# Variability-Driven Considerations in the Design of Integrated-Circuit Global Interconnects

Lei He[1], Andrew B. Kahng[2], King Ho Tam[1] and Jinjun Xiong[1]
EE Department, University of California, Los Angeles[1]
ECE Department, University of California, San Diego[2]
{lhe, ktam, jinjun}@ee.ucla.edu[1], {abk}@ucsd.edu[2]

## I. Introduction

Chemical-mechanical planarization (CMP) is an enabling technique to achieve uniformity of dielectric and conductor height in BEOL manufacturing processes. *Dummy fill* insertion improves the uniformity of metal feature density and enhances the planarization that can be obtained by CMP, but can also change the coupling and total capacitance of interconnects [1], [2]. Additionally, dishing and erosion phenomena change interconnect cross-sections [3], and hence affect interconnect capacitance and resistance.

The first contribution of this paper is a study of interconnect parasitic variations due to (i) different fill patterns that are nominally "equivalent" with respect to foundry rules, and (ii) dishing and erosion of conductors and dielectric similar to those predicted by ITRS [4]. We show that the fill *pattern*-dependent variation of coupling capacitance between adjacent wires can exceed 20X; variation of total interconnect capacitance can reach 12%. Dishing and erosion lead to interconnect resistance variations of up to 100%, but have limited impact on interconnect capacitance.

The second contribution of this paper is an evaluation of how CMP effects (fill insertion, dishing and erosion) impact the achievable bandwidth and delay of buffered global on-chip interconnects. We show that even in a regime of best-possible fill pattern solutions, *CMP-aware design* may improve bandwidth by up to 3% and reduce delay by up to 3%; improvements in today's context of suboptimal fill pattern solutions may well be higher. We also compare the effects of CMP-induced variation and random device variation on design performance, in order to assess the relative significance of CMP-related effects.

The remainder of this paper is organized as follows. Section II presents our study of interconnect capacitance variations due to choices among nominally "equivalent" fill patterns. Section III describes RC variations due to dishing and erosion. Section IV compares optimized buffered global interconnect designs based on CMP-aware and CMP-unaware RC modeling. We give conclusions and future directions for research in Section V.

## II. Modeling and Impact of Fill Patterns

### A. Modeling of Fill Patterns

We assume rectangular, isothetic fill features aligned horizontally and vertically as shown in Figure 1. In the figure, conductors $A$ and $B$ are *active* interconnects and the metal shapes between them are dummy fills. Each distinct *fill pattern* is specified by: (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; (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. We denote a fill pattern by $P(M, N, W_i, L_j, S_{x,i}, S_{y,j})$ for simplicity.

"Fixed-dissection" fill synthesis to meet foundry requirements [1], [2] typically results in a number of tiles (i.e., square regions of layout, usually several tens of microns on a side) wherein prescribed amounts of fill features are to be inserted. In each tile $T$, a total fill area $A$ must be inserted subject to at least two foundry-dependent constraints: (1) each fill feature dimension is within the bounds $[\overline{W_l}, \overline{W_u}]$, and (2) the spacing between any two neighboring fill shapes is at least $\overline{S_l}$. A *valid* fill pattern $P(M, N, W_i, L_j, S_{x,i}, S_{y,j})$ for a given tile $T$ achieves the required area of fill while respecting all design rules (e.g., minimum spacing between fill and interconnects). We address the following questions.

This paper is partially supported by NSF CAREER award CCR-0401682, the UC MICRO Program and the MARCO Gigascale Silicon Research Center.

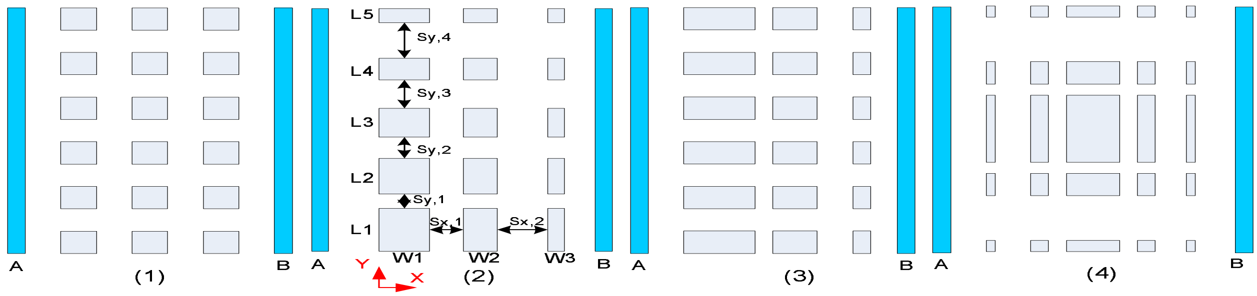



Fig. 1. Fill pattern examples.

- How much can inserted fills affect interconnect capacitance?
- How large is the range of interconnect capacitance impacts across the set of valid fill patterns?

### B. Fill Pattern Exploration

To answer the above questions, we require a methodology to explore a wide range of valid fill patterns within a given tile. Since all fills are aligned horizontally and vertically as in Figure 1, the total fill area is $A = \sum_i W_i \cdot \sum_j L_j = W_b \cdot L_b$, where $W_b$ and $L_b$ are the sums of fill widths and lengths, respectively. If the tile has width $W_t$ and length $L_t$, then the horizontal (resp. vertical) spacing budget is $S_{x,b} = \sum_j S_{x,i} = W_t - W_b$ (resp. $S_{y,b} = \sum_j S_{y,j} = L_t - L_b$). If we furthermore fix $M$ and $N$, then finding a valid fill pattern is equivalent to partitioning the budgets $W_b$, $L_b$, $S_{x,b}$, and $S_{y,b}$ among the respective series $\{W_i\}$, $\{L_j\}$, $\{S_{x,i}\}$, and $\{S_{y,j}\}$.

Enumeration of all combinations of partitions is infeasible when, e.g., capacitance extraction runtime is taken into account. Thus, we restrict our pattern exploration via the concept of a positive *distribution characteristic function* ($DCF$), denoted $f(z)$, where $z$ is an integer variable that takes the index of the element in the series. From the $DCF$ and the total budget, the $i^{th}$ element of the series is obtained as $f(i)$ plus the lower bound value as specified by design rules. For example, the value of the $i^{th}$ width $W_i = f(i) + \overline{W_l}$. If the width value $W_i$ thus obtained exceeds the upper bound $\overline{W_u}$, we take the upper bound value instead. In this way, the $DCF$ allows us to obtain a DRC-clean series under the given budget. We systematically explore different fill patterns by defining the respective $DCF$s for $\{W_i\}$, $\{L_j\}$, $\{S_{x,i}\}$, and $\{S_{y,j}\}$. Figure 2 depicts three $DCF$s for width, and their corresponding geometrical interpretation. If $f(z)$ is a constant, then all fills have uniform width. If $f(z)$ is linearly increasing, then the fills will have progressively increasing widths along the $x$-axis. And if we define $f(z)$ as a triangular function, then the center fill will have the largest width, while fills further away from the center will have progressively decreasing widths. In addition to defining different $DCF$s, we apply different $DCF$ combinations for $\{W_i\}$, $\{L_j\}$, $\{S_{x,i}\}$, and $\{S_{y,j}\}$ to explore a greater space of fill patterns; details of the methodology are given in [5].

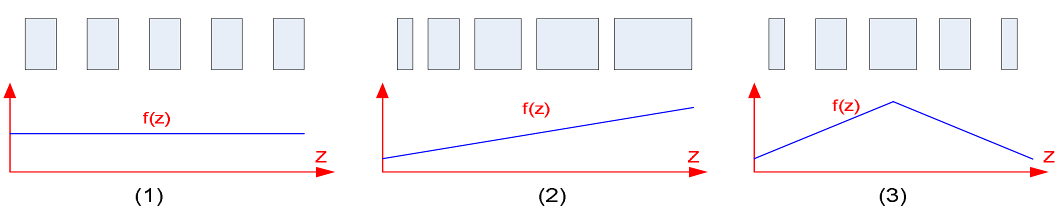
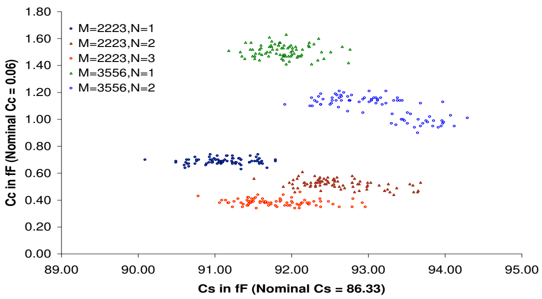


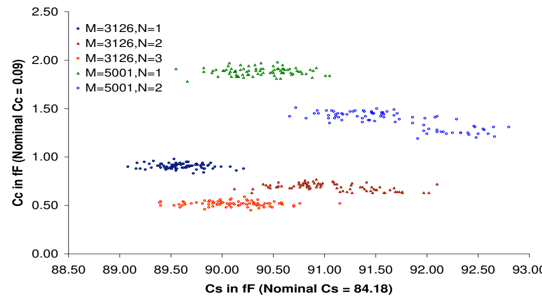

Fig. 2. Geometrical interpretation of $DCF$.

### C. Impact of Fill Patterns

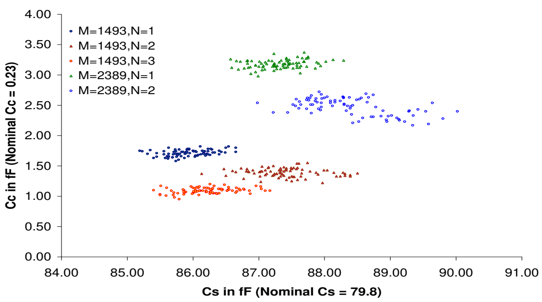
In the following, we examine the impacts of fills and fill patterns on interconnect capacitance. We consider the coupling capacitance ($C_c$) between active interconnects, and the total capacitance ($C_s$) of an individual interconnect. We use QuickCap [6], a commercial signoff-quality tool, to extract $C_c$ and $C_s$. The on-chip interconnect is modeled as a stripline where the interconnect layer is sandwiched between two ground planes (below, we call this structure a $GMG$ *stackup*; results for other stackup models may be found in [5]. We study global and intermediate (semi-global) interconnects in each of two technology nodes ($90nm$ and $65nm$), with conductor dimensions and spacing derived from the ITRS [4]. For each layout, the interconnect width is set to minimum width while the spacing between two active interconnects is set to 10X the minimum spacing. Interconnect length is $2000\mu m$ for all layouts. We assume a 50% metal density requirement for all layouts in our study.



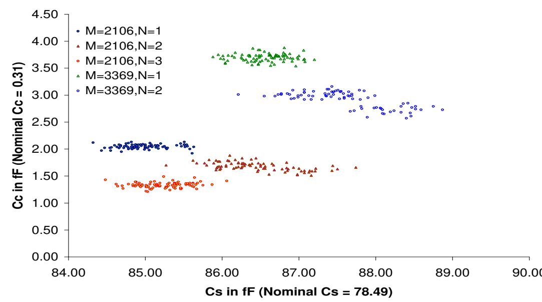

(a) $90nm$, intermediate interconnect




(b) $65nm$, intermediate interconnect




(c) $90nm$, global interconnect




(d) $65nm$, global interconnect

Fig. 3. Distribution of coupling capacitance ($C_c$) vs. total capacitance ($C_s$) over valid fill patterns for $GMG$ stackup.

For a given layout structure, we first extract the nominal $C_c$ and $C_s$ under the nominal geometries, without considering effects of either dishing and erosion or fill insertion. We then extract $C_c$ and $C_s$ under the same nominal geometry values but with fill insertion. Different valid fill patterns are enumerated as described above and detailed in [5]. Due to space constraints, the following discussion treats only our results for floating (ungrounded) fill insertion. Results for grounded fills can be found in [5].

Figure 3 plots coupling capacitance $C_c$ versus the total capacitance $C_s$ under the $GMG$ stackup model. In each plot, the y-axis is the $C_c$ value, and the x-axis is the $C_s$ value, all in $fF$. Nominal $C_c$ and $C_s$ values for each layout are indicated within the axis labels. Each point within a plot represents one valid fill pattern. From the figure, we observe that under the same constraints (e.g., metal density and filling rules), there are many potential valid fill patterns for a given layout. Inserted floating fills in the $GMG$ stackup can dramatically increase $C_c$ and $C_s$ over their respective nominal values, e.g., for $90nm$ intermediate interconnect the inserted fills can increase $C_c$ by more than 20X and $C_s$ by more than 9%; for $90nm$ global interconnect, $C_c$ can increase by more than 14X while $C_s$ increases by more than 12%. A similar trend is observed for $65nm$ technology. Therefore, to obtain robust designs that will meet requirements (delay, signal integrity, power, and parametric yield) after insertion of dummy fill, designers must be aware of the variation (increase) of both $C_c$ and $C_s$ throughout the implementation flow.

A second observation from Figure 3 is that the value of either $C_c$ or $C_s$ varies significantly across different fill patterns. For example, for the $90nm$ intermediate interconnect, the difference between the maximum and minimum $C_c$ values across all valid fill patterns can be 20X the nominal $C_c$. Variation of $C_s$ is less dramatic, but still we see a spread of more than 5% in relation to the nominal $C_s$.

To further examine the impact of different fill patterns on interconnect capacitance, we group all enumerated valid fill patterns according to their row number $M$ and column number $N$. The plots in Figure 3 show that the different fill pattern groups have very different impacts on $C_c$ and $C_s$, as they naturally fall into five different clusters in the plots. This observation strongly suggests us that there is an available lever in design, namely, the selection of fill pattern, which designers can exploit to obtain a better design. For example, if we seek a post-fill layout solution with minimum $C_c$, then from Figure 3, we can pose the following simple guidelines for selection of a "good" valid fill pattern.

- If the number of fill rows is fixed, then we should use as many fill columns as possible.
- If the number of fill columns is fixed, then we should use as few fill rows as possible.

As an example, data in Figure 3(a) indicate that for the lower three groups that have the same number of fill rows ($M$=2223), we can reduce the average $C_c$ from $1.71fF$ to $1.09fF$ by increasing $N$ from one to three. If a different design goal is sought, we may pose different design guidelines accordingly. For example, to minimize total capacitance $C_s$ in $65nm$ technology, reducing the number of fill columns seems to be a useful guideline according to the two plots on the right side of Fig. 3. (However, we note that there are still large variations for those fill patterns within a given group.)

In summary, fill insertion has a very substantial impact on $C_c$ and $C_s$; different fill patterns can result in widely varying $C_c$ and $C_s$ while meeting the same metal density requirements. Our ongoing research seeks to define useful design guidelines for selecting "good" fill patterns in various design contexts.

## III. Modeling the Impact of Dishing and Erosion on RC Parasitics

In Section II-C, we ignored the effects of dishing and erosion. We now study how dishing and erosion affect interconnect RC parasitics. Table I shows the RC parasitics for global interconnects at the $90nm$ and $65nm$ technology nodes under the $GMG$ stackup model. $R_0$ is the resistance computed from the nominal geometry values obtained from ITRS specifications of the respective technology nodes, i.e., dishing and erosion effects are not taken into account. $R_f$ is the resistance after "best" fill insertion under the 50% metal density constraint. Last, dishing and erosion effects are considered in computing $R_f$. We use the dishing and erosion model from [7] to calculate post-CMP interconnect geometries.

From Table I, we can see that resistance variation due to dishing and erosion is significant, and that resistance is always increasing. For $90nm$ technology, the resistance increases by around 50%, while for $65nm$ technology, the increase can be more than 100%. For any given technology node, as width becomes wider, the resistance variation becomes increasingly severe. For example, in $65nm$ technology, when conductor width increases from $0.24\mu m$ to $4.75\mu m$, the resistance variation goes from 74.58% to 105.63%. Because we assume the same metal density for all interconnects, the resistance is only a function of width; this is due to inherent limitations of the dishing and erosion models [7] we employ.

All capacitance values in Table I are extracted using QuickCap [6]. $C_{c,0}$ and $C_{s,0}$ are the nominal coupling capacitance and total capacitance without considering fill insertion or dishing and erosion effects. $C_{c,1}$ and $C_{s,1}$ are the coupling capacitance and total capacitance for the same stackup, taking geometry variations due to dishing and erosion effects (but no fill insertion) into account. Finally, $C_{c,f}$ and $C_{s,f}$ are the coupling capacitance and total capacitance when effects due to fills, dishing and erosion are all taken into consideration. All fills are floating, and the fill pattern is generated using the method discussed in Section II-B.

From Table I, we observe that dishing and erosion alone have negligible impact on capacitance for $90nm$ technology. As the technology shifts to $65nm$, capacitance variation due to dishing and erosion becomes more apparent, but remains at a level that is ignorable for most design contexts. In light of these results, we do not consider dishing and erosion effects in Section II-C. When dummy fills are inserted, the coupling capacitance variation becomes significant (e.g., it almost doubles for $90nm$ technology), yet the total capacitance still sees relatively small variation. portion (about 10%) of total capacitance. Whether this amount of variation is of concern will likely depend upon individual design methodologies and designers.

TABLE I

RC PARASITIC COMPARISON FOR GLOBAL INTERCONNECTS UNDER $GMG$ STACKUP.

| Width $\mu m$ | Space $\mu m$ | Nominal $R_0(K\Omega)$ | Real $R_f(K\Omega)$ | Nominal $C_{c,0}$ | Nominal $C_{s,0}$ | Dishing/Erosion $C_{c,1}$ (%) | Dishing/Erosion $C_{s,1}$ (%) | Fill+Dishing/Erosion $C_{c,f}$ (%) | Fill+Dishing/Erosion $C_{s,f}$ (%) |
|---|---|---|---|---|---|---|---|---|---|
| $90nm$ technology | | | | | | | | | |
| 0.34 | 1.34 | 39.2 | 56.35 (43.73) | 6.28 | 80.08 | 6.28 (0.00) | 80.08 (0.00) | 9.40 (49.68) | 82.82 (3.42) |
| 3.69 | 1.34 | 3.6 | 5.49 (54.03) | 6.49 | 278.43 | 6.49 (0.00) | 278.43 (0.00) | 8.77 (35.08) | 281.78 (1.20) |
| 6.70 | 1.34 | 2.0 | 3.20 (63.16) | 6.16 | 453.22 | 6.16 (0.00) | 453.22 (0.00) | 9.05 (47.04) | 457.59 (0.97) |
| 0.34 | 2.01 | 39.2 | 56.35 (43.73) | 2.02 | 79.04 | 2.02 (0.00) | 79.04 (0.00) | 5.39 (166.62) | 81.10 (2.60) |
| 3.69 | 2.01 | 3.6 | 5.49 (54.03) | 1.95 | 278.67 | 1.95 (0.00) | 278.67 (0.00) | 5.84 (199.68) | 280.72 (0.74) |
| 6.70 | 2.01 | 2.0 | 3.20 (63.16) | 2.03 | 450.63 | 2.03 (0.00) | 450.63 (0.00) | 5.46 (168.64) | 458.20 (1.68) |
| $65nm$ technology | | | | | | | | | |
| 0.24 | 0.95 | 78.0 | 136.18 (74.58) | 6.99 | 79.46 | 6.80 (-2.63) | 79.20 (-0.33) | 9.30 (33.06) | 79.38 (-0.11) |
| 2.61 | 0.95 | 7.1 | 13.50 (90.32) | 7.24 | 268.56 | 6.96 (-3.78) | 268.05 (-0.19) | 9.14 (26.33) | 264.92 (-1.35) |
| 4.75 | 0.95 | 3.9 | 8.02 (105.63) | 7.01 | 433.29 | 7.22 (2.97) | 436.25 (0.68) | 8.87 (26.51) | 432.29 (-0.23) |
| 0.24 | 1.43 | 78.0 | 136.18 (74.58) | 2.32 | 78.82 | 2.38 (2.54) | 78.72 (-0.13) | 5.63 (142.71) | 80.31 (1.88) |
| 2.61 | 1.43 | 7.1 | 13.50 (90.32) | 2.41 | 265.79 | 2.31 (-4.35) | 265.01 (-0.29) | 5.84 (141.81) | 266.76 (0.36) |
| 4.75 | 1.43 | 3.9 | 8.02 (105.63) | 2.17 | 437.34 | 2.34 (8.11) | 431.37 (-1.36) | 5.39 (148.81) | 434.32 (-0.69) |

## IV. IMPACT OF CMP ON INTERCONNECT DESIGN

The impact of CMP on interconnect design is two-fold. First, dishing and erosion decrease actual metal thickness and therefore increase resistance. Second, insertion of dummy fill to achieve uniform planarization may increase or decrease the total capacitance of interconnect. In this section, we optimize buffered bus structures under both the CMP-unaware and CMP-aware parasitic models, then compare the results to demonstrate the CMP effects on interconnect design.

### A. Experiment Setup

Applying the Berkeley Predictive Technology Model [9] for devices and the ITRS Interconnect Roadmap [4] for the interconnect stack, we perform the interconnect design for optimum bandwidth (BW) and optimum delay. We assume GMG structures defined in Section II for buses, and optimize buffered parallel buses with uniform width and space per line. We also assume that the entire bus operates at the same frequency, therefore the two edge lines of the bus, each of which has only one neighbor and hence less coupling capacitance, are assumed to operate at the same speed as the middle wires. With these assumptions, BW is defined as $BW = \frac{b_{line}}{T_{line}}$, where $b_{line}$ is the number of lines in the bus and $T_{line}$ is the delay of a wire with two neighboring wires on each side. Given the wire width $w$ and $s$, we obtain coupling capacitance $C_c$ and ground capacitance $C_g$ from a lookup table which is built using QuickCap [6]. Resistance is calculated by $R_{line} = \frac{\rho_{eff} \cdot l_{line}}{t_{line} \cdot w_{line}}$, where $\rho_{eff}$ is the bulk resistivity of copper and $l_{line}$, $t_{line}$ and $w_{line}$ are the length, thickness and width of the wire respectively. $t_{line}$ is appropriately adjusted in the CMP-aware design to reflect the loss of metal thickness due to dishing and erosion using the model in [7]. Using the delay model from [8], we obtain the delay and the optimum $k$, $h$, which are the number of buffers and buffer size (as a multiple of the minimum buffer size) respectively.

The experiment is performed twice with BW maximization and delay minimization as the design objectives, subject to different constraints. Each run of the experiment produces a combination of $w$, $s$, $k$ and $h$, which we call a *solution*. Each set of experiment produces three solutions with BW or delay:

- optimized under line resistance and capacitance with no CMP effect - Solution $\mathcal{N}$;
- optimized under line resistance and capacitance adjusted with the CMP effect - Solution $\mathcal{C}$; and
- optimized under random variation (from buffers) aware delay model - Solution $\mathcal{R}$.

Solutions $\mathcal{N}$ and $\mathcal{C}$ show how CMP affects the design. Solution $\mathcal{R}$ allows comparison of the CMP effects against those of random variation from devices, which is another important source of process variations. For Solution $\mathcal{R}$, the output resistance of the minimum-sized buffer $R_{d_m}$ is assumed to exhibit Gaussian random variation with mean $\overline{R}_{d_m}$ and standard deviation $\widehat{R_{d_m}}$ equal to 15% of $\overline{R}_{d_m}$. The delay that 99% of the instances of the interconnect with random device variation can achieve, i.e.., a "99% yield point", is

$$t_{99\%} = \frac{0.7}{h}\left(2.33\sqrt{k}\widehat{R_{d_m}} + k\overline{R}_{d_m}\right)\left(\frac{C_s}{k} + hC_{d_m} + 4.4\frac{C_c}{k}\right) + R_{line}\left(0.4\frac{C_s}{k} + 1.51\frac{C_c}{k} + 0.7hC_{d_m}\right) \quad (1)$$

TABLE II

BW MAXIMIZATION: SOLUTION $\mathcal{N}$ VS SOLUTION $\mathcal{C}$

| Technology | Layer | Solution $\mathcal{N}$ $w$ ($\mu m$) | $s$ ($\mu m$) | Estimate BW (Tbit/s) | Actual BW (Tbit/s) | $k$ | $h$ (x min) | Solution $\mathcal{C}$ $w$ ($\mu m$) | $s$ ($\mu m$) | Actual $\Delta$ BW (%) | $\Delta$ $k$ (%) | $\Delta$ $h$ (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total Buffer Area Unconstrained | | | | | | | | | | | | |
| 90nm | global | 0.335 | 0.335 | 0.931 | 0.857 | 3 | 168 | 0.335 | 0.335 | 0.233 | 0 | -7.74 |
| | interm | 0.225 | 0.225 | 4.38 | 3.70 | 1 | 96 | 0.225 | 0.225 | 0.804 | 0 | -15.6 |
| 65nm | global | 0.238 | 0.238 | 0.945 | 0.840 | 4 | 124 | 0.238 | 0.238 | 0.357 | 0 | -10.5 |
| | interm | 0.16 | 0.16 | 4.46 | 3.39 | 1 | 71 | 0.16 | 0.16 | 3.24 | 100 | -21.1 |
| Total Buffer Area Constrained to 50% | | | | | | | | | | | | |
| 90nm | global | 0.335 | 0.335 | 0.893 | 0.821 | 2 | 126 | 0.335 | 0.335 | 0 | 0 | 0 |
| 65nm | global | 0.238 | 0.238 | 0.889 | 0.821 | 3 | 82 | 0.238 | 0.238 | 0 | 0 | 0 |

TABLE III

BW MAXIMIZATION: SOLUTION $\mathcal{N}$ VS SOLUTION $\mathcal{R}$ (BUFFER AREA UNCONSTRAINED)

| Technology | Layer | Solution $\mathcal{N}$ $w$ ($\mu m$) | $s$ ($\mu m$) | Mean BW (Tbit/s) | BW @99% yield (Tbit/s) | $k$ | $h$ (x min) | Solution $\mathcal{R}$ $w$ ($\mu m$) | $s$ ($\mu m$) | $\Delta$ BW @99% yield (%) | $\Delta$ $k$ (%) | $\Delta$ $h$ (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 90nm | global | 0.335 | 0.335 | 0.931 | 0.843 | 3 | 168 | 0.335 | 0.335 | 0.237 | 0 | 9.52 |
| 65nm | global | 0.238 | 0.238 | 0.945 | 0.868 | 4 | 124 | 0.238 | 0.238 | 0.115 | 0 | 8.87 |

where $C_{d_m}$ is the minimum-sized buffer input capacitance, and $C_s = C_g + 2 \cdot C_c$ is the total wire capacitance.

### B. Experiments and Results

*1) Experiment I: Maximizing BW:* BW of the bus is maximized in this experiment. Total width of the bus (total wire width + total space between wires) is bounded by $100\mu m$, while the number of wires $N$, wire width $w$, wire space $s$, number of buffers per line $k$ and buffer size $h$ are all variables under this optimization. Global and intermediate interconnects are respectively 5mm and 1mm long. The fill insertion algorithm is tuned to minimize total interconnect capacitance $C_s$.

Table II shows the experimental results of BW maximization under the total bus width constraint. When the total buffer area is unconstrained, BW optimization always favors the minimum width/space configuration, which coincides with the conclusion in [8]. Since both Solutions $\mathcal{N}$ and $\mathcal{C}$ yield minimum spacing, no dummy fill is needed, so the only effect on design comes from dishing and erosion, which increase the wire resistance. From this experiment, the estimated BW from Solution $\mathcal{N}$ is always significantly larger than the actual BW when dishing and erosion are considered, ranging from 8% for 90$nm$ global interconnect to 24% for 65$nm$ intermediate interconnect. The relative difference of estimated BW and actual BW tends to be larger in 65$nm$ technology than that in 90$nm$ technology: for example, 8% and 16% in 90$nm$ for global and intermediate interconnects versus 11% and 24% in 65$nm$ correspondingly. Solution $\mathcal{C}$ slightly improves the actual BW over Solution $\mathcal{N}$: for example, Solution $\mathcal{C}$ achieves a 3.24% improvement of actual BW by having optimum $k$ and $h$ considering CMP.

In many situations, total area of buffers are constrained to save area and power while allowing a moderate increase in delay. Imposing a total buffer area constraint of $N \cdot k \cdot h \leq A_{total}$, where $A_{total}$ is a multiple of the minimum buffer size, we obtain another set of results presented in the lower part of Table II. The total buffer area is bounded by 50% of the total buffer area in Solution $\mathcal{N}$ without area constraint. This experiment is only performed on global interconnect due to their excessive buffer insertion requirement. Compared to the cases without area constraints, the large reduction in buffer area (and probably power) results in a moderate 4.2% and 2.3% drop in the actual BW for 90$nm$ and 65$nm$ technologies respectively. Adjustment to $k$ and $h$ by CMP-aware optimization is not possible since $N \cdot k \cdot h$ is very close to $A_{total}$. Consequently Solutions $\mathcal{N}$ and $\mathcal{C}$ use the same $k$ and $h$, which means that considering CMP effects does not improve the actual BW.

Table III presents the improvement that can be gained from Solution $\mathcal{R}$ over Solution $\mathcal{N}$ in terms of the 99% yield BW. For brevity, only global interconects are compared. As Solution $\mathcal{R}$ results in less relative improvement over

TABLE IV

DELAY MINIMIZATION: SOLUTION $\mathcal{N}$ VS SOLUTION $\mathcal{C}$

| Tech-nology | Layer | Solution $\mathcal{N}$ | | | | | | Solution $\mathcal{C}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $w$ ($\mu m$) | $s$ ($\mu m$) | Estimate Delay (ps) | Actual Delay (ps) | $k$ | $h$ (x min) | $w$ ($\mu m$) | $s$ ($\mu m$) | Actual $\Delta$ Delay (%) | $\Delta$ $k$ (%) | $\Delta$ $h$ (%) |
| Total Buffer Area Unconstrained | | | | | | | | | | | | |
| 90nm | global | 2.68 | 1.34 | 61.4 | 69.5 | 1 | 510 | 3.02 | 1.01 | -1.16 | 0 | 1.96 |
| | interm | 1.35 | 0.675 | 27.7 | 31.8 | 1 | 264 | 1.35 | 0.675 | -0.943 | 0 | -17.8 |
| 65nm | global | 2.14 | 0.95 | 83.9 | 95.8 | 2 | 404 | 2.14 | 0.95 | -0.522 | 0 | -13.1 |
| | interm | 0.96 | 0.48 | 34.0 | 43.0 | 1 | 190 | 0.8 | 0.64 | -2.56 | 0 | -36.8 |
| Total Buffer Area Constrained to 50% | | | | | | | | | | | | |
| 90nm | global | 2.01 | 2.01 | 66.2 | 73.5 | 1 | 255 | 2.01 | 2.01 | 0 | 0 | 0 |
| 65nm | global | 2.14 | 0.95 | 84.1 | 100 | 1 | 404 | 2.14 | 0.95 | -0.900 | 0 | -13.1 |

TABLE V

DELAY MINIMIZATION: SOLUTION $\mathcal{N}$ VS SOLUTION $\mathcal{R}$ (BUFFER AREA UNCONSTRAINED)

| Tech-nology | Layer | Solution $\mathcal{N}$ | | | | | | Solution $\mathcal{R}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $w$ ($\mu m$) | $s$ ($\mu m$) | Mean delay (ps) | Delay @99% yield (ps) | $k$ | $h$ (x min) | $w$ ($\mu m$) | $s$ ($\mu m$) | $\Delta$ Delay @99% yield (%) | $\Delta$ $k$ (%) | $\Delta$ $h$ (%) |
| 90nm | global | 2.68 | 1.34 | 61.4 | 71.7 | 1 | 510 | 2.68 | 1.34 | -0.558 | 0 | 16.1 |
| 65nm | global | 2.14 | 0.95 | 95.4 | 95.8 | 2 | 404 | 2.14 | 0.95 | -0.418 | 0 | 11.6 |

Solution $\mathcal{N}$ than $\mathcal{C}$ does, we conclude that design considering CMP is at least as important as design considering random device variation when BW maximization is the design objective.

*2) Experiment II: Minimizing Delay:* This experiment explores the impact of CMP on design with the objective of minimizing the delay subject to bounded total bus width and constant number of wires of the bus. This design objective is often used when the bus specification, for example, bit number and the routing area, is specified. In contrast to the optimized solution in Section IV-B.1, design for delay minimization tends to cause wires to become wider and sparser, and as a result suffer more due to dishing and dummy metal fill.

Several assumptions are made to the interconnect design in this sub-section. The optimal fill pattern is generated from the fill assumptions and algorithm in Section II to minimize the total interconnect capacitance. The total bus width is bounded by $130\mu m$ for $90nm$ technology and $100\mu m$ for $65nm$ technology. We also fix the number of bits in the bus at 32 for global interconnect and 64 for intermediate interconnect.

Table IV summarizes the results from this experiment. Solution $\mathcal{C}$ gives wire space of about 5x the minimum space, making it possible to insert dummy fill, and achieves slightly less actual delay than Solution $\mathcal{N}$ does. More delay reduction is observed for intermediate interconnects (up to 1.16% for global interconnects versus up to 2.56% for intermediate interconnects). Moreover, the relative impact of CMP-aware design in this Section is greater than that of the CMP-aware design in Section IV-B.1 due to the effects of dummy fill, which may either increase or decrease the total capacitance of the wire and provide an extra dimension for optimization. Similar to Section IV-B.1, the delay reduction is small or negligible when the total buffer area is constrained.

Table V compares Solution $\mathcal{N}$ with Solution $\mathcal{R}$ under the same objective and constraints as Table IV. Owing to the much larger buffer used in delay minimized designs, random variation from buffers becomes more significant, which presents more opportunity for optimization. This is confirmed by the larger relative improvement of 99% yield point delay of Solution $\mathcal{R}$ over Solution $\mathcal{N}$ than that of the actual delay of Solution $\mathcal{C}$ over Solution $\mathcal{N}$ in Table IV.

## C. Summary of Interconnect Design under CMP

From the results of the two sets of experiments, we conclude that

- failure to consider CMP effects on interconnect design can severely over-estimate the BW by up to 24% and under-estimate the delay by up to 26%;
- considering CMP effect during design always improves the design objective values (BW maximization and delay minimization) by up to about 3% under unconstrained buffer area assumption;

- considering CMP effect on intermediate interconnect has more impact than on global interconnect in terms of BW/delay improvement;
- considering CMP effect for newer technology (i.e. 65$nm$) tends to have more impact than for older technology (i.e. 90$nm$) in terms of BW/delay improvement;
- when buffer area is constrained, considering CMP effects has little impact on BW/delay in our experiment setting (and future study will be conducted to verify this in general);
- considering CMP makes a bigger impact on the design than considering random variation in BW maximization, but the opposite is true when delay minimization is the goal; and
- optimal fill pattern is important to achieve good design as sub-optimal fill causes excessive capacitance increase.

## V. Conclusions and Discussions

We have studied the optimal design of dummy fill patterns for CMP uniformity, as well as the optimal design of on-chip interconnects in the context of CMP-induced variability. Our simulations show that dummy fill can introduce variations of more than 20X and 12%, respectively, for coupling capacitance between adjacent wires and total interconnect capacitance. Dishing and erosion at the limits specified by the ITRS roadmap can cause interconnect resistance variations of up to 100%, but has limited impact on interconnect capacitance.

Overall, our results show that failure to consider changes in RC parasitics due to CMP can severely over-estimate interconnect bandwidth and under-estimate interconnect delay. Integrating an optimum choice of fill pattern, along with worst-case dishing and erosion effects, we apply a CMP-aware RC model to the design of global interconnects; the CMP-aware design improves bandwidth by up to 3% and reduces delay by up to 3%, compared to the conventional design that does not consider CMP effects. As process technology continues to scale, the CMP-induced RC variations become more significant and CMP-aware designs achieve more improvement in bandwidth and delay.

Finally, we note that our studies of CMP-aware global interconnect design assume fill pattern solutions that minimize total capacitance. Our enumeration of multiple fill pattern solutions - nominally "equivalent" with respect to the density criteria in foundry fill rules - shows that there can be substantial fill pattern effects on both coupling and total capacitance. Hence, we expect that today's unoptimized fill pattern solutions may entail bandwidth and delay penalties that are much larger than those presented in this paper. Our future studies will develop efficient algorithms for optimal coupling- and performance-aware fill insertion. We will also explore CMP-aware routing, wire sizing, etc. to account for CMP impacts early in the design cycle.

## References


[1] Y. Chen, P. Gupta, and A. B. Kahng, "Performance-impact limited area fill synthesis," in *DAC*, Jun 2003.

[2] P. Gupta and A. B. Kahng, "Manufacturing-aware physical design," in *ICCAD*, Oct 2003.

[3] R. Chang, *Integrated CMP Metrology and Modeling with Respect to Circuit Performance*. PhD thesis, University of California, Berkeley, 2004.

[4] Semiconductor Industry Association, *International Technology Roadmap for Semiconductors*, 2003.

[5] L. He, A. B. Kahng, K. H. Tam, and J. Xiong, "Variability-driven considerations in the design of integrated-circuit global interconnects," in *University of California, Los Angeles, Technical Report, http://eda.ee.ucla.edu*, 2004.

[6] "Quickcap user manual," in *http://www.magma-da.com/*.

[7] T. Tugbawa, T. Park, D. Boning, T. Pan, P. Li, S. Hymes, T. Brown, and L. Camilletti, "A mathematical model of pattern dependencies in cu cmp processes," in *CMP Symposium, Electrochemical Society Meeting*, Oct 1999.

[8] D. Pamunuwa, L. Zheng, and H. Tenhunen, "Optimising bandwidth over deep sub-micron interconnect," in *ISCAS*, 2002.

[9] "Berkeley predictive technology model," in *http://www-device.eecs.berkeley.edu/ ptm*.