A Model for Timing Errors in Processors with Parameter Variation *

Smruti R. Sarangi, **Brian Greskamp**, and **Josep Torrellas**
University of Illinois at Urbana-Champaign
{sarangi,greskamp,torrellas}@cs.uiuc.edu
http://iacoma.cs.uiuc.edu

Abstract

Parameter variation in integrated circuits causes sections of a chip to be slower than others. To prevent any resulting timing errors, designers have traditionally designed for the worst case. Unfortunately, this approach has the potential to nullify much of the upcoming gains of shrinking technologies.

To help understand this problem, we introduce a novel high-level and easy-to-apply model of how parameter variation affects timing errors in microprocessors. The model successfully predicts the probability of timing errors under different process and environmental conditions for both SRAM and logic units. Circuit designers can apply the model at design time to improve yield, and computer architects can use it to design processors that improve performance.

1 Introduction

As integration technology continues to scale relentlessly, designers of high-performance processors face the major challenge of parameter variation — the deviation of Process, Voltage, and Temperature (PVT) values from nominal specifications. It has been estimated that over the coming years, parameter variation may wipe out the performance gains of almost one full process generation. Therefore, it is necessary to design future circuits and computer architectures that can mitigate and tolerate the deleterious effect of variation.

Two broad classes of schemes to deal with parameter variation are circuit techniques and architecture techniques. Circuit techniques consist of schemes like adaptive body bias (ABB) and adaptive supply voltage (ASV). They mitigate variation by adjusting the body bias voltage and the supply voltage of different regions of the chip. Architecture techniques include a variety of schemes such as remapping slow memory rows to improve SRAM access time, and employing checker processors and related mechanisms [6] to tolerate variation-induced errors.

In this paper, we are interested in modeling how variation changes the latency of processor structures and, as a result, causes timing errors. The insights resulting from the model can be used to evaluate any of the schemes mentioned. To construct the model, we build on experimental latency measurements from a real processor pipeline [6] and analytical models of SRAM latency under random process variation [16]. We then develop a comprehensive model of timing errors due to variation, called VATS, for both logic and SRAM structures.

The key contributions of this paper are as follows:

Logic: We propose a model for the delay in logic paths that includes the effect of parameter (PVT) variation. It requires as input a histogram of nominal path delays, preferably obtained by a timing analysis tool. We outline a method to obtain the histogram from empirical measurements conducted by [6].

Memory: We extend the model proposed by [16], which only considered random process variation for V_t and used the Shockley model for transistor current. Our model adds the effects of systematic variation in V_t, as well as both random and systematic variation in L_{eff}. Additionally, we use the alpha-power current model [18], which is more representative of current and future technologies.

This paper is organized as follows. Section 2 introduces some background material; Section 3 presents the model of timing errors for logic and SRAM; Section 4 validates our model using empirical data obtained by [6, 11] and shows plots of error rates for different structures; and Section 5 presents related work.

2 Background

The parameters that we are interested in are the threshold voltage V_t and the effective channel length L_{eff}. They directly impact the delay and leakage power of a circuit. The parameter variation can be broken down into two major components: die-to-die (D2D) and within-die (WID). Moreover, WID variation can be broken up into random and system-

*This work was supported in part by the National Science Foundation under grants EIA-0072102, EIA-0103610, CHE-0121357, and CCR-0325603; DARPA under grant NBCH30390004; DOE under grant B347886; and gifts from IBM and Intel. Smruti R. Sarangi is now with Synopsys Research, Bangalore, India.

atic components. The former arises because of fluctuation in dopant density and lithographic phenomena like line edge roughness. The latter arises due to mask defects, lens aberrations, and sub-wavelength lithography. The variation Δ in any parameter, P (e.g., V_t or L_{eff}) can thus be represented as:

$$\Delta P = \Delta P_{D2D} + \Delta P_{WID} = \Delta P_{D2D} + \Delta P_{rand} + \Delta P_{syst}$$

We focus on WID variation, since the D2D component can be handled as an offset for the whole chip. We model the systematic component as a multivariate normal distribution [19] with a correlation matrix that is isotropic, position independent, and follows the Spherical model [4]. The random component is modeled as uncorrelated white Gaussian noise [19]. In our analysis, we assume that σ/μ for V_t variation is 9%. Based on the ITRS report [1], σ/μ for L_{eff} variation is roughly half of that value, or 4.5%. Moreover, we also assume that σ_{syst}/μ and σ_{rand}/μ are equal [12, 13]. Finally, based on [2], we assume that the systematic components of V_t and L_{eff} variations are perfectly correlated. In other words, for any given transistor, $\Delta V_{t\_syst} \propto \Delta L_{eff\_syst}$.

2.1 Gate delay

The delay of an inverter is given by the alpha-power model [18]:

$$T_g \propto \frac{L_{eff}V}{\mu(T)(V - V_t(T))^{\alpha}} \tag{1}$$

where α is typically 1.3 and μ is the mobility of carriers $\mu(T) \propto T^{-1.5}$. As V_t decreases, $V - V_t$ increases and the gate becomes faster. As T increases, $V - V_t(T)$ increases, but $\mu(T)$ decreases [10]. The second factor dominates and, with higher T, the gate becomes slower. The Shockley model occurs as a special case of the alpha-power model with $\alpha = 2$.

2.2 Transistor equations

The equations for transistor drain current I_d using the traditional Shockley model are as follows:

$$I_d = \begin{cases} 0 & \text{if } V_{gs} \le V_t \\ \beta(V_{gs} - V_t - \frac{V_{ds}}{2})V_{ds} & \text{if } V_{ds} < V_{gs} - V_t \\ \beta\frac{(V_{gs}-V_t)^2}{2} & \text{if } V_{ds} \ge V_{gs} - V_t \end{cases} \tag{2}$$

Here, $\beta = \mu C_{ox} W/L_{eff}$, where μ is the mobility and C_{ox} is the oxide capacitance.

In deep sub-micron technologies, these relationships are superseded by the alpha power law [18]:

$$I_d = \begin{cases} 0 & \text{if } V_{gs} \le V_t \\ \frac{W}{L_{eff}}\frac{P_c}{P_v}(V_{gs} - V_t)^{\alpha/2}V_{ds} & \text{if } V_{ds} < V_{d0} \\ \frac{W}{L_{eff}}\ P_c(V_{gs} - V_t)^{\alpha} & \text{if } V_{ds} \ge V_{d0} \end{cases} \tag{3}$$

In this equation, P_c and P_v are constants, and V_{d0} is given by:

$$V_{d0} = P_v(V_{gs} - V_t)^{\alpha/2}$$

2.3 Mathematical preliminaries

Single variable Taylor expansion
The Taylor expansion of a function $f(x)$ about x_0 is:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \tag{4}$$

where $f^{(n)}(x_0)$ is the n^{th} derivative of f at x_0.

μ, σ of a function of Gaussian random variables
Consider a function $Y = f(X_1, X_2, \ldots, X_n)$ of Gaussian random variables $X_1, \ldots, X_n$ with mean $\mu_1, \ldots, \mu_n$ and standard deviation $\sigma_1, \ldots, \sigma_n$. Multivariate Taylor series expansion [17] yields the mean and standard deviation of Y as follows:

$$\begin{aligned} \mu_y &= f(\mu_1 \ldots \mu_n) + \sum_{i=1}^{n}\left[\left.\frac{\partial^2 f(x_1 \ldots x_n)}{\partial(x_i)^2}\right|_{\mu_i} \times \frac{\sigma_i^2}{2}\right] \\ \sigma_y^2 &= \sum_{i=1}^{n}\left[\left(\left.\frac{\partial f(x_1 \ldots x_n)}{\partial(x_i)}\right|_{\mu_i}\right)^2 \times \sigma_i^2\right] \end{aligned} \tag{5}$$

Maximum of two independent Gaussian random variables
Let $Z = max(X, Y)$, where X and Y are independent Gaussian random variables with distributions $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$ respectively. Let:

$$\begin{aligned} \nu &= \sqrt{\sigma_1^2 + \sigma_2^2} \\ \eta &= (\mu_1 - \mu_2)/\nu \\ \varphi(x) &= (2\pi)^{-\frac{1}{2}} e^{-x^2/2} \\ \Phi(x) &= \int_{-\infty}^{x} \varphi(t)dt \end{aligned}$$

According to results in [3], Z can be approximated as a normal distribution with parameters:

$$\begin{aligned} \mu_z &= \mu_1\Phi(\eta) + \mu_2\Phi(-\eta) + \nu\varphi(\eta) \\ \sigma_z^2 &= (\mu_2^2 + \sigma_2^2)\Phi(-\eta)(\mu_1 + \mu_2)\nu\varphi(\eta) \\ &\quad + (\mu_1^2 + \sigma_1^2)\Phi(\eta) - \mu_z^2 \end{aligned} \tag{6}$$

Recursive application of Equation 6 yields the maximum of more than two Gaussian random variables.

3 Modeling timing errors

This section presents VATS, a novel model of variation-induced timing errors in processor pipelines. In the following, we first model errors in logic and then in SRAM memory.

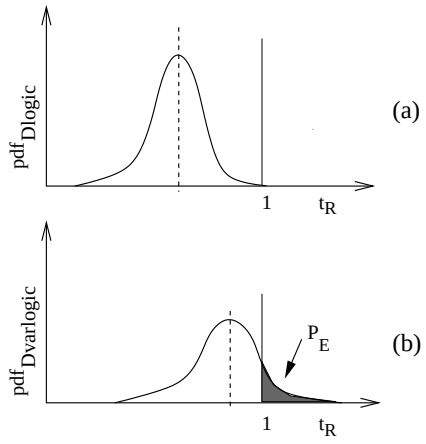

Figure 1. Example path delay distributions before (a) and after (b) variation, showing the introduction of timing errors.

3.1 General approach

A pipeline stage typically has a multitude of paths, each one with a different probability of being exercised on any given cycle. In our analysis, we make two simplifying assumptions.

Assumption 1: A path causes a timing fault iff it is exercised and its delay exceeds the clock period.

Assumption 2: The base clock period t_0 is set at the shortest time that admits error-free operation in the absence of process variation and at nominal temperature (85 °C).

In the following, path delay is normalized by expressing it as a fraction t_R of t_0. Our model begins with the probability density function (pdf) of the normalized path delays in the pipeline stage. Figure 1(a) shows an example pdf before variation effects. The right tail abuts the $X = 1$ abscissa and there are no timing errors.

As the pipeline stage paths suffer parameter variation, the pdf changes shape: the curve may change its average value and its spread (e.g., Figure 1(b)). All the paths that have become longer than 1 generate errors. Our model estimates the probability of error (P_E) as the area of the shaded region in the figure. Moreover, if we clock the processor with period t_R, the probability of error is:

$$P_E(t_R) = 1 - cdf(t_R)$$

In the event that race-through errors are also a concern, $cdf(t_h)$ gives the probability of violating the minimum hold time t_h. However, we will not consider hold-time violations in the rest of the paper.

3.2 Timing errors in logic

We start by considering a pipeline stage of only logic. We represent the logic path delay in the absence of variation as a random variable D_{logic}, which is distributed in a way similar to Figure 1(a). Such delay is composed of both wire and gate delay. For simplicity, we assume that wire accounts for a fixed fraction k_w of total delay. This assumption has been made elsewhere [9]. Consequently, we can write:

$$\begin{aligned} D_{logic} &= D_{wire} + D_{gates} \\ D_{wire} &= k_w \, D_{logic} \\ D_{gates} &= (1 - k_w) \, D_{logic} \end{aligned} \tag{7}$$

We now consider the effects of variation. Since variation typically has a very small effect on wires, we only consider the variation of D_{gates}, which has a random and a systematic component. For each path, we divide the systematic variation component (ΔD_{gates_sys}) into two terms: (i) the average value of it for all the paths *in the stage* ($\overline{\Delta D_{gates\_sys}}$) — which we call intra-stage systematic mean — and (ii) the rest ($\Delta D_{gates\_sys} - \overline{\Delta D_{gates\_sys}}$) — which we call intra-stage systematic variation.

Given the high degree of spatial correlation in parameter variation and the small size of a pipeline stage, the intra-stage systematic variation is small. For example, the distance at which the correlation between the V_t of two transistors becomes zero (i.e., the correlation range ϕ) has been measured to be 50% of the die width [7]. On the other hand, the length of a pipeline stage in a high-performance microprocessor is less than 10% of the width of a multicore die. Given that the stage dimensions are significantly smaller than ϕ, the transistors in a pipeline stage have highly-correlated V_t (and L_{eff}). Using Monte Carlo simulations with the parameters of Section 2, we find that the intra-stage systematic variation of $\overline{D_{gates}}$ has a $\sigma_{intrasys} \approx 0.004 \times \mu$, while the variation of $\overline{\Delta D_{gates\_sys}}$ across the pipeline stages of the processor has a $\sigma_{intersys} \approx 0.05 \times \mu$. Similarly, T varies much more across stages than within them.

The random component of D_{gates}'s variation is estimated from the fact that we model a path as n FO4 gates connected with short wires. Each gate's random component is independent. Consequently, for the whole n-gate path, D_{gates}'s σ_{rand} is $\sqrt{n} \times \sigma_{rand\_D_{FO4}}$, where D_{FO4} is the delay of one FO4 gate. If we take $n = 12$ as representative of high-end processors, the overall variation is small. It can be shown that D_{gates}'s $\sigma_{rand} \approx 0.01 \times \mu$. Finally, T has no random component.

We can now generate the distribution of D_{logic} with variation (which we call $D_{varlogic}$ and show in Figure 1(b)) as follows. We model the contribution of $\overline{\Delta D_{gates\_sys}}$ in the stage as a factor η that multiplies D_{gates}. This factor is the average increase in gate delay across all the paths in the stage due to systematic variation. Without variation, $\eta = 1$.

We model the contribution of the intra-stage systematic and of the random variations as D_{extra}, a small additive normal delay perturbation. Since D_{extra} combines D_{gates}'s intra-stage systematic and random effects, $\sigma_{extra} = \sqrt{\sigma^2_{intrasys} + \sigma^2_{rand}}$. For our parameters, $\sigma_{extra} \approx 0.011 \times \mu$. Like η, D_{extra} should multiply D_{gates} as shown in Equation 8. However, to simplify the computation and because D_{logic} is clustered at values close to one, we prefer to approximate D_{extra} as an additive term as in Equation 9:

$$D_{varlogic} = (\eta + D_{extra})\, D_{gates} + D_{wire} \quad (8)$$

$$\approx (1 - k_w)\,(\eta\, D_{logic} + D_{extra}) + k_w\, D_{logic} \quad (9)$$

After we compute the $D_{varlogic}$ distribution (shown in Figure 1(b)) we numerically integrate it to obtain $cdf_{D_{varlogic}}$. Then, the estimated error rate P_E of the stage cycling with a relative clock period t_R is:

$$P_E(t_R) = 1 - cdf_{D_{varlogic}}(t_R) \quad (10)$$

3.2.1 How to use the model

To apply Equation 9, we must calculate k_w, η, D_{extra}, and D_{logic} for the prevailing variation conditions. To do this, we produce a gridded spatial map of process variation using the model in Section 2 and superimpose it on a high-performance processor floorplan. For each pipeline stage, we compute η from the pipeline stage's *T* and the systematic L_{eff} and V_t maps (we neglect V variation). Moreover, by subtracting the resulting mean delay of the stage from the individual delays in the grid points inside the stage, we produce the intra-stage systematic variation. We combine this distribution with the effect of the random process variation to obtain the D_{extra} distribution. D_{extra} is assumed normal.

Ideally, we would obtain a per-stage k_w and D_{logic} through timing analysis of each stage. For our general evaluation, we assume that the LF adder in [8] is representative of processor logic stages, and set $k_w = 0.35$ [9]. Additionally, we derive $pdf_{D_{logic}}$ using experimental data from Ernst *et al.* [6]. They measure the error rate P_E of a multiplier unit as they reduce its supply voltage V. By reducing V, they lengthen path delays. Those paths with delays longer than the cycle time cause an error. Our aim is to find the $pdf_{D_{logic}}$ curve from their plot of $P_E(V)$ (a curve similar to that shown in Figure 2(a)).

Focusing on Equation 9, Ernst's experiment corresponds to an environment with no parameter variation, so $D_{extra} = 0$. Each V corresponds to a new average $\eta(V)$ and, therefore, a new $D_{varlogic}(V)$ distribution. We compute each $\eta(V)$ using the alpha-power model (Equation 1) as the ratio of gate delay at V and gate delay at the minimum voltage in [6] for which no errors were detected.

At a voltage V, the probability of error is equal to the probability of exercising a path with a delay longer than 1 clock cycle. Hence, $P_E(V) = P(D_{varlogic}(V) > 1)$. If we use Equation 9 and define $g(V) = 1/(k_w + \eta(V) \times (1 - k_w))$, we have $D_{varlogic}(V) = D_{logic}/g(V)$. Therefore:

$$\begin{aligned} P_E(V) &= P(D_{varlogic}(V) > 1) \\ &= P(D_{logic}/g(V) > 1) \\ &= P(D_{logic} > g(V)) \\ &= 1 - cdf_{D_{logic}}(g(V)) \end{aligned} \quad (11)$$

Letting $y = g(V)$, we have $cdf_{D_{logic}}(y) = 1 - P_E(V)$. Therefore, we can generate $cdf_{D_{logic}}$ numerically by taking successive values of V_i, measuring $P_E(V_i)$ from Figure 2(a), computing $y_i = g(V_i)$, and plotting (y_i,1-$P_E(V_i)$) — which is (y_i,$cdf_{D_{logic}}(y_i)$). After that, we smooth and numerically differentiate the resulting curve to find the sought function $pdf_{D_{logic}}$. Finally, we approximate the $pdf_{D_{logic}}$ curve with a normal distribution, which we find has $\mu = 0.849$ and $\sigma = 0.019$ (a curve similar to that shown in Figure 2(b)).

Strictly speaking, this $pdf_{D_{logic}}$ curve only applies to the circuit and conditions measured in [6]. To generate $pdf_{D_{logic}}$ for a different stage with a different technology and workload characteristics, one would need to use timing analysis tools on that particular stage. In practice, Section 4.1 shows empirical evidence that this method produces $pdf_{D_{logic}}$ curves that are usable under a range of conditions, not just those under which they were measured.

Finally, since D_{logic} and D_{extra} are normally distributed, $D_{varlogic}$ in Equation 9 is also normally distributed.

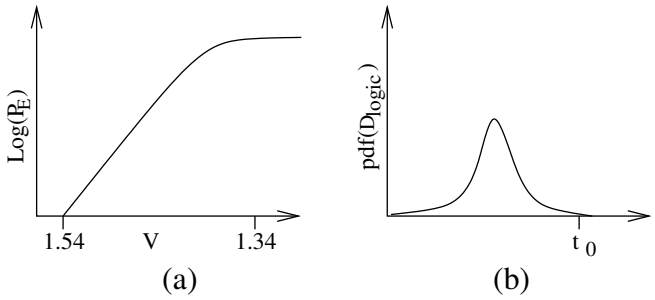

Figure 2. Error rate versus voltage curve from [6] (a) and corresponding $\mathbf{pdf_{D_{logic}}}$ (b).

3.3 Timing errors in SRAM memory

To model variation-induced timing errors in SRAM memory, we build on the work of Mukhopadhyay *et al.* [16]. They consider *random* V_t variation only and describe four failures in the SRAM cell of Figure 3: Read failure, where the contents of a cell are destroyed when the cell is read; Write failure, where a write is unable to flip the cell; Hold failure, where a cell loses its state; and Access failure, where the time needed to access

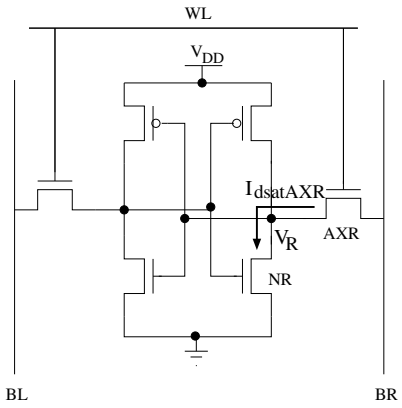

Figure 3. A read from a 6T SRAM cell, pulling the right bitline low.

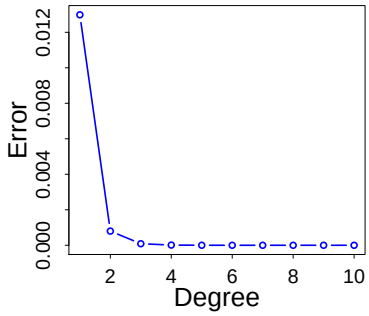

Figure 4. Error versus degree of expansion of z.

the cell is too long, leading to failure. The authors provide analytical equations for these failure rates, which show that for the standard deviations of V_t considered here, Access failures dominate and the rest are negligible.

Because Access failures are the dominant errors and have no clear remedy, they are our focus. In our analysis, we consider the effects of both systematic and random variation in both V_t and L_{eff}. Moreover, we use the alpha-power current model.

According to [16], the cell access time under variation on a read is:

$$\begin{aligned} T_{varacc} &\propto \frac{1}{I_{dsatAXR}} \\ &= h(V_{tAXR}, V_{tNR}, L_{AXR}, L_{NR}) \end{aligned} \tag{12}$$

where V_{tAXR} and L_{AXR} are the V_t and L_{eff} of the AXR access transistor in Figure 3, and V_{tNR} and L_{NR} are the same parameters for the NR pull-down transistor in Figure 3. We now discuss the form of this function h. We first briefly discuss the model of [16]. We then introduce our extension that uses the alpha-power model.

3.3.1 $\mathbf{I_{dsatAXR}}$ using the Shockley model

The model in [16] uses the traditional Shockley long channel transistor equations. Consider the case illustrated in Figure 3: a read operation where the bitline BR is being driven low. Transistor AXR is in saturation and transistor NR is in the linear range. Equating the currents using Kirchoff's current law:

$$\begin{aligned} I_{dsatAXR} &= \frac{K_1}{L_{AXR}}(V_{DD} - V_R - V_{tAXR})^2 \\ &= \frac{K_2}{L_{NR}}(V_{DD} - V_{tNR} - 0.5V_R)V_R \end{aligned} \tag{13}$$

In the Shockley model (Equation 2) we have replaced β with K/L_{eff}, where K is a constant and L_{eff} is the effective length of the respective transistor. Equation 13 is a quadratic equation in V_R. We can thus find I_{dsat} and subsequently the function h.

3.3.2 $\mathbf{I_{dsatAXR}}$ using the alpha-power model

We now use the more accurate alpha power law [18] to find $I_{dsatAXR}$. By equating currents as in Equation 13, we have:

$$\begin{aligned} I_{dsatAXR} &= \frac{K_1}{L_{AXR}}(V_{DD} - V_R - V_{tAXR})^{\alpha} \\ &= \frac{K_2}{L_{NR}}(V_{DD} - V_{tNR})^{\alpha/2}V_R \end{aligned} \tag{14}$$

As in Equation 13, constants have been folded into K_1 and K_2. To solve for V_R, perform the following transformation:

$$\begin{aligned} (V_{DD} - V_R - V_{tAXR})^{\alpha} = (V_{DD} - V_{tAXR})^{\alpha} \times \\ \left(1 - \frac{V_R}{V_{DD} - V_{tAXR}}\right)^{\alpha} \end{aligned} \tag{15}$$

Let $z = \frac{V_R}{V_{DD} - V_{tAXR}}$ and expand $(1-z)^{\alpha}$ using the Taylor series (Equation 4). Typical values of z are near 0.25, so we compute the expansion about that point. Figure 4 plots the error versus the degree of the expansion. Depending on the accuracy desired, we can choose the appropriate number of terms, but for most practical purposes, a degree of 2 is sufficient, making Equation 14 a quadratic equation in V_R:

$$(1-z)^{\alpha} \approx 1 - \alpha z + \alpha(\alpha - 1)\frac{z^2}{2}$$

Now, we can easily solve for V_R and find a closed form analytic expression for $I_{dsatAXR}$.

3.3.3 Error rate under process variation

We now have an analytic expression for the access time T_{varacc} using Equation 12. It is a function of four variables: V_{tAXR}, V_{tNR}, L_{AXR}, and L_{NR}. A six transistor memory cell is very small compared to the correlation range ϕ of V_t (Section 3.2). Therefore, we assume that the systematic component of variation is the same for all the transistors and even for the whole memory line. Now, using multivariate Taylor expansion (Equation 5), the mean $\mu_{Tvaracc}$ and standard deviation $\sigma_{Tvaracc}$ of T_{varacc} can be expressed as a function of the μ and σ of each of these four variables.

In reality, an SRAM array access does not read only one cell at a time but a line — e.g., 8-1024 cells. Consequently, we need to compute the distribution of the maximum access time of all the cells in a line. There is no exact analytical solution for the distribution of the maximum of n normally distributed variables, but we can use a normal approximation as shown in Equation 6. The resulting distribution has mean $\mu_{vararray}$ and standard deviation $\sigma_{vararray}$.

Finally, the access to the memory array itself takes only a fraction k of the whole pipeline cycle — the rest is taken by logic structures such as sense amplifiers, decoders, and comparators. Such logic delays are modeled according to Section 3.2. Consequently, the total path delay with variation D_{varmem} is the sum of the normal distributions of the delays in the line access and in the logic. It is distributed normally with:

$$\begin{aligned} \mu_{varmem} &= k\,\mu_{vararray} + (1-k)\,\mu_{varlogic} \\ \sigma_{varmem} &= \sqrt{k^2\,\sigma^2_{vararray} + (1-k)^2\,\sigma^2_{varlogic}} \end{aligned}$$

Then, the estimated error rate of a memory stage cycling with a relative clock period t_R is:

$$P_E(t_R) = 1 - cdf_{D_{varmem}}(t_R) \qquad (16)$$

3.3.4 Comparing the Shockley and alpha-power models

In Figure 5, we plot the mean access time ($\mu_{Tvaracc}$) for the Shockley model (dotted line) and for the alpha-power model (solid line). Access times are normalized to the one given by the Shockley model at 85 °C. From the figure, we see that the mean access time differs significantly for the two values of α. More importantly, it can be shown that $\sigma_{Tvaracc}$ is around 3.5% of the mean for the Shockley model and around 2% of the mean for the alpha-power model. Consequently, with decreasing α, the mean and standard deviation of the access time decrease.

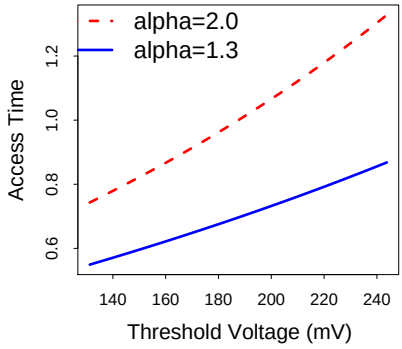

Figure 5. Relative mean access time ($\mu_{\mathbf{Tvaracc}}$) for α equal to 1.3 and 2.0. The latter corresponds to the Shockley model.

4 Evaluation

4.1 Empirical validation

To partially validate the VATS model, we use it to explain some error rate data obtained empirically elsewhere. We validate both the logic and the memory model components. For the former, we use the curves obtained by Das *et al.* [5], who reduce the supply voltage V of the logic units in an Alpha-like pipeline and measure the error rate in errors per cycle. They report curves for three different T: 45 °C, 65 °C, and 95 °C. Their curves are shown in solid pattern in Figure 6.

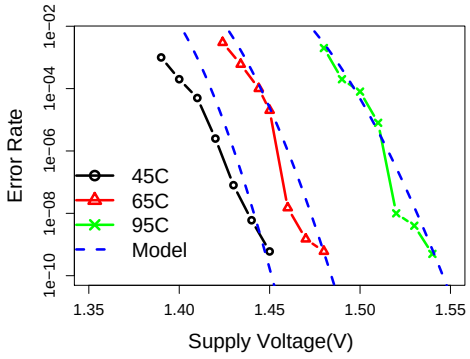

Figure 6. Validating the logic model by comparing the measured and predicted number of errors per cycle.

To validate our model, we use the 65 °C curve to predict the other two curves. We first determine D_{logic} from the 65 °C

curve through the procedure of Section 3.2.1. Recall that we generate the $pdf_{D_{logic}}$ numerically and then fit a normal distribution. We then use D_{logic} to predict the 95 °C and 45 °C curves as follows. We generate a large number of V_i values. For each V_i, we compute $\eta(V_i)$ as discussed in Section 3.2.1. Since there is no process variation, D_{extra} is zero. Knowing the D_{logic} distribution, we use Equation 9 for each $\eta(V_i)$ to compute the $D_{varlogic}(V_i)$ distribution. Finally, we plot the $(V_i, P_E(V_i))$ pairs from the model as dashed lines in Figure 6 along with the measured values (solid lines).

From the figure, we see that the predicted curves track the experimental data closely. The disagreement between the two comes largely from the normal approximation of D_{logic}, which is assumed for simplicity.

To validate the memory model, we use experimental data from Karl *et al.* [11]. They examine a 64KB SRAM with 32-bit lines comprising four different-latency banks, and measure the error rate as the supply voltage V changes. Since the SRAM is physically small, we assume that each cell in the array has the same value of the systematic process variation. Using the measured $P_E(V)$ for each bank, we find $D_{varmem}(t_R)$ using the method of Section 3.3 and fit a normal approximation. The original data is shown in solid pattern in Figure 7, and the prediction is displayed as a dashed line. From the figure, we see that the predicted and measured error rate are close.

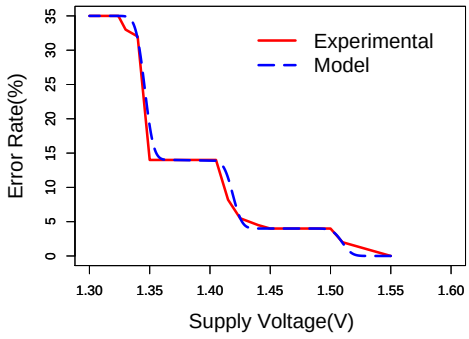

Figure 7. Validating the memory model by comparing the measured and predicted fraction of accesses that fail.

4.2 Example error curves

As one example of the uses of our model, we apply it to estimate the error rate of the logic and memory units of an AMD Opteron processor as we increase the frequency. After generating a V_t and L_{eff} variation map according to our variation model, we apply the timing error model to compute the error rate versus frequency for each pipeline stage. Figure 8 shows the results, where the frequency is normalized to the one that the processor without process variation can deliver.

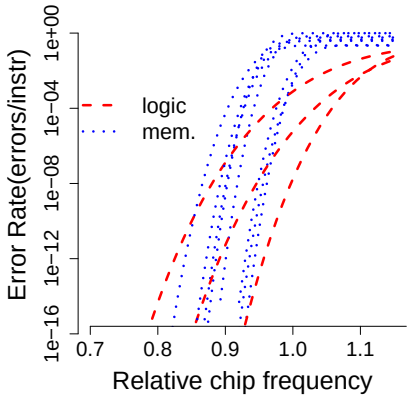

Figure 8. Estimated error rates of memory and logic pipeline stages in the AMD Opteron.

In the figure, each line corresponds to one pipeline stage. We see that memory stages have steeper error curves than the logic ones. This is because the paths in a memory stage are more homogeneous. We envision a situation where architects and circuit designers will use such error curves to design high performance or low power processors that can tolerate timing errors.

4.3 Tradeoffs in the model

Perhaps the main shortcoming of VATS is the loss of precision due to two main simplifying assumptions: (1) the use of normal approximations and (2) the assumption that wire delay is not affected by variation and accounts for a fixed fraction k_w of logic delay. The preceding section has argued that the loss of accuracy is small in practice. For logic circuits, better accuracy is possible by not using assumption (2). However, the approximations in VATS make it easy to apply it in the early stages of design, when architects must estimate variation effects at a high level.

5 Related work

Mukhopadhyay *et al.* [16] propose models for timing errors in SRAM memory due to random V_t variation. They consider several failure modes. As part of the VATS model, we extend their model of Access time errors by (i) also including systematic variation effects, (ii) also considering variation in L_{eff}, (iii) modeling the maximum access time of a *line* of SRAM rather than a single cell, and (iv) using the alpha-power model that uses an α equal to 1.3.

Memik *et al.* [14, 15] model errors in SRAM memory due to cross-talk noise as they overclock circuits. They use high degrees of overclocking — twice the nominal frequency and more. In the less than 30% overclocking regime that we consider, such cross-talk errors are negligible. For very small feature-size technologies, however, the situation may change.

Ernst et al. [6] and Karl et al. [11] measure the error rate of a multiplier and an SRAM circuit respectively by reducing the voltage beyond safe limits to save power. They plot curves for error rate versus voltage. In this paper, we outlined a procedure to extract the distribution of path delays from these curves, and validated parts of our model by comparing it against their curves.

6 Conclusions

We have presented a comprehensive timing-error model for both logic and memory in chips with parameter variation. For the logic, we provide formulas that incorporate results from timing analysis tools; for the memory, we extend models from other work. We validate our results with empirical data and find close agreement. We intend for the model to be used in evaluating high performance or low power processors that tolerate timing errors.

References

[1] International Technology Roadmap for Semiconductors (1999).

[2] Y. Cao and L. Clark. Mapping statistical process variation toward circuit performance variability: An analytical approach. In *Design Automation Conference (DAC)*, pages 658–663, 2005.

[3] C. E. Clark. The greatest of a finite set of random variables. *Operations Research*, 9:85–91, 1961.

[4] N. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 1993.

[5] S. Das, S. Pant, D. Roberts, S. Lee, D. Blaauw, T. Austin, T. Mudge, and K. Flautner. A self-tuning DVS processor using delay-error detection and correction. In *IEEE Symposium on VLSI Circuits*, June 2005.

[6] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Zeisler, D. Blaauw, T. Austin, K. Flautner, and T. Mudge. Razor: A low-power pipeline based on circuit-level timing speculation. In *International Symposium on Microarchitecture (MICRO)*, pages 7–18, 2003.

[7] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos. Modeling within-die spatial correlation effects for process-design co-optimization. In *International Symposium on Quality Electronic Design (ISQED)*, 2005.

[8] Z. Huang and M. D. Ercegovac. Effect of wire delay on the design of prefix adders in deep-submicron technology. In *Asilomar Conference on Signals and Systems*, 2000.

[9] E. Humenay, D. Tarjan, and K. Skadron. Impact of parameter variations on multicore chips. In *Workshop on Architectural Support for Gigascale Integration (ASGI)*, 2006.

[10] K. Kanda, K. Nose, H. Kawaguchi, and T. Sakurai. Design impact of positive temperature dependence on drain current in sub-1-V CMOS VLSIs. *IEEE Journal of Solid-State Circuits (JSSC)*, 36(10):1559–1564, 2001.

[11] E. Karl, D. Sylvester, and D. Blaauw. Timing error correction techniques for voltage-scalable on-chip memories. In *International Symposium on Circuits and Systems (ISCAS)*, 2005.

[12] T. Karnik, S. Borkar, and V. De. Probabilistic and variation-tolerant design: Key to continued moore's law. In *TAU Workshop*, 2004.

[13] X. Liang and D. Brooks. Latency adaptation of multiported register files to mitigate variations. In *Workshop on Architectural Support for Gigascale Integration (ASGI)*, 2006.

[14] A. Mallik and G. Memik. A case for clumsy packet processors. In *International Symposium on Microarchitecture (MICRO)*, pages 147–156, 2004.

[15] G. Memik, M. H. Chowdhury, A. Mallik, and Y. I. Ismail. Engineering over-clocking: Reliability-performance trade-offs for high-performance register files. In *International Conference on Dependable Systems and Networks (DSN)*, pages 770–779, 2005.

[16] S. Mukhopadhyay, H. Mahmoodi, and K. Roy. Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS. *IEEE Transactions on Computer Aided Design (TCAD)*, 24(12):1859–1880, Dec 2005.

[17] A. Papoulis. *Probability, Random Variables and Stochastic Processes*. McGrawHill, 2002.

[18] T. Sakurai and R. Newton. Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas. *IEEE Journal of Solid-State Circuits (JSSC)*, 25(2):584–594, 1990.

[19] A. Srivastava, D. Sylvester, and D. Blaauw. *Statistical Analysis and Optimization for VLSI: Timing and Power*. Springer, 2005.