Energy Efficient Design for Subthreshold Supply Voltage Operation

David Blaauw and Bo Zhai

{blaauw, bzhai}@umich.edu, University of Michigan, Ann Arbor, MI

***Abstract* - Subthreshold design has become an important area in low-power design due to its ultra-low power consumption and high energy efficiency. This is very useful in mobile applications where battery life is crucial. For most current DVS processor designs, the voltage range is limited from full V_{dd} to approximately half V_{dd} at most. Subthreshold design enables wide-range dynamic voltage scaling (DVS) by allowing circuits to operate in subthreshold voltages. In this paper we give an overview of several issues in energy-efficient subthreshold design. First, from a theoretical point of view, we show that, for subthreshold supply-voltages, leakage energy becomes dominant, making "just in time completion" energy inefficient. We derive an analytical model for the minimum energy-optimal voltage and study its trends with technology scaling. Second, we evaluate different low- power approaches such as MTCMOS (Multiple-Threshold CMOS) and DVS and show that wide-range DVS provides the best energy efficiency. Finally, we study the impact of process variation on subthreshold operation and discuss which sources of process variation are dominant in this regime and how they can be addressed.**

I. Introduction

Due to technology scaling, microprocessor performance has increased tremendously albeit at the cost of higher power consumption. Dynamic voltage scaling (DVS) was proposed as an effective approach to reduce energy use and is now used in a number of commercial low-power processor designs. In most current DVS processor designs, the voltage range is limited from full V_{dd} to approximately half V_{dd} at most. However, it is well known that CMOS circuits can operate over a very large range of voltage levels down to less then 200 mV. 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 this paper, we give an overview of the pressing issues in subthreshold design [8][9][10]. The first issue that needs to be addressed is the determination of a lower limit of the voltage range for optimal energy efficiency. First, we show that the quadratic relationship between energy and V_{dd} deviates as V_{dd} is scaled down into the subthreshold region of MOSFETs. In subthreshold operation the "on-current" takes the form of subthreshold current, which is exponential with V_{dd}, causing the delay to increase exponentially with voltage scaling. Since leakage energy is linear with the circuit delay, the fraction of leakage energy *increases* with supply-voltage reduction in the subthreshold regime. Although dynamic energy reduces quadratically, at very low voltages, where dynamic and leakage energy become comparable, the total energy can increase with voltage scaling due to the increased circuit delay. We derived an analytical model for this minimum-energy voltage and verify our model using SPICE. We also study its trends as a function of different design and process parameters. Furthermore, we analyze the energy efficiency of different low-power schemes, including MTCMOS, the standard DVS for a number of workload traces obtained from a processor running a wide range of applications.

The other issue is that subthreshold designs have a dramatically increased sensitivity to process variations since drive current becomes exponentially dependent on threshold voltage. 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 difficult to meet design specification predictably without dramatic overdesign which wastes energy efficiency. Therefore, we analyze the impact of process variation on subthreshold design and propose methods to mitigate its effect.

We show that random dopant fluctuations (RDF) [4] 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%. We then analyze the energy efficiency of subthreshold designs while capturing the impact of process variations. The nominal model of ignoring process variations can underestimate the minimum energy voltage by as much as 78mV for small devices, corresponding to a 40% underestimation.

The rest of the paper is organized as follows. Section 2 presents the minimum-energy voltage analysis. In Section 3, we evaluate the energy efficiency of different popular low-power approaches. Section 4 details the analysis of subthreshold circuit delay and power under process variations. Finally, Section 5 concludes the paper.

II. Minimum-Energy Voltage Analysis

We first illustrate the energy dependence on supply voltage using a simple inverter chain consisting of 50 inverters and then extend the analysis for more general circuits. A single transition is used as a stimulus and energy is measured over the time period necessary to propagate the transition through the chain. The energy-V_{dd} relation is plotted in Fig. 1. The dynamic energy component E_{active} reduces quadratically while the leakage energy E_{leak} increases with voltage scaling. The reason for the increase in leakage energy in the subthreshold operating regime is that as the voltage is scaled below the threshold voltage, the on-current (and hence, the circuit delay) decreases exponentially with voltage scaling while the off-current is reduced less severely. Hence, the leakage energy E_{leak} will rise and supersede the dynamic energy E_{active} at about 180mV. This effect creates a minimum energy point (referred to as V_{min}) in the inverter circuit that lies at 200mV, as shown in Fig. 1.

In the previous example, we are implicitly assuming that there is always one input transition per clock cycle. However, the switching activity varies in different circuits and therefore we include the input activity factor α, which is the average number of times the node makes a power consuming transition in one clock period. We now

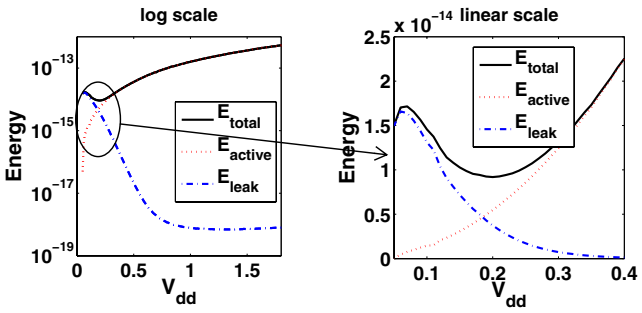

Figure 1. Energy as a function of supply voltage.

derive an analytical expression for the energy of an inverter chain as a function of the supply voltage. Suppose we have an *n*-stage inverter chain with an activity factor of α. The standard expression for subthreshold current is given by [5]:

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

where V_{th} is the threshold voltage of the MOSFET, μ_{eff} is the effective mobility, W is the transistor width, L_{eff} is the effective channel length, V_{gs} and V_{ds} are the gate-to-source and drain-to-source voltages respectively. We now express the total energy E_{nom} per clock cycle as the sum of dynamic and leakage energy:

$$\begin{aligned} E_{nom} &= E_{active} + E_{leak} \\ &= \alpha \cdot n \cdot \left(\frac{1}{2} \cdot C_S \cdot V_{dd}^2\right) + (n \cdot V_{dd} \cdot I_{leak}) \cdot (n \cdot t_p) \end{aligned} \quad (2)$$

where C_s is total switched capacitance of a single inverter, I_{leak} is leakage current of a single inverter, t_p is the delay of a single inverter.

Note that we ignore the short circuit component in energy modeling because we found that short circuit power is negligible in the regime in which we are interested. This is known to hold for well-designed circuits in normal (super-threshold) operation [6]. Using the method in [7], we measured the short circuit current for an inverter chain over a wide range of V_{dd} and have found that short circuit energy percentage is less than 9% at V_{min} and even lower as V_{dd} is further reduced, which is smaller than that at superthreshold. Although rise and fall time increases almost exponentially with the reduced V_{dd} in subthreshold operation, the average short circuit current also scales down almost exponentially with V_{dd}. Therefore, short circuit energy does not increase in subthreshold. In fact, it diminishes as a result of the leakage energy increase.

As the supply voltage reduces the total energy, consumption reaches a minimum at V_{min} Since the delay of the circuit increases dramatically, the circuit now leaks over a larger amount of time. This leads us to obtain the following expression for total energy:

$$\begin{aligned} E_{nom} &= \frac{1}{2} \cdot \alpha \cdot n \cdot C_S \cdot V_{dd}^2 + n \cdot V_{dd} \cdot I_{leak} \cdot n \cdot \frac{\eta C_S V_{dd}}{2 I_{on}} \\ &= \frac{1}{2} n C_S V_{dd}^2 \cdot \left(\alpha + \eta \cdot n \cdot e^{\frac{V_{dd}}{m V_T}}\right) \end{aligned} \quad (3)$$

where η is the delay factor arising from a non-step actual input [8]. Note that I_{on} here is subthreshold "on" current because we are focusing on the subthreshold region where V_{min} occurs. Drain-induced barrier lowering (DIBL) is not an issue in the above derivation because V_{th} term in I_{leak} and I_{on} cancels out. However, DIBL does come into picture when considering variation. This is investigated later in this paper.

Based on this simple expression of total energy, we can find the optimal minimum-energy voltage V_{min} by setting $\partial E / \partial V_{dd} = 0$. It is impossible to solve it analytically, therefore, we use curve-fitting to arrive at the following closed-form expression for the energy-optimal voltage V_{min}:

$$V_{min} = \left[1.587 \ln\left(\eta \cdot \frac{n}{\alpha}\right) - 2.355\right] \cdot m V_T \quad (4)$$

This closed-form formula is fitted for n and α values $20 < \frac{n}{\alpha} < 200$ and provides reasonable accuracy (<4.2% V_{min} relative error compared to the numerical approach) over the data range. (4) can have a different form for future technology when V_{th} reduces and V_{min} moves closer to V_{th} because subthreshold swing (and m) is no longer constant with respect to V_{dd}.

Note that in the presented model, the only parameters that are technology-dependent are η and m. As we switch from one technology to another, it is only required to determine these two parameters which can be easily accomplished. Interestingly, the total energy in (3) and the energy-optimal voltage V_{min} *does not depend on the threshold voltage* V_{th}, as verified using SPICE. This independence is caused by the fact that in subthreshold operation both leakage and delay have similar dependencies on V_{th}, and hence, the effect of V_{th} on the total energy cancels out. Also, we find that the energy-optimal voltage is strongly dependent upon the number of stages in the inverter chain. Active energy is linear with n whereas leakage energy is quadratic with n because in a longer inverter chain, more gates are leaking and these particular gates have more time to leak due to larger total propagation delay. Consequently, V_{min} occurs at a higher voltage in a longer chain. Finally, we point out that V_{min} is strongly related to the activity factor α. In a circuit with lower α, V_{min} occurs at a larger voltage than in circuits with higher α. This is because slower activity gives the circuit more time to leak and effectively increases the stage number. We therefore introduce the notation of *effective stage number* as $n_{eff} = n/\alpha$ to be used.

In order to verify the accuracy of the proposed model, we compared the results from (3) with SPICE simulations for inverter chains of different lengths. In Fig. 2, we compare the energy-V_{dd} relationship predicted by the proposed analytical model in the subthreshold

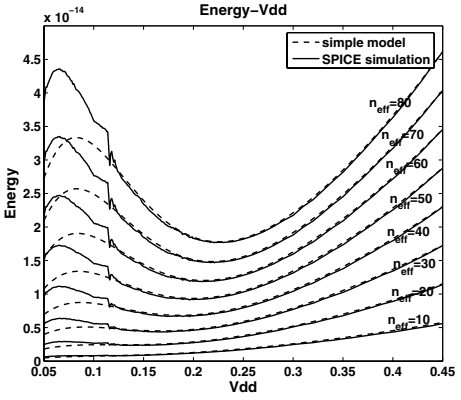

Figure 2. Inverter chain energy-V_{dd} (analytical model vs. SPICE)

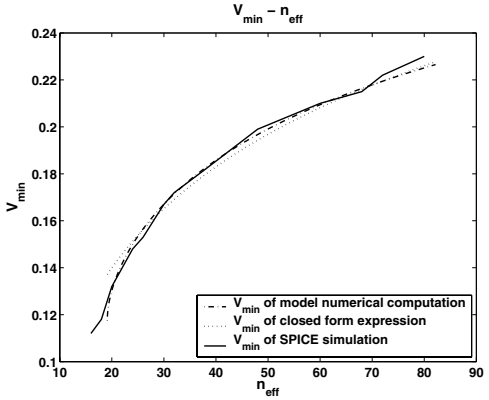

Figure 3. Minimal energy V_{min} with inverter effective stage number n_{eff}

region with SPICE simulation results for an industrial 0.18µm process. The plot shows a range of effective inverter chain lengths (n_{eff}). The analytical model matches SPICE well, except at voltages less than 100mV. In this region, this is not a severe problem since the important region of modeling around V_{min} shows good accuracy.

In Fig. 3, we compare the predicted minimum energy voltage V_{min} based on our model with that measured by SPICE simulation. The results using the fitted closed-form expression of (4) are shown, as well as the numerical solution of V_{min}. As can be seen, both match SPICE with a high degree of accuracy for a wide range of effective inverter chain lengths n_{eff}.

III. Energy Efficiency Evaluation of Different Low-Power Approaches

In this section, we compare the energy efficiency of different low-power design approaches. In the analysis we include the overhead that each specific low-power technique incurs, as well as the efficiency of the DC-DC voltage converter [9]. First, we define the following five different systems:

- S_{basic}, a basic system with clock gating but without power-gating or DVS.
- S_{mtcmos}, a system that employs power-gating during idle mode (clock gating is implied) but no DVS.
- S_{dvspg}, a partial DVS system with power-gating capability where the minimum scalable voltage is V_{limit}, set to be $V_{dd}/2$.
- $S_{dvsonly}$, a system similar to S_{dvspg} but without power-gating.
- S_{insom}, an Insomniac system with aggressive voltage scaling capability, down to the energy-optimal voltage.

We derived energy models [9] for various low-power schemes using the parameters from an existing Alpha processor design. We studied a number of different real applications running on Linux using Transmeta Crusoe TM5600 processors with dynamic voltage scaling and then recorded traces of the minimum necessary performance levels for each application using real-time monitoring. These real applications are selected based on typical activities of laptop computers and comprise both multimedia and interactive applications:

- *emacs* is a trace of user activity using the editor performing light text editing tasks
- *konqueror* and *netscape* are traces of web browsing sessions using the two browsers
- *fs* contains a record of filesystem-intensive operations
- *mpeg* is a trace using MPEG2 video playback

To make a fair comparison, we convert these traces to match the Alpha processor that we have used for physical parameter extraction.

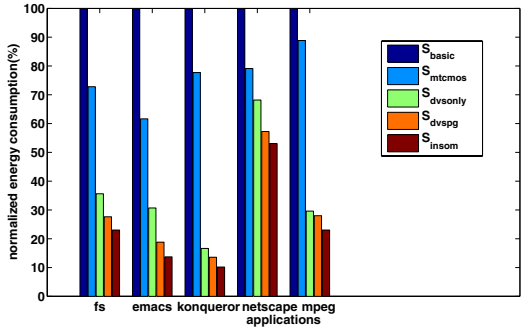

Figure 4. Comparison of energy consumption under different low-power schemes

By applying four different low-power schemes to these workloads, we computed the energy savings relative to S_{basic}. The results are shown in Fig. 4. As the bar graph shows, if the voltage cannot scale as low as the applications request ($S_{dvsonly}$), it is helpful to utilize power-gating (S_{dvspg}) to save leakage energy. However, the largest savings for all five applications is seen with the Insomniac system (S_{insom}). For instance, the energy savings of S_{insom} over S_{dvspg} is 27% for *emacs*, and 25% for *konqueror*.

IV. 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 DIBL, which is reduced 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 [4]. Therefore, I_{on} variation resulting from RDF becomes the dominating component as V_{dd} nears V_{th} as shown in Fig. 5.

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. Fig. 6 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. 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.

In order to estimate the energy consumption under process variation, we need to statistically model both delay and power. 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, implying from (1) that subthreshold on-current and propagation delay exhibit lognormal distributions.

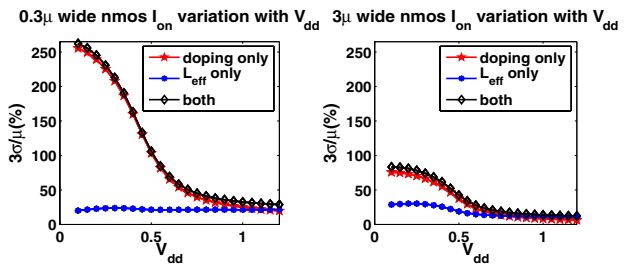

Figure 5. 3σ/µ of I_{on} due to different variation sources over a wide range of V_{dd}, showing the dominance of RDF in subthreshold operation.

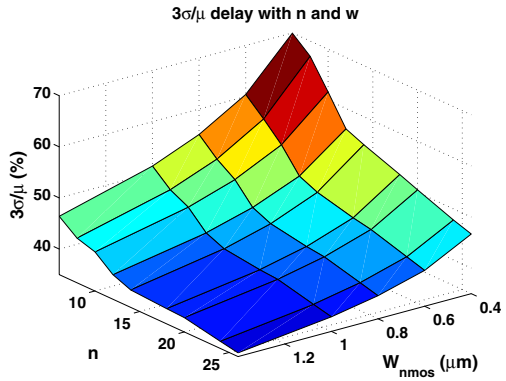

Figure 6. 3σ/μ of delay for an inverter chain with logic depth (*n*) and device sizing (*W*)

We first estimate the sum of lognormal gate delays to obtain the path delay. Then we can find the circuit delay by taking the maximum number of path delays.

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

$$t_{dm} = max(t_{d1}, t_{d2}, \ldots t_{dp}) \quad (5)$$

[10] details how to find the *pdf* of t_{dm} from t_{di} (i.e. the approach to estimate the greatest of lognormal RVs.) Then the worst-case propagation delay *tdly* can be derived from t_{dm} easily.

With the above delay models we can analyze the power/energy and $V_{min,stat}$ under process variation. Total energy consumption during signal propagation is the sum of active and leakage energy. In our energy modeling, we treat the switching energy deterministically since switching energy has only linear dependencies on process variation from C_S, which is much smaller compared to that of leakage energy variation. We consider worst-case leakage energy across all chips as the leakage energy. This is done by taking *tdly* and the worst case leakage power. If we use $I_{leak,M}$ to denote the worst-case leakage current, then the worst-case total energy across many dies is

$$E_{stat} = E_{active} + E_{leakM} = \frac{1}{2}N\alpha C_s V_{dd}^2 + I_{leak,M} \cdot V_{dd} \cdot tdly \quad (6)$$

Comparing (6) and (3), we see that the only difference lies in $I_{leak,M}$ and *tdly*. Therefore, we introduce a statistical adjustment factor A_{stat} to consider both statistical terms:

$$A_{stat} = \frac{I_{leakM} \cdot tdly}{N \cdot I_{leak0} \cdot t_{d,nom}} \quad (7)$$

By multiplying n with A_{stat} in (3), we can find $V_{min,stat}$ under process variation from $\partial E / \partial V_{dd} = 0$.

We simulate the circuit in SPICE using an industrial 130nm technology with a nominal V_{th} of ~350mV. The simulated and modeled results are shown in Fig. 7, demonstrating good fit. It is also shown in Fig. 7 that ignoring process variations results in an underestimation of V_{min}. In particular, the deterministic analysis does not predict a V_{min} (or V_{min}=0) for n<15 and α=1. And ΔV_{min}, the difference between V_{min} in deterministic and statistical models, shrinks with increasing logic depth. This is expected since larger logic depths enhance averaging, reducing the spread in timing and leakage energy. Thus, statistical analysis of subthreshold circuit design must be considered when targeting high energy efficiency.

V. Conclusions

In this paper, we developed analytical models for the most energy efficient supply voltage (V_{min}) for CMOS circuits. A number of interesting conclusions are drawn: 1) Energy shows a clear minimum in the subthreshold region since the time over which a circuit is leaking (delay) grows exponentially in this region while leakage current itself does not drop as rapidly with reduced V_{dd}. 2) V_{min} does not depend on V_{th} if V_{min} is smaller than V_{th}. 3) circuit logic depth and switching factor impact V_{min} since they relate to the relative contributions of leakage energy and active energy. The proposed analytical models are shown to match very well with SPICE simulations. We then compare the energy savings of different low-power schemes, namely, pure MTCMOS, partial-DVS, partial-DVS with MTCMOS, and Insomniac. The comparison for five application traces recorded on two commercial processors shows that Insomniac provides the best efficiency. For instance, it can provide 27% energy savings for *emacs* over the traditional DVS-with-MTCMOS design. Finally, we make several observations about the nature of variation in subthreshold operation and how it fundamentally differs from superthreshold operation. Based on our observation that random dopant fluctuations dominate variability in subthreshold operation, we suggest design strategies to maintain reasonable variability levels, e.g., <30%. We then derive statistical models of subthreshold circuit delay, power and energy efficiency. The nominal model underestimates V_{min} by up to 78mV for small devices, illustrating the need for statistical analysis of high energy efficiency subthreshold circuit design.

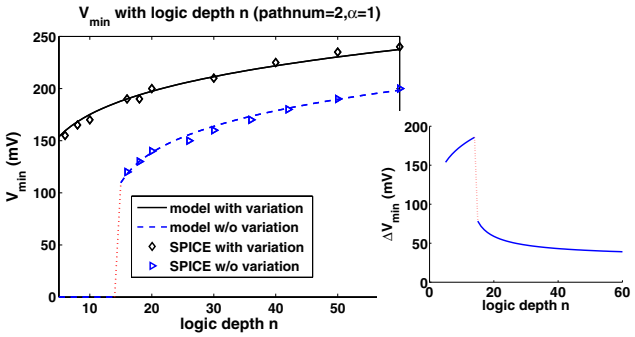

Figure 7. V_{min} results of models vs. HSPICE with logic depth *n*

Acknowledgements

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

References

[1] A. Wang, A. Chandrakasan, "A 180mV FFT Processor Using Subthreshold Circuits Techniques", Digest of Technical Papers, *IEEE ISSCC* Feb. 2004, Vol. 1, pages 292-529.

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

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

[6] J. Rabaey, "Digital Integrated Circuits: A Design Perspective", *Prentice Hall*, 1996.

[7] N. H.E. Weste, K. Eshraghian, "Principles of CMOS VLSI Design, a Systems Perspective", 2nd Edition, *Addison Wesley*, 2000.

[8] B. Zhai, *et al.* "Theoretical and Practical Limits of Dynamic Voltage Scaling", *DAC* 2004

[9] B. Zhai, *et al.* "Extended Dynamic Voltage Scaling for Low Power Design," *SOCC* 2004

[10] B. Zhai, *et al.* "Analysis and Mitigation of Variability in Subthreshold Design," *ISLPED 2005*