Analysis and Mitigation of Variability in Subthreshold Design

Bo Zhai, Scott Hanson, David Blaauw, Dennis Sylvester
{bzhai, hansons, blaauw, dennis}@umich.edu, University of Michigan, Ann Arbor, MI

ABSTRACT

Subthreshold circuit design is a compelling method for ultra-low power applications. However, subthreshold designs show dramatically increased sensitivity to process variations due to the exponential relationship of subthreshold drive current with V_{th} variation. In this paper, we present an analysis of subthreshold energy efficiency considering process variation, and propose methods to mitigate its impact. We show that, unlike superthreshold circuits, random dopant fluctuation is the dominant component of variation in subthreshold operation. We investigate how this variability can be ameliorated with proper circuit sizing and choice of circuit logic depth. We then present a statistical analysis of the energy efficiency of subthreshold circuits considering process variations. We show that the energy optimal supply voltage increases due to process variations and study its dependence on circuit parameters. We verify our analytical models against Monte Carlo SPICE simulations and show that they accurately predict the minimum energy and energy optimal supply voltage. Finally, we use the developed statistical energy model to determine the optimal pipelining depth in subthreshold designs.

Categories and Subject Descriptors

B.7.1 [Types and Design Styles]: Advanced technologies

General Terms Design, Algorithms, Reliability, Performance

Keywords max of lognormal RVs, subthreshold variability

1 Introduction

In subthreshold circuit design the supply voltage is less than the threshold voltage, allowing for ultra low power circuit operation. A number of successful subthreshold designs have been presented in the literature [1][2]. Using subthreshold design, it is expected that energy efficiency in the range of 1pJ / instruction can be achieved [3], hence enabling low performance applications powered by energy scavenging. In addition, wide range dynamic voltage scaling has been proposed [4] where processors can scale from high performance superthreshold operation to ultra low power subthreshold operation depending on workload.

In previous work [4][5], a minimum energy voltage (V_{min}) for CMOS subthreshold operation was demonstrated. Scaling the voltage supply below V_{min} ceases to reduce energy per operation due to the dominance of leakage in this voltage regime, combined with the exponential increase of circuit delay with supply voltage [4][5]. However, the proposed analyses do not account for the impact of process variation. It is well known that subthreshold designs have dramatically increased sensitivity to process variations since drive current is exponentially dependent on threshold voltage [2]. We observe that variations in gate delay can be as high as 300% from nominal, creating a significant challenge for subthreshold circuit design. It is therefore difficult to meet design specification predictably without dramatic overdesign which wastes energy efficiency. In this paper, we therefore analyze the impact of process variation on subthreshold design and propose methods to mitigate its effect.

We first analyze the impact of different sources of process variations on subthreshold circuit delay. We show that random dopant fluctuations (RDF) [6] become the dominant source of variation in subthreshold operation, in contrast to superthreshold operation where geometric variations (e.g., in L_{eff}) are equally important. Due to the independent nature of RDF variations it is possible to reduce their impact on circuit performance through averaging. Hence, we show how careful circuit sizing and choice of logic depth can reduce timing variability ($3\sigma/\mu$) to below 30% with appropriate design choices. We then analyze the energy efficiency of subthreshold designs while capturing the impact of process variations. We derive statistical expressions of circuit delay and static and dynamic power consumption and propose both analytical and numerically-derived expressions for the minimum energy and V_{min} as a function of circuit parameters. We show that the method in [4], which ignores process variations, can underestimate V_{min} by as much as 78mV for small devices, corresponding to a 40% underestimation.

Using the newly developed model, we then study the dependence of the minimum energy and V_{min} on design parameters such as the circuit logic depth, the number of critical paths in the circuit and the switching activity rate. Finally, we apply our model to a pipeline depth study. We show that the energy optimal pipeline depth in subthreshold designs increases from 10 fanout-of-four inverter (FO4) delays under nominal process conditions to 15 FO4 delays when process variations are considered.

The rest of the paper is organized as follows. Section 2 presents key observations in subthreshold circuit variability. In Section 3, we derive statistical models of circuit delay and power under process variations and use these to derive our analytical model of the minimum energy and V_{min}. In Section 4, we verify our analytical model against Monte Carlo SPICE simulations and examine some trends to provide useful insights on how to design subthreshold circuits efficiently. Section 5 explores energy efficiency from the perspective of pipelining, which includes process variation impact and latch overhead. Finally, Section 6 concludes the paper.

2 Variability Impact on Subthreshold Circuits

It is well known that process variability impact is magnified in subthreshold operation due to the exponential impact of V_{th} and L_{eff} on subthreshold drive current. However, little analysis has been performed to investigate the dominant components of variability in subthreshold circuits and other key trends. In this section we make several key observations about subthreshold circuit robustness based on SPICE simulations using an industrial 130nm technology. First, we point out that random dopant fluctuations (RDF) dominate geometric variations, particularly in channel length. This occurs since the channel length variation dependency of V_{th} stems from drain induced barrier lowering (DIBL), which reduces at low operating voltages. As a result, the magnitude of V_{th} variation arising from channel length uncertainty rapidly falls off as V_{dd} reduces. However, since on current (I_{on}) at low voltages becomes more sensitive to V_{th} fluctuations (exponentially dependent in subthreshold), the net result is that I_{on} variation due to DIBL remains roughly constant or slightly increases. On the other hand, the uncertainty in V_{th} due to RDF is independent of V_{dd} and solely a function of channel area [7]. Therefore, I_{on} variation resulting from RDF becomes the dominant component as V_{dd} nears V_{th} as shown in Figure 1.

ISLPED'05, August 8–10, 2005, San Diego, California, USA.
Copyright 2005 ACM 1-59593-137-6/05/0008...$5.00.

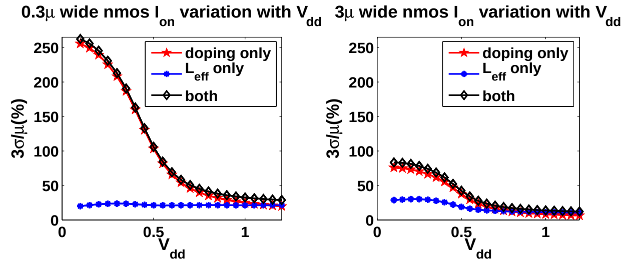

Figure 1. 3σ/μ of I_{on} due to different variation sources over a wide range of V_{dd}, showing the dominance of RDF in subthreshold operation.

We also observe that I_{on} variation due to RDF continues to increase even in the subthreshold region as seen in Figure 1. As EQ1 [8][10] shows, given a fixed amount of RDF V_{th} variation we expect a constant amount of I_{on} variation in the subthreshold region regardless of V_{dd}. In practice, however I_{on} uncertainty grows since subthreshold swing (S_S) improves in the subthreshold region as V_{dd} reduces.

$$I_{subvth} \propto \exp\left(\frac{V_{gs} - V_{th}}{mV_T}\right),\ m \propto S_S,\ 3\sigma_{I_{on}}/\mu_{I_{on}} = 3\sqrt{e^{\left(\frac{\sigma_{V_{th}}}{mV_T}\right)^2} - 1} \quad \text{(EQ 1)}$$

Considering that RDF dominates uncertainty in subthreshold circuits, we can address variability in this case through device sizing which reduces RDF. Furthermore, larger logic depths can serve to average out timing variations since stage delays are effectively independent. Figure 2 shows the 3σ/μ delay variation of an inverter chain versus the number of inverters (*n*) and inverter size (*W*) with Monte Carlo SPICE simulations. Interconnect loading for each stage is modeled by a lumped capacitance (50fF). As *W* or *n* increases, the relative variation becomes smaller, as expected. Figure 2 shows that by using sufficient logic depth and transistor sizing variability can be reduced to as little as 30%. In addition to selecting an appropriate logic depth, latch-based design (opposed to edge-triggered flip-flops) can enable time borrowing which gives more room to average out RDF variations, effectively increasing *n*. The impact of logic depth is further investigated in Section 5 which studies the optimal pipeline depth for energy efficiency in subthreshold circuits.

3 Subthreshold Statistical Analysis

In order to estimate the energy consumption under process variation, we need to statistically model both delay and power. In order to make the problem tractable, we choose to set up our target circuit with *p* identical inverter chains, each composed of *n* inverters. However, the analysis can be extended to more general gates as well. V_{th} typically follows a normal distribution; from EQ1 subthreshold on-current and propagation delay therefore exhibit lognormal distributions. Section 3.1 focuses on statistical subthreshold delay modeling. We first estimate the sum of lognormal gate delays to obtain the path delay (Section 3.1.1) and then find the circuit delay by taking the maximum of path delays (Section 3.1.2). This section also includes our mathematical formulation of the greatest of lognormal random variables (RVs). Section 3.2 details the power/energy and V_{min} analysis under process variation based on delay models in Section 3.1.

3.1 Subthreshold Propagation Delay Analysis

3.1.1 Subthreshold Delay Formulation

Let t_{di} be the delay of the i^{th} (*i=1,2,...p*) path. In this case, the final circuit delay t_{dm} can be expressed as

$$t_{dm} = max(t_{d1}, t_{d2}, \ldots t_{dp}) \quad \text{(EQ 2)}$$

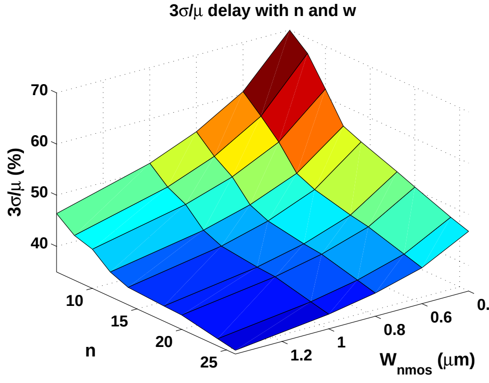

Figure 2. 3σ/μ of delay for an inverter chain with logic depth (*n*) and device sizing (*W*)

Using $t_{inv,j}$ as the j^{th} gate delay in a certain path, t_{di} can be further written as

$$t_{di} = \sum_{j=1}^{n} t_{inv,j} = \sum_{j=1}^{n}\left(\eta\frac{\frac{1}{2}C_s V_{dd}}{I_{on,j}}\right) = \frac{1}{2}\eta C_s V_{dd}\sum_{j=1}^{n}\frac{1}{I_{on,j}} \quad \text{(EQ 3)}$$

where η is the delay factor arising from a non-step actual input [4], and C_S is the switching load capacitance of each gate. Subthreshold current can be expressed as [8]

$$I_{sub} = \mu_{eff} C_{ox} \frac{W}{L_{eff}}(m-1)V_T^2 e^{\frac{V_{gs} - V_{th} - V_{off}}{mV_T}}\left(1 - e^{-\frac{V_{ds}}{V_T}}\right) \quad \text{(EQ 4)}$$

Let I_{s0} be $I_{s0} = \mu_{eff} C_{ox} \frac{W}{L_{eff}}(m-1)V_T^2 e^{-\frac{V_{off}}{mV_T}}$ (EQ 5)

then the on and off current can be rewritten as

$$I_{on} \approx I_{s0} e^{\frac{V_{dd}}{mV_T} - \frac{V_{th}}{mV_T}},\qquad I_{leak} \approx I_{s0} e^{-\frac{V_{th}}{mV_T}} \quad \text{(EQ 6)}$$

V_{th} has a normal distribution, hence $I_{on,j}$ has a lognormal distribution. Here we consider I_{s0} as deterministic assuming that L_{eff} variation in the denominator can be expressed in V_{th}. Substituting I_{on} into EQ3, we obtain

$$t_{di} = \frac{1}{2}\eta C_S V_{dd}\frac{1}{I_{s0}} e^{-\frac{V_{dd}}{mV_T}} \sum_{j=1}^{n} e^{\frac{V_{th,j}}{mV_T}} \quad \text{(EQ 7)}$$

We can see that t_{di} is the sum of several lognormally distributed RVs. From [9], the sum of several lognormal RVs can be approximated by another lognormal RV. We choose to match the first and second moment of the LHS and RHS in EQ7 as suggested in [9]. After some derivation, we arrive at

$$\mu(\ln t_{di}) = \ln\left(\frac{1}{2}\eta C_s V_{dd}\frac{1}{I_{s0}} e^{-\frac{V_{dd}}{mV_T}}\right) + \mu_{Vth} + \frac{1}{2}t^2 + \frac{1}{2}\ln\frac{n^3}{n - 1 + e^{t^2}} \quad \text{(EQ 8)}$$

$$\sigma(\ln t_{di}) = \sqrt{\ln\left(1 + \frac{1}{n}\left(e^{t^2} - 1\right)\right)} \quad \text{(EQ 9)}$$

where $t = \sigma_{V_{th}}/(mV_T)$.

This derivation finds the corresponding normal RVs (ln t_{di}) mean and standard deviation instead of $\mu(t_{di})$ and $\sigma(t_{di})$. We will need this information to find the max of t_{di}. In the next section we describe a method to estimate the mean and standard deviation of t_{dm} from t_{di}.

3.1.2 Greatest of t_{di} using lognormal RVs

We now seek the mean and standard deviation of the greatest of p lognormal RVs. Introducing notation, suppose u has a normal distribution with mean and standard deviation of μ_0 and σ_0. Then $X=exp(u)$ has a lognormal distribution. The mean and standard deviation of X are [10]

$$\mu(X) = e^{\mu_0 + \frac{1}{2}\sigma_0^2}, \quad \sigma(X) = e^{\mu_0 + \frac{1}{2}\sigma_0^2}\sqrt{e^{\sigma_0^2} - 1} \qquad \text{(EQ 10)}$$

We can easily solve for μ_0 and σ_0 from the above equation:

$$\begin{aligned} M(X) \equiv \mu_0 &= \frac{1}{2}\ln\left(\frac{\mu^4(X)}{\mu^2(X) + \sigma^2(X)}\right) \\ S(X) \equiv \sigma_0 &= \sqrt{\ln\left(\frac{\mu^2(X) + \sigma^2(X)}{\mu^2(X)}\right)} \end{aligned} \qquad \text{(EQ 11)}$$

Let X_Z be the max of X_A and X_B

$$X_Z = max(X_A, X_B) \qquad \text{(EQ 12)}$$

where X_A and X_B have lognormal distributions

$$X_A = e^{u_A}, \quad X_B = e^{u_B} \qquad \text{(EQ 13)}$$

Let μ_A and μ_B be the mean of u_A and u_B, σ_A and σ_B be the standard deviation of u_A and u_B. Then the pdf's of X_A and X_B are [10]

$$f_A(x) = \frac{1}{\sqrt{2\pi}x\sigma_A}e^{-\frac{(\ln x - \mu_A)^2}{2\sigma_A{}^2}}, \quad f_B(x) = \frac{1}{\sqrt{2\pi}x\sigma_B}e^{-\frac{(\ln x - \mu_B)^2}{2\sigma_B{}^2}}.$$

The probability that X_Z is smaller than x is

$$\begin{aligned} P(X_Z < x) &= P((X_A < x) \cap (X_B < x)) \\ &= P(X_A < x) \cdot P(X_B < x) \end{aligned} \qquad \text{(EQ 14)}$$

Since we are primarily concerned with delay variation due to RDF, we can assume that X_A and X_B are independent RVs, and the pdf of X_Z is

$$f_Z(x) = f_A(x)\left[\int_{-\infty}^{x} f_B(x)dx\right] + \left[\int_{-\infty}^{x} f_A(x)dx\right]f_B(x) \qquad \text{(EQ 15)}$$

Using $\mu_{z,k}$' to denote the k^{th} raw moment [10] of X_Z, we can finally obtain the following expression for $\mu_{z,k}$' after some derivation [see Appendix A]:

$$\begin{aligned} \mu_{z,k}' &= e^{k \cdot \mu_A + \frac{1}{2} \cdot k^2 \cdot \sigma_A^2}\Phi\left(\frac{\mu_A - \mu_B + k \cdot \sigma_A^2}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right) \\ &\quad + e^{k \cdot \mu_B + \frac{1}{2} \cdot k^2 \cdot \sigma_B^2}\Phi\left(\frac{\mu_B - \mu_A + k \cdot \sigma_B^2}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right) \end{aligned} \qquad \text{(EQ 16)}$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2}du$, $\sigma_A = S(X_A)$, $\sigma_B = S(X_B)$.

From the raw moments, the μ and σ of X_Z are given by

$$\mu(X_Z) = \mu_{z,1}', \quad \sigma(X_Z) = \sqrt{\mu_{z,2}' - (\mu_{z,1}')^2} \qquad \text{(EQ 17)}$$

There are usually many more than two critical paths in a well-designed circuit, thus we need to be able to estimate the max of an arbitrary number of lognormal RVs. One method is to apply the above approach iteratively.

ALGORITHM 1. detailed steps of iterative approach

```
MAX_OF_LOGNORMAL( ) {
    $X_M$=$X_1$;
    for(i=2;i<p+1;i++) {
        compute raw moments of X=max($X_M$, $X_i$);
        find $\mu$ and $\sigma$ of X;
        find $\mu$ and $\sigma$ of ln(X);
        $X_M$=X;
    }
}
```

$$\begin{aligned} X_M &= max(X_1, X_2, \ldots X_p) \\ &= max(\ldots max(max(X_1, X_2), X_3), \quad \ldots X_p) \end{aligned} \qquad \text{(EQ 18)}$$

This approach is based on the assumption that the max of two lognormal RVs is another lognormal; we will show the error of this approach later in this section. To summarize, the detailed steps in computing X_M are listed in ALGORITHM 1.

We find that the error incurred by the assumptions taken is very small for the random variables that we are concerned with. We show in Figure 3 that our proposed iterative method provides good accuracy over a wide range of p.

In the special case of identical paths, the greatest of p identical lognormal RVs as in our case can be approximated with a closed-form expression:

$$\begin{aligned} X_M &= max(X_1, X_2, \ldots X_p) \\ &= max(e^{u_1}, e^{u_2}, \ldots e^{u_p}) \\ &= e^{max(u_1, u_2, \ldots u_p)} = e^{u_M} \end{aligned} \qquad \text{(EQ 19)}$$

If u_M, the greatest of normal RVs, can be estimated with a normal RV, we can assume X_M has a lognormal distribution and find its μ and σ from EQ10. It then follows that we need to find μ and σ of u_M. Based on the expression of the maximum of two normal RVs [11], we can derive the maximum of p identical normal RVs for u_M as (see Appendix B):

$$\mu_{u_M} = \mu_0 + \frac{\sigma_0}{\sqrt{\pi}}\frac{\left[1 - \left(1 - \frac{1}{\pi}\right)^{\frac{r}{2}}\right]}{1 - \sqrt{1 - \frac{1}{\pi}}}, \quad \sigma_{u_M} = \sigma_0\left(1 - \frac{1}{\pi}\right)^{\frac{r'}{2}} \qquad \text{(EQ 20)}$$

$$\text{where } r = \log_2 p, \quad r' = -0.034r^2 + 0.85r + 0.28 \quad . \qquad \text{(EQ 21)}$$

In EQ20, r' is replaced by r in the original derivation. However this does not serve as a good approximation and r' is found via curve fitting to provide better accuracy. With μ_{uM}, σ_{uM}, and EQ10, we can find μ and σ expressions for X_M.

We plot the relative error in standard deviation using the iterative MAX_OF_LOGNORMAL and analytical methods in Figure 3 for a

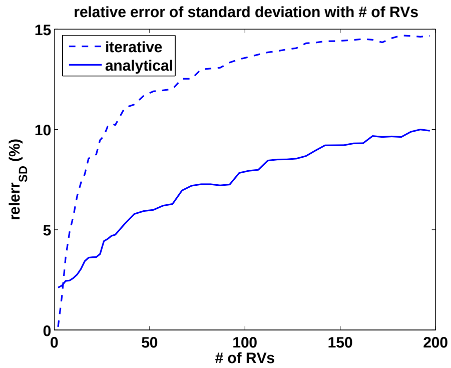

Figure 3. Relative error of two proposed approaches with # of RVs

value of σ_0 corresponding to minimum sized devices. Both methods provide sufficient accuracy in calculating standard deviation. For the error in computing the mean of X_M, the analytical approach typically yields errors on the order of 3-5% for the cases studied. The iterative method is superior in this respect, with error less than 1% (note that curve fitting of r in the μ term of EQ20 could be used to further improve the accuracy of the analytical approach). Since the mean is usually much larger than the standard deviation for X_M with large p, the iterative approach is preferable and is used in the remainder of the paper. Furthermore, while the analytical approach is only applicable to the greatest of *p identical* paths, the iterative approach is general. The analytical approach yields the same trends as those described below and thus provides a simple and efficient way to shed insight on the impact of variability on subthreshold circuits.

Combining the above analysis with Section 3.1.1, we can obtain μ and σ for t_{dm}. With μ and σ of t_{dm} in hand we can find the operating speed based on worst-case delay which is typical practice in ASIC designs. The worst delay *tdly* is

$$tdly = \mu(t_{dm}) + conf \cdot \sigma(t_{dm}) \qquad \text{(EQ 22)}$$

where *conf* is the confidence σ value. We use *conf*=3 in this paper unless otherwise specified.

3.2 Statistical Energy and V_{min} Modeling

Total energy consumption during signal propagation is the sum of active and leakage energy. In our energy modeling, we treat the switching energy deterministically. This is reasonable since switching energy has only linear dependencies and therefore smaller variation compared to leakage energy. Again, we consider worst case leakage energy across all chips as the leakage energy. This is done by taking the μ+*conf**σ of leakage power and *tdly*.

Total leakage current $I_{leak,total}$ can be expressed as

$$I_{leak,\,total} = \sum_{j=1}^{N} I_{leak,\,j} = I_{s0} \cdot \left(\sum_{j=1}^{N} e^{-\frac{V_{th,\,j}}{mV_T}} \right) \qquad \text{(EQ 23)}$$

where N=n*p is the total number of gates in the circuit. Then the worst case leakage current is:

$$\begin{aligned} I_{leakM} &= \mu(I_{leak,\,total}) + conf \cdot \sigma(I_{leak,\,total}) \\ &= I_{s0} e^{-\frac{\mu_{V_{th}}}{mV_T} + \frac{1}{2}\left(\frac{\sigma_{V_{th}}}{mV_T}\right)^2} \left(N + conf \sqrt{N\left(e^{\left(\frac{\sigma_{V_{th}}}{mV_T}\right)^2} - 1 \right)} \right) \end{aligned} \qquad \text{(EQ 24)}$$

The worst case total energy across many dies is

$$E = E_{act} + E_{leakM} = \frac{1}{2} N\alpha C_s V_{dd}^2 + I_{leak,\,M} \cdot V_{dd} \cdot tdly \qquad \text{(EQ 25)}$$

where α is the activity rate. The energy expression without considering variation is [4]

$$E_{nom} = \frac{1}{2} N\alpha C_s V_{dd}^2 + (N \cdot I_{leak0}) \cdot V_{dd} \cdot t_{d,\,nom} \qquad \text{(EQ 26)}$$

where $t_{d,nom}$ is the nominal delay of the inverter chain with n inverters and I_{leak0} is the leakage current per gate. Comparing EQ25 and EQ26, the only difference lies in $I_{leak,M}$ and *tdly*. Therefore, we introduce a statistical adjustment factor A_{stat} to consider both statistical terms:

$$A_{stat} = \frac{I_{leakM} \cdot tdly}{N \cdot I_{leak0} \cdot t_{d,\,nom}} \qquad \text{(EQ 27)}$$

Since subthreshold swing is a function of V_{dd} (a quadratic function serves as a good estimation), the closed-form expression for nominal $V_{min,nom}$ in [4] is no longer accurate. We empirically find the following equation to be a good approximation for $V_{min,nom}$ with η=2.7.

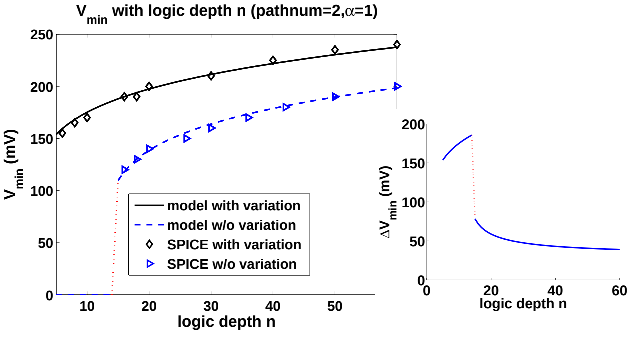

Figure 4. V_{min} results of models vs. HSPICE with logic depth n

$$\frac{V_{min,\,nom}}{V_T} = 3.055\left(\frac{n \cdot \eta}{\alpha} - 35\right)^{0.189} \qquad \text{(EQ 28)}$$

Multiplying n by A_{stat}, we find $V_{min,stat}$ under process variation. Using the analytical expression for *tdly* from Section 3.1, A_{stat} can be written as:

$$\begin{aligned} A_{stat} = {} & \frac{1}{N} e^{t^2} \left(N + 3\sqrt{N(e^{t^2} - 1)}\right) \sqrt{\frac{n}{n - 1 + e^{t^2}}} \\ & \cdot \exp\left(\frac{1 - \left(1 - \frac{1}{\pi}\right)^{\frac{r}{2}}}{1 - \sqrt{1 - \frac{1}{\pi}}} \cdot \sqrt{\frac{\ln\left(1 + \frac{1}{n}\left(e^{t^2} - 1\right)\right)}{\pi}} \right) \\ & \cdot \left(1 + \frac{1}{n}(e^{t^2} - 1)\right)^{\frac{1}{2}\left(1 - \frac{1}{\pi}\right)^{r}} \left(1 + 3\sqrt{\left(1 + \frac{1}{n}\left(e^{t^2} - 1\right)\right)^{\left(1 - \frac{1}{\pi}\right)^{r'}} - 1} \right) \end{aligned} \qquad \text{(EQ 29)}$$

where r and r' are the same as in EQ21, $t = \sigma_{V_{th}}/(mV_T)$.

We now clarify the modeling of V_{th} variation in EQ30. We model the total V_{th} variance as the sum of RDF and L_{eff} components. We neglect spatial correlation in $\sigma_{Vth,Leff}$ since the $\sigma_{Vth,RDF}$ component dominates in this application space, making the error incurred by ignoring spatial correlations small. $\sigma_{Vth,Leff}$ is modeled as the sum of intra-die and inter-die variation. The RDF component is proportional to the inverse of the square root of channel area [7]. Since NMOS and PMOS are sized differently, we take the average k_{Vth} of NMOS and PMOS with results showing good accuracy compared to SPICE.

$$\sigma_{V_{th}} = \sqrt{\sigma^2_{V_{th},\,RDF} + \sigma^2_{V_{th},\,L_{eff}}} = \sqrt{\left(\frac{k_{V_{th}}}{\sqrt{W \cdot L_{eff}}}\right)^2 + \sigma^2_{V_{th},\,L_{eff}}} \qquad \text{(EQ 30)}$$

4 Model Verification and Discussion

We simulate the circuit configuration of Section 3 (p identical paths of n inverters) in SPICE using an industrial 130nm technology with nominal V_{th} of ~350mV. Simulated and modeled results are seen in Figure 4 showing good fit. Inverters use 0.4µm wide NMOS with beta ratio of 1.4. Figure 4 shows that ignoring process variations underestimates V_{min}. In particular, deterministic analysis does not predict a V_{min} (or V_{min}=0) for n<15 and α=1. ΔV_{min} (the difference between V_{min} in deterministic and statistical models) shrinks with increasing logic depth. This follows from larger logic depths enhancing averaging, reducing the spread in timing and leakage energy.

After confirming the accuracy of our model, we apply it to determine the dependency of V_{min} on critical path count p. Results are shown in

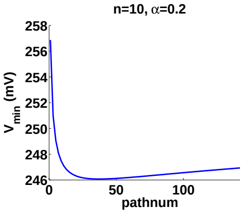

Figure 5. V_{min} versus path count

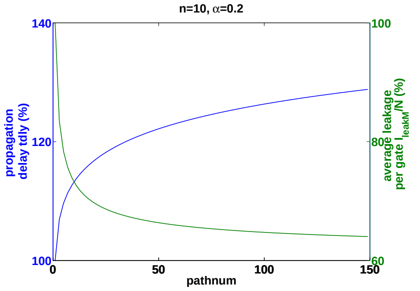

Figure 6. Average leakage and *tdly* versus path count

Figure 5 and Figure 6. Figure 5 shows that V_{min} first reduces and then rises slightly as critical path count increases. The reason for this is shown in Figure 6, which shows the worst case delay *tdly* and average worst case leakage per gate (I_{leakM}/N) versus critical path count. Note that *tdly* increases with more critical paths while average leakage becomes smaller. When p is not large, the reduction in average leakage dominates the increase in *tdly* and V_{min} decreases. However, when p becomes large the average leakage stabilizes while *tdly* continues to increase, yielding a small rise in V_{min}.

Figure 7 shows the statistical and nominal V_{min} with activity rate α. When α is high, the nominal V_{min} model predicts a V_{min} of zero. However, the new statistical model confirms that V_{min} exists. We also show that the V_{min} converges towards the nominal case when sizes increase. As shown, 1μm (W_{nmos}) inverters show smaller V_{min} since it has less variation from RDF. All three curves follow the same V_{min} trend as α decreases. This is due to the fact that α has no interaction with process variation and thus affects V_{min} in the same manner for both nominal and statistical cases.

With our statistical delay model for t_{dm}, some important results can be derived. Figure 8 shows the constant 3σ/μ delay contours in the logic depth and device size for a circuit block with 10 critical paths. Points farther from the origin exhibit less variation, whether due to device upsizing or increased logic depth. For the same amount of relative timing variability, we can compensate for small sizes with larger logic depth and vice versa. Notice that as the target 3σ/μ becomes smaller (<20%), it becomes difficult to achieve with reasonable sizes or logic depth. However, in Figure 8 the total number of paths is 10. With a larger number of critical paths, 3σ/μ naturally reduces as the tail of t_{dm} shrinks.

5 Optimal Pipeline Depth Investigation

Reference [4] showed that energy efficiency improves with increasing switching activity (α) since devices are then performing useful work and not simply contributing to leakage energy. This suggests that a subthreshold design should be aggressively pipelined to raise α and improve energy efficiency. As with any design optimization, increasing the number of pipeline stages will reach its limits. In particular, latch energy overhead will eventually overtake the advantage offered by high activity rates.

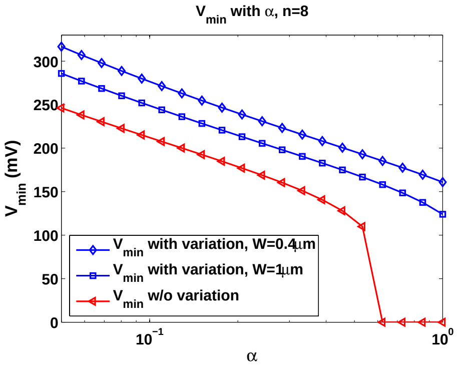

Figure 7. V_{min} with activity rate α

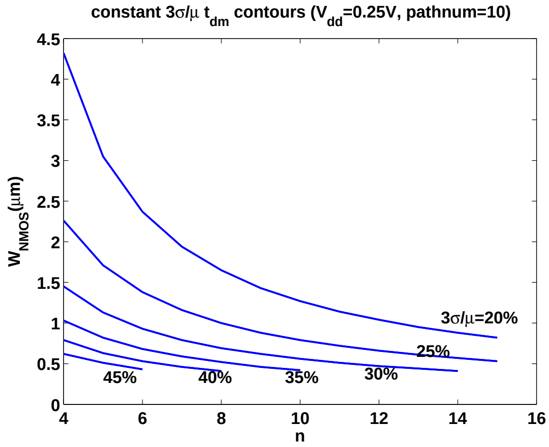

Figure 8. Logic depth/sizing (*n-W*) contours at constant 3σ/μ t_{dm}

The situation is further complicated when considering variability. As shown in Figure 4, V_{min} increases significantly when variation is included, but ΔV_{min} in Figure 4, and subsequently the change in energy, decreases as the logic depth increases. Designing longer paths therefore clearly reduces the effects of individual gate variations on total path delay and energy and we can limit delay and energy variation by increasing logic depth.

In light of process variation, a tradeoff exists between raising α through aggressive pipelining and reducing variation by increasing logic depth. In order to quantify this tradeoff, we examine a simple circuit consisting of 120 inverters with an FO4 load at each node, partitioned into a variable number of pipeline stages. As in Section 4, we use NMOS width of 0.4μm with a beta ratio of 1.4.

For each pipeline depth studied, we seek to minimize the energy consumed per operation. This is fundamentally different than typical superthreshold pipeline studies [12][13] since throughput is not a primary concern in subthreshold circuits. To find the minimum energy at each pipeline depth, we simulate one pipeline stage across a range of voltages. For each voltage, we find the smallest clock period that guarantees operation for the given pipeline stage and then simulate a single switching event during that clock period. The energy consumed during this switching operation is then multiplied

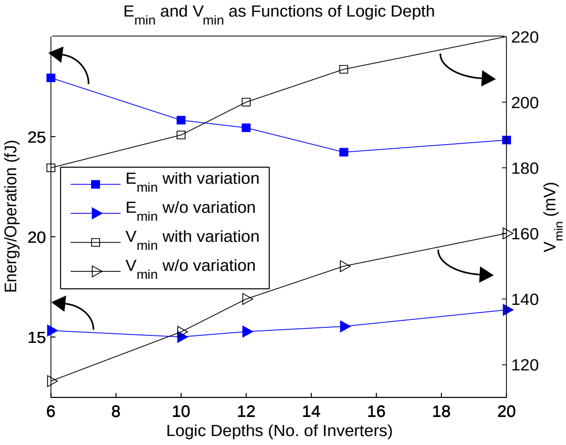

Figure 9. Minimum energy per operation versus logic depth

Table 1. Energy efficiency of different design strategies (*n*: logic depth)

Design Strategy	V_{min}	n	Energy/Op (w/o variation)	Energy/Op (w/variation)
Nominal	130mV	10	15.0 fJ	31.8 fJ
Variability-aware	210mV	15	-	24.2 fJ

by the number of pipeline stages to yield total pipeline energy. We perform this simulation with and without variation considered (both RDF and geometric variations). To account for variation, we perform a Monte Carlo analysis with 1000 iterations for each clock period and energy calculation, and find the minimum $\mu+3*\sigma$ value.

Figure 9 shows both minimum energy per operation (E_{min}), and the corresponding supply voltage (V_{min}), as a function of logic depth. We observe a large increase in E_{min} as a result of variation. Figure 9 highlights the fact that this energy penalty can be minimized by careful sizing of logic chains. For the nominal case, the energy minimum occurs at a logic depth of 10 inverters. The minimum when considering variation is noticeably shifted toward larger logic depths; approximately 15 inverters. Selection of V_{dd} is also critical to minimizing the effects of variation. V_{min} increases by roughly 60 mV across the range of logic depths presented in Figure 9 when variation is included. We begin to see that by designing longer logic paths and increasing supply voltage, the effects of variation can be minimized.

Now consider energy consumption when a circuit is designed without considering variation. For example, if a logic path is designed with 10 inverters per pipeline stage and a supply voltage of 130 mV, as suggested by the nominal results in Figure 9, a designer will expect the circuit to consume 15 fJ per operation. Monte Carlo SPICE simulations show that process variation causes the worst-case energy to be 31.8 fJ. If the circuit is instead designed with 15 inverters per pipeline stage and a supply voltage of 210 mV (conditions leading to minimal energy when variation is included), the worst-case energy consumption is 24.2 fJ, a 24% reduction in maximum energy per operation. This comparison is summarized in Table 1.

6 Conclusions

This paper considers the impact of process variation on subthreshold circuits. We first make several observations about the nature of variation in subthreshold operation and how it fundamentally differs from superthreshold operation. We then derive statistical models of subthreshold circuit delay, power and energy efficiency and verify these using SPICE. With a new statistical model for the minimum energy point V_{min}, we show that a previous nominal model underestimates V_{min} by up to 78mV for small devices. Based on the observation that random dopant fluctuations dominate variability in subthreshold circuits, we suggest design strategies to maintain reasonable variability levels, e.g., <30%. Finally, we explore the role of pipelining in the energy efficiency of subthreshold circuits. We observe a 24% energy reduction when properly considering process variation during microarchitectural planning.

Acknowledgements

This research was supported by NSF, GSRC/DARPA, Intel.

References

[1] A. Wang, A. Chandrakasan, "A 180mV FFT processor using subthreshold circuits techniques", *IEEE ISSCC* 2004

[2] C.H.-I. Kim, *et al.* "Ultra-low-power DLMS adaptive filter for hearing aid applications", *IEEE TVLSI*, Dec. 2003, pp. 1058 - 1067

[3] L. Nazhandali, *et al.*, "Energy optimization of subthreshold-voltage sensor network processors", *ACM ISCA* 2005.

[4] B. Zhai, D. Blaauw, D. Sylvester, K. Flautner, "Theoretical and practical limits of dynamic voltage scaling", *DAC* 2004

[5] B. H. Calhoun, A. Chandrakasan, "Characterizing and modeling minimum energy operation for subthreshold circuits," *ISLPED 2004*

[6] R.W. Keys, "Physical limitations in digital electronics," *Proc. IEEE*, vol. 63, pp. 740-766, 1975.

[7] M. J. M. Pelgrom, *et al.*, "Matching properties of MOS transistors," *IEEE JSSC*, vol. 24, no. 5, pp. 1433-1440, 1989.

[8] BSIM3, http://www-device.eecs.berkeley.edu/~bsim3/get.html

[9] N.C. Beaulieu, *et al.*, "Comparison of methods of computing lognormal sum distributions and outages for digital wireless applications", *IEEE Conf. Communications*, 1994

[10] Wolfram Research, www.mathworld.com

[11] C.E. Clark, "The greatest of a finite set of random variables", *Operations Research*, 1961.

[12] S. Heo and K. Asanovic, "Power-optimal pipelining in deep submicron technology", *IEEE ISLPED,* 2004

[13] A. Chandrakasan, *et al.*, "Low-power CMOS digital design", *IEEE JSSC*, vol. 27, no. 4, pp. 473-484, 1992

Appendix

A. Derivation of greatest of two independent lognormal RVs

In this section, we derive the maximum of two lognormal RVs, similar to [11] where the greatest of normal RVs is derived. By definition [10], $\mu_{z,k}' = \int_{-\infty}^{\infty} x^k f_Z(x)dx = H_1 + H_2$, where

$$H_1 = \int_0^\infty x^k \frac{1}{\sqrt{2\pi}x\sigma_A} e^{-\frac{(\ln x-\mu_A)^2}{2\sigma_A^2}} dx \int_0^x \frac{1}{\sqrt{2\pi}v\sigma_B} e^{-\frac{(\ln v-\mu_B)^2}{2\sigma_B^2}} dv \qquad \text{(EQ 31)}$$

where H_2 is H_1 with A and B interchanged. Since H_1 and H_2 are symmetric, we compute H_1 and then derive H_2 from H_1. Let

$$t = \ln x\,,\ u = (\ln v - \mu_B)/\sigma_B \qquad \text{(EQ 32)}$$

$$H_1(\mu_A) = \int_{-\infty}^{\infty} e^{kt} \left(\frac{1}{\sqrt{2\pi}\sigma_A} e^{-\frac{(t-\mu_A)^2}{2\sigma_A^2}} \int_{-\infty}^{\left(\frac{t-\mu_B}{\sigma_B}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \right) dt \qquad \text{(EQ 33)}$$

If we treat H_1 as a function of μ_B, the differential of H_1 w.r.t μ_B is

$$G(\mu_B) = \frac{dH_1}{d\mu_B} = -\int_{-\infty}^{\infty} e^{kt} \frac{1}{\sqrt{2\pi}\sigma_A} e^{-\frac{(t-\mu_A)^2}{2\sigma_A^2}} \frac{1}{\sqrt{2\pi}\sigma_B} e^{-\frac{(t-\mu_B)^2}{2\sigma_B^2}} dt\,. \qquad \text{(EQ 34)}$$

We can then find $H_1(\mu_B)$ by

$H_1(\mu_B) = \int G(\mu_B) d\mu_B$ with $H_1(\mu_B \to \infty) = 0$.

After some manipulation, we obtain

$$H_1 = e^{k \cdot \mu_A + \frac{1}{2} \cdot k^2 \cdot \sigma_A^2} \Phi\left(\frac{\mu_A - \mu_B + k \cdot \sigma_A^2}{\sqrt{\sigma_A^2 + \sigma_B^2}} \right) \qquad \text{(EQ 35)}$$

We can obtain H_2 by interchanging A and B and arrive at EQ16.

B. Analytical expression of greatest of *p* identical normal RVs

Clark [11] shows how to estimate the maximum of normal variables. The mean and standard deviation of y=max(u_1, u_2) are

$$\mu_y = \mu_0 + \frac{1}{\sqrt{\pi}} \cdot \sigma_0\,,\ \sigma_y = \sigma_0 \cdot \sqrt{1 - \frac{1}{\pi}} \qquad \text{(EQ 36)}$$

where μ_0 and σ_0 are the mean and standard deviation of u_1 and u_2. This implies that if $p=2^r$, we can group the RVs into pairs, then find the μ and σ of every two RVs with EQ36 to obtain 2^{r-1} RVs, denoted as level 1 RVs. Similarly we can continue this process and get level j RVs with

$$\mu_j = \mu_{j-1} + \frac{1}{\sqrt{\pi}} \cdot \sigma_{j-1}\,,\ \sigma_j = \sigma_{j-1} \cdot \sqrt{1 - \frac{1}{\pi}} \qquad \text{(EQ 37)}$$

where $j = 1, 2, \ldots r$. With this iterative approach, we can find the final analytical expression of μ_r and σ_r in EQ20. Note that EQ20 is derived under the assumption that p is a power of 2; it can be extended for any p.