Breakthrough curvilinear ILT enabled by multi-beam mask writing

Linyong (Leo) Pang,[a,*] Ezequiel Vidal Russell,[b] Bill Baggenstoss,[b] Michael Lee,[b] Jennefir Digaum,[b] Ming-Chuan Yang,[b] P. Jeffrey Ungar,[a] Ali Bouaricha,[a] Kechang Wang,[a] Bo Su,[a] Ryan Pearman,[a] and Aki Fujimura[a]

[a]D2S, Inc., San Jose, California, United States
[b]Micron Technology, Inc., Boise, Idaho, United States

Abstract

Background: Multi-beam mask writers have been one of the most significant additions to the semiconductor manufacturing equipment arsenal in over a decade. The ability of multi-beam mask writers to write masks with a constant write time regardless of mask shapes or complexity has made them an eagerly anticipated advancement to help write curvilinear mask shapes for both today's advanced 193i nodes and for extreme ultra-violet (EUV) lithography in the future. Perhaps the most obvious application for these new pixel-based mask writers is the production of curvilinear inverse lithography technology (ILT) masks. ILT has been seen as a promising solution to many of the challenges of advanced-node lithography, whether optical or EUV. However, the runtimes and mask writing times associated with this computational technique have limited its practical application. Until recently, it had been used for critical "hotspots" on chips, but had not been used for entire chips.

Aim: The introduction of multi-beam mask writing, along with the advent of graphics processing unit (GPU)-accelerated computing for mask and wafer, have enabled the introduction of a new approach to full-chip ILT using these new technologies. The goal was to produce full-chip, curvilinear ILT within the traditional turnaround times of mask shops.

Approach: The solution to the runtime problem for ILT has been particularly vexing, as the traditional approach to runtime improvement—partitioning and stitching—has failed to produce satisfactory results, either in terms of runtime or in terms of quality. In 2019, D2S introduced an entirely new, stitchless approach, systematically designed for ILT, multi-beam mask writers, and GPU acceleration, that makes full-chip ILT a practical reality in production for the first time.

Results: We present this new ILT approach, first introduced using a multi-beam mask writer to create the complex curvilinear mask shapes. We also review findings that targeting curvilinear mask shapes creates masks that are more resilient to manufacturing variation. Finally, we review the results of this new, stitchless full-chip curvilinear ILT as applied to memory chip making. We show mask making and wafer print results, including pattern fidelity and process window, to demonstrate the actual benefit of such technologies—a doubling in the wafer process window—for semiconductor manufacturing.

 [DOI: 10.1117/1.JMM.20.4.041405]

Keywords: multi-beam mask writer; photomask; graphics processing unit; single instruction multiple data; inverse lithography technology (ILT); curvilinear ILT; OPC; resolution enhancement technology.

Paper 21046SS received May 18, 2021; accepted for publication Oct. 22, 2021; published online Nov. 17, 2021.

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

1932-5150/2021/$28.00

Fig. 1 The original ILT mask patterns shown in the luminescent ILT paper are curvilinear.[9]

1 Introduction

1.1 *Curvilinear ILT Started over a Decade Ago*

Over the last two decades, for semiconductor manufacturers targeting advanced nodes—from 90 nm all the way to 5 nm now—the greatest challenge has always been lithography. This is because lithography is fundamentally constrained by basic principles of optical physics. It has long been known that the best lithography that is theoretically possible can be achieved by considering the design of photomasks as an inverse problem—and then solving the inverse problem to find the optimal photomask for a given process, using a mathematical approach. This approach has been explored for many years, starting with the pioneering work of Saleh et al. in the 1980s.[1–8] Then in 2005 and 2006, Luminescent Technologies (later acquired by Synopsys and KLA) introduced the industry's first commercial product and the author coined the term inverse lithography technology (ILT) for this approach.[9–17] ILT is a rigorous computational approach to determine the mask shapes that will produce the desired on-wafer results. Given a target wafer shape and models of the lithographic optics, an inverse calculation is made to arrive at the mask pattern that will supply the desired wafer result and the best process window. Since lithography optics is a band-limited system, the ILT solution tends to be curvilinear[9] (Fig. 1).

1.2 *Curvilinear ILT Produces the Best Process Window*

Since the late 1990s, the semiconductor industry has faced technical challenges posed by shrinking wafer geometries and the physical limitations of optical lithography to faithfully reproduce those geometries. ILT has shown great promise as a means of meeting these challenges. Numerous studies and wafer results have shown that ILT—in particular, unconstrained curvilinear ILT—can produce the best results in terms of wafer-pattern fidelity and process window.[18] In this study (Fig. 2),[18] authors looked at contact arrays with different pitches and generated five variations of ILT mask patterns for each pitch, masks were created with these ILT patterns, wafers were printed at different process conditions, and wafer images and CDs were captured

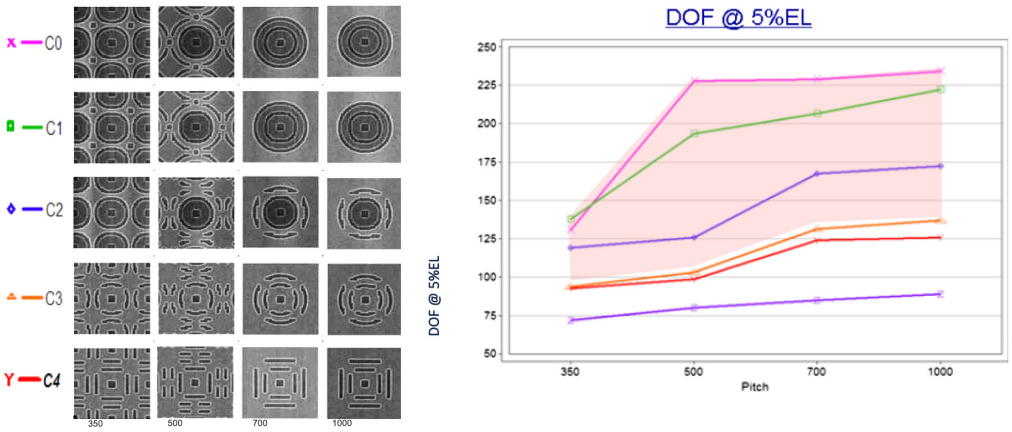


Fig. 2 Study of different complexities of ILT mask patterns and their corresponding process window shows unconstrained curvilinear ILT mask patterns produce the largest process window.[18]

and measured. This work showed that the unconstrained curvilinear ILT mask patterns produced the largest process window for all pitches.

Moving forward, ILT will be required by more and more masks, whether 193i or EUV. Optical lithography will rely more and more heavily on ILT for further progression in the roadmap to handle smaller nodes, more layers in the smaller nodes, and more aggressive design rules. With each new smaller geometry, more areas of masks become "critical" and need ILT to ensure resolution and preserve process windows. In addition, specific EUV effects (the non-normal, 6-deg incidence of the optical axis for the reflective optical system, as well as mask 3D effects such as mask shadowing), combined with tight lithography error budgets require curvilinear corrections for EUV, make curvilinear ILT the ultimate solution for EUV masks.

According to the eBeam Initiative Mask Makers Survey in 2020,[19] which also showed the data from 2017 to 2019, ILT has already been in use for critical layers in the leading technology nodes for several years. However, it has been mainly used in hot-spot fixer mode and has not been used for all critical layers.

Two major obstacles have kept ILT from being widely applied. One of these barriers—the ability to write curvilinear mask patterns—was removed recently by introduction of multi-beam mask writers, which can write any shape without time penalty. The other major barrier—ILT runtime—was still left to be overcome.

1.3 *Multi-Beam Mask Writer Enables Curvilinear ILT*

The most common mask writer for the leading semiconductor manufacturers, called the variable-shaped beam (VSB) mask writer, was invented to write Manhattan (rectilinear) patterns. It writes the mask using a single beam that can produce a rectangular shot with variable dimensions. The total write time for masks written by VSB mask writers is proportional to number of these rectangular shots that are required to produce the complete mask. This is an advantage when writing Manhattan patterns that can be produced with large rectangular shots, but to write a curvilinear pattern, the VSB writer has to break the curvy patterns into many small rectangular shots, resulting in write times that are too long and impractical for full-chip production use [Fig. 3(a)]. Some VSB mask writers added triangle shots and cell projection as a means of addressing the limitation of strictly Manhattan VSB. This helps for some special cases, such as 45-deg angled lines, but does not change the shot-count challenge for curvilinear patterns.

The mask industry recognized this challenge, and it became the major motivation to develop the new multi-beam mask writer. A multi-beam mask writer, instead of having a single, VSB, has an array of 256k beams that write in a single shot, with each individual beam controlled individually to turn on, off, or at partial (gray scale). Since the multi-beam mask writer writes in the pixel domain, write time is not affected by the patterns it writes, and it can write a mask with any shaped mask patterns—including curvilinear ILT mask patterns—in a constant write time, around 10 to 15 h[20,21] [Fig. 3(b)].

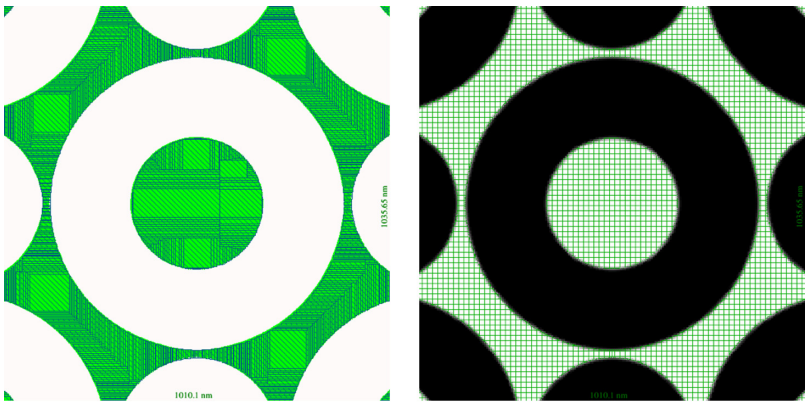


Fig. 3 (a) Conventional VSB mask writer: generates too many shots, takes too long to write. (b) Multi-beam mask writer: designed for curvilinear ILT, writes any shape in constant time.[22]

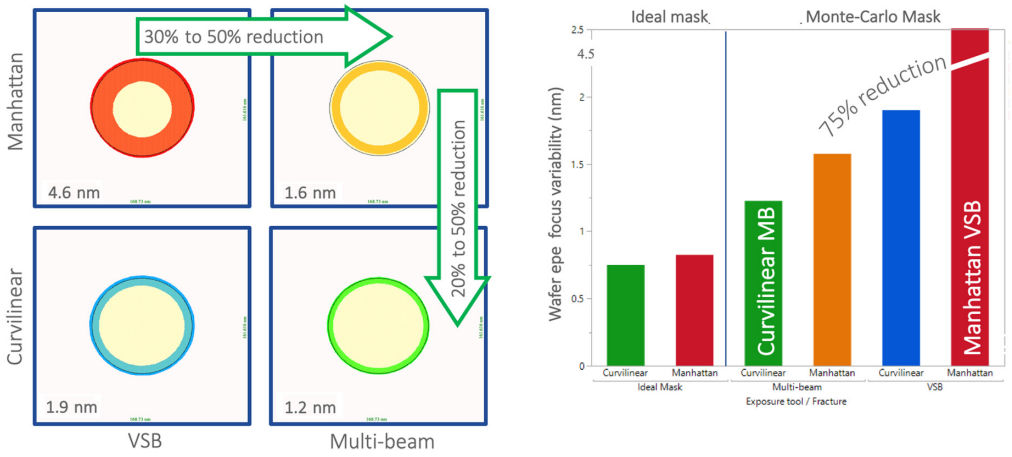


Fig. 4 Study[22] shows the mask variation PV bands can be reduced 75% by switching from writing Manhattan patterns on a VSB mask writer to writing curvilinear mask patterns on a multi-beam mask writer.

1.4 *Curvilinear Mask Shapes Are Much More Resilient to Manufacturing Variation than Manhattan Shapes*

Curvilinear ILT not only produces the best process window, but curvilinear mask shapes are also much more resilient to manufacturing variation than Manhattan shapes.[22] According to Pearman's study on EUV contacts using Manhattan and curvilinear mask shapes using both VSB and multi-beam mask writers, multi-beam mask writers can reduce mask process variation (PV) bands by 30% to 50% from VSB mask writer. Then by switching from target Manhattan patterns to target curvilinear mask patterns, mask PV bands are reduced by an additional 20% to 50%. So if we change from writing Manhattan mask patterns on VSB mask writers to writing curvilinear mask patterns on multi-beam mask writers, the mask PV bands can be reduced by a very significant 75% (Fig. 4).

1.5 *ILT Runtime Challenge: Conventional ILT Takes Weeks to Compute for Full Chip*

With the introduction of multi-beam mask writers, one of the major obstacles to full-chip ILT—excessive mask write time—was removed. This left another major barrier: ILT computation runtime. The computations and models required for accurate ILT have been established and refined over the last decade since the introduction of the concept. The problem has been the sheer volume of the computations required to perform full-chip ILT and the runtimes that result.

ILT's computation is already an order of magnitude higher than traditional optical proximity correction (OPC) due to the much larger solution space of ILT. On top of this, using the standard approach, the computations for full-chip ILT are too lengthy to be practical (Fig. 5). Since each CPU can handle ILT computation only for a small area, conventional approaches to ILT divide the task (or in this case, the chip) into partitions and have the computations for each partition run in parallel to save time. As each partition is passed to a CPU to process, the processor will first calculate the ideal ILT solution then the ideal ILT mask solution is cleaned up to meet mask rules. After that, it will go through a mask-shape modification called Manhattanization. This will simplify the ideal, curvilinear mask patterns into Manhattan shapes that a VSB mask writer can produce. Since the mask shapes are dramatically changed in this step, a reoptimization is required to ensure the new Manhattan mask pattern will meet the wafer pattern accuracy requirement and process window requirements. Then the partitions are "stitched" back together. However, because the physics of any mask feature are impacted by the features adjacent to it, but each partition is processed separately without knowing the changes on the adjacent partitions (even with a buffer region of overlap or "halo"), this approach will produce inconsistencies and discontinuities at the partition boundaries called "stitching errors." (Note that these

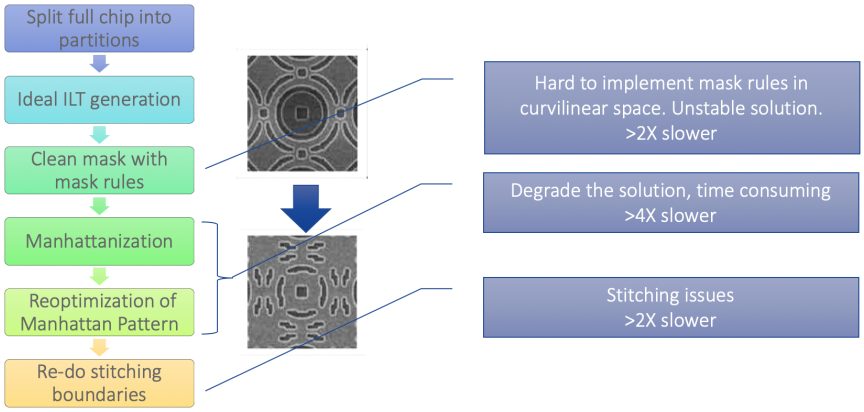


Fig. 5 In the conventional ILT flow, in order to generate a solution for full chip and for VSB mask writer to write, the runtime has increased order of magnitude due to all these extra steps.

stitching errors are generated by ILT, they have nothing to do with mask writer stitching error.) Such stitching errors must be corrected once the partitions are stitched together. The method to correct stitching errors is to take regions around the partition boundaries after merging, plus some buffer region, recalculate the ILT solution, and then put them back together. This method will fix the existing stitching errors, but 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 required due optical proximity effects is relatively large, these restitching areas are close to the size of the original partitions, effectively doubling the ILT runtime. All of these steps on top of ideal ILT generation increase the runtime significantly. In the end, the total runtime is an order of magnitude slower than the ideal ILT, which is already an order of magnitude slower than OPC.

As a result, commercial applications of ILT have been limited and have focused mainly on smaller, high-risk portions of masks, mainly used in hot-spot fixer mode to make runtime acceptable. A high-volume, full-chip ILT solution has been elusive.

2 Stitchless Full Chip Curvilinear ILT for the Multi-Beam Era

Multi-beam mask writers can write curvilinear masks in a constant write time, and curvilinear mask patterns are more resilient than Manhattan patterns, so it seemed the time had come to tackle the runtime obstacle and create an approach to curvilinear ILT for multi-beam era.

2.1 *Solution: Get Rid of the Stitches*

The rise, in the last decade, of the use of GPU-based computing for scientific applications has offered a new opportunity for bringing a practical full-chip ILT solution to market. GPU-accelerated computing excels at single-instruction, multiple data (SIMD) computation, in contrast to 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 (Fig. 6).

This is not a novel observation. Several attempts have been made to create commercial, full-chip ILT solutions by porting CPU-based solutions to a GPU-accelerated computing environment. However, these solutions have still fallen short in acceptable turnaround time.[23,24]

Partitioning/stitching has been the major culprit. Feeding chip partitions into a GPU-accelerated computing system can speed the processing of each partition. However, stitching errors and the recomputation required to address them are still show-stopping issues. D2S reasoned that what was needed was the ability to process the entire chip at once: a single, giant GPU/CPU pair that could optimize full-chip data seamlessly, without partitions.

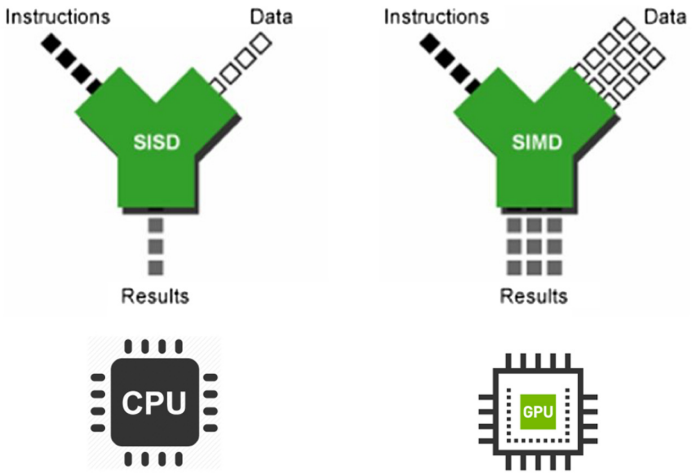


Fig. 6 Illustration of CPU's SISD and GPU's SIMD.

2.2 *Stitchless, Curvilinear Full-Chip ILT in a Day with GPU Acceleration*

Of course, such a giant GPU/CPU pair does not exist. However, by taking a "from the ground up," holistic approach, D2S was able to build an ILT-specific computing appliance that could emulate a giant GPU/CPU pair.

This approach did not stop with the hardware, but rather included every component of a holistically conceived, purpose-built system—hardware, software, models, visualization, and verification—that is designed and implemented from the ground up for GPU-acceleration and for full-chip ILT computation. Every aspect of the physics and chemistry of wafer lithography and processing, including litho simulations, mask and wafer models was examined and optimized synergistically throughout the system to reap the largest potential runtime benefits without compromising computational accuracy.

2.2.1 *Stitchless*

As discussed earlier, chip partitioning and parallel computing are the most common approaches to runtime reduction for full-chip computations. However, physics effects at advanced nodes are highly contextual, and partition boundaries naturally create contextual "disagreements" between items on either side of the boundaries. In addition, shifts that occur on mask can cause distortion of features that lay directly on the boundaries of a partition (think of misaligned sections of wallpaper). Handling these stitching errors—avoiding or correcting them—is one of the biggest hurdles for full-chip ILT (Fig. 7).

To avoid the time-consuming, recursive correction passes necessary to resolve these stitching errors, D2S built the GPU-accelerated hardware platform (called the computational design platform or CDP) and designed the software for ILT so that the entire chip could be optimized at

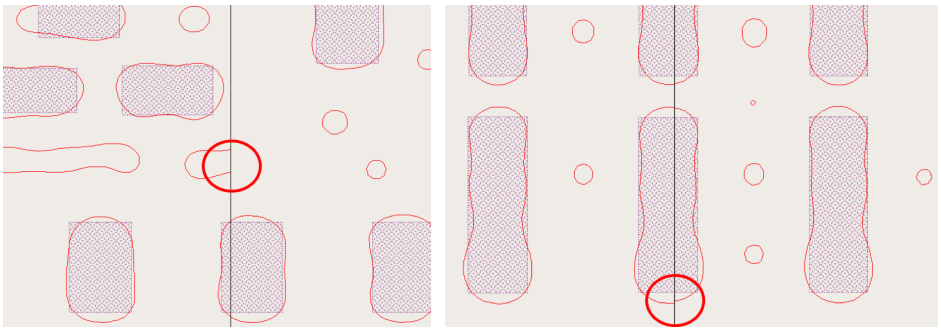


Fig. 7 Stitching errors occur when a chip is partitioned for parallel computing and reassembled.

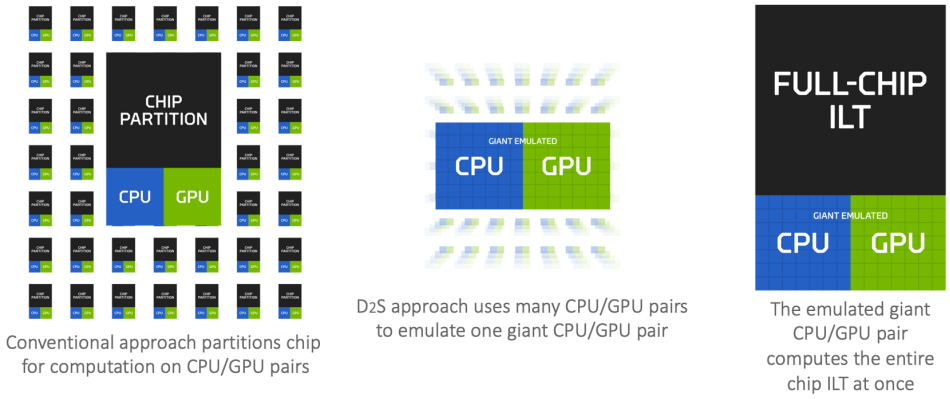


Fig. 8 Although comprising many GPU/CPU pairs, this solution has been holistically designed so that it behaves as a single GPU/CPU pair, iterating on the entire chip as a whole, and avoiding stitching errors.

once. The D2S CDP has been purpose-built specifically to address simultaneous full-chip optimization. Although it contains dozens of GPU-CPU pairs, this full chip curvilinear ILT solution, including the CDP and software, is designed to behave as though the whole system is a single, giant GPU–CPU pair that can process the mask for the entire chip simultaneously.

The system behaves as though there are no partitions. This means that each optimization iteration updates the entire chip as a whole, so that all optical proximity effects across the chip are accounted for with each update (Fig. 8). At a high level, the CDP holds the data for the entire chip in its memory across all computing nodes (with each computing node comprising a CPU/GPU pair). With each iteration, the CPU/GPU pairs process the chip in the CDP in parallel. At the end of each iteration, the full chip data held in the CDP is updated using the data sent back from each GPU. Then the next iteration will repeat this process.

Because there are no partition boundaries, the solutions everywhere are continuous, as shown in Fig. 9.

2.2.2 *Curvilinear*

Because nothing in nature (including the physics of semiconductor manufacturing) makes 90-deg corners, the shapes on manufactured masks and wafers are all curvilinear, even if the input geometries are rectilinear (see Fig. 10). In fact, as we showed in Sec. 1.4, curvilinear shapes with certain minimum curvatures of shapes and spaces have been shown to be more reliably manufacturable than rectilinear shapes.[20]

ILT is a mathematical approach that naturally produces curvilinear shapes. Traditionally, extra computation time has been needed to Manhattanize the curvilinear ILT shapes because VSB-based mask writing could not process curvilinear mask shapes within practical runtimes. However, with multi-beam mask writing now available, curvilinear shapes no longer require

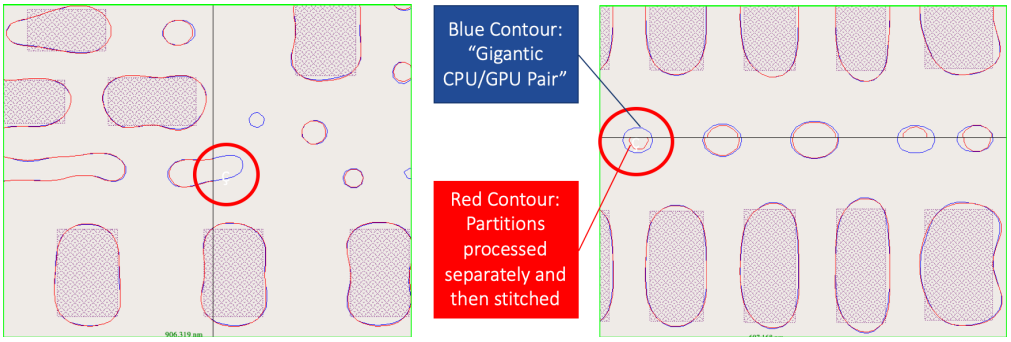


Fig. 9 No stitching errors occur in this full-chip approach.

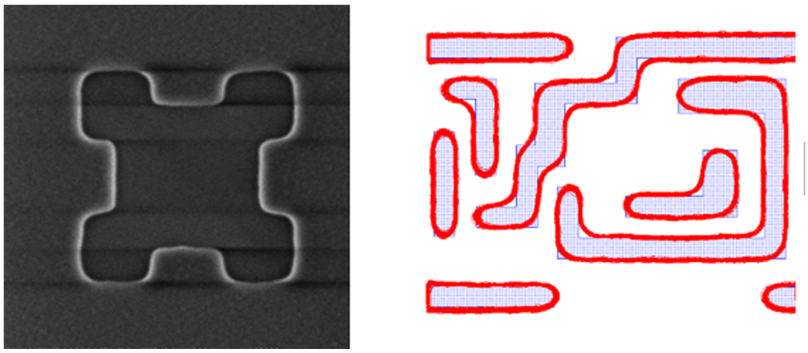


Fig. 10 All shapes on masks and wafers are curvilinear, even if the input geometries are rectilinear: (a) Manhattan OPC mask pattern with serif appears curvilinear on mask. (b) Wafer pattern designed as Manhattan appears curvilinear in simulation.

additional time to write. This full-chip, curvilinear ILT solution was built to leverage the power of this new world of multi-beam-based mask writing and is optimized for curvilinear mask output.

Curvilinear ILT does equally well on curvilinear input design shapes. As multi-beam mask writers and EUV move into volume production, designers will be able to target curvilinear designs that are more manufacturable, and curvilinear ILT will handle these designs with ease.

Uniquely, this curvilinear ILT solution is able to compute curvilinear shapes efficiently because of GPU acceleration. ILT inherently computes in the pixel domain; GPU-based computing was built for pixel-manipulation, so it is a perfect fit for this task. With its approach to emulate a giant GPU/CPU pair, this full chip curvilinear ILT computes, in essence, a rasterized image of the entire chip all at once, iterating on the full-chip ILT solution as a whole.

2.2.3 *Full-chip ILT*

Full-chip ILT has been the ultimate goal of ILT since its inception. It has been deployed only for "hotspots" and "critical areas" because the turnaround time for full-chip ILT was prohibitive. Ironically, however, stitching problems are more pronounced when "hotspot" ILT solutions need to be stitched into traditional OPC areas. There is no doubt that full-chip ILT is best, if runtime was not an issue. This unique approach to ILT makes full-chip ILT a practical reality.

2.2.4 *In a day*

For semiconductor companies, time is money, and time to market is critical for their revenue. This reality pushes semiconductor manufacturing companies, in particular, wafer fabs to tape out and deliver wafers in its shortest time possible, which commonly constrains the budgets for OPC and ILT process time to one day. This full-chip, curvilinear ILT is the first commercial ILT solution that delivers full-chip ILT within this time constraint.

2.2.5 *Integrates curvilinear mask rules to produce MRC clean results*

Curvilinear ILT still needs to meet mask rules, because mask processes, similar to lithography processes, are limited or affected by dose profile and contrast, resist resolution, and etching process. Mask rules for curvilinear patterns have turned out to be simpler than mask rules in Manhattan space: basically, they are minimum CD, minimum space, minimum area, and minimum curvature.[25] In this curvilinear ILT, such curvilinear mask rules are integrated into the ILT optimization, therefore, it produces mask-rule compliant (MRC)-clean results. Figure 11 shows an example of the mask patterns produced by this solution without and then with integrated MRC: when MRC is integrated, any features that violate minimum-feature rules do not appear in the final curvilinear ILT mask.

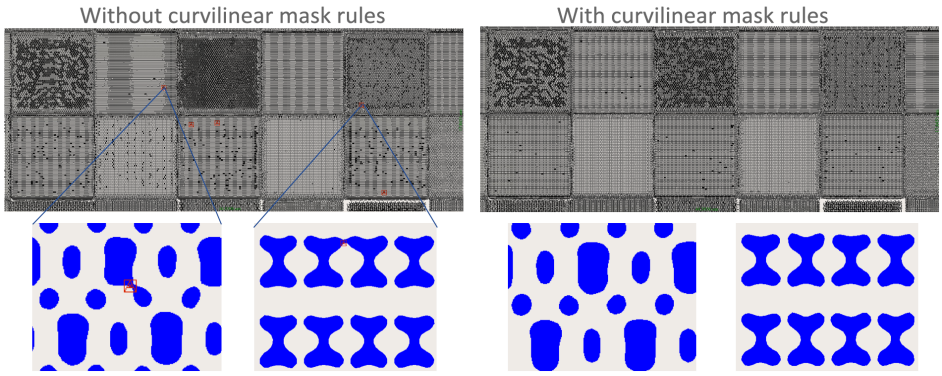


Fig. 11 Comparison of the full-chip ILT solution without and with integrated MRC. The top row shows the layout where the red marks are MRC violations detected by mask verification.

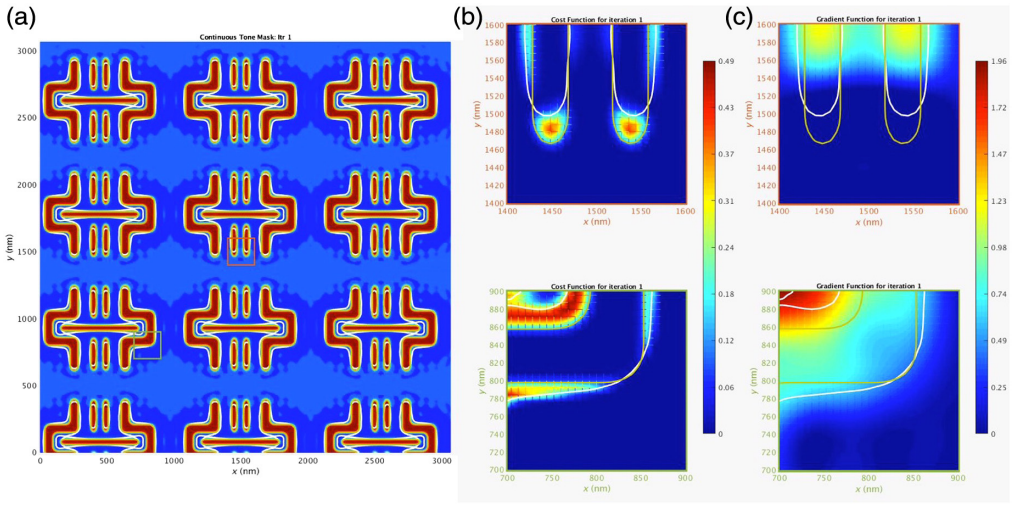


Fig. 12 Mask pattern, simulated wafer contour and its target, cost function, and cost gradient at the beginning of the ILT optimization. (a) Continuous tone mask, (b) cost function, and (c) gradient of the cost function.

2.2.6 *Meets EPE requirements*

Although this curvilinear ILT is a pixel-domain implementation, its optimization can directly drive edge-placement error to meet OPC requirements. Figure 12 shows a 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, cost function is not zero, and cost gradient is not flat. Figure 13 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.

2.2.7 *Continuous and symmetric*

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. This 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,[24] as opposed to the real domain (which is what is used by both Luminescent[9] and Intel[26]) with GPU acceleration.

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

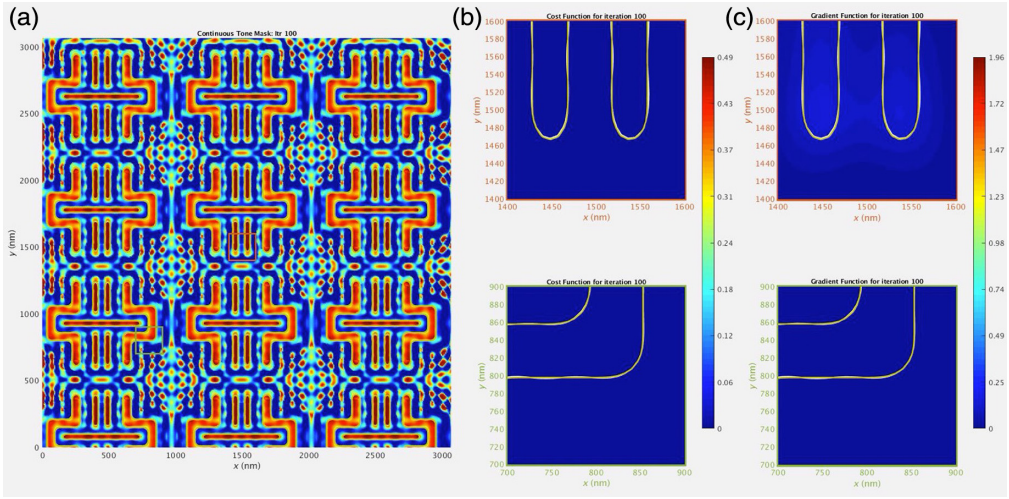


Fig. 13 Mask pattern, simulated wafer contour and its target, cost function, and cost gradient at the end of the ILT optimization. (a) Continuous tone mask, (b) cost function, and (c) gradient of the cost function.

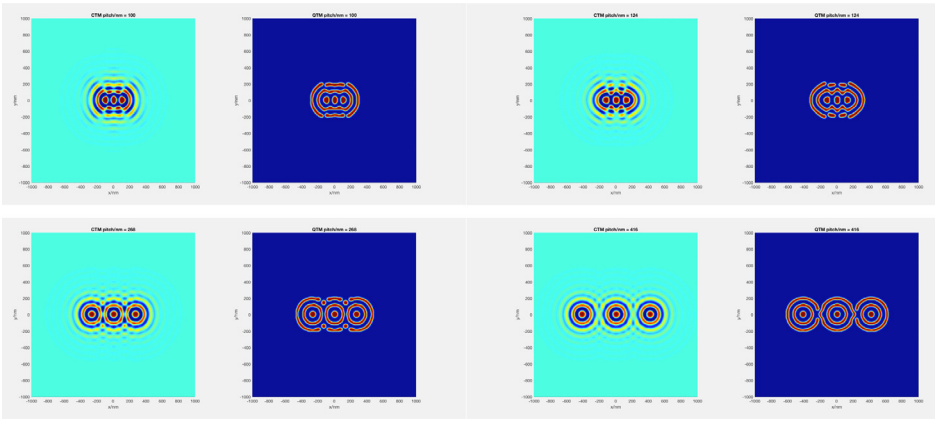


Fig. 14 Continues tone mask (CTM) and final ILT mask for three contacts in symmetric position at different pitches showing these ILT solutions are continuous and symmetric.

is that the same geometry (repeated patterns and symmetric patterns) in the real domain has 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 band-limited scanner optics, a mathematically rigorous approach to geometry selection is necessary to meet these requirements. 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 cutoff. 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.

The continuity and symmetry of this solution have been demonstrated. Figure 14 shows a symmetric three-contact configuration. As pitches change from small to large, the ILT solution gradually changes while maintaining the XY symmetry.

2.2.8 *On-grid and off-grid invariance*

Another challenge for most ILT approaches is the on-grid and off-grid invariance. The on-grid and off-grid invariance of this solution has been demonstrated. Figure 15 shows an equal-pitch contact array and its ILT solution. Figure 15(a) shows the on-grid case, whereas Fig. 15(b) shows the off-grid case. When pitches change from small to large, the ILT solution gradually changes

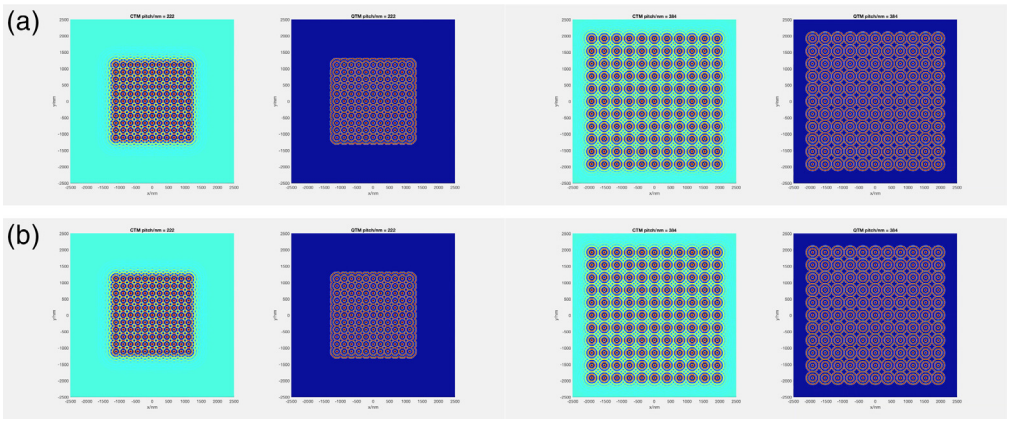


Fig. 15 CTM and final ILT mask for an equal-pitch contact array for on-grid and off-grid situation. (a) On-grid and (b) its corresponding off-grid configuration demonstrating the ILT solutions are grid invariants.

while maintaining the XY symmetry, and also the solution for off-grid case is the same as the on-grid case.

2.2.9 *Any angle*

The most challenging test for ILT is the combination of multiple pitches, on-grid and off-grid, with rotation. Figure 16 shows the same equal-pitch contact array and its ILT solutions with the pitch 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 this ILT solution.

3 Evaluation of Curvilinear ILT on MASK and Wafer

3.1 *Mask and Wafer: ILT Results on Memory Design with Free-Form Source Demonstrated*

To evaluate the benefits of this full-chip, stitchless, curvilinear ILT solution, masks were written and wafers were printed at Micron Technology using the process of record (POR). First, ILT model calibration was done using the D2S test-chip version 6. Then ILT was run at Micron on the D2S CDP to generate a curvilinear mask design.[27]

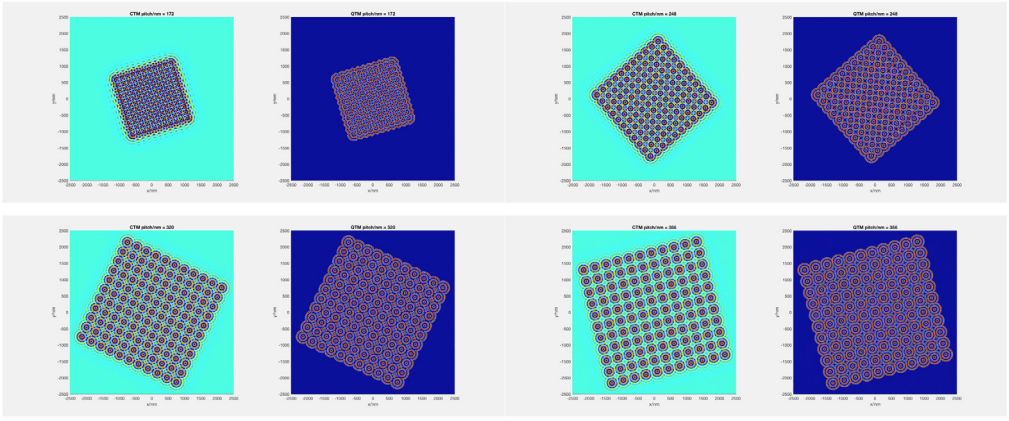


Fig. 16 CTM and final ILT mask for an equal-pitch contact array at on-grid and off-grid situation, pitch change, plus rotation demonstrating these ILT solutions are symmetric and rotation-invariant.

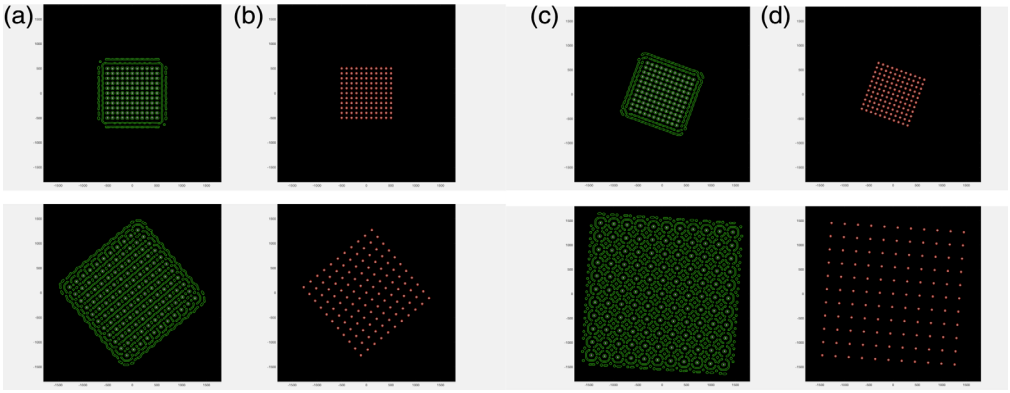


Fig. 17 In each pair, (a), (c) the ILT curvilinear mask designs for different pitches and orientations and (b), (d) corresponding wafer target and simulated wafer contours.

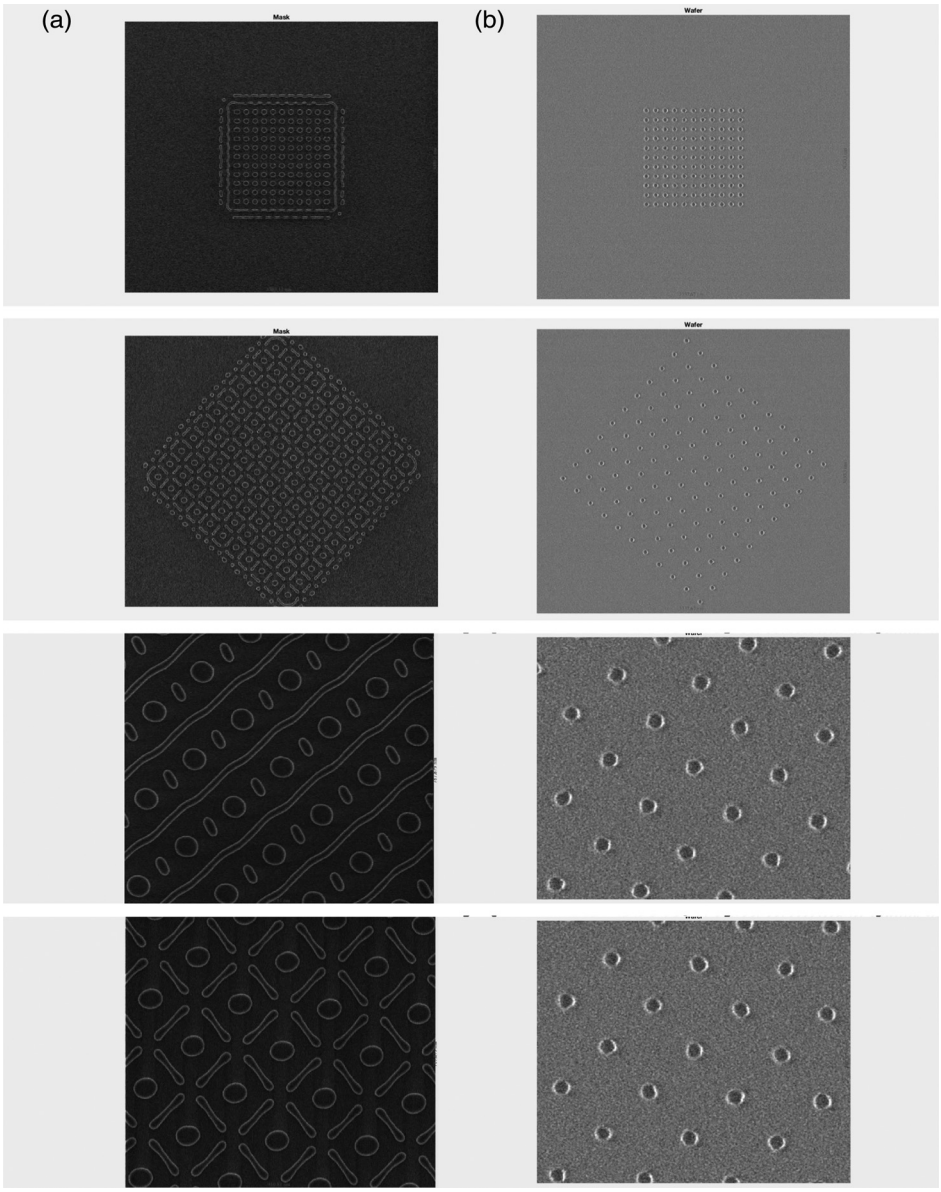


Fig. 18 In each pair, (a) the ILT curvilinear mask written by NuFlare multi-beam mask writer for different pitches and orientations and (b) the corresponding wafer prints using Micron process.

Figure 17 shows curvilinear mask patterns generated for a free-form source used in production. The contact array has 11×11 contacts with equal pitch. The contact size is 40 nm, and the tightest pitch is 85 nm. A total of 91 curvilinear ILT mask patterns were generated by varying the pitch and rotation angle.

Figure 18 shows SEM images of some instances of the actual curvilinear mask pattern written by the NuFlare multi-beam MBM 1000 and wafer print using the Micron POR. Mask patterns are resolved with high-pattern fidelity and very smooth profile. On the wafer print, all contacts are printed evenly from the center of array to edge of the contact array.

3.2 *Process Window: Wafer Results Show ILT Doubles the Wafer Process Windows Compared to OPC*

The ultimate goal for curvilinear ILT is to achieve the best process window, so in this evaluation, process windows were compared between OPC and the full-chip ILT solution using the same process.

Figure 19 shows the side-by-side wafer print comparison of OPC and this curvilinear ILT at different process conditions (different focuses and doses) from −60 nm defocus to +60 nm defocus and from 93.3% dose to 106.7% dose variation. Figure 19 randomly picked 6 process conditions from a total of 49 conditions. One can clearly see the ILT solution has a much bigger process window than OPC. In many of the OPC wafer prints, the contacts are not printed evenly from the center of the array to edge of the array, some have necking problems, some do not even print at all. In contrast, the ILT wafer images show very consistent print for all process conditions, for all contacts no matter their locations in the array, pitch, or angle of the rotation.

CDs were also measured to quantify the size of the process window for both the OPC and the ILT solution. This was done on another cut-layer type of design. Figure 20 shows the wafer prints for all process window matrices. The target CD is 62.8 nm, all dies with CD of 10% variations are considered within process window. Figure 21 shows the CD measurements, and the conditions within process window are highlighted in green. Notice the x axis is the focus, and y axis is the dose to be consistent with process window plot. Three wafer images at process center and two process corners are also shown in zoomed-in version. Compared to OPC, the ILT solution has increased the process window by over 100%.

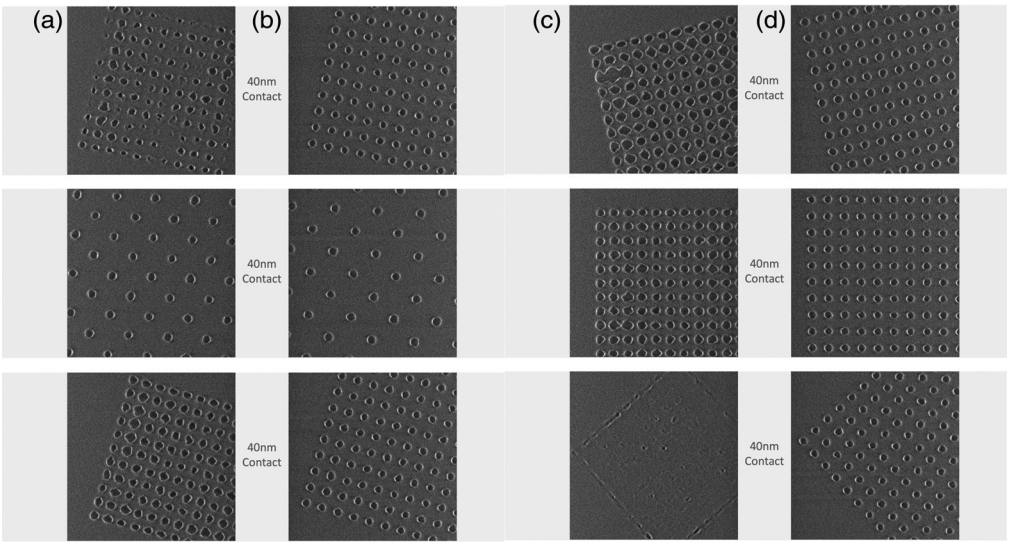


Fig. 19 In each pair, (a), (c) OPC wafer printed by Micron process at different process conditions and (b), (d) the ILT solution wafers printed at the same process condition.

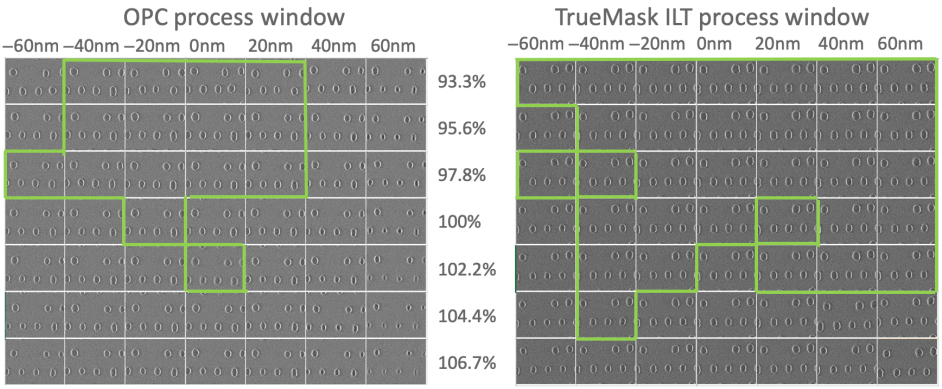
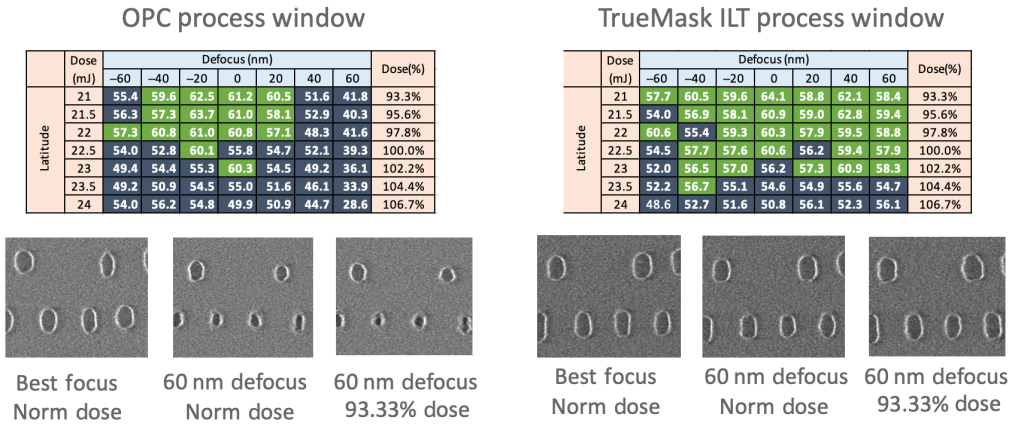


Fig. 20 Wafer print matrix for a cut-layer type of design. Highlighted regions are within process window.



OPC process window

Latitude	Dose (mJ)	Defocus (nm) –60	–40	–20	0	20	40	60	Dose(%)
	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

Latitude	Dose (mJ)	Defocus (nm) –60	–40	–20	0	20	40	60	Dose(%)
	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. 21 Process window CD measurements. The highlighted regions are within process window.

4 Summary and Conclusions

For several decades, the semiconductor industry has recognized the value of ILT in addressing the challenges of advanced-node lithography. However, major barriers of ILT computation runtime and curvilinear ILT mask write time have kept ILT from being adopted widely as a full-chip solution. The introduction of shape-agnostic multi-beam mask writers removed one of these critical barriers. Embracing a unique, holistically conceived, purpose-built system of GPU-accelerated hardware and software that emulates a single-giant GPU/CPU pair and iterates and optimizes the entire chip as a whole, made stitchless, curvilinear, full-chip ILT in a day a practical reality.

The benefit of this approach has been evaluated at Micron Technology. The ILT curvilinear mask was written by NuFlare multi-beam mask writer MBM-1000, the wafer was printed using Micron POR. The results first show ILT curvilinear mask pattern can be written by multi-beam mask writer with high-pattern fidelity. Most importantly, the study shows that this full-chip, stitchless, curvilinear ILT has much superior wafer print quality and enlarged the process window by over 100% compared to OPC.

Acknowledgments

The authors would like to thank NuFlare for their help to write the curvilinear masks with MBM-1000 and ASML for their help to acquire wafer images with eP5 platform.

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**, 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," *J. Micro/Nanolithogr. MEMS MOEMS* **1**(1), 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. Y. H. Oh and J.-C. Lee, "Resolution enhancement through optical proximity correction and stepper parameter optimization for 0.12-μm mask pattern," *Proc. SPIE* **3679**, 607–613 (1999).
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. D. S. Abrams and L. Pang, "Fast inverse lithography technology," *Proc. SPIE* **6154**, 61541J (2006).
10. L. Pang, Y. Liu, and D. Abrams, "Inverse lithography technology (ILT): what is the impact to the photomask industry?" *Proc. SPIE* **6283**, 62830X (2006).
11. Y. Liu et al., "Inverse lithography technology principles in practice: Unintuitive patterns," *Proc. SPIE* **5992**, 599231 (2005).
12. B. Lin et al., "Inverse lithography technology at chip scale," *Proc. SPIE* **6154**, 615414 (2006).
13. C.-Y. Hung et al., "Pushing the lithography limit: Applying inverse lithography technology (ILT) at the 65 nm generation," *Proc. SPIE* **6154**, 61541M (2006).
14. J. Ho et al., "Real-world impacts of inverse lithography technology," *Proc. SPIE* **5992**, 59921Z (2005).
15. A. Moore et al., "Inverse lithography technology at low k1: placement and accuracy of assist features," *Proc. SPIE* **6349**, 63494T (2006).
16. C. Y. Hung et al., "First 65-nm tape-out using inverse lithography technology (ILT)," *Proc. SPIE* **5992**, 59921U (2005).
17. C. W. Chu et al., "Enhancing DRAM printing process window by using inverse lithography technology (ILT)," *Proc. SPIE* **6154**, 61543O (2006).
18. B. G. Kim et al., "Trade-off between inverse lithography mask complexity and lithographic performance," *Proc. SPIE* **7379**, 73791M (2009).
19. "eBeam initiative luminaries survey results," 2020, http://www.ebeam.org (accessed 22 September 2020).
20. C. Klein and E. Platzgummer, "MBMW-101: World's 1st high-throughput multi-beam mask writer," *Proc. SPIE* **9985**, 998505 (2016).
21. H. Matsumoto et al., "Multi-beam mask writer MBM-1000 and its application field," *Proc. SPIE* **9984**, 998405 (2016).
22. 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).
23. J. Zhang et al, "GPU-accelerated inverse lithography technique," *Proc. SPIE* **7379**, 73790Z (2009).
24. I. Torunoglu et al, "A GPU-based full-chip inverse lithography solution for random patterns," *Proc. SPIE* **7641**, 764115 (2010).
25. R. Pearman et al., "Adopting curvilinear shapes for production ILT: challenges and opportunities," *Proc. SPIE* **11148**, 111480T (2019).
26. V. Singh et al, "Making a trillion pixels dance," *Proc. SPIE* **6924**, 69240S (2008).

27. 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**, 111480U (2019).

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 senior 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.

Ezequiel Vidal-Russell is the senior director of mask technology at Micron Technology. He joined Micron in 2002 in research and development working on photolithography masks for the manufacturing of microchips, with emphasis on resolution enhancement techniques (RETs) and optical proximity corrections (OPC). He earned a licenciado diploma in physics in 1996 and his PhD in physics in 2001, both from Instituto Balseiro (Bariloche, Argentina).

Jennefir Digaum received his PhD in optics and photonics from the University of Central Florida. He has a physics and electrical engineering undergraduate degree from Mindanao State University and MS degree in photonics from four different universities in Europe under the Erasmus Mundus consortium. Currently, he is a principal engineer – RET design lead at Micron where his group supports the patterning development of DRAM and other emerging memories, using both DUV and EUV photolithography technologies.

P. Jeffrey Ungar received his PhD in physics from Stanford University and his BSc degree in engineering physics from Queen's University. He is currently chief scientist for TrueMask ILT at D2S, Inc. He has 40 issued patents in many technologies, including computer graphics, power conversion circuitry, lithium ion batteries, and ILT algorithms.

Aki Fujimura is the founder and CEO of D2S, Inc. Previously, he served as CTO at Cadence Design Systems. He returned to Cadence for the second time through the acquisition of Simplex Solutions where he was president/COO and inside board member. He was also an inside board member and VP at Pure Software. Simplex and Pure both IPO'd during his tenure. He was a founding member of Tangent Systems in 1984, which was subsequently acquired by Cadence Design Systems in 1989. He was on the boards of HLDS, RTime, Bristol, S7, and Coverity, Inc., all of which were successfully acquired. He received his BS/MS degrees in electrical engineering from MIT.