MICROLITHOGRAPHY: TRENDS, CHALLENGES, SOLUTIONS, AND THEIR IMPACT ON DESIGN

WITH LITHOGRAPHY PARAMETERS APPROACHING THEIR LIMITS, CONTINUOUS IMPROVEMENT REQUIRES INCREASING DIALOGUES AND COMPROMISES BETWEEN THE TECHNOLOGY AND DESIGN COMMUNITIES. ONLY WITH SUCH COMMUNICATION CAN SEMICONDUCTOR MANUFACTURERS REACH THE 30-NM PHYSICAL-GATE-LENGTH ERA WITH OPTICAL LITHOGRAPHY.

Alfred K. Wong
University of Hong Kong

IC technology is the foundation of high-performance computers. The continuous increase in transistor density and decrease in gate length provide the fabric from which designers build ever-faster computers. The number of transistors on a chip has been increasing by 40 percent per year. This exponential increase—among many other trends in the semiconductor industry—approximates a straight line when plotted on semilog paper. Such an increase is known as Moore's law.[1]

Doubling the transistor count every two years has been the semiconductor industry's economic engine. Although a 40 percent increase per year requires an annual 15 percent increase in investment, the cost per transistor decreases by 25 percent each year. This cost reduction enables the production of ever more complex components at an approximately constant price. Historically, these factors result in an annual 15 percent increase in number of units sold and, hence, revenue. The industry reinvests part of the increased income to sustain the cycle.

The economic health of the semiconductor industry thus depends on the exponential increase in transistor count per chip, and creating these billions of circuit elements repeatedly and reliably is a daunting task. A successful microlithography technique should define small features whose size decreases with each technology generation. It must replicate patterns to their intended dimensions with little variation over the entire circuit area and place these patterns accurately with respect to previously defined patterns. The placement accuracy is approximately a third of the minimum feature size.

Optical lithography

Technical requirements aside, a requisite for production microlithography is low cost of ownership and operation. Because economics is the driving force behind the dramatic circuit miniaturization of the past few decades, a successful microlithography technique must have low cost of ownership. Productivity is critical. Because of its ability to process wafers quickly, projection optical lithography has become the preeminent microlithography technique, preferred over alternatives including x-ray, electron beam, and ion beam lithography.

Published by the IEEE Computer Society

Basics

Optical lithography comprises the basic components illustrated in Figure 1. A *photomask* (also called a *reticle*) holds the circuit patterns for delineation and is illuminated by a light source. Because of the Fourier transform property of lenses,[2,3] energy transmitted through the photomask forms a distribution in the pupil plane that is proportional to the mask spectrum. The pupil acts as a low-pass filter. Low-spatial-frequency components pass closer to the pupil's center, and higher-frequency components are nearer the pupil's periphery. The filter blocks the highest frequencies. The transmitted frequency components project onto a wafer coated with a layer of photoresist, which functions as the recording medium.

We can encapsulate the physics of the exposure system with three parameters: the light source's wavelength, λ; the size of the pupil that restricts angular extent θ of the image-forming rays; and σ, the ratio of the light source's size to that of the pupil. Analyses of optical imaging frequently use sin θ. This quantity is the numerical aperture of the exposure system: $NA = \sin\theta$.

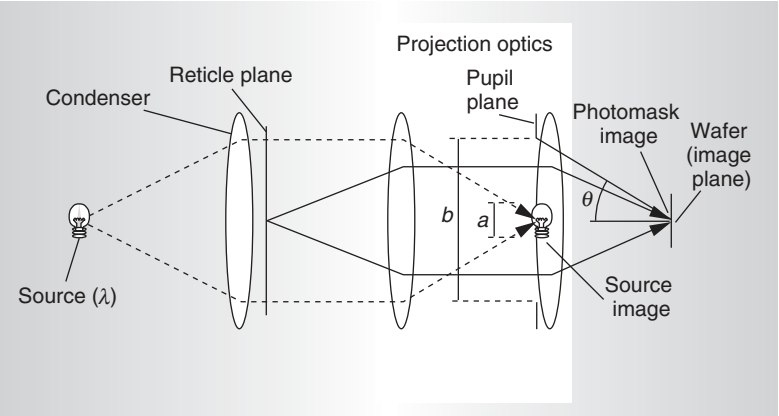

Figure 1. Basic components of optical lithography. In this figure, sin $\theta = NA$ and $\sigma = a/b$. Copyright SPIE, 2001. Reprinted with permission.[4]

Table 1. Changes in critical dimension.

Wavelength (nm)	Light source	Year of introduction	Decrease (percentage)
436	Mercury arc g-line	1970s	NA
365	Mercury arc i-line	1984	16
248	KrF laser	1989	32
193	ArF laser	1999	23
157	F_2 laser	After 2004	19
13	Plasma	Unknown	NA

Critical dimension and exposure system parameters

The critical dimension (*CD*) of optical lithography—the minimum size that it can define—is a function of three parameters: $CD = k_1 (\lambda / NA)$.

CD is proportional to wavelength λ of the exposure light and process-related factor k_1 (discussed in more detail later), but decreases with increasing *NA*. Optical lithography equipment can print smaller dimensions by decreasing the wavelength, increasing the numerical aperture, reducing k_1, or any combination of these actions. All three measures will be necessary to achieving the ultimate resolution of optical lithography.

Wavelength. Three factors limit wavelength reduction. First, few light sources can deliver adequate power to expose wafers at cost-effective throughput rates. Second, the atmosphere attenuates light significantly at wavelengths below 193 nm, requiring operation in an environment without oxygen and water. Suitable optical materials for making lenses are also rare. Third, associated with wavelength reduction are changes in process integration schemes, and the development of photoresist and optical materials. All are extremely involved tasks. These three factors preclude arbitrarily decreasing the wavelength. The wavelength used in projection optical lithography has decreased as shown in Table 1; each reduction decreases the critical dimension by approximately 20 percent. Projecting from current status, 157-nm processes might be available in a few years. Lithography at 13 nm—also called extreme ultraviolet (EUV) lithography—is still far from maturity, despite the substantial effort and investment over the past decade.

Numerical aperture. Because *NA* is the sine of an angle, its physical upper limit is one. In this limit, the imaging system captures light rays propagating in all directions. Although increasing *NA* improves the imaging system's resolution, it adversely impacts depth of focus—the maximum amount of focus variation that the exposure process can tolerate and still print the circuit patterns within

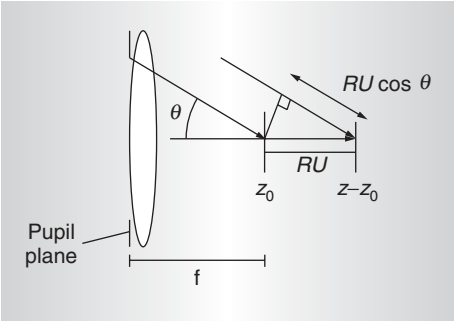

Figure 2. Rayleigh's depth of focus criterion.

specification. It is possible to estimate depth of focus using Rayleigh's criterion.

Consider the situation shown in Figure 2, where two rays—one from the pupil's center and the other from the aperture's edge—form an image. Suppose these rays interfere to form a sharp image at plane z_0; the image will gradually degrade as the observation plane moves farther from z_0. The relative phase change between the two rays at a plane a distance z away from z_0 is

$$z - z\cos\theta$$

where $\sin\theta = NA$. Rayleigh's criterion states that when the optical-path difference between the two rays grows to a quarter of the wavelength, it will blur the image. According to this criterion, depth of focus is distance $z = RU$ such that

$$\lambda/4 = RU(1-\cos\theta)$$
$$RU = \frac{\lambda}{8\sin^2(\theta/2)} = \frac{\lambda}{4NA^2}\left(1+\sqrt{1-NA^2}\right)$$

For low-NA imaging, Rayleigh unit RU is proportional to $\lambda/(2NA^2)$.

The dependence on $1/NA^2$ indicates that the depth of focus diminishes faster than the increase in NA. But the degradation is worse for high-NA imaging. As NA approaches one, RU approaches $(\lambda/4)$ rather than $(\lambda/2)$. This effectively reduces depth of focus by another factor of 2. For the 0.15-NA, 488-nm process typical in the 1970s, the depth of focus is over 20 μm. For a 157-nm system with $NA = 0.95$ (a system expected in the latter part of the 2000s), the focus tolerance is only 0.11 μm, a factor of 200 reduction.

Exposure systems. Despite technical difficulties, the wavelength and numerical aperture of exposure systems have been decreasing and increasing respectively, delineating transistors with ever-smaller dimensions. The price of these exposure systems reflects the escalating engineering challenges: a state-of-the-art system increased from $0.3 million in the early 1980s to $11 million in 2002, as Figure 3a shows. However, these systems have also

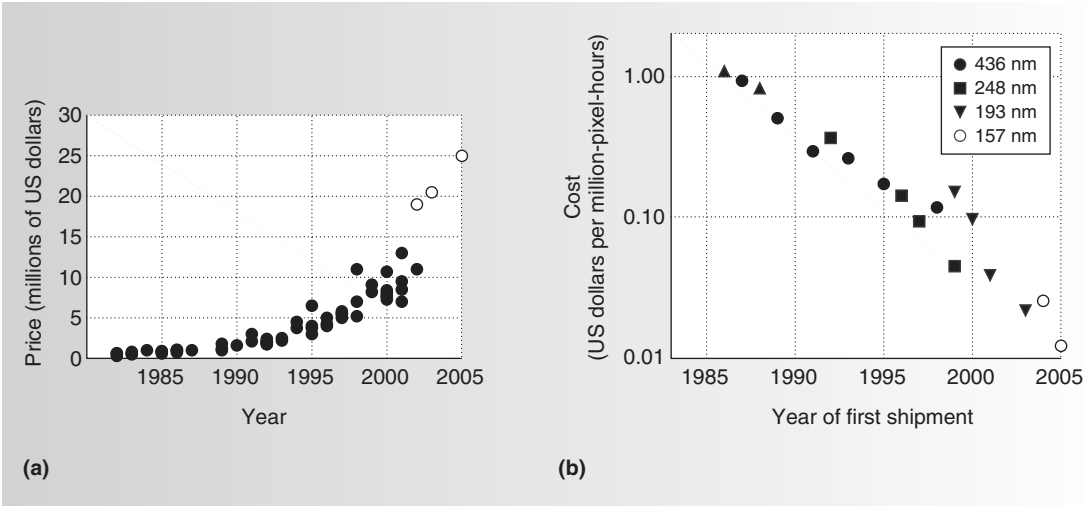

Figure 3. Although exposure system cost has skyrocketed (a), productivity has increased exponentially (b). Source: ASML.

exhibited an exponential improvement in productivity. This increase in productivity offsets the exposure system's growing base price, sustaining the semiconductor cycle. The cost of printing each pixel per unit time, shown in Figure 3b, will have decreased by a factor of 100 within 20 years, from $1 per million pixel per hour in 1985 to a projected $0.01 in 2005.

Figure 4 shows a plot of wavelength and *NA* trends for exposure systems. Wavelength decreased from 436 nm in the 1980s to 193 nm this year; *NA* increased from 0.15 in 1970 to 0.85 this year; and the k_1 factor (discussed in more detail later) fell below 0.5. *CD* (gate length) thus decreased exponentially, resulting in the familiar Moore's law.

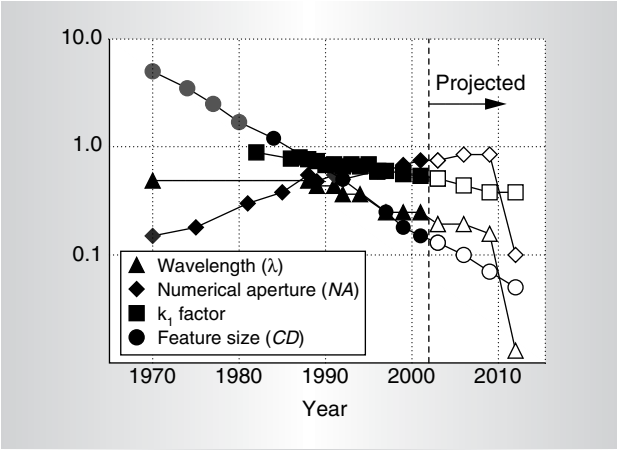

Figure 4. Evolution of optical lithography parameters. The numerical aperture has increased while the wavelength has decreased. The k_1 factor—a measure of lithography ease—has also decreased steadily. The sudden changes of wavelength and *NA* beyond 2010 assume the availability of EUV lithography. Copyright SPIE, 2001. Reprinted with permission.[4]

Low-k_1 photolithography

Before discussing the imaging challenges at decreasing k_1 values, let us examine the limits of optical lithography. It turns out that the limiting factors for pattern periodicity differ from those governing pattern dimension. Let us first focus on the period. Formation of an image requires two or more light rays, such that they interfere to form bright and dark fringes. The pattern pitch formed from two beams with an angle ϕ between them is

$$p = \lambda / [2\sin(\phi/2)]$$

The larger the ϕ, the smaller the period. For an optical lithography system, the largest angle between two image-forming rays is 2θ, representing the interference of rays from a pupil's opposite edges. This situation produces the minimum pitch

$$p_{\min} = 1/2 \times (\lambda/NA) = k_{p,\min} \times (\lambda/NA)$$

The theoretical limit on circuit pattern density depends only on exposure system parameters λ and *NA*, with $k_{p,\min} = 0.5$. For a 0.85 *NA*, 193-nm system, the tightest pitch is 113.5 nm; the shortest possible period imaged by a 0.95 *NA*, 157-nm system is 82.6 nm.

On the other hand, no theoretical lower limit exists for *CD*, that is, when $k_{1,\min} \rightarrow 0$. Once an exposure system projects an intensity modulation onto a photoresist-coated wafer, processing techniques such as overetching and ashing can arbitrarily produce small dimensions. For example, 248-nm lithography can define a 30-nm physical-gate-length transistor, as shown by the cross-sectional scanning electron micrograph (SEM) in Figure 5.

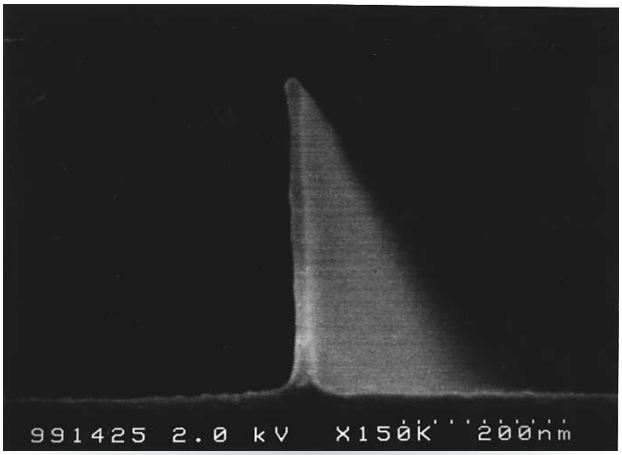

Figure 5. No theoretical limit exists for the minimum *CD*. This SEM image shows a 30-nm physical gate-length transistor defined using 248-nm lithography. Source: Nakao Shuji, Mitsubishi Corp.

Existence of a theoretical lower limit $k_{p,\min}$ and the lack of a lower limit for $k_{1,\min}$ means that it is possible (for a given exposure system) to continuously reduce *CD* without decreasing the period. Gate length shrinkage is not synonymous with transistor density increase. For some

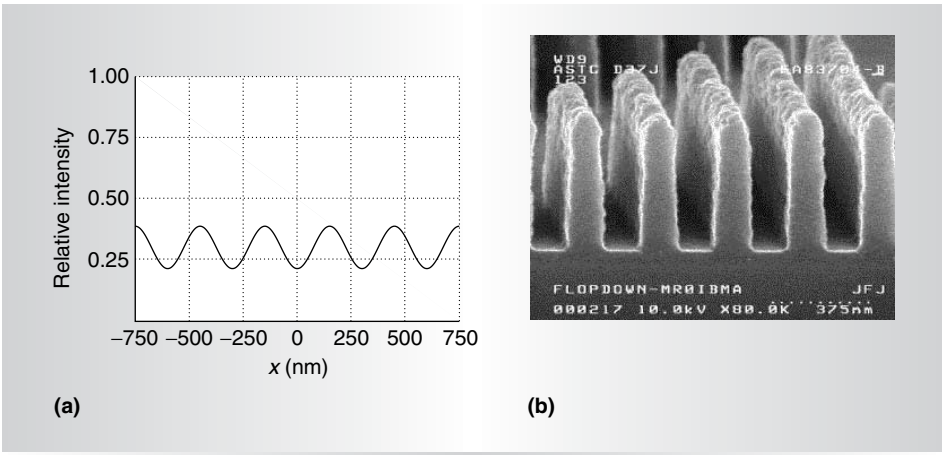

Figure 6. Photoresist chemistry can turn a sloppy image (a) into lines with vertical profiles (b).

low-k_1 imaging techniques such as phase-shifting mask and lithography-friendly design (discussed later), allowing a density decrease will permit delineation of shorter gate lengths. The desirability and ramifications of *CD* reduction without density increase are still open questions.

Returning to resolution limits, there are thus two categories of limits: theoretical and practical. The theoretical pitch limit is $0.5(\lambda/NA)$ while the minimum dimension has no lower limit. The practical limits are functions of process control and material robustness, determined by the need to control the sizes of all features to remain within specification over the entire chip. With lithography difficulty inversely proportional to the k_1 factor, I estimate the practical limits to be $k_{p,\min} \leq 0.7$ and $k_{1,\min} \leq 0.2$. These correspond to a minimum pitch of 116 nm and a *CD* of 33 nm for a 0.95 *NA* 157-nm exposure system.

Challenges and solutions

As k_1 falls below 0.75, image quality degrades noticeably. Consider delineation of 150-nm lines with a 300-nm period using a 0.6 *NA*, 248-nm exposure system. The k_1 factor is 0.363. Rather than well-defined bright and dark regions, the image is severely sloped because the pupil cuts off the high spatial-frequency components, as Figure 6a shows. Fortunately, photoresist chemistry can transform sloppy images into lines with vertical sidewalls, as Figure 6b shows.

Photoresist nonlinearity alone is insufficient for production-worthy low-k_1 lithography, because poor images result in lax *CD* control. Manufacturing at k_1 values less than 0.75 requires the use of resolution enhancement techniques (RETs) for image quality improvement.[4] These techniques include reconfiguring the light source, such as annular illumination; photomask methods, such as phase-shifting masks (PSMs) and optical proximity correction (OPC); and wafer coatings, such as antireflective layers and top-surface imaging. We focus on the first two techniques mentioned because they affect circuit design.

Phase-shifting mask. Use of phases on photomasks can improve image quality. Consider the imaging of a line, as illustrated in Figure 7a. If the bright regions bordering each side of the line differ in phase by 180°, the electromagnetic fields from these regions destructively interfere to yield a region of low field amplitude in between them. Because the image intensity is proportional to the square of the electric field, a sharp image results. Figure 7b, the image produced by a conventional mask, is less robust because of the lack of phase interaction.

Although alternating PSMs produce images with high fidelity, the image's dark regions are necessarily continuous because the boundaries of the 180° regions are continuous. Some applications (such as printing of the gate level) require the dark areas to be distinct. In these cases, a second exposure with a trim mask can remove the unwanted dark edges to provide a clean break between regions. For example, adding the images of a PSM (with a continuous dark region) and a non-phase-shifted trim mask, as shown in Figure 8, can define an isolated line. But the use of double exposure and phases on the mask require extra design rules. You can avoid double exposure by using three or more phases on the PSM. But the need for extra design rules is similar. Liebmann et al. offer a detailed discussion.[5]

Figure 9 shows a sample set of additional design rules. You must define critical feature size d_{crit} such that lines narrower than d_{crit} have borders of regions with a 180° phase difference

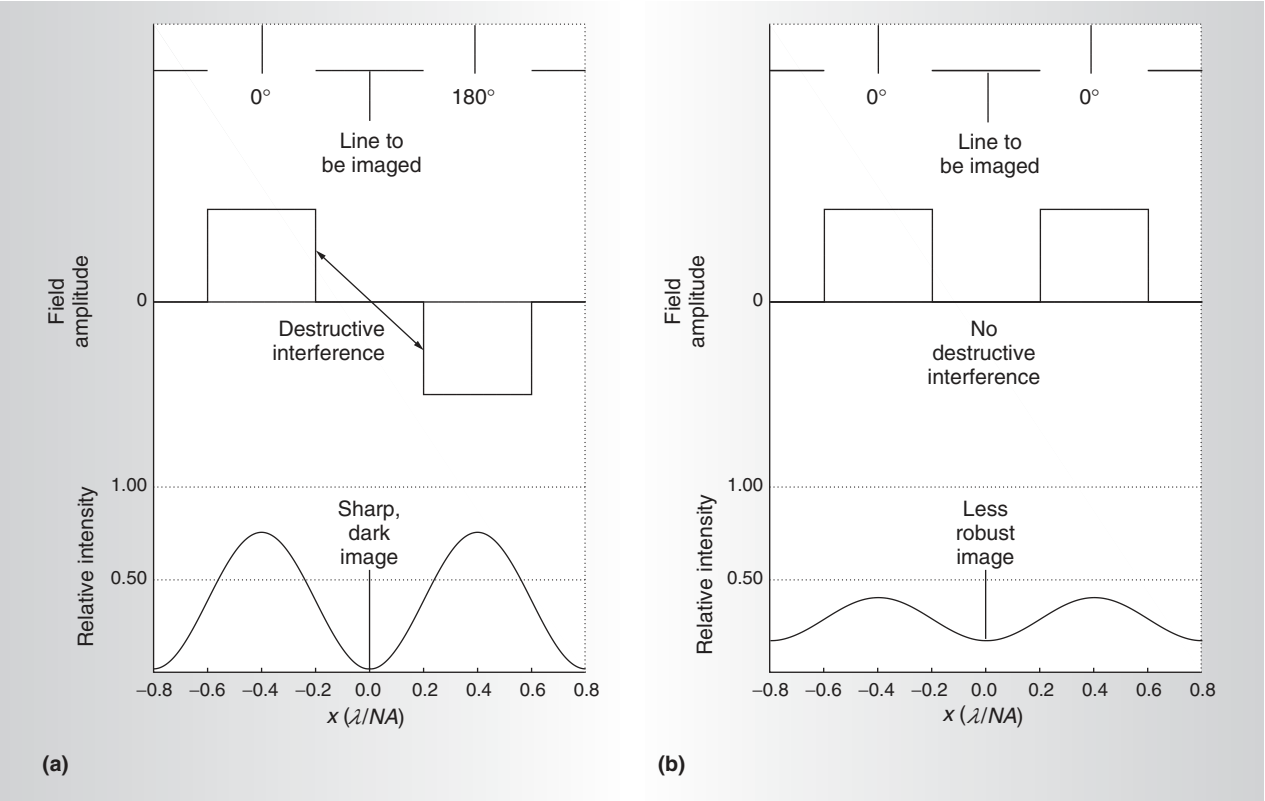

Figure 7. Destructive interference between two bright regions having a 180° phase difference (a) results in a sharp dark image. Using a conventional mask (b) produces light with too low an intensity. Copyright SPIE, 2001. Reprinted with permission.[4]

(rule 1). Pattern dimensions larger than d_{crit} do not require phase shifting (rule 2). Design rules must also specify the minimum width of the phase regions (rule 3), the minimum separation between phase regions (rule 4), and the minimum distance between phase and opaque areas (rule 5). There should also be a minimum separation between a critical line's end and another critical line (rule 6) to accommodate the trim exposure.

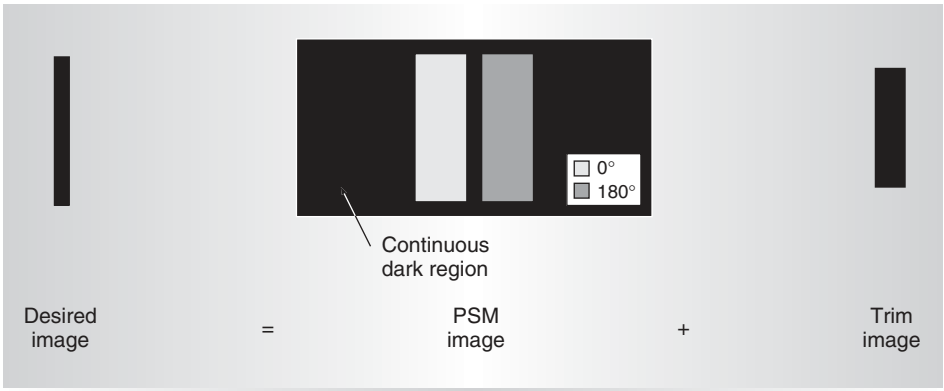

Figure 8. Second exposure trims unwanted phase transitions.

Because this technique requires phase shifting of circuit patterns smaller than d_{crit}, it cannot properly assign phases for configurations such as those shown in Figure 10. Phase conflict resolution is necessary. The conflict resolution approach depends on the level of involvement from designers. In the extreme where designers do not wish to know about phase-shifting lithography, design rules must be stringent enough to guarantee that all layouts are phase conflict free, a requirement that incurs unacceptable pattern density reduction. A practical, alternating PSM design process demands participation from physical designers to adopt practices that minimize the probability of phase conflicts and to understand alternatives to problematic configurations.[5]

The level of design hierarchy at which a

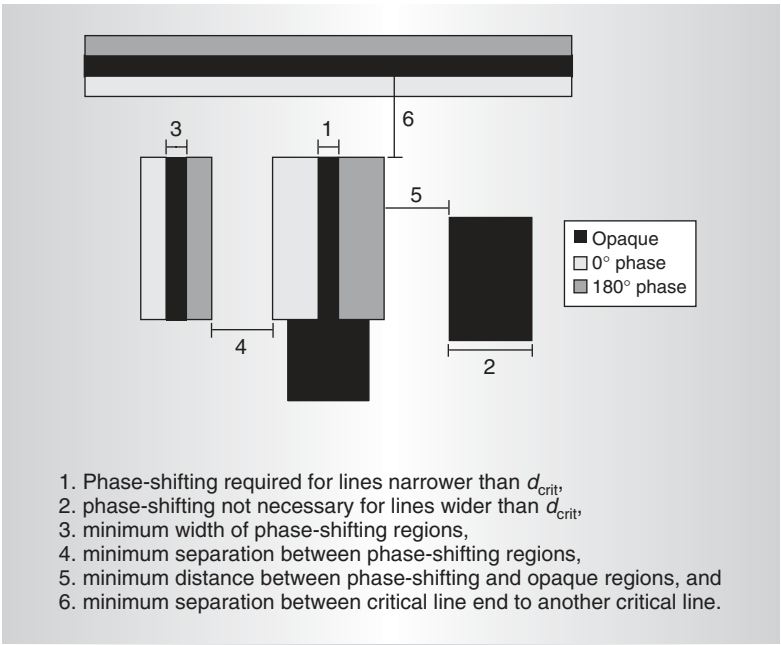

Figure 9. Use of alternating PSMs requires extra design rules. Copyright SPIE, 2001. Reprinted with permission.[4]

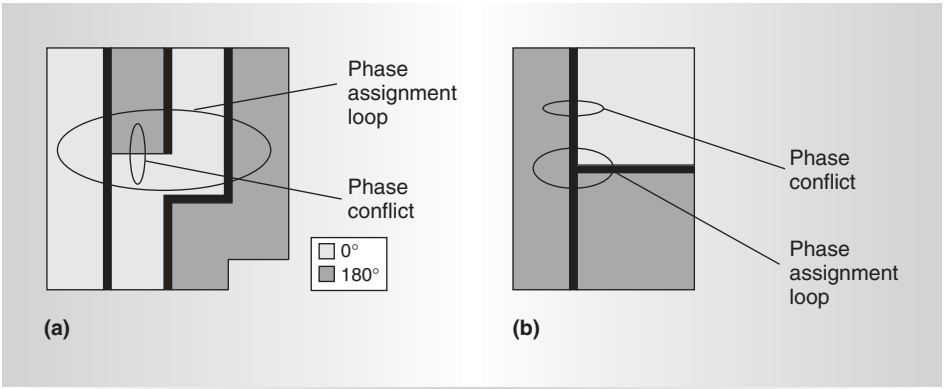

Figure 10. Configurations (a) and (b) result in phase assignment conflict because the number of regions in the phase assignment loop is odd.

design process introduces phase patterns also affects design quality.[5] If the process incorporates phase regions at a low level, it must include phase-shift-compliant rules regulating the layout and placement of basic circuit building blocks (such as standard cells) to contain the probability of phase conflict. Such rules are usually so restrictive that they dramatically increase circuit area. On the other hand, chip-level phase assignment is so computation intensive that algorithm execution time and data volume can be unacceptable. In addition, high-level phase conflict resolution is not straightforward because the solution might require nonlocal pattern reconfiguration. The optimal level for phase region insertion depends on the chip's composition (arrays, standard-cell and custom logic, IP blocks, and their combination) and its hierarchy. Implementation of alternating PSM involved tradeoffs between practicality, circuit density, and minimum transistor size.

Optical proximity correction. In low-k_1 imaging, beams corresponding to high-spatial-frequency components carry a sizable fraction of the light energy. Because the low-pass pupil does not capture these components, images are distorted from the original patterns.

Figure 11 shows three types of image distortion. Proximity effect refers to features with the same nominal *CD* that print differently because of environmental differences. Figure 11a shows the most typical scenario, which plots the printed dimension of a nominally $CD = 0.4(\lambda/NA)$ line as a function of pitch. As the period changes, the imaged line width varies by as much as 15 percent. This type of distortion results in increased across-chip line width variation (ACLV).

Another form of image distortion is line shortening. Illustrated by a rectangular pattern in Figure 11b, the rectangle's printed length is less than the drawn length when the width prints on target. This behavior is of critical concern because of its impact on overlay and circuit density. For instance, line shortening in dynamic random-access memory (DRAM) circuit patterns reduces the overlay budget between the capacitor and isolation levels.[6] For certain patterns, any potential increase in circuit density by decreasing the *CD* might be more than offset by the need to increase end-to-end spacing between the patterns.

Corner rounding is another type of image distortion. The pupil blocks high-frequency components of a sharp corner, resulting in a rounded image. Figure 11c shows an adverse effect of corner rounding. The gate of a metal-oxide-silicon transistor lies in close proximity

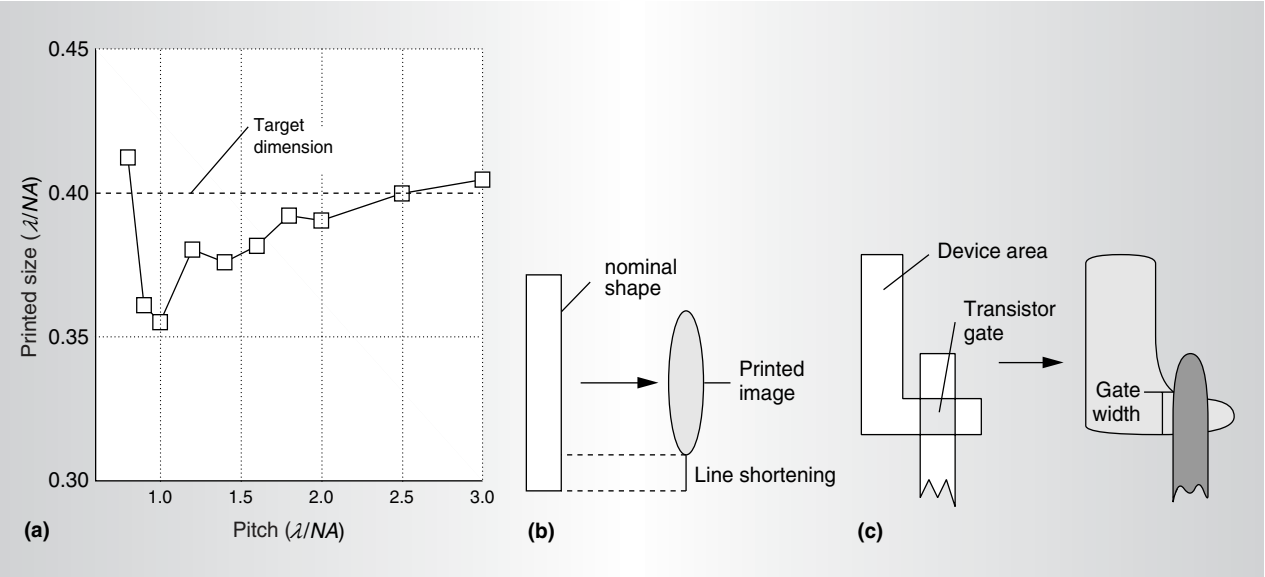

Figure 11. Various types of image distortion: proximity effect (a), line shortening (b), and corner rounding (c). Copyright SPIE, 2001. Reprinted with permission.[4]

to the elbow of an L-shaped device area. Rounding of the elbow results in a device whose effective width depends on the relative placement of the gate and active area.

OPC is a technique for mask pattern compensation that makes the printed features as close to the desired shapes as possible. Typical modifications include lengthening of a feature, displacing the edge of a pattern (biasing), introducing a nonprinting assist feature for image quality improvement, and creating serifs to reduce corner rounding.

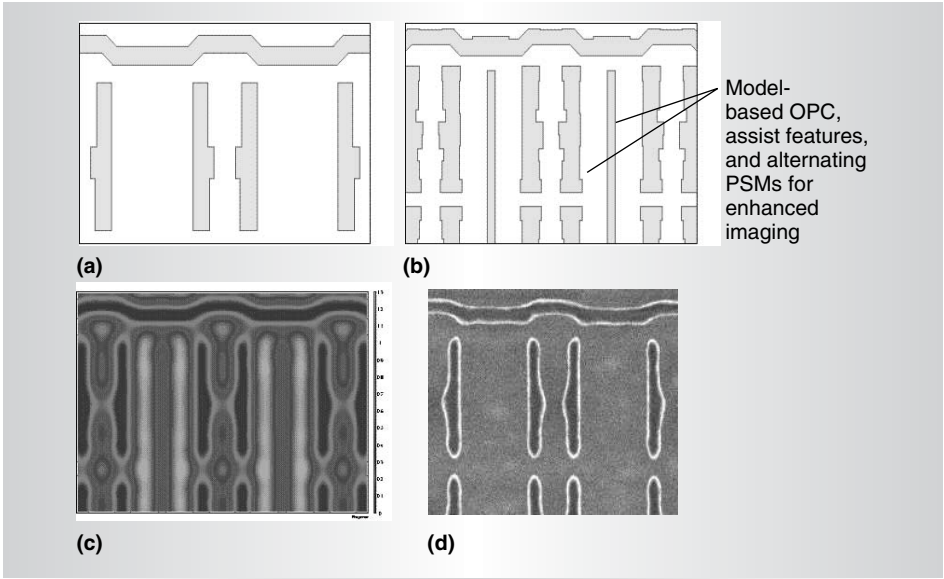

Figure 12. OPC at the gate level of a 100-nm SRAM design: original design (a), corrected layout with post-OPC mask pattern (b), simulated results of light intensity at the wafer plane (c), and exposure result showing the pattern in the photoresist (d). Source: Anthony Yen at TSMC/International Sematech.

Figure 12 shows an OPC example. The original design in Figure 12a is at the gate level of a 100-nm static random-access memory. The corrected layout, Figure 12b, includes crenelated patterns and assist features. Both simulation (Figure 12c) and the exposure result from top-down SEM measurement (Figure 12d) confirm the usefulness of OPC.

Corrections such as serifs, tiny crenelations, and assist features increase the time to produce photomasks and hence their cost. With a projected mask cost as high as $50,000 for the 50-nm technology (according to Walt Trybula of International Sematech) and a typical mask set consisting of 30 to 40 reticles, multimillion-dollar mask sets might become a requirement. Mask cost reduction is imminent.

Applying OPC judiciously is one approach for mask cost reduction. Rather than the current practice of using the same degree of predistortion for all shapes, the correction applied would depend on the circuit function of the patterns.[7] Geometries corresponding to circuit elements on critical nets should have the most aggressive correction, less-critical patterns receive a lower degree of correction, and noncritical shapes only need rudimentary correction. This graduated correction reduces mask cost because the masks contain fewer crenelations and serifs.

Frugal OPC requires an understanding of circuit functionality, and it is best applied prior to tape out, along with layout-versus-schematic verification and design rule checking.[8] Because corrections are sensitive to process changes, regular communication between physical designers and lithographers is necessary. Such communication can prove difficult, especially for foundries and their customers. If OPC should remain a post-layout activity, layout data formats (such as GDSII) should be extended to include criticality information.

Lithography-friendly design

At low-k_1 values, image quality depends not only on a pattern's size and shape but also on its environment. For example, you can improve the imaging of dense configurations of contact holes by using a light source that emphasizes oblique rays. But this illumination configuration is disadvantageous for sparse configurations of contact holes. Using light with a shallow angle of incidence is the best for imaging sparse contacts. This configuration unfortunately has an adverse effect on dense contacts. No illumination scheme allows optimal imaging of both dense and sparse contacts.

To improve manufacturability, it is necessary to limit the configurations of circuit patterns. For example, restricting contact placement to a grid such that all contacts align in both spatial directions can optimize lithography. Such restriction might at first suggest circuit density reduction. But the improved process robustness means that a design can use smaller contacts and pack them closer together.

My colleague and I have designed an add-compare-select unit in Viterbi decoders using three flavors of standard cells: normal, lithography friendly (on-grid contacts), and lithography friendly with the contact period shrunk by 10 percent. Compared with the circuit designed using normal standard cells, the circuit realized with lithography-friendly cells is 11 percent larger. However, if we use the third flavor of standard cells—those with contact period shrunk by 10 percent[9]—the circuit becomes 2 percent smaller than a traditional circuit. These initial results show that the tradeoffs between area, speed, and manufacturability deserve further investigation.

Optical lithography is an enabling technology for transistor miniaturization. With the wavelength and numerical aperture of exposure systems approaching their limits, the semiconductor industry needs continuous reduction of the k_1 factor. Challenges include image quality improvement, proximity effect correction, and cost control. An indispensable ingredient for future success is improvement in the design-manufacture interface. PSM and OPC implementation require lithographers to understand basic design principles and designers to have a basic appreciation of photolithography. Pattern configuration restriction and the possible advantages of *CD* reduction without density increase also require involved dialogues between the two traditionally separate communities.

Low-k_1 lithography does not allow the semiconductor industry to eat its cake and have it too. All stakeholders must compromise to sustain the semiconductor economic cycle, at least until the advent of self-assembling nanoelectronics and quantum computers. MICRO

Acknowledgments

I am grateful to Anthony Yen and Walt Trybula at International Sematech for insightful discussions and suggestions. I also acknowledge the data provided by ASML (Figure 3), Nakao Shuji at Mitsubishi Corp. (Figure 5), and Anthony Yen at TSMC/International Sematech (Figure 12), as well as the permission granted by SPIE for reproduction of Figures 1, 4, 7, 9, and 11. I also thank Keith Diefendorff at MIPS for the encouragement to write this article and the reviewers for thoughtful criticisms of the manuscript.

References

1. G.E. Moore, "Lithography and the Future of Moore's Law," *Proc. SPIE*, vol. 2440, 1995, pp. 2-17.
2. M. Born and E. Wolf, *Principles of Optics*, sixth ed., Pergamon Press, 1980.
3. J.W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, 1968.
4. A.K. Wong, *Resolution Enhancement Techniques in Optical Lithography*, SPIE Press, 2001.
5. L. Liebmann et al., "Enabling Alternating Phase Shifted Mask Designs for a Full Logic Gate Level," *J. Microlithography, Microfabrication, and Microsystems*, vol. 1, Apr. 2002, pp. 31-42.
6. A. Wong et al., "Level-Specific Lithography Optimization for 1Gb DRAM," *IEEE Trans. Semiconductor Manufacturing*, vol. 13, no. 1, Feb. 2000, pp. 76-87.
7. M. Rieger et al., "Enriching Design Intent for Optimal OPC and RET," *Proc. SPIE*, vol. 4754, 2002, pp. 132-137.
8. F. Schellenberg, "Optimal Insertion Points for OPC and PSM in Design Flows," *Solid State Technology*, Sept. 2001, pp. 63, 69, 70, 72, 74.
9. J. Wang and A. Wong, "Effects of Grid-Placed Contacts on Circuit Performance," *Proc. SPIE*, vol. 5043, 2003.

Alfred K. Wong is an associate professor in the Department of Electrical and Electronic Engineering at The University of Hong Kong. His research interests include IC design and fabrication, including resolution enhancement techniques in microlithography; ICs for communications and signal processing; and the synergies between circuit design and fabrication. Wong has a BS, an MS, and PhD in electrical engineering from the University of California, Berkeley. He received the 1990 IEEE Fortescue Fellowship, and serves as the chair of the SPIE Conference on Cost and Performance in IC Creation and as associate editor of the *Journal of Microlithography, Microfabrication, and Microsystems*.

Direct questions or comments about this article to Alfred K. Wong, Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong; awong@eee.hku.hk.