Case Study and Efficient Modeling for Variational Chemical-Mechanical Planarization

ABSTRACT

Chemical-mechanical planarization(CMP) is an enabling technique to achieve wafer planarity in backend manufacturing processes of integrated circuits. However, CMP also causes variations in metal and dielectric thicknesses due to the non-uniformity of metal feature density. In this paper, we first conduct a case study of CMP induced variations using an industrial CMP simulator and a widely used microprocessor hardcore fabricated in a 90nm technology with eight metal layers. We reveal a few interesting characteristics on thickness variations, and particularly vertical and horizontal correlations between variations while such correlations have been virtually ignored by the existing study. These characteristics may lead to better modeling and design optimization for CMP variations. As an example, we then propose a stochastic CMP model to efficiently incorporate CMP variations in the design flow, and develop two algorithms to reduce the CMP simulation runs by 7X and 3X respectively compared to generating the stochastic CMP model by detailed CMP simulations.

1. INTRODUCTION

CMP was invented by IBM in the late 80's to enable multi-metal layers in the integrated circuits (IC). It is now a commonly used technique in interconnect or inter-layer dielectric (ILD) planarization to ensure that interconnect or ILD thicknesses are uniform. Both chemical and mechanical processes are used in polishing metal and dielectric.

Basic process of ILD CMP is to deposit the silicon oxide thicker than the final thickness and polish the material back until the step heights are removed, which provides a flat surface for the next metal layer [1]. Copper CMP process is more complex and contains three steps: removal of the overburden copper, removal of the barrier material, and copper dishing and oxide erosion [2]. The above basic processes are repeated to add ILD and metal layers from bottom to top, and to obtain multi-level interconnections.

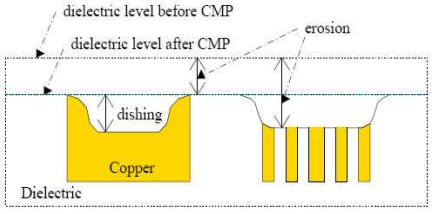

Figure 1: Dishing and Erosion in Copper CMP.

Fig. 1 from [3] illustrates dishing and erosion, two main sources of metal thickness variation. Dishing is the difference between copper heights in the trenches and around the trenches. Erosion is the difference between the dielectric thicknesses before and after CMP. CMP may also cause ILD variation. As demonstrated in Fig. 2 from [4] where (a) has no fill feature and (b) has fill feature to make metal density more uniform, non-uniformity in metal density results in different removal rate, and further causes variation of ILD thickness. However, CMP does not affect the interconnect width and spacing directly.

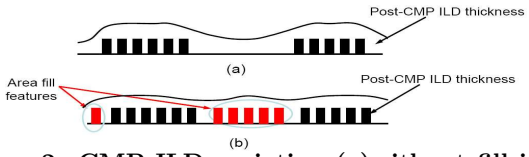

Figure 2: CMP ILD variation (a)without fill insertion (b)with fill insertion .

To model CMP procedure accurately, a number of simulators have been published. Pattern-density model from Preston's equation [5] is usually used in ILD thickness simulation. It convolves local density of wafer with density filters to get effective density, and finally calculate ILD thickness. Copper CMP process is much more complex than ILD process. The copper CMP process simulation in [6] models three steps, calculates the time it takes in each step, and obtains the amount of dishing and erosion in the end. In addition, industrial tools are also available such as DVIP (Designer Virtual Interconnect Predictor) from Cadence [7].

However, CMP simulators only produces raw data of CMP variations, but do not directly explicate any characteristics of CMP variations. The first contribution of this paper is to use a hardcore IP block fabricated in a leading technology and a mature industrial CMP simulator to reveal characteristics (called *observation* in this paper) of CMP variations. Particularly, we show that there exists strong vertical (i.e., between layers) and horizontal (i.e., within a same layer) correlations between CMP variations. These observations may lead to more accurate modeling and better design optimization for CMP variations. For example, existing work has studied CMP-induced RC extraction variations [8] and interconnect performance variations [3], buffering and wire sizing with fill insertion [9], and CMP aware global routing [10]. None of these study [8, 3, 9, 10] took into account the vertical correlation between multiple interconnect layers. However, this paper demonstrates that such correlation

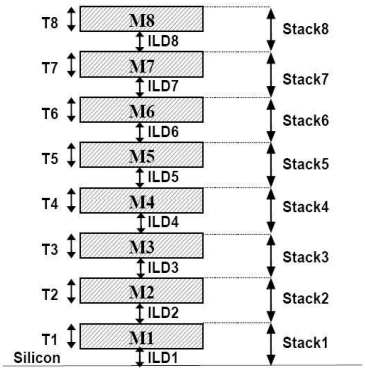

Figure 3: Diagram to define metal thickness T_1, T_2,...,T_8; ILD thickness ILD_1, ILD_2,...,ILD_8; stack $Stack_1$, $Stack_2$,...,$Stack_8$.

has a great impact on metal and dielectric thickness and should not be ignored.

The second contribution of this paper is that we propose a stochastic CMP model with spatial correlation in order to efficiently incorporate CMP variations in the modeling and design flow. We also develop two algorithms to reduce the CMP simulation runs by $7X$ and $3X$ respectively compared to generating the stochastic CMP model by detailed CMP simulations.

The rest of the paper is organized as follows. Section 2 discusses experiment setting and reveals CMP variation characteristics with vertical correlation. Section 3 introduces the stochastic CMP model with spatial correlation and runtime efficient algorithms to build this model. We conclude in Section 4.

2. CMP CHARACTERISTICS

2.1 Experimental Setting

To study CMP variations, we use a widely used microprocessor hardcore fabricated in a 90nm technology with eight metal layers. We show the detailed layer stacking in Fig. 3, where T1, T2, ..., T8 are metal thickness, and ILD1, ILD2, ..., ILD8 are ILD thickness. We also call metal layer and its underneath dielectric layer as one *stack*.

We run CMP simulation using DVIP (Designer Virtual Interconnect Predictor) from Cadence to obtain the thickness for every metal and dielectric layer, and the thickness is calculated for every location 10μm away. The thicknesses are stored in a two-dimensional matrix for each layer. Each row (or column) of thicknesses are those with same x-axis (or y-axis) locations. For simplicity of presentation, we report horizontal distance in unit of 10μm unless otherwise stated. Therefore, a distance of 1 means 10μm. For IP protection, we report all thickness and variation normalized with respect to the average thickness in each metal or dielectric layer. Note that metal fills have been inserted in the hardcore under study to satisfy the minimum metal density. We will show that the fill insertion makes metal and dielectric thickness more uniform and enables some desired CMP characteristics.

2.2 Thickness Variation

2.2.1 Edge Effects

We present the metal thickness T_1 and T_8 in Fig. 4 and Fig. 5. As shown in Fig. 4 (a), metal thickness at chip edges is distinctly larger than those in the center for M_1, but not for M_8 (Fig.4 (b)). We cut the edges within 20μm off the chip for a better observation on metal thickness variation in Fig. 5. As shown in this figure, the thickness variations exist in the center of both layers.

We have similar observation on ILD thickness (see Fig. 6 and Fig. 7). When cutting off the edges, we observe variations in both ILD_1 and ILD_8 layers. Thickness of ILD_1 at chip edges is significantly larger than those in the central area (Fig.6 (a)), but not for ILD_8 (Fig.6 (b)). While we only show figures for stacks 1 and 8, other stacks have similar trend and we conclude the following:

Observation 1 Both metal thickness and ILD thickness at edges are larger than those in the center for low stacks (stacks 1-3 in our case study), but not for other stacks.

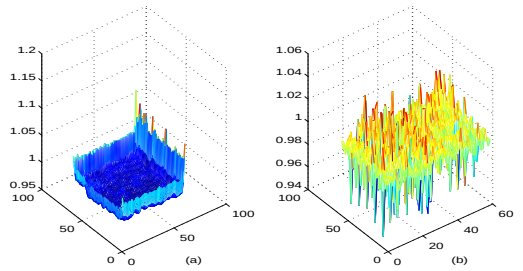

Figure 4: 3-D figure of thickness: (a) T_1 (b) T_8.

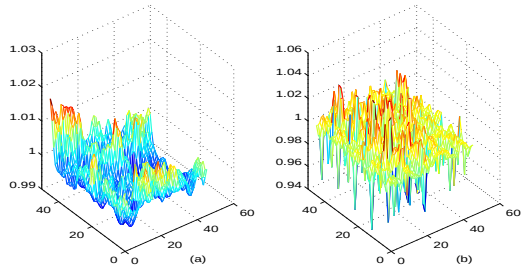

Figure 5: 3-D figure of thickness (Edges within 20μm are cut off): (a)T_1 (b) T_8.

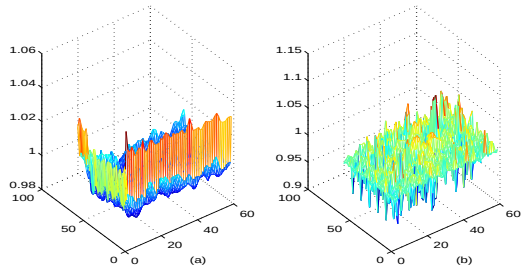

Figure 6: 3-D figure of thickness: (a) ILD_1 (b) ILD_8.

For more accurate modeling on thickness variation due to CMP for all layers, all the rest thickness calculation for

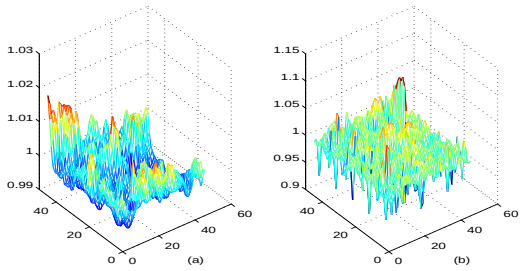

Figure 7: 3-D figure of thickness (Edges within 20μm are cut off): (a) ILD_1 (b) ILD_8.

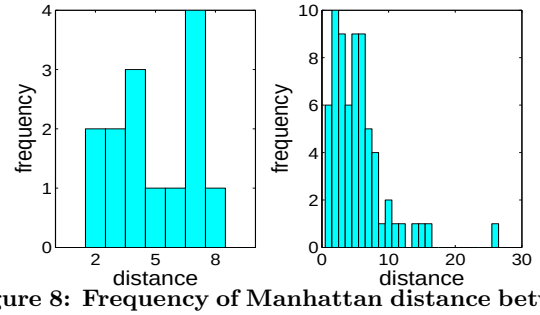

Figure 8: Frequency of Manhattan distance between valleys and peaks: (a) M_8 (b) ILD_8.

metal and ILD considers the central hardcore 20μm away from edges. More precisely, we use a 500μm×500μm selected layout window to show results in different layers after cutting off edges.

2.2.2 Roughness of Layers

Thickness variation in one layer results in the rough surface of this layer. To study roughness of layers, we first study neighbor difference and maximum difference within one layer for both metal thickness and ILD thickness. *Neighbor difference* (ND) is the thickness difference between two adjacent locations that are 10μm away. *Maximum difference* (MD) is the maximum thickness differences between any two locations within one layer.

As shown in Table 1, a higher stack has a higher variation in both metal and ILD thickness. Stacks 1-5 have similar mean and maximum neighbor difference. On the other hand, stacks 6-8 also have similar mean and maximum neighbor difference, but significantly larger than those in stacks 1-5. For maximum difference, stacks 1-4 have similar values while stacks 5-8 also have similar values but significantly larger. A relatively small neighbor difference but larger maximum difference in stack 5 suggests a locally smooth but globally rough layer. ILD thickness has a slightly larger variation compared to the corresponding metal thickness.

Second, we study the distribution of distance between neighbor valleys and peaks of thickness. *Peaks* are defined as locations with thickness larger than locations around them. *Valleys* are defined as locations with thickness smaller than locations around them. We only select those peaks and valleys, whose difference to neighbor valleys or peaks are greater than 4% of the average thickness of the current layer. Absolute frequency of Manhattan distance (in unit of 10μm) between peaks and valleys are shown in Fig.8. ILD layers have more peaks and valleys than correspondingly metal layers. In addition, ILD layers have a larger portion of peaks and valleys with small distances. For example, ILD_8 has more frequent peaks and valleys for short distances than M_8. In other words, it is more rough than M_8. Based on Table1 and Figure 8, we conclude:

Observation 2 Roughness increases for higher stacks , and ILD layers are slightly rougher than metal layers within same stacks.

2.3 Vertical Correlation

2.3.1 Preliminaries

In this section, we will discuss the correlation between metal thickness and ILD thickness across different layers. First, for two stochastic variables F_i and F_j, we define the correlation between them as follows:

$$\rho_{ij} = \frac{cov(F_i, F_j)}{\sigma_{F_i}\sigma_{F_j}} \quad (1)$$

where σ is the standard deviation and $cov(F_i, F_j)$ is the covariance. We use $MVx_{m,n}$ to represent the vertical correlation matrix between layers m and n, where x can be either metal or ILD, correspondingly representing the correlation matrix calculated by metal thickness or ILD thickness. $MVx_{m,n}(i,j)$ is the normalized covariance between the i_{th} row in layer m and j_{th} row in layer n. Therefore all thicknesses in the i_{th} row of layer m are samples of F_i in (1), and those in the j_{th} row in layer n are samples of F_j.

Second, we use correlation matrix image (CM image) in gray-scale to display vertical correlation matrix. Let 0 displayed as black and 1 displayed as white. Values between 0 and 1 are displayed as intermediate shades of gray. So the brighter a grid is, the higher correlation between rows in corresponding layers. We can calculate four correlation matrices between every two sets of thicknesses. One is a correlation matrix, one is a matrix of p-values for testing the hypothesis of correlation (each p-value is the probability of getting a correlation as large as the observed value by random chance, when the true correlation is zero), other two are lower and upper bounds for a 95% confidence interval for each coefficients [11]. In this paper, we only show correlation matrix images for the 95% upper bounds of confidence interval, which indicates the largest possible correlation between two rows at a 95% confidence.

Third, we define *complementary* and *correlated* to describe two extreme situations with high correlation. When two sequences have a high correlation, the absolute value of their correlation is close to 1. If its real value is close to -1, we call the two sequences *complementary*. If it is close to 1, we call they are *correlated*.

2.3.2 Observations

Observation 3 When routing is dense, metal and ILD layers in adjacent stacks have complementary thickness. The density threshold for complement is technology dependent.

Fig. 9 and Fig. 10 verify above observation. Fig.9 (a) and Fig.10 (a) plot thicknesses of M_1 and M_2 (as well as

	Metal Thickness			ILD Thickness		
Layer	mean ND	max ND	MD	mean ND	max ND	MD
1	0.13%	0.88%	2.80%	0.13%	0.94%	2.94%
2	0.07%	0.58%	1.11%	0.12%	0.92%	2.21%
3	0.15%	0.73%	1.91%	0.14%	0.64%	1.74%
4	0.14%	1.22%	1.88%	0.20%	1.12%	2.54%
5	0.20%	1.93%	6.51%	0.20%	2.00%	6.21%
6	0.62%	5.99%	9.50%	0.60%	5.04%	10.28%
7	0.84%	7.80%	11.94%	1.03%	8.61%	15.37%
8	0.68%	5.21%	10.30%	1.04%	10.08%	16.19%

Table 1: Neighbor difference and biggest difference comparison for metal and ILD thickness.

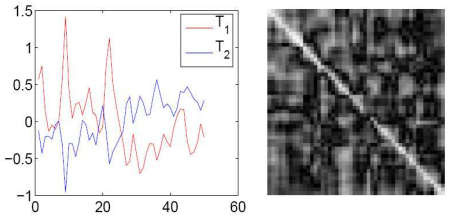

Figure 9: Thickness for M_1 and M_2 (a) Correlation of a row (b) Correlation matrix image.

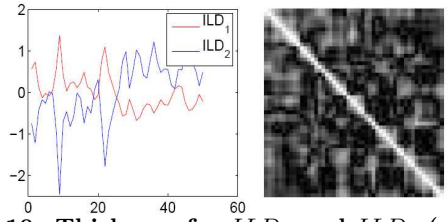

Figure 10: Thickness for ILD_1 and ILD_2 (a) Correlation of a row (b) Correlation matrix image.

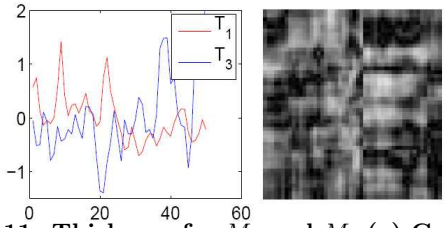

Figure 11: Thickness for M_1 and M_3 (a) Correlation of a row (b) Correlation matrix image.

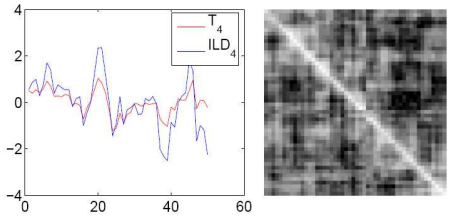

Figure 12: Thickness for M_4 and ILD_4 (a) Correlation of a row (b) Correlation matrix image.

ILD_1 and ILD_2). They are complementary in a selected row. M_2 (ILD_2) tends to be thinner where M_1 (ILD_1) is thicker for this row. As shown in Fig.9 (b) and Fig.10 (b), the diagonals of vertical correlation matrix images are bright. Thus absolute values of correlation coefficients on the diagonals are close to 1. In short, M_1 and M_2 (as well as ILD_1 and ILD_2) are complementary for the entire chip. This observation only holds for stacks 1 and 2 in our case study, where the metal density is high.

Observation 4 Layers in non-adjacent stacks are neither complementary nor correlated.

The above observation is verified by Fig.11. In Fig.11 (a), thickness of M_1 and M_3 is neither correlated nor complementary in the selected row. Correspondingly in Fig.11 (b), we cannot find a bright diagonal from vertical correlation matrix image any more. In short, T_1 in $Stack_1$ and T_3 in $Stack_3$, which belong to nonadjacent stacks, are neither complementary nor correlated over the entire chip. Observation 4 holds for all stacks in our experiments.

Observation 5 Within one stack, the metal layer and its underneath ILD layer have correlated thicknesses.

Fig.12 illustrates Observation 5. It shows metal thickness and ILD thickness in stack 4 are highly correlated, i.e. M_4 and ILD_4 tend to be thicker at the same location. This observation holds for all stacks. In contrast, a metal layer in general is not correlated with the ILD layer on the top stack.

3. EFFICIENT CMP MODELING

3.1 Stochastic Modeling with Spatial Correlation

Although CMP variation is deterministic for given multilayer layout, modeling its variation for each $10\mu m$ leads to explosive data amount and should be avoided. A viable alternative is to use a stochastic model. In this model, we divide each layer uniformly into, e.g., 10×10 regions, and then model the thickness in each region by a Gaussian variable with mean and variance, together with correlation between variances of different regions. We treat the thicknesses at different locations (e.g., $10\mu m$ away as in this paper) in each region as samples. Then mean, variance spatial correlation according to (1) can be calculated. Such correlation helps to improve accuracy in a statistic static timing analysis capable of dealing with spatial correlation.

Fig. 13 plots the spatial correlation between locations versus various distances between locations. One can easily see from the figure that the spatial correlation cannot be ignored and it is not isotropic with respect to the distance.

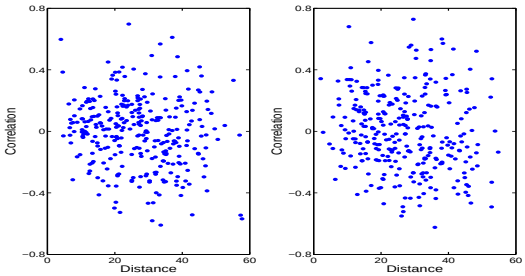

Figure 13: Correlation over distance: (a) M_5 (b) ILD_3

This is different from device variations which often has an isotropic spatial correlation with respect to the distance.

3.2 Speedup of Stochastic Modeling

Because using CMP simulation for every location $10\mu m$ away to build the above stochastic model is time consuming, we will propose two algorithms to speedup model generation and also verify the accuracy of the two algorithms.

3.2.1 Algorithm 1

The algorithm $Alg1$ is based on interpolation in the frequency domain and is shown in Table *Algorithm 1*. While the most steps are straightforward, combining 1D FFT of $\hat{D}_r$ and $\hat{D}_c$ and 2D FFT of $\hat{D}_g$ in Step 4 does not give us all elements in $\hat{W}$ for all locations $10\mu m$ away and we pad these missing elements by 0.

Algorithm 1:
1. Obtain matrix D_g by CMP simulations for a uniform grid $\tilde{K}$ coarser than K (the finest granularity)
2. Perform CMP simulations for every Kμm along 1_{st} row and 1_{st} column, and get two sequences D_r and D_c
3. Deduct mean value from D_g, D_r, D_c, and obtain $\hat{D}_g, \hat{D}_r, \hat{D}_c$.
4. Combine one-dimensional (1D) FFT of $\hat{D}_r$ and $\hat{D}_c$, and 2D FFT of $\hat{D}_g$ to obtain an estimated 2D FFT $\hat{W}$ for the unknown thickness matrix for $\hat{D}$ with the finest granularity.
5. Perform 2D IFFT on $\hat{W}$ and then add mean value of D_g, finally obtain estimated thickness $\hat{D}$
6. Use $\hat{D}$ to build stochastic model

3.2.2 Algorithm 2

Because we can locate and simulate peaks and valleys of thickness, we can simply use bicubic interpolation based on the above peaks and valleys. The resulting $Alg2$ in Table *Algorithm 2* is simple yet effective to be shown in the next Section 3.2.3.

Algorithm 2:
1. Locate and simulate peaks and valleys of thickness.
2. Perform bicubic interpolation to obtain $\hat{D}$.
3. Use $\hat{D}$ to build stochastic model.

3.2.3 Accuracy Comparison

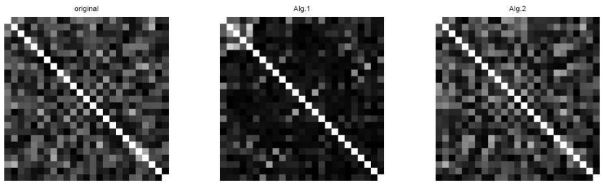

Figure 14: Correlation matrix image comparison for M_1: (a) Original data (b) Alg.1 (c) Alg.2

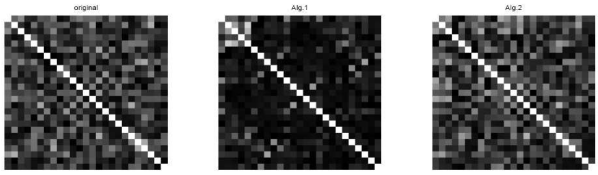

Figure 15: Correlation matrix image comparison for ILD_1: (a) Original data (b) $Alg1$ (c) $Alg2$

In this part, we compare $Alg1$ and $Alg2$ with the original model where CMP simulation is performed for each data point. In Tables 2 and 3, "Error on Mean" is the average error for the mean thickness for all regions in the stochastic model, "Relative Variance" is average of relative variance (defined as variance divided by mean value within each region) for all regions in the stochastic model, and "Simulation Number" is the number of CMP simulation runs needed in speedup algorithm divided by total simulation runs needed by the original model.

As shown in Table 2, both $Alg1$ and $Alg2$ are accurate enough in terms of mean thickness. $Alg2$ is $3X$ more accurate than $Alg1$ on average. In addition, not shown in the table, the maximum errors on mean thickness estimation for both $Alg1$ and $Alg2$ are less than 1%. However, both algorithms tend to under-estimate variances because the FFT interpolates in $Alg1$ and bicubic interpolation in $Alg2$ smooth (reduce) the variations of estimated thicknesses. Compared to $Alg1$, $Alg2$ is about $3X$ better in variances estimation as shown in the tables, especially in higher layers when relative variation is large. Figures 14 and 15 compare the correlation matrices for the original model and the models generated by $Alg1$ and $Alg2$. It is clear that $Alg2$ is reasonably accurate and is more accurate than $Alg1$ in terms of spatial correlation.

Although $Alg2$ achieves better accuracy, it needs to locate peaks and valleys of thickness, which can be done at some extra run time. On the other hand, $Alg1$ is easier to implement and has a constant number of simulation runs. The number of simulation runs for $Alg2$ changes with respect to different designs, and is $1.5X$ more than $Alg1$ in this case study. The two algorithms offers a tradeoff between accuracy and time complexity and may be suitable for different tool flows.

4. CONCLUSIONS AND DISCUSSIONS

In this paper, we have revealed several interesting characteristics (*Observations 1-5*) on thickness variations and vertical and horizontal correlations between the variations. We have also proposed a stochastic CMP model and developed two algorithms to reduce the CMP simulation runs by $7X$ and $3X$, respectively, to build this stochastic model.

It is easy to see that we can leverage the vertical cor-

	Error on Mean		Relative Variance			Simulation Number	
Layer	Alg.1	Alg.2	Accurate	Alg.1	Alg.2	Alg.1	Alg.2
1	0.16%	0.03%	0.22%	0.06%	0.15%	14%	30%
2	0.05%	0.01%	0.05%	0.01%	0.03%	14%	32%
3	0.13%	0.02%	0.19%	0.05%	0.13%	14%	32%
4	0.10%	0.02%	0.13%	0.03%	0.10%	14%	36%
5	0.18%	0.03%	0.55%	0.21%	0.42%	14%	37%
6	0.32%	0.17%	4.94%	0.89%	3.28%	14%	33%
7	0.24%	0.12%	4.97%	0.68%	3.75%	14%	34%
8	0.12%	0.09%	3.01%	0.45%	2.40%	14%	36%

Table 2: Comparison for metal thickness.

	Error on Mean		Relative Variance			Simulation Number	
Layer	Alg.1	Alg.2	Accurate	Alg.1	Alg.2	Alg.1	Alg.2
1	0.17%	0.04%	0.22%	0.06%	0.16%	14%	30%
2	0.11%	0.02%	0.27%	0.07%	0.19%	14%	31%
3	0.11%	0.02%	0.20%	0.05%	0.14%	14%	33%
4	0.13%	0.03%	0.38%	0.08%	0.24%	14%	33%
5	0.24%	0.03%	0.72%	0.27%	0.56%	14%	34%
6	0.34%	0.16%	6.55%	1.21%	4.50%	14%	32%
7	0.34%	0.16%	11.46%	1.74%	7.29%	14%	33%
8	0.21%	0.20%	8.10%	1.13%	5.69%	14%	36%

Table 3: Comparison for ILD thickness.

relation between thickness variations (Observations 3-5) to further speed up CMP modeling and likely to obtain another 2X-3X speedup. This is part of our future work. In addition, we will develop interconnect routing and optimization considering CMP variations with vertical (inter-layer) correlation which has been ignored by the existing work. Finally, the large amount of variation (up to 10%) on thickness may cause defocusing in lithography and in turn affect interconnect width and spacing. The interaction between CMP and lithography will be studied in the future as well.

5. REFERENCES

[1] D.Woodie, "hemical mechanical polishing (cmp)," in *Cornell Univ., http://www.cnf.cornell.edu/doc/CMP%20Primer.pdf*, 2001.

[2] V. Mehrotra, "Modeling the effects of systematic process variation on circuit performance," in *PhD thesis, Massachusetts Institute of Technology*, 2001.

[3] L. He, A. B. Kahng, K. Tam, and J. Xiong, "Design of ic interconnects with accurate modeling of cmp," in *International Society for Optical Engineering (SPIE) Symposium on Microlithograhpy, pp. 109-119*, 2005.

[4] Y. Chen, A. B. Kahng, G. Robins, A. Zelikovsky, and Y. Zheng, "Area fill generation with herent data volume reduction," in *Proc. European Design and Test Conf. (DATE)*, 2003.

[5] B. Tang and D. Boning, "Cmp modeling and characterization for polysilicon mems structures," in *Material Research Society Symposium, K7.5*, 2004.

[6] 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, pp. 605–615*, 1999.

[7] "http://www.cadence.com/," in *Cadence Design Systems*, 2006.

[8] L. He, A. B. Kahng., K. Tam, and J. Xiong, "Variability-driven considerations in the design of integrated-circuit global interconnects," in *IEEE VLSI Multilevel Interconnection Conference, pp. 214-221*, 2004.

[9] L. He, A. B. Kahng, K. Tam, and J. Xiong, "Simultaneous buffer insertion and wire sizing considering systematic cmp variation and random leff variation," in *Proc. Int. Symp. on Physical Design (ISPD)*, 2004.

[10] M. Cho, D. Pan, H. Xiang, and R. Puri, "Wire density driven global routing for cmp variation and timing," in *Proc. Int. Conf. on Computer Aided Design (ICCAD)*, 2006.

[11] "http://www.mathworks.com/," in *Math Works Inc.*, 2006.