One of the major challenges in nanoscale manufacturing is defect control because it is difficult to measure nanoscale features in-line with the manufacturing process. Optical inspection typically is not an option at the nanoscale level due to the diffraction limit of light, and without inspection high scrap rates can occur. Therefore, this paper presents an atomic force microscopy (AFM)-based inspection system that can be rapidly implemented in-line with other nanomanufacturing processes. Atomic force microscopy is capable of producing very high resolution (subnanometer-scale) surface topology measurements and is widely utilized in scientific and industrial applications, but has not been implemented in-line with manufacturing systems, primarily because of the large setup time typically required to take an AFM measurement. This paper introduces the design of a mechanical wafer-alignment device to enable in-line AFM metrology in nanoscale manufacturing by dramatically reducing AFM metrology setup time. The device consists of three pins that exactly constrain the wafer and a nesting force applied by a flexure to keep the wafer in contact with the pins. Kinematic couplings precisely mate the device below a flexure stage containing an array of AFM microchips which are used to make nanoscale measurements on the surface of the semiconductor wafer. This passive alignment system reduces the wafer setup time to less than 1 min and produces a lateral positioning accuracy that is on the order of \( \frac{1}{24} \mu\text{m} \).

**Introduction**

In the nanopatterning industry, metrology [1] is an important issue because it is needed to ensure the correct alignment and patterning of features. Integrating a sensing system which can instantaneously send back the dimensional information about manufactured products can reduce losses and defect rates. However, it is difficult to perform in-line metrology in nanofabrication systems because it requires not only real-time inspection but also nanoscale resolution of complex features.

Atomic force microscopy has the ability to make high resolution measurements (subnanometer-scale) and is widely utilized in scientific and industrial applications. However, there are two main setbacks to the use of AFM metrology in in-line manufacturing applications [2]. The first is the scanning speed of the AFM tip and the second is the time it takes to place and align samples in the AFM.

The limitations created by the long setup time and the low-speed scanning can be solved by using single-chip AFMs [3]. Single-chip AFMs use microelectromechanical system (MEMS)-based flexures, sensors, and actuators (as shown in Fig. 1) to scan the AFM tip across a surface in order to achieve high-speed measurements with small device dimensions. Each single-chip AFM MEMS device is approximately 2 mm wide by 1 mm tall and is capable of scanning a 10-\(\mu\text{m}\) by 10-\(\mu\text{m}\) area. Thermal actuators in the MEMS device are used to do the scanning in both the \(x\)- and \(y\)-directions as well as to excite the \(z\)-axis of the AFM so that it can be run in tapping mode. In the design of the inspection system presented in this paper, a precision XY-stage is used to position the single-chip AFM relative to the inspection wafer. The aim of this project is to enable atomic force microscopy to be integrated directly into wafer fabrication lines.

Flexure bearings are used in this system because of their precision and mechanical simplicity [4,5]. The flexure bearings are used to position the single-chip AFMs relative to the wafer that is
being inspected. The precision positioning stage is designed to be able to achieve millimeter-scale displacements with micron level precision. Multiple stages with multiple independent AFMs can be incorporated into the metrology system to increase the inspection speed and area as shown in Fig. 2. However, in order to be able to incorporate multiple stages into a single inspection system, the size of the positioning flexures must be minimized and the space utilization in the positioning system must be maximized.

System Design

The quick change in-line dimensional metrology system consists of three main parts: (1) a flexure stage with the AFM chips (2) a flexure-based z-axis approach mechanism, and (3) a passive wafer alignment stage as shown in Figs. 2(b)–2(d) [6].

Figure 2(e) is a flowchart showing how does this in-line metrology system works and operates. The silicon wafer is first placed into the passive alignment system in order to precisely position the wafer. The flexure stage is then placed down on top of the passive wafer alignment system and aligned to the wafer using a kinematic coupling. The flexure mechanisms on the top stage allow the AFM chip to be precisely located relative to the feature being measured on the silicon wafer. However, this location system only needs to be set the first time a wafer is loaded into the system. After the AFM chips are located relative to the silicon wafer, the passive alignment system and kinematic couplings...
are sufficient to ensure that the same spot is measured on each subsequent wafer that is loaded into the system.

Once the wafer is loaded into the system and the top plate is placed onto the kinematic couplings, an auto-approach mechanism is used to move the AFM chip into contact with the silicon wafer for measurement. After the measurement is taken, the AFM retracts and the top plate and wafer can be removed to allow another wafer to be placed into the system. These rapid alignment systems allow the time between measurements to be reduced to less than 1 min.

**Passive Wafer Alignment System**

In order to precisely align a silicon wafer under the AFM inspection stage a new passive alignment system was developed. The system consists of three pins that exactly constrain the wafer and a nesting force applied by a flexure to keep the wafer in contact with the pins. Kinematic couplings are used to precisely mate and a nesting force applied by a flexure to keep the wafer in contact with the pins. This level of positional accuracy was achieved by creating a new model for the optimal location of pins in a three pin alignment system as described below.

A wafer in contact with a flat surface has three degrees-of-freedom (3DOFs)—two translational and one rotational. Out of plane rotations and translation are thought to be sufficiently limited by the gravitational force on the wafer. Three pin constraints exactly constrain the wafer such that it has 0DOFs. Two pins with intersecting lines of action (LOAs) constrain the translational DOF of the wafer. A third pin is required to constrain rotation of the wafer and its location can have a great effect on wafer loading accuracy. Rigid bodies rotate about instantaneous centers located at the intersections of the LOAs of constraints [7]. Thus, in order to constrain a rotational DOF two of the three constraints must be parallel, forming an instantaneous center at infinity and thereby preventing rotation [7]. Conventional wafers possess a flat which provides a convenient location for the constraints with parallel LOAs. A nesting force applied by a flexure mechanism maintains contact between the wafer and the pins.

Determination of the optimal location for the third pin and nesting force is a major objective of this paper. Figure 3 illustrates naming conventions for the geometry of the constraint system. The flexure applies a nesting force \( F_n \) at an angle \( \phi \) with respect to the \( x \)-axis. The pins in contact with the flat of the wafer are referred to as “left pin” and “right pin.” The third pin is referred to as “third pin” and makes an angle \( \theta \) with respect to the \( x \)-axis.

\[
\begin{bmatrix}
  f_3 \\
  f_1 \\
  f_r
\end{bmatrix} = f_n \begin{bmatrix}
  -\cos \phi \sec \theta \\
  \frac{1}{2} (\sin \phi - \cos \phi \tan \theta) \\
  \frac{1}{2} (\sin \phi - \cos \phi \tan \theta)
\end{bmatrix}
\]  

(1)

The first step in optimizing the constraint location is analyzing the reaction forces and moments as a function of the unknown angles \( \theta \) and \( \phi \). Reaction forces at each pin are determined using matrix inversion to satisfy static equilibrium of forces in the \( x-y \) plane and equilibrium of moments about the \( z \)-axis [8]. Reaction forces at each pin are shown in Eq. (1).

Allowable combinations of \( \phi \) and \( \theta \) for a given wafer geometry are those resulting in compressive reaction forces (positive values) as this is a necessity for the wafer to be in static equilibrium. From this constraint, we find that \( \phi \) and \( \theta \) must be on opposite sides of an imaginary vertical line drawn between the left pin and the right pin. We arbitrarily chose the left side of this line for the nesting force and the right side for the third pin location.

Values for \( \phi \) and \( \theta \) are also limited by a “Nesting Force Window” [7]. In this paper, a numerical method of nesting force window analysis is developed from Blanding’s method [7] with the advantage of maximizing the utility of the nesting force. The intersections between pin lines of action constitute wafer instantaneous centers of rotation (ICRs). The moment of the nesting force about each ICR for a wafer in contact with only the pins with lines of action that intersect at that ICR must be in the direction of the pin that is not in contact with the wafer. Figure 4 illustrates the wafer instantaneous centers of rotation as well as the correct orientation of nesting force moments about the ICRs. Equation (2) shows the nesting force moment about ICR1 as a function of \( \theta \) and \( \phi \).

\[
M_{ICR1} = \frac{1}{2} F_n (r_x - r_z) \sec \theta \sin (\theta - \phi)
\]  

(2)

The nesting force moment is shown to be a function of the magnitude of the nesting force, the distance between the pins located on the wafer flat, and angles \( \phi \) and \( \theta \). By symmetry and choice of coordinate system, the two nesting force moments are equal in magnitude and opposite in direction. In order to increase the magnitude of the nesting force moments, the distance between the pins in contact with the wafer flat should be maximized. Maximizing the magnitude of the nesting force would also increase the magnitude of the nesting force moments. The effect of the location of the nesting force and third pin is less clear. A plot of the nesting force as a function of the angles \( \phi \) and \( \theta \) gives insight
The angle between the third pin and the \( x \)-axis is bound to the region between \(-45^\circ\) deg and \(80^\circ\) deg. The lower bound prevents the third pin from interfering with the right pin. The upper bound prevents over-constraint in the \( y \)-direction by maintaining a clearance between the vertical coordinate of the top of the wafer and the contact point between the wafer and the third pin. The lower bound for the nesting force angle is \(100^\circ\) deg to prevent over-constraint, and the upper bound for the nesting force angle is \(315^\circ\) deg in order to prevent interference with the left pin. Limiting the resulting moments to those with the appropriate orientation as shown in Fig. 4 allows for the visualization of the nesting force window as a function of third pin angle \( \theta \). The results of this exercise are shown in Fig. 5, and it is evident that the restoring moments about the ICRs are maximized as the third pin angle is moved toward its upper bound.

A prototype alignment mechanism was designed and fabricated to test the repeatability of the optimized design. Five-millimeter dowel pins were press fit into a block of 6061 aluminum to serve as constraints. The left pin and the right pin were permanently fixed such that they made contact with the wafer at a distance of 4 mm from the ends of the 32.5 mm flat on the wafer. Holes were drilled for the third pin at angles of \(-45^\circ\), 0, 45, 70, and \(80^\circ\) deg with respect to the \( x \)-axis. A prismatic flexure was designed to provide a nesting force of approximately 10 N to the wafer at an angle of \(135^\circ\) deg with respect to the \( x \)-axis. The nesting force angle was selected to strike a balance between maximizing the restoring moments about ICRs and generating approximately equal reaction forces at each of the pins. The surface of the alignment mechanism is recessed so that the central axis of the flexure comes into contact with the wafer in order to minimize torsion. Three vee-blocks were countersunk into the stage in an equilateral triangle configuration providing a stable kinematic interface for the half-spheres that extended from the mating translation stage. A photograph of the final prototype is shown in Fig. 2(d).

Experiments were performed to determine wafer placement repeatability as a function of third pin angle \( \theta \). Lateral repeatability was determined by measuring the distance from three capacitance probes to a reference block of aluminum bonded to a silicon wafer. Two capacitance probes separated by a known distance and with faces parallel to one face of the block measured translational repeatability in the \( x \)-direction and rotational repeatability about the \( z \)-axis. Translational repeatability in the \( y \)-direction was measured with an additional capacitance probe orthogonal to the two in the \( x \)-direction. On the first placement of the wafer, the capacitance probe measured distances were nulled such that each additional measurement was made relative to the first measurement. Between measurements the wafer was completely removed from the stage and then carefully hand-placed back on the stage. For each experiment, 50 measurements were recorded with the maximum and minimum values discarded. Repeatability was defined as the standard deviation of the trials. The square root of the sum of the squares of repeatability in the \( x \)- and \( y \)-directions was used as a measure of overall lateral repeatability. An image of the experimental setup is shown in Fig. 6.

Optimal repeatability was achieved with a coupling configuration that maximized the moments about the instantaneous centers of rotation. Figures 7 and 8 show the translational and rotational repeatability varied by location of the third pin. The experiment tests third pin in five locations, \(-45^\circ\) deg, 0 deg, \(45^\circ\) deg, \(70^\circ\) deg, and \(80^\circ\) deg. Positioning the third pin at an angle of \(80^\circ\) deg resulted in translational repeatability of 1.4 \(\mu\)m. Repeatability of other pin locations was skewed by a number of outliers with positioning errors that are an order-of-magnitude greater than the mean. Figures 9–11 give insight into the distribution of the repeatability data. The blue boxes bound results that fall between the 25th percentile (Q1) and 75th percentile (Q3) values of sample data. Median values for each pin location are indicated by a red line within the blue boxes. Statistical outliers are shown as red crosses.
and are values that are greater than \( Q_3 + 1.5 \times (Q_3 - Q_1) \) or smaller than \( Q_1 - 1.5 \times (Q_3 - Q_1) \). In a number of cases, rotation about the \( x \)- and \( y \)-axes was observed.

These results suggest that 10 N is the maximum safe level of force to apply as wafer damage occurred with thicker flexures. Repeatability without the nesting force was challenging to measure as results were often outside of the \( \pm 125 \mu m \) range of the capacitance probes used in the study.

**Flexure Stage Design**

Flexure mechanisms are commonly used in micropositioning \( XY \)-stages due to their superior isolation of motion between the \( X \)- and \( Y \)-axis and their great difference between in-plane and out-of-plane stiffness. There are many studies of double parallelogram flexure mechanisms (DPFM), which demonstrate extreme precision with mm-scale displacement range [9–11]. Figure 12 shows a simple model of double parallelogram flexure mechanism.

Because of the mm-scale deflection of the flexure beams, stiffness in the axial direction is affected by tangential displacement and tangential stiffness is affected by axial force [6]. That is,

\[
K_a \approx \frac{1}{\left( w^2 + \frac{9}{25} w^2 \right)} \frac{12EI}{L} \tag{3}
\]

and

\[
K_t \approx 12 \times \frac{3}{100} \left( \frac{F_a L^2}{EI} \right)^2 \frac{EI}{L^3} \tag{4}
\]

where \( w \), \( L \), and \( E \) are the flexure width, beam length, and Young’s modulus, respectively, and \( I \) is the second moment of area of the flexure beam.

Figure 13 shows a preliminary design for the \( XY \) positioning stage using the double parallelogram flexure mechanism. The image on the left is the top stage with the AFM chip and image on the right is the wafer sample stage.
The in-line inspection system concept with a flexure stage used to position the AFM chip and a sample stage used to hold the wafer for inspection. Kinematic coupling is used to quickly and precisely align the flexure stage to the sample stage.

The $X$ and $Y$-motions of the flexure are driven by a set of micrometer heads shown as green and red arrows in Fig. 13. Each axis of motion can be modeled as a spring system with two parallel sets of springs and two series sets of springs ($K_t$ and $K_a$). Since stiffness in the axial direction is much greater than stiffness in the tangential direction, total stiffness in the $X$-direction should be twice stiffness of the DPFM in the tangential direction. Sensitivity analysis for those multiple variables demonstrates that only variations in the beam length ($L$) and the flexure width ($w$) have a significant effect on the stiffness. We can plot the stiffness of the positioning as a function of the beam length and flexure width, as shown in Fig. 14.

In order to provide reasonable stiffness and to minimize the mechanism size, the flexure was designed to be cut from a 15-mm-thick block of 7075-T6 aluminum. A length of 20 mm and a width of 0.40 mm was selected for the flexures which resulted in a predicted 19.4 N/mm in-plane stiffness. The first-mode natural frequency of the stage was 133 Hz, which is two orders magnitude higher than largest frequencies generated by the environment. The actuating force in $X$- and $Y$-directions is 50 N in each axis, which results in 2.58 mm displacement to the chip-AFM positioning. Translation in the $Z$-direction and rotations about the $X$- and $Y$-axes are accomplished via actuation of three micrometers attached to the AFM stage. The spindle of each micrometer is press-fit to precision truncated balls, shown as Fig. 15. The balls interface with three vee-blocks, and when the micrometers are locked, the ball and vee-block coupling kinematically constrains all 6DOFs of the AFM stage relative to the wafer alignment stage.

Moreover, the kinematic coupling made by micrometers and truncated balls is capable of adjusting the $Z$ motion roughly. In order to avoid the damage of AFM scanning tip, it’s necessary to introduce another finely $Z$ motion to get AFM approach to sample slowly. This system utilizes another flexure set (Fig. 16) which is in symmetric design so that it can cancel most of the undesired parasitic motion. This flexure is pushed by a voice coil actuator which is controlled by an appropriate amplified circuit so that it could provide motion with nanoresolution.

The final flexure-based $XY$-precision stage was machined from a 15-mm thick 7075-T6 aluminum plate using a water jet cutting machine, shown as Fig. 17.

**Measurement**

Asymmetric arrangement of the flexure mechanisms and manufacturing error created in-plane yaw error motion. As a result, the motion of the stage differed from the cumulative input from the $X$- and $Y$-actuators. This parasitic motion was measured with two fiber-based optical displacements with 1.0-pm sensitivity and 100-mm working range. The setup is shown in Fig. 18.
Repeatability of the kinematic coupling between the AFM stage and the wafer stage was tested with capacitance probes. Three capacitance probes from LION Precision with 0.14-nm resolution and 2.0-mm working range were mounted on the optical table as the sensors for $X$, $Y$, and $Z$ displacements. The experimental setup is shown in Fig. 19. This setup was used to measure the repeatability in position during the repeated engagement of the kinematic couplings, as the two stages must be separated in order to change samples.

Results

In the flexure motion test, the stage exhibited 1.47-$\mu$m deviation in the $Y$-direction over 100-$\mu$m of actuation in the $X$-direction and 3.80-$\mu$m deviation in the $X$-direction over 100-$\mu$m of actuation in the $Y$-direction. The results indicated that the motion of the flexure-based stage is not in perfect orthogonal coordinate, as shown in Fig. 20. This error was caused by geometric asymmetry and manufacturing tolerances. However, the experiment discovered those errors are repeatable and therefore can be calibrated in the actuation system.

Finite element analysis was used to calculate the parasitic motion without manufacturing error. The result indicated 7.55-$\mu$m and −8.31-$\mu$m deviations in $X$- and $Y$-direction actuations within the 100-$\mu$m travel. The difference between parasitic motion test and simulation result suggests that the manufacturing tolerances contributed to the parasitic motion. Water jet cutting the stage yielded uneven flexure thicknesses which affected the motion of flexure.

Repeatability test constantly re-setup the flexure stage onto sample stage and measures position error within each trial. Results for the repeatability of the kinematic coupling were recorded for translation in $X$, $Y$, and $Z$ positions as well as rotated angle about the $Z$-axis of the stage. Table 1 summarizes the repeatability for each DOF. Repeatability is defined as the standard deviation calculated from the trials. In this study, the flexure-based stage has in-plane translational repeatability of 350–400 nm with in-plane rotation of 0.140 $\mu$rad and about 60-nm out-of-plane translational repeatability. Generally, AFM equipment scans in a couple of micron range, which is greater than the repeatable error. Therefore, an in-plane error of 400 nm is acceptable to be used in single-chip AFM operation.
Table 1  Repeatability performance in X, Y, Z, and rotation of X–Y plane for the flexure-based precision stage

<table>
<thead>
<tr>
<th>Repeatability</th>
<th>X (μm)</th>
<th>Y (μm)</th>
<th>Z (μm)</th>
<th>θ (μrad)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD (σ)</td>
<td>0.390</td>
<td>0.361</td>
<td>0.060</td>
<td>0.140</td>
</tr>
</tbody>
</table>

SD: standard deviation.

Fig. 21 AFM images taken using fast setup system: (a) One-dimensional grating sample which is 500 nm step structures with 3 μm pitch and (b) Two-dimensional grating sample, 100 nm height grid with 500 nm spacing. The units are shown in pixels.

Measurement Results

Using the fast alignment systems and chip-based AFMs described in this paper, several images were taken using the diffraction gratings. These measurements are shown in Fig. 21. Based on these scans, it is estimated that the z-height resolution of the AFM system is about 0.48 nm. Due to the fast setup design and the advantage provided by the chip-AFM, the total inspection cycle time is less than 1 min, and the preparation procedure is less than 15 s. Compared with most of the commercial AFM instruments which usually take tens of minutes for preparation and several minutes for imaging, this system is potentially capable of operating the in-line inspection in nanomanufacturing production.

Conclusions

Overall, this paper demonstrates that the time required to setup an AFM scan can be significantly be reduced by using chip-based AFMs and precision passive alignment systems. The passive alignment systems and flexure mechanisms integrated into the inspection system ensure that each time a wafer is loaded into the system it is in range of the AFM and no additional, time-consuming, alignment operations are required to take a measurement. The next step in this process is to incorporate multiple AFMs into the in-line inspection system in order to increase the area that is scanned for each wafer. By incorporating multiple AFMs into the system, it should be possible to analyze how nanomanufacturing processes vary over the entire processing wafer without needing to take the wafer offline for inspection. This type of inspection is therefore critical in order to do process control on nanomanufacturing systems.

References