Device Sizing Techniques for High Yield Minimum-Energy Subthreshold Circuits

Dan Holcomb and Mervin John
University of California, Berkeley
May 7, 2008

Abstract

Subthreshold CMOS logic can be used to provide energy-efficient computation in scenarios where performance is not a critical concern. Analysis in existing literature, and explored in this work, demonstrates that the use of minimum-sized devices can offer the greatest energy-efficiency, but in the face of variation, upsizing schemes must be considered to reduce gate failure rates. We explore the impact of variation on low-voltage functionality of gates and propose a novel methodology for determining the gate failure rates under different sizing schemes. We show that optimal sizing for min energy is dependent on loading.

1 Introduction

Emerging IC applications such as wireless sensor nodes, RFIDs, and medical devices place increasing emphasis on low energy operation instead of performance. Subthreshold circuits operate with a supply voltage set below the transistor threshold voltage and thus offer significant energy reduction, while operating at frequencies in the range of kHz up through a few MHz. A number of designs have successfully been presented using mature technology nodes [1, 2, 3]. It has been suggested that minimum size devices are theoretically optimum for minimizing energy [4]. However, subthreshold designs rely on leakage currents that are exponentially dependent on Vth. However, subthreshold designs rely on leakage currents that are exponentially dependent on Vth, and are far more sensitive to process variations than traditional above-threshold designs [3]. Therefore variability must be considered when trying to approach the energy-optimal operating point of circuits using advanced technologies.

We first present an analytical model of energy in subthreshold circuits. We address the major sources of variations in short-channel devices, and present a novel methodology for estimating circuit yield rates based on static noise margin. We use our yield rate estimation methodology to explore sizing techniques for subthreshold robustness. Our goal is a design capable of producing 80% yield for a 100k gate circuit, while minimizing energy per operation.

2 Model of Energy in Subthreshold Circuits

As leakage components such as gate current, gate-induced drain leakage (GIDL), and diode leakages are negligible in subthreshold, the analysis in existing literature equates total current to subthreshold drain-source current (Isub) for VDD below transistor threshold [4]. The well-established equation for modeling subthreshold current is

\[ I_{sub} = I_{o} e^{\frac{V_{GS} - V_{th}}{nV_{th}}} (1 - e^{-\frac{V_{DD}}{V_{th}}}) \]  

(1)

where \( n \) is the subthreshold slope factor \((1 + C_d/C_{ox})\), \( V_{th} \) is the thermal voltage \(kT/q\).

The propagation delay of a characteristic inverter with output capacitance \( C_g \) in subthreshold is

\[ t_d = \frac{KC_gV_{dd}}{I_{o,g} e^{\frac{V_{GS} - V_{th}}{nV_{th}}}} \]  

(2)

where \( K \), \( I_{o,g} \) and \( V_{th} \) are parameters used to fit these equations to the delay of a characteristic inverter. The equations for current (Eq. 1) and delay (Eq. 1) of an inverter can be used to derive the average energy per cycle [4] for arbitrary larger circuits (Eq. 5).

\[ E_T/\text{Cycle} = E_{DYN} + E_L \]  

(4)

\[ E_T/\text{Cycle} = C_{eff}V_{dd}^2 + W_{eff}L_{DP}KC_gV_{dd}^2 e^{-\frac{V_{th}}{V_{DD}}} \]  

(5)

Although Eq. 5 sacrifices accuracy for simplicity since the fitted parameters cannot account for the finer details such as \( \alpha \), it can provide a good estimate of the min-energy operating points for the average case and a basis for comparing designs when variation is included. Many previous work have addressed the dependence of the theoretical min-energy points on the major design parameters including supply voltage, logic depth, and sizing:

- Impact of Vdd: A key observation of the analysis of [4] is that the min-energy \( V_{DD} \) value is independent of frequency and \( V_t \) and is instead set by the relative significance of dynamic and leakage energy components. At higher supply voltages, the dynamic component dominates. As supply voltage decreases, the current decreases due to the DIBL effect but the delay increases exponentially in the subthreshold regime, leading to exponential increases in subthreshold leakage energy. In subthreshold circuits are more sensitive to variation at lower voltage due to the increase in \( V_{DD}/V_{th} \).
• **Impact of Logic Depth:** Previous work has shown that energy efficiency improves with increasing switching activity. The devices are performing useful work more often and not simply contributing to leakage energy [5]. This suggests that subthreshold circuits should be aggressively pipelined to reduce the logic depth, increase $\alpha$, and minimize energy-per-operation. However, variation complicates the situation considerably because short paths are more sensitive to individual device variations. Therefore, a tradeoff exists between raising $\alpha$ through aggressive pipelining and reducing variation through increased logic depth [5].

• **Impact of Sizing:** It has been suggested that minimum sized devices are theoretically optimal for reducing energy per operation [4]. However, process variation imposes a critical constraint on this rule. The standard deviation for threshold voltage is proportional to $1/\sqrt{WL}$, so minimum sized devices produce the worst case random $V_t$ mismatch. In addition, if the gates are driving substantially large loads, the increased delay can lead to higher leakage energy.

In the remainder of this paper, we will primarily use statistical analysis to confirm functionality of minimum sized devices in the face of process variation and explore the use of upsizing to ensure acceptable yield rates, at the expense of increased energy consumption.

### 2.1 Use of Ring Oscillator as Representative Circuit

A representative circuit is required to evaluate the energy cost of different sizing choices. In order to avoid influencing the energy consumption by driving devices from a particular waveform, we use free running oscillators. A beneficial consequence of using free running oscillator is that they will run at maximum frequency, allowing delay to also be determined.

Because the energy consumed by a circuit depends upon what types of gates are in the circuit, we use a distribution of 1/3 NAND2, 1/3 NOR2, and 1/3 INV gates as representative of general logic. This distribution is chosen because it contains both parallel and stacked NMOS and PMOS devices. When measuring power from simulation, we use 3 21-stage ring oscillators, with each one consisting exclusively of a single type of gate. We then find the period of oscillation of each (after three cycles, to ensure steady-state behavior), and measure the energy consumed by the circuit during exactly one cycle of operation, beginning and ending at the 50% rising edge of the same gate. The energy of the representative distribution of gate types is obtained by averaging the energy of the 3 oscillators.

### 3 Model of Variation

CMOS technology scaling over the 40 years has required reduction of the threshold voltage ($V_t$) along with the supply voltage ($V_{DD}$). However, $V_t$ scaling has not been accompanied by a corresponding reduction in threshold voltage variations ($\sigma_{V_t}$). This is primarily a result of random dopant fluctuations (RDF) that become more difficult to control in finer geometries [5]. $V_t$ variation is normally distributed and its $3\sigma$ variation is 30mV in a 180nm CMOS technology [6] but variations in sub-90nm technology are expected to increase. Schemes such as adaptive body bias can be used to compensate for die-wide variations but not for the within-die component of variations. Our goal is to explore sizing techniques to reduce the impact of variation at the gate level.

We use a Monte Carlo approach where the threshold voltages, transistor width variations, and length variations are varied according to the normal distributions given in Table 1. Because the standard deviation of $V_t$ is proportional to $1/\sqrt{W/L}$ [7], the minimum sized circuits that provide theoretical minimum energy have increased sensitivity to $V_t$ variation [8].

<p>| Table 1: The variation parameters used, and the scope of their correlation, if any. |
|---------------------------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>Param.</th>
<th>$\sigma$</th>
<th>Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vth</td>
<td>$30mV\sqrt{\frac{W_{min}L_{min}}{WL}}$</td>
<td>Random for each transistor</td>
</tr>
<tr>
<td>W</td>
<td>3nm</td>
<td>die-to-die</td>
</tr>
<tr>
<td>L</td>
<td>3nm</td>
<td>die-to-die</td>
</tr>
</tbody>
</table>

### 4 Yield Estimation

Because theoretical minimum energy is obtained at low supply voltages, our primary concern is circuit functionality at low voltages. We evaluate the impact of design choices on yield, and extend existing techniques to develop our own method of yield estimation.

#### 4.1 Use of Butterfly Plot Static Noise Margin

We quantify the functionality of a design at a given supply voltage using static noise margin (SNM), as was demonstrated by [8]. The use of a butterfly plot SNM provides a metric for evaluating whether a given gate can generate a logic level sufficient for controlling its fanout gate(s). Neglecting variation, the SNM depends on the choice of driving and fanout gate. We identify a worst-case circuit, and use the failure rate of this circuit as a representative statistic for the failure rates of logic gates in general.

#### 4.2 NAND-NOR Circuit as Worst-Case SNM

In the subthreshold domain, the difference between on current and off current (Eq. 1) are much smaller than in above threshold designs, and the impacts of threshold variation on the current are substantial. The functionality of a gate depends upon its on-current reliably exceeding its off-current; the importance of this is amplified in NAND and NOR gates due to two mechanisms:

- The off currents of the parallel devices will have additive impact
- The sizes of the stacked devices should be kept as small as possible for energy reasons, and thus the on current of the stack will kept minimal.
These two concerns lead to the observation that cross-coupled NAND and NOR gates represent a worst-case test circuit for exploring gate reliability [8], and we adopt this same cross-coupled NAND-NOR as a worst-case circuit for our analysis. The NAND $V_{OL}$ is compared against the NOR $V_{IL}$, and the NOR $V_{OH}$ is compared against the NAND $VIH$. The NAND $VIH$ and the NOR $VIL$ represent the most stringent high and low input margins, respectively. We drive both inputs of the NAND and NOR from the same node, which will increase leakage and decrease on current (Fig. 1). Although the analysis of [8] considers a gate to be failing when its SNM is 0, we conservatively consider any gate with an SNM of less than 15mV to be failing (Fig. 2).

Figure 1: Cross-coupled NAND-NOR circuit is used as a worst-case for verifying logic gate output levels.

Figure 2: VTC curve for 100 Monte Carlo runs of NAND-NOR circuits. Each color represent a single gate design/size and supply voltage. The boxes represent the 15mV SNM required.

4.3 Extrapolating Gate Failure Rate from SNM Distribution

A drawback to the use of SNM instead of a concise analytical model, is its reliance on Monte Carlo simulation, which inherently limits the ability to evaluate the gate failure rates with any realistic probability of failure. This is evident in literature in the analysis of Kwong [8], in which only 5k Monte Carlo runs are performed to evaluate failure rate- which gives rise to use of an allowed gate failure rate of 0.13% - far too high for any realistic design.

To overcome this limitation of SNM analysis we present our own novel method for performing gate failure rate analysis. We use Monte Carlo simulation to characterize the SNM as a Gaussian distribution (Fig. 3), and then infer failure rates from this.

For a given NAND-NOR design at a given supply voltage, the methodology is as follows:

1. Run 1k DC sweeps of NAND and NOR gate circuits using our Monte Carlo variation models (Tab. 1).
2. Determine SNM of each Monte Carlo run using VTC butterfly plots.
3. Use $\mu_{SNM}$ and $\sigma_{SNM}$ to estimate the probability ($F_{GATES}$) that a randomly chosen gate will have an SNM of less than 15mV using Eq. 6.

$$F_{GATE} = \frac{1}{\sigma_{SNM}\sqrt{2\pi}} \int_{-\infty}^{15mV} \exp\left(-\frac{(u-\mu_{SNM})^2}{2\sigma_{SNM}^2}\right) du \quad (6)$$

To highlight the advantages for our method for estimate gate failure rates using SNM, we refer back to the VTC of
two NAND-NOR circuits operating at different supply voltages (Fig. 2). Over the 100 Monte Carlo runs shown in the plot, we can see that no instances of either circuit ever fail to meet the SNM requirement of 15mV – furthermore, we can see that one design has a larger eye but larger variability, whereas the other has a smaller eye but reduced variability. While the typical VTC approach would provide no information to the relative failure rates of the two designs, our approach can overcome this limitation and indicate that the two designs have equivalent gate failure rates.  

To test our SNM-based failure rate methodology we compare it to the basic non-scalable method of counting the number of SNM failures, by applying both methods to 1k Monte Carlo runs of several different sizings of an otherwise identical NAND-NOR circuit across a range of supply voltages (Fig. 4). At low voltages the failure rates are high, and the two methods show good agreement. Once the voltage high is enough to make the failure rate drop below .1%, the approach of simply counting the failures does not find any failures (Fig. 4).

![Figure 4: Probability of a gate failure according to the VTC logic level test using 1k Monte Carlo runs. Three different gate sizings are used, and the supply voltage is swept for each design. The dotted lines are created using our own failure estimation method based on characterizing the SNM distributions. The circles are obtained by counting the number of Monte Carlo runs that have failing SNMs.](image)

Figure 4: Probability of a gate failure according to the VTC logic level test using 1k Monte Carlo runs. Three different gate sizings are used, and the supply voltage is swept for each design. The dotted lines are created using our own failure estimation method based on characterizing the SNM distributions. The circles are obtained by counting the number of Monte Carlo runs that have failing SNMs.

5 Sizing for Minimum Energy under Yield Constraints

We set a design goal of sizing to minimize energy, while hitting a target of 80% yield for a 100k gate circuit. Our methodology for estimating the gate failure rates provides a way to estimate the yield rate for the entire 100k circuit at a given voltage (Eq. 7).

\[ Yield_{CKT} = (1 - F_{GATE})^{N_{GATES}} \]

For a given design, a binary search is used to find the minimum supply voltage at which 80% of 100k-gate-circuits will have all gates be functional. To meet our target, we must have fewer than 2.23 failing gates per million, which translates into having an SNM distribution that satisfies the inequality of Eq. 8.

\[ \mu_{SNM} - 15mV \geq 4.59\sigma_{SNM} \]

5.1 Current and Delay Variability

In addition to maximizing the output swing using the SNM, another important metric is current variability as it directly sets propagation delay variation. Current and delay variability dramatically increases in subthreshold regime due to the exponential impact of \( V_{th} \) and \( L_{eff} \) on subthreshold drive current. It was derived in [5] that current variability is given by

\[ \frac{\sigma_{I_{sub}}}{\mu_{I_{sub}}} = \sqrt{\exp(\sigma_{V_T}/nV_{th})^2 - 1} \]

Considering that RDF dominates uncertainty in subthreshold circuits, we can address this uncertainty through device sizing which reduces its impact. Figure 6 shows current variability (\( \sigma/\mu \)) versus device sizing. Variability decreases with larger widths as expected and stacked devices display lower spreads in active current.

As circuit delay depends on active current, delay variability is also given by Eq. 9 and depends on \( \sigma_{V_T} \). The high delay

\[ Yield_{CKT} = (1 - F_{GATE})^{N_{GATES}} \]

For a given design, a binary search is used to find the minimum supply voltage at which 80% of 100k-gate-circuits will have all gates be functional. To meet our target, we must have fewer than 2.23 failing gates per million, which translates into having an SNM distribution that satisfies the inequality of Eq. 8.

\[ \mu_{SNM} - 15mV \geq 4.59\sigma_{SNM} \]

5.1 Current and Delay Variability

In addition to maximizing the output swing using the SNM, another important metric is current variability as it directly sets propagation delay variation. Current and delay variability dramatically increases in subthreshold regime due to the exponential impact of \( V_{th} \) and \( L_{eff} \) on subthreshold drive current. It was derived in [5] that current variability is given by

\[ \frac{\sigma_{I_{sub}}}{\mu_{I_{sub}}} = \sqrt{\exp(\sigma_{V_T}/nV_{th})^2 - 1} \]

Considering that RDF dominates uncertainty in subthreshold circuits, we can address this uncertainty through device sizing which reduces its impact. Figure 6 shows current variability (\( \sigma/\mu \)) versus device sizing. Variability decreases with larger widths as expected and stacked devices display lower spreads in active current.

As circuit delay depends on active current, delay variability is also given by Eq. 9 and depends on \( \sigma_{V_T} \). The high delay

\[ Yield_{CKT} = (1 - F_{GATE})^{N_{GATES}} \]

For a given design, a binary search is used to find the minimum supply voltage at which 80% of 100k-gate-circuits will have all gates be functional. To meet our target, we must have fewer than 2.23 failing gates per million, which translates into having an SNM distribution that satisfies the inequality of Eq. 8.

\[ \mu_{SNM} - 15mV \geq 4.59\sigma_{SNM} \]
variability of subthreshold circuits requires proper modeling for timing verification. Figure 7 shows the observed delay distribution of a min-sized and an upsized oscillator made of NAND2 gates. An analysis of delay distribution of subthreshold logic circuits is given in [8].

5.2 $\beta_{opt}$ – Optimal P/N Ratio

Due to the relative difference in saturation currents between PMOS and NMOS devices, they are often sized according to a fixed technology-dependent ratio $\beta$ to provide high performance. In subthreshold logic, priority is given to the ability to operate at low voltage, so we find a $\beta_{opt}$ that will maximize robustness to variation and enable our designs to produce 80% yield at low energy operating point.

We find $\beta_{opt}$ by using the SNM of a cross-coupled inverter pair to eliminate any stack effects from our consideration of $\beta$. Without any variation, we size the NMOS device at 400nm, and sweep the PMOS size to find the size that results in the maximum SNM. Contrary to above threshold design, we find an optimal of $\beta_{opt} = 0.5$ (Fig. 8); we believe that this is consistent with the PMOS device having a lower threshold than the NMOS in our 45nm PTM model. As would be expected, once the supply exceeds thresholds, $\beta_{opt}$ trends toward a value greater than 1; consistent with the saturation-current-per-unit-width of an NMOS being greater than that of the strained silicon PMOS.

We optimized the sizing of our NMOS (PMOS) stack transistors such that the equivalent resistance of the pull-down (pull-up) network matched that of the characteristic inverter found earlier. The simulation was conducted with two different load capacitances to remove the effects of parasitics capacitances on the delay.

5.3 Deviating from $\beta_{opt}$

Although the use of $\beta_{opt}$ allows supply voltage to be minimized, minimum energy operation depends largely on the ability to minimize capacitance as well (Eq. 5). Prior literature has shown accepting a higher supply voltage in exchange for reducing $c_{eff}$ can sometimes reduce energy [4]. Noting that the SNM was found to be fairly flat while sweeping $\beta$ (Fig. 8) to choose $\beta_{opt}$, we explore the tradeoff of $\beta_{opt}$ and robust stack sizing against using true minimum sized devices.

Our benchmark for comparison is the minimum sizing that uses $\beta_{opt}$, $W_n = 160nm$ and $W_p = 80nm$ and using stack upsizing. This is compared against two aggressively scaled sizing schemes which deviate from $\beta_{opt}$ and sacrifice min voltage for reduced capacitance. The first design uses $\beta = 1$, so that both PMOS and NMOS devices can be minimum sized, but keeping the stack upsizing by a factor of two. The second aggressively scaled design uses only minimum sized devices, with no stack upsizing, to drive the total capacitance down as much as is possible.
possible. We perform a rough estimation of the average gate capacitance by summing the total transistor width required to create a NAND, NOR and INV, and use our previously discussed algorithm for finding minimum supply voltage under a yield constraint. In the following section, we will consider the impact of wireloading on choosing an optimal sizing scheme. Table 2 shows simulated energy using these 3 schemes on our representative circuits.

Table 2: Abandoning $\beta_{OPT}$ and our optimal stack sizing in favor of minimum sized devices with stack upsizing, without stack upsizing, shows a potential for energy savings. The relative capacitance shown is the total transistor width (relative to $W_{min} = 80nm$) required for a NAND, NOR and INV.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{OPT}$ (normalized)</th>
<th>$\beta = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_p$</td>
<td>80nm</td>
<td>80nm</td>
</tr>
<tr>
<td>$W_n$</td>
<td>160nm</td>
<td>80nm</td>
</tr>
<tr>
<td>Stack Upsizing</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$V_{DD_{MIN}}$</td>
<td>237mV</td>
<td>265mV</td>
</tr>
<tr>
<td>$C_G$ (normalized)</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>$E_{no\ load}$ (fJ)</td>
<td>2.47</td>
<td>2.054</td>
</tr>
<tr>
<td>$E_{with\ load}$ (fJ)</td>
<td>26.06</td>
<td>31.48</td>
</tr>
</tbody>
</table>

5.4 Impact of Wire Loading

We included a 3-T model for wire loading based on estimated numbers for this process. We estimate a capacitance of ($C_{wire} = 20fF$) by scaling down the 50pF capacitance used in [5] for the 130nm technology node, and estimate $R_{wire} = 10\Omega$ based on the resistivity of aluminum, scaling down cross-sectional area from the 65nm technology node PTM model [9], and assuming a wirelength of $5\mu m$. The dynamic energy component of (Eq. 5) shows, a quadratic dependence on $V_{dd}$ and linear dependence on capacitance. When loading is included in the model, the wire capacitance increases the total $C_{eff}$ that needs to be switched. With this increased capacitance, the lower supply voltage due to upsizing shifts the minimum energy point to larger device sizes, as shown in figure 5. This implies that optimal transistor sizing for minimum energy should be chosen with consideration of the wireload and fanout of the design. Table 2 shows that minimum energy design of circuits with high loading favors sizing schemes that enable lower supply voltages over those that aggressively minimize only capacitance.

6 Conclusion

In this work, we considered minimum energy operation of sub-threshold circuits. We examined the effect of variation on minimum sized circuits and examined tradeoffs between low voltage and various sizing techniques for robustness. We introduced a novel methodology for estimating gate failure rates in arbitrarily large circuits, based on fitting a statistical model to a distribution of SNM obtained from Monte Carlo simulation of a worst-case test circuit. Neglecting variation, minimum sized devices will generally offer lowest energy consumption. However, when variability is considered, supply voltages must be increased to ensure functionality. Upsized devices are less susceptible to variability and can achieve high yield at lower supply voltages. By evaluating designs under various loading scenarios, we find that the optimum sizing for minimum energy depends on the amount of capacitive loading. This implies that device sizing must be carefully chosen in accordance with wirelength and fanout of the design.

References