Affordable and Process Window Increasing Full Chip ILT Masks

Guangming Xiao[*1], Dave Irby[1], Tom Cecil[1], David Kim[1]
[1] Luminescent Technologies, Inc., Palo Alto, CA 94303, USA

Shuichiro Ohara[2], Isao Aburatani[2]
[2] Nippon Control Systems Corporation, 2-7-9, Shin-Yokohama, Kouhoku-ku, Yokohama 222-0033, Japan

ABSTRACT

To enable Inverse Lithography Technology (ILT) for production as one of the leading candidates for low-k_1 lithography at 32nm and below, one major task to overcome is mask manufacturability including mask data fracturing, MRC constraints, writing time, and inspection. In prior publications[1,2], it has been shown that the Inverse Synthesizer (IS™) produces ILT full chip mask of contact layer with comparable mask write time with conventional OPC while maintaining the significant litho gains of ILT mask.

To fully integrate ILT masks into production for all layers including line and space layers such as poly layer, a number of areas were investigated to further reduce ILT mask complexity and total e-beam shot count. These areas include flexible controls of SRAF placements with respect to local feature sizes, improved Manhattan algorithm, topology based variable Manhattan segmentation, jog alignment and mask data fracture optimization. The impact of these approaches on e-beam shot count and lithography performance of ILT masks is presented in the paper.

Keywords: Inverse lithography technology (ILT), Sub-resolution assist feature (SRAF), Resolution enhancement technology (RET), e-beam shot count, e-beam mask write time, 32nm & below, low-k1 lithography, mask cost reduction

1. INTRODUCTION

Improvements in resolution of exposure systems have not kept pace with increasing density of semiconductor products. Alternatives such as EUV have not developed at the pace originally forecasted, so for the next few years we must rely on 193 nm systems using water immersion. In order to keep shrinking circuits using equipment with the same basic resolution, lithographers have turned to options such as double-patterning (DP), and resolution enhancement technology (RET) beyond model-based OPC in the search for optimal mask patterns. Inverse Lithography Technology (ILT) is becoming one of the strong candidates in 32nm and below in low-k_1 lithography regime. It enables computation of optimum mask patterns to minimize deviations of images from their targets not only at nominal but also over a range of process variations, such as dose, defocus, and mask CD errors. Masks computed through the use of ILT are known to provide significantly better lithographic performance than conventional rule-based and even model-based OPC [3-6].

In a previous paper [1], a simple method of optimizing segmentation length to bring down mask write time to conventional OPC level without compromising much of ILT mask litho performance was introduced. It concluded that ILT mask of contact layer shows no issue with full chip mask manufacturability of adopting in production, comparing its data volume, e-beam write time and inspection of ILT full-chip to traditional OPC results. However, for poly gate layer, ILT mask exhibited larger data volume and longer write time that were not desirable for mask shop's cost of ownership. It concluded that there are future challenges in implementing ILT mask for poly gate layers in production and thus further write time reduction for poly gate layers was needed for production implementation.

This paper extends further improvements in e-beam write time for poly gate layer by implementing improved Manhattan segmentation algorithm, selective SRAF placement with local CD awareness, topology based Manhattan conversion and jog alignment. To validate our work, NCS PATACON mask data fracturing tool was used to fracture the ILT masks for

* gxiao@luminescent.com

shot count calculations. As the number of shot counts of fractured mask is directly proportional to the e-beam write time[7], it provides a means of estimating mask write time. Results of up to 3X shot count reduction on poly gate layer, comparing to previous results, will be shown in this paper. Obviously, shot count reduction on contact layer will be an added benefit as well, further reducing its cost of ownership. Furthermore, NCS slide option fracturing technique can maintain overall shot counts of ILT masks and can reduce slivers during data sizing in mask shop.

2. MASK COMPLEXITY CONTROL WITH SELECTIVE SRAF ILT

2.1 Local CD width and space bitmaps

A chip design is typically not homogenous in pattern density, shape, feature width and spacing. To capture local environment variation in optimizing a mask solution can help achieve more desirable result by making RET aware of local environment. As discussed in previous section, custom topology can be defined based on many local properties and one of these properties is local critical dimension (CD) width and spacing. In order to obtain the local CD width and spacing information, two bitmaps are generated for input target during inversion.

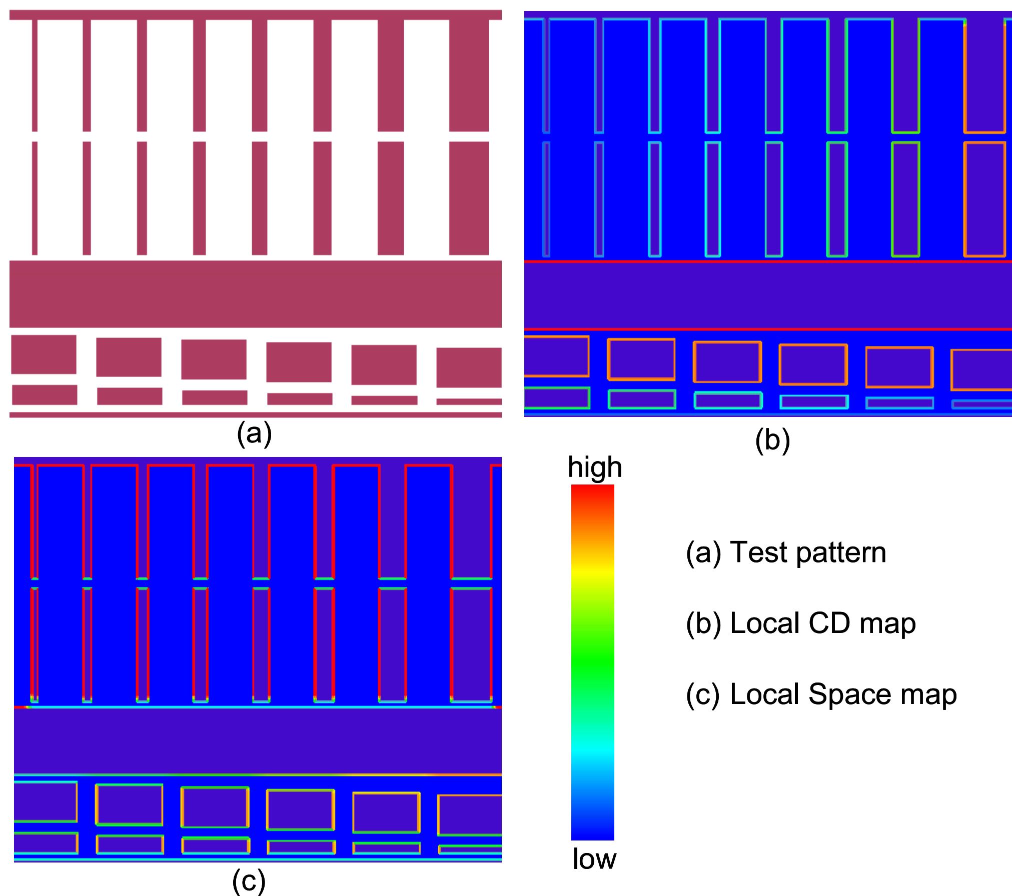

Figure 1. Local CD width and space mapping of a design target. (a) Test pattern design with varying feature sizes and spaces. (b) Local CD width map of test pattern shows color coded CD variation across the design. (c) Local space map displays color coded spacing between features across the design.

To demonstrate how local CD width and space information is generated, a test design was created with various feature width and spacing as shown in Figure 1 (a). Figure 1 (b) and (c) show color renderings of local CD width and space bitmaps of the test pattern. Clearly both local CD width and spacing information is captured visibly with color variations

on the bitmaps. This information provides the fundamental components for local environment aware ILT which will be detailed in next a few sections.

2.2 Local CD aware selective SRAF

As one of the most effective RET techniques, SRAF has been extensively applied to many design layers in low k_1 lithography and the complexity of SRAF has been increased tremendously over the years to keep up with the design shrink. ILT model based SRAF has demonstrated its significantly better performance in process window (PW) over conventional OPC approaches [2,4,5]. With increased performance with SRAF insertion, the mask write time is increased accordingly. One way to curb the increasing write time from aggressive SRAF is to apply SRAF only to regions where SRAF insertion is required in order to achieve required PW. For line and space layers such as poly gate or active layers, this is especially true due to its large variation of feature sizes. By not applying SRAF to larger non-critical patterns, it reduces the mask complexity and writing time without compromising wafer yield.

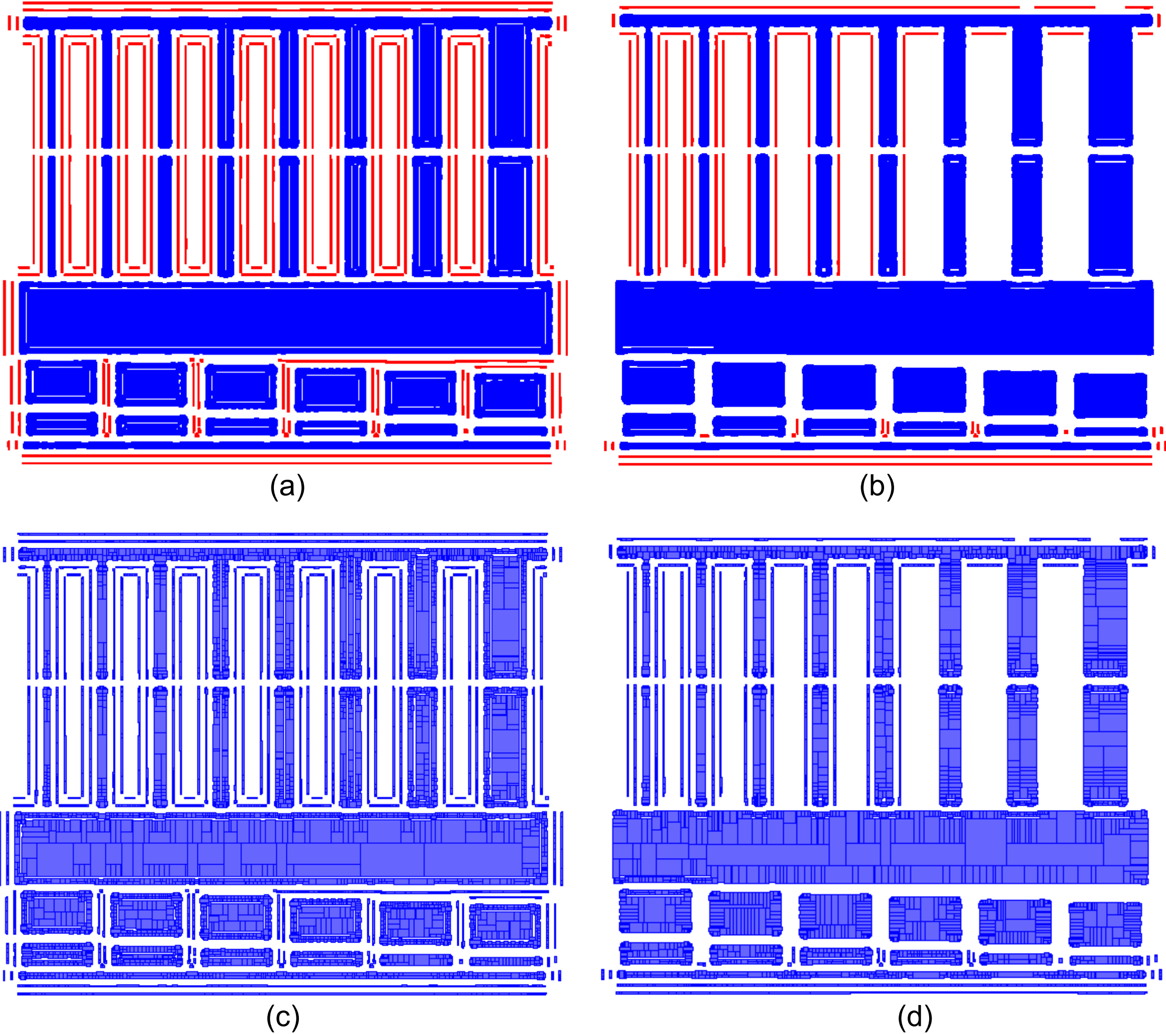

Figure 2. Comparison of SRAFs insertion with and without local CD awareness. (a) ILT mask with traditional ILT SRAF insertion where SRAFs appear at all possible places. (b) ILT mask with new local CD aware selective SRAF insertion. (c) Fractured VSB figure map of ILT mask in (a). (d) Fractured VSB figure map of ILT mask in (b).

Traditionally ILT generates SRAF at all places where there is room for insertion to maximize PW everywhere. At clip level for R&D, data volume and write time is not a major concern. At full-chip level, however, mask complexity and write time has serious impact on development cycle and turnaround time. Instead of allowing SRAF to be placed at all possible places as illustrated in figure 2(a), SRAF should be selectively placed only to these locations require SRAFs to achieve enough process window. Figure 2(b) demonstrates the result of ILT mask with local CD aware SRAF insertion. As shown in figure 2(b), two bands of exterior SRAFs are applied to small critical structure; with an increase of CD size, two-band SRAF gradually transitions into single band SRAF; eventually when feature size is large enough, no SRAF is generated at all. Similarly for interior SRAF, it only appears at locations of narrow critical space to counter bridge at defocus and hence satisfy PW requirements. Figure 2(c) and 2(d) show fractured VSB figures of mask in 2(a) and 2(b), respectively. Fractured VSB figure count directly corresponds to e-beam shot count on mask writer, hence mask write time[1, 7]. The total number of figure count in this example has 40% reduction with selective SRAF placement than original SRAF everywhere approach. This is a big step toward simpler mask for line/space layers. However, it alone is not enough to cut poly layer write time without acceptable level [1]. From figure 2(c) and 2(d), a large amount of small VSB figures are generated along relatively straight edges and corners. To further reduce to shot count, an improved Manhattan segmentation was implemented which will be introduced in section 3.

3. MASK WRITE TIME REDUCTION WITH SMART MANHATTAN CONVERSION

3.1 Manhattan conversion with improved handling of small jogs and corners

During mask pattern transfer process in mask shops, a sharp corner in the data will become a rounded corner on reticle after mask manufacturing flow due to e-beam resolution and effects from resist and etch processing. By taking into account of mask corner rounding, it allows coarser Manhattan corners when converting a curvy ideal mask from ILT without losing much of image fidelity. Figure 3(a) shows fractured VSB figure count from old Manhattan conversion and figure 3(b) shows the fracture figure count from improved Manhattan conversion. At the corners of line ends, new Manhattan algorithm uses much coarser segmentation than old Manhattan and results in significant figure count reduction. In addition to better handling of corners, new Manhattan eliminates tiny jogs often have negligible effects on wafer images. Along the long edges of main feature and SRAFs, the number of shots from new Manhattan in figure 3(b) is substantially less than that of old Manhattan in figure 3(a). It is most dramatically shown in SRAF where only one shot for SRAF in new Manhattan where the old Manhattan has far more than one shot. For a number of line/space design layers tested, a 2X to 4X reduction in total shot count can be achieved with new Manhattan than old Manhattan conversion. Figure 3(c) and 3(d) show nominal image contours from masks in 3(a) and 3(b), respectively. Both masks generate good nominal images meeting target. Figure 3(e) is a zoom-in image of line-ends region with overlay of both contours from new and old Manhattan masks. For smooth long edge and line ends, new and old Manhattan have little difference in nominal image contour; At corners regions, more pronounced difference can be seen. The new Manhattan mask exhibits slightly higher edge placement error (EPE) but still maintains well shaped contours. More detailed results on EPE and process window comparison between old and new Manhattan will be given in section 3.4.

3.2 Topology based Manhattan segmentation

In addition to fully automatically generated internal topology markers in Luminescent Inverse Synthesizer™, flexible custom topology markers can be created on top of internal topology markers for maximized user control for optimal mask solution [8]. Custom topology markers can be defined based on local feature width or space, design edge length, corner type, adjacent feature type, etc. Each custom topology can be assigned with its own weight and tolerance in inversion cost function for optimal mask solution. Furthermore, each custom topology marker can be used to control mask complexity during Manhattan conversion of ideal mask. For example, finer Manhattan segmentation can be used for regions marked with critical topology and coarser Manhattan segmentation can be used for regions marked with less critical topology. Consequently, Manhattan conversion with topology based variable degree of coarseness can maximize the simplification of curvy ideal mask at each location based on its criticality rather than treating everywhere equally across a full chip.

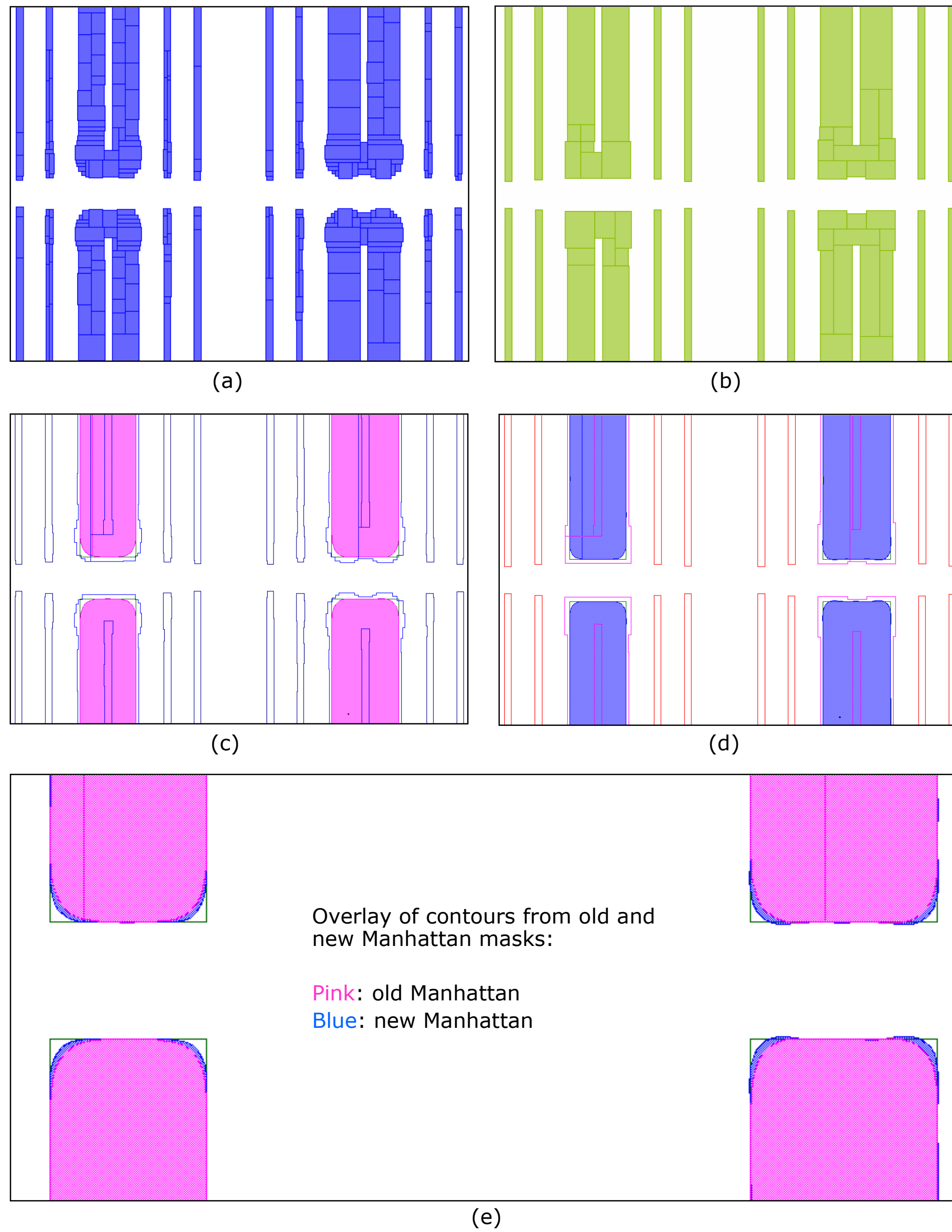

Figure 3. Comparison of old and improved Manhattan conversion. (a) Fractured VSB figures of old Manhattan mask. (b) Fractured VSB figures of new Manhattan mask. (c) Nominal image contour from old Manhattan mask in (a). (d) Nominal image contour from new Manhattan mask in (b). (e) Overlay of nominal image contours from both masks.

To demonstrate topology based Manhattan segmentation, a simple test pattern was design with 60nm lines and 150nm blocks as shown in figure 5. The tighter space regions have 60nm gap between neighboring features. Two types of custom topologies are defined here for demonstration purpose. One (TOPO1) is defined to mark regions with either 60nm line width or 60nm space. Another one (TOPO2) is defined to mark larger feature with width and space >140nm in this case. For legibility purpose, only custom markers are displayed in figure 5(a) and all other internal markers are hidden from view. Two different degrees of coarseness of Manhattan segmentation are applied independently for TOPO1 and TOPO2. Figures 5(b), (c), (e) and (f) show fractured VSB figures of masks from mixing different degree of coarseness between TOPO1 and TOPO2 during topology based new Manhattan conversion: 5(b) TOPO1=fine,

TOPO2=fine; 5(c) TOPO1 = fine, TOPO2 = coarse; 5(e) TOPO1 = coarse, TOPO2 = fine; 5(f) TOPO1 = coarse, TOPO2 =coarse. As shown in figure 5, the coarseness of a given mask edge is controlled locally by the topology type it is associated with. The shot count of regions covered by TOPO1 and TOPO2 varies independently according to the coarseness applied during Manhattan conversion.

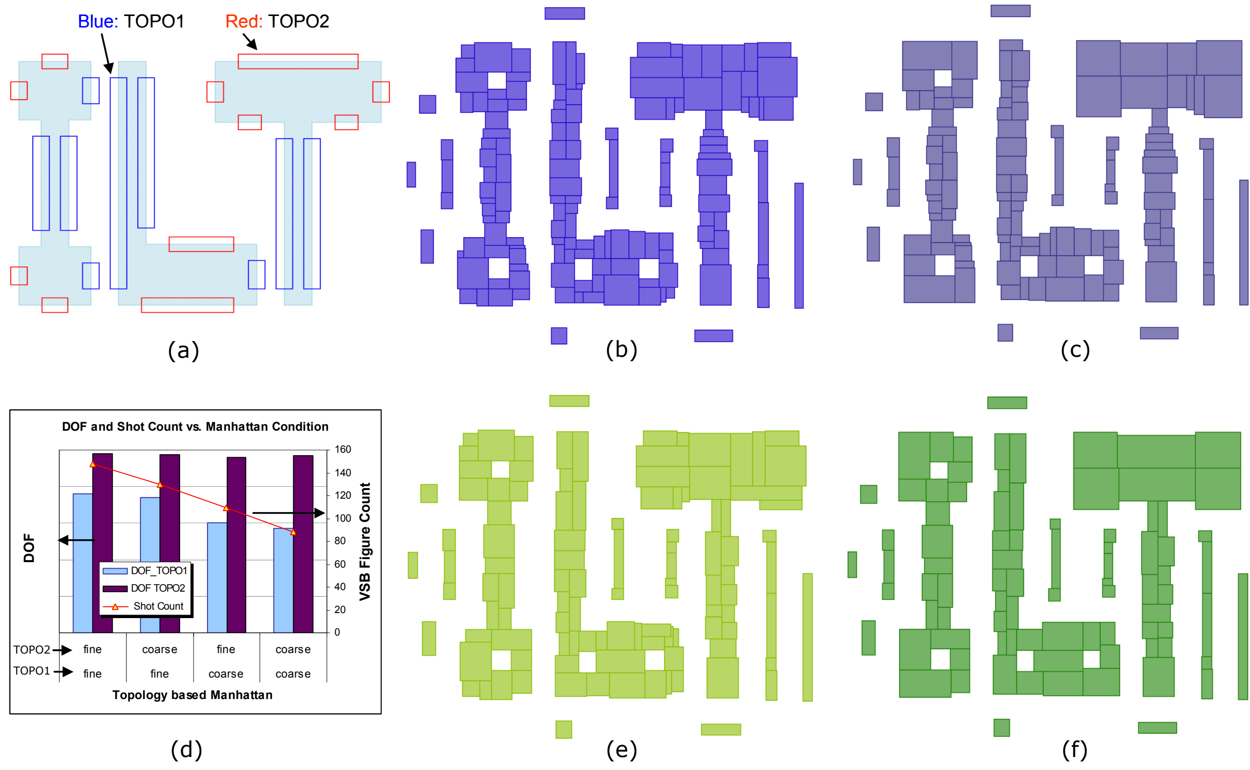

Figure 5. Illustration of topo-aware Manhattan conversion with an "ILT" test pattern. (a) Test pattern and custom topology markers: TOPO1 and TOPO2. (b), (c), (e) and (f) are fractured VSB figures of masks with different degree of coarseness of TOPO1 and TOPO2 during topology based Manhattan conversion: (b) TOPO1=fine, TOPO2=fine; (c) TOPO1 = fine, TOPO2 = coarse; (e) TOPO1 = coarse, TOPO2 = fine; (f) TOPO1 = coarse, TOPO2 =coarse. (d) DOF of TOPO1 and TOPO2 and VSB figure count of "ILT" pattern from each Manhattan conversion condition in (b), (c), (e) and (f).

As discussed earlier the idea of topology based Manhattan segmentation is to get maximum mask simplification without losing required process window. Figure 5(d) plots depth of focus (DOF) of regions covered by TOPO1 and TOPO2 with respect to different Manhattan conditions (b), (c), (e) and (f). In this example, TOPO1 covers regions with smaller line and tighter space while TOPO2 covers larger pattern and open space. The process window (PW) tolerance for TOPO1 is more stringent than for TOPO2. With fine Manhattan for both TOPO1 and TOPO2 (figure 5(b)), both regions have adequate process window. If coarse Manhattan being applied to all features (figure 5(f)), then it can cause too much process window loss on critical regions marked by TOPO1. By keeping fine Manhattan for critical TOPO1 region while using coarse Manhattan for less critical TOPO2 region (figure 5(c)), then both TOPO1 and TOPO2 regions can maintain enough process margin while overall mask complexity and shot count are reduced as shown in figure 5(d).

3.3 Shot count reduction with jog alignment

Several options of reducing mask write time have been presented in previous sections. Both selective SRAF placement and variable coarseness for Manhattan conversion will directly affect the mask shape and e-beam shot count. Jog alignment as yet another option, however, will not significantly change either the mask shape or mask gds file size. But it reduces the number of e-beam shots by aligning facing jogs across the pattern. Figure 6 shows an example of Manhattan mask with and without jog alignment option. Figure 6(a) is overlay of masks with and without jog alignment. The two masks are almost perfectly matched except minor variations especially around jogs. A close-up view (inset) of

small section of figure 6(a) points to the neighboring jogs with and without jog alignment. The impact of these sub-nanometer jog movements is mostly negligible to litho performance; however, the impact on shot count can be significant as shown in figure 6(b) and 6(c). It is clearly shown that more VSB figures are created from un-aligned facing jogs in figure 6(b) than mask with jog alignment in figure 6(c). Since jog alignment is a mild non-intrusive operation, it will have little impact on the process windows of final mask.

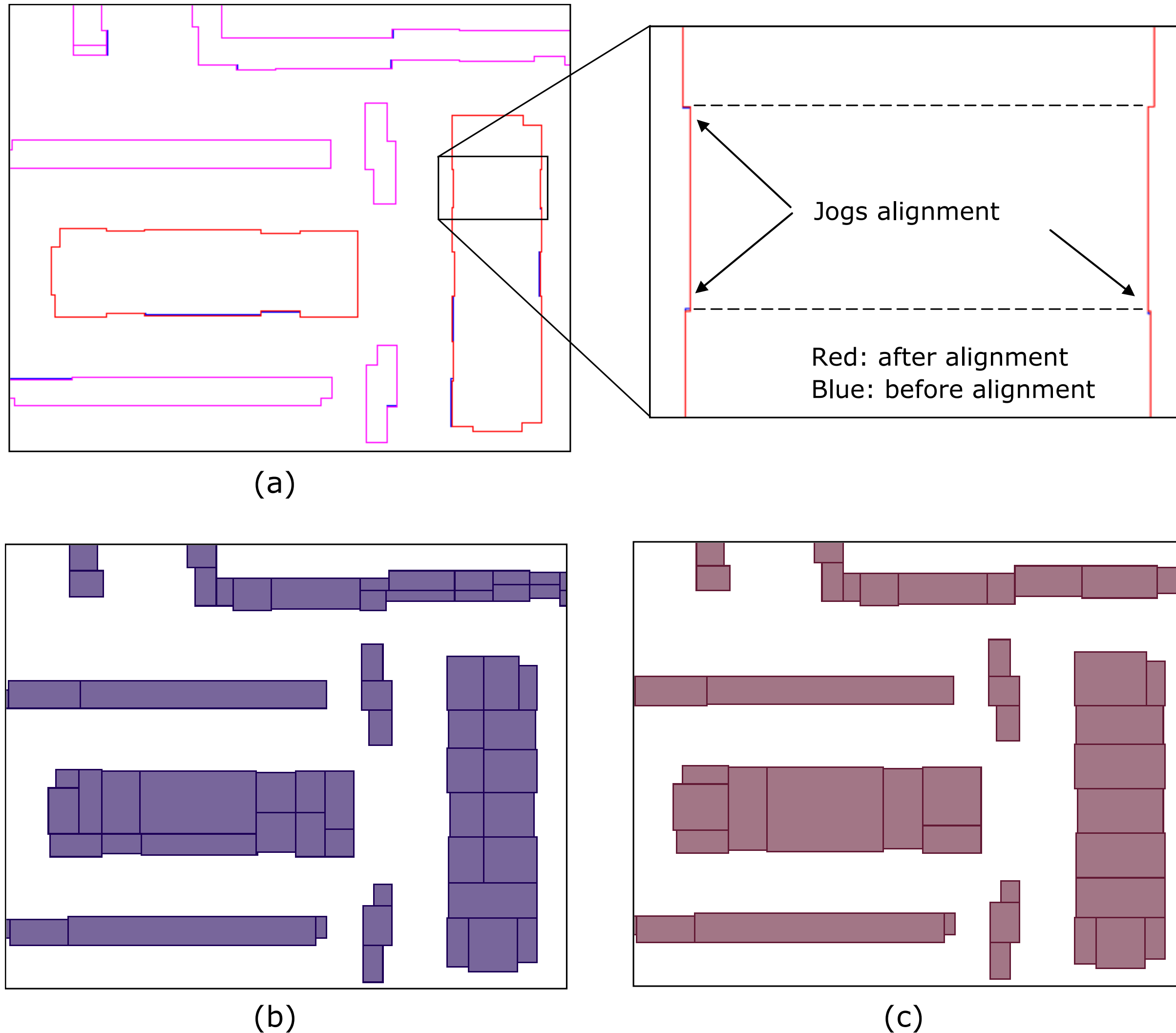

Figure 6. Shot reduction from jog alignment. (a) Overlay of ILT masks with and without jog alignment. A few locations marked by arrows display the effect of jog alignment option. (b) Fractured VSB figures from mask without jog alignment. (c) Fractured VSB figures from mask with jog alignment.

To evaluate jog alignment impact on total shot count at full chip level, several design layers are used to run ILT with and without jog alignment option in Manhattan conversion. Figure 7(a) is shot count result from five different designs. For easier comparison, total shot count of each design is normalized to the shot count of mask without jog alignment. In all cases, jog alignment reduces total shot count and the reduction is by amount over 10% on average. In addition to shot count reduction, another benefit from jog alignment is sliver shots reduction. As shown in figure 6(b), smaller shots are generated from unaligned facing jogs. Even though the fracture tool always tries to maximize each shot size, slivers creation is unavoidable in difficult scenarios. Figure 7(b) shows shot size histogram of Gate layer masks with and without jog alignment; note that only the tail of the distribution with shot size smaller than 200nm is shown to focus on slivers. For shots with size smaller than 50nm (at mask scale), jog alignment reduces the amount by more than half

comparing to no jog alignment. Since slivers in general are bad for e-beam mask writing and can potentially create pattern defects on mask, the reduction on slivers from jog alignment brings in additional benefit to mask manufacturing.

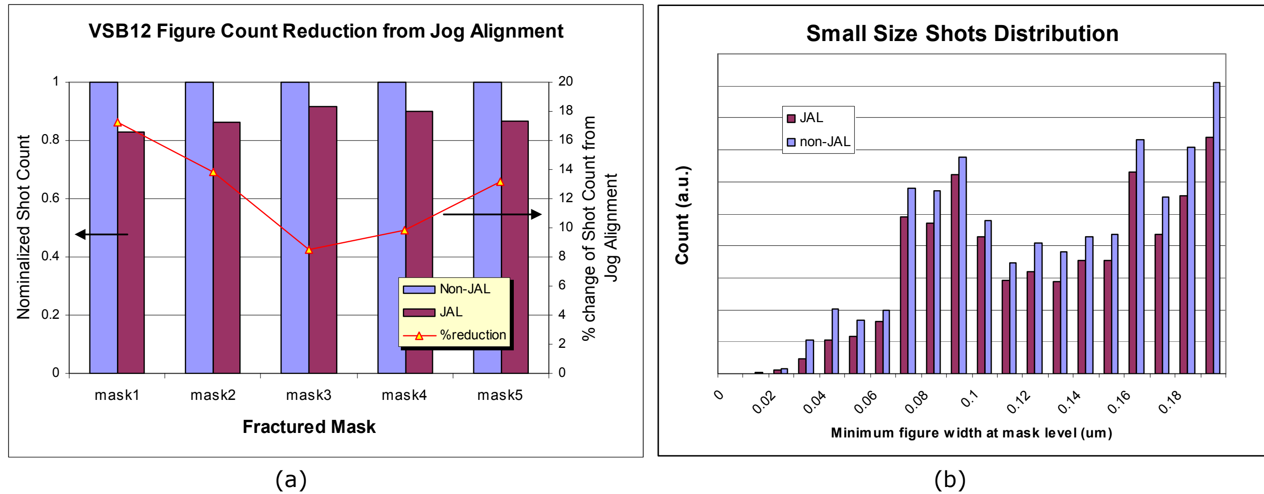

Figure 7. Shot and sliver count with and without jog alignment (JAL). (a) Fractured VSB figure count comparison w/o jog alignment of several masks. (b) Small figure distribution w/o jog alignment.

3.4 Litho performance versus Manhattan options

For the several Manhattan options presented in previous sections, each demonstrated significant reduction of mask total shot count hence write time. Ideally, any work in the efforts to reduce shot count should not compromise litho performance. The effect on image EPE (edge placement error) and PV band width (process variation band) requires detailed analysis at full chip level. Figure 8 shows the results on shot count reduction from new Manhattan with jog alignment and its impact on litho performance (EPE and PV Band) at full chip. With several line/space layer designs, new Manhattan exhibits 3X reduction in total shot count comparing with old Manhattan conversion as shown in figure 8(a). For litho performance, four critical topologies in target design of mask#1 were selected to evaluate EPE and PV band. Figure 8(b) plots the mean and sigma values of PV band width at regions covered by four critical topologies. The process corners used for PV band generation are nominal, +/-60nm defocus and +/-5% dose offset. For each topology, results from three masks are presented and compared: mask with old Manhattan as reference, mask with new Manhattan but without jog alignment and mask with new Manhattan plus jog alignment. The mean and sigma values with and without jog alignment are essentially identical. This is not surprising since jog alignment is a very mild non-intrusive process as discussed in section 3.3. In comparison with old Manhattan reference values, new Manhattan exhibits slightly worse PV band mean but almost the same sigma values for all four critical topologies.

Figure 8(c) and 8(d) are EPE charts of nominal and +60nm defocus images. Similar to PV band, jog alignment has little effect on mean and sigma values of EPE at either nominal or defocus conditions. New Manhattan although shows noticeably larger residual EPE at nominal than old Manhattan, at defocus condition the delta between new and old Manhattan becomes much less. This trend matches what is seen with PV band. At nominal condition even though new Manhattan gave higher EPE than old Manhattan, the EPE values are still very small and met customer specifications on nominal EPE tolerance. The calculated sacrifice of nominal EPE from new Manhattan brings benefit of almost 3X reduction in total shot count on mask#1 as shown in figure 8(a). Based on the results published in previous paper [1], the amount of shot count reduction from new Manhattan will enable ILT poly layer mask write time well within acceptable range for 30nm node and below.

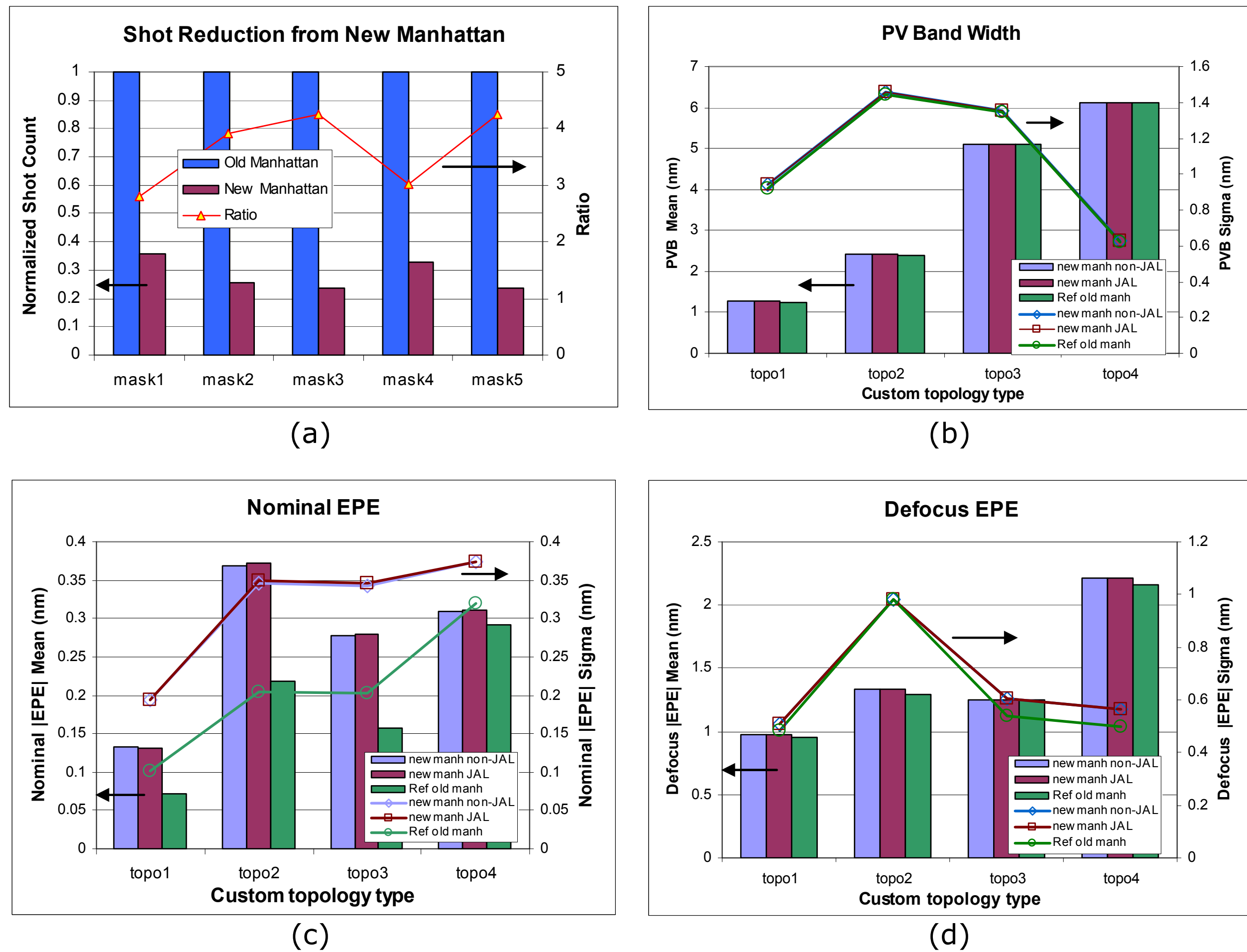

Figure 8. Shot count reduction from new Manhattan and its impacts on EPE and PV Band. (a) Total shot count comparison of masks from old and new Manhattan conversion on several designs. (b) Full chip PV Band statistics on four critical topologies of a Gate layer. (c) Full chip nominal EPE statistics. (d) Full chip off-nominal EPE (60nm defocus) statistics.

4. DATA FRACTURE OPTIMIZATION FOR MASK E-BEAM WRITE

Once optimized mask data is sent to mask shop, they often can not be directly used for mask writing due to non-zero process bias at mask shop. In other words, incoming OPC gds will have to go through sizing step before sending to mask writer in order to accommodate the process bias at mask shop. For simple rectangular type of OPC and SRAF, this sizing step will not introduce additional VSB figures during fracturing. However, for more complicated mask shapes from model based SRAF and OPC at 32nm node and below, data sizing can introduce additional VSB figures at regions with angled structures. An example of this effect can be shown in figure 9(d) at the regions where arrows point to. For ILT mask where jog alignment feature have already aligned neighboring facing jogs, data sizing will cause pre-aligned jogs becoming misaligned. This can substantially increase the number of e-beam shots. To address this issue, NCS mask data preparation system, NDE-MDP, provides various features to optimize total shot count and mask write time. One of the features NCS delivered is called “slide” option. It keeps jogs aligned during sizing to prevent fracture from generating new edges. Consequently, additional shots will not be introduced from data sizing at mask shop.

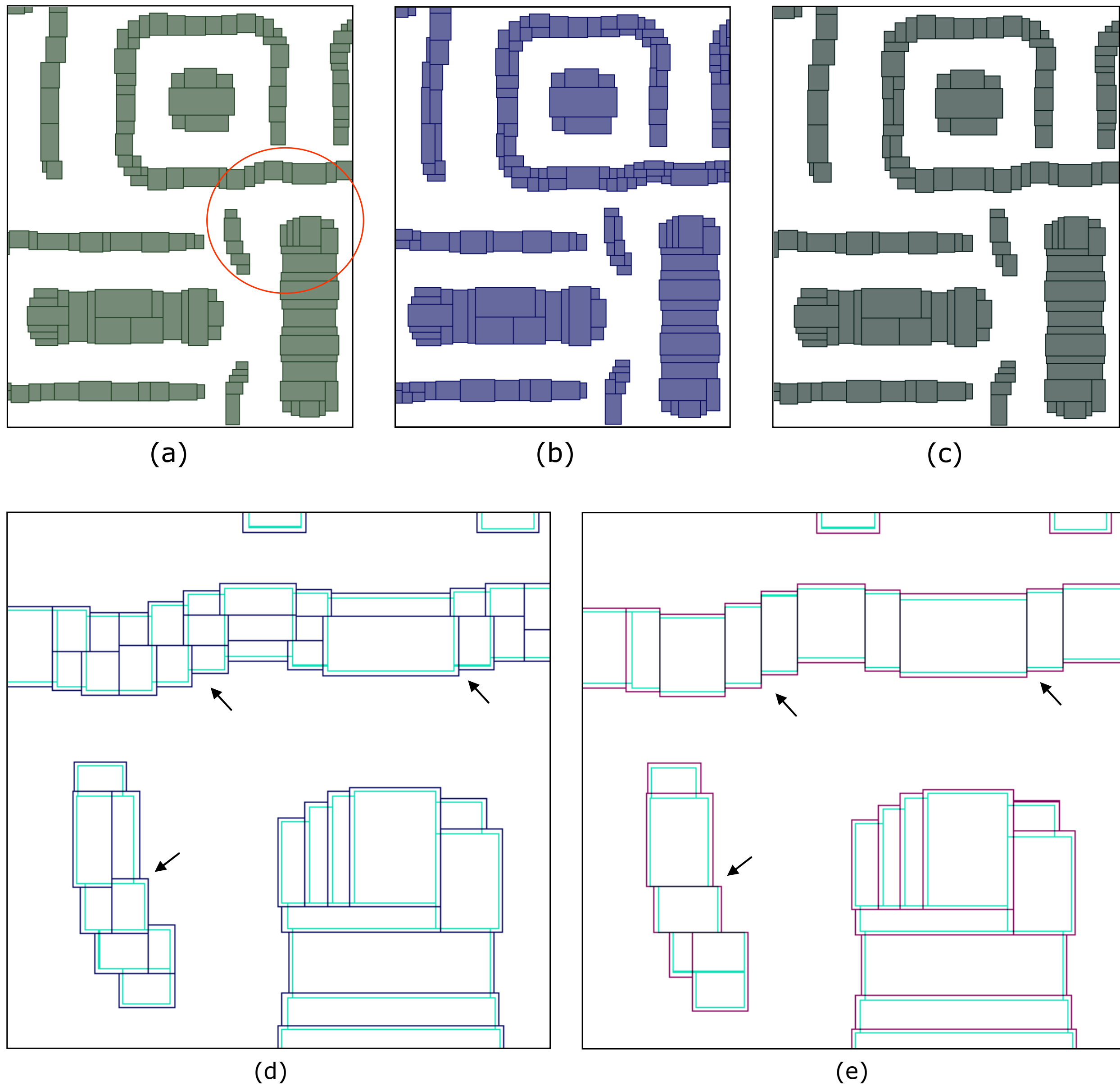

Figure 9. Fractured VSB figure count before and after data sizing and NCS slide option. (a) Fractured VSB figures of original gds without sizing. (b) Fractured VSB figures of sized-up gds. (c) Fractured VSB figures of sized-up gds with slide option. (d) Overlay of fractured VSB data between original (light color) and sized gds without slide option (dark color). (e) Overlay of fractured VSB data between original (light color) and sized gds with slide option (dark color).

The effect of slide option in NDE-MDP fracture engine is illustrated in Figure 9. Figure 9(a) is fractured data of original gds with sizing as a reference. After gds sizing, fractured VSB figures double in many locations as seen in figure 9(b). However, if slide option is applied during sizing the number of VSB figures is relatively unchanged as shown in figure 9(c). To see the effect better, figure 9(d) and 9(e) present a zoom-in view of the region circled in figure 9(a) with overlay of sized and original data. Figure 9(d) is without slide option where sizing grows data uniformly in all directions and causes previously aligned jogs becoming mis-aligned. Additional VSB figures are therefore introduced to avoid thin sliver shot otherwise be formed from jog misalignment. With slide option, data grows asymmetrically to keep already aligned staircases aligned during sizing hence no additional figure is introduced as shown in figure 9(e).

Figure 9 illustrates how slide option works with an example of very small clip. To assess the impact on shot count from data sizing at full chip level, three masks are fractured with and without slide option. The complexity of SRAF from mask 1 to 3 varies from simpler rectangular to more curvy shape. Figure 10(a) shows normalized total shot count to that of original data without sizing. 1~5% increase in total shot count is shown after data sizing without using slide option during fracturing. Since additional figures are introduced only at angled staircases, mask #3 with more complex SRAF exhibits more pronounced increase of shot count than others. If slide option is applied during fracture with sizing, the total shot count drops.

The more aggressively slide option applied, the more shot count decreases. In addition to total shot count, another factor to look at is sliver count. As discussed earlier, slivers can be introduced by jog misalignment from data sizing. Figure 10(b) shows sliver count of figure size less than 20nm at mask scale. The number of slivers is more than doubled for complex mask #3 after data sizing without slide option. With slide option, however, the number of slivers under 20nm is sharply reduced to the level not much more than reference value of original data without sizing.

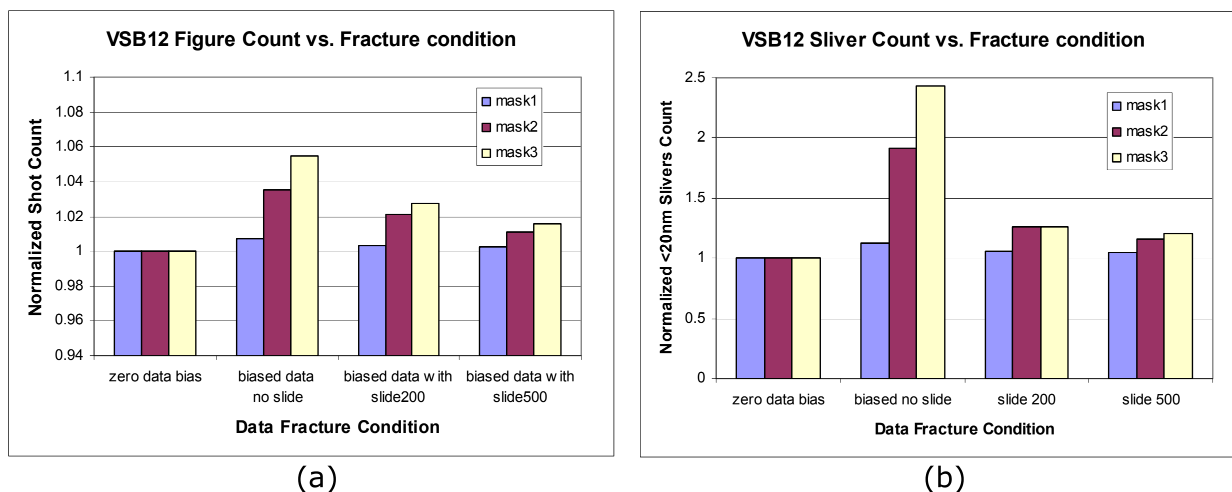

Figure 10. Total VSB figure count and sliver count versus different fracture conditions. (a) Normalized total shot counts of three masks with different fracture conditions. (b) Count of slivers with size less than 20nm.

5. SUMMARY AND CONCLUSIONS

In summary, several options of reducing mask complexity have been presented in this paper to improve ILT mask manufacturability for full-chip applications. Up to 3X reduction in shot counts of poly layers has been achieved with selective SRAF placement, jog alignment and improved ILT Manhattanization. Local CD aware selective SRAF can substantially reduce the mask complexity, especially for line and space layers like poly gate. Custom topology markers based Manhattan conversion provides maximum user control with localized segmentation for optimal Manhattanization of ILT mask. NCS slide option fracturing technique can maintain overall shot count of ILT mask and reduce slivers during data sizing in mask shop. Using the improved Manhattan solutions presented in this paper produces ILT masks with write times similar to conventional OPC, while maintaining superior litho performance, translating directly into a greatly reduced cost of ownership

REFERENCES

[1] Kim B.G., Suh, S.S., Woo S.G., Cho H.K., Xiao G., Son D., Irby D., Kim D., Baik K., "Inverse lithography (ILT) mask manufacturability for full-chip Device", Photomask Technology 2009, Proc. of SPIE Vol. 7488 (2009)

[2] Kim B.G., Suh S.S., Kim B.S., Tolani V., Dai G., Irby D., Wang K., Xiao G., Kim D., Baik K., Gleason G., "Trade-off between Inverse Lithography Mask Complexity and Lithographic Performance", Proc. of SPIE Vol. 7379 (2009)

[3] Abrams, S. and Pang, L., “Fast inverse lithography technology”, 31st International Symposium of Microlithography, Proc. SPIE 6154 (2006).

[4] Pang, L., et al, “Inverse Lithography Technology (ILT): Keep the balance between SRAF and MRC at 45 and 32nm”, Proc. SPIE Vol. 6730, (2007)

[5] Pang, L., Xiao, G., Tolani, V., Hu, P., Cecil, T., Dam, T., Baik, K. and Gleason, B., “Considering MEEF in inverse lithography technology (ILT) and source mask optimization (SMO)”, Photomask Technology 2008, Proc. of SPIE Vol. 7122 (2008)

[6] Farys V. et al, “SRAF Enhancement using Inverse Lithography for 32 nm Hole Patterning and Beyond”, Proc. of SPIE Vol. 7488 (2009)

[7] Zhang Y. et al, “Mask cost analysis via write time estimation”, Proc. of SPIE Vol. 5756, 313 (2005)

[8] Xiao, G., et al, “E-beam writing time improvement for Inverse Lithography Technology mask for full-chip”, Photomask and NGL Mask Technology XVII, 2010, Proc. of SPIE Vol. 7748 (2010)