2012

Critical Area Driven Dummy Fill Insertion to Improve Manufacturing Yield

Nishant Dhumane
University of Massachusetts Amherst

CRITICAL AREA DRIVEN DUMMY FILL INSERTION TO IMPROVE MANUFACTURING YIELD

A Thesis Presented

by

NISHANT DHUMANE

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
Of the requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL AND COMPUTER ENGINEERING

MAY 2012

Department of Electrical and Computer Engineering

CRITICAL AREA DRIVEN DUMMY FILL INSERTION TO IMPROVE MANUFACTURING YIELD

A Thesis Presented

by

NISHANT DHUMANE

Approved as to style and content by:

_____
SandipKundu, Chair

_____
Israel Koren, Member

_____
MaciejCiesielski, Member

_____
C.V. Hollot, Department Head
Department of Electrical and Computer Engineering

ABSTRACT

CRITICAL AREA DRIVEN DUMMY FILL INSERTION TO IMPROVE MANUFACTURING YIELD

MAY 2012

NISHANT DHUMANE, B.E., NAGPUR UNIVERSITY

M.S.E.C.E., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor SandipKundu

Non-planar surface may cause incorrect transfer of patterns during lithography. In today's IC manufacturing, chemical mechanical polishing (CMP) is used for topographical planarization. Since polish rates for metals and oxides are different, dummy metal fills in layout is used to minimize post-CMP thickness variability. Traditional metal fill solutions focus on satisfying density target determined by layout density analysis techniques. These solutions may potentially reduce yield by increasing probability of failure (POF) due to particulate defects and also impact design performance. Layout design solutions that minimize POF and also improve surface planarity via dummy fill insertions have competing requirements for line spacing. In this thesis, I present a formulation to balance these competing goals and provide a comparative study of greedy (or fixed spacing), variable spacing and LP formulation based fill insertions based on scalability and quality of solution.I extend the variable spacing fill to allow non-preferred direction routing of fill patterns in order to further improve the CA. Traditional fill solutions impact design performance due to increase coupling capacitance on signal nets. I present a fill insertion algorithm that minimizes this increase in coupling capacitance due to fill. Finally, I extend the critical area based solution to include SRAF insertion in order to account for optical diffraction in lithography.

Thus the proposed solution addresses both lithography and particulate related defects and minimizes the fill impact on design performance at the same time. Experimental results based on layout of ISCAS 85 benchmark circuits show that the variable spacing and the LP formulation based fill insertion techniques result in substantially reduced critical area while satisfying the layout density and uniformity criteria. The coupling capacitance minimization fill solution reduces the fill impact on coupling capacitance while at the same time minimizing the critical area.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

As the VLSI technology advances beyond 45nm, layout design in terms of interconnects that satisfy timing and physical design constraints have become increasingly complicated and difficult. Satisfying only these requirements is no longer sufficient for successfully working designs post-fabrication. As a result of this increased complexity and smaller layout geometries, vulnerability of the manufacturing process during critical dimension (CD) control, chemical-mechanical polishing (CMP) and lithography have led to an increased significance of the design for manufacturability (DFM) checks.

Chemical-Mechanical Polishing (CMP) refers to the topographical planarization of the oxide layers is one of the most important steps in manufacturing. Post-CMP local and global planarization depends on the layout pattern density. Non-uniform layout pattern can lead to uneven polished surfaces resulting from metal dishing or dielectric erosion as shown in Figure 1. This eventually results in a worsened lithography output. Such out-of-focus printed patterns can severely affect the performance and yield of the layout design.

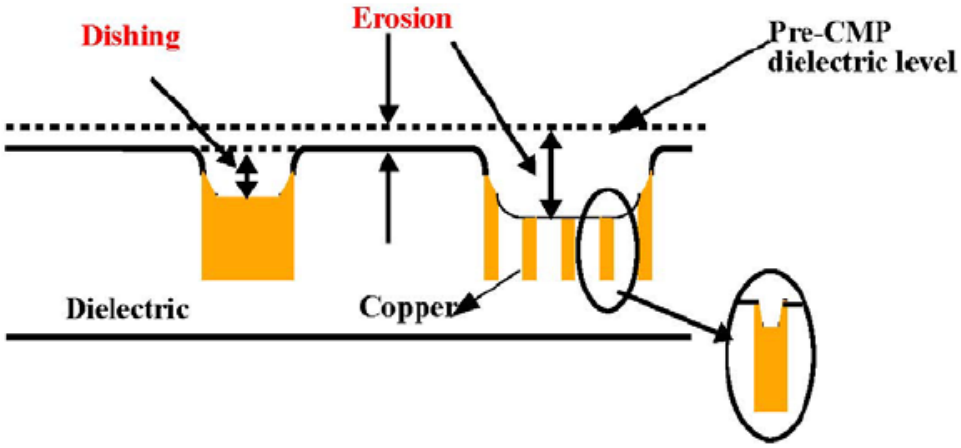


Figure 1. Metal dishing and dielectric erosion[22]

Also in recent technology, designs have high device density with multi-billion devices fabricated in less than 1 cm^2 area. The fabrication processes tend to use multiple levels of metal to support such high device density. As a result, the planarization and ILD thickness ofpolished surface of one layer has a cumulative effect on the planarization and ILD thickness subsequent higher layers [8]. Hence to ensure desired manufacturability output post-CMP process, dummy metal is inserted to maintain density uniformity in the layout.These dummy features are electrically isolated from the original layout features and insertion of dummy metal fill provides smoother surface planarity and better yield output as shown in Figure 2.

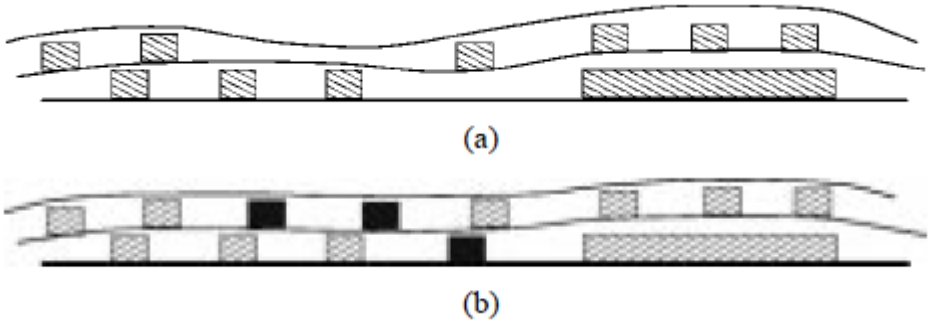
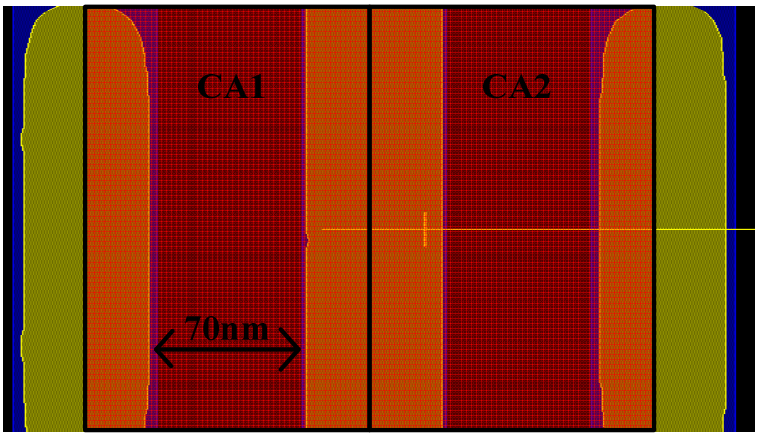


Figure 2. ILD planarization due to non-uniform layout density (a) before dummy metal fill insertion and (b) after dummy metal fill insertion (black tiles) [8]

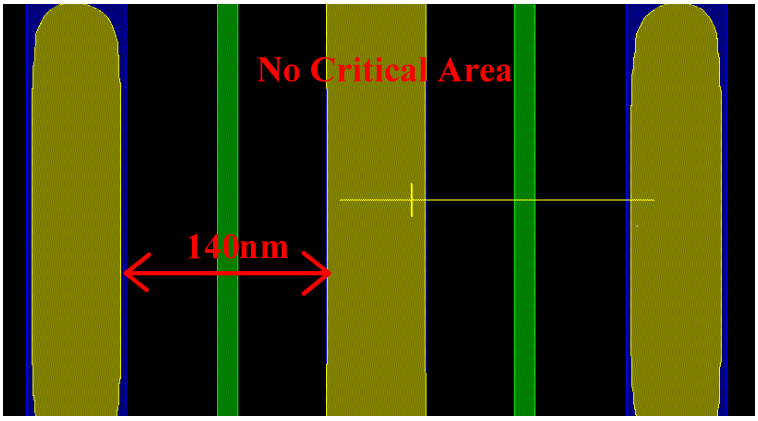
Methods for dummy metal fill insertion can be classified as: rule-based and model based. These methods are based on the fact that the layout should satisfy a density criteria for desired post-CMP output. This density criteria controls the amount of dummy fill to be inserted such that the layout density lies between allowed lower and upper bounds. Rule based approaches mainly focus on fill insertion in vacant layout spaces to satisfy the density rule for all overlapping or non-overlapping regions of a layout. Model-based methods on the other hand, rely on analytical expressions that are not necessarily just simplistic relations between the layout density and post-CMP output but also provide varying fill insertion solutions for surface planarity [8].

Intra-layer photolithographic defects arising due to the imperfections in the manufacturing process can not only result in circuit faults but also reduced yield [13]. Such spot

defects that lead to functional faults gives the probability of failure (POF). POF depends on the defect size and the area within which a defect must lie to cause a fault is known as the critical area (CA) for that defect size. Thus higher the critical area higher will be the POF. Techniques to calculate and analyze POF and the corresponding critical area have been long established [13]. Clearly maintaining uniform layout density for better printability and layout design for minimized spot defects play important roles in obtaining improved yield margins. Hence metal fill insertion focused on minimizing critical area inherently improves design yield. However they tend to have contradictory inter-feature spacing requirements. As shown in Figure 3(a), the two metal lines are spaced sufficiently to meet the physical design rules and the density bound for CMP. However, a particulate defect having size greater than their spacing can result in extra material being printed causing a functional failure. Thus placing the same metal lines with increased spacing can reduce the POF and help minimize critical area as shown in Figure 3(b). Usage of sub-resolution assist features (SRAFs) to improve the lithographic printability satisfies both the requirements.



(a) Part of critical area with dummy metal fill placed 70nm apart



(b) No critical area with dummy metal fill placed 140nm apart

Figure 3.CalibreWorkBench [20] simulation output. For a defect size of 140nm, the critical area is large (red colored rectangles) when the left and right dummy metal fills are placed 70nm from the middle rectangle in (a). No critical area is observed when they are moved 140nm away from the middle rectangle in (b). SRAFs added (green rectangles) maintain printability quality.

In this thesis, we present rule based and model based layout density analysis models which drive the various metal fill insertion techniques namely fixed spacing or greedy fill, variable spacing and LP formulation based fill insertion techniques. These techniques aim at minimizing the critical area while at the same time satisfying the layout density criteria. Based on the variable fill insertion technique, we implement a coupling capacitance minimization fill solution to minimize the coupling capacitance on critical nets.Finally we analyze and compare the critical area on the post-fill layouts and perform SRAF insertion for improved lithography output.

1.1 Thesis outline

The outline of the thesis is as follows: In chapter 2 we review some of the existing work in this area. In chapter 3 we describe the layout density analysis framework. In chapter 4 we discuss the various fill insertion techniques. We describe the algorithm to perform critical area analysis on the fill inserted layouts and the SRAF characterization and implementation in chapter

5. The experimental setup and results are presented in section 6. Chapter 7 concludes my thesis and includes the future work based on my thesis.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Layout Density Analysis

In chip designs, maintaining mask layout uniformity to minimize the variations during manufacturing has gained a lot of importance in lieu for achieving higher performance and yield benefits. In the previous section, we introduced the idea of how a non-uniform layout pattern can affect the CMP planarization output. Another problem with CMP is that the layout cannot have large stretches of metal or nonmetal regions. In order to avoid large portions of layout occupied by metal or large portions of it being nonmetal regions, foundries usually require an effective metal density to be satisfied across a layout. These metal density constraints are governed by specified minimum and maximum pattern-density values. To satisfy these density rules dummy metal fills are inserted into the layout to raise the layout density distribution across the layout meeting minimum layout density requirement and ensuring layout uniformity [10].

The layout pattern density over a region is defined as the ratio of the portion of region occupied by rectangular metal shapes to the total area of that region. Usually the layout pattern density range is between 30% and 70% [9] [19] [23]. Apart from CMP, the layout pattern density is also important in other applications. In optical lithography, resolution enhancement techniques (RETs) such as optical proximity correction (OPC) and off-axis illumination (OAI) are employed to improve printability of layout patterns. Issues such as increased mask design cost in OPC and the forbidden pitch observed in OAI reduce their effectiveness [12]. In 45nm and lower technologies with tighter critical dimension budgets, scattered light from the lens and other optical sources have a considerable effect on the quality of the lithography process output [9]. Given such stringent process requirements, these secondary effects need to be properly accounted for. Maintaining uniform layout density helps minimize these effects and maximizes the yield output.

2.1.1 Previous Work:

In the past, various approaches have been presented on layout density analysis [3] [6] [8]. Principally, these techniques referred to as fixed-dissection regime scan the layout by breaking it into grids of smaller rectangular or square boxes and analyzing the density over several overlapping or non-overlapping windows [3] [4] [5] [6] [7]. However these approaches do not check possible windows off the grid. This can potentially result in sub-optimal evaluation of density and the density bounds. Improvements discussed by Kahng et al. adopt a sliding window technique that eventually reports an extremal window from the layout. However this technique depends on the number of rectangles in a given region and involves re-computation of density for every window [3]. This makes the approach computationally expensive and not easily scalable for bigger layouts. Kahng et al. describe a recursive approach to find the extremal window depending on the number of rectangles contained in the tiles(r) [4] [7]. However this kind of recursive analysis and fill is suited only for deterministic fill requirement. For the problem at hand, recursion might prove to be ineffective and more time consuming than useful for dense layouts.

Chen et al. discuss a master-cell based hierarchical fill insertion approach where fill geometries are added in identical fashion to multiple copies of the same master-cell across the layout. This results in an increase in the data volume and additional number of constraints and variables due master-cell overlaps etc render this approach of scan and fill computationally expensive for an LP formulation under the current required framework [11]. Mukherjee and Chakraborty present a pixel based randomized greedy fill algorithm that performs a grid less layout density analysis [9]. However the pixel based filling constrains the layout scan to the pixelated array limiting the moving window scan. The layout density convergence given for sufficient moving window sizes does not clearly indicate the satisfaction of uniform metal density across the layout.In the early model based density analysis schemes discussed by authors, the ILD thickness predictions are good for length scales which are in few millimeters. However

for sub-millimeter range designs these models do not accurately predict the ILD thickness variation [24].

In the proposed work, we present two techniques for layout density analysis namely rule based and model based. These techniques focus on addressing the above issues such that

1) the density of any minimum sized partition (M) is computed only once

2) these are scalable to any layout following a simple partitioning given the size of M

3) our approaches make it easy to couple either of the density analysis techniques to any filler insertion approach be it greedy or variable spacing

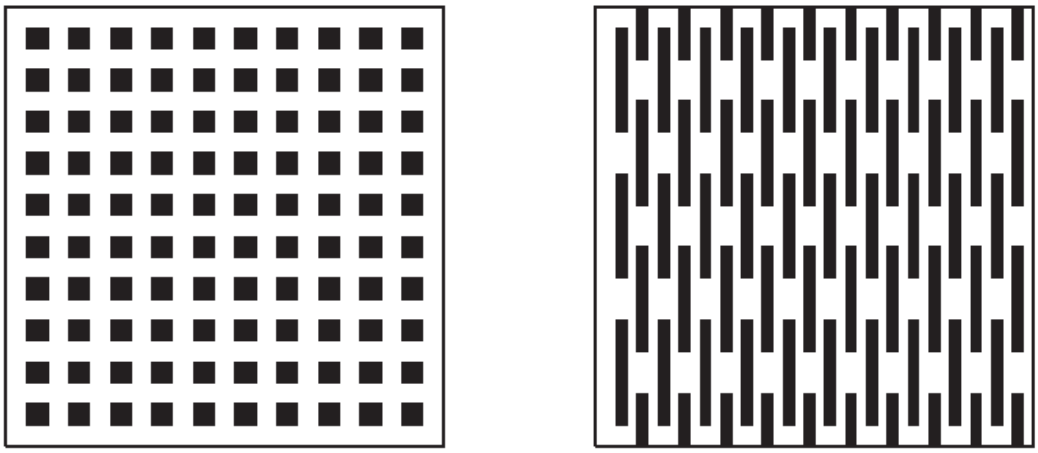
4) minimal number of window scans per M is required to conform to the density bound at the local partition level (surrounding partitions) in the rule based density analysis model as compared to the scanning techniques discussed earlier

5) the model-based solution makes use of the Lorentzian kernel suggested by Urbach and Rhezak [24] to accurately predict the ILD thickness variation for smaller technologies.

2.2Metal Fill Insertion

After the layout density scan and analysis is done to determine the regions of the layout violating density rules and the amount of fill required in each region, metal fill insertion is carried out. Traditionally metal fill geometries were inserted following a specified set of dimension and patterns. Figure 4 shows a symmetric arrangement of various fill patterns that have identical coupling capacitance to the adjacent long conductors in the layout [6]. Other techniques exist where the layout for fill insertion is pixilated and each pixel has the minimum allowed dimensions of the corresponding metal layer. Thus the presence of fill in a region is determined by whether the layout pixel is turned on or off. Fill geometries can also be classified

as grounded fill or floating fill. Grounded fill geometries are more robust and offer more predictability in aggressively timed designs as compared to the floating fill. Grounded fill metal regions are at known potentials and are easier to extract. These are more suited for microprocessor like designs. Floating fill on the other hand can be used in ASIC designs where timing is comparatively less critical as long floating fills can be potentially cause coupling effects.



(a) (b)

Figure 4. Examples of regular fill patterns. (a) 1 x 1 squares separated 1 unit apart and (b) 10 x 1 rectangles separated 1 unit apart [6]

2.2.1 Previous Work:

Various metal fill techniques exist which mainly aim at satisfying the density bound criteria and the min-var and min-fill [6] objectives. Kahng et al. discuss tile based metal fill insertion in symmetric regular or skewed grids that have varying effect on coupling to long conductors [5] [6]. Tian et al. use a linear programming (LP) formulation to insert fill once the amount of fill required and the spatial fill opportunities are determined using Boolean and other deterministic operations [8]. Mukherjee and Chakraborty propose a random greedy fill approach that inserts metal tiles on a pixel-by-pixel basis [9]. However the metal fill inserted in these fill

approaches, barely satisfies the physical design rules and does not involve any additional goal for further enhancing the performance or yield.

With increasing layout complexities and DFM requirements, metal fill insertion is implemented not only to satisfy the layout density criteria but is also targeted to improve additional yield objectives. For example, Deng et al. propose coupling aware fill insertion technique that minimizes coupling capacitance due to metal fill [12]. In their solution, sub-resolution assist features are used to replace the printable dummy features to minimize the coupling capacitance and at the same time maintain the lithography printability as show in figure 5. In recent technologies, uniform metal density layouts are strongly preferred. Thus the overlap capacitance due to upper/lower layers on metal features becomes significant (about 40% - 60%) of the total net capacitance. It becomes important forperformance aware metal fill solutions to consider overlap capacitance in order to obtain an effective improvement in the net delay.The fill solution proposed by Stine et. al [31] uses small rectangular floating fill geometries with increased spacing to minimize coupling capacitance. However it does not consider the impact of overlap capacitance and the capacitance comparison metric does not clearly indicate any possible improvement in the net delay values.

These approaches may result in reduced yield due to probability of failure from particulate defects. In the various approaches for dummy fill insertion proposed in this work, we heuristically utilize the available layout space for fill insertion to minimize the critical area.Also we extend the variable spacing fill approach to implement a coupling capacitance minimization technique that focuses on the above issues.

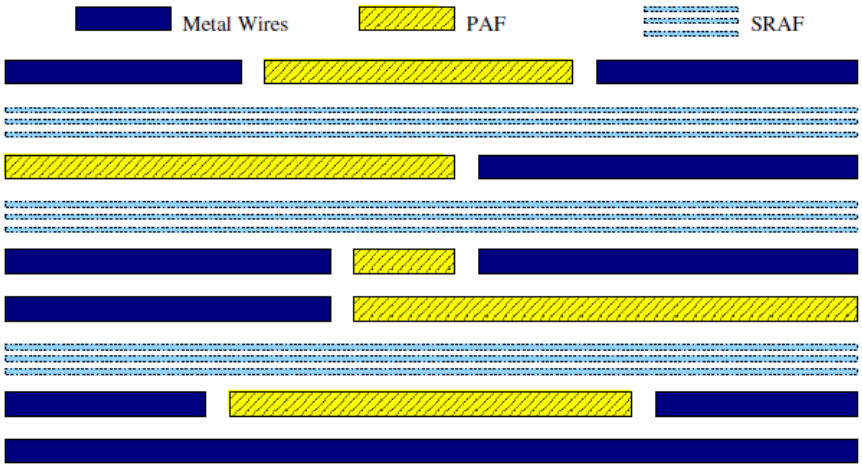


Figure 5.Coupling-aware metal fill insertion using regular dummy features (PAF) and SRAFs

2.3 Critical Area Analysis

Imperfections in the fabrication process result in yield-reducing manufacturing defects. The severity of these losses grows proportionally as the chip area and the device density [13]. These manufacturing defects can be classified as global defects and spot defects. Global defects arise as a result of mis-handling of wafers, over or under etching, mask misalignment etc. These defects can be minimized with well controlled manufacturing facility. Spot defects on the other hand, are random in nature mostly arising out of some chemical and airborne particles from the materials used in the process. Controlling these random spot defects is difficult as compared to global defects. Hence the amount of yield loss is also more [13]. Moreover the global defects are independent of the size of the chip while the spot defects increases with the chip size. Thus these are of greater importance in order to minimize the yield losses. For the purpose of our work, we focus on spot defects.

Intra-layer photolithographic defects arising due to the imperfections in the manufacturing process can not only result in circuit faults but also reduced yield [13]. Spot defects can result in extra material being printed (shorts) or missing patterns (opens) depending on the location of these defects. A defect that results in either a short or an open in the layout is

known as a structural fault or a physical defect. Figure 6 shows a scenario with three spot defects. However an open structural fault will occur only at the bottom defect location.

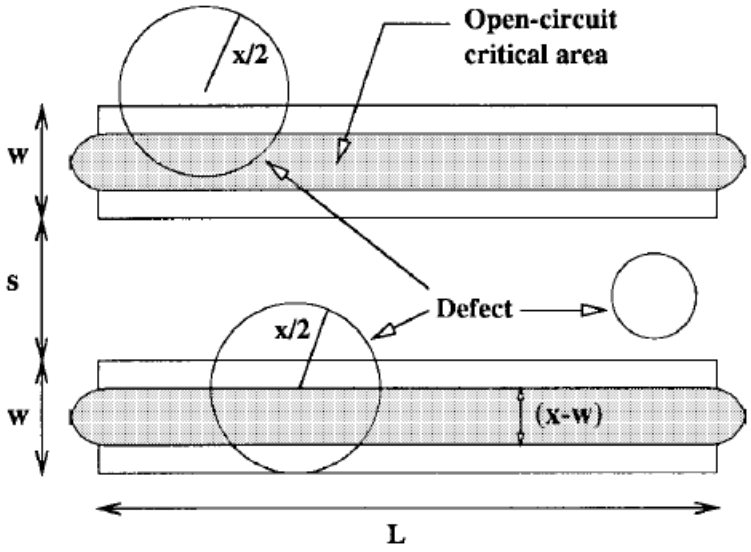


Figure 6. Critical area for two conductors for the metal defects [13]

Such spot defects are random in nature and the fraction of defects that lead to functional faults gives the probability of failure (POF). POF depends on the defect size. The area within which a defect must lie to cause a fault is known as the critical area (CA) for that defect size. Figure 6 shows the critical area for the two conductors for missing-metal defects. Thus higher the critical area higher will be the POF. It acts as a proxy to compute or analyze the POF due to particulate defects in a simple manner. There exist several methods of calculating the POF and critical area. For example, Monte Carlo type methods directly calculate the POF given a defect distribution while geometry based approaches calculate the critical area first [13]. We use the latter approach and calculate the POF for a defect size of x as follows:

$$\theta(x) = \frac{A(x)}{Achip}$$

where*Achip* is the total area of the layout. Averaging the critical area over the entire defect distribution gives the average POF for all defect sizes as

$$\theta = \frac{A}{Achip}$$

2.3.1 Previous Work:

Several models have been proposed for critical area computation. The pattern recognition technique proposed by Mattick et al. [16] and the layout expansion technique to perform fast CA calculation [15] use spatial overlap concept. The pattern match technique identifies a unique set of patterning rectangles (PRs) for which the critical area is already computed. These PRs are matched to various features present on the layout to obtain the total CA. The layout expansion technique suggested by Xue et al. expands the layout rectangles by defect_size/2 to obtain possible overlaps with the neighboring rectangles [15]. The overlapping areas so found give the critical area for the respective defect size. The authors propose an extra corner stitching plane (ECSP) structure to solve multiple overlaps during multi-level CA computation. However the ECSP structure can be complex for complex and multiple overlap patterns. They also suggest using pair-wise rectangle tagging to avoid the same rectangle being considered multiple times. For large layouts the pair-wise tagging of rectangles can be inefficient and unnecessary storage of data is required. In the proposed work, we present a pixel based critical area calculation that avoids the need to have an ECSP like structure. We use a forward linking rectangle list on a per partition basis that provides a simple and efficient approach to calculate the critical area and minimize data storage.

CHAPTER 3

LAYOUT DENSITY ANALYSIS

In chapter 2 we introduced the concept of layout density analysis followed by a brief discussion on some of the existing techniques to perform the density analysis of a given layout. In this chapter, we describetwo different implementations of layout density analysis models. These models primarily differ in their approach towards calculating the amount of fill to be inserted in the layout and subsequently driving the fill insertion algorithm to achieve the desired layout density.

These approaches are mainly classified into rule based and model based fill insertion algorithms. As discussed earlier, the foundry provides a density bound for metal densities on each of the various layers in order to have acceptable ILD thickness across the layout design. Based on this, the rule based fill insertion algorithms involve adding dummy metal features to satisfy a pre-determined density value within the bound. These techniques make use of Boolean operations [8] or involve greedily filling the layout space. Apart from meeting the physical design rules and satisfying the density bound, these approaches do not involve any optimization to improve the quality of the final fill solution. This leads to unevenly filled layout regions and subsequently result into uneven ILD thickness across the layout.

On the other hand, model based approaches rely on analytical expressions that relate layout density and ILD thickness and provide local layout pattern density dependent fill solutions. In other words the fill solutions varies in the way fill rectangles are inserted depending on the surrounding layout density. It has enough fill rectangles inserted such that apart from meeting the physical design rules and satisfying the density bound, the solution also results in a uniform metal density layout with minimized ILD thickness variation across the layout. Such

smoother metal surface topographies are highly desirable for modern design with multilayer metal stacks.

3.1. Rule based density analysis model

Initially a layout file containing the mask descriptions of all metal layers is read. Of these mask layer descriptions, we can choose any layer on which we wish to perform density analysis. The other layers may be simulated similarly. The features on the mask are read into a data structure which divides the mask into contours and rectangles. The data structure partitions the layout into an array of minimum sized partition (M) as shown in Figure 7. Figure 8 shows the pseudo-code for performing layout partitioning and density scan. Partitioning the layout decomposes the layout density computation problem into smaller sub-problems providing ease of computation and modularity. Density of each M is computed simultaneously during layout partitioning as indicated in lines 1 and 2. The density of any partition is obtained by calculating the ratio of the amount of area occupied by rectangles lying within the partition to the total area of the partition. This allows us to have an early knowledge of the partitions that violate the lower density threshold as required to be satisfied for effective CMP [5] [9] [19]. As shown in line 3, we focus our fill objective only on partitions that violate the lower bound (LB) of the density criteria to minimize the cost of inserting additional metal fill in terms of the critical area while satisfying the density criteria [19]. Hence, the partitions which have densities greater than LB are not considered for additional fill insertion.

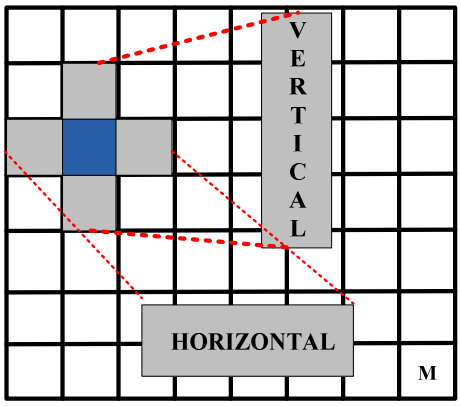


Figure 7.Density scan for a partition. Vertical and horizontal boxes shown as projection

Layout Density Scan & Analysis

INPUT: Layout , Min. sized partition (M) size , Min. density bound (LB)

OUTPUT: Filled Layout for M (if dummy fill is inserted)

```
1.  Decompose the layout into partitions each of size M
2.  Simultaneously scan the layout and populate list of violating
    partitions
3.  foreach violating partition M
4.      obtain neighboring partitions (Mn);
5.      foreach violating partition from (M + Mn)
6.          do:
7.              // perform fill insertion
8.          until (Density <= LB)
9.      perform uniform density checks on ver. & hor. boxes;
10.     redistribute & repeat for (M + Mn) if (density < LB)
11.     update densities for each (M + Mn);
12. Perform uniform density check for the entire layout
```

Figure 8. Pseudo-code for layout density scan and analysis

Once the list of all violating partitions is obtained, each partition from the list undergoes a density scan. As shown in Figure 7 the violating partition is scanned along with its orthogonal neighboring partitions (as in lines 4 & 5). After the neighboring partitions are identified and their pre-computed densities are obtained, fill insertion is performed in the central violating partition. At this stage, fill insertion is also performed in any neighboring partition if it is found to violate the density limit. This ensures that all the partitions satisfy the limit and their corresponding densities are updated to avoid any redundant fill insertion later on. After all partitions satisfy the density limit, a density check is performed on the vertical and horizontal boxes formed by merging upper-central-lower and left-central-right partitions. If any of these two boxes is found to violate the density limit, the additional amount of fill required to satisfy the density target

iscalculated and re-distributed over the individual partitions for fill insertion (as in line 9 & 10). This is repeated until the vertical and horizontal boxes satisfy the density limit. This ensures that the newly added dummy fill in one or more partitions under consideration, helps not only to meet the density target but also to maintain uniformity in layout density at a local level with respect to the violating partition.

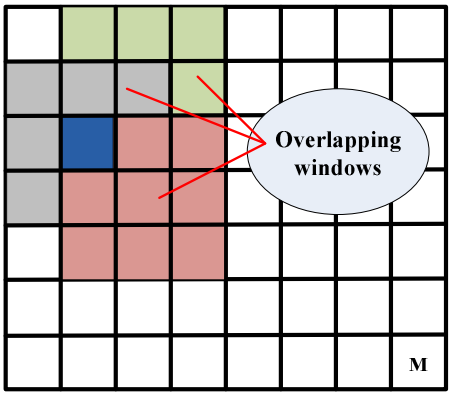


Figure 9. Multiple overlapping rectangular windows to scan a partition

From this discussion it is clear that any violating partition is required to be filled only once and also lesser number of scans per partition is required as compared to other rectangular window based scan and fill techniques discussed earlier. For example, Figure 9 shows a partition to be filled using a rectangular window scan format, checked using multiple overlapping windows to make sure that there is no window which violates after metal fill. After scanning and filling the partitions we perform density uniformity check for the entire layout (as in line 12). The following section describes in detail the approach and its proof for obtaining a uniformly filled layout.

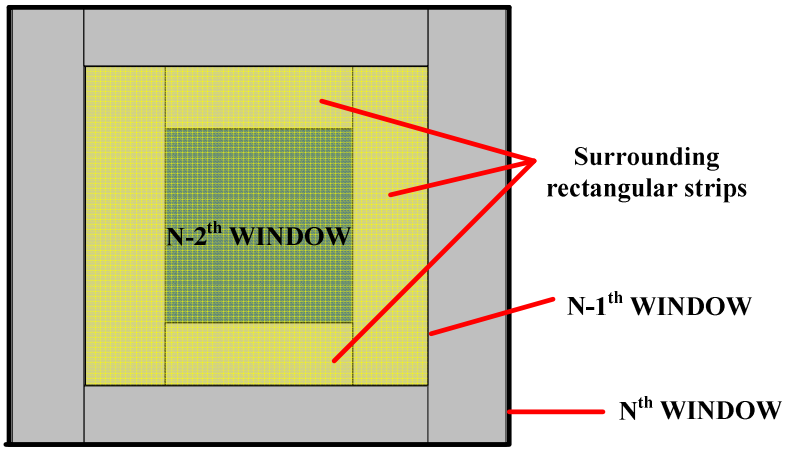


Figure 10: Concentrically growing windows (N-2, N-1 and N) and surrounding rectangular strips used for layout density uniformity check

3.2. Observation on layout density uniformity

Given a partitioned layout with all the partitions filled in the manner discussed in the earlier section, we observe that by using mathematical induction the entire layout conforms to the layout density criteria. The problem involves calculating the density of concentrically growing windows such that after *N* iterations, the window so formed is equivalent to the layout and it satisfies the density criteria for CMP.

In order to prove that the post-fill layout satisfies the density target, we start with a heuristically chosen rectangular window and calculate its density. At this point, we make sure that the window density satisfies the density target.This window forms the base case for the inductive proof (say N = 1 iteration)

Consider that after k^{th} iteration (N = k), we have a window that comprises of a central window surrounded by rectangular strips on its sides. An example for this is shown in Figure 10. To avoid repetitive density computation, we re-use the density value of the previously considered window *(N-k-1)*. Thus any stage involves only the density calculation for the surrounding stripes.Ateach step, the size of the surrounding strips chosen is small as compared to the

corresponding dimension for the central rectangular window. At this stage, let us assume that the window (N = k) density meets the density target.

In order to verify the density value for the next bigger window, we first obtain the surrounding strips with respect to k^{th} window. As a result, in this iteration we are actually calculating the density value for the N = $k+1^{th}$window.We assume that each of the stripes has sufficientpattern density and that it satisfies the target. However it should be noted that during this procedure, metal fill is inserted in any of the rectangular stripes if it is found to violate the density limit. Thus with the central window and all the surrounding strips satisfying the density target, we can say that the N = $K+1^{th}$ window satisfies the density target.

Therefore by using the principle of mathematical induction, we can safely show that any window *N*(say layout) will satisfy the density criteria if the previous window *(N-1)* and the surrounding rectangular strips that form the window *N* also satisfy the density criteria.

3.3. Model based layout density analysis model

The closed form analytical expression proposed by Stine et al. relates the ILD thickness variation post CMP to the pattern density across the layout [25] [26]. It is based on Preston's equation and is given by the following expression:

$$z = z_0 - \left(Kt \,|\rho_0(x,y)\right) \qquad Kt < \rho_0(x,y).z_1$$

$$z = z_0 - z_1 - Kt + \rho_0(x,y).z_1 \qquad Kt > \rho_0(x,y).z_1$$

where, Z is the ILD thickness at location (x,y) as shown in the figure below, Z_0 is the amount of dielectric before polishing, Z_1 is the initial step height, K is the blanket oxide removal rate, t is the polish time andρ_0 is the effective pattern density. Also all these values are constant for a specific CMP process. The above expression implies that if we polish the surface for sufficiently long time, the final ILD thickness is then directly related to the layout pattern density [26]. If the polish time is chosen to be t >(ρ_0 Z_1 / K), so from the above expression the final ILD

thickness is between 0 and (Z_0 – Z_1) and we say that local planarization at location (x,y) is achieved.

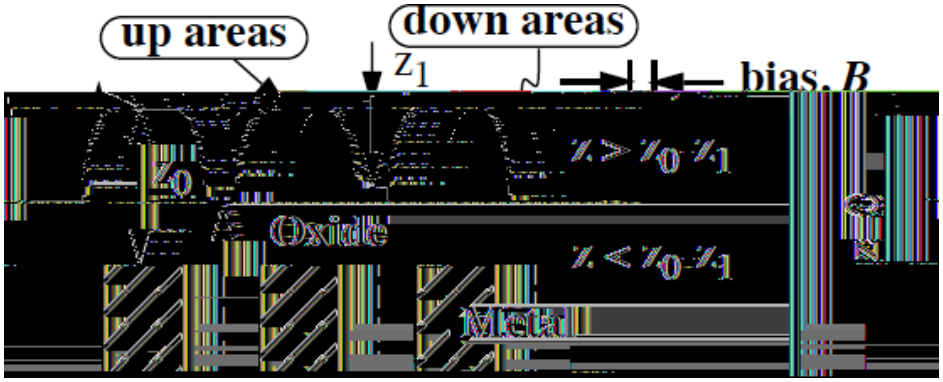


Figure 11. Variables used in the model [29]

Simplistic models have been proposed to achieve minimum ILD thickness variation have been proposed. Kahnget. al propose a min-variation based solution working on a fixed dissection regime that focuses on local pattern density. Thus the objective of minimizing the local pattern density variation translates into minimizing the ILD thickness [6]. This model however, does not consider the impact of polishing pad deformation on the ILD thickness resulting from the variation in pattern density. Oumaet. al consider the impact of polishing pad while calculating the oxide density [28]. The final local oxide density is calculated by adding the weighted local pattern densities within a weighing region and is then called as the effective pattern density from averaging [8]. If the local pattern density for each of the regions in a fixed dissection regime is given by $d(n_1, n_2)$ and the discretized weighing function by $f(n_1, n_2)$, then the effective discretized density is given by the convolution sum [28]:

$$\rho_0\,(n_1, n_2) = \sum_{k1=-\infty}^{\infty} \sum_{k2=-\infty}^{\infty} d(n_1, n_2) f(n_1 - k_1)(n_2 - k_2)$$

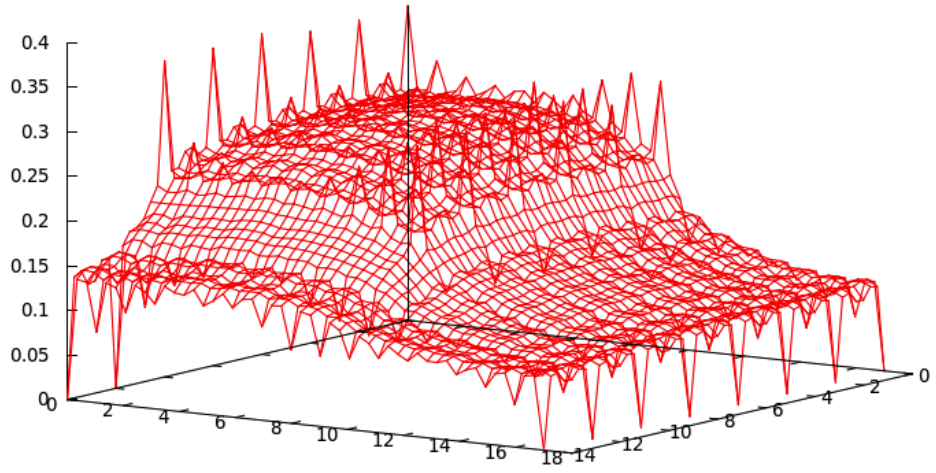
Zero padding is used to have the number of discretized regions in powers of 2, which helps to perform the computation in frequency domain using FFT techniques [8] [28]. The convolution sum of discretized regions and weighing function assumes that die is repeated across

the wafer and as a result local pattern densities near the die edges are affected by the layout patterns on the neighboring dies. By controlling the indices of these regions near the die edges, the convolution sum can be obtained. However, it is not required if we are considering only a single die. The weighing function *f*that models the behavior and impact of polishing pad in oxide thickness is taken to be an elliptic function in some formulations [8] [28]. The size and width of the weighing function is determined using the interaction distance or the planarization length. The interaction distance is defined as the length at which the relative weight of the function *f* becomes negligible. Typical values of the interaction distance are reported in few millimeters characterized from elastic properties of the polishing pad and other CMP process parameters.The model proposed by Wong et. al works well on the scale of the planarization length however, it is not accurate enough to predict the ILD thickness variations on the scales of the more recent technology nodes [24].Divechaet. aldiscuss the impact of surrounding layout features on the effective oxide pattern density [27]. They show that the dependency of effective density on metal pitch of surrounding layout features decreases for larger pitches. The smaller features exhibit substantial lateral deposition of the oxide, resulting in a smoother topography as compared to the larger features which more conformal deposition. Thus for smaller features sizes and narrower metal pitches pertaining to the more recent technologies, we apply appropriate biasing B as shown in the figure above. Biasing helps in averaging the effect of oxide deposition on original layout features while computing the oxide pattern density. We approximate the bias value B to be 20% of the feature size, so that all layout features are enlarged or shrunk from all sides by value B. As a result in order to accurately predict the oxide thickness for sub-millimeter ranges, we use Lorentzian kernel in the convolution sum given by the formula [24]:

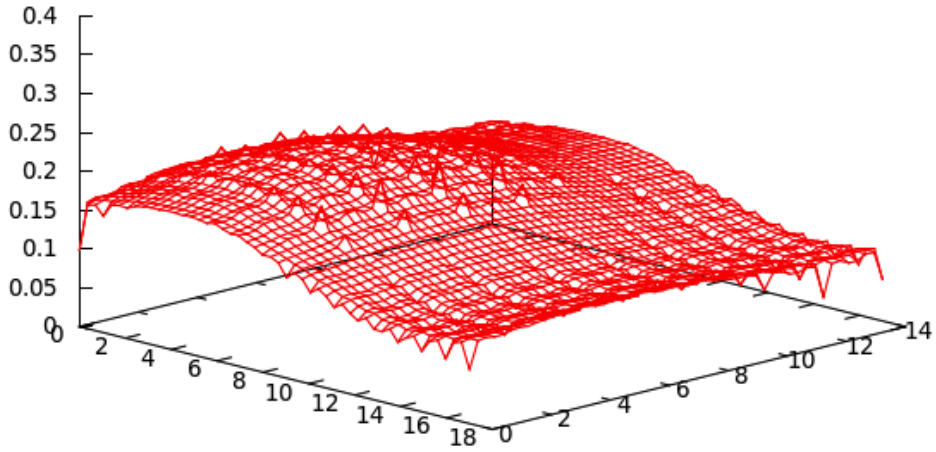
$$f(x_{n1,n2}, t) = \frac{1}{\pi} \frac{bt/2\pi}{x^2 + (bt/2\pi)^2}$$

This model of the weighing function *f* not only incorporates the polishing pad behavior but also takes into account the impact of the features on local pattern densities. The variable b is

constant for a given process and considering a polish time of 90sec [24], we obtain the full width at half maximum (FWHM) of the Lorentzian i.e. the planarization length to be approximately 28μm.



(a) Initial density plot for a partitioned layout



(b) Effective density plot using Lorentzian kernel

Figure 12. Post CMP topographies showing density (in Z direction), X, Y units are partition numbers scaled in μm, (a) Initial density plot for partitioned layout, (b) Effective density plot using Lorentzian kernel

The layout is thus partitioned such that the number of partitions in each Lorentzian kernel of size 28 sq.um is in powers of 2 for the convolution sum. Figure 12 (a) shows the intial density plot

for a partitioned test layout and (b) the effective density computed for all partitions using the Lorentzian kernel.

Once the layout is partitioned and the initial density of all partitions is determined as discussed in the earlier sections, we populate a list of violating partitions. Unlike the predetermined density cut-off of the rule based density analysis method, in this approach we first prepare a density based histogram of all partitions and then choose a minimum density cut off from the density space occupied by maximum number of partitions. This approach towards determining the density cut off helps in achieving a tradeoff between the ILD thickness variation across the layout and critical area. It avoids overfilling the layout to obtain a perfectly smooth surface and results in a better average ILD thickness while at the same time the critical area is minimized. Once the list of violating partitions is determined using the density cut-off, we apply the aboveLorentzian kernel to each of the violating partitionsand perform fill insertion such that the target effective density is achieved. Here instead of having an absolute min-variation objective while filling, we formulate a ranged variation approach similar to [8] where the fill objective is to achieve a target effective density within a range ε. The ranged variation helps in controlling the amount of fill to be inserted in order to satisfy the target density. Thus while computing the effective density of any violating partition, all violating partitions in the kernel window are filled up to a weighted tolerance limit of ε' with respect to the target density such that target effective density for the central partition is within ε of the target density. The density analysis approach discussed in this section is independent of the fill insertion methodology and hence provides flexibility in choosing and optimizing the design using appropriate fill insertion objective.

CHAPTER 4

DUMMY FILL INSERTION TECHNIQUES

In order to obtain higher yield benefit during dummy metal insertion, we heuristically utilize the inter-feature spacing to minimize the critical area. Given a partition M required to be filled, we first obtain a bounding box(R) that encloses all rectangles lying within M (see lines 3 to 5 in Figure 5). After fill is inserted in R using either of the fill insertion techniques described in this section, we concentrically grow this box and continue adding fill until it meets the required density limit (T). This is done in order have a uniform and regular metal pattern in the regions occupied by the rectangles. This follows the density uniformity observation by growing concentrically outwards from the centroid of the space occupied by rectangles. It is worth noting that even though the fill is inserted within the rectangle bounding box, the inter-feature spacing applied minimizes the critical area. The manner in which this inter-feature spacing is applied is what differentiates the fill insertion techniques. We discuss these techniques in detail along with the help of the pseudo-code in Figure 11.

4.1 Greedy (or fixed spacing) fill insertion technique

Before we start fill insertion, the bounding box and the amount of metal fill required to satisfy density limit are known. The space constraint (or space multiplier N) controls the inter-feature spacing and in turn minimizes critical area. It can be obtained from the defect size distribution or in multiples of minimum inter-feature spacing (S) from the design rules. In the greedy fill approach, the space constraint *min_space_req* is fixed at lower multiples of S as in line 5. Here we do not iterate over the inter-feature spacing multiplier (N), hence it becomes a greedy heuristic approach. In lines 6 to 8, the heuristic algorithm performs fill insertion in the

respective metal direction in a greedy manner at the first opportunity available that satisfies the spacing rule.

The process of fill addition continues until either the density limit is satisfied or the bounding box limit is reached (line 10). In the latter case (in lines 11 and 12), the next bigger sized box is taken for fill insertion. In the worst case, if the minimum sized partition is reached and density is still lower than the required limit, then using density redistribution method described earlier fill insertion is repeated (as in lines 13 to 15). Considering the marginal amount of additional fill required the density redistribution method ensures that density criteria is satisfied. However a fixed spacing constraint results in a limited critical area minimization. This serves as the motivation for variable spacing fill insertion technique described in the following subsection.

Metal Fill Insertion Technique

INPUT: Min. sized partition (M), Spacing multiplier (N), Target density (T)
OUTPUT: Filled Layout for M

```
1.  foreach violating partition M
2.    obtain neighboring partitions (Mn);
3.      foreach partition to fill from (M + Mn)
4.        do:
5.          obtain rectangle box (R) &  min_space_req = (N . S);
6.          do:
7.            start fill insertion from centroid of R moving outward;
8.            add fill shape if space available >= min_space_req;
9.            iterate over decreasing N for variable spacing
              technique;
10.         until (Density = T  or  box R limit reached)
11.         grow R concentrically;
12.       until ( R equals M )
13.     perform density checks (as in section III);
14.     redistribute & repeat for (M + Mn) if  (density < T)
15.     update densities for each (M + Mn);
```

Figure 13.Pseudo-code for fixed and variable spacing metal fill insertion techniques

4.2. Variable spacing fill insertion technique

In this technique, the fill insertion starts with a tighter space constraint *(min_space_req)* as compared to the greedy fill method. Initially (in line 5) the space constraint is chosen to be in higher multiples of the minimum feature spacing (*N.S*). Centered on the centroid of empty space, fill is inserted until it reaches margin of spacing constraint. Occasionally, this is insufficient to meet the density target. In such cases, the space constraint is relaxed iteratively and fill insertion is performed (in line 9). This space relaxation is in steps of integral multiples (N) of S. The iteration continues until the density target is met (in lines 10 to 15). At all times, the space margins are set to be highest possible under density constraints, allowing critical area to be minimized.

The varying space constraint is what differentiates this method from the greedy fill technique discussed earlier (line 9). The main advantage of spatially over constraining is thatit always provides more effective space utilization ascompared to the greedy technique.

4.3. Fill insertion using Linear Programming

The fill insertion problem using linear programming (LP) approach provides another perspective as to how the available layout space can be effectively utilized. As compared to the LP techniques [4] [5] [6] [8] implemented in the past that deal with optimizing fill insertion in order to satisfy the layout density bound, the proposed LP technique works on optimizing a dual objective function. The LP problem formulated in [5] [6] pre-computes the fill opportunities and these are fed in the form as constraints to the LP solver. Under the current problem at hand, such formulations would result in excessive number of constraints and variables making it intractable for the suggested window sizes in [6]. Hence in the proposed LP approach, we break the minimum sized partition M into a grid of pixels. The pixels have a binary declaration indicating the presence of metal. The pixel size can be chosen using either minimum width/spacing dimensions or their multiples.

Assuming the grid to have dimensions *X, Y*, we scan the grid using *w* x *w* windows where $w \leq X, Y$. Pixels which are already filled at the time of problem formulation are denoted by *pre_filled_pixels*. Constraints for pre-filled pixels are generated indicating the space occupied by the rectangles lying within the window. The number of pixels that need to be filled in order to satisfy the density limit is denoted by *total_pixels_req*. Thus knowing the *total_pixels_req* and *pre_filled_pixels*, the number of pixels that need to be filled can be easily determined as shown in constraint (2). Maximizing the distance between the numbers of pixels to be filled to satisfy the density limit forms the basis for our LP problem's objective function. The objective function can be generated by calculating the pair wise distances between pixels a_{ij} and a_{kl} that are turned *on* as indicated by variable Z_{ijkl}.

For sufficiently large window sizes, it is observed that generating *Z* variables for pixel pairs which are farther than 4 or more pitches apart does not improve the maximized output greatly. Hence for a pixel i,*j*, only pixels *k,l* falling within 4 pixel distance apart are considered as shown in constraints (1). Similarly turning *on* or *off* of pixels is also governed by the constraints generated to optimize the spacing between any pair of pixels turned *on*. These constraints are generated following the same reasoning, as shown in (3). These help satisfy the minimum physical design rules. However, a bigger window size can potentially result in a sub-optimal solution for dense layouts. Hence for space constraints generation, a separation of 2 or 3 pixel distances is considered. Spacing constraints are also generated for pixels lying along the edges of the sliding window, as shown in constraint (4). This takes into consideration pixels surrounding the sliding window and helps avoid closely placed metal features near abutting window boundaries. Note that since we are trying to maximize spacing between the *on* pixels, best results are obtained by choosing window sizes and spacings comparable to the defect size range considered. Thus with this objective function and constraints, the LP problem can be written as:

Maximize: $\sum_{i,j} Z_{ij\_kl} * d_{ij\_kl}$

Subject to:

$$\left(1 - Z_{ij\_kl}\right) + a_{ij} \geq 1 \quad (1)$$

$$\left(1 - Z_{ij\_kl}\right) + a_{kl} \geq 1$$

$$Z_{ij\_kl} + \left(1 - a_{ij}\right) + \left(1 - a_{kl}\right) \geq 1$$

$$\sum_{i,j} a_{ij} = Total\_required\_pixels - pre\_filled\_pixels \quad (2)$$

$$C1 * \left(1 - a_{ij}\right) - \left(1 - \sum_{p.q} neighboring\_pixels\right) \geq 1 (3)$$

$$C2 * \left(1 - a_{ij}\right) - \left(1 - \sum_{s,t} boundary\_pixels\right) \geq 1 \quad (4)$$

where $i = 1, \ldots, X$ *and* $j = 1, \ldots, Y$

$k = i-4, \ldots, i+4$ *and* $l = j-4, \ldots, j+4$

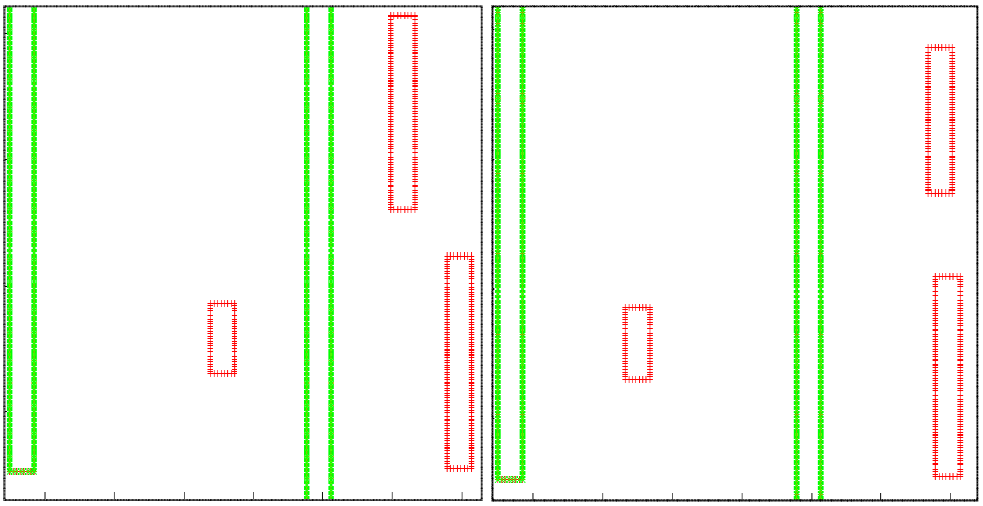
$p,s = i-2, \ldots, i+2$ *and* $q,t = j-2, \ldots, j+2$

$p,k,s \neq i$ *and* $q,l,t \neq j$

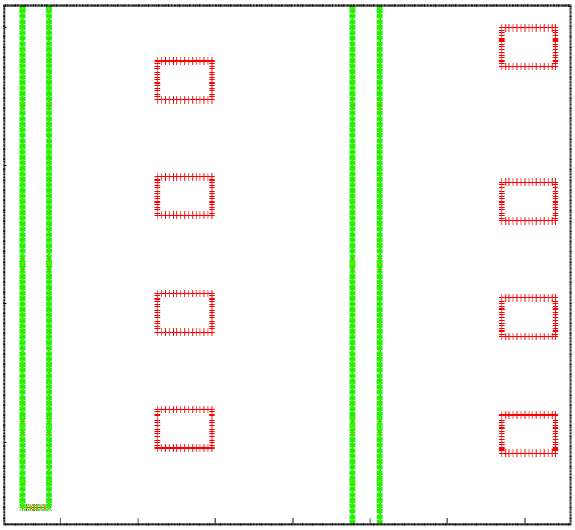
Here *C1* and *C2* are constants. The LP problem formulated above is solved using CPLEX LP solver [17] which gives an optimal solution with maximized metal feature distances within the *w* x *w* window. After all the windows in the partition M are solved, a combined optimized solution is obtained. The sliding window approach along with the easily generated space constraints has the benefit that it can be used with any sized layout without over burdening the solver with a bulky objective function.

Figure 14 below shows a post metalfill section of layout from ISCAS-85 C432 design for metal 2 obtained from each of the fill insertion techniques. After the metal fill insertion is

completed using either of the techniques described above, critical area analysis is performed on the layout.



(a) **(b)**



(c)

Figure 14. Post metalfill section of layout from ISCAS-85 C432 design for metal 2 using various fill insertion techniques (a) greedy fill insertion, (b) variable fill insertion and (c) LP based fill insertion techniques

4.4. Fill insertion in Non-preferred direction:

Traditionally the metal fill insertion techniques focus on satisfying the density goal and use tiling for dummy metal insertion. Other approaches use fixed dummy fill patterns in order to fill the layout space. Some of the recent techniques make use of printable assist features (regular dummy fills) and sub-resolution assist features to optimize the insertion and achieve additional goals to improve the manufacturability or design yield [12]. In the greedy and variable spacing fill insertion techniques discussed above, the dummy feature insertion depends on the available spacing and density target. These dummy features are routed in the regular routing directions specified in the design rules for the respective metal layer. In order to further optimize the fill insertion to have lesser critical area post fill insertion, we implement wrong way routing or non-preferred (NPF) direction routing during fill insertion.

Following the pseudo-code in Figure 13, the non-preferred direction fill is inserted in the layout prior to decrementing the spacing multiplier. Here the algorithm scans for free layout space in the non-preferred routing directions and inserts dummy fill features depending on the spacing multiplier and the density target. Figure 15below shows the updated pseudo-code for metal fill insertion with this implementation.

However in the more recent sub-millimeter technologies, the lithography process use off-axis illumination or dipole light sources in order to improve the printability. As a result only features aligned in regular directions for the metal layer mask being processed get printed on silicon. In order to be able to successfully print rectangular feature in non-preferred direction, we make use of wider metal features while filling the layout space.

Figure 16 shows a test layout with and without non-preferred direction fill. It is clear that by effectively utilizing the layout space in non-preferred direction, the critical area can be further minimized.

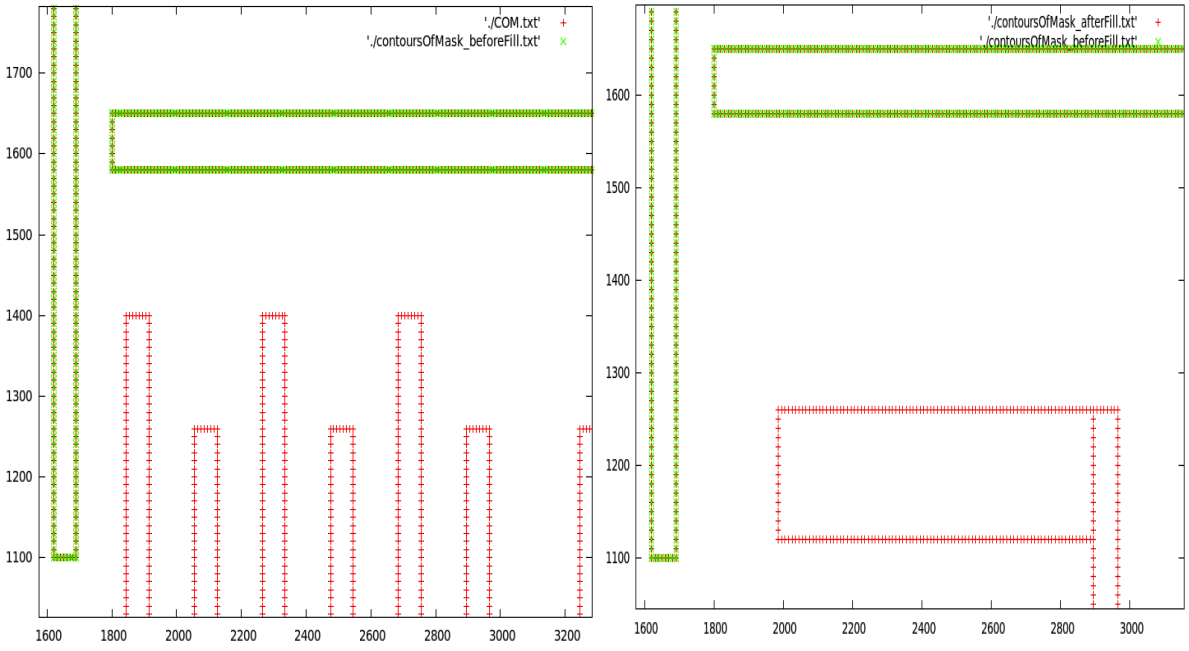
Metal Fill Insertion Technique

INPUT: Min. sized partition (M) , Spacing multiplier (N), Target density (T)

OUTPUT: Filled Layout for M

```
1.  foreach violating partition M
2.    obtain neighboring partitions (Mn);
3.      foreach partition to fill from (M + Mn)
4.        do:
5.          obtain rectangle box (R) &  min_space_req = (N . S);
6.          do:
7.            start fill insertion from centre of R moving outward;
8.            if (space available >= min_space_req)
9.              add fill shape in regular routing dir;
10.           else (scan layout in NPF dir)
11.             add fill shape in non-preferred routing dir;
12.           iterate over decreasing N for variable spacing
                technique;
13.         until (Density = T  or  box R limit reached)
11.         grow R concentrically;
12.       until ( R equals M )
13.     perform density checks (as in earlier section);
14.     redistribute & repeat for (M + Mn) if  (density < T)
15.     update densities for each (M + Mn);
```

Figure 15. Updated pseudo-code for metal fill insertion to include NPF direction fill



(a) **(b)**

Figure 16. Test layout comparing fill insertion techniques. (a) Regular routing fill (b) Non –preferred direction routing fill

4.5. Fill insertion to minimize coupling capacitance on critical nets:

In addition to satisfying the density criterion for metal filled layouts, it is desirable to have minimum impact on interconnect capacitance to maintain design performance.This problem of minimizing impact of capacitance can be approached in several ways depending on the design. One of most important fill decisions is to have the dummy fill features floating or connected to the nearest supply lines. As discussed earlier, floating fill and connected fill patterns have their own advantages and fallacies. Most of the existing solutions make use of floating fills and focus on minimizing the added coupling capacitance (Ccap) [30] [31]. However use of floating fills can not only lead to increased vulnerability to crosstalk delay and noise issuesbut also lead to having additional constraints while designing critical nets for expected performance. Also for dense designs it is a good practice to consider upper and lower layers as ground planes, while modeling the total capacitance for interconnects [32].Existing fill formulations primarily focus on inserting floating fill feature and do not consider the impact of overlap capacitances on performance [31]. Also maintaining uniform layout density and placing the fills for minimized capacitance has competing line spacing requirements.Thus placing dummy fill features with large inter-feature spacing can lead to greater ILD thickness variation across the layout.

The proposed coupling capacitance minimization technique focuses on these important issues. Given a layout, we first obtain all the critical nets in the design along with their initial net delays. For simplicity we consider any net longer than 10% of the maximum core dimension to be a critical net. Following the steps shown in the pseudo-code implementation in Figure 17, we use the model based layout density analysis model. The model based approach using Lorentzian kernel helps maintain smoother layout surface while the fill insertion tries to minimize the coupling capacitance on critical nets. Using this approach, we partition the layout and obtain the list of violating partitions based on the density target as indicated in lines 2 and 3. Following the earlier approach towards fill insertion, we start by populating the kernel window around each of

CCap Minimization Fill Insertion Technique

INPUT: Layout

OUTPUT: Filled Layout for Mimimised cap for critical nets

```
1.  Populate list of critical nets
2.  Use model based layout density analysis
3.  Partition the layout, populate list of violating partitions
4.      foreach partition to fill from (M + Mn)
5.      do:
6.          Shield critical nets with max spacing possible
7.          Calculate net delay for each critical net
8.          If delay worsens more than a limit, apply selective shielding
9.          Fill rest of the partition using variable spacing fill
10.     until (Density = T  or  Partition limit reached)
11.     redistribute & repeat for (M + Mn) if  (density < T)
12.     update densities for each (M + Mn)
```

Figure 17. Pseudo-code for Ccap minimization metal fill insertion

the violating partitions (as in lines 4 and 5). The spacing between two nets is inversely proportional to the coupling capacitance between them. As a result in order to minimize the coupling capacitance on critical nets, we first shield all critical net segments lying within the partition region with maximum inter-feature spacing possible. For each critical net, we then compare the initial delay and the net delay after shielding. If the delay worsens by more than pre-defined limit we implement selective shielding (as in lines 7 and 8). Here we calculate the net delay using interconnect model consisting of a driver (e.g. appropriately sized inverter), a 3π wire model and a FO4 unit sized inverter load. The 3π wire model gives an accurate estimate within 3% using the Elmore delay formulation [23]. The formulae for calculating coupling and ground capacitances are obtained from models proposed by Wong et. al [32] [34]. This limit is determined by the timing budget planned for the respective net. For example, on clock nets an uncertainty budget of about 50ps to 100ps is defined for each net, which basically includes the impact of possible clock jitter and other clock related issues. This pessimism becomes important

in case floating fill insertion as it increases the possibilities of higher coupling, clock jitter and noise issues. On the other hand, if we shield the fill features neighboring to such critical nets, then we minimize these probabilities and improve design performance.

For long nets, the ground capacitances are the major contributors to the total net capacitance. In such cases the net capacitance becomes less sensitive to inter-feature spacing variations due to neighboring fill features on the same layer as compared to the overlap capacitances. Thus we consider the worst-case impact of coupling capacitances in our delay calculations. Shielding critical nets over long run lengths can increase the overall ground capacitance and thus worsen the delay value. The pre-defined delay limit allows us to accommodate this extra delay in the net delay calculation and by controlling the amount of shielding we can relax the uncertainty budget to improve performance. After the critical net segments are appropriately shielded, we then fill the rest of the partition using variable spacing fill insertion technique. The algorithm repeats the above steps for all required partitions (as shown in lines 9 to 12).

CHAPTER 5

CRITICAL AREA ANALYSIS & SRAF INSERTION

In chapter 2 we discussed thatspot defects are random in nature and the size of these defects depends on the sanity of the manufacturing process. The probability that a defect will cause a structural fault depends on the size and shape of the defect and also the layout pattern geometry. As a result, as the layout geometries are scaled smaller and smaller, a given size defect has higher probability of causing a structural fault [14]. Thus in order to minimize the probability of failure for a given defect size distribution, performing critical areaanalysis of the layout design becomes important. We reviewed some of the existing techniques to perform critical area analysis on rectangular layout geometries. In this chapter we first describe the basic formulation to obtain the critical area for sample layout geometry. This is then followed by a detailed explanation on how this formulation is applied to calculate the critical area for the entire layout.

5.1 Basic critical area analysis formulation

The author in [14] describes the formulation to obtain the critical the for an open circuit structural fault. For the purpose of our work, we consider only the spot defects that cause short circuit faults. For simplicity we assume that the defects are rectangular in shape. Consider a simple case where we have two long parallel conducting segments of length L and width W as shown in Figure 18. Let S be the spacing between the two segments and A be the area of the die containing these segments. It is clear from the figure that a defect should have size equivalent to the spacing S to cause a structural fault. Also that any defect having a size x such that $S < X < S + A/L_A$ can cause a fault.

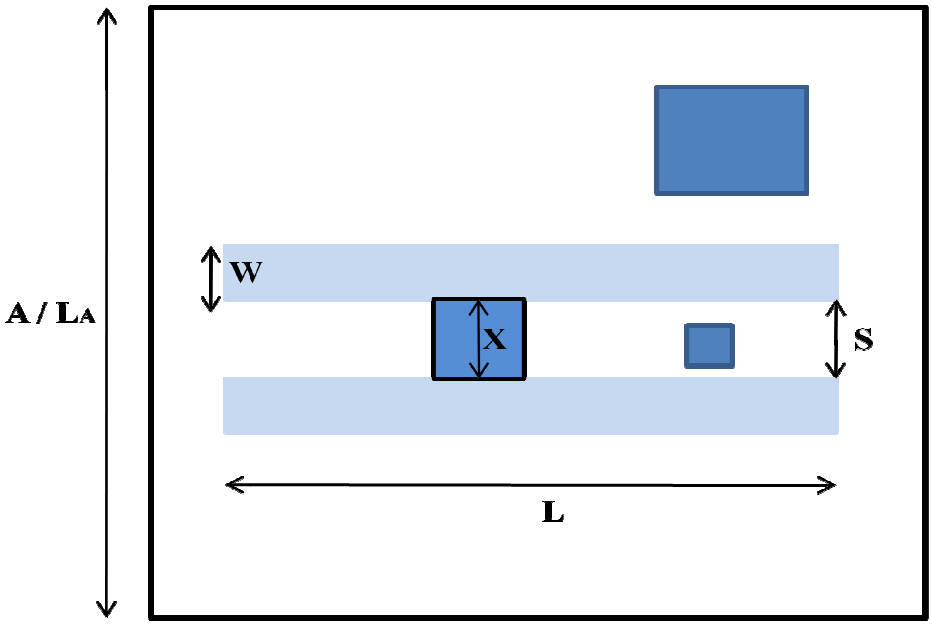


Figure 18. Sample layout showing the minimum size of the defect X (=S) that can cause a short circuit [18]

However for any defect size that has size in the range $S \leq X \leq S + A/L_A$, the critical area is computed as shown in Figure 19. The rectangular area in between the two rectangles is given

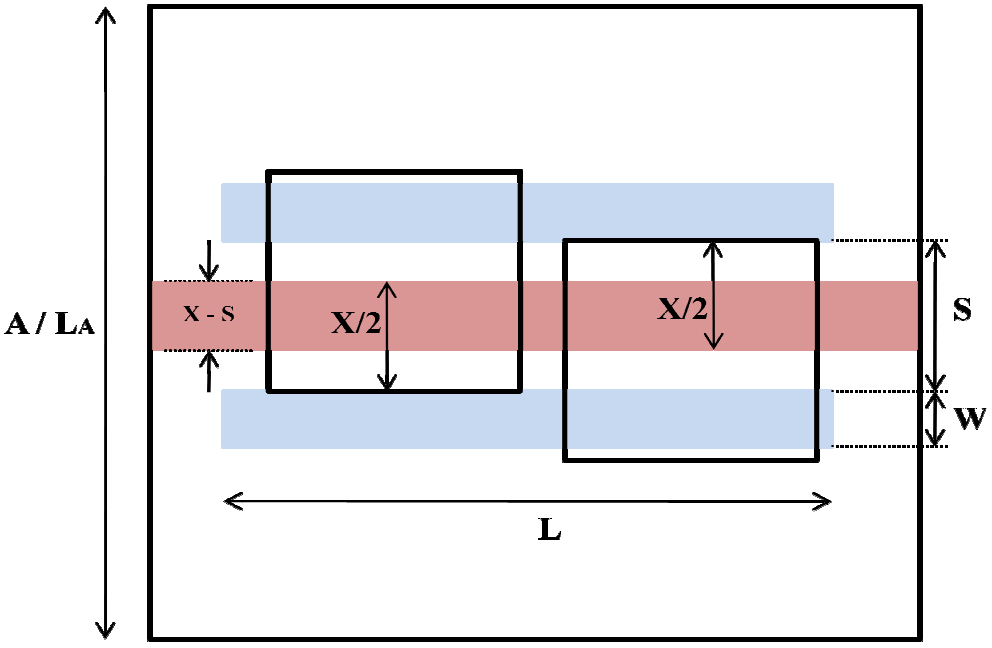


Figure 19. Critical area for a defect size of X between two rectangles (width W) separated by a distance S

by (X – S) * L plus the extension of this area beyond the length of the rectangles by an amount X/2 on both sides. Thus the total critical area obtained is

$$Critical\ Area = (X - S) * L + 2 * (X - S) * X/2$$

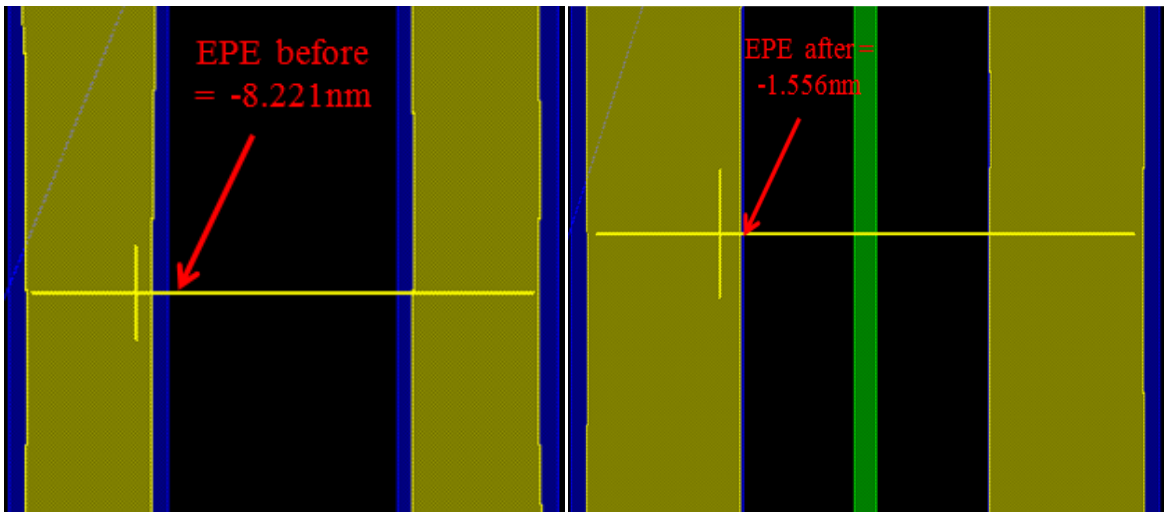
Generalizing the critical area computation for the entire layout given a defect size, we observe that the total critical is a linear sum of the individual component areas [14]. Thus as the defect size increases, the critical area increases. The probability of failure or fault occurrence being proportional to the critical area also increases.

5.2 Critical area analysis for entire layout

In order to calculate the critical area, we make use of the above expression. We compute the critical area using a pixel based layout. We make use of the layout partitions that are created during partitioning of the layout for density analysis to ensure scalability and each partition is represented as an array of Boolean pixels with a value '1' in constant time. Note that this description of pixels is different from the one used in the earlier chapters. Each rectangle is expanded on all sides by defect_size/2. Representing rectangles of the same net with similar identities maintains accuracy of computation. We maintain a forward linking list of rectangles that prevents a pair of rectangles from being considered twice during computation. Each expanded rectangle is checked for possible overlaps with its neighbors. These overlapping regions if any give the critical area between the corresponding rectangles. Multiple overlaps are taken care of automatically since all overlapping pixels hold the same Boolean value '0'. This avoids the need to have an ECSP structure. The final critical area is the total overlapping pixel region. This algorithm is linear to the number of rectangles present in the partition. By adding the critical area computed for all partitions, we obtain the total critical area for the layout.

5.3 SRAF Insertion

The dummy metal features discussed so far are additional features added to the mask that not only help maintain the topographical planarity but also assist in printing the required layout patterns. These assist patterns which also get printed on the mask are called printed assist features (PAFs). Similarly another set of features can be added to the mask to assist in printing the layout called as sub-resolution assist features. As the name suggests, these features are considerably smaller than the minimum feature size and hence they themselves don't get printed on the layout. However their presence improves the layout printability [21]. This improvement is characterized as reduction in the edge placement error (EPE). Edge placement error is the difference between the layout edge and the printed feature edge. Figure 20 shows the reduction in EPE on two rectangles on which lithography simulation was performed before and after the SRAF was inserted.



a) EPE before SRAF insertion **b) EPE after SRAF insertion**

Figure 20. Reduction in EPE post SRAF insertion

In the proposed SRAF insertion approach, we add single or multiple SRAFs depending on the spacing between the dummy metal features [21]. This spacing also controls the width the of the SRAF features being added. For the 45nm technology, we add SRAFs with varying width

from 4.5nm to 13.5nm in increments of 10% of the minimum feature size. Table I shows the varying SRAF width and count inserted depending on the inter-feature spacing and maximum EPE improvement observed.

Table 1.SRAF characterization

Feature spacing (nm)	SRAF Width (nm)	No. of SRAFs inserted
>=100 &<=140	9	1
> 140	9	2
> 210	13.5	2

CHAPTER 6

EXPERIMENTAL RESULTS

6.1 Experimental setup

For the purpose of our work, we chose from various ISCASC-85 benchmark circuits. The layout density scan and dummy fill insertion algorithms were implemented in C++. SRAF characterization is done in CalibreWorkBench [20]. Various circuits from the ISCAS-85 benchmark librarywere chosen for the study. The fill insertion techniques were implemented on metal layer 2 of the ISCAS-85 layout designs. Figure 21 shows a section of the layout for ISCAS-85 C432 design on which lithography simulation is performed in CalibreWorkBench [20] after dummy fill insertion. It also shows the inserted sub-resolution assist features with varying widths depending on the inter-feature spacing. For the purpose of critical area analysis, we assume a defect size range from 140nm to 210nm with a defect distribution suggested by Koren [13].

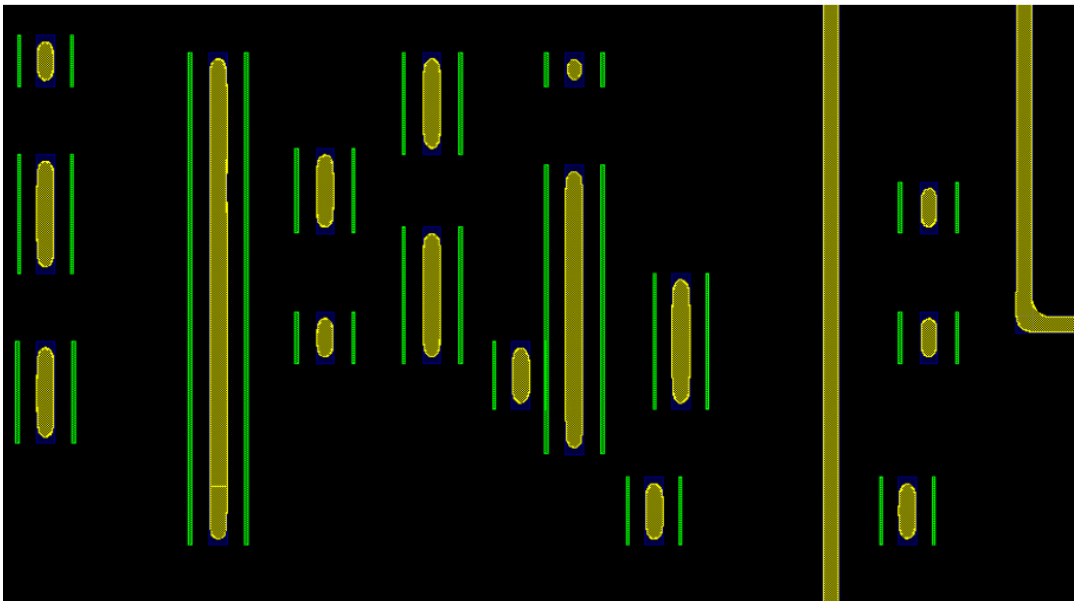


Figure 21. Section of c432 ISCAS-85 layout design metal layer 2 on which lithography simulation is performed post dummy fill and SRAF insertion

6.2 Rule based layout density analysis

In this section we present the results for rule based density analysis model. We implement the various fill insertion techniques using rule based density analysis modelfollowed by the critical area analysis and comparison. The rule based approach drives the fill insertion techniques to satisfy the lower limit (30%) of the density bound so as to have a density baseline for critical area analysis. However these techniques were also found to satisfy higher density requirements.

Figure 22 shows the increase in average critical area as the defect size increases for all fill insertion techniques implemented on C499 benchmark circuit.

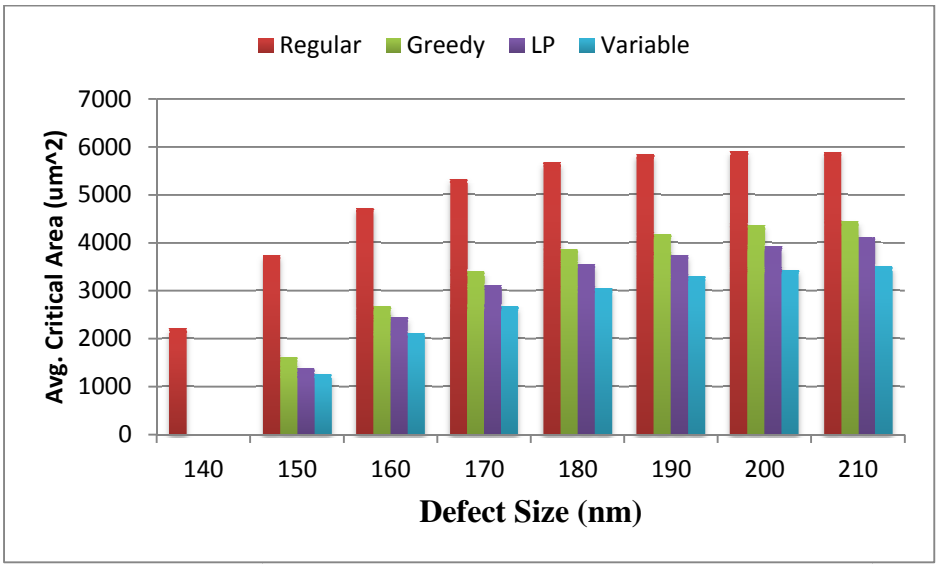


Figure 22.Varying defect size vs Critical area for C499 benchmark

Table 2 below shows the average critical area comparison for defect distributionfor various benchmark circuits with layout size on which the fill insertion techniques were implemented. The heuristic algorithms are compared to regular fill insertion technique that inserts dummy fill without any space utilization. Variable spacing and LP formulation based approaches utilize available space in a better way as compared to the greedy fill technique. It is observed that the average critical area for the fill insertion techniques proposed is significantly better as compared to the regular fill approach. From the table it can be seen that the fill insertion using space utilization in the variable spacing approach provides the least critical area followed by the LP

formulation and the greedy fill approaches. The greedy fill approach however gives a substantially better critical area number as compared to the regular fill technique.

Table 2. Critical area comparison between fill insertion techniques

Benchmark Circuits	Layout Area (um^2)	Avg. Critical Area (um^2)			
		Regular Fill	***Greedy (or fixed spacing) Fill***	***LP based Fill***	***Variable Spacing Fill***
C432	70.00 x 60.00	43914.51	29999.29	24216.34	21236.248
C499	66.12 x 59.28	39346.26	24521.58	22262.28	19343.09
C880	80.00 x 80.00	65014.03	44289.36	36481.71	32037.505
C1355	88.00 x 88.00	78963.78	54646.29	42315.27	38740.13
C2670	100.00 x 95.00	99504.88	64158.08	48682.64	47969.39
C3540	105.65 x 103.74	110609.05	65493.4	56046.83	54086.88
C5315	140.00 x 136.00	192704.39	123308.5	96213.52	93708.58

6.3 Fill insertion in non-preferred direction (NPF)

As discussed earlier, in order to obtain acceptable printability of metal features using off-axis or dipole illumination sources, we use wider metal pitches. For non-preferred direction routing of metal 2, we use double-width and double-spacing rule. This is based on the lithography simulations performed using off-axis light sources in Calibre Workbench [12] [20].

Table 3. Critical area comparison between variable fill and NPF fill insertion techniques

Benchmark Circuit	CA-Variable Spacing fill (um^2)	CA-Non-preferred fill (um^2)
C432	65.19	60.66
C499	57.21	52.19
C880	97.804	92.05
C1355	119.805	112.753
C2670	143.22	133.99
C3540	156.893	136.174
C5315	277.655	252.36

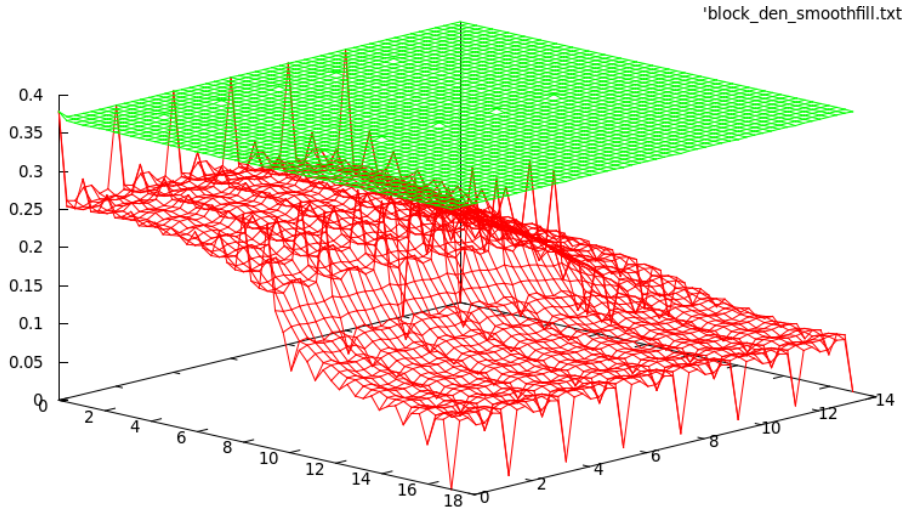
Table 3 shows a comparison of the critical area numbers for defect size of 150nm obtained from non-preferred direction fill insertion and the variable spacing fill insertion techniques both using rule based layout density analysis model. Here the fill contours are connected to the nearest supply line (i.e. VDD or GND). The non-preferred direction fill insertion algorithm improves the critical area substantially for the various benchmark circuits.

6.4 Model based layout density analysis

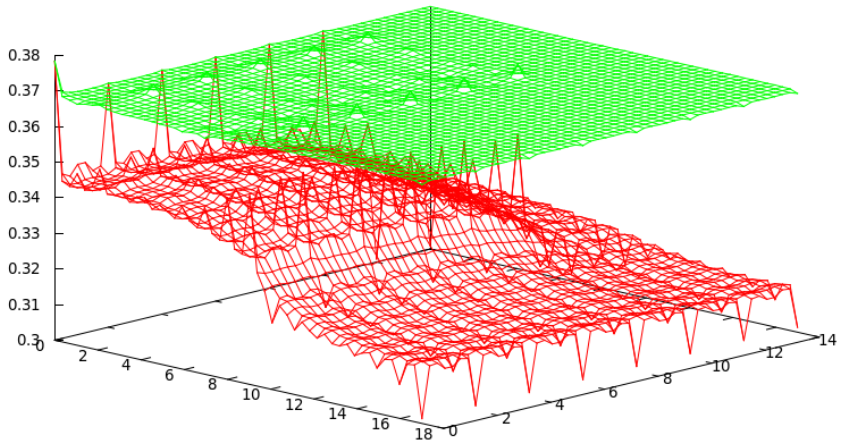
This section presents the results on the model based layout density analysis framework. Given the list of violating partitions, we use the Lorentzian kernel with a FWHM or with a planarization length of 28µm to perform fill insertion. The ranged-variation approach uses a tolerance of ε = 3% of the computed target effective density value for each partition. This variation is then converted into Z variation using the following equation:

$$Z = Z_{ILD} \,.\, [\max(\rho_0) - \min(\rho_0)]$$

whereZ_{ILD} is the step height (108nm) [35]. Figure 23 (a) below shows CMP topography in terms of the initial density for a partitioned test layout (red surface plot) and the post-fill effective density computed for all partitions using the Lorentzian kernel (green surface plot). While



(a) CMP topography comparison before fill insertion (red plot) and after fill insertion (green plot)



(b) CMP topography comparison between rule based (green plot) and model based density analysis (red plot)

Figure 23. CMP topography represented by partition density (Z axis) and partition numbers (X & Y axes) (a) with initial and post-fill insertion (b) post-fill insertion with rule based and model based density analysis on a test layout

Figure 23 (b) shows the similar CMP topography in terms of the post-fill insertion partition densities comparing rule based density analysis (green plot) and model based layout density analysis (red plot) approaches. Also the worst-case ILD thickness variation obtained post-fill insertion using rule based layout density analysis approach is 8.52nm as compared to 2.22nmobtained using model based layout density analysis. Thus from the above plots and the ILD thickness variation values obtained, it is clear that compared to the rule based density analysis, the model based layout density analysis approach coupled with NPF fill insertion technique provides us with the best surface topography along with minimized critical area.

Table 4. Comparing results of density analysis approaches

Approach	**Min (ρ_0)**	**Max (ρ_0)**	**Average (ρ_0)**	**Max Z (nm)**
Initial (pre-fill)	0.0	0.38	0.154	40.98
Rule-based	0.3	0.38	0.326	8.52
Model-based	0.36	0.38	0.3679	2.22

6.5Fill insertion to minimize coupling capacitance on critical nets

In this section we discuss the results obtained from the fill insertion solution to minimize coupling capacitance on critical nets presented in section 4.5. Table 5 below shows a comparison of the timing results obtained the regular fill insertion technique and the Ccap minimization technique. For the various benchmark circuits it gives the number of critical nets present in the design, the worst-case degradation in terms of percentage change and also net delay valuesfollowed bythe number of nets for which the timing improved post coupling capacitance minimization fill. The net delay calculation here is the total path delay including cell delays. It is observed that the delay degradation in approximately half the nets is because of the fact that the coupling capacitancecontributes to about 30% to 70% of the total capacitance. As a result post coupling aware fill insertion, for low ccap contribution nets the total ground capacitance now present worsensthe delay even though the coupling capacitance is minimized. However the

Table 5. Net delay results post Ccap minimization fill insertion

Benchmark Circuit	No. ofCritical Nets	% WC delay degradation	WC Delay Degradation	No. of nets improved
C432	48	8.55	1.61ps	27
C880	89	9.24	1.78ps	31
C1355	92	9.22	1.75ps	42
C2670	125	14.99	3.24ps	66
C3540	147	7.71	1.44ps	76
C5315	123	16.34	3.62ps	63

Table 6. Delay distribution comparingno. of nets in regular fill vsCcap min. fill insertion

Circuits	C432		C880		C1355		C2670		C3540		C5315	
Time (ps)	regFill	Ccap Fill	regFill	Ccap Fill	regFill	Ccap Fill	regFill	Ccap Fill	regFill	Ccap Fill	regFill	Ccap Fill
0 - 4.5	13	9	32	17	21	12	25	7	40	18	4	1
4.5 - 5	5	17	16	28	17	25	16	26	21	39	16	2
5 - 5.5	10	4	16	20	12	18	9	34	16	39	16	14
5.5 - 6	8	9	8	11	16	15	15	22	17	20	25	35
6 - 6.5	4	6	7	3	8	15	21	18	14	7	18	20
6.5 - 7	4	2	1	3	7	2	14	9	12	11	13	12
7.0 - 10	4	1	7	6	11	5	25	9	27	13	31	39

percentage delay degradation is less as compared to the total path delay and hence, the degradation in net delays is within an acceptable timing budget for these critical nets. Table 6 above shows the delay distribution comparison between the regular fill insertion and the Ccap minimization fill insertion techniques for various circuits. It can be observed that the net delay improves for nets having higher contribution of coupling capacitance. Consequently the number of nets having worst-case delays reduces post Ccap minimization fill. Figure 24 below shows a histogram for C3540 circuit indicating the change in the distribution of net delay for the critical nets.

The numbers of nets having delays greater than 6ps reduces post coupling minimization fill. It is clear that post coupling capacitance minimization fill insertion the net delay distribution does not change enough to impact the design performance. Thus by minimizing the coupling capacitance we are able to avoid any crosstalk related issues post-fill insertion without impacting the design performance and also minimize the critical area at the same time.

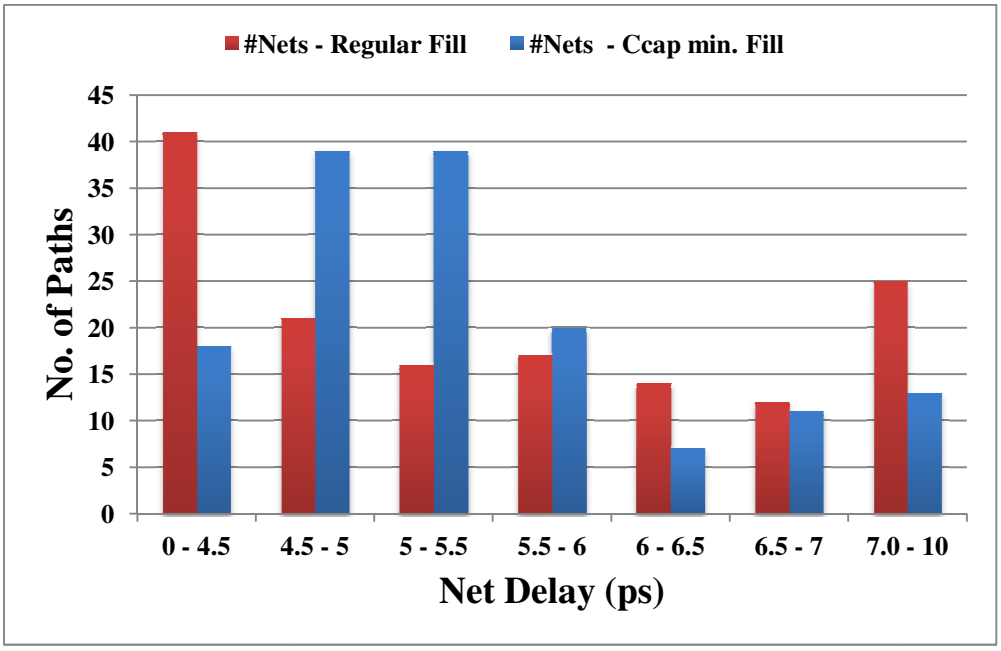


Figure 24. Histogram for C3540 benchmark comparing net delays for regular and Ccap minimization fill

CHAPTER 7

CONCLUSION & FUTURE WORK

Due to the increased lithographic printability and CMP planarity concerns, DFM checks are of critical importance. Satisfying the layout density bound for effective CMP by fill insertion alone is not enough to obtain higher yield benefit. Thus improving design yield by minimizing the critical area for random spot defects is equally important. In our proposed work, we present rule based and model based layout density analysis approaches for dummy fill insertion. The rule based density analysis model provides a fast and efficient solution while the model-based approach minimizes the ILD thickness variation. This improves the post CMP topography for the layout and improves manufacturability. These layout density analysis algorithms can be coupled with any of the various dummy fill insertion methods discussed. The non-preferred fill insertion helps us obtain a better fill insertion solution for a given spacing constraint with respect to critical area. The coupling capacitance minimization fill insertion solution minimizes the impact of inserted fill on critical nets in terms of coupling capacitance while at the same time minimizing the critical area. By performing critical area analysis on post-fill layout, it is observed that the proposed solutions provide effective space utilization to minimize critical area while satisfying the density limit criteria. We have not found any comparable work that considers lithographic printability and critical area minimization together.

A part of the thesis work presented in this document has been accepted and published in ISQED 2012. As part of the future work, we plan to enhance an existing double patterning lithography simulator to minimize the number of coloring conflicts by merging into it the presented dummy metal fill insertion solution.

BIBLIOGRAPHY

[1] SandipKundu and AswinSreedhar, Nanoscale CMOS VLSI Circuits: Design for Manufacturability, McGraw-Hill Professional, 2010.

[2] G. Nanz and L. Camilletti, "Modeling of Chemical-Mechanical Polishing: A Review," IEEE Trans. Semicon. Manufacturing, 8(4):382-389, November 1995.

[3] A. B. Kahng, G. Robins, A. Singh, H. Wang, and A. Zelikovsky, "Filling and Slotting: Analysis and Algorithms," Proc. International Symposium on Physical Design, Monterey, CA, pp. 95-102, 1998.

[4] A. B. Kahng, G. Robins, A. Singh, and A. Zelikovsky, "New multilevel and hierarchical algorithms for layout density control," in Proc. IEEE Asia South Pacific Des. Autom. Conf., 1999, pp. 221–224.

[5] A. B. Kahng, G. Robins, A. Singh, and A. Zelikovsky, "New and exact filling algorithms for layout density control," Proc. IEEE Int. Conf. VLSI Des., 1999, pp. 106–110.

[6] A. B. Kahng, G. Robins, A. Singh, and A. Zelikovsky, "Filling algorithms and analyses for layout density control," IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 18, no. 4, pp. 445–462, Apr. 1999.

[7] Y. Chen, A. B. Kahng, G. Robins, and A. Zelikovsky, "Closing the Smoothness and Uniformity Gap in Area Fill Synthesis," Proc. International Symposium on Physical Design, San Diego, CA, pp. 137-142, 2002.

[8] R. Tian, D. F. Wong, and R. Boone, "Model-based dummy feature placement for oxide chemical mechanical polishing manufacturability," Proc. ACM/IEEE Des.Autom.Conf., Los Angeles, CA, pp. 667–670, 2000.

[9] M. Mukherjee and K. Chakraborty, "A randomized greedy method for rectangular-pattern fill problems," IEEE Trans. on CAD, 27(8):1376– 1384, 2008.

[10] H. Xiang, K.-Y. Chao, R. Puri, andM. D. F.Wong, "Is your layout density verification exact?—A fast exact algorithm for density calculation," Proc. ACM/IEEE Int. Symp. Phys. Des., pp. 19–26, 2007.

[11]Y. Chen, A. B. Kahng, G. Robins, and A. Zelikovsky, "Hierarchical dummy fill for process uniformity," Proc. IEEE Asia South Pacific Des Autom. Conf., pp. 139–144, 2001.

[12] L. Deng, K. Chao, H. Xiang, and M. D. F. Wong, "Coupling-aware dummy metal insertion for lithography," Proc. IEEE Asia South Pacific Des Autom.Conf., 2007.

[13]I. Koren and Z. Koren, “Defect Tolerant VLSI Circuits: Techniques and Yield Analysis,” Proceedings of the IEEE, 1998.

[14] I. Koren, “The effect of scaling on the yield of VLSI circuits,” in Yield Modeling and Defect Tolerance in VLSI, W. Moore, W. Maly, and A. Strojwas, Eds. Bristol, UK: Adam Hillger, pp. 91–99, 1988.

[15]H. Xue, C. N. Di, and J. A. Jess ,“Fast multi-layer critical area computation,” Proc. of IEEE Int. Workshop on Defect and Fault tolerance on VLSI Systems, Oct. 1993.

[16]J. H. N. Mattick, R. W. Kelsall and R. E. Miles, “Improved Critical Area Prediction by Application of Pattern Recognition Techniques,” Microelectronics Robustness, vol. 36, no. 11-12, pp.1815-1818, Dec. 1996.

[17]ILOG Inc. ILOG CPLEX 6.5—user’s manual. ILOG Inc., 1999. http://www.ilog.com

[18]A. V. Ferris-Prabhu, “Role of defect size distribution in yield modeling,” IEEE Trans. Electron Devices, vol. ED-32, pp. 1727–1736, Sept. 1985.

[19]MOSIS CMP and Antenna Rules,
“http://www.mosis.com/Technical/Designrules/guidelines.html”

[20]Mentor Graphics, CalibreWORKbench Users Manual.

[21]Dhumane, Nishant; Srivathsa, Sudheendra K.; Kundu, Sandip, "Lithography Constrained Placement and Post-Placement Layout Optimization for Manufacturability," ISVLSI, vol., no., pp.200-205, 4-6 July 2011.

[22] A. B. Kahng, K. Samadi, R. O. Topaloglu, "Recent topics in CMP-related IC Design for Manufacturing,” Proc. Advanced Metallization Conference, 2008.

[23]N. Weste, D. Harris,”CMOS VLSI Design, A Circuits and Systems Perspective”, Addison--Wesley, 4th edition, 2010.

[24] J-P.Urbach and R. Rzehak “A Novel Approach to Analyze and Model Feature Size Effects in CMP”, *IEEE Trans. on Semiconductor Manufacturing,* 22(4) Nov. 2009, pp.566–571.

[25] Stine, B.E.; Ouma, D.O.; Divecha, R.R.; Boning, D.S.; Chung, J.E.; Hetherington, D.L.; Harwoo, C.R.; Nakagawa, O.S.; Soo-Young Oh; , "Rapid characterization and modeling of pattern-dependent variation in chemical-mechanical polishing," *Semiconductor Manufacturing, IEEE Transactions on* , vol.11, no.1, pp.129-140, Feb 1998

[26] B. Stine, D. Ouma, R. Divecha, D. Boning, J. Chung, D. Hetherington,I. Ali, G. Shinn, J. Clark, O. S. Nakagawa, and S.-Y. Oh, “A closedformanalytic model for ILD thickness variation in CMP processes,” inProc. CMP-MIC Conf., Santa Clara, CA, Feb. 1997, pp. 266–273

[27] R. Divecha, B. Stine, D. Ouma, J. Yoon, D. Boning, J. Chung, O.Nakagawa, and S.-Y. Oh, “Effect of fine-line density and pitch oninterconnect ILD thickness variation in oxide CMP processes,” inProc. 2nd Int. Conf. Chemical Mechanical Polishing for ULSI MultilevelInterconnect Conf., Santa Clara, CA, Feb. 1997, pp. 29–36

[28] Ouma, D.O., Boning, D.S., Chung, J.E., Easter, W.G., Saxena, V., Misra, S., Crevasse, A., "Characterization and modeling of oxide chemical-mechanical polishing using planarization length and pattern density concepts," *Semiconductor Manufacturing, IEEE Transactions on* , vol.15, no.2, pp.232-244, May 2002

[29] D. Ouma, D. Boning, J. Chung, G. Shinn, L. Olsen, and J. Clark, “An integratedcharacterization and modeling methodology for CMP dielectricplanarization,” in Proc. IEEE Int. Interconnect Technology Conf., Feb.1998, pp. 67–69

[30] Kahng A.B., TopalogluR.O., “Performance-aware CMP Fill Pattern Optimization”, Proc. International VLSI/ULSI Multilevel Interconnection Conference,2007

[31] Stine, B.E., Boning, D.S., Chung, J.E., Camilletti, L., Kruppa, F., Equi, E.R., Loh, W., Prasad, S., Muthukrishnan, M., Towery, D., Berman, M., Kapoor, A., "The physical and electrical effects of metal-fill patterning practices for oxide chemical-mechanical polishing processes," *Electron Devices, IEEE Transactions on* , vol.45, no.3, pp.665-679, Mar 1998

[32] Shyh-ChyiWong, Gwo-Yann Lee, Dye-Jyun Ma, "Modeling of interconnect capacitance, delay, and crosstalk in VLSI," *Semiconductor Manufacturing, IEEE Transactions*, vol.13, no.1, pp.108-111, Feb 2000

[34] Predictive Technology Model, ASU, www.ptm.asu.edu

[35] Datta. B, Burleson. W, "On temperature planarization effect of copper dummy fills in deep nanometer technology," *ISQED*, vol., no., pp.494-499, 16-18 March 2009