Albano, L.D.; et. al. “Engineering Design”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Engineering Design

Leonard D. Albano
Worcester Polytechnic Institute

Nam P. Suh
Massachusetts Institute of Technology

Michael Pecht
University of Maryland

Alexander Slocum
Massachusetts Institute of Technology

Mark Jakiela
Massachusetts Institute of Technology

Kemper Lewis
Georgia Institute of Technology

Farrokh Mistree
Georgia Institute of Technology

J.R. Jagannatha Rao
University of Houston

11.1 Introduction

Nam P. Suh

Traditionally, the design field has been identified with particular end products, e.g., mechanical design, electrical design, ship design. In these fields, design work is largely based on specific techniques to foster certain product characteristics and principles. Examples include the principles of constant wall thickness, lightweight construction, and shortest load path. Also, the design field has been subdivided by an increasing reliance on specialized knowledge and the division of labor. Precision engineering and robotics are examples of subfields that are distinguished by the accuracy and reliability that the product must have. In the field of precision engineering, for instance, the dimensions of interest are nanometers, which are often encountered in the semiconductor industry. Each one of these fields also has its specific know-how and paradigms to support effective design have also sub-divided the design field.

There are three branches of design. The traditional school, which still dominates, believes that design requires experience and cannot be taught. The second group deals with optimization as a subset of design, using computer-based tools such as genetic algorithms, fuzzy logic, and the like. The third school of thought believes that there are axioms that govern good design decisions. A good designer needs to use all three methodologies.

11.2 Elements of the Design Process

Nam P. Suh

All design activities must do the following:

1. *Know* the "*customers' needs*."
2. *Define the essential problems* that must be solved to satisfy the needs.
3. *Conceptualize the solution through synthesis*, which involves the task of satisfying several different functional requirements using a set of inputs such as product design parameters within given constraints.
4. *Analyze* the proposed solution to establish its optimum conditions and parameter settings.
5. *Check the resulting design solution* to see if it meets the original customer needs.

Design proceeds from abstract and qualitative ideas to quantitative descriptions. It is an iterative process by nature: new information is generated with each step, and it is necessary to evaluate the results in terms of the preceding step. Thus, design involves a continuous interplay between *the requirements the designer wants to achieve* and *how the designer wants to achieve these requirements*.

Designers often find that a clear description of the design requirements is a difficult task. Therefore, some designers deliberately leave them implicit rather than explicit. Then they spend a great deal of time trying to improve and iterate the design, which is time consuming at best. To be efficient and generate the design that meets the perceived needs, the designer must specifically state the users' requirements before the synthesis of solution concepts can begin.

Solution alternatives are generated after the requirements are established. Many problems in mechanical engineering can be solved by applying practical knowledge of engineering, manufacturing, and economics. Other problems require far more imaginative ideas and inventions for their solution. The word "creativity" has been used to describe the human activity that results in ingenious or unpredictable or unforeseen results (e.g., new products, processes, and systems). In this context, creative solutions are discovered or derived by inspiration and/or perspiration, without ever defining specifically what one sets out to create. This creative "spark" or "revelation" may occur, since our brain is a huge information storage and processing device that can store data and synthesize solutions through the use of associative memory, pattern recognition, digestion and recombination of diverse facts, and permutations of events. Design will always benefit when "inspiration" or "creativity," and/or "imagination" plays a role, but this process must be augmented by amplifying human capability systematically through fundamental understanding of cognitive behavior and by the development of scientific foundations for design methods.

11.3 Concept of Domains

Nam P. Suh

Design is made up of four *domains*: the customer *domain*, the *functional domain*, the *physical domain*, and the *process domain* (see Figure 11.3.1). The domain on the left relative to the domain on the right represents "what the designer wants to achieve," whereas the domain on the right represents the design solution, or "how the designer proposes to satisfy the problem." Therefore, the design process can be defined as mapping from the "what" domain to the "how" domain. During product design, the mapping is from the functional domain to the physical domain. In manufacturing process design, the designer maps from the physical domain to the process domain.

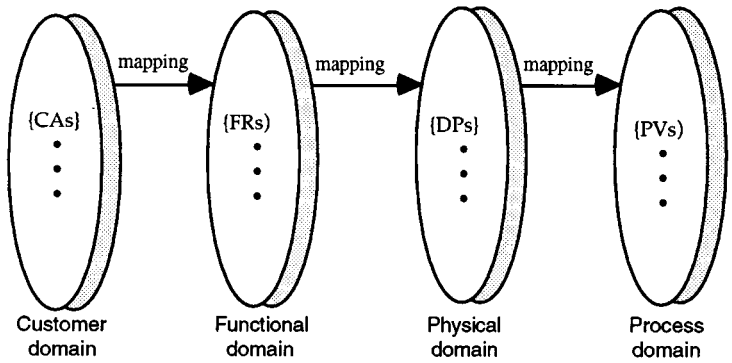


FIGURE 11.3.1 Four domains of the design world. {x} are characteristic vectors of each domain.

The customer domain is characterized by customer needs or the attributes the customer is looking for in a product, process, system, or material. In the functional domain, the designer formally specifies customer needs in terms of functional requirements (FRs). In order to satisfy these FRs, design parameters (DPs) are conceived in the physical domain. Finally, a means to produce the product specified in terms of DPs is developed in the process domain, which is characterized by process variables (PVs).

In mechanical engineering, design typically refers to product design and, often, hardware design. However, mechanical engineers also deal with other equally important designs such as software design, design of manufacturing processes and systems, and organizations. All designers go through the same thought process, although some believe that their design is unique and different from those of everyone else. In materials science, the design goal is to develop materials with certain properties (i.e., FRs). This is done through the design of microstructures (i.e., DPs) to satisfy these FRs, and through the development of material processing methods (i.e., PVs) to create the desired microstructures. To establish a business, the goals are {FRs}, and they are satisfied by structuring the organization in terms of its departments {DPs} and finding the human and financial resources {PVs} necessary to staff and operate the enterprise. Similarly, universities must define the mission of their institutions (i.e., FRs), design their organizations effectively to have an efficient educational and research enterprise (i.e., DPs), and must deal with human and financial resource issues (i.e., PVs). In the case of the U.S. government, the President of the United States must define the right set of {FRs}, design the appropriate government organization and programs {DPs}, and secure the resources necessary to get the job done {PVs}, subject to the constraints imposed by the Constitution and Congress. In all organizational designs the process domain represents the resources: human and financial.

Table 11.3 shows how seemingly different design tasks in many different fields can be described in terms of the four design domains. In the case of the product design, the customer domain consists of the customer requirements or attributes the customer is looking for in a product; the functional domain consists of functional requirements, often defined as engineering specifications and constraints; the

TABLE 11.3 Characteristics of the Four Domains of the Design World for Various Designs: Manufacturing, Materials, Software, Organizations, and Systems

Domains Character Vectors	Customer Domain {CAs}	Functional Domain {FRs}	Physical Domain {DPs}	Process Domain {PVs}
Manufacturing	Attributes which consumers desire	Functional requirements specified for the product	Physical variables which can satisfy the functional requirements	Process variables that can control design parameters (DPs)
Materials	Desired performance	Required properties	Microstructure	Processes
Software	Attributes desired in the software	Output	Input variables and algorithms	Subroutines
Organization	Customer satisfaction	Functions of the organization	Programs or offices	People and other resources that can support the programs
Systems	Attributes desired of the overall system	Functional requirements of the system	Machines or components, subcomponents	Resources (human, financial, materials, etc.)

physical domain is the domain in which the key design parameters {DPs} are chosen to satisfy the {FRs}; and the process domain specifies the manufacturing methods that can produce the {DPs}. It is indeed fortunate that all designs fit into these four domains, since in a given design task, mechanical design, software design, manufacturing issues, and organizational issues must often be considered simultaneously. Because of this logical structure of the design world, generalized design principles can be applied to all design applications, and the issues that arise in the four domains can be considered systematically and concurrently.

Customer needs are often difficult to define. Nevertheless the designer must make every effort to understand customer needs by working with customers to appreciate and establish their needs. Then these needs (or the attributes the customer is looking for in a product) must be translated into functional requirements (FRs). This must be done in a "solution neutral environment." This means that the FRs must be defined without bias to any existing or preconceived solutions. If the FRs are defined based on an existing design, then the designer will simply be specifying the FRs of that product and creative thinking cannot be done.

To aid the process of defining FRs, QFD (quality function deployment) has been used. In QFD customer needs and the possible functional requirements are correlated and important FRs are determined. Experience plays an important role in defining FRs, since qualitative judgment is often necessary for assessing customer needs and identifying the essential problems that must be solved.

11.4 The Axiomatic Approach to Design

Nam P. Suh

The creative process of mapping the FRs in the functional domain to DPs in the physical domain is not unique; the solution varies with a designer's knowledge base and creative capacity. As a consequence, solution alternatives may vary in their effectiveness to meet the customer's needs. The axiomatic approach to design is based on the premise that there are generalizable principles that form the basis for distinguishing between good and bad designs.

Suh (1990) identified two design axioms by *abstracting* common elements from a body of good designs, including products, processes, and systems. The first axiom is called the *Independence Axiom*. It states that the independence of *functional requirements* (FRs) must be always maintained, where FRs are *defined as the minimum set of independent functional requirements* that characterize the design goals. The second axiom is called the *Information Axiom*, which states that among those designs that satisfy the Independence Axiom the design that has the highest probability of success is the best design. During the mapping process (for example, mapping from FRs in the functional domain to DPs in the physical domain), the designer should make correct design decisions using the Independence Axiom. When several designs that satisfy the Independence Axiom are available, the Information Axiom can be used to select the best design.

Axioms are general principles or self-evident truths that cannot be derived or proven to be true; however they can be refuted by counterexamples or exceptions. Through axioms such as Newton's laws and the laws of thermodynamics, the concepts of force, energy, and entropy have been defined. One of the main reasons for pursuing an axiomatic approach to design is the generalizability of axioms, which leads to the derivation of corollaries and theorems. These theorems and corollaries can be used as design rules that precisely prescribe the bounds of their validity because they are based on axioms. The following corollaries are presented in Suh (1990).

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.

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

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

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.

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

The ultimate goal of axiomatic design is to establish a science base for design and improve design activities by providing the designer with a theoretical foundation based on logical and rational thought processes and tools.

The First Axiom: The Independence Axiom

The Independence Axiom may be formally stated as:

Axiom 1: The Independence Axiom
Maintain the independence of the functional requirements.

As stated earlier, functional requirements, FRs, are defined as the minimum set of independent requirements that the design must satisfy. A set of functional requirements is the description of design goals. The Independence Axiom states that when there are two or more functional requirements, the design solution must be such that each of the functional requirements can be satisfied without affecting any of the other requirements. This means that the designer must choose the correct set of DPs so that functional dependence or coupling is not introduced. When there is only one FR, the Independence Axiom is always satisfied. In this case, the given design alternatives should be optimized and the second axiom, the Information Axiom, is used to select the best design.

To apply the Independence Axiom, the mapping process from the design goals to the design solutions can be expressed mathematically. The set of functional requirements that define the specific design goals constitutes a vector {FRs} in the functional domain. Similarly, the set of design parameters in the physical domain that describe the design solution also constitutes a vector {DPs}. The relationship between the two vectors can be written as

$$\{\text{FRs}\} = [A]\{\text{DPs}\} \tag{11.4.1}$$

where [A] is the design matrix that characterizes the nature of the mapping. Equation (11.4.1) may be written in terms of its elements as $FR_i = A_{ij}DP_j$. Equation (11.4.1) is a design equation that may be used for the design of a product or the microstructure of a material. For the design of processes, the design equation may be written as

$$\{\text{DPs}\} = [B]\{\text{PVs}\} \tag{11.4.2}$$

where [B] is the design matrix that characterizes the process design.

Designs that satisfy the Independence Axiom must have either a diagonal or triangular design matrix (see Figure 11.4.1). When the design matrix [A] is diagonal, each of the FRs can be satisfied independently by adjusting one DP. Such a design is called an *uncoupled design*. When the matrix is triangular, the independence of FRs can be guaranteed if and only if the DPs are changed in a proper sequence. Such a design is called a *decoupled* or *quasi-coupled design*. Although the design matrix is a second-order tensor (note: stress, strain, and moment of inertia are also second-order tensors), the usual coordinate transformation technique cannot be applied to Equations (11.4.1) or (11.4.2) to find the invariants such as the diagonal matrix, since [A] and [B] typically involve physical phenomena and geometric relationships that are not amenable to coordinate transformation.

In addition to the Independence Axiom, the mapping of the design goals (FRs or DPs) to design solutions (DPs or PVs, respectively) is often subject to constraints, Cs. Constraints establish the bounds on the acceptable design solutions and differ from FRs in that they do not have to be independent. Cost, for example, is often considered a constraint since it is affected by all design decisions, but the design is acceptable as long as the cost does not exceed a set limit.

Example 1: Shaping of Hydraulic Tubes

In many applications (e.g., aircraft industry), steel tubes must be bent to complex shapes without changing the circular cross-sectional shape of the tube. To design a machine and a process that can achieve the task, the functional requirements can be formally stated as

FR1 = Bend the steel tube to prescribed curvatures.
FR2 = Maintain the circular cross section of the bent tube.

$$\begin{Bmatrix} FR_1 \\ FR_2 \\ FR_3 \end{Bmatrix} = \begin{bmatrix} X & 0 & 0 \\ 0 & X & 0 \\ 0 & 0 & X \end{bmatrix} \begin{Bmatrix} DP_1 \\ DP_2 \\ DP_3 \end{Bmatrix}$$

a) Uncoupled design

$$\begin{Bmatrix} FR1 \\ FR2 \\ FR3 \end{Bmatrix} = \begin{bmatrix} X & 0 & 0 \\ X & X & 0 \\ X & X & X \end{bmatrix} \begin{Bmatrix} DP1 \\ DP2 \\ DP3 \end{Bmatrix}$$

b) Decoupled design

$$\begin{Bmatrix} FR1 \\ FR2 \\ FR3 \end{Bmatrix} = \begin{bmatrix} X & 0 & X \\ X & X & 0 \\ X & X & X \end{bmatrix} \begin{Bmatrix} DP1 \\ DP2 \\ DP3 \end{Bmatrix}$$

c) Coupled design

FIGURE 11.4.1 Examples of uncoupled, decoupled, and coupled designs.

One mechanical concept that can do the job is shown schematically in Figure 11.4.2 for a two-dimensional bending case. It consists of a set of matching rollers with semicircular grooves on their periphery. These "bending" rollers can counterrotate at different speeds and move relative to each other to control the bending process. The centers of these two bending rollers are fixed with respect to each other, and the contact point of the bending rollers can rotate about a fixed point. A second set of "feed" rollers, which counterrotate at the same speed, feed the straight tube feedstock into the bending rollers. As the tubes are bent around the rollers, the cross-sectional shape will tend to change to a noncircular shape. The deformation of the cross section is prevented by the semicircular cam profile machined on the periphery of the bending rollers. The DPs for this design are

DP1 = Differential rotation of the bending rollers to bend the tube
DP2 = The profile of the grooves on the periphery of the bending rollers

The kinematics of the roller motion needs to be determined. To bend the tube, one of the bending rollers must rotate faster than the other. In this case, the tube will be bent around the slower roller. The forward speed of the tube is determined by the average speed of the two bending rollers. The motion of these rollers can be controlled digitally using stepping motors.

The design shown in Figure 11.4.2 is an uncoupled design, since each of the proposed DPs only affects one FR. Is this the best design? The only way this question can be answered is to develop alternative designs that satisfy the FRs, constraints (Cs), and the Independence Axiom. Then the information content for each of the proposed designs must be computed so as to use the Information Axiom to select the best among the proposed designs.

Decomposition, Zigzagging, and Hierarchy

In the preceding example the design was completed when the two FRs in the functional domain were mapped to two DPs in the physical domain. However, in many designs both the FRs and DPs must be decomposed into hierarchies because the high level conceptual design ideas need further design details before they can be implemented. To create the hierarchies for the FRs and DPs, the designer must return to the functional domain from the physical domain. As shown in Figure 11.4.3, the design process requires the designer to alternate or zigzag between the functional domain and the physical domain.

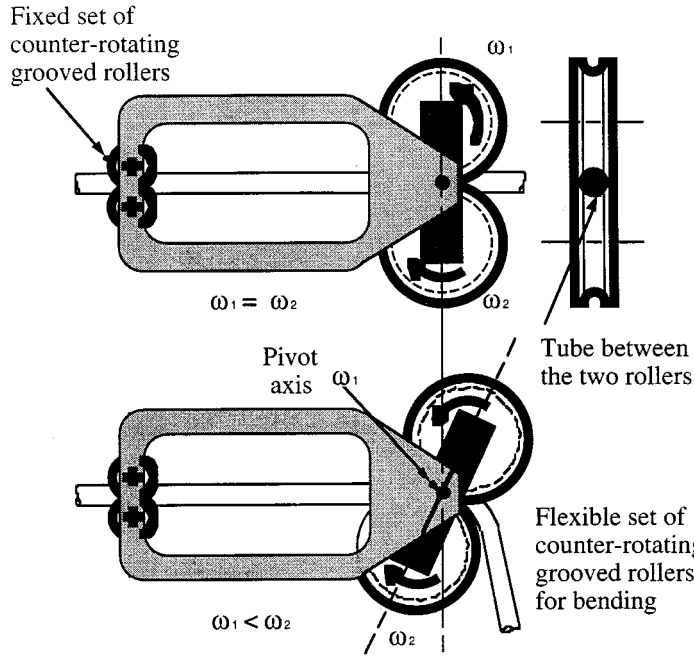


FIGURE 11.4.2 Concept for tube bending.

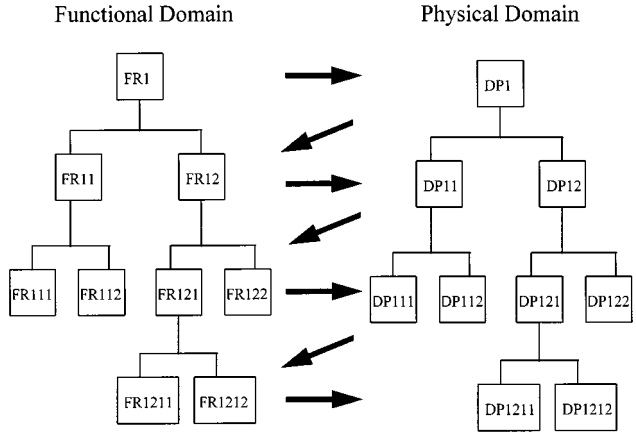


FIGURE 11.4.3 Hierarchical decomposition of FRs and DPs and zigzag mapping.

In many organizations, functional requirements or design specifications are decomposed without zigzagging by remaining only in the functional domain. This means that the designers are not working in a *solution-neutral environment*; they are specifying the requirements for an existing design. For example, assume that the design objective is to develop a vehicle that goes forward, stops, and turns. This vehicle has to satisfy these three FRs. These FRs cannot be decomposed until they are mapped to a set of DPs in the physical domain. This high-level mapping is often referred to as the *conceptual phase* of design. If, for instance, an electric motor is adopted as a means to move forward, then the FR "go forward" would be further decomposed in terms of this physical concept, and the evolving functional hierarchy will be different from that associated with the decision to use gas turbines. Therefore, to define the FRs in a solution-neutral environment, the designer needs to "zig" to the physical domain, and after proper DPs are chosen, the designer must then "zag" to the functional domain for further decomposition.

This process of mapping and zigzagging must continue until the design is completed, resulting in the creation of hierarachical trees for both FRs and DPs. This will be illustrated in the following example.

Example 2: Refrigerator Design

Historically, humankind has had the need to preserve food. Now consumers want an electrical appliance that can preserve food for an extended time. The typical solution is to freeze food for long-term preservation or to keep some food at a cold temperature without freezing for short-term preservation. These needs can be formally stated in terms of two functional requirements:

FR1 = Freeze food for long-term preservation.
FR2 = Maintain food at cold temperature for short-term preservation.

To satisfy these two FRs, a refrigerator with two compartments is designed. The two DPs for this refrigerator may be stated as

DP1 = Freezer section
DP2 = Chiller (i.e., refrigerator) section

To satisfy FR1 and FR2, the freezer section should only affect the food to be frozen and the chiller (i.e., refrigerator) section should only affect the food to be chilled without freezing. In this case, the design matrix will be diagonal. However, the conventional freezer/refrigerator design uses one compressor and one fan which turns on when the freezer section temperature is higher than the set temperature, and the chiller section is cooled by controlling the opening of the vent (see Figure 11.4.4). Therefore, the temperature in the chiller section cannot be controlled independently from that of the freezer section.

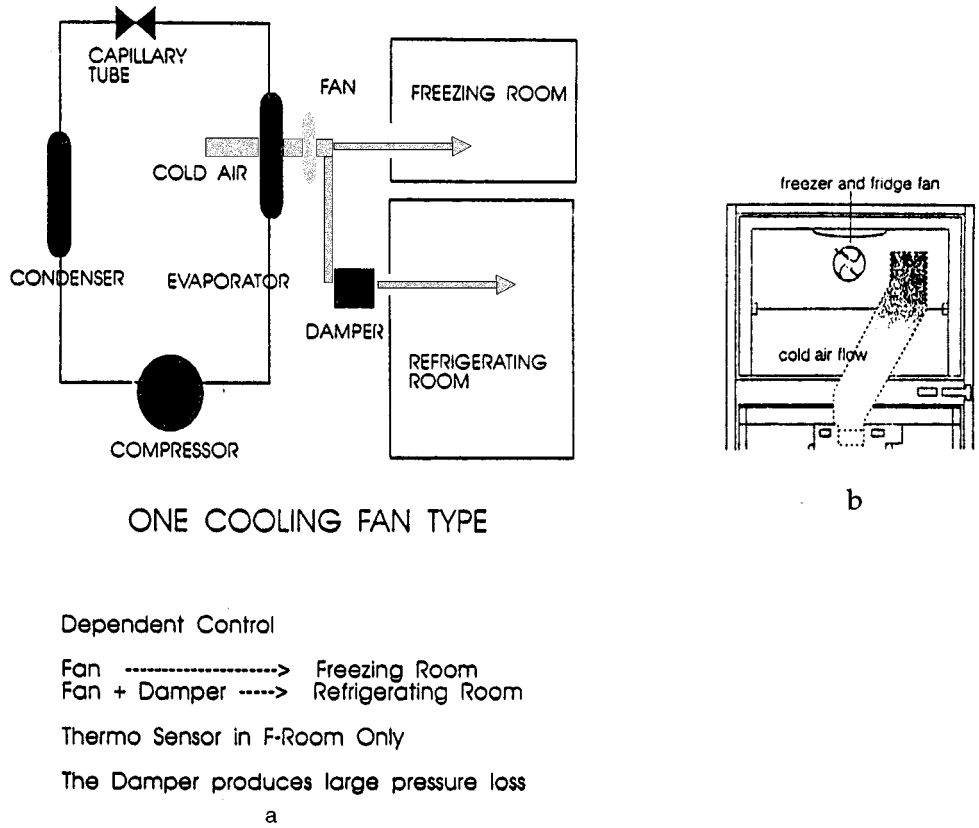


FIGURE 11.4.4 Conventional refrigerator/freezer cooling system.

Having chosen DP1, FR1 can now be decomposed as

FR11 = Control temperature of the freezer section in the range of –18°C.
FR12 = Maintain the uniform temperature throughout the freezer section at the preset temperature.
FR13 = Control humidity to relative humidity of 50%.

FR2 may be decomposed in the context of DP2 as

FR21 = Control the temperature of the chilled section in the range of 2 to 3°C.
FR22 = Maintain a uniform temperature throughout the chilled section at a preset temperature to within 1°C.

To satisfy the second level FRs, i.e., FR11, FR12, FR21, etc., the designer has to conceive a design and identify the DPs for this level of decomposition. Just as FR1 and FR2 were independent from each other through the choice of proper DP1 and DP2, the FRs at this second level must also be independent from one another.

Suppose that the requirements of the freezer section will be satisfied by pumping chilled air into the freezer section, circulating the chilled air uniformly throughout the freezer section, and monitoring the returning air for temperature and moisture in such a way that the temperature is controlled independently from the moisture content of the air. Then the second level DPs may be chosen as

DP11 = Turn the compressor on or off when the air temperature is higher or lower than the set temperature, respectively.
DP12 = Blow the air into the freezer section and circulate it uniformly throughout the freezer section at all times.
DP13 = Condense the moisture in the returned air when its dew point is exceeded.

The design equation is written as

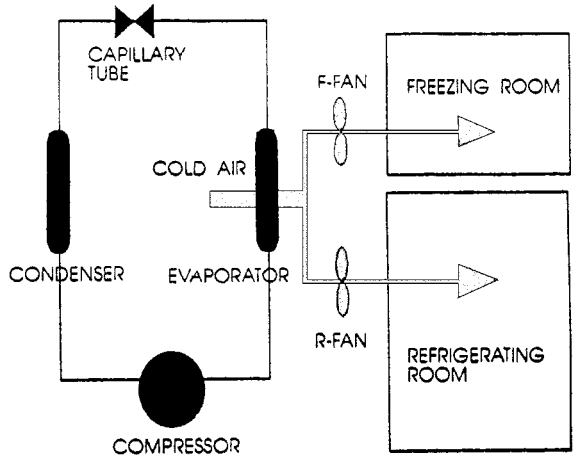
$$\begin{vmatrix} FR12 \\ FR11 \\ FR13 \end{vmatrix} = \begin{vmatrix} x & 0 & 0 \\ x & x & 0 \\ x & 0 & x \end{vmatrix} \begin{vmatrix} DP12 \\ DP11 \\ DP13 \end{vmatrix} \tag{11.4.3}$$

Equation (11.4.3) indicates that the design is a decoupled design.

The chilled section, where the food has to be kept in the range of 2 to 3°C, can now be designed. Here, again, a compressor may be used to control the air temperature within a preset range, and chilled air may be circulated to maintain a uniform temperature throughout the chilled section. This solution would result in a decoupled design as well. One of the design questions to be answered here is whether one compressor and one fan can be used to satisfy both {FR11, FR12, FR13} and {FR21, FR22}. Such a solution would minimize the information content without compromising functional independence (see Corollary 3). Most commercial refrigerators use only one compressor and one fan to achieve these goals (see Figure 11.4.4); however, many of these are coupled designs.

One can propose various specific design alternatives and evaluate the options in terms of the Independence Axiom. If a design allows the satisfaction of these FRs independently, then the design is acceptable for the set of specified FRs. If a solution that satisfies the Independence Axiom cannot be devised, then the designer must compromise the FRs by either eliminating some of the FRs or increasing the tolerance for temperature control, moisture control, etc.

One company has improved the preservation of food in their chilled section by providing one additional fan so as to be able to control the temperature of the chilled section more effectively (see Figure 11.4.5). This can be done since the evaporator was sufficiently cold to cool the air being pumped into the chiller section, even during periods when the compressor was not running. To maintain a uniform temperature distribution, extra vents were provided to insure good circulation of air. In this design, DP21 refers to



TWO COOLING FAN TYPE

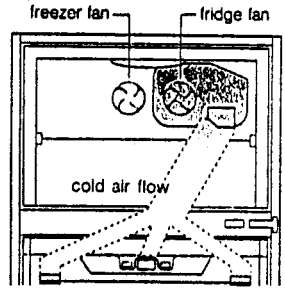
Independent Control

F-Fan ----------> Freezing Room
R-Fan ---------> Refrigerating Room

Thermo Sensor in F-Room and R-Room

No Damper is Required

a



b

FIGURE 11.4.5 New refrigerator/freezer cooling system.

the fan for the chiller section, and DP22 represents the vents. The design matrix for the {FR21, FR22}–{DP21, DP22} relationship is diagonal as shown in the design equation:

$$\begin{vmatrix} FR21 \\ FR22 \end{vmatrix} = \begin{vmatrix} X & 0 \\ 0 & X \end{vmatrix} \begin{vmatrix} DP21 \\ DP22 \end{vmatrix} \tag{11.4.4}$$

The company has compared the performance of the uncoupled design with that of competing coupled designs. For instance, the new design provides much more uniform temperature in the chiller section (Figure 11.4.6a) and much less temporal fluctuation than the existing coupled designs (Figure 11.4.6b). They also found that the uncoupled design saves electricity because air can be defrosted due to the air flow into the chiller section when the compressor is not operating. They also found that this new design enables the use of a quick refrigeration mode in the chilled section by turning on the fan of the chiller section as soon as food is put into the chiller section. To cool 100 g of water from 25 to 10°C it took only 37 vs. 58 min in a conventional refrigerator (Figure 11.4.7).

According to Corollary 3 (Integration of Physical Parts), the innovative idea of using two fans and uniformly positioned ducts may or may not be the best solution if the FRs can be satisfied independently using only one fan. If there is an alternative design that can satisfy the Independence Axiom, the Information Axiom and the detailed calculation of the information content must be used to determine the better of the two designs. As discussed in "The Second Axiom: The Information Axiom" (below), the best design is the one that has the minimum information content since it has the highest probability of success.

In the preceding example, the design matrices were formulated in terms of Xs and Os to model the nature of the relationship between each FR and each DP. In some cases, this simple, conceptual notation is sufficient to complete and implement the design. However, in many cases further steps should be taken to optimize the design. This means that the FR-DP relationships must be modeled more precisely after the conceptual design is represented in terms of Xs and Os.

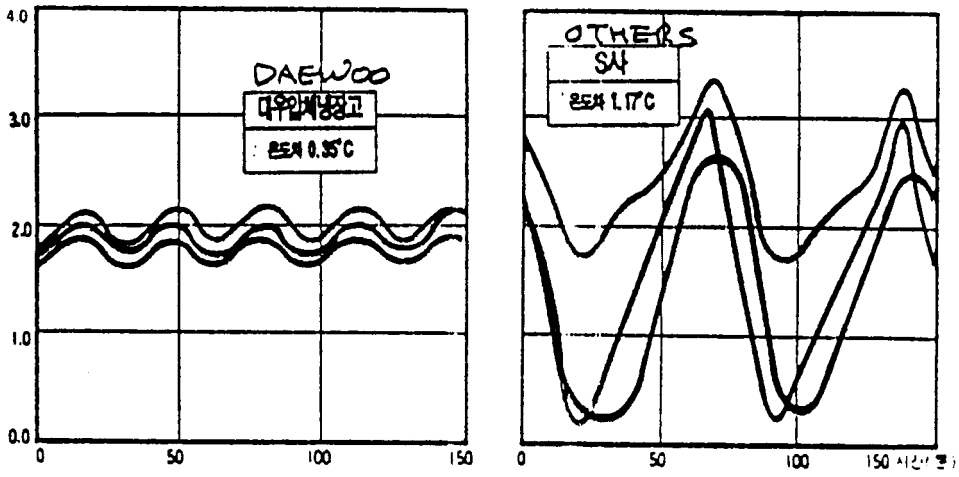
Concurrent Design and Manufacturing Considerations

The previous section, Decomposition, Zigzagging, and Hierarchy, demonstrates use of the Independence Axiom during mapping from the physical domain to the process domain, i.e., product design. In order to implement the chosen DPs, the designer has to map from the physical domain to the process domain (i.e., process design) by choosing the process variables, PVs. This process design mapping must also satisfy the Independence Axiom, although sometimes the solution may simply involve the use of existing processes. When existing processes must be used to minimize capital investment in new equipment, the corresponding PVs must also be used to complete the mapping from the physical domain to the process domain. Therefore, the PVs act as constraints in choosing DPs. Since early design decisions may determine 70 to 80% of manufacturing productivity in many enterprises, the product design and process design (or selection) should be performed at the same time in order to develop designs that can be manufactured without incurring cost overruns and schedule slippage. This is sometimes called *concurrent design*.

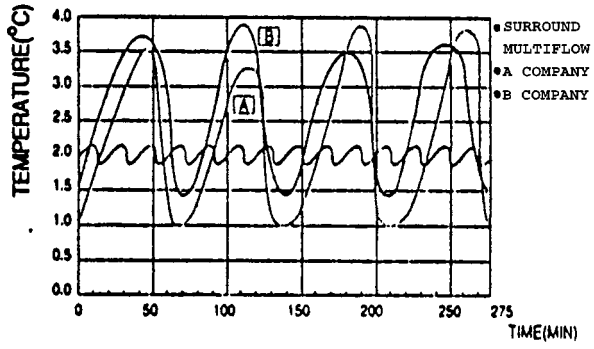
The Second Axiom: The Information Axiom

When there is only one FR, the Independence Axiom is always satisfied, and the only task left is to optimize the given design. Various optimization techniques have been advanced to deal with optimization problems involving one objective function. However, when there are two or more FRs, some of these optimization techniques do not work. In these cases, we must first develop a design that is either uncoupled or decoupled. If the design is uncoupled, it can be seen that each FR can be satisfied and the optimum points can be found. If the design is decoupled, the optimization technique must follow a set sequence.

For a given set of FRs, it is most likely that every designer will come up with different designs, all of which are acceptable in terms of the Independence Axiom. However, one of these designs is likely to be the superior alternative. The Information Axiom provides a quantitative means for establishing the merits of a given design, and this value is used to select the best solution. Specifically, the Information Axiom may be stated as



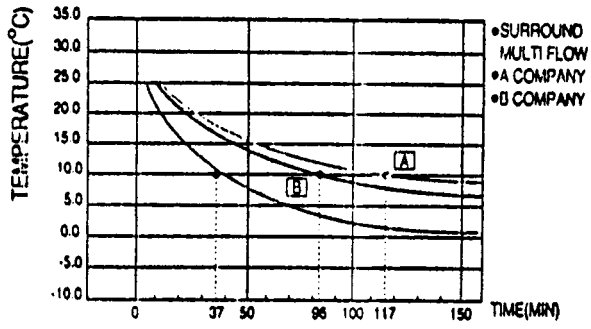
■ +2°C fixed temperature refrigeration
(No variation of temperature)



DIVISION	SURROUND MULTI AIR FLOW	CONVENTIONAL MULTI AIR FLOW	
	FR-820NT	A COMPANY	B COMPANY
TEMPERATURE VARIATION DEGREE	0.6°C	1.5°C	1.9°C

b

FIGURE 11.4.6 Comparison of temperature control.



DIVISION	SURROUND MULTI AIR FLOW	CONVENTIONAL MULTI AIR FLOW	
	FR-820NT	A COMPANY	B COMPANY
ΔT 15°C REFRIGERATE TIME	37 MINUTE	117 MIN	96 MIN

FIGURE 11.4.7 Quick refrigeration performance.

Axiom 2: The Information Axiom
Minimize the information content.

Information is defined in terms of the information content **I** that is related, in its simplest form, to the probability of satisfying a given set of FRs. If the probability of success is p, the information content **I** associated with the probability is defined as

$$\mathbf{I} = -\log_2 p \tag{11.4.5}$$

Equation (11.4.5) defines information content in the units of binary digits or bits. In the general case of an uncoupled design with *n* FRs, **I** may be expressed as

$$I = -\sum_{i=1}^{n} \log_2 p_i = \sum_{i=1}^{n} I_i \tag{11.4.6}$$

where p_i is the probability of DP_i satisfying FR_i. Since there are *n* FRs, the total information content is the sum of all the individual measures. When all probabilities p_i are equal to one, the information content is zero. Conversely, the information content is infinite when one or more probabilities are equal to zero.

A design is called *complex* when its probability of success is low. The quantitative measure for complexity is the information content: complex systems require more information to make the system function (see Equation 11.4.5). Thus, a large system that is comprised of many subsystems and components is not necessarily complex. Even a small system can be complex if its probability of success is low.

Example 3: Cutting a Rod to a Length

Suppose we need to cut rod A to the length 1 +/– 0.000001 m and rod B to the length 1 +/– 0.1 m. Which task is more complicated?

The answer depends on the cutting equipment available for the job! However, most engineers with some practical experience would say that the one that has to be cut within 1 μm would be more difficult, because the probability of success associated with the smaller tolerance is lower than that associated

with the larger tolerance. Therefore, the job with the lower probability of success is more *complex* than the one with higher probability. The Information Axiom links the notion of complexity with the specified tolerances for the FRs: the tighter the tolerance, the more difficult it may be to choose a design solution or a system that can satisfy the FRs.

In practice, the probability of success is a function of the intersection of the tolerances defined by the designer to satisfy the FRs and the ability of the chosen system to produce the part within the specified tolerances. This value can be computed by specifying the *design range* (*dr*) for an FR and by determining the *system range* (*sr*) that the proposed design can provide to satisfy the FR. Figure 11.4.8 illustrates these two ranges graphically. The vertical axis (the ordinate) is for the probability density function and the horizontal axis (the abscissa) is for either FR or DP, depending on the mapping domains involved. When the mapping is between the functional and the physical domains (as in product design), the abscissa is for FR, whereas for mapping between the physical and the process domains (e.g., process design), the abscissa is for DP.

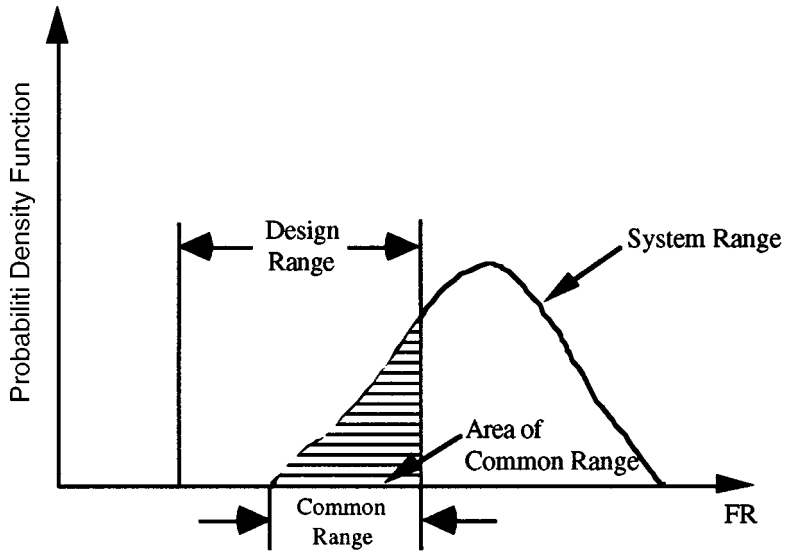


FIGURE 11.4.8 Design range, system range, and common range in a plot of the probability density function (pdf) of a functional requirement.

In Figure 11.4.8 the system range is plotted as a probability density function vs. the specified FR. The intersection of the design range and the system range is called the *common range* (*cr*), and this is the only region where the design requirements are satisfied. Consequently, the area under the common range divided by the area under the system range is equal to the design's probability of success. Then the information content may be expressed as

$$\mathbf{I} = \log_2\left(A_{sr}/A_{cr}\right) \tag{11.4.7}$$

where A_{sr} denotes the area under the system range and A_{cr} is the area within the common range. Furthermore, since A_{sr} =1.0 and there are *n* FRs to satisfy in most cases, the information content may be expressed as

$$I = -\sum_{i=1}^{n} \log_2\left(1/A_{cr}\right)_i \tag{11.4.8}$$

References

Suh, N.P. 1990. *The Principles of Design.* Oxford University Press, New York.

Suh, N.P. June 1995. Axiomatic design of mechanical systems. *Special 50th Anniversary Combined Issue of the Journal of Mechanical Design and the Journal of Vibration and Acoustics. Trans. ASME.* 117:1–10.

11.5 Algorithmic Approaches to Design

Leonard D. Albano

In addition to the axiomatic approach to design (see Section 11.4), there are many methods that are based on an algorithmic approach to design. In algorithmic design, the design process is identified or prescribed so that it leads the designer to a solution that satisfies the design goals. Algorithmic methods can be divided into several categories: pattern recognition, associative memory, analogy, experientially based prescription, extrapolation, interpolation, selection based on probability, etc. This section briefly discusses three common algorithmic methods for design: systematic design, the Taguchi method, and design for assembly (DFA).

Systematic Design

Systematic design methods prescribe how design should be done and provide standardized solutions for fulfilling certain functional requirements. Examples of systematic design methods include Pahl and Beitz (1988), Hubka and Eder (1988), and the German Standards Institute (VDI). This section describes the approach advanced by Pahl and Beitz.

The method of Pahl and Beitz (1988) divides the design process into a number of steps and prescribes methods for dealing with each step. Design starts with an appreciation of customers needs, which are then clarified in terms of an overall function. By definition, the overall function of an engineering system is to convert matter (e.g., energy, materials, signals). Energy, for example, can be converted from chemical energy to mechanical and thermal energy; materials can be shaped, finished, and coated to provide particular geometries and surface finishes; and signal conversion is often used to control the conversion of energy and materials. The overall function is an abstract formulation of customers needs and is defined in terms of inputs and outputs to a system.

Next, the overall function is decomposed into a hierarchy of generic subfunctions (i.e., energy, material, and signal conversions) that the product must perform in order to meet the overall function. The combination of subfunctions is termed a *function structure*. Use of the function structure facilitates the design task by breaking down the overall, complex function into smaller problems that can be divided among a number of experts. Construction of the function structure starts with the logical analysis of the functions that must appear in the solution if the overall function is to be satisfied. For example, a device for producing small, thermoplastic parts should include the following subfunctions: feed plastic pellets, melt plastic pellets, shape molten plastic into desired part geometry, cool part, dispense part, and control material flow. This function structure may be expanded during later steps of the design process as the designer gains additional insight.

Once a simple function structure is established, the designer searches for suitable solution principles to satisfy each of the subfunctions. Solution principles consist of some underlying physical principle or principles to effect the required conversion and some geometric form. Shear and torsion, for example, are physical principles for transferring mechanical energy, and the screw drive is a geometric form that embodies these effects. Published design catalogs, such as the VDI guidelines (VDI), provide a data base of standardized solution principles which facilitates the solution process. A design concept is obtained by integrating the solution principles for each subfunction.

The quality of the proposed design concept increases with the number of solution principles that are considered for each subfunction. Therefore, multiple solution principles should be identified for each subfunction, and these solution principles can be combined in various permutations to produce a rich solution field. Formulating the required subfunctions and the proposed principles in matrices provides a convenient scheme for organizing the results (see Figure 11.5.1). Each row of the matrix refers to a subfunction and the columns of the matrix contain combinations of solution principles.

The reduction of the solution field to a few promising proposals for further development involves a number of ad hoc strategies. One strategy relies on the experience of the designer to identify those

Sub-functions		Solution Concepts					
		1	2	...	j	...	m
1	F_1	S_{11}	S_{12}		S_{1j}		S_{1m}
2	F_2	S_{21}	S_{22}		S_{2j}		S_{2m}
:	:	:	:		:		:
i	F_i	S_{i1}	S_{i2}		S_{ij}		S_{im}
:	:	:	:		:		:
n	F_n	S_{n1}	S_{n2}		S_{nj}		S_{nm}

FIGURE 11.5.1 Matrix approach for developing solution concepts based on mapping different principles S_{xy} to each subfunction F_x.

combinations of solution principles that are not compatible with customers needs or with one another. A second strategy is to use value analysis to evaluate and compare the solution variants generated by the matrix combination. The general procedure for value analysis involves identifying evaluation criteria and assigning relative weights to each criterion. Each solution variant is then assigned a value by combining its weighted value for each criterion. The solution variants are then compared on the basis of their total weighted value, and the best solution is selected.

The selected solution concept is then developed in terms of its layout and detail specifications. Much of the design development relies on design rules and knowledge readily available in handbooks and design guidelines. In addition to the development of production drawings and specifications, detail design often involves optimization of the solution principles and their geometric forms. The reader is referred to Pahl and Beitz (1988) for additional information.

The Taguchi Method

The Taguchi method (1987, 1993) provides a mathematical basis for analyzing product robustness and is intended as a guide for improving design quality. According to Taguchi, higher quality products satisfy customers needs with less variation and are manufactured at lower cost. This definition of quality is based on a cost model that emphasizes the costs associated with product variability. Costs due to variation include those incurred during manufacturing (such as the cost of materials, machine adjustments, and scraps) and those assumed by the customer (e.g., the cost to repair and/or replace the deficient product). The significance of the Taguchi cost model is the fact that it forces the designer to focus on *offline quality control* as a means to improve product quality.

Offline quality control (QC) is a strategy for reducing variability and improving quality during product design and process planning. It is an attempt to take advantage of the fact that most of the cost is committed during the early stages of product development, while only a very small percentage of the cost is actually expended. In contrast, the concept of *online quality control* has been used historically to advance quality products during manufacturing operations. However, during this stage of the product life cycle, much of the cost has been expended and committed. Statistical process control is an example of online quality control.

The practice of offline QC divides the design process into three stages: system design, parameter design, and tolerance design. System design refers to the conceptual phase of design when customers needs are formulated into a design problem and solutions are generated and evaluated. Once a solution is established and defined in terms of its characteristic attributes or parameters, parameter design is used to establish appropriate parameter settings so as to reduce the design's sensitivity to uncontrollable sources of variation. Taguchi refers to these conditions as *noise factors*. Environmental factors (e.g., people and ambient temperature) and time-dependent phenomena (such as tool wear and material

shrinkage) are examples of noise factors. As a last step, tolerance design is used to enhance or fine-tune the quality improvements realized during parameter design. However, tolerance design often involves a trade-off between the improved quality and the increased production costs incurred by tightening tolerances.

Parameter and tolerance design rely on experimental and mathematical techniques to determine the best design, subject to various noise factors. Various strategies may be used to conduct the experiments. Consider a system that involves five design parameters and three levels of settings — high, medium, and low — for each parameter. The simplest strategy is to investigate all possible combinations of parameters and level settings, which would require $3^5 = 243$ sets of experiments. Instead, orthogonal arrays are an efficient and economical alternative to complete enumeration.

Figure 11.5.2 shows a portion of the L18 orthogonal array that enables a reduction in the work scope from 243 to 18 sets of experiments. The column headings A, B, C. D, and E correspond to each of the five design parameters or controllable factors, and the column entries refer to the three level settings 1, 2, and 3. Each row of the array defines an experiment as a unique combination of parameter settings. Any two columns have nine combinations of level settings — (1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), and (3,3) — and each combination appears with the same frequency. This balance is indicative of the array's orthogonality in a statistical sense because the influence of each control factor can be evaluated independently. The basic array can be extended to include the influence of noise factors and their level settings.

Experiment	Parameter Level Setting				
	A	B	C	D	E
1	1	1	1	1	1
2	2	2	2	2	2
3	3	3	3	3	3
4	1	1	2	2	3
5	2	2	3	3	1
6	3	3	1	1	2
7	1	2	1	3	2
8	2	3	2	1	3
9	3	1	3	2	1
10	1	3	3	2	2
11	2	1	1	3	3
12	3	2	2	1	1
13	1	2	3	1	3
14	2	3	1	2	1
15	3	1	2	3	2
16	1	3	2	3	1
17	2	1	3	1	2
18	3	2	1	2	3

FIGURE 11.5.2 Portion of L18 orthogonal array to investigate all combinations of five parameters with three level settings.

Analysis of variation techniques provide a mathematical basis for organizing and interpreting the experimental results. For each performance requirement, signal-to-noise (S/N) ratio is used to express how sensitive each design parameter is to uncontrollable noise (which is indicative of *design robustness*). A robust design has a high S/N ratio and performs well despite the influence of noise. Many equations are available for calculating S/N ratios, and the selection of the appropriate equation is a function of the type of cause–effect relationship under study. Examples include *higher-is-better, lower-is-better,* and *nominal-is-best*. The reader is referred to Taguchi (1987, 1993) for further information.

Design for Assembly

Design for assembly (DFA) supports the analysis and design of products for ease of assembly. Based on a collection of empirical time-study data, if–then rules, and design checklists, DFA methods and tools help the designer to focus on the relationship between the geometric features of a design and its

components and the effort and resources necessary to assemble these components into the desired product. Philips International, Hitachi, Xerox, Ford, and General Electric have been industry leaders in the adoption and dissemination of DFA. In this section the most widely known method, advanced by Boothroyd and Dewhurst (1985), is considered. The method gives the total assembly time and design efficiency, and these measures are then used to guide redesign.

The method of Boothroyd and Dewhurst (1985) starts with the description of a product's assembly sequence in terms of components, component features, and the basic assembly tasks of handling and insertion (which includes fastening). For each component involved in the assembly sequence, a catalog of generic part features is used to classify the component with respect to its difficulty for handling and insertion. These classifications are then used to determine the time required for each assembly task, and the task times are added to give the total operation time for the component. Therefore, the total assembly time (TM) for the design is the sum of the operation times for each component.

Design efficiency is the ratio of an estimated, ideal assembly time to TM. The ideal assembly time is based on a theoretical minimum number of parts (NM). The estimation of NM relies on a checklist for identifying redundant components or components that may be combined to reduce parts count. For manual assembly, the ideal assembly time is specified as 3 sec per part, and design efficiency EM is given by the equation

$$EM = 3 \times NM/TM \tag{11.5.1}$$

The resulting values for assembly time and design efficiency provide the designer with a basis for redesign. Suggested strategies for redesign include reduction in parts count, use of symmetry to facilitate handling, and use of alternative fastening mechanisms to facilitate insertion (e.g., snap-fit elements are faster to insert than parts requiring screw tightening). The reader is referred to Boothroyd and Dewhurst (1985) for further information.

References

Boothroyd, G. and Dewhurst, P. 1985. *Design for Assembly Handbook.* Boothroyd and Dewhurst Associates, Kingston, RI.

Hubka, V. and Eder, W.E. 1988. *Theory of Technical Systems.* Springer-Verlag, New York.

Pahl, G. and Beitz, W. 1988. *Engineering Design.* The Design Council, London.

Taguchi, G. 1987. *System of Experimental Design,* translated by Louise Watanabe Tung. Kraus International Publications, White Plains, NY.

Taguchi, G. 1993. *Taguchi on Robust Technology Development*. translated by S-C Tsai. ASME Press, New York.

VDI. 1987. *VDI Design Handbook 2221: Systematic Approach to the Design of Technical Systems and Products.* German Standards Institute, Dusseldorf.

11.6 Strategies for Product Design

Michael Pecht

Requirements

The design team must have a clear appreciation of user requirements and constraints before a design can begin. Requirements are established within the context of product effectiveness attributes and include limits on parameters such as speed, miles per gallon, computations per second, and accuracy, as well as constraints on size, weight, reliability, and cost. Often, contractual requirements and company policy can dictate a product type, such as the use of an in-house technology or product, or the exclusion of a technology or product.

Requirements must be addressed holistically in the design of a product and must not be limited to those that affect the product's immediate performance. For example, the supportability of the product is a critical product requirement for many products in terms of the ease of maintenance and accessibility to the products, spares, support equipment, the time and required personnel to repair the product, and the ease of discovering or detecting where a problem occurs. A poor definition of requirements could lead to a design where, for example, the air conditioning compressor has to be removed to replace a spark plug, or where special tools are required to replace the oil in a car. Note also that the design team should be aware of the impact of administrative policies on the reliability of products as reflected in the product's scenario for use. For example, outdated standards can contribute to failures that are not the result of either faulty design or human error.

The Life Cycle Usage Environment

The life cycle usage environment or scenario for use of a product goes hand-in-hand with the product requirements. The life cycle usage information describes the storage, handling, and operating stress profiles and thus contains the necessary "load" input information for effective assessment and development of design guidelines, screens, and tests. The stress profile of a product is based on the application profile and the internal stress conditions of the product. Because the performance of a product over time is often highly dependent on the magnitude of the stress cycle, the rate of change of the stress, and even the time and spatial variation of stress, the interaction between the application profile and internal conditions must be specified. Specific information about the product environment includes absolute temperature, temperature ranges, temperature cycles, temperature gradients, vibrational loads and transfer functions, chemically aggressive or inert environments, and electromagnetic conditions. The life cycle usage environment can be divided into three parts: the application and life profile conditions, the external conditions in which the product must be stored, handled, and operated, and the internal product-generated stress conditions.

The application and life profile conditions include the application length, the number of applications in the expected life of the product, the product life expectancy, the product utilization or nonutilization (storage, testing, transportation) profile, the deployment operations, and the maintenance concept or plan. This information is used to group usage platforms (i.e., whether the product will be installed in a car, boat, satellite, underground), develop duty cycles (i.e., on–off cycles, storage cycles, transportation cycles, modes of operation, and repair cycles), determine design criteria, develop screens and test guidelines, and develop support requirements to sustain attainment of reliability and maintainability objectives. The external operational conditions include the anticipated environment(s) and the associated stresses that the product will be required to survive. The stresses include temperature, vibrations, shock loads, humidity or moisture, chemical contamination, sand, dust and mold, electromagnetic disturbances, radiation, etc.

The internal operational conditions are associated with product-generated stresses, such as power consumption and dissipation, internal radiation, and release or outgassing of potential contaminants. If

the product is connected to other products or subsystems in a system, stresses associated with the interfaces (i.e., external power consumption, voltage transients, electronic noise, or dissipation) must also be included.

Characterization of Materials, Products, and the Manufacturing Processes

Design is intrinsically linked to the materials, products, interfaces, and manufacturing processes used to establish and maintain functional and structural integrity. It is unrealistic and potentially dangerous to assume defect-free and perfect-tolerance materials, products, and structures. Materials often have naturally occurring defects and manufacturing processes can induce additional defects to the materials, products, and structures. The design team must also recognize that the production lots or vendor sources for products that comprise the design are subject to change. Even greater variability in products characteristics is likely to occur during the fielded life of a product as compared to its design or production life cycle phases.

Design decisions involve the selection of products, materials, and controllable process techniques, using tooling and processes appropriate to the scheduled production quantity. Often, the goal is to maximize product and configuration standardization; increase package modularity for ease in fabrication, assembly, and modifications; increase flexibility of design adaptation to alternate uses; and utilize alternate fabrication processes.

Design Guidelines and Techniques

Generally, new products replace existing products. The replaced product can be used for comparisons (i.e., a baseline comparison product). Lessons learned from the baseline product can be used to establish new product parameters, to identify areas of focus in the new product design, and to avoid the mistakes of the past.

Once the products, materials, processes, and stress conditions have been characterized, design begins. In using design guidelines, there may not be a unique path to follow. Multiple branches may exist depending upon the input design constraints. The design team should explore enough of the branches to gain confidence that the final design is the best for the prescribed input information. The design team should also use guidelines for the complete design and not those limited to specific aspects of an existing design. This statement does not imply that guidelines cannot be used to address only a specific aspect of an existing design, but the design team may have to trace through the implications that a given guideline suggests.

Design guidelines that are based on physics of failure models can also be used to develop tests, screens, and derating factors. Tests can be designed from the physics of failure models to measure specific quantities and to detect the presence of unexpected flaws, manufacturing, or maintenance problems. Screens can be designed to precipitate failures in the weak population while not cutting into the design life of the normal population. Derating or safety factors can be determined to lower the stresses for the dominant failure mechanisms.

Preferred Products

In many cases, a product or a structure much like the required one has already been designed and tested. This is called a “preferred product or structure” in the sense that variabilities in manufacturing, assembly, and field operation that can cause problems have been identified and corrected. Many design teams maintain a list of preferred products and structures with acceptable performance, cost, availability, and reliability.

Redundancy

Redundancy permits a product to operate even though certain components and interconnections have failed. Redundant configurations can be classified as either active or standby. Elements in active redundancy operate simultaneously in performing the same function. Elements in standby redundancy are

designed so that an inactive one will be switched into service when an active element fails. The reliability of the associated function is increased with the number of standby elements (optimistically assuming that the sensing and switching products of the redundant configuration are working perfectly, and failed redundant components are replaced before their companion component fails).

A design team may often find that redundancy is a way to improve product reliability if time is of prime importance, if the products requiring redundancy are already designed and the products are known to be of poor reliability.

On the other hand, in weighing its advantages, the design team may find that redundancy will

- Prove too expensive, if the products and redundant sensors and switching products are costly
- Exceed the limitations on size and weight, particularly in avionics, missiles, and satellites
- Exceed the power limitations, particularly in active redundancy
- Attenuate the input signal, requiring additional amplifiers (which increase complexity)
- Require sensing and switching circuitry so complex as to offset the reliability advantage of redundancy

Protective Architectures

It is generally desirable to include means in a design for preventing a product, structure, or interconnection from failing catastrophically, and instead allow it to fail safely.

Fuses and circuit breakers are examples used in electronic products to sense excessive current drain and disconnect power from a failed circuit. Fuses within circuits safeguard products against voltage transients or excessive power dissipation and protect power supplies from shorted parts. Thermostats may be used to sense critical temperature, limiting conditions and shutting down the product or a component of the system until the temperature returns to normal. In some products, self-checking circuitry can also be incorporated to sense abnormal conditions and operate adjusting means to restore normal conditions or activate switching means to compensate for the malfunction.

Protective architectures can be used to sense failure and protect against possible secondary effects. In other cases, means can be provided for preventing a failed product or structure from completely disabling the product. For example, a fuse or circuit breaker can disconnect a failed product from a product in such a way that it is possible to permit partial operation of the product after the failure, in preference to total product failure. By the same reasoning, degraded performance of a product after failure of a product is often preferable to complete stoppage. An example is the shutting down of a failed circuit whose design function is to provide precise trimming adjustment within a deadband of another control product. Acceptable performance may thus be permitted, perhaps under emergency conditions, with the deadband control product alone.

Sometimes the physical removal of a part from a product can harm or cause failure of another part of the product by affecting load, drive, bias, or control. In some cases, self-healing techniques can be employed to self-check and self-adjust to effect changes automatically to permit continued operation after a failure.

The ultimate design, in addition to its ability to act after a failure, would be capable of sensing and adjusting for drifts to avert failures.

In the use of protective techniques, the basic procedure is to take some form of action after an initial failure or malfunction, to provide perhaps reduced performance, and to prevent additional or secondary failures. Such techniques can be considered as enhancing product reliability, although they also affect availability and product effectiveness. No less a consideration is the impact of maintenance, repair, and product replacement. If a fuse protecting a circuit is replaced, what is the impact when the product is reenergized? What protective architectures are appropriate for postrepair operations? What maintenance guidance must be documented and followed when fail-safe protective architectures have or have not been included?

Stress Margins

Products and structures should be designed to operate satisfactorily at the extremes of the parameter ranges, and allowable ranges must be included in the procurement or reprocurement specifications. To guard against out-of-tolerance failures, the design team must consider the combined effects of tolerances on products, subsequent changes due to the range of expected environmental conditions, drifts due to aging over the period of time specified in the reliability requirement, and tolerances in products used in future repair or maintenance functions.

Methods of dealing with product and structural parameter variations include statistical analysis and worst-case analysis. In statistical design analysis, a functional relationship is established between the output characteristics of the structure and the parameters of one or more of its products. In worst-case analysis, the effect that a product has on product output is evaluated on the basis of end-of-life performance values or out-of-specification replacement products.

Derating

The principle of derating is that there are no distinct stress boundaries for voltage, current, temperature, power dissipation, etc. above which immediate failure can occur and below which the product will operate indefinitely. Instead, the life of some products increases in a continuous manner as the stress level is decreased below the rated value. Practically, however, there are minimum stress levels below which increased derating will lower reliability, and the complexity required to step up performance will offset any gain in reliability.

Size and Weight Control

Methods to reduce product volume include:

- Eliminating unused space
- Selecting an exterior shape to require least possible volume when integrated with the intended installation structure
- Eliminating separable interconnections, using optimally shaped components and subassemblies
- Minimizing interconnection requirements
- Eliminating support components such as heat sinks, blowers, and special coolant flow passages
- Eliminating redundant components
- Using components and package elements to perform multiple functions

Methods to lower product weight include using high-strength lightweight materials, eliminating duplicate structures, and providing only that degree of strength required to reach a desired safety factor.

Potential Failure Sites and Failure Mechanisms

The design team must evaluate the potential failure mechanisms, failure stresses, failure sites, and failure modes, given a product architecture, the comprising products and materials, and the manufacturing processes. One approach to evaluation consists of the assessment of a list of candidate or "potential" failure mechanisms. The load conditions that cause the failure mechanisms to occur are then determined in light of the life cycle usage environment identified earlier. Table 11.6.1 presents various failure mechanisms and the associated load conditions.

Once the failure mechanisms and associated stresses have been identified, then failure sites are specified and a detailed reliability assessment is conducted for each location suspected to be a potential area of failure. Table 11.6.2 presents an example case of various failure sites, mechanisms, and load structures for a microelectronic product.

Failures can be categorized by their impact on end product performance. A critical failure is an event that reduces the performance of the end product to unacceptable levels. A noncritical failure, also called a fault, does not reduce performance to unacceptable levels. Failures can also be classified by the time

TABLE 11.6.1 The Load Conditions which Cause Failure Mechanisms

Failure Mechanism	Containment	Δ T (Temperature Cycle Magnitude)	RH (Relative Humidity)	T (Steady-State Temperature)	Vibration/Shock	Maintenance and Handling	Voltage
Brittle fracture		X		X	X	X	
Ductile fracture				X	X		
Yield		X		X	X		
Buckling		X		X	X	X	
Large elastic deformation		X			X		
Interfacial deadhesion	X	X	X	X	X		
Fatigue crack initiation	X	X	X	X	X		
Wear					X		
Creep		X		X	X		
Corrosion	X		X	X		X	
Dendritic growth	X		X			X	X
Fatigue crack propagation		X			X	X	
Diffusion				X			X

TABLE 11.6.2 Example: Failure Sites, Operational Loads, and Failure Mechanisms for a Microelectronic Product

Site	Containment	Δ T (Temperature Cycle Magnitude)	RH (Relative Humidity)	T (Steady-State Temperature)	Vibration/Shock	Voltage	Maintenance and Handling
Die		A, L, G	I		A, L, G		
Die attach		A, F, L, G			A, L, G, C		
Flip-chip solder joints	I, K	L, G	I	B, C, I, N	L, G	M	
Tape automated bonds	I, J, K	L, G	I, J	A, B, I, J, M	L, G	M, J	
Wire	I	G, L	I		G, L, E		
Leads	I, J	L	I, J	I	D, L, H	L, J	D, G
Substrate		L, G			A, B, L, G		
Substrate attach		F, L, G			C, L, G		

Note: A — brittle fracture; B — ductile fracture; C — yield; D — buckling; E — large elastic deformation; F — interfacial deadhesion; G — fatigue crack initiation; H — wear; I — corrosion; J — dendritic growth; K — interdiffusion; L — fatigue crack propagation; M — diffusion.

frame and manner in which the event occurs. The categories are catastrophic, intermittent, out-of-tolerance, and maladjustment.

1. **Catastrophic failures.** A catastrophic failure occurs when a product becomes completely inoperative or exhibits a gross change in characteristics. Examples are an open or shorted resistor or capacitor, a leaky valve, a stuck relay, or a broken switch.
2. **Intermittent failures.** Intermittent failures are nonperiodic failures that can occur within products, interconnects, or product interfaces including software. Examples include switch bouncing and poor transmission of signals. The design team must not only safeguard against the effects of intermittent failures, but also avoid creating possible modes of such failures (e.g., conditions permitting product-to-product or structural breakdowns).
3. **Out-of-tolerance failures.** Out-of-tolerance failures result from degradation, deterioration, drift, and wearout. Examples are the drifting of resistor and capacitor values and the wearing out of relays and solenoid valves and precision gears. The changes can arise as a result of time, temperature or temperature cycles, humidity, altitude, etc. When these changes, considered collectively, reach the point where product performance is below acceptable limits, we say that the product has failed.
4. **Maladjustment failures.** Maladjustment failures are often due to human error. Failures arise from improper adjustment of products, as well as abuse of adjusting products due to lack of understanding of the adjustments and the capabilities of the products. Such failures are hard to evaluate and difficult to avoid but must be considered if the reliability of the product is to be sustained.

Designing for the Application Environment

The design of modern products requires that the team be acquainted with the environmental factors affecting product performance over time. The environment is seldom forgiving and when a weak point exists, performance suffers. Design teams must understand the environment and its potential effects and then must select designs or materials that counteract these effects or provide methods to alter or control the environment within acceptable limits.

In addition, a product can create or perturb an environment. For example, many epoxies out-gas during cure, releasing corrosive or degrading volatiles or particulates into the product. Teflon may release fluorine and polyvinylchloride (PVC) may release chlorine. Certain solid rocket fuels are degraded into a jelly-like mass when exposed to aldehydes or ammonia, either of which can come from a phenolic nozzle cone. Common environmental considerations follow.

Temperature Control

Temperature is a powerful agent for electrical, chemical, and physical deterioration for two basic reasons:

- The physical properties of almost all known materials are modified by changes in temperature, temperature gradients, and temperature extremes.
- The rate of most chemical reactions is influenced by the temperature of the reactants.

Nevertheless, the effects of steady-state temperature and spatial and temporal temperature gradients must be clearly understood before a thermal control method is determined. A good practice is to develop a table of potential failure sites and failure mechanisms and the effects of temperature in its many forms.

Poor heat removal of dissipated I^2R losses, hysteresis, or eddy current losses can result in physical damage or accelerated chemical reaction rates. The latter occurrence, affecting certain types of materials, can cause general degradation and catastrophic or intermittent failures. But the fault is not always with the design; air intakes or outlets could be blocked because of faulty installation, or maintenance personnel might fail to clear or replace air filters.

Thermal management schemes for electronic equipment are selected, once the heat flux and the maximum temperature difference available for heat transfer have been identified. A complete thermal

management strategy involves selecting an appropriate heat removal scheme for each level of packaging. For low heat fluxes, passive thermal management techniques can be used. Such techniques do not require the expenditure of external energy for the heat removal. Interest in such techniques is currently very strong, due to their design simplicity, low cost, and high reliability.

Examples of passive cooling techniques include conduction cooling using high thermal conductivity substrates and/or heat sinks and natural convection air cooling. Conduction cooling can be employed when the heat source is not directly exposed to the coolant. This method is desirable in dense electronic packages, where radiation and convection are not effective cooling means. Natural convection cooling relies on the buoyancy-induced flows generated due to heating. For a maximum temperature rise of 85°C above the environment, heat fluxes of up to about 0.1 W/cm^2 can be removed by natural convection air cooling under nominal sea-level environmental conditions.

For higher heat fluxes, a combination of passive and active techniques can sometimes be used. This may include, for example, a high thermal conductivity substrate at the board level, along with forced air cooling at the box level. For this range of heat fluxes, the use of cold plate technology, thermoelectrics, and flow-through cooling are also possible; other common techniques are the use of heat pipe or thermosyphon (see Chapter 4).

Besides the out-gassing of corrosive volatiles when subjected to heat, almost all known materials will expand or contract when their temperature is changed. This expansion and contraction causes problems with fit between product interface and interconnections. Local stress concentrations due to nonuniform temperature are especially damaging, because they can be so high. A familiar example is a hot water-glass that shatters when immersed in cold water. Metal structures, when subjected to cyclic heating and cooling, may ultimately collapse due to the induced stresses and fatigue caused by flexing. The thermocouple effect between the junction of two dissimilar metals causes an electric current that may induce electrolytic corrosion.

Plastics, natural fibers, leather, and both natural and synthetic rubber are all particularly sensitive to temperature extremes as evidenced by their brittleness at low temperatures and high degradation rates at high temperatures.

Shock and Vibration Control

Shock and vibration can harmfully flex leads and interconnects, dislodge parts or foreign particles into bearings, pumps, and electronics, cause acoustical and electrical noise, and lead to structural instabilities. Protective measures against shock and vibration stresses are generally determined by an analysis of the deflections and mechanical stresses produced by these environmental factors. This generally involves the determination of natural frequencies and evaluation of the mechanical stresses within components and materials produced by the shock and vibration environment. If the mechanical stresses so produced are below the acceptable safe working stress of the materials involved, no direct protection methods are required. If, on the other hand, the stresses exceed the safe levels, corrective measures such as stiffening, reduction of inertia and bending moment effects, and incorporation of further support members are indicated. If such approaches do not reduce the stresses below the acceptable safe levels, further reduction is usually possible by the use of shock-absorbing mounts.

Products are sometimes specially mounted to counter the destructive effects of shock and vibration. Shock mounts often serve this purpose, but effective means for attenuating shock and vibration simultaneously are complex. Isolation of a product against the effects of vibration requires that the natural frequency of the product be substantially lower than the undesired frequency of vibration.

Three basic kinds of isolators are available:

- Elastomers made of natural or synthetic rubber, used in a shear mode or in a diaphragm to dampen the induced shock or vibration.
- Metallic isolators, such as springs, metal meshes, or wire rope. Springs lack good damping qualities, but meshes and rope provide smooth friction damping.

- Viscous dampers (similar to the type used on automobiles) which are velocity-sensitive, tend to become ineffective under high-frequency vibration. Resilient mounts must be used with caution, since they can amplify the intensity of shock and vibration if improperly placed. The ideal goal is to design a product to be resistant to shock and vibrations, rather than to provide complete isolation.

Another vibration protective technique is potting or encapsulation of small parts and assemblies. The potting material should be compliant enough to dampen vibrations. Some materials polymerize exothermically, and the self-generated heat may cause cracking of the casting or damage to heat-sensitive products. Another problem is shrinkage of the potting material. In some cases, molds are made oversized to compensate for shrinkage.

One factor that is not often considered is that the vibration of two adjacent components, or separately insulated subsystems, can cause a collision if maximum excursions and sympathetically induced vibrations are not accounted for in design. Another failure mode, fatigue (the tendency for a material to break under cyclic stress loads considerably below its tensile strength), includes high cycle fatigue, acoustic fatigue, and fatigue under combined stresses such as temperature extremes, temperature fluctuations, and corrosion.

In addition to using proper materials and configuration, it may still be necessary to control the amount of shock and vibration experienced by the product. Damping products are used to reduce peak oscillations and special stabilizers can be employed when unstable configurations are involved. Typical examples of dampeners are viscous hysteresis, friction, and air damping. Vibration isolators commonly are identified by their construction and material used for the resilient elements (rubber, coil spring, woven metal mesh, etc.). Shock isolators differ from vibration isolators in that shock requires a stiffer spring and a higher natural frequency for the resilient element. Some of the types of isolation mounting products are underneath, over-and-under, and inclined isolators. In some cases, however, even though an item is properly insulated and isolated against shock and vibration damage, repetitive forces may loosen the fastening products. If the fastening products loosen enough to permit additional movement, the product will be subjected to increased forces and may fail. Many specialized self-locking fasteners are commercially available.

Chemical Action Control

The earth's environment contains numerous deteriorators including oxygen, carbon dioxide, nitrogen, snow, ice, sand, dust, salt water spray, organic matter, and chemicals in general. Product specifications often state limits on temperature, humidity, altitude, salt spray, fungus, sunshine, rain, sand, and dust.

A material or structure can chemically change in a number of ways. Among them are interactions with other materials (i.e., metal migration, diffusion) and modifications in the material itself (recrystallization, stress relaxation, phase changes, or changes induced by irradiation). In addition to the deterioration problems associated with the external environments to which products are subjected, adhesives, batteries, and certain types of capacitors are susceptible to chemical aging and biological growths.

Materials widely separated in the electromotive series are subject to galvanic action, which occurs when two dissimilar metals are in contact in a liquid medium. The most active metal dissolves, hydrogen is released, and an electric current flows from one metal to the other. Coatings of zinc are often applied to iron so that the zinc, which is more active, will dissolve and protect the iron (a process known as *galvanization*). Galvanic action is known to occur within the same piece of metal, if one portion of the metal is under stress and has a higher free-energy level than the other. The part under stress will dissolve if a suitable liquid medium is present. Stress-corrosion cracking occurs in certain magnesium alloys, stainless steels, brass, and aluminum alloys. It has also been found that a given metal will corrode much more rapidly under conditions of repeated stress than when no stress is applied.

Corrosion-resistant materials should be used as much as possible. Although aluminum is excellent in this respect due to the thin oxide coating that forms when it is exposed to the atmosphere, it tends to

pit in moist atmospheres. Furthermore, aluminum may corrode seriously in a salt-laden marine atmosphere. Either anodization or priming with zinc chromate can provide additional protection.

Magnesium is sometimes used to minimize the weight of products but must be protected by surface additives against corrosion and electrolytic reaction. A coating of zinc chromate serves both purposes. Because magnesium is the most reactive metal normally used for structural purposes, the grounding of the magnesium structure requires careful selection of another metal for making the connection. Zinc- or cadmium-plated steel are the more commonly chosen connective materials.

Copper, when pure, is quite resistant to corrosion. However, under certain conditions, copper-made products will become chemically unstable. Transformers have occasionally failed due to impurities in the insulation and moisture. These conditions favor electrolytic reaction, which causes the copper to dissolve and eventually results in an open circuit.

Iron and steel are used in many products and structures because of their good magnetic and structural properties; however, only certain stainless steels are reasonably resistant to corrosion. Various types of surface plating are often added, but since thin coatings are quite porous, undercoatings of another metal such as copper are often used as a moisture barrier. Cadmium and zinc plating in some marine environments corrode readily, and often grow "whiskers", which can cause short circuits in an electronic product.

Proper design of a product therefore requires trade-offs in

- Selecting corrosion-resistant materials
- Specifying protective coatings if required
- Avoiding use of dissimilar metallic contacts
- Controlling metallurgical factors to prevent undue internal stress levels
- Preventing water entrapment
- Using high-temperature resistance coatings when necessary
- Controlling the environment through dehydration, rust inhibition, and electrolytic and galvanic protective techniques

Biological Growth Control

High humidity and warm temperatures often favor the growth of fungus, which can lead to deterioration of many electrical and mechanical properties. Products designed for use under tropical conditions can receive a moisture- and fungus-proofing treatment, which consists of applying a moisture-resistant varnish containing a fungicide. Air-drying varnishes generally are not as moisture-proof as are the baked-on types. This treatment is usually quite effective, but the useful life of most fungicides is less than that of the varnish, and retreatments may be necessary.

Moisture Control

Moisture, and impurities that may be contained with it, are known to cause corrosion of many metal systems. In addition, mated products can be locked together, especially when moisture condenses on them and then freezes. Similarly, many materials that are normally pliable at low temperatures become hard and brittle if moisture has been absorbed and subsequently freezes. Condensed moisture acts as a medium for the interaction between many, otherwise relatively inert, materials. Most gases readily dissolve in moisture. The volume increase from water freezing (i.e., converting from a fluid to solid state) can also physically separate components, materials or connections. The chlorine released by PVC plastic, for example, forms hydrochloric acid when combined with moisture.

Although the presence of moisture may cause deterioration, the absence of moisture also may cause reliability problems. The useful properties of many nonmetallic materials such as leather and paper, which become brittle and crack when they are very dry, depend upon an optimum level of moisture. Similarly, fabrics wear out at an increasing rate as moisture levels are lowered, and fibers become dry and brittle. Environmental dust can cause increased wear, friction, and clogged filters due to lack of moisture.

Moisture, in conjunction with other environmental factors, creates difficulties that may not be characteristic of the factors acting alone. For example, abrasive dust and grit, which would otherwise escape, can be trapped by moisture. The permeability (to water vapor) of some plastics (PVC, polystyrene, polyethylene, etc.) is related directly to their temperature. The growth of fungus is enhanced by moisture, as is the galvanic corrosion between dissimilar metals. Some design techniques that can be used separately or combined to counteract the effects of moisture are

- Elimination of moisture traps by providing drainage or air circulation
- Using desiccant products to remove moisture when air circulation or drainage is not possible
- Applying protective coatings
- Providing rounded edges to allow uniform coating of protective material
- Using materials resistant to moisture effects, fungus, corrosion, etc.
- Hermetically sealing components, gaskets, and other sealing products
- Impregnating or encapsulating materials with moisture-resistant waxes, plastics, or varnishes
- Separation of dissimilar metals or materials that might combine or react in the presence of moisture or of components that might damage protective coatings.

The design team also must consider possible adverse effects caused by specific methods of protection. Hermetic sealing, gaskets, protective coatings, etc. may, for example, aggravate moisture difficulties by sealing moisture inside or contributing to condensation. The gasket materials must be evaluated carefully for outgassing of corrosive volatiles or for incompatibility with adjoining surfaces or protective coatings.

Sand and Dust Protection

In addition to the obvious effect of reduced visibility, sand and dust can degrade a product by abrasion leading to increased wear, friction causing both increased wear and heat, and clogging of filters, small apertures, and delicate products. Thus, products having moving parts require particular care when designing for sand and dust protection. Sand and dust will abrade optical surfaces, either by impact when being carried by air or by physical abrasion when the surfaces are improperly wiped during cleaning. Dust accumulations have an affinity for moisture and when combined may lead to corrosion or the growth of fungus.

In relatively dry regions, such as deserts, fine particles of dust and sand can readily be agitated into suspension in the air, where they may persist for many hours, sometimes reaching heights of several thousand feet. Thus, even though there is virtually no wind present, the speeds of vehicles or vehicle-transport product through these dust clouds can also cause surface abrasion by impact.

Although dust commonly is considered to be fine, dry particles of earth, it also may include minute particles of metals, combustion products, and solid chemical contaminants. These other forms may provide direct corrosion or fungal effects on products, because this dust may be alkaline, acidic, or microbiological.

When products require air circulation for cooling or removing moisture, the question is not whether to allow dust to enter, but rather how much or what size dust can be tolerated. The problem becomes one of filtering the air to remove dust particles above a specific nominal size. The nature of filters, however, is such that for a given working filter area, as the ability of the filter to stop increasingly smaller dust particles is increased, the flow of air or other fluid through the filter is decreased. Therefore, the filter surface area either must be increased, the flow of fluid through the filter decreased, or the allowable particle size increased (i.e., invariably, there must be a compromise). Interestingly enough, for aircraft engines, the amount of wear is proportional to the weight of ingested dust, but inversely proportional to dust size.

Sand and dust protection must be planned in conjunction with protective measures against other environmental factors. For example, it is not practical to specify a protective coating against moisture if sand and dust will be present, unless the coating is carefully chosen to resist abrasion and erosion or is self-healing.

Explosion Control

Protection against explosion is both a safety and reliability problem. An item that randomly exhibits explosive tendencies is one that has undesirable design characteristics and spectacular failure modes. This type of functional termination, therefore, requires extreme care in design and reliability analyses.

Explosion protection planning must be directed to three categories (not necessarily mutually exclusive) of products:

- Items containing materials susceptible to explosion
- Components located near enough to cause the explosive items to explode
- Product that might be damaged or rendered temporarily inoperative by overpressure, flying debris, or heat from an explosion

The first category includes products containing flammable gases or liquids, suspensions of dust in the air, compounds that spontaneously decompose in certain environments, product containing or subjected to high or low extremes of pressure (includes implosions), or any other products capable of creating an explosive reaction. Keep in mind that even basically inert materials such as glass or metals can explode when subjected to rapid environmental changes or extremes such as temperature or stress.

The second category includes many variations on methods for providing an energy pulse, a catalyst, or a specific condition that might trigger an explosion. A nonexplosive component, for example, could create a corrosive atmosphere, mechanical puncture, or frictional wear on the side of a vessel containing a high-pressure and thereby cause the container to explode.

The third category is important because a potentially explosive product (such as a high-pressure air tank) can be damaged or made to explode from another explosion. Thus, some reasoning must be applied when considering products not defined by the first two categories. From a practical standpoint, explosion protection for items in the third category should be directed to product that might possibly be near explosions.

The possibility of an explosive atmosphere leaking or circulating into other product compartments must also be recognized. Lead-acid batteries, for example, create hydrogen gas that, if confined or leaked into a small enclosure, could be exploded by electrical arcing from motor brushes, by sparks from metallic impacts, or by exhaust gases. Explosive environments, such as dust-laden air, might be circulated by air distribution products. Dust from common ingredients such as wheat flour has been the source of massive, catastrophic explosions.

Explosion protection and safety are very important for design and reliability evaluations and must be closely coordinated and controlled. Just as a safe product is not necessarily reliable, neither is a reliable product necessarily safe; but the two can be compatible.

Electromagnetic Radiation Control

Protection against the effect of electromagnetic radiation has become a sophisticated engineering field of electromagnetic compatibility design. The radiation environment in space near the earth is composed primarily of Van Allen, auroral, solar flare, and cosmic phenomena. Of lesser importance are solar wind, thermal energy atoms in space, neutrons, naturally occurring radon gas, albedo protons, plasma bremsstrahlung, and man-made nuclear sources. In the electromagnetic spectrum are gamma rays, X-rays, ultraviolet, and Lyman-alpha radiation. Damage near the surface of the earth is caused by the electromagnetic radiation in the wavelength range from approximately 0.15 to 5 m. This range includes the longer ultraviolet rays, visible light, and up to about the midpoint of the infrared band. The most direct approach to protection is to avoid the limited region in which high radiation levels are found. When exposure cannot be avoided, shielding and filtering are important protective measures. In other cases, material design changes or operating procedural changes must be instituted in order to provide protection.

High Vacuum Control

In a high vacuum, materials with a high vapor pressure will sublimate or evaporate rapidly, particularly at elevated temperatures. In some plastics, the loss of plasticizer by evaporation will cause cracking, shrinking, or increased brittleness. Metals such as magnesium, which would normally evaporate rapidly (1 g/cm^2/year at 250°C), can be protected by inorganic coatings with low vapor pressures.

In a high vacuum, adjoining solid surfaces can become cold-welded after losing adsorbed gases. Some form of lubrication is therefore necessary. Conventional oils and greases evaporate quickly. Graphite becomes unsatisfactory (actually an abrasive) because of the loss of absorbed water. However, thin films of soft metals, such as lead, silver, or gold, are effective lubricants in a high vacuum. Thin films of molybdenum disulfide are often sprayed over chrome or nickel plating, forming easily sheared layers. The film also releases sulfur at interfaces during sliding, performing the same function as water vapor does for graphite.

Human Factors and Operability

Humans are active participants in the operating of many products, and can be traded as an external "stress." Consequently, the interaction of humans with products must be weighed against safety, reliability, maintainability, and other product parameters to assess product reliability, maintainability time and performance, system and subsystem safety analyses, and specific human engineering design criteria. For convenience and clarity, four types of human interactions with a product are identified:

- Design and production of a product
- Operators and repairers as mechanical elements (human engineering)
- Operators and repairers as decision elements (human performance reliability)
- Bystanders (this classification is not considered further because it is largely a safety matter, unless they consciously or inadvertently operate or affect the operation of a product)

Designing for Operability

While product design teams, planners, and managers would hope to have well-above-average people in every position associated with product operation, product design must accommodate the realities of the product's life cycle use. The complexity of the product that is presented to the operator, the array of components, including ancillary product, cable connectors, patch panels, switches, knobs, controls, panel markings, meters, oscilloscopes, horns, bells, panel lights, etc., must be of concern in design. The various components comprising the product should be clearly labeled for easy identification. Those items that must be manipulated in order to connect the product (cable connectors, patch panels, switches) should be arranged in a simple and logical order and should be plainly marked. Typical constraints are that

- An operator should be within the physical capabilities of the complete range of potential operators.
- A person should not be required to do something that his or her coordination will not allow him or her to do.
- In times of psychological stress, people cannot easily use, read, and respond to controls and displays.

Mock-ups under realistic conditions are very helpful in uncovering potential problems early in the design process. For example, if a product must be used at night in extremely cold weather, have a person try to use it in a freezing, poorly lit room for several hours.

Designing for Maintainability

Few operational products can be perfectly reliable while meeting other product trade-offs such as rust. Maintenance can thus be an important consideration in the long-term effectiveness of a product.

Maintainability is the probability that, when a specified maintenance action is initiated, a failed product will be restored to operable conditions in a specified downtime. Thus, design features that will expedite maintenance will enhance maintainability. Designing for maintainability means inclusion in the design of those features that can be conceived to assist the maintenance technician. Specific features include the degree of accessibility for product replacement, facilities for fault isolation, special tools or test-product requirements, the level of servicing skills required, servicing documentation requirements, and spare-part stocking requirements.

Critical rating factors include grouping of components by electrical function, use of integral fault indication for basic modules; components or functional assemblies removable without interruption of permanent electrical connections; elimination of tool requirements for mechanical disassembly; direct access to removable assemblies; products commonality; and identification of replaceable components.

In checking out a product, the maintenance technician must often also operate it. Therefore, the design considerations relating to operability are, in general, also applicable to maintainability. In addition, many design considerations relating to maintenance are more complex than those relating to operation.

The design team needs to consider how the product actually will be repaired in the field, perhaps under the pressures of understaffed maintenance crews, many of whom are inexperienced, or by field service personnel with limited knowledge about the product. The design team must always keep in mind that a product may be used and repaired by people who have other things in mind than "babying" the product. The design team must also realize the difference between what people will probably do and what the design team thinks they ought to do.

The design team, in acknowledging that a product could fail if not maintained, should provide means for ease of maintenance, ease of removal of a failed unit, trouble shooting, access to the failed unit, and repair or replacement. Such means, in addition to improving maintainability, will also improve reliability through the averting of subsequent failure due to human errors during maintenance. Many design details are important to the maintenance technician. A list of "Do's" and "Don'ts" in designing for maintainability is presented in Table 11.6.3.

TABLE 11.6.3 Warning to Design Teams: Some Common Design Errors Affecting Maintainability

Don't place products or maintenance structures (e.g., oil filter) where they cannot be removed without removal of the whole unit from its case or without first removing other products
Don't put an adjustment out of arm's reach
Don't neglect to provide enough room for the technician to get his or her gloved hand into the unit to make an adjustment
Don't screw subassemblies together in such a way that the maintenance technician cannot tell which screw holds what
Don't use chassis and cover plates that fall when the last screw is removed
Don't make sockets and connectors for modules of the same configuration so that the wrong unit can be installed
Don't provide access doors with numerous small screws or attachments
Don't use air filters that must be replaced frequently; don't place these filters so that it is necessary to shut down power and disassemble the product to reach them
Don't omit the guide from a screwdriver adjustment, so that while the technician is adjusting and watching a meter, the screwdriver slips out of the slot
Don't design frequent failure or highly adverse impact failure items in least accessible locations
Don't unnecessarily subject nonwear components to failure stress caused by maintenance actions

The Design Team

During the entire period of design, reliability should be continuously monitored by the design team. The membership of the design team and the technical matters requiring their attention are shown in Table 11.6.4. A generic guiding list is given in Table 11.6.5.

TABLE 11.6.4 Design Team

Members	Technical Considerations
Project engineer	Products lists and application
Electrical, mechanical, chemical, manufacturing, and design representations	Tolerance studies
Management representatives	Environmental and operational effects
Maintenance and logistics representatives	Drift, aging, and end-of-life parameters
User community representatives	Regression or worst-case analysis
	Reliability analysis
	Trade-off studies
	Maintenance factors
	Test data
	Availability and affordability analysis

TABLE 11.6.5 Design Reliability Check List Item

Are the requirements for performance, application, environment, and maintainability and reliability established for each product, interface, and structure?
Are the best available methods for reducing the adverse effects of operational environments on critical structures being utilized?
Have normal modes of failure and the magnitude of each mode for each component or critical product have been identified as to root cause?
Has shelf life of products chosen for final design been determined?
Have limited-life products been identified and inspection and replacement requirements specified?
Have critical products that require special procurement, testing, and handling been identified?
Are effective safety factors being used in the application of products, interfaces, and structures?
Have studies been made considering variability and degradation of parameters of products and structures?
Have all vital adjustments been classified as to factory, preoperational, or operator types?
Have adjustments been minimized?
Have stability requirements of all products and structures associated with each adjustment been established?
Are similar plugs and connectors adequately protected against insertion in wrong socket?
Are malfunction-indicating products being used extensively?
Are self-monitoring or self-calibration products installed in major products where possible?
Are mechanical support structures adequate?
Is there a concentrated effort to make the developmental model as near to the production model as possible?
Have packaging and mechanical layout been designed to facilitate maintenance and to reduce maintenance costs and downtime?

Summary

When a product is being designed, it is assumed that the required investment will be justified according to how well the product performs its intended function over time. This assumption cannot be justified when a product fails to perform upon demand or fails to perform repeatedly. It is not enough to show that a product can conduct a function but that it can do so repeatedly when needed by the product user.

The design of any product involves trade-offs, including, but not limited to, performance capability, size, weight cost, product maintenance activities, and other factors, depending upon the intended use. In cases where human life is at risk, reliability issues play a major role in design. However, it is equally important to ask how large can a safety factor be for a critical situation such as a spacecraft, where human life might be in jeopardy, yet weight is a governing element.

The demand for better performance over time, coupled with higher-life cycle costs and stricter legal liabilities, has made product reliability of great importance. As the attitude toward production of engineering products has changed, reliability has established itself as one of the key elements when designing and manufacturing a product. However, the application of reliability principles in the product design and manufacturing processes has not always kept pace with the evolution of highly complex and dynamic engineering products. The design team must understand reliability theory and techniques, in conjunction

with design and manufacturing processes as well as the scenario for use of the product. The goal is to provide a scientific basis for design decisions considering product effectiveness requirements and the scenario for use, create a product design that will satisfy those requirements, and document those design decisions for both the hardware product and the recommended support product for sustaining product effectiveness during the life of the product.

11.7 Design of Manufacturing Systems and Processes

Leonard D. Albano

Design of Manufacturing Systems

Section 11.6 provides strategies for the design of products that meet the customer's needs, based on findings derived from market studies and product research and development. The success of such designs depends on the enterprise's ability to manufacture it, without compromising product performance or incurring noncompetitive production costs. Thus, the design of a manufacturing system must be done with a clear understanding of the functional requirements and physical design parameters for a product and an appreciation for the business goals and constraints.

The structure of a manufacturing system is hierarchical. At the highest level, there are two distinct subsystems: the *direct manufacturing operations* and the *indirect overhead functions*. The direct manufacturing operations consist of the employees and equipment needed to convert the input resource materials into the final desired products. The indirect overhead functions refer to the people, vendors, and capital that support and supervise the direct operations. At the lowest level, the manufacturing system is defined by many basic manufacturing operations, such as metal cutting and plastic forming, that are used to create specific part types (see Chapter 13).

Manufacturing systems can be classified in terms of the volume and variety of products produced. One possible classification scheme is as follows:

1. *Continuous flow processes*: high volume production of a single product type, such as sugar and oil refineries
2. *Assembly lines*: high volume production of a limited product variety; relies on standardized components and interchangeable parts; common system for automotive industry
3. *Batch processing*: relatively low volume production of multiple product types; concepts of group technology used to batch workpieces to increase efficiency of process routing; example applications include the manufacture of aircraft and heavy construction equipment
4. *Job shop*: low volume production of a variety of nonstandard parts; focuses on customized products; early reliance on craftsmen has shifted to cellular manufacturing systems in order to maximize flexibility and throughput

The design of a manufacturing system may be defined in terms of the concept of domains, as presented in Section 11.3, and the integration of three or more design fields; they are product design, organization design, and software design.

1. *Product design*: The design of the direct manufacturing operations is based on mapping the product's design parameters in the physical domain to process variables in the process domain.
2. *Organizational design*: The human and financial resources, which contribute to both direct operations and the indirect overhead, are provided to satisfy the goals and structure of the business organization.
3. *Software design*: Computer-based tools are easily available to control production and inventory levels, support information flow for decision-making, and automate manufacturing operations.

Several important advances in manufacturing can be identified from the above design model. *Concurrent engineering* refers to the integration of product design and manufacturing system design. This approach involves continuous mapping from the functional domain to the physical domain to the process domain. If the organizational structure changes to adapt to new customer needs, the single enterprise may be replaced by a highly flexible, virtual enterprise that pools resources from several qualified organizations for rapid response. This virtual enterprise corresponds to *agile manufacturing*. With a rationalized approach to system design, the role of the computers, information technology, robots, etc. can be better understood (see Chapter 13). Mitchell (1991) defines *computer-integrated manufacturing*

as "the use of computers to achieve an integrated flow of manufacturing activities, based on integrated information flow that links together all organizational activities."

A designer may develop many different manufacturing systems that are technically feasible. The designer must have some basis for the selection of a system and the layout of the various elements that comprise it. The evaluation of economic feasibility is one criterion for decision-making. Economic feasibility depends on the planned production volume, fixed resources, total production costs, projected revenue stream, and the desired income. The latter case involves consideration of the investment strategy. Because of the interrelations among the system performance and the economic factors, the suggested design procedure is to consider economic feasibility as a design constraint. The ideal to be achieved is one involving the design and production of reliable products that satisfy the customer's needs, while providing the desired profit and return on investment. See Chapter 13 for information on manufacturing system and enterprise management.

Manufacturing Process Design

The principal objective of manufacturing process design is to produce an organized plan for converting raw materials into useful products. It involves the selection of timely and cost-effective methods to produce a product without compromising quality and reliability. As part of the product development process, good manufacturing process design contributes to the industrial competitiveness of a manufacturing enterprise, while poor process design contributes to cost and schedule overruns and the delivery of products that fail to meet some or all of the customers' needs.

Manufacturing process planning often requires the consideration of several manufacturing processes for a specific part. In addition to considering manufacturing costs and production time, rational planning should also involve evaluation of how well a particular process satisfies the design requirements, which are delineated within the engineering drawings and reference specifications. The suitability of a given process may be based on many factors, such as

1. The dimensions and geometric precision that can be obtained
2. The surface roughness that can be attained
3. The changes that may be produced in material properties and part performance

Unfortunately, there is seldom much time to conduct an exhaustive laboratory or computer-based study and evaluation of all solution alternatives. A systematic approach to design that provides fundamental principles for decision-making would facilitate process design while enabling the designer to consider all important factors. Section 11.4, for instance, describes the concept of axiomatic design as a scientific framework for design and decision-making.

Metals Processing

Metal shapes and components can be produced by various casting, forming, and metal-removal processes (see Chapter 13). In the case of metal removal, process planning involves selecting and sequencing the appropriate machine tools and operations so as to convert a solid piece part from its initial shape to a final, desired geometry. This involves matching machine and tool data to the design requirements, subject to certain constraints imposed by the manufacturing organization and facilities. In practice, it consists of five or more steps:

1. Interpretation of the engineering drawing and reference specifications
2. Selection of machining operations to form the specified surfaces
3. Selection of machine tools for each machining operation
4. Selection of jigs and fixtures to guide or facilitate machining
5. Sequencing the machining operations

The initial material shape may be one of any number of geometries, ranging from simple bar stock to a complex casting. Basic machine tool operations are discussed in Chapter 13, and they include shaping, planing, turning, grinding, sawing, and drilling.

Process planning is performed by either experienced planners or automated software systems. There are at least two major difficulties associated with relying solely on human planners. First, they often rely on intuition gained through experience, and this experience requires a significant period of time to accumulate. Consequently, the number of truly skilled planners in any given industry is limited from generation to generation. Second, the capacity of a manufacturing enterprise to adopt new processes and new systems is limited by the knowledge background and creativity of the process planner. In order to address these problems, computer-aided process planning (CAPP) systems (see Chapter 13) have been developed.

The two fundamental strategies for CAPP are the *variant approach* and the *generative approach*. The variant approach is closely related to the automation of manual process planning techniques, i.e., the process plan for a specific part is created by computer-based retrieval of an existing process plan from a data base. Thus, the role of the computer in a variant system is primarily to manage a data base of process knowledge. Generative systems synthesize manufacturing and planning knowledge to generate specific process plans for a specific part. The reasoning mechanism for generative systems varies from algorithms to decision trees to artificial intelligence techniques. The variant approach is the basis for the overwhelming majority of computer codes for process planning because of a number of practical computational advantages. Nevertheless, several generative systems have been developed by industry, and this area will continue to be the focus of research and development, especially for large companies that can support the considerable investment needed for software development and hardware procurement.

Polymer Processing

There are basically two major categories of plastics: *thermoplastics* and *thermosetting resins* (see Chapter 12). Each group contains thousands of specific formulations with a wide range of mechanical, physical, and chemical properties. The scope of applications includes the automotive, houseware, and packaging industries. For example, government regulations pertaining to energy conservation have forced the automotive manufacturers to reduce the weight of automobiles. As a result, lightweight plastics have replaced traditional steel bumpers and outer panels while fulfilling specific performance requirements for mechanical strength, thermal resistance, and environmental durability.

To design with plastics, the product engineer selects an appropriate plastic material by considering a number of technical and economic factors in concert with specific data compiled in manufacturer's literature and encyclopedic data sources, such as the *Modern Plastics Encyclopedia*. To select the appropriate manufacturing process, the process engineer must understand the impact of the processing techniques on the material structure and, ultimately, the desired technical performance, e.g., mechanical strength and physical properties. Although a considerable number of plastic material alternatives are available, the variety of forming operations is limited. The basic set of processes includes compression molding, transfer molding, injection molding, extrusion, cold molding, thermoforming, and blow molding. The reader is referred to Chapter 13 for specific information on the production of thermoplastics and thermosetting resins.

Because of the interrelations between processing techniques and material structure, the axiomatic approach to design (see Section 11.4) provides helpful information for concurrent engineering. In this context, the use of plastic parts involves the successive mapping of engineering properties in the functional domain, to a plastic material structure in the physical domain, to processing techniques in the process domain. This mapping can be written in the form:

$$\{\mathbf{FRs}\} = [\mathbf{A}]\{\mathbf{DPs}\} \tag{11.7.1}$$

$$\{\mathbf{DPs}\} = [\mathbf{B}]\{\mathbf{PVs}\} \tag{11.7.2}$$

in which the desired material properties are contained in the vector of functional requirements {**FRs**}, the material structure is described in {**DPs**}, and the manufacturing process is defined by {**PVs**}. Equations (11.7.1) and (11.7.2) display the general structure of the equations for materials design and process planning. The design matrices [**A**] and [**B**] are used to determine whether the proposed mapping satisfies the Independence Axiom: they must be either diagonal or triangular to allow for the manufacture of a plastic part with the desired properties (see Section 11.4.1, "The First Axiom: The Independence Axiom"). The production of microcellular plastics of MIT (Suh, 1990) is an example that illustrates design of a manufacturing process to yield the desired polymer structure.

References

Mitchell, F.H. 1991. *CIM Systems: An Introduction to Computer-Integrated Manufacturing.* Prentice-Hall, Englewood Cliffs, NJ.

Modern Plastics Enclycopedia. McGraw-Hill, New York.

Suh, N.P. 1990. *The Principles of Design.* Oxford University Press, New York.

11.8 Precision Machine Design

Alexander Slocum

Any machine, from a machine tool to a photocopier to a camera, is an assembly of components that are designed to work together to achieve a desired level of performance. Each machine has a budget for cost and performance, and achieving the best balance between the two, regardless of the function of the machine, is the essence of *precision machine design.*

In order to be able to effectively develop a design for a precision machine, the design engineer must simultaneously envision in his/her head the functions the machine must perform (e.g., milling, turning, or grinding) alongside a pictorial library of component technologies (e.g., bearings, actuators, and sensors), generic machine configurations (e.g., cast or welded articulated and/or prismatic structures), analysis techniques (e.g., back-of-the-envelope and finite element methods), and manufacturing methods (e.g., machine, hand, or replication finished). In addition, the machine design engineer must be aware of the basic issues faced by the sensor and electronics engineer, the manufacturing engineer, the analyst, and the controls engineer. Only by a simultaneous consideration of all design factors can a viable and effective design be rapidly converged upon. Awareness of current technological limitations *in all fields* can also help a design engineer to develop new processes, machines, and/or components.

The goal of the precision machine design engineer is to make all the components of a precision machine in proper proportion of each other, both in relation to their physical size and the capabilities of the servocontroller and power systems. If a component is oversized, it may increase the cost of the machine while performance may not be increased. If a component is too small, the rest of the machine's components may never reach their potential and machine performance will suffer. Note that size is a function of static and dynamic qualifiers. Components that behave well statically do not necessarily have good dynamic performance.

In today's world where rapid time to market is essential, the design of quality precision machines depends on the ability of the design and the manufacturing engineers to predict how the machine will perform before it is built. Kinematics of a machine are easily tested for gross functionality using mechanism synthesis and analysis software. Wear rates, fatigue, and corrosion are often difficult to predict and control, but for the most part are understood problems in the context of machine tool design. Hence perhaps the most important factors affecting the quality of a machine are the accuracy, repeatability, and resolution of its components and the manner in which they are combined. These factors are critical because they affect every one of the parts that will be manufactured using the machine. Accordingly, minimizing machine cost and maximizing machine quality mandate predictability of accuracy, repeatability, and resolution. As noted by Donaldson*: "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."

In this section, the design of precision machines will be studied by following a path of analysis of the overall system's potential for accuracy, as well as applying this perspective to the understanding of the operation of the components that make up a precision machine:

- Analysis of errors in a precision machine
- Structures
- Bearings

It is important to realize that the design of precision machines is like any other endeavor. The better one understands the fundamental principles and concepts, the better one will be able to integrate components into a precision system. In the light of understanding the science of design, one must always consider practical applied philosophy. A few broad design guidelines for the precision machine designer include:

* Donaldson, R. November 1972. The deterministic approach to machining accuracy. *SME Fabricat. Technol. Symp.* Golden, CO (UCRL Preprint 74243).

- Subject all decisions to an "is there a better way" value analysis based on system considerations.
- Always picture in your mind how the system will be manufactured, assembled, used, and maintained.
- Minimize the number of parts in an assembly and minimize their complexity.
- Maximize the number of instances where reference surfaces and self-locating "snap together" parts can be used.
- Whenever possible, take advantage of kinematic design principles.
- Utilize new materials and technologies to their fullest potential.
- Read continuously and familiarize yourself with technologies in many areas.
- Observe continuously and familiarize yourself with products in many areas.

Analysis of Errors in a Precision Machine

The first step in the design of a precision machine is to understand the basic definitions and principles that characterize and govern the design of all precision machines. Some of the more basic definitions and principles include

- The principle of reversal
- Modeling the errors in a machine
- Determination of the relative errors between the tool and the workpiece
- Linear motion system errors
- Estimating position errors from modular bearing catalog data
- Axis of rotation errors
- Error budgets

Accuracy, Repeatability, and Resolution

There are three basic definitions to remember with respect to how well a machine tool can position its axes: *accuracy, repeatability (precision),* and *resolution. Accuracy* is the maximum translational or rotational error between any two points in the machine's work volume. *Repeatability (precision)* is the error between a number of successive attempts to move the machine to the same position, or the ability of the machine to make the same motion over and over. Bidirectional repeatability is the repeatability achieved when the point is approached from two different directions. This includes the effect of backlash in a leadscrew. Accuracy is often defined in terms of the mean, and repeatability in terms of the standard deviation. For a set of N data points with a normal (Gaussian) distribution, the mean x_{mean} and the standard deviation σ are defined as

$$x_{mean} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - x_{mean}\right)^2} \tag{11.8.1}$$

The standard deviation is used in the determination of the probability of occurrence of an event in a system that has a normal distribution. Table 11.8.1 gives the percent chance of a value occurring within a number of standard deviations of its expected value.* One must be very careful not to confuse an offset of the mean with the allowed random variation in a part dimension. For example, precision-molded plastic lenses for cameras have their molds hand finished to make the mean size of lenses produced equal to the nominal size required.

* The equation for the computation of the percent chance was obtained from Drake, A. 1967. *Fundamentals of Applied Probability Theory.* McGraw-Hill, New York. p. 211. Values ϕ (k) were computed using Mathematics™. Another valuable reference to have is Natrella, M. *Experimental Statistics,* NBS Handbook 91.

TABLE 11.8.1 Chance of a Value Falling within kσ of its Expected Value

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

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

Resolution is the larger of the smallest programmable step or the smallest mechanical step the machine can make during point-to-point motion. Resolution is important because it gives a lower bound on the repeatability. When a machine's repeatable error is mapped, the resolution becomes important if the mapped errors are to be compensated for by other axes.

Amplification of Angular Errors

Perhaps the most overlooked error in machine design is the amplification of an angular error over a distance to create a large translational error at some point in the machine. Mathematically, this error has a magnitude equal to the product of the distance between the point and the rotation source and the sine of the included angle. Hence this type of error can be referred to as a *sine error.* This principle extends to locating bearing surfaces far from the workpiece area of the machine tool.*** Errors in the bearing's motion can be amplified by the distance between the bearing and the workpiece, and can be transmitted to the workpiece. This can result in horizontal and vertical straightness errors and axial position errors. The same is true for the effects of all other types of errors on machine components. A *cosine error*, on the other hand, is the error made when the measurement of the distance between a point and a line is not made along a path orthogonal to the line.

With respect to dimensional measurements. In the late 1800s, Dr. Ernst Abbe noted *"If errors in parallax are to be avoided, the measuring system must be placed coaxially with the axis along which displacement is to be measured on the workpiece."* The Abbe principle can be visualized by comparing measurements made with a dial caliper and a micrometer as shown in Figure 11.8.1. The dial caliper is often used around the shop because it is easy to use, the head slides back and forth to facilitate quick measurement. However, note that the measurement scale is located at the base of the jaws. When a part is measured near the tip of the jaws, the jaws can rock back slightly, owing to their elasticity and imperfections in the sliding jaw's bearing. This causes the caliper to yield a slightly undersize

* Bryan, J. The Power of Deterministic Thinking in Machine Tool Accuracy, 1st Int. Mach. Tool Eng. Conf. November 1984. Tokyo (UCRL Preprint 91531).

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

*** See Bryan, J.B. 1989. The Abbe principle revisited — an updated interpretation, *Precis. Eng.* 1(3):129–132. An extension of the Abbe principle to this type of situation is referred to as the *Bryan principle*, which is discussed in Section 5.2.1 along with a discussion on where to mount sensors for various types of measurements.

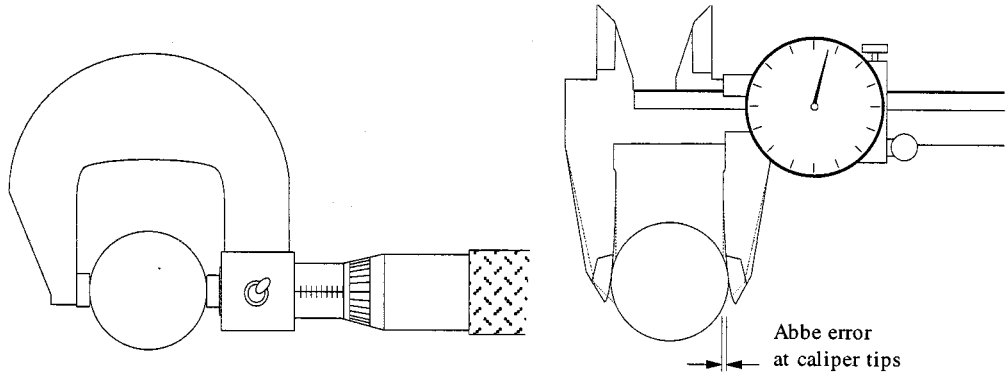


FIGURE 11.8.1 Abbe error illustrated through the use of a dial caliper and a micrometer.

measurement. The micrometer, on the other hand, uses a precision measuring device located in line with the part dimension, so there is no Abbe error. Both instruments are sensitive to how hard the jaws are closed on the part. A micrometer often has a torque limiting adjustment to provide a very repeatable measuring force. It is impossible to overstress the importance of Abbe errors.

The micrometer is more difficult to use and less versatile than the caliper, yet it has greater accuracy because of its more robust structural loop. So it is in the design of precision machines. Often one has to sacrifice ergonomics for performance. The goal, therefore, is to minimize the extent of the sacrifice.

Sensitive Directions

It is important to note that there are *sensitive directions* in a machine, and it is along these directions that the most effort must be expended to minimize errors. For example, as illustrated in Figure 11.8.2 an error motion ε of a tool tangent to the surface of a round part of radius r in a lathe results in a radial error in the workpiece of magnitude ε^2/r, which is much smaller than ε. Sensitive directions can be *fixed*, such as when the tool is stationary and the part is moving (e.g., a lathe), or *rotating*, such as when the tool is rotating and the part is fixed (e.g., a jig borer).

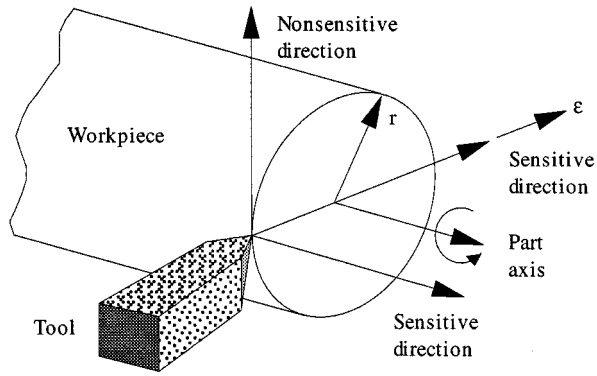


FIGURE 11.8.2 Illustration of sensitive directions.

The Reversal Principle

How were the first accurate machines developed? Machines were first made repeatable, which can be done by paying special detail to surface finish of bearings, prevention of lost motion (e.g., backlash), and by making the forces on the machine repeatable. Once a machine is made repeatable, it can be used to measure the accuracy of another component or machine using the principle of reversal. This is illustrated in Figure 11.8.3.

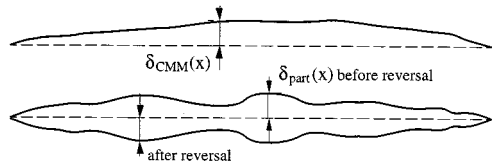


FIGURE 11.8.3 Illustration of the principal of reversal.

The machine is used to measure a part from one side, obtaining a reading of

$$Z_{\text{probe before reversal}}(x) = \delta_{CMM}(x) - \delta_{\text{part}}(x)$$

The part is then turned over, taking care to keep its axial position justified, then the machine is used to remeasure the part, obtaining a reading of

$$Z_{\text{probe after reversal}}(x) = \delta_{CMM}(x) - \delta_{\text{part}}(x)$$

By subtracting the measurements from each other, the repeatable error in the measuring machine is removed:

$$\delta_{\text{part}}(x) = \frac{-Z_{\text{probe before reversal}}(x) + Z_{\text{probe after reversal}}(x)}{2}$$

There are many variations (e.g., for roundness measurements), and the principle shows how repeatability is often the most important characteristic of a precision machine. In fact, the principle can be applied to the design of the machine itself. For example, two bearing rails can be ground side by side, and then placed end to end on a machine. If the carriage that rides on them has the same bearing spacing as the individual rail length, then as the carriage moves on the end-to-end rails, it will not pitch or yaw, it will only have a straightness error. This straightness error can then be more readily compensated for by the motion of an orthogonal axis via an error compensation algorithm in the controller.

Modeling the Errors in a Machine

In order to determine the effects of a component's error on the position of the toolpoint or the workpiece, the spatial relationship between the two must be defined. Figure 11.8.4 illustrates the issue: the tool is connected to the workpiece through the linkage that is the machine. Any errors between the machine's links create error between the tool and the workpiece. The essential goal of precision machine design is to be able to model and predict these errors, so they can be addressed during the design state of the machine.

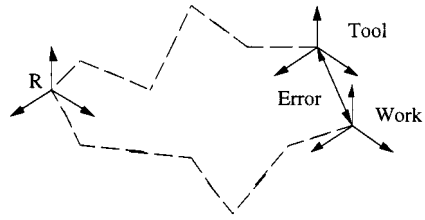


FIGURE 11.8.4 The goal is to be able to model the errors in the machine to predict the error between the tool and the work during the design phase so appropriate action can be taken before the machine is built.

To represent the relative position of a rigid body in three-dimensional space with respect to a given coordinate system, a 4×4 matrix is needed. This matrix represents the coordinate transformation to the reference coordinate system (X_R Y_R Z_R) from that of the rigid body frame (X_n Y_n Z_n), and it is called the *homogeneous transformation matrix* (HTM).* The first three columns of the HTM are direction cosines (unit vectors i, j, k) representing the orientation of the rigid body's X_n, Y_n, and Z_n axes with respect to the reference coordinate frame, and their scale factors are zero. The last column represents the position of the rigid body's coordinate system's origin with respect to the reference coordinate frame. P_s is a scale factor, which is usually set to unity to help avoid confusion. The presuperscript represents the reference frame you want the result to be represented in, and the postsubscript represents the reference frame from which you are transferring:

$$R_{\mathbf{T}_n} = \begin{bmatrix} O_{ix} & O_{iy} & O_{iz} & P_x \\ O_{jx} & O_{jy} & O_{jz} & P_y \\ O_{kx} & O_{ky} & O_{kz} & P_z \\ 0 & 0 & 0 & P_s \end{bmatrix} \tag{11.8.2}$$

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

$$\begin{bmatrix} X_R \\ Y_R \\ Z_R \\ 1 \end{bmatrix} = R_{\mathbf{T}_n} \begin{bmatrix} X_n \\ Y_n \\ Z_n \\ 1 \end{bmatrix} \tag{11.8.3}$$

For example, if the X_1 Y_1 Z_1 coordinate system is translated by an amount x along the X axis, the HTM that transforms the coordinates of a point in the X_1 Y_1 Z_1 coordinate frame into the XYZ reference frame is

$$XYZ_{\mathbf{T}_{X_1Y_1Z_1}} = \begin{bmatrix} 1 & 0 & 0 & x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11.8.4}$$

If the X_1 Y_1 Z_1 coordinate system is rotated by an amount θ_x about the X axis, the HTM that transforms the coordinates of a point in the X_1 Y_1 Z_1 coordinate frame into the XYZ frame is

$$XYZ_{\mathbf{T}_{X_1Y_1Z_1}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x & 0 \\ 0 & \sin\theta_x & \cos\theta_x & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11.8.5}$$

Machine structures can be decomposed into a series of coordinate transformation matrices describing the relative position of each axis and any intermediate coordinate frames that may assist in the modeling

* The HTM representation of structures has existed for many decades. See, for example, Denavit, J. and Hartenberg, R. June 1955. A kinematic notation for lower-pair mechanisms based on matrices. *J. Appl. Mech.* Perhaps the most often referenced work with respect to its application to manufacturing tools is Paul, R. 1981. *Robot Manipulators: Mathematics, Programming, and Control.* MIT Press, Cambridge, MA.

process, starting at the tip and working all the way down to the base reference coordinate system (n = 0). If N rigid bodies are connected in series and the relative HTMs between connecting axes are known, the position of the tip (Nth axis) in terms of the reference coordinate system will be the sequential product of all the HTMs:

$$ {}^{R}\mathbf{T}_{N} = \prod_{m=1}^{N} {}^{m-1}\mathbf{T}_{m} = {}^{0}\mathbf{T}_{1}\,{}^{1}\mathbf{T}_{2}\,{}^{2}\mathbf{T}_{3}\ldots \tag{11.8.6} $$

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

Determination of the Relative Errors Between the Tool and the Workpiece

Ideally, the HTM products (Equation 11.8.6) for the position of the point on the workpiece the tool contacts and the location of the toolpoint with respect to the reference frame will be identical. The relative error HTM $\mathbf{E}_{\text{rel}}$, representing position and orientation errors between the tool and workpiece, is determined for ${}^{R}\mathbf{T}_{\text{work}} = {}^{R}\mathbf{T}_{\text{tool}}\mathbf{E}_{\text{rel}}$:

$$ E_{\text{rel}} = {}^{R}\mathbf{T}_{\text{tool}}^{\;-1}\,{}^{R}\mathbf{T}_{\text{work}} \tag{11.8.7} $$

The relative error HTM is the transformation in the toolpoint coordinate system that must be done to the tool in order to be at the proper position on the workpiece. For implementation of error correction algorithms on numerically controlled machines, as illustrated in Figure 11.8.5 one must consider how the axes will be required to move in order to create the desired motion, specified by Equation 11.8.7 in the tool reference frame. For the general case of a machine with revolute and translational axes (e.g., a five-axis machining center), one would have to use inverse kinematic solutions such as those developed for robot motion path planning. Most machine tools and CMMs have only translational axes, and thus the error correction vector ${}^{R}\mathbf{P}_{\text{correction}}$ with respect to the reference coordinate frame can be obtained from

$$ {}^{R}\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix}_{\text{correction}} = {}^{R}\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix}_{\text{work}} - {}^{R}\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix}_{\text{tool}} \tag{11.8.8} $$

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

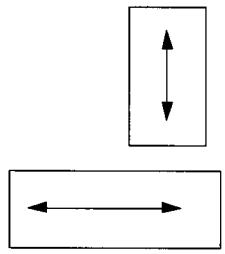


FIGURE 11.8.5 Repeatable error motions in one axis can be mapped and compensated for by motion of an orthogonal axis.

Compensating for errors in a machine can be a good thing; however, electronic compensation can only do so much. One has to start with a good design. The error in a machine can be minimized in general by maximizing the efficiency of the machine's *structural loop*. The structural loop is defined as the structure that joins the tool to the fixture to which the workpiece is attached. During cutting operations, the contact between the tool and workpiece can change the structural loop's characteristics. Maximizing the efficiency of the structural loop generally requires minimizing the path length of the mechanism. As can be seen from Abbe's principle or the HTM analysis method, the shorter the path length, the less the error amplification and the total end point error.

Linear Motion System Errors

Consider the case of an ideal linear motion carriage shown in Figure 11.8.6 with x, y, and z offsets of a, b, and c, respectively, All rigid bodies have three rotational (ε_X, ε_Y, ε_Z) and three translational (δ_X, δ_Y, δ_Z) error components associated with their motion. These errors can be defined as occurring about and along the reference coordinate system's axes, respectively. Often the errors will be a function of the position of the body in the reference frame.

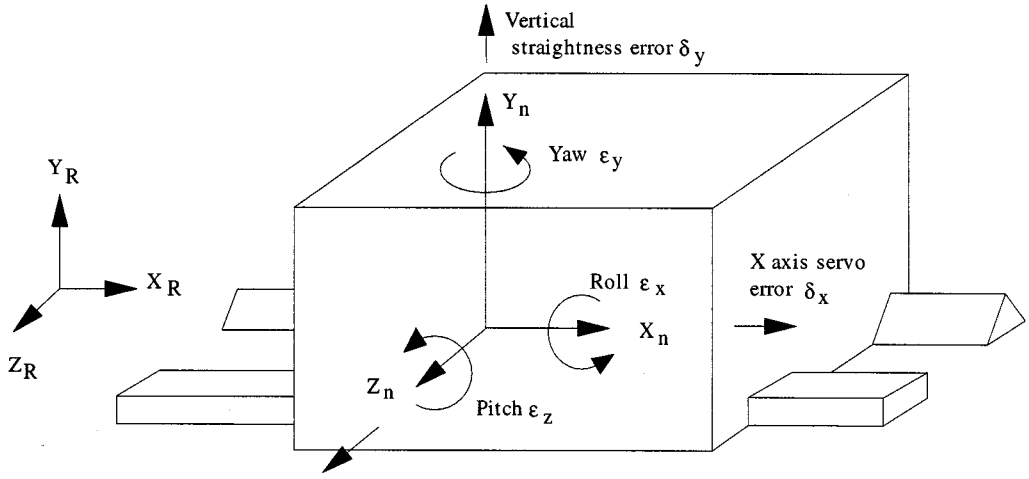


FIGURE 11.8.6 Motion and errors in a single-axis linear motion carriage (prismatic joint).

For the linear carriage, the HTM (see Equation (11.8.2)) that describes the effects of errors on carriage motion can be obtained in the following manner. Observing the right-hand rule, a rotation ε_X, about the X axis (*roll*) causes the tip of the Y-axis vector to move in the positive Z direction by an amount proportional to sin ε_X, and in the negative Y direction by an amount proportional to $1 - \cos \varepsilon_X$. Since the error terms are very small, at most on the order of minutes of arc, small-angle approximations are valid and will be used. Hence the element $o_{ky} = \varepsilon_X$. The roll error ε_X also causes the tip of the Z-axis vector to move in the negative Y direction, so that the element $o_{jz} = -\varepsilon_X$. A rotation ε_Y about the Y axis (*yaw*) causes the tip of the X-axis vector to move in the negative Z direction, so that $o_{kx} = -\varepsilon_Y$, and causes the tip of the Z-axis vector to move in the positive X direction, so $o_{iz} = \varepsilon_Y$. Similarly, a rotation ε_Z about the Z axis (*pitch*) causes the tip of the X-axis vector to move in the positive Y direction, so $o_{jx} = \varepsilon_Z$, and causes the tip of the Y axis vector to move in the negative X direction so that $o_{iy} = -\varepsilon_Z$. Translational errors δ_X, δ_Y, and δ_Z directly affect their respective axes, but care must be taken in defining them. Having neglected second-order terms, the HTM for the linear motion carriage with errors is

$$ {}^{R}\mathbf{T}_{\text{nerr}} = \begin{bmatrix} 1 & -\varepsilon_Z & \varepsilon_Y & a+\delta_X \\ \varepsilon_Z & 1 & -\varepsilon_X & b+\delta_Y \\ -\varepsilon_Y & \varepsilon_X & 1 & c+\delta_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11.8.9} $$

Estimating Position Errors from Modular Bearing Catalog Data

The HTM method is powerful, but from where does one get estimates of the errors? For the linear motion system, the HTM assumes that the errors occur at the center of stiffness of the carriage. The *center of stiffness* is the point at which, when a force is applied to the system, no net angular motion results. For a symmetrical system, it is located at the center of the system. For other systems, the center of stiffness can be found in the same way that the *center of mass* is found:

$$ x_{\text{center of stiffness}} = \frac{\sum_{i=1}^{N} K_i x_i}{\sum_{i=1}^{N} K_i} \tag{11.8.10} $$

Indeed, the *center of friction* can also be defined, and when locating the point, where the actuation force of an axis is to be applied, one can see that ideally, the center of mass, the center of stiffness, and the center of friction should all be coincident. In the best case, they will also be concurrent with external loads applied to an axis.

Once the center of stiffness has been found, the error motions about the center of stiffness can be determined using the dimensions of the system and estimates for the errors in individual bearing elements (e.g., linear ball bearing catalogs often include data on the vertical and horizontal error motions that will occur between the bearing block and the rail). Figure 11.8.7 illustrates the model of a linear motion system supported by modular bearings.

It can be assumed that the translational errors are based on the average of the straightness errors experienced by the bearing blocks:

$$ \begin{aligned} \delta_x &= \delta_{\text{servo}} \\ \delta_y &= \frac{\delta_{y1}+\delta_{y2}+\delta_{y3}+\delta_{y4}}{4} \\ \delta_z &= \frac{\delta_{z1}+\delta_{z2}+\delta_{z3}+\delta_{z4}}{4} \end{aligned} \tag{11.8.11} $$

The angular errors are based on the differences in the average straightness errors experienced by pairs of bearing blocks acting across the carriage:

$$ \begin{aligned} \varepsilon_x &= \frac{\frac{\left(\delta_{y2}+\delta_{y3}\right)}{2} - \frac{\left(\delta_{y1}+\delta_{y4}\right)}{2}}{W} \\ \varepsilon_y &= \frac{\frac{\left(\delta_{z3}+\delta_{z4}\right)}{2} - \frac{\left(\delta_{z1}+\delta_{z2}\right)}{2}}{L} \\ \varepsilon_z &= \frac{\frac{\left(\delta_{y1}+\delta_{y2}\right)}{2} - \frac{\left(\delta_{y3}+\delta_{y4}\right)}{2}}{L} \end{aligned} \tag{11.8.12} $$

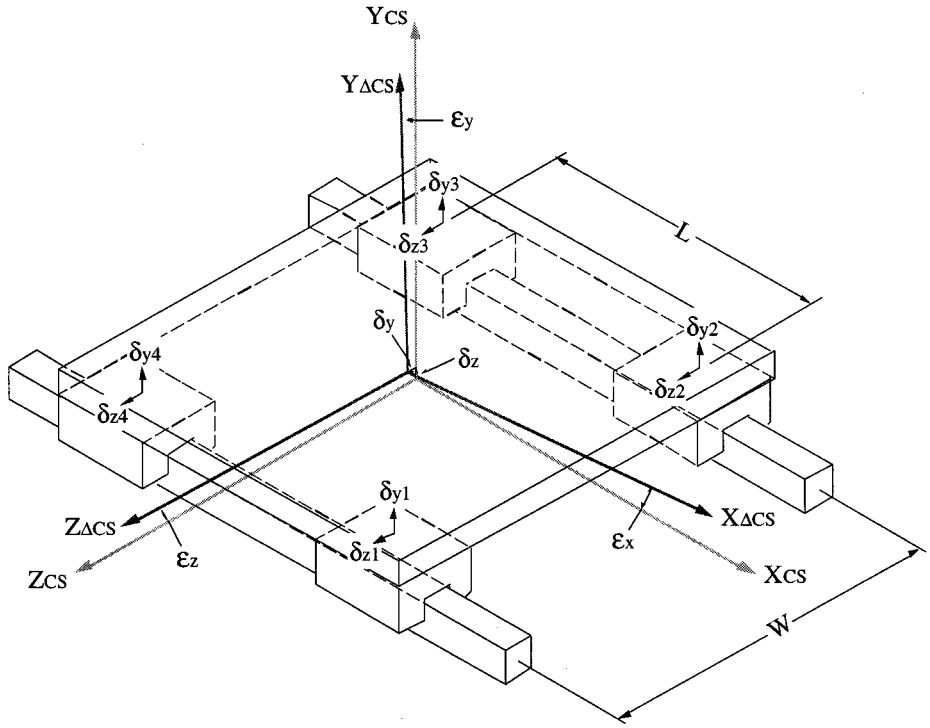


FIGURE 11.8.7 A linear motion system will have errors that act about the center of stiffness of the system.

When the bearing blocks are moving in unison to create the straightness error, they would not be creating the angular errors. However, to be conservative when modeling the errors in a machine, since it cannot be known which errors are going to occur, one should simultaneously incorporate both translational and angular error estimates into the error model of the machine.

Once a machine has been built, the error motions would be measured and analyzed to verify whether the machine meets its performance criteria. This is also a good time to evaluate the effectiveness of the model. For example, Figure 11.8.8 shows the straightness measurements for a linear motion axis supported by cam followers running on a vee and a flat. It is very difficult to tell what the source of the error is. Figure 11.8.9 shows a Fast Fourier Transform of the straightness data. Note that the FFT is plotted in terms of amplitude and wavelength (as opposed to the more commonly seen power and frequency plots used by electrical engineers). It can help identify the dominant sources of error, so design attention can be properly allocated. It can also show, as in this case, that the errors from a number of different sources are of similar magnitude, so if greater performance is required, one should probably seek an alternate design.

Axis of Rotation Errors*

Consider the rotating body shown in Figure 11.8.10. Ideally, the body rotates about its axis of rotation without any errors; however, in reality the axis of rotation revolves around an axis of the reference coordinate frame with radial errors δ_X and δ_Y, an axial error δ_Z, and tilt errors ε_X and ε_Y. All of these errors may be a function of the rotation angle θ_z. For a point in the spindle coordinate frame $X_nY_nZ_n$,

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

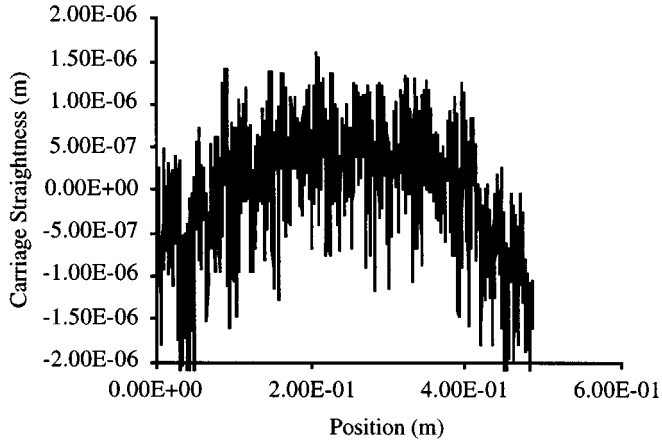


FIGURE 11.8.8 Strightness of a linear motion system.

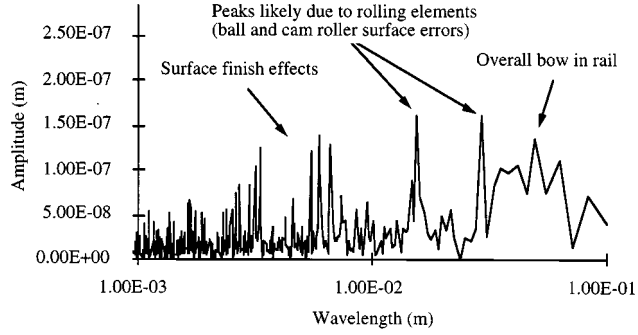


FIGURE 11.8.9 Fourier transform of the straightness of a linear motion system.

one would first use the rotation angle θ_z to transform the point roughly the reference frame. Then, since the other error motions are small, the order of multiplication of their HTMs would not be critical.

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

$$ {}^{R}\mathbf{T}_{\text{nerr}} = \begin{bmatrix} C\varepsilon_Y C\theta_Z & -C\varepsilon_Y S\theta_Z & S\varepsilon_Y & \delta_X \\ S\varepsilon_X S\varepsilon_Y C\theta_Z & C\varepsilon_X C\theta_Z - S\varepsilon_X S\varepsilon_Y S\theta_Z & -S\varepsilon_X C\varepsilon_Y & \delta_Y \\ -C\varepsilon_X S\varepsilon_Y C\theta_Z + S\varepsilon_X S\theta_Z & S\varepsilon_X C\theta_Z + C\varepsilon_X S\varepsilon_Y S\theta_Z & C\varepsilon_X C\varepsilon_Y & \delta_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11.8.13} $$

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

$$ {}^{R}\mathbf{T}_{\text{nerr}} = \begin{bmatrix} \cos\theta_Z & -\sin\theta_Z & \varepsilon_Y & \delta_X \\ \sin\theta_Z & \cos\theta_Z & -\varepsilon_X & \delta_Y \\ \varepsilon_X \sin\theta_Z - \varepsilon_Y \cos\theta_Z & \varepsilon_X \cos\theta_Z + \varepsilon_Y \sin\theta_Z & 1 & \delta_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11.8.14} $$

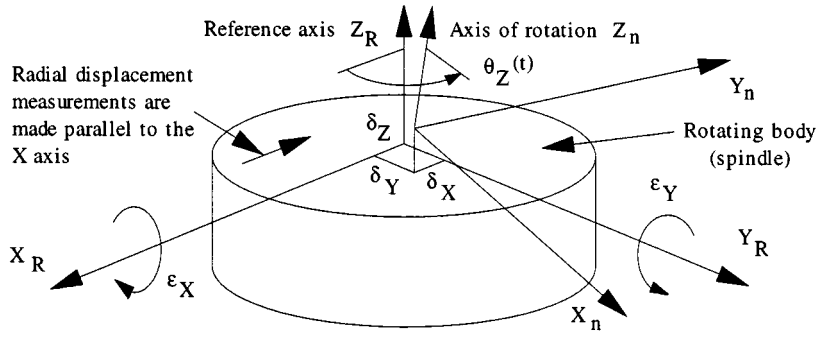


FIGURE 11.8.10 Motion and errors about an axis of rotation (revolute joint).

However, since this matrix is usually evaluated using a spreadsheet, it is best to use the exact form. When nanometer performance levels are sought, second-order effects can start to become important.

In the context of evaluating errors of rotating bodies, particularly when making measurements and discussing the results, it is necessary to consult the axis of rotation standard ANSI B89.3.4M. The "*error motion* of a spindle is the change in position, relative to the reference coordinate axes, of the surface of a perfect workpiece with its center line coincident with the axis of rotation." This definition excludes thermal drift errors. The *runout* is the total displacement measured by an instrument sensing against a moving surface or moved with respect to a fixed surface. The term total indicator reading (TIR) is equivalent to runout. Unfortunately, all too often an imperfect workpiece is eccentrically mounted to a spindle and used to evaluate the performance of a spindle; hence the runout can be a misleading measurement. The "*axial motion* is the error motion colinear with the Z reference axis." "Axial slip", "end camming", and "drunkenness" are nonpreferred terms which have been used in the past. The *tilt motion* is the error motion in an angular direction relative to the Z reference axis. Tilt motion creates sine errors on the spindle which is why the radial error motion is a function of Z position and face motion is a function of radius. Note that "coning", "wobble", and "swash" are sometimes used to describe tilt motion, but they are nonpreferred terms.

These errors are measured and then plotted using polar plots. The *error motion polar plot* is a polar plot of error motion made in synchronization with the rotation of the spindle. Error motion polar plots are often decomposed into plots of various error components. Some of the various types of error motion polar plots are shown in Figure 11.8.11. The *total error motion polar plot* is the complete error motion polar plot as recorded. The *average error motion polar plot* is the mean contour of the total error motion polar plot averaged over the number of revolutions. Note that asynchronous error motion components do not always average out to zero, so the average error motion polar plot may still contain asynchronous components. The average error motion value is a measure of the best roundness that can be obtained for a part machined while being held in the spindle (or the roundness of a hole the spindle is used to bore). The *fundamental error motion polar plot* is the best-fit reference circle fitted to the average error motion polar plots. The *residual error motion polar plot* is the deviation of the average error motion polar plot from the fundamental error motion polar plot. For radial error motion measurements, this represents the sum of the error motion and the workpiece (e.g., ball) out-of-roundness. The workpiece out-of-roundness can be removed using a reversal technique.* The *asynchronous error motion polar plot* is the deviation of the total error motion polar plot from the average error motion polar plot. Asynchronous in this context means that the deviations are not repetitive from revolution to revolution. Asynchronous

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

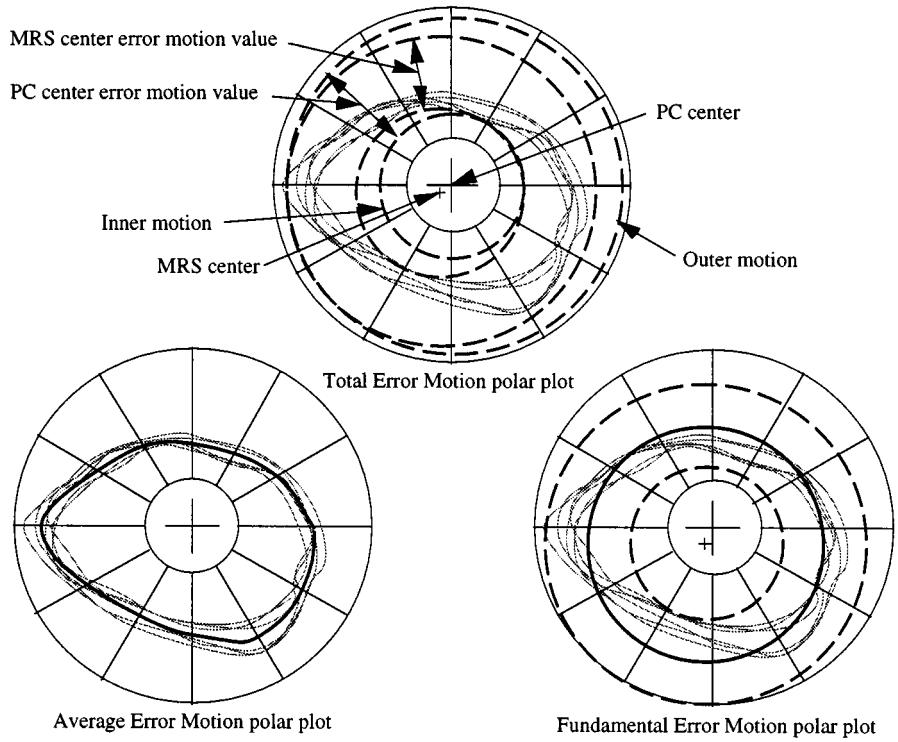


FIGURE 11.8.11 Examples of error motion and error motion component polar plots.

error motions are not necessarily random (in the statistical sense), but they do provide an indication of the attainable surface finish of the part.

Once again, the FFT is a vital tool for identifying the source of errors so that the designer can seek to minimize them. Figure 11.8.12 shows an FFT of a grinding spindle's error motion. The spindle speed was 1680 rpm (28 Hz), the bearing inner diameter was 75 mm, the outer diameter was 105 mm, the number of balls was 20, and the ball diameter was 10 mm. One of the dominant errors occurs at twice the rotation frequency, indicating that the bore in which the bearing is placed is most likely out-of-round.

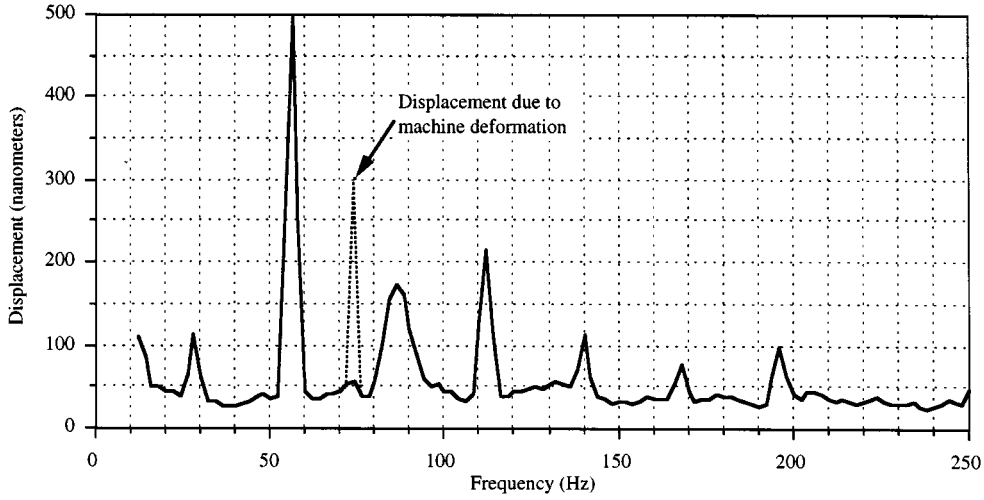


FIGURE 11.8.12 FFT of grinding spindle error motion.

Error Budgets

To define the relative position of one rigid body with respect to another, six degrees of freedom must be specified. To further complicate matters, error in each of the six degrees of freedom can have numerous contributing components. In fact, considering all the interacting elements in a typical machine tool, the number of errors that must be kept track of can be mindboggling. Therefore, the best way to keep track of and allocate allowable values for these errors is to use an *error budget*. An error budget, like any other budget, allocates resources (allowable amounts of error) among a machine's different components.* The goal is to allocate errors such that the ability of any particular component to meet its error allocation is not exceeded.

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

Perhaps the most important step in assembling the error budget for a machine is the placement of the coordinate frames and the assignment of linear and angular errors corresponding to the axes. Angular motion errors suffer no ambiguity in their definition since they are unaffected by other errors and therefore can be defined with respect to any set of axes. Linear motion errors, on the other hand, must be carefully defined in terms of linear motion caused directly and linear motion that is the result of an Abbe error. During the design phase, the coordinate systems must be located at the origin of the angular errors, where pitch, yaw, and roll errors are the center of stiffness as defined above.

Combinational Rules for Errors

Different types of errors do not mix unless handled properly. Once all error components are multiplied by their respective error gains, a final combination of errors can then be made to yield an educated guess as to the machine's expected performance. There are three common types of errors, which are defined as:**

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

Systematic and hysteresis errors can often be compensated for to a certain degree using calibration techniques. Random error cannot be compensated for without real-time measurement and feedback into a correcting servoloop. Thus, when evaluating the error budget for a machine, three distinct *subbudgets*

---

* See Donaldson, R. Error budgets. In *Technology of Machine Tools,* Vol. 5. *Machine Tool Accuracy,* R.J. Hocken, (ed.), Machine Tool Task Force.

** These definitions are from the CIRP Scientific Committee for Metrology and Interchangeability, 1978. A proposal for defining and specifying the dimensional uncertainty of multiaxis measuring machines. *Ann. CIRP.* 27(2):623–630.

based on systematic, hysteresis, and random errors should be kept. Often inputs for the error budget are obtained from manufacturers' catalogs (e.g., straightness of linear bearings), and they represent the peak-to-valley amplitude errors e_{PV}. A peak-to-valley error's equivalent random error with uniform probability of occurrence is given by $\varepsilon_{equiv.\ random} = \varepsilon_{PV}/K_{PV}$. If the error is Gaussian, then $K_{PVrms} = 4$ and there is a 99.9937% probability that the peak-to-valley error will not exceed four times the equivalent random error.

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

$$\varepsilon_{irms} = \left(\frac{1}{N} \sum_{i=1}^{N} \varepsilon_{irandom}^2 \right)^{1/2} \tag{11.8.15}$$

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

$$\varepsilon_{iworst\ case} = \sum \varepsilon_{isystematic} + \sum \varepsilon_{ihysteresis} + 4 \sum \varepsilon_{irandom} \tag{11.8.16}$$

The best-case error for the machine will probably be

$$\varepsilon_{ibest\ case} = \sum \varepsilon_{isystematic} + \sum \varepsilon_{ihysteresis} + 4 \left(\sum \varepsilon_{irandom}^2 \right)^{1/2} \tag{11.8.17}$$

In practice, the average of these two values is often used as an estimate of the accuracy the design is likely to achieve.

Type of Errors

There are many types of errors that occur in a machine including:

Straightness. *Straightness* is the deviation from true straight-line motion. One generally thinks of the straightness error as primarily dependent on the overall geometry of the machine and applied loads. As shown in Figure 11.8.13 straightness error can be considered the deviation from a straight line motion along a linear axis. The unofficial term *smoothness* can be used to describe straightness errors that are dependent on the surface finish of the parts in contact, the type of bearing used, and the bearing preload. In other words, the smoothness of motion would be the deviation from the best-fit polynomial describing the straightness of motion. Smoothness is not intended to be a descriptor of surface finish, but rather a descriptor of high-frequency straightness errors whose wavelength is typically on the order of the magnitude of the error, normal to the surface of two moving bodies in contact.

Kinematic Errors. Kinematic errors are defined here as *errors in the trajectory of an axis that are caused by misaligned or improperly sized components.* For example, kinematic errors include orthogonality (squareness or perpendicularity) and parallelism of axes with respect to their ideal locations and each other. Translational errors in the spatial position of axes are also a form of kinematic error. The dimensions of an axis's components can also cause the tool or workpiece to be offset from where it is supposed to be and is also classified as a kinematic error.

As shown in Figure 11.8.14 given two machine axes of motion in the XZ plane with one machine axis aligned with the X reference axis, the orthogonality error is defined as the deviation ε_y from 90° between the machine's other axis and the reference X axis. This is a straightforward definition which is simple to specify on a drawing, but is not necessarily simple to measure or control in production. Parallelism between two axes has horizontal and vertical forms that define the relative taper and twist

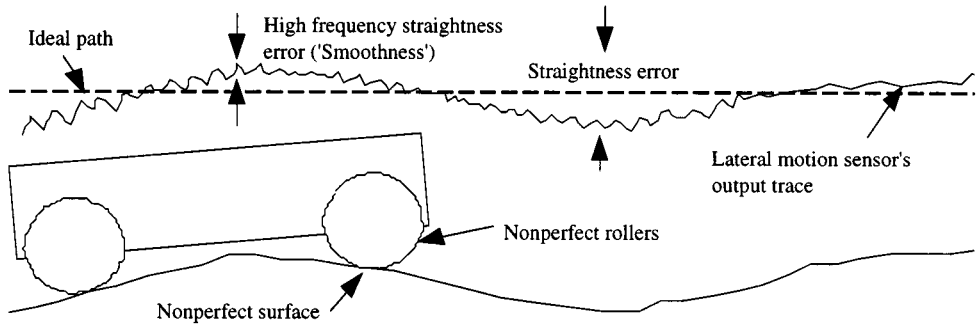


FIGURE 11.8.13 Strightness errors caused by surface form and finish errors.

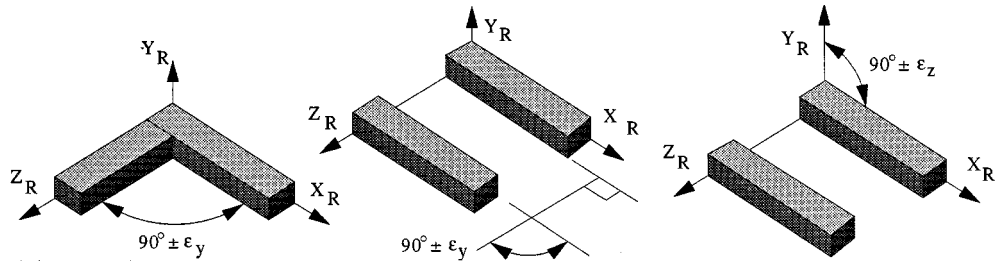


FIGURE 11.8.14 Orthogonality and horizontal and vertical parallelism errors.

between the two axes, respectively. An example of horizontal parallelism error (taper) would be an axis on a lathe that was not parallel to the spindle's axis of rotation. Using this axis to move a tool along the outer surface of a part would result in the part becoming tapered along its length. An example of vertical parallelism error is two axes used to support a milling machine bed. If one end of one of the axes is high, the bed itself will be warped when bolted to the axes, and parts machined while bolted to the bed will wobble.

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

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

In addition to increasing accuracy for enhancement of quality, machines are being required to move at greater speeds in order to increase productivity. Machine tools are usually thought of as big, bulky, slow-moving structures. The next generation of machine tools, however, will probably require axes to have acceleration capabilities in excess of 1 g. Deflections caused by accelerating the mass are acceptable for many drilling operations which would allow high spindle speeds and high axis feed rates to be used to increase productivity. Acceleration and deceleration rates of high-speed manufacturing equipment's

axes may one day routinely approach several g. In designing this type of equipment, accuracy along the path of motion may or may not be critical. However, final placement and settling time, where the maximum accelerations and inertial forces are present, will be important. Note that in the design of this type of machinery, cutting forces are often insignificant compared to inertial forces.

Another major contributor to load-induced errors in machine tools and some robots are cutting forces. Fortunately, high-speed cutting processes often generate low cutting forces, so the problems of high acceleration and high cutting force usually do not occur simultaneously. However, cutting forces are applied at the tool tip and act on every element in the machine. In order to estimate forces generated by the cutting process, actual cutting forces on machines with similar tools should be measured or appropriate handbooks consulted. In many cases, the rapid advance of new types of cutting tool materials and tool shapes will require the design engineer to consult with a tooling manufacturer or make experimental tests.

Load-Induced Errors from Machine Assembly. Even if all machine components are within required tolerance prior to assembly, additional load-induced errors can be introduced during assembly. The first type of error is forced geometric congruence between moving parts. An example is the bolting of a leadscrew to a carriage, where the nonstraightness of the leadscrew shaft creates a straightness error in the carriage with a period equal to the lead of the screw. A common example is mounting a mirror at its four corners, which often creates a visible distortion. The second type of error is the effect of the assembly process on the stiffness of the structure itself, and how the stiffness can be evaluated and incorporated into the HTM model. A third type of error, one that can be predicted, is the deformation of the machine when forces are applied to preload bearings and bolts. In addition, errors may also be caused by clamping or locking mechanisms.

Thermal Expansion Errors[*]. The need for ever-increasing accuracy and greater machine speeds makes thermal errors ever more important to control. Errors caused by thermal expansion are among the largest, most overlooked, and misunderstood form of error in the world of machine design. Thermal errors affect the machine, the part, and the tool. Even the warmth of a machinist's body can disrupt the accuracy of an ultraprecision machine. Figure 11.8.15 shows the thermal effects that must be accounted for in the design of a precision machine.[**] Thermal errors are particularly bothersome because they often cause angular errors that lead to Abbe errors.

Temperature changes induce thermal elastic strains ε_T, that are proportional to the product of the coefficient of expansion α, of the material and the temperature change, ΔT, experienced by the material:

$$\varepsilon_T = \alpha \Delta T$$

In addition, temperature gradients often cause angular errors, which lead to Abbe errors. If a machine, a tool, and a workpiece all expanded the same amount and could all be kept at the same temperature, then the system might expand uniformly with respect to the standard (always measured at 20°C) and everything would be within tolerance when brought back to standard temperature. However, different metals manufactured at different temperatures can experience serious dimensional metrology problems. Fortunately, standards have been developed (e.g., ANSI B89.6.2[***]) that define in great detail the effects of temperature and humidity on dimensional measurement and how measurements of these effects should be made. Although thermal strains can be minimized by using materials that do not expand very much,

[*] C° denotes temperature difference, whereas °C denotes an absolute temperature. This helps to avoid confusion, particularly when reference is made to temperature increases above ambient.

[**] J. Bryan, figure presented in the keynote address to the International Status of Thermal Error Research, *Ann CIRP*, Vol. 16, 1968. Also see McClure R. et al. 3.0 Quasistatic machine tool errors. October 1980. In *Technology of Machine Tools*, Vol. 5, *Machine Tool Accuracy*, NTIS UCRL-52960-5.

[***] Available from ASME, 22 Law Drive, Box 2350, Fairfield, NJ 07007-2350, (201)882-1167.

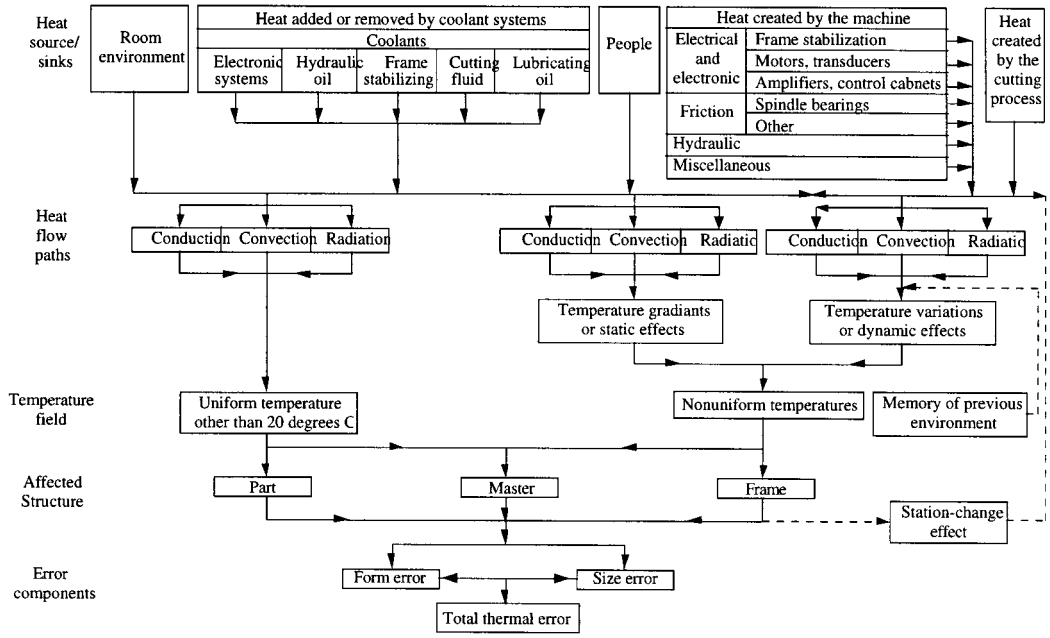


FIGURE 11.8.15 Thermal effects in manufacturing and metrology. (After Bryan.)

and by accurately controlling the environment; in practice, however, these can be difficult solutions to implement for economic reasons.

Structures

The detailed design of the machine structure requires an understanding of the errors in a machine and how they interact with the envisioned assemblage of components. When designing the structure, the engineer must have an overall philosophy in mind with respect to how to assemble the system and address principle types of errors.

For a precision machine, the engineer must first develop a strategy for attaining accuracy (these issues were addressed in the context of 11.8.1 “Analysis of Errors in a Precision Machine”):

- Accuracy obtained from component and assembly accuracy
  - Inexpensive once the process is perfected.
  - Accuracy is strongly coupled to thermal and mechanical loads on the machine.
- Accuracy obtained by error mapping a repeatable system
  - Inexpensive once the process is perfected.
  - Accuracy is moderately coupled to thermal and mechanical loads on the machine.
- Accuracy obtained from a metrology frame (measure the position of an inaccurate machine with respect to an accurate reference frame)
  - Expensive, but sometimes the only choice.
  - Accuracy is uncoupled to thermal and mechanical loads on the machine.

Next, a direction should be set regarding dynamic performance and how much is needed, by considering the following:

- System bandwidth requirements
- Effects of changing system parameters
- Methods to add damping
- Experimental modal analysis

Thermal errors are among the most difficult to predict and control, and an approach to address them should be established early on:

- Passive temperature control
- Active temperature control

With design approaches to the above, the next step is to consider structural materials and different machine configurations that are available:

- Materials
- Existing configurations

Finally, a decision should be made regarding the overall approach to be used with regard to how all the components are assembled:

- Elastically averaged design
- Kinematic design
- Bolted joint design
- Kinematic coupling design

These issues are discussed in greater detail below. Proper consideration of all the details takes a great deal of study* and practice, but these highlights should be useful as an introduction, or as a refresher.

Dynamic System Requirements

Engineers commonly ask "how stiff should the structure be?" A minimum specified static stiffness is a useful but not sufficient specification. For example, the machine can be made to not deflect too much under its own weight or the weight of a part. Fortunately, it can be predicted with reasonable accuracy. To obtain good surface finish or dynamic performance, the dynamic stiffness needs to be specified.

The dynamic stiffness of a system is the stiffness measured using an excitation force with a frequency equal to the damped natural frequency of the structure. The dynamic stiffness of a system is also equal to the static stiffness divided by the amplification at dynamic resonance. It takes a large amount of damping to reduce the amplification factor to a low level. There are several damping quantifiers that are used to describe energy dissipation in a structure. The quantifiers include:

η	Loss factor of material
η_s	Loss factor of material (geometry and load dependent)
$Q = A_r$	Resonance amplification factor
ϕ	Phase angle ϕ between stress and strain (hysteresis factor)
δ_{Ld}	Logarithmic decrement**
ΔU	The energy dissipated during one cycle
ζ	The damping factor associated with second-order systems

* See, for example, Slocum, A. 1997. *Precision Machine Design.* Society of Manufacturing Engineers, Dearborn, MI. Much of the contents of this section was derived from this book.

** Most texts on vibration refer to the log decrement as δ, however, to avoid confusion with discussions on displacement termed δ, the log decrement will be referred to here as δ_{Ld}.

The various damping terms are related (approximately) in the following manner:

$$\eta = \frac{1}{A_r} = \frac{\delta}{\pi} = \phi = \frac{\Delta U}{2\pi U} \tag{11.8.18}$$

The logarithmic decrement δ_{Ld} is an extremely useful measure of the relative amplitude between N successive oscillations of a freely vibrating system (one excited by an impulse):

$$\delta_{Ld} = \frac{-1}{N}\log_e\left(\frac{a_N}{a_1}\right) \tag{11.8.19}$$

The logarithmic decrement can also be related to the damping factor ζ, velocity damping factor b, mass m, and natural frequency ω_n of a second-order system model:

$$\zeta = \frac{\delta_{Ld}}{\sqrt{4\pi^2 + \delta_{Ld}^2}} \tag{11.8.20}$$

Note that the amplification at resonance of a second-order system is given by

$$Q = A_r = \frac{1}{2\zeta\sqrt{1-\zeta^2}} \quad (\zeta \le 0.707) \tag{11.8.21}$$

For example, Figure 11.8.16 shows the time decay of a fairly well-damped system. For this system, n = 4, $a_5 = 0.5$, $a_1 = 1.5$, $\delta_{ld} = 0.275$, $\zeta = 0.44$, and Q = 11.5.

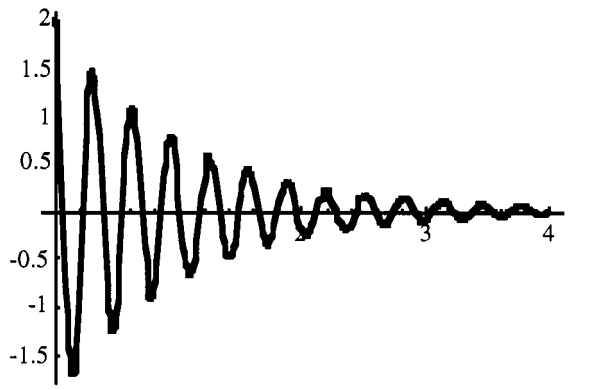


FIGURE 11.8.16 Time decay signal from a fairly well-damped system struck with an impulse.

Machines with good dynamic performance achieve damping by material selection and by the bolted joints and bearing interfaces in the structure. Unfortunately, material and joint damping factors are difficult to predict and are often too low. Thus the designer must consider the stiffness, mass, and damping of the system, and be prepared to alter them to achieve the desired level of performance. For high speed or high accuracy machines, damping mechanisms may have to be designed into the structure in order to meet realistic damping levels. The first step, however, is to try and determine the bandwidth required with an initial estimate of the stiffness, mass, and damping in the system. Then with these goals, one can try to change the parameter that has the greatest impact on performance.

System Bandwidth Requirements. An estimate of the system's servo bandwidth can be made by considering the motions the system is required to make. Most systems null high frequency disturbance forces with their own mass or added damping; however, lower frequency forces must be offset by forces from the controller/actuator. As a result, the servo bandwidth required is primarily a function of the types of moves that the system will be required to make. For start and stop moves (e.g., in a wafer stepper), the bandwidth should be on the order of

$$\omega(\text{hz}) = \frac{10}{2\pi t_{\text{settling time}}} \tag{11.8.22}$$

When contouring, the X and Y axis moves according to x = Rsinωt and y = Rcosωt, the frequency ω is a function of the linear velocity of the cutter through the material, and the radius of the contour:

$$\omega(\text{hz}) = \frac{v_{\text{linear}}}{2\pi r_{\text{path radius}}} \tag{11.8.23}$$

For example, for a large circle (e.g., 200 mm D) being cut at modest speeds (e.g., 0.1 m/sec), the bandwidth required is only 0.16 Hz. On the other hand, for a sharp turn in a corner, the radius of the cutter path trajectory may only be 1 mm, so ω > 16 Hz. The latter is a reasonable requirement for a large machine tool, where the spindle axis may weigh 1000 kg and thus require a critically damped system stiffness of 10 N/μm, which means a design stiffness on the order of 100 N/μm when one considers the limited damping that is obtainable.

In order to make a better assessment of what system parameters should be, a simple model can be used in the early stages of design. For example, a system with a motor driving a carriage can be modeled as shown in Figure 11.8.17.

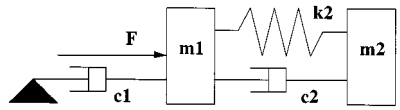


FIGURE 11.8.17 Effect of changing system mass on dynamic performance.

In this model:

- m_1 is the mass of a linear motor forcer or m_1 is the reflected inertia of the motor rotor and leadscrew (or just a linear motor's moving part):

$$M_{\text{reflected}} = \frac{4\pi^2 J}{l^2}$$

- m_2 is the mass of the carriage.
- c_1 is the damping in the linear and rotary bearings.
- c_2 is the damping in the actuator-carriage coupling and the carriage structure.
- k_2 is the stiffness of the actuator and actuator-carriage-tool structural loop.

The equations of motion of this system are

$$\begin{vmatrix} m_1 & 0 \\ 0 & m_2 \end{vmatrix} \begin{vmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{vmatrix} + \begin{vmatrix} c_1 + c_2 & -c_2 \\ -c_2 & c_2 \end{vmatrix} \begin{vmatrix} \dot{x}_1 \\ \dot{x}_2 \end{vmatrix} + \begin{vmatrix} k_2 & -k_2 \\ -k_2 & k_2 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} + \begin{vmatrix} F(t) \\ 0 \end{vmatrix} \tag{11.8.24}$$

The transfer function x_2/F (dynamic response of the carriage) is

$$\frac{x_2}{F} = \frac{k_2 + c_2 s}{c_1 s\left(k_2 + c_2 s + m_2 s^2\right) + \left(m_1 + m_2\right) s^2 \left(k_2 + c_2 s\right) + m_1 m_2 s^4} \tag{11.8.25}$$

Note that the product of the masses term tends to dominate the system. To illustrate the design implications, consider four different calculated design options for a linear motion system shown in Figure 11.8.18.

Actuator	ballscrew	lin. motor	lin. motor	lin. motor
Bearings	linear ball	linear ball	air	air
Structural damping	no	no	no	yes
material damping zeta	0.005	0.005	0.005	0.1
actuator to ground zeta	0.05	0.03	0	0
m1 (actuator) (kg)	50	5	5	5
m2 (carriage), kg	50	50	50	50
c1 (N/m/s)	187	355	0	0
c2 (N/m/s)	19	19	19	374
k1 (N/m)	1.75E+08	1.75E+08	1.75E+08	1.75E+08
Bandwidth (Hz)	25	100	30	100

FIGURE 11.8.18 Calculated parameters of four possible linear motion systems.

As a guideline, the servo bandwidth of the system is generally limited by the frequency at which the servo can drive the system without exciting structural modes. Without special control techniques, the servo bandwidth can be no higher than the frequency found by drawing a horizontal line 3 dB above the resonant peak to intersect the response curve. This method is used only to initially size components. A detailed controls simulation must be done to verify performance and guide further system optimization. As an example, consider the response of the ballscrew-driven carriage supported by rolling element linear bearings, as shown in Figure 11.8.19. In this case, since preloaded linear guides and a ballnut are used, damping to ground will be high. The inertia of the ballscrew will lower the system frequency considerably (note the $m_1 m_2 s^4$ term in the transfer function). Going up 3 dB from the resonance peak and projecting to the left to intersect the response curve gives an estimate of the maximum bandwidth the machine can be driven at, without danger of exciting a resonance, of about 25 Hz.

Effects of Changing the System's Mass. Decreasing the mass of the system will enhance the ability of the machine to respond to command signals. However, the trade-off is that the system will lose the ability to attenuate high frequency noise and vibration as shown in Figure 11.8.20.

A system with decreased mass offers a higher natural frequency, which means that higher speed controller signals can be used without compromising the accuracy of the system. A low mass system also shows improved damping, a result of the increased loss factor [the loss factor $\zeta = c/(2m)$], but a low mass system shows less noise rejection at higher frequencies. This suggests that the machine will be less able to attenuate noise and vibration. Considering all the issues, in conclusion, the mass should be minimized to reduce controller effort and improve the frequency response and loss factor (ζ).

Effects of Changing the System's Stiffness. A system with higher stiffness will give a flatter response at low frequencies and give smaller displacements for a given force input. More importantly, the compromise of decreased noise attenuation is not as dramatic as is the case with lowering the system mass. This is shown by the similar shapes in the three curves at high frequencies in Figure 11.8.21. However, acoustical noise may be worsened by adding stiffness. In conclusion, the stiffness of the machine structure should be maximized to improve positioning accuracy.

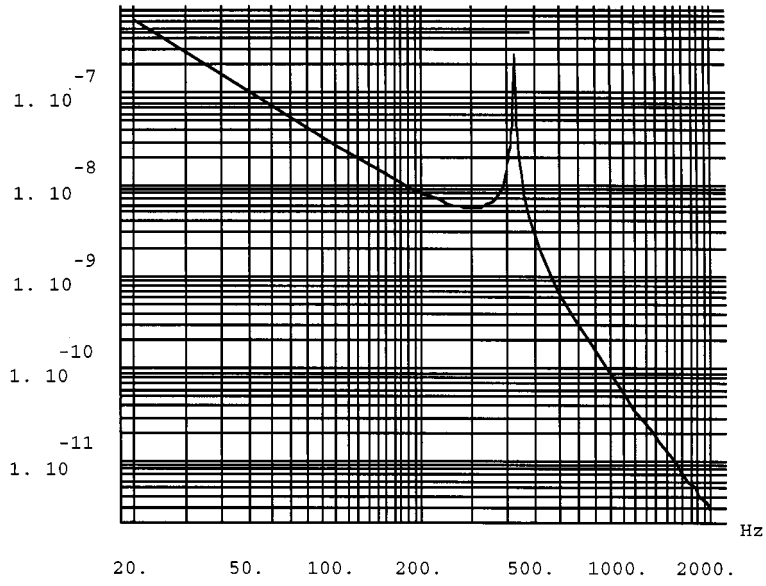


FIGURE 11.8.19 Frequency response of a ballscrew-driven system.

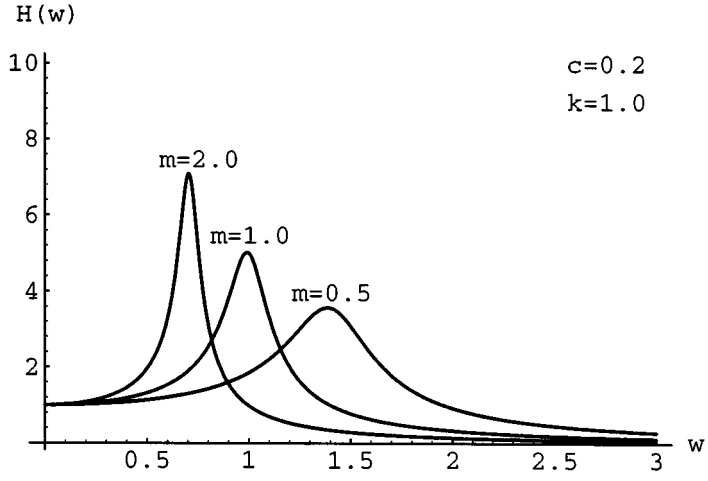


FIGURE 11.8.20 Effect of changing system mass on dynamic performance.

Effects of Changing the System's Damping. Increasing the system's damping can make a dramatic improvement in the system's response. The trend is for decreasing amplification of the output at resonance with increasing damping, although a damping coefficient of 0.4 may be difficult to obtain in practice. Figure 11.8.22 shows the dramatic improvement available by doubling the system's damping.

Methods of Achieving Damping

Although it has been extensively studied, the mechanism of damping in a material is difficult to quantify and one must generally rely on empirical results.* In fact, damping is highly dependent on alloy

* A discussion of the many different microstructural mechanisms that generate damping in materials is beyond the scope of this book. For a detailed discussion see Lazan, B.J., *Damping of Materials and Members in Structural Mechanics*. Pergamon Press, London.

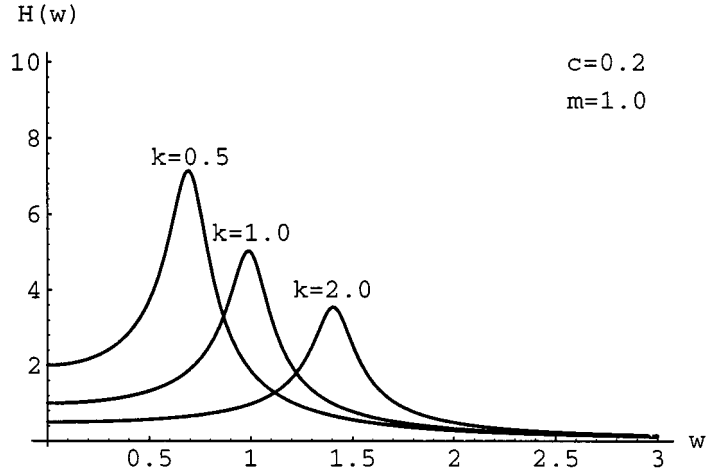


FIGURE 11.8.21 Effect of changing system stiffness on dynamic performance.

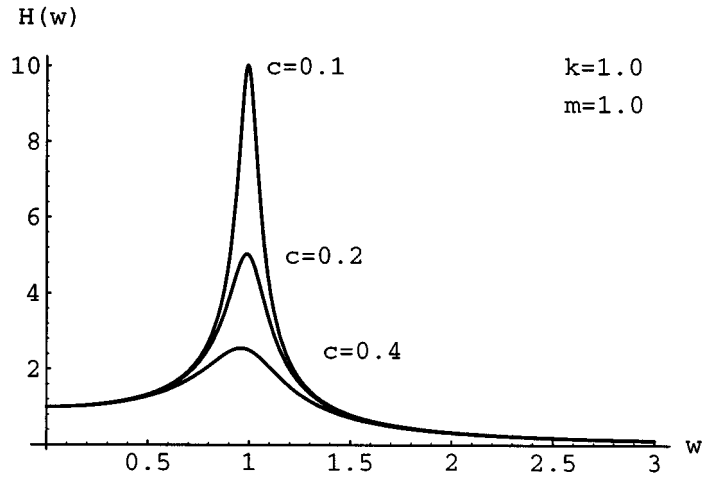


FIGURE 11.8.22 Effect of changing system damping on dynamic performance.

composition, frequency, stress level and type, and temperature. Structural damping levels are often quite low, and frequently the dominant source of damping is the joints in an assembly. In fact, one must be extremely wary of damping data that are presented in the literature, because often it is presented without a discussion of the design of the test setup.

The amount of damping one obtains from a material is very low compared to the amount of damping that one can obtain with the addition of a damping mechanism. Damping mechanisms can range from simple sand piles to more complex shear dampers or tuned mass dampers as discussed below. In general, cast iron is the best damped structural metal. Polymer concrete or granite make well-damped structures that tend to be more monolithic in nature; but ceramics have very poor damping.

Tuned Mass Dampers. In a machine with a rotating component (e.g., a grinding wheel), there is often enough energy at multiples of the rotational frequency (harmonics) to cause resonant vibrations in some of the machine's components. This often occurs in cantilevered components such as boring bars and some grinding wheel dressers. The amplification at a particular frequency can be minimized with the use of a tuned mass damper. A tuned mass damper is simply a mass, spring, and damper attached to a structure at the point where vibration motion is to be decreased. The size of the mass, spring, and damper

are chosen so they oscillate out of phase with the structure and thus help to reduce the structure's vibration amplitude.

Consider the single-degree-of-freedom system shown in Figure 11.8.23. The system contains a spring, mass, and damper. For a cantilevered steel beam, the spring would represent the beam stiffness, the mass would be that which combined with the spring yielded the natural frequency of the cantilevered beam, and the damper would be that which caused a 2% energy loss per cycle. As also shown in Figure 11.8.23, a second spring-mass-damper system can be added to the first to decrease the cantilevered beam's amplitude at resonance. The equations of motion of the system are

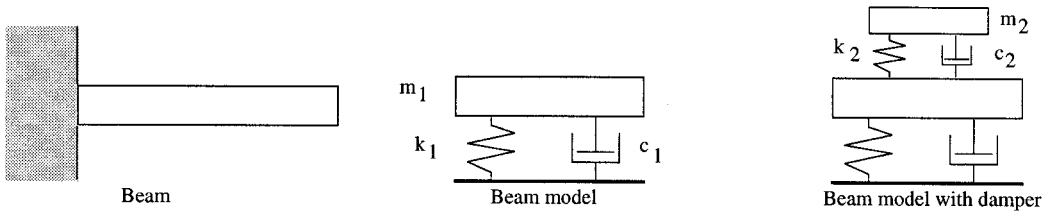


FIGURE 11.8.23 Cantilever beam, model, and model with tuned mass damper.

$$\begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \ddot{x}(t) + \begin{bmatrix} c_1 + c_2 & -c_2 \\ -c_2 & c_2 \end{bmatrix} \dot{x}(t) + \begin{bmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 \end{bmatrix} x(t) \tag{11.8.26}$$

In the frequency domain, in order to present a solution for the motion of the system, the following notation is introduced:

$$Z_{ij}(\omega) = -\omega^2 m_{ij} + i\omega c_{ij} + k_{ij} \qquad i, j = 1, 2 \tag{11.8.27}$$

The amplitudes of the motions of the component and the damper as a function of frequency are given by*

$$X_2(\omega) = \frac{-Z_{12}(\omega) F_1}{Z_{11}(\omega) Z_{22}(\omega) - Z_{12}^2(\omega)} \tag{11.8.28}$$

The design of a tuned mass damper system for a machine component may involve the following steps:

- Determine the space available for the damper and calculate the mass (m_2) that can fit into this space.
- Determine the spring size (k_2) that makes the natural frequency of the damper equal to the natural frequency of the component.
- Use a spreadsheet to generate plots of component amplitude as a function of frequency and damper damping magnitude (c_2).

Constrained Layer Dampers. The structural joints in a machine tool have long been known to be a source of damping by the mechanisms of friction and microslip. A study of structural joint damping has shown that numerous theories are available for predicting damping by these mechanisms,** however, the amount of damping obtained is still less than what is required for critical damping, and controlling the surface interface parameters at the joint to achieve uniform results from machine to machine is difficult. In addition, as far as the accuracy of the machine is required, it would be best if the joints behaved as

* Ibid., p 115.

** See, for example, Tsutsumi, M. and Ito, Y. September 1979. Damping mechanisms of a bolted joint in machine tools, *Proc. 20th Int. Mach. Tool Des. Res. Conf.*, pp. 443–448; and Murty, A.S.R. and Padmanabhan, K.K. 1982. Effect of surface topography on damping in machine joints. *Precis. Eng.* 4(4):185–190.

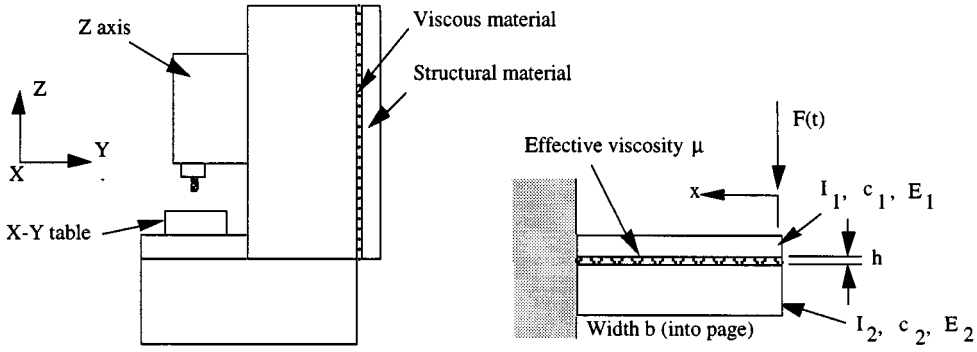


FIGURE 11.8.24 Increasing structural damping by adding alternate layers of viscous and structural materials.

a rigid interface which, as discussed below, can be obtained by bolting and grouting (or bonding) a joint. A better method is to add damping by applying alternate layers of viscoelastic and structural materials to a machine tool structure,* as illustrated in Figure 11.8.24.

Experimental Modal Analysis of a Machine's Dynamic Performance

The manner in which a machine behaves dynamically has a direct effect on the quality of the process. It is vital to be able to measure a machine's dynamic performance to enable engineers to determine exactly which parts of a machine need to be strengthened or made to be better damped.

Experimental modal analysis allows the study of vibration modes in a machine tool structure. An understanding of data acquisition, signal processing, and vibration theory is necessary to obtain meaningful results. For a modest-sized company (50 employees), it is feasible and reasonable to have one person dedicated to metrology, including experimental modal analysis. Companies routinely invest $100,000 or more in precision measurement devices (e.g., gauges, autocollimator, laser interferometer, ball bar, surface roughness, etc.). The dynamic performance of a machine is every bit as important, and requires only a modest investment (about $40K).

When a company's metrologist is trained in modal analysis, results can be rapidly obtained and presented to the design department. For example, suppose a new machine is being designed where axial stiffness of the drive system is of extreme importance (e.g., a centerless grinder). The design engineer wants to know if a rollerscrew will give better performance than a ballscrew. Not completely trusting catalog data, and not knowing the primary source of compliance (the mounting, the bearings, the shaft, or the nut), the engineer replaces a ballscrew — in a known good machine — with a rollerscrew. The metrologist does the modal analysis and the engineer learns in "real time" what the result is. Similarly, the engineer can look to other machines on the shop floor for physical models of elements to be used in a new design, and the metrologist can make the measurements in a short time.

All too often a modal analysis is done after a design is complete to help solve a problem. The results can often be used to help repair the problem; however, done before hand, modal testing of similar machines or test-subassemblies would have led to a better design approach that in the end would have resulted in a better integrated system. The results of a modal analysis yield important information regarding the machine's dynamic properties, including:

- Modal natural frequencies
- Modal damping factors
- Vibration mode shapes

* See, for example, Haranath, S., Ganesan, N., and Rao, B. 1987. Dynamic analysis of machine tool structures with applied damping treatment. *Int. J. Mach. Tools Manuf.* 27(1):43–55. Also see Marsh, E.R., Slocum, A.H. 1996. An integrated approach to structural damping. *Precis. Eng.*, 18(2,3):103–109.

This information is vital to a designer, who may use the information to

- Locate sources of compliance in a structure.
- Characterize machine performance.
- Optimize design parameters.
- Identify the weak links in a structure for design optimization.
- Identify modes which are being excited by the process (e.g., an end mill) so that structure can be modified accordingly.
- Identify modes (parts of the structure) which limit the speed of operation (e.g., in a coordinate measuring machine).

A manager may ask "why should we start doing this if we have not done it in the past, and were very successful?" The engineering answer is that faster and more accurate machines are needed in the future, and the materials being processed are getting harder and harder (e.g., ceramics). The business answer is that the problems are getting tougher to solve, and engineering time is expensive. Also, international competition is more widespread and much tougher. History has shown that application of new advanced tools virtually always guarantees an increase in performance and competitive edge.

Identification, Control, and Isolation of Heat Sources*

Heat that causes thermal errors is introduced into the machine from a number of sources including moving parts (e.g., high-speed spindles, leadscrew nuts, linear bearings, transmissions), motors, the material removal process, and the external environment (e.g., sunshine through a window, direct incandescent lighting, heating ducts, the floor, operators' body heat, etc.). Heat transfer mechanisms in a machine include conduction, convection, evaporation,** and radiation. They are executed by recirculating lubricating oil and cutting fluid, chips from the cutting process, conduction through the frame of the machine, convection of air in and around the machine, and internal and external sources of radiation.

There are three schools of thought regarding minimization of thermally induced errors. The first is to prevent thermal expansion in the first place. The second is to minimize the time it takes for the machine to reach its equilibrium temperature, or in some cases, bring the entire machine to a uniform temperature (e.g., circulate, temperature-controlled fluid through the machine), which can help to minimize differential expansion. The third is to disregard the effects of thermal errors and simply map them, which is a very difficult and often impractical thing to do. Regardless of the methodology followed, a thorough understanding of the effects of heat sources and transfer mechanisms is required, which is beyond the scope of this work.

The manner in which the temperature of a machine is to be controlled can have a large impact on the machine's design. In summary, there are many options available to the design engineer:

- Passive temperature control:
  - Minimize and isolate heat sources.
  - Minimize coefficient of thermal expansion.
  - Maximize thermal conductivity to minimize thermal gradients.
  - Maximize thermal diffusivity to quickly equilibrate transient thermal effects.
  - Minimize thermal emissivity of the structure to minimize radiant coupling to the environment, or maximize the emissivity to couple the structure to an environmental control enclosure.

* An important reference to have that discusses many of the issues discussed here in greater detail is *Temperature and Humidity Environment for Dimensional Measurement*, ANSI Standard B89.6.2-1973.

** Evaporation cooling occurs when aqueous fluids are used. Evaporative cooling is usually uneven and represents one of the biggest temperature control problems facing the precision machine design engineer.

- Active temperature control:
  - Air showers
  - Circulating temperature-controlled fluid within the machine
  - Oil showers
  - Thermoelectric coolers for localized temperature control of hot spots

Each conceptual design option must consider how the temperature within the machine will vary if each of these different temperature control strategies were applied.

Material Considerations

For the conceptual design phase, one should design the structure using differently available materials, and then also design a multimaterial hybrid. For example, cast iron can be made into virtually any shape, so the design engineer has greater freedom, but large sections are expensive to thermally anneal, which must be done to achieve material homogeneity and stability. Granite is usually used in the form of simple rectangular, circular, or planar shapes. Ships are made from welded steel plate and thus conceivably any size of machine tool could also be welded together. Polymer concrete can be cast into virtually any shape and requires no stress relief or prolonged aging cycle. With new ceramics and composite materials, the choices become even more varied, so one really must be alert and consider all options.

Cast Iron Structures

The good general properties of cast iron and the ease with which parts can be cast have made cast iron the foundation of the machine tool industry. Generally, when a machine tool component is smaller than a compact car, it is a candidate for being made of cast iron, although castings weighing hundreds of tons have been made. For larger parts or where economy is of prime importance, one should consider welding together plates and standard structural shapes (e.g., boxes, angles, I-beams, and channels), as discussed below.

Sand casting can be used to make virtually any size part, and it basically involves making a pattern out of a suitable material (e.g., foam, wood, or metal) and packing sand around it. Parting lines are used to allow the sand mold components to be disassembled from around the pattern and then reassembled after the part is removed. Sand cores are often inserted into the mold to form cavities inside the mold (e.g., the cylinders of an engine casting). Regardless of the design of the part, one must also consider that as the metal cools it shrinks (on the order of 5 to 10% for most metals), and that in order to remove the part from the mold without breaking the mold, a taper (draft) of about 1:10 is required. In addition, extra metal should be added to surfaces that will have to be machined (a machining allowance), and locating surfaces should be added so that the part can be fixtured to facilitate machining. Thus in order to specify a casting, there are a few basic guidelines one needs to know in order to minimize the work that a professional mold design engineer has to do to clean up your design. These guidelines are discussed below.[*]

Granite Structures[**]

Granite is used as a structural material in machines that are generally used in dry environments because granite can absorb moisture and swell. Thus it may not be appropriate to use granite in a machine where cutting fluid splashes all over it, although there is some debate as to how susceptible to swelling granite actually is. Granite can be sawed into a part of virtually any shape or size (deviations from round slabs

[*] See, for example, the handbook *Casting Design as Influenced by Foundry Practice.* Mechanite Metal Corp., Marietta, GA.

[**] A good source of design information about what shapes and sizes can be made is available from Rock of Ages Corp., Industrial Products Group, P.O. Box 482, Barre, VT 05641.

or rectilinear shapes can be expensive). Since it is a brittle material, sharp corners are not allowed, and most structures are built from pieces that are bolted (using inserts) and grouted or bonded together. Since granite is brittle, threaded holes cannot be formed, and thus threaded inserts must be glued or press-fit into round holes in the granite. Common applications of granite components in machine tools include structural members and air-bearing ways in coordinate measuring machines and other inspection machines. Note that the porosity of some granite makes it unsuitable for air bearings even after it has been polished.

The low thermal conductivity of granite makes it slow to absorb heat. This makes granite, particularly black granite, susceptible to thermal distortions caused by the top surface absorbing heat from overhead lights. Granite's coefficient of thermal expansion is less than that of most metals, so in the process of manufacture, shipping, and use, one must consider how differential thermal expansion will affect a machine's performance if metal components have been bolted to it. Although granite has been sitting in the ground for millions of years and thus may seem to possess the ultimate stability, one must consider the effects of relieving the stress upon the granite's dimensional stability after it has been quarried. There are suppliers of very stable granite components, so the design engineer must carefully shop around. A very desirable property of granite (or any other brittle material) is that it will chip when banged instead of forming a crater with a raised lip (Brinell). Granite is also relatively inexpensive to quarry, cut, and lap. Hence granite is often the material of choice for coordinate measuring machine tables.

Welded Structures*

Welded structures can be made from any weldable alloy (e.g., iron alloys such as 1018 structural steel or Invar). Welded structures (weldments) are commonly used, for example, where (1) the cost of a large structure is to be minimized, (2) a high degree of material damping is not needed or the structure will be filled with a damping material, and/or (3) the part is too large to be cast and thermally stress relieved economically. If welded properly, so that the weld material is in a metallurgically stable form, residual stresses can sometimes be removed using vibratory stress-relief methods.

A welded structure is similar to a cast structure, in that strength and stiffness are achieved through the use of sections that are reinforced with ribs; consequently, structural analysis can be difficult and beyond the conceptual design phase often requires the use of finite element methods. Damping and heat transfer characteristics across welded joints are also difficult to model because they are dependent on depth of weld penetration, composition of the weld material, and contact pressure between member surfaces at the joints where the weld does not penetrate. One solution is to specify full penetration welds. To minimize the cost of welded structures, the number of parts and linear meters of weld must be minimized. Furthermore, the more welds that are made, the greater the thermal distortion that is likely to occur as a result of the manufacturing process. However, if too few ribs are used, large plate sections can vibrate like drums. Damping and thermal performance of a welded structure can improve greatly if a viscous temperature-cooled fluid is recirculated throughout the structure, or if damping mechanisms are used as described above. A welded structure can also be used as a mold for a polymer concrete casting which creates a heavy but well-damped and stiff structure, but care must be taken to avoid bimaterial thermal deformation problems.

Ceramic Structures**

The first introduction of a machine that was made almost entirely of ceramics was in 1984 at the Tokyo machine tool show. Although all-ceramic machines generally perform admirably, they have yet to compete economically with machines made from cast iron or polymer concrete. However, in the future,

* A good reference to have is Blodgett, O. 1963. *Design of Weldments*. James F. Lincoln Arc Welding Foundation, P.O. Box 3035, Cleveland, OH 44104.

** See, for example, Ormiston, T. September 1990. Advanced Ceramics and Machine Design. *SME Tech. Paper* EM90-353.

as more and more ceramic components are made for use in consumer products (e.g., automobiles), it is likely that precision machines will contain more and more ceramic components.*

Hard materials (e.g., ceramics) offer advantages over conventional materials in terms of dimensional stability, strength, and stiffness over a wide range of temperatures. In applications ranging from adiabatic internal combustion engines for maximum efficiency to X-ray mirrors for X-ray photolithography, the ability to manufacture components from hard materials is clearly of vital importance to the future of the manufacturing industry. Unfortunately, most ceramic components are finish machined on machines built from cast iron and the abrasive nature of ceramics limits the life of these machines. Thus a new family of machine tools and machine tool components will have to be developed specifically for the manufacture of ceramic components. Consider several key properties of some ceramic materials that can help guide the design process for new machines:

- Most ceramic materials (e.g., aluminum oxide and silicon nitride) will not corrode in any fluid environment that might be used in the manufacture of ceramic components (i.e., fluids from oil to water).** Thus ceramics are key materials for water hydrostatic bearings.
- The more brittle a material is, the less plastic deformation that is generated during finish grinding or lapping; hence the surface is more likely to have a negative skewness. A surface with a negative skewness minimizes the need for a lubricant that has good wetting properties that allows for water to be considered as a lubricant. A surface with a negative skewness will also suffer less damage in the event that the lubricant is lost.
- The more brittle a material is, the less plastic deformation that is generated during finish grinding; hence the surface is less likely to contain high residual stress levels that can lead to dimensional instability. In addition, a precision machine should not be subjected to shock loads in the first place. For general machine tool applications, such as bearing rails, the element geometry generally provides more than enough strength to withstand impact loads generated when a machine crashes.
- Generous radii must be used in all corners, and threaded steel inserts must be used if parts are to be bolted to ceramic components. To bond two ceramic components together, conventional adhesives can be used, or for high-performance applications, ceramic parts can be frit bonded. In frit bonding, a glass powder is applied to the two mating surfaces and then fused by heating the parts above the melting point of the glass. The bond will not be as strong as the ceramic, but it will be almost as stiff.
- Unlike some metals, elements do not precipitate out of a ceramic material's microstructure with time (e.g., carbides do not form like in some iron alloys) so dimensional stability is enhanced. Thus ceramic components can also have virtually unmatched dimensional stability. For metrology masters (e.g., squares and straight edges), ceramics are much lighter than conventional materials, they are less likely to be damaged (dropped) in the first place, and they are less likely to become scratched or worn in everyday use.
- Most ceramics are pure and thus the achievable surface finish is limited only by the size of the grain structure formed during the sintering process; however, most advanced ceramic materials have submicron sizes and this effect is usually not a problem the way it can be with metals. Also note that metals contain discrete hard particles that are dragged across the softer surface during machining which degrades surface finish.

* Furukawa, Y. et al. 1986. Development of ultra precision machine tool made of ceramics. *Ann. CIRP.* 35(1): 279–282.

** Aluminum oxide components are subject to stress corrosion cracking in aqueous environments and thus care must be taken to minimize tensile stresses. Note that silicon-based ceramics, which have predominantly covalent molecules, are far less reactive with water. Only at high temperature and pressure do silicon-based ceramics become appreciably affected by aqueous environments.

- Ceramic materials have a high modulus that is good for machine stiffness, but some have poor thermal properties (e.g., alumina) that can lead to increased thermal deformations of the machine. Note that silicon carbide has very good thermal properties but it is considerably more expensive than alumina
- Ceramic materials in intimate contact will not gall or fret the way many metals often do; thus they make excellent bearing surfaces. However, when ceramic materials are in intimate sliding contact, traction forces can cause the local tensile strength to be exceeded, which produces surface cracks that lead to spalling. In such situations, it may be desirable to use a ceramic material with a high fracture toughness and tensile strength (e.g., silicon nitride). Thermal stresses can also initiate local or gross failure.
- Ceramic materials have a higher modulus of elasticity and lower density than do bearing steels; hence ceramic rolling elements have a smaller contact zone that leads to less heat generation. Hybrid bearings (e.g., metal rings and ceramic rolling elements) generate 30 to 50% less heat than do steel bearings. This means that grease can be used to lubricate the bearings at much higher speeds.* In general, for smaller diameter ultraprecision bearings, hybrid bearings can have up to three times the DN value of steel bearings (i.e., 4.5 million vs. 1.5 million).
- Aluminum oxide has good overall properties for structural applications such as fluidstatic bearing rails and CMM structures. Zirconia is very tough, and its coefficient of thermal expansion matches that of steel, so it can be used as a bearing rail liner for rolling element bearings without worrying about bimaterial expansion problems. However, most zirconias are multiphase materials and thus are not well suited for applications where high-dimensional stability is required (e.g., in precision bearings). Silicon nitride has the best overall properties including very high toughness, which makes it ideal for rolling element bearings, but it is too expensive to use for large structural components. Silicon carbide and tungsten carbide are extremely hard and wear-resistant and they are often used in cutting tool applications.
- Aluminum oxide components can be made by cold pressing followed by machining, firing, and grinding. Note that there is significant volume shrinkage during the firing process. Hence designing ceramic structural parts often requires assistance from the manufacturer. In general, the same shape design rules apply as for metal castings, and the wall thickness should not be greater than about 25 mm. Ceramics components (e.g., those made from silicon nitride) can also be made by hot pressing followed by grinding and lapping.

Polymer Concrete Castings**

Portland cement-based concrete is not dimensionally stable enough, due to its own internal structural variations with time and its hygroscopic nature, to allow it to be used for the main structure of a precision machine tool. Although in many applications, properly cured reinforced concrete on a stable, dry subgrade can provide a reasonably stable foundation for very large machines that are not self-supporting, unreinforced Portland cement concrete itself is not dimensionally stable due to (1) reaction shrinkage from cement hydration; (2) shrinkage due to loss of excess nonstoichiometric water, which leaves conduits for humidity-induced expansion or contraction; and (3) nonelastic dimensional changes (e.g., creep and microcracking in the inherent brittle/porous structure). Overall, strain variations with time may be as high as 1000 μm/m.

* Steel bearings require oil mist lubrication at higher speeds. Introducing an oil mist (oil dripped into a high-pressure air stream) into a bearing increases the chance that water and dirt particles might be introduced into the bearing, which can lead to premature failure. Actually, a precision bearing cooled by an oil mist should have its air cleaned and dried to a level usually associated with air bearings (e.g., 3-μm filter and H_2O < 50 to 100 ppm).

** See, for example, Capuano, T. September 10, 1987. Polymer concrete. *Mach. Des.* 133–135; Jablonowski, J. August 1987. New ways to build machine structures. *Am. Mach. Automat. Manuf.* 88–94; and McKeown, P.A. and Morgan, G.H. 1979. Epoxy granite: a structural material for precision machines. *Precis. Eng.* 1(4): 227–229.

Fortunately, a number of different types of polymer-based concretes have been developed which can be used to cast machine tool quality structures. For example, Fritz Studer AG, a prominent manufacturer of precision grinders in Switzerland, discovered that special polymers can be used to bind together specially prepared and sized aggregate to yield a stable, essentially castable, granite-like material with a damping coefficient much higher than that of cast iron.* By carefully controlling the manufacturing process and selection of binder and aggregate, properties can be varied somewhat to suit the user. The polymer concrete material and process developed by Studer is known as Granitan™ and its composition and manufacturing process was patented. Numerous companies have licensed the process and will make castings from Granitan™ to order. Other companies have developed their own proprietary polymer concretes with similar high performance properties.

For polymer concrete castings, the same rules for draft allowance apply as for metal castings if the mold is to be removed. Instead of ribs, polymer concrete structures usually use internal foam cores to maximize their stiffness-to-weight ratio. Unlike metal castings, a polymer concrete casting will not develop hot spots while curing even in thick, uneven sections. Polymer concrete castings can readily accommodate cast-in-place components such as bolt inserts, conduit, bearing rails, hydraulic lines, etc. It should be noted that a bolt will fail before a bonded-in-place insert.

With appropriate section design, polymer concrete structures can have the stiffness of cast iron structures and much greater damping than cast iron structures.** However due to polymer concretes' lower strength, heavily loaded machine substructures (e.g., carriages) are still best made from cast iron. Polymer concretes do not diffuse heat as well as cast iron structures and thus attention must be paid to the isolation of heat sources to prevent the formation of hot spots in a polymer concrete structure. In addition, their modulus of elasticity is about one fifth that of steel, and their strength is an order of magnitude less, so they are used primarily for machine bases.

Structural Configurations

For proper functioning of moving axes and operational stability, often it is important to minimize thermal and elastic structural loops. By this it is meant that the path length from the toolpoint to the workpiece through the structure should be minimal. The less material that separates the part of the structure that holds the tool and the part of the structure that holds the part, the more likely the entire system will quickly reach and maintain a stable equilibrium.

Open-Section (C or G) Frames

Most small machines are designed with an open frame, as shown in Figure 11.8.25 which greatly facilitates workzone access for fixturing and part handling. Note that the machine could be designed with the spindle oriented in the horizontal or the vertical direction. Unfortunately, open section frames are the least structurally and thermally stable. The lack of symmetry leads to undesirable thermal gradients and bending moments. The fact that a critical part of the structure is cantilevered means that Abbe errors abound; hence great care must be used when designing a precision machine with an open frame. Note that there are many different variations on this design for different types of machines (e.g., milling machines and lathes). The common feature of all is the fact that the structural loop is open.

Closed-Section (Bridge or Portal) Frames

Most large machines are designed with a closed frame as shown in Figure 11.8.26. When the Z motion is built into the bridge, a second actuator must often be slaved to the primary actuator that moves the

* Kreienbühl, R. September 19–20, 1990. Experience with synthetic granite for high precision machines. In *Proc. Symp. Mineralguss im Maschinenbau.* FH Darmstadt; and Renker, H.J. 1985. Stone based structural materials. *Precis. Eng.* 7(3): 161–164.

** See, for example, Weck, M. and Hartel, R. 1985. Design, manufacture, and testing of precision machines with essential polymer concrete components. *Precis. Eng.* 7(3): 165–170; and Salje, I. et al. 1988. Comparison of machine tool elements made of polymer concrete and cast iron. *Ann CIRP.* 37(1): 381–384.

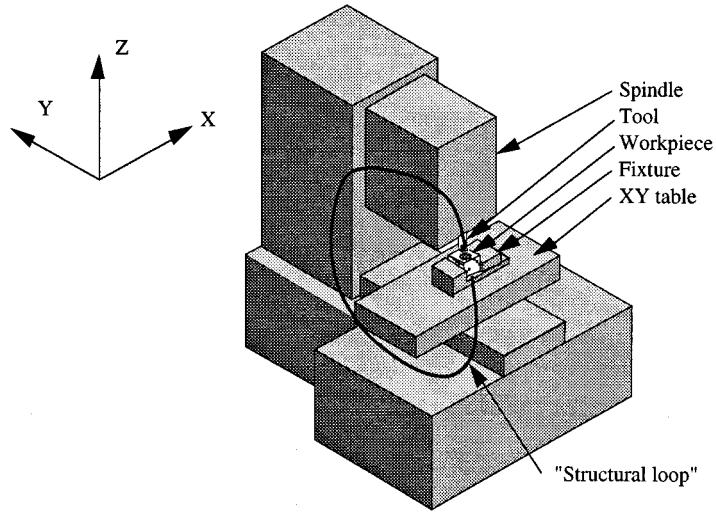


FIGURE 11.8.25 Structural loop in an open-frame machine tool.

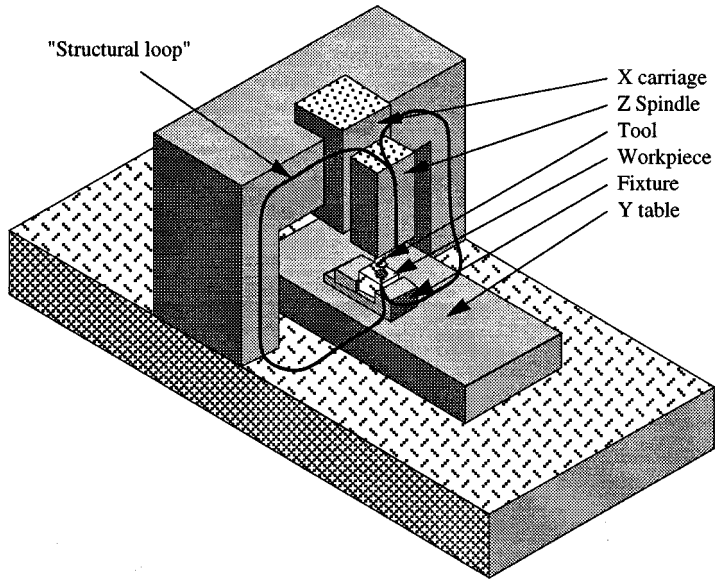


FIGURE 11.8.26 Structural loop in a closed-frame (bridge) machine tool.

bridge. This prevents the bridge from yawing (walking). Note that there are many different variations on this design for different types of machines (e.g., milling machines and lathes). The common feature of all is the fact that the structural loop is closed.

Tetrahedral Frames

Nature invented the tetrahedron and found it to be an immensely stable and powerful form (e.g., diamonds). In engineering and architecture, the tetrahedron represents the three-dimensional application of the age-old structure of stability, the triangle. Lindsey of NPL in England took these basic building blocks of nature and added well-engineered damping mechanisms to yield a major advancement in the

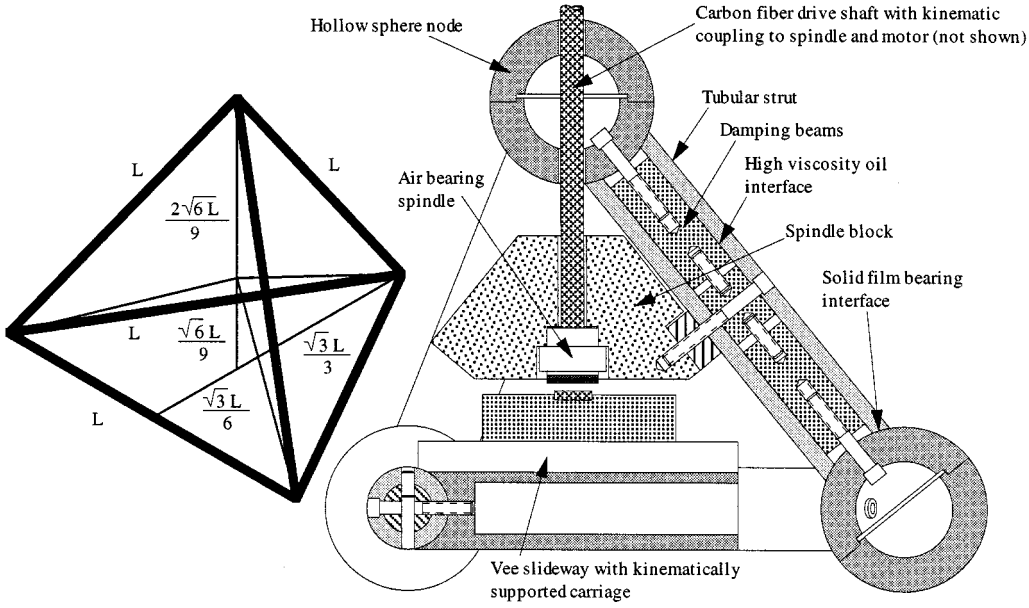


FIGURE 11.8.27 Tetraform structural concept for machine tools and instruments developed by Lindsey. (Courtesy of the National Physical Laboratory.)

structure of machine tools. The *Tetraform* machine tool concept is shown in Figure 11.8.27. The structures owes its exceptional dynamic performance to the following:*

- Damping in the legs is achieved through the use of inner cylinders which dissipate energy through viscous shear. Energy is dissipated via squeeze film damping and relative sliding motion damping.
- Damping at the joints is achieved by a sliding bearing (PTFE) interface.
- Microslip at the joints does not affect the dimensional stability of the machine because the minimum energy form of the tetrahedron wants to be preserved. Unlike a plane joint which can continue to slip and lead to dimensional instability, the tetrahedron's legs' spherical ends want to stay on the spherical joint nodes.

The latter point has even more profound consequences, in that it makes the use of composite materials in the structure an attractive alternative to metals or ceramics. Wound carbon fiber tubes can be designed to have a zero coefficient of thermal expansion along their length and they can have stiffness-to-weight ratios two times higher than are obtainable with metals. It would be difficult to design a conventional machine tool that made economical use of the desirable properties of carbon fibers.

Counterweights and Counterbalances

Counterweights, shown in Figure 11.8.28 can be used to minimize gravity loads. This helps minimize the holding torque required of servomotors, which in turn minimizes motor size and heat input to the machine. For a very high precision machine, however, the bearings used to guide the counterweights and support the pulley bearings should have negligible static friction. Counterweights also increase the mass of the system, which can decrease dynamic performance. In the case of quasistatic axes (e.g., large gantry-type surface grinders), the counterweight can increase the load on the structure that supports the axes, and if the structure is cantilevered, then as the counterbalanced axis moves, the deflection errors

* This concept is protected by worldwide patients. See, for example, U.K. patent 8,719,169, or contact the British Technology Group, 101 Newington Causeway, London, SE1 6BU, England.

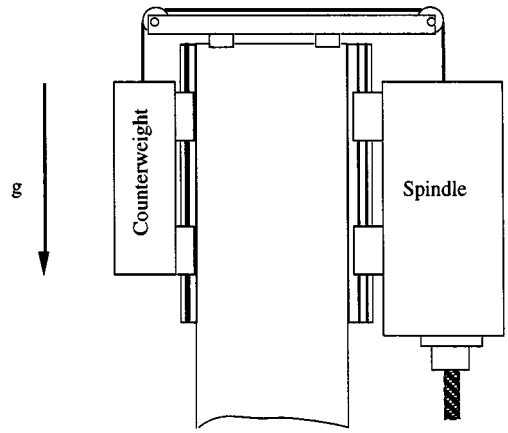


FIGURE 11.8.28 Use of a counterweight to balance the deadweight of an axis.

caused by the weight of the counterbalanced axis will change. Alternatively, a separate frame and low-accuracy slave carriage can be used to support the counterweights.

A counterbalance can be any passive means used to support the weight of an axis that moves in a vertical direction. Pistons have been used successfully but they can impart frictional and misalignment forces. A very effective counterbalance method is to use a float. This also provides viscous damping if the fluid used is a viscous oil.

Compensating Curvatures

All structures have finite stiffness, and when loads are applied, lateral and angular displacements will result. To compensate for these effects, it is possible to shape the otherwise straight ways of an axis so that the sum of the deflections and the intentional nonstraightness results in minimal net straightness error. This type of correction is known as a *compensating curvature*. Usually, it is very difficult to also compensate for angular errors. If the error budget is properly assembled, it can be used to plot the total error as a function of the position of an axis. Once the plot is found, it can be used to help design a shape (ideally, the inverse of the plot) that will cancel the lateral and hopefully also angular errors for the given bearing design. Compensating curvatures can easily be measured using an autocollimator.

When the load does not greatly change on parts moving along axes that are curvature compensated (e.g., axes that carry a measuring head), the method can be effective. If the loads do change greatly (e.g., a table carrying different weight parts and fixtures), the compensating curvature can sometimes decrease performance. One of its main attractions, however, is that once the correct compensated curvature and manufacturing process is found for a particular machine, it can be the least expensive way to correct for straightness errors.

With modern mapping techniques, compensating curvatures are used primarily on very large structures or when angular errors caused by deflection are too large and cannot be corrected for by another axis. For example, they might be used in a situation where otherwise the toolpoint would enter the workpiece at an angle rather than being perpendicular. A rotary axis would be required to compensate for this type of angular error, as opposed to just another linear axis, which can only be used to compensate for an Abbe error.

Structural Connectivity

The principle of kinematic design states that point contact should be established at the minimum number of points required to constrain a body in the desired position and orientation (i.e., six minus the number of desired degrees of freedom). This prevents overconstraint, and thus an “exact” mathematically continuous model of the system can be made. Kinematic locating mechanisms range from simple pins to Gothic arch-grooved three-ball couplings shown in Figure 11.8.29. Kinematic designs, however, are

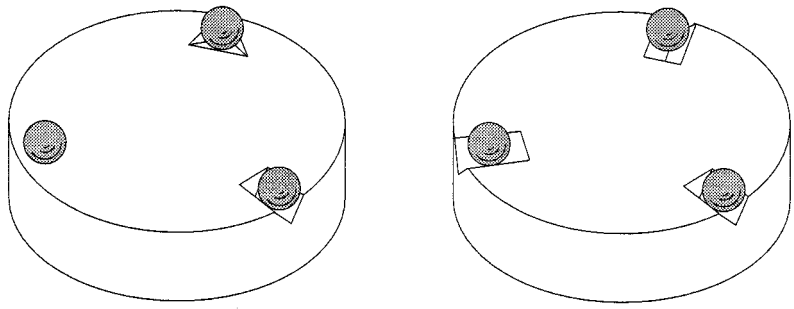


FIGURE 11.8.29 Flat-groove-tetrahendron (Kelvin clamp) and three-vee-groove kinematic couplings. In both cases, the balls are mounted to an upper plate (not shown) which is held in position with respect to the lower plate by the coupling.

subject to high-contact stresses, which often may require the use of ceramic components (e.g., silicon nitride balls and grooves) if the highest level of performance is to be achieved. If stress and corrosion fatigue are controlled, repeatability of a kinematic system can be on the order of the surface finish of the points in contact if the loads are repeatable or the preload high enough. Finite contact areas do exist, and they effectively elastically average out the local errors due to surface roughness. In addition, note that friction and microindentation will limit the accuracy of the kinematic model.

Kinematic support of a structure is often desired to ensure that the structure is not deformed by inaccuracies or instabilities of the mounting surface. For a small instrument, only one of the kinematic mounting points may be rigidly connected and the other two may consist of flexures that will allow for differential thermal growth between the instrument and the mounting surface. Friction does exist in kinematic couplings, and thus forces can be transmitted between a mounting surface and an instrument.

The principle of *elastic averaging* states that to accurately locate two surfaces and support a large load, there should be a very large number of contact points spread out over a broad region. Examples include curvic or Hirth couplings, which use meshed gear teeth (of different forms, respectively) to form a coupling. The teeth are clamped together with a very large preload. This mechanism is commonly used for indexing tables and indexing tool turrets. Figure 11.8.30 shows an indexing and clamping mechanism for a lathe tool turret. Note that many different types of clamping preload systems exist. However, this type of mechanism causes the system to be overconstrained; on the other hand, if an elastically averaged system is properly designed, fabricated, and preloaded, the average contact stress will be low, high points will wear themselves in with use, and errors will be averaged out by elastic deformation. The system itself will then have very high load capacity and stiffness. For a worn-in elastically averaged system, the repeatability is on the order of the accuracy of the manufacturing process used to make the parts divided by the square root of the number of contact points (i.e., teeth in a tooth coupling). Still, because of the large number of contact points, the chance of dirt contaminating the interface increases.

With respect to the connectivity between structural elements (e.g., a headstock and a bed), elements of a structure can either be connected together via:

- Kinematic design
  - Deterministic
  - Less reliance on manufacturing
  - Stiffness and load capacity limited
- Elastically averaged design
  - Nondeterministic
  - More reliance on manufacturing
  - Stiffness and load capacity not limited

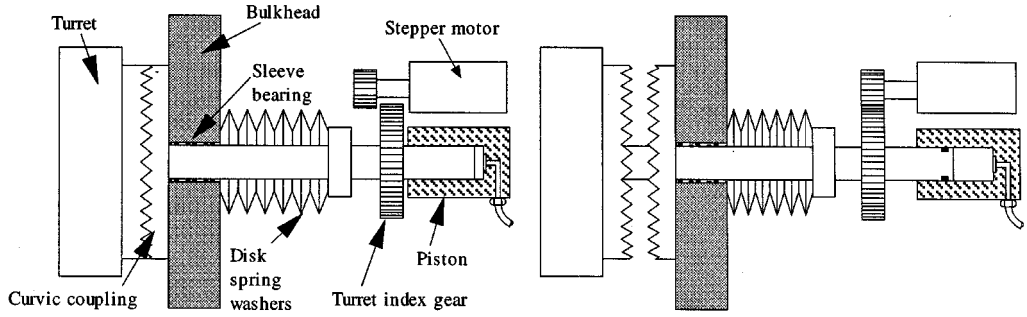


FIGURE 11.8.30 Typical turret indexing and locking mechanism in engaged and disengaged positions.

It should be noted that overconstrainted systems cannot accommodate differential thermal growth and hence are more prone to warping. Furthermore, finite deformation of the contact surface leads to micromechanical constraints that limit the true kinematicity of the structure. The use of hard materials (e.g., ceramics) helps to minimize the latter problem. For permanently connected components, one can align them using a kinematic location system and then inject epoxy or grout to create a bond between all the surfaces in close proximity, although with this method, one must make sure that the shrinkage of the bond material does not warp the components.

Bolted Joints*

Bolts can be used to prevent two parts from separating or sliding relative to one another. For the former, the tensile forces across the joint are transferred through the bolt shaft. For the latter, sliding motion is resisted by frictional forces generated from the normal load on the joint produced by tightening of the bolt and the coefficient of friction between the joint's parts. Because more than one bolt is usually used at a joint, it would be virtually impossible to ensure a tight fit of the bolt shafts in the holes, so it is not even worth trying. Sufficient lateral stiffness is usually provided by bolt preload and joint friction. For better resistance to shock loads, parts can be bolted in place and then holes drilled, reamed, and pinned with hardened steel dowels or roll pins. *In situ* drilling and reaming of the holes through both parts while they are bolted together maintains hole alignment, so multiple pins can be used.

Figure 11.8.31 illustrates the cross section of a typical portion of a bolted joint. A common bolt configuration used to bolt bearing rails for T-slides is shown in Figure 11.8.32. Many rails have a double row of bolts. In general, the cantilevered length should not exceed the bedded length. Ideally, the bedded length should be about 1.5 times the cantilevered length, but sadly this often takes up too much room. Furthermore, the stress cones under the boltheads should ideally overlap to maximize stiffness and minimize rail waviness. In practice, however, as shown in Figure 11.8.33 adequate stiffness can be obtained from a wider bolt spacing. The total stiffness for N segments of a bearing rail (e.g., N segments under a bearing pad) is given by**

$$K_{\text{Rail}} = \frac{N_{\text{Segments}}}{\dfrac{1}{K_{\text{Rail bend \& shear}}} + \dfrac{1}{K_{\text{Joint}} + \dfrac{1}{\dfrac{1}{K_{\text{Flange comp}}} + \dfrac{1}{K_{\text{Flange shear}}} + \dfrac{1}{K_{\text{Bed Shear}}} + \dfrac{1}{K_{\text{Bolt}}}}}}$$

* For a more detailed discussion, see, for example, Bickford, J.H. 1981. *An Introduction to the Design and Behavior of Bolted Joints.* Marcel Dekker, New York; as well as A. Blake's book.

** The theory and practical implementation implications of bolted joints for machine tool applications are beyond the scope of this work. These are discussed in greater detail in Slocum, A. 1997. *Precision Machine Design.* Society of Manufacturing Engineers, Dearborn, MI.

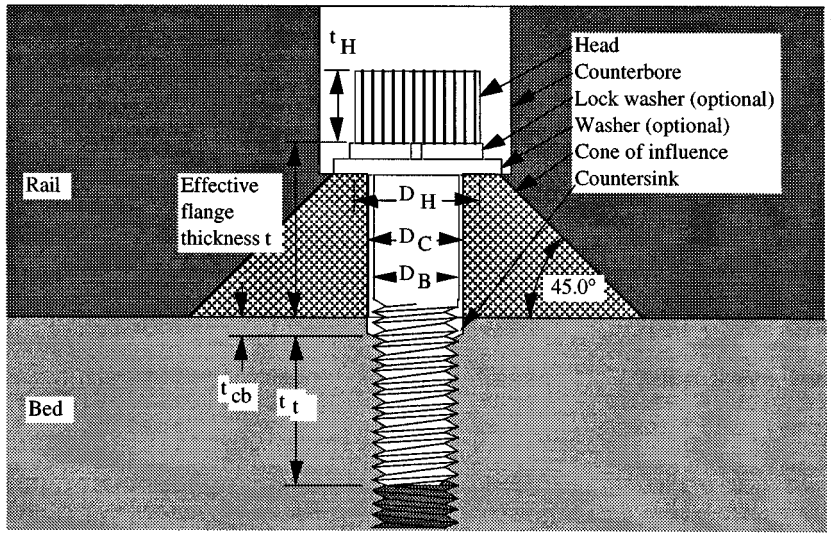


FIGURE 11.8.31 Typical components of a bolted assembly.

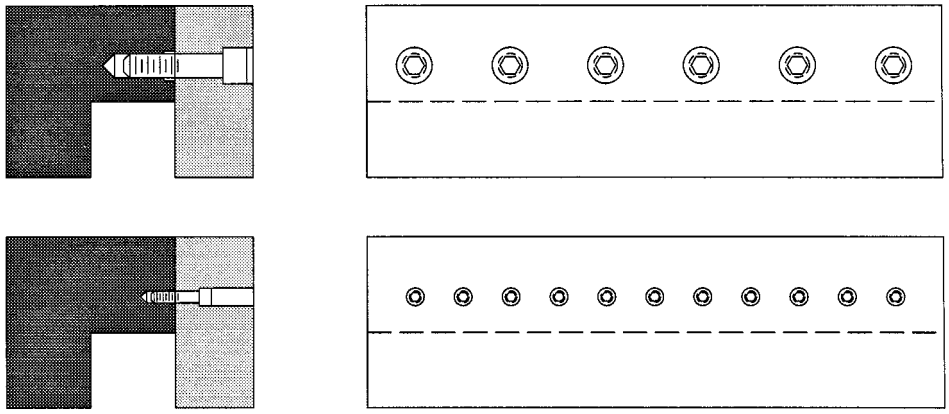


FIGURE 11.8.32 Counterbored and bolted bearing rails with equal stiffness.

Bolt diameter is not a sensitive parameter for stiffness when bolt spacing is made a function of the bolt diameter. When the length of the bolt and the cone are expressed as bolt lengths, bolt joint stiffness becomes linearly dependent on the bolt diameter. As a result, bolt diameter cancels out as shown in Figure 11.8.33.

Bearings

Since bearings are such a critical element in precision machines, one must *think* about the seemingly innumerable number of bearing design considerations that affect the performance of a machine, including:

- Speed and acceleration limits
- Range of motion
- Applied loads
- Accuracy
- Repeatability
- Sealability
- Size and configuration
- Weight
- Support equipment
- Maintenance

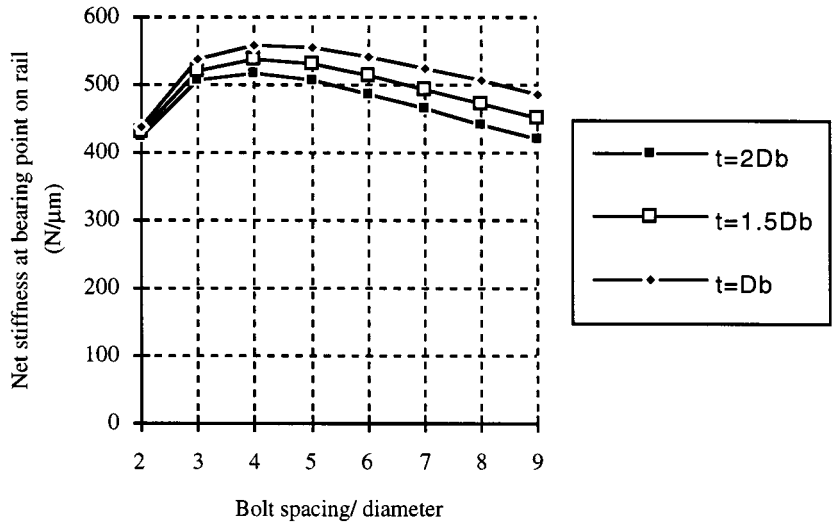


FIGURE 11.8.33 Relationships among stiffness and bolt spacing.

- Resolution
- Preload
- Stiffness
- Vibration and shock resistance
- Damping capability
- Friction
- Thermal performance
- Environmental sensitivity
- Material compatibility
- Mounting requirements
- Required life
- Availability
- Designability
- Manufacturability
- Cost

Because there are always design trade-offs in choosing a bearing for a precision machine, all of these factors must be considered simultaneously by the design engineer. In addition, there are several issues that are common to most types of bearings. These issues include surface roughness, preload, and replication of bearings in place.

Surface Roughness

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

- Sliding contact bearings tend to average out surface finish errors and wear less when the skewness is negative. A positive skewness (defined below) can lead to continued wear of the bearing.
- For rolling element bearings, if the contact area is larger than the typical peak-to-valley spacing, an elastic averaging effect will occur and a kinematic arrangement of rollers will produce smooth motion. If this condition is not met, the effect will be like driving on a cobblestone street. If numerous rolling elements are used, the effects of elastic averaging can help to smooth out the motion.
- Hydrostatic and aerostatic bearings are insensitive to surface finish effects when they are considerably less than the bearing clearance.
- Flexural and magnetic bearings are not sensitive surface finish.

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

$$R_q = \sqrt{\frac{1}{L}\int_0^L y^2(x)\,dx} \qquad R_a = \frac{1}{L}\int_0^L y(x)\,dx \qquad R_z = \frac{\sum_{i=1}^{5} y_{\text{peak}}(i) - y_{\text{valley}}(i)}{5} \tag{11.8.29}$$

Unfortunately, these measures do not provide any information as to the topographical characteristics of the surface. As shown in Figure 11.8.34 the surface topography can be characterized by the *skewness*. The skewness is the ratio of the third moment of the amplitude distribution and the standard deviation σ from the mean line drawn through the surface roughness measurements. Hence the skewness provides a measure of the shape of the amplitude distribution curve. For a contact-type bearing, valleys separated by wide flat planes may be acceptable. This form would have a negative skewness value and is typically in the range –1.6 to –2.0 for bearing surfaces. Sharp spikes would soon grind off, creating wear debris and more damage, and hence positive skewness values are unacceptable for contact-type bearing surfaces. The skewness is defined mathematically from:

$$\text{skew} = \frac{\mu_3}{\sqrt{\mu_2}} = \frac{\mu_3}{\sigma} \qquad \mu_n = \int_{-\infty}^{\infty} (y-\mu)^n f(y)\,dy \qquad \mu = \int_{-\infty}^{\infty} y f(y)\,dy \tag{11.8.30}$$

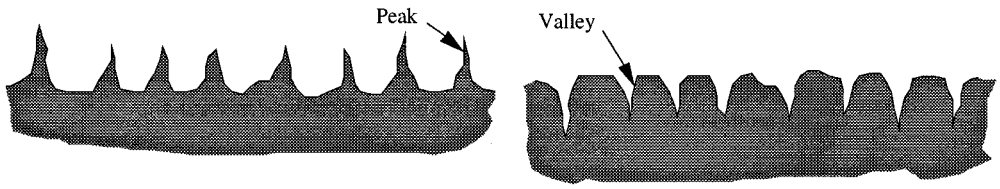


FIGURE 11.8.34 Surfaces with positive (left) and negative (right) skewness.

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

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

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

Bearing Preload

A preloaded bearing is one where one bearing pushes against another, thereby squeezing the bearing rail. This allows the bearing to resist bi-directional loads without motion nonlinearity (e.g., backlash) when the load reverses direction. As the force on a pair of opposing pads preloaded against each other is applied, one bearing pushes harder on the rail by an amount equal to the product of the pad stiffness and the carriage deflection, and one pushes less by the same amount.

Insufficient preload is synonymous with at least some of the bearing points periodically losing contact with the bearing surface. This results in difficult-to-map error motions and decreased stiffness, possibly leading to chatter of the tool. Tool chatter, in turn, degrades the surface finish of the part. However, with contact-type bearings, preload can accelerate wear and can lead to stiction, which decreases controllability. These issues lead to the desirability of externally pressurized fluid film bearings (air or fluid) where preload will not change with time.

Unless the axes of the machine are designed to utilize the weight of the machine itself as a preload, in many cases the preload will change with time. This is particularly true if contact-type bearing surfaces wear and if the structure relaxes, due to internal stress needed to maintain the preload. In an effort to counter these effects, the classic method has been to use a device known as a gib to generate the preload. In general, a gib can be defined as a mechanical device that can apply preload to a bearing. In the classical sense, a gib was a wedge-shaped part that was advanced by the action of a screw. When a gibbed machine wears, the gibs, bearings, and sometimes the ways must be rehand-finished and the gibs adjusted to provide the proper preload.

In many machines it is common to use modular rolling bearing components whose preload is set by using oversized balls or rollers, or by tightening bolts that push on a plate contacting the rollers. When the latter type of system wears, it is often discarded and replaced with a new unit. Hence large economies of scale can be obtained by some bearing component manufacturers.

Replicated Bearings*

Replication is a process whereby a very low shrinkage polymer is poured or injected around a master shape that has been coated with mold release. When the polymer cures, it has the shape of the master, which can then be removed and used again. This process can be used to form a bearing surface itself, such as a sliding bearing or a hydrostatic bearing with a pocket, or to form a mating surface between a modular bearing carriage and rough surface of a structure such as a casting.

Because hardening of many polymer resins is an exothrmic reaction, one of the important aspects of this process is to minimize the amount of polymer used, and to maximize the stiffness and thermal diffusivity of the master and the part. This is required to prevent the heat of polymerization from heating the structure, deforming it, and then the polymer hardening to the thermally deformed shape. Unlike bolted assemblies, once the polymer cures to form the desired shape (e.g., a trough in which a linear bearing rail is placed), alignment is no longer possible.

Figure 11.8.35 shows schematic details of surface preparation required to ensure the debonding does not occur. This sawtooth pattern can be obtained in a number of different ways. In addition to the rough surface finish required, the surface must also be thoroughly cleaned and a mold release applied to the master.

Sliding Contact Bearings

Sliding contact bearings are the oldest, simplest, least expensive bearing technology, and they still have a wide range of applications, from construction machinery to machines with atomic resolution. Sliding bearings are thus a very important element in the machine design engineer's tool kit. Sliding contact bearings utilize a variety of different types of lubricants between various interface materials. Lubricants

* See, for example, Devitt, A. October 1989. Replication techniques for machine tool assembly. In *East Manuf. Tech. Conf.* Springfield, MA. Available from Devitt Machinery Co., Twin Oaks Center, Suite G, 4009 Market Street, Aston, PA 19104.

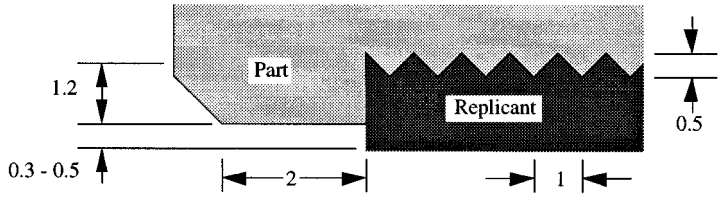


FIGURE 11.8.35 Details of surface and region for replication (dimensions are in mm).

range from light oil to grease to a solid lubricant such as graphite or a PTFE polymer. Because they often distribute loads over a large area, contact stresses and space requirements are often low, while stiffness and damping are usually high. In this section, general properties of sliding contact bearings are discussed followed by a discussion of design considerations.

There are numerous types of sliding contact bearings available and in the context of discussing their general properties, some specific categories will be discussed. In general, one should note that all sliding contact bearings have a static coefficient of friction that is greater (to some degree, no matter how small) than the dynamic coefficient of friction (static μ > dynamic μ). The difference between the static and dynamic friction coefficients will depend on the materials, surface finish, and lubricant.

Rolling Element Bearings

Figures 11.8.36 and 11.8.37 show representative types of ball and roller bearings. Typical linear rolling element bearing configurations are shown in Figure 11.8.38. Note that, in general, it is easier to make a ball spherical than a roller cylindrical; hence ball bearings are more commonly used in precision machines than are roller bearings, the exception being when very high loads must be withstood. A good roller bearing may be better than a moderate ball bearing, and hence in the end the most important thing is to compare manufacturers' specifications.

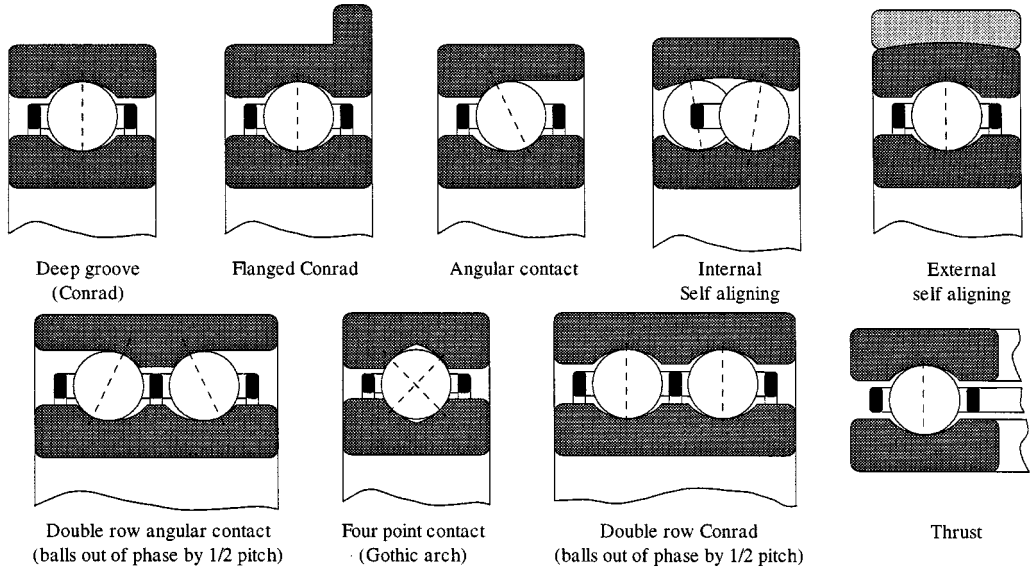


FIGURE 11.8.36 Typical ball bearing configurations for rotary motion bearings.

For linear motion applications, it is more difficult to maintain quality control of a curved surface a ball rides on than a planar surface a roller rides on because the latter can be self-checking. Note that machine-made linear bearing rails will have the same errors as the linear bearings on the machine that

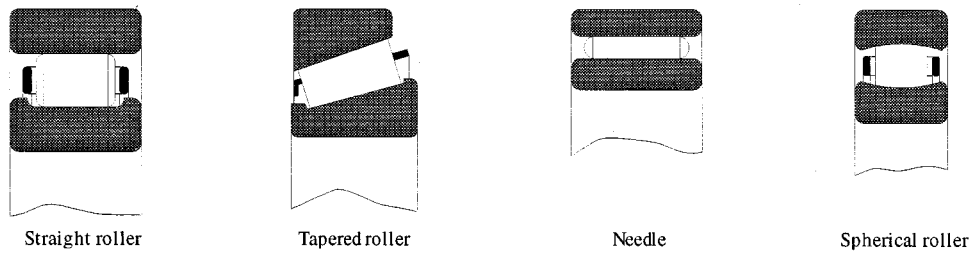


FIGURE 11.8.37 Typical roller bearing configurations for rotary motion bearings.

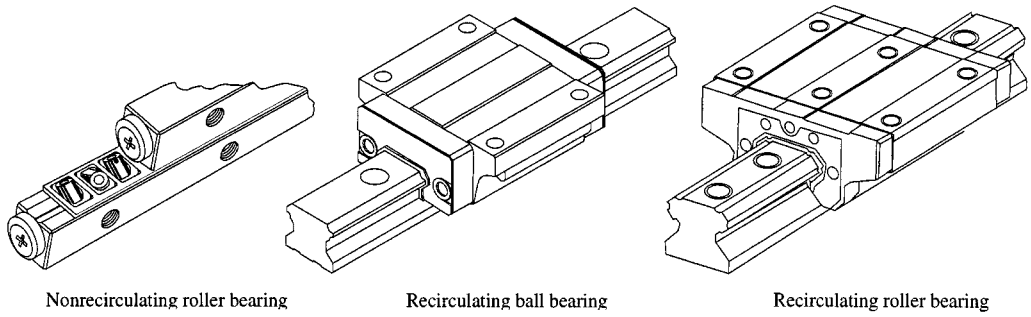


FIGURE 11.8.38 Typical linear rolling element bearing configurations.

was used to make them. For rotary axes, the raceway and the grinding tool can both be rotating at different speeds, which, combined with the random nature of precision spindle radial error motions, means that the form of the raceway can be uniform around is circumference. Thus rotary motion rolling element bearings can be made more accurate than linear motion rolling element bearings. Typical total radial error motion on a precision spindle is on the order of $^1/_4$ to 1 μm. For higher-performance linear or rotary motion bearings, one would typically move into the realm of aerostatic, hydrostatic, or magnetic bearings.

The design and production of a rolling element bearing requires careful analysis, materials selection, manufacturing quality control, and testing. Few companies other than bearing manufacturers have the resources for this type of effort. Whenever possible, one should use off-the-shelf bearing components. In addition, whenever possible, one should use modular components such as spindles and linear axes, particularly for nonsubmicron machines made in small lots (less than about 10 to 20 machines). The savings in design time, prototype testing, spare parts inventory, and repair and replacement costs often far outweigh the potential of saving a few dollars in manufacturing costs.

Designing with rolling element bearings can be intimidating because there are so many subtle details that can ruin a design if they are not considered. The best way to learn about how to handle these details is to think carefully about the physics of their design and the application and/or to work with others with experience, and if possible to experiment. In addition, manufacturers of precision bearing components are also usually willing to work with design engineers to integrate their bearing components into a design. Mapping and metrology frames can be used to increase accuracy given adequate repeatability, resolution, and controllability of the machine *if* the machine is designed with these error-reducing methods in mind.

Rolling Element Linear Motion Bearings

After the rotary motion rolling element ball bearing, the linear motion rolling element bearing ranks on the list of major inventions. Although not found in consumer products with anywhere near the frequency

of rotary ball bearings, linear rolling element bearings are vital for the manufacturing industry. They are critical elements in the design of today's high performance machine tools and factory automation systems.

Flexural Bearings*

Sliding, rolling, and fluid film bearings all rely on some form of mechanical or fluid contact to maintain the distance between two objects while allowing for relative motion between them. Since no surface is perfect and no fluid system is free from dynamic or thermal effects, all these bearings have an inherent fundamental limit to their performance. *Flexural bearings* (also called *fllexure pivots*), on the other hand, rely on the stretching of atomic bonds during elastic motion to attain smooth motion. Since there are millions of planes of atoms in a typical flexural bearing, an average effect is produced that allows flexural bearings to achieve atomically smooth motion. For example, flexural bearings allow the tip of a scanning tunneling microscope to scan the surface of a sample with subatomic resolution.** There are two categories of flexural bearings, monolithic and clamped-flat-spring, as shown in Figures 11.8.39 and 11.8.40.

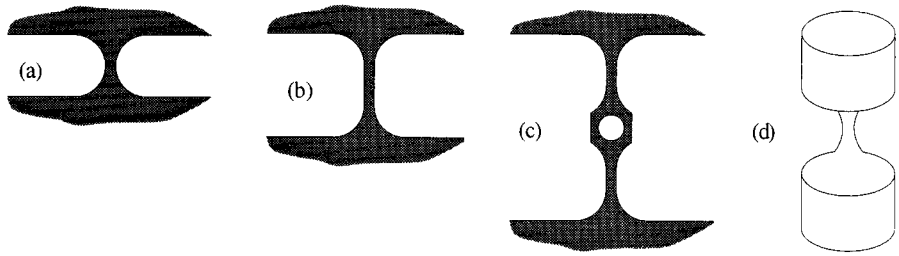


FIGURE 11.8.39 Monolithic flexural bearings.

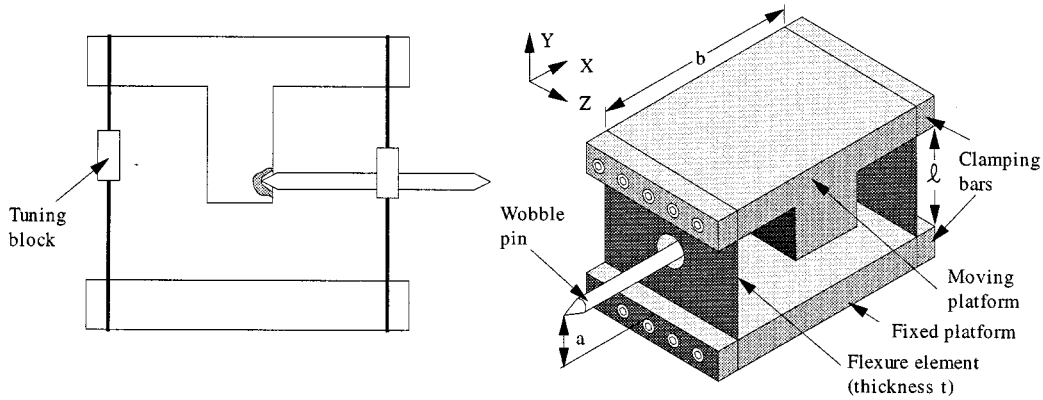


FIGURE 11.8.40 Clamped (blade) flexure.

* The reader may wish to consult the following references: Eastman, F.S. Nov. 1935. Flexure Pivots to Replace Knife Edges and Ball Bearings. *Univ. Wash. Eng. Exp. Sta. Bull.* No. 86; Eastman, F.S. Nov. 1937. The Design of Flexure Pivots. *J. Aerosp. Sci.* 5, 16–21; Jones, R.V. 1951. Parallel and Rectilinear Spring Movements. *J. Sci. Instrum.* 28, 38; Jones, R.V. and Young, I.R. 1956. Some Parasitic Deflections in Parallel Spring Movements. *J. Sci. Instrum.* 33, 11; Siddall, G.J. Sept. 1970. The Design and Performance of Flexure Pivots for Instruments. M.Sc. thesis, University of Aberdeen, Scotland, Department of Natural Philosophy.

** See, for example, Binnig, G. and Rohrer, H. 1982. Scanning Electron Microscopy, *Helv. Phys. Acta.* 55, 726–735.

Hydrostatic Bearings

Hydrostatic bearings utilize a thin film of high-pressure oil to support a load. In general, bearing gaps can be rather large, on the order of 5 to 100 μm. There are five basic types of hydrostatic bearings: single pad, opposed pad, journal, rotary thrust, and conical journal/thrust bearings. All operate on the principle of supporting a load on a thin film of high-pressure oil that flows continuously out of the bearing; hence a method is needed for supplying the pressurized oil and collecting and recirculating the oil that flows out of the bearing.

11.9 Robotics

Leonard D. Albano

The field of robotics is concerned with the design and development of mechanical devices that can be programmed to perform certain functions. Robots have been developed for a large variety of applications, such as material transport, automated assembly, and operations in controlled environments, such as high temperature and caustic surroundings. The semiconductor industry, for example, uses advanced robotic technology to fabricate IC chips. Many robots are simply relied on as a cost-effective alternative to human workers for certain highly repetitive tasks, such as spot welding and painting. For these applications, the robot must be able to undergo a range of accelerated motions while demonstrating the ability to position accurately its end effector. This must be done with minimum breakdowns and rapid repair. The control strategy involves solving the governing dynamic equations as the robot arm and end effector move through their operations. See Chapter 14 for further information.

11.10 Computer-Based Tools for Design Optimization

The range of computer applications in engineering design covers procedures from preliminary conceptual design to the production of manufacturing drawings and specifications (see Chapter 13). Most computer applications intended for production use can be classified into five or more major categories: analysis, computer-aided drafting and design, geometric modeling, data base management systems, and artificial intelligence. Traditional software for design optimization may be categorized as analytical applications, based on rational principles of mathematics and linear programming. Emerging computer-based tools for design optimization are an offshoot of research in artificial intelligence, capable of processing a variety of algorithmic, symbolic, deterministic, probabilistic, and fuzzy knowledge.

Design Optimization with Genetic Algorithms

Mark Jakiela

Introduction

One broad use of computer-based tools in design is to automate the design process itself. This is commonly done by modeling a design problem (or class of design problems) as a search problem. Such a model requires a representation and a search process. A representation is some set of parameters, either continuous or discrete, which can represent every possible design with appropriate values assigned to each parameter. We can call each possible parameter setting a design, and the set of all possible designs (possibly infinite in number) is called the design space. The search process is some method to investigate the design space and find some design that is acceptable or desired. When some acceptable designs are more desirable than others, the search process can be used to perform optimization.

Traditionally, the search and optimization technique that has been used most frequently for mechanical engineering problems has been a gradient-based search with continuous variables. When this representation and search method agrees with the problem being solved (the optimization model functions should be smooth and the objective preferably unimodal), this technique works exceedingly well. Currently, however, a variety of important problems are not amenable to this well-established technique because they cannot be modeled with only continuous variables (discrete or mixed discrete problems) and/or the objective and constraint functions do not exhibit appropriate characteristics. Other techniques are therefore required.

This subsection will provide an introduction to genetic algorithms ("GAs"), a search and optimization process that mimics the natural process of evolution. First, a brief tutorial on the fundamentals of GAs is given. This is followed by several examples drawn from a range of engineering problem domains. We hope that the variety of examples demonstrates the real strength and advantage of genetic algorithms: their versatility. They can be used to address almost any type of optimization problem. Requirements for their use are discussed in the tutorial. Some general and philosophical remarks conclude the subsection.

Genetic Algorithm Tutorial

Genetic algorithms are a simulation of a simplified process of evolution (Holland, 1975). The fundamental object of data used by a GA is referred to as a chromosome. Some number of chromosomes exist simultaneously in a population. The GA transforms one population into a succeeding population, creating a generation of children from a generation of parents. It does this with operators that are analogous to the biological operators of reproduction, recombination, and mutation. A measure of the quality of each chromosome, analogous to the fitness of a biological organism in an environment, is created and used to probabilistically select chromosomes that will serve as parents. With an increasing number of generations, the overall (e.g., average, single best) fitness of the population increases and the diversity represented in the population decreases.

For engineering purposes (see e.g., Goldberg, 1989), the chromosome is some type of encoding of the characteristics of a possible design. There is the notion, therefore, of a genotype (chromosome space) to phenotype (design space) mapping. Consider, as a simple example, a GA that is used to evolve optimal values of a set of parameters. A chromosome might be set up as a string of binary digits (i.e., bits), with concatenated substrings being binary encodings for each parameter. This is shown in Figure 11.10.1. Each bit position can be thought of as a gene and the value found there can be though of as an allele. The number of bits used to encode the parameter defines a resolution of the representation accuracy.

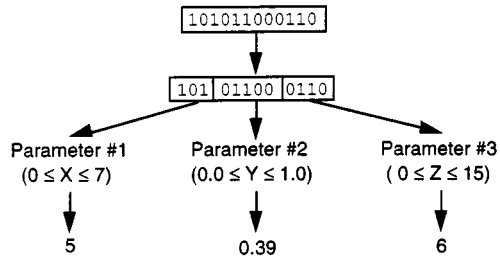


FIGURE 11.10.1 Binary chromosome representation.

Continuing with the explanatory parameter optimization example, the various processes of one iteration (i.e., generation) would proceed as follows. In a given generation, all chromosomes would be decoded to yield possible values for the parameters. These parameters would be used in an objective function of interest to yield a value measuring the quality of that set of parameters. This objective function value is used as the fitness measure of the chromosome. The probability of choosing a chromosome to serve as a parent is influenced by its fitness value: as in nature, the fittest are more likely to reproduce. Pairs of parents are randomly chosen with these weighting probabilities and mated. Copies of the parent chromosomes (i.e., reproduction) are used to build the chromosomes of the children. This is done by performing some type of recombination operation on them.

Figure 11.10.2 shows a simple one-point crossover recombination operation. First, a crossover point is randomly selected along the length of the chromosomes. Then, the front of one chromosome is appended to the back of the other chromosome and vice versa. This creates two children from two parents. Performing crossover with all the pairs of parents will create a tentative population of children that can replace the parents. The final step is to perform mutation on these children with a very low probability. One bit per 1000, for example, can be randomly chosen and inverted in value as an aid to maintaining diversity in the population. Finally, replacing the parents with the children brings the algorithm to the next generation.

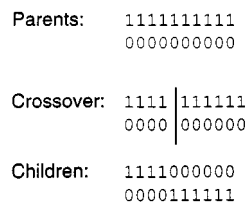


FIGURE 11.10.2 One-point crossover.

The procedure described above is commonly referred to as a "simple" GA (Goldberg, 1989) and can be thought of as a basic or canonical approach. In practice, several runtime parameters are required in its implementation. These include the probability of crossover (will two parents recombine, or will they proceed into the next generations unaltered?), the probability of mutation (on average, what fraction of genes are mutated?), a fitness scaling coefficient (an adjustment limiting the range of fitness values that

helps to prevent premature convergence), and population size (the number of chromosomes in a population). Note that a simple GA has the disadvantage of requiring a large number of function calls. This can be mitigated somewhat by using an "overlapping" population, in which a smaller number of parents will be chosen (i.e., fewer of the best) and the resulting children will replace the worst members of the parents' population. Regardless of the amount of overlap, note that a GA requires no derivative computations; the search is directed by the zeroth-order sampling of fitness function values. Because of this, a GA can perform well in ill-behaved multimodal search spaces (see Goldberg, 1989) and is relatively unlikely to become trapped in local suboptima. Equally important is the high degree of general flexibility of a chromosome representation. Although a binary encoding makes clear the general operation mechanisms and actually is used for a variety of problems, several other types of representations have been used and found to work well. In almost all cases, the evolutionary optimization strategy succeeds in producing improved designs. A variety of representations and fitness/objective functions will be described in the following examples.

Examples

Truss Parameter Selection. As a first example, we describe a problem that could readily be formulated as a continuous optimization problem (see Arora, 1989, pp. 23–31). The following genetic algorithm formulation, condensed from Wallace et al. (1995), will highlight the different approaches that are needed for a GA-based optimization.

Consider the symmetric truss shown in Figure 11.10.3. The truss is to be made of medium strength structural steel. We will assume that the load W and the truss height h are specified. The designer wishes to determine the truss base s and the inside and outside diameters of the members 1 and 2. A possible list of design variables and associated upper and lower limits for this problem are shown in Table 11.10.1. Note that the variables r_1 and r_2 allow the cross sections to vary continuously from solid rod stock to thin-walled pipe.

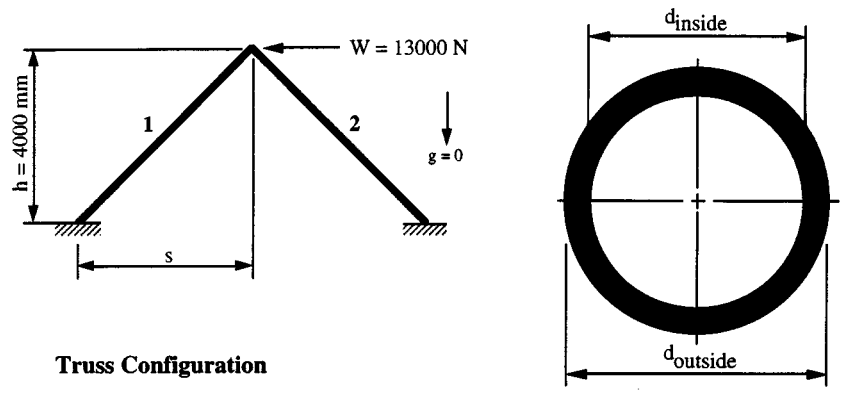


FIGURE 11.10.3 Definition of a simple two-member truss problem (gravity is assumed to be zero to simplify buckling calculations).

TABLE 11.10.1 Design Variables or Parameters to be Optimized in the Truss Problem

Design Variable	Limits
$d_{1(outside)}$	$0 < d_{1(outside)} < 100$ mm
$r_1 = d_{1(inside)}/d_{1(outside)}$	$0 < r_1 < 1$
$d_{2(outside)}$	$0 < d_{2(outside)} < 100$ mm
$r_2 = d_{2(inside)}/d_{2(outside)}$	$0 < r_2 < 1$
s	1000 mm < s < 3000 mm

In a continuous gradient-based formulation, the next step would be to model the relationships of interest as objective and constraint functions. Lagrange variables could then be introduced and the lagrangian would be subjected to the first- and second-order (constrained) optimality conditions to find a set of local optima. In addition to the upper and lower variable limits already listed, the other relationships of interest would include the normal stresses in members 1 and 2, the buckling stress in member 1, and the material cost as the objective function.

A genetic algorithm formulation does not use any derivatives. Instead, an objective function augmented with penalty terms for each potentially active constraint is used in an unconstrained optimization search. Wallace et al. (1995) formalize the notion of "design specifications" as a means to formulate such GA design optimization problems in a consistent and simplified way. The basic idea is that specifications indicate the probability that a particular performance level will be deemed acceptable. For the truss optimization the performance specifications considered are shown in Table 11.10.2. It is also possible to use other goal, objective, or preference formulations, such as the utility theory (Keeney and Raiffa, 1976).

The performance specifications on normal stress in the two members, $n_{\sigma 1}$ and $n_{\sigma 2}$, are modeled as safety factors. The specification distribution indicates that a safety factor for either member less than 1.5 is unacceptable (i.e., it will be accepted with a probability of zero, also see discussion below), and a safety factor greater than 2.0 is acceptable with certainty. To actually use this probabilistic specification, the design variables shown in Table 11.10.1 are decoded from a chromosome and used to compute values for $n_{\sigma 1}$ and $n_{\sigma 2}$, resulting in turn in two acceptance probabilities. A buckling safety factor for member 1 is treated in a similar manner (note that member 2 is always in tension). In addition to the upper and lower limits on r_1 and r_2, an additional performance specification is provided that indicates a strong preference for solid rod or thin-walled tube. This specification has, in effect, made the acceptable ranges of r_1 and r_2 discontinuous. Finally, the cost objective function is also provided to the problem as a probabilistic specification. If a full acceptance probability is achieved, the problem can be resolved with a lower material cost certainty point. In the solutions presented below, we assume that rod is 1/2 the price of pipe on a unit volume basis.

Once all individual acceptance probabilities are computed, a composite acceptance probability is used as the fitness function for the GA search. Intuitively, it seems as though the overall probability of acceptance is simply the product of the individual acceptance probabilities. Such a model does not work well in the GA-based search since a low score on a single characteristic will leave the overall design with a zero acceptance probability (in this case, the GA will consider "infeasible" designs with r_1 and/or r_2 between 0.2 and 0.7). The other characteristics of the designs could be very good and worthy of consideration in future generations. We therefore transform the multiplicative probability maximization into an additive information content minimization by taking the log of each probability term. The resulting objective is a generalized from of Suh's "Information Axiom" see Section 11.4.

$$I = \sum_{i=1}^{n} \log_2\left(\frac{1}{p_i}\right) \tag{11.10.1}$$

where

n = number of design specifications or criteria
p_i = probability of being acceptable as defined by the i^{th} specification*
I = design information content in bits

Results for the solved truss problem are shown in Figure 11.10.4. Member 1 is under compression and thus is tubular (for buckling stiffness), while the tensile member 2 is made of less expensive rod stock. Convergence to a solution typically occurred in 80 generations requiring 2.5 sec on a Silicon Graphics

* For small values of p_i, a minimum nonzero value is used in Equation 11.10.1 to prevent division by zero.

TABLE 11.10.2 Performance Specifications and Distributions for the Truss Problem

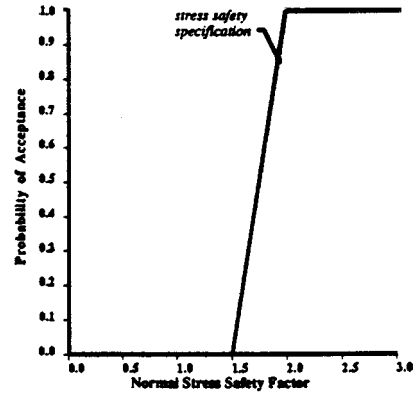
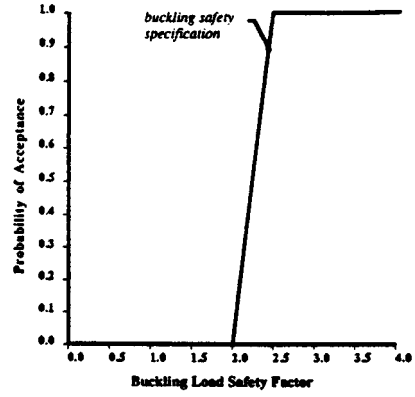
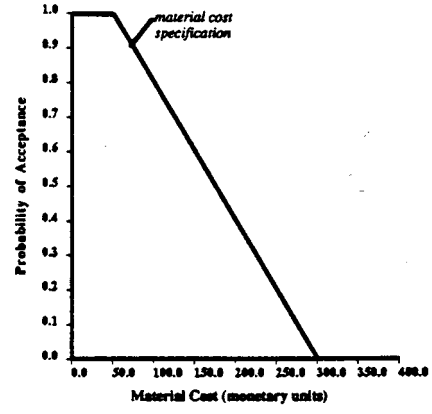
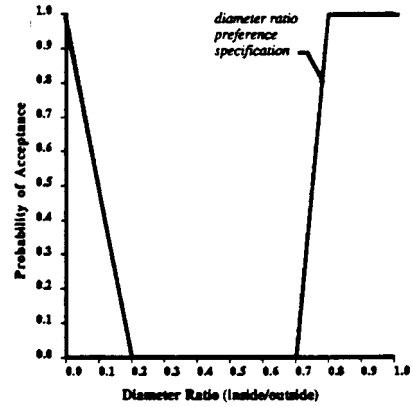
Performance Specifications	Specification Distribution
$n_{\sigma 1}$ (normal stress safety factor), member 1 $n_{\sigma 2}$ (normal stress safety factor), member 2	
$n_{\beta 1}$ (buckling safety), member 1	
C (cost)	

TABLE 11.10.2 Performance Specifications and Distributions for the Truss Problem (continued)

Performance Specifications	Specification Distribution
r_1 (diameter ratio), member 1	
r_2 (diameter ratio), member 2	

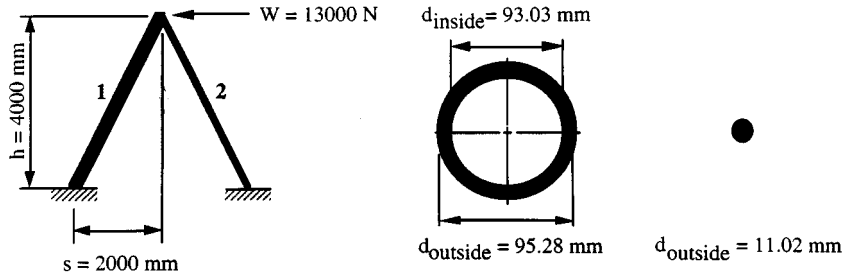


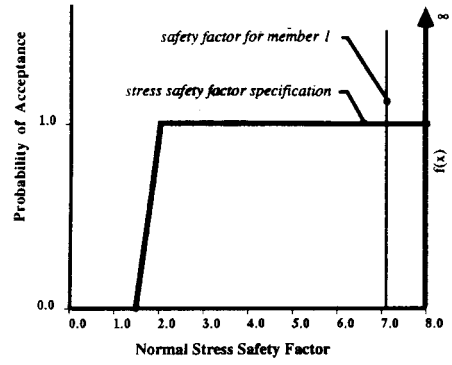
FIGURE 11.10.4 (a) Configuration of the optimized truss design.

Indigo[2] Extreme[2]. This search involved the evaluation of 2400 candidate designs (30 per generation). The search space for this problem contains $\approx 1.1 \times 10^{12}$ points (five parameters at 8-bit resolution gives 2^{40} points).

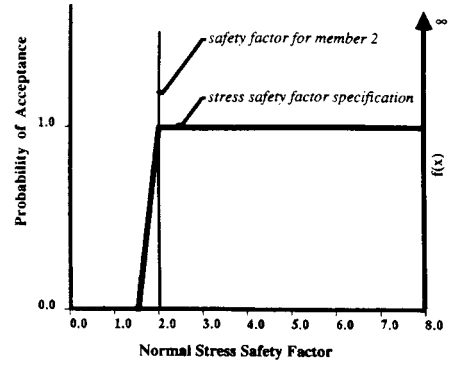
It must be noted that the search converged to this solution using rod stock for member 2 in only 17% of 60 consecutive optimizations. The more frequent solution shown in Figure 11.10.5 is nearly as good.

Topology Optimization. In this example, we will address planar stress/strain problems and the generation of optimal structural topologies. As shown in Figure 11.10.6, a design domain defines the allowable extents of any possible design. This domain is discretized into a rectangular grid, with each square element of the grid assigned a value of "material" or "void". Treating each grid element as a binary variable leads to a large combinatorial search space. The grid is used to create a finite element mesh for structural analysis. In our case, each material element contains four triangular finite elements made from the five nodes located at the four corners and the center of each material element. The genotype corresponding to this phenotype is a simple binary two-dimensional array chromosome, or a binary string, as described below.

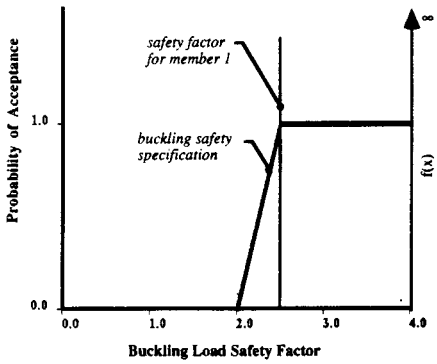
Phenotypically, certain material elements are required to always contain material, regardless of the bit value in the genome. We refer to these as material constraints. These are elements that serve as structural boundary conditions. In Figure 11.10.6, examples would be the highest and lowest elements bordering the wall, and the element that is used to apply the load. Thus, the genotype is not always a completely accurate representation of the actual structure.



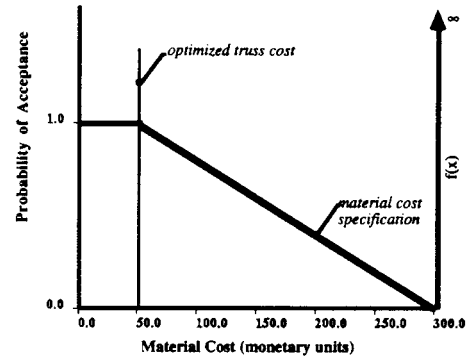
$n_{\sigma 1}$ = 7.09, $p_{acceptable}$= 1.0



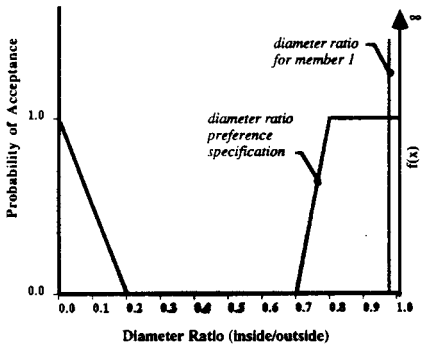
$n_{\sigma 2}$ = 2.03, $p_{acceptable}$= 1.0



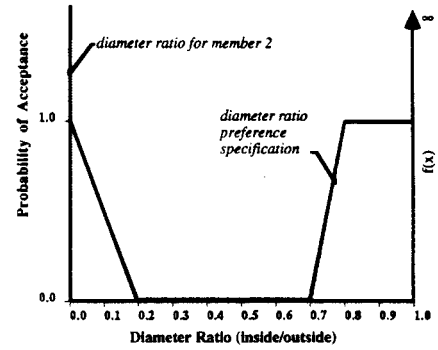
n_{Bl} = 2.50, $p_{acceptable}$= 1.0



C = 51.02 units, p = .9958



r_1 = 0.9764, $p_{acceptable}$ = 1.0



r_2 = 0.0, $p_{acceptable}$ = 1.0

FIGURE 11.10.4 (b) Comparison of truss performance variables to design specifications.

The genotype can also differ from the phenotype because of connectivity analysis. Connectivity analysis requires that all elements that are included in the structural analysis be connected to other elements by at least one edge–edge connection, as opposed to only a corner connection. In addition, all elements considered connected must be linked to one of the material constraint elements by a path of edge connections. Any elements not so connected are removed from the mesh for the purposes of structural analysis: they are considered to have no stiffness and no mass. Since we are addressing planar problems, the rationale for connectivity analysis is that elements connected at corners cannot transfer

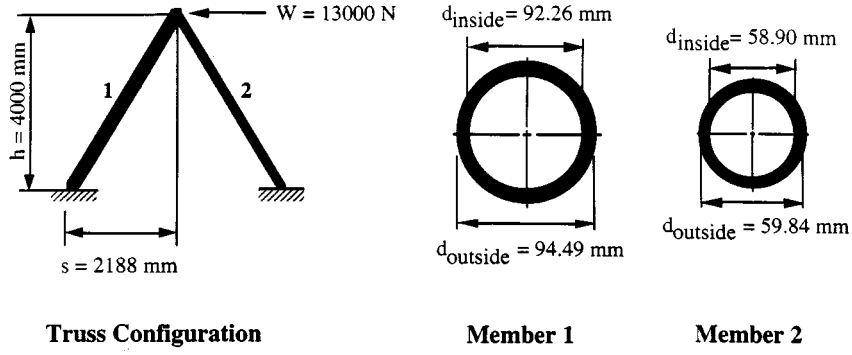


FIGURE 11.10.5 A result that did not find the better solution using cheaper rod stock for member 2. The cost is 56.73 units, giving $p_{acceptable}$ = 0.9728 (compared to 0.9958 for the solution using rod for member 2). $n_{\sigma 1}$ = 7.49, $n_{\sigma 2}$ = 2.01, $n_{\beta 1}$ = 2.50.

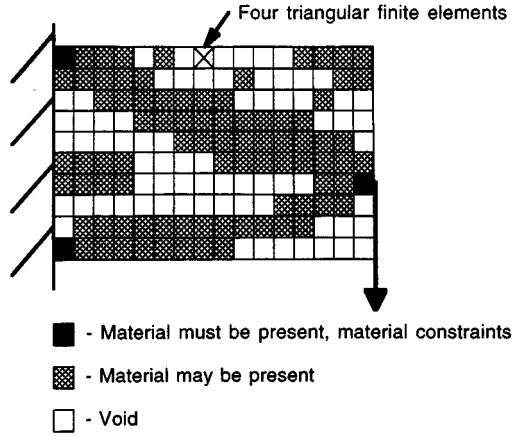


FIGURE 11.10.6 Typical design domain.

moments about those corners (since the finite element model treats corners as kinematic pin joints). It is less likely, therefore, that "disconnected" elements will contribute to enhanced structural performance. We have found that this is true empirically (see Chapman et al., 1993 for further discussion of this topic). Figure 11.10.7 shows the structure of Figure 11.10.6 after connectivity analysis. It is important to note that we do not penalize structures with more disconnected elements; we simply reward structures based upon the performance caused by their connected elements. This allows disconnected elements to in some sense be "recessive", in that crossover could cause them to combine with the (possibly disconnected) elements from another structure to yield a structure of connected elements of much improved performance.

As fitness, we are interested primarily in structural performance. In this study, this will be represented by two characteristics, mass and deflection. Generally, we wish the genetic algorithm to evolve structures that are both lightweight and deflect small amounts under a given load. The deflection, δ, will be measured with a finite element simulation and the mass m will be proportional to the number of connected material elements. These two characteristics can be used to create an unconstrained fitness, such as maximizing the following ratio:

$$\text{fitness} = \frac{1}{(\delta)(m)} \tag{11.10.2}$$

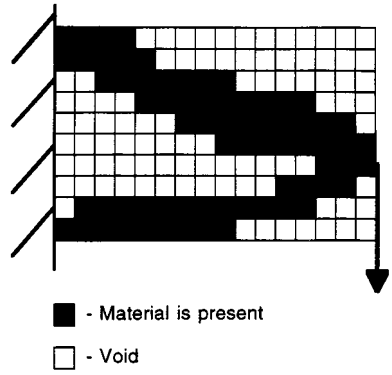


FIGURE 11.10.7 After connectivity analysis.

Alternatively, one of these characteristics can be optimized while treating the other as a constraint. Chapman and Jakiela (1995) provide some guidelines for using penalty functions for structural optimization that address many different characteristics (e.g., reducing the number of macroscopic holes in the topology).

In this example, we will use two different types of chromosomes to represent the material array. The first is a one-dimensional string chromosome made by concatenating the rows of the material array. The second is a two-dimensional binary array precisely matching the material array. For a two-dimensional array chromosome, a single-point crossover can be done as shown in Figure 11.10.8. A position is randomly selected along both dimensions, defining a point in the array. Complementary parts from both parents, diagonal about the point, contribute to the chromosomes of the children. Note that this approach can be extended to any dimension.

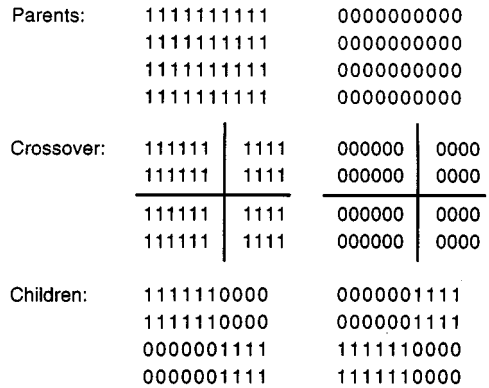


FIGURE 11.10.8 One-point crossover on 2D arrays.

Using the binary string chromosome, Figure 11.10.9 shows results for three different material discretizations. The loading and material constraints were as described in Figure 11.10.6. All three cases display a uniformly distributed porosity with an overall shape arising implicitly from the topological optimization.

Hierarchial Shape Packing. Effective material utilization is important to virtually all industries. A specific problem in this area that can be modeled as a combinatorial optimization problem is the arrangement of planar shapes to be cut out from a piece of material so as to minimize scrap. This problem has great economic significance in the garment, sheet metal, shipbuilding, and other industries. A common instance involves a “blank” of stock material of fixed width, such as would result from a roll of material.

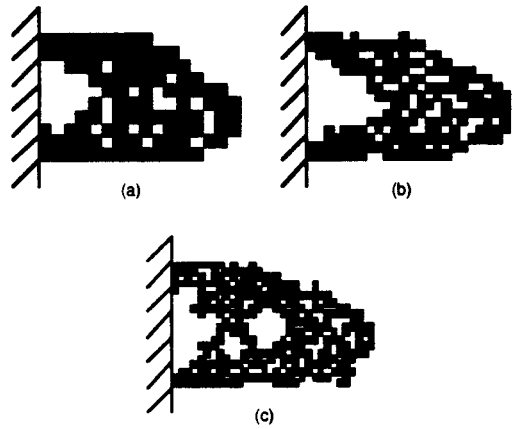


FIGURE 11.10.9 Results of (a) 10 × 16, (b) 15 × 24, and (c) 20 × 32 optimizations.

The part shapes must be arranged to minimize the length of material required from the roll. We have solved this problem with a two-level hierarchical genetic algorithm. A more complete description of this approach can be found in Dighe and Jakiela (1995a).

The approach of the lower level is shown in Figure 11.10.10. The blank of fixed width is considered analogous to a box into which the shapes will be packed. One by one, in a particular order, the shapes will be placed into the box so as to minimize the accrued height required. To pack a particular shape, a location and orientation is specified, as shown in Figure 11.10.10b, and the shape is "dropped" straight down until it contacts one of the previously packed shapes or the bottom of the box. The "dropping into a box" metaphor does not extend to include dynamic effects. Note, for example, that part D in Figure 11.10.10c would not rotate and slide to settle in a lower final position; it is in its final location for the values of x and θ chosen. Figure 11.10.10d shows a much better choice of parameters for x and θ. The lower level uses a genetic algorithm to optimize the values of x and θ and therefore is used each time a shape is packed.

The overall success of this packing strategy depends on the order in which the objects are placed. The higher-level GA searches the space of possible orders to find those that work well with the packing strategy of the lower-level GA. The chromosome representation for a packing order, and associated crossover and mutation operators are shown in Figure 11.10.11. The chromosome is simply a list of part labels, with the left-to-right order indicating the packing order. If we try to perform a single point crossover on these strings, Figure 11.10.11 shows that invalid offspring often result. Some labels are missing and others appear more than once. To remedy this problem, we use a simple order-based crossover as shown in Figure 11.10.11b. For positions to the left of the crossover point, each child receives the order from one of the parents. The order of the remaining parts is obtained from the order in which those remaining parts appear in the other parent's chromosome. Finally, mutation simply changes the location of a part label, usually by moving it to the left.

Note that this crossover operator tends to not disrupt the leftmost parts of the chromosomes. In particular, without mutation, the very leftmost position in any chromosome must have a label that was present in that position in the initial population. Hence, mutation is performed with a leftward bias.

Figure 11.10.12 shows some typical arrangements produced using this approach. Figures 11.10.12a and 11.10.12b show two configurations in the same optimization run. These ten objects were chosen such that their characteristic sizes varied by about a factor of 10. Figure 11.10.12 shows a jigsaw puzzle that we have used as a benchmark test for the packing system. Taking into account deliberate gaps between the parts, the density of the correct jigsaw arrangement is 0.92, not 1.00. Two arrangements from the same optimization run are shown in Figures 11.10.12d and 11.10.12e. It is clear that the simple "height-based" lower-level GA cannot match a human's visual processing and reasoning ability.

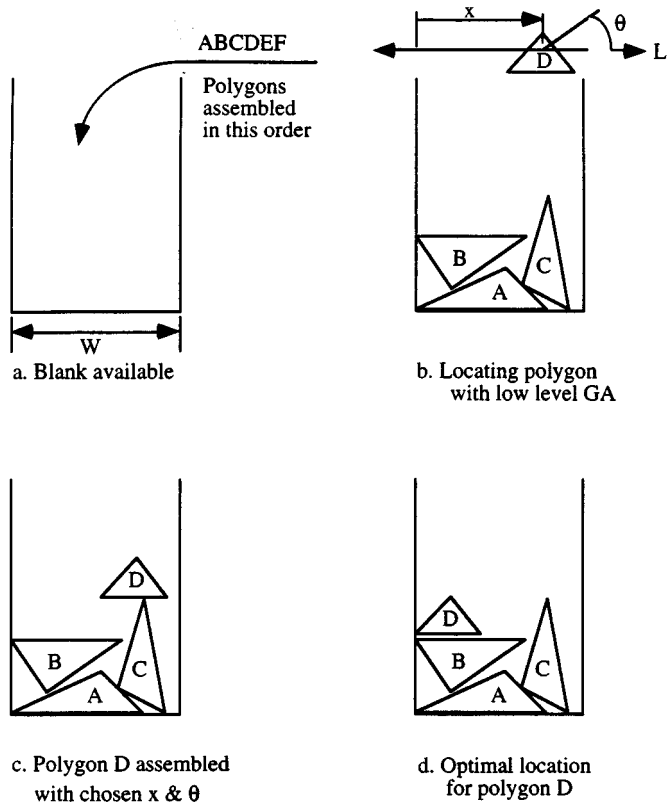


FIGURE 11.10.10 Lower-level GA for hierarchical shape packing.

Integrating a GA-based search with heuristics derived from observation of human experts is the subject of ongoing work in this area (see Dighe and Jakiela, 1995b).

References

Arora, J.S. 1989. *Introduction to Optimum Design.* McGraw-Hill, New York.

Chapman, C. and Jakiela, M. 1995. Genetic algorithm-based structural topology design with compliance and topology simplification considerations. *ASME J. Mech. Design.* to appear.

Chapman, C., Saitou, K., and Jakiela, M. 1993. Genetic algorithms as an approach to configuration and topology design. In *Proceedings of the ASME 19th Design Automation Conference: Advances in Design Automation,* Vol. 1. American Society of Mechanical Engineers, DE-Volume 65-1, New York, 485–498.

Dighe, R. and Jakiela, M.J. 1995a. Solving pattern nesting problems with genetic algorithms employing task decomposition and contact detection. In *Evolutionary Computation.* MIT Press. to appear.

Dighe, R. and Jakiela, M.J. 1995b. Automation of human patern nesting strategies. *ASME J. Mechan. Design.* submitted.

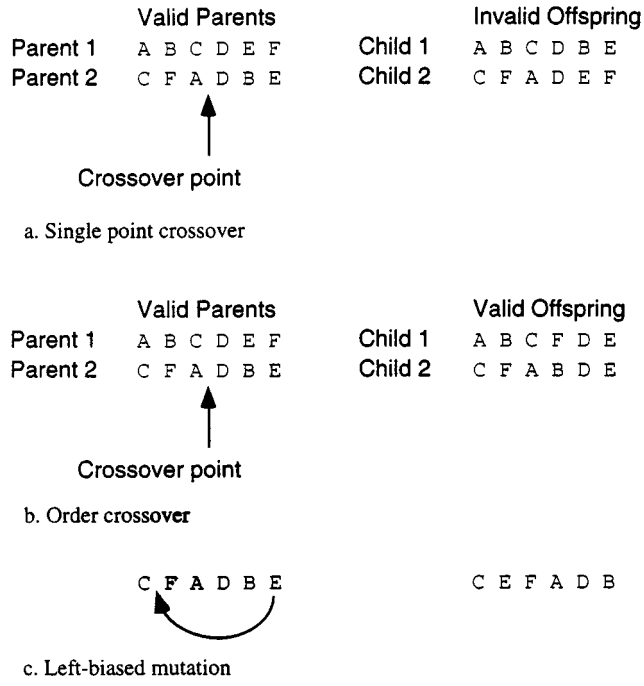


FIGURE 11.10.11 Crossover and mutation operators.

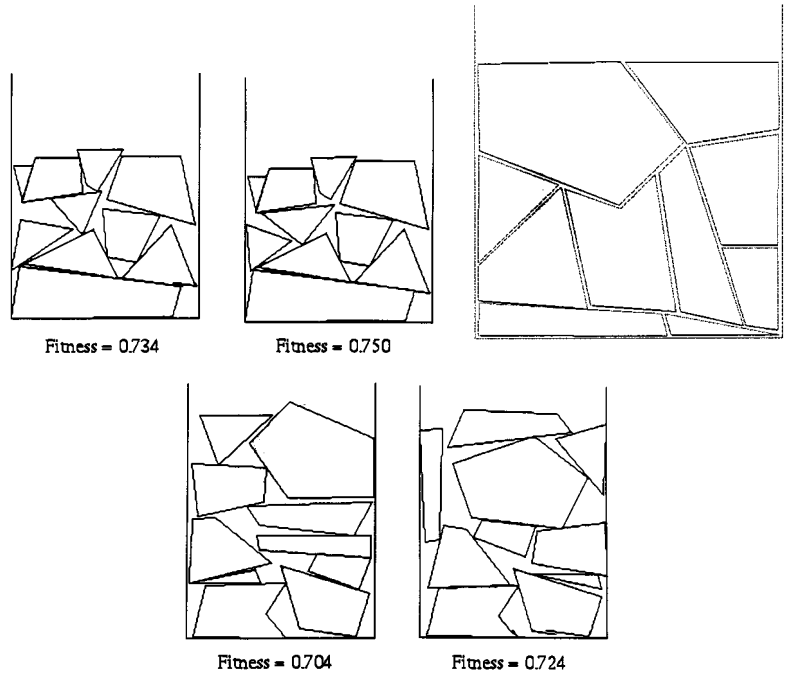


FIGURE 11.10.12 (a) Sample result from a GA run; (b) sample result from a GA run; (c) jigsaw puzzle attempted with sting-based representation; (d) layout of jigsaw puzzle using string-based representation and height-based fitness function; (e) layout of jigsaw puzzle using string-based representation and height-based fitness function.

Goldberg, D. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Reading, MA.

Holland, J. 1975. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor, MI.

Keeney, R.L. and Raiffa, H. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* John Wiley & Sons, New York.

Suh, N.P. 1990. *The Principles of Design.* Oxford University Press, New York.

Wallace, D.R., Jakiela, M.J., and Flowers, W.C. 1995. Design search under probabilistic specifications using genetic algorithms. In *Computer-Aided Design.* Butterworth-Heinemann, to appear.

Optimization in Multidisciplinary Design

Kemper Lewis, Farrokh Mistree, and J. R. Jagannatha Rao

Introduction

The design of complex systems is a difficult task of integrating disciplines, each with their own analysis, synthesis, and decision process. Optimizing such a system on a global scale is realistically impossible, but finding a solution that is "good enough" and robust is achievable. Several approaches to formulating and solving a multidisciplinary design problem have arisen in a rather ad hoc fashion since the inception of multidisciplinary design optimization (MDO). These approaches include single-level and multilevel formulations, hierarchical and nonhierarchical system decomposition methods, and numerous optimization and analysis processes and approaches at the system and subsystem levels. Designers are referred to as decision makes and objectives as goals or rewards. With only one decision maker, the problem becomes a scalar or vector optimization problem. But in MDO, many decision makers may exist, and each decision maker's strategy to optimize his/her rewards could depend on the strategies and decisions of other decision makers. The modeling of strategic behavior based on the actions of other individuals is known as a *game*. Therefore, the focus in this section is on problems characterized by

- Single decision makers who have multiple rewards
- Multiple decision makers who have multiple rewards

This focus in optimization theory is shown as the shaded region in Figure 11.10.13. Typical courses in optimization focus on the upper-left quadrant, namely, scalar optimization problems with one objective and one decision maker. In this section, the focus is on the other three quadrants as a means to expand the application of optimization theory to problems that frequently occur in engineering design.

Classification of MDO Formulations

Simultaneous and Nested Analysis and Design. With the advent of MDO and its various applications, a structure of problem approaches and formulations is necessary. In Cramer et al. (1994) and, more recently, Balling and Sobieski (1994) and Lewis and Mistree (1995), various classes and classifications of problem formulations are presented. Figure 11.10.14 is a generic representation of a coupled, three-discipline system.

Depending on the level of analysis, the modules in Figure 11.10.14 may refer to disciplines, components, or processes. It is the decomposition of the system, subsystem coupling and solution, and system synthesis that pose major research and application problems in MDO. The terms used in the figure, as well as other common terms, and are defined below.

s_1, s_2, s_3: disciplinary *state variables* which comprise the *state equations*
y_1, y_2, y_3: the *state equations*
r_1, r_2, r_3: *residuals* in the state equations
y_{12}, y_{13}, y_{21}, y_{23}, y_{31}, y_{32}: *coupling functions*, y_{ij} contains those functions computed in discipline i which are needed in discipline j

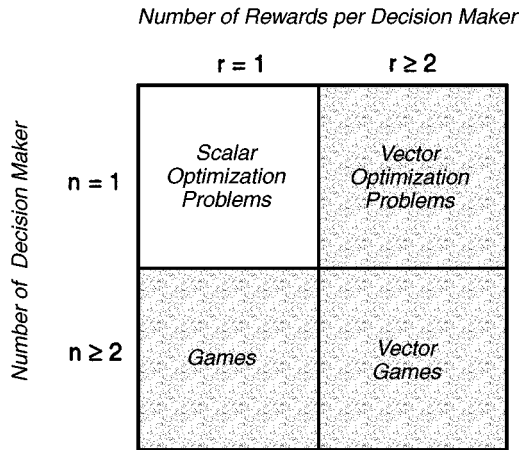


FIGURE 11.10.13 Various formulations in optimization theory.
t

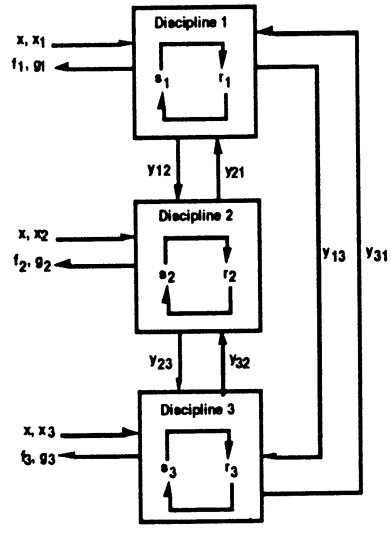


FIGURE 11.10.14 A three-discipline coupled system. (From Balling, R.J. and Sobieski, J. 1994. *5th AIAA/USAF/NASA/ISS/MD Symp. on Recent Advances in Multidisciplinary Analysis and Optimization.* Panama City, FL. 753–773.)

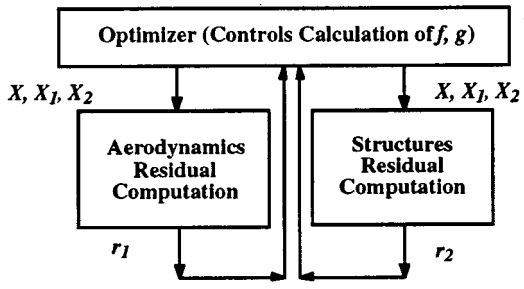


FIGURE 11.10.15 Single-SAND-SAND formulation.

$y^*_{12}, y^*_{13}, y^*_{21}, y^*_{23}, y^*_{31}, y^*_{32}$: *coupling variables*
x: *system design variables* needed by more than one discipline
x_1, x_2, x_3: *disciplinary design variables*
g_1, g_2, g_3: *design constraint functions*
f_1, f_2, f_3: *design objective functions*

The primary task at hand is summarized as follows: *Determine the values of the design, state, and coupling variables which satisfy the state equations, the coupling equalities, the design constraints, and the design objective functions.*

Based on this, six classifications for fundamental approaches to MDO problem formulation and solution are presented by Balling and Sobieski, which depend on three criteria:

1. System vs. multilevel decomposition
2. Simultaneous (SAND) vs. nested analysis and design (NAND) at the *system* level.
3. Simultaneous (SAND) vs. nested analysis and design (NAND) at the *subsystem* or *discipline* level.

At the discipline level, SAND implies that the disciplinary design and state variables are determined simultaneously by the optimizer, while NAND implies that the optimizer determines only the disciplinary design variables and requires determination of the state variables at each iteration. At the system level, SAND implies that the system design variables and coupling variables are determined simultaneously by the system optimizer, while NAND implies that the system optimizer determines only the system design variables and requires calls to a system analysis routine to determine the coupling variables at each iteration. The "optimizers" at the system level or discipline level could be gradient based or heuristic in nature, depending on the problem formulation. Further classifications can be generated if these approaches are combined or linked sequentially within one design problem.

Each approach has a three-part name consisting of the overall decomposition descriptor, the solution approach at the system level, and the solution approach at the subsystem level. The first part indicates whether the approach is a single-level or multilevel approach. The middle and last parts of the name indicate whether the SAND or NAND approach is used at the system and discipline levels, respectively.

Example

These terms are illustrated based on a simple example from aeroelastic MDO problem in Cramer et al. (1994). This aeroelastic problem involves two disciplines, aerodynamics and structures. This problem is extremely difficult, as the solution of either of these disciplines *alone* involves heavy analysis and computation. For simplicity, it is assumed that a *single-level* scheme is to be used. The first formulation choice is SAND or NAND at the system level. If a SAND formulation is used, the discipline-level formulation must be determined. These formulations are illustrated below.

Single-SAND-SAND. In this formulation, feasibility is *not* sought for the analysis problem in any sense until convergence in the solver is reached. The analysis "codes" for aerodynamics and structures in this formulation perform a simple function; they evaluate the *residuals* of the analysis equations, rather than solving some set of equations. This is illustrated in Figure 11.10.15. The *optimizer* controls the calculation of the objective functions, f, and the constraints, g, based on the residuals, r_i, from the disciplines. The optimizer sends system and disciplinary design variables, X and X_i, to the disciplinary analysis routines.

Single-SAND-NAND. In this formulation, feasibility *is* enforced for each individual discipline, while the optimizer drives the individual disciplines toward multidisciplinary feasibility and optimality by controlling the coupling functions between the disciplines. This is illustrated in Figure 11.10.16. The optimizer sends design and coupling variables, X and y^*_{ij}, to the disciplinary analysis routines, while these routines return coupling functions, y_{ij}, which are functions of the design variables and the disciplinary state variables, y_i.

Single-NAND-NAND. In this formulation, multidisciplinary feasibility is required at each iteration of the optimizer. Therefore, an aeroelastic analysis solver combines the information from the aerodynamics

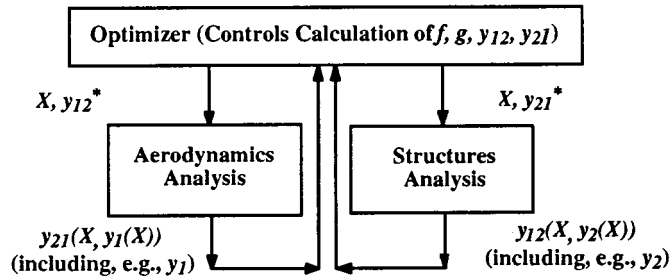


FIGURE 11.10.16 Single-SAND-NAND formulation.

and structures disciplines and ensures that the two disciplines are feasible concurrently. This is illustrated in Figure 11.10.17. The optimizer only controls the objective functions and constraints, while the coupling functions are handled by the aeroelastic analysis solver which becomes a "black box" from the perspective of the optimization code. In each discipline, an analysis code solves a set of equations. The resulting *state equations* are fed into the other disciplines. This process continues until convergence among the disciplines, and the optimizer is once again invoked.

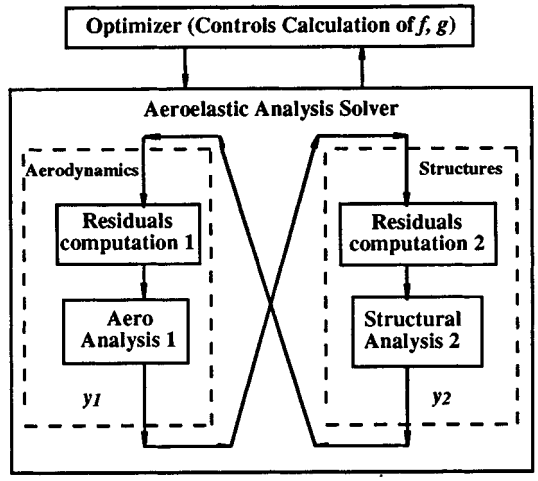


FIGURE 11.10.17 Single-NAND-NAND formulation.

Modeling Design Decisions: The Mathematical Programming Background

Decision-based design (DBD) is offered as a starting point for the creation of design methods that are based on the notion that the principal role of an engineer in the design of a product or process is to make decisions. Therefore, the process of design in its most basic sense is a series of decisions.

It is recognized that the implementation of DBD can take many forms; one implementation is the decision support problem (DSP) technique. Within the DSP technique there exist three types of decisions a designer can make: selection, compromise, or heuristic. In this section, the focus is on the compromise DSP, which is used to model multiobjective trade-offs in the solution of a mathematical model.

Compromise DSP. A compromise DSP is a hybrid formulation that incorporates concepts from both traditional mathematical programming and goal programming, and makes use of some new ones (Mistree et al., 1993). It is similar to goal programming in that the multiple objectives are formulated as system goals (involving both system and deviation variables) and the deviation function is solely a function of the goal deviation variables. This is in contrast to traditional mathematical programming where multiple objectives are modeled as a weighted function of the system variables only. The concept of system

constraints, however, is retained from the traditional constrained optimization formulation. Special emphasis is placed on the bounds on the system variables, unlike in traditional mathematical programming and goal programming. In effect the traditional formulation is a subset of the compromise DSP — an indication of the generality of the compromise formulation. The compromise DSP is stated in words as follows:

Given

An alternative that is to be improved through modification
Assumptions used to model the domain of interest
The system parameters
All other relevant information

n	Number of system variables
p + q	Number of system constraints
p	Equality constraints
q	Inequality constraints
m	Number of system goals
$g_i(X)$	System constraint function

$$g_i(\underline{X}) = C_i(\underline{X}) - D_i(\underline{X})$$

$f_k(d_i)$	Function of deviation variables to be minimized at priority level k for the preemptive case
W_i	Weight for the Archimedean case

Find

The values of the independent *system variables* (they describe the physical attributes of an artifact)

$$X_j \quad j = 1,\ldots, n$$

The values of the *deviation variables* (they indicate the extent to which the goals are achieved)

$$d_i^-, d_i^+ \quad i = 1,\ldots, m$$

Satisfy

The *system constraints* that must be satisfied for the solution to be feasible, there is no restriction placed on linearity or convexity

$$g_i(\underline{X}) = 0; \quad i = 1,\ldots, p$$

$$g_i(\underline{X}) \geq 0; \quad i = p+1,\ldots, p+q$$

The *system goals* that must achieve a specified target value as far as possible; there is no restriction placed on linearity or convexity

$$A_i(\underline{X}) + d_i^- - d_i^+ = G_i; \quad i = 1,\ldots, m$$

The lower and upper *bounds* on the system

$$X_j^{\min} \leq X_j \leq X_j^{\max}; \quad j = 1,\ldots, n$$

$$d_i^-, d_i \geq 0 \quad \text{and} \quad d_i^- \cdot d_i^+ = 0$$

Minimize

The *deviation function* which is a measure of the deviation of the system performance from that implied by the set of goals and their associated priority levels or relative weights:

Case a: Preemptive (lexicographic minimum)

$$Z = \left[f_1\left(d_i^-, d_i^+\right), \ldots, f_k\left(d_i^-, d_i^+\right) \right]$$

Case b: Archimedean

$$Z = \sum_{i=1}^{m} W_i\left(d_i^- + d_i^+\right); \qquad \sum W_i = 1; \quad W_i \geq 0$$

A graphical representation of a two-variable compromise DSP is shown in Figure 11.10.18. The difference between a system variable and a deviation variable is that the former represents a distance in the i^{th} dimension from the origin of the design space, whereas the latter has as its origin the surface of the system goal. The value of the i^{th} deviation variable is determined by the degree to which the i^{th} goal is achieved. It depends upon the value of $A_i(X)$ alone (Since G_i is fixed by the designer), which in turn is dependent upon the system variables X. The set of discrete variables can be continuous, Boolean, or a combination of types.

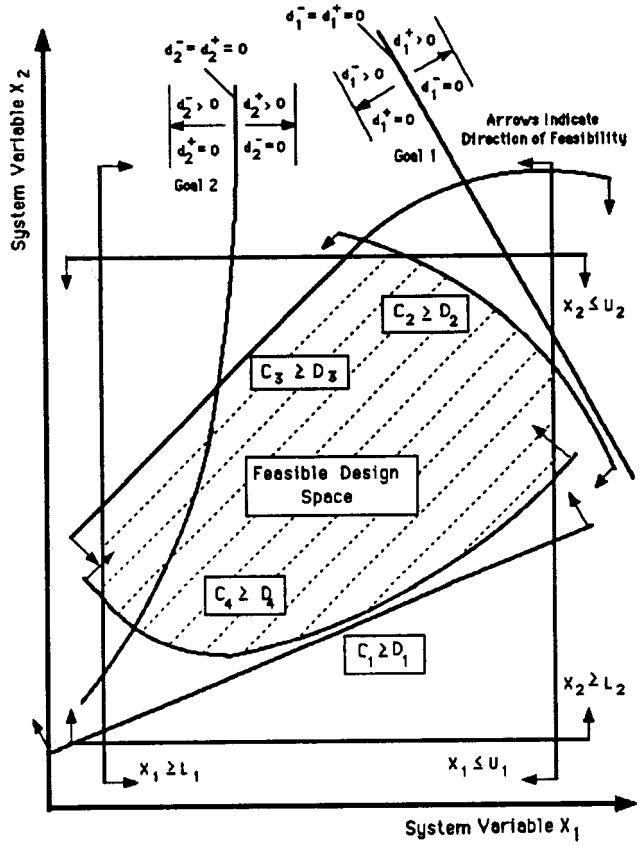


FIGURE 11.10.18 The compromise DSP.

The compromise DSP is fundamentally different from the traditional mathematical optimization model. This difference is depicted in Figure 11.10.19 where the optimization model on the left occurs when designer's aspirations can be met by the system achievement and therefore an optimum can be found. But when a designer's aspirations do not overlap with the actual achievements, a solution which is "good enough" or *satisficing*[*] must be found. This is depicted and modeled in the compromise DSP on the right of Figure 11.10.19. When the aspiration space overlaps with the feasible design space (center of Figure 11.10.19), the optimization model and compromise DSP are the same.

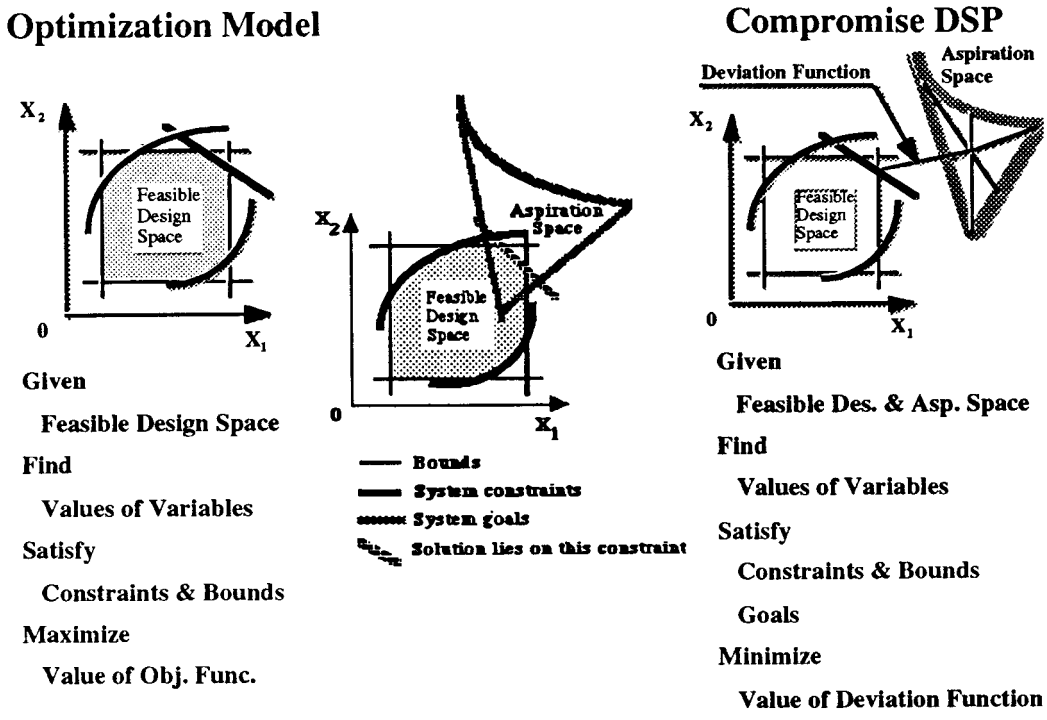


FIGURE 11.10.19 The optimizing and satisficing models.

Deviation Variables and Goals. In a **goal** one can distinguish the aspiration level, G_i, of the decision maker and the actual attainment, $A_i(X)$, of the goal. Three conditions need to be considered:

1. $A_i(X) \le G_i$; one wishes to achieve a value of $A_i(X)$ that is equal to or less than G_i.
2. $A_i(X) \ge G_i$; one wishes to achieve a value of $A_i(X)$ that is equal to or greater than G_i.
3. $A_i(X) = G_i$; one wishes to achieve a value of $A_i(X)$ equal to G_i.

Next, the concept of a **deviation variable** is introduced. Consider the third condition, namely, one would like the value of $A_i(X)$ to equal G_i. The deviation variable is defined as

$$d = G_i - A_i(\underline{X})$$

The deviation variable d can be negative or positive. In effect, a deviation variable represents the distance (deviation) between the aspiration level and the actual attainment of the goal. Considerable simplification

[*] Satisficing — not the "best", but "good enough" (the first use of this term, in the context of optimization, is attributed to Herbert Simon (Simon, 1982).

of the solution algorithm is affected if one can assert that all the variables in the problem being solved are positive. Hence, the deviation variable d is replaced by two variables:

$$d = d_i^- - d_i^+$$

where

$$d_i^- \cdot d_i^+ = 0 \quad \text{and} \quad d_i^-, d_i^+ \geq 0$$

The preceding ensures that the deviation variables never take on negative values. The product constraint ensures that one of the deviation variables will always be zero. The system goal becomes

$$A_i(\underline{X}) + d_i^- - d_i^+ = G_i; \quad i = 1, 2, \ldots, m \tag{11.10.3}$$

subject to

$$d_i^-, d_i^+ \geq 0 \quad \text{and} \quad d_i^- \cdot d_i^+ = 0 \tag{11.10.4}$$

If the problem is solved using an algorithm that provides a vertex solution as a matter of course, then the constraint is automatically satisfied making its inclusion in the formulation redundant. Since some algorithms use solution schemes which provide a vertex solution, it is assumed that this constraint is satisfied. For completeness this constraint is included in the mathematical forms of the compromise decision support problem given previously in this chapter and for brevity will be omitted from all subsequent formulations.

Note that a goal (Equation 11.10.3) is always expressed as an equality. When considering Equation (11.10.3), the following will be true:

if $A_i \leq G_i$ **then** $d_i^- > 0$ and $d_i^+ = 0$.
if $A_i \geq G_i$ **then** $d_i^- = 0$ and $d_i^+ \geq 0$.
if $A_i = G_i$ **then** $d_i^- = 0$ and $d_i^+ = 0$.

How are the three conditions listed using Equation (11.10.3)?

1. To satisfy $A_i(X) \leq G_i$, the positive deviation d_i^+ must be equal to zero. The negative deviation d_i^- will measure how far the performance of the actual design is from the goal.
2. To satisfy $A_i(X) \geq G_i$, the negative deviation d_i^- must be equal to zero. In this case, the degree of overachievement is indicated by the positive deviation d_i^+.
3. To satisfy $A_i(X) = G_i$, both deviations, d_i^- and d_i^+, must be zero.

At this point it is established what is to be minimized. In the next section means for minimization of the objective in goal programming are introduced.

The Lexicographic Minimum and the Deviation Function. The objective of a traditional single objective optimization problem requires the maximization or minimization of an objective function. The objective is a function of the problem variables. In a goal programming formulation, each of the objectives is converted into a goal (Equation 11.10.3) with its corresponding deviation variables. The resulting formulation is similar to a single objective optimization problem but with the following differences:

The objective is always to minimize a function.
The objective function is expressed using deviation variables only.

The objective in the goal programming formulation is called the achievement function. As indicated earlier the deviation variables are associated with goals and hence their range of values depend on the

goal itself. Goals are not equally important to a decision maker. Hence to effect a solution on the basis of preference, the goals may be rank-ordered into priority levels.

One should seek a solution which minimizes all unwanted deviations. There are various methods of measuring the effectiveness of the minimization of these unwanted deviations. The *lexicographic minimum* concept is a suitable approach, and it is defined as follows (see Ignizio, 1982; Ignizio, 1985).

Given an ordered array $\boldsymbol{f} = (f_1, f_2, \ldots, f_n)$ of nonnegative elements f_k, the solution given by $f^{(1)}$ is preferred to $f^{(2)}$ iff

$$f_k^{(1)} < f_k^{(2)}$$

and all higher-order elements (i.e., $f_1, \ldots, f_{k-1}$) are equal. If no other solution is preferred to $\boldsymbol{f}$, then $\boldsymbol{f}$ is the lexicographic minimum.

As an example, consider two solutions, $\boldsymbol{f}^{(r)}$ and $\boldsymbol{f}^{(s)}$, where

$$\boldsymbol{f}^{(r)} = (0,\ 10,\ 400,\ 56)$$

$$\boldsymbol{f}^{(s)} = (0,\ 11,\ 12,\ 20)$$

In this example, note that $\boldsymbol{f}^{(r)}$ is preferred to $\boldsymbol{f}^{(s)}$. The value 10 corresponding to $\boldsymbol{f}^{(r)}$ is smaller than the value 11 corresponding to $\boldsymbol{f}^{(s)}$. Once a preference is established, then all higher-order elements are assumed to be equivalent.

The Use of Coupled DSPs in Modeling and Solving Systems

Coupled compromise DSPs can be used to model multiobjective problems. These types of problems may involve the analysis and synthesis of problems from multiple disciplines, but in this section, the primary notion is *shared* design variable vectors. It is common in the design of complex systems, such as aircraft, for multiple design teams to include the same design variables in their design process. For example, consider the design of aircraft. The design variable, wing area, is used by the aerodynamics group to control the lift force. It is used by the structural group to control the structural framework of the wing. It is used by the propulsion group to control the amount of fuel needed. It seems advantageous to model systems this way, but this introduces an added level of complexity in a system model. Different disciplines are also dependent on the decisions made in other disciplines. Therefore, another focus is modeling the coupling of state variables between disciplines.

To begin the discussion, suppose that a designer (i.e., a decision maker) is investigating a typical engineering system such as a load-bearing structure, a high-speed mechanism, or an automotive transaxle. Assuming that this system is unambiguously described by a finite-dimensional variable vector $x \in \mathcal{R}^n$ (includes both the so-called "state" and "design" variables), then the general descriptive mathematical model can be characterized by a set of *parameters* p, which are fixed quantities over which the designer has no control, and the multifunction X, which represents the state equations and variational inequalities (derived from the natural laws such as the conservation principles) as well as the performance constraints for this system. Numerical approaches to such problems have been studied in Outrata (1990) and Outrata and Zowe (1995). If the only goal of the designer is to obtain a satisfying or a feasible design, then his/her task is simply to find a design x^o that satisfies the constraints and meets the goals as close as possible for a given p. On the other hand, if the designer wishes to find the design which minimizes the deviation between the goal achievement and aspiration with respect to several objective functions such as $f_i(x,p)$, i = 1, …, r, then the problem is to find a design x* that solves the following multiobjective formulation:

$$\underset{x \in X(p) \subset \Re^n}{\text{minimize}}\, d(x,p) = \left\{ d_1(x,p), \ldots, d_r(x,p) \right\} \tag{11.10.5}$$

where d_i represents the deviation from the objective f_i, to its target value, f_{iTV}. In other words, the problem is to minimize the deviation between what a designer wants and what a designer can achieve.

At this general level of discussion, one readily asserts that a model such as in Equation (11.10.5) is the typical starting point for much of the current education, research, and practice in mechanical systems modeling and applied optimization, and yet in specific design instances, this assertion should be boldly challenged. For example, since the designer only controls x and another foreign party controls p, how is p chosen? Can the designer assume that the foreign decision maker will always select the vector that is most advantageous to the design (as is tacitly assumed, it is argued later, by most justifications of existing engineering models)? If not, how should the designer respond to this conflict? A two-player strategic game has just been described (see, e.g., Von Neumann and Morgenstern, 1944; Aubin, 1979; Dresher, 1981) where one player controls x and the other player controls p and where p represents all decisions which are outside the scope of the designer. By the classical definition, a "game" consist of multiple decision makers or players (or designers, in this case) who each control a specified subset of system variables and who seek to minimize their own deviation functions subject to their individual constraints. To continue this argument further, consider a typical schedule of a product design and manufacture, as shown in Figure 11.10.20.

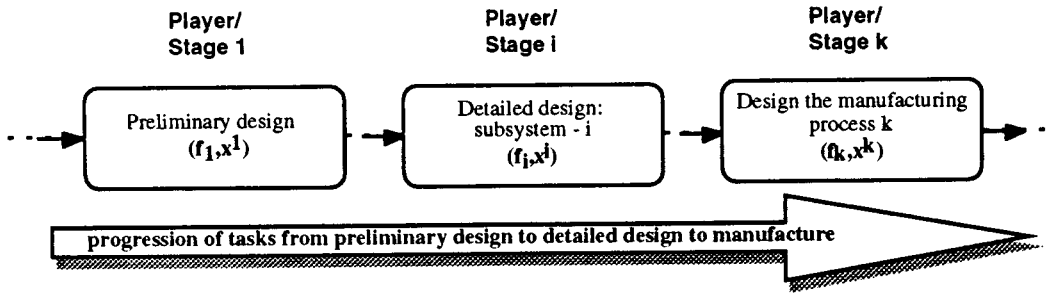


FIGURE 11.10.20 Different stages in a design time-line.

The decision makers at each of the several stages or stations in the time line are referred to as different players, even though they could be the same human designer. Each of the players controls and minimizes his own deviation function.

In the context where multiple decision makers collectively influence the final design, one must pose the following question: what mathematical formulation is appropriate for modeling the system in order to study the decisions made by *more than one player along the above time line?* Clearly, the conventional model of Equation (11.10.5) does not adequately represent the multiplayer scenario of the time line shown in Figure 11.10.20. This inadequacy is most acute when the players engage in strategic behavior which is anything other than total cooperation (e.g., dominance and noncooperation, among others). Thus, the major point of departure can be summarized as: (1) instead of formulating a single mathematical model at a single stage of the time line, as is done conventionally, these models represent one or more decision makers who determine the final design; and (2) each of the r decision makers has his own objective functions and one player's decision is influenced by the other. Thus, the final design could very well depend on whether the players compete, cooperate, or dominate one another — thus necessitating a study of the strategic interaction between the players. In short, a goal here is to model the design of Figure 11.10.20 as a *strategic game*, this figure perhaps being representative of a general engineering system design process at a high level of abstraction.

Mathematical modeling of such strategic behavior, where one decision maker's action depends on decisions by others, is well-established in wide-ranging applications from economics, business, and military (Aubin, 1979; Dresher, 1981; Fudenberg and Tirole, 1991; Mesterton-Gibbons, 1992). Such models are widely used in designing markets and auctions, and in predicting the outcome of encounters between competing companies or trading nations. As specific examples, consider automobile manufac-

turers who interact strategically in order to set prices and production schedules of cars; or the strategic behavior that occurs between drivers as they try to pass each other on a busy undivided highway.

References

Aubin, J.P. 1979. *Mathematical Methods of Game and Economic Theory.* North-Holland, Amsterdam.

Balling, R.J. and Sobieski, J. 1994. Optimization of coupled systems: a critical overview of approaches. In *5th AIAA/USAF/NASA/ISSMO Symp. on Recent Advances in Multidisciplinary Analysis and Optimization.* Panama City, FL. 753–773.

Cramer, E.J. Dennis, J.E., Frank, P.D., Lewis, R.M., and Subin, G.R. 1994. Problem formulation for multidisciplinary optimization. *SIAM J. Optimization.* 4(4): 754–776.

Dresher, M. 1981. *Games of Strategy.* Dover, New York.

Fudenberg, D. and Tirole, J. 1991. *Game Theory.* MIT Press, Cambridge, MA.

Ignizio, J.P. 1982. *Linear Programming in Single and Multi-Objective Systems.* Prentice-Hall, Englewood Cliffs, NJ.

Ignizio, J.P. 1985. *Introduction to Linear Goal Programming.* Sage University Papers. Beverly Hills, CA.

Ignizio, J.P. 1985. Multiobjective mathematical programming via the MULTIPLEX model and algorithm. *Eur. J. Operational Res.* 22: 338–346.

Lewis, K. and Mistree, F. 1995. On developing a taxonomy for multidisciplinary design optimization: a decision-based approach. In *The First World Congress of Structural and Multidisciplinary Optimization.* N. Olhoff and G.I.N. Rosvany, Eds., Elsevier Science, Oxford, UK: ISSMO. 811–818.

Mesterton-Gibbons, M. 1992. *An Introduction to Game-Theoretic Modeling.* Addison-Wesley, Redwood City, CA.

Mistree, F., Hughes, O.F., and Bras, B.A. 1993. The compromise decision support problem and the adaptive linear programming algorithm. In *Structural Optimization: Status and Promise*. AIAA, Washington, D.C., 247–286.

Outrata, J.V. 1990. On the numerical solution of a class of Stackelberg games. *Methods Models Operations Res.* 34: 255–277.

Outrata, J.V. and Zowe, J. 1995. A numerical approach to optimization problems with variational inequality constraints. *Math. Programming.* 68: 105–130.

Simon, H.A. 1982. *The Sciences of the Artificial.* MIT Press, Cambridge, MA.

Von Neumann, J. and Morgenster, O. 1944. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, NJ.