Simultaneous Flare Level and Flare Variation Minimization with Dummification in EUVL*

Shao-Yun Fang[1] and Yao-Wen Chang[1,2]
[1]Graduate Institute of Electronics Engineering, National Taiwan University, Taipei 106, Taiwan
[2]Department of Electrical Engineering, National Taiwan University, Taipei 106, Taiwan
yuko703@eda.ee.ntu.edu.tw; ywchang@cc.ee.ntu.edu.tw

ABSTRACT

Extreme Ultraviolet Lithography (EUVL) is one of the most promising Next Generation Lithography (NGL) technologies. Due to the surface roughness of the optical system used in EUVL, the rather high level of flare (i.e., scattered light) becomes one of the most critical issues in EUVL. In addition, the layout density non-uniformity and the flare periphery effect (the flare distribution at the periphery is much different from that in the center of a chip) also induce a large flare variation within a layout. Both of the high flare level and the large flare variation could worsen the control of critical dimension (CD) uniformity. Dummification (i.e., tiling or dummy fill) is one of the flare compensation strategies to reduce the flare level and the flare variation for the process with a clear-field mask in EUVL. However, existing dummy fill algorithms for Chemical-Mechanical Polishing (CMP) are not adequate for the flare mitigation problem in EUVL due to the flare periphery effect. This paper presents the first work that solves the flare mitigation problem in EUVL with a specific dummification algorithm flow considering global flare distribution. The dummification process is guided by dummy demand maps, which are generated by using a quasi-inverse lithography technique. In addition, an error-controlled fast flare map computation technique is proposed and integrated into our algorithm to further improve the efficiency without loss of computation accuracy. Experimental results show that our flow can effectively and efficiently reduce the flare level and the flare variation, which may contribute to the better control of CD uniformity.

Categories and Subject Descriptors

B.7.2 [**Integrated Circuits**]: Design Aids

General Terms

Algorithms, Design, Performance

Keywords

Extreme Ultraviolet Lithography, Flare, Dummification, Manufacturability

1. INTRODUCTION

Extreme Ultraviolet Lithography (EUVL) is one of the most promising Next Generation Lithography (NGL) technologies since the ten times reduction in wavelength in EUVL offers the capability of a continuation of Moore's law beyond the 22 nm technology node [1, 17]. However, the used light of 13.5 nm wavelength is not transmitted, but absorbed by most of materials, and thus only reflective optical components and masks can be used. Due to the surface roughness of the optical system, flare, which is undesired scattered light contribute to wafer exposure, is one of the most critical issues in EUVL, as illustrated in Figure 1(a).

For the process using a clear-field mask that is also made of reflective materials in EUVL, the layout patterns are formed by absorbers on the mask, as illustrated in Figure 1(b). Thus, during an exposure process on the wafer, the vacant regions not covered by layout patterns will be exposed by the light, and vice versa. However, the scattered flare reduces the contrast between bright regions (vacant regions) and dark regions (layout patterns), and may result in critical dimension (CD) distortion.

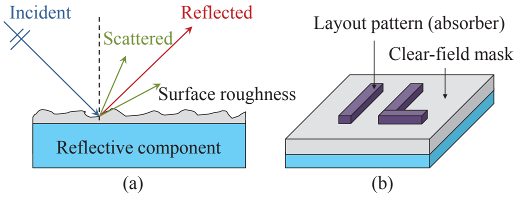


Figure 1: (a) Flare is undesired scattered light due to the surface roughness of reflective optical components and masks used in EUVL. (b) A clear-field mask on which layout patterns are formed by light absorbing materials.

Since the flare is proportional to the surface roughness of the optical system and inversely proportional to squared wavelength [12, 19], EUVL suffers from rather high level of flare compared to traditional lithography technologies. It is reported by the alpha demo tool (ADT) at IMEC that the intrinsic flare is about 16% [12]. On the other hand, the regions at the periphery of a chip receive much less flare compared to the regions in the center of a chip (assuming the regions outside the chip boundaries are dark-fields), causing a large flare variation [9]. We refer to the phenomenon as *the flare periphery effect*. In addition, the non-uniformity of layout patterns may contribute to the flare variation within a chip as well. For the process with a clear-field mask, regions with lower pattern density contribute to more flare distribution than those with higher pattern density [9, 10]. Since the high flare level causes CD distortion and the flare variation damages CD uniformity, flare compensation strategies are required.

There are two strategies for flare compensation. One is applying global CD resizing similar to optical proximity correction (OPC) on pattern features according to the flare value received by each feature [13, 19]. However, previous work has reported that one percent change in flare level may cause 10 nm change in CD (CD sensitivity on flare is 10 nm/% Δflare level) at the 22 nm technology node and may be considerably larger for more advanced technology nodes [10]. As a result, a large flare variation may not be fully compensated by applying

*This work was partially supported by IBM, SpringSoft, TSMC, and NSC of Taiwan under Grant No's. NSC 100-2221-E-002-088-MY3, NSC 99-2221-E-002-207-MY3, NSC 99-2221-E-002-210-MY3, and NSC 98-2221-E-002-119-MY3.

DAC'12, June 03 - 07, 2012, San Francisco, California, USA.
Copyright 2012 ACM 978-1-4503-1199-1/12/06 ...$10.00.

a global CD resizing and may cause the difficulty of controlling CD uniformity accurately. Another flare compensation strategy is dummification (i.e., tiling or dummy fill) [2, 9, 13]. By adding dummy patterns according to global flare distribution, intra-chip flare variation could be reduced. Although dummification may be limited to design rules and layout constraints, it has been shown that dummification can simultaneously reduce intra-chip flare level and flare variation in EUVL for the process with a clear-field mask, and thus may greatly simplify the flare compensation methodology with global CD resizing [10].

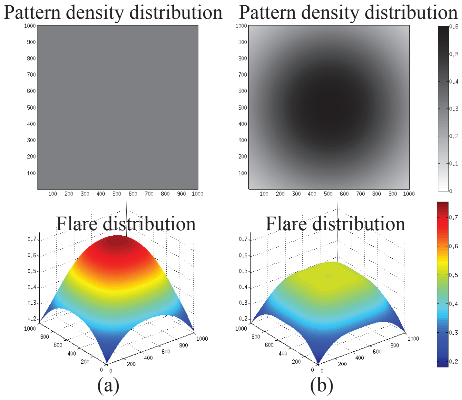


Figure 2: Flare comparison between layouts with different density distributions. (a) A layout with uniform density distribution may have large flare variation due to the flare periphery effect. (b) A layout with density distribution conforming to the global flare distribution has smaller flare variation.

There are many existing dummy fill algorithms for improving Chemical-Mechanical Polishing (CMP) quality [3, 4, 7, 18, 6]. The objectives of these algorithms mainly focus on density variation minimization or density gradient minimization for a layout. However, the flare variation of a layout with minimized layout density variation could be far from optimal; that is, achieving maximum pattern density uniformity is not equivalent to achieving maximum flare uniformity in EUVL. As shown in Figure 2(a), a layout with merely completely uniform layout density distribution may still suffer from large flare variation due to the flare periphery effect. Another layout with smaller flare variation is shown in Figure 2(b), in which the layout density distribution is less uniform but conforms to the global flare distribution (trend) of the layout. Thus, a more sophisticated dummification algorithm for flare variation minimization in EUVL is required.

In this paper, we propose the first work that solves the flare mitigation problem in EUVL for the process with a clear-field mask with a dummification algorithm considering global flare distribution. Since it is computationally expensive for deriving the global flare distribution of a layout, we first propose an error-controlled fast flare map computation approach to improve the efficiency of the algorithm without loss of computational accuracy. Then, we present a dummification process consisting of two stages: (1) the global dummification stage followed by (2) the local refinement stage. Experimental results based on the MCNC and the Faraday benchmarks show that our flow can effectively and efficiently reduce the flare level and the flare variation, which can contribute to the better control of the CD uniformity within a layout.

The rest of this paper is organized as follows: Section 2 gives some preliminaries and the problem formulation of this paper. In Section 3, an error-controlled fast flare map computation method is introduced. Section 4 details the proposed dummification algorithm for simultaneous flare level and flare variation minimization. Experimental results are reported in Section 5. Finally, we conclude our work in Section 6.

2. PRELIMINARIES

In this section, the preliminaries and the problem formulation of flare optimization with dummification are given.

2.1 Flare Map Computation

Practically, flare in EUVL can be modelled as a scattering point spread function (PSF), and the flare distribution can be obtained by convolving the PSF with the original image intensity I_0. Since flare in EUVL could result in significant change in the image intensity, accurate flare map computation with high resolution is required for implementing flare compensation. However, the flare PSF has a very large coverage. It has been reported that an area described by a radial distance of about 1000 μm accounts for only about 95% of the flare seen at a point on a wafer [15]. Due to the long-range effects of the flare PSF and the high complexity of I_0, directly convolving the PSF with I_0 could be very computationally expensive. To tackle this problem, previous work has shown that dividing a layout into suitably sized grids (e.g., $1\mu m \times 1\mu m$) and calculating the layout density for each grid can achieve a good approximation of I_0 [13, 19]. Then, by convolving the generated density map $I_D(x, y)$ at the coordinate (x, y) with the discrete PSF(x, y), we can derive the flare map $I_F(x, y)$ of a layout as follows:

$$I_F(x, y) = I_D(x, y) \otimes PSF(x, y). \tag{1}$$

Note that for a clear-field mask, flares are distributed from vacant regions without the coverage of patterns. Thus, in our work, density maps are referred to as *vacancy density maps*.

However, for full chip flare map generation, the computation process may still be too time-consuming. Therefore, other speed-up techniques without lose of computational accuracy are required for flare map computation.

2.2 Flare Reduction with Dummification

Dummification (tiling or dummy pattern insertion) is a simple method to reduce the flare effects in EUVL. Although dummification may be constrained by layout patterns, the technique can significantly mitigate the flare level and the flare variation for the process with a clear-field mask [13]. In addition, dummification may also greatly simplify the flare compensation methodology with global CD resizing [10].

Some previous work has proposed ideas to perform dummification for flare mitigation in EUVL. Singh et al. developed an iterative methodology that adds auxiliary patterns by utilizing commercial tools which are mainly driven by polishing requirements (e.g., CMP) [15]. However, as pointed out in Section 1, achieving maximum pattern density uniformity is not equivalent to achieving maximum flare uniformity in a dummification process due to the flare periphery effect in EUVL. Thus, the dummification algorithms for CMP are not suitable for flare mitigation in EUVL.

Another idea is to vary the size of dummies according to the location within a field [16]. This inspires an intuitive dummification method that varies dummy densities as a linear function of distance from the center of a layout. For regions far from the center of a layout, the flare values are expected to be smaller, and thus fewer dummy patterns are required. In contrast, more dummy patterns are required for those regions near the center of a layout. Although dummification with a linear function considers the flare periphery effect in EUVL, the solution space is limited and global flare distribution is not considered as well. Thus, a dummification method considering global flare distribution is desired.

2.3 Problem Formulation

Given an input layout, we first divide the layout into fine grids and analyze the vacancy density and the maximum available dummy density (the maximum areas dummy patterns can be inserted without conflicting with the original design) for each gird. A flare map is then computed by convolving the discrete flare PSF with the vacancy density map. After that, the dummification process can be formulated as a dummy value assignment problem considering global flare distribution. The problem formulation of simultaneous flare level and flare variation mitigation with dummification for the process with a clear-field mask in EUVL can be described as follows:

Problem 1. *Given a grid-based layout, assign a dummy density value to each grid such that the flare level and the flare variation in the flare map of the layout with dummification are simultaneously minimized.*

After deriving a dummy density value d_i for each grid g_i, we simply insert dummies in g_i with their total area equal to $d_i \times a_i$, where a_i is the area of a grid. In addition, dummies are inserted in the dummy available regions of a grid to ensure that the inserted dummies would not conflict with the original design.

3. ERROR-CONTROLLED FAST FLARE MAP COMPUTATION

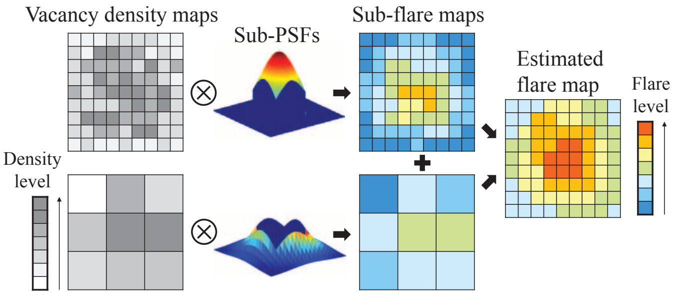
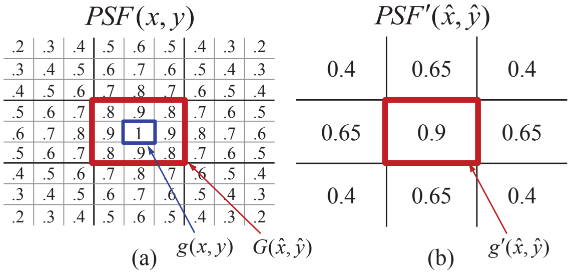


Figure 3: Estimated flare map computation with multiple convolutions between density maps and sub-PSFs.

As mentioned in Section 2.1, the full chip flare map generation is computationally expensive. Some previous work proposed a technique applying multiple convolutions with coarsened grids of different sizes instead of performing one convolution with very fine grids to speed up the computation process [10, 12, 15].

Figure 3 illustrates the method. The PSF is divided into several sub-PSFs with different, yet uniform, grid sizes. For a sub-PSF with larger variation, its grid size is chosen to be smaller; otherwise, the grid size is chosen to be larger. By convolving each sub-PSF and a density map with the same grid size and by summing the generated sub-flare maps, the estimated flare map is generated.



.2	.3	.4	.5	.6	.5	.4	.3	.2
.3	.4	.5	.6	.7	.6	.5	.4	.3
.4	.5	.6	.7	.8	.7	.6	.5	.4
.5	.6	.7	.8	.9	.8	.7	.6	.5
.6	.7	.8	.9	1	.9	.8	.7	.6
.5	.6	.7	.8	.9	.8	.7	.6	.5
.4	.5	.6	.7	.8	.7	.6	.5	.4
.3	.4	.5	.6	.7	.6	.5	.4	.3
.2	.3	.4	.5	.6	.5	.4	.3	.2

0.4	0.65	0.4
0.65	0.9	0.65
0.4	0.65	0.4

Figure 4: Illustration of deriving a sub-PSF from the original PSF. The values indicate the function values of the original PSF and the sub-PSF. (a) The original PSF with finest grids. (b) A sub-PSF with its grid size equal to 3.

However, the accuracy of an estimated flare map strongly depends on the division of the PSF and the resolution of each sub-PSF. No previous work mentioned how to control the accuracy of an estimated flare map. Therefore, we define a criterion for judging whether a sub-PSF is a good approximation of the original PSF or not. In addition, we propose an error-controlled sub-PSFs generation method such that the error between an original flare map and its estimated flare map can be controlled better. We first give some notations used as follows:

- PSF': a sub-PSF.
- $g(x,y)/g'(\hat{x},\hat{y})$: a grid in the PSF/$PSF'$.
- $PSF(x,y)/PSF'(\hat{x},\hat{y})$: a function value of the PSF/PSF' at $g(x,y)/g'(\hat{x},\hat{y})$.
- $G(\hat{x},\hat{y})$: a set of grids $g(x,y)$ contained in $g'(\hat{x},\hat{y})$.

Figure 4(a) illustrates a PSF and Figure 4(b) shows a sub-PSF PSF' with its grid size equal to 3. The values shown in Figure 4 are the function values of the PSF and the sub-PSF. As illustrated in Figure 4, the grids belonging to $G(\hat{x},\hat{y})$ (enclosed by the red/thicker window) in Figure 4(a) are contained in $g'(\hat{x},\hat{y})$ in Figure 4(b). Based on the notations, we give the following definitions as a criterion for judging how a sub-PSF is approximated to the original PSF.

Definition 1. *A function value $PSF'(\hat{x},\hat{y})$ in a PSF' is said to be ϵ-controlled if for each $g(x,y) \in G(\hat{x},\hat{y})$, the error between $PSF(x,y)$ and $PSF'(\hat{x},\hat{y})$ is less than ϵ%.*

Definition 2. *A PSF' is an ϵ-controlled sub-PSF if all function values $PSF'(\hat{x},\hat{y})$ are ϵ-controlled.*

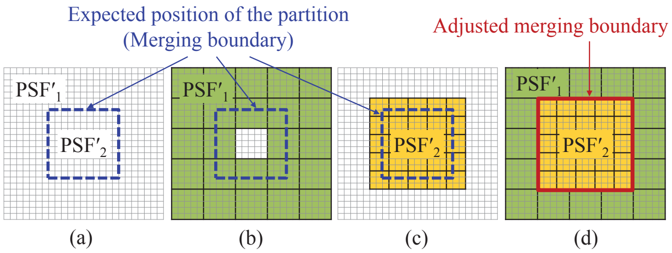

Expected position of the partition (Merging boundary)
Adjusted merging boundary
PSF'1
PSF'2
PSF'1
PSF'1
PSF'2
PSF'1
PSF'2
(a)
(b)
(c)
(d)

Figure 5: Illustration of the error-controlled sub-PSFs generation. (a) The original PSF is to be partitioned into two sub-PSFs according to the expected position of the partition. (b) PSF'_1 with its maximized grid size is error-controlled. (c) PSF'_2 with its maximized grid size is error-controlled. (d) The merging boundary is adjusted such that the two sub-PSFs are disjoint.

Thus, based on Definition 2, our objective is to find a set of ϵ-controlled sub-PSFs such that the union of the sub-PSFs is a good approximation of the original PSF. The set of sub-PSFs should be chosen such that the union of the sub-PSFs covers the original PSF and the PSFs are disjoint with each other. The ϵ-controlled sub-PSFs generation process is shown in Figure 5. Figure 5(a) shows an original PSF with finest grids. Suppose we want to partition the PSF into two sub-PSFs, PSF'_1 and PSF'_2. The blue (dashed) window in Figure 5(a) indicates the expected position of the partition. For each sub-PSF, we maximize its grid size and keep the sub-PSF to be ϵ-controlled to get better efficiency and accuracy. The maximized grid sizes of sub-PSFs may be different since the gradients of PSF values vary from the center to the periphery of the PSF. In addition, each sub-PSF needs to cover all function values in its corresponding region according to the partition positions. As illustrated in Figures 5(b) and 5(c), PSF'_1 and PSF'_2 have different maximized grid sizes, and they both cover function values in their corresponding regions. After separately deriving each sub-PSF, the sub-PSFs might be overlapped with each other on the merging boundaries. Therefore, as we merge the sub-PSFs, the merging boundaries are adjusted such that the sub-PSFs are disjoint with each other. As illustrated in Figure 5(d), the adjusted merging boundary is indicated by the red (solid) window, and the two sub-PSFs are disjoint. After getting the set of ϵ-controlled sub-PSFs, an estimated flare map can be efficiently computed with better error control.

4. FLARE LEVEL AND FLARE VARIATION MINIMIZATION WITH DUMMIFICATION

In this section, we present our dummification algorithm for simultaneous flare level and flare variation minimization. We first give the algorithm flow in Section 4.1, and two major stages in the flow, the global dummification stage and the local refinement stage, are then detailed in Section 4.2 and 4.3.

4.1 Algorithm Flow

Figure 6 shows our flare optimization algorithm flow with dummification. Given a grid-based mask layout as the input, we first generate its vacancy density map and compute the corresponding flare map. After that, dummy value assignment is performed with the guidance generated by using a quasi-inverse lithography technique to evaluate the dummy demands of different regions. The dummification process is composed of two major stages: the global dummification stage followed by the local refinement stage.

In the global dummification stage, we simultaneously optimize flare level and flare variation with a top-down dummy value assignment method. Since flare in EUVL has a long-range effect, the top-down

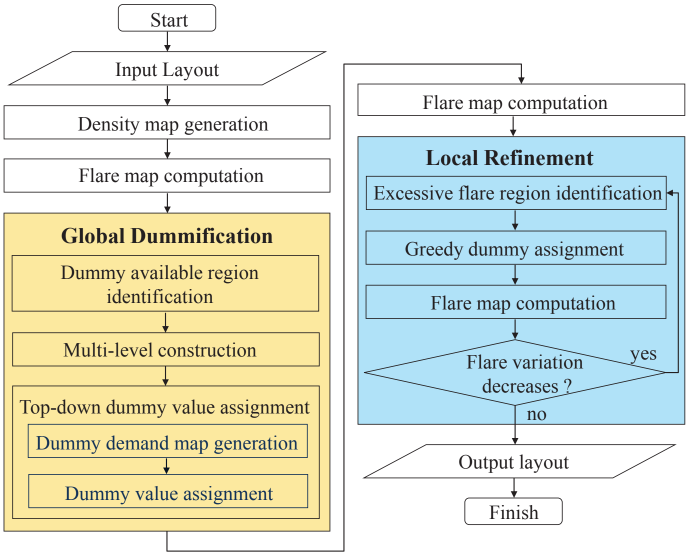


Figure 6: Overall flow of the proposed simultaneous flare level and flare variation minimization algorithm with dummification.

approach can capture the information of global flare distribution better than a bottom-up counterpart since the top-down one has a more global view. After identifying the available dummy region for each grid, we construct the multilevel structure by accumulating the flare map and the available dummy map. Then we repeatedly assign dummy values to subregions from the top level to sub-levels by applying a quasi-inverse lithography technique, which evaluates the demand of dummies for each subregion.

In the local refinement stage, the objective is to further minimize the flare variation of a chip such that the CD uniformity can be controlled better. We iteratively identify grids with excessive flare and greedily assign maximum available dummy values to grids with higher refinement demands by applying a similar quasi-inverse lithography technique. The optimization process terminates when no improvement in flare variation reduction can be made through this local refinement approach.

Note that flare map computation for a whole chip is needed for each stage in our algorithm flow. Especially in the local refinement stage, flare map computation is performed for each iteration for judging whether the flare variation is reduced or not. Consequently, efficient and accurate flare map computation is necessary. Thus, we accommodate the error-controlled fast flare map computation introduced in Section 3 to improve the program efficiency without loss of computational accuracy.

We detail the two stages in Section 4.2 and Section 4.3 respectively.

4.2 Global Dummification

As mentioned in Section 1, for the process in EUVL with a clear-field mask, it is possible to simultaneously reduce the flare level and the flare variation of a layout with dummification. In addition, even for a layout with uniform pattern density, the flare variation may be still large due to the flare periphery effect; that is, achieving maximum pattern density uniformity is not equivalent to achieving maximum flare uniformity in a dummification process. Furthermore, although the flare level of a chip can be minimized by inserting as many dummy patterns as possible, the flare variation should also be minimized to get better CD uniformity control. Therefore, a dummification algorithm considering global flare distribution for flare optimization is desirable.

From Equation (1), a flare map is computed by convolving the PSF and a vacancy density map $I_D(x,y)$. Thus, a flare map of a layout with dummification can be computed as follows:

$$I'_F(x,y) = (I_D(x,y) - I_{dummy}(x,y)) \otimes PSF(x,y), \tag{2}$$

where $I_{dummy}(x,y)$ is a dummy density map. Thus, the flare reduction due to dummification can be computed with the following equation:

$$I_F(x,y) - I'_F(x,y) = I_{dummy}(x,y) \otimes PSF(x,y). \tag{3}$$

To minimize flare variation, the regions with higher flare levels require more dummies than those with lower flare levels. Therefore, by Equation (3), we propose a quasi-inverse lithography technique to guide the dummy value assignment. After computing the original flare map of an input layout, we use a quasi-inverse PSF to propagate the flare reduction demand of a region to the dummy demands of neighboring regions, inspired by the work [5] for OPC optimization. The quasi-inverse PSF function is defined as follows:

$$Q(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} PSF(x-f, y-g)PSF(f,g)df dg, \tag{4}$$

This quasi-inverse PSF models the relation of multiple points on a wafer on one point of the mask. Since the flare of a region can be compensated by inserting dummies into neighboring regions, this quasi-inverse kernel function propagates the flare reduction demand of a region to the dummy demands of neighboring regions. Therefore, the dummy demand of a region is the sum of propagated dummy demands, and thus a dummy demand map can be obtained by convolving the quasi-inverse PSF with a flare map. A dummy demand map can be computed as follows:

$$D(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I_F(x-f, y-g)Q(f,g)df dg. \tag{5}$$

The larger value of a region in a dummy demand map indicates that the region requires more dummy patterns than those regions with smaller dummy demand values for flare compensation.

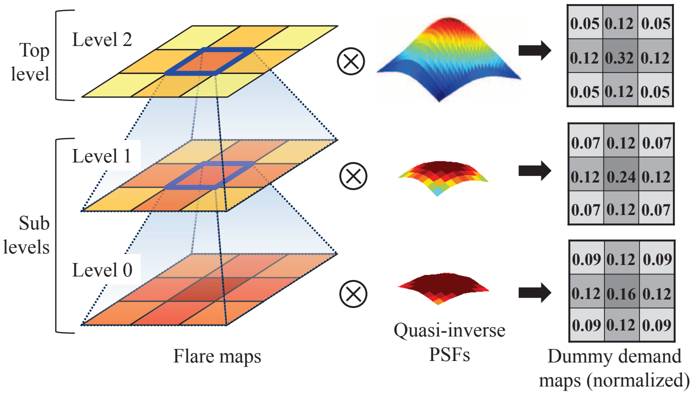


Figure 7: Illustration of the top-down dummy demand maps generation. The larger value of a grid in a dummy demand map indicates that the grid requires larger dummy value than those regions with smaller dummy demand values for flare compensation.

After identifying the available dummy area of each grid, we need to assign a dummy value not exceeding the available dummy value for each grid. However, performing dummy value assignment with a dummy demand map generated by directly using Equation (5) not only requires large computational effort, but also get an undesired dummy assignment solution, due to the lack of global information. Therefore, a top-down approach is applied which provides a more global view as solving the dummy value assignment problem. We propose a top-down framework using the quasi-inverse lithography technique with Equation (5) to derive dummy demand maps for each level. Figure 7 shows an example of generating multilevel dummy demand maps. First, the multilevel structure is constructed and $W \times W$ fine grids in level i are merged into one coarsened grid in level $i+1$. The flare values of level i are accumulated for constructing the flare map of level $i+1$, and the process is repeated until the top-level flare map is constructed. The multi-level quasi-inverse kernel functions are also constructed for each level according to the range covered by $W \times W$ grids. Then, the dummy demand maps are computed by convolving a flare map and a quasi-inverse kernel function from the top level to the bottom level. Note that each dummy demand map is normalized such that the summation of all demand values is one.

After deriving a dummy demand map for each level, the dummy assignment processes are also performed in the top-down manner. We solve the dummy assignment problems of the top-level and sub-levels by linear programming (LP), which are respectively detailed in Section 4.2.1 and Section 4.2.2.

4.2.1 *Top-Level Dummy Value Assignment*

For a top-level dummy value assignment problem, the inputs are a top-level dummy demand map and a top-level available dummy map, and the output is a top-level dummy value assignment map. For the dummy value assignment problem with a top-level dummy demand map, the objective is to maximize the total amount of assigned dummies and make the relative dummy values of any two grids conform to their relative dummy demand values. As illustrated in Figure 8(a), for the given top-level dummy demand map and the available dummy map, the assigned dummy values are maximized and conform to the dummy demand map. For example, the relative dummy values of grids i and j (10 and 1.6) conform to their relative dummy demand values (0.32 and 0.05).

We use an LP formulation to solve this top-level dummy value assignment problem. The notations used in the LP formulation are listed as follows:

- d_i: assigned dummy value of grid i.
- t_i: target dummy value of grid i.
- r_i: dummy demand value of grid i in the dummy demand map.
- a_i: available dummy value of grid i.
- m: the index of a grid with the maximum dummy demand value.
- α: user-defined parameter.

Based on the notations, the top-level dummy assignment problem can be formulated as follows:

$$\textit{maximize} \qquad \sum_i d_i - \alpha \sum_i |d_i - t_i|, \tag{6}$$

$$\textit{subject to} \qquad t_i = a_m \cdot (r_i / r_m), \forall i, \tag{7}$$

$$0 \leq d_i \leq a_i, \forall i. \tag{8}$$

In this formulation, the objective is to maximize each d_i constrained by a_i and to minimize the deviation between d_i and t_i for each grid i, where t_i is set to be a fraction of the a_m corresponding to the ratio r_i/r_m. Although the objective function is not linear due to the absolute values, the above formulation can be transformed into a linear model. We have the following theorem:

THEOREM 1. *The top-level dummy assignment problem can be solved by liner programming with linear numbers of constraints and variables.*

4.2.2 *Sub-Level Dummy Value Assignment*

Different from the top-level dummy value assignment problem, a sub-level dummy value assignment problem has one more input, the target total dummy value T of all grids in a subproblem. For a dummy value assignment problem in sub-level i, T is derived from an assigned dummy value in the previous level $i+1$. Thus, the objective of a sub-level dummy value assignment problem is to minimize the deviation between T and the sum of all dummy values and let the relative dummy values of all grids conform to a sub-level dummy demand map. As illustrated in Figure 8(b), a sub-level dummy value assignment map is generated with a sub-level dummy demand map, a sub-level available dummy map, and a target total dummy value T. In the dummy value assignment result, the sum of all dummy values equals T, and the dummy values conform to the dummy demand map. For example, the relative dummy values of grid i and grid j conform to their relative dummy demand values.

The sub-level dummy assignment problem can be formulated as follows:

$$\textit{minimize} \qquad |T - \sum_i d_i| + \alpha \sum_i |d_i - t_i|, \tag{9}$$

$$\textit{subject to} \qquad t_i = r_i \cdot T, \forall i, \tag{10}$$

$$0 \leq d_i \leq a_i, \forall i. \tag{11}$$

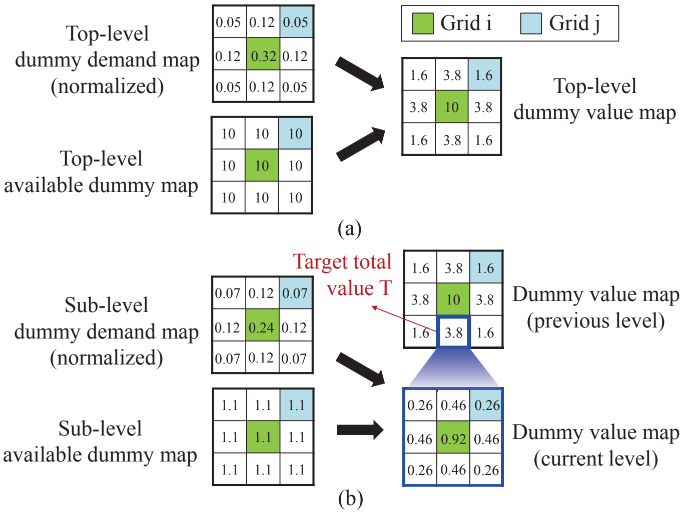


Figure 8: (a) Illustration of a top-level dummy value assignment problem. The assigned dummy values are maximized and conform to the dummy demand map. (b) Illustration of a sub-level dummy value assignment problem. The sum of all dummy values equals the target total value, and the dummy values conform to the dummy demand map.

In this formulation, the objective is to minimize the deviation between $\sum_i d_i$ and T, and the deviation between d_i and t_i for each grid i, where t_i is set to be a fraction of T according to its dummy demand value r_i. Note that a dummy demand map is normalized such that $\sum_i r_i = 1$, and thus $\sum_i t_i = T$. Similar to the top-level dummy assignment formulation, we have the following theorem.

THEOREM 2. *The sub-level dummy assignment problem can be solved by linear programming with linear numbers of constraints and variables.*

4.3 Local Refinement

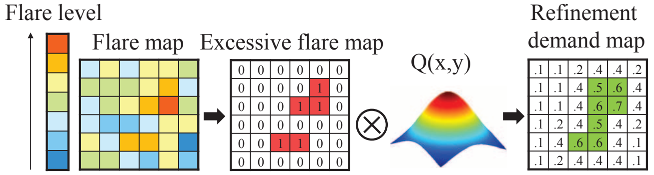


Figure 9: Local refinement. A refinement demand map is computed by convolving an excessive flare map and the quasi-inverse PSF.

After deriving a dummy value assignment solution in which the flare level and the flare variation are simultaneously optimized with our top-down framework, we try to further minimize the flare variation of a chip such that the CD uniformity can be controlled better in the local refinement stage. Figure 9 shows the local refinement process. Given the flare map of a current layout with dummification, we first identify grids with excessively large flare values in the map and construct the corresponding excessive flare map with 0/1 values. As shown in Figure 9, the red (shaded) grids marked as '1' have larger flare values than other grids. Then, we again utilize the quasi-inverse PSF (Equation (4)) to propagate the flare reduction demand of grids with value '1' in the excessive flare map to the refinement demands of neighboring regions. Then, the refinement demand of a grid is the sum of propagated refinement demands, and thus a refinement demand map can be obtained by convolving the quasi-inverse PSF with the excessive flare map (Equation (5)). For grids with refinement demand values exceeding a threshold value, we greedily assign dummy values to their maximum available dummy values and check if the flare variation of the refined layout is reduced or not. As illustrated in Figure 9, the larger refinement demand values of the green (shaded) grids indicate that the grids are desirable to be assigned with more dummies. The local refinement process is performed in an iterative manner, and the process terminates as no improvement in flare variation can be made by using this refinement approach.

5. EXPERIMENTAL RESULTS

Our algorithm was implemented in the C++ programming language on a 2.40 GHz Linux workstation with 16 GB memory. We used the lp_solve package as the LP solver [11] and used the FFTW library [8], which is a C subroutine library for computing the discrete Fourier transform (DFT), to further improve the efficiency of convolution operations. The experiments were based on two suites of benchmarks, the MCNC and the industrial Faraday benchmarks. We used the metal-1 of each circuit as the input layout.

We first pre-processed an input layout by dividing the layout into grids of size 1 $\mu m\times$ 1 μm, analyzing the pattern density of each grid, and identifying the available dummy region for each grid. The PSF was set to be a discrete Gaussian function. Since previous work has been reported that an area described by a radial distance of about 1000 μm accounts for about 95% of the flare seen at a point on a wafer [15], we set the PSF to be 2000 $\mu m\times$ 2000 μm with resolution identical to the grid-based layout, and set its standard deviation σ to be 500 μm to capture the flare distribution within 2σ. In addition, we assumed 1 unit of flare may be generated from a vacancy region of size 1 $\mu m\times$ 1 μm, and the density upper bound of each grid was set to be 0.6, which is the default value set in most commercial tools [6].

We compared the flare optimization results between two dummification approaches: (1) dummification according to a linear function of distance from the center of a layout and (2) our algorithm flow. For the first approach, the target dummy value t_i of each grid i was set according to a linear function as follows:

$$t_i = \left(1 - \frac{dist(i, center)}{dist_{max}}\right) \cdot a_{max}, \qquad (12)$$

where $dist(i, center)$ and $dist_{max}$ are the respective distances from the center of a chip to the grid i and to the farthest grid, and a_{max} is the maximum available dummy value of an empty grid (without any patterns), which is the density upper bound we set. The function tries to assign the maximum available dummy value to the central grid and make the dummy values of the farthest grids be 0. Then, we assigned a dummy value $d_i = \min(t_i, a_i)$ to each grid i, where a_i is the available dummy value of grid i.

Table 1 shows the comparison results. In the table, "Original", "Linear," and "Ours" respectively list the flare information of the original layout before dummification, the layout with dummification according to a linear function, and the layout with dummification by using our algorithm flow. "Avg." gives the average flare level, and "Var." gives the flare level variation of each circuit. Observing from the table, our algorithm achieves 36% average flare level reduction and 37% flare level variation reduction over the linear dummification mehtod. Furthermore, for the circuits with larger sizes (e.g., Struct, Primary1, and Primary2), the reductions in the flare level variation are about 50%. The significant improvements may result from the consideration of global flare distribution (trend) during the dummification process in our algorithm. These results show that our algorithm can effectively reduce the average flare level and the flare level variation, and thus the CD uniformity within a layout can be controlled better.

6. CONCLUSIONS

This paper has presented a simultaneous flare level and flare level variation optimization flow with dummification. Unlike the previous dummy fill algorithms for CMP and other heuristic methods, our algorithm performs dummification by considering global flare distribution (trend) in EUVL. In addition, the error-controlled fast flare map computation was integrated into the flow to further improve the algorithm efficiency without loss of computational accuracy. Experimental results based on two suites of benchmarks have shown that our algorithm can effectively and efficiently reduce the average flare level and the flare level variation, and thus the CD uniformity within a layout can be controlled better.

Table 1: Comparison of dummification results between a linear approach and our algorithm.

Circuit	Original		Linear			Ours		
	Avg.	Var.	Avg.	Var.	CPU (sec)	Avg.	Var.	CPU (sec)
Struct	0.850	0.743	0.606	0.535	66	0.382	0.288	155
Primary1	0.877	0.748	0.626	0.588	93	0.381	0.282	221
Primary2	0.898	0.745	0.645	0.627	157	0.388	0.283	417
S5378	0.061	0.007	0.045	0.005	16	0.029	0.003	23
S9234	0.054	0.005	0.040	0.004	17	0.025	0.003	22
S13207	0.129	0.031	0.094	0.022	18	0.061	0.015	37
S15850	0.143	0.038	0.104	0.028	17	0.068	0.018	39
S38417	0.289	0.161	0.207	0.110	19	0.139	0.077	78
S38584	0.322	0.205	0.230	0.138	21	0.159	0.099	95
Dma	0.101	0.016	0.073	0.011	19	0.043	0.007	31
Dsp1	0.258	0.102	0.184	0.071	20	0.115	0.045	59
Dsp2	0.224	0.076	0.160	0.054	19	0.098	0.033	52
Risc1	0.407	0.262	0.288	0.160	22	0.190	0.122	24
Risc2	0.386	0.235	0.274	0.161	22	0.183	0.109	100
Comp.	–	–	1.00	1.00	–	0.64	0.63	–

7. ACKNOWLEDGMENT

We would like to thank Mr. H.-C. Wang, Mr. Y.-C. Cheng and Mr. R.-G. Liu in TSMC for providing precious suggestions related to this work.

8. REFERENCES

[1] H. Aoyana et al., "Applicability of extreme ultraviolet lithography to fabrication of half pitch 35nm interconnects," *Proc. SPIE 7636*, Feb. 2010.

[2] M. Chandhok et al., "Determination of the flare specification and methods to meet the CD control requirements for the 32 nm node using EUVL," *Proc. SPIE 5374*, Feb. 2004.

[3] Y. Chen, A. B. Kahng, G. Robins, and A. Zelikovsky, "Practical iterated fill synthesis for CMP uniformity," *Proc. DAC*, pp. 671–674, Jun. 2000.

[4] Y. Chen, A. B. Kahng, G. Robins, and A. Zelikovsky, "Closing the smoothness and uniformity gap in area fill synthesis," *Proc. ISPD*, pp. 137–142, Apr. 2002.

[5] T.-C. Chen, G.-W. Liao, and Y.-W. Chang, "Predictive formulae for OPC with applications to lithography-friendly routing," IEEE TCAD, Vol. 29, No. 1, pp. 40–50, Jan. 2010.

[6] H.-Y. Chen, S.-J. Chou, and Y.-W. Chang, "Density gradient minimization with coupling-constrained dummy fill for CMP control," *Proc. ISPD*, pp. 105–111, Mar. 2010.

[7] L. Deng, M. D. F. Wang, K.-Y. Chao, and H. Xiang, "Coupling-aware dummy metal insertion for lithography," *Proc. ASP-DAC*, pp. 13–18, Jan. 2007.

[8] FFTW3 (Fastest Fourier Transform in the West). http://www.fftw.org/.

[9] C. Krautschik, M. Ito, I. Nishiyama, and S. Okazaki, "Impact of EUV light scatter on CD control as a result of mask density changes," *Proc. SPIE 4688*, Mar. 2002.

[10] J. Lee et al., "A study of flare variation in extreme ultraviolet lithography for sub-22nm line and space pattern," *Jpn. J. Appl. Phys.*, pp. 06GD09, Jan. 2010.

[11] lp_solve 5.5.2.0. http://lpsolve.sourceforge.net/5.5/.

[12] A. Myers et al., "Experimental validation of full-field extreme ultraviolet lithography flare and shadowing corrections," J. Vac. Sci. Technol. B, Vol. 26, No. 6, pp. 2215–2219, Nov. 2008.

[13] F. Schellenberg et al., "Layout compensation for EUV flare," *Proc. SPIE 5751*, Mar. 2005.

[14] M. Shiraishi, T. Oshino, K. Murakami and H. Chiba, "Flare modeling and calculation for EUV optics," *Proc. SPIE*, Feb. 2010.

[15] V. K. Singh et al., "US patent 6,625,802: Method for modifying a chip layout to minimize within-die CD variations caused by flare variations in EUV lithography," 2003.

[16] V. Singh, "The importance of layout density control in semiconductor manufacturing," *Proc. EDPW*, pp. 70-74, 2003.

[17] O. Wood et al., "Integration of EUV Lithography in the fabrication of 22 nm node devices," *Proc. SPIE 7271*, Feb. 2009.

[18] H. Xiang, L. Deng, R. Puri, K.-Y. Chao, and M. D. F. Wang, "Dummy fill density analysis with coupling constraints," *Proc. ISPD*, pp. 3–9, Mar. 2007.

[19] C. Zuniga et al., "EUV flare and proximity modeling and model-based correction," *Proc. SPIE7969*, Feb. 2011.