Handling Partial Correlations in Yield Prediction

Sridhar Varadan
Dept. of ECE
Texas A&M University
College Station, TX 77843, USA
sridhar@ece.tamu.edu

Janet Wang
Dept. of ECE
University of Arizona
Tucson, AZ 85721, USA
wml@ece.arizona.edu

Jiang Hu
Dept. of ECE
Texas A&M University
College Station, TX 77843, USA
jianghu@ece.tamu.edu

Abstract— In nanometer regime, IC designs have to consider the impact of process variations, which is often indicated by manufacturing/parametric yield. This paper investigates a yield model - the probability that the values of multiple manufacturing/circuit parameters meet certain target. This model can be applied to predict CMP (Chemical-Mechanical Planarization) yield. We focus on the difficult cases which have large number of partially correlated variations. In order to predict the yield for these difficult cases efficiently, we propose two techniques: (1) application of Orthogonal Principle Component Analysis (OPCA); (2) hierarchical adaptive quadrisection (HAQ). Systematic variations are also included in our model. Compared to previous work, the OPCA based method can reduce the error on yield estimation from $17.1\%-21.1\%$ to $1.3\%-2.8\%$ with $4.6\times$ speedup. The HAQ technique can reduce the error to $4.1\% - 5.6\%$ with $6\times -9.4\times$ speedup.

I. Introduction

When VLSI feature size shrinks to nanometer regime, manufacturing process variations are no longer negligible compared to corresponding nominal values [1]. Consequently, manufacturing and parametric yield need to be considered in circuit design stages. Roughly speaking, manufacturing yield refers to the probability that certain manufacturing spec is satisfied. Likewise, parametric yield is the probability that the target performance metrics, such as timing and power, are met. Accurate and efficient yield prediction models can guide circuits designs toward low variability and/or high tolerance to process variations.

In this work, we focus on a specific yield model that can handle the probability for m random variables within a given range. This model is applicable to either manufacturing variations or parametric variations. For example, if the m random variables represent metal thickness, this model can predict CMP (Chemical-Mechanical Planarization) yield [2]. As pointed out by previous research works, after the CMP procedure, metal thickness may have systematic variations depending on layout patterns and random variations due to CMP process fluctuations [3]. Figure 1 illustrates such variations. Too large variations may cause considerable performance deviations as well as the risk of open/short circuit. Therefore, it is highly desired that the thicknesses of all metals on the same layer are within certain range. In addition to CMP yield, this model can be extended to predict timing yield of sequential circuits [4] if the random variables correspond to the maximal combinational path delays.

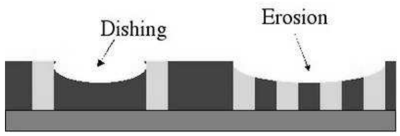

Fig. 1. Metal thickness variations after CMP (Chemical-Mechanical Planarization).

In general, process variations include systematic variations and random variations. Systematic variations can approximately be traced out according to circuit designs. Random variations can be further decomposed to inter-die and intra-die variations. For a manufactured chip, the inter-die variations are perfectly correlated and therefore can be treated as a single random component. Obviously, it is easy to estimate the yield for a single random variable. Intra-die variations, on the other hand, consist of independent parts and partially correlated parts. If the m random variables representing intra-die variations are independent of each other, the overall yield is the product of the probability for each individual variable within the given range. Compared to the perfectly correlated and independent random variables, the case of partially correlated random variables is much more difficult to handle. In theory, the yield of partially correlated variations can be obtained via numerical integration over the joint distribution function [5]. In practice, such computation can be quite expensive when the number of random variables is large.

Although partial correlation causes difficulties, it still allows the problem dimension m to be reduced to certain extent, as correlations are often spatially dependent. For instance, [2] clustered the random variables according to spatial proximity and treat the variables within each cluster as perfectly correlated. Then, it managed to reduce the problem dimension to the number of clusters. However, this heuristic faces a dilemma: cluster size influences the accuracy of the yield estimation and the associated computation cost in opposite ways. Large cluster size leads to over-estimated yield while small size results in high computational cost. In extreme cases, for example, when the number of variables m is very large, it is almost impossible to find a cluster size that leads to both high credibility and reasonable computation cost.

In order to handle partial correlations in the yield prediction efficiently, we propose two techniques. The first one is based on Orthogonal Principle Component Analysis (OPCA), which

978-1-4244-1922-7/08/$25.00 ©2008 IEEE

is a more formal treatment to partial correlations compared to the clustering based method [2]. The second one is a hierarchical adaptive quadrisection (HAQ) technique which can form heterogeneous cluster sizes according to systematic variations. Experiments are performed on large testcases, each of which has $360K$ variation locations. Compared to [2], which is the only previous work to the best of our knowledge, the OPCA based method can reduce the error on yield estimation from $17.1\% - 21.1\%$ to $1.3\% - 2.8\%$ with $4.6\times$ speedup. The HAQ technique can reduce the error to $4.1\% - 5.6\%$ with $6\times - 9.4\times$ speedup.

The rest of this paper is organized as follows. The variation and yield model in this work will be defined in Section II. In Section III, we will briefly review the previous work on this problem. Our new reduction techniques will be described in Section IV. Section V will provide the experimental results. In Section VI, we will give the conclusion and discuss future works.

II. Variation and Yield Model

The variation and yield models in this work are based on the case of metal thickness like in [2]. With small modifications, they can be applied to other cases, such as sequential timing yield.

Consider the metal thicknesses at m locations which are represented by an m-dimensional vector $\vec{p} = (p_1, p_2, ..., p_m)^T$. Then, each thickness $p_i \forall i \in \{1, 2, ..., m\}$ can be decomposed as:

$$p_i = \mu_i + \delta_i \quad \text{and} \quad \mu_i = \bar{\mu} + \Delta_i \tag{1}$$

where $\bar{\mu}$ is the nominal value, Δ_i denotes the systematic variation and δ_i is the random variation. The nominal value $\bar{\mu}$ is a constant and corresponds to the dashed line in Figure 2. The systematic variation Δ_i is a deterministic value depending on the layout pattern around metal segment i. We define $\mu_i = \bar{\mu} + \Delta_i$ as **deterministic thickness**, which is shown as black dots in Figure 2. The random variations, including inter-die variations, intra-die correlated variations and intra-die independent variations, are indicated by the vertical segments with double-arrow in Figure 2. Same as in [2], we assume that the random variations follow normal distributions with roughly equal variance. Then, the thickness vector $\vec{p}$ can be represented by a multivariate normal distribution $N(\vec{\mu}, \Sigma)$ where $\vec{\mu} = (\mu_1, \mu_2, ..., \mu_m)^T$ and Σ is an $m \times m$ covariance matrix. More specifically, the joint distribution can be described as:

$$\Phi(\vec{p}) = \frac{e^{-\frac{1}{2}(\vec{p}-\vec{\mu})^T \Sigma^{-1} (\vec{p}-\vec{\mu})}}{\sqrt{(2\pi)^m |\Sigma|}} \tag{2}$$

where $|\Sigma|$ is the determinant of the covariance matrix.

According to [2], **CMP yield** Y is defined as the probability that all values of $\vec{p}$ are within a given range $[L, U]$. In Figure 2, this corresponds to the probability that all thickness values are in the shaded region. In CMP yield prediction, an entire chip area is tessellated into an array of tiles. The metal thickness in each tile τ_i is represented by p_i. Since there could be sharp systematic variations, an accurate yield prediction requires fine-grained tessellation, i.e., a large number of small tiles.

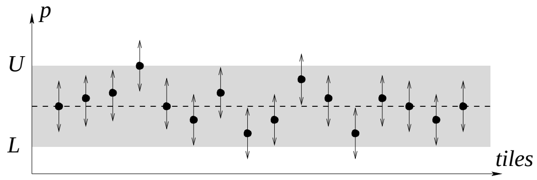

Fig. 2. Systematic and random variations of metal thickness.

III. Previous Work

In theory, the yield Y defined in the previous section can be obtained by computing

$$Y = \int_L^U \int_L^U ... \int_L^U \Phi(\vec{\mu}) dp_1 dp_2 ... dp_m \tag{3}$$

However, numerical integration on high dimensional problems is usually very expensive. In reality, the dimension m (or the number of tiles) for CMP yield prediction is in the order of 10^5 to 10^6 [2], which makes the numerical integration unaffordable.

In [2], a clustering based divide-and-conquer method was introduced. It first decomposes the CMP yield into upper yield and lower yield. Upper yield Y_U (lower yield Y_L) is the probability that all thickness values are no greater than the upper constraint U (no less than the lower constraint L). Then, the overall yield is

$$Y = Y_U + Y_L - 1 \tag{4}$$

The upper yield Y_U and lower yield Y_L can be computed separately yet in the same manner. It is observed that the correlation between the variations at two different locations decreases with the distance between them. The work of [2] defines perfect correlation circle (**PCC**) based on this observation. PCC is a circle where all variations inside can be assumed as perfectly correlated. This assumption is valid and accurate only when the radius of PCC is small.

When computing upper yield Y_U, the method used in [2] first finds the tile τ_i with the maximum deterministic thickness, i.e., the tile τ_i with the maximum value of μ_i. Then, a PCC is built centered at this tile. Since the variations in the PCC are assumed to be perfectly correlated, their impact to Y_U can be degenerated to p_i. For those regions not covered by this PCC, the tile τ_j with the maximum value μ_j is picked and the previous procedure is repeated. In the end, the entire chip is covered by these PCCs and Y_U is computed based on the tiles with the maximum μ in each PCC. This approach reduces the dimension of the problem to the number of PCCs. After the reduction, the upper yield Y_U is computed using Genz's algorithm [5], which is an efficient numerical integration method. The lower yield Y_L can be handled in the same manner.

In order to reduce computation runtime, a large radius is preferred for the PCCs. However, a large radius may result in overestimation of the yield. For example, consider the case that the center tile τ_i of a PCC has $\mu_i = 402nm$ and a boundary tile τ_j of the same PCC has $\mu_j = 400nm$. According to [2], if $p_i = \mu_i + \delta_i \le U$, then $p_j = \mu_j + \delta_j \le U$, as δ_i and δ_j are assumed to be perfectly correlated. However, when the radius is large, such assumption may cause errors. If δ_i and δ_j are in fact not perfectly correlated due to a large radius, it is likely that $p_j = \mu_j + \delta_j > U$ even when $p_i = \mu_i + \delta_i \leq U$. Therefore, the method of [2] may miss such violation and over-estimate the yield.

IV. New Reduction Methods

Given a yield prediction problem with large number of partially correlated random variables, our approach is to first reduce the number of variables and then perform numerical integration using Genz's algorithm [5] like in [2]. The focus of our work is on the variable reduction. We propose two reduction methods in order to improve the accuracy of the prediction as well as the computation efficiency.

A. Reduction with Orthogonal Principle Component Analysis (OPCA)

With the large number of correlations in the metal thicknesses due to the local variations, the resultant performance models have a number of random variables. The goal of using OPCA is to compute the most meaningful basis to re-express these correlated random variables into a set of independent and less number of random variables through an orthogonal base. Determining this orthogonal basis allows circuit designers to discern which random variables or variation sources are important, which are just redundant variables and which are just noise [6, 7].

In the CMP example, we treat m local metal thicknesses as m random variables. Let vector $\vec{\delta} = \{\delta_1, \delta_2, \cdots, \delta_m\}$ represent the m random parameters. The possible correlations among different metal thicknesses can be denoted by a correlation matrix

$$\Gamma(\vec{\delta}) = (\Gamma_{ij})_{m \times m} \tag{5}$$

Assume the variance of each metal thickness is σ_i^2, the covariance matrix Σ can be obtained as

$$\Sigma(\vec{\delta}) = (\Gamma_{ij}\sigma_i\sigma_j)_{m \times m} \tag{6}$$

It is easy to prove that this covariance matrix is symmetric. Each entry of this matrix is covariance. By definition, covariance must be non-negative, therefore the minimal covariance is zero. Covariance has important physical meanings. While the variance measures the perturbation of each random variable from its mean value, the covariance indicates the degree of the linear relationship between every two random variables. A small (large) value reveals low (high) redundancy or dependency.

Our goal is to find the dominant random variables by maximizing the main impact of these variables measured by variance while minimizing redundancy measured by covariance. This target can be achieved by performing eigenvalue decomposition on symmetric covariance matrix $\Sigma(\vec{\delta})$,

$$\Sigma(\vec{\delta}) = Q\Lambda(\vec{\delta})Q^T \tag{7}$$

where Q is a $m \times m$ matrix with column vectors representing eigenvectors. Here $\Lambda(\vec{\delta})$ is a diagonal matrix with eigenvalues λ_i at the diagonal locations.

$$\Lambda(\vec{\delta}) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix} \tag{8}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$. This eigenvalue decomposition on the covariance matrix serves two purposes: it provides the dominant directions of the covariance relationship by a diagonal matrix $\Lambda(\vec{\delta})$; it indicates, as an initial step, how we can map the original vector $\vec{\delta}$ to a new vector. Assume we would like to map the $m \times 1$ vector $\vec{\delta}$ to a new one $\vec{\xi} = \{\xi_1, \xi_2, \cdots, \xi_m\}$.

$$\vec{\delta} = B\vec{\xi} \tag{9}$$

where B is a $m \times m$ matrix and $\vec{\xi}$ is a $m \times 1$ vector. How to find out B and $\vec{\xi}$? Without loss of generality, we assume that the transferred variational sources have Gaussian distribution and can be standardized as

$$\begin{aligned} \mu(\vec{\xi}) &= \vec{0} \\ \Lambda(\vec{\xi}) &= I \end{aligned} \tag{10}$$

One can easily deduce that there exists a matrix J with dimension $m \times m$ such that

$$\Lambda(\vec{\delta}) = J\Lambda(\vec{\xi})J^T = JJ^T \tag{11}$$

And J can be obtained as

$$J = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \sqrt{\lambda_m} \end{bmatrix} \tag{12}$$

J together with Q gives the map $B = QJ$ that can transfer the original vector $\vec{\delta}$ to $\vec{\xi}$:

$$\vec{\delta} = B\vec{\xi} = QJ\vec{\xi} \tag{13}$$

It is obvious that the covariance matrix can be decomposed as

$$\Sigma(\vec{\delta}) = Q\Lambda(\vec{\delta})Q^T = QJ\Lambda(\vec{\xi})(QJ)^T \tag{14}$$

Figure 3 explains this mapping operation with a 2D case. By using eigenvalue decomposition on $\Sigma(\vec{\delta})$, we find a set of independent vectors (eigenvectors Q_i) to express the correlated random variable vector. Then we project this set of independent vectors to a set of orthogonormal vectors (vector $(QJ)_i$s). Thus, we successfully decompose the correlated random variable vector to a orthogonormal, independent one.

Remember that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$. It is possible that some eigenvalues are much smaller that the others. By neglecting the small eigenvalues, we reduce the number of variation sources. That is, we approximate $\Lambda(\vec{\delta})$ as

$$\Lambda(\vec{\delta}) \approx \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \tag{15}$$

Because

$$\Lambda(\vec{\delta}) = JJ^T \tag{16}$$

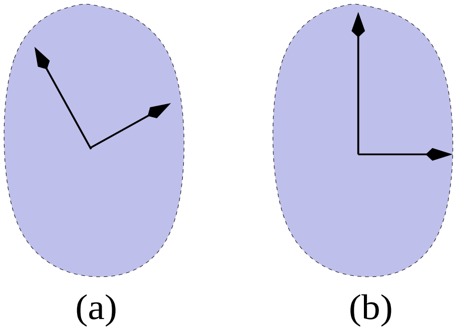

Fig. 3. OPCA finds out the dominant directions of sampling space for the random variables. (a) directions of the sampling space for the random variables; b) adjusted directions for the sampling space.

Then, $J_{m\times m}$ becomes $J_{k\times k}$ and can be formulated as

$$J = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \sqrt{\lambda_k} \end{bmatrix} \qquad (17)$$

The correspondent matrices in deriving the mapping matrix procedure $B = QJ$ follows the dimensions shown in Figure 4(a). This B matrix maps k variables in $\vec{\xi}$ to m variables in $\vec{\delta}$ by $\vec{\delta} = B\vec{\xi}$. Figure 4(b) demonstrates the dimensions in this case. Because OPCA approach in the current paper employs eigen-decomposition, the column vectors in B matrix are orthogonal to each other.

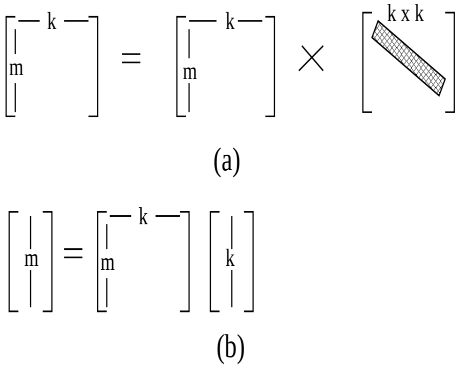

Fig. 4. Dimension demonstration for the reduction procedure, (a) dimension display for $B = QJ$; (b)dimension display for $\vec{\delta} = B\vec{\xi}$

In summary, performing OPCA requires four steps: 1) extract the correlation matrix of the metal thicknesses from CMP procedure; 2) organize the covariance matrix for random variables representing metal thicknesses; 3) perform eigen-decomposition for the covariance matrix; 4) construct mapping matrix B for dimension reduction. After the OPCA method, the large number of correlated random variables are transferred into a smaller set of independent random variables.

B. Reduction with Hierarchical Adaptive Quadrisection (HAQ)

In this method, we divide the entire chip area into a set of **basic subregions** and let the variations in each basic subregion be characterized by a single random variable. Similar as the PCCs in [2], each basic subregion usually contains many tiles. By this division, the total number of random variables in the yield prediction is largely reduced. In contrast to the uniform PCC size in [2], the sizes of the basic subregions may be different from each other. The size of each basic subregion is decided according to systematic variations. The purpose of the heterogeneous granularity is to minimize the number of basic subregions with limited side-effect on the accuracy of yield prediction.

We use the procedure of computing the upper yield Y_U to illustrate the hierarchical adaptive quadrisection method. The same method can be applied to compute the lower yield Y_L as well. At the beginning, we divide the entire chip area into an array of relatively large subregions, i.e., a coarse-grained array. For each of the large subregions, we perform hierarchical adaptive quadrisection as follows. For a subregion S, we first find the tile τ_i with the maximum deterministic thickness, i.e., $\mu_i = \mu_{max}$ in S. Next, the subregion S is temporarily quadrisected into four plates $\{P_1, P_2, P_3, P_4\}$ of the same size. The plate P_k containing tile τ_i is called **critical plate** and the others are called **non-critical plates**. For each non-critical plate P_j, we identify one of its tiles that has the maximum deterministic thickness $\mu_{j,max}$. Then, we compute the **critical difference** $d = \min_{\forall j\in\{1,2,3,4\},\ j\neq k}(\mu_{max} - \mu_{j,max})$. This is the minimum difference between μ_{max} and the maximum deterministic thickness of each non-critical plate. If $d > \theta$, where θ is a constant threshold, we keep subregion S as a basic subregion without further division. Otherwise, we divide S according to this temporary quadrisection and repeat this procedure for each of the four plates, recursively.

```
Procedure: $AdaptiveQuadrisection(S)$
Input: A layout region $S$ consisting of an array of tiles
Output: A set of subregions $P$ covering $S$
1. Find tile $\tau_i \in S$ with max deterministic thickness $\mu_{max}$
2. Temporarily quadrisect $S$ into plates $\{P_1, P_2, P_3, P_4\}$
3. Identify critical plate $P_k$ that contains $\tau_i$
4. Find the max deterministic tile thickness $\mu_{j,max}$
       for all plates except $P_k$
5. $d = \min_{\forall j\in\{1,2,3,4\},\ j\neq k}(\mu_{max} - \mu_{j,max})$
6. If $d$ is greater than a threshold $\theta$, $P \leftarrow S$
7. Else
8.    $P \leftarrow \emptyset$
9.    For $j = 1$ to $j = 4$
10.       $P \leftarrow P \cup AdaptiveQuadrisection(P_j)$
11. Return $P$
```

Fig. 5. Algorithm of hierarchical adaptive quadrisection.

The pseudo code of this algorithm is given in Figure 5. The key step is step 6. If the value of the critical difference d is large, the tile thicknesses in all non-critical plates are significantly smaller than μ_{max}. Hence, the impact of thickness variations in the non-critical plates are dominated by μ_{max}. In other words, if the thickness of critical tile $p_i \leq U$, then we can safely assume that the thickness in non-critical plates are no greater than U. Therefore, the probability of satisfying the upper constraint U is approximately decided by μ_{max} and the quadrisection on S is unnecessary.

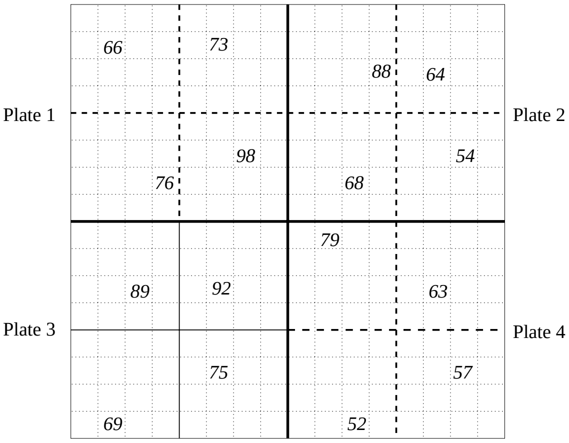

Fig. 6. An example of hierarchical adaptive quadrisection. The tiles are indicated by the dotted lines.

Figure 6 is an example of the hierarchical adaptive quadrisection procedure. At first, we find the maximum deterministic thickness, which is 98. Then, the given subregion is temporarily quadrisected into four plates. Since the maximum deterministic thickness 98 lies in plate 1, plate 1 is the critical plate and the others are non-critical plates. Next, the maximum deterministic thickness of each non-critical plate is found, 88 in plate 2, 92 in plate 3 and 79 in plate 4. One can see that the critical difference is 6, which occurs between 98 and the maximum deterministic thickness of plate 3. If $6 < \theta = 10$, we divide the subregion according to this quadrisection and repeat this procedure for all four plates. In the next level, the critical differences are 22, 20, 3 and 16 for the four plates, respectively. Only in plate 3, the critical difference $3 < \theta = 10$. Therefore, plate 3 is divided again according to quadrisection while the other plates are unchanged and become basic subregions. In Figure 6, temporary quadrisections are shown as dashed lines and the finally realized quarisections are indicated by solid lines. At the end, we have 7 basic subregions in this example. If we use uniform sizes as in [2], either we have 4 basic subregions which are too coarse, or 16 basic subregions which are too fine-grained. This example clearly shows the flexibility and advantage of our method.

V. Experimental Results

The experiment setup is similar as that of [2]. The size of entire chip is $4.8mm \times 7.5mm$. It is tessellated into a 480×750 array of tiles, i.e., each tile has size of $10\mu m \times 10\mu m$. The nominal thickness is $\bar{\mu} = 0.358\mu m$. The lower and upper thickness constraints are $L = 0.258\mu m$ and $U = 0.458\mu m$, respectively. For the random variations, the standard deviation is $0.03\mu m$. Same as [2], the correlation coefficient between the variations of two tiles is modelled as

$$\gamma = -\alpha \times 10^{-5} x + 0.9958$$

where x is the distance between the two tiles and α is a constant. When α increases, the correlation decreases faster with the distance. Based on this setup, we obtained three testcases, **case 1**, **case 2** and **case 3**, which have different systematic variations (or deterministic thickness profiles) and different values of α. The α values for the three cases are 2, 3 and 4, respectively.

We compared the following methods in the experiments:

- Monte Carlo (MC): 50K-run Monte Carlo simulation is performed for each case. In the simulation, the correlations are handled by OPCA, but without dimension reduction. In general, such results can serve as a baseline for evaluating the accuracy of yield prediction.
- Perfect correlation circle (PCC) based method [2]: To the best of our knowledge, this is the only previous work on this yield prediction problem. It first reduces the number of random variables using PCC and then obtains the yield based on Genz's algorithm [5]. We tested for the PCC approach using two different radii: PCC1 with radius of $150\mu m$ and PCC2 with radius of $250\mu m$. In [2], it is assumed that variations within $200\mu m$ are perfectly correlated. Therefore, we tested two radii around $200\mu m$.
- Orthogonal principle component analysis (OPCA): Variable reduction using OPCA followed by Genz's algorithm. We tried two different levels of reductions: OPCA1 which reduces the number of variables from $360K$ to 300, and OPCA2 which reduces the number to 200.
- Hierarchical adaptive quadrisection (HAQ): Variable reduction using HAQ followed by Genz's algorithm. We performed HAQ with two different values of threshold θ (see line 6 of Figure 5): HAQ1 where $\theta = 0.09\mu m$ and HAQ2 where $\theta = 0.075\mu m$. A larger threshold θ usually implies finer granularity of the basic subregions.

All of these methods are implemented in MATLAB. The experiments are performed on a Windows machine with 1.6GHz CPU and 1GB memory.

TABLE I
Results from Monte Carlo (MC) simulations.

		MC without OPCA		MC with OPCA	
Testcase	α	Yield	CPU(sec)	Yield	CPU(sec)
Case 1	2	60%	6165	76%	6516
Case 2	3	60%	6191	74%	6518
Case 3	4	60%	6158	71%	6472

The results from Monte Carlo simulations are summarized in Table I. For reference, we also include the results of Monte Carlo without considering correlations (without using OPCA) in the 3rd and 4th column. Since correlation is not addressed, the yield results in the 3rd column are independent of the value of α, which is a part of correlation model. One can see that neglecting correlation may significantly under-estimate the yield.

The yield results in column 5 include the effect of correlations and will serve as baselines for yield accuracy evaluation. The runtime of Monte Carlo simulations is nearly two hours for each case.

TABLE II
EXPERIMENTAL RESULTS FROM PCC [2], OPCA AND HAQ.

Testcase	Method	# variables	Yield	CPU(sec)
Case 1	PCC1	431/435	89%	2242
	PCC2	305/310	90%	1636
	OPCA1	300/300	77%	481
	OPCA2	200/200	78%	470
	HAQ1	175/178	80%	372
	HAQ2	153/155	82%	312
Case 2	PCC1	432/427	88%	2238
	PCC2	305/310	88%	1619
	OPCA1	300/300	76%	482
	OPCA2	200/200	77%	469
	HAQ1	80/79	77%	239
	HAQ2	61/61	79%	221
Case 3	PCC1	429/425	86%	2214
	PCC2	307/308	87%	1649
	OPCA1	300/300	73%	476
	OPCA2	200/200	74%	463
	HAQ1	172/170	75%	361
	HAQ2	148/143	76%	316

The results from PCC, OPCA and HAQ are shown in Table II. The 3rd column tells the number of variables after reduction. Since the overall yield is obtained from upper yield and lower yield according to Equation (4), the two numbers in the 3rd column correspond to the numbers of variables for computing the upper yield and the lower yield, respectively. For each method, the first variant (PCC1, OPCA1 and HAQ1) has less variable reduction. Consequently, they provide yield results closer to that of Monte Carlo simulation, i.e., more accurate results, compared to the second variant (PCC2, OPCA2 and HAQ2). Evidently, the runtime of the first variant is always larger than that of the second variant.

TABLE III
COMPARISONS AMONG PCC, OPCA AND HAQ.

Testcase	Method	Yield Error	Speed
Case 1	PCC1	17.1%	1×
	OPCA1	1.3%	4.7×
	HAQ1	5.3%	6.0×
Case 2	PCC1	18.9%	1×
	OPCA1	2.7%	4.6×
	HAQ1	4.1%	9.4×
Case 3	PCC1	21.1%	1×
	OPCA1	2.8%	4.7×
	HAQ1	5.6%	6.2×

Since the results from the first variant of these methods are more accurate, we compare these methods based on their first variant in Table III. The 3rd column shows the errors of these methods with respect to the results of Monte Carlo with OPCA. One can see that both of our methods lead to much less errors than the previous work of PCC [2]. At the same time, our methods are much faster than PCC. The speedup from our methods is listed in the rightmost column.

VI. CONCLUSION AND FUTURE WORK

In yield prediction, it is difficult to handle large number of random variables with partial correlations. We propose two reduction techniques for such difficult cases. Compared to previous work, our reduction techniques can significantly improve both the accuracy and the speed of yield prediction. In this paper, we applied our techniques for predicting CMP yield. In future, we will extend these techniques for estimating timing yield of sequential circuits.

REFERENCES

[1] S. R. Nassif. Modeling and analysis of manufacturing variations. In *Proceedings of the IEEE Custom Integrated Circuits Conference*, pages 223–228, 2001.

[2] J. Luo, S. Sinha, Q. Su, J. Kawa, and C. Chiang. An IC manufacturing yield model considering intra-die variations. In *Proceedings of the ACM/IEEE Design Automation Conference*, pages 749–754, 2006.

[3] T. Tugbawa, T. Park, D. Boning, T. Pan, P. Li, and S. Hymes. A mathematical model of pattern dependencies in Cu CMP process. In *Proceedings of the Electrochemical Society CMP Symposium*, 1999.

[4] M. Pan, C. C.-N. Chu, and H. Zhou. Timing yield estimation using statistical static timing analysis. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, pages 2461–2464, 2005.

[5] A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–149, 1992.

[6] J. Shlens. Tutorial on principal component analysis. Systems Neurobiology Laboratory, University of California at San Diego, December 2005.

[7] R. Jiang, W. Fu, J. M. Wang, V. Lin, and C. C.-P. Chen. Efficient statistical capacitance variability modeling with orthogonal principle factor analysis. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pages 683–690, 2005.