A Data-Driven Statistical Approach to Analyzing Process Variation in 65nm SOI Technology

Choongyeun Cho[1], Daeik Kim[1], Jonghae Kim[1], Jean-Olivier Plouchart[1], Daihyun Lim[2], Sangyeun Cho[3], and Robert Trzcinski[1]

[1]IBM, [2]MIT, [3]U. of Pittsburgh

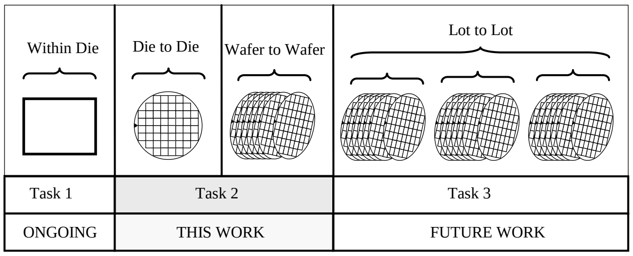
{cycho,dkim,jonghae,plouchar,rtrzcin}@us.ibm.com, daihyun@mit.edu, cho@cs.pitt.edu

Abstract

This paper presents a simple yet effective method to analyze process variations using statistics on manufacturing in-line data without assuming any explicit underlying model for process variations. Our method is based on a variant of principal component analysis and is able to reveal systematic variation patterns existing on a die-to-die and wafer-to-wafer level individually. The separation of die variation from wafer variation can enhance the understanding of a nature of the process uncertainty. Our case study based on the proposed decomposition method shows that the dominating die-to-die variation and wafer-to-wafer variation represent 31% and 25% of the total variance of a large set of in-line parameters in 65nm SOI CMOS technology.

1. Introduction

As the feature size of silicon technology is scaling down and the wafer size is getting larger, process variation is increasingly difficult to model and control, thus becoming a critical limiting factor of the performance and yield of integrated circuits [1].



(Within-die represents a process variation in the identical device or circuit within a die. Die-to-die means a process variation in different dies within a wafer. Wafer-to-wafer is a variation in different wafers within a lot. Lot-to-lot denotes a variation in different lots.)

Figure 1. Definition of four ranges of process variation

For fault detection and device characterization, in-line electrical measurements are typically performed off the manufacturing floor using a parametric tester. To keep track of the performance and DC characteristics of devices or other circuit elements, assorted test structures and conditions have been employed. For example, FET devices of different sizes and layouts are designed and fabricated for the purpose of regular monitoring of critical electrical parameters such as threshold voltage, drive current, and leakage current. However, the holistic view of the collection of heterogeneous in-line data has been largely neglected thus far. To the best of the authors' knowledge, there has been little research or practice exploiting in-line electrical measurement data of a large number of variation parameters from multiple wafers/lots to extract process variation in wafer-to-wafer and die-to-die levels, separately.

In this paper, we first propose a statistical method to analyze manufacturing in-line data to separate die variation and wafer variation. Using the proposed method and a set of pre-production manufacturing in-line data collected from test structures built with a 65nm SOI CMOS technology, we evaluate the relative amount of systematic die-to-die and wafer-to-wafer variations in the total in-line measurement data. Along with sensitivity analysis of circuit performance to the variation parameters, the contributions of systematic die-to-die and wafer-to-wafer variation can be evaluated separately. Our method also allows us to assess the effect of random variation, which is left as residual and cannot be explained by systematic variation components.

2. Proposed method

In this section, a multivariate statistical analysis technique is presented to separately monitor die-level and wafer-level systematic variations from the observation of in-line measurement data. Due to the complexity of semiconductor manufacturing processes and environmental factors, die and wafer variations are more or less inter-correlated depending on a specific fabrication recipe. The rationale to untangle die variation and wafer variation is to make it easier to conceptualize and analyze a given process variation and its physical mechanism, than otherwise leaving it as a lumped variation.

2.1 Principal component analysis

The principal component analysis (PCA) is a linear transformation of a set of random vectors to a new set of vectors, referred to as principal components (PC's) [2]. PC's are uncorrelated and are ordered so that a first few retain most of the variation present in all the original variables. The first PC is visualized as the direction on which the variance of the projection of the original vector is maximized as expressed in (1). Here, x is the original data vector, and w_i is the PC. The subsequent PC's are defined in the same way except that they need to be orthogonal to all the previous PC's.

$$\left.\begin{aligned} w_1 &= \arg\max_{\|w\|=1} \operatorname{var}(w^T x) \\ w_k &= \arg\max_{\|w\|=1,\, w\perp w_k \forall i=1,\ldots,k-1} \operatorname{var}(w^T x), \quad k \geq 2 \end{aligned}\right\} \qquad (1)$$

In reality, PCA is often implemented with the singular value decomposition (SVD) of a covariance matrix of a given data set. If the covariance matrix is not known *a* *priori*, it is often estimated based on ensemble of data samples. The PCA is a useful multivariate tool to reduce the dimension of data set, to reduce noise, or to visualize the representative features of the given multidimensional data. There has been a growing interest in PCA in semiconductor manufacturing industry for process failure analysis [3,4]; however, to the authors' knowledge, there was no previous work to treat die-to-die or wafer-to-wafer variations separately using PCA.

2.2 Constrained principal component analysis

The constrained principal component analysis (CPCA) is a method to extract constrained principal components (CPC's) which have the same properties with the original PC's but are constrained to a predefined subspace [5]. Similar to the ordinary PCA, the CPCA finds CPC's in a sequence of significance. It is useful to extract the PC's of die-to-die or wafer-to-wafer variations separately for better understanding. In the CPCA, PC's can vary only in a guided dimension which is consistent with the die-to-die or wafer-to-wafer variation.

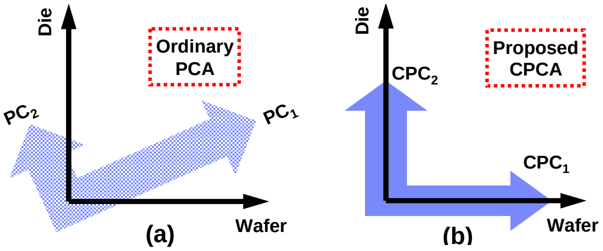


Figure 2. Comparison of the concept of the ordinary PCA (a) with the proposed CPCA (b): The ordinary PCA decomposes a vector into orthogonal spaces without considering die-to-die or wafer-to-wafer variations. The proposed CPCA finds orthogonal spaces that also coincide with either die-to-die or wafer-to-wafer variation.

Figure 2 visualizes the difference between the proposed CPCA in the right panel and the traditional PCA in the left panel. Conceptually, the PCA finds orthogonal coordinates which do not generally coincide with die and wafer variations. Therefore, understanding process variation using the ordinary PC's would be perceptually difficult. On the other hand, the CPCA guides the PC's to the die and wafer directions, leading to direct visualization of the variation in the die-to-die and wafer-to-wafer ranges. Only a few CPC's may be examined for this purpose because only a fraction of all the CPC's are sufficient to capture the most of the information as with the case of PC's.

2.3 Algorithm

Figure 3 illustrates how the CPC's can be iteratively obtained. Because the variability of the data is scale-dependent, the PCA is sensitive to the scaling of the data to which it is applied. Thus, at the preprocessing stage, the data set of each inline parameter is standardized to be zero-mean and unit-variance. The rationale for this standardization is to treat each in-line test parameter insensitive to arbitrary scaling (*e.g.*, different units) and bias (*e.g.*, systematic offset).

Subsequently, the data is screened for anomalies and insignificant values. In our implementation, all the in-line parameters which contain meaningless data points (*e.g.*, system default values for failed measurement) are filtered out.

Also, a simple Gaussianity test such as kurtosis analysis for the ensemble of each parameter can be applied to ensure the validity of the data set [6]. Kurtosis (the ratio of the fourth central moment to the square of the variance) is a measure of peakedness of a distribution. In our case, the parameters with the kurtosis, greater than 8 (having much fatter tail than the normal distribution which has kurtosis of 3) or less than -2 (having much thinner tail than the normal distribution) are flagged unusable. In the next step, PCA is performed to find the first PC for die and wafer variation. The PC of larger variance is selected. The data set is, then, transformed to be orthogonal to the space spanned by the selected PC. This routine is iterated for the residual data set until a given criterion is satisfied. This is a valid constrained PCA because the zero-mean die variation and wafer variation are orthogonal to each other.

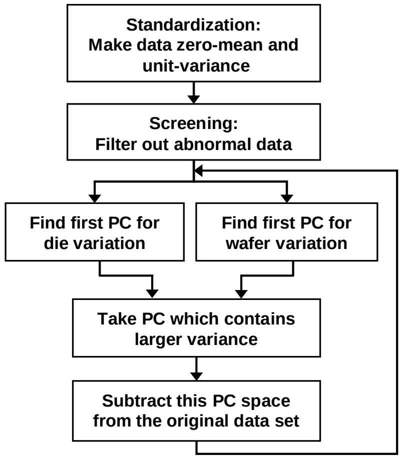

Standardization:
Make data zero-mean and unit-variance
Screening:
Filter out abnormal data
Find first PC for die variation
Find first PC for wafer variation
Take PC which contains larger variance
Subtract this PC space from the original data set

Figure 3. The flow of the proposed CPCA algorithm

The resulting PC is constrained to represent either wafer variation or die variation. A PC representing the die-to-die variation may serve as a snapshot of a given technology. Based on this systematic die-to-die pattern, the characteristic of a given lot(s), or generally a given technology can be monitored, thus allowing fast and critical feedback to manufacturing and technology. A PC for the wafer variation is also indicative of the technology and manufacturing used.

CPCA can be regarded as a source coding of the complex yet redundant data set: only a few PC's are sufficient to capture the most of the information contained in the whole data. Any lot or lots can be represented by the weighted sum of a few CPC's within some accuracy, thus effectively compressing all technology parameters into a handful of weighting factors. Hence, the CPCA can serve as a simple and effective way to keep track of the metrology of a given lot(s) or technology generations.

3. A case study: in-line parameters in 65nm SOI technology

3.1 The data set

For this study, 1109 in-line parameters in a pre-production 65nm SOI CMOS technology are used. A data set of each in-line parameter contains 520 samples (40 dies per wafer for 13 wafers) for each in-line parameter. Wafers used are 300mm, and belong to a same lot. There are various measurements from FET test structures (*e.g.*, Vt, Ion, Ioff), ring oscillators, SRAM's and capacitance as listed in Table 1.

Type	FET	Ring Oscillator	SRAM	Capacitance	**Total**
Number of parameters	759	83	159	108	**1109**

Table 1. The categories of in-line parameters used in the CPCA analysis

3.2 The CPCA results

Both the ordinary PCA and CPCA were performed on a given data set for the sake of comparison. The computation time was not more than one minute to obtain all the CPC's for this 1109-by-520 in-line data matrix using an ordinary PC. Figure 4 shows the variance which can be explained by the first 20 PC's and CPC's for the ordinary PCA and CPCA, respectively. A variance for each PC is shown as well as the cumulative variance. The first PC's in both the methods account for 31~34% of the total variance of the original data set. Using the first two CPC's, 57% can be explained, slightly less than 61% for the unconstrained PC's.

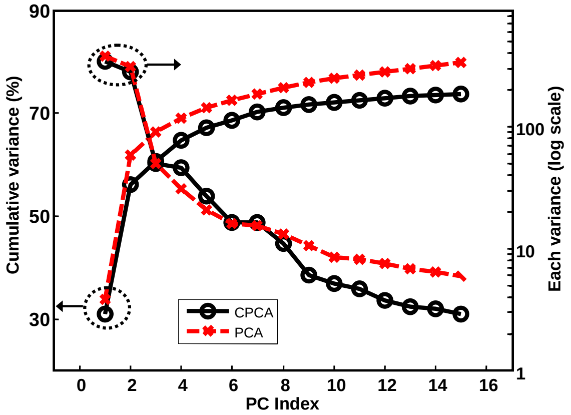


Figure 4. Cumulative variance of PCA and CPCA

It is also noted that the CPC's do not reach 100% asymptotically because the die or wafer variation alone cannot fully represent some intertwined relationship between the two. Nonetheless, the advantage of separating the die and wafer components of systematic variation justifies the slightly less coverage of variance by the same number of CPC's compared to the ordinary PC's.

	Die-to-die	Wafer-to-wafer	Variance explained	Cumulative Variance explained
1st CPC	√		31.0%	31.0%
2nd CPC		√	25.2%	56.2%
3rd CPC	√		4.5%	60.7%
4th CPC		√	4.2%	64.9%
5th CPC		√	2.4%	67.3%
6th CPC		√	1.5%	68.8%

Table 2. The type and variance of the first 6 CPC's

Table 2 lists the type (either die-to-die or wafer-to-wafer) and variance of the first six CPC's. This table also shows that the first and second CPC's capture the die and wafer variation, respectively. The die variation and wafer variation often alternate along the progression of CPCA iteration as expected: after one type of variation is subtracted, the other type is likely to be predominant in the residual data at the next CPCA iteration.

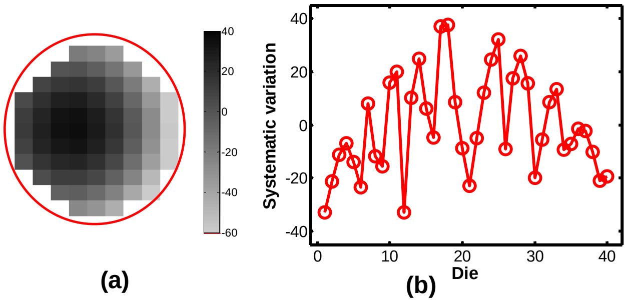


Figure 5. The first CPC (die variation) shown as a wafer map (a) and die-to-die plot (b)

Figure 5(a) shows the first CPC (die variation) image, fitted by the 2nd order polynomials on the 40 available values of the first CPC. The polynomial fitting was done to interpolate the missing values in some chip sites for the purpose of visualization. The slightly off-centered radial pattern is clearly visible in this PC. Figure 5(b) plots the first CPC with respect to a die index. This is the most prominent systematic variation by far, explaining about 31% of the variance of the whole data set.

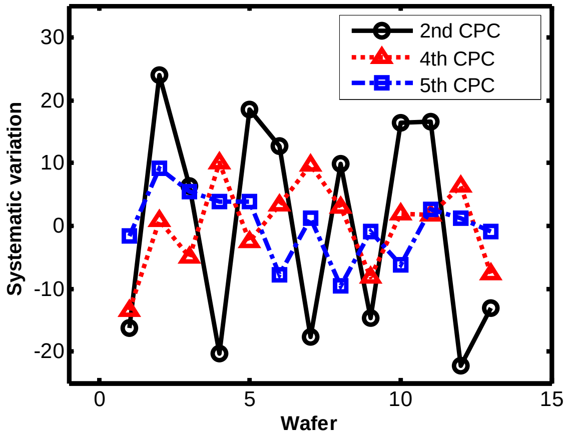


Figure 6. Second, fourth and fifth CPC which correspond to the first three wafer variations

Figure 6 exhibits the wafer variation captured by the second, fourth and fifth CPC corresponding to the first three wafer variations. The second CPC alone represents 25% of the total variance of the whole data set. It is observed that the dominant die variation (31%) is larger than the dominant wafer variation (25%), which is consistent with the recent trend that a die variation is increasingly important due to the larger wafer size (300mm) than before.

In Figure 6, the order of wafer indices is arbitrary, but the systematic pattern and the spread of wafer variations can be fed back to the technology development in order to further analyze the source of the this variation.

4. Applications

4.1 Process variation analysis on a new data set

CPC decomposition can be applied to a new data set of a totally different nature. For this study, we used a bench-tested RF self-oscillation frequency (Fso) for a static CML frequency divider. Fso was measured from the same dies and wafers on which the previous in-line parameters used for the CPCA reside. Figure 7 illustrates the sequence of CPCA in 3 dimensions to visualize how Fso can be reconstructed by adding one component at a time using an offset and the first four CPC's.

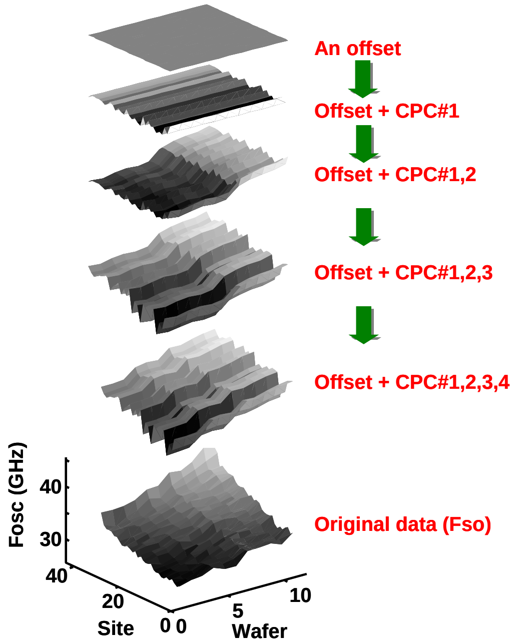


Figure 7. Self-oscillation frequency of a divider (bottom) represented by a global offset and the first four CPC's

The bottom surface shows Fso (z-axis) from different dies and wafers. A global offset shown on the top is an average of Fso over all dies and wafers, thus a constant. The second image is the first CPC plus the offset, having only die variation. The next image displays the added contribution of the second CPC (wafer variation) on top of the previous image. This figure demonstrates how original data can be successively reconstructed from or, equivalently, decomposed into a few CPC's. Note that these CPC's are calculated from the previous in-line DC test data and not from this Fso which is being analyzed. A weight for each CPC is obtained by projecting Fso data on to each CPC space. The first four PC's retain 66% of all the information of Fso variation, which is a significant amount especially because the test data (frequency of RF circuit) and the training data for CPC calculation (in-line DC measurement data) are quite different in nature. The physical mechanism of how each in-line device-level parameter affects complex RF circuitry such as the frequency divider is elusive and challenging to analyze. However, the proposed algorithm and experimental data show that the process variation is substantially systematic, and therefore, the CPC's obtained from in-line measurement can explain a significant portion of the process variation in complex RF circuits.

4.2 Efficient sampling for measurement and yield analysis

The most dominant die variation, the first CPC in our case, contains the most information (31%) about systematic within-wafer variations. Therefore, an intelligent sampling scheme can be proposed for cost-effective measurement and quick yield analysis, based on the first CPC; for example, if only two chips per wafer are allowed for measurement, it would be reasonable to sample the minimum and maximum points in first die-to-die CPC. One can also selectively measure some sensitive sites to effectively evaluate how much a wafer is compatible to the die variation pattern(s) without sacrificing a great deal of accuracy.

5. Conclusion

A statistical framework is proposed to separate die-to-die and wafer-to-wafer variations. Major advantages of the proposed method are as follows:

- This method allows effective visualization and analysis of systematic die-to-die and wafer-to-wafer variations using only an ensemble of manufacturing in-line data.
- This analysis can be implemented in near real-time to give rapid feedback to technology development. It can effectively complement the analytic or numerical modeling of process variation.

Our future work includes:

- Extending this framework to accommodate a within-die or lot-to-lot variation as another constraint in the CPCA algorithm.
- Relaxation of the Gaussian assumption for the in-line parameters and the usage of independent components in place of principal components which are uncorrelated but not necessarily independent.

6. References

[1] S. R. Nassif. "Modeling and Analysis of Manufacturing Variations," *Proc. IEEE Conf. on Custom Integrated Circuits*, pp. 223-228, May 2001.

[2] I. T. Jolliffe. *Principal Component Analysis*, Springer-Verlag New York, Inc., New York, 2002.

[3] G.A. Cherry and S.J. Qin. "Multiblock principal component analysis based on a combined index for semiconductor fault detection and diagnosis," *IEEE Trans. Semiconductor Manufacturing*, 19, May 2006, pp. 159-172.

[4] L. Yan. "A PCA-Based PCM Data Analyzing Method for Diagnosing Process Failures," *IEEE Trans. Semiconductor Manufacturing*, 19, Nov. 2006, pp. 404-410.

[5] Y. Takane, H. A. L. Kiers, and J. de Leeuw. "Component analysis with different sets of constraints on different dimensions," *Psychometrika*, 60:259-280, 1995.

[6] R. Pond. *Fundamentals of Statistical Quality Control*, Macmillan College Publishing, New York, 2004.