IEEE Asian Solid-State Circuits Conference
November 3-5, 2008 / Fukuoka, Japan

Cell Broadband Engine Performance and Yield Benchmark in 65nm SOI CMOS with Spatial, Temporal and Parametric Process Variability Model

Choongyeun Cho, Daeik D. Kim
IBM Semiconductor R&D Center
Hopewell Junction, NY
{cycho,dkim}@us.ibm.com

Jonghae Kim
Qualcomm Inc.
San Diego, CA
jonghaek@qualcomm.com

***Abstract*—This paper introduces a process variability model to determine the performance and yield of the Cell Broadband Engine (CBE) in 65nm SOI CMOS. The model incorporates spatial (die-to-die), temporal (manufacturing process drift), and parametric dimensions, and provides microprocessor performance tracking and comprehensive view on the process variability with embedded ring oscillator measurement at the wafer level. It extracts CBE performance regularity within die for the circuit design and models, and reveals the semiconductor manufacturing signatures in wafers and lots for process technology. The model reduces performance estimation testing requirements by surpassing conventional methods' accuracy by 28%.**

I. Introduction

Process variability has become a major challenge in designing high-performance microprocessor, especially in 65nm technology and beyond [1]. For a multi-core processor, within-die process variation results in individual cores in the chip to differ significantly in maximum supported frequency and the power consumption. Die-to-die variation is tied to the parametric and functional yield of a processor. In addition to these spatial variation sources, longer time-scale variation source induced by manufacturing process drift is measured and characterized in this work. Both short-term and longer-term variation signatures need to be considered for microprocessor design, model, simulation, and technology development.

There are two distinct perspectives when dealing with process variation (deviation of process from the nominal conditions) in microprocessor production: one from design point of view and the other from technology point of view. From design perspective, a host of circuit and system architecture techniques have been studied and exercised to ameliorate an effect of environmental variation in process, supply voltage, and temperature (PVT). These techniques include: dynamic voltage / frequency scaling and adaptive body biasing to make system more robust to threshold-voltage process variation; adaptive supply voltage to tackle supply voltage variation; and temperature-based voltage / frequency throttling to control the die temperature.

From technology or manufacturing perspective, all the key process variation sources need to be carefully modeled and measured. Only when a variation is reliably captured in a model, we can mitigate the variation in manufacturing, and also understand the variation mechanisms to further counteract using design techniques. It is, however, increasingly difficult to model or predict process variability reliably, partly due to the complicated nature of physical mechanisms for process variation. Also, test structures to monitor and extract process variation need to be carefully planned and designed to realistically represent process variation in a product circuit, using reasonable test time and cost.

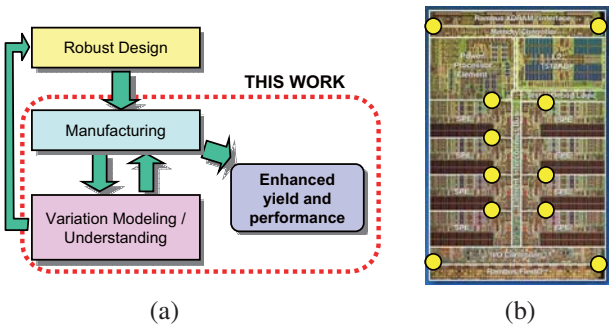
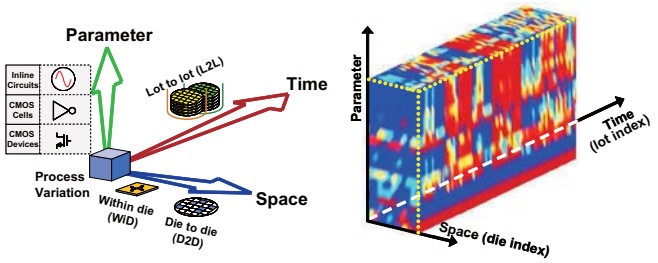


Fig. 1. (a) Illustration of process-variation-aware design flow and this work's contribution. (b) Die photo for Cell Broadband Engine processor. 11 within-die ring oscillators are scattered across a chip, marked in circles, as chip-variation and performance monitors.

To address the latter perspective, this paper proposes a general representation method for process-induced variation, exploiting spatial, temporal, and parametric information buried in measurement data. Fig. 1(a) illustrates a process-variation-aware design flow and the scope of this work. Specifically we examine the performance variation of the Cell Broadband Engine (CBE) in 65nm SOI CMOS – a nine-core microprocessor, jointly developed by IBM, Sony and Toshiba. Fig. 1(b) is a die photo of CBE and the locations of uniform embedded ring oscillators (ROs) to keep track of time trend and spatial variation of CBE.

There are three sources of correlations in any set of measurable parameters: (1) Spatial correlation: within-die (WiD) correlation in a given die, and die-to-die (D2D) correlation in a given wafer; (2) Temporal correlation: correlation between different wafers or lots, manufactured at different time frames,

978-1-4244-2605-8/08/$25.00 ©2008 IEEE



(a) (b)

Fig. 2. (a) Definition of three variation dimensions: spatial (D2D and WiD), temporal (manufacturing process drift or L2L variation), and parametric dimensions (different devices / test structures being measured). (b) CBE's embedded RO measurement data in 65nm SOI CMOS is represented as a 3D cube. (Each parameter is normalized to zero-mean and unit-variance.) For this data, 11 ROs are implanted across CBE chip to monitor WiD variation of the chip performance because ROs are known to have a high correlation with the microprocessor performance.

caused by manufacturing processes' drift and instability; (3) Parametric correlation: correlation between different physical parameters being measured, e.g. correlation between threshold voltage (V_{th}) and RO delay (τ_{RO}) in the same chip.

Conventionally, "D2D variation" is often used as an umbrella term which refers to "all" D2D variation components that come from within-wafer (WiW), wafer-to-wafer (W2W) and lot-to-lot (L2L). We differentiate these components in this paper because each has its own systematic pattern that can potentially be modeled and, thus, predicted. D2D in this work will specifically refer to within-wafer D2D variation.

Fig. 2(a) illustrates three variation dimensions. Three variation dimensions are treated as orthogonal coordinates. Using these coordinates, Fig. 2(b) shows a 3D representation of a set of RO measurements embedded in a CBE chip. From the sliced sections at the boundaries in Fig. 2(b), there exist correlations to different degrees in any pair from the three dimensions.

An RO is used extensively in order to monitor the processor performance and the core-to-core performance variability. Its primary advantage is the ability to estimate the processor performance and the parametric yield early in the manufacturing – well before packaging – with a simple and inexpensive test. It is known to well represent the actual maximum frequency of a processor [2].

In Section II, a current process-variation model is briefly explained, and a new model is proposed. Experimental results based on hardware data sets from CBE in 65nm SOI CMOS technology will be presented in Section III. An application is discussed in Section IV, followed by conclusion.

II. Variation Model

We denote x as a measurable parameter of interest. It can be a physical parameter such as channel length, oxide thickness, or a parametric quantity from device (e.g. V_{th}, I_{on}, I_{off}) or from a circuit (τ_{RO}, I_A, I_Q). Conventionally process variation is decomposed into a D2D and WiD components [1]:

$$x = x_0 + \Delta x_{D2D} + \Delta x_{WiD} + \Delta x_{res} \quad (1)$$

TABLE I
Number of samples for each variation dimension.

Dimension	*Data set #1*	*Data set #2*
Total lots	10	1
Total wafers	75	12
Dies per wafer	12	20
Within-die locations per die	11	1
Parameters	18	2,856
Total samples	$\sim 1.8 \times 10^5$	$\sim 6.8 \times 10^5$

where Δx_{D2D} is inter-chip global variation, Δx_{WiD} is intra-chip spatial variation, and Δx_{res} is the residual random component which is not captured by Δx_{D2D} and Δx_{WiD}. The time-dependent or L2L variation is not factored in to a conventional model because a manufacturing process drift is difficult to model.

We propose a variation model which captures time- and parameter-dependency in addition to spatial correlation:

$$\mathbf{x}(t) = \mathbf{A}(t)\mathbf{p}_0(t) + \Delta\mathbf{x}_{spatial}(t) + \Delta\mathbf{x}_{res}. \quad (2)$$

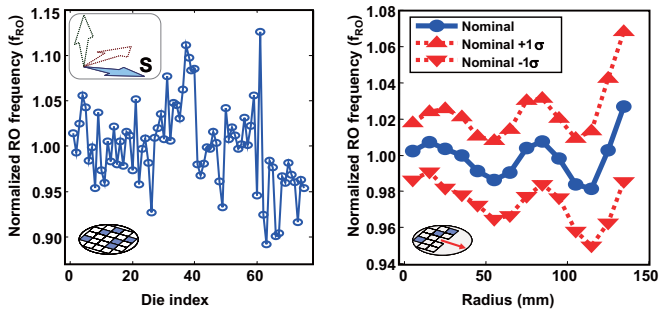
Here m-by-1 $\mathbf{x}(t)$ is a collection (vector) of all m parameters at time t. $\mathbf{A}(t)$ is a time-dependent mixing matrix, and $\mathbf{p}_0(t)$ is the k key underlying parameters where typically $k < m$. $\Delta\mathbf{x}_{spatial}(t)$ denotes *any* spatial variation which can be interpreted as sum of WiD and D2D variation. D2D and WiD variations are lumped together in our work because they are both spatial variations in nature. Here discrete time, t, refers to the index of a wafer or a lot that was fabricated at a certain time, or it can simply represent week or month index.

The intention of this paper is not to present an accurate model for each variation dimension, but to propose a general modeling approach and its benefits which accrue in addition to the conventional approach. Hence, each variation source (in space, time and parameter) is modeled briefly in the following section.

III. Process Variation Components

For our experiments, two sets of measurement data in 65nm SOI CMOS technology are used. *Data set #1* is measured from CBE (see Fig. 1(b)) using a manufacturing-inline tester. 11 uniform performance-screen ROs are embedded on a microprocessor chip to keep track of time trend and spatial variation of the host chip. ROs are known to be a robust monitor for microprocessor performance [2]. This data set contains relatively rich spatial information in terms of D2D and WiD, and temporal information (over 4 months) regarding the final product (microprocessor) performance.

In addition, for *data set #2*, one lot was thoroughly tested off from the manufacturing floor, using an automated parametric tester. This data set includes 2,856 parameters from various test structures including FETs, SRAM, and ROs. This set of measurements are primarily intended for technology development and device model-to-hardware closure. Table I arranges the number of samples in each variation dimension for both data sets.



(a) (b)

Fig. 3. Illustration of spatial variation from *Data set #1* in Table I. (a) CBE's RO frequency (f_{RO}) variation in a wafer. (b) A strong radius dependency is exhibited. It is important to characterize a systematic D2D pattern before dicing and packaging.

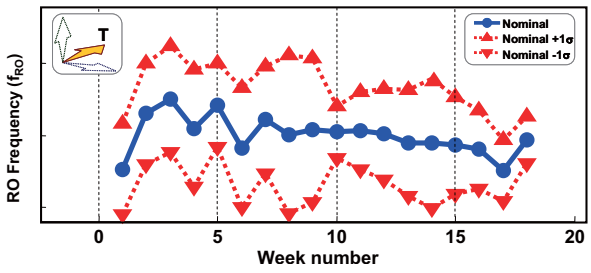
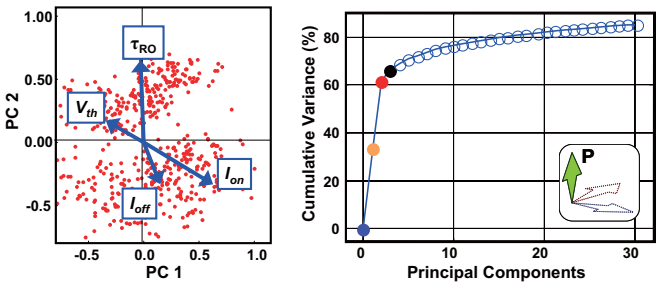


Fig. 4. Illustration of temporal variation: CBE's performance-monitor (f_{RO}) variation (normalized) as a function of time.



(a) (b)

Fig. 5. Illustration of parametric variation: (a) First two principal components and the directions of four physical parameters. (b) Cumulative variance explained by principal components. The data set used for this analysis is *data set #2* from Table I.

A. Spatial Performance

The spatial-variation aspect has received a significant attention recently in wide range of design applications from leakage power analysis, yield model, to SSTA [3]. In general, spatial variation is decomposed into WiD and D2D components. WiD variation is often induced by layout and topography interaction with the processes, such as chemical-mechanical polishing (CMP) and critical dimension variation in channel length or metal wire lines. WiD spatial correlation functions are extracted by posing and solving a constrained linear or nonlinear optimization problem [4]. Wafer-scale D2D variation is generally caused by equipment non-uniformity and other physical effects such as thermal gradients and loading effects, often exhibiting radial pattern or a slanted plane. Typically, D2D variation within a wafer contains low spatial-frequency component, and neighboring dies are likely to be highly correlated with each other. Fig. 3 illustrates CBE's RO frequency (f_{RO}) variation within a wafer (a) and the radius dependency of f_{RO} (b). The observed "W"-envelope pattern renders a donut-shaped wafer map. It is noted that die-index order can be arbitrary thus may fail to identify important D2D variation signature. Therefore, characterization of D2D variation before the dicing and packaging stage is important to capture a systematic D2D pattern.

B. Temporal Performance Tracking

A temporal variation originates from an accumulating drift in process equipment operation, and temporal correlation embodies a certain degree of redundancy among same parameters in different wafers or lots that were processed sequentially in time. From a manufacturing point of view, the temporal variation is monitored and controlled via statistical metrology and feedback, but it is not feasible to perfectly stabilize all the manufacturing tools. Temporal variation is not easily captured by simulation or model because of complicated nature of manufacturing equipment drift. There are generally two categories of temporal variation in terms of the correlation period. Short-term temporal variation originates primarily from non-uniformity of process equipment and processing environment. We will focus on the longer scale for temporal variation – the significant L2L variations due to process recipe or grade change, as seen in Fig. 2(b) as radical color changes from one epoch to an adjacent one. Fig. 4 exhibits CBE f_{RO} mean and standard variation as a function of time, over the period of four months. It is seen that as manufacturing processes mature after a number of yield-learning cycles, the nominal f_{RO} stabilizes, and the standard deviation generally decreases.

C. Performance by Parameters

Parametric correlations have been characterized by principal component analysis (PCA) and its variants without consideration of spatial or temporal correlations in the literature [5]. Different physical measurements carry some degree of correlation. For example, FET characteristics V_{th}, I_{off}, I_{on}, and ring gate delay τ_{RO} are correlated with each other. Fig. 5(a) is a scatter plot of the first principal component (PC) and the second PC weights for all 240 available chips from *data set #2*, and directions of four physical parameters in terms of the first and second PC's. Among 2,856 parameters, four important device/circuit characteristics (τ_{RO}, I_{off}, I_{on}, and V_{th}) are represented by vectors, and the direction and length of each vector indicates how each variable contributes to the two principal components (PC's) in the plot. For example, the first PC, represented in this plot as the horizontal axis, has positive coefficient for active current but very small coefficient for τ_{RO}. It is also seen that τ_{RO} and I_{off} are linear but in opposite directions, and so are V_{th} and I_{on}. Fig. 5(b) shows

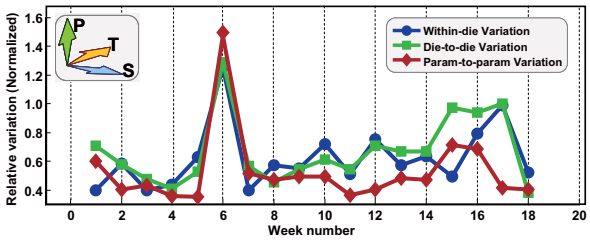


Fig. 6. Relative variations in all spatial, temporal, and parametric variation dimensions, calculated from *data set #1*.

cumulative variance explained by first 30 PC's. It is noted that 30 components explain approximately 85% of the total variance of the given data set of 2,856 variables.

The correlation or redundancy in different parameters, if properly captured in a model, can be exploited in a number of applications. For example, a product performance can be mapped to (or inferred from) a number of key device characteristics.

D. Comprehensive Performance Variation

Fig. 6 shows relative variation for WiD, D2D, and parameter-to-parameter (P2P) dimensions, as a function of time. Measurements from CBE (*Data set #1*) are used for this analysis. Each parameter is normalized to be zero-mean and unit-variance for all dies and wafers/lots. As a result, for any week, contributions from WiD, DID and P2P components do not sum to 100%. There is a strong time dependency for spatial, temporal, and parametric variation dimensions, and in this data set there appears relatively less inter-parametric variation than WiD and D2D variation. WiD and D2D variations are comparable in their contributions for the most weeks. It is worthwhile to note that one variation source does not dominate all the time: e.g. at week 1, D2D variation has the most contributions (70.7%), but at week 18, WiD variation has the most contributions (52.4%). If a temporal variation is not considered in a model, its contribution will be added to a random or residual variation component which is not explained by either WiD or D2D variation. Therefore, the temporal component in a process variation model needs to be monitored, characterized, and modeled for a robust microprocessor design and technology development.

IV. Chip Performance Reconstruction

The proposed process-variation model captures variation sources in space, time and parameter, thus opens up opportunities for various practical applications. We discuss the missing measurement estimation as an application of the proposed 3D process variation representation. Specifically the values of CBE's performance monitors are reconstructed using an actual hardware data.

Volume electrical test using a semiconductor parametric tester is prone to inaccuracies and data corruption induced by a number of non-ideal test conditions – for example, miscalibration and imperfect probe contact to pads. The proposed 3D representation can be utilized to accurately recover the corrupted measurements by exploiting the 3D redundancies carried by the existing measurements. In this experiment, we suppose some data points are unavailable in the CBE performance-tracking RO data, and they are estimated by 2D models (considering only two correlation dimensions) and the proposed 3D model. The first experiment is a case where one chip from twelve chips is presumed corrupted, and is estimated using 2D models and the proposed 3D model. The other case is that six chips (or half of all available chips) are corrupted for one wafer. The test data is recovered in the same fashion. To be fair, estimation was iterated for all the data points for one-chip estimation, or 10,000 randomized trials for six-chip estimation. The average RMS errors are presented in Table II. Each parameter is normalized to be unit-variance. Hence, RMS error of 18.6% for 3D model refers to the error of 18.6% of the standard deviation of original value being estimated. The 3D model is up to 43.0% more accurate than 2D models for one-chip estimation, and 28.5% for six-chip estimation.

TABLE II
RMS estimation error for CBE's performance-benchmarking RO frequency.

Model	One chip corrupted	6 chips corrupted (Half of all available chips)
Die-wafer 2D	30.3%	32.0%
Die-parameter 2D	32.5%	40.0%
Wafer-parameter 2D	25.7%	30.5%
Proposed 3D	18.6%	28.6%
% improvement	28% – 43%	6.2% – 28.5%

This experimental results show that the proposed 3D model outperforms 2D models in chip performance reconstruction, because it takes account of three variation dimensions simultaneously. A product performance can be, hence, reliably predicted with the 3D model at the wafer level.

V. Conclusion

Performance and variation of CBE in 65nm SOI CMOS are analyzed with the proposed comprehensive variation model incorporating spatial, temporal, and parametric dimensions. In addition to within-die and die-to-die variation, long-term lot-to-lot variation is a significant variation component, and needs to be considered for microprocessor design, model, simulation, and technology development.

References

[1] S. R.Sarangi *et al.*, "VARIUS: A model of process variation and resulting timing errors for microarchitects," *IEEE Trans. Semicond. Manufact.*, vol. 21, no. 1, pp. 3–13, Feb 2008.

[2] R.Berridge *et al.*, "IBM POWER6 microprocessor physical design and design methodology," *IBM J. Res. Dev*, vol. 51, no. 6, pp. 685–714, Nov 2007.

[3] H.Chang and S.Sapantnekar, "Full-chip analysis of leakage power under process variations, including spatial correlations," in *Proc. ACM/IEEE DAC*, 2005, pp. 523–528.

[4] J.Xiong, V.Zolotov, and L.He, "Robust extraction of spatial correlation," *IEEE Trans. Computer-Aided Design*, vol. 26, no. 4, pp. 619–631, Apr 2007.

[5] A.Singhee and R. A.Rutenbar, "Beyond low-order statistical response surfaces: Latent variable regression for efficient, highly nonlinear fitting," in *Proc. ACM/IEEE DAC*, 2007, pp. 256–261.