TrueMask® ILT Backgrounder
Stitchless Full-Chip ILT in a Day

Executive Summary

Inverse lithography technology (ILT) has been seen as a promising solution to many of the challenges of advanced-node lithography, whether optical or EUV. Curvilinear ILT, in particular, has been shown to achieve the best process window of any approach. However, the runtimes associated with this computational technique have limited its practical application, as full-chip ILT is measured in weeks when using conventional ILT. This is why, until now, ILT has been used for critical "hotspots" on chips, but has not been used for entire chips. 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. D2S has adopted an entirely new, stitchless approach, creating a holistically conceived, purpose-built system for ILT. This 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 novel approach, systematically designed for ILT and GPU acceleration, makes full-chip ILT a practical reality in production for the first time.

Background: The History of ILT

First introduced more than 10 years ago, ILT creates ideal lithography results through a mathematically rigorous inverse approach that determines 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 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 term of wafer-pattern fidelity and process window.

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. Specific EUV effects (the non-normal, 6-degree 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, making curvilinear ILT the best solution for EUV masks.

However, two major obstacles have kept ILT from being widely applied. One of these barriers – the ability to write curvilinear mask patterns – has been removed recently by introduction of

multi-beam mask writers, which can write any shape without time penalty. The other major barrier – ILT run time – was still left to be overcome.

Run Time: *The* Challenge for Full-Chip ILT

The biggest barrier to full-chip ILT has been 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 weeks-long runtimes that result.

The standard approach for computations that are too lengthy to be practical is to divide the task (or in this case, the chip) into partitions, and have the computations for each partition run in parallel to save time. Then, the partitions are "stitched" back together. However, because the physics of any mask feature are impacted by the features adjacent to it, and because perfect realignment of the patterns at the partition boundaries is difficult, partitioning introduces errors that must be corrected once the partitions are stitched together. The correction of these "stitching errors" then requires additional, often recursive computations, such that the time savings afforded by partitioning is substantially eroded.

As a result, commercial applications of ILT have been limited, and have focused mainly on smaller, high-risk portions of masks. A high-volume, full-chip ILT solution has been elusive.

The Solution: Get Rid of the Stitches

The rise, in the last decade, of the use of general-purpose graphics-processing unit (GP-GPU) 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. This is in contrast to central-processing unit (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.

Of course, 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 terms of acceptable turnaround time. Why?

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 re-computation 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.

TrueMask ILT: Stitchless, Curvilinear Full-Chip ILT In a Day
Of course, such a giant GPU/CPU pair does not exist. However, by taking a “from the ground up” approach, D2S was able to build an ILT-specific computing appliance that could *emulate* a giant GPU/CPU pair (see Figure 1).

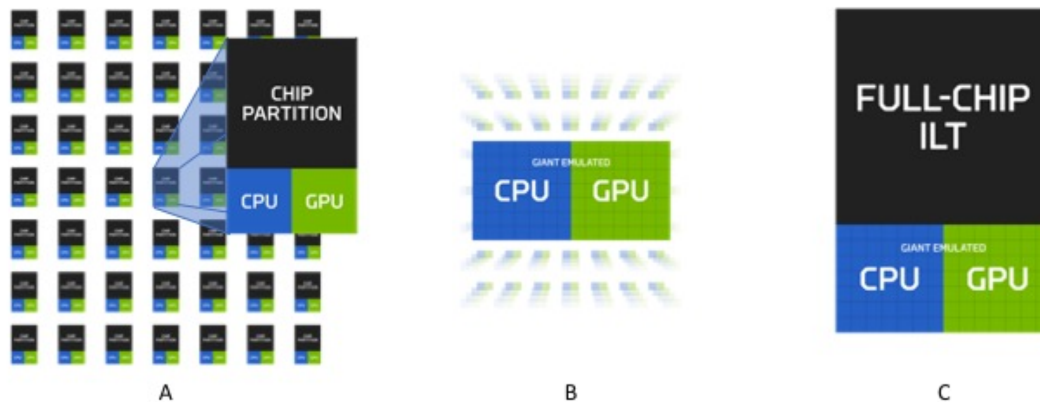

Figure 1: The conventional “partition and stitch” approach (A) is too slow and leads to errors. TrueMask ILT has been holistically designed so that, while it is composed of many CPU/GPU pairs, it emulates a single GPU/CPU pair (B). This approach enables TrueMask ILT to iterate on the entire chip as a whole, thus avoiding stitching errors (C).

This approach didn’t stop with the hardware, but rather included every component of a holistically conceived, purpose-built system – hardware, software, models, visualization, verification, etc. – that is designed and implemented specifically for GPU-acceleration and for full-chip ILT computation. Every aspect of the physics and chemistry of wafer lithography and processing, including litho simulations and mask and wafer models, was examined and optimized synergistically throughout the system in order to reap the largest potential run-time benefits without compromising computational accuracy.

The product of more than ten years of development, D2S TrueMask ILT is the first commercial product to create accurate, full-chip ILT in a single day.

<u>Stitchless</u>
Chip partitioning and parallel computing is the most common approach 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 (see Figure 2).

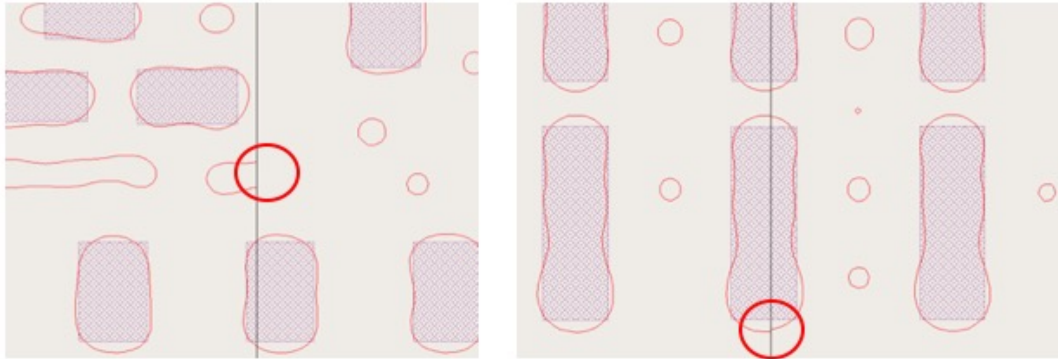

Figure 2: Stitching errors occur when a chip is partitioned for parallel computing and re-assembled.

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 TrueMask ILT so that the entire chip could be optimized at once. The D2S CDP has been purpose-built specifically to address simultaneous full-chip optimization. While it contains dozens of GPU-CPU pairs, TrueMask ILT, 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 proximity effects across the chip are accounted for with each update, and eliminates stitching errors (see Figure 3).

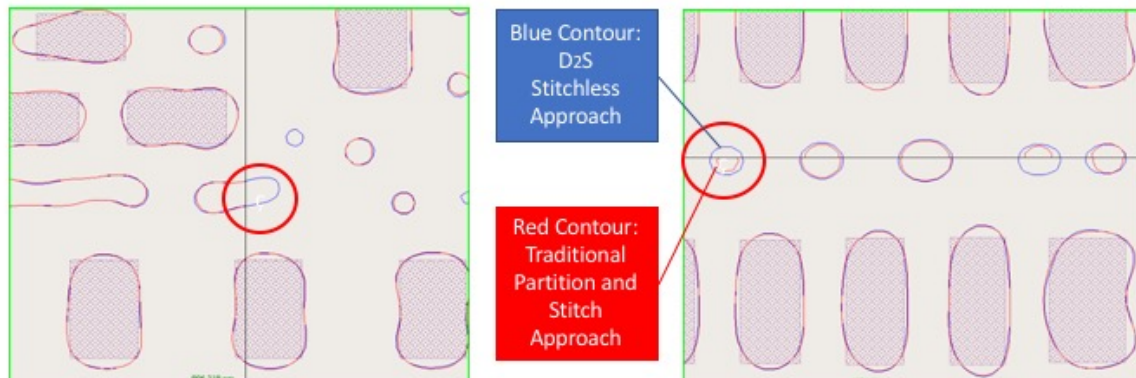

Figure 3: The red contours show the mask shapes produced with the conventional “partition and stitch” approach. Stitching errors are evident, as some shapes on boundaries are incomplete or distorted. The blue contours show the same mask shapes produced by TrueMask ILT and its stitchless approach. The masks shapes are complete and accurate.

Curvilinear

Because nothing in nature (including the physics of semiconductor manufacturing) makes 90-degree corners, manufactured masks and wafers *are all curvilinear*, even if the input geometries are rectilinear (see Figure 4). In fact, curvilinear shapes with certain minimum curvatures of shapes and spaces have been shown to be more reliably manufacturable than rectilinear shapes[1].

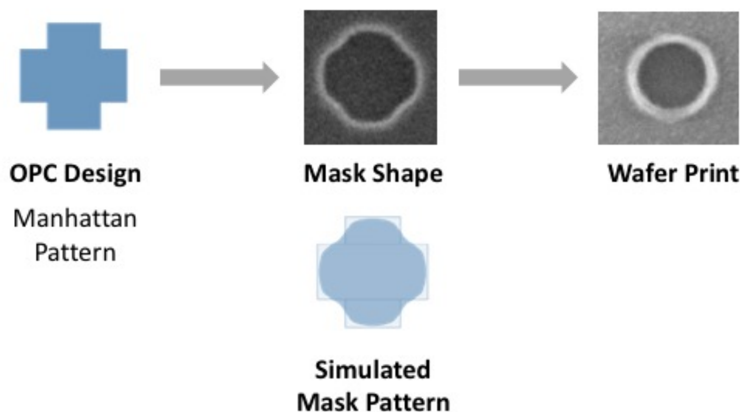

Figure 4: All shapes on masks and wafers are curvilinear, even if the input geometries are rectilinear.

ILT is a mathematical approach that naturally produces curvilinear shapes. Traditionally, extra computation time has been needed to Manhattanize the curvilinear ILT shapes because variable-shaped beam (VSB)-based mask writing could not process curvilinear mask shapes within practical runtimes. With multi-beam mask writing now available, curvilinear shapes no longer require additional time to write. TrueMask ILT was built to leverage the power of this new world of multi-beam-based mask writing and is optimized for curvilinear mask output.

TrueMask 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; TrueMask ILT will handle these designs with ease.

Uniquely, TrueMask ILT 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, TrueMask ILT computes, in essence, a rasterized image of the entire chip all at once, iterating on the full-chip ILT solution as a whole.

---

[1] Pearman, Ryan, et al, "How curvilinear mask patterning will enhance the EUV process window: a study using rigorous wafer+mask dual simulation," SPIE Photomask Japan, 2019.

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 weeks-long turnaround time for full-chip ILT was unacceptable. 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 run-time was not an issue. The unique approach of TrueMask ILT makes full-chip ILT a practical reality.

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 the shortest time possible, which commonly constrains the budgets for OPC and ILT process time to one day. TrueMask ILT is the first commercial ILT solution that delivers full-chip ILT within this time constraint.

Conclusion: ILT Vision Realized, At Last

For more than 10 years, the semiconductor industry has recognized the value of ILT in addressing the challenges of advanced-node lithography and improving ever-shrinking process windows. Until now, weeks-long runtime has been an insurmountable barrier to using ILT as a full-chip solution. Partitioning and stitching – the traditional approach to runtime improvement – have proven to be unsuccessful for full-chip ILT because of the computational time required to correct inevitable errors on partition borders.

By embracing a unique, holistically conceived, purpose-built system of GPU-accelerated hardware and software that emulates a single giant GPU/CPU pair, D2S TrueMask ILT iterates and optimizes the entire chip as a whole, making stitchless, curvilinear, full-chip ILT – and its undisputed process-window improvements – a practical reality.