Optimum mask and source patterns to print a given shape

Alan E. Rosenbluth
IBM T.J. Watson Research Center
Yorktown Heights, New York 10598-0218
E-mail: aerosen@us.ibm.com

Scott Bukofsky
Carlos Fonseca
IBM Semiconductor Research and
Development Center
Hopewell Junction, New York 12533-3507

Michael Hibbs
IBM Microelectronics Division
Essex Junction, Vermont 05452-4201

Kafai Lai
Antoinette Molless
IBM Semiconductor Research and
Development Center
Hopewell Junction, New York 12533-3507

Rama N. Singh
IBM T.J. Watson Research Center
Yorktown Heights, New York 10598-0218

Alfred K. K. Wong
Department of Electrical and Electronic
Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong

Abstract. New degrees of freedom can be optimized in mask shapes when the source is also adjustable, because required image symmetries can be provided by the source rather than the collected wave front. The optimized mask will often consist of novel sets of shapes that are quite different in layout from the target integrated circuit patterns. This implies that the optimization algorithm should have good global convergence properties, since the target patterns may not be a suitable starting solution. We have developed an algorithm that can optimize mask and source without using a starting design. Examples are shown where the process window obtained is between two and six times larger than that achieved with standard reticle enhancement techniques (RET). The optimized masks require phase shift, but no trim mask is used. Thus far we can only optimize two-dimensional patterns over small fields (periodicities of ~1 μm or less), though patterns in two separate fields can be jointly optimized for maximum common window under a single source. We also discuss mask optimization with fixed source, source optimization with fixed mask, and the retargeting of designs in different mask regions to provide a common exposure level. [DOI: 10.1117/1.1448500]

Subject terms: off-axis illumination; source optimization; RET; OPC; global optimization.

Paper JM3 001020 received Aug. 20, 2001; revised manuscript received Dec. 12, 2001; accepted for publication Dec. 13, 2001.

1 Introduction

An important synergy can be exploited in jointly optimizing mask and source to print a given shape. In many cases the resulting mask and source patterns fall well outside the realm of known design forms. For this reason it is desirable that the optimization algorithm provides good global performance; in particular, the algorithm should not be constrained to use a known starting design. Our work suggests that standard approaches to optical proximity correction (OPC) may have difficulty converging on the mask solution that is globally optimal.

Previous work on optimization of the source alone has described general algorithms[1] and specific implementations[2–4] for customizing illumination to print particular shapes. Enhancement techniques to customize masks (e.g., RET methods like assist features, serifs, phase tiling, etc.) are usually applied as adjustments or modifications to the nominal circuit patterns. In formal terms, one can say that the nominal patterns (or some simple extension of them) effectively serve as the starting solution when masks are optimized.

In this respect RET technologies are linked to classical lithography, wherein axially illuminated mask shapes that reproduce the target patterns are used to project a wave front with all attendant symmetries into the lens. The wave front section collected by the lens [whose finite numerical aperture (NA) acts as a cutoff filter] is likewise symmetrical under axial illumination, and as a result the input symmetry is transferred to the image. Wave front symmetry constraints include Hermitian radial symmetry (if the reticle phase is restricted to 0° or 180° to avoid distortions through focus), as well as any bilateral symmetries that the target pattern may have.

These constraints substantially reduce the number of truly independent orders that can be collected under axial illumination. Once a particular positive order is determined, the corresponding negative order is also fixed (to within an unimportant translational phase). From an optimization viewpoint, the quasisymmetry of typical wave fronts implies that the number of degrees of freedom in the lithographic image will be little larger than that corresponding to one quadrant of the NA, or half the NA if the mask shapes are highly nonsymmetric (but still restricted to 0° or 180° phase). Figure 1 illustrates this idea in schematic form.

This paper is a revision of a paper presented at the SPIE conference on Optical Microlithography XIV, Feb./March 2001, Santa Clara, CA. The paper presented there appears (unrefereed) in SPIE Proceedings Vol. 4346. The present version includes a new test case, and additional discussion.

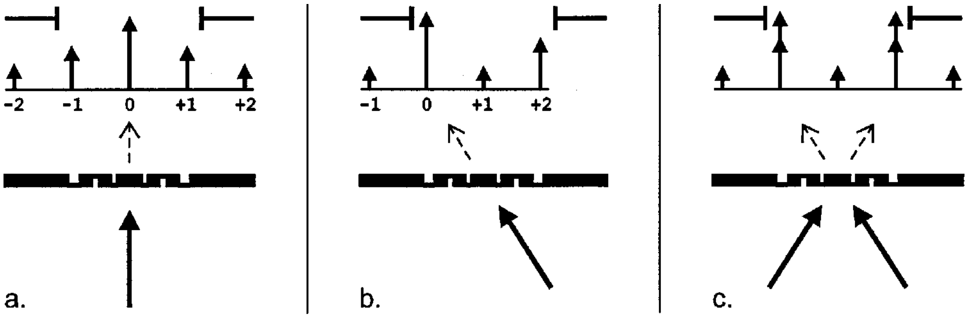

Fig. 1 Degrees of freedom in collected wave front using different illumination directions. Reticle phases other than 0° or 180° are ruled out to prevent distortions through focus. (a) Only two independent orders are collected under axial illumination, since $+1$ and -1 orders must be complex conjugates when reticle transmittance is real valued. (b) Three independent orders can be collected from (sufficiently oblique) illumination directions, aiding optimization. (c) Stability through focus is restored by illuminating reticle from mirrored directions.

However, when we illuminate the mask obliquely it is not necessary to impose a symmetry constraint on the decentered section of the wave front that is collected. In practice the illumination is limited to, e.g., $\sigma \lesssim 0.85$, so any feasible source direction will generally project both positive and negative diffraction orders into the lens. One typically finds that the number of truly independent orders that can be addressed from the dominant oblique directions in an optimized source will be of order $2\times$ larger than can be addressed with axial illumination. In many cases the availability of these extra degrees of freedom significantly enhances the quality of the optimized solution, and we can restore the required symmetries and focal insensitivity to the printed pattern by using a suitably symmetric source. The optimized diffraction pattern will therefore often be dominated by the way in which diffraction orders combine coherently from illumination directions that are strongly nonaxial, thereby forming the dominant image components of the incoherent sum.

The collected set of oblique orders usually has a more specialized structure after optimization than would be present with, e.g., the typical diffraction falloff from coarse mask rectangles. (For example, the latter non-optimized features will sometimes show a decreased depth-of-focus when the illumination is nonaxial, due to focus-runout of the strong zero-order. Such effects are usually much weaker in the optimized diffraction pattern.) This means that if the optimized mask were to be illuminated axially rather than obliquely, a completely different interference pattern would often be produced on the wafer (since the centered collection of orders would combine some new subset of the optimized oblique orders and corresponding negative orders in a qualitatively different and often undesirable way). In many cases the image produced under axial illumination would bear little resemblance to the optimized wafer image (while the optimized image will resemble the target pattern by design). It also follows that the optimized reticle pattern, which can be thought of as comprising a very large number of axially centered orders, can likewise differ substantially from the optimized image (or the target pattern).

This means that enhancement techniques which use the target patterns as a starting solution may not provide fully optimized reticles when the source shape can be freely adjusted. Note that most algorithms for nonlinear optimization are essentially local minimizers, and so are strongly dependent on the quality of the starting solution. Of course, lithographers face no explicit requirement to begin the design process using any particular trial layout. Indeed, independent of their direct utility, global algorithms are of interest as conceptual tools for bringing forward new design forms.

Casual experimentation with a local optimization routine suggests that changing the magnitude of individual orders by ~ 0.3 can move a trial solution into the vicinity of a new local minimum (in a test case where the average order intensity was set to about 1). This sensitivity reflects the oscillatory nature of the plane wave components that define the image. If we suppose that the orders typically span a range from about -3 to $+3$, and that the minimum field size needed to adequately bound the tails of the lens resolution (e.g., $\sim 2\lambda/\mathrm{NA}$) can be characterized by seven collected diffraction orders from a staggered array (allowing nonaxial illumination, but counting only truly independent orders), then if we wish to find globally optimal values for these amplitudes via the simple expedient of trying a large number of starting solutions, we would be required to run the optimizer from roughly (21^6)/2$=4\times 10^{7}$ different starting points in order to sample every potential local minimum. Inclusion of the source variables entails a further combinatorial explosion.

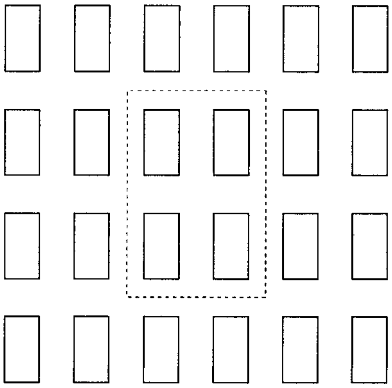

Fig. 2 Capacitor pattern. Horizontal period is 260 nm, vertical period 390 nm. Rectangles (130 nm×247 nm) are bright. Dashed boundary shows plot area for images in later figures.

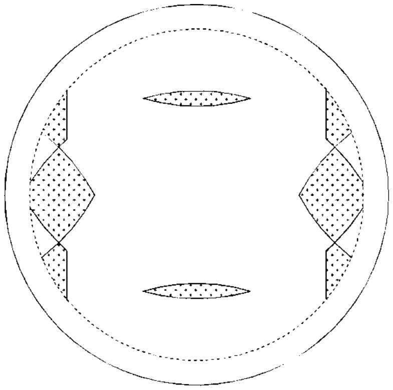

Fig. 3 Optimized source for Figure 2 capacitor pattern. λ =248 nm, NA=0.68 (solid circle). Integrated process window through focus is optimized. Hatched areas are bright. Dashed circle is 0.85 σ limit. Figure 4 shows mask.

This estimate is crude, but it demonstrates that even the most robust local convergence is insufficient for thorough RET optimization. To address this disadvantage we have devised global (i.e., non-local) algorithms that can optimize mask and source to print a given shape without using a starting design. The wave front from any individual off-axis direction is allowed to have arbitrary decentration (above and beyond that produced by the tilted illumination), and arbitrary lateral asymmetry. Focal tilt and bilateral asymmetries in the final image are removed by using symmetric illumination distributions. Several simplifying approximations are adopted, and full globality of the joint mask/source solution is not guaranteed. However, many of these approximations can be avoided in the subproblems of calculating the optimal mask for a given source, the optimal source for a given mask, and the most efficient mask to produce a given set of collected orders [yielding global solutions to these subproblems under the simplified formulations given later, as well as in a more general formulation where the merit function closely approximates integrated exposure/defocus (ED) window]. An optimized wave front generally requires 180° phase shift in the mask, which can be provided by either attenuating chrome, chromeless shifters, or phase-reversed openings in opaque chrome. No trim mask is used.

The present paper will describe in some detail a global optimization algorithm that uses exposure latitude as the merit function. However, we have also developed a preliminary version of an algorithm that optimizes against full process window through focus (using integrated area of the ED window as the merit function),[5] with the main limitation of the algorithm as thus far developed being a significant increase in processing time over the in-focus case. (In general, the scaling of processing time with problem size tends to be unfavorable when global solutions are sought.) We will show results from the more general algorithm, but will defer details of the method to a later publication.

Let us consider as an example the dynamic random access memory (DRAM) capacitor level shown in Figure 2. One critical dimension in this pattern is the width of the printed rectangles (bright for positive resist), which in this example we take to be 130 nm. Though difficult, it is also desired that the rectangles print with an aspect ratio of at least 1.9:1. At low k factor this elongated aspect ratio poses considerable difficulty for conventional RET methods. The DRAM cell uses a $2F\times 3F$ layout,[6] and the pitch ratio is only 1.5:1. Contrast in the dark gaps that separate the rectangle tips is poor, and the rectangles tend to print with considerable shortening. When shortening is compensated by narrowing the gaps, contrast degrades further. For example, at λ=248 nm and NA=0.68, even an ideal thresholded aerial image model predicts that we will only be able to print the array using an attenuated phase-shift mask ($T=6.5\%$) and annular illumination if we allow fairly relaxed critical dimension (CD) tolerances, and accept poor contrast in the dark separations between the tips of the rectangles. If we impose a requirement that the intensity at the center of the focused rectangle be at least three times larger than that midway between the tips (i.e., if we do not allow the feature to be biased beyond the point where max-to-min contrast in a vertical slice across the tips drops below 3:1), then the ED window achieves a depth of focus (DOF) of ±0.56 μm when tolerances of ±30 and ±15 nm are ap-

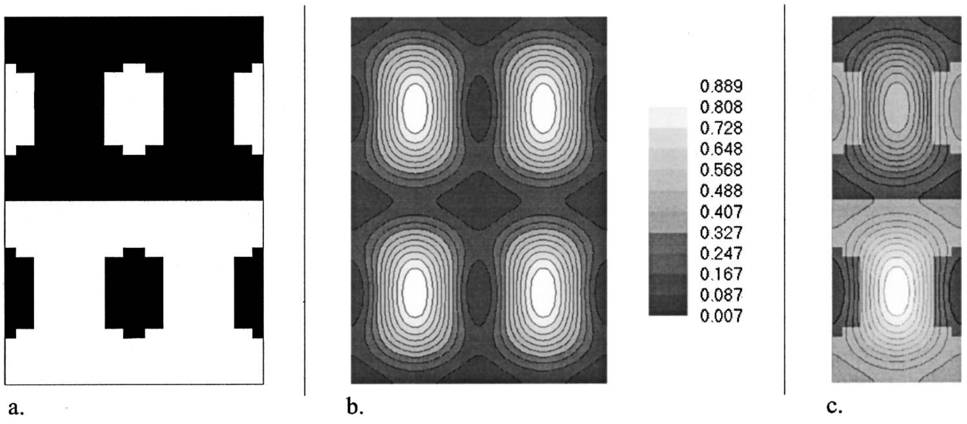

Fig. 4 (a) Optimized mask patterns (chromeless) for Figure 2 capacitor pattern. Black represents 0° phase shift, white 180°. Area shown corresponds to dashed region of Figure 2. (b) Aerial image [screen capture from Prolith (see Ref. 7) simulation]. (c) Superposition of mask and image. The "battery-shaped" mask features create dark horizontal separations in the image, and are positioned in between the bright image rectangles. Pattern layouts on mask and wafer are quite different.

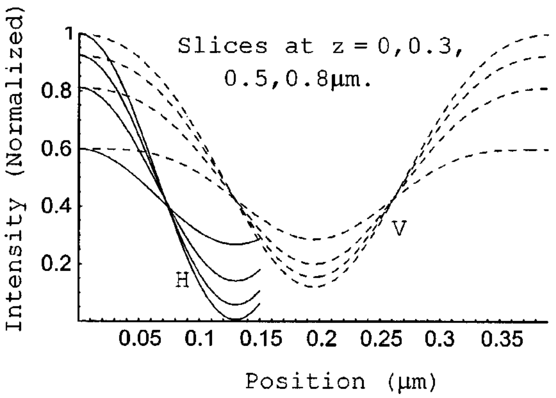

Fig. 5 Successively defocused image slices from Figures 3 and 4(a) solution, taken through centerlines of bright rectangles. Dashed are vertical slices, solid horizontal. Images are normalized against peak intensity of 89%.

plied to the length and width, respectively. The process window is 7%-μm (using integrated area under the two-sided ED curve as the process window metric).[5] If we remove all constraints on contrast, biasing can increase theoretical process window to 16%-μm, but contrast drops to 2.3:1 in the focused image. Experimentally, such low contrasts prove unusable, and printed resist images show zero common process window for length and width using conventional enhancement methods, unless separate exposures are used to print alternate rows of the array.[6]

Figures 3 and 4(a) show the result of optimizing mask and source to print the Figure 2 pattern (at $\lambda = 248$ nm, NA$=0.68$), using the algorithm that maximizes integrated process window through focus. Image slices are shown in Figure 5. A chromeless mask technology is used, though the same underlying solution can be realized in essentially any mask technology that provides 180° phase shift. Note that the bright rectangular features in the image actually fall in between the vaguely brick-like openings in the reticle, i.e., the direct resemblance of these reticle shapes to the image patterns is coincidental. Indeed, the reticle shapes in Figure 4 that are optimized for off-axis illumination have a distinctly different "topology" from the image shapes, i.e., their basic layout has a different internal connectedness. It would have been quite difficult for a conventional optimizer to have devised a path of smooth and continuous edge adjustments that reached the Figures 3 and 4(a) solution from starting shapes that matched Figure 2; moreover, even if such a path could be defined, a local algorithm would not follow it unless process window increased monotonically at every point. (Local algorithms that use pixel variables rather than edge variables might be more promising in this regard.)

Figure 6 plots the ED window obtained with the Figures 3 and 4(a) solution, using the same ±30 and ±15 nm tolerances on length and width considered earlier. Integrated process window is 45%-μm under a thresholded aerial image model (assuming no aberrations). This is between 3× and 6× better than the calculated performance of standard enhancement methods (see earlier). Max-to-min contrast across the rectangle tips is 8.2:1, also much improved over the conventional result. The solution in Figures 3 and 4(a) was obtained by direct optimization against process window; however, a similar solution with quite a good process window (37.6%-μm) is obtained by optimizing against exposure latitude in focus (algorithm P described in Sec. 3, with step 2 omitted). One caveat should be made regarding these process window comparisons: Our optimization algorithm does not use the so-called "obliquity factors" when calculating high-NA aerial images. (It does, however, implement the nonparaxial expression for defocus.) On the other hand, when optimizing patterns using conventional RET methods, we frequently employ software whose imaging core is a commercial program that uses both obliquity factors and nonparaxial defocus when set for high-NA operation; thus, the process windows we quote for conventionally optimized patterns usually include the former factor as well as the latter. The distinction is minor on the scale of the large improvement (generally >2×) that we find with our global algorithm.

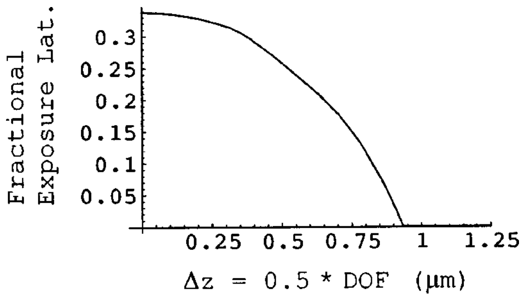

Fig. 6 Process window obtained with the solution of Figures 3 and 4(a). An aberration-free lens is assumed. CD tolerances are ±15 nm on width, ±30 nm on length. Curve is calculating from thresholded aerial images. Horizontal axis is single-side defocus, equal to half DOF. Integrated window (two-sided) is 45% μm.

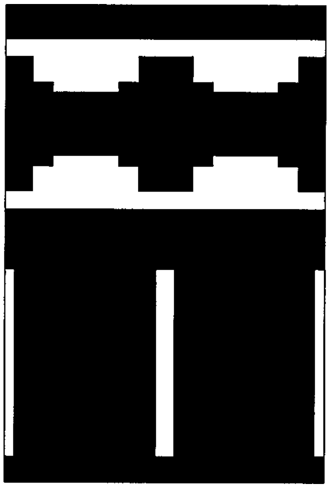

Fig. 7 Solution for Figure 2 pattern in attenuating phase-shift chrome. The area shown corresponds to Figures 4(a) and 4(b) and to dashed region of Figure 2. Mask openings are shown white. Chrome transmission is 6.5%, phase-shifted 180° (black shaded).

The optimized solutions can also be realized in attenuating phase-shift masks. The attenuating phase-shift solution in Figure 7 achieves the same large process window as the Figure 4(a) chromeless solution; however, overall intensity is quite low because the optimizer has realized the

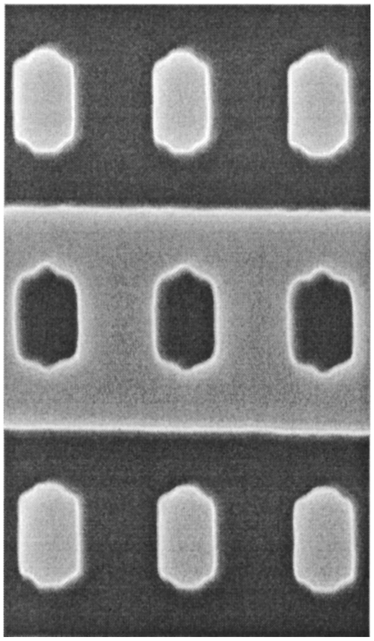

Fig. 8 SEM image of chromeless mask that implements Figure 4(a) solution.

solution by printing through the chrome as if it were a ‘‘hard’’ phase shifter.

Our global optimization approach provides novel design forms with high theoretical performance. Of course, in practice lithography cannot really be reduced to a purely formal optimization. After describing our method in more detail we will comment briefly on some issues of practical implementation. We will also discuss the prospect for extending our methods to optimize multiple patterns simultaneously. Global methods show promise for increasing the common process window of a suite of patterns. Indeed, in principle the common process window for a globally optimized set of patterns cannot be lower than that provided by conventional optimization methods. However, as with conventional methods, the common process window cannot be larger than is achieved for a single pattern that is optimized individually. Pattern diversity is necessarily limited within the field sizes that we can optimize at present (~1 μm), and source solutions for such small fields tend to be fairly specialized. Source directions at large-σ along the 45° azimuths tend to maximize the number of collected degrees of freedom, providing an advantage in optimizing a diverse set of patterns.

2 Experimental Test

Though the treatment in this paper is primarily theoretical, we felt it important to include an experimental demonstration of the theory. Figure 8 shows our implementation of the Figure 4(a) chromeless mask. To obtain results within a short deadline, we implemented the source of Figure 3 in the form of a simple illumination stop (located in a plane conjugate with the entrance pupil), and adopted the simplified hole pattern shown in Figure 9(a) for ease of fabrication. Figure 9(b) shows a measurement of the illumination pattern as realized in the exposure tool. The source apertures are sparsely filled because the input $\sigma = 0.85$ disk is realized by discrete multiple foldings within a homogenizing rod of rectangular cross section. The exposure tool uses a scanning slot field, so the input source appears striped in the pupil gram. In principle this kind of coarse discretization need not be present if source customization and uniformity are both provided by diffractive elements;[8] moreover, such discretization need not have a significant impact on the image, as may be seen in Figure 10. However, considerable source distortion was incurred in the present experiment [compare Figure 3 with Figure 9(b)].

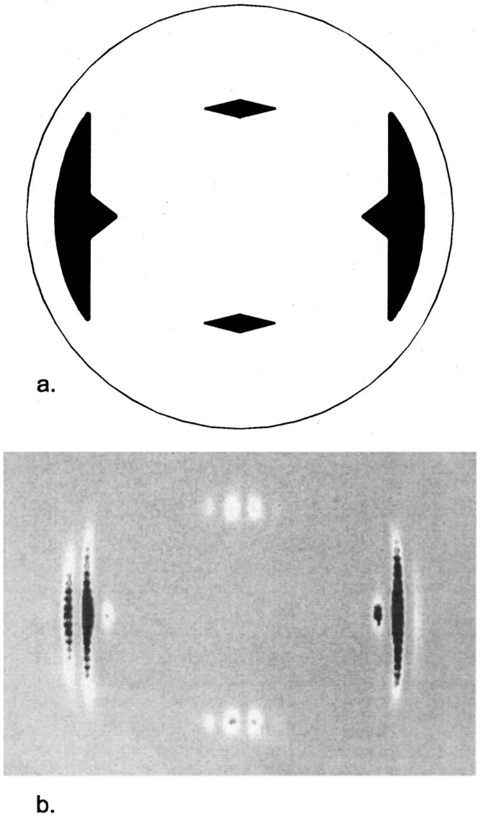

Fig. 9 Approximate realization of the Figure 3 source. (a) Simplified aperture pattern, designed to ease fabrication of stencil illumination stop in model shop. (b) Pupil gram (highly defocused image through mask pinhole) showing the illumination pattern actually realized in the exposure tool. Discretization from the light tunnel homogenizer is apparent.

Nonetheless, we achieved reasonable wafer images with this compromise source, as may be seen in Figure 11(a). Figure 12 shows focus/exposure data from the experiment [top-surface scanning electron micrographs (SEMs)]. Measured exposure latitude is about 14%, DOF approximately 0.7 μm, and process window roughly 7%-μm. This is quite a respectable result (though well below the ideal performance of the Figure 6 simulation), considering that in practice the pattern proves impossible to print within tolerance using conventional enhancement methods.[6] The investigations reported in Ref. 6 show that capacitor aspect ratio for 130 nm trenches is limited in practice to about 1.4:1 when annular illumination and phase-shift chrome are employed (versus 1.9:1 in the target pattern), even if the pitch is relaxed slightly to permit increased mask bias. Figure 11(b) shows the approximate limit of what can be achieved experimentally with the conventional approach [same NA and

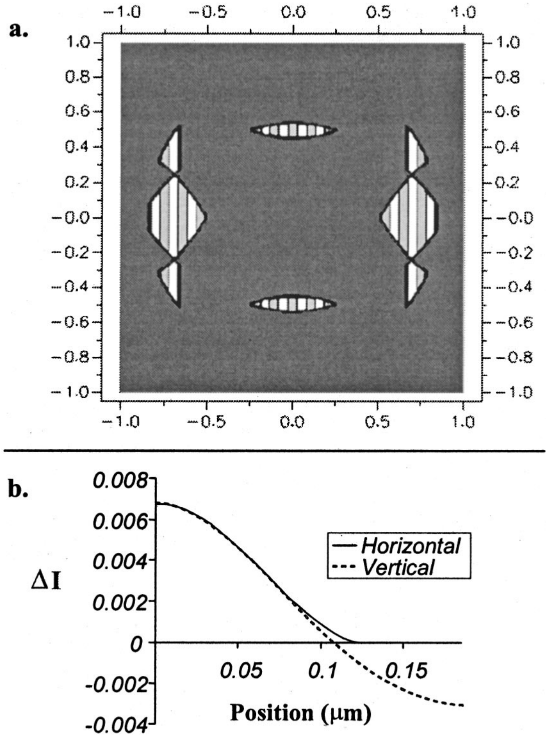

Fig. 10 Idealized model of source discretization by homogenizer. (a) Source pattern. The input $\sigma=0.85$ disk is sparsely filled, simulating the effect of homogenizing optics in a slot-field exposure tool. Plot shows source pattern after truncation by ideal Figure 3 aperture. (b) Difference between image with discretized Figure 10(a) source, and ideal image (continuous Figure 3 source).

λ as Figure 11(a), but different exposure tool and expanded pitch]. Because of the narrow vertical separation between adjacent capacitors, it is impossible to introduce a bias sufficient to meet tolerance unless every other row in the array is removed from the mask to free up more real estate; the array must then be printed in two separate microstepped exposures (see Figure 6 in Ref. 6).

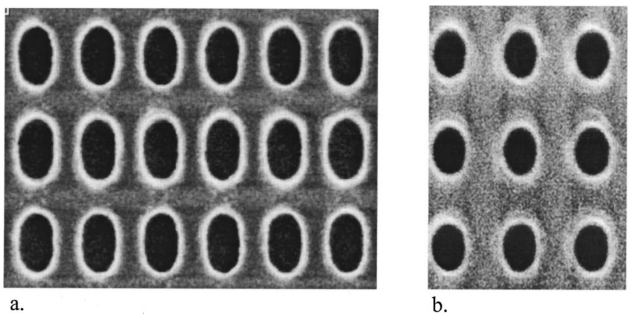

Fig. 11 Images of Figure 2 pattern in 5300 Å of UV82 resist, exposed at $\lambda=248$ nm, NA=0.68. (a) Exposures using the Figure 8 mask and Figure 9(b) source. Horizontal pitch is 260 nm, vertical pitch 390 nm, per Figure 2. (b) Attempt to print elongated capacitors of 130 nm width using conventional enhancement methods (annular illumination, phase-shift chrome, mask bias), and expanded pitch (relaxed to 300 nm horiz., 405 nm vert.). Adequate aspect ratio cannot be achieved in a single exposure. [Figure 11(b) image was scaled to same magnification as Figure 11(a) using graphics software.]

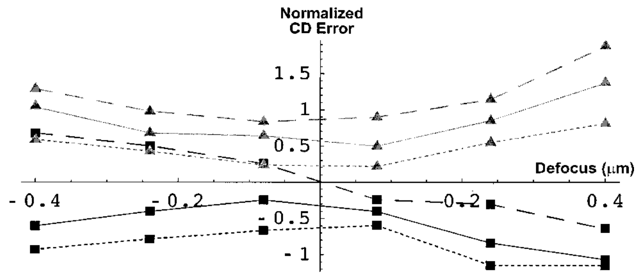

Fig. 12 Focus-exposure measurements using the Figure 9(b) source and Figure 8 mask. Each point represents the maximum CD error found in an adjacent pair of measurements. Errors are normalized, so that 1.0 represents the tolerance limit (±15 nm horiz., ±30 nm vert.). Gray triangles are width errors, black rectangles are length errors. Solid lines are nominal dose; dashed and dotted lines show the effect of increasing or decreasing dose by 4%.

3 Algorithm to Optimize Exposure Latitude

We now describe an algorithm for global optimization of mask and source against exposure latitude in focus. First, we note that highly efficient algorithms have been developed for local optimization;[9] these are available, for example, in packages like MATLAB,[10] Mathematica,[11] and IMSL.[12] Such algorithms can converge to local maxima in the merit function within polynomial time, even when the merit function is nonlinear. If one can model the system in the ''forward direction,'' and if one can devise a merit function to quantify the suitability of a given solution, then in most cases a nonlinear optimizer will efficiently refine a given starting design so that it converges to the nearest local maximum of the merit function.

In the case of global optimization, it has been proven that for a fully general merit function, no global algorithm can be guaranteed to perform better than simple exhaustive grid search of the parameter space (Nemirovsky and Yudin, as cited in Ref. 13). However, by exploiting the particular structure of the lithographic problem we can find solutions on a far more rapid basis. Knowledge of this special structure provides a strong advantage. For example, our tests of two general-purpose global optimization programs found them unable to solve even limited subproblems (e.g., source held fixed) of joint source/mask optimization problems that our specialized algorithms can handle.

The difficulty in lithographic problems is that the merit functions are usually not convex; indeed, the plane-wave orders that comprise the image are intrinsically oscillatory, giving rise to a great many local maxima. To achieve efficient global performance we adopt the following two-part strategy:

1. Seek the global solution to a simplified version of the problem; and
2. Use a local optimizer to refine the step 1 solution against a more complete model.

The robustness of widely available local optimization routines allows us to divert many detailed optimality criteria to step 2. Step 1 is solved under a scalar aerial image model.

The imaging solution determined in steps 1 and 2 is defined in the pupil plane (as a set of illumination and

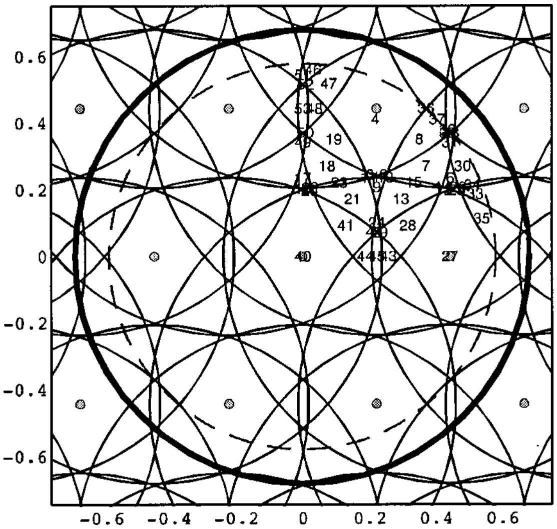

Fig. 13 Pupil diagram for array with staggered pitch. *x* pitch is 1120 nm, *y* pitch is 560 nm, and one basis vector is diagonal. Lens pupil radius (NA) is 0.68 (heavy circle). Dashed circle indicates $\sigma_{\text{Max}} = 0.85$. Diffraction orders (under axial illumination) are plotted as gray points. Circles of radius NA are erected about each order. Numbered overlap regions (53 in all) are source variables.

diffraction amplitudes), so to complete the solution we add a third step

3. Calculate a reticle pattern that provides the optimized wave front determined in step 2.

We later describe a simple approach to step 3 which exploits the linearity of the diffraction Fourier transform. As we have seen, step 2 can be handled by standard routines (given the limited field sizes considered here). For the more difficult step 1 global optimization we simplify the problem by considering only an aberration-free image (aberrations can be deferred to step 2). Further, the algorithm described in this section optimizes only the focused image during step 1, i.e., defocus aberration is also zero. Of course, the step 2 local refinement need not be restricted to optimization of exposure latitude.

With target patterns that are periodic (or to which we apply periodic boundary conditions), optimization of a focused image allows us to partition the continuous space of possible illumination directions into a fairly small number of distinct regions, since two illumination directions are equivalent (when aberrations including defocus are zero) if they direct the same set of diffraction orders into the collection pupil. This is illustrated in the *k*-space diagram of Figure 13. The entrance pupil (centered on the origin) has radius NA=0.68 in this example. $\sigma_{\text{Max}}=0.85$ is assumed as the illumination limit imposed by the stepper (shown as a dashed circle). The optimization program next divides the entrance pupil into independent source regions whose boundaries are formed by circles of radius NA centered on each diffraction order. The diffraction orders plotted in Figure 13 assume $\lambda = 248$ nm, and a staggered array with 1120 nm horizontal pitch and 560 nm vertical pitch.

Figure 14 shows a DRAM isolation pattern laid out on such an array. The rectangles (dark for positive resist) have width F equal to 140 nm. The vertical spacing of the rectangles is also F, and their length $6.5F$. The desired horizontal separation between the rectangle tips is $1.5F$.

The diffraction pattern shown in Figure 13 is produced by illumination on axis. (The diffraction orders are plotted as gray points.) The orders shift as the illumination is tilted, but the associated array of pupil-sized circles should be considered fixed in the lens aperture. Each circle then represents the range of illumination directions for which a given order can be collected, and each overlap region represents a range of illumination directions that provides the same set of collected orders. We can without loss of generality represent the fourfold symmetric source which optimizes any focused image (laid out on the Figure 13 pitch) using only 53 distinct variables, with each variable representing the illuminating intensity from one of the different pupil regions identified in the Figure 13 construction. We will denote these unknowns as a vector variable **s** (of length 53 in this example). Note that each element of **s** represents a set of 1, 2, or 4 equally intense illuminating beams that impinge on the mask from mirrored directions. If we assume that the illuminator fills all open illumination directions with a fixed power per unit solid angle, the variables must be constrained according to

$$0 \leqslant s_j \leqslant S_{\text{Max},j}, \tag{1}$$

where $S_{\text{Max},j}$ is the area of the jth illumination region in the pupil. If the source distribution is defined by diffractive elements it is more appropriate to constrain the summed intensity.

The m,nth diffraction order would ordinarily be defined as the amplitude $a_{m,n}$ that (under axial illumination) diffracts from the reticle in a direction $\mathbf{k}_{x,y} = k_0(m\lambda/p_x, n\lambda/p_y)$, with p_x and p_y the unit cell periodicities. However, for our problem it is desirable that the unknown amplitude variables represent independently adjustable components of the wave front, and (as conventionally defined) the collected orders are not entirely independent of one another (see earlier). For bilaterally symmetric patterns we adopt a notation in which m and n are nonnegative; $a_{m,n}$ then represents a single nonredundant unknown. Thus, in the Figure 13 example, three independent orders ($a_{0,0}, a_{1,1}, a_{2,0}$) are collected with axial illumination (source region 40), whereas seven are collected under illu-

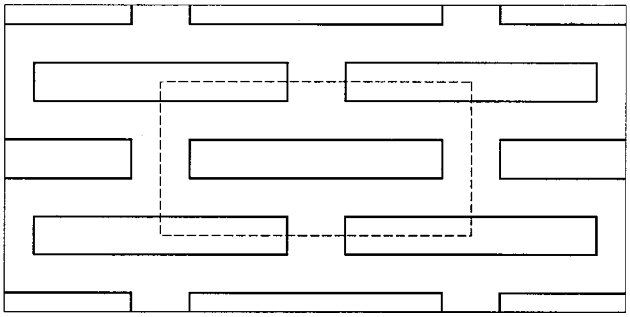

Fig. 14 Isolation pattern with periodicity matching Figure 13. Width of dark rectangles (denoted F) is 140 nm; separation between tips is 210 nm. Later figures plot optimized images over the region shown dashed.

mination from off-axis region 8 ($a_{0,0}, a_{1,1}, a_{2,0}, a_{3,1}, a_{2,2}, a_{0,2}, a_{4,0}$).

For a given source direction j, the normalized wafer-plane amplitude $b_{m,n,j}$ that is produced by an unknown amplitude $a_{m,n}$ may then include the result of interference between superimposed waves from the $\pm m, \pm n$ directions. In other words, $b_{m,n,j}$ may be given by

$$b_{m,n,j} = e^{2\pi i(mx/p_x + ny/p_y)},\ \text{ or }\ 2e^{2\pi imx/p_x}\cos\left(\frac{ny}{p_y}\right),$$

or

$$2e^{2\pi iny/p_y}\cos\left(\frac{mx}{p_x}\right),\ \text{ or }\ 4\cos\left(\frac{mx}{p_x}\right)\cos\left(\frac{ny}{p_y}\right), \tag{2}$$

depending on whether or not particular negative orders in the x,y mirror directions are collected simultaneously. It is convenient to write the $a_{m,n}$ and $b_{m,n,j}$ quantities as vectors; **a** for the unknown order amplitudes (including all orders that can be captured from at least one feasible illumination direction), and $\mathbf{c}_1$ and $\mathbf{c}_2$ for the real and imaginary parts, respectively, of **b**. To provide proper symmetry in the image we illuminate the reticle symmetrically from mirrored directions, which we distinguish with an index q. Using an index h to separate real and imaginary parts, we then have for the image intensity

$$I(x,y) = \sum_{q=1}^{4} \sum_{j=1}^{J_{\text{Max}}} \sum_{h=1}^{2} s_j (\mathbf{c}_{q,j,h} \bullet \mathbf{a})^2. \tag{3}$$

To optimize exposure latitude we now seek the global solution to the generalized fractional programming problem

$$\underset{\mathbf{s,a}}{\text{Maximize}}\ \ \Psi(\mathbf{s},\mathbf{a}),\ \ \text{where}\ \ \Psi(\mathbf{s},\mathbf{a}) \tag{4}$$

$$\equiv \underset{r}{\text{Min}}\left[\Delta\text{CD}_r \frac{\Sigma_{q=1}^{4}\Sigma_{j=1}^{J_{\text{Max}}}\Sigma_{h=1}^{2} s_j(\mathbf{c}_{q,j,h,r}\bullet\mathbf{a})(\nabla_\perp \mathbf{c}_{q,j,h,r}\bullet\mathbf{a})}{\Sigma_{q=1}^{4}\Sigma_{j=1}^{J_{\text{Max}}}\Sigma_{h=1}^{2} s_j(\mathbf{c}_{q,j,h,r}\bullet\mathbf{a})^2}\right]$$

subject to

$$\sum_{j=1}^{J_{\text{Max}}} s_j \geqslant S_{\text{Min}},$$

$$0 \leqslant s_j \leqslant S_{\text{Max},j} \qquad (\forall j | 1 \leqslant j \leqslant J_{\text{Max}}),$$

$$\sum_{q=1}^{4}\sum_{j=1}^{J_{\text{Max}}}\sum_{h=1}^{2} s_j(\mathbf{c}_{q,j,h,r}\bullet\mathbf{a})^2 = Q, \quad (\forall r | 1 \leqslant r \leqslant r_{\text{Max}}),$$

where Q is a nonpreset constant, independent of r, and

$$\sum_{q=1}^{4}\sum_{j=1}^{J_{\text{Max}}}\sum_{h=1}^{2} s_j(\mathbf{c}_{q,j,h,u}\bullet\mathbf{a})^2 \geqslant I_{\text{Bright}}Q \quad (\forall u | 1 \leqslant u \leqslant u_{\text{Max}}),$$

$$\sum_{q=1}^{4}\sum_{j=1}^{J_{\text{Max}}}\sum_{h=1}^{2} s_j(\mathbf{c}_{q,j,h,v}\bullet\mathbf{a})^2 \leqslant I_{\text{Dark}}Q \quad (\forall v | 1 \leqslant v \leqslant v_{\text{Max}}).$$

Here the index r refers to sample points (x_r, y_r) along the edges of the target patterns. $\nabla_\perp \mathbf{c}$ represents the derivative of c in a direction normal to the feature edge. Maximization of Ψ ensures that the shallowest log slope among feature edges is as steep as possible. The log slope at each edge is weighted in proportion to the local CD tolerance (denoted ΔCD). Indices u and v run over sample points that must be bright and dark, respectively. Constraints are imposed to (i) require achievement of minimum acceptable pupil fill, (ii) enforce geometric restrictions on the size of the s_j source regions, (iii) prevent line shortening and other CD errors in the printed pattern, (iv) require adequate exposure in bright areas, and (v) prevent excessive exposure in dark areas.

Techniques are reported in the literature for solving fractional optimization problems like Eq. (4), often reducing them to a parametric problem in the difference between numerator and denominator.[14] Equation (4) can also be approximated as a cubic polynomial optimization; a global optimum is then guaranteed in principle if a homotopy algorithm is used to solve the Lagrangian. However, we have found that problems of the size considered here pose considerable difficulty for homotopy algorithms reported in the literature.[15]

Our solution scheme for Eq. (4) exploits global solutions we have found for two simplified sub-problems in the equation. This decomposition method constitutes step 1 of our overall algorithm to optimize exposure latitude (denoted algorithm P), which is outlined in the following table:

Algorithm P

(0) Preliminary

(a) Problem definition; user specification of image sample points.

(b) Determine the J_{Max} source variables via Figure 13 construction.

(1) Global optimum

(a) Considering each source variable one at a time, calculate a global solution for $\mathbf{a}_j$ using simplified criteria.

(b) Initialize amplitude variables **a** to the best value obtained in previous step. Initialize S_{Min} to 0.

(c) Calculate the globally optimum source distribution **s** for the current values of **a** and S_{Min}.

(d) Use a local algorithm to optimize **s** and **a** together [per Eq. (4)].

(e) Increase S_{Min} by small increment and return to step c, until stopping criteria are met.

(2) Fix S_{Min} at desired final level and choose corresponding solution from step 1, then refine using local optimizer against more complex criteria.

(3) Calculate the optimum reticle pattern that produces wave front **a** with maximum intensity.

(a) Find global solution that produces wave front with maximum intensity.

(b) Refine step 3(a) solution using local optimizer to, e.g., satisfy mask CD tolerances, reduce shapes to Manhattan geometries, etc.

Let us now consider these steps in more detail. In calculating the step 1(a) amplitude sets $\mathbf{a}_j$, we defer constraints on equal feature bias and minimum pupil fill to step 1(c). Moreover, the overall intensity scaling of the amplitudes $\mathbf{a}$ is arbitrary until the step 3 mask calculation. This allows us to artificially set the intensity at active bright point constraints to 1 (the other bright points, usually including those away from feature edges, then being above 1). This indirect constraint eliminates the need to optimize against log slope per se until step 1(c), since slope and log slope are equalized at unit intensity. As a further simplification, we optimize in step 1(a) against the finite intensity difference across feature edges (i.e., between dark and bright points adjacent to the edges), rather than against a true derivative.

The step 1(a) optimization problem for the jth source direction is then to minimize intensity in dark points under these constraints, and we can write the problem in matrix form as

$$\text{Minimize} \quad \Phi_j(\mathbf{a}) = \mathbf{a}^T \mathbf{A}_0 \mathbf{a}$$

$$\text{subject to} \tag{5}$$

$$\mathbf{a}^T \mathbf{A}_u \mathbf{a} \geqslant 1 \qquad (\forall u | 1 \leqslant u \leqslant u_{\text{Max}}).$$

The symmetric $\mathbf{A}_0$, $\mathbf{A}_u$ matrices [obtained from Eq. (3)] take into account any orders that may be collected from negative directions, as well as the effect of mirroring the illumination. The $\mathbf{a}^T\mathbf{A}_u\mathbf{a}$ quadratic forms in the constraints of Eq. (5) represent the intensity at bright sample points, while the $\mathbf{a}^T\mathbf{A}_0\mathbf{a}$ term in the demerit function provides the average intensity within dark areas of the image. Algorithm P handles extended dark regions by the simple expedient of giving preferential weight to dark sample points that are adjacent to feature edges. The dark-region average is typically a very small quantity, since we are optimizing exposure latitude in focus. Proper polarity in all dark points is thus ensured, since conversion of even a single dark point to bright would drastically raise the average, i.e., Φ could not be minimal in such a case. [Note that we are free to suppose that only a limited number of dark points participate in this average, since points are not mutually constraining (in a direct way) if their separation greatly exceeds the lens resolution.] On the other hand, it is necessary that each bright point be entered as a separate constraint, since the optimizer can sometimes make an invalid improvement in the average bright-to-dark contrast by switching a few difficult bright points to dark.

Though the matrices in the Eq. (5) quadratic forms (ellipsoids) can be made positive definite, the problem is nonconvex because the inequality constraints are lower bounds (i.e., the region external to the $\mathbf{a}^T\mathbf{A}_u\mathbf{a}=1$ ellipsoids is not a convex domain). However, two aspects of the Eq. (5) structure allow the multiple local minima to be fully mapped in a very efficient way. First, Eq. (5) is already in homogeneous form, i.e., the Eq. (5) ellipsoids share a common center, and second, their principal axes (whose lengths are the reciprocal square roots of the matrix eigenvalues) must range between very small and very large amplitudes (since for feasible values of λ/NA it must be possible to print a wide range of image intensities on at least a subset of the sample points).

To exploit this property we first calculate the eigenvectors and eigenvalues of the black-region matrix $\mathbf{A}_0$. We then scale the eigenvectors by the square root of the reciprocal of the eigenvalues, thereby effectively scaling the diagonalized black region matrix to the identity matrix. The eigenvector basis can now be rotated into alignment with the eigenvectors of the matrix for mean intensity at bright points (average of the $\mathbf{A}_u$, denoted $\mathbf{A}_{\bar{U}}$). If we use the symbol $\mathbf{E}$ to denote eigenvector column matrices (i.e., $\mathbf{E}_Q$ denotes the column-matrix eigenvectors of a matrix $\mathbf{A}_Q$) then the transformation $\mathbf{W}$ from the new basis to the old is given by

$$\mathbf{W} = \mathbf{E}_0 \mathbf{D}_0^{-1/2} \mathbf{E}_B, \quad \text{where} \quad \mathbf{A}_B \equiv \mathbf{D}_0^{-1/2} \mathbf{E}_0^T \mathbf{A}_{\bar{U}} \mathbf{E}_0 \mathbf{D}_0^{-1/2}, \tag{6}$$

with the reciprocal square root of $\mathbf{D}_0$ denoting a diagonal matrix formed from the reciprocal square roots of the eigenvalues of $\mathbf{A}_0$. In basis $\mathbf{W}$, the summed squared amplitudes give the mean black-region intensity, and also the mean bright-region intensity when weighted by the eigenvalues of $\mathbf{A}_B$.

It is only possible to simultaneously diagonalize two matrices in this way (see treatment in Ref. 16), and no single eigenvector for the mean bright and dark region intensities is likely to provide high brightness at all bright sample points. Since the eigenvectors are only common to the mean intensities of the dark and bright regions, we cannot immediately calculate the relative eigenvector weightings that are required to provide an optimum image from the given source (e.g., region j, four-fold mirrored, or a more complex source). However, the solution vector must lie approximately within a subspace spanned by a limited number of these eigenvectors, namely the minimal set of eigenvectors such that for each of the bright sample points, at least one eigenvector in the set provides intensity above 1. (Of all sets that meet this condition, algorithm P chooses the set whose minimum bright-region eigenvalue is largest.)

Consider, for example, the amplitude eigenvectors shown in Figure 15 (these are the columns of $\mathbf{W}$), which correspond to illumination from region 8 in Figure 13 (four-fold mirrored). Each eigenvector has unit length, so each eigenvector provides unit mean intensity in dark regions of the image. The mean bright-region intensity (which is also the contrast) is given by the associated eigenvalue. The first two eigenvectors provide very high contrast, but do not allow the horizontal separations between the rectangle tips to be printed bright. Eigenvector 3 must also be employed in order to provide high intensity at all bright sample points, indicating that black region contrast is significantly impacted by the need to achieve high intensity between the rectangle tips. (Printing the isolation rectangles is thus more difficult than printing non-terminating lines and spaces.) Eigenvectors 4 through 7 degrade contrast in the image, and so can only contribute to the solution in small amounts.

To solve Eq. (5) we now need to find the point in basis $\mathbf{W}$ which is closest to the origin while remaining outside each of the individual ellipsoids representing unit intensity

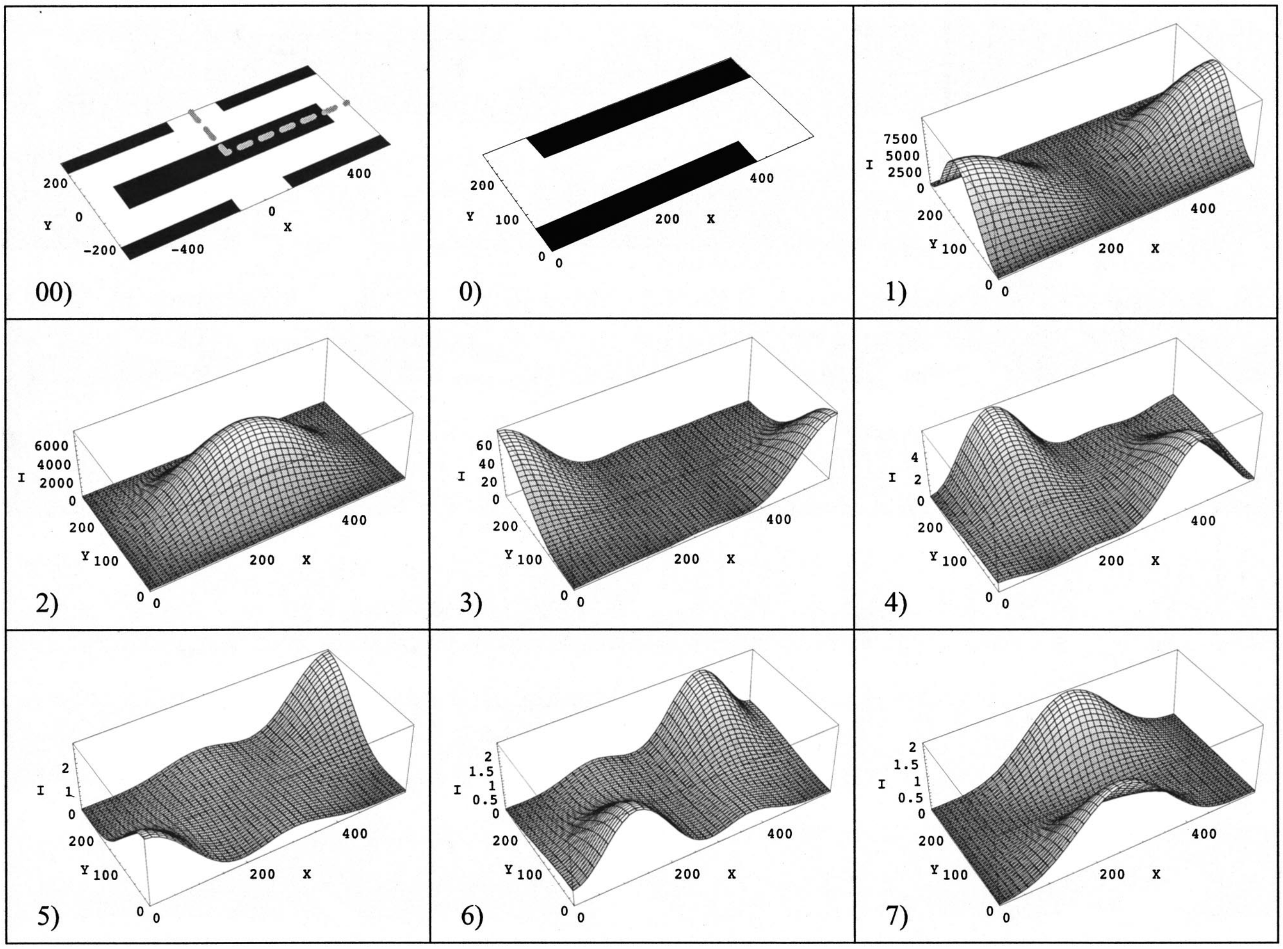

Fig. 15 Joint mean-intensity eigenvectors for bright and dark regions of Figure 14 isolation pattern, with illumination incident from source region 8 of Figure 13 (illumination is fourfold mirrored). As in Figure 13, the imaging conditions are $\lambda = 248$ nm, NA$=0.68$. Units for x and y axes are nm. (00) Perspective view of target pattern (central region of Figure 14). (0) Magnified view of target pattern (the dashed upper right quadrant of previous view). (1)–(7) The seven eigenvectors, plotted as images over upper right quadrant. Sorted in decreasing order of bright region intensity. All eigenvectors provide unit average intensity at dark sample points. Only eigenvectors 1, 2, and 3 can contribute significant amplitude to the optimal mask.

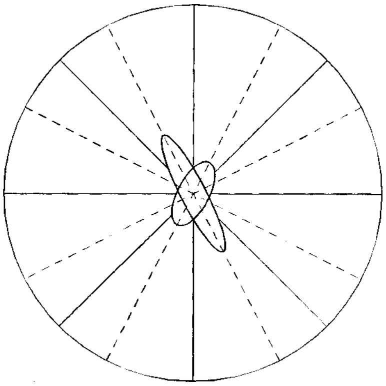

Fig. 16 Schematic of search space decomposition, for a pattern having two sample points in bright region (hence, two ellipsoids). Example in text yields three significant eigenvectors, but for ease of drawing we assume 2 in this figure (yielding a 2D subspace, where each Cartesian axis represents the amplitude of one of the basis **W** eigenvectors). In 2D the spherical triangles become arcs (bounded by dashed lines) whose midpoint radial vectors are shown solid. Note that by symmetry only half the arcs need be analyzed.

at particular bright points. We can consider the search to take place within the limited subspace spanned by the dominant eigenvectors for mean intensity (e.g., in the Figure 14 example, the three-dimensional subspace spanned by eigenvectors 1, 2, 3 of Figure 15). In order to fully probe the ‘‘nooks and crannies’’ of the intersecting ellipsoids in an efficient way, we organize the search space by erecting spherical triangles on the ‘‘celestial sphere’’ (i.e., a sphere where the intensities at all bright points are much higher than unity). The first set of vertex nodes for these bounding spherical triangles is defined by projecting the eigenvectors for individual bright points to the celestial sphere, i.e., by projecting vectors outward along the principal axes of the bright-point ellipsoids. (Of course, the algorithm must in general handle problems of arbitrary dimensionality. The number of vertices in each ‘‘triangle’’ is equal to the dimensionality of the subspace, and the ‘‘sphere’’ is a surface of dimensionality one less.) After this triangular mesh is formed on the celestial sphere, the other half of the node set is generated by splitting the triangles through the addition of a new vertex at the central coordinate of each. One can test for globality of the converged solution by further subdividing the triangular search mesh; our conclusion of globality is partly an empirical one, based on the observed sufficiency of the above single-midpoint mesh in such tests. We then proceed from each node by decreasing all amplitudes in a common proportion (i.e., along a radial vector to the origin) until we reach the outermost ellipsoid intersecting this trajectory. A local optimizer then settles into the nearest local minimum in the solution space (the innermost pocket of the intersecting ellipsoids in that region). Dark sample points away from feature edges can be omitted from the demerit function, subject only to the constraint that the intensity at such points lies below the punch-through threshold I_{Dark}. Our local optimizer uses the augmented Lagrangian algorithm in Bertsekas' textbook.[9] To exactly solve Eq. (5) during step 1(a), the local optimization should take place in the full vector space **W**. This decomposition is illustrated in Figure 16.

We should note that the method of Eqs. (5) and (6) allows the globally optimum mask to be determined for arbitrary fixed source, under the simplified formulation just described.

Once the step 1(a) subproblem is solved, algorithm P uses the solution to initialize **a**, and proceeds to the source optimization loop in step 1(c). Step 1(c) requires that we solve Eq. (4) for **s**, with **a** given. Even when **a** is fixed, Eq. (4) is nonlinear, since the merit function involves log slope. However, we can transform Eq. (4) to the linear program:

Minimize z_0

subject to

$$z_0+\mathbf{z}\cdot\sum_{q=1}^{4}\sum_{h=1}^{2}(\mathbf{c}_{q,j,h,r}\cdot\mathbf{a})(\nabla_{\perp}\mathbf{c}_{q,j,h,r}\cdot\mathbf{a})\geqslant 0 \quad (\forall r|1\leqslant r\leqslant r_{\mathrm{Max}}),$$

$$0\leqslant S_{\mathrm{Min}}z_j\leqslant S_{\mathrm{Max},j}\sum_{k=1}^{J_{\mathrm{Max}}}z_k \qquad (\forall j|1\leqslant j\leqslant J_{\mathrm{Max}}),$$

$$\mathbf{z}\cdot\sum_{q=1}^{4}\sum_{h=1}^{2}(\mathbf{c}_{q,j,h,r}\cdot\mathbf{a})^2=1 \qquad (\forall r|1\leqslant r\leqslant r_{\mathrm{Max}}), \tag{7}$$

$$\mathbf{z}\cdot\sum_{q=1}^{4}\sum_{h=1}^{2}(\mathbf{c}_{q,j,h,u}\cdot\mathbf{a})^2\geqslant I_{\mathrm{Bright}} \qquad (\forall u|1\leqslant u \leqslant u_{\mathrm{Max}}),$$

$$\mathbf{z}\cdot\sum_{q=1}^{4}\sum_{h=1}^{2}(\mathbf{c}_{q,j,h,v}\cdot\mathbf{a})^2\leqslant I_{\mathrm{Dark}} \qquad (\forall v|1\leqslant v\leqslant v_{\mathrm{Max}}).$$

Equation (7) is linear in the transformed set of $1+J_{\mathrm{Max}}$ variables $z_0,z_1,z_2,z_3,\cdots\equiv z_0,\mathbf{z}$, and so can be solved globally using standard linear programing algorithms. After Eq. (7) is solved, the step 1(c) source intensities that solve Eq. (4) are given by

$$\mathbf{s}=\frac{S_{\mathrm{Min}}\mathbf{z}}{\sum_{k=1}^{J_{\mathrm{Max}}}z_k}. \tag{8}$$

In general, the method of Eqs. (7) and (8) provides the globally optimum source to print a given mask, under the criteria of Eq. (4).

To complete our discussion of algorithm P we now describe the step 3 reticle calculation. (As noted earlier, it is straightforward to carry out the various local optimization steps in P using standard routines.) To begin with, we calculate the set of reticle patterns that provide the brightest possible image consistent with the step 2 solution for **a**. This initial layout must be then refined using standard criteria; for example, the optimized patterns must be rendered on the mask as polygons, preferably as a set of rectangles. The rectangles can be fairly coarse, e.g., of dimension only moderately smaller than the lens resolution. We use a local optimizer to do this refinement.

For the basic reticle calculation we approximate the Fourier diffraction integral as a summation over discrete sample points. The mask transmission function $T(x,y)$ is sampled on a two-dimensional (2D) grid, and then unraveled into a one-dimensional (1D) vector of unknowns **T** indexed by g:

$$\begin{aligned}\int_{-p_x/2}^{p_x/2}\int_{-p_y/2}^{p_y/2}dx\,dy\,T(x,y)e^{2\pi i(mx/p_x+ny/p_y)} &\cong\sum_{k=1}^{K}\sum_{l=1}^{L}T(x_k,y_l)e^{2\pi i(mx_k/p_x+ny_l/p_y)}\\ &\equiv\sum_{g=1}^{KL}T_g b'_{g,m,n}\\ &\equiv\sum_{g=1}^{KL}T_g b'_{g,w}.\end{aligned} \tag{9}$$

The symbol b' has been introduced in Eq. (9) as shorthand for the exponential, and an unraveled index w is introduced to represent the m,n indices of the wth captured amplitude in **a**. In replacing the integral in Eq. (9) by a simple sum, we are implicitly assuming small pixels. [When using the Eq. (9) formulation we generally choose a pixel size that is appreciably finer than the grid step actually used for mask fabrication.]

Step 3(a) now becomes a linear programing problem:

$$\text{Maximize}\quad \Omega(\mathbf{T})\equiv\mathrm{Sign}\left[\sum_{w=1}^{W_{\mathrm{Max}}}a_w\right]\sum_{g=1}^{KL}\sum_{w=1}^{W_{\mathrm{Max}}}T_g b'_{g,w},$$

subject to

$$\sum_{g=1}^{KL}T_g\left[\left(a_{w'}\sum_{w=1}^{W_{\mathrm{Max}}}b'_{g,w}\right)-\left(b'_{g,w'}\sum_{w=1}^{W_{\mathrm{Max}}}\right)\right]=0 \quad (\forall w'|1\leqslant w'\leqslant W_{\mathrm{Max}}), \tag{10}$$

$$T_{\mathrm{Min}}\leqslant T_g\leqslant T_{\mathrm{Max}}.$$

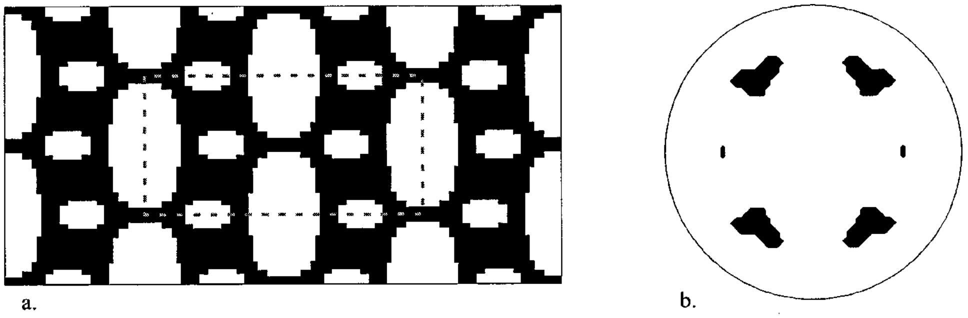

Fig. 17 Mask and source solution for Figure 14 isolation pattern using algorithm P (with step 2 omitted). (Results from a more sophisticated algorithm than P are shown in Figure 21.) (a) Chromeless (nonalternating) mask [$T_{\mathrm{Min}}=-1$ (shown black), $T_{\mathrm{Max}}=+1$ (shown white)]. Plotted region matches Figure 14. The mask features have a very different shape from the target patterns. (b) Binary source. Circle represents 0.68 NA. Illumination directions are shown dark.

Equation (10) forces the mask Fourier orders to be in the same ratio as the elements of the optimized diffraction order list **a** obtained in step 2. T_{Min} and T_{Max} are determined by the mask technology. T_{Max} would generally be $+1$, while T_{Min} would be, e.g., -1 for a chromeless mask, $-\sqrt{0.065}$ for an attenuating phase-shift mask with 6.5% chrome transmission, etc. In general we must set $T_{\mathrm{Min}}<0$ for Eq. (10) to provide a solution.

Equation (10) can be modified to adjust the exposure threshold of the printed pattern (e.g., to match its intensity with that provided by some other set of mask patterns) by adding the constraint

$$\mathrm{Sign}\left(\sum_{w=1}^{W_{\mathrm{Max}}} a_w\right)\sum_{g=1}^{KL}\sum_{w=1}^{W_{\mathrm{Max}}} T_g b'_{g,w}=\Omega_{\mathrm{Match}}. \tag{11}$$

This adjusts the intensity of the aerial image without changing its shape. Ω_{Match} must of course be smaller than the unmodified Eq. (10) maximum. To prevent excessively fine features in the returned solution, one can introduce a spatially smoothed version of the unmodified solution as a new objective vector. This gives preference to pixel adjustments near the edges of features, where the magnitude of the smoothed pattern passes through zero (so that correlation with the new objective vector is maximized when adjustments are made at the edges of existing features, rather than in newly introduced features). Alternatively such criteria can be enforced in the step 3(b) local optimization.

In the limit of an arbitrarily fine grid, the solution provided by Eqs. (10) and (11) will be "two-tone," in that (essentially) all pixels will be driven to either T_{Min} or T_{Max}. (Explicit discretization constraints are not needed.) To design a Levenson-type mask (i.e., a mask with 0° and 180° apertures opened in opaque chrome), we modify Eq. (10) with a change of variables and added constraints

$$\begin{aligned} &T_g \rightarrow T_g^+ - T_g^-, \\ &T_g^+ \geqslant 0,\ \ T_g^- \geqslant 0, \\ &\sum_{g=1}^{KL}(T_g^+ + T_g^-) \leqslant (1-G)KL. \end{aligned} \tag{12}$$

If parameter G were allowed to float, the change of variables in Eq. (12) would not revise the solution of Eq. (10) (assuming $T_{\mathrm{Min}}=-1, T_{\mathrm{Max}}=+1$), since the first two lines of Eq. (12) permit a transmission of ± 1 to be realized whenever the third line is not binding. This latter constraint is activated by setting G to a positive value; a fraction G of the reticle area is then driven to opaque chrome (i.e., T_g^+

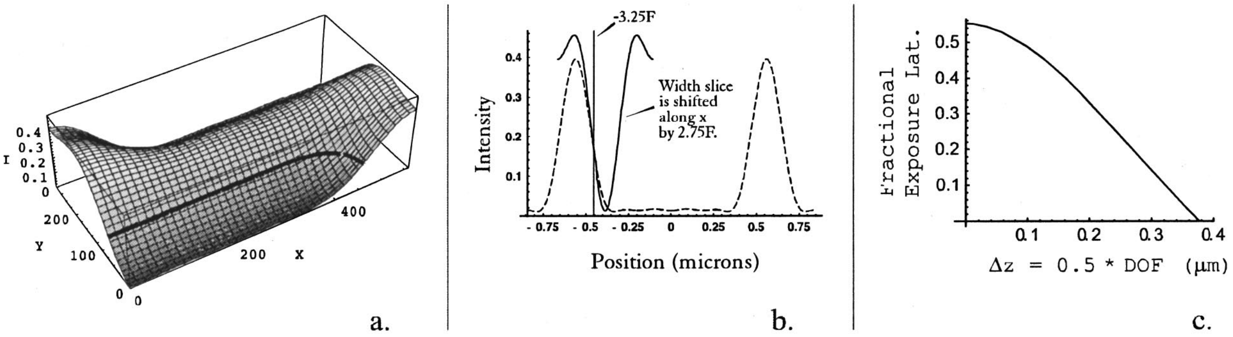

Fig. 18 (a) Focused aerial image from the Figure 17 solution (same perspective as Figure 15.0). Thick curve shows contour slice at nominal threshold. (Only the contour for the front rectangle of Figure 15.0 is visible.) (b) Horizontal (dashed) and vertical (solid) centerline slices through rectangle image. The vertical slice is shifted by the difference between the nominal length and width to show that the aerial image contour prints without line shortening. (c) Process window (thresholded aerial image model, assuming no aberrations). Exposure latitude is 55%, but DOF is small (less than ±0.4 μm), reducing process window to 24.7% μm (compare with Figure 23).

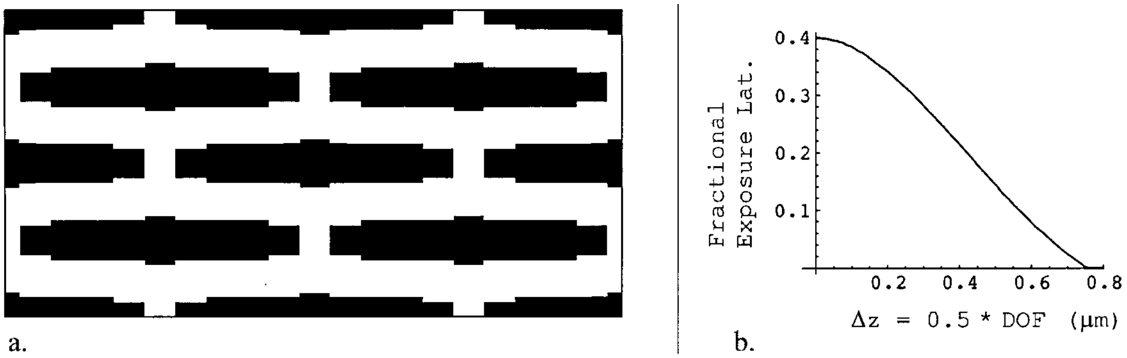

Fig. 19 Solution provided by conventional RET approach (using local optimizer to maximize integrated ED window, with nominal Figure 14 patterns as starting mask solution). Annular illumination parameters are optimized simultaneously, yielding $\sigma = 0.50$, 0.78. (a) Mask solution (phase shift chrome, $T = 6.5\%$), over same region as Figure 14. (b) Process window (thresholded aerial image model, assuming no aberrations).

$=T_g^- = 0$). As with Eq. (10) there is no need to impose explicit discretization constraints if the bitmap pixels are sufficiently fine.

Figure 17 shows the solution provided by algorithm P for the isolation pattern of Figure 14, in the simple case where the step 2 local optimization is omitted. Log slope across the narrow width of the rectangles is given 1.5× more stringent weighting than log-slope at the tips of the rectangles, corresponding to a tighter CD on the width than the length (tighter in absolute terms; relative tolerances are the same). Figure 18 shows the aerial image in focus. The intensity along the centerline of the dark rectangles is roughly 1/30th that at peak. When spacewidth tolerances of ±20% are applied to the bright horizontal and vertical separations between the rectangles, the exposure latitude is 55%. This is about a 1.4× improvement over the 40% exposure latitude achieved by a more conventional OPC approach, in which feature boundaries and source parameters are adjusted using a local optimizer (see next secion).

4 Optimization of Process Window Versus Exposure Latitude

Unfortunately, the depth of focus provided by the Figure 17 solution is not very large (±0.38 μm under the above ±20% CD tolerance), leading to an integrated process window of only 24.7% μm (using a thresholded aerial image model), despite the large exposure latitude in focus. This process window is considerably better than can be achieved with a simple opaque chrome mask incorporating the nominal patterns. However, standard OPC methods can do appreciably better. Figure 19 shows the result of using a local optimizer to adjust the shapes of mask openings in phase-shift chrome, with the nominal Figure 14 pattern serving as a starting solution. The inner and outer radii of annular illumination were adjusted simultaneously. Depth of focus is ±0.75 μm, substantially exceeding that of the Figure 17 solution, and a better process window overall is achieved (33.3% μm). Figure 20 shows plots of the aerial image.

We should emphasize that this decoupling of process window and exposure latitude does not always arise. Consider, for example, the optimization of mask and source to print the Figure 2 pattern: While the optimal Figures 3 and 4(a) solution was obtained using an algorithm that maximizes full process window, a very similar solution is provided by algorithm P (with the step 2 local optimization omitted). Process window with algorithm P is 37.6% μm, vs 45% μm for the solution of Figures 3 and 4(a). Indeed, the Figures 3 and 4(a) solution can be recovered exactly from algorithm P if process window is used as the merit function in step 2.

It is possible to attack the Figure 14 problem in the same way; i.e., by refining the step 1 solution (Figure 17) against process window using a local optimizer (step 2 of algorithm P). The solution found in this way yields a process window of 36.2% μm, slightly exceeding that of the more conventional Figure 19 approach. The step 2 refinement is found to improve depth of focus by 50% while decreasing exposure latitude only 2%, demonstrating again that pro-

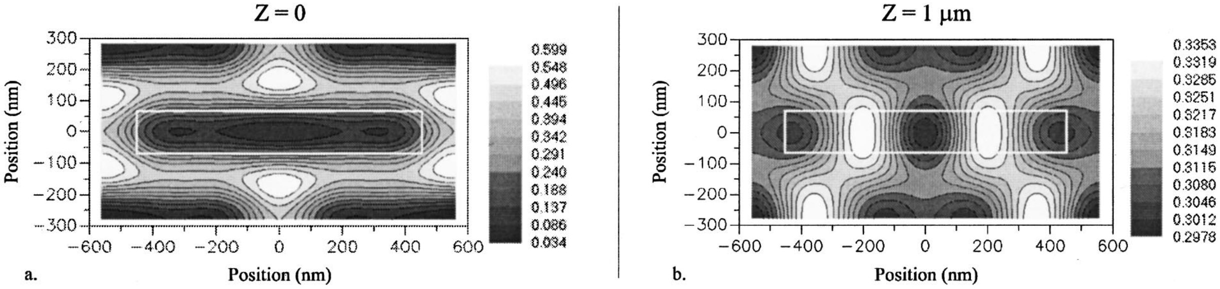

Fig. 20 Images from Figure 19 conventional RET solution. Plotted region matches dashed area of Figure 17. White insert shows nominal perimeter of the central dark rectangle. (a) Image in focus. (b) Defocused 1 μm. Image no longer shows useful modulation.

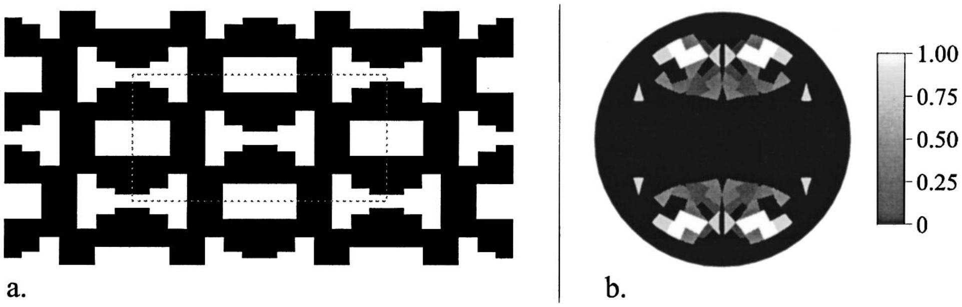

Fig. 21 Globally optimized solution to maximize process window for Figure 14 pattern. (See also Figure 17 solution, which only optimizes exposure latitude.) (a) Chromeless mask (nonalternating). Black represents 0° phase shift, white 180°. Plotted region matches that in Figures 14, 17, and 25. (b) Jointly optimized gray-scale source.

cess window and exposure latitude are not always strongly coupled. Clearly, it is preferable to have a global algorithm that can directly optimize the mask and source for maximum process window.

We have developed a preliminary version of such an algorithm. Integrated area under the ED window is maximized, assuming a thresholded aerial image model. Figure 21 shows the solution obtained by this method for the Figure 14 isolation pattern; Figures 22 and 23 show the resulting image and process window. (The solution of Figures 3 and 4(a) was also obtained with this algorithm, additionally imposing binary values on the source.) Integrated process window is 67% μm (see Figure 23), about double that obtained with the more conventional RET optimization of Figure 19 (and also about double that obtained by optimizing for process window in step 2 of algorithm P). The improvement in depth of focus may be seen by comparing Figures 20 and 22. (The tradeoff between exposure latitude and DOF that can be observed in Figure 23 is not unusual; in many cases we find that, in effect, our algorithm can achieve a larger increase in process window by increasing DOF than by increasing exposure latitude.) Figure 24 emphasizes the dramatic difference between the optimized mask shapes of Figure 21 and the resulting printed pattern.

Figure 25 shows an implementation in opaque chrome (i.e., a Levenson mask where features have unit transmittance and 0° or 180° phase shift). In general, Eq. (10) and related methods provide highest efficiency in chromeless technology, and Figures 5 and 22 demonstrate that reasonably high intensities can be achieved. We have found these methods to be quite successful in compensating the greater difficulty in maximizing intensity when a decentered wave front slice is optimized. Of course, exposure time will be significantly degraded if the optimized source is provided by an attenuating aperture rather than diffractive elements (as in exposure tools that provide software-selectable source distributions via a library of preloaded diffractive elements;[17] see also Ref. 8).

5 Conclusions and Future Directions

To achieve maximum process window one should not constrain reticle shapes to follow the inherent ‘‘topology’’ of an initial design form. By considering the implications of off-axis illumination in a detailed way, we have devised a design algorithm that is not encumbered by such restrictions. The theoretical improvement in performance from this global approach can be quite substantial. Further, our basic analytical approach allows many extensions; for example, our equations are little changed if certain of the mask source variables are made to contribute during separate exposures. This allows double-exposure printing to be globally optimized without reference to preconceived assumptions about how the target pattern should be divided.

Of course, many practical issues remain to be considered. The present paper focuses on development of the basic algorithm, but it is important that the solutions be compatible at a detailed level with practical constraints imposed by the illuminator and the mask-making process. For example, it is possible that the illumination will need to satisfy tighter requirements on directional uniformity when pattern symmetry is provided by the source rather than the collected wave front.

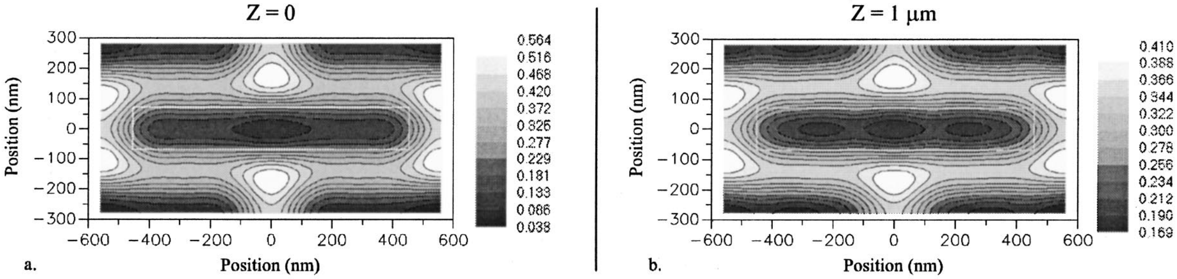

Fig. 22 Aerial images for the Figure 21 solution [screen captures from Prolith (see Ref. 7) simulations]. Plotted region matches dashed area of Figure 21 (also matches Figure 20). White insert shows nominal perimeter of the central dark rectangle. (a) Image in focus. (b) Defocused 1 μm. DOF is considerably larger than with conventional enhancement approach.

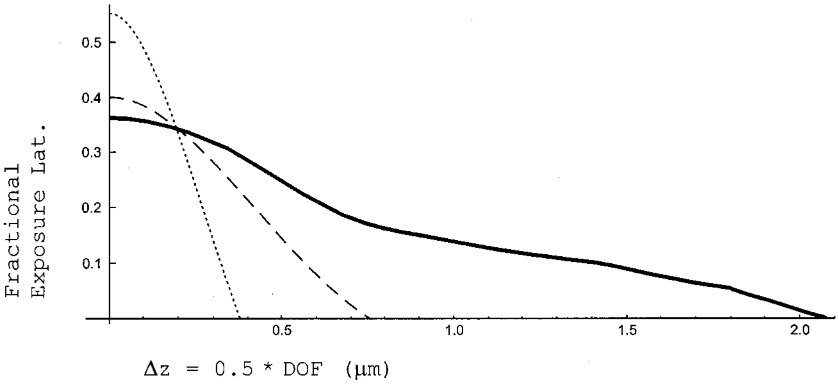

Fig. 23 The thick curve plots the process window for the Figure 21 solution, with ±20% CD tolerances on the bright horizontal and vertical separations between rectangles. A thresholded aerial image model is used, and an aberration-free lens assumed. The integrated window (two-sided) is 67% μm. The dotted curve superposes the window obtained by optimizing for exposure latitude in focus [repeats Figure 18(c)], while the dashed curve shows the performance of the conventional RET solution [repeats Figure 19(b)].

Global optimization must also be integrated into an overall strategy to print a given integrated circuit (IC) level. The field sizes considered earlier are sufficient for, e.g., separate exposure of the array region of a DRAM level, but for general purposes this is not adequate. Several approaches are available to accommodate larger sets of patterns. While globally optimized designs are often somewhat novel and unexpected, one can generally understand them ''after the fact'' in an intuitive way that is more compatible with a lithographer's ''bag of tricks'' than is possible for a purely mathematical result. Our discussion of global algorithms has been couched in terms of optimizing mask and source together; however, once the source has been optimized for critical patterns, it is possible to globally optimize less critical mask patterns with the source distribution held fixed [e.g., see Eqs. (5) and (6)]. The source can also be ''softened'' to improve compatibility with a wider range of shapes.[18]

Though the algorithm can be extended by such techniques, computational limitations make it necessary to interface the globally optimized solutions with neighboring patterns that are derived by other means. Periodic boundary conditions entail additional computational burden when target patterns are nonperiodic, e.g., to feather overlapping solutions across redundant buffer regions. Equation (11) allows the exposure threshold in a given aerial image to be adjusted up or down to maximize the common window with other patterns.

Though computational requirements make these hybrid approaches inevitable over full IC levels, it is interesting to speculate on how the benefits from global optimization might scale if no compromises were made, i.e., to assess the potential advantages of global optimization as the dimensional scale and pattern diversity of the simultaneously optimized feature set is increased. A key question is the extent to which we can preserve the synergy from joint optimization of mask and source when using the source to print a mix of critical and less critical patterns.

Off-axis illumination continues to provide access to more degrees of freedom when a pattern is optimized as a member of a group rather than individually, and, as we have seen, these degrees of freedom are in principle best optimized with a global algorithm. In general, the common process window for a group of features will usually be less than that of the features considered individually. Global optimization may prove a useful tool to bring to bear on this problem. On the other hand, the relative advantage of global optimization over conventional methods might decrease when a suite of patterns is optimized, since conventional methods already employ broader and more symmetric sources than are required for individual patterns. The Figure 13 construction implies that large-σ illumination directions along the 45° azimuths provide the largest number of independent collected orders when patterns are highly symmetric, potentially improving the prospects for optimiz-

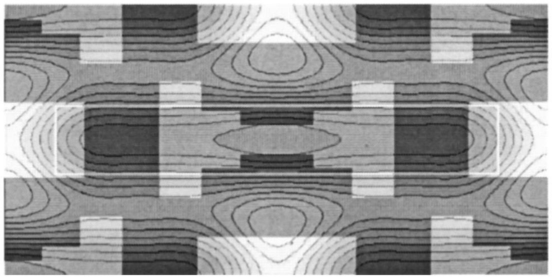

Fig. 24 Superposition of Figures 21(a) and 22(a). The dark image rectangles are centered on the bow-tie shapes. The centers of the rectangular mask features print bright. Plotted area corresponds to dashed regions of Figures 14, 17, 21, and 25.

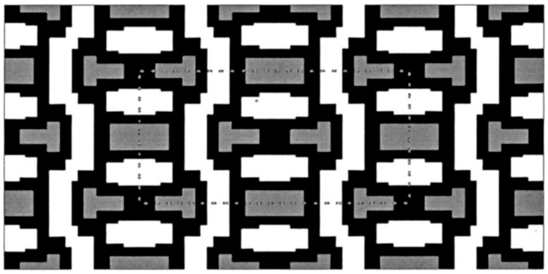

Fig. 25 Implementation of Figure 21 solution as Levenson mask. Opaque chrome is shown black; white and gray represent openings of 0° and 180° phase shift. (Mask is not alternating.) Plotted region is the same as Figure 14. Chrome coverage (low in this example) can be adjusted up or down [see Eq. (12)].

ing a broad set of patterns. Global optimization can theoretically allow the less critical patterns to be printed with a narrower and more discrete source than usual (i.e., a source optimized for critical patterns), but this may entail optimization of a great many shapes. While a fully global algorithm cannot in principle do worse than local optimization, it imposes a distinctly greater computational burden, which may force significant compromises. It remains to be established how these factors will trade-off when optimizing the pattern content of different IC levels.

To make a preliminary exploration of this question, we have extended our algorithm to optimize two independent mask regions under a single source (with the source and the two masks jointly optimized for maximum common window through focus).

For an initial test problem we optimized two separate line/space patterns with different pitches; first, a 2:1 line/space pattern, and second, equal lines and spaces. The CD of the space was $k=0.35$ in each case (specifically, 90 nm spaces, and 90 or 180 nm lines, at $\mathrm{NA}=0.75$, $\lambda=193\,\mathrm{nm}$, $\sigma_{\mathrm{Max}}=0.88$). The 2:1 pattern is in the so-called ‘‘forbidden pitch’’ region, where the line is too narrow for assist features to provide strong benefit. Oblique illumination is required since source points near the center of the pupil provide no useful modulation in the 1:1 pattern; thus, any solution for the common source will (for the most part) print the 2:1 pattern using only two-beam interference (since source points away from the center of the pupil only provide two collected orders at this pitch).

These difficulties apply with conventional enhancement methods as well as the global algorithm described here. For example, a conventional strategy that combines annular illumination, attenuated phase shift, mask bias, and assist features (with feature biases, assist widths, and illumination radii jointly optimized), can only achieve a common process window of 3.6% μm (for a ±9 nm tolerance on the 90 nm CDs).

Of course, more aggressive methods are available for patterns like those of our test problem; for example, we can employ gray-tone mask technology to equalize exposures in the two patterns. (The same effect can be obtained using a ‘‘dotted-line’’ mask, where the dot duty cycle is used to adjust exposure, and the dot pitch is too fine to be resolved.) In addition, there is a conventional specialized source that is known to be appropriate for patterns like these, namely a dipole source. If we simultaneously adjust the dipole position and the relative transmission of the two masks (each mask providing attenuated phase shift), while optimizing as before the assist widths and feature biases, we can achieve an integrated process window of 5.5%-μm.

By comparison, our global algorithm achieves an integrated process window of 13.6%-μm (without using graytone masks). Common process window is maximized, under constraints requiring that the source occupy at least 10% of the available pupil (with appropriate weighting for gray level source regions), that the exposure latitude in focus be at least 10%, and that the intensity in the minima of lines be no larger than 15% of peak. (A 10% pupil fill constraint was also imposed on the dipole solution, fixing the size of the individual poles.)

The optimized solution is shown in Figure 26; Figure 27 shows the resulting images (in focus) and process window.

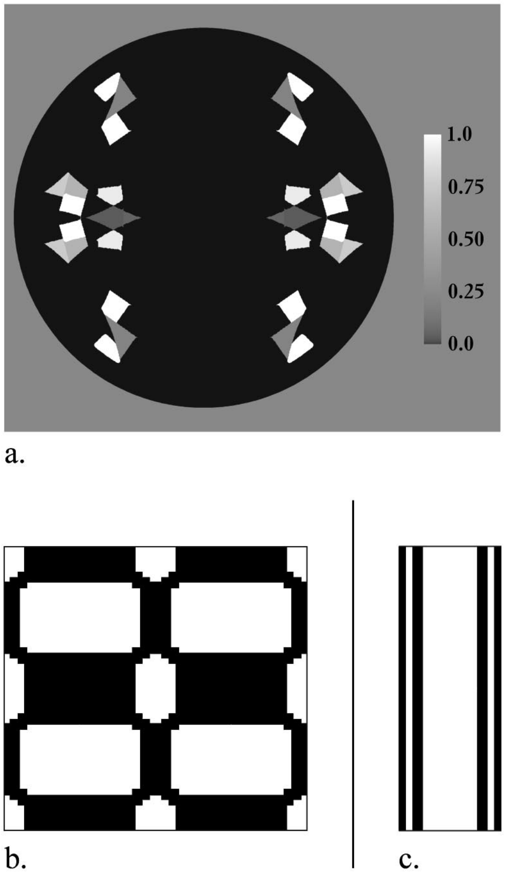

Fig. 26 Extension of the algorithm to maximize the common window of two independent mask patterns printed under a single source. In the example shown the two patterns are 2:1 and 1:1 line/space patterns (90 nm space CD with ±9 nm tolerance, alternated with 90 or 180 nm lines, printed at $\lambda=193\,\mathrm{nm}$, $\mathrm{NA}=0.75$, $\sigma_{\mathrm{Max}}=0.88$). (a) Optimized source. (b) Chromeless mask for (vertical) 2:1 lines and spaces. The mask shapes are quite different from the printed line/space pattern; two periods are shown, and the smaller battery-shaped mask features are aligned with the dark printed lines. (c) Chromeless mask for 1:1 lines and spaces. One period is shown, beginning at the center of a dark printed line.

The mask for the 2:1 patterns bears little resemblance to the printed lines and spaces; the mask features are in fact two dimensional. However, Figure 27(b) shows that with the Figure 26(a) source the image modulation is entirely 1D. [For clarity, Figure 26(b) shows 2×2 periods of the 2:1 mask; however, the amplitudes of the two bright fringes in the associated Figure 27(b) image have the same sign, i.e., the mask is not alternating, nor is the Figure 26(c) mask for the 1:1 line/spaces.] One drawback to the solution should be noted: While a peak intensity of 41% is obtained with the chromeless mask shown, peak intensity is only 6% when implemented in an attenuated phase shift mask (of 6.5% background transmission, as in our previous examples). Peak intensity with the dipole and annular solutions are 12% and 28%, respectively.

Also, we found a somewhat stronger tradeoff than usual between DOF and exposure latitude in-focus, hence, our constraint that exposure latitude be at least 10% in the focused image. (Thus, process windows above 13.6% μm can be achieved at the cost of lower in-focus latitude.)

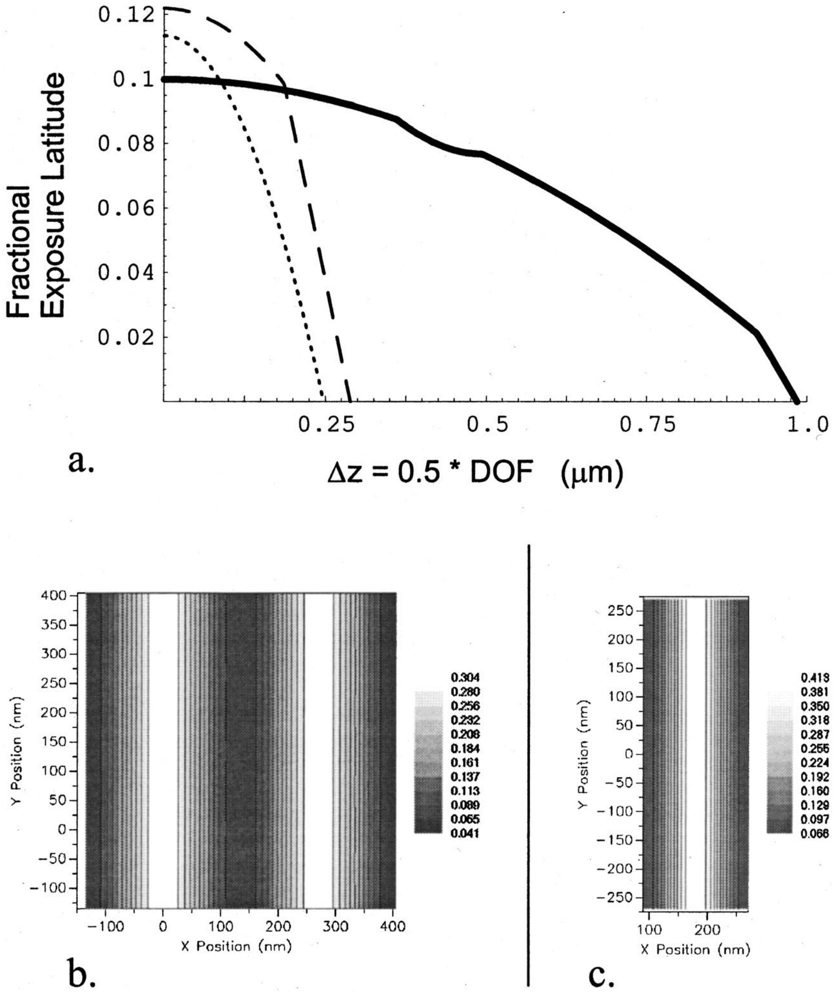

Fig. 27 (a) Comparison of common process window achieved by: Figure 26 solution (thick line, 13.6% μm window); a conventional solution using attenuated phase shift, annular illumination, feature bias, and assists, all optimized (dotted line, 3.6% μm window); and an aggressive conventional solution using attenuated phase shift, dipole source, gray-tone masks, feature bias, and assists, all optimized (dashed line, 5.5% μm window). (b) 2:1 line/space image provided by the Figure 26(b) mask with the Figure 26(a) source. Though the mask has a 2D structure, the image fringes show only a 1D modulation. The image region shown has a width of two periods, and the amplitude in adjacent bright spaces has the same sign, i.e., the mask is nonalternating. (c) 1:1 line/space image provided by the Figure 26(c) mask (also nonalternating) and the Figure 26(a) source. Images (b) and (c) are from Prolith (Ref. 7) screen captures.

Notwithstanding these limitations, it is quite encouraging that our algorithm can provide a 2.5× performance improvement over aggressive conventional solutions, even with patterns that have been the subject of very intensive prior study in the literature.

Acknowledgments

The authors have received help and support from many colleagues in the course of this work. Particular thanks are due to Tim Brunner, Donis Flagello and co-workers at ASML, Chris Progler, Nakgeuon Seong, and Tim Wiltshire.

References

1. J. Gortych and A. E. Rosenbluth, ‘‘Method and system for optimizing illumination in an optical photolithography projection imaging system,’’ US Patent No. 5,680,588 (1997).
2. A. E. Rosenbluth, T. Brunner, and D. G. Flagello, ‘‘Optical lithography as the NA=1 limit approaches,’’ in 1992 Optical Society of America Annual Meeting, Albuquerque, NM (1992), paper MVV-3.
3. M. Burkhardt, A. Yen, C. Progler, and G. Wells, ‘‘Illuminator design for the printing of regular contact patterns,’’ *Microelectron. Eng.* **41**, 91 (1998).
4. T.-S. Gau, R.-G. Liu, C.-K. Chen, C.-M. Lai, F.-J. Liang, and C. C. Hsia, ‘‘The customized illumination aperture filter for low k1 photolithography process,’’ *Proc. SPIE* **4000**, 271 (2000).
5. A. Wong, R. Ferguson, S. Mansfield, A. Molless, D. Samuels, R. Schuster, and A. Thomas, ‘‘Level-specific lithography optimization for 1-Gb DRAM,’’ *IEEE Trans. Semicond. Manuf.* **13**(1), 76–87 (2000).
6. S. Bukofsky, J. Mandelman, A. Thomas, C. Radens, and G. Kunkel, ‘‘A lithographically-friendly 6F^2 DRAM cell,’’ in *Proc. VLSI Technology, Systems, and Applications* (2001), p. 97.
7. See product description for Prolith at Finle website, http://www.finle.com.
8. M. D. Himel, R. Hutchins, and M. Poutous, ‘‘Design and fabrication of customized illumination patterns for low k1 lithography: a diffractive approach,’’ *Proc. SPIE* **4346**, 1436 (2001).
9. D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA (1995).
10. Matlab Optimization Toolbox, http://www.mathworks.com/products/optimization/8513v03_optim.pdf.
11. S. Wolfram, *The Mathematica Book*, 4th ed. Cambridge University Press, Cambridge (1999).
12. IMSL Online Documentation (Visual Numerics Inc.), http://www.vni.com/products/imsl/index.html.
13. S. A. Vavasis, ‘‘Complexity issues,’’ in *Handbook of Global Optimi-*

zation, Reiner Horst and Panos M. Pardalos, Eds., Kluwer Academic, Dordrecht (1995), p. 27.

14. A. I. Barros, J. B. G. Frenk, S. Schaible, and S. Zhang, ''A new algorithm for generalized fractional programs,'' *Math. Program.* **72**, 147–175 (1996).
15. A. Morgan, *Solving Polynomial Systems Using Continuation for Engineering and Scientific Problems*, Prentice-Hall, Englewood Cliffs, NJ (1987).
16. G. Hadley, *Linear Algebra*, Addison-Wesley, New York (1961).
17. D. Williamson, J. McClay, K. Andresen, G. Gallatin, M. Himel, J. Ivaldi, C. Mason, A. McCullough, C. Otis, J. Shamaly, and C. Tomczyk, ''Micrascan III, a 0.25 micron step and scan system,'' *Proc. SPIE* **2726**, 780–786 (1996).
18. J. Fung Chen (private communication).

Alan E. Rosenbluth, IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (e-mail: aerosen@us.ibm.com). Dr. Rosenbluth received the Ph.D. degree in optics from the University of Rochester in 1982 as a Hertz Foundation Fellow. In the same year he joined the IBM Thomas J. Watson Research Center. His principal research activities have been in soft x-ray optics, photolithography, display technology, and thin-film design. He is the author of many technical papers in these fields, and holds 26 US patents.