REVIEW

# Inverse lithography technology: 30 years from concept to practical, full-chip reality

**Linyong (Leo) Pang***
D2S, Inc., San Jose, California, United States

**Abstract.** In lithography, optical proximity and process bias/effects need to be corrected to achieve the best wafer print. Efforts to correct for these effects started with a simple bias, adding a hammer head in line-ends to prevent line-end shortening. This first-generation correction was called rule-based optical proximity correction (OPC). Then, as chip feature sizes continued to shrink, OPC became more complicated and evolved to a model-based approach. Some extra patterns were added to masks, to improve the wafer process window, a measure of resilience to manufacturing variation. Around this time, the concept of inverse lithography technology (ILT), a mathematically rigorous inverse approach that determines the mask shapes that will produce the desired on-wafer results, was introduced. ILT has been explored and developed over the last three decades as the next generation of OPC, promising a solution to several challenges of advanced-node lithography, whether optical or extreme ultraviolet (EUV). Today, both OPC and ILT are part of an arsenal of lithography technologies called resolution enhancement technologies. Since OPC and ILT both involve computation, they are also considered as part of computational lithography. We explore the background and history of ILT and detail the significant milestones that have taken full-chip ILT from an academic concept to a practical production reality.  [DOI: 10.1117/1.JMM.20.3.030901]

**Keywords:** inverse lithography technology; computational lithography; lithography; resolution enhancement technologies; curvilinear ILT; multibeam mask writer; VSB mask writer; photomask; graphics-processing unit.

Paper 21008V received Feb. 3, 2021; accepted for publication Jul. 26, 2021; published online Aug. 31, 2021.

## 1 Introduction: The Promise of ILT

In lithography, optical proximity and process bias/effects need to be corrected to achieve the best wafer print. Efforts to correct for these effects started with a simple bias, adding a hammer head in line-ends to prevent line-end shortening. This first-generation correction was called rule-based optical proximity correction (OPC). Then, as chip feature sizes continued to shrink, OPC became more complicated and evolved to a model-based approach. Some extra patterns were added to masks, to improve the wafer process window, a measure of resilience to manufacturing variation. Around this time, the concept of inverse lithography technology (ILT)—a mathematically rigorous inverse approach that determines the mask shapes that will produce the desired on-wafer results—was introduced. ILT has been explored and developed over the last three decades as the next generation of OPC, promising a solution to several challenges of advanced-node lithography, whether optical or extreme ultraviolet (EUV). Today, both OPC and ILT are part of an arsenal of lithography technologies called resolution enhancement technologies (RET). Since OPC and ILT both involve computation, they are also considered part of computational lithography.

The seeds for ILT were sown as early as the 1980s, with academic work.[1,2] By the 1990s, industrial research and development teams had started to work on ILT technologies that could be used in production,[3–8] with the first ILT product introduction in 2005/2006,[9–17] when the author first coined the term ILT. At this time, dry-process 45-nm lithography was the leading-edge target and was very challenging in terms of process window. The adoption of immersion

*Address all correspondence to Linyong (Leo) Pang, leo@design2silicon.com

lithography at 45 nm greatly enlarged the process window and delayed the immediate need for ILT, giving researchers more time to develop the technology. For the next two decades, both academia and industry continued to develop and refine ILT solutions, as the need for improvement in wafer process window became more pressing with each smaller process node.

Over this time, there have been numerous studies that demonstrate that curvilinear ILT mask shapes produce the best process windows.[18] However, there have been three significant roadblocks to broad production application of ILT. The first roadblock was that runtimes associated with this computational technique have limited its practical application to critical "hotspots" on chips.[19,20] The second roadblock has been that although ILT naturally produces curvilinear shapes, the variable-shaped beam (VSB) mask writers used to write the vast majority of production masks use rectilinear shapes to create mask shapes. While it has been possible to use very small rectilinear shapes to approximate curvilinear shapes, the process is time-consuming and expensive, as it requires many more VSB shots. Lastly, there was the roadblock of creating complex ILT masks that met mask-manufacturing rules, although admittedly, this had been more in the background, because the first two roadblocks had held quite firmly.

At this point, one might expect that this technology would fade away. However, the fact that ILT has been used throughout this time for hotspots suggests that there is general understanding that ILT produces superior process windows for the wafer. Despite these roadblocks, R&D teams in both academia and industry have continued to pursue ILT as an answer to advanced lithography woes for the simple reason that, even from the first, it has produced impressive improvements in wafer process window that have not been matched by any of the other advancements in RET.

Thankfully, the advent of two new technologies—general-purpose graphics-processing unit (GPGPU) computation and multibeam mask writers—paved the way in the past few years for some breakthroughs in these ILT roadblocks. In 2019, an entirely new approach, systematically designed for multibeam mask writers and GPU acceleration, made full-chip ILT a practical reality in production for the first time.[21] This new approach produced wafer results that confirmed a 100% improvement of the wafer process window versus OPC. Subsequent work in 2020, using a mask-wafer co-optimization (MWCO) technique, expanded these benefits and practical runtimes to ILT masks written by VSB mask writers.[22,23]

This paper will explore the background and history of ILT and detail the significant milestones that have taken full-chip ILT from an academic concept to a practical production reality.

## 2 Overview of Inverse Lithography Technology

ILT has been defined as follows: given a known forward transformation from mask patterns to images for a specified lithography process, compute an optimized mask that produces the desired wafer target with best pattern fidelity and/or largest process window.

By its nature, the optimized solution is not limited to simple heuristic modifications of the target pattern; in other words, it can explore regions of solution space that are very different from the original pattern.

To formulate the problem, we define the following mathematical functions and operators:[15]

Mask function: $\psi$

Target pattern: $\Phi$

Forward operator: $f$

Wafer pattern: $\omega$

The forward operator covers all elements of the transformation from mask to wafer: for example, electromagnetics of the three-dimensional (3D) mask, optics of illumination and the projection lens, behavior of the photo resist, dose and focus conditions, aberrations, etc. Thus,

$$\omega = f(\psi),$$

and we seek to find

$$\psi^* = f^{-1}(\Phi),$$

where $\psi^*$ is an optimal mask function.

The problem thus stated is ill-posed, however; because the forward operator $f$ is many-to-one (that is, many different masks will yield identical on-wafer results), the function has no well-defined inverse. Moreover, for typical target patterns $\Phi$ (e.g., a drawn layout with Manhattan geometry and sharp corners), there does not exist any mask function $\psi$ for which $\Phi = f(\psi)$.

These issues are addressed by recasting the inverse problem as an optimization problem. Optimization problems seek to find a solution as close to optimal as practical within the constraints of a reasonable computational time. Problems cast as optimization problems always find some answer, even if it is an inadequate answer. The degree of non-uniqueness is particularly high in lithography problems, because we are only interested in resist contours, discarding most of the information that is present in the corresponding grayscale images. An infinite number of mask solutions can produce resist contours with identical approximations to Manhattan targets.

We define a merit function, also called a cost function, energy function, or Hamiltonian (by analogy to quantum mechanics, where it would be an operator corresponding to the total energy of a system), and label it $H(\psi)$. This function indicates the quality of the solution or the "goodness" of the mask. A simple example would be

$$H = \iint |f(\psi) - \Phi|.$$

This Hamiltonian is the absolute value of the difference between the wafer image and the target pattern, integrated over the area of the region of the image. In practice, a number of additional elements may be included in the Hamiltonian, for example, the images at various operating conditions throughout the process window (i.e., over- or under-exposed and out of focus), normalized image log-slope of the image, robustness against mask error enhancement factor, or other factors as deemed appropriate. The actual functional form is more complicated than suggested by these simple relations. Elements that are not directly related to lithography may also be included. For example, simple masks may be preferred over complex masks, and terms to this effect may be included in the Hamiltonian. What is essential is that the Hamiltonian is a function of the mask function and that minimizing the Hamiltonian allows us to find an optimal mask according to the criteria we have chosen. In addition, a variety of constraints are imposed by the realities of mask manufacturing; for example, two disjoint chrome regions must be separated by a minimum distance, and a chrome line must have a minimum width. We address these constraints by defining a subspace of the full solution space of mask functions and restricting our solution to this subspace.

Even with restrictions for the sake of optimization, however, the key distinctive features of ILT are the absence of pattern-dependent heuristics and the ability to broadly explore wide areas of the solution space. This means that ILT algorithms frequently lead to mask patterns that are unanticipated by a knowledgeable lithographer. One example is the problem of placement of subresolution assist features (SRAFs). In the past, these were placed empirically, with great care, and frozen in place during the computation of the rest of the mask. In contrast, ILT can determine optimal SRAFs simultaneously with the rest of the mask.

One common misconception is that ILT always results in a unique global optimum. Currently, inverse lithography approaches are based on local search heuristics that find a solution close to a local minimum, even if it may not be globally optimum. Moreover, because the inverse problem often has several solutions that are nearly optimal, the algorithm (or algorithm designer) must determine which of many good solutions to select. Usually once all constraints are considered, such as that the features must be bigger than a certain size to meet mask manufacturing requirements and that we want the largest depth of focus (DoF), yet the SRAFs cannot print, solutions in ILT that meet the constraints end up being quite unique, at least from an optimization algorithm's gradient descent point of view.

Another common misconception is that ILT cannot handle the resist development and etching process modeling for which we only have empirical models valid over a limited range. ILT does not require a solution to be in closed form. Since ILT is cast as an iterative optimization problem, as long as the model for the forward transformation is known, ILT can find inverse solutions for lithography that cover resist development and etching processes in addition to optics.

Most ILT approaches optimize a cost function similar to the Hamiltonian shown earlier. The key is how to make such optimization finish in a reasonable time (in the order of hours or days) for a full-chip design, so it can be used in semiconductor manufacturing. In other words, the key is how to simplify the calculation by making certain approximations or reducing the number of variables.

Figures 1 and 2 show one implementation of such an optimization.[21] Figure 1 shows the mask pattern, its simulated wafer contour, cost function, and cost gradient at the beginning of the optimization. It is clear that the wafer contour does not hit the wafer target, the cost function is not zero, and the cost gradient is not flat. Figure 2 shows the situation at the end of the ILT optimization. Now, the simulated wafer contour hits the wafer target, the cost function approaches zero, and gradient of the cost function is flat.

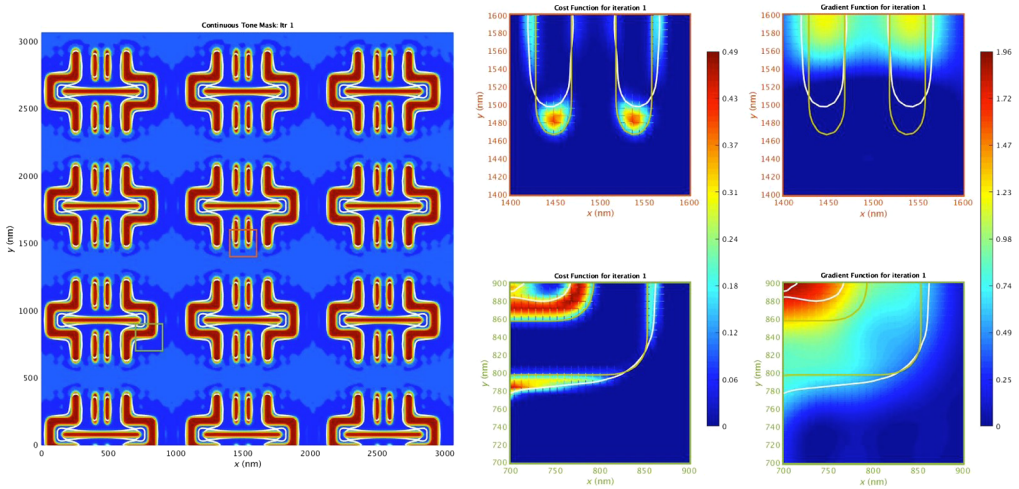


**Fig. 1** Mask pattern, simulated wafer contour and its target, cost function, and cost gradient at the beginning of the ILT optimization[21] (source: D2S).

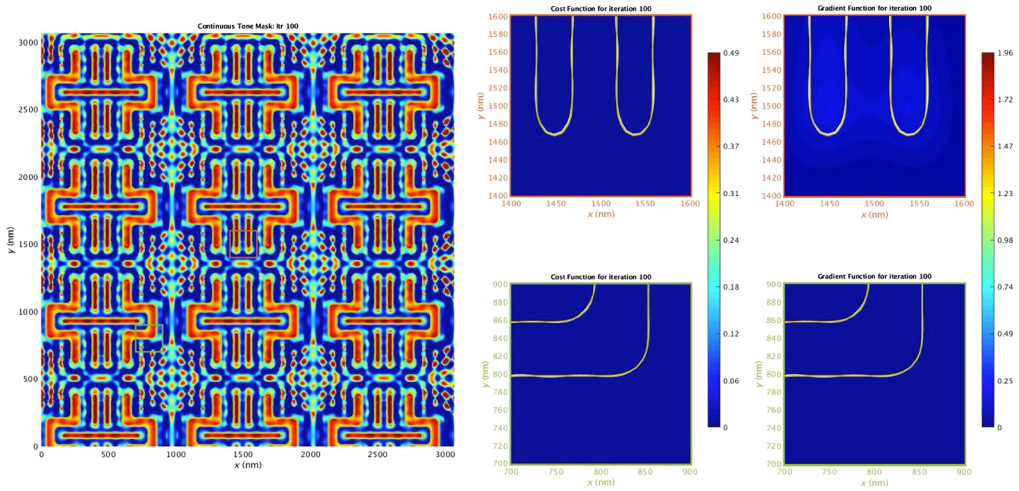


**Fig. 2** Mask pattern, simulated wafer contour and its target, cost function, and cost gradient at the end of the ILT optimization[21] (source: D2S).

## 3 History of Inverse Lithography

ILT was first proposed by B. E. A. Saleh and others at the University of Wisconsin-Madison. In 1981, Saleh and Sayegh[1] found optimized photomasks by a variation on simulated annealing or "pixel flipping." They started with an initial guess and randomly flipped individual pixels, accepting changes that improved the quality of the solution and rejecting changes that degraded it, and repeated this process until the solution converged on an optimal photomask. A few years later, Saleh and Nashold[2] described an algorithm using a sequence of projection operators to find a band-limited function, corresponding to a continuous-tone or gray-scale mask that optimized the desired image. Later, the same authors used a similar approach to find complex-valued functions that corresponded to continuous-tone phase masks.

In the early 1990s, Yong Liu and Avideh Zakhor at the University of California (UC) at Berkeley published a series of papers,[3,4] describing various approaches to ILT. In one case, they used branch-and-bound and the simplex method. In another, they used what they called a "bacteria" algorithm to satisfy mask constraints.

In 2001, Rosenbluth et al. at IBM described a source-mask optimization (SMO) algorithm that analyzed diffraction orders in an effort to jointly optimize the photomask and the stepper illumination.[5] In this approach, they first determined an optimum diffraction spectrum and then computed an ILT mask pattern to produce it.

Many other researchers made significant contributions to the development of ILT. These include Wang et al.[6] (first at Stanford University, and later at Numerical Technologies) and Jang et al. (at Wonkwang University in South Korea),[7] who developed the OPERA program. In addition, Fuhner and Erdmann[8] of the Fraunhofer Institute developed ILT using genetic algorithms.

As mentioned previously, these early approaches to ILT often produced solutions with good lithographic quality in simulation, as indicated by pattern accuracy and larger process windows. However, most of the methods required unreasonably long computation, so these were theoretical improvements that turned out to be impractical. For example, finding the optimal continuous-tone or grayscale mask is an easier mathematical problem than finding an optimal binary mask. However, only a binary mask or phase-shifting masks (PSMs) are practical for current production.

It was Intel that saw the potential of inverse lithography and sponsored a number of universities working on research in ILT, particularly methods using pixilated masks. This prompted a second wave of ILT development and initial commercialization.

The first push to commercialize ILT into real semiconductor manufacturing was started by Luminescent Technologies, Inc., in 2003. The key algorithm was based on established level-set methods,[24] invented by Osher (a Luminescent cofounder) and Sethian. Luminescent announced an ILT product at the 2005 Photomask Technology Conference. Six papers were presented by Luminescent[9,10] and its partners and customers, including UMC and Xilinx,[11] Cypress,[12] SMIC,[13] and Photronics.[14] The author of this ILT review paper, at that time working for Luminescent, was the first to formally name this method "inverse lithography technology" or "ILT," an acronym now universally used by the semiconductor industry. At SPIE Microlithography 2006, Luminescent CTO Dan Abrams and the author presented the milestone paper "Fast Inverse Lithography Technology,"[15] along with other joint papers with its customers.[9,16,17]

The Luminescent research and development on ILT attracted attention in the electronic design automation (EDA) and semiconductor industries, as well as in universities. For example, Yuri Granik from Mentor Graphics presented a pixel-based ILT development and applied it to place SRAFs.[25,26]

Intel continued to sponsor academic ILT research and has been a powerhouse of ILT development. Yan Borodovsky, then the Intel senior director on lithography, first showed their pixel-based, random, chromeless PSM at the Lithography Workshop in 2007.[27] The next year at SPIE Advanced Lithography 2008, Intel presented four papers on this subject, given by Borodovsky,[28] Singh et al,[29] Cheng et al.,[30] and Schenker et al.,[31] covering the subjects of modeling and computation, mask making and inspection, and integrating the technology to fabricate a working chip.

Gauda, another startup, also made a great contribution to ILT. They began their work in this area by working on OPC using GPU acceleration.[32] Later, they invented a new approach that solves the ILT problem in the frequency domain, in contrast to the Luminescent level-set method, which solves in the real domain.[33]

Amyn Poonawala was one of the PhD students at UC Santa Cruz sponsored by the Intel program. He developed an optimization framework for inverse lithography based on a pixel-based, continuous-function formulation, well-suited for the gradient-based search algorithms.[34–36]

While still at Luminescent, the author gave many talks at various conferences in China[37–41] that also stimulated ILT research there. Many papers were published by Yang and Shen from Zhejiang University.[42,43] They developed a pixel-based gradient approach to solve the inverse lithography problem. Another research group from Tsinghua University also developed an approach that does not depend on initial conditions[44] and explored hardware-accelerated ILT using GPUs.[45]

In addition to UC Santa Cruz and the two top research universities in China, ILT research also progressed in many other universities worldwide. Professor Edmund Lam and his students—in particular, Ningning Jia—from Hong Kong University did an extensive study in ILT,[46] particularly regarding the regularization of ILT mask solutions to meet mask manufacturing requirements.[47] He proposed automatic optimization of the mask and illuminator with a genetic algorithm. Ningning Jia was also the first to apply machine learning (ML) to ILT.[48] Xu Ma and Gonzalo R. Arce at Delaware University developed generalized, gradient-based RET optimization methods to solve the inverse lithography problem, for which the solution is not constrained to a finite-phase tessellation, but rather is found by arbitrary search trajectories in a complex space.[49] They extended this framework to solve the inverse lithography problem with partially coherent sources and binary[50] or attenuated PSMs.[51] Shanhu Shen in Professor David Pan's group in the University of Texas at Austin also presented work on ILT using two-dimensional (2D) discrete cosine transform of the target mask, for which the low-frequency components are used in the optimization. This method produced optimal patterns, similar in shape to those obtained with the level-set method.[52] Jue-Chin Yu from the National Jiaotong University in Taiwan also worked in this area and showed SRAF generation using ILT.[53]

By 2010, three companies had demonstrated and published the use of ILT to correct full-chip designs: Luminescent[54–58] (later acquired by Synopsys, Inc.), which employed the level-set method of ILT optimization, Intel[30] using pixelated PSM mask, and Gauda[33] (later acquired by D2S, Inc.), which presented a GPU-accelerated approach using a cost-function in the frequency domain. Other semiconductor manufacturing companies produced full-chip correction using ILT but never published their results. Later, ILT was combined with source optimization to gain further lithography improvement.[59–66]

As Luminescent was trying to drive ILT into production, it worked with its customers and partners and published numerous papers, many by the author, to demonstrate wafer process window benefits, to address mask-making-related issues, and to explore applications of ILT, such as design rule optimization.[10,18,67–97]

After the Luminescent ILT business was acquired by Synopsys in 2012, the rest of the decade was a little quiet in terms of new ILT development and publications. However, today all EDA companies that offer OPC products also offer ILT products of some kind, some used for fixing hotspots like Synopsys,[19,20,98] some for model-based SRAF generation like ASML Brion.[99–101] ILT was extended to EUV by Synopsys[102–105] and ASML Brion started exploring using deep learning (DL) in ILT for SRAF generation.[99–101] Despite steady, continuing research and development across academia and industry through the decade and demonstration of the use of ILT to correct full-chip designs, ILT was still seen as an advanced method for use in critical hotspots, rather than as a technique to be applied to full-chip mask generation. Excessive computational run-times continued to render full-chip ILT impractical in production settings.

Photomask industry, on the other hand, started to focus on ILT infrastructure. Luminescent continued working on computational inspection and metrology technologies to enable mask inspection, mask review, and mask repair ready for ILT masks, until it was acquired by KLA in 2012, and TSMC EBO presented a number of papers together with Luminescent showing such capabilities deployed in production.[92–96,106–114] ILT was the topic of panel

discussion at the SPIE Photomask Technology Conference for two consecutive years in 2015 and 2016. The industry also recognized the issue of using VSB mask writers to write curvilinear ILT masks. Both IMS[115] and NuFlare[116] developed multibeam mask writers that have shape-agnostic write times enabling production of curvilinear ILT masks.

At the 2019 SPIE Photomask Technology Conference, D2S, Inc. presented an entirely new, stitchless approach, described as "an extreme SIMD approach." This purpose-built ILT system includes a unique GPU-accelerated approach that emulates a single, giant GPU/CPU pair that can compute an entire full-chip ILT solution at once. This system could produce a full-chip ILT mask solution within a practical run-time—a day or two. The mask and wafer results demonstrated for standard memories showed the system produced continuous and symmetric masks that met all edge-placement error (EPE) requirements and yielded superior lithographic results with a process-window increase of over 100%. Full-chip, curvilinear ILT had finally become a practical reality.[21]

The 2019 paper utilized newly introduced multibeam mask-writing technology. In 2020, D2S presented an MWCO technique that enabled the ILT computation approach from the 2019 paper to be applied to a memory mask written by a VSB writer. The mask was written with a practical 12-h write-time and demonstrated similar mask quality and process-window expansion benefits.[22,23]

In the next section, we will take a more detailed look at the roadblocks encountered by these researchers and why it took more than 30 years of effort to find a practical, full-chip ILT solution.

## 4 Roadblocks to Wide Adoption of ILT

### 4.1 *Primary ILT Roadblock: Full-Chip ILT Runtime*

ILT computations themselves are extremely complex. The computation runtime to generate an ideal ILT solution is an order of magnitude longer than traditional OPC due to the larger solution space of ILT. There are three main components to a practical full-chip ILT solution. The first component is forward simulation with a lithographic model; the second component is the optimization of the ILT solution (time required for each iteration and number of iterations); the third component is how the full-chip solution is executed. Over the years, varied approaches have been taken to each of these aspects in an attempt to overcome this primary roadblock to the wide production adoption of ILT. We will discuss some of these approaches in Sec. 5. However, whichever approach(es) to forward simulation and optimization are taken, there are challenges associated with implementation of a full-chip solution.

Many, if not all, full-chip EDA tools, including full-chip OPC, solve full-chip runtime problems through a divide-and-conquer approach. This approach splits the full-chip into many small areas, called partitions, and feeds each partition to a CPU for computation. Each partition is run through a process, such as OPC, from the beginning to the end. Then, all the partitions are stitched together. Each partition includes overlaps with the surrounding areas, called halos, so the patterns from the borders of adjacent partitions are included.

Whether or not a desired wafer pattern prints as desired is affected by the physics of light by the features surrounding it. For 193i lithography, this effect is significant for features adjacent to a given feature, the features adjacent to those adjacent features, and so on, with a reach equivalent to multiple standard-cell rows. For EUV, the ambit is smaller, but still the effect is significant for multiple features beyond the nearest neighbors. OPC, having been invented in earlier days when the features were bigger compared with the wavelength of the light, is a local, minor modification of the target pattern, so the OPCed patterns on the partition boundary calculated from neighboring partitions are very close to the target pattern, so there are usually no big issues after stitching. When SRAFs are added to OPC, the SRAFs are generated first and fixed in place, therefore avoiding stitching errors at the boundary.

However, unlike OPC, which only makes small, local modifications to the target patterns, ILT mask patterns can be dramatically different from the target pattern, especially the SRAFs, because in ILT the main features and SRAFs are all computed/optimized at the same time. Even with a halo, the SRAFs produced for a given partition by end of the ILT optimization can be very

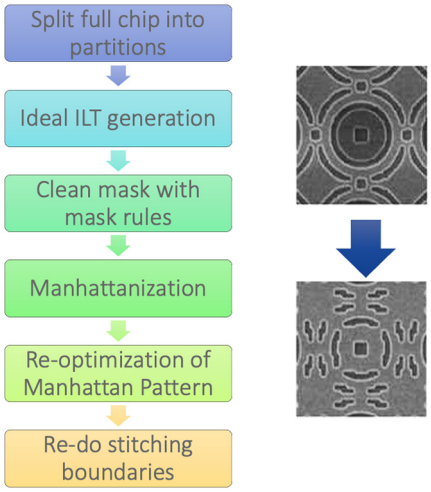


**Fig. 3** In the conventional ILT flow targeting a VSB mask writer, the extra steps are required to create a full-chip solution[22] (SEM Source: Samsung).

different from those produced for a neighboring partition, causing errors once the full-chip design is stitched back together. These "stitching errors" are what caused the conventional divide-and-conquer full-chip approach successfully used for OPC to break down when applied to full-chip ILT.

To illustrate the "total ILT runtime" challenge, Fig. 3 shows the conventional flow used to produce full-chip ILT masks. First, the chip is partitioned. Each partition is passed to a CPU to calculate the ideal ILT solution (which takes an order of magnitude longer than conventional OPC already). Then, the ideal ILT mask solution is cleaned up to meet mask rules. Next, the mask-rule-clean design, which is naturally curvilinear, is modified so it can be approximated with the rectilinear shapes created by VSB mask writers, a process called "Manhattanization." The mask shapes are dramatically changed in this step, so a reoptimization is required to ensure the new Manhattan mask pattern will meet the wafer pattern accuracy requirement and process window requirements. Once all partitions have gone through these processes, the partitions are "stitched" back together.

Now, the partitions must be verified to catch any stitching errors. The most common method to correct stitching errors is to recalculate the ILT solution for regions around the partition boundary after merging, plus some buffer region, and then restitch the partitions back together. While this method will fix the existing stitching errors, it may introduce new stitching errors at the new boundaries. In addition, because the partition size that can be handled by each CPU is relatively small and the buffer region necessary to account for optical proximity effects is relatively large, these recalculated areas are close to the size of the original partition. In a real implementation, dealing with stitching errors has been known to cause the run time to almost double.

At the end, the total runtime for the entire ILT flow is an order of magnitude slower than the generation of ideal ILT, which is already an order of magnitude slower than OPC. As a result, the commercial applications of ILT have been limited, focused on smaller, high-risk portions of chips, and have been mainly used in hot-spot correction mode. A high-volume, full-chip ILT solution has been elusive.

### 4.2 *Related Roadblock: VSB Mask Write Time*

Along with computational runtime, mask write time using VSB mask writers has been a major hurdle for wide adoption of ILT. As discussed earlier, ILT naturally creates curvilinear mask shapes, which must be converted, or Manhattanized, for writing by a VSB mask writer. The conventional approach to Manhattanizing curvilinear ILT masks requires a trade-off of accuracy for ILT runtime and the write time on the VSB mask writer. It is possible to get very close to a curvilinear target using many small rectilinear mask shapes to form curves with small "jogs" or "stair-steps." This approach creates fairly good curvilinear mask shapes using VSB writers.

However, the shot count involved in this approach would lead to impractical write-times if it were to be used on a full-chip ILT design. Alternatively, jog and step-sizes can be made larger, say 20 nm, to limit VSB shot count, but even then, if practiced at a full-chip scale, write times would be prohibitive using conventional fracturing without overlapping shots.[117] This practical reality, combined with the long runtimes of the traditional ILT software, even when running on a large bank of CPUs, has limited the use of ILT to small hotspots.

### 4.3 *Additional Worry: Mask Manufacturability*

Initially, some of ILT's advantage over conventional OPC in terms of larger wafer process window was canceled by the increase in mask variation due to the more complex mask shapes. Since mask aberration transfers as a systemic error to every chip, a mask that is more resilient to mask manufacturing variation is important. This was particularly true when curvilinear masks shapes, a natural output for ILT, were attempted using VSB writers. In addition to being inconvenient and expensive, masks that take a long time to write are naturally more susceptible to manufacturing variation. In addition, writing curvilinear shapes as a sequence of extremely small rectangles to mimic a curve is difficult to do accurately. So, both local critical dimension (CD) accuracy and global CD accuracy suffered in experiments that tried to write curvilinear ILT mask shapes with VSB writers.

Since the introduction of the first ILT product in 2005/2006, most companies that offer OPC/RET-related products offer some kind of solution for ILT, most focused on hotspots because of the roadblocks just discussed. In the next section, we briefly review the basic arc of ILT product development through a look at three varied approaches to ILT and how they tried to address these roadblocks.

## 5 ILT Product Development

One might say that the path of ILT product development has been less about path-finding and more about path-clearing. The roadblocks to wide adoption discussed in the last section needed to be cleared before any particular approach to ILT could hope to gain widespread application. Each of the three product development efforts reviewed later focused on streamlining ILT calculations in some way, addressing the primary runtime roadblock. In addition, both the Luminescent and the D2S efforts have sought to address the mask write time roadblock as well.

### 5.1 *Luminescent Level-Set Method*

The Luminescent ILT, which was the pioneering product first introduced in 2005/2006,[9–17] tried to solve the ILT runtime problem by reducing the number of variables using the level-set method. It also provides a mathematically elegant method for solving topology discontinuity during ILT optimization.

The level-set method is a branch of applied mathematics that was invented by Professor Stan Osher (Luminescent's co-founder) and James Sethian in 1980s.[24] It has been applied in many engineering fields and is regarded as one of the most efficient mathematical methods for solving problems involving dynamic change of 2D patterns with topology changes. ILT based on the level-set method was developed by Luminescent starting in 2003 to improve mask optimization efficiency and reduce complexity, and thus runtimes. Later, the same mathematical framework has also been applied to source optimization and SMO.[59–63]

The enabling technology in this form of ILT is the level-set representation of the design, mask, and wafer patterns. Representing the 2D design pattern, mask pattern, and wafer pattern by level sets is mathematically efficient and gives the mask patterns practically infinite degrees of freedom to change during the optimization. It also allows SRAFs to develop continuously to enable printing of wafer patterns with better CD uniformity and larger process margin.

In the level-set approach, the idea is to solve the ILT optimization problem in a higher dimension. One can take the original 2D curve and build it into a 3D surface by adding a mathematical function in the third dimension. For example, as shown in Fig. 4, one can define a distance

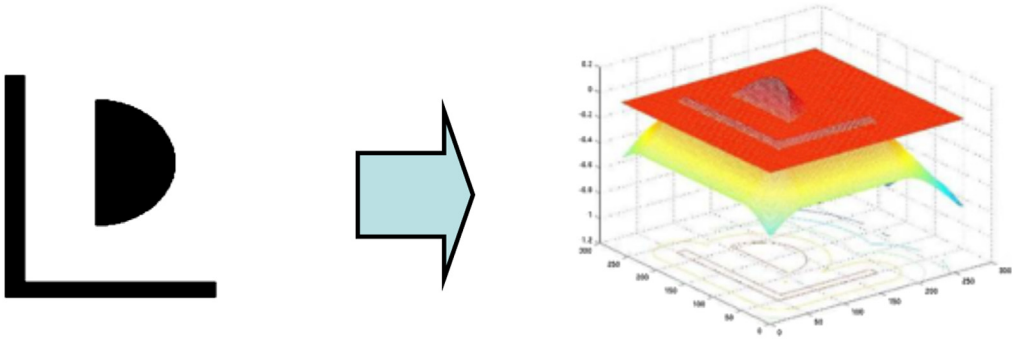

**Fig. 4** How to represent a 2D mask pattern by level set[68] (source: Luminescent/Synopsys).

function where the value $(z)$ of this function for any point $(x, y)$ is the shortest distance from the point to the nearest edge of the 2D curve. The surface representation will intersect the $xy$ plane in a 2D shape, and the 2D shape on the $xy$ plane (zero level set) is actually the original 2D curve.

This representation has several advantages. First, it will always be easy to track the surface as it evolves. Second, the curve may get wildly contorted, but the surface always remains well-behaved. Third, the complicated problems of breaking and merging are easily handled. All of these concepts also apply to higher dimensions. Fourth, it is easy to build accurate numerical schemes to approximate the equations of motion. Rather than tracking buoys that might end up colliding, the answer can be computed from a fixed point of reference on the $xy$ plane.

When adapted for numerical computing, the level-set representation is equivalent to a gray-scale pixel array, with each pixel's gray value corresponding to that of the level-set function.

Figure 5 shows how a mask pattern (a contact hole in this case) can be represented by a level set and how its shape and topology changes.[68] SRAFs are created from step (b) to step (c) by raising the surface around the main contact hole; when the surface passes through the plane corresponding to a level of zero (the $xy$ plane), the SRAFs appear. In 2D, such a change would introduce topology changes that make the function discontinuous. However, in the level-set representation (the 3D surface), it is still continuous. This makes the optimization easier to handle by numerical algorithms.

The Osher-Sethian level-set method tracks the motion of an interface by embedding the interface as the zero-level set of the signed distance function. The motion of the interface is matched with the zero-level set of the level set function, and the resulting initial value partial differential equation for the evolution of the level set function resembles a Hamilton–Jacobi equation.[24]

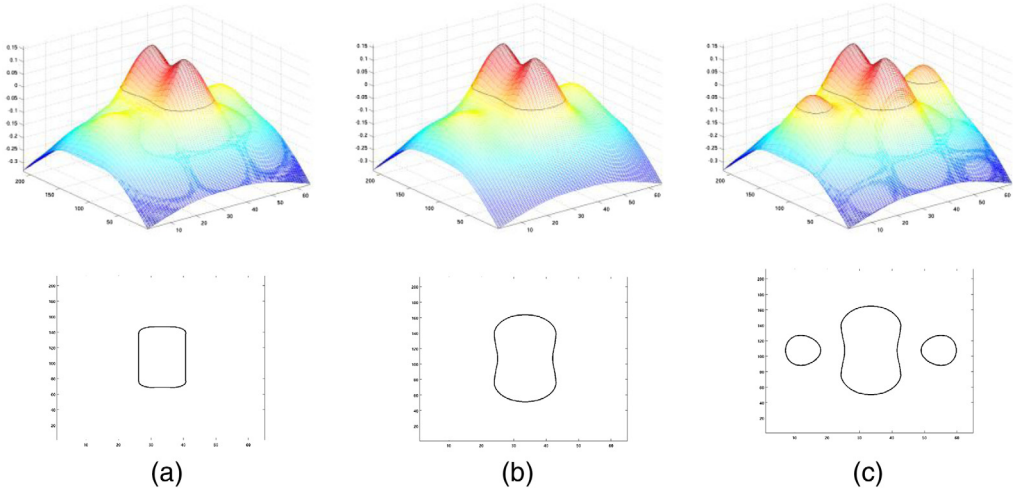


**Fig. 5** Illustration of a level-set representation of a mask during optimization. 2D topology changes, such as SRAF creation, correspond to changes of a continuous level-set function of two variables[68] (source: Luminescent/Synopsys).

**Fig. 6** ILT mask patterns shown in the 2005 Luminescent ILT shows the curvilinear ILT mask patterns and the Manhattanized mask patterns[15] (source: Luminescent/Synopsys).

In this setting, curvatures and normals may be easily evaluated, topological changes occur in a natural manner, and the technique extends trivially to three dimensions. This equation is solved using entropy-satisfying schemes borrowed from the numerical solution of similar equations that arise in numerous physical problems.

Once the mask patterns are represented by the zero-level set, the ILT optimization can be formulated as general, multiple-variable optimization problems and solved using standard optimization algorithms, such as the conjugate gradient method. The goal of the optimization is to minimize a cost function. The challenges of such an optimization problem have to do mainly with the scale of the optimization.

The Luminescent level-set method only addresses the optimization, not the lithography simulation. The level-set method also makes the mask contours continuous and tends to generate curvilinear ILT mask patterns (Fig. 6). In addition, because the level-set method is a real domain representation, it does not guarantee the pattern symmetry.

Figure 7 shows the ILT mask pattern design, the actual ILT mask pattern, and the corresponding wafer print for an static random-access memory (SRAM) published in the original Luminescent paper.[15]

Luminescent published many wafer results with multiple semiconductor companies between 2005 and 2010, a time when the dry 45-nm technology node was the leading-edge technology in development and very challenging in terms of lithography. Figure 8 shows the wafer results for a 45-nm SRAM poly layer by UMC with Luminescent.[54] A 193-nm dry scanner with NA 0.92, ½ annular (outer radii 0.95) illuminator, and AttPSM (6%) were used in this experiment. As shown in Fig. 8, the poly has a dense line/space pattern. The illuminator is tuned for such a dense line/space pattern, and therefore the CD through focus for the poly gate CD is flat, meaning the DoF is large. There would be no room for ILT correction to improve the DoF if the poly gate CD were the only concern. However, the ILT wafer shows larger exposure latitude, reducing the CD variation within the dose variation range from 30 to 20 nm.

The most significant improvement is seen for line ends: according to UMC, ILT correction demonstrated "remarkable" line-end shortening control in this case. Figure 9 shows the focus and exposure matrix for this poly layer. When considering poly gate CD only, ILT and OPC have

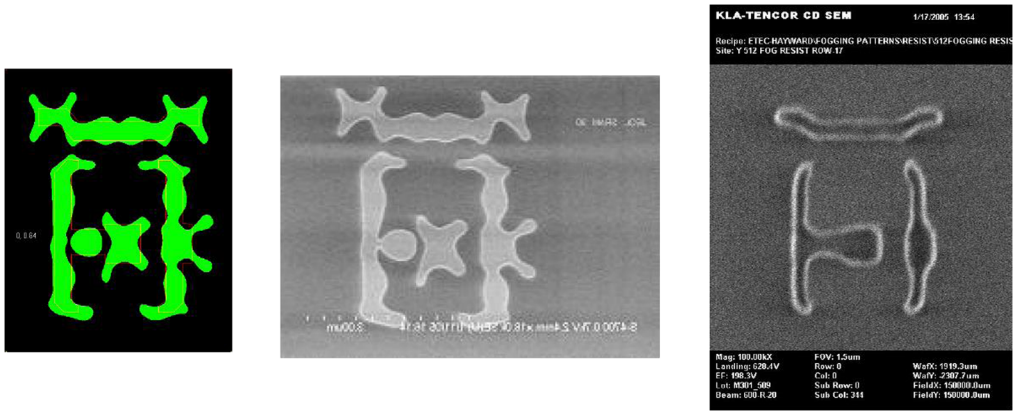


**Fig. 7** ILT mask pattern design, actual ILT mask pattern, and corresponding wafer print for a SRAM pattern[15] (source: Applied Materials).

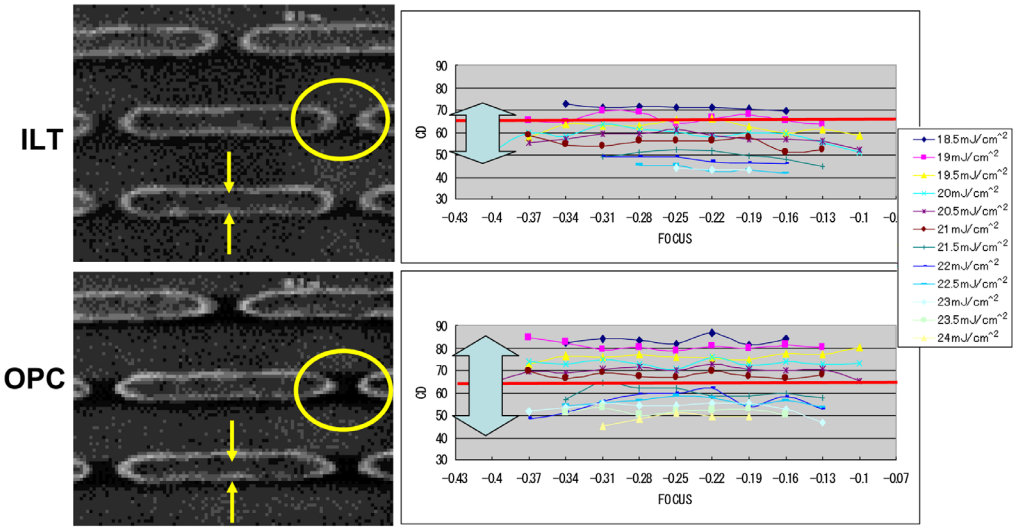


**Fig. 8** Wafer result of SRAM poly layer using ILT and OPC, showing ILT has a CD variation of 20 nm, whereas OPC has 30 nm[54] (Source: UMC).

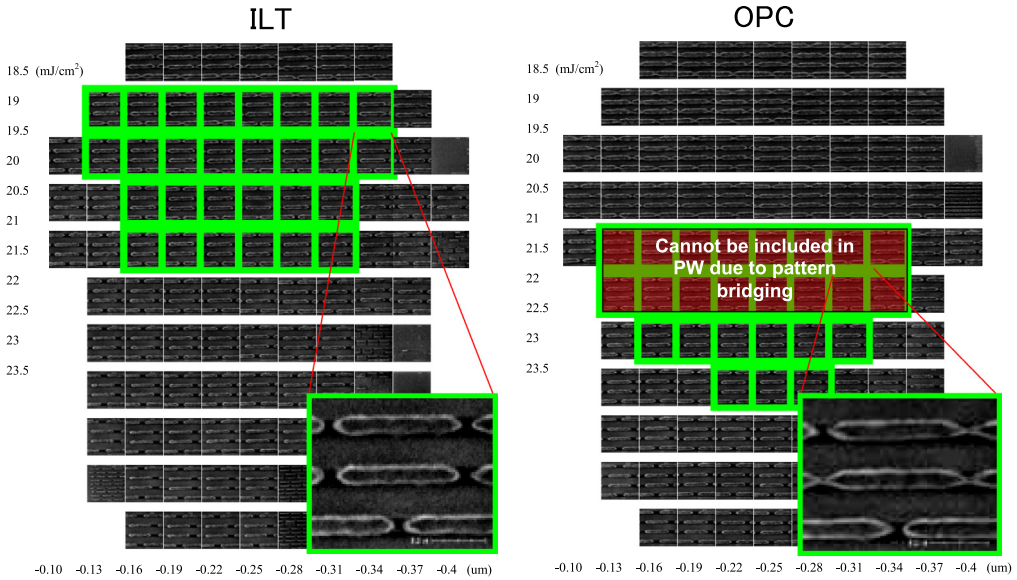


**Fig. 9** Wafer process window print of a 45-nm SRAM poly layer using ILT and OPC[54] (source: UMC).

similar process windows, but ILT correction shows a larger DoF at off-focus, because ILT considers images at multiple focuses in the inversion calculation (called process-window-based ILT). When the line-end distance is included into the common process window, many chips given standard OPC treatment show line-end bridging, unlike those patterned with ILT.[54] While this concept of using multiple process conditions in optimization was first introduced in ILT, it can be applied to OPC to improve its process window. However, in a case such as these dense line segments, OPC treatments at line-end are typically hammer-heads or serifs, so the OPC solution space is limited even if it uses multiple process conditions, while ILT is not. The benefit of this line-end shortening improvement is mainly due to the fact that ILT found more optimized solutions than OPC.

Working with the Luminescent solution, Hynix showed ILT can improve the pattern fidelity in DRAM design (Fig. 10).[67] In this case, when considering the DoF for dense line/space only,

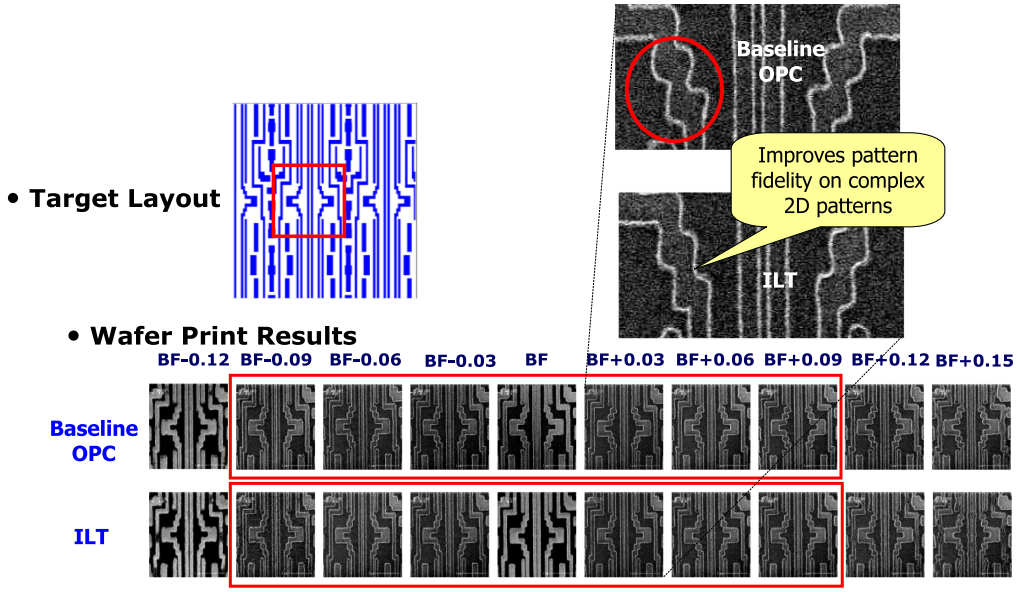


**Fig. 10** Wafer result of a 45-nm DRAM layer showing ILT improves pattern fidelity over OPC[67] (source: Hynix).

ILT and OPC have similar process windows. However, ILT produces 2D patterns with better pattern fidelity than OPC.

Samsung showed an example of applying Luminescent ILT correction to printing challenging patterns in a flash memory (Fig. 11).[54] In flash, the memory core is very dense, so there is no room to improve DoF with SRAFs. Usually, the illumination is tuned for the dense line/space pattern, because flash needs the finest resolution with any scanner. However, any array also has edges or connecting pads. Such irregular patterns become very difficult to print with adequate process window because the illumination is not tuned for them. ILT correction can help with such irregular sections in the dense line/space array by printing smaller irregular patterns and improving the pattern fidelity through focus. As shown in Fig. 11, the pad printed using ILT is more than 30% smaller than the smallest pad printable with OPC. Dense line/space patterns close to the pad show better fidelity through focus with ILT.[54]

Lithography for contact layers is the most difficult for any technology node. Luminescent worked with UMC and demonstrated its ILT could improve process windows for these layers, especially DoF, compared with OPC. Figure 12 shows the simulated images and SEM micrographs of a 45-nm SRAM contact layer printed using OPC and using ILT. These are critical contacts in two different configurations. ILT significantly improved the DoF compared with OPC — from 120 to 200 nm. A similar trend was observed on wafers.[54]

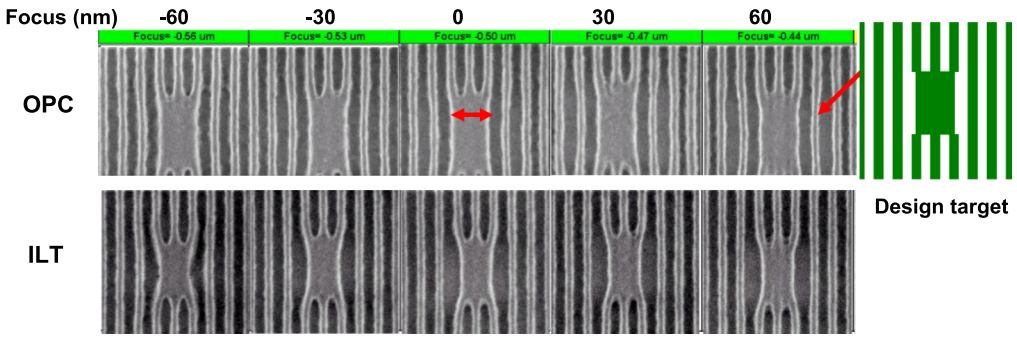


**Fig. 11** Wafer result of a FLASH design showing ILT having better fidelity than OPC on irregular pad patterns in dense line/space array[54] (source: Samsung).

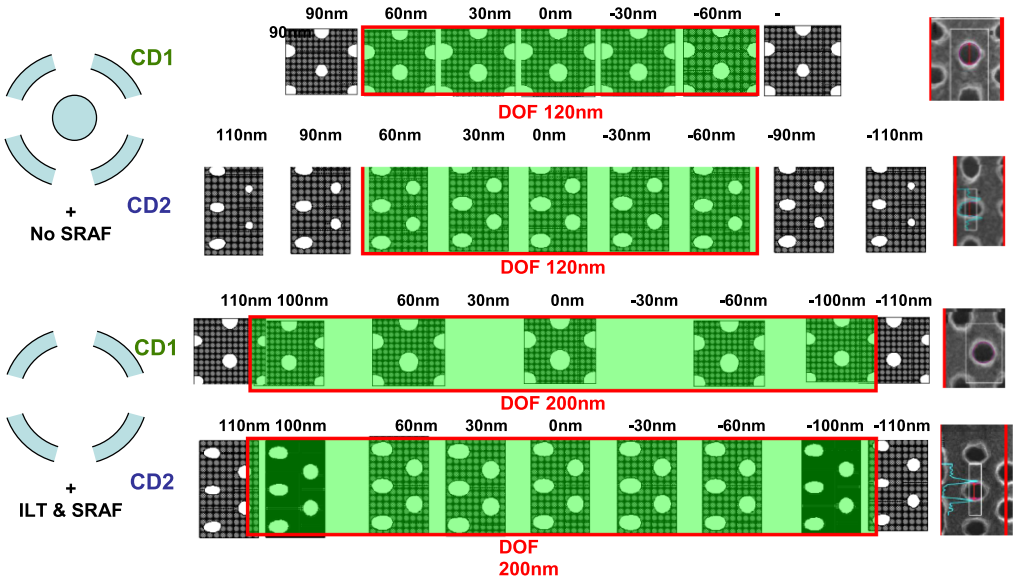


**Fig. 12** DoF for 45 nm SRAM contact with OPC and Luminescent ILT[54] (SEM source: UMC).

## 5.2 *Intel Pixelated ILT Method*

The Intel approach to ILT, presented by Vivek Singh and his team at the SPIE Advanced Lithography Conference in 2007,[29] used alternating phase-shift masks (AltPSMs) to improve the lithography resolution and also used a square shape, or "pixel," as the minimum mask feature to satisfy mask rules and to reduce the amount of ILT computation required. Unlike the Luminescent level-set method, which was trying to reduce the degree of freedom in solving the inverse problem to the edges of patterns, the Intel pixelated mask approach sought to reduce the degree of freedom to large pixels with only two degrees of freedom (0 degree of phase and 180 degree of phase). The entire design was mapped into these pixels.

To make the computation manageable, the pixel size was in the order of 100 nm in mask dimensions (25 nm in wafer dimension). Even with this 100 nm pixel size, the number of pixels for a full-chip design reached the order of trillions, which is why the 2007 paper was titled "Making a trillion pixels dance."[29] However, because 100-nm pixel size on mask is smaller than the wavelength at 193 nm, the mask 3D effect was strong, especially for AltPSM, where each phase is only 100 nm × 100 nm on mask. The Intel team had to develop a mask 3D model to accurately model the effect.

While using AltPSM improves the resolution, the limitation of this approach proved to be the 100 nm × 100 nm pixel size on mask or 25 nm × 25 nm on wafer, which is fairly large considering the design patterns are on a 1-nm grid or smaller. This limits the edge-placement accuracy. It is interesting to note that even through the Intel pixelated mask pixel is 100 nm × 100 nm square, the actual mask pattern on the mask is curvilinear (Fig. 13). This is due to mask writing resolution and mask process, such as resist resolution.

The Intel 2007 paper also discussed the stitching issues that arise for its full-chip ILT solution.[29] Like other traditional OPC and traditional ILT, Intel's ILT solution splits the full-chip design into small partitions and gives each partition to a CPU to compute its ILT solution. It uses a halo region that extends the partition into the neighboring partitions. However, when stitching the solution from each partition together, it was still seen that the ILT solutions from two neighboring partition were different at the partition boundary, causing stitching errors (Fig. 14). To solve such stitching errors, Intel used a technique called "stitch and heal," where it takes the region close to the stitching boundary and reoptimizes it. This increases the ILT runtime significantly, because this reoptimizing area is a fairly large percentage of the entire design. In addition, it does not guarantee to eliminate the problem since new partition boundaries—also vulnerable to errors—are created during this reoptimization.

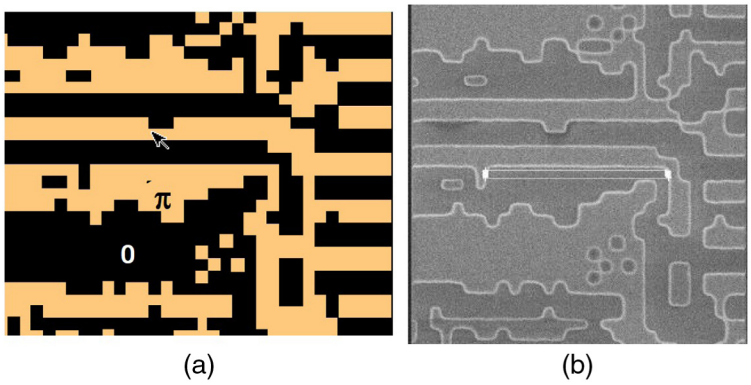


**Fig. 13** Intel pixelated mask pattern (a) the mask pattern with altPSM two phase pixels; (b) the actual mask SEM image, which shows curvilinear shapes[29] (source: Intel).

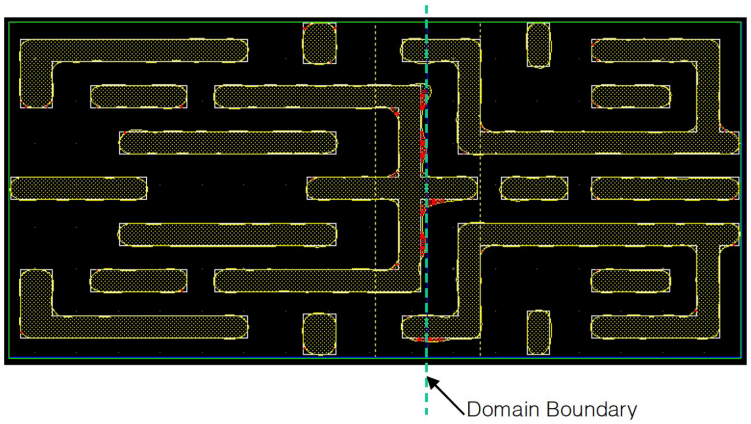


**Fig. 14** Example of stitching errors caused by inconsistent ILT solution at the partition boundary between the left partition and right partition[29] (source: Intel).

Intel also demonstrated its ILT on real masks and wafers. Its pixelated ILT mask designs were validated through the tapeout of an actual mask to pattern the most complex metal layer for a 65-nm-node microprocessor—the leading edge at that time—in high-volume manufacturing. This very first experimental tapeout resulted in wafer yield comparable to yields on mass-produced wafers made with production 65 nm technology. It was also shown that this technology can be used to eke out significantly more performance from steppers of a given generation.

The Intel ILT AltPSM is a totally transparent glass with etched trenches and holes (Fig. 15). This created huge challenges for mask inspection, since the industry-standard, high-resolution transmission image cannot see these patterns. Intel worked with Applied Materials to use the latter's aerial image mode to solve the problem, so the mask was inspected with the imaged wafer patterns.[29]

### 5.3 *Gauda/D2S GPU-Accelerated, Band-limited, Frequency-Domain Curvilinear ILT*

The D2S approach to solving the ILT problem expands and builds on the work initiated by Gauda (which D2S acquired in 2014) to solve the ILT optimization problem in the frequency domain,[33] as opposed to the real domain (which is what is used by both Luminescent and Intel) with GPU acceleration. In addition, the D2S approach seeks to address the "full-chip" aspects of the ILT runtime challenge, such as stitching errors, using a comprehensive hardware/software approach utilizing a GPU-accelerated computation platform (CDP) that is purpose-built for ILT.

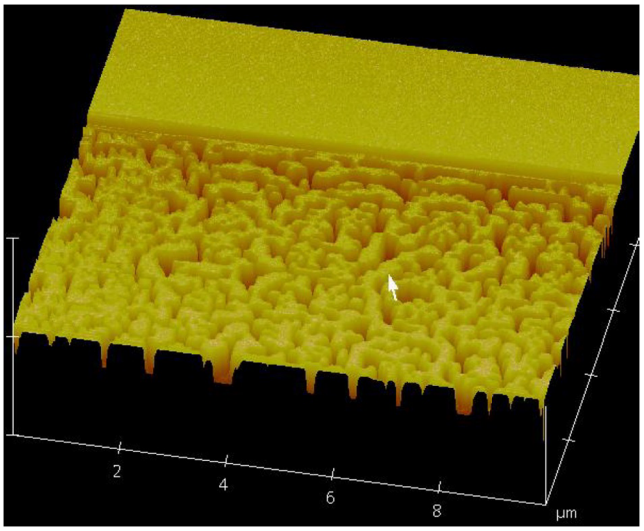


**Fig. 15** Atomic force microscope picture of pixelated phase mask showing the topography of the completely transparent AltPSM with etched trenches and holes[29] (source: Intel).

D2S ILT is based on a mathematically rigorous, band-limited, frequency-domain method, which naturally produces symmetric patterns and naturally avoids small features. The basic idea is that the same geometry (repeated patterns, symmetric patterns) in the real domain have the same frequency values/distributions in the frequency domain. If one modifies the cost function in the optimization in frequency domain, all the symmetric patterns and repeated patterns will be modified in the same way, and therefore, will naturally maintain the symmetry. With the band-limited scanner optics, a mathematically rigorous approach to geometry selection is necessary to produce these results. Another benefit of this approach is that because of these band-limited scanner optics, this band-limited function in the frequency domain has a clear cut off. By doing adjustments in the frequency domain, the band-limited nature is maintained easily, and the small features that are commonly seen in real-domain ILT methods are avoided.

Solution continuity and symmetry are always the most difficult things for most ILT approaches. That is why most ILT papers only show ILT patterns for random patterns to hide their symmetry issues. Figure 16 shows a symmetric three-contact configuration. When pitches change from small to large, the D2S ILT solution gradually changes while maintaining the $XY$ symmetry.

Another challenge for most ILT approaches is the on-grid and off-grid invariance. Figure 17 shows an equal-pitch contact array and its ILT solution. The top row is the on-grid case, while the bottom row is the off-grid case. When pitches change from small to large, the mathematically rigorous D2S ILT solution gradually changes while maintaining the $XY$ symmetry, and the solution for the off-grid case is identical to the on-grid case.

The most challenging test for ILT is the combination of multiple pitches, on-grid and off-grid situations, and rotation. Figure 18 shows the same equal-pitch contact array and its ILT solutions while the pitch is increasing, then also adding rotation. When pitches change from small to large, even with rotation, the ILT solutions gradually change while maintaining the symmetry. Since the source is an annular source, the ILT solutions are expected to be symmetric for any rotation angle, and we do see that from the D2S ILT solution.

The D2S ILT solution employs GPU acceleration to address the ILT runtime roadblock. GPU-accelerated computing excels at single-instruction, multiple data (SIMD) computation. This contrasts with CPU-based computing, which excels at logical (if-then-else) computation. Simulations of natural phenomena (such as weather or the physics effects inherent in semiconductor manufacturing) are SIMD computations, so GPU-accelerated computing is a natural fit for these operations, including ILT computations.

Several other attempts have been made to create commercial, full-chip ILT solutions by porting CPU-based solutions to a GPU-accelerated computing environment, including the original

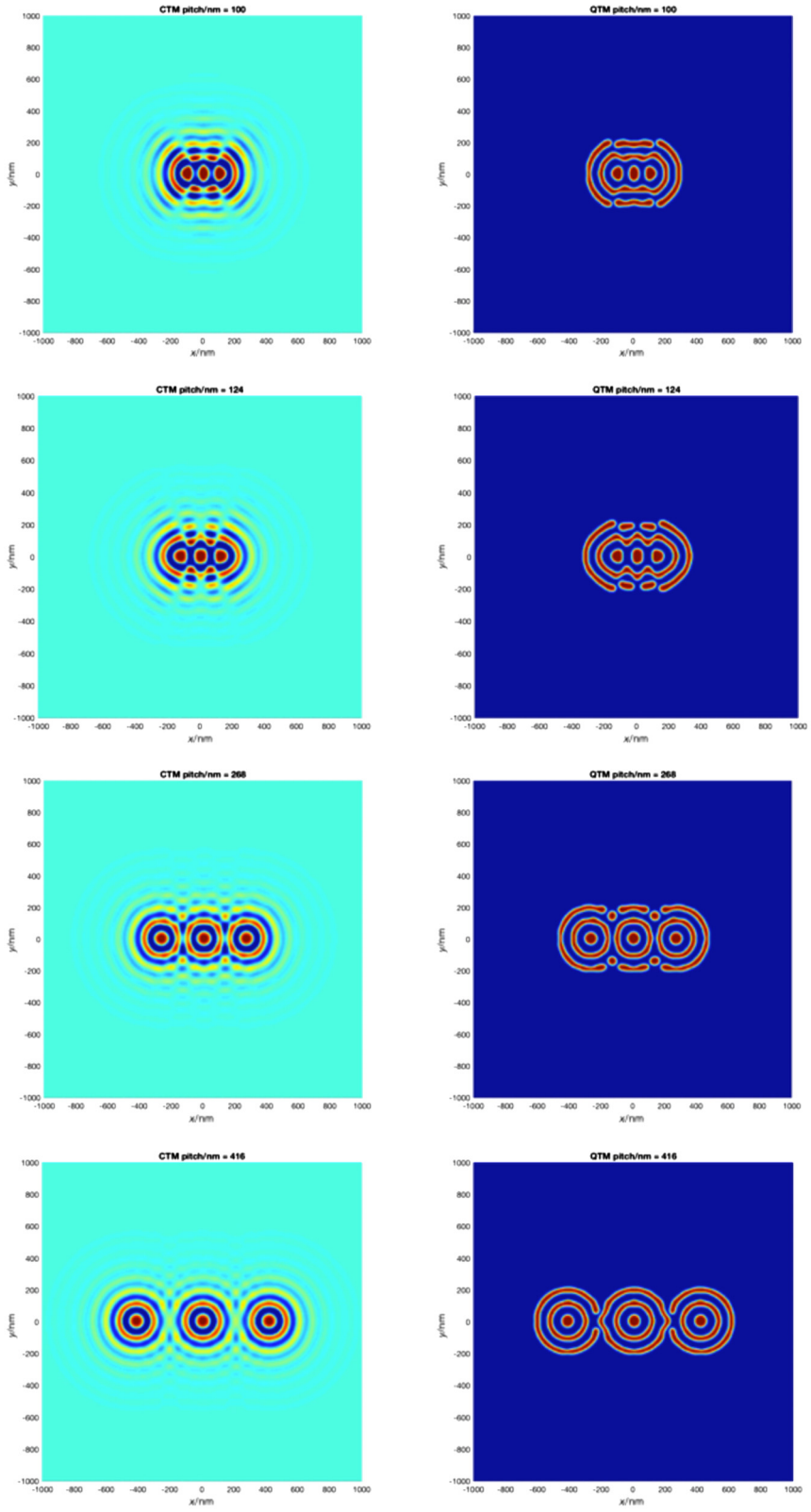


**Fig. 16** Continuous tone mask (CTM) and final ILT mask for three contacts in symmetric position at different pitches showing D2S ILT solutions are continuous and symmetric[21] (source: D2S).

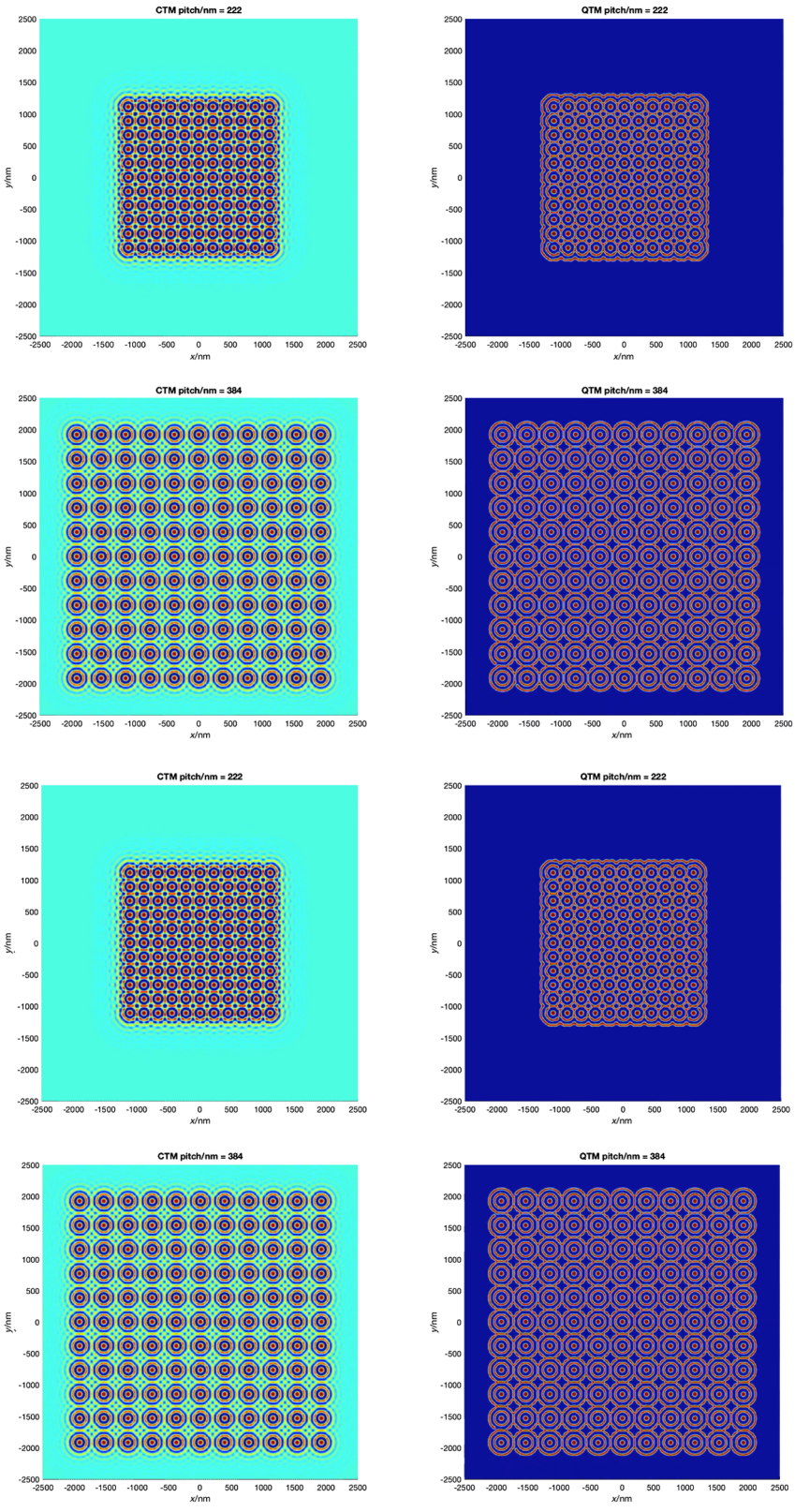


**Fig. 17** CTM and final ILT mask for an equal-pitch contact array for on-grid and off-grid situations. The top row is on-grid, whereas the bottom row is the corresponding off-grid configuration demonstrating D2S ILT solutions are grid-invariant[21] (source: D2S).

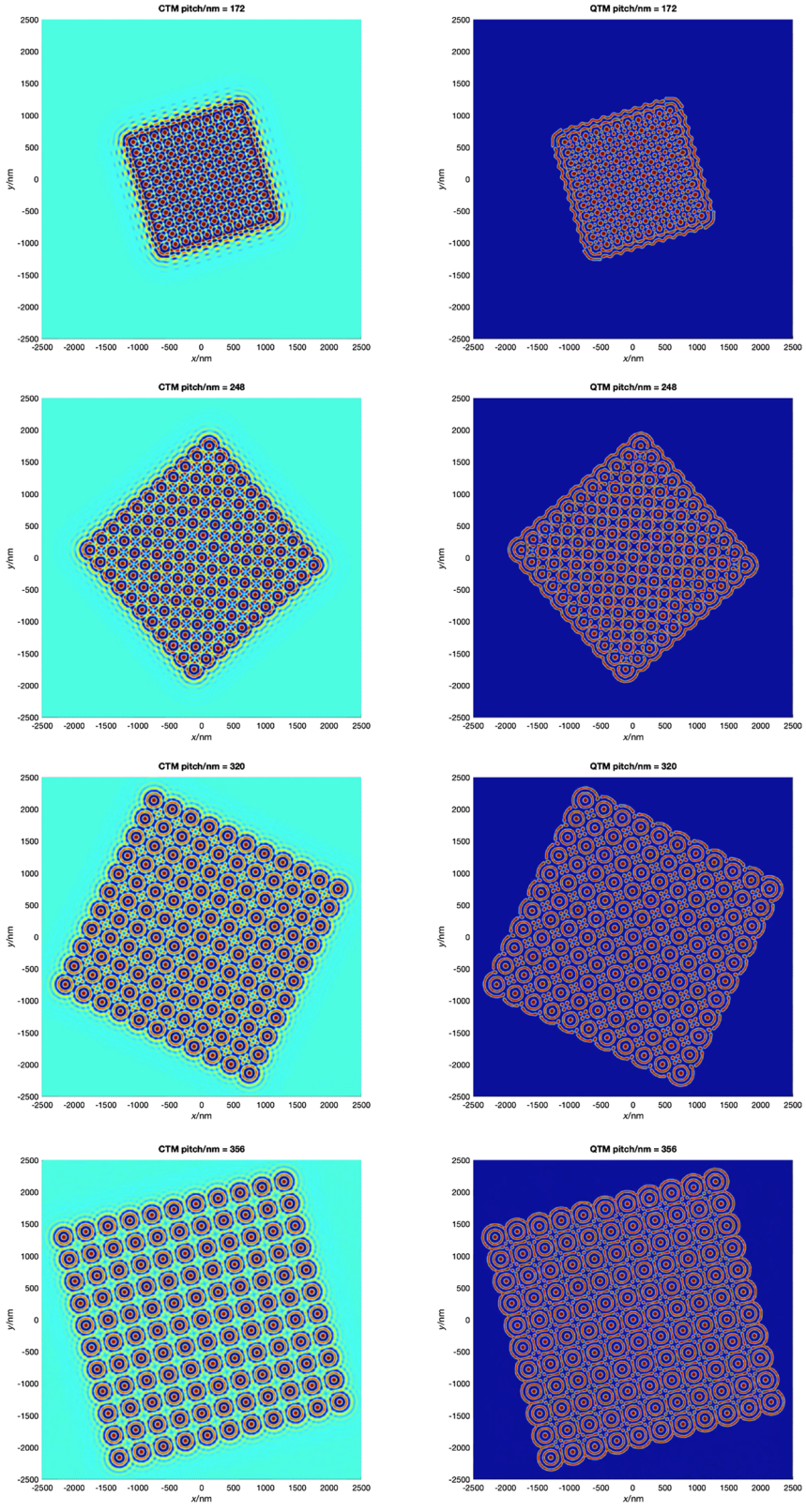


**Fig. 18** CTM and final ILT mask for an equal-pitch contact array at on-grid and off-grid situation, pitch change, plus rotation demonstrating D2S ILT solutions are symmetric and rotation invariant[21] (source: D2S).

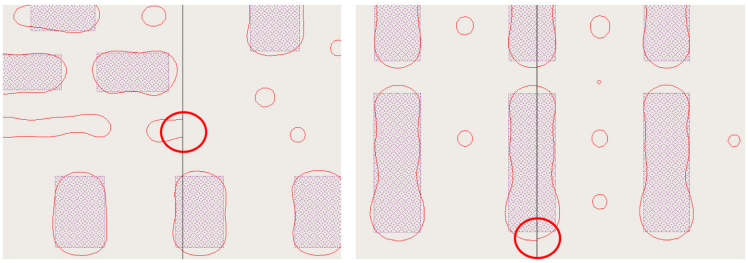

**Fig. 19** With traditional approaches, stitching errors occur when a chip is partitioned for parallel computing and reassembled[21] (source: D2S).

Gauda GPU-accelerated ILT.[33] However, these solutions still fell short in terms of acceptable turnaround time.

As we have discussed earlier, partitioning/stitching has been the major culprit for excess full-chip ILT runtime. D2S reasoned that what was needed was the ability to process the entire chip at once, using a single, giant GPU/CPU pair that could optimize full-chip data seamlessly, without partitions. Of course, such a giant GPU/CPU pair does not exist. However, by taking a "from the ground up," holistic approach, D2S built an ILT-specific computing appliance that could emulate a giant GPU/CPU pair[21] and designed its ILT solution so that the entire chip could be optimized at once. This speeds total runtime significantly by avoiding the time-consuming recursive correction passes necessary to resolve stitching errors, such as those shown in Fig. 19. The system behaves as though there are no partitions, so the solutions everywhere are continuous, as shown in Fig. 20.

The D2S holistic approach includes hardware, software, models, visualization, verification, etc., designed and implemented specifically for GPU-acceleration and for full-chip ILT computation.

Mask processes, similar to lithography processes, are limited or affected by dose profile and contrast, resist resolution, and etching process. Mask rules embody these limitations. Figure 21 shows an example of D2S curvilinear ILT output without and with integrated mask-rule checking (MRC). When MRC is integrated, the final curvilinear ILT mask corrects any features that violate minimum feature dimensions.[21]

Micron demonstrated 2X wafer process window improvement over its process of record (POR) OPC using the D2S ILT solution.[21] Figure 22 shows SEM images of some instances of the actual curvilinear mask pattern written by the NuFlare multibeam mask writer MBM 1000, and wafer images printed using the Micron POR. Mask patterns are resolved with high pattern fidelity and with an exceptionally smooth profile. On the wafer print, all contacts are printed evenly from the center of array to edge of the contact array.

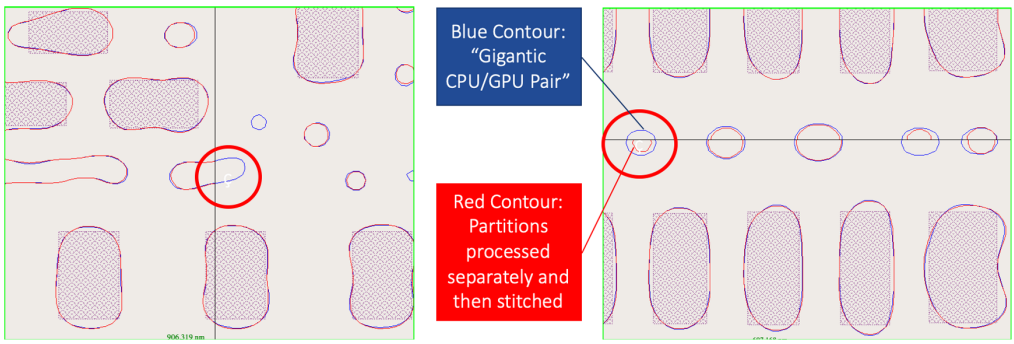


**Fig. 20** No partitions mean that no stitching errors occur in D2S stitchless curvilinear ILT[21] (source: D2S).

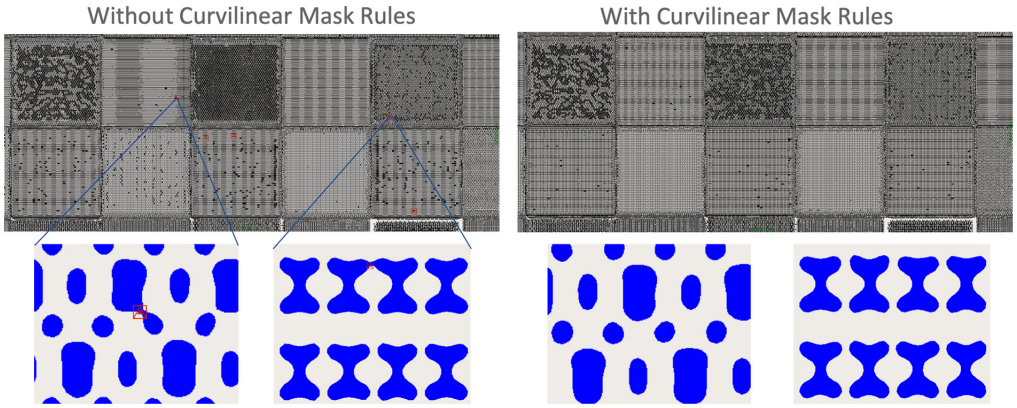


**Fig. 21** Comparison of D2S ILT without and with integrated MRC. The red marks are MRC violations detected by mask verification[21] (source: D2S).

Figure 23 shows the side-by-side wafer print comparison of OPC and D2S ILT prints for the entire process window matrix.[21] The target CD is 62.8 nm, all dies with CD with 10% variations are considered within process window.

Figure 24 shows the CD measurements; the conditions within process window are highlighted in green. Notice that the $x$ axis is the focus, $y$ axis is the dose to be consistent with the process window plot. Three wafer images at process center and two process corners are also shown, zoomed in. Compared with OPC, D2S ILT has increased the process window by over 100%.

In addition to GPU computation, the initial D2S solution introduced in 2019 also took advantage of another newly introduced technology: the multibeam mask writer. The mask industry had also recognized the challenge of writing curvilinear ILT mask patterns on VSB mask writer, and this became one of the motivations to develop a new multibeam mask writer.[115,116]

The multibeam mask writer, instead of having a single VSB, has an array of 256K beams that write in a single shot, and each individual beam can be turned on and off for a specified period of time up to a prespecified maximum exposure, providing grayscale exposure of each pixel location. Since multibeam mask writers write in the pixel domain, they are shape-agnostic in terms of write time and can write a mask with any-shaped mask patterns in a constant write time, around 10 to 24 h, including curvilinear ILT mask patterns (Fig. 25).

The D2S ILT solution combined purpose-built GPU-accelerated computation with multibeam mask writing to produce the first full-chip, curvilinear ILT solution with a practical runtime (around 48 h) and mask-write time (around 12 h), which is within the standard expectations for a traditional OPC/VSB implementation. This new development removes the runtime roadblock to wide adoption of ILT and assuages worries about mask manufacturability as well.

Using multibeam mask writing, which has a constant write time for all shapes, the VSB mask write time roadblock was circumvented. However, it is a practical reality that today, multibeam mask writers are in the early stages of deployment and the vast majority of production chips still are manufactured using VSB mask writers. For this reason, the advent of multibeam mask writing has not totally removed the roadblock (yet) to near-term adoption of ILT for production designs. In this interim period, finding a practical way to create full-chip, curvilinear ILT with VSB mask writers is still quite relevant. In the next section, we will review the efforts made by various product developers to address this challenge.

## 6 Curvilinear ILT on VSB Mask Writers

ILT naturally produces curvilinear mask shapes. As discussed previously, the fact that writing curvilinear ILT masks with VSB is difficult and slow has been one of the biggest roadblocks to wide adoption of ILT in production. In this section, we will review efforts made over more than 20 years to address the issue of creating curvilinear ILT shapes using VSB mask writers.

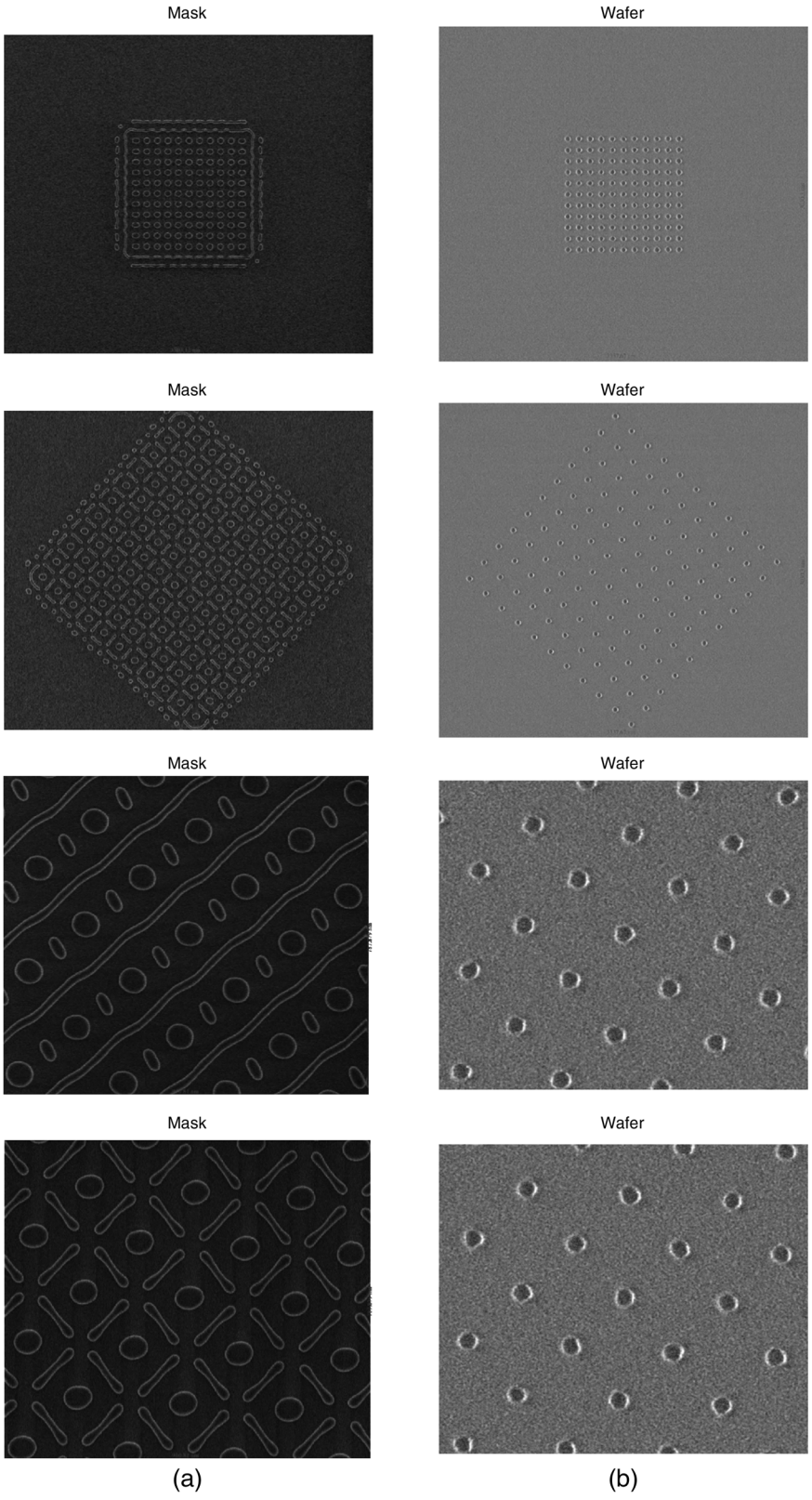


**Fig. 22** In each pair, (a) D2S ILT curvilinear mask patterns written by the NuFlare multibeam mask writer MBM 1000 for different pitches and orientations; (b) the corresponding wafer prints using the Micron POR[21] (source: Micron).

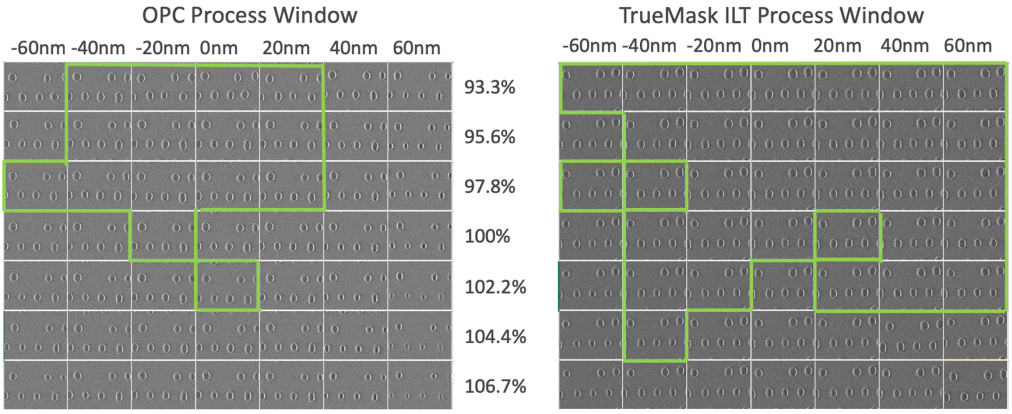


**Fig. 23** Wafer print matrix for a cut layer type of design. Highlighted regions are within process window[21] (source: Micron).

OPC process window

| | Dose (mJ) | Defocus (nm) -60 | -40 | -20 | 0 | 20 | 40 | 60 | Dose(%) |
|---|---|---|---|---|---|---|---|---|---|
| Latitude | 21 | 55.4 | 59.6 | 62.5 | 61.2 | 60.5 | 51.6 | 41.8 | 93.3% |
| | 21.5 | 56.3 | 57.3 | 63.7 | 61.0 | 58.1 | 52.9 | 40.3 | 95.6% |
| | 22 | 57.3 | 60.8 | 61.0 | 60.8 | 57.1 | 48.3 | 41.6 | 97.8% |
| | 22.5 | 54.0 | 52.8 | 60.1 | 55.8 | 54.7 | 52.1 | 39.3 | 100.0% |
| | 23 | 49.4 | 54.4 | 55.3 | 60.3 | 54.5 | 49.2 | 36.1 | 102.2% |
| | 23.5 | 49.2 | 50.9 | 54.5 | 55.0 | 51.6 | 46.1 | 33.9 | 104.4% |
| | 24 | 54.0 | 56.2 | 54.8 | 49.9 | 50.9 | 44.7 | 28.6 | 106.7% |

TrueMask ILT process window

| | Dose (mJ) | Defocus (nm) -60 | -40 | -20 | 0 | 20 | 40 | 60 | Dose(%) |
|---|---|---|---|---|---|---|---|---|---|
| Latitude | 21 | 57.7 | 60.5 | 59.6 | 64.1 | 58.8 | 62.1 | 58.4 | 93.3% |
| | 21.5 | 54.0 | 56.9 | 58.1 | 60.9 | 59.0 | 62.8 | 59.4 | 95.6% |
| | 22 | 60.6 | 55.4 | 59.3 | 60.3 | 57.9 | 59.5 | 58.8 | 97.8% |
| | 22.5 | 54.5 | 57.7 | 57.6 | 60.6 | 56.2 | 59.4 | 57.9 | 100.0% |
| | 23 | 52.0 | 56.5 | 57.0 | 56.2 | 57.3 | 60.9 | 58.3 | 102.2% |
| | 23.5 | 52.2 | 56.7 | 55.1 | 54.6 | 54.9 | 55.6 | 54.7 | 104.4% |
| | 24 | 48.6 | 52.7 | 51.6 | 50.8 | 56.1 | 52.3 | 56.1 | 106.7% |

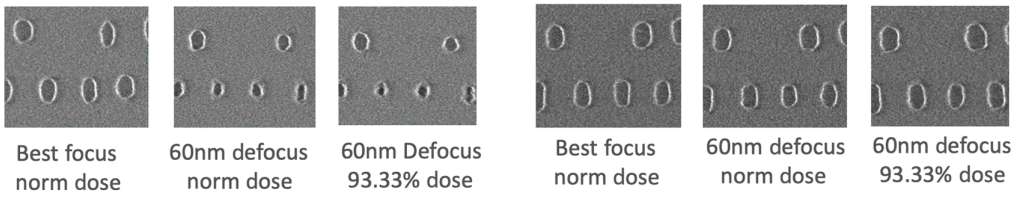


**Fig. 24** Process window CD measurements. The highlighted regions are within an acceptable process window[21] (source: Micron).

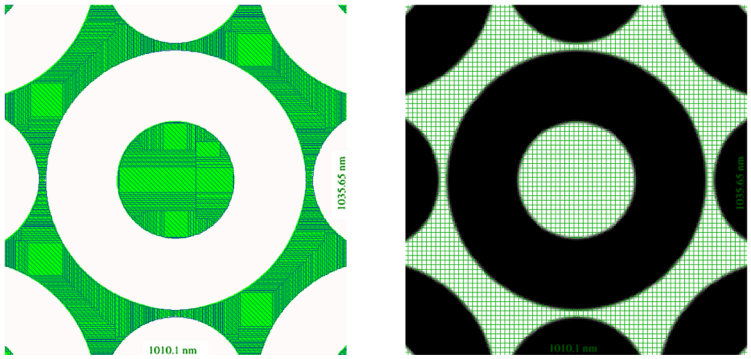


**Fig. 25** (a) Conventional VSB mask writers require curvilinear shapes to be fractured into many rectilinear shapes, which results in too many shots for a practical write time. (b) Multibeam mask writers, designed for curvilinear ILT, write any shape in constant time[21] (source: D2S).

### 6.1 *Early Results Demonstrated That Curvilinear Mask Shapes Produce the Largest Process Window*

Why spend so much time and effort trying to write curvilinear mask shapes with a mask writer that can only produce rectangles (or in some cases, triangles)? We have decades of evidence that curvilinear mask shapes produce the largest process window and better manufacturing resilience.[18] Because nothing in nature (including the physics of semiconductor manufacturing) makes 90-deg corners, manufactured masks and wafers are all curvilinear, even if the input geometries are rectilinear, as we saw with the Intel pixelated mask solution in Fig. 13 and in the example in Fig. 26.

In fact, curvilinear shapes with certain minimum curvatures of shapes and spaces have been shown to be more reliably manufacturable than rectilinear shapes.[118] These benefits have fueled decades of research and development, as teams have sought solutions to the problem of creating curvilinear mask shapes with a rectilinear mask-writing tool.

As far back as 2005, Luminescent, in conjunction with semiconductor manufacturing companies, demonstrated that full curvilinear ILT mask patterns produce the largest process window. For example, Samsung and Luminescent showed through-pitch contact arrays with different assist feature simplification schemes fractured for a VSB mask writer (Fig. 27).[18]

The mask SEM images of these patterns are shown in Fig. 28. The overall mask fidelity, even that for the complex mask C0, looks good.

Figure 29 shows the DoF and VSB shot count of these patterns. In general, both DoF and VSB shot counts decrease with decreasing SRAF complexity. The ILT solutions here have been

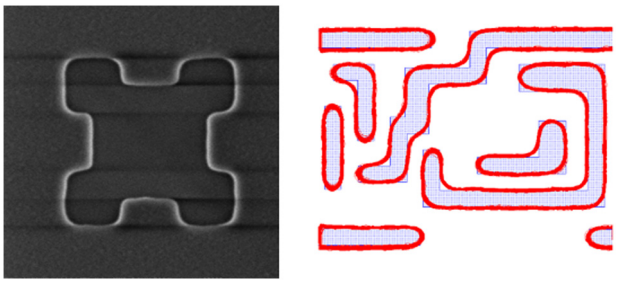


**Fig. 26** All shapes on masks and wafers are curvilinear, even if the input geometries are rectilinear: (a) Manhattan OPC mask pattern with serif are curvilinear on an actual mask. (b) Wafer pattern designed as Manhattan is curvilinear on an actual wafer[23] (source: D2S).

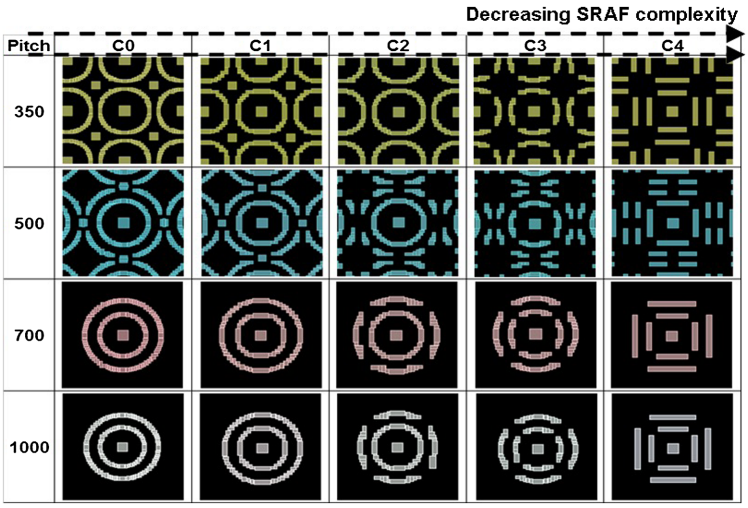


**Fig. 27** ILT solutions for through-pitch contact array with varying degrees of complexity[18] (source: Luminescent/Synopsys).

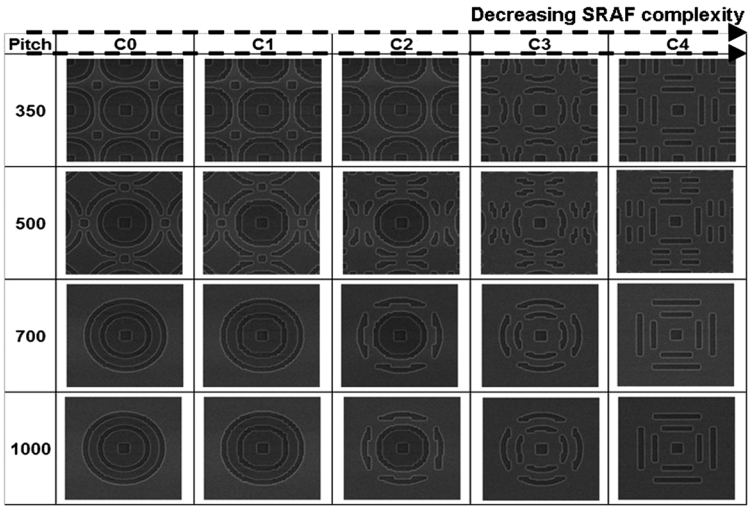


**Fig. 28** SEM images of through-pitch ILT contact array patterns with varying degrees of SRAF complexity[18] (source: Samsung).

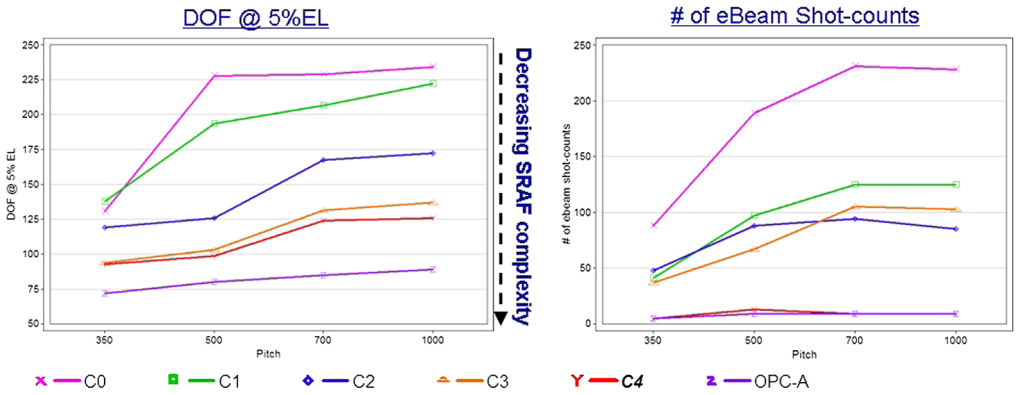


**Fig. 29** DoF and mask VSB shot-count performance for the through-pitch contact array with varying degrees of complexity[18] (source: Luminescent/Synopsys).

compared with a model-based OPC solution, designated as OPC-A. As can be seen from the graphs, the simple ILT solution C4 has approximately the same # of shots as OPC-A but provides substantially higher DoF performance.

As can be seen from Fig. 29, within the ILT mask solutions, the DoF performance of C0, which is the full curvilinear ILT solution, is the best. However, its shot count on a VSB mask writer is also much higher than the other solutions, making the mask write time for a full reticle not feasible. The work to improve VSB write time while preserving the benefits of curvilinear mask shapes continued.

## 6.2 *Adapting Curvilinear ILT to VSB Mask Writers: Strategic Simplification*

ILT mask manufacturability and write time on VSB mask writers are directly linked to pattern complexity. In the early days of ILT, many techniques were explored, mainly by Luminescent and its partners, to reduce the shot count while minimizing the loss of process window.[55,56] As we have discussed, the standard approach to adapting curvilinear ILT patterns to VSB mask writers is to Manhattanize, or fracture, the curved design into small rectilinear shapes that are grouped to approximate the curvilinear contours. As shown in Fig. 30, minimum fracture size affects mask patterns.[14] Resist processing improves smoothness of the edges.

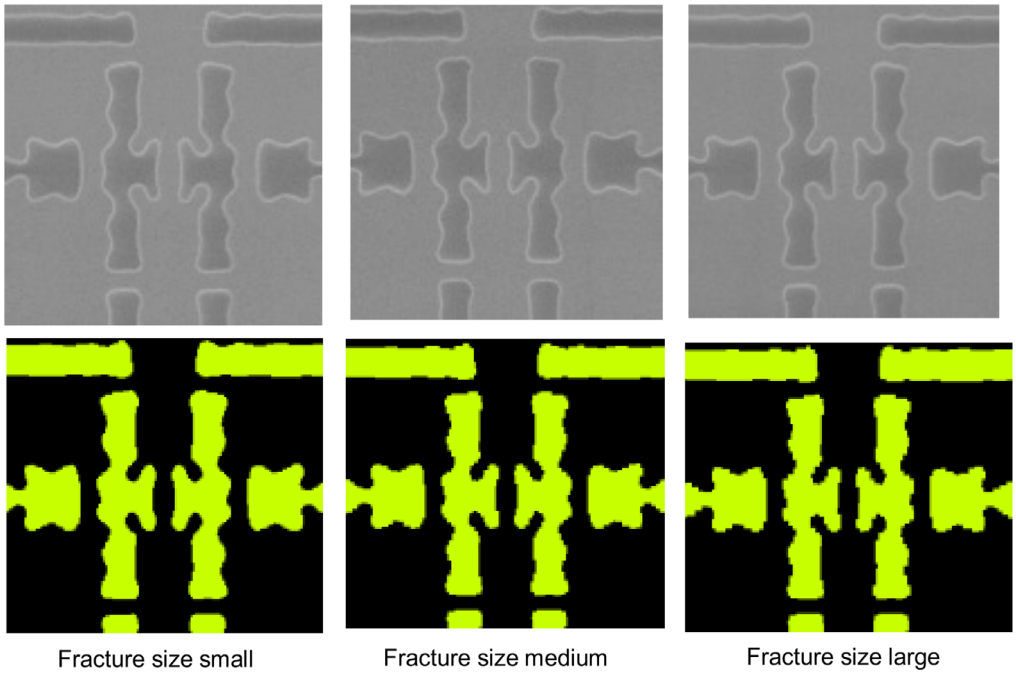


**Fig. 30** ILT mask designs at different minimum fracture size and their corresponding masks written by e-beam VSB writer. Beyond a certain fracture size, they all look the same[14] (source: Photronics).

Fairly early on, ILT with model-based SRAFs demonstrated significantly larger process windows compared with conventional OPC approaches. However, each SRAF in an ILT solution adds time to write the mask. One way to curb the increase of write time resulting from aggressive SRAFs is to limit SRAFs only to regions where they are needed to achieve the required process window. For line and space layers, such as poly gate or active layers, this is especially advantageous due to the large variation of feature sizes in these layers. If SRAFs were not applied to larger noncritical patterns, mask complexity and writing time could be reduced without compromising wafer yield.[55,56]

An illustration of this method is shown in Fig. 31. Figure 31(a) shows an implementation with SRAFs placed everywhere possible. Figure 31(b) shows an ILT mask with SRAFs restricted according to local feature sizes. Here, two bands of exterior SRAFs were applied to small critical features, but only a single SRAF or no SRAFs where the features are larger. Similarly, interior SRAFs were only used where needed to prevent bridging at defocus or to satisfy process window requirements. Figures 31(c) and 30(d) show fractured VSB figures of the masks in Figs. 31(a) and 31(b). The total number of VSB figures in Fig. 31(d), with selective SRAF placement, is 40% less than the implementation without restrictions based on feature sizes shown in Fig. 31(c). This was a big step toward simpler masks for line/space layers and can be used in combination with other methods for further reduction. As shown in Figs. 31(c) and 31(d), many small VSB figures are generated along relatively straight edges and corners. To further reduce shot count, an improved Manhattan segmentation was developed.

### 6.3 *Improving Manhattan Conversion of Small Jogs and Corners*

As we have seen, at advanced process nodes, the sharp corners in the Manhattan mask shapes become more rounded during the resist process and absorber etch, because while the corner rounding caused by eBeam short-range scattering stays about the same, with a reduced feature size, the effect is more apparent. By taking this into account, it is possible to obtain a good approximation of an ideal curved mask with relatively coarse Manhattan segments.[55,56] Figure 32 compares results from older [Fig. 32(a)] and newer [Fig. 32(b)] Manhattanization

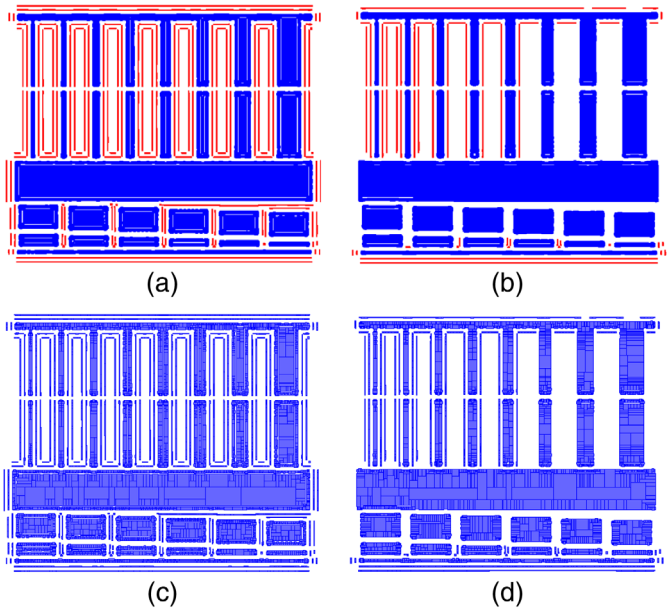


**Fig. 31** Comparison of SRAF insertion with and without local CD awareness. (a) ILT mask with traditional ILT SRAF insertion where SRAFs appear at all possible places; (b) ILT mask with local CD-aware selective SRAF insertion; (c) fractured VSB figure map of ILT mask in (a); (d) fractured VSB figure map of ILT mask in (b)[55] (source: Luminescent/Synopsys).

algorithms from Luminescent.[56] At the corners of line ends, the newer Manhattan algorithm uses much coarser segmentation than the old Manhattan algorithm, so it may be fractured with fewer VSB shots. Many of the small jogs that are eliminated have negligible effects on wafer images. The long edges of main features and SRAFs also show that the new algorithm eliminated about 50% of the shots used in mask shapes produced by the older algorithm. SRAFs show the largest reduction, so the net benefit of the new algorithm varies from about 2× to 4× depending on pattern density. Figures 32(c) and 31(d) show nominal image contours from masks in Figs. 32(a) and 32(b), respectively. Both masks generate nominal images on target. Figure 32(e) is a zoomed-in image of line-ends, showing an overlay of contours from both algorithms. The contours are nearly identical, except at corners, for which the older Manhattan algorithm with finer segmentation deviates less from the target, but this deviation is not very impactful.

## 6.4 *Jog Alignment for VSB Shot Count Reduction*

Selective SRAF placement and variable coarseness for Manhattan conversion will directly affect the mask shape and VSB shot count. Jog alignment is another option that does not significantly change either the mask shape or the size of the output file. It reduces the number of VSB shots by aligning facing jogs across the pattern.[55,56] Figure 33 shows an example of Manhattan masks with and without jog alignment. Figure 33(a) is an overlay of the two mask patterns, showing that jog alignment produces only minor differences. Figures 33(b) and 33(c) show the VSB shot patterns needed to write the two masks. The sub-nanometer movement needed to align the jogs reduces VSB shot count without significant compromise to EPE performance.

## 6.5 *Overlapping Shots and Mask Simulation for VSB Shot Count Reduction*

In 2010, D2S introduced the concept of using model-based mask data preparation (MB-MDP) that employs overlapping VSB shots to significantly reduce the curvilinear ILT mask shot count while improving dose margin. Fujimura et al.[70,71] showed an ideal ILT mask pattern for 22-nm contact holes, with curvilinear patterns written with about the same number of shots as a

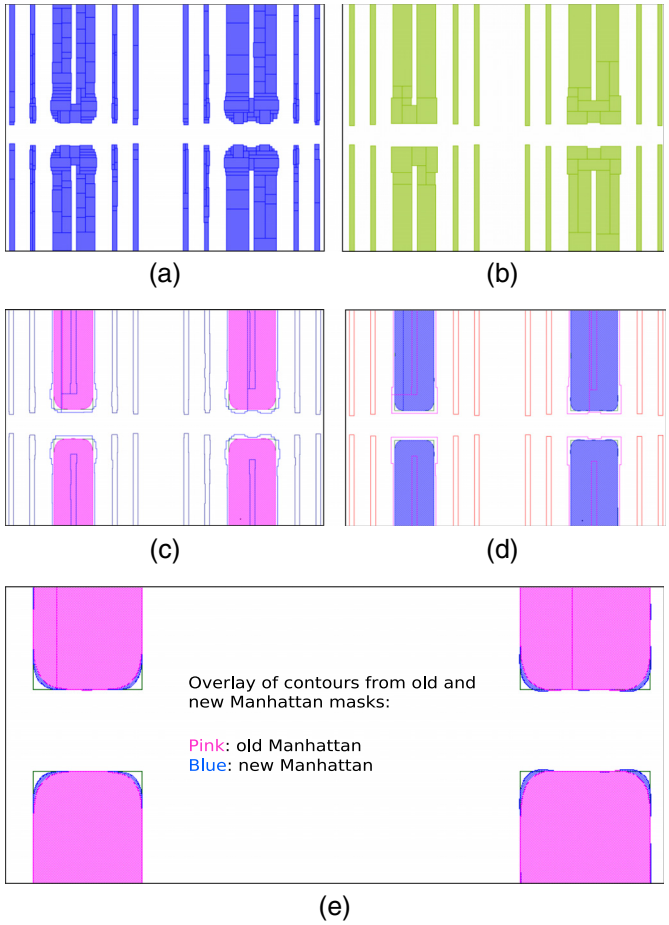


**Fig. 32** Comparison of old and improved Manhattan conversion. (a) Fractured VSB figures of mask created with legacy Manhattan fracturing. (b) Fractured VSB figures of mask created with new Manhattan fracturing. (c) Nominal image contour from old Manhattan mask in (a). (d) Nominal image contour from new Manhattan mask in (b). (e) Overlay of nominal image contours from both masks[56] (source: Luminescent/Synopsys).

simplified Manhattanized mask (Fig. 34). In this case, the VSB shots used an experimental circular aperture developed with JEOL, but the same techniques have also been applied to rectilinear VSB figures.

Figure 35 shows a typical curvilinear ILT mask pattern, fractured for a VSB mask writer. The pattern on the left uses conventional MDP for VSB; the pattern on the right employs MB-MDP with overlapping shots to create the same pattern. There are two observations from this example: first, overlapping shots can significantly reduce total shot count; and second, the majority of shots in this case (and in most production designs) are for the SRAFs, not for the main features. As we know, SRAFs have far less impact on the wafer EPE as compared with main features.[55,56] For any given target main feature in a contact layer, an overwhelming number of shots are used for the SRAFs in a conventionally fractured solution. Overlapping shots produce SRAFs that perform well without devoting so much of the VSB write-time to producing them.

However, these shot-reduction techniques notwithstanding, full-chip curvilinear ILT mask patterns remained impractical for VSB mask writers until very recently.[23]

ASML Brion and NCS also studied curvilinear ILT Manhattanization (which they call stair-casing) on VSB together with curvilinear mask process correction (MPC) on multibeam mask writer. They discovered the same thing: although VSB can write stair-cased ILT, the full-chip mask write time is not practical (95 h).[117]

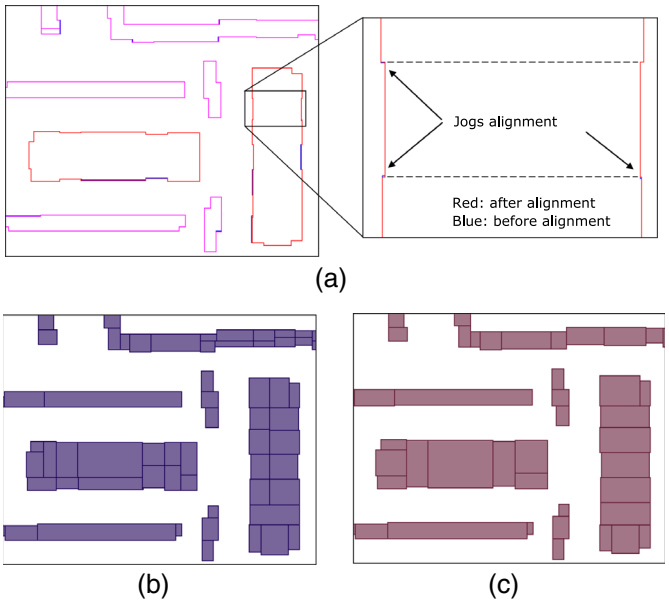


**Fig. 33** Shot reduction from jog alignment. (a) Overlay of ILT masks with and without jog alignment. A few locations marked by arrows display the effect of jog alignment option; (b) fractured VSB figures from mask without jog alignment; and (c) fractured VSB figures from mask with jog alignment[56] (source: Luminescent/Synopsys).

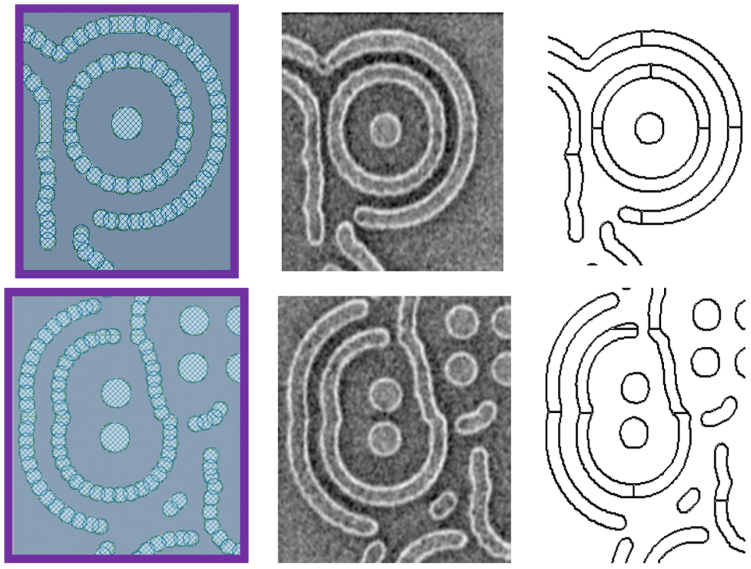

**Fig. 34** Example of MB-MDP: the smaller, overlapping circular VSB shots create "ideal" curvilinear ILT shapes with practical shot-count and runtimes[70] (SEM source: JEOL).

## 6.6 *Overlapping Shots and Mask Wafer Co-Optimization Solve the VSB-Written ILT Mask Problem*

In 2020, D2S introduced a technique called MWCO for 193i.[22,23] This approach shifts the OPC-to-mask-shop hand-off from mask shapes to mask shots and then uses simulated wafer EPE to guide shot reduction and placement decisions.

Today's semiconductor manufacturing process separates the responsibilities between the OPC/ILT shop and the mask shop, such that the OPC/ILT shop has the responsibility to specify

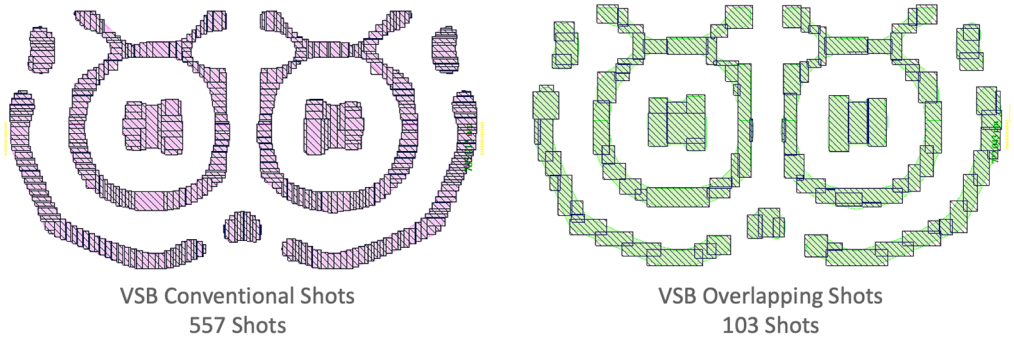


**Fig. 35** Example of a curvilinear ILT mask pattern written by VSB mask writer with conventional (fracturing) shots and overlapping rectilinear shots[23] (source: D2S).

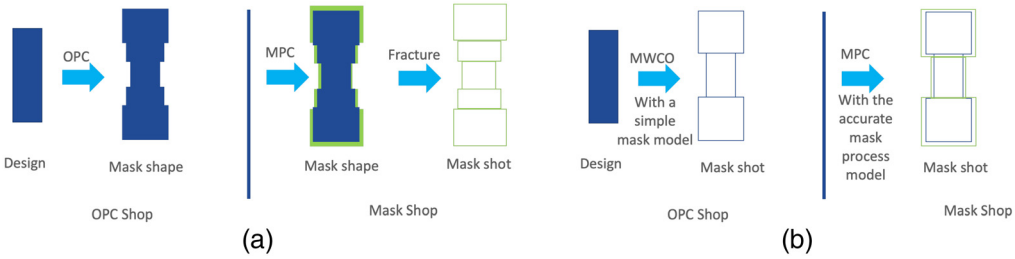


**Fig. 36** (a) Today, mask shape is the hand-off between OPC and mask shops. (b) MWCO shifts hand-off to mask shots[23] (source: D2S).

the desired mask shapes to achieve the best wafer results, as shown in Fig. 36(a). The mask shop has the responsibility to manufacture the masks as close as possible to the shapes specified by OPC/ILT. The mask shop, for VSB writing, fractures the mask shape into rectangles, where each rectangle is a VSB mask writer shot.

MWCO marries the D2S GPU-accelerated curvilinear ILT solution with the GPU-accelerated curvilinear MDP for VSB writers, using overlapping shots. MWCO incorporates overlapping shot generation and mask-wafer double simulation into the ILT process, so the output of the OPC shop is already optimized for shot count (Fig. 37). Using double simulation, wafer EPE is iteratively optimized while manipulating VSB shot edges to produce rectilinear target mask shapes that are known to be writable on a VSB writer with a known and acceptable shot count. In the MWCO flow, the OPC shop hands off mask shots to the mask shop, instead of mask shapes,

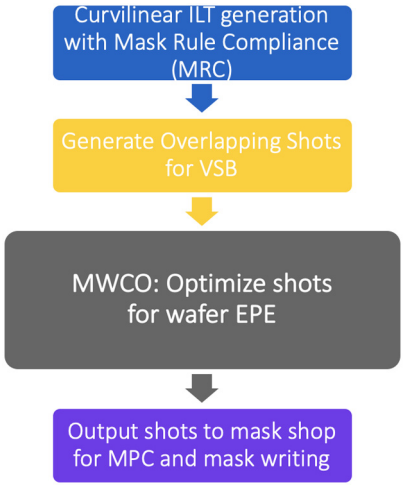


**Fig. 37** MWCO flow for full-chip, curvilinear ILT for VSB mask writers[23] (source: D2S).

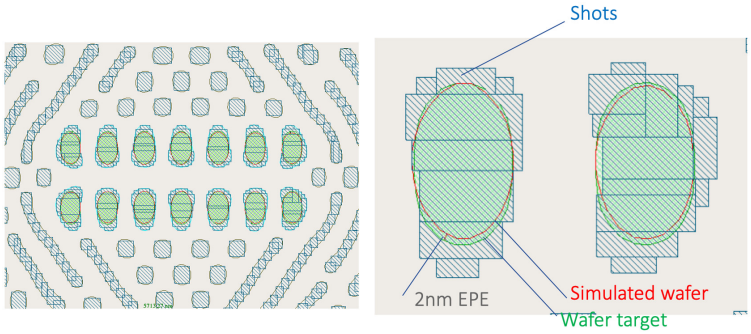


**Fig. 38** VSB shots generated to minimize mask EPE[23] (source: D2S).

as seen in Fig. 36(b). The mask shop will still run MPC, with its more accurate mask process models, but the mask shop does not need to fracture the mask shape—the file given by the OPC shop is a shot list that a VSB mask writer is known to be able to write.

Figure 38(a) shows an example contact array with curvilinear ILT for 193i producing desired curvilinear mask target shapes, then the VSB shots generated to produce it using overlapping shots.[70,71,119,120] In the figure, green lines show the wafer target, red lines show the wafer image simulated from mask images simulated from the VSB shots in a double simulation process. The VSB shots are shown as hatched blue rectangles. Overlapping shots are used to shoot SRAFs. For the SRAFs, thin brown lines reflect the target curvilinear mask shapes output by curvilinear ILT. Nonoverlapping shots shoot the main features, but with shot count just large enough to produce the target mask contour as specified by curvilinear ILT (not shown). MDP for overlapping shots is simulation-based, with an iterative optimization to produce a shot configuration that produces the desired mask contour with a low shot count, taking advantage of the natural corner-rounding in the mask process, which is especially prominent with SRAF dimensions. Figure 38(b) shows a zoomed-in picture of the two main features on the lower right of the contact array. Without using MWCO, the red contour of the simulated wafer image comes within 2-nm EPE after mask-wafer double simulation. Because this process first produces the target curvilinear mask shapes using curvilinear ILT and then separately optimizes the VSB shots to hit the desired mask contours, the trade-off with shot count inevitably results in accuracy loss, such as this 2-nm EPE.

The wafer results can be much improved with MWCO. Figure 39 shows the results when the shots to produce the mask contours are moved based on mask-wafer double-simulated wafer EPE. By taking this approach, without changing the shot count or shot configuration much, the wafer EPE is reduced from 2 to 0 nm at the same location and <1 nm in all the shapes. Iteratively optimizing VSB shot edges while optimizing for wafer EPE significantly improves the ability to target curvilinear mask shapes while minimizing impact on shot count.

Once the optimization target is changed from mask to wafer, MWCO can further reduce the shot count, since the scanner is a band-limited optical system that will filter out high-frequency

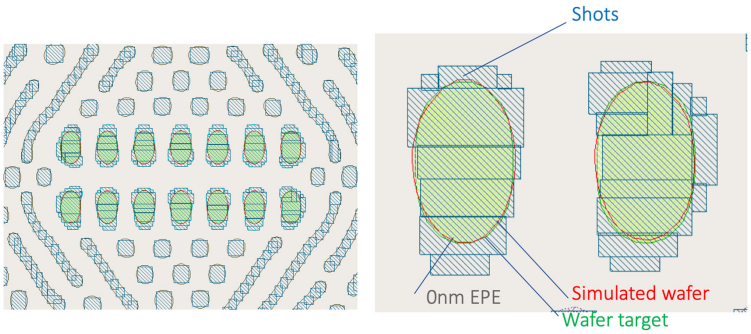


**Fig. 39** VSB shots generated to minimize wafer EPE[23] (source: D2S).

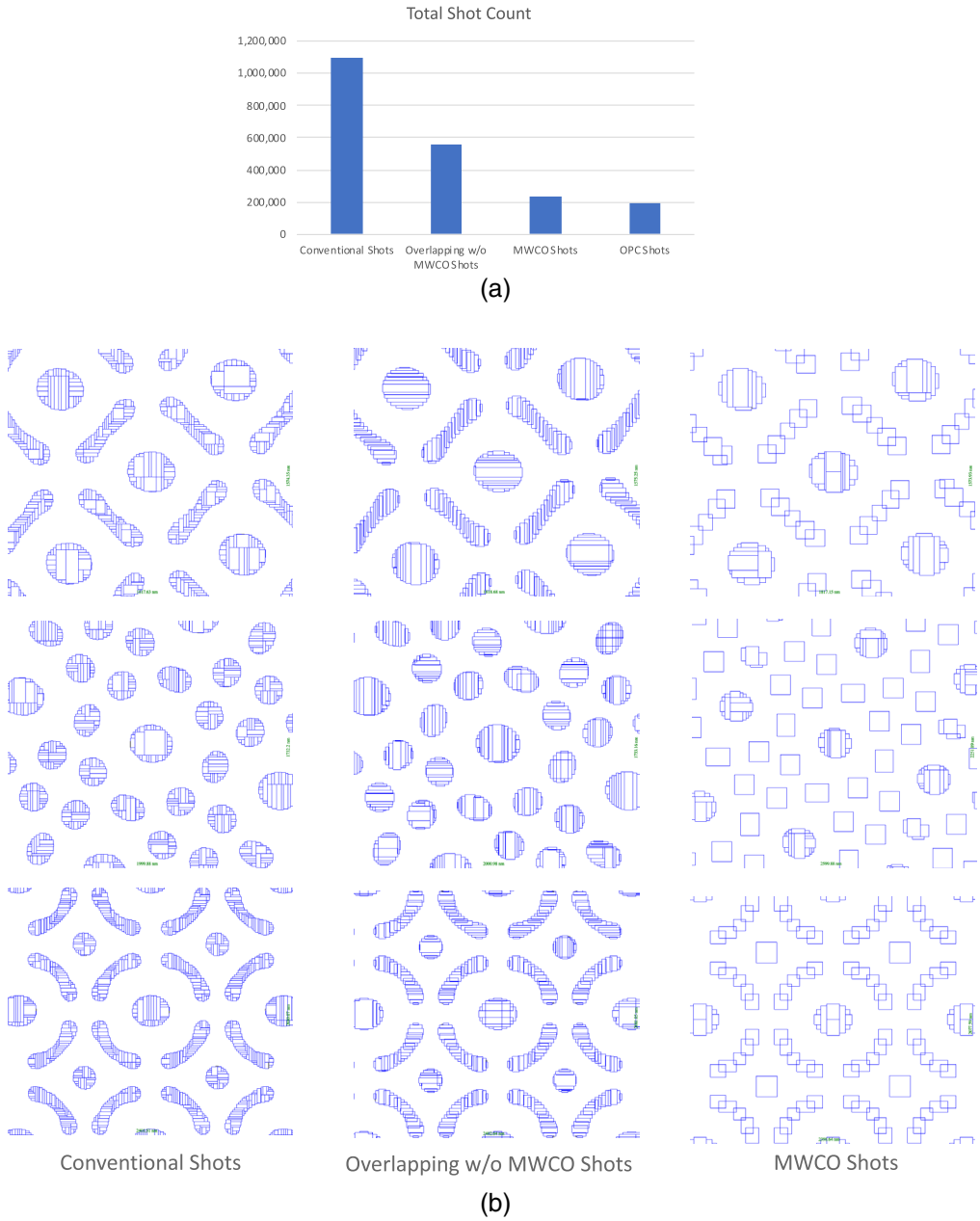


**Fig. 40** (a) VSB shot count and (b) shot configurations for three contact arrays. Note the POR OPC shot configurations are not shown in (b)[23] (source: D2S).

features on mask. Figure 40 shows three clips for the contact array.[23] It has 121 different configurations of an $11 \times 11$ contact array, each with slightly varied pitch and rotation angle. The contact array sequence includes features in a spectrum of placements, from dense placement all the way to nearly isolated features, with the contact array rotated to demonstrate the underlying curvilinear, all-angle, nature of this solution. The total shot count when using conventional fracturing is a little over 1 million shots. Overlapping shots without MWCO reduces the shot count by roughly half to a little over half a million shots. MWCO reduces the shot count by half again, to a little less than a quarter million. The OPC shot count is about 200K, so the MWCO shot count is comparable to OPC shot count, meaning the MWCO mask has about the same write time as OPC mask.

Micron wrote this mask using the NuFlare VSB mask writer EBM-9500 PLUS.[23] Figure 41 shows mask SEM images for the three configurations from Fig. 40. One can see that the

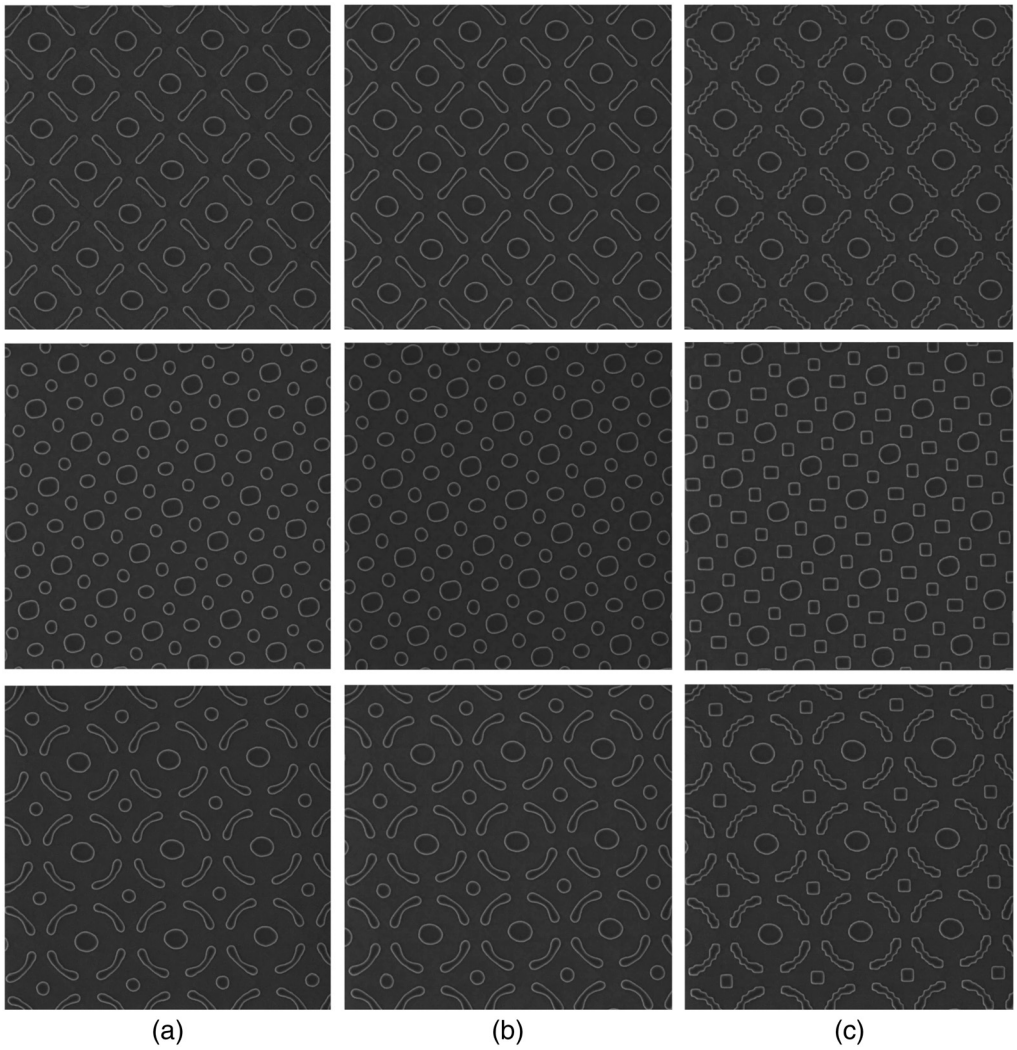


**Fig. 41** Mask SEM images of VSB shot for three contact arrays with (a) conventional shots, (b) overlapping without MWCO, and (c) MWCO[23] (source: Micron).

curvilinear ILT mask patterns were actually produced by the VSB mask writer. The curvilinear SRAFs in the MWCO do have staircase jogs but writing with a smaller number of larger shots makes placement and CD uniformity better, and using overlapping shots increases the dose applied without increasing write-times, improving dose margin.

The chart in Fig. 42 is a write-time comparison presented by NuFlare, comparing write times between their VSB writer and their multibeam machine.[116] Because, according to this NuFlare chart, VSB mask write time is proportional to the number of shots, it is only when shot count is >200 Gshots/pass that VSB write times exceed 12 h; below the 200-Gshots/pass level, VSB write times are faster than 12 h even at 75 $\mu C/cm^2$. When this number is converted into shot density per square micron, it turns out the magic number is 36 shots/$\mu m^2$. If the shot density is below this number, the mask write time using a VSB mask writer (i.e., NuFlare EBM 9500) will be <12 h. D2S demonstrated that for all 121 different contact configurations in this example, MWCO can produce ILT for 193i with a shot density lower than 36 shots/$\mu m^2$, and therefore enables the entire curvilinear ILT mask to be written within 12 h.

EUV masks, particularly EUV masks with curvilinear ILT, would have higher shot density than can be written reasonably even with MWCO and overlapping shots. EUV more faithfully reproduces mask aberrations on the wafer, and technology nodes that use curvilinear ILT for EUV demand even more accuracy on wafer. This means that even an overlapping shot solution

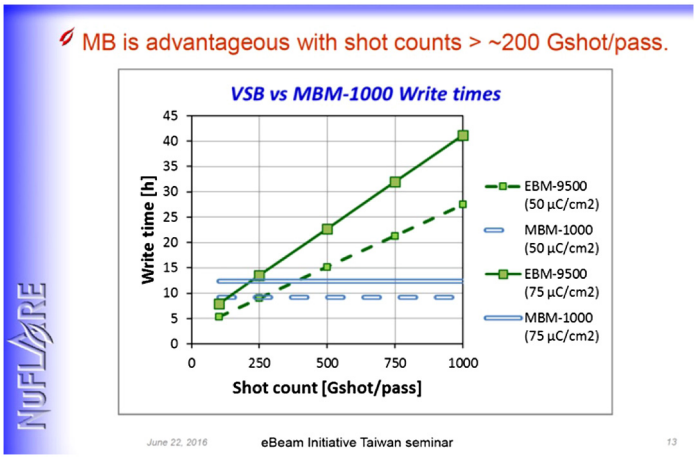


**Fig. 42** NuFlare's estimation of mask write time for their VSB mask writer and multibeam mask writer[116] (source: NuFlare).

for EUV would require more shot density than a proportionately scaled 193i mask shape. These factors, combined with EUV masks having more target shape density, reconfirm multibeam mask writing as most appropriate for EUV masks.

Micron produced VSB-written masks and wafers for 193i to validate the process window benefits of curvilinear ILT with MWCO. Figure 43 shows SEM images of the curvilinear ILT mask and its corresponding wafer print for an $11 \times 11$ contact array.[23] The contact ADI target is 40 nm, and the minimum pitch is 100 nm. One can see that the contact array was printed nicely with this MWCO curvilinear ILT mask: the contact holes are printed evenly across all the varied pitches and rotations, as well as from the center of the array all the way to the edge of the array, which is very challenging for OPC.

Micron together with D2S also demonstrated MWCO improved wafer process window by over 2× comparing to Micron POR OPC.[23] In this study, 61 different patterns—including some of the most challenging ones found in lithography and OPC—were selected. They were treated with conventional OPC VSB shots, curvilinear ILT using overlapping VSB shots without MWCO, and overlapping VSB shots with MWCO. All the patterns were written on the same mask using the NuFlare EBM-9500 PLUS VSB mask writer. The wafer was printed at seven different focuses and nine doses, for a total of 63 different process conditions.

Figure 44 shows the process window plot produced from this experiment. The $x$ axis is the focus, the $y$ axis is the dose change. Since there are 61 sites, the ratio of the number of sites meeting the process window requirement to the total number of sites was plotted.[23] A CD variation of 10% is used as the process window criteria. The pseudocolor from green to red represents process window from good to bad. Overall, curvilinear ILT using overlapping shots without MWCO and curvilinear ILT using overlapping shots with MWCO enlarged the green (or non-red) region by over 2×, especially the DoF. Comparing overlapping shots without MWCO and with MWCO, the MWCO is slightly better, showing the benefit of optimizing wafer EPE instead of mask EPE, while only using half the number of shots as the overlapping shots without MWCO case.

The use of MWCO along with GPU-accelerated curvilinear ILT removed the final roadblock to wide adoption of full-chip, curvilinear ILT.

## 7 Other Important Aspects of ILT and Future Development

With all the major roadblocks in the path of ILT adoption now cleared, we can now see some of the additional hurdles for research and development in the ILT arena.

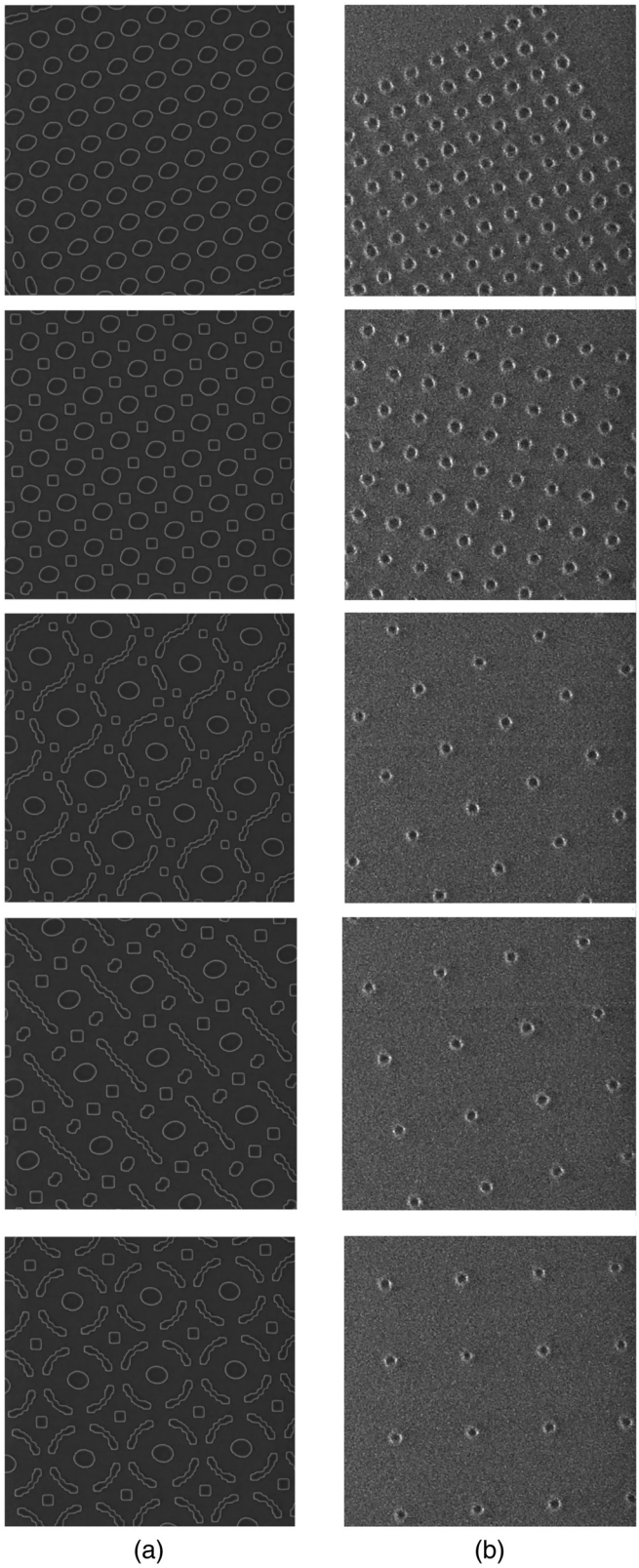


**Fig. 43** WCO results for an 11 × 11 contact array. In each pair, (a) MWCO VSB mask SEM images of curvilinear mask designs for different pitches and orientations; (b) SEM images of corresponding wafer print[23] (source: Micron).

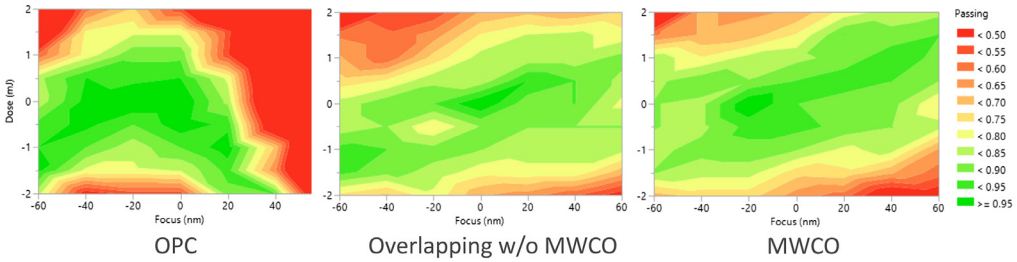


**Fig. 44** Process window plots for all 61 test patterns/sites at all 63 process conditions for OPC, Curvilinear ILT with overlapping shots without MWCO, and curvilinear ILT using overlapping shots with MWCO[23] (source: Micron).

## 7.1 *Litho Model for ILT*

No papers on lithography models for ILT have been published, because every company developing ILT solutions treats this subject as their trade secret. ILT models are similar to OPC models in terms of the optics and resist processes that must be simulated. However, there are some differences that make ILT modeling and simulation more challenging. The first one is the derivative of the cost function. Since ILT is an optimization, most optimization algorithms rely on a derivative of the cost function to make the optimization converge faster. The second difference is the all-angle nature of ILT: since curvilinear ILT models, unlike conventional OPC models, must consider all angles, not just horizontal and vertical edges, the fast-computing tricks often applied to processing Manhattan patterns cannot be applied.

Mask 3D models are another important aspect in lithography models for ILT. Intel pixelated mask ILT was first to use mask 3D models in ILT (Sec. 5.2). Because the mask is chromeless and each pixel is relatively small (100 nm × 100 nm), the mask 3D effect is strong. ASML Brion also introduced their freeform ILT engine with mask 3D models, and they have demonstrated it on a FLASH memory full-chip application.[121] Dr. Pearman from D2S also presented a fast mask 3D model for curvilinear ILT using DL.[122]

## 7.2 *Curvilinear Mask Rules*

There are certain physical and manufacturing requirements regarding the smallest features that a mask can have. For example, if a hole on chrome is too small, it may not be resolved completely in the mask writing and resist development/etching process. If a chrome piece is too small, it may peel off. For OPC or Manhattan mask shapes, the mask rules represent an approximation of these mask manufacturing constraints, mainly minimum line/space, minimum area, minimum corner-to-corner distance, and some combined rules for small features in certain aspect ratios.

Recently, after curvilinear full-chip ILT became a practical reality, mask rules for curvilinear ILT masks have become a hot topic.[118,123–129] One implementation of mask rules has been proposed by D2S, based on the idea of two circles.[118] As shown in Fig. 45, one small circle represents the minimum width/space checks, while one larger (or equal radius) circle represents the minimum radius of curvature for the 2D areas. Some mask shops may still want to separate checks into one-dimensional (1D) and 2D checks. 1D areas can still be defined by looking for long edges that have very low curvatures.

Conceptually, the curvature checks can be done by sliding circles around the edge of each boundary. Again, if there is any overlap between the circle and the contour of the pattern while rolling the circle, the curvature is too high and is not reliably manufacturable.

## 7.3 *Curvilinear Mask Inspection*

Mask inspection is another common concern for curvilinear ILT masks. For die-to-die inspections using algorithms based on intensity differences, details of the shapes of mask patterns do not really matter. In some respects, the images of smoothly curved ILT patterns may exhibit less noise due to the absence of sharp corners that modulate the intensity at high spatial frequencies.

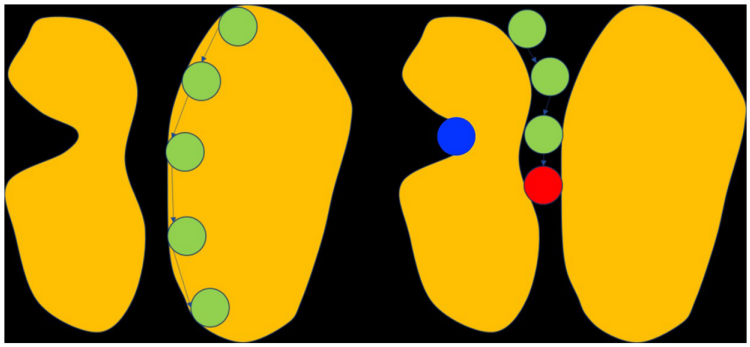


**Fig. 45** (a) Internal (width) checking. (b) External (space) checking. Conceptually, as long as the minimum circle can slide entirely within and around the features, it should meet the minimum width and space checks. This example fails on the space check as indicated in red and fails the minimum curvature test for the concave area as indicated in blue[118] (source: D2S).

In addition, die-to-database inspection is easier when the specified shapes are more manufacturable on mask. No "corner rounding" calculation is needed in the database preparation because the shapes are already rounded. This means that there will be fewer errors flagged because of the discrepancy between the corner-rounding models used versus the actual corner rounding.

The important consideration for ILT is defect disposition, which may become less intuitive when mask patterns are different than design targets. This is not unique to ILT masks, however; any mask with aggressive OPC or mask using SRAFs presents the same challenge. Smaller features on masks produce lower contrast on high-resolution inspection images and so are harder to differentiate from noise. There are two ways to solve this problem: improve the resolution of the mask-inspection image or perform defect disposition at the aerial or wafer plane instead of the mask plane. Numerical Technology pioneered this field with the i-Virtual Stepper System™.[74–83] (later acquired by Synopsys, also led by the author). The Applied Materials AERA™ inspection system is an aerial-image-based mask-inspection system.[84–89,130–134] KLA has introduced WPI/API to examine the impact on the wafer as a way to filter defects on mask.[135–137] Brion Technologies and NuFlare developed off-line lithography simulation software for the NuFlare mask inspection system.[90,91] Luminescent Technologies, after it sold its ILT business to Synopsys, also developed its LAIPH™ Litho Plane Reviewer system targeting for ILT mask defect dispositioning.[92–96]

## 7.4 *Applying Deep Learning to ILT*

Inspired by many success stories of ML in a broad range of artificial intelligence applications, both industrial and academic researchers are now actively developing ML solutions for challenging problems in computational lithography, including ILT.[99–101,121,138–140]

One of the first papers applying DL to ILT is from ASML Brion. Their 2017 paper shows how they use their freeform ILT engine to train an ILT DL model using a convolutional neural network (CNN) (Fig. 46)—a typical deep neural network commonly used in object recognition.[99] Once this DL ILT model is trained, it is used to create the SRAF initial-seeding map and also as the initial seeding of their ILT engine. Such DL-based SRAF generation has been demonstrated in both DRAM and logic full-chip applications[100,101] and FLASH memory full-chip application.[121]

While most research work has focused on using DL to accelerate ILT, Dr. Peng Liu from Synopsys showed DL can create ILT solutions all by itself, plus all models—including neural network-based 3D mask, imaging, and resist models—required for ILT.[139] He demonstrated a standalone mask synthesis flow that runs entirely on the TensorFlow® ML platform with a reinforcement learning (RL) approach and GPU acceleration. It is interesting to see that RL without optimization can actually generate ILT mask patterns that look quite similar to optimization-based ILT results (Fig. 47).

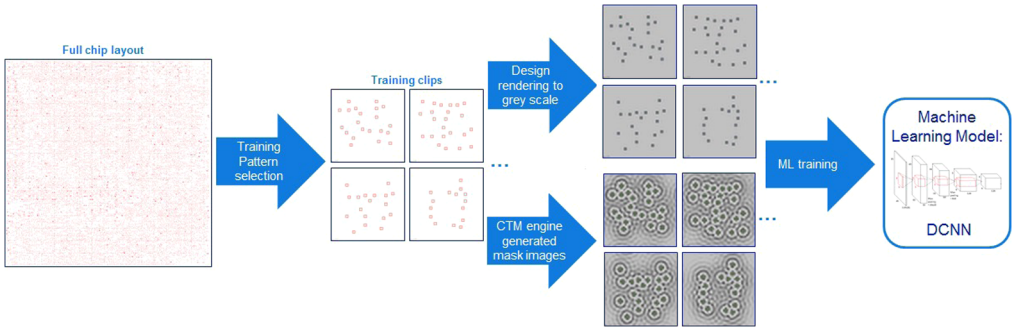


**Fig. 46** Using the input and output from a freeform ILT engine as the input and output of a CNN to train the DL model[99] (source: ASML Brion).

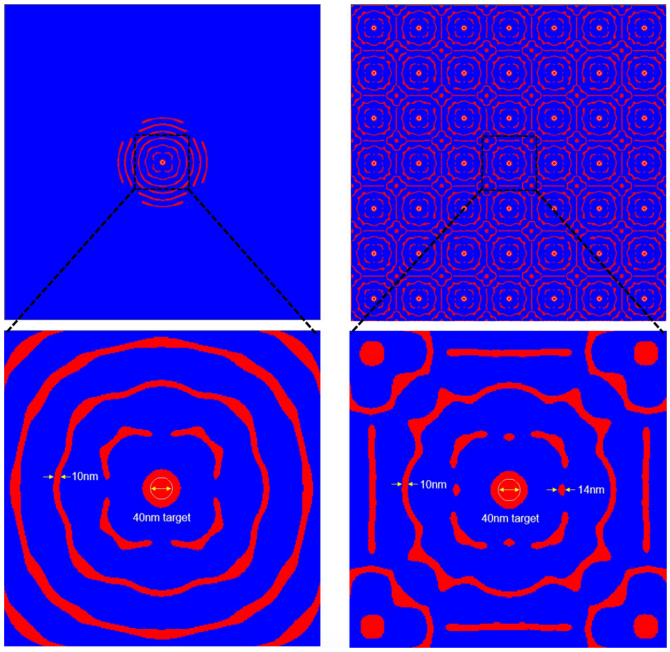


**Fig. 47** Curvilinear ILT mask patterns for both isolated contact and contact array generated by RL without using ILT data to train[139] (source: Synopsys).

Another area of DL applied to ILT is digital twins. A digital twin is a digital replica of a living or nonliving entity or a simulation-based system. By bridging the original (called "ground truth" in DL terminology) and the virtual world, data are transmitted seamlessly allowing the virtual entity to exist simultaneously with the original entity. For the whole mask and wafer ecosystem to be fully ready to handle curvilinear data, each step in the manufacturing process must test its equipment and processes using a large number of curvilinear mask designs to prepare for high-volume production. Since running ILT is compute-intensive for test-data generation for other process steps, a digital twin of ILT that runs quickly can be of benefit.

The author presented an ILT digital twin developed by D2S using its deep learning kit (DLK).[141] Figure 48 shows three examples of curvilinear ILT results from the D2S ILT solution (right) and results from its digital twins (left). Although the ILT digital twin results cannot be used for wafer printing because it may not produce results that meet EPE and process window requirements, its mask pattern shapes are very close to the curvilinear ILT result. It is definitely adequate for mask equipment to use for testing. For example, it can be used for suppliers of multibeam mask writers to test their writing capability and accuracy, or it can be used by suppliers of mask inspection tools to test their inspection algorithms for curvilinear masks.

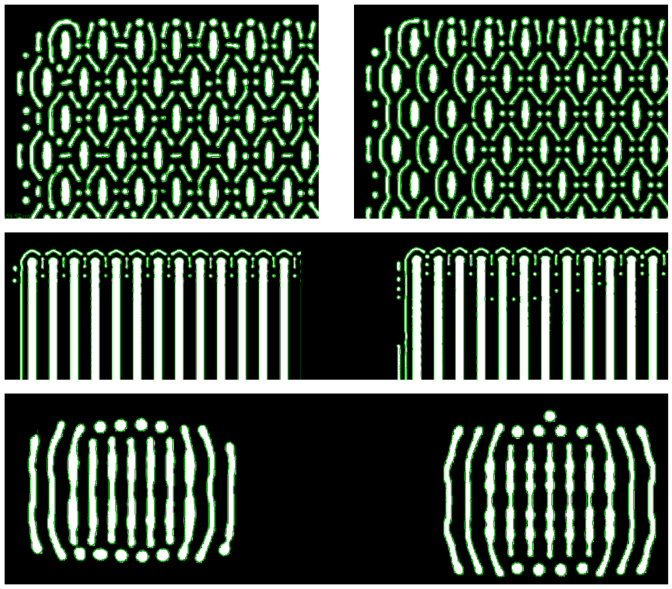

**Fig. 48** Examples of curvilinear ILT digital twin and the real curvilinear mask pattern generated from D2S ILT[141] (source: D2S).

Digital twins can also be used in a chain: for a DL project that requires curvilinear ILT mask SEM images to do training, one can use curvilinear ILT digital twin to generate curvilinear ILT mask patterns, then use mask SEM digital twins to generate the mask SEM images.[142]

## 7.5 *Curvilinear Mask Format*

Another area related to curvilinear ILT is the curvilinear mask format. Curvilinear ILT mask patterns can be represented by all-angle short segments in GDSII and OASIS. However, the data volume for a full-chip curvilinear ILT solution is high compared with Manhattan patterns. There have been three papers to date presented that propose modifications to the existing file format to fit curvilinear mask data better. Discussions thus far focus on splines and similar expressions of curves that are smooth to infinite resolution, while containing the data size to be better than a piecewise linear polygon that has, as an example, a vertex every 1 nm on mask dimensions. Dr. Frank Abboud from Intel first discussed curvilinear mask format in his invited paper in 2014.[143] Jin Choi from Samsung proposed to use control points to control the curve.[128] NuFlare proposed their second-generation mask data format for multibeam mask writers for curvilinear mask patterns.[144]

## 7.6 *Curvilinear Design*

IMEC presented a paper at SPIE Advanced Lithography 2020 showing the possibility of allowing curvilinear design to reduce the number of metal layers, shrink the design, and increase the transistor density.[145] Curvilinear ILT not only produces curvilinear mask shapes but also the approach fundamentally enables curvilinear target wafer designs. There are short-cuts that have been incorporated in approaches that assumed primarily Manhattan targets that will need to be avoided when targeting curvilinear wafer shapes. Now that curvilinear masks are practical because of multibeam mask writing, curvilinear ILT is practical. Because curvilinear ILT is practical, target wafer shapes that are curvilinear are practical to manufacture. Designing with large amounts of curvilinear shapes is not yet practical, however, because the CAD infrastructure does not support mass-scale use of curvilinear shapes. Selective use of curvilinear target shapes for critical areas may be practical as mentioned by Ezequiel Russell of Micron.[104] This has the potential to bring a paradigm shift to the entire EDA industry.

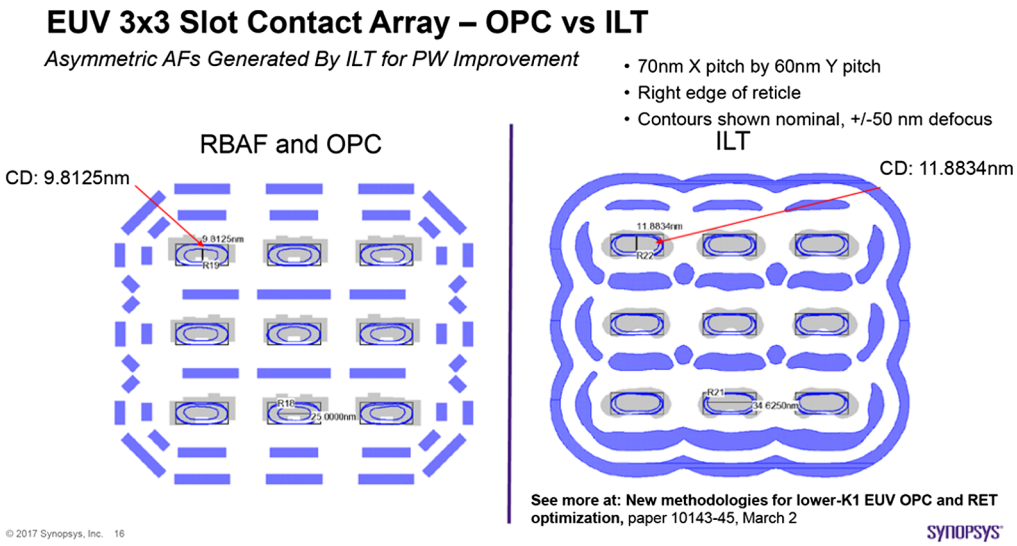


**Fig. 49** Simulation comparison of OPC with rule-based SRAF and ILT with asymmetric AF: ILT corrects the image contour unbalance problem and improves the PV band[105,146] (source: Synopsys).

### 7.7 *ILT for EUV*

Although all the work shown so far on ILT is for 193i lithography, ILT can be extended to EUV lithography. For the first and second generations of EUV, OPC is relaxed due to the large reduction of the wavelength from 193 nm immersion to the 13.5 nm wavelength of EUV. However, for processes at 3 nm and beyond, even EUV requires high numerical aperture (NA) or even multiple patterning. ILT will be helpful for EUV in these generations. In addition, line-edge roughness is a big concern for EUV, and multibeam mask writers are already required due to the high pattern density, and slower mask resists that are needed for resolution requirements. So curvilinear masks are required and curvilinear ILT is more desirable than Manhattan OPC. Tom Cecil from Synopsys presented a good example (Fig. 49), which shows that using curvilinear ILT with asymmetric assist features to correct the EUV contour unbalance problem greatly improved the process variation (PV) band compared with OPC with rule-based assist features.[105,146]

Since the computing grid for ILT in EUV would need to be between 2× and 2.5× denser in each dimension as compared to 193i, the runtime for ILT for the same given area will increase the computing needs (either amount of computing power or elapsed time) by 4× to ∼6×. With high NA EUV, only half the mask will be computed at a time. This may reduce the scaling to between 2× and ∼3×. When curvilinear ILT is required for EUV, computing will also scale. While CPU clock speed has not scaled, GPU computing bandwidth has scaled well and is anticipated to continue to scale because GPU computing scales by the number of cores doing the computing in the same chip versus clock speed. From 2011 to 2020, server-based general-purpose GPUs have gone from 4 TFLOPS to 37 TFLOPS in single-precision computing for a compound annual growth rate of 25%/year. A GPU-based ILT solution that scales well beyond a multiple-rack server will be capable of full-chip curvilinear ILT for EUV by the time it is needed.

## 8 Author's Predictions for the Future of ILT

This is primarily a review paper, but it will end with some predictions for ILT in the future.

### 8.1 *ILT Mask Datapath Will Stay in Pixel Space*

As discussed in Sec. 7.4, currently curvilinear mask formats are a hot topic. Multiple organizations have proposed similar ways to enhance the OASIS file format to accurately represent curves. All such efforts are good. In addition, there is the possibility, introduced by NuFlare's

Noriaki Nakayamada-san at Lithography Workshop,[147] that an alternative to a curvy format using spline-like expressions could be used with compressed-pixel data.

ILT is computed by pixels, curvilinear masks are written with pixels, all masks are inspected by pixel images, and all metrology is done with pixel images. Since practically everything in the mask-manufacturing flow is based on pixels, perhaps the industry should be gradually shifting over to converting from edge-based space to pixel-based space in the ILT step, then staying in the pixel space from there on. With this approach, mask data size is predictable regardless of design type and resolution is guaranteed to be sufficient (because mask-writer pixel size is known). Another benefit of this approach is that spurious subpixel artifacts will not print and so are not a concern. Just as multibeam mask writers write in constant time for any mask with a given resist sensitivity, regardless of shape or edge count, a pixel-based datapath would have a fixed runtime regardless of shape complexity. While shapes that do not require OPC or require only very simple OPC will take much longer to process, layers created with curvilinear ILT will benefit by staying in a pixel-based datapath. In the short term, it is likely that curvilinear ILT will be applied only to layers or hotspots that require it or that will reap the greatest benefit, so perhaps there might be a hybrid approach where only curvilinear ILT portions of the mask stay in pixel space.

Transmission of pixel data is another piece of the mask format puzzle. A multibeam mask writer with 10 nm pixels writing a 132 mm × 104 mm mask area needs 137 terapixels worth of pixel data per writing pass. Even if a 10× compression rate can be achieved, and 8 bits per pixel is sufficient, that is still more than 10 terabytes of data per writing pass. For 193i masks, an average chip size is much less than the full reticle size, so that helps to reduce the size. But for EUV masks, since lithographic effects act differently on each instance of the chip in one dimension, file sizes, and corresponding processing times in the mask shop datapath would still be an issue if mask-writer resolutions are used. However, because the information contained in the mask format is what was computed by ILT, the format expressing the mask shapes only requires the resolution of the ILT grid, not that of the mask writer grid. There is an opportunity for further work to explore transmission of pixel data compactly and without loss through the mask manufacturing datapath.

## 8.2 *Deep Learning Will Push GPU-Based Computing Platforms into the Mainstream*

Software for semiconductor manufacturing used to run on CPU farms. All semiconductor manufacturing companies have invested in computer farms with tens of thousands of CPUs or even hundreds of thousands of CPUs. They all want the flexibility to run OPC or ILT on their entire CPU farm when needed. This investment made contemplating a switch to GPU-based computing difficult. The recent boom in DL has changed this. GPU is the *de facto* platform for DL training, and as EDA software companies try to leverage DL, they are making the investment in GPU computing. Not only is GPU-based computing already 10 or even 100 times faster than CPU for SIMD-type computing but also its power increases faster than CPU-based computing according to Huang's Law.[148]

The existing CPU-based computer farms are good for many current tasks. But there is a clear need to invest in new computing hardware for the next-generation nodes. This is especially true for the leading-edge fabs, since the computing power required for EUV is several times more than DUV. This looming reality has made leveraging existing computing power less important than assuring sufficient computing power to support these fast-approaching changes. The semiconductor manufacturing industry is very conservative, but eventually it will follow the trend of high-performance computing and data centers where GPU-accelerated computing becomes the mainstream. It is just a matter of time.

However, we should be clear that "GPU-based computing" does not mean GPU only. The shift is to GPU-accelerated computing or GPUs paired with CPUs. The key is to keep GPUs busy at all times with SIMD tasks because GPUs are faster than CPUs at SIMD computing. Keeping GPUs busy requires optimizing algorithms to have as much of the computing that is suitable for GPUs stay in GPUs.

### 8.3 *Multibeam Mask Writing Will Accelerate the Adoption of Curvilinear ILT*

When we look at OPC versus ILT, we are really looking at an edge-based manipulation of mask shapes (OPC) versus pixel-based manipulation (ILT). OPC works the way it works because of the assumption of a Manhattan constraint. Its purpose is to defeat the natural corner-rounding and shortening of Manhattan shapes in advanced nodes. OPC has been used to approximate curvilinear shapes on mask; however, because it is edge-based, it requires many small rectilinear features to create a curved shape, which has led to prohibitive runtimes for this approach.

ILT is pixel-based and shape-agnostic. Its purpose is to find the best mask solution for the desired wafer shape. With the introduction of multibeam mask writers, which are also shape-agnostic in terms of write time, and with the even more recent introduction of MWCO,[23] which enables full-chip ILT with VSB mask writers within practical write times for 193i masks, the superior ILT solution becomes an easy choice.

OPC was the right overall tradeoff in the era of VSB writing and CPU-only computing to find the optimized mask solution to print the target wafer pattern, given Manhattan constraints. ILT is the best method for finding the optimized mask solution if you remove the Manhattan constraints. The evidence that curvilinear mask shapes are more reliably manufacturable[127,129,149] will lead the industry away from Manhattan assumptions and toward ILT. Dr. Ezequiel Russell from Micron showed that rectangular contacts have more visible variations on mask compared with oval contacts. By replacing rectangles with ovals for contacts, wafer CD uniformity was improved by 10% (Fig. 50).[149] In general, the curvilinear features produced by curvilinear ILT have a similar trend.

### 8.4 *193i ILT and Lithography Will Benefit Greatly from MWCO using Overlapping Shots*

There is no question that the multibeam mask writer is the right mask writer to write curvilinear ILT masks: it is clearly faster for complex shapes and has a constant write time regardless of shape. For EUV masks, where the additional need for accuracy requires lower-sensitivity resists, which in turn require longer exposure times in the mask writer, the multibeam mask writer is a necessity. Once the multibeam mask writer is chosen as the writer for EUV masks, mask write times, which dominate the mask shop's economics, will be the same for curvilinear masks or

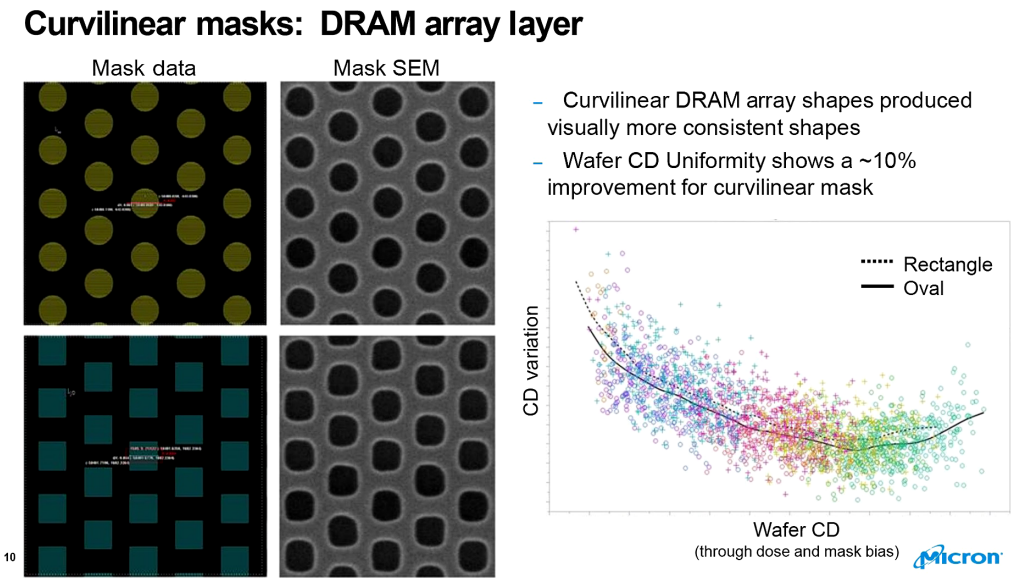


**Fig. 50** Mask and wafer results show oval contact shapes have more consistent shapes on both mask and wafer than rectangular contact shapes. CD variations were measured on many contacts over many local regions at different process conditions. The average shows the oval contacts show a 10% improvement in CD uniformity on wafer over rectangle contacts[149] (source: Micron).

masks dominated by Manhattan shapes. Since resilience to manufacturing variation on the wafer is superior with curvilinear masks,[127,129,149] there will be no reason not to use curvilinear masks for EUV.

However, VSB mask writers will still be the work horse for 193i lithography. MWCO using overlapping shots enables the VSB mask writer to write curvilinear ILT masks within a practical write time. The shapes on mask will not be as "curved" as what can be written with a multibeam mask writer, but for 193i lithography, that does not matter because 193i is blind to subresolution differences on mask. As far as 193i can "see," VSB using MWCO can create equal-quality masks with a reasonable VSB shot count, especially by taking advantage of overlapping shots, making the mask write times practical. MWCO embeds VSB shot determination during ILT to optimize the wafer quality that can be created with the mask shapes produced by overlapping VSB shots. The shapes that a VSB-based mask shop can practically produce in reasonable mask write times are optimized for wafer quality.

### 8.5 *Pixel-Based Inline Dose Correction on Multibeam Mask Writers Will Be the Ultimate MPC for Curvilinear ILT Mask*

Curvilinear ILT will also bring a paradigm shift to MPC. Conventional MPC is very similar to OPC, and it moves the edges of post-OPC Manhattan patterns to correct the mask proximity effects. Unlike OPC for lithography, where the only degree of freedom once the mask type is fixed, is the mask pattern shape, the mask writer has another degree of freedom in the dose or the amount of energy used to write the mask. Multibeam mask writers have the ability to use multilevel doses to define the edge position down to 0.1 nm addressability with an approximate 10 nm pixel size; therefore, multibeam mask writers can perform MPC using dose variation to move the edges of mask patterns. In addition, increasing dose can make the edge slope steeper. For small features, such as SRAFs, moving the edge will not provide sufficient correction, since the dose for such features is so small, but increasing dose can make an SRAF print more reliably. In addition, multibeam mask writers can improve the line-edge roughness and CD accuracy of features of all sizes by enhancing doses of pixel beams near the contour edge.

The ultimate MPC for curvilinear ILT masks is a pixel-based inline dose correction, because it eliminates off-line MPC from the mask house workflow, reduces precious mask making time, and produces superior MPC results over normal edge-based MPC. Such inline correction is already implemented in NuFlare's multibeam mask writers, where it is called pixel-level dose correction (PLDC): it does dose correction in the pixel domain inside the mask writer datapath in real time by using GPU-acceleration. PLDC does not increase the mask write time, significantly shortens the overall mask turnaround time by eliminating off-line MPC, and produces superior mask quality especially for curvilinear masks.[150–152]

## 9 Summary and Conclusions

The semiconductor industry has long recognized the value of ILT in addressing the challenges of advanced-node lithography. For more than 30 years, research and development teams across the industry have worked to create ILT solutions that would overcome the biggest roadblocks to wide adoption of this important lithography technology: ILT solution runtime, VSB mask write time, and mask manufacturing worries. Many teams from many companies, as well as academia, have played important roles in the development of ILT and in the removing these roadblocks.

The ability to support ILT in high-volume production will be key for the semiconductor industry moving forward. EUV will require curvilinear ILT at 3 nm node and beyond. The drive for design density may drive curvilinear design to replace Manhattan design eventually. The process window improvements afforded by ILT are welcome in any manufacturing flow.

The good news is that mask metrology, mask inspection, mask review, and mask repair equipment providers have all confirmed in panel discussions that their equipment is ready for ILT, at least in terms of technologies or principles. With the major roadblocks of runtime and mask writing removed for 193i, either with multibeam or VSB mask writers, there is a clear road ahead for ILT to develop a central role in lithography's future.

## References


1. B. E. A. Saleh and S. I. Sayegh, "Reductions of errors of microphotographic reproductions by optical corrections of original masks," *Opt. Eng.* **20**(5), 781–784 (1981).
2. K. M. Nashold and B. E. A. Saleh, "Image construction through diffraction-limited high-contrast imaging systems: an iterative approach," *J. Opt. Soc. Am. A* **2**, 635 (1985).
3. Y. Liu and A. Zachor, "Optimal binary image design for optical lithography," *Proc. SPIE* **1264**, 410–412 (1990).
4. Y. Liu and A. Zachor, "Binary and phase-shifting image design for optical lithography," *Proc. SPIE* **1463**, 382–399 (1991).
5. A. Rosenbluth et al., "Optimum mask and source patterns to print a given shape," *Proc. SPIE* **4346**, 13–30 (2002).
6. Y.-T. Wang et al., "Automated design of halftoned double-exposure phase-shifting masks," *Proc. SPIE* **2440**, 290–301 (1995).
7. S.-H. Jang et al., "Manufacturability evaluation of model-based OPC masks," *Proc. SPIE* **4889**, 520 (2002).
8. T. Fuhner and A. Erdmann, "Improved mask and source representations for automatic optimization of lithographic process conditions using a genetic algorithm," *Proc. SPIE* **5754**, 415–426 (2005).
9. Y. Liu et al., "Inverse lithography technology principles in practice: unintuitive patterns," *Proc. SPIE* **5992**, 599231 (2005).
10. L. Pang et al., "Laser and e-beam mask-to-silicon with inverse lithography technology (ILT)," *Proc. SPIE* **5992**, 599221 (2005).
11. J. Ho et al., "Real-world impact of inverse lithography technology," *Proc. SPIE* **5992**, 59921Z (2005).
12. A. Balasinski et al., "Inverse lithography technology: verification of SRAM cell pattern," *Proc. SPIE* **5992**, 599230 (2005).
13. C. Hung et al., "First 65 nm tape-out using inverse lithography technology (ILT)," *Proc. SPIE* **5992**, 59921U (2005).
14. P. Martin et al., "Manufacturability study of masks created by inverse lithography technology (ILT)," *Proc. SPIE* **5992**, 599235 (2005).
15. D. Abrams and L. Pang, "Fast inverse lithography technology," *Proc. SPIE* **6154**, 61541J (2006).
16. C. Y. Hung et al., "Pushing the lithography limit: applying inverse lithography technology (ILT) at 65 nm generation," *Proc. SPIE* **6154**, 61541M (2006).
17. B. Lin et al., "Inverse lithography technology at chip scale," *Proc. SPIE* **6154**, 615414 (2006).
18. B. G. Kim et al., "Trade-off between inverse lithography mask complexity and lithographic performance," *Proc. SPIE* **7379**, 73791M (2009).
19. W. Sim et al., "Hotspot fixing using ILT," *Proc. SPIE* **7973**, 79731L (2011).
20. K. Selinidis et al., "Resist 3D aware mask solution with ILT for hotspot repair," *Proc. SPIE* **10147**, 101470Q (2017).
21. L. Pang et al., "Study of mask and wafer co-design that utilizes a new extreme SIMD approach to computing in memory manufacturing: full-chip curvilinear ILT in a day," *Proc. SPIE* **11148** (2019).
22. L. Pang et al., "TrueMask ILT MWCO: full-chip curvilinear ILT in a day and full mask multi-beam and VSB writing in 12 hrs for 193i," *Proc. SPIE* **11327**, 113270K (2020).
23. L. Pang et al., "Enabling faster VSB writing of 193i curvilinear ILT masks that improve wafer process windows for advanced memory applications," *Proc. SPIE* **11518**, 115180W (2020).
24. S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: algorithms based O Hamilton–Jacobi formulations," *J. Comput. Phys.* **79**, 12–49 (1988).
25. Y. Granik et al., "Fast pixel-based mask optimization for inverse lithography," *J J. Micro/Nanolithogr. MEMS MOEMS* **5**(4), 043002 (2006).
26. Y. Granik, "On the uniqueness of optical images and solutions of inverse lithographical problems," *J. Micro/Nanolithogr. MEMS MOEMS* **8**, 031405 (2009).

27. Y. Borodovsky, "Enabling Moore's law through computational lithography," in *Lithogr. Workshop* (2007).
28. Y. Borodovsky, et al., "Pixelated phase mask as novel lithography RET," *Proc. SPIE* **6924**, 69240E (2008).
29. V. Singh et al., "Making a trillion pixels dance," *Proc. SPIE* **6924**, 69240S (2008).
30. W. Cheng et al., "Fabrication of defect-free full-field pixelated phase mask," *Proc. SPIE* **6924**, 69241G (2008).
31. R. Schenker et al., "Integration of pixelated phase masks for full-chip random logic layers," *Proc. SPIE* **6924**, 69240I (2008).
32. I. Torunoglu et al., "OPC on a single desktop: a GPU-based OPC and verification tool for fabs and designers," *Proc. SPIE* **7641**, 764114 (2010).
33. I. Torunoglu et al., "A GPU-based full-chip inverse lithography solution for random patterns," *Proc. SPIE* **7641**, 764115 (2010).
34. A. Poonawala and P. Milanfar, "OPC and PSM design using inverse lithography: a nonlinear optimization approach," *Proc. SPIE* **6154**, 61543H (2006).
35. A. Poonawala and P. Milanfar, "Double-exposure mask synthesis using inverse lithography," *J. Micro/Nanolithogr. MEMS MOEMS* **6**, 043001 (2007).
36. A. Poonawala et al., "ILT for double exposure lithography with conventional and novel materials," *Proc. SPIE* **6520**, 65202Q (2007).
37. L. Pang et al., "Inverse lithography technology (ILT) and its applications at 65 nm and beyond," in *Proc. ISTC* (2006).
38. L. Pang et al., "Inverse lithography technology (ILT) and its readiness for production in advanced technology nodes," in *Proc. ISTC* (2007).
39. L. Pang et al., "Applying inverse lithography technology (ILT) to develop the best lithography strategy and design rules for advanced technology nodes," in *Proc. ISTC* (2008).
40. L. Pang et al., "Inverse lithography technology (ILT) enabled source mask optimization (SMO)," in *Proc. ISTC* (2009).
41. L. Pang et al., "From computational lithography to computational inspection: inverse lithography technology (ILT) and inverse inspection technology (IIT)," in *Proc. CSTIC* (2010).
42. Y. Yang et al., "Hot-spots aware inverse lithography technology," in *Proc. CISTC* (2009).
43. Y. Yang et al., "Seamless-merging-oriented parallel inverse lithography technology," *J. Semicond.*, **30**(10), 106002 (2009).
44. J. Zhang et al., "A highly efficient optimization algorithm for pixel manipulation in inverse lithography technique," in *IEEE/ACM Int. Conf. Comput.-Aided Design*, IEEE, pp. 480–487 (2008).
45. J. Zhang et al., "GPU-accelerated inverse lithography technique," *Proc. SPIE* **7379**, 73790Z (2009).
46. A. Wong and E. Lam, "Computation lithography: virtual reality and virtual virtuality," in *Proc. ISTC* (2009).
47. E. Lam et al., "Regularization in inverse lithography: enhancing manufacturability and robustness to process variations," in *Proc. CSTIC* (2010).
48. N. Jia and E. Y. Lam, "Machine learning for inverse lithography: using stochastic gradient descent for robust photomask synthesis," *J. Opt.* **12**(4), 045601 (2010).
49. X. Ma and G. Arce, "Generalized inverse lithography methods for phase-shifting mask design," *Proc. SPIE* **6520**, 65200U (2007).
50. X. Ma and G. Arce, "Binary mask optimization for inverse lithography with partially coherent illumination," *Proc. SPIE* **7140**, 71401A (2008).
51. X. Ma and G. Arce, "PSM design for inverse lithography with partially coherent illumination," *Proc. SPIE* **7274**, 727437 (2009).
52. S. Shen et al., "Enhanced DCT2-based inverse mask synthesis with initial SRAF insertion," *Proc. SPIE* **7122**, 712241 (2008).
53. J. Yu et al., "Model-based sub-resolution assist features using an inverse lithography method," *Proc. SPIE* **7140**, 714014 (2008).
54. L. Pang et al., "Validation of inverse lithography technology (ILT) and its adaptive SRAF at advanced technology nodes," *Proc. SPIE* **6924**, 69240T (2008).

55. G. Xiao et al., "Affordable and process window increasing full chip ILT masks," *Proc. SPIE* **7823**, 78233T (2010).
56. G. Xiao et al., "E-beam writing time improvement for inverse lithography technology mask for full-chip," *Proc. SPIE* **7748**, 77481T (2010).
57. S. Jun et al., "Improvement of KrF contact layer by inverse lithography technology with assist feature," *Proc. SPIE* **7748**, 77481V (2010).
58. L. Pang et al., "Optimization from design rules, source and mask, to full chip with a single computational lithography framework: level-set-methods-based inverse lithography technology (ILT)," *Proc. SPIE* **7640**, 76400O (2010).
59. L. Pang et al., "Source mask optimization (SMO) at full chip scale using inverse lithography technology (ILT) based on level set methods," *Proc. SPIE* **7520**, 75200X (2009).
60. L. Pang et al., "Considering MEEF in inverse lithography technology (ILT) and source mask optimization (SMO)," *Proc. SPIE* **7122**, 71221W (2008).
61. G. Xiao et al., "Source optimization and mask design to minimize MEEF in low k1 lithography," *Proc. SPIE* **7028**, 70280T (2008).
62. V. Tolani et al., "Source-mask co-optimization (SMO) using level set methods," *Proc. SPIE* **7488**, 74880Y (2009).
63. T. Dam et al., "Source-mask optimization (SMO): from theory to practice," *Proc. SPIE*, **7640**, 764028 (2010).
64. S. L. Prins et al., "Inverse lithography as a DFM tool: accelerating design rule development with model-based assist feature placement, fast optical proximity correction and lithographic hotspot detection," *Proc. SPIE* **6925**, 69250E (2008).
65. S. Chang et al., "Exploration of complex metal 2D design rules using inverse lithography," *Proc. SPIE* **7275**, 72750D (2009).
66. J. Blatchford, "Litho/design co-optimization and area scaling for the 22-nm logic node," in *Proc. CSTIC* (2010).
67. B. U. Cho et al., "Evaluation of inverse lithography technology for 55 nm-node memory device," *Proc. SPIE* **6924**, 692438 (2008).
68. L. Pang et al., "Inverse lithography technology (ILT), what is the impact to photomask industry?" *Proc. SPIE* **6283**, 62830X (2006).
69. L. Pang et al., "Inverse lithography technology (ILT): keep the balance between SRAF and MRC at 45 and 32 nm," *Proc. SPIE* **6730**, 673052 (2007).
70. A. Fujimura et al., "Best depth of focus on 22-nm logic wafers with less shot count," *Proc. SPIE* **7748**, 77480V (2010).
71. A. Fujimura et al., "Writing 32 nm-hp contacts with curvilinear assist features," *Proc. SPIE* **7823**, 78230R (2010).
72. B. G. Kim et al., "Inverse lithography (ILT) mask manufacturability for full-chip device," *Proc. SPIE* **7379**, 73791M (2009).
73. L. Pang et al., "Defect printability analysis on alternating phase-shifting masks," *Proc. SPIE* **4754**, 614 (2002).
74. L. Cai et al., "Enhanced dispositioning of reticle defects using the virtual stepper with automated defect severity scoring," *Proc. SPIE* **4409**, 467 (2001).
75. C.-H. Chang et al., "Defect dispositioning using mask printability analysis on alternating phase-shifting masks," *Proc. SPIE* **4754**, 622 (2002).
76. L. Pang et al., "Simulation-based defect printability analysis on alternating phase shifting masks for 193 nm lithography," *Proc. SPIE* **4889**, 947 (2002).
77. S.-Y. Chiou et al., "Defect printability analysis study using virtual stepper system in a production environment," *Proc. SPIE* **4689**, 23 (2002).
78. Y. Morikawa et al., "Study of defect printability analysis on alternating phase shifting masks for 193 nm lithography," *Proc. SPIE* **4889**, 922 (2002).
79. L. Pang et al., "Enhanced dispositioning of reticle defects for advanced masks using virtual stepper with automated defect severity scoring," *Proc. SPIE* **5256**, 461 (2003).
80. J. Lu et al., "Application of simulation-based defect printability analysis at mask qualification control," *Proc. SPIE* **5038**, 33 (2003).

81. Y. Nagamura et al., "Photomask quality assessment strategy at 90-nm technology node with aerial image simulation," *Proc. SPIE* **5130**, 476 (2003).
82. L. Pang et al., "Simulation-based mask quality control in a production environment," *Proc. SPIE* **5375**, 1087 (2004).
83. L. Pang et al., "Photomask disposition based on simulated device performance," *Proc. SPIE* **5375**, 1183 (2004).
84. R. Liebe et al., "Aerial image-based mask inspection: a development effort to detect what might impact printing image quality on wafers," *Proc. SPIE* **5038**, 177 (2003).
85. S. Hemer et al., "Aerial-image-based off-focus inspection: lithography process window analysis during mask inspection," *Proc. SPIE* **5256**, 500 (2003).
86. A. Rosenbusch et al., "Aerial-image based inspection of AAPSM for 193-nm lithography generation," *Proc. SPIE* **5130**, 375 (2003).
87. Y. Zabar et al., "Aera193i: aerial imaging mask inspection for immersion lithography," *Proc. SPIE* **6518**, 65183G (2007).
88. D. Rost et al., "Qualification of aerial image 193 nm inspection tool for all masks and all process steps," *Proc. SPIE* **7028**, 70282Q (2008).
89. H. Baik et al., "Practical application of aerial imaging mask inspection for memory devices," *Proc. SPIE* **7028**, 70281G (2008).
90. G. Chen et al., "Defect printability analysis by lithographic simulation from high resolution mask images," *Proc. SPIE* **7488**, 74880A (2009).
91. G. Chen et al., "Mask-LMC: lithographic simulation and defect detection from high-resolution mask images," *Proc. SPIE* **7379**, 73791B (2009).
92. C. Y. Chen et al., "Mask defect auto disposition based on aerial image in mask product," *Proc. SPIE* **7379**, 73791F (2009).
93. J.-H. Park et al., "Mask pattern recovery by level set method based inverse inspection technology (IIT) and its application on defect auto disposition," *Proc. SPIE* **7488**, 748809 (2009).
94. L. Pang et al., "Computational inspection applied to a mask inspection system with advanced aerial imaging capability," *Proc. SPIE* **7638**, 76380V (2010).
95. L. Pang et al., "Computational lithography and inspection (CLI) and its applications in mask inspection, metrology, review, and repair," *Proc. SPIE* **7823**, 78232H (2010).
96. V. Tolani et al., "Lithographic plane review (LPR) for sub-32 nm mask defect disposition," *Proc. SPIE* **7823**, 78232G (2010).
97. J. Blatchford et al., "Improving 22 nm design space with source/design optimization," Solid State Technology (SST) (2010).
98. A. Tritchkov et al., "Use of ILT-based mask optimization for local printability enhancement," *Proc. SPIE* **9256**, 92560X (2014).
99. S. Wang et al., "Machine learning assisted SRAF placement for full chip," *Proc. SPIE* **10451**, 104510D (2017).
100. S. Wang et al., "Efficient full-chip SRAF placement using machine learning for best accuracy and improved consistency," *Proc. SPIE* **10587**, 105870N (2018).
101. K.-Y. Chen et al., "Full-chip application of machine learning SRAFs on DRAM case using auto pattern selection," *Proc. SPIE* **10961**, 1096108 (2019).
102. K. Hooker et al., "ILT optimization of EUV masks for sub-7 nm lithography," *Proc. SPIE* **10446**, 1044604 (2017).
103. K. Braam et al., "EUV mask synthesis with rigorous ILT for process window improvement," *Proc. SPIE* **10962**, 109620P (2019).
104. E. Russell, "ILT and curvilinear mask designs for advanced memory designs," *SPIE eBeam Initiative Lunch Event* (2020).
105. T. Cecil, "The resurgence of ILT," in *presented at eBeam Initiative SPIE 2017 Lunch* (2017).
106. L. Pang et al., "Computational metrology and inspection (CMI) in mask inspection, metrology, review, and repair," *Adv. Opt. Technol.* **1**(4), 299–321 (2012).
107. V. Tolani et al., "Computational defect review for wafer-fab reticle requal, part 1: mask plane inspections," *Proc. SPIE* **8522**, 85221O (2012).
108. C. Wang et al., "Improve mask inspection capacity with automatic defect classification (ADC)," *Proc. SPIE* **8880**, 88800C (2013).

109. A. Sagiv et al., "Computational inspection applied to a mask inspection system with advanced aerial imaging capability," *Proc. SPIE* **7748**, 77480P (2010).
110. L. Pang et al., "Bridging the gaps between mask inspection/review systems and actual wafer printability using computational metrology and inspection (CMI) technologies," *Proc. SPIE* **8522**, 85220Z (2012).
111. C. Y. Chen et al., "*In-situ* repair qualification by applying computational metrology and inspection (CMI) technologies," *Proc. SPIE* **8701**, 870108 (2013).
112. L. Pang et al., "From computational lithography to computational inspection: inverse lithography technology (ILT) and inverse inspection technology (IIT)," *ECS Trans.* **27**(1), 433 (2010).
113. L. Pang et al., "Expanding the applications of computational lithography and inspection (CLI) in mask inspection, metrology, review, and repair," *Proc. SPIE* **7971**, 79712A (2011).
114. T. Dam et al., "Enabling virtual wafer CD (WCD) using inverse pattern rendering (IPR) of mask CD-SEM images," *Proc. SPIE* **8166**, 81660O (2011).
115. C. Klein and E. Platzgummer, "MBMW-101: world's 1st high-throughput multi-beam mask writer," *Proc. SPIE* **9985**, 998505 (2016).
116. H. Matsumoto et al., "Multi-beam mask writer MBM-1000 and its application field," *Proc. SPIE* **9984**, 998405 (2016).
117. C. Spence et al., "Manufacturing challenges for curvilinear masks," *Proc. SPIE* **10451**, 1045104 (2017).
118. R. Pearman et al., "How curvilinear mask patterning will enhance the EUV process window: a study using rigorous wafer + mask dual simulation," *Proc. SPIE* **11178**, 1117809 (2019).
119. H. Zable et al., "Writing wavy metal 1 shapes on 22-nm logic wafers with less shot count," *Proc. SPIE* **7748**, 77480X (2010).
120. L. Pang et al., "Model-based MPC enables curvilinear ILT using either VSB or multi-beam mask writers," *Proc. SPIE* **10454**, 1045407 (2017).
121. Y. Feng et al., "Freeform mask optimization using advanced image based M3D inverse lithography and 3D-NAND full chip OPC application," *Proc. SPIE* **10587**, 105870G (2018).
122. R. Pearman et al., "Fast all-angle mask 3D for ILT patterning," *Proc. SPIE* **11327**, 113270F (2020).
123. V. W. Guo et al., "Lithographic benefits and mask manufacturability study of curvilinear masks," *Proc. SPIE* **10810**, 108100P (2018).
124. I. Bork et al., "CLMPC: curvilinear MPC in a mask data preparation flow," *Proc. SPIE* **10451**, 1045109 (2017).
125. B. Su et al., "Simulation-based MDP verification for leading-edge masks," *Proc. SPIE* **10454**, 1045409 (2017).
126. I. Bork et al., "MRC for curvilinear mask shapes," *Proc. SPIE* **11518**, 115180R (2020).
127. R. Pearman et al., "Adopting curvilinear shapes for production ILT: challenges and opportunities," *Proc. SPIE* **11148**, 111480T (2019).
128. J. Choi et al., "Requirements of data technology for EUV photomask," *Proc. SPIE* **11148**, 111480F (2019).
129. R. Pearman et al., "How utilizing curvilinear design enables better manufacturing process window," *Proc. SPIE* **11328**, 113280S (2020).
130. A. Sagiv et al., "What you see is what you print: aerial imaging as an optimal discriminator between printing and non-printing photomask defects," *Proc. SPIE* **7028**, 70281E (2008).
131. R. Nagpal et al., "Wafer plane inspection for advanced reticle defects," *Proc. SPIE* **7028**, 70281H (2008).
132. W.-S. Kim et al., "Implementation strategy of wafer-plane and aerial-plane inspection for advanced mask manufacture," *Proc. SPIE* **7379**, 73791C (2009).
133. W. S. Kim et al., "Aerial plane inspection for advanced photomask defect detection," *Proc. SPIE* **7488**, 74882Q (2009).
134. J. Kim et al., "Aerial image based die-to-model inspections of advanced technology masks," *Proc. SPIE* **7488**, 748808 (2009).

135. C. Hess et al., "Wafer plane inspection with soft resist thresholding," *Proc. SPIE* **7122**, 71221C (2008).
136. C. Hess et al., "High resolution inspection with wafer plane die: database defect detection," *Proc. SPIE* **7122**, 71221A (2008).
137. C. Hess et al., "A novel approach: high resolution inspection with wafer plane defect detection," *Proc. SPIE* **7028**, 70281F (2008).
138. X. Shi et al., "Physics based feature vector design: a critical step towards machine learning based inverse lithography," *Proc. SPIE* **11327**, 113270A (2020).
139. P. Liu, "Mask synthesis using machine learning software and hardware platforms," *Proc. SPIE* **11327**, 1132707 (2020).
140. T. Cecil et al., "Establishing fast, practical, full-chip ILT flows using machine learning," *Proc. SPIE* **11327**, 1132706 (2020).
141. L. Pang et al., "Making digital twins using the deep learning kit (DLK)," *Proc. SPIE* **11148**, 111480B (2019).
142. L. Pang et al., "How GPU-accelerated simulation enables applied deep learning for masks and wafers," *Proc. SPIE* **11178**, 111780A (2019).
143. F. E. Abboud et al., "Mask data processing in the era of multibeam writers," *Proc. SPIE* **9235**, 92350W (2014).
144. H. Matsumoto et al., "Multi-beam mask writer MBM-1000," *Proc. SPIE* **11518**, 115180A (2020).
145. A. Dounde et al., "A study of curvilinear routing in IN5 standard cells: challenges and opportunities," *Proc. SPIE* **11148**, 111481C (2019).
146. K. Hooker et al., "New methodologies for lower-K1 EUV OPC and RET optimization," *Proc. SPIE* **10143**, 101431C (2017).
147. N. Nakayama, "Proposal of a new data format for multibeam mask writer with curve expression," in *presented at Lithography Workshop* (2019).
148. Wikipedia, "Huang's law," https://en.wikipedia.org/wiki/Huang%27s_law (accessed August 2021).
149. E. Russell, "ILT and curvilinear mask designs for advanced memory nodes," in *eBeam initiative Lunch* (2020).
150. H. Zable et al., "GPU-accelerated inline linearity correction: pixel-level dose correction (PLDC) for the MBM-1000," *Proc. SPIE* **10454**, 104540D (2017).
151. H. Matsumoto et al., "Multibeam mask writer MBM-1000," *J. Micro/Nanolithogr. MEMS MOEMS* **17**(3), 109580J (2018).
152. H. Matsumoto et al., "Multi-beam mask writer MBM-2000 (conference presentation)," *Proc. SPIE* **11610**, 116100Y (2021).

**Linyong (Leo) Pang** received his PhD in mechanical engineering and an additional MS degree in computer science from Stanford University. Currently, he is the Chief Product Officer and Executive Vice President at D2S, Inc. Prior to D2S, he was the GM and Sr. Vice President of Luminescent Technologies. He is most widely known as the person who coined the term, "Inverse Lithography Technology" or "ILT," and who brought curvilinear ILT into the lithography and photomask world. Prior to joining Luminescent, he held several product development and marketing management positions at Numerical and Synopsys (after acquisition), and was a research scientist at Acuson. He has 38 issued patents, 27 pending patents, and 85 publications.