Statistical Framework for Technology-Model-Product Co-Design and Convergence

Choongyeun Cho[1], Daeik Kim[1], Jonghae Kim[1], Jean-Olivier Plouchart[2], and Robert Trzcinski[2]
[1]IBM Semiconductor Research and Development Center, Hopewell Junction, NY, USA
[2]IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
{cycho,dkim,jonghae,plouchar,rtrzcin}@us.ibm.com

ABSTRACT

This paper presents a statistical framework to cooperatively design and develop technology, product circuit, benchmarking and model early in the development stage. The statistical data-driven approach identifies device characteristics that are most correlated with a product performance, and estimates performance yield. A statistical method that isolates systematic process variations on die-to-die and wafer-to-wafer levels is also presented. The proposed framework enables translations of interactions among technology, product, and model, and facilitates collaborative efforts accordingly.

The proposed methodology has been applied to first three development generations of 65nm technology node and microprocessor product current-controlled oscillators (ICOs) for phase-locked loops (PLLs) that were migrated from 90nm. Automated manufacturing floor in-line characterization and bench RF measurements are used for the methodology. The ICO exhibits yield improvement of RF oscillation frequency from 47% to 99% across three different 65nm SOI technology generations.

Categories and Subject Descriptors
B.8.2 [**Performance and Reliability**]: Performance Analysis and Design Aids

General Terms
Measurement, Performance, Design, Economics, Verification.

Keywords
Technology-Model-Product Co-design, Statistical, Yield, Process Variation, Design for Yield (DFY)

1. INTRODUCTION

Currently product, model, and process are developed in a parallel fashion with not much consideration of their interplay. While this divide-and-conquer approach drives development of each component separately, it has limitation in addressing a complex interaction among components. The lack of systematic perspective on the whole IC development components hinders rapid yield learning of the target process technology node [1]. It is distinguished especially in the technology node ramp. Figure 1 exemplifies a typical technology node ramp-up timeline for product, model, benchmark, and process development for a new technology node. Here, a node refers to a technology associated with a MOSFET gate feature size (e.g. a 65nm node). A generation within a node uses a different set of masks (e.g. 65nm generations 1 and 2). An iteration employs same mask set but potentially different process recipes, at a different time. The initial benchmarking structures (B0) and product-driven circuits (P0) are commonly migrated from an earlier generation or an earlier node. On a process side, preparation for a new technology node usually begins in parallel with an initial target model (T0) development. B0 benchmarking and P0 product will be designed with T0 model. As front-end-of-the-line (FEOL) processes are done and most B0 devices are characterized, the output will be interpreted for model calibration that results in T1 model. When back-end-of-the-line (BEOL) processes end, the measurements on product circuits (P0) will provide feedback to process and model. Due to the high degree of complexity of product circuits and stringent timeline, a next generation or iteration begins before the feedback from P0 measurement is reflected in process and model. As in this illustration, the next generation (Gen 2) is developed concurrently with Gen 1.

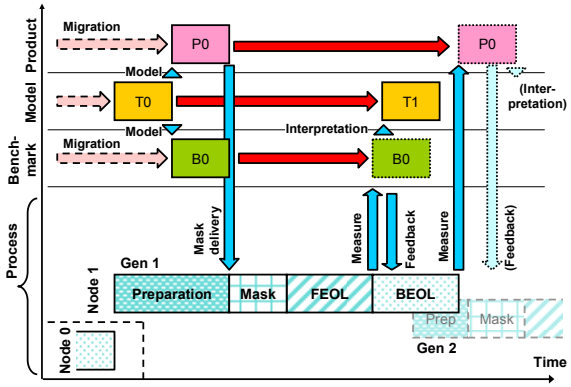


Figure 1. A new technology node development timeline for manufacturing process, benchmarking, model and product circuit design.

Fig. 2 illustrates feedback mechanism of process, benchmark, model and product.

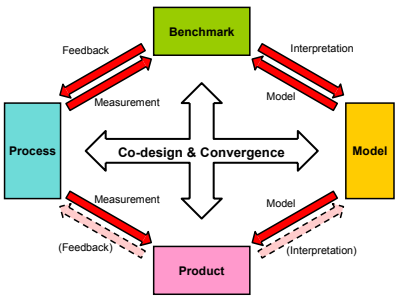


Figure 2. Interaction between process, benchmarking, model, and product.

The primary feedback to process and model comes from the test data which are experimented by benchmarking structures such as device macros, physical macros, and static ring oscillators (ROs). These test structures are designed to be sensitive to a certain set of process variations so that their impacts on circuit performance are directly monitored, and accommodated in the process development and model calibration. On the other hand, a complex and customized product circuit is difficult to provide such a feedback to either process and model because of its complicated

operation and interactions with process and model. In this paper, a methodology is pursued to link the performance of a complex product circuit (especially RF and high-speed circuits) to benchmarking which in turn is closely related to process and model. The goal of the proposed approach is to facilitate and expedite the co-design and convergence of all the development components, thereby reducing time-to-market and early development cost of new technology node before high-volume manufacturing.

The remainder of the paper is organized as follows. In section 2, we will review process-induced variability and its impact on circuit and technology. We will introduce a statistical framework for co-design and convergence of technology, model and product, in section 3. Applications and experiments of this methodology will be presented in section 4, using first three generations of 65nm process technology node and three-stage current-mode logic (CML) current-controlled oscillators (ICOs) for server product PLLs. Conclusions will follow in section 5.

2. PROCESS VARIATION

It is widely known that the dominating limiter of integrated circuit's performance yield is process variation [2,3]. A process parameter (p) has three contributions: a nominal value, a systematic variation that can be modeled or predicted, and a random variation that is left over as a residual. In general, systematic variation can be broken into four components in terms of its scope [4]: within-die, die-to-die, wafer-to-wafer, and lot-to-lot variations as shown in Fig. 3. Within-die means a process variation in the identical device or circuit within a die, which is defined by a new set of masks. Die-to-die represents a process variation in different dies within a wafer. Wafer-to-wafer is a variation in different wafers within a lot. The lot-to-lot denotes a variation in different lots. The suggested process variation decomposition is useful especially when the die-to-die variation is comparable to the wafer-to-wafer variation. This phenomenon becomes more noticeable in 300mm wafer processes and sub-100nm nodes [2].

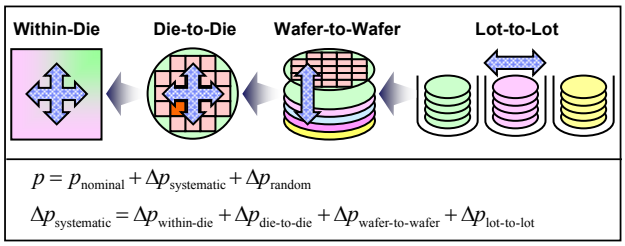


Figure 3. Definition of four ranges of process-induced variability. The systematic variation includes within-die, die-to-die, wafer-to-wafer, and lot-to-lot components.

In this work, only the die-to-die and wafer-to-wafer variations are considered although the presented methodology can be readily extended to within-die and lot-to-lot variations, which remains as future work. There are a number of key factors that impede fast yield learning as a technology node scales down. On a technology side, the process variation effectively increases relative to nominals (e.g. MOSFET threshold voltage variation increases). A circuit also must be tolerant to variation, and adhere to tighter design specifications. For example, a marginal supply voltage headroom emerges as a major analog design constraint. On a model side, the complex interaction between product circuits and manufacturing process cannot be fully captured by electrical models *a priori*, and is usually captured *a posteriori* by hardware data. A new technology node necessitates higher-degree statistical model calibration based on measurement data, especially in high-frequency and short-channel regime.

3. PROPOSED METHODOLOGY

There are two modules in the proposed methodology that coherently assist the co-design and rapid ramp-up of technology, model, and product development. One module is called the Parametric Statistical Analysis (PSA) that deals with interaction of parameters between process, product, and benchmarking. Cross-correlation analysis (elaborated in subsection 3.1) and yield estimation (subsection 3.2) constitute the PSA as illustrated in Fig. 4 (a). To the process side, the PSA feeds back relationship between benchmarking test structures and the product by cross-correlation analysis. To the model side, it provides model-to-hardware correlation (MHC) for both benchmarking structures and the product. It is noted that simple benchmarking devices and circuits (e.g. FETs, ROs, and SRAM) may not fully reflect the complex nature of a real product (e.g. voltage/current-controlled oscillator, PLL, and microprocessor). Thus, it is critical to link the benchmarking to a product via cross-correlation analysis and statistical yield estimation.

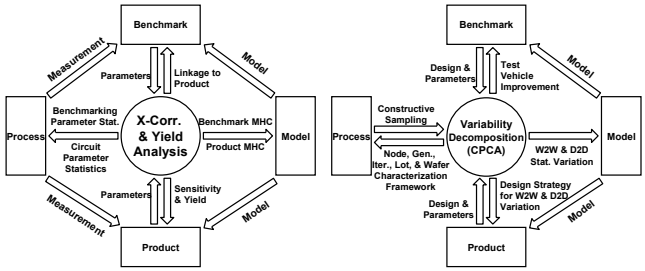


Figure 4. (a) Diagram of parametric statistical analysis (PSA), with enabling development component interactions (b) Diagram of variability decomposition using constrained principal component analysis (CPCA) and its enabling development activities.

The other module is variability decomposition using constrained principal component analysis (CPCA, subsection 3.3) that helps analyze and mitigate systematic wafer-to-wafer and die-to-die variability, as shown in Fig. 4 (b). CPCA separates a systematic variability to die-to-die and wafer-to-wafer components using only measurement data, leading to an effective characterization of a wafer, lot, technology iteration, generation or nodes. It also helps next model to accommodate the observed wafer-to-wafer and/or die-to-die variations. A constructive and adaptive sampling scheme for the purpose of measurement can be drawn based on the resulting die-to-die and wafer-to-wafer variation.

The proposed methodology can be iteratively applied in the early functional tests, thus, allowing quick learning to feed back into the next technology release. The proposed methodology can complement conventional approaches including first-order analytical prediction and simulation.

3.1 Cross-correlation Analysis

For fault detection and device characterization, in-line electrical measurements for benchmarking test structures (referred to as "in-line parameters" hereinafter) are typically performed off the manufacturing floor using an automatic parametric tester [5]. For example, turn-on current, off current, and capacitance, threshold voltage (Vth) of FET devices can be regularly measured, monitored and archived for all the wafers manufactured. A performance of circuit under consideration is more or less correlated with the in-line parameters. In our study, we use Pearson's sample correlation that measures linear relationship of two variables (N-sampled x and y) as expressed by

$$\rho_{x,y} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}}, \text{ where } \bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i,\ \bar{y}=\frac{1}{N}\sum_{i=1}^{N}y_i. \quad (1)$$

This correlation metric is based on linear relationship of two variables. The in-line electrical parameters and a circuit performance exhibit mostly linear or mildly non-linear relationship in our previous experiments. A different correlation metric may capture strong non-linearity (e.g. quadratic or exponential) when needed.

A true correlation (ρ_{CI}) within a confidence interval is calculated based on the sample correlation (ρ), sample size (N), and confidence interval (α) assuming Gaussianity of underlying variables:

$$\rho_{CI} = \frac{e^{2Z}-1}{e^{2Z}+1}, \text{ where } Z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \pm \sqrt{\frac{2}{N-3}}\, erf^{-1}(\alpha). \quad (2)$$

Here, *erf(·)* denotes the standard error function. For example, 90% sample correlation using 60 chip sites in a single wafer guarantees 84~94% correlation range within 95% confidence interval.

One can correlate a figure-of-merit of product with various (often thousands) in-line device-level characteristics, and sort the in-line parameters with decreasing order of their correlations. This cross-correlation analysis allows identifying which set of device parameters (threshold voltage, leakage current, oxide thickness, gate length, etc) are most correlated with a circuit performance. For a highly cross-correlated parameter, the slope of fitted regression line represents a sensitivity of a circuit performance with respect to a certain parameter given by (3). Here, x represents one of the process parameters and y is a circuit performance measure.

$$sensitivity = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\cdot\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}. \quad (3)$$

The cross-correlation and sensitivity information can be beneficially fed back to the product design, model, and technology development. On the design side, the information is used for design-for-yield (DFY): A circuit is modified to be more tolerant to a specific process parameter that has a high correlation, in the next release. Model-to-hardware correlation (MHC) on the correlation and sensitivity allows model calibration. To the technology, the dependency of a particular process parameter(s) on the product performance is recognized and monitored closely.

In the next technology iteration or generation, another set of benchmarking devices and/or circuits further improves the product yield by mitigating the sensitivity of the product performance to the residual process parameters which have not been accommodated in the previous cycle. Done in multiple iterations or generations, it helps understand and monitor the tolerance of a circuit performance to a certain process variations.

3.2 Statistical Yield Estimation

A performance yield is defined as the probability that a certain circuit is within pre-defined design specification(s). The performance yield is statistically estimated by fitting the histogram of measured performance values with a standard probability density function (PDF) such as normal, log-normal, or Weibull. The statistical nature of performance is inevitably from process variations. The resulting integration of PDF over design specification space is the estimated yield:

$$\text{Yield} = \Pr\left(\bigcap_{i=1}^{M} \text{design spec } i \text{ is met}\right) \approx \iint_{\forall x_i \in \text{design spec}} \cdots \int p(x_1, x_2, \cdots, x_M)\, dx_1 dx_2 \cdots dx_M \quad (4)$$

Here, x_i is each design space, and $p(x_i)$ represents a fitted PDF based on x_i samples. This method can accommodate more than one design parameters: for example, power and frequency in a ring oscillator (RO). Fig. 5 illustrates 2-d statistical yield estimation. Based on the 500 measurement samples of 101-stage ROs from 90nm technology, reciprocal of active current ($1/I_A$) vs. delay was fitted with bi-variate normal distribution as shown in the left panel. If the target maximum delay is 15ps and target maximum I_A is 1mA, then the estimated yield is 90.0% (right panel).

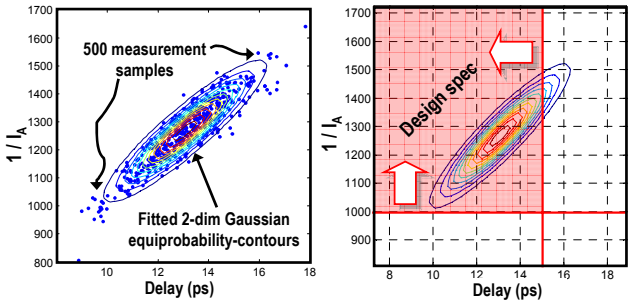


Figure 5. (a) 500 measurement points of reciprocal of active current ($1/I_A$) vs delay for a 101-stage RO in 90nm technology and the bi-variate normal distribution fit. (b) Illustration of the yield calculation using integration of volume under a set of design specifications (maximum active current=1mA and maximum delay=15ps).

This statistical yield estimation method is simple yet effective to provide fast yield prediction solely based on hardware data. The resulting yield curve can be then a basis for comparison for different technology iterations, generations or nodes.

3.3 Decomposition of Systematic Process Variation Space

Due to the complexity of semiconductor manufacturing processes, die-to-die and wafer-to-wafer variations are inter-correlated to some degree. However, the separation of process-induced variability to die-to-die and wafer-to-wafer levels offers insight and deeper understanding of root cause(s) of the variation. The principal component analysis (PCA) is a linear transformation of a set of random vectors to a new set of vectors called principal components (PC's) [6]. The first PC is the direction on which the variance of the projection of the original vector is maximized as expressed in (5).

$$w_1 = \arg\max_{\|w\|=1} \text{var}(w^T x)$$
$$w_k = \arg\max_{\|w\|=1, w\perp w_k \forall i=1,\ldots,k-1} \text{var}(w^T x), \quad k \geq 2 \quad (5)$$

Here, x is the original data vector, and w_i is the PC. The subsequent PC's are defined in the same way except they need to be orthogonal to all the preceding PC's. By definition PC's are uncorrelated and are ordered so that the first few contain most of the variation present in all of the original variables. The constrained principal component analysis (CPCA) is a method to extract constrained principal components (CPC's) which have the same properties with ordinary PC's but are constrained to a pre-defined subspace. In the context of analyzing process variability, it is useful to extract the PC's of die-to-die or wafer-to-wafer variations separately [7]. CPC's can vary only in a guided dimension that corresponds to either die-to-die or wafer-to-wafer variation.

Fig. 6 (a) illustrates how the CPC's can be obtained iteratively. Conventional PCA is sensitive to the scaling (e.g. different units) and offset to the data to which it is applied. Thus, at a preprocessing stage, the data set of each in-line parameter is standardized to be zero-mean and unit-variance. Subsequently the data is screened for anomalies and insignificant values. In the next step, CPCA's are performed to find the most dominating CPC for die-to-die and wafer-to-wafer variation separately. The CPC of larger variance is selected. The data set is, then, transformed to be

orthogonal to the space spanned by the selected CPC. This routine will be iterated for the residual data set until a given criterion is satisfied. Fig. 6 (b) demonstrates how an original data (oscillation frequency of a ring oscillator in 65nm technology) can be successively reconstructed from three most prevailing CPC's. In this illustration, CPC#1, #2, and #3 capture die-to-die, wafer-to-wafer, and die-to-die variation, respectively.

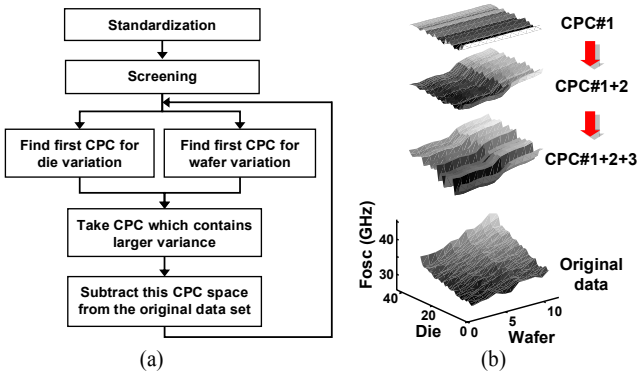
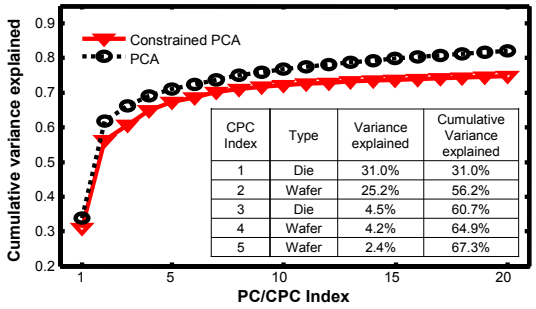


Figure 6. (a) The diagram of the proposed CPCA algorithm. (b) Original data (oscillation frequency of a ring oscillator with respect to die and wafer index for a 65nm technology node and 300mm wafer), CPC#1, CPC#1+#2 and CPC#1+#2+#3.

The CPCA method can allow monitoring of 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 generation/node can be monitored, thus allowing the fast and critical feedback to manufacturing and technology. Based on the results of this analysis, an efficient sampling can be proposed to represent a full wafer (or lot) measurements by a few samples within given accuracy.



CPC Index	Type	Variance explained	Cumulative Variance explained
1	Die	31.0%	31.0%
2	Wafer	25.2%	56.2%
3	Die	4.5%	60.7%
4	Wafer	4.2%	64.9%
5	Wafer	2.4%	67.3%

Figure 7. A Cumulative variance of PCA and CPCA for a typical in-line data (1109 parameters for 520 samples – 40 chips for 13 wafers). It shows the cumulative variance which can be explained by first 20 PC's and CPC's using the ordinary PCA and CPCA, respectively. The variation type (die-to-die or wafer-to-wafer) and variance corresponding to the first five CPC's are listed in the table.

Fig. 7 shows the cumulative variance which can be explained by first 20 PC's and CPC's using the ordinary PCA and constrained PCA, respectively, for a typical in-line data set (1109 parameters for 520 samples – 40 chips for 13 wafers). The first and second CPC corresponding to die-to-die and wafer-to-wafer variation, respectively, explains 31% and 25% of all information, only slightly less than ordinary PC counterparts. This typical data set, hence, justifies that the first few CPC's contain almost as much information as unconstrained PC's, but vary in only one dimension (either in die-to-die or wafer-to-wafer) leading to straight-forward analysis and visualization of variability.

4. APPLICATION

4.1 Experiment Background

The proposed statistical methodologies have been applied to first generations of three-stage CML ICOs. The ICOs were oscillator components in product PLLs for server processors in 65nm SOI technology. As described in Fig. 8, the ICO is a critical function block in microprocessor PLL. An ICO design should have reliable oscillation, target oscillation frequency, 2:1 tuning range, oscillation amplitude, specified power consumption, and acceptable phase noise against process variations. An I/O communication PLL's specification would be more stringent. Considering that an entire digital block's operation depends on the PLL, its physical and functional yields are decisive for a chip's pass or fail.

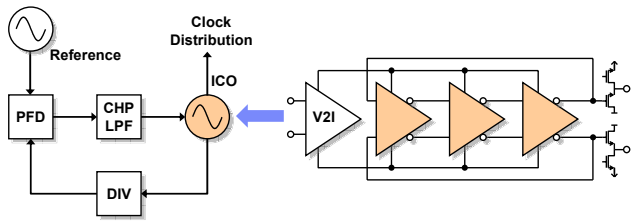


Figure 8. A diagram of ICO as a subcircuit of PLL.

Along the technology node ramp up to 65nm from 90nm, ICO designs were adopted from 90nm server PLL ICO, and developed over generations. Difficulties in migrating a 90nm ICO design to 65nm are the reduced supply voltage headroom, and increased nonlinearity, variation, and leakage current. For example, cascode current mirror reduces differential output swing, so that signal-to-noise ratio (SNR) is reduced as much. The ICO is composed of three-stage CML inverters, as shown in Fig. 9.

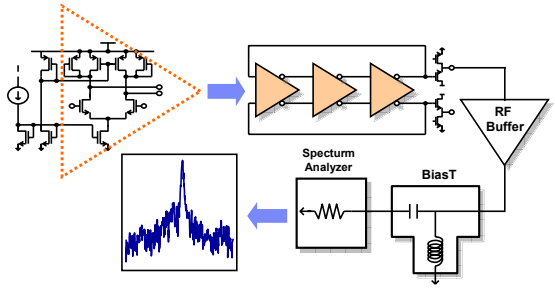


Figure 9. Diagram of a CML inverter in ICO, and measurement set-up for oscillation frequency (Fosc) acquisition.

A low-pass filtered charge pump output is converted to a current, and the current controls CML dc biasing conditions, or the ICO oscillation frequency. Current-controlled mirrors are implemented with body-contacted double-gate thick-oxide NFETs, along with PFET mirrored loads and replicas. The differential pair NFETs have attracted much of design considerations, and they changed over generations. For example, there were two version of ICO for device performance comparison – one with SOI floating-body (FB) and the other with body-contacted (BC) differential pair NFETs in the first generation. Details of ICO developments will be discussed in the following sections.

The ICO output is buffered by source followers, limiter, and divider in the PLLs. For the co-design and convergence efforts, ICO and source followers were implemented for each generation. The source follower output is amplified by a RF buffer, collected by a bias tee, and measured with a spectrum analyzer. An automated 300mm prober and tester database were used to collect ICO oscillation with given conditions. The power supply voltage was 1.2V, and a current biasing was given to have an equivalent tail current at each differential pair. Stand-by current was

measured when no CML biasing current is given to ICO. An acquisition time for an ICO oscillation frequency at one input biasing current is less than 10 seconds. A full 300mm wafer scan can be performed within half an hour. In practice, multiple ICO input biasings are provided to fully characterize frequency tuning range, power consumption, and performance yield.

4.2 Product Design Decision by Device Tuning and Statistical Measurements (Generation 1)

Since the introduction of SOI to digital processor manufacturing, design difficulties, performances, and other properties have been progressively evaluated in product analog blocks. In the 90nm PLL migration to 65nm efforts, both FB and BC NFETs were used in the ICO differential pairs. Layout schematics of partially-depleted FB NFET (cross-section) and BC NFET (top-view) are shown in Fig. 10 (a) and (b), respectively. For an FB device, a body is isolated from the substrate by buried oxide (BOX) layer. A gate of a BC device is typically T-shaped to make a contact to partially-depleted floating body vertically. Because of the body isolation from substrate, SOI FETs have lower junction capacitance, and because of the partially-depleted body, FB FETs have lower threshold voltage, which in turn increases transconductance gain, while they have history effect. Fig. 10 (c) and (d) show the statistical hardware results for FB and BC from the first generation CML ICO.

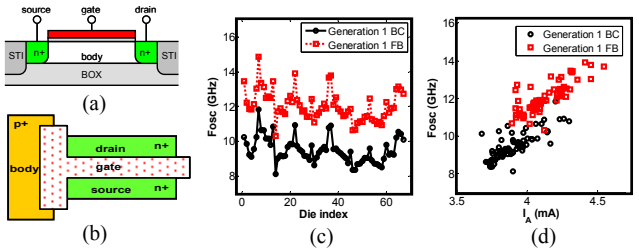


Figure 10. (a) Diagram of Floating body (FB) NFET device, (b) Diagram of Body contacted (BC) NFET device, (c) BC and FB CML ICO Fosc versus site index in Generation 1, (d) Fosc versus I_A for BC and FB CML ICO's in Generation 1.

FB-based and BC-based CML oscillators track well across the wafer with 98% correlation. It is proved that FB FET enhances oscillation frequency by 27% without loss of variation. (The standard deviation of oscillation frequency normalized by the average is 7.2% for FB FET-based ICO, and 7.6% for BC FET-based ICO.) The FB FETs have lower threshold voltage, high transconductance gain, and lower parasitic capacitance than BC FETs. It relieves the ICO design of the reduced voltage headroom in 65nm. Also transconductance gain is much higher with FB FET, while more power consumption is inevitable. Therefore for the yield improvement, the FB device was recommended for the subsequent generations along with device parameter sensitivity analysis.

4.3 Cross-correlation Analysis (Generations 2 and 3)

Fig. 11 (a) shows the scatter plot of Fosc versus threshold voltage (Vth) for generations 2 and 3, and Fig. 11 (b) displays histogram of Vth and Fosc for generation 3. Among many in-line parameters including FET on/off-current and capacitance, Vth is most correlated with Fosc. The correlation decreased from 94% in generation 2 to 13% in generation 3. The ICO tuning based on cross-correlation and sensitivity analyses resulted in a more robust design against process variation. Note that 95%-confidence-interval correlation range for 94% is 81%~98% using (2), representing statistically significant correlation value. It is also observed that the Vth variation itself tightened significantly as generation progressed due to the process improvement.

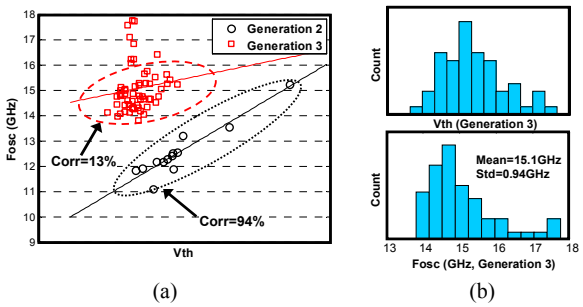


Figure 11. (a) Fosc vs Vth for generations 2 and 3, and their cross-correlations. Each line denotes a linear regression that fits sample points. (b) Histograms of Vth and Fosc in generation 3.

4.4 Variability Decomposition Using Constrained Principal Component Analysis (All Generations)

More than 650 manufacturing floor in-line parameters are used for each 65nm SOI technology generation. Each in-line parameter contains 255 samples (15 dies per wafer for 17 wafers). Wafers used are 300mm, and belong to a same lot for a given generation. Fig. 12 shows the dominating die-to-die CPC's for three technology generations, fitted by the 2nd-order polynomials on the 15 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 runtime for CPCA was less than one minute for each generation case. The first generation shows a highly irregular pattern on a wafer scale that is presumably from a certain process anomaly that is common in the first pre-production cycle. In the second and third generations, a much milder slightly off-centered radial pattern is observed.

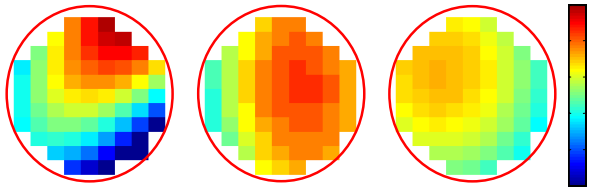


Figure 12. The dominating die-to-die CPC for three 65nm technology generations. (a), (b), and (c) correspond to Gen 1, 2, and 3, respectively. Gen 1 exhibits an irregular pattern on a wafer scale. Gen 2 and 3 show a much milder, radial pattern.

4.5 Experiment Summary

The proposed methodology was applied to a CML ICO that is a subcircuit of a microprocessor PLL, for first three 65nm technology generations. Fig. 13 summarizes the results from the proposed methodology for three generations. Note that the generation 1 was the first 65nm technology development, derived from an earlier 90nm technology node. Also, the CML ICO was migrated from 90nm counterpart. FB NFETs were selected for ICO differential pairs based on the statistical measurements. For generations 1 and 2, exactly same CML ICOs were tested and analyzed so that the impact of the technology to the circuit performance can be monitored. Improvement was observed for technology in terms of Vth variation and for product in terms of nominal Fosc and its relative variation. From generation 2 to generation 3, the technology was continually stabilizing, and also the CML ICO was modified to become more tolerant to process variations. As a result, generation 3 exhibits tightened Vth variation on the technology side, and faster Fosc and less variation on the product side.

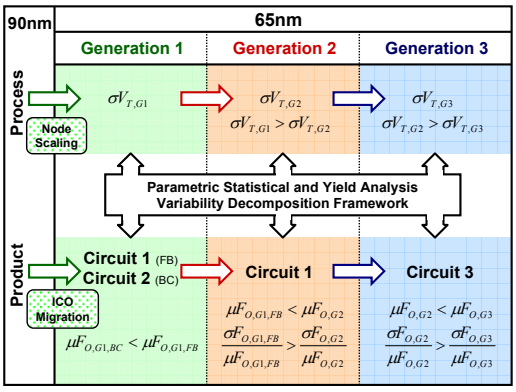


Figure 13. Summary of technology-product co-design results. The $\mu F_{O,G1,BC}$ denotes average Fosc of BC-device-based CML ICO for generation 1. The $\sigma F_x/\mu F_x$ denotes standard deviation of F_x normalized by its average.

Fig. 14 shows the results of the presented methodology in a different fashion using statistical yield calculation. Each curve represents estimated performance yield with respect to minimum Fosc specification via the method addressed in subsection 3.2. The yield at 12GHz target has improved from 47% to 99% between the generation 1 and 3 thanks to (1) SOI FET justification in the ICO circuit migration, (2) cross-correlation and sensitivity analyses between ICO generations (3) 65nm node process improvement through generations, and (4) ICO circuit tuning based on parametric statistical results.

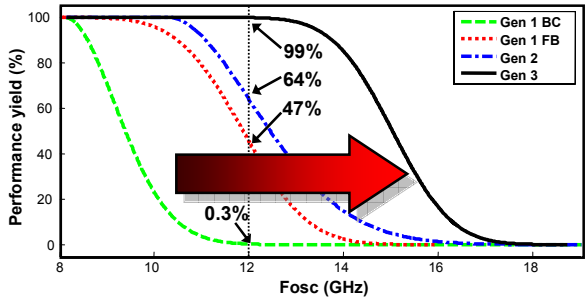


Figure 14. Fosc yield curves for three 65nm technology generations. Each curve represents estimated performance yield with respect to minimum Fosc target. The resulting yields with 12GHz target are shown for each generation. Both BC- and FB-device-based yields are shown for the first generation.

Fig. 15 shows Fosc average and normalized standard deviation for the target, generations 1, 2, and 3. The target Fosc was 12GHz, and the target standard deviation was 8%. Generation 1 did not meet both design targets. Due to the technology improvement, the generation 2 passed the nominal Fosc specification, but did not qualify the variation target. As a result of statistical co-design approach, generation 3 satisfied both design targets by far. The x-axis grid in this figure was not equally spaced to represent progressively diminishing development period as generation evolves and stabilizes. The proposed methodology was iteratively applied in the early technology generations and allowed rapid and successful yield learning.

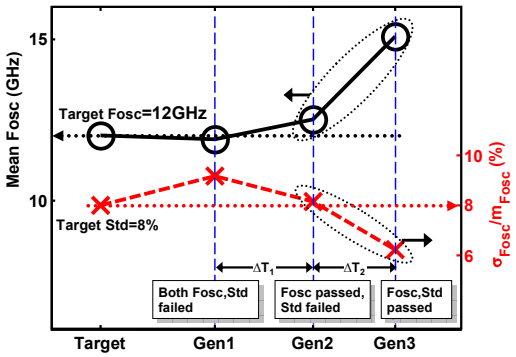


Figure 15. Mean Fosc and standard deviation normalized by mean, over three generations.

5. CONCLUSIONS

To expedite the co-design and convergence of technology, product, and model components, we exercised a statistical framework that includes the following tools:

- A cross-correlation analysis identifies what device-level characteristics are most related to the performance of a product. By linking technology parameters with a product figure-of-merit, the resulting information helps both product design and technology development.
- A performance yield of a product is statistically estimated for given design targets, based on hardware data. This analysis leads to a quick assessment and qualification of a product performance.
- A variant of principal component analysis allows perceptually orthogonal decomposition of a systematic variability. The systematic die-to-die and wafer-to-wafer variations are obtained and visualized using this method.

The aforementioned tools enable translations between technology, product, and model interaction, and foster pertinent collaborations from different components.

A current-controlled oscillator (ICO) of a 64-bit server processor has been developed using the proposed framework, across three different 65nm technology generations. As a result of our proposed methodology and collaborative effort of technology and product teams, the ICO enjoyed the rapid yield enhancement from 47% to 99% over three generations.

6. REFERENCES

[1] J.-A. Carballo and S. R. Nassif, "Impact of Design-Manufacturing Interface on SoC Design Methodologies," *IEEE Design & Test of Computers*, vol. 21, no. 3, pp.183-191.

[2] S. R. Nassif, "Modeling and Forecasting of Manufacturing Variations," *Proc. ASP-DAC*, 2001, pp.145-149.

[3] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, V. De, "Parameter Variations and Impact on Circuits and Microarchitecture," *IEEE Proc. DAC*, 2003, pp.338-342.

[4] D. S. Boning and J. E. Chung, "Statistical Metrology: Understanding Spatial Variation in Semiconductor Manufacturing," *SPIE Proc. Symp. on Microelectronic Manufacturing*, 1996, vol. 16, pp.16-26.

[5] M. Ketchen, *et al.*, "High Speed Test Structures for In-Line Process Monitoring and Model Calibration", *IEEE Proc. ICMTS*, 2005, pp.33-38.

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

[7] C. Cho, *et al.*, "A Data-Driven Statistical Approach to Analyzing Process Variation in 65nm SOI Technology", *IEEE Proc. ISQED*, 2007, pp 699-702.