Principles and Techniques for Designing Precision Machines

by

Layton Carter Hale

Case Western Reserve University, B.S.M.E., 1982

Massachusetts Institute of Technology, M.S.M.E., 1990

Submitted to the Department of Mechanical Engineering
In Partial Fulfillment of the Requirements
For the Degree of

Doctor of Philosophy in Mechanical Engineering
at the
Massachusetts Institute of Technology

February 1999

Signature of Author_____
Department of Mechanical Engineering

Certified by _____
Professor Alexander H. Slocum
Chairman, Thesis Committee

Accepted by _____
Professor Ain A. Sonin
Chairman, Department Committee of Graduate Studies

Principles and Techniques for Designing Precision Machines

by

Layton Carter Hale

Submitted to the Department of Mechanical Engineering
on January 8, 1999
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy in Mechanical Engineering
at the
Massachusetts Institute of Technology

Abstract

This thesis is written to advance the reader's knowledge of precision-engineering principles and their application to designing machines that achieve both sufficient precision and minimum cost. It provides the concepts and tools necessary for the engineer to create new precision machine designs. Four case studies demonstrate the principles and showcase approaches and solutions to specific problems that generally have wider applications. These come from projects at the Lawrence Livermore National Laboratory in which the author participated: the Large Optics Diamond Turning Machine, Accuracy Enhancement of High-Productivity Machine Tools, the National Ignition Facility, and Extreme Ultraviolet Lithography. Although broad in scope, the topics go into sufficient depth to be useful to practicing precision engineers and often fulfill more academic ambitions.

The thesis begins with a chapter that presents significant principles and fundamental knowledge from the Precision Engineering literature. Following this is a chapter that presents engineering design techniques that are general and not specific to precision machines. All subsequent chapters cover specific aspects of precision machine design. The first of these is *Structural Design*, guidelines and analysis techniques for achieving independently stiff machine structures. The next chapter addresses dynamic stiffness by presenting several techniques for *Deterministic Damping*, damping designs that can be analyzed and optimized with predictive results. Several chapters present a main thrust of the thesis, *Exact-Constraint Design*. A main contribution is a generalized modeling approach developed through the course of creating several unique designs. The final chapter is the primary case study of the thesis, the *Conceptual Design of a Horizontal Machining Center*.

Thesis Supervisor: Prof. Alexander Slocum, MIT

Thesis Committee: Prof. Carl Peterson, MIT
Prof. John Lienhard V, MIT
Dr. Robert Donaldson, LLNL

Acknowledgments

Many people contributed to this thesis in a variety of different ways, and I would like to thank and acknowledge these people for their knowledge, their ideas and their efforts. The thesis committee consisted of Professors Alexander Slocum, Carl Peterson and John Lienhard V all of MIT and Dr. Robert Donaldson who was my supervisor at the Lawrence Livermore National Laboratory (LLNL) from 1990 to 1993. If I please only one person with this thesis, I hope Bob Donaldson is that person. I thank Alex Slocum for motivating me to begin a doctoral program and for being a role model for professional achievement. My thesis committee has been unduly patient and tolerant of probably too few interactions drawn out over too long of a time. Their suggestions helped make this thesis better for you, the reader, for which I am very grateful. Leslie Regan and the staff at the ME Graduate Office helped me many times during my enrollment at MIT; thank you for being so good.

My thanks go to two well-known figures in Precision Engineering, Tyler Estler from NIST and James Bryan formally from LLNL, who provided many valuable comments on an early draft. They did so out of dedication to the field.

A number of my colleagues reviewed specific sections and/or collaborated on the projects used as case studies in the thesis. Debra Krulewich reviewed sections on error budgets, separation techniques, transformation matrices and least-squares fitting. Todd Decker reviewed the introduction to exact-constraint design. Eric Marsh reviewed the chapter on damping. Rick Montesanti reviewed the chapter on practical exact-constraint design. Terry Malsbury reviewed the EUVL examples of exact-constraint design. Jeff Klingmann reviewed the chapter on the conceptual design of a horizontal machining center. These colleagues contributed in various ways: Jeff Cardinal, Steve Jensen, Maggie Jong, Karen Lindsay, Stan Locke, John Parker, Hooman Tajbakhsh, Rick Thigpen, Don Yordy.

In addition, I would like to thank and acknowledge those who were influential in my career development as a design engineer. My father's interest in mechanical projects and machinery infused me and ultimately set my course. I learned volumes of practical design know-how from Bob Edwards and Dave Wood who were my first engineering supervisors. My thanks go to many unknown designers of machines that I have observed and studied. Special thanks go to Bob Donaldson, Steve Patterson and many others for creating and documenting the best model for precision machine design that I know, the Large Optics Diamond Turning Machine. It would be a national shame if that machine is not maintained to state-of-the-art condition.

Lastly, I would like to recognize the wonderful educational policy at LLNL and to thank Dennis Atkinson, Bill Ruvalcaba, Wendy Morris, Bob Donaldson and Bob Langland who created the opportunity for me to attend MIT full time for two semesters.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

Biography

Layton Hale demonstrated mechanical aptitude at an early age while assisting his father, an industrial arts teacher, with repairs on aging farm machinery and other mechanical projects such as a unique motor home. His father and mother instilled craftsmanship and creativity through everyday life. Natural ability in math and science lead to a career in mechanical engineering that began in 1978 as an undergraduate at Case Western Reserve University. Eager to apply the wonderful tools learned at Case, he joined Cincinnati Milacron in 1982 shortly after graduating with a B.S. in Mechanical Engineering. Milacron was a great place to develop a foundation in machine design and Mr. Bob Edwards and Mr. David Woods were particularly influential. Layton has been a registered professional engineer in Ohio since 1987.

Layton began graduate studies at the Massachusetts Institute of Technology in 1988. His master's thesis, Differential Feed Control of Flexible Materials, was conducted at the Charles Stark Draper Laboratory with guidance provided by Prof. Harry West of MIT and Mr. Edward Bernardon of Draper Labs. The MIT experience was very rewarding and two people were very influential. Prof. Alex Slocum inspired a specialty of sorts in Precision Machine Design, and later encouraged Layton to undertake a doctorate program. Prof. Carl Peterson hired Layton as a teaching assistant and recommended him for employment at the Lawrence Livermore National Laboratory (LLNL).

Layton joined LLNL in 1990 and worked at the Large Optics Diamond Turning Facility directed by Dr. Robert Donaldson. LLNL is a well-known center of excellence in the field of Precision Engineering due in large part by the contributions of Dr. Donaldson toward the Large Optics Diamond Turning Machine. Under the continuing education program at LLNL, Layton began his doctorate program in 1992 by attending MIT for two semesters to complete the course work. Much of the research for this dissertation was conducted on various precision engineering projects at LLNL.

Preface

I have been an engineer for the better half of my lifetime, and it is very rewarding to be in a profession where you can learn and grow throughout your career. I am just now completing my doctorate degree, and I wonder what the rewards will be for all of the effort that went into this thesis. Certainly the knowledge that I learned and created to prepare this document will make me more effective as an engineer with better tools to design new machines. I hope that many readers will study and use these tools for their own rewards. I hope that they too will experience satisfaction whenever a design solution is somehow better because of something learned in these pages.

I draw upon several projects conducted at the Lawrence Livermore National Laboratory (LLNL) as examples to show the application of precision engineering principles, and in many cases these projects motivated new developments and innovations. Through these projects, I have had the opportunity to work with some very intelligent and dedicated engineers, machinists and technicians with varying experience levels and backgrounds. I have observed that new engineers immersed in the *precision culture* at LLNL learn the precision engineering principles quickly and apply them to great advantage. Too often engineers battle errors without understanding the fundamental sources and how best to deal with them. This thesis is written to beginning and experienced engineers alike who are on their own to develop a precision culture. They should find the chapter *Precision Engineering Principles* particularly valuable as a comprehensive presentation of the key messages from the precision engineering literature.

The main theme of this thesis is machine design using fundamental precision-engineering principles. Design is such a huge topic that several recent books on precision machine design still leave ample room for new work in this area. In carving out a niche for this thesis, I thought of my colleagues as a model audience and addressed those issues that would best enhance our design skills, and in many cases we learned together. This thesis contains the information, knowledge and ideas that I want most to share with them and you. Although a thesis may not enjoy wide circulation, it is reasonable to believe that important new information will lead to papers and inventions that will reach and inspire a larger audience of researchers. Several papers have been published in advance of the thesis and others will follow. In addition, I intend to make this thesis and the many useful computer design tools available on the internet.

One further point: this thesis is centered about my work done predominantly at LLNL. It should not imply that LLNL is the only place where such work is being done or that others are not doing comparable work. I have attempted to be thorough in referencing the information borrowed from others, but certainly I have made less effort in providing complete surveys of applicable references. I apologize to those who may not be given their due credit. Please be understanding of the constraints necessary for a thesis of this breadth.

Table of Contents

List of Figures

intentionally blank

List of Tables

1 Introduction

Principles and Techniques for Designing Precision Machines is a specialized treatment of machine design from a precision engineer's perspective. A precision engineer may have been trained as an engineer (mechanical, electrical, controls, etc.), a physicist, a chemist, a tribologist, a material scientist, a computer scientist or even a machinist. A really good one is knowledgeable in all those fields in addition to knowing much about dimensional metrology. Precision crosses many disciplines and precision machine design often requires multidisciplinary teams. Although the term *precision engineer* is a recent one, precision engineers all through history have made important contributions that shaped our world in science, technology and manufacturing in particular. Clearly a significant development was the interferometer by Albert A. Michelson in 1881 for which he received the Nobel prize in 1907. The interferometer is the instrument that transfers the international standard of length into physical measurements. Today, interferometers are essential in ultraprecision lithography machines that now support a multibillion-dollar integrated circuits industry.

> Precision Engineering is the foundation of all modern Japanese Manufacturing.
>
> H. Takeyama, President
> Kanagawa Institute of Technology, May 1993

Precision machinery will play an even greater role in the future, which presents a tremendous challenge for builders and users alike to develop the expertise required to excel in a world-wide marketplace. Precision is a new frontier for much of corporate America, and frankly, there are not enough knowledgeable precision engineers to meet current demands. Until recently, very few engineering schools offered Precision Engineering and most practicing precision engineers never had such a class in college. Generally, they are self educated from a rich body of literature and from experiences in centers of precision excellence like the Lawrence Livermore National Laboratory (LLNL), where the original research for this thesis occurred. Precision Engineering has emerged as a discipline with a set of principles as fundamental as $F = m\ a$. Those who practice precision use these precision engineering principles instinctively, and through careful attention to detail, they generally succeed where others fail or profess the impossibility. Furthermore, new engineers immersed in a precision culture learn the principles very quickly and apply them to great advantage. This thesis helps address the obvious need to educate new precision engineers by extending knowledge beyond the boundaries of one precision culture.

This thesis is written to advance the reader's knowledge of precision-engineering principles and their application to designing machines that achieve both sufficient precision and minimum cost.

1.1 Contributions of this Thesis

This thesis describes the fundamental knowledge that I learned and/or contributed while working on several nationally recognized projects at LLNL. It is rare that a thesis can draw upon projects as significant in scope and funding as these. As a result, this is not a typical thesis that focuses on solving one particularly deep problem. Instead a rather broad undertaking, designing precision machines, is given foundation as a set of principles and techniques that are adaptable to the problems that you will face and solve. The words of the thesis title were chosen carefully to emphasize design as a creative process, whereas words like methodology and procedure imply strict adherence. The various projects serve two purposes, to demonstrate the foundation and to showcase approaches and solutions to specific problems that generally have wider applications. Although broad in scope, the topics go into sufficient depth to be useful to practicing precision design engineers and often fulfill more academic ambitions. We have the benefit of many ordinary and brilliant people alike who have contributed significantly to the field of Precision Engineering through new developments presented in the literature, commercial products (instruments, machines, etc.), and specialized equipment such as those described within. I encourage you to learn and contribute to the field as I have attempted through writing this thesis.

While the chapters in this thesis have a logical organization and taken together offer a certain synergy, most can be read independently with good understanding. As a reader's aid, the following paragraphs briefly describe the main contributions of each chapter.

Chapter 2 is a synopsis of significant principles and fundamental knowledge from the precision-engineering literature. Purposely concise and well referenced, this information provides the foundation in precision upon which to build new machine designs. This information should become instinctive to the practicing precision engineer.

In contrast, Chapter 3 presents engineering design techniques that are general and not specific to precision machines. Much has been written about engineering design and no attempt is made to cover the totality of opinions on the subject. Instead, this chapter includes a review of Prof. Nam Suh's *Axiomatic Design*, a reformulation of Dr. Thomas Saaty's *Analytic Hierarchy Process* for decision making, and my thoughts on creating design solutions. Subsequent chapters cover specific aspects of precision machine design.

Chapter 4 addresses a fundamental problem confronted by precision machine designers, namely, managing compliance in structures. Usually the problem arises from a need to make machine structures independently stiff and isolated from unstable foundations, relatively speaking. Examples from elementary Structural Mechanics build intuition and provide useful guidelines. Real structures are often sufficiently complex as to require finite element analysis or scale model testing. One example is a very useful study of openings in shear panels. A case study of a machining-center column demonstrates the importance of shear stiffness in structures, a point often overlooked by many designers.

Chapter 5 addresses a complementary issue, the damping (or imaginary stiffness) in structures. This chapter presents design techniques for achieving deterministic damping, in other words, damping designs that can be analyzed and optimized with predictive results. The primary emphasis here is in understanding the different damping mechanisms, their advantages and disadvantages, and in using analysis tools to evaluate options and to guide the design toward specific requirements. While much of the work presented is theoretical, damping experiments do confirm the calculations within reasonable accuracy.

Chapters 6 and 7 present a main thrust of the thesis, exact-constraint design. Chapter 6 presents practical design theory for kinematic couplings and flexures (Section 2.6 presents the basic exact-constraint theory). A main contribution of this thesis is a generalized modeling approach developed through the course of creating and applying this theory to several real and unique designs. Chapter 7 presents these designs as case studies. Not only are they useful examples to learn exact-constraint design, but they are useful machine components for future applications.

Preload is a central concept to the design of kinematic couplings, where a nesting force holds together six pairs of contact surfaces to exactly constrain six rigid-body degrees of freedom with zero backlash. The concept of preloaded constraints applies to more complicated mechanisms such as bearing systems and transmissions. Chapter 8 expands this concept and presents techniques for achieving practical anti-backlash transmissions.

Chapter 9 presents the primary case study of this thesis, the *Conceptual Design of a Horizontal Machining Center*. It was part of a three year project between Cincinnati Milacron, Inc. and LLNL to improve the precision of high-productivity machine tools. The conceptual design built upon this work with the opportunity to start with a clean sheet. While this study is only an example that may never be built, it generated a unique collection of ideas, strategies, analysis techniques and numerical examples that will be valuable to practicing machine tool designers as they develop next-generation products.

This thesis also includes several appendices that typically are more analytical in nature. They allow inclusion of very useful reference information that otherwise would be too specific and also help streamline the main chapters. Appendix A presents transformation matrices useful for modeling complex kinematic problems. Appendix B presents particular applications of least-squares fitting. Appendix C presents the basic Hertz equations for contact mechanics and more specialized equations that include effects of friction. These three appendices contain analysis programs based on the material presented within. The remaining appendices D through G break out particular details from the main chapters.

1.2 Introduction to Case Studies

These four case studies cover a diverse spectrum of precision design problems and solutions, and provide the basis for the original work in this thesis. This introduction provides the reader with basic background information, the main challenges and difficulties faced, and notable lessons learned. The case studies are presented here in generality so that specific points can be made as the topics arise throughout the thesis. The order chosen is chronological as the author participated and covers a span of about eight years.

1.2.1 The Large Optics Diamond Turning Machine (LODTM)

The Large Optics Diamond Turning Machine, shown in Figure 1-1, is an ultraprecision, vertical-axis lathe that is capable of machining a workpiece up to 64 inches in diameter by 20 inches high and 3000 pounds in weight. Two orthogonal axes in a stacked-slide arrangement move the cutting tool through an X-Z range of 37 by 20 inches along an arbitrary programmed path. Over this rather large work volume, the LODTM produces mirror surfaces in diamond-turnable materials to extraordinary accuracy. Its rated figure accuracy is 28 nm rms (nanometers root-mean-squared) with surface finish ranging from 5 to 10 nm rms depending on work materials, tool conditions and feed rate. Operational since the early 1980's, it continues to produce parts that typically no other facility can produce.

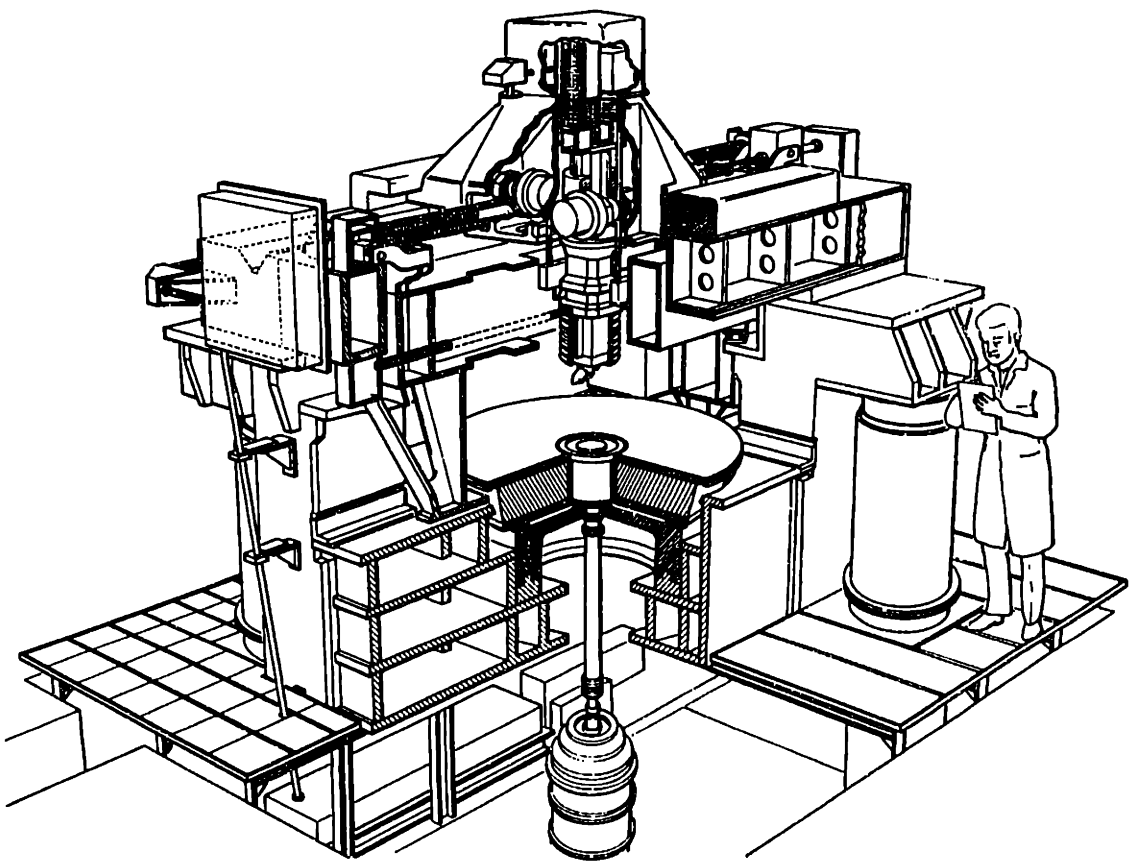


Figure 1-1 Cut-away view of the Large Optics Diamond Turning Machine.

At this level of accuracy, many error sources become significant and must be eliminated, isolated, controlled or compensated in some way. The LODTM design team pioneered the use of the error budget as a predictive tool for machine-tool accuracy. They considered every imaginable error source then made estimates (based on experience, analyses and experiments) of the effects on accuracy. They used the error budget to determined the priority and effort invested in managing the various error sources to acceptable levels. They realized that every component in the metrology loop[I] must be stable to a few nanometers over the several hours that finish cuts often require.[II] Furthermore, the LODTM had to be more accurate than any of its parts could be manufactured at that time. The optical straight edges, for example, must be calibrated and error compensated in software, and the accuracy of the error map must remain nanometer stable over days, months or perhaps years.

The metrology loop and the structural loop of the LODTM are physically separate to the greatest practical extent. The only components belonging to both loops are the spindle rotor, the workpiece and any work-holding fixtures, the cutting tool, the tool holder and the lower end of the tool bar. These parts are influenced by the dynamics and thermodynamics of the cutting process, although these effects are typically small compared to the error motion and heat generated in the hydrostatic spindle bearings. Spindle error motion is measured and compensated in software, but the heat cannot be fully captured by the embedded heat exchangers of the temperature control system. The spindle rotor is made from super invar, a low-thermal-expansion alloy of nickel, iron and cobalt, to reduce its sensitivity; however, to attain rated accuracy, the spindle requires a warm-up period lasting a few hours to stabilize at a slightly higher operating temperature.

Relative movement between the tool point and the workpiece in the sensitive X-Z plane is calculated by the machine controller using feedback from a system of sensors. The mechanical slideways control only the non-sensitive direction perpendicular to this plane. Furthermore, the machine is nearly symmetric in this plane and aligned to gravity so that out-of-plane structural deformation is minimally affected by the mass of the carriage as it moves horizontally along the X-axis. To prevent structural deformation from affecting the sensitive X-Z plane measurements, an isolated metrology frame provides a stable reference for capacitance gauges that measure movement of the spindle rotor and for laser interferometers that measure movement of the tool bar. The function of the metrology frame is to remain constant in size and shape to the nanometer level even though the main frame may deform several micrometers.

[I] The metrology loop refers to the system that determines the relative position of the cutting tool point to the workpiece.

[II] The spindle speed typically used for turning is only 60 rpm, and a typical feed is 200 μin per revolution. This works out to a feedrate of 0.72 inches per hour. It would take nearly two days to complete a facing cut over the full range of the machine.

Figure 1-2 shows a schematic of the machine components and sensors that make up the metrology loop. The metrology frame is a welded box structure constructed of super invar and supported with isolation flexures from the steel main frame. It is surrounded by noncontacting panels where temperature-controlled water circulates. Four interferometers measure at two elevations the X motion of vertical straightedges mounted on either side of the tool bar. Three more interferometers measure the difference in Z motion between the lower end of the tool bar and two horizontal straightedges mounted symmetrically about the X-Z plane on the metrology frame. The interferometers have resolution of 1.27 nm. This information together with error maps of the straightedges and measurements of X and Z axis alignments to the spindle axis is sufficient to calculate the motion of the tool point at its nominal location. Four capacitance gauges measure X and Z motion at two edges of the spindle rotor, and an absolute encoder measures the angle. The capacitance gauges have resolution of 0.32 nm. This information together with the error map of the spindle is sufficient to calculate the workpiece motion at the nominal tool point location. Clearly the computer algorithm that calculates the relative position between the tool point and the workpiece is an essential part of this metrology loop.

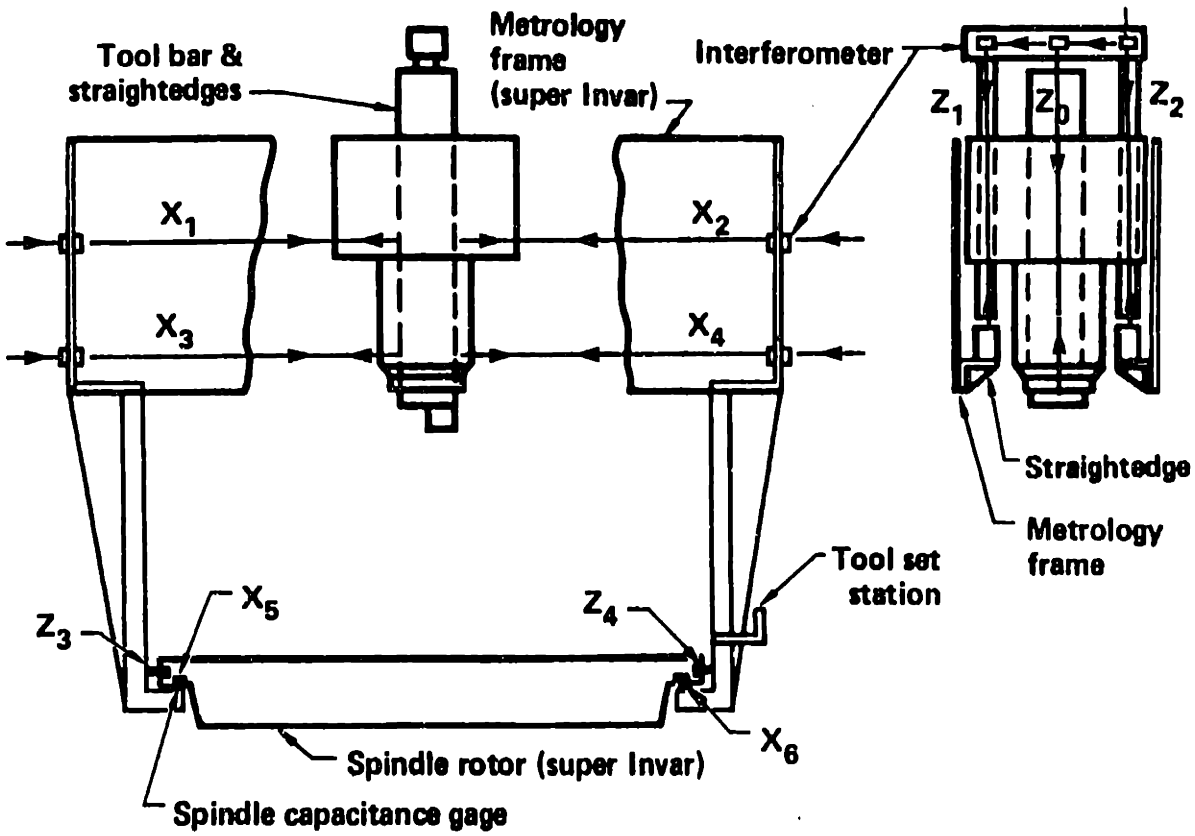


Figure 1-2 Schematic of the machine components and sensors that make up the metrology loop.

As light passes through air or any medium, temperature and pressure variations affect the wavelength and hence the distance measured by an interferometer. This problem is minimized to small air gaps between straightedges and windows at the ends of evacuated

beam pathways. Welded metal bellows allow the vacuum pathways to change length as the axes move, thus maintaining nearly constant air gaps. An evacuated bellows is self supporting like an extension spring, but the tension varies with barometric pressure and displacement. To cope with this problem, the metrology frame supports only the interferometer optics (each through a balanced pair of rolling diaphragms that provide the vacuum barrier) and the variable load transfers directly to the main frame or the carriage as the case may be. An additional bellows on each axis provides counterbalance by controlling the internal partial pressure with a servo regulator to minimize the current to the slide drive.

The least program increment for the computer control system is 0.1 μin (2.54 nm). It is quite impressive to see the machine faithfully execute such an unbelievably small move. There are several keys to achieving such fine motion control. One of course is high-resolution position feedback, but equally important is high-quality velocity feedback. Very high servo loop stiffness is possible by cascading the position and velocity loops, by compensating both loops with lead-lag filters and by using notch filters at mechanical resonances. The servo actuator consists of a brush-type DC servo motor driving through friction between a 2 inch steel roller and a 1 inch wide steel bar. A friction drive (or capstan) provides good mechanical stiffness, low parasitic motion and very low hysteresis, but lacks the mechanical advantage of a ball screw. The brush-type DC tachometer has a very high voltage constant and is tightly coupled mechanically to the servo motor yet electro-magnetically isolated. This package is thermally isolated from the machine with a double-pass heat exchanger of temperature-controlled water. The compensation circuitry is specially designed for low noise, and the servo drive amplifiers are linear rather than PWM (pulse width modulated). All the bearings for the slides and capstans are hydrostatic oil or air for the case of the tool bar. This eliminates all mechanical friction in the motion control system except for the motor and tachometer brushes.

The environmental and utility systems for the LODTM are also notable. As mentioned, temperature-controlled water flows though strategic areas of the machine. The water temperature varies less than 0.001° F from a nominal 68° F (20° C) at the distribution manifold. The air temperature within the machine enclosure (a room around the machine) varies less than 0.01° F if humans are excluded. Vibration induced by pumps and flowing fluid is a major concern. A gravity-feed system supplies water at a constant pressure, and the flows through the machine are laminar. A pair of air blow-down tanks alternately supplies very viscous oil to hydrostatic way bearings while a turbine supplies very low viscosity oil to the spindle bearings. Four pneumatic isolators symmetrically support the machine with a well-damped natural frequency of 0.5 Hz. Three pneumatic valves actively control the pressure in the isolators (two of the isolators share one valve) to maintain height and level. There is a resonance in the machine enclosure and air handling system that produces a slight signature in the machine but not enough to motivate a fix thus far.

Being a very complicated system of mechanisms, electro-magnetics, electronics, laser optics, computers, hydraulics and pneumatics, keeping the LODTM operational is a full-time job for a small team of people. The author was part of this team for approximately three years. Many components are original and approaching 20 years of age. Others have been upgraded over the years as commercial sources become available and/or technology improves. In addition, alignments and tests of stability and accuracy must be part of normal operations to achieve the full potential of the machine.

Fortunately, the technical work on the LODTM and other LLNL DTM's is well documented in papers and reports. To learn more, please see these references: [Donaldson, 1979] is the main LODTM report; [Donaldson and Patterson, 1983] and [Patterson, 1986] are general papers on the LODTM; [Bryan, 1979] is a general paper on the 84 inch DTM; [McCue, 1983] is a paper on the LODTM motion control system; [Donaldson and Maddux, 1984] is a paper on capstan slide drives; [Estler and Magrab, 1985] is a report on the calibration of the LODTM straightedges and [Estler, 1985] is a more general paper on straightedge calibration; [Patterson, 1988] is a dissertation on the stability of super invar; [Roblee, 1985] is a paper on precision temperature control.

1.2.2 Accuracy Enhancement of High-Productivity Machine Tools

Once the world leaders, the U.S. machine tool industry lost significant market share beginning in the early 1980's. This was due in large part to quality foreign-built equipment available for substantially lower cost combined with domestic complacency and a recessionary period. An increasingly important measure of quality is the precision of the parts produced by the machine tool. Manufacturers are recognizing that higher precision buys parts that assemble easier, function better and last longer. In many cases, precision manufacturing is an enabling technology that makes new products viable. The Department of Energy (DOE) recognized high-production, high-precision manufacturing as a national need and funded the Lawrence Livermore National Laboratory (LLNL) to work on a three year project with Cincinnati Milacron, a well-known domestic machine tool builder. Cincinnati Milacron supplied the project with a prototype Maxim™ 500 horizontal machining center, which was installed at LLNL for the duration. The stated goal of the project was to improve the accuracy of the Maxim by a factor of ten with minimal impact to productivity or machine cost. The intended goal was to merge technologies associated with precision and production, thus building a greater awareness of precision within Cincinnati Milacron and of production within LLNL.

The project was largely experimental; test how the Maxim performs, determine the sources of errors, modify the machine to reduce errors and evaluate the changes. Figure 1-3 shows a rendering of the Maxim from a Cincinnati Milacron advertisement. The machine tested at LLNL lacked some of the decorative sheet metal but otherwise was as shown. All tests were performed in an environmental enclosure where the temperature was controlled

to a set point or varied to test the temperature sensitivity of the machine. The first round of tests set a baseline for the machine as delivered, and we used an error budget to track the progress throughout the project. We typically use the tests described at the end of this section for assessing the accuracy of a machine and for locating error sources. We repeated tests as necessary to measure the effects of modifications made to the machine. The most significant gains in accuracy came through temperature control of the ambient air, the column structure, local heat sources and the part coolant. Squareness errors corrected by realignments accounted for the bulk of the geometric errors. To correct the rest, we developed a rapid method to map the geometric errors over the three-dimensional work volume using a new length measuring instrument, the laser ball bar [Krulewich, Hale and Yordy, 1995], [Krulewich, 1998].[1] The error map provides the information necessary for software-based error compensation. The combined result of this work was nearly an order of magnitude reduction in volumetric errors.

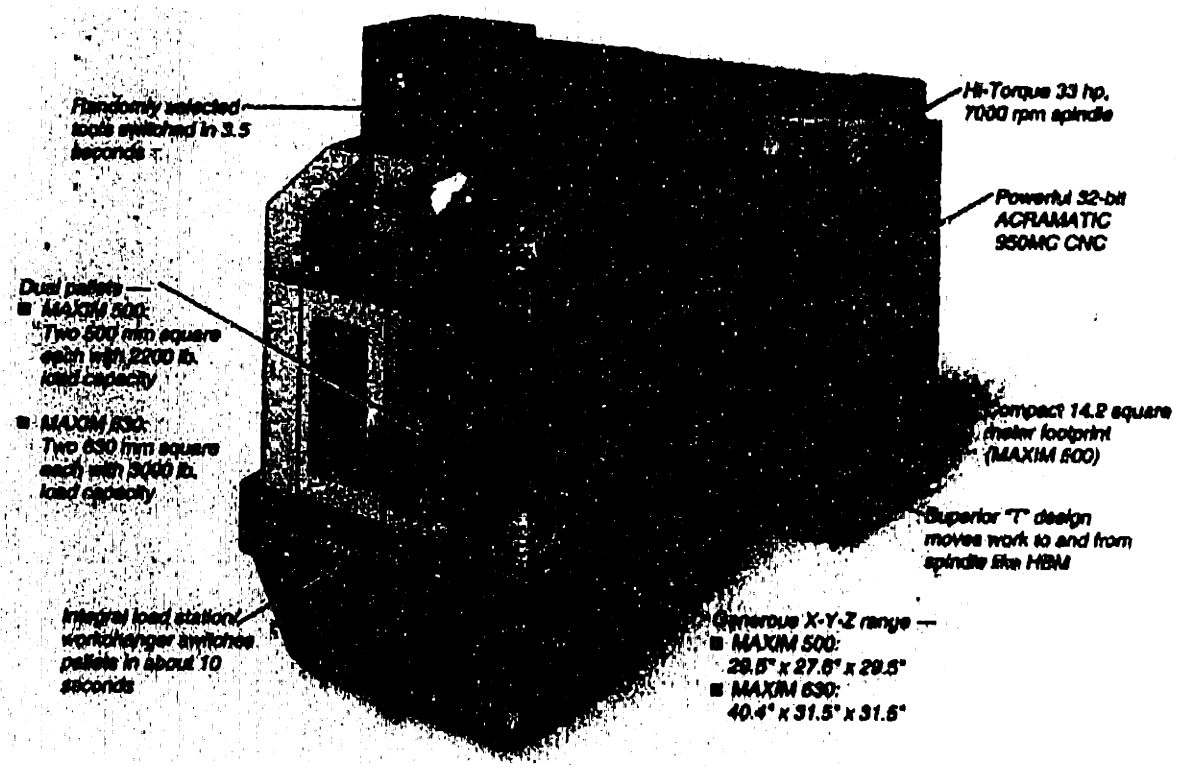


Figure 1-3 The Maxim 500 tested at LLNL has X, Y and Z servo axes with 1 micron resolution linear scales and an indexing B axis with 1° increments.

The Accuracy Enhancement project also included a broad-scope effort to develop a conceptual design of a new horizontal machining center. This approach released the constraint to work within the Maxim product on near-term solutions and allowed freer

[1] Prof. John Ziegert and students at the University of Florida developed the laser ball bar and market the instrument through Tetra Precision Inc., 4605 NW 6th St., Gainesville, FL 32609-1783, Phone: (352)335-7445.

thinking toward a next-generation product. Through this example, we could communicate methods and strategies for designing greater precision into high-production machine tools. The work on the conceptual design is primarily the author's, which is presented in Chapter 9 as the primary case study for this thesis.

Drift Test: Typically three capacitance probes measure the X, Y and Z dimensional changes over time between the spindle and the work table. The machine may be running with the servos actively holding position or it may be off to test more directly the structural changes. Often the ambient temperature is cycled through a defined profile to obtain sensitivities.

Dynamic Drift Test: Similar to a drift test, here the axes undergo a duty cycle between drift measurements. The dynamic drift test exposes self-heating effects of the axes.

Spindle Thermal Drift: The spindle cycles between measurements of drift.

Hysteresis or Step Test: The axis moves in a series of perhaps 10 equal increments in one direction then reverses over the same increments. The test should include several different sets of increments, the smallest being the program resolution. Usually the overall travel is small enough to measure the movement along the axis with a capacitance or LVDT probe.

Circle Test: The machine is programmed to follow a continuous circular path, usually full circles in the X-Y plane or partial circles in other planes. Usually a telescoping ball bar measures the relatively large circular path. This test exposes axis reversal problems and squareness errors. A pair of orthogonal capacitance probes can measure much smaller, high-frequency circles that expose dynamic problems with the control system and the machine structure.

Axis Positioning: The axis steps through its full range in both directions while a distance measuring interferometer (DMI) measures the position along an axis usually through the middle of the work volume of the machine. All modern machine controllers will accept this information for axis compensation.

Face Diagonals: The test should include a minimum of six positioning tests with DMI's performed along planar diagonals (two in each X-Y, Y-Z and Z-X plane) usually passing through the center of the work volume. This test exposes squareness errors.

Body Diagonals: The test includes four positioning tests with DMI's performed along diagonals between eight corners of the work volume. This test gives a good measure of the level of volumetric errors in the machine.

Parametric Errors: The six component error vector for each linear and rotary axis is measured throughout its travel. In addition, squareness errors between all axes are measured. These parametric errors can then be reassembled assuming rigid-body kinematics into a volumetric error map. Some modern machine controllers will accept this information for volumetric error compensation.

1.2.3 The National Ignition Facility (NIF)

The National Ignition Facility (NIF) when completed in 2003 will be the world's most powerful laser system and the first facility capable of achieving nuclear fusion and energy gain in a laboratory. Being able to create conditions more extreme than the center of the Sun, it will have far-reaching implications for national security (providing key data for nuclear weapons stockpile stewardship), fusion energy and a variety of scientific fields. LLNL is the site location and the lead laboratory in a team that includes Los Alamos National Laboratory, Sandia National Laboratory and the University of Rochester. Shown in Figure 1-4, the NIF is the size of a football stadium 704 feet long by 403 feet wide by 85 feet tall and will cost taxpayers $1.2 billion to design and build. Capable of delivering 500 billion kilowatts on target, the NIF is quite a bargain at 417 kW/$ compared to an ordinary electric light bulb at approximately 0.2 kW/$. Because the laser pulse lasts only a few nanoseconds, the energy delivered is very modest, only 1.8 megajoules or 0.5 kW-hr.

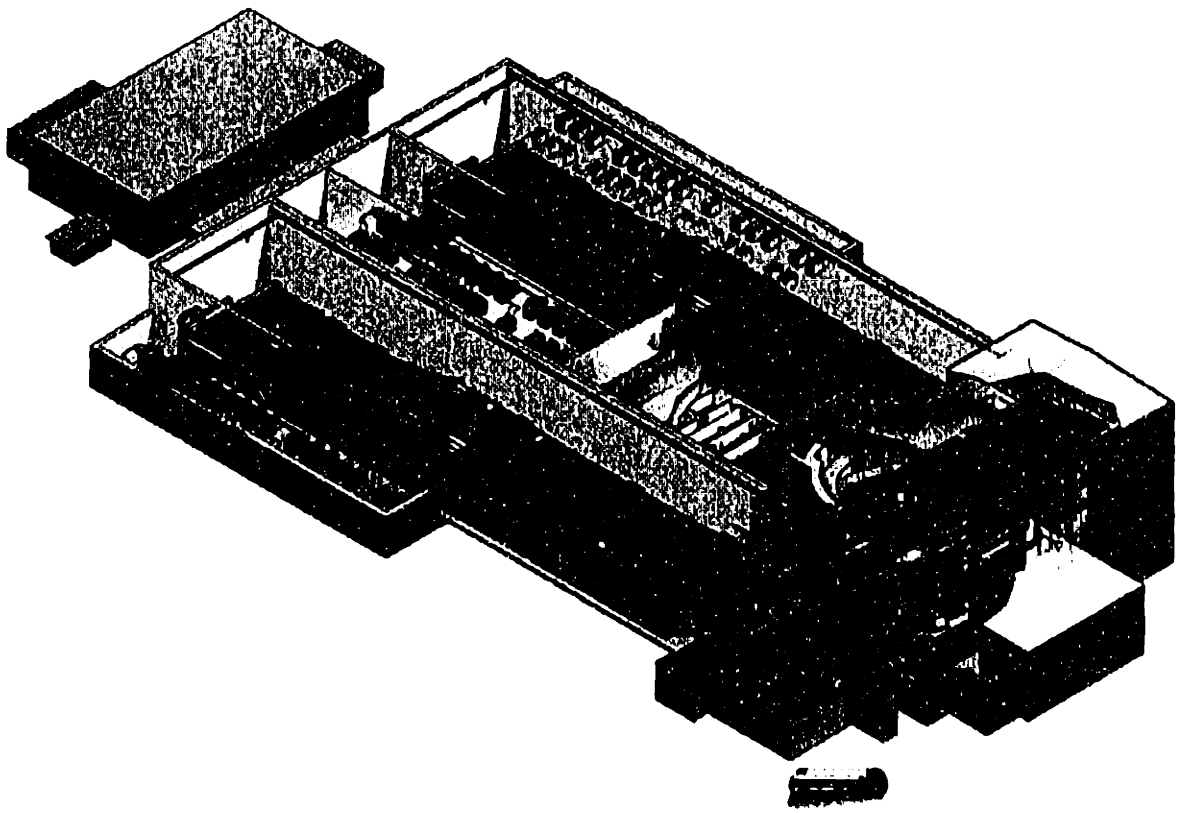


Figure 1-4 The NIF houses two separate laser bays that generate 192 laser beams each 38 cm square. The switchyard directs the beams through equal path lengths to a 10 m diameter spherical target chamber. Final optic assemblies mounted to the chamber each house a vacuum window, two frequency conversion crystals and the focus lens. Initially evacuated, the pressure in the target chamber will rise to approximately one-fifth of an atmosphere following a fusion ignition.

To achieve a self-sustained fusion reaction, the NIF will precisely focus 192 individual laser beams through the ends of a small gold cylinder that contains a BB-sized spherical capsule of deuterium and tritium fuel (two heavy isotopes of hydrogen). The

cylinder, called a hohlraum, converts ultraviolet laser light to X-rays and directs them to the capsule where an implosion occurs. This process of inertial confinement compresses the fuel to 20 times the density of lead and heats it to 100 million degrees Centigrade. This method of illuminating the capsule is called indirect drive. The NIF can also be configured for direct-drive illumination where the laser beams directly hit the capsule. Deuterium and tritium are chosen because they fuse into helium nuclei at the lowest ignition temperature of any other known fusion reactions. The initial fusion reaction occurs in the center 2% of the fuel and the energy of the helium nuclei produces a fusion *burn* wave that moves through the surrounding fuel. About 30% of the fuel will fuse releasing the energy equivalent of ten pounds of high explosives before the process extinguishes itself. The resulting radioactive waste produced at the NIF will be less than what a municipal hospital produces.

The future inertial-confinement fusion power plant will use advanced lasers or particle beams as the driver to ignite five to ten capsules per second. The glass laser driver used in the NIF can fire only once every couple of hours. Independent scientific review groups concluded that ignition and energy gain should be demonstrated first before developing a power-plant driver. They also concluded that the glass laser is the only driver capable of achieving ignition and energy gain within the next decade. The experiments conducted on the NIF over the next three decades will greatly expand scientific knowledge in the areas of nuclear fusion, plasma physics with applications to Astrophysics, dense-matter physics and Hydrodynamics (the study of fluid motion and fluid instabilities). In addition, the design, construction and operation of the NIF will greatly expand knowledge in laser science and engineering, provide new spin-off technologies and stimulate high-technology industries. For example, the NIF laser will contain 33,000 square feet of precision optics, more than all the world's telescopes put together.

There are many precision aspects to the NIF system, for example, control of pulse shape and timing to picoseconds, fast detection and diagnostic systems, beam control systems, target positioning and alignment systems, alignment of optic assemblies using advanced surveying techniques, facility temperature control, target manufacturing, and precision optics manufacturing. A particularly challenging aspect is diamond flycutting of KDP crystals. With regard to this thesis, the opto-mechanical designs conceived and developed by the author are presented. This refers to mechanical structures, mounts and actuators that support and position several types of large, reflective laser optics.

The architecture of the main laser system accommodates a dense packing of optics and provides access underneath for automated servicing of damaged optics. The constraints of this architecture and difficult requirements for stability, precise motion control, and cleanliness of optics created many design challenges. Despite the large budget, we also had cost constraints given the large numbers of items required. A good example is the precision linear actuator used in large-aperture tip-tilt mirror mounts. The NIF requires over 3200 of these actuators that have 25 nanometer resolution, and the budgeted cost is $1500 per unit.

1.2.4 Extreme Ultraviolet Lithography Projection Optics

In the opinion of John Carruthers, Director of Component Research at Intel Corporation, Extreme Ultraviolet Lithography (EUVL) is the only viable technology that will enable the microelectronics industry to continue rapid advancement over the next three decades. The financial commitment by an industrial consortium, the EUVL Limited Liability Company (LLC) led by Intel, Motorola, and Advanced Micro Devices, is $250 million over three years to bring EUVL technology to the production line by 2003. Over half of this goes to Sandia, Lawrence Berkeley and Lawrence Livermore National Laboratories, each having developed key technologies for EUVL. Sandia is responsible for advanced source development, resist development and integration of the Engineering Test Stand (ETS), a prototype EUVL tool. Lawrence Berkeley provides 13 nm wavelength metrology using their Advanced Light Source facility. LLNL specializes in visible light metrology, mask technology, multi-layer coatings and optical design. LLNL is responsible for delivering the projection optics system for the ETS and beyond. The immediate goal for the ETS is 90 nm feature sizes but the technology may eventually reach below 50 nm features.

EUVL optics are necessarily reflective in contrast to traditional lithography optics that typically are refractive. Reflective optics require more accurate surface figure than refractive optics because the wavefront changes by a factor of two rather than n - 1, where n is the index of refraction and is of order 1.5. Achieving diffraction-limited performance with 13 nm light requires Angstrom-level surface figure over the aperture. Simultaneously, the optics must be very smooth to achieve acceptable reflectivity and low scatter. On the positive side, EUVL requires far fewer optical surfaces that have lower numerical aperture (NA), and the homogeneity of optical glass is not an issue.[I] An enabling technology for optical manufacturing and system alignment is the phase-shifting, point-diffraction interferometer (PSDI) developed at LLNL. It eliminates the need for a reference optic by creating two arbitrarily good spherical wavefronts using diffraction. Eventually optics manufacturers will be able to reach 0.1 nm rms figure accuracy required for EUVL projection optics.[II]

There are many precision aspects to Extreme Ultraviolet Lithography, for example, metrology and control of scanning mask and wafer stages, multi-layer deposition processes, mask fabrication and optics manufacturing. The area where the author contributed is the mechanical design of the projection optics system and in particular the optic mounts and the actuated alignment mechanisms. Figure 1-5 shows the final design of

---

[I] Lower NA optics typically have less aspheric departure and are easier to manufacture.

[II] The typical aspheric fabrication method requires an iteration between the material removal process and fabrication metrology where the figure is measured. This process converges only if the optic mounts are sufficiently repeatable; in other words, they affect the optic figure the same way each time. The mounted figure accuracy goal is 0.5 nm rms for set 1 optics and 0.25 nm rms for set 2 optics.

the projection optics system including the ray bundle that emanates from the mask above, reflects through the system and then images at a 4:1 reduction on the wafer below.

The key challenges are: preserving the figure accuracy of the optics, aligning the aspheric, off-axis optics and meeting difficult stability requirements. The preferred approach for preserving figure is to support the optic during fabrication metrology in the same optic mount and orientation as used in the projection optics system. Then the optic mounts need only provide repeatable support rather than gravity-compensating support. The initial mechanical alignment of optic surfaces is accomplished on a CMM (coordinate measuring machine) to an accuracy of 10 to 20 microns. The final optical alignment requires a special PSDI to inspect the wavefront and distortion errors of light passing through the projection optics system. Then remotely actuated alignment mechanisms can move optics at nanometer resolution. The stability of the system over the time scale for printing must be the same order as the figure accuracy. Long-term alignment stability is a thousand times larger but still is very challenging for the length scales involved. For example, all the structural components are made from super invar.

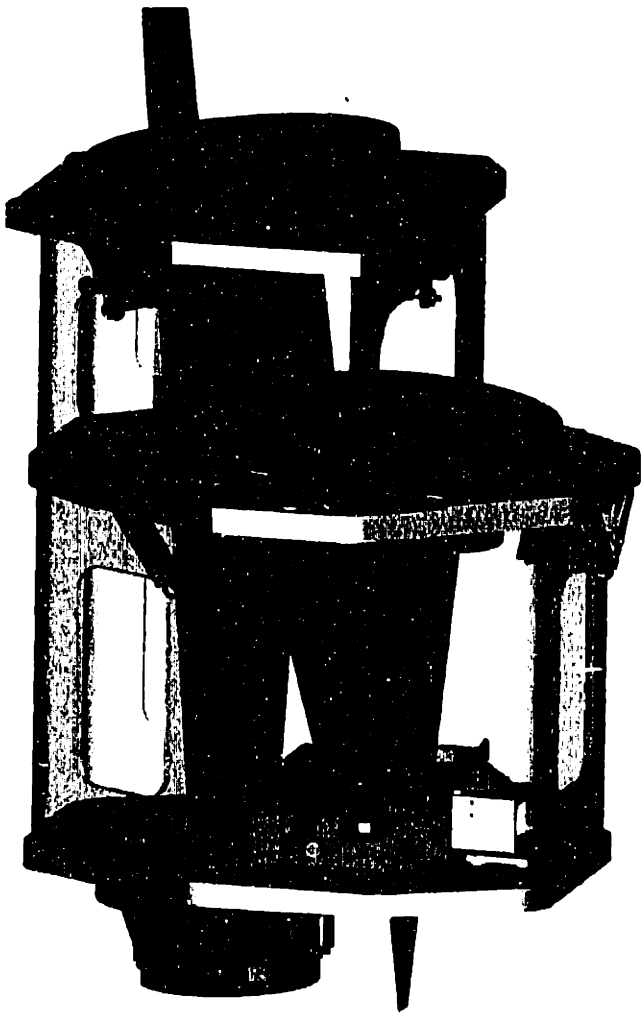


Figure 1-5 The EUVL projection optics system consists of four reflective optics, individual optic mounts, alignment mechanisms, the projection optics box and other ancillary devices.

2 Precision Engineering Principles

The Precision Engineering literature is rich in fundamental knowledge, useful techniques and strategies, studies of precision systems and processes, new ideas and developments, and valuable experiences. There are many hundreds of relevant papers and a good source for citations appears regularly in *Precision Engineering* (the Journal of the American Society for Precision Engineering). In addition, the following books cover a wide spectrum of Precision Engineering topics: *Gauges and Fine Measurements* [Rolt, 1929], *Foundations of Mechanical Accuracy* [Moore, 1970], *Precision Engineering: an Evolutionary View* [Evans, 1989], *Precision Machine Design* [Slocum, 1992], *Foundations of Ultra-Precision Mechanism Design* [Smith and Chetwynd, 1992].

This chapter presents significant principles and fundamental knowledge from the Precision Engineering literature. This information is the foundation in precision upon which to build new machine designs. The main topics in this chapter are organized into the following sections: Determinism, Alignment Principles, Symmetry, Separation of Metrology and Structural Loops, Separation of Systematic Errors, Exact Constraint Design, Elastic Averaging, Thermal Management, and Materials Selection. Some sections are long and detailed with examples and analyses. Others are brief, expressing basic ideas with simple examples. Several topics will echo throughout this thesis. Newcomers to Precision Engineering should find this chapter particularly valuable as an introduction to the key messages from the precision engineering literature.

2.1 Determinism

> A close study of an automatic manufacturing process shows that it is operating under the influence of natural laws and is not affected by the very wonderful, complicated, but slightly unpredictable mechanism known as a human being. If, as indicated above (there appears to be no known error in any natural law), all natural laws operate with 100% perfection, an automatic manufacturing process could be classified as operating perfectly. It may not be doing what is required, but if that is so it is because it has not been suitably arranged.
>
> [Loxham, 1970, unpublished]

> The basic idea is that machine tools obey cause and effect relationships that are within our ability to understand and control and that there is nothing random or probabilistic about their behavior. Everything happens for a reason and the list of reasons is small enough to manage.
>
> [Bryan, 1984]

> 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. Further, the causes are not esoteric and uncontrollable, but can be explained in terms of familiar engineering principles. These explanations are not simply educated (or uneducated) guesses, but are based on tests which are designed to isolate the sources of error. Once isolated, it is usually found that the source of error can be reduced to a satisfactory level by relatively simple and inexpensive means.
>
> [Donaldson, 1972]

Determinism is both a principle and a philosophy applicable to many systems and processes beyond automated manufacturing processes and machine tools in particular. The deterministic principle rests on the ability of physical laws to explain the behavior of systems and processes. The deterministic philosophy instills a belief that all aspects of a system or process can be understood and ultimately controlled as desired. The systematic method used to identify root sources of error and to bring them under control has become known as the deterministic approach. While often used in the context of existing machines, determinism applies to precision machines that have yet to be conceived, and so too is a principle and a philosophy for design.

The deterministic approach is an aspect of many topics in this thesis as one might expect. One topic, however, is unusual as it tests the practical limits of determinism. Appendix D *Determinism in Die Throwing and the Transition to Chaos* challenges conventional wisdom that holds die throwing as probabilistic. Specifically, it determines the level of precision required to control the trajectory of a die to a predictable state of rest. This required level of precision increases exponentially with the number of bounces until after about three bounces it transitions to chaos. Although technically deterministic, a system in chaos is unpredictable from a practical standpoint. In this regime, it is still possible to control the die's trajectory but not through initial conditions alone. It requires a deterministic approach as opposed to a statistical approach such as loading the die.

2.1.1 The Error Budget as a Deterministic Tool

The error budget is an important deterministic tool that provides a systematic way to predict and/or control the repeatable and nonrepeatable errors of a machine. The error budget is a model of the machine in its environment expressed in terms of cause-and-effect relationships. It may involve statistical quantities such as ground motion or predictable quantities such as deflections due to gravity or weights of moving axes and payloads. The error budget helps identify where to focus resources to improve the accuracy of an existing machine or one under development. It provides a useful format to specify subsystem precision requirements to achieve an overall balance in the levels of difficulty (risks and costs). Often the information required for a new machine may be estimated from tests or experiences gained through existing machines, or determined from analytical models such as finite element analysis (FEA).

The prevalent use of the error budget to tolerance optical systems inspired [Donaldson, 1979, 1980] to apply the technique to the design of the Large Optics Diamond Turning Machine. Most of the material in this section is based upon his work. In addition, [Slocum, 1992] has done much work on the geometric aspects of error budgeting using homogeneous transformation matrices, see Appendix A *Transformation Matrices*. The technical basis for the error budget rests on two premises; that the total error in a given direction is the sum of all individual error components in that direction, and that individual error components have physical causes that can be identified and quantified. A practical difficulty arises because we generally cannot quantify the errors in complete detail especially at the design stage. Although an error may vary spatially and temporally, usually the only estimate will be a bounding envelope and perhaps an approximate frequency of variation. This makes the combinatorial rule more speculative than we would like.

The first challenge in developing an error budget is to list all significant error sources that could affect the machine. Here, experience is definitely an asset but a thoughtful examination of the machine subsystems and the surrounding environment usually leads to a fairly complete list. Asking yourself and colleagues *what could possibly affect the accuracy of the machine* creates the proper frame of mind.[1] The second challenge comes in assessing the errors in the machine or process that the error sources cause. Implicit in this process is identifying the coupling mechanisms between the error sources and the point of concern. In general terms, the coupling mechanism is a time-and-space-varying dynamic system represented by a suitable transfer function, but often it simplifies to a purely static and/or stationary relationship.

As an example, ground motion, which can be represented as a power spectrum, transfers through the vibration isolation system and utility connections to the structure of a diamond turning machine leaving a recording of the error in the part being turned. In this case, the error source is ground motion and the coupling mechanism is the isolation system, utility connections and the machine structure. The transfer function could be determined experimentally or predicted using FEA. In addition, the coupling mechanism may involve the shape of the workpiece being turned since motion tangent to the surface, an insensitive direction, has negligible effect on the process. In some cases, the process may be part of the coupling mechanism, an error source or both. Processes that tend to average out errors such as polishing, honing, and grinding obviously affect the coupling mechanisms in a positive way.

It is useful to categorize workpiece errors since different error sources may have different effects. The dynamic nature of error sources and coupling mechanisms leads

---

[1] On my first visit to LLNL, Dan Thompson stated the variation of gravity due to the sun and the moon (the same cause that produces the effect of tides) was investigated as an error source for the LODTM and was determined to be insignificant but calculable.

naturally to a frequency approach. Some errors will vary as functions of time and others as functions of space, however they are related through the characteristic speeds of the process. For example, a turning process has a rather fast speed associated with rotation of the spindle and a much slower speed associated with the feedrate. Traditionally the whole frequency range has been partitioned into a few bins such as size, form, and surface finish. For an optic, the size may include the base radius of curvature and the outside diameter, both usually loose specifications. The form error, commonly called surface figure, is measured to nanometers with an interferometer for spatial wavelengths longer than a millimeter or so. Surface finish extends over the remaining shorter wavelengths (or high spatial frequency) and may be measured with a profilometer such as the AFM (atomic force microscope) or with optical techniques.[I] The trend in specifying optics (at least at LLNL) is to specify the full spectrum. This has prompted renewed interest at LLNL in a frequency-based error budget.[II]

The particular combinatorial rule to use is not obvious nor has it been rigorously determined. In the rare case that knowledge of all error components is complete, then simple addition is appropriate yielding a total error that may vary with time and space. When dealing with magnitudes, it is unlikely that all errors will occur at the same time or place and with the same sign along the direction being combined. This makes direct addition too conservative. Quadrature addition (or root sum of squares) gives the expected value of errors acting independently from one another. That is to say on average they must all be orthogonal. Experience has shown this to be somewhat optimistic. A usual tactic is to arithmetically average the conservative and optimistic estimates, although a geometric average seems more appropriate. Another sensible approach is to use the vector norm, given in Equation 2.1, with an exponent p between 1 and 2 (the values that give the conservative and optimistic estimates). For an error vector having up to 40 or so equal components, the vector norm with $p = 1.22$ closely matches the arithmetic average while $p = 1.33$ closely matches the geometric average. Using $p = \sqrt{2}$ has an intuitive appeal.

$$e_{norm} = \left[\sum_{i=1}^{n} |e_i|^p \right]^{\frac{1}{p}} \tag{2.1}$$

A further point is that all error components must be expressed in the same form before combining them. For example, it is inappropriate to combine rms values with P-V (peak-to-valley) values without first converting one set to the other. This requires an assumption of the probability density function for each error component. The probability density may be chosen as uniform, normal or some specific waveform such as sinusoidal

---

[I] EUVL optics require subnanometer rms figure and finish over 100+ mm apertures. This project is extending the state of the art in optics manufacturing and metrology.

[II] Debra Krulewich is leading this effort but there are no published results at this point in time.

depending on the nature of the error. Figure 2-1 shows the rms value corresponding to a unit P-V error for the probability density functions mentioned. Typically the uniform distribution is used unless further information dictates one of the others. If the error components have a mix of distributions, then it is appropriate to combine their rms values. The central limit theorem of probability theory states that the total error will tend to a normal distribution as the number of error components becomes large even though the individual distribution type may not be normal. An equivalent P-V error would be 4 or 6 times the total rms value for 95% or 98% confidence levels, respectively. [Shen, 1993] demonstrated this approach to agree well with Monte Carlo simulations for two published error budgets.

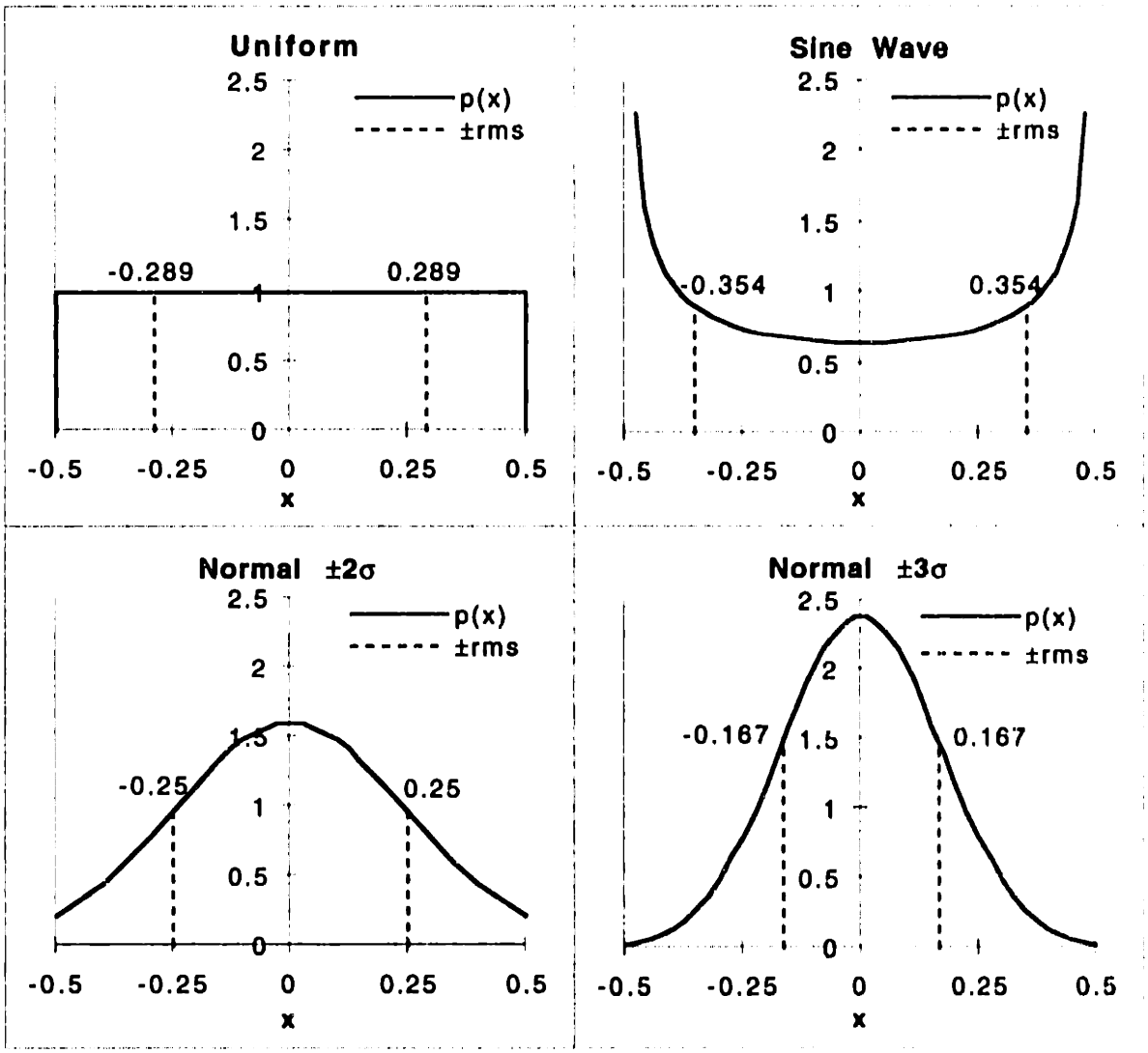


Figure 2-1 The probability density function p(x) gives the probability that the error will take on a specific value of x. The total probability that the error will fall within the bounds of ±0.5 is 100% for both uniform and sine wave, 95.4% for ±2σ normal and 97.7% for ±3σ normal. Said in terms of a unit rms value, the P-V value would be: $2\sqrt{3}$ for uniform distribution, $2\sqrt{2}$ for sine wave, 4 for ±2σ normal with 95.4% confidence, and 6 for ±3σ normal with 97.7% confidence.

Two practical examples show that error budgets can be quite simple to implement on a computer spreadsheet. The difficulties as mentioned are in developing a complete list

of error sources and applying appropriate values for error components. Table 2-1 shows the radial figure error budget for the Large Optics Turning Machine. Azimuthal figure errors and setup errors were budgeted separately. A similar error budget was developed for surface finish. Table 2-2 shows the error budget created for the accuracy enhancement project. We used it to represent and track the accuracy of the Maxim™ 500 horizontal machining center throughout the project.

Error Source	Peak-to-valley magnitudes			
	X-direction		Z-direction	
	(nm)	(μin)	(nm)	(μin)
Position interferometers				
Laser central frequency	3.0	0.12	5.5	0.22
Index of refraction	3.0	0.12	5.5	0.22
Optical, electronic factors	5.0	0.20	5.0	0.20
Straightness interferometers				
Path length difference	3.7	0.15	3.7	0.15
Optical flat difference	15.0	0.59	15.0	0.59
Index difference	3.5	0.14	3.5	0.14
Resolution-metrology frame drift			6.3	0.25
Control system				
Servo-controlled tool mount	10.0	0.39	10.0	0.39
Temperature control				
Metrology structural loop	10.0	0.39	7.5	0.30
Spindle growth			40.0	1.57
Workpiece thermal boundary layer	30.4	1.20	34.4	1.35
Spindle air supply			7.5	0.30
Gravitational loading	5.0	0.20		
Barometric pressure	4.0	0.16	5.0	0.20
Nonmachine factors				
Tool nose roundness	13.0	0.51	13.0	0.51
Workpiece fixture distortion	75.0	2.95	60.0	2.36
Workpiece body forces	25.0	0.98	20.0	0.79
Workpiece internal unbalance	25.0	0.98	20.0	0.79
Workpiece residual stress	25.0	0.98	20.0	0.79
RSS of P-V	95.5	3.76	91.5	3.60
RSS of RMS	27.6	1.09	26.4	1.04

Table 2-1 Radial figure error budget for the Large Optics Turning Machine. This error budget, developed in the conceptual phase of the project, represents a design slightly different from the actual machine. For example, the LODTM uses optical straightedges rather than straightness interferometers, the servo-controlled tool mount was never used and the spindle bearing is hydrostatic oil rather than air.

	POS-X (μm)	Y (μm)	Z (μm)	ANG μrad	ABBE X(m)	ABBE Y(m)	ABBE Z(m)	Notes
SOURCE OF ERROR				Work Volume: 600 X 600 X 600				
Z' AXIS KINEMATICS (Part motion)				Tool: 100 Dia. Max X 200 L Max				
Positioning Inaccuracy in Z, Dzz			5.0					
Nonstraightness of Z in X, Dzx	5.0							
Nonstraightness of Z in Y, Dzy		5.0						
Angular motion of Z about X (pitch), Ezx		3.0	3.0	5.0		.600	.600	y,z: Work Vol
Angular motion of Z about Y (yaw), Ezy	3.0		3.0	5.0	.600		.600	x,z: Work Vol + Bearing Offset
Angular motion of Z about Z (roll), Ezz	6.0	6.0		10.0	.600	.600		x,y: Work Vol + Bearing Offset
Y AXIS KINEMATICS								
Positioning Inaccuracy in Y, Dyy		5.0						
Nonstraightness of Y in X, Dyx	5.0							
Nonstraightness of Y in Z, Dyz			5.0					
Angular motion of Y about X (pitch), Eyx		1.5	0.5	10.0		.050	.150	y:Tool R, z: Max Tool L - Min Tool L
Angular motion of Y about Y (roll), Eyy	5.0		0.5	10.0	.050		.600	x:Tool R, z: Max Tool L + Bearing Offset
Angular motion of Y about Z (yaw), Eyz	1.0	0.0		10.0	.000	.100		x,y: Bearing Offset
Nonsquareness of Y to Z, Syz			9.0	15.0		.600		y: Work Vol
Nonsquareness of Y to X, Syx	9.0			15.0		.600		y: Work Vol
X AXIS KINEMATICS								
Positioning Inaccuracy in X, Dxx	5.0							
Nonstraightness of X in Y,Dxy		5.0						
Nonstraightness of X in Z, Dxz			5.0					
Angular motion of X about X (roll), Exx		8.0	8.0	10.0		.600	.600	y,z: Max Tool L + Bearing Offset
Angular motion of X about Y (yaw), Exy	0.8		0.3	5.0	.050		.150	x:Tool R, z: Max Tool L - Min Tool L
Angular motion of X about Z (pitch), Exz	3.0			5.0		.600		y: Work Vol
Nonsquareness of X to Z, Sxz			9.0	15.0	.600			x: Work Vol
Spindle KINEMATICS								
Radial motion in X, Dcx	1.0							
Radial motion in Y, Dcy		1.0						
Axial motion, Dcz			1.0					
Tilt motion about X, Ecx		1.5	0.3	5.0		.050	.300	y:Tool R, z: Max Tool L + Bearing Offset
Tilt motion about Y, Ecy	1.5		0.3	5.0	.050		.300	x:Tool R, z: Max Tool L + Bearing Offset
Nonsquareness of C to X, Scx	2.3		0.8	15.0	.050		.150	x:Tool R, z: Max Tool L - Min Tool L
Nonsquareness of C to Y, Scy		2.3	0.8	15.0		.050	.150	y:Tool R, z: Max Tool L - Min tool L
B' AXIS KINEMATICS								
Position Inaccuracy, Ebb	10.2		10.2	17.0	.600		.600	x,z: Work Vol + Bearing Offset
Radial motion in X, Dbx	5.0							
Radial motion in Z, Dbz			5.0					
Axial motion, Dby		5.0						
Tilt motion about X, Ebx		3.0	4.0	5.0		.600	.600	y,z: Work Vol + Bearing Offset
Tilt motion about Z, Ebz	4.0	3.0		5.0	.600	.600		x,y: Work Vol + Bearing Offset
Nonsquareness of B to X, Sbx	9.0			15.0		.600		y: Work Vol
Nonsquareness of B to Z, Sbz			9.0	15.0		.600		y: Work Vol
TOOL								
Z Tool Position (tool set or measure.)			2.0					
Nonrepeat. of Tool Change	1.0	1.0	2.0					
THERMAL								
due to spindle motor/ bearings	5.0	5.0	25.0					variation after 30 min warmup
due to influence of cutting fluid	25.0	25.0	25.0					75 F coolant, .5 m steel part
due to environmental air (scales)	30.0	40.0	30.0					drift test scale excursion from 68 F
due to environ. (structural effects)	10	40	20					dead stick test
due to local heat sources	5.0	15.0	15.0					dyn drift test
due to hydraulics								
BASE EFFECTS								
base-foundation interaction	10.0	5.0	10.0					
Algebraic Sum	163.7	180.3	208.5	=>	6	7	8	
Root Sum of Squares	45.6	52.1	54.1	=>	2	2	2	
Average	104.6	116.2 (μm)	131.3		4.1	4.6 (mils)	5.2	

Table 2-2 This error budget was created to represent and track the accuracy of the Maxim 500 horizontal machining center throughout the accuracy enhancement project. The numbers, however, are fictitious to protect proprietary information.

2.1.2 An Editorial Note on Determinism

For three decades, the precision community at LLNL, notably Bryan and Donaldson (both retired), have applied the deterministic approach to machine tool accuracy and precision manufacturing with remarkable success. The belief that nonrepeatable behavior in a

machine or process has underlying causes, that tests can be devised to find the causes and the relationships, and that ultimately those errors can be controlled, is truly a powerful point of view. The opposite view that nonrepeatable behavior is inherent and unavoidable leads naturally to statistical methods to describe the behavior. Not surprisingly, statistical methods have been associated with the defeatist point of view that rely on them. Statistics and determinism do coexist. We use statistical and other mathematical techniques frequently to uncover error sources from test data. There are several examples in this thesis. See for example, Section 2.4 *Separation of Systematic Errors* and Appendix B *Least-Squares Fitting*.

The last words on determinism go to Bryan as he argues against those who try to represent nonrepeatability as a statistically random event represented as a fixed value. This view gives up on the fundamental cause-and-effect relationships as being too hard to uncover, while the effort made to determine the statistic is a misdirected endeavor.

> A determinist will never agree that a fixed value of nonrepeatability can be assigned to a given machine. Such a value does not exist. Nonrepeatability depends primarily on the time, money, and skill (culture) of the user.
>
> A determinist will be happy to agree that some level of apparent nonrepeatability does exist for a given machine, on a given day, for a given series of tests, programmed in a given way, conducted by a given person, in a given environment consisting of a given level of temperature variation, vibration and dirt, using a given set of instruments with a given limitation on time and money.
>
> A determinist might also agree that the assumption of a Gaussian distribution and the calculation of sigma values, although technically invalid, might be useful as a statement of the level of the above variables existing at the time of a test. (A determinist would not want to waste the time necessary to make enough runs to calculate a three sigma value. He would first make a hysteresis test, and then a thermal drift test. He would then make a sufficient number of repeatability runs to determine that the range of nonrepeatable values is compatible with the resolution and quality of the machine and the time and money available. If not, he would stop and fix the problem. He would then proceed to measure the systematic errors concentrating on finding the worst points as soon as possible. If the systematic errors are close to the tolerance level, he might make additional repeatability runs at those points.)
>
> A determinist will have great doubt and anguish in ever agreeing that some level of nonrepeatability is inevitable regardless of the time, money, and skill available. This issue is negotiable, however.
>
> [Bryan, 1984]

2.2 Alignment Principles

The first principle of machine tool design and dimensional metrology is the Abbé Principle, which expresses the possibility, indeed the inevitability, of a *sine* error whenever the distance measurement and the scale or master do not lie along the same line but rather are separated by what has become known as an Abbé offset. The term sine error indicates that the error mechanism is the angular motion of the slide system acting through a lever arm (the Abbé offset). The original Abbé Principle was modified by Bryan to allow the use of modern techniques to reduce the sine error to an acceptable level either by controlling the slide to have zero angular motion or by measuring the angular motion and compensating for the sine error. Either approach can deal with a variable Abbé offset. Bryan also recognized that a sine error could occur in a straightness measurement, thus he proposed the Bryan Principle to address this problem.

> The measuring instrument is always to be so constructed that the distance being measured is a straight line extension of the graduations on the scale that serves as the reference. ... Should the measuring axis and that of the scale belong to two different axes, which are separated by a certain distance, then ... the length being read off will be identical to the length being measured in general only when the moving system ... undergoes pure parallel motion, with no rotation. If the system undergoes a rotation between the initial and final settings, then the scale reading and the measured length are different.
>
> [Abbé, 1890]

> The displacement measuring system should be in line with the functional point whose displacement is to be measured. If this is not possible, either the slideways that transfer the displacement must be free of angular motion or angular motion data must be used to calculate the consequences of the offset.
>
> [Bryan, 1979]

> The straightness measuring system should be in line with the functional point whose displacement is to be measured. If this is not possible, either the slideways that transfer the displacement must be free of angular motion or angular motion data must be used to calculate the consequences of the offset.
>
> [Bryan, 1979]

The Abbé and Bryan Principles emphasize the significance of sine errors in a measurement system. In the context of Abbé, the concern is a changing angle (or angular motion). In addition, a sine error may include a constant angular misalignment commonly referred to as a squareness error. Usually of lesser significance is the cosine error, which occurs when the distance measurement and the scale are not parallel. The cosine function is

second order for small angles while the sine function is first order, as indicated by Equation 2.2. Typically, the constant angular misalignment is most significant for a cosine error.

$$\varepsilon_{sine} = r\sin\theta \cong r\,\theta$$
$$\varepsilon_{cosine} = l(1-\cos\alpha) \cong l\frac{\alpha^2}{2} \qquad (2.2)$$

2.3 Symmetry

> **symmetry:** exact correspondence of form and constituent configuration on opposite sides of a dividing line or plane or about a center or an axis.
>
> [The American Heritage College Dictionary, third ed.]

There are many benefits to using symmetry in designs. Symmetrical designs are simpler to analyze, require less information to build, and often allow more accurate measuring and manufacturing methods. With regard to precision machines, symmetry can significantly reduce errors that occur in directions normal to the generator of the symmetry, for example, the line or plane. Not only must the material of the object be symmetric, but the loads (forces or heat) must also be symmetric for the errors to be self canceling. Symmetry about a single plane is usually easy to achieve in a design and if possible should be aligned to cancel errors in the most sensitive direction. A second plane of symmetry is more difficult or perhaps impractical to achieve globally, but efforts to create local symmetries are also beneficial.

The LODTM is an example of a design with nearly complete symmetry about two planes that intersect the spindle axis. Symmetry about the X-Z plane is most complete but the errors canceled are in the insensitive Y direction for turning. However, it is still important to keep those errors small since the Y direction is uncompensated. The main asymmetry is in the carriage bearing arrangement; two sets of two bearings ride in a vee way on one side and one bearing rides on a flat way on the other side. Symmetry about the Y-Z plane is as complete as it can be for a machine with an X-axis. That is, the LODTM is most symmetric about the Y-Z plane when the X-axis is at a position where the tool bar is centered on the spindle center line. At other positions, however, there is little consequence because the significant error sources remain symmetric and tend to cancel one another.

There are several points to make regarding the definition of symmetry given above. A higher degree of symmetry corresponds to a lower order generator. For example, an object that is symmetric to a line is also symmetric to a minimum of two planes that contain the line. This implies that a higher degree of symmetry occurs when the object is generated by a revolving operation rather than a mirroring operation. For example, a cone is more symmetric than a pyramid.

2.4 Separation of Metrology and Structural Loops

The metrology loop is a conceptual path through all the physical parts, sensors and controls in the machine that determine the location of the tool or probe with respect to the workpiece. A change to any of these parts due to a force or a temperature change will cause a measurement error. To reduce errors, certain loads (force or heat) may be eliminated entirely or partially from the metrology loop by providing an independent structural loop to carry those loads.[1] We may state the principle as follows:

> The metrology loop and the structural loop(s) should exist separately and independently to the greatest practical extent. Preferably the separation would be a physical one, but it could also be an informational separation, for example, through error compensation based on a force measurement and a structural model.

The LODTM is an excellent example of putting this principle into practice. The structural loop is much like a typical vertical-axis lathe except that the Z-axis tool bar is stacked on the X-axis carriage. All measurements in the sensitive X-Z plane between the tool bar and the spindle face are made with respect to a stationary metrology frame supported by flexures from the structural frame. The variable loads on the metrology frame and the system of non-contact sensors are negligible. The remaining path (tool bar, tool, workpiece, fixture and spindle) is common to both metrology and structural loops. Precise temperature control, low expansion materials, a stable metrology frame and error compensation combine to give long-term positioning accuracy on the order of one part in 30 million (one-millionth of an inch over an X-Z range of 32 by 20 inches).

2.5 Separation of Systematic Errors

> It is useful to separate machine tool errors into two categories, repeatable and nonrepeatable errors. A machine tool is repeatable, but not necessarily accurate, if it brings the tool into the same position relative to the work each time it is commanded to do so. ... A machine tool which is repeatable is far easier to deal with in terms of errors than a nonrepeatable machine. This is true whether one is following a policy of finding and eliminating error sources or of compensating for them. It is also frequently true that if the nonrepeatability can be eliminated from a machine, then the remaining repeatable errors are surprisingly small.
>
> [Donaldson, 1972]

It would seem that the idea of having repeatable and nonrepeatable errors is somehow inconsistent with the deterministic principle. Rather, an apparently nonrepeatable

---

[1] This viewpoint is a little unusual as often the structural frame is considered the primary or main frame, but this viewpoint gives proper emphasis to the function of the metrology loop.

error simply has not been correlated to a known error source and the responsible error source is simply not under control. An apparently nonrepeatable error can be made repeatable by determining the systematic relationship to the error source, or it may be reduced to a satisfactory level by bringing the error source under control and/or by isolating it from the system. Once the significant errors are systematic, then physically correcting them or compensating for their effects are effective ways to improve accuracy.

Separation of systematic errors expresses several key ideas, and a number of very useful techniques have been developed over the years to achieve separation.

- Apparently nonrepeatable errors have error sources that can be identified, for example, by correlating drift of the machine to temperature changes in the environment. Once identified, the error(s) may be reduced to a satisfactory level by various means.
- Particular systematic errors can be measured amidst other errors. This may require devising tests that are sensitive to particular systematic errors and insensitive to others, or subtracting out other errors based on well-known models. An example of the latter case occurs for laser interferometer measurements made in air where temperature-pressure compensation for index of refraction is common. See [Edlen, 1965], [Slocum, 1992] and [Birch and Downs, 1994].
- Errors that exist separately in the workpiece and the measuring machine become combined in the measurement data. The same situation occurs when an artifact is used to qualify a measuring machine or machine tool. Separation of the errors from the measurement data requires two or more tests and appropriate manipulation of the data.

The remainder of this section presents particular separation techniques that are important to understand for design since they enable more accurate measurements and manufacturing processes that converge to more accurate geometry. Except for the last topic, *Volumetric Error Mapping*, a recent survey paper by [Evans, Hocken and Estler, 1996] is a good reference for the techniques presented here plus several others.

2.5.1 Reversal Techniques

The idea of using reversal probably dates back to ancient Egypt and the pyramids, yet several reversal techniques have more recent origins and others wait to be invented. They have in common at least two separate comparative tests where one component is physically reversed (actually rotated) to change the sign of its error contribution to the measurement. Typically, the comparison is between a workpiece or artifact and a machine element. The term *workpiece* implies that it is the object to be measured while *artifact* implies that the machine is to be calibrated. This distinction is rather pointless since reversal gives both. Then the artifact may be used in many other positions to further calibrate the machine without having to do subsequent reversals. The instrument used to make the comparative measurement is often termed a probe or indicator. It must have sufficient resolution and be

traceable to a standard if the objective is quantitative measurement. Often in manufacturing, the objective is to drive the measurement to zero so that traceability is less of an issue. [Evans, Hocken and Estler, 1996] call this a null test rather than true reversal.

Symmetry is the key ingredient in all reversal techniques. The generator of the symmetry, typically an axis of rotation, effectively becomes a perfect artifact from which the measurement is referenced. Ideally the reversal of the workpiece and the instrument should be perfectly symmetric about the generator. Careful setup is important to minimize misalignments between tests, but typically the directions are insensitive so that the errors are second order. The sensitive direction errors are far more critical to the accuracy of the measurement. Reversal depends upon the machine and the instrument being repeatable and the workpiece remaining constant in shape over the same path of measurement. Usually the tests are arranged so that gravity deflection is in an insensitive direction, otherwise the effect must be reduced by better support or subtracted out by modeling the effect. These points will become clearer as particular reversal techniques are presented.

2.5.1.1 Straightedge Reversal

Straightedge reversal requires two comparative tests between a straightedge and a machine slideway. The straightedge and the machine have initially unknown straightness errors $S(x)$ and $M(x)$, respectively, both shown magnified in Figure 2-2. Notice that the reversed straightedge in the second test is shown in phantom lines. The indicator for the two tests measures linear combinations of the straightness errors as specified in Equation 2.3. This choice sets the sign convention for straightness according to the indicator reading rather than a particular coordinate system. To the extent that the machine slideway is repeatable and the straightedge is symmetric between the first and second tests, then reversal may be considered as one test between a straightedge and its rotated image, or between the machine slideway and the axis of symmetry. This point of view makes Equation 2.4 rather obvious.

$$I_1(x) = S(x) + M(x) \qquad I_2(x) = S(x) - M(x) \tag{2.3}$$

$$S(x) = \frac{1}{2}[I_1(x) + I_2(x)] \qquad M(x) = \frac{1}{2}[I_1(x) - I_2(x)] \tag{2.4}$$

The basics of straightedge reversal is just that simple, but a few subtle points are worth describing. There will inevitably be a setup error in each test that requires subtracting out best-fit lines from $I_1(x)$ and $I_2(x)$ or from $S(x)$ and $M(x)$ since they are linear combinations. The preferred indicator is a capacitance gauge rather than a contacting probe to greatly reduce sensitivity to dirt, surface roughness and differences between traces caused by slight misalignments. The straightedge should be supported to minimize gravity sag, typically close to the Airy points spaced 0.577 times the length of the straightedge. If gravity acts in the sensitive direction, then additional precautions may be required as described by [Estler, 1985]. Furthermore, the straightedge should be supported to

minimize twist along its axis, and the trace for each test should lie in the centroidal plane to reduce sensitivity to twist and Poisson effects from bending strain. Drift and hysteresis effects will be detected by tracing back over the same path. Usually the average is used to help reduce these error contributions in the straightness measurement.

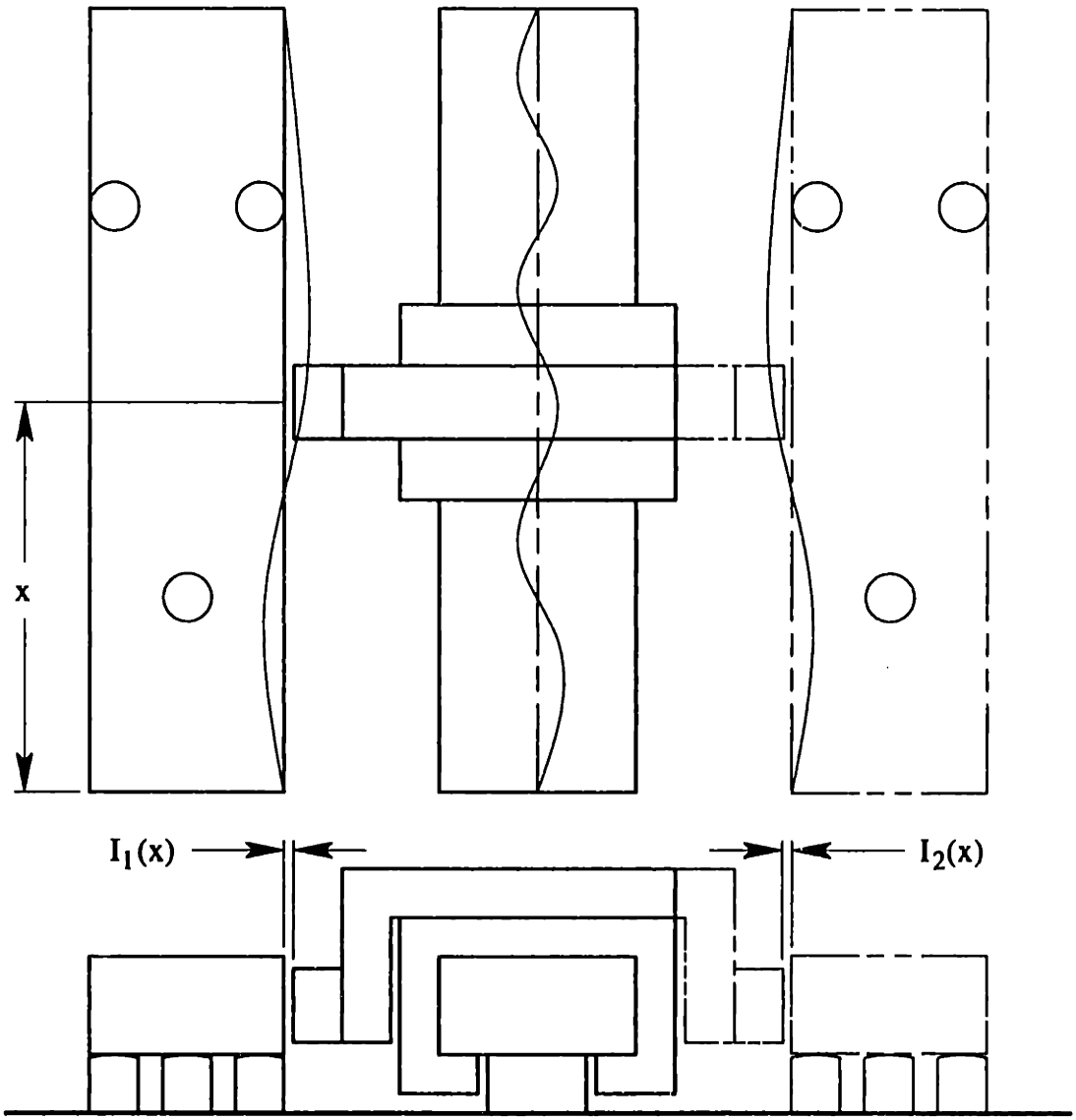


Figure 2-2 Top and end views show the basics of straightedge reversal. The first test of the straightedge occurs for the left-hand position (solid lines) with the gap measurement $I_1(x)$. For the second test, the straightedge is rotated about the center line to the right-hand position with the gap measurement $I_2(x)$.

2.5.1.2 Three-Flat Test

The three-flat test is a variant of straightedge reversal used commonly in the optics industry to overcome a limitation of Fizeau interferometry. This type of instrument measures the change in gap between two surfaces, which is ideal for straightedge reversal except that reversing one flat means that the second test would occur through one of the flats. This introduces a nonrepeatability between the two tests since light refracts in an unknown way

through the optical flat. This problem is overcome by using a third flat. The procedure will first be described in a way most analogous to straightedge reversal followed by the more common description. Figure 2-3 shows the arrangement of three tests involving three flats A, B and C. The reversal of flat B is apparent between tests 1 and 2. The problem is that flat B is compared against different flats A and C. This requires a third test to characterize the difference between flats A and C. Subtracting this difference from the sum of tests 1 and 2 leaves a comparison of flat B against its rotated image. The middle line of Equation 2.5 expresses this result, which applies only to the plane containing the generator of the symmetry. In other words, the three-flat test gives only straightness along the axis of rotation and not flatness over the surface as the name suggests. Simple algebraic manipulation of the same data gives the straightness for flats A and C.

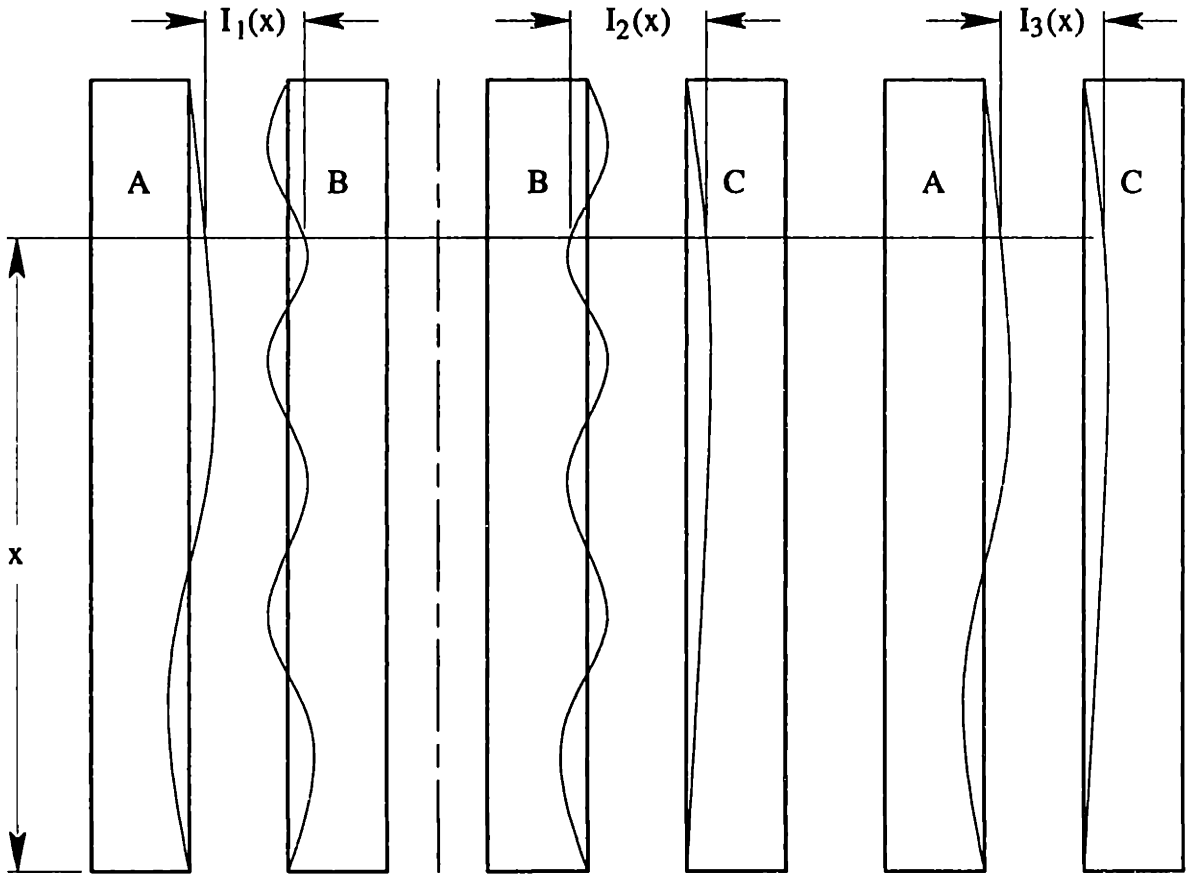


Figure 2-3 Flat B undergoes reversal and comparison to flats A and C in the first two tests. The third test characterizes the difference between flats A and C.

$$S_A(x) = \frac{1}{2}[I_1(x) - I_2(x) + I_3(x)]$$

$$S_B(x) = \frac{1}{2}[I_1(x) + I_2(x) - I_3(x)] \qquad (2.5)$$

$$S_C(x) = \frac{1}{2}[-I_1(x) + I_2(x) + I_3(x)]$$

When implemented on a Fizeau interferometer, flats A and B would be separate reference flats and C would be the flat under test (or A would be under test and B and C would be reference flats). Then the straightness for flat C would be the average of tests 2 and 3 minus the average straightness for reference flats A and B (contained in test 1). The third line of Equation 2.5 expresses this result. By repeating test 1 with flat B rotated an angle θ in plane, a new generator of symmetry is formed at $\theta/2$. This one test adds a new slice of straightness information to the surface data already collected. Continuing this process in multi-step fashion pieces together considerable information about absolute flatness for the three flats.

2.5.1.3 Square Reversal

Nonsquareness between linear axes of a machine is usually the single largest geometric error source, so a reversal test for squareness is obviously very important. Recall in straightedge reversal that the best-fit line is removed from the measurement. In square reversal, the slopes of best-fit lines to the sides of the square are of interest. The typical description of square reversal calls for just one reversal, for example, from position 1 to position 2 in Figure 2-4.[1] Notice that the vertical axis of the machine operates over the same range for positions 1 and 2 but the horizontal axis operates over two different ranges. Usually an assumption is made that side A is trammed perfectly to the horizontal axis at each position. However, if the axis has a bow, then the slope at position 1 will differ from the slope at position 2, which introduces an error in the angular measurement of the square. It still provides a reasonable measurement of the squareness between axes. The four-position squareness test proposed in Figure 2-4 is considerably more involved, but it will uncover straightness of both axes and provide an accurate angular measurement of the square. It introduces the technique called closure, which uses an obvious fact that angular subdivisions of one revolution must add to 360°.

Imbedded within the four-position squareness test are straightedge reversal tests for sides A and B. There is enough information to obtain two straightness estimates for each side as expressed in Equation 2.6. Notice that it includes the subtraction of best-fit lines, where a is the slope and b is the intercept for the corresponding indicator data. In Equation 2.7, this slope information is necessary to calculate the angle α between sides of the square and the angle β between the axes (as measured in the first or third quadrant). The angles are expressed in radians and the slopes are assumed small enough not to require converting with the arc tangent function. This is the main result from squareness reversal. As mentioned however, there is additional information about the straightness over the full extent of the tests. Equations 2.8 and 2.9 present the machine straightness for the x and y axes, respectively.

---

[1] See, for example, [Evans, Hocken and Estler, 1996].

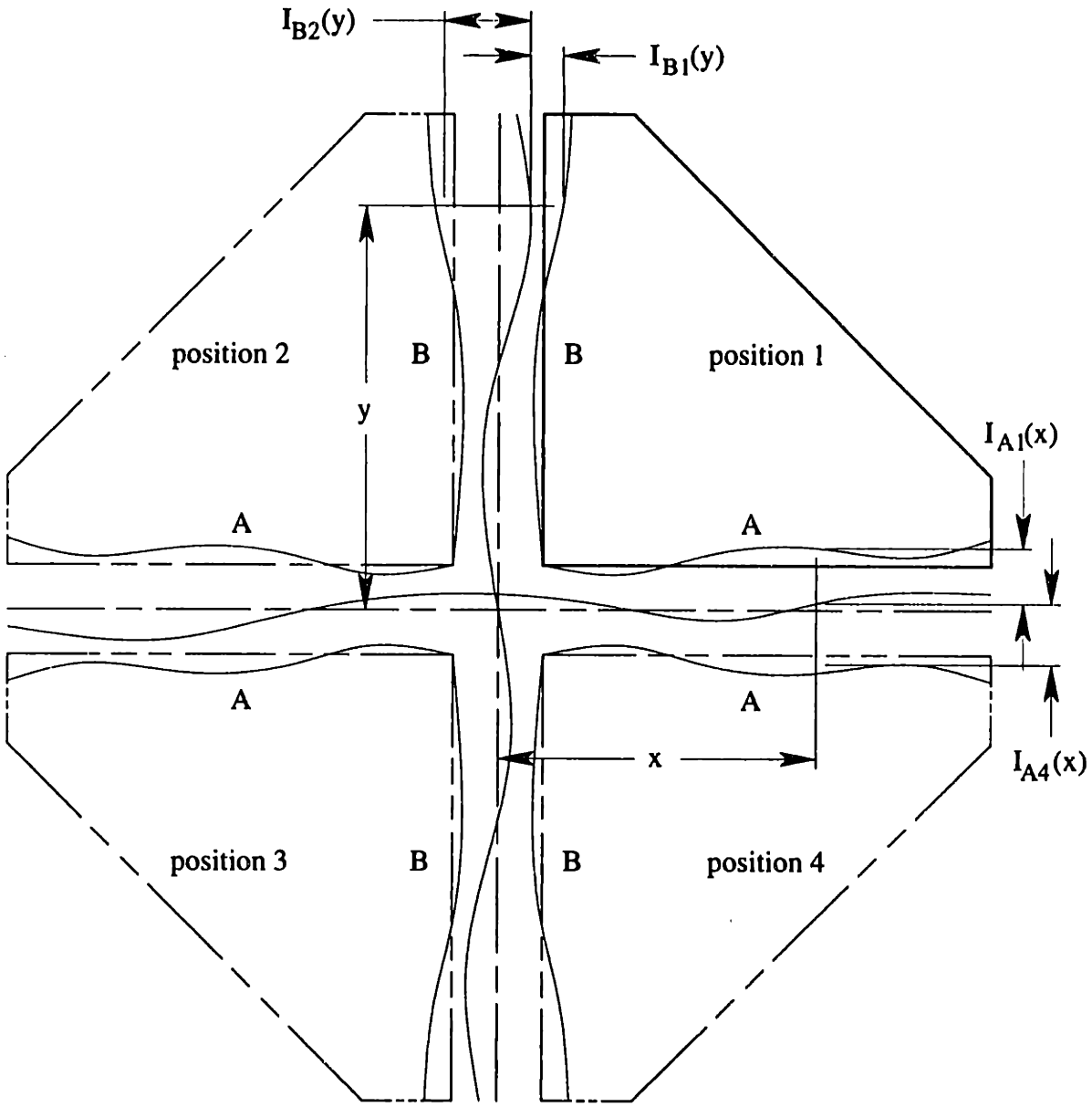


Figure 2-4 The four-position squareness test provides measurements of straightness and squareness for both the square and the machine axes.

$$
\begin{aligned}
2S_A(x) &= I_{A1}(x) + I_{A4}(x) - (a_{A1} + a_{A4})x - (b_{A1} + b_{A4}) \\
&= I_{A2}(-x) + I_{A3}(-x) - (a_{A2} + a_{A3})(-x) - (b_{A2} + b_{A3}) \\
2S_B(y) &= I_{B1}(y) + I_{B2}(y) - (a_{B1} + a_{B2})y - (b_{B1} + b_{B2}) \\
&= I_{B3}(-y) + I_{B4}(-y) - (a_{B3} + a_{B4})(-y) - (b_{B3} + b_{B4})
\end{aligned}
\tag{2.6}
$$

$$
\begin{aligned}
\alpha &= \frac{\pi}{2} - \frac{1}{4}(a_{A1} + a_{B1} + a_{B2} - a_{A2} - a_{A3} - a_{B3} - a_{B4} + a_{A4}) \\
\beta &= \frac{\pi}{2} + \frac{1}{4}(a_{A1} + a_{B1} - a_{B2} + a_{A2} - a_{A3} - a_{B3} + a_{B4} - a_{A4})
\end{aligned}
\tag{2.7}
$$

$$
\begin{aligned}
M(x)\big|_{x>0} &= \frac{1}{2}\left\{I_{A1}(x) - I_{A4}(x) - (a_{A1} - a_{A4})x - (b_{A1} - b_{A4})\right\} \\
&\quad + \frac{1}{4}(a_{A1} + a_{B1} + a_{B2} - a_{A2} + a_{A3} + a_{B3} + a_{B4} - a_{A4})x \\
M(x)\big|_{x<0} &= \frac{1}{2}\left\{I_{A2}(x) - I_{A3}(x) + (a_{A2} - a_{A3})x - (b_{A2} - b_{A3})\right\} \\
&\quad - \frac{1}{4}(a_{A1} + a_{B1} + a_{B2} - a_{A2} + a_{A3} + a_{B3} + a_{B4} - a_{A4})x
\end{aligned}
\tag{2.8}
$$

$$
\begin{aligned}
M(y)\big|_{y>0} &= \frac{1}{2}\left\{I_{B1}(y) - I_{B2}(y) - (a_{B1} - a_{B2})y - (b_{B1} - b_{B2})\right\} \\
&\quad + \frac{1}{4}(a_{A1} + a_{B1} - a_{B2} + a_{A2} + a_{A3} + a_{B3} - a_{B4} + a_{A4})y \\
M(y)\big|_{y<0} &= \frac{1}{2}\left\{I_{B4}(y) - I_{B3}(y) - (a_{B4} - a_{B3})y - (b_{B4} - b_{B3})\right\} \\
&\quad - \frac{1}{4}(a_{A1} + a_{B1} - a_{B2} + a_{A2} + a_{A3} + a_{B3} - a_{B4} + a_{A4})y
\end{aligned}
\tag{2.9}
$$

With so many slope terms used in different equations with different signs, a consistent sign convention is essential. Positive indicator readings correspond to the dimensions shown in Figure 2-4. Slopes are positive if the indicator reading increases with increasing x or y position. The easiest verification is to check the effect on α since it is obvious at each position. The effect on β is opposite from α for positions 1 and 3 and the same for positions 2 and 4.

2.5.1.4 Donaldson Ball Reversal

Ball reversal, attributed to [Donaldson, 1972], separates radial error motion of a spindle from the roundness of a test ball or other round artifact. Ball reversal is similar to straightedge reversal except that, as Figure 2-5 shows, the straightedge is curved around the spindle axis. We could think of this as straightness in polar coordinates. The axis of symmetry is nominally the same as the spindle axis. The reversal occurs by relocating the artifact and the indicator 180° with respect to the spindle, although it is often practical to set up two indicators rather than reversing one. The equations are the same as before except that Equations 2.10 and 2.11 use P for profile rather than S to avoid potential confusion with spindle. A few subtle points are worth reviewing. Like straightedge reversal, a setup error will exist between the artifact and the spindle. This can be removed to first order by subtracting out the mean and best-fit sin(θ) and cos(θ) terms from the indicator data.[1] To avoid problems with asynchronous spindle motion (particularly common with rolling

[1] The setup error could instead be removed from the separated data P and M but it is useful to know the setup error for each test.

element bearings), both tests should occur over the same spindle revolution. Once the profile of the artifact is known, it may be used to investigate asynchroneity and/or spindle motion in another plane.

$$I_1(\theta) = P(\theta) + M(\theta) \qquad I_2(\theta) = P(\theta) - M(\theta) \tag{2.10}$$

$$P(\theta) = \frac{1}{2}[I_1(\theta) + I_2(\theta)] \qquad M(\theta) = \frac{1}{2}[I_1(\theta) - I_2(\theta)] \tag{2.11}$$

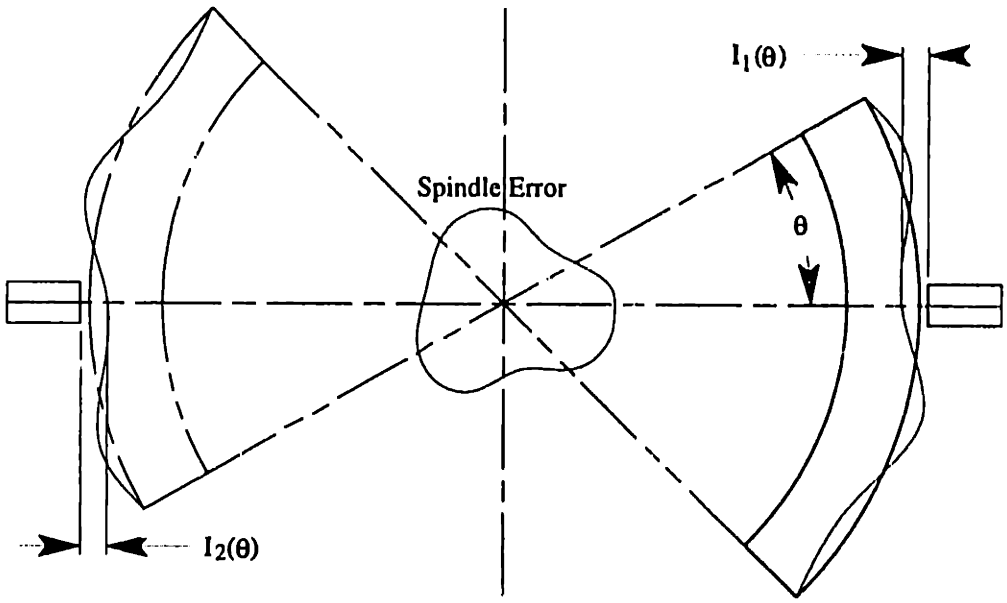


Figure 2-5 Ball reversal may be considered as one test between a ball and its rotated image, or between the spindle and the axis of symmetry.

Ball reversal may be extended to provide additional information about the angular motion of a spindle and its alignment to the machine. Figure 2-6 shows the first setup for a test using a cylindrical square. The indicator translates with the z-axis of the machine to allow two circular traces and one linear trace. Figure 2-7 shows the second setup where the indicator is repositioned to the opposite side and the spindle and square have rotated 180°. By taking a second linear trace, Equation 2.11 may be used to obtain the profile of the square (combined taper and nonparallelism to the spindle axis) and the motion of the z-axis with respect to the spindle axis. The slope of the data gives the nonparallelism between the spindle axis and the z-axis. Figure 2-8 shows the third setup where the cylindrical square is rotated 180° with respect to the spindle. This allows ball reversal at two axial locations and subsequent reduction to spindle angular motion.

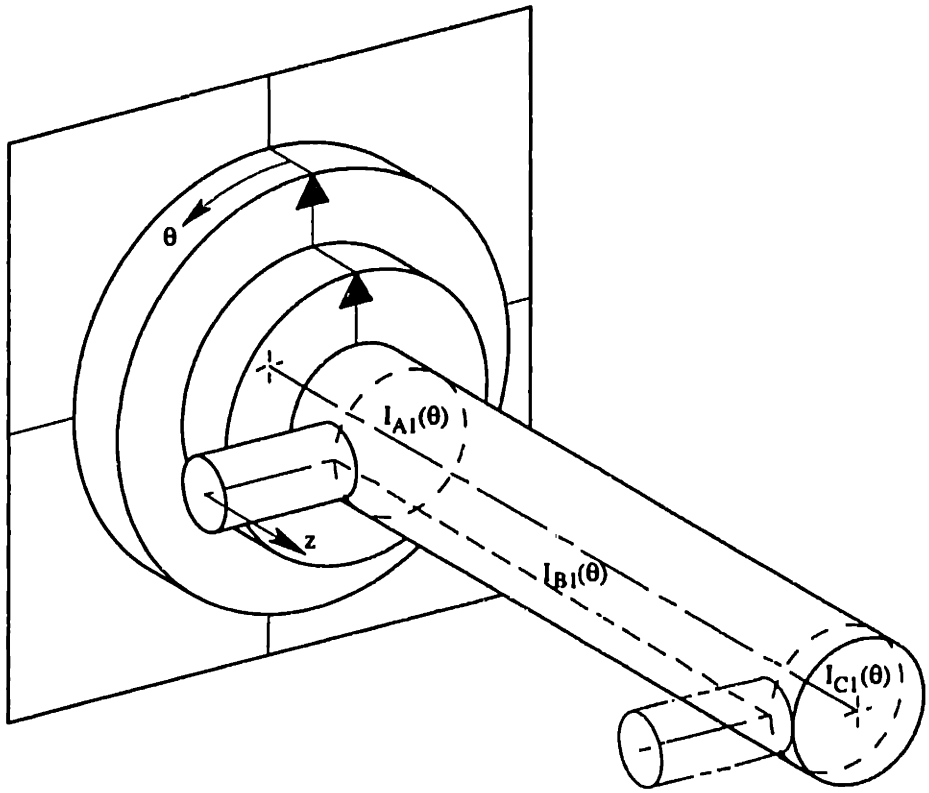


Figure 2-6 Two circular traces and one linear trace are recorded in the first setup with the cylindrical square. This is the first test for ball reversal at two axial locations and also straightedge reversal.

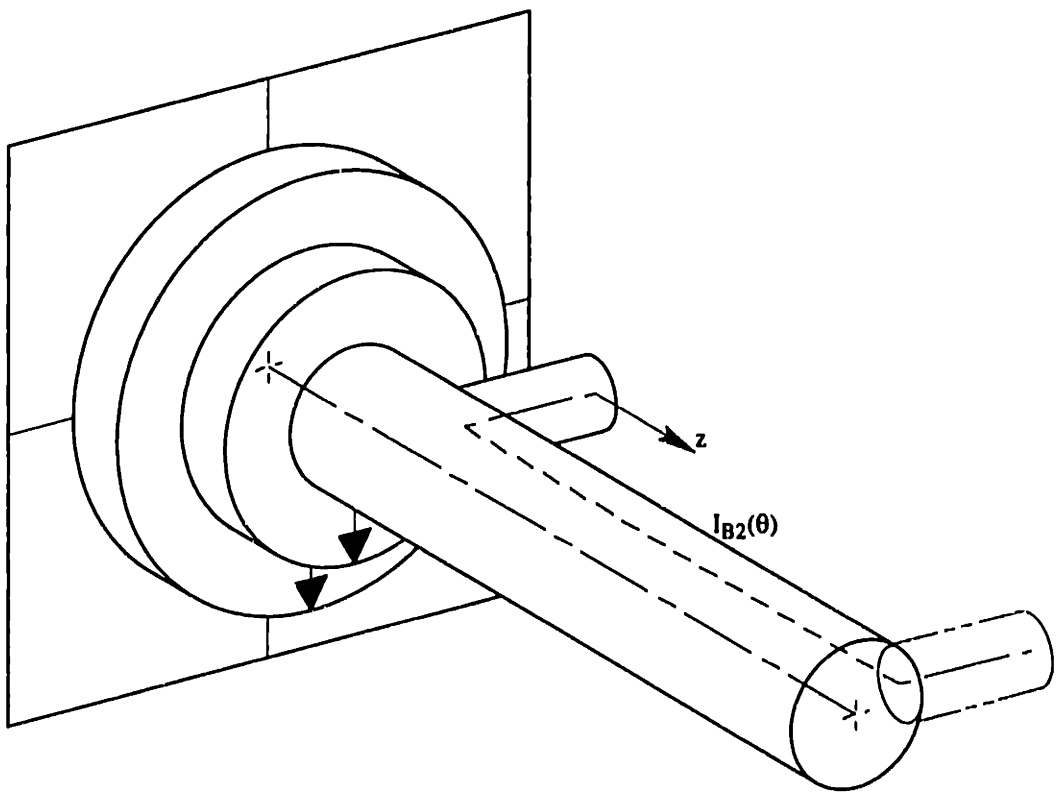


Figure 2-7 The second setup is for straightedge reversal. The spindle and cylindrical square are rotated 180°, and the indicator is relocated to record the second linear trace.

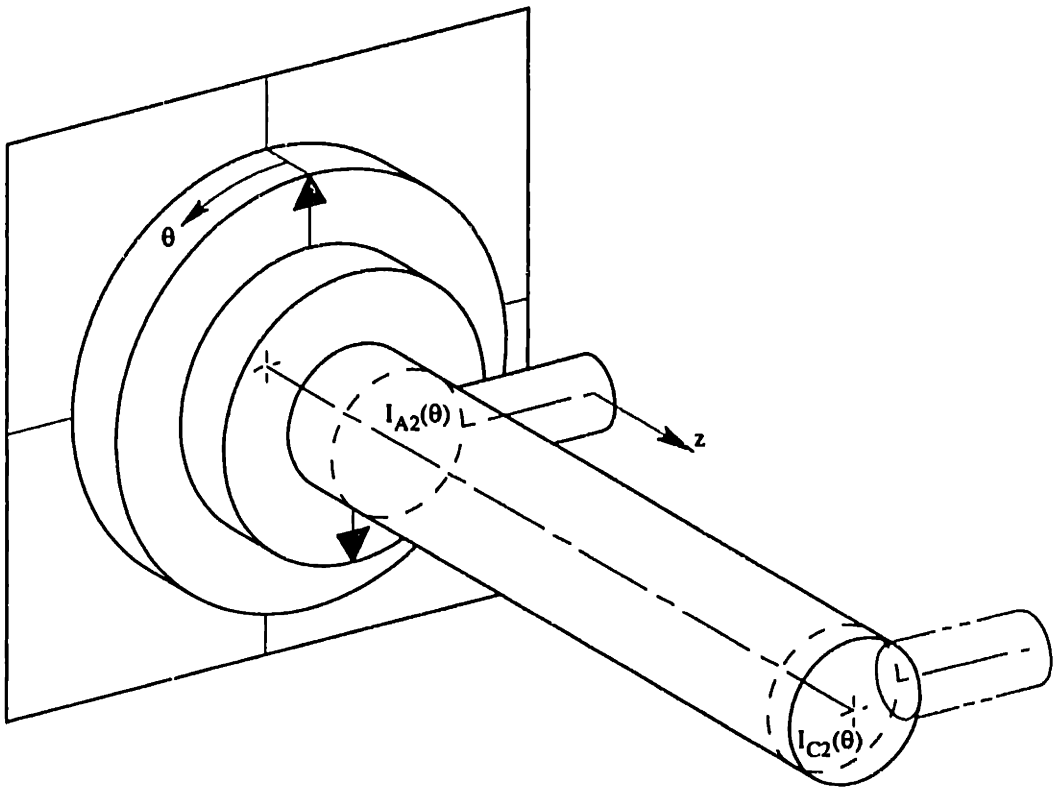


Figure 2-8 The third setup is for ball reversal. The cylindrical square is rotated 180° with respect to the spindle, and the indicator is already in position to record the second set of circular traces.

A difficulty with any reversal is ensuring that the action of physically reversing the artifact does not change the shape of the artifact or the repeatability of the machine. The three-vee kinematic coupling is a device well suited for ensuring repeatability between the artifact and the spindle. Although it can be made with six vees to allow 180° repositioning, the possibility of ball reversal using three 120° positions is appealing. It requires relocating the artifact and the indicator at 120° intervals with respect to the spindle. An additional benefit occurs because the indicators at 120° and 240° sense spindle motion orthogonal to the indicator at 0° or 180°. Equation 2.12 shows the errors that the indicators sense for the three tests. This matrix equation is easily solved for the profile of the artifact and the x and y spindle motion given in Equation 2.13.

$$\begin{bmatrix} I_1(\theta) \\ I_2(\theta) \\ I_3(\theta) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 2 & 2 & 0 \\ 2 & -1 & \sqrt{3} \\ 2 & -1 & -\sqrt{3} \end{bmatrix} \cdot \begin{bmatrix} P(\theta) \\ M_x(\theta) \\ M_y(\theta) \end{bmatrix} \tag{2.12}$$

$$\begin{bmatrix} P(\theta) \\ M_x(\theta) \\ M_y(\theta) \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 2 & -1 & -1 \\ 0 & \sqrt{3} & -\sqrt{3} \end{bmatrix} \cdot \begin{bmatrix} I_1(\theta) \\ I_2(\theta) \\ I_3(\theta) \end{bmatrix} \tag{2.13}$$

2.5.1.5 Estler Face-Motion Reversal

Face-motion reversal, attributed to Estler, separates combined axial and angular error motion of a spindle from the circular flatness of an artifact.[1] This separation is possible only if the axial error motion of the spindle is measured independently and subtracted from the indicator making the circular trace. Figure 2-9 shows the two tests required for face motion reversal. The indicator located on axis is sensitive only to axial spindle motion to first order. It allows the effect of axial spindle motion to be subtracted as described in Equation 2.14. These two equivalent indicators are then processes exactly the same as in ball reversal, including the removal of sin(θ) and cos(θ) terms from the indicator data. The result is the profile of the artifact and the axial spindle motion at the particular radius relative to the center. Subsequent reduction to angular spindle motion is obvious. Face-motion reversal extends to three tests at 120° rotations exactly the same as in ball reversal.

$$I_1(\theta) = I_{B1}(\theta) - I_{A1}(\theta) \qquad I_2(\theta) = I_{B2}(\theta) - I_{A2}(\theta) \tag{2.14}$$

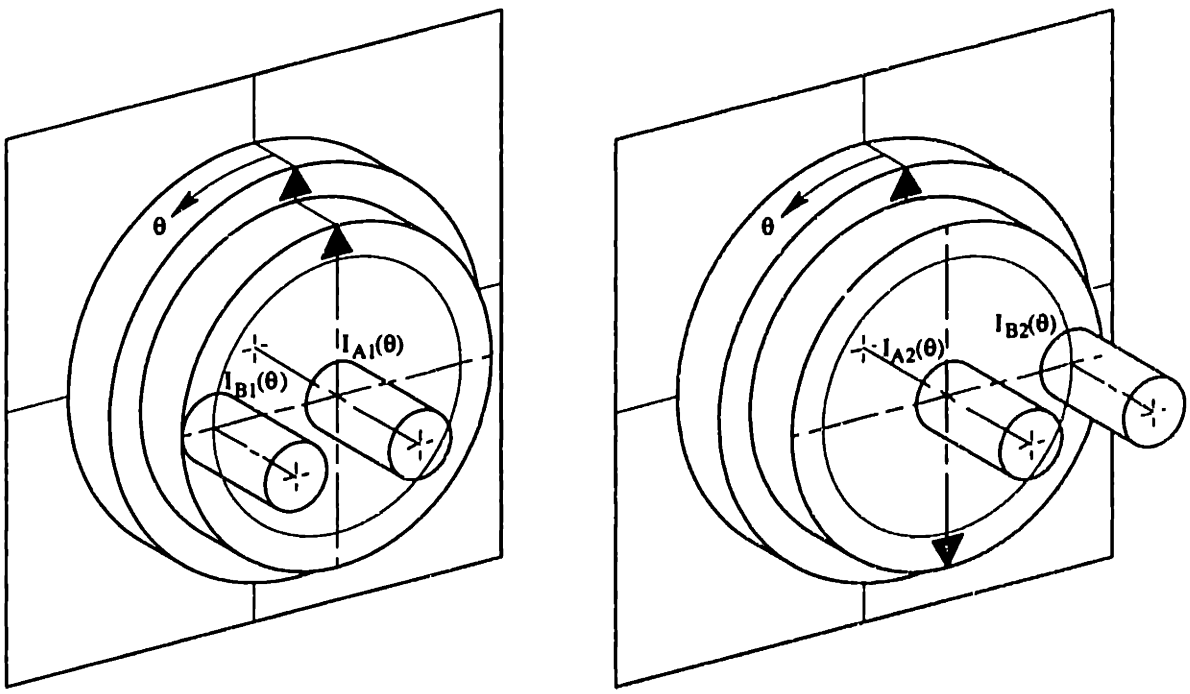


Figure 2-9 The generator of symmetry for face motion, like ball reversal, is an axis nominally at the spindle axis. However, the generator must be constrained axially by measuring its motion with the center indicator and subtracting the effect from the circular-trace data.

2.5.2 Multi-Step Averaging

Multi-step averaging is a separation technique that works for all spatial frequencies except integer multiples (or harmonics) of N, where N is the number of equiangular measurement

---

[1] [Evans, Hocken and Estler, 1996] state that face-motion reversal was first derived and described by W. Tyler Estler at the 1986 SME Precision Machining Workshop in Cambridge, MA.

steps. The technique may be applied to the inspection of roundness [Spragg and Whitehouse, 1967-68], circular flatness, or surface figure consisting of nonrotationally symmetric terms [Evans and Kestner, 1996]. Since reversal techniques are available for roundness and circular flatness, multi-step averaging is most valuable for interferometric testing of optical surface figure. However, a simpler example is more informative especially when contrasted against reversal. We will consider circular flatness since it does not appear in the references.

A conceivable application for multi-step averaging is in testing the ability of a diamond turning machine to produce rotationally symmetric (or circularly flat) optics. The spindle is the primarily limitation unless the axes are capable of tracking spindle motion. Suppose that a facing cut is performed on a test part followed by inspection with a tool-mounted indicator. To the extent that the spindle error motion is completely synchronous, circular traces of the part will appear perfect. The negative of the spindle's circular flatness has been recorded into the part's surface. The combined errors will become visible to the indicator only if they are moved out of (anti) phase by rotating the part relative to the spindle. The reversal technique requires only one 180° rotation, but the indicator must be moved to the far side of the part, perhaps requiring an extension of the tool holder. In addition, the axial spindle motion must be measured independently. Multi-step averaging avoids these two difficulties at the price of additional rotations of the part. The main technical problem is that harmonics of N are invisible to the indicator (because the part was faced) just like all frequencies were for the first inspection. If the part had not been faced on this spindle, then the indicator would sense harmonics of N but they would be indistinquishable between spindle motion or part profile.

The data processing for multi-step averaging is straightforward but no more so than for reversal techniques. The N steps of data are averaged to remove the part profile (less the harmonics of N) thus leaving the spindle motion. To get the part profile, each data set is rotated to the tested orientation of the part then averaged to remove the spindle motion. Both estimates contain the unseparated harmonics of N.

2.5.3 Closure and Subdivision

Closure is a simple idea that provides an elegant solution for extending the dynamic range of metrology instruments. Dynamic range is the physical range of the instrument divided by the resolution. Dynamic resolution is the inverse of dynamic range. Consider the historical problem of dividing a circle into 360°, the degree into 60 arc minutes or the meter bar into 1000 millimeters. The process begins with a unit of measure and an instrument that can detect differences in subdivisions with sufficient resolution but limited range. Presumably the unit is already subdivided at least crudely so that measurements can be made between all adjacent subdivisions. The instrument must be linear and have a known gauge factor unless it is solely used to achieve a null condition. Closure is an observation that the sum of all the

subdivisions must equal the unit being subdivided, obviously. The average reading for all the measurements is the ideal size of a subdivision.

Figure 2-10 shows an example of a step gauge being measured with a dial indicator. This instrument provides only a relative measurement; how different is one step from another. Equation 2.15 demonstrates the simple algebra involved to recover the true size for each step relative to the known length L and the calibration of the indicator's gauge factor. Once all steps are measured (the δ's plus an unknown offset X), then the sum is set equal to the known length L to recover X. This process may also be used to null the indicator to the ideal step size.

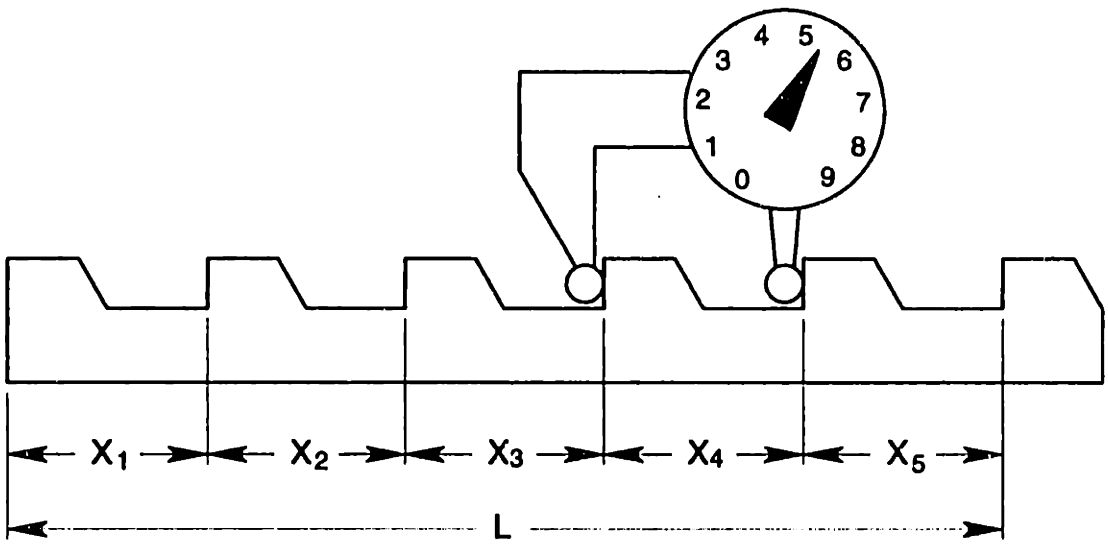


Figure 2-10 A step gauge with known length L but initially unknown step lengths. Closure enables the step lengths to be determined and the indicator to be nulled to the ideal step size.

$$\left.\begin{aligned} X_1 &= X+\delta_1 \\ X_2 &= X+\delta_2 \\ X_3 &= X+\delta_3 \\ X_4 &= X+\delta_4 \\ X_5 &= X+\delta_5 \end{aligned}\right\} \qquad L=\sum_{i=1}^{n} X_i=\sum_{i=1}^{n}(X+\delta_i) \qquad X=\frac{1}{n}\left(L-\sum_{i=1}^{n}\delta_i\right) \qquad (2.15)$$

Figure 2-11 shows a second way to measure the step gauge by comparing it with an exact copy. The copy may be realized with a repeatable measuring machine. As shown, the step gauge is measured before and after a physical translation X that is nominally the same as one step. The differences in measurements are use in Equation 2.15 to recover the true step sizes relative to the known length L and the calibration of the measuring machine over very small displacements. This basic idea extends to two- and three-dimensional self calibrations using grid-type artifacts. Self calibration using a 3D grid is cumbersome compared to volumetric techniques described later. Self calibration using a 2D grid is fairly

common and is described in the next section. The algebra becomes more complicated requiring coordinate transformations and a solution of equations.

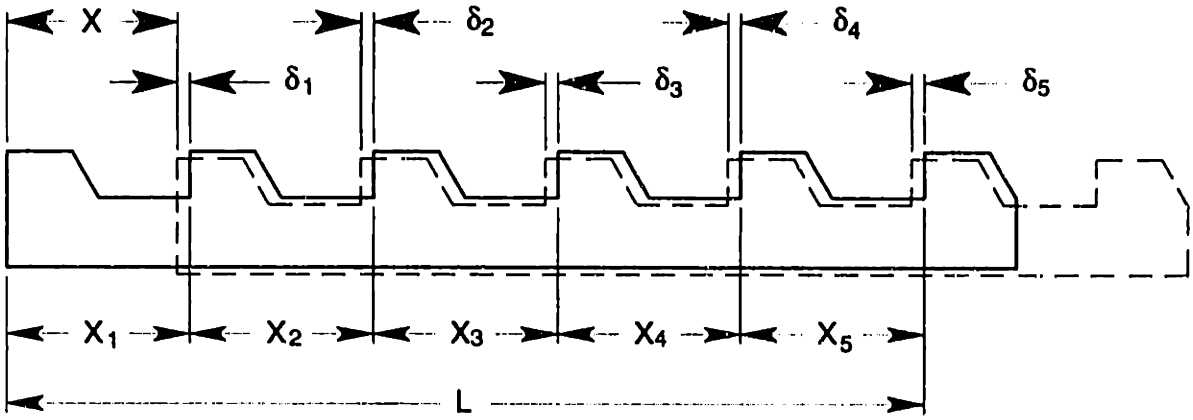


Figure 2-11 The comparison of a step gauge to its translated self provides the difference measurements required for closure.

2.5.4 Self Calibration of 2-D Artifacts

The separation techniques presented thus far have been one-dimensional in nature, although they extend readily to higher dimensions through repetition of 1-D tests. It may be more convenient, however, to calibrate a two-axis stage with a 2-D artifact such as a ball plate (a square array of balls mounted to a plate). Unfortunately most 2-D artifacts are accessible from only one side, which prevents the flip required for straightedge and squareness reversals.[1] The basic idea behind reversal and closure still applies: compare the artifact to itself after having been moved to symmetrical positions. In the literature, this type of separation is called self calibration [Raugh, 1984, 1985, 1997], [Takac, et al., 1996], [Ye, et al., 1997]. Although there are several approaches, we will focus on exact calibration of the artifact using the measuring machine as a repeatable comparator. This avoids (for now) the need to fit functions to the measuring machine error, but it requires an independent measurement of scale.

The references state that the artifact must be measured in at least three positions with the separations being either two rotations about different centers, or one rotation and one translation. We can show the need for one rotation by considering another example of closure. Figure 2-12 shows a nominally square plate compared to an exact copy rotated approximately 90°. The sides are straight but the angles are not quite square. A measuring machine is capable of measuring differences in angles very accurately even if its axes are out of square. These differences are expressed very clearly in the first four rows of Equation 2.16. The last row states the closure relationship, and the matrix would be singular otherwise. Closure and any three of the four difference measurements are

[1] A grid of thru holes is easy to probe from either side of the plate making reversal techniques practical. In addition, the measurements occur on the neutral axis of the plate to minimize errors due to bending.

sufficient to determine the four angles of the plate. However, the redundancy of the fourth measurement reduces the sensitivity to measurement errors. The pseudo inverse in Equation 2.17 solves the least-squares problem. This example demonstrates that absolute squareness can be determined without square reversal by rotating a four-sided square 90°.

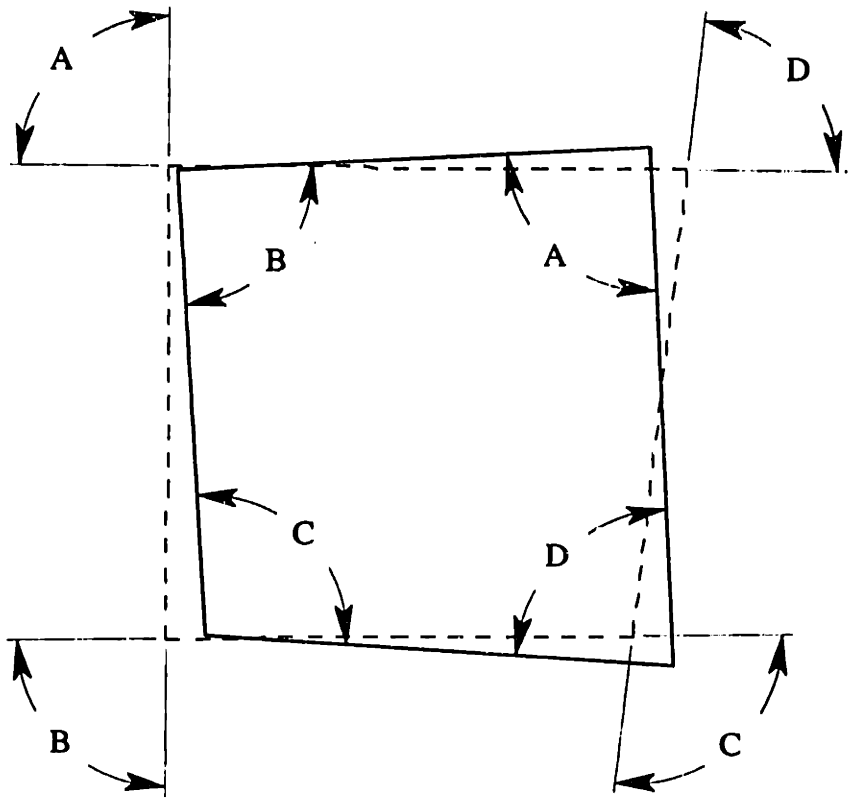


Figure 2-12 Closure applied to four-side square will determine the absolute angles between the sides.

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} = \begin{bmatrix} \alpha_{(A-B)} \\ \alpha_{(B-C)} \\ \alpha_{(C-D)} \\ \alpha_{(D-A)} \\ 2\pi \end{bmatrix} \tag{2.16}$$

$$\begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} = \frac{1}{8} \begin{bmatrix} 3 & 1 & -1 & -3 & 2 \\ -3 & 3 & 1 & -1 & 2 \\ -1 & -3 & 3 & 1 & 2 \\ 1 & -1 & -3 & 3 & 2 \end{bmatrix} \cdot \begin{bmatrix} \alpha_{(A-B)} \\ \alpha_{(B-C)} \\ \alpha_{(C-D)} \\ \alpha_{(D-A)} \\ 2\pi \end{bmatrix} \tag{2.17}$$

Figure 2-13 shows two cases of nonstraight sides that are visible to a comparator after a 90° rotation. Any deviation that is not four-fold symmetric (or multiples of) will be visible after a 90° rotation. The four-fold symmetric deviations in Figure 2-14 are exposed by a translation, but the spatial frequency corresponding to the translation and its multiples

are invisible. This is not a problem with grids since the spatial frequency is limited by the grid spacing. However, the sensitivity to certain modes (we will call them) may be low, depending on the positions measured. These modes will be more subject to measurement noise affecting the fidelity of the separation.

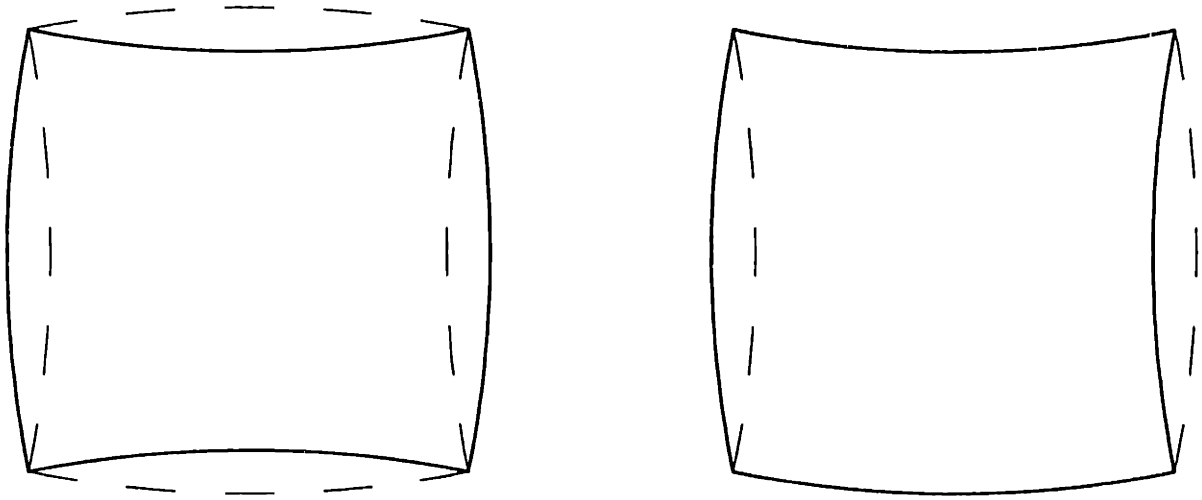


Figure 2-13 A 90° rotation exposes deviations from four-fold symmetry and multiples thereof.

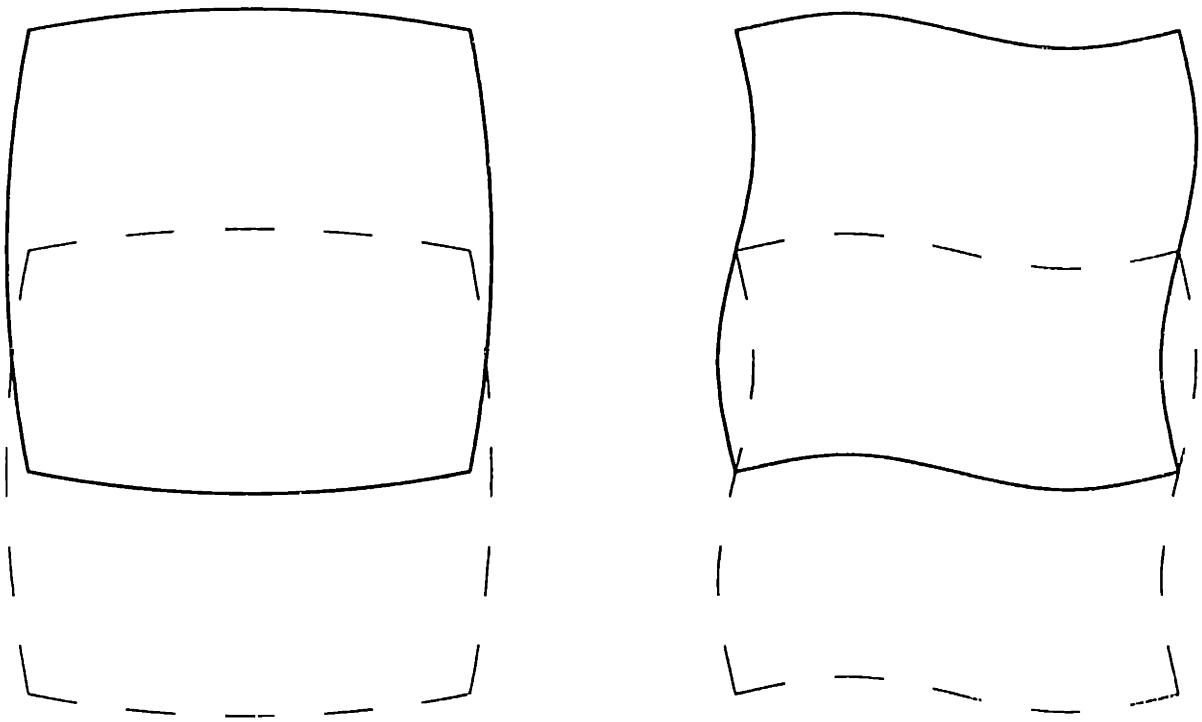


Figure 2-14 A translation exposes four-fold symmetric deviations not visible to a 90° rotation.

Figure 2-15 shows a typical 2-D artifact consisting of small circles overlaid on a nominal grid that represents the measuring machine. The four views show the artifact positioned differently with respect to the grid. The filled circles indicate the locations where

differences will be measured between the first setup and subsequent setups.[1] The symmetry of the artifact allows it to be repositioned on the measuring machine so that difference measurements will be much smaller than the spacing of features. Presumably the inaccuracy of the measuring machine over these small distances is negligible and both the artifact and the measuring machine are stable over the duration of the tests. Then the measured differences are attributable to the artifact and the rigid-body transformations between setups. In the manner of Equation 2.16, a system of 56 equations is set up for the 28 sets of X and Y difference measurements. There are nine unknowns associated with three planar transformations and 20 unknowns associated with x and y locations for 10 of the 12 circles. The x and y locations for 2 circles are specified to set the absolute length scale and to constrain the coordinate system. The size of the matrix prevents it from being shown here but the basics of setting up the equations are worth discussing further.

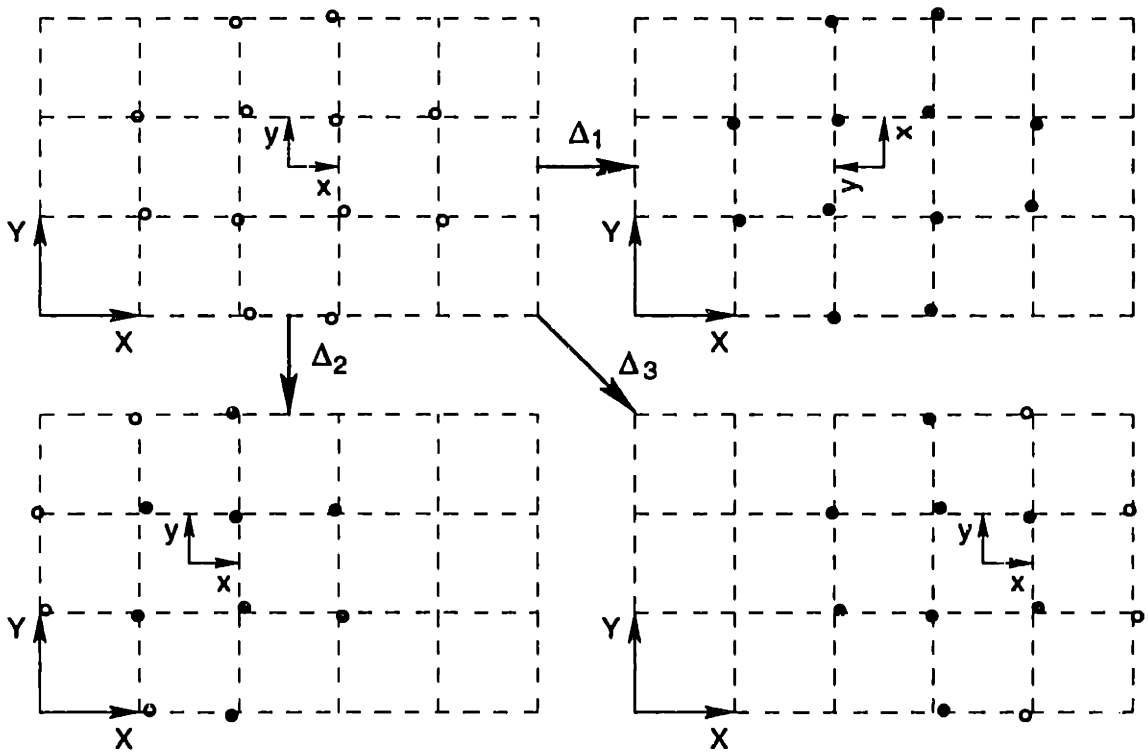


Figure 2-15 An artifact consisting of 12 circles is measured in four different positions with respect to a nominal grid. Differences in measurements are obtained for the filled circles and used to resolve the true shape of the artifact. An independent length measurement sets the scale of the artifact.

The first step is to mark all the feature locations on the artifact with numbers, 1 to 12 in this case. Make sure to duplicate the numbers in all setups accounting properly for the repositioning of the artifact. These numbers will identify the x and y variables in the equations. In addition, they can be used to identify the grid locations corresponding to the artifact's first setup. As an example, consider the first repositioning, a 90° rotation

[1] Difference measurements between other combinations of setups are also possible.

indicated by Δ_1. Suppose that feature 6 moves to the nominal position originally occupied by feature 1 at grid 1. Then Equation 2.18 relates the measured X-Y distance between the two features to the unknown variables. The notation is a little unusual; for example, $\Delta_1 X_1$ is one symbol representing the X difference after the first repositioning measured at grid 1. Similarly, the unknown variables for the rigid-body transformation are $\Delta_1 X$, $\Delta_1 Y$ and $\Delta_1 \theta$. Since the angle will be small, it is sufficient to use the nominal coordinate of feature 6, indicated by an overbar, to make the equations linear in the unknown parameters. The matrix that results from all 56 equations will be rather sparse with three 1's of appropriate sign and one nominal coordinate per row. The main difficulty will be getting the signs right. All that remains is computing the least-squares solution using the pseudo inverse.

$$\begin{aligned} \Delta_1 X_1 &= x_1 - \left\{-y_6 + \Delta_1 X - \Delta_1 \theta \cdot \bar{x}_6\right\} \\ \Delta_1 Y_1 &= y_1 - \left\{x_6 + \Delta_1 Y - \Delta_1 \theta \cdot \bar{y}_6\right\} \end{aligned} \tag{2.18}$$

2.5.5 Volumetric Error Mapping

Volumetric error mapping refers to any technique that measures and records the positional and/or angular errors of a machine moving throughout its working volume. The error map provides the information required to improve the accuracy of the machine either by physically correcting error sources or compensating their effects, for example, by changing the programmed path slightly. Usually the error map is intended to account for geometric errors inherent in the machine's manufacture and set up, but it might also include terms to account for short-term instability such as thermal growth. The error vector between the tool or probe and the workpiece usually has three to six components and the machine may have as many degrees of articulation. An error map can quickly become very complicated.

The traditional approach has been to measure the six-component error vector of each moving axis using a series of 1-D tests. This is the logical approach for physically correcting geometric errors in a serially arranged set of axes. Known as parametric error mapping, this approach provides a very dense data set along each axis of motion. This information is combined assuming rigid-body kinematics to provide an equally dense volumetric error map of the machine. It is a very efficient way to use the relatively modest amount of information. The disadvantages are the missing information about nonrigid-body coupling between axes and the number of specialized 1-D tests required for each axis. A number of researchers have described parametric error mapping, for example, [Domnez, et al., 1986], [Ehmann, Wu and DeVries, 1987] and [Kiridena and Ferreira, 1993].

A competing approach borrows the best idea of parametric error mapping, the efficient error model, and combines it with tests that measure the tool-to-work machine error rather than individual parametric errors. The tests vary among researchers but the basic characteristics are nearly opposite from parametric error mapping. The volumetric data is relatively sparse, nonrigid-body coupling is measured, and the tests are rather

simple. It is not clear whether there is an accepted name for this approach. [Sartori and Zhang, 1995] use the term *error function measurement* but then further classify this into the direct method (i.e., the parametric method) and the self-calibration method, which uses an artifact measured in a number of positions. This nomenclature does not fit well with the work of [Soons and Schellekens, 1992], [Soons, Theuws and Schellekens, 1992] whose method works with artifacts or direct measuring instruments. Our work at LLNL, [Krulewich, 1998], [Krulewich, Hale and Yordy, 1995], is very similar but we use the laser ball bar, a direct length measuring device. A term that we use and like is *functional error mapping*.

A full description of functional error mapping is well beyond the scope of this chapter, and the references provided are very good. However, the basic ideas are useful to know. The method begins with a kinematic model of the machine, either serial or parallel as the case may be. The model contains the same basic error terms that would be individually measured in the parametric method such as the straightness error as a function of the x-axis motion. In functional error mapping, those error terms are expressed as mathematical functions with coefficients that are later solved in one large least-squares problem. The functions are combined algebraically to yield a multi-variable error function with a relatively large number of fitting parameters. In parametric error mapping, the combination is strictly numerical and the error data may reside in look-up tables or in simple mathematical functions. The key distinction between the two methods is how the total error motion is separated into basic components. The parametric method requires a set of specialized tests that physically separates the total error into components. The functional method needs no physical separation although it provides an informational separation in the fitted parameters.

The type of mathematical function used to represent the error terms is rather arbitrary. Polynomials of order three to four are typical but other possibilities are sinusoids, piecewise polynomials and splines (see Appendix B.8). The model requires sufficient degrees of freedom to fit the measured errors accurately but not so much as to fit noise in the data. Currently, a definitive method does not exist for optimally choosing the order of the model. Krulewich has had some success using the Mallows C_p statistic to choose the best reduced-order model [Krulewich, Hale and Yordy, 1995], but often it comes down to judgment. The same problem exists for the parametric method but it is much easier to judge the fit to a fairly dense data set along one axis of motion.

2.6 Exact-Constraint Design

> By "dusting off" the principles of kinematics and applying them to machine design, we arrive at the method of *Exact Constraint*. The method of Exact Constraint has been developed to the point where it comprises a body of knowledge which can be used to routinely create new machine designs which are both high in performance and low in cost. The results are so excellent, yet so obvious; so elegant, yet so simple; that at once they seem

> both profound and trivial! Perhaps it is this duality which has kept these principles so well hidden. One may ask: "Of what value could anything so trivial be?" And so these principles have been overlooked. They have become disused.
>
> [Blanding, 1992]

The designers of mechanisms routinely use the principles of kinematics because overconstrained or underconstrained devices simply will not function. What the precision engineer must remember is that at some scale, everything is a mechanism. The component that must remain stable to nanometers will not if it is overconstrained to a structure that deforms by micrometers. This is often the most important motivation for exact-constraint or kinematic design in precision machines, that is, to isolate sensitive parts or systems such as a metrology frame from the influence of dimensionally changing supports and/or manufacturing tolerances.[I] Similarly, parts will fit together precisely and without backlash if they are exactly constrained, for example, kinematic couplings. This is why [Smith and Chetwynd, 1992] state that "a divergence from pure kinematic design results in increased manufacturing costs."[II]

The term exact constraint is very explicit and meaningful once the basic concept is understood. An unconstrained *rigid* object has six degrees of freedom usually identified as three translations and three rotations. A *nonrigid* object may have one or more degrees of flexibility that act as additional degrees of freedom, relatively speaking. For example, an open shoe box is torsionally flexible and so would have a total of seven degrees of freedom. The proper application of constraints would eliminate degrees of freedom in a one-to-one fashion. It is the objective of exact-constraint design to achieve some desired freedom of motion or perhaps no motion by applying the minimum number of constraints required. Often we conceptualize in terms of an *ideal* constraint, which is absolutely rigid against motion in one or more degrees of freedom and is absolutely free in the remaining degrees of freedom. A *real* constraint such as a small-area contact between surfaces, a link or a bearing, provides one or more degrees of constraint that are relatively much stiffer than the degrees of freedom and so approximates ideal behavior.[III]

The reference by Blanding is an excellent introduction to exact-constraint design and rigid structures. Because it is so basic and complete, it forms the basis for this

---

[I] The word kinematic is more often used, but exact constraint is more descriptive and thus is preferred.

[II] As a practical matter, purely kinematic designs are generally difficult to achieve; whereas, so called semi-kinematic designs (for example, Hertzian contact areas instead of frictionless point contacts) generally provide acceptable isolation characteristics, greater robustness and lower cost. The quotation might better read "a divergence from kinematic design theory may result in increased manufacturing costs." As an alternative, the process of replication produces a good fit and is relatively low in cost.

[III] In the case of a sliding bearing or small-area contact, the degree of freedom may be as stiff as the constraint until sliding occurs. Ideal behavior requires that the frictional force divided by the constraint stiffness be a negligible deflection.

introductory section. Several other references provide accounts of more specific exact-constraint designs such as [Slocum, 1992], [Smith and Chetwynd, 1992], and [Furse, 1981]. In addition, Chapter 6 provides a thorough treatment of exact-constraint design, and examples of exact-constraint designs appear in Chapter 7. In this section, however, the basic concepts are introduced through statements that appear in [Blanding, 1992]. Although these statements deal specifically with *ideal* constraints, they provide the essential understanding of kinematics required for the design of *real* constraint systems. Following each statement is an explanation to reinforce and sometimes extend its meaning. The point here is to understand and visualize basic kinematic techniques rather than to apply a bunch of rules that are easy to forget.

Statement 1: Points on the object along the constraint line can move only at right angles to the constraint line, not along it.

A single-degree constraint prevents motion in one direction, the constraint direction, represented by a line in space. The only component of motion allowed by the constraint is perpendicular to the constraint line. If the object is rigid, then all points of the object along the constraint line are so constrained. The initial or so called instantaneous motion is always perpendicular to the constraint direction (for ideal constraints). The constraint direction may change as the object moves in which case the constrained path is curved and the instantaneous motion is tangent to the curve. For example, any point on a wheel with a fixed axle is constrained to a circular path. The constraint direction is radial and the instantaneous motion is tangent to the circle.

Statement 2: Any constraint along a given constraint line is functionally equivalent to any other constraint along the same constraint line (for small motions).

By Statement 1, the instantaneous motion is always perpendicular to the constraint line irrespective of the actual constraint. It follows that any constraint on a given constraint line produces the same instantaneous motion. For small motions about an operating point, the curved path of motion produced by any constraint on the given line is approximately equal to the instantaneous motion (or tangent) at the operating point. Thus, any two constraints on the same constraint line are approximately equal for small motions.

Statement 3: Any pair of constraints whose constraint lines intersect at a given point, is functionally equivalent to any other pair in the same plane whose constraint lines intersect at the same point. This is true for small motions and where the two constraints lie on distinctly different constraint lines.[1]

---

[1] The final sentence in Statement 3 was changed from "This is true for small motions and where the angle between constraints does not approach 0° (180°)" to allow the possibility of parallel constraints that effectively intersect at infinity.

The intersection described by this statement is an instantaneous center of rotation, or simply an instant center. The reduction of a constraint pair (or triple) to an instant center is an important visual and conceptual aid in the field of kinematics. We must distinguish, however, between the stationary point that is on or relative to the constrained body and the instant center that lies momentarily at this point. The point on the body can have only instantaneous motion that is perpendicular to the plane formed by the two constraints. In this plane the point will appear stationary for small motions while the instant center will appear to move with the moving constraint directions. Any pair of constraints that lie in this plane and intersect the same point will allow the same instantaneous motion of that point on the body; however, the motion of the instant center may be quite different. Any other point on the body may have an additional, tangential component of instantaneous motion about the instant center, as shown in Figure 2-16 (a).

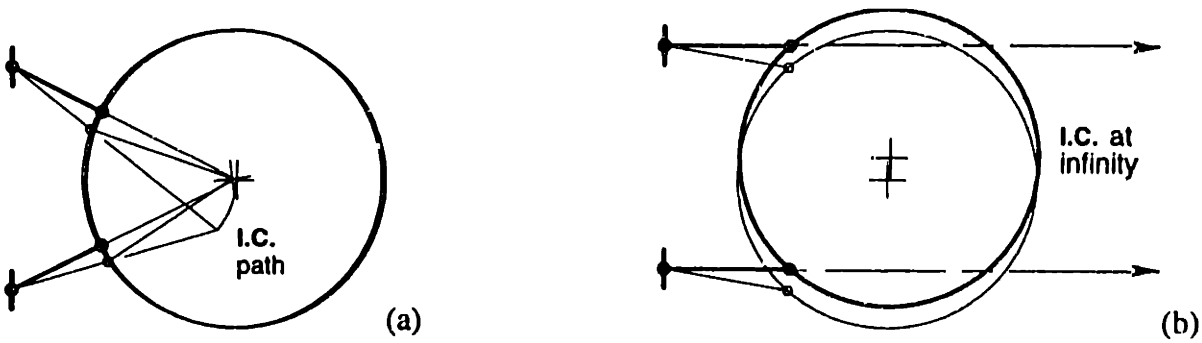


Figure 2-16 In (a), the instant center is momentarily located at the physical center of the circle (heavy lines). The instant center moves down (light lines) while the circle rotates approximately about its physical center. In (b), the instant center is off at infinity and the circle initially translates downward.

The condition that the two constraint lines be distinctly different requires further explanation as it leads to a key concept. If two constraints were to lie on a single line, then they would not define a plane and the statement would not make sense. The physical result would be one overconstrained degree of freedom rather than two constrained degrees of freedom. An acceptable case, however, is two parallel constraints that are separate and thus define a plane. As Figure 2-16 (b) indicates, we may consider parallel constraint lines to intersect at infinity such as railroad tracks appear at a distance. An object that rotates about a distant center appears to translate such as a ship appears as it rotates about the center of the Earth. With this background, we may conceptualize three translations and three rotations as being equivalent to six rotational degrees of freedom where three axes are at infinity.

Statement 4: The axes of a body's rotational degrees of freedom will each intersect all constraints applied to the body.[1]

[1] This is true if each axis provides uncoupled rotation. An example of coupling is the rotation of a lead screw with its associated translation. The consequence of violating Statement 4 is more complex motion.

This is a very powerful and comprehensive statement that uses explicitly the representation of translations as rotational axes located at infinity. It is a generalization of the instant center and is valuable as a visual aid to understanding a mechanism or in synthesizing the system of constraints for a new mechanism. The proof of this statement is quite trivial. If there exists an axis that intersects all applied constraint lines, then no constraint exists that can affect a moment about that axis because the lever arm is zero. Hence, the body is free to rotate about that axis, as demonstrated in Figure 2-17.

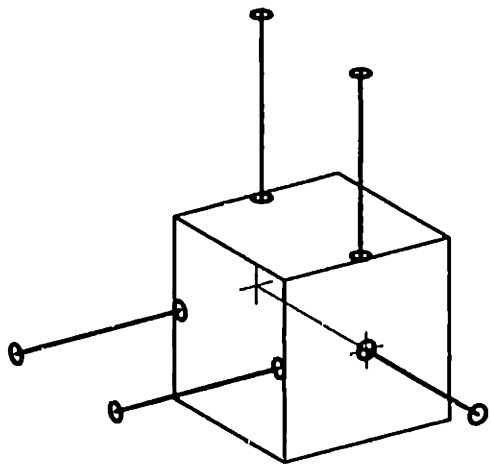


Figure 2-17 All five applied constraints intersect the only remaining degree of freedom, a rotation about the center of the cube.

Statement 5: A constraint applied to a body removes that rotational degree of freedom about which it exerts a moment.

In order to constrain a rotational degree of freedom (which includes translations by equivalency), the constraint must react with a moment about the axis of rotation. A constraint will satisfy this requirement if the constraint line does not intersect the axis of rotation and if the constraint line is not parallel to the axis of rotation.[1] An exception to the first condition will result in a zero-length lever arm. An exception to the second condition will result in a moment that has no component along the axis of rotation. Figure 2-18 shows the addition of a third constraint to prevent rotation of the circle about its center. Each constraint prevents rotation about the instant center formed by the other pair of constraints so that three degrees of freedom are exactly constrained. The axes of the remaining three rotational degrees of freedom will each intersect all three constraints per Statement 4. If the three constraints happen to lie in the plane of the figure, then so too will the axes of rotational freedoms.

[1] Just as parallel constraints intersect at infinity, an axis of rotation and a constraint line that are parallel to each other also intersect at infinity.

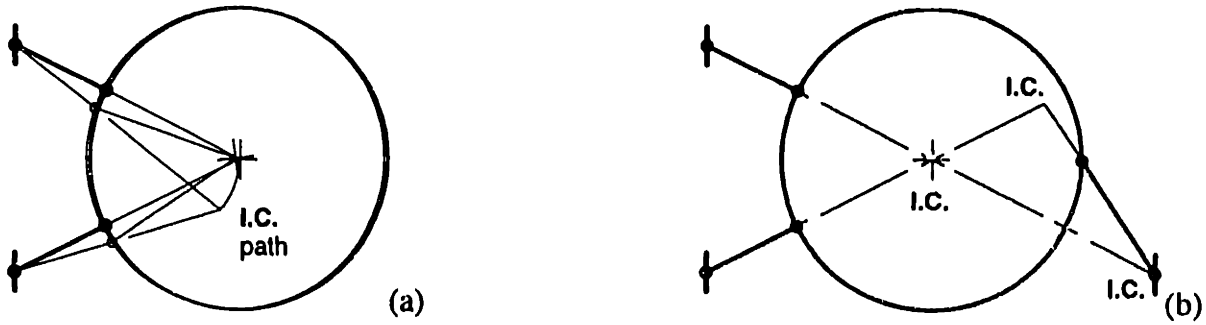


Figure 2-18 The rotational freedom of the circle about its center in (a) is constrained in (b) by the addition of a constraint that reacts with a moment about the center. Each constraint reacts with a moment about an instant center formed by the other pair of constraints.

The length of the lever arm, that is, a perpendicular drawn from the constraint line to the constrained rotational axis, is a relative measure of the effectiveness of that constraint. If one is seeking a balanced design, then a sensible approach is to seek lever arms having nearly equal length. This leads to a simple rule of thumb for planar problems; *arrange constraint lines to form an equilateral triangle*, as shown in Figure 2-19. Of course there may be valid reasons to choose different angles. Many three dimensional problems have a planar nature, which greatly helps visualization.

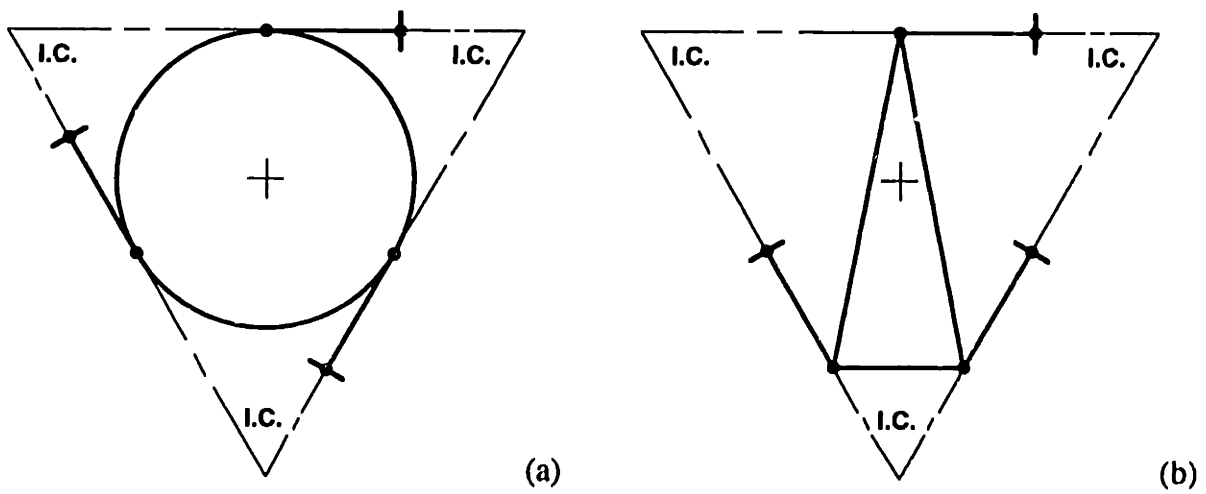


Figure 2-19 An equilateral arrangement of constraints often provides a better balance of stiffness. In (a) the center of stiffness lies at the center of the circle, and the vertical and horizontal stiffnesses are equal. The center of the triangular object in (b) is somewhat lower than the center of stiffness. A slightly wider angular spacing between the lower two constraints would lower the center of stiffness while increasing the horizontal stiffness and decreasing the vertical stiffness.

Statement 6: Any set of constraints whose constraint lines intersect a complete and independent set of rotational axes, is functionally equivalent to any other set of constraints whose constraint lines intersect the same or equivalent set of

rotational axes. This is true for small motions and when each set contains the same number of independent constraints.[I]

This is an extension of Statement 3, *the functional equivalence of any constraint pair having the same instant center*, based on Statement 4, *the generalization of the instant center to a rotational axis*. The statement is most meaningful and useful when the constraints total four or more, as at least two pairs of constraints are required to uniquely define an axis of rotation. The examples shown in Figure 2-20 are functionally equivalent (for small motions) to each other and to the example in Figure 2-17. Each has five constraints that uniquely define the same axis of rotation. It is natural to ask if there are any other arrangements of five constraints that will define the same axis of rotation. Blanding has developed a chart of all possible orthogonal constraints involving one to six constraints. Re-created in Figure 2-21, the orthogonal constraint chart provides an excellent starting point for any exact-constraint design.[II] An infinite number of nonorthogonal configurations is possible based on these basic configurations. Categories where a constraint arrangement does not exist usually require a series combination of constraint systems.

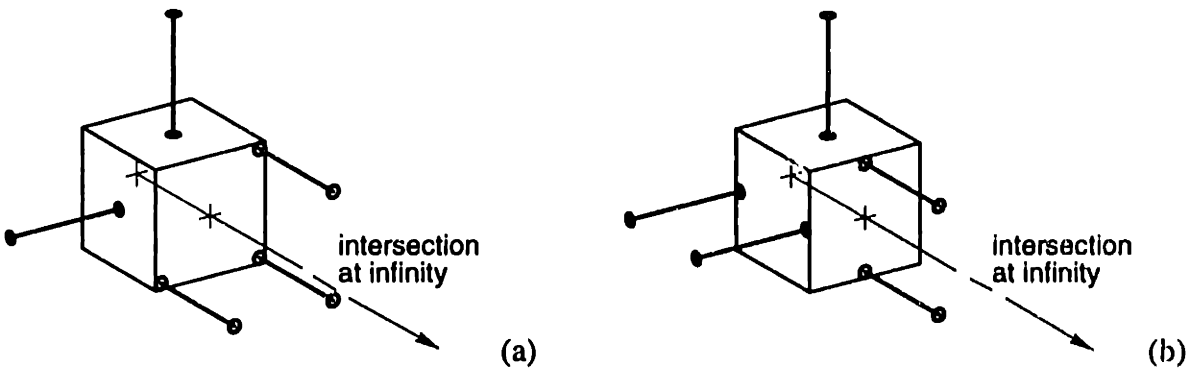


Figure 2-20 In both (a) and (b), all the applied constraints intersect the only remaining degree of freedom, a rotation about the center of the cube. These two constraint cases are functionally equivalent since all the constraints intersect the same rotational axis.

[I] Statement 6 was changed from "Each of a body's remaining rotational degrees of freedom is intersected by the line(s) of any applied constraint(s)" because it provided no new information beyond Statement 4. The definition of "equivalent sets of rotational axes" comes later in Statements 10 and 11.

[II] Blanding's original chart numbers the cells according to constraints .ather than degrees of freedom and does not show the centerlines and arrows to help in visualizing the degrees of freedom.

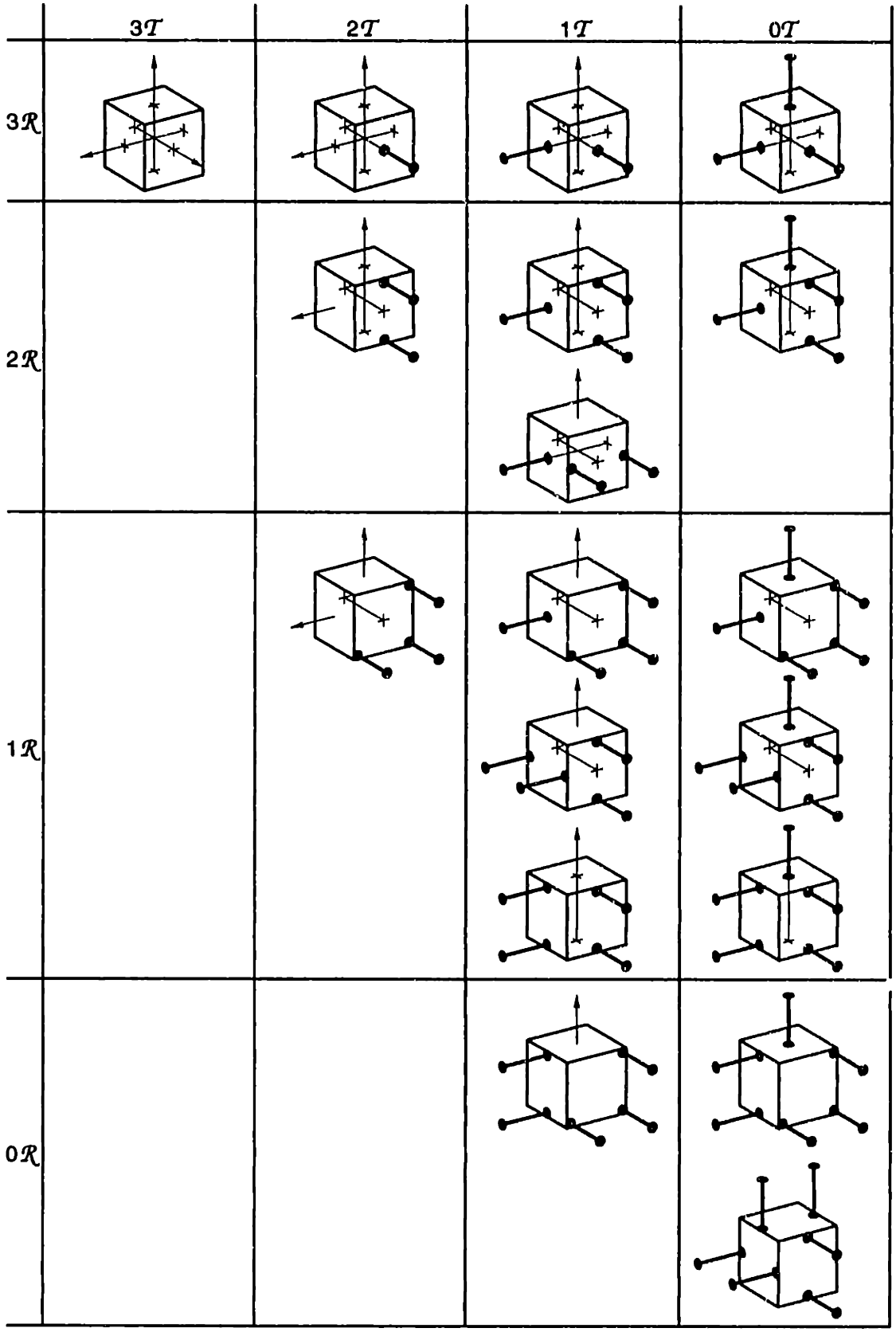


Figure 2-21 A matrix of desired rotational degrees of freedom (centerlines) and translational degrees of freedom (arrows) shows all possible orthogonal constraint arrangements, after [Blanding, 1992].

Statement 7: An *Ideal Sheet Flexure* imposes absolutely rigid constraint in its own plane (X, Y, and θ_Z), but it allows three degrees of freedom: Z, θ_X, and θ_Y.

A sheet flexure, more commonly called a blade flexure, is one of the most important constraint devices used in precision machines. A blade flexure allows out-of-plane motion while resisting in-plane motion as Figure 2-22 clearly shows. The equivalent three-constraint system superimposed on the flexure serves only as a conceptual aid. We know from experience that a thin blade is very compliant for out-of-plane bending. Equations 2.19 through 2.22 give the stiffnesses for the directions that are usually most relevant.[I] For a given axial stiffness, the moment stiffness varies as the square of the blade thickness. The desire to minimize blade thickness will invariably require much greater width than thickness so there is sufficient cross sectional area to carry the load and/or to provide axial stiffness. Generally, the size constraint on a blade flexure will be either the maximum width or the minimum thickness. Blade length affects moment stiffness and axial stiffness the same way and so is driven by other considerations. Usually the length to thickness ratio is limited to 10:1 for typical materials to avoid buckling. Equation 2.23 gives the condition required for the blade to yield before buckling. A blade sized to resist buckling usually is too short to have adequate translational freedom. In this case, two short blades spaced apart in the same plane and connected by a larger bar section will provide greater translational compliance by the square of the separation distance.[II] To the extent that the blade bends as a hinge, the bending stress is approximately constant over the length and given by Equation 2.24. Combining the stress and buckling relations leads to a bound on bending angle due only to material properties as expressed in Equation 2.25. An angle of 3° is reasonable for hardened steel but the resistance to buckling decreases with angle. The material parameters E, ν and σ_y represent elastic modulus, Poisson ratio and yield strength, respectively.

$$k_x = \frac{E \cdot t \cdot w}{a} \tag{2.19}$$

$$k_z = \frac{k_x t^2}{\left(1-\nu^2\right)a^2 + 2.4(1+\nu)t^2} \tag{2.20}$$

$$k_{\theta_y} = \frac{k_x t^2}{12\left(1-\nu^2\right)} \tag{2.21}$$

[I] Finite element studies of typical blade flexures reveal that plane stress better approximates axial stiffness while plane strain better approximates bending stiffness. As a result, these formulas lead to slighly more conservative designs. See Chapter 6.2 for further details and derivations.

[II] See Statements 11 and 12 for further explanation.

$$k_{\theta_x} = \frac{k_x t^2}{12}\left[\frac{1}{2(1+\nu)}\left(4+2.52\frac{t}{w}\right)+\frac{w^2}{\left(1-\nu^2\right)a^2+2.4(1+\nu)t^2}\right] \tag{2.22}$$

$$\frac{a}{t} < \pi\sqrt{\frac{E}{12\sigma_y}} \tag{2.23}$$

$$\sigma_b = \frac{E\cdot t}{2a}\theta = \frac{k_x}{2w}\theta \tag{2.24}$$

$$\theta < \pi\sqrt{\frac{\sigma_y}{3E}} \tag{2.25}$$

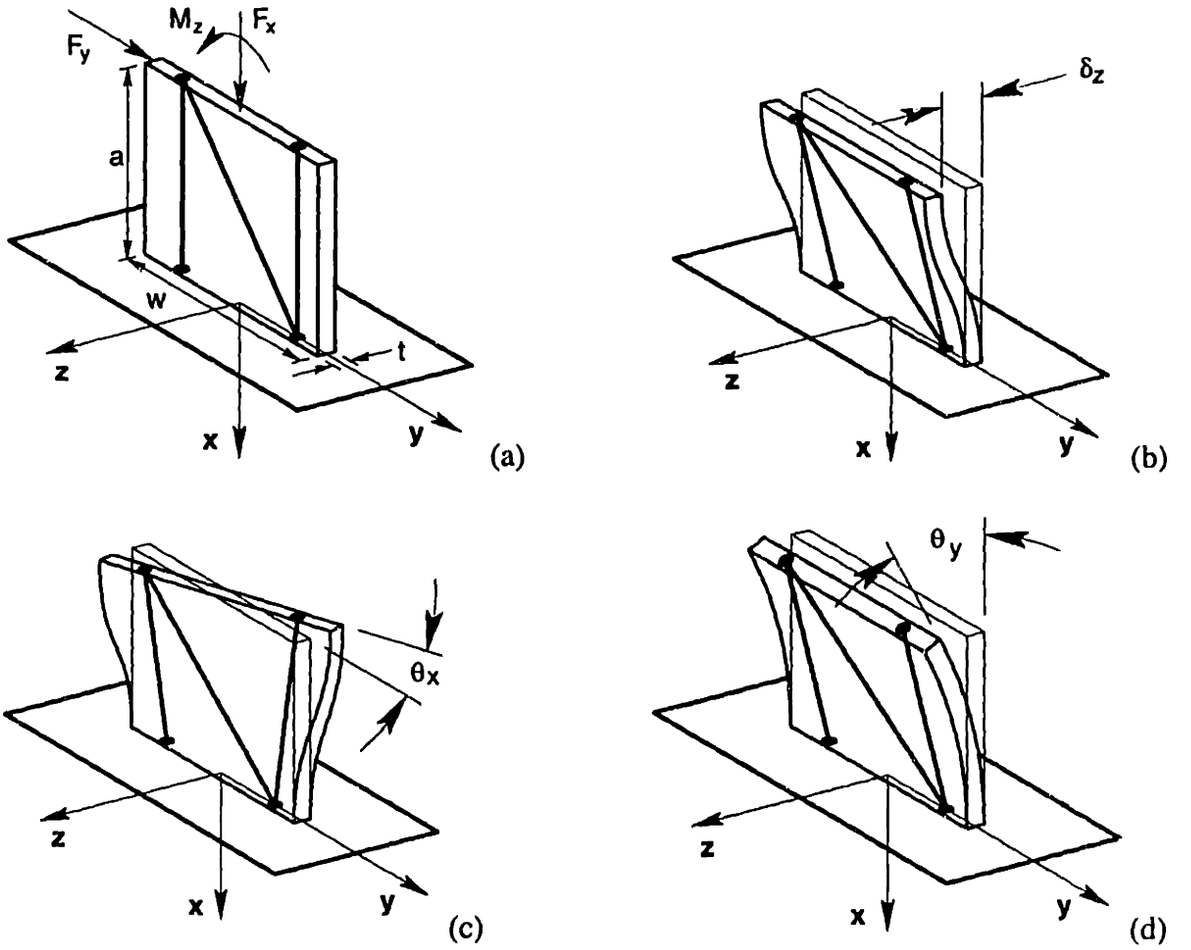


Figure 2-22 A blade flexure provides constraint against forces and moments in the plane of the blade (a). A blade flexure provides freedom to small motions in bending modes of the blade (b, c, d). The blade is represented equivalently by three single-degree constraints as shown.

Statement 8: An *Ideal Wire Flexure* imposes absolutely rigid constraint along its axis (X), but it allows five degrees of freedom: Y, Z, θ_x, θ_y, θ_z.

This is the ideal constraint that has become so familiar by now. Often two wires are used in opposition to constrain a single degree of freedom. Such a configuration resists buckling and doubles the axial stiffness if the wires have pretension. Pretension has the adverse effect of increasing lateral stiffness by the ratio of tension to length for each wire. A better alternative is simply to limit the length to diameter ratio to approximately 10:1, thereby stiffening translation to better resist buckling. To recover translational freedoms, two short wires spaced along the constraint line and connected by a larger section rod will provide greater translational compliance by the square of the separation distance.

> Statement 9: A constraint (C) properly applied to a body (i.e., without overconstraint) has the effect of removing one of the body's rotational degrees of freedom ($\mathcal{R}$'s). The $\mathcal{R}$ removed is the one about which the constraint exerts a moment. A body constrained by n constraints will have $6-n$ rotational degrees of freedom, each positioned such that no constraint exerts a moment about it. In other words, each $\mathcal{R}$ will intersect all C's.[1]

This is an extension of Statements 4 and 5 that provides a way to test for overconstraint and underconstraint. The first test is simply to count the number of constraints. We can generalize to nonrigid bodes by increasing the number of free-body degrees of freedom by f flexural degrees of freedom. The number of independent C's required to exactly constrain a body is $n=6+f-d$, where d is the number of desired degrees of freedom. The second test is required to determine whether the C's are independent. The removal of a redundant C will not affect the number of degrees of freedom that the remaining C's allow. The system is exactly constrained if the removal of any single constraint increases the number of degrees of freedom by one. Figure 2-23 provides specific examples of this and the orthogonal constraint chart, Figure 2-21, provides further examples by observing changes between adjacent cells.

Statement 9 as written has one small technical problem that may be discovered when performing the test for independent constraints. A single degree of freedom may consist of a coupled rotation and translation such as the motion described by the lead of a screw. This behavior occurs when a constraint does not intersect the axis of rotation (thus they are not parallel either), thereby introducing translation along the same axis. Figure 2-25 shows an example that demonstrates this behavior. Chapter 6 presents a flexure coupling for ball screws that is designed to have the same lead as the screw.

---

[1] Blanding uses the notation $\mathcal{R}$, $\mathcal{T}$, C throughout his book to represent rotational freedom, translational freedom and constraint, respectively. It just happens to show up in Statements 9 through 11.

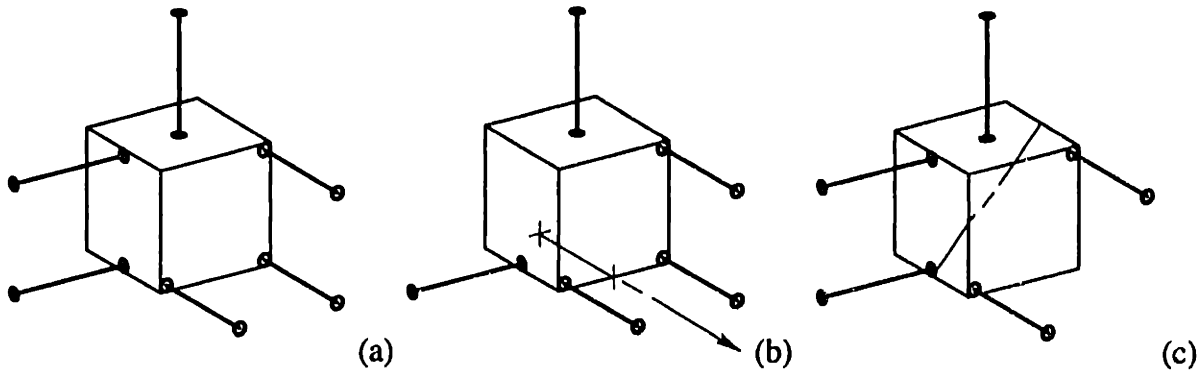


Figure 2-23 To test whether the six constraints in (a) are independent, remove any single constraint and see if a new degree of freedom results. The constraint removed in (b) can no longer exert a moment about the axis shown and neither can the remaining constraints. The same applies to (c) and the remaining cases.

Statement 10: Any pair of intersecting rotational degrees of freedom ($\mathcal{R}$'s) is equivalent to any other pair intersecting at the same point and lying in the same plane. This holds true for small motions.

Another way of stating this is that *a pair of intersecting $\mathcal{R}$'s can generate instantaneous rotation about any axis that lies in the plane and passes through that point.* As Figure 2-24 shows, the constraints that allow this motion must either lie in this plane or intersect the plane where the two $\mathcal{R}$'s intersect by Statement 4. The small motion requirement is necessary even for pure axes of rotation because rotation about one axis changes the orientation of the other. Usually this is not a problem and complicates only the algorithm that computes the angles. A familiar example is Hooke's coupling (the typical universal joint). It transmits shaft power through a bend but its transmission ratio is 2ω cyclic with an amplitude that increases with the square of the angle.

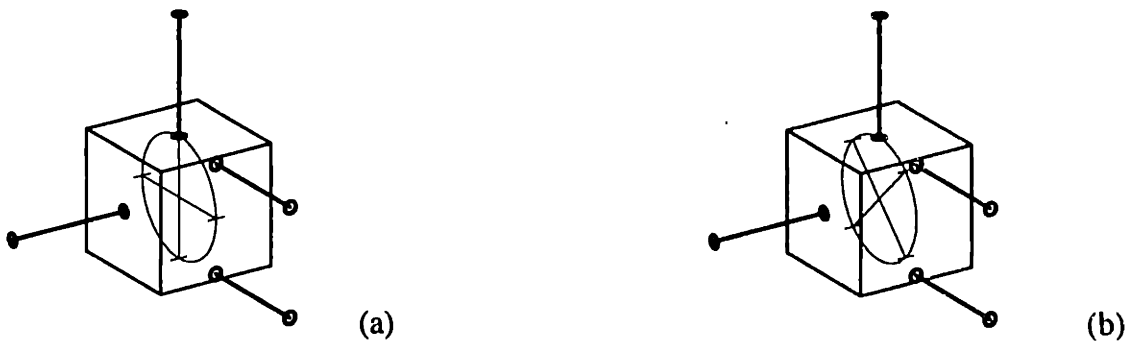


Figure 2-24 The circle lies in the plane of three constraints and the fourth constraint intersects its center. Any distinct pair of rotational axes that lie in this plane and intersect the center will represent the instantaneous motions allowed by the constraints.

Statement 10 extends to the intersection of three $\mathcal{R}$'s so long as each triple spans three-dimensional space. Three $\mathcal{C}$'s that intersect at the same point will allow instantaneous rotation about any axis that passes through that point. Figure 2-25 shows a flexure pivot that has three *nonintersecting* wire constraints. This flexure provides three $\mathcal{R}$'s that span

three-dimensional space, but the motion is more complicated since the $\mathcal{R}$'s do not intersect. This leads to an interesting coupling between rotations and translations, or equivalently between forces and moments. The easiest one to visualize is the rotation about the axis of symmetry causing a translation along the same axis. The consequence of constraints not intersecting the axis of rotation is the motion described by the lead of a screw.

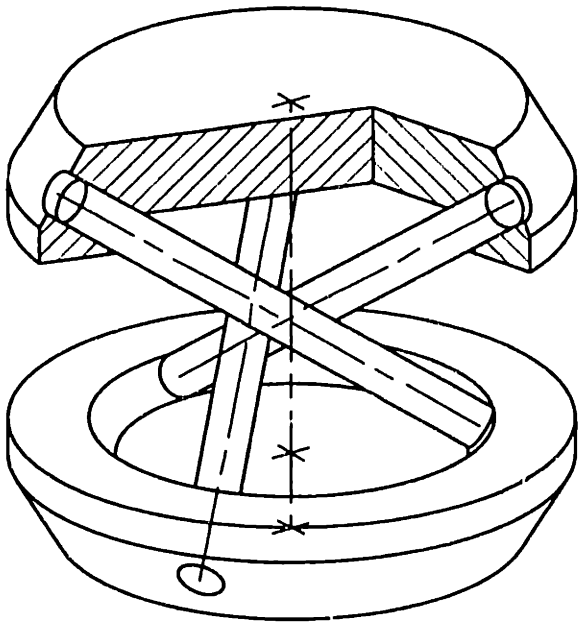


Figure 2-25 This flexure was created by shifting mutually orthogonal wires off center so that they clear one another. The angle of each wire with respect to the center axis is 54.736° (the arc tangent of $\sqrt{2}$). The offset of each wire from center is the desired distance between wire centers divided by $\sqrt{2}$.

Statement 11: Two parallel $\mathcal{R}$'s are equivalent to any two parallel $\mathcal{R}$'s, parallel to the first pair and lying in the same plane. They are also equivalent to a single $\mathcal{R}$ parallel to the first pair and lying in the same plane; and a $\mathcal{T}$ perpendicular to that plane.

This follows from Statement 10 where the point of intersection occurs at infinity. The small motion requirement has been dropped because rotation about one axis does not alter the orientation of a parallel axis, which is a first order effect for nonparallel axes. There is, however, a second order translation that depends inversely upon the distance between the parallel axes. Thus in the strictest sense, there is a small motion requirement.

Statement 12: When parts are connected in series (cascaded), add the degrees of freedom. When the connections occur in parallel, add constraints.[1]

---

[1] The series addition of functionally equivalent degrees of freedom results in an indeterminacy that may or may not present a problem.

We shall use two familiar rules to explain Statement 12. Rule 1: *The equivalent compliance of springs connected in series is the sum of their individual compliances.* Rule 2: *The equivalent stiffness of springs connected in parallel is the sum of their individual stiffnesses.*[I] We may recall these rules applied to single-degree-of-freedom springs, but they also apply to springs and structures of any dimension, where the spring constant becomes a symmetric matrix. We may consider a degree of freedom as being a dominant term in the compliance matrix whether its origin is the elasticity of a blade flexure or the motion of a bearing. Similarly, we may consider a constraint as being a dominant term in the stiffness matrix. The foundation for Statement 12 is that dominant terms in individual matrices remain dominant through the addition process.[II] Therefore, degrees of freedom dominate through series combinations while constraints dominate through parallel combinations. This exposes a subtlety that is not apparent in Statement 12, namely, how to deal with redundant constraints and degrees of freedom. We will work through these by example starting first with a series combination followed by a parallel combination.

Figure 2-26 (a) shows a series combination of two blades that share a common constraint line. Someone could misinterpret Statement 12 to mean that this series of blades, each with three degrees of freedom, would combine to have a total of six degrees of freedom and no constraints. The combined axial compliance along the constraint line is still orders of magnitude more rigid than the other directions, thus it remains a constraint. Likewise, the blades share a common rotational axis that results in a redundant degree of freedom. The combination may have twice the compliance but functionally remains a single degree of freedom. In the remaining directions, Statement 12 applies without confusion as four degrees of freedom combine with four constraints. In practice, we usually keep the blades short and duplicate another set further down the constraint line to provide much greater translational freedom. It is functionally equivalent to a wire flexure but with much higher axial stiffness and load capacity.

Figure 2-26 (b) shows a more elaborate flexure that has both series and parallel combinations of blades. This design has a hole down the center to prevent a *short circuit* in the desired degrees of freedom. Its symmetry leads to redundant constraints. It could function the same without the redundant constraints (by eliminating the symmetry) but this would sacrifice too much stiffness along its constraint line.

To summarize, the proper way to interpret Statement 12 requires an awareness of the directions involved. For a series combination of parts, the combined degrees of freedom will span the *union* of dimensional spaces spanned by the degrees of freedom of

---

[I] The stiffness of a spring is the load required to produce a unit displacement and the compliance is the inverse of stiffness.

[II] The individual matrices must be of the same size (generally 6 by 6) and with respect to the same coordinate system, which may require multiplication by a coordinate transformation matrix.

the individual parts; whereas, the combined constraints will span the *intersection* of spaces spanned by the constraints of the individual parts. For a parallel combination of parts, the combined constraints will span the *union* of dimensional spaces spanned by the constraints of the individual parts; whereas, the combined degrees of freedom will span the *intersection* of spaces spanned by the degrees of freedom of the individual parts. Fortunately this is easier to understand than it is to phrase.

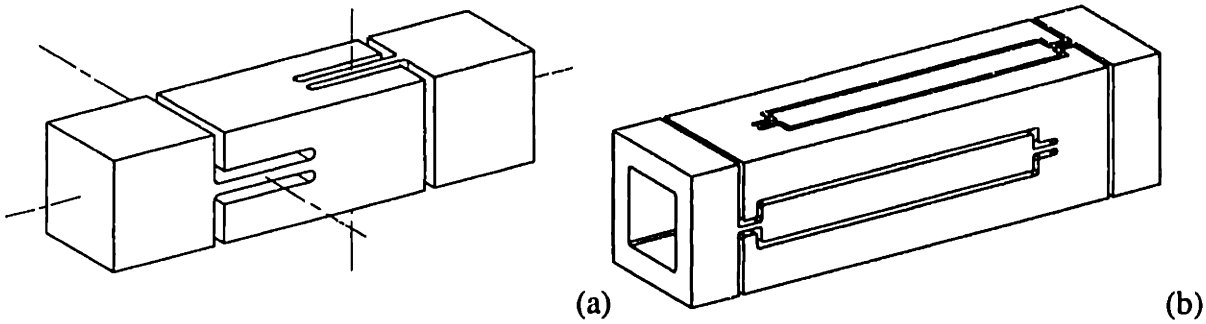


Figure 2-26 In (a), a series combination of two blades provides three independent rotational degrees of freedom and one axial constraint. Cutting these features into each end of a longer bar provides two translational freedoms. In (b), the blades are effectively longer to provide two translational freedoms but a center hole is required to prevent a short circuit around the series of flexures. The cross section could be round rather than square. These monolithic flexures can be manufactured using wire EDM.

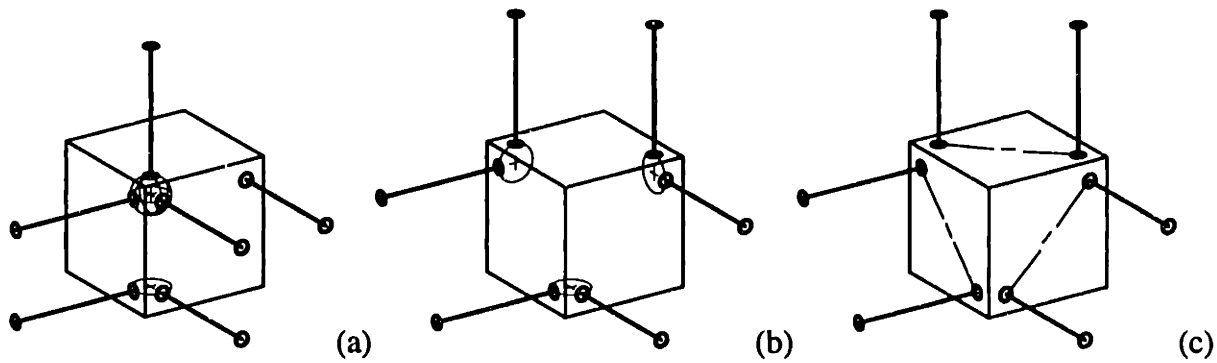


Figure 2-27 In (a), a sphere in a tetrahedral socket provides three constraints; a sphere in a vee provides two constraints; and a sphere on a plane provides the last constraint. In (b), three spheres in three vees each provide two constraints. In (c), three cylinders on three planes each provide two constraints.

The 12 Statements by Blanding are important to understand and to refer to when the time comes to design a constraint system or mechanism. In many cases it is possible to start with a basic design such as one of the familiar kinematic couplings, but as Figure 2-27 shows, any of these can be invented from the orthogonal constraint chart, Figure 2-21, along with a basic understanding of constraint devices such as spheres in vees. Be aware that nice orthogonal constraints and intersecting axes are artificial restrictions used to keep the chart simple and bounded. They should not restrict creativity as many examples in this thesis will demonstrate. Figure 2-28 is such an example of a two-axis gimbal where the rotational axes do not intersect. This design would be appropriate for supporting the principal load along the axis of symmetry such as required for a rocket motor.

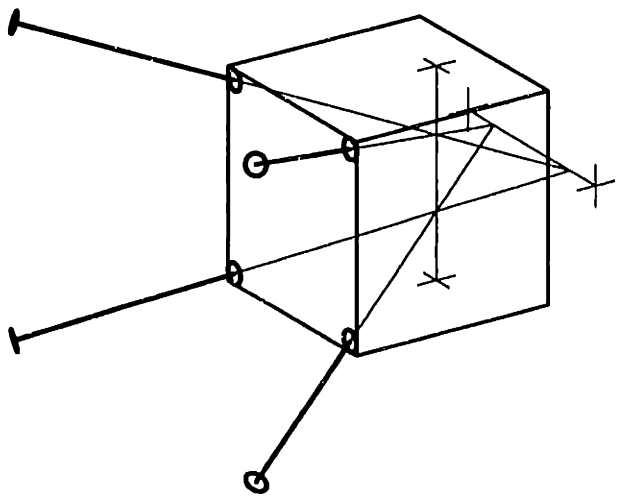


Figure 2-28 Four constraints intersect (and thus allow) two rotational degrees of freedom. The constraints prevent rotation about a third perpendicular axis and translations of those axes. If all four constraints intersected at a point, the result would be an overconstrained, three-axis pivot.

A key concept in exact constraint design is this one-to-one relationship between applied constraints and constrained degrees of freedom. You may choose to add redundant constraints to increase stiffness or to provide a nesting force for contact-type constraints; however, this usually will sacrifice one or more benefits of exact-constraint design. That is to say, the best design may not be the exactly constrained one, but you should begin there so that the implications of overconstraint are fully considered and expected.

2.7 Elastic Averaging

The term elastic averaging describes a condition where two objects are connected through many points of contact in a highly overconstrained manner. Elastic averaging seems so contrary to exact-constraint design that some people may argue one philosophy over the other rather than embracing their complementary virtues. Many kinematic designs rely on bearing systems that function by elastic averaging. A new class of machine tool based on the Stewart platform is a good example [Stewart, 1965-66]. These machines have become know as hexapods because they feature six actuators operating as a parallel-link mechanism to control the relationship between the tool and the workpiece. The actuators typically employ rolling-element bearing technology including ball screws, angular-contact thrust bearings and trunnion bearings. Since many balls share the load, irregularity of any particular ball has little influence over the net error motion. Being massively overconstrained, these devices require very accurate surfaces to fit and function properly, but once achieved, the result is further reduction of error motion due to the averaging of imperfections. Furthermore, the stiffness and load capacity multiply with the number of constraints sharing the load. Rolling-element bearings come in many types and sizes and are relatively inexpensive when mass produced in large quantities.

Elastic averaging is frequently used in the manufacturing process to increase the precision of certain components beyond the capability of available machine tools. Many such examples are described in the book *Foundations of Mechanical Accuracy* [Moore, 1970]. The lapping of a lead screw improves the uniformity of its pitch diameter and lead but not necessarily the accuracy. Symmetrical designs that allow physical reversal or closure techniques can often be manufactured with nearly zero error about the generator of the symmetry. [Slocum, 1992] refers to this property as self checking. This property is exploited in the manufacture of the Moore 1440, which is an angle standard that provides one-quarter degree increments with one-tenth arc second accuracy. It is a face coupling with 1440 teeth that have been self lapped by engaging and disengaging the coupling through all step angles over many hours with progressively finer lapping compound. This process converges the massively overconstrained surfaces to a conjugate shape that naturally satisfies the requirement of equal-angle increments.

2.8 Thermal Management

> By far the most common cause of nonrepeatability is temperature. More specifically, the cause is changing temperature with time, which causes thermal distortion of the structural loop (which consists of all the mechanical elements involved in holding the tool and work in a given relative position). Interestingly, temperature problems also seem to be the least widely appreciated error source in machine tools.
>
> [Donaldson, 1972]

The international standard temperature, where a solid object has its true size, is 20°C (68°F). This implies that the temperature distribution must be constant and uniform throughout the workpiece. By international agreement, metrology at any other temperature contains thermally induced errors, although the systematic components can be reduced through compensation. The thermally induced errors in the measuring machine or machine tool are usually more significant because of its size and complexity. [Donaldson, 1979] describes a series of methods to increase the thermal stability of a machine and the workpiece (specifically the LODTM). The methods generally follow an order of passive to active, which should be considered the preferred order for any machine design. These methods are presented in brief below.

1 *Reduce sensitivity*

   a) *Structural design*: Obtain symmetric temperature distributions in symmetric structures to reduce distortions.

   b) *Low-expansion materials*: Use low CTE materials to reduce variations in geometry due to variations in temperature.

2 *Manage heat sources*

a) *Eliminate*: Place necessary heat sources outside the controlled environment and eliminate unnecessary heat sources.

b) *Reduce*: Use components that dissipate less heat.

c) *Isolate*: Capture the heat near the source or prevent it from spreading into the structures.

d) *Avoid cascading*: Fluids that remove heat from an isolated source should return directly to the temperature control system rather than flowing over other sensitive parts of the system.

e) *Keep heat sources constant*: Necessary heat sources within the controlled environment such as lights should remain constant.

3 *Control machine environment*

a) *Control room air temperature*: Reduce temperature variations in the machine by controlling the air temperature around it.

b) *Isolate machine room*: Prevent heat leakage into or out of the machine room to reduce variations in the room air temperature.

c) *Isolate machine structure*: Control the temperature of the metrology loop.

d) *Control workpiece temperature*: Use a temperature controlled fluid flowing over the workpiece to control its temperature. Consider the effect of viscous heating in high-speed fluid flows.

e) *Isolate machine operator*: The human body represents a heat source of about 100 watts. Precision applications may require the use of insulating clothing such as gloves.

4 *Compensate for measured deviations*

a) *Spindle growth compensation*: Use a suitable displacement sensor such as a capacitance transducer to measure the axial spindle growth and compensate the z-axis position.

b) *Metrology frame temperature*: Measure the temperature of the metrology loop in sufficient detail to compute the error and compensate the slide positions.

Experience has shown that temperature control is the most reliable and effective means to reduce thermal errors. Further, it is usually one of the least expensive aspects of a precision machine tool and usually requires little additional hardware. The design challenge is figuring out how to provide sufficient control for the least costs. The cause-and-effect relationships can be calculated in considerable detail using modern computer software (finite element analysis and computational fluid dynamics) and empirical heat transfer formulas, but to do so requires considerable knowledge about the design and the

environment. Data obtained from existing systems or specific tests may be easier to obtain and just as valuable. A number of simple models are presented in the remainder of this section to gain greater understanding of the important parameters and their relationships to the thermal control problem. [Holman, 1976] was referenced in preparing these models.

Suppose a solid bar has a constant temperature gradient as expressed in Equation 2.26. We wish to calculate the local displacements and rotations for any point in the bar relative to the fixed y-z plane shown in Figure 2-29. The thermal strain gradient is related to the temperature gradient by the coefficient of thermal expansion α. Local rotations are obtained in Equation 2.27 by integrating the strain gradient from the fixed end.[1] Local displacements are obtained in Equation 2.28 by integrating the temperature difference along the x axis and by integrating the rotations about the y and z axes.

$$T(x,y,z) = T_0 + x\frac{\partial T}{\partial x} + y\frac{\partial T}{\partial y} + z\frac{\partial T}{\partial z} \tag{2.26}$$

$$\theta_y(x) = \int_0^x \alpha \frac{\partial T}{\partial z}\,dx = \alpha x \frac{\partial T}{\partial z} \qquad \theta_z(x) = -\int_0^x \alpha \frac{\partial T}{\partial y}\,dx = -\alpha x \frac{\partial T}{\partial y} \tag{2.27}$$

$$\begin{aligned} \delta_x(x,y,z) &= \int_0^x \alpha(T - T_0)\,dx = \alpha x\left(\frac{x}{2}\frac{\partial T}{\partial x} + y\frac{\partial T}{\partial y} + z\frac{\partial T}{\partial z}\right) \\ \delta_y(x) &= \int_0^x \theta_z\,dx = \frac{-\alpha x^2}{2}\frac{\partial T}{\partial y} \qquad \delta_z(x) = -\int_0^x \theta_y\,dx = \frac{-\alpha x^2}{2}\frac{\partial T}{\partial z} \end{aligned} \tag{2.28}$$

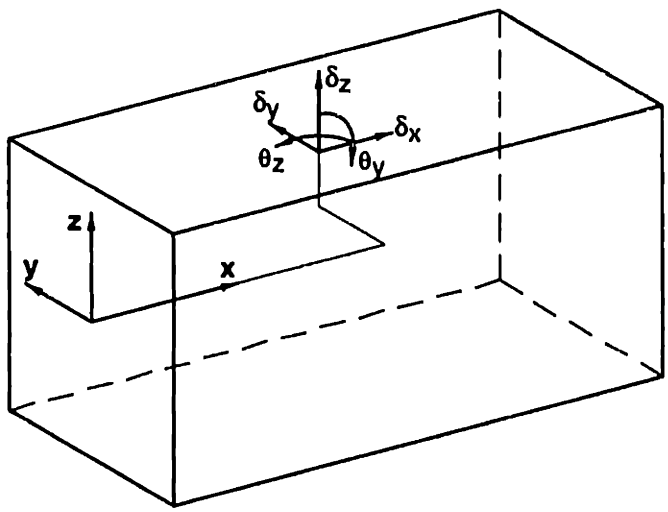


Figure 2-29 Displacements and rotations at a point (x, y, z) relative to the fixed x-y-z coordinate system subject to a constant temperature gradient.

[1] A similar expression exists for $\theta_x(y, z)$ but is insignificant for a relatively long bar. There is no x dependence because the bar does not twist.

Frequently it is more convenient to specify a steady, unidirectional heat flux (heat flow q divided by the area A) rather than a temperature gradient. For example, solar radiation is on the order of 1 kW/m^2. The heat flux through the material is equal to the temperature gradient multiplied by the thermal conductivity k. Assuming the heat flux is uniform and flows in either the x, y or z direction, the end point displacement of the bar has the same magnitude regardless of the direction of flow as Equation 2.29 and Figure 2-30 show. The important material property is the ratio of thermal expansion to conductivity α/k.

$$\frac{q_i}{A} = -k\frac{\partial T}{\partial i} \quad \Rightarrow \quad \delta_i = \frac{\alpha}{k}\frac{L^2}{2}\frac{q_i}{A} \qquad i = (x, y, z) \tag{2.29}$$

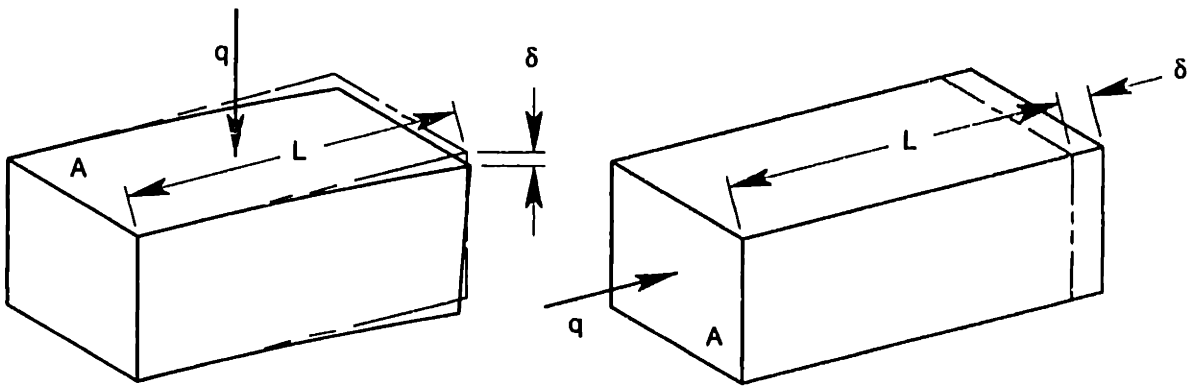


Figure 2-30 The end displacement due to a specified uniform heat flux through a solid bar.

Time dependent thermal problems involve the thermal mass of the structure interacting with conductive, convective and/or radiative heat transfer. Simplification of the problem leads to equations that describe each type separately. Beginning with conduction, the one-dimensional diffusion equation, shown in Equation 2.30, provides the important grouping of material properties that gives a measure of the time required for heat to diffuse or dissipate into the structure. The inverse of this group called diffusivity may be interpreted as a rate of dissipation. See Table 2-3 for the description of symbols.

$$\frac{\partial^2 T}{\partial x^2} = \frac{\rho C}{k}\frac{\partial T}{\partial t} \quad \Rightarrow \quad \Delta t \approx \frac{\rho C}{k}(\Delta x)^2 \tag{2.30}$$

The rate that the structure changes temperature in a bulk sense depends on the heat transfer to the surroundings. If the thermal resistance of the structure is low relative to the resistance of convection and/or radiation at its surface, then Equation 2.31 is a reasonable approximation that describes the system time constant τ. It expresses an isothermal lump of heat capacity exchanging heat to the surroundings in proportion to the heat-transfer coefficient h, the surface area A and the temperature difference. For plate-like structures, the volume V divided by the area is simply the plate thickness for single-side heat transfer or one-half the thickness for two sides. The system acts as a low-pass filter to attenuate temperature fluctuations in the surroundings that occur fast relative to the time constant.

Equation 2.32 is the transfer function that relates the thermally induced strain resulting from surrounding temperature fluctuations at a frequency ω. A check of the basic assumption is worthwhile and requires only comparing the times computed in Equations 2.30 and 2.31.

$$\tau \frac{\partial T}{\partial t} + T = T_\infty \qquad \tau = \frac{\rho C V}{h A} \tag{2.31}$$

$$\frac{|\varepsilon(\omega)|}{|T_\infty(\omega)|} = \frac{\alpha}{\sqrt{\tau^2\omega^2 + 1}} \cong \frac{\alpha}{\tau\omega} \quad \text{for} \quad \tau\omega > 1 \tag{2.32}$$

Table 2-3 shows a table of material properties and significant property groups for steel, aluminum, invar-36 and natural granite. The low coefficient of thermal expansion (CTE) for invar-36 is clearly an advantage, although the ability of aluminum to conduct heat nearly offsets this advantage under a steady heat flux. Granite structures attenuate temperature fluctuations very well due to low conductivity and massive construction.

Symbol	*Description*	*Units*	Steel	Aluminum	Invar 36	Granite
ρ	density	Mg/m^3	7.9	2.71	8.03	2.6
α	CTE	μm/m/C	12	23	1.2	6
k	conductivity	W/m/C	54	177	11	4
C	specific heat	kJ/kg/C	0.46	0.896	0.46	0.82
ρC	heat capacity	MJ/m^3/C	3.63	2.43	3.69	2.13
ρ C/k = 1/D	time to diffuse	s/mm^2	0.07	0.01	0.34	0.53
α/ k	exp. due to heat	μm/W	0.22	0.13	0.11	1.50
α/(ρ C)	attenuate by conv.	rel. steel	1.00	2.87	0.10	0.85
αk/(ρ C)	attenuate by cond.	rel. steel	1.00	9.40	0.02	0.06

Table 2-3 Properties and significant property groups for common structural materials.

The effectiveness of a fluid in a temperature control system depends primarily on its heat capacity, that is, the ability to carry away heat, and its convection heat-transfer coefficient. The heat-transfer ability increases with fluid velocity but there is a practical limit due to viscous heating of the fluid. Equation 2.33 provides an estimate of the velocity limit for an acceptable temperature rise ΔT in the fluid. Equation 2.34 shows a common formula for forced-convection heat transfer and the grouping of fluid properties that governs the coefficient. Equation 2.35 for natural-convection heat transfer is slightly different because the flow is buoyancy driven.[1] Usually the temperature differential between, say, an instrument and the walls of an environmental chamber, is relatively small making natural convection (for air) comparable to radiative heat transfer. Equation 2.36 describes radiation between an object at T_1 enclosed within a relatively large chamber at T_2. Assuming an emissivity of 0.5 and a mean temperature of 293° K, the radiation heat-transfer coefficient would be 2.85 W/m²/°C. Table 2-4 shows a table of properties, property groups and

[1] These equations are shown merely to identify the important fluid properties. Please refer to a heat transfer text when making calculations for a specific problem.

example heat-transfer coefficients for water, light weight oil and air. Water is clearly superior as a powerful temperature control medium, however corrosion and evaporative cooling are problems with water not usually encountered with oil or air.

$$u \approx \sqrt{\frac{k}{\mu}\Delta T} \tag{2.33}$$

$$Nu = const\ Re^m Pr^n \quad \Rightarrow \quad h \propto k\left(\frac{\rho}{\mu}\right)^m\left(\frac{C\mu}{k}\right)^n$$

$$m \approx \begin{cases} 1/2 \\ 4/5 \\ 0 \\ 4/5 \end{cases} \qquad n \approx \begin{cases} 1/3 & \text{external laminar} & Re_L < 5\cdot 10^5 \\ 1/3 & \text{external turbulent} & Re_L > 5\cdot 10^5 \\ 0 & \text{internal laminar} & Re_d < 2300 \\ 1/3 & \text{internal turbulent} & Re_d > 2300 \end{cases} \tag{2.34}$$

$$Nu = const\,(Gr\ Pr)^m \quad \Rightarrow \quad h \propto k\left[g\left(\frac{\rho}{\mu}\right)^2\left(\frac{C\mu}{k}\right)\right]^m \qquad m \approx \frac{1}{4} \tag{2.35}$$

$$h \equiv \sigma\,\varepsilon_1 \frac{T_1^4 - T_2^4}{T_1 - T_2} \equiv 4\sigma\,\varepsilon_1 T_m^3 \qquad \sigma = 5.669\cdot 10^{-8}\frac{W}{n.^2\,{}^\circ K^4} \tag{2.36}$$

Symbol	*Description*	*Units*	Water	Oil SAE 10	Air
ρ	density	kg/m^3	998	917	1.21
μ	viscosity	kg/m/s	0.001	0.08	1.9E-05
ν = μ/ρ	kinematic viscosity	m^2/s	1E-06	8.7E-05	1.6E-05
k	conductivity	W/m/C	0.602	0.145	0.0257
C	specific heat	J/kg/C	4179	1880	1006
Pr = Cμ/k	Prandtl number	nondim.	6.94	1037.24	0.75
ρC	heat capacity	kJ/m^3/C	4171	1724	1.22
u (ΔT=.001C)	velocity for heating	m/s	0.78	0.04	1.16
Eq. 2.34 (lam.)	forced conv. group	SI units	1147	157	5.87
Eq. 2.34 (turb.)	forced conv. group	SI units	72342	2595	162
Eq. 2.35	natural conv. group	SI units	1728	156	11
Re (L=1m)	Reynolds number	nondim.	7.7E+05	4.9E+02	7.3E+04
h (L=1m)	forced conv. coeff.	W/m^2/C	671.0	21.5	4.2
"	"	rel. water	1.000	0.032	0.006
Gr (ΔT=.1C)	Grashof number	nondim.	3.3E+09	4.4E+05	1.3E+07
h (L=1m)	natural conv. coeff.	W/m^2/C	138.5	12.5	0.85
"	"	rel. water	1.000	0.090	0.006

Table 2-4 All fluid properties and significant property groups are taken at 20°C. The Reynolds number is calculated for flow over a 1 meter long plate using the velocity u calculated to produce a very conservative viscous temperature rise of 0.001°C. Water in this example transitions to turbulent flow near the end, but the force-convection heat-transfer coefficient is calculated assuming laminar flow for consistency. The Grashof number is calculated for a temperature differential of 0.1° C. The natural-convection heat-transfer coefficient is calculated for a 1 m vertical plate, but the dependence on shape and size is relatively weak.

2.9 Materials Selection

Materials selection can involve some of the most important decisions made in the design of precision machines. Even when the obvious choice is steel, for example, there are more subtle choices to make regarding the type of steel or alloy, whether to forge, cast or use wrought stock, the type of heat treatment and hardness, and other material processing steps (joining, machining, surface treatment, etc.). The level of information goes from basic physical properties to practical knowledge to material science. This section addresses the more basic level, selecting materials based on physical properties that somehow optimize the design. It emulates the work of [Ashby, 1989, 1992] whose property groups are formulated to represent the optimal material requirement of the design. His material property maps show clearly which materials best satisfy the various property groups. [Chetwynd, 1987] presents a similar approach but with a definite bent towards precision applications. The approach proposed in this section uses the computer spreadsheet to allow the ultimate flexibility in creating and analyzing property groups.

Ashby's property maps are an effective way to visualize the characteristics of a wide variety of materials and classes of materials. Figure 2-31 shows one example from at least 18 such charts that cover the conceivable spectrum of useful properties. The properties in this particular chart were shown in the previous section to be important to thermal management. The dashed lines that appear on each chart are level curves for one or more property groups. The lines are straight because the chart is log-log scale and the property group always consists of products and quotients. This concept is an elegant example of traditional engineering, but it is not very convenient if the properties and groups of interest do not appear on the same (or any) chart. This points to the main advantage of a computer-spreadsheet approach, its flexibility. Properties may be grouped as necessary or groups may be combined or averaged in some way that represents the tradeoffs in the design. It is simple and convenient to sort the whole list of materials by the most pertinent property or group so that the best choices are obvious. The only problem is getting a spreadsheet of properties and groups. An example spreadsheet appears in Table 2-5 and Table 2-6.

The key idea is the property group; the property map and spreadsheet are just good ways to implement the idea. Ashby uses a performance-index approach to formulate the property group for a particular design problem. His procedure appears below.

a) Identify the *attribute* to be maximized or minimized (weight, cost, energy, stiffness, strength, safety, environmental damage, etc.).

b) Develop an equation for this attribute in terms of the functional requirements, the geometry and the material properties (the *objective function*).

c) Identify the *free* (unspecified) *variables*.

d) Identify the *constraints*; rank them in order of importance.

e) Develop *equations* for the constraints (no yield, no fracture, no buckling, maximum heat capacity, cost below target, etc.).

f) *Substitute* for the free variables from the constraints into the objective function.

g) *Group the variables* into three groups: functional requirements, F, geometry, G, and material properties, M, thus: ATTRIBUTE $\leq f(F,G,M)$

h) *Read off* the performance index, expressed as a quantity M, to be maximized.

i) *Note* that a full solution is not necessary in order to identify the material property group.

Ashby presents a number of case studies where the objective function is either separable, that is, $f = f_1(F)\, f_2(G)\, f_3(M)$, or nonseparable. The material selection for a separable case does not depend on the geometry or the values of the functional requirements.

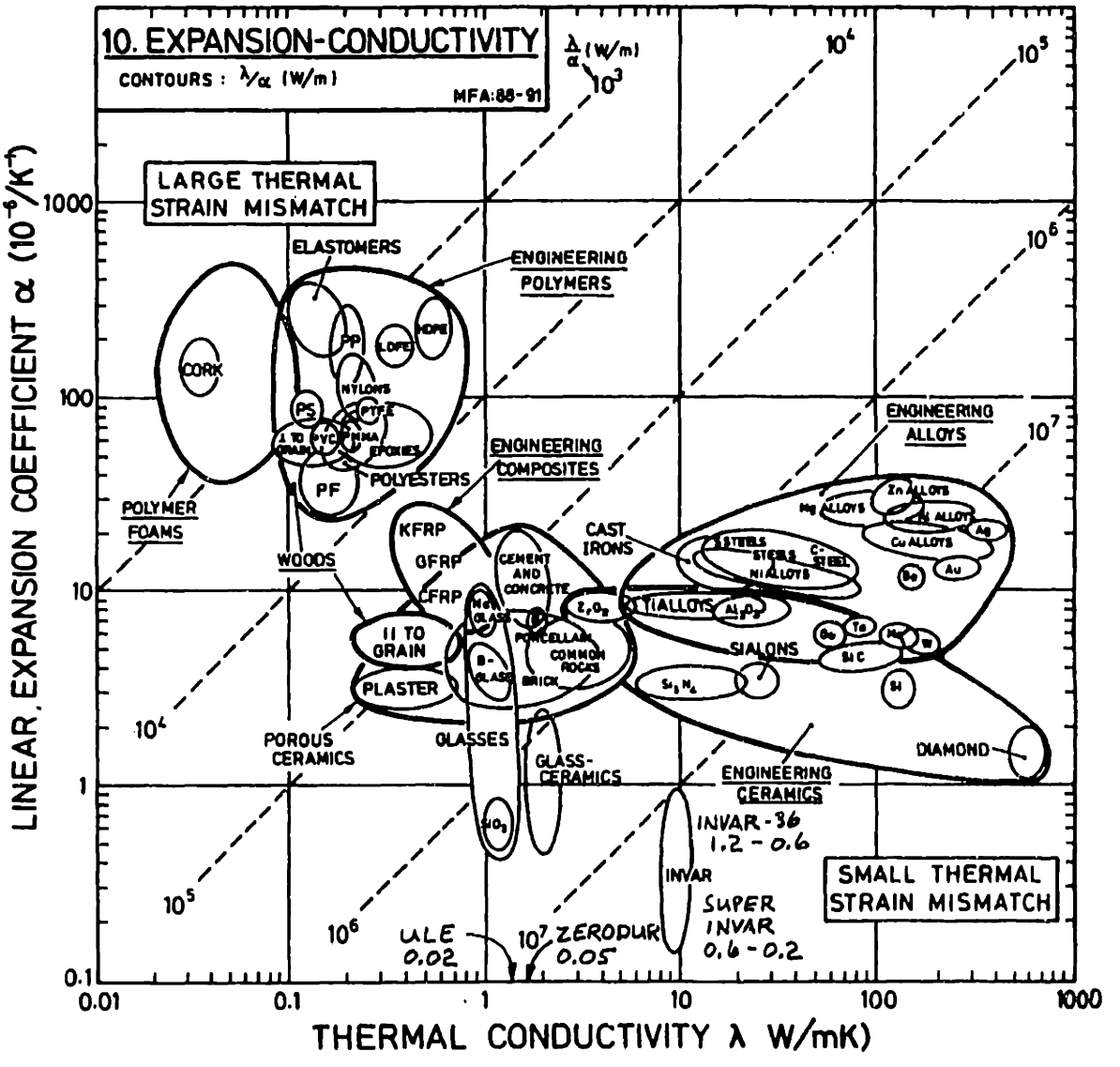


Figure 2-31 The thermal expansion coefficient plotted versus thermal conductivity, from [Ashby, 1992].

An example will serve to explain the steps in the procedure. Ashby presents an example of a large telescope mirror where the objective is to minimize the weight, thereby minimizing the cost of the support system. The problem is separable if the choice of material does not influence the basic construction of the mirror. In this example the mirror is a solid disk, which is not very realistic for a large telescope. It is, however, very realistic for most other optical systems. Rather than minimizing weight, we will minimize gravity-induced deflection (or sag). This completes step (a) in the procedure and identifies two important characteristics in this example. Equation 2.37 shows relationships for deflection δ and weight W in terms of physical dimensions and material properties.

Created by Layton Hale Last modified: 12/7/98		ρ Mass Density	E Elastic Modulus	ν Poisson Ratio	St Tensile Strength	Sc Compres. Strength	K Fracture Toughness	α Thermal Expansion	k Thermal Conduct.	C Specific Heat
SI Units ->		Mg/m^3	GPa	-	MPa	MPa	MPa √m	με/°C	W/m°C	kJ/kg/°C
U.S. Customary ->		lbm/in^3	Mpsi	-	ksi	ksi	ksi √in	με/°F	Btu/h/ft/F	Btu/lbm/F
Enter SI or US ->		SI	SI	-	SI	SI	SI	SI	SI	SI
Conversion factor ->		1	1	-	1	1	1	1	1	1
Aluminum 6061-T6	Al	2.710	71.00	0.33	310.00	310.00	23.00	23.00	177.00	0.896
Aluminum Oxide	AlO	3.900	380.00	0.22	300.00	3000.00	4.00	8.50	25.00	0.790
Beryllium I-70A	Be	1.850	304.00	0.06	350.00	500.00	4.00	11.60	180.00	1.900
Brass C36000	Bra	8.400	105.00	0.30	350.00	350.00	100.00	20.00	120.00	0.380
Bronze	Bro	8.400	120.00	0.30	350.00	350.00	100.00	19.00	85.00	0.380
Cast Iron (Gr 50)	CI	7.400	125.00	0.28	345.00	930.00	10.00	11.00	46.00	0.525
Copper	Cu	8.900	117.00	0.34	220.00	220.00	100.00	17.00	397.00	0.380
Copper-Beryllium	CuB	8.400	130.00	0.29	1000.00	1000.00		17.00	100.00	0.380
Epoxy	EP	1.300	3.80	0.30	60.00	60.00	1.00	80.00	0.30	1.900
Fused Silica	SiO2	2.200	72.00	0.17	60.00	1000.00	0.70	5.60	1.40	0.741
Granite (natural)	Gran	2.600	76.00	0.10	20.00	200.00	1.00	6.00	1.60	0.820
Invar-36	I-36	8.030	148.00	0.29	250.00	250.00	100.00	1.20	11.00	0.460
Magnesium AZ31B	Mg	1.850	45.00	0.35	250.00	250.00	15.00	26.00	76.00	1.000
Molybdenum	Mo	10.200	320.00	0.32	1000.00	1000.00		5.00	146.00	0.260
Nickel	Ni	8.900	210.00	0.36	350.00	350.00	100.00	13.00	86.00	0.450
Optical Glass (typ.)	OG	2.530	80.70	0.25	50.00	1000.00	0.80	7.10	1.12	0.879
PMMA	PM	1.200	3.30	0.30	85.00	85.00	1.40	70.00	0.20	1.500
Polyamide 6 (Nylon)	PA	1.150	2.60	0.30	80.00	80.00	4.00	80.00	0.25	1.900
Polyamide-imide	PAI	1.420	5.00	0.30	190.00	220.00		31.00	0.25	1.900
Polycarbonate	PC	1.200	2.20	0.30	70.00	80.00	2.00	120.00	0.20	1.900
Polyethylene HD	PE	0.950	1.00	0.30	35.00	35.00	2.00	120.00	0.50	2.100
Polyimide (Vespel)	PI	1.900	3.00	0.30	86.00	86.00		50.00	0.35	1.130
Polystyrene	PS	1.100	3.20	0.30	60.00	60.00	2.00	70.00	0.43	1.400
PTFE (Teflon)	PT	2.200	3.50	0.30	23.00	23.00	2.00	100.00	0.25	1.050
Silicon (single crys.)	Si	2.330	130.00	0.18	200.00	500.00	1.00	2.30	148.00	0.750
Silicon Carbide	SiC	3.200	410.00	0.20	300.00	2000.00	3.00	4.30	84.00	1.400
Silicon Nitride	SiN	3.200	310.00	0.27	550.00	1200.00	4.00	3.20	17.00	0.630
Stainless Steel 304	SSt	8.000	193.00	0.31	500.00	500.00	100.00	17.30	16.20	0.500
Steel (alloy 300HB)	St	7.800	210.00	0.29	1000.00	1000.00	100.00	12.00	54.00	0.460
Super Invar	SI	8.150	144.00	0.23	250.00	250.00	100.00	0.50	11.00	0.460
Titanium-6AL-4V	Ti	4.400	115.00	0.33	1000.00	1000.00	80.00	10.00	7.20	0.565
Tungsten	W	19.200	410.00	0.31	1350.00	1350.00		4.55	167.00	0.150
Tungsten Carbide	WC	14.500	550.00	0.24	2000.00	5000.00	15.00	5.10	108.00	0.230
Ultra low exp. glass	ULE	2.210	67.60	0.17	50.00	1000.00	0.70	0.02	1.30	0.767
Zerodure	ZD	2.530	91.00	0.24	80.00	1000.00	0.70	0.05	1.60	0.821
Zink-Alum. (ZA-12)	ZA	6.000	83.00	0.30	300.00	300.00		27.00	115.00	0.390
Zirconium Oxide	ZO	5.600	200.00	0.28	300.00	2000.00	8.00	10.50	1.50	0.670

Table 2-5 These material properties were compiled from a number of references. When references were inconsistent, I tried to choose the more reliable one or otherwise picked a median value. The strength of a material may very widely depending on its processing and/or heat treatment. I tried to choose typical values.

Created by Easton Hale Last modified 12/2/08		E/ρ	St/ρ	St/E	St^2/(ρE)	E^(1/3)/ρ	ρ C/k	α/k	α/(ρC)	αk/(ρC)
		Specific Modulus	Specific Strength	Flexible hinge	Low mass spring	Min. sag optic	Time to diffuse	Exp. due to heat	Attenuate by convec.	Attenuate by conduc.
SI Units ->										
U.S. Customary ->										
Enter SI or US ->										
Conversion factor ->										
Aluminum 6061-T6	Al	2.6E+01	1.1E+02	4.4E+00	5.0E+02	1.5E+00	1.4E-02	1.3E-01	9.5E+00	1.7E+03
Aluminum Oxide	AlO	9.7E+01	7.7E+01	7.9E-01	6.1E+01	1.9E+00	1.2E-01	3.4E-01	2.8E+00	6.9E+01
Beryllium I-70A	Be	1.6E+02	1.9E+02	1.2E+00	2.2E+02	3.6E+00	2.0E-02	6.4E-02	3.3E+00	5.9E+02
Brass C36000	Brs	1.3E+01	4.2E+01	3.3E+00	1.4E+02	5.6E-01	2.7E-02	1.7E-01	6.3E+00	7.5E+02
Bronze	Brz	1.4E+01	4.2E+01	2.9E+00	1.2E+02	5.9E-01	3.8E-02	2.2E-01	6.0E+00	5.1E+02
Cast Iron (Gr 50)	CI	1.7E+01	4.7E+01	2.8E+00	1.3E+02	6.8E-01	8.4E-02	2.4E-01	2.8E+00	1.3E+02
Copper	Cu	1.3E+01	2.5E+01	1.9E+00	4.6E+01	5.5E-01	8.5E-03	4.3E-02	5.0E+00	2.0E+03
Copper-Beryllium	CuB	1.5E+01	1.2E+02	7.7E+00	9.2E+02	6.0E-01	3.2E-02	1.7E-01	5.3E+00	5.3E+02
Epoxy	EP	2.9E+00	4.6E+01	1.6E+01	7.3E+02	1.2E+00	8.2E+00	2.7E+02	3.2E+01	9.7E+00
Fused Silica	SiO2	3.3E+01	2.7E+01	8.3E-01	2.3E+01	1.9E+00	1.2E+00	4.0E+00	3.4E+00	4.8E+00
Granite (natural)	Gran	2.9E+01	7.7E+00	2.6E-01	2.0E+00	1.6E+00	1.3E+00	3.8E+00	2.8E+00	4.5E+00
Invar-36	I-36	1.8E+01	3.1E+01	1.7E+00	5.3E+01	6.6E-01	3.4E-01	1.1E-01	3.2E-01	3.6E+00
Magnesium AZ31B	Mg	2.4E+01	1.4E+02	5.6E+00	7.5E+02	1.9E+00	2.4E-02	3.4E-01	1.4E+01	1.1E+03
Molybdenum	Mo	3.1E+01	9.8E+01	3.1E+00	3.1E+02	6.7E-01	1.8E-02	3.4E-02	1.9E+00	2.8E+02
Nickel	Ni	2.4E+01	3.9E+01	1.7E+00	6.6E+01	6.7E-01	4.7E-02	1.5E-01	3.2E+00	2.8E+02
Optical Glass (typ.)	OG	3.2E+01	2.0E+01	6.2E-01	1.2E+01	1.7E+00	2.0E+00	6.3E+00	3.2E+00	3.6E+00
PMMA	PM	2.8E+00	7.1E+01	2.6E+01	1.8E+03	1.2E+00	9.0E+00	3.5E+02	3.9E+01	7.8E+00
Polyamide 6 (Nylon)	PA	2.3E+00	7.0E+01	3.1E+01	2.1E+03	1.2E+00	8.7E+00	3.2E+02	3.7E+01	9.2E+00
Polyamide-imide	PAI	3.5E+00	1.3E+02	3.8E+01	5.1E+03	1.2E+00	1.1E+01	1.2E+02	1.1E+01	2.9E+00
Polycarbonate	PC	1.8E+00	5.8E+01	3.2E+01	1.9E+03	1.1E+00	1.1E+01	6.0E+02	5.3E+01	1.1E+01
Polyethylene HD	PE	1.1E+00	3.7E+01	3.5E+01	1.3E+03	1.1E+00	4.0E+00	2.4E+02	6.0E+01	3.0E+01
Polyimide (Vespel)	PI	1.6E+00	4.5E+01	2.9E+01	1.3E+03	7.6E-01	6.1E+00	1.4E+02	2.3E+01	8.2E+00
Polystyrene	PS	2.9E+00	5.5E+01	1.9E+01	1.0E+03	1.3E+00	3.6E+00	1.6E+02	4.5E+01	2.0E+01
PTFE (Teflon)	PT	1.6E+00	1.0E+01	6.6E+00	6.9E+01	6.9E-01	9.2E+00	4.0E+02	4.3E+01	1.1E+01
Silicon (single crys.)	Si	5.6E+01	8.6E+01	1.5E+00	1.3E+02	2.2E+00	1.2E-02	1.6E-02	1.3E+00	1.9E+02
Silicon Carbide	SiC	1.3E+02	9.4E+01	7.3E-01	6.9E+01	2.3E+00	5.3E-02	5.1E-02	9.6E-01	8.1E+01
Silicon Nitride	SiN	9.7E+01	1.7E+02	1.8E+00	3.0E+02	2.1E+00	1.2E-01	1.9E-01	1.6E+00	2.7E+01
Stainless Steel 304	SSt	2.4E+01	6.3E+01	2.6E+00	1.6E+02	7.2E-01	2.5E-01	1.1E+00	4.3E+00	7.0E+01
Steel (alloy 300HB)	St	2.7E+01	1.3E+02	4.8E+00	6.[illegible]E+02	7.6E-01	6.6E-02	2.2E-01	3.3E+00	1.8E+02
Super Invar	SI	1.8E+01	3.1E+01	1.7E+00	5.3E-01	6.4E-01	3.4E-01	4.5E-02	1.3E-01	1.5E+00
Titanium-6AL-4V	Ti	2.6E+01	2.3E+02	8.7E+00	2.0E+03	1.1E+00	3.5E-01	1.4E+00	4.0E+00	2.9E+01
Tungsten	W	2.1E+01	7.0E+01	3.3E+00	2.3E+02	3.9E-01	1.7E-02	2.7E-02	1.6E+00	2.6E+02
Tungsten Carbide	WC	3.8E+01	1.4E+02	3.6E+00	5.0E+02	5.7E-01	3.1E-02	4.7E-02	1.5E+00	1.7E+02
Ultra low exp. glass	ULE	3.1E+01	2.3E+01	7.4E-01	1.7E+01	1.8E+00	1.3E+00	1.5E-02	1.2E-02	1.5E-02
Zerodure	ZD	3.6E+01	3.2E+01	8.8E-01	2.8E+01	1.8E+00	1.3E+00	3.1E-02	2.4E-02	3.9E-02
Zink-Alum. (ZA-12)	ZA	1.4E+01	5.0E+01	3.6E+00	1.8E+02	7.3E-01	2.0E-02	2.3E-01	1.2E+01	1.3E+03
Zirconium Oxide	ZO	3.6E+01	5.4E+01	1.5E+00	8.0E+01	1.0E+00	2.5E+00	7.0E+00	2.8E+00	4.2E+00

Table 2-6 These property groups were calculated according to the relation at the top of each column. In the actual spreadsheet, these columns appear to the right of Table 2-5.

$$\delta = \frac{3}{4}\left(1-\nu^2\right)\frac{w\,r^4}{E\,t^3} \qquad W = m\,g = w\,\pi\,r^2 = \rho\,g\,t\,\pi\,r^2 \tag{2.37}$$

So far we have: developed the objective function, step (b), identified the only constraint, step (d), and developed the constraint equation, step (e). The free variable is t, the thickness, satisfying step (c). Step (f) requires the elimination of all free variables in the objective function. This is done by solving for the free variables among the constraint equations and substituting them into the objective function. Step (f) is simple in this example with just one constraint equation. Equation 2.38 shows, in addition, the result of step (g), putting variables in three groups. In group F, g is the acceleration of gravity or

some other acceleration from a mode of operation.[1] The mass m may enter as a design constraint. Group G contains only the radius r since the thickness is a free variable. Notice the strong dependence on the radius. This is why large telescopes are not made from solid disks. The last group M contains only material properties. In reading off the performance index, step (h), only the density ρ and the elastic modulus E are significant. The Poisson ratio ν does not change significantly among materials and its square is much less than one. Equation 2.39 shows M arranged as a quantity to maximize, although at times it may be more intuitive to arrange a quantity to minimize.

$$\delta(F,G,M) = \frac{3\pi^2}{4}\left(\frac{g}{m^2}\right)\left(r^8\right)\left(\left(1-\nu^2\right)\frac{\rho^3}{E}\right) \tag{2.38}$$

$$M \equiv \frac{E^{1/3}}{\rho} \tag{2.39}$$

Table 2-6 includes the property group from this example plus several other useful groups. As mentioned, it is simple to add more groups to a spreadsheet. The bigger challenge is coming up with meaningful groups for a given design problem. By following the procedure, many problems will be as simple as this example. Please refer to [Ashby, 1992] for guidance on more difficult problems or just to become better informed on materials selection.

---

[1] The secondary mirror for the Keck telescope is *chopped* for infrared viewing. The mirror tilts slightly causing an image shift between the object under study and an empty background, thus allowing subtraction of background light. This occurs at a rate of 30 times per second and follows closely an ideal square wave. This drove the engineers to use Beryllium, the most optimal material for acceleration, and to eliminate nearly all free board on the mirror (extra material left for polishing). Rather than removing the freeboard after polishing, the optical surface was diamond turned without freeboard using the LODTM. The mirror was first overcoated with electroless nickel-phosphorous alloy to present a diamond turnable surface and to avoid heath concerns with exposed Beryllium. Three mirrors have been produced to date. The secondary mirror for visible light is the more traditional polished optical glass.

3 Design Techniques

Everything should be made as simple as possible, but not simpler.

Albert Einstein

Engineering design is a broad subject in which people have very different opinions on: how to design, what distinguishes good design from bad design, whether design is more art than science and so forth.[1] It seems the harder the design problem is, the less likely two people will agree on the solution. This probably is a good thing as it leads to diversity of solutions, more thorough deliberation and usually better designs. Although our designer's pride may get in the way, peer review is an important element to a successful design. Self review of the design should occur all along the process. All aspects of a design should have a reason for being and we should be able to defend the logic or agree when it is arbitrary. Artificial constraints in our thinking (blinders we sometimes say) tend to obscure the simple and elegant design solutions that we strive to achieve. Peer and self reviews tend to uncover blinders that complicate the design or lead to unnecessary aspects.

All progress, all success springs from thinking.

Thomas Edison

In simplest terms, design is thinking. Ideas may come from nowhere or anywhere, but pulling all the ideas together into something that works requires thought more than anything else. This chapter presents techniques that help stimulate and/or give structure to our thinking. The first section is a brief exposition on conceptual design, which is based primarily on my own thoughts and experiences. It has been difficult to provide much substance beyond what may seem obvious to some. In addition, *The Design of a Design Engineer*, Section 1.4 in [Slocum, 1992] is recommended reading.

The two main topics of this chapter, *Axiomatic Design* and the *Analytic Hierarchy Process*, can be very useful in the design process, but neither really addresses the creative aspect of design. Axiomatic Design is a somewhat controversial formalism for engineering design pioneered by Prof. Nam Suh and colleagues at MIT. They have documented many successful examples of this science-based approach to design. The Analytic Hierarchy Process (AHP) is a decision making process developed by Dr. Thomas Saaty. It is used later in Chapter 9 to select the best configuration of horizontal machining center out of many possible options. A notable contribution of this thesis is a reformulation of the AHP keeping most fundamental aspects and simplifying others for easy spreadsheet manipulation. The changes are given sound mathematical justification.

---

[1] Much has been written about engineering design and no attempt is made to cover such a wide range of opinions on the subject.

3.1 Conceptual Design

How does someone go about solving a new design problem? It depends on the problem, his/her state of knowledge, available resources, various physical and business constraints, and so forth. Engineering students are taught many techniques for solving analytical problems. Sometimes a design problem is inherent in the assignment. In many cases, however, the conceptual design is already solved and only an analysis task remains, for example, compute the bending stress in a gear tooth, or synthesize the profile of a cam. These are very different problems from designing an automatic transmission or an engine valve train. Through practice and exposure to a great variety of problems and solutions, we become adept at solving new problems including design problems. This is a great ability of the human mind that we may never quite understand. Accepting our natural ability to conceive new designs from acquired information, the emphasis becomes how to cultivate our conceptual design skills and simulate our creative thinking.

3.1.1 Understanding the Problem

> The engineer's first problem in any design situation is to discover what the problem really is.
>
> unknown author

Engineers are paid to solve problems. It helps one's career and peace of mind to solve the right problems. This could apply to many situations and professions but the context here is a specific assignment or research interest requiring engineering design. For example, the assignment may come from someone in management who does not really understand the true technical problem(s) and is more interested in the cost of the solution and the delivery date. It is up to the design engineer to root out the fundamental problems so the appropriate solutions can be devised. Further, it may not be the fashionable approach within the organization (e.g., to question the way things were done before) but it is the conscientious approach. This step is the *problem definition* identified by Suh. The following ideas may help in defining the problems.

- List the goals, requirements and constraints. Answer what the consequences are for eliminating or failing to meet any one. This helps separate ones that are real and important from artificial ones.
- Become knowledgeable in the field by reviewing technical literature, products and patents. Identify the first principles that might apply to the application.
- If there are predecessors and/or competitors, answer what aspects of the designs are good or bad. Answer what can be improved. Reverse engineer the designs; answer what problems were being solved. Identify the problems that are likely to be in common with the new design.

- Trace apparent problems to more fundamental sources. For example, problems in manufacturing may result from a mistake in the design. Problems attributed to bad design may result from poor assembly procedures.
- List the challenges (technical, economic, logistics, etc.) to achieving a solution. Some of these may lie outside the design solution but not necessarily outside the process.

3.1.2 Generating Concepts

> I have been fortunate throughout my career that my first design concept is the best one.
>
> Alvis Farris

Mr. Farris, a mechanical designer at LLNL for many years, has a talent for conceiving new machine designs then creating detailed layout drawings usually on the first attempt. His quotation motivates two points about conceptual design. First, there is some truth to the statement that the first concept is frequently the best. Often the conceptual design process comes full circle back to the original concept. Perhaps the first concept is naturally simple and uncluttered by preconceptions. The stronger point is a warning. Designers who are satisfied not to look further than the first idea must be satisfied to fail against their competitors. Those who do look further must guard against preconceptions. It is all too easy to adopt artificial constraints imposed by the original concept that somehow block progress toward simpler ones. The following ideas may help in generating concepts.

- Develop an understanding of how things work and why things don't work. Consider what distinguishes good ideas from poor ones. Develop an extensive mental inventory.
- Observe the designs around us and condition the mind through practice: design forwards (how would I solve those problems) and design backwards (what problems were they trying to solve).
- Conceptual design is nothing more than a puzzle of idea pieces. Most of the ideas are readily available and the challenge is fitting them together. Often it takes just a few key ideas for the puzzle to come together.
- Avoid creating more problems when attempting to solve one problem. The best way to solve a problem is to eliminate it.[1]
- Seek out someone with which to share, exchange and stimulate ideas. Learn to use yourself in the same way. What if I did x instead of y?
- Be of the mind set *what else could be done*. Start fresh with the *real* requirements and constraints. Any idea is fair game to build a new concept (or an empire a la Bill Gates).

---

[1] Bob Edwards would say the best weldment is one with no welds.

- Design and manufacturing are intimately linked. Consider new or different manufacturing processes to stimulate ideas for design.
- Subdivide the problem and consider solutions for the simpler parts individually. Consider permutations of solutions and merge the best combinations of ideas into a few good concepts.

3.1.3 Visualization Techniques

> When I get an idea I start at once building it up in my imagination. I change the construction, make improvements and operate the device in my mind. It is absolutely immaterial to me whether I run my turbine in my thought or test it in my shop. I even note if it is out of balance.
>
> Nikola Tesla[1]

The great inventor Nikola Tesla had an exceptional ability to visualize complex problems and solutions within his mind. Those of us with lesser abilities must compensate with techniques to extend our visualization skill, which certainly is a key element to conceptual design. The use of so many figures, graphs and pictures in this thesis expresses the importance placed on visualization. The following ideas may help for better visualization.

- Use pencil and paper to note thoughts and sketch ideas; doing so will make room for more.
- The lure of precision drawings can hinder progress when roughing out concepts on computer. Instead, use the computer to draw to scale those things that probably will not change, then sketch over prints.
- Use analogies to make unfamiliar problems more familiar. Model simpler cases to obtain order of magnitude estimates and relationships among parameters.
- Exaggerate the deflections of structures as if made from rubber. Simplify motions of mechanisms about instant centers or consider equivalent mechanisms that are simpler to visualize.
- Consider variables at their possible extremes to understand the influence on the system.
- Invert the mechanism: turn it upside down and inside out, reverse the roles of parts, view it from different perspectives -- all to shake out different ways to see it function.
- Use computer models as an aid to good visualization skills, not a replacement!
- Make physical models out of simple materials such as paper, cardboard, wood, clay or straws. It not only helps you visualize but helps in explaining to others.

---

[1] From "Tesla: Man Out of Time" by Margaret Cheney, original source: Nikola Tesla "My Inventions" Electrical Experimenter, May, June, July, October, 1919.

3.2 Design Axioms

> Design may be formally *defined* as the creation of synthesized solutions in the form of products, processes or systems that satisfy perceived needs through the *mapping* between functional requirements (FRs) in the functional domain and the design parameters (DPs) of the physical domain, through the proper *selection* of DPs that satisfy FRs. ... The design axioms provide the principles that the mapping technique must satisfy to produce a good design, and offer a basis for comparing and selecting designs.
>
> [Suh, 1990]

Suh's contribution to design has been to provide an intellectual foundation as two design axioms (or principles) and numerous corollaries and theorems that are derivable from the axioms. His goal is to make design a science-based discipline with a set of "governing principles or laws that describe the underlying thought process and reduce a seemingly complex array of facts and observations into consistent and explicitly stated knowledge." Without a set of principles, design knowledge exists only as intuition and experience in the minds of *experts* who may not understand what they know or how to communicate that knowledge to others.

There are several purposes for presenting the design axioms. They state the fundamentals of good design practice and design engineers should understand them on a basic level. Just because so many successful designs exist, presumably without the use of the axioms, does not relegate them to an academic curiosity. Good designers may have instincts that are consistent with the axioms. Surely enough bad designs exist to warrant improvements in design methods. Hopefully this presentation will cause the reader to relate his/her design process to the axioms and to further investigate Suh's work. The impact on my work has been a greater awareness of the functional aspects of the design problems driving the design solutions.[1]

Suh identifies four distinct aspects of design: the *problem definition* from a notion of the needs and wants to a coherent statement of the problem; the *creative process* of proposing a physical solution to the problem; the *analytical process* of evaluating the proposed solution; and the *ultimate check* of whether the design product satisfies the original perceived needs. Suh notes that the analytical process serves as a feedback loop for the creative process and that several iterations may occur to solve a particular problem. Once that problem or set of problems is solved, the process repeats usually with a more detailed set of problems. In this way, the design process is hierarchical, beginning with the

---

[1] While Axiomatic Design has been an active research topic producing two books and numerous papers, it lacks wide-spread acceptance in the design community and seems difficult to apply except in mathematically well-defined design problems.

most fundamental issues and moving toward the details. The ultimate check is also a feedback loop but it may occur less frequently, for example, as a prototype test phase.

The problem definition clearly lies in the functional domain, that is, what the design must do. The creative process or rather the synthesis process (as creativity is important all through the design process) takes information from the functional domain and produces information in the physical domain in such forms as drawings, mathematical models, tolerances and so forth. The analytical process takes place in the physical domain but the information generated is often used back in the functional domain, either for a design iteration or in the definition of more detailed problems in the design hierarchy.

The design axioms state what the creative process must satisfy to produce a good design and how to choose the best design solution among several options.

Axiom 1 *The Independence Axiom*
Maintain the independence of functional requirements (FRs).

Axiom 2 *The Information Axiom*
Minimize the information content.

These statements may seem neither profound nor useful, but they state concisely what good designers should do. The significance of these statements should become clearer with the explanation of functional requirements and information content, but the essence of the message is to solve the design problem as simply and directly as possible.

The objectives of the design exist in the functional domain and are to be stated as a complete and independent set of functional requirements.[1] The FRs are complete and independent if all are required to fully state the objectives and none are redundant. This is the point where the synthesis process should begin. It may, however, lead to a particular design solution that cannot independently satisfy all the functional requirements. This is a coupled design that, according to the Independence Axiom, is not good. The synthesis process should continue until the design is decoupled or other uncoupled solutions are found. In addition to functional requirements, there usually will be a set of constraints that limit the possible design solutions in the physical domain. Therefore, constraints are physical limits or requirements rather than FRs. Often constraints result from design decisions made earlier in the design hierarchy.

A given design solution may be characterized by a set of design parameters (DPs). The DPs are the physical knobs in the design that determine the functional performance of

---

[1] Often a formalism can impede creative progress; therefore, the designer should create a list of the essential objectives in a manner that is intuitive and natural without particular regard to independence. Then the list can be studied to determine whether some objectives are too specific for the present level of detail and should occur later in the design hierarchy, or whether some are redundant and should be combined. The remaining objectives may then be stated as the set of functional requirements. This effort will be worthwhile if it leads to a better understanding of the objectives.

the design, and so a relationship exists through the design between the FRs and the DPs. Suh presents an interesting image of the design process as separate hierarchies of FRs and DPs. The two hierarchies are identical because they are built simultaneously starting at the beginning with the most fundamental decisions and branching out with time to the most detailed decisions. At each level in the hierarchy (where the axioms are applied) there is a design solution that relates FRs and DPs.

Suh uses the notion of information content as a measure of the knowledge required to satisfy a given FR at a given level of the FR hierarchy. The information content depends on the particular design and so it is an effective way to evaluate competing design options. A design is less expensive to produce and has less risk of failure if the knowledge required to satisfy the FRs is less than other options, according to Suh. Calculating information content is not essential to this presentation, but briefly it is inversely related to the probability of satisfying the FR (or directly related to the probability of not satisfying the FR). Suh uses the base 2 logarithm of the ratio of achievable tolerance to required tolerance, and then simply adds the results for all FRs. For comparison purposes, this is equivalent to simply multiplying the achievable FR tolerances for any given design and choosing the design with the lowest product. The design option with the minimum information content is the one with the minimum expense and/or risk.

The implications of the design axioms on good design practice will be presented in the remainder of this section using corollaries and theorems that Suh derived from the axioms. Those that are not relevant to this discussion are excluded.

Corollary 1 (Decoupling of Coupled Designs)
Decouple or separate parts or aspects of a solution if FRs are coupled or become interdependent in the designs proposed.

The relationship that exists through the design solution between the FRs and the DPs can be represented by a sensitivity matrix, where variations in FRs and DPs are mapped by the sensitivity (or design) matrix as **FR = [A] DP**. The design matrix indicates whether the design is coupled, decoupled or uncoupled. A diagonal matrix (or a square matrix with insignificant off-diagonal terms) indicates that a one-to-one correspondence exists between FRs and DPs. This is an uncoupled design. A triangular matrix (or approximately triangular) requires only a substitution process in mathematical terms to affect the FRs as desired with the DPs. This is a decoupled design. A full matrix (not necessarily square) represents a coupled design. Special cases may be decoupled through a matrix inversion process. A good example is a robot manipulator that can generate arbitrary trajectories from coordinated axis moves.

As an example, consider two FRs of a precision machine tool: that the tool moves with respect to the workpiece, and that their spatial relationship be accurately known. A traditional machine tool is a coupled design because moving axes distort the frame of the machine, which adversely affects the accuracy. If the positioning error can be determined,

then it is possible to restore accuracy by compensating the motion of the axes. This would be a decoupled design. An uncoupled design is realized by incorporating a metrology frame that is unaffected by moving axes. A metrology frame is a notable feature of the LODTM.

Corollary 2 (Minimization of FRs)
Minimize the number of FRs and constraints.

Any number of FRs greater than the minimum number makes the design problem harder to solve. Remember to use the minimum set of complete and independent FRs. In the same way, extra constraints make the design problem harder to solve.

Corollary 3 (Integration of Physical Parts)
Integrate design features in a single physical part if FRs can be independently satisfied in the proposed solution.

A similar philosophy was used in the design of the LODTM:

> Where possible, individual machine components serve a single function. This allows the component to be optimized without compromise between competing requirements.
>
> [Patterson, 1986]

Corollary 4 (Use of Standardization)
Use standardized or interchangeable parts if the use of these parts is consistent with FRs and constraints.

Corollary 5 (Use of Symmetry)
Use symmetrical shapes and/or components if they are consistent with the FRs and constraints.

Corollary 6 (Largest Tolerance)
Specify the largest allowable tolerance in stating FRs.

Standardization, symmetry and largest tolerance go towards reducing the information content in a design.

Corollary 7 (Uncoupled Design with Less Information)
Seek an uncoupled design that requires less information than coupled designs in satisfying a set of FRs.

Theorem 1 (Coupling Due to Insufficient Number of DPs)
When the number of DPs is less than the number of FRs, either a coupled design results, or the FRs cannot be satisfied.

Theorem 2 (Decoupling of Coupled Design)
When a design is coupled due to the greater number of FRs than DPs (i.e., $m > n$), it may be decoupled by the addition of new DPs so as to make the number of FRs and DPs equal

to each other, if a subset of the design matrix containing n by n elements constitutes a triangular matrix.

Theorem 3 (Redundant Design)
When there are more DPs than FRs, the design is either a redundant design or a coupled design.

Any redundancy in a design indicates that it may be simplified to reduce costs. This does not preclude the use of redundant back-up systems often required to ensure safety.

Theorem 4 (Ideal Design)
In an ideal design, the number of DPs is equal to the number of FRs.

Theorem 5 (Need for New Design)
When a given set of FRs is changed by the addition of a new FR, or substitution of one of the FRs with a new one, or by selection of a completely different set of FRs, the design solution given by the original DPs cannot satisfy the new set of FRs. Consequently, a new design solution must be sought.

Theorem 6 (Path Independency of Uncoupled Design)
The information content of an uncoupled design is independent of the sequence by which the DPs are changed to satisfy the given set of FRs.

Theorem 7 (Path Dependency of Coupled and Decoupled Design)
The information content of coupled and decoupled designs depend on the sequence by which the DPs are changed and on the specific paths of change of these DPs.

Theorem 8 (Independence and Tolerance)
A design is an uncoupled design when the designer-specified tolerance is greater than

$$\left(\sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\partial \mathrm{FR}_i}{\partial \mathrm{DP}_j} \Delta \mathrm{DP}_j \right)$$

in which case the nondiagonal elements of the design matrix can be neglected from design consideration.

Theorem 9 (Design for Manufacturability)
For a product to be manufacturable, the design matrix for the product, **[A]** (which relates the **FR** vector for the product to the **DP** vector of the product) times the design matrix for the manufacturing process, **[B]** (which relates the **DP** vector to

the **PV** vector of the manufacturing process) must yield either a diagonal or triangular matrix. Consequently, when any one of these design matrices; that is, either [**A**] or [**B**], represents a coupled design, the product cannot be manufactured.

The manufacturing process is also a design process governed by the same axioms. The DPs of the design become the FRs for the process. The DPs of the process are renamed the process variables (PVs).

Theorem 13 (Information Content of the Total System)
If each DP is probabilistically independent of other DPs, the information content of the total system is the sum of information of all individual events associated with the set of FRs that must be satisfied.

Theorem 14 (Information Content of Coupled versus Uncoupled Designs)
When the state of FRs is changed from one state to another in the functional domain, the information required for the change is greater for a coupled process than for an uncoupled process.

Theorem 15 (Design-Manufacturing Interface)
When the manufacturing system compromises the independence of the FRs of the product, either the design of the product must be modified, or a new manufacturing process must be designed and/or used to maintain the independence of the FRs of the products.

Theorem 16 (Equality of Information Content)
All information contents that are relevant to the design task are equally important regardless of their physical origin, and no weighting factor should be applied to them.

Suh does not provide an argument for Theorem 16, although it avoids having to determine weighting factors. This is a point of inconsistency with the decision making process presented in the next section where weights are allowed.

Suh's book, *Principles of Design*, contains many interesting examples of product and process design that serve to reinforce the theory. Usually the number of FRs is relatively small (3 to 5) and apply to the first level of the design hierarchy. This may indicate the best way to use the method or that the method is still immature. Nevertheless, the concepts presented should have a significant impact on the way we design and think about design.

3.3 The Analytic Hierarchy Process

> Everything we do in design has a *hierarchical* nature to it. That is, decisions must be made in order of importance by decomposing the problem into a hierarchy. ... When such a hierarchical nature of decision making is not utilized, the process of decision making becomes very complex. ... Proficient use of the hierarchy is a prerequisite for design or organizational success.
>
> [Suh, 1990]

The Analytic Hierarchy Process (AHP) is a decision making tool that was developed by Thomas Saaty and his colleagues between 1971 and 1978 [Saaty, 1980]. They applied the method to large and complex problems in government, industry and society. More recently [Slocum, 1992] and [Marsh, 1994] applied the method to design. Marsh finds the AHP to be more comprehensive than other published methods and shows it to be consistent with the Independence Axiom and the Information Axiom required by [Suh, 1990]. This indicates an often overlooked utility of the method. The hierarchy of design criteria developed for evaluation purposes is also an aid to generating conceptual designs. Just as a project schedule forces the manager to identity tasks and to organize resources, the AHP forces the designer to identify and weigh design criteria. With the criteria framework in place, the developed options are more likely to meet the goals for the design. The selection of the best option is made as objectively as possible by inter comparing the options based on the design criteria. The process may even draw attention to ways to combine the best features of the options into a superior design.

The benefits of using the AHP become more significant as the design becomes more interdisciplinary and/or interdepartmental. Such design decisions require participation from many people and the hierarchy provides a global picture while focusing attention to specific details. When the participants make a decision using the same set of criteria (as a parallel combination), Saaty recommends reaching a consensus decision, which will include the effects of the group dynamics in the decision. This can be good as it better represents participants having greater expertise or commitment or it can be bad as it also errs toward dominant personalities. As either an aid or an alternative to consensus decisions, the AHP can be extended to include the decision makers in the hierarchy. This way the opinion of each participant can be tallied much like a vote but with the possibility that the vote can carry a weight depending on the participant's department or expertise. Furthermore, some criteria may clearly lie within the domain of one authority (person or department) who decides independently from other domains (as a serial combination). Since the AHP provides a document of the decision-making process, a corporate data base would develop over time, that properly managed, would make future applications of the AHP easier and more accurate.

The AHP is simply a model of the design requirements (e.g., Suh's functional requirements and other design goals) organized as a set of hierarchical criteria with corresponding weights. The design criteria are those characteristics that an acceptable design should possess to satisfy the perceived needs for the design. These characteristics are important to consider during the design process to help guide decisions and to evaluate the merits of design options. Some design criteria may be difficult to quantify but still weigh heavily in the decision process while others may conflict and force compromise solutions. The AHP model represents these relationships, provides a format to systematically evaluate the design options and contains the algorithm to determine the best option. Developing a decision model may seem unfamiliar and unnecessary at first, but consider it a challenge to communicate your decision process to colleagues, management or posterity. If a framework on which to base a decision cannot be specified, then how can a decision be made with any confidence that it is correct?

3.3.1 A New Formulation of the AHP

Certain aspects of Saaty's AHP are cumbersome and confusing, in my opinion, and warrant simplification. This new approach uses Saaty's basic concepts of hierarchy and weighting but lacks redundancy in the intercomparisons of options or criteria. The approach by Marsh is more concordant with Saaty's but also lacks redundancy. The elaborate spreadsheet that Marsh developed for AHP gives the user a visual check of all redundant intercomparisons without actually having to make them. It is worth giving up the notion of redundancy, in my opinion, to gain a simpler, more flexible spreadsheet that the user can develop from a few simple rules. The full mathematical justification comes in the next two sections for the interested reader. This section presents the mechanics of the new AHP formulation. Section 9.3 *Selecting a Configuration of Axes* shows how the AHP was used in the conceptual design of a horizontal machining center.

The first and most difficult step in developing the AHP model is to identify and to organize the design criteria into hierarchical relationships. This step is also the most valuable because it requires up-front thought about the purpose and requirements of the design. The process begins with a problem statement about the decision, which becomes the start of an inverted tree (or root) system. The first branch may contain perhaps three to five of the most fundamental criteria. These in turn may branch to sub criteria at the next level and so on until the problem is broken down into sufficient detail to effectively evaluate the design options. As an example, Figure 3-1 shows a hypothetical hierarchy having three levels of breakdown. It just happens to have terminal criteria at all three levels indicated by (A-A-A), (A-A-B), (A-B) and (B). In a real hierarchy, the criteria would have descriptive names to make them more familiar. A spreadsheet provides an ideal environment to organize the hierarchical relationships and to implement the algorithm of the AHP model. Table 3-1 presents the same hierarchy in a spreadsheet that resembles the tree structure. The outline feature of the spreadsheet, indicated by bars at the top, accents the hierarchy as well

as being extremely useful to hide or show levels of detail. In developing a hierarchy, it is often useful to consider the problem from both top-down and bottom-up perspectives. This has a way of uncovering criteria that would not be obvious from only one perspective.

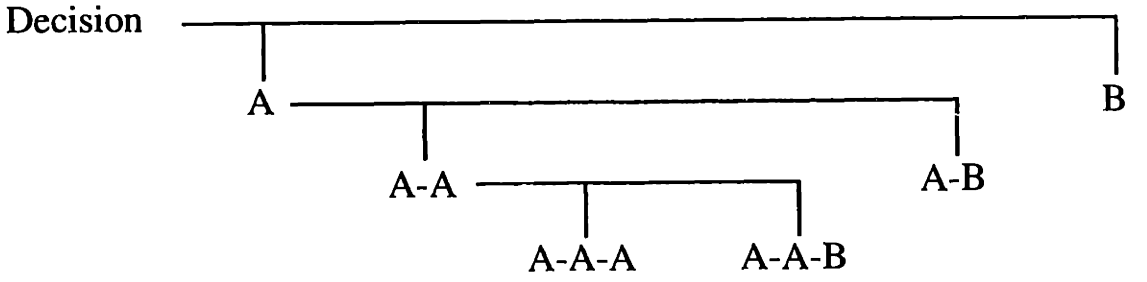


Figure 3-1 This three-level hierarchy has terminal criteria at all three levels indicated by (A-A-A), (A-A-B), (A-B) and (B).

	Decision	7					2
Criteria Level 1 >	1.00	A	5			2	B
Criteria Level 2 >		0.78	A-A	2	3	A-B	0.22
Criteria Level 3 >			0.71	A-A-A	A-A-B	0.29	
				0.40	0.60		
Design Option 1 >	4.68	4.02	3.68	5	3	5	8
Design Option 2 >	7.03	7.75	8.59	8	9	6	5
Design Option 3 >	4.65	5.27	6.60	10	5	3	3
Design Option 4 >	2.48	3.21	2.35	3	2	7	1

Table 3-1 This spreadsheet shows the AHP applied to the hierarchy in Figure 3-1. Notice that the arrangement of criteria (A, B, A-A, etc.) matches the tree structure in Figure 3-1.

The second step in developing the AHP model is to determine the relationships among the design criteria. This requires numerically ranking the criteria within each hierarchical level and then entering the numbers into the spreadsheet directly above the corresponding names. For example, Criteria Level 1 has two criteria *A* and *B* that are ranked 7 and 2, respectively. The scale is arbitrary since the ranking represents only a relative measure of significance. Thus, *A* is seven-halves more significant to the decision than *B*. Directly below each name is the normalized weight calculated simply by dividing the particular ranking number by the sum of ranking numbers across that criteria level. Thus, *A* has a weight of seven-ninths and *B* has a weight of two-ninths. This step is not absolutely necessary but leads to a more intuitive algorithm and easier interpretation of results. The basis of the ranking should be whatever is most intuitive and comfortable for the user under the circumstances. The basis could be empirical data, engineering calculations or simply judgment as is often the case.

The AHP model is now ready to use to evaluate the various design options. The design options are listed, as shown, then ranked against each other or against an ideal standard for each terminal criterion (indicated in the spreadsheet by dots over the columns). Again, the scale is arbitrary and should represent the factor by which one option is better than another. For example, the ranking based on the first terminal criterion A-A-A is entered as 5, 8, 10 and 3 indicating that the third design option is best in this particular respect. Independently for each design option, the spreadsheet combines the ranking numbers using the respective weights as exponents and multiplying these together as shown in Equation 3.1 for Design Option 2. The algorithm is easier to interpret by substituting Level 3 into Level 2 then Level 2 into Level 1. The result is a weighted geometric average of the ranking numbers given by Equation 3.2.[1] The best decision is the design option with the highest average, which is Design Option 2 in this example. In addition, it is useful to check the consistency at intermediate level criteria by comparing the ranking that the algorithm calculates to your own perception.

$$\begin{aligned} 8.59 &= 8^{0.4} \cdot 9^{0.6} && \text{Level 3} \\ \text{Design Option 2} \qquad 7.75 &= 8.59^{0.71} \cdot 6^{0.29} && \text{Level 2} \\ 7.03 &= 7.75^{0.78} \cdot 5^{0.22} && \text{Level 1} \end{aligned} \tag{3.1}$$

$$\begin{aligned} \text{Design Option 2} \qquad 7.03 &= \left[\left(8^{0.4} \cdot 9^{0.6}\right)^{0.71} \cdot 6^{0.29}\right]^{0.78} \cdot 5^{0.22} \\ &= 8^{0.22} \cdot 9^{0.33} \cdot 6^{0.22} \cdot 5^{0.22} \end{aligned} \tag{3.2}$$

A large number of design options may be considered by taking a hierarchical approach to presenting all combinations of optional technologies. Some may be dismissed immediately as being obviously poor choices for the application. The remaining options are then developed far enough to evaluate using the AHP without spending undue time on the poor ones. A clearly superior concept may not emerge out of the AHP in which case the good concepts will require further development and re-evaluation. The AHP may also be useful to track the progress of design changes.

3.3.2 The Reciprocal Matrix and Redundancy

Saaty would probably object most to the absence of redundancy in the new formulation proposed here. He recommends making pairwise comparisons, that is, comparing all

---

[1]This formulation of the AHP is trivial to code in a spreadsheet. Most of the time and effort will go into organizing the design criteria followed by ranking the design criteria and the design options. The original AHP formulation by Saaty uses a weighted arithmetic average to combine ranking numbers for each design option. This requires a normalization of the ranking numbers within criteria. Saaty chose to normalize by the sum, although normalizing by the vector length is more appropriate. Besides being easier to code, the weighted geometric average correctly represents the design option that fails to meet a criterion and is more consistent with Suh's approach for calculating information content.

combinations in the group taken two at a time. He expresses each comparison as a ratio of importance and enters each ratio into the upper (or lower) triangular half of a square matrix with its rows and columns corresponding to the items being compared. The other triangular half represents the same information but in reciprocal form, and ones occur automatically on the diagonal. Any full row or column is sufficient to represent a decision but obviously the reciprocal matrix contains more information. The redundancy in this process leads to a more accurate decision than would a simple ranking, according to Saaty. However, since the AHP distributes the decision over a number of criteria, a simple ranking may be just as effective. Regardless of this issue, the reciprocal matrix and ways to combine redundant information are useful to study (at least academically) because they help establish a mathematical basis for the new AHP formulation.

An example used by Saaty is the relative brightness of four objects identified as A, B, C and D. The results of six pairwise comparisons (hypothetical judgments rather than measurements), A to B, A to C, etc., are presented as a reciprocal matrix in Equation 3.3. In my opinion, the most intuitive way to interpret the matrix is as individual decision vectors in column form. For example in the first column, object A is the basis to compare all other objects: B is one-fifth the brightness of A, C is one-sixth of A and D is one seventh of A. Likewise in the second column, object B is the basis to compare all other objects. Since the A-to-B comparison was previously made, Saaty forces the B-to-A comparison to be consistent, that is to say, reciprocal. So then, A is five times brighter than B; however, C being one-fourth the brightness of B and D being one-sixth of B is inconsistent with the first decision vector. A consistent decision would have C being five-sixths the brightness of B and D being five-sevenths of B, but it may not be any better choice. Only one more comparison is required to fill the matrix, which is D being one-fourth the brightness of C in the third column. A choice consistent with column 1 would have D being six-sevenths the brightness of C. Reciprocity completely determines the last column. All told there are six comparisons in this example and three are redundant. In general for n items or options, there are $n(n\text{-}1)/2$ entries to the reciprocal matrix and $(n\text{-}1)(n\text{-}2)/2$ are redundant.

$$\mathbf{W} := \begin{bmatrix} 1 & 5 & 6 & 7 \\ \frac{1}{5} & 1 & 4 & 6 \\ \frac{1}{6} & \frac{1}{4} & 1 & 4 \\ \frac{1}{7} & \frac{1}{6} & \frac{1}{4} & 1 \end{bmatrix} \begin{matrix} A \\ B \\ C \\ D \end{matrix} \tag{3.3}$$

The inconsistencies between decision vectors are clearly visible in Equation 3.4, where each column is normalized to have unit length as indicated by the underline.[1] Somehow the redundant information must be reduced into a single, best decision vector. Saaty uses the principal eigenvector of the reciprocal matrix, and the result calculated in Equation 3.5 is certainly plausible when compared to the columns of $\underline{W}$. Realizing that the eigenvector may be inconvenient to calculate, Saaty also recommends two methods to approximate the eigenvector and both overcome the limitation that the matrix be square and reciprocal. Equation 3.6 shows the first method is simply a geometric average of columns. Equation 3.7 shows the second method is a sum of columns each normalized to the sum of its components. The most intuitive way to reduce the redundant information is simply to sum the columns of $\underline{W}$ giving a resultant decision vector. The sum of unit-length columns shown in Equation 3.8 is later proven to solve an optimal problem.

$$\underline{W}^{\langle j \rangle} := \frac{W^{\langle j \rangle}}{\left| W^{\langle j \rangle} \right|} \qquad \underline{W} = \begin{pmatrix} 0.959 & 0.979 & 0.824 & 0.693 \\ 0.192 & 0.196 & 0.549 & 0.594 \\ 0.16 & 0.049 & 0.137 & 0.396 \\ 0.137 & 0.033 & 0.034 & 0.099 \end{pmatrix} \begin{matrix} A \\ B \\ C \\ D \end{matrix} \tag{3.4}$$

$$\lambda := \text{eigenvals}(W) \qquad \underline{w} := \text{eigenvec}\left(W, \lambda_1\right) \qquad \underline{w} = \begin{pmatrix} 0.922 \\ 0.351 \\ 0.15 \\ 0.067 \end{pmatrix} \begin{matrix} A \\ B \\ C \\ D \end{matrix} \tag{3.5}$$

$$w := \overrightarrow{\left[\prod_j \left(W^{\langle j \rangle} \right)^{\frac{1}{n}} \right]} \qquad \underline{w} := \frac{w}{|w|} \qquad \underline{w} = \begin{pmatrix} 0.919 \\ 0.357 \\ 0.154 \\ 0.067 \end{pmatrix} \begin{matrix} A \\ B \\ C \\ D \end{matrix} \tag{3.6}$$

$$w := \sum_j \frac{W^{\langle j \rangle}}{\Sigma W^{\langle j \rangle}} \qquad \underline{w} := \frac{w}{|w|} \qquad \underline{w} = \begin{pmatrix} 0.907 \\ 0.375 \\ 0.177 \\ 0.076 \end{pmatrix} \begin{matrix} A \\ B \\ C \\ D \end{matrix} \tag{3.7}$$

---

[1] The equations in this section come directly from the Mathcad™ program used to compute this example. Usually the symbols follow general mathematical convention but some may be unfamiliar. The bracketed superscript <*j*> indicates the jth column of a matrix where *j* is a range variable from 1 to 4 in this example. The absolute value symbol |w| computes the length of the vector w. The overhead arrow is the vectorize operator. It overrides normal matrix algebra and computes a vector of scalar equations.

$$\mathbf{w} := \sum_j \frac{\mathbf{W}^{<j>}}{|\mathbf{W}^{<j>}|} \qquad \underline{\mathbf{w}} := \frac{\mathbf{w}}{|\mathbf{w}|} \qquad \underline{\mathbf{w}} = \begin{pmatrix} 0.894 \\ 0.396 \\ 0.192 \\ 0.078 \end{pmatrix} \begin{matrix} A \\ B \\ C \\ D \end{matrix} \tag{3.8}$$

The method that determines the best decision vector from the redundant set of decision vectors will minimize an error function of a logical choice. By thinking of a decision as a vector direction, clearly the error function must involve a measure of directional error. One measure is the difference between two unit-length vectors, which may be thought of as the chord length between two points on a unit hypersphere. An appropriate error function would be the sum of squared chord errors between $\underline{\mathbf{w}}$ and columns of $\underline{\mathbf{W}}$, or the rms equivalent given by Equation 3.9. A slightly better error function would incorporate the arc length between two points on the hypersphere. This is obtained in Equation 3.10 by using the arc cosine of the inner product of unit-length vectors. Table 3-2 shows the results of applying these error functions to the four reduction methods presented. For this example, the principal eigenvector and the geometric average are nearly the same but are not quite as optimal as the two methods that sum columns. The best choice is the sum of unit-length columns, which is shown next to be the optimal solution.

$$\text{error}_{\text{chord}}(\underline{\mathbf{w}}) := \left[\frac{1}{n}\left[\sum_j \left(|\underline{\mathbf{w}} - \underline{\mathbf{W}}^{<j>}|\right)^2\right]\right]^{\frac{1}{2}} \tag{3.9}$$

$$\text{error}_{\text{arc}}(\underline{\mathbf{w}}) := \left[\frac{1}{n}\cdot\left(\sum_j \text{acos}\left(|\underline{\mathbf{w}}^T\cdot\underline{\mathbf{W}}^{<j>}|\right)^2\right)\right]^{\frac{1}{2}} \tag{3.10}$$

Reduction Method	*Chord Error*	*Arc Error*	*Eq #*
Principal eigenvector	0.271	0.272	3.5
Geometric average of columns	0.269	0.270	3.6
Sum of columns each normalized by sum	0.264	0.265	3.7
Sum of unit-length columns	0.262	0.263	3.8

Table 3-2 The four methods presented to reduce redundant decision vectors into a single decision vector are compared based on chord error and arc length error.

It is quite trivial to show that the sum of unit-length columns minimizes the chord error. Minimizing the arc length error is more interesting because it involves the arc cosine function and it requires a constraint equation. Equation 3.11 shows the power series approximation for the cosine function, which provides a simpler relationship for the

squared angle between two unit-length vectors $\underline{\mathbf{w}}$ and $\underline{\mathbf{W}}^{<j>}$. As approximations go, this one is very good since the error is fourth order. For example, the error in calculating the cosine of 45° is only about 2 percent. Since the maximum value of the inner product is one, minimizing the squared angle is equivalent, within the approximation, to maximizing the inner product subject to the constraint that $\underline{\mathbf{w}}$ is unit length. Rather than minimizing the error function, we will maximize its dual function, which is the sum of columns of $\underline{\mathbf{W}}$ projected onto $\underline{\mathbf{w}}$. Equation 3.12 shows the function to optimize, which contains the dual function and the Lagrange multiplier λ times the constraint equation. The optimum occurs when the derivatives are zero, as indicated by Equations 3.13 and 3.14. After eliminating λ, the optimal solution given by Equations 3.15 is the same as Equation 3.8, the sum of unit-length columns.

$$1-\frac{\theta_j^{\,2}}{2}\cong\cos\left(\theta_j\right)\equiv\underline{\mathbf{w}}^{\mathrm{T}}\cdot\underline{\mathbf{W}}^{\langle j\rangle}\quad\Rightarrow\quad\frac{\theta_j^{\,2}}{2}\cong 1-\underline{\mathbf{w}}^{\mathrm{T}}\cdot\underline{\mathbf{W}}^{\langle j\rangle} \tag{3.11}$$

$$J(\underline{\mathbf{w}},\lambda)=\sum_j\underline{\mathbf{w}}^{\mathrm{T}}\cdot\underline{\mathbf{W}}^{\langle j\rangle}+\frac{\lambda}{2}\left(\underline{\mathbf{w}}^{\mathrm{T}}\cdot\underline{\mathbf{w}}-1\right) \tag{3.12}$$

$$\frac{\partial J}{\partial\underline{\mathbf{w}}}=\mathbf{0}=\sum_j\underline{\mathbf{W}}^{\langle j\rangle}+\lambda\,\underline{\mathbf{w}} \tag{3.13}$$

$$\frac{\partial J}{\partial\lambda}=0=\underline{\mathbf{w}}^{\mathrm{T}}\cdot\underline{\mathbf{w}}-1\quad\Rightarrow\quad|\underline{\mathbf{w}}|=1 \tag{3.14}$$

$$\underline{\mathbf{w}}=\left|\sum_j\underline{\mathbf{W}}^{\langle j\rangle}\right|^{-1}\cdot\sum_j\underline{\mathbf{W}}^{\langle j\rangle} \tag{3.15}$$

This work has implications on other aspects of the AHP besides the reciprocal matrix. The whole premise of the AHP depends on being able to split up the final decision, one that is too complicated to think about all at once, into many smaller decisions. Then they are recombined in an appropriate way to get the right decision. Another aspect is combining the input from a number of people to get the right decision. Both of these problems eliminate the principal eigenvector as a solution because the matrix is not necessarily square and reciprocal. Here the concept of the decision vector becomes very useful. The direction that the vector points represents the decision and the magnitude represents the value of that decision. For decision vectors that have equal value, like the columns of $\underline{\mathbf{W}}$, combining them either by the geometric average or by the sum of unit-length vectors has merit. Both methods are easy to extend to decision vectors that have unequal value. The other method, the sum of vectors each normalized by the sum of its components, is technically incorrect because it affects their values, although the effect is usually small.

3.3.3 Duality in the Decision Vector

A choice remains as to the method used to combine decision vectors. It seems obvious from the previous discussion that the sum of unit-length vectors is the best method. This method is intuitive and simple to extend to the AHP; the weights in the AHP scale the lengths of the decision vectors and the best decision is the sum or resultant decision vector. However, this method lacks a certain symmetry that is important to be technically correct. That symmetry is expressed in a duality principle as follows.

> It is arbitrary whether the better choice is assigned a larger value or a smaller value; rather, the ratio is the significant quantity. A decision vector and its reciprocal equally represent the same set of ratios. This implies that the method used to combine decision vectors must preserve this equality.

Comparing Equation 3.16 to Equation 3.8 demonstrates that the sum of unit-length vectors does not preserve the equality of reciprocal decision vectors. In contrast, the geometric average does preserve this equality and fortunately it extends well to the AHP. One difference is that the length of the decision vector no longer has any meaning; it simply multiplies through. Decision vectors that have unequal levels of importance are combined as a weighted geometric average, where the exponents reflect the value of the respective decisions. For example, if vector A has twice the value as vector B, then vector A should appear twice in the geometric average, which is accomplished by giving A an exponent of two-thirds and B an exponent of one-third.

$$\mathbf{w} := \sum_j \frac{\left(\mathbf{W}^T\right)^{<j>}}{\left|\left(\mathbf{W}^T\right)^{<j>}\right|} \qquad \underline{\mathbf{w}} := \frac{\overrightarrow{\mathbf{w}^{-1}}}{\left|\overrightarrow{\mathbf{w}^{-1}}\right|} \qquad \underline{\mathbf{w}} = \begin{pmatrix} 0.924 \\ 0.332 \\ 0.172 \\ 0.081 \end{pmatrix} \begin{matrix} \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \end{matrix} \tag{3.16}$$

This concludes the mathematical basis for the new AHP formulation. The geometric average, recommended by Saaty as an approximation to the principal eigenvector, was shown to be the proper method to combine decision vectors of equal importance. A simple extension to weighted exponents accommodates decision vectors that have unequal importance. None of the other reduction methods preserve the equality of reciprocal decision vectors.

intentionally blank

4 Structural Design

A fundamental problem confronted by precision machine designers is the management of compliance to yield tolerable deflection errors or conversely to provide certain degrees of flexibility to release overconstraints. While every structure has compliance, its rigidity or flexibility is relative to other components in the system. Therefore, it is important to model compliances and to direct the design toward the desired goals. Finite element analysis (FEA) is the proper tool for the design engineer to use to model a machine structure, but often FEA occurs too late in the design process to provide significant guidance. The design engineer needs good intuition and rapid modeling tools to converge to a baseline structure that is worthwhile to model and optimize using FEA. Simple models using FEA, beam theory, lumped parameters or even construction paper can be very enlightening early in the design process. This section provides information in the form of guidelines, parameter studies, modeling tools, analysis techniques and numerical examples, which will help build intuition and bring analysis earlier into the design process. This section is presented in four parts: Guidelines from Structural Mechanics, Modeling Complex Structures, Shear Panel Models and A Case Study of the Maxim Column.

4.1 Guidelines from Structural Mechanics

Structural Mechanics is a basic component of engineering education and is practiced frequently by design engineers, for example, in calculating beam deflections and stresses. A key to understanding Structural Mechanics is to be able to visualize how loads flow through a component. Then familiar formulas may be applied or new ones derived to calculate the stresses and displacements. The ability to visualize the flow of stress or strain and the resulting displacement in a structure is a key to structural design intuition. In this respect, FEA can be a great teacher and a review of strain-energy plots and mode shapes of similar structures is a good place to start a new design. Another key is understanding the relationships among the design parameters, which are expressed in the basic formulas.

Five statements from [Blanding, 1992] provide the basic ideals to strive for in achieving a stiff structure. In the quotations that follow, the terms *tension* or *compression* could be interchanged with *shear* by virtue of Mohr's circle.

Statement 1: When designing a structure for optimal stiffness, it's important that all its members (bars and plates) be used in stretching or compressing rather than bending.

Statement 2: Loads applied to a *rigid* structure result in forces being carried along the length of its members, causing the members to be loaded in *tension* or *compression*, not in bending. Loads applied to a *flexible* structure result in *bending* deflections.

Statement 3: A three-dimensionally rigid structure rigidly resists deflection by torsional loads in addition to satisfying the requirements of 1D and 2D rigidity, (e.g., a bar and a flat sheet, respectively).

Statement 4: The minimum structure required for three-dimensional rigidity is a structurally closed polyhedral shell or the bar equivalent structure.

Statement 5: Each and every face of a polyhedral shell must be two-dimensionally rigid in order for the entire structure to be three-dimensionally rigid. In other words, the polyhedral shell must be "structurally closed."

He goes on to say that a nonrigid structure becomes rigid when connected to a rigid structure, that is, its degrees of flexibility are constrained by the rigid structure. This may be a valid reason to use overconstraint in a bearing system.

The old adage *a chain is only as strong as its weakest link* applies surprisingly well to the stiffness of most structures because the members of a structure generally are in series or act in series. Models developed from series and parallel combinations of springs are so basic to the present discussion that some review is warranted. Recall that the stiffness of a spring is the force divided by the displacement and the compliance is the displacement divided by the force (the inverse of stiffness). We are most familiar with single-degree-of-freedom springs, but the same ideas also apply to springs and structures of higher dimension, where a symmetric matrix represents the linear relationship between forces and displacements for all combinations of degrees of freedom. A combination of springs may be reduced to a single equivalent spring by applying the following rules.

Rule 1: The equivalent compliance of springs connected in series is the sum of their individual compliances.

Rule 2: The equivalent stiffness of springs connected in parallel is the sum of their individual stiffnesses.

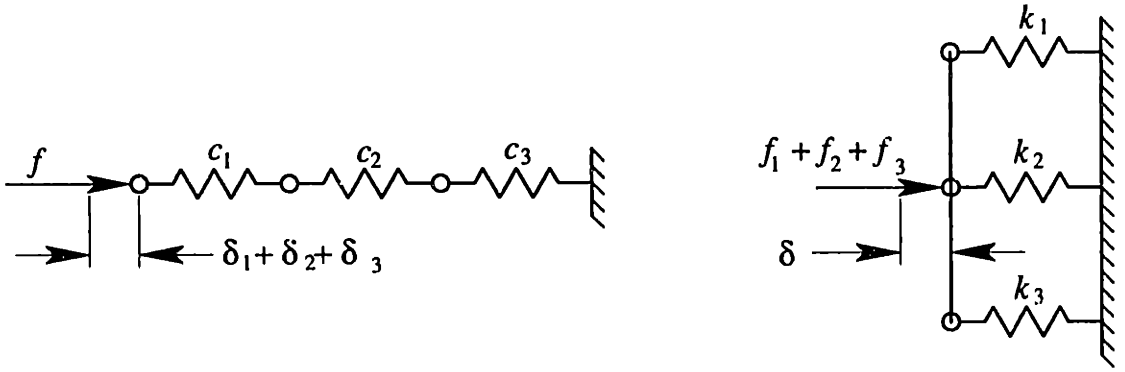
These rules are evident from Figure 4-1, where the force is the same through a series of springs and the deflections add, whereas the displacement is the same for parallel springs and the forces add. Equations 4.1 and 4.2 correspond to the three-spring combinations shown in the figure. Notice that the formula for series stiffness k_{series} or parallel compliance $c_{parallel}$ is rather difficult to remember compared to c_{series} or $k_{parallel}$, which correspond to the two rules.

$$c_{series} = \frac{\delta_1 + \delta_2 + \delta_3}{f} = c_1 + c_2 + c_3$$

$$k_{series} = \left(k_1^{-1} + k_2^{-1} + k_3^{-1}\right)^{-1} = \frac{k_1\, k_2\, k_3}{k_2\, k_3 + k_1\, k_3 + k_1\, k_2} \tag{4.1}$$

$$k_{parallel} = \frac{f_1 + f_2 + f_3}{\delta} = k_1 + k_2 + k_3$$

$$c_{parallel} = \left(c_1^{-1} + c_2^{-1} + c_3^{-1}\right)^{-1} = \frac{c_1\, c_2\, c_3}{c_2\, c_3 + c_1\, c_3 + c_1\, c_2} \tag{4.2}$$



Series Combination Parallel Combination

Figure 4-1 Spring combinations are more convenient to represent as compliances *c's* for series combinations and stiffnesses *k's* for parallel combinations, where $c = 1/k$.

Frequently, a combination of springs may be connected through a lever arm or lever ratio, which must be accounted for in the equivalent spring calculation. The force and the displacement of a particular spring must be *reflected* through the lever arm or ratio between it and the equivalent spring. Once all springs are reflected to a common reference point, then Rules 1 and 2 apply. The important point to remember is that both force and displacement are affected by the lever arm or ratio. Rule 3 expresses the consequences for a compliance or a stiffness, although the same relationship applies to inertia and damping.

Rule 3: A lever arm or lever ratio always appears as a squared term in a compliance or a stiffness.

A thin-walled box structure is a good example to demonstrate these statements and rules. Figure 4-2 represents a structurally closed box where each face is a thin plate. Individually, each plate has negligible torsional stiffness. The box carries an x-direction moment load T_x applied to two edges as a force couple. Each face of the box reacts to the load by developing a shear flow f (force per unit length) that is equal over all faces. This is more convenient than using the shear stress τ since it will vary with face thickness t. The relationship between the shear flow and the moment is simple to calculate using a cutting plane approach. Equation 4.3 shows that the shear flow is proportional to the moment and inversely proportional to the area A_1 over which the moment is applied.

$$f = \tau_i\, t_i = \frac{T_x}{2\, L_y\, L_z} = \frac{T_x}{2\, A_1} \qquad i = 1 \ldots 6 \tag{4.3}$$

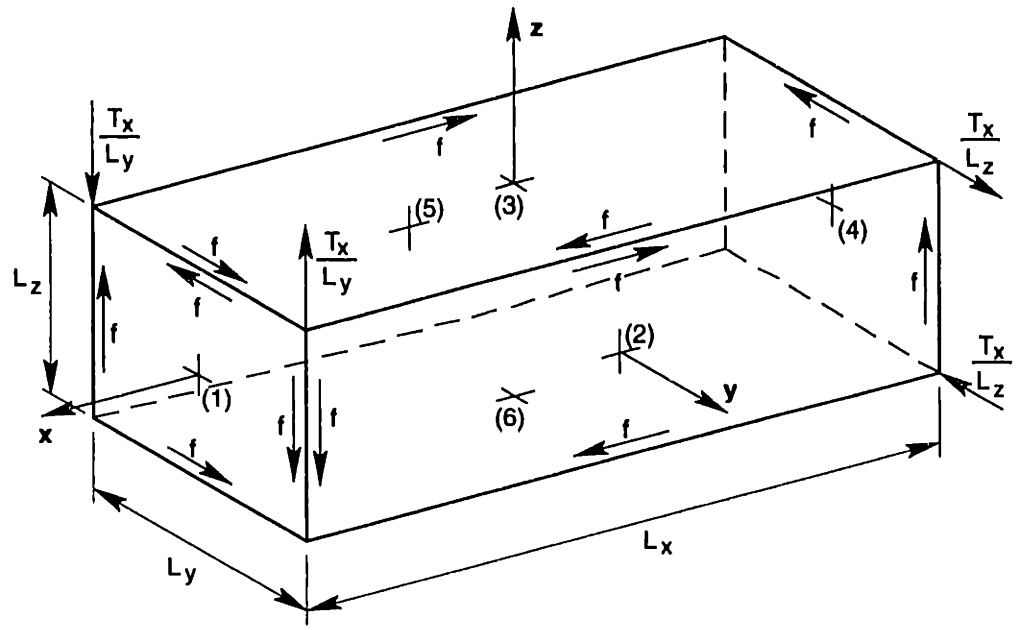


Figure 4-2 Each face of a closed box reacts to a moment load with in-plane shear stress to provide rigid behavior.

The simplest approach to calculate the deflection of the box is to apply Castigliano's second theorem to the total strain energy in the box. For a given face, the strain energy distribution is uniform if the thickness is also uniform. Equation 4.4 gives the strain energy summed over all six faces, where each face may have its own shear modulus G, uniform thickness t and area A. Equation 4.5 results from applying Castigliano and gives the torsional compliance as a summation of compliances indicating that all faces act in series. If any one face were left open, that is, $t = 0$, then the remaining faces would not carry any shear force and only the negligible torsional stiffnesses of the individual faces acting in parallel would provide any torsional stiffness.

$$u = \frac{1}{2}\left(\frac{T_x}{2A_1}\right)^2 \sum_{i=1}^{6} \frac{A_i}{G_i t_i} \tag{4.4}$$

$$c_x = \frac{\theta_x}{T_x} = \frac{1}{T_x}\frac{\partial u}{\partial T_x} = \frac{1}{4A_1^2}\sum_{i=1}^{6} \frac{A_i}{G_i t_i} \tag{4.5}$$

Usually in design we want to minimize the compliance and the mass at the same time, or equivalently, we want to maximize the stiffness-to-mass ratio known as the specific stiffness. Equation 4.6 shows the mass-compliance product for the box structure, which is equivalent to the inverse specific stiffness. The box is more mass efficient when A_1 and G_i are maximum and when A_i and ρ_i are minimum. The optimal wall thicknesses are not obvious since t_i appears in both the numerator and the denominator. If an optimum exists, it will occur when the partial derivatives are simultaneously zero as Equation 4.7 shows. Performing this step results in an equations for each side, where the common terms

are grouped as a constant. The optimal thickness t_j is a function of the material properties for face j. The constant is arbitrary meaning that all thicknesses may be scaled the same without affecting the optimum. This becomes apparent when the thicknesses are substituted back into (4.6) and the constant, which can come outside the summations, cancels leaving Equation 4.8. This analysis confirms our inclination to construct a box with equal thickness walls for a given material.[1]

$$c_x\, m = \frac{1}{4\, A_1^{\,2}} \sum_{i=1}^{6} \frac{A_i}{G_i\, t_i} \sum_{i=1}^{6} \rho_i\, t_i\, A_i \tag{4.6}$$

$$\frac{\partial}{\partial t_j}(c_x\, m) = 0 \quad \Rightarrow \quad t_j^{\,2} = \frac{const.}{G_j\, \rho_j} \qquad const. = \sum_{i=1}^{6} \rho_i\, t_i\, A_i \Big/ \sum_{i=1}^{6} \frac{A_i}{G_i\, t_i} \tag{4.7}$$

$$[c_x\, m]_{optimal} = \frac{1}{4\, A_1^{\,2}} \left[\sum_{i=1}^{6} \sqrt{\frac{\rho_i}{G_i}}\, A_i \right]^2 \tag{4.8}$$

Although Equation 4.5 was derived for a box, it is applicable to more general shapes, types and numbers of faces provided that the cross sectional area is constant along the length. Consider a closed tube of arbitrary cross section with a uniform wall thickness and shear modulus so they may be taken outside the summation. If the ends of the tube are kept from deforming, then the summation in (4.5) reduces to the wall area of the tube or simply the length L times the perimeter p_{tube}. Equation 4.9 is the result that gives the torsional compliance for an arbitrary closed-section tube. Equation 4.10 gives the mass-compliance product for an arbitrary tube (top), a rectangular tube (middle) and a circular tube (bottom). The specific shear modulus (G/ρ) is approximately 10 million (m/s)² for both aluminum and steel, which leaves the proportions of the tube as the only significant design parameters.

$$c_{tube} = \frac{A_{wall}}{4\, A_{tube}^{\,2}\, G\, t_{wall}} = \frac{L\, p_{tube}}{4\, A_{tube}^{\,2}\, G\, t_{wall}} \tag{4.9}$$

$$\begin{aligned} [c\, m]_{tube} &= \frac{\rho}{4\, G} \left(\frac{A_{wall}}{A_{tube}} \right)^2 = \frac{\rho}{4\, G} \left(\frac{L\, p_{tube}}{A_{tube}} \right)^2 \\ &= \frac{\rho}{G} \left(\frac{L}{a} + \frac{L}{b} \right)^2 \qquad a, b = \text{sides of a rectangle} \\ &= \frac{\rho}{G} \left(2\, \frac{L}{d} \right)^2 \qquad d = \text{diameter of a circle} \end{aligned} \tag{4.10}$$

---

[1] We found earlier that the shear flow is equal over all faces. If all faces have the same thickness, then the shear stress is uniform. In addition if all faces have the same shear modulus, then the strain energy density is also uniform. Strain energy density is a good indicator of how efficiently material is used in a structure.

A box structure may require openings in one or more faces for access or clearance reasons. If the box must be torsionally rigid, then the open faces must support shear loads. Diagonal braces are effective shear members but partially obstruct the opening. Often the only acceptable reinforcement is a perimeter frame that carries shear through the bending stiffness of the frame members. The shear compliance of the perimeter frame shown in Figure 4-3 may be analyzed using beam theory. Since the frame is symmetrical, only one quadrant needs to be analyzed, say, the lower right quadrant. The shear force across the horizontal beam is constant and equal to the shear flow f times $L_2/2$. The moment distribution that results is linear and is maximum at the corners as the moment diagram shows. The shear across the vertical beam is greater because $L_1 > L_2$ but the maximum moment is the same. The shear angle γ that describes the frame's distortion is calculated by treating the endpoint deflections of the two cantilever beams as rotations about their intersection. The only subtle point is how to account for the compliance of the corner joint in the lengths of the cantilever beams. The diagonal gusset stiffens the corner by transferring the tension in the inner flanges to the compression in the outer flanges. A gusset $\sqrt{2}$ times thicker than the flanges is equivalent to the inner flanges continuing on to an infinitely stiff outer corner. The webs provide additional stiffness and the best correlation to a finite element model occurred by using lengths of $L_1/2$ and $L_2/2$.

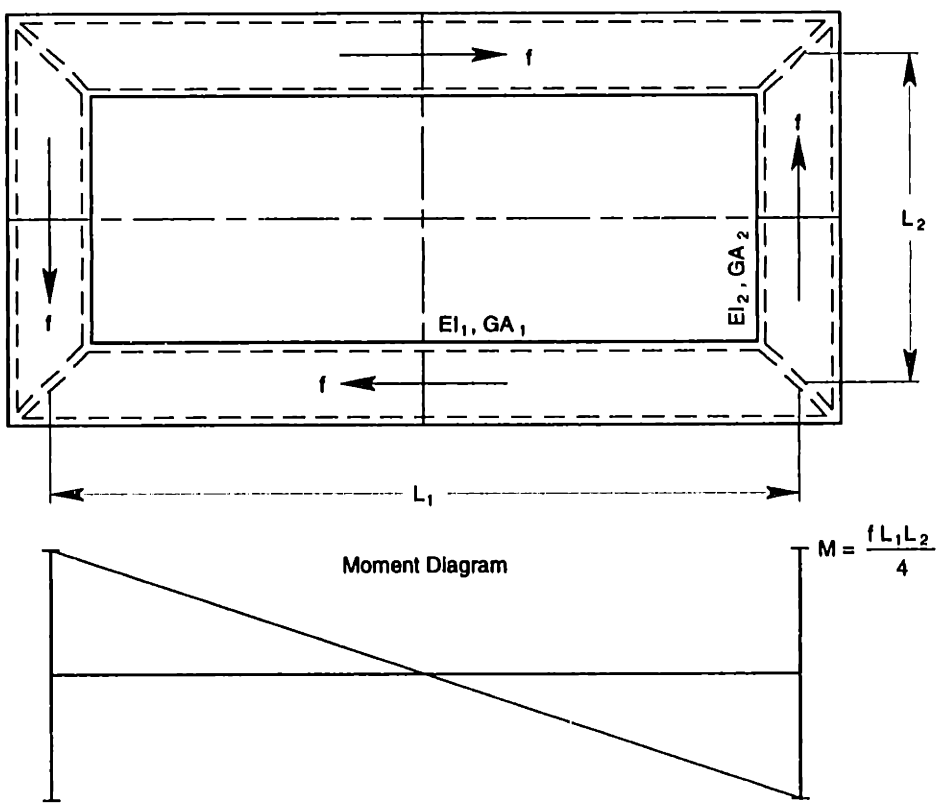


Figure 4-3 The perimeter frame supports shear loads, represented as shear flow f, by the bending stiffness of the frame members. The frame contains constant cross section beams represented by the elastic modulus E, the moment of inertia I, the shear modulus G, and the effective shear area A. The moment diagram is typical of all frame members.

Equation 4.11 gives the shear compliance and a flat plate equivalent for this perimeter frame with the shear applied at the centroid of the beams. The finite element analysis revealed that the frame is considerably stiffer when the shear force is applied to the outside walls and conversely less stiff for the inside walls. The position of the shear force changes the bending moment and the stiffness. The model becomes more complex for either case since the walls of the box tends to stiffen the perimeter frame and results in the need to model with finite elements. See Section 4.2 *Shear Panel Models* for a finite element study on the effect of various opening shapes and sizes through a shear panel.

$$\frac{\gamma}{f} = \frac{L_2}{2}\left[\frac{(L_1/2)^2}{3\,EI_1} + \frac{1}{GA_1}\right] + \frac{L_1}{2}\left[\frac{(L_2/2)^2}{3\,EI_2} + \frac{1}{GA_2}\right] = \left(\frac{1}{G\,t}\right)_{eq} \tag{4.11}$$

The previous example demonstrates that bending cannot always be avoided in preference to tension, compression and shear. A stiff beam, however, will consist of plates or shells that act largely as tension, compression or shear members. These members act as series springs so that any one could limit the equivalent stiffness of the beam. The beam deflection equations available from handbooks rarely treat the shear compliance that is in series with the bending compliance. Shear becomes significant in short beams typical of machine-tool frames and must be considered in the design. As a design aid, the following analysis demonstrates the optimal proportion of a beam cross section such as an I-beam into web area, which carries all the shear load and some of the bending moment, and flange area, which carries only bending moment but does so three times more efficiently. The analysis applies to cantilever beams and other beams that can be reduced to equivalent cantilever beams such as those shown in Figure 4-4.

The proportions of the web and flange areas may be expressed in terms of the total area A and a single, nondimensional parameter α, which will be optimized for minimum bending compliance. The area moment of inertia I may be developed from the web and flange areas and the height of the beam h. The relationship in Equation 4.12 simplifies by the introduction of α. The bending compliance of a cantilever beam is well known, and the shear compliance is calculated by assuming the shear stress is approximately constant across the web area. Equation 4.13 gives the combined bending and shear compliance normalized to the axial compliance of the beam. This eliminates the total area and all material properties except the Poisson ratio v, which does not vary significantly.

$$\begin{aligned} A_w &= \alpha \cdot A \qquad A_f = (1-\alpha)\,A \\ I &= \frac{A_w\,h^2}{12} + \frac{A_f\,h^2}{4} = \frac{A\,h^2}{12}(3-2\alpha) \end{aligned} \tag{4.12}$$

$$\frac{c_{beam}}{c_{axial}} = \left(\frac{L^3}{3\,E\,I} + \frac{L}{G\,A_w}\right)\frac{E\,A}{L} = \frac{4}{(3-2\alpha)}\left(\frac{L}{h}\right)^2 + \frac{2(1+\nu)}{\alpha} \tag{4.13}$$

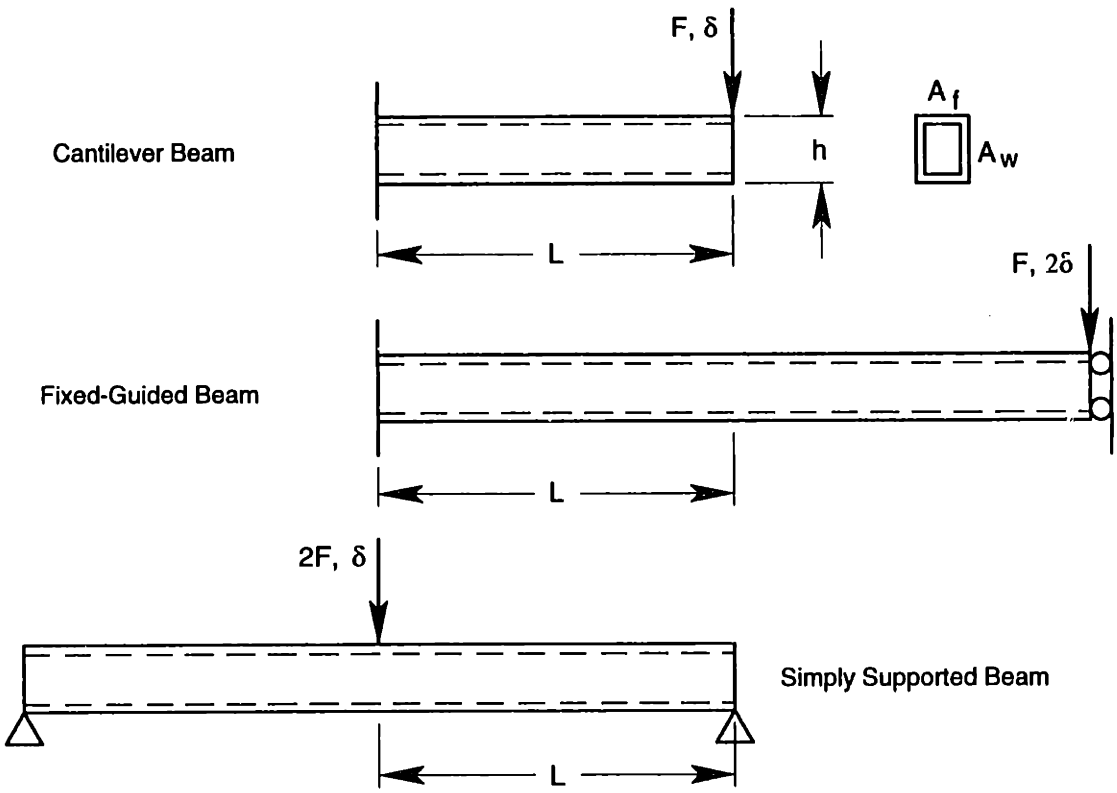


Figure 4-4 The parameters that describe the beam cross section are the height h, the flange area A_f, and the web area A_w. The relationship between force F and deflection δ for a cantilever beam may be applied to a fixed-guided beam, which is equivalent to two cantilever beams in series, or to a simply supported beam, which is equivalent to two cantilever beams in parallel.

The optimal web proportion α is found by differentiation in Equation 4.14. Figure 4-5 shows that the optimal web proportion increases with the height-to-length ratio, which indicates the increasing significance of shear. Substituting the optimal proportions back into (4.13) shows that the optimal beam compliance, given by Equation 4.15, is quadratic in the length-to-height ratio. Figure 4-6 compares the compliance of a beam with optimal web proportions to a beam with equal web and flange areas. This shows that the optimum is not very sensitive and that a square tube provides a good compromise for most applications.

$$\frac{\partial}{\partial \alpha}\left(\frac{c_{beam}}{c_{axial}}\right)=0 \quad \Rightarrow \quad \alpha_{optimal}=\frac{3\sqrt{1+\nu}}{2\left(\frac{L}{h}+\sqrt{1+\nu}\right)} \tag{4.14}$$

$$\left.\frac{c_{beam}}{c_{axial}}\right|_{optimal}=\frac{4}{3}\left(\frac{L}{h}+\sqrt{1+\nu}\right)^2 \tag{4.15}$$

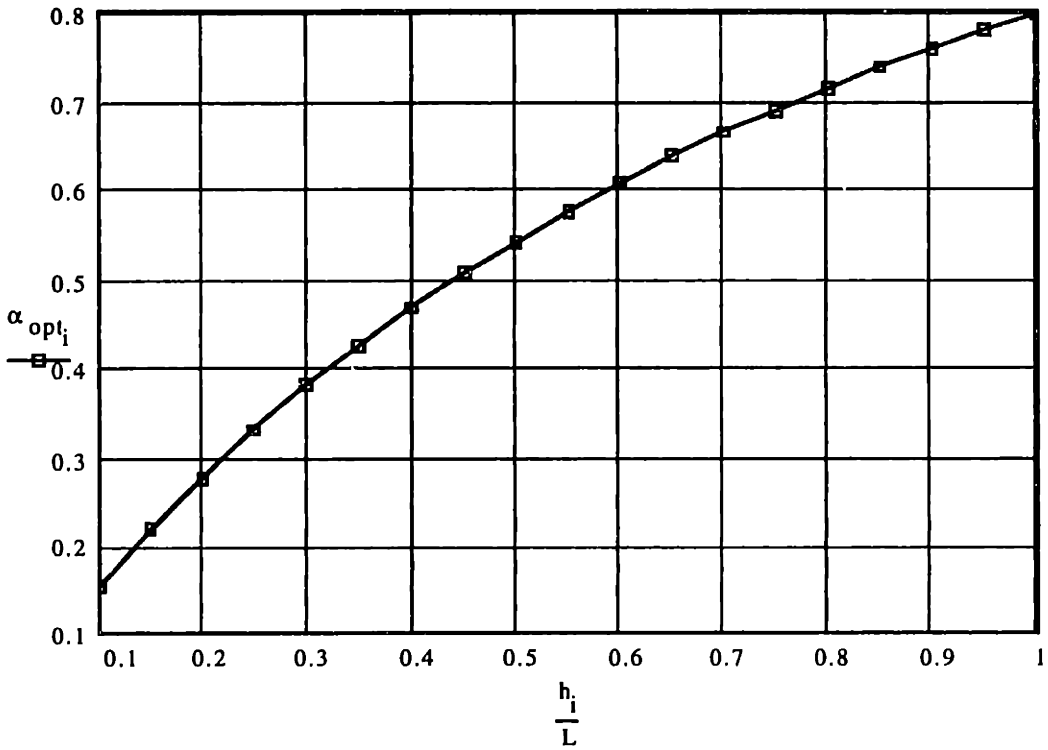


Figure 4-5 The optimal proportion of web area increases with the height-to-length ratio. This curve for the cantilever beam depends mildly on the Poisson ratio and was calculated for steel (ν = 0.29).

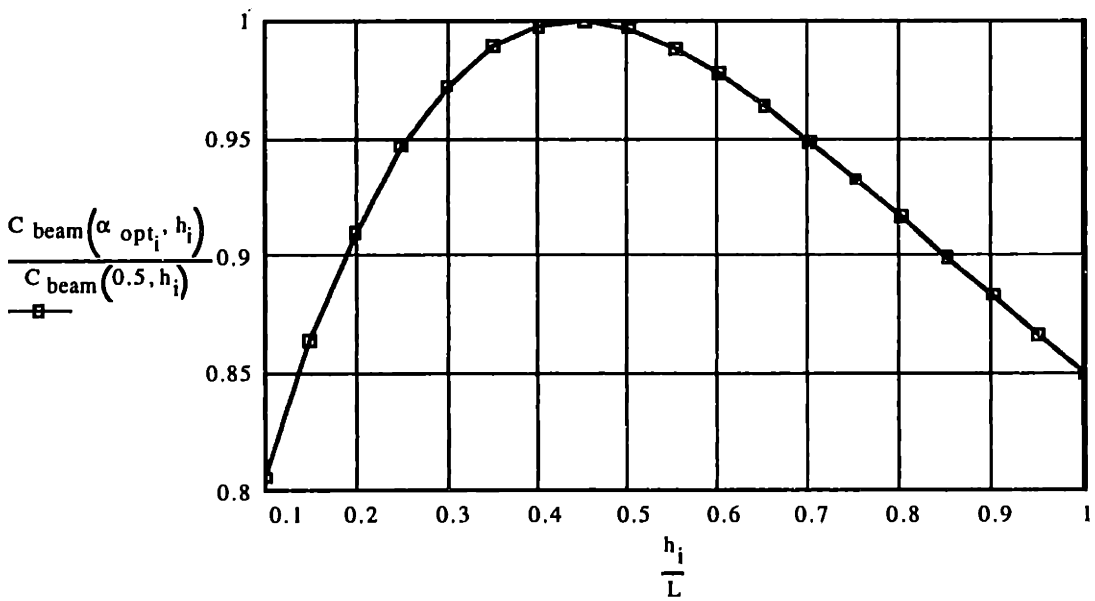


Figure 4-6 The compliance of a beam with optimal web proportions is somewhat less than a beam with equal web and flange areas. This curve was calculated for steel (ν = 0.29).

The previous analysis demonstrated that a stiff beam requires maximum separation between the flanges in tension and compression and requires an optimal shear connection between the flanges. Diagonal ribs in a truss arrangement provide a useful alternative to web-type shear members because they allow access through the beam and may be easier to

cast or fabricate in certain cases. Assuming infinitely stiff flanges in tension or compression, Figure 4-7 shows a truss and the equivalent structure obtained by unfolding the diagonal ribs into a single tension member. Equation 4.16 gives the shear compliance, where the term A_r is the rib area cut by a vertical plane. Also given is the minimum shear compliance that occurs at the optimal angle $\beta = 45°$. For steel with a Poisson ratio $\nu = 0.29$, the diagonal ribs are only 64.5% as effective in shear as the same material used as a web. In addition, the ribs do not contribute appreciable bending stiffness as does a web. For these reasons, the optimal proportions of the rib and flange areas will be slightly different from the previous analysis.

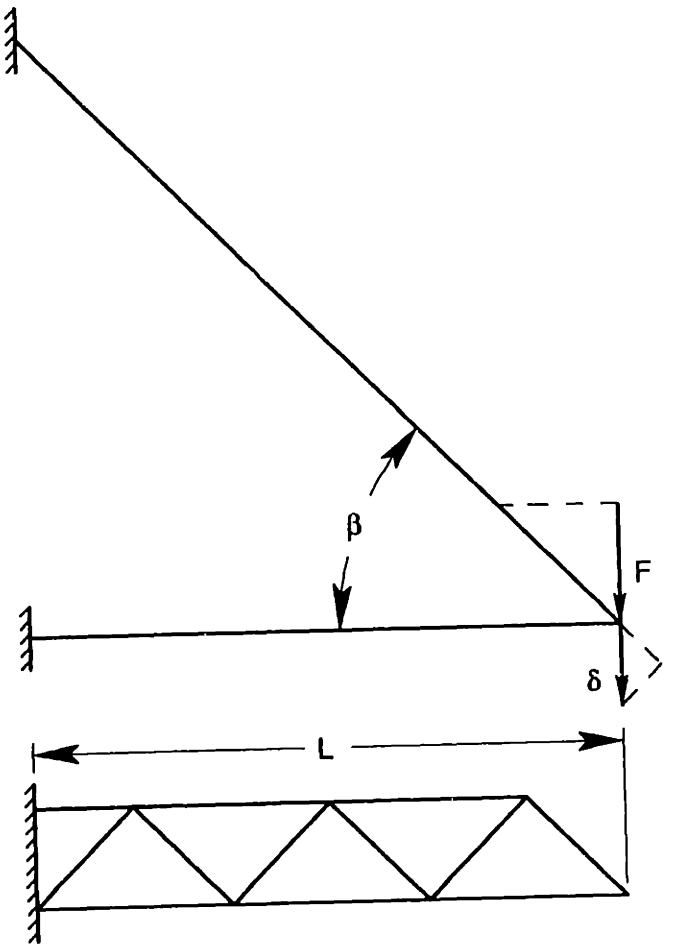


Figure 4-7 The shear compliance of the truss (below) is perhaps easier to visualize as a single diagonal tension member (above).

$$c_{shear} = \frac{\delta}{F} = \frac{L/\cos(\beta)}{E\,A_r \cos(\beta)} \frac{1}{\sin^2(\beta)} \qquad (4.16)$$

$$c_{shear}(45°) = \frac{4\,L}{E\,A_r} = \frac{2}{1+\nu} \frac{L}{G\,A_r}$$

The analysis of the optimal proportions for the rib and flange areas of a truss follows closely the previous analysis for the web with minor changes to account for the

reduced effectiveness of the ribs. To represents the mass consistently, the combined bending and shear compliance is normalized to the axial compliance calculated using the total area rather than just the flanges. As a result, the normalized compliance for the truss beam is consistently greater than for the webbed beam. For a cantilever beam with $L/h = 2$, the truss beam would be 1.42 times more compliant than a webbed beam of identical mass.

$$A_r = \alpha \cdot A \qquad A_f = (1-\alpha)\,A$$
$$I = \frac{A_f\, h^2}{4} = \frac{A\, h^2}{4}(1-\alpha) \tag{4.17}$$

$$\frac{c_{beam}}{c_{axial}} = \left(\frac{L^3}{3\,E\,I} + \frac{4\,L}{E\,A_r}\right)\frac{E\,A}{L} = \frac{4}{3(1-\alpha)}\left(\frac{L}{h}\right)^2 + \frac{4}{\alpha} \tag{4.18}$$

$$\frac{\partial}{\partial\alpha}\left(\frac{c_{beam}}{c_{axial}}\right) = 0 \;\Rightarrow\; \alpha_{optimal} = \frac{\sqrt{3}}{\frac{L}{h} + \sqrt{3}} \tag{4.19}$$

$$\left.\frac{c_{beam}}{c_{axial}}\right|_{optimal} = \frac{4}{3}\left(\frac{L}{h} + \sqrt{3}\right)^2 \tag{4.20}$$

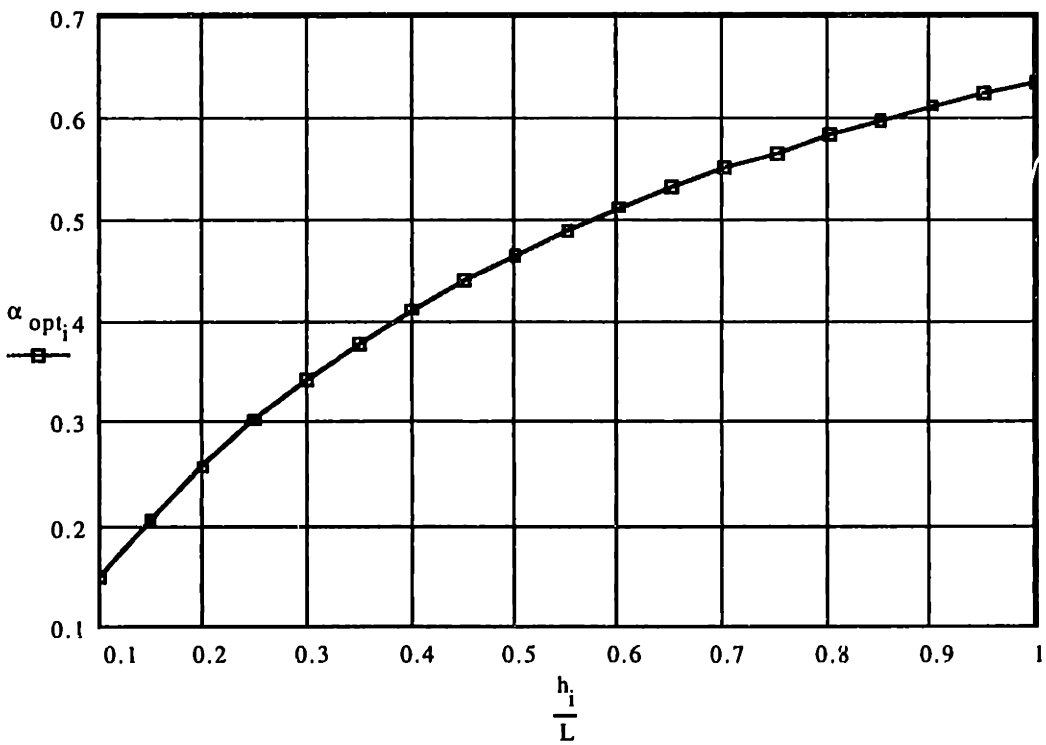


Figure 4-8 The optimal proportion of rib area increases with the height-to-length ratio. This curve for the cantilever truss beam has no material dependence.

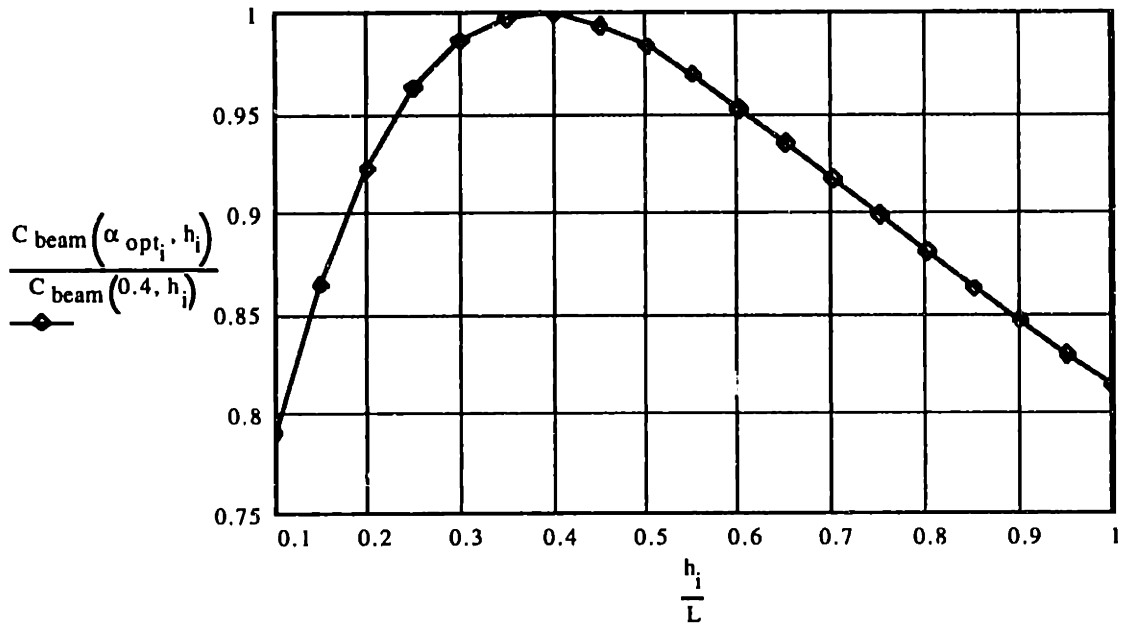


Figure 4-9 The compliance of a beam with optimal rib-flange proportions is somewhat less than a beam with a rib area proportion of 40%. This curve has no material dependence.

A wide truss beam also has significant torsional stiffness since the diagonal ribs act as shear faces to *close the section.* Although the analysis involves simple theory, visualizing the loads and setting up the problem is challenging. Figure 4-10 shows one section of the truss along with its dimensional parameters. The torsional load on the truss develops a constant shear flow in the flanges as shown. Since there is no shear load on the sides of the truss, a linear moment distribution also develops in the flanges as shown by the moment diagrams. The top and bottom flanges are mirror images that are out of phase by one-half the cell length. Since the truss as a whole is loaded only in torsion, the net bending moment in the vertical direction must be zero. The rib moment M_2 may be determined from the flange moment M_1 by using the moment vector polygon. The resultant of these moments M_3 turns out to be one-half of the applied torque regardless of the rib angle β. The other half is carried by shear in the flanges as Equation 4.22 indicates. Since we will be interested in the optimal proportions of material into flanges and ribs, the sectional properties are presented in Equation 4.23 in terms of the nondimensional parameter α and the cross sectional area A.

$$M_1 = \frac{f\,a\,b}{\tan(\beta)} \qquad M_2 = \frac{f\,a\,b}{\sin(\beta)} \qquad M_3 = f\,a\,b \tag{4.21}$$

$$T = f\,a\,b + M_3 = 2\,f\,a\,b \tag{4.22}$$

$$A_r = \alpha\,A \qquad A_f = (1-\alpha)\,A \qquad f_s = 6/5$$
$$I_r = \frac{\alpha\,A\cos(\beta)\,a^2}{12} \qquad I_f = \frac{(1-\alpha)\,A\,a^2}{12} \tag{4.23}$$

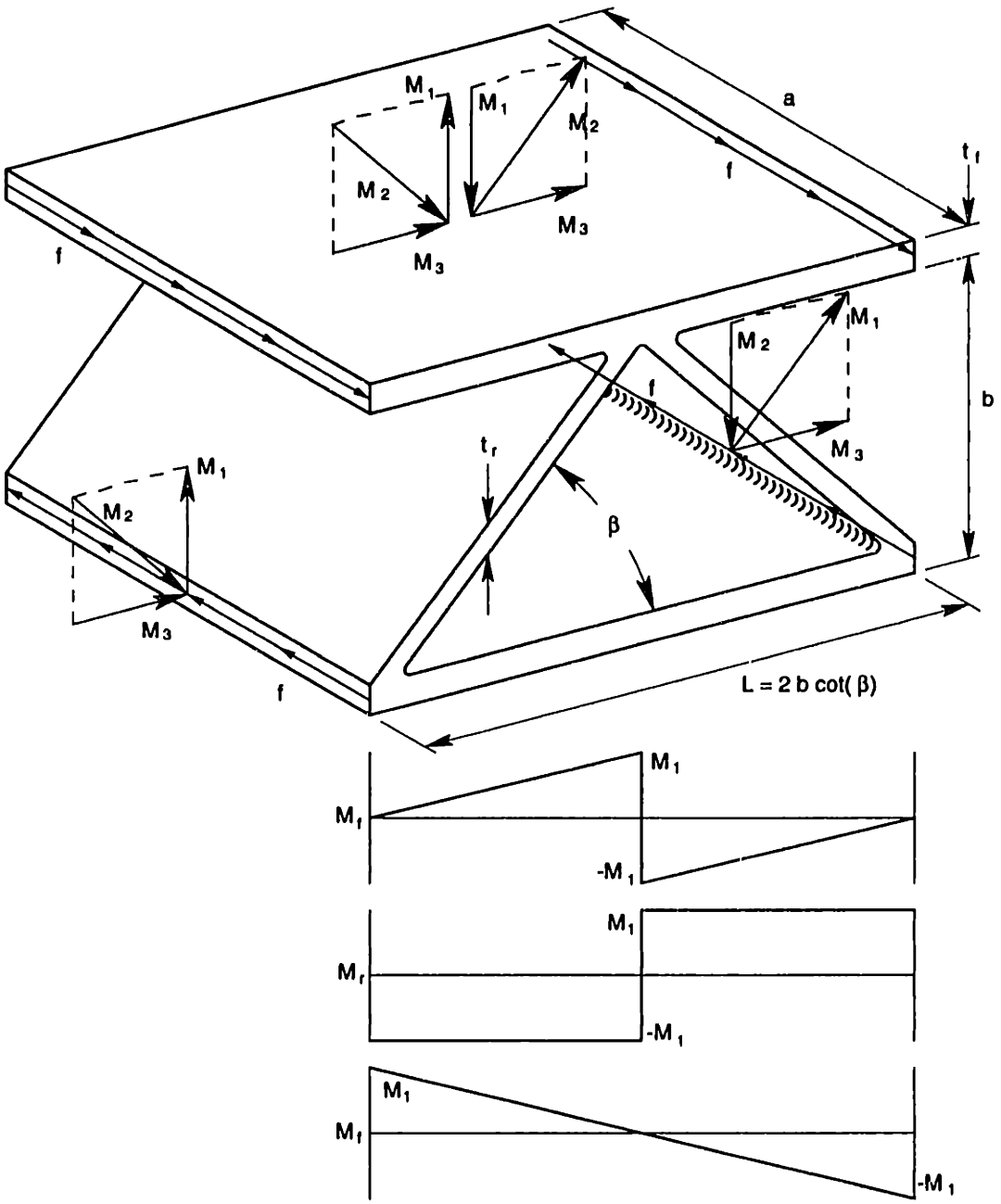


Figure 4-10 One section (or cell) of a truss that carries a torsional load (top) and the corresponding moment diagrams for the top and bottom flanges and the rib (bottom).

The torsional compliance is calculated as before by applying Castigliano's second theorem to the total strain energy. The strain energy comes from three sources, shear and bending in the flanges and bending only in the ribs. The flange area represents both flanges so that $A_f = 2\ a\ t_f$, which requires that the loads on both flanges be accounted for in the strain energy expression developed in Equation 4.24. The torsional compliance is given by Equation 4.25 in terms of a nondimensional parameter ψ, which may be optimized by adjusting α, β and the aspect ratio R. The equations for the optimal β and the optimal R depend on α and are too cumbersome to present, but instead are plotted in Figure 4-11 and

Figure 4-12. The rib angle turns out not to be very sensitive, and a good compromise is 50°. The aspect ratio plot may be used either to determine the best R given α or the best α given R. Figure 4-13 and Figure 4-14 show the effects on ψ of operating away from the optimal rib angle or aspect ratio, respectively. Figure 4-15 is perhaps the most revealing, which shows ψ for a rectangular tube and a reasonable truss design versus a range of aspect ratios. The truss is clearly most efficient as a torsion member when the aspect ratio is in the range of two to four where it is about half as good as a rectangular tube.

$$u = 2\left\{ f_s \frac{(2\,f\,a)^2}{2\,G\,A_f} \int_0^{L/2} dx + \frac{(2\,M_1)^2}{2\,E\,I_f} \int_0^{L/2} \left(\frac{x}{L/2}\right)^2 dx + \frac{M_2{}^2}{2\,E\,I_r} \int_0^{L/(2\cos(\beta))} dx \right\}$$

$$= \frac{(2\,f\,a\,b)^2\,L}{2} \left\{ \frac{f_s}{G\,A_f\,b^2} + \frac{1}{3\,E\,I_f \tan^2(\beta)} + \frac{1}{4\,E\,I_r \sin^2(\beta)\cos(\beta)} \right\} \tag{4.24}$$

$$= \frac{T^2 L}{2\,G\,A} \left\{ \frac{6/5}{(1-\alpha)\,b^2} + \frac{1}{2(1+\nu)\,a^2} \left[\frac{4}{(1-\alpha)\tan^2(\beta)} + \frac{3}{\alpha \sin^2(\beta)\cos^2(\beta)} \right] \right\}$$

$$c_{truss} = \frac{\theta}{T} = \frac{1}{T}\frac{\partial u}{\partial T} = \frac{L}{G\,A}\frac{\psi}{a\,b} \qquad R \equiv \frac{a}{b}$$

$$\psi \equiv R\frac{6/5}{(1-\alpha)} + \frac{1}{R}\frac{1}{2(1+\nu)} \left[\frac{4}{(1-\alpha)\tan^2(\beta)} + \frac{3}{\alpha \sin^2(\beta)\cos^2(\beta)} \right] \tag{4.25}$$

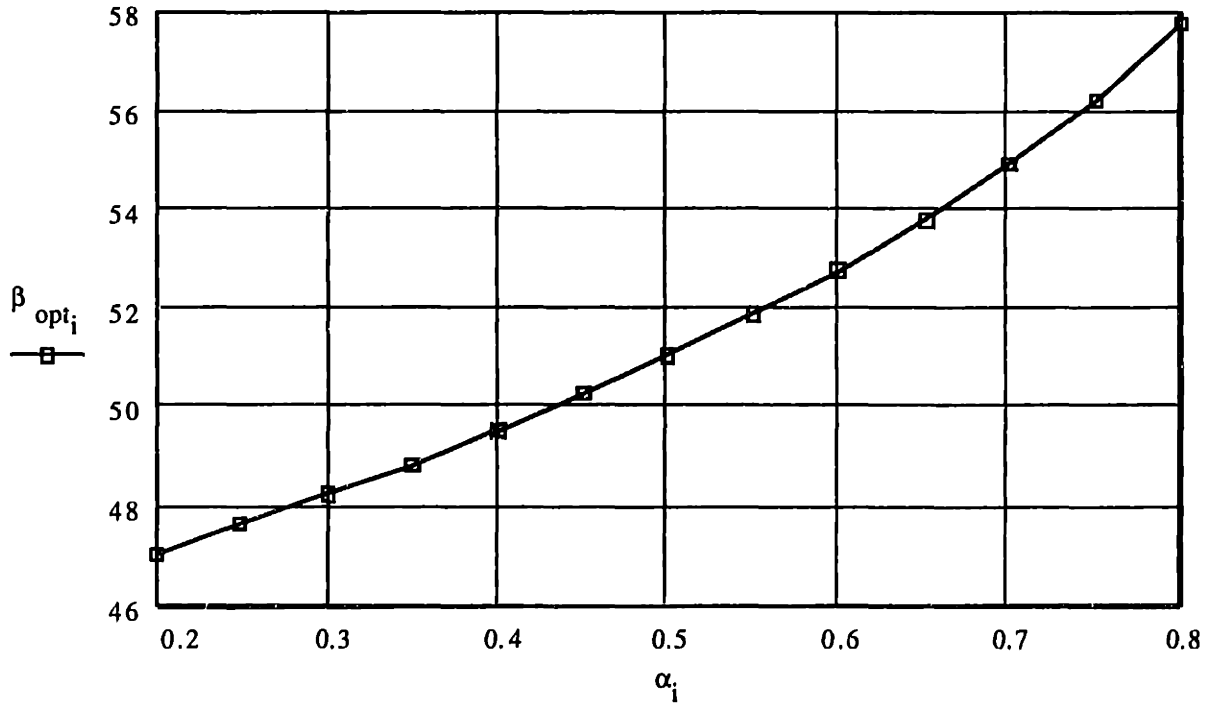


Figure 4-11 The optimal rib angle versus the rib proportion of the truss.

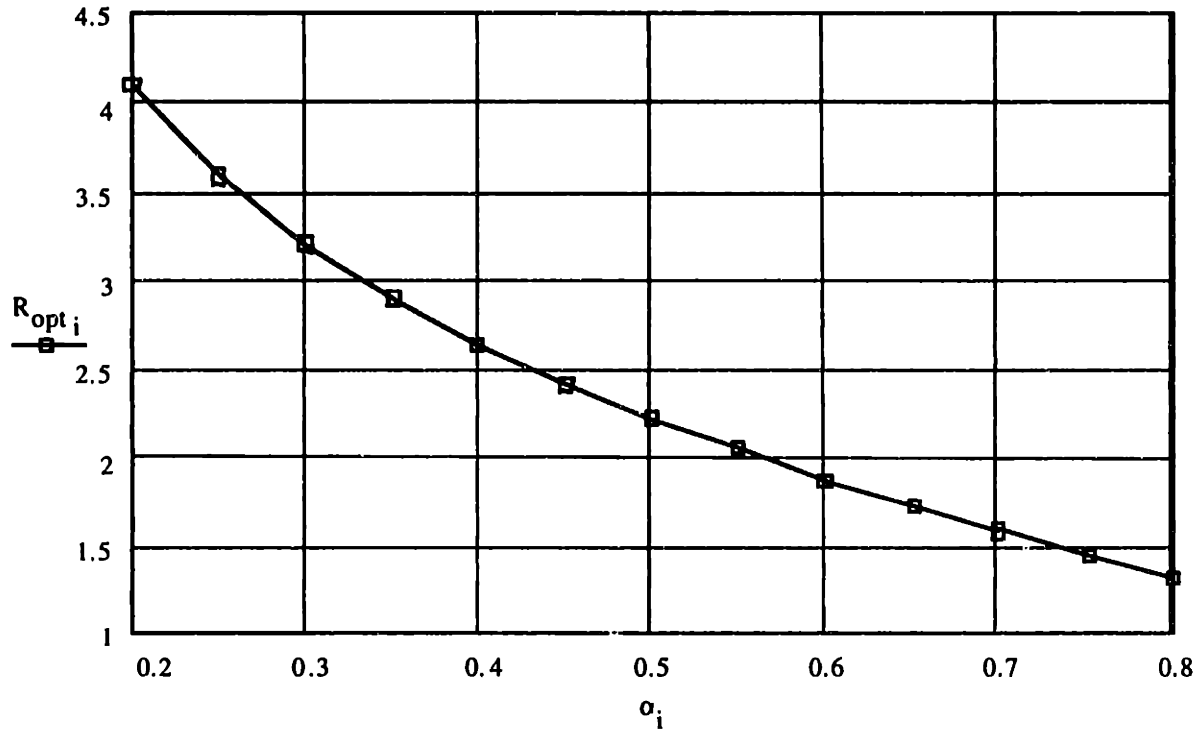


Figure 4-12 The optimal aspect ratio (*a/b*) versus the rib proportion of the truss. This curve may also be used to determine the optimal rib proportion for a given aspect ratio.

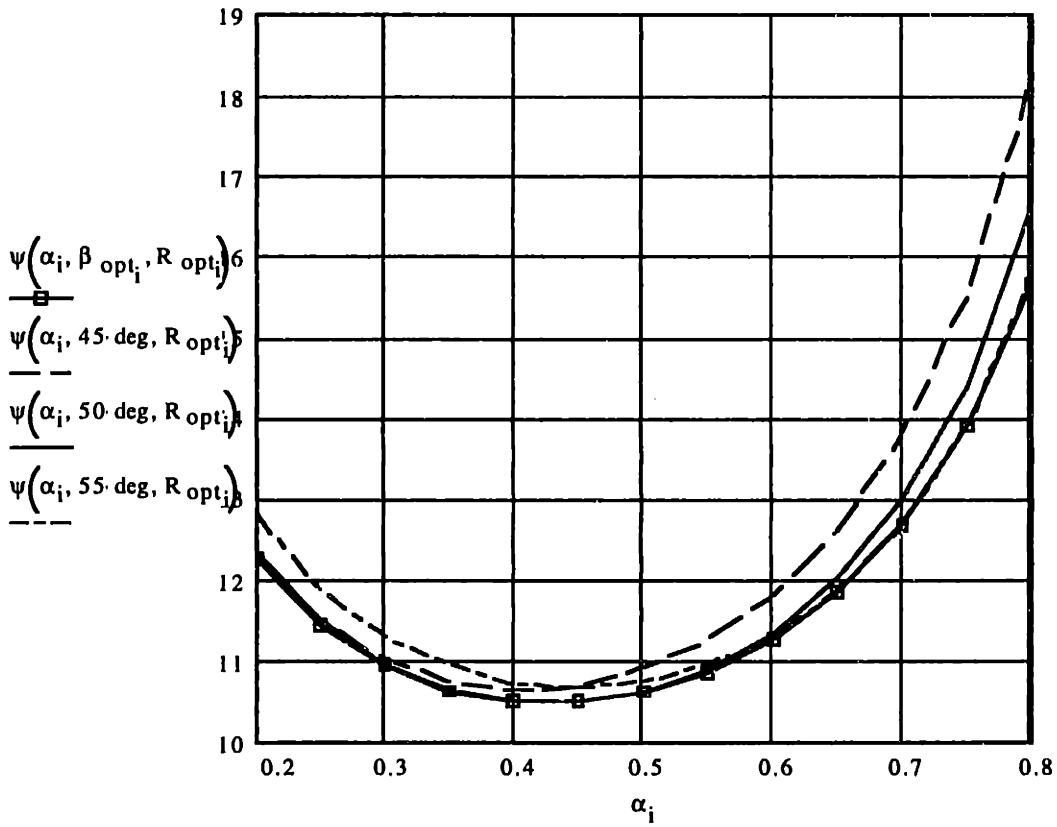


Figure 4-13 Effect of a non-optimal rib angle on the torsional compliance parameter.

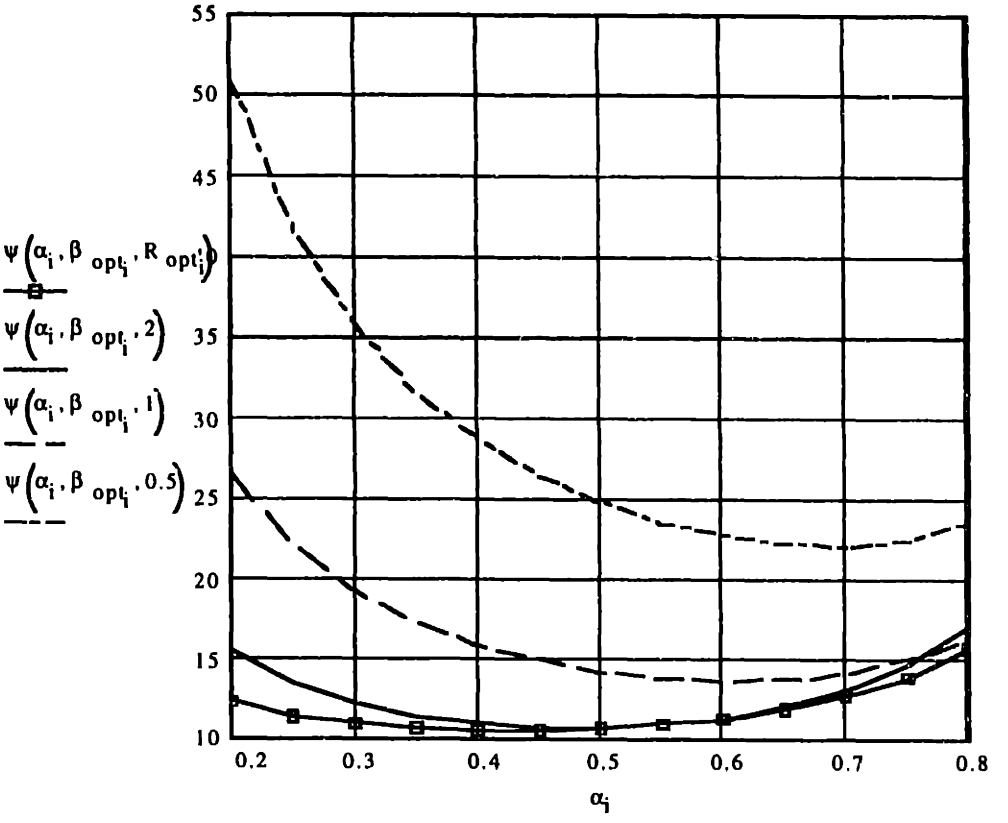


Figure 4-14 Effect of a non-optimal aspect ratio on the torsional compliance parameter.

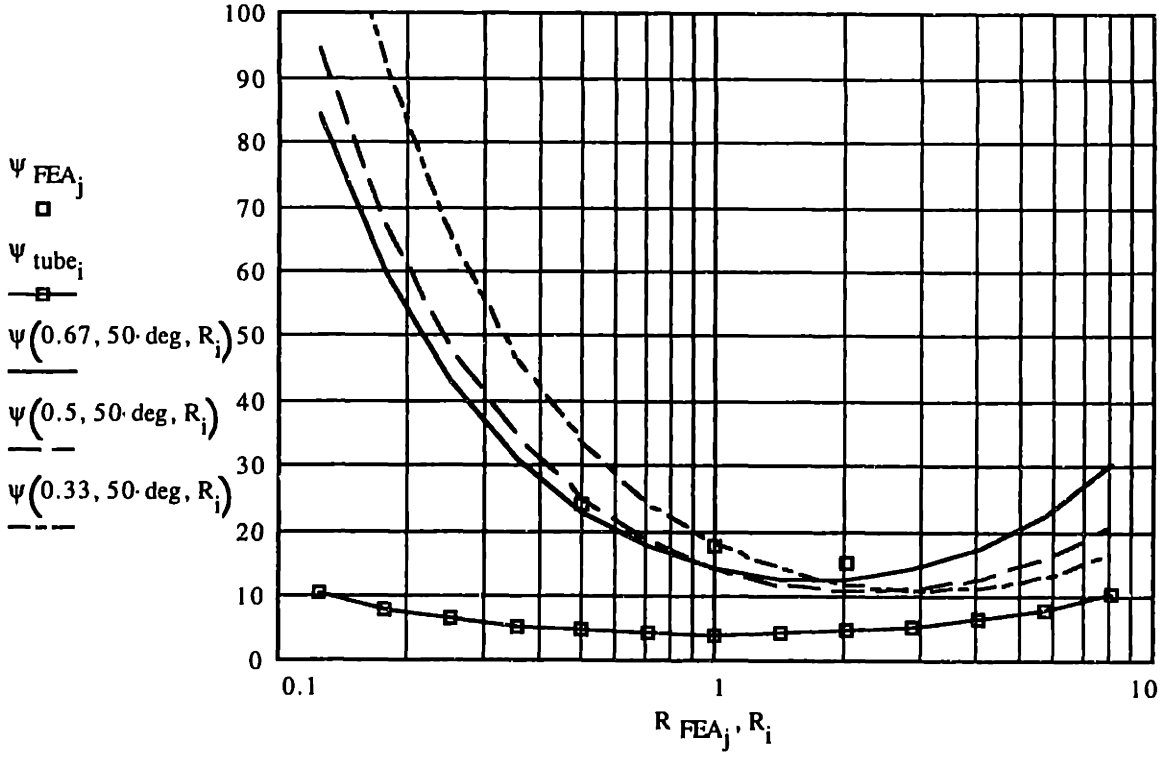


Figure 4-15 Torsional compliance parameter for a tube and a truss for several rib proportions versus the aspect ratio. The results from finite element analysis ($\alpha = 0.5$) agree reasonably well with Equation 37.

The use of normalized stiffness or compliance has been useful in setting up optimization problems and in showing how to make efficient use of the material and/or the space provided. This is particularly important for inertial loads, for example, when the structure is subject to an acceleration. In this case, the mass of the structure constitutes an inertial load and simply increasing the amount of material will not significantly reduce deflections. Most engineering materials have about the same specific modulus E/ρ, and the high-performance materials that offer four to six times higher specific modulus are generally not practical for large machine structures. Of the common materials shown in Table 4-1, steel has the highest specific modulus and cast iron has the lowest at about 65% that of steel. That leaves the designer with the placement of material as being the most significant design parameter for inertial-loaded structures.

Material	Density	Modulus	Strength	Spec. Mod.	Spec. Str'th	CTE
Properties	Mg/m^3	GPa	MPa	M(m/s)^2	K(m/s)^2	µm/m/C
1020 HR	7.9	210	400	26.58	50.63	12
6061-T6	2.71	71	275	26.20	101.48	23
Cast Iron	7.4	125	350	16.89	47.30	11
Invar 36	8.03	148	245	18.43	30.51	.64 - 1.3
Super Invar	8.15	144	245	17.67	30.06	.23 - .56
Ti-6Al4V	4.4	115	900	26.14	204.55	8.8
Mg-AZ31B	1.85	45	160	24.32	86.49	26
Be (I-70)	1.85	304	500	164.32	270.27	11.2
SiC	3.2	410	500	128.13	156.25	4.3
Si3N4	3.2	310	850	96.88	265.63	3.2
Al2O3	3.9	380	300	97.44	76.92	8.5
WC	14.5	550	2000	37.93	137.93	5.1
Granite	2.6	60	20	23.08	7.69	6

Table 4-1 Table of material properties for typical engineering and high-performance materials.

The relationship between mass, stiffness, acceleration and displacement is easiest to understand for a single degree-of-freedom mass-spring system. When the system is subject to a constant acceleration, the resulting displacement is given by Equation 4.26, which is proportional to the dynamic force $m\ a$ divided by the stiffness k. The displacement is minimum when the specific stiffness k/m or equivalently the square of natural frequency ω_n is maximum. This formula is surprisingly useful to estimate displacements or to set minimum frequency requirements for a complex structure. Figure 4-16 shows the graph of this formula relating displacement to natural frequency and acceleration.

$$\delta = \frac{f}{k} = \frac{m \cdot a}{k} = \frac{a}{\omega_n^{\,2}} \tag{4.26}$$

The approximate relationship between an acceleration induced displacement and the natural frequency is easy to verify for beams. Equation 4.27 gives the natural frequency for an Euler beam (neglecting shear effects), where λ is a nondimensional constant that depends on the end conditions and the mode number as Table 4-2 shows. The maximum

displacement of a beam loaded uniformly by an acceleration may be calculated *exactly* using a standard beam deflection equation. Together with (4.26), the beam deflection equation may be cast in the form of (4.27) to show that agreement is reasonable for a distributed mass structure, which is evident by comparing the λ's from the Uniform Load column to the First Mode column. The premise is that (4.26) applies approximately to structures more complex than beams and that the square of natural frequency is the correct figure of merit for more general structures under inertial loads.

$$\omega_n = \left(\frac{\lambda}{L}\right)^2 \sqrt{\frac{E\,I}{\rho\,A}} \tag{4.27}$$

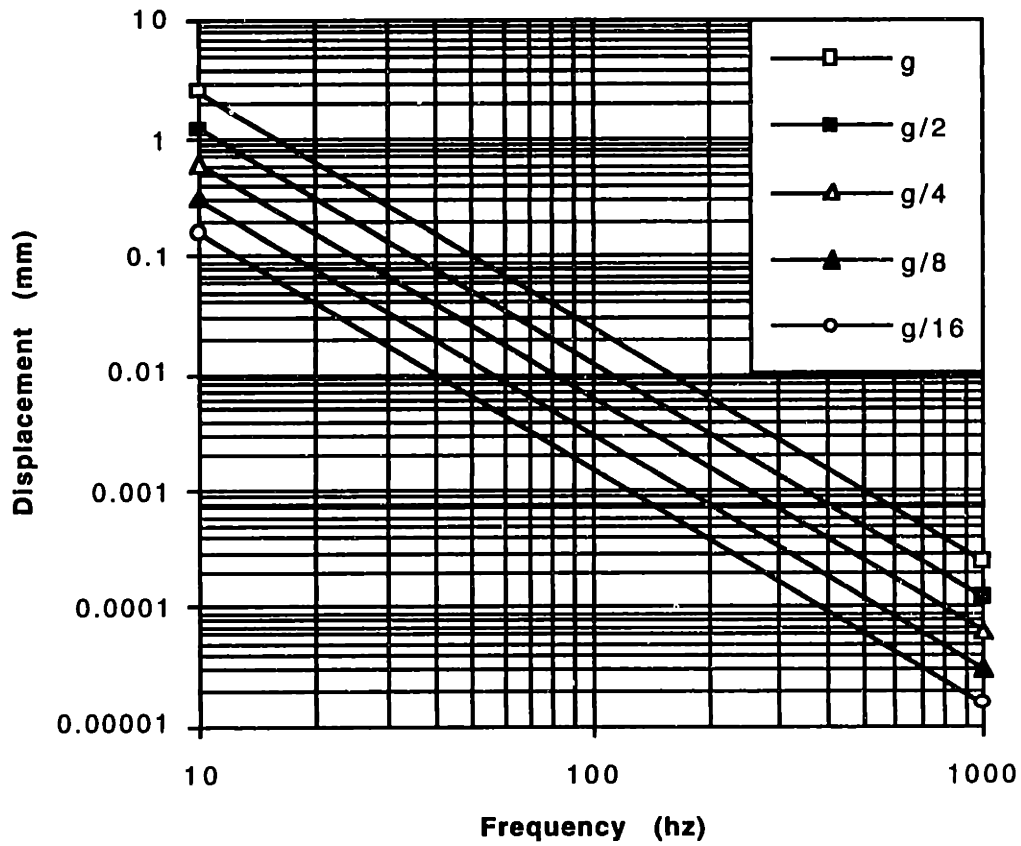


Figure 4-16 This graph shows the relationship between displacement and natural frequency for a simple mass-spring system subject to a constant acceleration expressed as a standard gravity *g*.

End Conditions	Uniform Load	First Mode	Second Mode	Third Mode
Fixed-Free	1.68	1.875	4.69	7.85
Pinned-Pinned	2.96	3.14	6.28	9.43
Fixed-Fixed	4.43	4.73	7.85	11

Table 4-2 This table gives the coefficient λ in Equation 4.27 for a uniform cross section beam, which depends on the end conditions and the bending mode number. The column labeled Uniform Load is calculated using the deflection equation for a uniformly loaded beam and shows the approximate relationship between acceleration induced displacements and natural frequency.

Figure 4-17 represents a lumped-parameter model of a rigid machine tool structure mounted on relatively compliant bearings. We wish to determine the effect of the bearing stance, described by the aspect ratio *a/b*, on the natural frequency of the system. The system has three degrees of freedom with natural frequencies given by Equation 4.28, but the lowest frequency, which corresponds to a combined horizontal and angular motion, is of primary interest. Figure 4-18 shows the rather significant effect that the aspect ratio has on the lowest natural frequency and its square, which have been appropriately normalized. For example, an aspect ratio of 3:1 causes a factor of ten greater displacement (to an inertial load) compared to the structure supported through its center of mass.

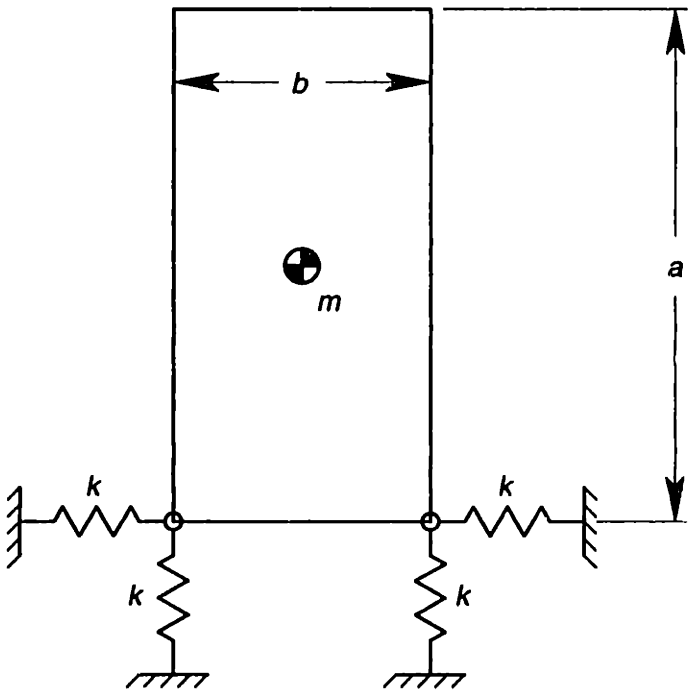


Figure 4-17 Lumped-parameter model of a rigid machine tool structure on compliant bearings.

$$\begin{bmatrix} \omega_1^2 \\ \omega_2^2 \\ \omega_3^2 \end{bmatrix} = \frac{2k}{m} \begin{bmatrix} 2 - \dfrac{\sqrt{4a^4 + 5a^2b^2 + b^4}}{a^2 + b^2} \\ 1 \\ 2 + \dfrac{\sqrt{4a^4 + 5a^2b^2 + b^4}}{a^2 + b^2} \end{bmatrix} \tag{4.28}$$

The large size and varied content of this section may obscure some of the main messages; therefore, a brief summary is in order before moving on to the next section on complex structures. Blanding's Statement 1 is the main message to design by. It is important that the designer sees how loads flow through the structure and that the flow is in plane rather than in bending.

Statement 1: When designing a structure for optimal stiffness, it's important that all its members (bars and plates) be used in stretching or compressing (or shear) rather than bending.

Statement 5 follows from Statement 1, but the message of *structurally closed* is too critical to simply infer. The example of the closed box demonstrated that every face is critical to torsional stiffness and the optimal design occurs when the faces are equal in thickness (for the same material).

Statement 5: Each and every face of a polyhedral shell must be two-dimensionally rigid in order for the entire structure to be three-dimensionally rigid. In other words, the polyhedral shell must be "structurally closed."

Rule 3 may seem inconsequential but it appears in every equation for compliance in this section. Designing the lever arm or lever ratio to be favorable is certainly an important message. The last example of the rigid machine tool structure on compliant bearings demonstrated that a non favorable aspect ratio is very costly to performance.

Rule 3: A lever arm or lever ratio always appears as a squared term in a compliance or a stiffness.

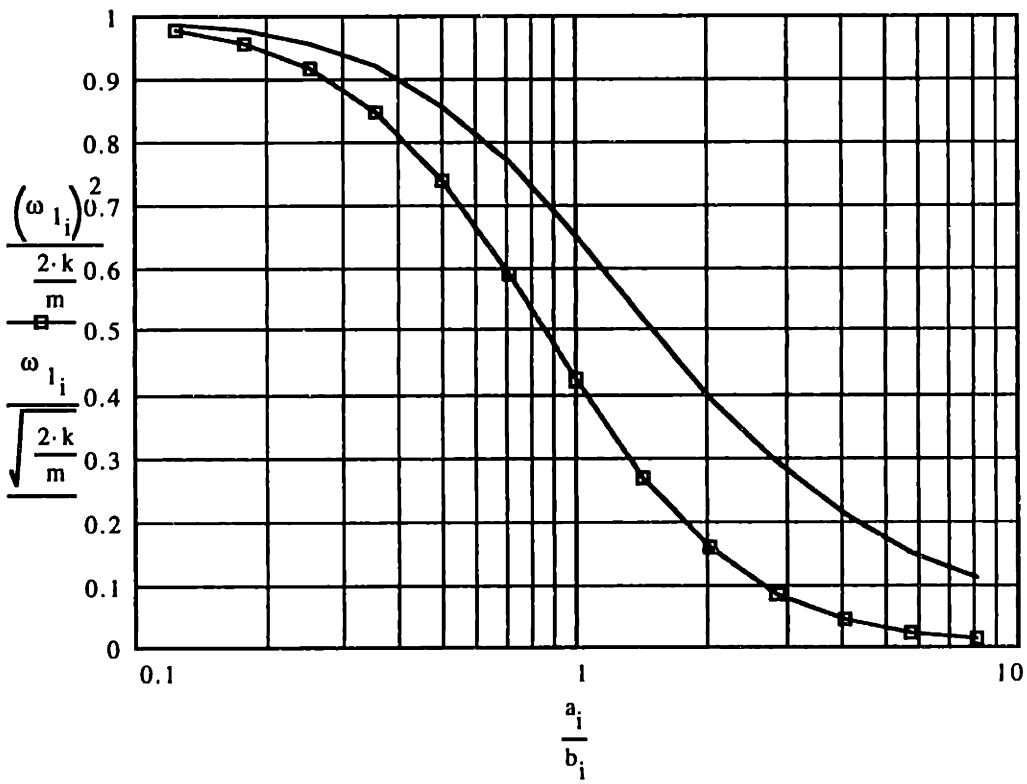


Figure 4-18 The figure of merit for the bearing stance (lower curve) falls rapidly as the aspect ratio becomes less favorable.

4.2 Modeling Complex Structures

Machine-tool structures and assemblies of structural components are generally too complex to accurately model using the simple methods of the previous section. However, the designer would be wise initially to trade off accuracy for near-term feedback of general trends or obvious problems that simple models would provide. Three basic techniques for modeling structures have been useful through the course of this thesis. These are physical models, finite element models and engineering models. The particular model may range in detail from simple to elaborate with proportionate costs and levels of accuracy.

The most direct technique is the physical model. Although seemingly old fashioned, simple models constructed of paper or wood are very enlightening and inexpensive to build. Stereo lithography machines can generate plastic models directly from solid model files. Intermediate in the range of detail would be testable scale models or prototype subassemblies. Full-scale prototype machines are obviously very expensive to build and test, but are essential to uncover subtle systems problems. To emphasize the need for physical models, Dr. Donaldson noted that "Mother Nature makes no simplifying assumptions."

Finite element analysis (FEA) has become a powerful analytical tool that is now available on the desktop. System costs range from a few thousand dollars to tens of thousands of dollars; however, the cost of the learning curve is often more significant. A well-known professor of finite element analysis emphasizes that while FEA may solve the mathematical problem very accurately, its approximation to the physical problem depends upon the skill and knowledge of the analyst.[1] The effective use of FEA in the design process also requires skill and knowledge. Developing simple finite element models concurrently with the design of key structures is very effective early in design. As the design matures, the finite element model will become progressively more accurate since the analyst already knows a good deal about the structures. Once the design and the finite element model are beyond a few iterations, a sensitivity analysis of key figures of merit to various design parameters provides further understanding important for optimizing the design.

An engineering model is the modern extension of pencil-and-paper analysis to the computer spreadsheet or mathematics program such as Mathcad™. Examples of engineering models pervade this thesis. For structures, an engineering model cannot match the detail of a finite element model and developing the engineering model requires considerable engineering knowledge and some proficiency with a mathematics program. However, the engineering model provides greater flexibility for custom problems and the whole process helps develop valuable intuition and understanding.

---

[1] Professor Bathe of M.I.T.

The main purpose for including this section is to share some ideas and thoughts on how to make finite element modeling more effective during the structural design process. Perhaps the greatest asset for design is the visualization that FEA provides. Animated mode shapes (or deflection shapes for strictly static analysis) are valuable for identifying areas that have significant distortions (i.e., relative displacements) and thus require reinforcement or perhaps new structural members. Conversely, material should be removed from areas where the absolute displacements are large but the distortions are small. These changes will increase the modal frequency of the particular mode shape but a quantitative indicator is needed to know best where to add and remove material. An indicator based on the Rayleigh quotient is theoretically simple to incorporate into a finite element program, although the feature may be difficult to obtain from a vendor. Equation 4.29 shows the Rayleigh quotient for a vibrating system represented by a stiffness matrix **K** and a mass matrix **M**, where ω is the modal frequency and Φ is the mode shape. The numerator in (4.29) is the modal strain energy and most finite element programs can generate contour plots of strain energy density. The denominator when multiplied by ω^2 is the modal kinetic energy of the vibrating system. Rayleigh recognized that the strain energy and kinetic energy are equal in amplitude at resonance and provide a fairly accurate estimate of the modal frequency using only an approximate mode shape. We can use the Rayleigh quotient as a sensitivity indicator to show how to change the finite element model to increase the frequency.

$$\omega^2 = \frac{\Phi^T \mathbf{K} \Phi}{\Phi^T \mathbf{M} \Phi} \tag{4.29}$$

The finite element analysis generates **K** and **M** and then calculates a number of modes and frequencies. Assuming for the moment that only one mode is problematic, we can consider what happens to that modal frequency when a small change is made to the model, for example, a 10% change to the wall thickness of a group of elements. Rather than re-solving the model, we can use the Rayleigh quotient for those elements to indicate whether to increase or decrease the material. If the local Rayleigh quotient is greater than the average, then additional material will increase the modal stiffness faster than the modal mass, thus increasing the modal frequency. Conversely, a local Rayleigh quotient that is below average indicates that less material is better. To make this method truly effective, we need a contour plot of the modal frequency distribution given by the square root of the Rayleigh quotient. This plot would clearly show how best to move material from low-frequency regions to high-frequency regions. Probably the best way to extend this method to multiple mode shapes is to plot each one and look for common areas where changes will provide the most improvement.

Unfortunately the idea of modal frequency distribution has not been demonstrated on a FEA system. A potential platform is a third-party software product that works with

many FEA products to provide additional flexibility in the presentation of FEA data.[1] This company is capable and willing to provide special features to their product for a reasonable fee.

When the analysis is static rather than modal, the previous method reduces to a familiar strain-energy approach. Assuming that the deflection shape does not change, strain energy is proportional to stiffness. Then the best place to stiffen a structure is in regions of high strain energy density. It may be confusing that the total strain energy will decrease when the structure is stiffened because the load on the structure is specified, thus making strain energy proportional to compliance. From this point of view, the stiffener is placed in a region of high compliance. Strain energy also provides a convenient accounting measure of compliance throughout the parts of a machine.

A strain-energy-density plot of the structure clearly shows the effectiveness of existing material but it does not show very well the relative movement between members that require a better connection. One approach to show relative movement is to fill the voids in the structure with a low-modulus material that will not influence the structure. Plots through the filler material alone will show relative displacements as high strain energy and indicate where to put new structural members.

---

[1] Visual Kinematics, Inc. 883 North Shoreline Blvd. Suite B210, Mountain View, CA 94043. 650-961-3928, Fax. -9286.

4.3 Shear Panel Models

A shear panel is an important structural component that frequently requires an opening for functional or manufacturing reasons, for example, to clean core sand from a casting. The opening shape and size can greatly affect the shear stiffness of the panel as demonstrated by this finite element study of several shear panel designs.[1] Figure 4-19 shows how significantly the compliance varies as a function of hole area and for different designs each represented by a curve. The next several figures show the strain energy distribution for each design. A strain energy plot clearly shows the effectiveness of the existing material. Areas of high strain energy have the most influence on the overall stiffness and indicate where reinforcement is the most beneficial.

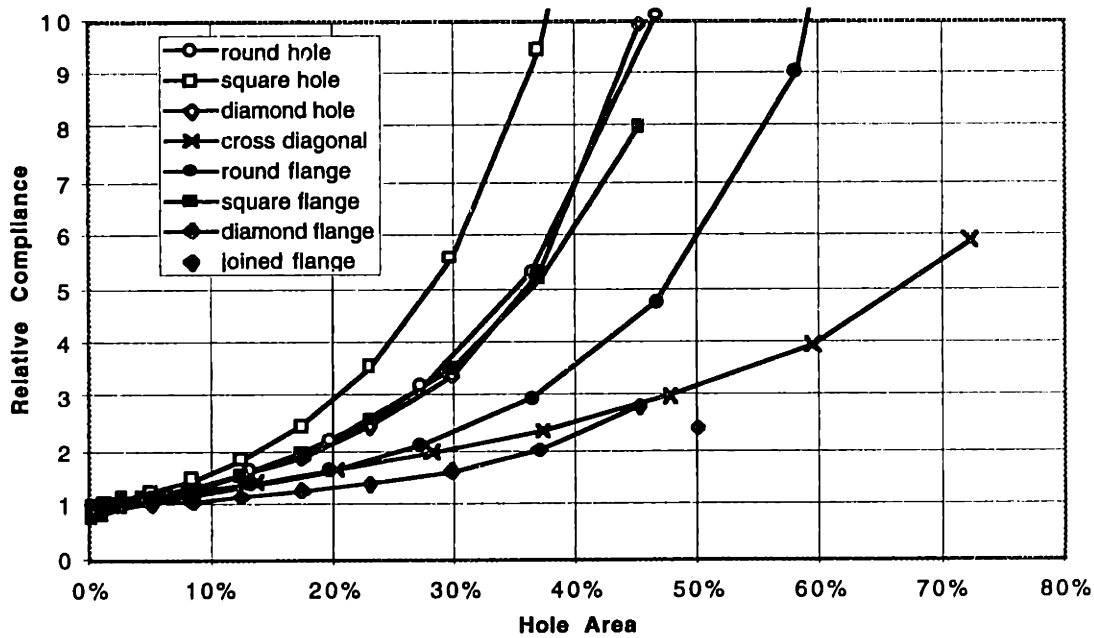


Figure 4-19 This plot shows the compliance of a shear panel having various opening designs and sizes. As indicated, each curve represents a different design and shows the compliance as a function of the hole area relative to the compliance of a solid panel. The next several figures show the different designs. Each shear panel is 400 mm square with 100 mm wide exterior flanges and 10 mm thick walls.

It is useful to notice that the markers on the curves resemble the hole shapes and the filled markers indicate designs with reinforcing flanges around the holes. The extra flange material significantly stiffens the round, square and diamond holes, respectively, 2.12, 2.16 and 3.5 times for hole areas at or near 45%. The cross diagonal design is best in terms of material efficiency because shear loads are equivalent to axial loads applied diagonally across opposite corners. This is why the diamond hole and round hole are significantly better than a square hole of equal hole area, and why the diamond flange provides the most

[1] Pro/MECHANICA™ by Parametric Technology Corp. is the finite element software used in this study.

improvement. At a hole area of 50%, the diamond flange becomes joined to the exterior flange to provide a complete load path through the flanges.

Referring to the strain energy plots in Figure 4-20, the local concentrations in the corners of the diamond or square could be reduced by adding fillets, and in the limit, would approach a round hole. When unobstructed access through the center is required, the round hole usually provides the better compromise. Although the numbers are difficult to read, the maximum strain energy for the cross diagonal is less than one-tenth that of the others and the distribution is the most uniform. The cross diagonal is clearly best when an obstruction is allowed across the center.

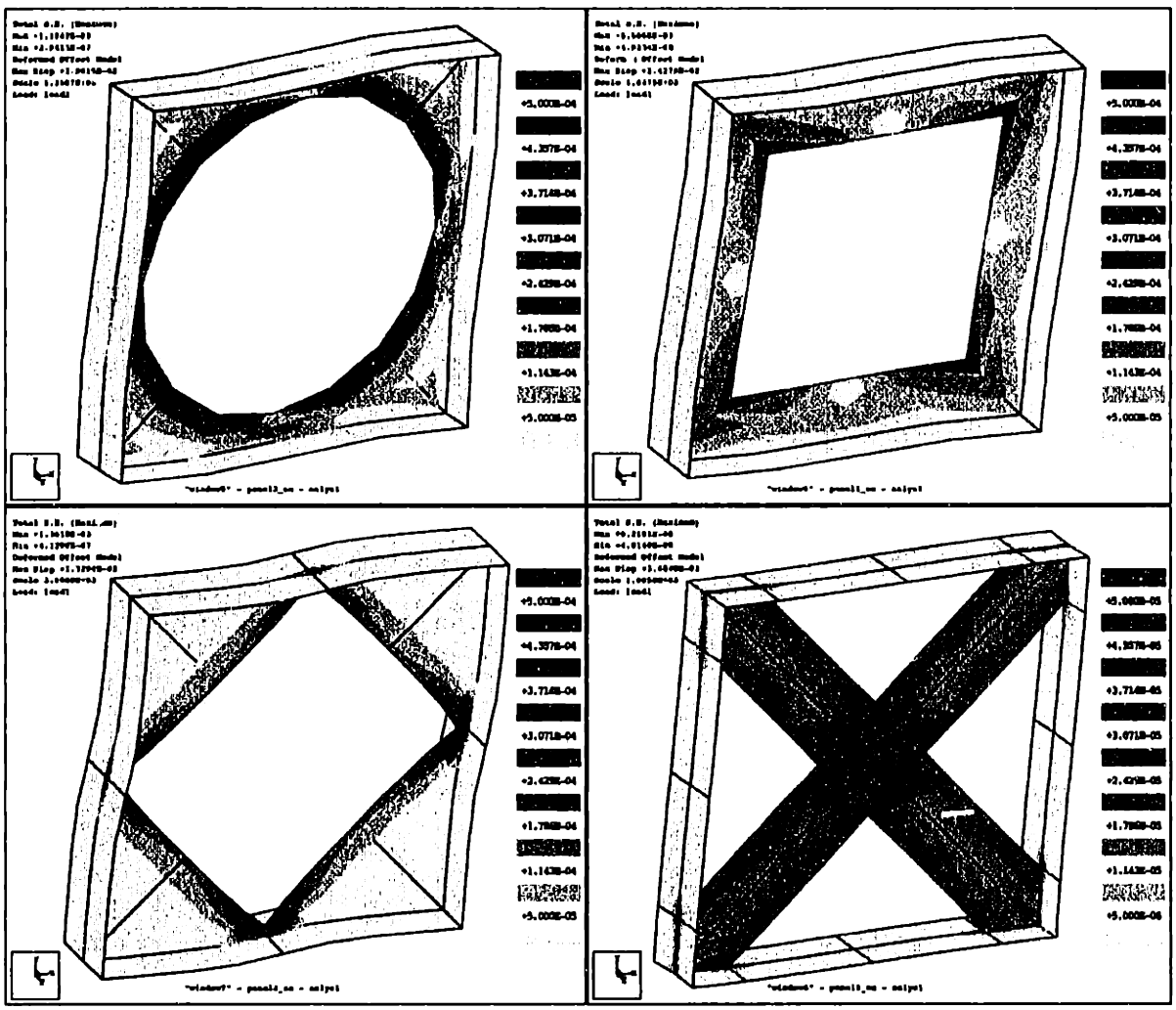


Figure 4-20 Plots of strain energy distribution for round hole, square hole, diamond hole and cross diagonal designs.

Referring now to Figure 4-21, the diamond flange, although stiffer than the round and square flanges, shows high concentrations of strain energy in the corners. Each corner displaces relative to the outer flange as the web deforms in shear. The axial load carried uniformly across the flange near its mid length must transfer into the web near the corner. The better way to arrange the diamond flange is to join the corners to the outer flanges as a truss. The uniform distribution of strain energy in this case indicates that the flange material is more efficient. The round flange is not quite as effective because the outer edges bend inward to partially relieve tensile loads or bend outward to partially relieve compressive loads. To be efficient, the round flange should not be too thin relative to its width.

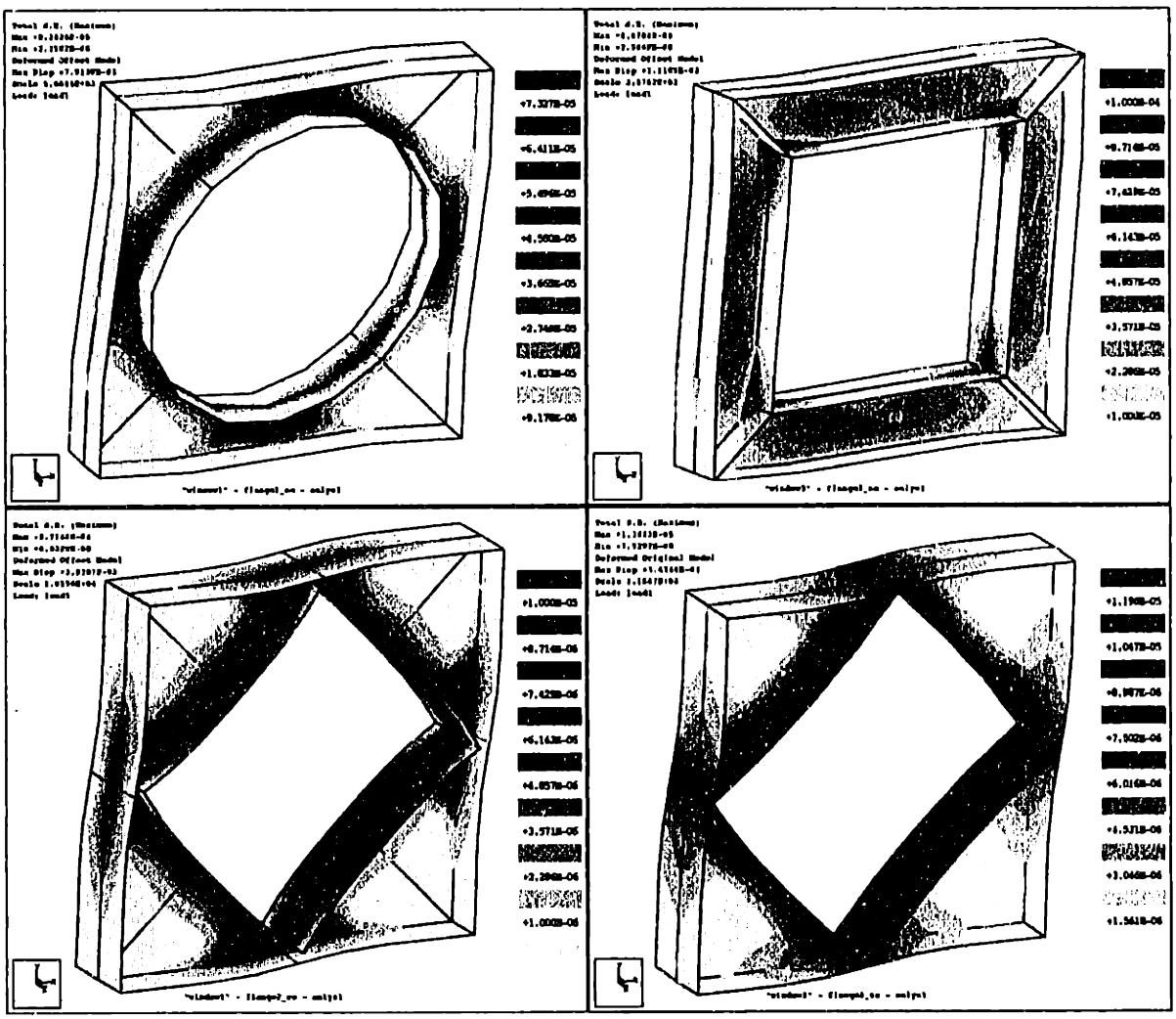


Figure 4-21 Plots of strain energy distribution for round flange, square flange, diamond flange and joined flange designs.

4.4 A Case Study of the Maxim™ Column

The column on the Cincinnati Milacron Maxim™ machining center is a ring-shaped structure that supports the Y-axis ways and accommodates the spindle carrier within the ring. The column in turn travels on the X-axis ways that are mounted to the machine base. Figure 4-22 shows the basic structure of the model B column and several of its design features are worth noting. Because this casting has partial inner walls rather than fully enclosed uprights, it can be cast with one main core rather than an assembly of three cores. This also makes core sand easy to remove from the inside of the column. A second core is required to form the bottom box. The large openings in the front and back faces of this column are naturally compliant in shear and torsional stiffness suffers as a consequence. Shear stiffness particularly in the front face is important for the accuracy of any machine that has the Y-axis ball screw located off center. Any hysteretic moment that occurs in the Y-axis deforms the column in a racking or shear mode. This manifests reversal errors in both the X and Y axes. Thus, any improvement made to shear stiffness directly improves the accuracy of the machining center. In this case study, the aim is to improve accuracy without adversely affecting the cost or the envelope of the column.

Figure 4-23 shows the first approach to form a continuous ring using a structurally closed cross section. The rear of the structure is an open truss to allow access for removing sand. Initially, we envisioned the diagonal ribs running full depth so they could be cast from the back without cores. Discussions with Cast-Fab Engineers indicated that shallow ribs formed by a ring-shaped core would not be particularly expensive and would be easier to cast and clean.[I] Shallow ribs provide a better structural design since their purpose is to structurally close the rear (the front is closed by a web). Full-depth ribs are an advantage in the corners where the flanges require support to transfer tension or compression around a bend, and this is the reason for having extra ribs in the front corners.

The (full-depth) truss design was compared to the Maxim column using static FEA for three load conditions and two boundary conditions.[II] Table 4-3 shows the results for the Maxim (slightly larger to match the envelope of the others), two versions of the truss design (55° and 45°) and a completely solid structure to show an upper bound for stiffness. The load was applied to the front face of the column either as an X-direction force (1000 N) or a Y-direction moment (700 N-m) or the sum of the two, which is equivalent to an X-direction tool load. The boundary conditions were either pinned supports to directly compare the column designs or spring supports to include the compliance of the bearings.

[I] David Knapp and Roger Fleckenstein of Cast-Fab Technologies, Inc. 3040 Forrer St. Cincinnati, OH 45209-1016, Tel: 513-758-1101.

[II] Pro/MECHANICA™ by Parametric Technology Corp. is the finite element software used in this study.

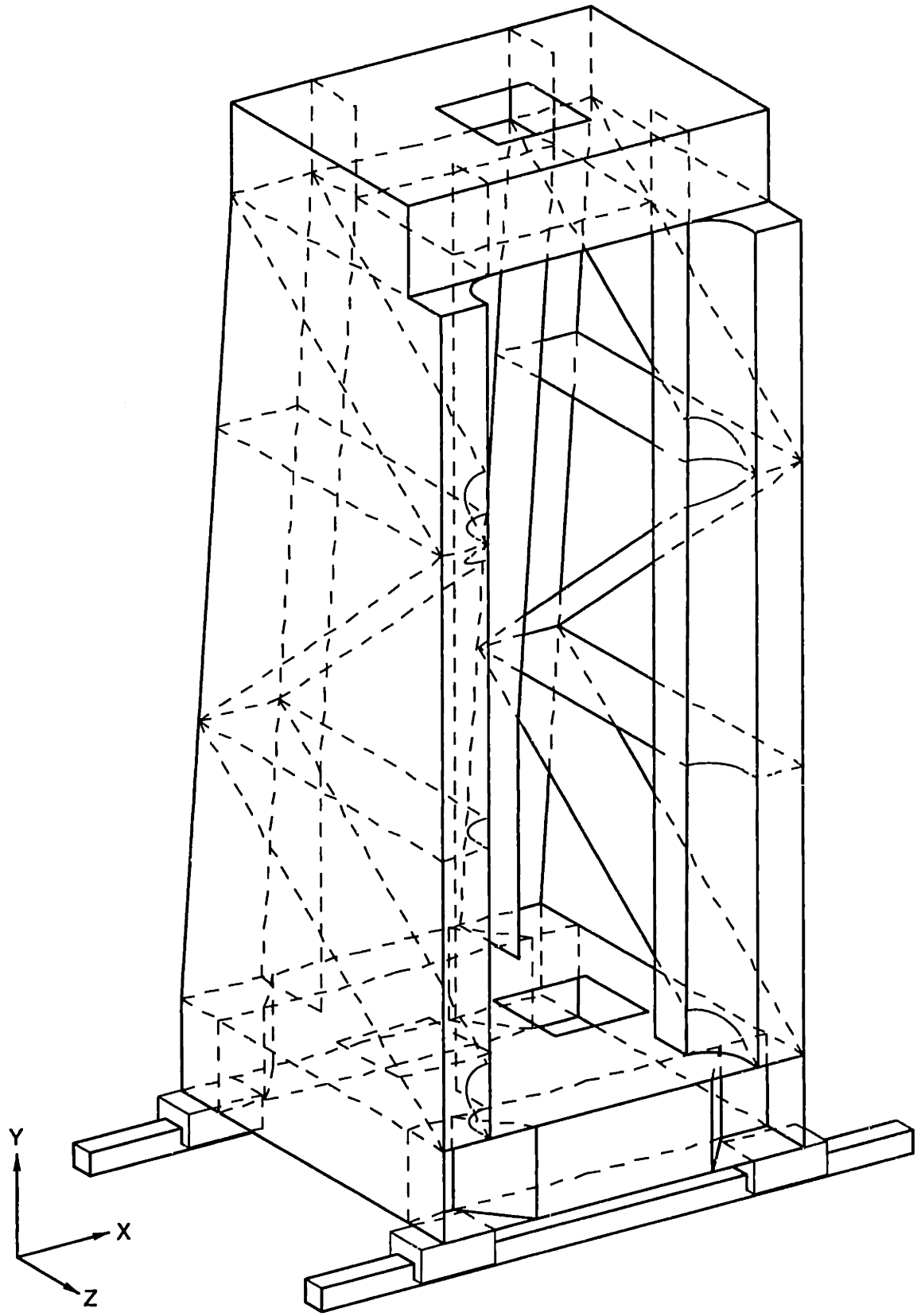


Figure 4-22 The standard column casting for the Maxim machining center. The top corner joints are fairly stiff but the lower corners are not well connected due to the pockets above the bearing blocks. The diagonal ribs contribute only about 6% of the total torsional stiffness and 9% of the mass.

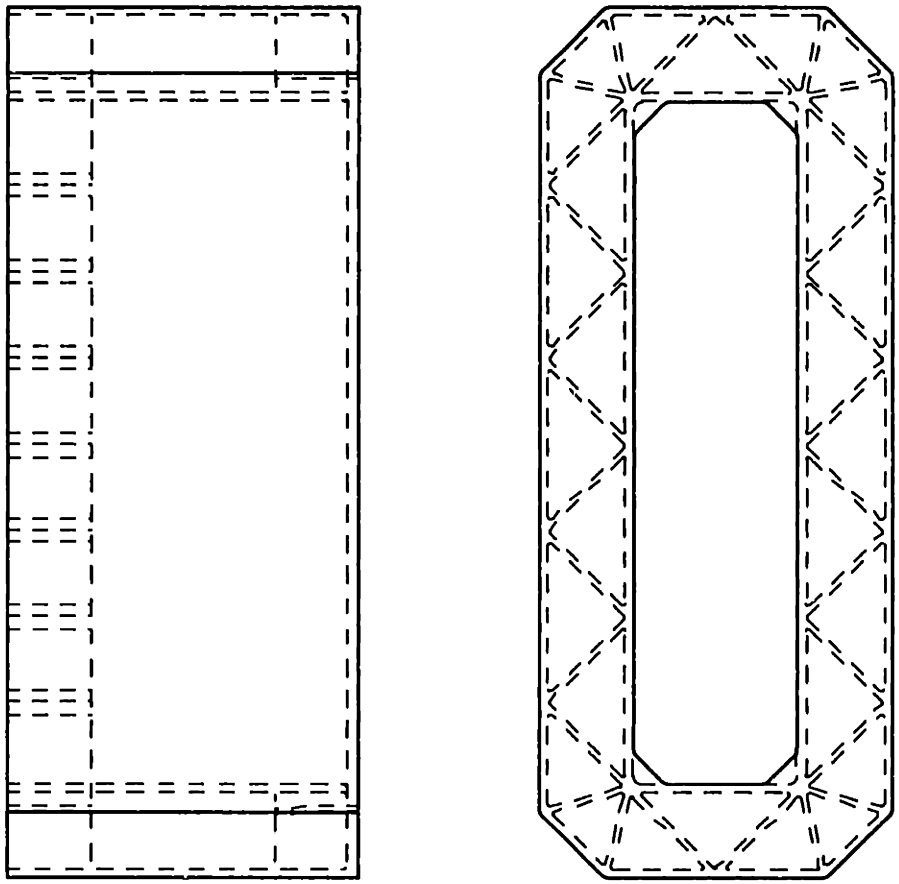


Figure 4-23 A torsionally stiff ring structure may be cast with an open face (for cleaning out the sand mold) provided that diagonal ribs are used to carry the shear load. Rather than diagonal ribs running full depth, Cast-Fab engineers recommended shallow ribs formed by a single-piece, ring-shaped core.

Support	Measure	Loading	Units	Column Design			
	(absolute)			Maxim	55 truss	45 truss	Solid
	mass		kilogram	1848	2263	2200	6028
pinned	x displ. at tool	1000 N	micrometer	29.12	12.09	12.79	4.92
pinned	x displ. at tool	700 N-m	micrometer	5.71	4.75	4.67	1.92
pinned	x displ. at tool	Combined	micrometer	34.83	16.84	17.46	6.84
spring	x displ. at tool	1000 N	micrometer	31.22	17.85	18.21	11.58
spring	x displ. at tool	700 N-m	micrometer	8.66	6.86	6.84	4.08
spring	x displ. at tool	Combined	micrometer	39.88	24.71	25.05	15.66

Table 4-3 Finite element analysis results comparing the Maxim column to a truss design. Fairly significant improvements are possible within the envelope of the standard Maxim column.

The analysis that best represents the reversal problem is the pinned support with the 1000 N load. Compensating for the difference in mass, the 55° truss design is almost two times better than the Maxim column (1.97). Unexpectedly, the solid column is nearly as mass efficient being 1.81 times better. This may indicate that areas of concentrated strain energy exist where walls should be thickened. The case that best indicates the tool point compliance is the spring support with the combined load. Compensating for the difference in mass, the 45° truss design is 1.34 times better than the Maxim column. The shallow rib

design was not modeled, but the results should be somewhat better because the rib material is optimally placed at the rear of the column.

A rather curious result of the analysis is that the Maxim column is nearly as stiff in torsion as the truss design whose members are torsionally stiff. At the time, we did not fully appreciate the significance of having front and rear faces that support shear as opposed to side members that are torsionally stiff. Certainly the side members of the truss design are stiffer in torsion than the sides of the Maxim column with rather long diagonal ribs. Perhaps the lower box of the Maxim column is more effective than the equivalent truss structure? To test this hypothesis, we constructed a paper model with upper and lower boxes connected by open-section sides similar to the B column but without diagonal ribs. The model felt very stiff and the design could be cast with a single-piece core. This seemed to be a breakthrough but there was one more surprise looming. After cutting a face on each box to make it structurally open, there was no apparent decrease in stiffness. The whole structure is torsionally stiff because the perimeter frames on the front and rear faces support shear, and the benefit from torsionally stiff members is rather insignificant. We demonstrated this technique in a new model constructed similarly to Figure 4-24.

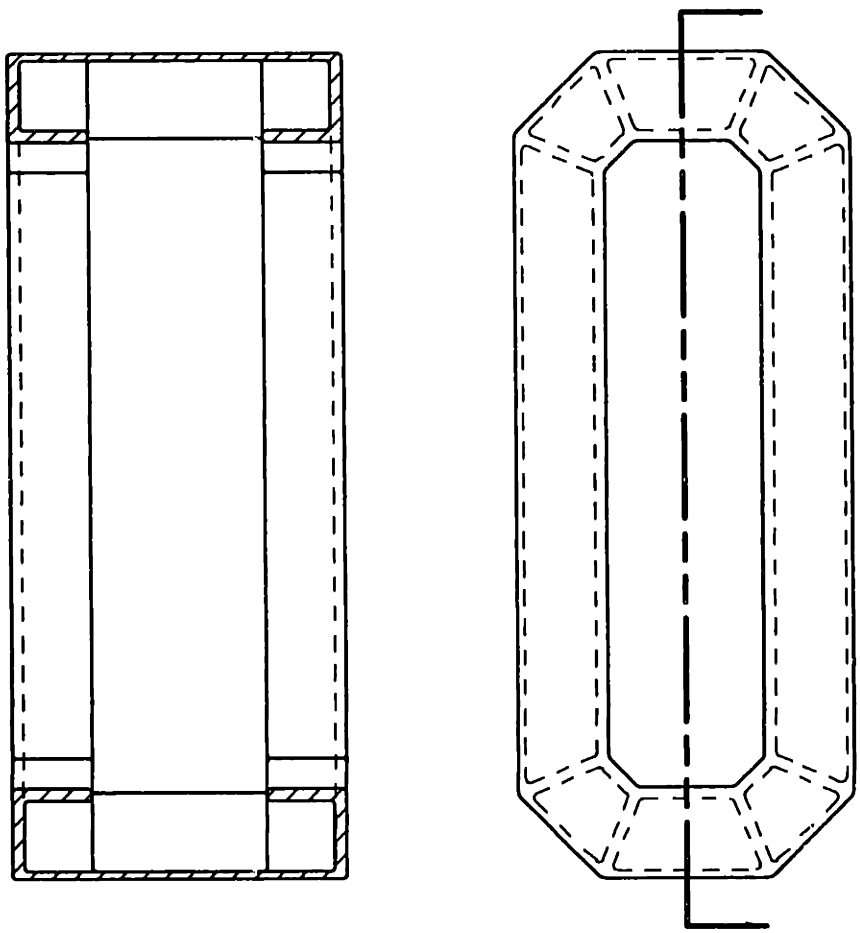


Figure 4-24 This ring structure is torsionally stiff even though the side members are open sections that individually are very flexible in torsion. The framework around each open face supports shear and effectively closes the structure for torsional loads. A single-piece core and large openings make this easy to cast.

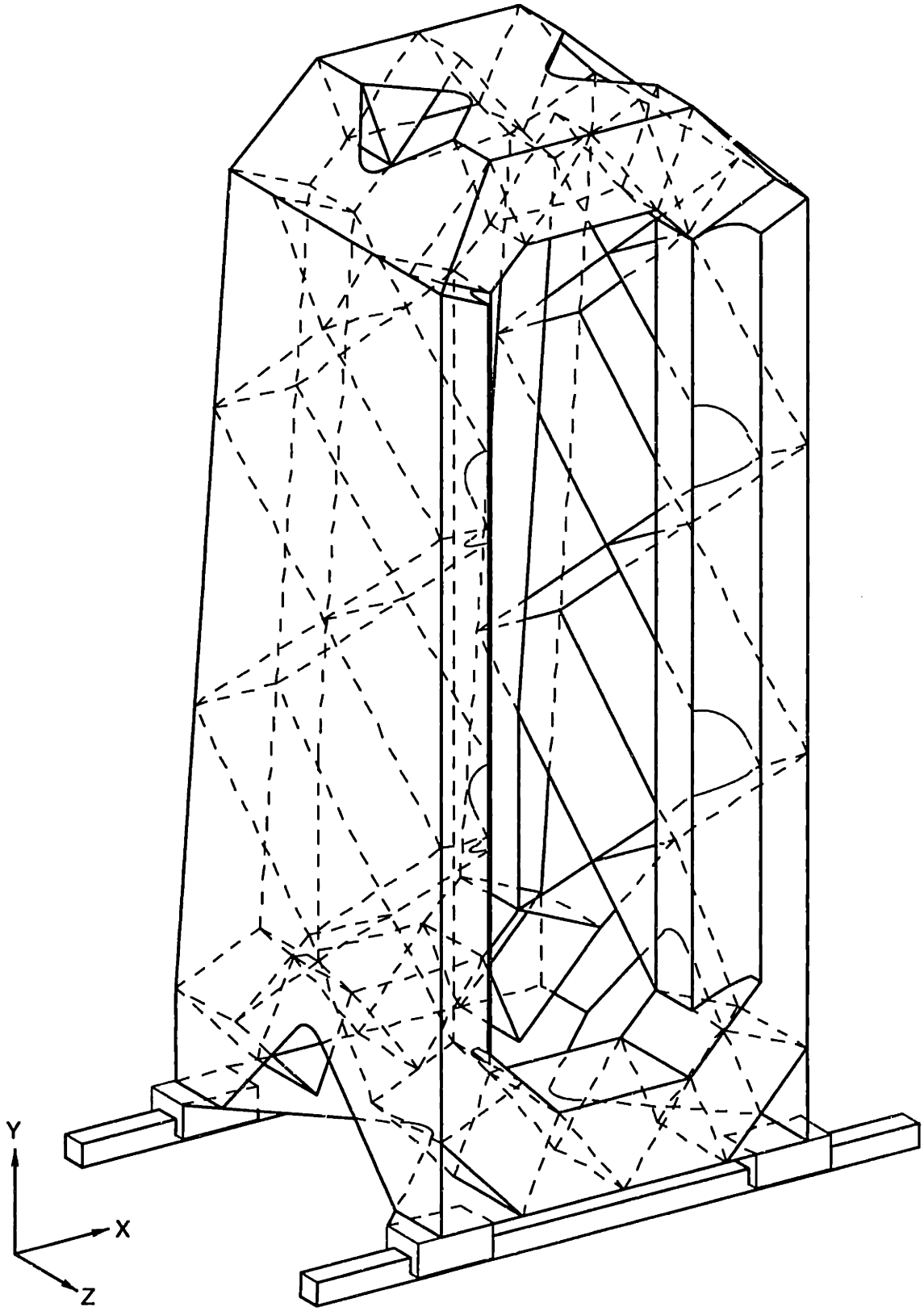


Figure 4-25 The column for the Maxim machining center will benefit from improved corner joints that increase the shear stiffness of the open faces. The cross-pattern of ribs provides twice the torsional stiffness as diagonal ribs; however, a sensitivity analysis using FEA may show that the material would be more effective if used in the framework around the openings.

The Maxim column is fairly easy to modify along the lines of Figure 4-24. The main structural improvement shown in Figure 4-25 is in the corners where the flanges are continuous and well supported around the bends. Also shown is an extra set of diagonal ribs in the sides to replace the horizontal ribs. The cross-pattern will provide approximately twice the torsional stiffness as one set of diagonal ribs; however, this material may be more effective if used in the framework around the openings. Determining this would require a sensitivity analysis using FEA. Certainly the casting would be easier to clean without the ribs, but even so it is more open than the original Maxim column.

Another curious result of the analysis on the Maxim column is that an X-direction tool load causes greater Y-direction reaction loads in the rear bearings than in the front bearings! Since the tool point is not at the shear center (near the physical center), the column twists and causes the unexpected loading. What if we designed the column to place the shear center at the tool point? Then a torsionally compliant column, which is exactly constrained by four bearings, would not twist at all under a tool load.

Figure 4-26 shows a free-body diagram of a simple column constructed of sheets. The tool load F produces shear flow in the sheets and ultimately bearing reaction loads R_f and R_r. If the front and rear bearings have equal stiffness, the column will not twist when R_f and R_r are equal, assuming an infinitely stiff base. Equation 4.30 shows the condition required for zero twist, which is independent of the tool height h.

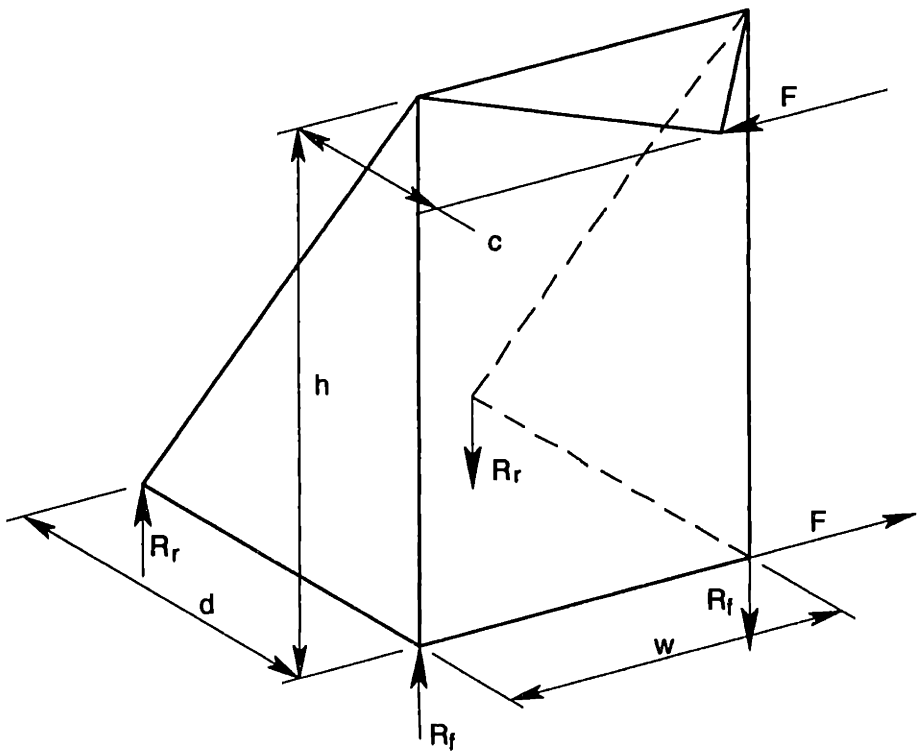


Figure 4-26 This free-body diagram of a simple column shows the bearing reactions that result from a tool load F. The condition required to place the shear center at the tool point is $d = 2c$.

$$R_r = F\frac{c}{w}\frac{h}{d} \qquad R_f = F\left(\frac{h}{w} - \frac{c}{w}\frac{h}{d}\right)$$
$$R_r = R_f \quad \Rightarrow \quad d = 2c \tag{4.30}$$

It is important to remember, however, that the shear center analysis applies to static loads. A dynamic load may well interact with a torsional mode and produce sizable displacements. A better design approach is to locate the bearings so that the mode shape has a node at the tool point. Then a tool load would not excite the mode and the tool point would not displace if the mode becomes excited. To get a mental picture of how this might work, imagine the column in Figure 4-26 in a torsional mode such that the front and rear bearings are moving the same amplitude. It so happens that the shear center does not translate for this hypothetical mode shape and therefore is a node (a paper model shows this clearly). A real column would behave in a similar way but would require a unique bearing spacing that is predictable using FEA.

Figure 4-27 shows a new column designed to place the shear center near the tool point. The rear face is open to provide torsional compliance. This column appears to be lighter and more stable but derives virtually all the torsional stiffness through four bearings to the base. Unmounted it is rather flimsy and requires special care during manufacturing operations. In this regard, X-axis vertical straightness errors couple directly into the Y-axis but three-dimensional compensation can handle such coupling. The top, bottom and side walls would be minimum thickness over most of the depth and thickened in the front as part of the perimeter frame. The Y-axis rails would bolt as close as possible to the side walls so that the loads transfer directly into the structure. There are several changes to note about the perimeter frame. The lower part of the perimeter frame is not nearly as tall as the upper part. This reflects the desire to raise the X-axis rails and ball screw closer to the tool point. Also in these regions, the flanges join at the center rather than being parallel (as shown in Figure 4-24). This improves the shear stiffness in these regions without sacrificing the bending stiffness since the bending moment tapers to zero at the center. The new column design is more open than the Maxim column and access to pockets and corners that require cleaning of sand is very good. The side walls may require some thin stiffening ribs but these probably would be external for easy access.

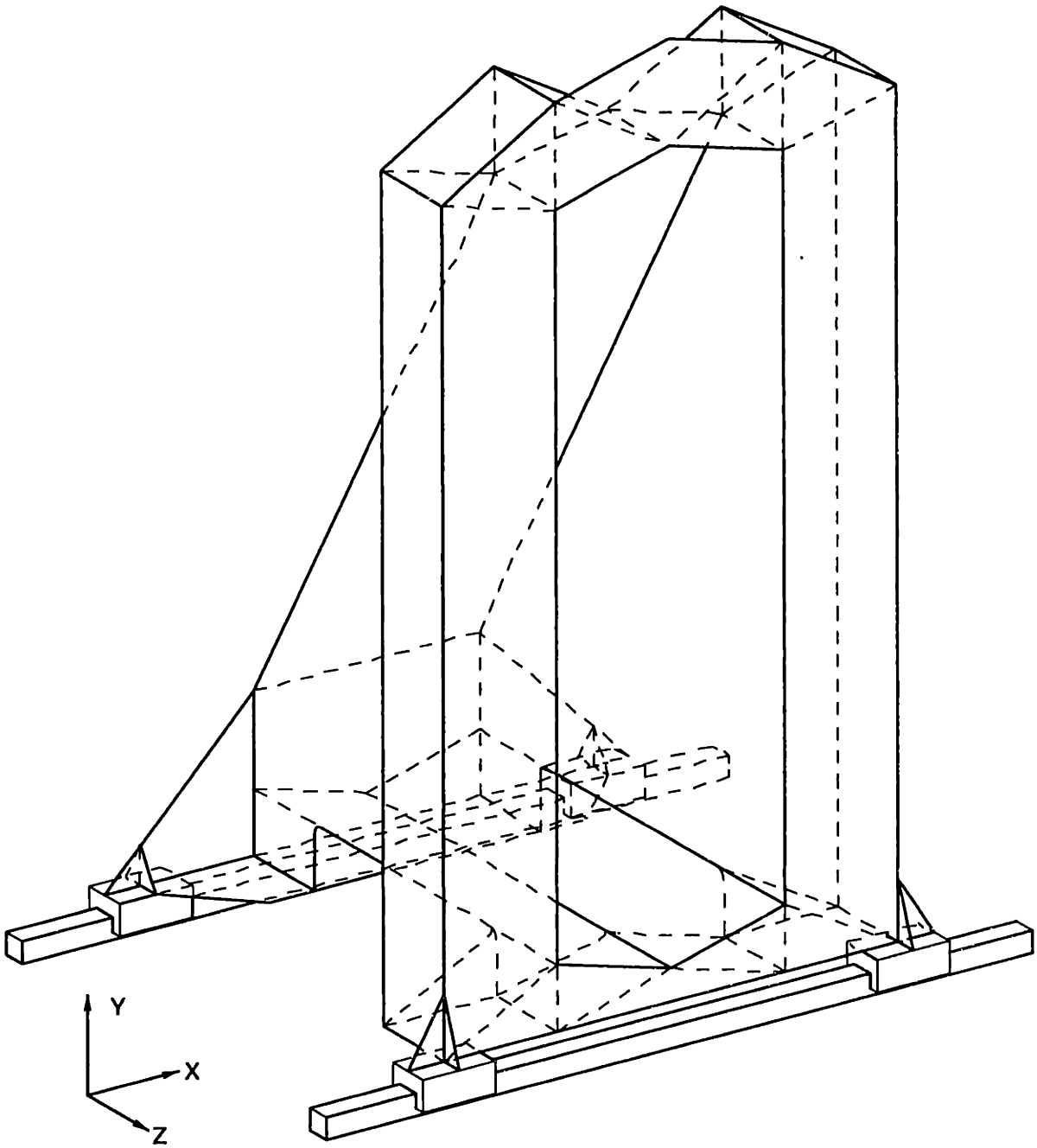


Figure 4-27 This torsionally compliant column requires four bearings to approximate an exact-constraint system. The perimeter frame in front provides X-direction shear stiffness and the sloping rear face is open.

5 Deterministic Damping

Damping is recognized as being very important in machine tools to reduce levels of vibration thus improving surface finish. The vibration in the machine may result from an external source transmitted through the foundation, from an internal imbalance or roughness, or from the cutting process. In an extreme case, self-excited chatter may limit the productivity of the machine. Vibration can be reduced at the source by isolating the transmission path or by making the machine more robust with damping. An inverse measure of damping known as the quality factor Q is the amplification of the system vibrating at resonance. The loss factor $\eta = 1/Q$ is a direct measure of damping and is approximately twice the damping ratio ζ in a second order system. In a machine, material damping usually plays a very small role compared to other mechanisms such as micro slip in bolted joints and viscous effects in bearings. Dampers based on sliding friction are generally undesirable because they require a certain level of excitation to cause movement and may cause nonrepeatable behavior of the machine. The deterministic damping mechanisms presented in the following sections are viscous based and can significantly increase structural damping when the system is properly engineered. In each case, the damping treatment is most effective when it is impedance matched to the system, which requires analysis and/or measurement to achieve optimal performance.

5.1 Viscoelastic Constrained-Layer Damping

The use of viscoelastic materials (VEM) to damp structural vibration is quite prevalent and successful in many industries. This demand has led to a variety of materials that offer high loss factors (typically of order one) over a particular frequency-temperature range. The material is typically available in sheet form and ranges in thickness from a few thousandths to a few tenths of an inch and in shear modulus from a few hundreds to tens of thousands of psi. The sheet of VEM is usually applied between the structure and a stiffener so that relative motion between them shears the VEM and dissipates energy. Figure 5-1 shows an example of VEM applied between a beam in bending and two constraining layers.

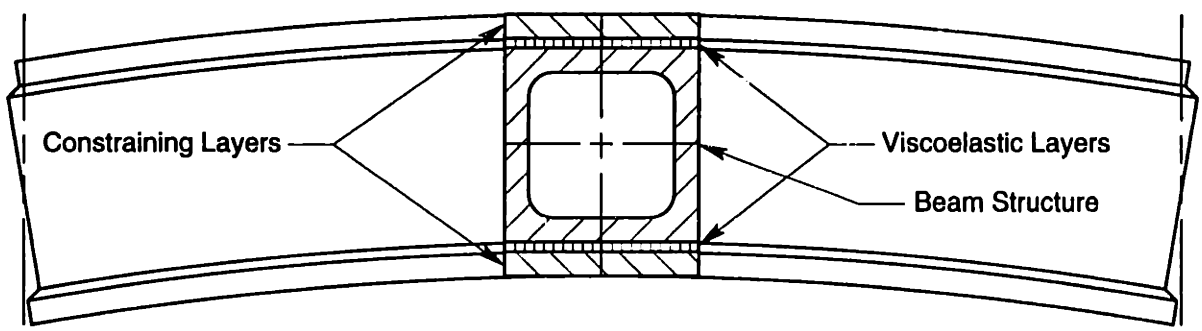


Figure 5-1 The viscoelastic layer provides the shear connection between the structure and the constraining layer that acts as a stiffener for the beam in bending.

When the structure is a flat plate, the constrained-layer damper is straightforward to model using the theory developed by [Kerwin, 1959] and [Ross, Ungar and Kerwin, 1959], which has become known as the RKU theory. More complex structures generally require finite element analysis (FEA) to model, and the approach developed by [Johnson and Kienholz, 1982] using modal strain energy is both accurate and simple. More recently, [Slocum, et al., 1994] and [Marsh, 1994] have developed replication techniques to embed constraining layers inside machine tool structures. Their theory, although approximate, applies to arbitrary beam cross sections. An analytical optimization of a damped structure is a result from this thesis [Marsh and Hale, 1995, 1998]. The development of the optimization offers insight into the relationships among key design parameters and leads to an estimate for optimal damping. This approach is applicable to beam-like structures but also has utility in conjunction with FEA of more complex structures.

A modal analysis of a structure separates the complicated dynamic behavior into a set of simple mass-spring-damper oscillators called modes. A continuous structure requires an infinite set of modes to exactly represent the behavior, but usually the lowest few modes are of practical interest. Then for a particular mode, the damped structure is equivalent to a single mass-spring-damper system that is conceptually simple to model. When the damping element is viscoelastic rather than strictly viscous, the traditional model used is the complex stiffness, where the loss factor is equal to the imaginary part divided by the real part. The assumption is that the damping force is proportional to the displacement rather than the rate, which implies that it is independent of frequency. However, both the loss factor and the shear modulus of a real viscoelastic material are frequency dependent, but the changes are small over the damping-dominated region of a particular mode. This representation is convenient because the model reduces to the combination of three springs shown in Figure 5-2. The combined mass of the structure and the damper becomes part of the excitation force and does not directly affect the damping in the system; rather it indirectly affects the damping through the frequency dependence of the real VEM.

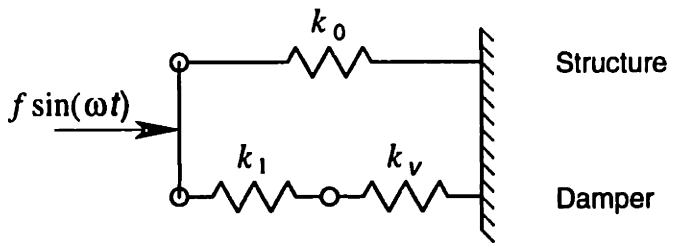


Figure 5-2 Spring model of a structure enhanced with a constrained-layer damper. The bare structure has a stiffness k_0 that acts in parallel with the damper. The damper stiffness is a series combination of the constraining layer k_1 and the viscoelastic sheet k_v. The viscoelastic stiffness is a complex number to represent damping, which makes the equivalent stiffness of the composite structure complex as well.

A constrained-layer damper acts to stiffen the structure as a stiffness added in parallel. The *optimal* damping increases with the ratio r of the constraining layer stiffness k_1 to the structural stiffness k_0, where the word *optimal* indicates the proper choice of the

viscoelastic stiffness k_v to maximize damping. A ratio greater than one is usually difficult to obtain due to other constraints, and the mode shape is likely to change in a way to lessen the benefit of a larger ratio. The usual design strategy is to make r as large as practical and to select a high-loss viscoelastic material with the appropriate shear modulus and thickness to maximize some indicator of the damping. For the three-spring system, Equation 5.1 indicates the nature of the optimization problem, where α is a nondimensional parameter that indicates the relative viscoelastic stiffness and η is the material loss factor. The damping in the system approaches zero as α approaches either zero or infinity, since both k_0 and $k_\infty = k_0 + k_1$ are assumed to be real. Without an engineering approach to damper design, it is easy to miss the optimal damping by a factor of three or more.

$$k_{eq} = k_0 + \frac{k_1 k_v}{k_1 + k_v} \qquad \frac{k_v}{k_1} \equiv \alpha(1 + i\eta) \qquad r \equiv \frac{k_1}{k_0}$$
$$k_{eq}\Big|_{\alpha \to 0} \to k_0 \qquad k_{eq}\Big|_{\alpha \to \infty} \to k_0 + k_1 \equiv k_\infty \tag{5.1}$$

The first step in the development of the optimal damper design requires separating k_{eq} into real and imaginary parts. Equation 5.2 shows this step, and the relationship for each part is plotted in Figure 5-3 as a function of α for $\eta = 1$ and $r = 1$. As expected, the real part of the stiffness increases with α from k_0 to k_∞ and the imaginary part, which represents the damping, has a maximum that coincides with the real part being midway between k_0 and k_∞. It is common to use the loss factor as the damping indicator to maximize, and indeed this was the original approach described in the aforementioned papers. Instead there are compelling reasons to use just the imaginary part rather than the loss factor (the imaginary part divided by the real part). Although the differences are not that significant, the benefits include greater dynamic stiffness, faster settling time and slightly simpler equations.[1] Therefore, we will derive the optimal damper based on the maximum imaginary part of the equivalent stiffness.

$$\frac{k_{eq}}{k_0} = \left[1 + r\frac{\alpha(1+i\eta)}{1+\alpha(1+i\eta)}\right]\frac{1+\alpha(1-i\eta)}{1+\alpha(1-i\eta)}$$
$$= \frac{1 + (2+r)\alpha + (1+r)\alpha^2(1+\eta^2) + i\eta r\alpha}{1 + 2\alpha + \alpha^2(1+\eta^2)} \tag{5.2}$$

[1] The maximum loss factor occurs at a point before (smaller α) the maximum imaginary part since the real part is an increasing function, see Figure 5-3. A better indicator of the damping treatment is the imaginary part of the damped structure divided by the real part of the undamped structure ($\alpha = 0$). When the imaginary part is maximum, the resonance peak is minimum and transients settle faster. When the loss factor is maximum, transients settle in fewer cycles but the time per cycle is longer because the real part is smaller.

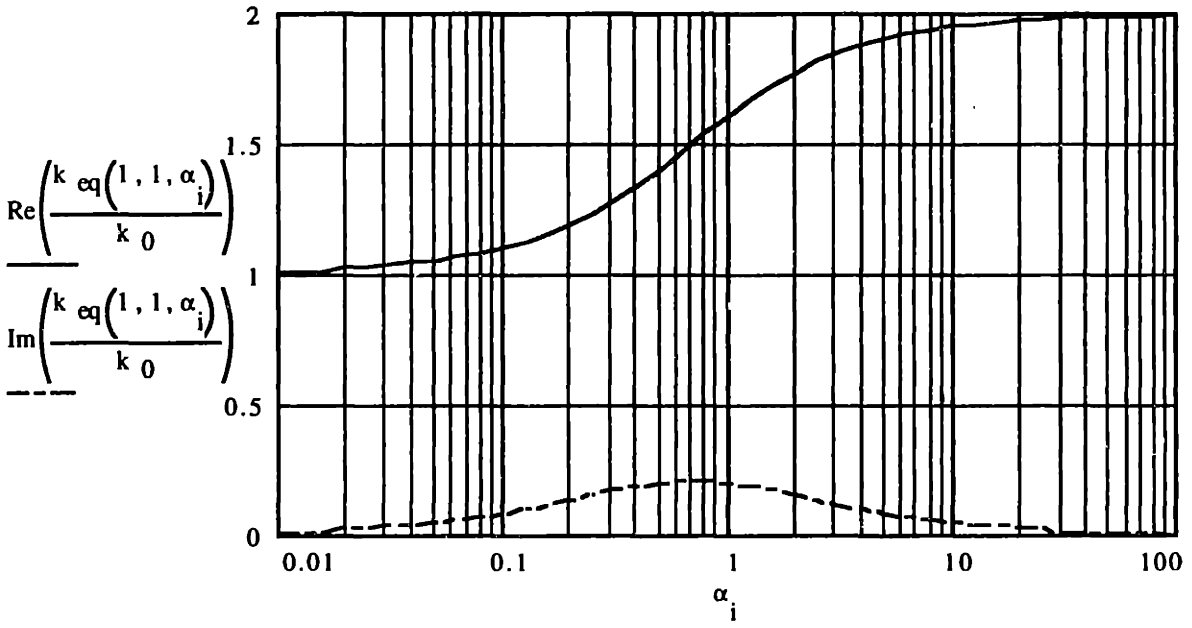


Figure 5-3 Real and imaginary parts of the equivalent stiffness $k_{eq}(\eta, r, \alpha)$ versus α for $r = 1$ and $\eta = 1$.

The maximum imaginary stiffness occurs when the derivative with respect to α is zero and solving this equation gives Equation 5.3 for α_{opt}. The optimal equivalent stiffness, given by Equation 5.4, results from substituting α_{opt} into (5.2). The maximum imaginary stiffness increases with r and η as Figure 5-4 and Figure 5-5 show, but the practical limits for both are approximately one, which results in a respectable loss factor (based on the undamped structure) of 20% or approximately 10% of critical damping.

$$\mathrm{Im}\left(\frac{\partial k_{eq}}{\partial \alpha}\right) = 0 \quad \Rightarrow \quad \alpha_{opt} = \left(1+\eta^2\right)^{-\frac{1}{2}} \tag{5.3}$$

$$\left.\frac{k_{eq}}{k_0}\right|_{opt} = \frac{2+r}{2} + \frac{i\,\eta\, r}{2+2\sqrt{1+\eta^2}} \tag{5.4}$$

The difficulty in applying this model to a real system lies first in determining r for the particular application and then determining the viscoelastic stiffness to achieve α_{opt}. This is not too difficult for a damped beam structure by assuming a deflection shape and integrating the strain energies that correspond to k_0, k_1 and k_v. More complicated structures generally require a finite element method such as the modal strain energy method. However, the equations developed from the spring model are useful to augment and double check the modal strain energy approach. Since the real stiffness also varies with α, the real part of (5.4) can serve as an indicator for the optimal design point, and the imaginary part provides an estimate of the achievable damping. This technique is presented next.

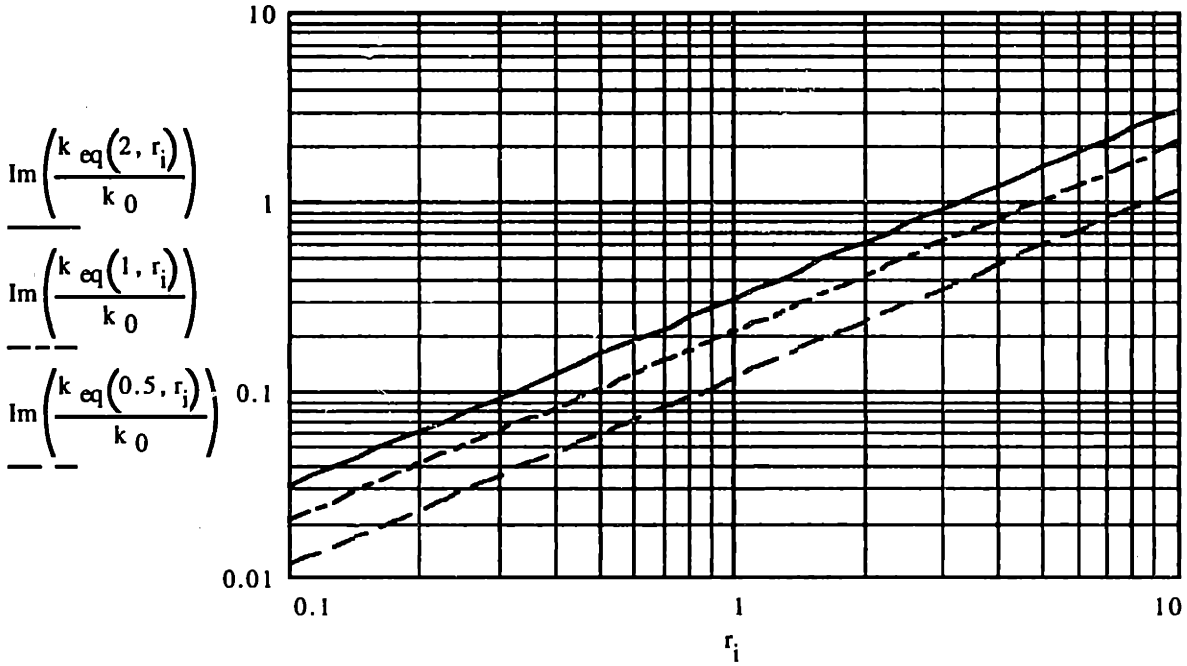


Figure 5-4 Imaginary part of the equivalent stiffness $k_{eq}(\eta, r)$ at α_{opt} versus r for $\eta = 2$, 1 and 0.5.

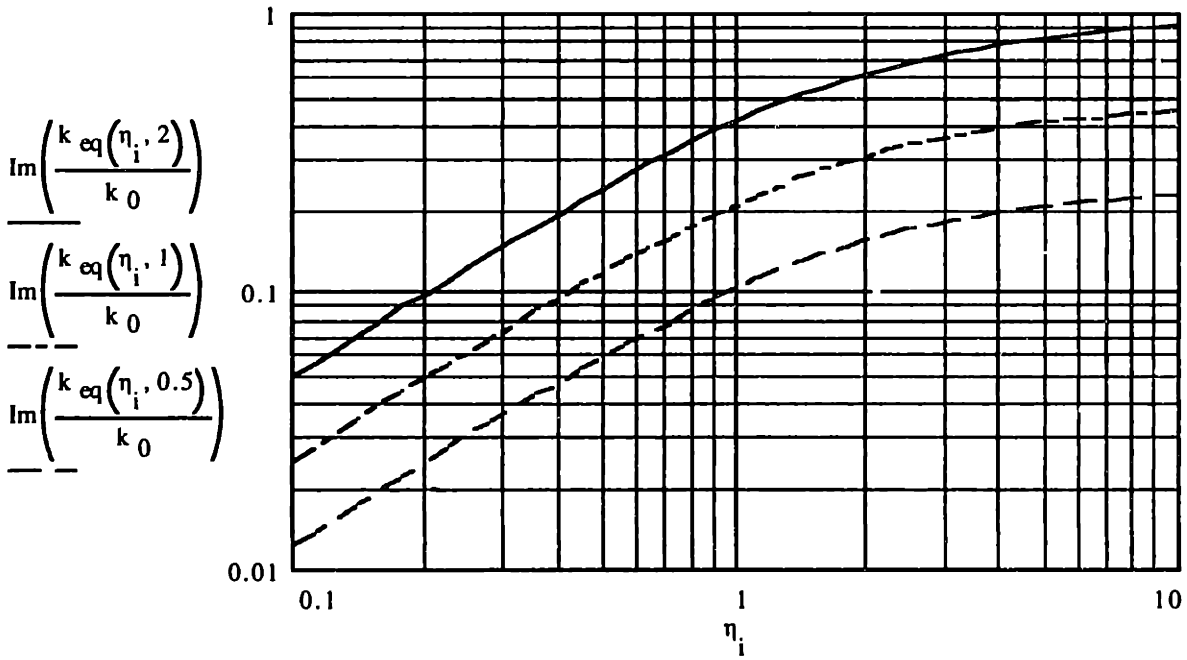


Figure 5-5 Imaginary part of the equivalent stiffness $k_{eq}(\eta, r)$ at α_{opt} versus η for $r = 2$, 1 and 0.5.

The modal strain energy method requires a finite element model of the damped structure, but there is no need for damping elements or complex moduli. Typically, the structure and the constraining layer are modeled with shell elements. The viscoelastic sheet may be modeled with solid elements or spring elements between nodes. The analysis may be static or modal depending on the application. The design goal is to find the viscoelastic shear modulus and thickness that maximize the proportion of strain energy in the VEM,

which is analogous to finding α_{opt}. The method is somewhat brute force in that the analysis must run many times to obtain a plot like Figure 5-3 and nothing ensures that the optimum will be good enough once it is found. This is where Equation 5.4 is effective in combination with FEA. Specific examples of FEA-based design of damped structures occur later in this chapter (see Section 5.3 *Damping Experiments*). A generic description of the steps involved should clarify the method and show that it is conceptually simple and easy to implement in finite elements.

- The first step is to develop a concept for the structure and how the damper(s) might be incorporated. An analysis of the bare structure is useful to identify the particular mode shapes and frequencies that require damping. It will also show areas of high strain energy that are good locations to reinforce with constrained-layer dampers.
- The next step is to add the constraining and viscoelastic layers to the model using a very low shear modulus for the VEM, perhaps 10 psi. Most FEA codes require the elastic modulus E and the Poisson ratio rather than the shear modulus G. The correct Poisson ratio is 0.49 to approximate an incompressible material, which gives $E = 3\ G$. The results of the analysis should show that the constraining layer elements move with the structure but are practically unstressed due to the low shear modulus of the VEM. This analysis will help show any problems with the model and will give a stiffness proportional parameter for k_0, either inverse strain energy or squared frequency.
- The next step requires repeating the analysis with a very high shear modulus for the VEM, perhaps 10 million psi. This analysis will give a stiffness proportional parameter to use for k_∞. Using k_0 and k_∞ to determine r, estimates can be found from Equation 5.4 for the maximum loss factor and the stiffness parameter that indicates the optimum. This step provides quick feedback on the effectiveness of the constraining layer.
- The next step requires changing the shear modulus of the VEM to locate the optimal point where the proportion of strain energy in the VEM is maximum. This is a trial-and-error procedure, which converges more rapidly with the aid of the stiffness indicator. The loss factor of the damped structure is simply the material loss factor of the VEM times the proportion of strain energy in the VEM.
- The last step is to determine the appropriate VEM to use based on the analysis. Manufacturers of VEM will provide temperature-frequency plots of the loss factor and the real part of the complex shear modulus. The best material has the highest loss factor and the correct shear modulus for the expected temperature and frequency. The shear modulus found using FEA corresponds to the real part of the complex shear modulus. It is also useful to know that the shear stiffness of the viscoelastic sheet is proportional to the area and inversely proportional to the thickness. Often a change to these other parameters will allow a better match between the analysis and the available material.

Damped structures that are beam-like can be designed without FEA by relating k_0, k_1 and k_ν to the strain energy in an assumed deflection shape. This approach is similar to the RKU theory for three-layer beams composed of a structural layer, viscoelastic sheet and constraining layer, and the approach by [Ungar, 1962] for more general three-component beams.[I] The basic approach is to consider the composite beam in two extreme conditions where the viscoelastic shear modulus is either zero or infinity. Although the actual deflection shape will change somewhat with viscoelastic stiffness, the analysis requires the assumed deflection shape to be constant so that stiffness and strain energy remain proportional. Then for each extreme, the strain energy calculated for the composite beam is used in place of k_0 or k_∞, respectively.[II] Similarly, k_ν is proportional to the strain energy in the VEM when no extension takes place in the constraining layer(s) or the beam. Applying these same steps to the spring model verifies that the different strain energies are equivalent to the corresponding stiffnesses.

As Equations 5.5 and 5.6 show, the strain energy for either extreme, u_0 or u_∞, depends on the curvature of the assumed deflection shape $\phi(x)$ and differs only by the bending stiffness term EI_0 or EI_∞ for the composite beam.[III] The ratio r calculated to estimate the maximum damping is independent of the deflection shape and the beam length. The expression for r in Equation 5.7 results from using the familiar formulas for built-up beams, where the index i indicates the beam ($i = 0$) and one or more constraining layers ($i = 1, 2, 3$, etc.). The variable c_i is the location of an individual centroid (or neutral axis) with respect to an arbitrary reference such as the beam's centroid. It is a significant advantage to spread the constraining layers as far as possible from the beam's centroid because of the squared effect on r.

$$k_0 \propto u_0 = \frac{1}{2} EI_0 \int_0^L \left(\frac{d^2\phi}{dx^2} \right)^2 dx \tag{5.5}$$

$$k_\infty \propto u_\infty = \frac{1}{2} EI_\infty \int_0^L \left(\frac{d^2\phi}{dx^2} \right)^2 dx \tag{5.6}$$

$$r = \frac{EI_\infty - EI_0}{EI_0} = \frac{\sum_i E_i A_i (c_i - \bar{c})^2}{\sum_i E_i I_i} \qquad \bar{c} = \frac{\sum_i E_i A_i c_i}{\sum_i E_i A_i} \tag{5.7}$$

[I] This approach differs in that imaginary stiffness is optimized rather than loss factor and that any number of constraining layers and viscoelastic layers is possible.

[II] For a linear spring, the strain energy as a function of displacement is proportional to the stiffness, whereas the strain energy as a function of load is proportional to the compliance.

[III] The section and material properties of the beam are assumed to be uniform along the length of the beam. If this is not the case, then the x dependent properties must go inside the integral.

As Equation 5.8 shows, the strain energy in the VEM depends on the difference between the slope of the deflection shape and the slope where shear is zero in the VEM, which occurs naturally at a point of symmetry or at a boundary condition where the constraining layer is rigidly attached to the beam. Such a condition occurs, for example, when the constraining layer of a cantilever beam also attaches to the support. It is generally advantageous for the point of zero shear to coincide with the point of maximum bending moment. Without an obvious point of symmetry or a physical attachment, the point of zero shear is found by minimizing the integral in (5.8) or simply by guessing. The squared relationship of the constraining layer spacing is similar to but subtly different from (5.7) in that the reference is to c_0 rather than $\bar{c}$. The distinction is not so significant considering the approximate nature of this analysis; rather it eliminates having to account for the viscoelastic layers once for each constraining layer and again for the beam. The nondimensional aspect ratio R_i in (5.8) describes the size of the viscoelastic layer between the beam and a particular constraining layer. The aspect ratio is the width in shear divided by the thickness and is analogous to the area of the constraining layer in (5.7). Usually the same VEM is used on all constraining layers in which case the thickness and the shear modulus G_v come outside the summation.

$$k_v \propto u_v = \frac{1}{2}\sum_i G_v R_i (c_i - c_0)^2 \int_0^L \left(\frac{d\phi}{dx} - \left.\frac{d\phi}{dx}\right|_0 \right)^2 dx \tag{5.8}$$

Recall that the design goal is to make r as large as possible within the design constraints or as large as necessary to achieve the required damping. Then the optimal point on the curve indicated by α_{opt} must be calculated and related back to the physical design parameters. Equation 5.9 gives this relationship found by substituting the strain energies into the definition for α. The deflection shape is an important part of this relationship as it determines the effective length of the viscoelastic layer in shear and the constraining layer in tension. The proper deflection shape to use depends on the inertial load and the boundary conditions. When the mass is uniform along the length of the beam, the proper deflection shape is the theoretical mode shape for an Euler beam having the appropriate boundary conditions. Equation 5.10 gives the mode shape for a simply supported Euler beam and the integrated expression for the effective length, where n is the mode number. Euler mode shapes for other boundary conditions are readily available but are sufficiently complex that numerical integration is more practical for calculating the effective length. Table 5-1 gives some common Euler mode shapes and the effective length computed for each. When there is significant mass concentrated at one or more locations along the length of the beam, then the static deflection shape gives a better approximation to the actual mode shape. The effective length is calculated by superimposing the deflection shape proportional to each mass load and the distributed mass of the beam. Then it is only a matter of adjusting the shear modulus and/or the aspect ratio in (5.9) until α is equal to α_{opt} calculated with (5.3).

$$\alpha = \frac{\mathrm{Re}(u_v)}{u_\infty - u_0} = \frac{\mathrm{Re}(G_v)\sum_i R_i \left(c_i - c_0\right)^2}{\sum_i E_i A_i \left(c_i - \bar{c}\right)^2} {L_{eff}}^2 \tag{5.9}$$

$${L_{eff}}^2 \equiv \int_0^L \left(\frac{d\phi}{dx} - \left.\frac{d\phi}{dx}\right|_0 \right)^2 dx \Big/ \int_0^L \left(\frac{d^2\phi}{dx^2} \right)^2 dx$$

$$\phi(x) = \sin\left(n\,\pi \frac{x}{L} \right) \quad \Rightarrow \quad L_{eff} = \frac{L}{n\,\pi} \tag{5.10}$$

End Condition	*Fundamental Mode Shape*	*Zero Shear*	*Effect. Lgth*
Fixed-Free	$\cosh\left(1.875\frac{x}{L}\right) - \cos\left(1.875\frac{x}{L}\right) - 0.734\left(\sinh\left(1.875\frac{x}{L}\right) - \sin\left(1.875\frac{x}{L}\right)\right)$	Fixed end	0.613 *L*
"	"	Free end	0.314 *L*
"	"	0.4 *L*	0.229 *L*
Pinned-Pinned	$\sin\left(\pi\frac{x}{L}\right)$	Center	0.318 *L*
Fixed-Fixed	$\cosh\left(4.73\frac{x}{L}\right) - \cos\left(4.73\frac{x}{L}\right) - 0.983\left(\sinh\left(4.73\frac{x}{L}\right) - \sin\left(4.73\frac{x}{L}\right)\right)$	Center	0.158 *L*
Free-Free	$\cosh\left(4.73\frac{x}{L}\right) + \cos\left(4.73\frac{x}{L}\right) - 0.983\left(\sinh\left(4.73\frac{x}{L}\right) + \sin\left(4.73\frac{x}{L}\right)\right)$	Center	0.314 *L*

Table 5-1 This table gives the effective length to use in Equation 5.9 for an Euler beam with various end conditions and points of zero shear.

5.2 Squeeze-Film Damping

Squeeze-film damping occurs in fluid-film bearings such as hydrostatic, hydrodynamic and well-lubricated plane ways when the fluid film changes thickness in response to a changing load. Fluid-film bearings have a natural advantage in this regard over rolling-element bearings, but often the advantage is not fully realized because the bearing is too stiff, loosely speaking, compared to the structure. An optimal relationship exists between the structure and the squeeze-film damper, which is very similar to the constrained-layer damper of the previous section. There, the relationship was between the constraining layer and viscoelastic layer, which act in series. In this case the structure and the damper act in series as represented in Figure 5-6 and Equation 5.11.

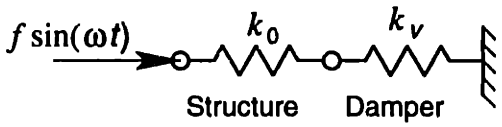


Figure 5-6 Spring model of a structure supported on a viscoelastic damper. The structure has a stiffness k_0 that acts in series with the damper. The damper stiffness is a complex number to represent the viscous damping mechanism in parallel with an elastic component.

$$k_{eq} = \frac{k_0\, k_v}{k_0 + k_v} \qquad \frac{k_v}{k_0} \equiv \alpha(1 + i\,\eta) \tag{5.11}$$

As before, k_0 represents the structural stiffness and k_v is a complex stiffness that represents the viscoelastic behavior of a fluid-film bearing or perhaps a viscoelastic support pad. The nondimensional parameter α indicates the relative stiffness of the damper and the loss factor η is frequency dependent.[1] Equation 5.12 shows the equivalent stiffness k_{eq} separated into real and imaginary parts, and Figure 5-7 shows each part plotted as a function of α for $\eta = 1$. As before, the real part of the stiffness increases with α, and the imaginary part, which represents the damping, has a maximum that coincides with the real part being midway between zero and k_0 in this case.

$$\frac{k_{eq}}{k_0} = \left[\frac{\alpha(1 + i\,\eta)}{1 + \alpha(1 + i\,\eta)}\right]\frac{1 + \alpha(1 - i\,\eta)}{1 + \alpha(1 - i\,\eta)} = \frac{\alpha + \alpha^2(1 + \eta^2) + i\,\eta\,\alpha}{1 + 2\,\alpha + \alpha^2(1 + \eta^2)} \tag{5.12}$$

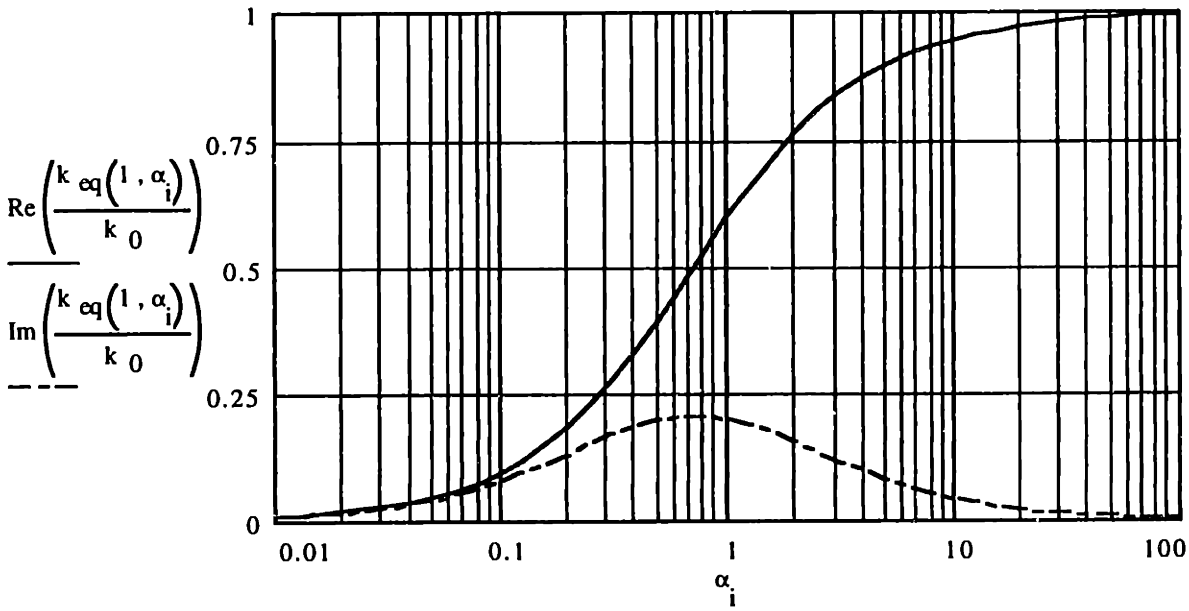


Figure 5-7 Real and imaginary parts of the equivalent stiffness $k_{eq}(\eta, \alpha)$ versus α for $\eta = 1$.

We will derive two optimal points for α because there is a compromise situation that exists between static stiffness and damping. The first optimum designated α_{opt1} is the point of maximum imaginary stiffness and the result given by Equation 5.13 is identical to the optimum for the constrained-layer damper. The equivalent stiffness at α_{opt1} given by Equation 5.14 is similar to Equation 5.4 for the constrained-layer damper. The second optimum designated α_{opt2} comes from maximizing the product of the imaginary stiffness times the static stiffness, as Equation 5.15 shows. This optimum given by Equation 5.16

[1] The loss factor is approximately proportional to the excitation frequency until elastic effects cause a roll off at higher frequencies. This behavior is similar to a typical viscoelastic material.

should achieve a reasonable balance between the two. As Figure 5-8 shows, the second optimum requires approximately twice the stiffness in the damper, which results in greater static stiffness in Figure 5-9 but less damping in Figure 5-10. In addition, Figure 5-11 shows that an overdamped squeeze film ($\eta > 2$) is undesirable since the product of imaginary stiffness times static stiffness for either optimum decreases beyond this point.

$$\mathrm{Im}\left(\frac{\partial k_{eq}}{\partial \alpha}\right)=0 \quad \Rightarrow \quad \alpha_{opt1}=\left(1+\eta^2\right)^{-\frac{1}{2}} \tag{5.13}$$

$$\left.\frac{k_{eq}}{k_0}\right|_{opt1}=\frac{1}{2}+\frac{i\,\eta}{2+2\sqrt{1+\eta^2}} \tag{5.14}$$

$$\frac{\partial}{\partial \alpha}\left(\frac{\eta\,\alpha}{1+2\,\alpha+\alpha^2\left(1+\eta^2\right)}\cdot\frac{\alpha}{1+\alpha}\right)=0 \tag{5.15}$$

$$\alpha_{opt2}=\left(1+\eta^2\right)^{-\frac{1}{2}}\left[\left(\sqrt{1+\eta^2}+\eta\right)^{\frac{1}{3}}+\left(\sqrt{1+\eta^2}-\eta\right)^{\frac{1}{3}}\right] \tag{5.16}$$

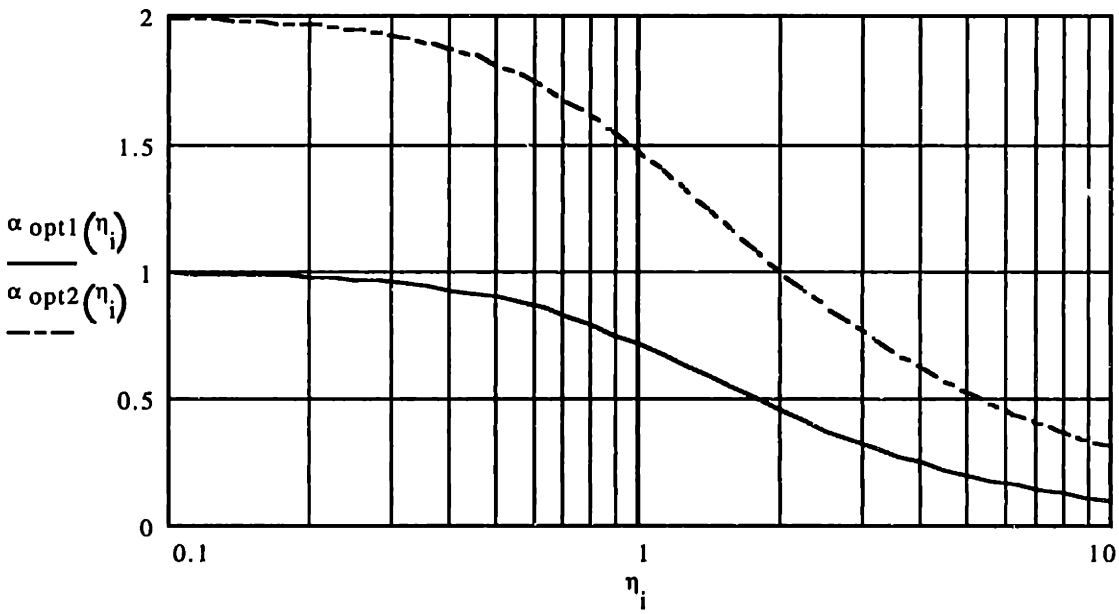


Figure 5-8 The first optimum α_{opt1} maximizes the imaginary part of the equivalent stiffness. The second optimum α_{opt2} maximizes the product of the imaginary stiffness times the static stiffness.

This model provides the optimal stiffness to design into the physical damper mechanism as a function of its loss factor, but the loss factor usually depends on the design parameters of the damper. This situation may require the designer to estimate the loss factor and apply an iterative approach. In addition, the designer must determine whether to optimize damping alone (α_{opt1}) or a combination of damping and static stiffness (α_{opt2}).

An example where static stiffness is not important is a machine tool supported kinematically on viscoelastic pads. A reasonable material loss factor for this case is $\eta = 0.5$, which gives $\alpha_{opt1} = 0.89$ and a system loss factor of 0.24. An example where damping and static stiffness are both important is a slide system where the bearings act as dampers. Rolling element bearings have relatively little damping so that $\alpha_{opt2} \approx 2$. Hydrostatic bearings designed for a loss factor of $\eta = 2$ (critical damping) are optimal when $\alpha_{opt2} = 1$, which results in a system loss factor of approximately 0.5.

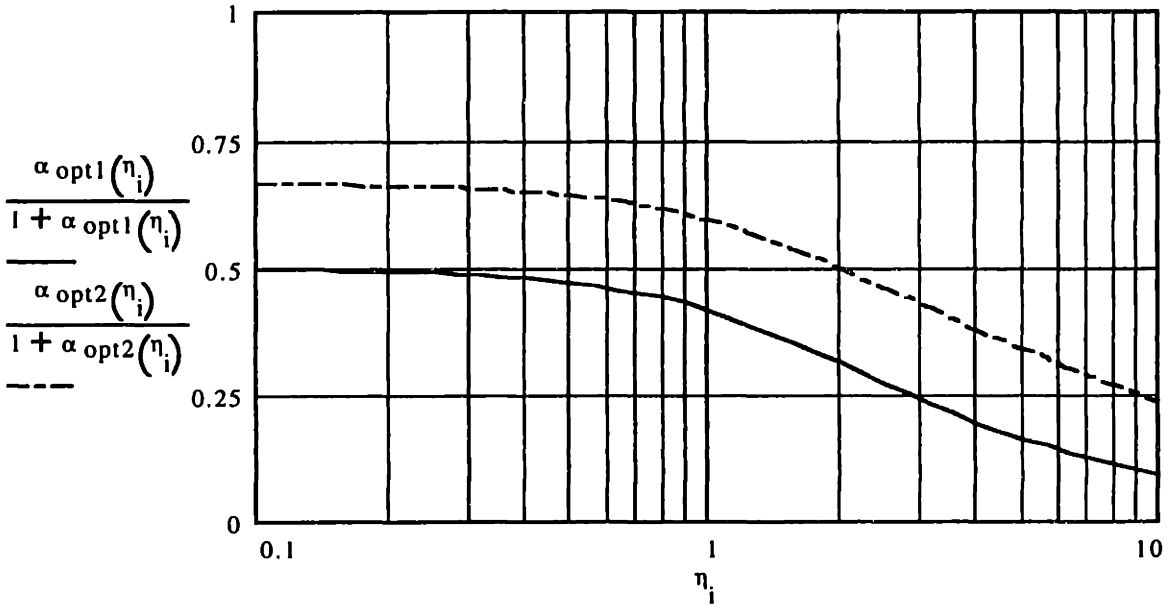


Figure 5-9 Static stiffness of the system for α_{opt1} and α_{opt2} versus η. This plot is normalized to the stiffness of the structure k_0.

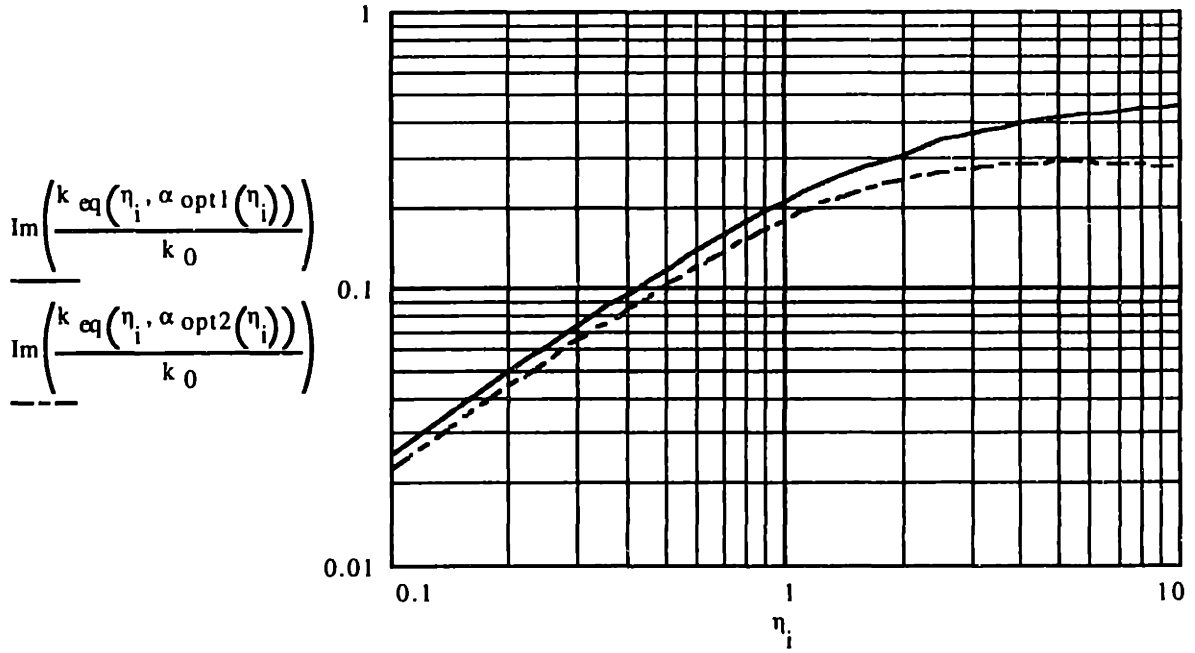


Figure 5-10 Imaginary part of the equivalent stiffness $k_{eq}(\eta, \alpha)$ at α_{opt1} and α_{opt2} versus η.

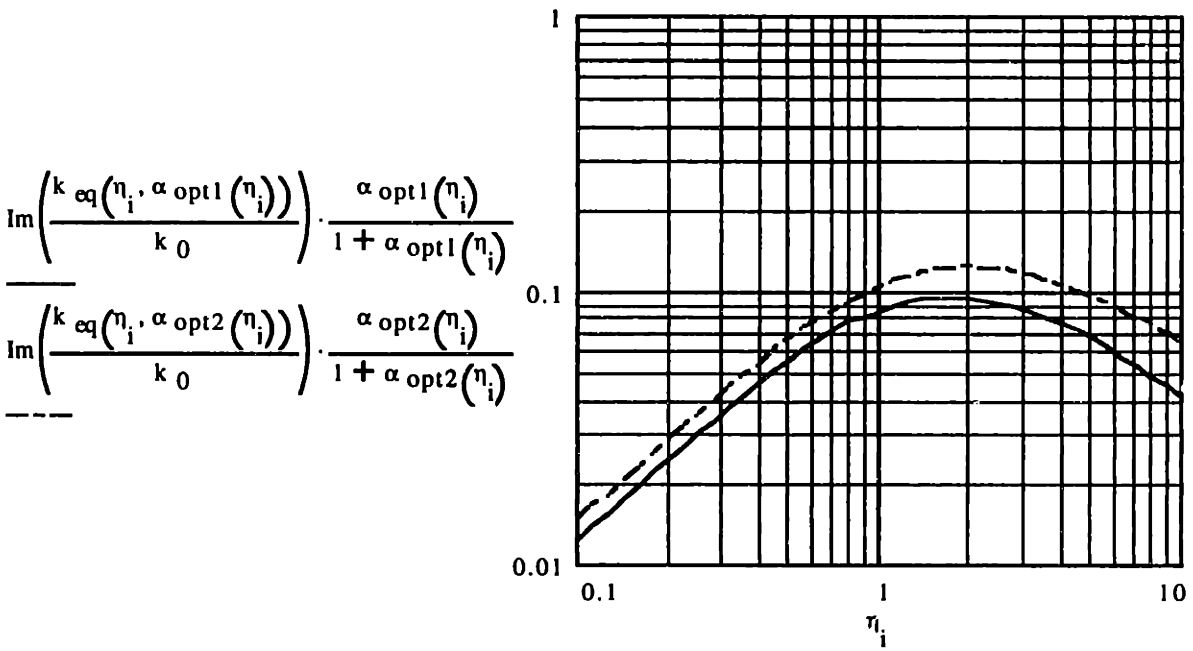


Figure 5-11 Product of the imaginary stiffness times the static stiffness at α_{opt1} and α_{opt2} versus η.

It is an unusual practice to model the damping in a fluid-film bearing system even though, as previously shown, there is considerable damping to gain from an optimal design. To aid the design of squeeze-film dampers, models are developed for viscous-dominated flow between two plates. The models are not difficult to develop assuming the flow is unidirectional with a fully developed parabolic velocity profile between the plates. In areas where the flow is two directional, one-dimensional solutions are superimposed in preference to deriving two-dimensional solutions. The accuracy of this approach should be adequate since the flow over most of the area is one dimensional. Each model is represented as a complex stiffness, where the real part represents the stiffness and the imaginary part represents the damping. In every case the damping is proportional to the excitation frequency ω and the fluid viscosity μ, and inversely proportional to the cube of the film thickness h.

Figure 5-12 shows the geometric parameters for a squeeze-film damper in the shape of an annulus. The fluid flows radially whenever the film thickness changes and energy is dissipated in the viscous flow. Equation 5.17 gives the exact expression (top) and the approximation (bottom), which shows that the damping is proportional to the mean circumference and the cube of the annular width. Equation 5.18 gives the exact expression for the special case of a circular pad where $r_1 = 0$. Equation 5.18 may be applied to an annulus with a closed pocket by subtracting the damping lost when the pocket is removed from a solid pad, as in the manner of calculating the area or the moment of inertia.

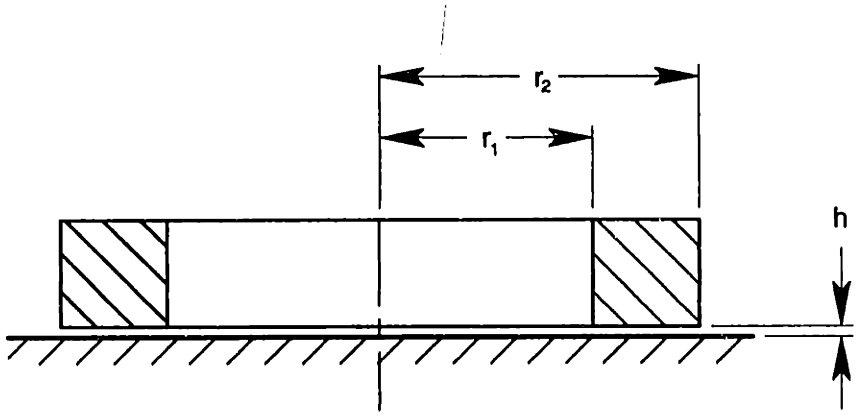


Figure 5-12 Annular squeeze-film damper.

$$k(\omega) = i\,\omega\,\mu\,\frac{3\pi}{2h^3}\left[r_2^{\,4} - r_1^{\,4} - \frac{\left(r_2^{\,2} - r_1^{\,2}\right)^2}{\ln(r_2/r_1)}\right] \tag{5.17}$$

$$\cong i\,\omega\,\mu\,\pi\left(r_2 + r_1\right)\left(\frac{r_2 - r_1}{h}\right)^3 \qquad \text{for} \quad r_2 \sim r_1$$

$$k(\omega) = i\,\omega\,\mu\,\frac{3\pi}{2h^3}\,r_2^{\,4} \qquad \text{for} \quad r_1 = 0 \tag{5.18}$$

Figure 5-13 shows the geometric parameters for a rectangular damper. The fluid flows across the width except near the ends where it becomes two dimensional. Equation 5.19 gives the exact solution for one dimensional flow, which is a reasonable approximation for a long rectangle. As the rectangle becomes square, the flow becomes approximately radial. Therefore, Equation 5.18 may be used with a radius that gives the same area as the square and Equation 5.20 shows this result. For cases near square where the flow is clearly two-dimensional, a series combination of stiffnesses calculated for unidirectional flow in two orthogonal directions gives a more reasonable estimate. Equation 5.21 shows the case for a general rectangle, which closely matches the expression for a square and asymptotically matches the expression for a long rectangle.

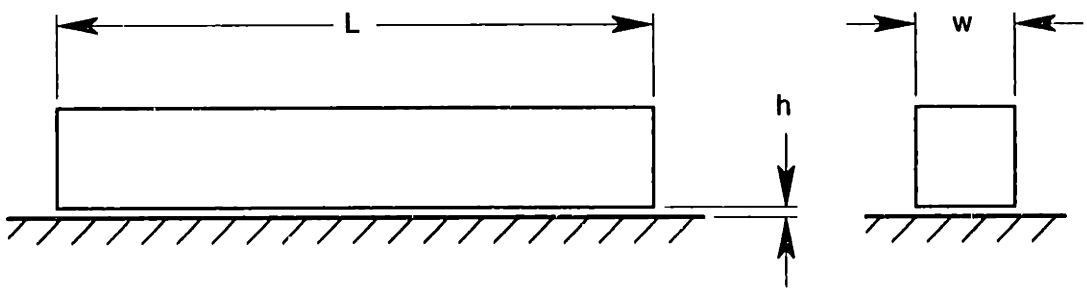


Figure 5-13 Rectangular squeeze-film damper.

$$k(\omega) \cong i\,\omega\,\mu\,L\left(\frac{w}{h}\right)^3 \qquad \text{for} \quad L >> w \tag{5.19}$$

$$k(\omega) = i\,\omega\,\mu\,\frac{3}{2\pi h^3}\,w^4 \qquad \text{for} \quad L = w \tag{5.20}$$

$$k(\omega) \cong i\,\omega\,\mu\left(\frac{L\,w}{h}\right)^3 \frac{1}{L^2 + w^2} \tag{5.21}$$

Figure 5-14 shows the geometric parameters for a cylindrical damper. Fluid will flow predominantly in either an axial or a tangential direction if the dimensions are disparate or if the flow is constrained by a seal. Equation 5.22 gives the exact solution for one-dimensional flow in the axial direction. Equation 5.23 gives the exact solution for one-dimensional flow in the tangential direction. When unidirectional flow is not a good approximation, the technique using a series combination of stiffnesses, as discussed for the rectangular squeeze film, is appropriate.

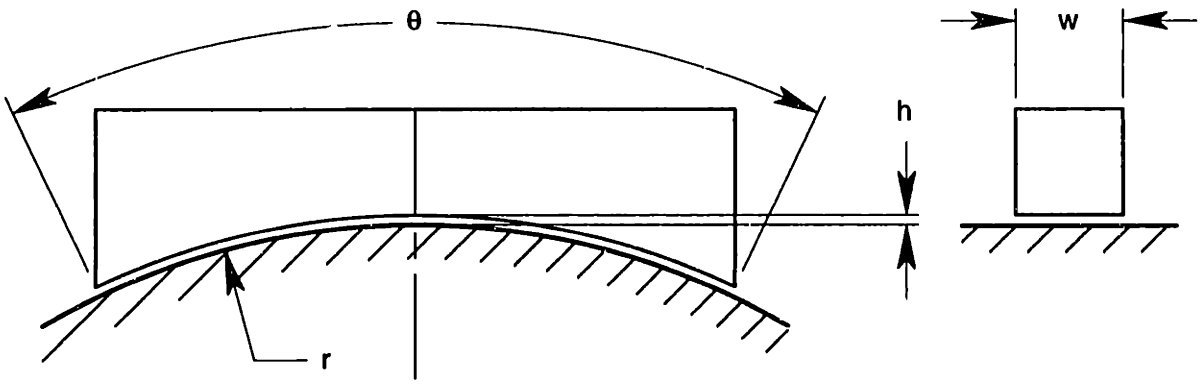


Figure 5-14 Cylindrical squeeze-film damper.

$$k(\omega) = i\,\omega\,\mu\,r\left(\frac{w}{h}\right)^3 \frac{\theta + \sin(\theta)\cos(\theta)}{2} \cong i\,\omega\,\mu\,r\,\theta\left(\frac{w}{h}\right)^3 \tag{5.22}$$

$$k(\omega) = i\,\omega\,\mu\,w\left(\frac{r}{h}\right)^3 6\left[\theta - \sin(\theta)\right] \cong i\,\omega\,\mu\,w\left(\frac{r\,\theta}{h}\right)^3 \tag{5.23}$$

Figure 5-15 shows the geometric parameters for a circular hydrostatic bearing. An external supply pressure P_s causes fluid to flow into the bearing through a laminar-flow restriction intended to regulate the volume flow rate. It is common practice to size the restriction so that the nominal pressure in the annular groove is a particular ratio γ of P_s, typically one-half for maximum stiffness. The fluid flows radially outward through the bearing gap h as the pressure falls to zero at the outer edge. The pressure under the inner pad is nominally at the groove pressure and differs due to the squeeze-film effect. Equation 5.24 gives the exact expression for the dynamic stiffness. The real part is the familiar expression for static stiffness. The imaginary part comes from three sources: the trapped volume captured by the bearing, the squeeze film under the annular land, and the squeeze film of the inner pad unless there is a full pocket. Typically the trapped volume is more significant than the annular squeeze film and less significant than the squeeze film of the

inner pad if it exists. The static load carried by the bearing is simply the groove pressure times the effective area as indicated in Equation 5.25.

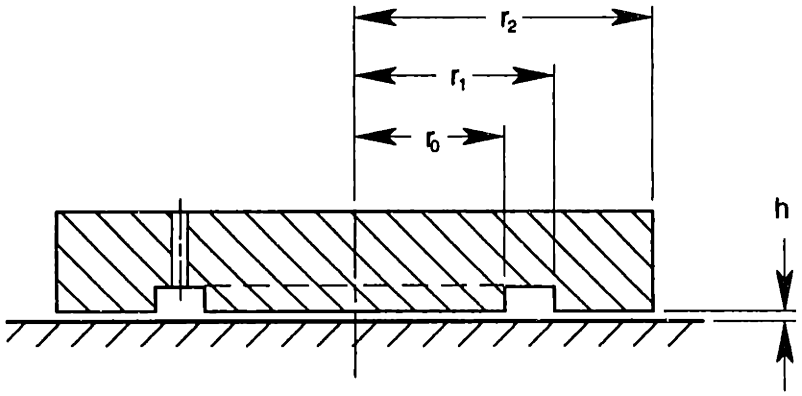


Figure 5-15 Circular hydrostatic bearing. The restriction has a fixed resistance set so that the nominal pressure in the groove is γP_s.

$$k(\omega) = P_s \gamma(1-\gamma)\frac{3\pi\left(r_2^2 - r_1^2\right)}{2h\ln(r_2/r_1)} + i\omega\mu\frac{3\pi}{2h^3}\left[r_2^4 - r_1^4 - \gamma\frac{\left(r_2^2 - r_1^2\right)^2}{\ln(r_2/r_1)} + r_0^4\right] \tag{5.24}$$

$$f = \gamma P_s A_{eff} = \gamma P_s \frac{\pi\left(r_2^2 - r_1^2\right)}{2\ln(r_2/r_1)} \tag{5.25}$$

It is more common for a hydrostatic bearing to have a full pocket rather than a groove. Assuming a full pocket, Table 5-2 shows a set of reasonable parameters that results in a critically damped bearing.

ω	Modal frequency	628 r/s	100 Hz
μ	Viscosity (water)	0.001 kg/m/s	1.45e-7 lb-s/in^2
P_S	Supply pressure	3.0 MPa	435 psi
γ	Pressure ratio	0.5	0.5
h	Fluid film thickness	10 μm	0.000 4 in
r_2	Outer radius	50 mm	1.97 in
r_1	Inner radius	45 mm	1.77 in
f	Static load	10 000 N	2400 lbf
Re(k)	Static stiffness	1600 N/μm	9 lbf/μin
Im(k)	Imaginary stiffness	3200 N/μm	19 lbf/μm
η	Calculated loss factor	2.0	2.0

Table 5-2 Numerical example of a circular hydrostatic bearing.

Figure 5-16 shows the geometric parameters for a rectangular hydrostatic bearing. It operates in the same manner as the circular hydrostatic bearing. Equation 5.26 for the dynamic stiffness reflects the different geometry of the rectangular hydrostatic bearing.

$$k(\omega) \cong P_s \gamma(1-\gamma)\frac{3}{h}\left(\frac{L_2 w_2 + L_1 w_1}{2}\right) + i\omega\mu\frac{6(1-\gamma)}{h^3}\frac{(L_2 - L_1)(w_2 - w_1)}{L_2^2 - L_1^2 + w_2^2 - w_1^2}\left(\frac{L_2 w_2 + L_1 w_1}{2}\right)^2$$
$$+ i\omega\mu\frac{1}{h^3}\left[\left(\frac{L_2 - L_1}{2}\right)^3 (w_2 + w_1) + (L_2 + L_1)\left(\frac{w_2 - w_1}{2}\right)^3 + \frac{(L_0 w_0)^3}{L_0^2 + w_0^2}\right] \tag{5.26}$$

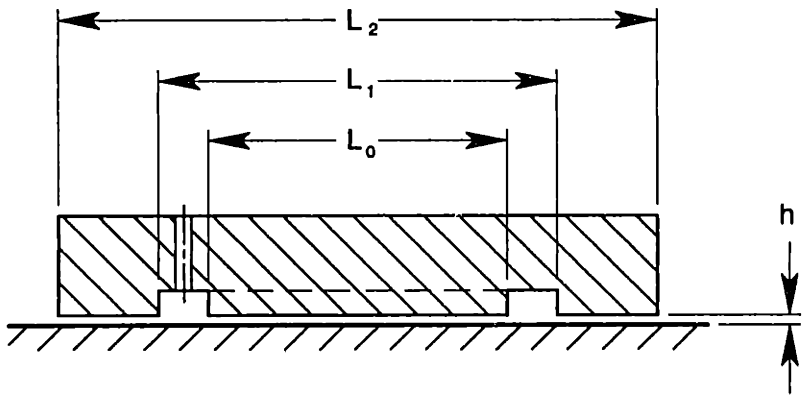


Figure 5-16 Rectangular hydrostatic bearing. The restriction has a fixed resistance set so that the nominal pressure in the groove is γP_s.

No experiments were conducted during this project to verify the models for squeeze-film damping. The general experience with hydrostatic bearings is that the theory is remarkably accurate but errors result from uncertainties in parameters such as the fluid viscosity and the bearing clearance. At the very least these models should provide order-of-magnitude estimates for the design of experiments and help in the interpretation of test data.

5.3 Tuned-Mass Damping

A tuned-mass damper (or vibration absorber) acts to dynamically stiffen a structure over a relatively narrow frequency range near its resonance, where the damper vibrates and dissipates energy that otherwise would excite the structural resonance. For this reason the frequency of the damper must closely match the particular resonance of the structure to be most effective. Unlike the dampers previously discussed, the tuned-mass damper has an optimal loss factor that is relatively modest and depends on the relative size of the tuned mass. Figure 5-17 shows the mass-spring model used in the design of a tuned-mass damper, where the springs are complex to represent structural damping. A simple design approach results by recognizing that the structure and the damper are separate dynamic systems that react in parallel to the excitation force. The dynamic properties of the structure are often determined through experimental modal analysis while the mass and stiffness of the damper are design parameters usually determined analytically. Therefore, being able to treat them as separate systems is a definite advantage over a coupled-system approach.

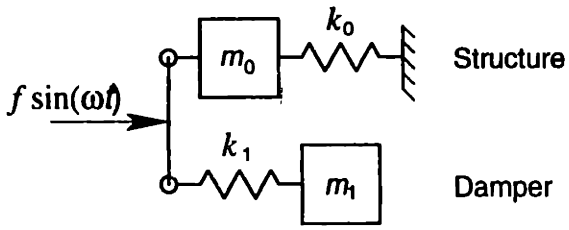


Figure 5-17 The mass-spring model of a structure enhanced with a tuned-mass damper shows that the dynamic stiffness of the structure and the damper react in parallel against an excitation force *f*. Complex springs represent structural damping in the structure and the damper.

Equation 5.27 gives the dynamic stiffness for the single mass-spring system that represents a mode of the structure. This equation is derived in most vibration texts by assuming a steady sinusoidal excitation force applied to the mass. Equation 5.28 gives the dynamic stiffness for the damper derived by assuming a steady sinusoidal excitation force applied to the spring. This equation is similar to the equation for support motion found in most vibration texts. Figure 5-18 shows that the damper behaves somewhat inversely from the structure. When combined in parallel, these dynamic systems complement each other giving much improved dynamic stiffness in the vicinity of the resonance.

$$k_s(\omega) = k_0\left[1-\left(\frac{\omega}{\omega_0}\right)^2 + i\,\eta_0\right] \tag{5.27}$$

$$k_d(\omega) = -k_1(1+i\,\eta_1)\left(\frac{\omega}{\omega_1}\right)^2\left[1-\left(\frac{\omega}{\omega_1}\right)^2 + i\,\eta_1\right]^{-1} \tag{5.28}$$

$$k_d\big|_{\omega>>\omega_1} \cong k_1(1+i\,\eta_1) \qquad k_d\big|_{\omega<<\omega_1} \cong -m_1\,\omega^2$$

As indicated, tuning the resonance of the damper to match the resonance of the structure is the only critical aspect of the design. A poor match leads to poor results as Figure 5-19 plainly shows. An exact match, however, is not quite optimal. The resonant frequency of the damper should be slightly less than that of the structure. An optimum also exists for the loss factor in the damping element. Finding the optimal frequency and loss factor is quite easy by adjusting the parameters in a spreadsheet model. [Den Hartog, 1956] provides analytical derivations of these optimums assuming that the structural resonance has no inherent damping.[1] Equations 5.29 and 5.30 express his results using the symbols from Figure 5-17. Equations 5.31 gives the minimum dynamic stiffness calculated for the optimal frequency and loss factor. Clearly, the damper becomes more effective as its mass and stiffness become larger in proportion to the modal mass and modal stiffness of the

[1] In addition, viscous damping is used rather than complex stiffness but the difference is minor.

structure. A further benefit is apparent in Figure 5-20. The frequency range of the damper is wider since the optimal loss factor is greater.

$$\frac{\omega_1}{\omega_0} = \frac{m_0}{m_0 + m_1} \tag{5.29}$$

$$\eta_1 = \left(\frac{3}{2}\frac{m_1}{m_0}\right)^{\frac{1}{2}}\left(1+\frac{m_1}{m_0}\right)^{\frac{-3}{2}} \tag{5.30}$$

$$\min\left(k_{eq}\right) = k_0\left(\frac{m_1}{m_1 + 2m_0}\right)^{\frac{1}{2}} \tag{5.31}$$

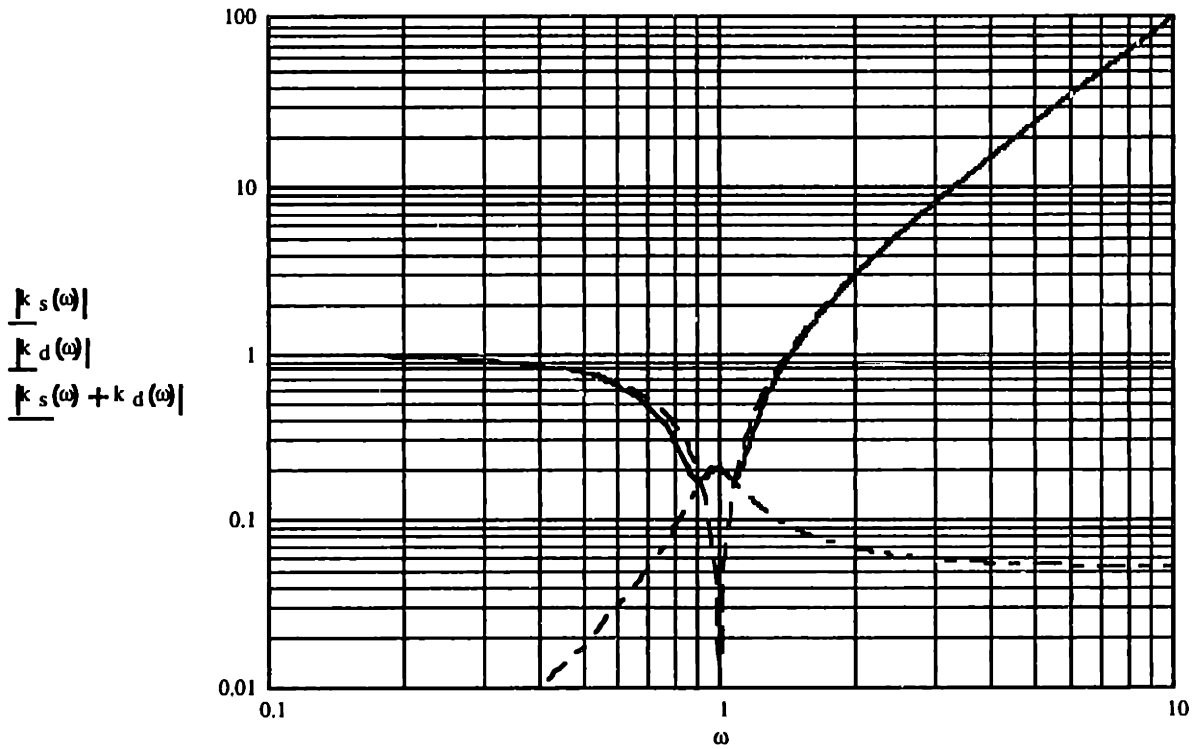


Figure 5-18 The dynamic stiffness of the structure (dash curve) and the dynamic stiffness of the damper (dot-dash curve) combine as indicated by the solid curve. The resonant frequency of the damper is slightly less than the structure, 0.95 versus 1, to cause the two intersection points to occur at approximately the same level of stiffness. The minimum dynamic stiffness for the combined system is 0.169 for an optimal loss factor in the damper of 0.28.

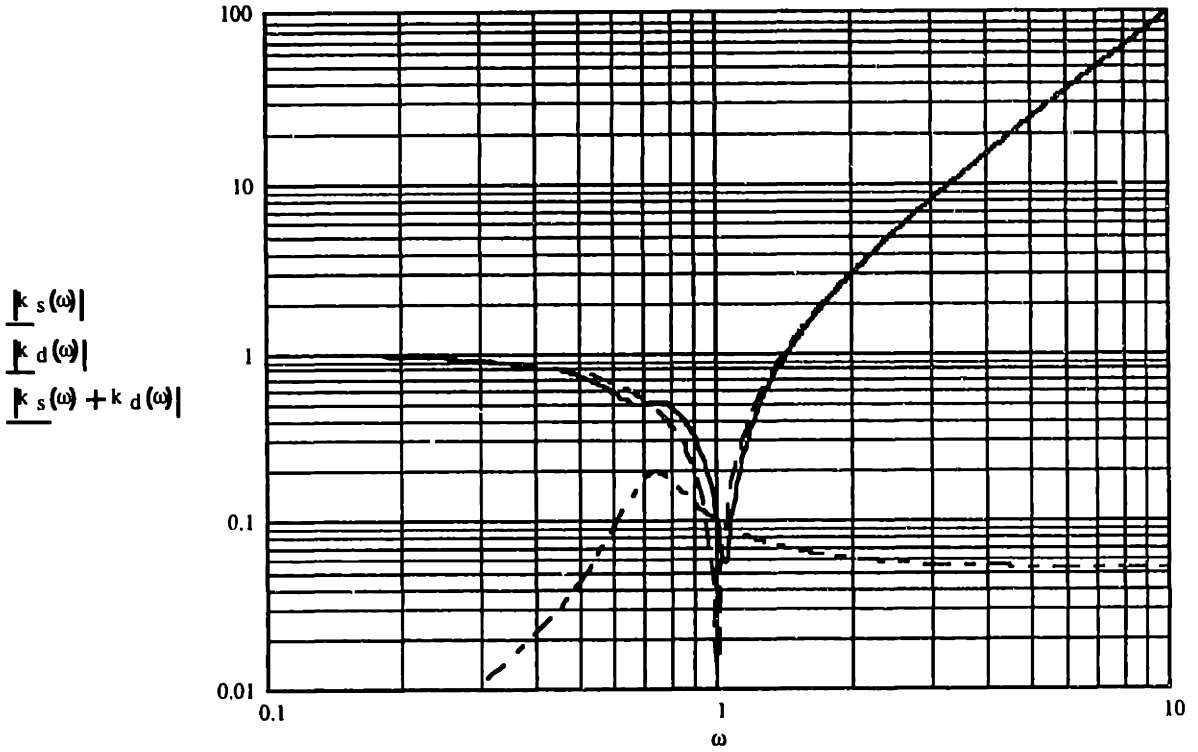


Figure 5-19 With identical parameters except for the 0.707 resonant frequency of the damper, the minimum dynamic stiffness for the combined system falls by a factor of three to 0.057.

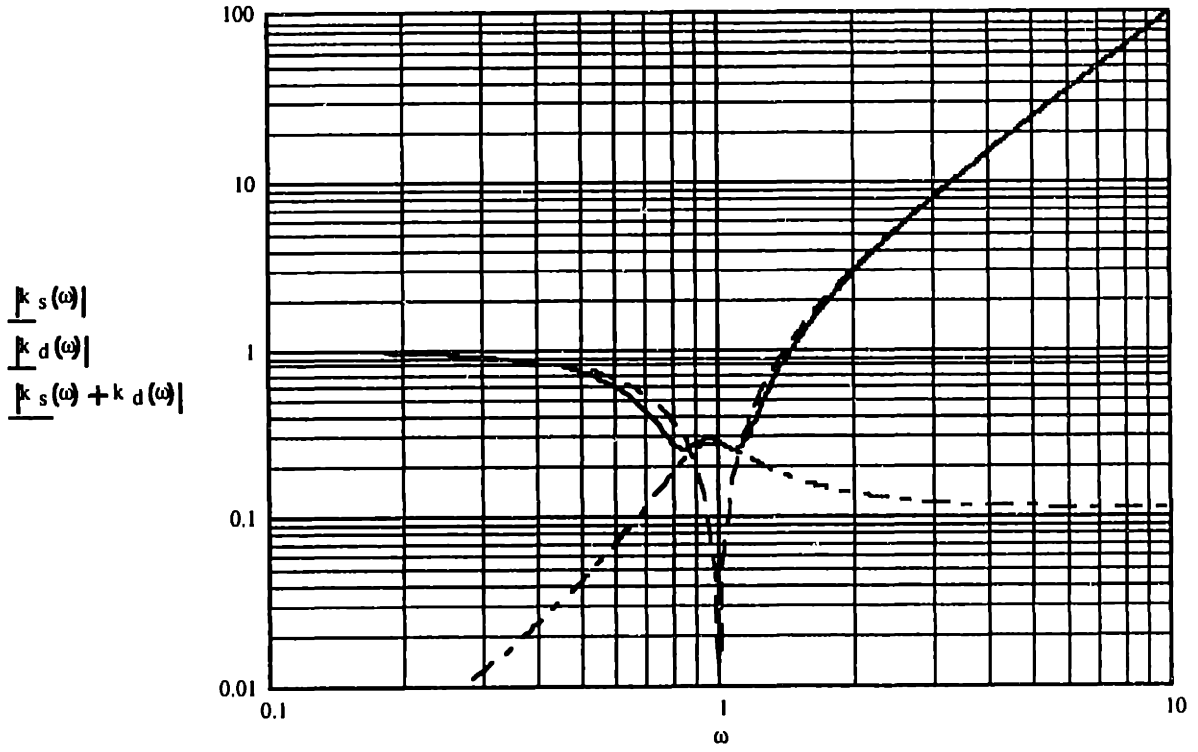


Figure 5-20 A damper twice as large as before has an optimal resonant frequency of 0.885 and an optimal loss factor of 0.45. The minimum dynamic stiffness for the combined system is 0.247, an increase of 46%.

An intriguing way to incorporate tuned-mass damping into a machine tool is by using massive components that for other purposes attach to the structure but have no critical relationship to the metrology of the machine. A number of components might be considered, for example, the spindle motor, tool changer equipment, transformer, hydraulic power unit and guarding. The design effort may be considerably more challenging since the components usually have distributed mass and stiffness, limitations on the location or method of attachment and so forth. Determining the compliance requirements for the mount(s) would require experimenting perhaps with the physical machine or more likely with a dynamic system model in finite elements. Developing separate dynamic stiffness curves as in Figure 5-18 seems a reasonable approach.

5.4 Damping Experiments

Damping is perceived to be an important ingredient for high-precision, high-productivity machine tools, although a minimum requirement is difficult to define. The more-is-better philosophy works up to the point where costs begin to increase faster than the benefit of additional damping. Judging where this point lies is also difficult to define. Certainly any advancement in the performance-to-cost ratio would change the equation and be an advantage worth investigating. In addition, the possibility that inherent damping mechanisms could be lost in the shift toward precision design philosophies is of particular concern; therefore, having ready solutions is a prudent position to take. For example, supporting a machine tool at three points to eliminate stability problems associated with an overconstrained support is a significant departure from the industry norm. Fortunately in this case, the exact-constraint support can incorporate a viscoelastic material to provide damping comparable to a support involving multiple leveling screws.

Several experiments were conducted to gain experience in developing damping treatments for machine tool structures. They served two general purposes: to experimentally verify the analytical design methods and to estimate the benefit of the damping treatment on the machine tool.

5.4.1 Constrained-Layer Damping on the Maxim™ Column

The Cincinnati Milacron Maxim™ tested at LLNL came equipped with a tuned-mass damper designed to improve the first two modes of the column. Presumably the column needs a damping treatment and therefore would be a good test bed to experiment with the constrained-layer damping technique. That is not to say that constrained-layer damping is better than tuned-mass damping but rather is another tool with advantages and disadvantages. For example, a constrained-layer damper is not nearly as frequency sensitive but it reacts only to column distortions rather than column displacements (including bearing deflections) as does the tuned-mass damper.

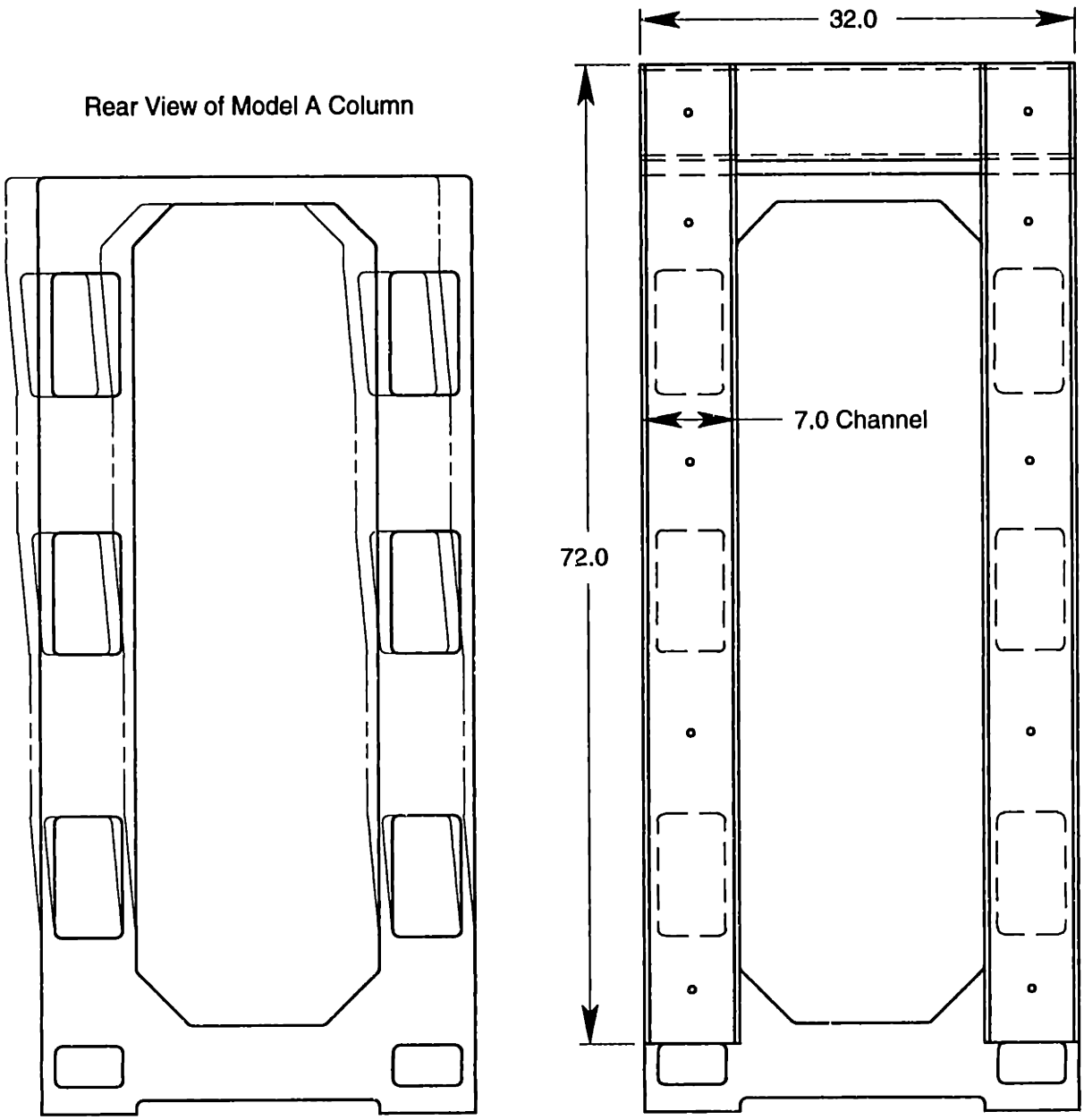


Figure 5-21 Rear view of the Model A column with the shear mode superimposed (left) and the constraining layers applied (right).

The main constraints placed on the damper design have to do with the need to retrofit to an existing column with minimal modifications. In addition we wanted to show that the technique can be low cost. The mode shapes of the column are also constraints as they define where the constraining layer(s) should go. In examining the mode shapes, we determined that shearing of the rear face was the most flexible and therefore would benefit most from constraining layers. Figure 5-21 shows the column with the shear mode superimposed (left) and the constraining layers (right) intended to stiffen the shear mode. The shear mode is not a dynamic mode but rather is a significant compliance in the torsional mode and to a lesser extent in the rocking mode. Constraining layers were constructed from channels and fit to the rear face of the column with the flanges facing away. They were

bonded to the column with a layer of viscoelastic in between and secured with quarter-inch screws near zero-shear points. The horizontal channel at the top was attached to the vertical channels in the same way to provide additional shear area.

We used an FEA-based strain energy method to determine the proper viscoelastic shear modulus and to estimate the loss factor.[I] Our understanding of the FEA program at that time limited the analysis to static deflection shapes rather than modal, and the viscoelastic strain energy was integrated manually from contour plots. In addition, the constraining layers used in the analysis were half inch plates rather than channels and there was no cross piece. The change to channels, however, should increase the dynamic stiffness. We ran four cases for this configuration varying only the viscoelastic shear modulus. Each case had two separate load cases applied to the front face of the column: a Y-moment load to simulate the torsional mode and an X-force to simulate the rocking mode. Table 5-3 shows a summary of the results. The cases for zero and infinite shear modulus provide the information to estimate the optimal parameters in the column labeled *optimal* using Equation 5.4. For the other two cases, 2000 and 5000 psi, the loss factors were calculate by taking the ratio of viscoelastic strain energy to total strain energy, which implies a material loss factor of one. Comparing these results to the estimated optimum, the optimal shear modulus should be somewhat greater than 5000 psi.

Y-Moment applied to front face					
VEM Shear Modulus	zero	infinite	optimal	2000	5000
Total Strain Energy	8.66	5.49	6.72	7.60	7.14
VEM Strain Energy	0	0	0.62	0.58	0.58
Calc'd Loss Factor	0	0	0.093	0.077	0.082
X-Force applied to front face					
VEM Shear Modulus	zero	infinite	optimal	2000	5000
Total Strain Energy	17.29	15.17	16.16	16.69	16.47
VEM Strain Energy	0	0	0.44	-	-
Calc'd Loss Factor	0	0	0.027	-	-

Table 5-3 Finite element results for the Model A column with a varying viscoelastic shear modulus.

The thickness of the viscoelastic layer used in the finite element analysis is somewhat arbitrary because the shear stiffness scales inversely with the thickness. The 0.020 inch (0.5 mm) thick layer used in the analysis turned out to be the same thickness used for the damping treatment. The product that we chose to use is called DYAD 606 and is available from Soundcoat.[II] It is available in 0.020 or 0.050 inch thickness and costs, respectively, 6 or 11 dollars per square foot. They also supply B-Flex epoxy, which seems

[I] ProMechanica from Parametric Technologies, Inc. is the finite element software used in this study.

[II] Soundcoat, 3002 Croddy Way, Santa Ana, CA 92799-9202.

to work very well. Figure 5-22 shows the data sheet for DYAD 606. To use this chart, first locate the intersection between the operating temperature (diagonal lines) and the excitation frequency. From the intersection, draw a vertical line to intersect the shear modulus curve and the loss factor curve. Using a temperature of 20° C and a frequency of 60 Hz, the shear modulus is approximately 41 MPa or 6000 psi and the loss factor is close to one. This material appears to be a good choice based on the analysis.

The application of the damping treatment to the column was very simple once the constraining layers had been fit and the column tapped for screws. The mating surfaces were well cleaned with solvent but the paint was left on the column. The VEM was cut with excess that could be trimmed later. Epoxy was painted onto the column and the VEM was applied over top. Epoxy was applied to the VEM so that the channels could be handled without mess. The channels were put into place and secured with screws to force excess epoxy out the gaps. Each channel also had a jacking screw installed near the bottom to aid in removing the dampers after the tests were complete. The bond was quite strong and required a concerning level of torque on a 5/16 screw and several minutes before a loud pop signaled a rupture in the bond.

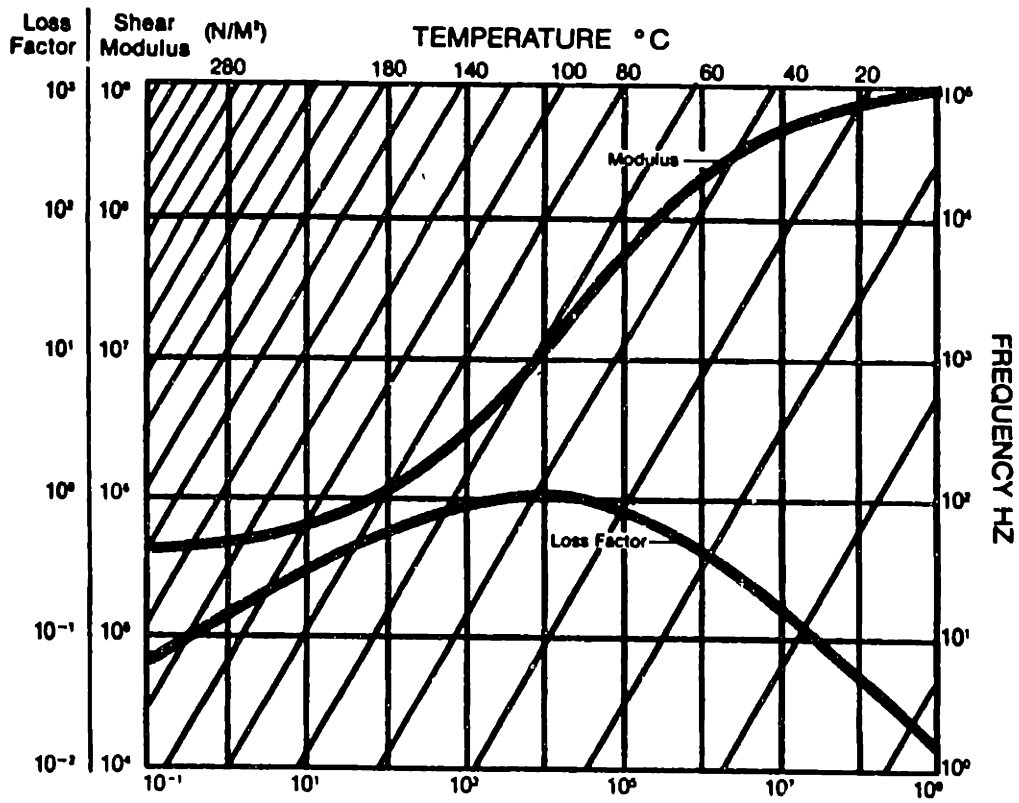


Figure 5-22 Shear modulus and loss factor curves for DYAD 606 from Soundcoat. Locate the intersection between the operating temperature (diagonal lines) and the excitation frequency. From the intersection, draw a vertical line to intersect the shear modulus curve and the loss factor curve.

5.4.2 Dynamic Compliance Tests on the LLNL Maxim™ Column

The Engineering Measurements and Analysis Section of LLNL conducted dynamic compliance tests on the Maxim to judge the benefits of the tuned-mass damper and the viscoelastic constrained-layer damper previously described.

The dynamic compliance was measured in the X-Y plane of the machine. The apparatus consisted of an electromagnetic shaker, two capacitance gages and a number of accelerometers.[1] The shaker was mounted to the work table in either the X or Y direction and attached to a one-inch bar in the spindle. Two capacitance gauges were supported in a separate fixture relative to the work table and they measured the X and Y motions of the one-inch tool holder near the spindle face. Accelerometers were placed on the column, spindle, work table and base. For the configuration with the tuned-mass damper, the dynamic compliance was measured in both the vertical and lateral directions. In addition, the cross compliance was measured but there was little interaction. Only the lateral compliance was measured in the other configurations. Compared to the undamped configuration, the tuned-mass damper improved the first rocking mode approximately 20% and the second torsional mode approximately 13%. It also introduced a much higher mode that was not present in the other configurations. The spindle mode became approximately 5% worse. The constrained-layer damper had no appreciable effect on the rocking mode but it improved the torsional mode approximately 13%. While both damping treatments show moderate improvement, it would appear that the greatest benefit would come from a damping treatment applied to the spindle!

[1] The accelerometers were used to determine mode shapes.

intentionally blank

6 Practical Exact-Constraint Design

The basic concepts of kinematics and exact-constraint design are presented in Section 2.6 following the 12 statements from [Blanding, 1992]. This chapter brings those concepts closer to reality by considering various constraint devices and the many ways that constraints may be arranged. Several analytical studies on flexures provide deeper understanding into the particulars of flexure design. A new approach for kinematic-coupling design optimizes the ability of the coupling to overcome friction and become centered. The approach is most useful for unusual, nonsymmetric configurations where intuition is inadequate. The sometimes complex spatial relationships between constraints, whether for flexures or in couplings, soon become insurmountable unless systematic, matrix-algebra techniques are used to manage all the terms. Working through many such problems has culminated in generalized kinematic modeling software. Written in Mathcad™ Plus 6, programs for flexure systems and kinematic couplings appear in Section 6.3.

6.1 Useful Constraint Devices and Arrangements

Kinematic devices serve many applications that generally require one or more of the following features: 1) separation and repeatable engagement as with a kinematic coupling, 2) defined motion along or about one or more axes, and 3) minimum influence that an imprecise or unstable foundation has on the elastic stability of a precision component. A device is kinematic if it provides the proper number of constraints required for the intended purpose. For example, a supported object should have $n = 6 + f - d$ independent constraints to exactly constrain six rigid-body degrees of freedom plus f flexural degrees of freedom minus d desired axes of motion. In addition to the proper number of constraints, a kinematic design is free of overconstraint.

A purely kinematic design may be difficult (or expensive) to achieve in practice. The term *semi-kinematic* has been used to describe designs that are impure to some extent. That should not imply something is wrong; rather, there are tradeoffs to make in almost every design. It is important to understand the advantages and limitations of various constraint types so tradeoffs can be made to best satisfy the application. I strictly avoid classifying designs as kinematic, semi-kinematic or non-kinematic because there will always be ambiguity. Instead, I advocate applying kinematic design principles within the limits of practical constraint devices; there will almost always be some benefit in doing so. This approach is not limited to precision design but applies to more general machine and mechanism design. See, for example, [Kamm, 1990] and [Reshetov, 1982].

The constraint devices common to precision applications tend to fall into three categories: 1) relatively short-travel flexural bearings (e.g., blade flexures), 2) relatively long-travel bearing components, and 3) repeatable connect-disconnect couplings (e.g.,

kinematic couplings). This chapter focuses on blade flexures and kinematic couplings. Chapter 8, *Anti-Backlash Transmission Design*, presents several common bearing components. See [Slocum, 1992] for more extensive treatment of bearing components.

6.1.1 Basic Blade Flexures

This section presents several common arrangements of blade flexures that provide one axis of motion over a short range of travel. These arrangements: parallel blades, cross blades, and axial blades, are well documented in the literature perhaps with slightly varying names. See, for example, [Jones, 1951, 1962], [Weinstein, 1965], [Siddall, 1970], [Vukobratovich and Richard, 1988] and [Smith, 1998]. A key concept to learn from this section is summarized in the following statement. Several blades connected together as parallel constraints (as opposed to serial constraints) will retain the degrees of freedom that the individual blades have in common. This concept will become clearer after examining the arrangements in this section.

Two parallel blades, connected as shown in Figure 6-1, share a common translational degree of freedom. The rotational degrees of freedom of the individual blades occur about axes that are not in common, thus the combination of two blades constrains those degrees of freedom. Both blades redundantly constrain rotation about the translational axis. A displacement δ_z in the direction of freedom has an associated second-order displacement δ_x given by Equation 6.1. This behavior is a general concern for all flexure designs. All other constraint directions have nominally zero error, although geometric tolerances lead to very small errors that are first order with δ_z.

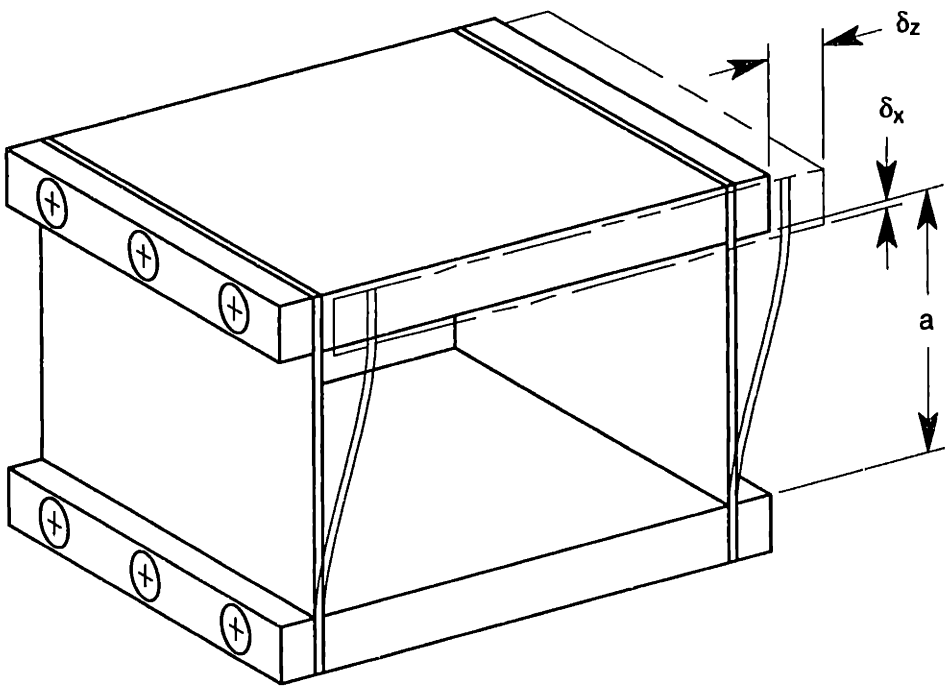


Figure 6-1 Two parallel blades allow one translational degree of freedom and constrain all others. This example shows bolted construction but monolithic designs are also common.

$$\delta_x = \int_0^a \left\{ 1 - \sqrt{1 - \left(\frac{dz}{dx}\right)^2} \right\} dx \cong \int_0^a \frac{1}{2}\left(\frac{dz}{dx}\right)^2 dx = \frac{3\delta_z^2}{5a} \tag{6.1}$$

Two cross blades, connected as shown in Figure 6-2, share one rotational degree of freedom. One blade constrains the degrees of freedom of the other that are not in common. Both blades redundantly constrain translation along the rotational axis. For a given rotation θ, each blade contributes a second-order radial displacement δ_r given by Equation 6.2. The first term inside the braces is the chord across a deflected blade while the second term is the comparable dimension produced by an ideal hinge. The net result is an extension rather than foreshortening as in parallel blades. The total error is the vector sum from both blades.

$$\delta_r = a\left\{ \frac{2}{\theta}\sin\left(\frac{\theta}{2}\right) - \cos\left(\frac{\theta}{2}\right) \right\} \cong \frac{a\theta^2}{12} \tag{6.2}$$

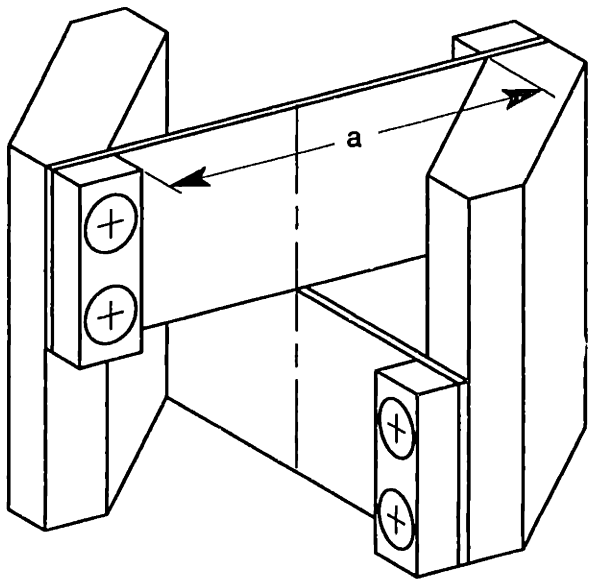


Figure 6-2 Two cross blades allow one rotational degree of freedom and constrain all others. This example shows bolted construction but brazed connections are common in commercial products.

Blades arranged axially, as shown in Figure 6-3, share one rotational degree of freedom. Two blades in different planes are sufficient to constrain the remaining degrees of freedom, but a symmetrical design with four blades is more common. This arrangement has nominally zero radial error motion in contrast to the cross-blade flexure. However, foreshortening of the blades in the axial direction presents an interesting compromise. Equation 6.3 shows the condition required to maintain equal foreshortening across the width of the blades. This condition is satisfied by joining the blades to parabolic-shaped flanges as Figure 6-3 (a) shows. Equation 6.4 shows the condition required to maintain equal bending stress across the width of the blades. The usual compromise solution is to join the blades to conical end caps as Figure 6-3 (b) shows. Making the blades relatively narrow improves this compromise but reduces the stiffness and load capacity of the

flexure. In addition, these equations suggest making the ratio *a/r* large, but they soon become invalid as the geometry diverges from normal beam theory. Finite element analysis is helpful in computing the bending stresses, but a linear code will not represent foreshortening in the blades and the axial stress that may result.

$$\delta_a(r) \cong \frac{3(r\theta)^2}{5a} \quad \rightarrow \quad a \propto r^2 \tag{6.3}$$

$$\sigma_b(r) \cong \frac{3Et(r\theta)}{a^2} \quad \rightarrow \quad a^2 \propto r \tag{6.4}$$

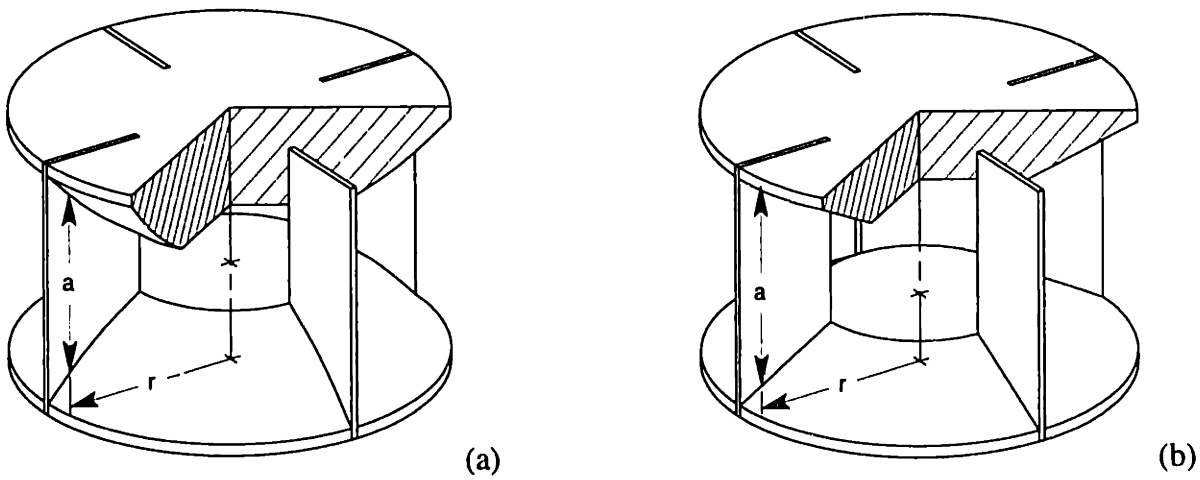


Figure 6-3 Axial blades allow one rotational degree of freedom and constrain all others. The shape of the flanges is important to the behavior of the flexure. In (a), two paraboloids that share a common vertex satisfy equal axial displacement across the blades. In (b), two cones that share a common vertex provide a reasonable compromise between equal axial displacement and equal bending stress across the blades.

6.1.2 Basic Kinematic Couplings

A kinematic coupling provides rigid and repeatable connection between two objects through usually six local contact areas. This is the case for the two traditional configurations shown in Figure 6-4: (a) the three-vee coupling and (b) the tetrahedron-vee-flat coupling (also known as the Kelvin clamp). The weight of the object being supported or some other consistent nesting force holds the surfaces in contact. A spring or compliant actuator may apply the nesting force, but ideally it should allow all surfaces to engage freely with minimum friction and wear. Otherwise, the coupling will not become *centered* as precisely as it should or perhaps not at all. Friction between the contacting surfaces acting on the compliance of the coupling is a main contributor to nonrepeatability as experimentally determined by [Slocum and Donmez, 1988].

The symmetry of three vees offers several advantages: better distribution of contact forces, better centering ability, thermal expansion about a central point and reduced manufacturing costs due to identical features. Conversely, the tetrahedral socket offers a

natural pivot point for angular adjustments. Many tip-tilt mirror mounts operate in this fashion. The three-vee coupling is the natural choice for adjustments in six degrees of freedom or when there is no need for adjustment.

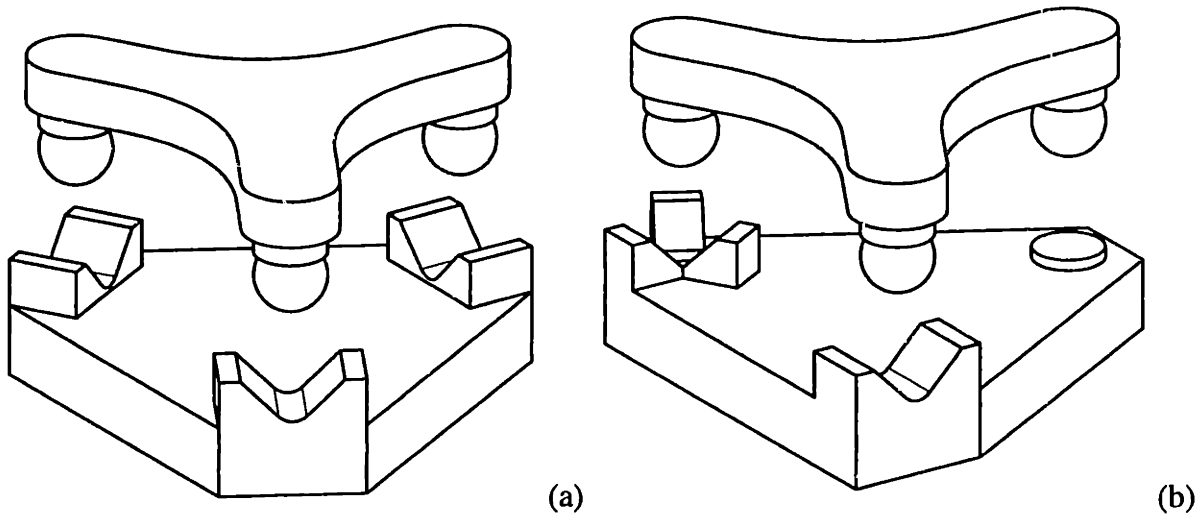


Figure 6-4 In (a), the three-vee coupling has six constraints arranged in three pairs. In (b), the tetrahedron-vee-flat coupling has six constraints arranged in a 3-2-1 configuration. Often for manufacturing reasons, the tetrahedron is replaced with a conical socket, hence the more familiar name cone-vee-flat.

The local contact areas of the traditional kinematic couplings are quite small and require a Hertzian analysis to ensure a robust design for the chosen material pair (see Appendix C, *Contact Analysis*). Greater durability is achieved by better curvature matching between contacting surfaces. Rather than use a full sphere against a flat surface, a partial sphere of much larger radius may be used instead. The same applies to cylindrical surfaces contacting with crossed axes. Another approach is to use a full sphere against a concave spherical or cylindrical surface. Figure 6-5 compares these two approaches for a vee constraint. Both constraints have the same relative (or effective) radius but the sphere in a gothic arch has less capture range.

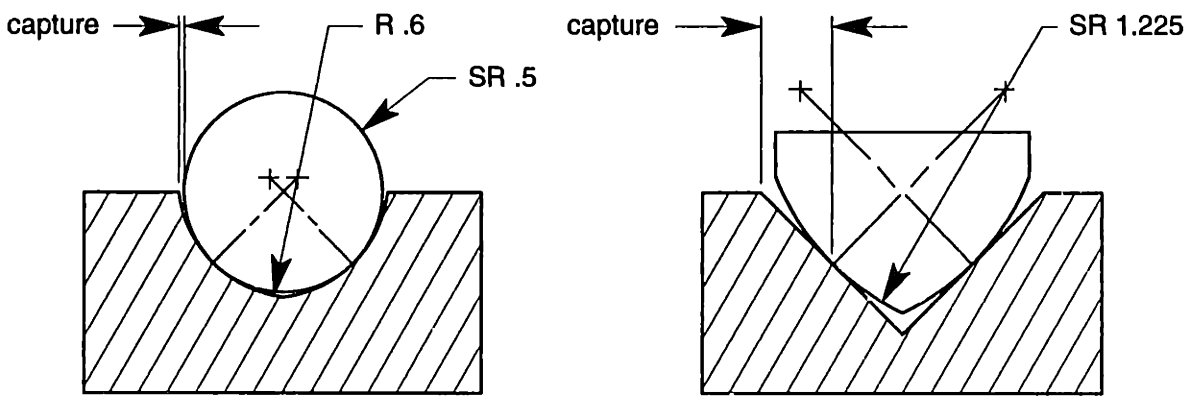


Figure 6-5 A vee constraint showing two ways to increase the area of contact. Capture is the maximum distance off center that the constraint will engage with tangency.

Designs based on line contact rather than point contact offer a significant increase in load capability and stiffness. For example, line contact forms between a precisely made, heavily loaded sphere and conical socket. The kinematic equivalent to three vees is a set of three sphere-cone constraints with either the spheres or the cones supported on radial-motion flexures. The upper member in Figure 6-6 (a) has six rigid-body plus three flexural degrees of freedom that three cones exactly constrain. Alternatively in (b), the three-tooth coupling forms three theoretical lines of contact between cylindrical teeth on one member and flat teeth on the other member. Each line constrains two degrees of freedom giving a total of six constraints. Manufactured with three identical cuts directly into each member, the teeth must be straight along the lines of contact but other tolerances may be relatively loose. Both of these kinematic couplings are being used on the EUVL project to overcome the limited hardness of super invar.

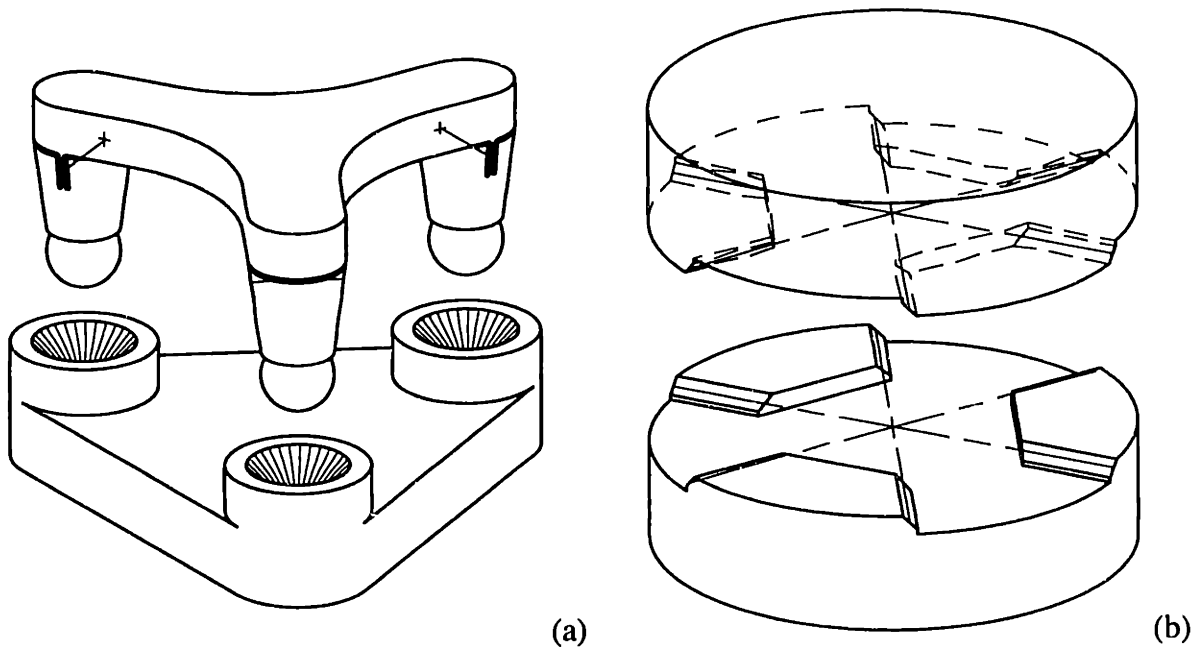


Figure 6-6 In (a), flexure cuts in the upper member allow each sphere limited radial freedom to seat in the conical sockets of the lower member. In (b), the three-tooth coupling forms three theoretical line contacts between cylindrical teeth on one member and flat teeth on the other member.

6.1.3 Extensions of Basic Types

Arranging constraints is a design process that requires a basic understanding of kinematics and the mechanics of constraint devices. The blade flexures and kinematic couplings presented thus far are good examples from which to learn and start new designs. This section presents several interesting and useful extensions based on three vee constraints. The examples range from fairly direct implementation on a touch trigger probe to a less obvious flexure stage with three degrees of freedom. In my experience, thinking of six constraints as three pairs has been a valuable and simplifying conceptual construct.

6.1.3.1 Touch Trigger Probe

Touch trigger probes are commonly used on coordinate measuring machines to indicate precisely where in the travel of the machine axes that contact is made with the workpiece. It is sufficient if the probe signal occurs with a known position lag as this is easy to correct in software. A common design studied by [Estler, et al., 1996, 1997] employs a three-vee kinematic coupling that acts as the electrical switch and the mechanical registration. The problem that Estler addresses through modeling and compensation is the variation in position lag depending upon the direction of travel, the orientation of the surface and other effects. A dominant error term, referred to as probe lobing, results from a three-fold variation in the trigger force acting on the compliance of the probe shaft.

The heart of the problem is the orientation of the vee constraints. Figure 6-7 shows the probe mechanism studied by Estler (a) and a new design (b) that solves probe lobing, at least in theory. In (a), the probe side of the coupling is preloaded down by a compression spring into three vee constraints represented by angled cylinders. The probe will not trigger until there is sufficient moment imparted to the coupling for any of the constraints to become unloaded, thus breaking electrical continuity. Although the preload is constant, the lever arm may vary up to a factor of two depending whether the coupling pivots about one vee or two. In (b), the new vee orientation requires a torsional preload to seat the coupling. In addition, the spring would be set to off-load the weight of the probe coupling. In this configuration, any applied moment (orthogonal to the preload) equally unloads one side of each vee; there is no directional preference. The downside will be a greater influence of friction since any pin must now slide up or down a vee rather than simply lifting out.

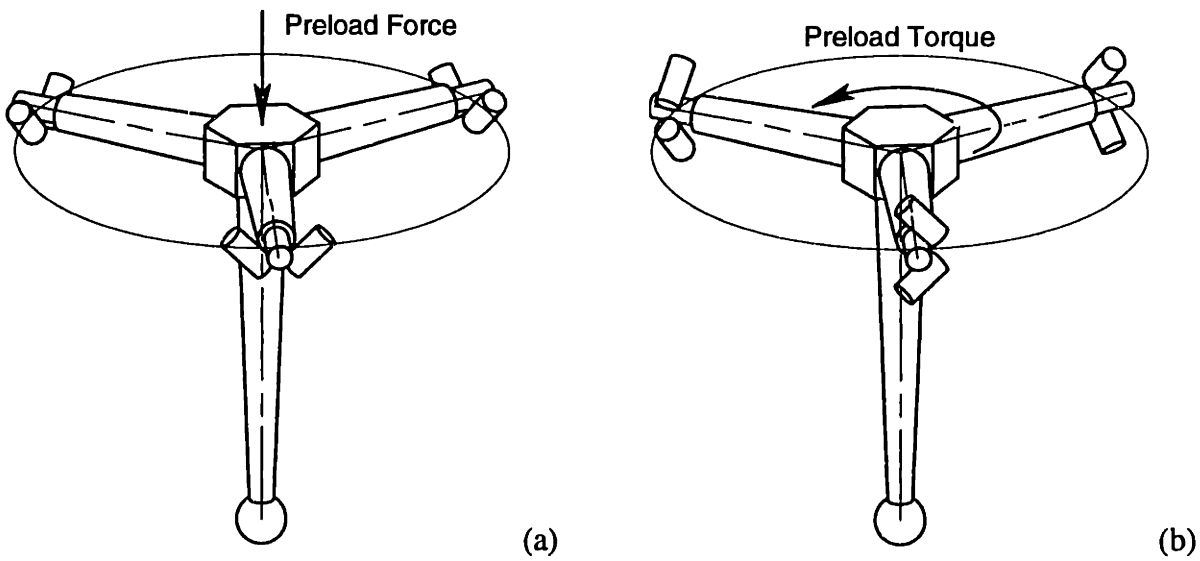


Figure 6-7 In (a), the moment required to unseat one vee while pivoting about the other two vees is a factor of two less than the moment required to unseat two vees while pivoting about the third vee. In (b), a moment applied about any axis in the plane of the vees produces equal reaction at all vees.

6.1.3.2 The NIF Diagnostic Inserter

The NIF requires a number of diagnostic instruments near the center of the 10 m diameter target chamber. Each instrument is transported approximately 6 m into the chamber by a telescoping diagnostic inserter. Since only the end position of travel requires submillimeter positioning, a kinematic coupling is being considered to provide repeatable registration at the end of a rather imprecise telescoping stage. However, the long, skinny geometry of the inserter presents an unfavorable aspect ratio for a traditional kinematic coupling. The configuration shown in Figure 6-8 was proposed to work within the geometric constraints yet provide acceptable moment stiffness and capacity. It was conceived by splitting the vees of a three-vee coupling and axially separating the odd-numbered constraints from the even-numbered constraints. The odd-numbered constraints act like a right-hand screw while the even-numbered constraints act like a left-hand screw. An applied axial preload force translates the cylinder until all constraints are engaged and an axial torque is established between the two sets of three constraints.

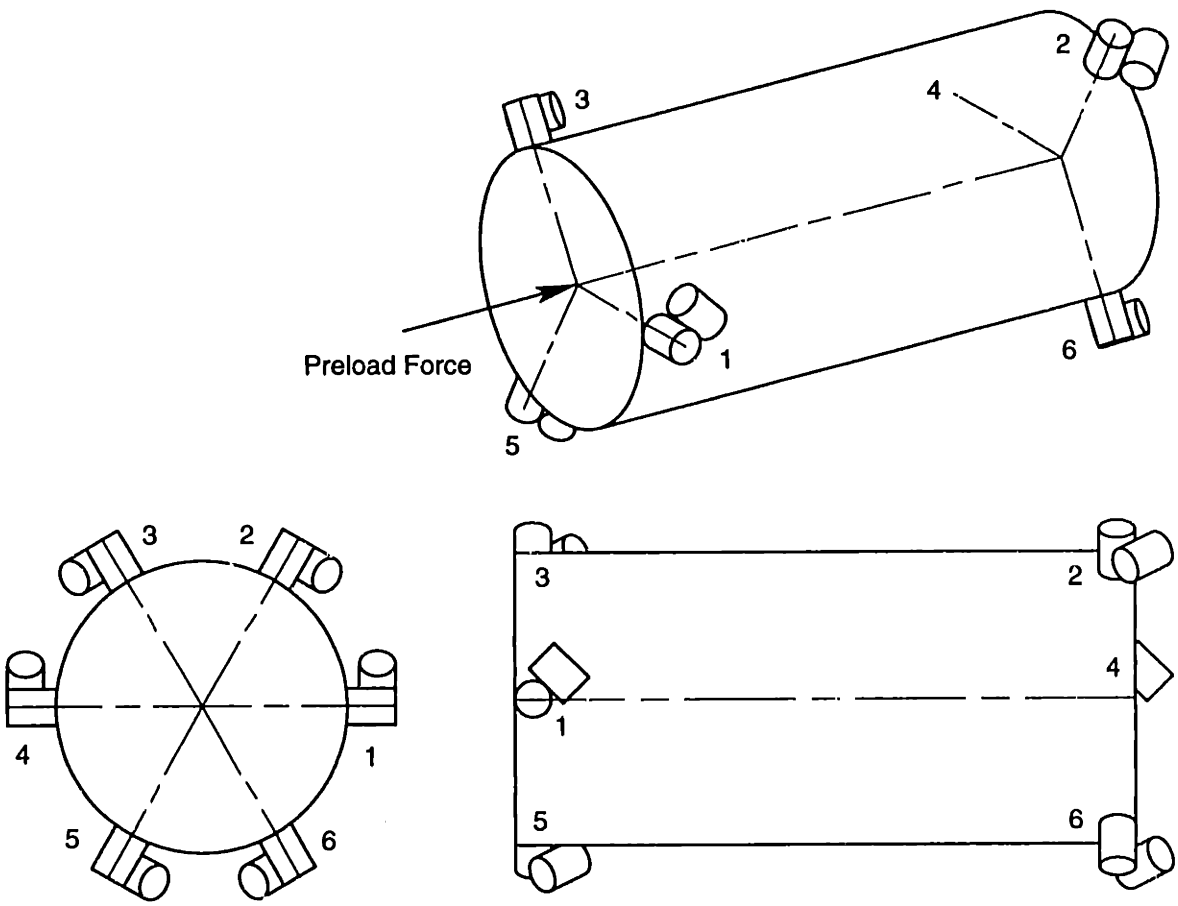


Figure 6-8 The end view looks much like a three-vee coupling with constraint pairs 2-3, 4-5 and 6-1 apparently forming three vees. The side view shows the significant separation between odd- and even-numbered constraints. As in Figure 6-7, the angled cylinders are constraints fixed to an unseen structure.

6.1.3.3 The NIF optics assembly

The NIF requires many hundreds of kinematic couplings to support large, replaceable optics assemblies. There are several types of kinematic couplings used throughout the system, but one in particular demonstrates a simple evolution from a basic three-vee coupling to a more novel configuration well suited for tall assemblies. Figure 6-9 shows the evolution in three simple steps. The horizontal configuration is convenient because gravity provides the preload. Rotating the coupling to the vertical configuration has obvious consequences, which motivates the next step to rotate the lower vees to carry the gravity load. It is important that the centroid of the supported object be offset from the lower vees in a direction that preloads the upper vee. The next step of spreading the upper vee has a particular advantage for NIF optics assemblies. The widely spaced vee provides frictional constraint that stiffens the torsional vibration mode of the optics assembly. This example appears again in Section 6.3.3 and Chapter 7.

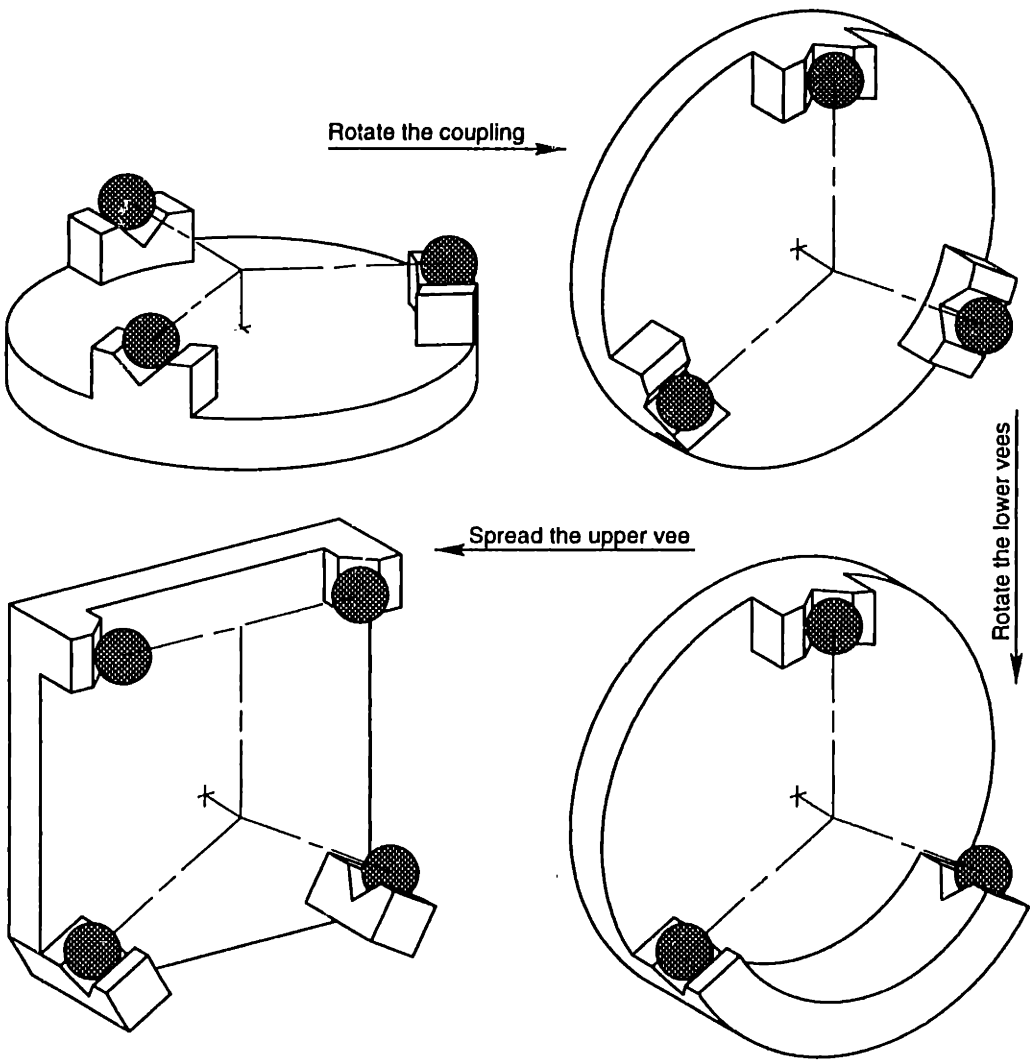


Figure 6-9 The evolution from a horizontal three-vee coupling to the configuration used for many NIF optics assemblies. The spheres in each configuration attach to the object being supported.

6.1.3.4 EUVL Mirror Mount

Friction between the contacting surfaces of a kinematic coupling is a disadvantage when it causes significant distortion in the precision component being supported. A common approach for mounting super-precision optics is an arrangement of three vee-flexures that [Vukobratovich and Richard, 1988] refer to as bipods. Figure 6-10 shows the bipod design used for EUVL mirror mounts. Each leg consists of four blades in series to provide one constraint and five degrees of freedom. One bipod provides the same constraint as a sphere and vee but without friction. Three bipods fully constrain the supported object with six constraints connected in parallel. Notice too that the top of the bipod has the features for a three-tooth coupling. There are mating features on the optic to provide the connect-disconnect function. The kinematic repeatability of the couplings ensure repeatable forces imposed by the bipod flexures on the optic, leading to a repeatable distortion between optic manufacturing and final use. This example appears again in Chapter 7 in greater detail.

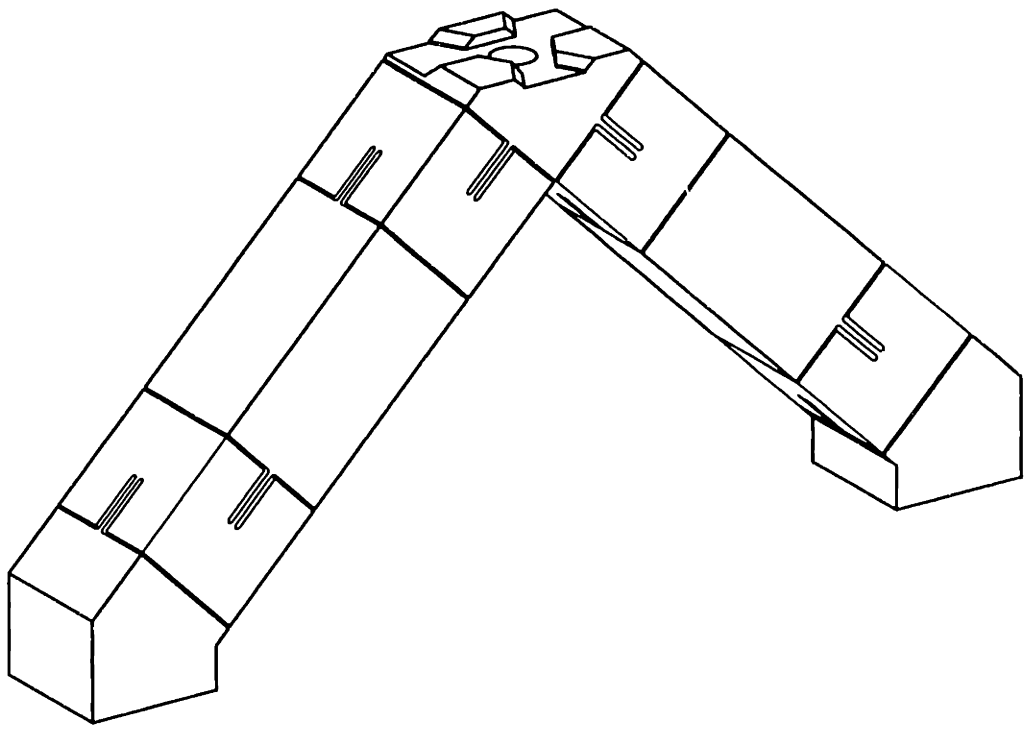


Figure 6-10 A single bipod flexure constrains two degrees of freedom in the plane of the vee. Usually the center section connects to the precision component and the ends connect to the support. At times it may be advantageous to reverse this role for better weight distribution provide by six supports.

6.1.3.5 X-Y-θ_z Flexure Stage

A flexure stage that provides pure planar motion (X-Y-θ_z) satisfies a number of applications found particularly in microelectronics and opto-mechanical systems. One approach to this problem is to serially connect single-axis flexure stages, for example, two sets of parallel blades and one set of cross blades. Besides being an awkward design, good stiffness in each constraint direction is difficult to obtain. When possible, it is better to

arrange constraints in parallel. An obvious example is a set of three single-constraint flexures arranged physically parallel to each other. This arrangement is rigid and simple, but has second-order, out-of-plane error motion that can only be reduced with longer constraints. A better arrangement appears in Figure 6-11. It consists of three folded-hinge flexures arranged as parallel constraints. This arrangement provides pure planar motion except for errors arising from geometric tolerances.

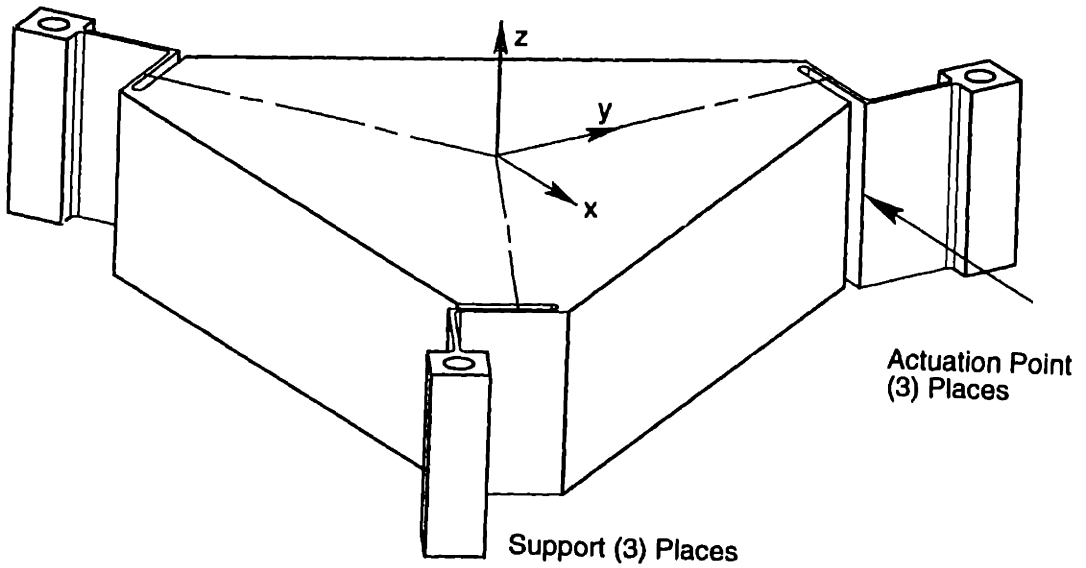


Figure 6-11 Three folded hinge flexures constrain motion within a plane and provide convenient points with which to actuate the stage.

The folded hinge provides one constraint although it appears compliant in all directions. Rather, the two blades have one degree of freedom in common so only five of the six (2 x 3 DOF) are independent. One nice feature of the folded hinge is the convenient point to apply actuation, for example, with a micropositioner. This was the approach used for an EUVL X-Y stage that appears in Chapter 7. [Ryu, Gweon and Moon, 1997] designed an X-Y-θ_z wafer stage that uses piezoelectric actuators driving folded hinges.

6.2 Analytical Design of Flexures

Much has been written about the analysis of flexures, so much so that the papers are seemingly saturated with the same information. There has been little new understanding presented in resent years. The emphasis in this section is in providing new information and understanding. This is accomplished using both beam theory and finite element analysis. A fundamental contribution is a matrix-algebra technique for modeling flexure systems. The equations for a blade flexure are contained in a compliance matrix and a stress matrix, both of which consider column effects. This is a sophistication not found in the formulas of Section 2.6. Computer software written for specific configurations such as the bipod flexure has proved very valuable in the case studies for this thesis. A general-configuration program for flexure systems, written in Mathcad™ Plus 6, appears in Section 6.3.

6.2.1 Comparison of Flexure Profiles

The blade flexures presented thus far have had constant thickness except perhaps near the ends where small fillets are typical. Another common flexure profile is the circular hinge, which typically is manufactured by drilling two adjacent holes to form the flexure and then by relieving other material as necessary to allow freedom of motion. The primary reference for the circular hinge is [Paros and Weisbord, 1965]. More recently, the elliptical hinge was studied by [Xu and King, 1996] and [Smith et al., 1997]. When the thickness of the flexure is small compared to the circle or ellipse, both of these profiles are well approximated by a parabola. A parabolic profile leads to simpler equations and better understanding. Since these three profiles have effectively the same performance, the circular hinge is the obvious choice for ease of manufacturing (whether by drilling holes or using circular interpolation). The interesting comparison is between the (circular, elliptical or parabolic) hinge flexure and the blade flexure because each has particular advantages.

To remain the most general, the presentation uses the elliptical profile described by the major and minor diameters a and b, respectively. For a circular profile, simply replace both a and b with the diameter d. Equation 6.5 gives the thickness profile for the ellipse and its approximate parabolic profile. Figure 6-12 compares these two profiles for an example that is near the limit for a good approximation. The approximation is better for a circular profile and of course when the minimum thickness t_0 is thinner. The straight lines in the figure have to do with the comparison to the equivalent blade flexure discussed later.

$$t_e = t_0 + b\left[1 - \sqrt{1 - \left(\frac{2x}{a}\right)^2}\right] \cong t_0 + \frac{b}{2}\left(\frac{2x}{a}\right)^2 = t_p \tag{6.5}$$

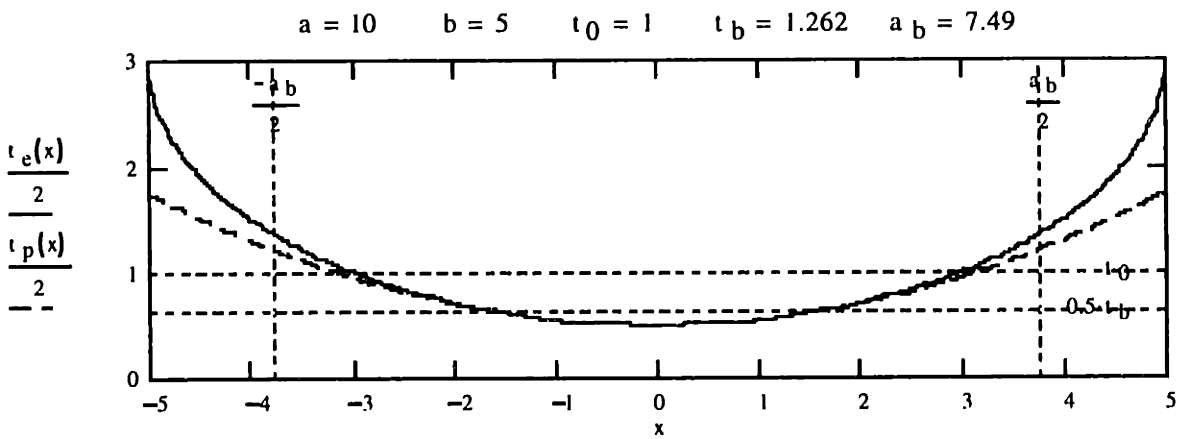


Figure 6-12 The solid line indicates the profile for one side of an elliptical hinge flexure. The horizontal axis is a plane of symmetry. The parabola (dash line) provides a good approximation in the thin region of the flexure that governs both the axial and moment compliance. The equivalent blade flexure is bounded by straight lines indicated by a_b and t_b.

Equations 6.6 and 6.7 give simplified expressions for axial compliance and moment compliance, respectively, for the parabolic profile. For this example, the axial compliance

of the ellipse is underestimated by 3% and the moment compliance is overestimated by 5% compared to *exact* solutions using the elliptical profile. However, the use of beam theory in the derivation is itself an approximation. These solutions are similar to those for the blade flexure. If the blade were taken to be of length a and thickness t_0, then the term in square brackets would represent the factor by which the hinge flexure was different.

$$c_x = \frac{1}{E\,w}\int_{-a/2}^{a/2} t_p^{-1}\,\mathrm{d}\,x \cong \frac{a}{E\,w\,t_0}\left[\frac{\pi}{2}\sqrt{\frac{2t_0}{b}} - \frac{2t_0}{b}\right] \tag{6.6}$$

$$c_\theta = \left(1-\nu^2\right)\frac{12}{E\,w}\int_{-a/2}^{a/2} t_p^{-3}\,\mathrm{d}\,x \cong \left(1-\nu^2\right)\frac{12a}{E\,w\,t_0^{\,3}}\left[\frac{3\pi}{16}\sqrt{\frac{2t_0}{b}}\right] \tag{6.7}$$

It is instructive to consider the length and thickness of a blade that is equivalent to the hinge flexure in terms of axial and bending compliance. The solution to two equations with two unknowns appears in Equation 6.8, where the subscript b indicates the equivalent blade parameters. This explains the straight lines in the figure marked with either a_b or t_b. The line marked t_0 indicates the part of the parabola that has the greatest slenderness ratio for buckling. The usual definition for slenderness ratio is the length divided by the minimum radius of gyration. Here it is more convenient to use the length divided by the thickness. It is obvious from the figure that the equivalent blade being both longer and thinner is more likely to buckle under a compressive load. Equation 6.9 gives the condition required for the hinge flexure to yield before buckling and the factor by which the equivalent blade flexure is more likely to buckle. For this example the factor is 1.88.

$$\frac{a_b}{a} \cong \sqrt{\frac{16}{3\pi}\frac{2t_0}{b}\left[\frac{\pi}{2} - \sqrt{\frac{2t_0}{b}}\right]^3} \qquad \frac{t_b}{t_0} \cong \sqrt{\frac{16}{3\pi}\left[\frac{\pi}{2} - \sqrt{\frac{2t_0}{b}}\right]} \tag{6.8}$$

$$SR_p \cong \frac{a}{\sqrt{2t_0\,b}} < \pi\sqrt{\frac{E}{12\sigma_y}} \qquad \frac{SR_b}{SR_p} \cong \pi - 2\sqrt{\frac{2t_0}{b}} \tag{6.9}$$

The hinge flexure clearly has the advantage over the blade flexure for buckling resistance. The bending stress appears to be slightly higher for the thinner hinge flexure, since both equivalently require the same bending moment for a given rotation, but stress concentrations in the fillets of the blades can be just as high. The main advantage for the blade flexure comes when there is need for rotational flexibility about the axis of the blade, so as to twist. Of course the hinge flexure is better if the application calls for resisting twist. This is also true if the flexure is to be used as a secondary constraint in shear.

6.2.2 A Study on Fillets for Blade Flexures

Beam theory works well for blade flexures except at each end where there is a transition to some larger cross section. Monolithic blade flexures are usually manufactured with small corner radii know as fillets. Clamped blades, on the other hand, usually have very abrupt transitions that are difficult to model with any certainty. The edges of the clamping surfaces may have partial radii to transition the clamping force, but it is assumed that microslip will relieve the theoretically high stress concentration for axial and moment loads. Naturally the subject of this study is the behavior of fillets as a function of radius size. This is accomplished through a parameterized finite element model for one end of the blade.[I] In the study, the radius varies from one-half blade thickness to twice the blade thickness, and the model is subject to either axial or moment loading. Curve fits to the finite-element results are useful to supplement the limitations of beam theory.

An assumption is made in beam theory that either the out-of-plane stress or the out-of-plane strain is zero. The same is true for 2D FEA. The two choices bracket the range of a 3D model, plane stress for zero width and plane strain for infinite width. Both types were compared to a 3D model of varying widths. Consistent with the general practice in flexure design, plane stress is most appropriate in the calculation of axial stiffness and stress due to axial and moment loads. However, the results indicate that plane strain is more appropriate for moment stiffness, contrary to general practice. The effect is rather small, only 5 to 10 percent but in the nonconservative direction. Hence, the equations found in this thesis for bending of flexures have a factor $(1 - \nu^2)$ to account for stiffening due to the Poisson effect.[II] The same is true for the finite-element results that appear later in this section.

The shape of the transition region and the range of fillet radii are apparent in Figure 6-13. In this case, an axial load is applied to the left end of the 20 x 20 block, and the right end of the 2 x 10 blade is constrained. In Figure 6-14, opposite forces on the top and bottom of the block generate a moment load. Both figures show the deflected shape of the model and contours of von Mises stress. Although difficult to see, the maximum stress for axial loading occurs approximately at the quarter point of the fillet closest to the blade and occurs very near the start of the fillet for moment loading. The node on the lower right corner of the block is the displacement location used for the compliance calculations. All the results presented are normalized to the blade thickness t and calculations from beam theory.

[I] Pro/MECHANICA by Parametric Technology Corp. is the finite element software used in this study.

[II] The bending stress across the thickness of the flexure changes from tension to compression over a very short distance. The blade would bow if not connected on each end to a stiff structure. The plane-strain assumption does not allow any bowing so the calculation underestimates the desired quantity, bending compliance. The blade has some opportunity to bulge in width when axially loaded. The plane-stress assumption freely allows bulging so the calculation underestimates the desired quantity, axial stiffness. The maximum stress due to axial and moment loads occurs on the sides where the plane-stress assumption is valid.

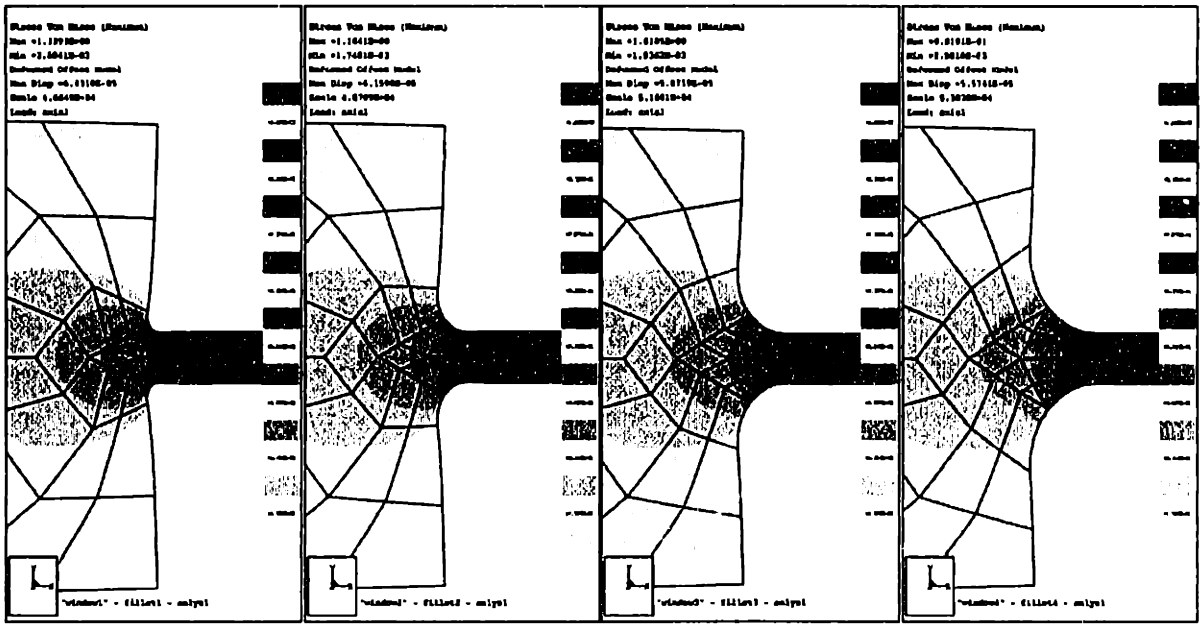


Figure 6-13 The fillet radii shown here are one-half, one, three-halves and two times the blade thickness. The deflected shape and the contours of von Mises stress result from an axial load.

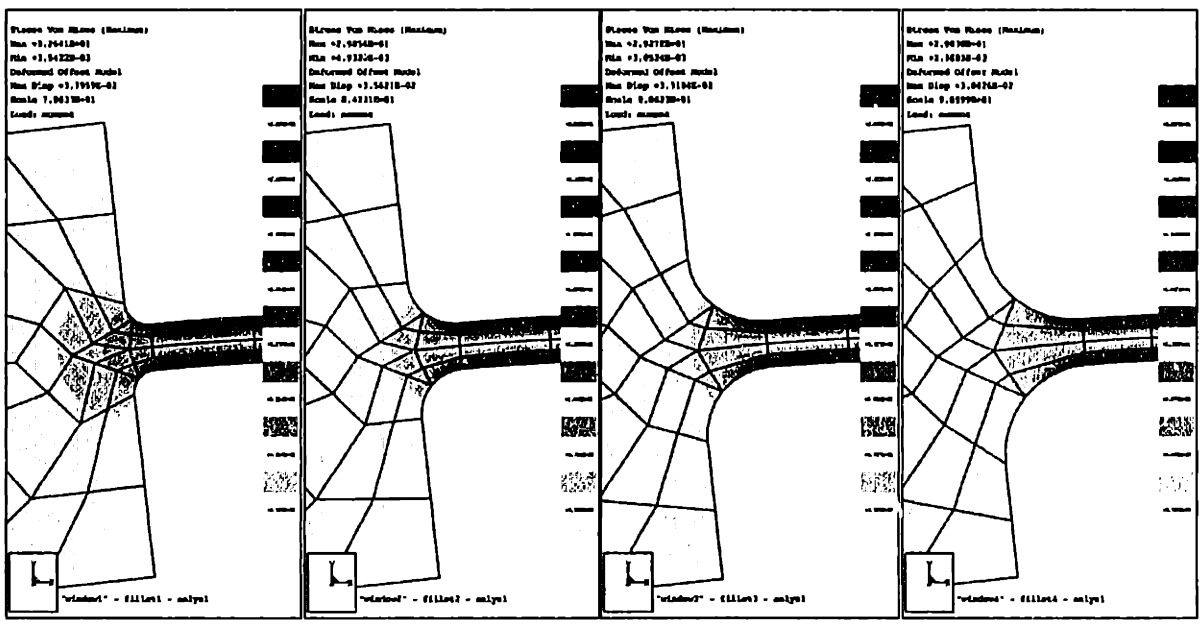


Figure 6-14 The fillet radii shown here are one-half, one, three-halves and two times the blade thickness. The deflected shape and the contours of von Mises stress result from a moment load.

The maximum stress from the 2D plane-stress model divided by the stress calculated from beam theory is the stress concentration factor plotted in Figure 6-15 for axial loading and Figure 6-16 for moment loading. In each graph, the solid line is a fitted curve to discrete results from the finite element model. The equation at the top of each graph may be used to calculate the stress concentration factor for any radius-to-thickness ratio between one-half and two. The knee in the curve appears to be at a ratio near one.

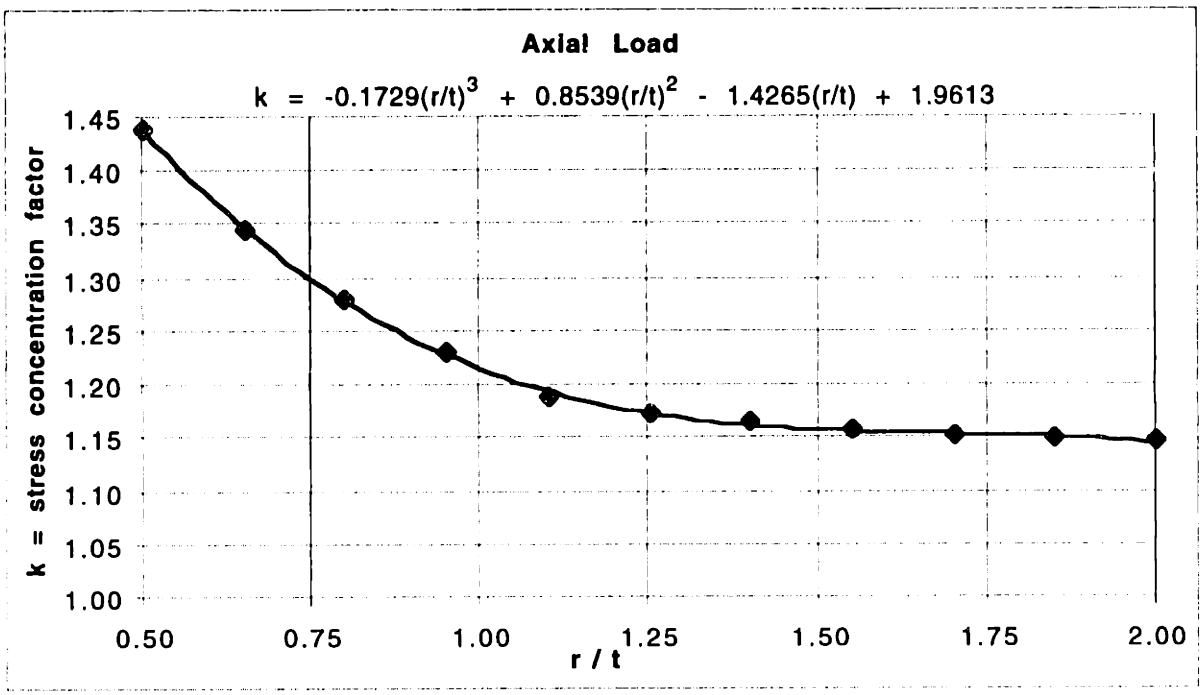


Figure 6-15 The stress concentration factor for axial loading is closely approximated by a cubic polynomial, where *r/t* is the ratio of fillet radius to blade thickness.

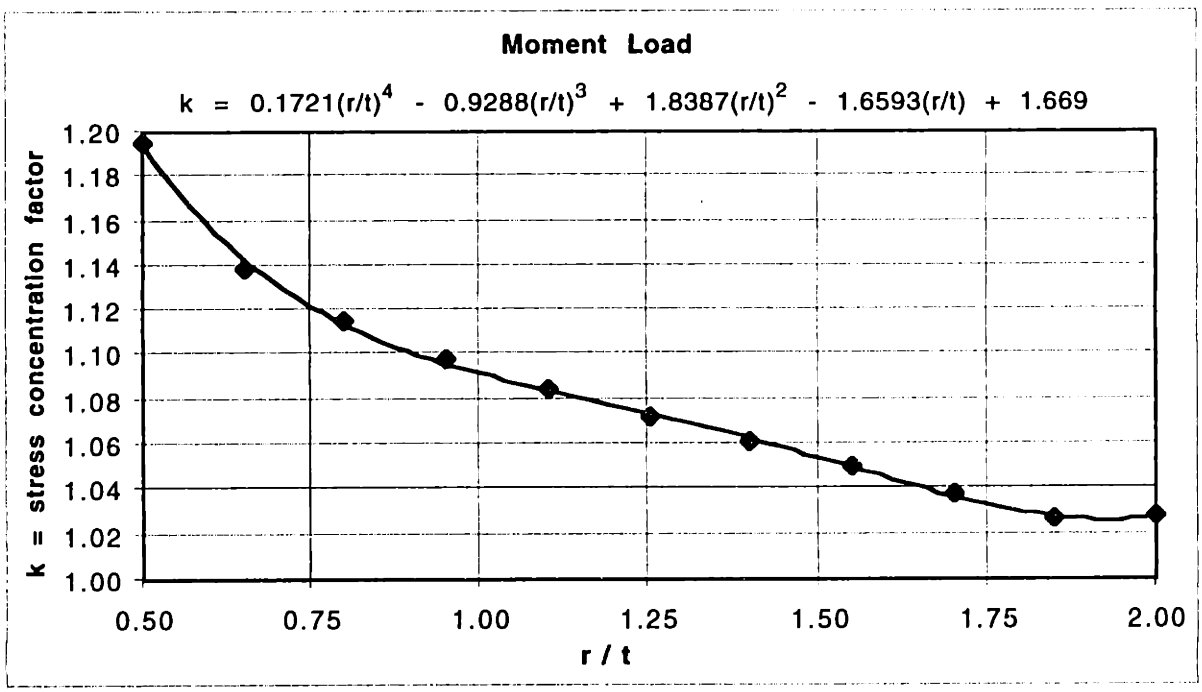


Figure 6-16 The stress concentration factor for moment loading is closely approximated by a fourth-order polynomial, where *r/t* is the ratio of fillet radius to blade thickness.

The size of the fillet radius also has an effect on the amount of deflection under load. A larger fillet shortens the effective length of the blade assuming that the end structures remain separated by a constant distance *a*. This effect on blade length is apparent in Figure 6-17 for axial loading and Figure 6-18 for moment loading. The curves give the

additional length of blade required to make beam theory match the displacement predicted from the finite element model. As might be expected, the compliance due to the elasticity of the end structures is significant for axial loading. For moment loading, beam theory matches the finite element model for a ratio $r/t = 0.62$.

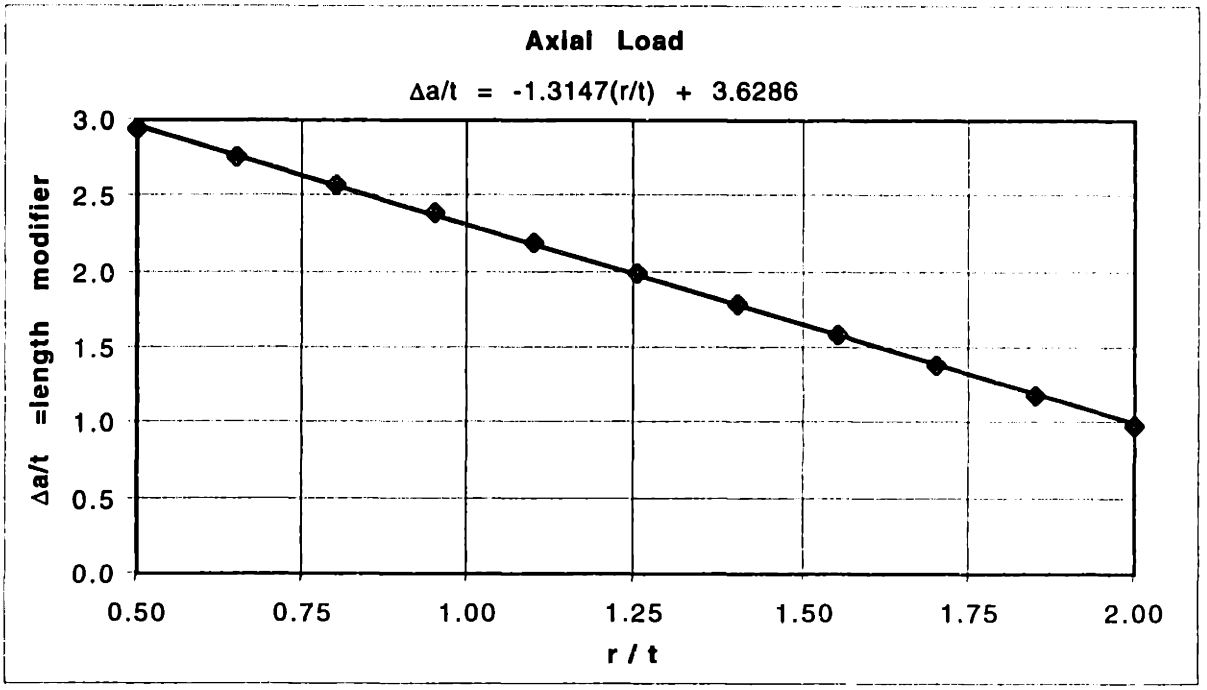


Figure 6-17 The length modifier for axial loading is closely approximated by a linear curve, where *r/t* is the ratio of fillet radius to blade thickness.

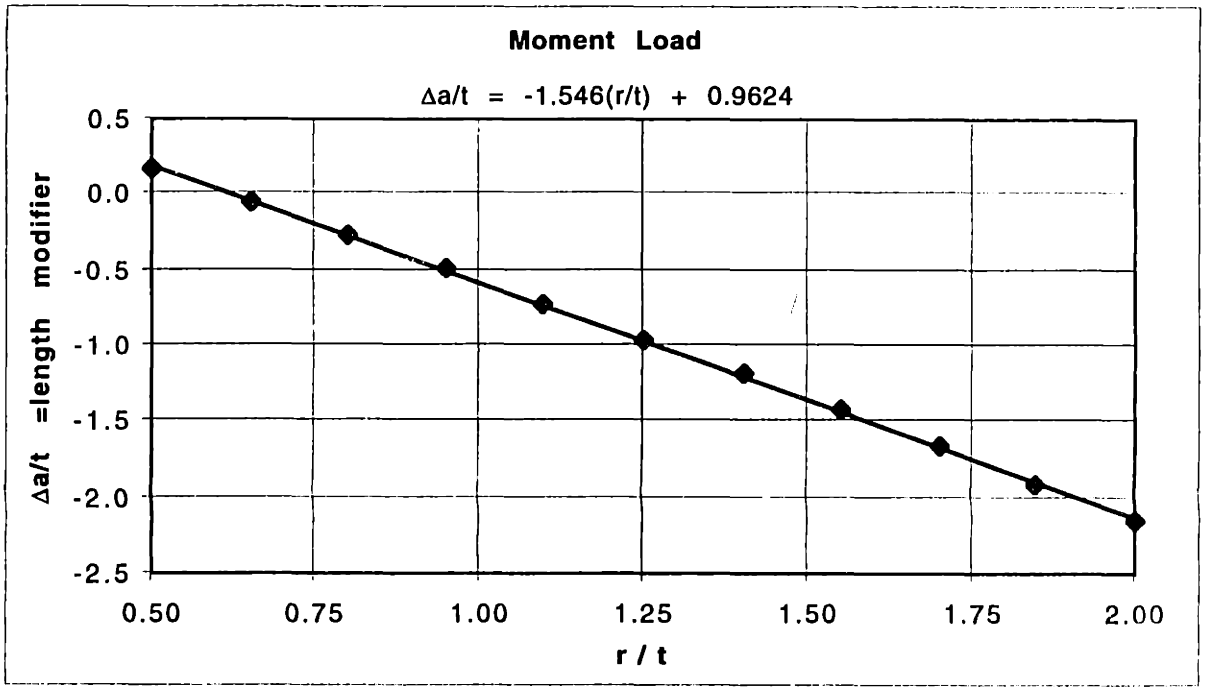


Figure 6-18 The length modifier for moment loading is closely approximated by a linear curve, where *r/t* is the ratio of fillet radius to blade thickness.

6.2.3 The Compact Pivot Flexure

An axial arrangement of two blades in series is a useful single-constraint device that provides angular freedom about three axes. Since the two translational degrees of freedom are rather stiff for short blades, it is common to duplicate another set of two blades some distance along the axial constraint direction. See, for example, the bipod flexure in Figure 6-16. The same cuts used to make the bipod flexure appear more clearly in Figure 6-19 (a). This basic design is being used on the NIF and EUVL projects. An advantage of this design becomes apparent when compared to the more common design in (b), where the axial compliance introduced at the junction between blades is significant. It clearly shows the compromise between axial stiffness and how closely spaced the blades can be. The design shown in (a), with much deeper end sections, greatly relieves this compromise. Even so, it starts to become an issue again when the blade is wider than four times its length. This three-dimensional behavior is best studied with 3D finite element analysis. As before, finite-element results are displayed so as to extend the usefulness of simple theory.

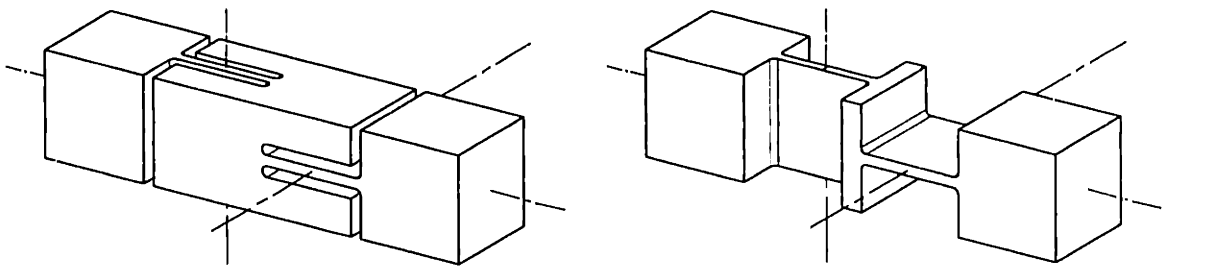


Figure 6-19 The design in (a) allows the minimum spacing of blades and maintains good axial stiffness. In addition, the gaps may be controlled to provide over-flexion protection. In order for the design in (b) to have good axial stiffness, the junction between blades would have to be lengthened.

Since only axial displacement is of interest in this study, the use of symmetry boundary conditions at two midplanes simplifies the model to just one-quarter the physical pivot flexure. This model, shown in Figure 6-20, also aids in viewing contours of von Mises stress through the blades. The variable parameter in this study is the blade width w, which varies from one to four times the length a. The blade length is ten times the thickness and the fillet radius is one-half the blade thickness.

Although the blades become stiffer with increasing width, the aspect ratio of the junction becomes less favorable and contributes a larger proportion to the total compliance. This is the reason in Figure 6-21 that the axial displacement when normalized to theory increases with blade width. This behavior is also apparent in von Mises stress as gradients that increase with blade width. Figure 6-22 shows how stress varies across the half-width taken through the center of the blade (length and thickness). Figure 6-23 shows how stress varies along the axis of symmetry. Notice that these stresses are away from the stress concentrations caused by fillets. A practical maximum for blade width is two times the length partly because the torsional stiffness increases rapidly with the ratio w/a.

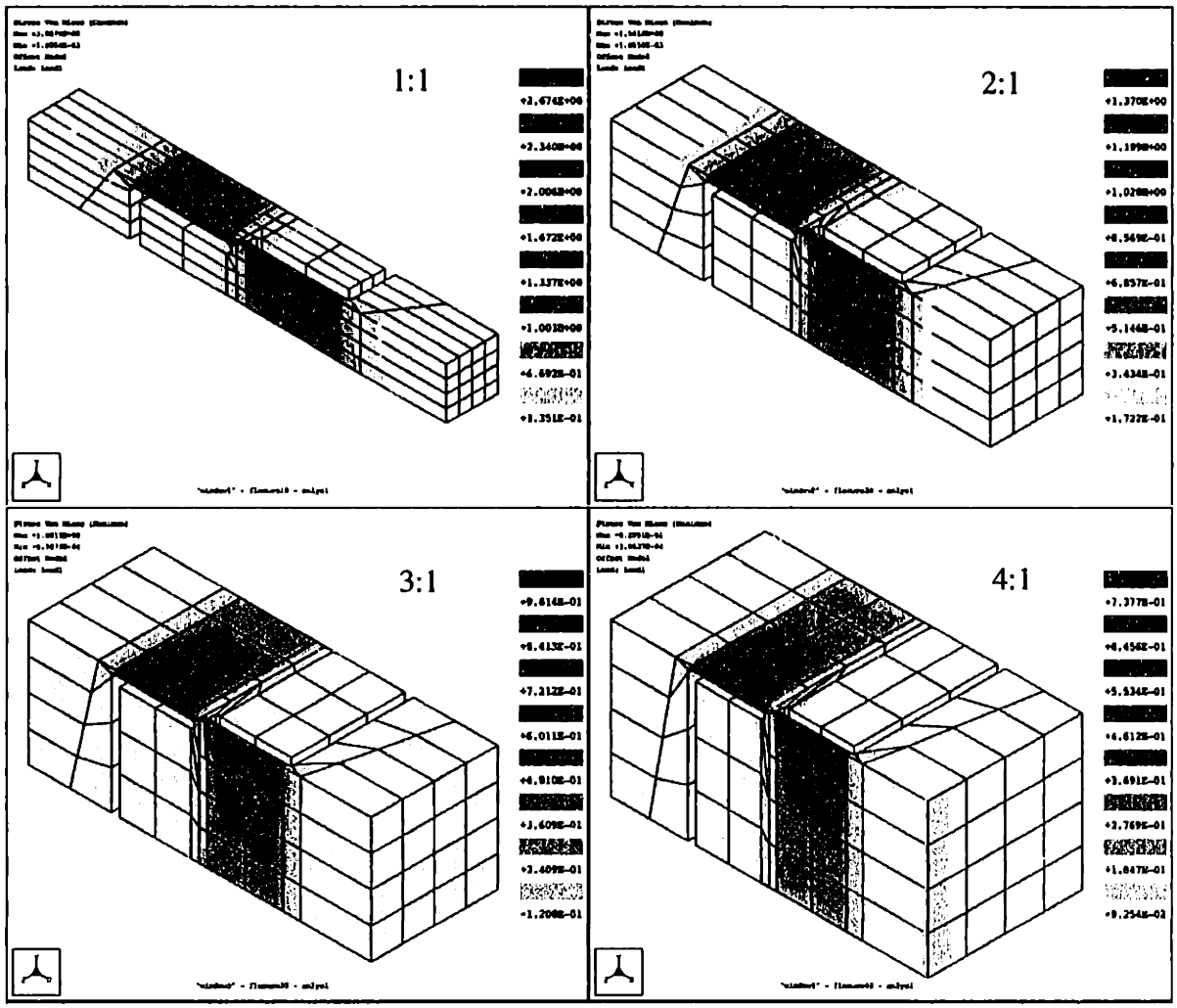


Figure 6-20 Contours of von Mises stress for blade widths from one to four times the blade length.

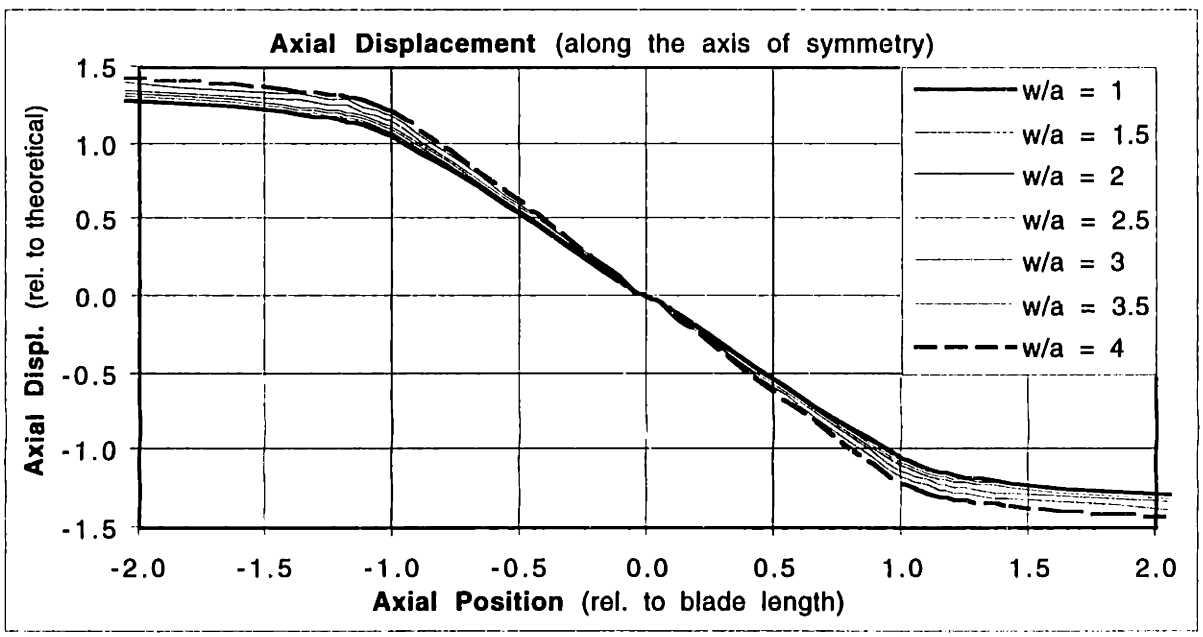


Figure 6-21 Axial displacement versus axial position along the axis of symmetry.

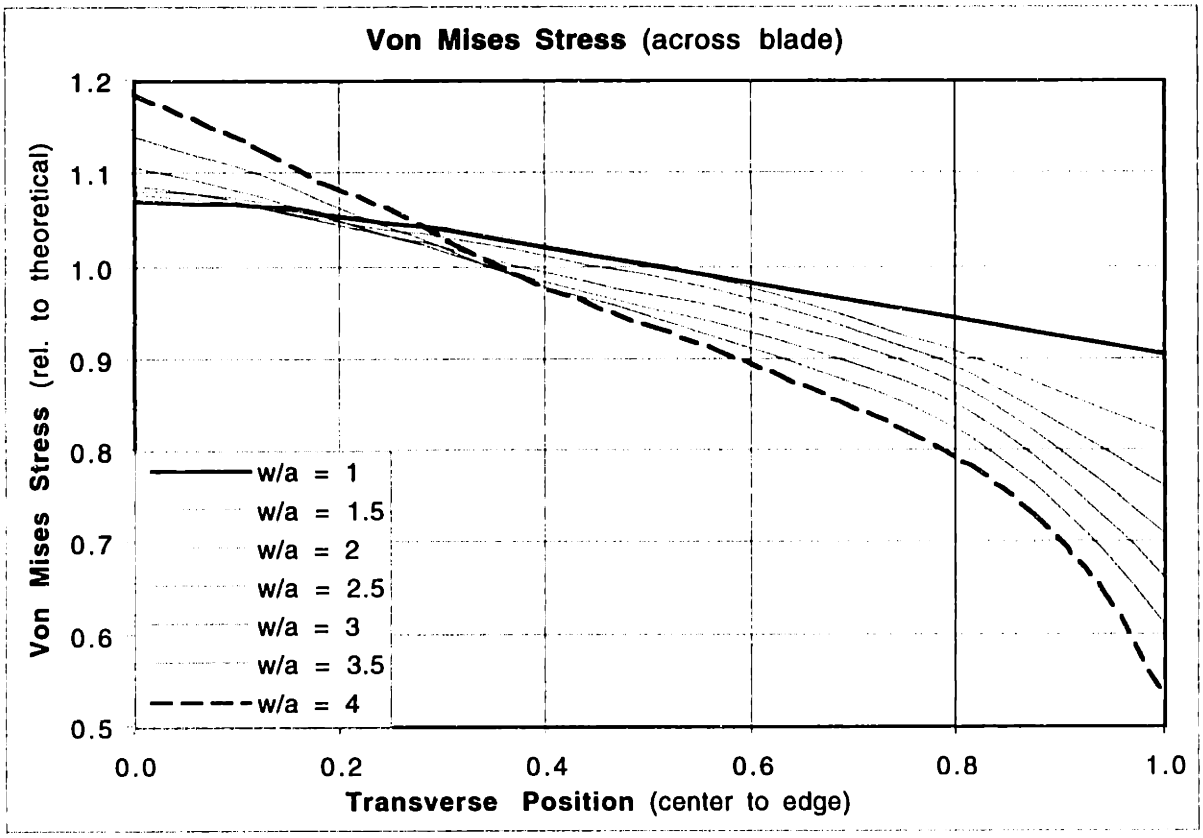


Figure 6-22 Von Mises stress versus position across the half width of the blade (taken at the mid length).

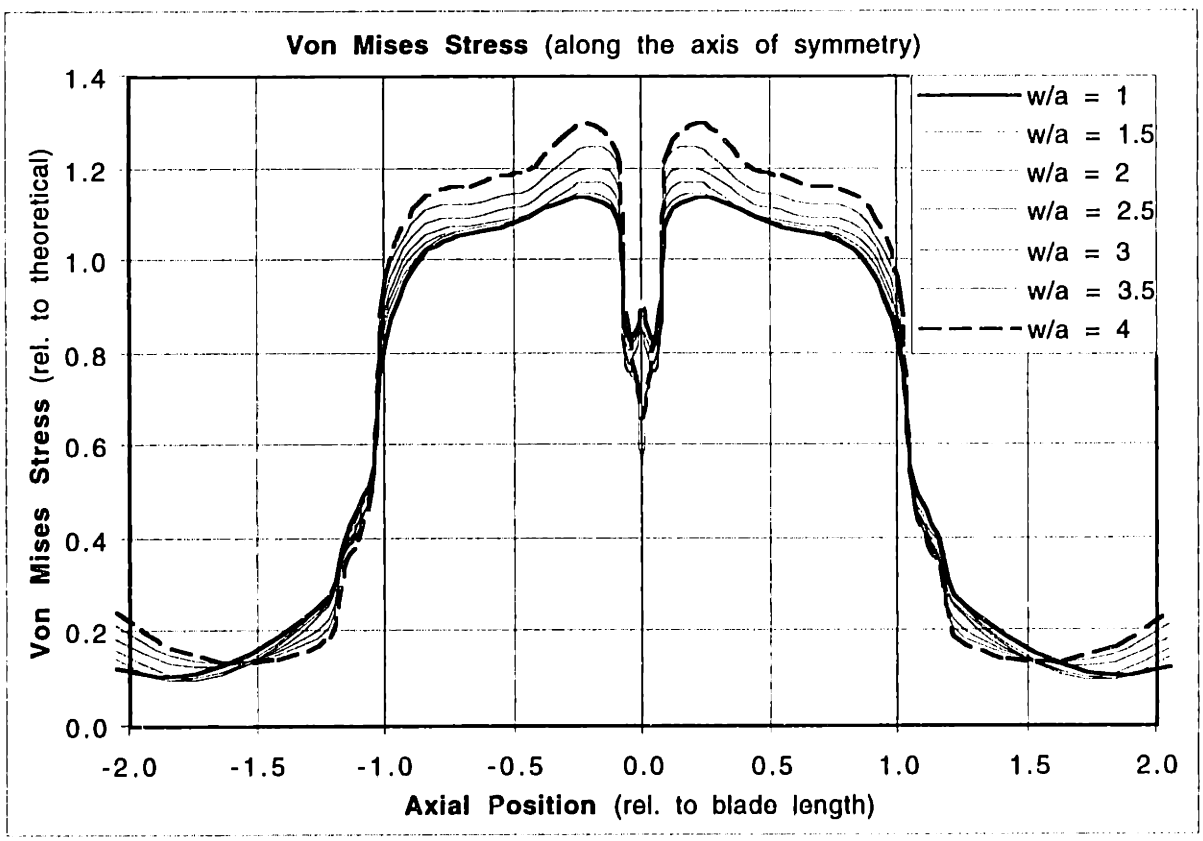


Figure 6-23 Von Mises stress versus position along the axis of symmetry.

6.2.4 Helical Blades for a Ball-Screw Isolation Flexure

The rather specialized application of a ball-screw (or leadscrew) isolation flexure motivated the development of an analytical model for a mildly helical blade flexure. Isolation flexures or bearing systems are commonly used on ultra-precision machines to couple only the desirable degrees of freedom between a ball screw and the carriage it drives [Slocum, 1992]. For different reasons, a recent paper demonstrated a clever way to improve the resolution of a leadscrew by effectively placing a flexural leadscrew in series with the mechanical leadscrew [Fukada, 1996], although no words to this effect are mentioned. While the screw-nut interface requires some level of torque before sliding takes place, the flexural leadscrew responds to arbitrarily small torque to give arbitrarily small resolution. Both of these valuable functions, exact constraint and smaller resolution, can be achieved in one simple device using a set of helical blade flexures. This idea is being used on the NIF precision linear actuator (see Chapter 8.5).

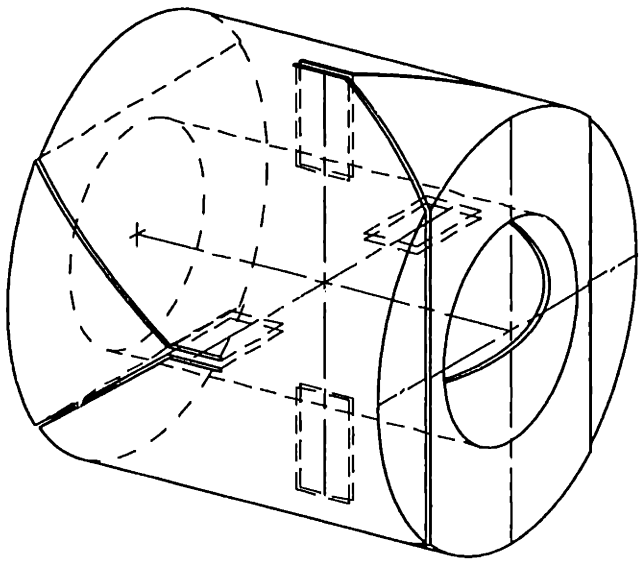


Figure 6-24 The ball-screw flexure for the NIF precision actuator requires two rotational degrees of freedom, a primary constraint against translation along the screw, and secondary constraints for the remaining degrees of freedom. Note, some hidden lines were removed to better show the main features.

Figure 6-24 shows the basic flexure design used for the NIF actuator. It resembles the compact pivot flexure of the last section except that it is hollow to allow the ball screw to pass through. In addition, the two hinge axes intersect to maximally condense the overall length, but this is not required in general. On the NIF actuator, there is another pivot some distance beyond the end of the screw so that the pivot pair provides free translation. Ordinarily the ball-screw flexure would have two pivots to provide free translation. Although it is not apparent from the figure, the blades are manufactured with a slight helix angle. Conceptually, if the blades were concentrated at the pitch diameter of the screw, then the proper helix angle would be perpendicular to the helix of the screw. Since the blades must lie outside the screw, then the actual helix angle must be somewhat smaller. This

condition does two things simultaneously: it aligns the blades to the reaction force; and it aligns out-of-plane motion of the blades to an insensitive direction of the screw. Effectively this creates a flexural screw with the same lead as the mechanical screw. Either one or both can be active since they appear the same to the system.

Before getting into the nature of helical blades, it is instructive to look over the finite-element results in Figure 6-25 for the one-quarter model. The von Mises stress plot in (a) highlights the blades but the point to notice is the gradient across the width. This is an indication of the relative flexibility of the annular junction between the two blades. The axial displacement plot in (b) shows a gradient along the length of each blade and nearly as significant a change in shading through the annular junction. Again this points to the annular junction as being a challenge to make much stiffer than the blades. There is some advantage to using a square cross section rather than an annulus if space permits.

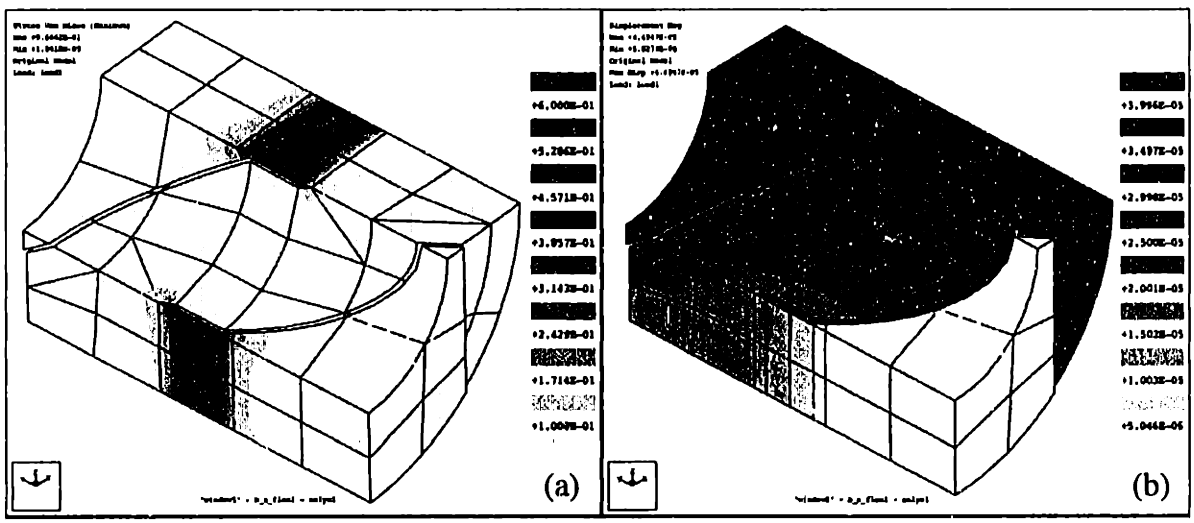


Figure 6-25 Contours of von Mises stress (a) and axial displacement (b) show the basic behavior.

Usually a blade flexure is initially straight so that a small out-of-plane displacement gives only a second-order axial displacement. The helical blade flexure is deliberately inclined with respect to the axis so that a small component of the out-of-plane displacement is along the axis. The angle of inclination varies with radius approximately as $r\,\theta/a$ using the parameters in Figure 6-26. The out-of-plane displacement also varies with radius in a way that compounds with the inclination. Using similar triangles, it is simple to relate the differential motion at any radius to the parameters of the flexure. This relation is then applied to the one particular radius $\bar{r}$ that has no net strain along the blade. The result is the effective lead of the flexure given by Equation 6.10.

$$L \equiv 2\pi \frac{dx}{d\theta} = 2\pi \bar{r}^2 \frac{\theta}{a} \tag{6.10}$$

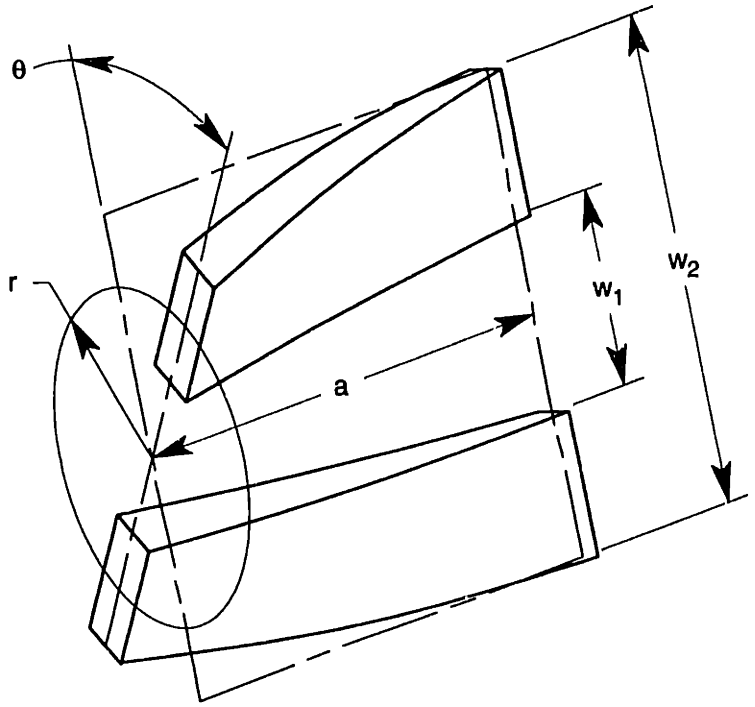


Figure 6-26 The effective lead of a helical blade flexure is governed by the parameters a, w_1, w_2, and θ.

To go further requires an assumption that the ends are constrained to remain parallel. Again using similar triangles, it is simple to determine the axial strain at any radius. Since the flexure is in equilibrium, the axial strain integrated over the cross section must be zero as indicated in Equation 6.11. Upon solving, Equation 6.12 gives the radius for zero strain, which then may be substituted back into Equation 6.10.

$$0 = \int_{r_1}^{r_2} \frac{\partial \varepsilon}{\partial \theta}\, dr = \int_{r_1}^{r_2} \left\{ \frac{\left(r^2 - \bar{r}^2\right)\theta}{a^2 + (r\theta)^2} \right\} dr \cong \frac{\theta}{a^2 + (\bar{r}\theta)^2} \int_{r_1}^{r_2} \left(r^2 - \bar{r}^2\right) dr \tag{6.11}$$

$$\bar{r}^2 \cong \frac{1}{3}\left({r_1}^2 + r_1 r_2 + {r_2}^2\right) = \frac{1}{6}\left({w_1}^2 + w_1 w_2 + {w_2}^2\right) \tag{6.12}$$

The axial strain that results from the helical shape also acts to stiffen the flexure in torsion beyond that for a flat blade. The approach used in Equation 6.13 to compute this effect is essentially a strain energy method (Castigliano's first theorem). The additional torsional stiffness due solely to the helix correctly goes to zero as the lead of the screw goes to zero.

$$\begin{aligned} k_{\theta_{helix}} &= 2E\,t\,a \int_{r_1}^{r_2} \left(\frac{\partial \varepsilon}{\partial \theta}\right)^2 dr \cong 2E\,t\,a \left\{ \frac{\theta}{a^2 + (\bar{r}\theta)^2} \right\}^2 \int_{r_1}^{r_2} \left(r^2 - \bar{r}^2\right)^2 dr \\ &\cong \frac{E\,t\left(w_2 - w_1\right)^3}{12a} \left(\frac{4{w_1}^2 + 7w_1 w_2 + 4{w_2}^2}{5{w_1}^2 + 5w_1 w_2 + 5{w_2}^2} \right) \left\{ \frac{2\pi\bar{r}}{L} + \frac{L}{2\pi\bar{r}} \right\}^{-2} \end{aligned} \tag{6.13}$$

6.2.5 A General Approach for Analyzing Flexure Systems

The most general approach for analyzing a flexure system is finite element analysis. Arbitrarily large, complex systems to very simple systems are readily modeled with commercial FEA software. For example, blade flexures typically modeled with shell elements are connected as necessary to other shell and/or solid elements to represent the whole system. It is hard to imagine a more flexible way to accurately analyze deflections and stresses in some spatially complex arrangement of flexures. Yet in several respects the approach presented here is more flexible than FEA especially early in the design cycle. The model is completely parametric and represents only the elements of interest, usually the flexures. It reports the stiffness and compliance matrices for the constrained system, and it includes column effects that a linear FEA code cannot. The main drawback is that the user must understand the basics of matrix algebra and transformation matrices, which is transparent to the user of an FEA code.

The basic assumption is that the flexure system can be modeled as parallel and series combinations of springs and that an equivalent spring for the system represents useful information, for example, the stiffness matrix. If desired, that information can be propagated back to individual springs, for example, to obtain local forces and moments. The key formalism in this approach is the six-dimensional vector used to succinctly represent three linear degrees of freedom and three angular degrees of freedom. We will deal strictly with force-moment vectors and differential displacement-rotation vectors. These vectors are related through the stiffness matrix or the compliance matrix of the spring. The concept of a three-dimensional stiffness matrix as expressed in Equation 6.14 may be more familiar. The six-dimensional stiffness matrix is assembled as blocks of three-dimensional matrices as Equation 6.15 shows. At times it may be easier to start by building the compliance matrix as in Equation 6.16. Converting from one to the other requires inverting the whole matrix rather than inverting separate blocks.

$$\mathbf{f} = \begin{bmatrix} f_x \\ f_y \\ f_z \end{bmatrix} = \begin{bmatrix} k_{xx} & k_{xy} & k_{xz} \\ k_{xy} & k_{yy} & k_{yz} \\ k_{xz} & k_{yz} & k_{zz} \end{bmatrix} \cdot \begin{bmatrix} d\delta_x \\ d\delta_y \\ d\delta_z \end{bmatrix} = \mathbf{K}_{f/\delta} \cdot d\delta \tag{6.14}$$

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{m} \end{bmatrix} = \left[\begin{array}{c|c} \mathbf{K}_{f/\delta} & \mathbf{K}_{f/\theta} \\ \hline \mathbf{K}_{m/\delta} & \mathbf{K}_{m/\theta} \end{array}\right] \cdot \begin{bmatrix} d\delta \\ d\theta \end{bmatrix} \qquad \mathbf{K}_{m/\delta} = \mathbf{K}_{f/\theta}{}^{T} \tag{6.15}$$

$$\begin{bmatrix} d\delta \\ d\theta \end{bmatrix} = \left[\begin{array}{c|c} \mathbf{C}_{\delta/f} & \mathbf{C}_{\delta/m} \\ \hline \mathbf{C}_{\theta/f} & \mathbf{C}_{\theta/m} \end{array}\right] \cdot \begin{bmatrix} \mathbf{f} \\ \mathbf{m} \end{bmatrix} \qquad \mathbf{C}_{\theta/f} = \mathbf{C}_{\delta/m}{}^{T} \tag{6.16}$$

Once the stiffness matrix or the compliance matrix is formed in one coordinate system (CS), it is a simple matter using the [6 x 6] transformation matrix to reflect it to any another (CS). Once expressed in the same CS, stiffness matrices are added to represent

parallel combinations, and compliance matrices are added to represent series combinations. This process is expressed in Equations 6.17 and 6.18 (also A.35 and A.36). Mixed combinations of parallel and series springs require like groups to be combined first then inverted as necessary to complete the combination. See Appendix A for a complete discussion of transformation matrices and parallel and series combinations.

$$\mathbf{K}_0 = \sum_i \mathbf{T}_{0/i} \cdot \mathbf{K}_i \cdot \mathbf{T}_{0/i}^{\mathrm{T}} \tag{6.17}$$

$$\mathbf{C}_0 = \sum_i \mathbf{T}_{0/i}^{-\mathrm{T}} \cdot \mathbf{C}_i \cdot \mathbf{T}_{0/i}^{-1} \tag{6.18}$$

The remainder of this section focuses on the details that make this approach truly useful. The first task is to derive the compliance matrix for the blade flexure. The solution depends on the CS so naturally we will choose the simplest one. Similarly, the stress matrix is derived so that maximum stresses in the blade are easy to calculate. Finally the details of constructing parallel-series spring models are presented.

6.2.5.1 The Compliance Matrix for a Blade Flexure

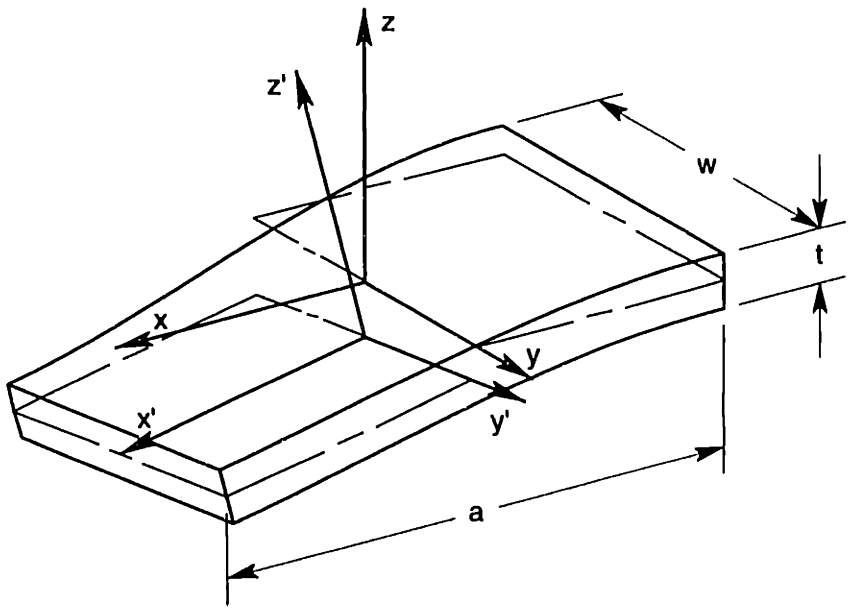


Figure 6-27 Imagining the CS's as rigid links to the ends of the blade, the application of forces and moments at these CS's results in respective displacements and rotations with no coupling between axes.

The flexion of a blade may be represented as movement between two CS's that attach to opposite ends of the blade. It would be more common to place a CS at each end but it is more convenient to place them initially coincident with the principal axes of the blade. Then a blade undergoing flexion appears as slightly displaced CS's in Figure 6-27. In similar fashion, the forces and moments may be represented about these CS's rather than at the ends where they are physically applied. This choice of CS's diagonalizes the compliance

matrix. Four of the six diagonal elements appear already in Chapter 2.6 as stiffnesses but without derivation or explanation. Here we go through each element with due care and add more generality with column effects and side-by-side blades. As in Figure 6-26, w_2 is the outside dimension of blades and w_1 is the inside dimension. In addition, it is necessary to represent the width of the individual blades or of a single blade as w.

We begin with the in-plane constraint directions. Axial compliance is the first diagonal element in the compliance matrix (6.19) and is so familiar that it needs no explanation. It is linear in all the key parameters and therefore is convenient for normalizing the other compliances. The second diagonal element (6.20) corresponds to a y-direction force, which produces combined bending and shear in the blade. It comes from the familiar fixed-guided beam equation and has an added term for the shear deformation. The section properties have been simplified for the side-by-side blade geometry. The last constraint direction is also the last element in the compliance matrix (6.21). Here the blade is in pure bending due to a moment about the z axis.

$$\mathbf{C}_{1,1} = c_x = \frac{a}{E\,t(w_2 - w_1)} \tag{6.19}$$

$$\mathbf{C}_{2,2} = c_y = c_x\left\{\frac{a^2}{{w_1}^2 + w_1 w_2 + {w_2}^2} + 2.4(1+\nu)\right\} \tag{6.20}$$

$$\mathbf{C}_{6,6} = c_{\theta_z} = \frac{12\,c_x}{{w_1}^2 + w_1 w_2 + {w_2}^2} \tag{6.21}$$

The remaining diagonal elements are out-of-plane directions usually considered as degrees of freedom. The elements corresponding to z and θ_y directions are similar to y and θ_z directions from before and differ mainly in a factor t^2 in the denominator rather than w^2. However, there are two subtle distinctions. As noted earlier in Section 6.2.2, the equations for bending compliance should include a factor $(1 - \nu^2)$ to account for the Poisson effect. The other factor is the effect of an axial force, either compressive or tensile. The solutions for fixed-guided and cantilever boundary conditions are available from numerous sources, for example, [Vukobratovich and Richard, 1988], [Young, 1989] and [Smith, 1998]. Unfortunately the equations are not particularly convenient or intuitive, involving trigonometric functions for compression and hyperbolic functions for tension. Upon substituting power series approximations, the two types of functions become the same when expressed with a positive or negative axial force. The number of terms required in the power series depends upon how closely the flexure operates to the critical buckling load. We will keep enough terms to have good accuracy through one-half the critical load.

The third diagonal element for the z-direction force (6.22) follows from the series approximation of the axially loaded, fixed-guided beam. The effect of the axial force is contained within the square brackets. As expected, the compliance decreases for a positive

tensile force and increases for a negative compressive force. The approximation differs from the exact solution by only 6.6% at one-half the critical load. Usually we would not venture this close to buckling unless the intent is a zero-stiffness condition. Then it is necessary to use the exact solution from the references.

$$\mathbf{C}_{3,3} = c_z \cong c_x\left\{\left(1-\nu^2\right)\left(\frac{a}{t}\right)^2\left[1-\frac{6}{5}\gamma+\frac{17}{35}\gamma^2-\frac{62}{315}\gamma^3\right]+2.4(1+\nu)\right\} \qquad (6.22)$$

$$\gamma \equiv \frac{f_x c_x a}{t^2} \qquad \gamma_{cr} = \frac{\pi^2}{12}$$

The fifth diagonal element corresponds to a moment load applied about the *y*-axis. When an axial force also exists, we must be careful to apply it through the CS rather than at the end of a cantilever beam as in the published solution. Figure 6-28 shows the subtle difference in the way the axial force is applied to the end in (a) and through the CS in (b). The differential equation for the model in (b) has one additional term, $f\ \theta\ (a - x)$. The solution proceeds in a similar way and simplifies for compression to a single sine function rather than a tangent function or the equivalent hyperbolic function for tension. Applying the series approximation results in the fifth diagonal element (6.23). In this case the compliance increases under tension and decreases under compression.

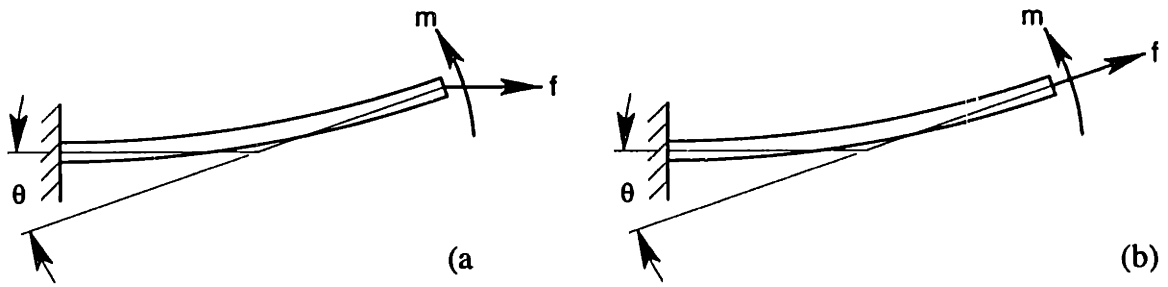


Figure 6-28 The tensile force stiffens the beam in (a) while making it more compliant in (b).

$$\mathbf{C}_{5,5} = c_{\theta_y} = \frac{12\,c_x}{t^2}\left(1-\nu^2\right)\left[1+2\gamma+\frac{6}{5}\gamma^2+\frac{12}{35}\gamma^3\right] \qquad (6.23)$$

The fourth diagonal element (6.24) corresponds to twisting of the blade. This relation comes from the parallel combination of two effects. The first term within the braces is simple twist with no end effects. This solution is given in several references, for example, [Timoshenko and Goodier, 1951] and [Young, 1989]. The second term brings in the end effects by considering the blade as a series of thin fixed-guided beams distributed across the width. The deflection varies as the radius from the twist axis, and an integration provides the cumulative effect. Since c_z appear in this relationship, it accounts for the Poisson and axial-force effects. This relation agrees well with a parameterized finite-element study.

$$\mathbf{C}_{4,4} = c_{\theta_x} = 12\left\{\frac{1}{2(1+\nu)}\left(4+2.52\frac{t}{w}\right)\frac{t^2}{c_x}+\frac{{w_1}^2+w_1 w_2+{w_2}^2}{c_z}\right\}^{-1} \tag{6.24}$$

6.2.5.2 The Stress Matrix for a Blade Flexure

The distribution of stress through a blade varies greatly depending on the direction of the applied force or moment. For purposes of sizing the blade, however, it is sufficient to consider only the maximum absolute value of stress resulting from applied forces and moments. This information may be represented in a diagonal matrix similar to the compliance matrix. The judicious choice of the CS causes symmetric distributions of stress due to individual components of the force-moment vector. Each distribution will have the maximum absolute value of stress always extending to the eight corners of the blade. There is one worse-case corner where all the stresses are in general alignment (perfectly aligned if not for shear stresses). Then a somewhat conservative estimate of the combined maximum absolute principle stress is the simple sum of the individual maximum's as calculated from the stress matrix multiplied by the applied force-moment vector.

In keeping with the order presented for the compliance matrix, the three terms that correspond to constraint directions are considered first for the stress matrix. The first diagonal element for axial loading (6.25) is simply the inverse of the cross sectional area. It is convenient to use for scaling the other terms. For the second diagonal element (6.26), the y-direction force produces combined bending and shear in the blade. A somewhat conservative approach treats the maximum absolute value of principle stress as if aligned with the x-axis. This simplifies the final step of combining all the stress components but yields a slightly higher combined stress than a more rigorous analysis. The last constraint direction and also the last element in the stress matrix (6.27) is much simpler because the blade is in pure bending from a z moment.

$$\mathbf{S}_{1,1} = s_x = \frac{1}{t(w_2 - w_1)} \tag{6.25}$$

$$\mathbf{S}_{2,2} = s_y = s_x\left\{\frac{3aw_2}{2\left({w_1}^2+w_1 w_2+{w_2}^2\right)}+\sqrt{\left(\frac{3aw_2}{2\left({w_1}^2+w_1 w_2+{w_2}^2\right)}\right)^2+1}\right\} \tag{6.26}$$

$$\mathbf{S}_{6,6} = s_{\theta_z} = s_x\frac{6w_2}{{w_1}^2+w_1 w_2+{w_2}^2} \tag{6.27}$$

The stresses due to the remaining three elements usually result due to required motions of the flexure but they will be written for applied loads. The third diagonal element (6.28) corresponds to a z-direction force. Although the blade experiences combined bending and shear loading, the shear stress is usually not significant given normal blade

proportions. It is excluded in preference of keeping a simple expression for the column effect from an axial force. As expected the bending stress increases with a negative compressive force and decreases with a positive tensile force. The fifth diagonal element corresponds to a *y*-moment (6.29) and exhibits slightly more complicated behavior. The axial force causes the bending moment to be nonsymmetric along the length of the blade.[1] The bending moment is maximum at the fixed end for tension and minimum for compression. The use of the singularity function in the equation turns off the effect of a compressive force when γ is negative, leaving just the applied moment as the maximum.

$$\mathbf{S}_{3,3} = s_z = s_x \frac{3a}{t}\left[1 - \gamma + \frac{6}{5}\gamma^2 - \frac{51}{35}\gamma^3\right] \tag{6.28}$$

$$\mathbf{S}_{5,5} = s_{\theta_y} = s_x \frac{6}{t}\left[1 + 6\langle\gamma\rangle^1 + 6\langle\gamma\rangle^2 + \frac{12}{5}\langle\gamma\rangle^3\right] \tag{6.29}$$

As described in the previous section, there are two effects to consider when a blade twists on axis. The first is simple twist with no end effects. The shear stress produced is maximum far away from the corners and therefore may be safely ignored. The other effect is bending of the blade as a fixed-guided beam with the maximum stress occuring at the corners. The fourth diagonal element (6.30) represents this effect by using s_z as the reference rather than s_x. As a result it also represents column effects.

$$\mathbf{S}_{4,4} = s_{\theta_x} = s_z \frac{6\,w_2}{{w_1}^2 + w_1 w_2 + {w_2}^2} \tag{6.30}$$

The stress matrix multiplied by the force-moment vector gives a vector of stress components that may be positive or negative depending on the loading. It is useful to look at each component to understand which loads are most significant. The worse-case stress is conservatively estimated by summing the absolute values of the stress components.

6.2.5.3 Parallel-Series Spring Models

The use of parallel and series combinations of springs is fairly common in engineering analysis. Working in 6-D is certainly less common but the basic concept of parallel-series combinations is no different. Most of the gritty details are in the matrix algebra carried out by the computer. The most challenging aspect usually is in setting up the transformation matrices. Perhaps the best way to understand the modeling aspect is to work through an example step by step. The X-Y-θ_z flexure stage, shown in Figure 6-11 on page 184, is a good example to represent several levels of parallel and series combinations of springs.

---

[1] The asymmetry of the bending moment is a consequence of the way the blade is loaded along the *x* axis of one CS. A symmetric bending moment would result if the axial force bisected the *x* axes of both CS's. This symmetric case is more appropriate for the cross blade flexure but not for more general arrangements.

Figure 6-29 shows the spring model of the same flexure stage with three actuators. The steps required to set up and analyze the model are enumerated below.

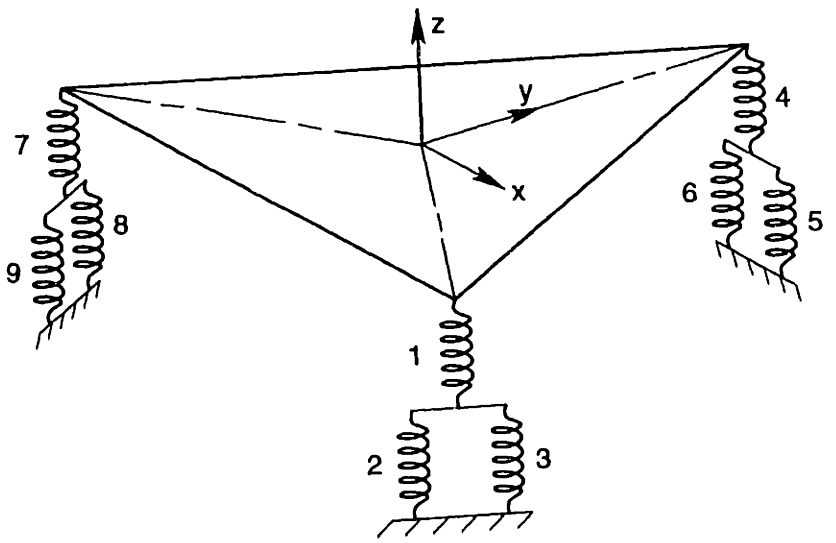


Figure 6-29 The X-Y-θ_Z flexure stage appearing in Figure 6-11 is modeled with nine springs in parallel and series combinations as shown.

1) Identify the main member (the thing being moved or supported) and attach the base coordinate system CS_0 at a convenient location. It is simple to reflect the results to a different CS if desired.

2) Identify all the separate paths from the main member to ground. There are six in the example, three flexure supports and three actuators.

3) Identify and number each spring in each path. There are nine in the example, six blades and three actuators.

4) Assign a unique CS to each spring. Number these CS_1 to CS_n. Then create [6 x 6] transformation matrices to represent the spatial relationships between the local CS's and the base CS_0.

5) Create as many compliance matrices as required to represent all the springs with respect to their own local CS's. There are only two for the example, one matrix for six identical blades and one matrix for three identical actuators.

6) Reflect the compliance matrix for each spring to the base CS_0 using the transformation matrices, thus creating *n* unique compliance matrices. Number these $\mathbf{C}_1$ to $\mathbf{C}_n$.

7) Identify the springs that form either series or parallel combinations. When reflected to the same CS, series springs experience the same load while parallel springs experience the same deflection. The example has three sets of parallel springs, 2-3, 5-6 and 8-9.

8) Add the stiffness matrices of springs in parallel and add the compliance matrices of springs in series. Indicate these new equivalent springs by the spring numbers they represent. The equivalent springs for the example are $\mathbf{K}_{2\text{-}3}$, $\mathbf{K}_{5\text{-}6}$, and $\mathbf{K}_{8\text{-}9}$.

9) Repeat steps 7 and 8 using the equivalent springs in place of the ones they represent. Stop when there is only one equivalent spring remaining. The example requires a total of three combination steps before reaching the equivalent spring for the system. The second step is a series combination resulting in $\mathbf{C}_{1\text{-}2\text{-}3}$, $\mathbf{C}_{4\text{-}5\text{-}6}$, and $\mathbf{C}_{7\text{-}8\text{-}9}$. The last step is a parallel combination resulting in $\mathbf{K}_0$, the equivalent spring for the system.

There are a number of uses for the system stiffness and compliance matrices. Presumably there is some requirement that drives the design to have a certain level of stiffness in the constraint directions and certain freedoms in other directions. Specific load cases may be applied to ascertain deflections or certain motions may be specified to determine resulting reaction forces. The sizes and locations of blades are easily modified to evaluate design changes. Details about individual blades such as stresses or reaction forces require the applied load or specified motion to be propagated back through the combination process, being careful to apply loads to springs in series or motions to springs in parallel. A clearly labeled sketch of the model for each step in the combination process will help avoid confusion and mistakes.

The fine details of these analysis steps appear in the flexure system analysis program in Section 6.3. The program documents the example of the X-Y-θ_Z flexure stage discussed here. In particular it shows how to set up the [6 x 6] transformation matrices and reflect compliance matrices to the base CS_0. It also shows how easily a fairly complex system of blades is modeled with parallel and series combinations of springs.

A slightly more advanced topic is the modeling of column effects in a system of flexures. It was not an important effect in the X-Y-θ_Z flexure stage so it was not introduced then. The compliance matrix and the stress matrix for individual blades account for local column effects, but it is necessary to include the system effects at the system level. Fortunately this is straightforward to do with additional springs that represent the column behavior. This behavior may occur when a series combination carries a significant axial force, for example, when two blades lie in the same plane so as to act like one much longer blade. The effect of the axial force is modeled with a new spring placed in parallel with the series combination. The stiffness of the new spring depends only on the axial force and the distance between the two CS's of the series combination, as in Equation 6.31. It may be positive or negative for a tensile or compressive force, respectively. When reflecting this stiffness to the base CS_0, the transformation matrix should be the same as the most mobile blade in the series combination. Then it may be combined as any of the other springs.

$$\mathbf{K}_{3,3} = \frac{f_x}{L} \tag{6.31}$$

6.3 Friction-Based Design of Kinematic Couplings[I]

Friction affects several aspects important to the design of kinematic couplings, but particularly the ability to reach the centered position is fundamental. It becomes centered when all constraints are fully engaged even though a small uncertainty may exist about the ideal center where potential energy is minimum. For many applications, centering ability is a good indicator for optimizing the coupling design. Typically, the coupling design process has been largely heuristic based on a few guidelines [Slocum, 1992]. Symmetric forms of the basic kinematic couplings are simple enough to develop closed-form equations for centering ability. More general configurations lacking obvious symmetries are difficult to model in this way. A unique kinematic coupling for large, interchangeable optics assemblies in the NIF motivated the development of computer software to optimize centering ability. The program, written in Mathcad™ Plus 6, appears in Section 6.3.

6.3.1 Friction Effects in Kinematic Couplings

Friction affects at least four important characteristics of a kinematic coupling as indicated by order-of-magnitude estimates that all include the coefficient of friction μ. These estimates are listed in Equations 6.32 through 6.35 and are described below.

Repeatability
$$\frac{f}{k} \approx \mu \left(\frac{2}{3R}\right)^{1/3} \left(\frac{P}{E}\right)^{2/3} \tag{6.32}$$

Kinematic support
$$|f_t| \le \mu f_n \tag{6.33}$$

Stiffness
$$k_t = k_n \frac{2-2\nu}{2-\nu}\left(1-\frac{f_t}{\mu f_n}\right)^{1/3} \approx 0.83\, k_n \tag{6.34}$$

Centering ability
$$\frac{f_c}{f_n} \approx 0.5 - 1.3\mu \tag{6.35}$$

Tangential friction forces at the contacting surfaces may vary in direction and magnitude depending how the coupling comes into engagement. This affects the repeatability of the coupling and the kinematic support of the precision component. The estimate for repeatability is the unreleased frictional force multiplied by the coupling's compliance. The estimate is derived as if the coupling's compliance in all directions is equal to a single Hertzian contact carrying a load P and having a relative radius R and elastic modulus E. The frictional force acts to hold the coupling off center in proportion to the compliance. This estimate will underestimate the repeatability if the structure of the coupling is relatively compliant compared to the contacting surfaces.

[I] This material was previously published in condensed form [Hale, 1998].

Kinematic support is important for stability of shape of the precision component being supported by the coupling. The estimate for kinematic support simply gives a bound on the magnitude of friction force acting at any contact surface. A sensitivity analysis of the precision component will determine a tolerable level of friction that the coupling can have. This may drive the design to include flexure elements and/or procedures to release stored energy. If repeatable engagement is not so important, then constraints using rolling-element bearings offer very low friction. For example, a pair of cam followers that contact with crossed axes is equivalent to a sphere on a flat but with twenty or more times less friction.

In some cases frictional overconstraint is valuable for increasing the overall system stiffness. Provided the tangential force is well below what would initiate sliding, the tangential stiffness of a Hertzian contact is comparable to the normal stiffness [Johnson, 1985]. This motivated the widely spaced vee used to stiffen the first torsional mode of NIF optics assemblies (see Figure 6-9).

Centering ability can be expressed as the ratio of centering force to nesting force. The estimate provided is typical for a symmetric three-vee coupling. A larger ratio means the coupling is better at centering in the presence of friction. It is also convenient to express centering ability as the coefficient of friction where this ratio goes to zero. For the estimate, the limiting coefficient of friction is 0.5/1.3 = 0.38. The coupling will center if the real coefficient of friction is less than the limiting value.

6.3.2 Centering Ability of the Basic Kinematic Couplings

We begin by studying the centering ability of the basic kinematic couplings because their geometry is relatively simple and familiar. When brought together initially off center, the constraints engage sequentially as the coupling seeks a path to center.[I] This path becomes better defined as more constraints engage. For example, five constraints allow the coupling to slide along one well-defined path. Four constraints allow motion over a two-dimensional surface of paths and so forth. Although there are infinitely many paths to center, only the limiting case is of practical interest for determining centering ability. Further, it is reasonable to expect the limiting case to be one of six possible paths that have five constraints engaged.[II] This point will be demonstrated using examples and simple logic.

The simplest example is the symmetric three-vee coupling. Figure 6-30 shows two ways that the coupling may slide to center depending on its initial misalignment. In (a), two vee constraints are fully engaged and the third is off center giving a total of five constraints. The two vees define an instantaneous center of rotation. The off-center vee transforms its share (one-third) of the nesting force into a centering moment about the instant center.

[I] It is a rare possibility that initial contact can occur at two places simultaneously.

[II] The exception to this statement is the ball-cone constraint since the cone provides only one constraint until the ball is fully seated. A tetrahedral socket remedies this situation.

Equation 6.36 describes the centering force at the center of the coupling due to this moment, where α is the angle to each surface from the plane of the vees. This path is the limiting case along with five other symmetrically identical paths.[1] In (b), two off-center vees transform their share (two-thirds) of the nesting force into a centering force given by Equation 6.37. This causes the coupling to translate along the fully engaged vee. With only four constraints, the coupling is also free to roll but there is no moment in this direction. Given freedom like this, the coupling will move in the general direction of the centering force until the next constraint engages and forces it along a more resistive path to center.

$$\frac{f_c}{f_n} = \frac{\sin\alpha - \mu\cos\alpha}{2(\cos\alpha + \mu\sin\alpha)} - \frac{\sqrt{3}\,\mu}{3\cos\alpha} \tag{6.36}$$

$$\frac{f_c}{f_n} = \frac{\sqrt{3}\sin\alpha\sqrt{4+3\tan^2\alpha} - 4\mu}{3\left(\cos\alpha\sqrt{4+3\tan^2\alpha} + \sqrt{3}\mu\tan\alpha\right)} - \frac{\mu}{3\cos\alpha} \tag{6.37}$$

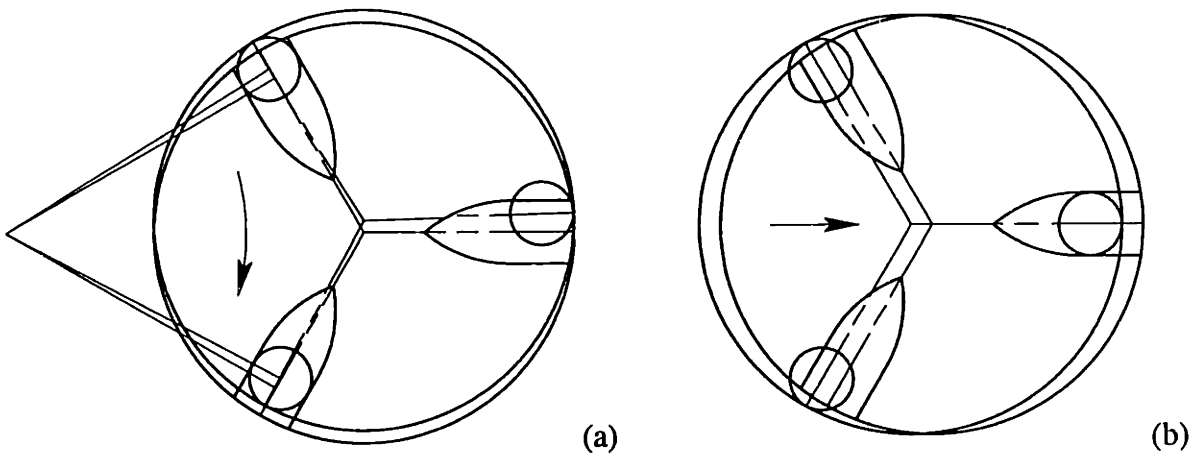


Figure 6-30 In (a), the three-vee coupling slides on five constraints producing rotation about an instant center. In (b), the coupling slides on four constraints in the general direction of the centering force.

The three-tooth coupling behaves similarly to the three-vee coupling, but the centering force with five constraints engaged is very difficult to model in closed form. Table 6-1 shows the limiting coefficient of friction for both three-vee and three-tooth couplings as calculated by the kinematic coupling analysis program. There is negligible difference between the two types over the range 45° to 55°. The three-vee coupling has slightly better centering ability at the nearly optimal angle of 60°. With only four constraints engaged, the centering force causes simple translation of the three-tooth coupling, which leads to a reasonable closed-form solution given by Equation 6.38. Graphs of Equations 6.36, 6.37 and 6.38 versus the coefficient of friction μ appear in Figure 6-32 (a). The point to notice is that the centering force decreases when the fifth constraint engages.

---

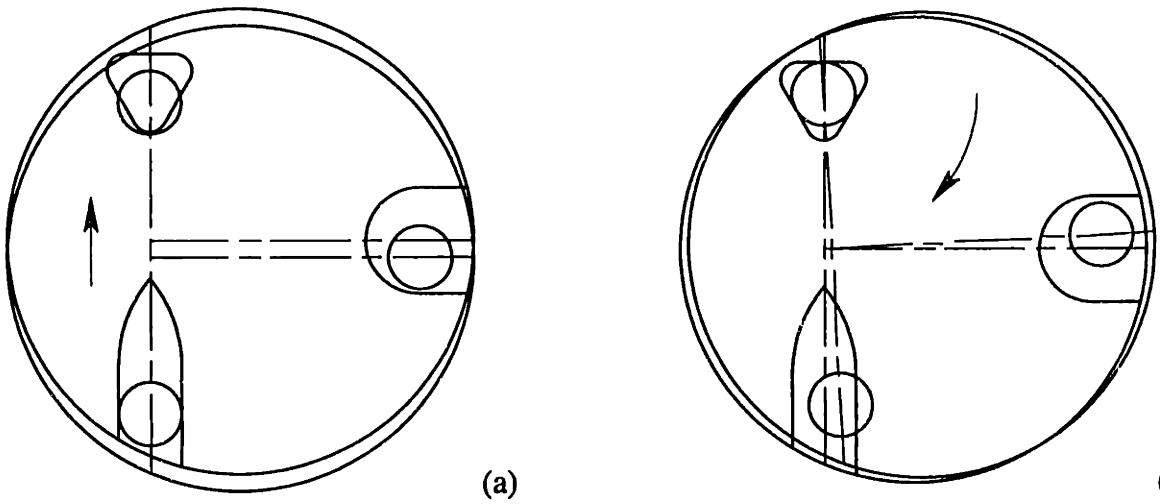
[1] The coupling may rotate clockwise or counterclockwise about any of three instant centers.

Angle of Inclination α	45°	50°	55°	60°	65°
Three-Vee Coupling	0.317	0.338	0.354	0.364	0.365
Three-Tooth Coupling	0.319	0.339	0.351	0.352	0.339

Table 6-1 The limiting coefficient of friction versus angle for three-vee and three-tooth couplings.

$$\frac{f_c}{f_n} = \frac{\sin\alpha\sqrt{4+\tan^2\alpha}-4\mu}{2\left(\cos\alpha\sqrt{4+\tan^2\alpha}+\mu\tan\alpha\right)} \tag{6.38}$$

An aspect hidden by the symmetry in the previous examples is the possibility that a path with five constraints engaged may have greater centering force than a different path with only four constraints. However as the coupling continues toward center, the centering force cannot increase as the fifth constraint engages. The tetrahedron-vee-flat coupling exhibits behavior of the type shown in Figure 6-32 (b). Usually the centering ability will be limited by the path shown in Figure 6-31 (a). The centering force for this path is given in Equation 6.39, where α is the vee angle and β is the tetrahedron angle. The opposite-direction path has one less constraint and greater centering force as described by Equation 6.40. This solution is also representative of the cone-vee-flat coupling. However in Figure 6-31 (b), a different path with five constraints has typically greater centering force as described by Equation 6.41. The main point is that all six paths having five constraints engaged must be considered to determine centering ability.



(a) (b)

Figure 6-31 The tetrahedron-vee-flat coupling has six unique paths having five constraints engaged. The path in (a) usually limits the centering ability of the coupling. The path in (b) could limit the centering ability if the vee angle is too shallow.

$$\frac{f_c}{f_n} = \frac{\sin\beta\sqrt{4+\tan^2\beta}-4\mu}{6\left(\cos\beta\sqrt{4+\tan^2\beta}+\mu\tan\beta\right)} - \frac{\mu}{3\cos\alpha} - \frac{\mu}{3} \tag{6.39}$$

$$\frac{f_c}{f_n} = \frac{\sin\beta - \mu\cos\beta}{3(\cos\beta + \mu\sin\beta)} - \frac{\mu}{3\cos\alpha} - \frac{\mu}{3} \tag{6.40}$$

$$\frac{f_c}{f_n} = \frac{\sqrt{3}}{3}\left\{\frac{\sin\alpha - \mu\cos\alpha}{\cos\alpha + \mu\sin\alpha} - \mu\right\} \tag{6.41}$$

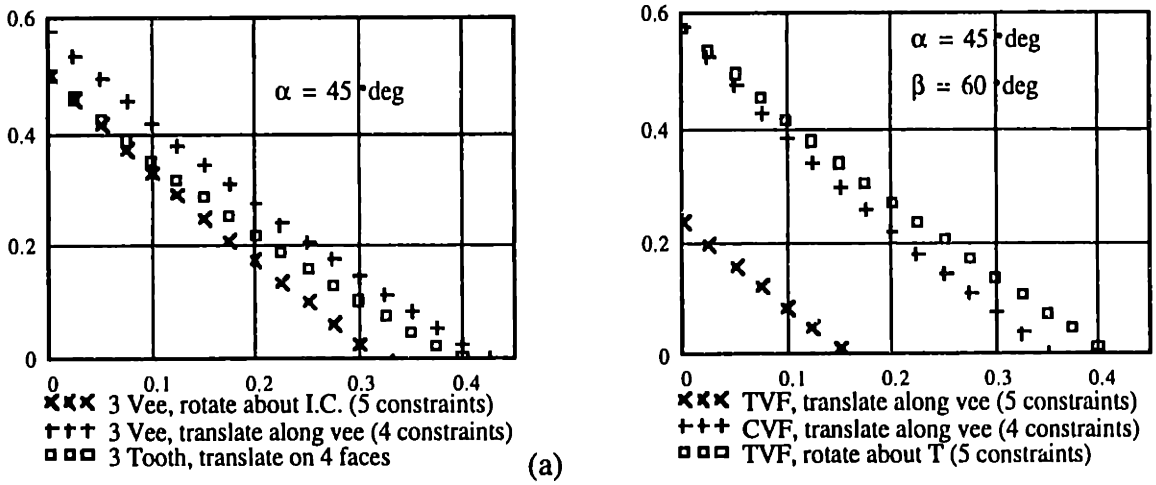


Figure 6-32 Normalized centering force versus coefficient of friction. In (a), the three-vee and three-tooth couplings being symmetric always have five-constraint paths with less centering force than all four-constraint paths. In (b), the tetrahedron-vee-flat coupling may have five-constraint paths with more centering force than some four-constraint paths.

The difficulty encountered with more general configurations of kinematic couplings comes first in determining the six possible paths to center then in developing compatibility and equilibrium equations for complex geometry. Even when the coupling has relatively simple geometry, the equations are rather tedious to develop. Compound this with the problem of optimizing the design and it becomes obvious that a systematic, computer-based approach is essential for designing more general configurations of kinematic couplings.

6.3.3 A General Approach for Optimizing Centering Ability

There are three basic steps required for optimizing the ability of a kinematic coupling to become centered.

- Represent the geometric arrangement of six constraints with model parameters. Some parameters will vary during the optimization while others may be fixed by strict design constraints. This step requires user knowledge and input following the general format provided by the program.
- Determine the six paths to center and the coefficient of friction for each path that just impedes sliding (i.e., zero centering force). This step is purely algorithmic.
- Vary the model parameters to maximize the minimum coefficient of friction among the six paths. This step could be strictly algorithmic but experience has shown the user's

intelligence to be very valuable. Keeping the user in the loop builds intuition and understanding of the tradeoffs involved.

These three steps are explained through the example that motivated this work, the NIF optics assembly.

Orthographic and isometric views of one NIF optics assembly, discussed previously in Section 6.1.3.3 and again in Chapter 7, appear in Figure 6-33. The six constraints that kinematically support the optics assembly are represented in the figure by reaction force vectors numbered 1 to 6. In the model, each constraint is represented by a simple spring stiffness defined along the z-axis of a local x-y-z coordinate system (CS). The use of six-dimensional vectors and [6 x 6] transformation matrices completely automates the development of compatibility and equilibrium equations, but the local CS's can be challenging to set up even for experienced users. See Appendix A for a complete discussion of transformation matrices.

In the program, two equations completely define the local CS's for any arrangement of six constraints. Equation 6.42 defines the overall transformation matrix as a series of simple transformations: x-y-z translation, z rotation, y rotation and x rotation. The user must decide the proper number and order of simple transformations for the application, but usually there will be six parameters, p_1 through p_6. These parameters come from the column of Equation 6.43 that corresponds to the particular constraint under consideration. In the example, the CS definition for constraint 1 is: rotate θ_2 about global x, rotate $-\theta_1$ about global y, and translate in global x-y-z (-0.385, -0.315, -0.8). An equivalent definition is: translate in local (initially global) x-y-z (-0.385, -0.315, -0.8), rotate $-\theta_1$ about local y, and rotate θ_2 about local x. During the optimization, constraints 1 and 3 symmetrically change orientation according to θ_1 and θ_2. Similarly, constraints 2 and 4 change according to θ_1 and θ_3, and constraints 5 and 6 change according to θ_4. The user determines the number of variable parameters used in the optimization. In addition, the user must supply the nesting force vector and a vector of spring stiffnesses, one for each constraint.

$$T(p) := T_{xyz}(p_1, p_2, p_3) \cdot R_z(p_6) \cdot R_y(p_5) \cdot R_x(p_4) \tag{6.42}$$

$$P(\theta) := \begin{bmatrix} -0.385 & -0.385 & 0.385 & 0.385 & -0.397 & 0.397 \\ -0.315 & -0.315 & -0.315 & -0.315 & -0.241 & -0.241 \\ -0.8 & -0.8 & -0.8 & -0.8 & 0.85 & 1.31 \\ \theta_2 & -\theta_3 & \theta_2 & -\theta_3 & 90 \cdot \text{deg} & 90 \cdot \text{deg} \\ -\theta_1 & -\theta_1 & \theta_1 & \theta_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\theta_4 & \theta_4 \end{bmatrix} \tag{6.43}$$

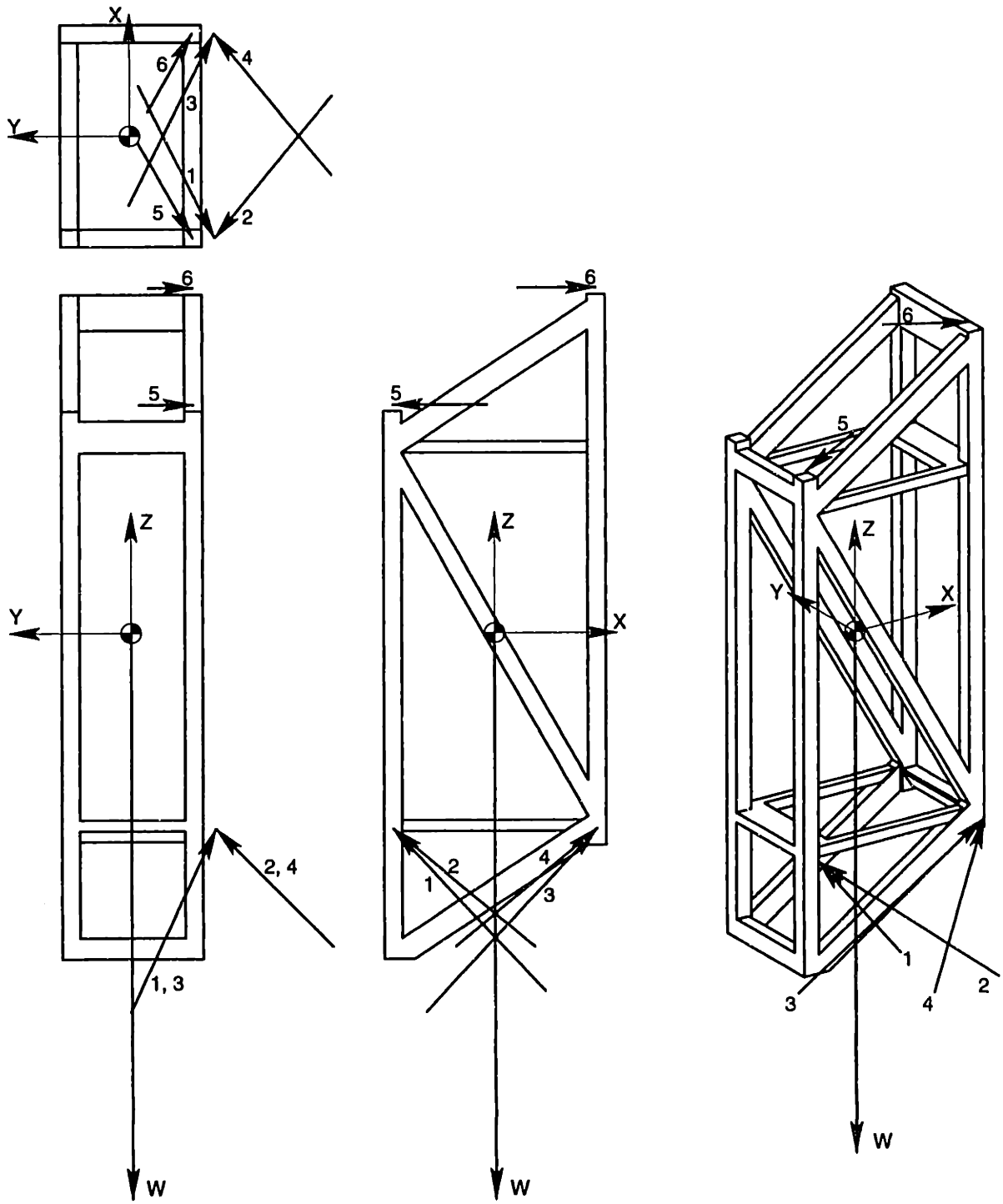


Figure 6-33 Orthographic and isometric views of one NIF optics assembly. The arrows numbered 1-6 represent the six constraints that support the assembly. The arrows are proportional to the reaction forces.

The second step uses the constraint arrangement, the nesting force vector and constraint stiffnesses to determine the six paths to center and the coefficient of friction for each path that yields zero centering force. If the constraint arrangement is truly kinematic, then the parallel combination of six constraint stiffnesses will form a [6 x 6] stiffness matrix for the coupling that is full rank. Equation 6.17 (also A.35) on page 198 expresses

the process of reflecting the stiffness matrix from a local CS_i to a global CS_0 then summing all the reflected stiffness matrices. Since there is only one nonzero term in $\mathbf{K}_i$, the matrix equation reduces to an outer product of the third column of $\mathbf{T}_{0/i}$ times the simple stiffness.

The technique used to find the vector direction for sliding is rather elegant. A degenerate stiffness matrix is computed for only five constraints engaged. This matrix has five nonzero eigenvalues (if the coupling is truly kinematic). The eigenvector corresponding to the zero eigenvalue gives the direction in the global CS that the coupling will slide for that particular set of five constraints. This procedure is executed six times, once for each constraint not engaged.

The procedure used to determine the coefficient of friction for zero centering force is coded in just three expressions but becomes rather complicated to explain. It should help to review the comments and equations in the program. Starting with one sliding vector in the global CS, the same transformation matrices are used to create six local sliding vectors, one for each constraint. Then a six-dimensional, force-moment vector is calculated for each constraint using the Coulomb law of friction and a unit normal force. Transforming back to the global CS, the vectors are assembled into a matrix that when multiplied by a vector of normal forces (i.e., the contact force for each constraint) gives the resultant force-moment vector for the coupling. The inverse of this matrix is useful because it takes the applied force-moment vector (i.e., the nesting force) and gives back the contact forces for a given coefficient of friction. Finally, the row of this matrix equation corresponding to the constraint not engaged is solved for the coefficient of friction that makes its normal force zero. This gives the coefficient of friction that just impedes sliding. This procedure is executed six times, once for each constraint not engaged. The minimum of six results is the limiting coefficient of friction for the particular constraint arrangement.

The last step is to adjust the constraint arrangement in a manner that maximizes the limiting coefficient of friction. The user presumably has set up the program to include some number of variable parameters given by the vector θ. The optimization approach is a graphical technique that provides the user with sufficient visual feedback to find the optimum within a few iterations. The user must enter three vectors: the nominal vector θ and the range for the graph, θ_{min} and θ_{max}. For each parameter, the program generates one curve over its range while holding all other parameters at their nominal values. All curves appear in one graph versus a normalized range of model parameters. The user adjusts the nominal values attempting to maximize the limiting coefficient of friction.

Figure 6-34 shows two graphs for the NIF optics assembly with slightly different nominal parameters. The horizontal dashed line indicates the limiting coefficient of friction for the nominal parameter set. The optimal parameter set is apparent in (a) because a change to any one parameter reduces the limiting coefficient of friction. It is useful to observe the suboptimal parameter set in (b). By adjusting the nominal values towards the peaks in the curves, the user soon converges to the optimal parameter set.

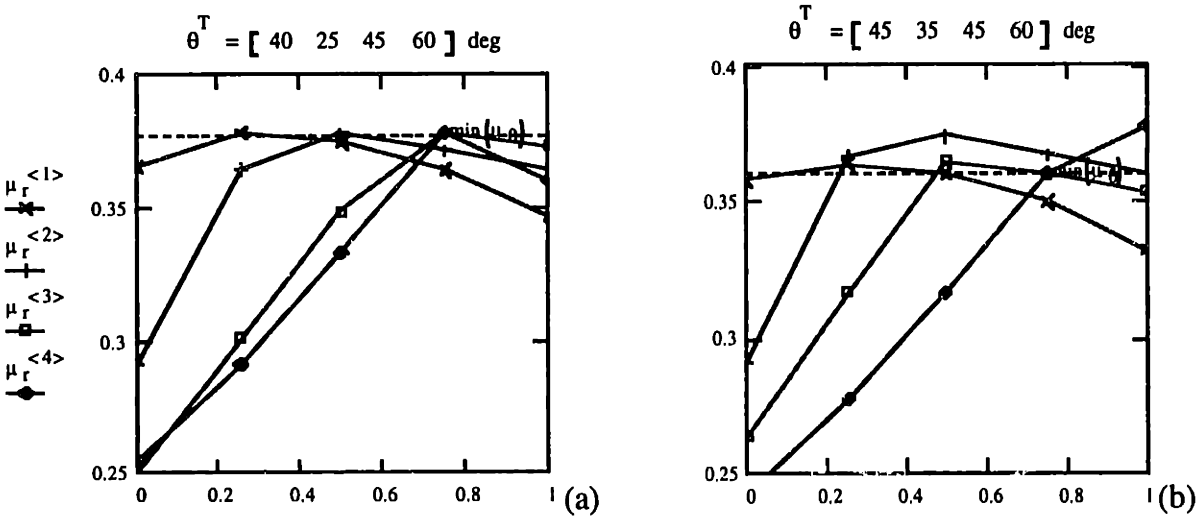


Figure 6-34 Limiting coefficient of friction versus a normalized range of model parameters. Each curve corresponds to one individually varying parameter with the others held constant at nominal values. The horizontal dashed line indicates the nominal parameter set. The graph in (a) shows the optimal configuration while (b) is suboptimal. In this example, the markers occur in 5° increments of the angle parameters.

6.4 Mathcad™ Documents for Generalized Kinematic Modeling

The Mathcad computing environment is ideal for engineering analysis where it is important for the user to understand the mechanics of the program. All the code is in plain sight and in familiar mathematical format. It is easy to add text and drawings to better explain how to use the program, usually referred to as a document.

6.4.1 Flexure System Analysis Program

This program helps the user develop a parameterized model of a flexure system so that sizes and locations of blades may be chosen to achieve key performance requirements. Each blade in the system is represented by the compliance matrix of a spring in a local coordinate system (CS). All blades are then reflected to a common CS to allow simple addition of compliance matrices for springs in series or of stiffness matrices for springs in parallel. The end result is the equivalent compliance matrix for the system and its reciprocal, the stiffness matrix. Applied loads or specified motions may then be applied to these matrices to determine the resulting behavior. Loads may be propagated back to local blades to determine the maximum stresses.

This program uses six dimensional vectors and [6 x 6] matrices to deal simultaneously with both linear and angular measures. The first three components correspond to linear measures such as force and displacement, and the last three correspond to angular measures such as moment and rotation. The same arrangement applies to the rows and columns of the matrices used for coordinate transformation, stiffness and compliance. In addition, the positions and orientations for an arbitrary number n of flexure blades are conveniently stored in an [m x n] matrix. Each column describes a sequence of moves (typically m = 6) required to go from the global X-Y-Z CS to a local x-y-z CS as shown in the figure. This matrix may also contain parameters that the user may vary to optimize the configuration.

Forces and displacements for the flexure system are represented in the global CS while forces and displacements at a particular flexure are represented in its local x-y-z CS.

The transformations between the X-Y-Z CS and x-y-z CS's are done as a sequence of moves. The number and order may vary to suit the application, but they must be the same for each flexure since they are stored in a matrix. The basic moves are: translation in x-y-z, rotation about x, rotation about y and rotation about z. The sequence may be interpreted as moves relative to the base CS or in the opposite order as moves relative to the local CS. Two rules express these interpretations for a sequence of transformation matrices: 1) post multiply to transform in the local CS and 2) pre multiply to transform in the base CS. A typical sequence relative to the base CS is: rotate about the X axis, rotate about the Y axis, rotate about the Z axis, and translate (X, Y, Z). The equivalent sequence relative to the local CS is: translate (x, y, z) (initially coincident with the base CS), rotate about the z axis, rotate about the y axis, and rotate about the x axis. The transformation matrix for this sequence is $T_{XYZ}\, R_Z\, R_Y\, R_X$.

Enter each blade parameter in a row vector with as many columns as unique blades.

$E := 200000$ Elastic modulus

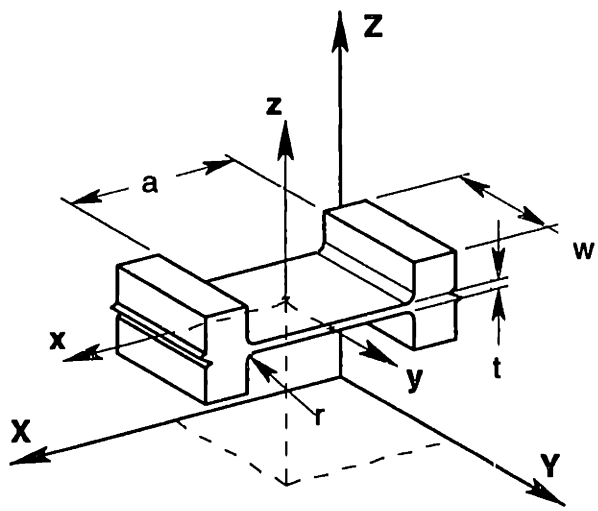
$\nu := 0.3$ Poisson ratio

$a := 20$

$w := 40$

$t := 0.5$

$k_{act} := 0$



Define below the transformation matrix T(p) for the particular sequence of moves to be used. For each local x-y-z CS, enter values of moves that define it in the corresponding column of the position matrix P below. Include any parameters in the position matrix as desired to aid in changing the configuration.

$T(p) := R_z(p_6) \cdot T_{xyz}(p_1, p_2, p_3) \cdot R_z(p_5) \cdot R_x(p_4)$ $b := 100$

$$P := \begin{bmatrix} 0 & 0.5\cdot a & 0 & 0 & 0.5\cdot a & 0 & 0 & 0.5\cdot a & 0 \\ b & b+\frac{a}{2} & b & b & b+\frac{a}{2} & b & b & b+\frac{a}{2} & b \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 90\cdot deg & 90\cdot deg & 0 & 90\cdot deg & 90\cdot deg & 0 & 90\cdot deg & 90\cdot deg & 0 \\ 0 & 90\cdot deg & 0 & 0 & 90\cdot deg & 0 & 0 & 90\cdot deg & 0 \\ 240\cdot deg & 240\cdot deg & 240\cdot deg & 0 & 0 & 0 & 120\cdot deg & 120\cdot deg & 120\cdot deg \end{bmatrix}$$

Expressed in the base CS, the rotation matrix transforms either a force-moment vector or a differential displacement-rotation vector in the new CS to the base CS. Being orthonormal, the transposed rotation matrix gives the inverse transformation. The translation matrix is not orthonormal and transforms only a force-moment vector in the new CS to the base CS. The transposed translation matrix transforms only a differential displacement-rotation vector in the base CS to the new CS. The same applies to a general [6 x 6] transformation matrix.

$$R_x(x) \equiv \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \cos(x) & -\sin(x) & 0 & 0 & 0 \\ 0 & \sin(x) & \cos(x) & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cos(x) & -\sin(x) \\ 0 & 0 & 0 & 0 & \sin(x) & \cos(x) \end{bmatrix}$$

[6 x 6] rotation matrix for rotation about the x-axis.

$$R_y(y) \equiv \begin{bmatrix} \cos(y) & 0 & \sin(y) & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ -\sin(y) & 0 & \cos(y) & 0 & 0 & 0 \\ 0 & 0 & 0 & \cos(y) & 0 & \sin(y) \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -\sin(y) & 0 & \cos(y) \end{bmatrix}$$

[6 x 6] rotation matrix for rotation about the y-axis.

$$R_z(z) \equiv \begin{bmatrix} \cos(z) & -\sin(z) & 0 & 0 & 0 & 0 \\ \sin(z) & \cos(z) & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cos(z) & -\sin(z) & 0 \\ 0 & 0 & 0 & \sin(z) & \cos(z) & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

[6 x 6] rotation matrix for rotation about the z-axis.

$$T_{xyz}(x, y, z) \equiv \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -z & y & 1 & 0 & 0 \\ z & 0 & -x & 0 & 1 & 0 \\ -y & x & 0 & 0 & 0 & 1 \end{bmatrix}$$

[6 x 6] translation matrix for translation in x, y and z.

Define the diagonal elements of the blade's compliance matrix in the local CS.

$$C_{1,1} := \frac{a}{E \cdot t \cdot w} \qquad C_{2,2} := C_{1,1} \cdot \left[\left(\frac{a}{w} \right)^2 + 2.4 \cdot (1 + \nu) \right] \qquad C_{6,6} := C_{1,1} \cdot \frac{12}{w^2}$$

$$C_{3,3} := C_{1,1} \cdot \left[(1 - \nu^2) \cdot \left(\frac{a}{t} \right)^2 + 2.4 \cdot (1 + \nu) \right] \qquad C_{5,5} := C_{1,1} \cdot (1 - \nu^2) \cdot \frac{12}{t^2}$$

$$C_{4,4} := 12 \cdot \left[\frac{1}{2 \cdot (1 + \nu)} \cdot \left(4 + 2.52 \cdot \frac{t}{w} \right) \cdot \frac{t^2}{C_{1,1}} + \frac{w^2}{C_{3,3}} \right]^{-1}$$

$$C = \begin{bmatrix} 5 \cdot 10^{-6} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.69 \cdot 10^{-5} & 0 & 0 & 0 & 0 \\ 0 & 0 & 7.3 \cdot 10^{-3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 4.04 \cdot 10^{-5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.18 \cdot 10^{-4} & 0 \\ 0 & 0 & 0 & 0 & 0 & 3.75 \cdot 10^{-8} \end{bmatrix}$$

Define the diagonal elements of the blade's stress matrix in the local CS.

$$S_{1,1} := \frac{1}{t \cdot w} \qquad S_{2,2} := S_{1,1} \cdot \left[\frac{3 \cdot a}{2 \cdot w} + \sqrt{\left(\frac{3 \cdot a}{2 \cdot w} \right)^2 + 1} \right] \qquad S_{6,6} := S_{1,1} \cdot \frac{6}{w}$$

$$S_{3,3} := S_{1,1} \cdot \frac{3 \cdot a}{t} \qquad S_{5,5} := S_{1,1} \cdot \frac{6}{t} \qquad S_{4,4} := S_{3,3} \cdot \frac{6}{w}$$

$$S = \begin{bmatrix} 0.05 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7.5 \cdot 10^{-3} \end{bmatrix}$$

Reflect the compliance and stiffness matrices from the local CS j to the base CS.

$$K_b(j) := T\left(P^{<j>}\right) \cdot C^{-1} \cdot T\left(P^{<j>}\right)^T \qquad C_b(j) := K_b(j)^{-1}$$

$$K_a(j) := T\left(P^{<j>}\right)^{<1>} \cdot k_{act} \cdot T\left(P^{<j>}\right)^{<1>T} \qquad C_a(j) := K_a(j)^{-1}$$

Combine parallel and series combinations.

$$K_{2\_3} := K_b(2) + K_a(3) \qquad C_{2\_3} := K_{2\_3}^{-1}$$

$$K_{5\_6} := K_b(5) + K_a(6) \qquad C_{5\_6} := K_{5\_6}^{-1}$$

$$K_{8\_9} := K_b(8) + K_a(9) \qquad C_{8\_9} := K_{8\_9}^{-1}$$

$$C_{1\_2\_3} := C_b(1) + C_{2\_3} \qquad K_{1\_2\_3} := C_{1\_2\_3}^{-1}$$

$$C_{4\_5\_6} := C_b(4) + C_{5\_6} \qquad K_{4\_5\_6} := C_{4\_5\_6}^{-1}$$

$$C_{7\_8\_9} := C_b(7) + C_{8\_9} \qquad K_{7\_8\_9} := C_{7\_8\_9}^{-1}$$

$$K_0 := K_{1\_2\_3} + K_{4\_5\_6} + K_{7\_8\_9} \qquad C_0 := K_0^{-1}$$

$$C_0 = \begin{bmatrix} 3.89 \cdot 10^{-3} & 1.78 \cdot 10^{-15} & 0 & 0 & 0 & 0 \\ 0 & 3.89 \cdot 10^{-3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.37 \cdot 10^{-5} & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.72\ 10^{-9} & 0 & 0 \\ 0 & 0 & 0 & 0 & 2.72 \cdot 10^{-9} & 0 \\ 0 & 0 & 0 & 0 & 0 & 3.73\ 10^{-7} \end{bmatrix}$$

$$K_0 = \begin{bmatrix} 256.94 & -1.18\ 10^{-10} & 0 & 1.53\ 10^{-9} & 1.25\ 10^{-9} & 4.66\ 10^{-9} \\ -5.48 \cdot 10^{-11} & 256.94 & 0 & -1.25 \cdot 10^{-9} & 1.53 \cdot 10^{-9} & 3.86 \cdot 10^{-8} \\ 0 & 0 & 7.28\ 10^{4} & -7.73 \cdot 10^{-7} & -6.01\ 10^{-6} & -3.06 \cdot 10^{-9} \\ 1.53 \cdot 10^{-9} & -1.25 \cdot 10^{-9} & -7.72 \cdot 10^{-7} & 3.68 \cdot 10^{8} & 3 \cdot 10^{-4} & 0 \\ 1.25\ 10^{-9} & 1.53\ 10^{-9} & -6.01\ 10^{-6} & 3\ 10^{-4} & 3.68\ 10^{8} & 0 \\ -3.48 \cdot 10^{-9} & 3.56 \cdot 10^{-8} & -3.06 \cdot 10^{-9} & 0 & 0 & 2.68 \cdot 10^{6} \end{bmatrix}$$

6.4.2 Kinematic Coupling Analysis Program

This program helps the user develop a parameterized model of a kinematic coupling so that its ability to reach a seated position may be optimized. The coupling reaches a seated position when all six constraint surfaces are contacting. It approaches the seated position from one of several directions determined by the surfaces already in contact. The limiting case will always occur with five surfaces in contact, and there are only six such directions. The program determines the six directions then solves for the corresponding coefficients of friction that just impede sliding. The smallest coefficient of friction among the six is the limiting coefficient of friction. The program varies parameters between limits set by the user and plots the limiting coefficient of friction. This visual feedback helps the user find the global optimum for this highly coupled, nonlinear problem. In addition, the program calculates the following items for the optimized coupling assuming zero friction: stiffness and compliance matrices, coupling displacement, and contact forces. Due to lengthy calculations, it is advisable to set automatic calculation off.

This program uses six dimensional vectors and [6 x 6] matrices to deal simultaneously with both linear and angular measures. The first three components correspond to linear measures such as force and displacement, and the last three correspond to angular measures such as moment and rotation. The same arrangement applies to the rows and columns of the matrices used for coordinate transformation, stiffness and compliance. In addition, the positions and orientations for six constraint surfaces are conveniently stored in a matrix with six columns. Each column describes a sequence of moves required to go from the global X-Y-Z CS (coordinate system) to a local x-y-z CS as shown in the figure. This matrix also contains the parameters that the program varies in the optimization process. Forces and displacements for the kinematic coupling are represented in the global CS while forces and displacements at a particular constraint are represented in its local x-y-z CS. The z direction represents the outward facing normal of the fixed surface so that a negative displacement is movement of the coupling into contact.

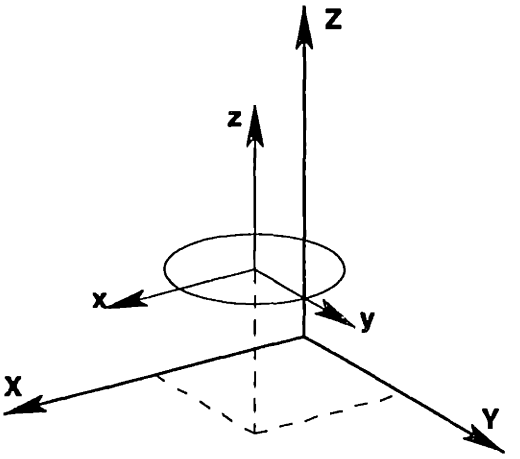
The transformations between the X-Y-Z CS and x-y-z CS's are done as a sequence of moves. The number and order may vary to suit the application, but they must be the same for each constraint since they are stored in a matrix. The basic moves are: translation in x-y-z, rotation about x, rotation about y and rotation about z. The sequence may be interpreted as moves relative to the base CS or in the opposite order as moves relative to the local CS. Two rules express these interpretations for a sequence of transformation matrices: 1) post multiply to transform in the local CS and 2) pre multiply to transform in the base CS. A typical sequence relative to the base CS is: rotate about the X axis, rotate about the Y axis, rotate about the Z axis, and translate (X, Y, Z). The equivalent sequence relative to the local CS is: translate (x, y, z) (initially coincident with the base CS), rotate about the z axis, rotate about the y axis, and rotate about the x axis. The transformation matrix for this sequence is $T_{XYZ}\, R_Z\, R_Y\, R_X$.

Define below the transformation matrix T(p) for the particular sequence of moves to be used. For each local x-y-z CS, enter values of moves that define it in the corresponding column of the position matrix P(θ) below. Include components of the parameter vector θ in the position matrix to allow optimization of the coupling.

$$T(p) := T_{xyz}(p_1, p_2, p_3) \cdot R_z(p_6) \cdot R_y(p_5) \cdot R_x(p_4)$$

$$P(\theta) := \begin{bmatrix} -0.385 & -0.385 & 0.385 & 0.385 & -0.397 & 0.397 \\ -0.315 & -0.315 & -0.315 & -0.315 & -0.241 & -0.241 \\ -0.8 & -0.8 & -0.8 & -0.8 & 0.85 & 1.31 \\ \theta_2 & -\theta_3 & \theta_2 & -\theta_3 & 90 \cdot deg & 90 \cdot deg \\ -\theta_1 & -\theta_1 & \theta_1 & \theta_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\theta_4 & \theta_4 \end{bmatrix}$$

Enter a vector k of contact stiffnesses and a vector f of forces and moments applied to the coupling at the origin of X-Y-Z. Note, a transformation matrix may be used to apply the force elsewhere.



$$k := \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad f := \begin{bmatrix} 0 \\ 0 \\ -450 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Expressed in the base CS, the rotation matrix transforms either a force-moment vector or a differential displacement-rotation vector in the new CS to the base CS. Being orthonormal, the transposed rotation matrix gives the inverse transformation. The translation matrix is not orthonormal and transforms only a force-moment vector in the new CS to the base CS. The transposed translation matrix transforms only a differential displacement-rotation vector in the base CS to the new CS. The same applies to a general [6 x 6] transformation matrix.

$$R_x(x) \equiv \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \cos(x) & -\sin(x) & 0 & 0 & 0 \\ 0 & \sin(x) & \cos(x) & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cos(x) & -\sin(x) \\ 0 & 0 & 0 & 0 & \sin(x) & \cos(x) \end{bmatrix}$$

[6 x 6] rotation matrix for rotation about the x-axis.

$$R_y(y) \equiv \begin{bmatrix} \cos(y) & 0 & \sin(y) & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ -\sin(y) & 0 & \cos(y) & 0 & 0 & 0 \\ 0 & 0 & 0 & \cos(y) & 0 & \sin(y) \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -\sin(y) & 0 & \cos(y) \end{bmatrix}$$

[6 x 6] rotation matrix for rotation about the y-axis.

$$R_z(z) \equiv \begin{bmatrix} \cos(z) & -\sin(z) & 0 & 0 & 0 & 0 \\ \sin(z) & \cos(z) & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \cos(z) & -\sin(z) & 0 \\ 0 & 0 & 0 & \sin(z) & \cos(z) & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

[6 x 6] rotation matrix for rotation about the z-axis.

$$T_{xyz}(x, y, z) \equiv \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -z & y & 1 & 0 & 0 \\ z & 0 & -x & 0 & 1 & 0 \\ -y & x & 0 & 0 & 0 & 1 \end{bmatrix}$$

[6 x 6] translation matrix for translation in x, y and z.

A coupling with only five surfaces in contact is free to slide in one direction. In this direction, the stiffness matrix is singular, assuming zero friction. The eigenvector of the stiffness matrix corresponding to the zero eigenvalue provides the direction of sliding, although it may have the wrong sign. The stiffness matrix K for the coupling is a summation of contact stiffnesses transformed to the global CS. The definition of K below excludes the jth contact surface from the sum. Use j = 0 to calculate K with all surfaces in contact. The matrix Δ contains the six sliding directions in its columns, all relative to the global CS.

$$K(P, j) := \sum_{i=1}^{6} T\left(P^{\langle i \rangle}\right)^{\langle 3 \rangle} \cdot k_i \cdot (i \neq j) \cdot T\left(P^{\langle i \rangle}\right)^{\langle 3 \rangle^T}$$

$$\Delta(P) := \left| \begin{array}{l} \text{for } i \in 1..6 \\ \quad \Delta^{\langle i \rangle} \leftarrow \text{eigenvec}(K(P, i), 0) \\ \Delta \end{array} \right.$$

The coefficient of friction that just impedes sliding causes the coupling to be in static equilibrium. Obviously the non contacting surface has zero normal force. Therefore, static equilibrium equations can be solved to determine the coefficient of friction that makes the normal force zero at the surface intended to be noncontacting. This first requires a relationship for the friction force. Given a sliding direction δ and a coefficient of friction μ, the vector ϕ is a normalized force-moment vector that includes tangential friction. Both δ and ϕ are defined in the local x-y-z CS. Transformed to the global CS, ϕ's for all six contact surfaces are assembled in matrix A. Matrix A multiplied by a six-component vector of surface normal forces equals the resultant force-moment vector for the coupling.

$$\phi(\delta, \mu) := \left[\frac{-\mu \cdot \delta_1}{\left| \begin{bmatrix} \delta_1 & \delta_2 & \delta_3 \end{bmatrix}^T \right|} \quad \frac{-\mu \cdot \delta_2}{\left| \begin{bmatrix} \delta_1 & \delta_2 & \delta_3 \end{bmatrix}^T \right|} \quad 1 \quad 0 \quad 0 \quad 0 \right]^T$$

$$A(P,\delta,\mu) := \left| \begin{array}{l} \text{for } i \in 1..6 \\ \quad A^{\langle i \rangle} \leftarrow T\left(P^{\langle i \rangle}\right) \cdot \phi\left(T\left(P^{\langle i \rangle}\right)^{T} \cdot \delta, \mu\right) \\ A \end{array} \right.$$

Since matrix A is invertible, the surface normal forces are linearly solved from the applied force-moment vector f. The only one of interest is the surface intended to be non contacting, which is zero at equilibrium. The root of this nonlinear equation provides the coefficient of friction that just impedes sliding. The coefficients of friction for all six sliding directions are assembled in vector μ.

$\mu := 0$ Starting value for the solver.

$$\mu(P,\Delta) := \left| \begin{array}{l} \text{for } i \in 1..6 \\ \quad \mu_i \leftarrow \left| \text{root}\left(\text{lsolve}\left(A\left(P, \Delta^{\langle i \rangle}, \mu\right), f\right)_i, \mu\right) \right| \\ \mu \end{array} \right.$$

The program plots curves of the limiting coefficient of friction over ranges of parameters that vary individually. Each curve is a function of one parameter while the others are held at nominal values. The user will adjust the nominal set towards the optimal set. Enter minimum and maximum values for the parameter ranges to plot. Enter the nominal parameter set. Enter the number of points to plot over the range (typically $n_r := 5$). Enter the number of parameters (typically $n_p := \text{rows}(\theta)$ but may be fewer). Start the calculation (command =) if in manual mode.

$$\theta_{min} := \begin{bmatrix} 35 \\ 15 \\ 30 \\ 45 \end{bmatrix} \cdot \text{deg} \qquad \theta_{max} := \begin{bmatrix} 55 \\ 35 \\ 50 \\ 65 \end{bmatrix} \cdot \text{deg} \qquad \theta := \begin{bmatrix} 40 \\ 25 \\ 45 \\ 60 \end{bmatrix} \cdot \text{deg}$$

$$\frac{\theta - \theta_{min}}{\theta_{max} - \theta_{min}} \rightarrow \begin{bmatrix} 0.25 \\ 0.5 \\ 0.75 \\ 0.75 \end{bmatrix}$$

$n_r := 5$ $\quad i := 1..n_r$ $\quad P_\theta := P(\theta)$

$n_p := 4$ $\quad j := 1..n_p$ $\quad \text{rank}\left(K\left(P_\theta, 0\right)\right) = 6$

$$\alpha_i := \frac{i-1}{n_r - 1}$$

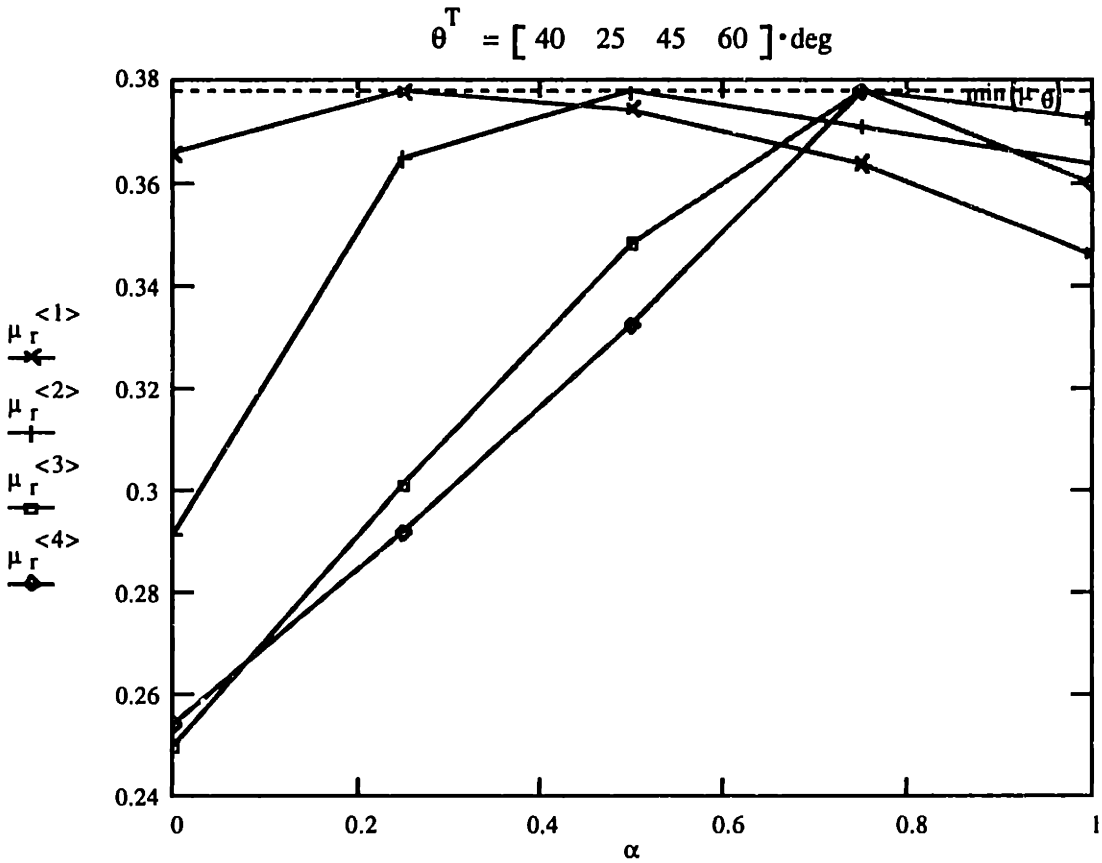
$$\theta_r(i,j) := \left| \begin{array}{l} \theta_r \leftarrow \theta \\ \theta_{r_j} \leftarrow \left(1 - \alpha_i\right) \cdot \theta_{min_j} + \alpha_i \cdot \theta_{max_j} \\ \theta_r \end{array} \right.$$

$$\mu_\theta := \mu\left(P_\theta, \Delta\left(P_\theta\right)\right)$$

$$\mu_\theta = \begin{bmatrix} 0.391 \\ 0.445 \\ 0.53 \\ 0.441 \\ 0.378 \\ 0.387 \end{bmatrix}$$

$$\mu_{r_{i,j}} := \min\left(\mu\left(P\left(\theta_r(i,j)\right), \Delta\left(P\left(\theta_r(i,j)\right)\right)\right)\right)$$

If necessary, change the number of curves to match the number of parameters. Since the parameters may vary over different ranges, it is convenient to plot the curves over a normalized range between 0 and 1. The horizontal dashed line is the limiting coefficient of friction for the nominal parameter set. It crosses each curve at the nominal parameter value. The goal is to adjust the nominal set so that the horizontal line crosses near the peaks of the curves. Another indication of being near the optimum is a good balance among components in μ_θ, the coefficients of friction for the six sliding directions.



Display the optimized position matrix of constraints $P(\theta)$, the sliding directions $\Delta(P)$ and the corresponding coefficients of friction μ_θ. Each column corresponds to a particular constraint.

$$P_\theta = \begin{bmatrix} -0.385 & -0.385 & 0.385 & 0.385 & -0.397 & 0.397 \\ -0.315 & -0.315 & -0.315 & -0.315 & -0.241 & -0.241 \\ -0.8 & -0.8 & -0.8 & -0.8 & 0.85 & 1.31 \\ 0.436 & -0.785 & 0.436 & -0.785 & 1.571 & 1.571 \\ -0.698 & -0.698 & 0.698 & 0.698 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1.047 & 1.047 \end{bmatrix}$$

$$\Delta(P_\theta) = \begin{bmatrix} -0.322 & 7.232\cdot10^{-3} & -0.341 & -0.034 & 0.392 & 0.378 \\ -0.121 & 0.2 & 0.222 & 0.203 & 0.525 & -0.531 \\ 0.325 & 0.276 & -0.361 & 0.275 & -0.207 & 0.209 \\ -0.229 & 0.193 & 0.097 & 0.189 & -0.656 & 0.664 \\ 0.318 & -0.02 & 0.295 & -1.307\cdot10^{-3} & 0.311 & 0.3 \\ 0.789 & -0.92 & 0.779 & 0.92 & 0 & 0 \end{bmatrix}$$

$$\mu_\theta^{T} = [\,0.391 \quad 0.445 \quad 0.53 \quad 0.441 \quad 0.378 \quad 0.387\,]$$

Display a vector of surface normal forces, the couplings displacement, and the stiffness and compliance matrices.

$f_{norm} := \text{lsolve}(A(P_\theta, 0, 0), f)$ $\quad K_\theta := K(P_\theta, 0)$ $\quad C_\theta := K_\theta^{-1}$

$$f_{norm}^{T} = [-180.844 \quad -169.289 \quad -203.765 \quad -168.51 \quad -83.822 \quad -68.812] \qquad C_\theta \cdot f = \begin{bmatrix} 1.809 \\ 57.505 \\ -317.867 \\ -94.622 \\ 17.728 \\ 23.375 \end{bmatrix}$$

$$K_\theta = \begin{bmatrix} 2.592 & 0 & 0 & 0.199 & 0.245 & 0.42 \\ 0 & 1.857 & 0.179 & 0.489 & -0.199 & 0 \\ 0 & 0.179 & 1.551 & -0.345 & 0 & 0 \\ 0.199 & 0.489 & -0.345 & 1.542 & 0.43 & 2.349\cdot10^{-3} \\ 0.245 & -0.199 & 0 & 0.43 & 3.559 & -0.487 \\ 0.42 & 0 & 0 & 2.349\cdot10^{-3} & -0.487 & 0.346 \end{bmatrix}$$

$$C_\theta = \begin{bmatrix} 0.544 & -0.01 & -4.02\cdot10^{-3} & -0.023 & -0.155 & -0.878 \\ -0.01 & 0.625 & -0.128 & -0.249 & 0.084 & 0.132 \\ -4.02\cdot10^{-3} & -0.128 & 0.706 & 0.21 & -0.039 & -0.052 \\ -0.023 & -0.249 & 0.21 & 0.815 & -0.133 & -0.165 \\ -0.155 & 0.084 & -0.039 & -0.133 & 0.419 & 0.778 \\ -0.878 & 0.132 & -0.052 & -0.165 & 0.778 & 5.048 \end{bmatrix}$$

7 Examples of Exact-Constraint Designs

This chapter presents particular exact-constraint designs that are being used for the NIF and EUVL projects. These designs use both flexural elements and contacting surfaces as constraint devices. All the designs have been thoroughly analyzed using various techniques, but this chapter is about design rather than analysis. The intent is to present the thinking behind the designs. Where it is of interest, analytical or experimental results will be mentioned. Most of the figures in this chapter are photographs of hardware. All efforts have been made to ensure good quality, but photographs sometimes do not reproduce well.

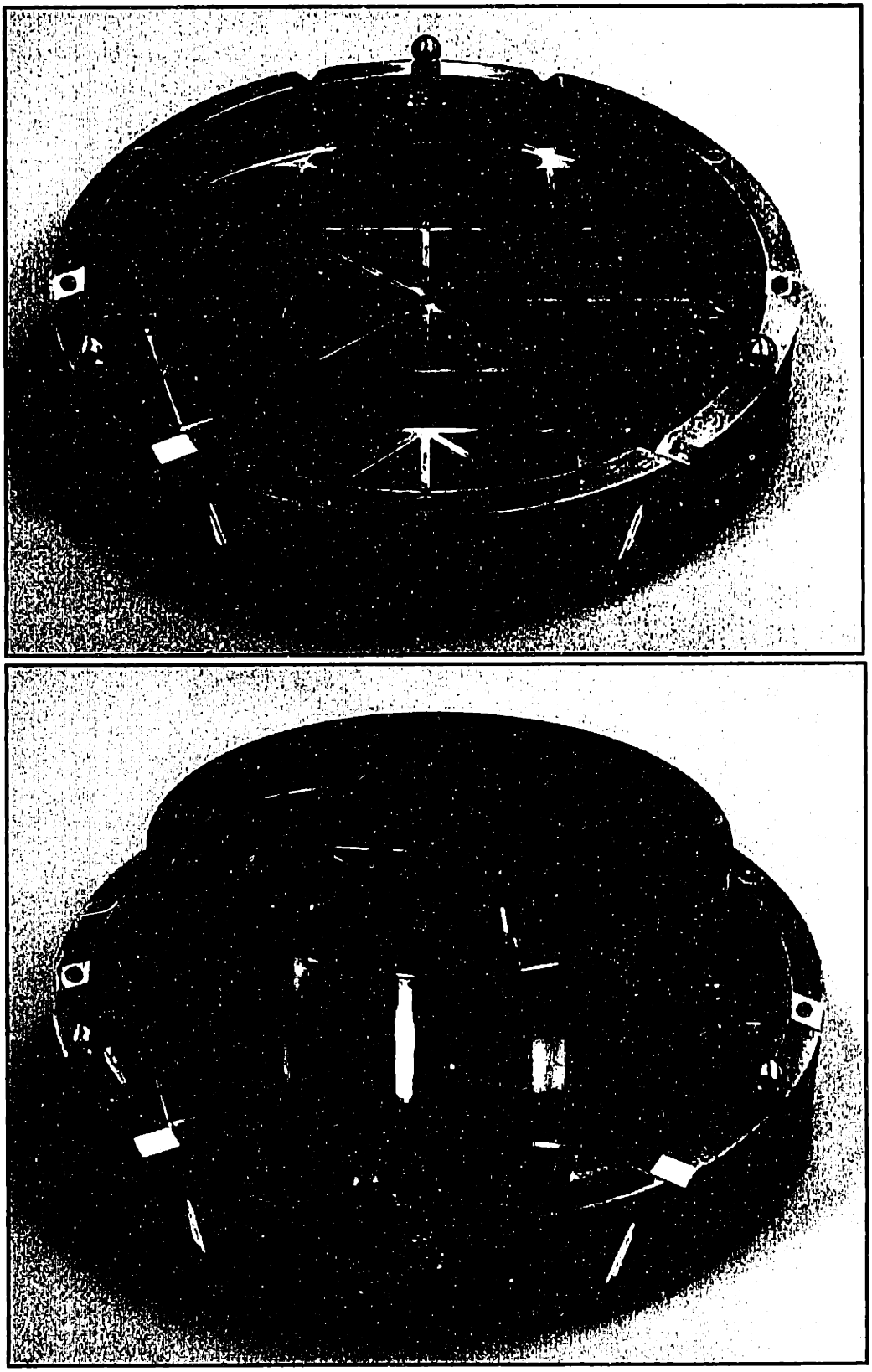
7.1 Optic Mounts for EUVL Projection Optics

The key requirements for EUVL projection-optic mounts are as follows:

- Provide rigid and stable location of the optic with respect to the main structural component, called the projection optics box.
- Apply negligible forces and moments to the optic other than those required to support the weight of the optic.
- Provide a repeatable connect-disconnect capability so that the optic can be processed separately from the mount and returned to a repeatable location and state of strain.

Each projection optic in the system has an additional requirement that is satisfied at a higher system level by a unique design, but all those designs incorporate the same basic mount design described here. Later sections describe those higher-level requirements and design solutions. These requirements common to all the optics are satisfied by a structural element called an optic cell and an ideally kinematic coupling between it and the optic.

The desire for symmetry and the minimum number of physical connections to the optic led to a general configuration using three pairs of constraints. The sphere-vee constraint is conceptually simple but high stiffness in the unconstrained directions (i.e., before sliding occurs) is a concern due to differences in thermal expansion between the optic and the cell. Another concern is the contact stress that would result on the low-thermal-expansion materials being considered such as super invar and Zerodur™. The configuration that appears in Figure 6-6 (a), three sets of sphere-cone constraints each with a radial-motion flexure, solves the contact stress problem, but the sphere-cone constraint is still stiff in rotational degrees of freedom. For example, a portion of the loads applied to the cell transfers through the optic mount and into the optic. This problem is greatly reduced by adding three more flexural degrees of freedom. The result is the bipod flexure, a two-constraint device. A set of three bipods can provide nearly ideal kinematic support for super-precision optics and other sensitive instruments. Figure 7-1 shows a typical mount design with and without the optic installed.



(a)

(b)

Figure 7-1 The optic cell and three bipod flexures (a) present three small kinematic couplings that engage three lugs bonded to the optic (b). A coil spring held with a shoulder screw provides the preload for each coupling. The lugs are boned to the optic while attached to relaxed bipod flexures, and subsequently the assembly becomes matched for life. The individual three-tooth couplings are repeatable to the micron level.

While the decision to use three bipod flexures was fairly obvious and unconstrained, the opposite was true for the type of connection to make between each bipod and the optic. One design constraint was the need to order the optic substrate material before there was time to design the mounts. This and the desire to keep the optics as simple as possible led to the decision to epoxy bond mounting features to the semifinished optic that would arrive months later. The other primary decision was the type of connect-disconnect device to use at each bipod. The sphere-cone constraint is simple and effective, but it comes with the risk that significant *noise* moments can exist in the bipods depending how the optic is installed in the mount. A significant noise moment for these optics is a few newton-millimeters. Ultimately we chose to use the three-tooth coupling, a fully constraining connection. The positional repeatability of the coupling ensures very good elastic repeatability of the flexures. Further, the technique used to bond lugs to the optic puts the bipod flexures very nearly in a relaxed stress state.

The decisions made for this projection optics system, being a fast-track experimental tool, are not necessarily appropriate for future production tools. For example, an issue with the epoxy bond is long-term dimensional stability. Measurements of loaded samples extrapolated in time indicate that the system may drift out of optical alignment in perhaps 6 to 12 months (thus requiring an off-line realignment of optics).[1] This is obviously not acceptable for a production tool and the positional requirements probably will become more stringent in the future. It becomes apparent that future designs will require direct connections between the optic and the mount. An epoxy bond could be used as a fastener, for example, in conjunction with a compliant element to apply a preload, but the interface that determines stability should not be a polymer.

Given this insight, the sphere-cone constraint becomes a favorite because three cones are easily manufactured directly into the optic. The cell would present three small spheres mounted on bipod flexures, for example. There are techniques available to make the optic less sensitive to noise moments at the constraints, and the installation process can be made more deterministic to reduce noise moments. In addition, there is no particular limitation with this design from being able to process the optic in one cell and use it in another. This would be very difficult to do with fully constraining couplings. Other aspects may also come into play such as combining actuated alignment mechanisms within the optic mount. Presently the cells are manipulated by alignment mechanisms, but it is difficult to achieve the very high resonance frequencies desired for such systems with a series arrangement of constraints.

---

[1] Experiments were performed at the University of North Carolina Charlotte [Patterson, et al., 1998] and LLNL.

7.2 A Gravity-Compensating Optic Mount for EUVL

Three of the projection optics are manufactured using vertical-axis fabrication metrology to match the orientation of the projection optics system. These optics will have the proper figure when mounted rather than in the free state, assuming of course that the mount is absolutely repeatable. One optic in the system is an exception because its large radius of curvature makes a vertical-axis interferometer impractical. With the optic supported in a sling, horizontal fabrication metrology produces an optic with the proper figure nearer the free state, especially when using multi-step averaging. It then becomes necessary to compensate for the different mount and the change in gravity orientation.

A finite element model of the optic supported at three perimeter points, as with three bipods, revealed unacceptable gravity-induced figure error. Figure 7-2 (a) shows the figure error within the clear aperture after removing the best-fit sphere.[1] The dominant error is trifoil produced from the three supports. By distributing the weight over nine perimeter points, the error becomes primarily spherical, which the system will tolerate with a focus adjustment. The remaining nonspherical error in (b) is more than an order of magnitude less than in (a). For comparison, the P-V error goes from 1.62 nm to 0.107 nm and the rms error goes from 0.314 nm to 0.027 nm.

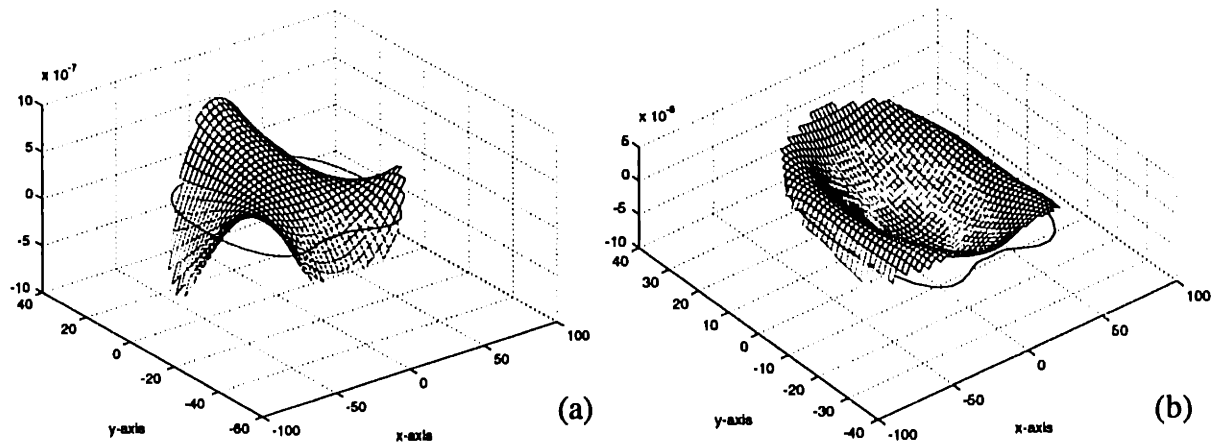
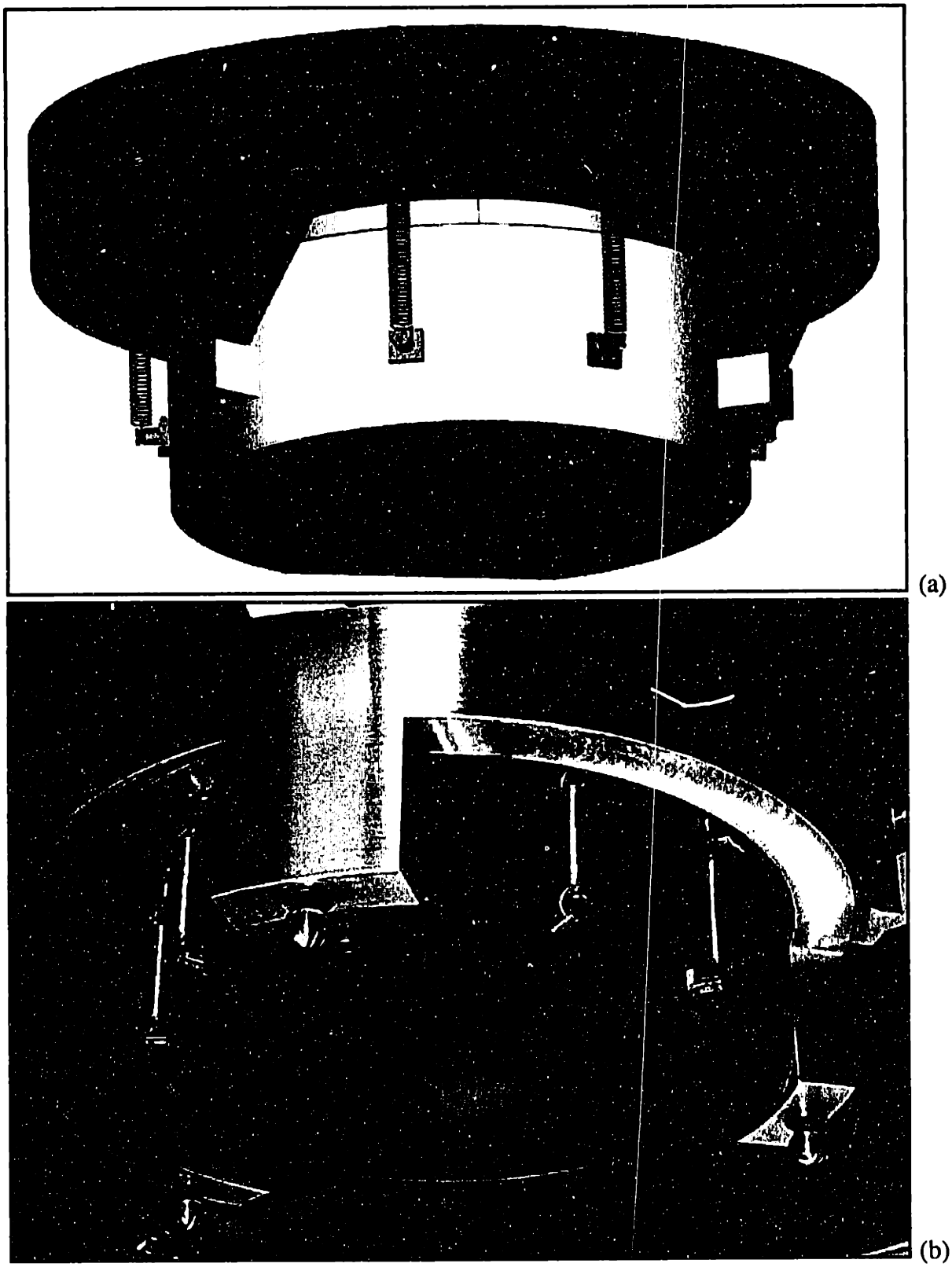


Figure 7-2 Figure error (in mm) over the clear aperture with three supports (a) and nine supports (b).

The gravity-compensating optic mount supports the weight of the optic at nine locations around the perimeter. Three of the nine supports are rigid bipod constraints as used in the other three optic mounts. The remaining six supports are compliant springs that provide weight relief but no constraint. The tension is set according to modeling results that minimize the nonspherical error over the clear aperture. Figure 7-3 shows the design for the projection optics system in (a) and a demonstration test optic and mount in (b).

---

[1] Pro/MECHANICA™ by Parametric Technology Corp. is the finite element software used in this study. MATLAB™ by The MathWorks, Inc. is the computing environment used to process the finite-element results and create the surface plots.



(a)

(b)

Figure 7-3 The gravity-compensating optic mount supports the weight of the optic at nine locations around the perimeter using three rigid bipod constraints and six compliant springs. The model in (a) is the design for the projection optics system. A demonstration test optic and mount appear in (b).

7.3 θ_x-θ_y-Z Flexure Stage for EUVL Projection Optics[1]

Two of the projection optics require remotely actuated alignment degrees of freedom in θ_x, θ_y and z, where z is the optical axis. This out-of-plane motion may be accomplished by translating any three perimeter points appropriately in the z direction. Effectively, these three translations are adjustable constraints. Three more passive constraints provide exact constraint for the suspended object, in this case for the optic cell. Although a 3-2-1 constraint arrangement is possible, the natural choice is three identical pairs of constraints, where one direction is actuated and the other is passive. This is the arrangement shown in Figure 7-4 for a prototypical optic, cell and actuation system.

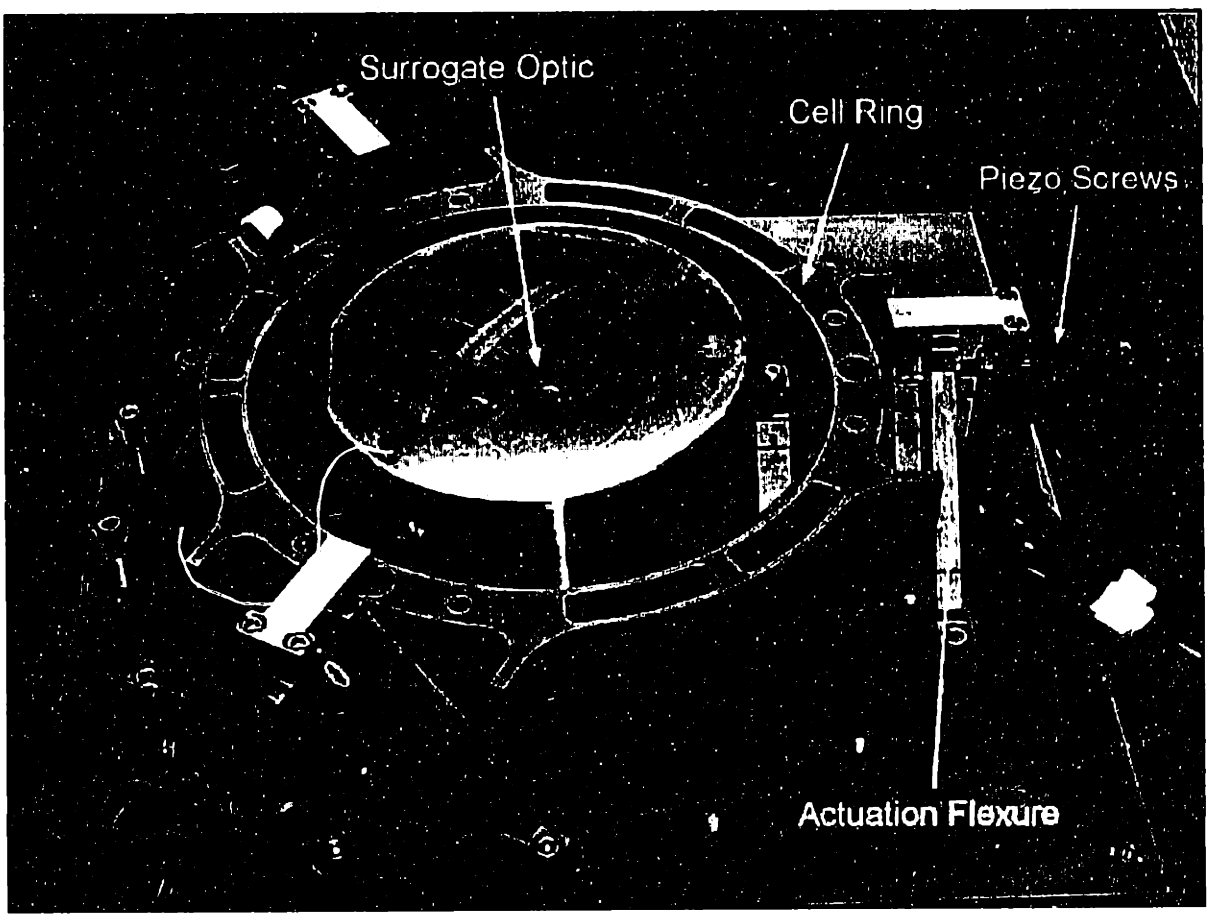


Figure 7-4 Three sets of actuation flexures support the optic cell relative to a base plate. Each actuation flexure transfers radial motion from a piezo-driven screw to axial (z) motion of the cell through a 5:1 reduction ratio. The cell then supports the optic passively through three bipod flexures.

The actuation system consists of three actuation flexures that support the optic cell relative to the projection optics box and a set of three piezo-driven screws that actuate the flexures. The actuation flexure provides both passive constraint tangent to the cell and a 5:1 transmission ratio between the screw and z motion at the cell. The screw provides the

[1] This material was previously published in greater detail [Tajbakhsh, et al., 1998].

actuated constraint through the 5:1 reduction ratio, thereby providing better resolution to the cell. Figure 7-5 shows more clearly the functional parts of the actuation system and the flexure cuts that provide the necessary freedoms.

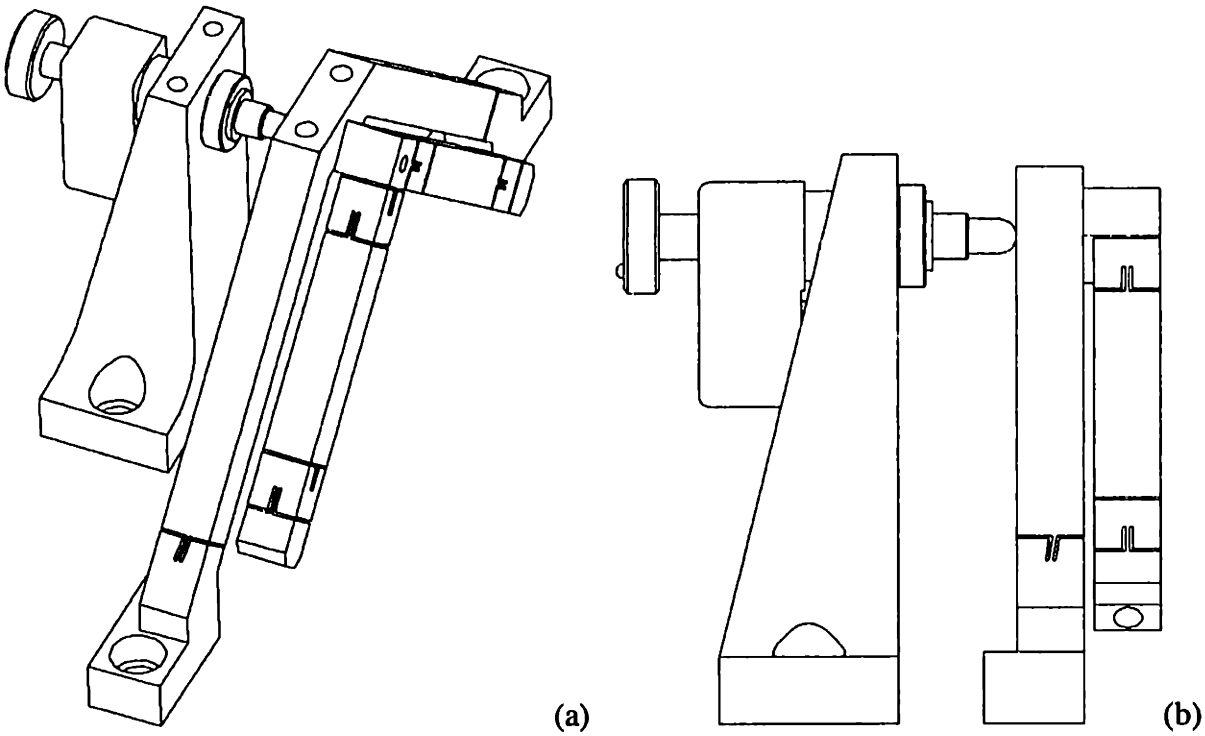


Figure 7-5 The actuation flexure consists of a supporting hinge flexure connected to a bipod. In (b), the 5:1 lever ratio is more apparent between horizontal motion of the screw and vertical motion of the cell.

The actuation flexure is physically one piece of super invar that has been cut out as shown using wire EDM (electro-discharge machining). Functionally, it consists of the constraint side that attaches to the cell and a transmission side that connects to the base structure. The constraint flexure is the same basic bipod design that is used for optic mounts. It provides local angular freedom to the cell while transferring two constraints from the transmission flexure. The instant center (where the two constraints intersect) is placed in the plane of the optic so that the rotation axis will also be in this plane. The transmission flexure is a simple hinge axis that is controlled by the screw position. A subtle aspect is the angle of the blade whose plane intersects the instant center of the constraint flexure. Effectively the forces in the bipod flexure transfer to the transmission flexure through this point. If the screw also acts through this point, then there is no out-of-plane force on the blades. The angle of the plane defined by the hinge axis and the instant center determines the transmission ratio between the screw and the cell.

There are two main design issues with this actuation flexure. The first is the moment imparted to the cell during actuation. This could be alleviated by placing the hinge point adjacent to the instant center, but this would require a cross flexure rather than a simple hinge. The finite element model shows no problems with this design because the

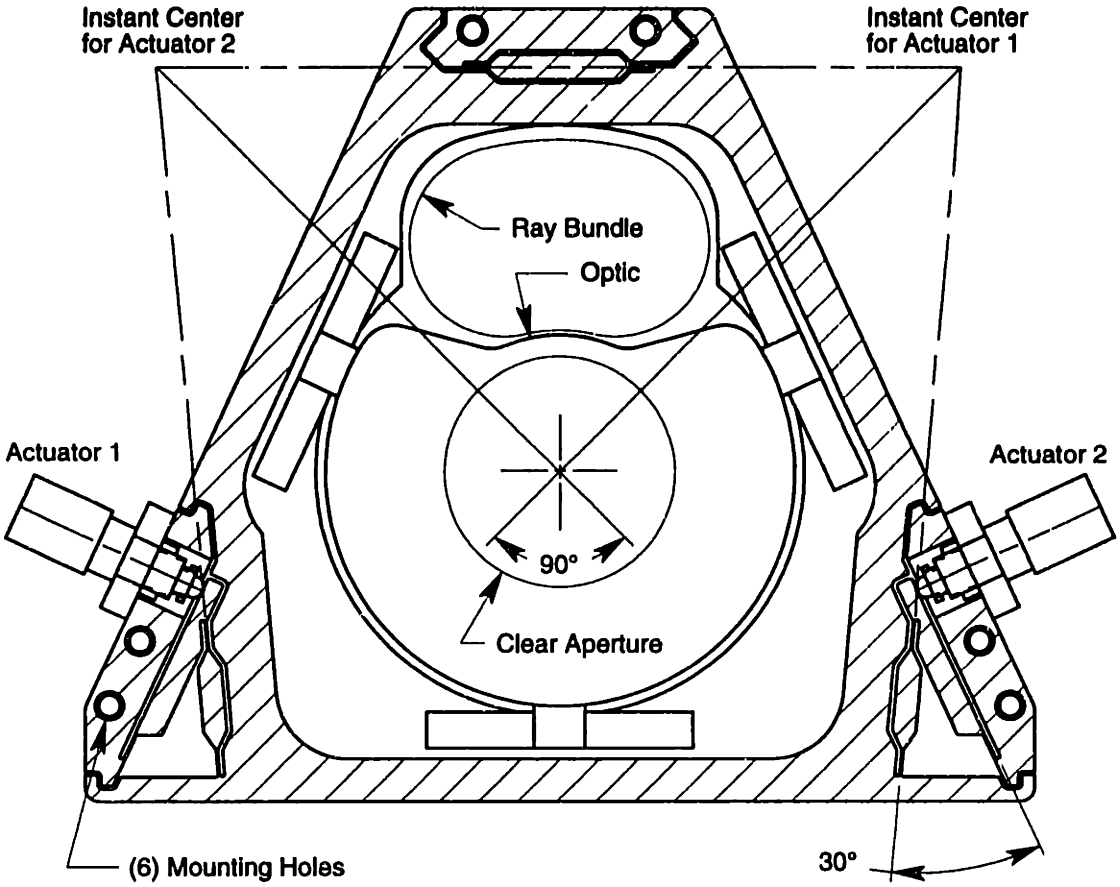
cell is sufficiently stiff. It would become a problem if the cell were eliminated in a future design with the optic directly supported on actuation flexures. The second issue is bending compliance in the transmission side since much of its structure lies outside the plane between the hinge axis and the instant center. This is the main compliance in the prototype system whose first constrained mode is 133 Hz. In an effort to push towards 200 Hz, the transmission side is now substantially thicker and joined across the legs.

It is interesting to state some of the experimental results of the prototype system. As mentioned the first constrained mode is 133 Hz, but it is remarkable and perhaps coincidental that FEA and experimental modal analysis gave the same number. Actually there are two modes at this frequency corresponding to any two orthogonal translations in the x-y plane. The remaining modal frequencies are 200 Hz and above. The dynamics of the optic with respect to the cell is rather insignificant. The positioning resolution demonstrated by this system is 2 to 3 nanometers as determined by capacitance gauge feedback. The screws are controlled in a low-bandwidth loop until the final position is reached, then they are turned off. All the main structural components are super invar except for the commercially available piezo-driven screws.

7.4 X-Y Flexure Stage for EUVL Projection Optics

One projection optic requires remotely actuated alignment degrees of freedom in x and y, where z is the optical axis. This optic is spherical and physically smaller than the other three aspheric optics. Being spherical, it then is acceptable for the optic to rotate as it translates. This allows the possibility of a rotational axis being used to translate the optic. Being smaller, it becomes practical to build the alignment mechanism directly into the optic cell. The X-Y-θ_z stage that uses three folded-hinge flexures is a natural starting point for this design (see Chapter 6.1.3). It merely requires one less actuated degree of freedom.

Figure 7-6 shows the design for the optic cell and integrated alignment mechanism. As with the other optics, three small bipod flexures support the optic relative to the cell. The cell has three areas that attach directly to the projection optics box as indicated by mounting holes. The rest of the cell articulates on the flexures. For motion in the x-y plane, each folded-hinge flexure provides one actuated constraint along the constraint side of the flexure (using the same terminology as the previous section). The single flexure at the top provides one passive constraint in the x-y plane. The constraint lines for these flexures (indicated by centerlines) form the instant centers about which the optic rotates. Extending actuator 1, for example, moves the optic down and to the right. If both actuators extend equally, then the optic translates downward without rotation. If both actuators extend equal amounts in opposite directions, then the optic rotates about the center of the passive constraint. The 30° angle between the transmission and constraint flexures provides a 2:1 reduction ratio between the actuator and the constraint line. There is another reduction ratio of approximately 1.4:1 resulting from rotation about either instant center.



(a)

(b)

Figure 7-6 In (a), a cross section through the optic cell shows the details of the flexures and the instant centers about which the optic rotates. In (b), the optic cell interfaces to the interferometer through a three-vee coupling. The black ears with tooling balls are removed from the cell after interferometry is complete.

The three out-of-plane degrees of freedom are overconstrained in this design by one extra constraint. Each folded-hinge flexure provides one z constraint while the single flexure provides two constraints, z and θ_y. The consequences of overconstraint will be some residual stress in the cell, flexures and the supporting structure. Although the surfaces that bolt together are precisely machined, they may require some hand fitting at assembly. This overconstraint will not affect the freedom of motion in the x-y plane.

The modal frequencies as predicted by FEA are quite high for this design. The first mode at 289 Hz results primarily from the flexibility of the folded-hinge flexures in their passive constraint direction. The next two modes are in the general direction of the actuated motions. Rotation about the passive constraint occurs at 354 Hz and translation (up and down) occurs at 436 Hz.

7.5 Kinematic Mounts for NIF Optics Assemblies

The basic kinematics and the optimization analysis for this example are treated in Chapter 6. This section presents the design solution and explains the main decisions made through the design process. The key requirements for the NIF kinematic mounts are as follows:

- Provide stiff, repeatable location in six degrees of freedom.
- Allow straight-line installation of the assembly from underneath.
- Provide the maximum amount of clearance and capture range within the space constraints of the closely packed laser beams.
- Provide secondary safety support and seismic restraint.
- Operate in accordance with class 100 clean room requirements.

The last two requirements are not discussed other than to say that a separate mechanism provides seismic restraint, effectively holding the kinematic mounts together, and that all the parts are corrosion resistant and easy to clean. Since the operation is infrequent and the mechanisms operate under essentially no load, particle generation is not expected to be problem. The first three requirements govern the kinematic mount design described here.

The architecture for the NIF laser system is heavily influenced by the need to routinely replace damaged optics in a very clean and inert atmosphere. A concept was developed where any optics assembly, known as a line replaceable unit or LRU, could be installed or removed from underneath the beam line of the laser.[1] The LRU being installed is transported in a clean canister from the clean assembly area to the beam line. The canister docks to the laser structure using a kinematic coupling, establishes a pressure-tight seal,

[1] As the design matured, some types of LRU's were easier to load from the top or the side, but most still load from below.

removes an access panel, and installs the LRU with a straight-line lift.[I] The interface between the LRU and the canister lift platform is also kinematic and preloaded by gravity.[II] Yet another kinematic coupling supports the LRU from the laser structure. This is the kinematic mount that must satisfy the requirements stated above. A prototype LRU for periscope optics appears in Figure 7-7 along with close-up views of the kinematic mounts.

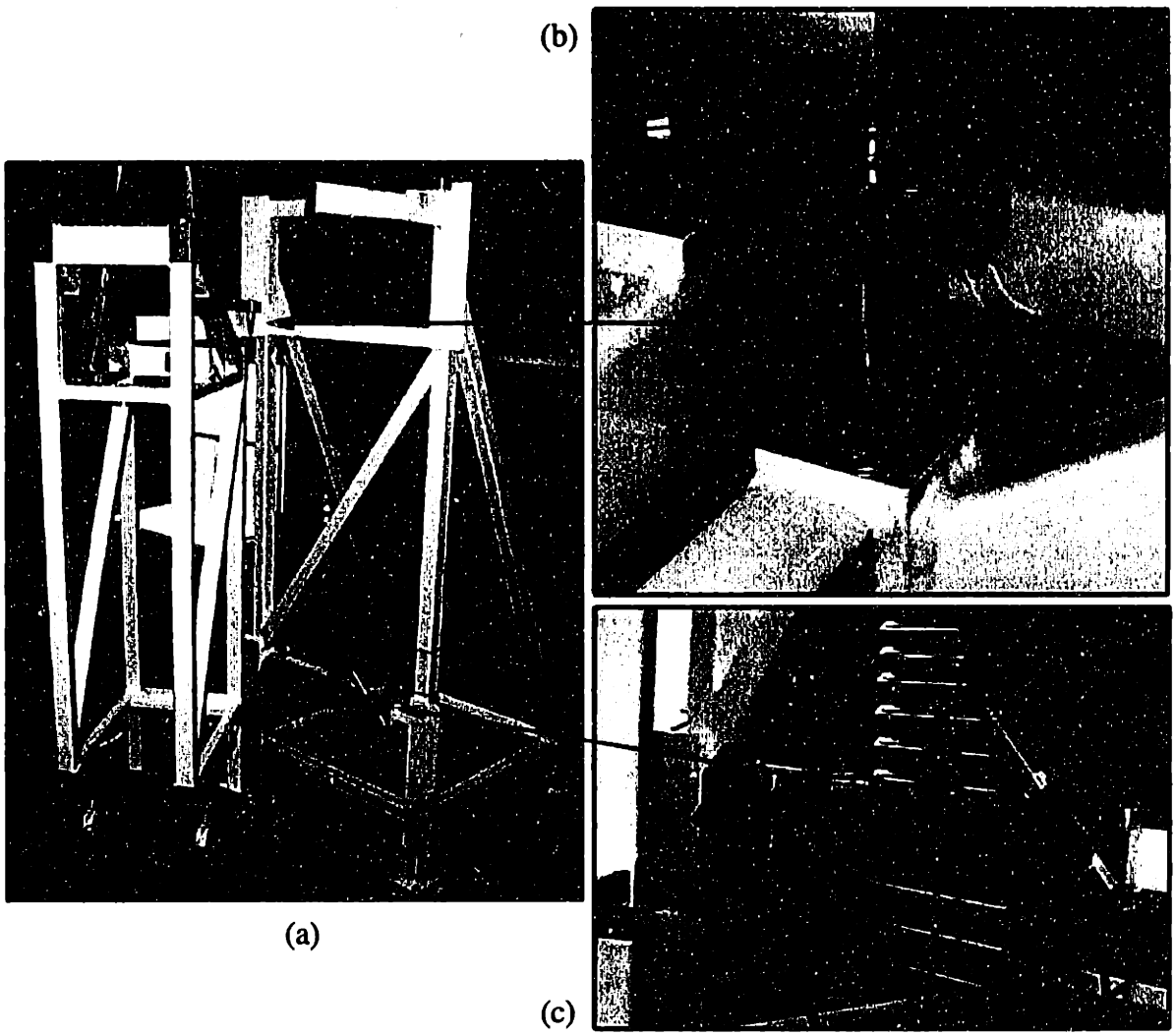


Figure 7-7 The prototype LRU in (a) is temporarily supported by a crane to better show the kinematic mounts. The LRU is over 2.5 m tall, and when mounted in the NIF periscope structure, the lowest point will be 3.6 m to 5.5 m above the floor. The upper mount has two pin-slot constraints like the one shown in (b) partially engaged. The pin attaches to the structure and is Ø 35 mm. In (c), two vee blocks on the LRU engage two Ø 32 mm actuated pins on the structure to form the lower mount.

[I] The kinematic coupling consists of three conical seats in the bottom of the laser structure and three spheres attached to the top of the canister each with a hinge axis to release the radial constraints. The upward preload is provided by the transporter that carries the canister.

[II] The bottom of the LRU has conical sockets that engage spheres on the canister lift platform each having the appropriate flexural freedom.

The upright shape and dense packing of LRU's combined with demanding stability requirements (0.6 μr rms at the optics from all sources) constrained the mounting points to lie in a vertical plane to give the most favorable aspect ratio. Further, FEA showed that the torsional mode of the LRU frame would be a limitation to achieving the vibrational part of the stability budget. This assumes that the one end is torsionally constrained, say with two vees, and the other end is free to rotate about the third vee.[1] The modal frequency can be increased somewhat by placing the instant center of the vee near the principal axis, thereby reducing the mode's moment of inertia. This naturally leads to a widely spaced vee, which has the more significant benefit of adding a stiff, frictional constraint against rotation. For example, the stiffness of this frictional constraint is an order of magnitude stiffer than the LRU frame. On the other hand, tangential friction forces in the constraints could potentially twist the frame up to 40 μr from the free state, but this is much less than the requirement for initial alignment. It is only important that the twist remains constant.

One very basic choice made early in the design process was to use gravity to preload the kinematic mount. Frankly, the alternative (a latching mechanism that would preload the LRU up against a passive kinematic coupling) was not explored as thoroughly as it should have been. Either case requires a mechanism that allows a straight-line insertion from underneath and then restrains both gravity and potential seismic loads. In the beginning, probably not enough consideration was given to the seismic load. A mechanism that works in concert with gravity would seem to be simpler than one that must defeat gravity. It is easy to become convinced this way when you have a workable concept in mind and have not carefully thought through the alternatives. Another aspect that seems to favor a gravity-loaded kinematic mount is the freer transfer of the LRU from the canister lift platform, also a gravity-loaded kinematic coupling. The weight transfers in a smooth hand off from one coupling to another as the platform rises or lowers.

The basic configuration of the LRU kinematic mount is a three-vee coupling with one widely spaced vee at the top and two vees near the bottom. The upper vee is passive with two pin-slot constraints that engage as the LRU lifts into place. The lower mount is active and formed by two vee blocks on the LRU that can pass by retracted pins on the structure. The pins extend to receive the vee blocks and support the weight of the LRU. The details of these will be presented later. This design is an inversion of the original concept that had two actuated pins mounted near the top of the LRU. There were differing opinions as to the better arrangement. The final decision was made by a fairly diverse group of twelve people using the Analytic Hierarchy Process. The results through the first criteria level appear in Table 7-1. Some of the key factors in this decision were: 1) the preference to place potential particle generators below optics, 2) better personnel access in

[1] Had there been space around the LRU, a reasonable approach would be to place three vees in a horizontal plane at the middle elevation. It still may have been difficult to meet the stability budget.

case the mechanism failed to release, 3) avoid remotely breaking pneumatic lines between the LRU and the canister, and 4) better capture range provided by a passive upper mount.

AHP Design Spreadsheet				
Created by L. Hale 2/27/97				
	Decision:	10	10	10
Criteria Level 1 >	1.00	Functionality	Design Issues	Maintenance
Criteria Level 2 >		0.33	0.33	0.33
Design Options				
Active mount location				
> LRU - upper	5.23	3.62	6.26	6.32
> LRU - lower	6.77	3.91	7.94	10.00
> Structure - upper	5.01	3.83	7.88	4.16
> Structure - lower	7.42	5.54	10.00	7.37

Table 7-1 The AHP helps provide a global picture of the decision while focusing attention to specific details. The design options were evaluated at Criteria Level 2. These are: Loading, Centering and Cleanliness under Functionality; Clearances, Seismic and Pneumatics under Design Issues; and Reliability, Release Access and Repair Access under Maintenance.

The pin-slot constraints of the upper mount provide a simple, passive engagement upon inserting the LRU into position. As Figure 7-8 shows, each constraint consists of a tapered pin attached to the structure and a slotted receiver at the top of the LRU. The combination provides the top of the LRU with approximately 15 mm of radial capture range. The upward-facing receiver also tends to catch any wear particles generated from the siding surfaces. The vee angle formed by the slots was determined to optimize centering ability as discussed in Chapter 6. This angle varies among different types of LRU's but the worse-case load governs the design of these parts that are in common. Managing the contact stress is the primary design problem for relatively heavy loads, and the need for capture range makes it difficult to use closely conforming surfaces. After working through the compromises, the results of the contact analysis appear in Table 7-2. The materials are the same as used on the lower mount and will be discussed later.

Analysis for Upper Mount		
Pin's principal radii of curvature	R_{xx} = 215 mm	R_{yy} = 17.5 mm
Slot's principal radii of curvature	R_{xx} = 45 mm	R_{yy} = inf. (straight)
Load cases: nominal and 4x nominal	P = 90 kgf	P = 4 (90) kgf
Contact pressure (compressive stress)	p = 223 ksi	p = 353 ksi
Maximum shear stress (no sliding)	τ = 72.5 ksi	τ = 115 ksi
Equivalent tensile stress $\sigma = \sqrt{3}\,\tau$	σ = 125 ksi	σ = 199 ksi
Approach of distant points	δ = 11 μm	δ = 27 μm
Stiffness at the nominal load	k = 0.70 Mlb/in	

Table 7-2 Two load cases are provided for the upper mount at full engagement. Four times the nominal load represents a dynamic overload that might occur in an earthquake. A nominal load at initial engagement has nearly identical stress as the overload case.

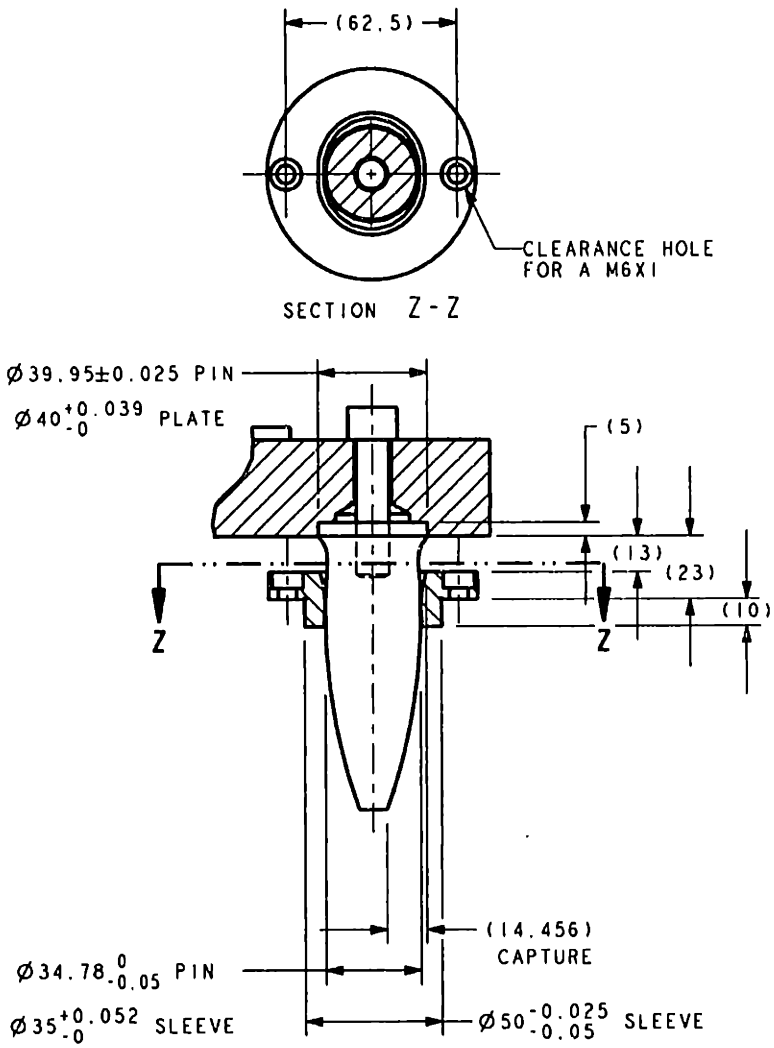


Figure 7-8 The slotted receiver at the top of the LRU engages the tapered pin attached to the structure.

The lower active mount consists of right- and left-hand assemblies of identical parts. For example, Figure 7-9 shows the left-hand assembly. Each vee block bolts to the LRU with the contact surfaces pointing down and towards the center. These surfaces are revolved so that contact with the pin occurs at two local areas. Each pin mechanism bolts to the structure and is actuated by a pneumatic cylinder. For safety reasons, the 1.50 inch bore cylinder operating at 60 psig is incapable of retracting under the weight of the lightest LRU. For control purposes, a pair of photodetectors in the canister sense retroreflective tape on the vee blocks. When the pins have extended far enough to block the return beams, it then is safe for the LRU to be lowered onto the pins. This avoids placing many hundreds of limit switches within the beam line. The angles of the pins and the surfaces of the vees were determined to optimize centering ability as discussed in Chapter 6. Due to the angle of the pin, the load is primarily compressive across its 32 mm diameter. In other words, the pin bridges the gap between the vee block on the LRU and the bore of the housing that guides the pin. The pin diameter, capture ranges and clearances between parts are as large as possible given the available space.

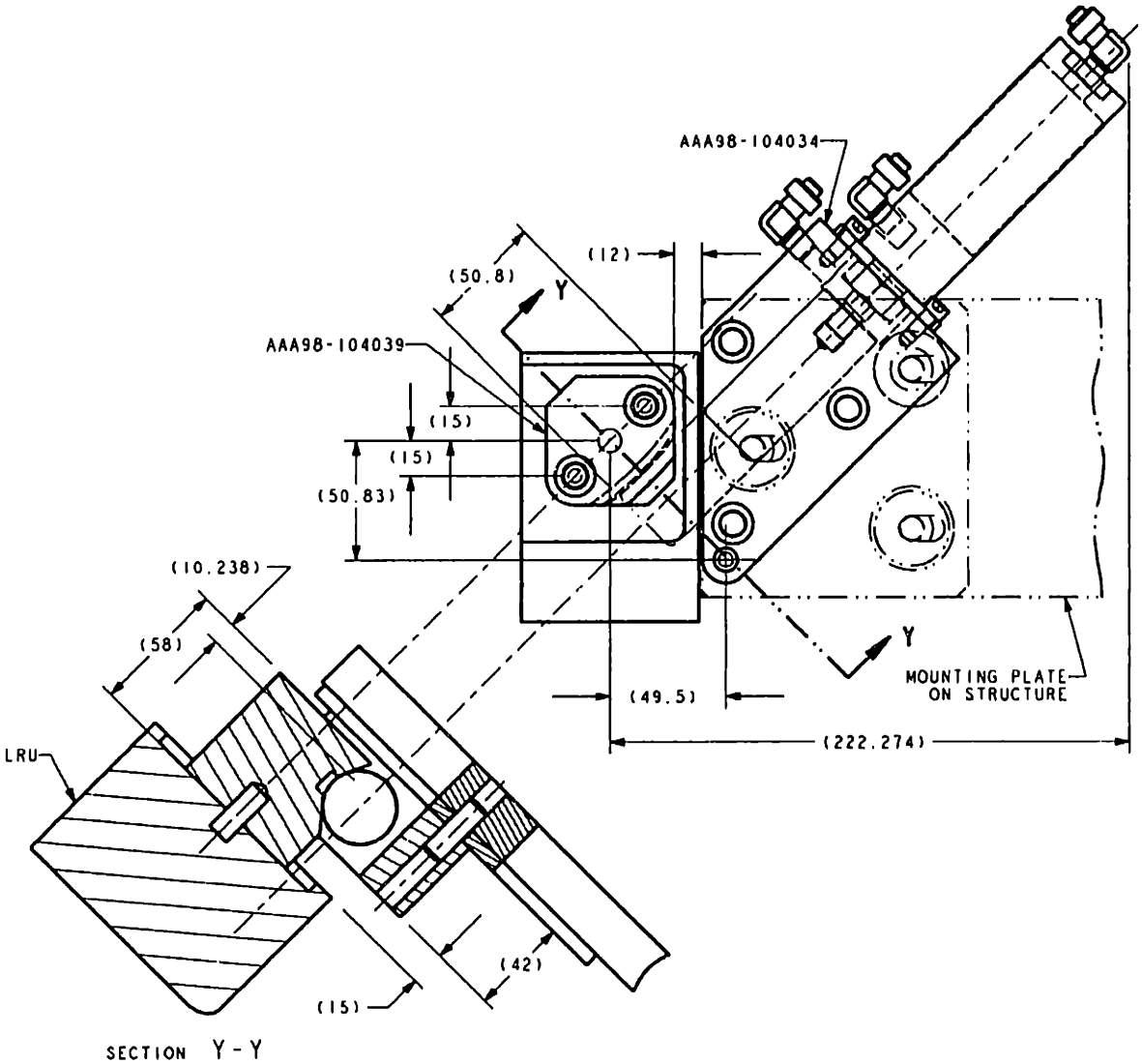


Figure 7-9 The pneumatically actuated pin extends underneath the vee block to support the LRU.

Table 7-3 shows the results of the contact analysis between the pin and vee block. These results are virtually the same as for the upper mount so the same materials, heat treatments and finishes are being used for both upper and lower mounts. The pins are made from 52100 steel to a 0.4 μm surface finish, then heat treated to 58-62 Rc, and flash chrome plated 2.5 to 5 μm in thickness. The slotted receivers and vee blocks are made from 440C stainless steel to a 0.4 μm surface finish and then heat treated to 57-60 Rc. Chrome is a hard, dissimilar material that provides greater wear resistance and lower friction. High-strength steel that has been chrome plated requires post heat treating to avoid hydrogen embrittlement. The temperature of this process is 375°F, which is about the same as the bake-out temperature being proposed to clean small parts for NIF. Many alloy steels would lose hardness at this temperature, but 52100 steel and 440C stainless steel retain full hardness. In addition, these steels are cleaner (metallurgically speaking) than most alloy steels since their predominant use is for rolling-element bearings.

Analysis for Lower Mount		
Pin's principal radii of curvature	R_{xx} = inf. (straight)	R_{yy} = 16 mm
Vee's principal radii of curvature	R_{xx} = 150 mm	R_{yy} = inf. (straight)
Load cases: nominal and 4x nominal	P = 250 kgf	P = 4 (250) kgf
Contact pressure (compressive stress)	p = 225 ksi	p = 358 ksi
Maximum shear stress (no sliding)	τ = 71.8 ksi	τ = 114 ksi
Equivalent tensile stress $\sigma = \sqrt{3}\,\tau$	σ = 124 ksi	σ = 198 ksi
Approach of distant points	δ = 16 μm	δ = 41 μm
Stiffness at the nominal load	k = 1.29 Mlb/in	

Table 7-3 Two load cases are provided for the lower mount. Four times the nominal load represents a dynamic overload that might occur in an earthquake.

Relatively high contact stress also exists between the pin and the bore of the housing at the very edge. The close fit, 75 to 25 μm diametral clearance, picks up area rapidly around the edge so it is acceptable and perhaps preferable that the housing be relatively soft and malleable compared to the pin. Therefore, the housing is made from free-machining 303 stainless steel. A Hertz analysis is not really valid for this problem because the contact area forms an arc around the pin and the edge is not a well-defined surface. Still it is possible to obtain a rough estimate of the elastic-plastic behavior by assuming an edge radius that produces a contact pressure equal to the Brinell hardness (kgf/mm^2) of the softer material. Table 7-4 shows the results of the contact analysis using the three principal radii from the geometry and the edge radius chosen to allow yielding, R_{xx} = 8 mm. This analysis predicts a significant wrap around the pin, which is why a round bore is used rather than some other shape to act as a vee constraint. The repeatability should be more than adequate since the centering of reflective optics is not very demanding.

Analysis for Pin in Housing		
Pin's principal radii of curvature	R_{xx} = inf. (straight)	R_{yy} = 15.975 mm
Bore's principal radii of curvature	R_{xx} = 8 mm	R_{yy} = -16.0125 mm
Nominal load case	P = 410 kgf	
Contact pressure (compressive stress)	p = 233 ksi	p = 164 HBn
Arc length of contact	2 (a) = 21 mm	2 (a/r) = 74°
Width of contact	2 (b) = 0.23 mm	
Approach of distant points	δ = 9.5 μm	
Stiffness at the nominal load	k = 3.6 Mlb/in	

Table 7-4 The elastic-plastic behavior between the pin and housing of the lower mount is estimated by an elastic Hertz analysis using an edge radius that makes the contact pressure equal to the material hardness.

7.6 Tip-Tilt Mounts for NIF Large-Aperture Optics

All 192 beam lines in the NIF require up to eight large-aperture laser mirrors, LM1 through LM8, and a polarizer optic to direct light through the system. The mount for each reflective optic requires tip-tilt actuation with a program step size of the order 0.1 μr over a range of 10 mr. In addition, the mount must support the optic sufficiently well to meet the wavefront error budget and have sufficient rigidity to meet the stability budget. There are three basic mount designs being used for NIF reflective optics. One is unique to LM1 because it is a deformable mirror and not particularly challenging to mount. The second type is the topic of this section. LM2 and the polarizer require full-aperture light to pass through the mount. LM3 is very similar to the polarizer since the pair forms the periscope optics. The third type supports LM4 through LM8 from the back side where there is free access. This design features a tripod flexure with the instant center placed at the centroidal plane of the optic. The attachment points for the tripod and two actuators are chosen to minimized wavefront error. A fourth flexure constrains in-plane rotation about the tripod.

The traditional approach to a full-aperture optic mount would be a bezel that clamps the faces of the optic in one of several possible ways. Clamps at three local areas would be the most kinematic, or four clamps would be acceptable if the bezel were torsionally flexible and the clamps were initially coplanar. It is also common to use a full-length compliant element such as an o-ring. In hind sight, the bezel mount with four clamps may have posed the fewest problems, but a preconception that the optic must fit within the LRU frame left too little space for a bezel. There would have been space problems but not insurmountable ones if the bezels were external. The attractive feature of a bezel is greater freedom in how the tip-tilt mechanism fastens to the bezel rather than directly to the optic. However, with a direct mounting solution in mind, it is compelling not to use a bezel.

Figure 7-10 shows the tip-tilt mount being used for NIF periscope optics. These optics are fairly large (807 x 417 x 90 mm for the polarizer and 740 x 417 x 80 mm for LM3) and inclined 33.6° from a horizontal plane. A similar mount is being used for LM2 except the optic is 412 x 412 x 80 mm and in a vertical plane. There are three support points that lie in the centroidal plane of the optic. The mount provides two constraints at each point giving a total of six. As noted, two support points are actuated in the out-of-plane direction to provide tip-tilt motion. The three constraint lines and instant centers (one is off the page) help in visualizing the in-plane constraints. The physical connection between the optic and each support is a separable spherical joint. Each support must release one degree of freedom of the three defined by the spherical joint to give the required two constraints at each support. A simple flexure hinge at the passive support releases the optic along its centerline. The arm of the actuated support is free to rotate about the same bearing that provides the actuated motion. This releases motion of the optic about the upper instant center, for example. Try standing with your legs apart to simulate this motion. Your hips are equivalent to the spherical joints and your ankles are equivalent to the bearings.

The optic has three conical sockets machined into the edges to receive three plastic bearing inserts. The shape of the insert is slightly toroidal to form an annular contact with a hardened stainless steel ball. The thickness of the insert is only 0.5 mm between the conical socket and the 12 mm diameter ball. The included angle of the cone and the annular contact area is 40°. An axial force is required to maintain engagement of the ball, insert and cone. A compression spring provides the preload for the passive support, and the weight of the optic preloads the active supports. In addition, seismic restraints prevent the supports from completely disengaging if there is not adequate preload. Several plastic materials were tested and Torlon produced the least creep and provided relatively low friction.

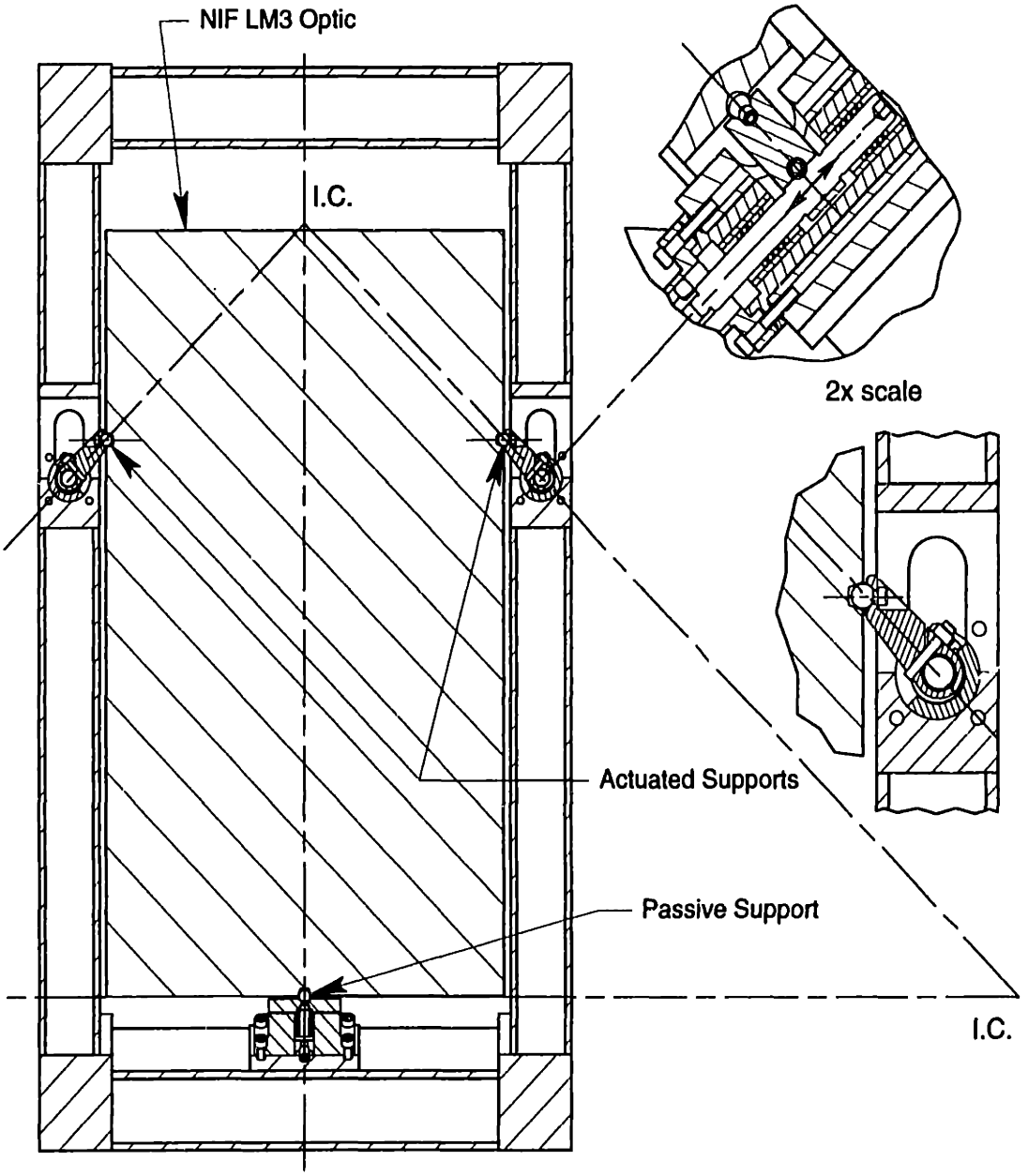


Figure 7-10 The NIF edge-style mount supports the optic at three conical sockets machined directly into the glass. Two actuated supports provide tip-tilt motion of the optic about the third passive support.

8 Anti-Backlash Transmission Design: *a natural extension of exact-constraint design*

Servo mechanisms used in precision machines provide constrained motion along some desired path. The path may be a simple translation or rotation, or a coordinated combination of several simple motions. In simplest terms, a servo mechanism is a constraint that provides controllable motion along a degree of freedom. The servo mechanism may be used in combination with various other types of constraints described elsewhere. The Stewart Platform is an example of six linear-motion servos arranged as exact constraints to control the motion of six degrees of freedom [Stewart, 1965-66]. Machine tools based on the Stewart Platform have been an active research area in recent years [Pritschow and Wurst, 1997].

The actuator of choice for high dynamic range applications is the electric motor.[1] Table 8-1 list several types and forms of electric motors that are applicable to servo mechanisms. Since electric motors are available with rotary, linear or planar motion, the use of a transmission at all must be considered carefully. Real transmissions present several undesirable behaviors that in excess are incompatible with precision servo mechanisms. Compliance, inertia (leading to mechanical resonances), error motion, friction (leading to hysteresis), wear, heat generation, and backlash can often be reduced to acceptable levels by careful design and precision manufacturing. In particular, backlash can be eliminated by preloading the mechanical elements, but some level of hysteresis will remain due to the interaction between friction and compliance. In addition, there is a dollar cost associated with these improvements. The primary motivations to use a transmission are to obtain a high velocity ratio and/or to change rotary motion to linear motion. Both are described by a transmission ratio *R*.

Type\Form	*Rotary*	*Linear*	*Planar*
C-L, PM Brush	common	uncommon	uncommon
C-L, PM Brushless	common	common	uncommon
C-L, AC Induction	less common	common	uncommon
O-L, Hybrid Stepper	common	common	common
C-L, Hybrid Stepper	less common	less common	common

Table 8-1 Most rotary motors are used with transmissions, although direct-drive applications are becoming more common. Linear and planar motors are almost always used in direct-drive applications. C-L and O-L refer to closed-loop or open-loop control, respectively.

[1] Hydraulic servos were common in the machine tool and robotics industries into the late 1970's but today they are rare. However, recent research demonstrates hydraulic servos can provide nanometer resolution and N/nm stiffness [Kanai, et al., 1991].

Electric motors usually operate most efficiently at much higher speed and deliver much lower torque than required by the servo mechanisms in typical precision machine tools.[I] Furthermore, the motor's weight, cost and heat generation increase in proportion to its torque output. This situation favors a relatively large transmission ratio over a directly coupled motor. As a further benefit to the closed-loop control system, a large transmission ratio reduces the load inertia reflected to the motor by R^2 and increases the servo stiffness reflected to the load by R^2. This effect makes the system easier to control but not necessarily better due to the non ideal behaviors of real transmissions.[II]

The decision whether to use a transmission or to directly coupled the motor depends upon the application, the available motors (or motor technology) and the available transmissions (or transmission technology). As a general recommendation for precision applications, use a transmission only when absolutely necessary, for example, when the motor cannot provide enough force or torque, the heat produced cannot be captured sufficiently, or the cost is prohibitive. Assuming that the decision is to use a transmission, then the goal becomes: *design (or select) a transmission that provides acceptable behavior.*

This chapter presents several types and configurations of transmission designs that provide high ratios and have zero backlash. It provides the basic understanding (design theory, case studies, test results) required to design a new anti-backlash transmission or to select suitable commercial devices. To motivate why backlash is a problem, please see Appendix F *Friction and Backlash in Servo Mechanisms* for a dynamic simulation of a servo mechanism that includes backlash and friction in the model.

8.1 Preloaded Rolling-Element Bearings

The technique of preloading rolling-element bearings in back-to-back or face-to-face configurations is quite common and effective for increasing stiffness and removing backlash in a spindle, for example. Preload is generated by fitting the parts together with interference rather than clearance.[III] The same technique applies to transmissions such as ball screws and gear trains (as discussed in the next section). Rolling element bearings are tolerant to preload because they are accurately made and locally compliant in the Hertz

---

[I] High-torque, variable-reluctance servo motors are specifically designed for direct-drive applications.

[II] Many servo designers follow a rule of thumb that the transmission ratio should make the reflected load inertia equal to the motor inertia. This ratio will maximize the acceleration of the servo assuming it is unloaded. If acceleration is not important, then neither is this rule of thumb. If the servo must also overcome a static load such as gravity, then the maximum acceleration occurs for a ratio

$$R = \frac{T_L}{e \cdot T_M} + \sqrt{\left(\frac{T_L}{e \cdot T_M}\right)^2 + \frac{J_L}{e(J_M + J_T)}}$$

where T_L = load torque, T_M = motor torque, J_L = load inertia, J_M = motor inertia, J_T = transmission inertia at the motor, and e = transmission efficiency.

[III] Preloaded angular-contact ball bearings are manufactured such that the faces of the inner and outer races are flush under the applied preload.

contact regions. The mechanisms work and last because sliding friction is minimal. However, they will not work very well or last very long if: the lubrication is inadequate, the bearing surfaces become contaminated with hard particles, or the as-mounted geometry is too inaccurate (bores and shafts not round and concentric, shoulders not square and parallel, the fits too tight, etc.).

The mechanical interference that generates the preload is shared among the components. For example, each bearing in Figure 8-1 tolerates approximately equal interference and must conform approximately the same to any geometric inaccuracies in the parts, mounts or alignments. If the preload is to remain fairly constant, then the parts, mounts and alignments must be several times more accurate than the interference required for preload. Since the components are accurately made to the micron level, commensurate care in mounting and alignment is necessary to ensure fairly constant preload.

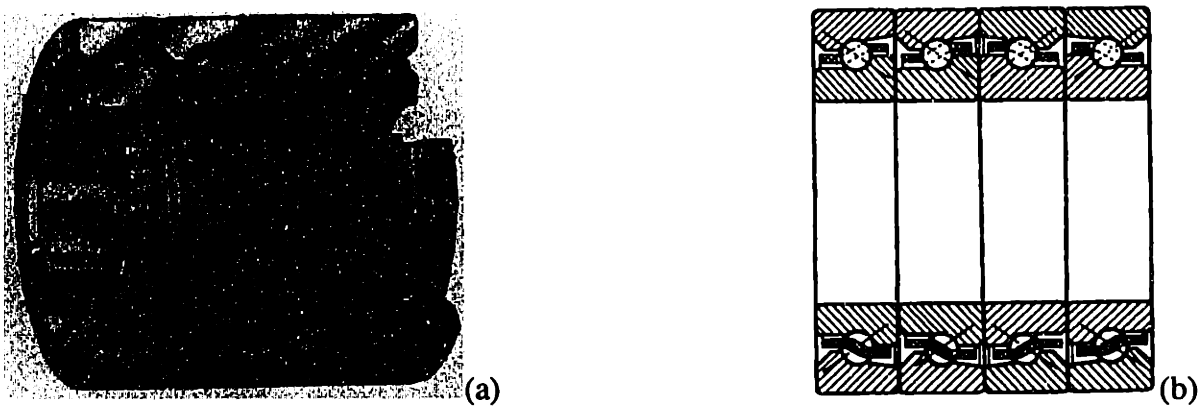


Figure 8-1 This type of angular-contact thrust bearings is specifically manufactured for ball-screw applications requiring high axial stiffness. The arrangement shown is face-to-face since each bearing faces inward as indicated by where the lines of contact intersect. Often angular-contact bearings are manufactured as DU (duplex universal) so they can be used back-to-back or face-to-face and with any number of bearings. Reproduced from a catalog courtesy of Barden Precision Bearings.

Measuring the preload and especially the variation is challenging in an assembled system. Lack of preload shows up clearly as backlash but too much preload is not readily apparent in its freedom of movement. Indirect measures are useful but are difficult to quantify. For example, the operating temperature of a spindle increases with preload and misalignments. Measurements of error motion may indicate inaccuracies that also cause the preload to vary. Since stiffness increases with preload, measurements of stiffness can indicate the level of preload with a prior knowledge of force-defection curves. However, a direct measurement usually requires some invasion of the assembly.

Figure 8-2 provides a clear example of the issues discussed so far. The ball screw in (a) has two nuts under a preload that is controlled by the thickness of a spacer (there must be a pair of keys or pins that register the relative angular orientation). The preload diagram in (b) shows the initial displacement or interference δ_{ao} required for each nut to provide the desired preload F_{ao}. An external axial force F_a produces an axial displacement

of both nuts by δ_a. As long as both nuts carry load, even though in opposite directions, the stiffness of each nut contributes to the total axial stiffness. We expect the preload to vary as the nut pair travels up and down the ball screw due primarily to size and lead variations in the screw. This would be represented in the preload diagram by the curves shifting horizontally to one another. One way to measure the variation would be to instrument the spacer perhaps with a piezoelectric device to measure load directly. Another way would be to measure axial stiffness at points along the screw.[1] The concern about fairly constant preload results from having too much compliance or possibly backlash at the low extreme and premature failure or too much heat generation at the high extreme.

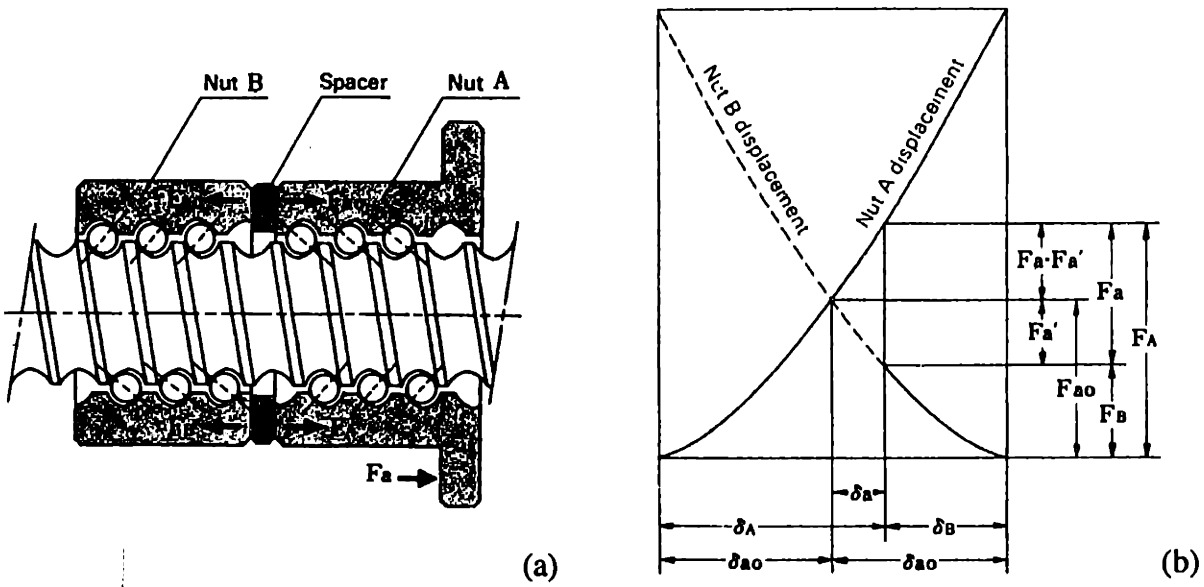


Figure 8-2 A common way to preload a ball screw (a) is to load one nut against the other. The preload diagram (b) represents the forces and displacements that take place between the two nuts. Reproduced from a catalog courtesy of THK.

Rolling-element bearing technology has been applied to a number of practical devices that make excellent low to zero backlash constraint devices. The proper use depends on understanding the primary and secondary constraint directions (usually the degree(s) of freedom is obvious). Figure 8-3 show a ball spline designed primarily to transmit torque along an axial degree of freedom. Under a torque load, three raceways act together to self center the spline, constraining all but the axial degree of freedom. In the opposite direction, a different set of three raceways constrains the spline to a slightly different center and orientation. The manufacturer must accurately grind all raceways in the spline parallel and at the proper radius and angular spacing so that the preload can be set with the size of balls installed. A preloaded ball spline makes a very good linear bearing for applications requiring simply supported rails. Moment load capacity and stiffness (as to

[1] A local displacement measurement between the nut and screw avoids the column compliance of the screw influencing the stiffness measurement.

cause bending of the shaft) are a good deal less than those for torsion, thus they are considered secondary constraints.

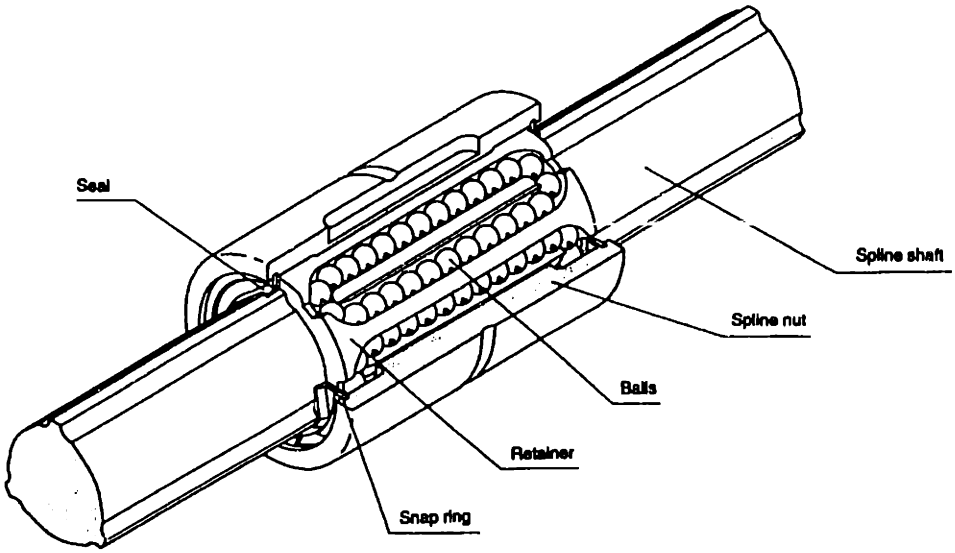


Figure 8-3 Although designed to transmit torque, the ball spline is an effective linear bearing if the bending compliance of the shaft is acceptable. Reproduced from a catalog courtesy of THK.

Figure 8-4 shows what has become known as a linear guide. Specifically designed as a linear bearing, four raceways provide sufficient redundancy to constrain all but axial travel with or without preload. This design is a face-to-face configuration having somewhat less moment capacity and stiffness about the rail axis as a back-to-back configuration. There are many variations on this basic scheme available from several manufacturers.

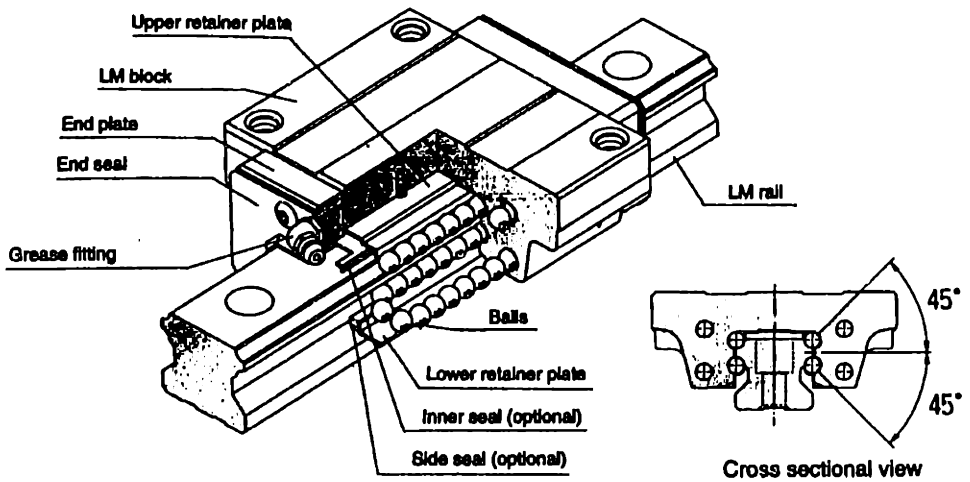


Figure 8-4 The linear guide has four raceways and continuous support along the length of the rail. Reproduced from a catalog courtesy of THK.

Figure 8-5 shows better the inner workings of a typical ball screw. Usually the screw rotates and the nut translates but several other combinations are possible. Although two directions of motion occur, they are coupled by the lead of the screw and count for only one degree of freedom. The primary constraint direction (as typically used) is axial translation. The ball screw provides secondary constraint in the remaining directions. The issue with secondary constraints is particularly significant for ball screws. Ideally, the nut should be free to follow the path defined by the screw and the only rigid connection between the nut and the slide system will be axial translation and rotation. More typically, the nut is fully constrained to the slide, which requires accurate alignment to prevent destructive loads on the ball screw. This cause of failure is prevalent enough to make people believe that a ball screw cannot function as a linear bearing. It may have relatively large cyclic error motion but its secondary constraints can carry modest loads. Interestingly, the usual motivation for decoupling a ball screw's secondary constraints from the slide axis is to prevent the screw's cyclic errors from influencing the slide motion rather than protecting the screw from overload.

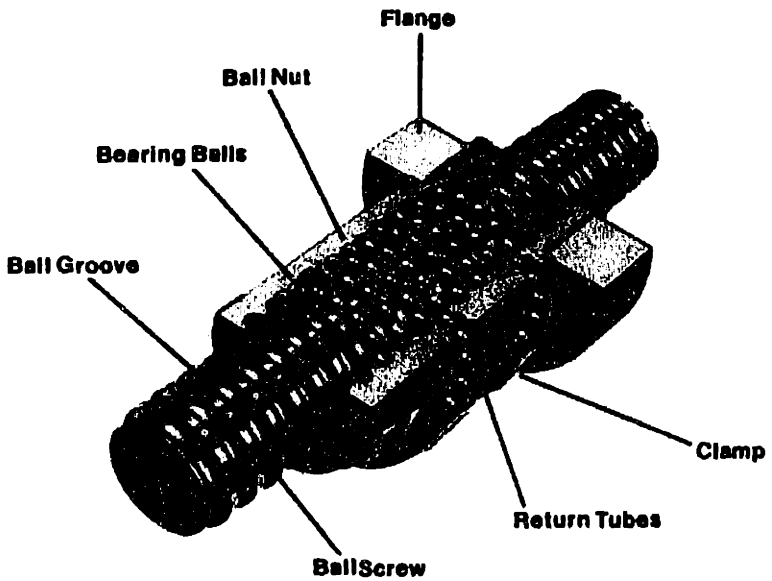


Figure 8-5 Ball screws come in sizes ranging from a few millimeters to a few hundred millimeters and leads ranging from sub millimeter to several tens of millimeters. There are several basic configurations from which to choose. Rolled-thread screws are less expensive and less accurate than ground threads. Reproduced from a catalog courtesy of Warner.

Figure 8-6 shows an interesting extension of the ball screw to a worm-gear speed reducer [Fengler, 1970]. Traditional worm gears depend on fully developed hydrodynamic lubrication to have good efficiency, which is difficult to maintain in most servo mechanisms. The U.S. Patent 3,494,215 for this unique anti-friction worm gear was assigned to Fordson Drive of Dearborn Michigan but I am unaware of a commercial source.

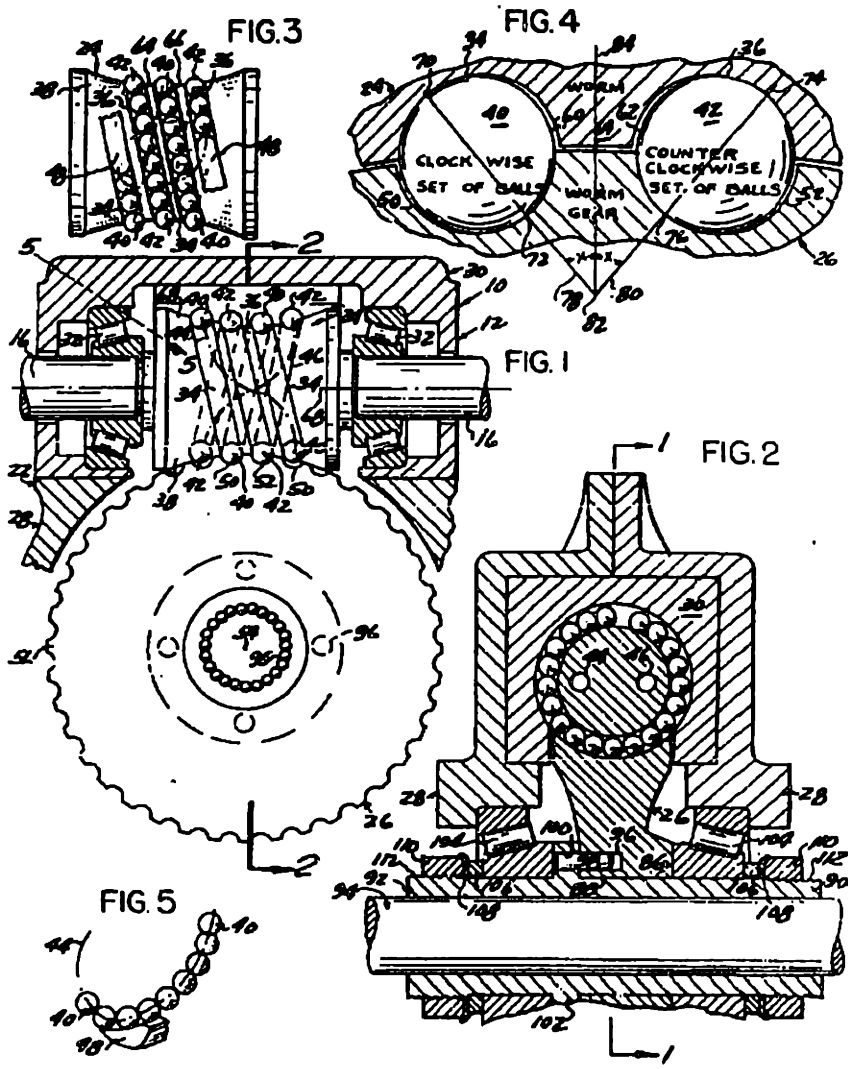


Figure 8-6 An anti-friction, anti-backlash worm drive is analogous to a ball screw (from U.S. Patent 3,494,215).

8.2 Preloaded Gear Trains[1]

Gear trains used in precision servo mechanisms must have very low or preferably zero backlash and be high in stiffness. The technique of preloading, which works well for rolling-element bearings and ball screws, is also applicable to gear trains provided that sufficient compliance exists to tolerate manufacturing inaccuracies. This section presents design theory for two basic methods of mechanically preloading gear trains. Named simply *Type 1* and *Type 2*, these methods are predominant in the patents for anti-backlash gears. Two case studies, a robot revolute joint and a machine tool axis of rotation, help to reinforce the theory and raise practical issues.

References on anti-backlash gear trains are very few [Michalec, 1966], but a number of patents exist for applications primarily in robotics and machine tools. Most of the patents are Type 1, meaning there is one stiff path between the input and the output. Preload is applied by a second path that has built-in compliance to tolerate manufacturing inaccuracies. A Type 1 design is applicable when loads are light or primarily in the direction carried by the stiff path. Figure 8-7 illustrates this concept with two configurations that are functionally equivalent. Part (a) shows two distinct transmission paths from the input (Item 10) to the output (Item 11) [Baumgarten and Schloeglmann, 1990]. Part (b) also has two paths but they are concentric. It requires preloaded bearings since the loads may reverse.

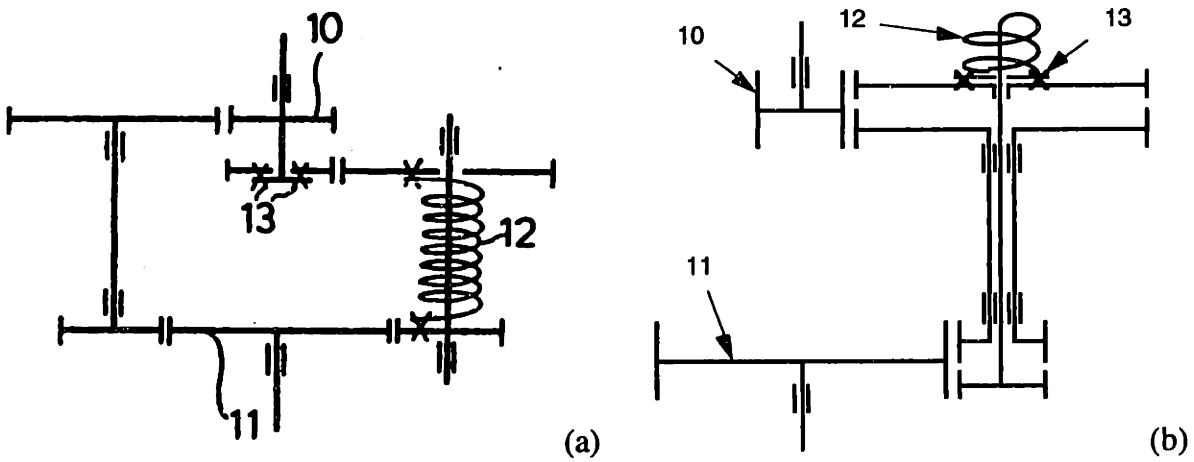


Figure 8-7 A Type 1 anti-backlash gear train has one stiff path and one path that includes a compliance element (Item 12). Part (b) has concentric paths and requires less space. By means of a preload adjustment device (Item 13), the paths can be set to oppose one another to eliminate all backlash.

Figure 8-8 shows a Type 2 design for a large tracking antenna [Cresswell, 1972]. A Type 2 design has two stiff paths between the input and the output. Compliance is introduced by a preload device usually located at the common input to the two paths. In this

[1] This material was previously published [Hale and Slocum, 1994].

example, the preload device consists of a compliant spring (Item 23) acting through a differential (Item 11) which causes the bevel gears (Items 20) to counter rotate, thus removing backlash in each path and generating preload. This particular design is too complicated but clearly illustrates a Type 2 design. A Type 2 design is applicable when stiffness is important and moderate to heavy loads are expected in either direction.

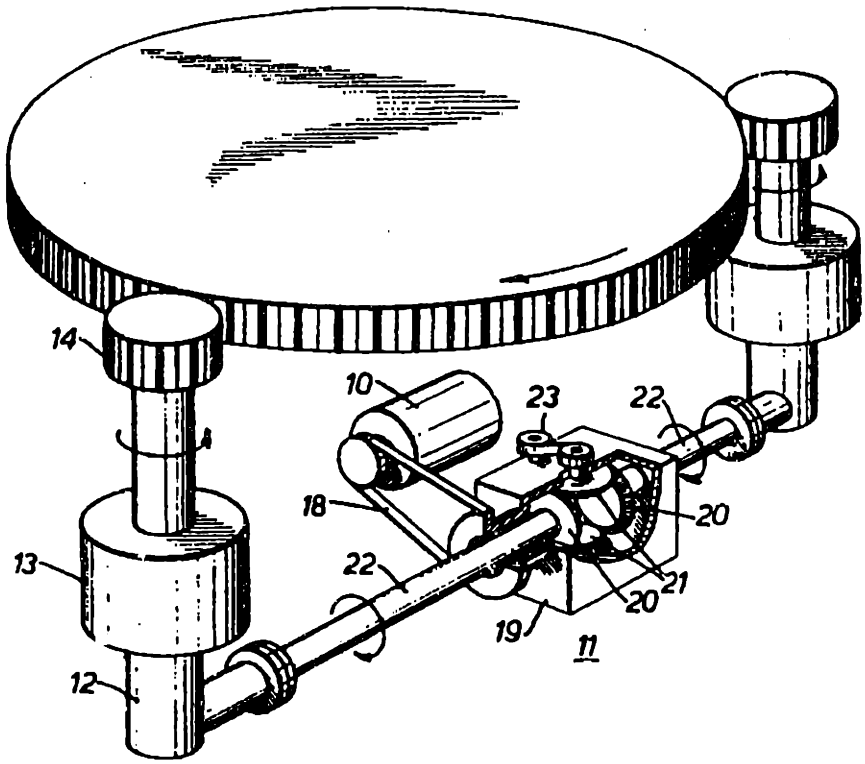


Figure 8-8 Both paths of this large tracking antenna contribute their full stiffness to the drive, making it a Type 2 design (from U.S. Patent 3,665,482).

8.2.1 Modeling Preloaded Gear Trains

Even though nonlinear effects exist, simple linear spring models are sufficient to explain the characteristics of Type 1 and Type 2 preloaded gear trains. The force-deflection curve for a transmission with ball bearings will be nonlinear due to Hertzian contact, but often one path hardens as the other softens giving a nearly linear curve. More significant is the effect of the teeth moving through the mesh. The number of teeth in actual contact changes as the gears rotate, thus the mesh stiffness changes rather abruptly by a factor two or so for spur gears. However, shaft and bearing compliances are usually larger thereby masking much of this behavior. The linear models remain useful for design calculations.

The details of calculating path stiffnesses are not essential to the development of the models; however, four facts are important to keep in mind:

- The stiffnesses of components in parallel add.
- The compliances (reciprocal of stiffness) of components in series add.

- A stiffness reflected through a transmission ratio R increases by a factor R^2.
- Similarly, a linear stiffness acting through a lever arm r requires multiplication by r^2 to give the equivalent torsional stiffness.

Finding the total stiffness of a path becomes an assembly of component stiffnesses. Since most components are in series, it is often easier to work with compliance reflected to a common location and type, for example, torsional compliance at the output shaft. Most bearing manufacturers will supply force-deflection curves for their bearings, the derivative of which is stiffness. Beam theory gives the compliance of most other components. Particularly valuable is the compliance of a typical gear mesh along its line of action, given by Equation 8.1, where E is the elastic modulus and w is the face width [Hale and Slocum, 1994].

$$c_{mesh} = \frac{6.2}{E \cdot w} \tag{8.1}$$

The preload diagram is a good visual aid for understanding what happens to forces and deflections in each path of a preloaded device. The two diagrams in Figure 8-9 are force-deflection curves for two quite different designs. In (a), the paths have equal stiffness ($k_1 = k_2$) like a duplex pair of angular-contact bearings. In (b), the path represented by k_1 is more complaint than the path represented by k_2. The curves are linear except for backlash at zero force represented by b in the figure. Preloading one path against the other to a force F_L requires an interference represented by the total deflection δ_L around the loop formed by the two paths. The overlapping curves have, at the point of intersection, zero net drive force F_D and zero drive deflection δ_D by convention. The vertical line of length F_D represents the drive force developed in the transmission when it is deflected a distance δ_D from the intersection point along the horizontal axis. The total drive stiffness k_D follows directly from the figure as simply the parallel sum of path stiffnesses, as shown in Equation 8.2. Should one path enter a region of backlash, then the drive stiffness reduces to the stiffness of the other path.

$$k_D = \frac{F_D}{\delta_D} = k_1 + k_2 \tag{8.2}$$

Each transmission path will have errors from such sources as tooth form error, pitch-line run out and bearing run out. These errors cause δ_L to vary during operation and F_L varies as a result, where the constant of proportionality is the loop stiffness k_L. As Equation 8.3 shows, the loop stiffness is the series sum of path stiffnesses. Typically the compromise is to achieve maximum drive stiffness for an allowable variation in preload, perhaps 25%. An appropriate indicator is the ratio of drive stiffness to loop stiffness, given by Equation 8.4. The least optimal ratio occurs when $k_1 = k_2$, and it gets better as k_1 and k_2 become different. This leads to a Type I design where one path has extra compliance to

make the transmission more tolerant to errors. Equal path stiffness may be appropriate if the application requires equal bi-directional stiffness outside the preload region, and control of transmission errors is very good as compared to δ_L.

$$\frac{1}{k_L} = \frac{\delta_L}{F_L} = \frac{1}{k_1} + \frac{1}{k_2} = \frac{k_1 + k_2}{k_1 \cdot k_2} \tag{8.3}$$

$$\frac{k_D}{k_L} = \frac{(k_1 + k_2)^2}{k_1 \cdot k_2} \tag{8.4}$$

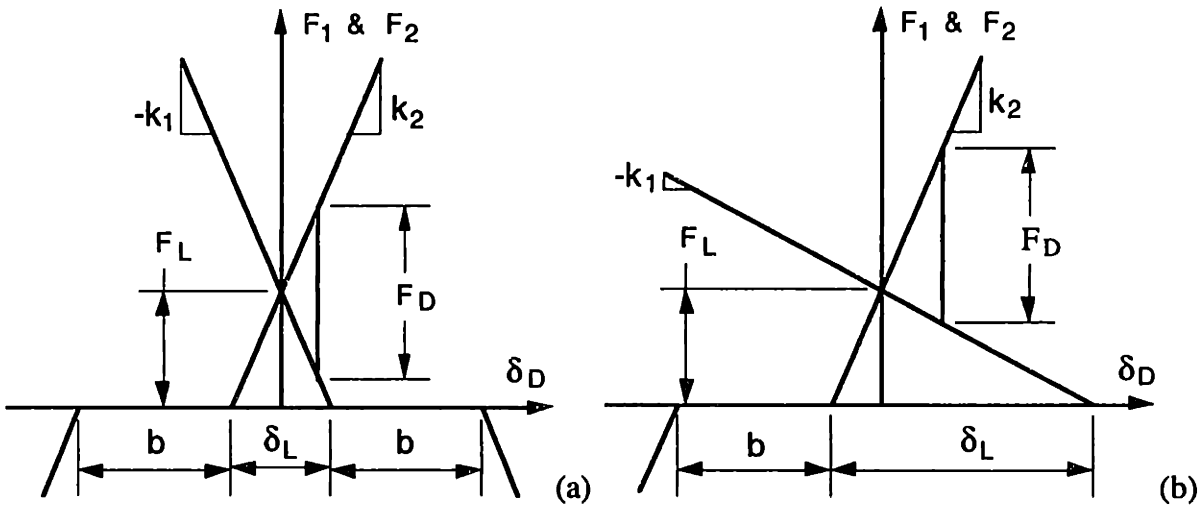


Figure 8-9 Preload diagrams show cases of equal path stiffnesses in (a) and significantly different path stiffnesses in (b). The case represented in (b) is more tolerant of transmission errors within the loop.

A Type 2 design is different because the added compliance is sensitive to variations in preload but is insensitive to variations in drive load. To accomplish this, a mechanism like the differential (Item 11) in Figure 8-8, divides the drive force evenly between the two transmission paths. Then it becomes a simple matter to add a spring to the mechanism that preloads both paths. Figure 8-10 demonstrates this with a linear spring model, where the springs k_1 and k_2 represent the two transmission paths, and k_3 is the preload spring. In (a), the forces in both paths, F_1 and F_2, add to equal the drive force F_D (actually the torque applied to the input shaft). The fact that F_1, F_2 and F_3 are in balance gives a second equation. By representing forces in terms of spring deflections and stiffnesses, and by eliminating the deflection of k_3, the two equations will combine to give the drive stiffness shown in Equation 8.5. In Figure 8-10 (b), only a loop force results from a loop displacement. As before, the loop stiffness k_L is the series sum of path stiffnesses plus the preload spring factored by a lever ratio (squared), as Equation 8.6 shows.

$$k_D = \frac{F_D}{\delta_D} = (k_1 + k_2) - \frac{(k_1 - k_2)^2}{(k_1 + k_2 + k_3)} \tag{8.5}$$

$$\frac{1}{k_L} = \frac{\delta_L}{F_L} = \frac{1}{k_1} + \frac{1}{k_2} + \frac{4}{k_3} \tag{8.6}$$

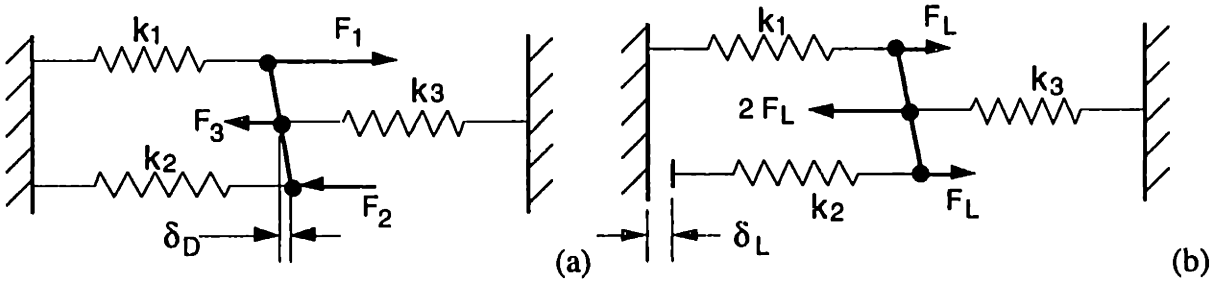


Figure 8-10 A displacement of the drive results in forces given in (a). A displacement in the loop due to component errors results in forces given in (b).

Usually the transmission paths are identical so that $k_1 = k_2$. Then the ratio of drive stiffness to loop stiffness becomes very simple. Equation 8.7 shows that k_3 has a significant effect on the ratio. If on the other hand k_3 were infinitely large (zero compliance), then Equations 8.5 and 8.6 would reduce, respectively, to Equations 8.2 and 8.3 for the Type 1 model. The benefit of a Type 2 design is much lower loop stiffness with no sacrifice to drive stiffness.

$$\frac{k_D}{k_L} = 4 + 8\frac{k_1}{k_3} \qquad \text{for} \qquad k_1 = k_2 \tag{8.7}$$

8.2.2 Designing Preloaded Gear Trains

While it is entirely possible to remove all mechanical backlash between the servo motor and the final drive, design constraints and other factors may lead to a compromise where one or more gear meshes near the motor have minimum backlash. For example, two case studies that follow each had a well-defined space constraint and a *best effort* was made to provide high stiffness, minimum backlash and *reasonable cost.*[1] The goal of this section is to develop an understanding of important design issues and to share some valuable techniques that illustrate practical implementation of the theory presented.

The distributions of inertia, compliance, friction and backlash throughout the gear train are important to understand when developing and evaluating design options. Each gear mesh i in the gear train contributes a factor R_i to the total transmission ratio. Each gear mesh contributes inertia, compliance, friction and backlash that vary from mesh to mesh. Furthermore, the values are different when reflected to other shafts in the gear train,

[1] A servo mechanism can tolerate backlash that occurs outside the position loop. Backlash that occurs between the servo motor and the position sensor is destabilizing. Both case studies had the encoder directly coupled to the servo motor to avoid this problem.

typically the input or the output shaft. It is useful to consider these properties in invariant forms so that all the parts can be compared regardless of location in the gear train.[1]

The invariant form for inertia is the kinetic energy of a gear train operating at an arbitrary speed, say unit speed at the output. Each component in motion contributes to the total kinetic energy. The invariant form for compliance is the strain energy of a gear train under an arbitrary load, say unit torque at the output. Each component under load contributes to the total strain energy. Kinetic and strain energy can easily be tabulated in a spreadsheet to provide a distribution among components. Another useful distribution is the square root of the strain energy to kinetic energy ratio (see Section 4.2 *Modeling Complex Structures* on using the Rayleigh quotient as a sensitivity indicator). You may think of this as a frequency distribution. To get better performance, stiffen components with high frequency and lighten components with low frequency. A balanced design will have a fairly uniform frequency distribution and sufficient margin on stresses throughout the gear train.

The invariant form for friction is the efficiency, which is the ratio of power output to power input. Usually each mesh will have about the same efficiency. The total efficiency for a conventional single-path gear train is the product of efficiencies for all meshes. Clearly using fewer stages is better but this generally requires more space for a given transmission ratio. A preloaded gear train is more complicated because the power going through the preloaded gears is greater than the input power. For a Type 2 design, the combined friction in both paths is constant and depends on the preload setting rather than the drive load. The situation is different for a Type 1 design because it depends on the direction of the load. This is best understood from the preload diagram, Figure 8-9 (b).

Backlash does not have a nice intuitive invariant form. We get one by expressing backlash at each mesh in terms of a common unit such as the number of encoder counts required to take up the clearance. If some backlash is tolerable, then a good strategy is to preload only the meshes near the output where clearance is most detrimental. Each preloaded mesh experiences increased friction, wear and fatigue, in addition to costing more and requiring extra space. Any other mesh without preload should have the minimum backlash possible by tolerance or by adjustment. This strategy eliminates almost all backlash, adds inertia where it has the least impact, and adds stiffness where it is most effective. The two case studies follow this strategy.

8.2.2.1 Type 1 Design Issues

A Type 1 design can be simpler, more compact and less expensive than a Type 2 design. It is most appropriate when the loads are light or predominately in the direction carried by the stiff path. Figure 8-11 shows a Type 1 design with concentric paths to reduce space. The inner path is naturally compliant with a shaft size determined to accommodate anticipated

[1] This approach seems more intuitive but is no different from reflecting all terms to a common point.

loop errors. The maximum load carried by the inner path is considerably less and accordingly has less face width than the outer path. A frictional interface at the right end of the inner path allows a phase adjustment using a special tool to apply the preload. The amount of preload set should be slightly greater than the maximum expected load tending to unload the stiff path.

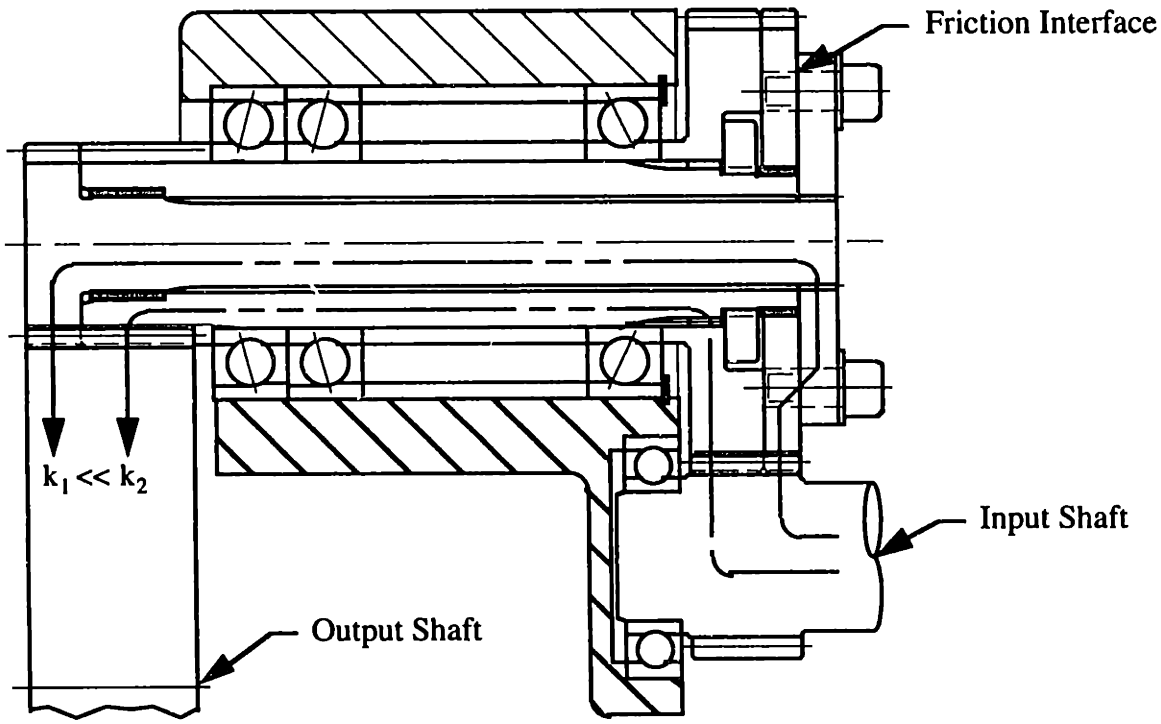


Figure 8-11 A concentric Type 1 design requires fewer anti-friction bearings but they require preload since the direction of load may change.

8.2.2.2 Type 2 Design Issues

Applications requiring high stiffness and equal bi-directional load capability favor the Type 2 design. The distinguishing feature for the Type 2 design is a preload device usually located at the common input to the two transmission paths. This device has two degrees of freedom; one that rotates the two paths in normal fashion and another that counter rotates the two paths to apply preload. It splits the drive torque equally between the two paths without loss of stiffness. It applies preload compliantly and with minimal friction. In addition, there must be a means to set or adjust the preload. The following are three ways that a device could generate preload in each path independently from the drive motion:

- Radial motion of a gear common to both paths and in the direction having equal components along each line of action, as shown in Figure 8-12.
- Axial motion of a pair of right and left handed helical gears, or of one herringbone gear.
- Opposite angular motion of two gears related by a differential as shown in Figure 8-8.

Radial motion preload is simple and compact. Figure 8-12 shows the optimum configuration where the two lines of action are parallel and opposite to the externally applied preload force F_p. This configuration minimizes the lateral force on the linear motion bearing that supports the input gear. The optimum is fairly flat within several degrees of the operating pressure angle. Since both case studies use this configuration, further discussions will follow in those sections.

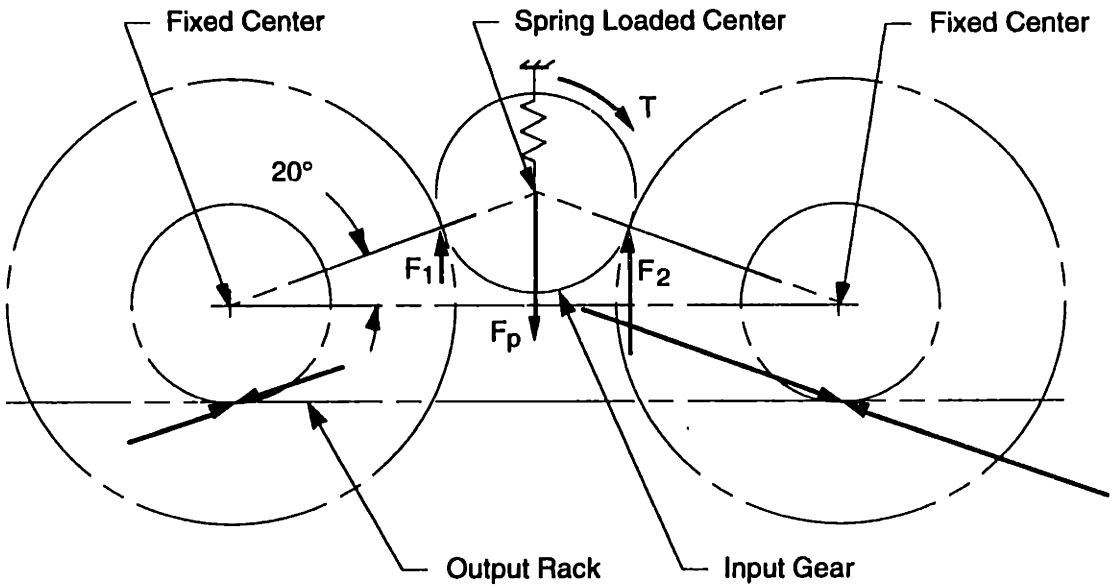


Figure 8-12 The sum of contact forces $F_1 + F_2$ balances the external preload force F_p. The difference between contact forces balances the drive torque T.

The diagram for axial motion preload is like Figure 8-12 except that the external preload force acts along the axis of the input shaft. Axial motion produces counter rotation of left and right handed helical gears in mesh with the input shaft. The easiest way to provide axial motion is by letting the bearings for the input shaft slide in their housing. Clearance in the bearings is usually not a problem, but the slide will friction lock if the helix angle is too small.

Angular motion preload requires a differential to divide the drive torque between two paths, as is so common in automobiles. Referring to Figure 8-8, a spring-loaded torque applied to Items 21 causes counter rotation of Items 20, thus removing backlash and generating preload in the loop. The input to the differential is a belt in this case, Item 18. The differential makes this method by far the most costly.

For any of these preload devices, the external preload force usually comes from a spring deflected through a measured distance. Its stiffness should be several times smaller than either path stiffness so that the loop stiffness is essentially all due to the preload spring. A frictional interface somewhere in the loop is a common method for setting the preload in the spring. This also solves the *phasing* problem of getting the last assembled gear to go into mesh. The problem is particularly significant to the radial preload method

because it controls the center location of the input gear. The first case study required a different solution to this problem.

The preload in the loop should be at least one half of the maximum expected load to prevent one path from becoming unloaded. If that happens, the drive stiffness reduces to the series sum of the loaded path and the preload spring. This relatively high preload seems to be a disadvantage because it increases friction and wear; however, a Type 1 design would require approximately twice this preload if used for bi-directional loads.

8.2.2.3 Case Study: A Robot Revolute Joint

The robot in this study has four degrees of freedom. Two identical links house the wrist and elbow joints, both with vertical axes of rotation. A third revolute joint at the shoulder gives the robot full planar motion. The shoulder mounts to a vertical slide to give the robot volumetric range. Only the joint located in each link was the subject of a re-design effort to decrease backlash and to double the gear ratio. The envelope of the link did not change.

Figure 8-13 shows the joint end of the robot link as it would appear from above with its cover removed. The bevel gear mesh has minimal backlash when properly set. The other gears are in the optimal configuration for radial motion preload. In this design, the servo motor, bevel gear assembly and input gear move as a unit on three parallel blade flexures. Two compression springs provide the external preloading force.

A difficulty arose because the envelope did not allow room for an angular adjustment interface. Without one, the input gear would not mesh properly with the two cluster gears, except by accident or by design. The correct angular relationship between gears would have to be built into the assembly. There are two ways to proceed but only one makes sense. Start by manufacturing each cluster gear with the same angular orientation between its large and small gears. One method is to lightly press and bond the large gear to the shaft of the small gear using a jig for angular alignment. Then it becomes a mathematics problem to find the location of the gear centers to make the teeth mesh properly.

Figure 8-14 shows the construction used to set up the mathematical model. Initially place the gears in a straight line with their centers crowded to remove all backlash. Since both cluster gears are identical, this configuration guarantees that all gears will mesh together. Rolling the cluster gear up the output gear, as shown in (a), causes the input gear to rotate in a clockwise direction through an angle determined by the angle A_2, the gear ratios R_1, R_2, and by the angular backlash B_2 present at the operating center C_2. In similar fashion, rolling the input gear back to the original centerline, as shown in (b), causes an additional clockwise rotation determined by A_1, R_1 and B_1 at C_1. Any half-tooth increment i of the input gear is also an acceptable configuration for meshing with the mirror image path, since it contributes exactly the opposite rotation, thus making two half increments whole. This construction gives Equation 8.8 for two unknowns A_1 and A_2. Equation 8.9 with the same unknowns comes directly from the geometry.

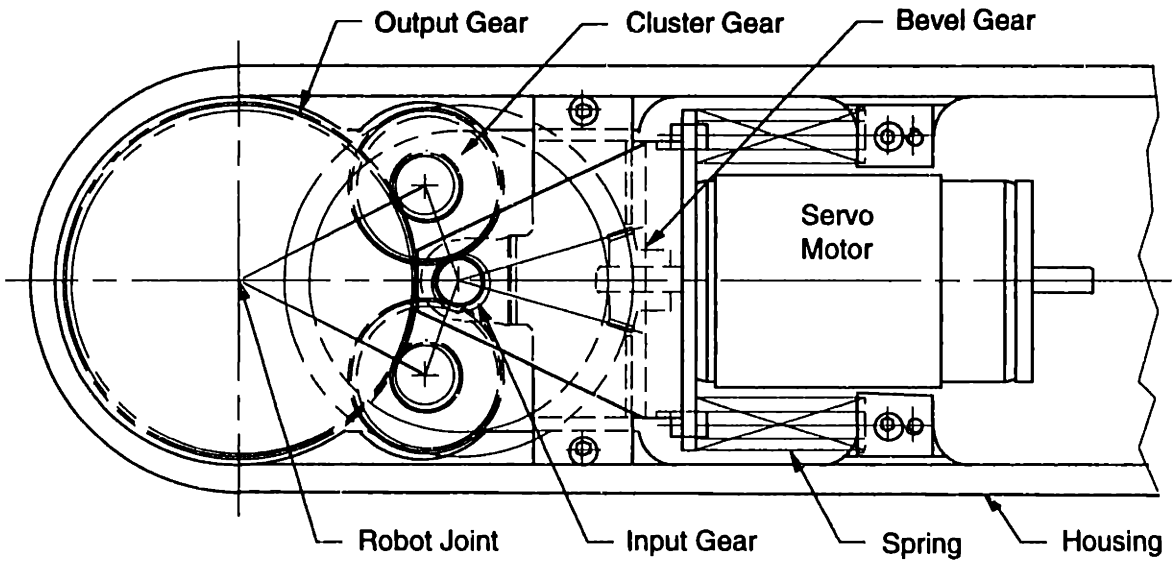


Figure 8-13 A Type 2 configuration can be compact. The servo motor and bevel gear assembly move on three parallel blade flexures to provide radial motion preload. The drawing is half size.

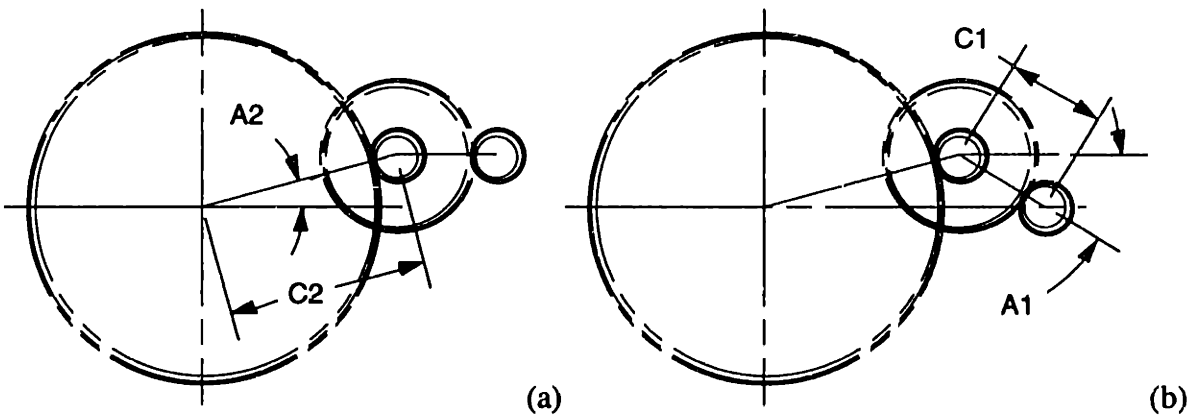


Figure 8-14 Construction of the mathematical model begins with the gears initially in a straight line. Moving the gears to the configuration in (a) causes clockwise rotation of the input gear. The configuration in (b) causes additional clockwise rotation. There are many possible configurations where the mirror image path would also mesh with the input gear.

$$\left[A_2 \cdot (R_2 + 1) - \frac{B_2}{2}\right] \cdot R_1 + A_1 \cdot (R_1 + 1) - \frac{B_1}{2} = \frac{180}{N} i \tag{8.8}$$

$$C_1 \sin A_1 = C_2 \sin A_2 \tag{8.9}$$

Initially, the number of half-tooth increments i is unknown. Choosing a nominal configuration such as $A_1 = 70°$, A_2 is solved from Equation 8.9 and then i is determined by solving Equation 8.8 and rounding to the nearest integer. Using that integer, a numerical solution of Equations 8.8 and 8.9 gives the proper geometry represented by A_1 and A_2. A spreadsheet with a built in solver is very convenient for manipulating various errors and tolerances that factor into the equations. From estimates of manufacturing tolerances, the calculated range of travel for the input gear is 0.012 inches, which corresponds to a change

in center distance of 0.004 inches. This change must be considered in the calculations to ensure that each mesh operates with backlash despite the tolerances.

8.2.2.4 Case Study: A Machine Tool C-Axis

A common option for a CNC turning center (or lathe) is the capability to use rotating tools in addition to the normal turning tools. Such a machine may operate either in a turning mode or in a milling mode. The machine could conceivably switch modes several times during a single piece part. While in the milling mode, the lathe spindle becomes the C-axis, which is equivalent to a rotary table on a milling machine. The C-axis must have very stiff and accurate angular position control. The lathe spindle requires only velocity control. The maximum speed required for each mode is vastly different, 15 rpm versus 4000 rpm. The C-axis in this study was to be an add-on option to the standard, two-range headstock. As a result, the primary constraint was the envelope. The machine controller constrained the gear ratio to be either 180:1 or 360:1 from the encoder to the spindle. Time and money constraints dictated the use of proven techniques and off-the-shelf equipment.

A few things were evident from the start. The C-axis would have to disengage from the spindle during turning mode. The spindle bull gear, used for low-range operation, was the best connection point to the spindle. Cincinnati Milacron, where this work was performed, had good success with *dual-pinion* designs for similar applications such as rotary tables, but no one had designed one that completely disengaged from mesh.[1] Previous designs used axial motion preload, but this was incompatible with the best way to disengage from the bull gear, which was axial motion of the dual pinions. After much thought, a new preload method emerged, which is recognized from Figure 8-15 as the optimal configuration for radial motion preload. The conceptual step from axial motion to radial motion was not smooth or deliberate. Rather it was a sudden inspiration that now seems obvious.

Two particularly interesting features of the design are worth explaining in detail. Referring to Figure 8-15, the extra length of the input gear provides the added compliance necessary to limit variations in preload. It is a simply supported beam spring designed to have zero slope at the mesh. Not shown in the figure are frictional interfaces in each cluster gear. With the C-axis engaged by hydraulic pressure, an adjustment to the interface causes the spring to deflect, which gives a measure of preload. Now the problem becomes how to engage and disengage a preloaded gear train. Fortunately the standard headstock used helical gears. Relative axial motion between the two pinions also causes relative rotation due to the helix angle. When sliding into or out of engagement, one pinion leads the other to provide backlash until both pinions come hard against travel stops. The amount of lead is

[1] Dual-pinion designs are usually Type 2 but some are Type 1 (see Figure 8-7), probably out of ignorance.

easy to calculate from estimates of backlash and other geometry. The method is very simple and works perfectly.

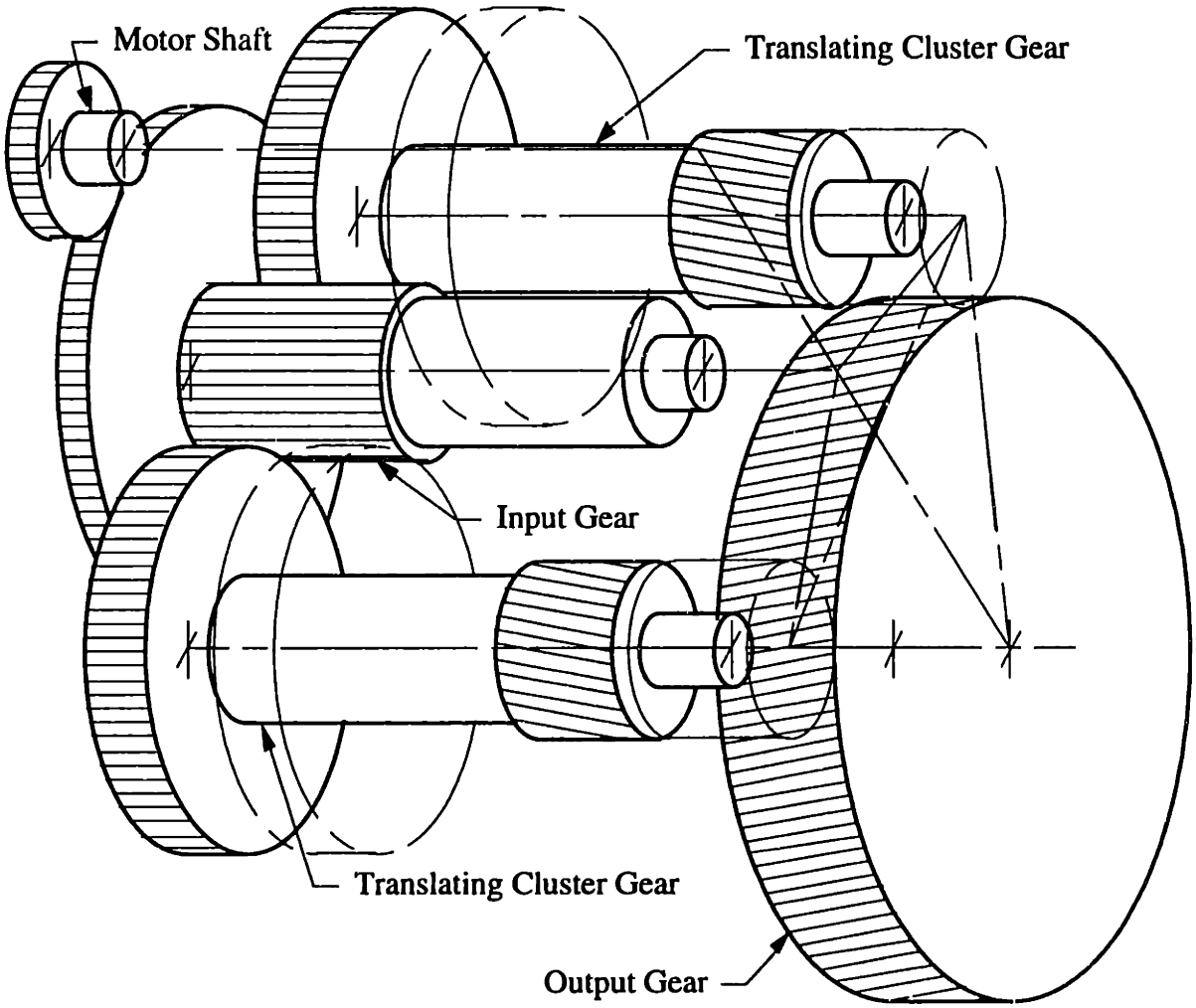


Figure 8-15 The first mesh at the motor end is not preloaded. The long pinion at the common input to the two preloaded paths serves as a beam spring for radial motion preload. The dual pinions engage the spindle bull gear by axial travel, one leading the other to release the preload during engagement.

8.3 Dual-Motor Drives

A very effective and flexible method for eliminating backlash uses two identical motors and transmissions in mesh with a common output gear. The two motors are controlled in one of several possible ways to accomplish the same task that the preload mechanism does in preloaded gear trains. Figure 8-16 shows two methods for controlling a dual-motor drive system [Kelling, 1974]. The first method uses unidirectional torque at each motor so that neither transmission experiences a reversing load (assuming of course that the deceleration is low enough). This method does not take full advantage of both motors but is suitable for position feedback located at either motor. The second control method simply applies an opposing bias signal to each motor. This method has regions where both motors work

together, but since their torques may reverse, the position feedback device should be at the output. Limit cycling is not a problem if at least one motor is driving through a stiff path.

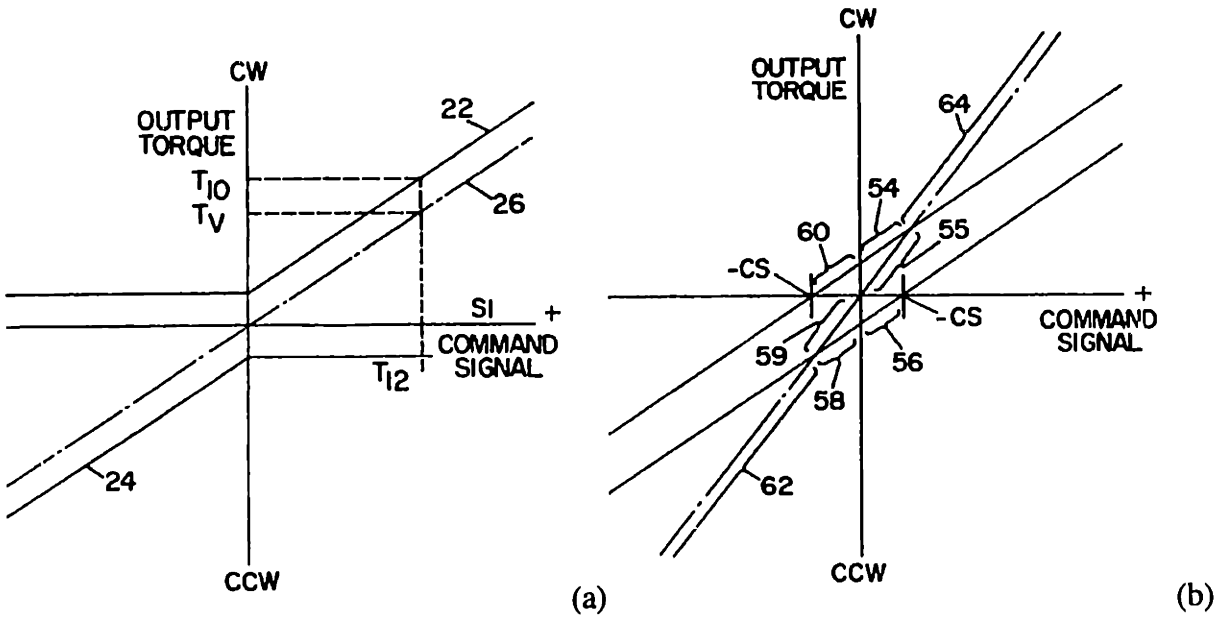


Figure 8-16 The coordination of two drive motors can be very simple. Motor torques (solid lines) add to give a net output torque (center line). In (a), each motor applies an opposite unidirectional torque. In (b), the motors have a constant bias torque between them (from U.S. Patent 3,833,847).

A third, more advanced method comes as an attempt to smooth out the transitions through backlash and the momentary reduction in gain that it causes. The bias signal does not have to be constant and the proper function will maintain constant total gain and provide smooth transitions through backlash. Equation 8.10 gives the function for each motor in terms of the bias function. The sum of these two functions is a linear function completely described by the gain k. The bias function proposed here is a piecewise cubic described in Equation 8.11. It satisfies the following conditions that are desirable for a smooth bias function.

- At $x = 0$, the first and second derivatives are zero.
- At the first transition point x_1, which is an inflection point, the curve is tangent to a line through the origin with slope $k/2$.
- At the second transition point x_2, the function and the first derivative are zero.

In addition, the second transition point x_2 should be chosen according to Equation 8.12 to ensure that the function for each motor never has a negative slope. Figure 8-17 shows these functions plotted for a gain $k = 1$.

$$f_1(x) = \frac{k}{2}x + f_b(x) \qquad f_2(x) = \frac{k}{2}x - f_b(x) \tag{8.10}$$

$$f_b(x) = \frac{k}{2}\begin{cases} \frac{1}{3}\left[2+\left(\frac{x}{x_1}\right)^3\right] & x \le x_1 \\ \left(\frac{x_2 - x}{x_2 - x_1}\right)^2 \frac{x(x_1+x_2)-2{x_1}^2}{x_2 - x_1} & x_1 \le x \le x_2 \\ 0 & x \ge x_2 \end{cases} \tag{8.11}$$

$$x_2 \ge \frac{2+3\sqrt{2}}{2} x_1 = 3.12132\, x_1 \tag{8.12}$$

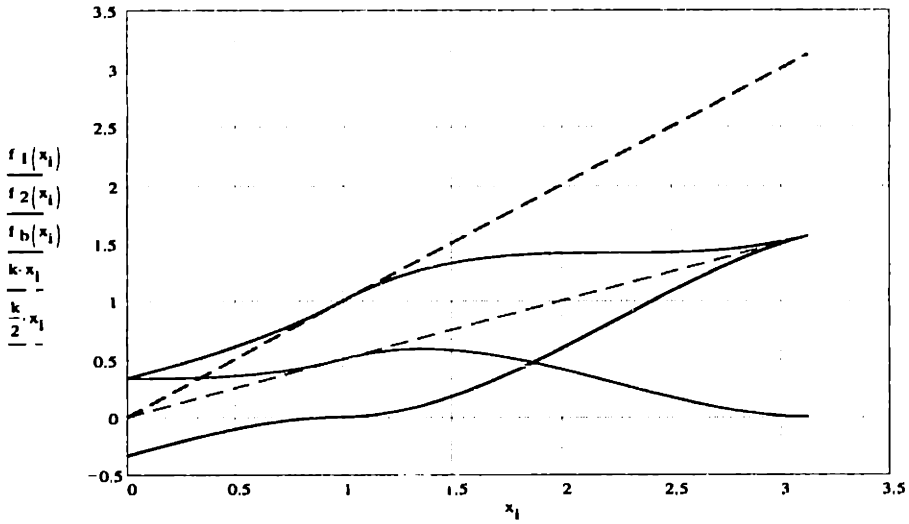


Figure 8-17 The motor curves f_1 and f_2 sum to give the line with unit slope. The line with one-half slope plus or minus the bias curve f_b gives f_1 or f_2, respectively. These curves are calculated for $k = 1$.

An important practical issue is where to apply the bias signal in the physical control system. We shall consider a constant bias first since this is the easiest and most likely choice. Figure 8-18 shows a dual motor drive each with a typical control system consisting of a current loop within a velocity loop within a position loop. The position command signal and the encoder feedback are common to both motor controllers. A constant bias signal could be injected just about anywhere in the system provided that it does not become integrated, causing the amplifier to saturate. This should not be a problem if the bias comes in after the integrator or if the integrated loop can zero out its error signal. For example, an integrator in the current loop (which is typical in most amplifiers) will cause the current to rise just enough to balance the bias. An integrator in the velocity loop is a problem because both motors, being geared together, cannot respond to a velocity bias. In addition, any

difference in tachometer gain between the two motors looks like a variable velocity bias.[I] The safest and most convenient approach is to work within the current loop using the balance adjustments built in to the typical amplifier.

The other two methods require nonlinear functions of current output for a given commanded current signal. This limits the location for intervention to just ahead of the summing junction of each current loop. The signal generated by each velocity loop should be approximately equal to each other and proportional to the total torque required at the output, provided the tachometer gains are balanced and the transmissions are of sufficient quality. The unidirectional method would be simple to implement using a diode circuit. The variable bias method requires a digital processor to generate the piecewise cubic function. It is not unreasonable to build this capability into modern digital drives that have processors.

It is critical that the transmissions are free to back drive for stability reasons. Some transmissions such as high-ratio worm gears will not back drive well enough or not at all. The situation arises where the motor responsible for resisting the motion must instead drive the transmission to allow motion. This is very unstable much like sliding down stairs.

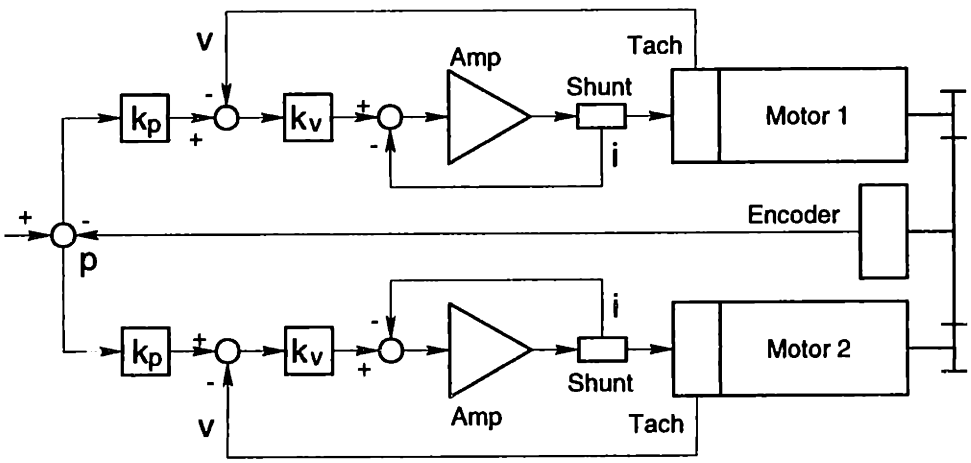


Figure 8-18 The two transmission paths are preloaded through the control system rather than a mechanical preload device. A disadvantage is the additional motor heating to generate the preload.

8.4 Commercial Differential Drives

Several commercial sources are available for compact, high-ratio transmissions based on the principle of differential motion. The so called *perpetual wedge*, invented by James White over two hundred years ago, provides the basis for the differential drives used today [White, 1989].[II] Using involute gears arranged differentially, the device outputs the

---

[I] It would be best to integrate the sum of velocity errors and send this signal to both loops.

[II] In an interesting coincidence of names, G. White writes about James White's early work with differentials published in *A New Century of Inventions* (being designs and descriptions of one hundred machines, relating to arts, manufacturers and domestic life), 1822 by J. White. The paper by G. White provides a theoretically sound and practical discussion of differential epicyclic gears and their applications.

difference between two slightly different gear meshes operating at the speed of the input shaft. Since the gears also operate under output-sized loads, losses can be overwhelming. A transmission ratio greater than 50:1 is easy to achieve but the efficiency is typically below 50% making a single-stage device impossible to back drive. This may be useful for the device in Figure 8-19 but it is not very practical for power transmissions or servo mechanisms. The differential drives available commercially, namely, epicyclic drives, harmonic drives and cycloidal drives, achieve higher efficiency through clever ways to reduce friction. Backlash control for these devices usually comes through control of tolerances and selective fitting.

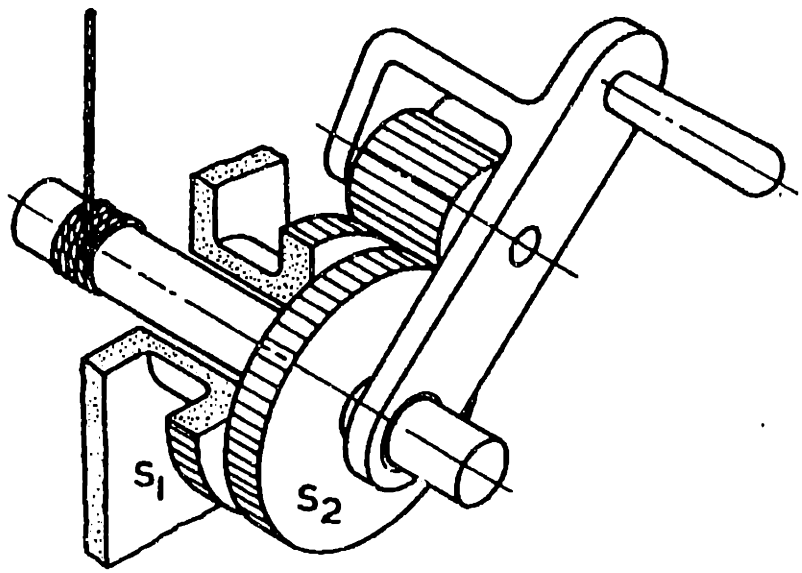


Figure 8-19 Gears S_1 and S_2 have different numbers of teeth but engage the same pinion gear. One orbit of the pinion causes S_2 to index by the difference in the number of teeth, S_1 being fixed. The gear teeth appear to work like wedges, hence the name, but those familiar with involute action will prefer the name perpetual lever (reproduced from White).

As part of the accuracy enhancement project with Cincinnati Milacron, three commercial differential drives were evaluated for use on machine tools, particularly for rotary axes such as the A- or B-axis on a machining center. All three drives have sufficient transmission ratio and are reasonably compact to fit well in the overall machine design. In addition, they are available off the shelf with zero backlash, which avoids the complication of preloaded gear trains or dual motor drives. Each drive uses different technology and the best is not nearly as clear for rotary axes as ball screws are for linear axes of moderate length. Although vendor catalogs provide useful design and performance information, first hand experience gained from comparative testing seemed prudent since these devices are uncommon in the machine tool industry.

We developed a bench-top test apparatus that provided position control for the input shaft and torque control for the output shaft. Small angular motion of the output shaft was measured with an LVDT (linear variable differential transformer) sensing a long lever arm.

Torque at the input shaft was measured indirectly from current to the servo motor. Stiffness, hysteresis and efficiency in both forward and backward torque directions (with respect to motion) were the primary tests performed. Transmission errors were not directly measured since a positioning axis should use output position feedback. Stiffness was measured by holding the input shaft stationary and applying a variable torque to the output shaft. A plot of the output-shaft deflection as a function of the output-shaft torque shows any backlash and frictional hysteresis present in the speed reducer. The slope of the curve is the compliance (the inverse of stiffness). Hysteresis is measured by approaching the same commanded point from opposite directions. This may be extended to a trajectory of points by moving the input shaft through a small sinusoidal angle and measuring the resulting angle at the output shaft. A plot of the output angle versus the input angle will clearly show a hysteresis loop due solely to the speed reducer. In the normal mode of operation, efficiency is the power out at the output shaft divided by the power in at the input shaft. This reverses for the back-drive mode so that efficiency is the power out at the input shaft divided by the power in at the output shaft. Efficiency becomes a meaningless notion for a transmission that will not back drive since it consumes power at both ends.

8.4.1 The Cycloidal Drive

The cycloidal drive is commercially available in a wide range of sizes.[1] It features a high reduction ratio in a single compact stage, smooth motion, zero backlash and a reasonable efficiency. These benefits are the result of pure rolling contact between a hypocycloidal cam and a number of rollers as shown in Figure 8-20. The components (cam, rollers, housing, etc.) require precision manufacturing to achieve good load sharing among the rollers. Given this precision, it is possible to eliminate backlash using an interference fit between the cam and the rollers.

The motion of the hypocycloidal cam within a fixed set of rollers is identical to the motion of the inner pitch circle rolling against the stationary pitch circle without slipping. Since the cam has one lobe less than the number of rollers, one orbit causes a rotation of the cam by one lobe relative to the rollers. The cycloidal drive achieves a cascade reduction effect by using a double cam that engages a fixed set of rollers on one end and a second set of rollers on the other end. The second set connects to the output shaft and has a different number from the fixed set. The input shaft causes the dual cam to orbit, which drives the output rollers relative to the fixed rollers. The transmission ratio is $(N_{fixed} - 1) \cdot N_{output}$ for the usual case where the difference between sets of rollers is only one. Figure 8-21 shows the lever analogy of a cycloidal drive and the high transmission ratio that is possible using a differential drive. The cycloidal drive has the identical differential motion as White's perpetual wedge but improves upon the efficiency by using rolling elements.

---

[1] Dojen™ by Mectrol Corp., 9 Northwestern Dr., Salem, NH 03079, (603) 890-1515.

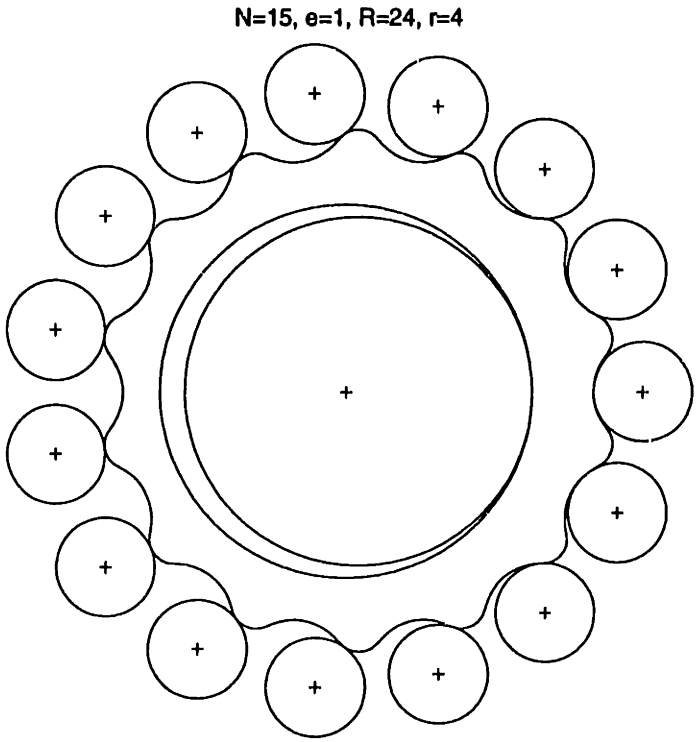


Figure 8-20 The point where the two pitch circles contact is an instantaneous center meaning that all points on the cam instantaneously rotate about this center. In addition, each surface in contact with a roller is moving tangentially about the instant center. The normal force at each contact acts through the instant center. That point on the cam is instantaneously constrained against translations; only rotation is allowed.

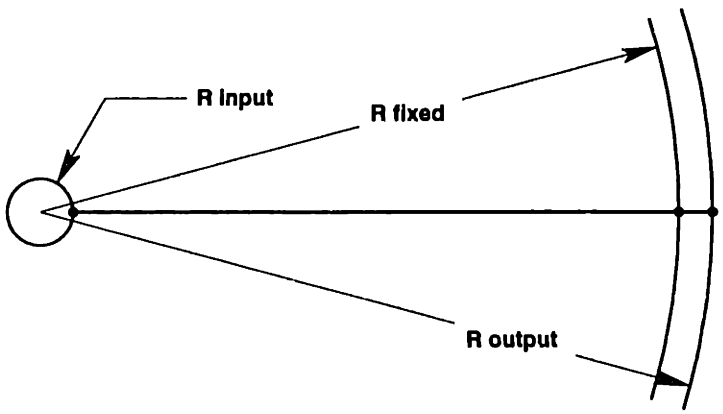


Figure 8-21 The cam of the cycloidal drive acts like a lever with an instantaneous pivot at R_{fixed}. The center of the dual cam makes an orbit given by R_{input} while it rolls simultaneously against the pitch circles R_{fixed} and R_{output}.

8.4.2 The Harmonic Drive

The harmonic drive is also commercially available in a wide range of sizes.[I] The differential motion occurs between a circular spline (or internal gear) and a flex-spline (thin-walled external gear) shown in Figure 8-22. Usually the circular spline is stationary and the flex-spline rotates with the output shaft, but the reverse could also be true. The flex-spline has usually two fewer teeth than the circular spline. The two splines will mesh properly only when the flex-spline is forced into an elliptical shape by an elliptical ball bearing device called a wave generator. The wave generator rotates with the input shaft and forces the elliptical shape of the flex-spline to rotate at the same rate, although the flex-spline rotates at a much slower rate. The resulting velocity ratio is $N_{output}/(N_{output} - N_{fixed})/2$ for a two-lobed wave generator. Since the flex-spline is nearly as large as the circular spline and because it flexes in a way that reduces the contact ratio, an unusually efficient rolling action occurs in the mesh. In addition, power crosses only one mesh, which is uncommon for a differential drive. Backlash is reduced or eliminated by controlling the fit in the mesh.

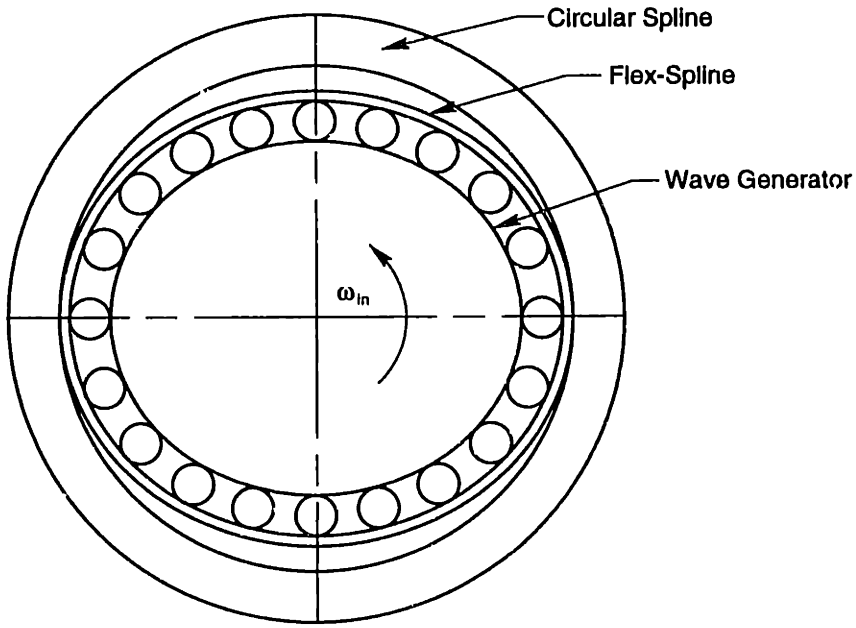


Figure 8-22 The three basic components of a harmonic drive are the circular spline, the flex-spline and the wave generator.

8.4.3 The Epicyclic Drive

The epicyclic drive that we tested is commercially available through a Russian manufacturer, however sizes are very limited.[II] Since it is based on conventional spur gear technology and the basic invention is over two hundred years old, nothing fundamental

---

[I] H D Systems, Inc., 89 Cabot Court, Happauge, NY 11788, (516) 231-6630.

[II] Mansis Corp., 519 South Austin St., Seattle, WA 98108, (206) 768-1970.

stands in the way of producing any size by any gear manufacturer chosen. This particular design may be unique since it combines a differential with a conventional planetary as Figure 8-23 indicates. The planet carrier on White's perpetual wedge is driven by the input shaft. This design achieves additional gear ratio by driving the planet gears directly with a sun gear attached to the input shaft. The gear ratio provided by the planetary reduces the ratio typically needed in the differential thereby reducing the total drive losses. It is also possible to eliminate the planet carrier and bearings since the planet gears are naturally self supporting, as we demonstrated by modifying the unit tested. Backlash is reduced or eliminated by controlling the fit in all meshes.

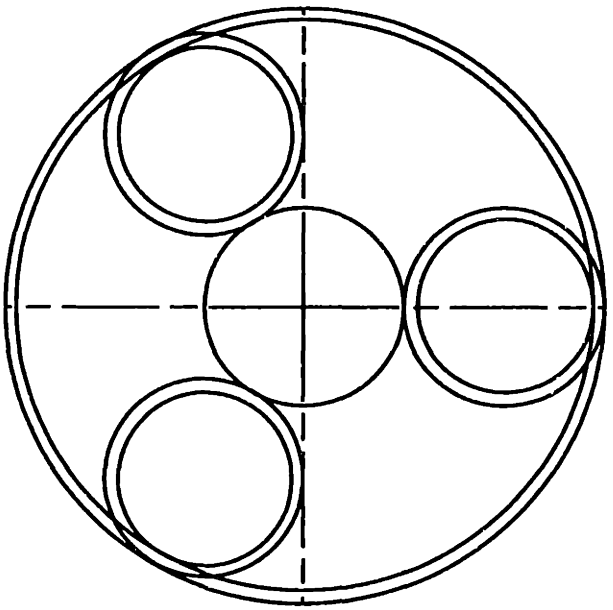


Figure 8-23 The circles represent the operating pitch diameters of the various gears. The planet gears have two operating pitch diameters because they mesh simultaneously with two different sized internal gears. One internal gear is fixed and the other is the output shaft. The sun gear is the input shaft.

8.4.4 Experimental Results

The three commercial speed reducers were tested for stiffness, hysteresis and efficiency, which are discussed separately. The epicyclic drive was physically much smaller and weaker than the others, which made testing particular difficult. Twice we stripped gear teeth from the ring gear. This forced us to operate at much lower torque levels, which caused resolution problems with the efficiency tests. Some peculiar results in the stiffness test may be the result of damaged parts that were not replaced along with the ring gear.

Of the three speed reducers tested, the harmonic drive has the greatest stiffness and the most linear deflection curve. The measured stiffness correlates well with the mid-range stiffness K_2 = 1060 in-lb/mr published by HDS.[1] The stiffness of the cycloidal drive,

[1] HDS series R, model HDUC-2AG-R, size 40, 160:1 ratio.

which is physically larger, is about one-half that of the harmonic drive when loaded and even less when lightly loaded. The published stiffness of 600 in-lb/mr is consistent with the loaded case. The low-load stiffness and the considerable hysteresis are not consistent with the published information.[1] The stiffness of the epicyclic drive is much lower, but in fairness we should compensate for its small size. The ring gear in the epicyclic is about one-half that of the harmonic drive. Using a cubic scaling function, a comparable sized epicyclic drive would have eight times the stiffness, which is nearly as stiff as the harmonic drive. Oddly, the stiffness in one direction was so low that the deflection measurements went off scale. We have no logical explanation why this should happen, and there is no published information on the epicyclic drive to compare these results.

Figure 8-24 shows the results of the positional hysteresis test, where the output shaft motion is plotted versus the input shaft motion divided by the velocity ratio. An ideal speed reducer would plot on a straight line having a slope of one. The hysteresis plot shows the combined effects of friction and compliance. If either were zero, the hysteresis would also be zero. The harmonic drive has the lowest hysteresis followed by the cycloidal drive. This is consistent with the stiffness tests since the harmonic drive is both stiff and low in friction. The epicyclic drive shows relatively high hysteresis due to its higher compliance and friction. The trend for lower hysteresis at higher loads is likely due to higher stiffness with load rather than lower friction. Hysteresis directly affects the positioning accuracy when the position feedback device is located at the input shaft since the drive becomes part of the metrology loop. Locating the position feedback at the output shaft places the hysteresis within the feedback loop where it is much less significant.

The harmonic drive has consistently higher efficiency than the cycloidal drive or the epicyclic drive. This is consistent with the stiffness and hysteresis tests, which indicated lower friction in the harmonic drive. The efficiency depends on speed and torque with the highest efficiency occurring at lower speed and higher torque. The best efficiency measured for the harmonic drive varied between 75 and 85 percent in the forward direction, and between 65 and 75 percent in the back-drive mode. The best efficiency measured for the cycloidal drive varied between 60 and 70 percent in the forward direction, and between 30 and 40 percent in the back-drive mode. The epicyclic drive would not back drive for the limited torque that it could withstand. The best efficiency measured was between 45 and 50 percent. This is not unexpected since the differential stage has a ratio of 25:1 (the planetary stage has a 4:1 ratio). If there were no losses, then the power passing through the differential stage would be 25 times the input power. A loss of a couple percent from 25 times the input power is a 50% loss.

---

[1] Dojen™ size 05, 168:1 ratio.

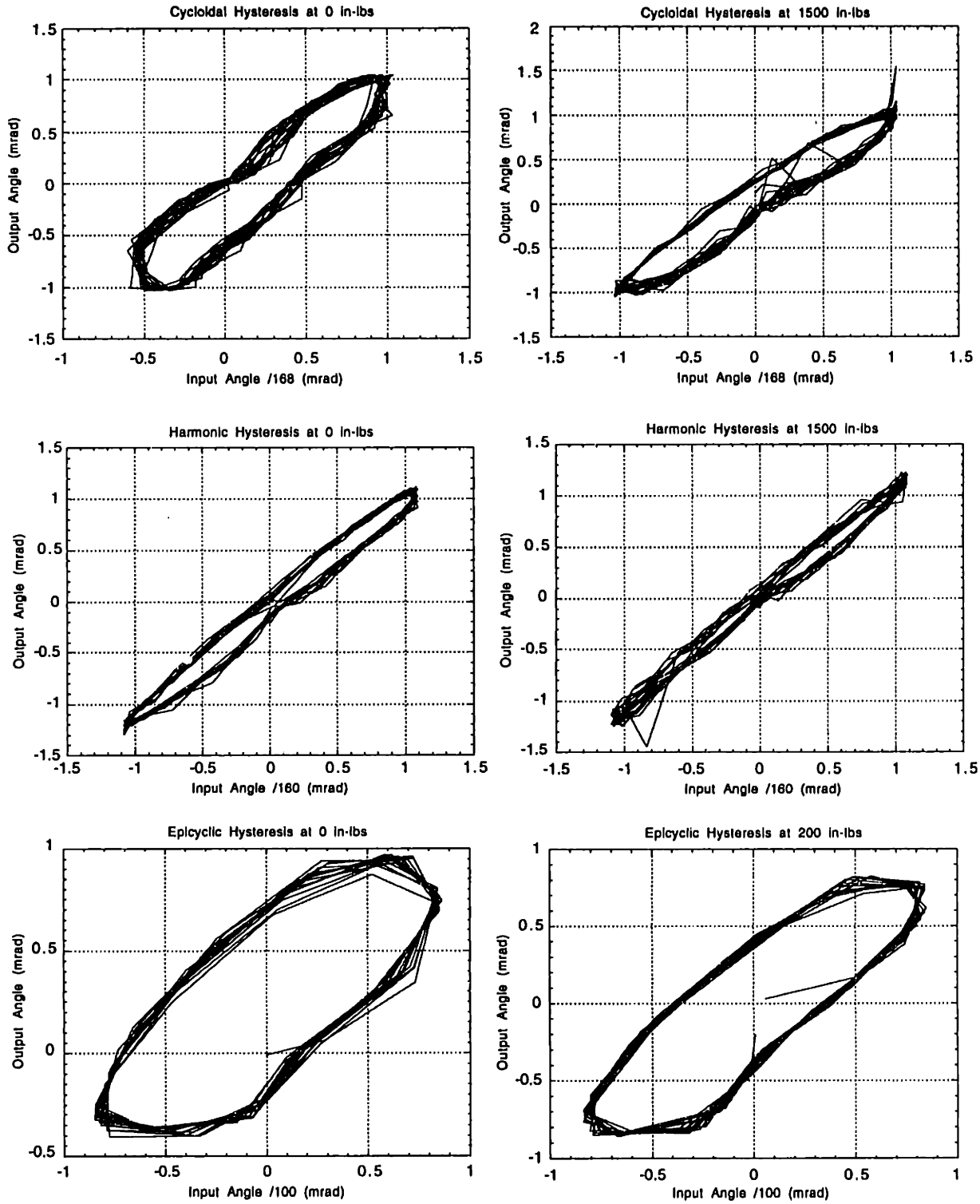


Figure 8-24 Hysteresis plots for zero torque (left) and maximum torque (right) each for cycloidal, harmonic and epicyclic speed reducers (top to bottom). The output shaft angle was measured and plotted as a function of the input shaft angle (divided by the gear ratio).

8.4.5 Conclusions

The three commercial differential drives each provide a high reduction ratio in a compact, concentric package. Each demonstrated zero backlash, although the epicyclic drive behaved suspiciously as if it were damaged or poorly made. The harmonic drive had the highest stiffness, the highest forward and backward efficiencies and the lowest positional hysteresis. The cycloidal drive is the least compact for a given stiffness or load capacity. The epicyclic drive, when corrections are made for its smaller size, is nearly as stiff as the harmonic drive. The units tested are physically too small to directly drive a machine tool rotary axis. The largest available cycloidal and harmonic drives border on being too small for the 500 mm class machining center under consideration. The remaining discussion centers on how to extrapolate each technology to a large enough size.

Choosing a target torsional stiffness involves a certain amount of judgment but framing the problem in more familiar terms is helpful. For example, the total loop stiffness for a 500 mm class machining center is on the order of 0.4 lb/μin. At a 10 inch radius, this is 40 in-lb/μr or 4.5 N-m/μr. The stiffness of the transmission alone should be four to five times greater. Taking the largest payload, a transmission stiffness of 25 N-m/μr gives a locked-rotor natural frequency of 100 Hz. Several other estimates based on cutting requirements also lead to a stiffness goal of 20 to 25 N-m/μr. The goal for the rotary axis speed, determined by making the tangential speed at the edge of the pallet equivalent to the linear axis speed, is 25 rpm at the axis or 3000 rpm at the input assuming a 120:1 ratio.

The largest cycloidal drive in the published data is the Size 10 with a housing diameter of 381 mm including the support bearing for the output shaft. The circle inscribed by the rollers is the appropriate size measure for scaling purposes, which is approximately 216 mm for the Size 10. A cubic fit of the published data agrees well for both stiffness and torque capacity. An inverse relationship agrees well for the input speed. Extrapolating to a 400 mm inscribed circle gives a stiffness of only 2.8 N-m/μr, which is an order of magnitude too low. The output torque capacity is more than adequate at 10,000 N-m, but the input speed is low at 1400 rpm. The speed of the rollers is approximately the same as the input shaft, thus it should be possible to satisfy the speed requirement. The stiffness requirement is more difficult to satisfy but may be possible since there is an apparent design flaw in the Dojen™ brand of cycloidal drives. The forces transferred by the cam between sets of rollers produce a significant couple on the cam that is unsupported within the clearances between the sides of the cam and the housing. The movement of the cam and the associated friction would explain the low stiffness and the high hysteresis apparent in the

deflection plots. A significant improvement in stiffness may be possible simply by providing a bearing system to support the cam against rocking.[1]

The largest harmonic drive in the published data is the Size 100 with a housing diameter of 330 mm including a support bearing for the output shaft. The pitch diameter is the appropriate size measure for scaling purposes, which is approximately 260 mm for the Size 100. A cubic fit of the published data agrees well for both stiffness and torque capacity. An inverse relationship agrees well for the input speed. Extrapolating to a 400 mm pitch diameter gives a stiffness of 12 N-m/μr, which is a factor of 2 lower than the goal. The output torque capacity is more than adequate at 20,000 N-m, but the input speed is low at 1700 rpm. The DN value of the wave generator and the pitch-line velocity are fairly high at 680,000 mm-rpm and 7000 ft/min, which leaves little room for improvement. It may be possible to meet the stiffness goal by using a three-lobed wave generator and by carefully optimizing the flex-spline. The engagement of the flex-spline and circular spline at three equally spaced points provides full radial constraint, which may eliminate the need for a radial bearing depending whether the radial run out is acceptable.

The epicyclic speed reducer has the potential to be the stiffest of the three commercial speed reducers because the main load path, from the fixed internal gear to the output gear, is short and through three gear meshes in parallel. The planet gears are self-supporting, which makes it possible to eliminate the planet carrier and support bearings. The fundamental problem is frictional losses in the main load path, which usually will be the same order as the useful power out. Speed is not a problem because the pitch line velocity scales with the size of the input pinion rather than the internal gear. In terms of manufacturing, the epicyclic drive is most conventional, consisting of internal and external spur gears. The circular spline and the flex-spline of the harmonic drive are fairly conventional parts except that a special tooth form is now being used by HDS. The elliptical wave generator requires precision cam grinding machinery. The hypocycloidal cam of the cycloidal drive also requires precision cam grinding machinery, and the housing requires precision boring to locate the cam followers.

8.5 The NIF Precision Linear Actuator

The NIF requires over 3200 ultra-precision linear actuators for large-aperture, tip-tilt mirror mounts. One design constraint is that it be powered by a conventional stepper motor without position feedback on the actuator. This allows the possibility of multiplexing several actuators to one controller to minimize the overall cost of the system. Diagnostics on the laser beams provide the feedback to converge to the proper alignment. The specification for the actuator's step resolution is 25 nanometers or approximately one

[1] It is theoretically possible to locate the eccentric driver in a location that exactly balances the couple, but the length of the speed reducer would become impractical.

microinch over a range of 10 mm. The specification for incremental accuracy is ±3% of the distance traveled for moves greater than or equal to 100 steps or 2.5 μm. For shorter moves, the incremental accuracy is ±3 steps or ±75 nanometers non accumulating. This assumes unidirectional operation with a constant load up to the maximum gravity load of 250 newtons. The specification for hysteresis error, which includes any backlash, is 6 steps or 150 nanometers. Hysteresis error encompasses bi-directional repeatability (the physical difference between two moves to the same commanded point from opposite directions) and reversal error (the command distance required to generate a physical movement).

A search for commercial actuators yielded only one that was close to meeting the stated requirements, but it was too expensive. This led to the decision to design a custom actuator that uses components that are either commercially available or manufacturable with conventional equipment. Then several vendors could bid to produce the design to print. Figure 8-25 shows the current NIF actuator design. It has a 400 step/rev motor, 100:1 ratio speed reducer and 8 mm diameter by 1 mm lead ball screw to provide 25 nm/step resolution. The ball-screw manufacturer customizes the length and end detail of an otherwise standard screw to satisfy this design. This cost turns out to be rather insignificant for quantities greater than about ten units. Other notable components are the 25° angular-contact, duplex thrust bearings and the ball-screw flexure. A description of the ball-screw flexure appears in Chapter 6. This actuator design depends on the connection to the mirror mount to prevent rotation of the ball nut via the flexures.

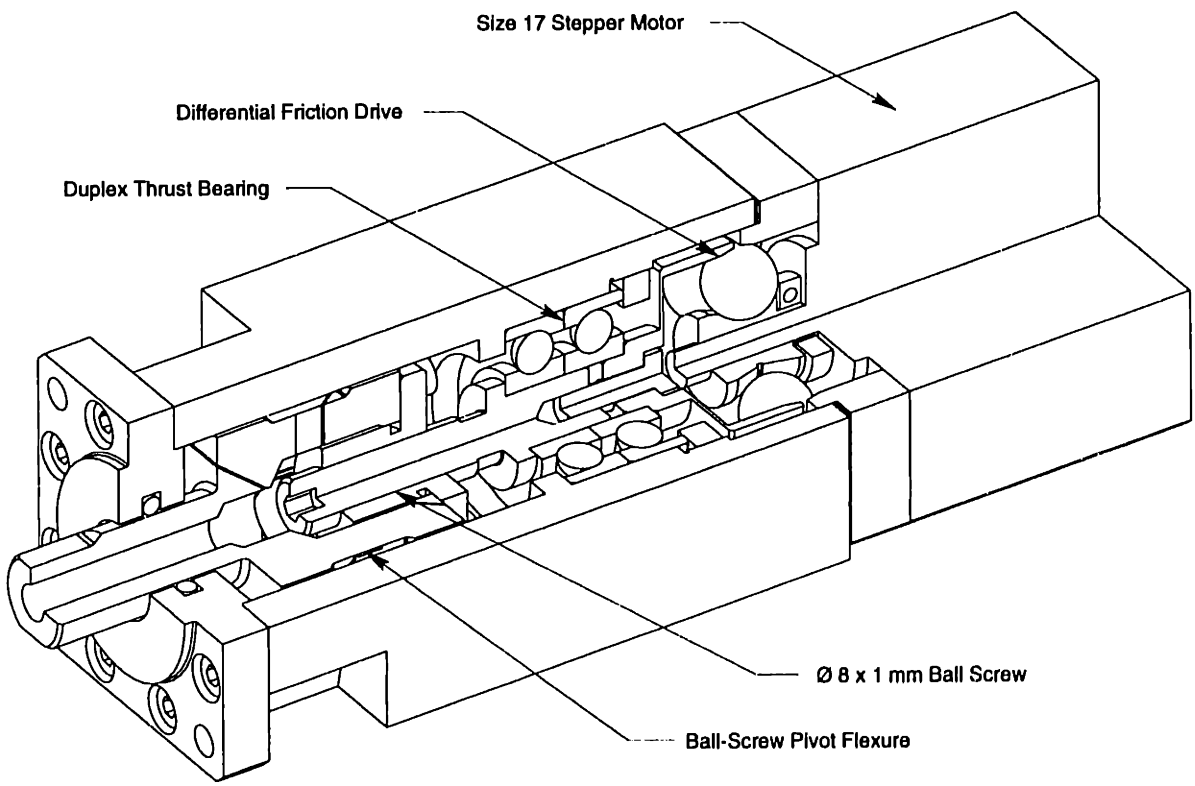


Figure 8-25 Cut-away isometric view of the NIF ultra-precision linear actuator.

A suitable speed reducer is the harmonic drive but the desire to reduce cost led to the investigation of the differential friction drive that appears in the design. The cost of the friction drive is approximately one-half the cost of a harmonic-drive component set. In addition, the friction drive acts as a slip clutch to protect the ball screw from overload if it drives against a hard stop. A disadvantage is the variation in velocity ratio that depends on manufacturing tolerances and the load on the actuator. The next section elaborates on these peculiarities of the differential friction drive. A combination of larger variations than expected and a new understanding of the control strategy caused us to change plans and use a harmonic drive as the speed reducer. However, the test results presented in a later section are for the actuator with a differential friction drive.

8.5.1 The Differential Friction Drive

The differential friction drive functions exactly like the epicyclic drive described in Section 8.4.3 except that torque is transmitted through traction between balls and races rather than through gears. Nevertheless, a patent was granted for such a device to provide coarse/fine adjustment of a dial [Everman, 1995]. Figure 8-26 shows a cross section through the device. It is an inverted version of the differential friction drive used for NIF. Figure 8-27 shows a simplified cross section of the key parts and the parameters required to compute the velocity ratio. In principle, any relatively large velocity ratio is achievable but the ratio has a tolerance associated primarily with the tolerance on the race angles. Consider an instructional case where the angles are all equal. Then the device operates as a four-point contact bearing where the balls spin about a horizontal axis and no relative motion occurs between the fixed race and the output race. This case has an infinite velocity ratio. Changing the angle of any one race changes the circumference traced out by the balls. This difference results in motion of the output race. The angles used to obtain the 100:1 ratio for NIF are 19.275° for θ_1 and 15° for the rest. A rather surprising result is that neither the ball diameter d nor the orbit diameter D affect the ratio when the fixed race and the output race have identical angles.

A rather extensive Mathcad™ program was developed during the design of the NIF actuator to compute such things as the velocity ratio and contact stresses. The more significant aspects are reproduced here. It contains the equations and explanations required to make new designs. It also provides a real example that shows the torque limitation and resulting slip that is characteristic of friction drives. In particular, the program underestimates the amount of slip by a factor of approximately four compared to test results. I suspect that the discrepancy is due to a component of spin between the balls and the races that is not accounted for in the theory. The main advantages of friction drives are zero backlash, low hysteresis and very smooth motion due to the precision with which the components can be made. Although the differential friction drive will not be used for NIF, I believe there are many applications that would benefit from its use.

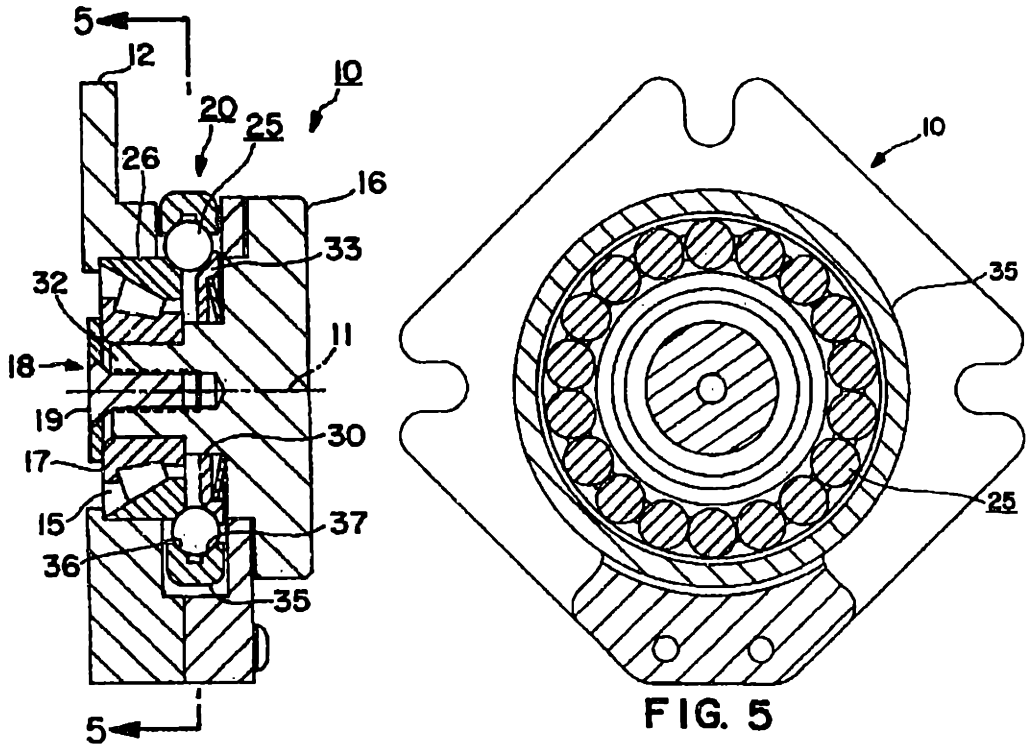


Figure 8-26 Coarse/fine adjustment mechanism featuring a differential friction drive (from U.S. patent number 5,435,651). Coarse adjustment may be applied to item 16, which causes sliding between the balls and races. Fine adjustment may be applied to item 35, which drives item 16 differentially through friction.

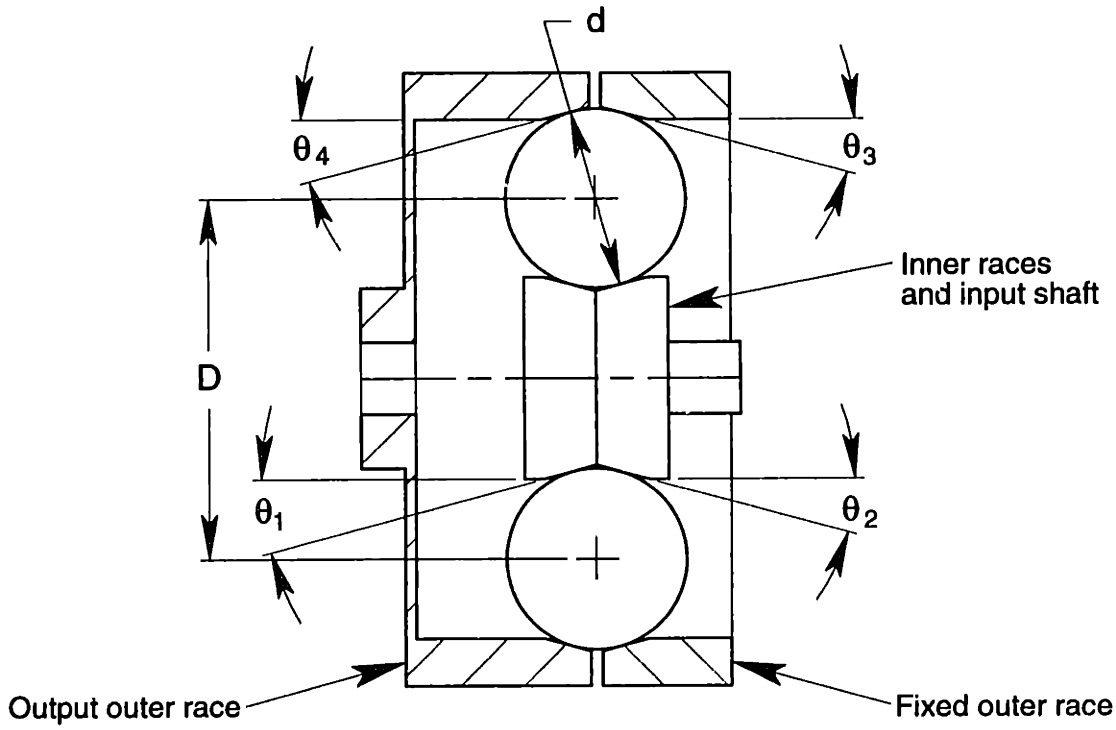


Figure 8-27 The differential friction drive has four races and at least three balls. The input shaft connects to and drives two races. Differential action occurs between the fixed race and the output race. These parameters are sufficient to compute the velocity ratio. The ratio is infinite if all the angles are identical.

8.5.1.1 Analysis Program for the Differential Friction Drive

This program performs design analysis for the differential friction drive shown below. The drive consist of three or more balls (equivalent to planet gears) and four tapered races. Two inner races turn with the input shaft (equivalent to a sun gear), one outer race is fixed (right), and the other turns the output shaft (left). The races are identified 1, 2, 3 and 4 as indicated by the race angles (see the Figure 8.27 above). Entered as components in the vector θ, they are the key parameters that determine the velocity ratio. In fact, the velocity ratio has no dependence on the ball diameter **d** and the orbit diameter **D** when the outer race angles are equal. In addition to the parameters mentioned, enter the input angular velocity ω, the desired velocity ratio **R**, the number of balls **n**, and the axial preload $\mathbf{f_a}$. The program will change one or more race angles to achieve the desired ratio. Also enter the elastic properties for the balls and races.

$$\theta := \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \cdot 15 \cdot deg$$

$d := 10 \cdot mm$ $\quad D := 20.1 \cdot mm$ $\quad \omega_{in} := 1000 \cdot rpm$

$R := 100$ $\quad n := 6$ $\quad f_a := 10 \cdot lbf$

$E_b := 30 \cdot 10^6 \cdot psi$ $\quad \nu_b := 0.29$

$E_r := 30 \cdot 10^6 \cdot psi$ $\quad \nu_r := 0.29$

Define the transformation from race angles θ to angles α from the ball center to each contact point. Also define the diameter where the balls contact the race.

$$\alpha(\theta) := \left[\left[\begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} \cdot \theta\right] + \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix}\right] \cdot 90 \cdot deg \qquad d_c(\alpha) := D + d \cdot \sin(\alpha)$$

The angular velocity for each race is related linearly to three parameters that define the motion of the balls. Since races 1 and 2 are constrained to turn with the input shaft and race 3 cannot turn, the inverse relationship provides the solution to the three parameters, axial and radial components of the ball's spin vector and the orbit angular velocity.

Define the angular velocity vector ψ, where the first two components are axial and radial spin and the third is the orbit velocity. Also define the output shaft angular velocity (race 4).

$$\psi(\alpha) := \begin{bmatrix} d \cdot \sin(\alpha_1) & d \cdot \cos(\alpha_1) & D \\ d \cdot \sin(\alpha_2) & d \cdot \cos(\alpha_2) & D \\ d \cdot \sin(\alpha_3) & d \cdot \cos(\alpha_3) & D \end{bmatrix}^{-1} \cdot \begin{bmatrix} d_c(\alpha_1) \\ d_c(\alpha_2) \\ 0 \cdot d \end{bmatrix} \cdot \omega_{in}$$

$$\omega_o(\alpha) := \frac{\left[\begin{bmatrix} d \cdot \sin(\alpha_4) & d \cdot \cos(\alpha_4) & D \end{bmatrix} \cdot \psi(\alpha)\right]_1}{d_c(\alpha_4)}$$

Solve for the change in race angle(s) to achieve the correct output angular velocity as defined by the velocity ratio. The change may be made to any single race or combination of races (but be careful of signs). Also calculate the new race angles, the output angular velocity and the ball's angular velocities.

Initial value

$\Delta\theta := 0 \cdot deg$

$\Delta\theta := root\left[R \cdot \omega_o\left[\alpha\left[\theta + \begin{bmatrix}\Delta\theta\\0\\0\\0\end{bmatrix}\right]\right] - \omega_{in}, \Delta\theta\right]$

$\theta := \theta + \begin{bmatrix}\Delta\theta\\0\\0\\0\end{bmatrix}$

$\Delta\theta = 4.275 \cdot deg$

$\theta = \begin{bmatrix}19.275\\15\\15\\15\end{bmatrix} \cdot deg$

$\alpha := \alpha(\theta)$

$\omega_o(\alpha) = 10 \cdot rpm$

$\psi(\alpha) = \begin{bmatrix}-540.453\\-57.49\\267.123\end{bmatrix} \cdot rpm$

For each race, calculate the diametral taper, contact diameter and the space between balls.

$\overrightarrow{taper := 2 \cdot \tan(\theta)}$

$space := D \cdot \sin\left(\frac{\pi}{n}\right) - d$

$taper = \begin{bmatrix}69.94\\53.59\\53.59\\53.59\end{bmatrix} \%$

$\overrightarrow{d_c(\alpha)} = \begin{bmatrix}10.661\\10.441\\29.759\\29.759\end{bmatrix} mm$

$space = 0.05 \cdot mm$

Calculate the sensitivity of the output angular velocity to a variation in each angle. Note, for the case where the outer races have equal angles, there is no sensitivity to variations in **d** or **D**.

$i := 1..4 \qquad \delta\alpha := identity(4) \cdot 10^{-6}$

$S_i := \frac{\omega_o\left(\alpha + \delta\alpha^{\langle i \rangle}\right) - \omega_o(\alpha)}{\delta\alpha_{1,1} \cdot \omega_o(\alpha)}$

$S = \begin{bmatrix}-1.342\\-1.326\\0.281\\0.665\end{bmatrix} \cdot \frac{\%}{mrad}$

For each race perform a Hertz analysis to calculate the size of the contact area and the contact stress. Note, the races are assumed to be conical (or straight sided).

For each race, calculate the diameter of an equivalent cylindrical contact surface (negative radii are internal surfaces).

$\overrightarrow{d_r := -\left[\frac{d_c(\alpha)}{\sin(\alpha)}\right]}$

$d_r = \begin{bmatrix}11.294\\10.809\\-30.809\\-30.809\end{bmatrix} \cdot mm$

Calculate the relative radii and the contact radius (equivalent to a ball on a flat).

$$R_a := \left(\frac{2}{d} + \frac{1}{d_r} - \left|\frac{1}{d_r}\right|\right)^{-1} \qquad R_b := \left(\frac{2}{d} + \frac{1}{d_r} + \left|\frac{1}{d_r}\right|\right)^{-1} \qquad R_c := \sqrt{R_a \cdot R_b}$$

$$R_a = \begin{bmatrix} 5 \\ 5 \\ 7.403 \\ 7.403 \end{bmatrix} \text{mm} \qquad R_b = \begin{bmatrix} 2.652 \\ 2.597 \\ 5 \\ 5 \end{bmatrix} \text{mm} \qquad R_c = \begin{bmatrix} 3.641 \\ 3.604 \\ 6.084 \\ 6.084 \end{bmatrix} \text{mm}$$

Calculate the radial load per ball for each race.

$$P_r := \frac{f_a}{n} \cdot \begin{bmatrix} \left(\cot(\theta_3) + \cot(\theta_4)\right) \cdot \left(\cos(\theta_1) + \sin(\theta_1) \cdot \cot(\theta_2)\right)^{-1} \\ \left(\cot(\theta_3) + \cot(\theta_4)\right) \cdot \left(\cos(\theta_2) + \sin(\theta_2) \cdot \cot(\theta_1)\right)^{-1} \\ \csc(\theta_3) \\ \csc(\theta_4) \end{bmatrix} \qquad P_r = \begin{bmatrix} 5.717 \\ 7.292 \\ 6.44 \\ 6.44 \end{bmatrix} \cdot \text{lbf}$$

Calculate the major and minor radii of the elliptical contact area.

$$E_c := \left[\frac{1 - \nu_b^2}{E_b} + \frac{1 - \nu_r^2}{E_r}\right]^{-1} \qquad G_c := \frac{1}{2}\left[\frac{2 + \nu_b - \nu_b^2}{E_b} + \frac{2 + \nu_r - \nu_r^2}{E_r}\right]^{-1}$$

$$a := \left[\left(\frac{3 \cdot P_r \cdot R_a}{4 \cdot E_c}\right)^{\frac{1}{3}} \cdot \left(\frac{R_a}{R_b}\right)^{\frac{1}{6}}\right] \qquad b := \left[\left(\frac{3 \cdot P_r \cdot R_b}{4 \cdot E_c}\right)^{\frac{1}{3}} \cdot \left(\frac{R_b}{R_a}\right)^{\frac{1}{6}}\right]$$

$$a = \begin{bmatrix} 105.063 \\ 114.334 \\ 119.669 \\ 119.669 \end{bmatrix} \cdot \mu\text{m} \qquad b = \begin{bmatrix} 68.84 \\ 73.881 \\ 92.122 \\ 92.122 \end{bmatrix} \cdot \mu\text{m}$$

Calculate the maximum contact pressure located at the center of each contact area.

$$p := \frac{3 \cdot P_r}{2 \cdot \pi \cdot a \cdot b} \qquad p = \begin{bmatrix} 243.505 \\ 265.912 \\ 179.934 \\ 179.934 \end{bmatrix} \cdot \text{ksi} \qquad p = \begin{bmatrix} 171.201 \\ 186.955 \\ 126.506 \\ 126.506 \end{bmatrix} \cdot \text{HBn}$$

A rolling contact pair that carries a tractive force experiences slip over a portion of the contacting area near the trailing edge. This manifests an overall slip, termed creep in the literature, that is proportional to the nominal distance rolled and is dependent on the tractive force, the coefficient of friction and other parameters already entered. The theory by Johnson is approximate for a ball on a plane and does not include spin. The effect of spin is to increase the area that slips, so this estimate may be factors too low.

Enter the coefficient of friction and the drive torque.

$\mu := 0.075$ $\quad T := \frac{52 \cdot lbf \cdot 1 \cdot mm}{2 \cdot \pi}$ $\quad T = 0.326\ in \cdot lbf$

In general, each race carries a slightly different traction due to differences in contact angles. Calculate the traction **tr** per ball at each race corresponding to the drive torque at the output race (4).

$$tr := \frac{2 \cdot T}{n \cdot d_c(\alpha_3)} \cdot \left[\begin{bmatrix} 1 & 1 & 1 & 1 \\ \cos(\alpha_1) & \cos(\alpha_2) & \cos(\alpha_3) & \cos(\alpha_4) \\ \sin(\alpha_1) & \sin(\alpha_2) & \sin(\alpha_3) & \sin(\alpha_4) \\ 0 & 0 & 0 & 1 \end{bmatrix}^{-1} \right]^{\langle 4 \rangle} \qquad tr = \begin{bmatrix} -0.081 \\ 0.081 \\ -0.092 \\ 0.093 \end{bmatrix} lbf$$

Calculate the percentage slip using Johnson's theory.

$$\xi := \left[\frac{3 \cdot \mu \cdot P_r}{16 \cdot a \cdot b \cdot G_c} \cdot \left[1 - \left(1 - \frac{tr}{\mu \cdot P_r} \right)^{\frac{1}{3}} \right] \right] \qquad \xi = \begin{bmatrix} -0.013 \\ 0.012 \\ -9.306 \cdot 10^{-3} \\ 0.011 \end{bmatrix} \cdot \%$$

Calculate the ratio of material velocity moving through each contact to the velocity at the output race.

$$v_r := \frac{\overrightarrow{d \cdot \sin(\alpha)} \cdot (\psi(\alpha)_1 - \psi(\alpha)_3) + \overrightarrow{d \cdot \cos(\alpha)} \cdot \psi(\alpha)_2}{d \cdot \sin(\alpha_4) \cdot \psi(\alpha)_1 + d \cdot \cos(\alpha_4) \cdot \psi(\alpha)_2 + D \cdot \psi(\alpha)_3} \qquad v_r = \begin{bmatrix} 26.254 \\ 25.712 \\ -26.712 \\ -25.712 \end{bmatrix}$$

Calculate the percentage of slip at the output shaft. $\quad \overrightarrow{\Sigma \left| \xi \cdot v_r \right|} = 1.16 \cdot \%$

Define units: $\quad mrad \equiv 10^{-3} \cdot rad$ $\quad ksi \equiv 1000 \cdot psi$ $\quad \mu m \equiv 10^{-6} \cdot m$

$$rpm \equiv \frac{2 \cdot \pi \cdot rad}{60 \cdot sec} \qquad HBn \equiv \frac{kgf}{mm^2}$$

8.5.2 Test Results for the NIF Actuator

Tests were conducted to determine the important characteristics of the NIF actuator and its main components. This section presents the more significant results of this testing. The tests included measurements of torque to cause sliding in the friction drive, positioning capability and axial stiffness. After millions of cycles, repeat tests demonstrated a sufficient level of reliability. The overall level of performance was good, but the velocity ratio varied due to slip and part tolerances more significantly than anticipated.

The torque required to cause sliding in the friction drive was measured with a torque meter for two conditions: 1) with the input shaft stationary and the output shaft driven by external means, and 2) with the output shaft stationary and the input shaft driven by a stepper motor. In both tests a torque meter measured the output shaft. This information may be used to set the proper preload in the friction drive to protect the ball screw from over loading against a positive stop. The static load capacity for the ball screw is 130 kgf, which is equivalent to 1.8 in-lb of torque for the 1 mm lead screw. The measured drive torque was 1.7 in-lb for an axial preload of approximately 10 lbs. The coefficient of friction was calculated to be 0.13 for the driven output shaft and 0.075 for the stationary output shaft. The smaller value apparently results from combined rolling and sliding since the balls continue to roll as the input shaft rotates. When the output shaft is driven, pure sliding can occur between the balls and the outer races. The steel balls and races were grease lubricated for both cases, which is necessary for long life.

The actuator is expected to provide up to 52 lbs of actuation force. The calculated torque for this load is only 0.33 in-lb, which is well below the limiting torque of the friction drive. However, a slip phenomenon occurs with traction-type mechanical drives that effectively makes the velocity ratio load dependent. This behavior is evident in Figure 8-28, where under different loads the actuator follows different forward and backward paths. The theory dates back to the 1920's but it is difficult to apply to this problem precisely because the balls have a component of spin in addition to rolling. The slip predicted for this torque and axial preload is 1.16 %, which is a factor of four less than measured. The measured slip is fairly consistent (within ± 0.5%) over various ranges of travel and number of cycles on the unit. This behavior would be of minor consequence for a real-time control system, but the NIF control system provides only supervisory control. In addition, the system requires fairly high dynamic range. Each mirror mount would require mapping so that the control system could correct for the uncertain velocity ratio.

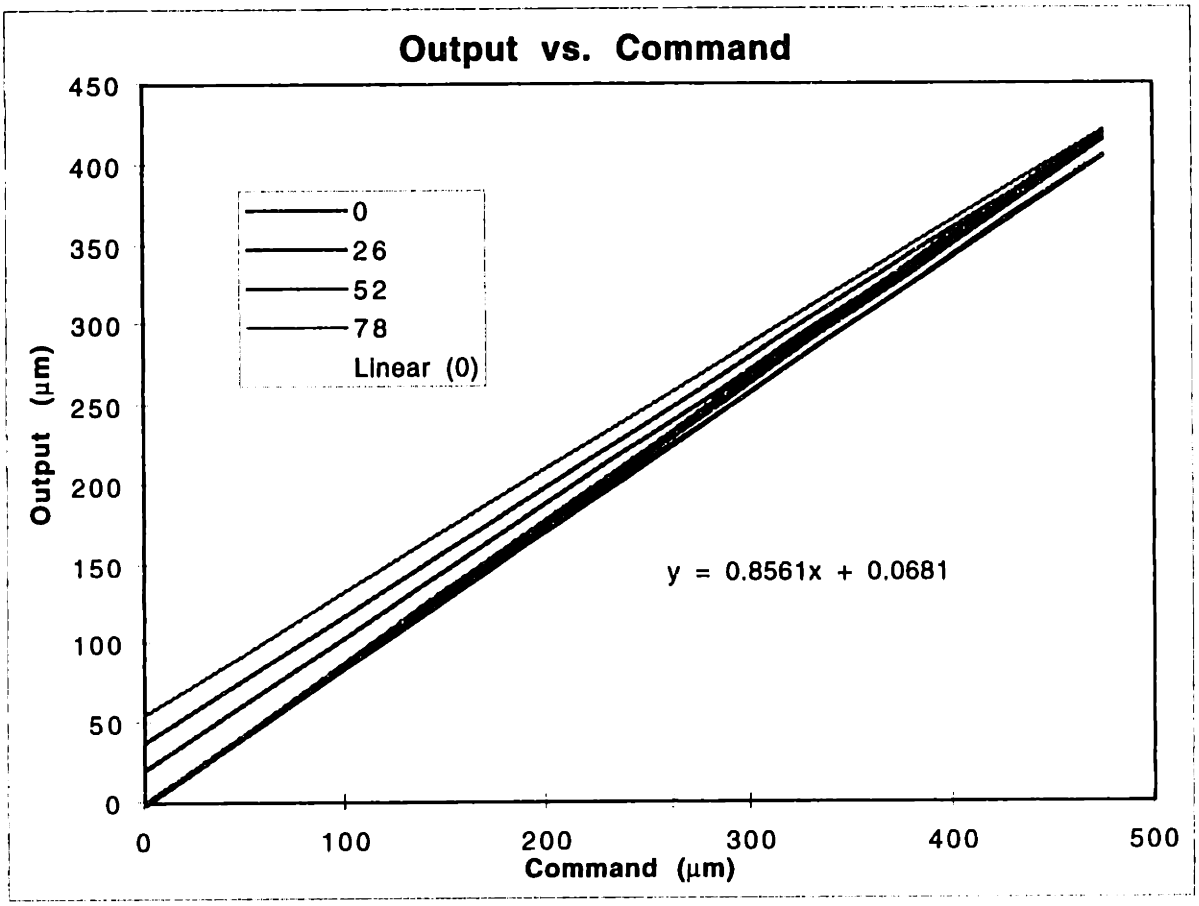


Figure 8-28 These curves show the load-dependent behavior of the friction drive. The actuator is commanded in 47.5 µm steps from zero to 475 µm then back to zero. This move would map to a straight line with a unit slope if the friction drive had a perfect 10C:1 ratio. Instead, the slope is 0.8561 for zero load and it varies slightly depending on load and direction as indicated.

The actuator demonstrated exceptional positioning capability even below the theoretical 25 nm step size. Figure 8-29 shows a cycle where the motor is commanded to move 40 half-step increments in one direction, then is commanded to return over the same increments. The fact that the slope changes with directions is due to slip in the friction drive as previously discussed. Each marker represents one half step of the 200 step/rev motor but the curves are generated using the angular motion of the motor as measured by an encoder. The uneven spacing of the markers reflects the inability of the motor to make accurate half steps. This may be due to improper tuning of the controller, but the problem will not exist with 400 step/rev motors operated with full steps. Figure 8-30 shows the same test with the effect of 4% slip subtracted from the data. This value corresponds to the 52 lbs axial load case shown in Figure 8-28. The remaining difference between the forward and backward curves gives an estimate for the hysteresis to expect for a fixed-ratio speed reducer. The average hysteresis is approximately 40 nm, which is well below the specification. Part of the hysteresis is due to friction in the test stand acting on the compliance of the actuator, but the level is probably very similar to the mirror mounts. Backlash at zero axial load is not present since all the mechanical components are preloaded.

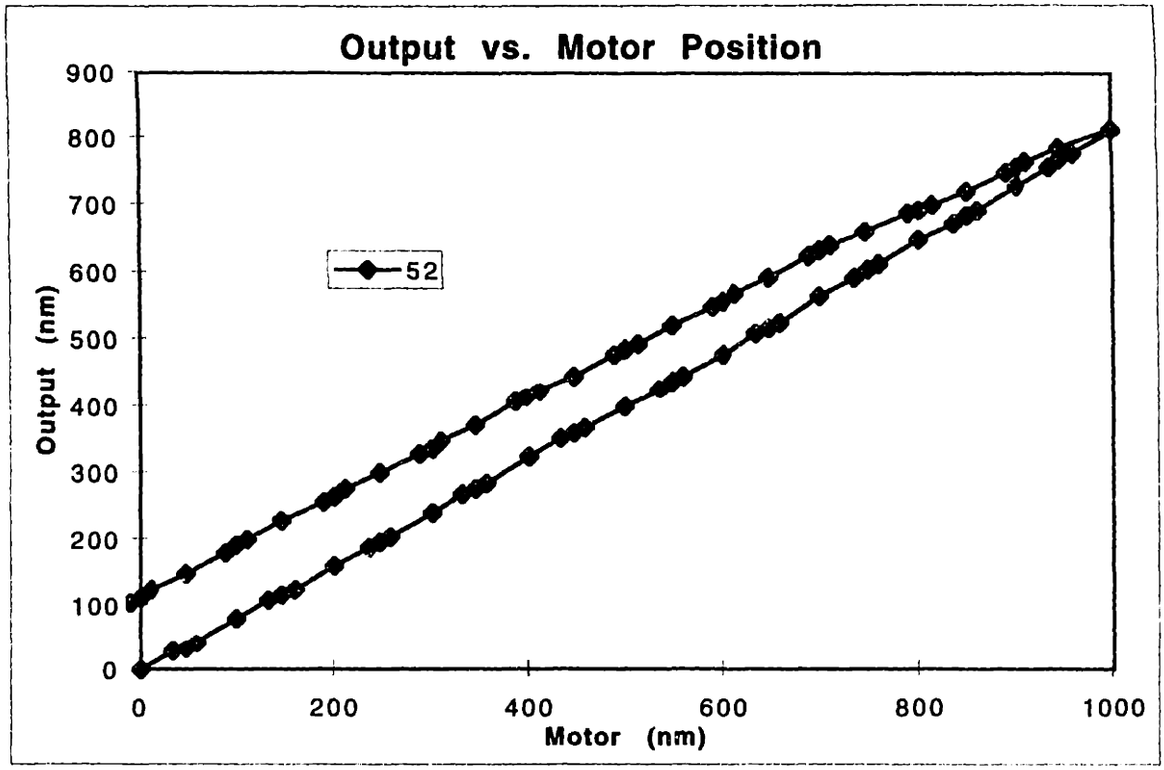


Figure 8-29 Linear travel versus motor angular position under a 52 lb load. The horizontal scale is computed for a 100:1 velocity ratio and 1 mm lead of the screw. Each marker indicates one half step of the 200 step/rev motor.

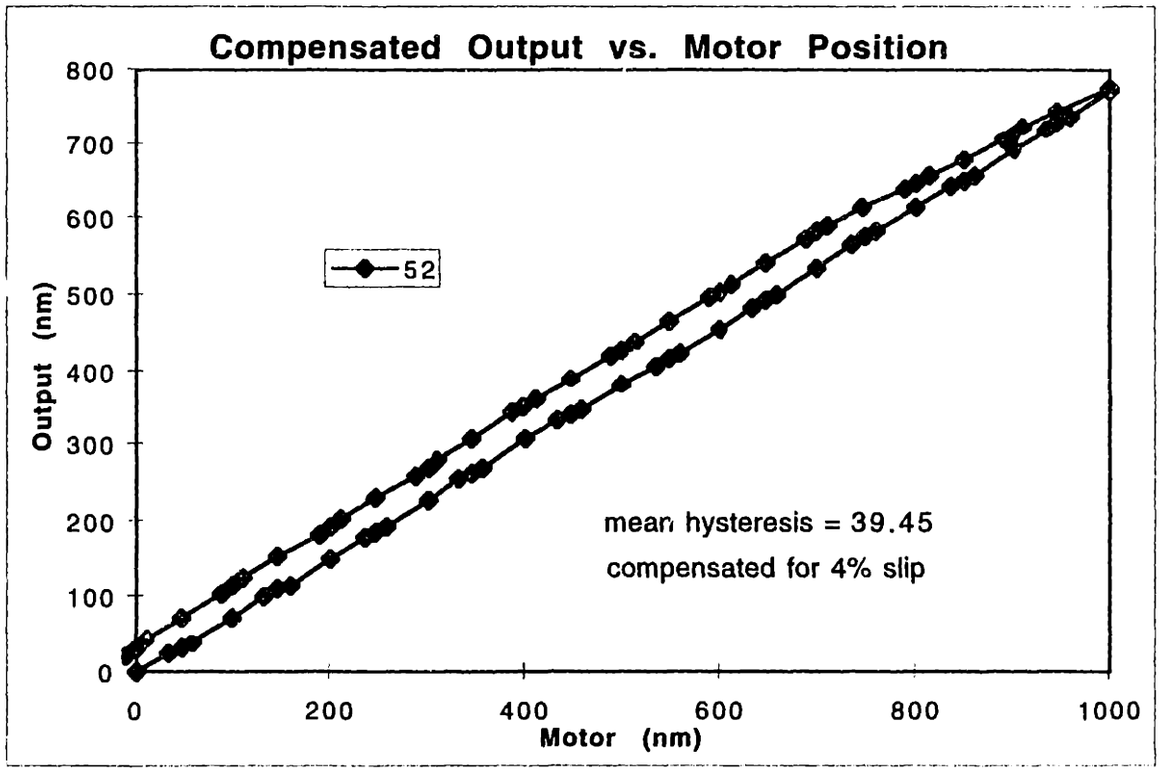


Figure 8-30 The same data presented in Figure 8-29 is compensated for 4% slip in the friction drive.

Axial compliance tests of an assembled actuator and its primary components were somewhat inconsistent and indicate that approximately 20% of the total compliance is unidentified. An overly optimistic compliance goal was set at 0.25 μm/kgf for the actuator, which at the time did not include the compliance of the built-in pivot flexure. This goal gives a natural frequency of 200 Hz for the suspend mass of 25 kg, realizing that other compliance sources would lower the frequency. Figure 8-31 shows axial displacement of the actuator versus axial load as measured on the test stand. In this test, the load was applied in increments of 26 lbs up to 78 lbs then removed incrementally. The compliance is the slope of the curve, 0.2744 μm/lb or 0.60 μm/kgf. Tests conducted on different actuators assembled to mirror mounts are consistent with this value. This is 2.4 times more compliant than the goal but probably is of little consequence given the original optimism.

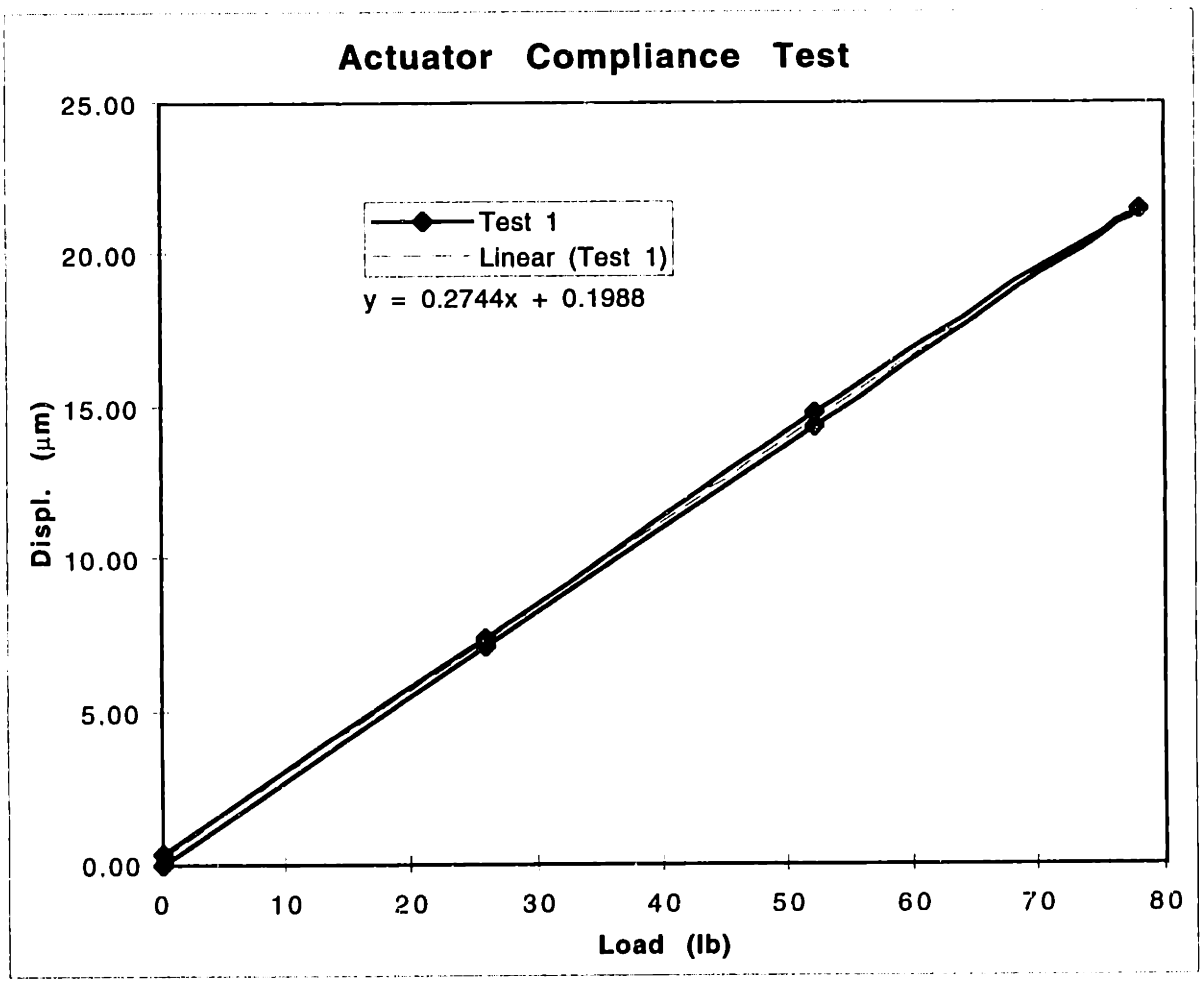


Figure 8-31 Axial displacement of the actuator versus axial load as measured on the actuator test stand.

Additional tests were constructed to separately measure the compliance of the three main components in the actuator: the ball screw, the thrust bearing and the pivot flexure. The load in each test was applied manually with a force gauge. The measured compliance of the ball screw is 0.13 μm/kgf. Two types of duplex angular-contact thrust bearings were measured: 40° contact angle without preload, and 25° contact angle with medium preload.

Preloaded 40° bearings are not available as a standard product. The 40° bearing was difficult to measure, but it can be interpreted as being somewhat stiffer than the 25° bearing. However, the 25° bearing appears to be adequate with a compliance of 0.15 μm/kgf, and it is available off the shelf with medium preload and higher precision than the 40° bearing. It is interesting to compare the original goal (0.25) with the sum compliance of the ball screw and thrust bearing (0.28) since these would have been the main sources at the time that the goal was set.

The most significant compliance source in the actuator is the pivot flexure measured at 0.20 μm/kgf. The finite element prediction of the compliance is only 0.11 μm/kgf, 1.8 times stiffer. The sum of the three main actuator components is 0.48 μm/kgf or 80% of the assembled actuator. Additional compliance may come from bolted joints, but I suspect the flange connection between the ball nut and the flexure interacts in a way that is difficult to duplicate in separate tests. The discrepancy of the finite element result is also surprising.

The actuator was disassembled after 4 million cycles to inspect for wear and other signs of use. The friction drive races showed visible tracks produced by the balls. These were most noticeable on the inner spool because the contact stress is higher there and the path length is much shorter than for the outer races. The indentation on the inner spool was measured and found to be of the same order as the ground surface finish. The track is concave with less roughness than surrounding areas but not significantly deeper. There is no indication after 4 million cycles that the friction drive is nearing its useful lifetime. An unusual discovery was that 18 balls were found clinging to the end of the screw that apparently worked their way out of the nut. A careful disassembly of the ball screw revealed that each of the three ball tracks contained 21 balls and all three ball returns were working flawlessly. In addition, the balls within the ball returns appeared slightly rougher under magnification than the ones found outside. This indicates that the balls worked loose very early in testing. The manufacturer confirmed that the total number of balls recovered is very close to the prescribed number. The most likely explanation is that the unit was assembled improperly, but this showed no apparent degradation in performance except perhaps the compliance. The grooves in the screw and the nut showed no signs of wear under magnification. No attempt was made to reassemble with 81 0.8 mm balls.

The one negative result of the testing is the potential variability of velocity ratio from one unit to the next. Although this is a simple error to compensate in software, the decision has been made to use a harmonic drive rather than the friction drive. The precision and life of the harmonic drive are expected to be adequate for this application. Figure 8-32 shows a layout of the actuator with a harmonic drive rather than the friction drive. The assembly is slightly longer but there is no clearance problem with any of the mirror mounts. A slip clutch is incorporated into this design by using a stack of four belleville washers to preload the interface between the flex-spline of the harmonic drive and the bearing cap.

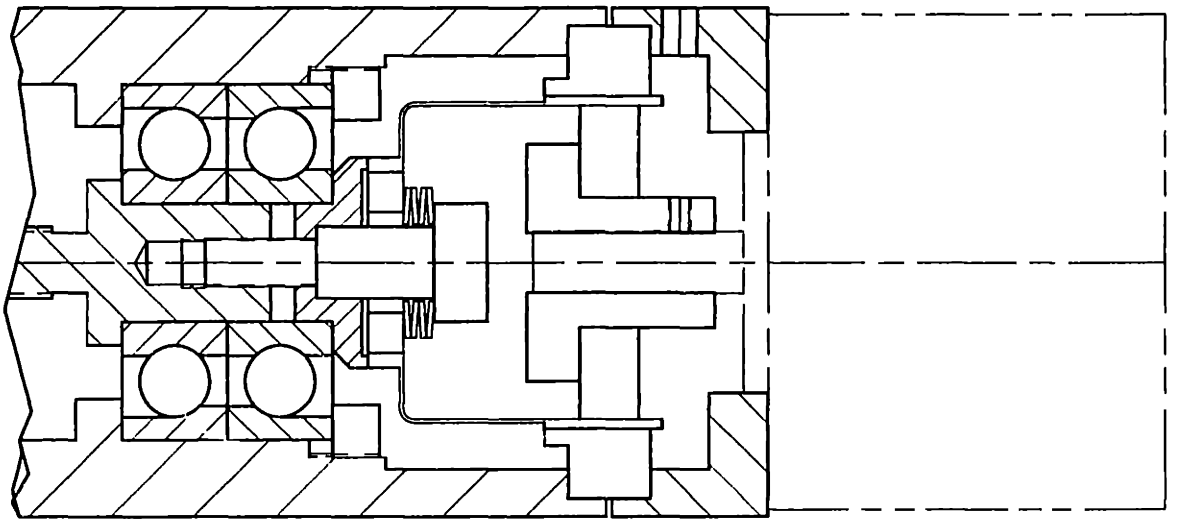


Figure 8-32 Layout of the actuator (motor end) with an HDUC 11-100-2AR harmonic drive from HDS.

9 Conceptual Design of a Horizontal Machining Center

The purpose of this design study is to demonstrate the use of methods and strategies for designing higher-accuracy machine tools for general industry. The challenge is to achieve the productivity and the precision that industry needs for a price that they will pay. As a specific example, we will use a 500 mm class horizontal machining center based loosely on the Maxim™ 500 by Cincinnati Milacron. The machine is intended to be medium duty, to be highly productive and to have exceptional contouring accuracy. While many of the decisions made will be unique to this example, the process and logic behind the decisions should be quite general and applicable to other types and sizes of machine tools.

The purpose for designing a new machine tool is to best satisfy the needs of the customer who expects to profit from the investment in the new machine. The best possible machine will completely satisfy the customer needs for the minimum lifetime cost to the customer. The design team must be conscious of this purpose throughout the design process and a well-defined set of design goals will help guide them to the best solution. At any time during the design process, the design goals should represent the best vision of the end product or indicate the knowledge of how to get there. Some goals will be dynamic with the design process until they become discarded, implemented or refined in some way. Others may be relatively static such as the functional requirements or a list of desirable features. Still others will be strategies for satisfying goals, for example, using temperature controlled liquid to stabilize critical metrology structures. The design goals for this example are developed and presented in the following sections.

The functional requirements provide the highest level of guidance for the design and must properly represent the customer needs.

- The machine must be capable of removing material from a workpiece using a milling process where the tool rotates in a single spindle.
- The most general configuration requires the five-degree-of-freedom relationship between the rotating tool and the workpiece to be fully programmable under computer control. The standard configuration requires three linear axes of travel.
- The spindle must have access to all exposed sides of the workpiece as provided, for example, by an index table. In all cases this relationship must be sufficiently accurate and rigid to meet the needs of general industry.
- The work table must be horizontal to facilitate the loading of a large workpiece or fixturing.

- The working range of the machine must be roughly equal in the three spatial directions X, Y and Z.
- The work volume must be enclosed to prevent chips and coolant from escaping and must encourage the flow of chips and coolant to a removal system.
- The machine must be configurable to interface with automatic workpiece loading systems.
- The machine must provide a buffer of tools and be capable of automatically loading tools in any order into the spindle.
- The machine must be readily transportable to the customer site and operate on standard utilities.

9.1 Developing Specifications

The specifications are a set of numbers that potential customers will use to judge the capability of the proposed machine versus competing machines. Although a comparative market survey is important to do, simply improving the numbers over the competition probably will not best satisfy the customer needs. Ideally, the specifications chosen should maximize the ratio of the value added by the machine to its total lifetime costs. In practice, mathematical functions for the value and the cost in terms of the specifications are difficult to develop, but it should be possible to estimate the sensitivities of value and cost to any one or combination of specifications. For example, if better volumetric accuracy reduces scrap, improves part quality or eliminates an operation on a boring mill or an inspection machine, then considerable value has been added. Only when the cost of the solution, say geometric error compensation and temperature control, is smaller than the increase in value added will the better specification be an advantage. Some specifications may require iteration since the decisions depend on the design solutions that evolve over the course of the project. This is why the specifications are considered design goals at this stage.

The specifications presented in the following sections are developed from a rather general knowledge of their value to potential customers. A machine tool builder such as Cincinnati Milacron would have considerably more knowledge and access to information in this area to develop better specifications. Consider this a recommended model for developing the specifications rather than a set of recommended specifications. Spindle speed and axis force are examples that depend on the materials and tooling expected to run on the machine. A *machining process model* is helpful in this regard. Appendix E presents the development of a multiple-tooth, rotating-tool cutting model based on orthogonal cutting. This was coded into a spreadsheet and used to develop a *machining study* for several materials (steel, cast iron and aluminum), a reasonable range of tool sizes, and moderate to high cutting speeds consistent with modern cutting tools. The machining study will help guide the decisions on the requirements presented in following sections.

9.1.1 Machining Study

The machining process model relates different cutting conditions to the corresponding requirements for the machine. Table 9-1 shows the case for moderate cutting speeds on tough steel, which represents a lower bound on axis and spindle speeds and an upper bound on axis forces and spindle torque. Table 9-2 shows high cutting speeds on mild steel, which is similar to moderate cutting speeds on aluminum. Table 9-3 shows high cutting speeds on aluminum, which represents an upper bound on axis and spindle speeds and a lower bound on axis forces and spindle torque. There are four different machining operations for each spreadsheet. The first is a slotting operation using a one-inch diameter end mill. The second is a slab milling operation using a two-inch diameter end mill. The third and fourth are facing operations using four- and eight-inch diameter face mills. Each operation has the same area of cut but most other parameters are necessarily different.

Working with the spreadsheet, that is, by entering the parameters and reviewing the calculated values, is a judgment process based on anticipated machine capabilities and the expected operating conditions. The variety of operations and materials represented in the spreadsheets will tax the machine in different ways. The specifications developed using this approach should result in a machine that is versatile and well balanced.

9.1.2 Spindle Power, Torque and Speed

Although a spindle will not be designed for this example, it is important to balance the capabilities of the machine to the spindle. The extremes of the machining study span an incredible range to try to satisfy with one or even two spindle designs. The spindle speed ranges from 30,000 to 240 rpm. The spindle torque ranges from 690 to 6 ft-lb. The spindle power ranges from 108 to 18 hp. Reaching 30,000 rpm or greater spindle speeds requires either a physically smaller spindle than is typical or an advanced technology such as active magnetic bearings. A magnetic bearing spindle with a directly coupled motor is an excellent option for a dedicated high speed machine, but it would not be capable or at least cost effective for general-purpose machining. Neither is a smaller spindle consistent with general-purpose machining. A reasonable solution, however, is to provide a smaller spindle with a built-in overdrive transmission as a piece of tooling. For example, one or more spindle tools complete with cutters could be stored in the tool chain and used only for high speed work where the loads and the process stiffness are low. This strategy should greatly increases the value of the machine for the least additional cost.

A reasonable maximum spindle speed is 7500 rpm, which enables high speed machining of aluminum with a small face mill as Table 9-3 indicates. A four-to-one overdrive with a spindle tool gives 30,000 rpm, which enables high speed machining with an end mill. This spindle speed is only slightly greater than the standard 7000 rpm spindle on the Maxim 500 and the Ram 500 by Giddings and Lewis (G&L) so the cost differential should be negligible.

Parameters	*Units*	*Enter Values*	*Enter Values*	*Enter Values*	*Enter Values*
Tool Mat'l		Carbide	Carbide	Carbide	Carbide
Work Mat'l		4140-300HB	4140-300HB	4140-300HB	4140-300HB
Shear Str'th	ksi	88	88	88	88
Cutting Speed	ft/min	300	300	500	500
Max Chip Th'k	in	0.005	0.005	0.01	0.01
Av. Cutter Dia	in	1	2	4	8
Entry Point	in (to center)	-0.50	-1.00	-2.00	-4.00
Exit Point	in (to center)	0.50	-0.75	2.00	4.00
Depth of Cut	in	1.00	4.00	0.25	0.13
No. of Teeth		4	6	6	12
Lead Angle	degree	0	0	45	45
Rake Angle	degree	15	15	15	15
Friction Angle	degree	45	45	45	45
Parameters	*Units*	*Calc'd Value*	*Calc'd Value*	*Calc'd Value*	*Calc'd Value*
Spindle Speed	rpm	1146	573	477	239
Tooth Freq.	Hz	76	57	48	48
Tooth Spacing	in	0.785	1.047	2.094	2.094
Entry Angle	degree	-90.00	-90.00	-90.00	-90.00
Exit Angle	degree	90.00	-48.59	90.00	90.00
Arc of Contact	in	1.571	0.723	6.283	12.566
Teeth in Cut		2.000	0.690	3.000	6.000
Feed/Tooth	in	0.0050	0.0076	0.0141	0.0141
Av. Chip Th'k	in	0.0032	0.0026	0.0064	0.0064
Min Chip Th'k	in	0.0000	0.0000	0.0000	0.0000
Feed Rate	in/min	22.92	25.99	40.51	40.51
Shear Angle	degree	30	30	30	30
Radial Press.	ksi	176	176	176	176
Tang'l Press.	ksi	305	305	305	305
Spec. Power	hp-min/in^3	0.77	0.77	0.77	0.77
Area of Cut	in^2	1.00	1.00	1.00	1.00
Removal Rate	in^3/min	22.92	25.99	40.51	40.51
Spindle Power	hp	17.64	20.00	31.19	31.19
Spindle Torque	ft-lbs	80.86	183.38	343.07	686.13
X-Force	lbs	880	2501	660	660
Y-Force	lbs	1524	2109	1617	1617
Z-Force	lbs	0	0	840	840
Resultant	lbs	1760	3272	1938	1938
Process Stiff.	lb/mil	352	654	194	194

Table 9-1 Milling model results for several cutter sizes on tough steel at moderate speeds.

Parameters	*Units*	*Enter Values*	*Enter Values*	*Enter Values*	*Enter Values*
Tool Mat'l		Carbide	Carbide	Silicon Nitride	Silicon Nitride
Work Mat'l		Steel-140HB	Steel-140HB	Steel-140HB	Steel-140HB
Shear Str'th	ksi	40	40	40	40
Cutting Speed	ft/min	1000	1000	3000	3000
Max Chip Th'k	in	0.005	0.005	0.01	0.01
Av. Cutter Dia	in	1	2	4	8
Entry Point	in (to center)	-0.50	-1.00	-2.00	-4.00
Exit Point	in (to center)	0.50	-0.75	2.00	4.00
Depth of Cut	in	1.00	4.00	0.25	0.13
No. of Teeth		4	6	6	12
Lead Angle	degree	0	0	45	45
Rake Angle	degree	15	15	15	15
Friction Angle	degree	45	45	45	45
Parameters	*Units*	*Calc'd Value*	*Calc'd Value*	*Calc'd Value*	*Calc'd Value*
Spindle Speed	rpm	3820	1910	2865	1432
Tooth Freq.	Hz	255	191	286	286
Tooth Spacing	in	0.785	1.047	2.094	2.094
Entry Angle	degree	-90.00	-90.00	-90.00	-90.00
Exit Angle	degree	90.00	-48.59	90.00	90.00
Arc of Contact	in	1.571	0.723	6.283	12.566
Teeth in Cut		2.000	0.690	3.000	6.000
Feed/Tooth	in	0.0050	0.0076	0.0141	0.0141
Av. Chip Th'k	in	0.0032	0.0026	0.0064	0.0064
Min Chip Th'k	in	0.0000	0.0000	0.0000	0.0000
Feed Rate	in/min	76.39	86.62	243.09	243.09
Shear Angle	degree	30	30	30	30
Radial Press.	ksi	80	80	80	80
Tang'l Press.	ksi	139	139	139	139
Spec. Power	hp-min/in^3	0.35	0.35	0.35	0.35
Area of Cut	in^2	1.00	1.00	1.00	1.00
Removal Rate	in^3/min	76.39	86.62	243.09	243.09
Spindle Power	hp	26.73	30.31	85.06	85.06
Spindle Torque	ft-lbs	36.76	83.35	155.94	311.88
X-Force	lbs	400	1137	300	300
Y-Force	lbs	693	959	735	735
Z-Force	lbs	0	0	382	382
Resultant	lbs	800	1487	881	881
Process Stiff.	lb/mil	160	297	88	88

Table 9-2 Milling model results for several cutter sizes on mild steel at high speeds.

Parameters	*Units*	*Enter Values*	*Enter Values*	*Enter Values*	*Enter Values*
Tool Mat'l		Carbide	Carbide	Carbide	Carbide
Work Mat'l		6061-T6	6061-T6	6061-T6	6061-T6
Shear Str'th	ksi	26	26	26	26
Cutting Speed	ft/min	8000	8000	8000	8000
Max Chip Th'k	in	0.005	0.005	0.01	0.01
Av. Cutter Dia	in	1	2	4	8
Entry Point	in (to center)	0.00	-1.00	-2.00	-4.00
Exit Point	in (to center)	0.50	-0.88	2.00	4.00
Depth of Cut	in	1.00	4.00	0.13	0.06
No. of Teeth		4	6	6	12
Lead Angle	degree	0	0	45	45
Rake Angle	degree	15	15	15	15
Friction Angle	degree	45	45	45	45
Parameters	*Units*	*Calc'd Value*	*Calc'd Value*	*Calc'd Value*	*Calc'd Value*
Spindle Speed	rpm	30558	15279	7639	3820
Tooth Freq.	Hz	2037	1528	764	764
Tooth Spacing	in	0.785	1.047	2.094	2.094
Entry Angle	degree	0.00	-90.00	-90.00	-90.00
Exit Angle	degree	90.00	-61.04	90.00	90.00
Arc of Contact	in	0.785	0.505	6.283	12.566
Teeth in Cut		1.000	0.483	3.000	6.000
Feed/Tooth	in	0.0050	0.0103	0.0141	0.0141
Av. Chip Th'k	in	0.0032	0.0026	0.0064	0.0064
Min Chip Th'k	in	0.0000	0.0000	0.0000	0.0000
Feed Rate	in/min	611.15	946.80	648.23	648.23
Shear Angle	degree	30	30	30	30
Radial Press.	ksi	52	52	52	52
Tang'l Press.	ksi	90	90	90	90
Spec. Power	hp-min/in^3	0.23	0.23	0.23	0.23
Area of Cut	in^2	0.50	0.50	0.50	0.50
Removal Rate	in^3/min	305.58	473.40	324.11	324.11
Spindle Power	hp	69.50	107.67	73.72	73.72
Spindle Torque	ft-lbs	11.95	37.01	50.68	101.36
X-Force	lbs	-13	500	98	98
Y-Force	lbs	142	386	239	239
Z-Force	lbs	0	0	124	124
Resultant	lbs	143	632	286	286
Process Stiff.	lb/mil	29	126	29	29

Table 9-3 Milling model results for several cutter sizes on aluminum at high speeds.

The maximum spindle torque required influences whether to use a transmission. The high torque provided by a transmission is important when taking heavy cuts in tough materials with large diameter face mills. However, productivity may not suffer in many cases by using smaller face mills. Heavy roughing cuts are usually made to unhardened parts and then lighter finish cuts are made after heat treatment. In addition, advancements in tooling technology tend toward higher speeds and reduced cutting loads provided by higher rake angles. Recent advancements in motor and drive technologies make direct drive systems practical, available and cost effective. For these reasons, the value of high torque may not justify the cost of a transmission versus the simplicity of a direct drive system.

A heavy cut with a large face mill in mild steel requires approximately 300 ft-lb of torque. A heavy cut with a small face mill (or a light cut with a large face mill) in tough steel requires approximately 350 ft-lb. This is an acceptable target for a direct drive motor. Reuland Electric makes a wide range of frameless AC induction motors for spindle applications. A frame size 1500-2475 (15.00" OD by 24.75" long) is available in this torque range with four or six poles. The six-pole motor would give better low speed control and its maximum speed of 8000 rpm is adequate. The maximum power rating at 60 Hz is 102 hp at 1800 rpm for the four-pole motor and 69 hp at 1200 rpm for the six-pole motor, which translates to 300 ft-lb maximum torque for each. This is a continuous rating assuming liquid cooling and presumably a higher machine-tool-duty rating could be negotiated with the manufacturer. The power available at higher speeds is proportionally higher and is limited mainly by the power electronics. For example, the six-pole motor would produce approximately 100 hp at 90 Hz or 140 hp at 120 Hz. The inside diameter of the rotor can be up to 6 inches, which is adequate to house the draw-bar mechanism.

This information looks favorable enough to set the spindle requirements in line with a direct drive system. The maximum speed for the main spindle will be 7500 rpm. A four-to-one spindle tool for high-speed end milling will extend the speed range to 30,000 rpm. A power rating of 100 hp (75 kW) will satisfy high-speed machining needs but costs may drive it down for the standard product. The power is constant down to a base speed of 1800 rpm where the torque is 300 ft-lb. The machine-tool-duty torque targeted at 350 ft-lb is available below a base speed of 1543 rpm. For comparison, the 7000 rpm spindle on the Maxim has a continuous torque of 421 ft-lb below the base speed of 334 rpm and a continuous power of 27 hp above base speed.

9.1.3 Axis Velocity, Acceleration and Thrust

The machining study provides the basis for determining the velocity requirement for the slide axes. The maximum feed rate in the study is 947 in/min or 24 m/min. In comparison, the maximum velocity for the Maxim 500 is 25 m/min and for the Ram 500 is 38 m/min. It seems that the speed of the Maxim is adequate and that there is no additional value provided by the Ram. However, the disparity here is too significant to ignore. Greater speed is

valuable for rapid traverse and perhaps worth some additional cost. A reasonable compromise is 30 m/min, which is a pleasing round number in any units: 1200 in/min, 100 ft/min, 500 mm/s or 20 in/s. This speed is reasonably achieved using a ball screw with either a 12 mm lead at 2500 rpm or a 0.5 in lead at 2400 rpm.

The value of high axis accelerations relates to the time saved during stop-and-start moves and relatively sharp contours found in pockets. In addition, tool life may suffer in proportion to the time when the chip load is suboptimal during deceleration, especially for high speed machining where the chip carries away most of the heat. Table 9-4 shows the effect of varying the axis acceleration on a number of relevant parameters. The minimum tool path radius and the distance to accelerate are related by a factor of two, proportional to the square of the velocity and inversely proportional to the acceleration. The specific power is equal to the acceleration times the velocity and is one measure of the cost. The time to accelerate is proportional to the velocity and inversely proportional to the acceleration. From a productivity standpoint, an acceleration rate of 5 m/s^2 provides a very reasonable time of 100 ms to reach the maximum speed of 500 mm/s. Of course this time is lower for more typical feedrates and starts to approach the time delays typical of the CNC (computer numerical control). Although some customers request one or two *g*'s acceleration for high speed machines, the value is minimal at this feedrate and too expensive for a general purpose machine. The specification stands at 5 m/s^2 acceleration on each linear axis and a comparable value for rotary axes.

Parameters	*Units*	*Results for a terminal velocity of*			**0.5**	*m/s*
Axis acceleration	m/s^2	5	10	15	20	25
Min. radius at speed	mm	50.0	25.0	16.7	12.5	10.0
Distance to accelerate	mm	25.0	12.5	8.3	6.3	5.0
Specific power	W/kg	2.5	5.0	7.5	10.0	12.5
Time to accelerate	sec.	0.100	0.050	0.033	0.025	0.020
Time to 50 mm	sec.	0.200	0.150	0.133	0.125	0.120
Time to 100 mm	sec.	0.300	0.250	0.233	0.225	0.220
Time to 200 mm	sec.	0.500	0.450	0.433	0.425	0.420
Time to 400 mm	sec.	0.900	0.850	0.833	0.825	0.820
Time to 800 mm	sec.	1.700	1.650	1.633	1.625	1.620

Table 9-4 The acceleration study covers a range from approximately 0.5 to 2.5 times the acceleration of gravity (9.81 m/s^2). The other values are calculated using the specified axis velocity.

The axis force should be consistent with the specified spindle torque and power. The maximum spindle torque of 350 ft-lb acting at a 2 inch tool radius results in 2100 lb of force. The maximum power of 100 hp acting at a cutting speed of 1500 ft/min results in 2200 lb of force. These numbers are consistent with the machining studies for tough steel, and only slab milling results in greater force, where the maximum axis force is 2500 lb and the resultant force is 3300 lb. In comparison, the Maxim 500 has 4000 lb of Z-axis thrust and somewhat less for X and Y axes, perhaps 3000 lb. The Ram 500 has 3000 lb of Z-axis force and 2500 lb of X- and Y-axis force. The higher Z-axis force is generally regarded as

important for drilling operations. A reasonable compromise for the specification is 15 kN (3372 lb) for the Z axis and 12 kN (2698 lb) for the X and Y axes.

9.1.4 Part Size and Weight

By definition, a 500 mm class horizontal machining center has a 500 mm by 500 mm square pallet. Since the machine must be capable of machining to the corner of the square, there is no extra cost associated with allowing a circular part that circumscribes the square, approximately 700 mm (which matches the Maxim 500). A little extra room around the part could be of value for customers with odd shaped parts and the associated cost should not be too significant. For this reason, a maximum part diameter of 750 mm is chosen to match the G&L Ram 500. A round pallet of 650 mm will circumscribe the pallet of both the Maxim 500 and the Ram 500 and will provide additional pallet area for negligible additional cost. However as Figure 9-1 shows, this increases the volume of interference between the spindle and the pallet. A round pallet with two flats is a good compromise.

The height of the part should be the same order as the 750 mm diameter. There is value in increasing the height to allow more parts on a tombstone fixture or to accommodate taller parts. The costs associated with a longer travel is not very significant but the machine's stiffness and accuracy specifications will be more costly to meet. The choice of 750 mm part height is short of the Ram 500 by 100 mm.

The part weight is based on the maximum size and the expected density of the part. A solid steel part would weigh 25 kN (2550 kg). A solid aluminum part would weigh 8.8 kN (900 kg). The specification for both the Maxim 500 and the Ram 500 is 1000 kg. A more realistic density would be an iron casting with 25 % solid volume, which gives 5.8 kN (600 kg). The weight of the part will adversely affect the accuracy, the axis acceleration, the natural frequencies of the structure and the servos, and the cost of the pallet changer. The maximum part weight of 5.8 kN (600 kg) will be a soft specification realizing that some customers will manage to exceed this value. Bearing systems, for example, must be designed to carry much higher weights.

9.1.5 Ranges of Travel

As Figure 9-1 shows, the ranges of travel chosen for the X, Y and Z axes are 750, 650 and 600 mm, respectively. These are somewhat smaller than the ranges for the Maxim 500 (750, 700 and 750 mm) and for the Ram 500 (750 mm on all axes). The Z-axis range allows a boring operation through a 500 mm part and allows a 200 mm tool to reach the center of the pallet. The X and Y ranges allow a facing operation over the maximum sized part. The range for the B axis or rotary table is unlimited. The range for the A axis, which should be considered as an optional axis, is highly dependent on the design solution and is arbitrarily chosen at $\pm 30^{\circ}$.

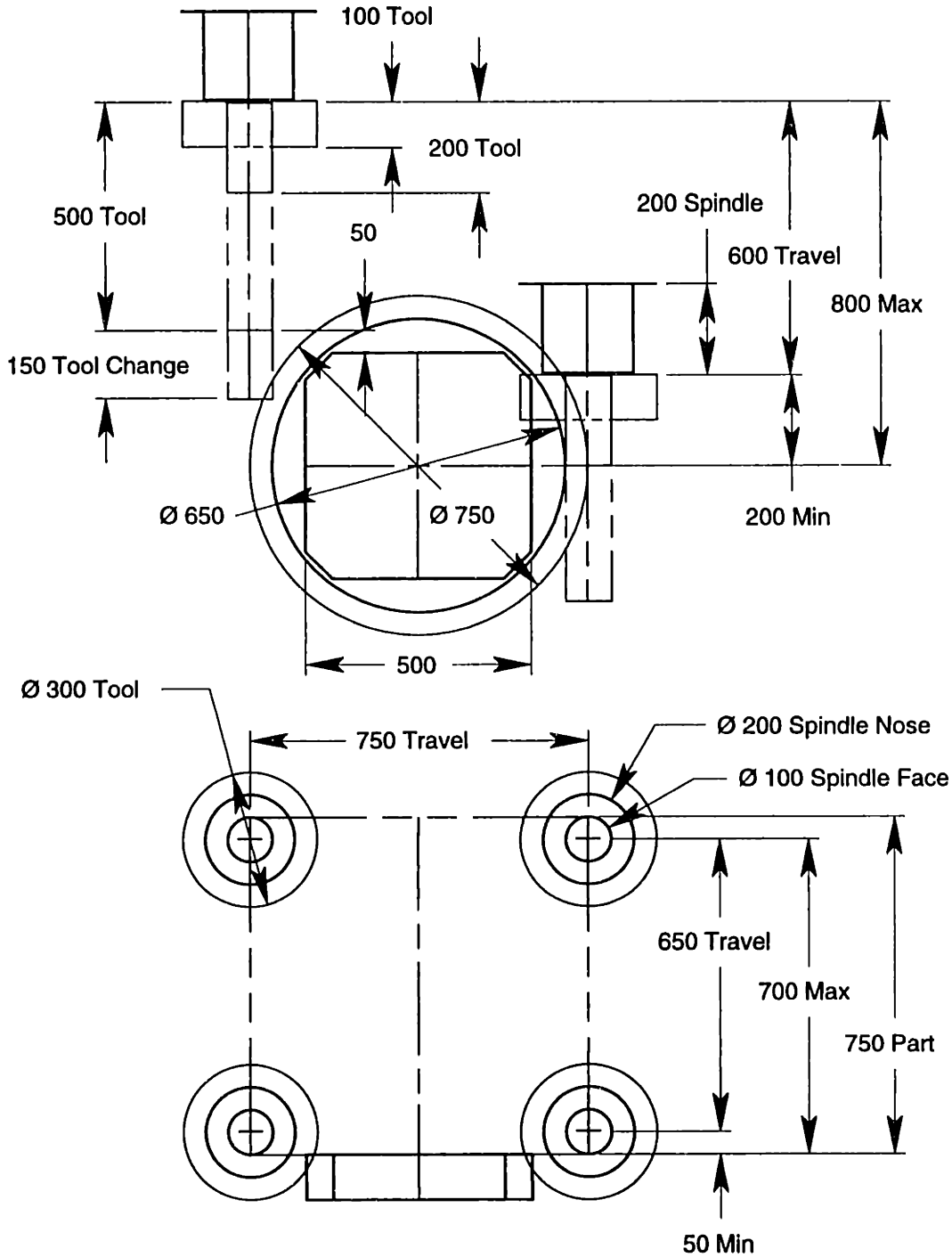


Figure 9-1 Range drawing showing the ranges of travel for the X, Y and Z axes.

9.1.6 Volumetric Accuracy

The ability to produce accurate work depends on the ability of the machine to control the relationship between the tool and the workpiece and on the cutting process, which influences both the machine and the workpiece. Therefore, the accuracy specification must also contain a corresponding task specification to be meaningful. The accuracy most often cited for machine tools is linear accuracy, which only applies to the line where the axis

position was mapped with a laser interferometer. In addition, there is no tool load or process variation included in the linear accuracy specification. The repeatability of a machine is also commonly specified as an accuracy, but this is only an indicator of the potential accuracy using a suitable compensation system such as post-process gauging or volumetric error mapping. The repeatability specification applies to the tool-to-work relationship for any given position in the volume approached from a given direction at a given speed and without a tool load or process variation. Seldom will the repeatability specification include the very real dimensional changes to the machine due to changing temperatures and gradients. Under these very special conditions, the repeatability of a machine with sufficiently low hysteresis should approach the resolution of the feedback device, and the linear accuracy should approach the repeatability.

The Maxim 500 has a specified linear accuracy of ± 3 μm and a specified repeatability of ± 1 μm using optional 1 μm linear-scale feedback. The Ram 500 has a specified linear accuracy of ± 2.5 μm and a specified repeatability of ± 1 μm. Our data on the Maxim suggests that these specifications are fairly realistic, however they have almost no functional significance regarding the accuracy of a workpiece. The volumetric accuracy gives a much better description of the machine accuracy and includes the effects of all error sources except those that are a direct consequence of the particular process or operation (for example, tool loads, inertial loads, process heat, theoretical surface finish and the tool geometry). The included error sources are geometric errors, hysteresis and contouring errors, variable part weight, internal heat sources that move or vary and environmental influences such as changing temperature.

To define volumetric accuracy, imagine a test where the machine is programmed to follow an arbitrary trajectory throughout the work volume for a specified set of conditions such as environmental temperature, part weight, etc. All along the trajectory, measurements are taken of the position and orientation of the spindle axis average line relative to the pallet. After removing setup errors with a rigid body translation and rotation of the measured data, the difference from the programmed trajectory is the volumetric error. The angular volumetric error would simply be the spindle squareness error if there are no angular moves. The squareness error could be large if the spindle were not parallel to the Z axis of the rotated coordinate system. The B5.54 standard recommends the body diagonal test and the circle test using a telescoping ball bar to measure positional volumetric accuracy. Parametric error mapping is more rigorous and expensive but gives a complete description of volumetric accuracy. A sufficiently distributed and redundant set of length measurements using the laser ball bar, for example, provides a more cost effective measurement of volumetric accuracy. This is the basis for the rapid volumetric error mapping procedure developed at LLNL [Krulewich, Hale and Yordy, 1995], [Krulewich, 1998].

The body diagonals measured for unidirectional, point-to-point moves indicate that the Maxim could have a volumetric accuracy of 6 μm (± 3 μm) with error compensation and

well-controlled environmental temperature. Hysteresis present in the slides would degrade the volumetric accuracy of bi-directional, point-to-point moves to approximately 10 μm. Contouring moves were not measured. A factor in such good results was the machine remaining thermally stable from the time it was mapped to the time the body diagonals were measured. This raises an important point about the stability of the machine, the error map and our definition of volumetric accuracy. A change or drift in the metrology loop of the machine that occurs over the short duration of a test or a part cycle directly affects the volumetric accuracy. Take for example axial spindle growth that is most significant at startup. A change in growth affects the flatness of an end-milled surface, but once stabilized, the only effect is a shift in the location of the surface relative to the work-holding fixture. This error is correctable by probing the surface and a reference on the fixture with a tool-mounted probe, or by checking the end mill in a tool set station. Only the change in growth falls within our definition of volumetric accuracy. A long-term drift occurring over many part cycles affects primarily the accuracy of the error map. The compensation that results will be in error but this is easy to fix by periodically re-mapping the machine. This may entail a potentially simple and rapid exercise using a tool-mounted probe with a calibrated artifact or a length-measuring mapping procedure.

The value of high accuracy depends on the part and the types of features being produced. Bored holes typically require tight control on position, parallelism, squareness of shoulders, etc. This type of operation allows the use of unidirectional point-to-point trajectories for greater accuracy. Probing is another high-accuracy operation where the trajectory is more general. From a marketing stand point it would be a great advantage to take credit for as much accuracy as the machine will deliver for any particular operation, trajectory, environmental condition, etc. Imagine a virtual machining program, perhaps running on the salesperson's portable computer, that would load the customer's part file and environmental parameters, generate an optimized part program, calculate the production rate and estimate the errors in the virtual part based on actual performance data.[1] A reasonable step in this direction is an error budget, which provides a breakdown of errors for different conditions and operations. The accuracy specification then becomes a table of numbers that the customer can use to evaluate the machine for particular types of operation and environmental conditions.

Table 9-5 shows the accuracy specification selected for the conceptual machining center. It is based on the error budget for the Maxim including geometric and limited thermal error compensation and reasonable improvements to the design of the machine. Eventually a table like this would comprise test data gathered from prototype experiments and/or standard runoff tests.

---

[1] [Frey, 1997] describes a mathematical framework and computer simulation that relates known error profiles through the machine's kinematics and a specified part program to obtain a virtually machined surface.

Volumetric Accuracy for Optimal Conditions	X,Y,Z Trans. (uncomp'ed)	X,Y,Z Trans. (x,y,z comp)	Spindle Sq. (uncomp'ed)	Spindle Sq. (x,y,z comp)	Spindle Sq. (a,b comp)
Feedback Resolution	0.5 μm	0.5 μm	n.a.	n.a.	2 μr
Geometric Comp. Accuracy	n.a.	5 μm	n.a.	20 μr	10 μr
Unidirectional Positon Accuracy	96 μm	6 μm	72 μr	22 μr	12 μr
Bidirectional Position Accuracy	98 μm	8 μm	76 μr	26 μr	16 μr
Contour Accuracy < 0.5 m/s^2	100 μm	10 μm	80 μr	30 μr	20 μr
Additional Errors for Suboptimal Conditions	X,Y,Z Trans. (uncomp'ed)	X,Y,Z Trans. (x,y,z comp)	Spindle Sq. (uncomp'ed)	Spindle Sq. (x,y,z comp)	Spindle Sq. (a,b comp)
Day/Night Temp. Cycle, 10 C	25 μm	5 μm	10 μr	10 μr	10 μr
Spindle Duty Cycle, 5000 rpm	50 μm	5 μm	5 μr	5 μr	5 μr
Axis Duty Cycle, 500 ipm	5 μm	5 μm	5 μr	5 μr	5 μr
Part Weight, 5 kN (1120 lbs)	10 μm	2 μm	5 μr	5 μr	5 μr
Tool Load, 500 N (112 lbs)	10 μm	10 μm	10 μr	10 μr	10 μr
High Acc. Contour, 5 m/s^2	20 μm	20 μm	20 μr	20 μr	20 μr

Table 9-5 Volumetric accuracy specification (total band) for various tasks and conditions. The uncompensated accuracies are before compensation. An uncompensated machine would require better mechanical accuracies than specified. The compensation assumed here would compensate for geometric errors, part weight, quasi-static operating temperature (ΔCTE) and z spindle growth errors.

9.1.7 Static and Dynamic Stiffness (or Compliance)

The machine requires sufficient stiffness in the structural loop between the tool and the workpiece so that variable loads result in insignificant deflections. The entire structural loop is exposed to process loads while other types of loads may act on portions of the structural loop. While stiffness may be more intuitive, working with compliance is more convenient since the loop is serial in nature and most tests report compliance. The term *dynamic compliance* implies a frequency variation but generally only the resonance peaks are of interest. Lower resonant peaks (indicating greater damping) and higher resonant frequencies equate to smaller amplitudes and better finish, and allow greater servo stiffness. Quasi-static loads, such as the reactions calculated from the machining process model or inertial loads from accelerating masses, act on the static compliance to generate position errors. Thus for high accuracy, good finish and heavy cuts, the machine requires low static compliance, well-damped resonances and high resonant frequencies.

The structural loop includes the tool, the workpiece and fixturing in addition to the compliance due strictly to the machine. The non-machine components contribute significant compliance and cannot be dismissed from the specification. We will assume that the specification refers to a relatively stiff arrangement where the machine contributes two-thirds to three-fourths of the total compliance. In the compliance budgeting process, the finite element model should include reasonable models for the tooling. Compliance testing usually includes brackets and fixtures similar to stiff tooling so the results should be consistent with the model. Still, the placement of the load (the tool length and the height above the pallet) can significantly affect the measured compliance.

The machining process model provides an estimate of the process stiffness for a particular set of cutting parameters and the material shear strength. The process stiffness

represents the force variation that occurs due to a change in the uncut chip thickness. Dynamic instability can occur because there is a feedback loop formed by undulations left from the previous cut affecting the dynamics of the present cut through the chip thickness and the process stiffness. Like too much gain in a feedback system, regenerative chatter can occur when the process stiffness is too large compared to the loop stiffness of the machine. How large is a complicated question. We attempted to reach the chatter limit of the Maxim with relatively heavy face milling cuts in tough steel. The cuts generated much noise and vibration but no apparent self-excited chatter. The calculated process stiffness was 0.278 lb/ μin (or 3.6 μin/lb as a compliance), which is much stiffer than the resonance that the test would excite. The theory would indicate that the Maxim is too compliant dynamically to perform this cut but obviously this is false. The discrepancy is probably due to a very complex problem being analyzed too simply.

The process stiffnesses calculated for the machining study are shown in Table 9-6. With a couple of exception, the Maxim should be stiff enough to perform these operations based on our cutting test. In particular, slotting and slab milling tough steel may be tooling limited rather than machine limited. Using the Maxim as a reference point, we will specify the highest dynamic compliance peak at 10 μin/lb for a 4 inch (100 mm) tool length and 10 inches (250 mm) above the pallet. This is consistent with the Y direction compliance measured on the Maxim but represents a significant improvement in the X direction. Achieving this goal may require damping treatments to the column and spindle.

Material	Units	1" Slot	2" Slab Mill	4" Face Mill	8" Face Mill
4140-300 HB	lb/ μin	0.352	0.654	0.194	0.194
Steel-140 HB	lb/ μin	0.160	0.297	0.088	0.088
6061-T6	lb/ μin	0.29	0.126	0.029	0.029

Table 9-6 Process stiffnesses calculated for the machining study.

The static compliance was given indirectly in Table 9-5 as a tool point error of 10 μm for a tool load of 500 N, which equates to 20 nm/N or 3.5 μin/lb. It represents a reasonable improvement over the Maxim. Table 9-5 also specifies a tool point error of 20 μm for an axis acceleration of 5 m/s^2 or approximately one-half *g*. A useful figure of merit for an accelerating structure is the lowest natural frequency in the direction of travel (derived from the deflection caused by an inertial load acting on a mass-spring system). As Figure 4-16 shows, an 80 Hz mass-spring system will deflection 20 μm under a one-half *g* acceleration. This specification will drive the configuration of the conceptual machining center to be substantially different from the Maxim.

9.2 Design Strategies

The strategies for satisfying the design goals should clearly guide the design team to the basic technologies to use and how to approach key problems. This section is particularly important to this study because many of the ideas may not show up or be obvious in the conceptual design drawings.

9.2.1 Accuracy

The basic approach proposed for this concept design is to obtain accuracy by designing for mechanical repeatability and long-term stability and to use software-based volumetric compensation to correct for repeatable errors. This is the same approach used to enhance the accuracy of the Maxim with good success. Still, people criticize compensation saying that it does nothing for angular errors. This is only partly true assuming that the machine does not have A and B angular axes. Squareness errors are static angular errors that are compensated by a simple coordinate rotation (except B-C squareness) or corrected by mechanical alignments. Axis compensation significantly reduces sine errors that result from angular errors acting through Abbé offsets. Volumetric compensation reduces to a minimum those sine errors associated with the x-y-z position and the tool length but not the tool diameter as this requires angular axes. In this respect, the gain in using volumetric compensation, assuming neither A nor B axes, is approximately the size of the work volume compared to the diameter of a critical shoulder or face. For many customers, volumetric error compensation will provide a significant improvement in accuracy for little or no additional cost if it eliminates costly mechanical alignments.

The accuracy of a machine, whether software compensated or mechanically corrected, is fundamentally limited by the repeatability of the axes and the stability of the metrology loop. Friction, backlash and changing temperatures are the main sources of nonrepeatability and instability. Test results on the Maxim indicate that its ball screws, linear guides and servo system are adequate regarding repeatability. Friction present in other devices such as the counterbalance and the X-Y shield are more serious especially when the friction acts away from the actuator as to produce a frictional moment that results in a sine error. The best approach is to eliminate sliding contact wherever possible. For example, the counterbalance might use a hydrostatically balanced labyrinth seal or a vacuum bellows.[1] When friction is unavoidable, it should act close to or symmetrically about the actuator. To provide greater repeatability and robustness to tool loads, the axis should have widely spaced bearings for greater moment stiffness and a minimum lever arm between the tool and the position feedback device. To provide greater dynamic stiffness and to reduce dynamic moments, the actuator should act through the center of mass.

---

[1] Eliminating the counterbalance is an option, but there is evidence that the servo system has greater force ripple when it operates under a significant load.

Preloaded rolling-element bearings such as ball screws, thrust bearings and linear guides have zero backlash, low frictional hysteresis and relatively low heat generation. Elastic averaging over a large number of contact surfaces results in high load capacity, high stiffness and smooth repeatable motion. Rolling-element bearings have a finite lifetime that depends on loads, cycles and contamination in particular, but they require nearly zero maintenance until failure. By comparison, hydrostatic bearings offer zero static friction, much higher damping, resistance to contamination and potentially infinite life. Water-based hydrostatic bearings provide an attractive alternative to rolling-element bearings, especially for precision grinding applications. Research is currently underway at MIT[I] to develop modular hydrostatic bearings that are interchangeable with industry-standard linear rolling-element bearings. As these become available, they will provide a reasonably simple upgrade path to greater precision. The cost and maintenance associated with the pumping system remain a disadvantage but not a limitation. Currently, the ball screw is the best actuator for this application so it makes sense to use the same technology throughout the machine.[II] Therefore, we will concentrate efforts on achieving the greatest repeatability and stability using rolling-element bearing technology.

The ball screws on the Maxim are pretensioned with thrust bearings at each end. The thrust bearings consist of a triplex set having two bearings acting in the tension direction and the third bearing in opposition. Initially, a combination of too much pretension and misalignment between the nut and the spindle carrier caused considerable heat generation and thermal stability problems in the column. These problems were corrected through mechanical adjustments. In contrast, the G&L Ram has a main thrust bearing and a secondary thrust bearing that is hydraulically coupled to relieve low-frequency axial loads such as thermal growth. The more traditional approach is to use one thrust bearing and one support bearing. The basic approach used on the Maxim has distinct advantages that we wish to retain for this concept design. A screw supported at each end has nearly constant axial stiffness throughout the range of travel compared to the traditional approach. Constant dynamic characteristics contribute to better servo performance. Another advantage for non-scale feedback machines is that thermal expansion is approximately one-half as great and centered in the middle of the work volume. The main disadvantage is the axial load that thermal expansion of the screw creates on the thrust bearings, which produces even more heat. The axial load also distorts the frame but we found this to be less significant for the Maxim column than the heat produced by the lower thrust bearing. G&L attacked the thermal growth problem with a hydraulic coupling. However, the direct

---

[I] Professor Alex Slocum, MIT Rm 35-010, 77 Massachusetts Ave, Cambridge, MA 02139, phone: 617-253-0012, email: slocum@mit.edu

[II] Since the ball screw must be protected, lubricated and sized for life, rolling-element linear guides can enjoy the same protection, lubrication and life.

approach, reducing the heat generated and controlling the temperature of the screw and thrust bearings, provides the best stability.

We will consider two different ball screw configurations having supports at both ends. The first is a rotating screw with thrust bearings at each end like the Maxim. The second is a non-rotating screw with rigid end supports and a rotating nut supported by thrust bearings. Since the second configuration is uncommon, a prototype should be built and tested before committing to this design. Heat will be reduced by minimizing misalignment, increasing the tolerance to misalignment and using the minimum number of bearings consistent with the maximum load and balanced compliances. Temperature-controlled liquid flowing over or through the screw will provide thermal stability and minimize the differential axial expansion between the screw and the frame.

Ideally the assembled ball screw, ball nut and thrust bearings should operate with small residual (moment and radial) loads compared to the built-in preload and the anticipated axial load. Severe misalignment leads to premature failure, excessive heat generation and significant error motion especially near the ends of travel where the loop is stiffest. A stress-free assembly requires: 1) the axis of the screw and the travel of the slide to be parallel; 2) the bores for the thrust bearings to be centered and parallel to the screw; 3) the thrust bearing faces to be square to the screw; 4) the nut flange to be square to the screw; and 5) the screw to be set for zero tension at the average operating temperature.[I] The flange mount naturally allows for centering the nut. There are several ways to achieve these requirements. The traditional method requires skilled hand labor to fit by scraping or by grinding fitting spacers. Where possible, use planar and/or spherical mounting surfaces to allow simple adjustments. Requirements 1, 2 and 3 can be machined directly into the frame using a line-boring operation set up parallel to the slideways.[II] A second way is by replicating bearing cartridges to the frame using an alignment fixture. To increase the tolerance to misalignment, bearing arrangements should be face-to-face to provide greater angular compliance, and the screw should have extra length at each end to provide greater bending compliance.[III] An alternative to extra length is to neck down the screw somewhat near the bearings. Either approach must be examined for possible buckling or whirl instability.

[I] This set of requirements is written toward a rotating screw configuration. The non-rotating screw configuration requires similar alignments worded somewhat differently.

[II] The two fixtures that support each end of the arbor are themselves line bored to provide identical spacing from the way surfaces.

[III] Face-to-face duplex bearings are known to be thermally unstable at high speed because greater axial growth of the rotor leads to greater preload, heat generation and more growth. This is not anticipated to be a problem for the size of bearing, speed and duty cycle of the screw. Axial growth of the screw between two thrust bearings is of greater concern and is one motivation for controlling the temperature of the screw.

An obvious way to control the temperature of a screw is with a shower of temperature-controlled lubricating oil. Collecting the runoff and preventing contamination are key issues. In addition, too much oil in a rolling-element bearing leads to viscous heating at high speeds. It would be less messy to flow water-based coolant through the center of the screw. This is trivial for the non-rotating screw since it does not require a rotary coupling and both ends are accessible. For a rotating screw, the connection would be at the end opposite from the motor. A square or fluted tube on the inside of the screw would transport the water to the motor end and then back to the rotary coupling. However, it may be sufficient to control the temperature around the thrust bearings and the nut and let the screw come to a higher operating temperature. These sorts of issues and decisions are best determined through testing.

The following analysis shows the characteristics of the two proposed ball screw configurations, where each has the total compliance equally distributed among the screw, the nut and the thrust bearings. Figure 9-2 shows the parameters used in the first model. For this example, the screw is steel, the diameter is 40 mm and the length is 1 m. Equation 9.1 gives the pertinent relationships in the model. For the rotating screw configuration, Figure 9-3 shows the total compliance as a function of the nut position, which is fairly constant and approximately three-fourths of the full-length screw compliance. Stiffer thrust bearings would reduce the total compliance but the relative variation would increase. Figure 9-4 shows that each thrust bearing shares differently in the axial load depending on the nut position. The difference is reduced by operating further from the ends or by lowering the stiffness of the thrust bearings. A temperature differential of 1° C will generate a force in the thrust bearings of 1.5 kN, which is 12% of the specified X or Y axis force. Doubling the thrust bearing compliance would increase the total compliance by one-third and would reduce the force by one-third. Figure 9-5 shows the total compliance for the non-rotating screw. The variation is greater but still is very acceptable. For the same temperature differential, the force in the screw would be twice as great but it would not react against the thrust bearings.

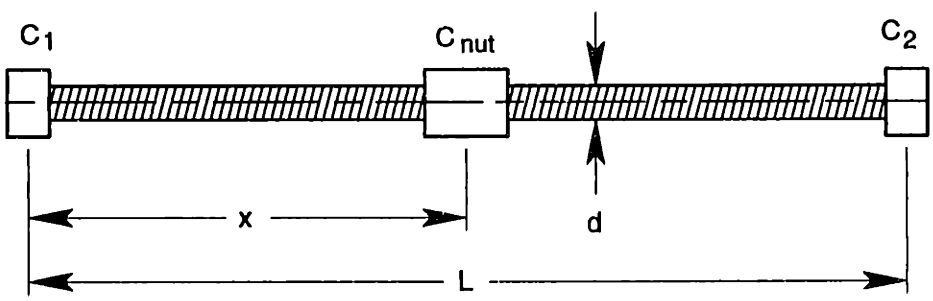


Figure 9-2 Parameters used in the model to describe the compliances of the screw, nut and thrust bearings.

$$C_{s1_i} := C_1 + C_{screw}\cdot\frac{x_i}{L} \qquad C_{s2_i} := C_2 + C_{screw}\cdot\frac{L - x_i}{L}$$

$$C_{total_i} := \left[\left(C_{s1_i}\right)^{-1} + \left(C_{s2_i}\right)^{-1}\right]^{-1} + C_{nut}$$

$$F_{1_i} := \left[C_{s1_i}\cdot\left[\left(C_{s1_i}\right)^{-1} + \left(C_{s2_i}\right)^{-1}\right]\right]^{-1}$$

$$F_{2_i} := \left[C_{s2_i}\cdot\left[\left(C_{s1_i}\right)^{-1} + \left(C_{s2_i}\right)^{-1}\right]\right]^{-1}$$

$$F_{\Delta T} := \frac{\alpha\cdot L}{C_1 + C_2 + C_{screw}} \qquad F_{\Delta T} = 1.521\cdot 10^3\cdot\frac{newton}{C} \tag{9.1}$$

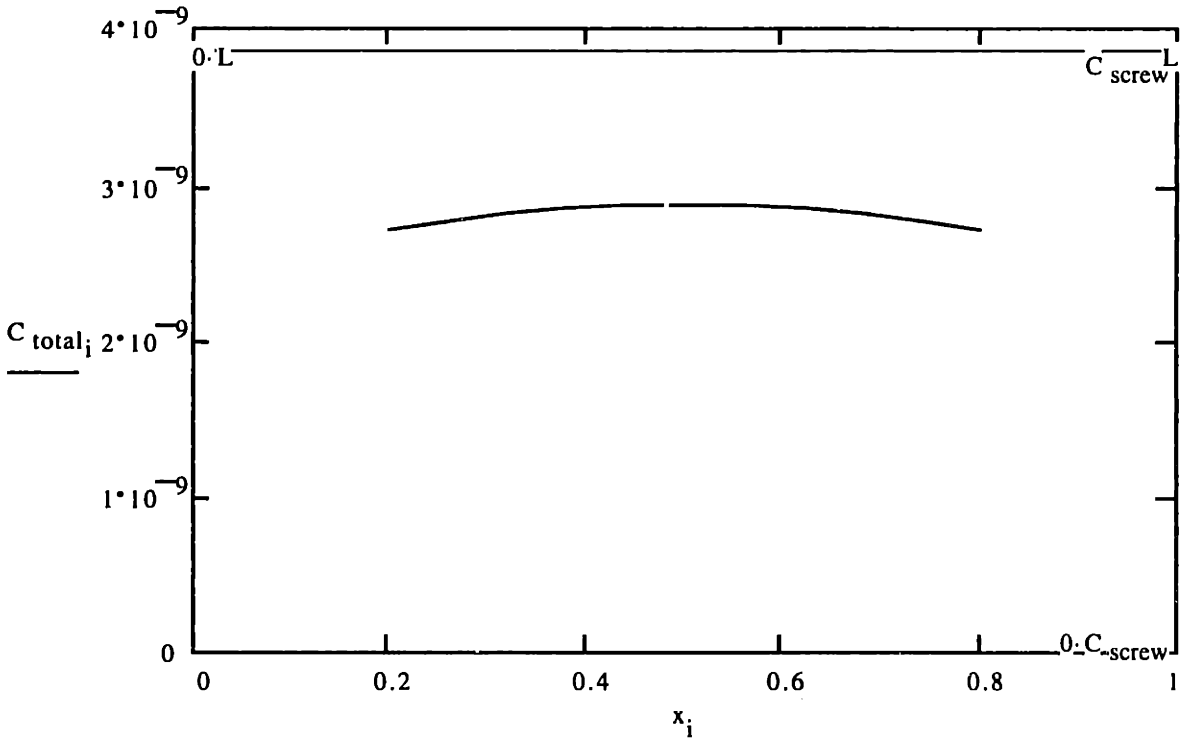


Figure 9-3 The total compliance in meters per Newton is a function of the nut position in meters. The screw is 40 mm in diameter by 1 m long. The axial compliance of the thrust bearing at each end of the screw is equal to one-half the full-length screw compliance. The axial compliance of the nut is equal to one-quarter the screw compliance.

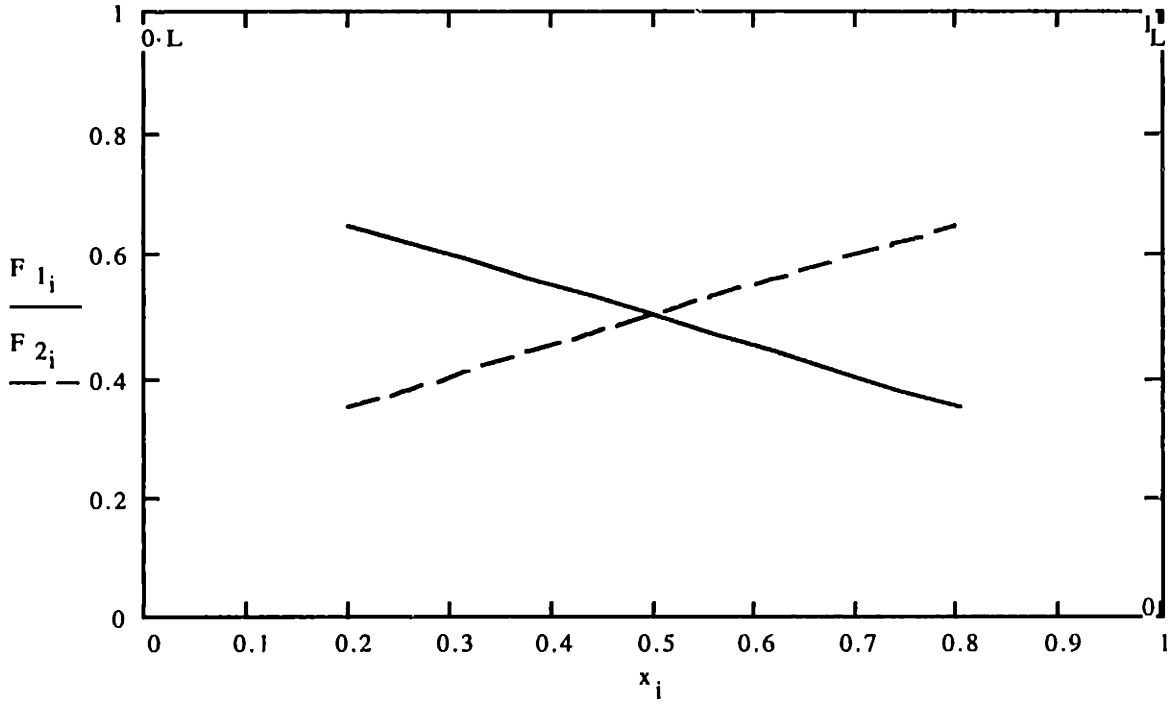


Figure 9-4 Reaction force at each thrust bearing for a unit axial load applied to the nut is a function of the nut position. The same conditions apply as described in Figure 9-3.

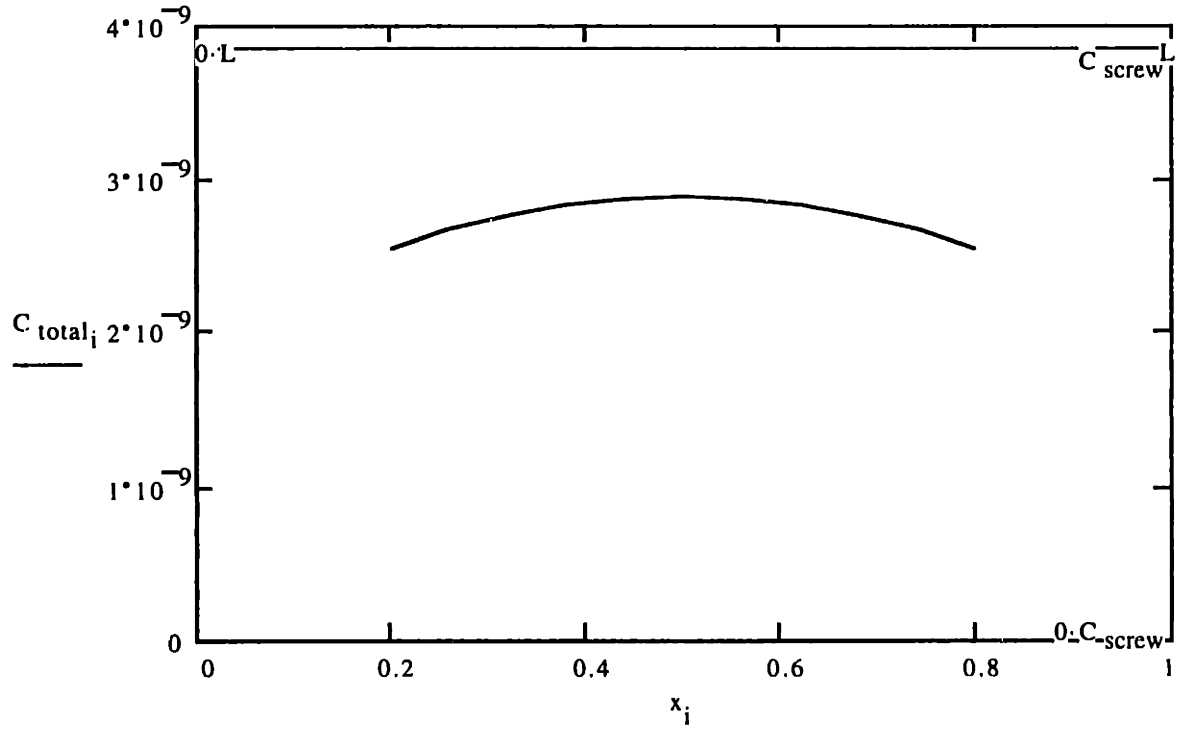


Figure 9-5 The total compliance in meters per Newton is a function of the nut position in meters. The screw is 40 mm in diameter by 1 m long. The axial compliance of the thrust bearing supporting the nut is equal to one-quarter the full-length screw compliance. The axial compliance of the nut is equal to one-quarter the screw compliance.

Misalignment of overconstrained linear guides is less of a thermal problem because the speeds are usually much lower than ball screws and the heat spreads out over the frame. This condition still causes unnecessary loads that may lower bearing life. Linear guides seem to be tolerant to misalignment and overconstraint as evidenced by so many such designs. The equivalent sliding-bearing system would have excessive clearance in some places and would stick in other places. Hydrostatic bearings could touch down with the small clearances involved if the structures are too stiff to conform to misalignment. The tolerance to misalignment in a linear-guide system results from a combination of compliances in the bearings and the structures, low friction and accurate geometry typical of machine tools. Applying ideas from exact constraints to release overconstraint in linear-guide slide systems can improve the accuracy and life for potentially lower cost.

Linear guides are available with rollers or with balls as rolling elements. In general, balls will provide greater tolerance to misalignment and slightly better precision, while rollers will provide greater load capacity and stiffness. The frictional characteristics will be similar unless the balls run in a gothic arch groove. As a rule, gothic arch designs should be avoided due to higher friction and wear unless it carries sufficient load as to cause the balls to roll against only one side of the arch.

A single linear guide usually constrains three rotational degrees of freedom, which ideally should be free to approximate exact-constraint behavior. Because the geometry of the rails must be relatively straight for machine accuracy and because the bearing blocks will be widely spaced for moment stiffness, angular error motion will be relatively insignificant. Initial misalignment of the mounting surfaces to the bearing blocks will be the only significant concern. The pitch and yaw directions are more critical than the roll direction since the bearing block is longer than the width of the rail. This is also consistent with our ability to control the tolerances in these directions. Since shims could foul these angular alignments, the control of squareness between axes (for example the X-Y squareness) should be machined into the mounting surfaces rather than shimming. Instead of shims, it is more cost effective to control the tolerances during machining by the timely feedback of inspection information and repair if necessary. If adjustments at assembly are deemed essential, then the alternatives are replicating or scraping the structure to match the bearings, or designing in adjustment interfaces elsewhere.

The remaining modes of overconstraint are in-plane and out-of-plane parallelism of the rails and bearing blocks. The rails must be held parallel and straight to the order of 5 μm. There are several approaches to controlling the in-plane spacing of the rails. One depends on the mounting surfaces being manufactured sufficiently straight and parallel to function as the guides for both rails. A light lapping operation is effective in removing localized machining imperfections that could affect the rail alignments. In another approach, one rail is set to a sufficiently straight mounting surface and becomes the master used to set the second rail. This is done by translating a pair of bearing blocks (one on each rail set to a

fixed separation) just ahead of the fastener being tightened on the second rail. This technique eliminates the small tolerance stack up inherent in the rails. A third approach requires a master gauge to set the straightness and parallelism of both rails in one step. This gauge could also set the squareness with respect to another axis. In this way, precision is put into the gauge and replicated to numerous machine tools. Parallelism errors that remain and thermal expansion are accommodated either by making the structure compliant in this direction (perhaps three to four times the bearing compliance) or by using one linear guide with freedom in this direction. For example, a gravity load causes the linear guides in Figure 9-6 to approximate a vee-and-flat slideway.

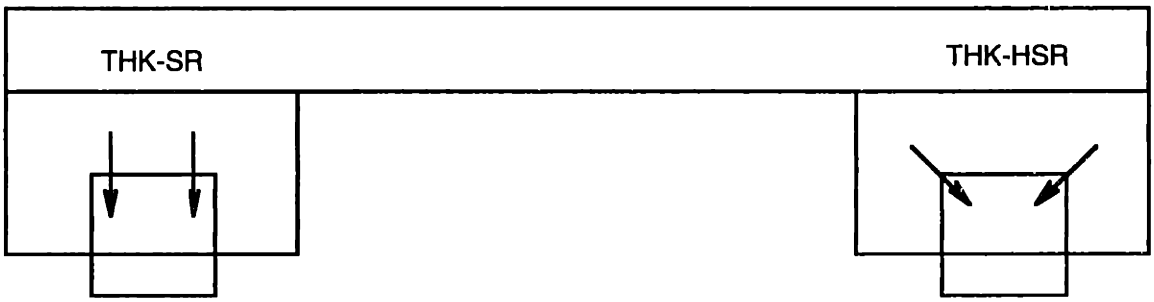


Figure 9-6 A linear guide designed primarily for vertical load capacity has greater horizontal freedom under load. This arrangement improves the tolerance of the bearings to in-plane parallelism misalignment.

Out-of-plane parallelism is typically more difficult to achieve but the sensitivity can be greatly reduced by using three bearing blocks to define the plane. This requires the structures to be independently rigid in torsion, which may be very difficult to achieve. Four bearing blocks may work better when the carriage has four-fold symmetry or has insufficient torsional stiffness. In that case, it is advantageous to make the carriage torsionally compliant to relax somewhat the overconstraint between it an the four bearings blocks. This is a case where shimming can relax misalignment. Each slide design will require analysis to determine the compliance required to achieve a tolerable bearing load due to expected alignment and operating errors.

9.2.2 Thermal Stability

The international standard temperature, where a solid object has its true size, is 20° C or 68° F. This implies that the temperature distribution must be constant and uniform throughout the workpiece. Metrology at any other temperature requires compensation that depends on the temperatures and the particular coefficients of thermal expansion (CTE) for the workpiece and the measuring machine. Suppose for example that the workpiece is steel with a CTE of 11.7 με/°C ± 10% and the machine has a slightly different effective CTE, say 10 με/°C. The error that would exist at 30° C or 86° F is 8.5 ± 5.9 μm for a 500 mm part. If instead the part were aluminum with a CTE of 23.6 με/°C ± 10%, the error would be 68 ± 12 μm under the same conditions. These calculations were made assuming the temperatures of the workpiece and the machine were identical and steady, which with care

can be approached. Temperature compensation under these conditions is straightforward and reasonably accurate if the CTE's are well known, as cautioned by [Bryan and McClure, 1967].[1] Although the proper solution is to maintain both the workpiece and the machine at a steady 20° C, it may not be the solution that the customer is willing to buy. Therefore, we will consider strategies for various levels of customer commitment to temperature control.

The workpiece may deviate from standard by being at a steady uniform temperature other than 20° C, which affects its size, or by having a steady or transient temperature gradient, which affects it shape. It is currently impractical to compensate for temperature gradients in the part. Significant gradients would occur if a warm part, say 30° C, were placed in a flood of 20° C coolant with insufficient time to thermally stabilize or if heavy cuts were made without the benefit of coolant. This situation may require the workpiece to cycle through a thermal wash, perhaps at the part load station. The machine may also deviate from standard in the same ways but the consequences are different. In some ways the problem is less severe because the machine can be at any steady temperature with any steady gradient as long as the state is identical to when the machine was mapped and the stability of the part is unaffected. In addition, the next development step in volumetric error compensation and rapid mapping using the laser ball bar will include transient thermal errors. In other ways, however, the problem is more severe because the machine is physically much larger than the workpiece and contains varying and moving heat sources. The ability of thermal compensation to deal effectively with the whole problem of thermal stability is questionable and certainly comes with more risk and development effort than the direct approach, which is to control temperatures and gradients within acceptable limits. As a basic strategy, we will use thermal compensation only where it is reliable and easy to implement, namely, for size scaling based on CTE and operating temperature, and for well-defined localized phenomena such as axial spindle growth. We will use the direct approach to ensure that the machine and the workpiece are operating close to the same uniform temperature that is not necessarily 20° C or steady.

The level of environmental temperature control that the customer might supply varies from a steady 20° C to virtually nonexistent. A customer at the latter extreme may not care as much about high accuracy but certainly about better accuracy. The goal of temperature control as stated above is to maintain the workpiece and the machine at a uniform temperature, which indicates our preference to keep all gradients small rather than

---

[1] According to Richard Kirby of the National Bureau of Standards, "The accuracy of a tabulated value of a coefficient of thermal expansion is about ± 5% if the heat and mechanical treatment of the steel is indicated. The precision of the coefficient (a) among many heats of steel of nominally the same chemical content is about ± 3 %, (b) among several heat treatments of the same steel is about ± 10%, and (c) among samples cut from different locations in a large part of steel that has been fully annealed is about ± 2 % (hot or cold rolling will cause a difference of about ± 5 %)." From [Bryan, 1984].

strictly constant. For a well-controlled environment near 20° C, the best strategy is fairly clear; control the part coolant to 20° C and use it to stabilize the workpiece and to capture heat from local internal sources. The remainder of the machine tends to follow the controlled air temperature, which has a surprisingly strong influence even on machines as massive as the Maxim. This strategy applies to any steady temperature other than 20° C if the part coolant is controlled to the same temperature and the machine is mapped at that temperature. This strategy fails if the environmental temperature varies because some parts of the machine change while other parts and the workpiece remain stable.

The question is whether to make the workpiece and the machine follow the changing environment or to add temperature control to the entire machine (or at least the entire metrology loop). Since it is impossible to achieve thermal time constants that are the same for all critical parts in the metrology loop, any rate of temperature change will cause gradients and non uniform thermal expansion. Therefore, a limit exists where the environmental temperature changes too fast for the application and the only solution is complete temperature control of the machine. In addition, there is the uncertainty in the CTE of the part and the machine, and the possibility of thermal hysteresis in the machine structure. We may decide on a case-by-case basis to track the environmental temperature; this requires experimentation. However, the baseline strategy must be to stabilize the whole metrology loop and workpiece to a steady temperature, preferably 20° C. This will require additional equipment such as air conditioned enclosures, water-cooled heat exchangers built into the structures and/or water-cooled panels surrounding the structures. It may turn out through some clever designs that this cost is not so significant compared to the money that the customer would spend for a stable environment.

9.2.3 Structural Stability

Structural stability means that the metrology loop must remain invariant through time to obtain accuracy either by error compensation or geometric correction. Thermal stability is an aspect of this requirement that was previously addressed. Here we are concerned with eliminating variable loads on the metrology structures and/or making those structures robust to ones that are unavoidable. Short-term load sources are moving and variable masses within the system, friction in slides, over constraints between moving components, inertial loads and process loads. Long-term load sources result from movement in the foundation due to settling or hygroscopic action and whenever a friction joint slips and releases built-up stress.

The separation principle requires that variable forces be eliminated from the metrology loop; however, this is difficult to apply to the machining center in the purest sense. A separate metrology frame and system of sensors become awkward for three or more axes and the expense is unwarranted for this level of accuracy. However, we can apply the principle in other ways. For example, the effect of a moving mass can be

separated out through compensation if the mass is constant or if an accurate model exists and the change in mass is measured. In applying the separation principle, the first step is to identify the metrology loop and the parts within it. The parts list for the Maxim includes the main structures (base, column, etc.), the linear guides, the scales, the spindle, the tooling, the workpiece, the fixtures, the pallet, and due to overconstraint, the foundation and 14 leveling screws.[I] The next step is to identify and classify the variable loads as separable, repeatable (or compensable) and nonrepeatable. Large separable loads or ones that are easy to separate should be removed from the metrology loop while others that are small relative to the process load may be ignored. Compensation can reduce the effects of repeatable loads but simply stiffening the structures may be the better solution since the effects of nonrepeatable loads are also reduced.

The machine base is the primary metrology structure for the Maxim and most other machine tools. It supports and is influenced by the weight of moving and non moving structures, the part weight and variable loads such as process and inertial loads. The undesirable influence from foundation movement is mitigated by eliminating overconstraint in the machine support. The simplest approach is to use three support points aligned with gravity and incorporating elastic or viscoelastic interfaces to avoid frictional hysteresis. The number of supports can be increased up to six (the number required for exact constraint) to distribute more uniformly the weight of the machine but this requires the constraints to be angled appropriately. The additional support points will not affect the stiffness of the machine base but they may affect the mode shapes and the degree of damping that can be gained from viscoelastic supports.[II] For considerably larger machines that may require multiple base sections, each section should have exact constraint support relative to the foundation and adjacent sections so that movement coupled from the foundation is detectable using sensors between sections. This information can be used to calculate compensation or to automatically level the machine with appropriately placed actuators.

The conceptual machining center will have a single-piece base with sufficient stiffness so that uncompensable deflections are insignificant. Any significant weight that does not have a role in the metrology loop such as the tool chain or the part loading station will be separated by supporting the weight directly on the foundation and by coupling compliantly to the machine base. The placement of supports for the base will minimize the deflections due to moving and variable masses that are not separable. Viscoelastic material incorporated in the mounts will provide passive damping and also will reduce frictional

---

[I] The objection to many supports applies only to the metrology loop. A structural frame with many supports may work well if there is a separate metrology frame that is exactly constrained and uninfluenced by foundation movement.

[II] An alternative is to use weight-relief points, in addition to three support points, that are statically compliant and also aligned with gravity. They can be used as traditional leveling screws to correct the static geometry of the machine, but they do not reduce deflections from moving or changing masses.

hysteresis. The geometry of the base will be machined in using the same supports with additional clamping loads applied to approximate the weight of structures added during assembly. Geometric error compensation will reduce the effect of manufacturing errors, deflections due to moving masses and variable part weight if necessary.

Concern is often expressed about recovering from a wreck if the geometry cannot be adjusted with leveling screws. A wreck of sufficient energy to permanently deform the base of the machine would probably damage a number of other components to the extent that re-leveling the machine is no fix. A wreck of lesser energy may cause certain alignments to slip in which case they can be realigned. The simplest repair would require only a re-mapping of the machine and update to the compensation. The mapping software could be fashioned to alert that certain physical alignments or repairs are required first. It is possible but more expensive to provide compliant leveling screws that would provide adjustable weight relief with minimal coupling of the base to the foundation, but this is not being proposed for the conceptual machining center.

9.2.4 Stiffness

Chapter 4 provides the fundamental ideas and numerous tools for designing stiff structures. For most large machine structures, the economics guide the design toward typical engineering materials such as steel or cast iron. The specific modulus of cast iron is only 65% that of steel or aluminum yet it remains the preferred material for general-purpose machine tools. If the particular structure moves as an axis, then the value of steel may outweigh its extra manufacturing cost. Aluminum castings are less expensive to manufacture but the material cost exceeds cast iron even though the weight required is less. Its higher coefficient of thermal expansion is a disadvantage too, but a network of steel tubes can be cast in to provide internal passages to flow temperature controlled water. In addition, aluminum can be used as the matrix in a composite with ceramic fibers. The use of a structural ceramic such as aluminum oxide may have application for smaller components that have difficult design constraints. For example, a quill made of aluminum oxide would be 80% stiffer than steel for one-half the weight.

Usually the choice of material plays a relatively small role in the stiffness of a machine-tool structure compared to the placement of the material. It is a matter of using the material most efficiently in the space that is available. Clearly, the use of large, structurally closed sections and favorable aspect ratios are priorities in the layout of the machine. At the component level, it is important that loads carried by a structural member be in plane to produce tension, compression or shear rather than bending. This is particularly critical in areas of concentrated loads such as bearing supports. For structures that see bending loads such as the machine base, the members in tension and compression must have adequate shear members connecting them. For structures that see torsional loads such as the column, all faces must support shear loads and should enclose the maximum volume. The open face

of a bifurcated column requires a stiff perimeter frame around the opening to support the shear load. The shear load produces a bending moment in the frame that is maximum at the corners. The tension and compression members of the frame require extra support around the corners to prevent localized bending and the resulting loss of stiffness.

Finite element analysis is the proper tool to provide visual and numerical information required to evaluate competing structural designs and/or to improve the baseline design. A plot of strain energy density is useful in optimizing the thicknesses of structural members by thickening high energy regions and thinning low energy regions, thus trying to achieve a uniform distribution. A sensitivity analysis, obtained by varying a design parameter and observing its effect on a figure of merit, is useful to assess the impact of a design change. A parameter study is similar except that it extends over a range that hopefully encompasses the optimum. These techniques are most applicable after the conceptual stage of design. Ideally, simple finite element models should be developed concurrently with the design concepts since much can be learned about the general behavior from simple models.

9.2.5 Productivity

Several aspects of productivity have been addressed in the requirements and strategies for machining performance. The machine will have ample power, spindle speed and axis speed for high-speed cutting and will have ample dynamic stiffness, spindle torque and axis force for heavy cuts in steel. The issues left to address are associated with nonproductive time such as tool changes and pallet exchanges, which are usually quoted in the advertising material. For the Maxim, Cincinnati Milacron advertises 3.5 seconds tool-to-tool and 7.5 metal-to-metal for tool changes and 10 seconds for a pallet exchange. For the Ram machine, G&L advertises 3 seconds tool-to-tool and 4.5 metal-to-metal for tool changes and 8 seconds for a pallet exchange. The difference in this case is significant and could easily exceed 15 seconds per part. There are several other areas where significant time could be saved but the gains are harder to quantify. Poor part programming by the user can waste several seconds per part. An optimizer software package that fixes inefficiencies could be developed and advertised as a control feature. Reliability is a critical factor for the machine to be productive. In addition, down time from nuisance alerts to outright failures negatively impacts the cost of ownership and the reputation of the equipment manufacturer. Ergonomics is a critical factor for the machine operator to be productive. The dictionary defines ergonomics as *the applied science of equipment design, as for the workplace, intended to maximize productivity by reducing operator fatigue and discomfort.* Improvements that make the machine simpler, easier and faster to use will usually be worthwhile.

The tool change operation is the most significant nonproductive time (unless the machine is unreliable) simply because many can occur per part. Therefore we will

concentrate on strategies to provide fast and reliable tool changes. The following steps occur in the typical tool exchange that affect the nonproductive time: 1) the spindle moves to the exchange position; 2) guarding is retracted to provide an access path; 3) the exchange arm moves from its home position to simultaneously grip the old tool and the new tool (assumed to be waiting at the load station); 4) the tools are released from the tapers of the spindle and the load station; 5) the exchange arm extracts the tools; 6) the exchange arm interchanges the old and new tools; 7) the exchange arm reinserts the tools into the spindle and the load station; 8) the spindle and the load station clamp the tools; 9) the exchange arm returns to its home position; 10) the guarding returns to a closed position; and 11) the spindle resumes cutting with the new tool. The movement of the guarding may overlap in time with the movement of the spindle so as not to add nonproductive time. There are three basic ways to reduce the nonproductive time: a) reduce the number of steps, b) overlap the timing so the steps appear shorter, or c) speed up the moves. For example, steps 3 and 9 can be eliminated if the home position of the exchange arm is at the spindle. Unfortunately, this is not very practical as it complicates the off-line task of delivering the new tool into a moving load station, among other things. Since the other steps appear to be essential, only timing and speed remain as avenues for improvement.

The surest way to speed and timing is to servo control the motion of the exchange arm and coordinate it with the motion of the spindle, that is, as two additional axes of motion within the machine controller. Computer control allows several interesting features that otherwise would be impractical. The velocity and acceleration of the arm can vary depending on the presence of a tool or perhaps on the tool weight measured at the load station. The arm and the spindle can move simultaneously into position so that engagement occurs moments after the spindle is in position. The extract-interchange-insert steps can occur as a continuous motion profile taking advantage of the extra clearance due to the taper. The servo gains can be controlled such that the arm acts compliantly as the tool engages or disengages the spindle. Since all moves follow defined paths and those paths are easy to program, the timing and speed can be optimized for the best combination of performance and reliability.

The typical exchange arm requires rotary and linear motion preferably but not necessarily supported by a common bearing system. Figure 9-7 shows one of several mechanisms that provides the desired motion. This mechanism is somewhat unique because the gear connected to the output shaft has both right and left hand helix angles of 45°. It meshes with two pinions of opposite hand so they rotate the gear when driven in the same direction and translate the gear when driven in the opposite direction. This scheme has the advantage that the two servo motors work together over most of the motion profile to drive the load. A similar but more conventional system has a single pinion driving a spur gear for rotary motion and a ball screw for linear motion. Another alternative is a combination ball screw and ball spline manufactured by THK and NSK. The motion of the exchange arm can also drive a revolving door in the guarding that is concentric with the

axis. The sequence is quite simple; the arm engages the door while extracting the tool, rotates the door with the interchange and disengages while inserting the tool. Counter to engaging and disengaging with the arm, the door must disengage and re-engage with a stop that prevents unwanted rotation of the door. Then the arm can return home to a shielded position against the door.

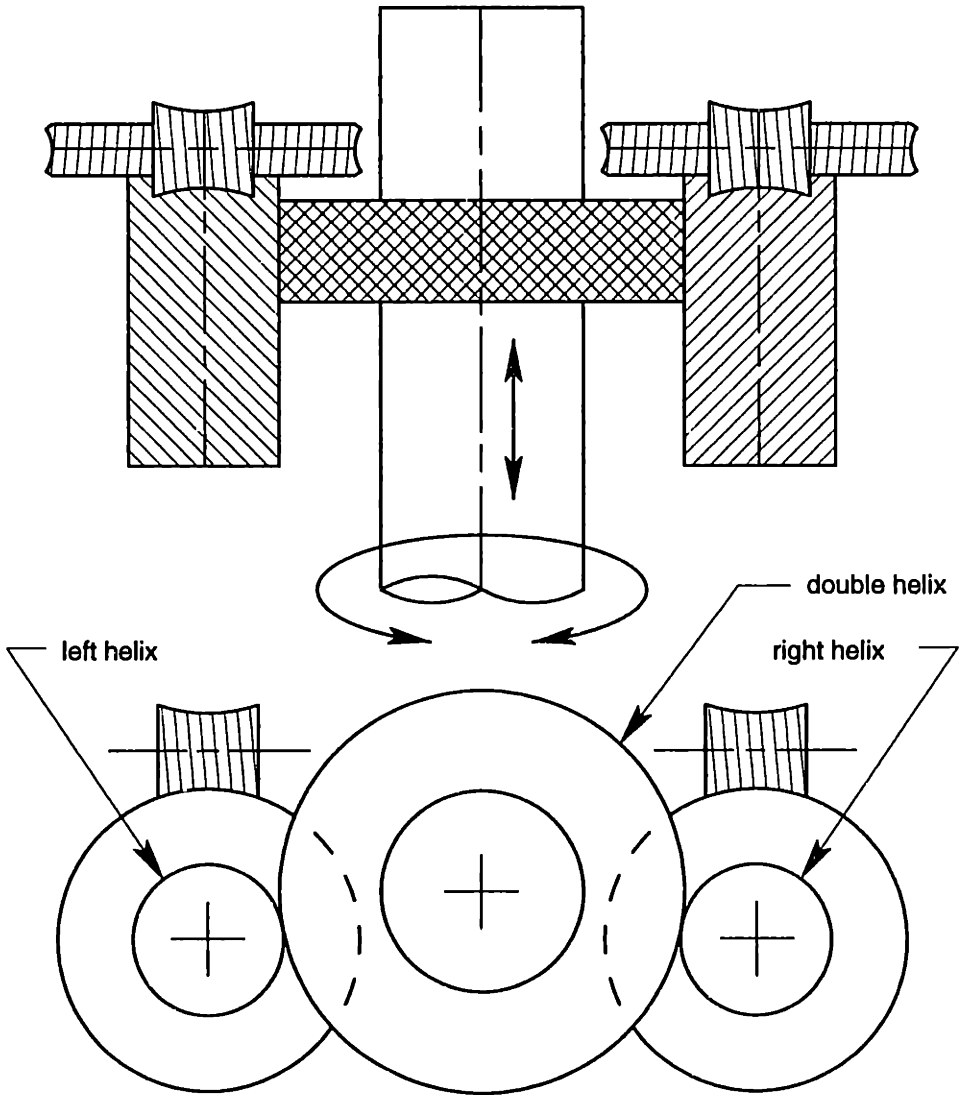


Figure 9-7 Coordinated motion by two helical pinions control the motion of the single output gear that is in mesh. The pinions must be of opposite hand, and the gear must be doubly helical. Not shown are two servo motors that each drive one pinion through a reduction gear shown here as a worm gear.

The work change operation is less critical because it happens only once per cycle (twice if there is an intermediate thermal wash). The large inertia of two fully loaded pallets presents an interesting motion control problem. The pallet clamp and lift mechanisms will likely require hydraulic actuators, and since the pallet exchange has only two positions (180° moves), a hydraulic index mechanism is preferable. A rack and pinion mechanism is simple but adequate acceleration and deceleration in the motion profile is difficult to achieve. The mechanism in Figure 9-8 naturally provides acceleration and deceleration by

simply controlling the flow rate to the cylinder with a flow control valve. The shape of the slot that the cam follower engages determines the motion profile.

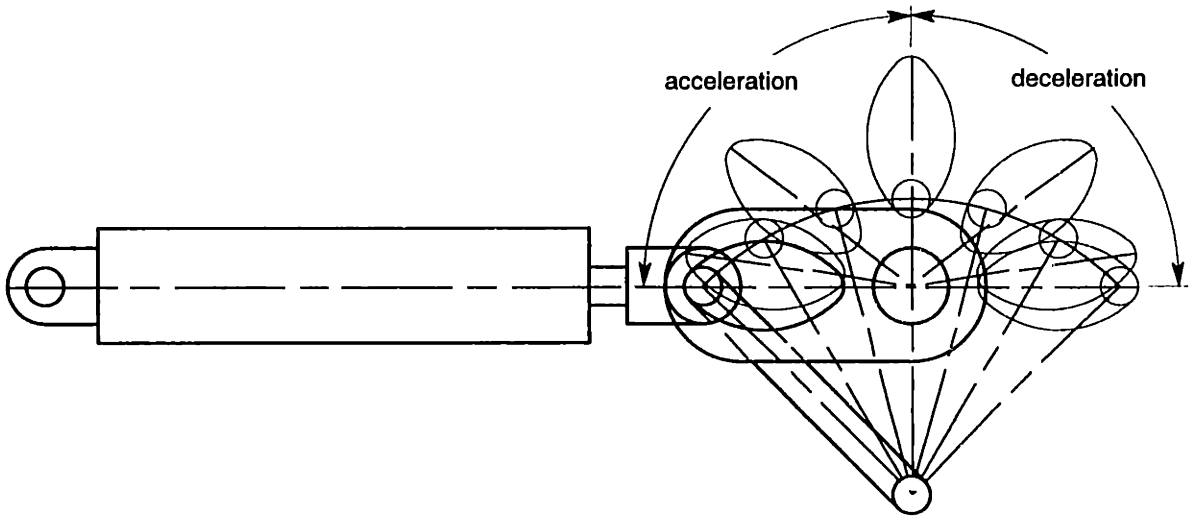


Figure 9-8 Nearly constant velocity of the cylinder generates smooth acceleration and deceleration of the crank as indicated by smaller angular change near the ends of travel compared to the mid travel.

The only recommendation regarding ergonomics is to interview a number of machinists who use similar equipment from various manufacturers. They are the experts in this area and often have very good ideas for improvements.

9.2.6 Manufacturability

A machine that is optimally manufacturable will require the minimum effort to build and qualify. The labor required to fabricate, machine, assemble and qualify the machine represents a significant expenditure that is directly affected by the design of the machine. The cost of commercial components and raw materials, although not really a part of manufacturability, is also affected by the design of the machine. An obvious strategy is to eliminate nonessentials, that is, only add parts or features that are required for the machine to function or that simplify some other aspect of the machine. The principles of exact constraint provide guidance in determining the essential structural supports for and connections between components. The use of pre-qualified modules and subassemblies, while costing more per unit, reduces the time required to build the machine, reduces the chances of rework and increases the flexibility to reconfigure the machine for faster delivery or field retrofits. Scraping, shimming and other fitting techniques should be avoided in preference to machining the components to adequate precision. This method may require in-process inspection of certain features and may become impractical if too many tolerances stack up in the design. Such cases require correction either through simple adjustments or software compensation.

The machine base is a good example to present specific ideas on obtaining precision at reasonable cost. The primary machining requirement for the base is the ability to produce

straight and parallel surfaces to support the rails. The typical base carries the X and Z axes so a squareness is also required. Squareness may require an in-process inspection or pre-qualification of the machine tool to achieve process control. The correction, if required, may come from an adjustment to the part program, the part fixture or the machine tool. Other machining operations can be relatively imprecise and may be more cost effective if done on a different machine tool.[1] A key aspect of the machining process is the proper support of the machine base so that the geometry is correct when the machine is assembled and used. If the machine base is designed to kinematically mount to its foundation, then the same mounting points should be used throughout manufacture to ensure consistency. A reasonable way to account for the weights of assembled components is to apply duplicate clamping forces to the base during machining. A clamping force is different from a constraint and ideally should have zero stiffness. In addition, it is important to consider the effects of friction at the constraints.

9.3 Selecting a Configuration of Axes

There are many possible ways to configure X, Y and Z linear axes and A and B rotary axes for a horizontal machining center, but only a few of these are typically used in practice. Since there may be an uncommon configuration that has significant advantages for precision, we will consider all possible combinations of X, Y and Z axes and require that the B axis be located directly under the workpiece. The total number of combinations is 24 because the order of support is significant. The requirement to provide an optional A axis is factored into the process of selecting the best standard configuration rather than considering all possible combinations of X, Y, Z and A axes. The Analytic Hierarchy Process (AHP) is used to aid and document the selection process. It requires determining criteria, developing designs and evaluating options as discussed in Chapter 3.

Table 9-7 shows the AHP spreadsheet for 24 configurations, which are identified by number and a descriptive code. Axes that support the workpiece are listed under the heading *Work* in order of support from the base to the workpiece. In the same way the axes that support the tool are listed under the heading *Tool*. The first-level criteria appear across the top and the lower-level criteria, where ranking takes place, are hidden from view to reduce the size of the spreadsheet. The selection process began by making first-cut sketches of feasible designs. Some configurations, however, were not feasible and thus were not evaluated further. Appendix G shows the 14 feasible designs and the entire AHP spreadsheet used to evaluate them. The following sections present the AHP criteria and the results of the decision process. Section 9.3 *Design Layouts* continues the development of the selected configuration into the conceptual design.

---

[1] The ball screw supports require precise location relative to the rails, however as mentioned previously, bores can be produced relative to the rails in a number of ways.

AHP1			Decision	10	10	8	6	4
	Criteria Level 1 >		1.00	Req'd sys.	Accuracy	Productivity	Manuf. cost	Ergonomics
	Criteria Level 2 >			0.26	0.26	0.21	0.16	0.11
	Criteria Level 3 >							
I.D.	*Work*	*Tool*						
1	A, B	X, Y, Z	6.22	5.53	6.16	6.25	6.34	8.19
2	A, B	X, Z, Y	5.53	5.19	5.17	5.12	6.72	6.68
3	A, B	Y, X, Z	5.46	4.15	7.01	7.22	5.73	3.13
4	A, B	Y, Z, X	4.56	2.76	6.58	6.25	4.96	3.00
5	A, B	Z, X, Y	5.34	4.47	5.55	4.88	7.06	5.97
~~6~~	~~B~~	~~Z, Y, X~~						
7	X, A, B	Y, Z	6.75	4.97	7.29	8.14	6.67	8.41
8	X, B	Z, Y, A	8.14	7.85	8.26	8.53	7.90	8.18
9	Y, B	X, A, Z	7.49	7.91	7.72	7.94	6.20	7.13
10	Y, B	Z, X, A	7.53	8.44	8.00	6.93	6.73	6.83
11	Z, B	X, Y, A	7.93	8.54	7.70	7.48	7.50	8.71
12	Z, A, B	Y, X	5.78	3.51	7.77	8.02	6.15	4.52
~~13~~	~~X, Y, B~~	~~Z~~						
14	X, Z, B	Y, A	7.56	8.63	7.10	7.18	6.95	7.97
15	Y, X, B	Z, A	7.55	7.66	8.29	7.56	7.03	6.41
~~16~~	~~Y, Z, B~~	~~X~~						
17	Z, X, B	Y, A	7.50	8.63	7.10	7.00	6.95	7.79
~~18~~	~~Z, Y, B~~	~~X~~						
~~19~~	~~X, Y, Z, B~~							
~~20~~	~~X, Z, Y, B~~							
~~21~~	~~Y, X, Z, B~~							
~~22~~	~~Y, Z, X, B~~							
~~23~~	~~Z, X, Y, B~~							
~~24~~	~~Z, Y, X, B~~							

Table 9-7 All combinations of X, Y, and Z axes that support the tool and the workpiece. The most common configurations in industry are T-based (8 and 11). The G&L Ram 500 corresponds to (7) and the C-M Maxim 500 to (11). The traditional knee-and-column horizontal mill corresponds to (22).

9.3.1 AHP Criteria

The AHP Criteria and the weights assigned to them become the decision model used to make a comparative evaluation of design options. When the options and/or the factors that affect the decision are numerous, the decision process becomes cumbersome unless it is broken down into manageable pieces. A tournament style approach, where design options are compared pairwise to eliminate losers, is a simple way to deal with a large number of options especially if the pair differs in only a few respects. The AHP approach is different in that decisions are made one factor at a time and then the results are compiled based on the weights assigned to the criteria. The simplest AHP has no hierarchy; all criteria appear in one level and carry equal weight. More often the criteria are not equal in importance and thus require weights to compensate. Organizing the hierarchy as an outline is a natural approach to make the decision model more manageable.

The following outline presents the important factors affecting the choice for the best configuration. It reiterates many of the ideas and thoughts presented throughout this thesis. The AHP spreadsheet is an exact duplicate of this outline with the addition of weights that

indicate the relative importance of the criteria. The criteria are arranged in a generally decreasing order of importance.

1 Required systems: These systems enhance the value of the machine and must be incorporated into its design. Their effectiveness and complexity depend on the configuration of the machine.

1.1 A-axis: Locate the optional A-axis under the tool if possible to avoid changing the orientation of the workpiece. Placing the A-axis under the workpiece increases the size of the enclosure, reduces the effective range and makes compensation for part weight more difficult. In addition, minimize the changes required to the standard machine to add an A-axis.

1.2 Tool changer: Locate the tool exchange arm as close as possible to the spindle to minimize index time and swept volume. Some configurations may require the tool changer to be more complex to achieve similar performance. For example, the exchange arm may have to mount to the column that travels in Z.

1.3 Work changer: Locate the work changer as close as possible to the workpiece to minimize index time and swept volume. Strive for a work changer that is self supporting and easily installed on the machine.

1.4 Chip removal: Provide a simple and effective chip removal system using, for example, sloping way covers and other surfaces or devices that enhance the movement of chips to a chip conveyer.

2 Accuracy: The highest attribute that a machine tool can have is accuracy. Accuracy, unlike production rate, cannot be improved by increasing the number of machines. It requires control of the process in a deterministic way.

2.1 Static stiffness: Maximize static stiffness to reduce errors caused by process forces, moving masses and variations in part weight.

2.1.1 Cross sections: Maximize section depths for greater specific stiffness. Provide shear members (webs or diagonal ribs).

2.1.2 Stiffness loop: Minimize the length of the stiffness loop from the tool to the work.

2.1.3 Aspect ratio: Provide a favorable ratio of bearing spacing to moment arm.

2.2 Alignment principles: Minimize sine errors by minimizing lever arms.

2.2.1 Abbé: Align position feedback sensors (linear scales or ball screws with encoders) as collinear as possible with the tool point.

2.2.2 Bryan: Align straightness references (linear bearings) as coplanar as possible with the tool point.

2.3 Stability: Metrology structures are considered stable if they remain invariant, vary in a symmetric way as to be canceling or vary only in nonsensitive directions. Separate the sources of instabilities from the metrology loop physically and/or informationally using error compensation. Favor concepts that require minimal compensation or are easier to compensate.

2.3.1 Moving loads: Position errors caused by the weight of moving axes can be mapped and compensated. Compensation of angular errors requires servo A and B axes and is more time consuming to map.

2.3.2 Variable loads: A variable part weight can be included in the compensation scheme but requires knowledge of the weight and additional time for mapping.

2.3.3 Thermal: Thermal errors are more difficult to compensate but can be reduced through temperature control.

3 Productivity: Maximize material removal rate and minimize unproductive time.

3.1 Dynamic stiffness: Design structures to have approximately the same stiffness as the bearings and servos so that their damping couple into the structures.

3.2 Natural frequencies: Maximize natural frequencies (high stiffness to mass) to allow faster precision contour moves.

3.3 Axis acceleration: Minimize the weight of moving structures and locate actuators as close as possible to the mass centroid.

3.4 Reliability: Consider the protection of linear guides and ball screws from chips and coolant. Make moving shields, covers or other protection devices self cleaning and simple.

4 Manufacturability: Design for manufacturability to reduce product cost.

4.1 Modularity: Use pre-qualified modules and subassemblies to reduce build time and chances of rework and to increase the ability to reconfigure or field retrofit.

4.2 Exact constraint: Use the minimum number of constraints to reduce the number and the required precision of machined features.

4.3 Machining: Optimize the design for machining the precision surfaces. Minimize the number of setups and operations. Machining fixtures should duplicate the loads and constraints that the structures will have in use.

4.4 Assembly: Avoid fitting at assembly to obtain alignments. Machine in precision, prescribe fitting spacers from inspection data or provide simple adjustments with rapid feedback of results.

4.5 Floor space: Minimizing floor space has a ripple effect on cost from the component level to the customer's plant.

5 Ergonomics: Design for easier setup, operation and maintenance.

5.1 Visibility: Provide good visibility of the workspace for setup and monitoring the process.

5.2 Pallet access: Provide clear access to the pallet for load or unload and setup of fixtures. Provide overhead crane access.

5.3 Spindle access: Provide clear access to the spindle for manual tool change or inspection.

5.4 Maintenance and repair: Design the workspace to be easily cleaned and maintained (covers for example). Provide crane access for repair or replacement of subassemblies or field installation of options.

5.5 Safety: Eliminate or guard crush zones. Eliminate sharp edges and corners. Prevent the possibility of slides falling due to a power failure or other reasons.

9.3.2 Discussion of Results

Applying the Analytic Hierarchy Process has been an experiment in decision making that has generated data as a score card. Simply using the data (as intended) to select a configuration would overlook two important aspects of the effort; that is, understanding the technical basis for the decisions and learning how to better use the AHP for complex design decisions. We can plainly see the scores in Figure 9-9 and yet be completely oblivious to the thought and deliberation that went into making 364 entries to the AHP spreadsheet (26 terminal criteria times 14 design options). Furthermore, it is difficult to learn much by examining the numbers in the spreadsheet. Therefore, the discussion necessarily centers on the decision process in addition to the outcome.

It is important to bear in mind that the comparisons made were for first-cut designs of different configurations. A poor score may simply mean that the particular design is poor rather than the configuration it represents. In addition, the comparisons were subjective interpretations of the designs rather than being based on engineering calculations. This reflects the early conceptual stage of the design process and the large number of choices. Indeed, nothing restricts the AHP from being used to compare a few competing designs of the same configuration or for tracking the progress of design iterations. The spreadsheet is particularly easy to modify for additional design options; however, the number being compared should remain manageable, perhaps six to eight. Comparing fourteen designs at once proved to be difficult and probably would work better if divided into smaller groups.

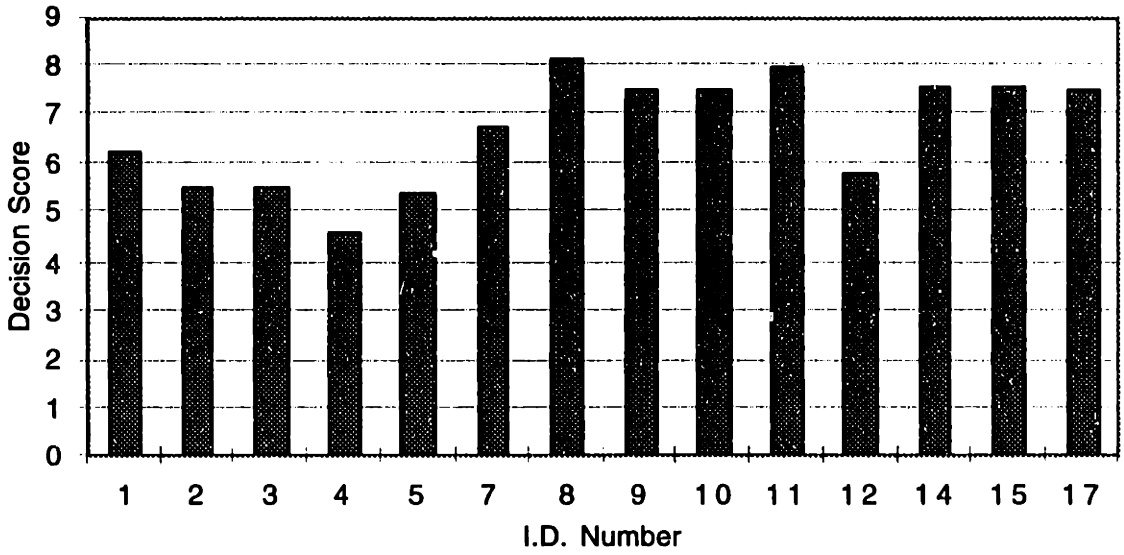


Figure 9-9 The overall scores for the 14 ranked configurations represented as a bar chart.

In reviewing the scores in Figure 9-9, the first thing to notice is the parity among the top seven or eight configurations. The seven configurations that scored above 7 all have the tool located over the A-axis while the remaining configurations all have the workpiece located over the A-axis. Had the A-axis not been a requirement, the scores would be even closer. In particular, configuration number 7, which is the configuration of the G&L Ram machine, would be a top contender. Another important factor is the number of stacked slides. Configurations 1-6 all have the tool located over X, Y and Z axes, which is the reason that the workpiece is located over the A-axis. Configurations 19-24 were not ranked because of the difficulty in placing the workpiece over X, Y, Z and B axes. Five of the top seven configurations are nearly equal with scores near 7.5. Three of these (9, 10 and 15) have the workpiece located over the Y-axis. The main advantages are a stocky base, a simpler pallet exchange mechanism (because the vertical motion is inherent in the Y-axis) and less influence from a moving or variable part mass. Complications with the A-axis and chip removal under the Y-axis are the main disadvantages. The other two configurations with scores near 7.5 (14 and 17) have the tool located over the Y-axis on a stationary column and the workpiece moves horizontally in both X and Z directions. While the column can be made relatively stiff, the depth of the base is limited by the stacked slides. This arrangement also poses difficulties in protecting the ways from chips and coolant. The top two configurations (8 and 11) are T-based meaning that the horizontal X and Z axes are supported on the base as to form the letter tee. Certainly these two configurations are the most prevalent among manufacturers and deserve more detailed consideration before choosing one.

Figure 9-10 shows the comparison of the two T-based configurations through the first criteria level. Although the two are nearly equivalent in overall score, they each have notable advantages over the other. When the column moves in X (configuration 11), the

operator has better access to the spindle and better visibility, and the tool changer is easier to package. A column that moves in Z (configuration 8) can be supported centrally with a deep-section base, which presumably results in higher natural frequencies for column modes and less error motion as the column translates. In addition, the X-axis carriage naturally accommodates a symmetric three-bearing arrangement under the workpiece. The advantages of accuracy and productivity are fundamental to the value of a machine tool whereas the others are advantages of convenience. Thus the first level scores more clearly support the decision to use configuration 8. Ideally, we would carry both configurations through a preliminary design and analysis to allow a more thorough comparison, but fortunately, comparisons can be drawn to the Maxim, a configuration 11 machine.

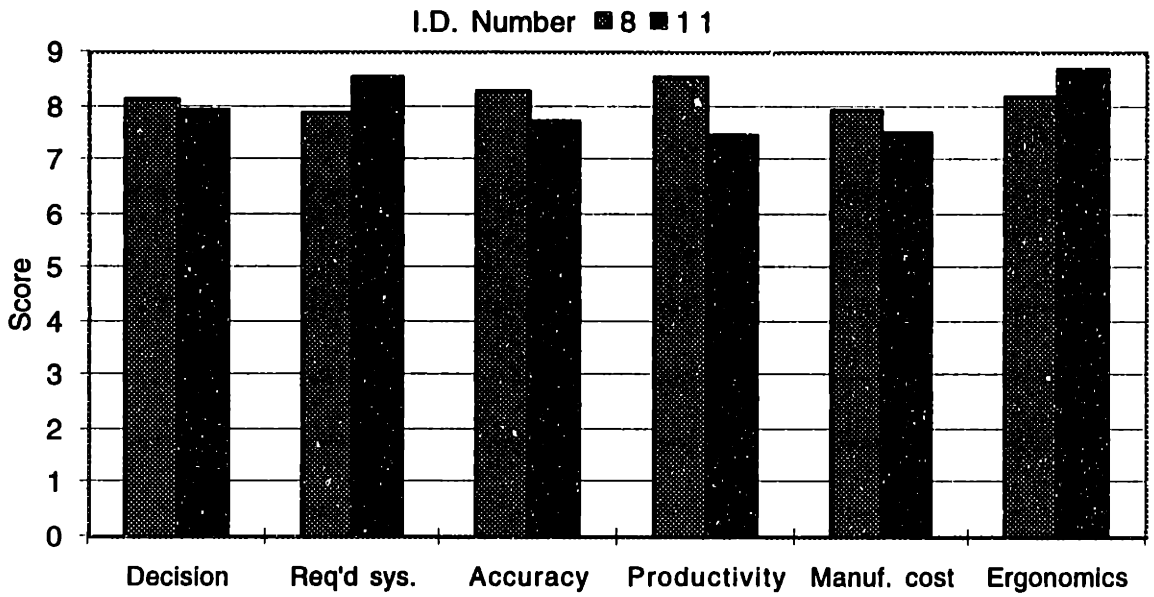


Figure 9-10 The scores for the top two configurations through the first criteria level.

9.4 Design Layouts

This section continues the development of the selected configuration into the conceptual design. The design-analysis process is iterative starting with the selected configuration, the range drawing (Figure 9-1) and actual or estimated envelopes for components like the spindle motor. In this study, the geometry was created first in 3D wireframe and exported to a finite element program for analysis.[1] A number of iterations occurred to improve and/or simplify the design, but only the final version is presented. Where appropriate, past trends or future improvements will be discussed. Figure 9-11 through Figure 9-14 show views of the wireframe model. Finite element models shown in following sections give better

[1] Vellum™ 3D by Ashlar Inc. is the design/drafting software used in this study. Pro/MECHANICA™ by Parametric Technology Corp. is the finite element software used in this study. File transfer can be in either dxf or iges file formats.

definition of the construction. Since the wireframe geometry was created for the finite element model, a shell model, the lines generally represent midplanes of walls. As a result, the gaps between linear bearings and the mounting surfaces may look peculiar. The model also lacks ball screws, guarding and other details that would be time consuming to include, but ideas for them are discussed.

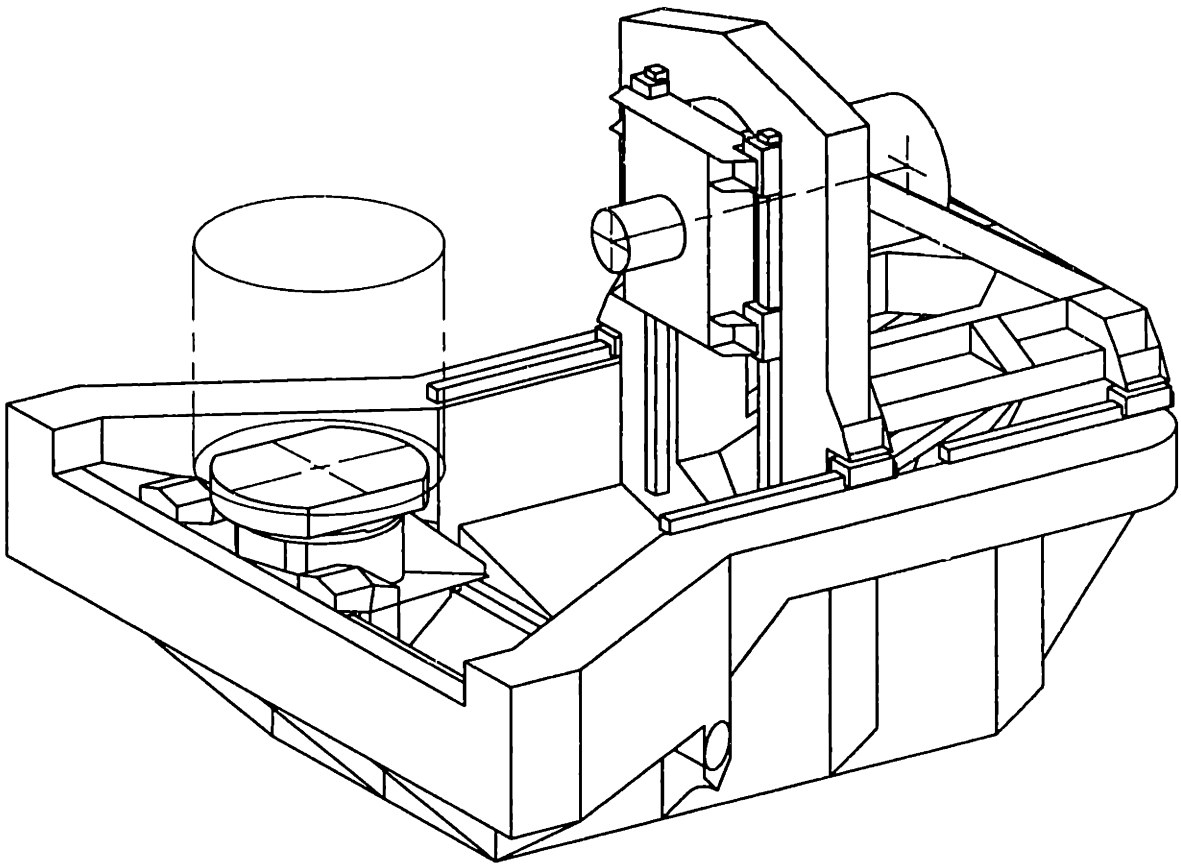


Figure 9-11 The 3D wireframe model of the conceptual design shown at maximum X, Y and Z travel.

This configuration leads naturally to first designing the column around the spindle motor and carrier and then designing the base around the column and work carriage. Four bearings support the spindle carrier in a symmetric arrangement. The column is bifurcated with a substantial perimeter frame for shear stiffness at the Y-axis bearing plane. Four Z-axis bearings support the column at an elevation approaching its centroid. The designs of both the column and spindle carrier are torsionally compliant in an effort to release the planar overconstraint of four bearings. The spindle carrier housing has a large opening in the front that is covered with a constrained-layer damper. Although designed as a casting, the spindle carrier would be simple to fabricate from steel plate. Initially a casting, the column was changed to a tubular steel structure to allow interior water circulation (a trickle down approach to minimize the volume of water inside the column) and to demonstrate a built-in constrained-layer damper. The casting design had nearly the same modal frequencies as the tubular structure because material can be more optimally placed in a

casting. The column is very open behind the perimeter frame to allow clearance for the spindle motor as the optional A-axis articulates. The column requires good shear stiffness in three planes, X-Y, Y-Z and X-Z.

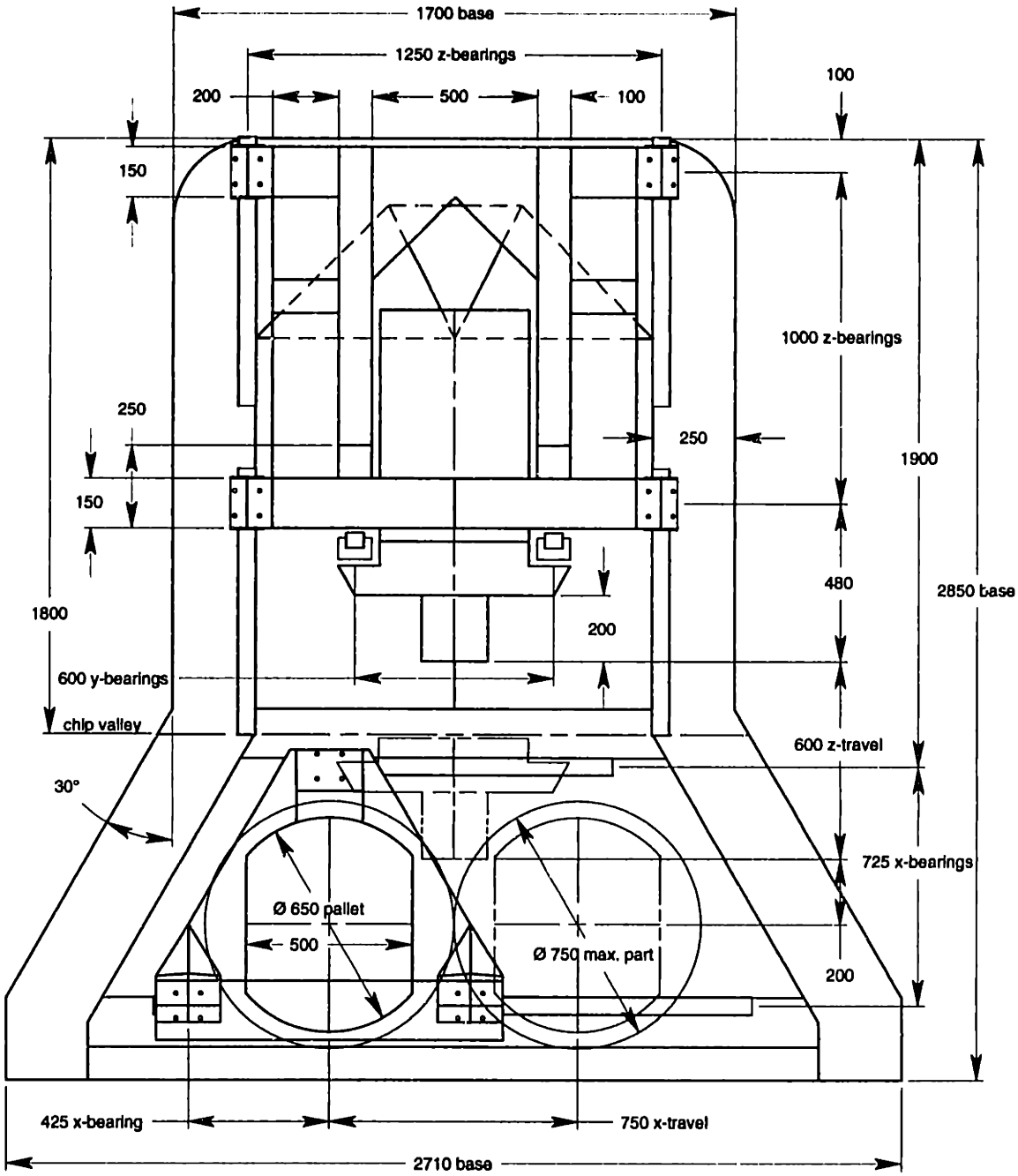


Figure 9-12 Plan view of the conceptual design.

The work carriage rides on three bearings equally spaced about the B-axis on a 15° incline (to promote chip flow to an auger). It too is a casting but would be simple to fabricate from steel plate. The carriage is the weak link in the system as indicated by its activity in the lowest mode shapes. The principal compliances are the bearings, the cantilever bearing mounts and the pallet coupling (a three-tooth kinematic coupling). The

manner in which the pallet coupling was modeled probably underestimates its stiffness. Improvement to the bearing mounts is necessary and may require the bearing rail to move down somewhat so that the mounts have deeper sections.

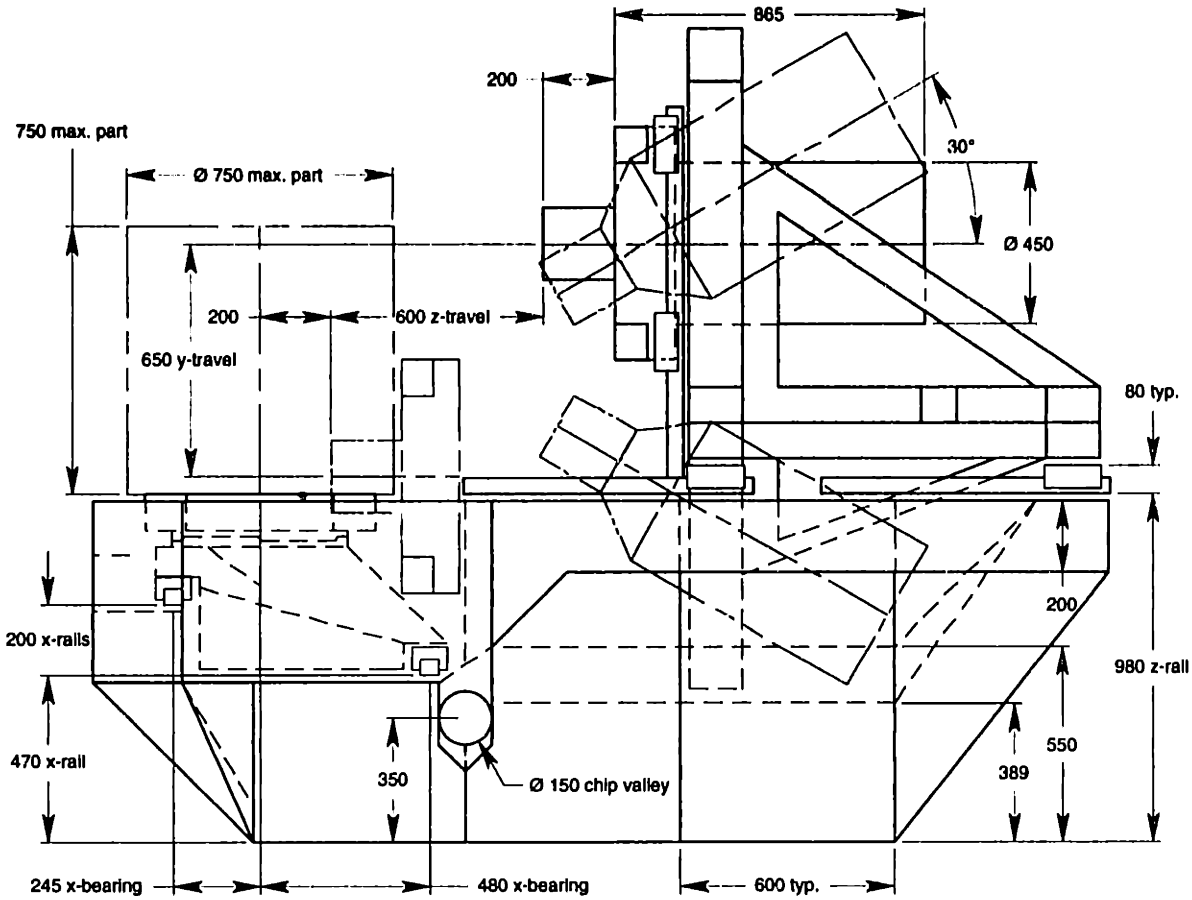


Figure 9-13 Right elevation view of the conceptual design.

The base is a one-piece casting shaped to fully contain coolant flow used for temperature control. Sloping way covers over the X-axis and scrapers between the column and base promote chip flow to a crosswise central valley. An auger propels chips from the valley out the side of the machine. The preferred location for the X-axis ball screw is in the plane of the bearings and intersecting the B-axis. The details of the B-axis servo may impact the location somewhat. Twin ball screws drive the column with a combined thrust force acting very near its centroid. This reduces pitch and yaw of the axis during acceleration. Each servo would have local velocity control to provide damping against shear and rotation of the column in the X-Z plane. An interesting choice remains regarding the position control. The column was designed with sufficient shear stiffness in the X-Z plane to allow a single position loop (each servo would receive identical position feedback). If instead each servo had its own position control, then angular error compensation about the Y-axis would be possible without a B-axis servo, and the column would be simpler without need for such large X-Z shear stiffness.

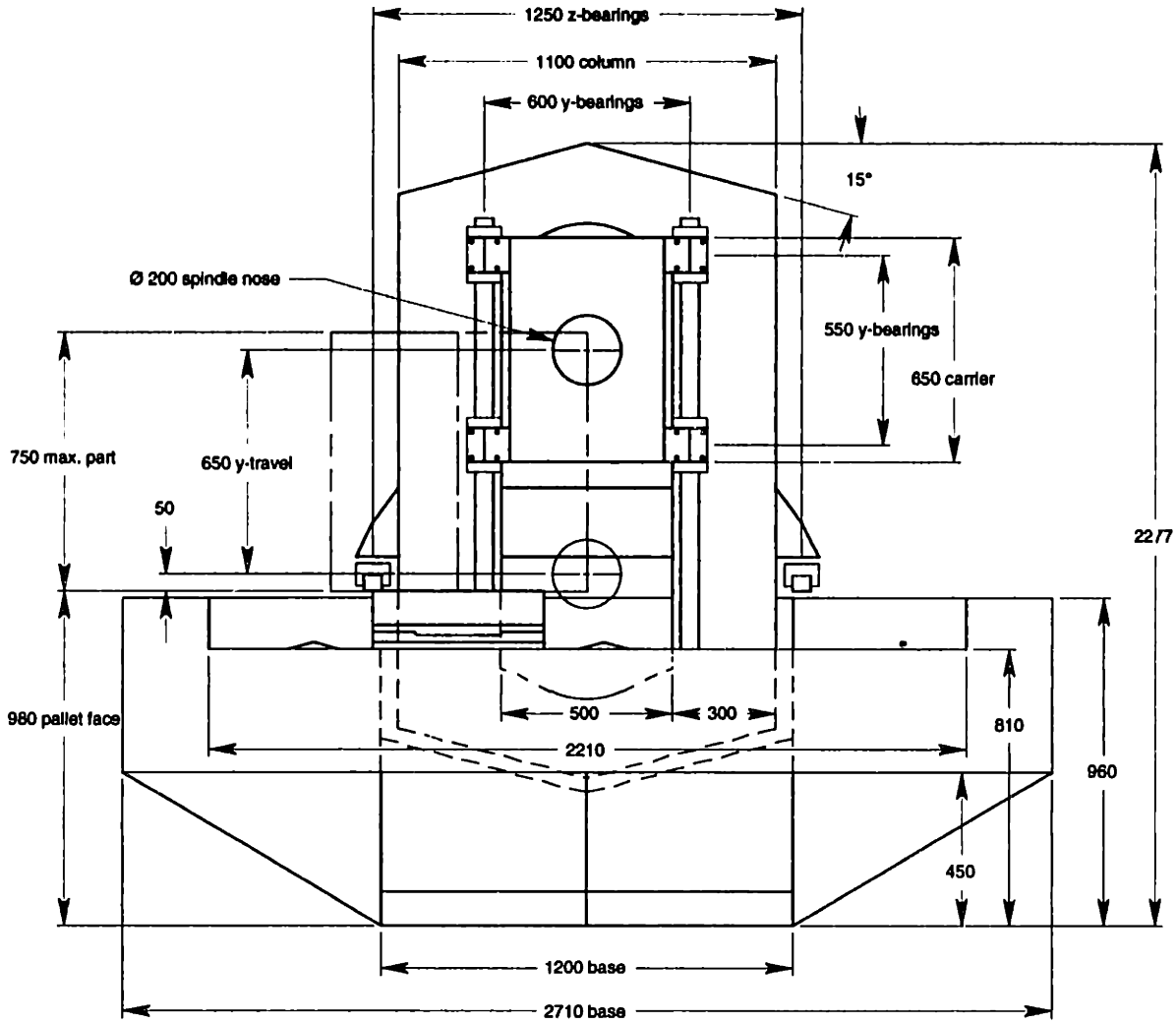


Figure 9-14 Front elevation view of the conceptual design.

The base has four supports to the floor but the two rear supports are angled ±45° in the X-Y plane to form a wide vee constraint. The instant center of the vee is approximately 600 mm above the floor and centered on the Y-Z plane of symmetry. The node of the torsional mode of the base typically lies close to the instant center. Placing the node nearer the centroid reduces the inertia of this mode and increases its frequency. The two supports at the front of the base could be vee constraints but it is more common to use flat pads with anchor bolts, where friction provides the horizontal constraint. In either case, friction also causes overconstraint between the base and the floor. Overconstraint is good for static and dynamic stiffness but bad for precision. Ultraprecision machines typically use air isolators that are very compliant in the lateral direction so that overconstraint is not a problem nor is it a benefit. Air isolators lack the very high viscous damping required to dynamically couple the base to the floor. A reasonable and practical approach for this machine is to use a viscoelastic material like DYAD 606 that develops its maximum damping in a range from 10 to 100 Hz. The static shear stiffness of the material is several orders of magnitude lower than the dynamic shear stiffness but never zero. Therefore, an initial setup procedure is

necessary to approach zero friction force in the sliding joint. This procedure may be as simple as removing the weight from the pads so that the viscoelastic material can relax. The details of the supports were not studied but the basic idea, to use the floor as a constraining layer, has been demonstrated to be effective.

The figures do not show the work changer, tool changer, guarding, operator control station and various utilities. These have not been designed but strategies for the tool changer and work changer were discussed in Section 9.2.5 *Productivity*. The plan is to support all these items except the tool exchange arm from the floor rather than the base. The work changer is an option as some customers will prefer different work handling systems. There will be reference features machined on the base to set up critical alignments but they will not transmit loads. The guarding is free standing on casters so as to be easily removed for transportation or any major repairs on the machine. Completely enclosing the machine structure, it becomes an effective thermal enclosure. The operator control station is simply a table with a commercial keyboard and monitor. In addition, there will be a hand-held pendant that can plug into several locations around the machine. Most utilities will be at the rear of the machine, also mounted on casters, and have quick disconnects to the machine.

It is easy to get the impression that this design is large and massive. Rather, it is smaller and lighter than the Maxim (see Table 9-8). A key difference is the Maxim base being long enough to support the work changer. The spindle motor is the only component that is heavier on the conceptual design. The decision to use a direct-drive spindle should be reconsidered more carefully. Its inertia and centroid location will degrade the dynamic response of the A-axis and possibly the Y-axis. Counterbalance is necessary and easy to accomplish by lifting the spindle motor at a location that balances both axes. Perhaps the simplest counterbalance is a hydraulic cylinder mounted horizontally above the spindle motor and connected symmetrically on each side with a cable and pulley. A reasonable alternative to articulating the spindle motor with the A-axis is to use a constant-velocity universal joint, but this poses difficulties for spindle coolant and tool changing.

Component Name	*Typical Wall*	*Concept Mass*	*Maxim Mass*
Base casting	20 mm	3487 kg	4857 kg
Column structure	9.525 mm	795 kg	2100 kg
Spindle carrier casting	25 mm	218 kg	348 kg
Spindle motor	60 mm	535 kg	213 kg
Spindle carrier assembly		835 kg	701 kg
Column-carrier assembly		1630 kg	2801 kg
X-axis carriage casting and pallet	20 mm	572 kg	1078 kg
Machine assembly less guarding, etc.		5689 kg	8736 kg

Table 9-8 Masses for the conceptual design compared to the Maxim 500.

9.5 Analytical Results

The concept is sufficiently developed at this point to present analytical predictions and to compare them to design goals. The presentation begins at the component level showing finite element models and results of static and modal analyses. The component models are later integrated into an assembly model. The spring stiffness used to represent each slide bearing is a key assumption in the assembly model. The bearings used on the Maxim are size 55 linear guides by NSK, and the advertised stiffness of 883 N/μm is used in the model. The tool-to-work loop compliance is computed for extreme ranges of the Y and Z axes. The range with the shortest loop is the only one to meet the goal and the model did not include the compliance of the spindle. The work carriage is the *weak link* in the chain and ways to improve its stiffness are suggested. Finally, the volumetric errors that result from two sources, the weights of moving slides and a moment applied to the base, are calculated throughout the work volume. The results indicate that there is no need to include part weight in the geometric error compensation and that a static moment of order 800 N-m is acceptable. Several ways to improve the design of the base are suggested.

9.5.1 The Spindle Carrier Model

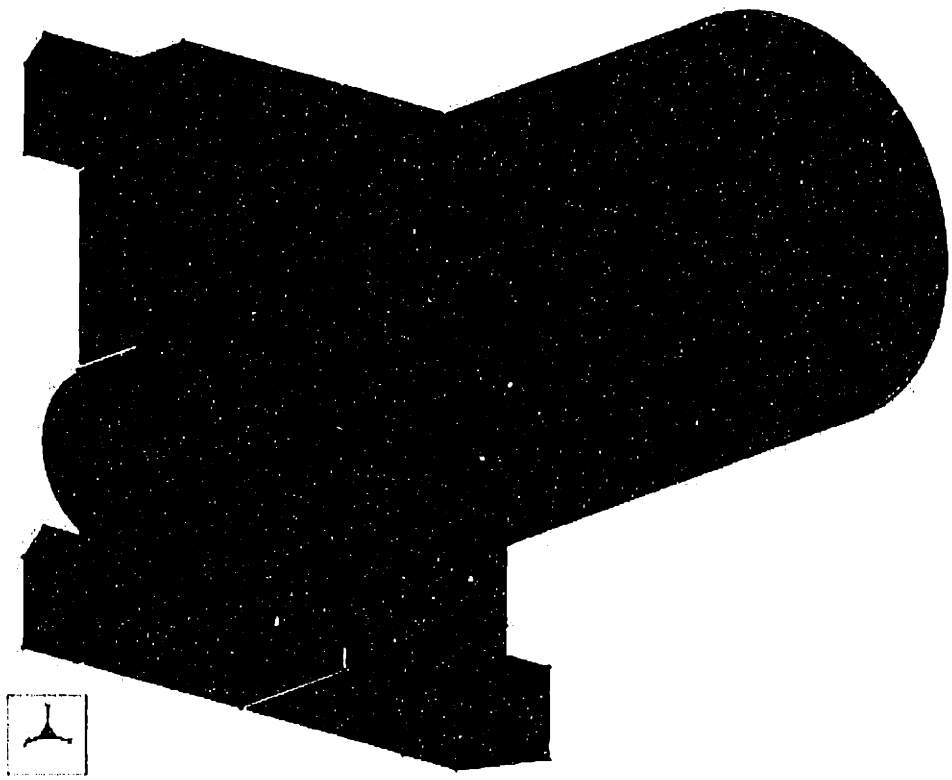


Figure 9-15 Spindle carrier model with constraining layer removed.

Shown in Figure 9-15, the spindle carrier housing is a cast iron structure that is open in the front to release the planar overconstraint of four cantilever bearing mounts. A constrained-layer damper (not shown) covers the opening to enhance dynamic stiffness. The damper

consists of a 12 mm steel plate and a viscoelastic layer between it and the housing. It is best to epoxy the damper together, although a bolted connection would allow occasional access inside. Without a constraining layer and with each bearing mount constrained at a point for X-Y-Z translations, the first modal frequency exceeds 100 Hz. Removing the constraint at one mount reduces the first mode to 52 Hz. Applying a unit Z-force to the unconstrained mount results in a torsional stiffness about the Y-axis of 14 N-m/μr. A four-bearing mount with the equivalent torsional stiffness would require bearings with 157 N/μm stiffness, which is 5.6 times more compliant than the actual bearings. This indicates that the housing is relatively flexible compared to the planar constraint of the bearings. This ratio would be much greater if the motor were connected at four quadrants rather than around its perimeter.

Table 9-9 shows the results of a second analysis to investigate the benefit of the constrained-layer damper. Four springs each with 883 N/μm stiffness support the spindle carrier in X and Z. The damping estimate k_{Im}/k_0 was calculated with Equations 5.1 and 5.4 using the square of modal frequency as the stiffness-proportional parameter. Although the damping estimates shown are modest, the effect on the spindle mode may be the most significant to cutting performance. The modal strain energy method, used to determine the VEM thickness and shear modulus, should include details of the spindle design.

Mode	*Mode Shape*	$G = 0$	$G = \infty$	k_{Im}/k_0
1	Rigid body Y-translation	0	0	0
2	Y-rotation of housing, motor and spindle	100 Hz	114 Hz	0.062
3	X- rotation of housing, motor and spindle	135 Hz	149 Hz	0.045
4	Piston mode in Z-direction	185 Hz	204 Hz	0.045
5	Z-rotation of housing, motor and spindle	286 Hz	300 Hz	0.021

Table 9-9 Spindle carrier modes with a 12 mm steel constraining layer and limiting values of viscoelastic shear modulus G. The damping estimate k_{Im}/k_0 is approximately equal to the loss factor.

9.5.2 The Column Model

Figure 9-16 and Figure 9-17 show the column model with the spindle carrier assembled at the upper end of the Y travel. As mentioned previously, the column is constructed of steel tubing, predominately 4 x 8 x 3/8 inch wall. The design can accommodate a viscoelastic layer between parallel tubes in the vertical members of the perimeter frame and also between tubes running front to back. The tubes would be welded together near their ends but the rest of the span would be bonded through the viscoelastic layer. Intuitively, this design should be highly damped but it has not been analyzed.[1] The manufacturing problem

[1] Since the welded structure should go through a stress-relieving heat treatment, the viscoelastic layer must be applied afterwards in-between tubes. A pourable viscoelastic may be more convenient than filling gaps on either side of a viscoelastic sheet with epoxy. See Chapter 5 for references on constrained layer damping.

of applying the viscoelastic layer requires additional thought and experimentation. Furthermore, the value to performance needs to be quantified and compared to the costs.

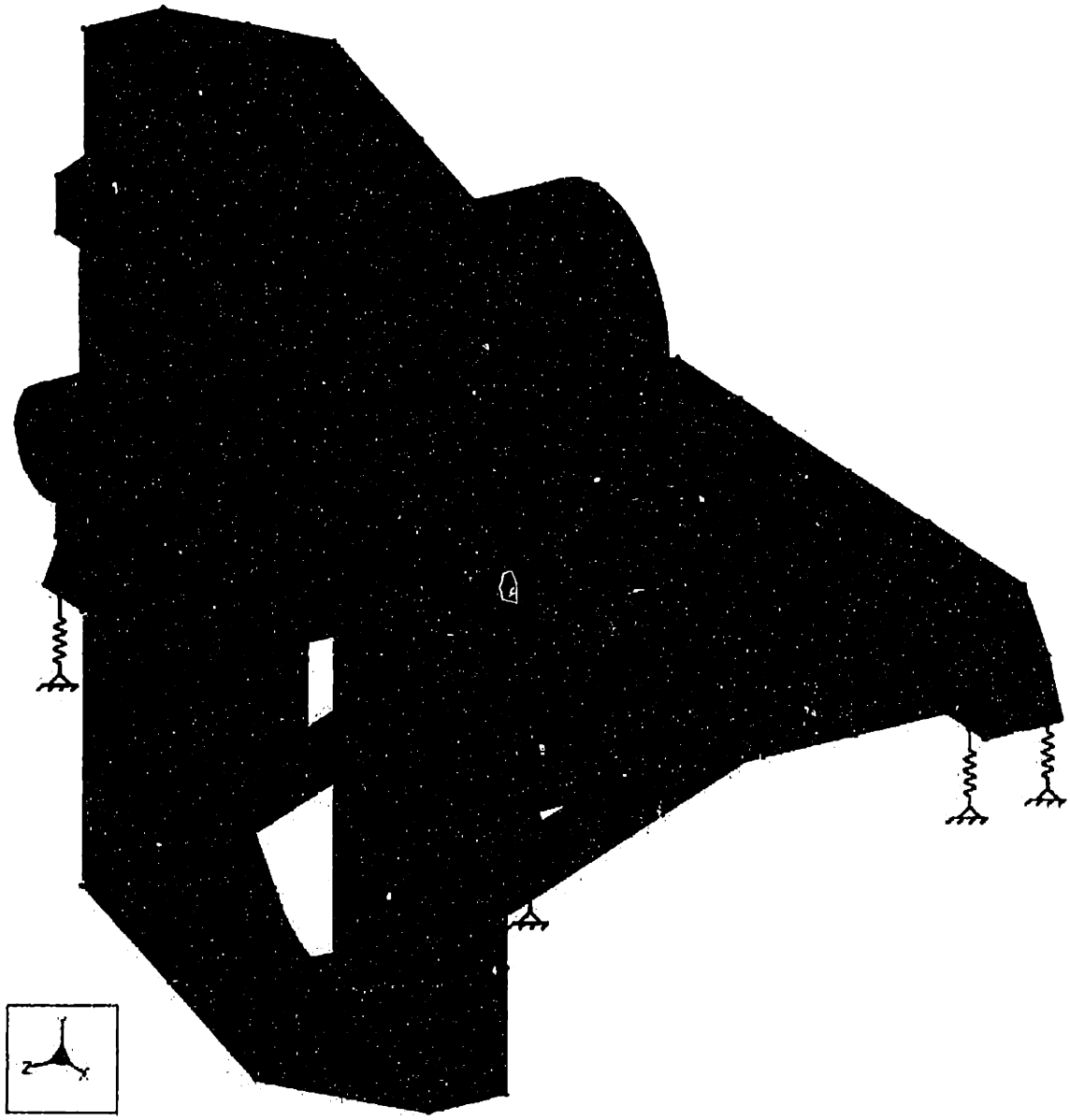


Figure 9-16 The column model with the spindle carrier in the upper position is supported on four pairs of grounded springs each pair having stiffness in X and Y of 883 N/μm. Four springs each having 883 N/μm stiffness in X, Y and Z attach the spindle carrier to the column.

Four pairs of grounded springs support the column in X and Y, where each pair has a stiffness of 883 N/μm. The Z direction is left unconstrained so as not to artificially stiffen the column in the X-Z plane. The four springs that attach the spindle carrier to the column each have 883 N/μm stiffness in X, Y and Z. A modal analysis was performed with the spindle carrier at three different Y-positions, 0, 325 and 650 mm (the one shown). Table 9-10 presents the results of the modal analysis. Motion of the spindle carrier on its bearings dominates modes 2, 3 and 4 while motion of the column on its bearings dominates

mode 5. The results show that the column is relatively stiff with only slight deformation for mode 2 and moderate deformation for mode 4. The relatively small decrease in frequency as the spindle carrier moves up is another indication that the column has ample stiffness.

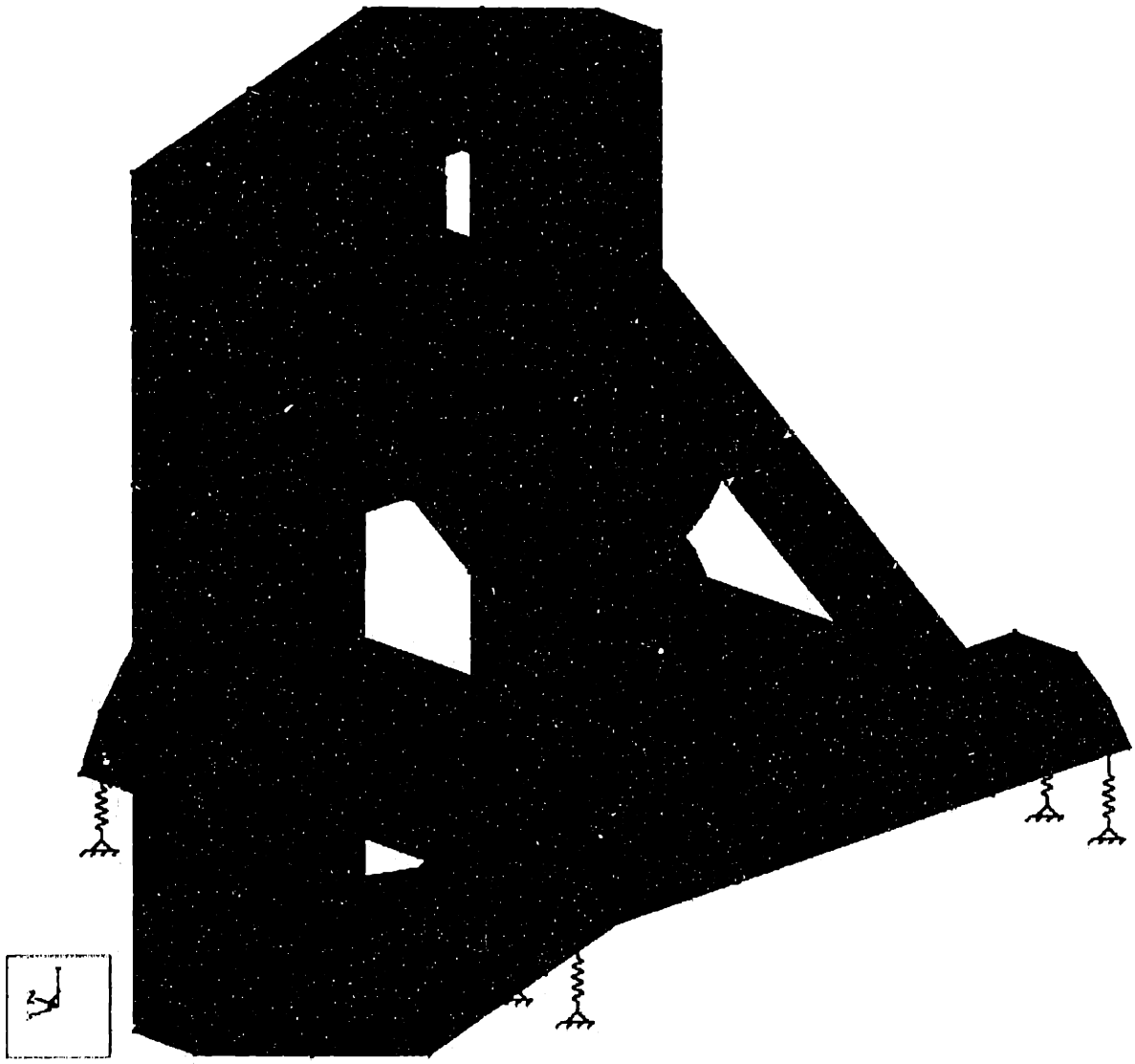


Figure 9-17 Column model with the spindle carrier in the upper position.

Mode	*Mode Shape*	*Y = 0*	*Y = 325*	*Y = 650*
1	Rigid body Z-translation	0	0	0
2	Y-rotation of spindle carrier	76.9 Hz	74.0 Hz	68.5 Hz
3	X- rotation of spindle carrier	85.4 Hz	86.5 Hz	84.3 Hz
4	Y-rotation of spindle carrier, shear and twist of column in phase with carrier	127 Hz	128 Hz	113 Hz
5	X-rotation of carrier and column in phase	136 Hz	132 Hz	123 Hz

Table 9-10 Column-carrier mode shapes and frequencies for three positions of the Y-axis.

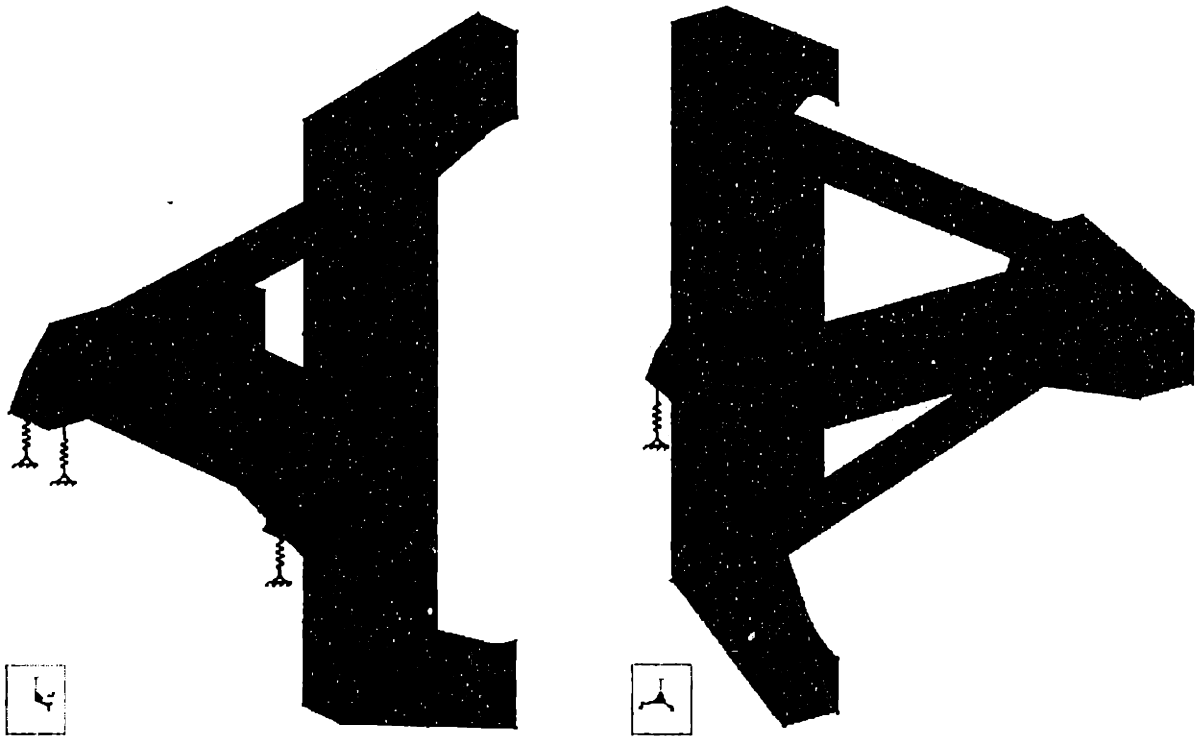


Figure 9-18 Left half of the column model, shown strictly for visualization purposes.

A static analysis was performed to assess the torsional stiffness of the column and the spindle error at three Y-positions caused when the column twists. Table 9-11 presents these results. The column is 9.4 times more compliant in torsion than the four bearings supporting it. The principal error caused by twisting the column is X-displacement and it increases with tool length. The magnitude is not very significant and amounts to a lever arm approximately 100 mm long.

Column Twist	*Unit*	*Y = 0*	*Y = 325*	*Y = 650*
Z-moment stiffness	N-m/μr	36.83	37.07	36.16
Equiv. bearing stiffness	N/μm	94.30	94.91	92.57
X-displacement ratio	μm/μr	-0.079	-0.008	0.062
Y-rotation ratio	μr/μr	-0.202	0.008	0.219
Z-rotation ratio	μr/μr	-0.014	-0.016	-0.033

Table 9-11 Column torsional stiffness and spindle error caused by twisting the column about the Z-axis.

A second static analysis was performed to assess the error motion of the moving Y-axis due to gravity. Figure 9-19 shows displacement and rotation at the spindle versus the Y-axis position. The centroid of the spindle carrier is well behind the plane of the bearings (which constrain Y in this model) resulting in a gravity moment that translates with the Y-axis. Even without the benefit of counterbalance, which would greatly reduce the magnitude of the gravity moment, the peak-to-valley errors are quite small. This indicates again that the column is relatively stiff compared to the bearings. An accelerating Y-axis produces the same effect, however the errors being dynamic are measured from zero. For example, an acceleration of one g would cause approximately 8 μm of error at the spindle

face. Compensating for dynamic errors is well beyond our ability today and poses considerable difficulty in the mapping process.

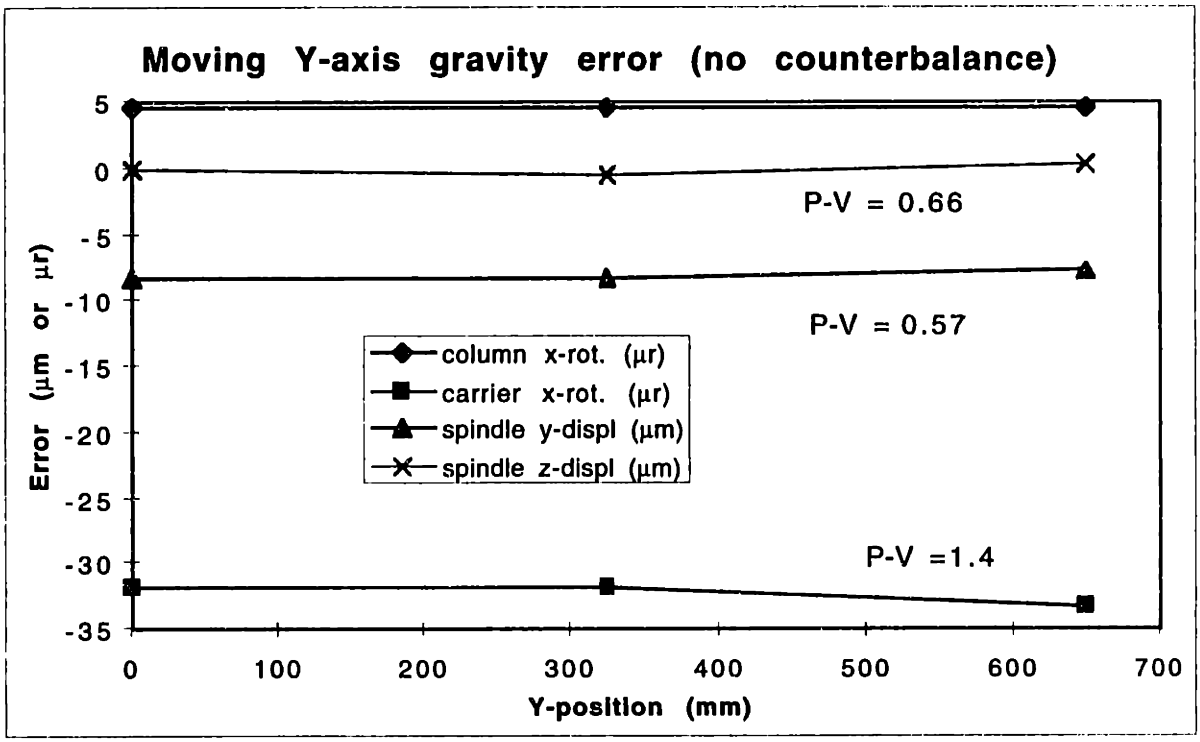


Figure 9-19 The gravity moment of the spindle carrier causes very little peak-to-valley error motion as the Y-axis travels. Proper counterbalance will reduce the gravity moment to a negligible concern, but an accelerating axis will produce an inertial moment and potentially significant errors.

9.5.3 The Work Carriage Model

The work carriage model, shown in Figure 9-20, consists primarily of shell elements except the solid elements in the pallet and cantilever bearing mounts. The housing has 20 mm thick cast iron walls. Two 20 mm thick steel rings connected at six node points represent the three-tooth pallet coupling; however, the B-axis compliance is not represented. For the analysis, three grounded springs each having 883 N/µm stiffness in Y and Z support the carriage and a Ø 500 by 700 mm cylindrical part weighing 600 kg.

Table 9-12 presents the results of the modal analysis. As indicated, the bearings and the bearing mounts play a large role in the system compliance. If the bearings were the only compliance source, then the frequency of a translational mode would be 239 Hz, or about twice as great as mode 4. Thus for mode 4, the bearings account only for one-fourth the total compliance. Some compliance is an artifact of the model where springs connect to single nodes, but still there is obvious need to stiffen the bearing mounts, housing and pallet coupling. With the addition of a ball screw constraint, modes 1 and 3 would combine to form a new first mode with a frequency probably in the 50 Hz range.

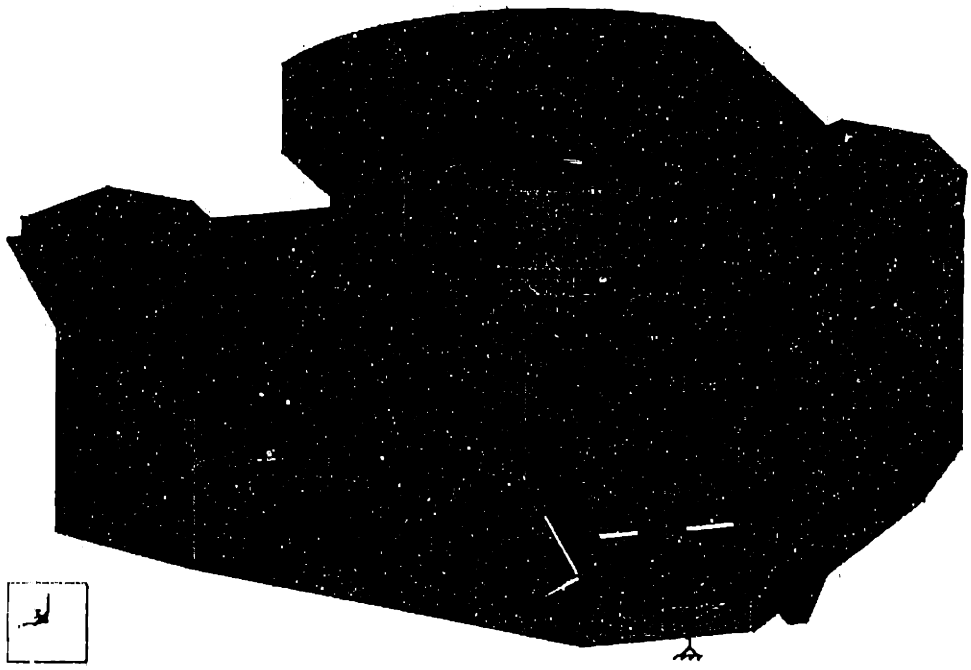


Figure 9-20 The work carriage model is supported on three grounded springs each having stiffness in Y and Z of 883 N/μm.

Mode	*Mode Shape*	*Frequency*
1	Rigid body X-translation	0
2	X-rotation, flexibility in bearings, mounts and coupling	59.0 Hz
3	Z-rotation, flexibility in bearings, mounts and coupling	72.8 Hz
4	Y-translation, flexibility in bearings mounts and housing	120 Hz
5	Y-rotation, flexibility in bearings	180 Hz

Table 9-12 Mode shapes and frequencies for the work carriage with a Ø 500 by 700 mm cylindrical part weighing 600 kg.

9.5.4 The Base Model

The base model went through many iterations to investigate different design ideas to increase stiffness or to simplify casting. Usually these goals conflicted and the design presented is neither the stiffest nor the simplest, but it provides a good compromise. Furthermore, this design has not been reviewed by a casting authority or a manufacturing authority. These reviews are essential for the final design.

The base model shown in Figure 9-21 and Figure 9-22 consists entirely of 20 mm shell elements using cast iron as the material (E = 131,000 MPa). Representative masses for other machine components are not in this model. Table 9-13 shows the modal analysis results run with a kinematic support using 175 N/μm (1 lb/μin) stiff springs. The first and most significant flexible mode of the base is the torsional mode (mode 7 at 119 Hz). Several of the various models had torsional modes around 130 Hz. These models had features that complicated the casting in some way and therefore were removed from the

final concept design. For example, a pair of plates across the chip valley helps to brace this area. The crosswise chip valley turned out to be a weakness and a complication in this design. An approach that now seems obviously better is to run the chip valley down the center and out the back of the machine. This would open up options in the design that were not previously considered and would positively affect the torsional and higher modes.[1]

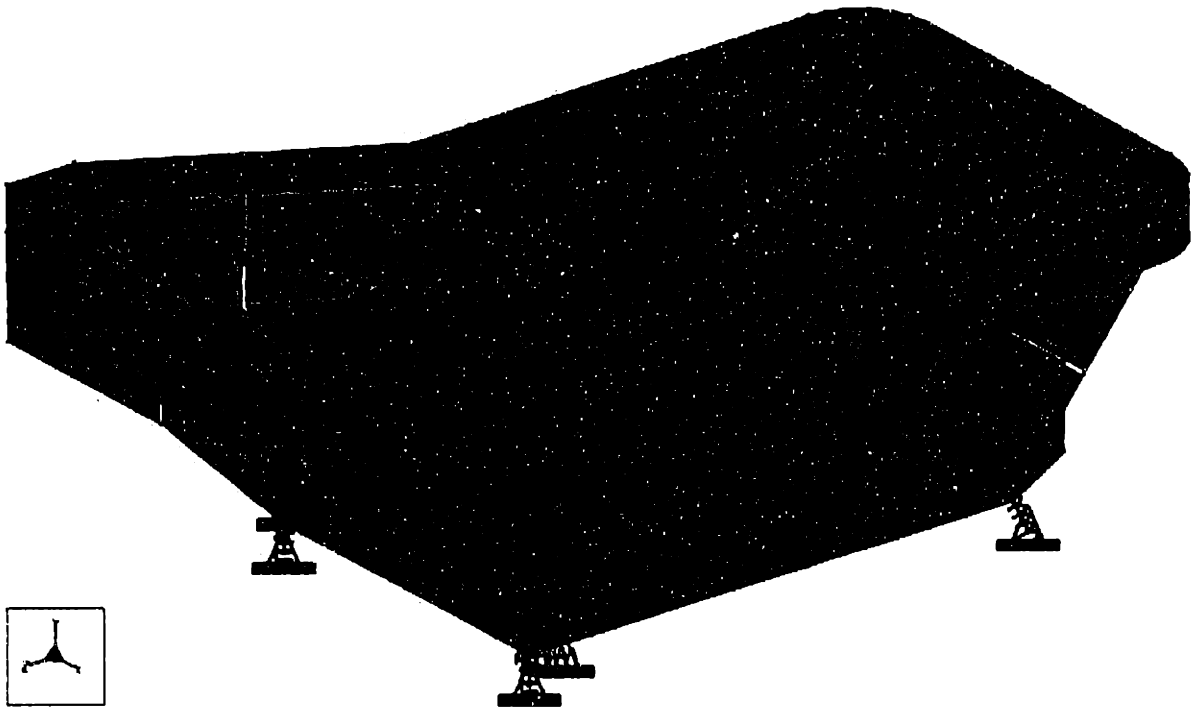


Figure 9-21 A downward view shows a tub-like base model supported on kinematic springs. The high walls support the column near its centroid and provide a very deep section to reduce gravity deflections.

Another problem area is where the sides flare outward to provide room for the work carriage and X-axis way covers. The sides and the long span that connects them across the front are very active in mode 8. Moving the support points wider helps this mode but makes gravity deflections larger for the X-axis. The need to exchange pallets limits the height of the front of the machine to its present value, but further optimization of the section is possible. The change recommended to the chip valley should also help this problem since from a side view mode 8 appears as a bending mode in this area. The present design has space provided for a single way cover per side over both X-axis rails and the ball screw. Each side wall makes a step from inside (to support the column) to outside (to provide the space), which contributes to the flexibility of mode 8 and also complicates the casting. Changing to a center chip valley would favor individual way covers that could easily protrude through small holes in the sides, thus eliminating the need for the walls to step

[1] While I would like to investigate this idea fully, it would require far too much rework (principally analysis work). At this point, I can only note possible ways to improve upon the concept design presented here.

outward. The lower X-axis rail would have to rise about 75 mm in order to extend the 15° sloping surfaces from the column area to the front underneath the work carriage. To stiffen the carriage bearing mounts, the upper X-axis rail should lower, probably to the same level as the other rail. A sheet of flowing coolant introduced around the inside perimeter of the base would promote chip flow to the valley and provide effective thermal control.

Mode	*Mode Shape*	*Frequency*
1	Rigid-body rotation parallel to X and 2 meters below the base	29.1 Hz
2	Rigid-body rotation about the rear I.C. and front supports	33.5 Hz
3	Rigid-body rotation parallel to Y and 1 meter in front of the base	43.8 Hz
4	Rigid-body rotation parallel to X and at the upper front edge	52.5 Hz
5	Rigid-body rotation about 45° Y-Z axis through the base center	59.6 Hz
6	Rigid-body rotation parallel to X and 600 mm behind center	63.8 Hz
7	Torsion parallel to Z axis through the base center	119 Hz
8	A saddle shaped with bending across the front end	156 Hz
9	Local bending of rear ribs out of phase	172 Hz
10	Local bending of rear ribs in phase	174 Hz

Table 9-13 Mode shapes and frequencies for the base supported on 1 lb/µin kinematic springs without the mass of the column or carriage.

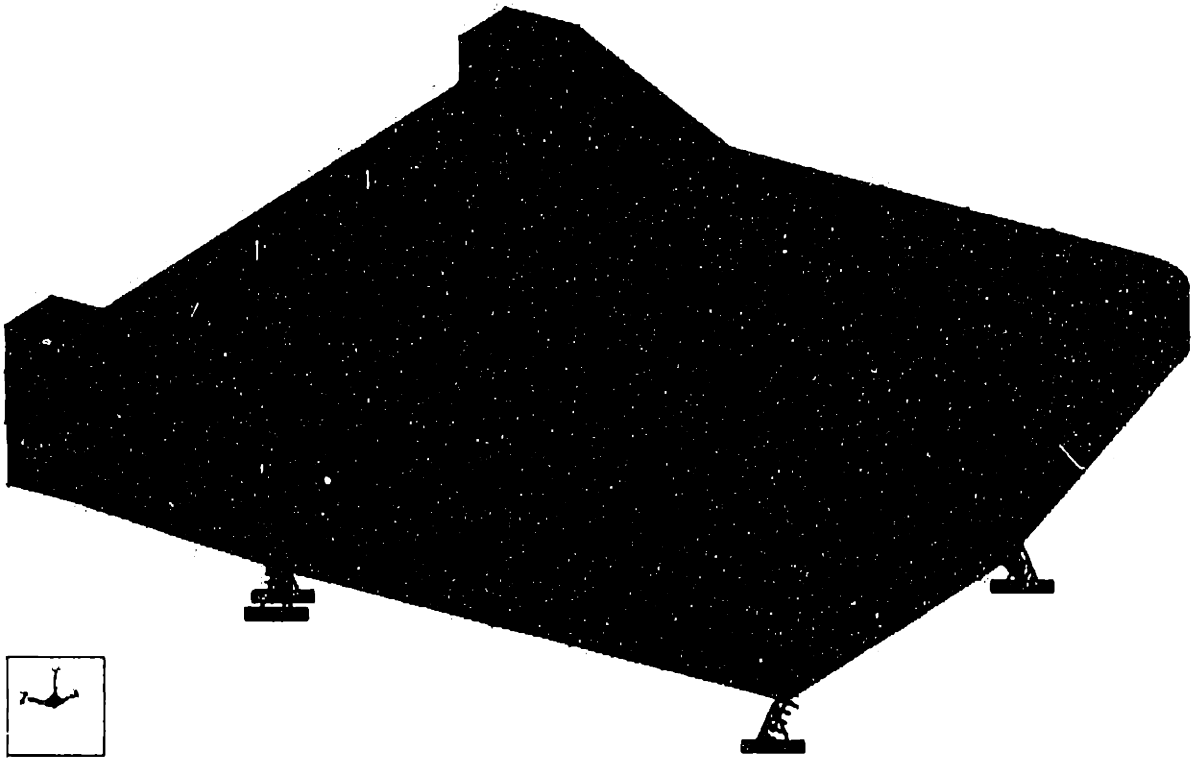


Figure 9-22 An upward view shows a diagonal pattern of core holes, which provide access to remove sand with minimal loss of shear stiffness. Exterior ribs and spring supports are more clearly visible.

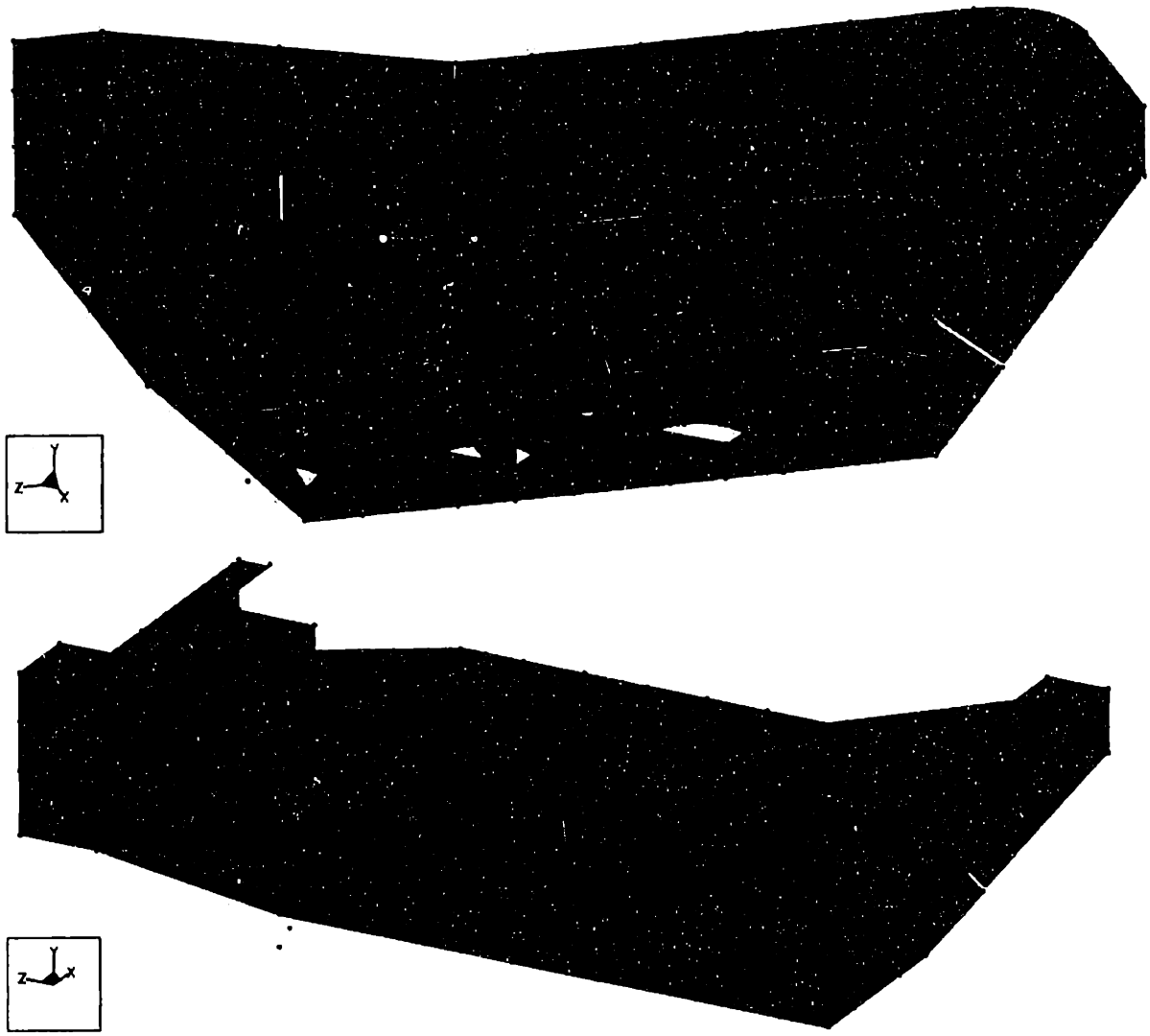


Figure 9-23 Views of half a base model show details of the interior important to its construction. Diagonal ribs on the lower plate help to restore shear stiffness lost to core holes and prevent local plate modes. Interior diagonals join to exterior walls where there are exterior ribs to provide moment connections.

Any base mounted kinematically requires outstanding torsional stiffness. The figures show the importance placed on making the bottom structurally closed. Several patterns of openings were modeled but all featured diagonal members for shear stiffness. The difference between small round or square holes in a diagonal pattern and larger triangular openings was not significant. The pattern shown works well with the particular proportion of the rectangular bottom and the external ribs. A static analysis of the base predicts a torsional stiffness between front and rear supports of 159 N-m/μr. A four-point mount with the equivalent torsional stiffness would require supports with 441 N/μm (2.5 lb/μin) stiffness. While this sounds respectable, the frictional overconstraint in a wide vee could generate 5 μr of twist between front and rear supports using a friction coefficient of 0.05 (Teflon). The plan to dynamically couple the base, particularly the torsional mode, to the foundation must include a way to limit the static frictional constraint to an acceptable

level. The use of a viscoelastic material and a low-friction material or lubricant seems promising. It is also possible to adjust the twist in the base by providing an adjustment to the location of the instant center such that a gravity moment exists on the base. This might be useful to correct manufacturing errors or nonsymmetric loads.

9.5.5 The Assembly Model

The assembly models consist of the previously discussed component models arranged in four different configurations, where the Y and Z axes are at either extreme position with the X axis centered. Figure 9-24 shows the configuration identified as (x, y, z) = (0, 650, 600). There is an inconsistency between the coordinate system shown in the figure and the general convention used in the machine tool industry. Displacements are reported in the coordinate system shown with positive Z towards the front of the machine. Reference to the axis position is opposite this with positive Z being a move away from the workpiece.

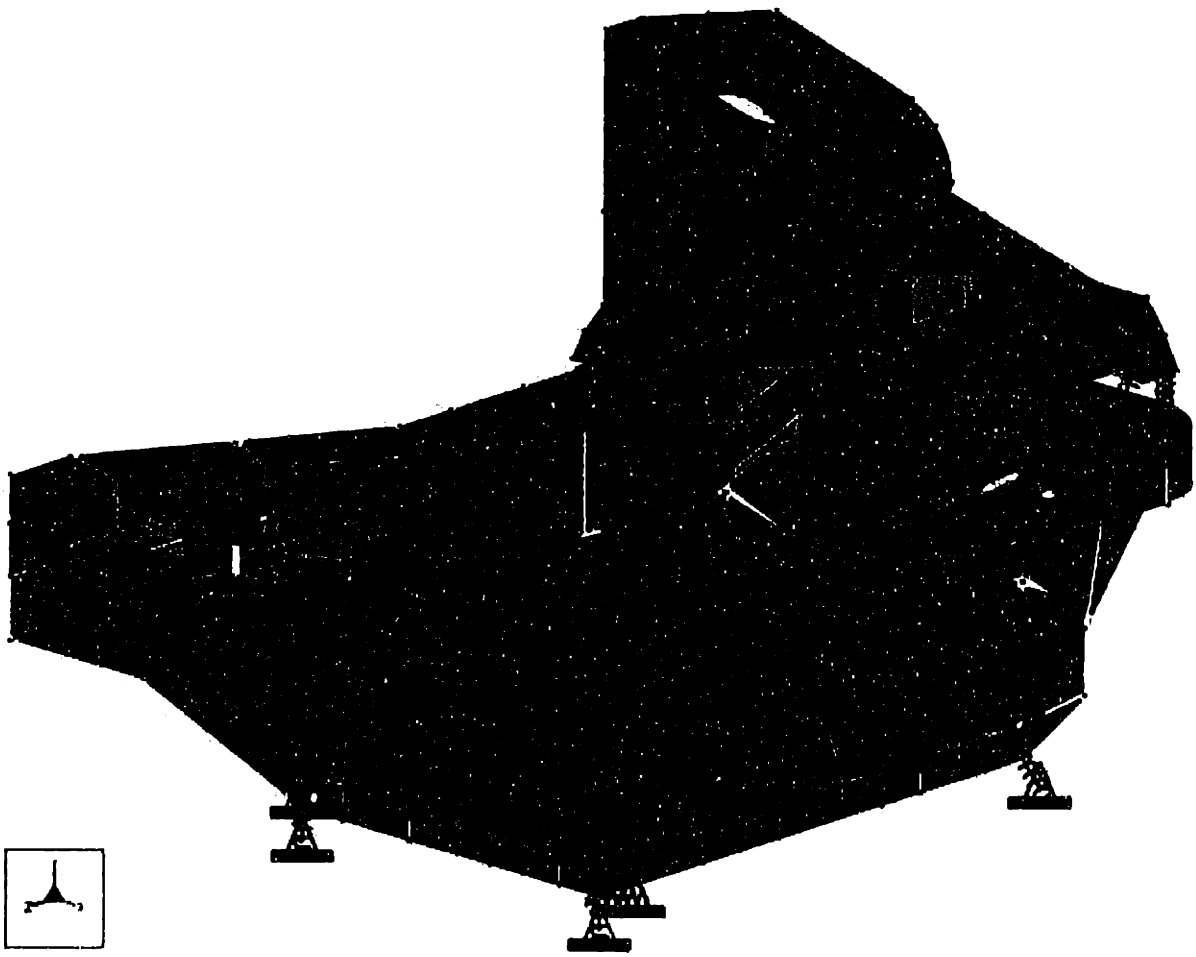


Figure 9-24 Assembly of the component models in a configuration (x, y, z) = (0, 650, 600).

The spring elements used in the assembly are the same as the respective springs used in the component models with the exception that those representing bearings have equal stiffness in X, Y and Z. Thus there are no purely rigid-body modes although the first

few support modes show very little deformation in the machine structure. We should expect the four configurations to share very similar modes but also to have unique modes. This is evident in the mode shapes and frequencies in Table 9-14.

The first several flexible modes occur between 45 and 57 Hz, and the principal flexibility in each is the work carriage. This analysis includes a Ø 500 by 700 mm cylindrical part weighing 600 kg. As mentioned previously, the flexibility in the carriage requires additional work to bring those modes to higher frequency. It requires stiffening so that twisting of the base becomes the principal flexibility. The next higher mode involves the spindle carrier with frequency between 82 and 89 Hz. The principal flexibility is the Y-axis bearings and it is driven by the motor's large moment of inertia. Still higher modes bring in movement of the column on its bearings and more complicated movement of multiple components.

Mode Shape	x = y = z =	0 0 0	0 0 600	0 650 0	0 650 600
Rigid-body Z-translation and X-rotation		20	20	20	19
Rigid-body Z-rotation		22	23	21	21
Nearly rigid-body Y-rotation		39	35	40	36
Nearly rigid-body X-rotation			38		38
Nearly rigid-body Y-translation		41		41	
X-rotation of carriage in phase with rigid base and column		48	48		
Z-rotation of carriage out of phase with base and column		51	48	47	45
X-rotation of column and base in phase, carriage is nearly still				47	47
X-rotation of work carriage		55	54	52	51
Z-rotation of work carriage		56	57	56	57
Y-rotation of spindle carrier		85	89	82	85
X-rotation of carriage and column in phase, base out of phase			94		85
X-rotation of spindle out of phase with column and carriage, Y-translation of column and carriage out of phase, base is steady				89	
X-rotation of spindle out of phase with base, Y-translation of carriage and column out of phase with base		90			

Table 9-14 Assembly mode shapes and frequencies.

9.5.6 Tool-to-Work Compliance

In addition to the modal analysis performed on the assembly, a static analysis was performed to estimate the tool-to-work compliance of the machine in the same four configurations. Separate loads were applied to the spindle face and the pallet face and later combined in such a way as to simulate a force at various tool lengths. Likewise, the displacements measured at the spindle face and the pallet face were combined to provide the

displacement at the tool point relative to the pallet. The effect of this approach is to ignore any compliance that would exist in the tooling and fixturing. In addition, the model does not account for compliance in the spindle rotor and spindle bearings.

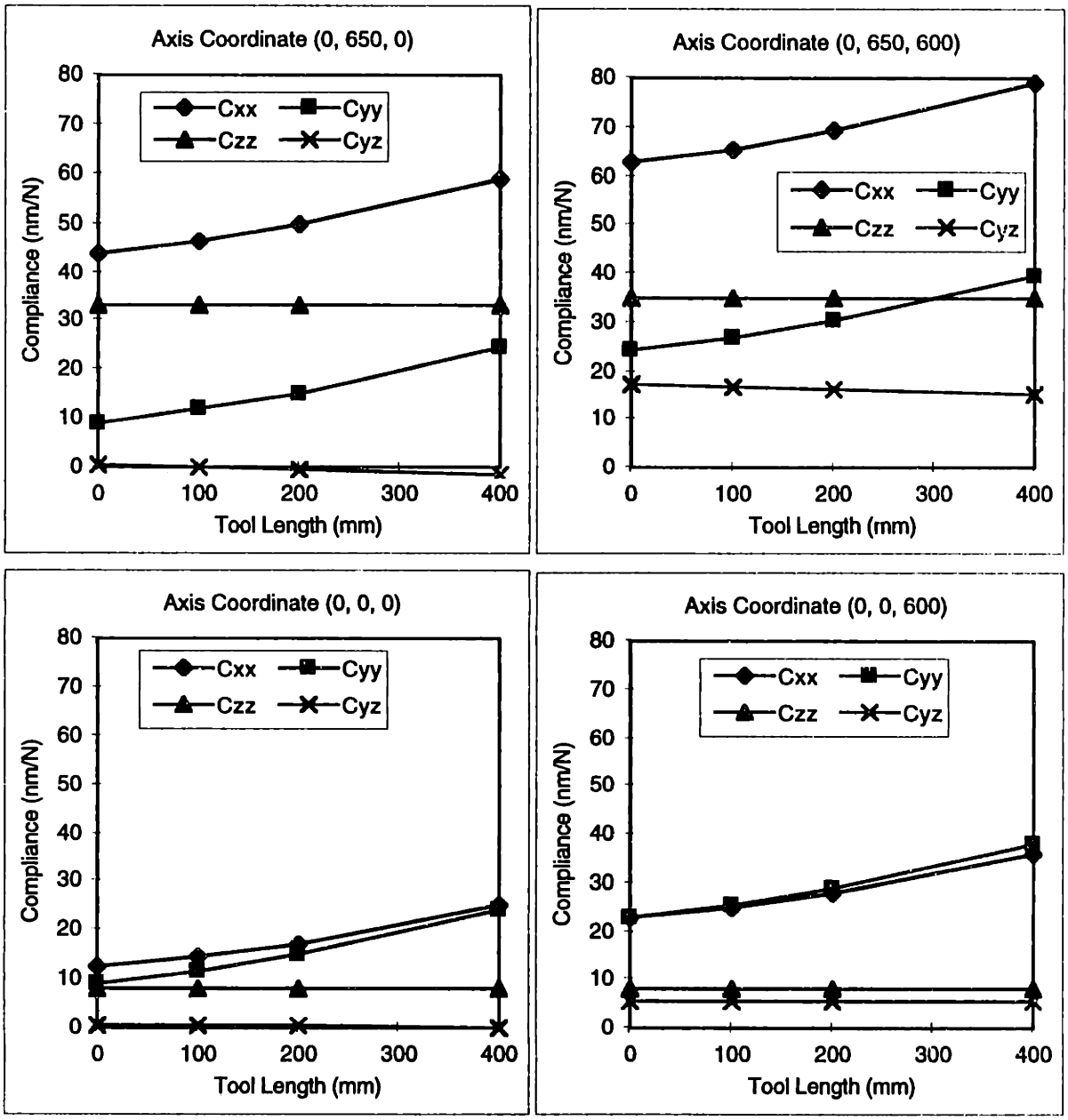


Figure 9-25 Tool-to-work compliance for the assembly at extreme y-z ranges.

Figure 9-25 shows the results for tool lengths between 0 and 400 mm. Only one cross compliance C_{yz} is nonzero when the X-axis is centered in its travel. Away from center, the other two cross compliances C_{xz} and C_{xy} should be similar in magnitude to C_{yz}. The normal compliances C_{xx}, C_{yy} and C_{zz} are generally of greater interest. The goal from the design requirements is 20 nm/N, which includes stiff tooling 100 mm long. Thus to be consistent, we should compare the results at 100 mm to a goal of 15 nm/N. Only the configuration (0, 0, 0) meets this goal. In other configurations, the compliance in the work

carriage is exaggerated by the distance between the tool and the pallet. Reinforcing the carriage will go a long way toward meeting the goal at extreme ranges of travel.

9.5.7 Predicted Error Motion

The error motion caused by the weight of moving axes and/or a twisting moment to the base was determined by applying sets of loads to the base model and processing the resulting deflections to get equivalent tool-to-work displacements and rotations. This method becomes complicated if a moving axis such as the column is overconstrained to the base because coupling effects need to be included. The advantage is that a simpler finite element model needs only to be solved once for the different load cases. With sufficient computer capability, it is faster and more accurate to generate and solve models for all assembly configurations (48 in this example).

With all this information, presenting the results in a meaningful and simple way becomes a challenge. After considering many different ways to plot error motion, the most intuitive approach was to plot only angular error motion as a function of the axis travel. Plots of angular error motion tell most clearly how the machine deforms as it articulates. Furthermore, the angular errors are more difficult or expensive to compensate in software. The complete error vector (δx, δy, δz, θx, θy, θz) over the entire range of travel is provided in tabular form. Table 9-15 distills this information into peak-to-valley (P-V) and root-sum-squared (RMS) errors. Concentrating on P-V errors, the first three rows are the maximum excursions of error components taken over the working volume due respectively to separate load cases: a 1.2 kN-m moment applied to the base, the moving work carriage with a 600 kg part, and the moving Z-axis column. Movement of the Y-axis does not affect the load on the base and the effect on the column is negligible with counterbalance. However, the error measured at the spindle may well change throughout the range of the Y-axis. The fourth and fifth rows are combined load cases as indicated. This is what actually would be measured if the machine were articulated throughout its volume.

P-V error over volume	δx (μm)	δy (μm)	δz (μm)	θx (μr)	θy (μr)	θz (μr)
1 kN x 1.2 m z-moment	6.19	1.84	2.30	1.81	3.10	2.73
x-axis with 600 kg part	6.24	1.75	0.41	0.79	1.92	5.98
z-axis	5.39	9.38	14.33	10.56	3.74	3.16
x and z axes without part	7.91	10.13	14.24	10.62	4.71	6.13
x and z with 600 kg part	10.65	10.95	14.17	11.02	4.71	9.12
RMS error over volume	δx (μm)	δy (μm)	δz (μm)	θx (μr)	θy (μr)	θz (μr)
1 kN x 1.2 m z-moment	3.27	0.56	0.72	0.74	1.67	3.98
x-axis with 600 kg part	1.65	0.72	0.10	0.36	0.75	2.43
z-axis	1.54	3.47	4.07	3.85	0.97	1.20
x and z axes without part	2.33	3.50	4.08	3.86	1.33	2.33
x and z with 600 kg part	3.14	3.57	4.09	3.87	1.33	3.54

Table 9-15 Peak-to-valley and root-mean-square errors over the working volume due to translating axes or a moment applied to the base as indicated. Errors are measured at the spindle face relative to the pallet face.

Figure 9-26 shows how the angular error varies as the work carriage and 600 kg part translate. Whether the spindle face is at Z = 0 or Z = 600 has little effect on the measured error as one might expect. The X component of angular error or roll error is characteristic of a three-bearing slide. Typically the support under the single bearing deflects most at the center position; however, the mechanism here appears to be deflection of the two-bearing rail near the ends of travel. The Y component of angular error or yaw error is rather unexpected because it comes from movement of the column instead of the carriage. The Z component of angular error or pitch error is the largest of the three but has the least consequence unless the A-axis is at an extreme angle. Its sign is consistent with the two-bearing rail deflecting more at the ends of travel than at the center. This indicates a problem in the base that should be easy to fix by reinforcing directly under the rail. The present external rib shown vertically in Figure 9-22 should be in line with the end wall to directly brace the problem area.

Figure 9-27 shows how the angular error varies as the weight of the column and spindle carrier translate along the Z-axis. As expected, the X component of angular error has the greatest magnitude and very little dependence on the X-axis position. Due to symmetry, the Y and Z angular components are zero when the X-axis is on center. Off center the symmetry is broken and the work carriage reacts to the deflection of the base with Y and Z angular components in addition to the primary X component. The proposed change to reorient the chip valley down the center of the base should reduce the deflection substantially. A factor of two reduction in the X angular component would be very pleasing.

Table 9-16 through Table 9-18 provide the complete error vector (δx, δy, δz, θx, θy, θz) taken over the entire range of travel. The P-V and RMS errors given in Table 9-15 were calculated from these tables. The numbers are quite reasonable and well under the uncompensated accuracy specification, see Table 9-5. Specifically, there is no need to include part weight in the geometric error compensation. Two recommended changes to the base should further reduce the larger angular errors that are difficult to compensate. The only caution to mention is the issue concerning a frictional moment applied to the base through the wide vee constraint. Recall that the purpose of the wide vee and viscoelastic pads is to increase the dynamic stiffness of the torsional mode. If it supports a static moment different from when the machine was mapped, then the error map will be inaccurate to some level. It is a matter of managing the static moment to an acceptable level, which based on this error analysis is of order 800 N-m.

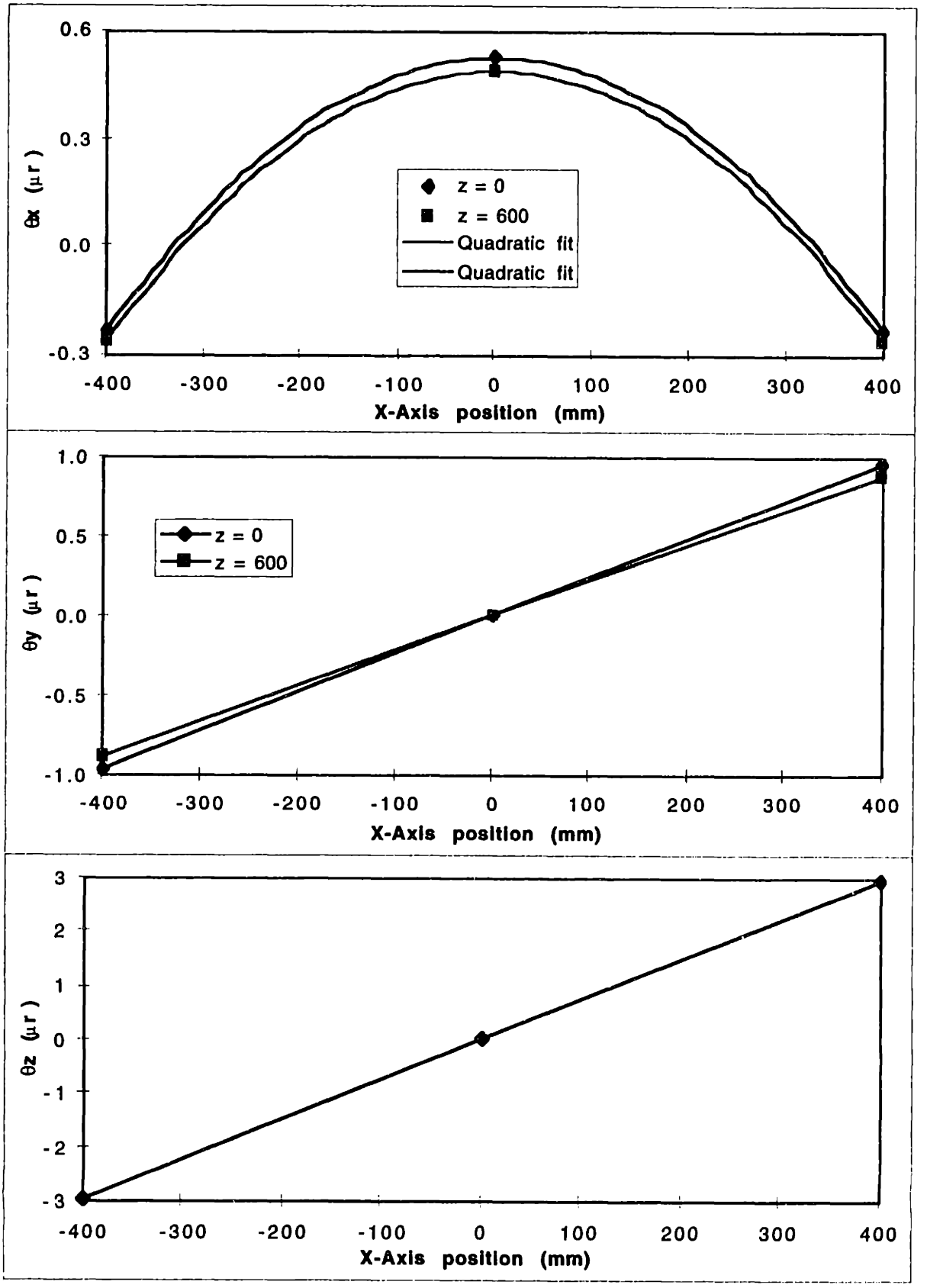


Figure 9-26 Angular error motion caused by the work carriage and 600 kg part translating along the X-axis. Errors measured at the spindle face are relative to the pallet face.

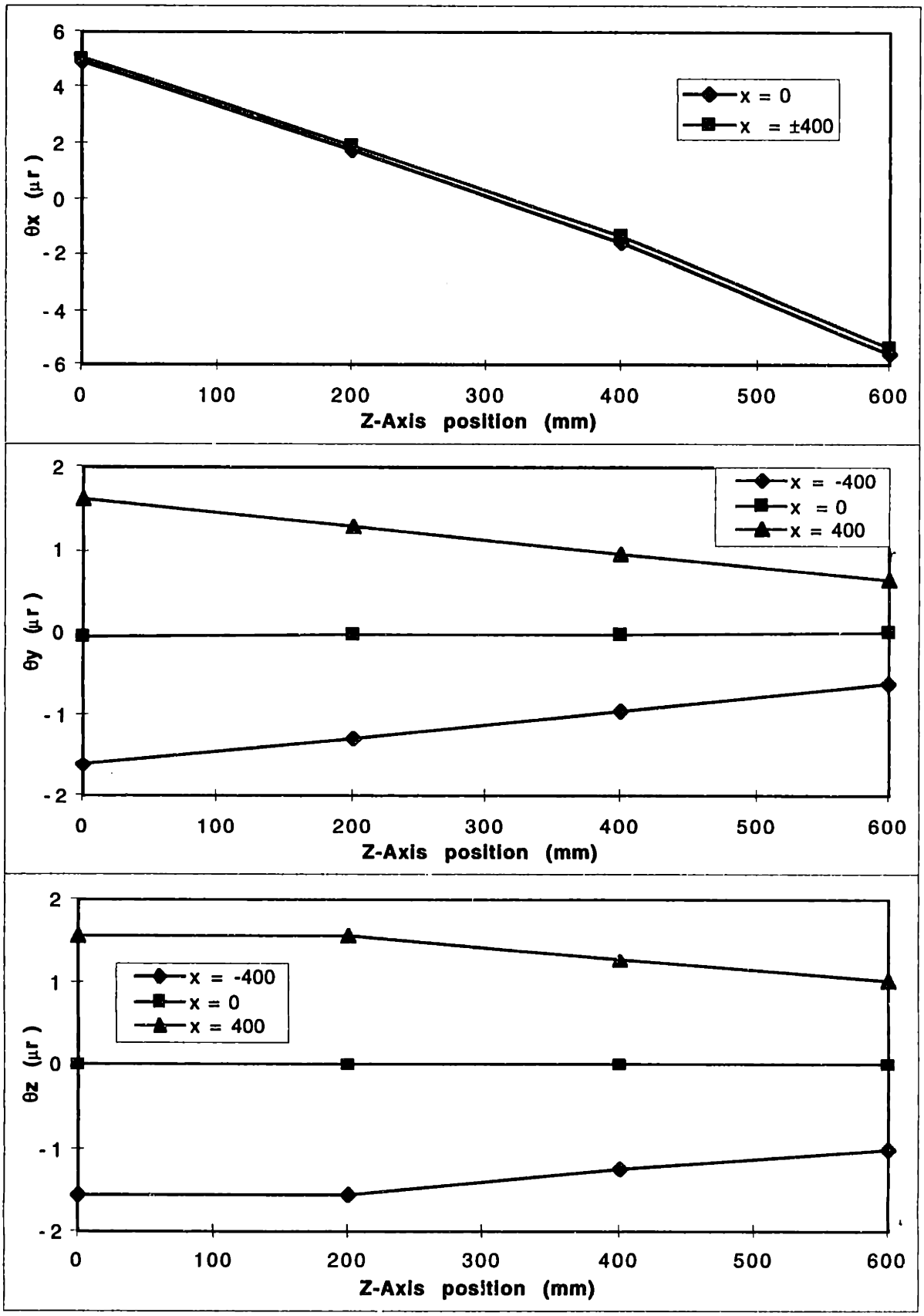


Figure 9-27 Angular error motion caused by the column translating along the Z-axis. Errors measured at the spindle face are relative to the pallet face.

Spindle error motion caused by 1.2 kN-m z-moment at rear supports				
δx(x,y,z) (μm)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	2.42	3.53	4.75	6.00
(x, y) = (-400, 325)	1.34	2.20	3.18	4.19
(x, y) = (-400, 0)	0.26	0.87	1.61	2.39
(x, y) = (0, 650)	2.80	3.93	5.17	6.45
(x, y) = (0, 325)	1.61	2.50	3.50	4.54
(x, y) = (0, 0)	0.43	1.07	1.83	2.64
(x, y) = (400, 650)	2.43	3.54	4.75	6.00
(x, y) = (400, 325)	1.35	2.20	3.18	4.20
(x, y) = (400, 0)	0.26	0.87	1.61	2.39
δy(x,y,z) (μm)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	-0.38	-0.56	-0.74	-0.92
(x, y) = (-400, 325)	-0.38	-0.56	-0.74	-0.92
(x, y) = (-400, 0)	-0.38	-0.56	-0.74	-0.92
(x, y) = (0, 650)	0.00	0.00	0.00	0.00
(x, y) = (0, 325)	0.00	0.00	0.00	0.00
(x, y) = (0, 0)	0.00	0.00	0.00	0.00
(x, y) = (400, 650)	0.38	0.56	0.74	0.92
(x, y) = (400, 325)	0.38	0.56	0.74	0.92
(x, y) = (400, 0)	0.38	0.56	0.74	0.92
δz(x,y,z) (μm)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	-1.14	-1.14	-1.15	-1.15
(x, y) = (-400, 325)	-0.84	-0.84	-0.86	-0.86
(x, y) = (-400, 0)	-0.55	-0.55	-0.56	-0.56
(x, y) = (0, 650)	0.00	0.00	-0.01	-0.01
(x, y) = (0, 325)	0.01	0.01	-0.01	-0.01
(x, y) = (0, 0)	0.01	0.01	-0.01	-0.01
(x, y) = (400, 650)	1.14	1.14	1.13	1.13
(x, y) = (400, 325)	0.85	0.85	0.84	0.84
(x, y) = (400, 0)	0.56	0.56	0.54	0.54
θx(x,y,z) (μr)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	-0.91	-0.91	-0.91	-0.91
(x, y) = (-400, 325)	-0.91	-0.91	-0.91	-0.91
(x, y) = (-400, 0)	-0.91	-0.91	-0.91	-0.91
(x, y) = (0, 650)	-0.01	-0.01	-0.01	-0.01
(x, y) = (0, 325)	-0.01	-0.01	-0.01	-0.01
(x, y) = (0, 0)	-0.01	-0.01	-0.01	-0.01
(x, y) = (400, 650)	0.90	0.90	0.90	0.90
(x, y) = (400, 325)	0.90	0.90	0.90	0.90
(x, y) = (400, 0)	0.90	0.90	0.90	0.90
θy(x,y,z) (μr)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	-1.28	-1.01	-0.51	0.08
(x, y) = (-400, 325)	-2.08	-1.77	-1.22	-0.57
(x, y) = (-400, 0)	-2.88	-2.53	-1.92	-1.23
(x, y) = (0, 650)	-1.41	-1.14	-0.64	-0.05
(x, y) = (0, 325)	-2.21	-1.90	-1.35	-0.71
(x, y) = (0, 0)	-3.02	-2.66	-2.05	-1.36
(x, y) = (400, 650)	-1.28	-1.01	-0.51	0.08
(x, y) = (400, 325)	-2.09	-1.77	-1.22	-0.58
(x, y) = (400, 0)	-2.89	-2.53	-1.92	-1.23
θz(x,y,z) (μr)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	-2.63	-3.44	-4.22	-4.98
(x, y) = (-400, 325)	-2.57	-3.38	-4.16	-4.93
(x, y) = (-400, 0)	-2.56	-3.37	-4.15	-4.92
(x, y) = (0, 650)	-2.94	-3.75	-4.53	-5.29
(x, y) = (0, 325)	-2.88	-3.69	-4.47	-5.24
(x, y) = (0, 0)	-2.87	-3.68	-4.46	-5.23
(x, y) = (400, 650)	-2.64	-3.44	-4.22	-4.99
(x, y) = (400, 325)	-2.58	-3.38	-4.17	-4.94
(x, y) = (400, 0)	-2.57	-3.37	-4.16	-4.93

Table 9-16 Error motion caused by 1.2 kN-m twist applied to the base through the supports.

Spindle error motion caused by moving column and carriage without part				
δx(x,y,z) (μm)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	3.93	3.96	3.66	3.24
(x, y) = (-400, 325)	2.87	2.93	2.74	2.43
(x, y) = (-400, 0)	1.80	1.90	1.82	1.61
(x, y) = (0, 650)	-0.05	-0.04	-0.03	-0.01
(x, y) = (0, 325)	-0.05	-0.04	-0.02	-0.01
(x, y) = (0, 0)	-0.05	-0.03	-0.02	-0.01
(x, y) = (400, 650)	-3.93	-3.96	-3.66	-3.24
(x, y) = (400, 325)	-2.87	-2.93	-2.74	-2.43
(x, y) = (400, 0)	-1.80	-1.90	-1.82	-1.61
δy(x,y,z) (μm)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	-5.54	-1.17	1.73	3.78
(x, y) = (-400, 325)	-5.54	-1.17	1.73	3.78
(x, y) = (-400, 0)	-5.54	-1.17	1.73	3.78
(x, y) = (0, 650)	-4.52	-0.23	2.59	4.58
(x, y) = (0, 325)	-4.52	-0.23	2.59	4.58
(x, y) = (0, 0)	-4.52	-0.23	2.59	4.58
(x, y) = (400, 650)	-5.54	-1.17	1.73	3.78
(x, y) = (400, 325)	-5.54	-1.17	1.73	3.78
(x, y) = (400, 0)	-5.54	-1.17	1.73	3.78
δz(x,y,z) (μm)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	7.18	2.06	-2.14	-6.87
(x, y) = (-400, 325)	5.60	1.47	-1.67	-5.09
(x, y) = (-400, 0)	4.02	0.89	-1.20	-3.31
(x, y) = (0, 650)	7.07	1.89	-2.34	-7.06
(x, y) = (0, 325)	5.40	1.24	-1.91	-5.33
(x, y) = (0, 0)	3.72	0.59	-1.49	-3.59
(x, y) = (400, 650)	7.18	2.06	-2.14	-6.87
(x, y) = (400, 325)	5.60	1.47	-1.67	-5.09
(x, y) = (400, 0)	4.02	0.89	-1.20	-3.31
θx(x,y,z) (μr)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	4.87	1.80	-1.45	-5.47
(x, y) = (-400, 325)	4.87	1.80	-1.45	-5.47
(x, y) = (-400, 0)	4.87	1.80	-1.45	-5.47
(x, y) = (0, 650)	5.15	1.99	-1.30	-5.33
(x, y) = (0, 325)	5.15	1.99	-1.30	-5.33
(x, y) = (0, 0)	5.15	1.99	-1.30	-5.33
(x, y) = (400, 650)	4.87	1.80	-1.45	-5.47
(x, y) = (400, 325)	4.87	1.80	-1.45	-5.47
(x, y) = (400, 0)	4.87	1.80	-1.45	-5.47
θy(x,y,z) (μr)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	-1.87	-1.62	-1.34	-1.08
(x, y) = (-400, 325)	-2.11	-1.74	-1.40	-1.08
(x, y) = (-400, 0)	-2.35	-1.85	-1.46	-1.08
(x, y) = (0, 650)	-0.04	-0.03	-0.02	0.00
(x, y) = (0, 325)	-0.04	-0.03	-0.02	0.00
(x, y) = (0, 0)	-0.04	-0.04	-0.02	0.00
(x, y) = (400, 650)	1.87	1.62	1.34	1.08
(x, y) = (400, 325)	2.11	1.74	1.40	1.08
(x, y) = (400, 0)	2.35	1.85	1.46	1.08
θz(x,y,z) (μr)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	-3.07	-3.06	-2.77	-2.49
(x, y) = (-400, 325)	-3.05	-3.05	-2.77	-2.49
(x, y) = (-400, 0)	-3.04	-3.05	-2.77	-2.49
(x, y) = (0, 650)	0.01	0.02	0.01	0.01
(x, y) = (0, 325)	0.01	0.02	0.01	0.00
(x, y) = (0, 0)	0.01	0.02	0.01	0.00
(x, y) = (400, 650)	3.07	3.06	2.77	2.49
(x, y) = (400, 325)	3.05	3.05	2.77	2.49
(x, y) = (400, 0)	3.04	3.05	2.77	2.49

Table 9-17 Error motion caused by the column and work carriage without a part translating in Z and X.

Spindle error motion caused by moving column and carriage with 600 kg part				
δx(x,y,z) (μm)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	5.17	5.33	5.12	4.80
(x, y) = (-400, 325)	3.63	3.81	3.72	3.50
(x, y) = (-400, 0)	2.08	2.30	2.32	2.21
(x, y) = (0, 650)	-0.08	-0.07	-0.05	-0.04
(x, y) = (0, 325)	-0.07	-0.06	-0.04	-0.03
(x, y) = (0, 0)	-0.06	-0.05	-0.03	-0.03
(x, y) = (400, 650)	-5.17	-5.33	-5.12	-4.80
(x, y) = (400, 325)	-3.63	-3.81	-3.72	-3.50
(x, y) = (400, 0)	-2.08	-2.30	-2.32	-2.21
δy(x,y,z) (μm)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	-5.77	-1.41	1.47	3.49
(x, y) = (-400, 325)	-5.77	-1.41	1.47	3.49
(x, y) = (-400, 0)	-5.77	-1.41	1.47	3.49
(x, y) = (0, 650)	-4.11	0.25	3.12	5.17
(x, y) = (0, 325)	-4.11	0.25	3.12	5.17
(x, y) = (0, 0)	-4.11	0.25	3.12	5.17
(x, y) = (400, 650)	-5.77	-1.41	1.47	3.49
(x, y) = (400, 325)	-5.77	-1.41	1.47	3.49
(x, y) = (400, 0)	-5.77	-1.41	1.47	3.49
δz(x,y,z) (μm)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	7.16	2.01	-2.20	-6.93
(x, y) = (-400, 325)	5.62	1.47	-1.69	-5.11
(x, y) = (-400, 0)	4.07	0.92	-1.17	-3.29
(x, y) = (0, 650)	7.19	1.98	-2.26	-6.99
(x, y) = (0, 325)	5.43	1.25	-1.92	-5.33
(x, y) = (0, 0)	3.67	0.52	-1.57	-3.68
(x, y) = (400, 650)	7.16	2.01	-2.20	-6.93
(x, y) = (400, 325)	5.62	1.47	-1.69	-5.11
(x, y) = (400, 0)	4.07	0.92	-1.17	-3.29
θx(x,y,z) (μr)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	4.75	1.67	-1.58	-5.60
(x, y) = (-400, 325)	4.75	1.67	-1.58	-5.60
(x, y) = (-400, 0)	4.75	1.67	-1.58	-5.60
(x, y) = (0, 650)	5.41	2.24	-1.05	-5.09
(x, y) = (0, 325)	5.41	2.24	-1.05	-5.09
(x, y) = (0, 0)	5.41	2.24	-1.05	-5.09
(x, y) = (400, 650)	4.75	1.67	-1.58	-5.60
(x, y) = (400, 325)	4.75	1.67	-1.58	-5.60
(x, y) = (400, 0)	4.75	1.67	-1.58	-5.60
θy(x,y,z) (μr)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	-1.87	-1.62	-1.34	-1.08
(x, y) = (-400, 325)	-2.11	-1.74	-1.40	-1.08
(x, y) = (-400, 0)	-2.35	-1.85	-1.46	-1.08
(x, y) = (0, 650)	-0.04	-0.03	-0.02	0.00
(x, y) = (0, 325)	-0.04	-0.03	-0.02	0.00
(x, y) = (0, 0)	-0.04	-0.04	-0.02	0.00
(x, y) = (400, 650)	1.87	1.62	1.34	1.08
(x, y) = (400, 325)	2.11	1.74	1.40	1.08
(x, y) = (400, 0)	2.35	1.85	1.46	1.08
θz(x,y,z) (μr)	z = 0	z = 200	z = 400	z = 600
(x, y) = (-400, 650)	-4.55	-4.56	-4.27	-3.98
(x, y) = (-400, 325)	-4.53	-4.55	-4.26	-3.98
(x, y) = (-400, 0)	-4.53	-4.55	-4.26	-3.98
(x, y) = (0, 650)	0.03	0.03	0.02	0.02
(x, y) = (0, 325)	0.03	0.03	0.02	0.02
(x, y) = (0, 0)	0.03	0.03	0.02	0.02
(x, y) = (400, 650)	4.55	4.56	4.27	3.98
(x, y) = (400, 325)	4.53	4.55	4.26	3.98
(x, y) = (400, 0)	4.53	4.55	4.26	3.98

Table 9-18 Error motion caused by the column and carriage with a 600 kg part translating in Z and X.

9.6 Recommendations

This project, to develop a conceptual design for a high-precision, high-productivity machining center, was a mechanism to communicate knowledge and ideas about precision to a community of machine tool engineers that often focus first on productivity concerns. This conceptual design may never become hardware and if developed would likely evolve into something quite different. However, if no aspect of this work finds its way into production, then an incredible amount of work will be wasted. Perhaps the best way to close is by reviewing the most significant aspects as encouragement to put them into production equipment. This will include a few remaining thoughts on the conceptual design and the steps required for it to become a product.

1) Determinism: "Machine tools obey cause-and-effect relationships and do not vary randomly for no reason." Always be mindful of the metrology loop and think about the many possible error sources that could disturb it. Have good reasons for every aspect of the design and back them up with analysis.

2) Exact Constraint: Always begin a design thinking exact constraint. Additional constraints can often cost more so have a good reason for and fully understand the consequences of overconstraint.

3) Thermal Management: Adopt strategies for thermal management, particularly the simple ones. Usually the costs are incremental and the results are remarkable.

4) Structural Design: The placement of material in a structural design usually has the greatest impact on stiffness. Use material most effectively by placing it in tension, compression or shear rather than bending. Shear deflections are usually significant in machine tool structures yet most handbooks neglect them in beam bending formulas. Use simple finite element models early in the design process because the design will change as you come to understand its behavior.

5) The Maxim Column Study: Much thought went into improving the bifurcated column design on the Maxim. The main finding was the need for shear stiffness in the plane of the Y-axis bearings. This requires a substantial perimeter frame for a bifurcated column with particular attention placed on the corner joints where bending moments are highest. All faces must support shear if the column is to be independently stiff in torsion. Torsionally stiff side members (diagonal ribs on the Maxim) have little influence compared to the full cross section of a well-designed column. A completely different approach is to design the column to be torsionally compliant and dependent on the base for stiffness through a four-bearing support. The concept of the shear center may be used to arrange the bearings so that the column does not twist under a tool load.

6) Damping Techniques: Motivated by the column work but much broader in scope, we investigated viscoelastic constrained-layer damping both theoretically and

experimentally. Several other damping techniques were presented and found to require the proper impedance match between the structure and the damping mechanism.

7) Machining Process Model: A simple orthogonal cutting model was adapted for milling and coded into an Excel spreadsheet. It proved useful in developing specifications for the conceptual design.

8) High-Speed Spindle Tool: A possible solution to the problem of attaining very high spindle speeds on a medium to heavy duty spindle is to develop a high-speed spindle tool that steps up the speed, for example, from 7500 rpm to 30,000 rpm. A smaller, light-duty spindle is compatible with cutters that operate at very high speeds and generally is power limited rather than stiffness limited. Using the power of the main spindle, the spindle tool would be loaded by the tool changer with the proper cutter already installed. A stationary key is required to prevent rotation of the step-up transmission but this feature is valuable for other types of spindle attachments. A right-angle tool is common and a tool with an adjustable angle would be useful as a pseudo A-axis.

9) Strategies for the Conceptual Design: A number of useful thoughts and ideas are presented in Section 1.1. Included in this section is an analysis showing the nearly constant stiffness of a ball screw supported at each end. Temperature rise in a rotating screw becomes an issue because it changes the load between the two thrust bearings. Several ways to manage this problem are discussed. Another possibility is to rotate the nut rather than the screw but this has its own set of issues.

10) Configuration of Axes: Despite the attempt to be objective and complete, at least two biases were present in the selection of the configuration. First is the requirement of an optional A-axis. None of the main-stream horizontal machining centers have a built-in A-axis. For good reasons, an A-axis under the work ranked worse than under the tool. This led to the second bias that the spindle and motor must pivot about the A-axis as a unit. A constant-velocity universal joint is a reasonable way to decouple the motion of the spindle from the motor. Without these biases the selected configuration would be a little better but other configurations could be much better. In particular, configuration 10 (YB-ZXA) could become a winner with just a few changes. Referring now to the sketches in Appendix G, discard the overhead bridge that supports the X-axis and replace it with the Z-X combination used in configuration 5 (AB-ZXY).

11) Conceptual Machine Design: Several structural changes are recommended based on FEA results and intuition. The crosswise chip valley creates several complications in the design that are easy to avoid by instead running it down the center of the machine. This would allow the cross section of the base to blend smoothly from the column area to the work carriage area. The lower X-axis rail must rise as a result and the upper X-axis rail must lower to stiffen the bearing mounts on the work carriage. With these

changes, the two supports at the front end of the base should widen and the side walls should be continuous from top to bottom. None of the wall thicknesses have been optimized yet. Any of the structural components could be fabricated steel or cast iron. If there is little cost difference, then steel should be used for moving components. Several areas need further work before this becomes a really credible conceptual design. The spindle, A-axis if used and the B-axis need definition, and then the ball screws and way covers should be incorporated. In addition, concepts for the tool changer, work changer and guarding must be created. It then should be subjected to a thorough design review before starting detailed design.

intentionally blank

Bibliography

Abbé, Ernst, Journal for Instrumental Information, Vol. X, pp 446-8, 1890 46

Ashby, M. F. "Materials Selection in Mechanical Design" Acta Metallurgica et Materialia, Vol. 37, No. 5, pp. 1273-1293, 1989 89

Ashby, M. F. "Materials Selection in Mechanical Design" Pergamon Press, Oxford, UK, 1992 ..... 89, 90, 93

Baumgarten K. and Schloeglmann K. "Mechanical Gear Drive." U.S. Patent 4,953,417, Sept. 4, 1990 249

Birch, K. P. and Downs, M. J. "Correction to the Updated Edlen Equation for the Refractive Index of Air" Metrologia, Vol. 31, pp. 315-316, 1994 49

Blanding, Douglass. L. "Principles of Exact Constraint Mechanical Design" Eastman Kodak Company, Rochester, New York, 1992 ..... 68, 69, 74, 114, 174

Bryan, J. B. and McClure, E. R. "Heat vs Tolerances" American Machinist, Special Report No. 605, June 5, 1967 308

Bryan, James B. "Design and Construction of an Ultraprecision 84 inch Diamond Turning Machine" Precision Engineering, Vol. 1, No. 1, pp 13-17, 1979 (UCRL-81010 Preprint) 31

Bryan, James B. "The Abbé Principle Revisited -- An Updated Interpretation" Precision Engineering, Vol. 1, No. 3, pp 129-132, 1979 (UCRL-82591 Preprint) 46

Bryan, James B. "The power of deterministic thinking in machine tool accuracy" First International Machine Tool Engineers Conference, Tokyo, Japan, November 1984 (UCRL-91531 Preprint) 38, 45, 308, 438

Chetwynd, D. G. "Selection of structural materials for precision devices" Precision Engineering, Vol. 9, pp. 3-6, 1987 89

Cresswell R.C. "Tracking Antenna With Anti-Backlash Spring in Gear Train" U.S. Patent 3,665,482, May 23, 1972 249

Den Hartog, J. P. "Mechanical Vibrations" Fourth Edition, McGraw-Hill, 1956 165

Deresiewicz, H. "Oblique Contact of Nonspherical Elastic Bodies" Journal of Applied Mechanics, Trans. ASME, pp. 623-624, December, 1957 425

Domnez, Bloomquist, Hocken, Liu and Barash, "A General Methodology for Machine Tool Accuracy Enhancement by Error Compensation" Precision Engineering, Vol. 8, No. 4, pp. 187-196, 1986 66

Donaldson, Robert R. "A Simple Method for Separating Spindle Error from Test Ball Roundness Error" Annals of the CIRP, Vol. 21, pp. 125-126, January, 1972 55

Donaldson, Robert R. "The deterministic approach to machining accuracy" Society of Manufacturing Engineers Fabrication Technology Symposium, Golden, Colorado, November 1972 (UCRL-74243 Preprint) 39, 48, 83

Donaldson, Robert R. "Large Optics Diamond Turning Machine, Volume 1" Lawrence Livermore National Laboratory report UCRL-52812, Vol. 1, September, 1979 ..... 31, 40, 83

Donaldson, Robert R. "Error Budgets" Sec. 9.14, Vol. 5 of the Technology of Machine Tools, Lawrence Livermore National Laboratory report UCRL-52960-5, pp. (9.14) 1-14, October, 1980 40

Donaldson, R. R. and Patterson, S. R. "Design and Construction of a Large, Vertical Axis Diamond Turning Machine" Proceedings of the SPIE, Vol. 433, pp. 62-67, August, 1983 (UCRL-89738 Preprint) 31

Donaldson, R. R. and Maddus, A. S. "Design of a High-Performance Slide and Drive System for a Small Precision Machining Research Lathe" Annals of the CIRP, Madison, WN, August, 1984 (UCRL-90475 Preprint) 31

Edlen, Bengt "The Refractive Index of Air" Metrologia, Vol. 2, No. 2, pp. 71-80, 1965 49

Eman (or Ehmann), Wu and DeVries "A Generalized Geometric Error Model for Multi-Axis Machines" Annals of the CIRP, Vol. 36, No. 1, pp. 253-256, 1987 66

Ernst, H. and Merchant, M.E. "Chip Formation, Friction and High Quality Machined Surfaces," Surface Treatment of Metals, Transactions of the American Society of Metals, Vol. 29, pp. 299-378, 1941. 454

Estler, W. T. and Mcgrab, E. B. "Validation Metrology of the Large Optics Diamond Turning Machine" National Bureau of Standards report NBSIR 85-3182(R), 1985 ... 31

Estler, W. Tyler "Calibration and use of optical straightedges in the metrology of precision machines" Optical Engineering, Vol. 24, No. 3, pp. 372-379, May-June, 1985 .. 31, 50

Estler, Phillips, Borchardt, Hopp, Witzgall, Levenson, Eberhardt, McClain, Shen and Zhang "Error compensation for CMM touch trigger probes" Precision Engineering, Vol. 19, No. 2/3, pp. 85-97, 1996 180

Estler, Phillips, Borchardt, Hopp, Levenson, Eberhardt, McClain, Shen and Zhang "Practical aspects of touch-trigger probes" Precision Engineering, Vol. 21, No. 1, pp. 1-17, 1997 180

Evans, Chris J. "Precision Engineering: An Evolutionary View" Cranfield Press, Cranfield, England, 1989 38

Evans, Chris J. and Kestner, Robert N. "Test Optics Error Removal" Applied Optics, Vol. 35, No. 7, pp. 1015-1021, March, 1996 60

Evans, Hocken and Estler "Self-Calibration: Reversal, Redundancy, Error Separation, and Absolute Testing" Annals of the CIRP, Vol. 45, No. 2, pp. 617-634, 1996 ..... 49, 50, 59

Everman, Michael R. "Backlash Free Rotationally Adjustable Mount in the Nature of a Transmission" U.S. patent number 5,435,651, July 25, 1995 274

Fengler, W. H. "Anti-Backlash Speed-Reduction Gearset." U.S. Patent 3,494,215, Feb. 10, 1970 248

Flanders, H. and Price, J. J. "Calculus with Analytic Geometry" Academic Press, Inc. New York, NY, Ch. 15.6, pp. 751-759, 1978 397

Frey, Daniel, D. "Using Product Tolerances to Drive Manufacturing System Design" Ph.D. Thesis, M.I.T. Cambridge Massachusetts, 1997 297

Fukada, Shigeo "Microscopic behavior of leadscrew-nut system and its improvement" Proceedings of the American Society for Precision Engineering, Vol. 14, pp. 314-317, 1996 194

Furse, J. E. "Kinematic design of fine mechanisms in instruments" J. Phys. E:Sci. Instrum., Vol. 14, 1981 69

Hale, L. C. and Slocum, A. H. "Design of Anti-Backlash Transmissions for Precision Position Control Systems" Precision Engineering, Vol. 16, No. 4, pp. 244-258, October, 1994 (UCRL-JC-115115 Preprint) 249, 251

Hale, Layton C. "Determinism in Dice Throwing and the Transition to Chaos" Proceedings of the American Society for Precision Engineering, Vol. 12, pp. 272-275, 1995.... 438

Hale, Layton C. "Friction-Based Design of Kinematic Couplings" Proceedings of the American Society for Precision Engineering, Vol. 18, pp., 1998 (UCRL-JC-131297 Preprint) 205

Holman, J. P. "Heat Transfer" Fourth Edition, McGraw-Hill, New York, NY, 1976....85

Johnson, C. and Kienholz, D. "Finite Element Prediction of Damping in Structures with Constrained Viscoelastic Layers" American Institute of Aeronautics and Astronautics, Vol. 20, pp. 1284-1290, 1982 149

Johnson, K. L. "The Effect of a Tangential Contact Force Upon the Rolling Motion of an Elastic Sphere on a Plane" Journal of Applied Mechanics, Trans. ASME, Vol. 80, pp. 339-346, September 1958 ..... 425, 426

Johnson, K. L. "Contact Mechanics" Cambridge University Press, 1985.....206, 414, 420

Jones, R. V. "Parallel and rectilinear spring movements" J. Sci. Instrum., Vol. 28, pp. 38-41, 1951 175

Jones, R. V. "Some uses of elasticity in instrument design" J. Sci. Instrum., Vol. 39, pp. 193- 203, 1962 175

Kalker, J. J. "Rolling with Slip and Spin in the Presence of Dry Friction" Wear, Vol. 9, pp. 20-38, 1966 425

Kamm, Lawrence J. "Designing Cost-Efficient Mechanisms" McGraw-Hill, Inc., New York, 1990 174

Kanai, Yoshioka and Miyashita, "Feed and Frictional Load Characteristics of Carriage on Plain Bearing Guideways in the Range of Nanometer Positioning" Proceedings of the American Society for Precision Engineering, Vol. 4, pp. 80-84, 1991 242

Kelling, L. U. C. "Anti-Backlash Servomotor Drive System." U.S. Patent 3,833,847, Sept. 3, 1974 260

Kerwin, Edward, "Damping of Flexural Waves by a Constrained Viscoelastic Layer" Journal of the Acoustical Society of America, 21 (7), pp. 952-962, 1959 149

Kiridena, V. and Ferreira, P. "Mapping the Effects of Positioning Errors on the Volumetric Accuracy of Five-Axis CNC Machine Tools" Int. J. Mach. Tools Manufact., Vol. 33, No. 3, pp. 417-437, 1993.....66

Krulewich, Hale and Yordy, "Rapid Mapping of Volumetric Errors" Proceedings of the American Society for Precision Engineering, Vol. 12, pp. 408-411, 1995 (UCRL-JC-121688 Preprint).....32, 67, 296

Krulewich, D. A. "Rapid mapping of volumetric machine errors using distance measurements" Proc. of the CIRP International Seminar on Improving Machine Tool Performance, Vol. 2, pp. 487-496, July, 1998 (UCRL-JC-130154 Preprint)..... .....32, 67, 296

Loxham, J. "The commercial value of investigations into repeatability" Cranfield Unit for Precision Engineering, Bedford, England, 1970 (quoted in Bryan, 1984).....38

Marsh, Eric R. "An Integrated Approach to Structural Damping" Ph.D. Thesis, M.I.T. Cambridge Massachusetts, 1994..... 104, 149

Marsh, E. R. and Hale, L. C. "Damping Machine Tools with Imbedded Viscoelastic Materials Constrained by a Shear Tube" Proc. ASME Design Engineering Tech. Conf., ASME DE, Vol. 84-3, pp. 9-14, Boston, MA, 1995 149

Marsh, E. R. and Hale L. C. "Damping of Flexural Waves with Imbedded Viscoelastic Materials" ASME Journal of Vibration and Acoustics, Vol. 120, No. 1, pp. 188-193, 1998 149

McCue, Howard K. "The Motion Control System for the Large Optics Diamond Turning Machine (LODTM)" Proceedings of the SPIE, Vol. 433, pp. 68-75, August, 1983 (UCRL-89362 Preprint) 31

Michalec, George W. "Precision Gearing: Theory and Practice" John Wiley & Sons, New York, NY, Ch. 6, pp. 252-286, 1966 249

Mindlin, R. D. "Compliance of Elastic Bodies in Contact" Journal of Applied Mechanics, Trans. ASME, Vol. 71, pp. 259-268, September, 1949 425

Mindlin, Mason, Osmer and Deresiewicz, "Effects of an Oscillating Tangential Force on the Contact Surfaces of Elastic Spheres" Proceedings of the First U.S. national Congress of Applied Mechanics, Chicago, IL, pp. 203-208, 1951 425

Mindlin, R. D. and Deresiewicz, H. "Elastic Spheres in Contact Under Varying Oblique Forces" Journal of Applied Mechanics, Trans. ASME, Vol. 75, pp. 327-344, September, 1953.....423, 425

Moore, Wayne R. "Foundations of Mechanical Accuracy" The Moore Special Tool Company, Bridgeport, Connecticut, 1970.....38, 83

Nash, J. C. "Compact Numerical Methods for Computers (linear algebra and function minimization), Second Edition" Adam Hilger, Bristol and New York, 1990..... 384

Paros, J. M. and Weisbord, L. "Flexure Hinges" Machine Design, pp. 151-156, November 25, 1965 185

Patterson, Steven R. "Development of Precision Turning Capabilities at Lawrence Livermore National Laboratory" Third Biennial International Machine Tool Technical Conference, Chicago, IL, September, 1986 (UCRL-94719 Preprint) 31, 101

Patterson, Steven R. "Interferometric Measurement of the Dimensional Stability of Superinvar" Lawrence Livermore National Laboratory report UCRL-53787, February, 1988 31

Patterson, Badami, Lawton, Tajbakhsh "The Dimensional Stability of Lightly-loaded Epoxy Joints" Proceedings of the American Society for Precision Engineering, Vol. 18, pp. 384-386, October, 1998..... 226

Paul, Richard P "Robot Manipulators" The MIT Press, Cambridge, MA, 1981..... 360

Press, Teukolsky, Vetterling and Flannery "Numerical Recipes in C, Second Edition" Cambridge University Press, 1992.....384, 388

Pritschow, G. and Wurst, K.-H. "Systematic Design of Hexapods and other Parallel Link Systems" Annals of the CIRP, Vol. 46/1, pp. 291-295, 1997 242

Raugh, Michael R. "Absolute 2-D sub-micron metrology for electron beam lithography" SPIE, Vol. 480, Integrated Circuit Metrology, pp. 145-163, 1984.....62

Raugh, Michael R. "Absolute two-dimensional sub-micron metrology for electron beam lithography: A theory of calibration with applications" Precision Engineering, Vol. 7, No. 1, pp. 3-13, 1985 62

Raugh, Michael R. "Two-dimensional stage self-calibration: Role of symmetry and invariant sets of points" Journal of Vacuum Science and Technology B, Vol. 15, No. 6, pp. 2139-2145, 1997 62

Reshetov, L. "Self-Aligning Mechanisms" Translated from Russian by Leo M. Sachs, Mir Publishers, Moscow, 1982, 1986..... 174

Roblee, Jeffrey W. "Precision Temperature Control for Optics Manufacturing" 2nd International Technical Symposium on Optical and Electro-Optical Applied Science and Engineering, Cannes, France, November, 1985 (UCRL-93540 Preprint).....31

Rolt, F. H. "Gauges and Fine Measurements" Macmillan and Co. Limited, London, 1929 38

Ross, Ungar and Kerwin "Damping of Plate Flexural Vibrations by Means of Viscoelastic Laminae" Proc. Colloq. Structural Damping, ASME, 1959..... 149

Ryu, Gweon and Moon "Optimal design of a flexure hinge based $XY\theta$ wafer stage" Precision Engineering, Vol. 21, No. 1, pp. 18-28, July, 1997..... 184

Saaty, Thomas "The Analytic Hierarchy Process" McGraw-Hill, New York, 1980 104

Sartori, S. and Zhang, G. X. "Geometric error measurement and compensation of machines" Annals of the CIRP, Vol. 44, No. 2, pp. 1-11, 1995.....67

Shen, Y. L. "Comparison of Combinatorial Rules for Machine Error Budgets" Annals of the CIRP, Vol. 42, No. 1, pp. 619-622, 1993 42

Siddall, Graham J. "The Design and Performance of Flexure Pivots for Instruments" M. Sc. Thesis, University of Aberdeen, Scotland, 1970 175

Slocum, Alexander H. "Kinematic Couplings for Precision Fixturing - Part I - Formulation of Design Parameters" Precision Engineering, Vol. 10, No. 2, pp. 85-91, April, 1988

Slocum, A. H. and Donmez, A. "Kinematic Couplings for Precision Fixturing - Part II - Experimental Determination of Repeatability and Stiffness" Precision Engineering, Vol. 10, No. 3, pp. 115-121, July, 1988 177

Slocum, Alexander H. "Precision Machine Design" Prentice Hall, Englewood Cliffs, New Jersey, 1992..... 38, 40, 49, 69, 83, 94, 104, 175, 194, 205

Slocum, Marsh and Smith "A New Damper Design for Machine Tool Structures: the Replicated Internal Shear Damper" Precision Engineering, Vol. 16, No. 3, pp. 174-183, July, 1994 149

Smith, S. T. and Chetwynd, D. G. "Foundations of Ultra-Precision Mechanism Design" Gordon and Breach, 1992 ..... 38, 68, 69

Smith, Badami, Dale and Xu "Elliptical flexure hinges" Rev. Sci. Instrum, Vol. 68, No. 3, pp. 1474-1483, 1997..... 185

Smith, Stuart T. "Flexures" not yet published, supplied as course notes for tutorial, ASPE Conference, October, 1998..... 175, 199

Soons, J. A. and Schellekens, P. H. "On the Calibration of Multi-Axis Machines Using Distance Measurements" Proc. ISMQC, pp. 321-340, 1992..... 67

Soons, Theuws and Schellekens "Modeling the errors of multi-axis machines: a general methodology" Precision Engineering, Vol. 14, No. 1, pp. 5-19, January, 1992 67

Spragg, R. C. and Whitehouse "Accurate Calibration of Surface Texture and Roundness Measuring Instruments" Institution of Mechanical Engineers Proceedings, Vol. 182 Pt 3K, pp. 397-405, 1967-68 60

Stewart, D. "A Platform with Six Degrees of Freedom" Proc. Instn Mech Engrs, pp. 371-386, 1965-66..... 82, 242

Stewart, Ian "Does God Play Dice?: the Mathematics of Chaos" Basil Blackwell Inc., New York, NY, 1989 ..... 438, 451

Strang, Gilbert "Linear Algebra and Its Applications, Second Edition" Academic Press, Inc., New York, NY, 1980 384

Strang, Gilbert "Introduction to Applied Mathematics" Wellesley-Cambridge Press, Wellesley, MA, 1986..... 371, 384

Suh, Nam P. "The Principles of Design" Oxford University Press, New York, NY, 1990.. ..... 98, 104

Tajbakhsh, Hale, Malsbury, Jensen, Parker "Three-Degree-of-Freedom Optic Mount for Extreme Ultraviolet Lithography" Proceedings of the American Society for Precision Engineering, Vol. 18, pp. 359-362, October, 1998 229

Takac, Ye, Raugh, Pease, Berglund and Owen "Self-calibration in two-dimensions: the experiment" SPIE, Vol. 2725, pp. 130-146, 1996 62

Timoshenko, S. and Goodier, J. N. "Theory of Elasticity" McGraw-Hill, New York, 1951 200

Ungar, Eric E. "Loss Factors of Viscoelastically Damped Beam Structures" J. Acoust. Soc. Am. Vol. 34, No. 8, pp. 1082-1089, August, 1962 154

Vukobratovich, Daniel and Richard, Ralph M. "Flexure mounts for high-resolution optical elements" SPIE Vol. 959, Optomechanical and Electro-Optical Design of Industrial Systems, 1988 ..... 175, 183, 199

Weinstein, Warren D. "Flexure-Pivot Bearings, Part 1" Machine Design, pp. 150-157, June 10, 1965 175

Weinstein, Warren D. "Flexure-Pivot Bearings, Part 2" Machine Design, pp. 136-145, July 8, 1965 175

White G. "Early Epicyclic Reduction Gears." Mech. Mach. Theory, Vol. 24, No. 2, pp. 127-142, 1989 263

Whitney, D. E. "The Mathematics of Coordinated Control of Prosthetic Arms and Manipulators" Journal of Dynamic Systems, Measurement and Control, pp. 303-309, December, 1972 366

Xu, W. and King, T. "Flexure hinges for piezoactuator displacement amplifiers: flexibility, accuracy, and stress considerations" Precision Engineering, Vol. 19, No. 1, pp. 4-10, July, 1996 185

Ye, Takac, Berglund, Owen and Pease "An exact algorithm for self-calibration of two-dimensional precision metrology stages" Precision Engineering, Vol. 20, No. 1, pp. 16-32, January, 1997 62

Young, Warren C. "Roark's Formulas for Stress and Strain, Sixth Edition" McGraw-Hill, New York, 1989 ..... 199, 200

intentionally blank

A Transformation Matrices

Transformation matrices have been essential for much of the work described in this thesis. They help organize complex relationships into simple expressions that computers can solve with ease, yet they can be confusing to use especially when used occasionally. Thus a concise reference is important to have available. Much of this presentation is based on the work of [Paul, 1981] who treats transformation matrices in the context of robot manipulators. In addition, there are some useful formulations not usually found elsewhere.

Transformation matrices come in several types and sizes for doing different kinds of operations over different domains. In all cases the matrix is square and usually invertible. Coordinate transformation matrices are the most familiar and three types are presented. Each type has a particular use that will become clear through the presentation. A coordinate transformation matrix changes the representation of the same vector from one coordinate system (CS) to another. The inverse of the matrix gives the reverse transformation between the same CS's. A transformation matrix is expressed with respect to its base CS. Whether the transformation takes a vector in the new CS back to the base CS or vice versa is a matter of convention. The former is the more popular convention and the one used here. This choice expresses the angular orientation of the new CS as unit vectors in the columns of the [3 x 3] rotation matrix.

Sequential coordinate transformations are easy to represent by multiplying transformation matrices in the proper order. Equation A.1 shows a sequence from the base CS_0 to the newest CS_n. The fractional subscripts are useful to keep the proper order and sense of the matrices. Taken in the direction from left to right, each transformation matrix is expressed in the most current CS. From right to left, each transformation matrix is expressed in the base CS. These two views of the same process may be confusing at first but have important practical value when dealing with sequential transformations. Later examples will reinforce the two rules stated below.

$$\begin{aligned} \mathbf{v}_0 &= \mathbf{T}_{0/1} \cdot \mathbf{T}_{1/2} \cdots \mathbf{T}_{n-1/n} \cdot \mathbf{v}_n \\ &= \mathbf{T}_{0/n} \cdot \mathbf{v}_n \end{aligned} \tag{A.1}$$

Rule 1: Post multiply to transform in current coordinates.

Rule 2: Pre multiply to transform in base coordinates.

Transformation matrices are also useful for changing one type of vector to another within the same coordinate system. The first example of this is the cross product matrix, which for example, transforms a force vector acting on a lever arm to a moment vector in the same CS. Ironically though, its main use is as part of a coordinate transformation matrix for six-dimensional vectors. A six-dimensional vector describes three linear degrees of freedom and three angular degrees of freedom. As we shall see, the [6 x 6] coordinate

transformation matrix is useful in developing the stiffness matrix, which transforms displacement vectors to force vectors within the same CS.

A.1 The Rotation Matrix

The [3 x 3] rotation matrix unambiguously describes the angular orientation of a new CS with respect to the base CS. It provides the rotation transformation for any kind of three-dimensional vector (position, angular velocity, force, etc.) between two CS's as if they share a common origin.[1] The rotation matrix is orthonormal giving it the useful property that the transpose of the matrix is equal to the inverse. This is not true for the other types of transformation matrices presented later. The orthonormal property places six constraints on the nine elements of the rotation matrix: three orthogonality conditions and three unit-length conditions (the unit vectors of an orthogonal CS). Therefore, three independent parameters are sufficient to represent a given angular orientation, usually as a series of three angular dimensions taken in a specific order. There are multiple ways to arrive at the unique rotation matrix, and three such methods are presented here.

The rotation matrix is easiest to derive when the rotation occurs about any one coordinate axis. This is evident in Figure A-1 by contrasting the three separate rotations (a, b and c) to three sequential rotations (d). The columns of the rotation matrix express the axes of the rotated CS (represented by x', y', z') with respect to the base CS. Equations A.2, A.3 and A.4 give, respectively, the rotation matrices for separate rotations about x, y and z axes. Since the columns are unit length and orthogonal, the rotation matrix is orthonormal by definition. In addition, these single-axis rotation matrices are skew symmetric; so the inverse transformation is simply the reverse rotation. Sequential rotations are easy to compute by multiplying single-axis rotation matrices in the proper order.

$$\mathbf{R}_x(\theta_x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix} \tag{A.2}$$

$$\mathbf{R}_y(\theta_y) = \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y \\ 0 & 1 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y \end{bmatrix} \tag{A.3}$$

$$\mathbf{R}_z(\theta_z) = \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 \\ \sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{A.4}$$

---

[1] Two CS's that do not have a common origin may or may not require accounting for the offset depending on the quantity being rotated.

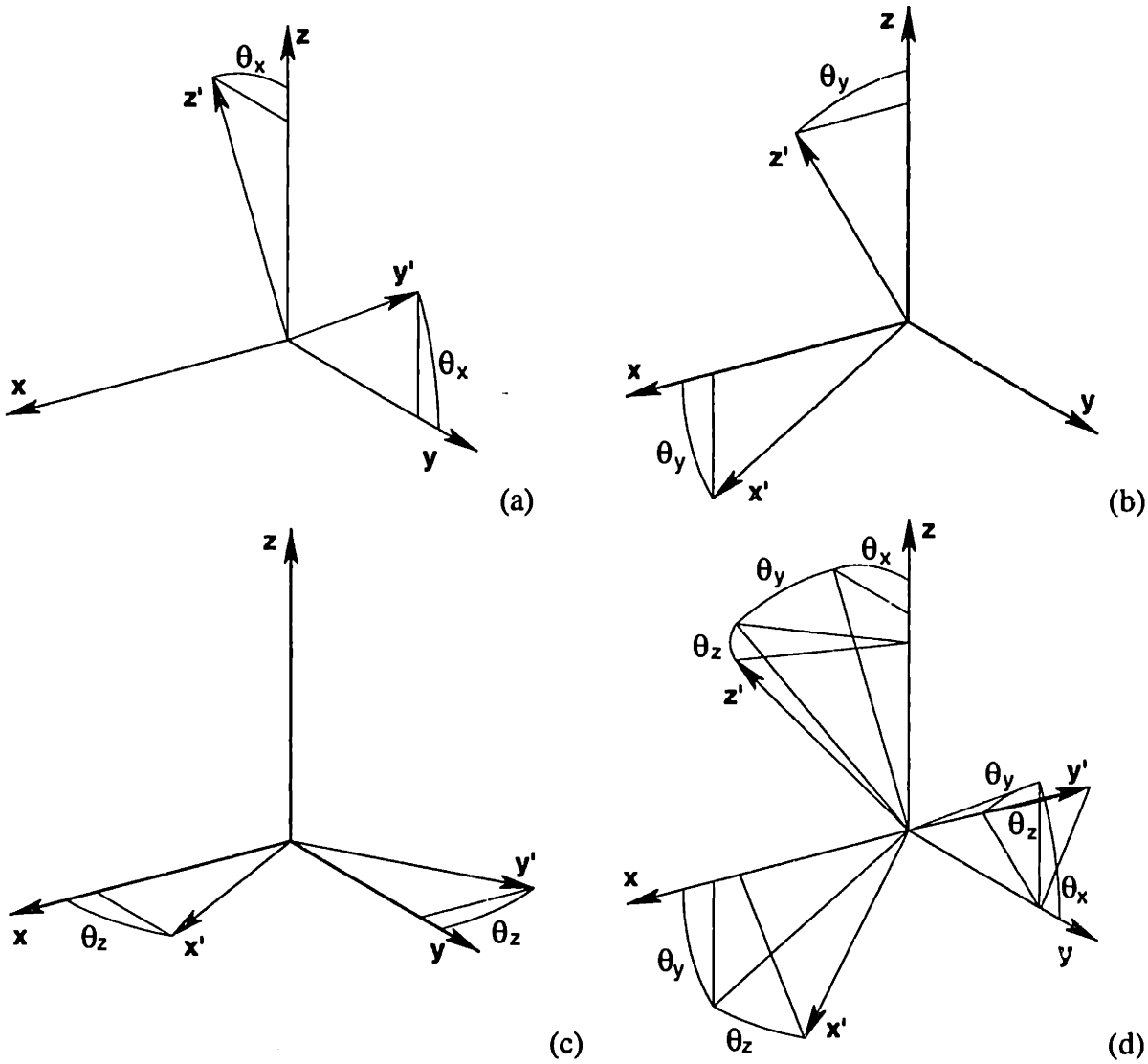


Figure A-1 A single rotation about x, y or z is easy to visualize, and the corresponding rotation matrix is simple to derive from the geometry. Sequential rotations about base axes x, y and z are more difficult to visualize, and the rotation matrix is much harder to derive directly from the geometry.

A natural sequence of rotations and the one depicted in Figure A-1 is rotation about the base axes in the order x, y, z. Invoking Rule 2 to pre multiply for base coordinates, Equation A.5 shows that the proper order goes from right to left. Rule 1 gives an entirely different interpretation of the same transformation going left to right: *roll* about the z body axis, *pitch* about the y body axis, and *yaw* about the x body axis. In practice, either view could turn out being easier to visualize for a particular case.

$$\mathbf{R}_{xyz}(\theta) = \mathbf{R}_z(\theta_z) \cdot \mathbf{R}_y(\theta_y) \cdot \mathbf{R}_x(\theta_x) \tag{A.5}$$

Equation A.6 shows that the inverse transformation reverses the order of rotations, as required by the rules of linear algebra. The transpose operator properly reverses the order and effectively changes the direction of each single-axis rotation (due to skew symmetry). The inverse transformation will *undo* the forward transformation. It is like

following directions in reverse; you have to reverse the order of turns in addition to changing right hand turns to left hand turns.

$$\begin{aligned}\mathbf{R}_{xyz}(\theta)^{-1} &= \mathbf{R}_x(\theta_x)^{-1}\cdot\mathbf{R}_y(\theta_y)^{-1}\cdot\mathbf{R}_z(\theta_z)^{-1}\\ &= \mathbf{R}_x(\theta_x)^{\mathrm{T}}\cdot\mathbf{R}_y(\theta_y)^{\mathrm{T}}\cdot\mathbf{R}_z(\theta_z)^{\mathrm{T}} = \mathbf{R}_{xyz}(\theta)^{\mathrm{T}}\\ &= \mathbf{R}_x(-\theta_x)\cdot\mathbf{R}_y(-\theta_y)\cdot\mathbf{R}_z(-\theta_z) \neq \mathbf{R}_{xyz}(-\theta)\end{aligned} \tag{A.6}$$

Euler chose a different set of three sequential rotations to represent an arbitrary orientation. Figure A-2 shows the Euler angles ϕ, θ and ψ taken about the coordinate axes z, y' and z''. Invoking Rule 1 to post multiply for current coordinates, Equation A.7 shows the proper order from left to right: rotate about the z axis, rotate about the y' axis, and rotate about the z'' axis. It differs from the roll-pitch-yaw representation only in the last rotation.

$$\mathbf{R}_{\text{Euler}}(\phi,\theta,\psi) = \mathbf{R}_z(\phi)\cdot\mathbf{R}_{y'}(\theta)\cdot\mathbf{R}_{z''}(\psi) \tag{A.7}$$

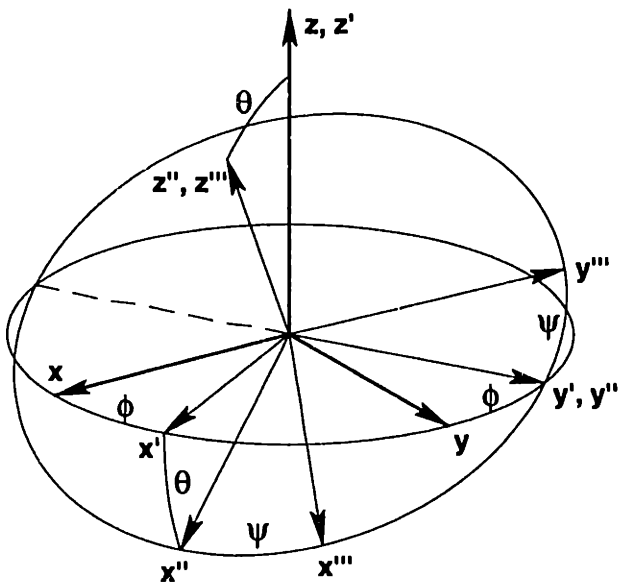


Figure A-2 Euler angles describe an arbitrary orientation with sequential rotations ϕ, θ, ψ about z, y', z''.

A third way to described an arbitrary orientation is with a single angle rotation about a fixed axis. The vector direction of the axis and the magnitude of the rotation can be expressed as an angle vector in the base CS. This transformation equivalently represents simultaneous rotations about the base x, y and z axes. It also may be constructed from a sequence of rotations as follows. The first two rotations establish a temporary CS that has its z axis aligned to the angle vector θ. Equation A.8 shows this transformation represented by $\mathbf{T}$ (a rotation matrix).[1] Now the desired rotation is simple to compute using $\mathbf{R}_z(\|\theta\|)$, but

[1] Use the two-argument (four-quadrant) arc tangent function to properly handle negative vector components and to avoid possible division by zero.

the effect of the first transformation **T** must be reversed after the rotation as Equation A.9 shows.[1] An algebraic expansion of this sequence eventually simplifies to a manageable result given in Equation A.10. Further simplifications are possible assuming the rotation is small. The approximation in Equation A.11 has third-order error for simultaneous rotations and somewhat greater error for sequential rotations of any order.

$$\mathbf{T} = \mathbf{R}_z\left(\arctan\frac{\theta_y}{\theta_x}\right)\cdot\mathbf{R}_y\left(\arctan\frac{\|(\theta_x,\theta_y)\|}{\theta_z}\right) \tag{A.8}$$

$$\mathbf{R}_{\text{axis}}(\theta) = \mathbf{T}\cdot\mathbf{R}_z(\|\theta\|)\cdot\mathbf{T}^{\mathrm{T}} \tag{A.9}$$

$$\mathbf{R}_{axis}(\theta) = \cos(\|\theta\|)\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix} + \frac{\sin(\|\theta\|)}{\|\theta\|}\begin{bmatrix}0&-\theta_z&\theta_y\\\theta_z&0&-\theta_x\\-\theta_y&\theta_x&0\end{bmatrix} + \frac{1-\cos(\|\theta\|)}{\|\theta\|^2}\begin{bmatrix}\theta_x\\\theta_y\\\theta_z\end{bmatrix}\cdot\begin{bmatrix}\theta_x&\theta_y&\theta_z\end{bmatrix} \tag{A.10}$$

$$\mathbf{R}_{axis}(\theta) \cong \begin{bmatrix}1&-\theta_z&\theta_y\\\theta_z&1&-\theta_x\\-\theta_y&\theta_x&1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}-(\theta_y^2+\theta_z^2)&\theta_x\theta_y&\theta_x\theta_z\\\theta_x\theta_y&-(\theta_x^2+\theta_z^2)&\theta_y\theta_z\\\theta_x\theta_z&\theta_y\theta_z&-(\theta_x^2+\theta_y^2)\end{bmatrix} \tag{A.11}$$

The rotation matrix is useful for dealing with the complexities of large-angle rotations between CS's. The addition of a translation or offset between CS's leads to more elaborate coordinate transformations such as the homogeneous transformation matrix and the [6 x 6] transformation matrix. Both of these have the rotation matrix embedded within their structures. Understanding the rotation matrix is requisite for understanding the remainder of this chapter.

A.2 The Inverse Problem, Finding the Angles of Rotation

On occasion it may be useful to recover a particular angle representation from an existing rotation matrix. This is a simple procedure for sequential rotations since the angles of rotation can be determined in reverse order. The first example is for sequential rotations about the base axes in the order x, y, z. It is helpful in following the derivation to refer to Figure A-1d. Recall that the columns of the rotation matrix express the axes of the rotated CS (represented by x', y', z') with respect to the base CS. Since the x' axis was transformed by just two rotations about y then z, θ_z is straight forward to recover using Equation A.12. Transforming the rotation matrix backwards by θ_z, as indicated by $\mathbf{R}'$ in

[1] This formulation satisfies the necessary condition that zero and full revolutions about the axis result in an identity matrix. Additional tests using orthogonal rotations are simple to verify through graphical means.

Equation A.13, simplifies the recovery of θ_y. A second transformation backwards by θ_y simplifies the recovery of θ_x from either y'or z' in Equation A.14.

$$\theta_z = \arctan\left(\frac{a_y}{a_x}\right) \qquad \mathbf{R} = \begin{bmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{bmatrix} \tag{A.12}$$

$$\theta_y = \arctan\left(\frac{-a'_z}{a'_x}\right) \qquad \mathbf{R}' = \mathbf{R}_z(\theta_z)^{\mathrm{T}} \cdot \mathbf{R} \tag{A.13}$$

$$\theta_x = \arctan\left(\frac{b''_z}{b''_y}\right) \qquad \mathbf{R}'' = \mathbf{R}_y(\theta_y)^{\mathrm{T}} \cdot \mathbf{R}' \tag{A.14}$$

The same method applies to recovering the Euler angles from a given rotation matrix; however, the rotations must occur about the base axes rather than the body axes. From this point of view, Figure A-3 shows the Euler angles now in opposite order: ψ about the z axis, θ about the y axis, and ϕ about the z axis. The first angle to recover is ϕ using the coordinates of z', as Equation A.15 shows. Transforming backwards by ϕ, as indicated by $\mathbf{R}'$ in Equation A.16, θ is recovered using z' again. A second transformation backwards by θ allows ψ to be recovered from either x'or y' in Equation A.17.

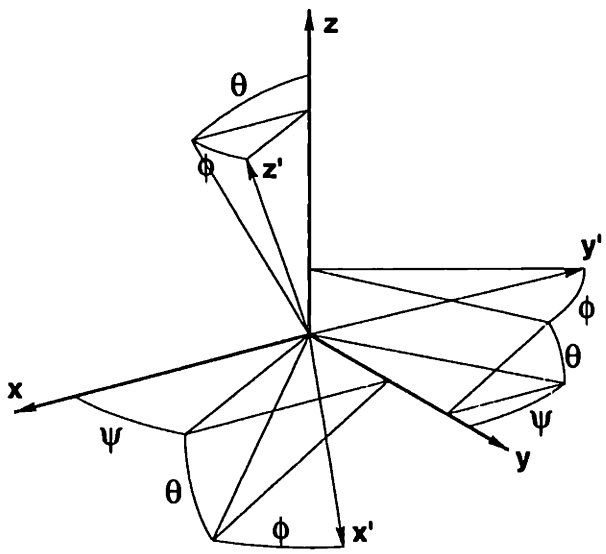


Figure A-3 Euler angles describe an arbitrary orientation with sequential rotations ψ, θ, ϕ about z, y, z.

$$\phi = \arctan\left(\frac{c_y}{c_x}\right) \qquad \mathbf{R} = \begin{bmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{bmatrix} \tag{A.15}$$

$$\theta = \arctan\left(\frac{c'_x}{c'_z}\right) \qquad \mathbf{R}' = \mathbf{R}_z(\phi)^{\mathrm{T}} \cdot \mathbf{R} \tag{A.16}$$

$$\psi = \arctan\left(\frac{a''_y}{a''_x}\right) \qquad \mathbf{R}'' = \mathbf{R}_y(\theta)^{\mathrm{T}} \cdot \mathbf{R}' \tag{A.17}$$

Recovering the angle vector for a fixed-axis rotation (or equivalently simultaneous rotations about base axes x, y, z) is more challenging since the axis could be in any direction. From geometry, the axis must be perpendicular to three vectors: ($\mathbf{x}'$ - $\mathbf{x}$), ($\mathbf{y}'$ - $\mathbf{y}$) and ($\mathbf{z}'$ - $\mathbf{z}$), however one is redundant. Equation A.18 expresses this requirement that the axis, represented by the vector direction of θ, must be perpendicular to these three vectors. As observed by [Whitney, 1972], this vector must be the eigenvector of $\mathbf{R}$ corresponding to a unit eigenvalue since it does not change direction or magnitude when transformed by $\mathbf{R}$. The eigenvector algorithm is the most robust way to find the axis of rotation. The magnitude is found as before by transforming back to the temporary CS as indicated by $\mathbf{R}'$ in Equation A.19, where $\mathbf{T}$ is defined in Equation A.8 and $\mathbf{v}$ is the eigenvector.

$$\begin{bmatrix} a_x - 1 & b_x & c_x \\ a_y & b_y - 1 & c_y \\ a_z & b_z & c_z - 1 \end{bmatrix}^{\mathrm{T}} \cdot \begin{bmatrix} \theta_x \\ \theta_y \\ \theta_z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{A.18}$$

$$\|\theta\| = \arctan\left(\frac{a'_y}{a'_x}\right) \qquad \mathbf{R}' = \mathbf{T}\left(\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix}\right)^{\mathrm{T}} \cdot \mathbf{R} \cdot \mathbf{T}\left(\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix}\right) \tag{A.19}$$

An alternative method for recovering the angle vector involves only algebraic manipulation of Equation A.10. Sine and cosine terms are extracted as shown in Equation A.20 and used in the two-argument arc tangent function to determine the angle of rotation. If the angle is zero, then the angle vector is also zero. If the angle is $\pm\,\pi$, then $\mathbf{R}$ has no skew symmetric part and the angle vector is simple to recover from Equation A.10. Otherwise, the angle vector is found from the skew symmetric part ($\mathbf{R}$ - $\mathbf{R}^{\mathrm{T}}$)/2 as shown in Equation A.21. This method exhibits minor round-off error for angles very near $\pm\,\pi$.

$$\cos(\|\theta\|) = \frac{\mathrm{trace}(\mathbf{R}) - 1}{2} \qquad \sin(\|\theta\|) = \frac{1}{2}\left\| \begin{Bmatrix} b_z - c_y \\ c_x - a_z \\ a_y - b_x \end{Bmatrix} \right\| \tag{A.20}$$

$$\begin{bmatrix} \theta_x \\ \theta_y \\ \theta_z \end{bmatrix} = \frac{1}{2} \begin{Bmatrix} b_z - c_y \\ c_x - a_z \\ a_y - b_x \end{Bmatrix} \frac{\|\theta\|}{\sin(\|\theta\|)} \tag{A.21}$$

A.3 The Homogeneous Transformation Matrix (HTM)

The homogeneous transformation matrix (HTM) is an augmentation to the rotation matrix to include a translation between CS's but is applicable only to position vectors, changing the representation from one CS to another. In the field of computer graphics, the HTM is doubly useful because it can represent a change in scale or a perspective view, for example. Those features are not presented here since they have no value in Mechanics.

Equation A.22 represents what the HTM must do, namely to change the representation of the position vector $\mathbf{p}$ from CS_1 to CS_0. The position vector $\mathbf{r}_0$ expressed in CS_0 locates the origin of CS_1. The rotation matrix describes the angular orientation of CS_1 with respect to CS_0. Equation A.23 expresses the same relationship using HTM's. The first matrix describes translation and the second matrix describes rotation. The fourth column as it multiplies through performs the additive operation in the previous equation. This feature makes the HTM representation more convenient to use for sequential coordinate transformations.

$$\mathbf{p}_0 = \mathbf{r}_0 + \mathbf{R}_{0/1} \cdot \mathbf{p}_1 \tag{A.22}$$

$$\begin{bmatrix} \mathbf{p}_0 \\ \hline 1 \end{bmatrix} = \left[\begin{array}{ccc:c} 1 & 0 & 0 & \\ 0 & 1 & 0 & \mathbf{r}_0 \\ 0 & 0 & 1 & \\ \hdashline 0 & 0 & 0 & 1 \end{array}\right] \cdot \left[\begin{array}{ccc:c} & \mathbf{R}_{0/1} & & 0 \\ & & & 0 \\ & & & 0 \\ \hdashline 0 & 0 & 0 & 1 \end{array}\right] \cdot \begin{bmatrix} \mathbf{p}_1 \\ \hline 1 \end{bmatrix} \tag{A.23}$$

Equation A.24 shows the result of multiplying the two HTM's in (A.23) to obtain a single HTM for both rotation and translation. Notice the first form of the HTM is expressed in CS_0 and the second is expressed in CS_1. Equation A.25 gives the inverse transformation obtained simply by reversing the subscripts. The inverse is easy to prove by multiplying to get the identity matrix. In practice, we usually use the first form of Equation A.24 and numerically invert to get the inverse transformation.

$$\mathbf{T}_{0/1} = \left[\begin{array}{c:c} \mathbf{R}_{0/1} & \mathbf{r}_0 \\ \hdashline \mathbf{0} & 1 \end{array}\right] = \left[\begin{array}{c:c} \mathbf{R}_{1/0}{}^{\mathrm{T}} & -\mathbf{R}_{1/0}{}^{\mathrm{T}} \cdot \mathbf{r}_1 \\ \hdashline \mathbf{0} & 1 \end{array}\right] \tag{A.24}$$

$$\mathbf{T}_{1/0} = \left[\begin{array}{c:c} \mathbf{R}_{1/0} & \mathbf{r}_1 \\ \hdashline \mathbf{0} & 1 \end{array}\right] = \left[\begin{array}{c:c} \mathbf{R}_{0/1}{}^{\mathrm{T}} & -\mathbf{R}_{0/1}{}^{\mathrm{T}} \cdot \mathbf{r}_0 \\ \hdashline \mathbf{0} & 1 \end{array}\right] \tag{A.25}$$

A.4 The Cross Product Matrix

The cross product is a three-dimensional vector operation that, in the context of Mechanics, relates through a lever arm either a force to a moment or an infinitesimal rotation to an infinitesimal translation. A point to notice is that each process must go in the direction

stated; that is, the inverse process is ambiguous. For example, there are infinitely many forces that can act on the same lever arm to give the same moment. However, only the component of force orthogonal to the lever arm is active in producing the moment. In addition, the order that the vectors appear in the cross product is important. Equation A.26 gives the order for forces and moments and Equation A.27 gives the order for infinitesimal rotations and translations.[1]

$$\begin{bmatrix} m_x \\ m_y \\ m_z \end{bmatrix} = \begin{bmatrix} r_x \\ r_y \\ r_z \end{bmatrix} \times \begin{bmatrix} f_x \\ f_y \\ f_z \end{bmatrix} = \begin{bmatrix} 0 & -r_z & r_y \\ r_z & 0 & -r_x \\ -r_y & r_x & 0 \end{bmatrix} \cdot \begin{bmatrix} f_x \\ f_y \\ f_z \end{bmatrix} \tag{A.26}$$

$$\mathbf{m} = \mathbf{r} \times \mathbf{f} = \mathbf{C}(\mathbf{r}) \cdot \mathbf{f}$$

$$\begin{bmatrix} d\delta_x \\ d\delta_y \\ d\delta_z \end{bmatrix} = \begin{bmatrix} d\theta_x \\ d\theta_y \\ d\theta_z \end{bmatrix} \times \begin{bmatrix} r_x \\ r_y \\ r_z \end{bmatrix} = \begin{bmatrix} 0 & r_z & -r_y \\ -r_z & 0 & r_x \\ r_y & -r_x & 0 \end{bmatrix} \cdot \begin{bmatrix} d\theta_x \\ d\theta_y \\ d\theta_z \end{bmatrix} \tag{A.27}$$

$$d\delta = d\theta \times \mathbf{r} = \left(d\theta^{\mathrm{T}} \cdot \mathbf{C}(\mathbf{r})\right)^{\mathrm{T}} = \mathbf{C}(\mathbf{r})^{\mathrm{T}} \cdot d\theta$$

The definition of the cross product matrix **C** could come from either equation since no technical basis exists for this choice. The convention, chosen strictly as a memory aid, uses Equation A.26 so that the cross product and its matrix representation occur in the same order. The transpose operator reverses the order in Equation A.27, which amounts to a sign change since the cross product matrix is skew symmetric. The non-invertible cross product matrix properly expresses the unidirectionality of the three-dimensional transformation involving only the lever arm. When extended to six dimensions in a later section, the transformation becomes invertible.

A.5 Equations of Compatibility and Equilibrium

In a structures problem, compatibility and equilibrium equations express the same geometric relationship; compatibility as a consistent relationship among displacements and equilibrium as a balance of forces. The two are linked by a constitutive law such as Hooke's law. Taken together, they express the structures problem as an input-output relationship. If the structure is linear, then the relationship is a stiffness matrix for a displacement input or a compliance matrix for a force input. The form of the equilibrium and compatibility equations depends on the context of the problem. Take the finite element problem for example; the compatibility equations relate nodal displacements to distortions of the finite elements (strain for solid elements). The equilibrium equations relate force distributions in the finite elements back to nodal forces. The problem and the equations

[1] Many applications involve displacements that are small enough for a linear approximation.

simplify if the structure can be lumped into a few spring and/or inertia elements. This level of complexity is reasonable to attempt without a finite element program. An example is the compliance matrix for a kinematic coupling due solely to compliance at contact points. In this case, the compatibility and equilibrium equations merely require the coordinate transformations between the coupling's CS and the CS at each contact point.

Figure A-4 shows one spring in what could be a parallel and/or series combination with several other springs. One end of the spring is grounded and the other is connected to the base CS through a rigid link. Movement of the base CS will cause a related deflection of the spring as described by Equation A.28, the compatibility equations. It is not obvious in Equation A.29 but the equilibrium equations express the same geometric relationship in terms of forces and moments developed in the spring and transferred to the base CS through the link. This equivalence will become obvious in the following section once the equations are expressed in terms of six-dimensional vectors. This leads to the [6 x 6] coordinate transformation matrix, which is useful for representing a collection of springs in the same CS so that parallel or series combinations can be added together.

$$\begin{aligned} d\delta_1 &= \mathbf{R}_{1/0} \cdot (d\delta_0 + d\theta_0 \times \mathbf{r}_0) \\ d\theta_1 &= \mathbf{R}_{1/0} \cdot d\theta_0 \end{aligned} \tag{A.28}$$

$$\begin{aligned} \mathbf{f}_0 &= \mathbf{R}_{0/1} \cdot \mathbf{f}_1 \\ \mathbf{m}_0 &= \mathbf{R}_{0/1} \cdot \mathbf{m}_1 + \mathbf{r}_0 \times (\mathbf{R}_{0/1} \cdot \mathbf{f}_1) \end{aligned} \tag{A.29}$$

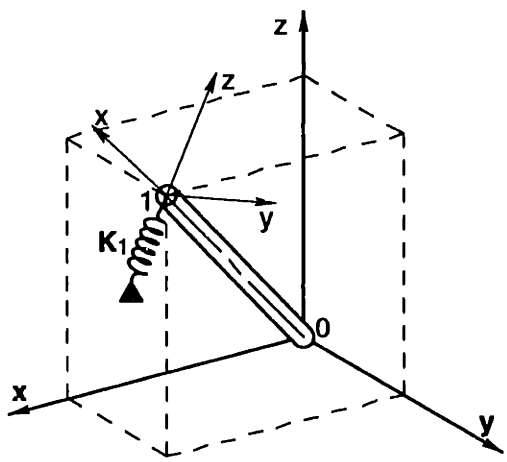


Figure A-4 The stiffness matrix $\mathbf{K}_1$ of the spring is expressed in terms of the local CS_1. The goal is to express the stiffness matrix in terms of the base CS_0. A collection of springs once represented in the same CS can be added together in series or parallel combinations.

A.6 The [6 x 6] Transformation Matrix

The [6 x 6] transformation matrix is an augmentation of the rotation matrix and the cross product matrix to change the representation of a six-dimensional vector from one CS to another. A six-dimensional vector describes three linear degrees of freedom and three angular degrees of freedom. Expressed in this form, Equations A.30 and A.31 show the previously discussed compatibility and equilibrium relationships as matrix equations. The definition of the [6 x 6] transformation matrix could come from either equation; however, the choice represented in Equation A.32 is logical since the transpose operator is consistent among the submatrices and the full matrix. This choice is also consistent with the HTM, where the first matrix in Equation A.31 describes translation and the second matrix describes rotation. Different only by the transpose operator, it is now apparent that compatibility and equilibrium equations clearly express the same geometric relationship.

$$\begin{bmatrix} d\delta_1 \\ --- \\ d\theta_1 \end{bmatrix} = \left[\begin{array}{ccc:ccc} & \mathbf{R}_{0/1}{}^T & & 0 & 0 & 0 \\ & & & 0 & 0 & 0 \\ & & & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & & & \\ 0 & 0 & 0 & & \mathbf{R}_{0/1}{}^T & \\ 0 & 0 & 0 & & & \end{array}\right] \cdot \left[\begin{array}{ccc:ccc} 1 & 0 & 0 & & \mathbf{C}(\mathbf{r}_0)^T & \\ 0 & 1 & 0 & & & \\ 0 & 0 & 1 & & & \\ \hdashline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \cdot \begin{bmatrix} d\delta_0 \\ --- \\ d\theta_0 \end{bmatrix} \tag{A.30}$$

$$\begin{bmatrix} \mathbf{f}_0 \\ --- \\ \mathbf{m}_0 \end{bmatrix} = \left[\begin{array}{ccc:ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \hdashline & & & 1 & 0 & 0 \\ & \mathbf{C}(\mathbf{r}_0) & & 0 & 1 & 0 \\ & & & 0 & 0 & 1 \end{array}\right] \cdot \left[\begin{array}{ccc:ccc} & & & 0 & 0 & 0 \\ & \mathbf{R}_{0/1} & & 0 & 0 & 0 \\ & & & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & & & \\ 0 & 0 & 0 & & \mathbf{R}_{0/1} & \\ 0 & 0 & 0 & & & \end{array}\right] \cdot \begin{bmatrix} \mathbf{f}_1 \\ --- \\ \mathbf{m}_1 \end{bmatrix} \tag{A.31}$$

$$\begin{bmatrix} \mathbf{f}_0 \\ --- \\ \mathbf{m}_0 \end{bmatrix} = \mathbf{T}_{0/1} \cdot \begin{bmatrix} \mathbf{f}_1 \\ --- \\ \mathbf{m}_1 \end{bmatrix} \qquad \begin{bmatrix} d\delta_1 \\ --- \\ d\theta_1 \end{bmatrix} = \mathbf{T}_{0/1}{}^T \cdot \begin{bmatrix} d\delta_0 \\ --- \\ d\theta_0 \end{bmatrix} \tag{A.32}$$

Equation A.33 shows the result of multiplying the two matrices to obtain $\mathbf{T}_{0/1}$. Notice the first form of $\mathbf{T}_{0/1}$ is expressed in CS_0 and the second is expressed in CS_1. Equation A.34 gives the inverse transformation obtained simply by reversing the subscripts. The inverse is easy to prove by multiplying to get the identity matrix. Notice that each transformation matrix is invertible simply by transposing the submatrices; however in practice, we usually numerically invert the first form of Equation A.33.

$$\mathbf{T}_{0/1} = \left[\begin{array}{c:c} \mathbf{R}_{0/1} & \mathbf{0} \\ \hdashline \mathbf{C}(\mathbf{r}_0) \cdot \mathbf{R}_{0/1} & \mathbf{R}_{0/1} \end{array}\right] = \left[\begin{array}{c:c} \mathbf{R}_{1/0}{}^T & \mathbf{0} \\ \hdashline \left(\mathbf{C}(\mathbf{r}_1) \cdot \mathbf{R}_{1/0}\right)^T & \mathbf{R}_{1/0}{}^T \end{array}\right] \tag{A.33}$$

$$\mathbf{T}_{1/0} = \left[\begin{array}{c|c} \mathbf{R}_{1/0} & \mathbf{0} \\ \hline \mathbf{C}(\mathbf{r}_1)\cdot\mathbf{R}_{1/0} & \mathbf{R}_{1/0} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{R}_{0/1}{}^{\mathrm{T}} & \mathbf{0} \\ \hline \left(\mathbf{C}(\mathbf{r}_0)\cdot\mathbf{R}_{0/1}\right)^{\mathrm{T}} & \mathbf{R}_{0/1}{}^{\mathrm{T}} \end{array}\right] \tag{A.34}$$

The [6 x 6] transformation matrix is useful for representing a collection of springs in the same CS so that parallel or series combinations can be added together. The block diagram in Figure A-5 shows the required flow of the process for parallel springs (a) and series springs (b). The subscript i indicates there are as many loops as there are springs. Equation A.35 results from summing the stiffness matrices of springs in parallel. Equation A.36 results from summing the compliance matrices of springs in series. Mixed combinations of parallel and series springs require like groups to be combined first then inverted as necessary to complete the combination. This technique also works for inertia matrices, where the first three diagonal elements contain the mass, the lower diagonal block matrix contains the inertia tensor and zero are elsewhere.

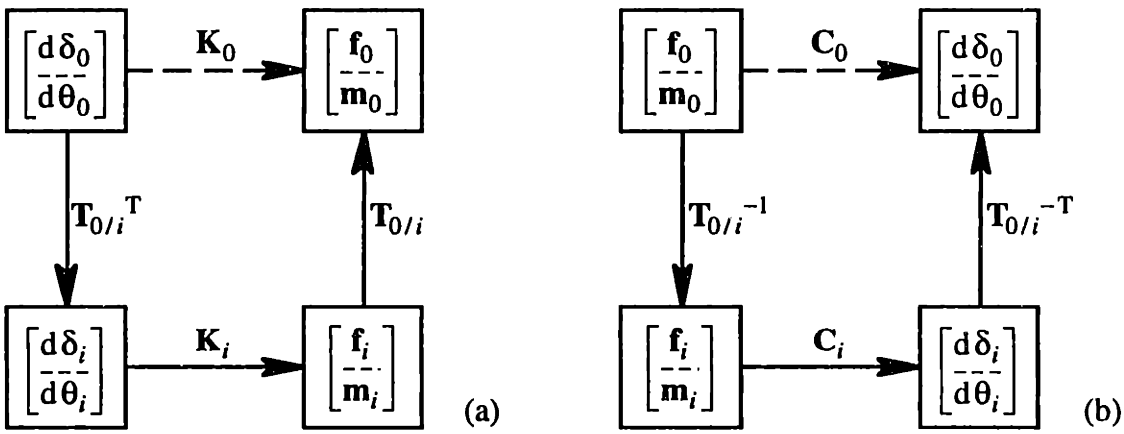


Figure A-5 The diagram shows the required process flow for parallel springs (a) and series springs (b), after [Strang, 1986].

$$\mathbf{K}_0 = \sum_i \mathbf{T}_{0/i}\cdot\mathbf{K}_i\cdot\mathbf{T}_{0/i}{}^{\mathrm{T}} \tag{A.35}$$

$$\mathbf{C}_0 = \sum_i \mathbf{T}_{0/i}{}^{-\mathrm{T}}\cdot\mathbf{C}_i\cdot\mathbf{T}_{0/i}{}^{-1} = \sum_i \mathbf{T}_{i/0}{}^{\mathrm{T}}\cdot\mathbf{C}_i\cdot\mathbf{T}_{i/0} \tag{A.36}$$

A.7 Dynamic Simulations Involving Large-Angle Motion

Simulating a dynamic system by numerically integrating state equations may require tracking potentially large angular motion in three dimensions. For example, the die-throwing simulation motivated this development. The rotation matrix is invaluable for transforming vectors between the moving body CS_1 and the fixed base CS_0. The difficulty lies in integrating angular velocity to get a meaningful time history of the angular orientation as expressed by the rotation matrix. For the typical vector quantity, numerical integration is simply vector addition of a sequence of small steps beginning at some initial state. Of course a sequence of small rotations requires matrix multiplication, but standard ODE

(ordinary differential equation) solvers such as ode45 in Matlab™ support only addition. This section presents three approaches to this problem. The first two methods work with any standard ODE solver but have limitations that warrant a better approach, one that requires a more flexible solver. This last method works with a specially developed Matlab™ function called rk4, a fourth-order, Runge-Kutta integration algorithm.

A typical state vector for a single rigid body has 12 terms made up of four subvectors. Linear velocity, angular velocity and position are conventional three-dimensional vectors represented in the base CS_0. The fourth subvector contains information describing the orientation of the rigid body. The focus of this presentation is the form and processing of this information, particularly how the rotation matrix is kept current. The other state equations may have complexities associated with a particular dynamic system but otherwise are straight forward to integrate. To begin, Equation A.37 shows the state equation for angular orientation used in the die-throwing simulation. Using sequential rotations about the base x, y, z axes, the time rate of change of θ_0 is linearly related to the angular velocity vector through $\mathbf{J}(\theta_0)$, the Jacobian matrix.[I] Although the Jacobian matrix is nonlinear and requires inversion, this computation occurs in the user function called by the solver. This Jacobian approach works with a standard solver but requires safeguards to avoid singularities in the matrix inversion (a significant problem discussed a little later).

$$\frac{\mathrm{d}}{\mathrm{d}t}(\theta_0) = \mathbf{J}(\theta_0)^{-1} \cdot \omega_0 \tag{A.37}$$

Deriving the Jacobian matrix requires a physical relationship between the angle parameters of the rotation matrix and the angular velocities. Equation A.38 shows two ways to represent the velocity at any point on the rigid body due solely to angular motion.[II] On the left, the time derivative of the position vector to any arbitrary point on the body is equal to the cross product between angular velocity and the same position vector. Notice that $\mathbf{r}_1$ is arbitrary and unchanging in the body CS_1, and the rotation matrix transforms it to the base CS_0. Being arbitrary, $\mathbf{r}_1$ can be removed to obtain Equation A.39. Alternatively, Equation A.40 shows two ways to express the angular orientation after an infinitesimal time step. On the right, the small-angle rotation matrix pre multiplies the nominal rotation matrix because the angular velocity is expressed in the base CS_0. The definition of the derivative allows either Equation A.39 or Equation A.40 to be expressed as the other.

[I] In the context of robot manipulators, the Jacobian matrix transforms the velocities at joints to the velocity at the end effector. It provides the continuity relation for the manipulator. Thinking of the rotation matrix as a mechanism, the Jacobian matrix provides the transmission matrix.

[II] CS_1 is attached to the moving body typically with the origin at the mass centroid. The rotation matrix $\mathbf{R}_{0/1}$ represents the rotation of CS_1 with respect to CS_0 as if they have common origins. CS_0 could be the stationary base CS or an intermediate CS translating with CS_1; it would not change the equations.

$$\frac{d}{dt}\{\mathbf{R}_{0/1}(\theta_0)\cdot\mathbf{r}_1\}=\omega_0\times\{\mathbf{R}_{0/1}(\theta_0)\cdot\mathbf{r}_1\} \tag{A.38}$$

$$\frac{d}{dt}[\mathbf{R}_{0/1}(\theta_0)]=\begin{bmatrix}0 & -\omega_z & \omega_y\\ \omega_z & 0 & -\omega_x\\ -\omega_y & \omega_x & 0\end{bmatrix}\cdot\mathbf{R}_{0/1}(\theta_0) \tag{A.39}$$

$$\mathbf{R}_{0/1}(\theta_0+d\theta_0)=\begin{bmatrix}1 & -\omega_z dt & \omega_y dt\\ \omega_z dt & 1 & -\omega_x dt\\ -\omega_y dt & \omega_x dt & 1\end{bmatrix}\cdot\mathbf{R}_{0/1}(\theta_0) \tag{A.40}$$

Having arrived at Equation A.39, there are several reasonable avenues to pursue. Perhaps the simplest one uses the equation directly without forming the Jacobian matrix. This is the other method that works with a standard solver. Putting the Jacobian approach aside for the moment, this other method uses the nine elements of the rotation matrix as the state subvector for orientation. This only requires the Matlab™ command *reshape* to change between a vector and a matrix within the user function. However, there are two problems with this method: the elements are oscillatory for a spinning body, and the rotation matrix looses its orthogonality over a number of steps. Approximate orthogonality is maintained by conditioning the rotation matrix inside the state equation as Equation A.41 shows.[1] Alternatively, the rotation matrix may be converted to angle parameters then back to a perfectly orthogonal rotation matrix, but it may not be any more accurate.

$$\frac{d}{dt}[\mathbf{R}_{0/1}]=\begin{bmatrix}0 & -\omega_z & \omega_y\\ \omega_z & 0 & -\omega_x\\ -\omega_y & \omega_x & 0\end{bmatrix}\cdot\frac{\mathbf{R}_{0/1}+\mathbf{R}_{0/1}^{-T}}{2} \tag{A.41}$$

The Jacobian approach has no orthogonality problem since the rotation matrix is computed from integrated angle parameters, but the possibility of a singularity is far more serious. The condition where it becomes singular depends on the type of rotation matrix used in the derivation. For example, the rotation matrix described by sequential rotations about fixed x, y and z axes makes the derivation relatively simple, but rotations about x and z are equivalent if the rotation about y is ±90°. Figure A-6 demonstrates this condition of gimbal lock, which effectively sacrifices one degree of freedom. This problem was avoided in the die-throwing simulation by initially aligning the z axis of an intermediate CS to the spin axis of the die after each bounce. This tends to keep the x and y angles small and less

[1] The exponent -T indicates that the matrix is inverted and transposed. The result on an orthogonal matrix would be no change.

oscillatory, but does not necessarily prevent singularities. This safeguard also introduces the complication of the intermediate CS.

Equation A.42 shows how to compute the Jacobian matrix for sequential rotations about fixed x, y and z axes. The matrices **A**, **B** and **C** come from applying the chain rule to Equation A.39. These matrices are skew symmetric so the appropriate elements can be matched with ω_x, ω_y and ω_z to form **J**. The same technique would apply to Euler angles, but then the singularity would occur when the second rotation is zero or a full revolution.

$$\mathbf{J} = \begin{bmatrix} \mathbf{A}_{3,2} & \mathbf{B}_{3,2} & \mathbf{C}_{3,2} \\ \mathbf{A}_{1,3} & \mathbf{B}_{1,3} & \mathbf{C}_{1,3} \\ \mathbf{A}_{2,1} & \mathbf{B}_{2,1} & \mathbf{C}_{2,1} \end{bmatrix} \qquad \mathbf{B} = \mathbf{R}_z \cdot \frac{d\mathbf{R}_y}{d\theta_y} \cdot \mathbf{R}_y{}^{\mathrm{T}} \cdot \mathbf{R}_z{}^{\mathrm{T}} \tag{A.42}$$

$$\mathbf{A} = \mathbf{R}_z \cdot \mathbf{R}_y \cdot \frac{d\mathbf{R}_x}{d\theta_x} \cdot \mathbf{R}_x{}^{\mathrm{T}} \cdot \mathbf{R}_y{}^{\mathrm{T}} \cdot \mathbf{R}_z{}^{\mathrm{T}} \qquad \mathbf{C} = \frac{d\mathbf{R}_z}{d\theta_z} \cdot \mathbf{R}_z{}^{\mathrm{T}}$$

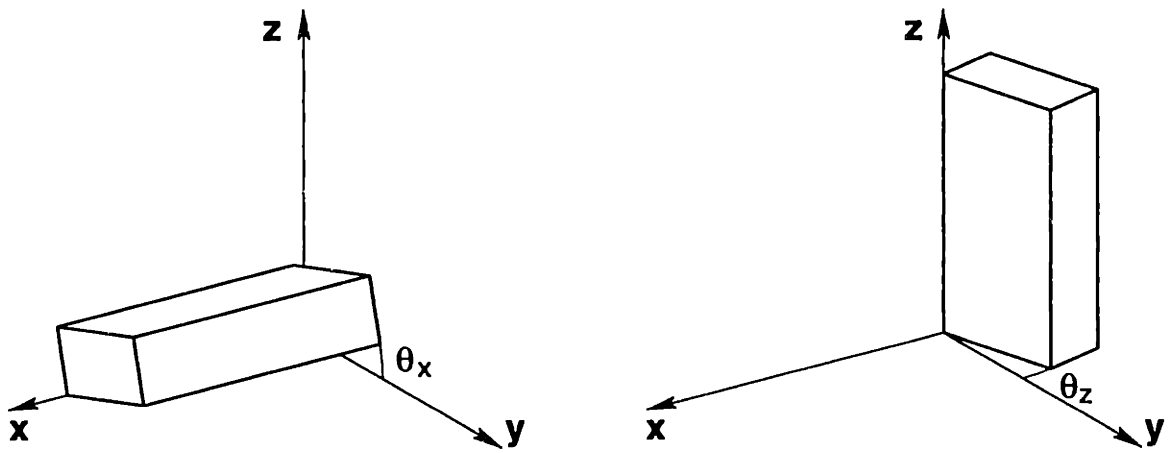


Figure A-6 After a ±90° rotation about y, a rotation about z is equivalent to a previous rotation about x. It is not possible to rotate differentially about x from this position without large changes to θ_x and θ_z.

It seems intuitive that simultaneous rotations would avoid the problems of singularity and oscillation. Unfortunately, the Jacobian matrix based on simultaneous rotations becomes singular at full revolutions.[1] Why not do the obvious and simply integrate angular velocity to get a trajectory or sequence of small simultaneous angle vectors? The problem is that a standard ODE solver passes only the endpoint of the trajectory to the user function, which means nothing unless the trajectory happens to be a straight line. The trajectory must be processed as a product of rotation matrices rather than a summation of angle steps. This requires a parallel computation to update the rotation matrix at each step of the integration. The integration algorithm rk4 was specially developed for this purpose. In addition to the state variables, rk4 can track other variables and pass them

[1] At full revolutions, the rotation matrix is equal to the identity matrix and the rotation axis is arbitrary. Thus a differential rotation through this point may require a large jump in the rotation axis.

to and from the user function. This enables the user function to compute the current rotation matrix given the rotation matrix and angle vector at the previous step of the integration, as shown in the first line of Equation A.43. The last line resets the variables for the next step.

$$\begin{aligned} \mathbf{R}_k &= \mathbf{R}_{\text{axis}}(\theta_k - \theta_{k-1}) \cdot \mathbf{R}_{k-1} \\ \frac{d\theta_k}{dt} &= \omega_k \\ \theta_{k-1} &= \theta_k \qquad \mathbf{R}_{k-1} = \mathbf{R}_k \end{aligned} \tag{A.43}$$

A.8 Matlab™ Functions for Transformation Matrices

This section contains Matlab™ functions for computing transformation matrices and other matrix techniques presented throughout this chapter. The reader is granted permission to use these routines for his/her personal research. These routines may not be used commercially without permission from the author.

```
function R = Rxyz(theta)
%Rotation Matrix about fixed x, y, z axes
%
%  R = Rxyz(theta)
%
%  Generates a 3 by 3 rotation matrix for sequential rotations about the
%  base x, y, z axes in that order. This is equivalent to roll, pitch, yaw
%  about the body z, y, x axes, respectively. A rotation matrix transforms
%  a three-component vector in the body coordinate system back to the base
%  coordinate system. The transpose gives the inverse transformation.
%
%  Input:   'theta' is a three element vector containing the sequential
%        rotation angles about x, y, z axes, respectively.
%        Angles are expressed in radians.
%  Output:  'R' is the 3 by 3 rotation matrix

%  Layton C. Hale 8/1/98
%  Copyright (c) 1998 by Layton C. Hale

m_n = size(theta);
if (max(m_n)~= 3) | (min(size(m_n))~= 1)
   error('vector theta must have 3 elements')
end
s = sin(theta);
c = cos(theta);
R1 = [1 0 0; 0 c(1) -s(1); 0 s(1) c(1)];      % rotate about x
R2 = [c(2) 0 s(2); 0 1 0; -s(2) 0 c(2)];      % rotate about y
R3 = [c(3) -s(3) 0; s(3) c(3) 0; 0 0 1];      % rotate about z
R = R3*R2*R1;

function R = R_Euler(theta)
%Rotation Matrix using Euler angles
%
%  R = R_Euler(theta)
```

```
%
%  Generates a 3 by 3 rotation matrix for sequential rotations about the
%  body z, y', z" axes in that order. This is the sequence now known as
%  Euler angles. This is equivalent to the opposite order about the base
%  z, y, z axes. A rotation matrix transforms a three-component vector
%  in the body coordinate system back to the base coordinate system.
%  The transpose gives the inverse transformation.
%
%  Input:   'theta' is a three element vector containing the sequential
%        rotation angles about z, y', z" axes, respectively.
%        Angles are expressed in radians.
%  Output:  'R' is the 3 by 3 rotation matrix

%  Layton C. Hale 8/1/98
%  Copyright (c) 1998 by Layton C. Hale

m_n = size(theta);
if (max(m_n)~= 3) | (min(size(m_n))~= 1)
   error('vector theta must have 3 elements')
end
s = sin(theta);
c = cos(theta);
R1 = [c(1) -s(1) 0; s(1) c(1) 0; 0 0 1];      % rotate about z
R2 = [c(2) 0 s(2); 0 1 0; -s(2) 0 c(2)];      % rotate about y'
R3 = [c(3) -s(3) 0; s(3) c(3) 0; 0 0 1];      % rotate about z"
R = R1*R2*R3;

function R = Raxis(axis,theta)
%Rotation Matrix about an angle vector
%
%  R = Raxis(axis,theta)
%
%  Generates a 3 by 3 rotation matrix for a rotation about a fixed axis.
%  This is equivalent to simultaneous rotations about base x, y, z axes.
%  A rotation matrix transforms a three-component vector in the body
%  coordinate system back to the base coordinate system. The transpose
%  gives the inverse transformation.
%
%  Input:   'axis' is a three element vector giving the direction of the axis.
%        'theta' is the rotation angle in radians (default theta = |axis|).
%  Output:  'R' is the 3 by 3 rotation matrix

%  Layton C. Hale 8/23/98
%  Copyright (c) 1998 by Layton C. Hale

m_n = size(axis);
if (max(m_n)~= 3) | (min(size(m_n))~= 1)
   error('vector theta must have 3 elements')
end
mag = norm(axis);
if nargin < 2
   theta = mag;
end
if mag > eps
```

```
   axis = axis/mag;            % axis must be unit length
   c = cos(theta);
   s = sin(theta);
   R = eye(3)*c + (1 - c)*axis(:)*axis(:)'...
      + s*[0 -axis(3) axis(2); axis(3) 0 -axis(1); -axis(2) axis(1) 0];
else
   R = eye(3);
end

function theta = rotation_angles(R,i)
%Rotation Angles for a rotation matrix
%
%  theta = rotation_angles(R,i)
%
%  Recovers the angles of rotation for a particular rotation matrix.
%
%  Input:   'R' is the 3 by 3 rotation matrix used in recovering the angles.
%        'i' indicates the type of rotation. Use: 1 for rotation about base
%        x, y, z axes; 2 for Euler angles (body z, y', z" axes);
%        3 for rotation about the angle vector.
%  Output:  'theta' is the angle vector corresponding to the type of rotation.

%  Layton C. Hale 8/1/98
%  Copyright (c) 1998 by Layton C. Hale

[m,n] = size(R);
if (m ~= 3) | (n ~= 3)
   error('matrix R must be 3 by 3')
end
if i == 1                              % rotation about base x, y, z axes
   theta(3) = atan2(R(2,1),R(1,1));
   s = sin(theta(3));
   c = cos(theta(3));
   Rz = [c -s 0; s c 0; 0 0 1];
   R1 = Rz'*R;
   theta(2) = atan2(-R1(3,1),R1(1,1));
   s = sin(theta(2));
   c = cos(theta(2));
   Ry = [c 0 s; 0 1 0; -s 0 c];
   R2 = Ry'*R1;
   theta(1) = atan2(R2(3,2),R2(2,2));
elseif i == 2                          % rotation about body z, y', z" axes
   theta(1) = atan2(R(2,3),R(1,3));     % equivalent to base z, y, z axes in
   s = sin(theta(1));                   % opposite order of rotations
   c = cos(theta(1));
   Rz = [c -s 0; s c 0; 0 0 1];
   R1 = Rz'*R;
   theta(2) = atan2(R1(1,3),R1(3,3));
   s = sin(theta(2));
   c = cos(theta(2));
   Ry = [c 0 s; 0 1 0; -s 0 c];
   R2 = Ry'*R1;
   theta(3) = atan2(R2(2,1),R2(1,1));
elseif i == 3                          % rotation about the angle vector
```

```
    v = [R(3,2)-R(2,3), R(1,3)-R(3,1), R(2,1)-R(1,2)]/2;
    s = norm(v);
    c = (trace(R) - 1)/2;
    ang = atan2(s,c);
    if abs(ang) < eps                    % no rotation
       theta = zeros(1,3);
    elseif abs(ang) < pi-eps                % calculate from skew sym. part
       theta = v*ang/s;
    else                                 %'calculate from symmetric part
       V = ((R + R')/2 + eye(3))/2;
       [max_diag, j] = max(diag(V));
       v = V(j,:)/sqrt(V(j,j));
       theta = ang*v;
    end
else
    error('i must be 1, 2 or 3')
end
```

```
function T = HTM(R,p,inverse)
%Homogeneous Transformation Matrix
%
%  T = HTM(R,p,inverse)
%
%  Generates a 4 by 4 coordinate transformation matrix used to transform a
%  position vector in the body coordinate system (CS) to the base CS.
%  The position vector must be augmented with a fourth row set equal to 1.
%
%  Input:   'R' is a 3 by 3 rotation matrix that describes the orientation
%         of the body CS with respect to the base CS.
%         'p' is a 3 element position vector that describes the position
%         of the origin of the body CS with respect to the base CS.
%         'inverse' is an optional argument that when present causes the
%         calculation of the inverse transformation matrix.
%  Output:  'T' is the 4 by 4 homogeneous transformation matrix.

%  Layton C. Hale 8/2/98
%  Copyright (c) 1998 by Layton C. Hale

[m,n] = size(R);
if (m ~= 3) | (n ~= 3)
   error('matrix R must be 3 by 3')
end
if length(p) ~= 3
   error('vector p must have 3 elements')
end
if nargin < 3
   T = [R, p(:); 0 0 0 1];
else
   T = [R', -R'*p(:); 0 0 0 1];
end
```

```
function C = CPM(p)
%Cross Product Matrix
%  C = CPM(p)
```

```
%
%  Generates a 3 by 3 skew symmetric matrix that performs the cross product
%  operation when multiplied by a column vector.
%  CPM(p)*q is the same as p x q. CPM(p)'*q is the same as q x p.
%
%  Input:    'p' is a 3 component vector.
%  Output:   'C' is the [3 x 3] cross product matrix.

%  Layton C. Hale 8/29/98
%  Copyright (c) 1998 by Layton C. Hale

C = [0 -p(3) p(2); p(3) 0 -p(1); -p(2) p(1) 0];
```

```
function T = T6x6(R,p,inverse)
%[6 x 6] Transformation Matrix
%
%  T = T6x6(R,p,inverse)
%
%  Generates a 6 by 6 coordinate transformation matrix used to transform a
%  six-component, force-moment vector in the body coordinate system (CS)
%  to the base CS. The transpose of the matrix transforms a differential,
%  displacement-rotation vector in the base CS to the body CS.
%
%  Input:    'R' is a 3 by 3 rotation matrix that describes the orientation
%        of the body CS with respect to the base CS.
%        'p' is a 3 element position vector that describes the position
%        of the origin of the body CS with respect to the base CS.
%        'inverse' is an optional argument that when present causes the
%        calculation of the inverse transformation matrix.
%  Output:   'T' is the [6 x 6] transformation matrix.

%  Layton C. Hale 8/9/98
%  Copyright (c) 1998 by Layton C. Hale

[m,n] = size(R);
if (m ~= 3) | (n ~= 3)
   error('matrix R must be 3 by 3')
end
if length(p) ~= 3
   error('vector p must have 3 elements')
end
C = [0 -p(3) p(2); p(3) 0 -p(1); -p(2) p(1) 0];
if nargin < 3
   T = [R, zeros(3,3); C*R, R];
else
   T = [R', zeros(3,3); (C*R)', R'];
end
```

```
function [T,varargout] = rk4(f,ts,tol,key,varargin)
%Runge-Kutta integration of state equations
%
%  [T,P1,P2,...,Pm] = rk4(f,ts,tol,key,p1,p2,...,pn)
%
%  Integrates a system of first order differential equations of the form
```

```
%   dx/dt = f(t,x) using the fourth-order Runge-Kutta algorithm. Use to
%  simulate state equations over time starting from an initial condition.
%
%  Similar to Matlab functions ode23 and ode45, this function is more
%  flexible in passing variables to and from the state equations and in
%  recording their time history through the integration.
%
%  Variables are passed back and forth to the state equations using cell
%  arrays whose cells can contain different data types. Functions recognize
%  varargin and varargout as special cell arrays used for passing a variable
%  number of arguments. The cells hold state variables and any other
%  variables the problem requires. By and large, the use of cell arrays is
%  transparent to the user unless modifications are made to this function.
%
%  INPUT:
%  f  A string containing the name of the function to integrate.
%     Ex:   f = 'state' calls [q1,q2,...,qn] = state(t,p1,p2,...,pn)
%        t  The present time in the integration (scalar).
%        p1 The first in a series of input arguments that are state
%           variables or auxiliary variables. For example, several
%           arguments might be state variables (position vector,
%           velocity vector, etc.) while others might keep track of
%           auxiliary information (rotation matrix).
%        q1 The first in a series of output arguments that must
%           correspond in number and kind to the input arguments.
%  ts Time sequence (a vector from start to finish). It controls the
%     maximum time steps. Set tol={} to force exact time steps.
%  tol   A cell array containing the tolerances allowed in the integration.
%     Arrange it to match the number and order of state variable arguments.
%     Uses only the nonzero elements for adapting the time step.
%  key   A vector with as many elements as arguments p1,...,pn. It tells rk4
%     how to process the different arguments. Choose from the following:
%     0 do nothing,
%     1 integrate state variables w/o recording their time history,
%     2 integrate state variables and record their time history,
%     3 update auxiliary variables for the next time step (p's = q's),
%     4 update auxiliary variables and record their time history.
%  p1 The initial condition for argument p1 in 'f', and so on.
%
%  OUTPUT:
%  T  A column vector of the actual time sequence used in the integration.
%  P1 The first in a series of output arguments that contain the time
%     history of a variable as directed by key. Each row corresponds to
%     one time increment of the variable.

%  Layton Hale, 8/31/98.
%  Copyright (c) 1998 by Layton C. Hale

% Error checking
if ~isstr(f)
   error('the function name must be a string')
end
nts = length(ts);
if nts < 2
```

```
   error('ts must contain at least the start and end times')
end
if any(ts(1:nts-1) >= ts(2:nts))
   error('time must increase through the integration')
end
narg = length(varargin);
if length(key) ~= narg
   error('key is inconsistent with the number of input arguments')
end

% Find indices of key that correspond to integrate, update and record
int = find((key == 1) | (key == 2));
upd = find((key == 3) | (key == 4));
rec = find((key == 2) | (key == 4));

tols = length(tol);
if tols & (tols ~= length(int))
   error('tol is inconsistent with the state variable arguments')
end
% Initialize
i = 1;            % integration increment
ii = 2;           % increment for next ts
t = ts(i);
tspan = ts(nts) - t;
if    tols                          % use adaptive time step
   dt = min((ts(nts) - t)/128, ts(ii) - t);
   for j=1:tols
      nztols{j} = find(tol{j});        % indices of non zero tolerances
   end
else
   dt = ts(ii) - t;
end
x1 = varargin;
x2 = x1;
x3 = x1;
x4 = x1;
args = 1:narg;
[s1{args}] = feval(f, t, x1{:});

% Allocate a chunk of space to record history
chunk = 128;
T = zeros(chunk,1);
T(1,1) = t;                         % record time
k = 1;
for j=rec
   X{k} = zeros(chunk,length(x1{j}));
   X{k}(1,:) = x1{j}(:)';              % record variables
   k = k+1;
end
% The main loop
while 1
   dt2 = dt/2;
   for j=int
      x2{j} = x1{j} + s1{j}*dt2;       % forward Euler integration
```

```
   end
   [s2{args}] = feval(f, t+dt2, x2{:});
   for j=int
      x3{j} = x1{j} + s2{j}*dt2;        % backward Euler integration
   end
   [s3{args}] = feval(f, t+dt2, x3{:});
   for j=int
      x4{j} = x1{j} + s3{j}*dt;           % trapezoidal integration
   end
   [s4{args}] = feval(f, t+dt, x4{:});
   if tols                               % calculate the relative error
      k = 1;
      for j=int
         error = abs(s1{j} - s2{j} - s3{j} + s4{j})/12;
         re(k) = tspan*max(error(nztols{k})./tol{k}(nztols{k}));
         k = k+1;
      end
      max_re = max(re);
   else
      max_re = 0;
   end
   if max_re <=1                         % OK to proceed with next step
      i = i+1;
      t = t + dt;
      for j=int                          % integrate variables
         x1{j} = x1{j} + (s1{j} + 2*(s2{j} + s3{j}) + s3{j})*dt/6;
      end
      [s1{args}] = feval(f, t, x1{:});    % evaluate f at new time step
      for j=upd                          % update variables
         x1{j} = s1{j};
      end
      flag = 0;
      if i > length(T)                   % allocate more space
         T = [T;zeros(chunk,1)];
         flag = 1;
      end
      T(i,1) = t;                    % record time
      k = 1;
      for j=rec
         if flag                         % allocate more space
            X{k} = [X{k};zeros(chunk,length(x1{j}))];
         end
         X{k}(i,:) = x1{j}(:)';          % record variables
         k = k+1;
      end
   end
   if tols                               % calculate the next time step
      if t >= ts(ii)
         ii = ii+1;
         if ii > nts, break, end         % exit loop
      end
      max_dt = ts(ii) - t;
      dt = min(0.9*dt/sqrt(max_re), max_dt);
   else
```

```
      ii = ii+1;
      if ii > nts, break, end          % exit loop
      dt = ts(ii) - t;
   end
end
T = T(1:i,1);                          % truncate record
for k = 1:length(rec)
   varargout{k} = X{k}(1:i,:);
end
```

B Least-Squares Fitting

The problem of fitting a mathematical model to a set of measurements is common among many disciplines. In precision engineering, the model may provide compensation for errors in a machine, separate feature errors from a part's ideal shape, or help diagnose systematic errors in a machine and/or process. The model could be static or dynamic and consist of nearly any mathematical function chosen by the user. The fitting process determines the values of model parameters that minimize a scalar measure of the residual errors between the fit and the measurements. For least-squares fitting, the measure is the sum of squared errors otherwise known as the L^2 norm of the error vector. Other measures of practical interest are the sum of absolute values (the L^1 norm) and the peak-to-valley error (the L^∞ norm). The least-squares technique is the most tractable mathematically and arguably provides the best fit to a set of measurements.[I] For example, if the model is a constant, then the least-squares solution is the average of the measurements. Every linear algebra book treats the topic of least-squares fitting, but seldom found are the particular applications that you will likely need.[II] The applications presented here should cover most needs, or new applications can be derived from the examples presented.

A direct solution to the least-squares fit is possible if the model is linear in the parameters, for example, the coefficients of a polynomial function. Nonlinear least squares is a simple extension of linear least squares but requires an iterative solution. The most familiar example is linear regression where the model is the equation of a line, $y = a\,x + b$. Arranged in matrix form, Equation B.1 shows m such equations for m measurements. It requires a vector of errors to account for the y-distance that the (x, y) data points differ from the straight line.[III] The matrix for a general model may contain several columns computed from nonlinear functions of the data. The symbol **A** is chosen to emphasize that the matrix is constant with respect to the parameters and that all terms are determined before solving the *linear* least-squares problem. The symbol **x** is chosen to emphasize that the parameter vector is the variable solved to minimize the sum-squared error in the equations. The way that the equations are written determines the direction that the errors act. Usually it is desirable to measure the errors perpendicular to the curve or surface being fitted.

[I] The least-squares solution gives the maximum likelihood estimation of the fitted parameters when the measurement errors are independent and normally distributed with equal standard deviations [Press, et al., 1992]. Unequal standard deviations lead to chi-square fitting, a particular weighted least squares to be discussed later. Likelihood refers to the probability that the data set lies within an arbitrary but constant error band about the model. Thus, the maximum likelihood estimation gives the best fit statistically.

[II] Several books found useful in developing this section are [Press, et al., 1992], [Nash, 1990] and [Strang, 1980, 1986]. Many other references may be found under linear algebra and numerical methods.

[III] Since a line has constant slope, the error vector may be multiplied by a constant after fitting to change the alignment of errors as desired, typically perpendicular to the line.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix} \qquad \text{(B.1)}$$

$$\mathbf{y} = \mathbf{A} \cdot \mathbf{x} + \mathbf{e}$$

The goal of the fitting process is to determine the values of parameters in the model that minimize the sum-squared error written succinctly as $\mathbf{e}^T\mathbf{e}$. Replacing $\mathbf{e}$ in this expression with $[\mathbf{y} - \mathbf{A}\,\mathbf{x}]$, Equation B.2 expands the result into a form that is readily differentiable. The minimum sum-squared error occurs when the derivatives with respect to the parameters are simultaneously zero as indicated by Equation B.3, which leads to the so called normal equations. However, a simpler derivation using geometry rather than calculus is more compelling. The best combination of columns of $\mathbf{A}$, namely $\mathbf{x}$, will result in an error vector $\mathbf{e}$ that has no component along any column of $\mathbf{A}$; otherwise, the error could be reduced further. This requires only an orthogonality condition $\mathbf{A}^T\,\mathbf{e} = \mathbf{0}$ and hence the name normal equations. Confined to the column space of $\mathbf{A}$, the solution to the normal equations eliminates the maximum possible error in the fit (without arguing for a least-squares measure of error).

$$\mathbf{e}^T\mathbf{e} = [\mathbf{y} - \mathbf{A}\,\mathbf{x}]^T[\mathbf{y} - \mathbf{A}\,\mathbf{x}] = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{A}\,\mathbf{x} + \mathbf{x}^T\mathbf{A}^T\mathbf{A}\,\mathbf{x} \qquad \text{(B.2)}$$

$$\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{e}^T\mathbf{e}\right) = \mathbf{0} \quad \rightarrow \quad \mathbf{A}^T\mathbf{A}\,\mathbf{x} = \mathbf{A}^T\mathbf{y} \qquad \text{(B.3)}$$

The algebraic solution to the normal equations is obvious in Equation B.4, but it requires $\mathbf{A}^T\,\mathbf{A}$ to be full rank to solve the inverse. In practice, the overdetermined equations $\mathbf{y} = \mathbf{A}\,\mathbf{x}$ may be poorly conditioned making $\mathbf{A}^T\,\mathbf{A}$ very nearly singular.[I] This occurs if the model is redundant or when the data set does not properly span its degrees of freedom. Numerical round-off error in a procedure like Gauss elimination operating on $\mathbf{A}^T\,\mathbf{A}$ may well lead to nonsensical results if a solution is found at all. Instead the references recommend using an orthogonalizing procedure on $\mathbf{A}$ as being more stable.[II] Even so it is advisable to fix the problem of the ill-conditioned matrix rather than worrying much about numerical errors. The diagnostic information included in a solution using singular value decomposition makes it an appealing algorithm in addition to being very stable.

[I] The condition number of $\mathbf{A}^T\,\mathbf{A}$ (the ratio of minimum to maximum eigenvalues) is the square of the condition number of $\mathbf{A}$.

[II] The backslash operator in Matlab™ computes the solution to an over-determined system of equations using QR decomposition. The pseudo-inverse function in Matlab™ uses singular value decomposition where any singular values less than a tolerance are eliminated from the inverse. The least-squares solution may be computed using either routine: x = A\y or x = pinv(A)*y.

$$\mathbf{x} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y} \tag{B.4}$$

It is also useful at times to weight each equation to account for differences in significance that occur, for example, with nonuniform distributions in the data or with different levels of confidence in the measurements. To accomplish this, Equation B.5 includes a weighting matrix **W** to create weighted terms ($\mathbf{A}_W$ and $\mathbf{y}_W$) from which **x** is solved as before. Usually **W** is diagonal but it may be full (and symmetric) for chi-square fitting, which uses the covariance matrix as weights according to $\mathbf{W}^2 = \mathbf{V}^{-1}$. Chi-square fitting is statistically motivated using a premise called maximum likelihood estimation. When normally distributed measurement errors can be assigned a covariance matrix, chi-square fitting maximizes the probability that the data set lies within an arbitrary but constant error band about the model. Even when the errors are not normal, assigning weights according to confidence in the measurements is a reasonable approach.

$$\mathbf{e}_w \equiv \mathbf{W}\mathbf{e} = \mathbf{W}[\mathbf{y} - \mathbf{A}\,\mathbf{x}] \equiv \mathbf{y}_w - \mathbf{A}_w\,\mathbf{x} \tag{B.5}$$

Given this elegant theory to the linear least-squares problem, there remain several practical issues that really motivated this presentation. The first is setting up models for several useful applications. A main difficulty lies in arranging the equations so that the residual errors lie in a desirable direction, typically perpendicular to the curve or surface. This may not be possible within the constraints of linear least squares, which motivates the second issue of nonlinear least squares. The simple approach presented here will work well for most applications, but more elaborate algorithms are available in the references.[1] The third topic is singular value decomposition, which has special utility for model fitting. The presentations of these topics occur in reverse order.

B.1 Solution by Singular Value Decomposition

Singular value decomposition (SVD) is an orthogonalizing procedure that picks out a set of directions where the matrix becomes decoupled. SVD is very similar to eigenvalue analysis and they are equivalent when the matrix is symmetric and positive semi-definite. SVD is useful in least-squares problems because it factors **A** into three separately invertible matrices **U**, Σ and **V**. These matrices are trivial to invert since **U** and **V** have orthonormal columns and Σ is diagonal (the number crunching takes place in the factorization). Computationally, SVD is more expensive than solving the normal equations by Gauss elimination, but its robustness and particularly the information in the orthogonalized model are main motivations for its use.

---

[1] The optimization toolbox in Matlab™ has several nonlinear minimization routines available. The function called *leastsq* is most applicable to this discussion.

Equation B.6 shows the compact definition of SVD either as a matrix triple product or a sum of outer (vector) products.[I] The latter form indicates that **A** can always be constructed from a sum of rank-one matrices and the singular values (the σ's) indicated the significance of each.[II] Equation B.7 helps in visualizing the shapes of the matrices and it emphasizes the importance of columns in the definition. As a matter of convention, the singular values are positive and ordered in descending value (although repeated values are possible), and the left and right singular vectors appear in corresponding positions. For least-squares problems, **A** and therefore **U** will have many more rows than columns, since typically $m > n$ by one or two orders of magnitude, where m is the number of data points and n is the number of parameters in the model. Alternatively, the model, represented in the **a**'s, may be constructed from combinations of n orthogonal functions, represented in the **u**'s, where the columns of (Σ $\mathbf{V}^T$) provide the correct combinations.

$$\mathbf{A} = \mathbf{U}\,\Sigma\,\mathbf{V}^T = \sum_{i=1}^{n} \sigma_i\,\mathbf{u}_i\,\mathbf{v}_i^T \tag{B.6}$$

$$\underset{m\,x\,n}{\left[\left\{\mathbf{a}_1\right\} \cdots \left\{\mathbf{a}_n\right\}\right]} = \underset{m\,x\,n}{\left[\left\{\mathbf{u}_1\right\} \cdots \left\{\mathbf{u}_n\right\}\right]} \cdot \underset{n\,x\,n}{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}} \cdot \underset{n\,x\,n}{\left[\left\{\mathbf{v}_1\right\} \cdots \left\{\mathbf{v}_n\right\}\right]^T} \tag{B.7}$$

Although not required for least squares, Equation B.8 and Equation B.9 show the matrix products $\mathbf{A}^T\,\mathbf{A}$ and $\mathbf{A}\,\mathbf{A}^T$ and the relationship of SVD to eigenvalues and eigenvectors. Both products have the same nonzero eigenvalues, as they must, but their eigenvectors (the left and right singular vectors) become increasingly different from one another as **A** departs from symmetry. This also explains why the normal equations are more prone to numerical round-off errors; the condition number, equal to the ratio σ_n/σ_1, is squared for $\mathbf{A}^T\,\mathbf{A}$. Although not nearly as efficient, it is straightforward to compute the SVD using an eigenvalue algorithm on the matrix products.

$$\mathbf{A}^T\mathbf{A} = \mathbf{V}\,\Sigma\,\mathbf{U}^T\,\mathbf{U}\,\Sigma\,\mathbf{V}^T = \mathbf{V}\,\Sigma^2\,\mathbf{V}^T = \mathbf{V}\,\Lambda\,\mathbf{V}^T \tag{B.8}$$

$$\mathbf{A}\,\mathbf{A}^T = \mathbf{U}\,\Sigma\,\mathbf{V}^T\,\mathbf{V}\,\Sigma\,\mathbf{U}^T = \mathbf{U}\,\Sigma^2\,\mathbf{U}^T = \mathbf{U}\,\Lambda\,\mathbf{U}^T \tag{B.9}$$

[I] The compact definition of SVD, given here, is certainly more prevalent than the formal definition that has **U** as m x m and Σ as m x n (**V** remains n x n). Since only the diagonal terms in Σ can be nonzero, only the first n columns of **U** carry through the multiplication.

[II] SVD provides a simple way to approximate a matrix by eliminating its insignificant singular values.

After processing **A** through SVD, the pseudo-inverse of **A** (and the least-squares solution $\mathbf{x} = \mathbf{A}^+ \mathbf{y}$) is simple to calculate by inverting **U**, Σ and **V** as described by Equation B.10. Comparing it to Equation B.6, the pseudo-inverse is constructed from the transpose of the same outer products but is weighted inversely by the singular values. If the condition number of **A** approaches the precision of the computer, it is advisable to approximate **A** by neglecting insignificant singular values so that their reciprocals do not contaminate the pseudo-inverse with noise. [Press, et al., 1992] recommend setting the threshold ε to *m* times the machine precision, which is the default tolerance in the Matlab™ function pinv(A). While this approach makes the computation more reliable, the need to use it usually indicates that the model is redundant or the data set does not properly span its degrees of freedom. Either problem requires further investigation to ensure a reliable fit.

$$\mathbf{A}^+ = \mathbf{V}\,\Sigma^+\,\mathbf{U}^T = \sum_{i=1}^{n} \sigma_i^+\,\mathbf{v}_i\,\mathbf{u}_i^T \qquad \sigma_i^+ = \begin{cases} \sigma_i^{-1} & \text{if } \sigma_i > \varepsilon\,\sigma_1 \\ 0 & \text{if } \sigma_i \le \varepsilon\,\sigma_1 \end{cases} \tag{B.10}$$

Equation B.11 shows the least-squares solution computed from the SVD of **A**. While this is the end goal of the calculation, more information of value is available in the orthogonalized model. For instance, the columns of **U** hold the shape of the orthogonal functions. Plotting the orthogonal functions may reveal ways to reduce the coupling in the original basis functions. Equation B.12 gives the exact recipe for obtaining this particular set of orthogonal functions (they depend on the original set). The participation of the orthogonal functions in the least-squares solution does not necessarily correspond to the singular values but rather to $(\Sigma^+\,\mathbf{U}^T\,\mathbf{y})$, since **V** is normalized. This will tell if any of the orthogonal functions can be eliminated and still provide a good fit. Since the original basis functions are generally not orthogonal, it may not be obvious if any can be eliminated simply by looking at the solution **x**. Normalizing **A**, as indicated in Equation B.13, will reduce the underlying bias in **x** making simplifications of the model more obvious. Polynomial functions are simple to normalize as indicated in Equation B.14. In addition, it is helpful if the origin is in the center of the range to take advantage of symmetry.

$$\mathbf{x} = \mathbf{V}\,\Sigma^+\,\mathbf{U}^T\,\mathbf{y} \tag{B.11}$$

$$\mathbf{U} = \mathbf{A}\left(\mathbf{V}\,\Sigma^+\right) \tag{B.12}$$

$$\mathbf{A}_N = m\begin{bmatrix} \|\mathbf{a}_1\| & & \\ & \ddots & \\ & & \|\mathbf{a}_n\| \end{bmatrix}^{-1} \cdot \mathbf{A} \tag{B.13}$$

$$p_k(x) = \left(\frac{x}{\bar{x}}\right)^k \qquad \bar{x} \equiv \left\{\frac{x_{max}^{\,2k+1} - x_{min}^{\,2k+1}}{(2k+1)(x_{max} - x_{min})}\right\}^{\frac{1}{2k}} \tag{B.14}$$

Finally, it is interesting to see what constitutes the residual error. Equation B.15 is reminiscent of an eigenvalue problem, but the square matrix product $\mathbf{U}\,\mathbf{U}^T$ is special. It has n unity eigenvalues and m - n zero eigenvalues by its construction from n orthogonal vectors. However, there are still m unique eigenvectors that span the full space of $\mathbf{y}$ and are in effect components of $\mathbf{y}$. Those components that correspond to unity eigenvalues will produce no error because they satisfy (B.15) as an eigenvalue problem. The remaining components go directly into the error because the eigenvalue problem that they satisfy has $\mathbf{I}$ multiplied by zero. The error is minimum because all directions that can be affected by the model are zeroed out while the rest are unaffected in the fitting process.

$$\mathbf{e} = \mathbf{y} - \mathbf{A}\,\mathbf{x} = \left(\mathbf{I} - \mathbf{U}\,\mathbf{U}^T\right)\mathbf{y} \tag{B.15}$$

B.2 Nonlinear Least Squares

The least-squares problem becomes nonlinear when some or all parameters appear as nonlinear functions in the model. The model evaluated at each data point can no longer be represented as a matrix of numbers. Instead, Equation B.16 represents the model as an m-dimensional vector of nonlinear functions each evaluated at a corresponding data point. Solving this problem requires an iterative technique such as Newton's method. Here the approach is to linearize the model with respect to the parameters and then to use the linear least-squares solution on this approximate model. Repeated solutions of the linearized model will converge to the solution of the nonlinear problem, provided that the neglected higher-order terms are indeed negligible. Convergence is apparent when the solution vector stops changing.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} f_1([x_1 \;\; x_2 \;\; \ldots \;\; x_n]) \\ f_2([x_1 \;\; x_2 \;\; \ldots \;\; x_n]) \\ \vdots \\ f_m([x_1 \;\; x_2 \;\; \ldots \;\; x_n]) \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix} \tag{B.16}$$

$$\mathbf{y} = \mathbf{F}(\mathbf{x}) + \mathbf{e}$$

Equation B.17 shows a Taylor series expansion of $\mathbf{F}(\mathbf{x})$ about an operating point $\mathbf{x}_{(k)}$ in parameter space, where k indicates a particular step in the iteration. However in this arrangement, the (linear least-squares) solution is obvious. The iteration algorithm follows directly in Equation B.18. One point to notice is that the error minimized by this routine includes the approximation error in addition to the fitting error. Since the approximation error is second order, it converges rapidly to zero leaving only the fitting error that occurs at the nonlinear least-squares solution. As in the linear problem, it is simple to weight the measurements using a weighting matrix $\mathbf{W}$ in the same manner as Equation B.5.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} - \begin{bmatrix} f_1\left(\mathbf{x}_{(k)}\right) \\ f_2\left(\mathbf{x}_{(k)}\right) \\ \vdots \\ f_m\left(\mathbf{x}_{(k)}\right) \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}_{(k)} \cdot \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \vdots \\ \Delta x_n \end{bmatrix}_{(k)} + \begin{bmatrix} e'_1 \\ e'_2 \\ \vdots \\ e'_m \end{bmatrix}_{(k)} \tag{B.17}$$

$$\mathbf{y} - \mathbf{F}\left(\mathbf{x}_{(k)}\right) \equiv \Delta\mathbf{y}_{(k)} = \mathbf{A}_{(k)} \cdot \Delta\mathbf{x}_{(k)} + \mathbf{e}'_{(k)} \qquad \Delta\mathbf{x}_{(k)} \equiv \mathbf{x} - \mathbf{x}_{(k)}$$

$$\mathbf{x}_{(k+1)} = \Delta\mathbf{x}_{(k)} + \mathbf{x}_{(k)} \tag{B.18}$$

B.3 Planar Fit

The description of a plane requires three independent parameters but the equation is more generally written with four parameters. For example, Equation B.19 has four parameters (a, b, c, d) with a constraint equation to create one dependency. With (a, b, c) constrained as a unit gradient vector, d is the perpendicular distance from the origin to the plane. The plane intercepts the x, y, z axes at d/a, d/b, and d/c, respectively. Although this particular constraint equation is nonlinear, it does not need to be satisfied exactly until after fitting.

$$\begin{bmatrix} x & y & z \end{bmatrix} \cdot \begin{bmatrix} a \\ b \\ c \end{bmatrix} = d \qquad \text{subject to} \qquad a^2 + b^2 + c^2 = 1 \tag{B.19}$$

An alternative approach creates three new independent parameters by dividing through by one parameter, effectively eliminating it from the fit. The difficulty is that any of the four parameters could have zero value. For example, the plane could pass through the origin ($d = 0$) or be parallel to any axis (a, b or $c = 0$). Effectively, both approaches require a crude estimate of the gradient to begin. Given the estimate, the approach using a constraint equation is cleaner to code; therefore, it is the approach presented here.

Equation B.20 shows the matrix equation for m data points plus the constraint equation. In many applications the gradient will be known to lie in a general direction; the closest coordinate axis is sufficient. A very good numerical estimate comes from using the cross product on three points that are sufficiently spread compared to the size of errors. This information goes in the last row of the matrix. Notice that $\mathbf{y}$ would be all zeros if not for the constraint equation and would not give a unique solution. In addition, the constraint is solved with zero error but the solution vector will require scaling to achieve a unit gradient vector since the estimate will generally not be exact.

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \hline 1 \end{bmatrix} = \left[\begin{array}{cccc} x_1 & y_1 & z_1 & -1 \\ x_2 & y_2 & z_2 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ x_m & y_m & z_m & -1 \\ \hline a & b & c & 0 \end{array}\right] \cdot \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \\ \hline 0 \end{bmatrix} \tag{B.20}$$

Notice that the planar fit reduces to a linear fit of x-y data. Although similar to the first example, Equation B.1, this one fits the equation $d = a\,x + b\,y$.

B.4 Spherical Fit

The description of a sphere requires four independent parameters, the location of its center (a, b, c) and its radius r. On the left, Equation B.21 expresses the distance computed along Cartesian coordinates from the center of the fitted sphere to any data point i near the surface; while on the right, the radial distance includes the fitted radius and the radial error. The first step in the derivation is to expand the squares and segregate terms as shown in Equation B.22. A key observation is that the squared parameters are constant and can be represented as a new variable d.[1] Also notice that a squared error term has been neglected; it just means the error vector minimized is slightly different from the ideal. Now linear in a, b, c and d, equations for all surface points are assembled into Equation B.23 for a linear least-squares analysis. Finally, r is computed from the fitted parameters in Equation B.24.

$$(x_i - a)^2 + (y_i - b)^2 + (z_i - c)^2 = (r + e_i)^2 \tag{B.21}$$

$$x_i^2 + y_i^2 + z_i^2 \cong 2\left[a\,x_i + b\,y_i + c\,z_i\right] + \underbrace{r^2 - a^2 - b^2 - c^2}_{2d} + 2\,r\,e_i \tag{B.22}$$

$$\frac{1}{2}\begin{bmatrix} x_1^2 + y_1^2 + z_1^2 \\ x_2^2 + y_2^2 + z_2^2 \\ \vdots \\ x_m^2 + y_m^2 + z_m^2 \end{bmatrix} \cong \begin{bmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_m & y_m & z_m & 1 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} + r \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix} \tag{B.23}$$

$$r = \sqrt{a^2 + b^2 + c^2 + 2d} \tag{B.24}$$

Notice that the spherical fit reduces to a circular fit of x-y data, or to a cylindrical fit of x-y-z data by dropping the component of the data in the direction of the cylinder's axis. This cylinder is constrained to be parallel to the chosen axis, unlike the cylindrical fit presented later.

---

[1] Suggested by Terry Malsbury, my colleague at LLNL.

B.5 Linear Fit

A line through three-dimensional space may be represented as a parametric curve where its x, y, z coordinates vary as functions of a single variable t, usually referred to as the parameter. It may be confusing that the constants in the parametric curve are the parameters used for curve fitting. We will distinguish them by using the term *fitting parameter*. For a line, it is intuitive to think of starting and ending points each represented by a position vector. Any point on the line is a linear combination of the two vectors as determined by the parameter t in Equation B.25. Although usually designed to vary uniformly between 0 and 1, the parameter could vary nonuniformly and/or between different limits. Given a discrete set of (x, y, z) points dispersed evenly in time or in a known sequence, then Equation B.26 shows the points and the model of a line cast in the form appropriate for linear least squares.[I] This model has six fitting parameters, two more than required for a line. With the range of the parameter t defined, the extra degrees of freedom allow the line to shift and scale so that each error aligns perpendicular to the line, provided that the parameter sequence corresponds to the point dispersion.

$$\begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}(1-t) + \begin{bmatrix} u \\ v \\ w \end{bmatrix} t \tag{B.25}$$

$$\left[\begin{array}{c} x_1 \\ y_1 \\ z_1 \\ \hline x_2 \\ y_2 \\ z_2 \\ \hline \vdots \end{array}\right] = \left[\begin{array}{ccc:ccc} 1-t_1 & 0 & 0 & t_1 & 0 & 0 \\ 0 & 1-t_1 & 0 & 0 & t_1 & 0 \\ 0 & 0 & 1-t_1 & 0 & 0 & t_1 \\ \hline 1-t_2 & 0 & 0 & t_2 & 0 & 0 \\ 0 & 1-t_2 & 0 & 0 & t_2 & 0 \\ 0 & 0 & 1-t_2 & 0 & 0 & t_2 \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right] \cdot \left[\begin{array}{c} a \\ b \\ c \\ \hline u \\ v \\ w \end{array}\right] + \left[\begin{array}{c} e_{x_1} \\ e_{y_1} \\ e_{z_1} \\ \hline e_{x_2} \\ e_{y_2} \\ e_{z_2} \\ \hline \vdots \end{array}\right] \tag{B.26}$$

The situation becomes more difficult if the timing of the points is unknown because each point requires a unique parameter value for its error to be perpendicular. The unique value corresponds to the point on the line that is closest to the particular data point. This is easy to find by projecting the data point onto the line, but this first requires knowing the fitting parameters of the line. The problem is circular and requires an iterative solution; first finding an approximate solution to the linear fit then refining the model with updates to the parameter values and so on.[II] Equation B.27 provides the parameter value where the error

---

[I] A different order of equations is certainly possible. For example, the x components for all points could occur first followed by y components, etc.

[II] This problem could be given the full formalism of nonlinear least squares but this iterative approach results in a simpler algorithm.

is perpendicular to the line. In this case, the two extra degrees of freedom require constraints on the start and end of the sequence, say 0 and 1. This is easy to do by shifting the sequence by the minimum value calculated, then by scaling the shifted sequence by its maximum value.

$$t_i = \frac{(x_i - a)(u - a) + (y_i - b)(v - b) + (z_i - c)(w - c)}{(u - a)^2 + (v - b)^2 + (w - c)^2} \tag{B.27}$$

Starting the iteration requires either creating a time sequence to use in Equation B.26 or estimating fitting parameters to use in Equation B.27. The latter option is most robust since the extreme data points are easily found and are very close to the endpoints of the line. Once started, the iteration converges quickly as indicated by the change in the solution going to zero.

B.6 Fitting Surfaces of Revolution

Surfaces of revolution have characteristics similar to the plane, the sphere and the line, and fitting them naturally builds upon that work. The techniques allow for an arbitrary and initially unknown orientation of the axis, and the minimized errors are perpendicular to the surface. The basic approach maps three-dimensional data to a radial-axial CS where the function is fitted. While the function may be linear in the local CS, the mapping is nonlinear and necessitates an iterative solution. In a later section, *Fitting Quadratic Surfaces*, the surface has two arbitrary axes (the third is always orthogonal) and requires a full coordinate transformation within the fitting routine.

B.6.1 Cylindrical Fit

The description of a cylinder requires five independent parameters, the radius and four parameters describing the axis. It is convenient, as in the case of the linear fit, to describe the axis with six parameters and two constraint equations. In Figure B-1, the unit-length vector (u, v, w) describes the direction of the axis and the point (a, b, c) near the centroid of the data describes it location. The right triangle formed by the axis, point (a, b, c) and any data point (x_i, y_i, z_i) near the surface is the basis for Equation B.28, which expresses the lengths of its sides. The choice of (a, b, c) near the centroid reduces coupling between the parameters. After expanding squares and collecting terms, Equation B.29 differs from the spherical fit by only the projected distance p_i between the data point and (a, b, c). This extra term makes the equation nonlinear. In addition, the constraint equations are nonlinear.

$$\begin{aligned} (x_i - a)^2 + (y_i - b)^2 + (z_i - c)^2 &= p_i^2 + (r + e_i)^2 \\ p_i^2 &= \left[(x_i - a)u + (y_i - b)v + (z_i - c)w\right]^2 \end{aligned} \tag{B.28}$$

$$x_i^2 + y_i^2 + z_i^2 \cong 2\left[a\,x_i + b\,y_i + c\,z_i\right] + \underbrace{r^2 - a^2 - b^2 - c^2}_{2d} + p_i^2 + 2\,r\,e_i \tag{B.29}$$

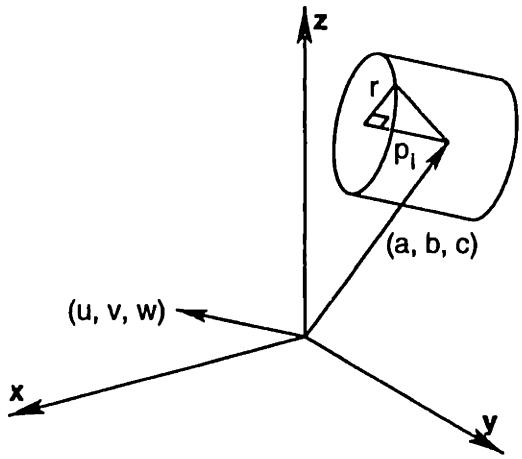


Figure B-1 A cylinder may be constructed from a point (a, b, c) on the axis, a unit vector (u, v, w) in the direction of the axis and the radius r. A point (x_i, y_i, z_i) projected to the axis is a distance p_i from (a, b, c).

When assembling the set of $m + 2$ equations, it is useful to segregate the linear part, represented by the matrix in Equation B.30, from the nonlinear part that changes with each iteration. Following the nonlinear least-squares procedure, Equations B.31 and B.32 show the linearized model obtained by differentiating $\mathbf{F}$. Of course the last two rows correspond to the constraint equations. The iteration algorithm is the same as Equation B.18.

$$\mathbf{F}\left(\mathbf{x}_{(k)}\right) = \left[\begin{array}{cccc} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_m & y_m & z_m & 1 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \cdot \left[\begin{array}{c} a \\ b \\ c \\ d \end{array}\right]_{(k)} + \frac{1}{2}\left[\begin{array}{c} p_1^2 \\ p_2^2 \\ \vdots \\ p_m^2 \\ \hline u^2 + v^2 + w^2 \\ 2\left[(a - a_c)u + (b - b_c)v + (c - c_c)w\right] \end{array}\right]_{(k)} \tag{B.30}$$

$$\mathbf{A}_{(k)} = \left[\begin{array}{cccc|ccc} x_1 & y_1 & z_1 & 1 & p_1 x_1 & p_1 y_1 & p_1 z_1 \\ x_2 & y_2 & z_2 & 1 & p_2 x_2 & p_2 y_2 & p_2 z_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_m & y_m & z_m & 1 & p_m x_m & p_m y_m & p_m z_m \\ \hline 0 & 0 & 0 & 0 & u & v & w \\ u & v & w & 0 & a - a_c & b - b_c & c - c_c \end{array}\right]_{(k)} \tag{B.31}$$

Getting the iteration started requires an estimate for both (a, b, c) and (u, v, w). The user is best able to estimate the direction of the axis from a plot of the data or knowledge of the experiment. The closest coordinate axis is convenient and usually sufficient. The centroid of the data is simple to calculate for (a, b, c).

$$\frac{1}{2}\begin{bmatrix} x_1^2+y_1^2+z_1^2 \\ x_2^2+y_2^2+z_2^2 \\ \vdots \\ x_m^2+y_m^2+z_m^2 \\ \hline 1 \\ 0 \end{bmatrix} - \mathbf{F}\left(\mathbf{x}_{(k)}\right) \cong \mathbf{A}_{(k)} \cdot \begin{bmatrix} \Delta a \\ \Delta b \\ \Delta c \\ \Delta d \\ \Delta u \\ \Delta v \\ \Delta w \end{bmatrix}_{(k)} + r \begin{bmatrix} e_1' \\ e_2' \\ \vdots \\ e_m' \\ \hline 0 \\ 0 \end{bmatrix}_{(k)} \tag{B.32}$$

B.6.2 Conical Fit

The description of a cone requires six independent parameters, one more than a cylinder. As with the cylinder, six parameters describe the axis but only the direction vector (u, v, w) requires a constraint equation. In Figure B-2, the point (a, b, c) is the vertex of the cone and the origin of the radial-axial CS. Formed from the right triangle between the axis and any data point, Equation B.33 expresses the mapping from (x, y, z) space to (r, p) space.[1]

$$\begin{aligned} r_i^2 &= (x_i-a)^2+(y_i-b)^2+(z_i-c)^2-p_i^2 \\ p_i^2 &= \frac{\left[(x_i-a)u+(y_i-b)v+(z_i-c)w\right]^2}{u^2+v^2+w^2} \end{aligned} \tag{B.33}$$

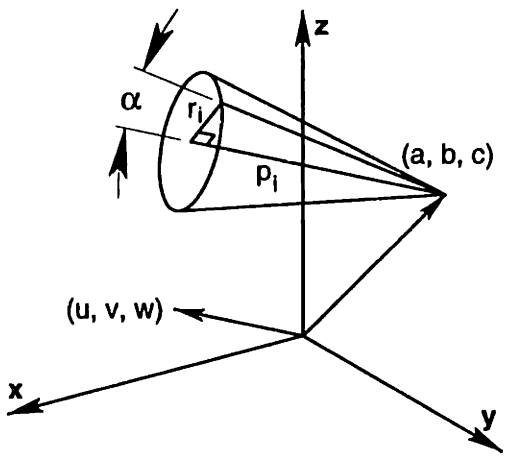


Figure B-2 A cone may be constructed from its vertex (a, b, c), a unit vector (u, v, w) in the direction of the axis and the cone angle α. A point (x_i, y_i, z_i) projected on the axis is a distance p_i from (a, b, c).

The remaining parameter is the half cone angle α. Equation B.34 expresses the perpendicular error between the fitted cone and the data point; however, a problem arises with using the first power of r_i. It introduces a square root term that when differentiated puts r_i in the denominator, thus allowing a possible divide by zero. Squaring each side

[1] Notice that Equation B.33 does not rely on (u, v, w) being constrained to unit magnitude, in contrast to Equation B.28. This is necessary to prevent a null solution where $(u, v, w) = (0, 0, 0)$ and $\alpha = \pi/2$.

solves the problem but effectively weights the error by the radius. This may be appropriate if the data is equally spaced in angle about the axis; otherwise the fit is not quite ideal.

$$\begin{aligned} p_i \sin(\alpha) &= r_i \cos(\alpha) + e_i \\ p_i^2 \sin^2(\alpha) &\cong r_i^2 \cos^2(\alpha) + 2 r_i \cos(\alpha) e_i \end{aligned} \tag{B.34}$$

Equation B.35 expresses the same relationship in a form slightly more convenient for the nonlinear least-squares procedure. It makes geometric sense, as a point projected to the axis must equal the hypotenuse times the cosine of the half cone angle.

$$\begin{aligned} 0 &= \underbrace{\frac{1}{2}\left(h_i^2 \cos^2(\alpha) - p_i^2\right)}_{f_i(\mathbf{x})} + r_i \cos(\alpha) e_i \\ h_i^2 &= (x_i - a)^2 + (y_i - b)^2 + (z_i - c)^2 \end{aligned} \tag{B.35}$$

Equation B.36 provides the vector of seven partial derivatives for the nonlinear function f_i. It must be evaluated for all m data points and assembled into Equation B.37 for the least-squares solution. The last row is the constraint on (u, v, w) to be unit length. Starting the iteration requires an initial estimate for all seven parameters. Again the user is best able to make the estimate from a plot of the data.

$$\begin{aligned} \partial f_i &= \begin{bmatrix} -h_i^2 \cos(\alpha)\sin(\alpha) \\ -(x_i - a)\cos^2(\alpha) + q_i u \\ -(y_i - b)\cos^2(\alpha) + q_i v \\ -(z_i - c)\cos^2(\alpha) + q_i w \\ -q_i (x_i - a) + q_i^2 u \\ -q_i (y_i - b) + q_i^2 v \\ -q_i (z_i - c) + q_i^2 w \end{bmatrix}^{\mathrm{T}} \cdot \begin{bmatrix} \partial\alpha \\ \partial a \\ \partial b \\ \partial c \\ \partial u \\ \partial v \\ \partial w \end{bmatrix} = \mathbf{a}_i \, \partial\mathbf{x} \\ q_i &= \frac{(x_i - a)u + (y_i - b)v + (z_i - c)w}{u^2 + v^2 + w^2} \end{aligned} \tag{B.36}$$

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} - \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \\ u^2 + v^2 + w^2 \end{bmatrix}_{(k)} \cong \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_m \\ 0 \;\; 0 \;\; 0 \;\; 0 \;\; 2u \;\; 2v \;\; 2w \end{bmatrix}_{(k)} \cdot \begin{bmatrix} \Delta\alpha \\ \Delta a \\ \Delta b \\ \Delta c \\ \Delta u \\ \Delta v \\ \Delta w \end{bmatrix}_{(k)} + \cos(\alpha) \begin{bmatrix} r_1 e_1' \\ r_2 e_2' \\ \vdots \\ r_m e_m' \\ 0 \end{bmatrix}_{(k)} \tag{B.37}$$

B.6.3 Fitting a General-Form Revolution

Building on the development of the conical fit, any surface of revolution may be fit to the appropriate radial function $g(r)$. Equation B.38 provides the general expression for the perpendicular error between the general surface and any data point. For the cone, $g(r)$ is equal to $r\cot(\alpha)$ and the general expression simplifies to Equation B.35. As a final example, Equation B.39 shows the fitting function for a paraboloid. Although a simple function, the partial derivatives in Equation B.40 are lengthy and probably mark the point to consider numerical derivatives. For the cone and paraboloid, $g(r)$ has only one parameter α describing the geometry of each surface. Higher order surfaces will require additional parameters and the partial derivatives will become more complicated.

$$0 = \underbrace{\left[g(r_i) - p_i\right]\left(1 + g'(r_i)^2\right)^{-1/2}}_{f_i(\mathbf{x})} + e_i \qquad g'(r) \equiv \frac{dg}{dr} \tag{B.38}$$

$$f_i(\mathbf{x}) = \left[\alpha r_i^2 - p_i\right]\psi \qquad \psi \equiv \left(1 + (2\alpha r_i)^2\right)^{-1/2} \tag{B.39}$$

$$\mathbf{a}_i = \begin{bmatrix} r_i^2\psi - f_i(2r_i\psi)^2\alpha \\ u\eta\psi - \left\{2\alpha\psi - f_i(2\alpha\psi)^2\right\}(x_i - a - q_i u) \\ v\eta\psi - \left\{2\alpha\psi - f_i(2\alpha\psi)^2\right\}(y_i - b - q_i v) \\ w\eta\psi - \left\{2\alpha\psi - f_i(2\alpha\psi)^2\right\}(z_i - c - q_i w) \\ -\left\{(1+2\alpha p_i)\psi - f_i\psi^2(2\alpha)^2 p_i\right\}(x_i - a - q_i u)\eta \\ -\left\{(1+2\alpha p_i)\psi - f_i\psi^2(2\alpha)^2 p_i\right\}(y_i - b - q_i v)\eta \\ -\left\{(1+2\alpha p_i)\psi - f_i\psi^2(2\alpha)^2 p_i\right\}(z_i - c - q_i w)\eta \end{bmatrix}^T \tag{B.40}$$

$$\eta \equiv \left(u^2 + v^2 + w^2\right)^{-1/2}$$

B.7 Fitting Quadratic Surfaces

Ellipsoids, hyperboloids and other quadratic surfaces are expressible in terms of a quadratic polynomial such as Equation B.41. This polynomial is sufficiently general to represent any quadratic surface that is orthogonal to and centered on its coordinate system. Any particular type requires only a subset of the parameters listed.[1] For example, an ellipsoid requires

[1] See, for example, [Flanders and Price, 1978] or another text on analytic geometry for further descriptions and examples.

three parameters A, B and C with k set to -1. The difficulty in fitting this simple polynomial is the coordinate transformation required to take the data from the global CS_0 to the local CS_1 where the polynomial representation is simplest.[1] In Equation B.42, a rotation matrix and a shift is equivalent to an HTM and is more convenient in this case for transforming between CS_0 and CS_1. Equation B.43 shows the required transformation transposed conveniently to work with columns of data.

$$f(x,y,z) = Ax^2 + By^2 + Cz^2 + px + qy + rz + k = 0 \tag{B.41}$$

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_0 = \mathbf{R}_{0/1}\begin{bmatrix} x \\ y \\ z \end{bmatrix}_1 + \begin{bmatrix} a \\ b \\ c \end{bmatrix} \qquad \mathbf{R}_{0/1}{}^{\mathrm{T}}\left\{\begin{bmatrix} x \\ y \\ z \end{bmatrix}_0 - \begin{bmatrix} a \\ b \\ c \end{bmatrix}\right\} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}_1 \tag{B.42}$$

$$\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_m & y_m & z_m \end{bmatrix}_1 = \left\{\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_m & y_m & z_m \end{bmatrix}_0 - [a \quad b \quad c]\right\}\mathbf{R}_{0/1} \tag{B.43}$$

$$\mathbf{y} - \mathbf{F}\left(\mathbf{x}_{(k)}\right) = \mathbf{A}_{(k)} \cdot \Delta\mathbf{x}_{(k)} + \mathbf{e}'_{(k)} \tag{B.44}$$

$$\mathbf{A}_{(k)} = \left[\begin{array}{ccc|ccc|ccc|ccc} x_1{}^2 & y_1{}^2 & z_1{}^2 & x_1 & y_1 & z_1 & \frac{\partial f_1}{\partial a} & \frac{\partial f_1}{\partial b} & \frac{\partial f_1}{\partial c} & \frac{\partial f_1}{\partial \theta_x} & \frac{\partial f_1}{\partial \theta_y} & \frac{\partial f_1}{\partial \theta_z} \\ x_2{}^2 & y_2{}^2 & z_2{}^2 & x_2 & y_2 & z_2 & \frac{\partial f_2}{\partial a} & \frac{\partial f_2}{\partial b} & \frac{\partial f_2}{\partial c} & \frac{\partial f_2}{\partial \theta_x} & \frac{\partial f_2}{\partial \theta_y} & \frac{\partial f_2}{\partial \theta_z} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_m{}^2 & y_m{}^2 & z_m{}^2 & x_m & y_m & z_m & \frac{\partial f_m}{\partial a} & \frac{\partial f_m}{\partial b} & \frac{\partial f_m}{\partial c} & \frac{\partial f_m}{\partial \theta_x} & \frac{\partial f_m}{\partial \theta_y} & \frac{\partial f_m}{\partial \theta_z} \end{array}\right]$$

$$\Delta\mathbf{x}_{(k)} = \left[\Delta A \quad \Delta B \quad \Delta C \,\vdots\, \Delta p \quad \Delta q \quad \Delta r \,\vdots\, \Delta a \quad \Delta b \quad \Delta c \,\vdots\, \Delta\theta_x \quad \Delta\theta_y \quad \Delta\theta_z\right]^{\mathrm{T}}$$

In forming Equation B.44, it will be necessary to remove unused terms from the polynomial and possibly move a constant term into **y**. Although **F** appears linear in the parameters, the parameters a, b, c, θ_x, θ_y and θ_z, are embedded in the transformed data points. This is apparent in **A** where their partial derivatives are calculated numerically to avoid a terrible mess. Avoiding a further complication, this fit lacks any scaling that would make the minimized errors perpendicular to the surface.

---

[1] Fitting to a more general polynomial is no problem but interpreting the fit or converting it to a pure type (such as an ellipsoid) is usually impossible.

B.8 Fitting Cubic Splines

A spline is a smooth, piecewise polynomial usually of order three with continuous first and second derivatives. The usual applications are interpolating data and representing free-form curves and surfaces. The interpolating spline passes exactly through the points that define it, and the slopes at those points are solved to maintain continuity of first and second derivatives along its length. With origins in naval architecture, the spline curve is the shape of a thin beam constrained at a number of *knot* locations along its length. The B-spline is somewhat different because the curve only passes close to a number of control points. Continuity of first and second derivatives is inherent in its basis functions giving it the useful property that the points affect only local control of the curve. In the application presented here, least-squares fitting a spline to data is somewhat unusual but is very useful since the curve has better local flexibility than polynomials and sinusoids.

There are at least two basic approaches to fitting splines. The obvious approach is to fit B-spline basis functions to the data over their local extent. Except for the ends, setting up the least-squares problem is straightforward. The least-squares solution finds the control points that put the curve closest to the data regardless whether the spline passes through them. The same problem could be developed much like the interpolating spline where the parameters are knot locations and slopes. An obvious problem occurs when the data is sparse over one or more sections; the spline may have an unrealistic shape or be completely unsolvable (a singular matrix). The approach taken here solves this problem and provides a way to stiffen the spline without need to alter the number or spacing of knots. A spline that is too flexible does not adequately average out noisy data. It raises the issue of initially selecting a stiffness parameter, but this is not much different from selecting the number of sections. Granted this is very subjective and some trial and error is inevitable.

In this approach, the minimized function includes the sum-squared error plus a measure of bending energy in the spline. Although similar to least squares, this minimization does not lead to the normal equations and the pseudo-inverse is not the correct solution. The two terms compete in the minimization process since a straight spline has zero bending energy and the most flexible spline has the least fitting error. The compromise found by the minimization depends on the stiffness parameter entered by the user and on other factors such as the number of sections in the spline and the particular data set. Nothing in this approach forces the spline to have a continuous second derivative as is typical. This allows the curve somewhat greater freedom to fit the data at the expense of being slightly less smooth. Physically, it is like a beam with both forces and moments applied to a number of knots.[1] Since the data points *pull* at locations different from the

[1] It seems fairly straightforward to shift knots along the spline so that the moments tend to zero, but that is not part of this development.

knots, the spline responds in the only way it can, with equivalent forces and moments applied to knots. The interpolating spline is a unique case where the data pull exactly at knots. This particular curve minimizes bending energy and just happens to have a continuous second derivative since there are no applied moments.

Turning now to the derivation, we begin in the typical way with Hermite cubics defined in Equation B.45 and shown in Figure B-3. Combined as a composite curve between 0 and 1, each piece satisfies one of four boundary conditions, either position or slope at either end. Equation B.46 demonstrates how to apply them to an arbitrary interval between x_i and x_{i+1}. The resulting cubic curve has endpoint positions p_i and p_{i+1} and endpoint slopes s_i and s_{i+1} but requires a correction term β when $\Delta x_i \neq 1$.

$$\mathbf{h}(u) = \begin{bmatrix} (u-1)^2(2u+1) \\ u(u-1)^2 \\ u^2(3-2u) \\ u^2(u-1) \end{bmatrix} \qquad \frac{d\mathbf{h}}{du} = \begin{bmatrix} 6u(u-1) \\ (3u-1)(u-1) \\ 6u(1-u) \\ u(3u-2) \end{bmatrix} \qquad \frac{d^2\mathbf{h}}{du^2} = \begin{bmatrix} 6(2u-1) \\ 2(3u-2) \\ 6(1-2u) \\ 2(3u-1) \end{bmatrix} \tag{B.45}$$

$$y(x) = \mathbf{h}(u)^T \cdot \underbrace{\begin{bmatrix} 1 & & & \\ & \Delta x_i & & \\ & & 1 & \\ & & & \Delta x_i \end{bmatrix}}_{\beta_i} \cdot \begin{bmatrix} p_i \\ s_i \\ p_{i+1} \\ s_{i+1} \end{bmatrix} \tag{B.46}$$

$$u = \frac{x - x_i}{\Delta x_i} \qquad \Delta x_i = x_{i+1} - x_i \qquad x_i \le x \le x_{i+1}$$

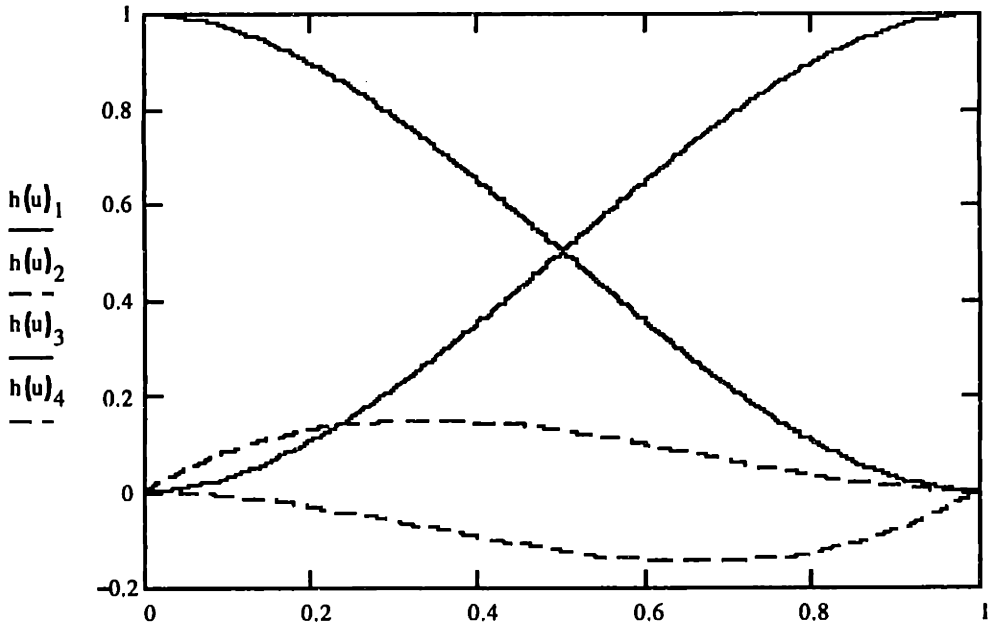


Figure B-3 Each Hermite cubic has one end with either unit position or unit slope; all others are zero.

Equation B.47 shows the formation of the least-squares problem where the Hermite cubics are evaluated at data points along the various sections of the spline and the correction terms are evaluated for the corresponding sections. Since no control is placed on the second derivative, this is the most flexible spline possible for the given number and location of knots. However, the matrix is likely to be singular for sparse data. Notice that the fitting error is measured in the y-direction rather than normal to the spline. This is adequate when the slopes are modest, for example, when used for error mapping.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_k \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1{}^{\mathrm{T}}\beta_1 & & & \\ \vdots & & & \\ \mathbf{h}_j{}^{\mathrm{T}}\beta_1 & & & \\ & \ddots & & \\ & & \mathbf{h}_k{}^{\mathrm{T}}\beta_{n-1} & \\ & & \vdots & \\ & & \mathbf{h}_m{}^{\mathrm{T}}\beta_{n-1} & \end{bmatrix} \cdot \begin{bmatrix} p_1 \\ s_1 \\ \vdots \\ p_n \\ s_n \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_j \\ \vdots \\ e_k \\ \vdots \\ e_m \end{bmatrix} \tag{B.47}$$

Equation B.48 shows the formulation of the bending energy term, where k is user defined and controls the bending stiffness of the spline.[1] The matrix $\mathbf{K}$ is only positive semidefinite because any set of parameters for a straight spline results in zero bending energy. As discussed, minimizing the sum-squared error plus the bending energy in the spline leads to a matrix equation similar to the normal equations. The solution given in Equation B.49 is different from the least-squares solution only by the presence of $\mathbf{K}$ within the inverse. The combination $\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathbf{K}$ is positive definite (meaning that a unique minimum exists) provided there are at least two separate data points in the fit. This special case results in a straight spline passing exactly through the two points.

$$E = \sum_{i=1}^{n-1} \left\{ \frac{k}{2} \int_{x_i}^{x_{i+1}} \left(\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} \right)^2 \mathrm{d}x \right\} = \begin{bmatrix} p_1 \\ s_1 \\ \vdots \\ p_n \\ s_n \end{bmatrix}^{\mathrm{T}} \cdot \underbrace{\sum_{i=1}^{n-1} \begin{bmatrix} \ddots & & & & \\ & 0 & & & \\ & & k\dfrac{\beta_i \mathbf{H} \beta_i}{\Delta x_i^3} & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix}}_{\mathbf{K}} \cdot \begin{bmatrix} p_1 \\ s_1 \\ \vdots \\ p_n \\ s_n \end{bmatrix} \tag{B.48}$$

$$\mathbf{H} \equiv \frac{1}{2} \int_0^1 \frac{\mathrm{d}^2 \mathbf{h}}{\mathrm{d}u^2} \cdot \frac{\mathrm{d}^2 \mathbf{h}^{\mathrm{T}}}{\mathrm{d}u^2} \, \mathrm{d}u = \begin{bmatrix} 6 & 3 & -6 & 3 \\ 3 & 2 & -3 & 1 \\ -6 & -3 & 6 & -3 \\ 3 & 1 & -3 & 2 \end{bmatrix} \qquad k \sim \frac{1}{24} \left(\frac{x_n - x_1}{m-1} \right)^3$$

[1] Although in this derivation k is constant all along the spline, it could vary with x or be different for each section of the spline.

$$\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{e}^{\mathrm{T}}\mathbf{e}+\mathbf{x}^{\mathrm{T}}\mathbf{K}\,\mathbf{x}\right)=\mathbf{0} \quad \rightarrow \quad \mathbf{x}=\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}+\mathbf{K}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y} \tag{B.49}$$

The example in Figure B-4 demonstrates how the fit depends on the stiffness of the spline. In the plot, k is a relative measure of stiffness based on k in Equation B.48. Each spline has the same number of knots, indicated by circles, and fits the same four data points, indicated by asterisks. The choice of stiffness is ultimately up to the judgment of the user, just as with the order for a polynomial fit. The main advantages of the spline over a polynomial are locally flexibility and better extrapolation properties. The spline tends to be straight, due to its stiffness, except where the data pulls it into shape. The other approach to fitting a spline is flexible but ineffective for extrapolation because it requires sufficient data on each section to avoid singularities.

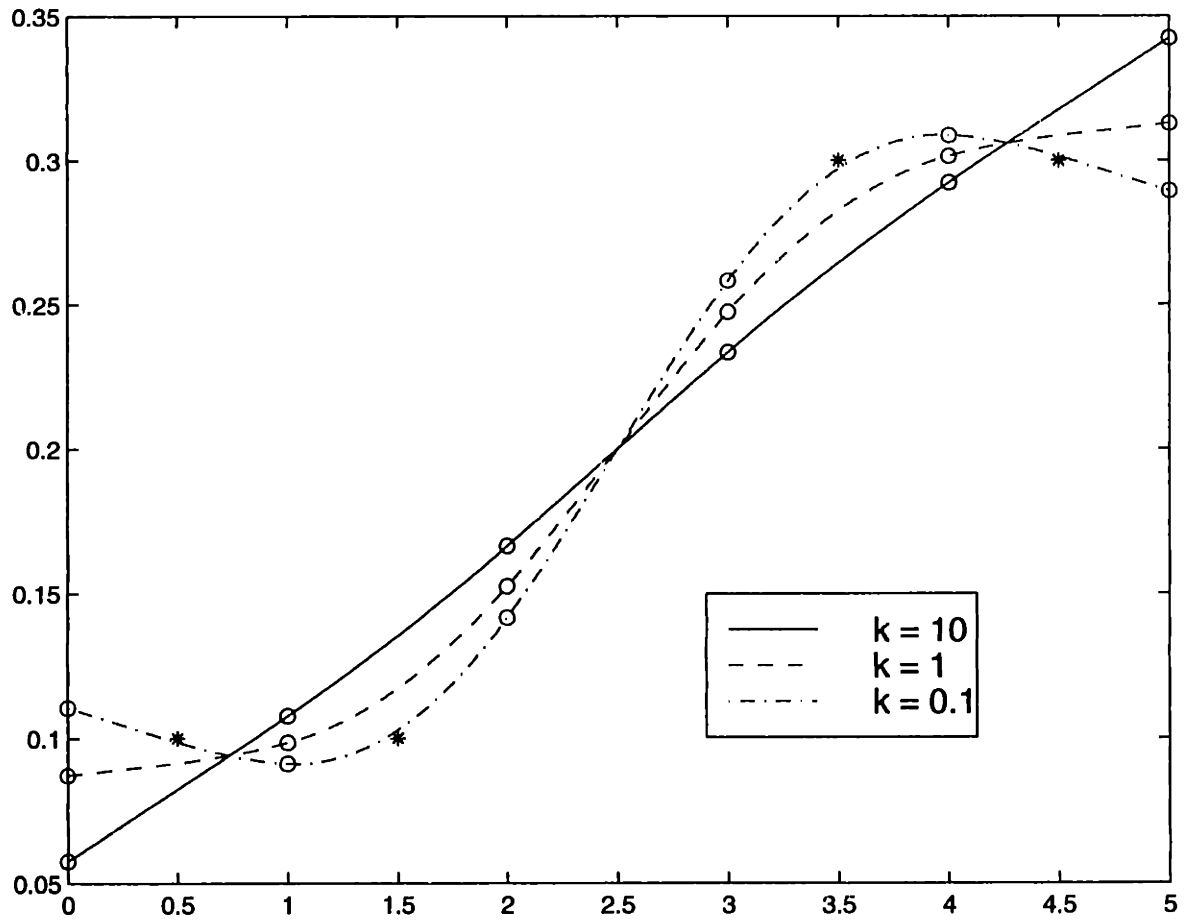


Figure B-4 The stiffness is proportional to k and controls how closely the spline fits the data, indicated by asterisks.

B.9 Matlab™ Functions for Least-Squares Fitting

This section contains Matlab™ functions for least-squares fitting data to the particular models presented throughout this chapter. The reader is granted permission to use these routines for his/her personal research. These routines may not be used commercially without permission from the author.

```
function [grad,d,rms,e] = planar_fit(xyz,cs,w)
%Planar Fit to (x,y,z) data points
%
%  [grad,d,rms,e] = planar_fit(xyz,cs,w)
%
%  Fits a plane to (x,y,z) data points such that the sum-squared error
%  normal to the plane is minimum. The plane may be in any orientation.
%  Fits a line to (x,y) data points.
%
%  Input:   'xyz' is a matrix m rows by 3 columns (2 for linear fit).
%        'cs' indi  tes the axis of the coordinate system in the
%        general direction of the gradient, where 1 = x, 2 = y, 3 = z
%        (default cs = 0, direction found using the cross product).
%        'w' is an m-dimensional vector of weights (default w = 1).
%  Output:'grad' is a unit gradient vector (or surface normal).
%        'd' is the perpendicular distance from the origin to the plane.
%        'rms' is the rms error of the fit (using normalized weights).
%        'e' is the fitted error vector (using normalized weights).

%  Layton C. Hale 6/8/98
%  Copyright (c) 1998 by Layton C. Hale

[m,n] = size(xyz);
if ~((n == 2) | (n == 3))
   error('matrix xyz must have 2 or 3 columns')
end
if nargin < 2                              % unknown orientation
   cs = 0;
end
if cs == 0                                 % find an approx. gradient
   c = mean(xyz);
   xyz_c = xyz - ones(m,1)*c;              % center the data
   [d1,p1] = max(sum(xyz_c.^2, 2));           % the most distant point
   if n == 2
      grad = [-xyz_c(p1,2), xyz_c(p1,1)]/sqrt(d1);
   else                                 % use the max. cross product
      cp = cross(xyz_c, ones(m,1)*xyz_c(p1,:));
      [d2,p2] = max(sum(cp.^2, 2));
      grad = cp(p2,:)/sqrt(d2);
   end
elseif (cs == 1) | (cs == 2) | (cs == 3)
   if cs > n
      error('cs exceeds the number of columns')
   end
   grad = zeros(1,n);
```

```
   grad(cs) = 1;
else
   error('cs must be 0, 1, 2 or 3')
end
if nargin < 3                              % no weights
   nzw = m;                               % number of nonzero weights
   A = [xyz, -ones(m,1);
       grad, 0];
else
   nzw = nnz(w);                           % number of nonzero weights
   w = w(:)*nzw/sum(w(:));                 % normalized weights
   A = [xyz.*(w*ones(1,n)), -w;
       grad, 0];
end
x = A\[zeros(m,1); 1];                     % fit the constrained model
x = x/norm(x(1:n));                        % normalize for unit gradient
grad = x(1:n)';
d = x(n+1);
e = A(1:m,:)*x;
rms = sqrt(e'*e/nzw);
```

```
function [c,r,rms,e] = spherical_fit(xyz,w)
%Spherical Fit to (x,y,z) data points
%
%  [c,r,rms,e] = spherical_fit(xyz,w)
%
%  Fits a sphere to (x,y,z) data points such that the sum-squared radial
%  error is minimum. Fits a circle to (x,y) data points.
%
%  Input:'xyz' is a matrix m rows by 3 columns (2 for circular fit).
%        'w' is an m-dimensional vector of weights (default w = 1).
%  Output:'c' is a position vector to the center.
%           'r' is the best-fit radius.
%        'rms' is the rms error of the fit (using normalized weights).
%        'e' is the fitted error vector (using normalized weights).

%  Layton C. Hale 6/8/98
%  Copyright (c) 1998 by Layton C. Hale

[m,n] = size(xyz);
if ~((n == 2) | (n == 3))
   error('matrix xyz must have 2 or 3 columns')
end
if nargin < 2                              % no weights
   nzw = m;                               % number of nonzero weights
   A = [xyz, ones(m,1)];
   y = sum(xyz.^2, 2)/2;
else
   nzw = nnz(w);                           % number of nonzero weights
   w = w(:)*nzw/sum(w(:));                 % normalized weights
   A = [xyz.*(w*ones(1,n)), w];
   y = sum(xyz.^2, 2).*w/2;
end
x = A\y;                                   % fit the model
```

```
c = x(1:n)';
r = sqrt(x(1:n)'*x(1:n) + 2*x(n+1));
e = (A*x - y)/r;
rms = sqrt(e'*e/nzw);
```

function [a,b,rms,e] = linear_fit(xyz,w)

```
%Linear Fit to (x,y,z) data points
%  [a,b,rms,e] = linear_fit(xyz,w)
%
%  Fits a parametric line to (x,y,z) data points such that the
%  sum-squared error normal to the line is minimum. The line may be
%  in any orientation. Performs an iterative solution starting with
%  a line through the extreme points of the data.
%
%  Input:   'xyz' is a matrix m rows by 3 columns.
%         'w' is an m-dimensional vector of weights (default w = 1).
%  Output:  'a' is a position vector to the start of the line.
%         'b' is a position vector to the end of the line.
%         'rms' is the rms error of the fit (using normalized weights).
%         'e' is the fitted error vector (using normalized weights).

%  Layton C. Hale 6/11/98
%  Copyright (c) 1998 by Layton C. Hale

[m,n] = size(xyz);
if n ~= 3
   error('matrix xyz must have 3 columns')
end
if nargin < 2                              % no weights
   nzw = m;                              % number of nonzero weights
   w = ones(m,1);
   y = xyz(:);
else
   nzw = nnz(w);                           % number of nonzero weights
   w = w(:)*nzw/sum(w(:));                 % normalized weights
   y = xyz(:).*[w;w;w];
end
% find the extreme points
c = mean(xyz);
xyz_c = xyz - ones(m,1)*c;                 % center the data
[d1,p1] = max(sum(xyz_c.^2, 2));
a = xyz(p1,:);                             % first extreme point
xyz_p1 = xyz - ones(m,1)*a;                   % offset the data
[d2,p2] = max(sum(xyz_p1.^2, 2));
b = xyz(p2,:);                             % second extreme point
x = [a, b]';
A = zeros(3*m,6);
for i=1:10
   b_a = b - a;
   t = (xyz - ones(m,1)*a)*b_a'/(b_a*b_a');
   t = t - min(t);
   t = t/max(t);                  % ideal parameters between 0 and 1
   tw = t.*w;
```

```
   A(1:m,1) = w - tw;          A(1:m,4) = tw;
   A(m+1:2*m,2) = w - tw;      A(m+1:2*m,5) = tw;
   A(2*m+1:3*m,3) = w - tw;    A(2*m+1:3*m,6) = tw;
   last_x = x;
   x = A\y;                              % fit the model
   a = x(1:3)';
   b = x(4:6)';
   dx = x - last_x;
   if dx'*dx/(x'*x) < 1e-30, break, end
end
e = A*x - y;
rms = sqrt(e'*e/nzw);

function [a,u,r,rms,e] = cylindrical_fit(xyz,a,u,w)
%Cylindrical Fit to (x,y,z) data points
%
%  [a,u,r,rms,e] = cylindrical_fit(xyz,a,u,w)
%
%  Fits a cylinder to (x,y,z) data points such that the sum-squared radial
%  error is minimum. Performs an iterative solution starting with a
%  center and an axis specified by the user.
%
%  Input:    'xyz' is a matrix m rows by 3 columns.
%         'a' is a vector to a point near the center of the cylinder.
%         'u' is a vector in the general direction of the cylinder's axis.
%         'w' is an m-dimensional vector of weights (default w = 1).
%  Output:  'a' is a position vector to the center of the cylinder.
%         'u' is a unit vector in the direction of the cylinder's axis.
%            'r' is the best-fit radius.
%         'rms' is the rms error of the fit (using normalized weights).
%         'e' is the fitted error vector (using normalized weights).

%  Layton C. Hale 6/20/98
%  Copyright (c) 1998 by Layton C. Hale

[m,n] = size(xyz);
if ~(n == 3)
   error('matrix xyz must have 3 columns')
end
if length(a) ~= 3
   error('vector a must have three components')
end
if length(u) ~= 3
   error('vector u must have three components')
end
if nargin < 4                              % no weights
   nzw = m;                               % number of nonzero weights
   w = ones(m,1);
   xyz_w = xyz;                            % weighted data
   y = [sum(xyz.^2, 2);1;0]/2;
else
   nzw = nnz(w);                           % number of nonzero weights
   w = w(:)*nzw/sum(w(:));                 % normalized weights
   xyz_w = xyz.*[w,w,w];                      % weighted data
```

```
   y = [sum(xyz.^2, 2).*w;1;0]/2;
end
d = -sum(a.^2)/2;
x = [a(:); d; u(:)/norm(u)];                % initial solution
a = x(1:3)';
u = x(5:7)';
A = zeros(m+2,7);
A(1:m,4) = w;
a0 = mean(xyz);
for i=1:12
   x_a = xyz - ones(m,1)*a;
   p = x_a*u';                          % points projected to the axis
   pw = p.*w;
   F = [[xyz_w, w]*x(1:4) + p.^2.*w/2;     % nonlinear model
       u*u'/2;
       (a0 - a)*u'];
   A(1:m,1:3) = xyz_w - pw*u;           % linearized model
   A(1:m,5:7) = x_a.*[pw,pw,pw];
   A(m+1,5:7) = u;
   A(m+2,1:7) = [-u, 0, a0 - a];
   dx = A\(y - F);                      % fit the linearized model
   x = dx + x;                          % iterate
   a = x(1:3)';
   u = x(5:7)';
   if dx'*dx/(x'*x) < 1e-30, break, end
end
if i == 12, warning('the solution did not converge'), end
r = sqrt(x(1:3)'*x(1:3) + 2*x(4));
e = (y(1:m) - F(1:m))/r;
rms = sqrt(e'*e/nzw);

function [a,u,alpha,rms,e] = conical_fit(xyz,a,u,w)
%Conical Fit to (x,y,z) data points
%
%  [a,u,alpha,rms,e] = conical_fit(xyz,a,u,w)
%
%  Fits a cone to (x,y,z) data points such that the sum-squared error
%  normal to the cone is minimum. Note, the errors are weighted by
%  the local radii to avoid singularity problems near zero radius.
%  Performs an iterative solution starting with a vertex and an axis
%  specified by the user.
%
%  Input:   'xyz' is a matrix m rows by 3 columns.
%        'a' is a vector to a point near the vertex of the cone.
%        'u' is a vector in the general direction of the cone's axis.
%        'w' is an m-dimensional vector of weights (default w = 1).
%  Output:  'a' is a position vector to the vertex of the cone.
%        'u' is a unit vector in the direction of the cone's axis.
%           'alpha' is the half angle of the cone.
%        'rms' is the rms error of the fit (using normalized weights).
%        'e' is the fitted error vector (using normalized weights).

%  Layton C. Hale 6/20/98
%  Copyright (c) 1998 by Layton C. Hale
```

```
[m,n] = size(xyz);
if ~(n == 3)
   error('matrix xyz must have 3 columns')
end
if length(a) ~= 3
   error('vector a must have three components')
end
if length(u) ~= 3
   error('vector u must have three components')
end
if nargin < 4                              % no weights
   nzw = m;                               % number of nonzero weights
   w = ones(m,1);
else
   nzw = nnz(w);                           % number of nonzero weights
   w = w(:)*nzw/sum(w(:));                 % normalized weights
end
x = [a(:); u(:)/norm(u); pi/4];              % initial solution
a = x(1:3)';
u = x(4:6)';
A = zeros(m+1,7);
y = [zeros(m,1); 0.5];
for i=1:15
   x_a = xyz - ones(m,1)*a;
   x_au = x_a*u';
   u2 = u*u';
   q = x_au/u2;
   p2 = x_au.^2/u2;                        % squared projection to axis
   h2 = sum(x_a.^2, 2);                    % squared hypotenuse
   s = sin(x(7));
   c = cos(x(7));
   c2 = c^2;
   F = [(h2*c2 - p2); u2]/2;                % nonlinear model
   A(1:m,1:3) = -x_a*c2 + q*u;              % linearized model
   A(1:m,4:6) = -x_a.*[q,q,q] + q.^2*u;
   A(1:m,7) = -s*c*h2;
   A(m+1,4:6) = u;
   if nargin == 4                          % include weights
      F = F.*w;
      for j=1:7
         A(1:m,j) = A(1:m,j).*w;
      end
   end
   dx = A\(y - F);                         % fit the linearized model
   x = dx + x;                            % iterate
   a = x(1:3)';
   u = x(4:6)';
   if dx'*dx/(x'*x) < 1e-30, break, end
end
if i == 15, warning('the solution did not converge'), end
alpha = x(7);
r_rms = sqrt(mean(h2 - p2));
e = -F(1:m)/(r_rms*c);
```

```
rms = sqrt(e'*e/nzw);

function [a,u,alpha,rms,e] = parabolic_fit(xyz,a,u,alpha,w)
%Parabolic Fit to (x,y,z) data points
%
%  [a,u,alpha,rms,e] = parabolic_fit(xyz,a,u,w)
%
%  Fits a paraboloid to (x,y,z) data points such that the sum-squared
%  error normal to the paraboloid is minimum. Performs an iterative
%  solution starting with a vertex, an axis and a parabolic constant
%  specified by the user.
%
%  Input:   'xyz' is a matrix m rows by 3 columns.
%        'a' is a vector to a point near the vertex of the parabola.
%        'u' is a vector in the general direction of the parabola's axis.
%           'alpha' is the approximate parabolic constant (y = alpha*x^2).
%        'w' is an m-dimensional vector of weights (default w = 1).
%  Output:  'a' is a position vector to the vertex of the parabola.
%        'u' is a unit vector in the direction of the parabola's axis.
%           'alpha' is the parabolic constant (y = alpha*x^2).
%        'rms' is the rms error of the fit (using normalized weights).
%        'e' is the fitted error vector (using normalized weights).

%  Layton C. Hale 6/20/98
%  Copyright (c) 1998 by Layton C. Hale

[m,n] = size(xyz);
if ~(n == 3)
   error('matrix xyz must have 3 columns')
end
if length(a) ~= 3
   error('vector a must have three components')
end
if length(u) ~= 3
   error('vector u must have three components')
end
if nargin < 5                              % no weights
   nzw = m;                               % number of nonzero weights
   w = ones(m,1);
else
   nzw = nnz(w);                           % number of nonzero weights
   w = w(:)*nzw/sum(w(:));                 % normalized weights
end
x = [a(:); u(:)/norm(u); alpha];           % initial solution
a = x(1:3)';
u = x(4:6)';
A = zeros(m+1,7);
y = [zeros(m,1); 0.5];
for i=1:60
   x_a = xyz - ones(m,1)*a;
   x_au = x_a*u';
   u2 = u*u';
   u1 = sqrt(u2);
   q = x_au/u2;
```

```
   p = x_au/u1;                              % projection to axis
   r2 = sum(x_a.^2, 2) - p.^2;                  % radius
   x_a_qu = x_a - q*u;
   psi2 = 1./(1 + 4*x(7)^2*r2);
   psi1 = sqrt(psi2);
   f = (x(7)*r2 - p).*psi1;
   F = [f; u2]/2;                            % nonlinear model
   A(1:m,1:3) = (4*x(7)^2*f.*psi2 - 2*x(7).*psi1)*[1,1,1]...
            .*x_a_qu + psi1*(u/u1);
   A(1:m,4:6) = (-(1 + 2*x(7)*p).*psi1 + 4*x(7)^2*f.*psi2.*p)...
            *[1,1,1].*x_a_qu/u1;
   A(1:m,7) = r2.*psi1 - 4*x(7)*f.*psi2.*r2;
   A(m+1,4:6) = u;
   if nargin == 5                            % include weights
      F = F.*w;
      for j=1:7
         A(1:m,j) = A(1:m,j).*w;
      end
   end
   dx = A\(y - F);                           % fit the linearized model
   x = dx + x;                             % iterate
   a = x(1:3)';
   u = x(4:6)';
   if dx'*dx/(x'*x) < 1e-30, break, end
end
if i == 60, warning('the solution did not converge'), end
alpha = x(7);
e = -F(1:m);
rms = sqrt(e'*e/nzw);
```

```
function [Y_fit,y,s,ys] = spline_fit(X,Y,x,k)
%Spline Fit to (X,Y) data points
%
%  [Y_fit,y,s,ys] = spline_fit(X,Y,x,k)
%
%  Fits a cubic spline to (X,Y) data points such that the sum-squared
%  error plus the bending energy in the spline are minimum. Calls "pcp".
%  The spline is a piecewise cubic polynomial defined by (x,y) junction
%  points and slopes s at the junctions. The relative stiffness of the
%  spline is controlled with k (default = 1).
%
%  Input:   X, Y and x are vectors (rows or columns). X and Y must have
%         the same number of elements. X must lie within the range
%         defined by x, and x must be monotonic. k is a real number.
%  Output:  Y_fit is a column vector with the same number of elements as
%         X and Y. y and s are column vectors with the same number of
%         elements as x. ys is a column vector formed from y and s

%  Layton C. Hale 6/6/98
%  Copyright (c) 1998 by Layton C. Hale

m = length(X);
n = length(x);
if m ~= length(Y)
```

```
   error('X and Y must have the same number of elements')
end
if nargin < 4, k = 1; end

A = pcp(X,x);              % matrix of Hermite polynomials

K = zeros(2*n,2*n);
H = [6 3 -6 3; 3 2 -3 1; -6 -3 6 -3; 3 1 -3 2]/24;
dX = (x(n) - x(1))/(m-1);

for i=1:n-1                % calculate the spline's stiffness matrix
   dx = x(i+1) - x(i);
   b = [1 dx 1 dx];
   sub = 2*(i-1) + (1:4);
   K(sub,sub) = K(sub,sub) + diag(b)*H*diag(b)*(dX/dx)^3;
end
Y = Y(:);
ys = (A'*A + k*K)\(A'*Y);  % fit the spline
Y_fit = A*ys;
y = ys(1:2:2*n-1);
s = ys(2:2:2*n);
rms_error = norm(Y-Y_fit)/sqrt(m)
bending_energy = ys'*K*ys
```

```
function A = pcp(X,x)
%Piecewise Cubic Polynomial
%
%  A = pcp(X,x)
%
%  Calculates a matrix A of Hermite polynomials. Called by "spline_fit".
%  Use to find (X,Y) points that lie on a piecewise cubic polynomial
%  defined by (x,y) junction points and slopes s at the junctions.
%
%  Y = A*ys(:); where ys = [y(:),s(:)]';
%
%  Input:  X, x are vectors (rows or columns). X must lie within the
%        range defined by x, and x must be monotonic.
%  Output: A is a matrix with as many rows as elements in X and
%        with twice as many columns as elements in x.

%  Layton C. Hale 6/6/98
%  Copyright (c) 1998 by Layton C. Hale

if (max(X) > max(x)) | (min(X) < min(x))
   error('Xi must lie within the range defined by x')
end
m = length(X); n = length(x);
dx = x(2:n) - x(1:n-1);
A = zeros(m,2*n);

if all(dx > 0)             % x increases monotonically
   dir = 1;
elseif all(dx < 0)         % x decreases monotonically
   dir = -1;
```

```
else
   error('x must be monotonic')
end
j = 1;                % index for (n-1) cubic sections
for i=1:m
   while n > 2        % find the right section
      if dir*X(i) > dir*x(j+1)
         j = j+1;
      elseif dir*X(i) < dir*x(j)
         j = j-1;
      else break
      end
   end
   u = (X(i) - x(j))/dx(j);
   A(i,2*j-1) = (u-1)^2*(2*u+1);
   A(i,2*j)   = u*(u-1)^2*dx(j);
   A(i,2*j+1) = u^2*(3-2*u);
   A(i,2*j+2) = u^2*(u-1)*dx(j);
end

function [M,v] = norm_cols(M)
%Normalize the columns of a matrix
%
%  [M,v] = norm_cols(M)
%
%  Changes the columns of a matrix to have unit norms. Used in fitting
%  routines to reduce round-off errors.
%
%  Input:    'M' is the matrix to normalize.
%  Output:   'M' is the normalized matrix.
%         'v' is a vector containing the norms of the columns

%  Layton C. Hale 6/28/98
%  Copyright (c) 1998 by Layton C. Hale

[m,n] = size(M);
for i=1:n
   vnorm = norm(M(:,i));
   if vnorm > eps
      v(i) = vnorm;
      M(:,i) = M(:,i)/vnorm;
   else
      v(i) = 1;
   end
end
```

intentionally blank

C Contact Mechanics[I]

The theory developed by Hertz in 1880 remains the foundation for most contact problems encountered in engineering. It applies to normal contact between two elastic solids that are smooth and can be described locally with orthogonal radii of curvature such as a toroid. Further, the size of the actual contact area must be small compared to the dimensions of each body and to the radii of curvature. Hertz made the assumption based on observations that the contact area is elliptical in shape for such three-dimensional bodies. The equations simplify when the contact area is circular such as with spheres in contact. At extremely elliptical contact, the contact area is assumed to have constant width over the length of contact such as between parallel cylinders. The first three sections present the Hertz equations for these distinct cases. In addition, Section C.4 presents the equations for a sphere contacting a conical socket. Hertz theory does not account for tangential forces that may develop in applications where the surfaces slide or carry traction. Extensions to Hertz theory approximate this behavior to reasonable accuracy. Section C.5 presents the more useful equations. The final section C.6 contains Mathcad™ documents that implement these equations for computer analysis.

The Hertz equations are important in the engineering of kinematic couplings particularly if the loads carried are relatively high. For a particular choice of material and contact geometry, the pertinent calculations reduce to families of curves that are convenient for sizing purposes. The typical material used is hardened bearing steel, for example, 52100 steel or 440C stainless steel heat treated to 58-62 Rockwell C. A typical contact geometry consists of a sphere against a cylindrical groove, for example, one side of a gothic arch. The graphs that follow use the elastic properties for steel and are scaled relative to the radius of the ball R_{ball} rather than its diameter. Figure C-1 shows the relationship between the load P and the maximum shear stress τ that occurs just below the surface. The allowable shear strength is approximately 58% the allowable tensile strength with $\tau = 150$ ksi being reasonable for the steels mentioned. The curves show the positive effect of curvature matching indicated by the ratio of ball radius to groove radius approaching one. A ratio of zero indicates that the groove radius is flat with infinite radius. Figure C-2 shows the normal displacement δ versus P and Figure C-3 shows the normal stiffness k versus P.

Greater load capacity and stiffness are possible by spreading the load along a line of contact, for example, with a sphere and conical socket. Three spheres and three sockets with either the spheres or the sockets supported on radial-motion blade flexures duplicate the kinematics of a three-vee coupling. The graphs that follow use the elastic properties for steel and the optimal cone angle of 45° with respect to the centerline. Figure C-4 shows the

[I] The principal reference for this chapter is Contact Mechanics by [Johnson, 1985].

relationship between the axial load P and the shear stress τ that occurs just below the surface of the contact circle. Figure C-5 shows the axial displacement δ versus P and Figure C-6 shows the axial stiffness k versus P.

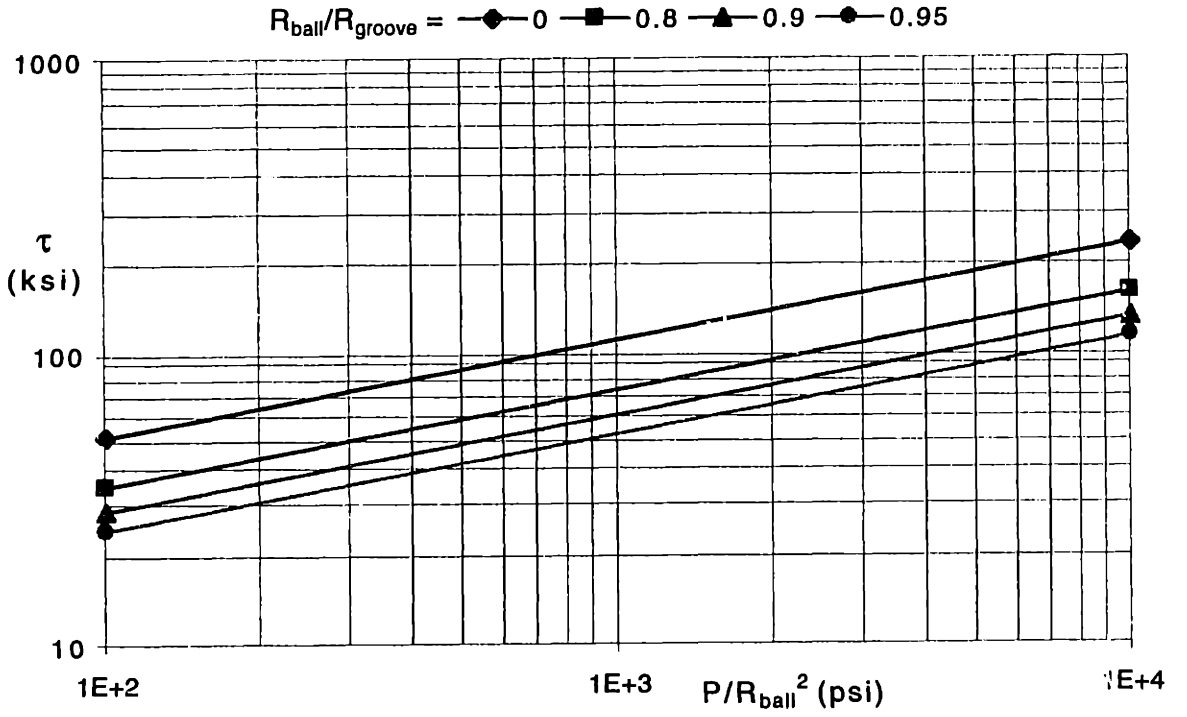


Figure C-1 Shear stress τ versus the load P for a ball against a cylindrical groove. Use this graph to determine τ, P or R_{ball} from the other two.

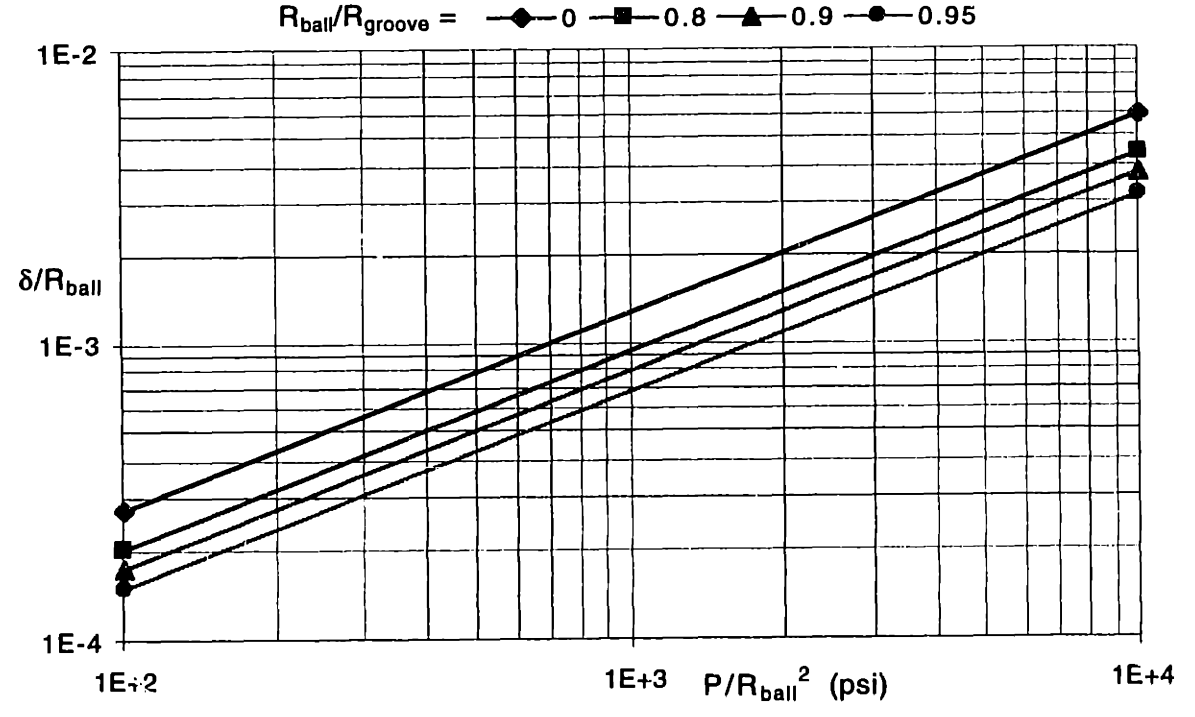


Figure C-2 Normal displacement δ versus the load P for a ball against a cylindrical groove.

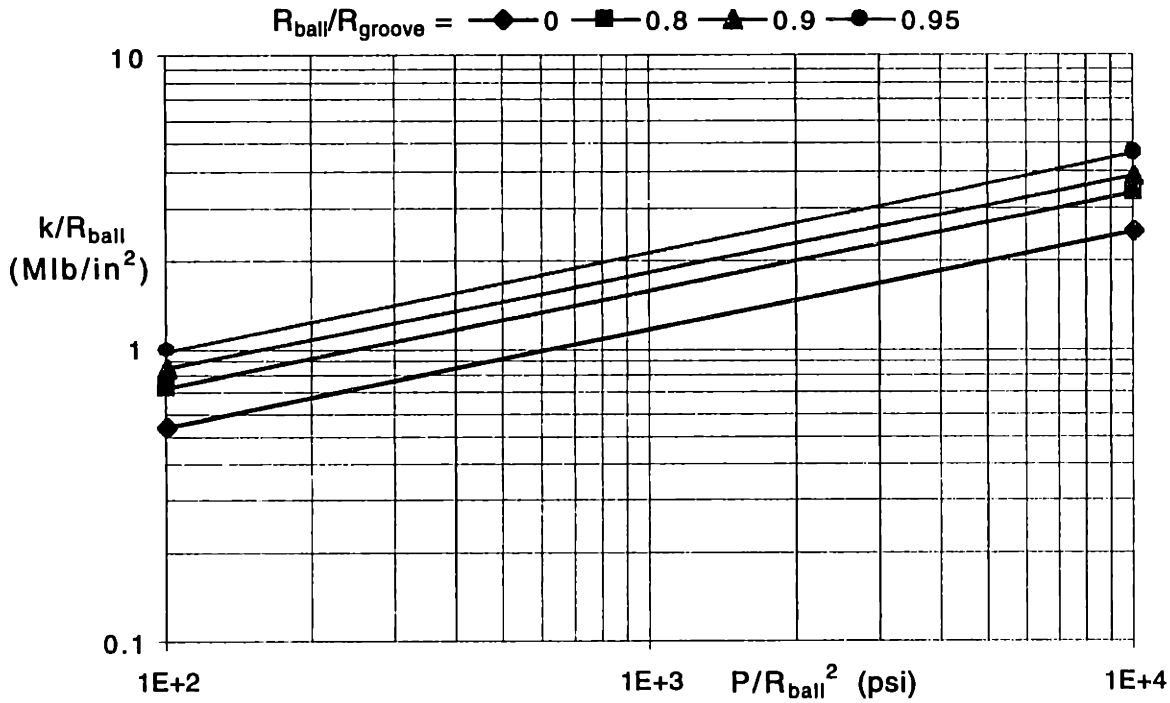


Figure C-3 Normal stiffness k versus the load P for a ball against a cylindrical groove.

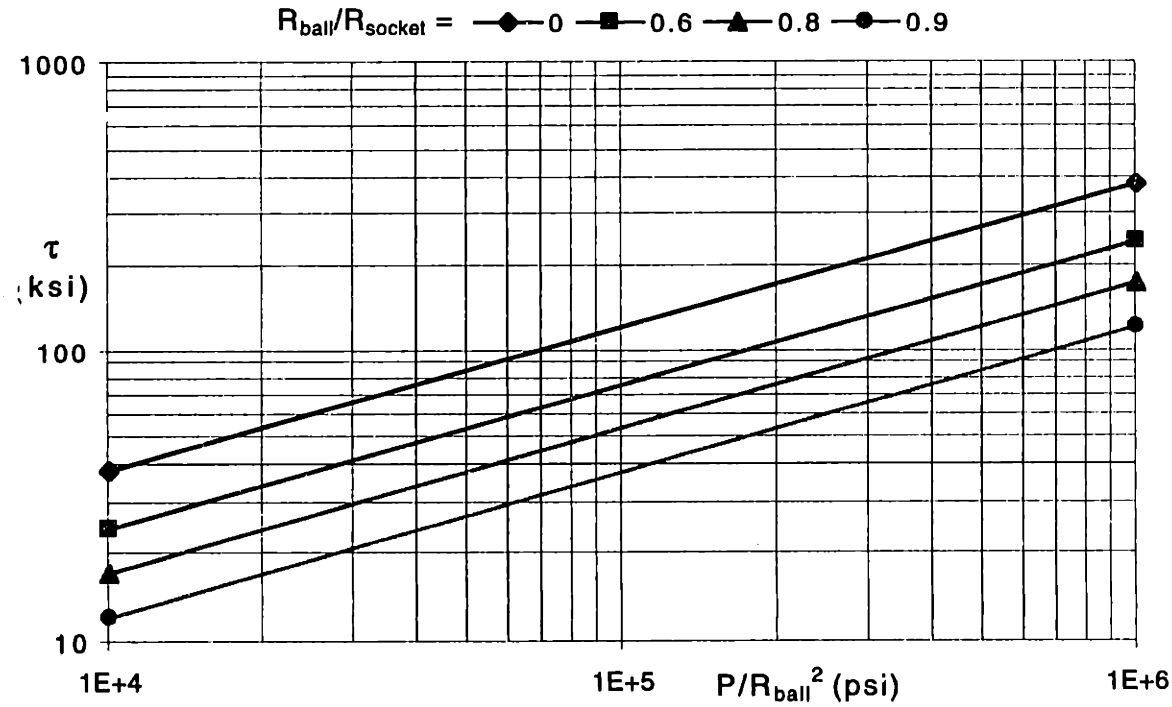


Figure C-4 Shear stress τ versus the axial load P for a ball in a conical socket. Use this graph to determine τ, P or R_{ball} from the other two.

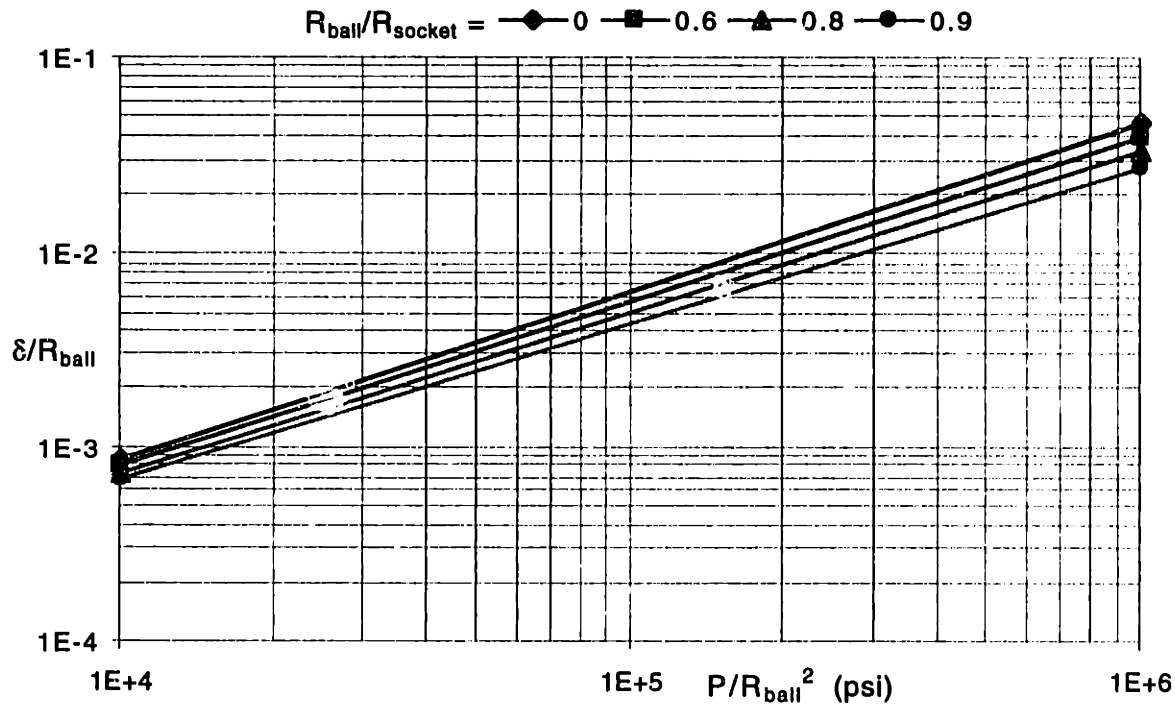


Figure C-5 Axial displacement δ versus the axial load P for a ball in a conical socket.

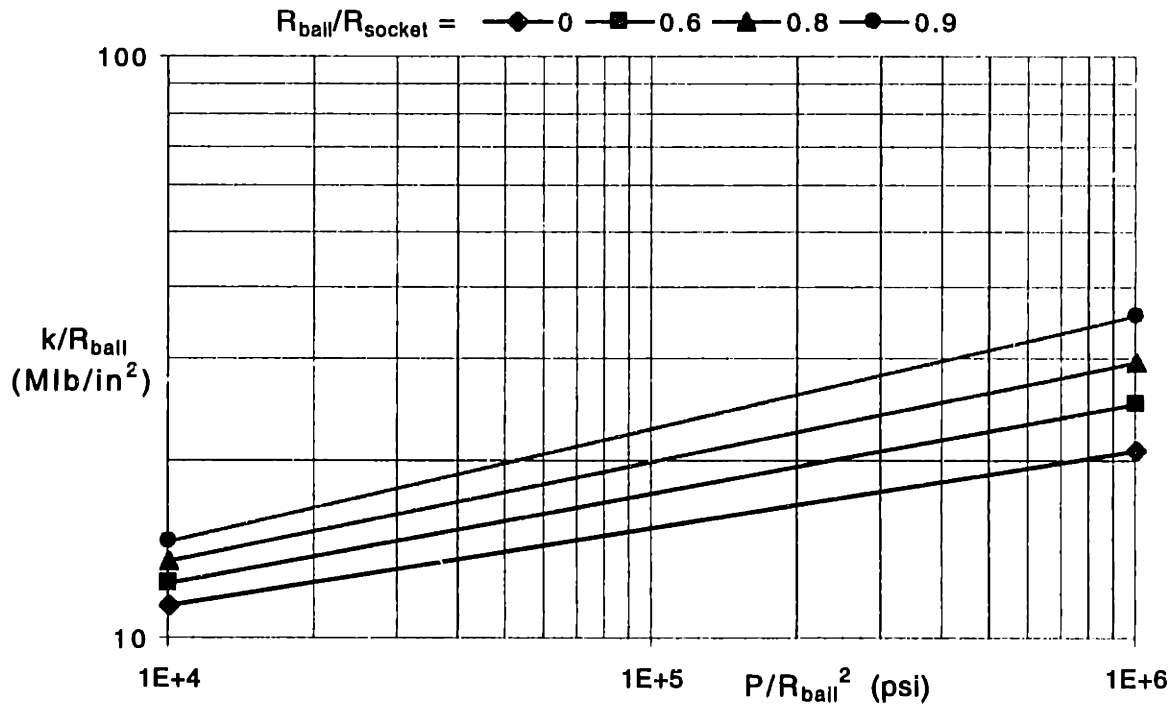


Figure C-6 Axial stiffness k versus the axial load P for a ball in a conical socket.

Applications with unusual geometry or with materials that deviate significantly from the elastic properties of steel will require analysis using the appropriate Hertz equations directly. This is not difficult given the explanations that follow and the Mathcad™ documents that implement these equations for computer analysis.

C.1 Circular Contact

A circular contact area forms when two spheres come into contact or when two cylinders of equal radius contact with 90° crossed axes. Both are special cases of elliptical contact where symmetry simplifies the equations. It is instructive to consider circular contact first to present the main concepts without undo complication.

Equations C.1 through C.8 are the Hertz equations for circular contact. The contact modulus (C.1) expresses the elastic properties of both bodies 1 and 2 effectively as a series combination of springs since stiffness is proportional to the elastic modulus for plain strain. The relative radius (C.2) expresses a summation of curvatures (or inverse radii). Note that curvature is positive for a convex surface and negative for a concave surface. Either radius may be positive or negative so long as the relative radius is positive since it represents an equivalent sphere in contact with a plane. Quite different sets of contacting surfaces behave identically if they have identical contact modulii and relative radii.

Contact Modulus

$$\frac{1}{E_c} = \frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2} \tag{C.1}$$

Relative Radius

$$\frac{1}{R_c} = \frac{1}{R_1} + \frac{1}{R_2} \tag{C.2}$$

The size of the circular contact (C.3) increases weakly with increasing load P and relative radius but decreases weakly with increasing contact modulus. The maximum pressure (C.4) is 1.5 times the mean pressure and occurs at the center of the contact area. Both surfaces experience the same pressure profile, which is hemispherical going to zero of course at $r = c$. Due to hydrostatic stress in the contact region, materials can endure substantially higher pressure than their tensile yield strength. Ductile materials first yield at the point of maximum shear stress (C.5) just below the surface. The allowable shear strength is approximately 58% the allowable tensile strength. Brittle materials fail by fracture at the edge of the contact where the tensile stress is maximum (C.6).

Radius of Contact Circle

$$c = \left(\frac{3PR_c}{4E_c}\right)^{\frac{1}{3}} \tag{C.3}$$

Maximum Pressure

$$p = \frac{3P}{2\pi c^2} = \frac{1}{\pi}\left(\frac{6PE_c^2}{R_c^2}\right)^{\frac{1}{3}} \tag{C.4}$$

Maximum Shear Stress

$$\tau = 0.31p \quad \text{at} \quad z = 0.48c \tag{C.5}$$

Maximum Tensile Stress

$$\sigma = \frac{p}{3}(1-2\nu) \quad \text{at} \quad r = c \tag{C.6}$$

The normal displacement (C.7) refers to the approach of distant points on the two bodies due primarily to deflection in the region of contact. It is obtained by integrating strain from the contact point to distant points in the bodies. The strain goes rapidly to zero thus allowing an improper integral to be bounded. The normal stiffness (C.8) is obtained by differentiating the deflection with respect to load to get compliance, then inverting.

Normal Displacement

$$\delta = \frac{c^2}{R_c} = \left(\frac{3P}{4E_c}\right)^{\frac{2}{3}} \left(\frac{1}{R_c}\right)^{\frac{1}{3}} \tag{C.7}$$

Normal Stiffness

$$k = 2E_c c = \left(6PR_c E_c^{\,2}\right)^{\frac{1}{3}} \tag{C.8}$$

C.2 Elliptical Contact

An elliptical contact area forms when two three-dimensional bodies, each described locally with orthogonal radii of curvature, come into contact. In addition, the orthogonal coordinate system of one body may be rotated relative to the other by an arbitrary angle α. Any radius may be positive (convex) or negative (concave) so long as all three relative radii (C.9) are positive. To first order, R_c represents an equivalent sphere in contact with a plane, while R_a and R_b represent an equivalent toroid in contact with a plane. The contact modulus remains unchanged from circular contact (C.1). Quite different sets of contacting surfaces behave identically if they have identical contact modulii and relative radii.

Relative Radii

$$R_c = \sqrt{R_a R_b} \tag{C.9}$$

$$R_a = \frac{1}{(A+B)-(B-A)} \qquad R_b = \frac{1}{(A+B)+(B-A)}$$

$$A+B = \frac{1}{2}\left(\frac{1}{R_{1xx}} + \frac{1}{R_{1yy}} + \frac{1}{R_{2xx}} + \frac{1}{R_{2yy}}\right)$$

$$B-A = \frac{1}{2}\left\{\left(\frac{1}{R_{1xx}} - \frac{1}{R_{1yy}}\right)^2 + \left(\frac{1}{R_{2xx}} - \frac{1}{R_{2yy}}\right)^2 \cdots\right.$$

$$\left. +2\left(\frac{1}{R_{1xx}} - \frac{1}{R_{1yy}}\right)\left(\frac{1}{R_{2xx}} - \frac{1}{R_{2yy}}\right)\cos(2\alpha)\right\}^{\frac{1}{2}}$$

The approximate expression for the eccentricity of the contact ellipse (C.10) is sufficient for practical geometries; however, the full solution is implemented in Section C.6. The radius of an equivalent circular contact (C.11) contains a correction factor F_1 that gradually decreases from one as the contact becomes more elliptical. The major and minor

radii of the contact ellipse (C.12) follow from the eccentricity and the equivalent radius. The maximum pressure (C.13) differs from circular contact only in that the pressure profile is semiellipsoidal going to zero of course at the edge of the contact ellipse. The maximum shear stress and depth below the surface (C.14) are very similar to circular contact. The equations provided are curve fits to Table 4.1 in [Johnson, 1985]. The radially oriented tensile stress at the major and minor contact radii (C.15) becomes increasingly different from one another (and the tensile stress for circular contact) as the contact becomes more elliptical. The tensile stress at the major radius σ_a is maximum.

Eccentricity of Contact Ellipse

$$e^2 = 1 - \left(\frac{b}{a}\right)^2 \cong 1 - \left(\frac{R_b}{R_a}\right)^{\frac{4}{3}} \tag{C.10}$$

Equivalent Radius of Contact

$$c = \sqrt{ab} = \left(\frac{3PR_c}{4E_c}\right)^{\frac{1}{3}} F_1 \tag{C.11}$$

Major and Minor Contact Radii

$$a = c\left(1-e^2\right)^{-1/4} \qquad b = c\left(1-e^2\right)^{1/4} \tag{C.12}$$

Maximum Pressure

$$p = \frac{3P}{2\pi c^2} = \frac{3P}{2\pi ab} \tag{C.13}$$

Maximum Shear Stress

$$\tau \cong p\left\{0.303 + 0.0855\frac{b}{a} - 0.808\left(\frac{b}{a}\right)^2\right\}$$
$$z \cong b\left\{0.7929 - 0.3207\frac{b}{a}\right\} \tag{C.14}$$

Tensile Stress at a and b

$$\sigma_a = p(1-2\nu)\frac{b}{ae^2}\left\{\frac{1}{e}\tanh^{-1}(e) - 1\right\}$$
$$\sigma_b = p(1-2\nu)\frac{b}{ae^2}\left\{1 - \frac{b}{ae^2}\tan^{-1}\left(\frac{ea}{b}\right)\right\} \tag{C.15}$$

The normal displacement (C.16) and the normal stiffness (C.17) differ from circular contact only by the correction factors F_1 and F_2. Equation C.18 provides curve fits to the exact values as calculated in Section C.6.

Normal Displacement

$$\delta = \frac{c^2}{R_c}\frac{F_2}{F_1^2} = \frac{ab}{R_c}\frac{F_2}{F_1^2} \tag{C.16}$$

Normal Stiffness

$$k = \frac{2E_c c}{F_1 F_2} \tag{C.17}$$

$$F_1 \cong 1-\left[\left(\frac{R_a}{R_b}\right)^{0.0602}-1\right]^{1.456} \qquad F_2 \cong 1-\left[\left(\frac{R_a}{R_b}\right)^{0.0684}-1\right]^{1.531} \tag{C.18}$$

C.3 Line Contact

The contact between parallel cylinders away from end effects is well represented by two-dimensional Hertz theory. Then the contact area is assumed to have constant width $2b$ over the length of contact $2a$. These symbols are chosen to be consistent with elliptical contact; however to be consistent with the references, P is the line load (load per length of contact) rather than the actual load. The contact modulus (C.1) and the relative radius (C.2) remain the same as circular contact. The remaining Hertz equations for line contact are similar to but somewhat different from those of elliptical contact. The half width of contact (C.19) varies faster with load, 1/2 versus 1/3, but the contact area varies slower with load, 1/2 versus 2/3. The maximum pressure (C.20) is somewhat closer to the mean pressure, $4/\pi$ versus 3/2. The maximum shear stress (C.21) is nearly the same ratio to the maximum pressure but occurs deeper. There is no tensile stress for line contact.

Half Width of Contact
$$b=\left(\frac{4PR_c}{\pi E_c}\right)^{\frac{1}{2}} \tag{C.19}$$

Maximum Pressure
$$p=\frac{2P}{\pi b}=\left(\frac{PE_c}{\pi R_c}\right)^{\frac{1}{2}} \tag{C.20}$$

Maximum Shear Stress
$$\tau = 0.30p \quad \text{at} \quad z = 0.78b \tag{C.21}$$

The normal displacement (C.22) and the normal stiffness (C.23) now depend on the distance that the reference points d_1 and d_2 are from the contact point. This occurs because two-dimensional theory only allows the load to spread out in one direction. These expressions are approximate if the elastic properties are different between the two bodies; however, the exact expressions are implemented in Section C.6.

Normal Displacement
$$\delta \cong \frac{P}{\pi E_c}\left\{\ln\left(\frac{4d_1}{b}\right)+\ln\left(\frac{4d_2}{b}\right)-1\right\} \tag{C.22}$$

Normal Stiffness
$$k \cong \frac{\pi E_c 2a}{\ln\left(\frac{4d_1}{b}\right)+\ln\left(\frac{4d_2}{b}\right)-2} \tag{C.23}$$

C.4 Sphere and Cone Contact

The theoretical contact between a sphere and a conical socket forms a circle whose radius depends on the radius of the sphere R and the cone angle θ with respect to the axis. Hertz theory for line contact may be extended to sphere and cone contact rather simply by assuming the line load (C.24) consists of a constant term due to the axial load f_a and a sinusoidal variation around the circle due to the radial load f_r. The maximum and minimum line loads occur at $\phi = 0$ and π radians, respectfully. Then maximum and minimum values of contact width, pressure and so forth are simple to calculate using the Hertz equations for line contact. The axial stiffness (C.25) and the radial stiffness (C.26) come from integrating the local contact stiffness (reflected to the proper angle) around the circle. Since the local contact stiffness does not usually change significantly around the circle, it is adequate to use the mean value to simplify the integration.

Line Load

$$P(\theta,\phi) \cong \frac{1}{2\pi R\cos(\theta)}\left\{\frac{f_a}{\sin(\theta)} + \frac{2\cos(\phi)f_r}{\cos(\theta)}\right\} \tag{C.24}$$

Axial Stiffness

$$k_a \cong \frac{2\pi^2 E_c R\cos(\theta)\sin^2(\theta)}{\ln\left(\frac{4d_1}{b_{max}}\right) + \ln\left(\frac{4d_2}{b_{min}}\right) - 2} \tag{C.25}$$

Radial Stiffness

$$k_r \cong \frac{\pi^2 E_c R\cos^3(\theta)}{\ln\left(\frac{4d_1}{b_{max}}\right) + \ln\left(\frac{4d_2}{b_{min}}\right) - 2} \tag{C.26}$$

C.5 Tangential Loading of an Elliptical Contact

The normal pressure that exists between two three-dimensional bodies in contact has a semiellipsoidal profile that is maximum at the center and zero at the boundary of the contact area. It is reasonable to expect the same profile to exist for surface traction if the two bodies also slide while held in contact. The traction in this case is related to the normal pressure through the coefficient of friction. If it were possible to perfectly adhere the surfaces in some way, then a tangential force would cause the reciprocal profile to develop over the contact. In this case the traction is minimum at the center and rises to infinity along the edge. This is plausible behavior since the joint is equivalent to a very deep, sharp crack. The friction connection obviously cannot support infinite traction where the normal force is zero. A region of slip forms at the edge of contact and extends toward the center until the

normal pressure is sufficient to carry the traction by friction.[I] As the tangential force increases, the slip region encroaches further into the adhered region until nothing remains to prevent sliding, and the tangential stiffness goes to zero. The useful tangential constraint device typically operates well below the point of incipient sliding.

This theory has practical applications in the design of friction drives for precision machines and also in kinematic couplings where frictional constraints act to stiffen the coupling and the flexible modes of the supported object. Slip between two elastic bodies is a dynamic process that requires something to change. This section treats the two most practical possibilities: when the contact is stationary and the tangential force varies, or when the force is constant and the contact moves across rolling bodies.[II] For example, slip that occurs from an oscillating tangential force results in a hysteresis loop in the force-displacement curve. Slip that occurs between rolling bodies in contact results in a small differential velocity called creep in the literature. These nonideal behaviors are usually very small and can be estimated with reasonable accuracy using the equations provided.

C.5.1 Stationary Elliptical Contact, Variable Tangential Force

A tangential force applied to a stationary, elliptical contact produces a relative tangential displacement governed principally by elastic deformation in the contact. Typically small inelastic behavior results from slip that always accompanies the elastic deformation. In the normal direction, all the Hertz equations for elliptical contact still apply. Since traction at the surface produces shear stress in the material, it is convenient to define the contact shear modulus (C.27) to simplify the main equations. The transition from an adhered region to a slip region occurs theoretically on an ellipse that is smaller in size but with the same proportions as the contact ellipse. The transition ellipse (C.28) shrinks in size as the tangential force increases until the point of incipient sliding occurs when $T = \mu P$. The subscript in the equation indicates that the traction profile may have more than one transition depending on the history of the tangential force. The tangential displacement between distant points (C.29) is valid when there is one transition in the traction profile. This occurs if the contact initially has zero traction and the tangential force increases monotonically. Corresponding to this condition is the tangential stiffness (C.30), which applies only when the force increases. A decreasing force causes the entire contact to momentarily adhere then establish a second transition ellipse outside the first.

Contact Shear Modulus

$$\frac{1}{G_c} = \frac{2-\nu_1}{G_1} + \frac{2-\nu_2}{G_2} \tag{C.27}$$

[I] It is customary to distinguish between the terms *slip* and *sliding*. Slip refers to small relative displacement resulting from differing strain fields in the two contacting surfaces. Sliding is arbitrarily large movement between two contacting surfaces.

[II] [Mindlin and Deresiewicz, 1953] treat combinations of varying normal and tangential force.

Transition Ellipse
$$\frac{a_1}{a} = \frac{b_1}{b} = \left(1 - \frac{T_1}{\mu P}\right)^{\frac{1}{3}} \tag{C.28}$$

Tangential Displacement
$$\delta = \frac{3\mu P}{16 a G_c}\left\{1 - \left(1 - \frac{T}{\mu P}\right)^{\frac{2}{3}}\right\}\Phi \tag{C.29}$$

Tangential Stiffness
$$k = 8 a G_c \left(1 - \frac{T}{\mu P}\right)^{\frac{1}{3}} \frac{1}{\Phi} \tag{C.30}$$

The equation for the second transition ellipse (C.31) depends on the maximum tangential force T_1 and the relaxed or reversed tangential force T_2. If the magnitude of T_2 becomes equal to or greater than T_1, then the second transition ellipse encroaches on the first and effectively eliminates its record. It is possible in theory to have several transitions if there are several reversals and each is smaller than the previous one. The tangential displacement for a decreasing tangential force (C.32) has an added term to account for the second transition. Setting T to zero gives the half width of the hysteresis loop between $\pm T_1$. An approximate expression of the hysteresis half width (C.33) is primarily quadratic in T_1. The tangential stiffness for a decreasing tangential force (C.34) is momentarily maximum, as expected, then it decreases as the tangential force relaxes. Notice that all the equations for displacement and stiffness share the correction factor Φ. This factor accounts for the ellipticity of the contact and whether the tangential force is parallel to the major radius a or the minor radius b. The approximate correction factor (C.35) provides good agreement with the more complicated exact expressions plotted in Figure C-7. The exact expressions are used in Section C.6.

Transition Ellipse
$$\frac{a_2}{a} = \frac{b_2}{b} = \left(1 - \frac{T_1 - T_2}{2\mu P}\right)^{\frac{1}{3}} \tag{C.31}$$

Tangential Displacement
$$\delta_d = \frac{3\mu P}{16 a G_c}\left\{2\left(1 - \frac{T_1 - T}{2\mu P}\right)^{\frac{2}{3}} - \left(1 - \frac{T_1}{\mu P}\right)^{\frac{2}{3}} - 1\right\}\Phi \tag{C.32}$$

Hysteresis Half Width
$$\delta_h \cong \frac{3\mu P}{16 a G_c}\left\{\frac{1}{18}\left(\frac{T_1}{\mu P}\right)^2 + \frac{1}{27}\left(\frac{T_1}{\mu P}\right)^3\right\}\Phi \tag{C.33}$$

Tangential Stiffness
$$k_d = 8 a G_c \left(1 - \frac{T_1 - T}{2\mu P}\right)^{\frac{1}{3}} \frac{1}{\Phi} \tag{C.34}$$

Correction Factor

$$\Phi \equiv \begin{cases} 1+(1.4-0.8\nu)\log\left(\frac{a}{b}\right) & T \,//\, b \\ 1 & a = b \\ 1+(1.4+0.8\nu)\log\left(\frac{a}{b}\right) & T \,//\, a \end{cases} \qquad \text{(C.35)}$$

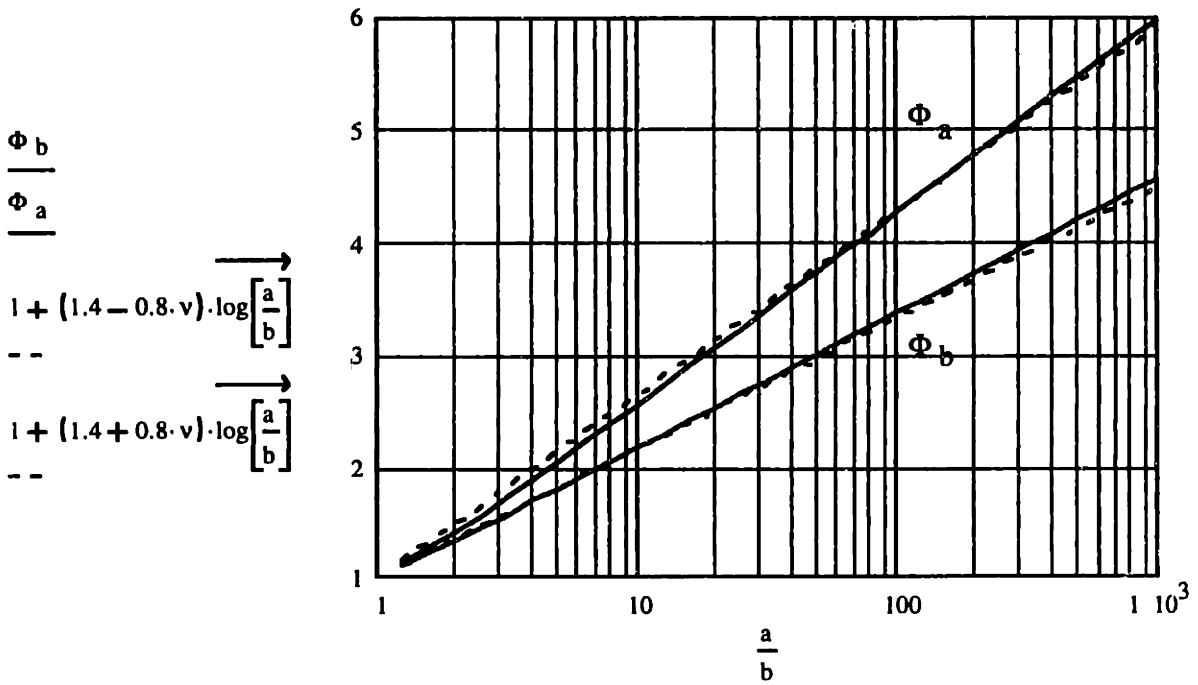


Figure C-7 Correction factors for elliptical contact calculated for ν = 0.3. The tangential force may be parallel with the major radius *a* or the minor radius *b*.

The following references present this theory in greater detail [Mindlin, 1949], [Mindlin and Deresiewicz, 1953], [Deresiewicz, 1957]. One reference [Mindlin, et al., 1951] provides experimental data that agrees reasonably well with theory except for the energy dissipation for low-amplitude cycles. The theory predicts that the energy loss is cubic in force amplitude whereas the experiment shows quadratic behavior.

C.5.2 Rolling Circular Contact, Constant Tangential Force

The equations for slip in a rolling circular contact due to a constant tangential force are very similar to those for stationary contact. This is understandable since the theory developed by [Johnson, 1958] assumes that the transition between the adhered region and the slip region is the same proportion described in Equation C.28. He further assumes that the adhered region is tangent to the leading edge of contact whether the tangential force is parallel to or perpendicular to the direction of travel. This is intuitive behavior since the material entering the contact area is in a relaxed state and acquires increasing traction until it exceeds the threshold for slip towards the trailing edge of contact. A numerical approach developed by [Kalker, 1966] predicts a lemon-shaped adhered region bounded by the leading edge of the

contact circle and an arc shifted from the trailing edge. The slip region then appears as a crescent. Although the numerical approach closely matches observed results, Johnson's theory yields simple equations and reasonable estimates of creep (the ratio of the slip velocity to the rolling-contact velocity).

The direction of the tangential force governs the direction of slip but it plays a minor role in the magnitude of the slip ratio. Johnson uses the term *longitudinal* to indicate that the tangential force is in the direction of rolling along the x-axis. A *transverse* tangential force is in the y-direction. As before, it is convenient to group the elastic properties of both bodies into one. The longitudinal and transverse shear modulii (C.36 and C.37) differ only about ±10 percent from the contact shear modulus (C.27). Given the approximate nature of this theory, it would be adequate to use the contact shear modulus (C.27) for any direction of tangential force. Other than accounting for differences in directions, the longitudinal and transverse slip ratios (C.38 and C.39) are identical.

Longitudinal Shear Modulus

$$\frac{1}{G_x} = \frac{2-1.5\nu_1}{G_1} + \frac{2-1.5\nu_2}{G_2} \tag{C.36}$$

Transverse Shear Modulus

$$\frac{1}{G_y} = \frac{2-0.5\nu_1}{G_1} + \frac{2-0.5\nu_2}{G_2} \tag{C.37}$$

Longitudinal Slip Ratio

$$\xi_x = \frac{3\mu P}{16c^2 G_x}\left\{1-\left(1-\frac{T_x}{\mu P}\right)^{\frac{1}{3}}\right\} \tag{C.38}$$

Transverse Slip Ratio

$$\xi_y = \frac{3\mu P}{16c^2 G_y}\left\{1-\left(1-\frac{T_y}{\mu P}\right)^{\frac{1}{3}}\right\} \tag{C.39}$$

Although this theory was developed for circular contact, its similarity to stationary contact suggests that it could be extended to elliptical contact using the same correction factor Φ. The use of Φ_a or Φ_b and particularly whether to use $c^2 = a\,b$ or a^2 in the denominator is strictly conjecture. The recommendation is to use $a\,b$ when rolling is along the minor axis, and conversely to use a^2 when rolling is along the major axis. The rationale is that one factor of a belongs with Φ and that the slip ratio scales inversely with the contact width in the direction of rolling, either a or b. The choice of Φ_a or Φ_b depends solely on whether the tangential force is along the major or minor axis. The use of longitudinal or transverse shear modulus is optional.

Experimental work by [Johnson, 1958] shows reasonable agreement with his theory for rolling contact of a sphere on a plane. The agreement is good for low levels of slip but it underestimates creep by 20 to 30 percent for $T > \mu P/2$.

C.6 Mathcad™ Documents for Contact Mechanics

This program uses Hertz equations to calculate deflection and stress between two elastic spheres, including the effects of tangential force if it exists. Ref. K.L. Johnson, 1958, 1985, Mindlin and Deresiewicz, 1953.

Elastic modulus: $E_1 := 210 \cdot GPa$ $E_2 := 210 \cdot GPa$

Poisson ratio: $\nu_1 := 0.29$ $\nu_2 := 0.29$

Radius of curvature: $R_1 := 5 \cdot mm$ $R_2 := -6 \cdot mm$

Normal and tangential forces: $P := 100 \cdot kgf$ $T := 4 \cdot kgf$

Coefficient of friction: $\mu := 0.1$ $\mu \cdot P = 10 \cdot kgf$

Calculate the contact modulii.

$$E_c := \left[\frac{1 - \nu_1^2}{E_1} + \frac{1 - \nu_2^2}{E_2} \right]^{-1} \qquad G_c := \frac{1}{2} \left[\frac{2 + \nu_1 - \nu_1^2}{E_1} + \frac{2 + \nu_2 - \nu_2^2}{E_2} \right]^{-1}$$

$E_c = 114.641 \cdot GPa$ $G_c = 23.8 \ GPa$

Calculate the relative radius equivalent to a ball on a flat.

$$R_c := \left(R_1^{-1} + R_2^{-1} \right)^{-1} \qquad R_c = 30 \ mm$$

Calculate the half width of the circular contact.

$$c := \left(\frac{3 \cdot P \cdot R_c}{4 \cdot E_c} \right)^{\frac{1}{3}} \qquad c = 0.577 \cdot mm \qquad 2 \cdot c = 1.155 \ mm$$

Calculate the max. contact pressure, the max. shear stress and its z location below the contact.

$$p := \frac{3 \cdot P}{2 \cdot \pi \cdot c^2} \qquad p = 203.721 \cdot ksi \qquad p = 143.23 \cdot HBn$$

$\tau := 0.31 \cdot p$ $\tau = 63.153 \ ksi$ Equivalent tensile:

$z := 0.48 \cdot c$ $z = 0.277 \cdot mm$ $\tau \cdot \sqrt{3} = 109.385 \cdot ksi$

Calculate the maximum tensile stress (radial) located at the edge of the contact zone.

$$\sigma := \frac{1}{3}\cdot\left(1 - 2\cdot\nu_1\right)\cdot p \qquad \sigma = 28.521\cdot ksi$$

Calculate the normal and tangential displacements.

$$\delta_n := \frac{c^2}{R_c} \qquad \delta_t := \frac{3\cdot\mu\cdot P}{16\cdot c\cdot G_c}\cdot\left[1 - \left(1 - \frac{T}{\mu\cdot P}\right)^{\frac{2}{3}}\right]$$

$$\delta_n = 11.112\cdot\mu m \qquad \delta_t = 0.386\cdot\mu m$$

Calculate the normal and tangential stiffness.

$$k_n := 2\cdot E_c\cdot c \qquad k_t := 8\cdot c\cdot G_c\cdot\left(1 - \frac{T}{\mu\cdot P}\right)^{\frac{1}{3}}$$

$$k_n = 132.381\cdot\frac{newton}{\mu m} \qquad k_t = 92.719\cdot\frac{newton}{\mu m}$$

Calculate the half width of the hysteresis loop due to a cyclic tangential force.

$$\delta_h := \frac{3\cdot\mu\cdot P}{16\cdot c\cdot G_c}\cdot\left[2\cdot\left(1 - \frac{T}{2\cdot\mu\cdot P}\right)^{\frac{2}{3}} - \left(1 - \frac{T}{\mu\cdot P}\right)^{\frac{2}{3}} - 1\right] \qquad \delta_h = 0.016\cdot\mu m$$

Calculate the slip ratio for a rolling contact due to a steady tangential force.

$$\xi := \frac{3\cdot\mu\cdot P}{16\cdot c^2\cdot G_c}\cdot\left[1 - \left(1 - \frac{T}{\mu\cdot P}\right)^{\frac{1}{3}}\right] \qquad \xi = 3.629\cdot 10^{-4}$$

Define units: $GPa \equiv Pa\cdot 10^9$ $\quad kN \equiv newton\cdot 10^3$ $\quad HBn \equiv \frac{kgf}{mm^2}$

$\mu m \equiv m\cdot 10^{-6}$ $\quad mils \equiv in\cdot 10^{-3}$

$ksi \equiv psi\cdot 10^3$ $\quad \mu in \equiv in\cdot 10^{-6}$

This program uses Hertz equations to calculate deflection and stress between two elastic toroids, including the effects of tangential force if it exists. The angular orientation between the axes for body 1 and body 2 is given by α. Ref. K.L. Johnson, 1958, 1985, Mindlin and Deresiewicz, 1953.

Elastic modulus: $E_1 := 210 \cdot GPa$ $E_2 := 210 \cdot GPa$

Poisson ratio: $\nu_1 := 0.29$ $\nu_2 := 0.29$

Principle radius of curvature: $R_{1xx} := 10 \cdot mm$ $R_{2xx} := 10^6 \cdot m$

Principle radius of curvature: $R_{1yy} := 10^6 \cdot m$ $R_{2yy} := 500 \cdot mm$

Angle between 1&2 axes: $\alpha := 0 \cdot deg$

Normal and tangential forces: $P := 200 \cdot kgf$ $T := 5 \cdot kgf$

Coefficient of friction: $\mu := 0.1$ $\mu \cdot P = 20 \cdot kgf$

Calculate the contact modulii.

$$E_c := \left[\frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2}\right]^{-1} \qquad G_c := \frac{1}{2}\cdot\left[\frac{2+\nu_1-\nu_1^2}{E_1} + \frac{2+\nu_2-\nu_2^2}{E_2}\right]^{-1}$$

$E_c = 114.641 \cdot GPa$ $G_c = 23.8\ GPa$

Calculate the relative radii equivalent to a toroid on a flat.

$$A\_plus\_B := \frac{1}{2}\left(\frac{1}{R_{1xx}} + \frac{1}{R_{1yy}} + \frac{1}{R_{2xx}} + \frac{1}{R_{2yy}}\right)$$

$$B\_minus\_A := \frac{1}{2}\cdot\sqrt{\left(\frac{1}{R_{1xx}} - \frac{1}{R_{1yy}}\right)^2 + \left(\frac{1}{R_{2xx}} - \frac{1}{R_{2yy}}\right)^2 \ldots + 2\cdot\left(\frac{1}{R_{1xx}} - \frac{1}{R_{1yy}}\right)\cdot\left(\frac{1}{R_{2xx}} - \frac{1}{R_{2yy}}\right)\cdot\cos(2\cdot\alpha)}$$

$R_a := (A\_plus\_B - B\_minus\_A)^{-1}$ $R_a = 500 \cdot mm$

$R_b := (A\_plus\_B + B\_minus\_A)^{-1}$ $R_b = 10 \cdot mm$

Calculate the relative radius equivalent to a ball on a flat.

$R_c := \sqrt{R_a \cdot R_b}$ $R_c = 70.711\ mm$ $\frac{R_a}{R_b} = 50$

Define complete elliptic integrals required for remaining calculations.

$$E(k) := \int_0^{\frac{\pi}{2}} \left(1 - k^2 \cdot \sin(\phi)^2\right)^{\frac{1}{2}} d\phi \qquad K(k) := \int_0^{\frac{\pi}{2}} \left(1 - k^2 \cdot \sin(\phi)^2\right)^{\frac{-1}{2}} d\phi$$

Solve for the eccentricity of the elliptical contact.

Approximate eccentricity: $e := \sqrt{1 - \left(\frac{R_b}{R_a}\right)^{\frac{4}{3}}}$ $e = 0.997$

$$e := \text{root}\left[\frac{R_a}{R_b} \cdot (K(e) - E(e)) + K(e) - \frac{E(e)}{1 - e^2}, e\right] \qquad e = 0.997$$

Calculate correction factors.

$$F_1 := \text{if}\left[e^2 = 0, 1, \left(1 - e^2\right)^{\frac{1}{4}} \cdot \left(\frac{R_a}{R_b}\right)^{\frac{1}{6}} \cdot \left[\frac{4 \cdot (K(e) - E(e))}{\pi \cdot e^2}\right]^{\frac{1}{3}}\right] \qquad F_1 = 0.853$$

$$F_2 := \frac{2}{\pi} \cdot \left(1 - e^2\right)^{\frac{1}{4}} \cdot \frac{K(e)}{F_1} \qquad \nu := \frac{1}{2} \cdot \left(\nu_1 + \nu_2\right) \qquad F_2 = 0.834$$

$$\Phi_a := \frac{4}{\pi \left(2 - \nu\right)} \cdot \left[(1 - \nu) \cdot K(e) + \frac{\nu}{e^2} \cdot (K(e) - E(e))\right] \qquad \Phi_a = 2.676$$

$$\Phi_b := \frac{4}{\pi \left(2 - \nu\right)} \cdot \left[K(e) - \frac{\nu}{e^2} (K(e) - E(e))\right] \qquad \Phi_b = 2.266$$

Calculate the half widths a and b of the elliptical contact and the equivalent half width c.

$$c := \left(\frac{3 \cdot P \cdot R_c}{4 \cdot E_c}\right)^{\frac{1}{3}} \cdot F_1 \qquad a := c \cdot \left(1 - e^2\right)^{-\frac{1}{4}} \qquad b := c \cdot \left(1 - e^2\right)^{\frac{1}{4}}$$

$c = 0.826 \cdot mm$ $a = 2.869 \cdot mm$ $b = 0.238$ mm

$2 \cdot c = 1.651$ mm $2 \cdot a = 5.738$ mm $2 \cdot b = 0.475$ mm

Calculate the maximum contact pressure located at the center of the contact point.

$$p := \frac{3 \cdot P}{2 \cdot \pi \cdot a \cdot b} \qquad p = 199.218 \cdot ksi \qquad p = 140.064 \cdot HBn$$

Calculate the maximum tensile stress (radial) and compressive stress (tangential) located at the edge of the contact zone.

$$\sigma_a := p \cdot \left(1 - \nu_1 - \nu_2\right) \cdot \frac{b}{a \cdot e^2} \cdot \left(\frac{1}{e} \cdot \operatorname{atanh}(e) - 1\right) \qquad \sigma_a = 15.307 \cdot ksi$$

$$\sigma_b := p \cdot \left(1 - \nu_1 - \nu_2\right) \cdot \frac{b}{a \cdot e^2} \cdot \left(1 - \frac{b}{a \cdot e} \operatorname{atan}\left(\frac{a \cdot e}{b}\right)\right) \qquad \sigma_b = 6.116 \cdot ksi$$

Calculate the maximum shear stress and its z-location below the contact point using curve fits to Table 4.1 in [Johnson].

$$\tau := p \cdot \left[0.303 + 0.0855 \cdot \sqrt{1 - e^2} - 0.0808 \cdot \left(1 - e^2\right)\right] \qquad \frac{\tau}{p} = 0.31$$

$\tau = 61.664 \cdot ksi$ Equivalent tensile: $\tau \cdot \sqrt{3} = 106.805 \cdot ksi$

$$z := b \cdot \left(0.7929 - 0.3207 \cdot \sqrt{1 - e^2}\right) \qquad z = 0.182 \ mm$$

Calculate the normal and tangential displacements. Note, use the correction factor Φa or Φb, respectively, when the tangential force is along the major radius a or the minor radius b.

$$\delta_n := \frac{a \cdot b}{R_c} \cdot \frac{F_2}{F_1^2} \qquad \delta_t := \frac{3 \cdot \mu \cdot P}{16 \cdot a \cdot G_c} \cdot \left[1 - \left(1 - \frac{T}{\mu \cdot P}\right)^{\frac{2}{3}}\right]$$

$\delta_n = 11.054 \cdot \mu m$ $\delta_t \cdot \Phi_a = 0.252 \cdot \mu m$ $\delta_t \cdot \Phi_b = 0.213 \cdot \mu m$

Calculate the normal and tangential stiffnesses.

$$k_n := \frac{2 \cdot E_c \cdot c}{F_1 \cdot F_2} \qquad k_t := 8 \cdot a \cdot G_c \cdot \left(1 - \frac{T}{\mu \cdot P}\right)^{\frac{1}{3}}$$

$$k_n = 266.149 \ \frac{newton}{\mu m} \qquad \frac{k_t}{\Phi_a} = 185.425 \ \frac{newton}{\mu m} \qquad \frac{k_t}{\Phi_b} = 218.964 \ \frac{newton}{\mu m}$$

Calculate the half width of the hysteresis loop due to a cyclic tangential force.

$$\delta_h := \frac{3\cdot\mu\cdot P}{16\cdot c\cdot G_c}\cdot\left[2\cdot\left(1-\frac{T}{2\cdot\mu\cdot P}\right)^{\frac{2}{3}}-\left(1-\frac{T}{\mu\cdot P}\right)^{\frac{2}{3}}-1\right]$$

$$\delta_h\cdot\Phi_a = 0.021\cdot\mu m$$

$$\delta_h\cdot\Phi_b = 0.018\cdot\mu m$$

Calculate the slip ratio for a rolling contact due to a steady tangential force. Note, use a^2 or c^2, respectively, in the denominator when the rolling direction is along the major radius a or the minor radius b.

$$\xi := \frac{3\cdot\mu\cdot P}{16\cdot c^2\cdot G_c}\cdot\left[1-\left(1-\frac{T}{\mu\cdot P}\right)^{\frac{1}{3}}\right]$$

$$\xi\cdot\Phi_a = 5.546\cdot 10^{-4}$$

$$\xi\cdot\Phi_b = 4.697\cdot 10^{-4}$$

Define units:

$GPa \equiv Pa\cdot 10^9$ $\quad kN \equiv newton\cdot 10^3$ $\quad HBn \equiv \frac{kgf}{mm^2}$

$\mu m \equiv m\cdot 10^{-6}$ $\quad mils \equiv in\cdot 10^{-3}$

$ksi \equiv psi\cdot 10^3$ $\quad \mu in \equiv in\cdot 10^{-6}$

This program uses Hertz equations to calculate deflection and stress between two elastic cylinders. Ref. K. Johnson, 1985.

Elastic modulus: $E_1 := 210 \cdot GPa$ $E_2 := 210 \cdot GPa$

Poisson ratio: $\nu_1 := 0.29$ $\nu_2 := 0.29$

Radius of curvature: $R_1 := \frac{10}{2} \cdot mm$ $R_2 := 10^6 \cdot m$

Depth of reference pt for deflection: $d_1 := 10 \cdot mm$ $d_2 := d_1$

Load f and half length of contact a: $f := 200 \cdot kgf$ $a := \frac{10}{2} \cdot mm$

Calculate the line load. $P := \frac{f}{2 \cdot a}$ $P = 20 \cdot \frac{kgf}{mm}$

Calculate the contact modulus.

$$E_c := \left[\frac{1 - \nu_1^2}{E_1} + \frac{1 - \nu_2^2}{E_2} \right]^{-1} \qquad E_c = 114.641 \ GPa$$

Calculate the relative radius equivalent to a cylinder on a flat.

$$R_c := \left(R_1^{-1} + R_2^{-1} \right)^{-1} \qquad R_c = 5 \cdot mm$$

Calculate the half width of contact.

$$b := \left[\frac{4 \cdot P \cdot R_c}{\pi \cdot E_c} \right]^{\frac{1}{2}} \qquad b = 0.104 \ mm \qquad 2 \cdot b = 0.209 \ mm$$

$\frac{a}{b} = 47.91$ $\frac{b}{a} = 0.021$

Calculate the max. contact pressure, the max shear stress and its z location below the contact.

$p := \frac{2 \cdot P}{\pi \cdot b}$ $p = 173.527 \cdot ksi$ $p = 122.002 \cdot HBn$

$\tau := 0.3 \cdot p$ $\tau = 52.058 \cdot ksi$ Equivalent tensile:

$z := 0.78 \cdot b$ $z = 0.081 \cdot mm$ $\tau \cdot \sqrt{3} = 90.167 \cdot ksi$

Calculate the normal displacement of each cylinder (approach of distant points).

$$\delta_1 := P \cdot \frac{1 - \nu_1^2}{\pi \cdot E_1} \cdot \left(2 \cdot \ln\left(\frac{4 \cdot d_1}{b}\right) - 1\right) \qquad \delta_1 = 2.967 \ \mu m$$

$$\delta_2 := P \cdot \frac{1 - \nu_2^2}{\pi \cdot E_2} \cdot \left(2 \cdot \ln\left(\frac{4 \cdot d_2}{b}\right) - 1\right) \qquad \delta_2 = 2.967 \cdot \mu m$$

$$\delta_1 + \delta_2 = 5.935 \ \mu m$$

Calculate the normal stiffness.

$$k := \pi \cdot 2 \cdot a \cdot \left[\frac{1 - (\nu_1)^2}{E_1} \left(2 \cdot \ln\left(\frac{4 \cdot d_1}{b}\right) - 1\right) + \frac{1 - (\nu_2)^2}{E_2} \cdot \left(2 \cdot \ln\left(\frac{4 \cdot d_2}{b}\right) - 1\right) - \frac{1}{E_c}\right]^{-1}$$

$$k = 363.885 \cdot \frac{newton}{\mu m} \qquad k = 2.078 \cdot \frac{lbf}{\mu in}$$

Define units:

$GPa \equiv Pa \cdot 10^9$ $\quad kN \equiv newton \cdot 10^3$ $\quad HBn \equiv \frac{kgf}{mm^2}$

$\mu m \equiv m \cdot 10^{-6}$ $\quad mils \equiv in \cdot 10^{-3}$

$ksi \equiv psi \cdot 10^3$ $\quad \mu in \equiv in \cdot 10^{-6}$

This program uses Hertz equations to calculate deflection and stress for a ball (1) in a cone (2). Based on the Hertz equations for cylinders.

Elastic modulus: $E_1 := 210 \cdot GPa$ $E_2 := 210 \cdot GPa$

Poisson ratio: $\nu_1 := 0.29$ $\nu_2 := 0.29$

Radius of curvature: $R_1 := \frac{10}{2} \cdot mm$ $R_2 := 10^6 \cdot m$

Axial and radial loads: $f_a := 200 \cdot kgf$ $f_r := 80 \cdot kgf$

Depth of ref. pt for deflection: $d_1 := R_1$ $d_2 := d_1 \cdot 10$

Cone angle from the center line: $\theta := 45 \cdot deg$

Calculate the contact modulus.

$$E_c := \left[\frac{1 - \nu_1^2}{E_1} + \frac{1 - \nu_2^2}{E_2} \right]^{-1} \qquad E_c = 114.641 \cdot GPa$$

Calculate the relative radius equivalent to a cylinder on a flat.

$$R_c := \left(R_1^{-1} + R_2^{-1} \right)^{-1} \qquad R_c = 5 \ mm$$

Define the maximum and minimum line loads due to the axial and radial loads (assuming a sinusoidal asimuthal variation).

$$P_{max}(\theta) := \frac{1}{2 \cdot \pi \cdot R_1 \cdot \cos(\theta)} \cdot \left(\frac{f_a}{\sin(\theta)} + \frac{2 \cdot f_r}{\cos(\theta)} \right) \qquad P_{max}(\theta) = 22.918 \cdot \frac{kgf}{mm}$$

$$P_{min}(\theta) := \frac{1}{2 \cdot \pi \cdot R_1 \cdot \cos(\theta)} \cdot \left(\frac{f_a}{\sin(\theta)} - \frac{2 \cdot f_r}{\cos(\theta)} \right) \qquad P_{min}(\theta) = 2.546 \cdot \frac{kgf}{mm}$$

Calculate Pmax and Pmin for the optimal cone angle.

Given $P_{max}(\theta) = 0 \cdot \frac{kgf}{mm}$ $\theta > 0$ $P_{min}(\theta) > 0 \cdot \frac{kgf}{mm}$ $\theta_{op} := \text{Minerr}(\theta)$

$\theta_{op} = 34.338 \cdot deg$ $P_{max}(\theta_{op}) = 21.138 \cdot \frac{kgf}{mm}$ $P_{min}(\theta_{op}) = 6.198 \ \frac{kgf}{mm}$

Calculate the maximum and minimum half widths of contact.

$$b_{max} := \sqrt{\frac{4 \cdot P_{max}(\theta) \cdot R_c}{\pi \cdot E_c}} \qquad b_{min} := \sqrt{\frac{4 \cdot P_{min}(\theta) \cdot R_c}{\pi \cdot E_c}}$$

$b_{max} = 0.112 \cdot mm$ $\qquad$ $b_{min} = 0.037 \cdot mm$

Calculate the maximum contact pressure, the maximum shear stress and its location.

$$p := \frac{2 \cdot P_{max}(\theta)}{\pi \cdot b_{max}}$$

$p = 185.756 \cdot ksi$ $\qquad$ $p = 130.6$ HBn

$\tau := 0.3 \cdot p$ $\qquad$ $\tau = 55.727 \cdot ksi$ $\qquad$ Equivalent tensile:

$z := 0.78 \cdot b_{max}$ $\qquad$ $z = 0.087 \cdot mm$ $\qquad$ $\sqrt{3} \cdot \tau = 96.522 \cdot ksi$

Calculate the local deflection of the ball and cone normal to the cone angle.

$$\delta_{max} := \frac{P_{max}(\theta)}{\pi \cdot E_c} \cdot \left(\ln\left(\frac{4 \cdot d_1}{b_{max}}\right) + \ln\left(\frac{4 \cdot d_2}{b_{max}}\right) - 1 \right) \qquad \delta_{max} = 7.287 \ \mu m$$

$$\delta_{min} := \frac{P_{min}(\theta)}{\pi \cdot E_c} \cdot \left(\ln\left(\frac{4 \cdot d_1}{b_{min}}\right) + \ln\left(\frac{4 \cdot d_2}{b_{min}}\right) - 1 \right) \qquad \delta_{min} = 0.962 \cdot \mu m$$

Calculate the axial and radial stiffnesses.

$$k_a := \frac{2 \cdot \pi^2 \cdot E_c \cdot R_1 \cdot \cos(\theta) \cdot \sin(\theta)^2}{\ln\left(\frac{4 \cdot d_1}{b_{max}}\right) + \ln\left(\frac{4 \cdot d_2}{b_{min}}\right) - 2} \qquad k_a = 1.94 \cdot \frac{lbf}{\mu in}$$

$$k_r := \frac{\pi^2 \cdot E_c \cdot R_1 \cdot \cos(\theta)^3}{\ln\left(\frac{4 \cdot d_1}{b_{max}}\right) + \ln\left(\frac{4 \cdot d_2}{b_{min}}\right) - 2} \qquad k_r = 0.97 \cdot \frac{lbf}{\mu in}$$

Define units: $GPa \equiv Pa \cdot 10^9$ $\qquad$ $kN \equiv newton \cdot 10^3$ $\qquad$ $HBn \equiv \frac{kgf}{mm^2}$

$\mu m \equiv m \cdot 10^{-6}$ $\qquad$ $mils \equiv in \cdot 10^{-3}$

$ksi \equiv psi \cdot 10^3$ $\qquad$ $\mu in \equiv in \cdot 10^{-6}$

intentionally blank

D Determinism in Die Throwing and the Transition to Chaos[I]

James Bryan's rather philosophical paper on determinism [Bryan, 1971, published 1984] inspired this study on the dynamics of die throwing. Governed by perfect natural laws, he described the process as being completely deterministic; it appears random because no one has managed to measure and control the variables. The process variables describe macroscopic events that do not require quantum mechanics to explain.[II] Then as a matter of principle, determinism applies. Replace the unpredictable human being with an automated system, according to Loxham, and the process will be deterministic.

As a practical matter, the effects of errors compound with each bounce so that seemingly insignificant variations in a throw could perpetuate into unpredictable behavior. In addition, there are unstable states of rest (on an edge or a corner) where an infinitesimal influence could topple the die onto one of two or three faces. Such states are unpredictable and corresponding trajectories must be avoided. Is this system an example of chaos and is chaos an exception to the philosophy of determinism? After all, the mathematics of chaos was generally unknown until 1976, well after Bryan wrote his paper in 1971.[III]

I believe the die problem is a good metaphor for many difficult precision problems and that the deterministic philosophy applies into the realm of chaos far enough to be useful. Often a particular design solution is pushed to the limit of practicality and to go reasonably further begs for a different solution having a higher level of sophistication or some natural advantage. A key first step in the design process is determining the level of difficulty so that proper solutions are used. A deterministic design rests on analytical predictions and/or experimental evidence that the system will function as expected. We will demonstrate the deterministic method by determining the level of precision required to control the trajectory of a die to a predictable state of rest. Perhaps an interested reader will go on to build a machine capable of throwing a die to a designated outcome.

[I] This material was previously published in condensed form [Hale, 1995].

[II] Bryan discusses three physical phenomena that currently require a probabilistic explanation. The first is the interaction of subatomic particles where the Heisenberg uncertainty principle is important. The uncertainty of momentum times the uncertainty of position is given by Plank's constant 6.63×10^{-34} J-s. The second is Brownian movement where the thermal noise of atoms and molecules affects the motion of small particles. Boltzman's constant gives the order of magnitude of the effect 1.38×10^{-23} J/°K. The third is shot noise where very small current flows become discontinuous due to the motion of discrete electrons. The phenomenon becomes significant for currents on the order of 10^{-19} amperes.

[III] Over one-hundred books have been written on the topic of chaos, for example, [Stewart, 1989]. A common view of chaos is that some nonlinear dynamic systems exhibit such complex behavior as to be unpredictable without perfect knowledge of the initial conditions and of the coefficients in the differential equation. This view does not specifically conflict with the deterministic principle.

The best way to proceed is with a perfect dynamic computer model to which known imperfections are applied. The model will not exactly duplicate nature, but for a sensitivity analysis, it must capture the influence of all significant sources of error. We will develop a computer model to simulate the die's trajectory and use it to develop an average sensitivity model that relates the final outcome to variations in the initial state of the die and other parameters describing the die-table interaction. Using the sensitivity model, we will estimate the acceptable errors for a case having some reasonable number of bounces. Simulations will demonstrate that the process is largely deterministic provided that the number of bounces is not too high and the error sources are adequately controlled. Otherwise the system transitions into a state of chaos and becomes unpredictable.

D.1 Developing the Dynamic Model

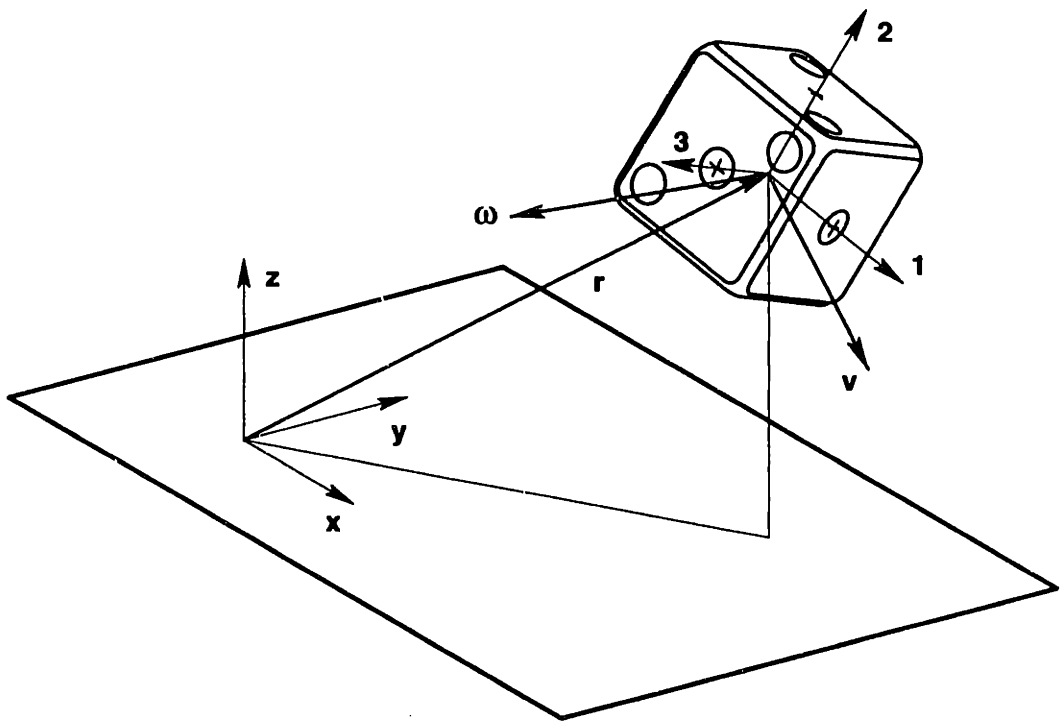


Figure D-1 The system of coordinates consists of an inertial x-y-z system fixed to the table and a moving 1-2-3 system fixed to the die.

The system of coordinates shown in Figure D-1 consists of an inertial x-y-z system fixed to the table and a moving 1-2-3 system fixed to the die. The 1-2-3 system corresponds to the 1-2-3 faces and the 4-5-6 faces are arbitrarily chosen to be opposite from the 1-2-3 faces, respectively.[1] The motion of the die is represented by twelve state variables consisting of the position vector to the mass centroid **r**, the angular orientation of the 1-2-3 coordinate system, the velocity vector **v** and the angular velocity vector ω. The angular orientation is

[1] I learned after completing this work that the opposite faces of a proper die add to seven. Thus the 4-5-6 faces should be opposite from the 3-2-1 faces, respectively.

uniquely defined by a 3 by 3 rotation matrix that can be developed from sequential rotations about three axes, for example, θ_x then θ_y then θ_z, or from a single rotation about an angle vector θ. These representations are different from one another and different from simultaneous rotations about three axes that result from the time integration of angular velocity. The rotation matrix is indispensable for transforming point locations, velocities and angular velocities from one coordinate system to the other. Please see Appendix A for details on transformation matrices used for this dynamic model.

It is important in a large and complicated problem to reduce the parameter space for the test. For example, the symmetry of the table makes the initial conditions for r_x, r_y and θ_z arbitrary and naturally we choose zero. In addition, any possible trajectory can start from an initial condition where v_z is zero (due to gravity in the z-direction). With minimal loss in understanding the sensitivity of the system, we will set the remaining *nominal* velocities to zero, which leaves only three state variables r_z, θ_x and θ_y in the parameter space. The symmetry of the cube reduces the parameter space even further. As Figure D-2 shows, each face has eight-fold symmetry and we need to test only one face, thus compressing the parameter space by 48 times. Although completely arbitrary, it is most convenient to work in a positive angular range $0 \leq \theta_x \leq 45°$ and $0 \leq \theta_y \leq \arctan(\sin(45°))$, which places the gravity vector in octant 7 as shown.

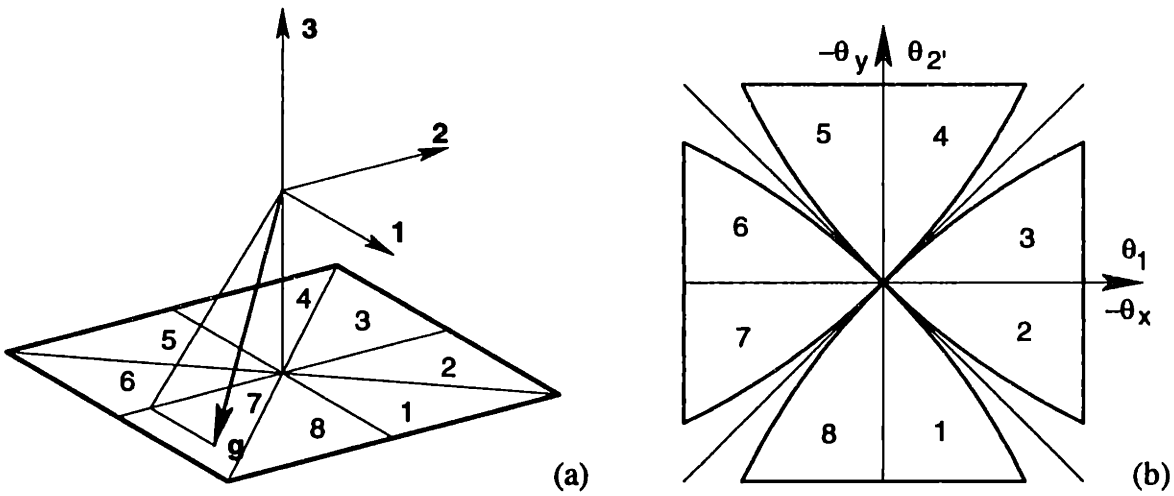


Figure D-2 The orientation of the gravity vector through any face has an eight-fold symmetry represented in 3D (a) or projected into angular coordinates (b).

The trajectory of the die through space is calculated by integrating the system of nonlinear state equations, given by Equation D.1, using a fourth-order Runge-Kutta algorithm. Rows 1 and 2 are the kinematic equations. Row 2 contains the Jacobian matrix **J**, which relates the die's angular velocity to the rates of change of sequential angles that describe the die's orientation. With a little effort, the Jacobian can be derived from Equation D.2, where each side describes the die's changing orientation in terms of the rotation matrix

$\mathbf{R}(\theta)$.[I] Rows 3 and 4 are the kinetic equations and include the acceleration of gravity -$\mathbf{g}$ and quadratic terms for aerodynamic drag.[II]

$$\frac{d}{dt}\begin{bmatrix}\mathbf{r}\\ \theta\\ \mathbf{v}\\ \omega\end{bmatrix}=\begin{bmatrix}\mathbf{v}\\ \mathbf{J}(\theta)^{-1}\cdot\omega\\ -\mathbf{g}-\frac{b}{m}\|\mathbf{v}\|\,\mathbf{v}\\ -\frac{c}{I}\|\omega\|\,\omega\end{bmatrix} \tag{D.1}$$

$$\frac{d}{dt}[\mathbf{R}(\theta)]=\begin{bmatrix}0 & -\omega_z & \omega_y\\ \omega_z & 0 & -\omega_x\\ -\omega_y & \omega_x & 0\end{bmatrix}\cdot\mathbf{R}(\theta) \tag{D.2}$$

Numerical integration of the contact phenomenon is computationally expensive and numerically sensitive. Instead, a nonlinear solver based on Newton iteration terminates the integration when the die makes contact with the table. At this instant an impact model calculates new velocity states for the die and determines if there is sufficient energy to reach another face. If so, the integration will start again from this new state and end at the next impact. The impact model is admittedly simplistic being based on the coefficient of restitution and the coefficient of friction, but is appropriate for a sensitivity study. The table is assumed to be rigid since this effect can be lumped into the coefficient of restitution.

The impact model provides the relationship between an impulse δ_a applied to the contact point and the change in the velocity state of the die. Equation D.3 provides the kinematic relationship between the velocity $\mathbf{v}_a$ at the contact point and the velocity states of the die. Equation D.4 provides the kinetic relationship between the impulse and the change in velocity states. Derived either from equilibrium or compatibility, Equation D.5 defines the transformation matrix $\mathbf{A}$ that appears in both equations. Combining these equations in Equation D.6 results in an invertible relationship between $\Delta\mathbf{v}_a$ and δ_a, where $\mathbf{C}$ is a diagonal matrix involving reciprocals of the mass m and the moment of inertia I.

$$\mathbf{v}_a=\mathbf{A}\cdot\begin{bmatrix}\mathbf{v}\\ \cdots\\ \omega\end{bmatrix} \tag{D.3}$$

[I] To avoid orientations where the Jacobian matrix becomes singular and to speed the numerical integration, an intermediate coordinate system, aligned after each bounce to the spin axis of the die, was used in the simulation to track the orientation.

[II] The fourth row is simplified to take advantage of a cube having equal principal moments of inertia; however, the simulation is more general and accepts three separate principal moments of inertia.

$$\begin{bmatrix} m\,\Delta\mathbf{v} \\ \text{----} \\ I\,\Delta\omega \end{bmatrix} = \mathbf{A}^{\mathrm{T}} \cdot \delta_a \tag{D.4}$$

$$\mathbf{A} = \left[\begin{array}{ccc:ccc} 1 & 0 & 0 & 0 & r_{a_z} & -r_{a_y} \\ 0 & 1 & 0 & -r_{a_z} & 0 & r_{a_x} \\ 0 & 0 & 1 & r_{a_y} & -r_{a_x} & 0 \end{array}\right] \tag{D.5}$$

$$\Delta\mathbf{v}_a = \mathbf{A} \cdot \mathbf{C} \cdot \mathbf{A}^{\mathrm{T}} \cdot \delta_a \tag{D.6}$$

The first step in the impact algorithm uses Equation D.6 to calculate the impulse required to change the velocity at the contact point to zero. Slip will occur in the plane of the table when the angle between the impulse vector and the table normal is greater than the friction angle tan(μ), where μ is the coefficient of friction. The algorithm handles slip by directing the impulse vector along the friction angle and by changing its magnitude so that the normal velocity at the contact point remains zero. The elasticity of the die and the table is modeled using the coefficient of restitution e. The effect is to enlarge the impulse by a factor $(1 + e)$, where $e = 0$ for a perfectly plastic material and $e = 1$ for a perfectly elastic material. The new velocity states are then calculated using Equation D.4.

Figure D-3 and Figure D-4 show the behavior of a typical simulation. The die is 15 mm in size with 1.5 mm edge and corner radii, and the density corresponds to Delrin plastic. The die drops straight down from a height of 0.2 meters with no initial velocity. The coefficients of friction and restitution are 0.2 and 0.6, respectively. The first impact develops appreciable angular momentum and causes significant lateral movement. There appear to be three bounces before going into the final *rattle* mode after about 0.6 seconds. The simulation seems fairly realistic of dice on semi-hard tables. Hard, slick tables produce considerably more action and longer simulations.

D.2 Developing the Sensitivity Model

An average sensitivity model is useful for this problem because certain trajectories are many times more sensitive than others. The average better determines the levels of difficulty for controlling each error source. The sensitivity model will relate variations in the initial state of the die and other parameters that affect the die-table interaction to accumulated variations after a given number of bounces. It can be used in reverse to budget the errors for deterministic behavior, on average. We will use $r_z = 0.2$ m, $\mu = 0.2$ and $e = 0.6$ (the same as before) as *reasonable* values to conduct the sensitivity analysis realizing that the results are highly dependent on these parameters. The angular orientation will range over one octant to obtain the average sensitivity model.

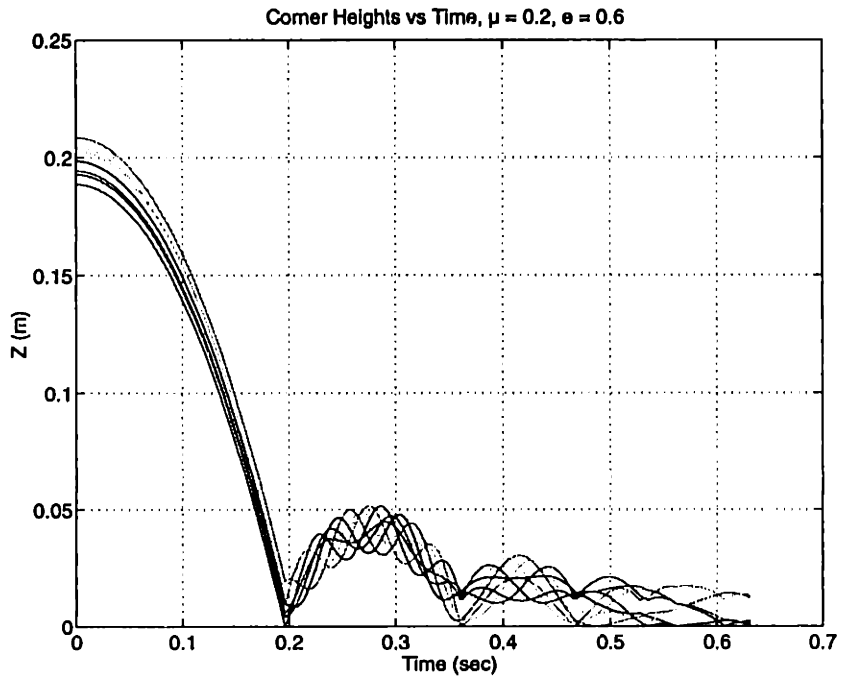


Figure D-3 The heights of all corners of the die are plotted versus time.

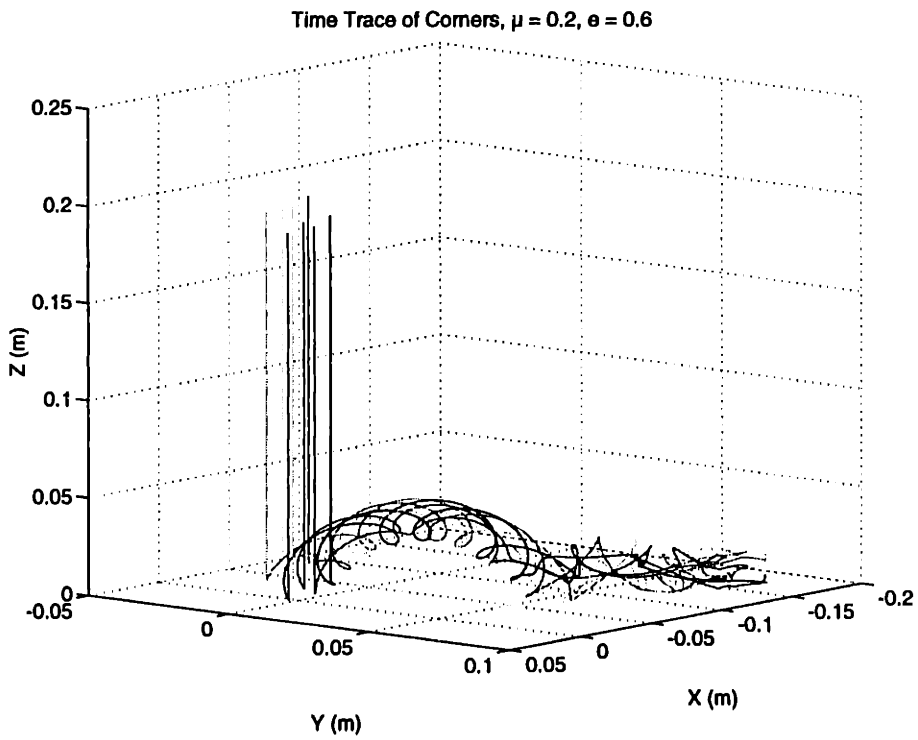


Figure D-4 The time history of all corners of the die is plotted versus spatial dimensions x, y and z.

The usual way to determine the sensitivity of a parameter is to run the full simulation with a slight error and compare the effect to a run with no error. This is repeated for each parameter times the number of samples in the average. There are several problems with this method, which prompted the alternative used here. Instead of running the full simulation, only the first bounce is simulated. The effect of multiple bounces is

accomplished by raising the sensitivity matrix for one bounce to the corresponding power. One bounce consists of one impact followed by one trajectory. The velocity states just after the first impact are influenced by small variations in the prior velocity states, the angular orientation, the coefficient of friction and the coefficient of restitution.[1] Once the die breaks contact with the table, there are no significant error sources that affect the velocity states. In the end, we are only interested in the variation in the angular orientation as this determines whether the die lands on a different face.

Calculating the sensitivity matrix requires running the impact algorithm once for each error source and comparing that to one with zero error. Then this set is repeated many times over the angular range to obtain the rms average of the sensitivity matrix given by Equation D.7. In addition, Equation D.8 gives the rms average of the velocity states resulting after the first impact. On average, the die develops significant spin, 25.4 revolutions per second (or 1524 rpm).

$$\begin{bmatrix} dv_x \\ dv_y \\ dv_z \\ \hline d\omega_x \\ d\omega_y \\ d\omega_z \end{bmatrix}_1 = \left[\begin{array}{ccc:ccc:ccc:cc} 0.9096 & 0.0577 & 0.1964 & 0.0010 & 0.0018 & 0.0004 & 0.4420 & 0.9661 & 2.0564 & 0.2365 \\ 0.0577 & 0.8409 & 0.1062 & 0.0021 & 0.0010 & 0.0007 & 0.9186 & 0.5038 & 1.1115 & 0.1279 \\ 0.0835 & 0.0448 & 0.3333 & 0.0025 & 0.0043 & 0.0000 & 1.2709 & 2.4147 & 1.9466 & 1.4762 \\ \hdashline 17.064 & 59.949 & 43.262 & 0.4594 & 0.1752 & 0.1964 & 249.70 & 75.352 & 203.43 & 52.232 \\ 48.675 & 17.064 & 70.426 & 0.1752 & 0.5016 & 0.1062 & 76.129 & 347.27 & 392.60 & 84.955 \\ 10.038 & 18.227 & 0.0000 & 0.1964 & 0.1062 & 0.9067 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{array}\right]$$

$$\cdot \left[\begin{array}{ccc:ccc:cc:cc} dv_x & dv_y & dv_z & d\omega_x & d\omega_y & d\omega_z & d\theta_x & d\theta_y & d\mu & de \end{array}\right]_0^T \tag{D.7}$$

$$\begin{bmatrix} v_x \\ v_y \\ v_z \\ \hline \omega_x \\ \omega_y \\ \omega_z \end{bmatrix}_1 = \left[\begin{array}{cc} 0.3784 & \text{m/s} \\ 0.2047 & \text{m/s} \\ 0.6414 & \text{m/s} \\ \hline 83.5716 & \text{r/s} \\ 135.9279 & \text{r/s} \\ 0.0000 & \text{r/s} \end{array}\right] \tag{D.8}$$

So far the sensitivity matrix is incomplete since it must be square to multiply itself. The key missing terms are the variations in angles at the next impact. Essentially a differential of a time integration to the next impact, Equation D.9 provides a reasonable approximation to the angular variations, where $2v_z/a$ is the approximate time in flight. In Equation D.10, this information is assembled into a second sensitivity matrix for the trajectory phase of the bounce. Equation D.11 gives the full sensitivity matrix for one bounce, obtained by multiplying the two sensitivity matrices and augmenting additional

[1] A variation in the height of the die can be expressed as a variation in the vertical velocity component.

rows for $d\mu$ and de. The rows corresponding to $d\theta_x$ and $d\theta_y$ are of most interest as noted before, and their terms are of the same magnitude, usually within a factor of 2.

$$\begin{bmatrix} d\theta_x \\ d\theta_y \end{bmatrix} \cong \frac{2\,dv_z}{a}\begin{bmatrix} \omega_x \\ \omega_y \end{bmatrix} + \frac{2v_z}{a}\begin{bmatrix} d\omega_x \\ d\omega_y \end{bmatrix} \tag{D.9}$$

$$\begin{bmatrix} dv_x \\ dv_y \\ dv_z \\ \hline d\omega_x \\ d\omega_y \\ d\omega_z \\ \hline d\theta_x \\ d\theta_y \end{bmatrix}_1 = \left[\begin{array}{ccc:ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \hdashline 0 & 0 & 17.038 & 0.1308 & 0 & 0 \\ 0 & 0 & 27.712 & 0 & 0.1308 & 0 \end{array}\right] \cdot \begin{bmatrix} dv_x \\ dv_y \\ dv_z \\ \hline d\omega_x \\ d\omega_y \\ d\omega_z \end{bmatrix}_0 \tag{D.10}$$

$$\begin{bmatrix} dv_x \\ dv_y \\ dv_z \\ \hline d\omega_x \\ d\omega_y \\ d\omega_z \\ \hline d\theta_x \\ d\theta_y \\ \hline d\mu \\ de \end{bmatrix}_1 = \left[\begin{array}{ccc:ccc:cc:cc} 0.9096 & 0.0577 & 0.1964 & 0.0010 & 0.0018 & 0.0004 & 0.4420 & 0.9661 & 2.0564 & 0.2365 \\ 0.0577 & 0.8409 & 0.1062 & 0.0021 & 0.0010 & 0.0007 & 0.9186 & 0.5038 & 1.1115 & 0.1279 \\ 0.0835 & 0.0448 & 0.3333 & 0.0025 & 0.0043 & 0.0000 & 1.2709 & 2.4147 & 1.9466 & 1.4762 \\ \hdashline 17.064 & 59.949 & 43.262 & 0.4594 & 0.1752 & 0.1964 & 249.70 & 75.352 & 203.43 & 52.232 \\ 48.675 & 17.064 & 70.426 & 0.1752 & 0.5016 & 0.1062 & 76.129 & 347.27 & 392.60 & 84.955 \\ 10.038 & 18.227 & 0.0000 & 0.1964 & 0.1062 & 0.9067 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ \hdashline 3.6551 & 8.6045 & 11.338 & 0.1025 & 0.0963 & 0.0257 & 54.314 & 50.998 & 59.775 & 31.983 \\ 8.6815 & 3.4733 & 18.449 & 0.0918 & 0.1850 & 0.0139 & 45.178 & 112.34 & 105.29 & 52.019 \\ \hdashline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \cdot \left[\begin{array}{ccc:ccc:cc:cc} dv_x & dv_y & dv_z & d\omega_x & d\omega_y & d\omega_z & d\theta_x & d\theta_y & d\mu & de \end{array}\right]_0^T \tag{D.11}$$

The sensitivity model must account for energy decay beyond the first bounce, which raising the sensitivity matrix to a higher power does not capture. As Equation D.12 shows, the square of the coefficient of restitution represents the energy decay per bounce, where n is the total number of bounces and $\mathbf{S}$ is the sensitivity matrix. Equation D.13 shows the sensitivity for two bounces and Equation D.14 shows it for three bounces. Beyond three the problem is quite hopeless. These equations can be used in reverse to estimate the average precision required for deterministic behavior. Table D-1 shows the estimates calculated using an angular change of 45° after three bounces as the criterion for average deterministic behavior. The first column is a worst case assuming all errors add and contribute equally. The second column is optimistic assuming a quadrature sum. Often used as a more realistic estimate, the third column is the average of the first two columns.

$$[\mathbf{S}_n] \cong e^{2(n-1)}[\mathbf{S}_1]^n \tag{D.12}$$

$$\begin{bmatrix} d\theta_x \\ d\theta_y \end{bmatrix}_2 = \left[\begin{array}{ccc|ccc|cc|cc} 234.98 & 237.85 & 566.39 & 3.7327 & 5.3269 & 0.7793 & 1911.9 & 3087.2 & 3158.7 & 1603.5 \\ 417.87 & 285.15 & 939.60 & 5.4312 & 9.1222 & 0.9998 & 2734.7 & 5418.0 & 5322.0 & 2660.8 \end{array}\right] \cdot \left[\begin{array}{ccc|ccc|cc|cc} dv_x & dv_y & dv_z & d\omega_x & d\omega_y & d\omega_z & d\theta_x & d\theta_y & d\mu & de \end{array}\right]_0^T \quad \text{(D.13)}$$

$$\begin{bmatrix} d\theta_x \\ d\theta_y \end{bmatrix}_3 = \left[\begin{array}{ccc|ccc|cc|cc} 12382 & 9987 & 28594 & 174.39 & 274.20 & 33.934 & 88450 & 161340 & 160990 & 80969 \\ 20899 & 15536 & 47616 & 282.78 & 459.46 & 53.574 & 142920 & 271620 & 268910 & 134840 \end{array}\right] \cdot \left[\begin{array}{ccc|ccc|cc|cc} dv_x & dv_y & dv_z & d\omega_x & d\omega_y & d\omega_z & d\theta_x & d\theta_y & d\mu & de \end{array}\right]_0^T \quad \text{(D.14)}$$

Error Term =	45°/10	45°/ $\sqrt{10}$	average
dv_x (m/s)	4.72 e-06	1.49 e-05	9.82 e-06
dv_y (m/s)	6.15 e-06	1.95 e-05	1.28 e-05
$5\, dr_z = dv_z$	2.06 e-06	6.52 e-06	4.29 e-06
$d\omega_x$ (r/s)	3.44 e-04	1.09 e-03	7.15 e-04
$d\omega_y$ (r/s)	2.14 e-04	6.77 e-04	4.46 e-04
$d\omega_z$ (r/s)	1.80 e-03	5.68 e-03	3.74 e-03
$d\theta_x$ (r)	6.79 e-07	2.15 e-06	1.41 e-06
$d\theta_y$ (r)	3.63 e-07	1.15 e-06	7.55 e-07
$d\mu$	3.65 e-07	1.16 e-06	7.60 e-07
de	7.28 e-07	2.30 e-06	1.51 e-06

Table D-1 These errors contribute equally to non deterministic behavior either as an arithmetic sum in the first column or a quadrature sum in the second column. The third column is an average of the two.

It turns out that these error estimates are at least an order of magnitude smaller than what works for the full simulation in the vicinity of $\theta_x = 30°$, $\theta_y = 20°$. Had this been an actual project, rather than a fun example, the wiser path to more accurate sensitivities would be brute-force computer power running thousands of simulations.

D.3 Simulation Results

The simulations shown in the following figures have the same nominal parameters as the sensitivity analysis, but the angular range is much smaller and centered about $\theta_x = 30°$ and $\theta_y = 20°$. Each grid point in a figure represents one simulation and its outcome is indicated on the vertical scale. Figure D-5 shows simulations over the full range of an octant with zero errors. The grid is so coarse at 2.25° increments that the results appear random, but a closer look at Figure D-6 reveals the start of deterministic behavior with 0.05° increments. Still finer grids show more pronounced plateaus. Figure D-7 with 0.01° increments shows one nice plateau and several thin ridges and valleys. Figure D-8 with 0.0025° increments

shows several nice plateaus and indicates the angular precision required. The introduction of uniform random errors to the model parameters degrades the nice deterministic behavior. Figure D-9 shows noticeable irregularity near transition boundaries from errors of magnitude $dz = 0.000\ 03$ m, $d\mu = 0.000\ 02$ and $de = 0.000\ 05$. Figure D-10 shows non repeatable behavior over most of the plateau from errors of magnitude $dz = 0.0001$ m, $d\mu = 0.000\ 06$ and $de = 0.000\ 15$. Figure D-11 shows completely non repeatable behavior from errors of magnitude $dz = 0.0003$ m, $d\mu = 0.0002$ and $de = 0.0005$.

The sensitivity analysis and the results of the simulations indicate that it is impractical to control the trajectory of a die to a predictable state of rest as an open-loop process. We have the ability to control the initial conditions well enough but the die-table interaction probably has two to three orders of magnitude greater variability than required for a repeatable process. We could seek to find the right combination of materials with sufficient control of their properties, surface textures, contamination levels, dimensions and so forth to achieve the tolerances required for μ and e. Alternatively, we could apply closed-loop feedback to control the process through any one of several variables that affect the die-table interaction, for example, slight z-movement of the table or a controllable material that affects the coefficient of restitution. This solution requires a much higher level of sophistication but it has no natural limitations.

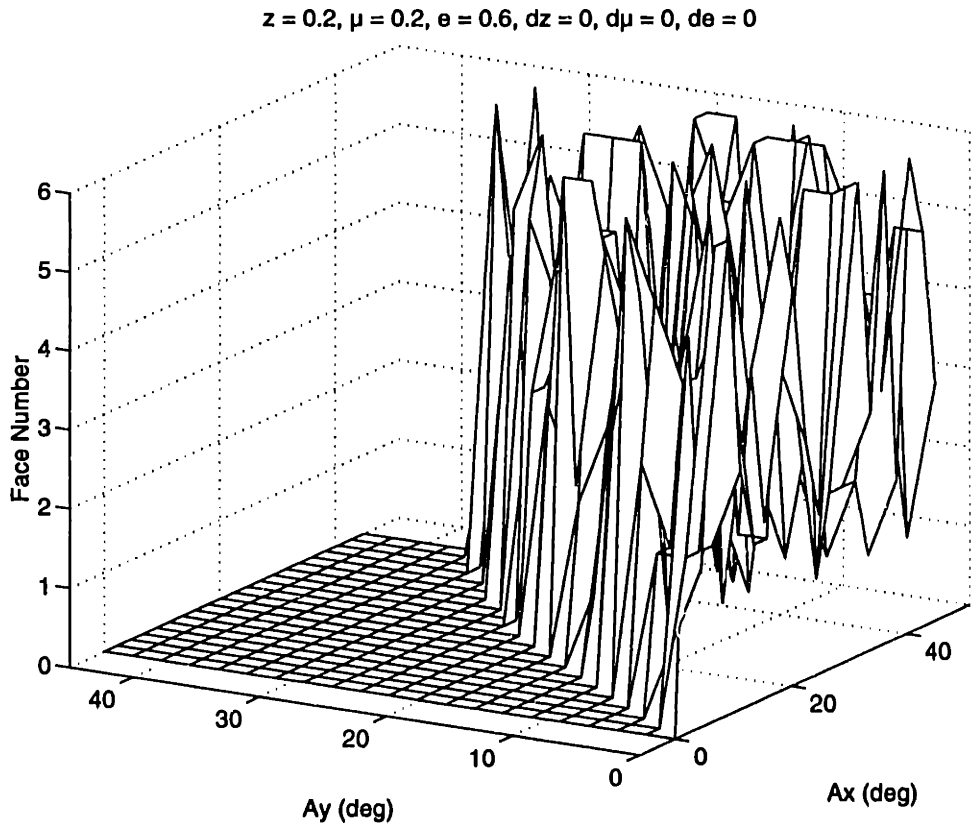


Figure D-5 The simulation appears random for 2.25° grid increments over the range of one octant.

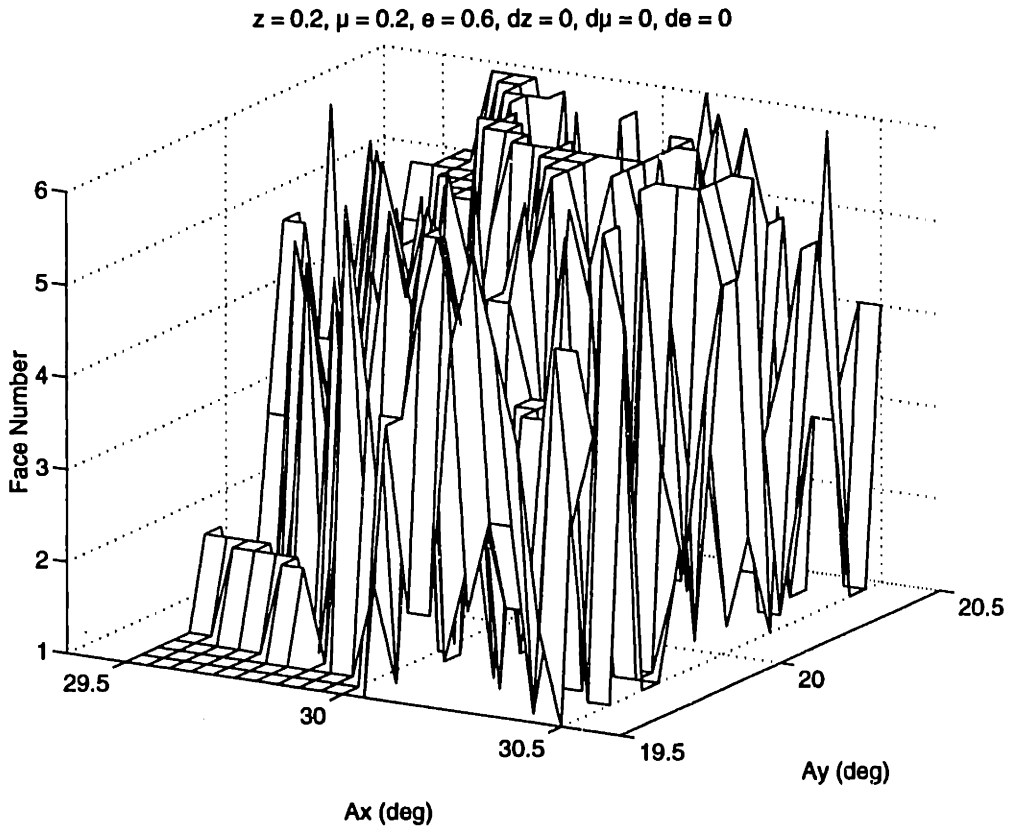


Figure D-6 The simulation starts to show pockets of deterministic behavior for 0.05° grid increments.

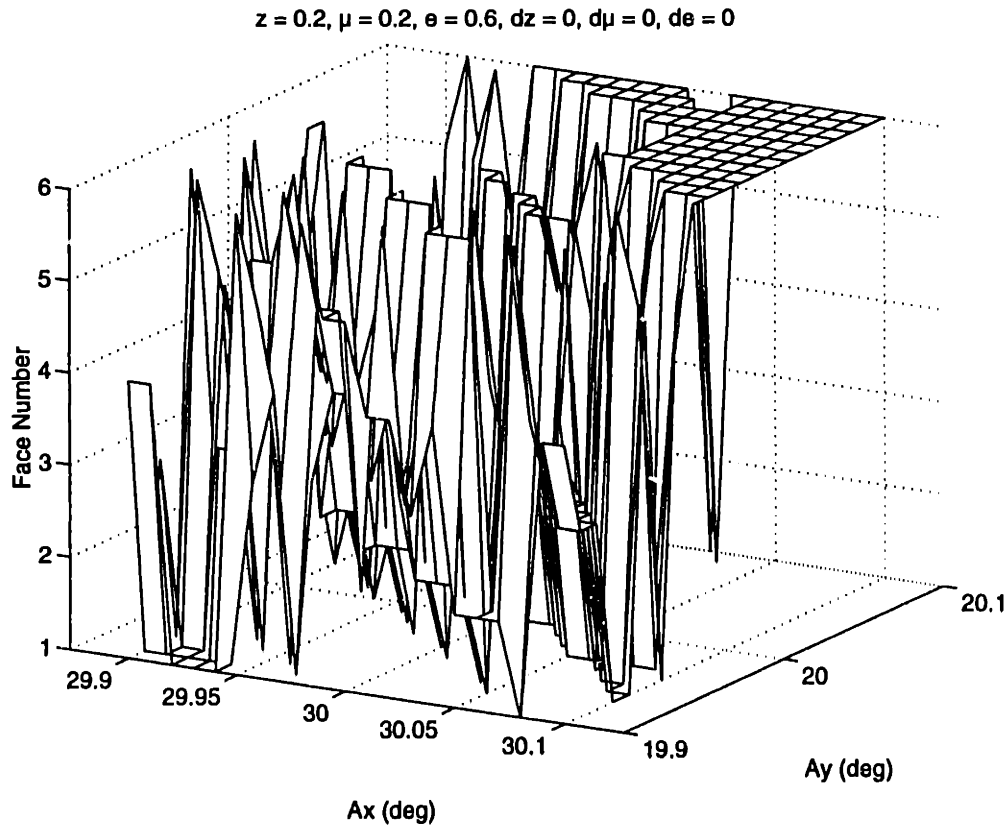


Figure D-7 The wide plateau apparent with 0.01° grid increments shows near the center of Figure D-6.

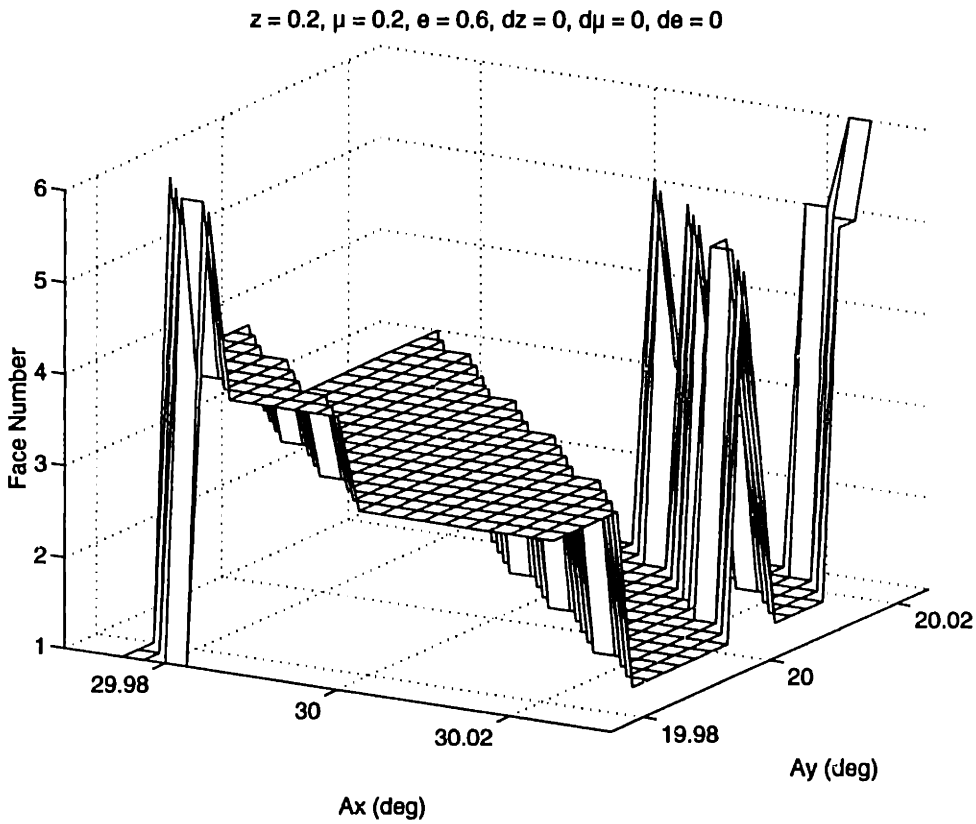


Figure D-8 This simulation with small 0.0025° grid increments indicates the angular precision required for deterministic behavior.

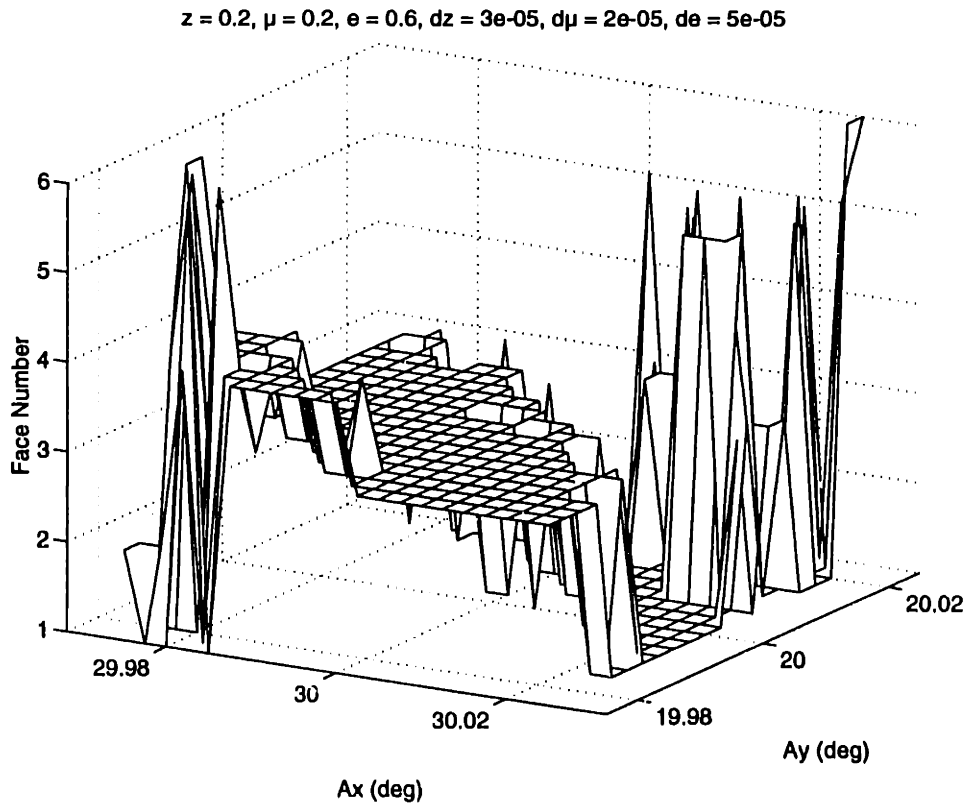


Figure D-9 Uniform random errors as indicated cause non repeatable behavior near the transitions.

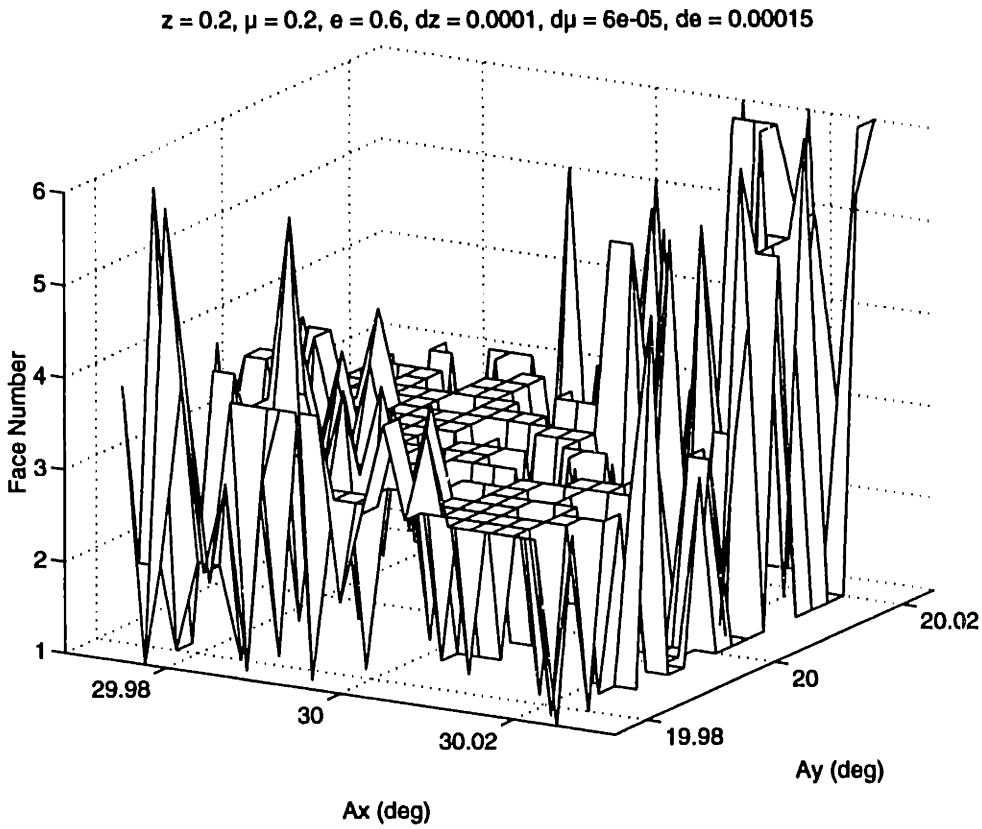


Figure D-10 Uniform random errors as indicated cause non repeatable behavior over most of the plateau.

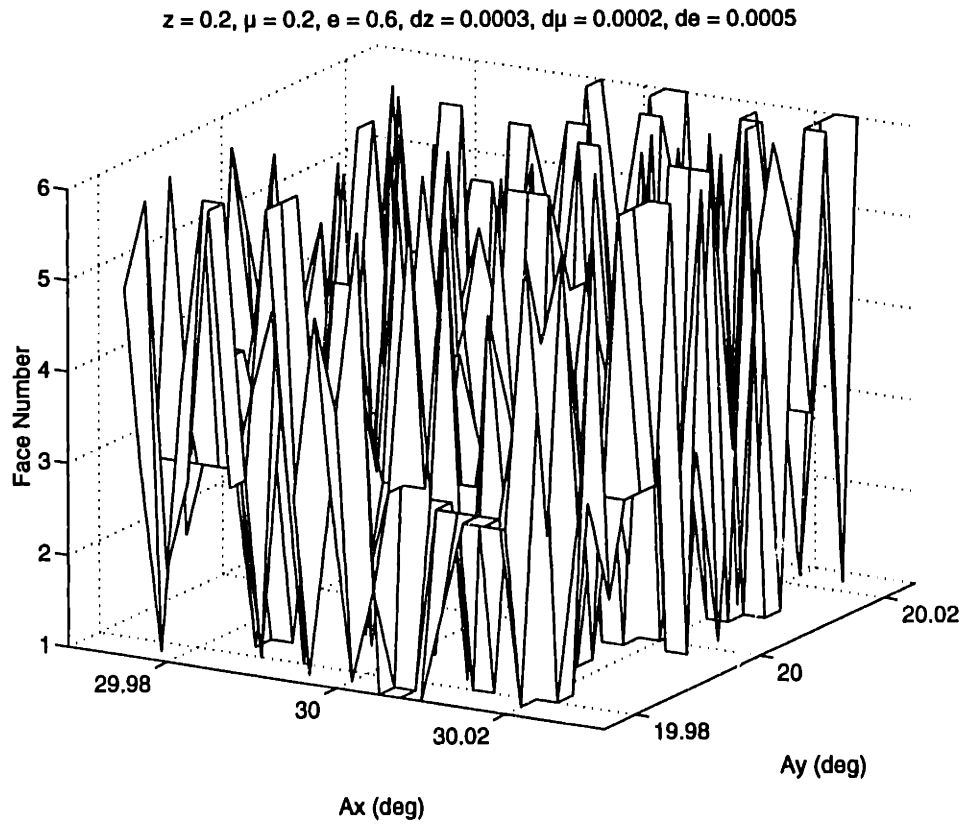


Figure D-11 Uniform random errors as indicated cause completely non repeatable behavior.

D.4 An Example from the Chaos Literature

Chaos has existed physically amidst the order of Nature since the beginning. Early humans probably first applied chaos when they mixed pigments for paint, crafted clay pots or kneaded dough. We have a long history of building machines that rely on chaotic processes for mixing but only recently have we come to understand the stretch-and-fold order of chaos from mathematical and scientific perspectives. Mitchell Feigenbaum's study of the logistic mapping effectively established chaos as a field of study, although others made important prior discoveries. This and many other historical examples of chaos theory appear in a well-written book by [Stewart, 1989]. The purpose for including this example is to offer a much different perspective of deterministic systems.

Although interested in turbulence, Feigenbaum was wise to start with a simple nonlinear difference equation rather than the complicated Navier-Stokes equations. When iterated, Equation D.15 describes a discrete dynamic system whose behavior can change drastically for small changes to its only parameter k. In Figure D-12 the cobweb diagrams conveniently show the discrete time history of the system for different values of k. The cobweb is created by alternately projecting a line vertically to the parabola, which is the same as evaluating the equation, then projecting a line horizontally to the 1:1 sloping line, which is the iteration step. When $k < 3$, the system converges to one stable point determined simply by solving for $x = 1 - 1/k$. Convergence slows to zero at $k = 3$ because the slope of the parabola at the stable point x = 2/3 is -1; the nearly square web can barely close on the point. When $k > 3$, the point becomes unstable and pushes the web away toward two new points called stable attractors. This is an example of bifurcation, which is fairly common in dynamic systems. However as k increases further to about 3.4495, the two points become unstable and four new attractors form. Bifurcation occurs over and over and ever faster as k increases until apparently random chaos. Yet for any reasonable number of bifurcations n and a value of k in that range, the attractive states can be computed by solving the system of nonlinear equations in Equation D.16 or by iterating Equation D.15 until it converges.

$$x_{n+1} = k\, x_n(1 - x_n) \tag{D.15}$$

$$\begin{bmatrix} x_2 \\ \vdots \\ x_{2n} \\ x_1 \end{bmatrix} = k \begin{bmatrix} x_1(1-x_1) \\ x_2(1-x_2) \\ \vdots \\ x_{2n}(1-x_{2n}) \end{bmatrix} \tag{D.16}$$

An example of this behavior is the limit cycle, which is most noticeable in servo-mechanisms with sliding bearings. Static friction is the source of nonlinearity that causes the linear control system to alternately step from one stable attractor to another, never

reaching the goal in between. Machine systems may be too complicated to display all the intricacies of chaos, but the simple logistic mapping exhibits a complexity that no one could imagine. A fundamental constant falls out of its chaos, which has been observed in natural phenomena. The interval from one bifurcation to the next decreases by a factor approaching 4.6692016090 each time. The small pockets of order that emerge from the chaos are also remarkable. This behavior is apparent in Figure D-13, where the attractive states of the logistic mapping are plotted as a function of k. You can see the relationship to the cobweb diagrams after convergence; the last cycle tracing out the order in which the iteration visits each attractor. You can imagine a movie of patterns developing as the height of the parabola increases with time. It would be simple to program but difficult to play in a thesis.

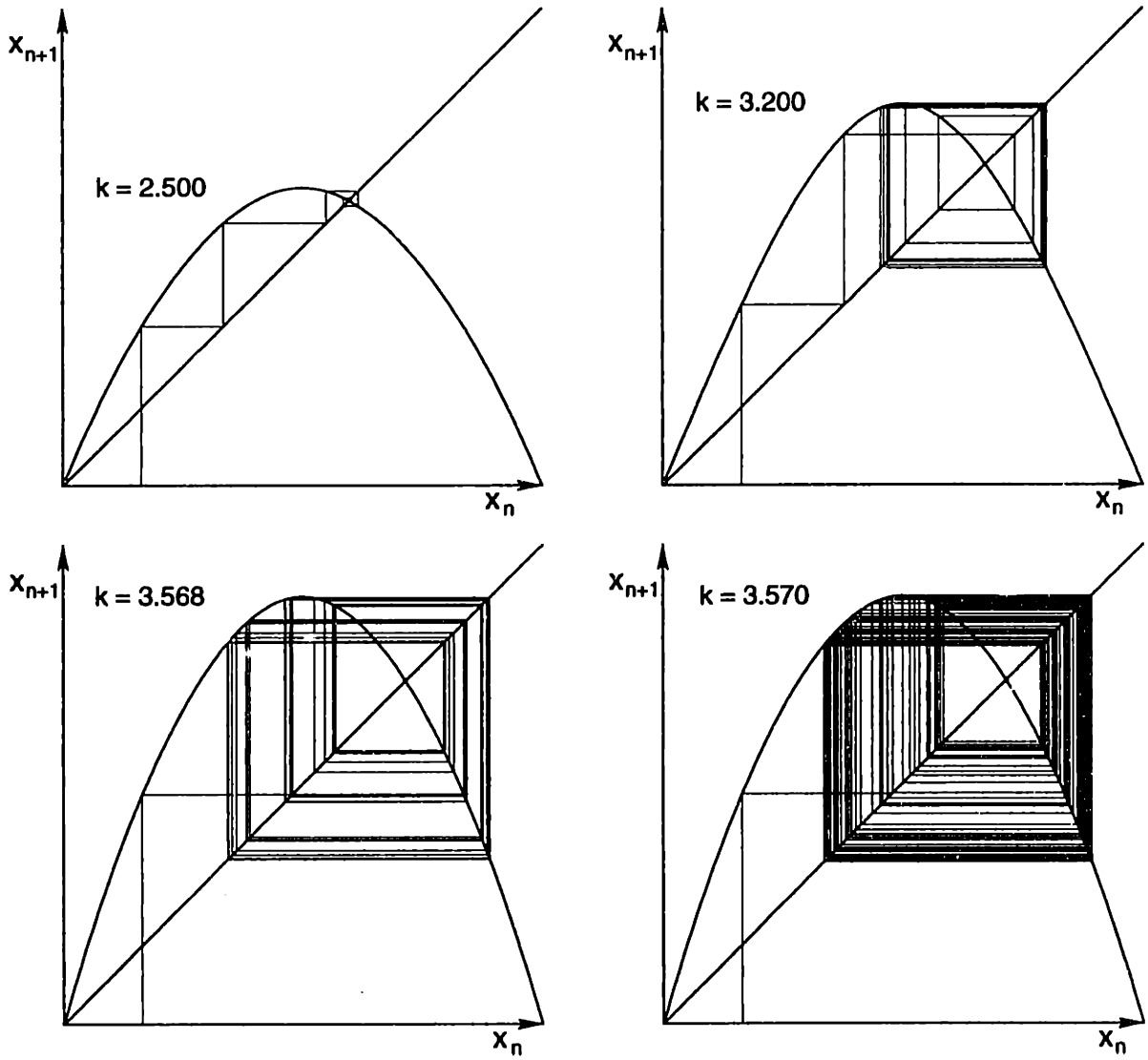


Figure D-12 Cobweb diagrams conveniently show the discrete time history of the logistic mapping when iterated. The value of the k parameter determines the number and locations of the attractive states.

This brief look at chaos may hold only one lesson for the precision machine designer: try to avoid friction and other nonlinear processes that could manifest chaotic

behavior. When this is unavoidable, explore the parameter space and operate away from problem areas. There may be ways to modify the system with linear elements that delay nonlinear problems.

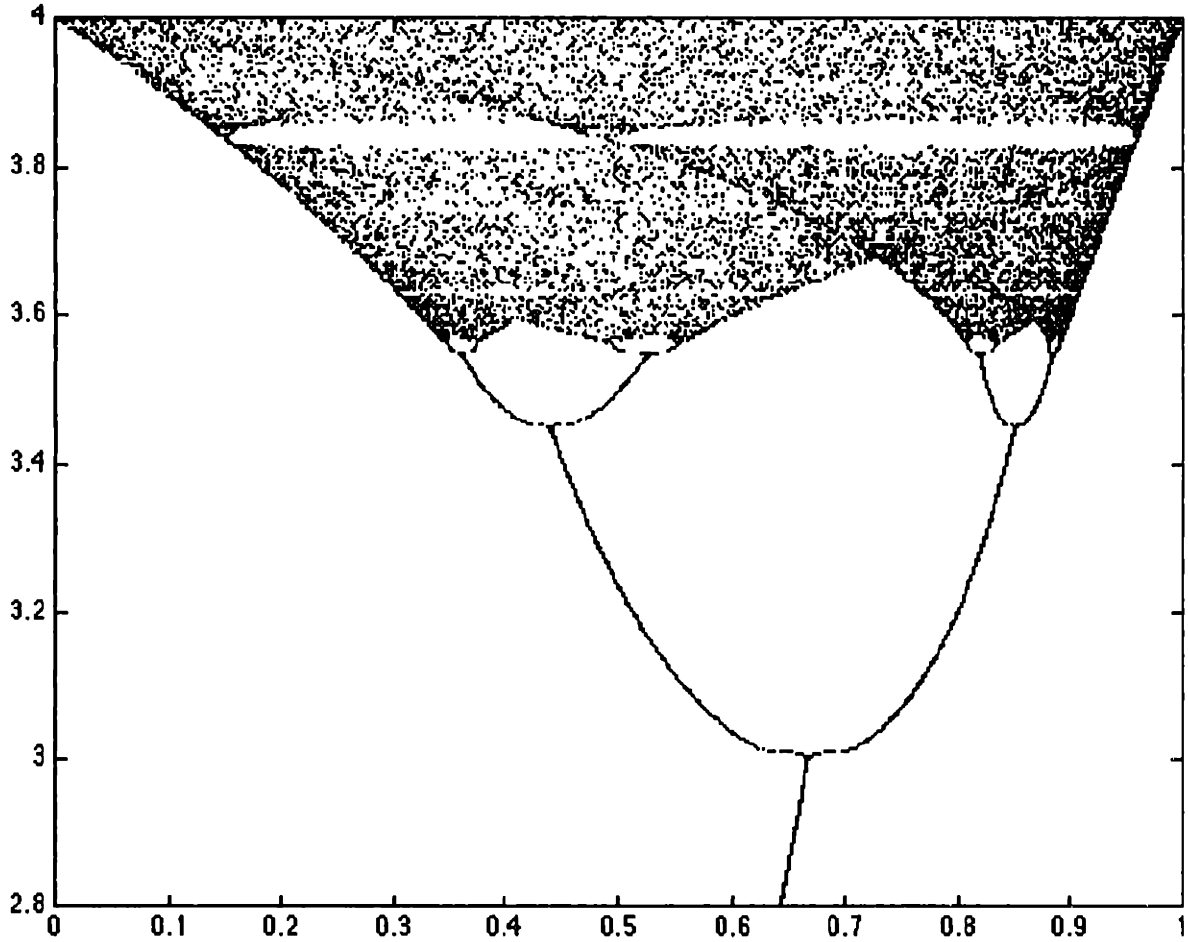


Figure D-13 The bifurcation diagram of the logistic mapping shows the attractive states along the horizontal axis as the *k* parameter increases up the vertical axis.

E Orthogonal Machining Model

The simple orthogonal cutting model developed by [Ernst and Merchant, 1941] is the basis for the machining model developed here.[1] Their model relates the radial and tangential cutting forces F_r and F_t, shown in Figure E-1, to the shear strength τ_s of the material being cut, the uncut chip area A_u, the tool rake angle α and the friction angle β between the chip and the tool. Equation E.1 presents this model as a pressure vector as a function of the angle ϕ to the shear plane. If the rake angle is chosen to be zero, which is a common assumption, then both components are approximately equal to 4.8 times the shear strength or 2.8 times the tensile strength. The tangential component of pressure is equivalent to the specific power (or unit power or power constant) in the appropriate units. For example, steel having a Brinell hardness of 300 would produce a tangential pressure of 410 ksi or 1.04 hp-min/in^3, which is very close to published experimental results.

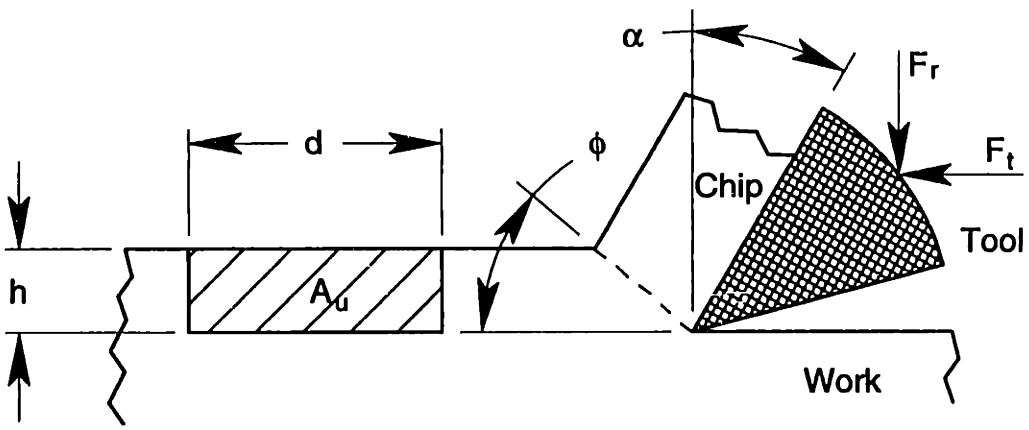


Figure E-1 Orthogonal cutting model showing the key cutting parameters.

$$\begin{bmatrix} p_t \\ p_r \end{bmatrix} = \frac{1}{A_u}\begin{bmatrix} F_t \\ F_r \end{bmatrix} = \frac{2\tau_s}{\tan(\phi)\sin(2\phi)}\begin{bmatrix} \sin(2\phi) \\ \cos(2\phi) \end{bmatrix} \qquad (E.1)$$

$$2\phi = 90^\circ + \alpha - \beta \cong 45^\circ + \alpha$$

A multiple-tooth, rotating-tool cutting model is developed from the orthogonal model by integrating the effects of the variable chip thickness and the number of teeth over the arc of contact. Figure E-2 shows the geometry of the milling cutter and the variables of interest. Given these cutting parameters, we must calculate the three orthogonal reaction forces (F_x, F_y, F_z) applied to the spindle and the required spindle torque T_s. Equation E.2 gives the variable chip thickness in terms of the instantaneous tooth angle θ, the lead angle λ, the circular spindle velocity (or surface speed) V_s, the feed velocity V_f and the circular

[1] For more recent references, see for example, Analysis of Material Removal Processes by W.R. DeVries or Fundamentals of Machining and Machine Tools by G. Boothroyd and W.A. Knight.

tooth spacing s. This is rearranged in Equation E.3 to show that the feed velocity depends on the maximum allowable chip thickness h_{max}, which occurs at the minimum angle θ_{min}. Combining (E.2) and (E.3) simplifies the variable chip thickness given by Equation E.4.

$$h(\theta) = \frac{s}{V_s} V_f \cos(\theta)\cos(\lambda) \tag{E.2}$$

$$V_f = \frac{V_s}{s} \frac{h_{max}}{\cos(\theta_{min})\cos(\lambda)} \tag{E.3}$$

$$h(\theta) = h_{max} \frac{\cos(\theta)}{\cos(\theta_{min})} \tag{E.4}$$

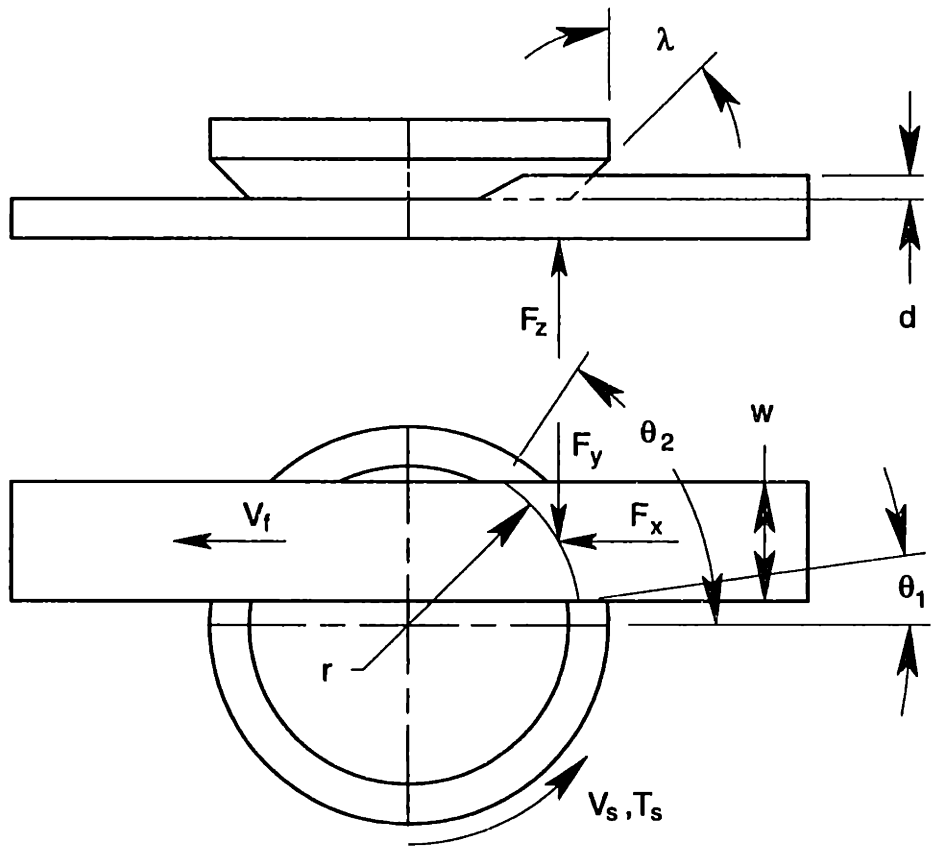


Figure E-2 Milling model showing the cutter and the workpiece.

Equation E.5 shows the development of the spindle torque, which consists of the mean radius of cut r, the tangential pressure, the depth of cut, the average chip thickness and the average number of teeth in cut. The torque is proportional to the radius if the circular tooth spacing is constant or to the number of teeth n. The spindle power P_s given by Equation E.6 is easy to calculate from the spindle torque and the spindle speed given as a circular velocity and a radius or the rotational speed N.

$$\begin{aligned} T_s &= r \cdot p_t \frac{d}{\cos(\lambda)} \frac{\int_{\theta_1}^{\theta_2} h(\theta)\,d\theta}{\theta_2 - \theta_1} \frac{r(\theta_2 - \theta_1)}{s} \\ &= r \cdot p_t \frac{d}{\cos(\lambda)} \frac{h_{max}\left(\sin(\theta_2) - \sin(\theta_1)\right)}{\cos(\theta_{min})} \frac{r}{s} \\ &= \frac{r}{s} p_t \frac{d \cdot w \cdot h_{max}}{\cos(\lambda)\cos(\theta_{min})} = \frac{n}{2\pi} p_t \frac{d \cdot w \cdot h_{max}}{\cos(\lambda)\cos(\theta_{min})} \end{aligned} \tag{E.5}$$

$$P_s = \frac{V_s}{s} p_t \frac{d \cdot w \cdot h_{max}}{\cos(\theta_{min})\cos(\lambda)} = n \cdot N \cdot p_t \frac{d \cdot w \cdot h_{max}}{\cos(\theta_{min})\cos(\lambda)} \tag{E.6}$$

The development of the force vector is similar to but more complicated than the spindle torque due to the changing orientation of the tangential and radial tooth forces. For this reason, only the solution is given in Equation E.7. If the width of the workpiece is significantly smaller than the cutter and near the spindle centerline, then small angle approximations lead to a simpler, more intuitive result given by Equation E.8.

$$\begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} = \frac{d \cdot w \cdot h_{max}}{s \cos(\theta_{min})} \begin{bmatrix} p_r \dfrac{2(\theta_2 - \theta_1) + \left(\sin(2\theta_2) - \sin(2\theta_1)\right)}{4\left(\sin(\theta_2) - \sin(\theta_1)\right)} - \dfrac{p_t}{\cos(\lambda)} \dfrac{\sin(\theta_2) + \sin(\theta_1)}{2} \\ p_r \dfrac{\sin(\theta_2) + \sin(\theta_1)}{2} + \dfrac{p_t}{\cos(\lambda)} \dfrac{2(\theta_2 - \theta_1) + \left(\sin(2\theta_2) - \sin(2\theta_1)\right)}{4\left(\sin(\theta_2) - \sin(\theta_1)\right)} \\ p_r \tan(\lambda) \end{bmatrix} \tag{E.7}$$

$$\begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} \cong \frac{d \cdot w \cdot h_{max}}{s} \begin{bmatrix} p_r - \dfrac{p_t}{\cos(\lambda)} \dfrac{\theta_2 + \theta_1}{2} \\ p_r \dfrac{\theta_2 + \theta_1}{2} + \dfrac{p_t}{\cos(\lambda)} \\ p_r \tan(\lambda) \end{bmatrix} \tag{E.8}$$

We can use the machining process model to estimate the dynamic stiffness of the machine required for chatter-free machining. Regenerative chatter occurs when, through the response of the structure, undulations left on the part by the previous tooth cause greater undulations left by subsequent teeth. Chatter generally occurs near a modal frequency where the dynamic stiffness is governed by the damping in that mode. Instability occurs when the tool-to-work dynamic stiffness of the structure and the process add to zero, which can occur for dynamic systems. Another way to think about the phenomenon is that the process adds negative damping to the structure. A detailed model of the phenomenon is

quite complex because the direction of the force variation is not aligned to the vibration of the mode and neither one may align to the principal axes of the machine. As a simple estimate, we will assume that the vibration of the mode is in the direction of the surface normal where the chip thickness is maximum. Then the process stiffness is simply the derivative of (E.7) or (E.8) with respect to h_{max}. Using the simpler of the two for example, Equation E.9 shows that the process stiffness is independent of the chip thickness.

$$\begin{bmatrix} k_x \\ k_y \\ k_z \end{bmatrix} \cong \frac{d \cdot w}{s} \begin{bmatrix} p_r - \frac{p_t}{\cos(\lambda)} \frac{\theta_2 + \theta_1}{2} \\ p_r \frac{\theta_2 + \theta_1}{2} + \frac{p_t}{\cos(\lambda)} \\ p_r \tan(\lambda) \end{bmatrix} \tag{E.9}$$

To develop greater intuition about the parameters involved, the previous relationships are further simplified and summarized in Equation E.10. Any of the terms on the left hand side may limit the productivity of the machine, but preferably the spindle power should limit the metal removal rate through the specific power. The axis force will not be a limitation if it is sized for the maximum power at the minimum surface speed expected for production machining. The surface speed is dependent on the materials, the tooling and the culture where the machine will be used. It might range from 100 m/min (300 ft/min) for conventional milling of steel to 1000 m/min (3000 ft/min) for high speed milling of steel and cast iron and up to 5000 m/min (15000 ft/min) for aluminum. The spindle torque below base speed is constant and will limit production for large radius tools that cannot run at higher speeds where the power is constant (power decreases with speed below base speed). For electric motors, it is more expensive to buy torque than it is to buy speed and often a transmission ratio is required to provide adequately high torque and low base speed. Equation E.11 provides a way to estimate the base speed from the minimum surface speed and the maximum tool radius expected for production machining. A direct drive motor starts to be practical for a base speed of order 2000 rpm or approximately 1000 ft/min surface speed per inch of cutter radius.

$$\begin{aligned} k_i &\cong 4.8\tau_s \frac{d \cdot w}{s} \qquad i = x, y, z \\ F_i &\cong k_i \cdot h_{\max} \\ T_s &\cong F_i \cdot r \\ P_s &\cong F_i \cdot V_s \end{aligned} \tag{E.10}$$

$$\omega_{base} = \frac{P_s}{T_s} \cong \left(\frac{V_s}{r} \right)_{\min} \tag{E.11}$$

F Friction and Backlash in Servo Mechanisms

Friction and backlash are common sources of nonlinearity in a servo mechanism. Backlash is usually the more serious problem associated with geared transmissions and lead screws. Friction and backlash may lead to: limit cycling for systems with output position feedback, unacceptable position errors for systems with motor position feedback, uneven following error, and possibly dynamic instability problems. Computer simulation is a convenient way to test the effects of nonlinearity in a dynamic system. Using this approach, we will simulate a typical servo mechanism that includes friction and backlash in the model. We will optimize the controller assuming linear behavior and then compare its performance with varying degrees of friction and backlash.

F.1 Developing the Dynamic Model

Figure F-1 shows the structure of the control system having a current loop within a velocity loop within a position loop. To simplify the model, we will assume that the electrical dynamics are sufficiently fast to safely ignore. We will develop the system model using state representation to facilitate simulating the nonlinear system.[1] This also eliminates much of the algebra associated with Laplace transform techniques for linear systems. The most convenient variables to represent the current state of the system are the positions and velocities of the motor and the axis. Please refer to Table F-1 for a description of all the symbols used in the system model.

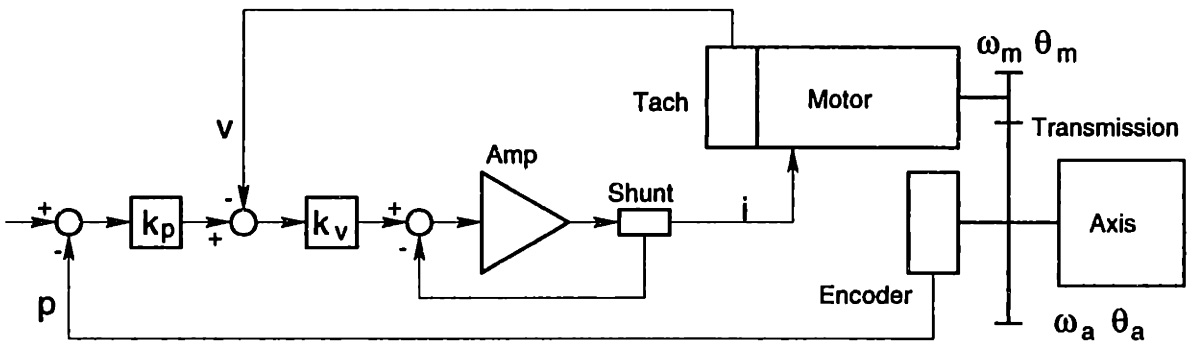


Figure F-1 This block diagram shows the components, signals and state variables used in the simulation. The motor and the axis both contain inertia and viscous friction terms. The transmission has a velocity ratio, compliance, backlash and friction.

Equation F.1 is the equation of motion for the motor with inertia and viscous drag terms on the left and torque terms on the right. The over dot represents differentiation with respect to time. Equation F.2 is the equation of motion for the axis. The motor and axis are physically connected by the transmission, and Equation F.3 is the linear model that

[1] The state representation of a dynamic system is expressed as a system of first order differential equations in terms of the state variables. The number of equations is always equal to the number of state variables.

describes its compliance. Later, we will alter this model to include nonlinear effects. Although trivial, Equation F.4 provides the relationship between state variables required for the simulation. These four equations are the state equations for the open-loop linear model. Equation F.5 is the control law that provides the feedback shown in Figure F-1.

$$\left(J_m + J_t\right)\dot{\omega}_m + b_m\,\omega_m = k_m\,i + T_m \tag{F.1}$$

$$J_a\,\dot{\omega}_a + b_a\,\omega_a = T_a + T_{in} \tag{F.2}$$

Symbol	Value	Description
θ_m	(state variable) r	Motor angular position
ω_m	(state variable) r/s	Motor angular velocity
θ_a	(state variable) r	Axis angular position
ω_a	(state variable) r/s	Axis angular velocity
θ_{in}	(input variable) r	Commanded input angle to the controller
T_{in}	(input variable) N-m	Disturbance input torque applied to the axis
T_m	(variable) N-m	Transmission torque at motor
T_a	(variable) N-m	Transmission torque at axis
p	(signal) count	Position feedback signal
v	(signal) V	Velocity feedback signal
k_p	0.0143 V/count	Position loop gain
k_v	25	Velocity loop gain
k_c	1 A/V	Current loop gain
k_{tach}	0.3 V-s/r	Tachometer gain
k_e	572,958 count/r	Encoder gain
k_m	0.34 N-m/A	Motor torque constant
J_m	0.001 kg-m^2	Motor mass moment of inertia
b_m	0.0008 N-m-s/r	Motor viscous drag constant
J_a	62.5 kg-m^2	Axis mass moment of inertia (maximum)
b_a	10 N-m-s/r	Axis viscous drag constant
R	180	Transmission ratio, ω_m/ω_a
k_t	5e7 N-m/r	Transmission stiffness
J_t	0.001 kg-m^2	Transmission mass moment of inertia at the input
T_f	(parameter) N-m	Transmission frictional torque at the output
θ_b	(parameter) r	Transmission backlash at the output

Table F-1 List of symbols and parameter values used in the system model. Parameter values used for the amplifier, tachometer, encoder and motor are typical but inconsequential to the simulations.

$$T_a = k_t\left(\frac{\theta_m}{R} - \theta_a\right) \qquad T_m = \frac{-T_a}{R} \tag{F.3}$$

$$\omega_m = \dot{\theta}_m \qquad \omega_a = \dot{\theta}_a \tag{F.4}$$

$$i = k_c\, k_v\left[k_p\, k_e\left(\theta_{in} - \theta_a\right) - k_{tach}\,\omega_m\right] \tag{F.5}$$

It is convenient for design purposes to assemble the state equations for the linear open-loop system into the matrix equation given by Equation F.6. The two inputs to the open loop system are the motor current i and the disturbance torque T_{in} applied to the axis. By applying the control law (F.5), the terms that multiply state variables become part of the closed loop system matrix **A** and the others go into the input vector **B**. These are defined in Equation F.7. Only θ_{in} and T_{in} remain as inputs to the closed loop system.

$$\frac{d}{dt}\begin{bmatrix}\theta_m\\ \omega_m\\ \theta_a\\ \omega_a\end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0\\ \dfrac{-k_t}{R^2 J_{m+t}} & \dfrac{-b_m}{J_{m+t}} & \dfrac{k_t}{R\,J_{m+t}} & 0\\ 0 & 0 & 0 & 1\\ \dfrac{k_t}{R\,J_a} & 0 & \dfrac{-k_t}{J_a} & \dfrac{-b_a}{J_a}\end{bmatrix}\cdot\begin{bmatrix}\theta_m\\ \omega_m\\ \theta_a\\ \omega_a\end{bmatrix} + \begin{bmatrix}0\\ \dfrac{k_m}{J_{m+t}}\,i\\ 0\\ \dfrac{1}{J_a}\,T_{in}\end{bmatrix} \tag{F.6}$$

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0\\ \dfrac{-k_t}{R^2 J_{m+t}} & \dfrac{-b_m - k_m\, k_c\, k_v\, k_{tach}}{J_{m+t}} & \dfrac{k_t - k_m\, k_c\, k_v\, k_p\, k_e\, R}{R\,J_{m+t}} & 0\\ 0 & 0 & 0 & 1\\ \dfrac{k_t}{R\,J_a} & 0 & \dfrac{-k_t}{J_a} & \dfrac{-b_a}{J_a}\end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix}0\\ \dfrac{k_m\, k_c\, k_v\, k_p\, k_e}{J_{m+t}}\,\theta_{in}\\ 0\\ \dfrac{1}{J_a}\,T_{in}\end{bmatrix} \tag{F.7}$$

The eigenvalues of the system matrix are the roots of the characteristic equation (or poles of the transfer function). The eigenvalues describe the dynamics of the system. Calculating the eigenvalues numerically is a trivial operation in a mathematics program such as Matlab™. Plotting the locations of the roots in the complex plane is a useful technique to visualize the behavior of a system as a function of one parameter. Known as a root locus diagram, Figure F-2 shows the root locations as the velocity loop gain varies from 2.5 to

50 with zero position loop gain. The complex pole pair represents the mechanical resonance and the single real pole represents the main dynamics of the velocity loop. The maximum damping in the resonance occurs when the complex pole pair is the farthest left of zero, which corresponds to $k_v = 25$. We will use this value for the nonlinear simulations.

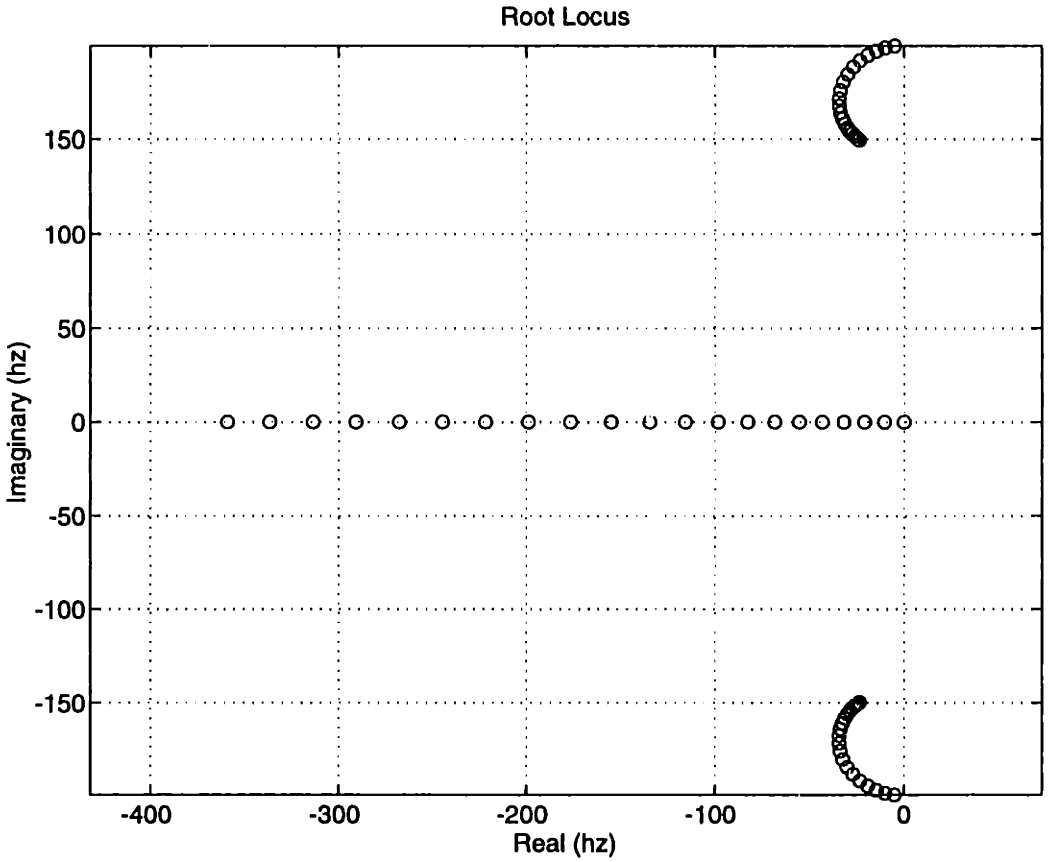


Figure F-2 Root locations in the complex plane as a function of k_v for $k_p = 0$. The maximum damping occurs in the resonance when the complex pole pair is the farthest left of zero ($k_v = 25$).

The root locus as a function of the position loop gain (with constant $k_v = 25$) indicates a range of behavior for the closed loop system. Figure F-3 shows the behavior up to the point of instability when the locus crosses into the right half plane, which corresponds to $k_p = 0.057$ V/count. Another way to interpret the instability is as negative damping where the servo adds energy to the mechanical resonance rather than dissipating energy. The point of instability occurs when the servo motor responds to the position feedback with the exact amount of torque required to remain rigid. For this locked-rotor condition, the static stiffness of the system is equal to the transmission stiffness.[1]

[1] To determine the static stiffness of the system, assume a unit deflection and calculate the resulting torque due to the feedback loop, which is ($k_e\ k_p\ k_v\ k_c\ k_m\ R$).

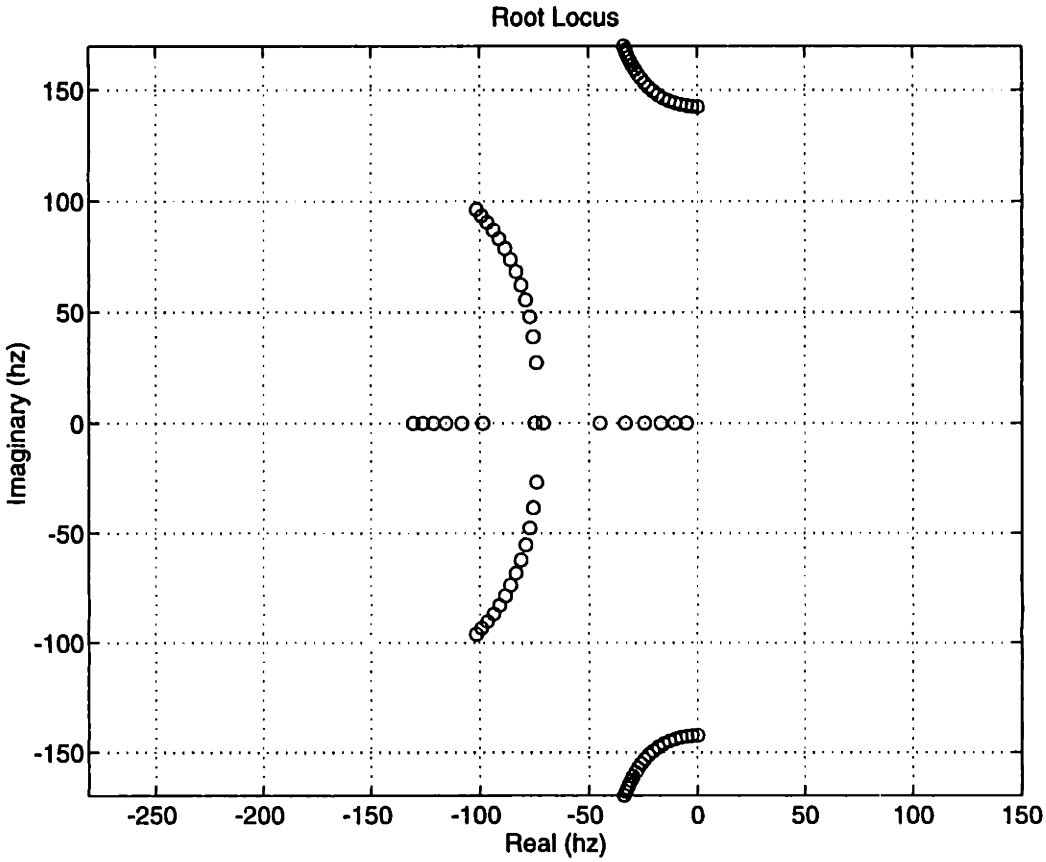


Figure F-3 Root locations in the complex plane as a function of k_p for $k_v = 25$. Instability occurs when the complex pole pair corresponding to the mechanical resonance crosses zero ($k_p = 0.057$ V/count).

The choice of position loop gain is a compromise since the static stiffness increases with k_p while stability and damping decrease. To make a good judgment, we need to know how the dynamic stiffness varies throughout a range of frequencies. By exciting the input T_{in} with a unit sinusoid $e^{j\omega}$, the complex amplitude of the axis position will represent the dynamic compliance of the system for any given position loop gain k_p. Equation F.8 gives the complex amplitude of the state vector **x** as a function of excitation frequency ω. Figure F-4 shows the magnitude and phase of the dynamic compliance for $k_p = 0.0143$ V/count. Although the static compliance is four times that of the transmission, this is easy to reduce with additional compensation such as a lag filter. Decreasing k_p further does not significantly improve the resonance. For example, using $k_p = 0.0057$ V/count results in a static compliance of 10 and a compliance at the resonance of approximately 1.8. The former gain seems to be a little better compromise so we will use $k_p = 0.0143$ V/count for the nonlinear simulations. This system has two real poles at -08 and -33 Hz and a complex pole pair at $-31 \pm 163j$ Hz. The poles of an unloaded axis would be different. By reducing the axis moment of inertia by a factor of 10, the real poles move to -163 and -28 Hz and the complex pole pair moves to $-6 \pm 464j$ Hz.

$$\mathbf{x}(j\omega) = (j\omega\,\mathbf{I} - \mathbf{A})^{-1} \cdot \mathbf{B} \tag{F.8}$$

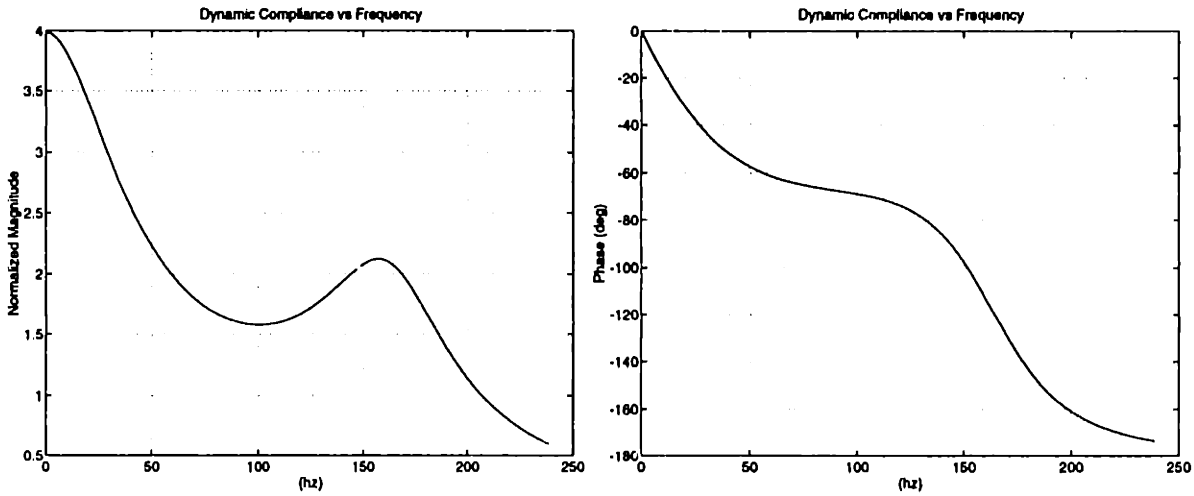


Figure F-4 Magnitude and phase plots of the dynamic compliance for k_p = 0.0143 V/count. The magnitude is normalized to the compliance of the transmission.

The final step in developing the system model is to add friction and backlash to the model of the transmission. Figure F-5 shows a graphical representation of backlash and compliance in the transmission. The three line segments in the graph are represented by three conditions in Equation F.9. It is easy to show that this reduces to a simple compliance for the special case of zero backlash. Given by Equation F.10, the model used for friction is a constant, resistive torque applied directly to the motor shaft.[1] This model is most applicable to single stage transmissions where the friction occurs on the motor side of the compliance. A conventional gear train would have friction distributed at each gear mesh.

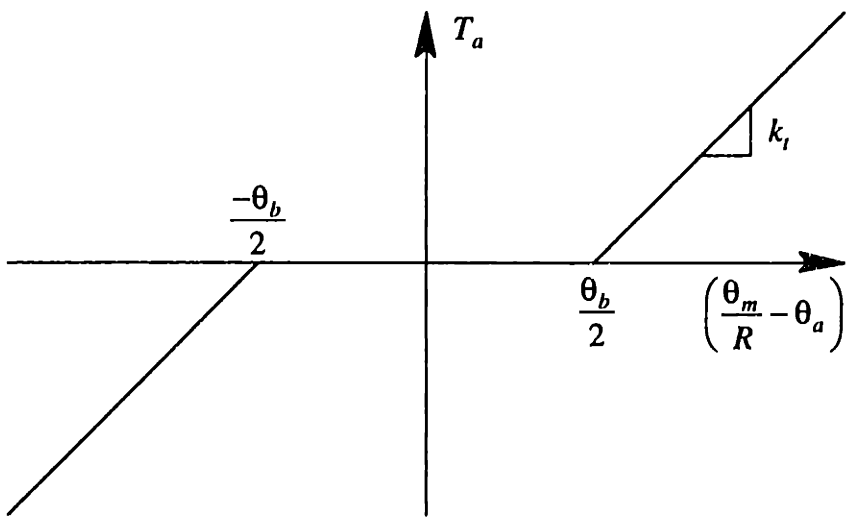


Figure F-5 This graph shows the output shaft torque for a transmission including compliance and backlash. The torque is a function of the angle difference between the input shaft, scaled by the ratio, and the output shaft. The backlash parameter θ_b is with respect to the output shaft.

[1] A saturation function was used instead of the signum function to improve the speed of numerical integration. The slope of the saturation function increases the effective velocity loop gain by a large factor.

$$T_a = \begin{cases} k_t\left(\frac{\theta_m}{R} - \theta_a - \frac{\theta_b}{2}\right) & if\left(\frac{\theta_m}{R} - \theta_a\right) > \frac{\theta_b}{2} \\ 0 & if\left|\frac{\theta_m}{R} - \theta_a\right| \le \frac{\theta_b}{2} \\ k_t\left(\frac{\theta_m}{R} - \theta_a + \frac{\theta_b}{2}\right) & if\left(\frac{\theta_m}{R} - \theta_a\right) < \frac{-\theta_b}{2} \end{cases} \quad \text{(F.9)}$$

$$T_m = \frac{-T_a - T_f \,\text{sign}(\omega_m)}{R} \quad \text{(F.10)}$$

F.2 Simulation Results

The nonlinear system model was simulated in Matlab using a built-in Runga-Kutta algorithm (ODE45). The input to the system model was a sinusoidal disturbance torque applied to the axis. The amplitude used was 100 N-m at three different frequencies, 2.5, 20 and 160 Hz, to explore different regimes of the response. Figure F-6 shows the simulations for two levels of backlash, 1 μr (left) and 10 μr (right). The simulation shows that backlash within the control loop is destabilizing and ultimately will limit the performance. Backlash outside the control loop only causes an uncertain output position. Friction, on the other hand, is normally dissipative.[1] Figure F-7 shows simulations for two magnitudes of friction. On the left side, the friction is one-half of the excitation and the effect is to stiffen the servo. On the right side, the friction is equal to the excitation. The result should be zero response at the servo motor, but the saturation function used in the simulation (rather than the signum function in Equation F.10) allowed the motor to turn slowly. The middle plot shows nicely that damping is below optimum when friction locks up the servo. This occurs at low frequency too but the effect is less pronounced (observe the tiny ripple). The bottom plot shows a condition near resonance where friction is insufficient to prevent the motor response (the saturation function becomes saturated).

As a result of the nonlinear simulation, we see that backlash should be minimized whether inside or outside the control loop. Although friction appears not to be a serious issue from a stability perspective, it will limit the positioning accuracy. We can calculate the following error required to overcome a static frictional torque. Using the assumptions in this model and T_f = 100 N-m, the following error required is 6 μr (the servo has three times more static compliance than the transmission from Figure F-4).

[1] Stick slip is an exception caused when friction decreases with increasing velocity and when the frictional interface is moving, for example, brake squeal and violin strings.

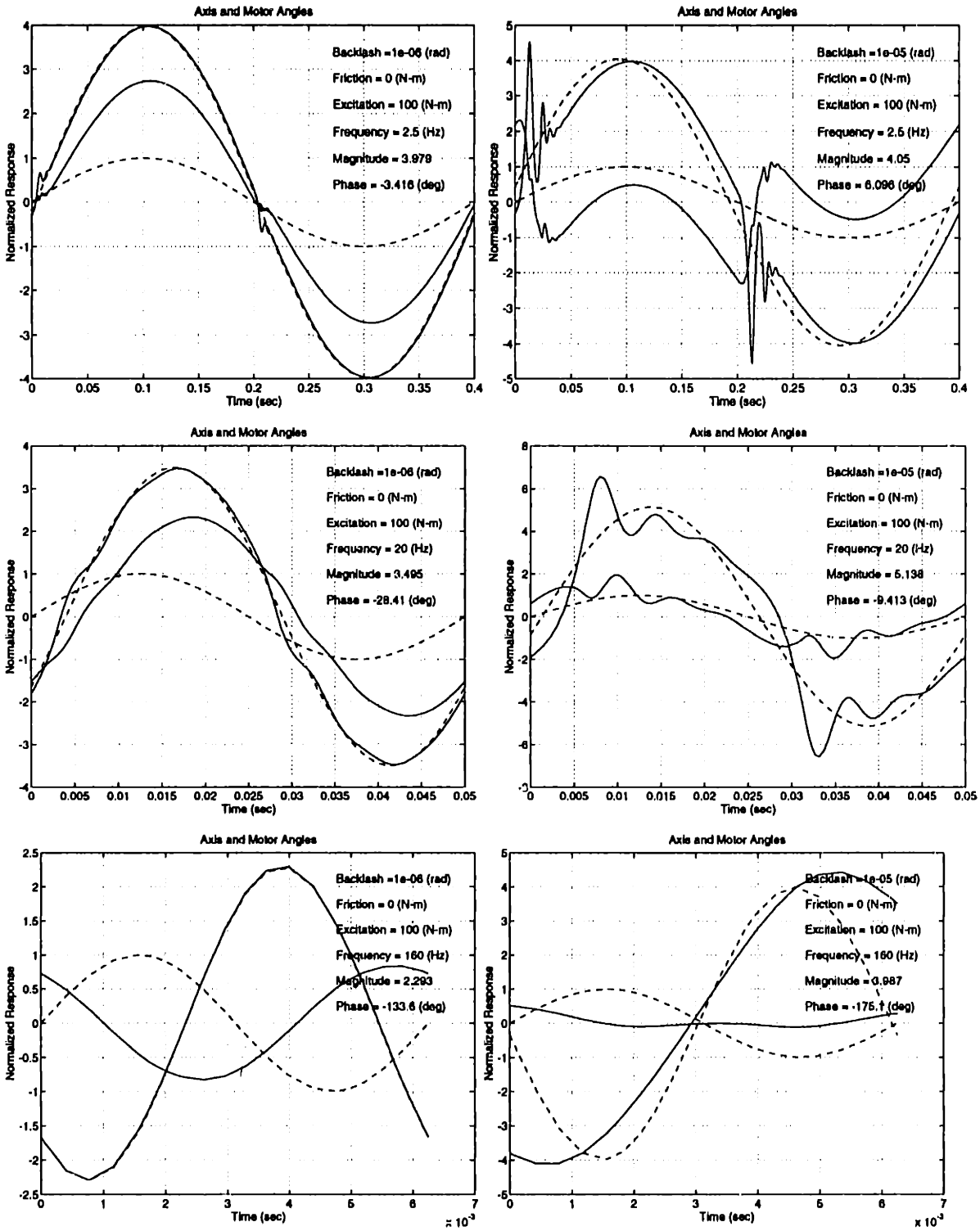


Figure F-6 Simulations for increasing backlash (left to right) and increasing frequency (top to bottom). Solid lines are the axis and motor responses to a sinusoidal torque excitation at the axis. These are normalized to the response of a spring equal to the transmission stiffness (unit sinusoid, broken line). The second broken line is the best-fit sinusoid to the axis response.

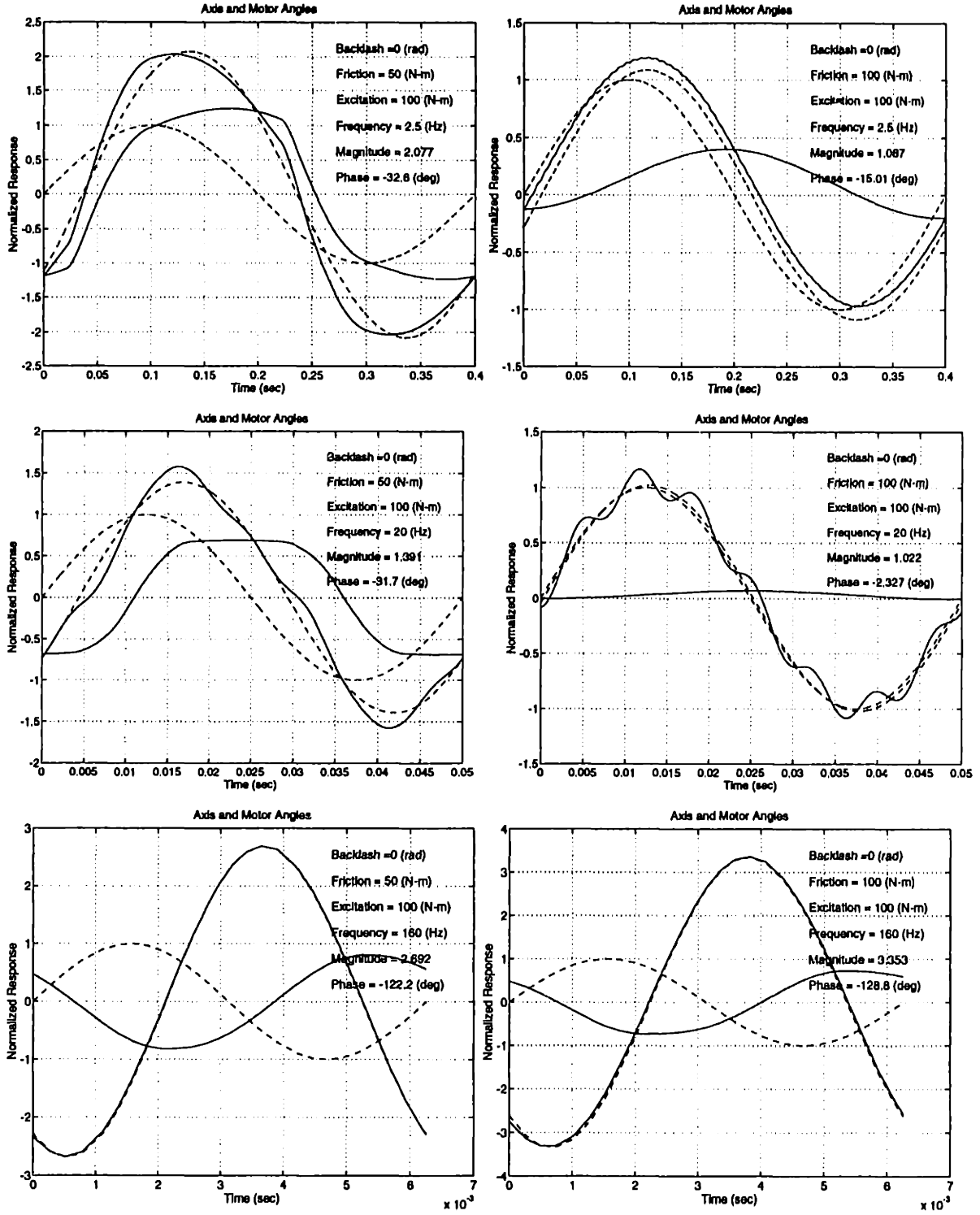


Figure F-7 Simulations for increasing friction (left to right) and increasing frequency (top to bottom). Solid lines are the axis and motor responses to a sinusoidal torque excitation at the axis. These are normalized to the response of a spring equal to the transmission stiffness (unit sinusoid, broken line). The second broken line is the best-fit sinusoid to the axis response.

intentionally blank

G AHP Spreadsheet and Configuration Drawings

The following tables show the full detail of the AHP spreadsheet used to evaluate the different configurations considered for the conceptual design of a horizontal machining center. Table G-1 shows the level 1 criteria and rankings computed from lower level input. Table G-2 shows the terminal criteria and the rankings entered for the level 1 criterion *Required Systems*. Table G-3, Table G-4 and Table G-5 shows the terminal criteria and the rankings entered for the level 1 criterion *Accuracy*. Table G-6 shows the terminal criteria and the rankings entered for the level 1 criterion *Productivity*. Table G-7 shows the terminal criteria and the rankings entered for the level 1 criterion *Manufacturing Cost*. Table G-8 shows the terminal criteria and the rankings entered for the level 1 criterion *Ergonomics*.

Figure G-1 through Figure G-14 show the first-cut design sketches that represent the different configurations considered for the conceptual design of a horizontal machining center. The rankings entered into the AHP spreadsheet are based on these designs.

AHP1			Decision	10	10	8	6	4
	Criteria Level 1 >		1.00	Req'd sys.	Accuracy	Productivity	Manuf. cost	Ergonomics
	Criteria Level 2 >			0.26	0.26	0.21	0.16	0.11
	Criteria Level 3 >							
I.D.	*Work*	*Tool*						
1	A, B	X, Y, Z	6.22	5.53	6.16	6.25	6.34	8.19
2	A, B	X, Z, Y	5.53	5.19	5.17	5.12	6.72	6.68
3	A, B	Y, X, Z	5.46	4.15	7.01	7.22	5.73	3.13
4	A, B	Y, Z, X	4.56	2.76	6.58	6.25	4.96	3.00
5	A, B	Z, X, Y	5.34	4.47	5.55	4.88	7.06	5.97
~~6~~	~~B~~	~~Z, Y, X~~						
7	X, A, B	Y, Z	6.75	4.97	7.29	8.14	6.67	8.41
8	X, B	Z, Y, A	8.14	7.85	8.26	8.53	7.90	8.18
9	Y, B	X, A, Z	7.49	7.91	7.72	7.94	6.20	7.13
10	Y, B	Z, X, A	7.53	8.44	8.00	6.93	6.73	6.83
11	Z, B	X, Y, A	7.93	8.54	7.70	7.48	7.50	8.71
12	Z, A, B	Y, X	5.78	3.51	7.77	8.02	6.15	4.52
~~13~~	~~X, Y, B~~	~~Z~~						
14	X, Z, B	Y, A	7.56	8.63	7.10	7.18	6.95	7.97
15	Y, X, B	Z, A	7.55	7.66	8.29	7.56	7.03	6.41
~~16~~	~~Y, Z, B~~	~~X~~						
17	Z, X, B	Y, A	7.50	8.63	7.10	7.00	6.95	7.79
~~18~~	~~Z, Y, B~~	~~X~~						
~~19~~	~~X, Y, Z, B~~							
~~20~~	~~X, Z, Y, B~~							
~~21~~	~~Y, X, Z, B~~							
~~22~~	~~Y, Z, X, B~~							
~~23~~	~~Z, X, Y, B~~							
~~24~~	~~Z, Y, X, B~~							

Table G-1 AHP spreadsheet showing only level 1 criteria and rankings computed from lower level input.

AHP1			Decision	10				
	Criteria Level 1 >		1.00	Req'd sys.	10	10	8	6
	Criteria Level 2 >			0.26	A-axis	Tool change	Work change	Chip flow
	Criteria Level 3 >				0.29	0.29	0.24	0.18
I.D.	*Work*	*Tool*						
1	A, B	X, Y, Z	6.22	5.53	5	8	3	8
2	A, B	X, Z, Y	5.53	5.19	5	7	3	7
3	A, B	Y, X, Z	5.46	4.15	5	3	3	8
4	A, B	Y, Z, X	4.56	2.76	5	1	3	5
5	A, B	Z, X, Y	5.34	4.47	5	7	3	3
~~6~~	~~B~~	~~Z, Y, X~~						
7	X, A, B	Y, Z	6.75	4.97	3	8	4	7
8	X, B	Z, Y, A	8.14	7.85	8	7	8	9
9	Y, B	X, A, Z	7.49	7.91	7	8	9	8
10	Y, B	Z, X, A	7.53	8.44	10	7	9	8
11	Z, B	X, Y, A	7.93	8.54	8	10	8	8
12	Z, A, B	Y, X	5.78	3.51	3	3	4	5
~~13~~	~~X, Y, B~~	~~Z~~						
14	X, Z, B	Y, A	7.56	8.63	10	10	7	7
15	Y, X, B	Z, A	7.55	7.66	10	5	10	7
~~16~~	~~Y, Z, B~~	~~X~~						
17	Z, X, B	Y, A	7.50	8.63	10	10	7	7
~~18~~	~~Z, Y, B~~	~~X~~						
~~19~~	~~X, Y, Z, B~~							
~~20~~	~~X, Z, Y, B~~							
~~21~~	~~Y, X, Z, B~~							
~~22~~	~~Y, Z, X, B~~							
~~23~~	~~Z, X, Y, B~~							
~~24~~	~~Z, Y, X, B~~							

Table G-2 AHP spreadsheet showing the terminal criteria and rankings entered for Required Systems.

AHP1			Decision	10				
	Criteria Level 1 >		1.00	Accuracy	10			
	Criteria Level 2 >			0.26	Stiffness	10	10	10
	Criteria Level 3 >				0.33	Cross sect.	Stiff. loop	Aspect ratio
						0.33	0.33	0.33
I.D.	*Work*	*Tool*						
1	A, B	X, Y, Z	6.22	6.16	6.54	5	7	8
2	A, B	X, Z, Y	5.53	5.17	5.94	5	6	7
3	A, B	Y, X, Z	5.46	7.01	7.88	7	7	10
4	A, B	Y, Z, X	4.56	6.58	7.96	9	7	8
5	A, B	Z, X, Y	5.34	5.55	6.60	4	9	8
~~6~~	~~B~~	~~Z, Y, X~~						
7	X, A, B	Y, Z	6.75	7.29	7.65	7	8	8
8	X, B	Z, Y, A	8.14	8.26	9.32	9	10	9
9	Y, B	X, A, Z	7.49	7.72	7.65	7	8	8
10	Y, B	Z, X, A	7.53	8.00	7.65	8	7	8
11	Z, B	X, Y, A	7.93	7.70	8.62	8	10	8
12	Z, A, B	Y, X	5.78	7.77	9.32	9	9	10
~~13~~	~~X, Y, B~~	~~Z~~						
14	X, Z, B	Y, A	7.56	7.10	8.28	7	9	9
15	Y, X, B	Z, A	7.55	8.29	8.65	8	9	9
~~16~~	~~Y, Z, B~~	~~X~~						
17	Z, X, B	Y, A	7.50	7.10	8.28	7	9	9
~~18~~	~~Z, Y, B~~	~~X~~						
~~19~~	~~X, Y, Z, B~~							
~~20~~	~~X, Z, Y, B~~							
~~21~~	~~Y, X, Z, B~~							
~~22~~	~~Y, Z, X, B~~							
~~23~~	~~Z, X, Y, B~~							
~~24~~	~~Z, Y, X, B~~							

Table G-3 AHP spreadsheet showing the terminal criteria and rankings entered for Accuracy/Stiffness.

AHP1			Decision	10				
	Criteria Level 1 >		1.00	Accuracy	10			
	Criteria Level 2 >			0.26	Align. princ.	10	10	
	Criteria Level 3 >				0.33	Abbé	Bryan	
						0.50	0.50	
I.D.	*Work*	*Tool*						
1	A, B	X, Y, Z	6.22	6.16	6.00	6	6	
2	A, B	X, Z, Y	5.53	5.17	4.00	4	4	
3	A, B	Y, X, Z	5.46	7.01	6.48	7	6	
4	A, B	Y, Z, X	4.56	6.58	6.48	6	7	
5	A, B	Z, X, Y	5.34	5.55	4.47	4	5	
~~6~~	~~B~~	~~Z, Y, X~~						
7	X, A, B	Y, Z	6.75	7.29	7.48	8	7	
8	X, B	Z, Y, A	8.14	8.26	8.00	8	8	
9	Y, B	X, A, Z	7.49	7.72	7.48	8	7	
10	Y, B	Z, X, A	7.53	8.00	8.49	9	8	
11	Z, B	X, Y, A	7.93	7.70	7.00	7	7	
12	Z, A, B	Y, X	5.78	7.77	6.93	6	8	
~~13~~	~~X, Y, B~~	~~Z~~						
14	X, Z, B	Y, A	7.56	7.10	7.00	7	7	
15	Y, X, B	Z, A	7.55	8.29	8.00	8	8	
~~16~~	~~Y, Z, B~~	~~X~~						
17	Z, X, B	Y, A	7.50	7.10	7.00	7	7	
~~18~~	~~Z, Y, B~~	~~X~~						
~~19~~	~~X, Y, Z, B~~							
~~20~~	~~X, Z, Y, B~~							
~~21~~	~~Y, X, Z, B~~							
~~22~~	~~Y, Z, X, B~~							
~~23~~	~~Z, X, Y, B~~							
~~24~~	~~Z, Y, X, B~~							

Table G-4 AHP spreadsheet showing the terminal criteria and rankings entered for Accuracy/Alignment Principles.

AHP1			Decision	10				
	Criteria Level 1 >		1.00	Accuracy	10			
	Criteria Level 2 >			0.26	Stability	10	8	6
	Criteria Level 3 >				0.33	Moving load	Variable load	Thermal
						0.42	0.33	0.25
I.D.	*Work*	*Tool*						
1	A, B	X, Y, Z	6.22	6.16	5.97	6	8	4
2	A, B	X, Z, Y	5.53	5.17	5.80	4	8	7
3	A, B	Y, X, Z	5.46	7.01	6.73	7	8	5
4	A, B	Y, Z, X	4.56	6.58	5.53	5	8	4
5	A, B	Z, X, Y	5.34	5.55	5.80	4	8	7
~~6~~	~~B~~	~~Z, Y, X~~						
7	X, A, B	Y, Z	6.75	7.29	6.76	8	6	6
8	X, B	Z, Y, A	8.14	8.26	7.57	7	8	8
9	Y, B	X, A, Z	7.49	7.72	8.02	8	10	6
10	Y, B	Z, X, A	7.53	8.00	7.88	7	10	7
11	Z, B	X, Y, A	7.93	7.70	7.57	7	8	8
12	Z, A, B	Y, X	5.78	7.77	7.27	8	6	8
~~13~~	~~X, Y, B~~	~~Z~~						
14	X, Z, B	Y, A	7.56	7.10	6.19	7	4	9
15	Y, X, B	Z, A	7.55	8.29	8.24	8	8	9
~~16~~	~~Y, Z, B~~	~~X~~						
17	Z, X, B	Y, A	7.50	7.10	6.19	7	4	9
~~18~~	~~Z, Y, B~~	~~X~~						
~~19~~	~~X, Y, Z, B~~							
~~20~~	~~X, Z, Y, B~~							
~~21~~	~~Y, X, Z, B~~							
~~22~~	~~Y, Z, X, B~~							
~~23~~	~~Z, X, Y, B~~							
~~24~~	~~Z, Y, X, B~~							

Table G-5 AHP spreadsheet showing the terminal criteria and rankings entered for Accuracy/Stability.

AHP1			Decision	8				
	Criteria Level 1 >		1.00	Productivity	10	8	6	4
	Criteria Level 2 >			0.21	Dyn. stiff.	Nat'l freq.	Axis accel.	Reliability
	Criteria Level 3 >				0.36	0.29	0.21	0.14
I.D.	*Work*	*Tool*						
1	A, B	X, Y, Z	6.22	6.25	6	6	6	8
2	A, B	X, Z, Y	5.53	5.12	8	4	4	4
3	A, B	Y, X, Z	5.46	7.22	6	8	8	8
4	A, B	Y, Z, X	4.56	6.25	7	7	6	4
5	A, B	Z, X, Y	5.34	4.88	7	4	4	4
~~6~~	~~B~~	~~Z, Y, X~~						
7	X, A, B	Y, Z	6.75	8.14	7	9	8	10
8	X, B	Z, Y, A	8.14	8.53	9	9	7	9
9	Y, B	X, A, Z	7.49	7.94	7	9	9	7
10	Y, B	Z, X, A	7.53	6.93	6	8	8	6
11	Z, B	X, Y, A	7.93	7.48	8	7	7	8
12	Z, A, B	Y, X	5.78	8.02	9	9	9	4
~~13~~	~~X, Y, B~~	~~Z~~						
14	X, Z, B	Y, A	7.56	7.18	8	7	7	6
15	Y, X, B	Z, A	7.55	7.56	9	8	8	4
~~16~~	~~Y, Z, B~~	~~X~~						
17	Z, X, B	Y, A	7.50	7.00	8	7	7	5
~~18~~	~~Z, Y, B~~	~~X~~						
~~19~~	~~X, Y, Z, B~~							
~~20~~	~~X, Z, Y, B~~							
~~21~~	~~Y, X, Z, B~~							
~~22~~	~~Y, Z, X, B~~							
~~23~~	~~Z, X, Y, B~~							
~~24~~	~~Z, Y, X, B~~							

Table G-6 AHP spreadsheet showing the terminal criteria and rankings entered for Productivity.

AHP1			Decision	6					
	Criteria Level 1 >		1.00	Manuf. cost	10	10	8	6	4
	Criteria Level 2 >			0.16	Modularity	Exact constr.	Machining	Assembly	Floor space
	Criteria Level 3 >				0.26	0.26	0.21	0.16	0.11
I.D.	*Work*	*Tool*							
1	A, B	X, Y, Z	6.22	6.34	7	6	6	6	6.89
2	A, B	X, Z, Y	5.53	6.72	6	7	7	7	6.94
3	A, B	Y, X, Z	5.46	5.73	7	5	5	5	7.81
4	A, B	Y, Z, X	4.56	4.96	5	6	4	4	6.41
5	A, B	Z, X, Y	5.34	7.06	6	8	7	8	6.51
~~6~~	~~B~~	~~Z, Y, X~~							
7	X, A, B	Y, Z	6.75	6.67	8	6	6	6	7.94
8	X, B	Z, Y, A	8.14	7.90	7	8	8	8	9.92
9	Y, B	X, A, Z	7.49	6.20	7	6	5	6	8.06
10	Y, B	Z, X, A	7.53	6.73	7	6	6	7	9.62
11	Z, B	X, Y, A	7.93	7.50	7	7	8	8	8.42
12	Z, A, B	Y, X	5.78	6.15	7	6	4	7	9.26
~~13~~	~~X, Y, B~~	~~Z~~							
14	X, Z, B	Y, A	7.56	6.95	7	6	7	7	9.62
15	Y, X, B	Z, A	7.55	7.03	7	6	7	8	8.74
~~16~~	~~Y, Z, B~~	~~X~~							
17	Z, X, B	Y, A	7.50	6.95	7	6	7	7	9.62
~~18~~	~~Z, Y, B~~	~~X~~							
~~19~~	~~X, Y, Z, B~~								
~~20~~	~~X, Z, Y, B~~								
~~21~~	~~Y, X, Z, B~~								
~~22~~	~~Y, Z, X, B~~								
~~23~~	~~Z, X, Y, B~~								
~~24~~	~~Z, Y, X, B~~								

Table G-7 AHP spreadsheet showing the terminal criteria and rankings entered for Manufacturing Costs.

AHP1			Decision	4					
	Criteria Level 1 >		1.00	Ergonomics	10	10	8	10	8
	Criteria Level 2 >			0.11	Visibility	Pallet access	Spind. access	Maintenance	Safety
	Criteria Level 3 >				0.22	0.22	0.17	0.22	0.17
I.D.	*Work*	*Tool*							
1	A, B	X, Y, Z	6.22	8.19	10	5	10	8	10
2	A, B	X, Z, Y	5.53	6.68	8	4	8	7	8
3	A, B	Y, X, Z	5.46	3.13	1	3	4	5	6
4	A, B	Y, Z, X	4.56	3.00	1	4	5	3	5
5	A, B	Z, X, Y	5.34	5.97	7	7	5	5	6
~~6~~	~~B~~	~~Z, Y, X~~							
7	X, A, B	Y, Z	6.75	8.41	10	5	10	9	10
8	X, B	Z, Y, A	8.14	8.18	7	9	8	9	8
9	Y, B	X, A, Z	7.49	7.13	5	10	9	8	5
10	Y, B	Z, X, A	7.53	6.83	5	10	7	8	5
11	Z, B	X, Y, A	7.93	8.71	8	9	10	8	9
12	Z, A, B	Y, X	5.78	4.52	2	6	5	5	7
~~13~~	~~X, Y, B~~	~~Z~~							
14	X, Z, B	Y, A	7.56	7.97	8	9	8	7	8
15	Y, X, B	Z, A	7.55	6.41	5	10	7	6	5
~~16~~	~~Y, Z, B~~	~~X~~							
17	Z, X, B	Y, A	7.50	7.79	8	9	7	7	8
~~18~~	~~Z, Y, B~~	~~X~~							
~~19~~	~~X, Y, Z, B~~								
~~20~~	~~X, Z, Y, B~~								
~~21~~	~~Y, X, Z, B~~								
~~22~~	~~Y, Z, X, B~~								
~~23~~	~~Z, X, Y, B~~								
~~24~~	~~Z, Y, X, B~~								

Table G-8 AHP spreadsheet showing the terminal criteria and rankings entered for Ergonomics.

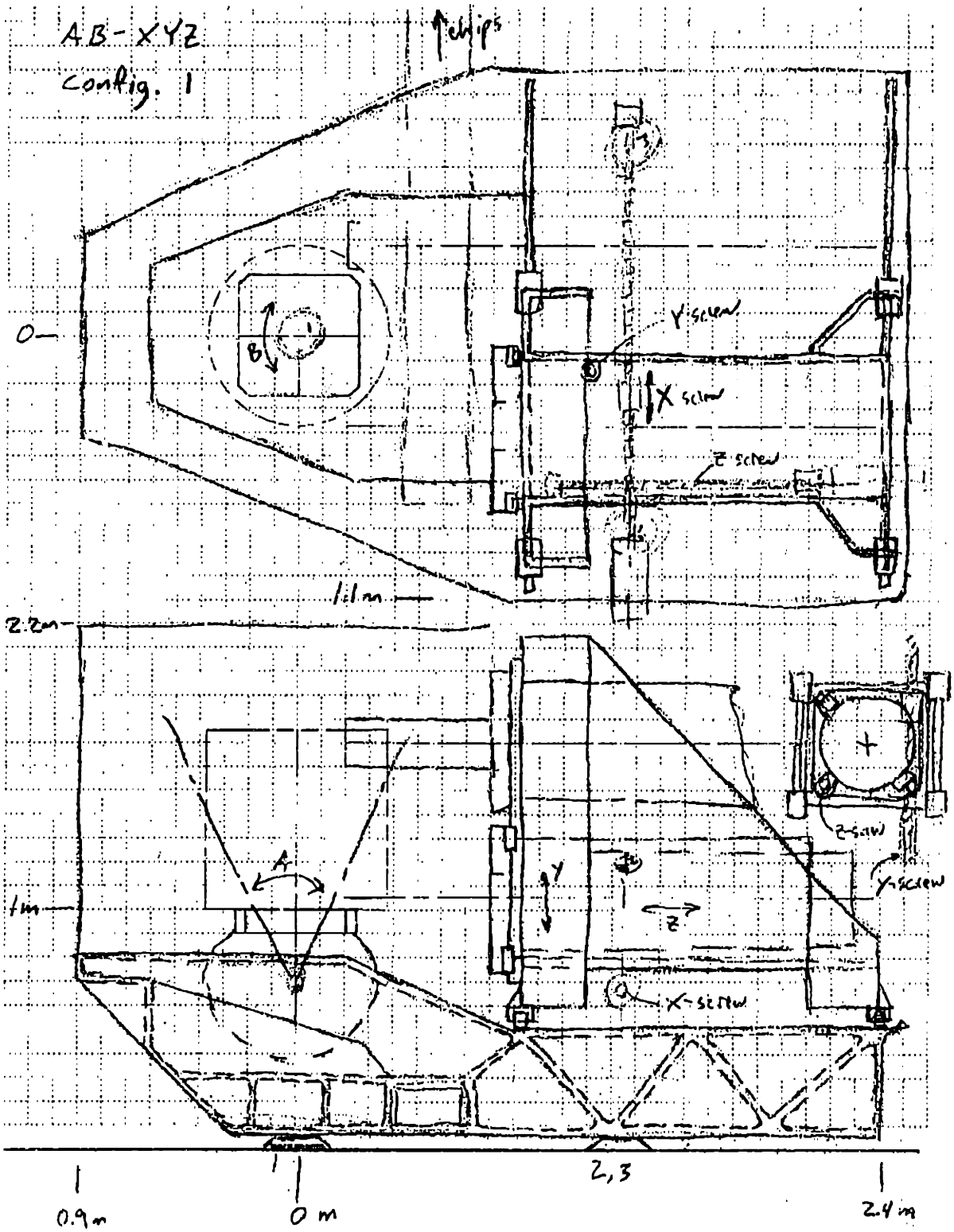


Figure G-1 Configuration 1, work over A and B, and tool over X, Y and Z.

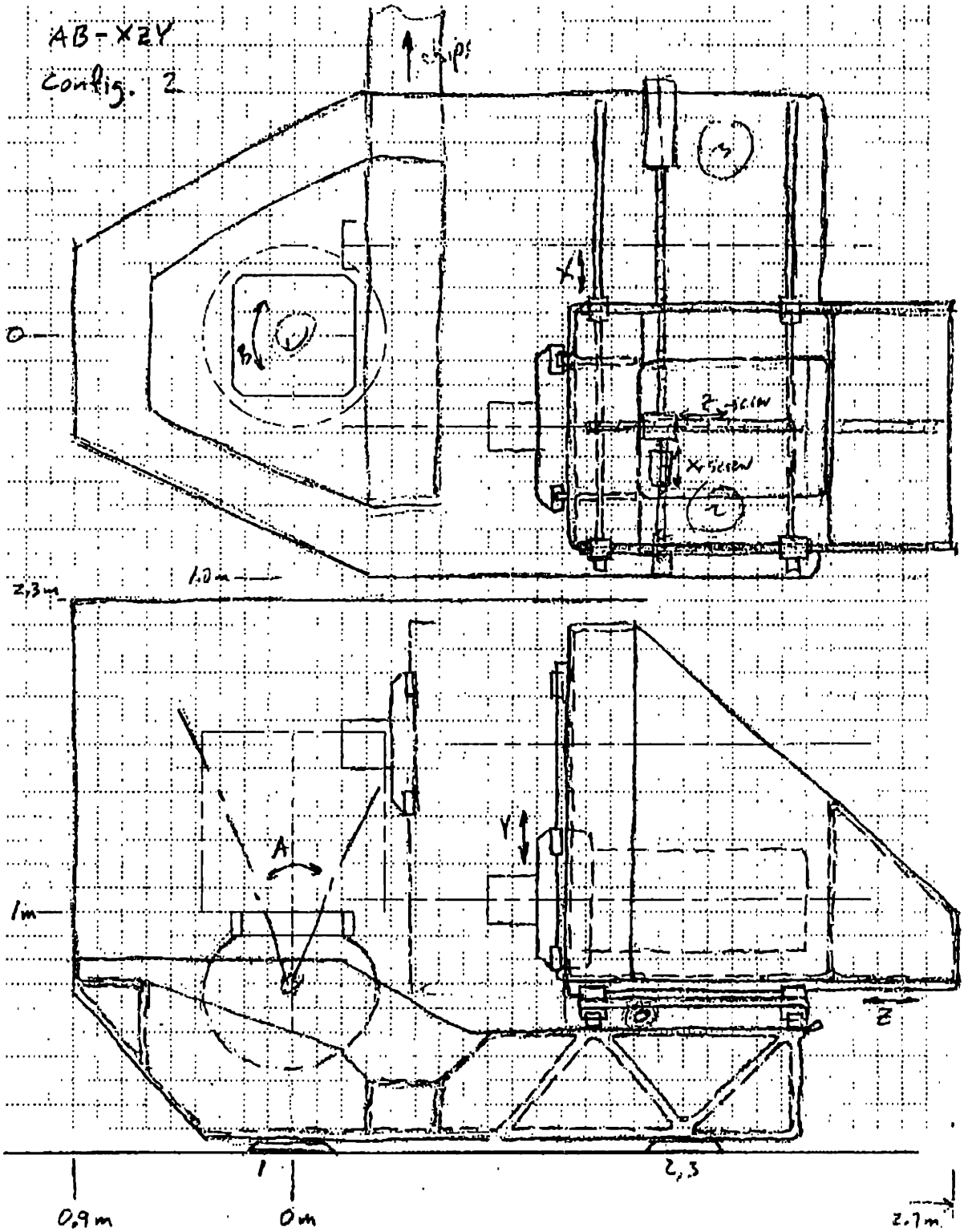


Figure G-2 Configuration 2, work over A and B, and tool over X, Z and Y.

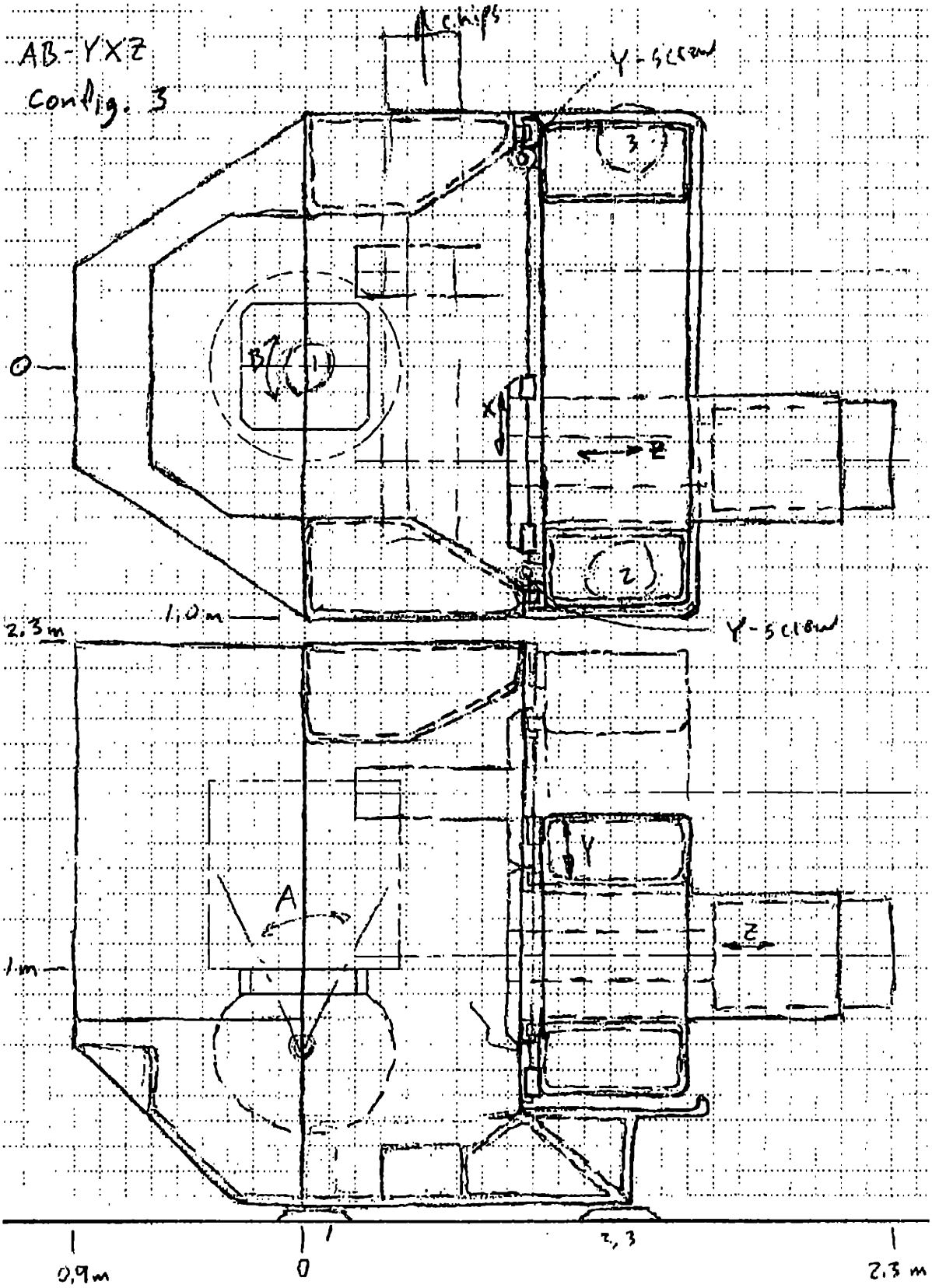


Figure G-3 Configuration 3, work over A and B, and tool over Y, X and Z.

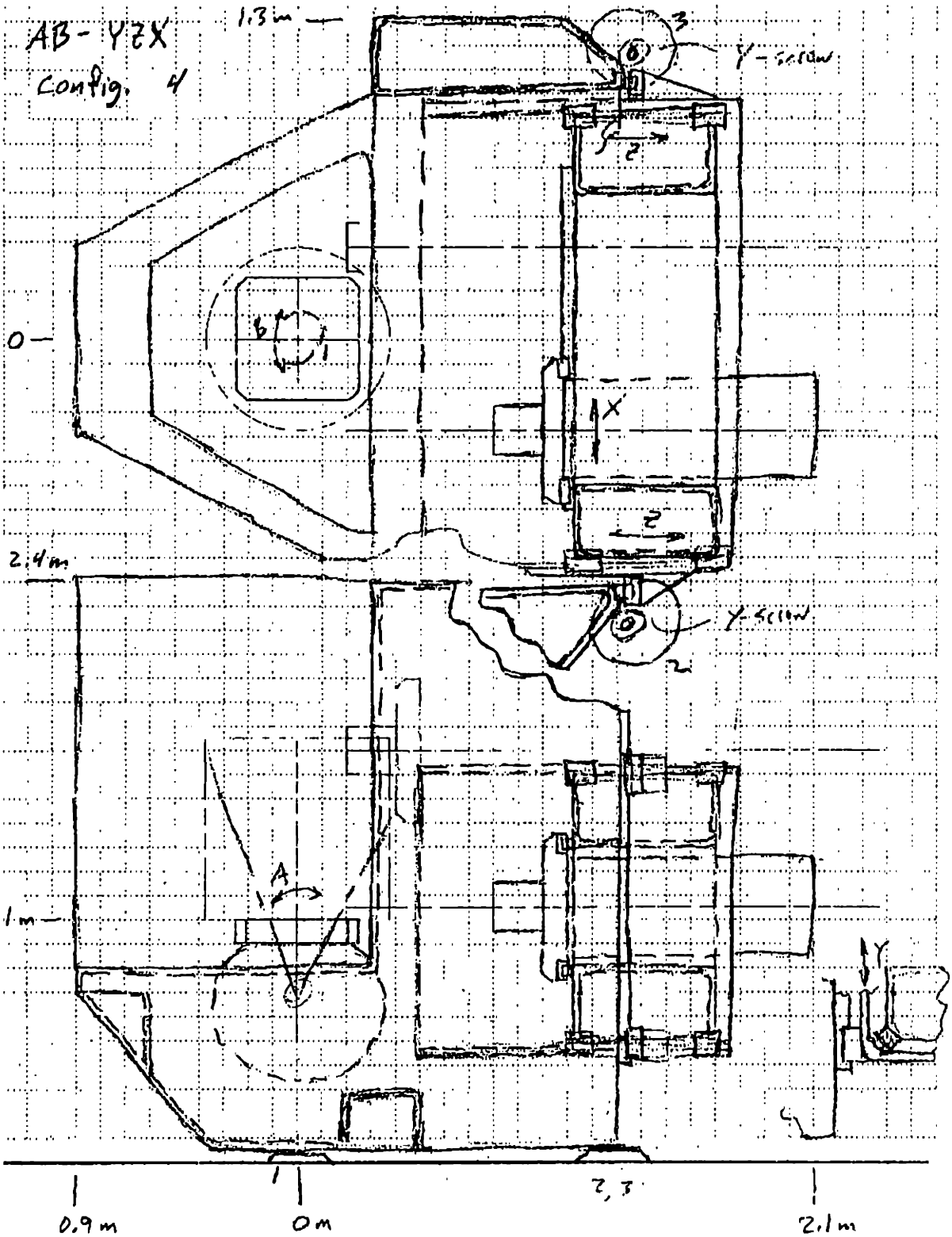


Figure G-4 Configuration 4, work over A and B, and tool over Y, Z and X.

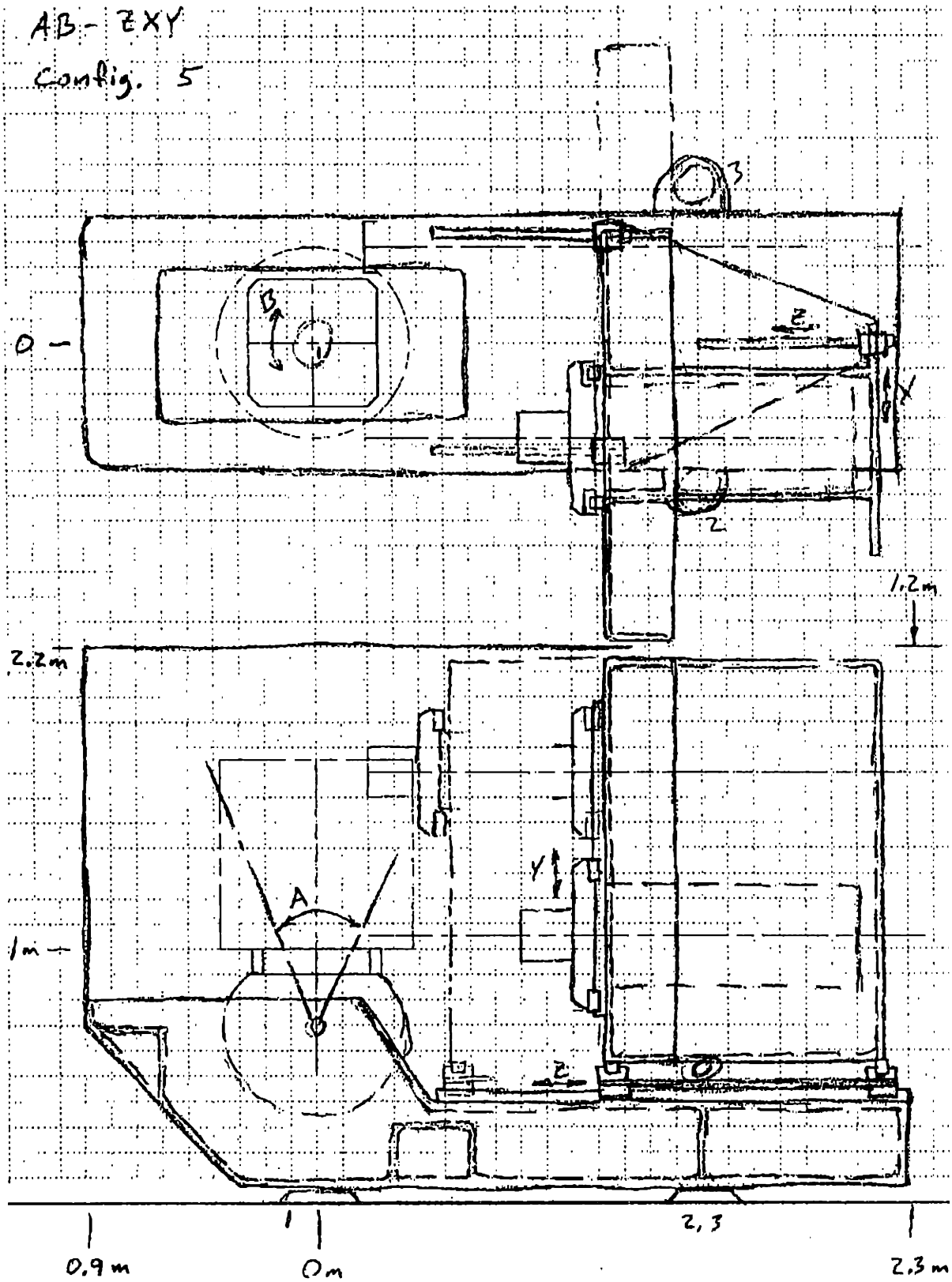


Figure G-5 Configuration 5, work over A and B, and tool over Z, X and Y.

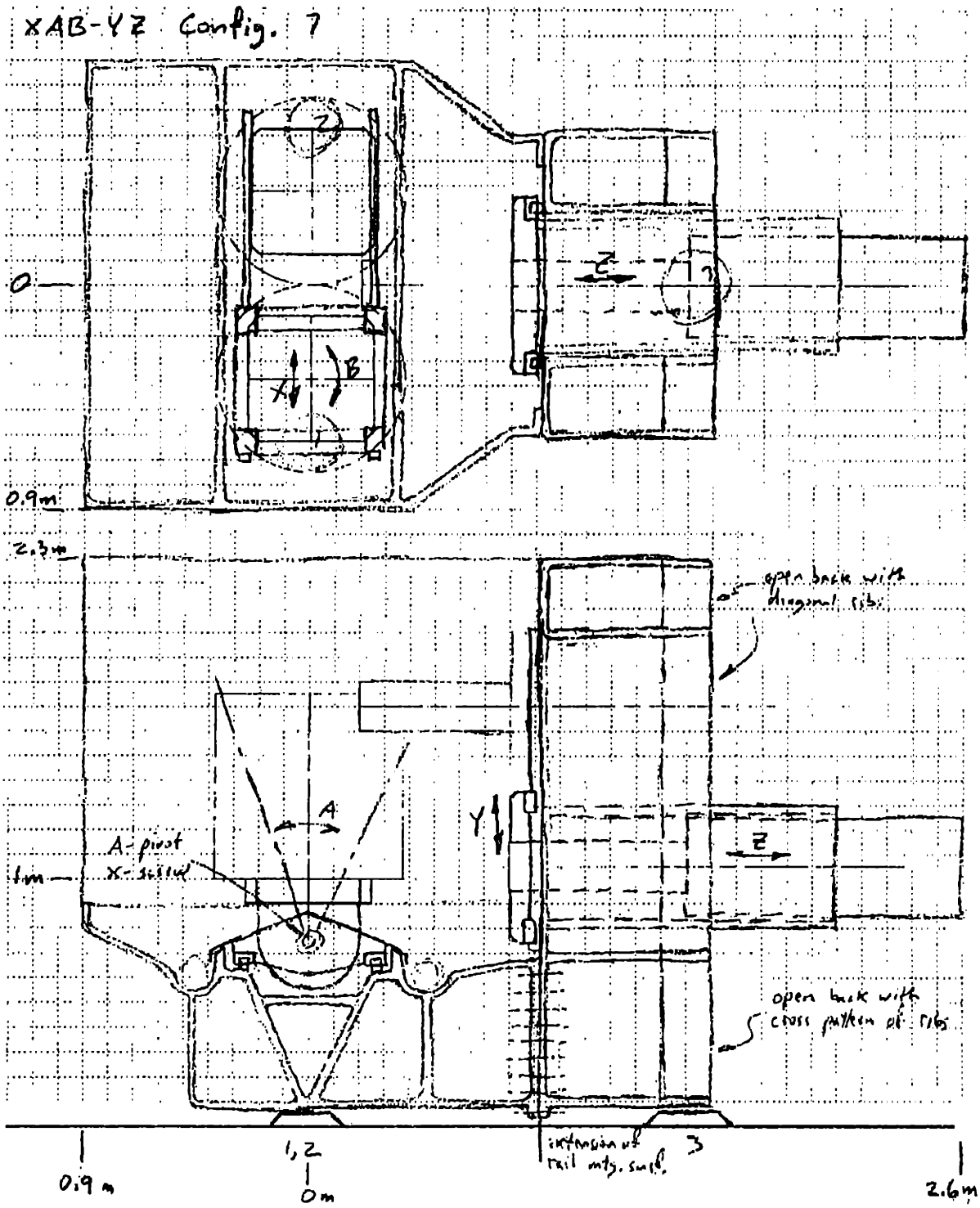


Figure G-6 Configuration 7, work over X, A and B, and tool over Y and Z.

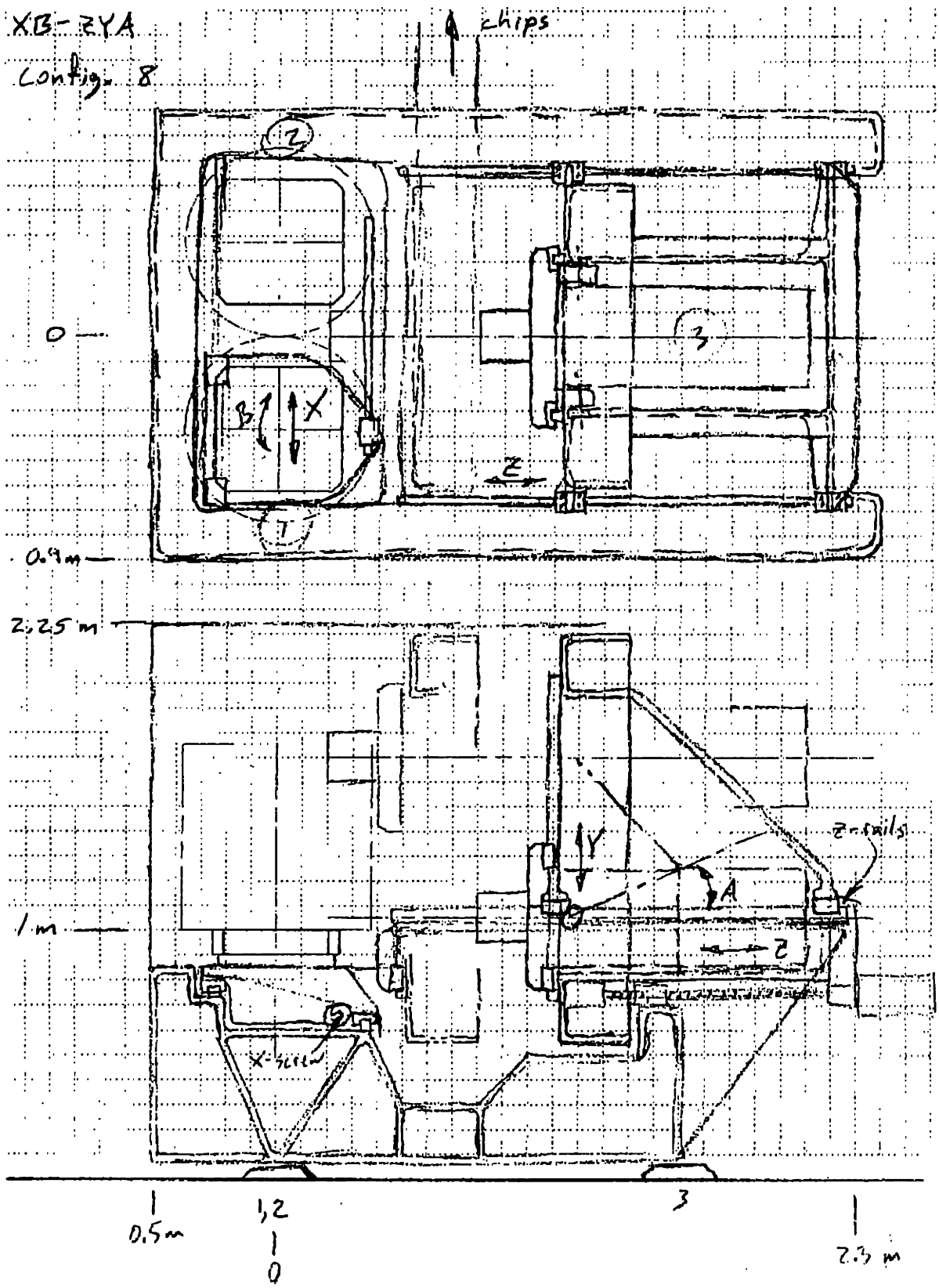


Figure G-7 Configuration 8, work over X and B, and tool over Z, Y and A.

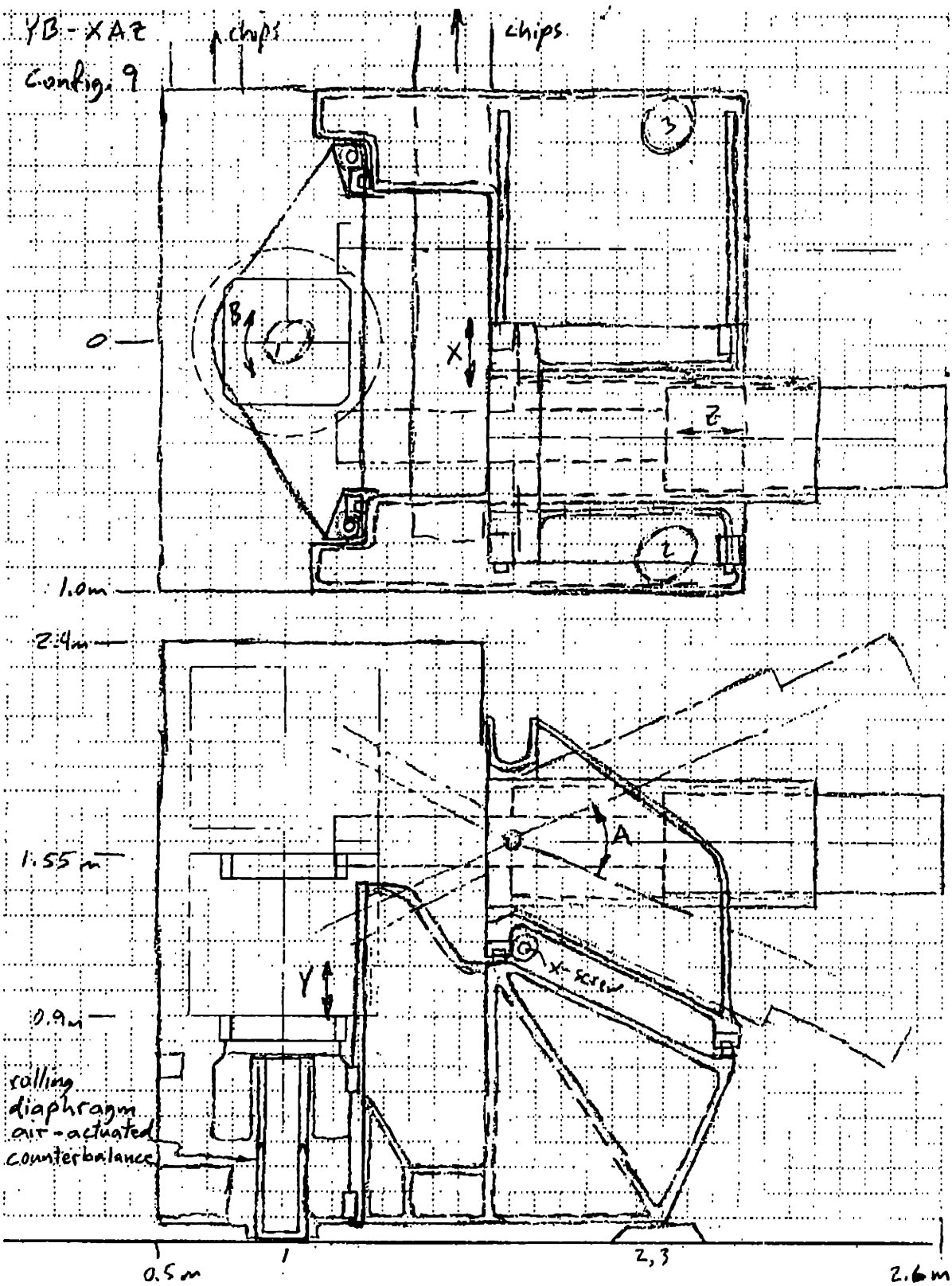


Figure G-8 Configuration 9, work over Y and B, and tool over X, A and Z.

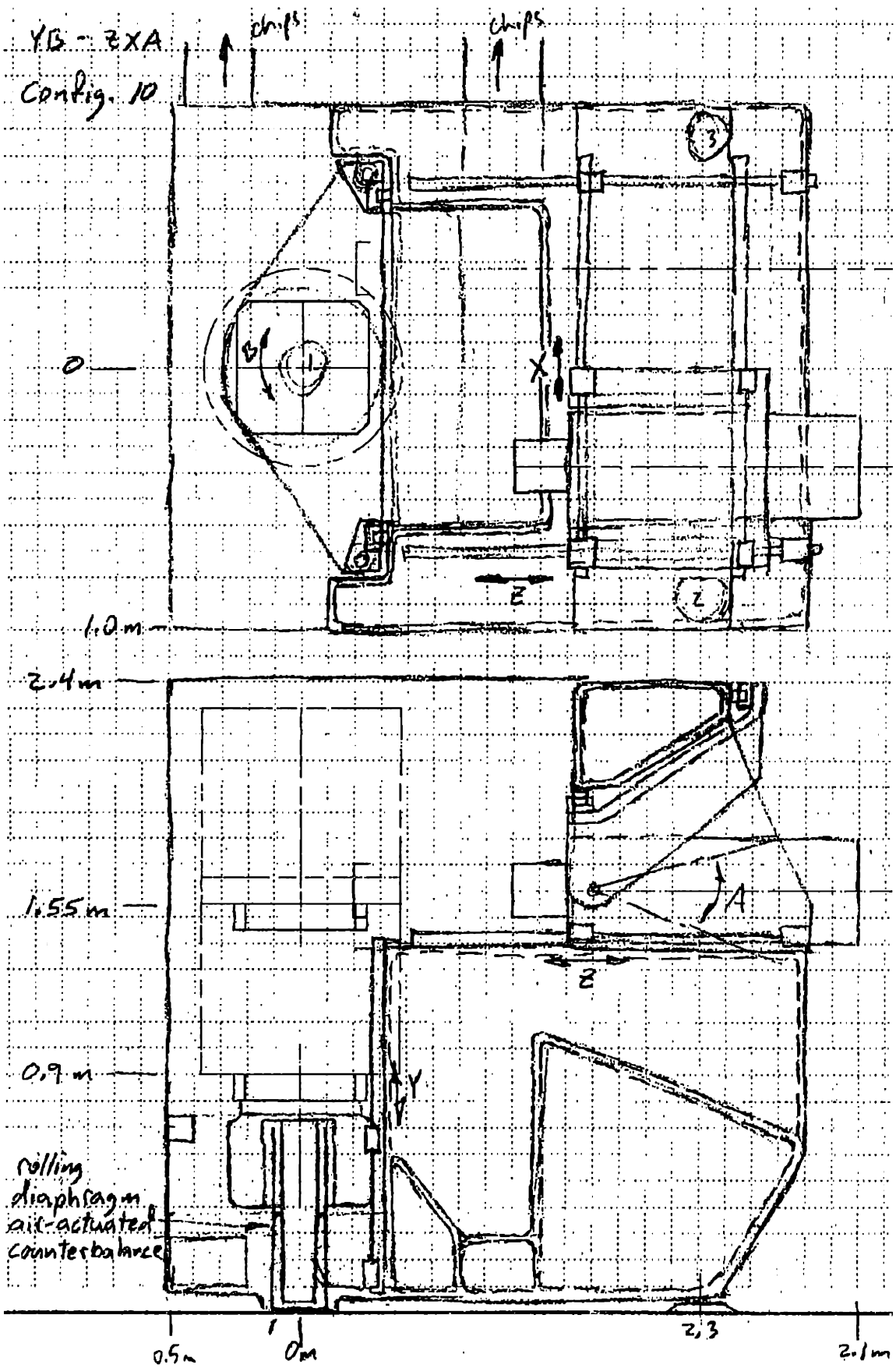


Figure G-9 Configuration 10, work over Y and B, and tool over Z, X and A.

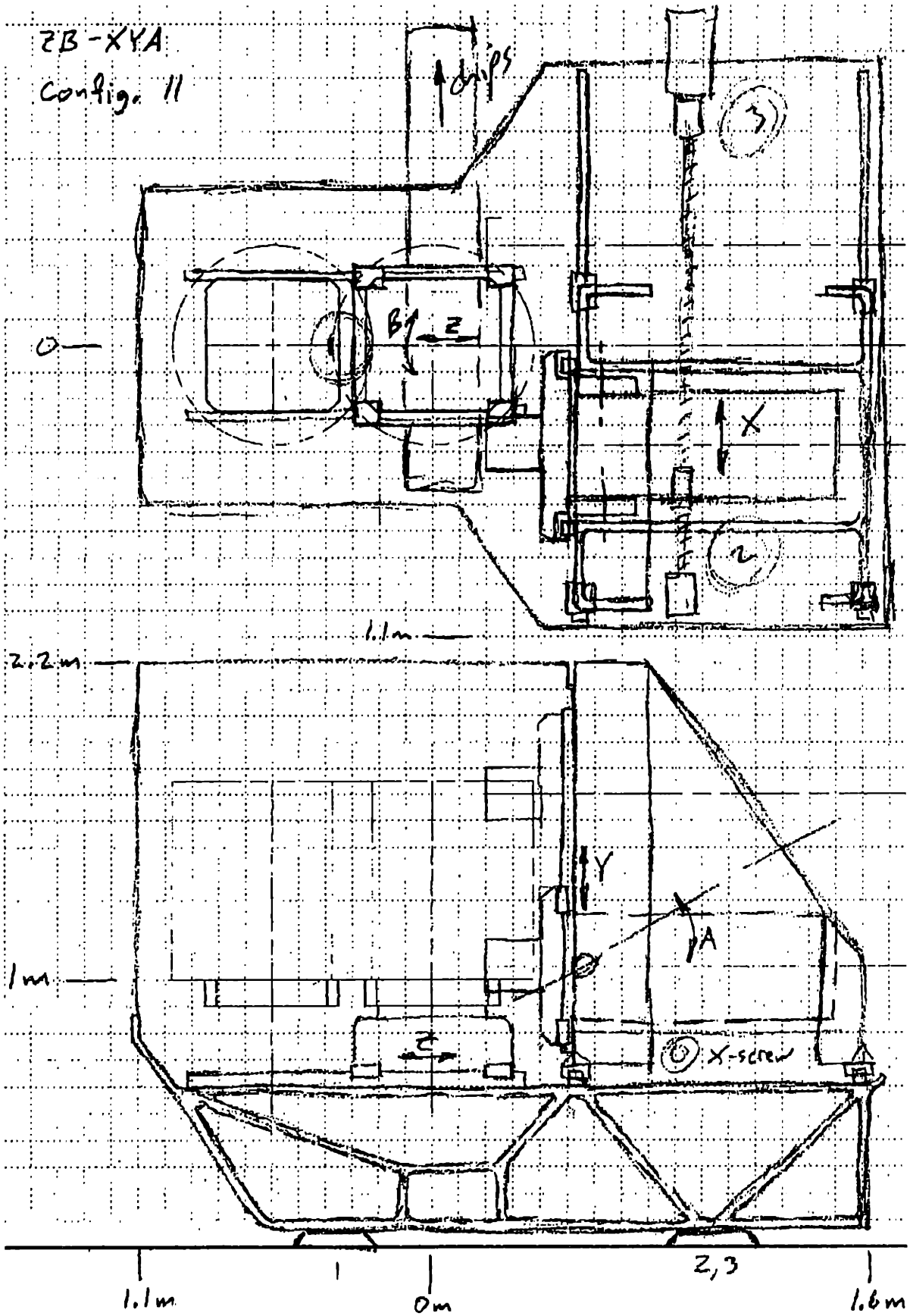


Figure G-10 Configuration 11, work over Z and B, and tool over X, Y and A.

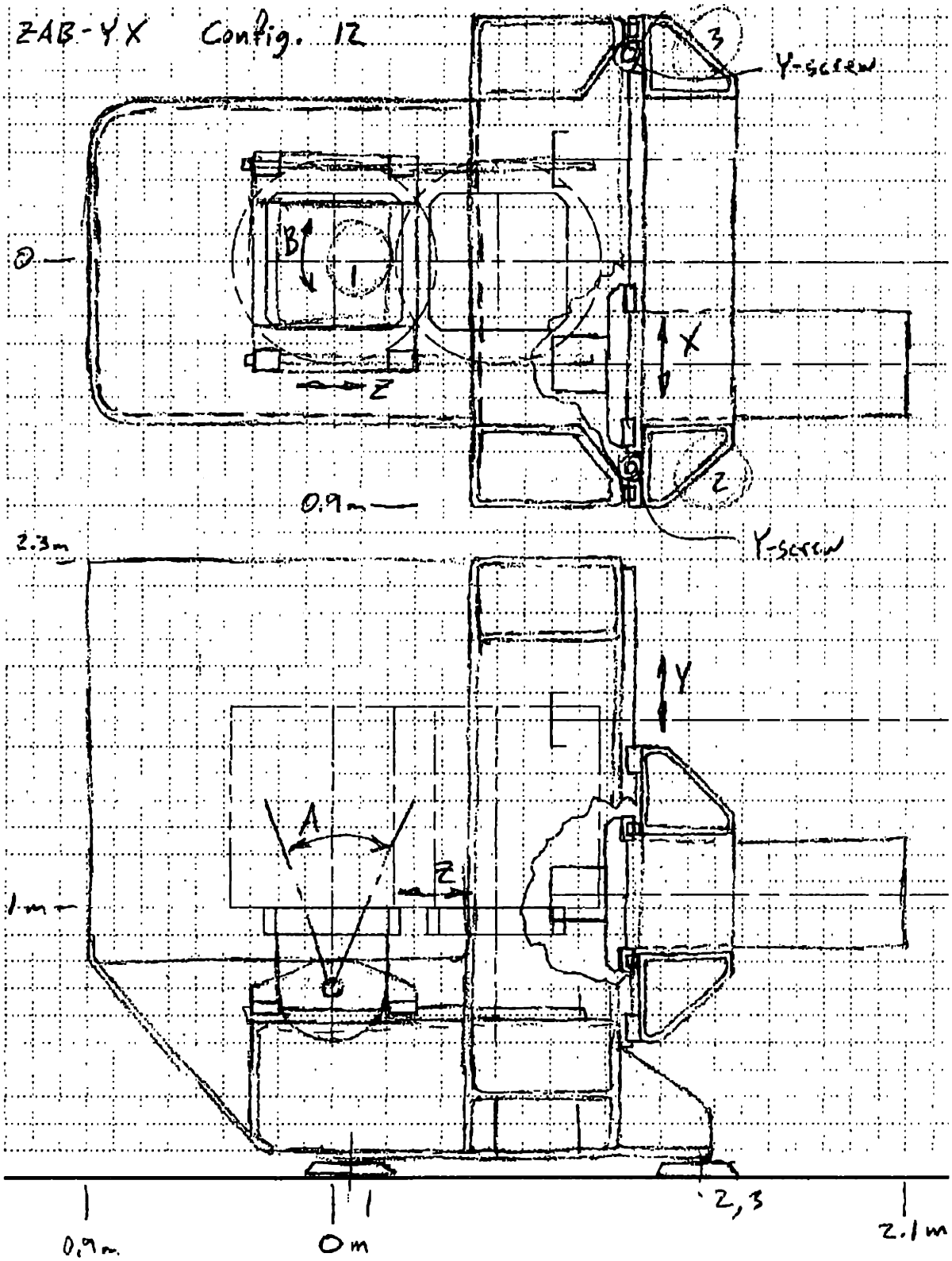


Figure G-11 Configuration 12, work over Z, A and B, and tool over Y and X.

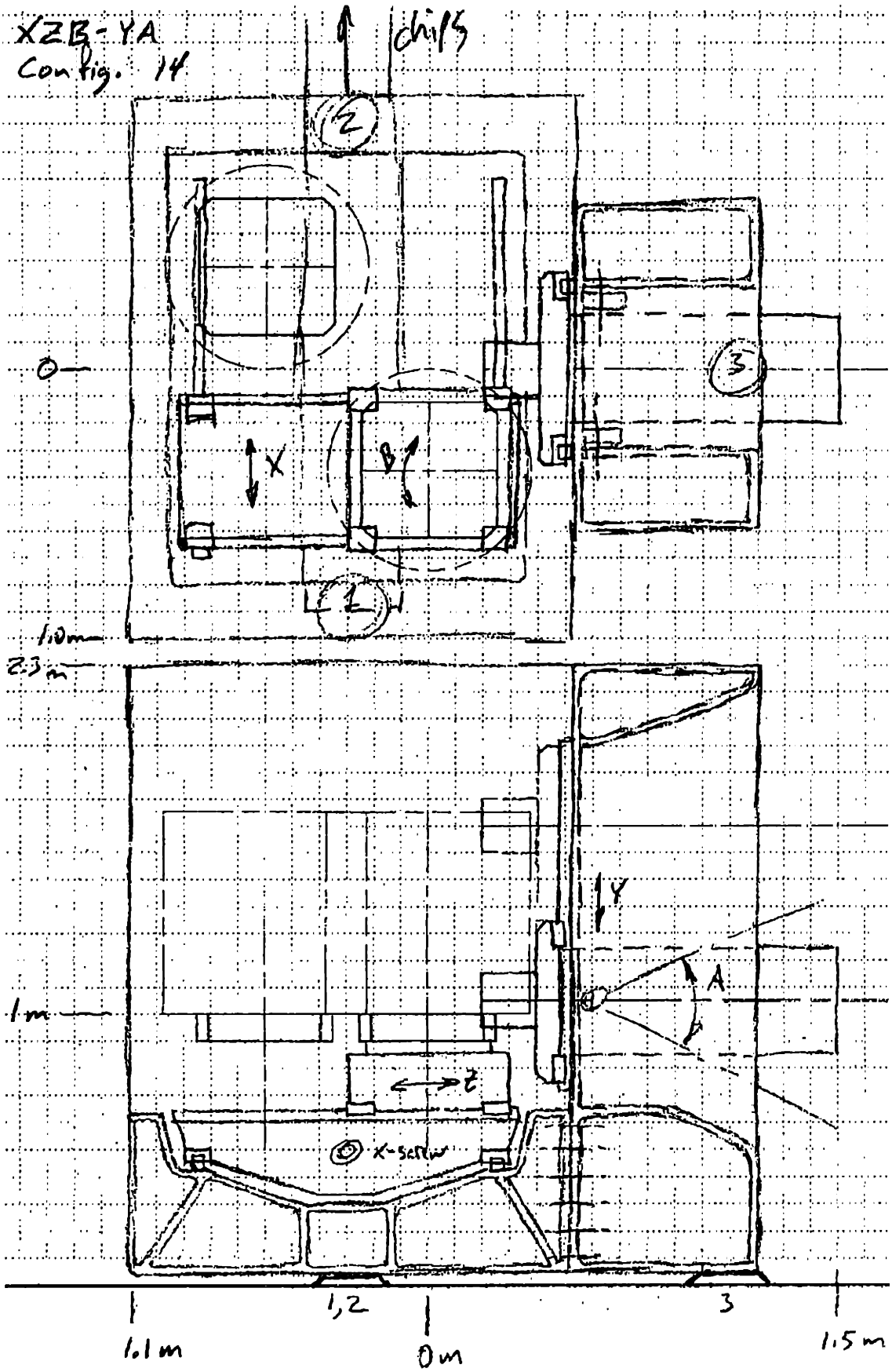


Figure G-12 Configuration 14, work over X, Z and B, and tool over Y and A.

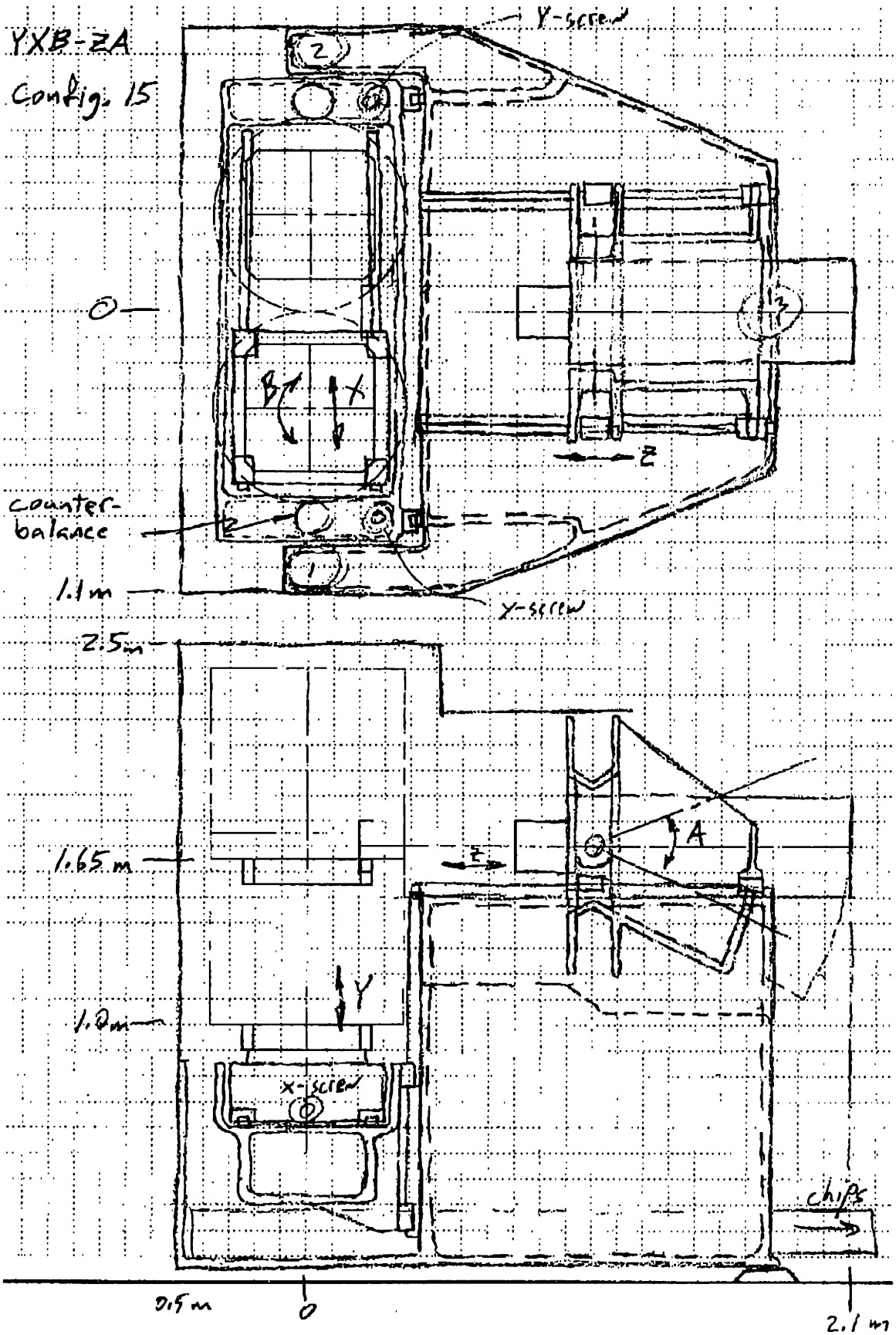


Figure G-13 Configuration 15, work over Y, X and B, and tool over Z and A.

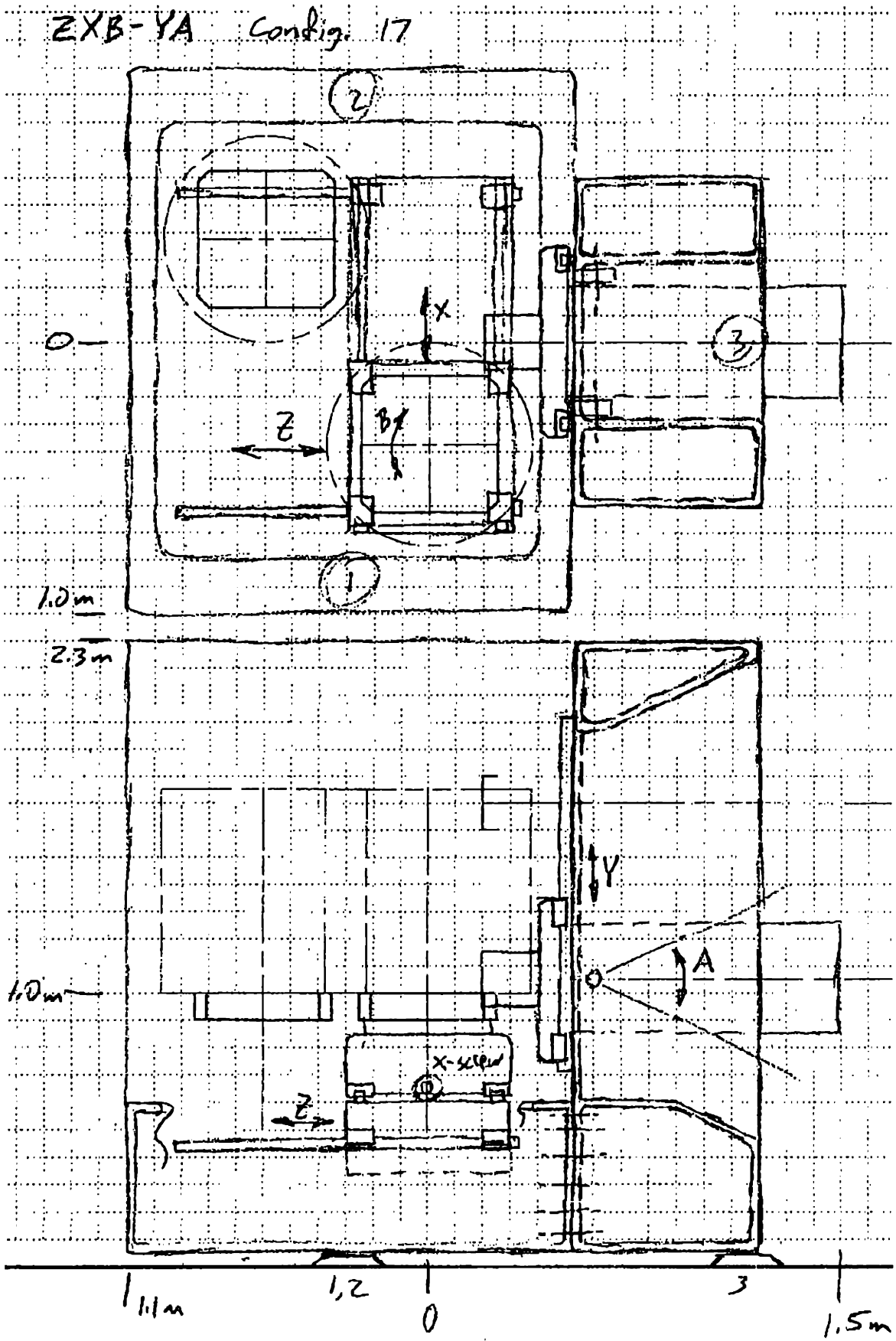


Figure G-14 Configuration 17, work over Z, X and B, and tool over Y and A.