REVIEW

Ultra-precision engineering in lithographic exposure equipment for the semiconductor industry

BY ROBERT-H. MUNNIG SCHMIDT*

Precision and Microsystems Engineering, Delft University of Technology, Mekelweg 2, Delft 2628CE, The Netherlands

The developments in lithographic tools for the production of an integrated circuit (IC) are ruled by ‘Moore’s Law’: the density of components on an IC doubles in about every two years. The corresponding size reduction of the smallest detail in an IC entails several technological breakthroughs. The wafer scanner, the exposure system that defines those details, is the determining factor in these developments. This review deals with those aspects of the positioning systems inside these wafer scanners that enable the extension of Moore’s Law into the future. The design of these systems is increasingly difficult because of the accuracy levels in the sub-nanometre range coupled with motion velocities of several metres per second. In addition to the use of feedback control for the reduction of errors, high-precision model-based feed-forward control is required with an almost ideally reproducible motion-system behaviour and a strict limitation of random disturbing events. The full mastering of this behaviour even includes material drift on an atomic scale and is decisive for the future success of these machines.

Keywords: lithography; precision positioning; mechatronics; metrology; stability; control

1. Introduction

The discovery of the semiconductor diode effect in 1897, by the German physicist Karl Ferdinand Braun, started a cycle that would change the technological face of the world. In the continuous search to unveil the secrets of semiconductor materials, the next significant milestone was the invention of the transistor by William Shockley, Walter Brattain and John Bardeen at Bell Laboratories in 1947 [1]. The real breakthrough in electronics came after the independent invention in 1958 of the integrated circuit (IC) by Jack Kilby, working at Texas Instruments, and by Robert Noyce, of Fairchild Semiconductor [2]. Since then, in almost all of the peopled locations in the world, several electronic products serve to enhance their environment.

*r.h.munnigschmidt@tudelft.nl

One contribution of 16 to a Discussion Meeting Issue ‘Ultra-precision engineering: from physics to manufacturing’.
An important factor in this worldwide proliferation of electronics is the efficient production process of large numbers of ICs that are realized on wafers (0.7 mm thin discs of mono-crystalline silicon) by means of the process that is visualized in figure 1. In principle, an IC is nothing else than a multitude of active and passive electronic elements, such as transistors, resistors and capacitors, arranged together on one single piece of semiconductor material, interconnected by a network of metal wiring in different layers.

The best-known semiconductor material is silicon. When applied in ICs, the silicon used has a mono-crystalline structure that is created by the Czochralski process. In this process, a large single crystal, the ingot, is grown out of molten silicon, starting with a small seed crystal. This ingot can have a diameter up to 300 mm, and thin wafers of approximately 0.7 mm thickness are cut from this ingot by means of very thin grinding discs. Silicon is a very brittle ‘stone-like’ material. Apart from grinding and other abrasive techniques, it can be machined only by chemical etching in a process called lithography, which is ancient Greek for writing on stone.

(a) Lithographic manufacturing process

The lithographic manufacturing process of an IC consists of up to approximately 30 cycles. Each cycle defines a specific layer and starts with a chemical pre-treatment such as the application of an oxidation layer that is needed to isolate the different electronic components. This is followed by the application of a photosensitive layer, called resist. An optical imaging process exposes the resist with the pattern of the specific layer. The resist areas that received light are removed during the development of the resist, and the material beneath that

Phil. Trans. R. Soc. A (2012)
is no longer covered can be treated in a subsequent chemical treatment. This treatment can, for instance, be a selective removal of the oxide layer by plasma-etching, doping of the material by implantation, depositioning of another material or other process steps that determine the functional electronic device.

Most of the layers are used for the interconnecting metal wiring that have to connect the functional parts, defined in the first layers, without interfering with any of the wiring in the other layers. Especially, the composition of the active first layers of the device requires an extreme level of accuracy in the relative positioning of the layers. This relative positioning is called overlay and is one of the primary technology drivers in the precision engineering of the wafer scanner used in the definition of the pattern.

While all other process steps work on the wafer as a whole, the exposure-process step by the wafer scanner determines, to a large degree, the details and relative position of the pattern structures. Continuing developments in these machines have thus been of overriding importance for the technology roadmap of the semiconductor industry with ever faster and more powerful microprocessors, memory and other devices.

(b) The wafer scanner

The original pattern of a specific layer of the IC is defined in a chromium layer on a reticle, a quartz plate of six-inch square that is made unique for every layer. In the early days, direct shadow exposure with a contact-mask was sufficient to define the pattern. Owing to contamination and lack of resolution, optical imaging methods were soon preferred over this simple contact method. The resolution of an optical system is determined by the wavelength ($\lambda$) of the light, the numerical aperture (NA) of the imaging system and a $k$-factor that is mainly determined by the illumination system. This relation is shown in figure 2, where the ‘critical dimension’ (CD) defines the smallest details that can be imagined reliably according to the following equation:

$$\text{CD} = k \frac{\lambda}{\text{NA}}.$$  

From this relation, it is clear that a short wavelength, a high NA and a low $k$-factor are needed to achieve a small CD. The need for a short wavelength has resulted in a continuous development of powerful light sources with ever smaller wavelengths, starting with mercury arc-discharge lamps with 436, 405 and 365 nm wavelength, followed by excimer lasers with a wavelength of 248 and 193 nm. The next step will be the extreme ultraviolet (EUV) source with 13 nm wavelength, in which case only optics with mirrors can be used for the exposure. The necessary improvements to the $k$-factor have resulted in the development of several optical techniques that allow a minimum of captured diffractive orders of the image in the aperture of the lens.

It was the required NA that forced the industry to cover the wafer in parts and not as a whole. The NA is defined as

$$\text{NA} = n \sin \vartheta,$$

where $n$ equals the refractive index of the medium between the lens and the image and $\vartheta$ equals the angle of incidence between the outer rays and the normal on the image surface. An NA of just less than one can be achieved in air ($n = 1$) when the opening angle would approach 90°. Exposing a large surface with this opening angle would be feasible only with extremely expensive lenses with a very large diameter. For that reason,
Review. Ultra-precision lithography equipment

Figure 2. The smallest details (CD) that can be imaged by a wafer scanner are determined by the numerical aperture (NA) of the imaging system.

\[ \text{NA} = n \sin \theta \]

\[ \text{CD} = \frac{k \lambda}{\text{NA}} = 0.5 \, \mu \text{m} \]

Figure 3. State-of-the-art wafer scanner from ASML indicating the main functional modules for exposing a wafer. The illuminator (1) directs the exposure light to the reticle on the reticle stage (2). The pattern of the reticle, defined in a chromium layer on a quartz plate, is imaged by the projection lens (3) onto the wafer (4) located on the wafer stage (5). The wafer handler (6) exchanges the wafer after each exposure while the reticle handler (7) exchanges the reticle for each different layer that is required for a functioning IC. (Courtesy of ASML.)

it was assumed more affordable to use a smaller lens and expose the wafer in steps. This was first performed by means of a wafer stepper, but later the exposure was carried out by a scanning motion. Figure 3 shows the main modules of a modern wafer scanner, whereas figure 4 explains the scanning principle. For practical reasons, the NA in air is maximized at around 0.93, and because of that limitation,
most modern wafer scanners use water as intermediate medium with a refractive index of $\approx 1.44$ at 193nm. This has increased the practical value of the NA to $\approx 1.35$. In a wafer scanner, the reticle is positioned by a reticle stage that performs a scanning motion, where the illumination system illuminates only one part of the mask, the ‘slit’, a light stripe. This scanning motion has several advantages in respect to stepping with a stationary exposure as used in wafer steppers, because of the larger attainable size of the image field and the averaging of small image position errors. The projection lens images the mask on the silicon wafer. This wafer is positioned by means of the wafer stage that moves synchronously with the reticle stage. Because of the four-times demagnification of the projection lens, smaller details can be imaged from a relatively large reticle. The disadvantage of this demagnification is that the reticle moves four times as fast in the opposite direction as the wafer stage with corresponding higher acceleration levels and reaction forces. This demagnification is however necessary to be able to produce reticles with a sufficient quality by means of electron beam pattern generators.

(c) Requirements on precision

The overlay and the CD are the two main drivers of precision in a wafer scanner. The CD is often defined as ‘half-pitch’ ($0.5 \times p$), half the distance between the centres of parallel lines as indicated in Figure 2. The continuous decrease of the CD is almost like a law of nature determined by Moore’s Law [3]. Gordon E. Moore, co-founder of Intel, defined his law based on the observation that the number of transistors on an Intel processor doubled every 2 years. Although not a real law of nature, this postulation has predicted the continuous exponential growth rate over the last three decades, with a corresponding ‘shrink’ of the smallest details. The industry has organized itself so well in this respect that it even uses a jointly defined roadmap, the International Technology Roadmap for Semiconductors, ITRS [4], that describes in full detail the expected developments of the different parameters that rule this market for a 15 year forecast.

Overlay is the relative position of any pattern layer with respect to the other layers. As can be observed in Figure 18, all layers need to be connected by a multitude of small conductive pillars, the so-called vias. Overlay errors impact...
both the electrical properties of the contacts and the insulation and might even create short circuits. The required overlay value is strongly related to the CD. Originally, an overlay value of the order of 30 per cent of the CD could be sufficient, but actual values are more close to 15 per cent, because of special exposure methods such as double patterning or double exposure, where thin lines are exposed in between the lines from a previous exposure cycle [5]. These methods enable IC manufacturers to extend the use of 193 nm wavelength light to ever smaller dimensions, but it has simultaneously resulted in a strong increase in the overlay requirements beyond Moore’s Law, as can be seen in figure 5.

The third factor of importance for precision is the productivity, which indicates the amount of successful exposures per unit of time. A derived and more frequently used factor is the throughput in exposed wafers per hour that does not include the exposure quality. The enormous cost of these machines, mainly driven by the optical parts, requires an ever increasing speed of operation. Presently, more than 180 wafers can be exposed per hour, which corresponds to 20 s per wafer, including loading and unloading. The related extreme velocities and accelerations of the stages and other robotic motion systems cause strong reaction forces with vibration levels that easily could impair the CDs by fading, lack of contrast by vibrations during exposure.

All these factors result in a large set of requirements for the wafer scanner. The list in table 1 gives an overview of a selection of the most important requirements that directly determine the precision positioning systems, the wafer and the reticle stage. The values are approximated as they are meant only to give a reasonably realistic idea of the order of magnitude.

On the basis of these requirements, this study focuses on the following important aspects that have played a role in the achievement of these extreme levels in performance and will pose new challenges for future developments [6].

— Dynamic architecture, preventing interfering vibrations of the sensitive parts.

Phil. Trans. R. Soc. A (2012)
Several measures are taken to avoid dynamic problems in modern wafer scanners. The most important is the area around the lens and the metrology frame. (Courtesy of ASML.)

Table 1. Main requirements on precision of the positioning systems in a wafer scanner.

<table>
<thead>
<tr>
<th>requirement</th>
<th>approximate value</th>
</tr>
</thead>
<tbody>
<tr>
<td>critical dimension (CD)</td>
<td>&lt;40 nm</td>
</tr>
<tr>
<td>overlay</td>
<td>&lt;4 nm</td>
</tr>
<tr>
<td>wafer stage velocity, stepping</td>
<td>&gt;2 m s(^{-1})</td>
</tr>
<tr>
<td>wafer stage velocity, scanning</td>
<td>&gt;0.5 m s(^{-1})</td>
</tr>
<tr>
<td>reticle stage velocity</td>
<td>&gt;2 m s(^{-1})</td>
</tr>
<tr>
<td>wafer stage acceleration</td>
<td>&gt;30 m s(^{-2})</td>
</tr>
<tr>
<td>reticle stage acceleration</td>
<td>&gt;120 m s(^{-2})</td>
</tr>
<tr>
<td>wafer stage metrology error</td>
<td>&lt;0.5 nm over 20 s</td>
</tr>
<tr>
<td>wafer stage in-plane impact on overlay (average position)</td>
<td>&lt;1 nm</td>
</tr>
<tr>
<td>wafer stage in-plane impact on fading (vibrations)</td>
<td>&lt;10 nm</td>
</tr>
<tr>
<td>deviation of the wafer surface level from ideal focus</td>
<td>&lt;100 nm</td>
</tr>
<tr>
<td>settle time to reduce position errors before exposure</td>
<td>&lt;10 ms</td>
</tr>
</tbody>
</table>

— Zero-stiffness stage actuation with Lorentz actuators to fulfil the demands on dynamic performance.
— Indirect relative position measurement between image and wafer with alignment marks and long-range sensors.
— Motion control by a combination of calibrated model-based feed-forward control and a well-tuned proportional–differential–integrating (PID)-feedback controller.
— Thermal control and structural stability as the last frontier of precision.
2. Dynamic architecture

(a) Balance masses

One major rule in precision engineering is to first prevent problems by fully mastering the physics-based intrinsic dynamics (eigendynamics) of the system, before attempting to control its position. In the wafer scanner, the dynamics of the optical imaging system is most sensitive to dynamic excitations and needs to be protected from any interfering forces. Sources of interference include vibrations from other equipment such as the large air-conditioning, purifying and processing equipment in a wafer fab, the usual name for a semiconductor factory. Other important sources can be found within the wafer scanner itself. The immense acceleration levels of the quite heavy stages create reaction forces of the order of several kilonewtons. In the first generations of wafer scanners, these reaction forces were directed towards heavy frames, mounted separately of the machine to the ground. This caused much problems, for the dynamics of the fab floor, for any other equipment nearby, and for the sensitive parts inside the machine that indirectly received the vibrations owing to these forces. For that reason, these reaction forces are absorbed in the recent generations of wafer scanners by very heavy balance masses that are guided on air bearings or compliant mechanisms (figure 6). These measures have helped in reducing the remaining reaction forces to below sufficiently harmless levels, because most of the forces are directed in a plane parallel to ground.

(b) Vibration isolation

In spite of the balancing measures, the area around the metrology frame and the projection lens needs high-grade vibration isolation measures to prevent any remainder of vibration from entering into this sensitive area. The first line of defence against interfering vibrations is determined by the air mounts. In principle, they consist of air cylinders with a large volume that are sealed with air-bearings to prevent transmission of vibrations via the sealing. Their stiffness can be of the order of $10^5 \text{ N m}^{-1}$ and with a combined mass of approximately 2500 kg of the metrology frame and the projection lens, the first natural frequency of these air mounts is of the order of 1 Hz. A problem occurs however owing to the relatively low stiffness. The total sag of the spring by the mass would be 25 cm and a small variation of this mass would already determine an unacceptable change in the position of the lens in respect to the wafer stage. For that reason, the height is actively controlled by changing the amount of air in the air cylinders.

The next problem with this system is the need for damping. At the natural frequency of 1 Hz, the system will resonate when excited with that frequency by a force or a movement of the floor. A normal viscous damper would introduce an unwanted connection with the vibrating other parts, increasing the transmission of these vibrations to the sensitive part as shown in figure 7. For that reason, a so-called skyhook active damper is applied. This principle is based on the measurement of the ‘absolute’ velocity, relative to a real quiet reference, imitating the sky. In practice, this quiet reference consists of an elastically suspended seismic mass inside the sensor. A suitable sensor with this principle is the geophone that measures the velocity relative to the seismic mass. The velocity
Figure 7. Active ‘skyhook’ damping in a vibration isolation system avoids the transmission of external vibrations through the damper. An absolute velocity detector measures the velocity of the sensitive body relative to an inert seismic mass, and its output is used to exert a force opposite to the velocity. The benefits are shown in the magnitude Bode plot of the transmissibility.

The velocity signal is used to create a proportional damping force, opposite to the velocity, by means of an actuator of which the force is determined only by the velocity signal and not by the position. As a consequence, this actuator has no stiffness as that would add to the stiffness of the supporting mechanical spring $k$, increasing the transmissibility.

Although this method of active damping is often used, its performance is limited by the noise of the sensor. In principle, the sensor needs to detect only very small movements and any noise source will insert an interfering signal in the damping loop. This is especially problematic at very low frequencies. Many investigations are being carried out to improve the active part of these vibration isolation systems, including actively reducing the stiffness of the connection, but owing to the sensor noise, it is in practice still preferred to have the main part of the vibration isolation based on a heavy mass and a compliant spring.

(c) Eigendynamics

A third important dynamic property of the system is related to the eigendynamics of the stages. Every mechanical system can be dynamically modelled as an infinite amount of bodies, interconnected by springs. The related mass and stiffness values determine several resonating eigenmodes, modes of elastic deformation that are distinguished by a specific mode-shape with a corresponding eigenfrequency where the resonance occurs. One can think as an example of a tuning fork where the first eigenmode is described by the alternating in-and-out swinging movement of the legs as mode-shape and the 440 Hz tone (note A) as the eigenfrequency. From a motion control point of view, these eigendynamics are the main determining factor in the behaviour of

Phil. Trans. R. Soc. A (2012)
the closed-loop feedback-controlled system. For that reason, the eigenfrequencies need to be higher than the required bandwidth frequency of the feedback system. The bandwidth determines the maximum frequency that a system can follow and is a measure for the capability of the control system to reduce lower frequency errors.

3. Zero-stiffness stage actuation

With the active vibration isolation, an actuator was introduced to exert the damping forces without introducing a stiff connection between the sensitive part and the vibrating part. In a similar way, also the requirements on the accuracy of the stages, with their high level of velocity and acceleration, require the use of these so-called ‘zero-stiffness actuators’ to an extreme level, aiming to completely avoid any elastic connection between the stage and the outside world. In essence, the positioning principle is based on Newton’s Second Law that states that the acceleration of a body is proportional to the force and inversely proportional to the mass of the body. As long as the forces are known and controlled, a body with a known mass can be accurately positioned by doubly integrating the accelerations, first to velocity and then to position. In the design of a positioning stage for a wafer scanner, most of the effort is invested in full control of the forces acting on the moving body, consisting either of the wafer and its support or of the reticle and its support. As long as these bodies behave like a solid body with predictable and reproducible dynamics, the movement of the body is determined only by the forces acting on it.

With this reasoning, the term ‘zero stiffness’ is related only to the mechanical connection to sources of interfering vibrations and does not refer to the control stiffness. When zero stiffness is achieved, the body will behave solely like an inertial mass, and the task of the control system is to connect the stages as stiffly as possible with respect to the image and object positions of the optical system. More precisely stated, the control system needs to connect the reticle stage as stiffly as possible to a planned scanning trajectory, relative to the optical system while simultaneously connecting the wafer stage as stiffly as possible to the scanning image of the moving reticle.

(a) The actuators

As stated earlier, this zero-stiffness requirement implies for the actuators that they are not allowed to transmit vibrations over the connection between the moving and stationary part of the actuator. From all possible actuator types only one electromagnetic version, working according to the Lorentz principle, almost ideally complies with this zero-stiffness criterion. Piezoelectric actuators are stiff by definition and reluctance-based actuators have a negative stiffness that induces a comparable transmission of vibrations as a positive stiffness.

Figure 8 shows the basic working principle of such a Lorentz actuator in a loudspeaker-type moving-coil configuration. A permanent magnet creates a magnetic field $\mathbf{B}$ in an air gap with flux density $B$. A coil, consisting of a number of windings with a total length $\ell$, carrying a current $I$, will experience a force $F = B I \ell$, when the entire coil is surrounded by an ideally uniform magnetic field.
In this ideal situation, the force is linearly related only to the current and not to the position. In reality, however, there is only a small area where this zero-stiffness situation is realized and also parasitic nonlinear reluctance-type forces play a role. As long as the range of movement is kept small, in practice ±0.1 mm around the mid-position of the coil in the magnetic field, stiffness values below $10^3$ N m$^{-1}$ can be realized with actuators that deliver more than 3 kN force, when the coil is connected to the stationary part, and the permanent-magnet system is connected to the moving part. If the magnet and coil would be reversed, then the wires and water-cooling tubes to the coil would result in a much higher connecting stiffness.

A second actuator, the long-stroke actuator, is added because of the small zero-stiffness range of the Lorentz short-stroke actuator. This long-stroke actuator is connected to the stationary part of the short-stroke actuator and needs to position only this part such that the permanent-magnet system of the short-stroke actuator remains in the ±0.1 mm range of the coil with the lowest stiffness value. This configuration allows the long-stroke actuator to work with a reduced precision of only this same level of ±0.1 mm, even though the precision of the short-stroke actuator needs to be on sub-nanometre levels. Because of the low stiffness of the short-stroke actuator, the movements and vibrations of the long-stroke actuator are hardly transferred to the sensitive wafer position. For that reason, strong electronically commutated three-phase actuators can be used in the long-stroke actuation system that are able to transport the cables and tubing for the short-stroke coils. A schematic of this principle is shown in figure 9. Fortunately, the in-plane motions are large in a wafer scanner, so those motions require only a corresponding long-stroke actuator. The out-of-plane motions for focus adaptation are realized only by the short-stroke actuators.

This H-configuration was originally designed as an improvement to the often-applied, stacked single-directional positioning stages, to avoid tilting torques at acceleration. With an H-configuration, all forces act in one plane close to the centre of gravity of the moving body, the wafer table, also called the wafer chuck.
Figure 9. A long-stroke, short-stroke positioning stage in an H-configuration, where the actuator forces all act in the plane of the centre of mass of the wafer table.

Figure 10. The forces acting on a coil, positioned in the magnetic field above an array of alternating permanent magnets. Depending on the position in the field, the forces can be in-plane or perpendicular or a combination.

(b) Full magnetic levitation

Until recently, the wafer stage in the H-configuration was mainly guided by means of air bearings. In future wafer scanners with EUV light of 13 nm, air is not allowed anymore in the light path, so these machines are operated in vacuum. Although in principle air bearings can be realized in vacuum, by adding special measures to locally extract the inserted air, it is better to avoid them. For that reason, a long-stroke stage has been developed that uses an electromagnetic planar drive. Its working principle is shown in figure 10. A magnetic platform, consisting of a plurality of alternating permanent magnets, creates a magnetic field with different field directions. A current-carrying coil will experience a force that can be in any direction, depending on its position on the magnetic table and the current direction. By using different coils at different locations, a configuration
can be created that is able to exert forces and torques in all six orthogonal directions at any position of a free moving stage, when combined with a suitable electronic amplifier system.

(c) Future challenges in actuation

The described stage concept with long- and short-stroke actuators has been the standard in this industry for many years. The continuous increase of acceleration levels, however, poses several challenges that are increasingly not trivial. The main issue is that both actuators are essentially placed in series, which means that they both need to deliver the same force to the wafer table.

Next to its almost ideal zero-stiffness behaviour, a Lorentz actuator has fundamental limitations in its force and power capability that are thermally determined. This limitation can be understood from the following reasoning sequence.

When the acceleration of a known body with a certain mass needs to increase, the current in the actuator needs to increase proportionally. The power dissipation in the resistance of the windings is however proportional to the current squared \(P = I^2R\), and when the motor initially would be sufficiently cooled, it will no longer be at the increased current level.

Measures to increase the cooling capability and simultaneously reduce the dissipated power both automatically lead to an increase of the mass. A reduced resistance, needed for a reduced power dissipation, implies a larger volume of the coil and that requires also a larger permanent-magnet system. At a certain force-level per unit of mass, this increase of mass cancels the benefit of the increase of force and the resulting acceleration will no longer increase. This acceleration ceiling is not absolutely firm, as it depends on heat flow and other factors, but it is a serious issue, where research is required to enable further improvements. Part of this research focuses on the reduction of the moving mass by hollow structures with thin plates. As a side effect of the reduced mass, the sensitivity for external interfering forces by transmission is increased. This subject is called the mass dilemma of precision positioning systems and requires improved system dynamics and control to achieve new levels of precision. It will return further on in this study when presenting the motion control system.

4. Relative indirect position measurement

It is very important to be aware that the requirements on overlay of a few nanometres are only strictly relative. There is no traceability relation with the standard metre other than in global terms. Traceability is not important, and all calibrations are done relative to the previous layers on the wafer.

One other important aspect is the impossibility to directly measure the image relative to the wafer. This means that the measurement can only be carried out indirectly in a sequence of different steps.

— The wafer is positioned on the wafer stage by means of the wafer handler with a maximum position error of approximately ten micrometres.
The wafer position and previous layer is referenced, aligned, to the wafer stage by means of alignment marks on the wafer and the fiducials on the wafer stage.

At exposure, the fiducials of the wafer stage are aligned to the reticle through the projection lens.

During these process steps, an incremental measurement system with a resolution of approximately 0.1 nm consistently relates the position of the wafer stage to the lens.

The principle of alignment with the alignment marks is shown in figure 11. It is a method related to the precision position encoder that will be presented further on, where phase gratings are used to achieve a relative position measurement. By illuminating the grating with a laser, the phase steps create diffraction orders of which only the first order is captured. By combining these first orders, their interference pattern will move over the photodetector in the lateral direction. The phase shift by the movement in the $+1$ order is opposite and equal in magnitude to the phase shift of the $-1$ order, resulting in a doubling of the combined differential phase shift, thereby increasing the measurement resolution by a factor of two [7].

(a) Dual-stage measurement and exposure

It is necessary to actively position the wafer within the focal area of the projection lens at the region of the exposed die. This focusing is further complicated by the modification of the surface of the wafer during the manufacturing process. Over the years, this has made it increasingly difficult to guarantee that the entire surface is sufficiently well aligned within the depth.
of focus. Figure 12 shows the relation between the depth of focus and the CD, and it is clear that a flatness of less than 25 nm over a die of 25 mm is a challenging requirement. This problem is tackled from two sides, flattening the surface by additional means and active control of the average wafer surface relative to the image plane. In the manufacturing process, the deposited metal layers for the wiring are often treated by a chemical mechanical polishing process to keep the surface non-planarity locally within acceptable levels.

The active positioning of the first wafer scanners used real-time measurement of the surface in the moving direction ahead of the lens. The levelling control was done with a servo-tracking feedback action to keep the surface always parallel levelled to the image in the focal plane. Feedback control can never prevent some remaining focal error owing to the positioning system dynamics and the limited control gain. This remaining error can be reduced only by measuring of the surface profile before the scan, to be able to position the wafer stage by feed-forward control. This was one of the reasons to apply a dual-stage configuration [8], with a separate measurement position and an exposure position. The other reason for such a configuration is the possibility to increase the utilization of the expensive lens. With a dual wafer stage, the lens can expose one wafer during the time that the other wafer is measured, and as a consequence the lens is used continuously and productivity (throughput) is increased.

Figure 13 shows the process steps of a dual-stage wafer stage. The measurement cycle is synchronous with the exposure cycle and as a result of the measurement, the exposure starts with a wafer that is fully known in respect to the position of the alignment markers to the wafer stage and the height profile of the surface. This enables a maximum use of feed-forward control where the remaining error is determined only by the measurement error in the measurement cycle.

(b) Wafer curvature correction on a die

As a next step to achieve a better focus performance, research to use a curved focal plane (instead of a flat focal plane) in wafer scanners, as shown in figure 14, is ongoing. The double telecentric projection optics in these systems allows us to achieve the desired curvature correction by active deformation of an optical element within the exposure system [9,10]. The importance of this kind of adaptive optics is expected to increase in the following years to keep track of the continuous demands on focus.
Figure 13. The process steps in a dual stage wafer scanner between loading and unloading consist of two cycles. In the measurement cycle, the wafer is aligned to the wafer stage and the surface is mapped for focus. In the exposure cycle, the wafer stage is aligned to the reticle and the wafer is exposed. (Courtesy of ASML.)

(c) Long-range incremental measurement system

The long-range measurement systems in the wafer and reticle stages need to be able to track the position with the specified accuracy and velocity levels. Until not so long ago, this kind of performance was achievable only by means of laser interferometer measurement systems. Especially, the heterodyne interferometers, using a dual-wavelength orthogonally polarized laser and a polarized Michelson interferometer, have served this industry as of the very beginning. The main problem with these systems is that they measure over a long distance through air. Air will always show small variations in the light velocity that have a significant influence on the measurement. The variation of the light velocity in air is mainly determined by the temperature, but also the humidity and the pressure play a role. For the temperature, the sensitivity of $\Delta L/L = 10^{-6}$ per kelvin requires the maximum change in temperature in the measurement beam to be less than 1 millikelvin for a distance of 500 mm with a maximum allowable position error of 0.5 nm.

Only recently, it became possible to design a planar encoder system with a sufficient resolution to operate as the long-range measurement system of a wafer scanner as schematically shown in figure 15 [11]. Its main advantage is the short path through air of less than 15 mm, reducing the thermal sensitivity by a factor of 100 to $\Delta L/L = 10^{-8}$ per degree Celsius. As a comparison, figure 16 shows the different drive and measurement configurations.
Figure 14. By curving the image plane, the image can be made to better conform to the local curvature of the wafer. This can be achieved by deforming an optical element in the projection optics by means of piezoelectric elements.

Figure 15. Two methods for long-range incremental position measurement. (a) The laser interferometer has long been the standard but is partly replaced by (b) the plane encoder system. (Courtesy of ASML.)
Figure 16. The combination of (a) the H-configuration with the laser interferometer measurement system and (b) the plane encoder with the planar actuation system clearly illustrate the differences. (Courtesy of ASML.)

Figure 17. The position control of the wafer stage is a SISO six-axis PID-control system. The gain scheduling matrix (GS) transforms the control forces ($F_c$) in six orthogonal directions, according to the machine coordinate system into the forces ($F_{com}$) of the wafer stage coordinate system around the centre of mass. The gain balancing matrix (GB) transforms these forces into the actuator forces ($F_a$). (Courtesy of ASML.)

5. Motion control

Figure 17 shows the basic principle of stage control. The control expert will miss elements like multiple input multiple output control, observer-based control and other mathematics-based ‘modern’ control approaches. Indeed, the controller is only a straightforward single input single output (SISO) controller for all six directions (six-axis) separately and the feedback controller uses a well-tuned...
classical PID-control algorithm. The six-axis machine coordinate system for the position measurement and the controller has its orientation at the centre of the image position. This is not the place where the centre of mass of the wafer stage, nor the actuators of the wafer stage are located. For that reason the six control forces $F_c$ are transformed into the required forces on the actuators by two matrix transformations. The reason for this rather simple and straightforward approach in control is twofold. First of all, this methodology allows for a more direct investigation of errors in the system. This shortens the time for troubleshooting at the customer site, when the system failed, or when parts have been exchanged and the system should be tuned again. The second reason is that the feedback part only plays a limited role in the control of these systems. An often forgotten fact of feedback control is that it needs an error to act anyway. The related time to reduce the remaining error is called the ‘settle-time’.

(a) The importance of feed-forward control

Mainly because of the impact of settle-time on throughput, a lot of attention in the design of a wafer scanner is given to create a dynamic system that behaves as systematically and deterministically as possible and feedback should only be applied to correct the part of the position error that is caused by random interference or real uncertainties in the different elements of the system.

A small example can place this requirement in perspective.

Imagine a wafer stage with the specifications as mentioned in the list of table 1, a mass of 20 kg, accelerating with $30 \text{ m s}^{-1}$ to a maximum speed of $0.5 \text{ m s}^{-1}$. With these numbers the required force during acceleration becomes $600 \text{ N}$ and the acceleration time takes $\approx 17 \text{ ms}$.

Assume that this situation is the result of a pure, open-loop, inertia-based, feed-forward-controlled system with a maximum error in the force of only $0.6 \text{ N} (= 0.1\%)$. Then, a small calculation gives the following position error after the $17 \text{ ms}$ acceleration time:

$$
\varepsilon = 0.5at^2 = 0.5 \cdot \frac{0.6}{20} (17 \times 10^{-3})^2 \approx 4 \times 10^{-6} \text{ m.} \quad (5.1)
$$

This is still a factor of 400 above the required 10 nm level. Improving the error with this factor would mean a reduction of the deviation of the force to around 0.25 ppm ($1:10^6$) and that relative level can presently be realized only in electronics. This all clearly indicates the need for an additional feedback-control action to reduce the remaining error. A sufficiently high feedback loop gain will create an increased stiffness to the target position and this reduces the remaining error to the specified values when allowing a settle-time of 10 ms. This seemingly small value is however already a very costly sacrifice. An average 300 mm wafer has 100 dies so 100 times acceleration and deceleration. The 10 ms settle-time per scan would mean approximately 1 s time-loss from the 20 s total exposure time of the machine. This 5 per cent loss of time reduces directly 5 per cent of the customer value of the wafer scanner, with an average cost of many million euros each!

There are only two ways to further reduce this settle-time. The first is to increase the control loop gain. This however requires further improvements in the mechanical dynamics that were already assumed to be optimal. The second
possibility is an improved accuracy of the feed-forward control system. A force error of less than $10^{-5}$ for $\approx 400\text{nm}$ position error is even $10^{-6}$ for $\approx 40\text{nm}$ position error is then necessary, requiring the accuracy of the dynamic model to be at the same level. The errors in that model are mainly determined by changing parameters of the actuators by temperature and in practice such a low error level could possibly only be maintained when the total system is continuously calibrated.

(b) The mass dilemma

With all measures taken to a maximum, still these last errors have to be corrected in only a few milliseconds by the feedback control system and there the mass dilemma returns in its full magnitude! A reduced moving mass is strongly preferred in order to reduce the electrical power necessary for the high levels of acceleration and limit the reaction forces. As mentioned already, a reduced mass will however proportionally increase the sensitivity for external interference forces by transmission. This means that an increased feedback control loop gain is needed to correct that increased error. Fortunately, the reduced mass also results in higher eigenfrequencies, allowing this increased bandwidth, but that is a first-order approximation. In reality, a reduced mass, without reduced overall dimensions, can be realized only by a combination of thin structures and those are not by definition sufficiently stiff in all directions to guarantee high eigenfrequencies. Research is performed at several places to compensate these oscillating eigenmodes by means of over-actuation with more than one actuator per orthogonal direction [12,13]. With this method, the excitation of the lower-frequency eigenmodes can be avoided.

6. Thermal control

The measurement and positioning accuracy of the wafer needs to be stable only within a time frame of about 20s per exposure because of the continuous calibration of all relative measurement parameters between each exposure cycle. The extreme requirements on accuracy however even pose problems on stability in this short time. For that reason, the interferometer position-measurement system always relied on an extremely stable monolithic structure of the wafer stage between the wafer and the measurement mirror. The metrology loop of that system however also included the mounting of the interferometers to the metrology frame and the metrology frame itself. Several measures have been taken to safeguard the stability such as the choice of Invar for the material and active stabilization of the temperature of the metrology frame.

With the change to the encoder-based long-range measurement system, the architecture changed fundamentally. From that moment, the sensors were positioned on the moving stage, and the monolithic structure was abandoned. The grating is directly connected to the metrology frame and made from Zerodur, a ceramic glass with an extremely low thermal expansion coefficient. This almost completely cancels measurement errors owing to thermal drift in the metrology frame. All effort now concentrates on the sensors in the wafer stage. These sensors consist of different parts and that alone almost guarantees stability problems owing to drift in glue and other mountings. These problems have been solved by
a fast in-line correction principle, using eight sensors for six orthogonal directions. With the well-defined and calibrated two-dimensional grating scale, the two additional sensors produce correlation data with the other sensors that enable an instantaneous correction of any drift that might have occurred in that time frame. In spite of these measures, the influence of temperature will remain important and in depth research is performed to further reduce these effects in the applied materials and connections [14].

7. Future developments

Already at the time of the first wafer stepper at Philips Research some 30 years ago, experts were doubting the future of optical lithography systems, owing to the limitations in wavelengths and materials for lenses. Also then, people were already wondering when it would become impossible to create isolated structures in silicon at ever smaller sizes. Even nowadays, in spite of the continuous proof of Moore’s Law, this continues to be a key question in the community, and figure 18 shows as an example how next steps are taken at IBM to still make this miracle happen.

Will Moore’s Law continue? The most probable answer is: ‘Yes, when necessary!’

This answer is based on the observation that another more technological law, one might call it Moore’s Second Law, seems to be hidden behind his first law and that is:

When an invention is needed for a next step on the semiconductor roadmap, it will be done!

Phil. Trans. R. Soc. A (2012)
Knowing that a single atom of silicon is about 0.25 nm in diameter, it seems still a long way before that size becomes a problem, but already now other materials like ‘graphene’ [15] and photon switches [16] start to open up new options for further miniaturization and increased performance of electronics. One of those related interesting questions is whether we will ever be capable of using the full three-dimensional space of a silicon wafer. Although many circuits can be seen as a two-dimensional+ structure, where sometimes new active layers are built on top of the other layers, essentially ICs are still mainly two-dimensional. When this limitation is taken away, Moore’s Law could continue without the need for smaller details.

In view of the related difficulties, it is however not to be expected that the present race into ever refined structures will end in the near future. The introduction of EUV lithography will enable further steps until far into the next decade, and inventions will be done to make this all possible. Undoubtedly, precision engineering and mechatronics continue to play an important role in that future.

**Glossary**

critical dimension (CD): the smallest imageable detail of an IC pattern
overlay: relative position accuracy between different layers
wafer: thin disc of semiconductor material
reticle: object with the pattern of one layer of an IC
productivity: amount of successful exposed wafers per unit of time
throughput: amount of exposed wafers per hour
stage: positioning system

I thank all sources of information and especially ASML for the use of several illustrations and their kind cooperation in reviewing this study.

**References**


