ACCURACY IN POSITIONING SYSTEMS

Kevin McCarthy
New England Affiliated Technologies

The state of the art in precision positioning systems has undergone continuing improvement, with the result that modern positioning systems can now achieve unprecedented levels of accuracy. These gains have come about due to specific technical advances (most notably, the availability of coherent light sources) as well as inexorable pressure from high-tech applications which depend on dimensional accuracy for their existence. Notwithstanding the gains that have been made, there are gaps between levels of accuracy which are perceived as achievable, and those levels which can actually (and/or affordably) be met. This paper will attempt to address the realistic accuracy levels which various positioning technologies can meet, as well as the nature of the limitations which restrict accuracy.

WHAT IS "ACCURACY"?

Dimensional accuracy is simply the degree to which displacements executed by a positioning system match agreed upon standards of length. Ultimately, all length measurements are tied to the meter, as defined by the Committee Consultif pour Definition du Meter. Its current value is the distance which light in a vacuum travels in 1/299,792,458 of a second. When describing accuracy, we employ a variety of units considerably smaller than a meter. These include the familiar millimeter (10^{-3} meter), micron (10^{-6} meter), nanometer (10^{-9} meter), Angstrom (10^{-10} meter) and picometer (10^{-12} meter). For comparison purposes, a human hair is about 100 microns in diameter, semiconductor line widths are about 1 micron, and an atom is about 1 Angstrom.

"FUZZ" vs. "BUNK"

The heading, while somewhat jocular in nature, reflects a widespread lack of seriousness with respect to accuracy claims. Positioning system purchasers prefer that accuracy be summarized in a single, easily digestible number (and the smaller, the better). Positioning system vendors, in turn, comply; the unfortunate results include a recent full page ad which claimed to extract "tenth micron accuracy" from an open loop stepper based system. When questioned, an applications engineer responded that they were using a 1 mm leadscrew, and a divide-by-50 microstepper; hence, "tenth micron accuracy". Examples such as these reflect either a profound lack of awareness of the meaning and limitations of high accuracy systems ("fuzz"), or an overly aggressive marketing of "small numbers" for competitive advantage ("bunk"). We regularly find that our tables improve dramatically (were the literature to be believed) upon their incorporation into other firms' products. Common practices include defining table accuracy as equal to that of the purchased leadscrew incorporated in the table, ignoring thermal factors and Abbé error; mentioning the accuracy of multi-axis systems without a "per axis" qualifier; providing accuracy values which reflect only the no-load value, etc. The fact of the matter is that accuracy is a global parameter, which is affected by a combination of positioning table attributes; control and feedback systems; application specific details (e.g., the height above the table of the point of interest); as well as the operating environment. A meaningful characterization of system accuracy is better achieved by a complete analysis than by an attention grabbing "number".

THE PRIOR ART

Many of today's applications for high accuracy positioning systems are tied to the requirements of the semiconductor industry and inspection systems for ultra-precise machined parts. Over a hundred years ago, however, scientists and technicians were busy creating X-Y tables with surprising accuracy, given the tools at their disposal. At that time, the challenge was the ruling of large precise diffraction gratings for spectroscopy, and the positioning tables were referred to as ruling engines.

The design and fabrication of these ruling engines was a herculean effort, and the history of their development is replete with decade-long attempts which met with failure. Henry Rowland produced several engines capable of ruling acceptable four inch gratings in the 1880's; Professor Michelson (of interferometer fame), labored unsuccessfully from 1900 to 1930 to extend the useful travel to twelve inches. Colleagues who sought the ruling engine designs of H.J. Grayson upon his death were shocked to learn that his widow had promptly burnt them, perhaps in response to the all-consuming monomania to which ruling engine refinement drove its designers. Albert Ingalls has written an article chronicling the development of these instruments.[1]

Many of the physical factors which tormented ruling engine developers live on to harass present day positioning equipment vendors. Among these are temperature effects, friction, wear, internal stress-warpage, flexure, and vibration. Moreover, few customers are content with delivery times quoted in terms of decades (if then)! Fortunately, high accuracy feedback systems available today avoid the need for much of the obsessive mechanical design required of the open loop ruling engines. As an example of the pains which were taken to produce acceptable gratings, consider that the ruling engine John Anderson operated at Johns Hopkins University required 2½ hours to achieve thermal stability, and an additional 15 hours for the lubricant films to become uniform before ruling could commence. Many of the process and design principles (for example, techniques for ultra-precise lapping of lead screws) found in these ruling engines have since been incorporated into modern high accuracy positioning equipment. In fact, one large wafer-stepper firm was a direct descendent of a ruling engine manufacturer.

The development of replication processes led to low cost replica gratings, and sounded the death knell to the fledgling ruling engine market.

WAY ACCURACY

Positioning system accuracy can be conveniently divided into two categories: 1) the accuracy of the way itself, and 2) the linear positioning accuracy along the way. The former describes the degree to which the ways (ball and rod, crossed roller, air bearing, etc.) provide an ideal single axis translation, while the latter is concerned with the precision of incremental motion along the axis (typically related to the leadscrew, linear encoder, or other feedback device).

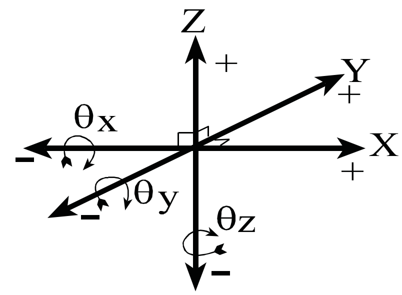


Figure 2: Six Degrees of Freedom

Any moving object has six available degrees of freedom (Fig. 2). These consist of translation, or linear movement along any of three perpendicular axes X, Y, and Z, as well as rotation around any of those axes (Ox, Oy, Oz). The function of a linear positioning way is to precisely constrain the movement of an object to a single translational axis (typically described as the X axis). Any deviations from ideal straight line motion along the X axis are the result of inaccuracy in the way assembly.

There are five possible types of way inaccuracy corresponding to the five remaining degrees of freedom (Fig. 3): translation in the Y axis; translation in the Z axis; rotation around the X axis (roll); rotation around the Y axis (pitch); and rotation around the Z axis (yaw). Since there are interrelations between these errors (angular rotation, for example, produces a transitional error at any point other than the center of rotation), it is worthwhile to carefully examine the effects of each type of error and its method of measurement.

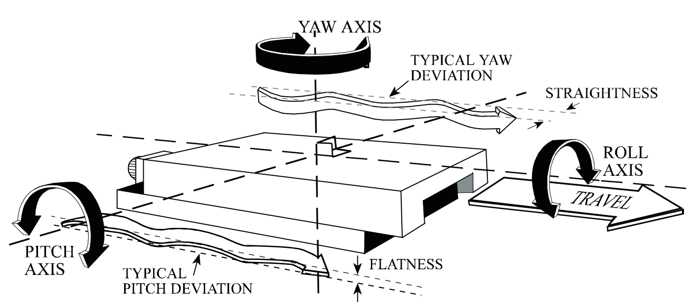


Figure 3: Way Errors

Since all useful methods of producing linear motion average over a number of points (due to multiple balls or rollers, or the area of an air bearing), "pure" transitional errors from straight line motion (that is, without any angular error) are usually minor.

Positioning tables do, nonetheless, exhibit some vertical and horizontal run out (typically referred to as errors of flatness and straightness, respectively), as can be measured by placing a sufficiently sensitive indicator on a table and measuring the vertical or horizontal displacement along its travel. With the following exception, however, these transitional errors are the consequence of underlying angular errors, as described below.

In the example of figure 4, the ways are perfectly straight and allow only translation along a single axis. Since, however, our desired X axis of motion is usually defined as parallel to the base of the table, and the ways are inclined relative to that base, the indicator will see a rise and fall as the table travels back and forth. While the ways may be ideal, their orientation within the stage can result in translation along the Z axis (also called vertical runout, or an error of flatness). There is no basis for a corresponding effect in the Y axis since the exterior sides of positioning tables are not commonly assumed to include a reference surface.

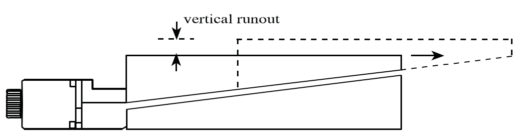


Figure 4: Vertical Runout

The angular errors of roll, pitch, and yaw (Ox, Oy, and Oz, respectively) are always present at some level in positioning tables and degrade performance in several ways. Their direct effect is to vary the angular orientation of a user payload. Due to the relative care with which these errors can be maintained at low levels (2-40 arc seconds), they are of little consequence 'in many applications. Certain optical positioning tasks, however, may be directly impacted by angular errors.

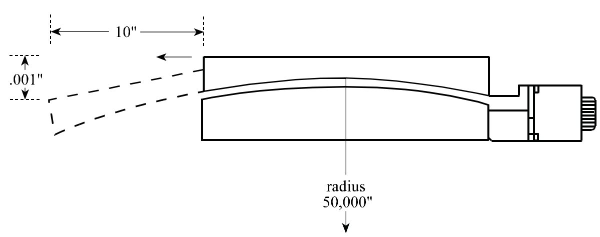


Figure 5: Pitch Error

Of somewhat greater concern are the translational errors resulting from underlying angular errors. The simple pitch error shown in Fig. 5, corresponding to a radius of curvature of 50,000 inches, will produce a Z axis translation of .001" in a 20" travel stage at either end of travel, relative to its centered position. Such simple pitch errors are typically found in non-recirculating table designs, due to the overhanging nature of the load at both extremes of travel. More complex curvatures involving roll, pitch, and yaw, as well as multiple centers of curvature, can also be encountered.

The worst impact of angular errors is the resulting Abbé (offset) error which affects linear positioning accuracy. Unlike the simple translational error described in the above example, Abbé error increases as the distance between the precision determining element and the measurement point increases. This effect is described in detail below.

RESOLUTION AND REPEATABILITY

Together with accuracy, these three terms are the fundamental parameters of positioning systems. Unfortunately, they are often used synonymously with resulting confusion on the part of users and vendors alike.

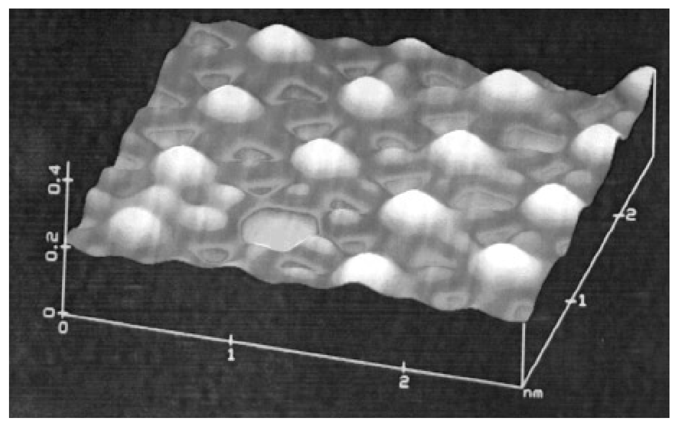
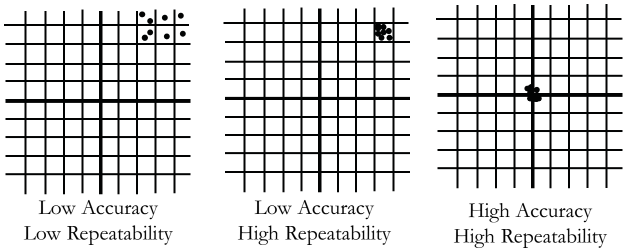


Figure 6: STM Image of Iodine Atoms

Resolution is frequently defined as the smallest positional increment which can be commanded of a system; a more rigorous definition would modify this to reflect the smallest positional increment which can be realized. Open loop or rotary-encoded servo systems are capable (depending on leadscrew pitch) of providing useful resolutions of as low as 0.1 micron. The use of a linear feedback transducer, together with a servo loop incorporating an integrator (the "I" in P-l-D), allows useful resolutions below 0.01 microns (10 nanometers).

Perhaps the ultimate level of positioning resolution has been achieved in the Scanning Tunneling Microscope for which a Nobel Prize in Physics was awarded in 1986. In this device, piezoelectric technology and elaborate vibration isolation measures were used to achieve better than .1 Angstrom resolution (<0.00001 micron, or 0.0000000004!), allowing detailed pictures of surface atomic structures to be viewed. Our X-Y tables are used as coarse positioners in such a system. Fig. 6 shows a beautiful picture of iodine atoms forming a monatomic layer on a palladium substrate. Can you find the missing iodine atom?

The repeatability of a positioning system is the extent to which successive attempts to move to a specific location vary in position. A highly repeatable system (which may or may not also be accurate) exhibits very low scatter in repeated moves to a given position, regardless of the direction from which the point was approached. Figures 7a, 7b, and 7c illustrate the difference between repeatability and accuracy.



Figures 7A, B, and C: Accuracy vs. Repeatability

A distinction can be drawn between the variance in moves to a point made from the same direction (unidirectional repeatability) and moves to a point from opposing directions (bidirectional repeatability). In general, the positional variance for bidirectional moves is higher than that for unidirectional moves. Quoting unidirectional repeatability figures alone can mask dramatic amounts of backlash.

Our repeatability testing is performed in the following sequence: The table is indexed to a point from one direction (say from 10.000 mm to 0.000 mm). The measuring instrument (typically a laser interferometer) is then "zeroed". The table then continues in the same direction to +10.000 mm, returns to 0.000, and continues on to -10.000 mm. The move sequence is then repeated for 3 cycles, with positional data acquired at each approach to "zero". Successive measurements alternately display the unidirectional and bidirectional values, and the worst case deviations are recorded as the respective repeatabilities. There is a natural tendency to want to collect data from a large number of cycles, and statistically process these to prepare a 3 sigma value of repeatability. While this can be done to characterize complete, closed loop positioning systems, the repeated move sequences tend to generate some fractionally induced leadscrew heating, with consequent thermal expansion and positional change. Accordingly, repeatability figures for open loop or rotary-encoded positioning tables are short-term measurements which reflect the intrinsic properties of the leadscrew and nut. The short-term nature of the repeatability test also eliminates any influence due to ambient temperature or air refractive index changes.

High resolution and repeatability are both far easier to achieve than accuracy. Synonymous use of these terms can be very expensive for positioning system specifiers. A quick look at three systems should help illustrate the distinctions. In system #1, a user is manipulating an object on an X-Y table with 10 micron resolution, and is viewing the result on a video microscope with a 100 micron field of view. The object will exhibit an annoying "hopping" motion since the travel has been quantized at the 10 micron level. This user needs more resolution. System #2 also has 10 micron resolution and must insert pins in a PGA socket on a 0.100" gridpoints within ±0.002" (± 50.8 microns). The target socket field has been mapped to eliminate leadscrew error. However, the system fails to fulfill the application requirements due to a non-preloaded rolled ballscrew with 150 micron repeatability. This system needs a higher repeatability. In system #3, an X-Y table must move a resist-coated glass plate under an electron beam to produce a reference grid plate capable of inspecting production runs of X-Y tables. This application will require high accuracy.

LEADSCREW BASED SYSTEMS

Leadscrews serve as the linear actuating mechanism in the majority of positioning systems and function as the accuracy determining element in low to moderate accuracy systems. Most lead screws use either recirculating ball nuts or anti-backlash friction nuts, with a small percentage using planetary roller nuts. The quality of the leadscrew determines the overall accuracy while the nut design, if properly executed, will eliminate backlash. The intrinsic accuracy is usually represented by two terms: a cumulative component, which is caused by minute but monatomic pitch errors, and the periodic component, which varies cyclically over each revolution. Low cost, medium accuracy leadscrews can be produced by the thread rolling process which is capable of holding cumulative error in the range of 25 to 75 microns/250 mm, and periodic errors in the range of 8-16 microns. Thread grinding is a slower and more costly process, but produces leadscrews with cumulative accuracies in the 8 to 20 micron/250 mm range, and periodic errors in the 3 to 8 micron range. Lapping is a process in which a long split nut and abrasive slurry are used to rework a ground leadscrew; it permits cumulative lead errors as low as 2 microns/250 mm, and periodic errors as low as 0.3 micron. A duplex, preloaded angular contact bearing set usually serves to constrain axial motion of the leadscrew; this introduces thrust plane errors of between 0.5 and 2 microns.

The nut should perform as a faithful follower, averaging over multiple threads and eliminating backlash upon direction reversal. Friction nuts usually incorporate two or more flexural sectors, together with a spring preload to positively engage the leadscrew. These designs can provide unidirectional repeatability of under 0.1 micron, and bidirectional repeatability (approaching the zero point from opposing directions) of 0.1 to 0.5 microns. The positive preload also automatically compensates for wear as the system ages. Ballscrews achieve backlash reduction through elastic oversizing of balls, helical cut nut bodies, or the use of two opposing ball nuts in a thread phase preload. While it is commonly assumed that ballscrews are considerably more efficient than friction nuts, their operating torque, if preloaded for high repeatability, will often exceed those of friction nuts (especially if contaminant seals are installed). In addition, the entry and exit of balls from the active race region produces torque fluctuations. Ballscrews are usually specified for applications with axial loads of high repetition rates, while ground and lapped friction nut leadscrews are best for high accuracy, light duty applications.

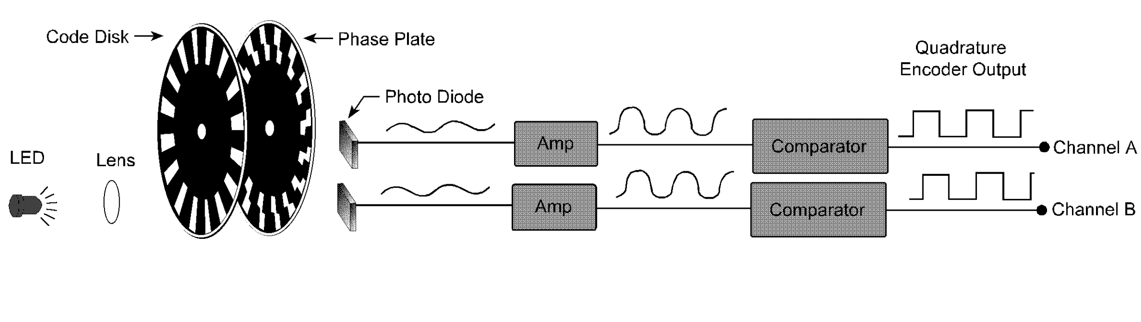
In addition to the difficulties imposed by stringent grinding and lapping tolerances, attempts to wring increasing accuracy from leadscrews run into additional barriers. Chief among these is friction induced thermal expansion: as the leadscrew spins within the nut, its temperature rises and it expands. Depending on the duty cycle and traversing velocity, leadscrews can operate at 3-10 degrees C above ambient. Together with a thermal expansion coefficient of 12 ppm/degree C (12 microns/meter per degree), this effect can result in errors of up to 120 ppm, swamping the leadscrews' intrinsic accuracy. Ruling engines were fortunate in that their duty cycle was continuous and the system stabilized after a lengthy warm-up period. Many modern systems must perform moves of various lengths, settle, acquire and process data, and move again, with no clearly defined duty cycle. Should there be any axial loads in the system, the relatively compliant nut and thrust bearings define additional error sources. The net result is that an "extremely accurate" leadscrew is somewhat of a contradiction in terms; while adequate for low to moderate accuracy systems, additional expense is better targeted at a feedback system which can sense the actual payload position, than in increasingly higher tolerance leadscrews. Leadscrews are also subject to potentially large amounts of Abbé error (see below).

THE ROLE OF FEEDBACK

The early ruling engines could be said to have had a sort of feedback: machines which produced acceptable gratings were highly accurate, and the errors of other machines were all too obviously recorded in their gratings. This information did not clearly point out areas for design improvement and served mostly as evidence of success or failure. It was appreciated at the time, however, that if an accurate "real time" position feedback system could be developed, then many of the extremely exacting mechanical requirements could be relaxed. A "servo" system could then be employed to force the payload to the desired position, irrespective of non-idealities in the mechanical drive train. The lack, at that time, of light sources possessing both high luminance and high coherence frustrated efforts along these lines.

A rudimentary form of feedback utilizes rotary encoders in conjunction with a leadscrew. The operating principle is illustrated in Fig. 8; as the code disk rotates, quadrature (90° phase shifted) signals are produced, which are then totalized in external counting circuitry. This scheme can be used with either stepping or servo motors; in the former case, it provides warning should the system lose steps or stall. Short of this advantage in stepper based systems, however, rotary encoder feedback provides no intrinsic advantage over systems based on leadscrews alone. Leadscrew bearing runout, periodic error, cumulative error, thermal expansion, Abbé error,

FIG. 8 ENCODER OPERATING PRINCIPLE



nut compliance, and nut backlash remain unchanged as error sources. To function effectively, a feedback system should sense the actual position of the payload throughout its travel, as opposed to the angular position of the rotary actuator (motor).

LINEAR ENCODERS

Linear encoders provide an accurate, cost-effective means of improving accuracy over that attainable with leadscrew-based systems. They are compact in cross section and are available in travel lengths of up to several meters. The operating principle is similar to that shown in Fig. 8, except that the code disk is now a long glass spar with chrome graduations, and the read head is a linear equivalent of the phase plate shown in the illustration. Linear encoders can be conveniently categorized as having either digital or analog output signals, and as being of either contacting or non-contacting design. Digital output encoders provide square-wave quadrature signals directly from the read-head avoiding the need for bulky and expensive interpolation boxes. Digital output models are now available with resolutions as low as 0.25 microns. Analog output encoders provide quadrature low-level sinusoidal signals which must be externally converted to digital format. While this requires additional cabling and expense, the analog signals can then be interpolated (subdivided) to achieve resolutions as low as 0.05 microns (50 nanometers), with one manufacturer (Futaba) offering a 10 nanometer resolution unit. In all cases, light transmission through the glass spar and phase plate relies on zero-order (ray) optics; diffraction limits the practical spacing of graduations on the spar to about 100 lines per millimeter (10 micron spacing). Due to the space requirements of interpolation circuitry, most high resolution systems are of the analog output type.

The intrinsic accuracy of linear encoders depends on their design; contacting models, while convenient and forgiving in their mounting tolerances, are typically capable of ± 1 to ± 5 micron base accuracy, with an additional cumulative component of between 2 and 5 microns per meter. Non-contacting designs which consist of a separate read-head and glass spar, are capable of achieving much better accuracies; several manufacturers offer units capable of ± 0.5 micron accuracy over 500 mm, and ± 0.3 micron accuracy over 200 mm.

Despite the high intrinsic accuracy of linear encoders, a number of factors conjoin to reduce the overall system accuracy. Since the linear encoder cannot be located in the same position as the object undergoing translation, there is a resulting offset between the point of interest and the point of measurement. Together with the inevitable presence of angular errors in the ways, this leads to Abbé error (see below). Depending on the encoder location, and the type of way used to define the translation axis, this error source can reach levels of some ten's of microns. With a thermal expansion coefficient of approximately 10ppm/C, ambient temperature changes can easily exceed the intrinsic accuracy of a linear encoder: a 500 mm long encoder will expand by 5 microns per degree C. In some applications, however, it is only the differential expansion between the encoder and work piece that is of interest. In this case, their expansion coefficients should be matched (soda-lime glass, silicon, and most steels are within 1-2 ppm/C of each other). Linear encoders are an inherently "one per axis" transducer; accordingly, they do not record opposite axis error and orthogonality errors in multi-axis systems (see below). Additional error sources are present due to read-head windup (approximately 0.1-0.3 microns in contacting encoder designs); interpolation errors (0.05 to 0.3 microns), and the least-significant-error jitter due to the resolution quantization (up to 1 count). Properly specified, linear encoders can significantly improve positioning system accuracy, particularly if mapping (see below) is employed, but their limitations are frequently understated.

GRATING INTERFEROMETERS

As the spacing between graduations on a linear encoder decreases, more and more of the light energy is shifted away from the zero order and diffracted into higher orders. This leads to impracticably small read-head gaps as linear encoders line spacings go below 10 microns. Although the defining patent goes nearly two decades[2], in recent years a series of grating interferometers designed to exploit this "limitation" have become commercially available. Current vendors include Holograf, Canon, Mititoyo, Heidenhain, and Sony Magnascale. Early models were transmissive in nature, required a He-Ne gas laser for operation, and achieved a 0.5 micron grating pitch by the interference of two argon laser beams on a resist-coated substrate. Recent variants use electron beam writing on a resist coated master to produce grating pitches of approximately 1.6 microns, are reflective in operation, and incorporate a compact single or multi-mode diode laser. A five axis focused ion-beam system employing grating interferometers is shown in Fig. 9.

When monochromatic light is incident upon a grating, the light diffracted from adjacent slits interferes

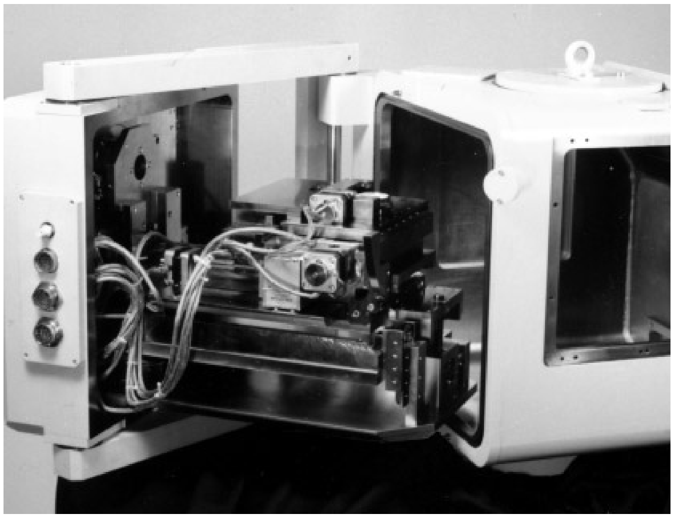


Figure 9: Five Axis System with Grating Interferometers

to form intensity maxima at particular angles. While the exit angle is fixed for any given wavelength, grating pitch, and order, the optical phase is a function of the total path length from the source. Moving the grating by one pitch interval produces an optical phase shift of exactly one cycle. With appropriate polarizing optics (Fig. 10), the light diffracted to either side of normal can be made to interfere, and the resultant intensity variations will provide quadrature signals with a period equal to one half the grating pitch. Reflective versions encounter the grating twice, resulting in a quadrature period one-quarter that of the grating. Since both rising and falling edges of each channel can be counted, the non-interpolated resolution of a 1.6 micron pitch grating would be 0.1 micron; the use of 10x or 25x interpolation yields 0.01 or 0.0025 micron resolution (10 or 2.5 nanometers, respectively).

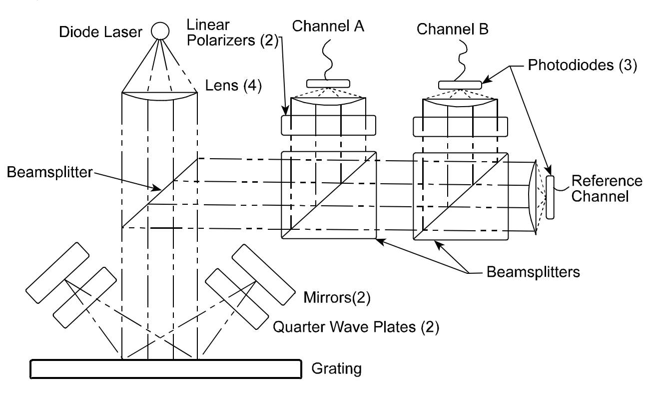


Figure 10: Grating Interferometer Operating Principle

Since the grating masters are generated on fairly conventional e-beam equipment intended for I.C. mask lithography, the available travels have been limited to 150 mm or 200 mm. One manufacturer has recently offered a 400 mm version (presumably generated by butting two units end to end in a phase-controlled manner), and plans to announce an 800 mm version. All grating interferometers provide a comfortable working gap of 3 mm to 9 mm, and reasonable alignment tolerances. Their accuracy is a function of the mask-making machine which generates the master; current claims range from 0.2 to 0.6 microns over 150 mm. Since both optical legs of the grating interferometers are equal, these devices are totally free from the effects of air index changes due to temperature, pressure, humidity, and trace gases; they are similarly free from errors due to laser wave length shift. The 90 degree phase shift between the quadrature signals is also of higher quality, and more tolerant of misalignment, than that of linear encoders, simplifying interpolation requirements. There are two camps regarding the thermal expansion coefficient of the substrate material; one generates the grating on fused quartz (thermal expansion coefficient 0.5 ppm/degree C), while the other utilizes conventional soda-lime glass (10 ppm/degree C), or steel (12 ppm/degree C). The former is superior for "pure" dimensional metrology or on work pieces maintained at 20.0 degrees C, while the latter embraces a pragmatic approach that emphasizes a feedback device which tends to "track" the workpiece.

In addition to questions regarding appropriate substrate expansion coefficients, grating interferometers are subject to Abbé error (see below) in amounts which can substantially exceed their intrinsic error. Due to the fact that they are "one per axis" devices, they fail to detect opposite axis error and orthogonality in multi-axis systems (see below). Their cost is significantly higher than that of most linear encoders, although the lower cost models approach the cost of similar accuracy and resolution linear encoders, and are well below the costs incurred when laser interferometers are required. When cost concerns allow their consideration, they constitute a welcome addition to the feedback tools at the disposal of positioning system designers.

ABBÉ ERROR

Abbé error (pronounced ab-a') can be a significant source of error in positioning applications. Named after Ernst Abbé, a noted optical designer, it refers to a linear error caused by the combination of an underlying angular error (typically in the ways which define the motion) and a dimensional offset between the object being measured and the accuracy determining element (typically a leadscrew or encoder). In open loop systems (or closed loop systems employing rotary feedback), the accuracy is nominally determined by the precision of the leadscrew. Similarly, in systems with linear encoders or interferometers, it is that device which determines the accuracy. It is important, however, to recall exactly what information these devices provide: leadscrews really tell us nothing but the relative position of the nut and screw, and encoders tell us only the position of the read-head relative to the glass scale. Extrapolating this to include the position of an item of interest, despite its firm mechanical connection to the nut or encoder read-head, is ill-founded.

To illustrate this, consider Fig. 11 which shows a single axis stage with a linear encoder. The stage carries an offset arm which positions a probe over a sample. The apparent distortion in the stage is intentional; it is intended to illustrate, in exaggerated fashion, a stage whose ways have a curvature (in this case, yaw). Someone using this stage, and in possession of appropriate test instruments, would measure an error between the stage position, as determined by the encoder read-head, and the actual linear position of the probe.

Suppose the curvature is sufficient to produce an angle a' b in Fig. 11 of 40 arc-seconds (a' is drawn parallel to a). If the stage moves forward 250 mm, the probe will be found to have moved 250.100 mm, resulting in an X axis error of +100 microns. If the ways were, in fact, curved in a circular arc as shown, there would also be a Y-axis shift of +25 microns. This Y-axis error would be eliminated (while the X-axis error would remain) if the angular error were a purely local property of the ways at the +250.000 mm location.

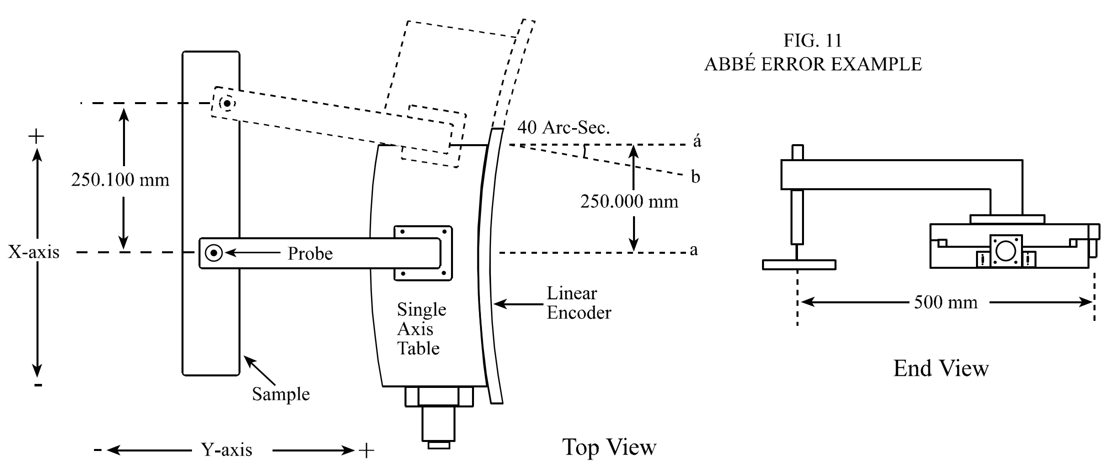
Abbé error is insidious, and can best be countered by assuming the presence of angular error in a system and then working to minimize both the underlying error and its effect through design optimization and appropriate placement of leadscrews, encoders, etc. The best tool to analyze angular error is the laser interferometer which, when used with special dual path optics, can measure pitch or yaw with 0.05 arc-second resolution. Roll can be measured using a video autocollimator and rectangular optical flat, or by performing multi-point surface measurements with LVDT's.

Sources of angular error include the following:

1) Curvature of ways
2) Entry and exit of balls or rollers in recirculating ways
3) Variation in preload along a way
4) Insufficient preload or backlash in a way
5) Contaminants between rollers and the way surface
6) Torsional compliance in a way due to:
    a. external forces acting on the load
    b. overhang torques due to the load's travel

In the example shown in Fig. 11, Abbé error could be lessened by moving the encoder to the left side of the stage. Reducing the arms length, or mounting the encoder at the edge of the sample (with the read-head connected to the arm), would be more effective. Virtual elimination of Abbé error could be achieved by using a laser interferometer and mounting the moving retroreflector on the probe assembly. Note that the component positions shown in Fig. 11 effectively control Abbé error due to the pitch error of the stage since the height of the probe and encoder are roughly equal. While the stage might exhibit a pitch error (rotation around the Y-axis), there is no corresponding vertical (Z-axis) offset needed to produce Abbé error. The third degree of rotational freedom, roll, corresponds in the illustration to the rotation around the axis of motion (X-axis). This would result in the gap between the probe and the sample varying as the stage moved.

FIG. 11
ABBÉ ERROR EXAMPLE



In general, try to estimate or measure the magnitude of all three possible angular errors (roll, pitch, and yaw) in any given system under actual load bearing conditions. Then, look for any offsets between driving or measuring devices and the point of interest on the load. Calculate the Abbé error and, if it proves unacceptable, optimize the design to reduce either the offset or the underlying angular error. In general, systems built using precision lapped granite and air bearings which do not extend the load beyond the table base at any point in the travel, are best at minimizing angular errors.

To determine the magnitude of Abbé error, simply multiply the offset by the tangent of the angle. In the example, this was: 500 mm x tan (40 arc-seconds) = 500 x tan (.011 degrees) = 500 x .000194 = 97 microns. If the angle is known in radians instead of degrees, the problem is that much easier: the Abbé offset is simply equal to the angle x offset. Finally, a helpful rule of thumb is that the Abbé error will equal about 5 nanometers per mm of offset and arc-second of angular error. Once again, 40 x 500 x 5 = 100,000 nanometers, or 100 microns. The chart in Fig. 12 may prove helpful in determining which offsets produce Abbé error for a given angular error.

Angular Error	Offset Axis	Error Axis
Ox (roll)	X	none
Ox	Y	Z
Ox	Z	Y
Oy (pitch)	X	Z
Oy	Y	none
Oy	Z	X
Oz (yaw)	X	Y
Oz	Y	X
Oz	Z	none

Figure 12: Offset Axis vs. Error Axis

COSINE ERROR

Cosine error results from an angular misalignment between the motion of a positioning table and the accuracy determining element (leadscrew, encoder, or laser interferometer beam path). Under most circumstances, it has a negligible effect on overall accuracy, owing to the significant degree of misalignment needed to influence accuracy. Consider, for example, a case of a 250 mm travel positioning table with a linear encoder. The encoder is pitched so as to be inclined to the direction of motion and the encoder will accordingly measure a larger move than has actually occurred. Pythagoras's theorem ($a^2 + b^2 = c^2$) yields the magnitude of the error. At a 0.1 mm misalignment, the encoder path equals $250^2 + 0.1^2 = 62500.01$, or 250.00002 mm; the error is only 20 nanometers. If the misalignment is specified in terms of angle, then the error will equal: travel * (l-cos O) - hence the name: cosine error. In the above example, the angle was 83 arc-seconds, and cos O = 0.999999920.

If the encoder resolution is one micron, then a misalignment of 71 microns would be necessary to generate a cosine error equivalent to a single count. Typical stage design, fixturing, and inspection procedures can hold way and encoder alignment to levels far below this value, rendering cosine error of negligible consequence in most positioning stages. In systems using laser interferometers for positional feedback, however, simple visual alignment with a reduced aperture can introduce cosine error on the order of several ppm. This is significant when compared with the intrinsic interferometer accuracy of <0.1 ppm, and may necessitate careful adjustment of the beam angle in pitch and yaw to maximize the measured distance. Note that with laser interferometers, cosine error results in a distance measurement smaller than the actual move; this is opposite to the effect of cosine error for a linear encoder.

MAPPING

Mapping can be an effective tool to reduce errors in positioning systems. Sources of error amenable to correction via mapping include those due to leadscrew cumulative error, leadscrew periodic error, Abbé error, nut backlash, cosine error, and deviations from orthogonality in multiple axis systems. Essentially, mapping consists of measuring and recording the actual position of a stage, for later use in returning to that point. In most cases, the measuring instrument is used only to acquire data on the stage and is not present during actual operation. Common calibration sources include laser interferometers and precision "low-E" glass grid plates. The positioning system must have sufficient resolution to implement a corrective move to the desired degree of accuracy. As an example, consider a positioning table with one micron resolution. Nominally, a 40.000 mm move would require 40,000 steps or counts. In this case, due to a cumulative leadscrew error, 40,000 counts actually results in a 40.009 mm move. Programming a move of 40,000 ¸ 40,009 x 40,000, or 39,991 counts, will produce the desired 40.000 mm move.

Mapping is especially effective when a relatively small number of positions are required; in this case, a unique measured value can be used for each location. In other cases, one or more points can be recorded and subsequent points inferred, or "interpolated", from the nearest measured values. In the above example, a 20.000 mm move would require 19,996 counts, under the assumption that the screw error is linear. Compensation for leadscrew periodic error requires several points for each revolution, substantially increasing the storage requirements. Leadscrew or encoder thermal expansion often sets a limit on the level of accuracy worth reducing by mapping techniques.

LASER INTERFEROMETERS

Laser interferometers (Fig. 13) provide the ultimate in position feedback combining very high resolution, non-contact sensing, high update rates, and intrinsic accuracies of 0.02 ppm. They can be used in positioning systems as either passive position readouts, or as

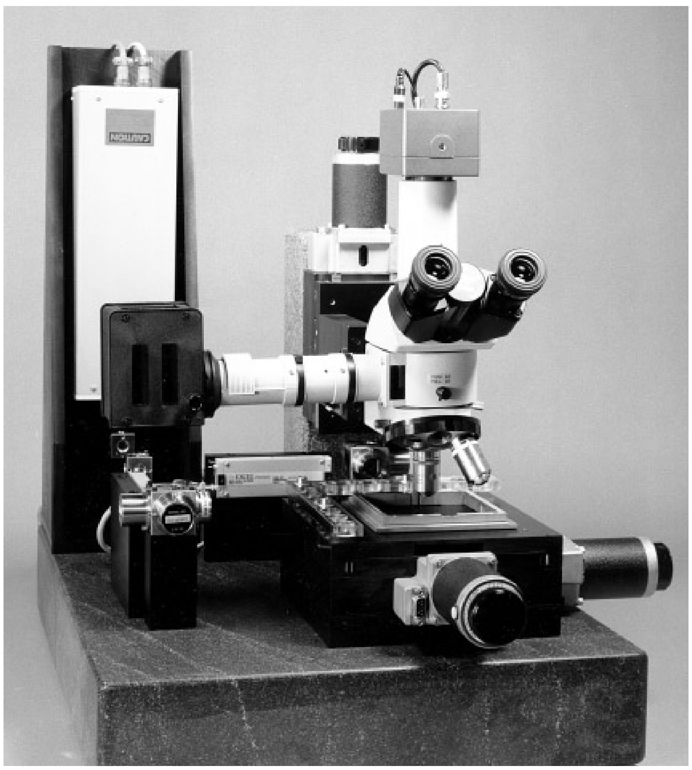


Figure 13: Interferometer Feedback System

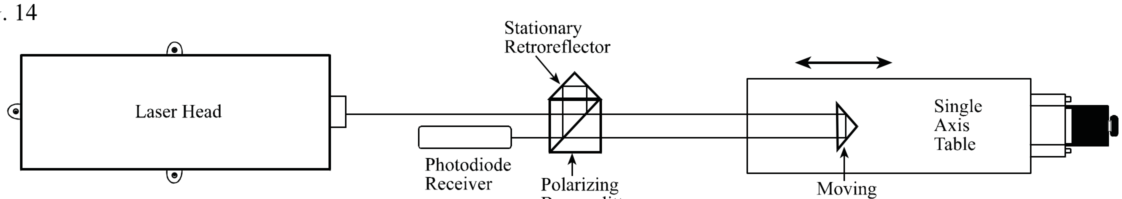
active feedback components in a position servo loop. Unlike linear encoders, the interferometer beam path can usually be arranged to coincide with the item or point being measured, eliminating or greatly reducing errors due to Abbé error.

Laser interferometers can be divided into two categories: fringe counting and two-frequency systems. The former is similar in operation to a Michaelson interferometer, while the latter uses two closely spaced frequencies, one of which experiences a Doppler shift from the moving reflector. Upon recombination, the two frequencies are heterodyned to generate a beat frequency within the range of counting eletronics. The two frequency design, while more costly to implement, is considered the higher performance system, especially for velocity feedback. In both cases polarization selective optics are used to route one beam to and from the moving workpiece, while retaining a fixed path for the reference beam.

Single axis systems utilize a beam path (as shown in Fig. 14) and consist of the laser head, polarizing beam splitter with retroreflector, the moving retroreflector, and a photo diode receiver. XY systems (Fig. 15) replace the moving retroreflector with a plane mirror and add a quarter-wave plate and an additional retroreflector to the separation optics. The quarter wave plate circularly polarizes the workpiece beam causing it to perform two passes with a corresponding doubling of resolution and halving of achievable top speed. This configuration eliminates errors due to Abbé offset, yaw, pitch (to a first order), and opposite axis horizontal run out, and ignores orthogonality errors in the X-Y table (the plane mirrors, however, must be precisely square to each other). The reflectors can consist of two "stick mirrors" in adjustable mounts, or a single "L mirror" (as shown in the photo). The latter eliminates concerns over stick mirror misadjustments, but carries cost penalties which grow rapidly with increasing travel.

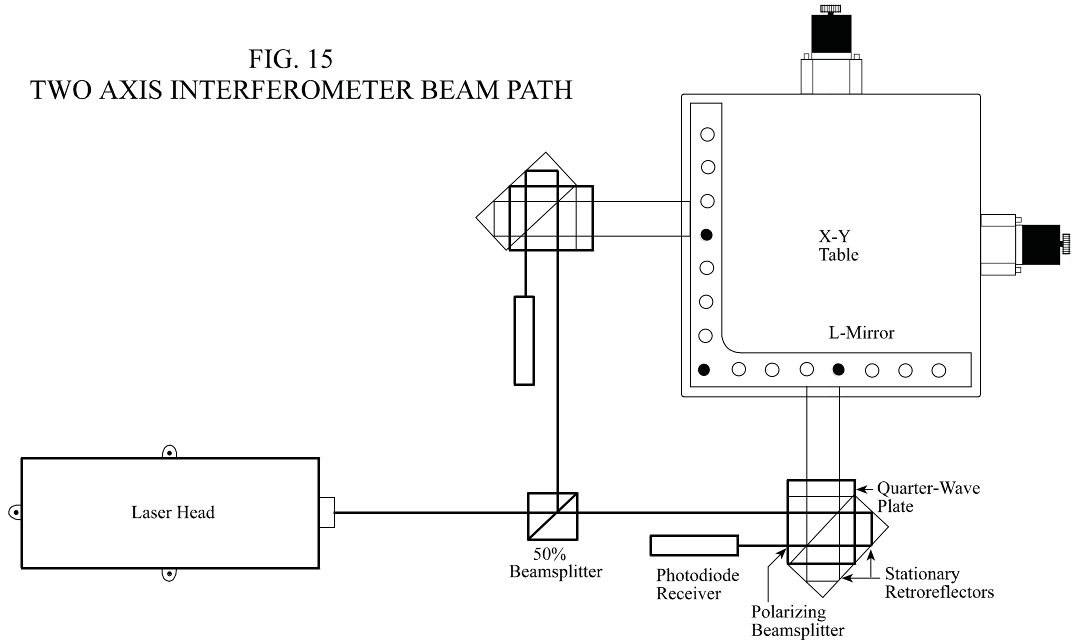
The double-pass plane mirror interferometer mentionedabove attains a resolution of 10 nanometers. A variant upon this design (Fig. 16) produces four passes along the measurement path, providing a resolution of 5 nanometers; similar schemes with higher electronic interpolation reach 0.625 nanometers, the highest value offered by commercial interferometers. To simplify following the beam path in Fig. 16, note that two passes through the quarter-wave plate rotate the polarization vector by 90 degrees with the result that a beam, whose

FIG. 14



SINGLE AXIS INTERFEROMETER BEAM PATH

FIG. 15
TWO AXIS INTERFEROMETER BEAM PATH



initial polarization was transmitted through the beam splitter diagonal will now be reflected, and vice versa. On the academic front, Dr. Robert Reasenberg of the Smithsonian Astrophysical Observatory has developed a

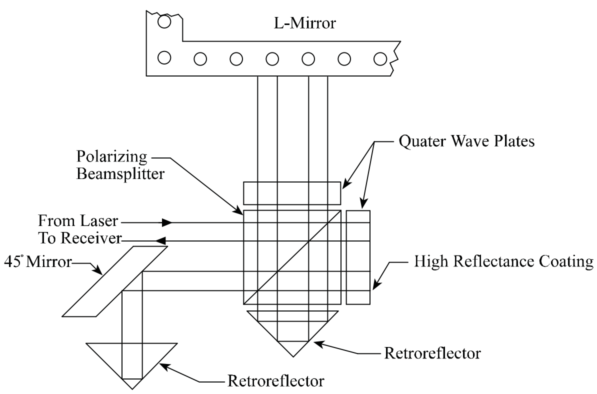


Figure 16: Four Pass Interferometer Beam Path

15 picometer null sensing interferometer for use in a future orbiting 5 micro-arc second stellar interferometer, (P.O.I.N.T.S.)[3]. Professor Ray Weiss of MIT has developed a 30 pass interferometer system for L.I.G.O. (Laser Interferometer Gravitational Wave Observatory) that will monitor displacements between suspended masses at the ends of a buried 8 km long "L" shaped vacuum tunnel with sensitivity below 3 x 10^{-16} cm/ Hz from 100 Hz to 1 KHz. Way to go, Ray!

As the N.B.S. (now N.I.S.T.) pointed out in the mid seventies, any He-Ne laser provides frequency stability equal to, or better than, 1 part in 10^{-6} (any greater error would inhibit the lasing process due to the narrow neon line-width). Frequency stabilization systems can improve this, achieving long term accuracies of as little as 2 parts in 10^{-8} (0.02 ppm). The following error sources, however, conjoin to degrade this very high intrinsic accuracy:

1) Speed of light variations due to temperature, pressure, etc.
2) Pressure, temperature, and humidity sensor accuracy
3) Plane mirror squareness and flatness
4) Thermal expansion of workpiece, positioning table, base, plate, and optics
5) Cosine error
6) Accuracy of workpiece thermal expansion coefficient
7) Differential flexure of positioning table top through its travel
8) Edlen and Jones equation accuracy
9) Deadpath correction accuracy

It is often assumed that once the cost increments associated with laser interferometers have been justified, high accuracy can be assumed. As the above list of error sources should indicate, shifting to an interferometer based system also reveals a new regime of low level errors, the aggregate effect of which may be serious. We have seen that the laser wavelength accuracy and stability itself is on the order of 0.02 ppm (5 nanometers over 250 mm). It is helpful to compare each of the above error sources to this quite high intrinsic accuracy.

Item #1 reflects the variation in atmospheric refractive index due to temperature, pressure, and humidity. If uncompensated, the laser wavelength in air will vary by 1 ppm per degree C, 0.4 ppm per mmHg pressure change, and 0.1 ppm per 10% change in R.H.

On a low pressure, muggy summer day, this can total 15 ppm (787 mmHg, 25 degrees C, and 70% R.H.), a factor of 750 times the laser's intrinsic accuracy. This is clearly unacceptable and, accordingly, sensors are used per item #2 and the Edlen equation[4] (item #8) to compensate for the air index variability. The question now becomes the absolute accuracy and drift of these sensors for which commercially available compensation systems achieve 1.5 ppm after calibration.

While these sensors compensate for air temperature, pressure, and humidity variations, they fail to detect index changes due to excess CO_2, oil and diesel vapors, operator flatulence, etc. The wavelength tracker shown in Fig. 17 employs a differential interferometer (see below) to measure the "change" in distance between mirrors formed by the ends of a Zerodur bar (0.1 ppm/degree C expansion coefficient). Since the ends of the Zerodur bar are, for all intents and purposes, "not going anywhere", the measured dimensional change is due entirely to refractive index changes. One limitation of this approach is that it only tracks changes from initial measurements obtained with conventional sensors. The optical path of an absolute refractometer is shown in Fig. 18; it incorporates two interferometers, one of which is maintained in vacuum, and another whose ratio of air path to vacuum path is varied over 20 mm by a linear actuator. This is a tricky design which requires strong, vacuum tight bellows but provides absolute atmospheric compensation at the 0.5 ppm level.

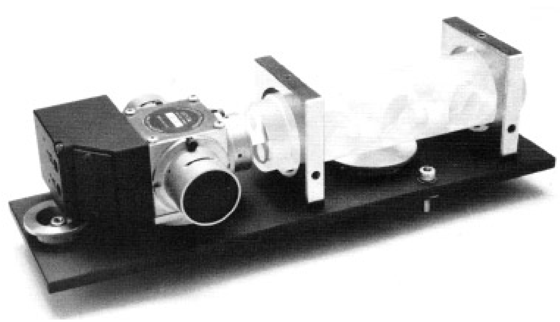


Figure 17: Wavelength Tracker
(photo courtesy of Hewlett Packard)

A related factor is the deadpath value (item #9). In general, the system layout should minimize the distance between the positioning table zero position and the polarizing beam splitter/reference retroreflector. As the refractive index of any intervening air changes, there is an effective offset of the "zero" position of the table. This distance must be carefully measured and air index changes applied to it to compensate for this zero point shift.

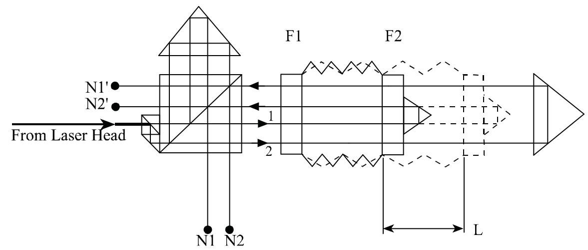


Figure 18: Absolute Refractometer
(diagram courtesy of Spindler & Hoyer GMBH)

Item #3 relates to the orientation and surface quality of the plane mirrors in two axis systems. These mirrors may be either a single "L" mirror or individual "stick" mirrors. Optical vendors are unwilling to quote upon and certify "L" mirrors below 1 arc-second of squareness, and 2 seconds is a more easily achievable value. One arc-second of squareness error alone will produce errors of 5 ppm, or 1.0 micron, over 200 mm. In the case of vendor aligned "stick" mirrors, the ability to align the mirrors presents a risk of accidental or eventual misalignment. The method used to determine squareness should be examined carefully; in addition, shipping trauma, mounting stresses, and thermal expansion of the substrate may alter the initial squareness. These mirrors are typically fabricated from Zerodur; while this retains an excellent surface figure over changing temperature, it exacerbates differential expansion with the metal to which it is mounted (in some cases, the entire XY top section is made from Zerodur with integral mirrors). Finally, the surface flatness constitutes an error source; in practice a surface error of $\pm$ 0.1 wave (0.1 micron total) is the best achievable, and this requires a substantial thickness to length ratio.

The optics thermal expansion error mentioned in item #4 takes place because the reference beam has a path length within the beam splitter and retroreflectors which is half that of the measurement beam. As the ambient temperature changes, the glass expands, and the difference in beam paths produces an error which is typically 0.5 microns/degree C. By substituting a highly reflective quarter-wave plate for one of the retroreflectors (Fig. 19), this effect can be reduced by more than tenfold.

Simple visual beam alignment can produce cosine error (item #5) of several ppm, which can be reduced in retroreflector-based systems to under 1 ppm with more exacting procedures. Plane mirror systems can use auto-reflection alignment techniques to reduce cosine error to below 0.1 ppm.

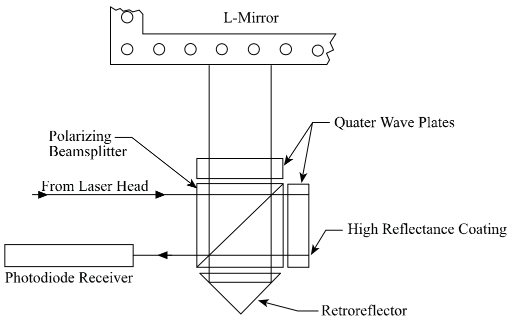


Figure 19: Low Thermal Drift Interferometer

As mentioned above, plane mirror interferometers on X-Y tables compensate for yaw errors in the table as well as (to a first order) pitch errors. Should the table top region carrying the plane mirror sag differentially from the workpiece area, however (item #7), a positional error will result. Such flexure is encountered on overhanging table designs, and recalculating or air bearing designs are accordingly preferred.

Since the interferometer only measures distance variations between the stationary optics and plane mirror, or retroreflector, there are a number of thermal expansion possibilities that can corrupt measurements. In many cases, the workpiece is moved under a stationary function (microscope, e-beam, laser axis, etc.) which defines the point of interest. This problem, referred to as column reference, clearly requires that we measure the workpiece position relative to the column center point. One tool for such work is the differential interferometer (Fig. 20) which measures only positional variations between the stage plane mirror and a separate mirror

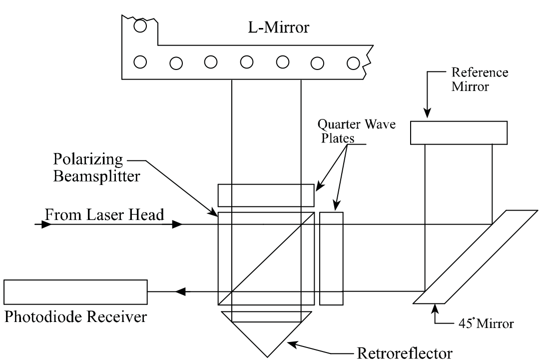


Figure 20: Differential Interferometer

which can be column mounted. This eliminates errors due to thermal expansion of the column support bridge. Differential interferometers also allow more compact vacuum chamber dimensions for high vacuum positioning applications. When used in air, proper correction for the deadpath (distance between reference mirror and stage mirror) must be performed. Additional complicating factors include errors due to workpiece thermal expansion, inability of the optics to perfectly separate orthogonal polarizations (5-10 nanometers) and phase interpolation electronic errors (one to two times system resolution). As the preceding should indicate, laser interferometers provide the highest attainable system accuracy but still require careful attention to error sources as part of an overall error budget.

MULTI-AXIS SYSTEMS

Most of the preceding discussion has dealt with single axis systems; an optimistic viewpoint might conclude that multi-axis systems would generate errors in accuracy describable by a square (two axes) or cube (three axes) of side dimensions equal to the error produced by a comparable single axis system. Alas, no such luck. A dominant error source in multi-axis systems is the degree of orthogonality between axes; in addition to static errors, dynamic (flexural) effects can occur as the axes move relative to each other. A squareness error of 20 arc-seconds will produce linear errors of 100 ppm, or 25 microns over 250 mm. Merely measuring squareness at the center of one axis of travel is misleading; a comprehensive squareness measurement should incorporate yaw errors on each axis and be the result of a grid of measurement points. Precision granite reference squares, or a grid plate with microscope, can be used to measure squareness; in the latter case inverting the grid plate provides a simple stratagem that can allow squareness measurement accuracy to exceed that of the grid plate itself. In three axis systems, a sphere bar (an Invar bar with precision balls at each end) can be used to determine accuracy over a three dimensional workspace; the result of such tests rapidly converted a number of early "tenth micron" coordinate measurement machines to "tens microns" systems.

As previously mentioned, leadscrews, linear encoders, and grating interferometers are inherently single axis devices; should any axis exhibit horizontal run out, the encoder on that axis will not detect it, nor will the encoder of any other axis; this effect is referred to as "opposite axis error". Two axis laser interferometer systems substitute mirror squareness for axis squareness; this is equally challenging, and additional interferometer axes encounter traditional squareness requirements.

POSITIONING SYSTEM DESIGN

A number of design factors influence the accuracy of positioning systems. Among rolling element tables, two fundamental categories are recirculating and non-recirculating designs. The former (Fig. 21a) incorporate recirculating races of balls or rollers, and permit a

Figure 21A: NEAT's HMS-1000-SM

smaller "shuttle" payload carrier to move along a fixed base. As balls or rollers enter and exit the ways, force fluctuations and small angular errors are produced. Non-recirculating designs (Fig. 21b) make use of a full sized top, together with a set of balls or rollers, which move along the ways at one-half the speed of the table top. As the table traverses, it overhangs the base, resulting in a torque moment and consequently some angular error. A variant upon the latter design uses a set of balls or rollers greater than, or equal to, in length to the base and table top. This provides a higher degree of support, but introduces force and angular perturbations as balls enter and exit the ways and may require additional space into which the retained ball compliments may extend.

Figure 21B: NEAT's HM-1800-SM

Air bearings (Fig. 22) provide an alternate way design and are the most effective means of constraining free movement to a single axis of translation. Air bearings have an inherently "averaging" nature which results in linear and angular errors significantly below those of the surfaces which define their motion. They can achieve linear run outs below 2 microns/250 mm, and hold roll, pitch, and yaw below 5 arc-seconds/250 mm. Air bearing designs are usually of "shuttle" design, avoiding angular errors due to overhung loads. Their deficits include higher cost, additional support apparatus in the form of compressors, filters, etc., and a lower torsional and linear stiffness than that found in rolling element bearings. Air bearings often incorporate precision lapped granite to define way surfaces; one design variant allows

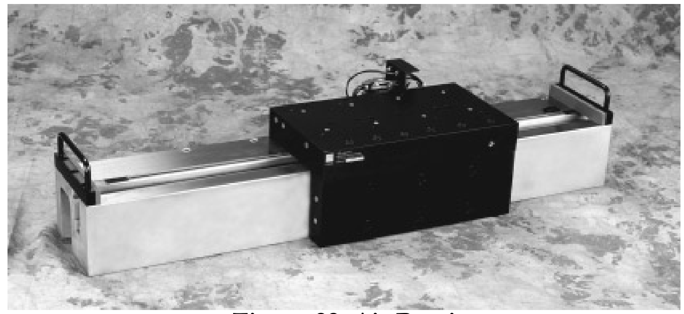


Figure 22: Air Bearing

a single-piece platform to move in both X and Y axes while fully supported on an ultra-flat granite base. Other designs employ an airbearing X axis translating beneath a moving Y with Z axis gantry. An example (Fig. 23) utilizes non-contacting linear servo motors with 0.5 micron encoder feedback.

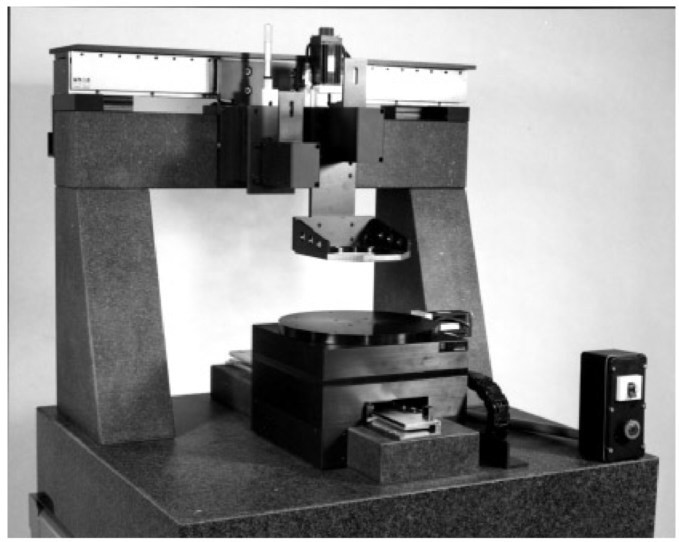


Figure 23: Air Bearing X, Y, Z System

The role of the linear actuator in high accuracy, high resolution systems merits careful consideration. Leadscrews remain effective as linear actuators, but may lead to servo loop stability problems in high resolution systems, depending on the payload mass and nut or coupling compliance. Stiff, lapped nuts and fine pitch leadscrews improve stability conditions, as does a "dual loop" approach in which a tightly coupled rotary servo operates in conjunction with a high resolution linear feedback device. Piezo-electric actuators offer exceptional resolution and linearity, but are restricted to travels below 200 microns unless "inchworm" or resonant devices are employed. Linear stepping motors can function as actuators, but are limited by their poor damping and stiffness. Recently, brushless linear servometers (Fig. 23) have gained acceptance; they translate current directly into force without the backlash, friction, and decoupling associated with leadscrews. In most cases, the goal is to move and settle to within one resolution element of the target position in as little time possible. As accuracies and resolution requirements increase, this continues to present challenging design problems.

CONCLUSION

In summation, high resolution and high repeatability are positioning systems parameters which are attainable with moderate effort and can be described in many cases by a simple pair of "specs". High accuracy proves to be a much more elusive goal, with rapidly escalating cost and system complexity, as higher and higher levels are sought. Despite customer preference (and vendor willingness) to simply "pin a number" on accuracy, it is, in reality, a global parameter which requires a comprehensive approach to the specific positioning components, control and feedback systems, functional application, and operating environment. When approached in such a realistic fashion, both positioning system purchasers and vendors benefit from meaningful and defensible accuracy ratings.

REFERENCES:

1) Ingalls, Albert, Scientific American, pages 45-52, June, 1952

2)Huntley, Wright, U.S. Patent #3,738,753, Interferometer Using a Grating to Measure Motion, June 12, l973

3) Reasenberg, Robert, "Progress Report on a Picometer Null-sensing Distance Gauge," JPL Workshop on Space Interferometry, 30 April-2 May 1990

4) Edlen, B., "The Refractive Index of Air," Metrologia, Vol. 2, #3, 1966

Reprinted from The Motion Control Technology Conference Proceedings, March 19-21, 1991.

Five Axis Air Bearing Gantry System

ABOUT THE AUTHOR...

Kevin McCarthy is the President of New England Affiliated Technologies in Lawrence, Massachusetts. He holds a B.S. degree in Physics (1975) from M.I.T.. Your comments and criticism are welcome; please contact the author at 1-800-227-1066.

Kevin McCarthy, President
New England Affiliated Technologies