Keynote Paper

CMP Fill Synthesis: A Survey of Recent Studies

Andrew B. Kahng, *Senior Member, IEEE*, and Kambiz Samadi, *Student Member, IEEE*

***Abstract*—We survey recent research and practice in the area of chemical–mechanical polishing (CMP) fill synthesis, in terms of both problem formulations and solution approaches. We review the CMP as the planarization technique of choice for multilevel very large-scale integration metallization processes. Post-CMP wafer topography varies according to pattern density. We review density-analysis methods and density-control objectives that are used in today's fill-synthesis algorithms. In addition, we discuss the concept of design-driven fill synthesis that seeks to optimize CMP fill with respect to objectives beyond mere density uniformity. Design-driven fill synthesis minimizes the impact of CMP fill on design performance and parametric yield while still satisfying the density criteria that are motivated by manufacturing models. We conclude with a discussion of where CMP fill synthesis may be naturally integrated within future design and manufacturing flows.**

***Index Terms*—Chemical–mechanical polishing (CMP), CMP fill, design-driven fill, topography.**

I. Introduction

CHEMICAL–MECHANICAL polishing (CMP) is the planarizing technique of choice to satisfy the local and global planarity constraints imposed by today's advanced lithography methods [36], [54]. As device geometries scale, there is an inevitable need for better planarization of the multilevel-interconnect structures. Early planarization approaches are of two types: spin-on-glass (SOG) and reverse etchback (REB).

SOG is a method which consists of coating the surface of a particular layer with SOG materials. These materials can be categorized in three different groups: 1) silicate-based compounds; 2) organosilicon compounds; and 3) dopant-organic compounds. In SOG, the silicon wafer must be cleaned before coating. The wafer is placed on a spinner, and approximately 1 ml of SOG material is dropped on the center of the wafer. Then, the carrier holding the wafer is rotated at several thousand cycles per minute to create a thin layer of the SOG material on the wafer. In most cases, a film thickness between 50 and 500 nm will result. Controlling the thickness is a matter of controlling the solution viscosity. Among the various techniques for planarization, the SOG method is advantageous because of the simplicity of the process, the good adhesion characteristics, and the low level of stress and shrinkage in the SOG material [63].

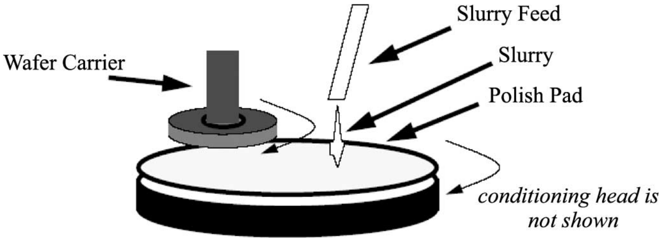


Fig. 1. Typical CMP tool [36].

REB uses a second mask to etchback raised areas to lower the pattern density. The etchback mask is created by shrinking all features on a given layout by a fixed amount called "etchback bias." This results in removal of the majority of the raised material if the features are large. Selective REB uses customization of the etchback mask to reduce the amount of material that is etched away [37].

In modern submicrometer processes, SOG and REB techniques no longer work, because implementation of the SOG technique requires thorough understanding of the glass and the stability of the remaining material, which leads to considerable variation in the practicality of this technique [50]. In addition, REB suffers from complexity and significant cost due to extra masking steps. In addition, REB requires a significant amount of monitoring to control the level of defects caused by the process [50]. CMP uses both mechanical and chemical means to planarize the surface of the wafer. In a typical CMP tool, the wafer is held on a rotating holder, as shown in Fig. 1. The surface of the wafer being polished is pressed against the polishing pad (i.e., a resilient material), which is mounted on a rotating disk. A slurry composed of particles suspended in a chemical solution is also deposited on the pad as the chemical abrasive.

The material-removal mechanism of silicon-dioxide (oxide) CMP is similar to the removal found in glass polishing. First, a chemical reaction softens the deposited film surface and, then, a mechanical surface abrasion aided by slurry particles removes the material [13], [36]. The chemical reaction between the slurry and the surface of the wafer creates a hydroxilated-form material. The new material has weaker atomic bonds. It is, therefore, more easily removed during the polishing process [45]. The second step involves the removal of the weakened film surface through abrasion.

Manuscript received November 1, 2006; revised February 20, 2007. This work was supported by MARCO Gigascale Systems Research Center. This paper was recommended by Associate Editor D. Z. Pan.

A. B. Kahng is with the Department of Computer Science and Engineering and the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093-0404 USA (e-mail: abk@ucsd.edu).

K. Samadi is with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093-0409 USA (e-mail: ksamadi@ucsd.edu).

Digital Object Identifier 10.1109/TCAD.2007.907061

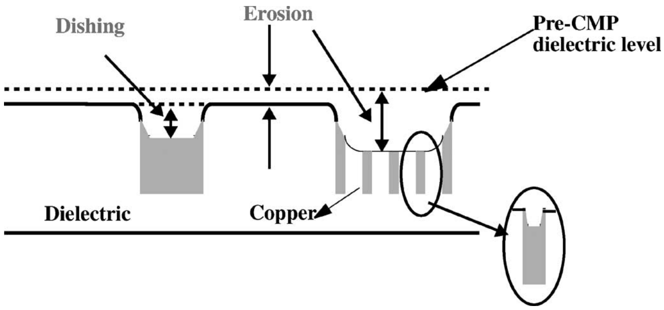


Fig. 2. Dishing and erosion [64].

In the very deep-submicrometer, very large-scale integration-regime manufacturing steps, including optical exposure, resist development, and etch, and CMP have varying effects on device and interconnect features depending on local properties of the layout. Foundry economics dictate that the process-window volumes be maximized, which in turn requires that device and interconnect features be fabricated as predictably and uniformly as possible. To achieve this goal, the layout must be made uniform with respect to a certain density parameter. The physics of semiconductor processing make predictable and uniform manufacturing difficult [5], [14], [35], [53]. In particular, the quality of post-CMP depends on the pattern density of the layer beneath a given dielectric layer.

In the past decade, CMP has emerged as the predominant planarization technique for multilevel metallization processes. However, significant surface-topography variation can still exist for some layout patterns; this impacts depth of focus (DOF) in lithography, which in turn leads to variations in critical dimension (CD) (Fig. 13). Two other major defects caused by CMP are metal dishing and oxide erosion. In the copper CMP process, metal dishing is defined as the difference between the height of the oxide in the spaces and that of the metal in the trenches. Oxide erosion is defined as the difference between the oxide thickness before and after CMP [65]. In this paper, dishing and erosion refer to metal dishing and oxide erosion, respectively. Fig. 2 shows metal dishing and oxide erosion in copper CMP process. These two phenomena impact the performance of the circuit, since variation in interlayer-dielectric (ILD) thickness profile and interconnect height lead to variations in interconnect capacitance and resistance. This variation will increase the timing uncertainty of the circuit; hence, it is crucial to minimize dishing and erosion. However, due to CMP nonidealities, there will always be some amount of dishing and erosion. It is important to model the effect of these variations during parasitic extraction to obtain a more accurate estimation of the circuit performance [52].

Even though pattern density is the major cause of the CMP defects, there are other factors, such as slurry flow rate and pad-conditioning temperature, that contribute to the amount of dishing and erosion. The slurry acts as a coolant material at the interface of the pad and wafer contact and takes away a significant part of the heat through convective heat transfer [42], [55], [56], [66]. The dissipated heat changes the chemical kinetics and the physical properties of the polishing pad [42], [55], [66]. As the amount of dissipated heat increases, the polishing pad tends to become softer, which results in an increase in the area contact at the interface. In addition, pad conditioning has a major impact on the removal rate during the CMP process, as underconditioned pads will lose their surface roughness, which eventually leads into removal-rate reduction [44] (for more details on effects of slurry flow rate and pad-conditioning temperature on metal dishing and dielectric erosion, the reader is encouraged to look into [44]).

To increase the fabrication-process uniformity and predictability, the layout must be made uniform with respect to a certain density parameter. One solution for designers and manufacturers is to use techniques like CMP fill[1] insertion and slotting to increase and decrease the pattern density [26]. CMP fills are dummy features that do not directly contribute to the functionality of the circuit and can either be grounded or left floating. CMP fill insertion reduces the amount of dishing and erosion by increasing the pattern-density uniformity. However, it is well known that CMP fill insertion can increase the coupling and total interconnect capacitance and, consequently, deteriorate circuit performance [38], [57]. If not modeled appropriately, this can directly affect yield and time-to-market.

The rest of this paper is organized as follows. In Section II, we propose a taxonomy for the CMP fill and review its benefits and tradeoffs in current designs. In Section III, different density-analysis methods are studied. Section IV first discusses characterization and modeling approaches for back-end-of-line (BEOL) CMP processes and then reviews different corresponding fill-synthesis approaches; it reviews both the problem formulations, as well as the solutions. In Section V, we first discuss characterization and modeling techniques for front-end-of-line (FEOL) CMP processes and then discuss the shallow-trench-isolation (STI) fill-insertion problem. Section VI introduces the concept of design-driven fill synthesis, which tries to optimize the fill-insertion technique by reducing its impact on the design performance while satisfying the required density constraints. Section VII gives the summary and discusses the future CMP fill flows.

II. CMP Fill Taxonomy, Benefits, and Tradeoffs

As the number of layers increases and linewidths shrink, tolerance for topographical imperfections decreases. This is due to tight DOF variation requirements and high sensitivity of resistance to metal thickness. Despite improvements in CMP technology, layout-pattern sensitivities are significant, causing certain regions to have higher topographies than others, due to differences in underlying densities [59]. CMP-fill-insertion technique is employed by the designers and manufacturers to decrease the metal-density variation [26]. Dummy fills are non-functional features that do not directly contribute to the logic implementation and can either be grounded or left floating.

A. Grounded Versus Floating

Traditionally, foundry-supplied design rules have been used by the designers to meet density requirements while not sig-

[1] In this survey, CMP fill and fill have been used interchangeably.

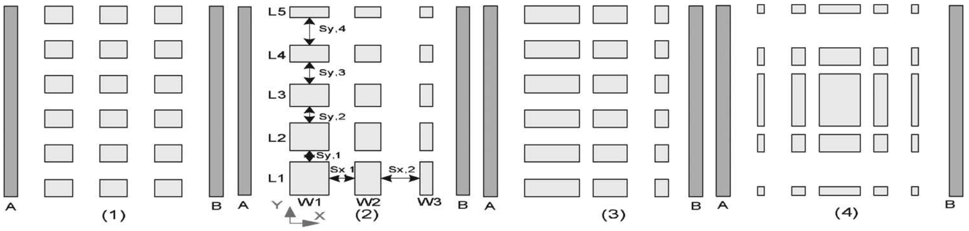


Fig. 3. Examples of fill pattern [17].

nificantly increasing the interconnect capacitance. While fill-insertion design rules have sufficed until now, they are overly conservative and arguably at the end of their lifecycle. Specifically, buffer-distance (i.e., no fill shape is less than this distance of any layout feature) rules have been used to limit the impact of fill on total and coupling capacitance. As crosstalk analysis gains importance and interconnect delay increases, both coupling and total capacitance must be accurately modeled. In the absence of reliable fill extraction tools, buffer distance (or keepoff distance) must be increased. Unfortunately, this is not feasible since small density variation may not be achievable if buffer distance is large [30].

Floating fill, in comparison to grounded fill, offers smaller increase in total capacitance and does not require power/ground routes to the fill geometry. However, floating fill increases coupling capacitance that can lead to signal-integrity issues. In the absence of fast and reliable floating-fill extraction tools,[2] floating fill is cautiously used or not used at all (e.g., in analog circuits). Grounded fill, despite its larger impact on total capacitance and high routing costs that often lead to engineering change orders, is used as a substitute [30].

B. *Impact on RC Parasitics*

In this section, we review the impacts of CMP fill on both interconnect resistance and capacitance. CMP fill insertion can change both coupling and total capacitance of interconnect. In addition, metal dishing and dielectric erosion change interconnect cross section and, therefore, affect interconnect resistance. He *et al.* [18] report an increase of more than 30% in interconnect resistance due to dishing and erosion, while the impact on interconnect capacitance is insignificant. He *et al.* [19] also propose a wire-sizing approach to lessen the amount of interconnect-resistance variation due to the CMP process. Increased wire size compensates for the increased resistance caused by dishing and erosion and also reduces the effect of the large R_{eff} (i.e., driver-output resistance) variation on delay.

1) CMP Fill Pattern: CMP fill insertion, even as it contributes to layout pattern-density uniformity, increases the coupling and total interconnect capacitance. Therefore, it is important to assess the impact of CMP fill on interconnect capacitance to reduce the uncertainty in circuit-timing calculations. He *et al.* [17] explore a space of different fill patterns that are equivalent from the foundry perspective (i.e., respecting all the minimum design rules, etc.) and their respective impact on interconnect capacitance. All the fill features are assumed to be rectangular and are aligned horizontally and vertically, as shown in Fig. 3. Using the notation from [17], conductors A and B are active interconnects, and the metal shapes between them are CMP fills. Each distinct fill pattern is specified by the following: 1) the number of fill rows (M) and columns (N); 2) the series of widths $\{W_i\}_{i=1,\ldots,N}$ and lengths $\{L_j\}_{j=1,\ldots,M}$ of fills; and 3) the series of horizontal and vertical spacings $\{S_{x,i}\}_{i=1,\ldots,N-1}$ and $\{S_{y,j}\}_{j=1,\ldots,M-1}$ between fills.

Enumeration of all the possible combinations of the aforementioned parameters is not feasible. Therefore, to restrict the space of exploration, He *et al.* [17] propose a positive-distribution-characteristic function, denoted as $f(k)$, where k is an integer variable that takes the index of the element in the series. For example, the value of the ith element of the width is calculated as $W_i = f(i) + \overline{W_l}$, where $\overline{W_l}$ is the minimum-width design rule. Based on the results of the experiments, He *et al.* [17] propose two guidelines as to what a "good" fill pattern might be among all the possible valid fill-pattern combinations. The criteria for this assessment are based on the impact of the pattern on interconnect capacitance. According to these guidelines: 1) in a fixed-length budget, the number of fill columns should be maximized; and 2) in a fixed-width budget, the number of fill rows should be minimized.

In addition to the parameters covered in the previous experiments, Gupta *et al.* [15] add four more parameters in its space of exploration. These parameters are metal width, metal height, dielectric constant, and keepoff distance. The trend of the changes in interconnect capacitance was observed for the corresponding parameters. A recent work by Kahng *et al.* [30] systematically studies the impact of various floating-fill configuration parameters, such as fill size, fill location, interconnect size, separation from interconnect edges, multiple fill columns and rows, etc., on coupling capacitance. On the basis of their studies, Kahng *et al.* [30] propose certain guidelines for fill insertion to reduce their impact on coupling capacitance while achieving the prescribed metal density.

Fig. 4 shows an application of the proposed guidelines for a represented fill/wire configuration. Increase in coupling capacitance is 27%, and 11% when fill is inserted in a regular pattern and with the proposed guidelines, respectively. Kahng *et al.* [30] report that, on average, 53% reduction in coupling-capacitance increase is achieved through the application of the guidelines for fill insertion.

In addition to the aforementioned guidelines, different fill shapes have been proposed. Fill patterns should be devised such that all long conductors on adjacent layers have identical

[2] Recent full-chip 3-D extraction tools support floating-fill extraction. Some of these tools, however, implicitly assume regular fill patterns and large buffer distances. Reliable extraction of floating fill arranged in arbitrary patterns is still, to our knowledge, a topic of active research.

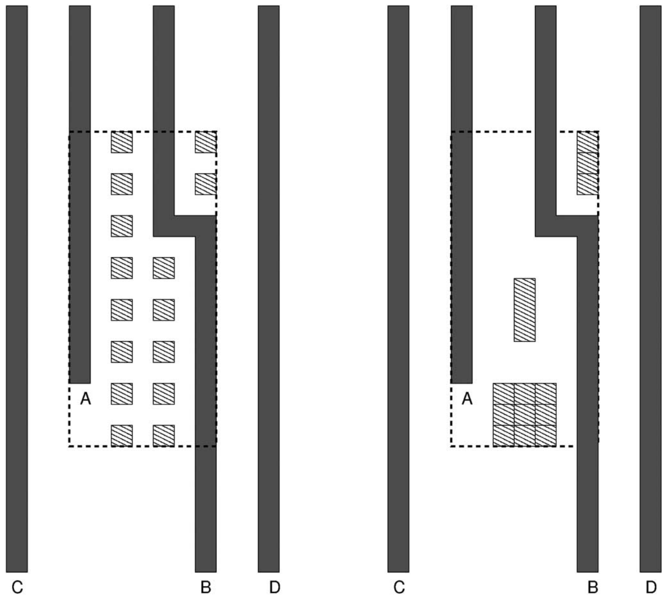


Fig. 4. (a) Regular fill pattern. (b) Fill insertion with guidelines.

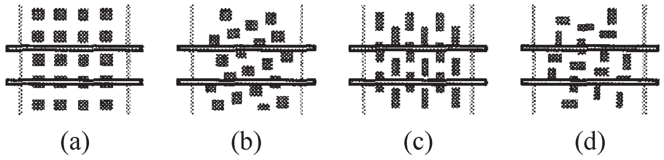


Fig. 5. Various patterns of CMP fill. (a) Traditional. (b) Staggered. (c) Alternate rectangles. (d) Basket-weave [34].

coupling capacitance to the inserted fill [Fig. 5(b)–(d)]. Several practical ways of achieving this is to use “staggered,” “alternate rectangles,” and “basket-weave” fill patterns, as shown in Fig. 5 [29]. In other words, the fill pattern should not consist of regular grid geometries but, instead, have some internal offsets that “skew” the pattern.

Among the aforementioned patterns, “staggered” is more favorable due to its impact on microloading. Microloading is the relation between local-etch rate and pattern density. Staggered fill pattern tends to make the fluid dynamics more uniform and, hence, reduce the competition for etch reactants leading to a more uniform etch rate. In addition, reflectance from the underlying layer of metal is a source of flare and, in general, causes lithographic variation. This provides a similar rationale for staggering as the goal of uniform coupling of fill to conductors on adjacent layers.

Fill today is hardly “rectangular.” There is a wide range of fill shapes, often inserted by multipass heuristics (e.g., litho protection “track fill,” then large rectangles, then medium rectangles, then small rectangles). Fill shapes are distinguished by whether they require optical-proximity correction (OPC) or no OPC; typically, any fill shape within 0.3 μm of a functional wire may be a candidate for OPC treatment; remaining fill shapes can be “raw.” Another reason that fill shapes are nonrectangular in practice is that large default shapes must often be subjected to “cutouts” due to actual layout features. For example, the end of a 0.2-μm-wide wire, with a fill-to-wire keepoff distance of 0.3 μm, will create a 0.8-μm notch in the fill shape.

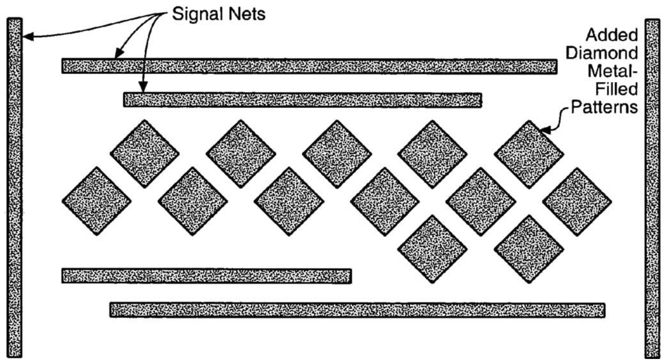


Fig. 6. Diamond CMP-fill features [25].

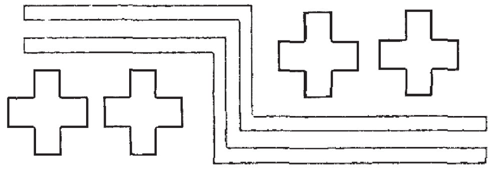


Fig. 7. Plus-sign pattern fill [46].

Hung [25] also presented diamond CMP fill patterns to reduce the effective coupling capacitance. The diamond fill features are square shapes, which are oriented at 45° from the metal lines, as shown in Fig. 6.

In addition to the aforementioned rectangular patterns, Nelson [46] proposes the use of plus-sign-shaped fill features (Fig. 7). Plus-sign fill patterns fill approximately as effectively as the square patterns but have significantly less capacitive impact, particularly intralayer.

2) CMP Fill and Interconnect Capacitance: CMP fill features, despite their role in uniforming layout pattern density, have a significant impact on coupling and total interconnect capacitance. There is a body of work that addresses different issues regarding the estimation or optimization of the capacitance impact of the CMP fill.

Park *et al.* [49] briefly describe a model-library-based approach to extract floating fill. Results demonstrating the accuracy of the approach and characterization time are, however, not presented. Lee *et al.* [40] present a methodology for full-chip extraction of total capacitance in the presence of floating fill and [39] extended their analysis. Their approach adjusts the permittivity and sidewall thickness of dielectric to account for the capacitance increase due to fill. According to Kurokawa *et al.* [34], the capacitance of a configuration is directly proportional to the charge accumulated on one of the electrodes $(Q = CV)$. The charge density on an electrode depends on the electric field close to the electrode $(E = \sigma/A)$. Therefore, the electric field close to an electrode determines the capacitance of a configuration. When a floating plate of thickness $t\,(t < d)$ and the same size as the conductor plates is inserted in the space between the conductors, the capacitance increases to $\varepsilon A/(d - t)$, where d is the distance between the conductor plates.

In addition, Batterywala *et al.* [2] propose an extraction methodology, where fills are eliminated one by one using a graph-based random-walk algorithm while updating the coupling capacitances. In this method, a network of capacitors

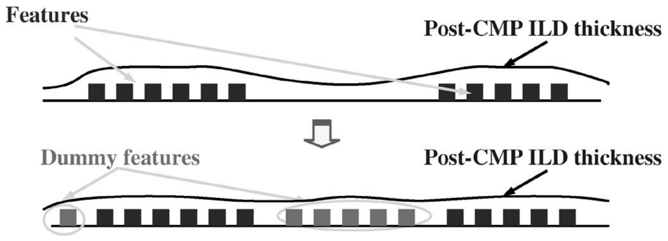


Fig. 8. Insertion of CMP fill to decrease post-CMP variation of ILD thickness [24].

is collapsed into the equivalent capacitance between two nets. In addition, Yu *et al.* [70] propose enhancements to the current-field solvers by taking into account floating fills and their conditions in the direct-boundary-element equations. The basic idea in their approach is to add additional equations about the floating-CMP fill features to generate a solvable system of linear equations. In the conventional approach, the field solver is called as many times as the number of conductors and floating fill features, whereas in the proposed method, the field solver is only called as many times as the number of conductors. Hence, the proposed method significantly reduces the computation runtime of the field-solving process as compared to traditional methods. Chang *et al.* [6] present a charge-based capacitance-measurement methodology to analyze the impact of fills. Recently, Kim *et al.* [32] propose guidelines that simplify the development of compact analytical models for capacitance increments. Their capacitance-increment models are a function of density, design rules, metal thickness, and the amount of CMP fill in the neighboring layers. Based on the experimental results, these models are very fast and accurate as compared to current methods built into parasitic extractors. Kurokawa *et al.* [33] propose three techniques of fill insertion in order to reduce the interconnect capacitance and the number of fills inserted. It also provides an estimation of the required number of fill geometries for each of the proposed techniques. However, it fails to report the accuracy and reliability of the methods and estimations for densities greater than 30%. Finally, to reduce the inaccuracies in floating fill capacitance extraction, Kahng *et al.* [72] propose a set of design of experiments (DOE), which will be used in addition to what is available in the extraction tools for regular interconnects and grounded fills. The proposed DOE, enables extensive analyses of the CMP fill impacts on coupling capacitance.

C. Data-Volume Considerations

Post-CMP planarization quality depends on the density of the features in the layer beneath. The surface of a particular layer will not be flat unless the layout geometries in the previous layer of material exhibit uniform spatial density, i.e., every "window" of given size in the chip layout contains roughly the same total area of copper wiring. Therefore, millions of "CMP fill" features (typically, small squares on a regular grid) are introduced into sparse regions of the layout to equalize the spatial density, as shown in Fig. 8. The disadvantage is that the layout-data-file size is dramatically increased. A small layout data file is desired to quickly transmit the chip design to the manufacturing process.

TABLE I
TERMINOLOGY

Symbol	Description
B	fill data file, an m by n binary matrix
D	data block, a b_1 by b_2 sub-matrix of B
R	reference block, a b_1 by b_2 binary matrix
C	cover block, a b_1 by b_2 binary matrix
$\mathcal{C}$	cover, a set of C's "close" to a set of data blocks
$s(D)$	# of 1's in D
$w(D)$	# of bits of D allowed to change from 1 to 0
k, f	proportional loss ratio, fixed speckle loss
m,n	# of rows, columns of B
b_1,b_2	# of rows, columns of D, R, C
h	Total size of the compressed items ($h = \sum_x O_x$)

Kahng *et al.* [24] propose several heuristic algorithms with the general outline, as shown in Algorithm 1. The terminology used is summarized in Table I. A layout containing CMP fill features can be expressed as a binary (0–1) matrix.[3] CMP fill compression takes as input an $m \times n$ binary matrix B and outputs compressed items (D's, R's, I, etc.) with total size h. The objective of fill compression is to minimize h. The compression ratio is defined as $r = mn/h$. Kahng *et al.* [24] develop a number of compression heuristics based on Joint Bilevel Image Processing Group methods [20], [22], [23]. The algorithms can be lossless when fill must appear exactly as specified or lossy when a "close" approximation of the fill-pattern suffices. Different loss tolerances can be applied according to the application context. Lossy-compression algorithms allow both proportional loss, a prescribed upper bound on the fraction of ones in a single data block which may be changed to zeros, and fixed loss, a prescribed upper bound on the absolute number of ones in a single data block which may be changed to zeros. In the former context, for a given data block D, most $w(D) = \lfloor k \cdot s(D) \rfloor$ of its ones can be changed to zeros.

```
Algorithm 1 (General compression scheme):
1) Segment data matrix $B$ into blocks $D$.
2) If lossy compression is desired,
    (a) generate a cover $\mathcal{C}$ for data blocks $D$ (every $D$ must match, modulo possible loss, a cover block $C \in \mathcal{C}$), and
    (b) replace $B$ with lossy matrix $B'$ by replacing each data block $D$ with its matching cover block $C$.
3) Perform lossless compression on $B$.
```

Data-volume reduction is also achieved by, e.g., multipass fill that inserts large arrayed shapes first, then "fills in the cracks" with smaller shapes, etc. The OASIS data standard [11] facilitates data-volume reduction by proving compression operators that are specifically targeted at compression of "tilted" or staggered fill patterns.

III. Layout Density-Analysis Methods

Traditionally, only foundries have performed the post-processing needed to achieve pattern-density uniformity using

[3] This is true of all major commercial fill-insertion tools, such as Mentor Calibre, Avant! Hercules, and Cadence Assura, even when operating in modes that output "tilted fill" or tiled "fill cells."

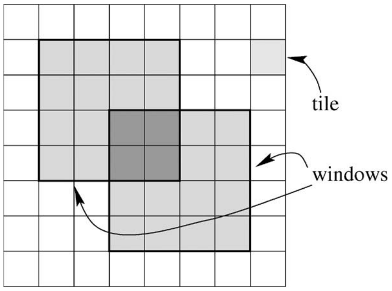


Fig. 9. Layout is partitioned by $r^2 (r = 4)$ fixed dissections into $(nr/w) \times (nr/w)$ tiles. Each $w \times w$ window (light gray) consists of r^2 tiles. A pair of windows from different dissections may overlap [27].

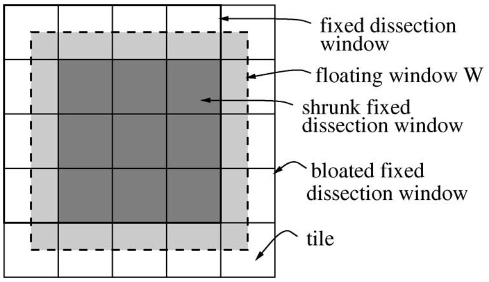

fixed dissection window
floating window W
shrunk fixed dissection window
bloated fixed dissection window
tile

Fig. 10. Any floating $w \times w$-window W always contains a shrunk $(r-1) \times (r-1)$-window of a fixed r-dissection and is always covered by a bloated $(r+1) \times (r+1)$-window of the fixed r-dissection [28].

insertion "filling" or partial deletion "slotting" of features in the layout [26]. However, layout pattern density must be calculated prior to addressing the filling or slotting problem. Regions that are violating the lower and upper area density bounds are identified using density-analysis methods.

A. *Fixed Dissection Versus Continuous*

To verify (or enforce) upper and lower density bounds for $w \times w$ windows, a very practical method is to check (or enforce) these constraints only for $w \times w$ windows of a fixed dissection of the layout into $(w/r) \times (w/r)$ tiles, i.e., the set of windows having top-left corners at points $(i \cdot (w/r), j \cdot (w/r))$, for $i, j = 0, 1, \ldots, r((n/w) - 1)$, as shown in Fig. 9. Here, r is an integer divisor of w.

To analyze all the eligible $w \times w$ windows takes a significant amount of time, while the analysis of fixed dissections can be done much faster. Simply, an array of $(n/w) \times (n/w)$ counters will be associated with all the dissection windows, and then for each rectangle R, the counters of windows intersecting R will be incremented by the area of intersection. In general, the aforementioned procedure must be repeated r^2 times to check all the $(r \cdot (n/w))^2$ windows [26].

Today's standard fill approaches verify density constraints only for windows in a "fixed dissection" of the layout region (e.g., 100-μm windows shifted by multiples of 50 μm). Unfortunately, density constraints can be violated by other 100-μm windows within the layout. The discrepancy between "fixed dissection" and "continuous" density control is exactly characterized in the following two theorems [29].

Theorem 1: Suppose all $(w/r) \times (w/r)$ fixed r-dissection tiles with bottom-left corners at points $(i \cdot (w/r), j \cdot (w/r))$, $i, j = 0, 1, \ldots, r((n/w) - 1)$, have an area density of at least L and at most U. Then, the exact lower bound on the area density of $w \times w$ windows equals

$$\frac{(r-1)^2}{r^2} \cdot L + \frac{4(r-1)}{r^2} \max\{L - 0.5, 0\} + \frac{4}{r^2} \max\{L - 0.75, 0\}$$

and the exact upper bound equals

$$\frac{(r+1)^2}{r^2} \cdot U - \frac{4(r-1)}{r^2} \max\{U - 0.5, 0\} - \frac{4}{r^2} \max\{U - 0.25, 0\}.$$

Theorem 2: Suppose all $w \times w$-sized windows with bottom-left corners at points $(i \cdot (w/r), j \cdot (w/r))$, for $i, j = 0, 1, \ldots, r((n/w) - 1)$, have an area density of at least L and at most U. Then, any $w \times w$ window has density at least $L - (1/r) + (1/4r^2)$ and at most $U + (1/r) - (1/4r^2)$, and these bounds are tight.

In follow-up to Theorem 2, Lin [43] meticulously partitioned the general case into subcases and found better bounds for some of them, leaving the global bound the same. From a mathematical standpoint, Lin [43] made a reasonable analysis of subcases although incorrectly claimed that the [29] bound was not tight. From the engineering point of view, it is not clear if one needs such analysis at all.

In a recent work, Xiang *et al.* [68] propose a fast exact algorithm for density calculation. In their work, they first present the traditional fixed-dissection method limitation. Then, using a series of data preprocessing and efficient data structures, they propose a fast algorithm, on the basis of fixed-dissection method, for density calculation.

B. *Multiwindow and Multilayer Density Analysis*

Even though the fixed-dissection analysis can be performed quickly, it can underestimate the maximum floating-window density worst case.[4] Kahng *et al.* [28] propose a new multilevel density-analysis approach which, as opposed to the techniques presented in [26] and [27], has the efficiency of the fixed-dissection analysis without sacrificing the accuracy for the floating-window worst case analysis. The multilevel density analysis is based on the following simple observation.

Observation: Given a fixed r-dissection, any arbitrary floating $w \times w$ window will contain some shrunk $w(1 - 1/r) \times w(1 - 1/r)$ window of the fixed r-dissection and will be contained in some bloated $w(1 + 1/r) \times w(1 + 1/r)$ window of the fixed r-dissection, as shown in Fig. 10.

The first implication of the aforementioned observation is that the floating-window area can be upper bounded by the area of bloated windows and lower bounded by the area of shrunk windows. A fixed r-dissection regime can be recursively subdivided into smaller dissections until the number of tiles in each dissection is small. Then, the floating density analysis can be applied without significant runtime complexity. In addition, the recursion can be terminated once the floating density analysis is within some user-defined criteria, for example, $\varepsilon = 1\%$ [28].

[4] In general, when all the eligible windows are being examined and filled, it is referred to as the floating-window regime.

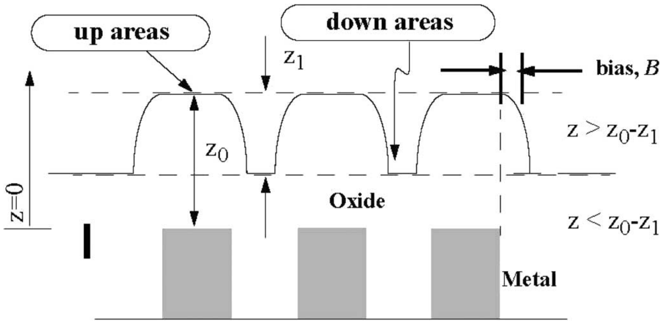


Fig. 11. Dishing and erosion in copper CMP process [47].

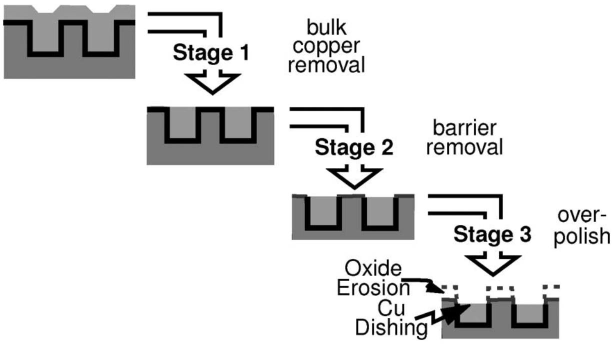


Fig. 12. Three intrinsic stages in copper CMP processes [64].

IV. BEOL CMP Fill

A. BEOL CMP Characterization and Modeling

1) Oxide CMP: Pattern density is a significant contributor to oxide-CMP-process quality. The Preston equation shows that the material-removal rate is a linear function of the pressure, which is affected by the pattern density at the interface between polishing pad and wafer. However, pattern-density calculation is not trivial. In fact, the effective density at a particular point on the die depends on the size of the neighboring area over which density is averaged. The weighting function is also a major factor since it captures the influence of the surrounding area on the local pressure.

Modeling of CMP for oxide planarization is reduced to accurately calculate the local pressure and, hence, the pattern-density distribution across every die [47]. As described in the previous section, there are several models that have been proposed to account for pattern effects in CMP, but their applicability has been limited.

The basic model in [47] is based on the work by Stine *et al.* [58]. In this model, the ILD thickness z at location (x, y) is calculated as

$$z = \begin{cases} z_0 - \left(\frac{Kt}{\rho_0(x,y)}\right), & t < (\rho_0 z_1)/K \\ z_0 - z_1 - Kt + \rho_0(x,y)z_1, & t > (\rho_0 z_1)/K. \end{cases} \quad (1)$$

The constant K is the blanket wafer-removal rate (i.e., where the density is 100%). The important element of this model is the determination of the effective initial pattern density $\rho_0(x, y)$. Fig. 11 defines the terms used in (1).

In (1), when $t < (\rho_0 z_1)/K$, the local step height has not been completely removed. However, when features are planarized for a long enough time ($t > (\rho_0 z_1)/K$), the local step height is completely removed and a linear relationship between pattern density and ILD thickness exists [58].

The planarization length, which captures pad deformation during the CMP process, determines the amount in which neighboring features affect pattern density at a spatial location on the die. Thickness profile of any arbitrary mask pattern, under the same process conditions, can be determined using the effective local density and an analytic thickness model. This reduces the characterization step into a single phase where only the planarization length of the process is determined. Planarization length is also a useful metric in oxide-CMP-process optimization since it reduces the investigation of the entire die to smaller regimes according to the planarization length [47].

Ouma [47] proposes a characterization methodology for oxide-CMP processes that includes the following: 1) the use of an elliptic pattern-density-weighting function which has better correspondence to the polish-pad deformation; 2) a three-step effective pattern-calculation scheme, which uses fast Fourier transforms for computational efficiency; and 3) the use of layout masks with step densities which facilitate the determination of the characteristic length (defined as the planarization length) of the elliptic function by introducing large abrupt post-CMP-thickness variations.

2) Copper CMP: Unlike oxide CMP which only involves the removal of oxide material, the copper CMP involves simultaneous polishing of three materials: copper, dielectric (oxide), and barrier. Barrier is a very thin layer (Tan, Ti, etc.) that prevents the copper from diffusing into the dielectric. The goal in copper CMP is to remove the excess copper (also called overburden copper) and to polish the barrier on top of the dielectric regions, isolating the adjacent interconnect lines. This is required to prevent electrical connection between adjacent interconnect lines. Due to the heterogeneous nature of copper CMP, a specific set of process parameters, as well as a consumable set, are required to achieve the particular removal rate for each corresponding material [64].

The goal in copper CMP is to remove the excess copper and the unwanted barrier layer. Ideally, this process should be fast without incurring extra dishing, erosion, or other defects. Due to heterogeneous nature of copper CMP, different materials are polished simultaneously. Initially only overburden copper is polished, followed by the polishing of both copper and barrier film. Finally, copper, barrier, and dielectric are polished at the same time. As stated in [64], to model a copper-CMP process, three stages of polish are identified: excess-copper removal, barrier-film removal, and overpolish stage, as shown in Fig. 12. In the excess-copper-removal stage, the evolution of the copper-thickness profile across the chip and the time it takes to remove the excess copper are of interest. The time to polish the overburden copper varies across the die depending on the pattern density at the location of interest.

In the second stage, copper and barrier film are polished simultaneously. The time to clear the barrier film, as well as the dishing that results when barrier is removed at any location on the die, are of interest. Due to process and deposited-copper-thickness variations across the wafer and different pattern densities across the die, the removal rates of the three materials (copper, barrier, and dielectric) are different. This difference in removal rates results in different polish times across the wafer for each stage. For example, by the time the excess copper and barrier are cleared at a point on the die, they might have already been cleared at another point. Hence, some points on the die are overpolished. In copper CMP, overpolishing is defined as polishing beyond the time it takes to remove the overburden copper and barrier at any spatial location. During the overpolishing stage, the dielectric is eroded [64].

In addition, the dishing that might have started during the barrier-clearing stage can worsen during overpolishing. This overpolishing is identified as the third intrinsic stage in the copper-CMP process. The dishing and erosion that occur during this stage are of interest. In computing the amount of dishing during the overpolish stage, the dishing that occurs during the barrier-clearing stage is used as an initial condition. It is important to note that the term overpolishing is used loosely in the CMP literature, and in the CMP industry [64].[5]

B. Density-Driven Fill Synthesis

1) Linear-Programming (LP)-Based and Monte-Carlo-Based Methods: Kahng *et al.* [27] propose the first min-variation formulation using an LP approach. In a fixed r-dissection regime, for any given tile $T = T_{ij}, i, j = 1, \ldots, (nr/w)$, the total feature area inside T and the maximum fill amount that can be placed within T without violating the density upper bound U in any window containing T are denoted as area(T) and slack(T), respectively. The following is the filling problem as described in [27].

Filling Problem for Fixed r-Dissection: Suppose a fixed r-dissection of the layout with tiles of size $(w/r) \times (w/r)$, as well as an area(T) and slack(T) for each tile in the dissection. Then, for each tile T_{ij}, the total fill-pattern area $p_{ij} = p(T_{ij})$ to be added to T_{ij} must satisfy

$$0 \leq p_{ij} \leq \text{slack}(T_{ij})$$

and

$$\sum_{T_{ij} \in W} p_{ij} \leq \max\left\{U \cdot w^2 - \text{area}(W), 0\right\} \tag{2}$$

for any fixed-dissection $w \times w$-window W.

Then, the min-variation formulation seeks to maximize the minimum window density

$$\text{maximize}\left(\min_{ij}(\text{area}(T_{ij}) + p_{ij})\right).$$

The LP approach seeks the optimum fill area $p(T_{ij})$ to be inserted into each tile T_{ij}. Recall that the fill area $p(T_{ij})$ cannot exceed slack(T_{ij}), which is the area available for filling inside the tile T_{ij} computed during density analysis. The auxiliary variable M is the lower bound on all window densities. The first LP for the min-variation objective [27], [29] is

$$\begin{aligned}
&\text{Maximize}: M \\
&\text{subject to}: \\
&\qquad 0 \leq p(T_{ij}) \leq \text{slack}(T_{ij}) \\
&\qquad M \leq \rho(M_{ij}) \leq U, \qquad i, j = 1, \ldots, \frac{nr}{w} - 1.
\end{aligned}$$

An important step in the aforementioned LP approach is to determine slack values. To calculate the total area of all the possible overlapping rectangles, the approach of "measure of union of rectangles" sweep-line-based technique [51] has been used. In a follow-up work, Tian *et al.* [62] proposed the following min-fill LP formulation:

$$\begin{aligned}
&\text{Minimize}: \quad \sum_{i,j} p(T_{ij}) \\
&\text{subject to}: \\
&\qquad 0 \leq p(T_{ij}) \leq \text{slack}(T_{ij}) \\
&\qquad L \leq \rho(M_{ij}) \leq U, \qquad i, j = 1, \ldots, \frac{nr}{w} - 1.
\end{aligned}$$

Tian *et al.* [62] incorporate the elliptical weighting function in the planarization model of Ouma [47], which has better correspondence to the pad deformation. In addition, the CMP process is modeled as a low-pass filter through which the local pattern density not only contributes to immediate but also short-range ILD thickness within the weighting region. Tian *et al.* [62] also observe that manufacturability does not require the extreme min-variation formulation, i.e., given a target-window density M, a variability budget ε can be minimized

$$\begin{aligned}
&\text{Minimize}: \quad \varepsilon \\
&\text{subject to}: \\
&\qquad 0 \leq p(T_{ij}) \leq \text{slack}(T_{ij}) \\
&\qquad M - \varepsilon/2 \leq \rho(M_{ij}) \leq M + \varepsilon/2, \qquad i, j = 1, \ldots, \frac{nr}{w} - 1.
\end{aligned}$$

The aforementioned formulation is an example of the min-fill formulation where the manufacturability is guaranteed by the constraints, and the total amount of CMP fill inserted is minimized by the objective.

In addition to the LP approaches, Chen *et al.* [10] introduce the Monte Carlo method for min-variation objective. In the Monte Carlo approach, a tile is chosen randomly and its content is incremented with a predetermined fill amount. Tiles are chosen based on their priority, which is the probability of choosing a particular tile T_{ij}. The priority of a tile T_{ij} is zero if and only if either T_{ij} belongs to a window which has already achieved the density upper bound U or the slack of T_{ij} is equal to the already-inserted fill area. As described in [10], the priority of a tile T_{ij} is chosen to be proportional to $U - \text{MinWin}(T_{ij})$, where $\text{MinWin}(T_{ij})$ is the minimum density over windows containing the tile T_{ij}. The only drawback of the Monte Carlo method is that it may insert an excessive amount of total fill.

[5] In the CMP industry, overpolishing means polishing beyond the endpoint time.

A variant of the Monte Carlo approach is the greedy algorithm. At each step, the min-variation greedy algorithm adds the maximum possible amount of fill into a tile with the highest priority which causes the priority of that particular tile to become zero.

In the presence of two objectives, namely, min-variation and min-fill, the intuitive approach would be to first find a solution that optimizes one of the objectives then modify the solution with respect to the other objective. The min-fill objective tries to delete as much of the previously inserted fill as possible while maintaining the density criteria. Min-variation and min-fill can be thought of as objectives for manufacturability and design, respectively.

To optimize the min-fill objective with the Monte Carlo approach, a fill geometry from a tile randomly chosen according to a particular priority is iteratively deleted. Priorities are chosen symmetrical to the priority in the min-variation Monte Carlo algorithm, i.e., proportional to $\mathrm{MinWin}(T_{ij}) - L$. Again, symmetrically, no filling geometry can be deleted from the tile T_{ij} (i.e., T_{ij} is locked) if and only if it either has zero priority or else all fill previously inserted into T_{ij} have been deleted. Thus, the min-fill Monte Carlo algorithm deletes fill geometries from unlocked tiles, which are randomly chosen according to the aforementioned priority scheme. Similarly, the min-fill greedy algorithm iteratively deletes a filling geometry from an unlocked tile with the current highest priority.

A variant of the Monte Carlo approach is the deterministic greedy algorithm where, at each step, the greedy min-variation algorithm adds the maximum possible amount of fill into a tile with the highest priority. The runtime for this approach is slightly higher than Monte Carlo because of finding the highest priority tile rather than random ones [8].

2) Iterated Monte Carlo and Hierarchical Methods: Both Monte Carlo and greedy approaches are suboptimal for the min-variation objective resulting in a minimum window density that may be significantly lower than the optimum. Chen *et al.* [8] propose a new iterative technique, alternating between the min-variation and min-fill objectives, to narrow the gap between the upper window density bound U and the minimum window density bound L. As described in [8], the iterated Monte Carlo and greedy filling algorithms are modified as follows: 1) Interrupt the filling process as soon as the lower bound L on window density is reached, i.e., when $M = L$, instead of improving the minimum window density (while possible) for the min-variation objective. 2) Continue iterating but without changing the lower density bound $M = L$. An improved solution can typically be obtained by keeping track of the best solution observed over all iterations. All the filling methods aforementioned were proposed for flat designs. However, the filling problem for hierarchical (standard cell) layouts is similar to the problem for flat layouts. The constraints for the hierarchical filling problem [9] are as follows.

1) Filling geometries are added to master cells.
2) Each cell in a filled layout is a filled version of the original master cell.
3) Layout data volume should not exceed a given threshold.

The proposed method by Chen *et al.* [9] first computes the slack value for all the master cells (i.e., the original copy of the cell within the library and not an individual instance). Then, a keep-off distance around master cells is created to avoid overfilling the regions near master-cell boundaries. Then, master cells are filled using a Monte Carlo method where master cells that are more underfilled are assigned a higher priority. This process is continued until either all the master cells are filled above their minimum density lower bound or the slack in the underfilled master cells becomes zero.

However, due to overlaps between different instances of master cells and features or the interactions among the "bloat" regions in the vicinity of the master cells, pure hierarchical filling may result in some sparse or unfilled regions. This could result in high-layout-density variation. An intuitive solution would be to apply a postprocessing phase, i.e., apply a standard flat-fill approach. However, this will greatly increase the resultant data volume and runtime and diminish the benefit of the hierarchical approach. Chen *et al.* [9] propose a three-phase hybrid hierarchical flat-filling approach as follows:

1) a purely hierarchical phase;
2) a split-hierarchical phase, where certain master cells that were considered underfilled in phase 1) would be replicated so that distinct copies of a master cell may be filled differently than other copies of the same master cell;
3) a flat-fill "cleanup" phase (i.e., LP, Monte Carlo, etc.), which will fill any remaining sparse or underfilled regions which were not satisfactorily processed during the first two phases.

3) Timing-Driven Fill Synthesis: One of the largest concerns in fill synthesis, apart from meeting the CMP-design rules, is the impact of fill insertion on the interconnect capacitance. An excessive increase in wire capacitance can cause a net to violate its setup-timing constraint. A large value for keepoff distance (i.e., minimum distance from fill to wire) reduces the impact but it erodes into available areas to insert fills and sometimes makes it impossible to meet the minimum-density constraint. Chen *et al.* [7] propose the first formulation of the performance-impact-limited-fill (PIL-Fill) problem with the objective of either minimizing total-delay impact or maximizing the minimum slack of all nets, subject to a given predetermined amount of fill. They also developed simple capacitance models to be used in their delay calculations. The PIL-Fill synthesis formulation has two objectives: 1) minimizing layout-density variation and 2) minimizing the CMP fill features' impact on circuit performance (e.g., signal delay and timing slack). Since it is difficult to satisfy both objectives simultaneously, practical approaches tend to optimize one objective while transforming the other into constraints. Using the terminology in [7], the two problem formulations proposed are as follows (note that these formulations are for fixed-dissection regimes).

1) Given tile T, a prescribed amount of fill is to be added into T, a size for each fill feature, a set of slack sites (i.e., sites available for fill insertion) in T per the design rules for floating square fill, and the direction of current flow and the per-unit length resistance for each interconnect

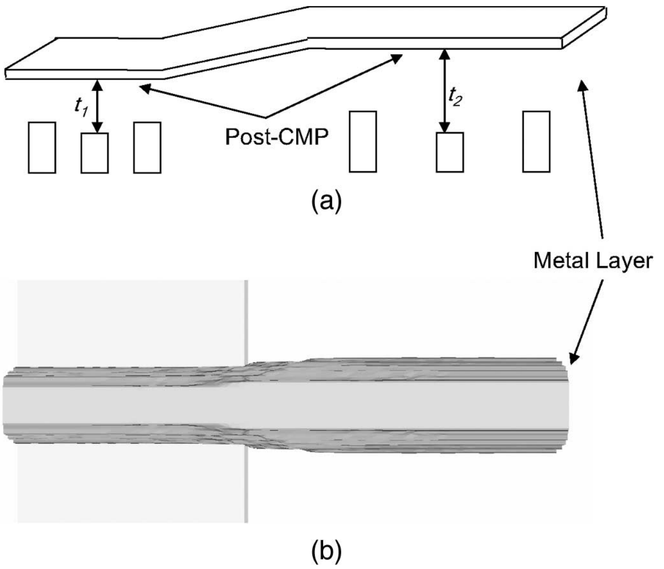


Fig. 13. (a) Side view showing thickness variation over regions with dense and sparse layout. (b) Top view showing CD variation when a line is patterned over a region with uneven wafer topography, i.e., under conditions of varying defocus [16].

segment in T; insert fill features into T such that total impact on delay is minimized.

2) Given a fixed-dissection routed layout and the design rule for floating-square fill features, insert a predetermined amount of fill in each tile such that the minimum slack over all nets in the layout is maximized.

The first formulation corresponds to minimum delay with fill-constrained formulation, while the second one is the maximum min-slack with fill-constrained formulation. A weakness with the first formulation is that it minimizes the total-delay impact independently for each tile. Hence, the impact due to fill features on signal delay of the complete timing path is not considered. The second formulation, therefore, has been proposed to alleviate this problem by maximizing the minimum slack of all nets, subject to a constraint of inserting a predetermined amount of fill in every tile of the layout. Chen *et al.* [7] propose two integer LP methods and a greedy approach for the minimum delay and maximum min-slack formulations, respectively. However, the capacitance models used in delay calculations are not accurate, as they do not consider the presence of fill features on the neighboring layers. This incurs inaccuracy in the estimated capacitance values and eventually causes uncertainty in the timing analysis. In addition, they do not account for signal-flow direction which causes layout nonuniformity (i.e., as fills are pushed to the receiver edge, the driver edge becomes less dense).

In a recent work, Topolaglu [71] proposes a CMP fill synthesis framework which uses coupling-aware fill insertion guidelines to reduce the impact of CMP fill shapes on the design's performance. In another work, Xiang *et al.* [67] propose a coupling-aware constrained CMP-fill analysis algorithm which identifies feasible locations for fill features such that the fill-induced coupling capacitance can be bounded within the given coupling threshold of each wire segment. The algorithm also utilizes ground capacitance for more robustness and predictability.

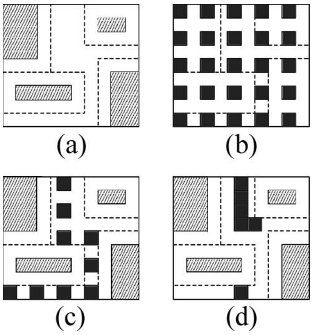


Fig. 14. (a) Example layout with features lightly shaded and exclusion zone in dashed lines. (b) 25% density fill insertion prior to Boolean operations. (c) Rule-based fill insertion after the application of the Boolean operations. (d) Possible model-based fill insertion [61].

4) Multipurpose Fill Synthesis: In this section, we review two other fill-synthesis techniques that satisfy nontraditional objectives: wafer topography and IR-drop reduction.

Wafer Topography: DOF is the major contributor to lithographic-process margin. One of the major causes of focus variation is imperfect planarization of fabrication layers. Currently, OPC methods are oblivious to the predictable nature of focus variation arising from wafer topography. As a result, designers suffer from manufacturing-yield loss, as well as loss of design quality through unnecessary guardbanding. Fig. 13 shows how post-CMP-thickness variation results in loss of CD control. Fig. 13(a) shows how post-CMP thickness in copper-oxide polishing will predictably change with the region pattern density. The DOF variation corresponding to the thickness variation severely affects metal patterning of the subsequent upper layer, as shown in Fig. 13(b). In this figure, t_1 and t_2 are post-CMP-thickness variations over dense and sparse regions, respectively. Hence, to minimize the impact of pattern-dependent effects of the CMP process, the OPC methods should be aware of the post-CMP topography to assign appropriate defocus value for all the features with the same topography. In a recent work, Gupta *et al.* [16] propose a flow and methodology to drive OPC with a topography map of the layout that is generated by CMP simulation. The experimental results showed that the proposed topography-aware OPC can yield up to 67% reduction in edge-placement errors at the cost of little increase in mask cost.

a) IR-Drop: An increasing challenge in 90-nm (and beyond) designs is the increase in IR-drop. The tolerance for IR-drop is becoming smaller as the voltage source scales. If the wire resistance is too large or the cell current is too high, a significant voltage drop may occur, leading to timing-degradation or signal-integrity issues. In the worst case, the voltage drop may be large enough that transistors fail to switch correctly, resulting in catastrophic chip failure. In his work, Leung [41] addresses these issues by utilizing CMP fill to simultaneously satisfy metal-density requirements and reduce IR-drop of the power network. According to the experimental results, the proposed method achieves an average IR-drop reduction of 62.2%.

C. Model-Based Fill Synthesis

Methods for fill insertion can be categorized into two groups: rule- and model-based. Rule-based fill synthesis is based on concepts such as density or keepoff distance rules, which are applied to wiring segments that have less than certain threshold amounts of timing slack. Model-based fill-insertion approach is based on analytical expressions that define the relationship between local pattern density and ILD thickness. Fig. 14 shows possible rule- and model-based fill-insertion approaches. In addition, model-based fill synthesis, on the other hand, would use CMP models to identify regions where planarity is important (next to heavily loaded critical segments and below-critical segments). The model-based approach has implicit tight coupling to a timer and models the impact of fill on coupling capacitance.

The model-based fill-insertion approach, given a CMP-process model, is to find the amount and the location of the fill features to be inserted in the layout so that certain electrical- and physical-design rules are preserved and certain post-CMP-topography variation is met. Tian *et al.* [61] propose a two-step solution with consideration of both single- and multiple-layer layouts in the fixed-dissection regime. The first step uses LP to compute the necessary amount of fill to be inserted in each of the dissection's tiles. In the second step, the amount of fill calculated by the first step will be placed into each tile such that certain local properties (i.e., electrical, physical, etc.) are preserved. Experimental results with the single-layer formulation (i.e., the cumulative variation of underlying layers is ignored) show reduction of post-CMP-topography variation from 767 to 152 Å.

V. FEOL CMP FILL

A. STI CMP Characterization and Modeling

STI is the isolation technique of choice in CMOS technologies. In STI, trenches are etched in silicon substrate and filled with silicon dioxide to electrically separate active devices [31]. The previously used isolation technique, local oxidation of silicon, suffers from lateral growth, which causes the isolation region to widen beyond the etched spaces. This lowers the integration density. It also complicates device fabrication and introduces device-functionality problems, such as high parasitic capacitances [47].

As described by Lee [36], the typical STI-process flow initially involves growing a thin pad oxide and then depositing a blanket nitride film on a raw silicon wafer. The isolation trenches are etched such that the desired trench depth (i.e., depth from silicon surface) is achieved. The CMP process is used to polish off the overburden dielectric down to the underlying nitride, where the nitride serves as a polishing stop layer. After CMP, the nitride layer is then removed via etch, resulting in active areas surrounded by field trenches. A typical STI-process flow is shown in Fig. 15.

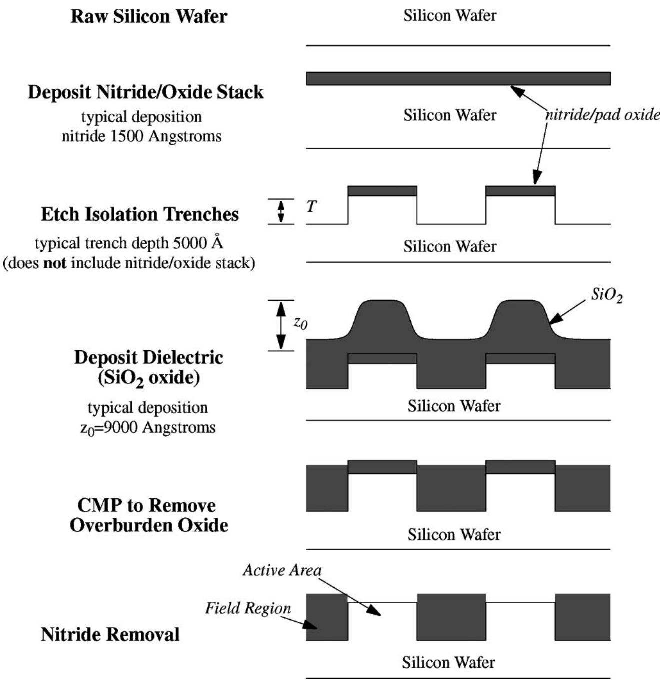
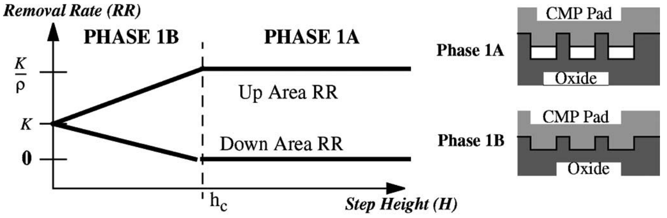


Fig. 15. Typical STI process [36].



Phase 1A indicates polish before the CMP pad contacts the down areas.
Phase 1B indicates polish after down area has been initially contacted.

Fig. 16. Removal-rate diagrams for STI CMP polish (oxide overburden phase) [36].

Lee [36] identifies two major phases in the STI CMP process. The first phase is the polish of overburden oxide. The second phase is the overpolish into the nitride layer. The second phase is due to the different pattern densities across the die, e.g., CMP pad contacts the nitride layer at different locations at different times. The first phase can be further broken down into two subphases. The first subphase happens between the start of the polish and before the CMP pad contacts the down areas (i.e., areas with lower height than their surroundings). The second subphase occurs from the time CMP pad contacts the down areas until the up-area overburden oxide has been completely cleared to nitride.

The first subphase has a homogeneous nature in that only one material is being polished at each moment. Lee [36] uses a removal-rate diagram to represent the polish of a single material. In this analysis, the assumption is that the initial starting point is a spatial location on the dielectric layer with a fixed step height. The feature densities for each point vary depending on the location on the die. Thus, any spatial location with a fixed effective pattern density can be expressed using a removal-rate diagram. Fig. 16 shows the removal-rate diagram for phase one. For a significantly large step height, the CMP pad only contacts the up areas, and the down-area removal rate is zero. This is the first subphase denoted as Phase 1A, as shown in the figure. The up areas polish at a patterned removal rate K/ρ, as shown in

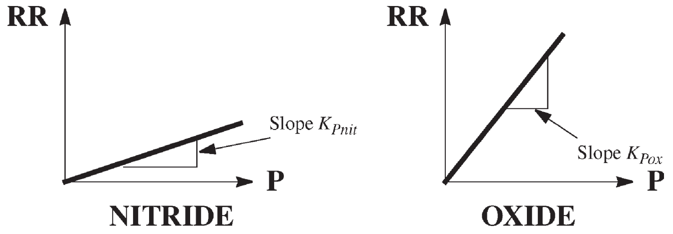


Fig. 17. Removal rate versus pressure, for oxide and nitride [36].

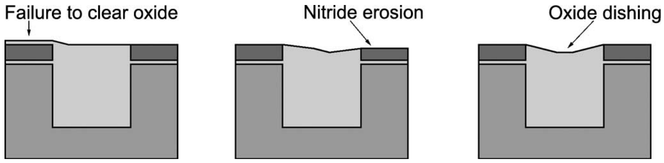


Fig. 18. Three main defects caused by STI CMP process [31].

the removal-rate diagram. As CMP process progresses, the step height reduces, and eventually, the polishing pad contacts the down areas. This is when the second subphase starts, which is denoted as Phase 1B in the figure. The up and down removal rates linearly approach each other until the step height is zero; after which, the entire oxide film is polished at the blanket oxide-removal rate K [36].

Due to the heterogeneous nature of the second STI CMP phase, a different removal diagram is used to express the polish of the two separate materials of silicon dioxide and silicon nitride. Fig. 17 shows the two removal-rate-versus-pressure curves for nitride and oxide. Assuming a Prestonian relationship, these are linear curves [36].

Dishing and erosion equations can be derived from the amount removal equations. These equations are more useful since it is the dishing and erosion phenomenon that is of most interest in STI CMP. The dishing and erosion equations are also more useful because they isolate key model parameters, making simpler equations from which to extract out model parameters. Dishing is simply the step height as a function of time, and the erosion can be computed as the amount of nitride removed. Therefore, dishing and erosion can be fully specified and predicted if the Phase 1 and Phase 2 STI CMP model parameters are known. These model parameters are characteristic of a given CMP process (tool, consumable set, etc.), and the model equations can be used to predict dishing and erosion on wafers patterned with arbitrary layouts that are subjected to a specific characterized CMP process [36]. In the next section, density-analysis methods are introduced. In order to assess the post-CMP effect, the pattern-density parameter must be computed.

B. STI CMP Fill Synthesis

There are three types of failures due to STI CMP. First, the CMP process may fail to completely remove the excess oxide. If the overburden oxide is not completely removed, it will prevent the stripping of the underlying nitride, resulting in a circuit failure. Second, even if CMP does remove the excess oxide completely, it may cause erosion of the underlying nitride, which exposes the underlying active devices and causes device failure. Finally, CMP may remove an excessive amount of oxide within the trenches, causing oxide dishing [4]; this results in poor isolation. These failures, due to the CMP process, are shown in Fig. 18. Traditionally, CMP imperfections have been mitigated by either REB or fill insertion. However, the etchback process incurs extra processing cost (i.e., mask cost and throughput loss) and, hence, is not economically attractive. Fill insertion for STI is the alternative and involves the addition of dummy nitride features to increase the nitride (and, hence, oxide) density.

Hence, the primary goal of fill insertion is to maximally reduce causes for the three key-manufacturing failures due to imperfect CMP. The second goal would be to control STI-induced stress, a significant component of which is unmodeled due to the size of STI wells. STI stress is due to the following: 1) size of diffusion regions and 2) size of the STI well isolating the diffusion region. Stress due to diffusion size is already included in today's SPICE models. However, stress due to the STI-well size is not modeled and can be a significant source of variation [31]. Typical power/performance characterization considers wells of smallest or largest size for the best and worst case analysis. However, when nitride density is higher, then devices get smaller STI wells around them which reduces the difference between these estimates. Hence, it improves the power/performance predictability of the devices [31].

Postplanarization topography in STI CMP depends on the overburden-oxide density, which is affected by the underlying nitride density. Due to the high-density-plasma process, which is widely used for oxide deposition, the oxide is deposited with slanted sidewalls. Hence, features on the oxide layer are a shrunk version of the nitride features [3], [48], [69]. For example, a square feature on the nitride layer with sides of $5x$ could correspond to a square with sides of $3x$ when deposited on the oxide layer—i.e., each edge is brought in by distance x. In this scenario, features with sides less than $2x$ will not appear on the oxide layer. Therefore, cleverly chosen fill shapes and sizes can be used to simultaneously control the pattern densities on both the nitride and oxide layers.

Failure to completely remove the overburden oxide is the main cause of failure in the oxide-CMP process. This phenomenon happens over regions where oxide density is higher than average. In higher density regions, the CMP-pad pressure is reduced, and hence, the removal rate is less than that of the regions with lower density [47]. Oxide dishing and nitride erosion can be significantly reduced by increasing the nitride density. In fact, since nitride is used as a polish stop, higher nitride density makes the detection of the nitride more accurate. Based on the aforementioned analyses, Kahng *et al.* [31] propose a new fill-insertion methodology for STI CMP processes. They give the following prioritization of fill-insertion objectives: 1) minimize oxide-density variation and 2) maximize nitride density. A bicriteria formulation is introduced in [31].

Given:

1) a set of rectilinear nitride regions contributed by the devices in the design;

2) parameter α by which nitride features shrink on each side to give oxide features;
3) design rules: minimum nitride width, maximum nitride width, minimum nitride space and notch, minimum nitride area, minimum enclosed area by nitride.

Find:

1) locations for fill insertion.

Such that:

1) Oxide density variation is minimized.
2) Nitride density is maximized.

For the first objective, Kahng *et al.* [31] use the same LP formulation proposed in [29] (see Section IV-B). The fill slack in the STI method is the maximum oxide density due to fill insertion, and the maximum oxide-density contribution is made by maximum fill insertion on the nitride layer. Using the terminology of [31], the maximum fill region, the union of all regions where fill can be inserted subject to design-rule constraints, is denoted by $\mathrm{Nitride}_{\max}$ and its density is denoted as $|\mathrm{Nitride}_{\max}|$. Maximum oxide density can be achieved by shrinking $\mathrm{Nitride}_{\max}$ by x on all sides for any polygon. In addition, $|\mathrm{Oxide}_{\max}|$ is used to denote the oxide density due to $\mathrm{Nitride}_{\max}$, and it is the highest oxide density achievable by fill insertion. To solve the second objective, Kahng *et al.* [31] first consider the following cases of $|\mathrm{Oxide}_{\mathrm{target}}|$.

1) $|\mathrm{Oxide}_{\mathrm{target}}| = |\mathrm{Oxide}_{\max}|$.
2) $|\mathrm{Oxide}_{\mathrm{target}}| = 0$.
3) $0 < |\mathrm{Oxide}_{\mathrm{target}}| < |\mathrm{Oxide}_{\max}|$.

Case $|\mathrm{Oxide}_{\mathrm{target}}| = |\mathrm{Oxide}_{\max}|$. This is the trivial case. Fill is inserted at $\mathrm{Nitride}_{\max}$ to attain oxide density of $|\mathrm{Oxide}_{\max}|$ and nitride density of $|\mathrm{Nitride}_{\max}|$. The maximum-nitride-size design rule is typically over 100 μm, which is significantly larger than typical lengths of polygons in $\mathrm{Nitride}_{\max}$. Therefore, the maximum nitride-size design rule for computing $\mathrm{Nitride}_{\max}$ can be ignored; any design-rule-check (DRC) violations are fixed postfill.

Case $|\mathrm{Oxide}_{\mathrm{target}}| = 0$. Due to the nature of the problem, there is no need to increase the oxide density of most tiles, and this case is very frequent. For this case, nitride fill features that do not contribute to the oxide density must be inserted. Fill rectangles that have one side smaller than 2α do not contribute to the oxide density due to shrinkage by α on each side. Unfortunately, rectangular fill features are suboptimal in offering the highest nitride density. To have zero oxide density, all points on inserted fill shapes must be within a distance α from the nearest edge of the shape.

General Case $0 < |\mathrm{Oxide}_{\mathrm{target}}| < |\mathrm{Oxide}_{\max}|$. Due to the nature of the LP solution [27], tiles which require density increase get an $|\mathrm{Oxide}_{\mathrm{target}}| = |\mathrm{Oxide}_{\max}|$, and this case is very infrequent. Hence, first, fill insertion is performed in $\mathrm{Nitride}_{\max}$, and then, holes of minimum size are created since they offer high nitride density with zero or small oxide density.

Fig. 19 shows a small section of OpenRisc8. Fig. 19(a) is the unfilled layout with nitride in the shaded rectilinear regions and trenches everywhere else. The same section after tiling-based fill insertion (fill size $= 0.5\ \mu$, fill spacing $= 0.5\ \mu$) performed with Mentor Calibre v9.3_5.9 is shown in Fig. 19(b). Fill regions are illustrated in gray. In Fig. 19(c), the same section with fill insertion performed with the proposed methodology is shown. As is evident, nitride density is substantially higher with the proposed fill approach. Holes created in fill regions to control the oxide density are also visible.

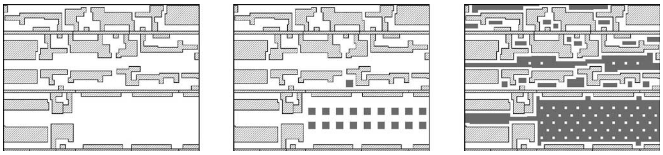


Fig. 19. Layout with fill inserted using tiling-based method and with the proposed method. Unfilled layout, layout with tile-based fill inserted, and layout with fill inserted with the proposed method are shown. Fill is shown in gray and the shaded regions represent nitride due to CMOS devices (i.e., diffusion regions) [31].

Experimental results show that using the proposed method, oxide-density variation is reduced by 63% and minimum nitride density is increased by 79% as compared to typical tiling-based fill-insertion practices. In addition, post-CMP topography becomes more uniform: maximum final step height decreases by 9% with a 17% increase in the planarization window (i.e., process window for duration of CMP application) [31].

Tian *et al.* [60] propose a model-based STI fill-insertion methodology by deriving a time-dependent relation between post-CMP topography and layout pattern density for CMP in STI. Based on their derivations, they used a non-LP formulation for the CMP-fill insertion. The assumption in the derivation of the CMP model for STI is that the profile of oxide deposited is conformal to the underlying trench profile and side walls of both the oxide deposited and underlying trenches are straight. In addition, CMP-fill features in the method by Tian *et al.* [60] are limited to square shapes, whereas in [31], they can take up any rectilinear shapes.

VI. Design-Driven Fill Synthesis

As the industry moves into the 65-nm node and beyond, traditional fill-synthesis methods reach their limits of usefulness. One indication of this is the emergence of "recommended rules," e.g., "it is better to have a small difference between the density values of adjacent windows" (a smoothness objective), or "it is better to maximize the overlap of fill shapes on adjacent layers to enable dummy via insertion" (since dummy vias enhance low-k dielectric-material stability). The impact of fill synthesis on parasitics and timing also continues to be a key concern for the designer. It is increasingly difficult for a DRC platform to obtain an optimal design-driven fill-synthesis solution that meets all basic CMP design rules and as many recommended rules as possible while minimizing impact on timing. With this in mind, we now sketch the capabilities of more sophisticated "design-driven" fill syntheses. Such techniques can potentially reduce engineering effort while enhancing manufacturability through increased process and design latitudes. Design-driven fill embodies such features as the following.

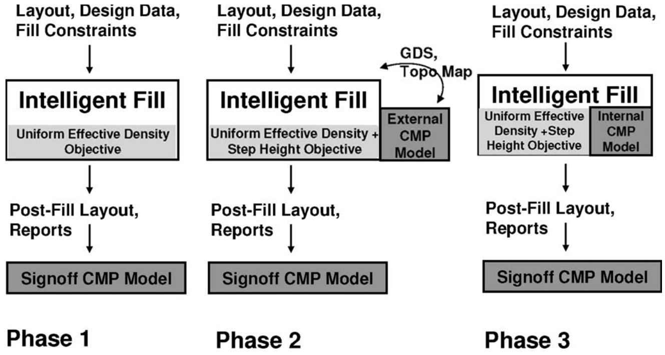

Layout, Design Data, Fill Constraints
Intelligent Fill
Uniform Effective Density Objective
Post-Fill Layout, Reports
Signoff CMP Model
Phase 1
Layout, Design Data, Fill Constraints
GDS, Topo Map
Intelligent Fill
Uniform Effective Density + Step Height Objective
External CMP Model
Post-Fill Layout, Reports
Signoff CMP Model
Phase 2
Layout, Design Data, Fill Constraints
Intelligent Fill
Uniform Effective Density +Step Height Objective
Internal CMP Model
Post-Fill Layout, Reports
Signoff CMP Model
Phase 3

Fig. 20. Evolution of intelligent fill with CMP modeling.

1) Global optimization: Typical foundry rules specify upper and lower bounds on density in windows of the layout region. These windows are "stepped" to improve density uniformity, e.g., a Calibre deck might verify density in 200 μm × 200 μm windows, stepped in increments of 50 μm (recall the illustration of Fig. 9). In this scenario, windows have $w = 200$ μm, each window is divided into $r = 4$ "steps," and the step distance is $w/r = 50$ μm. A large application-specified integrated circuit 20 mm on a side would have 160 000 "tiles" on each layer, and a ten-layer metal process implies 1.6 million tiles in the chip. Future design-driven "intelligent" fill must compute optimal amounts of fill to be added into each tile, simultaneously, for all tiles on all layers of the chip. This is enabled by, e.g., highly scalable nonlinear-optimization technology.
2) Direct optimization of total variation and smoothness: For improved manufacturability, it is necessary to explicitly minimize the difference between minimum and maximum postfill window densities, i.e., we should optimize the min-var objective. It is also necessary to create "smooth" fill, e.g., through explicit upper bounds on the density difference between any pair of adjacent windows.
3) Model-based fill synthesis: Rule-based fill synthesis is based on concepts such as density or keepoff distance rules, which may be applied to wiring segments belonging to nets with less than the given threshold amounts of timing slack. Model-based fill synthesis, on the other hand, would use CMP models to, e.g., identify regions where maintaining planarity is important (e.g., next to timing-critical segments and below-critical segments). The model-based approach has implicit tight coupling to extraction and timing engines, particularly to model the impact of fill on coupling capacitance. Fig. 20 shows the evolution of model-based fill synthesis from today's Phase 1 practice through a mature Phase 3 portrait. In Phase 3, the CMP model is very tightly coupled into the design-driven fill synthesis.
4) Timing-driven fill synthesis: One of the largest concerns in fill synthesis, apart from meeting the CMP design rules, is the impact of fill insertion on capacitances of signal nets. Excessive increase in wire capacitance can cause a net to violate setup timing constraint (or, on the other hand, added capacitance can increase hold timing slack). A large value for keepoff distance reduces the setup slack danger but reduces fill slack and, therefore, can make it impossible to meet minimum density constraints. With timing-driven intelligent fill, the impact of inserting fills on timing is continually assessed, and the minimum keepoff distance for each net to meet the setup-time constraint can be computed to avoid a wastefully large "one size fits all" keepoff distance. In a more advanced intelligent timing-driven fill flow, the impact of fill insertion on both wafer topography and timing would be analyzed and optimized concurrently. One additional advantage of timing-driven fill is that it can improve the hold-time slack of a net by deliberately and selectively introducing capacitance to that net (Fig. 21).

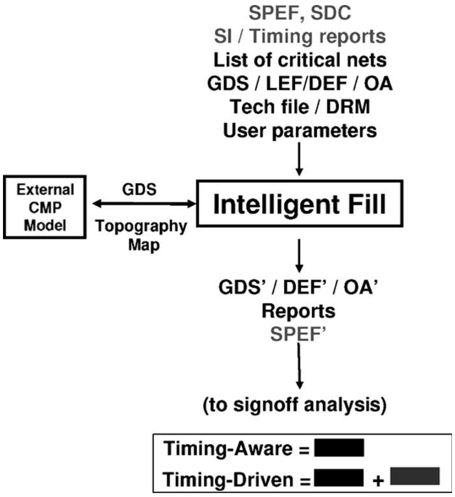
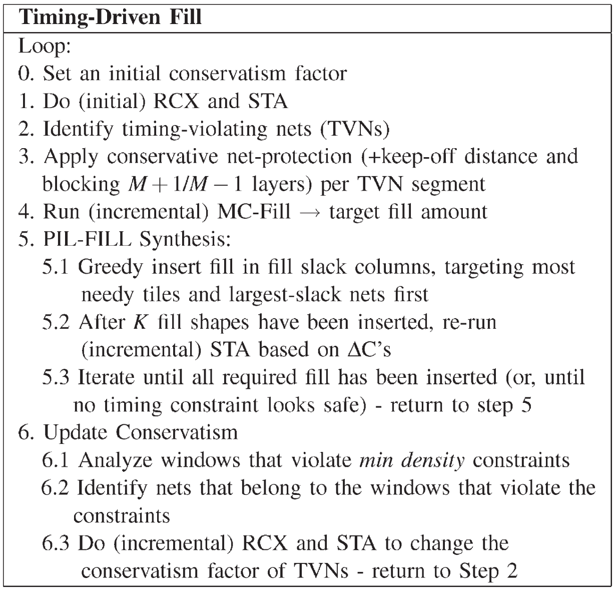


Fig. 21. Timing-aware and timing-driven use models.

5) Flexible fill grid: Global optimization of fill synthesis will return better solution quality if the fill "slack" per tile is maximized. Accurate computation of fill slack in any tile entails solving a maximum-fill problem. This requires flexible adaptation of the underlying "grid" of arrayed fill shapes so as to maximally exploit available space (and without compromising peak memory or output data volume).
6) Symmetric analog fill: To satisfy requirements of analog and other mismatch-sensitive circuits, intelligent fill must respect user-defined axes of symmetry within specified regions per layer.
7) Maximization of via fill: To improve dielectric mechanical stability and other objectives, intelligent fill maximizes the vertical alignment of inserted fill shapes to enable maximum insertion of via fill. The via-fill insertion should also be performance-driven.

Fig. 22 shows a practical approach to timing-driven fill. After all the required fill has been inserted, windows that still violate the minimum density criteria are identified, and all the nets belonging to these windows are selected. To meet the density criteria, the conservatism factor of timing-violating nets (TVNs) must be updated so as to allow additional fill to be inserted. This is done in accordance with the results of



```
Timing-Driven Fill
Loop:
0. Set an initial conservatism factor
1. Do (initial) RCX and STA
2. Identify timing-violating nets (TVNs)
3. Apply conservative net-protection (+keep-off distance and
   blocking $M+1/M-1$ layers) per TVN segment
4. Run (incremental) MC-Fill → target fill amount
5. PIL-FILL Synthesis:
   5.1 Greedy insert fill in fill slack columns, targeting most
       needy tiles and largest-slack nets first
   5.2 After $K$ fill shapes have been inserted, re-run
       (incremental) STA based on $\Delta$C's
   5.3 Iterate until all required fill has been inserted (or, until
       no timing constraint looks safe) - return to step 5
6. Update Conservatism
   6.1 Analyze windows that violate min density constraints
   6.2 Identify nets that belong to the windows that violate the
       constraints
   6.3 Do (incremental) RCX and STA to change the
       conservatism factor of TVNs - return to Step 2
```

Fig. 22. Timing-driven fill-synthesis approach [15].

an incremental RC extractor and static-timing analyzer (i.e., basically to update the timing slacks of TVNs).

Intelligent design-driven fill supports both timing-aware and timing-driven use models, in both the place-and-route and postplace-and-route contexts (Fig. 21). With the timing-aware use model, (black color) for the postplace-and-route context, intelligent fill honors net-class information (e.g., according to timing and noise margins) and user-specified fill keepoff rules for each net class. With the timing-driven use model (black + gray colors), intelligent fill reads standard delay constraints and standard parasitic exchange format (SPEF), as well as golden-signoff-analysis reports, and delivers a new SPEF output in addition to the timing- and severity-index (SI)-correct fill. The supporting infrastructure should provide the incremental extraction and timing/SI analyses that are needed to support timing-driven fill synthesis. Interfacing to external topography simulation (CMP model) should also be supported.

VII. Summary and Future CMP-Fill-Synthesis Flows

We have reviewed the role of CMP fill in reducing manufacturing variation (specifically, postpolishing wafer topography).[6] A number of techniques for density analysis and global optimization of fill patterns have been noted. We have also pointed out the recent trend toward design-driven "intelligent" fill synthesis that integrates scalable global optimization, handling of a rich array of constraint types, CMP process modeling and simulation, and incremental extraction and timing analysis. Such techniques will eventually offer the capability to produce globally optimized design-driven CMP fill that satisfies difficult fill-pattern and density constraints.

To conclude this paper, it is worth considering which "platform" will deliver future CMP fill syntheses that enhance manufacturability, as well as parametric yield, of the IC product. The first possibility is that fill synthesis and timing are integrated together within the detailed router. This is intuitively reasonable for such reasons as follows.

1) Routers lay down geometries and close timing and, so, are the natural candidates to perform fill synthesis.
2) Timing closure will be more certain for the design team before handoff to manufacturing.
3) Grounded fill, which reduces timing uncertainty and improves IR drop, is a natural extension of power/ground-routing capability.

Cho *et al.* [12] also propose the first wire-density-driven global routing that considers CMP variation and timing. On the other hand, it is also reasonable to suggest that the router should not perform detailed fill insertion due to the following reasons.

1) Complicated (wraparound, full-chip, width-distribution-dependent, etc.) density analyses that support high-quality CMP modeling are not easily performed by a router.
2) Routers optimized for batch solution of full-chip detailed routing cannot deliver high-quality fill without a significant runtime hit.
3) With the possible exception of hold-time slack and coupling-induced delay uncertainty issues, grounded fill is a bad idea from a performance standpoint, and the more preferable floating-fill approach is less natural for a router.
4) Because reticle enhancements (RETs), including CMP-fill-expose process requirements, foundries may wish to control more of the postlayout RET; at the same time, extraction, area density, and fill-pattern design rules provide the foundry with great leverage to avoid any need to solve fill in the router.
5) Better passing of design intent from design to manufacturing can reduce the need to solve the problem in the router, as noted in [7].

Future CMP-fill methodologies will likely involve four potential elements: 1) CMP simulation, 2) topography-aware RC extraction, 3) timing and signal integrity awareness, and 4) multilayer fill synthesis. An important observation is that a "smart" fill synthesis, if it accurately optimizes planarity and accounts for multilayer topographic effects, minimizes the need for 2). Another observation is that today's efforts toward 1) and 2) will not of themselves provide any design solutions; they are only analyses. On the other hand, as illustrated in Phase 2 of Fig. 20, a combination of 1), 3), and 4), with planarity assumptions validated by best possible CMP simulation, may provide a very strong replacement for today's physical-verification-based floating fill or router-based grounded fill.

References

[1] A. Balasinski and J. Cetin, "Intelligent fill pattern and extraction methodology for SoC," in *Proc. Int. Workshop System-on-Chip for Real-Time Appl.*, 2006, pp. 156–159.

[2] S. Batterywala, R. Ananthakrishna, Y. Luo, and A. Gyure, "A statistical method for fast and accurate capacitance extraction in the presence of floating dummy fills," in *Proc. IEEE Int. Conf. VLSI Des.*, 2006, pp. 129–134.

[6] In a recent paper [1], silicon data are presented, which shows the correlation between density uniformity and post-CMP-topography uniformity.

[3] P. Beckage, T. Brown, R. Tian, E. Travis, A. Phillips, and C. Thomas, "Prediction and characterization of STI CMP within-die thickness variation on 90 nm technology," in *Proc. Chemical Mech. Polish for VLSI/ULSI Multilevel Interconnection Conf.*, 2004, pp. 267–274.

[4] D. Boning and B. Lee, "Nanotopography issues in shallow trench isolation CMP," *MRS Bull.*, vol. 27, no. 10, pp. 761–765, 2002. Materials Gateway.

[5] L. E. Camilletti, "Implementation of CMP-based design rules and patterning practices," in *Proc. IEEE/SEMI Adv. Semicond. Manuf. Conf.*, 1995, pp. 2–4.

[6] Y. W. Chang, H. W. Chang, T. C. Lu, Y. King, W. Ting, J. Ku, and C. Y. Lu, "A novel CBCM method free from charge injection induced errors: Investigation into the impact of floating dummy fills on interconnect capacitance," in *Proc. Int. Conf. Microelectron. Test Struct.*, 2005, pp. 235–238.

[7] Y. Chen, P. Gupta, and A. B. Kahng, "Performance-impact limited area fill synthesis," in *Proc. ACM/IEEE Des. Autom. Conf.*, 2003, pp. 22–27.

[8] Y. Chen, A. B. Kahng, G. Robins, and A. Zelikovsky, "Practical iterated fill synthesis for CMP uniformity," in *Proc. ACM/IEEE Des. Autom. Conf.*, 2000, pp. 671–674.

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

[10] Y. Chen, A. B. Kahng, G. Robins, and A. Zelikovsky, "Monte-Carlo algorithms for layout density control," in *Proc. IEEE Asia South Pacific Des. Autom. Conf.*, 2000, pp. 523–528.

[11] Y. Chen, A. B. Kahng, G. Robins, A. Zelikovsky, and Y. Zheng, "Compressible area fill synthesis," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 8, pp. 1169–1187, 2005.

[12] M. Cho, D. Z. Pan, H. Xiang, and R. Puri, "Wire density driven global routing for CMP variation and timing," in *Proc. IEEE Int. Conf. Comput.-Aided Des.*, 2006, pp. 487–492.

[13] L. M. Cook, "Chemical processes in glass polishing," *J. Non-Cryst. Solids*, vol. 520, no. 1–3, pp. 152–171, May 1990.

[14] W. B. Glendinning and J. N. Helbert, *Handbook of VLSI Microlithography: Principles, Technology, and Applications*. Park Ridge, NJ: Noyes, 1991.

[15] P. Gupta, A. B. Kahng, O. S. Nakagawa, and K. Samadi, "Closing the loop in interconnect analyses and optimization: CMP fill, lithography and timing," in *Proc. Int. VLSI/ULSI Multilevel Interconnection Conf.*, 2005, pp. 352–363.

[16] P. Gupta, A. B. Kahng, C.-H. Park, K. Samadi, and X. Xu, "Wafer topography-aware optical proximity correction for better DOF margin and CD control," in *Proc. Photomask Next-Generation Lithography Mask Technol. X Conf.*, 2005, vol. 5853, pp. 844–854.

[17] L. He, A. B. Kahng, K. H. Tam, and J. Xiong, "Variability-driven considerations in the design of integrated-circuit global interconnects," in *Proc. Int. VLSI/ULSI Multilevel Interconnection Conf.*, 2004, pp. 214–221.

[18] L. He, A. B. Kahng, K. H. Tam, and J. Xiong, "Design of IC interconnects with accurate modeling of CMP," in *Proc. SPIE—Conf. Design Process Integration for Microelectronic Manufacturing*, 2005, pp. 109–119.

[19] L. He, A. B. Kahng, K. H. Tam, and J. Xiong, "Simultaneous buffer insertion and wire sizing considering systematic CMP variation and random Leff variation," in *Proc. ACM/IEEE Int. Symp. Phys. Des.*, 2005, pp. 78–85.

[20] P. Howard, F. Kossentini, B. Martins, S. Forchhammer, W. Rucklidge, and F. Ono, "The emerging JBIG2 standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 7, pp. 838–848, Nov. 1998.

[21] *International Technology Roadmap for Semiconductors*, 2004.

[22] "Information technology—Coded representation of picture and audio information—Progressive bi-level image compression," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Recommendation T.82—ISO/IEC 11544:1993, 1993. (commonly referred to as the JBIG1 standard).

[23] *JBIG2 Final Draft International Standard*, Dec. 1999. ISO/IEC.

[24] A. B. Kahng, R. Ellis, and Y. Zheng, "Compression algorithms for dummy fill layout data," in *Proc. SPIE—Conf. Design Process Integration for Microelectronic Manufacturing*, 2003, pp. 233–245.

[25] C.-J. Hung, "Diamond metal-filled patterns achieving low parasitic coupling capacitance," U.S. Patent 6 998 716, Feb. 14, 2006.

[26] A. B. Kahng, G. Robins, A. Singh, H. Wang, and A. Zelikovsky, "Filling and slotting: Analysis and algorithms," in *Proc. ACM/IEEE Int. Symp. Phys. Des.*, 1998, pp. 95–102.

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

[28] 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.

[29] 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.

[30] A. B. Kahng, K. Samadi, and P. Sharma, "Study of floating fill impact on interconnect capacitance," in *Proc. IEEE Int. Symp. Quality Electron. Des.*, 2006, pp. 691–696.

[31] A. B. Kahng, P. Sharma, and A. Zelikovsky, "Fill for shallow trench isolation CMP," in *Proc. IEEE Int. Conf. Comput.-Aided Des.*, 2006, pp. 661–668.

[32] Y. Kim, D. Petranovic, and D. Sylvester, "Simple and accurate models for capacitance increment due to metal fill insertion," in *Proc. IEEE Asia South Pacific Des. Autom. Conf.*, 2007, pp. 456–461.

[33] A. Kurokawa, T. Kanamoto, T. Ibe, A. Kasebe, C. W. Fong, T. Kage, Y. Inoue, and H. Masuda, "Dummy filling methods for reducing interconnect capacitance and number of fills," in *Proc. IEEE Int. Symp. Quality Electron. Des.*, 2005, pp. 586–591.

[34] A. Kurokawa, T. Kanamoto, A. Kasebe, Y. Inoue, and H. Masuda, "Efficient capacitance extraction method for interconnects with dummy fills," in *Proc. Custom Integr. Circuits Conf.*, 2004, pp. 485–488.

[35] H. Landis, P. Burke, W. Cote, W. Hill, C. Hoffman, C. Kaanta, C. Koburger, W. Lange, M. Leach, and S. Luce, "Integration of chemical–mechanical polishing into CMOS integrated circuit manufacturing," *Thin Solid Films*, vol. 220, no. 1/2, pp. 1–7, Nov. 1992.

[36] B. Lee, "Modeling of chemical–mechanical polishing for shallow trench isolation," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA, 2002.

[37] B. Lee, D. S. Boning, D. L. Hetherington, and D. J. Stein, "Using smart dummy fill and selective reverse etchback for pattern density equalization," in *Proc. Chemical Mech. Polish for VLSI/ULSI Multilevel Interconnection Conf.*, 2000, pp. 255–258.

[38] W.-S. Lee, K.-H. Lee, J.-K. Park, T.-K. Kim, and Y.-K. Park, "Investigation of the capacitance deviation due to metal-fills and the effective interconnect geometry modeling," in *Proc. Int. Symp. Quality Electron. Des.*, 2003, pp. 354–357.

[39] W.-S. Lee, K.-H. Lee, J.-K. Park, T.-K. Kim, Y.-K. Park, and J.-T. Kong, "Investigation of the capacitance deviation due to metal fills and the effective interconnect geometry modeling," in *Proc. Int. Symp. Quality Electron. Des.*, 2003, pp. 373–376.

[40] K.-H. Lee, J.-K. Park, Y.-N. Yoon, D.-H. Jung, J.-P. Shin, Y.-K. Park, and J.-T. Kong, "Analyzing the effects of floating dummy fills: From feature scale analysis to full-chip *RC* extraction," in *IEDM Tech. Dig.*, 2001, pp. 31.3.1–31.3.4.

[41] K.-S. Leung, "SPIDER: Simultaneous post-layout IR-drop and metal density enhancement with redundant fill," in *Proc. Int. Conf. Comput.-Aided Des.*, 2005, pp. 33–38.

[42] Z. Li *et al.*, "Effect of slurry flow rate on tribological, thermal, and removal rate attributes of copper CMP," *J. Electrochem. Soc.*, vol. 151, no. 7, pp. G482–G487, 2004.

[43] R.-B. Lin, "Comments on "Filling algorithms and analyses for layout density control"," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 21, no. 10, pp. 1209–1211, Oct. 2002.

[44] S. Mudhivarthi, N. Gitis, S. Kuiry, M. Vinogradov, and A. Kumar, "Effects of slurry flow rate and pad conditioning temperature on dishing, erosion, and metal loss during copper CMP," *J. Electrochem. Soc.*, vol. 153, no. 5, pp. G372–G378, 2006.

[45] G. Nanz and L. E. Camilletti, "Modeling of chemical–mechanical polishing," *IEEE Trans. Semicond. Manuf.*, vol. 8, no. 4, pp. 382–389, Nov. 1995.

[46] M. M. Nelson, "Optimized pattern fill process for improved CMP uniformity and interconnect capacitance," in *Proc. Univ./Government/Ind. Microelectron. Symp.*, 2003, pp. 374–375.

[47] D. Ouma, "Modeling of chemical–mechanical polishing for dielectric planarization," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA, 1998.

[48] J. T. Pan, D. Ouma, P. Li, D. Boning, F. Redecker, J. Chung, and J. Whitby, "Planarization and integration of shallow trench isolation," in *Proc. Int. VLSI/ULSI Multilevel Interconnection Conf.*, 1998, pp. 467–472.

[49] J.-K. Park, K.-H. Lee, J.-H. Lee, Y.-K. Park, and J.-T. Kong, "An exhaustive method for characterizing the interconnect capacitance considering the floating dummy fills by employing an efficient field solving algorithm," in *Proc. Int. Conf. Simul. Semicond. Processes Devices*, 2000, pp. 98–101.

[50] K. A. Perry, "Chemical mechanical polishing: The impact of a new technology on an industry," in *Proc. Symp. VLSI Technol.*, 1998, pp. 2–5.

[51] P. F. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*. New York: Springer-Verlag, 1985.

[52] S. Raghvendra and P. Hurat, "DFM: Linking design and manufacturing," in *Proc. Int. Conf. VLSI Des.*, 2005, pp. 705–708.

[53] *Handbook of Microlithography, Micromachining, and Microfabriation*, SPIE Opt. Eng. Press, Bellingham, WA, 1997.

[54] S. Sivaram, H. Bath, R. Legegett, A. Maury, K. Monning, and R. Tolles, "Planarizing interlevel dielectrics by chemical mechanical polishing," *Solid State Technol.*, vol. 35, no. 5, pp. 87–91, May 1992.

[55] J. Sorooshian *et al.*, "Effect of process temperature on coefficient of friction during CMP," *Electrochem. Solid-State Lett.*, vol. 7, no. 10, pp. G222–G224, 2004.

[56] J. Sorooshian *et al.*, "Arrhenius characterization of ILD and copper CMP processes," *J. Electrochem. Soc.*, vol. 151, no. 2, pp. G85–G88, 2004.

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

[58] B. Stine, D. Ouma, R. Divecha, D. Boning, J. Chung, D. L. Hetherington, I. Ali, G. Shinn, J. Clark, O. S. Nakagawa, and S.-Y. Oh, "A closed-form analytic model for ILD thickness variation in CMP processes," in *Proc. Chemical Mech. Polish for VLSI/ULSI Multilevel Interconnection Conf.*, 1997, pp. 266–273.

[59] B. Stine, D. O. Ouma, R. R. Divecha, D. S. Boning, J. E. Chung, D. L. Hetherington, C. R. Harwood, O. S. Nakagawa, and S.-Y. Oh, "Rapid characterization and modeling of pattern-dependent variation in chemical–mechanical polishing," *IEEE Trans. Semicond. Manuf.*, vol. 11, no. 1, pp. 129–140, Feb. 1998.

[60] R. Tian, X. Tang, and M. D. F. Wong, "Dummy-feature placement for chemical–mechanical polishing uniformity in a shallow-trench isolation process," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 21, no. 1, pp. 63–71, Jan. 2002.

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

[62] R. Tian, D. F. Wong, R. Boone, and A. Reich, "Dummy feature placement for oxide chemical–mechanical polishing manufacturability," Dept. Comput. Sci., Univ. Texas, Austin, TX, pp. 9–19, 1999. Tech. Rep.

[63] N. N. Toan, "Spin-on glass materials and applications in advanced IC technologies," Ph.D. dissertation, Universiteit Twente, Enschede, The Netherlands, 1999.

[64] T. Tugbawa, "Chip-scale modeling of pattern dependencies in copper chemical mechanical polishing processes," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA, 2002.

[65] T. Tugbawa, T. Park, D. Boning, T. Pan, P. Li, S. Hymes, T. Brown, and L. Camilletti, "A mathematical model of pattern dependence in Cu CMP process," in *Proc. Int. Chemical-Mech. Polishing Symp.*, 1999, pp. 605–615.

[66] D. White, "Characterization and modeling of dynamic thermal behavior in CMP," *J. Electrochem. Soc.*, vol. 150, no. 4, pp. G271–G278, 2003.

[67] H. Xiang, L. Deng, R. Puri, K.-Y. Chao, and M. D. F. Wong, "Dummy fill density analysis with coupling constraints," in *Proc. ACM/IEEE Int. Symp. Phys. Des.*, 2007, pp. 3–9.

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

[69] X. Xie, T. Park, D. Boning, A. Smith, P. Allard, and N. Patel, "Characterizing STI CMP processes with an STI test mask having realistic geometric shapes," in *Proc. Chemical-Mech. Polishing Symp., MRS Spring Meeting*, 2004.

[70] W. Yu, M. Zhang, and Z. Wang, "Efficient 3-D extraction of interconnect capacitance considering floating metal fills with boundary element method," in *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, Jan. 2006, vol. 25, pp. 12–18.

[71] R. O. Topaloglu, "Energy-minimization model for fill synthesis," in *Proc. IEEE Int. Symp. Quality Electron. Des.*, 2007, pp. 444–451.

[72] A. B. Kahng and R. O. Topaloglu, "A DOE set for normalization-based extraction of fill impact of capacitance," in *Proc. IEEE Int. Symp. Quality Electron. Des.*, 2007, pp. 467–474.

Andrew B. Kahng (SM'07) received the A.B. degree in applied mathematics (physics) from Harvard College, Cambridge, MA, and the M.S. and Ph.D. degrees in computer science from the University of California at San Diego (UCSD), La Jolla.

He joined the University of California, Los Angeles (UCLA) Computer Science Department as an Assistant Professor in July 1989, and became Associate Professor in July 1994 and Full Professor in July 1998. In January 2001, he joined UCSD as a Professor in the CSE and ECE Departments. He served as Associate Chair of the UCSD CSE Department from 2003 to 2004. In October 2004, he co-founded Blaze DFM, Inc. and served as CTO of the company until resuming his duties at UCSD in September 2006. He has published over 300 journal and conference papers. Since 1997, his research in IC design for manufacturability has pioneered methods for automated phase-shift mask layout, variability-aware analyses and optimizations, CMP fill synthesis, and parametric yield-driven, cost-driven methodologies for chip implementation.

Prof. Kahng was the founding General Chair of the 1997 ACM/IEEE International Symposium on Physical Design, co-founder of the ACM Workshop on System-Level Interconnect Prediction, and defined the physical design roadmap as a member of the Design Tools and Test technology working group (TWG) for the 1997, 1998, and 1999 renewals of the International Technology Roadmap for Semiconductors. From 2000 through 2003, he was Chair of both the U.S. Design Technology Working Group, and of the Design International Technology Working Group, and continues to serve as co-chair of the Design ITWG. He has been an executive committee member of the MARCO Gigascale Systems Research Center since its inception in 1998. He has received NSF Research Initiation and Young Investigator awards, 11 Best Paper nominations, and six Best Paper awards.

Kambiz Samadi (S'01) received the B.S. degree in computer engineering from the California State University, Fresno, in 2004. He has been working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, since 2004.

His research interests include *RC* extraction, chemical–mechanical-polishing fill, and design for manufacturing.

Mr. Samadi was the recipient of the 2004 Honorable Mention Outstanding Undergraduate Award by the Computing Research Association.