

Rings are very popular devices for actuators. Electrodes are placed on the inner and outer surfaces to be used in the transverse ring, transverse width, and thickness modes. The electrodes can also be placed on the top and bottom of the rings and have the ring operate in a thickness extension mode which allows a transmission bar to pass through the stack. Rings can also be fabricated, like cylinders, in sections. The longitudinal axis in this case is the 2-axis. The polarization axis is in the θ-direction. This shape operates in the parallel mode as a rotary motion actuator.

A piezoelectric actuator is typically made by epoxying together a stack of piezoelectric materials and electrodes. The stack must then be attached to a structure. Typically, the stack is bonded to a metal case which has bolt holes in it so that the actuator can be bolted in place. However, because of the many different modes of operation that a piezoelectric crystal can have, it is very important not to overconstrain the stack itself,[52] and thus many existing designs may not be suitable for angstrom-motion-level applications.

10.6 Fluid Power Systems[53]

Only hydraulic systems are considered here because the compressibility of air in general makes it unsuitable for precision positioning systems. Although one does not normally think of hydraulic power systems as being able to provide precise position control, this is perhaps more of a cultural bias associated with people most often seeing hydraulic systems on heavy construction equipment. In many precision design applications, hydraulic actuators are superior to other types because hydraulic actuators can be designed with zero friction and nearly backlash free transmission effect (e.g., metal bellows actuators). In this section, the following aspects of hydraulic power systems as they pertain to precision machine design will be discussed:

- Linear actuators
- Rotary actuators
- Poisson actuators
- Servovalves

A detailed discussion of the fundamental dynamic properties of these systems, which is necessary to model and predict performance accurately before being manufactured, is beyond the scope of this book. However, some basic rules of thumb for expected performance will be given along with references that should allow the design engineer to pursue the subject in-depth.

10.6.1 Linear Motion Hydraulic Actuators

The basic principle behind the operation of hydraulic cylinders is that the force produced is equal to the product of the net pressure of the hydraulic fluid and the area of the surface on which the fluid

[52] Conversations with Dr. Anthony Gee of Cranfield Institute of Technology, Cranfield, Bedford MK43 0AL, England.

[53] A general discussion of fluid power systems, although not focused on precision machine design, is given in Machine Design's annual Fluid Power Reference issue. A trade journal dedicated to fluid power which one may want to subscribe to if one is going to be doing a lot of design that involves fluid power systems is Hydraulics and Pneumatics, published by Penton Publishers. Other references include Hydraulic Handbook, 7th Ed., Trade & Technical Press, Surrey, England, 1979.

acts. The speed of the surface is determined by its size and the flow volume into the actuator. There are two fundamental types of linear motion hydraulic actuators: cylinders and bellows.

Cylinders

Cylinders have rods that slide in and out of a cylinder as fluid is supplied to the cylinder. There are a seemingly innumerable number of variations on this basic idea. In almost all cases, servovalves are used to control the flow of fluid in and out of the cylinder for precise positioning applications. Cylinders are generally used where large ranges of motion (up to many meters) and moderate resolution are required. With special low-friction Teflon® seals and hydrostatic bearings to support the piston and rod in the cylinder, enhanced performance and micron resolution can be obtained. The latter type of system, however, must often be custom designed.

Single-action cylinders provide force during the stroke in only one direction. The piston is returned to its starting position by allowing the hydraulic fluid to drain from the cylinder. This is usually done by having an external force, such as gravity or a spring, push the piston back to its starting position. Single-action cylinders are also often used in opposed pairs. The piston and rod may be replaced with a ram, which is a rod that is at or near full piston diameter. Large-diameter rams are more resistant to buckling when column loads are extremely high or when side loads on the rod are large. Ram cylinders are frequently used for large press applications and for jacking. Single-action cylinders may be used in servo applications where the weight of the object being moved is used to provide the return force. This is common in some large materials testing machines.

Double-action cylinders provide bidirectional motion control and are available with a seemingly infinite variety of mounting options. Double-action cylinders have a fluid chamber on each side of the piston that allows for differential pressure control across the piston; hence by controlling the differential pressure (or flow), the position of the rod can be controlled. The effective working area of the rod side of the piston is less than that of the other side unless a double-rod piston is used. Therefore, more force is applied from the fluid when the rod is extending than retracting, and rod speed is faster during extension. This makes the gain of the system dependent on direction of motion, a fact that is easily compensated for with a digital servo. Some double-action cylinders have rods extending out from each side of the piston; hence the working areas on both sides of the piston are equal. Equal areas allow for equal speed and equal forces in both directions.

Many variations of the conventional double-action cylinder may be found, including rodless cylinders. A cable is used in place of the rod and a pulley is attached to each end of the cylinder. The cable is connected to one side of the piston, fed around both pulleys, and connected again to the other side of the piston. The motion of object to be controlled runs parallel to the cylinder and axial space required for the cylinder is minimized. However, the stiffness of the cable is less than that of a solid rod, but the cable does act as a flexible coupling.

Telescoping cylinders provide a long stroke from a short body by using a telescoping tube to act as the piston and rod. The total stroke length may typically be as much as four times the length of the collapsed cylinder. Because the piston working area will vary with rod extension, force output also varies, with the highest force at the beginning of the stroke.

Rotating cylinders have special seals that allow an axial force to be applied to a rotating device. The most common application of this type are actuators for clamping chuck jaws or collets on lathes. It should be noted that seal friction from this type of actuator can be a primary source of heat in a spindle, even greater than the heat from the spindle bearings. It is often better to use a disk spring washer system to provide a constant force to a rotating member similar to the one shown for actuating curvic couplings in Figure 6.1.5. The force from the springs can be released by compressing the spring stack with a hydraulic or pneumatic piston which makes contact with the stack after rotation has stopped. Therefore, pistons are used as clamping devices only when precise force control, via pressure regulation, is required.

Metal Bellows Actuators

Metal bellows actuators are a common type of hydraulic actuator used for precision limited-range-of-motion servo applications.[54] Instead of a piston that slides in and out of a cylinder and requires a seal, a bellows uses a deformable structure to allow for motion of the piston without

[54] Note that pneumatically operated rubber bellows are also commonly used as actuators and vibration isolation mounts for precision equipment. See, for example, D. Grass, "Flexible Air Springs Do Multiple Duty," Mach. Des., April 7, 1988, or Firestone Industrial Products' Airstroke actuator literature.

requiring a seal. Bellows are generally used where limited range of motion (up to a millimeter or so) and high resolution (e.g., submicron) are required, all in a very compact space at the point where the output force must be applied.

A common bellows configuration for precision positioning is simple as shown in Figure 10.6.1. A small-diameter, large-length (master) bellows is connected by a fluid line to a large-diameter, short-length (slave) bellows, and the entire system is filled with fluid and hermetically sealed.[55] A linear actuator (e.g., an electric motor-driven leadscrew) compresses the small bellows and forces fluid into the large bellows. This system increases the resolution of the leadscrew drive by the ratio of the large bellows to the small bellows' cross-sectional areas. Because the system is hermetically sealed, there are none of the leakage problems often associated with hydraulic systems in general. The bellows system itself has no hysteresis, backlash, or friction problems; however, if a leadscrew is used to compress the master bellows, then all the problems associated with a leadscrew will be present, although they will be decreased by a factor equal to the bellows' area ratio.

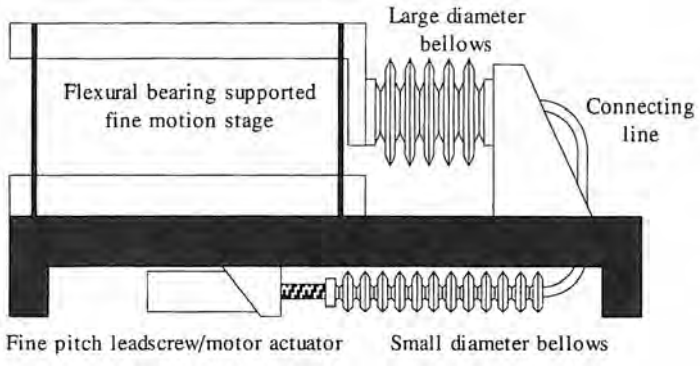


Figure 10.6.1 A *master-slave* bellows design for linear transmission reduction.

There are four basic types of bellows construction, as shown in Figure 10.6.2: flat plate, nesting ripple, single sweep, and torus. Each is formed from a series of *diaphragms*. The diaphragms are welded at the inside diameter to form a *convolution*. Numerous convolutions are then welded together at the outside diameter to form a *capsule*.

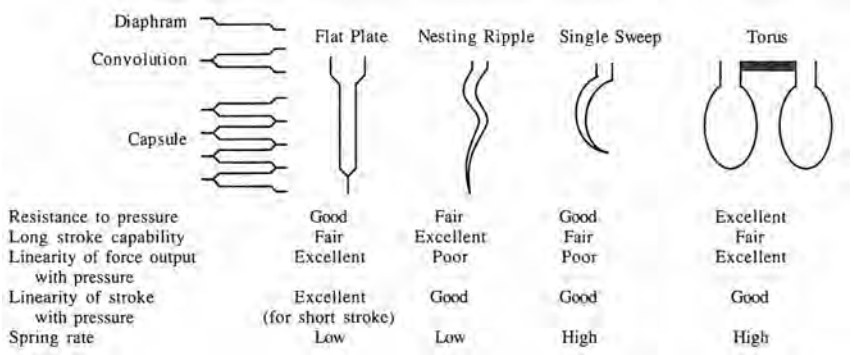
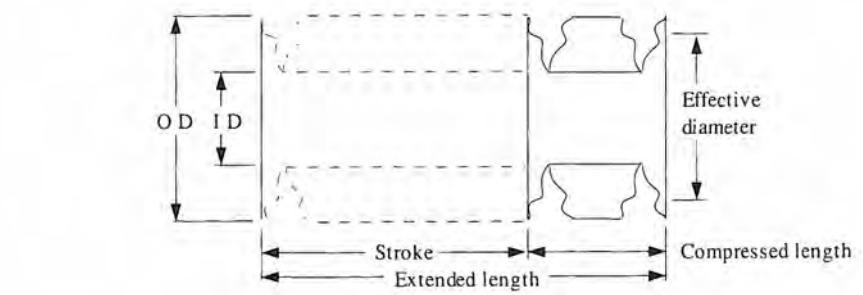


Figure 10.6.2 Types of bellows convolutions. (Courtesy of Metal Bellows Division, Sharon Parker Bertea Aerospace, Parker Hannifin Corporation.)

The following terminology is applicable to metal bellows:

[55] These types of systems are available as a made-to-order system directly from a metal bellows manufacturer.



OD Code	OD (mm)	ID (mm)	$A_{eff.}$ cm^2	$P_{max\ ext}$ kPa	Stroke/capsule mm	Ext. length mm	Comp. length mm	Spring rate N/mm
05	9.5	3.2	0.316	689	3.6	5.3	1.8	2.3
10	12.7	4.8	0.60	1034	8.4	11.7	3.3	9.6
20	19.0	6.4	1.26	345	7.6	9.9	2.3	4.2
30	26.2	14.0	3.16	207	13.5	16.8	3.3	4.4
35	38.1	24.6	7.68	276	7.4	10.9	3.6	3.9
40	41.4	19.0	7.10	207	7.9	10.9	3.0	2.1
50	48.0	35.3	13.61	310	21.8	26.7	4.8	2.6
55	57.2	38.1	17.74	345	12.7	18.3	5.6	3.7
60	64.8	44.4	23.42	345	18.0	24.6	6.6	4.7
70	75.9	50.8	31.55	276	23.9	29.7	5.8	5.1
80	101.3	68.3	56.52	276	25.4	31.8	6.4	8.8
85	108.0	81.3	70.97	310	20.3	28.7	8.4	13.1
90	126.2	101.6	101.87	345	20.3	29.2	8.9	13.1

Figure 10.6.3 Typical characteristics per capsule of commercially available metal bellows. (Courtesy of Metal Bellows Division, Sharon Parker Bertea Aerospace, Parker Hannifin Corporation.)

OD	Outside diameter of capsule.
ID	Inside diameter of capsule.
Span	Convolution width = (OD - ID)/2; OD/span should be less then 3.
P	Pitch (height) of a convolution.
NP	Nested pitch, compressed height of a convolution.
T	Diaphragm metal thickness.
FL	Free length of the bellows with no load applied.
N	Number of convolutions.
K	Spring rate of a bellows.
MD	Mean diameter = (OD + ID)/2.
EA	Effective area on which pressure acts to produce thrust:

$$EA = \frac{\pi\left(\frac{OD + ID}{2}\right)}{4} \tag{10.6.1}$$

Typically available metal bellows are shown in Figure 10.6.3. Conventional and high-vacuum end fittings are available. Take note that most bellows are custom manufactured: a single banana-size metal bellows may cost several thousand dollars in small lots.

10.6.2 Rotary Motion Hydraulic Actuators

There are numerous types of hydraulic rotary actuators that can provide continuous and limited range of motion. Of the many types of hydraulic rotary motors available, most have too much backlash, which makes them noncompetitive with respect to electric motors. Thus most common types of hydraulic motors (e.g., geroler, piston, etc.) are not discussed here. On the other hand, limited-range-of-motion hydraulic actuators are used in machine tools and robots and include vane actuators and rack and pinion actuators. Both can provide very high forces with reasonable motion resolution.

Vane Actuators

Vane actuators can have one or two vanes which are connected radially to the drive shaft. Hydraulic fluid enters the cylinder and creates a torque on the drive shaft. A stationary barrier limits the range of motion. Typical single-vane actuators can have up to 280 ^{o} of motion. Note, however, that there can be large radial forces on the rotating member. This contributes to the difficulty often encountered in controlling vane actuators. Double-vane actuators have twice the torque (because the working area is twice as great), can rotate only about 100 ^{o}, but do not have unbalanced radial forces.

Vane actuators are capable of supplying tremendous torques while requiring very little space; however, there are two main problems with their design:

1. It is difficult to obtain a good seal where the corners of the vane meet the corners of the housing so they have a fair amount of leakage compared to cylinders. This leakage can create a deadband when directions are reversed, which contributes to the difficulty in servocontrolling vane actuators.
2. Fluid pressure on one side of single-vane actuators produces large radial loads on the shaft which supports the vane. These potentially large radial loads, combined with a finite coefficient of friction for the actuator's rotary bearings, can create large starting torques under load.

Vane actuators are commonly used in heavy industrial equipment where it is impractical to generate rotary motion with a cylinder and a linkage. Vane actuators have also been used in hydraulic robots. They are not commonly found in precision machine tools other than as indexing devices because of difficulty in controlling their motion.

Rack-and-Pinion Rotary Actuators

A rack-and-pinion rotary actuator transforms linear motion into rotation motion. Fluid pressure moves a piston that is connected to a gear rack, which rotates a pinion and the output shaft. Sometimes the rack is machined directly into the piston rod. To eliminate high radial loads on the output shaft, one piston may be placed on opposite sides of the pinion. Standard rotation ranges of 90 ^{o}, 180 ^{o}, and 360 ^{o} are available. Output torques can range from the very small (N-m) to the very high (MN-m). If not for the age-old problem of gear backlash, this type of actuator may have found more use in precision rotary tables as a servoactuator, although backlash can be compensated for with two opposing cylinder driven pistons in contact with a single piston. Rack-and-pinion rotary actuators are commonly used on rotary transfer machines to index large rotary tables. However, the angular position of the table is usually established with the use of a hirth or curvic coupling.

10.6.3 Poisson Actuators

The Poisson effect described by Equations 7.3.3 gives rise to an interesting form of actuator that can also be classified as a monolithic cylinder. A Poisson actuator is an actuator that attains axial motion by stressing the shaft in an orthogonal direction. When high axial stiffness is required but a reasonable axial displacement and low operating pressure are also desired, the ID and OD of a steel column can be pressurized.[56] The radial stress will be equal to the pressure, and hence the axial strain will be equal to the pressure divided by Poisson's ratio.

Similarly, if a cylinder with closed ends is internally pressurized, then radial and axial strains will be generated. This effect can be utilized to make a short-stroke, (on the order of nanometers to microns), high-force actuator. A servovalve, or alternatively, a linear actuator (e.g., a ballscrew or voice coil actuator), can displace a diaphragm to control the pressure inside a cylinder and hence the Poisson displacement. This type of actuator can have many forms and it is often simple to machine the device itself into a monolithic flexure. After the bore is finished, it can be plugged with the pressure control device. The amount of axial displacement depends on the length of the cylinder and cross-sectional areas of the pressurized region and the cylinder wall. Note that the latter affects both the axial stress and strain and the circumferential stress and strain. The latter is linked to the former by the Poisson effect described by Equations 7.3.3.

[56] The author was first shown this type of actuator by Kevin Lindsey of NPL. It is in a sense the mechanical analog of a piezoelectric actuator.

10.6.4 Servovalves[57]

A servovalve is the fluid flow control device that is the heart of a precision electrohydraulic servo system. Electrohydraulic servo systems typically are used where electromechanical systems are impractical due to the power density or frequency requirements. There are many different types of servovalves, but this discussion will be limited to the two-stage type most commonly found in machine tool servos. Servovalves are very sensitive to manufacturing tolerances and contamination. The former are readily controlled, although at a price. A typical servovalve may cost $1500-$2000. Contamination is a very serious concern, as it can cause wear or even catastrophic failure; thus extreme cleanliness of the fluid is required.

A typical two-stage electrohydraulic servovalve operates as shown in Figure 10.6.4. Upon receiving a command signal, the torque motor rotates the flapper blade about the pivot point. When the nozzle is displaced between the orifices, it changes the resistance to flow out of the orifices and creates a differential pressure in the lines leading to each end of the spool. This differential pressure causes the spool to move axially, which shifts the flow between the ports. A feedback spring is moved by the spool and generates a torque on the flapper opposite to the one generated by the torque motor. Hence the flapper and spool soon reach an equilibrium position. This type of valve is called a *nozzle flapper valve*. A servovalve's closed-loop electrohydraulic system allows it to achieve very high bandwidths. A servovalve's dynamics can usually be custom tuned, during manufacturing, for the intended application. However, should one of the orifices become plugged with a dirt particle, the spool will move all the way over, directing flow out of one port only. This could lead to a dangerous situation should the valve be controlling a critical component such as the control surfaces on an airplane.

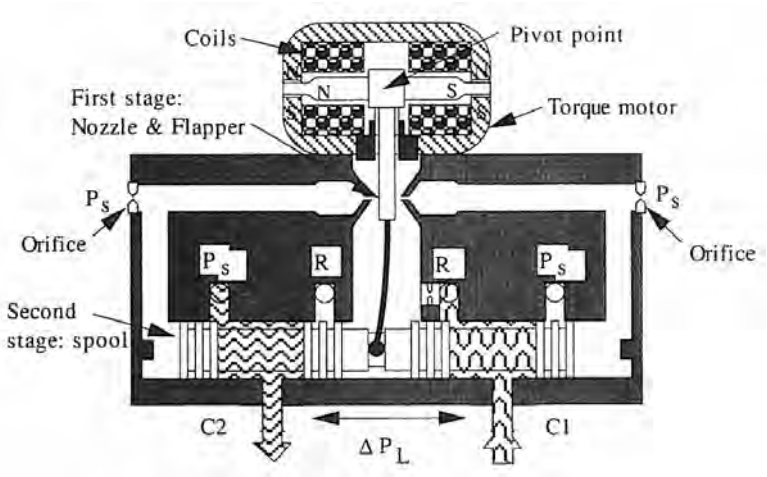


Figure 10.6.4 Schematic drawing of a two-stage nozzle flapper servovalve. (Courtesy of Atchley Controls Inc.)

To make servovalves less sensitive to this type of failure mode, the jet pipe servovalve was invented, which is shown in Figure 10.6.5. The principle of closed-loop operation of the two valves is the same, but the jet pipe valve creates a differential pressure on the spool by means of directing the flow into passageways. In the event of flow to the jet pipe nozzle becoming clogged, the spool will maintain its position or possibly return to the home position if current to the coils is shut off. Hence the failure mode is not catastrophic. Jet pipe valves are also less sensitive to erosion wear, as shown in Figure 10.6.6, but in general they do not have the dynamic bandwidth of nozzle flapper valves.

Both types of valves are usually interchangeable as far as bolt and port patterns are concerned. There are numerous suppliers of both types (and other types also), and often a supplier will sell both

[57] Technical references that provide detailed analysis methods for electrohydraulic servo-controlled systems include T. Viersma, Analysis, Synthesis and Design of Hydraulic Servosystems and Pipelines, Elsevier Science Publishers, Amsterdam, 1980; Blackburn et al., Fluid Power Control, MIT Press, Cambridge, MA; H. E. Merritt, Hydraulic Control Systems, John Wiley & Sons, New York; and J. Watton, Fluid Power Systems, Prentice Hall, Englewood Cliffs, NJ, 1989.

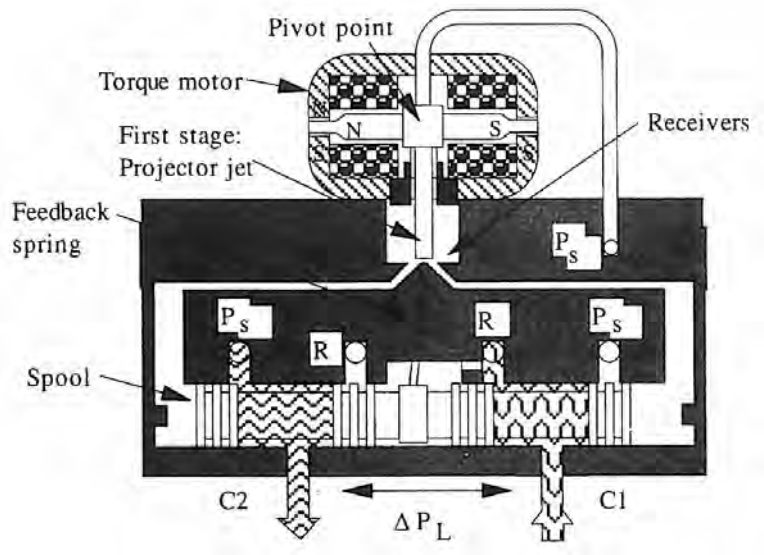


Figure 10.6.5 Schematic drawing of a jet pipe servovalve. (Courtesy of Atchley Controls Inc.)

types. A typical servovalve used in factory automation requires the space allocation of a cube 10 cm on a side. Pressure, return, and control ports are usually arranged in a diamond pattern on the bottom of the valve. A 0-dB frequency response to 50-100 Hz is not uncommon, and most manufacturers readily provide transfer functions and other dynamic modeling information for their valves. A servovalve should be sized to provide just enough flow to enable the actuator to move at the maximum desired speed. If the servovalve is too large, system resolution will be compromised because only a small portion of the control current the valve is able to receive will ever be utilized. The choice between the two types of valves should be made based on the following:

- If someone can get hurt if the valve fails, use a jet pipe valve.
- For extreme reliability, use a jet pipe valve.
- For maximum bandwidth, use a nozzle flapper valve.

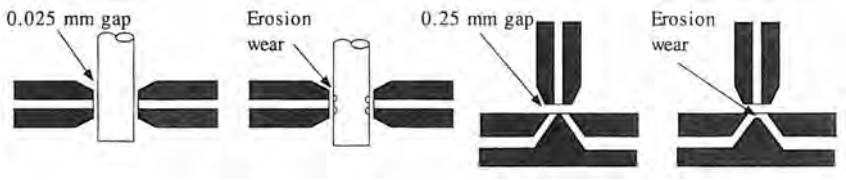


Figure 10.6.6 Comparison of wear in jet pipe and nozzle flapper servovalves. (Courtesy of Atchley Controls Inc.)

A servohydraulic cylinder system's components are sized in the following manner: First, the maximum force the system can exert at stall (no motion) is specified and the area of the cylinder is determined:

$$A_{cylinder} = \frac{F_{stall}}{P_{supply}} \tag{10.6.2}$$

Given the maximum velocity at load, the loaded flow rate and load pressure drop are determined:

$$Q_{load} = A_{cylinder} V_{load} \tag{10.6.3}$$

$$P_{load} = \frac{F_{load}}{A_{cylinder}} \tag{10.6.4}$$

From these values the no-load flow rate is computed[58]:

$$Q_{\text{no load}} = Q_{\text{load}} \sqrt{\frac{P_{\text{supply}}}{P_{\text{supply}} - P_{\text{load}}}} \tag{10.6.5}$$

The rated flow for the servovalve at the rated pressure drop across the valve is given by

$$Q_{\text{rated flow}} = Q_{\text{no load}} \sqrt{\frac{P_{\text{rated pressure drop}}}{P_{\text{supply}}}} \tag{10.6.6}$$

The next step is to select a servovalve with the rated flow. Typically, the valve is oversized by 10% to make sure that it can provide enough flow. Commonly available servovalves have rated flows of 4-60 L/min (1-15 gpm) at 6.9 MPa (1000 psi) pressure drop across the valve.

For ensured good dynamic response, the valve's frequency at its 90 $^{\text{o}}$ phase lag point should be three times the load natural frequency. To calculate the load natural frequency, the system mass and stiffness must be determined. For a well-designed precision machine, the mass will be dominated by the mass of the carriage and the stiffness dominated by the compressibility of the fluid in the cylinder and the line. Note that for a high-performance system the servovalve is always placed near the cylinder, without a flexible hose between the two. Steel tubing from the valve manifold to the cylinder can be considered rigid. The displacement of the piston in the cylinder due to an applied force compressing the fluid of bulk modulus β in the cylinder will be

$$\delta_{\text{cylinder}} = \frac{FL_{\text{cylinder}}}{A_{\text{cylinder}}\beta} \tag{10.6.7a}$$

Hence the stiffness of the fluid column will be:

$$K_{\text{cylinder}} = \frac{A_{\text{cylinder}}\beta}{L_{\text{cylinder}}} \tag{10.6.7b}$$

For the fluid in the line a similar calculation can be made, with the note that the change in volume in the line must equal the change in volume of the cylinder; hence the stiffness of the fluid in the line as it acts on the piston in the cylinder will be

$$K_{\text{line}} = \frac{A_{\text{cylinder}}\beta}{L_{\text{line}}} \tag{10.6.8}$$

The cylinder and line stiffnesses add in series, and hence with the system mass M, yield a system natural frequency in hertz of

$$f_n = \frac{1}{2\pi} \sqrt{\frac{A_{\text{cylinder}}\beta}{(L_{\text{line}} + L_{\text{cylinder}})\ M}} \tag{10.6.9}$$

This frequency should be evaluated for the extremes in line and cylinder length.

These equations suffice for a first-order sizing and selection of system components. In order to determine the dynamic characteristics of the closed-loop mechanical system actuated by an electrohydraulic system, many other factors must be considered which are beyond the scope of this book. The references cited earlier, as well as literature available from most reputable servovalve manufacturers, will enable the design engineer to assemble the desired dynamic model.

10.6.5 Support Equipment

Hydraulic systems require specialized hardware and support equipment, including fluid lines, filters, flow control devices, seals, and pumps.

[58] From Moog Inc. servovalve selection literature.

Fluid Lines[59]

Fluid must be delivered to the actuator, and this can be accomplished using hoses, steel tube, or manifolds. Hoses provide an economical flexible conduit, but their elasticity decreases system stiffness slightly. Improperly routed, hydraulic hoses can quickly fatigue, look bad, and create snag problems; thus care must be taken[60] when routing hydraulic lines, which often includes providing attachment points for the lines to the structure.

The fluid pressure will affect the stiffness of a hydraulic hose; in fact, hose forces can be one of the biggest force inputs to an ultraprecision machine tool carriage. Hose forces can be relieved by using a balanced design (i.e., two hoses mounted so that their forces cancel) or a linear hydraulic commutator. There are two main designs of hydraulic commutators: (1) those that use a long tube, with a radial hole in the middle, that passes through the carriage, or (2) a pressurized groove on the bearing surface.[61] For the tube to be effective and not impart any friction forces onto the carriage, it must be supported with respect to the carriage by hydrostatic bearings. PTFE sliding bearings will not work for this application; the friction forces will be greater than the hose forces will be. The pressurized groove design will work well, but often the leakage flow will be very high, which leads to too much pumping power and heat generation.

Filters and System Cleanliness

Perhaps the most critical elements in a fluid power system are the filters. Some manufacturers routinely suggest using fine filters (those which remove 98.7% of particles 3 μm or larger). However, if the application does not warrant the use of such a fine filter, a coarser filter can actually give better performance in terms of reduced maintenance costs. Choosing a filter for the system you design is like choosing any other component where you rely on the vender's expertise. You must check with several vendors and then choose the system type that at least two venders agree on.[62] Small in-line screen filters are available[63] and can be used as a last-chance filter before a critical component.

Whenever a component is machined, there are traces of cutting fluid and chips left behind. If the component is to be used in a hydraulic system, it must be thoroughly cleaned. Otherwise, when the system is turned on, contaminants can be forced into valves, flow restrictors, and the like, and cause early failure. This is particularly true for hydrostatic bearing flow restrictors. To clean components, one cannot merely use a cotton swab and some alcohol. A pickling process should be used to dissolve most particles and then the system flushed with high-pressure hydraulic oil before critical components are installed. Note that pickling will not change component dimensions, at least on the micron level, but it does so thoroughly clean the component that the surface is left very reactive. Hence afterwards, the component must be thoroughly protected with grease or oil. Note that it is advisable, if possible, to avoid the use of a thin-plated layer (e.g., chrome or nickel) on components subject to high-pressure impinging flow. High-velocity fluid may cause local erosion of the plating layer, particularly in corners. The eroded plating material can then clog sensitive hydraulic components.

Flow Control Components

When designing a fluid power system, often there is a need for small components such as plugs, last-chance filter screens, orifices, flow restrictors, and so on.[64] Perhaps the most commonly used element is the plug, which allows one to turn a block of metal into a block of Swiss cheese for internal porting and then easily plug hole openings. Care should be taken, however, to use plugs with straight threads and O ring or bonded seals for holes that have components in them which may need maintenance. There are a number of suppliers of these types of components and all design engineers should take it upon themselves to build up and maintain an appropriate catalog file. One of the best places to do this, of course, is at the annual fluid power design show that is advertised in most trade journals.

---

[59] It is recommended that any designer of hydraulic systems that are to be used on submicron-accuracy machinery be concerned with hydraulic line dynamics and methods for filtering out noise. An excellent reference on this subject is T. Viersma, Analysis, Synthesis, and Design of Hydraulic Servosystems and Pipelines, Elsevier Science Publishers, Amsterdam, 1980.

[60] See, for example, P. Lee, "Routing High-Pressure Hose," Mach. Des., March 24, 1988.

[61] See U.S. Patent 4,865,465 by Sugita et al. of Citizen Watch Co.

[62] See J. Drennen "Shooting Holes in Filtration Myths," Mach. Des., Feb. 11, 1988. Note that some companies have computer programs that model typical hydraulic systems and aid in determining the best type of filter system to use.

[63] See Technical Hydraulic Handbook a catalog available from Lee Co., Westbrook CT 06498-0424, (203) 399-6281.

[64] See, for example, K. Korane, "Small Hydraulics Solves Big Problems," Mach. Des., July 7, 1988.

Seals[65]

Seals are a most vital part of any hydraulic actuator because they keep fluid loss to a minimum and keep dirt out. Not surprisingly, one of the most important factors affecting the obtainable resolution from a hydraulic piston is the type of seal used. Most seals depend on the oil forcing the seal material against the piston bore to form a seal (self-help). Thus friction increases with pressure. For servo applications, looser-fitting Teflon seals are available, and pistons that have these seals are often referred to as hydraulic servo cylinders. The decrease in friction results in a decrease in sealing efficiency. When pressure is reversed on the piston in order to reverse direction, as is often the case when servo positioning, the leakage has to stop before the direction changes. The result is a deadband which contributes to limit cycling. Diaphragms may also be placed in with the piston for applications that require low friction, no leakage across the piston, or extremely sensitive response to small pressure variations. In the food and drug industry diaphragms are frequently used with pneumatic actuators because they require no lubrication and do not exhaust a contaminating oil mist.

Hydraulic Pumps

There are numerous types of hydraulic pumps and for most applications, the pump type should be chosen to meet the pressure and flow requirements with the maximum efficiency. Regardless of the pump type chosen for servo systems, if possible it should be variable stroke so that it works only as hard as needed to supply the fluid the system needs. Otherwise, unwanted heat will be generated as excess fluid is vented back to the tank. With some small precision gear pumps used for hydrostatic bearings this is not possible, but the flow required by the bearing should be relatively constant. Furthermore, because pumps can be mechanically noisy (to some degree), they should be isolated from the machine and an accumulator should be placed as near as possible to the pump. Other vibrations in the line from the pump to the machine can be filtered out using notch filters.[66]

10.7 Rotary Power Transmission Elements

Rotary power sources (e.g., electric motors) are commonly used in many types of machines. One generally finds that when purchasing a motor with a given power level, it is less expensive to buy a low-torque, high-speed motor than a high-torque, low-speed motor. Unfortunately, many applications require slow, steady, controlled motion; thus methods are required to reduce the speed and increase the torque of motors. In everyday nonprecision applications, this is usually accomplished without trouble. However, for precision applications, there are numerous factors which contribute to the degradation of accuracy and controllability of a rotary power transmission system:

- Form error in the device components
- Component misalignment
- Hysteresis
- Backlash
- Friction

Each of these factors can itself be manifested in many forms in many different components in the device. One way to help minimize these effects is to specify simple designs that employ easy to manufacture and assemble components.

Component form error and misalignment create nonlinearities in motion. Form error also causes preload variations in preloaded mechanical speed reduction devices. These variations create small drive torque fluctuations which cause the input motor speed to vary slightly and hence vary the output speed. Speed variations due to form error in preloaded systems can be on the order of 0.1-1.0%. Hand-lapped systems can of course perform much better. The amount of correction that a closed-loop servo system can provide depends on the actuator, controller, and power supply. For rotary tables used to control motion while cutting contours, finish cuts are usually made at slow speed, so most controllers are able to maintain proper contour accuracy. Form error speed variation problems occur to some extent in any preloaded mechanical speed reduction device.

Hysteresis results from the nonlinear behavior of contact between curved surfaces. When a load is reversed, the deflection changes in a nonlinear manner which results in nonlinear system

[65] Sealing systems were discussed in Section 7.6.6.

[66] T. Viersma, Analysis, Synthesis, and Design of Hydraulic Servosystems and Pipelines, Elsevier Science Publishers, Amsterdam, 1980.

response. Often, hysteresis effects can be accounted for by the nonlinear load-deflection characteristics predicted by Hertz theory.

Backlash results from gaps between components and subsequent temporary lack of motion upon torque reversal. For many constant-speed applications this is not a problem, but backlash can cause servoed systems to limit cycle. Many systems, such as those which contain gears, require some backlash to allow for manufacturing and assembly errors in components. Other systems that use flexible components, such as belt-drive systems, can be made without backlash.

Friction is present in all systems with mechanical contact between moving components and creates thermal and control problems. Heat sources are almost always a problem for precision machines, and thus drives with the highest efficiency should be sought. Controllability also requires low static friction. Most companies are aware of these issues and there are many high efficiency drive systems available.

In most rotary motion servo-system designs, it is best to minimize or eliminate backlash (as discussed for various systems below) and mount the rotation sensor on the output shaft. It used to be that one would use the transmission system as a means of increasing the resolution of sensors. For modern high-accuracy systems, it is best to mount a high-accuracy sensor directly to the output shaft and use good-quality preloaded zero backlash components so the servo system will not limit cycle. This is generally a less expensive and more reliable design than one that uses a moderate accuracy sensor and extra, super-quality transmission components which are bound to wear anyway.

There are numerous types of rotary power transmission systems available, but only those predominantly used in precision machines (or with the potential for use) are discussed herein, including gears, modular speed reducers, belts and chains, cams, and couplings. For each of these device types there are immense volumes of catalogs and design information and this section is only intended to provide an introductory exposure. For greater details, the reader is referred to typical references.[67]. Clutches and brakes are also a vital part of many systems but are discussed in detail in most machine element design texts.

10.7.1 Gears[68]

There are a seemingly infinite variety of gear types that have been developed for seemingly innumerable purposes. Deviations from the basic spur and worm gear type have been developed to minimize noise, maximize power transmission, minimize cogging torque, increase load-carrying capability, and transfer power from oddly intersecting shafts. However, as the gear shape becomes more complex (e.g., helical gears), it becomes more difficult to manufacture and produce with the precision required for high-accuracy machines. Hence herein only spur and worm gears will be discussed, the former for their simplicity and the latter for their high-reduction-ratio capability. The references cited provide a complete and thorough discussion of other types of gears available and gear strength and life calculation methods.

Gear catalogs will also usually provide all the information necessary to complete a design (e.g.. torque, speed, and backlash limits). Stiffness data is usually not provided and if not available from the manufacturer, it can be easily estimated using bending and shear deformation analysis methods. Gears for precision machines are often sized based on tooth and shaft stiffness criteria. When stiffness criteria for a machine tool are met, the stress criteria are also usually met; hence for the discussion of gears and gear systems, typical catalog data are not provided in this book. With the plethora of different gear manufacturers, the precision machine design engineer will rarely have to design the details of the gears he or she may require. The design engineer must only specify the size of the gear and the number of teeth. If the required gears cannot be found in a catalog somewhere, a gear manufacturer somewhere is usually more than happy to provide the necessary design assistance.

[67] Two good general-purpose references that provide general information and extensive vendor information are the Thomas Register of American Manufacturers and Machine Design Annual Mechanical Drives Reference Issue (Penton Publ.)

[68] There are innumerable references available for the design and application of all types of gears. A good general-purpose reference is Machinery's Handbook, published by Industrial Press. Other references include, for example, F. Jones, Gear Design Simplified, Industrial Press, New York; D. Dudley, Handbook of Practical Gear Design, McGraw-Hill Book Co., New York; M.F. Spotts, Design of Machine Elements, 6th Ed., Prentice Hall, Englewood Cliffs, NJ, 1985; and J. Shigley and C. Mischke, Standard Handbook of Machine Design., McGraw-Hill Book Co, New York.

Thus this section will focus on what other references do not tend to provide, which is how to estimate gear accuracy for precision machine applications.

The transmission ratio between a single set of gears is just the ratio of the gears' pitch diameters. The teeth are designed such that as the gears rotate, there is rolling contact between the teeth. The shape of the tooth that allows for this condition is called an *involute*. The pressure angle is the angle of the contact force between the the gear teeth relative to the tangent to the pitch circles (angle of inclination of the teeth in a rack). The larger the pressure angle, the greater the thickness of the tooth at its base, and hence the stronger and stiffer the gear teeth are. However, the greater the pressure angle, the larger the reaction forces that tend to displace the gears. For precision servo-controlled cutting applications, static stiffness is often of prime concern, so a large pressure angle is desired with a large shaft and support bearings to withstand the larger radial forces. For measuring applications where smoothness of motion is desired, fine teeth (a small pressure angle) are desired. The quality of a gear is ascertained by its AGMA number.[69] The higher the number, the better the gear and the higher the cost.

Nonlinear motion is caused by errors in the circular pitch and tooth or arc thickness. Variations in circular pitch or tooth thickness create the effect of a sinusoidal variation on the gear ratio superimposed on the nominal gear ratio. Both also are a source of backlash. Small gear ratio variations are not too difficult for a robust control system to compensate for, but backlash can cause limit cycling in the control system.

In many applications, a certain amount of backlash is desired to allow for manufacturing errors, deflection under load, and thermal expansion. For precision machines and instruments where loads are low and thermal control is good, essentially zero backlash can be specified, although it is difficult to achieve. In any case, it is vital to ensure that the gears are not forced to mesh or else rapid wear will result. Backlash is affected by:

- Tooth thickness error
- Tooth profile error
- Deflection under load
- Tooth wear
- Center distance error
- Pitch line radial error motion
- Gear axis parallelism
- Thermal expansion

The first four factors on the left affect primarily the circumferential position of the contact point. The latter four factors affect primarily the effective center distance.

Circumferential accuracy of gear dimensions is affected primarily by how the gear is manufactured. When extreme accuracy and minimal backlash are a must, the gears can be ground. In order to grind gear teeth, a method is needed for indexing the gear as each tooth is ground. The angular error $\varepsilon_{\text{gear}}$ in a gear is a function of the Abbe error in tooth location of the gear on the index table $r_{\text{gear}}\varepsilon_{\text{index table}}$, the grinding process error δ_{grind}, and the gear radius r_{gear}:

$$\varepsilon_{\text{gear}} = \frac{r_{\text{gear}}\ \varepsilon_{\text{index table}} + \delta_{\text{grind}}}{r_{\text{gear}}} \tag{10.7.1}$$

Precision indexing tables are commonly available that are accurate to 1 arcsecond (4.8 μrad), and 0.1 arcsecond tables are available. Figure 10.7.1 illustrates the angular error (e.g., backlash) that a gear is likely to experience as a result of manufacturing errors. This error is for each gear in a gear set and is cumulative. Note that gears can be specially ground or lapped to achieve a few arcseconds of accuracy, but the cost can be very high.

An error in center distance location between gears will also affect their accuracy, predominantly by creating backlash. Consider two parallel lines which are at the pressure angle ϕ with respect to a vertical line that connects two gears' centers. As the gear centers move apart by an error δ_{center}, the distance δ_{tooth} the center distance moves in a direction tangent to its contact point increases. An approximation for the resulting angular error in the gear set $\varepsilon_{\text{gear set}}$ is just δ_{tooth} divided by the radius of the output gear r_{gear}, which fortunately is usually that of the larger gear:

$$\varepsilon_{\text{gear}} = \frac{r_{\text{center}} \tan\phi}{r_{\text{gear}}} \tag{10.7.2}$$

[69] See American Gear Manufacturers' Association, Gear Classification Manual 390-02.

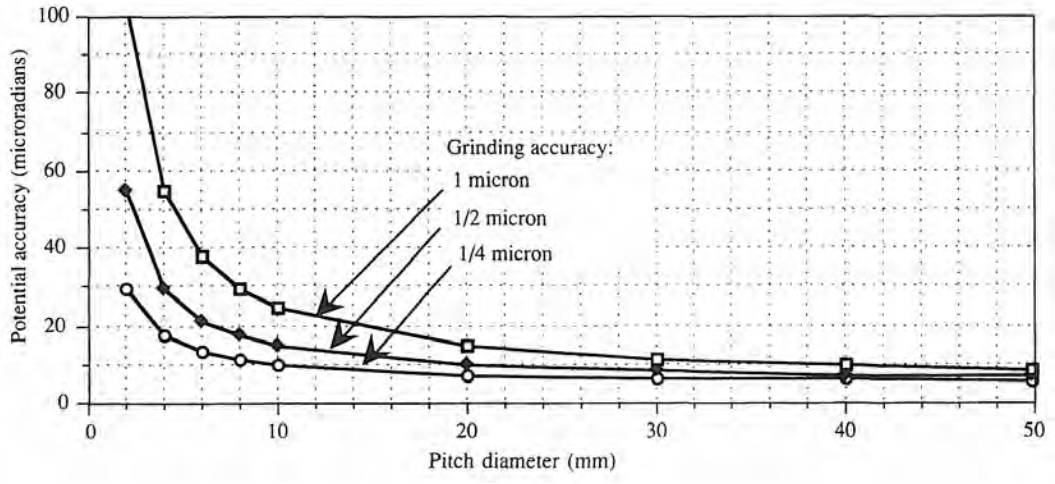


Figure 10.7.1 Potential accuracy of gears ground on an index table with an accuracy of 1 arcsecond.

For the greatest insensitivity to center distance errors, one can see that a small pressure angle is desired. Figure 10.7.2 shows the effect of center distance mounting errors on gear backlash for various common pressure angles and a 10 μm center distance error.

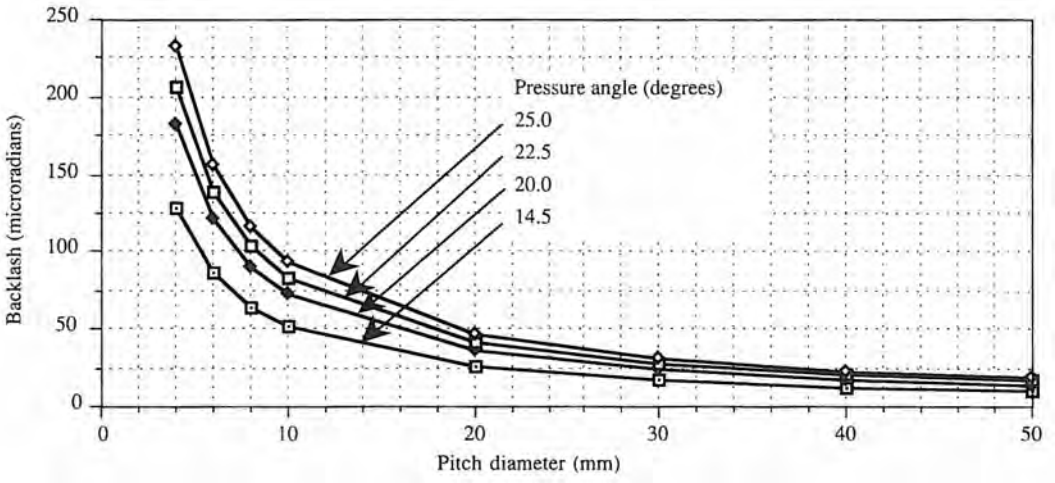


Figure 10.7.2 Effect of a 10 μm center distance error on gear backlash.

If a machine's rotary axis is servocontrolled, one might surmise that gear accuracy is not as important as eliminating backlash, because it is the latter which causes limit cycling. To minimize backlash without the expense of grinding gears to very close tolerances there are two methods that can be used: (1) use a constant force or torque spring or hanging weight to keep the torque on the gears always acting in one direction, and/or (2) use antibacklash gears.

Antibacklash gears are made by taking one gear in a pair and making it from at least two gears. These two gears are rotated relative to each other using a setscrew or constant-torque spring so that one gear transmits torque in one direction and the other gear transmits torque in the other direction. Alternatively, two drive motors can be used with two gears to drive one large gear. The preload torque between the two gears ensures that tooth contact is maintained regardless of the direction of rotation. This is an economic, effective way of virtually eliminating backlash, but it does increase the wear on the gears. This method can induce a moment on the gear support shaft if the antibacklash gear is composed of only two gears. Also, it takes some skill to adjust antibacklash

gears whose preload is fixed by a setscrew. Antibacklash gears preloaded by a constant-force spring are easier to install but still have a small amount of hysteresis in their response upon torque reversal.

Worm gears have a more complicated tooth form than spur gears, but they can achieve much higher transmission ratios with a single set of gears. Spur gears' transmission ratio is a function of relative gear diameters. Worm gears' transmission ratio, on the other hand, is a function of the pitch of the worm gear and the diameter of the driven gear. As the worm gear rotates, its thread helix continuously circumferentially displaces the teeth on the driven gear. Given the lead l_{worm} of the worm and the diameter D of the driven gear, the transmission ratio of a single worm gear set is just $TR = D/2l_{worm}$. The same type of accuracy and backlash considerations apply to worm gears as to spur gears.

With worm gears sliding contact exists between the teeth, so friction and wear levels are higher than for spur gears; however, the simplicity of a worm gear system makes it ideal for precision machines where high transmission ratios are desired, servo-controlled angular position rates are generally slow, and the worm gear assembly can be submerged in oil. Worm gears are commonly used to control the position of rotary index tables and rotary servo-controlled axes on machining centers.

10.7.2 Modular Speed Reducers

There are many different types of speed reducers that can be purchased in a modular bolt-on form. Most are intended for industrial applications where if the torque ever reverses, backlash is not very critical; hence only speed reducers which are more commonly found in precision machines and robots will be discussed here, including planetary speed reducers, harmonic speed reducers, cycloidal drives, traction roller drives, and wire capstan drives. Spur and worm gears as discussed above are also commonly used in modular speed reducers.

Many drives come equipped with an integral servomotor and resolver. Extreme caution should be used when purchasing such a drive system for a position control application. Often the servomotor, resolver, and controller are meant for speed control only. One should not take a salesman's word that he sees no reason why his system won't work in a new application. Salespeople are not necessarily design engineers, and they often do not understand the subtle differences that make a system work in one situation and fail in another. If possible, one should arrange to bench test a unit before a particular unit is specified unless it has a history of success in similar applications. A reputable company should not object to this request. *Caveat emptor!*

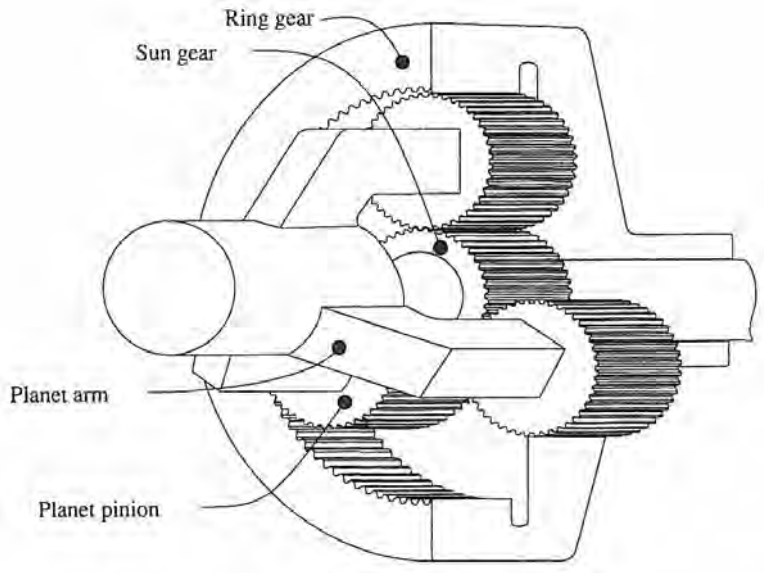


Figure 10.7.3 Main components of a planetary speed reducer designed for cascaded series operation.

Planetary Speed Reducers

There are two common arrangements for planetary or epicyclic speed reducer that a machine tool design engineer is likely to encounter.[70] The first is illustrated in Figure 10.7.3. The ring gear is stationary and the motor drives the sun gear. As the sun gear rotates, the small planet gears on the planet arm rotate as they roll around meshing with the ring and sun gears. The distance the planet gears roll on the inside of the ring gear is the product of the rotation angle of the planet arm and the pitch diameter of the ring gear D_{ring}. The planets also mesh with the sun gear; however, the sun gear has a smaller pitch diameter D_{sun} than the ring gear. In order to prevent the planet gears from not rolling due to geometric overconstraint (the planet gears cannot roll on two different stationary diameters at once), the amount the planet carrier gear rotates must equal the difference in distances the planet gears would rotate if they only made contact with the ring and sun gears, respectively. The transmission ratio of the gear train is the ratio of input to output rotation angles (sun gear to planet arm rotation angle):

$$TR = \frac{D_{sun}}{D_{ring} - D_{sun}} \tag{10.7.3}$$

This type of transmission is among the most compact available, since the ring gear can be broached on the inside of a tube and the sun gear used to drive a planet arm which has a sun gear on it, which in turn drives another planet arm, and so on. In this manner it is easy to build up a series gearbox with a tremendously high gear ratio. In addition, the sun gear is making contact with several teeth; hence the contact ratio is much higher than for a conventional mullti-stage gearbox, and thus planetary speed reducers can carry very large loads. Although economical to manufacture, this type of series gear train can be noisy at high speeds because the radial position of inner planet assemblies is usually not fixed with bearings. It often must rely to some extent on the meshing of the three planets with the ring gear to centralize the planet arm.

Another type of planetary or epicyclic gear train is illustrated in Figure 10.7.4 and is sometimes referred to as a *perpetual wedge*. In this design, gears 1 and 3 are attached to a common shaft that is supported in the planet arm by bearings. Usually three planet arms exist and are attached to the input shaft. Gear 1 meshes with the fixed gear 2, and gear 3 meshes with gear 4, which is attached to the output shaft. Tracing through the amounts each gear rotates and travels circumferentially, it is not difficult to show that the ratio of the input to output rotation is

$$TR = \frac{D_1 D_4}{D_1 D_4 - D_2 D_3} \tag{10.7.4}$$

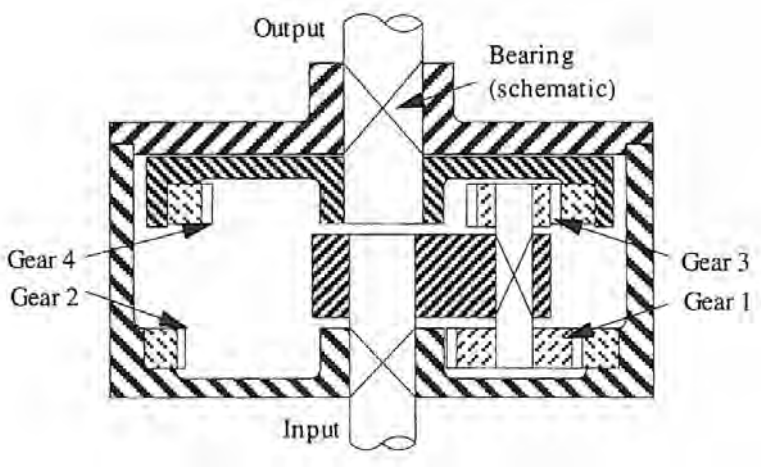


Figure 10.7.4 Single-stage high-transmission-ratio planetary speed reducer.

It is not difficult to make the product of D_1 and D_4 approach that of D_2 and D_3, thereby obtaining a very high transmission ratio with a single stage. All the gears can be supported by

[70] There are actually 12 different types of planetary gear trains. These are shown schematically and their transmission ratios tabulated in J. Shigley and C. Mischke, Standard Handbook of Machine Design, McGraw-Hill Book Co, New York.

precision bearings and the input and output shafts' bearing bores can be line bored. It is also possible to use antibacklash gears for the planets. These factors tend to minimize the amount of backlash one is likely to encounter and makes this type of planetary speed reducer acceptable for use in some precision machines. However, note that this design creates very high force and velocity at the mesh which can easily make the efficiency fall below 50%. It is left as an exercise to the reader to determine what the backlash of a planetary speed reducer is, given the backlash in each individual gear.

Harmonic Speed Reducer

A harmonic speed reducer operates in much the same way as a planetary gear speed reducer does, except that the components are different as shown in Figure 10.7.5. The device still has a large fixed ring gear with internal teeth, but the three planets and output ring gear are typically replaced with two cam rollers and a flexible external tooth spline (flex-spline), respectively. The flexible spline is attached to the output shaft through a rigid back plate. As the input shaft rotates the wave generator, the flexible spline orbits within the ring gear while slowly rotating in the opposite direction as the teeth mesh. In a sense this is a device like a planetary gearbox shown in Figure 10.7.4, the transmission ratio is derived from Equation 10.7.4 by setting D1 = D3, and where N_{ring} is the number of teeth on the ring gear and N_{spline} is the number of teeth on the spline:

$$TR = \frac{N_{spline}}{N_{ring} - N_{spline}} \tag{10.7.5}$$

In order to achieve high ratios the gear teeth must be very small which limits their stiffness and strength. There may be 202 teeth on the ring and 200 teeth on the spline, which gives a transmission ratio of 100:1. Note that harmonic drives subject to heavy overloads have been known to strip their gear teeth.

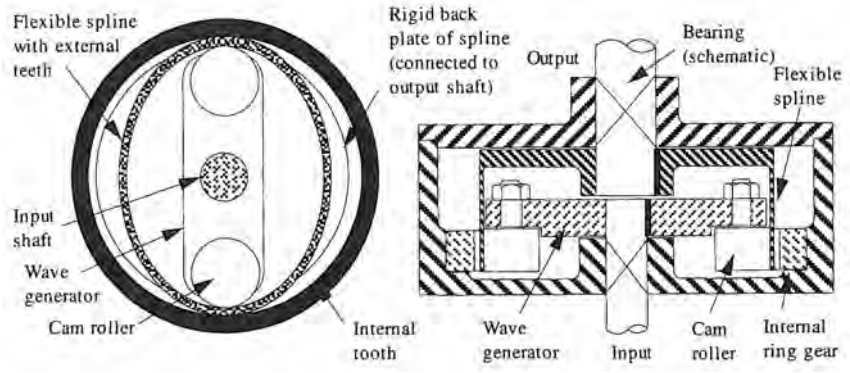


Figure 10.7.5 Harmonic speed reducer (teeth are too small to show).

Harmonic speed reducers can be made very compact and lightweight and thus have been popular with robot manufacturers and in other applications where weight is critical. For most precision machine tools weight is not a factor; one is probably better off using a system that is more able to withstand high loads and abuse such as the cycloidal speed reducer described below.

Cycloidal Speed Reducer

A cycloidal drive (or an epitrochoidal drive) operates on a similar principle to that of the planetary gear train shown in Figure 10.7.4. As shown in Figure 10.7.6, a cycloidal drive uses cam rollers on a fixed housing instead of a ring gear, a dual trochoidal-shaped cam instead of a planetary system, and cam rollers on the output housing also instead of a ring gear. The cam is made to orbit inside the input and output housing by an eccentric cam attached to the input shaft.

There is one less lobe on each track of the epitrochoidal-shaped[71] cam than on the input and output housings, respectively; thus as the cam orbits, it also rotates. The shape of the dual track cam allows it to be in contact with all rollers at all times, with each roller contacting the cam at a different

[71] A sinusoid superimposed on the circumference of a circle.

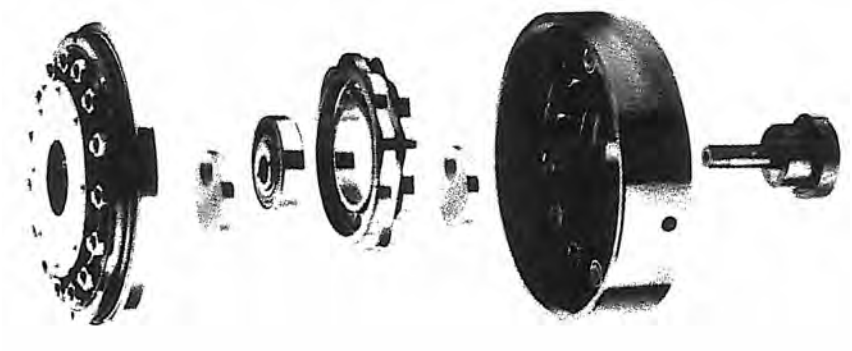


Figure 10.7.6 Trochoidal speed reducer. (Courtesy of Dojen Div., Lenze USA, L.P.)

point on the cam profile. As a result, the device has tremendous stiffness and overload capacity. The transmission ratio for this type of drive is

$$\mathrm{TR} = \frac{(\mathrm{N_{input}} - 1)\ \mathrm{N_{output}}}{\mathrm{N_{input}} - \mathrm{N_{output}}} \tag{10.7.6}$$

where N refers to the number of followers on the input or output. Typically, there may be 11 input rollers and 10 output rollers, which gives a transmission ratio of 100:1. Ratios from 10:1 to 225:1 are commonly available. Although this type of drive is physically much more complicated than a worm gear drive, it can achieve greater transmission ratios with higher efficiency and is thus becoming more commonly used. Figure 10.7.7 shows characteristic parameters of this type of drive. There are many variations on this type of drive technology and many different manufacturers exist.[72]

Traction Drives

One of the biggest problems with gears is the difficulty in manufacturing the exact tooth shape desired. When loads are low and the gears are needed primarily for the purposes of power transmission and transmission reduction, one may wish to consider the use of a traction drive.[73] A traction drive replaces gears with round rollers that are preloaded against each other. If the output shaft is overloaded, the rollers merely slip. The load capacity of a traction drive is readily calculated from the geometry of the system, the preload, and the coefficient of friction. The latter is typically assumed to be 0.1. The stiffness is dependent primarily on the stiffness of the system's input and output shafts. Calculation of the effect of the deformation at the contact interface on the torsional stiffness can be accomplished using the equations in Section 5.6.

To enhance tractive effort while providing lubrication, tractive fluids have been developed.[74] These fluids thicken under high pressure and thus form a temporary instantaneous polymer bond at the rolling interface for contact pressures above 104 MPa (15,000 psi). This allows the effective coefficient of friction to be about what an unlubricated system would have, on the order of 0.1 for polished steel on polished steel. Yet because of the presence of the fluid, reduced wear and slip damage at the contact interface is generated because there is always a fluid layer between the components.

Tractive fluids were developed originally for use in continuously variable transmissions (CVT) for the automobile industry and are often used in traction drives for industrial applications. Tractive fluids are also used in ball bearing-supported systems subject to high accelerations to help keep the balls from skidding. Tractive fluids cost significantly more than conventional oils and greases, so they are used only when needed.

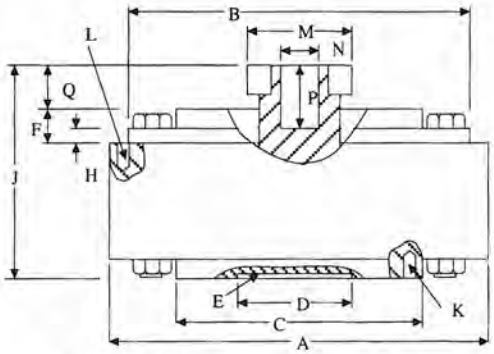
Round rollers can be made far more accurately than gears can be properly ground; hence traction drives have the potential to achieve near perfect accuracy with zero backlash. However, small dimensional variations can greatly alter preloads, so one must be careful when specifying tolerances for a traction drive system and, if possible, build in eccentric shaft supports and other adjustment devices.

---

72 See, for example, M. Seneczko, "Gearless Speed Reducers," Mach. Des., Oct. 14, 1984.

73 See, for example, D. Cameron, "Traction Drives for High Speed Reduction", Power Transm. Des., Vol. 22, No. 9, 1979, pp. 46–47.

74 For example, Monsanto Corporation's Santotrac® fluid.



Parameter	Units	M02	M03	M04	M05	M06	M08	M10	M12
A ± 0.010	inch	4.000	5.25	7.125	8.125	10.125	12.250	15.000	19.250
B	inch	3.030	3.999	5.499	6.499	7.999	9.874	11.749	15.874
		3.028	3.997	5.497	6.497	7.997	9.871	11.746	15.870
C ± 0.010	inch	2.000	2.625	4.000	4.750	5.500	7.000	8.500	11.50
D	inch	1.375	1.437	1.625	1.812	2.187	2.812	3.625	4.437
		1.376	1.438	1.626	1.814	2.189	1.814	3.627	4.440
E	inch	0.06	0.08	0.09	0.09	0.09	0.12	0.12	0.12
F	inch	0.48	0.53	0.62	0.59	0.75	0.94	1.19	1.31
H	inch	0.17	0.17	0.24	0.21	0.25	0.32	0.44	0.50
J	inch	2.59	2.86	3.12	3.70	4.19	4.94	5.94	8.38
K		10-32	1/4-28	1/4-28	5/16-24	5/16-24	3/8-24	1/2-20	5/8-18
K_{diam}	inch	1.693	2.000	3.562	4.250	5.000	6.375	7.750	10.500
L		10-32	10-32	1/4-28	1/4-28	5/16-24	3/8-24	1/2-20	1/2-20
L_{diam}	inch	3.329	4.438	6.000	7.125	8.625	10.750	13.000	17.500
M	inch	1.31	1.75	2.06	2.38	2.75	3.00	3.00	3.25
N_{max}	inch	0.551	0.75	1.00	1.38	1.38	1.62	1.62	1.75
P	inch	0.88	0.93	0.94	1.28	1.34	1.50	1.62	2.25
Q	inch	0.56	0.62	0.62	0.75	0.81	0.94	0.94	1.38
Weight	lbf	5	9	19	30	50	83	152	364
Rated torque	in-lbf	100	500	1000	2000	4000	7000	14000	25000
Inp. inertia	in-lbf-sec^2	0.000107	0.000212	0.000686	0.00229	0.00468	0.0150	0.0316	0.190
$K_{torsion\ out}$	in-lbf/rad	100000	250000	475000	750000	1175000	2250000	3750000	5575000
Nom. Efficiency	%	50	55	60	64	68	72	76	80
Max. Inp. speed	rpm	8000	6000	5000	4000	3600	3000	2400	2000
Output shaft capacity:									
Radial	lbf	560	860	1630	1820	4020	4630	8490	10100
Thrust	lbf	1410	2130	4120	4540	10060	11530	21100	25200
Moment	in-lbf	930	1890	4900	6360	17600	24800	55200	85550

Figure 10.7.7 Properties of Dojen® cycloidal speed reducers. (Courtesy of Dojen Div., Lenze USA, L.P.)

WireCapstan Drives[75]

For limited ranges of motion, an alternative to gears is the wire capstan drive illustrated in Figure 10.7.8. The wire capstan drive provides the advantages of traction drives without the accompanying large radial bearing loads and close manufacturing tolerances. The amount of motion obtainable with a wire capstan drive depends on the number of cable windings on the large-diameter shaft. The figure 8 cable wrap essentially eliminates radial bearing load. Like most belt drives (discussed below), there is no backlash. The transmission ratio is directly proportional to the diameters of the input and output shafts and can be as high as 50:1. The maximum torque that can be transmitted and the stiffness depend on the cable diameter, number of cables, shaft diameters, and coefficient of friction. Since wire capstan drives are designed and built especially for each application, details of the analysis required to calculate transmittable torque and stiffness will be discussed here.

The cable diameter D_{cable} is chosen such that the sum of the preload and dynamic cable loads is less than 10% of the cable's breaking strength. Typically, the preload on the cable is set equal to the dynamic load, which is the output torque divided by the number of cables and the radius of the output shaft. Thus as the cable begins to apply a torque on the output shaft, the tension is

[75] Marketed as a *Roto-Lok* drive from Sagebrush Technology Inc., Albuquerque, NM (505) 299-6623.

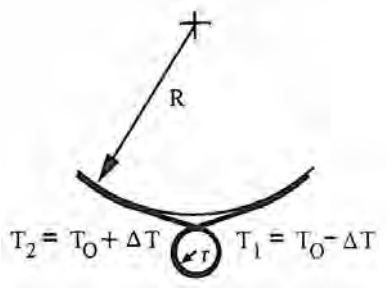


Figure 10.7.8 Wire capstan drive speed reducer.

released from one side and increased on the other. The result is that the maximum working load is 1.5 times the dynamic load.[76] It is also important that the diameter of the smallest shaft (generally the input shaft) be at least least 25 times the uncoated cable diameter, $D_{input} = 25D_{cable}$. The output shaft diameter is then the product of the input shaft diameter and the transmission ratio, $D_{output} = TRD_{input}$. For a typical steel-stranded cable,[77] the required cable diameter in millimeters is:

$$D_{cable} = 1.12\left(\frac{\Gamma_{out}}{NTR}\right)^{1/3} \tag{10.7.7}$$

where Γ_{out} is the output torque (N-m) of the device and N is the number of cables. This diameter should give essentially infinite lifetime (several million cycles).

To keep the cable from slipping on the input (small) shaft, it must have a sufficient wrap angle θ. In terms of the coefficient of friction μ and the tension ratio, the cable wrap angle is found using the analysis technique for a capstan:

$$\theta_{max} = \frac{\text{Log}_e \dfrac{T_{preload} + T_{load}}{T_{preload}}}{\mu} \tag{10.7.8}$$

From the condition of the preload equaling the dynamic load, the maximum ratio between input and output cable tensions on the small shaft will be 3. The coefficient of friction is usually about 0.15 for plastic-coated cable on a steel drum, so typically the cable must wrap 7.3 rad or about 1.2 wraps, regardless of the size of the input shaft. The cable can only make an integral number of wraps on the input shaft, so where any slip near maximum load cannot be tolerated or where maximum stiffness is desired, 2 wraps per cable strand should be used. On the output shaft the tension ratios are 1.5 for the tightening side and 2 for the slackening side, so the wrap angles are at most only 155° and 264°, respectively.

Cable stiffness is calculated from the free length of the cable and the portion under tension in the wrap angle on the input and output shafts. As shown in Figure 10.7.8, in terms of the distance λ between the outer circumferences of the two shafts, the free cable length on each side of the input shaft spans an angle ϕ:

$$\phi = \cos^{-1}\left(\frac{1}{1 + \dfrac{2\lambda}{D_{input} + D_{output}}}\right) \tag{10.7.9}$$

The free length of the cable on either side of the input shaft is

$$\ell_{free} = \frac{D_{input} + D_{output}}{2}\tan(\phi) \tag{10.7.10}$$

In order to calculate the equivalent torsional stiffness, one must realize that the preloaded cables act like a set of preloaded bearings.[78] Thus the total stiffness is the sum of the equivalent stiffnesses

[76] The torque is formed by the couple $0.5T_{dyn}$ separated by the shaft diameter.
[77] For example, a 1 mm (0.040 in.) steel stranded cable has a breaking strength of about 860 Newtons (200 lbf).
[78] See Section 8.2, Equation 8.2.2.

of each cable. The stiffness of each cable is dependent on its wrap angle on the input and output shafts and on the free length. Initially, the tension is everywhere equal to the preload tension. As the torque is generated, one cable's tension changes from the preload tension on the output shaft to the preload tension - (+) load tension in the free length zone, and back to the preload tension on the input shaft. As the torque increases to the maximum, the tension zones on the input drum may overlap, but generally enough of a safety factor exists in the design so that this effect can be ignored here.

The total stretch $\delta_{\text{cable shaft}}$ in a cable wrapped around a shaft through an angle large enough to prevent slip due to the different tensions on the ends of the cables is found from

$$\begin{aligned} \delta_{\text{cable shaft}} &= \int_{-\theta_{\max}}^{0} \frac{R_{\text{shaft}} \left(T_{\text{preload}} + T_{\text{load}}\right) e^{-\mu\theta} \, d\theta}{Æ} \\ \delta_{\text{cable shaft}} &= \frac{R_{\text{shaft}} T_{\text{load}} \left(T_{\text{preload}} + T_{\text{load}}\right)}{Æ \, \mu \, T_{\text{preload}}} \end{aligned} \tag{10.7.11}$$

where T_{preload} and T_{load} are cable tensions caused by the preload and output torque (one-half the dynamic load), respectively. Herein Æ is the effective product of the cable's cross-sectional area and modulus of elasticity and is available from cable manufacturers. The axial stiffness of the cable in this region is found by taking the inverse of the compliance ($\partial\delta/\partial T_{\text{load}}$) found from Equation 10.7.11:

$$K_{\text{cable shaft}} = \frac{Æ \, \mu \, T_{\text{preload}}}{R_{\text{shaft}} \left(T_{\text{preload}} + 2T_{\text{load}}\right)} \tag{10.7.12}$$

This axial stiffness is a minimum when the load is a maximum. The maximum and minimum load tensions will be $T_{\text{load}} = 1.5T_{\text{preload}}$ and $T_{\text{load}} = 0.5T_{\text{preload}}$, respectively, so Equation 10.7.12 yields the following expressions for the tensioned and slackened cables' axial stiffnesses, respectively:

$$K_{\text{tensioned}} = \frac{Æ \, \mu}{4R_{\text{shaft}}} \qquad K_{\text{slackened}} = \frac{Æ \, \mu}{2R_{\text{shaft}}} \tag{10.7.13}$$

The tension cable has a lower stiffness because the higher load increases the effective wrap angle and hence the effective length. Each cable wraps on the input and output shafts, respectively. Each cable also has an axial stiffness associated with the free-length section:

$$K_{\text{free length}} = \frac{Æ}{\ell_{\text{free}}} \tag{10.7.14}$$

Each cable's total axial stiffness is thus comprised of three (input, free, output) stiffness terms in series. The total axial stiffness for each cable is thus

$$K_{\text{tensioned}} = \frac{Æ}{\frac{2(D_{\text{input}} + D_{\text{output}})}{\mu} + \ell_{\text{free}}} \tag{10.7.15a}$$

$$K_{\text{slackened}} = \frac{Æ}{\frac{(D_{\text{input}} + D_{\text{output}})}{\mu} + \ell_{\text{free}}} \tag{10.7.15b}$$

The total torsional stiffness of the preloaded cable system is thus

$$K_{\text{torsion}} = \frac{\left(K_{\text{tensioned}} + K_{\text{slackened}}\right) D_{\text{output}} \, N}{2} \tag{10.7.16}$$

In practice it is not uncommon for the torsional stiffness of the cable system to be greater than the torsional stiffness of the input shaft.

10.7.3 Belts and Chains

Most people are familiar with the operating principles of belts and chains. The types of belts and chains commonly found in precision machines include vee belts, timing belts, flat belts, and roller chain. Most traditional machine element design texts discuss load and life design factors for all these

types of belts and chains. The discussion here will focus on how they affect the performance of a precision machine.

Remember that with any belt or chain system, a means to adjust the tension is usually required. This can be accomplished by enabling one of the pulley shafts to be movable, or through the use of an idler pulley. Spring-loaded idler pulleys are an effective way to maintain constant belt tension even as the belt stretches with use.

Vee Belts

Vee belts are the least expensive way to transmit power between parallel shafts and thus are the most commonly used belt. Their vee shape wedges into the pulley with increasing loads and contributes to the belt's ability to transmit power. A fair amount of preload is initially required to seat the belt, and thus radial shaft loads are induced. In addition, their vee shape also means that contact is made at more than one diameter so there is relative slip between the belt and the pulley and heat is generated. Vee belts are often used to transmit power to spindles.

Timing Belts

Timing belts have teeth that transmit power to toothed pulleys much like the chain on a bicycle does; however, because they are flexible, timing belts bend over a smooth continuous pitch diameter and thus do not induce any appreciable cogging effect. Timing belts are often reinforced with Kevlar® and thus can have high effective torsional stiffnesses when the pulleys are not far apart. Timing belts require an initial preload to prevent backlash upon load reversal. The preload tension also makes all the teeth engage the pulley, which helps create an elastic averaging effect. This makes timing belts an effective accurate means of transferring power from a servomotor to a shaft when the motor cannot be directly coupled to the shaft.

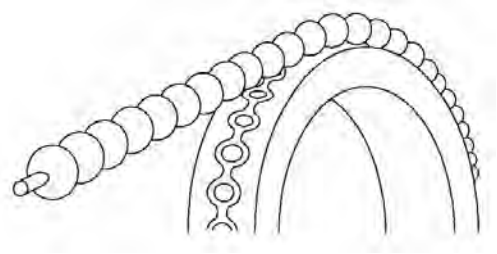


Figure 10.7.9 Bead belt drive.

Timing belts are most often seen as flat belts with teeth. These belts must be purchased in fixed lengths. Timing belts are also available in a form that looks like a strand of rope with ladder rungs or beads as the drive teeth as shown in Figure 10.7.9. This type of timing belt can be bought as a long strand and cut and spliced to the length desired. A bead belt can be routed over various pulley configurations to transmit power between nonparallel shafts.

Flat Belts[79]

Flat belts conjure up images of nineteenth-century factories with forests of belts transmitting power from a single overhead main shaft to individual machines below. Flat belts are still used but materials like Kevlar and spring steel have replaced leather. Flat belts can have higher efficiencies than vee belts and are generally preferred for high speed applications. For low speed high-torque applications, vee belts or chains are more often used. To help lessen alignment requirements, flat belts ride on pulleys that have a slight crown. Reverse bends are also possible with flat belts. In some applications, mostly on small machines or instruments, the belt is perforated and it rides on a toothed pulley. Cables can also be thought of as flat belts.

Roller Chain

Roller chain can transmit significantly higher loads than a belt system and is the least expensive type of chain drive. The problem with roller chain is that its links can engage the sprocket only in an incremental fashion. As the sprocket turns there is an effective sinusoidal variation in the pitch diameter. This variation is unacceptable for a precision machine drivetrain. Also, the rigidness

[79] See R. Morf, "Flat Belts Shed the Leather Strap Image," Mach. Des., March 9, 1989.

of the links means that only a few teeth on the sprocket will be transmitting load to the chain and hence there is far less of an elastic averaging effect than there is with timing belts. Roller chain has been used to transfer loads between counterweights and machine components (e.g., spindle assembly counterweights on vertical spindle machines), but it is better to use a cable to avoid any potential cogging torque problems.

Silent chain looks like a timing belt in cross section, although it is made up of chain links and it is capable of significantly higher speeds and power transmission than roller chain. However, because it uses links, it still has an effective varying pitch diameter, which makes it unsuitable for precision machine servos. Silent chain is generally used for high-speed, high-power applications.

10.7.4 Cams

Cams are used to transform continuous rotary motion into nonlinear motion. There are a great many different types of cams, some of which are illustrated in Figure 10.7.10.[80] Classic machine design texts discuss cam design in detail, and there are a number of software programs available for cam design. Cams used to be a vital part of automatic screw machines and machines for turning or grinding nonround parts. However, in today's age of numerical control cams are used less and less as control elements in precision machine tools. Cams, however, are still a vital part of many machines and consumer products (e.g., automobile engines).

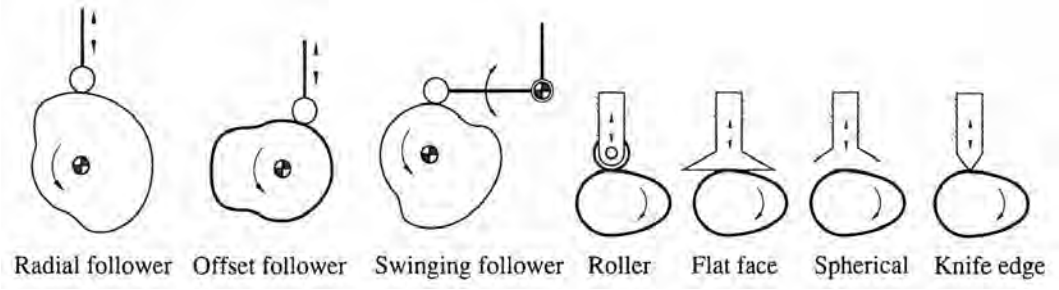


Figure 10.7.10 Types of cams.

10.7.5 Couplings

There are two basic types of couplings: power couplings and servo couplings. Both types are meant to allow for lateral and angular misalignment. Power couplings are designed to transmit torque while also often acting as a shock absorber to smooth out load variations and vibrations. Servo couplings, on the other hand, are designed to respond instantly without storing any appreciable energy or having any backlash or friction. Based on a consideration of bearing types discussed in Chapter 8, flexural bearings are the best choice for use in servo couplings.

In precision machines, the use of modular motors and sensors often means that a coupling is needed to allow for all the types of alignment errors that can occur between shafts. In order to avoid the periodic errors described in Section 5.3.3, a flexible coupling should be used, which provides essentially constant-velocity power transmission regardless of shaft angle. Flexural couplings also typically have zero backlash. Note that manufacturing tolerances prevent any flexural coupling from acting like a perfect flexure, so there will always be some parasitic motion, as discussed in Section 8.6, which can cause very small velocity errors that are negligable for most applications. For example, when a motor is coupled to a precision ballscrew to move a carriage, if a linear position sensor is used, coupling errors may only be on the order of a few parts-per-million which will be easily compensated for by the control system. If a coupling manufacturer does not provide *actual data* on the error produced by the couplings they sell, a conservative estimate of the error can be

[80] Other types of cams, linkages and mechanisms are described in wonderful detail in Ingenious Mechanisms (4 vols.), Industrial Press, New York.

obtained from:[81]

$$\varepsilon_{\text{flexible coupling}} = \left[\frac{\text{shaft eccentricity}}{\text{small shaft radius}}\right]\left[\frac{\text{coupling bore eccentricity}}{\text{small bore radius}}\right] \tag{10.7.17}$$

For example, consider coupling a 10 mm diameter shaft to a 30 mm diameter shaft, where the shaft eccentricity is 0.1 mm. The coupling's bores are assumed to have a 0.025 mm eccentricity due to manufacturing error, so the total maximum coupling error for the system can be estimated to be (0.1/10)(0.025/10) = 25 μrad.

There are six basic types of couplings for precision applications that can handle bidirectional torques:[82]

- Metal bellows couplings
- Helical couplings
- Link couplings
- Diaphragm (flexible disk) couplings
- Center of percussion couplings
- Belt couplings

Whichever type of coupling is chosen, a setscrew should never be used to clamp a coupling to a precision shaft. A split ring that squeezes the shaft upon tightening of a bolt should always be used.

Metal Bellows Couplings

A metal bellows coupling is shown in Figure 10.7.11. Metal bellows were discussed in Section 10.6.1 in the context of actuators. Metal bellows couplings cost substantially less than metal bellows actuators because metal bellows couplings are made in standard sizes and they do not need to be able to hold pressure or vacuum. Metal bellows couplings provide perhaps the greatest coupling action of any flexible coupling but have a commensurate lower torsional stiffness. They are generally used in applications where torque levels are less than about 10 N-m (90 in.-lbf). For most precision machine applications, speed is not usually a limiting factor in metal bellows coupling design and they can be sized to undergo millions of cycles without failure.

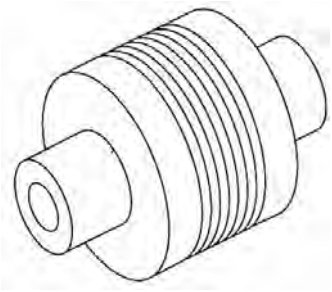


Figure 10.7.11 Metal bellows coupling.

Helical Beam Couplings

Helical beam couplings come in a large variety, all of which incorporate one or more curved beams that extend from one end of the coupling to the other. The beam is in the shape of a helix, or spiral, generated in a hollow cylinder as shown in Figure 10.7.12. This type of coupling relies on bending of the helical beam to accommodate misalignment.

Helical couplings work well in most types of rotating machinery. Typical applications include ball and leadscrews, encoders, gearboxes, pumps, conveyor systems, and rollers usually driven by electric motors. Helical beam couplings accommodate radial and angular misalignments. Axial motion between the two shaft ends is absorbed by the coils. They can be manufactured with a specific torsional stiffness and a wide range of torque capacities. Since they are usually one piece couplings, there is no backlash. By design, the rotational output motion of the driven end can be assumed to be the same as the input motion, making it a constant velocity coupling.

[81] Sadly, the author could not find a proof for this rule-of-thumb equation. It would be nice to see some experiments done on this subject.

[82] Other types of couplings for general power transmission are discussed, for example, in J. Shigley and C. Mischke, Standard Handbook of Machine Design, McGraw-Hill Book Co, New York.

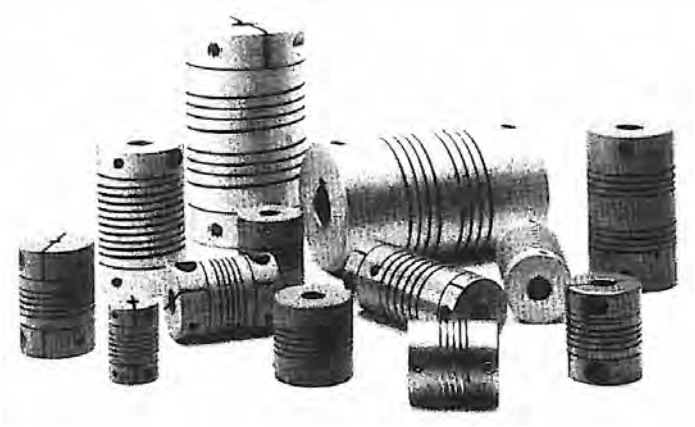


Figure 10.7.12 Helical flexible couplings. (Courtesy of Helical Products Company, Inc.)

Basic part #	Length (in.)	Outside diameter (in.)	Bore diameter (in.)	Max. torque (lbf-in.)	Torsional stiffness (deg/lbf-in)
MCAC100	1.75	1.00	0.313	23	0.370
MCAC125	2.38	1.25	0.375	47	0.170
MCAC150	2.63	1.50	0.500	88	0.100
MCAC200	3.00	2.00	0.625	164	0.049
MCAC225	3.50	2.25	0.750	262	0.032

Figure 10.7.13 Characteristics of commercially available helical couplings: 7075-T6 aluminum, integral squeeze clamp, 0.030 in. lateral misalignment capability, 5 ^{o} angular misalignment, $\pm$0.010 in. axial motion. Note that each end of the coupling can have its own specified bore diameter. (Courtesy of Helical Products Company, Inc.)

Many types of materials may be used to manufacture helical beam couplings. High strength aluminum is often used for encoder coupling applications. High strength stainless steel is usually used where high torque capacity, stiffness, and corrosion resistance are required. Figure 10.7.13 shows a family of helical couplings typically used in small-to-medium sized machine tools. Various end attachments are easily incorporated. Other bore sizes are also available.

Link Couplings

A link-type coupling uses flexural elements (e.g., the flexural pivot shown in Figure 8.6.10). Link couplings can be made to have even greater coupling action then either of the previous couplings because the links can be made from very thin pieces of spring steel. Remember, the greater the coupling action, the less the torsional rigidity.

Diaphragm-type (Flexible Disc) Servo Couplings

A diaphragm-type coupling is shown in Figure 10.7.14. Typical characteristics of this type of coupling are shown in Figure 10.7.15. Diaphragm-type couplings typically generate one-third less radial load on the shaft and have twice the torsional stiffness of other types of flexible servo couplings. Like most types of flexural couplings, this type of coupling can typically be used at several thousand rpm; however, speed limitations depend on load, misalignment, size, and so on.

Center-of-Percussion Couplings[83]

Most couplings will transmit radial forces from one shaft to another; however, when coupling a motor shaft to a precision spindle, it is important to prevent vibrational forces from the motor shaft from being transmitted to the spindle. These vibrational forces can be caused by an imbalanced rotor or unsymmetrical magnetic field lines. On the micron level, brushless direct-drive spindle motors may seem best, but on the microinch level, the latter effect can cause significant asynchronous

[83] This coupling was first described to the author by Eugene Dahl of Professional Instruments, Inc., which patented the concept in 1974: Vibration Attenuation Coupling Structure, US Patent 3,800,555, April 2, 1974.

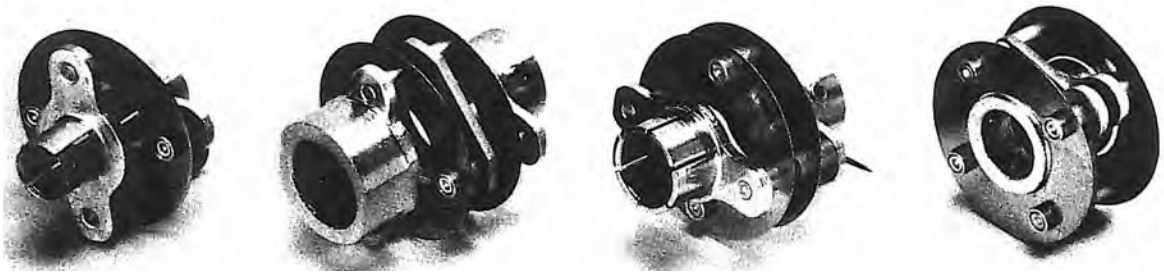


Figure 10.7.14 Fleximite® diaphragm (flexible disk) type flexible servo coupling. (Courtesy of Renbrandt, Inc.)

Size	O.D. (in.)	Bore (in.)	Torque (in-oz)	Angle (deg.)	Lateral* (in.)	I (alum.) (oz.-in^2)	Weight (alum.) (oz.)	Torsional stiffness [min/(oz-in.)] B	C	E	Radial force (oz./0.001") B	C	E
1	3/4	1/16-1/4	30	3	0.010	0.006	0.19	0.71	0.44	1.9	2	0.8	0.34
2	1	1/16-3/8	60	3	0.015	0.013	0.29	0.64	0.37	2.8	2	0.8	0.4
3	1.5	1/16-5/8	250	3	0.018	0.044	0.44	0.21	0.17	0.39	4.9	1.8	0.2
5	2.5	1	20ft-lb	2	0.010	4.15	5.71	-	0.006	-	-	7	-

* Allowable shaft TIR is twice this value.

Figure 10.7.15 Characteristics of Fleximite® couplings. (Courtesy of Renbrandt, Inc.)

motion of the spindle. Center-of-percussion couplings can prevent the transmission of vibrational forces from the motor to the spindle while maintaining a high degree of torsional rigidity.

If you take a pencil laid flat on the table and hit one end of it, you will notice that the pencil rotates about a point which is known as the *center of percussion*. For example, baseball players know that unless the ball hits the bat at the right spot, the impact will sting their hands. If the pencil is now a drive shaft with one end being attached to the output shaft of the motor, the driveshaft should be attached to the spindle at the center-of-percussion point. It is impractical to have a long shaft protruding beyond the coupling point, so instead, one uses a shaft with a heavy weight near the coupling point. The location of the center of percussion is determined as follows:

The restoring torque due to gravity that acts on a body that is free to rotate about a point like a pendulum is

$$\Gamma = -Mg\ell \sin\theta \tag{10.7.18}$$

This torque equals the product of the rotational inertia of the body about the center of rotation and the angular acceleration:

$$I_{cr}\frac{d^2\theta}{dt^2} + Mg\ell \sin\theta = 0 \tag{10.7.19}$$

For small angles, the period of oscillation of this system, referred to as a physical pendulum, is

$$\tau = 2\pi\sqrt{\frac{I_{cr}}{Mg\ell}} \tag{10.7.20}$$

For a simple pendulum (i.e., a ball on the end of a string) l = L, M = m, and $I_{cr} = mL^2$, so the period is

$$\tau = 2\pi\sqrt{\frac{L}{g}} \tag{10.7.21}$$

The length of a simple pendulum that has the same period as a physical pendulum is

$$L = \frac{I_{cr}}{M\ell} \tag{10.7.22}$$

This point on the body is called the *center of oscillation* or *center of percussion*. If a force impulse is applied at the center of oscillation, the body will undergo translational and angular accelerations. At the axis of rotation, the translational acceleration will be a = F/M. The rotational acceleration about

the center of mass will be $\alpha = \Gamma/I_{cm}$ = -F(L - l)/I_{cm}. The equivalent translational acceleration at the axis of rotation is just -F/M and hence there is no motion (or force) at the point a distance L from the center of oscillation. Therefore, the center of oscillation is also known as the center of percussion. Likewise, by Maxwell's law of reciprocity (or simple force balance calculations) a force applied at the axis of rotation (not the center of mass) due to, for example, bearing vibration will not cause a force to be felt at the center of percussion.

This is a very powerful result and means that if one adds an overhanging weight to one end of a coupling so that the coupling's center of percussion coincides with the point of coupling action, radial forces will not be transmitted between the shafts. This analysis holds true for couplings that have essentially rigid shafts, such as drive shafts with a universal joint at each end or flexible couplings with a long rigid section between flexible sections at each end. Note that a single Hooke's joint like in an automobile driveshaft is not a constant-velocity joint; however two shafts joined by a shaft with a Hooke's joint at each end will rotate without speed variations between them (if they are in phase and operating at the same angle).

The Hooke's joints on each end of the shaft can be made with precision ball or needle bearings to minimize backlash and have low friction and high torsional stiffness. A parallel flexure machined integral with the shaft can allow the shaft to accommodate length changes. A more economical alternative that is well suited for devices which are servo velocity controlled primarily in one direction (e.g., spindles) is to replace the rolling element bearings with sliding plastic parts which can also allow axial growth of the shaft, thereby also eliminating the need for the flexural bearings. Thus a motor could be located on a vibration absorbing mount near a precision machine and coupled to the machine without having large radial loads transmitted to the precision machine's input shaft as the motor vibrated on its vibration absorbing mounts. Note that a conventional flexible coupling would require the motor shaft to be held in close alignment to the input shaft, which would not be conducive to mounting the motor on a vibration absorbing material. Systems with center of percussion couplings have been used for some time by Professional Instruments Corp. on many of their modular motor-driven spindle product lines. Note that as brushless motor technology advances and motors are made that are precisely dynamically balanced, direct-drive systems will begin to dominate in more and more applications.

Belt Couplings

If a long springy continuous fiber belt is used to transmit power from one shaft (e.g., a motor shaft) to another (e.g., a spindle), small radial motions of the motor shaft will insignificantly affect the belt tension; hence vibrational forces from the motor will not be transmitted to the spindle through the coupling device. When the length of the belt is parallel to a nonsensitive direction (e.g., the vertical direction in a T-base lathe), the spindle will not experience any sensitive direction error motions caused by the motor. It is important that the belt be continuous, so there is no belt joint to cause a periodic error. A continuous flat fiber belt (e.g., made from wool) also has a greater deal of damping than, for example, an elastomer belt. The belt also serves to help thermally isolate the motor from the spindle. As spindle accuracy requirements head toward the microinch and better realm, even the magnetic asymmetries in electric motors can cause forces that generate unacceptable levels of asynchronous spindle motion. Hence ironically precision machines may see a return to the days of old, where long belts coupled motors to spindles.

10.8 Linear Power Transmission Elements

There are many types of devices which convert rotary power to linear power, and these are discussed in this section. As with rotary power transmission components, there are five principal error sources that affect linear power transmission elements' performance:

- Form error in the device components
- Component misalignment
- Hysteresis
- Backlash
- Friction

Each of these error components were discussed in detail in Section 10.7 and apply here equally well. In addition, one must also consider all of the error sources and their effect on the carriage

being actuated, as discussed in Chapter 2. Specific examples are discussed for leadscrews in Section 10.8.3 but are also applicable to other rotary-to-linear power transmission systems. Linear power transmission elements which are discussed in this section include:

- Rack and pinion drives
- Friction drives
- Leadscrew drives
- Coupling elements

10.8.1 Rack-and-Pinion Drives

A rack-and-pinion drive provides one of the least expensive methods of generating linear motion from rotary motion. Its main feature as far as machine tools are concerned is gear racks can be placed end to end for as great a distance as one can provide a secure base on which to bolt them. Hence rack-and-pinion drives are commonly used on very large machines such as gantry robots and machining centers used in the aircraft industry where lengths of travel are greater than about 3-5 m. Rack-and-pinion drives' biggest drawback is that they do not provide a mechanical advantage the way a leadscrew system does. It is difficult to obtain the "optimal transmission ratio" and a direct-drive motor for a rack-and-pinion system will often run at low speed and high torque; thus a speed reducer is sometimes used with a motor that drives the pinion. The exception to this is a worm gear rack, but unless a hydrostatic design is used as shown in Figure 10.8.33, friction will be too great.

The characteristics of gears discussed in Section 10.7.1 apply here equally well, including the use of antibacklash pinions. There is no elastic averaging effect, so tooth form errors and backlash are translated directly into the linear positioning errors the system will experience. Also, the nature of gear teeth making and braking contact creates a sinusoidally varying lateral force. This causes a sinusoidally varying straightness error in a carriage driven by a rack-and-pinion drive. The straightness error, in turn, changes the gear center distance that causes small nonlinearities in axial position. Most precision closed-loop control systems should have little difficulty in dealing with these axial errors.

Rack-and-pinion drives are most often used for machines with long travel ranges, and the motor and pinion are most often mounted on the carriage. In some instances, for short travel where low cost is desired and a fine motion stage is used for final positioning accuracy, the rack may be mounted on the carriage and the motor and pinion on the machine base. For a carriage-mounted motor, consider a simple T slide. The motor needs to be mounted on the outboard side of the carriage, so its heat can be most easily dissipated without warming up the carriage. The rack should be mounted near the center of the carriage so the actuation forces do not cause yaw errors in the carriage, and a shaft used to transmit power from the motor to the pinion.

An interesting variation on rack-and-pinion drives is a cam roller drive. This type of drive has a sinusoidal linear track with four or more vertically actuated cam rollers mounted on the carriage that press on the track. The rollers are such that one is always pressing at near the 45^{o} slope of the sine wave for maximum sensitivity in forward or reverse motion. Hence every downward newton of force is translated approximately into an axial newton of force. As the carriage is moved along, the motion of the cam rollers looks like a trumpet being played. In fact, typically the up-and-down motion of the cam rollers is controlled by pneumatic or hydraulic power. The cam rollers can be controlled by individual servovalves, so they can be preloaded against each other or with a single servovalve that has internal porting and flow control devices. This drive can have effectively zero backlash and very good repeatability and resolution to the micron level. One of this drive's most noteworthy applications has been on some gantry robots. For large, precision machines, however, the lateral forces generated by the cam rollers would cause too large straightness errors unless opposing systems that pushed against each other were used.

10.8.2 Friction Drives

Linear friction drives can be configured as a wheel (capstan) driving a flat bar supported by a backup roller as illustrated in Figure 10.8.1. When fluidstatic bearings are used in the system, the backup

roller is usually replaced with a flat fluidstatic bearing. Friction drives[84] are commonly used as linear actuators on high-precision machines. Friction drives' desirable properties include:

- Minimal backlash and deadband (due to elastic deformation).
- Low drive friction.[85]
- Uncomplicated design.

Their undesirable properties include:

- Low drive force capability.
- Low to moderate stiffness and damping.
- Minimal transmission gain.
- High sensitivity to drive bar cleanliness.[86]

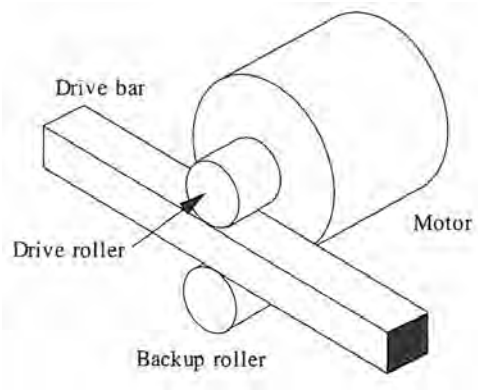


Figure 10.8.1 Friction (capstan) drive.

An ideal friction drive would use hydrostatic bearings to support the drive roller shaft, and a hydrostatic flat pad bearing instead of a backup roller. High surface finish of the contact surfaces is required to minimize wear and to ensure that the drive does not act like a wheel on a cobblestone street, and accurate rollers are required to maintain a constant preload, transmission ratio, and tare torque. With an appropriate sensor and servo system, a properly designed and manufactured friction drive can achieve nanometer resolution of motion. For high-accuracy applications, position repeatability and accuracy should be obtained with carriage position measurement by a linear encoder or laser interferometer. Because the components of a friction drive system can be made so accurately but their tractive force is limited, friction drives are found primarily on high-precision machines that are subject to relatively light cutting forces (e.g., the LODTM discussed in Section 5.7.1, and the OAGM discussed in Section 7.8. See Figure 7.8.5). In these applications, the maximum speed of the carriage is rather low and the drive roller is driven directly by a motor to avoid introducing transmission nonlinearities into the system.

The effective stiffness of the contact interface of a friction drive can be estimated using Equation 5.6.30 as discussed below. Tractive fluids can be used to prevent fretting corrosion and increase damping and make stiffness less a function of bar cleanliness. Note that if the tractive force is not applied through the drive bar's neutral axis, bending of the bar will occur. With a compliant coupling, this may not be a problem; however, if the bar is long, it may vibrate, due to the servo loop acting as an excitation source for the bar.[87] If a drive bar is made as a channel, the roller should make contact in the web and the flanges chosen in thickness and width, so the neutral axis lies along

[84] Also known as *capstan* or *traction* drives.

[85] Because a flat spot forms at the contact interface due to elastic interface, true rolling motion is not achieved; hence some friction exists, although in most cases it will be negligible.

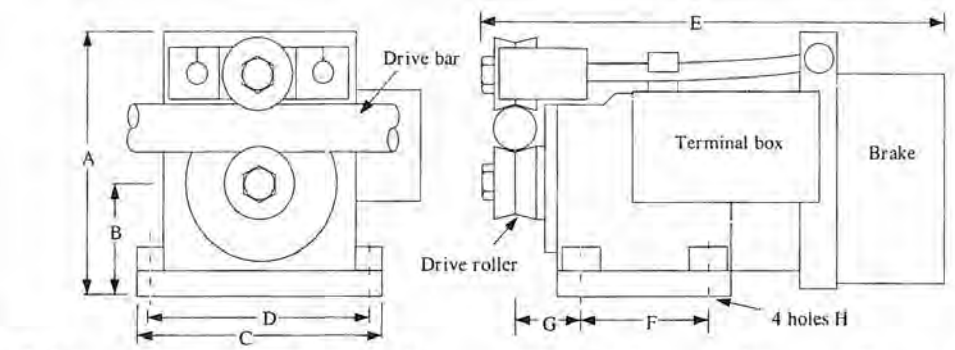
[86] The tangential stiffness is a function of the coefficient of friction. If properly sized, however, friction drives can be run with the drive bar coated in oil.

[87] Jim Bryan described to the author a case where the drive bar generated a high-pitched audible whine when the machine was being servo-controlled. In this case, the problem was solved by altering the control algorithm and loop time, but it would have been nice if the problem had never occurred.

the contact point. In this manner the contact point will also be at the shear center of the beam, so twisting will not occur if, for example, two backup rollers were used with the drive roller to fully support the drive bar. It is left as an exercise for the reader to design this type of drive bar and the method by which it can be manufactured.

With a direct-driven drive roller, it is generally very difficult to choose a large enough diameter capstan to keep the contact stresses low while meeting the optimal transmission requirements discussed in Section 10.2.2. As the mass of the carriage increases, the optimal drive roller diameter must decrease; however, with increasing carriage mass, more force is required to move the mass, so the capstan diameter must increase to keep preload contact pressures at an acceptable limit. The drive roller cannot simply be made longer to lower contact stresses because then manufacturing becomes more difficult.[88] The use of a conventional speed reducer for the motor is also not practical because then many of the benefits of a direct-drive system would be lost. If a speed reducer is required, a traction drive (rotary friction drive) speed reducer can be used. Fortunately, most machines that need a friction drive move in a quasi-static mode anyway and have good thermal control of heat-generating components.

Since friction drives require high preload forces, typically 10 times the actuation force, considerations of shaft bending are very important. Note, however, that the preload force needed for a friction drive can also serve to preload the carriage bearings on some designs (e.g., if a friction drive was used to actuate the carriage design shown in Figure 8.5.9). Because alignment of friction drive components is so critical and small geometric variations can cause the capstan to become misaligned or change its preload, it is best to preload a friction drive's backup roller using flexural bearings that act as large springs to provide the preload force. In addition, to accommodate misalignment, which will always be present to some degree, between the bar and axis of motion and the drive roller and the bar, a flexural mounting system can be used as discussed in detail in Section 8.6.5. For short-stroke applications, the drive roller assembly may be stationary and the drive bar may be attached to the carriage by an hourglass-shaped link which acts as a flexible coupling (see Figure 10.8.40).



Model	A (mm)	B (mm)	C (mm)	D (mm)	E (mm)	F (mm)	G (mm)	H (mm)
1656/2	145	62	140	120	291	65	38	11
1656/3	340	140	265	230	450	75	100	13

	Max torque (N-m)	Max speed (rpm)	Vel. ripple (±%)	Vel. linearity (% of F.S.)	Typ. drive roller diameter (mm)	Typ. drive rod diameter (mm)
1656/2	6.8	300	2.0	0.5	40	20
1656/3	54	416	0.1	0.2	60	50

Figure 10.8.2 Characteristics of Cranfield Precision Engineering's friction drive. (Courtesy of CPE Ltd.)

Figure 10.8.2 shows a commercially available direct-drive friction actuator[89] that uses a round drive bar and grooved rollers to minimize alignment requirements. Note that the backup (preload)

[88] The use of a spreadsheet program will help tremendously in evaluating design parameters. As with any design problem, sizing capstan drive components is a balancing act.

[89] Cranfield Precision Engineering Ltd., Cranfield, Bedford MK43-0AL, England.

Portion of max load at which stiffness is evaluated				50%		Roller modulus of elasticity (GPa)				310	
Usable force portion of maximum load				75%		Drive bar modulus of elasticity (GPa)				204	
Coefficient of friction				0.1		Roller Poisson ratio				0.27	
Maximum desired contact stress (GPa)				2.07		Drive bar Poisson ratio				0.29	
Roller D_{minor} (mm)	25	30	35	40	45	50	55	60	65	70	75
Roller D_{major} (mm)	2500	3000	3500	4000	4500	5000	5500	6000	6500	7000	7500
Preload force (N)	5963	8587	11687	15265	19320	23852	28861	34347	40309	46749	53666
Usable (75%) F (N)	447	644	877	1145	1449	1789	2165	2576	3023	3506	4025
Semiaxis c (mm)	3.95	4.74	5.53	6.32	7.10	7.89	8.68	9.47	10.3	11.1	11.8
Semiaxis d (mm)	0.348	0.418	0.488	0.558	0.627	0.697	0.767	0.836	0.906	0.976	1.05
K along c (N/μm)	243	291	340	388	437	486	534	583	631	680	728
K along d (N/μm)	245	293	342	391	440	489	538	587	636	685	734

Figure 10.8.3 Spreadsheet used for the design of a friction drive with a silicon nitride roller and steel drive bar.

roller is preloaded against the drive roller with stiff flexures. A friction drive that uses a kinematic arrangement of cam rollers to support a round bar with a flat edge that is driven by a crowned roller through a friction type transmission is also commercially available.[90] Because a transmission is provided, it is possible to choose an optimal transmission ratio. A cylinder and a flat plane can perhaps be made more accurately and smoothly than any other shapes with the exception of a sphere, and thus some manufacturers use designs with rectangular drive bars and cylindrical drive rollers even though greater alignment care is required.

Another variation on the friction drive theme is that of a cable that wraps around a motor driven capstan and a freely rotating capstan, and is used to pull a carriage back and forth. This application is seen in many types of wheel and dot matrix printers. One advantage of this type of drive is that the cable is usually long enough to make the system self-coupling. In addition, the drive motor and all the heat it produces can be kept far away from the carriage. The primary disadvantage is that the wire is very compliant, so system stiffness will be very low. As discussed in Section 10.9, there is a way to control this system that increases the apparent stiffness of the wire.

In many instances, one will want to custom design a friction drive for a particular application. The Hertz equations, discussed in Section 5.6, can be used to help develop the design. For example, using the Hertz equations, one can easily develop a spreadsheet which characterizes the properties of a friction drive. Figure 10.8.3 shows a portion of the output from a spreadsheet used to design friction drives.

Hysteresis and Damping Caused by Slip at the Contact Interface

Deresiewicz[91] originally derived Equation 5.6.27a and then considered the effect of unloading from a load T* to arrive at the expression for the displacement during the unload cycle, which is given by Equation 5.6.27b. Once again, this equation must be evaluated for the displacement on the roller and the drive bar, and then the displacements summed to obtain the total displacement of the system. Using Equations 5.6.27 and 5.6.31, the load/displacement curve is obtained for the 50 mm roller friction drive described in Figure 10.8.3. The hysteresis is too small to see between the load and unload curves, so it is plotted along with the load/displacement curves in Figure 10.8.4.

As the tangential force level increases, so do the velocity and hysteresis. The area of the hysteresis loop represents the energy dissipated by slip. If a time period is established between incremental changes in force, then the acceleration of the mass by the force can be used to compute the velocity. The force rises and then falls, which corresponds to a parabolic increase in velocity. If the energy dissipated in the hysteresis loop is divided by the time and the average velocity squared, then an average damping figure can be obtained. Figure 10.8.5 shows that for the 50 mm diameter roller friction drive discussed above, the damping will be proportional to the velocity.

Because the damping is proportional to speed, a control system tuned to provide "optimal" response at a visible speed may be underdamped at a low speed. If the controller is tuned to provide an "optimal" response at a slow speed, it may be overdamped at a high speed. Of course the amount of over- or underdamping depends on the proportion of total damping provided for by the slip.

[90] Available from Euro Precision Technology, P.O. Box 126, Tylersport, PA 18971, or P.O. Box 11072, Rotterdam 3004 EB, The Netherlands.

[91] H. Deresiewicz, "Oblique Contact of Nonspherical Bodies," J. Appl. Mech., Vol. 24, 1957, pp. 623–64.

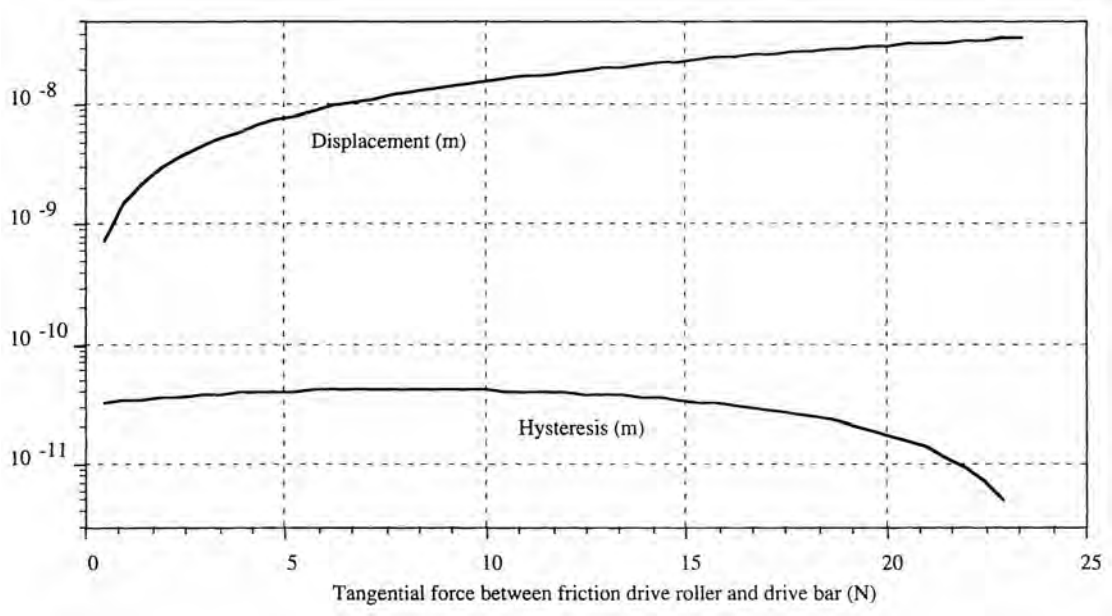


Figure 10.8.4 Load/displacement curve for the 50 mm diameter friction drive roller of Figure 10.8.3. The carriage mass was 100 kg, the time over which the force was applied was 0.1 s, and the maximum velocity was 11 mm/s.

$F_{tan}/\mu F_n$	Average damping (N/m/s)
1.0000E-04	1.9645E-06
2.0000E-04	3.9292E-06
4.0000E-04	7.8593E-06
6.0000E-04	1.1790E-05
8.0000E-04	1.5722E-05
1.0000E-03	1.9655E-05
2.0000E-03	3.9335E-05
4.0000E-03	7.8766E-05
6.0000E-03	1.1829E-04
8.0000E-03	1.5792E-04
1.0000E-02	1.9764E-04

Figure 10.8.5 Effect of velocity on average damping for the 50 mm diameter roller friction of Figure 10.8.3. The carriage mass was 100 kg, the time over which the force was applied was 0.1 s, and the maximum velocity was 11 mm/s.

Usually, most of the damping will be provided by the controller (i.e., the velocity term for a position loop or the acceleration term for a velocity loop), so the effect of the mechanical damping will probably only be of significance on the nanometer level.

In order to investigate the general effects of velocity on damping, the slip is assumed to be equal to the product of the displacement determined by Equation 5.6.27 and $F_{tan}/\mu F$, which provides a lower estimate of the amount of slip than would be obtained if a fractional power of $F_{tan}/\mu F$ were used. Finite element analysis or experiments could be used to determine the proper exponent. The incremental power dissipated is the product of the incremental change in slip displacement with the force. The damping effect felt by the control is the incremental power dissipation divided by the velocity squared. Figure 10.8.6 show the results of these calculations for the 50 mm diameter roller friction drive of Figure 10.8.3.

Because the damping is a function of velocity, with a simple control system there will be an error in the calculated velocity if it is assumed that the damping is constant. This error is on the order of parts per billion and thus may be insignificant; however, the error may become important if the friction drive is used on a machine where slow contouring cuts are made and nanometer accuracy is desired. In this case, a digital control algorithm with good filters may be needed to determine higher

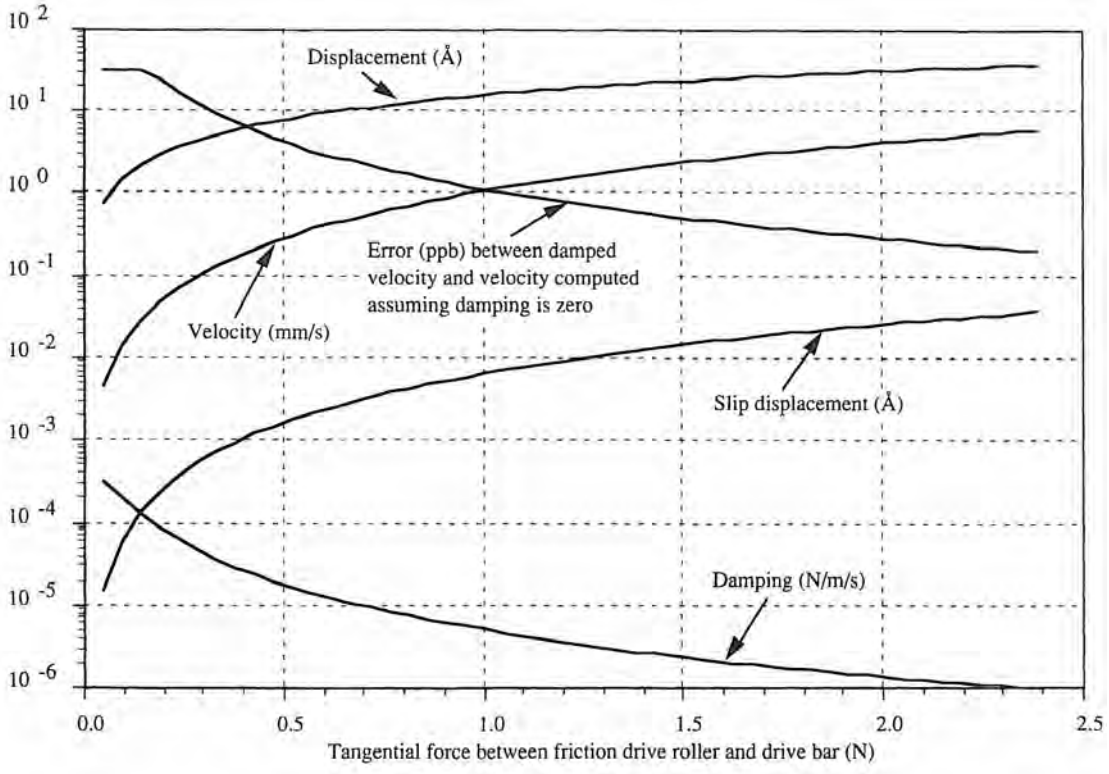


Figure 10.8.6 Energy dissipation factors for the 50 mm diameter roller friction drive of Figure 10.8.3. The carriage mass was 100 kg and the time over which the force was applied was 0.05 s.

order derivatives (e.g., acceleration and jerk) which would allow for very accurate velocity control in the presence of the nonlinearities caused by nanoslip. It may even be necessary to make the control coefficients functions of the velocity.

10.8.3 Leadscrews

The principle of a leadscrew and nut has been used for centuries to provide a means for converting rotary motion into linear motion. By turning a leadscrew and holding a nut so that it does not rotate, the nut moves along the length of the leadscrew. Alternatively, the shaft can be held and the nut turned. This provides an effective means for attaining linear motion that has been used in countless machines. The introduction and the continued success of the leadscrew is due to the fundamental fact that rotary motion motors are easier to produce and are often more efficient than linear motion motors.

The first known application of a screw thread to do useful work was perhaps Archimedes' screw pump, which converted rotary power of a screw into an elevator for raising water from a river to an irrigation ditch. The first leadscrew cutting lathes were introduced in the fifteenth century and were used to manufacture wooden screws. Wooden screws led to metal screws. Screwthreads could be increased in accuracy with hand finishing and it was just a matter of time before they developed into useful tools for metalworking.

It was found that even though leadscrews were prone to manufacturing errors, the effect of many threads in a nut that was made somewhat compliant (e.g., leather threads), and forced to engage the leadscrew simultaneously, caused some of the errors to average out. Thus, by snugly fitting a nut to a leadscrew and wearing it in, the accuracy to which leadscrews could be manufactured steadily increased. Each more accurate screw was used to make an even more accurate screw. As early as 1800, Henry Maudslay was credited with developing a leadscrew with four threads per millimeter and an accuracy of 25 μm. In 1855, Joseph Whitworth developed a leadscrew-driven machine that

could compare differences in the size of parts to within 1 μin. Eventually averaging reached its limits, and elaborate mechanical corrector cams evolved to correct lead errors. Fortunately, with modern sensor and servo systems, corrector cam mechanisms are a thing of the past.

In the remainder of this section, the basic physics of operation common to most leadscrews will be discussed in detail. The results shed insight to the operating properties of perhaps the most common actuators in use today on precision machine tools. With these results one can determine not only the drive torque, but the efficiency and the magnitude of "noise" moments created in the leadscrew nut, which can cause small pitch and yaw errors in a leadscrew-driven carriage.

There are many types of leadscrews that are available including:

- Sliding contact thread leadscrews
- Traction-drive leadscrews
- Oscillatory motion leadscrews
- Nonrecirculating rolling element leadscrews
- Planetary roller leadscrews
- Ballscrews
- Hydrostatic leadscrews

10.8.3.1 Leadscrew Operating Principles

A leadscrew is one of the simplest and best known power transmission elements. In this section the following will be discussed:

- Error sources
- Force and moment analysis
- Backdriveability
- Efficiency
- Noise moments

Where applicable, later in the discussion of specific types of leadscrews additional comments will be made. Note that the stiffness is dependent on the type of thread interface (e.g., rolling or sliding) and is discussed in the context of each type of leadscrew.

Error Sources

All leadscrews may be subject to many different types of error sources, including[92]:

1. Lack of squareness between the thrust collar and the thrust bearing will produce a periodic error in the system.
2. Eccentricity of the support journals and the screw shaft will cause periodic errors. Unless the nut is correctly coupled to the carriage, journal eccentricity will also cause straightness and angular errors in the carriage motion.
3. Lateral and angular misalignment between the screw and nut, and nut and carriage causes straightness and angular errors in the carriage motion as well as small periodic errors. In addition, the straightness of the leadscrew shaft and the amount it bends under its own weight will affect carriage motion accuracy.[93]
4. A varying pitch diameter will cause periodic errors and can contribute to backlash.
5. Mating thread form profile errors will cause periodic errors, backlash, and limit resolution.
6. Thread form errors (drunkenness) cause periodic errors. In the case of a multistart thread, which is used to increase the load capacity but does not change the transmission ratio, the relation between the threads' relative angular orientation also causes periodic errors.
7. Journal support bearings can be sources of periodic error, lateral motion, and backlash.

The net result of many of these types of error sources can be seen in Figure 6.4.6. Periodic errors can readily be mapped or obviated through the use of linear position sensors. Resolution can be increased by polishing components. Backlash, however, continues to be a difficult problem to deal with. One of the most common methods for dealing with backlash is to use two nuts that are preloaded against each other. With sliding contact acme threads, the nut can be split and circumferentially clamped so that the threads are wedged into each other. With ballscrews, oversized balls can be used, but this leads to increased rolling friction, because the groove shape is a Gothic arch and

[92] Remember, choosing a component is like farming. No matter how good the component (or crop), it will always be subject to various types of bugs.

[93] See Section 2.5 for a detailed discussion of the effects of forced geometric congruence.

four-point contact results. With care all errors can be dealt with; the task is to identify the source of error. In addition, as discussed below, in generating an axial force, a leadscrew also generates moments about axes orthogonal to the shaft, thereby creating other forms of error. When selecting a leadscrew and incorporating it into the design of a precision machine, one must be very careful to consider all of these effects. *Caveat emptor!*

Force and Moment Analysis[94]

The mechanical advantage provided by a leadscrew is a function of the lead and efficiency. The lead l of a leadscrew thread is defined as the linear distance the nut travels in one revolution of the nut relative to the shaft. For a rotation angle ϕ (in radians), the distance x traveled is

$$x = \frac{\ell\phi}{2\pi} \tag{10.8.1}$$

The lead angle θ is found by unwrapping a single turn of the thread which rises a distance equal to the lead l, at a pitch radius R:

$$\theta = \tan^{-1}\left(\frac{\ell}{2\pi R}\right) \tag{10.8.2}$$

It is assumed here that the thread depth is small compared to the pitch radius, so the lead angle is essentially constant. For deep thread screws, θ is a function of R and the analysis becomes more complicated, but of the same form, as follows.

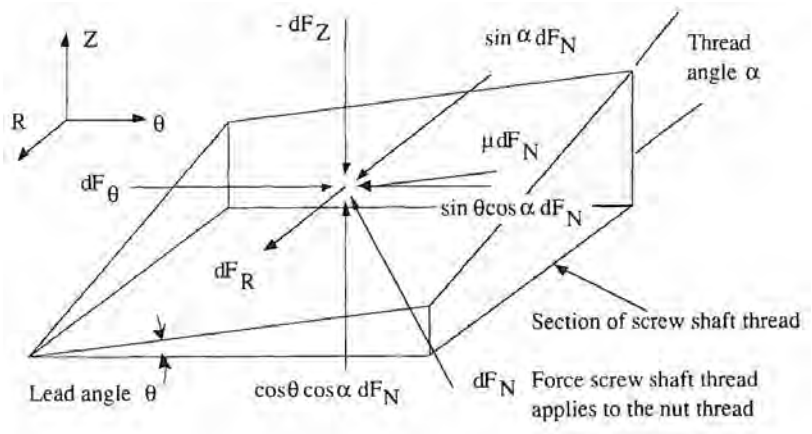


Figure 10.8.7 Forces a section of leadscrew thread apply to the nut thread when lifting (working against) a load. The thread normal force has been decomposed into its components.

Figure 10.8.7 shows a section of a leadscrew thread that is being used to lift (work against) a load. The relation between the differential axial force dF_Z and the differential circumferential and radial forces are found from a summation of the forces at a point:

$$dF_\theta = dF_Z\left\{\frac{\ell\cos\alpha + 2\pi R\mu}{2\pi R\cos\alpha - \mu\ell}\right\} \tag{10.8.3}$$

$$dF_R = \frac{-dF_Z\sin\alpha}{\cos\alpha\cos\theta - \mu\sin\theta} \tag{10.8.4}$$

For the case where the screw is used to lower a load

$$dF_\theta = dF_Z\left\{\frac{2\pi R_\mu - \ell\cos\alpha}{2\pi R\cos\alpha + \mu\ell}\right\} \tag{10.8.5}$$

$$dF_R = \frac{-dF_Z\sin\alpha}{\cos\alpha\cos\theta + \mu\sin\theta} \tag{10.8.6}$$

94 "*Mathematics is the alphabet with which God has written the universe.*" Galileo Galilei

The presence of the radial force is a hint that one must look at what else is happening to the nut in addition to the usual calculation for the required drive torque.

Since the depth of thread is assumed small compared to the diameter of the screw, the axial force can be considered to be distributed around a helical line through a thread contact angle Ψ. The differential axial force dF_Z is thus

$$dF_Z = \frac{F_Z d\Psi}{\Psi} \tag{10.8.7}$$

For a right-hand coordinate system superimposed at the center of the leadscrew shaft, the Cartesian coordinates of any point on the helix at an angle Ψ will be

$$X = R\cos\Psi \qquad Y = R\sin\Psi \qquad Z = \frac{\Psi\ell}{2\pi} \tag{10.8.8}$$

We are interested in the case where gravity does not help the leadscrew to move the load (i.e., *raising* a load versus *lowering* a load). For convenience, let the following constants be introduced:

$$C_{\theta R} = \frac{\ell\cos\alpha + 2\pi R\mu}{2\pi R\cos\alpha - \mu\ell} \qquad C_{RR} = \frac{-\sin\alpha}{\cos\alpha\cos\theta - \mu\sin\theta} \tag{10.8.9}$$

For the case of a load being raised, the differential X and Y force components are, respectively,

$$dF_X = dF_Z\,\{C_{RR}\cos\Psi - C_{\theta R}\sin\Psi\} \tag{10.8.10}$$

$$dF_Y = dF_Z\,\{C_{RR}\sin\Psi - C_{\theta R}\cos\Psi\} \tag{10.8.11}$$

Integrating the differential moments about the X, Y, and Z axes, caused by the differential forces over the entire helix angle Ψ, yields the moments imposed on the screw shaft while raising a load:

$$M_X = F_Z\left\{\frac{\ell}{2\pi}\left[C_{RR}\left(\cos\Psi - \frac{\sin\Psi}{\Psi}\right) + C_{\theta R}\left(\frac{1-\cos\Psi}{\Psi} - \sin\Psi\right)\right] + \frac{R(1-\cos\Psi)}{\Psi}\right\} \tag{10.8.12}$$

$$M_Y = F_Z\left\{\frac{\ell}{2\pi}\left[C_{\theta R}\left(\cos\Psi - \frac{\sin\Psi}{\Psi}\right) + C_{RR}\left(\sin\Psi + \frac{\cos\Psi - 1}{\Psi}\right)\right] - \frac{R\sin\Psi}{\Psi}\right\} \tag{10.8.13}$$

$$M_Z = F_Z\,C_{\theta R}\,R = F_Z R\left\{\frac{\ell\cos\alpha + 2\pi R\mu}{2\pi R\cos\alpha - \mu\ell}\right\} \tag{10.8.14}$$

M_Z is the torque required to turn the leadscrew shaft. Some leadscrews have multistart threads (i.e., more than one thread helix), but in most cases where the engagement length of each helix is equal, this would not change the moment equations. The purpose of multistart threads is to distribute the load over more surface area so that the screw can carry more load. This is particularly true when high leads are required and the nut length must be kept reasonable. Note that if the thread depth was large, the differential force would have the form

$$dF_Z = \frac{F_Z d\Psi\, dR}{(R_o - R_i)\Psi} \tag{10.8.15}$$

Equations 10.8.8-10.8.15 would have to be expanded to include the relation between θ and R given in Equation 10.8.2. The moment equations would then be of the form $dM = dF_z\{\ \}$. The total moments could then finally be obtained by integrating over the height of the thread. This is a straightforward but tedious process that produces equations too long to print here, and they are rarely needed. Note that for the case of a common ballscrew, the contact is essentially at a constant radius.

A similar analysis can be performed for the case of lowering a load. The torque required to lower the load is

$$M_Z = F_Z\,R\left\{\frac{2\pi R\mu - \ell\cos\alpha}{2\pi R\cos\alpha + \mu\ell}\right\} \tag{10.8.16}$$

There are three interesting results from this analysis that are discussed in detail below.

Backdriveability

From Equation 10.8.16, it can be seen that if the lead satisfies the following relationship, no torque will be required to prevent the screw from turning regardless of the magnitude of the axial load:

$$\ell \leq \frac{2\pi R\mu}{\cos\alpha} \tag{10.8.17}$$

A leadscrew which satisfies this condition is known as *non-backdriveable* or *self-locking*. For industrial leadscrew applications, backdriveability is often an important feature because it minimizes the holding power and brake size required. For precision servo-controlled leadscrews it is not really an important requirement although it does help to prevent force perturbations from being reflected back through the drivetrain.

Efficiency

The efficiency e is defined as the actual work divided by the ideal work ($\mu = 0$). Setting $\beta = 2R/\ell$, the worst-case efficiency occurs when lifting a load:

$$e = \frac{\cos\alpha\,(\pi\beta\cos\alpha - \mu)}{\pi\beta\cos\alpha(\cos\alpha + \pi\beta\mu)} \tag{10.8.18}$$

Figure 10.8.8 shows efficiencies for various types of screws with different diameter/lead ratios and thread angles α as defined in Figure 10.8.8. Note that a standard Acme thread has an angle between the threads of 29° or α=14.5°. The coefficient of friction is difficult to predict, but one can obtain reasonable estimates for the purposes of choosing the type of leadscrew and sizing the drive motor.

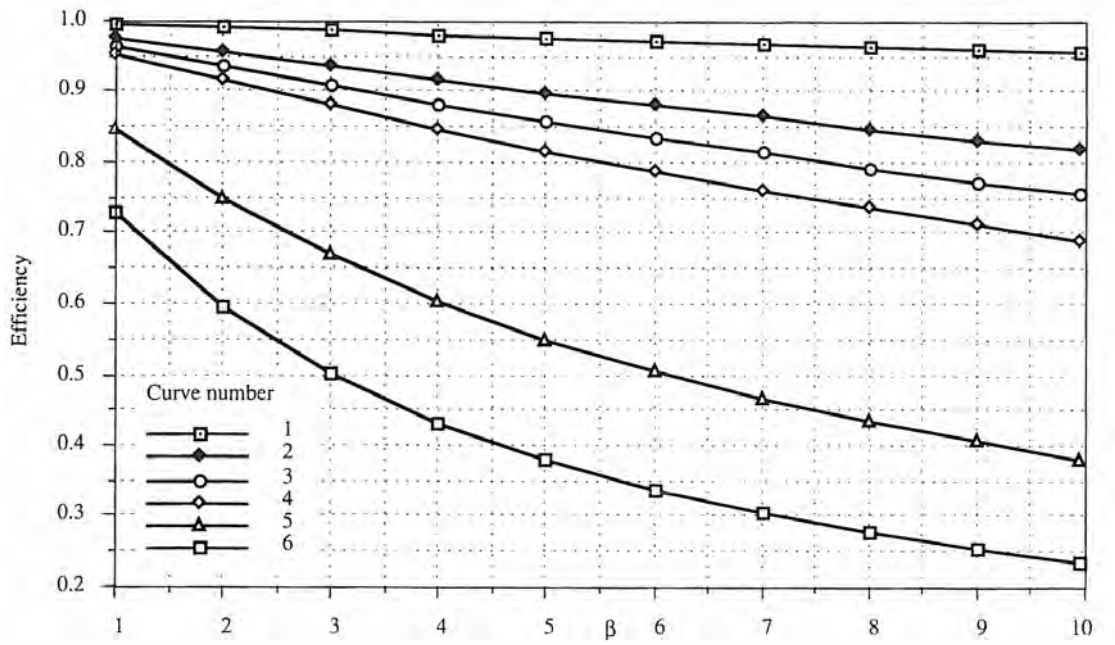


Figure 10.8.8 Leadscrew efficiency: (1) light preload special finish ballscrew α = 45° and μ = 0.001, (2) light preload ballscrew α = 45° and μ = 0.005, (3) lubricated lapped lightly loaded Acme thread α = 14.5° and μ = 0.01, (4) heavy preload ballscrew α = 45° and μ = 0.01, (5) lubricated ground Acme thread α = 14.5° and μ = 0.05, (6) lubricated tapped Acme thread α = 14.5° and μ = 0.1.

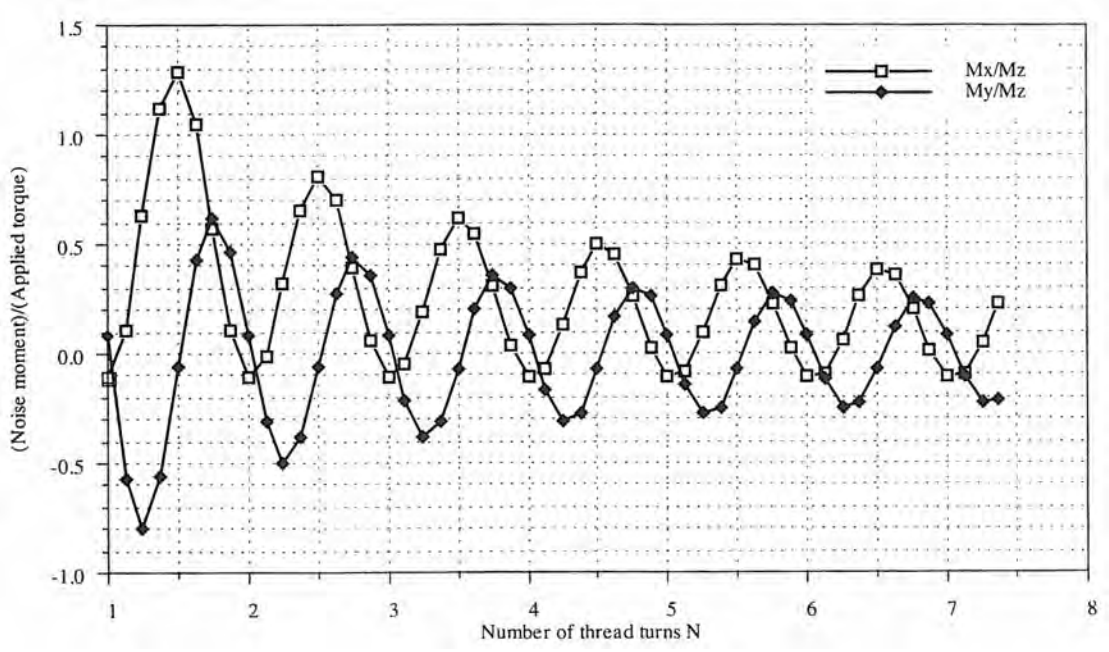


Figure 10.8.9 Noise moments on a leadscrew nut with 14.5° thread angle, coefficient of friction = 0.1, lead = 10 mm, and diameter = 40 mm.

Noise Moments

The analysis above showed that in the process of creating an axial force with a moment about the leadscrew axis (shaft torque), off-axis moments are also created about axes normal to the shaft. These moments are referred to here as *noise moments*[95] not because they are caused by random effects (they are not) but because they are unwanted resultant moments. These moments are plotted in Figure 10.8.9 for a lapped leadscrew with forced light oil lubrication. Similar results are obtained for other types of leadscrews.

Figure 10.8.9 shows the noise moment functions to be decaying sinusoids. The magnitude of the sinusoid can be large and thus requires extreme care on the part of the leadscrew manufacturer to choose a helix angle to minimize the two moments. The sinusoids are out of phase so there is no point where both moments are zero. As shown in Figure 10.8.10, the rms of the noise moments is minimized at an integer number of turns of the helix angle. This also causes the net lateral forces to be zero. A problem arises when manufacturing errors effectively change the amount of contact between the nut and screw and hence the magnitude of the noise moment. As can be seen from the graph, a small variation in number of effective thread turns can sometimes mean a large increase in noise moments. Although the sinusoids are decaying, there is no point where the M_X and M_Y moments are both zero. The more thread turns that are used, however, the less the chance of a variance in contact area, causing a large disturbance. The noise moments are why no leadscrew can be directly bolted to a carriage without causing pitch and yaw errors, even though in most cases they are negligible, to be imposed on the carriage's motion. In many cases the noise moments will seem to vary randomly as the process of varying contact zones may itself be random. In other cases the noise moments may vary sinusoidally as balls enter and leave a ballscrew raceway. The noise moments also contribute to the wear of the leadscrew.

The minimum rms noise moment ratio is independent of the number of thread turns; however, the fewer thread turns, the steeper the rms noise moment slope, so if the exact desired number of thread turns is not achieved, the consequences will be greater. Figure 10.8.11 shows the effect of the coefficient of friction on the minimum rms noise moment ratio. In the limit when μ is infinite, a torque applied to the shaft only produces a torque on the nut. This helps to explain why sliding contact leadscrews are quieter, in addition to their lack of rolling balls, than ballscrews. Figure

[95] "Noise moment" is not a standard term, so sometimes you may have to explain that they are moments about axes orthogonal to the screw axes. Noise moment *is* a simple term that describes what they are to the design engineer.

10.8.12 shows that a decreasing lead also yields a quieter leadscrew. In the limit when β goes to infinity, a rotation applied to the shaft yields no linear motion and hence no noise moments. Figure 10.8.13 shows the effect of varying thread flank angle on the minimum rms noise moment ratio. The ideal threadform would have square threads ($\alpha = 0$). Note that ballscrews, which have a large contact angle of 45°, thus typically have large noise moment ratios.

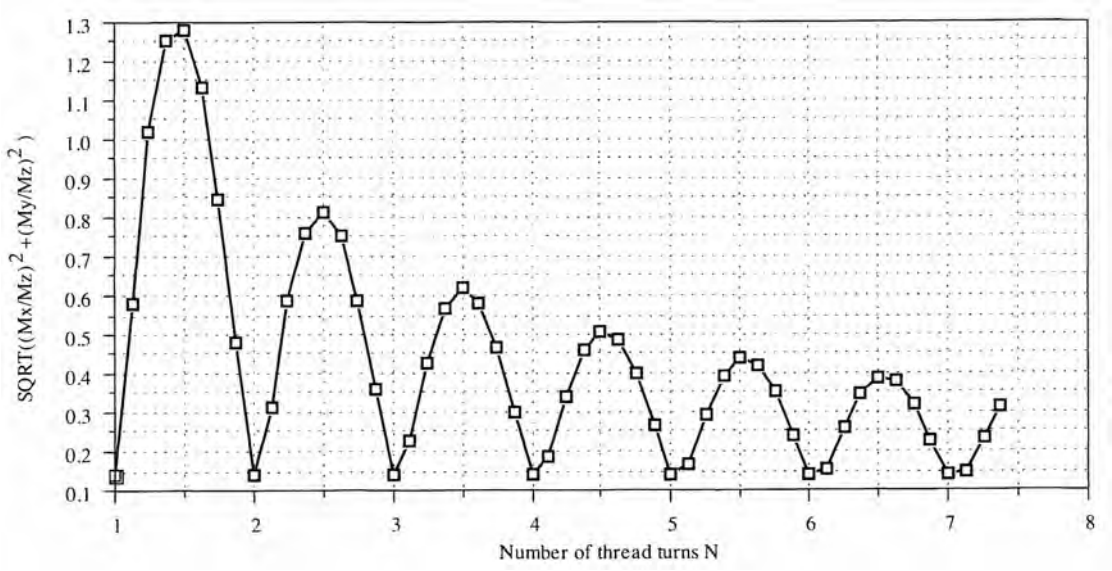


Figure 10.8.10 Root mean square of noise moments in Figure 10.8.9

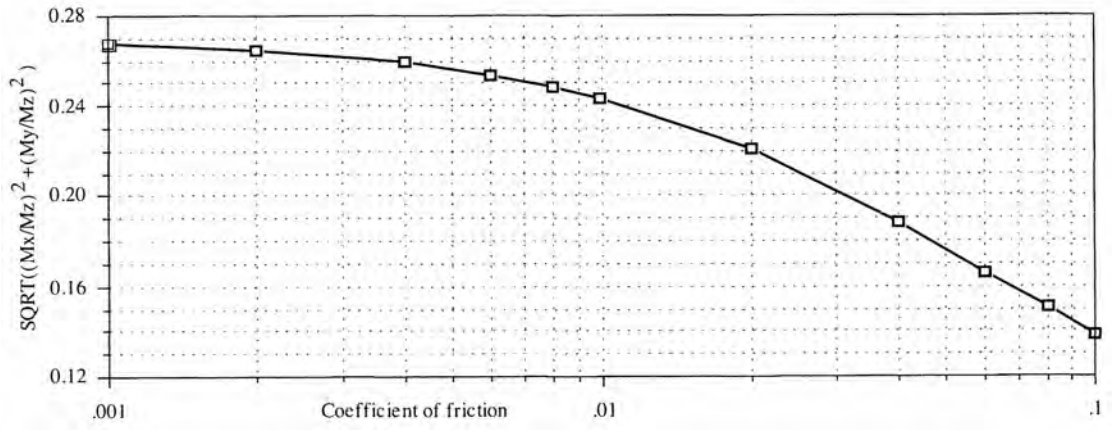


Figure 10.8.11 Effect of coefficient of friction on minimum rms noise moments for a 40 mm diameter, 10 mm lead leadscrew with 14.5° thread angle α.

The maximum rms noise moment decays asymptotically with the number of thread turns. Figure 10.8.14 shows the effect of the diameter-to-lead ratio β on the maximum rms noise moment ratio. Changing β changes the lead angle, which affects the maximum noise moment ratio in a nonlinear manner. Still for a large number of thread turns, the noise moment ratio decreases with increasing β. For a small number of thread turns, one wants β to be small to avoid the possibility of experiencing large maximum noise moments if the exact desired number of thread turns is not achieved. Figure 10.8.15 shows the effect of the coefficient of friction on the maximum rms noise moment ratio. Once again, the higher the coefficient of friction, the lower the noise moments. This is not to say that one should purposefully increase the coefficient of friction, because that could lead to other control problems. Figure 10.8.16 shows that the maximum rms noise moments decrease with decreasing thread angle α and increasing number of thread turns.

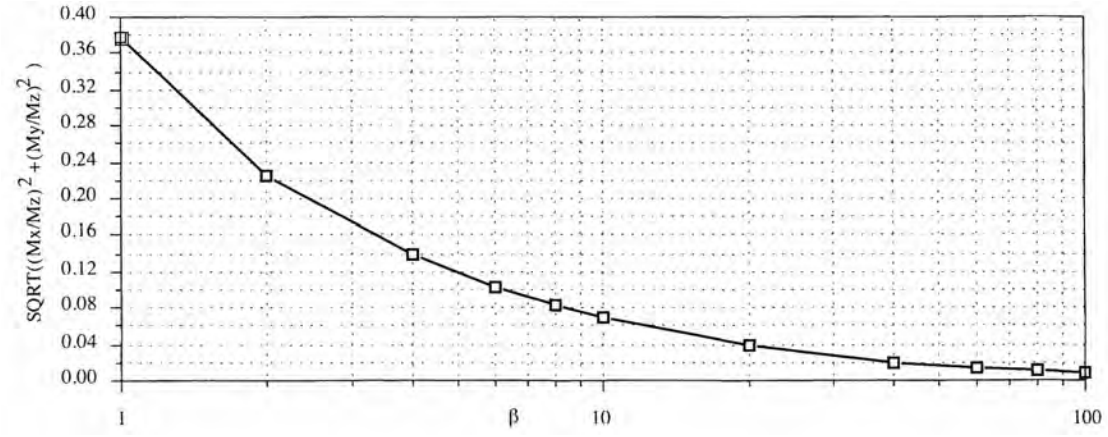


Figure 10.8.12 Effect of diameter-to-lead ratio β on minimum rms noise moments for a leadscrew with coefficient of friction of 0.1 and a 14.5° thread angle α.

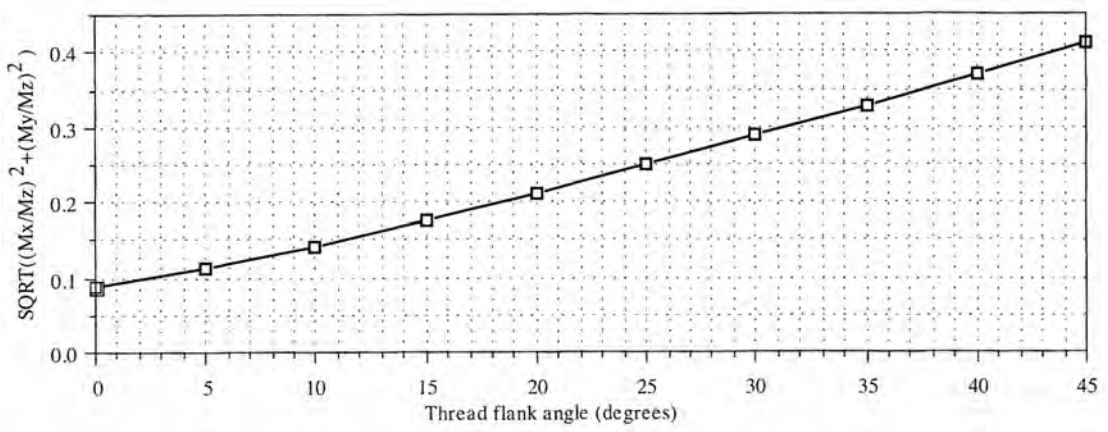


Figure 10.8.13 Effect of thread angle α on minimum rms noise moments for a leadscrew with coefficient of friction of 0.1, 40 mm diameter, and 10 mm lead.

A high-quality leadscrew will have minimal noise moment variations. This is often achieved (in general) by increasing the number of thread turns, which also increases load capacity and stiffness. Increasing the number of thread turns decreases the slope of the noise moment curve, so if the exact number of desired turns is not achieved (and the number may oscillate slightly when balls are used), the effects will be minimized. Unfortunately, many manufacturers are not aware of the existence of noise moments because noise moments usually only manifest themselves as problems when submicron performance is sought. Some manufacturers will provide noise moment data if they have ever bothered to make the measurements. In many designs, such as in the design of a turning center with 1 μm resolution, the noise moments create insignificant errors. In other cases, such as a diamond turning machine, the error due to noise moments can be significant. Sometimes one just has to buy a whole box of leadscrews and test each one until one is found that meets the desired specifications.

As discussed in Section 10.8.4, there are various types of coupling systems one can use to help minimize the magnitudes of loads imposed on the carriage driven by a leadscrew. Ideally, a leadscrew would have no mechanical contact between the threads. This would give it a much lower noise moment ratio and no variation in contact area; thus it would have very consistent performance, which could readily be mapped. This type of leadscrew is discussed in Section 10.8.3.8.

Application of the Analysis

The drive torque and efficiency results of this analysis are within engineering tolerance for use with most types of leadscrews. The results for noise moments can be used to provide an estimate

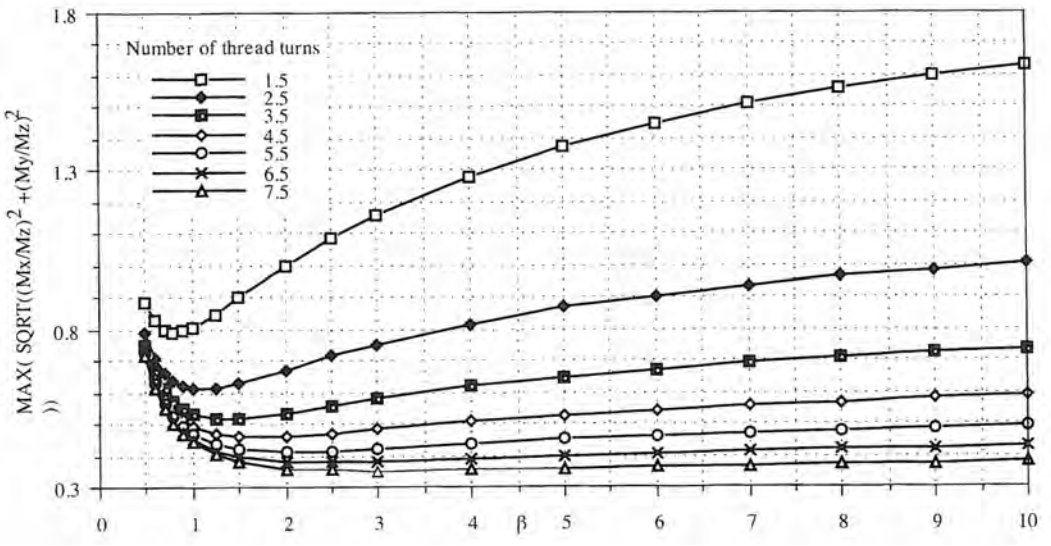


Figure 10.8.14 Effect of diameter-to-lead ratio β on maximum rms noise moment envelope for a leadscrew with coefficient of friction of 0.1 and a 14.5° thread angle α.

of the noise moments that will be generated. These estimates can then be used in the machine's error budget to help guide the sizing of bearings and carriage structural components. As will be seen below, some screws do not have continuous contact around the thread helix. These screws would require individual summation of the forces at the contact points. For example, a ballscrew manufacturer would use a computer to provide a summation of the differential moments caused by all the balls contacting the thread surfaces. This would allow the manufacturer to choose just the right number of balls for the recirculating path to carry the desired proper load and to minimize noise moments.

The efficiency (Equation 10.8.18) can be used to obtain a much simpler equation for the relation between drive torque, axial force, and lead. The rotary power into the system is the product of the drive motor torque M_Z and the rotation angle θ. This power equals the linear power out, which is the product of the force F_Z generated, the axial distance moved, and the efficiency of the system. The force generated by a leadscrew can therefore be expressed as

$$F_Z = \frac{2\pi e\, M_Z}{e} \tag{10.8.19}$$

If a leadscrew has a fine lead, a low-torque, high-speed motor can be used to drive the system. In addition, if the lead is accurate, a lower-resolution sensor can be used to measure the rotation of the screw. Until recently, mechanical systems were more accurate than electronic systems. Today, the motivation for using fine leads is to minimize drive-torque requirements and to allow for the use of stepper motors.

With the relations above, it is also not difficult to show that the equivalent linear stiffness of a rotary stiffness reflected through a leadscrew is

$$K_{\text{linear equivalent}} = \frac{4\pi^2 K_{\text{rotary}}}{e^2} \tag{10.8.20}$$

As was shown in Section 5.2, the axial stiffness of a leadscrew is almost always (except for screws with very large leads) much less than the equivalent linear stiffness of the rotary stiffness. Thus a leadscrew can increase the apparent stiffness of an electric motor. This is one reason why linear electric motors may never replace leadscrews in machine tool applications. The same magnet technology that advances the state of the art in linear motors can always be rolled into a circle and used with a leadscrew to achieve a higher performance level.

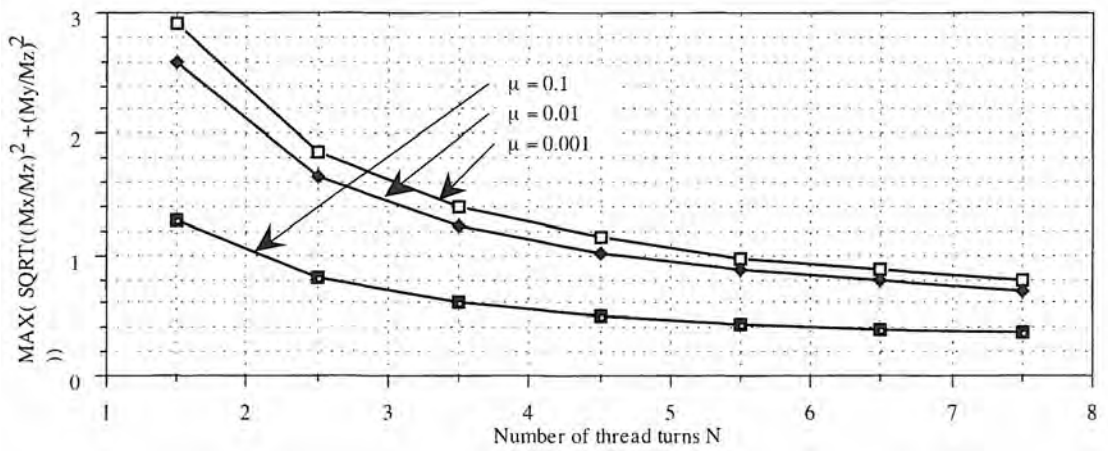


Figure 10.8.15 Effect of coefficient of friction on maximum rms noise moment envelope for a 40 mm diameter, 10 mm lead leadscrew with 14.5° thread angle α.

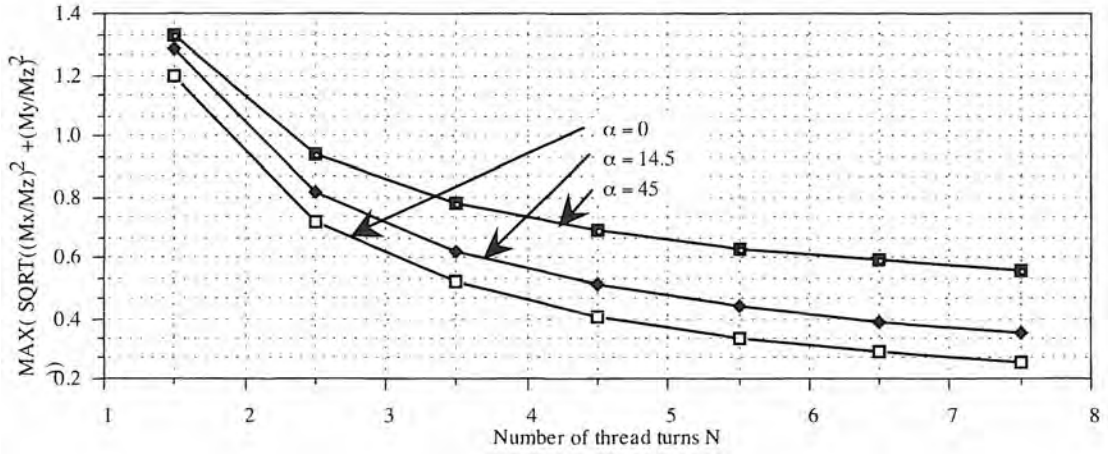


Figure 10.8.16 Effect of thread angle α on maximum rms noise moment envelope for a leadscrew with coefficient of friction of 0.1, 40 mm diameter, and 10 mm lead.

10.8.3.2 Sliding Contact Thread Leadscrews

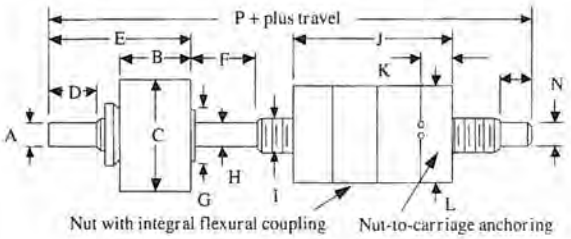
Sliding contact thread leadscrews represent the range from least expensive (machine finished) to most expensive (hand lapped) leadscrews. Usually, the nut is made of a bearing brass or bronze but can also be made from PTFE. The nut can also be formed by casting a polymer of babbit bearing material around the shaft. For low force applications (e.g., for instrument carriages) the nut can be bored without threads and then have axial slits cut into it. The nut is then placed over a fine pitch leadscrew (e.g., 100 threads per inch) and O rings used to clamp the nut circumferentially. The fine thread screw will then make its own impressions into the nut.

Molded plastic nuts are often split and preloaded by an O ring, which puts circumferential pressure on the nut. A molded plastic nut running on a rolled shaft may have an accuracy of about about 1 mm/m. Molded plastic nut leadscrew assemblies can have leads as high as four times the pitch diameter of the screw. They are typically used in applications where the loads and shaft diameters are low, less than 500 N and 20 mm, respectively. Molded plastic nut leadscrew assemblies may only cost on the order of $10-$100.

Commercially available thread ground and lapped sliding contact thread leadscrew assemblies may have nuts preloaded against each other or they may have split nuts that are preloaded with a circumferential spring. Some even have built-in flexural couplings. Figure 10.8.17 shows size and load capacity data for a commercially available ground and lapped leadscrew with an integral flexible coupling in the nut. These ground and lapped leadscrews are comparably in cost to precision

ballscrews. One of the larger units with X accuracy grade may cost about $1800. An XXX accuracy grade may cost three times this amount. With a closed-loop servo-control system that uses a linear position sensor, there is generally no need for specifying an accuracy grade higher than X.

The coefficient of friction between a sliding thread contact leadscrew can range from 0.1 for a greased nut to 0.01-0.05 for a lightly loaded lapped thread with forced lubrication. Load capacity is an order of magnitude less than for a ballscrew. However, the lapped continuous contact thread provides much greater smoothness of motion. Smoothness of motion between the threads easily allows the ground and lapped screw to achieve submicron resolutions once initial stick-slip has been overcome. To achieve this high resolution from a ballscrew requires great care and testing of many screws, which would make them as expensive as ground and lapped thread leadscrews. Smoothness of motion also implies very repeatable and minimal noise moments. For the leadscrew shown, these moments are generally born by the leadscrew shaft and prevented from being transmitted with any appreciable magnitude to the carriage by the flexible coupling.



Nom. Diam.	1/4	1/2	3/4
Max. travel	6	20	30
Stand. leads	0.025, 0.050, 0.100, 0.200 1mm, 2mm	0.025, 0.040, 0.050, 0.100 0.125, 0.200, 0.400, 0.800 1mm, 2mm, 3mm, 4mm, 5mm	0.025, 0.050, 0.100, 0.200 1mm, 2.5mm, 5mm
Max. friction torque	4 oz-in	8 oz-in	12 oz-in
Max. load	30 lbf	50 lbf	100 lbf
Working load	8 lbf	10 lbf	20 lbf
Nut stiffness	40,000 lbf/in	50,000 lbf/in	80,000 lbf/in
A	0.1870 - 1865	0.2495 - 0.2490	0.4995 - 0.4990
B	0.5512	0.7086	0.8662
C	0.8661	1.1811	1.3780
D	0.375	0.81	1.00
E	1.125	1.750	2.187
F	0.475	0.60	1.25
G	0.40	0.50	0.75
H	0.195	0.37	0.625
I	0.265	0.500	0.750
J	1.25	1.50	3.00
K	0.25	0.31	0.50
L	0.7495 - 0.7490	0.8745 - 0.8740	1.2495 - 1.2490
M	0.1966 - 0.1963	0.394 - 0.391	0.5903 - 0.5900
N	0.25	0.50	0.562
P	3.00	4.14	8.00

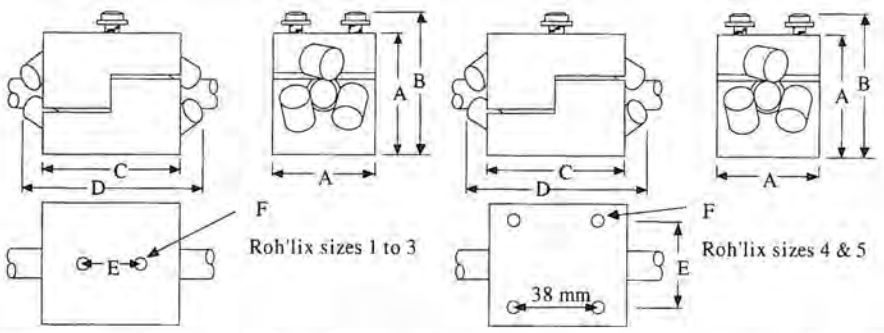
Figure 10.8.17 Characteristics of a commercially available lapped sliding contact leadscrew for precision instruments. (Courtesy of Universal Thread Grinding Co.)

10.8.3.3 Traction Drive Leadscrews

A traction drive leadscrew establishes rolling contact between the nut and a smooth shaft with the use of cam rollers. A commercial leadscrew of this type is shown in Figure 10.8.18. The cam rollers' axes are inclined to the axis of the shaft. The angle of inclination determines the lead. The efficiency of this type of screw is generally on the order of 0.9 (90%) and load capacity is moderate. If overloaded, the nut will slide along the shaft, but this can mar the shaft and is undesirable from an accuracy standpoint.

The preload force is established by preloading the top roller against the lower two rollers. Due to manufacturing tolerances in the positioning of the rollers in the nut and the structure of the cam rollers, backlash is on the order of 20-30 μm. The commercial grade of this actuator is intended for

applications requiring moderate accuracy and load capability with high efficiency and low cost. One big advantage of this type of leadscrew is that the shaft is smooth and round, so it is exceptionally easy to seal. Traction-drive leadscrews are low in cost and have wide applicability in many types of industrial packaging and transfer line equipment.



Size	Model	Shaft Diam. (mm)	Lead (mm)	Thrust (N)	A (mm)	B (mm)	C (mm)	D (mm)	E (mm)	F Tapped mount holes
1	1901	8	1.3	22	28.6	42.9	41.3	57.2	20.0	M3
1	1902	8	2.5	22	28.6	42.9	41.3	57.2	20.0	M3
2	2901	8	2.5	133	38.1	50.8	50.8	71.4	25.4	M5
2	2902	8	15.0	133	38.1	50.8	50.8	71.4	25.4	M5
2	2903	12	5.0	133	38.1	50.8	50.8	71.4	25.4	M5
2	2904	12	15.0	133	38.1	50.8	50.8	71.4	25.4	M5
2	2905	12	25.0	133	38.1	50.8	50.8	71.4	25.4	M5
3	3901	12	2.5	266	50.8	65.1	63.5	87.1	32.0	M6
3	3902	12	10	266	50.8	65.1	63.5	87.1	32.0	M6
3	3913	16	2.5	266	50.8	65.1	63.5	87.1	32.0	M6
3	3914	16	15.0	266	50.8	65.1	63.5	87.1	32.0	M6
3	3915	16	25.0	266	50.8	65.1	63.5	87.1	32.0	M6
4	4901	25	2.5	444	76.2	84.1	63.5	88.9	64.0	M6
4	4902	25	5.0	444	76.2	84.1	63.5	88.9	64.0	M6
4	4903	25	25.0	444	76.2	84.1	63.5	88.9	64.0	M6
5	5901	40	10.0	889	114.3	114.3	69.9	118.1	100.0	M6
5	5901	50	5.0	889	114.3	114.3	69.9	118.1	100.0	M6
5	5901	50	50.0	889	114.3	114.3	69.9	118.1	100.0	M6

Figure 10.8.18 Cam roller-type traction drive leadscrew. (Courtesy of Zero-Max Inc.)

10.8.3.4 Oscillatory Motion Leadscrews

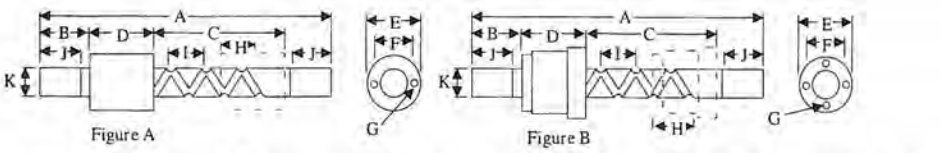
Many manufacturing applications require linear oscillatory motion over a fixed path. For this type of application it is often desirable to design a system with a simple analog velocity servo. The leadscrew shown in Figure 10.8.19 was designed for this type of application. Various turnaround curve profiles are available, which allows the dwell to be chosen for the application. This type of leadscrew is not intended for precision applications (submicron) but is a very useful industrial actuator worth noting. A typical application would be in photocopying machines.

10.8.3.5 Nonrecirculating Rolling Element Leadscrews

One of the primary features of the Rollnut® is that the rolling elements are fixed and can pass over discontinuities in the shaft. This means that a very long shaft can be spliced together and suspended by shaft hangers. The nonrotating nut can thus achieve travels of tens of meters without having to worry about shaft deflection and critical speeds except for regions between hangers. Figure 10.8.20 shows typically available sizes and capacities of Rollnut® actuators. This type of leadscrew is not intended for precision applications (submicron) but is a very useful industrial actuator worth noting.

10.8.3.6 Planetary Roller Leadscrews

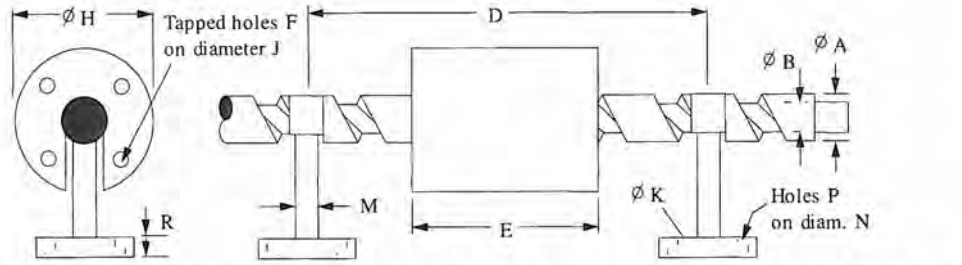
The planetary roller leadscrew (*rollerscrew*) shown in Figure 10.8.21a has a nut with as many thread starts (leads) as the leadscrew itself, which typically is from three to eight leads. The planetary



	1600 (Fig A)	1700 (Fig A)	1800 (Fig A)	1900 (Fig B)	2000 (Fig B)	2100 (Fig B)
K	9.53	12.7	19.05	31.75	44.45	63.50
I	12.70	19.05	25.40	31.75	50.80	76.20
B	38.10	50.80	99.06	154.94	198.12	198.12
J	-	-	108.70	152.40	177.80	203.20
D	35.05	47.50	65.02	104.90	127.00	171.45
E	24.89	30.99	41.15	60.20	82.55	113.54
H	18.80	25.40	34.04	57.66	64.77	89.41
F	19.81	23.88	33.27	49.28	68.33	95.25
G	M3	M4	M6	M8	M8	M12
$A_{min.}$	122	166	291	438	578	644
$A_{max.}$	427	623	900	1200	1797	2473
$C_{min.}$	11.2	16.6	22.3	28.5	44.5	63.5
$C_{max.}$	316	474	632	791	1264	1892
$Load_{max.}$	53	98	173 or 434	534 or 1254	1068 or 2638	1899 or 4893

Units are mm and Newtons

Figure 10.8.19 Characteristics of a leadscrew that provides oscillating motion. (Courtesy of Norco, Inc.)



Model	A	B	l_{Min}	l_{Max}	l_{std}	D	K	L	M	N	P	R	E	F	H	J	Dyn load	Static load
25L06	25	17.1	10	40	25	1000	50	55	12.7	35	7	6	205	M6	70	60	2000	3000
38L11	38	26.2	20	50	35	1200	80	70	20	60	9	12	292	M8	102	88	6000	9000
50L14	50	35.3	25	80	50	1500	80	90	20	60	9	15	350	M10	130	110	10000	15000
75L19	75	55.8	40	110	75	2000	90	120	25	70	12	15	505	M12	180	150	20000	30000

Units are mm and Newtons

Figure 10.8.20 Metric Rollnuts® for very long range of motion actuation. (Courtesy of Norco, Inc.)

rollers (*rollers*) have a single thread with a pitch equal to the apparent pitch (real pitch divided by the number of leads) of the leadscrew. The rollers mesh with both the screw and the nut and the rollers are spaced apart with spacer rings at their ends which act like the cage in a roller bearing. As the nut rotates relative to the leadscrew, the rollers rotate and orbit in the nut like rollers in a roller bearing. To eliminate any relative axial motion between the rollers and the nut, there are critical relations between the pitch diameters of the screw, rollers, and the nut:

Nut pitch diameter = leadscrew pitch diameter + 2 × roller pitch diameter
Roller pitch diameter = nut pitch diameter divided by the number of leads

When these relations are met, the advance of the rollers in the nut due to orbiting is canceled by their rotation. To compensate for minute inaccuracies in the actual relations between the pitch diameters, the ends of the rollers have external gears which mate with an internal gear in the nut. Rollerscrews are typically produced with a diameter ranging from 3.5 to 120 mm, a pitch ranging from 1 to 40 mm, and a single nut dynamic load capacity ranging from 5300 N to 753 kN. The nut can be made in one piece without preload, or in two pieces if preload is required. The maximum speed of a planetary

roller leadscrew is 6000 rpm, assuming that shaft whip, as discussed in Section 10.8.3.9, does not occur.

A recirculating rollerscrew is shown in Figure 10.8.21b. The leadscrew has one or two leads and the rollers have circular grooves. To eliminate the relative axial motion between the rollers and the nut, the nut is provided with a longitudinal recess enabling the rollers to disengage from both the screw and the nut. At this particular position, the rollers are brought back to their original position by a cam. This design does not require a special relationship between roller and nut diameters; thus it is possible to obtain up to six times finer pitch for a given screw diameter. However, this design limits the speed of rotation to about 1000 rpm because of the noise produced by the action of the cam.

With either design, the large number of contacting threads creates a large averaging effect and also gives rollerscrews load and stiffness capabilities many times higher than those of similar diameter ballscrews. In addition, the thread shape of the rollers can be optimized (e.g., a large radius of curvature) to minimize the contact stresses.[96] Rollerscrews are manufactured with accuracies on the order of ballscrews (see ballscrews below). Nonrecirculating rollerscrews are also much quieter than ballscrews, whose recirculating balls can make considerable audible noise at high speed. With a highly finished lubricated rollerscrew, the effective coefficient of friction between the threads can be on the order of 0.01. For a typical rollerscrew with $\beta = 4$, the efficiency is on the order of 88-90%. The diameter/lead ratio (β) is usually greater than 2 because the planetary rollers are of small diameter, and large leads would create helix angles that are too large. On the other hand, it is easier to have a very fine lead with a large-diameter rollerscrew than with a ballscrew. Because rollerscrews can have such small leads and large load capacities, they are well suited to achieving an optimal transmission ratio for heavily loaded systems. If shaft whip is not a concern, a rollerscrew shaft can be driven at up to three times the rpm of a ballscrew. For slow speed, high force, and stiffness applications, such as in creep feed grinders, rollerscrews are often preferred over ballscrews. Figure 10.8.22 shows typical characteristics of rollerscrew nuts. Depending on the screw, a rollerscrew may cost one to three times as much as a ballscrew.

10.8.3.7 Ballscrews

Ballscrews are perhaps the most common type of leadscrew used in industrial machinery and precision machines. Ballscrews can easily be used to achieve repeatability on the order of 1 μm, and specially manufactured and tested ballscrews can attain submicroinch motion resolution. Typical ballscrew designs are shown in Figure 10.8.23. Ballscrews achieve very high efficiency by using rolling steel balls with smaller spacer balls in between to transfer loads from the screw shaft to the nut threads. No rolling element is perfect and under load balls lose their sphericity; thus even ballscrews have finite efficiencies, as was shown in Figure 10.8.8.

The selection procedure for ballscrews is straightforward, as shown in Figure 10.8.24, and could be used as a model for the selection of other types of leadscrews as well. Figure 10.8.25 shows typical ballscrew applications as a function of accuracy grade. Many ballscrew manufacturers have computer programs to help guide the design engineer in selecting the proper ballscrew. Before talking to a manufacturer, however, the prudent designer will use the information in this text and information supplied in the manufacturer's catalog to select a ballscrew. This helps the design engineer gain experience, ask the manufacturer the right questions, and increase the chances of obtaining the correct ballscrew. Always remember, *caveat emptor!*

Accuracy

In addition to the factors generally affecting leadscrew accuracy discussed earlier, factors to consider when selecting a ballscrew include:

- Roundness and size uniformity of the balls
- Design of the recirculating entrance and exit paths
- Mechanical compensation for thermal expansion
- Lead accuracy
- Preload
- Mounting accuracy

[96] See, for example, P. Munn, "A Rollerscrew with Special Qualities," Proc. 22nd Int. Machine Tool Design & Research Conf, Sept. 1981.

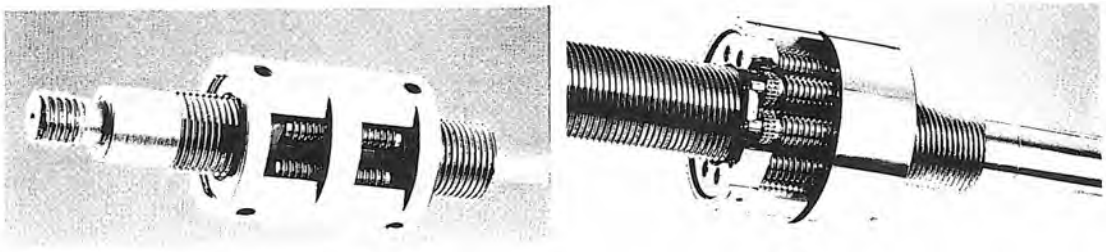
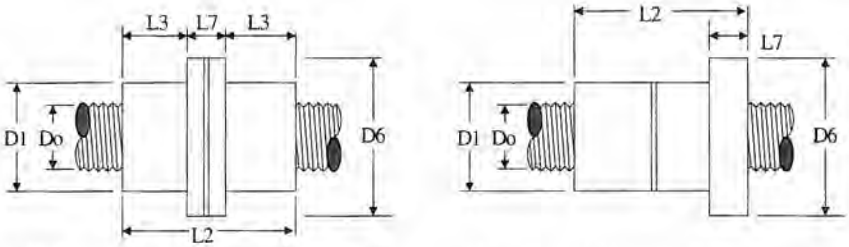


Figure 10.8.21 Construction of a planetary roller leadscrew. (Courtesy of ROLLVIS S.A., a member of the FAG-Group.)



Model	Lead (mm)	Do (mm)	D1 (mm)	D6 (mm)	L2 (mm)	L3 (mm)	L7 (mm)	C_{dyn}* (N)	C_{stat} (N)
RV5x1	1	5	19	39	41	14.0	13	4440	4960
RV8x1	1	8	21	41	41	14.0	13	4980	6200
RV8x2	2	8	21	41	41	14.0	13	4110	8780
RV8x3	3	8	21	41	41	14.0	13	3744	7910
RV8x4	4	8	21	41	41	14.0	13	3318	9135
RV8x5	5	8	21	41	41	14.0	13	2976	10215
RV12x1	2	12	30	46	41	14.0	13	5088	5705
RV12x2	2	12	26	46	41	14.0	13	6024	13100
RV12x4	4	12	26	46	41	14.0	13	5250	13600
RV12x5	5	12	26	46	41	14.0	13	4890	15200
RV15x2	2	15	34	56	51	16.5	18	7926	15560
RV15x4	4	15	34	56	51	16.5	18	7026	16180
RV15x5	5	15	34	56	51	16.5	18	6588	18090
RV18x5	5	18	34	56	51	16.5	18	7446	20595
RV18x10	10	18	34	56	51	16.5	18	6120	29125
RV20x2	2	20	42	64	65	22.5	20	19416	32355
RV20x4	4	20	42	64	65	22.5	20	17376	33645
RV20x5	5	20	42	64	65	22.5	20	16356	37615
RV23x2	2	23	45	67	65	22.5	20	22230	34755
RV23x4	4	23	45	67	65	22.5	20	20052	36140
RV23x5	5	23	45	67	65	22.5	20	18936	40405
RV27x2	2	27	53	83	69	23.5	22	26178	38070
RV27x4	4	27	53	83	69	23.5	22	23826	39590
RV27x5	5	27	53	83	69	23.5	22	22590	44260
RV30x4	4	30	62	92	69	23.5	22	26190	41730
RV30x5	5	30	62	92	69	23.5	22	24888	46655
RV30x10	10	30	62	92	69	23.5	22	20724	65980
RV36x5	5	36	74	110	84	29.5	25	38022	64300
RV36x10	10	36	74	110	84	29.5	25	31428	90935
RV39x5	5	39	80	116	84	29.5	25	40806	66925
RV39x10	10	39	80	116	84	29.5	25	34506	94645
RV48x10	10	48	86	122	104	39.5	25	60492	162275
RV48x12	12	48	86	122	104	39.5	25	57906	177765
RV60x6	6	60	110	150	124	47.0	30	101700	169610
RV60x8	8	60	110	150	124	47.0	30	96066	195850
RV60x10	10	60	110	150	124	47.0	30	91644	218965
RV60x12	12	60	110	150	124	47.0	30	87990	239865
RV80x6	6	80	138	180	158	61.5	35	178686	240615
RV80x8	8	80	138	180	158	61.5	35	169440	277840
RV80x10	10	80	138	180	158	61.5	35	162204	310635
RV80x12	12	80	138	180	158	61.5	35	156240	340280

* For a life of 10^6 revolutions. For single unpreloaded nuts, divide by 0.6 and 0.5 respectively.

Figure 10.8.22 Representative characteristics of preloaded flanged rollerscrew nuts. (Courtesy of ROLLVIS S.A., a member of the FAG-Group.)

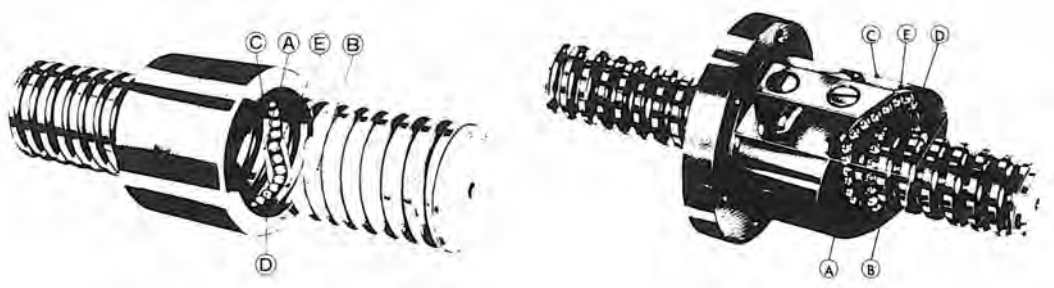


Figure 10.8.23 Return tube and internal deflector type ballscrew nuts. (Courtesy of NSK Corp.)

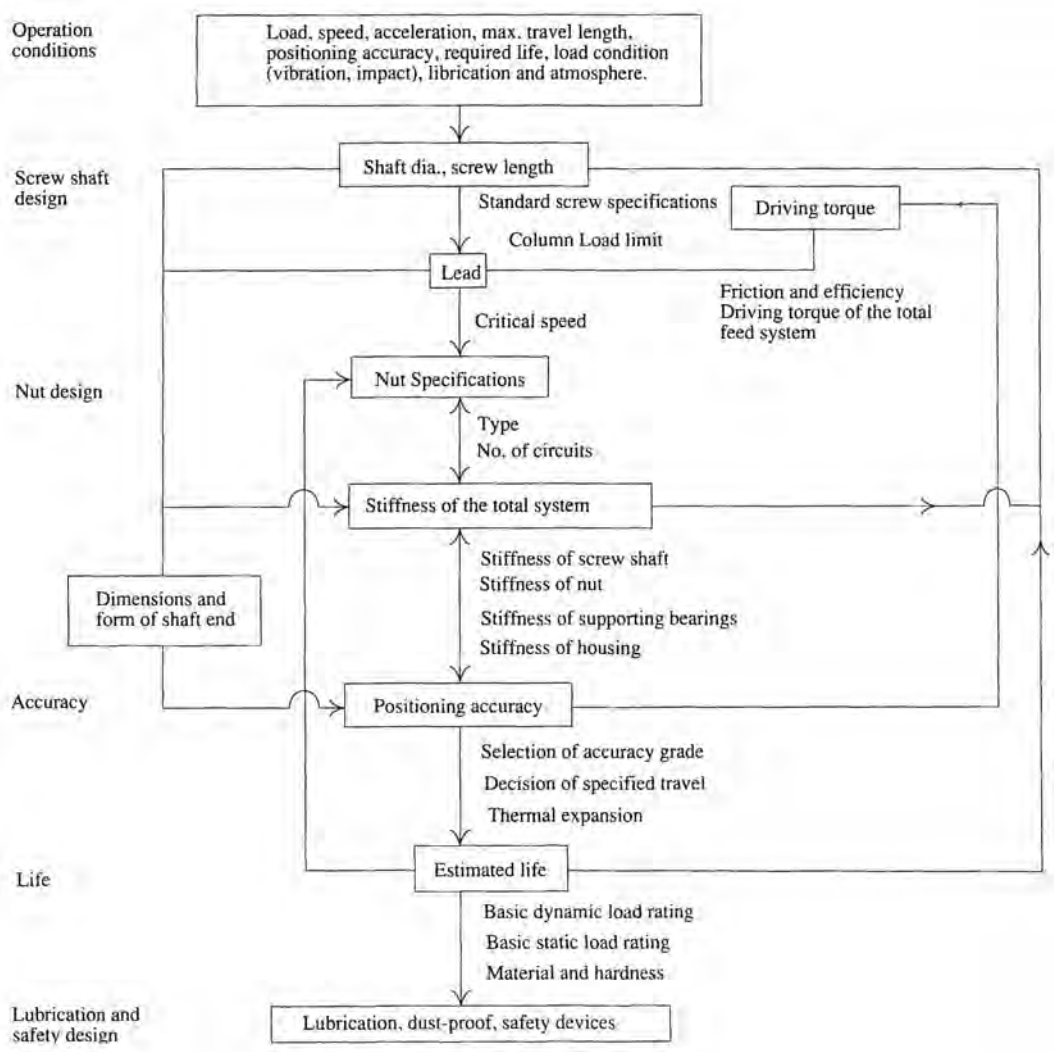


Figure 10.8.24 Ballscrew selection procedure. (Courtesy of NSK Corp.)

	C0	C1	C2	C3	C5	C7	C10
Boring machine			•	•	•		
CMMs	•	•	•				
Drilling machine				•	•	•	
EDM		•	•	•	•		
Grinding machine	•	•	•	•			
Jig borer	•	•					
Lathe	•	•	•	•	•		
Laser cutting machine				•	•		
Milling machine		•	•	•	•		
Machining center	•	•	•	•	•		
Punching press				•	•		
Robots:							
Cartesian-assembly		•	•	•	•		
Cartesian-material handling					•	•	•
Revolute-assembly		•	•	•	•		
Revolute-material handling				•	•	•	
Semiconductor equipment							
Insertion machine			•	•	•	•	
PCB driller		•	•	•	•	•	
Prober	•	•	•				
Steppers	•	•					
Wire bonder		•	•				
Wood working machines					•	•	•

Figure 10.8.25 Ballscrew applications by accuracy class. (Courtesy of NSK Corp.)

Balls are easy to make and sort, so it is not difficult to specify the accuracy grade of the balls used. In general, because balls can be made so well so easily, problems lie in the thread and not the balls.

As the balls leave the thread helix on their way to be recirculated, they can either leave suddenly or gradually roll out. The former condition yields to a noisy ballscrew with a roughness of motion; however, it is the easiest way to manufacture the nut. A ballscrew manufacturer concerned with precision will taper the entrance and exit paths. Even when the entrance and exit paths are tapered, the flow of balls in the return tubes and the continuous flow of discrete elements (balls) compressing and decompressing contributes to the audible noise of a fast-moving ballscrew,[97] and creates axial disturbance forces that can typically be seen on the submicron level even in many a high-precision screw.

Because all leadscrews with mechanical contact between their threads have less than perfect efficiency, they can generate considerable heat in fast-moving, high-cycle machines. Heat transfer out of the screw is difficult since the screw is held to the rest of the machine only by point contacts. As the screw expands, an error in lead can result. This lead error due to thermal expansion can be compensated for in a number of different ways or combinations thereof:

1. The screw can be made with a deliberate negative lead error. A lead offset may be from 20 to 50 μm/m.
2. The screw can be pretensioned in its bearing mounts, which stretches the lead until the screw thermally expands.
3. The leadscrew shaft can be made hollow and oil forced to flow through it, producing a cooling effect, as shown in Figure 10.8.26. Oil forced through the nut and journal support bearings also increases lubrication, decreases wear and noise, and removes heat. However, this requires an oil distribution, collection, and temperature control system. Unless the oil is carefully kept separate from cutting fluid, it should be used only as a coolant and not forced through the nut.

[97] See T. Igarashi et al., "Studies on the Sound and Vibration of a Ball Screw," JSME, Series III, Vol. 31, No. 4, 1988.

4. Use a larger lead, which increases efficiency and decreases rotation speed. Unfortunately, a larger lead may not always be desirable from an optimal transmission ratio point of view.
5. Monitor temperature and make software-based error corrections.
6. Use a linear position sensor (e.g., linear encoder) for closed-loop control. One still has to be concerned with how the heat generated in the screw affects the rest of the machine.

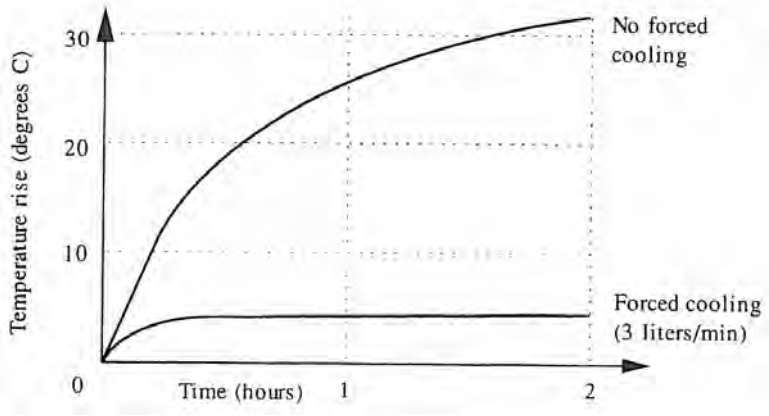


Figure 10.8.26 Effect of forced cooling of a hollow ball screw with 32 mm shaft diameter, 10 mm lead, and 1500 N preload. (Courtesy of NSK Corp.)

Many design engineers rely on a precision ballscrew to allow them to use an encoder or resolver to determine linear position. As discussed in Section 8.2.1, for systems with high friction (i.e., some sliding contact bearings) a rotary sensor on a ballscrew is sometimes needed to minimize jerk when the servo starts from a standstill. Hence lead accuracy is of prime concern in many applications. The lead accuracy of a ballscrew is the correlation between turns of the screw shaft and the theoretical versus actual travel of the nut. Figure 10.8.27 illustrates the terminology used to define lead accuracy. Terms used include:

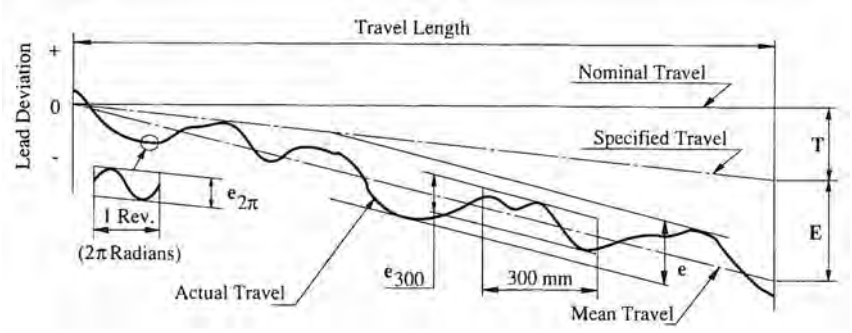


Figure 10.8.27 Definition of lead accuracy. (Courtesy of NSK Corp.)

- *Specified Travel T:* Difference between specified and nominal travel within the travel length. T is negative if it is desired to compensate for thermal expansion of the screw shaft.
- *Actual Travel:* Axial displacement of the nut relative to the screw shaft.
- *Mean Travel:* Best-fit straight line of the actual travel. See Figure 3.1.1 for line types (e.g., best line or least squares line).
- *Mean Travel Deviation E:* Difference between the mean travel and the specified travel within the total travel length.
- *Travel Variations:* Defined by the two lines on either side of the mean travel line that envelop the actual travel line:

Mean Travel Deviation (**E**) and Travel Deviation (**e**)

Travel length Over	Incl.	C0 ±E	C0 ± e	C1 ±E	C1 ± e	C2 ±E	C2 ± e	C3 ±E	C3 ± e	C5 ±E	C5 ± e
-	100	3	3	3.5	5	5	7	8	8	18	18
100	200	3.5	3	4.5	5	7	7	10	8	20	18
200	315	4	3.5	6	5	8	7	12	8	23	18
315	400	5	3.5	7	5	9	7	13	10	25	20
400	500	6	4	8	5	10	7	15	10	27	20
500	630	6	4	9	6	11	8	16	12	30	23
630	800	7	5	10	7	13	9	18	13	35	25
800	1000	8	6	11	8	15	10	21	15	40	27
1000	1250	9	6	13	9	18	11	24	16	46	30
1250	1600	11	7	15	10	21	13	29	18	54	35
1600	2000	-	-	18	11	25	15	35	21	65	40
2000	2500	-	-	22	13	30	18	41	24	77	46
2500	3150	-	-	26	15	36	21	50	29	93	54
3150	4000	-	-	30	18	44	25	60	35	115	65
4000	5000	-	-	-	-	52	30	72	41	140	77
5000	6300	-	-	-	-	65	36	90	50	170	93
6300	8000	-	-	-	-	-	-	110	60	210	115
8000	10000	-	-	-	-	-	-	-	-	260	140
10000	12500	-	-	-	-	-	-	-	-	320	170

Variation per 300 mm ($\mathbf{e_{300}}$) and Wobble error ($\mathbf{e_{2\pi}}$)

	C0	C1	C2	C3	C5
$\mathbf{e_{300}}$	3.5	5	7	8	18
$\mathbf{e_{2\pi}}$	2.5	4	5	6	8

Figure 10.8.28 Typical accuracy ratings for various classes of ballscrews. (Courtesy of NSK Corp.)

e	Maximum width of variation over total travel length.
$\mathbf{e}_{300}$	Variation over any 300 mm section of travel.
$\mathbf{e}_{2\pi}$	Variation for one revolution.

Figure 10.8.28 shows accuracy ratings for various classes of ballscrews. The accuracy class descriptors (e.g., C0, C1, etc.) are particular to the manufacturer, but the values are typical for precision ballscrew manufacturers. Figure 10.8.29 shows typical available shaft lengths for these accuracy grades. Unfortunately, ballscrew manufacturers cannot provide data on achievable resolution because it is so highly dependent on the user's servo system.

Nut Design

There are two main types of ballscrew nuts, as were shown in Figure 10.8.23. The tube type uses an external tube to gather the balls as they exit the nut's thread helix and direct them back to the beginning of the thread helix. The internal ball deflector type performs the same operation internally. The main difference between the two types is that the tube type recirculates the balls after an odd multiple of turns (e.g., 1.5 or 2.5 turns). The internal deflector type recirculates the balls after essentially 1 turn. Internal ball deflectors can only be used on ballscrews with small to moderate leads (β 4), and they make for a quieter ballscrew that is more easily mounted. Internal ball deflector nuts may also be more expensive than tube-type nuts. Both types for precision applications usually use spacer balls to prevent neighboring loaded balls from rubbing against each other.

The number of circuits represents the number of return paths and should be selected so that the product of the number of turns and number of circuits is an integer number. This will minimize the magnitude of the rms noise moments, as shown in Figure 10.8.8. Note that the manufacturer may have oriented the individual circuits with respect to each other so that the noise moments are minimized even if the product is not an integer number. If the product is not an integer number, then ask to speak to an engineer and question him or her about it.

Standard ballscrew nuts have a circular flange with or without a large flat. The large flat allows the screw shaft to be placed closer to the carriage, as shown in Figure 10.8.30. To minimize noise in ballscrews with external return tubes, they should be mounted with the tubes facing down.

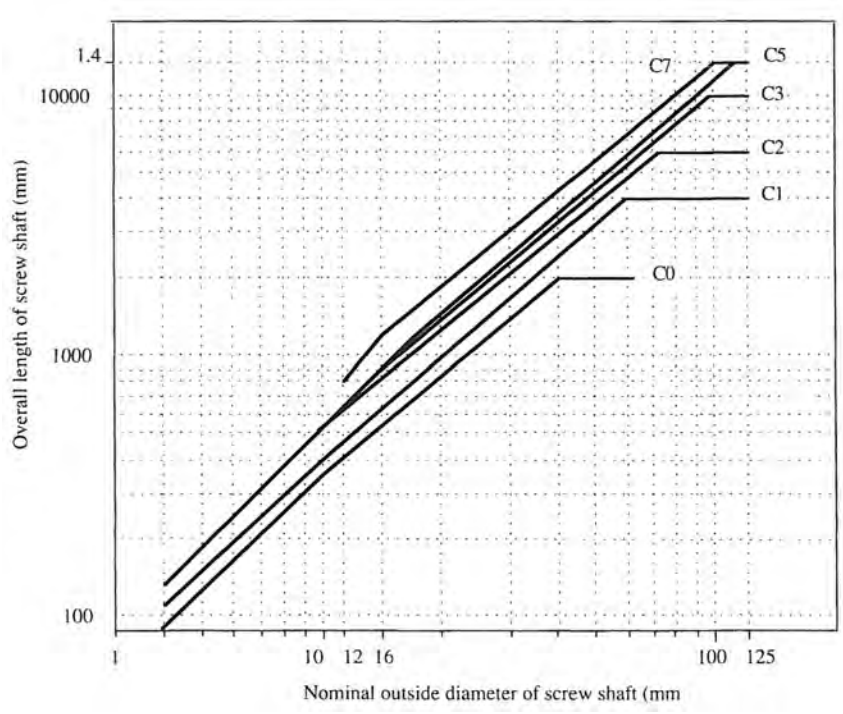


Figure 10.8.29 Typical production scope of ball screws. (Courtesy of NSK Corp.)

To minimize torque variations at low speed, the return tubes should be mounted facing up. The wise user will try both orientations to see which is the best for the intended application. If the ballscrew passed through the center of mass of a device, a full circular nut flange would be used. It is very important to specify the perpendicularity of the mounting surface with respect to the carriage travel direction in order to prevent placing a moment on the ballscrew when it is bolted to the carriage. These moments can decrease the life of the ballscrew and increase carriage motion errors. Tolerances are typically perpendicularity to 500 μrad (target 200 μrad) and lateral misalignment of less than 25 μm (100 μin.). The effect of the tolerance can be calculated using the methods described in Section 2.5.

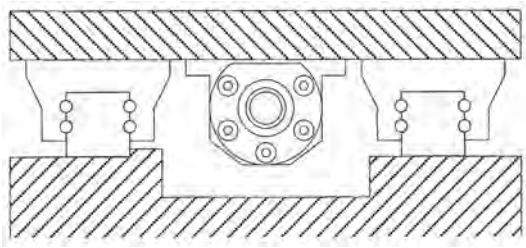


Figure 10.8.30 Mounting a nut with a partial circular flange to a carriage. (Courtesy of NSK Corp.)

There are many different nut designs that can be used to achieve a preloaded condition, as shown in Figure 10.8.31. Tensile preloading is created by inserting an oversized spacer between two nuts and then clamping the nuts together. One nut takes loads in one direction and the other takes loads in the other direction. This creates a back-to-back mounting effect that is thermally stable for a rotating shaft design. A rotating shaft will generally be hotter than the nut because the shaft is not attached to a large heat sink. As the shaft temperature increases, ball contact points on the shaft spread apart axially, which lowers the preload; however the shaft diameter increases which tends to increase the preload. Ideally, the two effects balance, and preload remains constant.

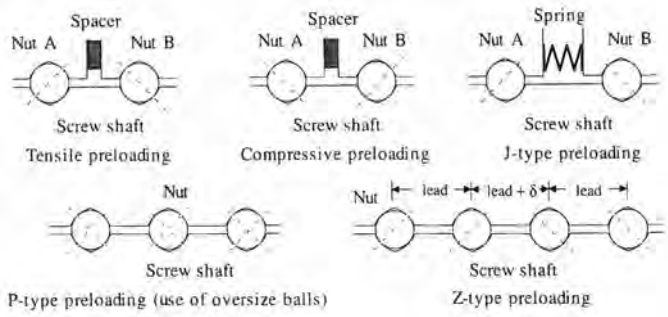


Figure 10.8.31 Common ballnut preload methods. (Courtesy of NSK Corp.)

Compressive preloading uses an undersized spacer between two nuts. This creates a face-to-face mounting situation and is therefore only thermally stable if the nut is likely to be hotter than the leadscrew. In addition, if the nut has a moment placed on it due to mounting errors, the ball loads will be higher with the face-to-face mounting, thereby decreasing nut life.

P-type preloading uses oversize balls to obtain preloading by four-point contact between the balls and the Gothic arch thread shape of the shaft and the nut. This greatly increases the amount of skidding the ball is subjected to, as was illustrated in Figure 8.5.2. This can decrease life and position controllability. Hence this type of preloading is only appropriate for light preloads. When the forces generated by the screw are large, the balls are forced to one side of the groove, so full-four point contact does not occur and skidding is minimized. It is generally less expensive to specify P-type preloading because only one nut is required.

Z-type preloading is also obtained with a single nut by shifting the lead between ball circuits. This creates two-point contact between the balls and the grooves. Z-type preloading is suitable for medium preloads, but is probably not as accurate or easily controllable as tensile preloading.

J-type preloading uses a spring (e.g., disk spring) between the nuts to establish a constant preload. This provides the most constant preload, and hence the most uniform drive torque is obtained. However, the carriage can only be mounted rigidly to one of the nuts. Thus the load capacity and stiffness will vary with direction, and larger hysteresis effects may occur upon load reversal. For some axes on some machines (e.g., some types of turning centers) this may not be a problem. This is the best way to ensure that the preload changes minimally with time.

Preload can also be obtained by hanging a weight by a cable over the edge of the machine, running the cable over a pulley, and connecting it to a carriage that moves in a horizontal plane. However, this is suitable only for applications where the load is much less than the preload. One experimental setup used a carpenter's tape measure as a constant-force spring. Vertically moving axes can use a portion of their weight to provide a constant preload to a ballscrew nut.

Most ballscrew manufacturers use Gothic arch groove threads. When the balls are contacting the arches at one point on the screw and nut, respectively, there is low friction. In the event of overloads, the balls on the portion of the nut which normally would not feel the load can make contact and bear some of the load. Spacerballs are usually used in high-speed or precision ballscrews to prevent the balls from rubbing against each other. For ballscrews subject to high loads or shock loads, spacer balls may not be used, so the load capacity of the ballscrew will be increased. For a high-precision machine, where loads are usually low and better then 1 μm resolution is desired, spacer balls should be specified.

Efficiency is greatly affected by the type of preloading method used. For ballscrews, the coefficient of friction depends on the amount of preload and how it is applied and whether spacer balls are used. It is assumed here that a ballscrew used for a precision machine tool application will have spacer balls. A ballscrew preloaded with oversized balls will have a high coefficient of friction because the balls make four-point contact. A ballscrew preloaded by using two nuts forced against each other will have its balls make essentially two-point contact with the thread groove and thus have a much lower coefficient of friction.

The higher the preload of a ballscrew nut, the greater the stiffness, heat generation, and wear rate. So what is the "optimum preload"? For a compressive or tensile preloaded nut, ideally one

nut will only just lose contact when the maximum force is applied. Recall that the displacement of a system governed by Hertzian contact is proportional to the force to the two-thirds power. If the displacement of each half of the nut under preload is δ_P, when a force is applied to the nut a displacement δ will result and the force balance equation for the nut becomes

$$F_Z = \left[\frac{\delta p + \delta}{C}\right]^{3/2} - \left[\frac{\delta p - \delta}{C}\right]^{3/2} \tag{10.8.21}$$

where C is a constant here. The desired condition is $F_Z = F_{Zmax}$ when $\delta = \delta_P$. This is achieved when the preload force equals 0.35 of the maximum (basic dynamic rated load) force:

$$F_p = 2^{-3/2} F_{Zmax} \tag{10.8.22}$$

Although a preload of 0.35 times the maximum load is the "optimal" value from a stiffness point of view, it leads to excessive heat generation and wear. Since Hertzian contact stiffness acts as a hardening spring, a good compromise for the maximum preload is 10% of the basic dynamic rated load. When the preload is a portion χ ($0 \leq \chi \leq 1$) of the maximum load, $F_P = \chi F_{Zmax}$, the preload is lost with an applied force of

$$F_Z = \chi 2^{3/2} F_{Zmax} = 2.83\chi F_{Zmax} \tag{10.8.23}$$

The deflection of a ballscrew nut is governed by Hertzian contact and thus varies with the two-thirds power of the applied force. The geometry of a ballscrew nut also greatly influences the stiffness, but the deflection will in general still be dominantly proportional to the two-thirds power of the applied force. The compliance is just $d\delta/dF$ and the stiffness is the inverse of the compliance. From a catalog given a ballscrew's stiffness K_{nutmfg} at a preload F_{Pmfg}, the stiffness K_{nut} at a preload force $F_P = \chi F_{Zmax}$ will be:

$$K_{nut} = 0.8 K_{nutmfg} \left[\frac{\chi F_{Zmax}}{F_{Pmfg}}\right]^{1/3} \tag{10.8.24}$$

The factor of 0.8 is added to include the compliance of the nut structure. The total axial stiffness of the leadscrew system will be a function of the nut, the screw shaft, the journal bearings, and the equivalent axial stiffness of the torsional stiffnesses.

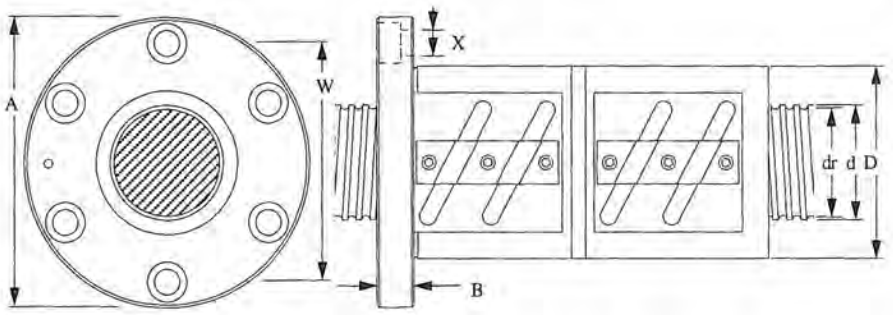
The torque required to overcome the friction caused by the preload force can be found from Equation 10.8.14 by substituting the preload force F_P for F_Z and choosing an appropriate friction coefficient (e.g., $\mu = 0.005$ for a high-precision lightly preloaded ballscrew nut).

There are seemingly innumerable different ballscrew and nut configurations available. A small representative sample of available ballscrews is shown in Figure 10.8.32. Many manufacturers have standard off-the-shelf designs, which means that the end journals are of a specific design and all the customer can vary is the length. Most manufacturers will also sell ballscrews with the ends custom made to the customer's requirements.

10.8.3.8 Hydrostatic Leadscrews

Ideally, a leadscrew would not have mechanical contact between the threads. This would give it a much lower noise moment ratio and no variation in contact area; thus it would have very consistent performance which could readily be mapped. In addition, no torque would be required to overcome friction between the nut and the shaft. The only heat which would be generated during motion of the nut would be due to viscous shear of the media between the threads, but this would also help to increase damping. In addition, squeeze film damping would also be present if a liquid lubricant existed between the threads. An obvious way to achieve this goal is to use a hydrostatic bearing interface between the threads. Hydrostatic leadscrews have been around for many years,[98] and their design is based on the traditional leadscrew design principle of elastic averaging to achieve lead accuracy. This means that many threads are used on the nut and the nut is kept centered on the leadscrew shaft by an Acme thread shape. In order to achieve radial centering, multiple pads are

[98] See, for example, J. Lombard and A. Moisan, "Caractéristiques Statiques et Dynamiques d'un Systéme Vis-écrou Hydrostatique," Ann. CIRP, Vol. 18, 1970, pp. 521–525; anonymous, Mach. Des., April 11, 1968, pp. 219–224; and M. Weck, Handbook of Machine Tools, Vol. 2, John Wiley & Sons, New York, 1984.



Shaft d	Lead l	Root dr	Dyn. Ca	Stat. Coa	Nut K	Shaft K	D	A	B	L	W	X
mm	mm	mm	daN	daN	daN/μm	daN/μm/m	mm	mm	mm	mm	mm	mm
16	4	13.8	515	1050	32	3.0	34	57	11	85	45	5.5
	5	13.2	1360	2750	61	2.7	40	63	11	107	51	5.5
	6	13.2	875	1650	38	2.7	40	63	11	110	51	5.5
20	4	17.8	875	2190	62	5.0	40	63	11	93	51	5.5
	5	17.2	1520	3500	74	4.6	44	67	11	106	55	5.5
	6	16.4	1310	2580	46	4.2	48	71	11	110	59	5.5
	8	16.4	1310	2580	46	4.2	48	75	13	120	61	6.6
25	4	22.8	975	2780	75	8.2	46	69	11	92	57	5.5
	5	22.2	1690	4460	89	7.7	50	73	11	105	61	5.5
	6	21.4	2280	5460	91	7.2	53	76	11	122	64	5.5
	8	20.5	1880	3880	57	6.6	58	85	13	133	71	6.6
	10	20.5	2150	4500	66	6.6	58	85	15	147	71	6.6
28	5	25.2	1780	4980	98	10.0	55	85	12	106	69	6.6
	6	25.2	1780	4980	98	10.0	55	85	12	123	69	6.6
	10	23.5	1990	4380	63	8.7	60	94	15	152	76	9
32	4	29.2	1070	3580	91	13.4	54	81	12	93	67	6.6
	5	29.2	2670	8580	159	13.4	58	85	12	136	71	6.6
	6	28.4	2520	7080	111	12.7	62	89	12	123	75	6.6
	8	27.5	3230	8360	113	11.9	66	100	15	154	82	9
	10	26.4	4720	11000	117	11.0	74	108	15	190	90	9
	12	26.4	3040	6610	72	11.0	74	108	18	181	90	9
36	5	33.2	2800	9690	176	17.3	65	100	15	139	82	9
	6	32.4	3830	12000	181	16.5	65	100	15	162	82	9
	10	30.4	5030	12500	129	14.5	75	120	18	193	98	11
40	5	37.2	2920	10800	191	21.7	67	101	15	139	83	9
	6	36.4	3990	13400	197	20.8	70	104	15	162	86	9
	8	35.5	3550	10500	135	19.8	74	108	15	154	90	9
	10	34.4	5300	14000	141	18.6	82	124	18	193	102	11
	12	34.1	6220	15800	144	18.3	86	128	18	225	106	11
	16	34.1	4010	9490	89	18.3	86	128	22	214	106	11
50	5	47.2	2060	8040	139	35.0	80	114	15	128	96	9
	6	46.4	4370	16700	234	33.8	84	118	15	164	100	9
	8	45.5	5600	20000	240	32.5	87	129	18	205	107	11
	10	44.4	8340	26700	251	31.0	93	135	18	253	113	11

Figure 10.8.32 Typically available ballscrew sizes. (Courtesy of NSK Corp.)

used, which also give the nut a high stiffness about axes orthogonal to the leadscrew axis. This type of design is still in use on large machines and works very well in many applications. For very long travel, high-capacity machines, sag of the leadscrew can be prevented with the use of intermediate radial supports.

A variation on this design is the linear hydrostatic worm gear and rack, sometimes called a *Johnson drive*, which is shown in Figure 10.8.33. The rack is made up of sections that have been replicated using a master leadscrew. The worm is a section of the same pitch leadscrew with a slightly thinner thread to provide the bearing gap needed by the hydrostatic pads. Pressurized oil is supplied to the worm by a rotary commutator which directs oil only to the restrictors in the worm that have rotated to be over a pad pocket in the rack. The worm is turned by a motor either with direct drive or through a rotary transmission system. The transmission system can be in the form of a geartrain that turns the worm shaft, or the outer circumference of the worm can have gear teeth formed along the axial direction so that the worm itself acts as a giant spur gear. The worm can be supported radially and axial by conventional or hydrostatic bearings. With this design concept, moments and lateral forces exist on the worm; however, these forces are negligible given the size and stiffness of the machines that these types of drives are normally used on. The design is also analogous

to using a leadscrew with the shaft fixed and the nut turning to minimize rotary inertias. Like a rack and pinion, there is no fundamental limit to the length of travel that can be provided; however, unlike a rack and pinion, a linear hydrostatic worm drive can have an appreciable transmission ratio.

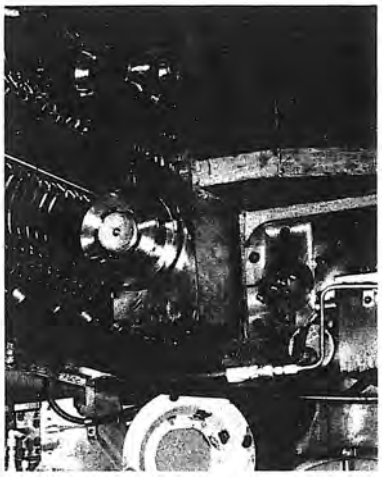


Figure 10.8.33 Linear hydrostatic worm rack made from replication of the rack from a long leadscrew. The "worm" is made from a section of the leadscrew. (Courtesy of Ingersoll Milling Machine Co.)

For very high precision applications, forced geometric congruence between the leadscrew and carriage can cause substantial errors. Hence it would be desirable if the nut could be made to be stiff only along the axis of motion and be free to move along and about all other axes. Such a nut would be called *self-coupling*. Figure 10.8.34 shows the typical cross section of a patented hydrostatic leadscrew (hydroscrew), designed to be self-coupling.[99] By using only one thread turn, manufacturability and self-coupling capability are maximized. Because there is no friction in the system, with a linear position sensor the closed-loop controllability of this leadscrew will be greater than with any other type.

The thread width is chosen based on the lead, maximum force, and stiffness desired. A single turn of the thread allows the noise moments to be minimized, as shown in Figure 10.8.10, while maximizing the pitch and yaw error capability of the nut. The rectangular thread (thread flank angle = 0^o) also minimizes noise moments. When rotated about its center, a pitch or yaw error equal to the hydrostatic bearing gap divided by the screw OD will cause the gap to close by about one-half, depending on the diameter-to-lead ratio. Radial motion capability is typically 5-10 times the hydrostatic bearing gap. For the radial error motion to cause a gap closure of 33%, the amount of self coupling obtainable is:

$$\delta = \frac{h}{3\sin\left(\tan^{-1}\left(\frac{\ell}{\pi D_i}\right)\right)} \tag{10.8.25}$$

When the hydroscrew is supported by hydrostatic or aerostatic bearings and coupled to a similarly supported carriage, the following system properties are obtainable:

- There will be no wear and no static friction in the leadscrew or slide.
- Axial resolution will be limited only by the performance of the servo system.
- Geometric errors in the leadscrew system will not cause error motions in the carriage.
- Axial stiffness can be as high as 1.8×10^8-1.8×10^9 N/m (10^6-10^7 lb/in).
- One meter range of motion can easily be achieved.
- Axial damping will be very high, due to the squeeze film effect.

[99] See A. Slocum, A System to Convert Rotary Motion to Linear Motion, U.S. Patent 4,836,042, June 6, 1989 (and corresponding foreign patents), assigned to AESOP Inc., Concord, NH, (603) 228-1541. Also see A. Slocum, "A Replicated Self-Coupling Hydrostatic Leadscrew for Sub-micron Applications," SME Techn. Paper MS90-320.

The hydroscrew is truly self-coupling in that error motions of the screw shaft will not affect motion of the carriage. There will still be noise moments, but they will be smooth functions that are a small percentage of the drive torque (<10%). Note that the errors produced by these moments will typically be an order of magnitude less than geometric errors in the system. Hence a machine whose performance is degraded by these small moments will have to have a metrology frame anyway. However, because the moments will be smooth functions, not the noisy type produced by mechanical contact-type leadscrews, errors caused by the noise moments will be readily measurable and correctable in real time by the metrology frame and servo system respectively.[100]

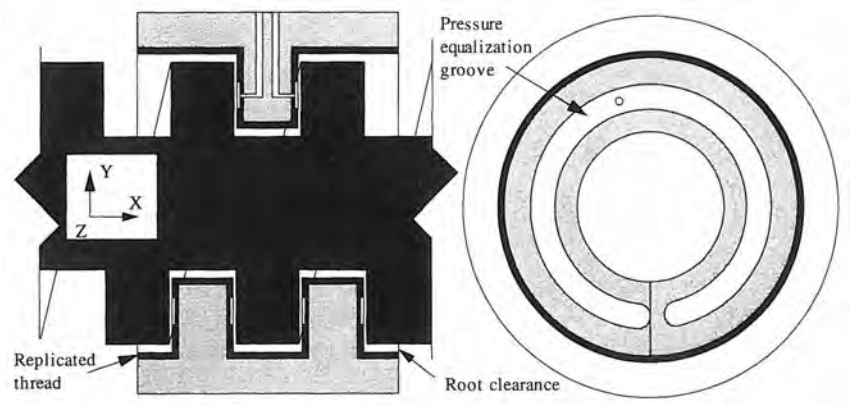


Figure 10.8.34 Cross section of a self-coupling hydrostatic leadscrew, and an end-view of the nut. (Courtesy of AESOP Inc.)

10.8.3.9 Leadscrew Mounting Methods

There are two primary methods for mounting a leadscrew with a carriage: (1) Nonrotating shaft and rotating nut, or (2) rotating shaft and nonrotating nut. The former is most often used when the critical speed limit of a long shaft is below the maximum desired shaft speed. It can also be useful for systems where the carriage length is longer then the travel. A rotating nut is supported by bearings and can be rotated with the use of gears, a timing belt, or a direct-drive motor that uses the OD of the nut as the shaft for the motor rotor. Some manufacturers offer ballscrew nuts with integral ball bearings for rotating nut applications.

Figure 10.8.35 shows common mounting methods for leadscrews. Note that trunions are sometimes used to increase misalignment tolerance. Also note that when the leadscrew is stretched to compensate for thermal expansion, it is possible to achieve a condition whereby the shaft is always in tension and therefore cannot buckle. Regardless of the method used, it is important to minimize the moment and radial force placed on the nut. A precision ballscrew should be mounted with less than 20 μm radial and 200 μrad angular misalignment, respectively, so as not to decrease life substantially. When the shaft is rotating, the primary factor to consider is the critical shaft speed. If the rotational frequency equals the natural frequency of the shaft in bending, an unstable condition known as *shaft whip* occurs. No rotating shaft is perfectly balanced; hence it is important to keep the rotational frequency below that of the first bending natural frequency of the shaft. If this condition is not met, shaft whip can occur. Particular care should be taken when designing leadscrews where the ratio of shaft length to diameter exceeds 70.

Figure 10.8.36 shows the natural frequency and buckling factors for common mounting cases. When evaluating the cross-sectional area and moment of inertia, one should use the values of the leadscrew radius that give the most conservative result. One should also check the compressive and tensile stresses in the leadscrew shaft. For very long leadscrews, intermediate shaft supports can be designed which retract when the carriage approaches. In addition to the critical shaft speed and buckling loads, the dN value for the balls must be determined. This value is the product of the pitch

[100] If no moments on the carriage are tolerable whatsoever, a hydrostatic leadscrew could be used with a paddle coupling, as discussed in Section 10.8.4. The benefit of the hydrostatic leadscrew would be that there would be no drive friction in the system.

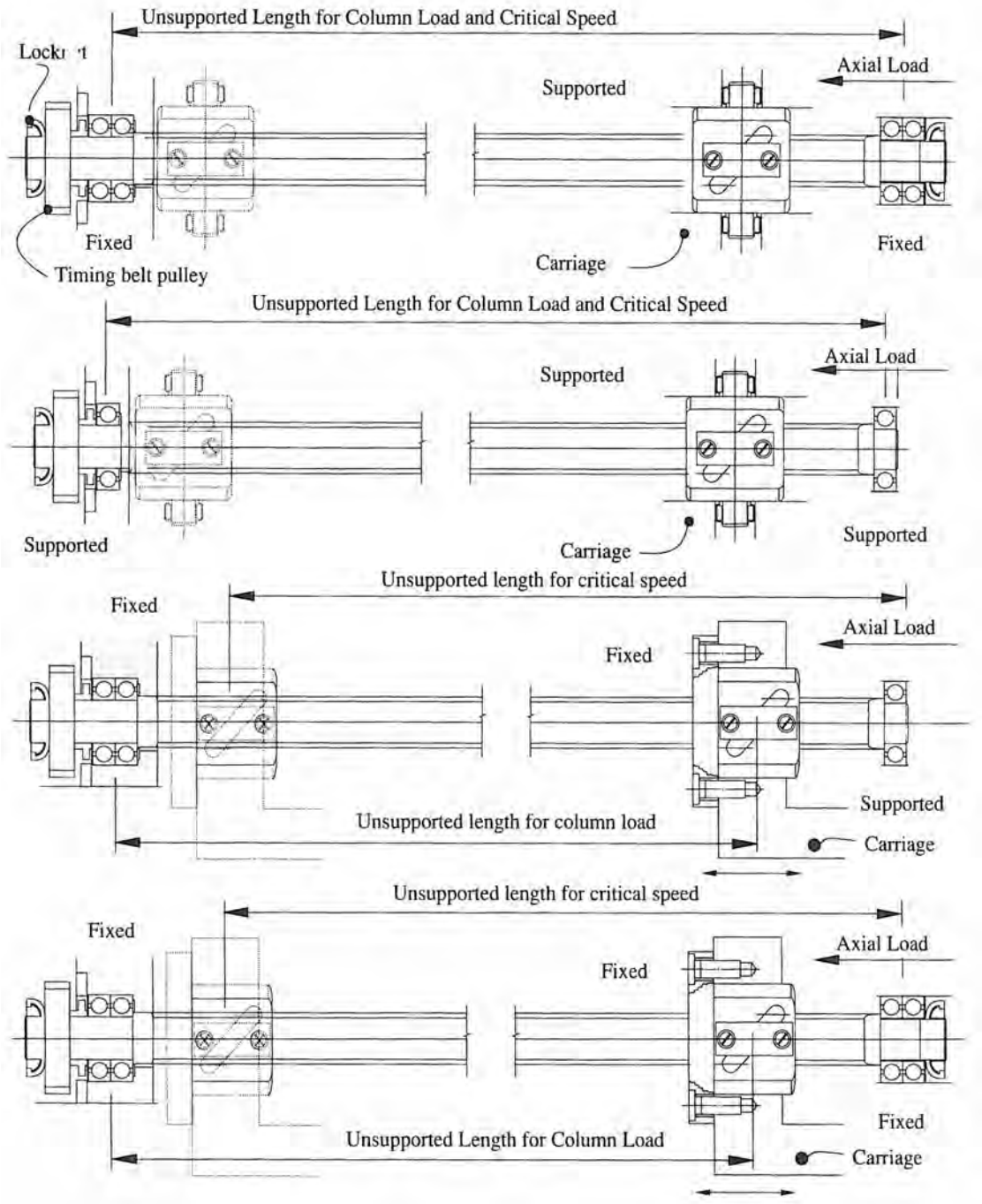
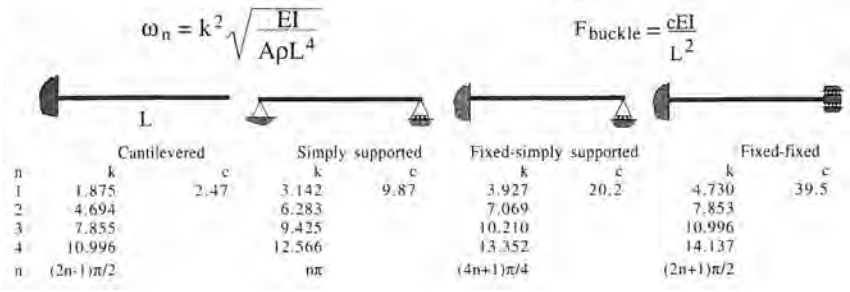


Figure 10.8.35 Examples of leadscrew mounting methods. (Courtesy of NSK Corp.)

diameter (mm) and the speed (rpm) of the screw shaft. For precision-ground ballscrews, dN should be less then 70,000. At higher values, the centrifugal forces on the balls become too great.

$$\omega_n = k^2\sqrt{\frac{EI}{A\rho L^4}} \qquad F_{buckle} = \frac{cEI}{L^2}$$



	Cantilevered		Simply supported		Fixed-simply supported		Fixed-fixed	
n	k	c	k	c	k	c	k	c
1	1.875	2.47	3.142	9.87	3.927	20.2	4.730	39.5
2	4.694		6.283		7.069		7.853	
3	7.855		9.425		10.210		10.996	
4	10.996		12.566		13.352		14.137	

Figure 10.8.36 Bending natural frequencies and column buckling loads for various leadscrew mounting configurations.

When a fixed-fixed mounting is used, it allows the leadscrew to be stretched for thermal compensation. However, stretching increases the load on the journal bearings, which in turn generates more heat. In addition, it is one more item which must be controlled during the assembly operation. Stretching the leadscrew does prevent the leadscrew from being solely in compression under some loads and thus increases the buckling load. On the other hand, as discussed in Section 2.5, a simply supported leadscrew is the best way to minimize carriage errors due to forced geometric congruence. A simply supported leadscrew also minimizes lateral forces and moments on the leadscrew nut, thereby increasing life. To achieve a simply supported mounting, angular contact bearings at opposite ends of the shaft would have to be used to preload each other. This would require the use of shims, which as noted below is undesirable. A duplex pair of bearings at one end and a radial contact bearing at the other will provide a fixed-simply supported mounting condition. This design can be consistently manufactured with high quality and minimal effort and cost.

In some applications, a stroke of only a few decimeters is required. In these applications, the best possible alignment of the leadscrew can be achieved using a monolithic mount shaped somewhat like a bathtub. First the overall shape is machined. The mount is then stress relieved and the outside surfaces are ground flat and perpendicular to each other. The mount is then fixtured to a precision rotary index table and one shaft support bearing bore is jig bored. The index table is rotated 180° and the other bore is jig bored. When designing a leadscrew mount, it is important to consider how the leadscrew support bearings will be preloaded. Leadscrew support bearings must withstand primarily high thrust loads. Thus special high-contact-angle angular contact bearings are often used, as shown in Figure 8.4.7. For absolute best performance in terms of minimizing the starting torque and heat generation, aerostatic or hydrostatic bearings should be used. Few machines, however, warrant the latter options. As discussed in Section 8.9, to minimize thermal effects on preload of bearings, a back-to-back mounting configuration should be used with angular contact bearings that support leadscrew shafts. In order to preload angular contact bearings, shims can be used to define the amount of preload or the bearings can be purchased with the preload displacement included in the race width. Using shims is a sure way to create difficulty in getting the preload just right so that there is no variation in tare torque. The reason is that a difference in bearing preload caused by tens of microns can be too much. Controlling shim thickness to this level is extremely difficult. The proper way to preload an angular contact bearing is for the manufacturer to hang the preload weight from the inner race and then grind the races flush. When the bearings are installed and forced together, the preload will be perfect. If a manufacturer tries to sell you leadscrew support bearings

that require you to make measurements, determine the required shim thickness, and then use shims, tell them that you do not want their bearings. There are several bearing manufacturers that sell ABEC 7 and better angular contact leadscrew support bearings that have the preload displacement included in the width of the bearing.

In order to evaluate the axial stiffness of a leadscrew-actuated system, the stiffness of the journal support bearings, leadscrew shaft, and leadscrew nut (discussed above for ballscrews) must be added in series. This stiffness would itself be in series with the structural stiffnesses of the journal and leadscrew nut mounting brackets. For the shaft support bearings, if the manufacturer does not give axial stiffness values, the bearing stiffness can easily be estimated using Hertz contact theory, discussed in detail in Section 5.6. One need only be careful to account for the contact angle and number of balls and how it affects the loading. The stiffness of the leadscrew shaft depends on the type of mounting and the location of the nut on the shaft. When the shaft is not pretensioned, the stiffness is that of a column whose length is from the nut to the bearing that supports a thrust load. When the shaft is pretensioned, the stiffness is the sum of the stiffnesses of the shaft lengths on either side of the nut. The stiffness of the leadscrew nut could be determined considering the type of thread and contact properties. Fortunately, most manufacturers provide stiffness data for their leadscrew nuts.

In most cases the leadscrew should be protected from dirt and slime by bellows or protective way covers, as shown in Figure 10.8.37. Wipers should be used as a last-chance dirt filter. In most applications the leadscrew nut is greased for life. In ultraprecision applications a light spindle oil is preferred, and means to apply the oil periodically or continuously is required. Also, as mentioned earlier, hollow screw shafts with forced cooling oil are sometimes used; hence means to deliver and collect the oil must be provided.

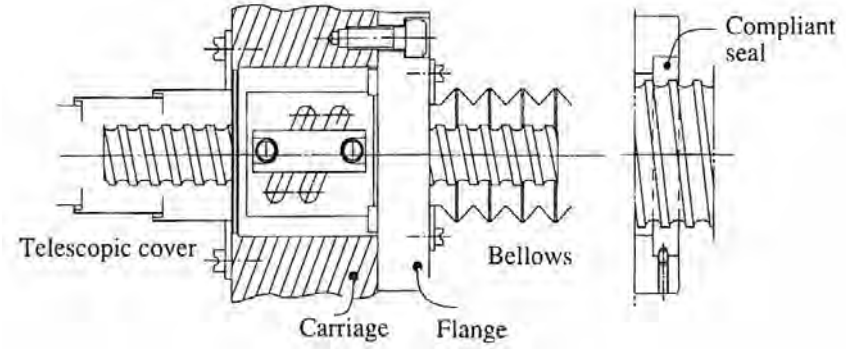


Figure 10.8.37 Leadscrew sealing methods. (Courtesy of NSK Corp.)

10.8.3.10 Choosing the Optimal Lead

In order to choose the "optimal" lead for a leadscrew, the shaft diameter is first chosen using Equation 10.2.6 in conjunction with other axial stiffness controlling components. Together with the expected motor inertia, this will establish the rotational inertia of the system. This is used with the inertia matching criteria given by Equation 10.2.23 or Equation 10.2.33 when external loads are substantial. Quite often a small lead results from these specifications, so the critical shaft speed often becomes the dominant lead selection factor.

Given a maximum desired carriage velocity V_{max} (in meters per second) and the maximum shaft speed ω_n (in radians/second from Figure 10.8.36), the minimum leadscrew lead in meters is

$$\ell = \frac{2\pi V_{max}}{\omega_n} \tag{10.8.26}$$

The value of this lead is then used in Equation 10.2.26 or Equation 10.2.33 with the leadscrew, coupling, and load inertia, to find the "optimal" motor inertia. One can then search the catalogs for a motor with the required torque and speed characteristics. Sometimes this will lead to the selection of a large-diameter *pancake* motor. This type of motor is often a more efficient producer of torque

than long, small-diameter motors. If a motor with a high enough inertia cannot be found, one just has to live with the situation. It does not make sense to add rotating mass to the system unless one does it to increase the system stiffness simultaneously.

10.8.3.11 Life Calculations

Leadscrew life can be defined in terms of revolutions, hours, or meters of travel. Variables used to determine life include:

$\mathcal{L}$	Fatigue life in revolutions.	F_a	Axial load in newtons
$\mathcal{L}_t$	Fatigue life in hours.	n	Rotation speed in rpm
$\mathcal{L}_s$	Fatigue life in kilometers nut travel.	ℓ	Lead in meters
Ca	Basic dynamic rated load.		
f_w	Load factor:		
	Smooth without impact	1.0-1.2	
	Normal operation	1.2-1.5	
	Impact and vibration	1.5-3.0	

The load-life equations are[101]

$$\mathcal{L} = \left(\frac{C_a}{F_a f_w}\right)^3 \times 10^6 \tag{10.8.27}$$

$$\mathcal{L}_t = \frac{\mathcal{L}}{60n} \tag{10.8.28}$$

$$\mathcal{L}_s = \frac{\mathcal{L}\ell}{10^6} \tag{10.8.29}$$

Very few applications place a leadscrew under a constant load; thus a method is needed to formulate an equivalent constant load. This is done using a form of Miner's rule used for fatigue analysis. Given a series of N axial loads F(i), each applied at a rotating speed of $n(i)$ for a duration t(i), the equivalent axial load F_a is

$$F_a = \left(\frac{\sum_{i=1}^{N} f(i)^3 n(i) t(i)}{\sum_{i=1}^{N} n(i) t(i)}\right)^{1/3} \tag{10.8.30}$$

Similarly, the equivalent rotational speed is given by

$$n = \frac{\sum_{i=1}^{N} n(i) t(i)}{\sum_{i=1}^{N} t(i)} \tag{10.8.31}$$

Rarely is a leadscrew designed for "infinite life." Typical machine tool applications have a design life of 20,000 h.

The maximum static load must also be checked where the maximum static load is typically one-half of the rated static load for normal operation and two to three times less for operation with impact or vibration. Note that for ballscrews, the rated static load will produce a permanent deformation at the contact points equal to 0.01% of the ball diameter. Once deformed, the leadscrew will be noisy. For very slow operation or infrequent operation of a machine or some of its axes, one needs to be concerned with fretting corrosion. In these applications, the maximum load has to be seriously derated, sometimes by a factor of 2-10. If possible, the design engineer should make sure that the machine's owner's manual specifies that all axes of the machine be moved through their full range of motion (or what the tooling will allow) several times a day. Ultimately, the leadscrew size should be the larger of the one chosen by the minimum stiffness criteria or by the load-life criteria.

[101] Precision Machine Parts: Linear Motion Products, NSK Corp., Chicago, IL.

10.8.4 Coupling Methods

Perhaps the most difficult task in designing and building a precision machine or instrument is accounting for misalignment of actuators and bearings.[102] Specifically, coupling a linear actuator to a linear bearing so as to not induce wear and error[103] into the system can often be achieved only via the use of hand finishing operations, error mapping and software, coarse-fine actuation systems, or kinematic transmission elements.

Hand Finishing Operations

Hand finishing operations (e.g., lapping or scraping) have traditionally been employed to make sure that a leadscrew is parallel to an axis of motion and that the leadscrew nut flange bolts to the slide without imposing any stresses on the system. Hand lapping has also enabled instruments with sliding tables to be developed with resolution on the order of angstroms.[104] Hand finishing operations are also the key to successful manufacture of diamond turning machines, which are critical to the manufacture of many optical components, computer memory disks, and precision scientific instruments. However, hand finishing techniques are difficult to employ on a large production basis because of an increasing lack of skilled craftspeople.

The alternative to hand finishing operations is to place strict tolerances on machined parts so that during assembly the actuator (e.g., leadscrew) will be aligned with the axis of motion. In order to determine what the tolerances should be with respect to generating minimal errors in the system, a procedure should be used such as that described in Section 2.5.

Mapping and Error Correction Algorithms

Regardless of the technique used to manufacture a machine, software-based error mapping and compensation techniques can be used to increase accuracy to near the level of the repeatability, as discussed in detail in Chapter 6. However, these techniques can only provide a limited factor (typically 10) of improvement and thus are most effective when applied to machines that have already been made mechanically as good as economically possible. In addition, any errors due to deformation of components caused by forced geometric congruence can change with time due to wear. Thus machines whose errors are compensated for with software-based error correction techniques may have to be remapped periodically.

Coarse-Fine Motion Systems

Cost and difficulty of attainment of accuracy are often proportional to the "parts per million" accuracy required. Thus coarse-fine (macro-micro) systems have evolved. This type of system was shown in Figure 8.8.8. It uses a stage manufactured with conventional precision grinding techniques and a flexural bearing fine motion stage to correct for errors. Extensive work has been done on these types of systems, and they are used in precision devices such as wafer steppers for the manufacture of integrated circuits.[105] In different configurations they have also found application in robotics[106] and experimentally as "fast tool servos" for single-axis final positioning of diamond tools.[107]

A coarse-fine system is a proven approach that compensates for errors in one system by actively servoing them out using another system. Coarse-fine systems typically use a leadscrew or rack-and-pinion driven stage supported by rolling or sliding element bearings. The fine motion stage is usually supported by flexural bearings and actuated with microhydraulics, voice coils, or piezoelectrics. The principal problems with coarse fine systems, however, are the complexity of the mechanical design, difficulty in design and implementation of control algorithms, and the need for additional multiple sensors and servo-control hardware.

---

[102] Section 2.5 illustrates in detail how to calculate errors associated with coupling linear slides to actuators when the two are not exactly aligned.

[103] See Equation 2.2.29 for a discussion on the center of stiffness of a bearing system.

[104] K. Lindsey and P. Steuart, "NPL Nanosurf 2: A Sub-nanometer Accuracy Stylus-Based Surface Texture and Profile Measuring System with a Wide Range and Low Environmental Susceptibility," 4th Int. Precis. Eng. Semin, Cranfield, England, May 11–14, 1987, p. 15.

[105] For example, GCA Corporation's DSW® wafer stepper.

[106] A. Slocum, "Design and Implementation of a Five Axis Robotic Micromanipulator," Int. J. Mach. Tool Des., Vol. 28, No. 2, 1987, pp. 131–139.

[107] E. Magrab and S. Patterson, "Design and Testing of a Fast Tool Servo for Diamond Turning Machines," Precis. Eng., Vol. 7, No. 3, 1985, pp. 123–128.

Kinematic Transmissions

The purpose of a kinematic transmission system is to prevent nonaxial motion components of the actuator from causing any motion of the carriage. An added benefit is that they often help to reduce the amount of heat transferred between the actuator to the slide. A kinematic transmission system is thus one in which only one degree of freedom between the actuator and the bearing supported slide is restrained. A kinematic transmission system filters out error motions by allowing members to slide or deflect in nonsensitive directions. There are two categories of kinematic transmission elements, active and passive.

Active Kinematic Transmissions

Active kinematic transmissions have sliding members and can achieve a true kinematic condition so that only an axial force is transmitted through them. There are several types of active kinematic transmission elements. One type of active kinematic transmission is the paddle type shown in Figure 10.8.38. The actuator is directly connected to a slave carriage which has a U-shaped yoke attached to it. A paddle attached to the master carriage fits into the U. Single-entry fluid film thrust bearings, which cannot resist lateral or angular motions, cause the paddle to remain axially centered in the yoke. In this manner only axial forces are transmitted from the slave to the master carriage.[108]

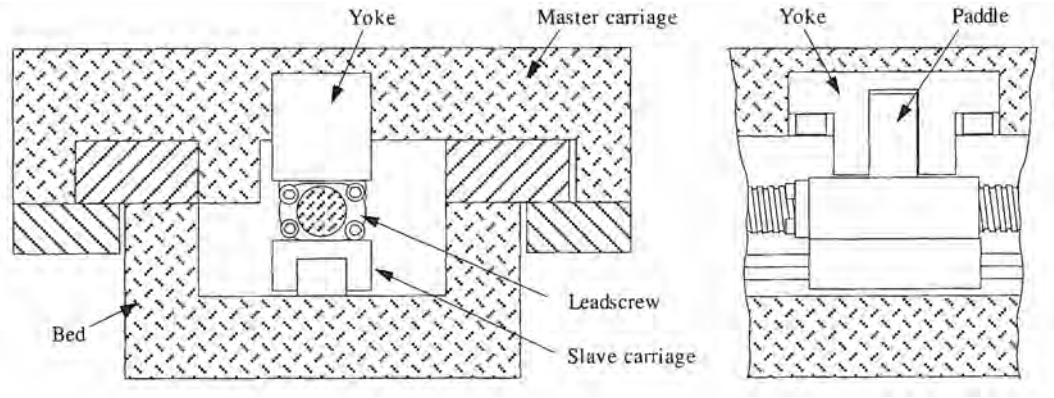


Figure 10.8.38 Paddle-type active kinematic transmission.

The center of the paddle is most often located near the center of gravity of the master carriage, or where it is most of the time in the case of carriages that have greatly varying part loads. In this manner, pitch and yaw motions of the carriage due to an actuation force applied far from the center of gravity are minimized. The slave carriage can also act as the anchor point for the moving end of a cable carrier. Then wires and fluid lines can run from the slave to the master carriage with only a small strain relief required. This is an extremely desirable situation because often on a precision slide without a slave carriage, the force of the cable carrier on a carriage can be enough to cause just enough nanometers of deflection to ruin a critical part. The disadvantage of a paddle-type kinematic transmission is the need for an extra carriage and the bearings between the paddle and the yoke. If the actuator is a conventional leadscrew or ballscrew (i.e., not a hydrostatic leadscrew), one still has to reckon with backlash and friction.

A paddle transmission with a hydrostatic or aerostatic bearing paddle can be designed using several simple steps:

1. Lateral motions are virtually unrestricted, and thus one only needs to determine the maximum angular errors that must be accommodated.
2. Using the equations in Section 8.7 or the design charts in Chapter 9, design a fluidstatic bearing with the appropriate load and stiffness, making sure that one-half the bearing gap divided by the bearing diameter is greater than the maximum error angle the coupling must accommodate. This design may require some iteration.

[108] See P. A. McKeown, "High Precision Manufacturing and the British Economy," Proc. of IMechE, 1986, Vol. 200, No. 76, pp. 1–19. Also see D. H. Youden, "The Design and Construction of a Sub-microinch Resolution Lathe," Ultraprecision in Manufacturing Engineering, M. Weck and R. Hartel (eds.), Springer-Verlag, New York, 1988.

3. Design the paddle and yoke structure to have the appropriate bending compliances. The compliance of each element in the drive train should be approximately equal. The sum of the compliances should then be equal to the maximum allowable axial compliance for the system (see Equation 10.2.6).

Remember that the bearing interface between the paddle and the yoke must only transmit axial forces and not be able to support moment loads.

A partial kinematic transmission is the crossed yoke coupling shown in Figure 10.8.39. This device is intended primarily for use with leadscrews and consists of two linear/rotary motion bearings at 90 ^{o} to each other. The leadscrew nut is bolted to the first block, which has pins on its ends that are free to slide and rotate in the arms of a yoke. The yoke itself has pins which are free to slide and rotate in a second yoke that is bolted to the carriage. Note that the torque from the leadscrew is transmitted to the carriage, which can cause roll errors. This type of device requires four linear/rotary motion bearings and does not provide an intermediate point to which a cable carrier can be attached. Note that this configuration could be used as a coupling between a ram type device such as a hydraulic piston rod and a carriage.

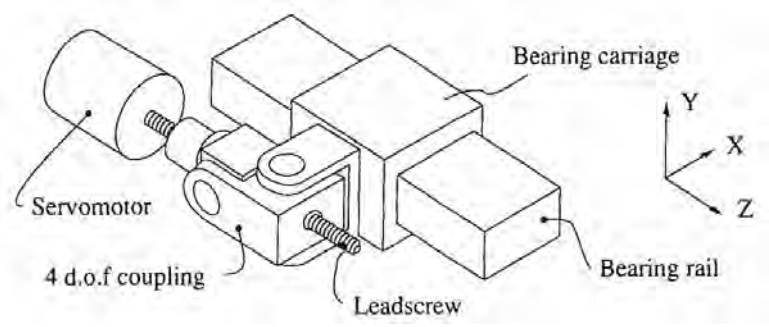


Figure 10.8.39 Four degree-of-freedom active kinematic transmission.

Flexural Couplings

Flexural couplings have members that easily and elastically deform in order to accommodate error motions while being relatively rigid along the direction of the platten's axial motion. There are numerous types of flexural couplings which act as effective kinematic transmission elements, some of which are shown in Figure 10.8.40. Unfortunately, providing compliance in the direction of error motions almost always leads to a stiffness in the axial direction that is far less (e.g., typically a factor of 10 less) than if the actuator were coupled to the carriage rigidly or through a paddle-type element. As a result, a machine built with flexural couplings will have a slower (e.g., one-third) response time, which can lead to decreased productivity and inability to servo out noise inputs. Still, the simplicity and economics of flexural couplings make them extremely useful to design engineers who recognize their limitations and include them in their system models before building a prototype. In addition, with the use of an added position sensor and a slightly more complicated control algorithm, one can increase the apparent axial stiffness of a flexural coupling as discussed in Section 10.9.

There are many different types of flexural couplings that provide varying degrees of coupling action (i.e., axes of compliance). Examples range from a leadscrew with an integral coupling in its nut shown in Figure 10.8.17 to the wire, membrane, and beam type elements shown in Figure 10.8.40. Also recall the flexural pivot discussed in Section 8.6 which in addition to serving as a bearing often is used as a coupling. Of these, the wire type is probably most easily placed at the center of mass of the platten with the drive bar connection reaching out to a slave bearing that is then driven by the actuator.

Wire-Type Flexural Coupling Design

A wire coupling, shown in Figures 10.8.40 and 10.8.41, is the simplest form of flexural kinematic transmission system. The axial stiffness of the wire coupling held at its midpoint is unaffected by the tension. For a wire cross-sectional area A and wire length 2l, assuming the region of length 2c where the wire is grabbed is rigid with respect to the wire, the axial stiffness is K_{axial} = Æ/a. As the

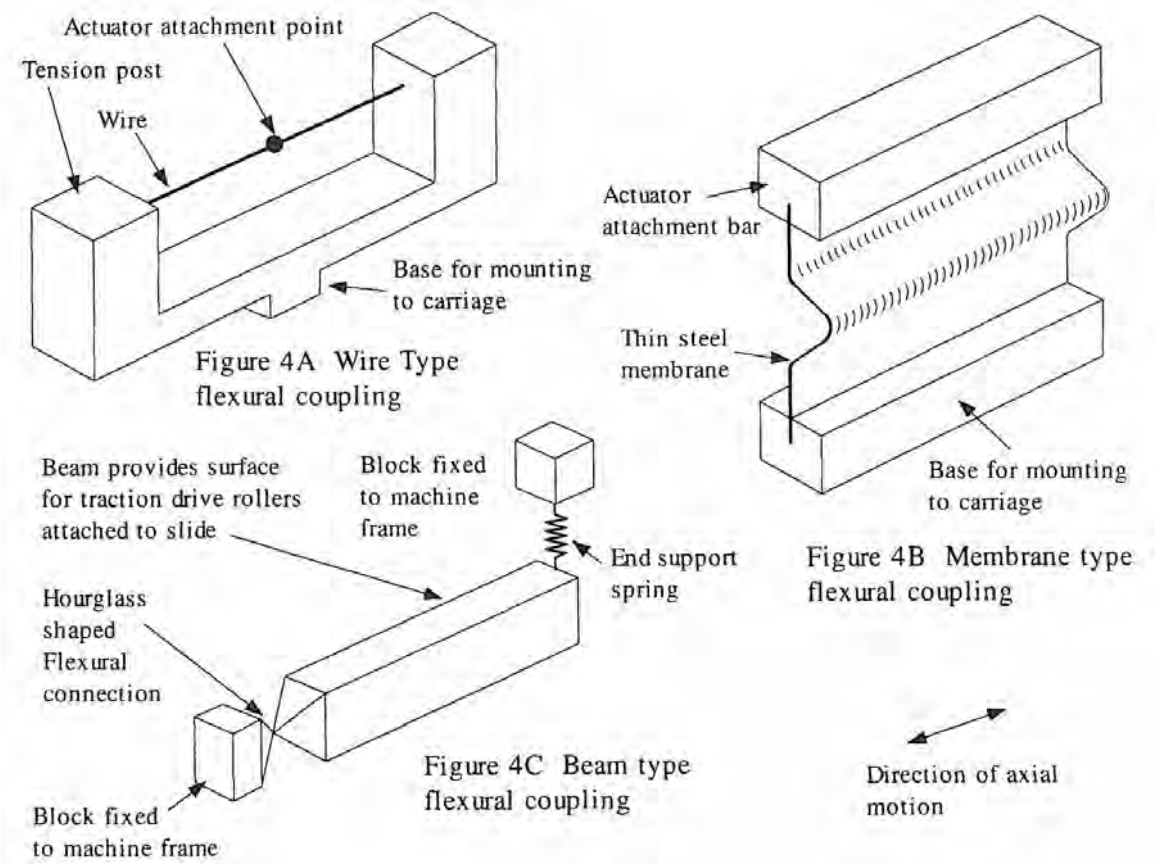


Figure 10.8.40 Types of flexural couplings.

wire is displaced laterally, forces are generated to resist this motion by the initial tension in the wire and the stretching of the wire. The change in tension caused by the wire stretching as it is deflected δ laterally is

$$\Delta T = EA\left\{\left(1+\frac{\delta^2}{a^2}\right)^{1/2}-1\right\} \tag{10.8.32}$$

For an initial wire tension T and change in tension ΔT caused by lateral motion δ of the wire, the resultant force on the end supports of the wire is

$$F_{lateral} = 2(T+\Delta T)\sin\gamma \approx \frac{2\delta\,(T+\Delta T)}{a} \tag{10.8.33}$$

The effective lateral stiffness of the wire is found[109] by substituting for ΔT in $F_{lateral}$ and taking the partial derivative $\partial/\partial\delta$ of Equation 10.8.33:

$$K_{lateral} = \frac{2T}{a}+\frac{3EA\delta^2}{a^3} \tag{10.8.34}$$

When an axial force is applied, the tension on one side increases by the same amount it decreases on the other side. Thus the above is still a good approximation of the lateral stiffness in the presence of axial forces that do not reduce the tension on one side to zero. A moment applied to the center of the wire about an axis orthogonal to its length causes the wire to be displaced in opposite lateral directions and hence causes a force couple at the ends of the wire. From Figure 10.8.41, the elongation Δa of the wire is found using the law of cosines:

$$\Delta a \approx \frac{c\ell(1-\cos\theta)}{a} \tag{10.8.35}$$

[109] Note that for $\varepsilon < 0.01$, $(1+\varepsilon)$

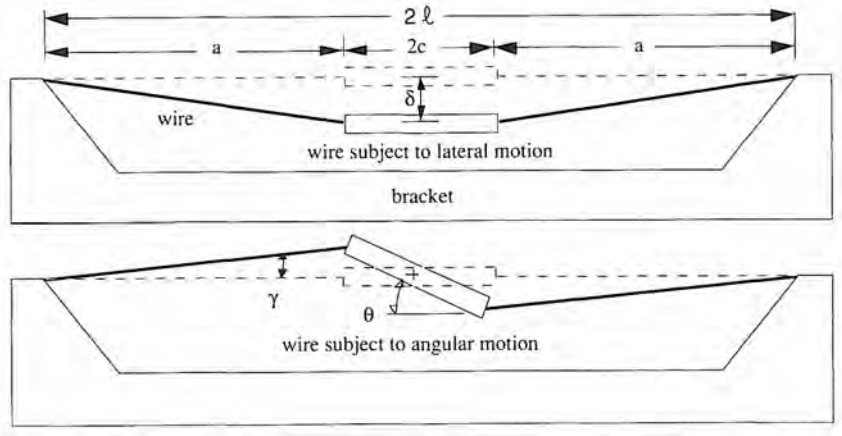


Figure 10.8.41 Geometry associated with wire-type flexural coupling.

The lateral force exerted by this segment of the wire on the attachment point is thus

$$F_{lateral} \approx \frac{c^2\, EA\theta\ell(1-\cos\theta)}{a^3} + \frac{Tc\theta}{a} \tag{10.8.36}$$

The force is nonlinear, and hence the lateral stiffness caused by an angular displacement is found from $\partial/\partial\theta$ of Equation 10.8.36:

$$K_{lat-ang} \approx \frac{c^2 EA\ell(1-\cos\theta+\theta\sin\theta)}{a^3} + \frac{Tc}{a} \tag{10.8.37}$$

For small angles, the effective lateral stiffness is essentially just Tc/a.

If the applied axial force is greater than the initial tension in the wire, one side of the wire will become slack. Upon reversal of the force, the wire will have to be displaced by an amount equal to the slack generated and will thus behave like there is backlash in the system. Hence the initial tension must be carefully controlled. Too much tension and the lateral stiffnesses are too high. Too little tension and backlash occurs. A system with a wire coupling is shown in Figure 10.8.42.

For an instrument with a positioning platten (e.g., a scanning microscope) driven by a steel wire coupling, the following parameters may exist:

- Wire tension = 10 N (2.25 lbf)
- Slide stiffness = 10^8 N/m (570,000 lbf/in.)
- Length 2l = 11 cm, 2a = 10 cm (4 in.)
- Wire diameter = 0.25 mm (0.010 in.)
- $K_{lateral} \approx 402$ N/m (2.288 lbf/in.)
- Error motion = 100 μm (0.004 in.)
- Resultant maximum lateral force = 0.0402 N (0.0090 lbf)
- Resultant slide lateral error motion = 4 Å
- K_{axial} = 203 kN/m (1160 lbf/in.)

Even if the wire tension were increased by an order of magnitude, the lateral error motion would only be 40 Å. Note that if a flat metal band is twisted and the coupling point located in the middle, then although some lateral flexibility will be lost, far greater axial coupling capability will be obtained.

Membrane and Beam-type Kinematic Transmissions

A membrane-type kinematic transmission element, as shown in Figure 10.8.40b, uses thin plates to provide compliance for error motions while maintaining some degree of axial stiffness. In order to conservatively (under) estimate the axial stiffness of this type of coupling, one assumes that the membrane is a beam of length equal to the contour length along one edge of the coupling. Depending on the angular stiffness of the linear bearings, the beam will behave somewhere between that of a beam that is built into a wall at both ends and a cantilever beam. One must be sure to account

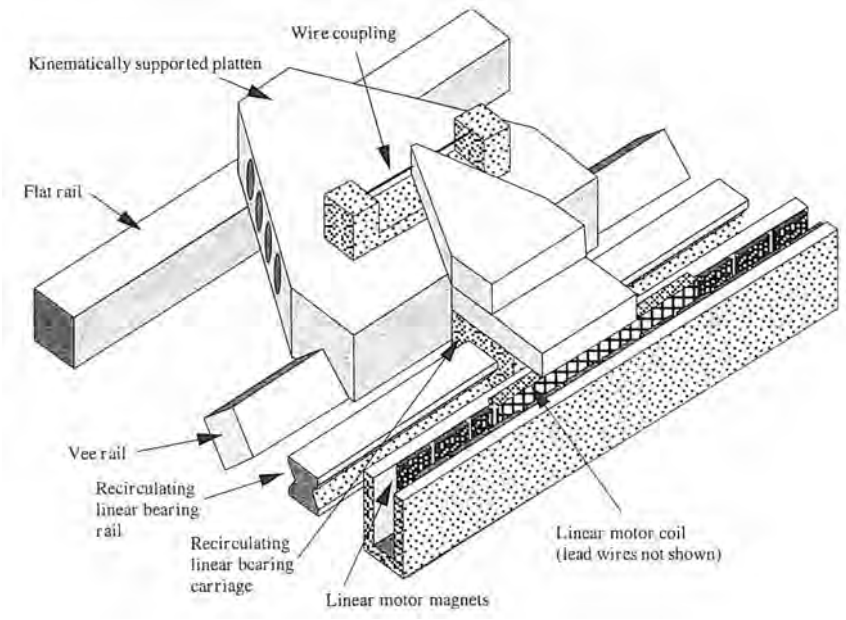


Figure 10.8.42 Instrument platten with sliding contact bearings, linear motor, and wire-type flexural coupling.

for bending and shear deformations.[110] There are also many other different geometric configurations for membrane-type couplings, including one that is manufactured integral with a precision leadscrew nut, as was shown in Figure 10.8.17.

A beam-type kinematic transmission system, as shown in Figure 10.8.40c, uses a long rigid beam to attain high axial stiffness while being supported at its ends by a flexural hourglass shaped spring and a soft spring respectively. This type of flexural coupling is not as good as a wire for filtering out errors, but is much stiffer axially. Some companies have used a kinematically supported beam and friction drive for the drive system of their coordinate measuring machines. Some CMMs, however, use touch trigger probes to "measure on the fly," and thus overtravel and dynamic performance of the linear axes are not always as important as they are for a machine tool.

The axial compliance of a beam-type coupling is easily computed by evaluating the integral of the differential compliance element over the length of the beam where dc = dx/EA. As one integrates over the length of the beam, one needs only to be careful to include the effect of the area changing with position near the attachment point. The stiffness is then just the inverse of the compliance. The lateral compliance of the coupling will depend on the axial position along the beam. One can model the beam as a rigid beam attached at one end to a wall with a torsional spring and at the other end with a linear spring. The equivalent angular stiffness of the hourglass region can be found using methods described in Section 8.6.4.

Coupling Summary

Most machine tools couple the actuator (e.g., ballscrew) directly to the carriage in order to maximize axial stiffness. With careful grinding and fit-up of components, the machine may typically experience coupling errors in axis straightness on the order of 5-10 μm. When the components are hand finished during assembly (e.g., scraped), the error can drop to 0.5-5 μm. With hand lapping or the use of a coupling element, an order-of-magnitude increase in performance can be expected. As shown in Section 10.9, when high accuracy is needed but one cannot afford the lapping process or the room for a paddle-type coupling, sensors and software can make up for the lower axial stiffness of flexural couplings.

[110] See Section 8.6.5 for a detailed analysis of this type of coupling.

10.9 Design Case Study: Increasing Axial Stiffness Through Control[111]

This section investigates how the axial stiffness of a flexural coupling can effectively be increased through the use of added sensors and control software. This method was developed for use with a wire-type coupling on the atomic resolution measuring machine (ARMM)[112] developed by MIT and the Precision Engineering Division of the National Institute of Standards and Technology. The method requires the system to measure the position of the platten at the points where the coupling is attached to the platten and where the actuator is attached to the coupling. The stretch of the coupling can then be determined and compensated for with the motion of the actuator if the latter has suitable bandwidth.

The ARMM's design goal is a spatial resolution of 10^{-10}m over a range of 0.1 m by 0.1 m, and has potential uses in many areas such as molecular biology, integrated circuit fabrication, and material science. A single-axis prototype of the ARMM with a kinematic arrangement of magnetic bearings is shown in Figure 8.8.10. It consists of a slide supported kinematically by magnetic bearings and actuated by an incremental motion piezoelectric actuator that is coupled to the slide via a wire-type flexural coupling. An alternative would be to use a linear electric motor whose moving coil was supported by an air bearing slide. A beam from the air bearing carriage would then be attached to the center of the wire. In this manner, the straightness errors in the slide and heat generated from the motor could be isolated from the ARMM platten.

With the use of a flexible wire coupling, lateral errors from the actuator will be greatly attenuated; however, the axial responsiveness of the platten will also be greatly reduced. The physical analogy to this problem is using a bending credit card to push a book across a table and stopping it at a desired point. If the observer cannot see how much the card is bent, it is difficult to tell when to stop pushing. When the observer stops moving one end of the card, the energy stored in the bent card keeps pushing the book forward for a bit. The result is a classic limit cycling problem. With the use of a frictionless bearing, the friction component of limit cycling is removed, but the same overshoot problem remains. Only when the observer can see how much the card is bending can he or she more accurately tell when to stop pushing. This crude experiment was the motivation for the development of the simple algorithm described below.

Control System Design

Much work has been done on control of flexible systems, but this work has primarily addressed control of structural mode shapes and not necessarily increasing resolution of precision systems.[113] Research on precision axial control of mechanical slides has focused on coarse-fine systems. However, to decrease complexity and increase reliability, a method for controlling a single actuator/transmission/slide system is desired. Even if a fine motion stage is to be added, the better the coarse motion stage can be controlled, the greater the accuracy of the combined coarse-fine system.

Figure 10.9.1 shows a fourth-order model of the actuator, wire coupling, and platten. The motor is modeled as a force source that acts on a mass dampened by friction. The wire transmission is the dominant spring which connects the motor (mass) to the platten, which is also modeled as a mass and damper. The position of the slide cannot be controlled directly as the force exerted on the wire (and the slide) by the motor is unknown. The only way to determine the force in the spring with high resolution is to accurately measure the axial deflection of the spring (wire).

Since the force of the motor acting on the slide is unknown, current feedback from the motor cannot be used to determine accurately the force being exerted on the platten. Both the position of the platten and the motor must be used, in combination with the spring length, to determine the force. When the platten arrives at the desired point, the motor must *pull back* so that the wire spring is no longer compressed. Otherwise the force stored in the spring will cause the platten to continue

[111] This case study represents a condensed section of D. Thurston's (my wonderful wife) M.Sc. thesis Design and Control of High Precision Linear Motion Systems, MIT, Electrical Engineering Department, April 1989. Also see Section 8.8.2.

[112] A. Slocum and D. Eisenhauer, "Design Considerations for Angstrom Resolution Machines (ARMs)," NASA Conf. Magn. Suspens.Technol., Hampton, VA, Feb. 2-5, 1988.

[113] See for example W. J. Book et al., "Feedback Control of Two-Beam Two-Joint System with Distributed Flexibility," ASME J. Dyn. Syst. Meas. Control, Dec. 1975, pp. 424–431, and R. C. Burrows and T.P. Adams, "Control of a Flexibly Mounted Stabilized Platform," ASME J. Dyn. Syst. Meas. Control, Sept. 1977, pp. 174–182. Much work has since been done on this subject.

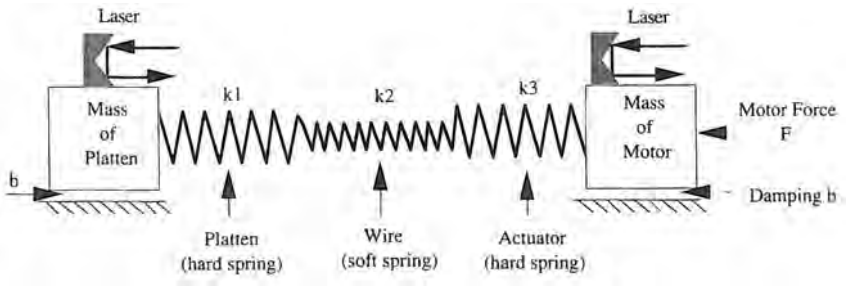


Figure 10.9.1 Dynamic system model of a wire-type kinematic transmission.

to move. The problem is to determine how much to tell the motor to pull back when the platten stops moving.

It turns out that the pullback can be obtained simply by implementing a PID loop around the position feedback from the platten and a second loop (e.g., PD or PID) using the position feedback from the motor carriage. The outputs from the PID loops are then added and used as an input to the motor. Thus the system is a single-input (voltage to motor), multiple-output (two position measurements) system. The logic used in arriving at this double PID algorithm is simple: If the distance between the actuator and platten at rest is a stable constant, then after a move, the stable distance must again be established. Thus a closed-loop algorithm is used around the position of the platten and the motor so that the difference between the two will be the equilibrium separation.

For example, where x_1 and x_2 are the positions of the platten and motor carriage and u is the control force output to the motor, a section of the digital servo algorithm would look like:

1 Output u to DAC
Read Lasers
$e_1 = x_1\text{desired} - x_1\text{actual}$
$e_2 = x_2\text{desired} - x_2\text{actual}$
$u_1 = a_{11}*e_1+a_{12}*e_{1old1}+a_{13}*e_{1old2}+a_{14}*u_{1old1}+a_{15}*u_{1old2}$
IF ($u_1 > u_{max}/2$) THEN $u_1 = u_{max}/2$
$u_2 = a_{21}*e_2+a_{22}*e_{2old1}+a_{23}*e_{2old2}+a_{24}*u_{2old1}+a_{25}*u_{2old2}$
IF ($u_2 > u_{max}/2$) THEN $u_2 = u_{max}/2$
{ store old values }
$u = u_1 + u_2$
Wait for timing, then GOTO 1

The primary point here is that the position of the motor and the platten must both be measured and used in the digital controller difference equations. One could try a PID loop around the platten and a PI loop around the motor to avoid overconstraining the system.

The equations of motion for the mechanical system are

$$m_1 \ddot{x}_1 + k(x_1 - x_2) = F - b_1 \dot{x}_1 \tag{10.9.1}$$

$$m_2 \ddot{x}_2 + k(x_2 - x_1) = -b_2 \dot{x}_2 \tag{10.9.2}$$

For this model the system has a single input, the motor force, and single output, the position of the platten. The platten and motor masses are assumed to be $m_1 = 2$kg and $m_2 = 5$kg. It is assumed that the wire coupling described in Section 10.8.1 is used. The damping effects are assumed to be linear and uncoupled for this model. The viscous damping b_1 is calculated assuming a damping coefficient[114] ζ of 0.7 and as if each mass were attached to a wall by a spring instead of to each other.

[114] A damping coefficient $\zeta = 0.7$ represents a well-damped system with relatively fast response time and no overshoot. This could be realized in a magnetic bearing system with viscous (oil tub) damping or with a platten supported by sliding bearing pads.

Field Search for Best Controller Coefficients

A linearized model and linear control design methods can be used to obtain the controller coefficients on the order of the best ones, using standard canned controller design programs. For fine-tuning to achieve ppm or ppb resolution, the coefficients for the system need to be tested in real time and then fine tuned accordingly. A field search program provides the best method for finding the best coefficients first numerically and then in real time. Field searches are usually done on the machine itself by turning the dials on an analog decade box until the system response is satisfactory. For this study, a program was written which conducts a numerical field search for the "optimal" coefficients while incorporating all the non-linearities (e.g., saturation, resolution, and deadband effects) in order to find the best coefficients for the controller.[115] Figure 10.9.2 shows a flowchart of the program. This same program's search algorithm can then be used with the hardware implementation to fine tune the coefficients for the physical system.

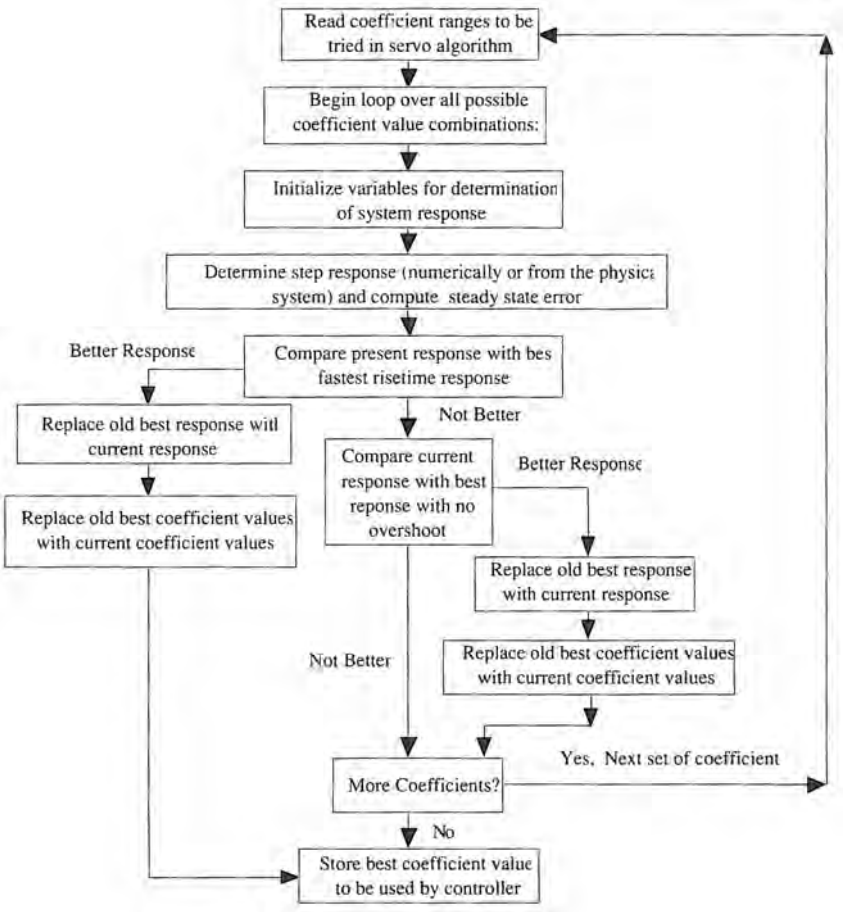


Figure 10.9.2 Search algorithm flowchart.

The program is designed to select the best step response with no overshoot and the fastest response with minimal overshoot and settling time. Note that this program is not intended to be a "canned program" but a program specific to a particular application. The user enters the parameters of the system, including the describing system equations and nonlinearities. By making the program less general, the program has fewer inputs, can run faster, and can incorporate more nonlinearities.

Results and Conclusions

The program was used to find "optimal" coefficient values for the step response with best rise time with no overshoot and best rise time with maximum specified allowable overshoot, for the following cases:

[115] A similar search algorithm was used for a hydraulic robot. See A. Slocum, "Design and Implementation of a Five Axis Robotic Micromanipulator," Int. J. Mach. Tool Des., Vol. 28, No. 2, 1987, pp. 131–139.

- PID for the ideal system with no nonlinearities
- PID for the ideal system with nonlinearities
- PIDPID for the ideal system with nonlinearities

The numerically simulated step responses are shown in Figure 10.9.3, and significant improvement is obtained with the use of the PIDPID control algorithm.

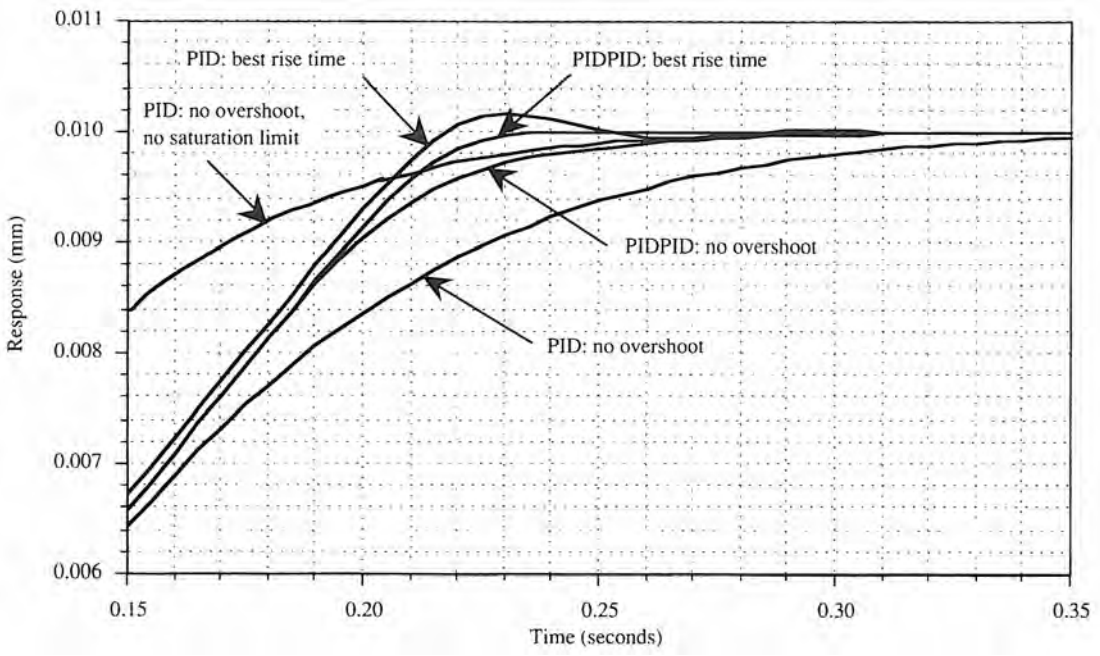


Figure 10.9.3 Two degree-of-freedom system simulated fastest rise time with saturation and resolution nonlinearities included in model.