13

Precision Positioning of Rotary and Linear Systems

13.1 Introduction

Precision positioning systems have historically been a key part of successful industrial societies. The need to make something requires the ability to move something with a very high level of accuracy. This has not changed in the Information Age, but instead has become even more important as global competition forces manufacturers to hold ever tighter specifications, with increased throughput and reduced costs. Automation is the equalizer that allows manufacturers in countries with high labor rates to compete globally with developing countries. The definition of a precision machine continues to evolve as technology advances. The components used to build machines become more precise, and so the machines themselves improve. As loosely defined here, precision machines are those that repeatably and reliably position to within a tolerance zone smaller than is possible in commonly available machines. Designing machines to best use the available components and manufacturing techniques requires specialized skills beyond those of general machine designers. It is interesting to note that many of the fundamental rules for designing precision machines have not changed for hundreds of years. Evans [9] tracks the evolution of precision machines and provides a historical context to the present state-of-the-art.

Modern precision-positioning systems are largely mechatronic in nature. Digital computers interface with electronic sensors and actuators to affect the motion of the system mechanics. Auslander and Kempf [2] review the basic elements required to interface between system mechanics and control electronics, while Aström and Wittenmark [24] and Franklin et al. [11] present the required discrete-time control theory and implementation. Kiong et al. [17] describe how these elements are specifically combined...
in the context of precision motion control. The performance of the overall system depends on all of its mechanical, electronic, and software components, and so a multi-disciplinary design team is usually needed to undertake the project development. Precision machines are increasingly being used as a part of an overall manufacturing system and, thus, are required to share process information over a network with many other systems. This information may be high- or low-speed and synchronous or asynchronous to the processes. Decisions on the control hardware and network structure are often very expensive to change later in the design and should be evaluated in parallel with the more-traditional electromechanical elements of the machine design.

This chapter focuses on issues important to engineers involved in the specification and evaluation of commercially available motion control systems. The emphasis is primarily on presenting precision machine design principles and comparing the uses and specifications of different existing technologies; a thorough review of bottom-up precision machine design is beyond the scope of this work, but the interested reader is referred to Shigley and Mischke [30], Slocum [34], Smith and Chetwynd [35], and Hale [15] for a more in-depth look at undertaking a precision machine design. The American Society for Precision Engineering [1] also publishes a journal and maintains a list of recommended readings.

We will limit the discussion of precision machines to those features most common in modern designs. These include discrete-time microprocessor-based control, ballscrew or direct-drive mechanisms, brushless motors, and quantized position feedback. This is not to suggest that all precision machines contain these elements or that any device lacking them is not precise, rather it is an observation that a designer looking to build his own motion control system, or select from available subsystems, will generally find these to be economical alternatives. Kurkes and Jenkins [18] detail precision control techniques as applied to systems with predominantly analog feedback sensors and compensation. The fundamental analog and continuous-time control design techniques still apply, but adding microprocessor control and quantized feedback adds an additional set of advantages and limitations that will be explored here.

13.2 Precision Machine Design Fundamentals

Precision machines vary widely in their design and application but follow several common principles. Machines are generally not accurate by accident but because of the effort taken in design throughout the process. Determinism is a proven machine design philosophy that guides engineers in these developments. Error motions and imperfections in the components used to build precision machines have different effects on the overall performance depending on their location in the system. Some errors are relatively inconsequential; other errors heavily influence the overall performance. One class of errors occurs when an angular error is multiplied by a lever arm into a linear error. These errors, referred to as Abbé errors, should be minimized in all precision machine designs. Applying forces to a machine strains the mechanical elements. These dimensional changes may affect the accuracy of the measurements. Machines should, therefore, be designed such that the load-bearing elements are as separate as possible from the elements used for metrology. Machines are all made of interconnected elements, and the method of connecting these elements affects both the precision of the design and the ease with which it can be analyzed. Exact-constraint designs are based on having contact at the minimum number of points necessary to constrain the required degrees-of-freedom of the free part. Elastic-averaging is the opposite and relies on contact over so many points that errors are averaged down to a low level. Problematic designs are those that occur somewhere in the middle of the two extremes. Finally, some often-neglected elements of a precision machine design include cable management, the environment that the system will operate in, analysis of heat generation and flow, and maintenance. Each of these items will be further detailed here.

13.2.1 Definitions of Precision

The terms accuracy, repeatability, and resolution are frequently misused when applied to precision machines. Accuracy is the nearness of a measurement to the standard calibrated value. Absolute accuracy is expensive, and there are few shortcuts to achieving it. The position readout of a system may indicate a 100 mm move,
Poor Repeatability
Good Average Accuracy

Poor Accuracy
Good Repeatability

FIGURE 13.1 The figure on the left shows a measurement with poor repeatability and poor accuracy on any single measurement but good accuracy when the measurements are averaged together. The figure on the right shows a measurement with poor accuracy but excellent repeatability. Accuracy can usually be improved by mapping techniques, and so a machine with high repeatability is usually preferred.

but the accuracy of the system quantifies how closely this 100 mm compares with the standard measure of a meter. In applications that require absolute accuracy (e.g., semiconductor wafer mask writers, machine tools, and metrology instruments), the designer should be prepared to invest in a temperature-controlled environment, temperature-compensated mapping, and high attention to details of sensor location relative to the workpiece. In some applications, accuracy can be achieved statistically by taking many measurements and averaging the results together. In many other applications, absolute accuracy is less important than repeatability.

Repeatability is the nearness of successive samples to each other, and it is the goal of a deterministic machine design to produce as repeatable a machine as is possible. Figure 13.1 illustrates the difference between repeatability and accuracy. If the 100 mm move on our hypothetical machine is truly 99.95 mm, but is the same 99.95 mm each time it makes the move, then the machine is inaccurate but repeatable. Repeatability is a broad term that captures the overall uncertainty in both the positioning ability of a system and the measurement of this positioning. Repeatability can be measured as uni-directional (in which the same point is approached repeatably from the same direction) and bi-directional (in which the point is approached from both directions). Systems can have excellent repeatability when approached from a single direction, but terrible when approached from both (consider the case of a backlash in a leadscrew).

Know which specification you are looking at. Depending on the level of modeling required, a statistical distribution of the measurements can be constructed to provide confidence intervals on the positioning ability of the system. The repeatability measurement itself can be separated into short-term and long-term repeatability. Short-term apparent nonrepeatability is generally caused by small variations in the size or roundness of the rolling elements in the bearings or ballscrew. Deterministic machine design holds that even these errors could be predicted with enough modeling, but the effort required to do so is forbidding. Temperature variations are usually the cause of longer-term apparent nonrepeatability. Nonuniform drag from any cables or hoses can also cause different performance over time. The goal of a precision machine design, therefore, becomes to design a system such that the level of nonrepeatability is small enough to be unimportant to the application.

The other frequently misused term regarding machine performance is resolution. Resolution (by itself) can be taken to mean mechanical resolution, which is the minimum usable mechanical displacement, or electrical resolution, which is the least significant digit of the displacement sensor. These two items are only loosely related, and the designer should again be clear about which value is given in the specification. Mechanical resolution is the ultimate requirement and is limited by characteristics of the control system, vibration in the environment, electrical noise in the sensors, and mechanical static friction effects. Fine mechanical resolution is required for good in-position stability, implementing error-correcting maps, and achieving tight velocity control.
13.2.2 Determinism

Researchers at the Lawrence Livermore National Laboratory developed and championed the philosophy of determinism in precision machine design. Much of this work was performed in the context of designing large-scale high-precision lathes, referred to as diamond-turning machines. The philosophy of determinism maintains that machine behavior follows and can be predicted by familiar engineering principles [4]. The emphasis here is on familiar, since esoteric models are generally not required to predict behavior even at the nanometer level. A careful and thorough accounting of all elements is often sufficient. Determinism maintains that machine behavior is highly repeatable. Apparent randomness in machine behavior is simply a result of inadequate modeling or lack of control over the environment. Machines must behave exactly as designed but may simply not have been designed adequately for the required task. This is not to suggest that creating a mathematical model that captures all the nuances of the machine performance is straightforward, or even practical, given limited resources. Rather, determinism guides the selection and design of systems in which the apparent randomness or nonrepeatability is below a certain level and for which models can be developed economically. Central to the deterministic machine design is the error budget in which all of the machine errors are tabulated and translated into a final error between a tool and workpiece through the use of transformation matrices. Slocum [34] and Hale [15] provide details of these calculations. A properly designed error budget allows the engineer to quickly assess the impact of component changes on the machine performance and to identify the elements that have the largest impact on overall error.

13.2.3 Alignment Errors (Abbé Principle)

Alignment errors occur when the axis of measurement is not precisely aligned with the part to be measured. There are two basic types of alignment errors, named cosine and sine based on their influence on the measurement. Examples of cosine errors include nonparallelism between a linear stage and a linear scale, or taking a measurement with a micrometer when the part is seated at an angle. Figure 13.2 illustrates a typical cosine error. The true displacement \( D \) and the measured value \( D_c \) are related by simple geometry through

\[
D_c - D = D_c (1 - \cos \theta)
\]  

(13.1)

where \( \theta \) is the misalignment angle. For small angles of \( \theta \), the error is well-approximated by

\[
D_c - D \approx D_c \frac{\theta^2}{2}
\]  

(13.2)

Cosine errors are typically small but should not be dismissed as insignificant at sub-micron levels. They can usually be compensated for with a simple scale factor correction on the feedback device.

Sine errors, also known as Abbé errors, occur when an angular error is amplified by a lever arm into a linear error. Minimizing these errors invokes a principle derived by Dr. Ernst Abbé who noted that

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 [34].

FIGURE 13.2 Example of cosine error in a typical machine setup.
Abbé errors typically occur on precision machines when a scale or measuring device is offset from the tool tip on a stage. The true displacement $D$ and the measured value $D_s$ in this case are related by

$$D_s - D = R \sin \theta$$  \hspace{1cm} (13.3)

where $R$ is the length of the lever arm, and $\theta$ is again the angular error (Figure 13.3). For small angles, the error is approximated by

$$D_s - D \approx R \theta$$  \hspace{1cm} (13.4)

In contrast to cosine errors, Abbé errors are almost always significant in the overall performance of a machine and should be avoided to the greatest extent possible. As an example of the effect of sine errors, consider a linear slide having a linear accuracy of 1 $\mu$m at the center of the table and a yaw error of 20 $\mu$rad (approximately 4 arc-sec). At the edge of a 300 mm diameter payload ($R = 150$ mm), the yaw error contributes 3 $\mu$m to the linear error for a total of 4 $\mu$m. A secondary position sensor, usually a laser interferometer, can often be used to supplement the existing stage sensors and measure near to the point of interest. The Abbé Principle is perhaps the easiest to understand but most violated rule in precision machine design. Economy, serviceability, and size requirements often dictate that a multi-axis machine be made in a stacked arrangement of individual axes. It is not uncommon for angular errors on the lowest axis to be multiplied by a lever arm of several hundred millimeters before reaching the tool tip. In other cases, the geometry may offer only limited places to position the sensor, and this is often not at the ideal location. Deterministic machine design does not require strict adherence to the Abbé Principle, but it does require the designer to understand the consequences of violating it through the use of an error budget.

### 13.2.4 Force and Metrology Loops

Most precision machines are closed-systems, meaning that forces are applied and measurements are taken between machine elements. Force and metrology loops are conceptual tools used to guide a structural design and location of sensors and actuators. A metrology loop is the closed path containing all the elements between the sensor and the workpiece that affect the measurement. The metrology loop is not necessarily all mechanical, but can also consist of electrical or optical elements. These include sensor electronics or even the air in the beam path of a laser interferometer. Changes to the temperature, pressure, or humidity of the air change its index of refraction and thus also change the wavelength of the laser. This drift in the sensor readings is interpreted as stage motion. By definition, any change to elements in the metrology loop is impossible to distinguish from actual movement of the tool or workpiece. The force (or structural loop) is similarly defined as the closed path showing the conceptual flow of force in a loop around the structure. Applying force to structural elements generates stress, strain, and distortion of the elements. If any of the distorted elements is shared with the metrology loop, the result will be an inaccurate measurement. Precision machines should be designed to separate the force and metrology loops to the greatest possible extent. In extreme cases, this can be accomplished with a separate metrology frame with structural elements that are completely independent of load-bearing elements. The large optics diamond turning machine at
Lawrence Livermore is an example of a machine designed in this manner [7]. This separation may not be economical in all instances, but it demonstrates the general guideline of monitoring the flow of force in a machine, identifying the elements that influence the measurement, and separating these to the greatest extent practical.

13.2.5 Constraint

13.2.5.1 Exact-Constraint or Kinematic Design

Exact-constraint or kinematic design techniques create interfaces between machine elements in such a way that only the required number of the degrees of freedom are constrained leaving the element free to move in the remaining ones. A free-body has six degrees of freedom (three translation and three rotation), and each constraint reduces the available motion. For example, the ideal linear slide is kinematically constrained with only five points of contact leaving the stage with one remaining translational degree of freedom. Kinematic design is often used in precision engineering because of the relative ease with which geometric errors occurring at the points of constraint can be translated into errors at the workpiece through rigid-body coordinate transformations. There is nominally no structural deformation of the components. Blanding [3], Hale [15], and Schmiechen and Slocum [28] present quantitative analysis techniques for use with kinematic designs. Because the workpiece is exactly constrained, the structural elements behave fundamentally as rigid bodies. When a system is overconstrained, structural elements must deform in order to meet all of the constraints (consider the classic example of the three-legged vs. four-legged stool sitting on uneven ground), and this deformation often requires a finite-element model to analyze properly.

The design of kinematic couplings is a subset of exact-constraint design for items that are meant to be constrained in precisely six degrees of freedom. These fixtures contain exactly six points of contact and are often used for mounting components without distortion. Precision optical elements are often mounted using these techniques. Properly designed kinematic couplings are also highly repeatable, often allowing a component to be removed and replaced within micron or better tolerances. The classic kinematic couplings are the three-ball, three-groove coupling and the flat-vee-cone coupling (where the “cone” should more formally be a trihedral hole making three points of contact with the mating ball). Three-ball, three-groove couplings have the advantage of common elements in their construction with a well-defined symmetry about the center of the coupling. Any thermal expansion of the parts would cause growth about the center. The flat-vee-cone coupling has the advantage of allowing the entire coupling to pivot about the cone should any adjustments be necessary. The cone and mating ball are fixed together under thermal expansion. Slocum [32,33] presents analysis tools useful for designing kinematic couplings. In practice, kinematic couplings must be preloaded in order to achieve any meaningful level of stiffness. This stiffness can be determined quantitatively through an analysis of Hertzian contact stress. In general, larger radii on the contact points allow for the use of higher levels of preload, and greater stiffness. Under preload, however, the contact points become contact patches, and the coupling no longer has precisely six “points” of contact. Friction and wear at the contact patch decrease the repeatability of the coupling, and practical designs often include flexures that allow for this motion [14,29].

13.2.5.2 Elastic Averaging

Elastic averaging is the term used when contact occurs over such a large number of points that errors are averaged down to a low level. The classic hand-scraped machine way is one such example. All of the high points on the mating surfaces have been scraped down to lie in essentially the same plane, and contact between elements occurs uniformly over the entire area. Another example of elastic averaging in machine design is commonly used in precision rotary indexing tables. Two face gears are mated and lapped together. There is contact over so many teeth that individual pitch errors from one to the next are averaged out. This type of mechanism is often referred to as a Hirth coupling. A further example of elastic averaging occurs in recirculating-ball type linear bearings. A number of balls are constrained in a “racetrack” to maintain contact with the rail. Any individual ball will have size and roundness errors, but the average contact of many balls reduces the overall effect of the error.
13.2.6 Thermal Management

In all but the rarest cases, machines must be built with materials that expand and contract with temperature. The site where the machine is to be used should be prepared with appropriate HVAC systems to limit the temperature change in the environment, but it is rarely economical, or even possible, to control the temperature well enough that the machine accuracy will not be compromised to some level. The machine itself usually has some heat sources internal to it. Motors are essentially heating coils, friction in bearings and drivescrews creates heat, and even the process itself may lead to some heat generation. The precision machine design engineer must consider the flow of heat through the system and the effect that it will have on performance and accuracy. There are several techniques available for managing heat flow. One technique is to isolate the heat sources from the rest of the system and provide a well-defined path to transfer the heat away to a cooling sink. This could be as simple as a closed air or fluid path around a motor back out to a radiator, or strategic placement of materials with high thermal conductance. The entire system could also be placed in an air or fluid shower. A separate system keeps a tight control on the shower temperature. Note here that it is usually easier to heat and maintain a fluid or system above ambient temperature than it is to cool below. These techniques are not inexpensive but are often absolutely required for high-accuracy systems.

There are also mechanical design techniques that can be used to limit thermal-growth problems. Linear growth is usually much less of a problem than bending is, so the designer should try to maintain symmetry wherever possible in the design. Symmetric heat flow means equal temperatures and so equal growth about machine centerlines. Closed structural loops are preferable to open C-shapes. A C-shape will open under heating while a ring will expand uniformly. The designer must pay very careful attention to points where dissimilar materials are attached to each other. In each case, there is the possibility for a “bi-metallic strip” effect that can lead to bending. Where material mismatch and nonuniform growth is unavoidable, the designer should add specific compliant elements (i.e., expansion joints) so that the location of the displacement is known.

Finally, it is possible to map the errors induced by thermal growth of a machine. Given a series of measurements taken at different temperatures, and a set of temperature sensors properly located around the system, the control software can either move the axis to the correct position or at least present a corrected position measurement. As is generally the case with mapping and error correction techniques, there are usually a few key contributors to the overall error that are readily addressed, but identifying and correcting beyond these terms is challenging at best.

13.2.7 Cable Management

Cable management is a critical but often overlooked part of the overall machine design. It is often the most unreliable part of a precision machine. The cable management system (CMS) consists of the electrical cables themselves, pneumatic or hydraulic tubing, fittings and connectors, and often a carrier system. Common problems include conductor breakdown, insulation breakdown (shedding), connector reliability, and the influence of the drag force on the stage motion. The cable lifetime should be experimentally verified as early in the design as possible because the lifetime calculations are highly dependent on the actual implementation and mounting techniques. The magnitude of the cable drag force is generally position-dependent and can include a periodically varying component. This is particularly true when a chain-type cable carrier consisting of multiple links is used. In the highest-precision applications, a separate cable-carrier axis can be used to take up the main cable drag force while a short link connects it to the main stage with a relatively-constant force. There are two general techniques available for designing the CMS. The most-exercised portion of the CMS can be connectorized to allow for easy field replacement as part of a preventive maintenance schedule. This is appropriate when the duty cycle suggests that regular cable replacement will be required. However, the extra connectors required in this arrangement are themselves a possible failure point. In light-duty applications, it may be preferable to run continuous cable from the machine elements (motors, encoders, limits, and so on) through to a single junction block. In all cases, the CMS must be designed in parallel with the overall mechanics of the system.

Copyright © 2005 by CRC Press LLC
13.2.8 Environmental Considerations

The design of a motion control system is often largely influenced by the environment in which it operates. Some systems, such as traditional metal-cutting machine tools, must contend with chips and coolant spray. Other forms of manufacturing, such as laser machining, form finer particulates that require additional sealing techniques. Different types of bellows and sliding seals are available for each application, and the disturbance forces that they impart can affect system repeatability and dynamic performance. Other systems must operate in clean-room environments where the emphasis is on preventing the stage from contaminating the process. In this case, special lubricants and seals must be used that are chosen specifically for low-levels of particulate generation.

Semiconductor processing technologies, such as extreme ultraviolet lithography (EUVL) and electron-beam inspection, often require precision motion control in a high-vacuum environment. The key considerations when designing or specifying a motion control system for operating in a vacuum are material selection and heat transfer. Most materials, when subjected to low pressures, will outgas into the vacuum environment. This outgassing rate will limit the achievable vacuum level, and the compounds released may even contaminate the process. In general, uncoated aluminum and stainless steel are acceptable materials for the main body of the structure. Polymers and epoxies used in connectors, cabling, and purchased components must be thoroughly reviewed. Care must be taken in the mechanical design to eliminate trapped volumes that can slowly outgas into the vacuum environment, increasing pumpdown times. This is generally done through the use of cross-drilled holes, machined channels, and vented fasteners. Systems that are designed for use in vacuum environments are generally cleaned and baked out at elevated temperature to eliminate as many contaminants as possible before use. Bakeout temperatures of over 100°C are not uncommon, and the motion system must be designed to survive this temperature and to maintain performance after a return to normal operating temperature. As will be described in the next section, serviceability is critical to the design of systems for use in a vacuum environment. Systems placed in a vacuum chamber generally are difficult (i.e., expensive) to remove after installation, and access to the stage may be limited to just a few windows through the chamber walls.

13.2.9 Serviceability and Maintenance

Machines must be designed to allow for routine preventive maintenance and for field-replacement of the items determined to be most-likely to fail. Many motion control systems are part of larger overall structures, and the designer should carefully monitor access points to the system after full integration. Preventive maintenance usually includes cleaning and bearing relubrication. It may also include replacing part of the cable management system in high-duty-cycle applications. It is important in the design to ensure that all required access is available in the fully assembled system, rather than just in the individual components.

13.3 Mechatronic Systems

13.3.1 Definition of Mechatronic Systems

Mechatronic systems are those systems whose performance relies on the interdependence between mechanical, electrical, and control components. This definition covers most modern precision linear and rotary motion systems. Designing in a mechatronic framework allows the designer to trade complexity between the various disciplines. Software-based error correction tables can be used to improve the accuracy of repeatable mechanics rather than requiring high accuracy of the components themselves. Torque ripple in motors and nonlinearities in feedback sensors can likewise be mapped. It is wrong to assume that all mechanics and electronics can be economically improved by the addition of software and control, but the mechatronic viewpoint shows that gains and tradeoffs are often possible. The emphasis in mechatronic design shifts heavily to the need for repeatability in designs, which agrees well with the deterministic machine design philosophy. Repeatability is the goal, and any non-repeatability or apparent randomness
in the behavior is simply an inadequacy in the model. The philosophy is the same regardless of whether the system is fundamentally mechanical, electrical, or algorithmic in origin.

13.3.2 Discrete-Time System Fundamentals

Feedback compensation, and digital controllers in particular, are largely the element that distinguishes mechatronic systems from traditional electromechanical devices. Analog control continues to be used, and correctly so, in many applications. A compensator can often be built for just a few dollars in components. However, microprocessor control allows for greater flexibility in designing the controller, assures consistent operation, and may be the only method available for implementing some required functionality. Since microprocessor-based control will likely be encountered in some manner in precision system design, we will address it here. Two main features that distinguish analog from digital control are sampling (measurements taken only at fixed time intervals) with the associated aliasing problems and quantization (measurements have only a fixed number of values).

13.3.2.1 Sampling and Aliasing

Sampling refers to the process of taking measurements of system variables at a periodic rate. The system controller can then perform some action based on these measurements at that instant in time. The unavoidable problem is that no measurements of system behavior are available during the inter-sample period. If the frequency content of the signal to be sampled is low enough, then it can be proven that the samples alone are sufficient to exactly and uniquely reconstruct the original waveform. The criteria for this reconstruction is quantified with the Nyquist Sampling Theorem [22]. The Nyquist Sampling Theorem states that if \( x_c(t) \) is a continuous-time bandlimited signal having a frequency content of

\[
X_c(j\omega) = 0 \text{ for } |\omega| > \omega_N
\]

then \( x_c(t) \) is uniquely determined by its samples \( x[n] = x_c(nT), n = 0, \pm 1, \pm 2, \ldots \), if

\[
\omega_s = \frac{2\pi}{T} > 2\omega_N
\]

The frequency \( \omega_N \) is referred to as the Nyquist frequency, and the sampling rate \( (1/T) \) must be at least double this frequency in order for the samples to uniquely characterize the signal. More simply stated, a continuous-time signal can be exactly reconstructed from its samples if the signal contains no components at frequencies past half the sampling rate. In practice, it is usually preferable to sample several times faster than the Nyquist rate to ensure that several points are taken for each period of a sinusoid.

The consequence of sampling a signal at slower than the Nyquist rate is a phenomenon called aliasing in which the sampled waveform appears at a lower frequency than the original continuous-time signal. Figure 13.4 illustrates how any number of continuous-time frequencies can generate the same discrete-time sampled data points. Aliasing maps all continuous-time frequencies onto the range from zero to the Nyquist frequency (i.e., half the sampling rate). One way to understand aliasing is by viewing it in the frequency domain. Oppenheim et al. [23] provide a detailed derivation of the process that is summarized here. Sampling a continuous-time signal can be modeled as convolving the signal with a train of impulses evenly spaced at the sample time. The resulting multiplication in the frequency domain places a copy of the frequency spectrum of the sampled signal at each integer multiple of the sampling frequency. If \( x_c(t) \) is the original continuous-time waveform having a frequency content of \( X_c(\omega) \), then the sampled signal will have a spectrum given by

\[
X_P(\omega) = \frac{1}{T} \sum_{k=-\infty}^{+\infty} X_c(\omega - k\omega)
\]

If there is any frequency content in the original signal past the Nyquist frequency, then there will be overlap between the repeated copies. This overlap, or mapping of higher-frequency signals into lower ones, is the phenomenon of aliasing. Figure 13.5 illustrates this operation. In the discrete-time frequency domain,
FIGURE 13.4 Multiple continuous-time frequencies can generate the same set of discrete-time samples. The lowest frequency sinusoid that fits all of the points is the fundamental frequency, and all higher-frequency sinusoids are aliased down to this lower frequency between zero and half the sampling frequency.

All frequencies are mapped into the region of \( \pm \pi \text{ rad/sample} \). If the sampled signal is frequency-limited such that its highest component is less than half the sampling frequency, then the signal contained in the discrete-time frequency region of \( \pm \pi \text{ rad/sample} \) exactly captures the original continuous-time waveform. An ideal reconstruction filter passing only the initial copy of the components will reconstruct the original signal exactly. It is more common that the original waveform contains some components at frequencies above the Nyquist frequency, and the copies of these higher frequencies have wrapped down into the \( \pm \pi \) band and added to the spectral content there. When reconstructed the higher-frequency components modify the original waveform.

Aliasing alters the apparent frequency of a signal but does not change the amplitude. This is an important distinction and is a reason why aliasing can be so problematic. As an example, consider measurements taken of a system variable at 1000 Hz. An apparent 300 Hz component in that measured waveform could be at 300 Hz, but could also be at 700, 1300, 1700 Hz, etc. There is no way to identify an aliased waveform, and so appropriate anti-aliasing precautions must be taken before sampling the waveform. A properly designed system will guard against aliasing, but as sampled-data systems are inherent in computer control (whether as part of the feedback system or simply diagnostic monitoring), the design engineer should be familiar with its effects.

Discussions of sampling are usually limited to time-based systems, but other indices can be used as well. For example, position-based sampling can be used to link sampling events in multi-axis systems to the

FIGURE 13.5 The frequency-domain view of aliasing shows that sampling creates repeated copies of the continuous-time frequency spectrum at each multiple of the sampling frequency. Aliasing occurs at the overlap between the copies.
position of a particular axis. This position-based sampling may be referred to as position-synchronized output (PSO), position event generation (PEG), or alternate terms. The fundamental idea is that actions are based in an external non-time-based signal. The same rules apply for sampling and aliasing whether the trigger signal is time-based, position-based, or otherwise. Sampling must be performed at least twice as fast as the highest frequency component (or shortest period for position-based systems) in the signal.

There are some instances, particularly in metrology, when aliasing can be used to simplify a measurement. This is because aliasing changes frequency but does not alter magnitude. As an example, consider an accuracy measurement that is to be taken of a ballscrew-based system. Mapping short-wave errors (occurring within one rotation of the screw) would seem to require taking many points per revolution of the screw. This can be time-consuming, and even counter-productive when longer term temperature drifts in the environment are considered. Sampling at exactly the period of the screw (once per revolution) will alias the short-wave errors into constant values. This measures only the long-wave accuracy errors of the screw. Sampling even slower, longer than the period of the screw, aliases the higher-frequency intercycle errors into lower-frequency (longer period) values. This measurement shows the long-wavelength accuracy errors and the short-wavelength intercycle errors both with the same economical test setup.

### 13.3.2.2 Quantization

The term quantization refers to the process of taking continuous signals (with infinite resolution) and expressing them as a finite resolution number for processing by digital systems. There is always some error inherent in the conversion, and at some level, the error can make a meaningful difference in the accuracy of the measurement or performance of the servo loop. The common sources of quantization in measurements are reviewed further here.

#### 13.3.2.2.1 Analog-to-Digital Conversion

Most analog position sensors will be used with a digital control loop, and so the signal must be discretized with the use of an analog-to-digital (A/D) converter. To prevent aliasing, an analog low-pass filter should be placed before the A/D conversion that attenuates the signal content in all frequencies past one-half the sampling frequency. Typically the cutoff frequency of a practical analog filter is placed somewhat lower than this. The phase lag of the analog filter should always be included in the controller model. Analog-to-digital converters are rated in terms of bits of resolution where the minimum quanta size is given by

$$\Delta = \frac{\text{Range}}{2^{\text{bits}}}$$

The two most common in precision control systems are 12-bit (having 4096 counts) and 16-bit (having 65536 counts). The total range of travel divided by the number of counts gives the fundamental electrical resolution. The true analog value is always somewhere between counts, and so the quantization process adds some error to the measurement. This error can be approximated as a random signal, uniformly distributed over a single quanta with a variance of $\Delta^2/12$.

The fundamental A/D resolution can be improved by oversampling and averaging the A/D converter signal. Typically an A/D converter can sample at several hundred kilohertz while the position control loop only updates at several thousand Hertz. Rather than taking one sample per servo cycle, the software can therefore average several hundred readings if sufficient processing power is available. Averaging $N$ samples reduces the standard deviation of the signal by $\sqrt{N}$. This technique only works when the signal is sufficiently noisy to toggle between counts, otherwise a dither signal can be added to the sensor reading [5]. Typically however, the wiring in most systems absorbs enough electrical noise that adding anything additional is not required.

#### 13.3.2.2.2 Encoders and Incremental Position Sensors

Some types of position sensors produce inherently quantized output. Digital encoders are one example of these. For a square-wave output encoder, the resolution is set simply by the period of the lines. Encoders used in precision machines are often supplied with electronics that interpolate the resolution much lower than
the grating period. Typical quadrature decoders increase the resolution to four times the grating period, but other electronics are available that increase this to 4096 times or higher in compatible encoders. Encoders that have an amplified sinusoid output ideally have analog signals modeled by

\[
    v_a = V \sin \left( \frac{2\pi}{\lambda} \Delta x \right) \tag{13.9} \\
    v_b = V \cos \left( \frac{2\pi}{\lambda} \Delta x \right) \tag{13.10}
\]

where \( \lambda \) is the period of the encoder, and \( \Delta x \) is the displacement. When plotted vs. each other, they form a perfect circle,

\[
    v_a^2 + v_b^2 = V^2 \tag{13.11}
\]

The displacement within a single period of the scale (i.e., angle around the circle) can be found by taking an inverse-tangent,

\[
    \Delta x = \tan^{-1} \left( \frac{v_a}{v_b} \right) \tag{13.12}
\]

which is usually done through the use of a software lookup table parametric on \( v_a \) and \( v_b \). The resolution of the A/D converter and size of the lookup table determine the new size of the position quanta.

There are errors in the practical application of encoder multiplication. The two analog sinusoid signals will have offsets in their DC level, gains, and phase. A more realistic representation of the encoder signals are

\[
    v_a = V_a \sin \left( \frac{2\pi}{\lambda} \Delta x \right) + a_0 + \text{noise} \tag{13.13} \\
    v_b = V_b \cos \left( \frac{2\pi}{\lambda} \Delta x + \phi \right) + b_0 + \text{noise} \tag{13.14}
\]

where each signal can have a DC offset, a different gain, and a nonorthogonal phase shift. Depending on the quality of the encoder, there may also be higher-order sinusoids present in the signal. When plotted against each other, these signals no longer form a perfect circle, but the shape is instead primarily a fuzzy off-axis ellipse. Higher-order sinusoids can create multi-lobed shapes. These signals are usually functions of alignment and analog circuit tuning and may change with position along the encoder scale. Higher-quality scales that are installed with careful attention to mounting details usually will have patterns that are more stable and circular, allowing higher levels of interpolation.

Interpolation errors are modeled as a sinusoidal noise source at speed-dependent frequencies. Offset errors in the multiplication occur at a frequency equal to the speed divided by the grating period, and errors in gain result in twice the frequency. For example, poor multiplier tuning in a stage with a 4 \( \mu \)m grating period traveling at 10 mm/s results in noise disturbances at 2500 and 5000 Hz. On typical systems, these disturbance frequencies can easily be several times the sampling frequency of the controller, meaning that they will alias down to below the Nyquist frequency. If this frequency falls well above the servo bandwidth of the stage, the noise source will appear in the frequency spectrum of the position error while scanning, if it falls within the servo bandwidth, it will not appear at all since the stage is tracking it. It will only show when the stage position is measured with a secondary sensor. There is no equivalent to the analog anti-aliasing filter used with A/D converters here. The best method to reduce this error is by carefully tuning the multiplier. Alternatively, the encoder signal can be sampled several times faster than the servo bandwidth and averaged for each update. This is more computationally intensive but has the effect of increasing the Nyquist frequency for this process.
13.3.2.2.3 Digital-to-Analog Conversion

A reconstruction filter must be used to generate a continuous-time waveform from discrete-time samples, and the characteristics of this filter affect the accuracy of the reconstruction. The ideal reconstruction filter is a low-pass filter that exactly passes all frequency components less than the Nyquist frequency and exactly blocks all frequencies higher. Such a filter is not realizable. The most common method to reconstruct a continuous-time waveform is through a digital-to-analog converter that holds an output at a discrete level for the duration of a sample period and changes at the next sample instant. This type of filter is referred to as a zero-order-hold (ZOH) reconstruction filter, and this filter adds some error to the reconstructed signal [22]. The ZOH reconstruction filter has an equivalent continuous-time frequency response of

\[ H_{ZOH}(j\omega) = Te^{-j\pi} \frac{\sin(\omega T/2)}{\omega T/2} \]  \hspace{1cm} (13.15)

with magnitude and phase as shown in Figure 13.6. The sharp edges in the “staircase” output are the higher-frequency components in the signal at amplitudes indicated by the sidelobes of the frequency response.

The effect of the ZOH reconstruction filter is to include higher-frequency components in the spectrum of the output signal. As an example, consider the case of a 200 Hz sinusoid written through a D/A converter at a 1 kHz rate. Figure 13.7 shows the resulting waveform. The ideal reconstruction filter would create a waveform that is exactly

\[ v_{\text{ideal}}(t) = \sin(2\pi 200t) \]  \hspace{1cm} (13.16)

However, according to Equation (13.15), the signal produced by the ZOH reconstruction filter is instead

\[ v_{\text{ZOH}}(t) = 0.9355 \sin(2\pi 200t - 36^\circ) + 0.2339 \sin(2\pi 800t - 144^\circ) + H.F.T \]  \hspace{1cm} (13.17)

\[ \begin{align*}
|H_{ZOH}(j\omega)| & \approx 1 & \text{for } |\omega| \leq \frac{\pi}{T} \\
\angle H_{ZOH}(j\omega) & \approx -90^\circ & \text{for } |\omega| \leq \frac{\pi}{T}
\end{align*} \]

FIGURE 13.6 The magnitude and phase of the zero-order-hold reconstruction filter are significantly different from the ideal reconstruction filter, particularly as the frequency approaches the Nyquist frequency.
FIGURE 13.7 The zero-order-hold reconstruction filter creates a “stairstep” version of a signal that can be significantly distorted as compared with a continuous-time signal.

The difference between the target and actual outputs can be significant for precision machines, and a compensation filter may be required to pre-condition the discrete-time signal before sending it to the D/A converter. Note also in this case that the 800 Hz component of the output signal is above the Nyquist frequency. This command signal (presumably sent to the amplifiers) will generate motion at 800 Hz, but the 1000 Hz sampling frequency will alias 800 Hz down to 200 Hz, corrupting the measurement of actual motion of the stage at 200 Hz. A clear understanding of all sampling and reconstruction filters is usually needed to interpret the frequency content of signals that transfer between analog and digital domains.

13.3.3 Precision Mechanics

Precision mechatronic systems must begin with precision mechanics. Controls and electronics can be used to correct for some errors, but the controls problem is always easier when the mechanics are well-behaved (meaning repeatable and readily modeled). Some of the main components of the mechanics are the bearings, the machine structure itself, and the vibration isolation system.

13.3.3.1 Linear and Rotary Bearings

The bearings of a precision machine are the most critical element that defines the performance of a machine. Most bearings (or sets of bearings) are designed to constrain all but one degree of freedom of motion, and bearings of all types are used in precision machines. Many of the types of bearings are available in linear and rotary versions. The designer must choose a bearing based on its load carrying capability, stiffness, repeatability and resolution (the ability to move in small increments), friction, size, and cost. Slocum [34] provides a good overview of different bearing types, their advantages, limitations, and preferred uses.
13.3.3.2 Machine Structure

The overall machine structure and vibration isolation system must be considered as part of a precision machine design and installation. The primary concern in designing the structure of the machine is to provide a dimensionally-stable base with well-characterized, well-damped resonant modes. Vibrations of these modes can enter the feedback loop and either destabilize the system or, more typically, add an extra mode to the response that extends the settling time for a move. Achieving high stiffness is relatively quantitative, particular with the use of finite element models for analysis. However, achieving high damping is equally important in attenuating the influence of the structural modes. Most engineering materials (metals usually) have relatively little internal damping, and so damping must be explicitly added. Riven [26], Nayfeh [21], and Marsh and Slocum [20] detail methods for designing specific damping elements into structures. Jones [16] concentrates in a specific family of damping techniques using viscoelastic polymers. A common use of these materials is in constrained layer dampers in which a sheet of damping material is sandwiched between a structural member and a stiff constraining layer. Any vibration of the structural member shears and strains the viscoelastic material, thus creating a loss mechanism for the energy at the vibration frequency. These damping techniques are often applied in an attempt to fix existing problematic designs, but with mixed success. It is preferable to address damping in the mechanical elements at the earliest possible stage of the design.

13.3.3.3 Vibration Isolation

Vibration isolation systems are used primarily to attenuate the influence of ground-borne vibrations on the position stability of a precision machine. In other cases, the vibration isolation is to present movements of the machine itself from detrimentally influencing surrounding processes. Riven [25] and DeBra [6] provide detailed overviews of vibration isolation of precision machines. It should be noted that this is an active research area with significant publication activity. The two general types of isolation, passive and active, differ mainly on whether direct measurement of the vibration is used to help attenuate it. Passive isolation vibration isolation systems are usually chosen based on their natural frequency and damping level. Riven [27] details the design and application of passive isolation devices. Active isolation systems are generally more complex and require one of more vibration sensors (usually accelerometers or geophones) to measure payload and ground vibrations then apply forces to the payload to oppose this motion [12]. For cost reasons, passive systems are almost always preferable, but all-passive systems have a fundamental limitation that active systems can overcome.

The fundamentals of the problem can be seen with a conceptual single-degree-of-freedom model. Consider the case of a free mass, representing the machine base, attached to the ground with a combination spring-damper. To provide isolation from ground-based vibrations, the spring and damper should be made as soft as possible. However, there are also disturbance forces applied directly to the machine base. These forces are usually reaction forces created by motion of the devices attached to the base. Keeping the base stationary under these forces requires that the spring-damper system be as rigid as possible. The two requirements are in opposition and can be expressed mathematically as sensitivity and complementary sensitivity functions. This means they always add to unity, and any improvement in rejection of disturbance forces comes exactly at the cost of reduced rejection of ground vibrations. There is no way to adjust the impedance of a passive mount to improve the isolation from both sources. Isolating from ground vibrations requires a soft mount; isolating from disturbance forces requires a stiff mount. Active isolation systems do not have this same limitation since they are able to measure ground and payload vibration directly and (in-effect) adjust the instantaneous impedance of the mount as conditions require.

Most sources of vibration can be characterized as a summation of single-frequency sinusoids (from various pumps and motors in a facility) and random vibrations with a given spectral density. One common set of standards for characterizing the level of seismic vibration in a facility is the Bolt Beranek & Newman (BBN) criteria [13]. Their study presents vibration levels as a series of third-octave curves plotting RMS velocity levels vs. frequency. The curve levels range from VC–A, suitable for low-power optical microscopes, to VC–E, the highest level presumed to be adequate for the most-demanding applications. Most precision machines are designed to be placed in existing facilities, and so whenever possible, it is preferred to take a
sample measurement of ground vibrations directly. Note that the ground does not just vibrate vertically, but laterally as well. Several measurements are usually necessary since vibration levels can vary greatly at different locations and at different times. These measured-vibration levels can be characterized and used as the input to a model of the isolation system. The response to the deterministic (single-frequency) elements of the disturbance can be predicted directly, but only a statistical characterization of the motion is possible given a random input. Wirsching et al. [38] detail techniques required to model the response of dynamic systems to random vibrations.

13.3.4 Controller Implementation

Dedicated electronics and hardware are required to implement the algorithms that control precision machines. The hardware usually consists of a top-level controller, power amplifiers, actuators, and position feedback sensors. Appropriate software and control algorithms complete the system design. These elements of a mechatronic design are as critical to the overall performance of a system as are the mechanics themselves.

13.3.4.1 Feedback Control Hardware

Microprocessor-based controllers implement the feedback compensation algorithms for most precision machines. Position is usually the controlled variable, but force, pressure, velocity, acceleration, or any other measured variable may also be controlled. The controller may be a standalone unit or may require a host PC. Commercially available systems differ in the number of axes they are able to control, the amount of additional I/O they can read and write to, their expandability, methods of communication to other electronic hardware, and as always, cost. In general, the analog circuitry in these controllers is being replaced by digital or microprocessor-based circuits. Analyzing and troubleshooting these systems require dedicated software routines rather than a simple oscilloscope, and the availability of such an interface should also factor into the choice of a controller. The selection of an appropriate motion controller may be driven by the need to be compatible with existing equipment, existing software, or existing expertise within a group. As is generally the case with microprocessor-based systems, the features available increase so rapidly that the system design engineer should always consult the controller manufacturers frequently. For unique applications, the best alternative may be to select a general-purpose processor and write the low level code required for implementing the controller, but this should usually be viewed as a last resort.

13.3.4.2 Power Amplifiers

Power amplifiers convert the low-level signals from the feedback compensation into high-power signals suitable for driving the actuators. Power amplifiers for electromagnetic actuators usually take the form of a current loop in which the command is a scaled representation of the desired actuator current. A proportional or proportional-integral control loop adjusts the voltage applied to the motor as required to maintain the commanded current. The voltage available to drive the current is limited by the amplifier bus rail and the speed-dependent back-emf voltage of the motor. Most higher-power actuators and electronics are multi-phase, typically three-phase, and so there will be three such compensators inside each amplifier. Power amplifiers have traditionally been designed with analog compensation for the current loop. Passive elements (resistors and capacitors) used to set the loop gains may be mounted to a removable board, also known as a personality module, to accommodate motors with different values of resistance and inductance. Potentiometers may also be included to allow fine-tuning of the control, but setting these can be difficult to quantify, and the setting may drift over time. Power amplifiers are increasingly moving to digital control loops, in which a microprocessor implements the control algorithm. A key advantage here is that the gain settings can be rapidly and repeatably matched for the desired performance of the current loop to the exact motor used.

Two basic types of power amplifiers are available for a machine designer to select from. Linear amplifiers are typically very low-noise, but larger and less power-efficient than comparable switching amplifiers. Switching amplifiers, also known as pulse-width-modulation (PWM) amplifiers operate by rapidly switching the voltage applied to the actuator between zero and full-scale. The duty cycle of the switching sets
the average level of current flowing through the actuator, and the switching typically occurs at 20–50 kHz. PWM amplifiers rely on motor inductance to filter the rapidly changing voltage to a near-continuous current. Some level of ripple current will almost always be present in PWM systems.

In either case, the system designer should characterize the performance of the power amplifier. The amplifier should have a high-enough bandwidth on the current loop to track the required commands with an acceptable level of phase lag, but not so high that excessive high-frequency noise amplification occurs. The system designer should also monitor the fidelity of the amplified signal as it passes through zero amps. This is typically a handoff point in the amplifiers, where conduction passes from one set of transistors to another, and is a potential source of error.

### 13.3.4.3 Actuators

Precision systems use a variety of actuators to convert the amplifier commands into forces and motion. Brushless permanent-magnet motors, in linear and rotary forms, are the most common. These motors are usually three-phase systems and require a supervisory controller for commutation to ensure continuous motion. The system designer can select between iron-core and ironless motors in an application. Iron core motors offer higher force (or torque) density than do ironless models, but the magnetic lamination leads to cogging that affects the smoothness of the applied force. As usual, there is no best choice for all applications. Other types of actuators include voice-coil type motors, stepper motors, and piezoelectric drives.

### 13.3.5 Feedback Sensors

The selection, mounting, and location of the sensors are critical to the successful design of a motion control system. A controller with poor feedback devices will be unable to perform well regardless of the care taken in the mechanical design or the sophistication of the control algorithms. The most common sensors used in precision machines are encoders, laser interferometers, and a variety of analog feedback devices.

#### 13.3.5.1 Rotary Encoders

Rotary encoders are used to measure the angular displacement of rotary tables and motors. They are inherently incremental devices and usually need some form of a reference marker to establish a home position. The primary source of errors in rotary encoders is misalignment between the optical center of the grating pattern and the geometric center of the axis of rotation. The angular errors caused by decentering can be estimated by

$$
\epsilon = \pm 412 \frac{e}{D}
$$

where $\epsilon$ is the angular error in arc-sec (1 arc-sec $\approx 4.85$ $\mu$rad), $e$ is the centering error in microns, and $D$ is the diameter of the grating pattern in millimeters. A 5 $\mu$m centering error creates an angular error of 20 arc-sec on a 100 mm diameter encoder. This error is repeatable and can be mapped and corrected for in software. When used with a ballscrew, a rotary encoder can be used to indicate linear position, but allowances must be made in the error budget for changes to the lead of the screw over the range of operating temperatures. An alternative to software correction for the decentering error is to use additional encoder readheads. A second readhead, placed 180° opposite the first, reverses the decentering error. Averaging the signals from the two readheads cancels it. Reversal techniques abound in precision machine design [10], and the system design engineer should be familiar with them.

#### 13.3.5.2 Linear Encoders

Linear encoders are somewhat easier to mount and align than are rotary encoders because the fundamental mounting error is to misalign the encoder with the direction of travel. This results in a cosine error and a resulting linear scale factor error. Once this factor is known, it is a simple scaling to correct in software. The more difficult problem is to reliably mount the scale to still perform properly under thermal changes. Scales made of zero-expansion glass are available, but they are typically mounted to stages that do expand and contract. In this case, the designer should compensate for the mismatch by fixing a point on the scale,
preferably at some home signal, and allowing the scale to expand from there. The opposite case is to use a
scale that adheres to the substrate firmly enough to expand and contract with the structure. Allowing for
growth is not a problem, but it can be difficult to choose a fixed point about which growth occurs to use
in thermal error mapping.

13.3.5.3 Laser Interferometers
Laser interferometers measure displacement by monitoring the interference between a reference beam, and
one reflected from the target. Most interferometers use a stabilized Helium-Neon source with a wavelength
of 632.2 nm as their basis. Linear displacement measuring interferometers are generally classified as
either homodyne or heterodyne types. Homodyne interferometers use a single-frequency laser with optics
that interfere a reference beam with one reflected from a moving target. The resulting constructive and
destructive interference effectively creates a position-dependent periodic intensity that can be converted
into an electrical signal with a set of photodiodes. To the end-user, the resulting sinusoidal signals (in
quadrature) are treated no differently from the output of an analog encoder. Heterodyne interferometers
use a two-frequency laser. One of the frequencies is diverted to a stationary reference, and the other
frequency is reflected off the moving target. The frequency of the measurement beam is Doppler shifted
by the velocity of the target on its return. This frequency shift is measured as a phase difference in the beat
frequency produced when the measurement and reference beams are recombined. This phase difference
can be accurately measured and converted to a velocity and displacement. Both categories, homodyne
and heterodyne, have their fervent critics and supporters. In either case, sub-nanometer resolutions with
sub-micron absolute accuracy are possible.

There are several major advantages to laser interferometers over linear encoders. The measurement is
potentially very accurate due to its basis in the frequency of the laser beam. Calibration with a laser interfer-
ometer is generally used as the final confirmation of the accuracy of a machine. The measurement can also
be made very close to the tool or workpiece, thus reducing or even eliminating Abbé errors. The noncontact
nature of the measurement makes it easier to separate the force and metrology loops in the design.

The performance of a laser interferometer is usually limited by the environment that it operates in.
Changes in air temperature, pressure, and humidity affect the index of refraction in the air and, thus,
the wavelength of light and accuracy of measurement. Bulk changes can be compensated for with Edlens
Equation [8,36], but localized changes (due to air turbulence) are nearly impossible to track. The best
environment in which to operate a laser interferometer is a vacuum. When a vacuum is not practical, the
beams should be routed through tubes that guard against air currents, and the free path of the beam kept
as short as possible. The beam path of a laser interferometer is part of the metrology loop, and so errors
in this measurement are not discernable in the feedback loop. In other cases, particularly when the stage
is moving, uncertainty in the time of the measurement limits the overall accuracy. The interferometer can
say where the stage was, but not precisely when it was there or where it may be now. Systems attempting
for nanometer-level accuracy must incorporate this data age uncertainly into the error budget. Correctly
applying a laser interferometer feedback system to a precision machine requires careful attention to detail,
but in most cases, it provides the highest-resolution, highest-accuracy measurement possible.

13.3.5.4 Analog Sensors
Analog position sensors are frequently used in limited-travel applications and can resolve motion to less
than 1 nm with appropriate electronics. Common types of sensors include capacitance gages, eddy-current
sensors, and LVDTs. Their relatively small size and limited range make them a natural match to flexure-
based mechanisms. Resolution is a function of the measurement bandwidth in analog systems, and the
sensor noise floor should never be quoted without also noting the bandwidth at which it applies. Assuming
that the noise floor of an analog sensor is relatively constant over all frequencies (i.e., white noise), then
the standard deviation of this noise reduces with the square root of the bandwidth of the measurement.
Figure 13.8 demonstrates the results of this operation. An analog sensor with a 3σ resolution of 10 nm
at a 1 kHz bandwidth will have a resolution of 3.2 nm at 100 Hz, and 1 nm at 10 Hz. This increase
in resolution increases the phase shift of the feedback system and possibly will require a reduction in
controller loop bandwidth. If the noise present in the sensor measurement is essentially random, the sensor resolution can be expressed as a spectral density and quoted as nm/√Hz. The designer can then match the appropriate filter bandwidth to his resolution requirements. However, a caution here is that many sensors have inherently higher levels of noise at the lower frequencies. The amount of filtering required to achieve a resolution may be higher than expected. Analog sensors are usually supplied with signal conditioning electronics that convert the natural output of the sensor to a standard form (typically ±10 V). These electronics should be placed as close as possible to the sensor because the low-level signals in the raw sensor output are often highly susceptible to electronic noise. The conditioned signal is usually somewhat less sensitive but should still be shielded and routed as far as possible from high-power conductors. Analog sensor accuracy and linearity is entirely distinct from the resolution specification. Sensors may be highly linear about an operating point but often become increasingly nonlinear over their full range of travel. This nonlinearity can be compensated for with a table or curve fit in software. Sensor manufacturers may also provide a linearized output from their electronics package. This is very useful for metrology, but a designer should confirm the update rate of this output before using for feedback. Often this additional output is no longer a native analog signal, but it may have been quantized, corrected, and re-synthesized through a digital-to-analog converter. All of these processes add time and phase lag to the measurement making it less suitable for use as a feedback device.

13.3.6 Control Algorithms

The application of control algorithms separates mechatronics systems and mechatronic design from traditional electromechanical systems. Modern controllers do more than simply close a PID loop. They provide additional options on feedback filtering and feedforward control, allow for coordination between multiple axes, and should provide an array of tuning and data collection utilities. An overview of control algorithms used with precision machines is provided here.

13.3.6.1 System Modeling

Dynamic models of the system should be developed for creating the controller and verifying its performance. Each controller design should begin with a simple rigid-body model of the overall structure.
This simplified model is justified when, at a typical closed-loop bandwidth for mechanical systems of 10–100 Hz, friction is low and any structural resonances are located at significantly higher frequencies. While not complete, the simplified model shows the bulk rigid-body motion that can be expected given the mass of the stage, the selection of motors and amplifiers, and the types of commanded moves. It can also be used to generate a more-accurate estimation of motor currents (and power requirements) for typical machine moves. Figure 13.9 shows a block diagram of a simplified linear-motor-driven system. The voltage from the analog output of the controller is converted into current by the amplifiers and into a force by the linear motors. In this first model, we assume no dynamics in the amplifiers and that commutation is implemented properly to eliminate force ripple. The motor force accelerates the stage mass, and an encoder measures the resulting stage position. The continuous-time transfer function from control-effort to stage position (in counts) is therefore

$$P(s) = \frac{k}{s(ms + b)}$$  \hspace{1cm} (13.19)

and this transfer function can be used to generate controller gains. Scaling factors will always be specific to the supplier of the control electronics, and it may take some research to find these exactly.

The simplified model is adequate for the initial generation of gain parameters, but determining performance at the nanometer level always involves augmenting the model with additional information. This advanced model should include the discrete-time controller, quantized position feedback and control effort, higher order dynamics in the amplifiers, higher-frequency mechanical modes (obtained either from a finite-element study or experimentally through a modal analysis), force ripple in the motors, error in the feedback device, and friction and process-force effects. Some can be predicted analytically before building the system, but all should be measured analytically as the parts of the system come together to refine the model.

### 13.3.6.2 Feedback Compensation

Commercially available controllers generally use some form of classical proportional-integral-derivative (PID) control as their primary control architecture. This PID controller is usually coupled with one or more second-order digital filters to allow the user to design notch and low-pass filters for loop-shaping requirements. In continuous-time notation, the preferred form of the PID compensator is

$$C(s) = K_p \left(1 + T_D s + \frac{1}{T_I s}\right) = K_p \frac{T_D T_I s^2 + T_I s + 1}{T_I s}$$  \hspace{1cm} (13.20)

Separating the proportional term by placing it in front of the compensator allows the control engineer to compensate for the most common system changes (payload, actuator, and sensor resolution) by changing a single gain term ($K_p$). Techniques for designing continuous-time PID loops are well documented in almost any undergraduate-level feedback controls textbook. The preferred method for designing, analyzing, and characterizing servo system performance is by measuring loop-transmissions, otherwise known as open-loop transfer functions. This measure quantitatively expresses the bandwidth of the system in terms of the crossover frequency and the damping in terms of the phase margin.
13.3.6.3 Feedforward Compensation

Feedforward compensation uses a quantitative model of the plant dynamics to prefilter a command to improve tracking of a reference signal. Feedforward control is typically not studied as often as feedback control, but it can cause tremendous differences in the ability of a system to follow a position command. Feedforward control differs from feedback in that feedforward control does not affect system stability. If a perfect model of the plant were available with precisely known initial conditions, then the feedforward filter would conceptually consist of an inverse model of the plant. Preshaping the reference signal with this inverse model before commanding it to the plant will effectively lead to unity gain (perfect tracking) at all frequencies. If properly tuned, the feedback system is required only to compensate for modeling inaccuracies in the feedforward filter and to respond to external disturbances. In practical implementations, the plant cannot be exactly inverted over all frequencies, but the feedforward gain will instead boost tracking ability at low frequencies.

Formal techniques exist for creating feedforward filters for generalized discrete-time plants. Tomizuka [37] developed the zero phase error tracking controller (ZPETC) that inverts the minimum-phase zeroes of a plant model and cancels the phase shift of the nonminimum phase components. Nonminimum phase zeroes are a common occurrence in discrete-time systems. The ZPETC technique begins with a discrete-time model of the closed-loop system (plant and controller),

\[ G_p(z^{-1}) = \frac{z^{-d} B_c^-(z^{-1}) B_c^+(z^{-1})}{A_c(z^{-1})} \]  

in which the numerator is factored in minimum-phase zeros \( B_c^+ \) and non-minimum-phase zeros \( B_c^- \). The \( z^{-d} \) term is the \( d \)-step delay incurred due to computation time or any phase lag. The ZPETC filter for this system becomes

\[ G_{ZPETC}(z^{-1}) = \frac{z^d A_c(z^{-1}) B_c^-(z)}{B_c^+(z^{-1}) B_c^-(1)^2} \]  

which cancels the poles and minimum-phase zeros of the system, and cancels the phase of the nonminimum phase zeros. This is a noncausal filter, but it is made causal with a look ahead from the trajectory generator. In multi-axis machine, it is generally allowable to delay the position command by a few servo samples provided the commands to all axes are delayed equally and the motions remain synchronized.

Feedforward control techniques are also available that modify the input command to cancel trajectory-induced vibrations. One such technique, trademarked under the name Input Shaping was developed by Singer and Seering [31] at MIT. Input Shaping and related techniques, convolve the input trajectory with a series of carefully timed and scaled impulses. These impulses are timed to be spaced by a half-period of the unwanted oscillation and are scaled so that the oscillation produced by the first impulse is cancelled by the second impulse. Much of the research work in this area has involved making the filter robust to changes in the oscillation frequency. The downside to this technique, in very general terms, is that the time allowed for the rigid-body move must be lengthened by one-half period of the oscillation to be cancelled.

Specialized feedforward techniques can be applied when the commanded move is cyclic and repetitive. In this case, the move command can be expressed as a Fourier series and decomposed into a summation of sinusoidal components [19]. The steady-state response of a linear system to a sinusoid is well defined as a magnitude and phase shift of the input signal, and so each sinusoidal component of the command can be pre-shifted by the required amount such that the actual movement matches the required trajectory.

13.4 Conclusions

Precision positioning of rotary and linear systems requires a mechatronic approach to the overall system design. The philosophy of determinism in machine design holds that the performance of the system can be predicted using familiar engineering principles. In this chapter, we have presented some of the components that enter into a deterministic precision machine design. Concepts such as alignment errors,
force and metrology loops, and kinematic constraint guide the design of repeatable and analytic mechanics. Often-neglected elements in the mechanical design are the cable management system, heat transfer and thermal management, environmental protection (protecting system from the environment it operates in and vice-versa), and inevitable serviceability and maintainance. Mechatronic systems require precision mechanics; high-quality sensors, actuators, and amplifiers; and a supervisory controller. This controller is usually implemented with a microprocessor, and the discrete-time effects of sampling, aliasing, and quantization must be accounted for by the system design engineer. Finally, the feedback control algorithms must be defined, implemented, and tested. The elements used to create precision mechatronic systems are widely available, and it is left to the designer to choose among them trading complexity and cost between mechanics, electronics, and software.

References

[1] The American Society for Precision Engineering, P.O. Box 10826, Raleigh, NC 27605-0826 USA.