Cost-Driven Mask Strategies Considering Parametric Yield, Defectivity and Production Volume

Kwangok Jeong[1], Andrew B. Kahng[1,2] and Christopher J. Progler[3]
[1] ECE Department, University of California at San Diego
[2] CSE Department, University of California at San Diego
[3] Photronics, Inc.
kjeong@vlsicad.ucsd.edu, abk@ucsd.edu, cprogler@photronics.com

ABSTRACT

In this paper, we provide new yield-aware mask strategies to mitigate emerging variability and defectivity challenges. To address *variability*, we analyze critical dimension (CD) variability with respect to reticle size, and its impact on parametric yield. With a cost model that incorporates mask, wafer, and processing cost considering throughput, yield, and manufacturing volume, we assess various reticle strategies (e.g., single layer reticle (SLR), multiple layer reticle (MLR), and small and large size) considering field-size dependent parametric yield. To address *defectivity*, we compare parametric yield due to EUV mask blank defects for various reticle strategies in conjunction with reticle floorplan optimizations such as shifting of the mask pattern within a mask blank to avoid defects being superposed by performance-critical patterns of a design.

1. INTRODUCTION

Photomask cost is a highly critical issue in manufacturing. Semiconductor manufacturers have long sought cost-effective photomask strategies. Multiple copies of a single layer of one project (IC) are

---

Further author information: (Send correspondence to Kwangok Jeong)

Kwangok Jeong: E-mail: kjeong@vlsicad.ucsd.edu, Telephone: 1 858 822 5003

patterned in a full-size mask blank to obtain a *single-layer reticle* (SLR), used in most high-volume products. In a *multi-project reticle*, the same layer of several different projects (ICs) is implemented on a single reticle; this allows sharing of mask costs between individual project owners. Beyond a "single-layer-per-reticle" strategy, multi-layer and multi-product strategies is also implemented on a single reticle,[4] and an algorithm to enable the layer placement and quality check procedure according to a parameterized cost function is proposed.[5] In addition, reticle size and number of dies per reticle are other knobs that can be tweaked by manufacturers or designers.

IC manufacturing traditionally uses the maximum possible reticle size. This is commonly believed to maximize lithography tool throughput and minimize manufacturing cost. However, as reticle size increases, the mask cost (write, inspection, defect disposition, repair, etc.) also increases. For high-volume products, mask cost can be disregarded, but for low-volume products – in light of shuttle-based prototyping, design revisions and respins, market competition, and other factors – mask cost has significantly impact on overall cost per die. Also, larger reticles can result in larger CD variation in silicon, leading to parametric yield loss that potentially increases manufacturing cost. Hence, a new cost model is required to comprehend reticle size-dependent cost changes.

Besides the issue of variability, defectivity (notably, mask blank defects in extreme ultraviolet lithography (EUVL)) looms as a critical issue for mask generation and product yield. EUVL uses reflective masks instead of the traditional optical transmission masks. EUVL mask blanks contain a stack of 40 to 50 Mo-Si alternating layers, to maximize reflection at 13.5nm wavelength. Each of these layers requires a discrete processing step, hence defects at each layer can accumulate.[6] Defects in multi-layer EUVL blanks are difficult to detect and they cannot currently be repaired. An EUVL buried mask defect is known to cause critical dimension (CD) change.[10] Such CD changes may not cause catastrophic defects in the IC product, but can cause parametric yield loss through timing failures. Since EUVL mask blanks are not anticipated to be completely defect-free, a new reticle floorplan method is required to deal with defective mask blanks. Burns et al.[6] propose mask pattern translation and rotation in a mask blank to avoid the placement of critical mask patterns on defects. Such freedom in reticle floorplanning also depends on reticle size.

In this paper, we compare the manufacturing cost and yield of various reticle strategies considering parametric yield changes from field size and mask defectivity. The remainder of this paper is organized as follows. Section 2 presents various reticle strategies to be discussed throughout this paper, and Section 3 discusses a manufacturing cost model considering mask and lithography costs, and yield. Section 4 analyzes defect-aware parametric yield for EUVL, again with various reticle strategies. Finally, Section 5 gives our conclusions.

2. RETICLE STRATEGIES

In this section, we describe various reticle strategy assumptions that underlying our cost comparison. A reticle contains one or more dies, and all dies in a reticle are printed at the same time. We study the following strategies.

- *Single-layer reticle on large field (SLR-L):* traditional mask strategy. A reticle contains one processing layer for many copies of a die as shown in Figure 1(a)
- *Single-layer reticle on small field (SLR-S):* a reticle contains one processing layer for one or small number of dies as shown in Figure 1(b). Lithography throughput may be reduced, but mask cost can be reduced.
- *Multi-layer reticle on large field (MLR):* a reticle contains multiple layers (e.g., M1, M2, etc.) of a design as shown in Figure 1(c). When printing one layer, the other region (i.e., other layers) of the reticle is blocked. The number of reticles for a design can be reduced.

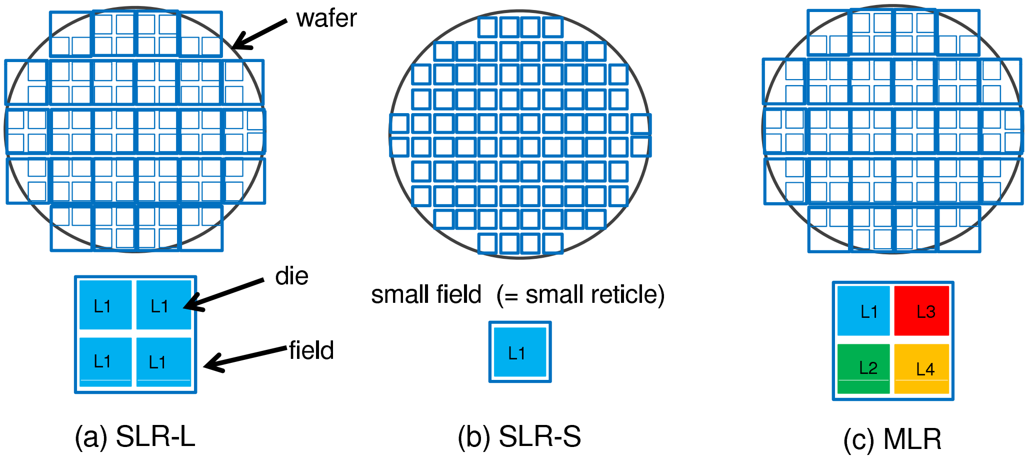


Figure 1. Examples of mask strategies: (a) single-layer reticle on a traditional large field, (b) single-layer reticle on a small field, and (c) multi-layer reticle on a large field.

3. COST MODEL

SEMATECH has for many years provided guidance on mask costs and their potential effects on product cost. The 1996 SEMATECH included actual manufacturing process steps. A 2000 revision added mask processing time to the cost model. In 2001, the mask cost of ownership (CoO) model was revised to reflect technology acceleration, i.e., a 2-year cycle of technology improvement, instead of the previously assumed 3-year cycle. Mask set cost is obtained as the sum of costs for all masks in the set; mask set costs are rising due to the increase in individual mask cost as well as an increase in the total number of masks in a mask set. Trybula[3] reviewed the methodology developed at SEMATECH to ensure that projected mask costs reflect the geometries being planned. Grenon[2] observed that the largest mask cost improvement came from higher defect repair yields, and proposed a new mask cost projections considering new mask repair technologies; improvements in focused ion beam (FIB), nano-machining and femtosecond laser repair. Pramanik et al.[11] analyzed cost of reticle strategies based on the SEMATECH cost of ownership model. Although Pramanik et al. propose the cost of mask and lithography with respect to the field size, they mainly focus on the mask generation cost including mask yield and stepper cost, and do not consider the parametric yield variation of silicon dies. Our present work extends that of Pramanik et al. by integrating the impact of field size on CD variation in silicon wafer observed from recent 65*nm* and 45*nm* foundry data, and then reevaluating the manufacturing cost of various reticle strategies with the scaled 45*nm* mask and lithography cost from those of 90*nm* technology.

3.1 Mask Cost Model

90*nm* mask cost model.[11] Each reticle strategy is differentiated by the number of dies per field. To represent mask cost considering the number of dies per field, we use the following parameters.

- w_f: field width on wafer in *mm*
- h_f: field height on wafer in *mm*
- M: mask reduction factor (in general, 4)
- n_{row}: number of rows of dies per field
- n_{col}: number of columns of dies per field

- $n_{m,vc}$: number of masks for very critical layers (e.g., 193nm)
- $n_{m,c}$: number of masks for critical layers (e.g., 248nm)
- $n_{m,nc}$: number of masks for noncritical layers (e.g., I-line)
- n_m ($= n_{m,vc} + n_{m,c} + n_{m,nc}$): total number of masks

The key contributors to mask costs are time-dependent cost (i.e., mask writing/inspection time) and yield-dependent cost. Mask writing/inspection time is proportional to the mask area and the mask resolution. Mask area is calculated based on how many dies are in a mask. To reflect cost difference due to the mask resolution, scaling factors are used. The writing/inspection times of very critical and critical layers are respectively assumed to be 4× and 2× larger than that of noncritical layers. The combined time-dependent cost is calculated as

$$c_{time} = r_{res} \cdot T_{min} \cdot A$$

where r_{res} is the cost scaling factor for mask resolution, T_{min} is the writing/inspection time for noncritical layers normalized to a unit area, and A is the mask field area calculated as $w_f \cdot h_f$.

Mask yield is affected by critical dimension (CD) (Y_{cd}), image placement error (Y_{pl}), random defects (Y_{def}), and some other uncertainties (Y_{misc}). The overall yield of a mask layer is calculated as

$$Yield = Y_{cd} \cdot Y_{pl} \cdot Y_{def} \cdot Y_{misc} \tag{1}$$

The *baseline mask yields* (Y^*) with full-size reticle for 90nm technology are assumed as Y^*_{cd} = 90%, Y^*_{pl} = 90%, Y^*_{def} = 80%, and Y^*_{misc} = 90%, with these values obtained from the third-year production yield of a typical 180nm node technology. The cumulative baseline mask yield is 58% from Equation (1). From the baseline yield values, yields for various reticle sizes are calculated, considering corner protrusion impacts p from different reticle sizes and a yield correction factor b. Corners of a square mask suffer from resist film thickness non-uniformity, which causes CD and image placement errors. Corner protrusion is the extension of a square field beyond the circular "stable region" in a mask, and is proportional to the diagonal of the mask field (i.e., $\sqrt{w_f^2 + h_f^2}$). The yield correction factor b is based on the idea of "bucketing" of yield-loss sources. Pramanik et al.[11] assume that a third of mask CD yield loss is from field size-dependent random variation, and another third from the corner protrusion effect. Each component of mask yield is then calculated from the baseline yield as

$$Y_{cd} = (Y^*_{cd})^{\left(1+w_f/w^*_f+p/p^*\right)/b},\ Y_{def} = \left(Y^*_{def}\right)^{A/A^*},\ \text{and}\ Y_{pl} = \left(Y^*_{pl}\right)^{p/p^*}$$

where A^*, w^*_f, and p^* are the area, mask field width and corner protrusion of the $100\times100mm^2$ reference mask, respectively.

From time-dependent cost and yield-dependent cost, overall mask cost is calculated. Let the calculated cost of very critical, critical and noncritical layers in a mask set be $c_{m,vc}$, $c_{m,c}$, and $c_{m,nc}$, respectively, and the number of masks for corresponding mask layers be $n_{m,vc}$, $n_{m,c}$, and $n_{m,nc}$, respectively. The total mask set cost $Cost_{maskset}$ is calculated as

$$Cost_{maskset} = c_{m,vc}\cdot n_{m,vc} + c_{m,c}\cdot n_{m,c} + c_{m,nc}\cdot n_{m,nc}.$$

Table 1 summarizes 90*nm* mask cost with respect to the field size shown in Table 4 of Pramanik et al.[11] The numbers of very critical, critical and non-critical layers for 90*nm* were assumed as 8, 8 and 12, respectively.

Table 1. 90*nm* mask cost from Pramanik et al.'s work.[11]

Mask field size (*mm*×*mm*)	100×100	64×96	64×64	32×64	32×32
Wafer field size (*mm*×*mm*)	25×25	16×24	16×16	8×16	8×8
Mask die size (*mm*×*mm*)	32×32	32×32	32×32	32×32	32×32
Wafer die size (*mm*×*mm*)	8×8	8×8	8×8	8×8	8×8
Number of dies in a field	9	6	4	2	1
Mask cost per layer (very critical) ($)	112,000	59,000	41,000	24,000	19,000
Mask cost per layer (critical) ($)	28,000	20,000	15,000	11,000	9,000
Mask cost per layer (non-critical) ($)	10,000	8,000	7,000	6,000	6,000
Mask set cost (very critical) ($)	896,000	472,000	328,000	192,000	152,000
Mask set cost (critical) ($)	224,000	160,000	120,000	88,000	72,000
Mask set cost (non-critical) ($)	120,000	96,000	84,000	72,000	72,000
Overall mask set cost ($)	1,240,000	728,000	532,000	352,000	296,000

Scaled 45*nm* mask cost. We estimate mask set cost for 45*nm* technology from the 90*nm* cost model, based on the following assumptions.

- Mask cost doubles at the introduction year of every technology node.

- Mask cost decreases by 20% per year.
- The introduction years of 90*nm*, 65*nm* and 45*nm* are 2003, 2005, and 2007, respectively.
- The number of mask layers for 45*nm* is 33 predicted in ITRS 2007.[1]
- The portion of very critical, critical and non-critical layers is equal (i.e., 11 layers for each).

These assumptions give 45*nm* mask cost as $4 \times (0.8)^{(2011-2003)}$ of 90*nm* initial mask cost; the factor 4 is from the two technology generations, and the mask cost is continuously reduced by 20% since the 90*nm* technology introduction year 2003. Table 2 shows the calculated mask set cost for 45*nm*. We observe that this cost is similar to the 90*nm* mask set cost; this matches our rough observations of mask cost trends across several recent technology nodes.

Table 2. 45*nm* mask cost scaled from 90*nm* mask cost.

Mask field size ($mm \times mm$)	100×100	64×96	64×64	32×64	32×32
Mask die size ($mm \times mm$)	32×32	32×32	32×32	32×32	32×32
Number of dies in a field	9	6	4	2	1
Mask cost per layer (very critical) ($)	75,162	39,594	27,515	16,106	12,751
Mask cost per layer (critical) ($)	18,790	13,422	10,066	7,382	6,040
Mask cost per layer (non-critical) ($)	6,711	5,369	4,698	4,027	4,027
Mask set cost (very critical) ($)	826,781	435,537	302,661	177,167	140,258
Mask set cost (critical) ($)	206,695	147,640	110,730	81,202	66,438
Mask set cost (non-critical) ($)	73,820	59,056	51,674	44,292	44,292
Overall mask set cost ($)	1,107,296	642,232	465,064	302,661	250,987

3.2 Litho Cost Model

Total manufacturing cost depends on the throughput. A smaller field is expected to cause lower throughput, since it requires a greater number of exposures. We calculate the lithography cost as a function of mask field size.

Parameters that affect lithography cost are number of exposures per wafer n_e, cost of a single exposure c_e, and number of mask layers n_m. The number of exposures is inversely proportional to the mask field size, and is calculated as the total number of dies per wafer divided by the number of dies per field. Then, multiplied by the number of wafers developed n_w, the total lithography cost $Cost_{litho}$ is calculated as

$$Cost_{litho} = c_{e,vc} \cdot n_{e,vc} \cdot n_{m,vc} + c_{e,c} \cdot n_{e,c} \cdot n_{m,c} + c_{e,nc} \cdot n_{e,nc} \cdot n_{m,nc}$$

where subscripts *vc*, *c*, and *nc* denote very critical, critical and noncritical layers, respectively.

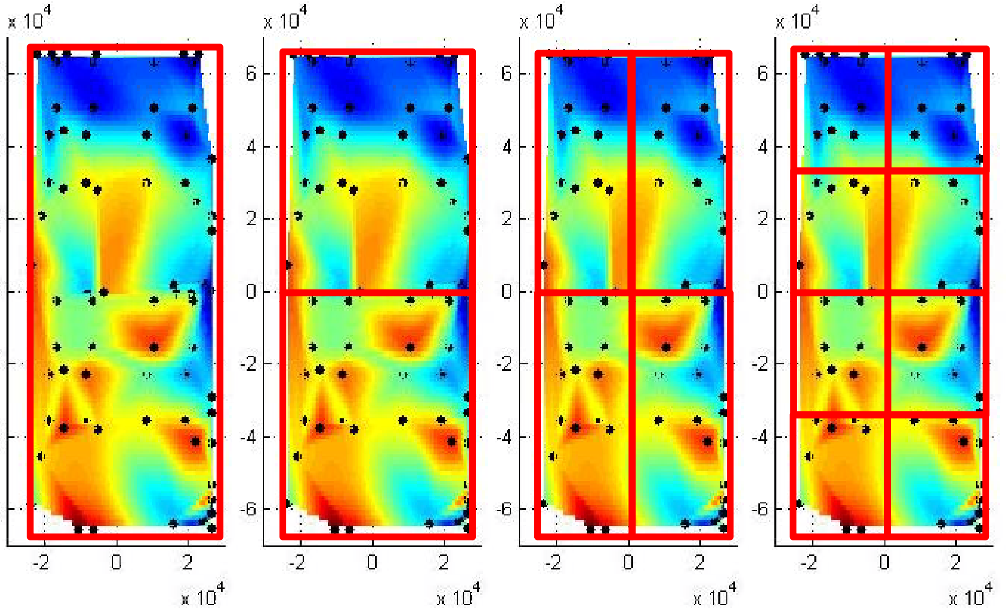
For 45*nm* lithography cost, we study three scenarios.

- *Scenario 1: constant lithography cost.* Cost of an exposure for very critical ($c_{e,vc}$), critical ($c_{e,c}$) and non-critical ($c_{e,nc}$) layers is assumed as \$2.5, \$1.5 and \$0.5, respectively, based on 90*nm* lithography cost estimation.[11]
- *Scenario 2: scaling by the lithography tool cost ratio.* Lithography tool cost is assumed as \$40M, \$49M, and \$52M, for 45*nm*, 32*nm* and 22*nm* technologies. From curve-fitting of lithography tool cost with respect to technology generation, the 90*nm* single exposure tool cost is estimated as \$29M. Then, scaling 90*nm* exposure cost by 1.38 (= \$40M / \$29M) gives \$3.45, \$2.07 and \$0.69 as the 45*nm* exposure cost for critical, critical and non-critical layers, respectively.
- *Scenario 3: doubling at every technology generation.* We also study a pessimistic lithography cost scenario to see the impact of high lithography cost on mask strategy. 45*nm* exposure cost for very critical, critical and non-critical layers is assumed as \$13.79, \$8.28, and \$2.76, respectively.

3.3 Parametric Yield Cost

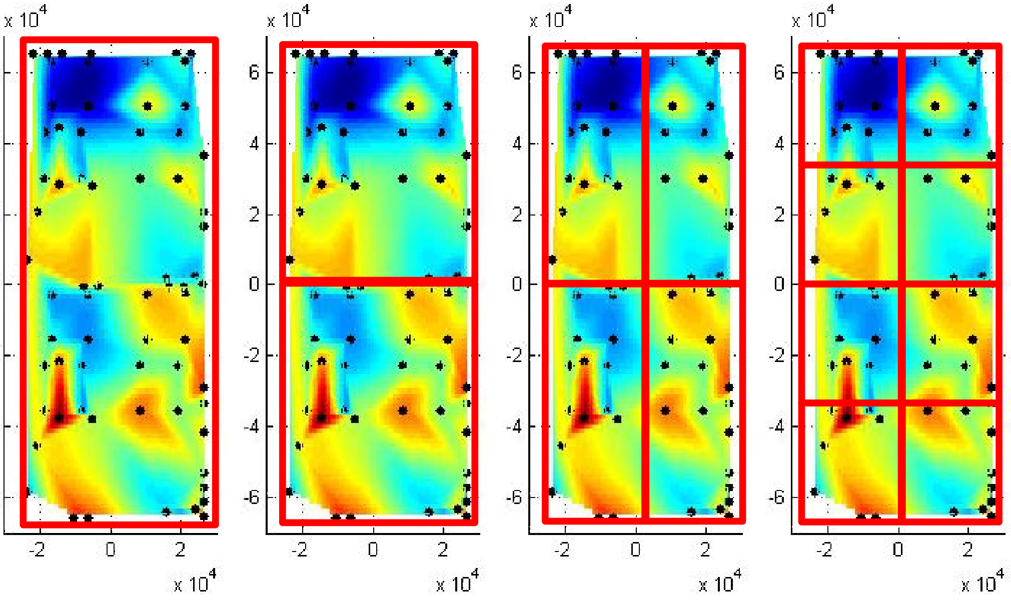
In addition to the mask yield, mask size affects the parametric yield of the manufactured dies. We have analyzed how CD variation changes with respect to mask size. Figure 2(a) (respectively, Figure 3(a)) shows mask CD variability of an example very critical layer from a 90*nm* (respectively, 65*nm*) industry product. The original mask size is approximately 52*mm*×132*mm* for both masks.

From the given CD measurement data, we analyze CD variations while decreasing field size, as shown in parts (b), (c), and (d) of Figures 2 and 3. Figures 4 and 5 show resulting 3σ CD variations with respect to field size, for the 90*nm* and 65*nm* masks, respectively. As we decrease field size from 52*mm*×132*mm* to 26*mm*×33*mm*, the average of 3σ CD variations of the small subfields is continuously reduced from 2.04*nm* (resp. 2.21*nm*) to 1.37*nm* (resp. 1.77*nm*) for the 90*nm* (resp. 65*nm*) mask. Furthermore, if we are allowed to choose the minimum CD variation field out of all subfields, 3σ CD variation can be reduced to 1.10*nm* (resp. 1.08*nm*) for 90*nm* (resp. 65*nm*) mask, as shown in the red-dotted traces in Figures 4 and 5.



(a) 52mm x 132mm (b) 52mm x 66mm (c) 26mm x 66mm (d) 26mm x 33mm

Figure 2. Mask CD variation map for a 90*nm* product.



(a) 52mm x 132mm (b) 52mm x 66mm (c) 26mm x 66mm (d) 26mm x 33mm

Figure 3. Mask CD variation map for a 65*nm* product.

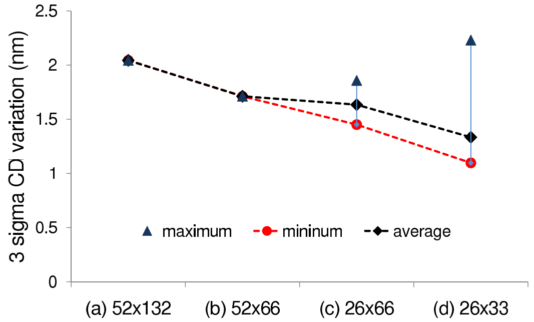


Figure 4. 3σ CD variation (*nm*) versus field size (*mm*^{2}) for the 90*nm* mask CD map in Figure 2.

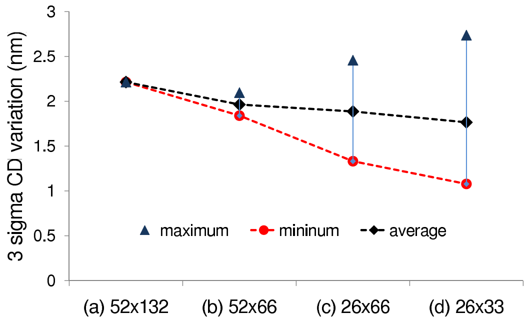


Figure 5. 3σ CD variation (*nm*) versus field size (*mm*^{2}) for the 65*nm* mask CD map in Figure 3.

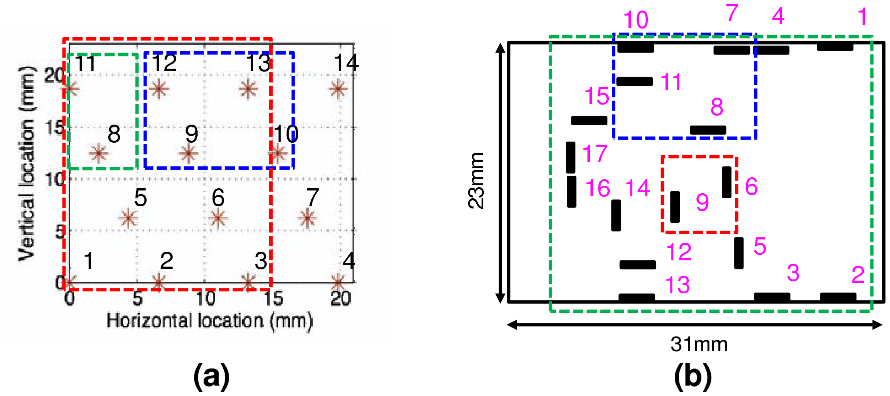


Figure 6. Measurement point locations in fields from Foundry A data in (a) and Foundry B data in (b).

Reduced variation in mask CD from a small-field strategy would contribute to reduced variation in electrical characteristics on the manufactured wafer. We do not have access to a unified CDU data set for both mask and wafer of a single design. However we have been able to analyze variations of (on-wafer) electrical characteristics in two industry data sets with respect to the mask size. Our first data set consists of measured ring oscillator delay in a 65*nm* test chip from Foundry A. There are 14 measurement points regularly placed in a 20×20*mm*^{2} field as shown in Figure 6(a). The number of measured fields is 36,727. From this data, we calculate delay variation (σ/μ) changing the size of sampling window to account for the impact of small field, as shown by the dotted boxes in Figure 6(a). The size and location of a sampling window together determine the measurement points included. Table 3 summarizes σ/μ with respect to the sampling window height and width. Given the yield of the full-size field, we normalize the delay variation of different field sizes to that of the full-size field, and calculate corresponding parametric yields. For instance, the number of standard deviations resulting in 90% yield is 1.645. The delay variation of the 400*mm*^{2} full-size field of Foundry A is 2.995 as shown in Table 3. Then, 3.126 of delay variation from 266.66*mm*^{2} field is equivalent to 1.576 (=

1.645×(2.995/3.126)) of standard deviation, which gives 88.5% yield. Column 4 (resp. 5) of Table 3 shows the parametric yield assuming that the parametric yield of a full-size field is 90% (resp. 80%). The parametric yield improves as window size decreases.

Table 3. Delay variation and parametric yield with respect to field size in 65*nm* test chip from Foundry A.

Width (μm)	Height (μm)	Area (mm^2)	Delay variation σ/μ (%)	$Y_{p,90}$ (%)	$Y_{p,80}$ (%)
20000	20000	400.00	2.995	90.0	80.0
13333	20000	266.66	3.126	88.5	78.0
8888	6500	57.77	2.749	92.7	83.7
2222	6500	14.44	1.897	99.1	95.7

Our second data set consists of measured $I_{d,sat}$ in a 45*nm* test chip from Foundry B. There are 17 measurement points in a $23\times31mm^2$ field, as shown in Figure 6(b). We again calculate $I_{d,sat}$ variation while changing sampling window size. Table 4 summarizes σ/μ of $I_{d,sat}$ variation with respect to the sampling window height and width. We again assume that the yield of a full-size field is 90% (resp. 80%), then normalize delay variation of different field sizes to that of the full-size field, and calculate parametric yields.

Table 4. $I_{d,sat}$ variation and parametric yield with respect to field size in 45*nm* test chip from Foundry B.

Width (μm)	Height (μm)	Area (mm^2)	$I_{d,sat}$ variation σ/μ (%)	$Y_{p,90}$ (%)	$Y_{p,80}$ (%)
22941	20418	468.42	3.209	90.0	80.0
16088	12937	208.13	2.945	92.7	83.7
6881	9239	63.57	2.421	97.1	91.0
7548	8312	62.74	1.687	99.8	98.5
6222	4855	30.20	2.266	98.0	93.0
5999	4385	26.31	2.368	97.4	91.7
5617	3698	20.77	1.501	100.0	99.4
3640	2994	10.90	2.153	98.6	94.4
2360	4385	10.35	1.085	100.0	100.0
3172	75	0.24	1.042	100.0	100.0

Finally, Figure 7 shows the relationship between parametric yield and field area, for both data sets. From linear regression, we obtain a parametric yield model with respect to the normalized field area F_{area}. The obtained parametric yield model is reflected in the final cost model as a denominator to the lithography cost, assuming that more wafers will be processed as parametric yield decreases. The linear parametric yield model is

$$Y_p(F_{area}) = (1 - \alpha F_{area})$$

where α is 0.1296 (resp. 0.2657) when the yield of full-size mask is assumed as 90% (resp. 80%).

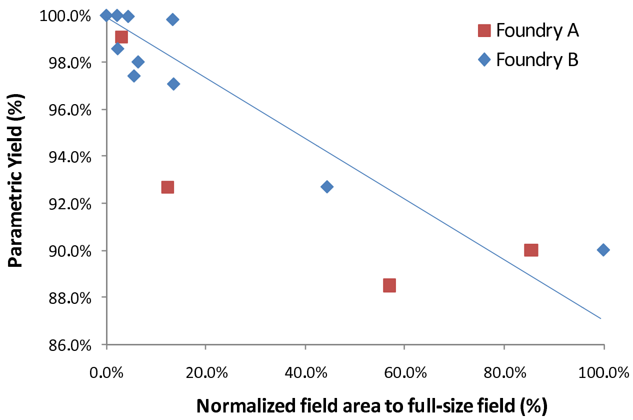


Figure 7. Field size normalized to a full-size field in x-axis, versus parametric yield assuming 90% for full-size field in y-axis.

3.4 Overall Manufacturing Cost Comparison

Finally, the overall manufacturing cost considering parametric yield is calculated as

$$Cost_{all} = n_{rebuilt} \cdot Cost_{maskset} + Cost_{litho}/Y_p$$

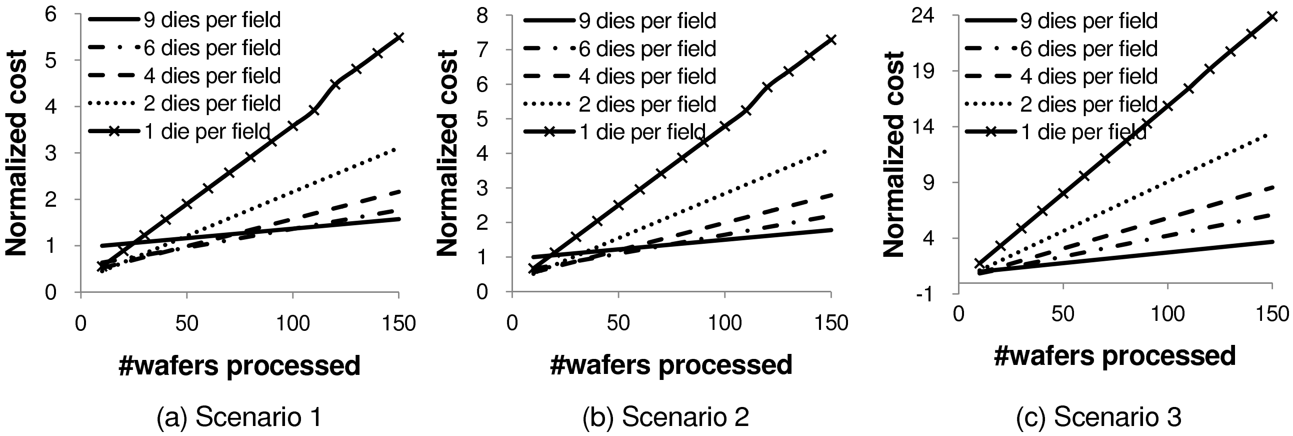
where $n_{rebuilt}$ is the number of mask sets rebuilt due to mask wearout.*

Figure 8 shows overall manufacturing cost with respect to varying number of dies per field, as the number of wafers processed is increased. This comparison assumes 90% parametric yield for a full-size field. All values are normalized to the cost of processing 10 wafers with a $100\times100mm^2$ field.

For Scenario 1 shown in Figure 8(a), we can observe that below 100 wafers, fewer dies per field can have lower cost than the full-size field (i.e., 9 dies per field) case: up to 20 wafers, 2 dies per field has best cost; between 20 and 40 wafers, 4 dies per field has best cost; and between 50 wafers and 100 wafers, 6 dies per field has best cost. The benefit of small-size field (i.e., SLR) is reduced as lithography cost increases. This is seen in Figure 8(b)-(c): for Scenario 2, the full-size field has best

---

*We assume that the mask set must be rebuilt at every 86,000 exposures, which is the number of exposures for 1,000 $300mm$ wafers with $25\times25mm^2$ full-size field. With a small field, $n_{rebuilt}$ increases due to the increase of the number of exposures per wafer. For instance, for the smallest mask size (= $32mm \times 32mm$ mask) in Column 6 in Table 1, the number of exposures per wafer is 774. When the number of wafers processed exceeds 111 (= 86,000 / 774), the mask set needs to be rebuilt.



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

Figure 8. Overall manufacturing cost in y-axis versus the number of wafers processed. Cost values are normalized to the cost of processing 10 wafers with a $100 \times 100 mm^2$ field.

cost when more than 70 wafers are processed; and for Scenario 3, the full-size field has best cost when more than 10 wafers are processed.

4. RETICLE STRATEGIES FOR EUVL

In this section, we compare parametric yield due to EUVL mask defects for various reticle strategies.

4.1 Defect-Aware Parametric Yield Calculation

To calculate defect-aware parametric yield, we first randomly distribute defects on a mask blank. At the same time, we extract timing-critical regions from a design using signoff timing analysis and placement information. We then check whether any defect in a mask blank overlaps with any timing-critical region. The overlapping of defects and timing-critical regions varies with respect to the reticle strategy and the location of the field on a mask blank. We estimate measure the yield from Monte Carlo simulation.

4.1.1 Assumptions

We consider the following assumptions and scenarios to calculate defect-aware parametric yield.

Defect density and distribution. Burns et al.[6] assume 10 to 55 defects per mask; Heuvel et al.[7] use a mask with 0.72 known defects/cm^2 in their experiments, and find around 200 defects from inspection. Early EUVL mask blanks contain thousands of defects. With steady improvements in blank generation, the detectable defect count, with first-generation mask-blank inspection tools limited to detecting 80*nm* defect size, is reduced to hundreds in 2007. However, the number of defects increases again by more than an order of magnitude when detectable defect size is reduced from 80*nm* to 50*nm* by advances in mask-blank inspection technology.[6] Among the detectable defects, defects that change feature size by more than 10% are regarded as critical defects in the ITRS.[1] Burns et al.[6] assume defect size of 146*nm* to 3,690*nm* in their defect-avoiding mask alignments, while the ITRS specifies that the critical defect size for EUVL masks is 41*nm* in 2009 and reduces to 16*nm* in 2024.[1]

Our experiments focus on the substrate defects which are the majority (e.g., 75% in Rastegar et al.[9]) of EUVL mask defects. These substrate defects are randomly placed in a typical 152*mm*×152*mm* mask blank. Our test design has 8*mm*×8*mm* area, and we assume that 16 (4×4) dies can be fit into the full-size reticle. The defect density in our experiments is summarized in Table 5. Up to 2.222 defects/cm^2 in a mask blank in Table 5 may be realistic, but we also examine much larger defect densities to account for future inspection technology improvements and/or early stages of technology introduction.

Table 5. Defect density in our experiments.

Field size (1×) (*cm* × *cm*)	#Defects per mask blank	Mask blank size (4×) (*cm* × *cm*)	Defect density (/cm^2) in a mask blank (4×)
3.2 × 3.2	10	15.0 × 15.0	0.044
3.2 × 3.2	50	15.0 × 15.0	0.222
3.2 × 3.2	100	15.0 × 15.0	0.444
3.2 × 3.2	500	15.0 × 15.0	2.222
3.2 × 3.2	1000	15.0 × 15.0	4.444
3.2 × 3.2	5000	15.0 × 15.0	22.222

For a given defect density, we distribute the defects in two ways.

- *Uniform random.* The number of defects per mask blank is calculated from the given defect density, and defect location coordinates are determined by uniformly random number generation between 0 and mask blank size in x and y respectively.

Table 6. Surface defect height, CD variation (ΔL), and resulting timing variation (ΔT) from a 45*nm* open-source design kit.

Height (*nm*)	ΔL (*nm*)	ΔT (ps)
1	1.03	2.00
2	3.06	5.87
4	7.11	13.41
8	15.22	28.27

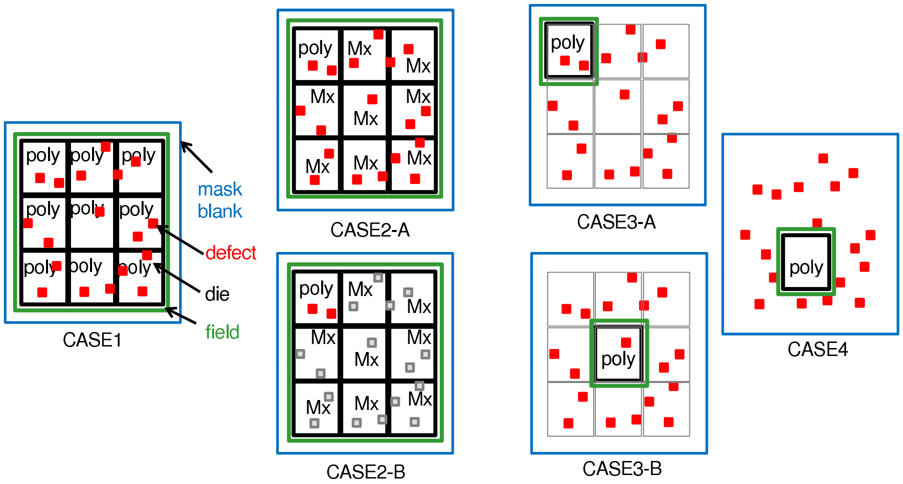
- *Decentered Gaussian.* The number of defects per mask blank is calculated from the given defect density, and defect locations are sampled from a decentered Gaussian distribution. The decentered Gaussian distribution is composed of two Gaussian distributions: for x-(y-)coordinates, one mean is located at the left (bottom) boundary of the mask blank, and the other mean at the right (top) boundary of the mask blank. We take one-sixth of the mask blank width (height) as the sigma of the Gaussian distribution.

Defect and impact on circuit timing. Clifford et al.[10] show that square defects at the substrate with widths varying from 60*nm* to 90*nm* all result in around 50-60*nm* width defects at the final multilayer (ML) surface with heights varying from 1.5*nm* to 5.5*nm*. They also show that critical dimension (CD) varies mainly with the defect height at the top of the ML surface. We calculate CD variation (ΔL) from a range of reasonable surface defect heights, i.e., 1*nm* to 8*nm*, based on Clifford et al.'s CD variation model. Table 6 summarizes the defect heights assumed in our experiments, and their respective impacts on CD and timing. To quantify the impact on timing, we measure delay variation (ΔT) from a nominal worst-case delay of a most frequently used cell (i.e., 2-input NAND gate) in our testcase with respect to transistor gate length, using a 45*nm* open-source design kit.[12]

From the delay variation due to defects, we can estimate parametric yield. When a defect is located on a timing-critical cell whose slack is less than ΔT of the defect, the die will fail due to timing errors and be counted as a yield loss.[†]

Reticle strategies. We consider various reticle strategies as illustrated in Figure 9.

[†] We ignore the fact that multiple defects on a timing path (i.e., the sum of timing variations from multiple defects) can cause a timing failure.



(a) SLR-L **(b) MLR** **(c) SLR-S**

Figure 9. Reticle strategies: (a) SLR-L, (b) MLR with same weight for all layers in top and different weights for different layers in bottom, (c) SLR-S with random location in top-left and lowest defect location in a gridded mask blank in bottom-left, and with optimal location in right.

- CASE1: *SLR-L*
- CASE2-A: *MLR*. Defects on every layer (i.e., region) have the same impact on timing
- CASE2-B: *MLR*. Defects only on critical layers (e.g., poly) affect timing
- CASE3-A: *SLR-S*. Mask location is selected randomly in a 2-D lattice of available locations in a mask blank
- CASE3-B: *SLR-S*, with mask generated at the lowest defect-density region in a 2-D lattice of available locations in a mask blank
- CASE4: *SLR-S*, with mask generated at the lowest defect-density region with no restriction in the location. CASE3-B maximizes the number of available masks per mask blank (e.g., 9 fields), but CASE4 may not.

Intuitively, one can expect that CASE1 and CASE2-A should have the same yield when all layers have the same sensitivity to defects, since overall yield is a cumulative yield of all layers for both cases. CASE1 and CASE2-B should have the same yield when only one critical layer (e.g., poly) is sensitive to defects. In both cases, yield only depends on the yield of the critical layer. In addition, CASE2-B and CASE3-A should have the same yield, since both cases use the same region in the mask blank. Clearly, CASE4 will have better yield than CASE3-B, since CASE4 has no constraints for the location of the mask. CASE3-B will have better yield than CASE3-A, since CASE3-B can the region with lowest defect density out of nine available regions in a mask blank. Hence, assuming that only defects on critical layers have impact, there are four distinct cases; CASE1, CASE2-B, CASE3-B and CASE4.

4.1.2 Yield Calculation

We calculate timing-critical regions in a design from a signoff static timing analysis. We find a list of timing-critical cells whose timing slack is less than the timing variation due to defects (ΔT), and obtain a list of bounding boxes of timing-critical cells from placement information (e.g., Design Exchange Format (DEF)[15]). Using the timing-critical regions in a die and randomly-placed defect regions in a mask blank, we check whether any defect region overlaps with any timing-critical regions of dies in a field. If there is an overlap, the die is regarded as failed. This geometric manipulation reduces simulation time required to perform actual timing analysis with defect-induced linewidth variation.

We note that CASE4 shows zero yield loss with the reasonable defect densities that we assumed. Although CASE4 can have a yield loss with very high defect densities, the runtime for the overlap checking increases excessively. Hence, for CASE4, we calculate a lower bound for defect density which incurs a yield loss, instead of performing the overlap checking.

To calculate a lower bound of defect density, we define the following sets of regions.

- S_C: set of timing-critical regions in a field. List of bounding boxes of timing-critical cells that would result in parametric failure when intersecting with defect locations.
- S_F: set of forbidden regions in a mask blank where mask origin should not be located, to avoid overlap of defect regions with S_C.
- S_P: set of feasible regions in a mask blank where mask origin can be located with no overlaps between S_F and S_C. S_P is calculated by subtracting S_F from the bounding box of the entire mask blank.

Figure 11 illustrates a simple example of the forbidden region calculation for a single point defect and a timing-critical cell. When a defect p is located at (p_x, p_y) in a mask blank, and there is one timing-critical region r at (r_x, r_y) with width of r_w and height of r_h, the mask origin should not be placed in the red region defined by the lower-left corner at $(p_x - r_x - r_w,\ p_y - r_y - r_h)$ and the upper-right corner at $(p_x - r_x,\ p_y - r_y)$ as shown in Figure 11. If the mask origin is placed in the red region, timing-critical region r must be overlapped by the defect.[‡] Figure 10 shows the procedure to calculate S_F for a single defect. Each defect defines $|S_C|$ rectangular regions in S_F, and we iterate the procedure for all defects in the mask blank to obtain S_F.

```
Procedure: FORBIDDEN_REGION
Inputs: defect $p$ at $(p_x, p_y)$
Outputs: forbidden region $S_F(p)$

$S_F(p) \leftarrow \emptyset$
foreach timing-critical region $r \in S_C$
    calculate a defect region $f\ (x_1, y_1, x_2, y_2)$ by
        $x_1 \leftarrow p_x - (r_x + r_w)$
        $y_1 \leftarrow p_y - (r_y + r_h)$
        $x_2 \leftarrow p_x - r_x$
        $x_2 \leftarrow p_y - r_y$
    $S_F(p) \leftarrow S_F(p) \cup f$
end
```

Figure 10. Procedure to calculate forbidden regions due to a single defect.

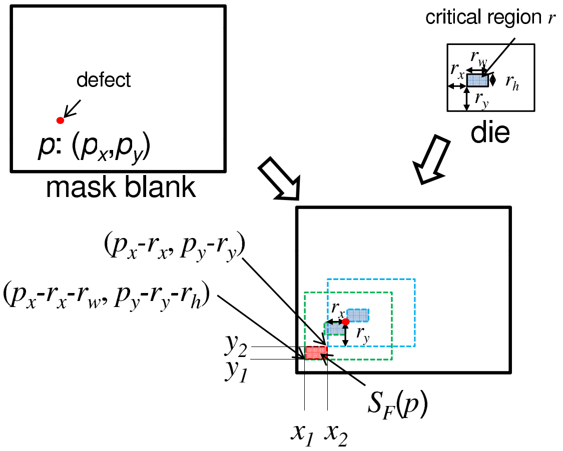


Figure 11. An example of forbidden region calculation for a single defect.

With a pessimistic assumption that no forbidden regions due to different defects intersect each other, the area of S_F, which is a union of all forbidden regions, is simply calculated as the area of S_C multiplied by the number of defects. As the number of defects increases, the area of S_F increases and the area of S_P decreases. When the area of S_F is equal to the area of the mask blank, the area of S_P reaches zero and CASE4 must have a yield loss regardless of the choice of the mask location. Hence, the lower bound of the number of defects is calculated as $Area(S_{P0})/Area(S_C)$ where S_{P0} is the area of feasible region without defects. S_{P0} is calculated as $(width_{field} - width_{die}) \times (height_{field} - height_{die})$. If the number of defects does not exceed the lower bound, then there must exist a nonempty subset of the feasible region within which a die can be located, and hence CASE4 has 100% yield.

[‡] For nonzero-area defects, the calculation method is similar, with the dimensions of a forbidden region expanded by the width and height of defects.

4.2 EUVL Parametric Yield Comparison

We calculate parametric yield due to EUVL defects for a given number of mask sets, i.e., 1,000 sets. We furthermore evaluate the parametric yield sensitivity to defect parameters, such as defect density d, defect height h, defect influence distance r, and defect distribution method m.

Parametric yield versus defect density. Our first experiment compares the parametric yield changes due to defect density. For this experiment, other parameters are fixed in a reasonable ranges. Defect height is assumed as $4nm$ and defect influence distance is assumed as $30nm$ on wafer ($120nm$ on reticle), which is 2× a typical FWHM reported by Clifford et al.[10] Defects are assumed to have a uniform random distribution. Figure 12 compares parametric yields of various reticle strategies. CASE2-B has the worst yield, since it assumes that a possible problematic mask for a critical layer in *MLR* is used for all dies. CASE1 has better yield than CASE2-B, but still has lower yield than other two cases, since several of the dies in a field can be affected by defects. CASE4 shows perfect yield, since there is large flexibility to place a critical layer on a mask blank avoiding defects,[§] and CASE3 shows the second best yield. While the yield trends are clear, we note that the differences between cases are not significant in the range of reasonable defect densities.

Parametric yield versus defect height. Our second experiment assesses parametric yield changes due to defect height. Since defect height determines the CD variation, timing impact and hence the timing-critical area in a design (i.e., S_C) are affected. For this experiment, defect density is fixed at a reasonable range, i.e., 0.444-2.222 defects/cm^2, defect influence distance is assumed as $30nm$, and defects are assumed to have a uniform random distribution. Figure 13 compares parametric yields of various reticle strategies. We observe that parametric yield is not significantly changed due to the defect height. The reason is that the timing-critical region is relatively small compared to the entire field area, and this swamps even the assumption of a pessimistic defect height, e.g., 8*nm*.

[§] $Area(S_C)$ of our testcase is 22,443.4μm^2 with $250nm$ defect influence distance on wafer ($1000nm$ on reticle), and $Area(S_{P0})$ is $576mm^2$ (= $(32mm$ - $8mm)^2$). The lower bound of the number of defects $Area(S_{P0})/Area(S_C)$ is 25,665. As long as the number of defects is less than 25,665, CASE4 has 100% yield.

Parametric yield versus defect influence distance. Our third experiment assesses parametric yield impact of the defect influence distance. We examine zero influence defect distance (i.e., point defect), a reasonable influence distance (i.e., 2× typical FWHM), and a very large influence distance (i.e., 1,000*nm* on reticle (250*nm* on wafer)), with 4*nm* defect height and uniform defect distribution.[¶] Figure 14 compares parametric yields for various reticle strategies. We see that the yield sensitivity to defect influence distance is negligibly small. For 0 and 120*nm* distance, there is almost no difference. With larger defect influence, i.e., 1,000*nm* on reticle, yield is reduced, but the yield loss is still insignificant. This again may be attributed to the relatively small timing-critical region in a design.

Parametric yield versus defect distribution. Finally, Figure 15 assesses the yield difference between uniform and decentered-Gaussian defect distributions. CASE4 still shows perfect yield. In addition, CASE3-B with decentered Gaussian distribution has also shows perfect yield, since the center dies in 16 possible locations have low defect probability due to the construction of the decentered Gaussian distribution. However, the worst case of CASE2-B, where field location is chosen along the boundary of the mask blank, shows a sharp yield loss. Except for CASE2-B, yield with the decentered Gaussian defect distribution is higher than yield with the uniform defect distribution.

4.3 Significance of EUVL Defectivity

From the experiments, our observations are summarized as follows.

- As defect density increases, parametric yield decreases.
- As defect height increases, parametric yield decreases.
- As defect influence distance increases, parametric yield decreases.
- Decentered Gaussian random distribution assumption even reduce the parametric yield loss. Especially, when we are looking for a best location for critical layer to be placed in a mask blank, it allows lower defects density near the center of the mask blank.

[¶] Although typical mask defect size is as small as less than 100*nm*, the map of defect locations produced by the inspection process may not be accurate (e.g., around 500*nm* resolution in *x*- and *y*-coordinates, respectively). Hence, the case of 1,000*nm* defect influence distance may not be overly pessimistic.

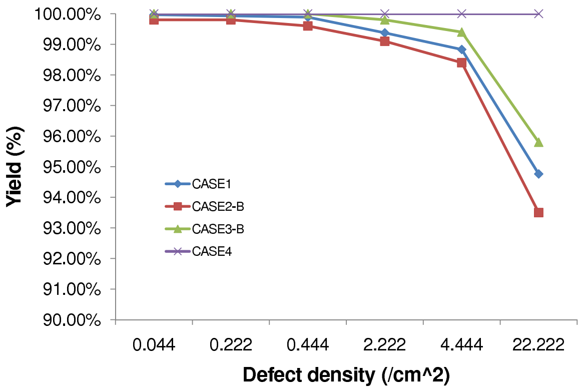


Figure 12. Defect density versus yield for various reticle strategies. 4*nm* defect height and 120*nm* (4×) defect influence distance are assumed, with defects uniformly distributed.

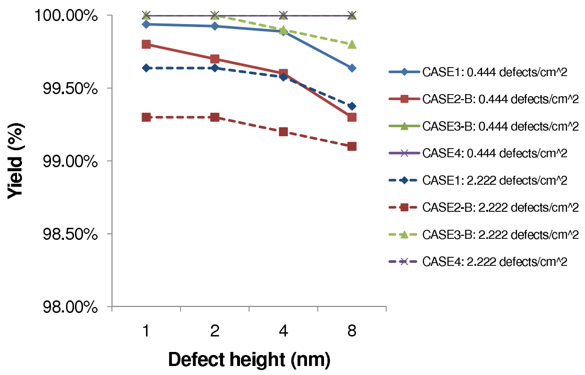


Figure 13. Defect height versus yield for various reticle strategies. 0.444-2.222 defects/cm^2 defects are uniformly distributed and defect influence distance is assumed as 120*nm* (4×).

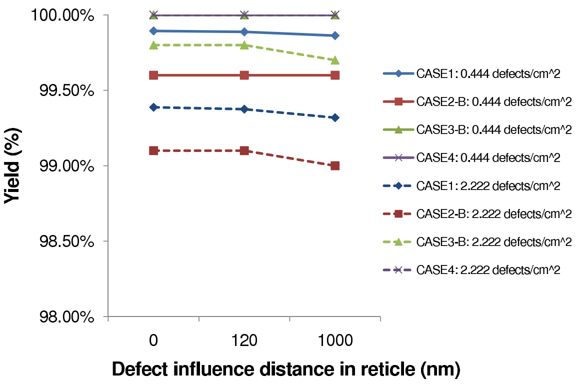


Figure 14. Defect influence distance versus yield for various reticle strategies. 0.444-2.222 defects/cm^2 defects with 4*nm* height are uniformly distributed.

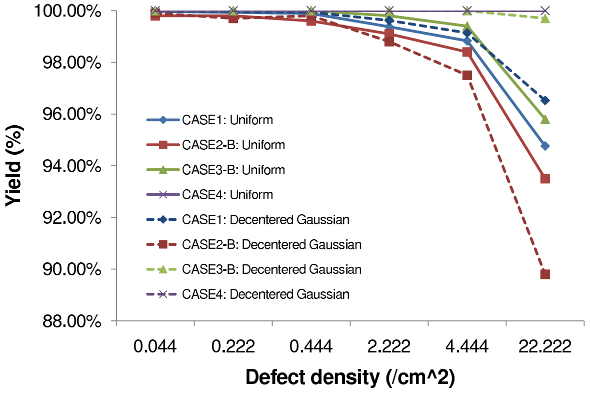


Figure 15. Defect distribution methods versus yield for various reticle strategies. Defects with 4*nm* height and 120*nm* (4×) influence distance are distributed.

These observations are fairly intuitive, and they support the notion that defect should be accurately identified and cleaned as much as possible to mitigate potential defect-induced parametric yield loss. Interestingly, however, our studies indicate that the parametric yield loss due to mask blank defects may not be as significant as has been recently thought by most EUVL researchers. The main reason is that in typical designs the timing-critical region that can be affected by mask blank defects is quite small relative to the entire design area. Table 7 shows the relative size of the timing-critical region of several real designs implemented in 65*nm* and 45*nm* technologies. (In the table, the testcase used for yield calculation is based on an MPEG2 core.) Hence, as long as the relative size of the timing-critical region does not increase significantly, mask blank defectivity may not be the most critical issue for near-term EUVL adoption, and more concerns and efforts can be devoted to other technical hurdles for EUVL.

Table 7. Portion of timing-critical regions in real designs. The timing-critical area is calculated as the sum of areas of cells for which timing slack is less than 20*ps*.

65*nm*			45*nm*		
Design	Timing-critical area (%)	Source	Design	Timing-critical area (%)	Source
MPEG2	1.077	Opencores[13]	AES	2.068	Opencores[13]
AES	1.746	Opencores[13]	JPEG	0.187	Opencores[13]
JPEG	0.442	Opencores[13]			

5. CONCLUSION

We have provided new yield-aware mask strategies to mitigate emerging variability and defectivity challenges. Our study has analyzed CD variability with respect to reticle size, and quantified its impact on parametric yield. We have also integrated parametric yield depending on field size with a cost model that incorporates mask, wafer, and processing cost considering throughput, yield, and manufacturing volume. This enables assessment of various reticle strategies (e.g., single layer reticle (SLR), multiple layer reticle (MLR), and small and large size) considering field-size dependent parametric yield. Another aspect of our study addresses defect-induced parametric yield in EUVL, where we assess the sensitivity of parametric yield to several defect parameters, i.e., defect density, height, distribution and influence distance. We then compare parametric yields of various reticle strategies. Our study confirms a clear cost benefit from use of small-field rather than traditional full-field reticles when the volume size is small. Furthermore, our study shows that small-size field in EUVL can have significantly higher parametric yield in light of EUVL mask blank defectivity. Our ongoing work seeks to update the cost model for future technologies with various mask and patterning technologies (DPL, EUVL, imprint, etc.), and include more data for various design types (SoC, MPU, ASIC, etc.) and design sizes in order to derive realistic design-dependent defectivity requirements.

REFERENCES

[1] *International Technology Roadmap for Semiconductors*, http://public.itrs.net/.

[2] B. J. Grenon, “Mask Costs, A New Look”, *Proc. SPIE European Mask and Lithography Conference*, Vol. 6281, 2006, pp. 628101-1–628101-7.

[3] W. Trybula, “A Common Base for Mask Cost of Ownership”, *Proc. SPIE Photomask Technology*, Vol. 5256, 2003, pp. 318–323.

[4] A. Balasinski, “Multi-Layer and Multi-Product Masks: Cost Reduction Methodology”, *Proc. SPIE Photomask Technology*, Vol. 5567, 2004, pp. 351–359.

[5] A. Balasinski, J. Cetin, A. B. Kahng and X. Xu, “A Procedure and Program to Calculate Shuttle Mask Advantage”, *Proc. SPIE Photomask Technology*, Vol. 6349, 2006, pp. 63492B-1–63492B-8.

[6] J. Burns and M. Abbas, “EUV Mask Defect Mitigation Through Pattern Placement”, *Proc. SPIE Photomask Technology*, Vol. 7823, 2010, pp. 782340-1–782340-5.

[7] D. van den Heuvel, R. Jonckheere, J. Magana, T. Abe, T. Bret, E. Hendrickx, S. Cheng and K. Ronse, “Natural EUV Mask Blank Defects: Evidence, Timely Detection, Analysis and Outlook”, *Proc. SPIE Photomask Technology*, Vol. 7823, 2010, pp. 78231T-1–78231T12.

[8] U. Okoroanyanwu, A. Tchikoulaeva, P. Ackmann, O. Woord, B. La Fontaine, K. Bubke, C. Holfeld, J. H. Peters, S. Kini, S. Watson, I. Lee, B. Mu, P. Lim, S. Raghunathan and C. Boye, “Assessing EUV Mask Defectivity”, *Proc. SPIE Extreme Ultraviolet Lithography*, Vol. 7637, 2010, pp. 76360J-1–76360J-12.

[9] A. Rastegar, “Overcoming Mask Blank Defects in EUV Lithography”, *SPIE Newsroom*, 2009.

[10] C. H. Clifford and A. R. Neureuther, “Smoothing Based Model for Images of Isolated Buried EUV Multilayer Defects”, *Proc. SPIE Emerging Lithography Technologies XII*, 2008, pp.692119-1–692119-10.

[11] D. Pramanik, H. H. Kamberian, C. J. Progler, M. Sanie and D. Pinto, “Cost Effective Strategies for ASIC Masks”, *Proc. SPIE Cost and Performance in Integrated Circuit Creation*, Vol. 5043, 2003, pp. 142–152.

[12] *NANGATE*, http://www.nangate.com/

[13] *OPENCORES.ORG*, http://www.opencores.org/.

[14] *Sun OpenSPARC Project*, http://www.sun.com/processors/opensparc/.

[15] *Cadence Design Exchange Format*, http://openeda.si2.org/projects/lefdef/.