Design Dependent Process Monitoring for Back-end Manufacturing Cost Reduction

Tuck-Boon Chan*, Aashish Pant*, Lerong Cheng†, Puneet Gupta*
{tuckie,apant,puneet}@ee.ucla.edu, lerong.Cheng@sandisk.com,
*Department of Electrical Engineering, University of California, Los Angeles
†Sandisk Inc.

***Abstract*—Short-loop process monitoring structures (usually simple device $I-V$, $C-V$ measurements made after M1 fabrication) are commonly put in wafer scribe-lines. These test structures are almost always design independent and measured/monitored by the foundry to keep track of process deviations. We propose a design-dependent process monitoring strategy which can accurately predict design performance based on simple I_{eff}-based delay and I_{off}-based leakage power estimates. We show that our strategy works much better (0.99 correlation vs. 0.87) compared to conventional design-independent monitors. Further, we use the predicted delay and leakage power for early yield estimation for pruning bad wafers to save test and back-end manufacturing costs We show that wafer pruning based on our approach can achieve upto 98% of the maximum achievable benefit/profit. We design the measurement and prediction schemes so as to minimize data as well as computation that needs to be kept track of during wafer fabrication. Such design-dependent process monitoring can help target process control/optimization effort, enable quicker yield ramp besides saving test and manufacturing costs.**

I. Introduction

Modern manufactured chips exhibit wide power/performance spread which necessitates careful screening. Frequency and power tests done after packaging to screen defective chips are expensive and time consuming. Moreover, the defective chips till this point have already incurred large manufacturing and packaging costs. Therefore, there is an incentive to prune bad wafers and chips during early stages of manufacturing wherever possible using simple wafer level tests.

An example of post-silicon diagnosis is shown in [6,19], where delay values of ring oscillators (RO) are used as references for delay defects screening. There is an inherent error in the estimation because every critical path has different sensitivities to process variations. A configurable critical path monitor is introduced in [4], which tracks critical delay within small error. Note that directly measuring path delays is not applicable for early wafer pruning because measuring cannot be done until all metal layers are fabricated.

In [8], parametric test structures are placed in scribe-lines (i.e. empty space between dies) to detect process variations for circuit performance evaluation without sacrificing wafer area. The test structures are designed to be similar to the circuit to ensure high correlation with critical path(s). Due to area constraints, test structures placed in scribe-lines are limited. Therefore, they are unlikely to capture all critical path delays variation of a circuit, which have different sensitivities to process variations.

Cho et. al. [9] measure electrical parameters from manufacturing in-line benchmark structures (MIBS) to train a neural network for product performance prediction. The method is made capable of targeting multiple critical paths and other specification constraints. Since the neural network is trained or fitted based on a set of training data and output performances, characteristics of target circuit are not modeled explicitly. This can lead to unpredictable errors when process variations are different from the chosen set of training data.

Alternatively, Liu et. al. [7] proposes a method to synthesize a representative critical path for post-silicon delay prediction. Although the synthesized critical path is designed to have maximum correlation to all critical paths, it may fail to match circuit performance variations. This happens whenever process variation is not evenly distributed as predicted but amplified by a particular process parameter.

In this work, we propose a design dependent approach of accurately estimating circuit performance and leakage power after the metal-1 stage of manufacturing. As opposed to adding new test structures, we try to leverage existing process monitoring I-V, C-V measurements which are very commonly used in modern silicon foundries. We derive the design-specific sensitivities of leakage power and delay of critical paths to changes in *off current* (I_{off}) and *effective drive current* (I_{eff}) [10]–[13] respectively. These sensitivities help in concisely modeling the behavior of the design to process variations, which can then be used by the foundry for delay and leakage power estimation. Our work is different from the ring oscillator guided performance measurements. Though ring oscillator guided testing strategies [6,19] are common, we have not seen any work dealing with designing scribe-line ring oscillators which is design specific. Moreover, due to area constraints, only a small number of such ring oscillator based test structures can be embedded in the scribe-lines. In contrast, simple scribe-line based current and capacitance measurements are almost universally done and in this work, we attempt to use these measurements for delay prediction.

A. Device I-V and C-V measurement

In this paper, we assume that device parameters are obtained from the compact scribe-line test structures (e.g. [3]). These test structures are design independent and placed in scribe-line and capable of measuring individual device currents

and capacitance. We assume that following parameters are measured from scribe-line test structures [1]:

- I_h=I_{ds} at $V_{gs} = V_{dd}$, $V_{ds} = V_{dd}/2$
- I_l=I_{ds} at $V_{gs} = V_{dd}/2$, $V_{ds} = V_{dd}$
- I_{off}=I_{ds} at $V_{gs} = 0$, $V_{ds} = V_{dd}$
- C_{gate} at $V_{gs} = V_{dd}$, $V_d = V_s = 0$

In reality, many sources can alter measurement values. For example, random local (within-die) variation, voltage supply and temperature fluctuations, probe contact resistance, etc can induce uncertainties in measured values. To reduce the uncertainties, it is common to have multiple devices under test connected in parallel and carry out the measurement repeatedly. We assume every measurement is repeated N_e times and the scribe-line test structure has N_d devices connected in parallel. Thus, only the sum of device currents and capacitance of every chip are measured, i.e., the mean I_l,I_h,I_{off} and device capacitance per unit width are obtained.

There are various practical scenarios where the proposed methodology can prove beneficial. The computed design specific sensitivities can be used for process monitoring and optimization. As delay and leakage power estimation can be done after metal-1, it can also be used for wafer pruning. In this work, our contributions are the following:

- We propose a scribe-line based design dependent approach for circuit performance and leakage power estimation using I_{eff} and I_{off} and present methods for statistically incorporating within die variation and measurement noise effects.
- We show how the above information can be used to accurately identify bad wafers and help in wafer pruning and yield estimation.

The overview of our approach is depicted in Figure 1.

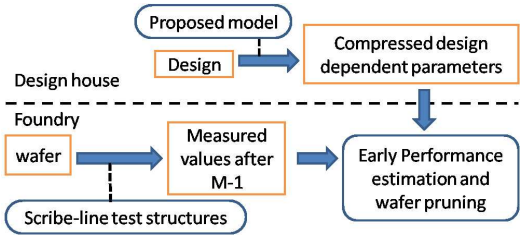


Fig. 1. Overview of proposed approach.

Rest of this paper is organized as follows. In section II, we discuss our I_{eff} current based path delay estimation model. In section III, we describe our I_{off} current based leakage power estimation model. In section IV, we describe how our analysis can be used for early wafer pruning. In section V, we present the results using our detailed wafer level simulation setup. We conclude in section VI.

II. Delay Estimation Using I_{eff}

I_{eff} is the average current that charges or discharges a circuit node during a logic transition. The charging or discharging delay can be expressed as

$$delay \propto \frac{CV}{I_{eff}}\ ,\ I_{eff} = \frac{I_h + I_l}{2} \tag{1}$$

where C is the node capacitance that is being charged (or discharged), V is the voltage swing and I_{eff} is the effective drive current. While I_{eff} cannot be physically measured, several works propose approximations using device level I-V characteristics [11]–[13]. In this work, we use I_{eff} from [11], where I_h and I_l are defined in Section I-A. Though more complex models (e.g. [12]) can be used as well, our experiments indicate that (1) suffices for our device models and libraries.

A. Cell Delay Model

Using (1), we can express the propagation delay of a cell type (c) (for example, INV, NAND etc) as

$$d_{cell}(c) = \sum_{t \in T} \frac{K_{cell}(c,t)CV}{I_{eff}(t)} \tag{2}$$

where T is the set of all device types [2]. $K_{cell}(c,t)$ is the cell and device type specific delay scaling coefficient, which is fitted for different input slew and output load combinations. We do not show the explicit dependence on slew and load for notational convenience. Also, note that these coefficients are specific to a rise or fall transition. This fact is implicit and we do not show it for notational convenience. Expanding (2) using Taylor series with respect to $I_{eff}(t)$ for all $t \in T$ and ignoring the crossing and higher order terms, we get

$$\begin{aligned} d_{cell}(c) &= d_{cell\_0}(c) \\ &- \sum_{t \in T} \frac{K_{cell}(c,t)CV}{I_{eff\_0}(t)} \left(\frac{\Delta I_{eff}(t)}{I_{eff\_0}(t)} - \frac{\Delta I^2_{eff}(t)}{2I^2_{eff\_0}(t)}\right) \end{aligned} \tag{3}$$

where d_{cell_0} and I_{eff_0} denotes the corresponding quantity under nominal process conditions. $K_{cell}(c,t)$ is the sensitivity of cell delay to $I_{eff}(t)$ and these coefficients are fitted for every cell using (3) by varying process conditions for different input slew and output load points. *This model fitting can be done very efficiently as it can use existing process specific timing libraries which are available for various corners*. Since most cells consist of single V_{th} devices, they have two non-zero $K_{cell}(c,t)$ coefficients out of four device types. In our experiments, we do not have access to a sufficient number of these libraries. Therefore, we fit the model using spice simulations on individual cells.

B. Path Delay Model

The delay of path j under process variations can be expressed as

$$d_{path}(j) = d_{path\_0}(j) + \Delta d_{path}(j)$$

[1]The bias points are derived from commercial device data sheets.

[2]In this work, we consider design with four device types: {high V_{th}, low V_{th}}×{ pmos, nmos }

where $d_{path\_0}(j)$ refers to nominal delay of path j. $\Delta d_{path}(j)$ is the delay change due to process variation, which is equal to the sum of delay changes of every cell in the path,

$$\Delta d_{path}(j) = \\ -\sum_{i\in G_j}\sum_{t\in T}\frac{K_{cell}(i,t)C(i)V}{I_{eff\_0}(t)}\left(\frac{\Delta I_{eff}(t)}{I_{eff\_0}(t)} - \frac{\Delta I^2_{eff}(t)}{2I^2_{eff\_0}(t)}\right)$$

where G_j is the set of all cell instances which belongs to path j The sensitivity of delay of path j to changes in $I_{eff}(t)$ can therefore be expressed as

$$K_{path}(j,t) = \sum_{i\in G_j} K_{cell}(i,t)C(i) \tag{4}$$

Note that $K_{cell}(i,t)$ is instance-dependent as input slew and output load may vary with instance. The total path delay can now be written as

$$d_{path}(j) = d_{path\_0}(j) - \\ \sum_{t\in T}\frac{K_{path}(j,t)V}{I_{eff\_0}(t)}\left(\frac{\Delta I_{eff}(t)}{I_{eff\_0}(t)} - \frac{\Delta I^2_{eff}(t)}{2I^2_{eff\_0}(t)}\right) \tag{5}$$

C. Handling Load Capacitance Variation

In (4), the path specific delay sensitivities to I_{eff} depend on the nominal value of output load, which is seen by the cells. However, with process variations, this output load also changes. Therefore we scale the estimated delay by the ratio of actual device capacitance to nominal capacitance. i.e.

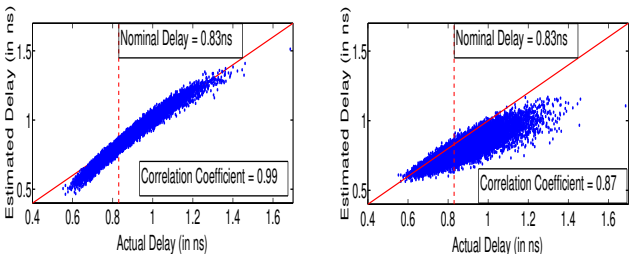
$$d'_{path}(j) = (d_{path}(j) - d_{path-interconnect}(j))\frac{C_{gate}}{C_{nom}} + \\ d_{path-interconnect}(j) \tag{6}$$

where C_{gate} is process variation affected capacitance (measured by scribe-line monitors), C_{nom} is its nominal value and $d_{path-interconnect}(j)$ is total interconnect delay of a critical path.

Figure 2 shows the benefits of the proposed design dependent delay estimation technique as tested on *C432 ISCAS85* benchmark. The delay estimated using (6) tracks the actual delay well. The correlation coefficient is found to be 0.99 as against 0.87 for a design independent approach (in which delay is estimated to be inversely proportional to the mean I_{eff} of all device types). This is because the design independent methodology is oblivious of the exact nature, topology and the structure of the cells that make up the critical paths in the design while our strategy effectively captures this dependence in the $K_{path}(j,t)$ form.

D. Effect of Within Die Variation on Delay

I_{eff} values measured from test structures are typically different from the ones on critical paths due to within die variation. Since the variation is usually random, it is expressed as a normally distributed random variable with zero mean and standard deviation, $N(0,\sigma_{wd})$. The distribution can be estimated by making multiple measurements per die or from pre-existing characterization. Considering the first order term



(a) Proposed Delay Model (b) Design Independent Model

Fig. 2. Scatter plot (C432 Monte-Carlo timing simulations) that shows how the delay estimated by (a) proposed delay model and, (b) a design independent approach, compared with actual delay for an ISCAS85 C432 benchmark.

in (5) $d'_{path}(j)$, the path delays are rewritten in concise matrix form as,

$$\mathbf{D} = \begin{bmatrix} d'_{path}(1) \\ \vdots \\ d'_{path}(z) \end{bmatrix} + \mathbf{W}.\mathbf{I}_{wd}, \mathbf{W} = \begin{bmatrix} w_{11}\ldots w_{1n} \\ \vdots\ddots\vdots \\ w_{z1}\ldots w_{zn} \end{bmatrix},$$

$$w_{ji} = \begin{cases} K_{cell}(i,t) & \text{if cell instance } i \text{ is on path } j \\ 0 & else \end{cases}$$

where z is the total number of paths, n is the total number of cells. Every entry in $\mathbf{I}_{wd}$ is an independent normal random variable, $N(0,\sigma_{WD})$. Performance of the circuit is given by

$$delay_{max} = \max_{j=1}^{z}(d'_{path}(j)). \tag{7}$$

Since the path delays are correlated, we need to evaluate the covariance of critical paths, ($\mathbf{WW}^T$). Due to the large number of critical paths and cells, keeping the entire covariance matrix on test machines is not practical. To reduce the size of $\mathbf{W}$, we extract and use its v largest principle components (PC). This reduces the total data size by a factor of v/n but some correlation information is lost and the variance of each path is less than the exact correlation value. To ensure that we do not underestimate the variance of path delays, a residue term $\mathbf{R}$ is introduced. This residue is assumed to be uncorrelated such that it is unlikely to underestimate the path delay. Therefore, the path delays can be expressed as

$$\mathbf{D} = \begin{bmatrix} d'_{path}(1) \\ \vdots \\ d'_{path}(z) \end{bmatrix} + \mathbf{W'}.\mathbf{I}_{wd} + \mathbf{R}, \\ \mathbf{R}^T = [r_1.i_{res\_1}, \ldots, r_z.i_{res\_z}], \tag{8}$$

where $\mathbf{W'}$ is the compressed matrix with v principle components and i_{res} are normal random variables. Each residue element in $\mathbf{R}$ is given by

$$r_j = \sum_{i=1}^{n} w_{ji} - \sum_{x=1}^{v} w'_{jx},$$

where w_{jy} and w'_{jx} are the entries of $\mathbf{W}$ and $\mathbf{W'}$, respectively. Though part of the correlation information is not captured,

Figure 3 shows that our method is efficient in reducing pessimism in delay estimation in contrast to assuming that all paths are completely independent. Moreover, this method is flexible as it provides for a trade-off between accuracy and data size by choosing a suitable number of principle components. The size of correlation matrix is O(v * number of paths). Based

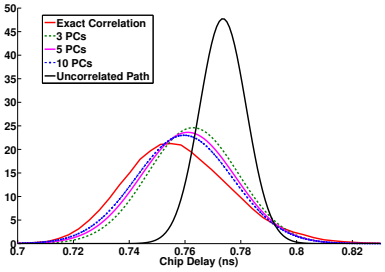


Fig. 3. Comparison between delay distributions for circuit C432.

on (8), the delay of a critical path can be written as,

$$d_j = d'_{path}(j) + y_j.i_{wd} + r_j.i_{res}, \tag{9}$$

where y_j is j^{th} row of **W'**, i_{wd} and i_{res} are independent normal random variables. Equation (9) is in the canonical form for tightness probability calculation to solve (7). By using the method proposed in [5], the mean and variance of the maximum of two or more timing quantities can be obtained. Thus, we can express the maximum delay of z critical paths as a normal distribution. We implemented the calculation in [5] hierarchically. I.e., we recursively calculate maximum delays of distinct critical path pairs. This reduces the error in mean and variance estimation as the number of times $maximum$ operation is performed increases logarithmically with the number of critical paths.

E. Dealing with Measurement Noise

The mean of measured I_{eff} for a chip is given as

$$\hat{I}_{eff} = \frac{1}{N_e}\sum_{m=1}^{N_e}\frac{\tilde{I}_{eff}(m)}{N_d} \tag{10}$$

where $\tilde{I}_{eff}(m)$ is the total I_{eff} from m^{th} measurement. Considering measurement noise,

$$\tilde{I}_{eff}(m) = (1+F_m)\sum_{s=1}^{N_d}[I_{eff} + I_{wd}(s)] \tag{11}$$

where I_{eff} is the exact value, I_{wd} is the effect of within die variation and F_m is measurement noise. Combining (10) and (11),

$$I_{eff} \approx \hat{I}_{eff}(1 + \sum_{m=1}^{N_e} F_m/N_e) - \frac{1}{N_d}\sum_{s=1}^{N_d} I_{wd}(s)$$

since $\sum_{m=1}^{N_e} F_m/N_e \ll 1$.

Since I_{wd} and F are Gaussian random variables, I_{eff} is also a Gaussian random variable with its mean and variance given by

$$\begin{aligned}\mu_{I_{eff}} &= \hat{I}_{eff} \\ \sigma^2_{I_{eff}} &= \frac{\hat{I}^2_{eff}\sigma^2_F}{N_e} + \frac{\sigma^2_{I_{wd}}}{N_d}\end{aligned}$$

where $\sigma^2_{I_{wd}}$ and σ^2_F are the variance of within-die variation and measurement noise, respectively. Note that, the variance of I_{eff} is inversely proportional to the number of measurements and total devices in the test structure. In this paper, unless otherwise mentioned, we assume 5 measurements are taken every time ($N_e = 5$) and there are 10 devices in each test structure ($N_d = 10$). We assume 3σ of measurement noise to be 5% of nominal I_{eff} value. I_{wd} is obtained by running Monte-Carlo simulation over variation ranges specified in Table I.

F. Interconnect Delay Variation

The proposed model cannot handle the delay variation because of variations in interconnect metal layers. The effect of interconnect variation is however less pronounced due to following reasons [20]:

- Delay change averages out across all metal wires in a path.
- Width variation changes wire resistance and capacitance of in opposite ways, thus reducing the net effect on RC.

Nonetheless, we include this effect in our experiments and analyze the error incurred in estimation of delay because of variation in interconnect metal layers.

III. Leakage Power Estimation Using I_{off}

A. Leakage Power Model

We model cell leakage power of an instance as linear function of I_{off} [3],

$$P(i) = \sum_{t\in T}\alpha_c(t)I_{off}(i,t)$$

where $\alpha_c(t)$ is the leakage power fitting coefficient for cell type (c) and device type (t). The full chip leakage power of a design is therefore,

$$P_{chip} = \sum_{t\in T}\sum_{c\in\Gamma}\sum_{i=1}^{N_c}\alpha_c(t)I_{off}(i,t) \tag{12}$$

where N_c is the total number of instances of cell type c in the design, Γ is the set of all cell types and $I_{off}(i,t)$ is leakage current of device type (t) in cell instance i.

B. Off Current Variation Model

To estimate leakage power variation, we use a similar approach as that in [22] whereby I_{off} is modeled as an exponential function of variation sources.

$$I_{off}(i,t) = I_{off\_0}(t)e^{Y(i,t)}$$

[3]In this paper, we only consider subthreshold leakage but the model can be easily extended to consider gate leakage.

where I_{off_0} is the nominal I_{off} and Y represents the impact of variation sources. In this work, we assume Y to be the linear combination of all variation sources and model it as a Gaussian random variable with zero mean. Moreover, variation sources are decomposed into inter-die and within-die variation:

$$I_{off}(i,t) = I_{off\_0}(t)e^{Y_g(t)+Y_r(i,t)} \tag{13}$$

where Y_g denotes total inter-die variation and Y_r is the total within-die variation. Combining (12) and (13), we have

$$P_{chip} = \sum_{t\in T} I_{off\_0}(t)e^{Y_g(t)} \sum_{c\in\Gamma} \alpha_c(t) \cdot N_c \cdot \mu_r(t) \tag{14}$$

Since N_c is large, we can approximate the sum of e^{Y_r}'s as the sum of their mean [22], i.e.,

$$\sum_{i=1}^{N_c} e^{Y_r(i,t)} \approx N_c \cdot \mu_r(t)$$

where $\mu_r(t)$ is the mean of $e^{Y_r(i,t)}$. In this work, $\mu_r(t)$ is obtained by running Monte-Carlo simulations. In practice, foundry can use historical data to estimate $\mu_r(t)$.

C. Dealing with Measurement noise

Equation (14) shows that we need to know Y_g to estimate total leakage power which is derived from measurements. As mentioned earlier, we take N_e measurements of the current of N_d devices in test structures. Considering measurement noise and within die variation, the m^{th} measured I_{off} is modeled as

$$\begin{aligned}\tilde{I}_{off}(m) &= \sum_{s=1}^{N_d} I_{off\_0}(t)e^{Y_g+Y_r(s)}(1+Z_m) \\ &\approx N_d I_{off\_0}(t)\mu_r e^{Y_g}(1+Z_m),\end{aligned} \tag{15}$$

where Z is a unitless scalar to model measurement noise. From (13) and (15), the estimated value of Y_g is given by

$$\hat{Y}_g = Y_g + \frac{1}{N_e}\sum_{m=1}^{N_e} ln(1+Z_m) \tag{16}$$

where Y_g denotes the exact value. Since measurement noise Z_m is much smaller than 1, (16) can be simplified as

$$Y_g = \hat{Y}_g - \frac{1}{N_e}\sum_{m=1}^{N_e} Z_m$$

From the above equation, we observe that the exact inter-die variation Y_g is a random variable centered at $\hat{Y}_g$. Since Z_m's are Gaussian random variables, Y_g is a Gaussian random variable given $\hat{Y}_g$ is a Gaussian random variable. The mean and variance of Y_g is

$$\begin{aligned}\mu_{Yg} &= \hat{Y}_g \\ \sigma^2_{Yg} &= \sigma^2_Z/N_e\end{aligned} \tag{17}$$

Since each $Y_g(t)$ is a Gaussian random variable, $e^{Y_g(t)}$ is a lognormal distribution. From Equation (14), we find that P_{chip} is the sum of lognormal distribution. Thus, we can apply Wilkinson's approach [22] to approximate the sum of lognormal random variables as another lognomal random variable by matching the mean and variance. Note that, the uncertainties in $Y_g(t)$ are caused by within-die random variation and measurement noise, which are mutually independent. Therefore, the mean and variance of P_{chip} can be calculated as the sum of mean and variance of $e^{Y_g(t)}$.

IV. Early Wafer Pruning Analysis

Often, accurate circuit performance becomes available only after dicing and packaging. Therefore, any failed chip at that stage incurs losses due to unneeded fabrication, packaging and testing costs. This can be avoided by using M1-testable scribe-line test structures to do wafer pruning, which can save back-end (layers beyond M2) processing costs in addition to wafer sort test cost.

A. Passing Probability for a Chip

In previous sections, we have shown that given the measured currents and capacitance, the distribution of delay and leakage power can be estimated. Based on design specifications, the probability of a chip meeting timing constraint is given by

$$\Pr\{\text{chip delay} \le D_{\text{spec}}\} = \Phi(\frac{D_{\text{spec}} - \mu_{\text{delay}}}{\sigma_{\text{delay}}}).$$

where D_{spec} is the maximum allowed delay for a design, μ_{delay} and σ_{delay} are the mean and standard deviation of maximum delay distribution. On the other hand, the probability of a chip meeting leakage power constraint is given by

$$\Pr\{P_{\text{chip}} \le P_{\text{spec.}}\} = \Phi[\frac{ln(P_{\text{spec}}) - \mu_L}{\sigma_L}],$$

where μ_L and σ_L is the mean and varaince of $ln(P_{chip})$, respectively. Given the measured values of every chip (I_{eff},I_{off} and capacitance), uncertainties in delay estimation are due to within die variation and measurement noise while uncertainty in leakage power estimation is only induced by measurement noise (within die leakage power is modeled as a mean shift). Since the measurements of I_{eff} and I_{off} are different, the probability distributions of the estimations are independent. Thus, the passing probability of a chip is

$$\Pr\{P_{\text{chip}} = \text{pass}\} = \Pr\{P_{\text{chip}} \le P_{\text{spec.}}\} \cdot \Pr\{\text{chip delay} \le D_{\text{spec}}\} \tag{18}$$

Therefore, the expected number of good chips in a wafer can be estimated as

$$EG_w = \sum_{all\ \ chips\in w} \Pr\{P_{\text{chip}} = \text{pass}\} \tag{19}$$

B. Cost Analysis

After fabricating Metal-1, current and capacitance values are extracted. Then, we can decide to scrap a wafer or continue back-end-of-line processes based on projected profit. Let M_f and M_b be the front-end-of-line and back-end-of-line manufacturing cost, M_t be the full-chip testing cost and M_s the scribe-line testing cost per wafer,

$$\begin{aligned}\text{Additional Cost} &= (M_b + M_t), \text{ and} \\ \text{Expected profit} &= \text{Expected good chips} \times \text{Chip price} \\ &\quad - (M_f + M_b + M_t + M_s).\end{aligned} \tag{20}$$

If the final number of working chips is close to the expected number of good chips, it is profitable to continue processing

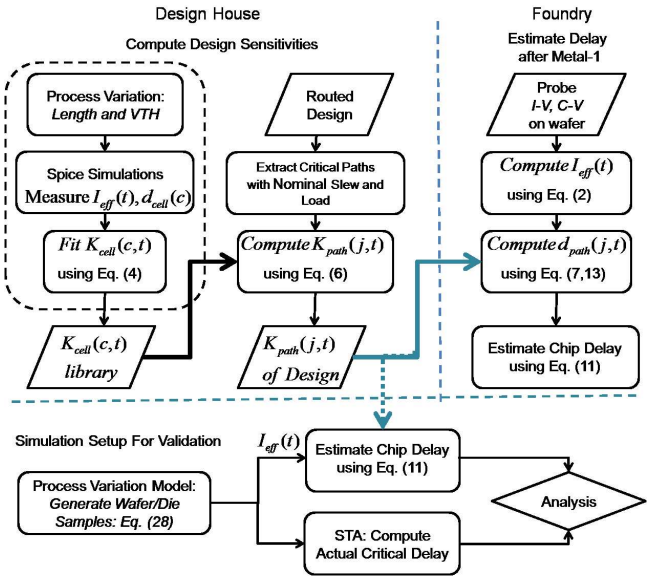


Fig. 4. The proposed critical delay estimation strategy. The left part of the figure shows how the compressed design dependent parameters are computed, while the right part indicates how delay is estimated using these parameters at the foundry. The bottom part of the figure shows our simulation setup for validation of our method. The corresponding flow for leakage power estimation is similar and we do not show it for brevity.

the wafer as long as *expected profit* is larger than additional cost. M_s is usually negligibly small compared to other costs. Thus, we assume measurements are taken for every chip rather than sampling the measurements using [14,15]. Note that the cost for front-end process is not added in *additional cost* because the process has been carried out and incurred processing cost regardless of the decision.

V. Experiments and Results

Figure 4 summarizes the proposed critical delay estimation strategy. The left part of the figure summarizes the extraction of K_{cell} library by doing Monte-Carlo Spice simulations for all cells. Note that, this extraction can also be done using timing libraries at various process corners. Using the K_{cell} library, the design specific K_{path} coefficients are computed for every critical path. We consider all those paths with nominal delay within 5% of the nominal critical path delay[4]. The designs are synthesized using 45nm Nangate Open Cell library [21]. The right side of the figure shows how these compressed design dependent delay coefficients are used to estimate chip delay during manufacturing after Metal-1. To verify the proposed delay estimation method, we build an elaborate simulation setup as shown in the bottom part of Figure 4. The variation model used to generate the process variation samples on the wafer/die is described in the next subsection. The wafer diameter is $300mm$ and the chip dimensions are assumed to be $10mm$x$10mm$. 250 wafers with 657 chips are simulated for every design and the expected number of good chips and good wafers is estimated using (18)[5]. In our experiments, the timing constraint is taken to be 110% of nominal critical path delay of the respective designs. The leakage power constraint is taken to be $5X$ the nominal leakage power.

A. Variation Model

We model five independent variation sources for transistors and they are summarized in Table I. V_{th} variations are modeled by Gaussian distributed random variables with no spatial variation [18]. Channel length is modeled as [16] to include systematic across-wafer variation:

$$D_{sys} = ax^2 + by^2 + cx + dy + exy, \tag{21}$$

where x and y represent the coordinates of a chip's centroid. The values of a, b, c, d and e are obtained by matching systematic delay variation across wafer to 65nm silicon data[6]. Other variation parameters indicated in Table I are extracted from the same silicon data.

Interconnect variation is modeled as random Gaussian distributed die to die variation [17]. In our experiments, this is implemented by perturbing resistance and capacitance values in LEF.

TABLE I
SUMMARY OF VARIATION PARAMETERS

Variation Source	$Wafer-Wafer_{ran}\%$	$Die-Die_{sys}\%$	$Die-Die_{ran}\%$	$Within-Die_{ran}\%$
Channel length	$N(0, 2.13)$	$ax^2+by^2+cx+dy+exy$	$N(0, 1, 29)$	$N(0, 1.56)$
Devices' V_{th}	$N(0, 6.4)$	–	$N(0, 6.08)$	$N(0, 4.7)$
Wire width	–	–	$N(0, 6.08)$	–
Wire thickness	–	–	$N(0, 10)$	–

B. Results

For wafer pruning analysis, we define *wafer passing threshold* (WPT) as the minimum percentage of good chips that a wafer should have in order to be considered a good wafer. WPT can be derived from (20) such that a good wafer always has a positive expected profit, i.e., a wafer passing WPT is likely to have a larger profit compared to back-end-of-line manufacturing cost. Therefore, we define β_w as the passing decision of the wafer w

$$\begin{aligned} \beta_w = &\ 0 \ \text{ if } \ EG_w < (\text{WPT * Total Chips in Wafer}) \\ = &\ 1 \ \text{ otherwise} \end{aligned} \tag{22}$$

To quantify the quality of our wafer pruning approach, we define *Wafer Pruning Benefit* (WPB) of a wafer as

$$WPB_w = \beta_w * (AG_w - \text{WPT * Total Chips in Wafer})$$
$$WPB = \sum_{w=1}^{\text{All Wafers}} WPB_w$$

where AG_w is the actual number of good chips on wafer w. Note that, any wrong selection, either picking a bad wafer or

[4]Many improved critical path selection algorithms have been proposed in literature. This is beyond the scope of our work.

[5]We use 5 principal components for I_{eff} of each device type.

[6]For our model, $a = 7.7e{-}4$, $b = 1.0e{-}3$, $c = -1.6e{-}2$, $d = -7.8e{-}3$, $e = 1.6e-4$

TABLE II
COMPARISON OF TOTAL WPB OF DIFFERENT WAFER PRUNING STRATEGIES. THE WPB IS NORMALIZED W.R.S TO THE IDEAL WAFER PRUNING SETUP.

bench mark	WPT = 25%		WPT = 40%		WPT = 50%	
	Dep%	Indep%	Dep%	Indep%	Dep%	Indep%
c432	94.26	75.60	95.69	61.76	97.97	62.40
s15850	97.97	89.55	97.12	83.07	97.46	82.43
s38584	95.26	91.83	94.26	76.89	92.80	75.66
mips789	96.97	83.17	94.24	72.07	92.04	64.95
c432 (low V_{th} device only)	99.84	88.20	99.88	73.94	99.59	73.26

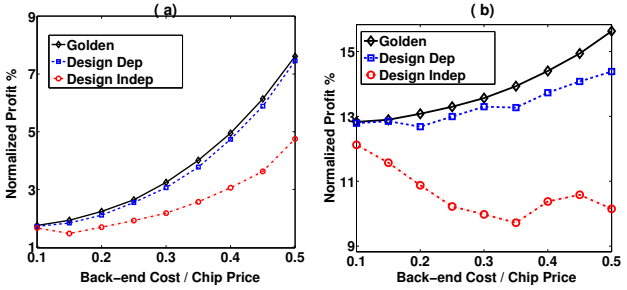


Fig. 5. Profit for benchmark design (a) c432, (b)mips789 using different wafer pruning strategies. The profit is normalized to the total selling price of all die assuming 100% yield. X-axis is the ratio between back-end-of line manufacturing cost to the chip price. (Total number of wafers is 250.)

dropping a good wafer affects the total WPB. Picking a bad wafer (false escapes) contributes to a negative WPB, while dropping a good wafer (yield loss) does not increase the WPB.

In Table II, we compare the WPB values of our method to design independent wafer pruning approach for a combination of ISCAS85 and OpenCores benchmarks. The design independent approach is implemented with delay and leakage power model which has equal proportion of high V_{th} and low V_{th} cells. The values in Table II are normalized (expressed as percentage) to the WPB of the ideal wafer pruning setup which is guided by the exact delay and leakage power of every chip (AG_w). From the table, we observe that the total WPB of our approach is always above 90% while the WPB for design independent approach ranges from 60% to 90%. We also compare the net profit of different wafer pruning approaches. Assuming the total manufacturing cost of a chip to be 60% of chip's selling price, we calculate the net profit after wafer pruning for different split-ups between front-end and back-end-of-line manufacturing costs. The results are shown in Figure 5 for $C432$ and $Mips789$ benchmark designs. From the figure, we observe that the net profit realized through our strategy is very close to the ideal wafer pruning approach and is always higher than the design independent methodology. In Section II and III, we discussed the impact of test-structure design on measurement noise. Table III shows that the WPB of our strategy is insensitive to the measurement count as well as number of devices in test structures.

VI. CONCLUSIONS

In this work, we have presented a novel approach for design-dependent process monitoring. Such process monitors are on wafer scribe-lines and can be tested after M1 fabrication. This allows for early die performance and wafer yield estimation

TABLE III
C432 WPB FOR DIFFERENT MEASUREMENT/TEST STRUCTURE SETUP.

N_e	N_d	WPT = 25%	WPT = 40%	WPT = 50%
1	1	94.18	92.98	95.23
5	10	94.26	95.69	97.97
100	100	94.26	95.69	97.97

dependent on the current process snapshot (as opposed to long term statistics). We use this for cutting short the production of obviously bad wafers, where the wafer yield is too low to cover manufacturing/test costs. The wafer pruning approach based on our method can achieve upto 98% of the maximum achievable benefit. The monitoring strategy is chosen so as to minimize information exchange between the design and the foundry as much as possible.

VII. ACKNOWLEDGMENTS

This work was supported in part by IMPACT UC Discovery Grant (http://impact.berkeley.edu) contract number ele07-10291, SRC and NSF CAREER Award number 0846196.

REFERENCES

[1] M. Bhushan, A. Gattiker, M. B. Ketchen and K.K. Das, "Ring Oscillators for CMOS Process Tuning and Variability Control," in IEEE Trans. on Semiconductor Manufacturing, vol. 1, no. 1, pp. 10-19, Feb. 2006.
[2] M.B. Ketchen, M. Bhushan and D. Pearson, "High Speed Test Structures for In-Line Process Monitoring and Model Calibration," in Intl. Conf. on Microelectronic Test Structures, pp.33-38, Apr. 2005.
[3] R. Lefferts and C. Jakubiec, "An Integrated Test Chip for the Complete Characterization and Monitoring of a 0.25um CMOS Technology that fits into five scribe line structures 150um by 5000 um," in ICMTS,Mar. 2003.
[4] A.J. Drake, R.M. Senger, H. Singh, G.D. Carpenter and N.K. James, "Dynamic Measurement of Critical-Path Timing," in IEEE ICICDT, pp. 249-252,June 2008.
[5] C. Visweswariah, K. Ravindran, K. Kalafala, S.G. Walker, S. Narayan, D.K. Beece et. al., "First-order Incremental Block-Based Statistical Timing Analysis," in IEEE trans. on Computer-Aided Design of Integrated Circuits and Systems, vol. 25, issue 10, pp. 2170-2180, Oct. 2006.
[6] S. Mitra, E. Volkerink, E.J. McCluskey and S. Eichenberger, "Delay Defect Screening using Process Monitoring Structures," in Proc. of VLSI Test Synposium, pp. 43-48, Apr. 2004.
[7] Q. Liu, S.S. Sapatnekar, "Synthesizing a Representative Critical Path for Post-Silicon Delay Prediction," in IEEE/ACM ISPED, Mar. 2009.
[8] F. Rigaud, J.M. Portal, H. Aziza, et. al. ,"Test Structure for Process and Product Evaluation," IEEE Intl. Conf. on Microelectronic Test Structures, Mar., 2007.
[9] C.Y. Cho, D.D. Kim, J.H. Kim D.Y. Lim and S.Y. Cho,"Early Prediction of Product Performance and Yield Via Technology Benchmark," in IEEE Custom Intergrated Circutis Conference, pp. 205-208, Sept., 2008.
[10] K.K. Das, S.G. Walker and M. Bhushan, "An Integrated CAD methodology for Evaluating Mosfet and Parasitic Extraction Models and Variabilitiy," in Proc. of IEEE, vol. 95, issue 3, pp.670-687, Mar., 2007.
[11] M.H. Na, E.J. Nowak, W. Haensch and J. Cai, "The Effective Drive Current in Cmos Inverters," in IEEE IEDM, pp. 121-124, 2002.
[12] K. v. Arnim, C. Pacha, K. Hofmann, T. Schulz, K. Schrfer and J. Berthold, "An Effective Switching Current Methodology to Predict the Performance of Complex Digital Circuits," in IEDM,Dec 2007.
[13] S.-J. Han, X. Yu, N. Zamdmer et. al., "Improved Effective Switching Current (I^{+}_{EFF}) and Capacitance Methodology for Cmos Circuit Performance Prediction and Model-to-Hardware Correlation," in IEEE IEDM, pp. 1-4, Dec., 2008.
[14] S. Reda, S. R. Nassif, "Analyzing the Impact of Process Variations on Parametric Measurements: Novel Models and Applications," in IEEE DATE, 2009.
[15] X. Li, R. Rurenbar and S. Blanton, "Virtual Probe: A Statistically Optimal Framework for Minimum-Cost Silicon Characterization of Nanoscale Integrated Circuits," in IEEE ICCAD, Nov. 2009.
[16] L. Cheng, P. Gupta, C. Spanos, K. Qian and L. He ,"Physically Justifiable Die-Level modeling of Spatial Variation in View of Systematic Across Wafer Variability," IEEE Design Automation Conference, pp. 104-108, July, 2009.
[17] Y. Cao, P. Gupta et. al. ,"Design Sensitivities to variability: Extrapolation and assessments in nanometer VLSI," IEEE ASIC/SoC Conference, Sept., 2002.
[18] W. Zhao, Y. Chao, F. Liu, K. Agarwal, D. Acharyya, S. Nassif and K. Nowka, "Rigorous extraction of process variations for 65nm CMOS design," in European Solid State Device Research Conf.,Sept., 2007.
[19] M. Ketchen, M. Bhushan and D. Pearson, "High Speed Test Structures for In-Line Process Monitoring and Model Calibration," in ICMTS, April 2005
[20] T.B. Chan, R.S. Ghaida and P. Gupta, "Electrical Modeling of Lithographic Imperfections," in Proc IEEE/ACM VLSI Design Conference,pp. 423-428, Jan. 2010
[21] Nangate Open Cell Library,https://www.nangate.com/ .
[22] R. Rao, A. Devgan, D. Blaauw and D. Sylvester, "Parametric Yield Estimation Considering Leakage Variability", Design Automation Conference, June 2004.