PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Leaping into the curvy world with GPU accelerated O(p) computing

Abhishek Shendre, Aki Fujimura

Abhishek Shendre, Aki Fujimura, "Leaping into the curvy world with GPU accelerated O(p) computing," Proc. SPIE 12751, Photomask Technology 2023, 1275108 (21 November 2023); doi: 10.1117/12.2689299

SPIE.

Event: SPIE Photomask Technology + EUV Lithography, 2023, Monterey, California, United States

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 27 Nov 2023 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Leaping into the Curvy World with GPU Accelerated $O(p)$ Computing

Abhishek Shendre[a], Aki Fujimura[a]
[a]D2S Inc., 4040 Moorpark Ave., San Jose, CA, USA 95117

ABSTRACT

With the advent of curvilinear mask enabled by multi-beam mask writing [1] and curvilinear ILT [2], full reticle curvilinear mask processing is emerging as one of the new challenges in electronic design automation and specially in the mask data preparation (MDP) domain. Whether for 193i or for EUV, curvilinear masks provide superior wafer results from larger process windows. Although the curvilinear photomask designs can provide an excellent opportunity to improve mask process window [3,4] compared to traditional Manhattan designs, they put a strain on the MDP data path [5] due to the increasing complexity of data representation. The mask industry is tackling this issue using a Bezier and B-spline based "Multigon" format [6] to replace the traditional piecewise linear polygon-based formats. Pixel-based computing and consequently a representation of curves that is aware of the mask writer pixel size [7] can further assuage the problems of data path and computational overhead in using curvilinear photomasks.

This paper demonstrates the inherent advantages of pixel-based computing for curvilinear photomasks, when using a GPU-based platform, through comprehensive analysis and empirical evidence. GPU acceleration has played a very important role in making the full chip curvilinear mask correction for shapes represented using piecewise linear polygons [8]. Since CPU-based algorithms perform better with piecewise linear polygons, this approach to GPU acceleration is necessary and important to the industry [9]. By taking a different approach that assumes the presence of GPUs in a compute node, however, pixel-based computations are enabled, taking advantage of the Single Instruction Multiple Data (SIMD) nature of GPUs. This paper studies the advantages of using GPU acceleration for pixel-based computing in various mask processing and verification steps. The paper highlights the natural runtime predictability of pixel-based computing, which is in the order of number of pixels, or $O(p)$, irrespective of the complexity of the mask shapes. The paper also emphasizes that pixel dose equivalence and information theory [7] provide a mathematical basis for the practical accuracy of pixel-based approach towards MDP.

Pixel-based computing has been the backbone of various fields in computer science and computer-aided design (CAD) tools. However, it is still a relatively unexplored computational paradigm for the photomask industry, especially in the mask verification and processing steps. GPU performance scales by bit-width rather than by clock speed. The continued scaling of GPU processing speed has enabled the shift in perspective towards GPU-based computing [10]. This paper concludes that the $O(p)$ approach for GPU acceleration enables accurate and practical data processing for curvy masks governed by information theory, as we leap into an increasingly complex curvy world.

Keywords: Photomask, GPU, Curvilinear Masks, Information Theory, Pixel Based Computing, Mask Data Preparation, Mask Process Correction, Mask Verification.

1. INTRODUCTION: TRANSITION TO CURVY MASKS

The progression of technology nodes with Moore's Law has been a big challenge for the semiconductor industry. Since the critical dimensions in technology nodes have reached beyond the wavelength of light used in the lithography, we need Optical Proximity Correction (OPC), creating complex photomasks, to ensure that designs print as intended on wafer. Initial OPC methodologies were rule based, but as the technology nodes progressed further, a paradigm shift in OPC methodology was required to get better pattern fidelity. This was provided by Inverse Lithography Technology (ILT) which led to the emergence of curvilinear masks.

Photomask Technology 2023, edited by Ted Liang, Proc. of SPIE Vol. 12751,
1275108 · © 2023 SPIE · 0277-786X · doi: 10.1117/12.2689299

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 27 Nov 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

1.1 Curvilinear masks are more manufacturable

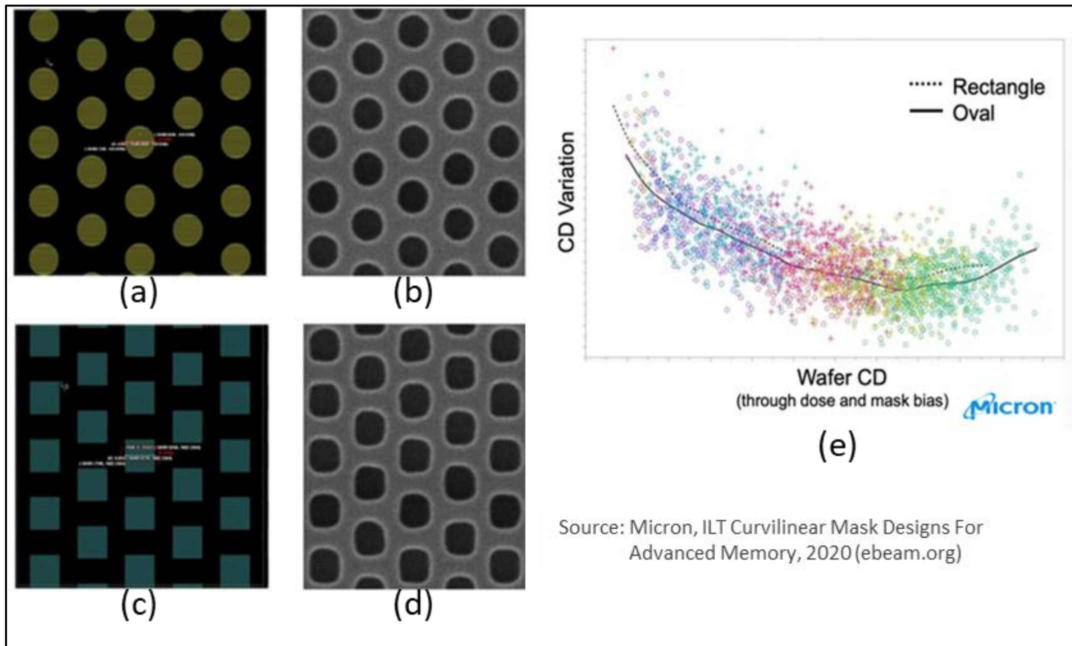

Figure 1. (a) Oval DRAM array mask data. (b) Mask SEM of oval DRAM array. (c) Rectangular DRAM array mask data. (d) Mask SEM of rectangular DRAM array. (e) Wafer CD variation across different nominal wafer CDs.

Study [11] by Micron shows oval shown in figure 1 performs comparison between rectangular DRAM Array Figure 1(c) and oval shaped DRAM array Figure 1(a). The corresponding mask SEM images, i.e. Figure 1(d) and Figure 1(b) respectively, show that oval masks data produces visually more consistent shapes compared to rectangular mask data. Figure 1(e) show there is less variance in wafer CD uniformity for oval masks compared to rectangular masks.

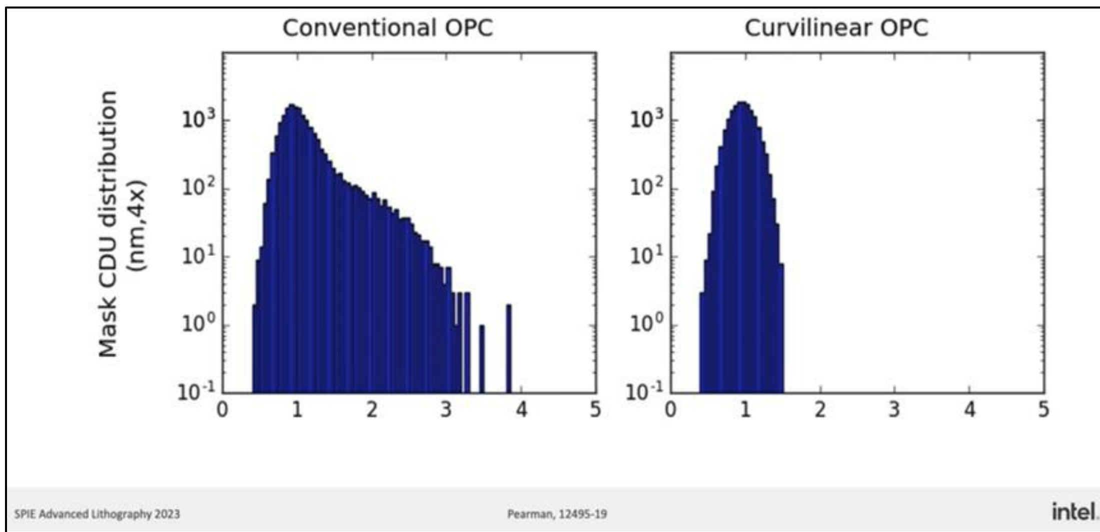

Figure 2. Mask EPE variance (1σ) for identical features.

Study [12] by Intel a shown in Figure 2 does a comparison between conventional OPC and curvilinear OPC. The mask CDU distribution chart shows a much smaller variance for curvilinear OPC compared to conventional OPC. Thus, it can be concluded that curvilinear masks are more manufacturable based on actual mask write data from these studies.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 27 Nov 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

1.2 Multi-beam writers enable writing curvy masks

Multi-beam mask writers write using pixels. As shown in figure 3, an input shape like a circle will first be rasterized into pixels. These pixels will have a value that is equivalent to the percentage of pixel covered by the shape as shown by the grayscale image in Figure 3. Here white represent 100% coverage and black represents 0% coverage. The exposure time of time of the multi-beam writer at a given locations, also known as dose, is governed by the rasterized gray-scale value.

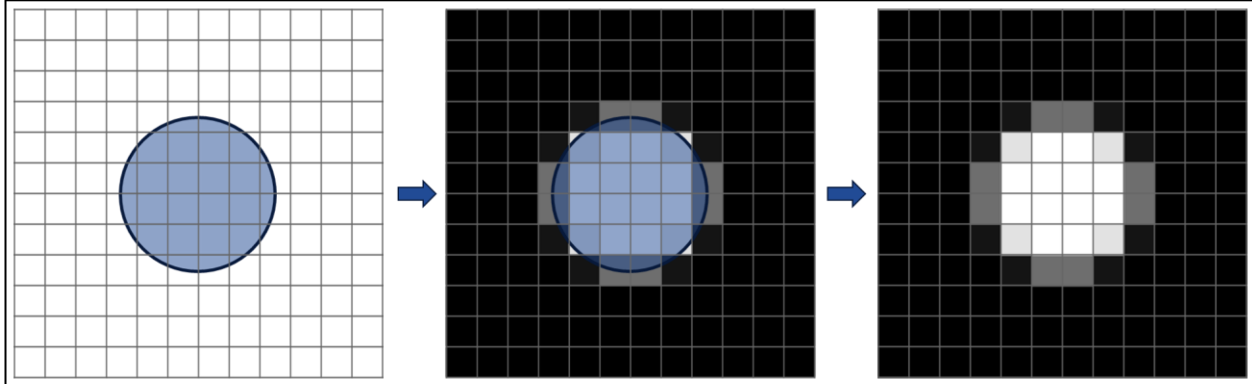

Figure 3. Rasterization by Multi-beam writers to write the curvy mask

Thus, any curvy shape can be rasterized and written on mask based on the dose governed by the pixel value.

1.3 Curvilinear data representation mitigates data path bottlenecks

Representing curvy shapes is challenging if we continue using traditional piecewise linear formats. The paper [7] talks about how using a curve format provide us similar information content with less data if we us a pixel-based computing aware piecewise curvy format. As shown in Figure 4, the red piecewise linear representation requires a lot of vertices to represent the curve. On the other hand, the dark blue piecewise curvy data, which can be based on Bezier or B-splines, can represent the same curve using much less control points. As long as they rasterize to the same dose values, they can be called pixel dose equivalent and thus they will produce the same mask written by the multi-beam mask writers.

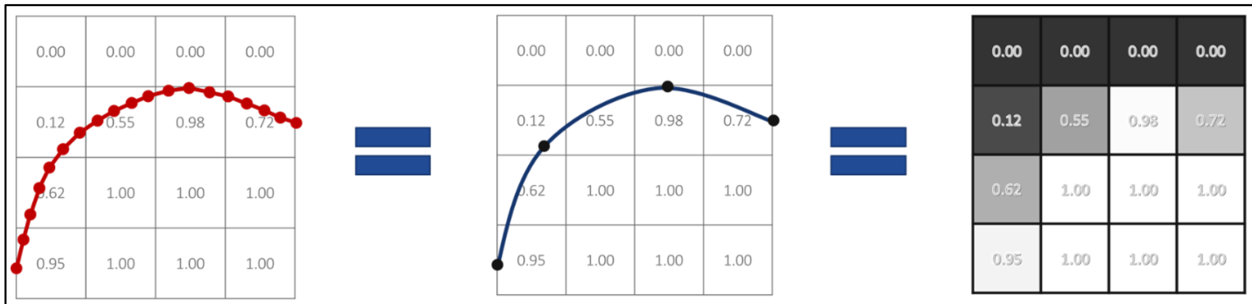

Figure 4. Pixel dose equivalence between piecewise linear and piecewise curvy data representation

Thus, pixel-based computing allows multibeam mask writers to write any mask and pixel dose equivalence allows limiting the data necessary to faithfully represent the desired mask.

2. PIXEL-BASED COMPUTING FOR CURVY MASKS PROCESSING

Mask writers write curvy masks using pixels. However, pixel-based computing can also be used in other mask processing steps. This enables mask processing for any curvy shape that can be sampled into pixels while satisfying the Nyquist-Shannon Sampling Theorem.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 27 Nov 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

2.1 Information theory in pixel domain mask processing

The paper [7] talks about how the intuition behind Nyquist criteria in pixel domain can be seen from grid alignment. In fact, if the input mask shape is sampled into pixels with sufficient resolution, the resulting shapes on masks are always preserved. This can be proved with the example in Figure 5. Figure 5(a) represents a curvy shape (i.e., oval) at grid alignments. Even though the ovals produce completely different sets of pixels after rasterizations, as shown in Figure 5(b), they produce similar shapes after simulation, as shown in Figure 5(c). This implies that as long as the data is sampled with sufficient sampling rate, it would faithfully represent the unique information in the desired curvy shape, after mask processing, across different pixel grid alignment.

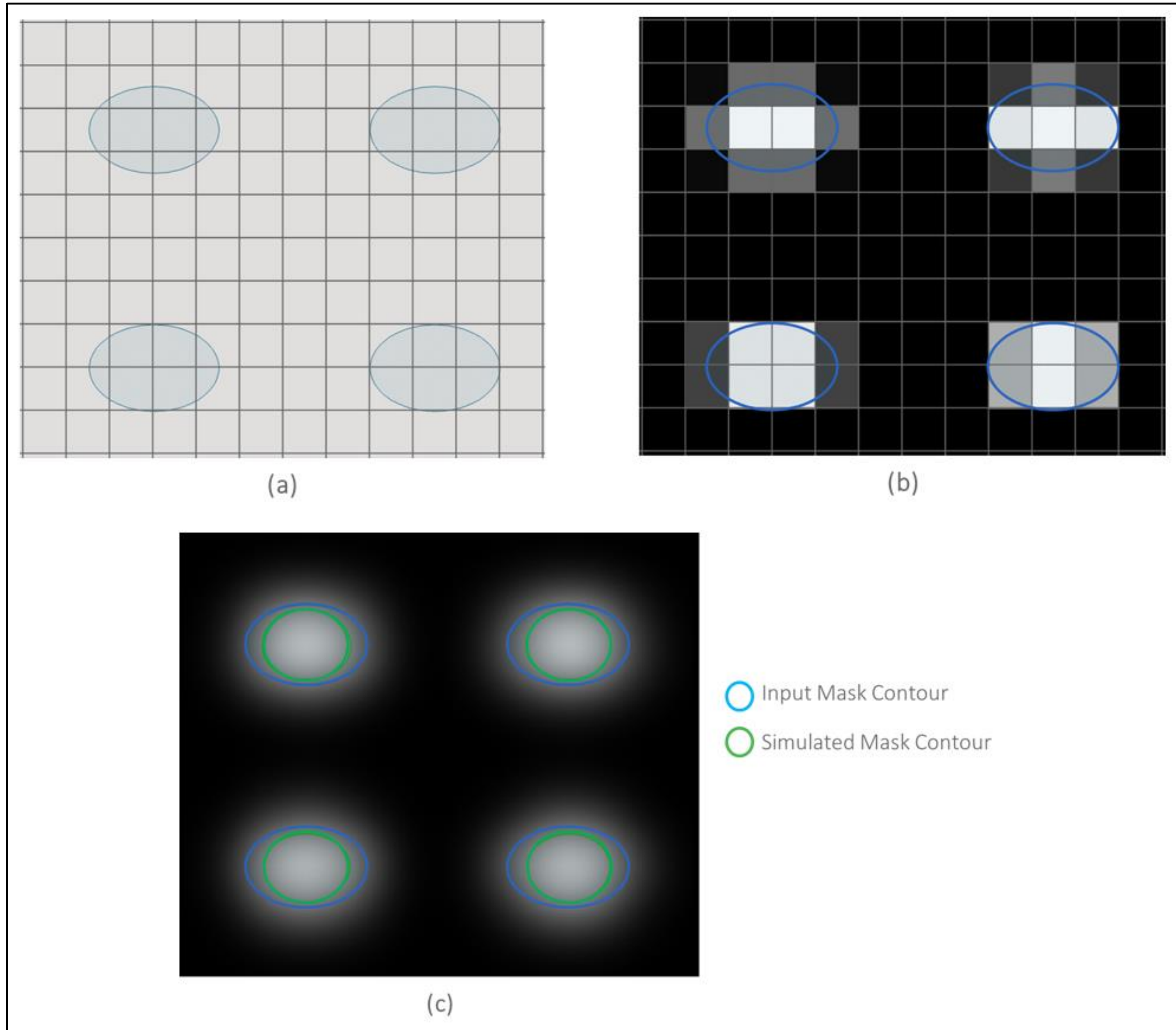

Figure 5. (a) Ovals at different grid alignments. (b) Pixels generated by rasterizing the ovals. (c) Simulation of the rasterized ovals showing equivalent simulated contours.

2.2 Mask rules for pixel-based mask processing

Mask writing process has certain limitations. Each generation of multi-beam mask writers have their own pixel size that governs the smallest feature representable using Nyquist criteria purely based on data representation and information theory. In addition to this, the printability of a feature is also governed by the physics and chemistry involved in the mask making processes like develop, etch, etc. A common methodology to estimate printability of a feature is to use a physical

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 27 Nov 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

model that can have multiple gaussians and calibrating it using some test patterns. This calibrated model is then used to simulate the mask shapes to verify printability.

Figure 6(a) shows a simple gaussian model simulation on test patterns of varying sizes. Here it can be seen that some of the shapes are missing in the simulation contour. This is an indication to the fact that the smaller shapes were beyond the limit of what is printable. In fact, such shapes are considered a Mask Rule Check (MRC) violation. Curvy shapes may require different set of MRC [13]. Figure 6(b) shows the dose-margin calculated based on the model for the shapes in Figure 6(a). Here dose margin can be defined as the slope of the simulated doses at a given point of interest. It can be computed at any point, but it is only interesting to know the dose slope near the threshold amount defining the simulated contour. So, Figure 6(b) is only showing dose-margin near the threshold dose amounts. Steeper dose-slope are considered good since a steeper slope means that a larger dose variation is required for unit change in the edge placement of simulated contour. In general, a region of feature with bad dose-margin is not reliably manufacturable as the mask shapes vary more with small changes in dose value. The manufacturable curves are band-limited as the mask processes act like a low-pass filter. Therefore, in order to ensure reliable manufacturing of mask features, we need to look at MRC violations like width, spacing and area violations. A band-limited shape is more likely satisfy MRC rules, and thus is more reliably manufacturable.

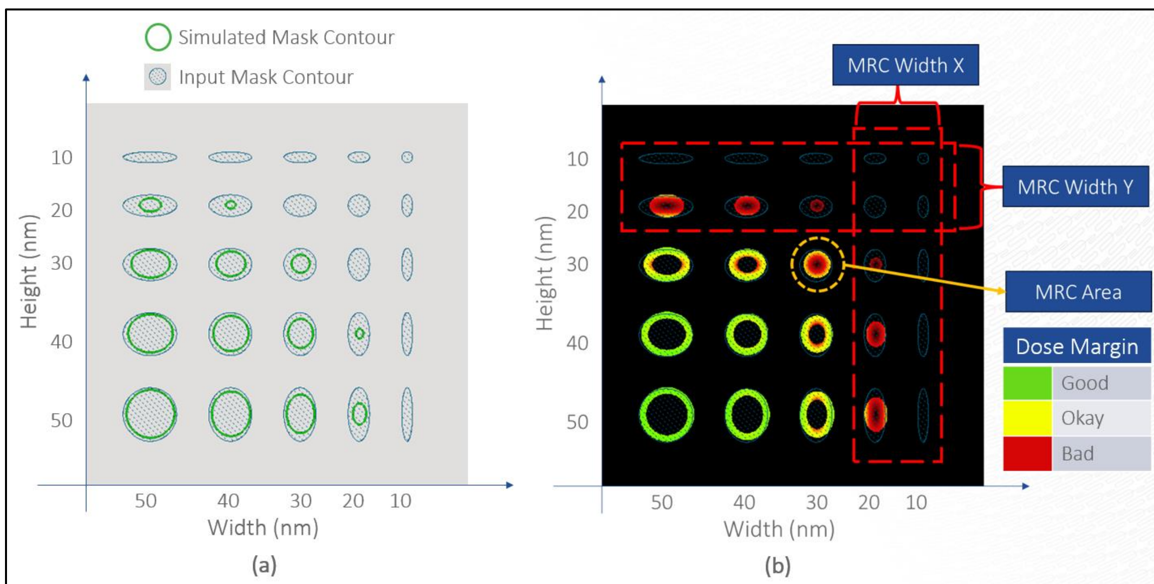

Figure 6. (a) Input test mask contour and corresponding simulated contour. (b) Simulation based dose-margin check and MRC checks

2.3 GPUs accelerated $O(p)$ computing

Mask processing in pixel domain is relatively niche technique in MDP softwares. Legacy mask data formats were mostly Manhattan, which does allow fast edge-based processing algorithms that can efficiently run on CPUs. However, with the advent of curvy masks, the legacy algorithms require modifications and they pose huge computing challenges. The runtimes of such algorithms are typically proportional to number of edges which can be represented as $O(\mathcal{E})$ in big O notations. Pixel-based computing puts an upper bound on the runtime since the algorithm runtimes are proportional to number of pixels or $O(p)$ in big O notations. Since the number of pixels covering the entire mask data is fixed for a given pixel size, the runtime is very predictable.

$O(p)$ may still be slower than $O(\mathcal{E})$ if we purely talk about solving using a single threaded system. So, it may appear that pixel-based computing is eventually going to be slower, especially for sparse data. This constraint can be resolved using GPUs instead of CPUs. As shown in Figure7, CPUs are designed for single instruction single data (SISD) applications, while GPUs are designed for single instruction multiple data (SIMD) applications [14]. GPUs are uniquely suitable for Pixel domain calculations due to the SIMD nature of pixel-based computing. GPUs have orders of magnitude more threads that can works in parallel as shown in Figure 7(b). The recently released H100 GPU has 16,896 cores [15] that allows massive parallelism. Most pixel algorithms involve performing the same operation on all the pixels. Thus, they can take

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 27 Nov 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

full advantage of GPUs threads. This may not be true for the $O(\mathcal{E})$ algorithms. The edge-based algorithms tend to be more SISD than SIMD and thus can run efficiently with CPUs. Thus, GPUs accelerated $O(p)$ computing enables fast and efficient pixel-based mask processing.

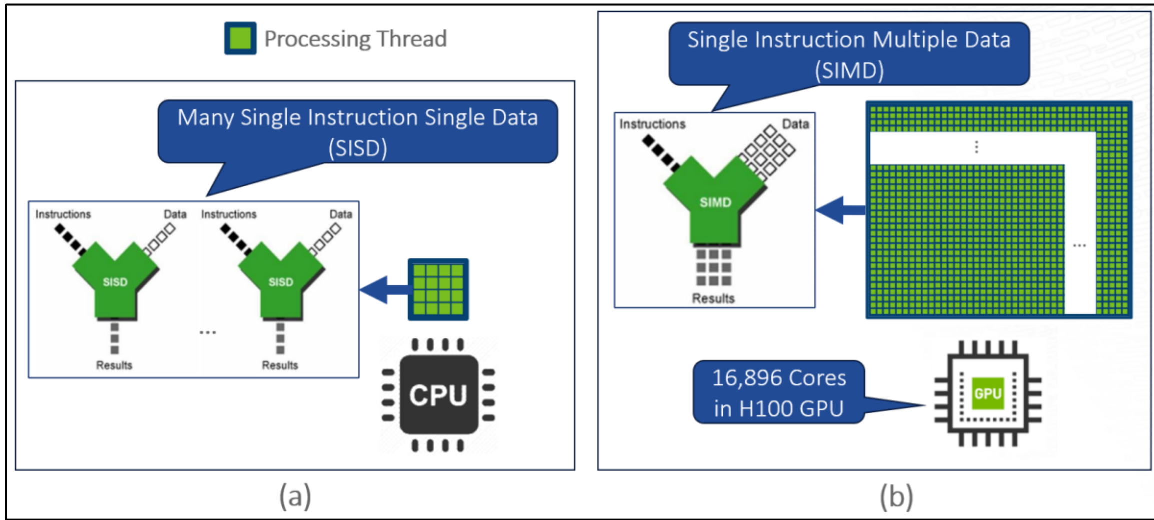

Figure 7. (a) SISD representation of CPU threads. (b) SIMD representation of GPU threads

3. CASE STUDY: SMOOTH SHAPES ARE MORE MANUFACTURABLE

The lithography process acts like a low-pass filter which tend to block all high frequency components in the mask data. This means that any data in photomasks that has a frequency higher than Nyquist Rate would not appear on the wafer. The Rayleigh Criteria [16][17] dictates the natural sampling resolution of the lithography process (SR_{litho}) is a function of the process constant (k_1), wavelength (λ) and numerical aperture (NA) as shown in equation (1).

$$SR_{litho} = k_1 \frac{\lambda}{NA} \tag{1}$$

Hence, effective Nyquist Rate (NR_{litho}) can be computed using equation (1) to get the following equation (2)

$$NR_{litho} = \frac{2}{SR_{litho}} = \frac{2\,NA}{k_1\lambda} \tag{2}$$

Figure 8 shows 2 different masks that would create the similar wafer as seen by the wafer simulated shape. This is indeed due to the low-pass filter impose by the lithography process. We can call both the masks shown in Figure 8 as equivalent nominal mask contours as that produce similar nominal wafer contours. Thus, band-limited curves are generally smooth and non-smooth curves will only pass band-limited information based on Nyquist limit.

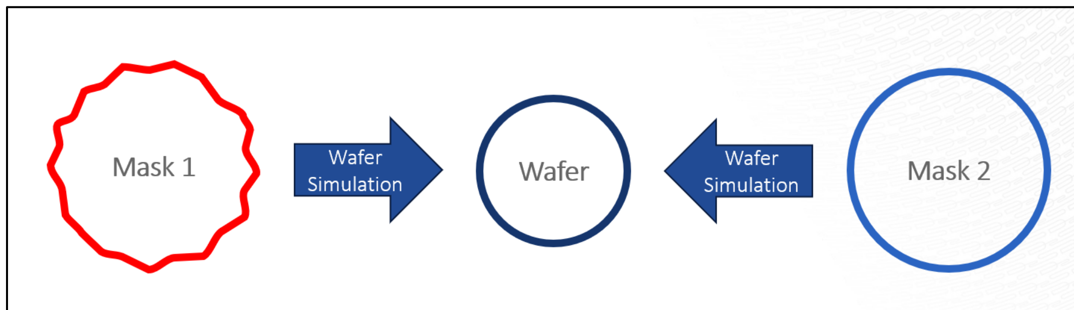

Figure 8. Equivalent nominal mask contours producing same simulated nominal wafer contour

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 27 Nov 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

3.1 Manhattanized Shapes Vs Smooth Shapes

The photomasks that are written using variable shaped beam (VSB) mask writers [18] require special considerations. Since, the write time of VSB writers is dependent on number of shots and since the VSB writers can only allow rectangular beams, the MDP softwares would produce Manhattanized mask data even for simple diagonal shapes as shown in Figure 9(a). With the advent of multi-beam mask writers, it is possible to write smoother shapes like Figure 9(b) without the need to Manhattanize the mask data. This allow reducing unnecessary complexity in the mask representation and MDP.

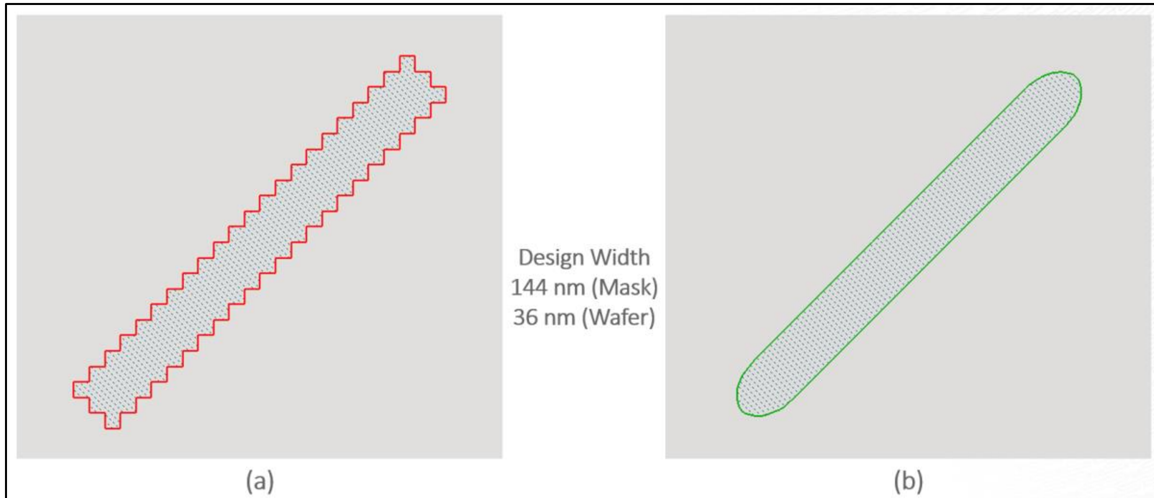

Figure 9. (a) Manhattan diagonal (b) Smooth diagonal

3.2 Mask process variation band (MPVB) is better for smooth shapes

Analyzing the mask data to predict the effect of mask process variation is important. This can be done through real test writes where the same mask shape can be written at various regions of the photomask and analyzed using scanning electron microscope (SEM) imaging and metrology. However, this methodology is time consuming and expensive. A model-based simulation can also be used to predict how the mask would appear physically after all the mask processes like develop, etch, etc. A mask process variation band (MPVB) can then be computed to predict the mask reliability. Here MPVB is defined as the process variation band due to 10% variation in dose but seen after full mask simulation which includes eBeam resist blur and etching models. The MPVB is measured in nanometers (nm) and smaller the magnitude of MPVB, the better it is. As shown in Figure 10, the MPVB is wider (4.8 nm) for Manhattan diagonal test while it is narrower (3.4nm) for smooth diagonals.

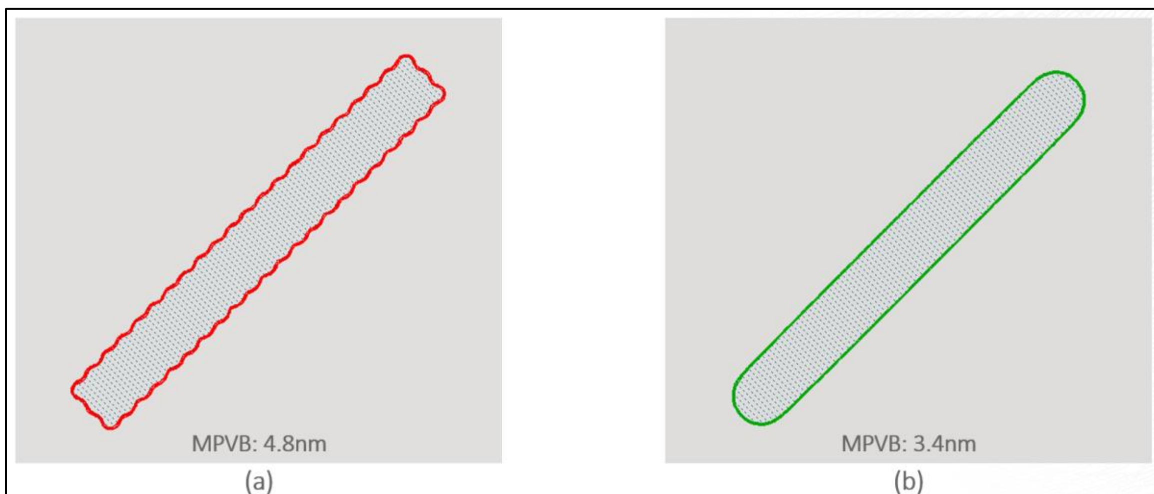

Figure 10. (a) MPVB measurement for Manhattan diagonal. (b) MPVB measurement for smooth diagonal

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 27 Nov 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

3.3 Mask error enhancement factor (MEEF) is better for smooth shapes

The impact of mask process variation on the wafer is also an important metric to measure. There have been many studies [4][19][20] that look into mask error enhancement factor (MEEF) to measure the amplification of mask error on the wafer. We can use equation (3) to measure MEEF

$$MEEF = \frac{\Delta CD_{Wafer}}{(\Delta CD_{Mask}/F)} \tag{3}$$

Where, ΔCD_{Wafer} is the amount of variation of the shape's critical dimension (CD) on wafer, ΔCD_{Mask} is the amount of variation of the shape's CD on photomask and F is the optical imaging reduction factor which is typically 4.

Numerically, a higher MEEF means that small errors on mask get amplified to larger errors on wafers. Thus, it is desirable to have a numerically smaller number for MEEF. Figure 11 shows the comparison of MEEF between the Manhattan diagonal and smooth diagonal. The MEEF is also compared near the line-end region and the mid-edge region. It can be observed that the MEEF is better for smooth diagonals compared to Manhattan diagonals. It can also be seen that the MEEF at mid-edge region is better than line-end region.

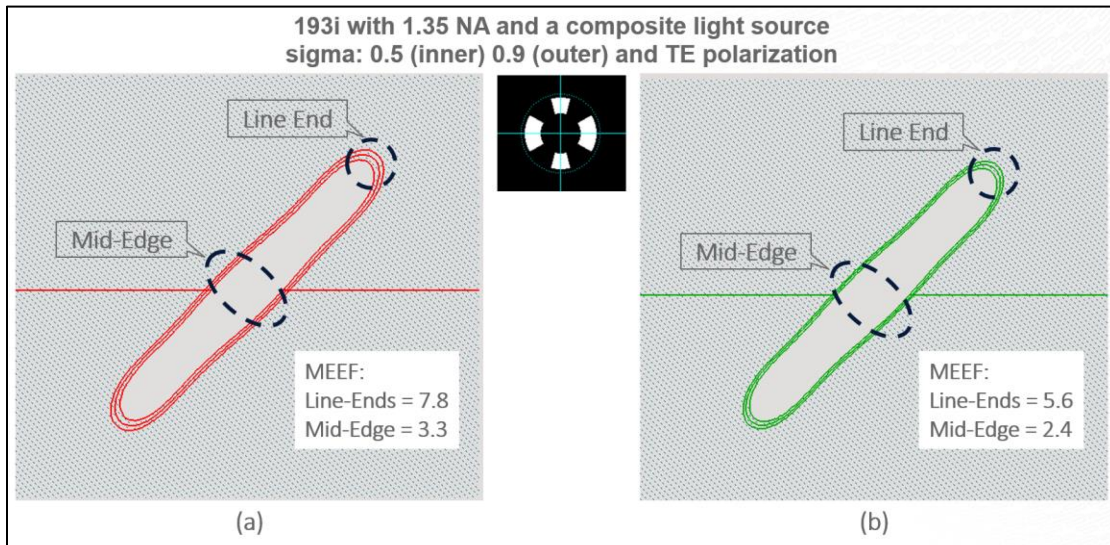

Figure 11. (a) MEEF measurement for Manhattan diagonals. (b) MEEF measurement for smooth diagonal

In fact, MEEF depends on the shape's perimeter & area. Mid-edge sees more local area compare to line-end so a unit of change in mask edge placement causes less change in area. This leads to less MEEF. Smooth diagonals have less perimeter which means that a unit change in edge placement contributes to less area change compared to Manhattan diagonals, thus resulting in less MEEF. So, we can say, MEEF is directly proportional to local perimeter (P_{Local}) and inversely proportional to local area (A_{Local}) as shown in equation (4).

$$MEEF \propto \frac{P_{Local}}{A_{Local}} \tag{4}$$

Band-limited shapes provide better MEEF as they tend to have less perimeter for given area. Therefore, it is desirable to use a bandlimited shape like the smooth diagonal instead of the Manhattan diagonals to get better MEEF.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 27 Nov 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

4. CONCLUSION

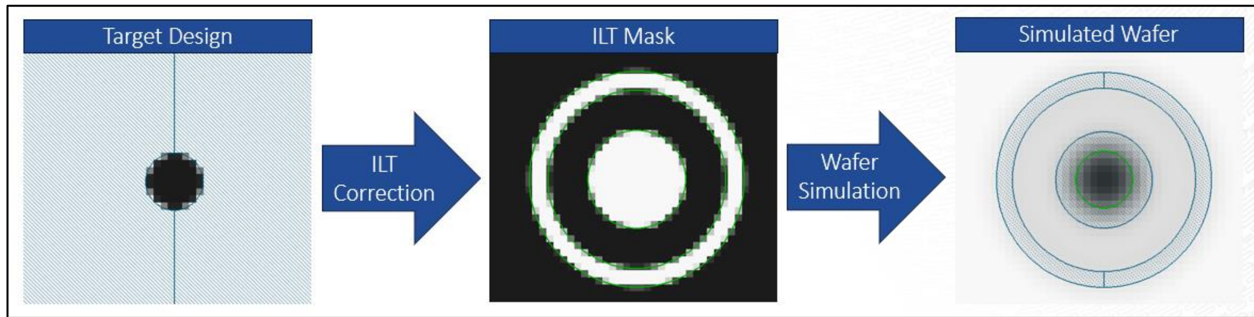

Figure 12. Pixel Based Curvy ILT Flow

Curvy design is enabled by pixel-based ILT. As shown in figure 12, curvy target design representing a circle can generated on wafer using curvy ILT masks using pixel based ILT computation. On the other hand, curvy mask like Figure 13(a) is enabled by multi-beam writer (Figure 13(b)) and $O(p)$ computing. $O(p)$ computing is enabled by GPU acceleration as pixel algorithms are more efficient on GPUs compared to CPUs due to the SIMD nature of massively parallel GPU threads (Figure 13(c)). This helps improve the throughput and total turnaround time for MDP. Apart from the runtime advantages, the cases study from section 3 also proves that if we can start from curvy designs, then with help of Pixel-Based ILT enabled by multi-beam writers and $O(p)$ computing, we can manufacture more reliable masks and good quality wafers.

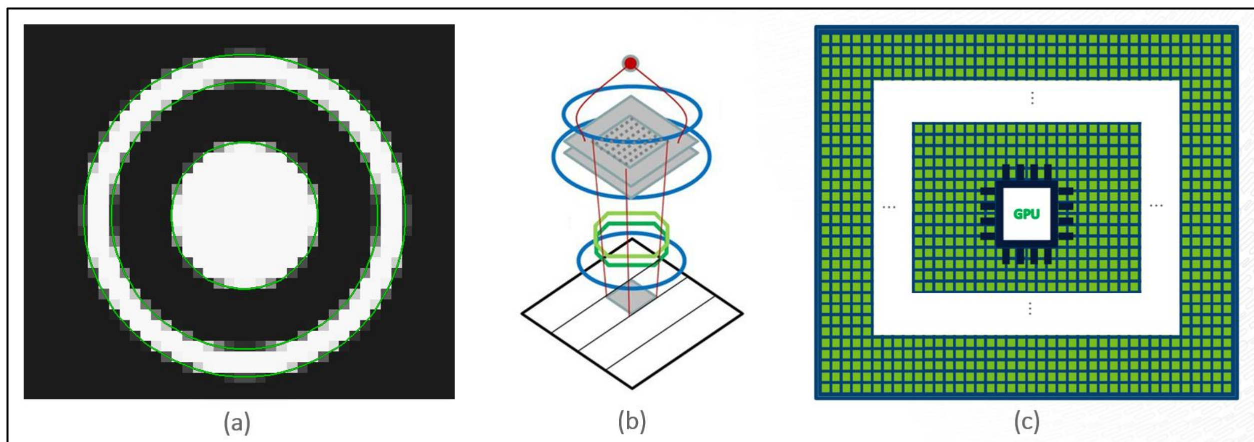

Figure 13. (a) Pixel based ILT Mask. (b) Multi-beam mask writer. (c) GPU with massively parallel computations

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 27 Nov 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

REFERENCES

[1] Matsumoto, H., Yamaguchi, K., Kimura, H. and Nakayamada, N., "Multi-beam mask writer, MBM-2000," Proc. SPIE 11908, Photomask Japan 2021: XXVII Symposium on Photomask and Next-Generation Lithography Mask Technology, 119080L (2021)

[2] Pang, L., "Inverse lithography technology: 30 years from concept to practical, full-chip reality," J. Micro/Nanopattern. Mats. Metro. 20(3) 030901 (2021)

[3] Choi, Y., Fujimura, A., Shendre, A., “Curvilinear masks: an overview”, Proc. SPIE 11855, Photomask Technology 2021, 118550U (2021)

[4] Pearman, R., Ungar, J., Shirali, N., Shendre, A., Niewczas, M., Pang, L., Fujimura, A., "How curvilinear mask patterning will enhance the EUV process window: a study using rigorous wafer+mask dual simulation," Proc. SPIE 11178, Photomask Japan 2019: XXVI Symposium on Photomask and Next-Generation Lithography Mask Technology, 1117809 (2019)

[5] Choi, J., Ryu, S., Lee, S., Kim, M., Park, J., Buck, P., Bork, I., Durvasula, B., Gharat, S., Rao, N., Pai, R., Koranne, S., and Tritchkov, A., "Study on various curvilinear data representations and their impact on mask and wafer manufacturing," J. Micro/Nanopattern. Mats. Metro. 20(4) 041403 (2021)

[6] Choi, J., Ryu, S., Lee, S., Kim, M, Lee, S., Buck, P., Durvasula, B., Gharat, S., Liubich, V., "Status of curvilinear data format working group", Proc. SPIE 12325, Photomask Japan 2022: XXVIII Symposium on Photomask and Next-Generation Lithography Mask Technology, 1232508 (2022)

[7] Shendre, A., Fujimura, A., Niewczas, M., Kronmiller, T., “You don’t need 1nm contours for curvilinear shapes: pixel-based computing is the answer,”, Proc. SPIE 12293, Photomask Technology 2022, 1229307 (2022)

[8] Pearman, R., Shendre, A., Syrel, O., Zable, H., Bouaricha, A., Niewczas, M., Su, B., Pang, L. and Fujimura, A., "Full-chip GPU-accelerated curvilinear EUV dose and shape correction", Proc. SPIE 10451, Photomask Technology 2017, 1045108 (2017)

[9] Singh, V., "Computational lithography: accelerating the future", Proc. SPIE 12052, DTCO and Computational Patterning, 1205207 (2022)

[10] Baji, T., "GPU: the biggest key processor for AI and parallel processing", Proc. SPIE 10454, Photomask Japan 2017: XXIV Symposium on Photomask and Next-Generation Lithography Mask Technology, 1045406 (2017)

[11] Ezequiel, R., ILT curvilinear mask design for advanced memory, Micron, https://www.ebeam.org/docs/ilt-curvilinear-mask-designs-for-advanced-memory.pdf (2020)

[12] Pearman, R., Venkatesan, R., Sundaramurthy, R., Straney, P., Rice, Z., Conner, R., Grunes, H.,"A CD uniformity study comparing MRC-constrained all-angle to Manhattan OPC corrections on EUV", Proc. SPIE 12495, DTCO and Computational Patterning II, 124950F (2023)

[13] Pearman, R., Ungar, J., Shirali, N., Shendre, A., Niewczas, M., Pang, L., Fujimura, A., “Adopting curvilinear shapes for production ILT: challenges and opportunities”, Proc. SPIE 11148, Photomask Technology 2019, 111480T (2019)

[14] Flynn, M. J., and Rudd, K. W., “Parallel architectures”, ACM computing surveys (CSUR), 28(1), 67-70. (1996)

[15] Choquette, J. "Nvidia hopper h100 gpu: Scaling performance." IEEE Micro (2023).

[16] Ma, X., Arce, G.R., [Computational lithography], John Wiley & Sons, 10-14 (2011)

[17] Yen, A., "Rayleigh or Abbe? Origin and naming of the resolution formula of microlithography," J. Micro/Nanolith. MEMS MOEMS 19(4) 040501 (2020)

[18] Matsui, H., Kamikubo, T., Nakahashi, S., Nomura, H., Nakayamada, N., Suganuma, M., Kato, Y., Yashima, J., Katsap, V., Saito, K. and Kobayashi, R., "Electron beam mask writer EBM-9500 for logic 7nm node generation," Proc. SPIE 9985, Photomask Technology 2016, 998508 (2016);

[19] Mack, C. A., "Mask linearity and the mask error enhancement factor", Microlithography World, 8(1), 11-12. (1999)

[20] Granik, Y., "Generalized mask error enhancement factor theory," J. Micro/Nanolith. MEMS MOEMS 4(2) 023001 (2005)

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 27 Nov 2023
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use