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 $I_{off}$-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 up to 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:

- $I_{th} = I_{ds}$ 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 $V_{gs} = V_{dd}$, $V_{d} = V_{d} = 0$

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 $N_c$ 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 $I_{l}, I_{th}, I_{off}$ 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

$$\text{delay} \propto \frac{CV}{I_{eff}}, \quad I_{eff} = \frac{I_{b} + I_{l}}{2}$$  (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 $I_{b}$ 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

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

$$d_{cell}(c) = d_{cell, 0}(c) - \sum_{t \in T} K_{cell}(c, t) CV \left( \frac{\Delta I_{eff}(t)}{I_{eff}(t)} - \frac{\Delta I_{eff}^2(t)}{2I_{eff}(t)} \right)$$  (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

$$d_{path}(j) = d_{path, 0}(j) + \Delta d_{path}(j)$$

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

2In this work, we consider design with four device types: {high $V_{th}$, low $V_{th}$} $\times$ {pmos, nmos}
where $d_{\text{path} \circ j}(j)$ refers to nominal delay of path $j$, $\Delta d_{\text{path}}(j)$ is the delay change due to process variation, which is equal to the sum of delay changes of every cell in the path,

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

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 $I_{\text{eff}}(t)$ can therefore be expressed as

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

(4)

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

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

C. Handling Load Capacitance Variation

In (4), the path specific delay sensitivities to $I_{\text{eff}}$ 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.

$$
d'_{\text{path}} (j) = \left( d_{\text{path}}(j) - d_{\text{path} \sim \text{interconnect}}(j) \right) \frac{C_{\text{gate}}}{C_{\text{nom}}}
$$

(6)

$$
+ \sum_{i \in G_j} K_{\text{cell}}(i, t) C(i)
$$

where $C_{\text{gate}}$ is process variation affected capacitance (measured by scribe-line monitors), $C_{\text{nom}}$ is its nominal value and $d_{\text{path} \sim \text{interconnect}}(j)$ 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 $I_{\text{eff}}$ 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 $K_{\text{path}}(j, t)$ form.

D. Effect of Within Die Variation on Delay

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

$$
D = \begin{bmatrix}
d'_{\text{path}}(1) \\
\vdots \\
d'_{\text{path}}(z)
\end{bmatrix} + W' I_{\text{wd}}, W = \begin{bmatrix}
w_{11} & \cdots & w_{1n} \\
\vdots & \ddots & \vdots \\
w_{zn} & \cdots & w_{nn}
\end{bmatrix},
$$

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

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

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

(7)

Since the path delays are correlated, we need to evaluate the covariance of critical paths, $(W' W)^T$. 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 $W$, 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 $R$ 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

$$
D = \begin{bmatrix}
d'_{\text{path}}(1) \\
\vdots \\
d'_{\text{path}}(z)
\end{bmatrix} + W' I_{\text{wd}} + R,
$$

(8)

$$
R^T = [r_{1, i_{\text{res}1}}, \ldots, r_{z, i_{\text{res}z}}],
$$

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

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

where $w_{kj}$ and $w'_{jx}$ are the entries of $W$ and $W'$, 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 $O(v^* \text{number of paths})$. Based on (8), the delay of a critical path can be written as,

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

where $y_j$ is $j^{th}$ row of $W^*$, $i_{\text{wd}}$ and $i_{\text{res}}$ 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 $I_{\text{eff}}$ for a chip is given as

$$\bar{I}_{\text{eff}} = \frac{1}{N_c} \sum_{m=1}^{N_c} \frac{I_{\text{eff}}(m)}{N_d} \quad (10)$$

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

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

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

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

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

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

$$\mu_{I_{\text{eff}}} = \bar{I}_{\text{eff}}$$

$$\sigma^2_{I_{\text{eff}}} = \frac{\bar{I}_{\text{eff}}^2 \sigma^2_{I_{\text{wd}}}}{N_c} + \frac{\sigma^2_{I_{\text{wd}}}}{N_d}$$

where $\sigma^2_{I_{\text{wd}}}$ and $\sigma^2_{I_{\text{eff}}}$ are the variance of within-die variation and measurement noise, respectively. Note that, the variance of $I_{\text{eff}}$ 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 ($N_c = 5$) and there are 10 devices in each test structure ($N_d = 10$). We assume $3\sigma$ of measurement noise to be 5% of nominal $I_{\text{eff}}$ value. $I_{\text{wd}}$ 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 $I_{\text{off}}$

#### A. Leakage Power Model

We model cell leakage power of an instance as linear function of $I_{\text{off}}$,

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

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,

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

where $N_c$ is the total number of instances of cell type $c$ in the design, $\Gamma$ is the set of all cell types and $I_{\text{off}}(i, t)$ 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 $I_{\text{off}}$ is modeled as an exponential function of variation sources.

$$I_{\text{off}}(i, t) = I_{\text{off}}(0) e^{Y(i, t)}$$

1In this paper, we only consider subthreshold leakage but the model can be easily extended to consider gate leakage.
where $I_{off}$ 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:

$$I_{off}^i(t) = I_{off} + Y_i(t) + r$$

(13)

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

$$P_{chip} = \sum_{t \in T} I_{off}(t)e^{Y_i(t)}\sum_{c \in C} a_c(t) \cdot N_c \cdot \mu_r(t)$$

(14)

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

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

where $\mu_r(t)$ is the mean of $e^{Y_i}$. 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 $Y$ to estimate total leakage power which is derived from measurements. As mentioned earlier, we take $N_c$ measurements of the current of $N_d$ devices in test structures. Considering measurement noise and within die variation, the $n^{th}$ measured $I_{off}$ is modeled as

$$I_{off}^n = \sum_{s=1}^{N_d} I_{off}(t)e^{Y_i+s}(1+Z_m)$$

(15)

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

$$Y = Y + \frac{1}{N_c} \sum_{m=1}^{N_c} ln(1+Z_m)$$

(16)

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

$$Y = Y - \frac{1}{N_c} \sum_{m=1}^{N_c} Z_m$$

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

$$\mu_Y = Y$$

(17)

$$\sigma_Y^2 = \sigma_Z^2/N_c$$

Since each $Y_i(t)$ is a Gaussian random variable, $e^{Y_i(t)}$ 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 lognormal random variable by matching the mean and variance. Note that, the uncertainties in $Y_i(t)$ 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 $e^{Y_i}$.

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

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

where $D_{spec}$ is the maximum allowed delay for a design, $\mu_{delay}$ and $\sigma_{delay}$ 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

$$Pr\{P_{chip} \leq P_{spec}\} = \Phi(\frac{ln(P_{spec}) - \mu_{L}}{\sigma_{L}}),$$

where $\mu_L$ and $\sigma_L$ is the mean and variance of $ln(P_{chip})$, respectively. Given the measured values of every chip ($I_{off}, I_{off}$ 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 $I_{off}$ and $I_{off}$ are different, the probability distributions of the estimations are independent. Thus, the passing probability of a chip is

$$Pr\{P_{chip} \leq P_{spec}\} = Pr\{P_{chip} \leq P_{spec}\} \cdot Pr\{chip \ delay \leq D_{spec}\}.$$

(18)

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

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

(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,

$$Additional \ Cost = (M_b + M_s),$$

Expected profit =Expected good chips $\times$ Chip price

(20)

If the final number of working chips is close to the expected number of good chips, it is profitable to continue processing
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 5.0X 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:

\[
D_{sys} = ax^2 + by^2 + cx + dy + exy,
\]

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.

### V. Experiments and Results

Figure 4 summarizes 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.

\(^5\)Many improved critical path selection algorithms have been proposed in literature. This is beyond the scope of our work.

\(^6\)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.T. TO THE IDEAL WAFER PRUNING SETUP.

<table>
<thead>
<tr>
<th>bench mark</th>
<th>WPT = 25%</th>
<th></th>
<th>WPT = 40%</th>
<th></th>
<th>WPT = 50%</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dep%</td>
<td>Indep%</td>
<td>Dep%</td>
<td>Indep%</td>
<td>Dep%</td>
<td>Indep%</td>
</tr>
<tr>
<td>c432</td>
<td>97.25</td>
<td>88.60</td>
<td>95.09</td>
<td>87.76</td>
<td>97.37</td>
<td>84.80</td>
</tr>
<tr>
<td>s15850</td>
<td>97.97</td>
<td>89.55</td>
<td>97.12</td>
<td>83.07</td>
<td>97.46</td>
<td>82.43</td>
</tr>
<tr>
<td>s38584</td>
<td>95.26</td>
<td>91.83</td>
<td>94.26</td>
<td>76.89</td>
<td>92.80</td>
<td>75.66</td>
</tr>
<tr>
<td>mips789</td>
<td>96.97</td>
<td>83.17</td>
<td>94.24</td>
<td>72.07</td>
<td>92.08</td>
<td>69.49</td>
</tr>
<tr>
<td>c432 (low Vth device only)</td>
<td>99.84</td>
<td>88.20</td>
<td>99.88</td>
<td>75.94</td>
<td>99.59</td>
<td>73.26</td>
</tr>
</tbody>
</table>

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.)

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 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 up to 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
  E T)

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

<table>
<thead>
<tr>
<th>N</th>
<th>N</th>
<th>WPT = 25%</th>
<th></th>
<th>WPT = 40%</th>
<th></th>
<th>WPT = 50%</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dep%</td>
<td>Indep%</td>
<td>Dep%</td>
<td>Indep%</td>
<td>Dep%</td>
<td>Indep%</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>94.18</td>
<td>92.98</td>
<td>95.26</td>
<td>93.59</td>
<td>96.97</td>
<td>97.97</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>94.26</td>
<td>95.69</td>
<td>99.26</td>
<td>95.69</td>
<td>99.26</td>
<td>99.26</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>94.26</td>
<td>95.69</td>
<td>99.26</td>
<td>95.69</td>
<td>99.26</td>
<td>99.26</td>
</tr>
</tbody>
</table>