Yield-Driven Minimum Energy CMOS Cell Design

Max A. Korbel, Dylan C. Stow, Chris R. Ferguson, David Money Harris
Harvey Mudd College
301 Platt Blvd.
Claremont, CA 91711
{Max_Korbel, David_Harris}@hmc.edu

***Abstract*— CMOS circuits operating near or below threshold offer the lowest energy per computation. Previous work reduces the total energy by using minimum sizing and lowering the voltage without concern for yield. To achieve better yield, the voltage or size must increase. The minimum energy point for minimum-sized NAND2 gates in a 65 nm process is 0.475 V consuming 0.0275 fJ/cycle with a gate failure rate of 2×10^{-4}. However, to achieve a failure rate of 10^{-6}, minimum energy is achieved by widening pMOS transistors by 50%, increasing total energy by 11.9%, which is 7.2% better than minimum width devices and higher voltage.**

I. Introduction

A growing body of applications including wireless sensor networks and energy-scavenging systems has modest computational requirements but demands extremely low energy per operation. Such systems typically operate at a relatively low voltage while actively computing, and then turn off to stop leakage when the computation is complete. The best operating voltage is a balance of dynamic and leakage energy; operating at a lower voltage reduces dynamic energy but increases the computation time and hence the total leakage energy [1].

Previous work has aimed to minimize the total energy of near-threshold and sub-threshold circuits. Early work in this field has concluded that minimum energy E_{min} is achieved by using minimum sizing [2]. Higher activity factors favor lower supply voltage because dynamic energy becomes more important. Previous work suggests that the minimum energy operation point under nominal process conditions is near V_{min} = 250-380 mV for logic [3-4] and higher for SRAM [5]. A Pentium-class processor has been demonstrated to operate at as low as 280 mV, but the minimum energy point was 450 mV, all the gates were at least 2× minimum width, and the yield was not cited [6].

Near-threshold circuits are prone to functional failure due to shifts in the voltage transfer characteristics from process variation, especially when small devices are used. To achieve yield constraints, the operating voltage or transistor widths must be increased [5-7]. A previous study of logic gates considered a gate-level failure rate of 1.3×10^{-4} [7], which is too high to build complex systems. Circuits are also subject to timing failures, but these are outside the scope of this study.

This work explores the sizing and voltage space to find the minimum energy point for logic gates and latches under more stringent yield constraints. We demonstrate that widening the pMOS transistors reduces the total energy for a given failure rate. This finding deviates from conventional wisdom, which suggests that minimum sizing is always energy optimal.

This work is supported by the National Science Foundation under grant 910606 and by the Clay-Wolkin Fellowship at Harvey Mudd College.

II. Simulation Methods

Monte Carlo simulations of logic gates and latches were performed in an IBM 65 nm bulk CMOS process to determine energy and failure rate. The process has a nominal V_t of approximately 300 mV for both nMOS and pMOS [8]. High- and low-V_t devices are about 100 mV higher and lower than nominal, respectively. The global process was set to TT and only local process variation was considered in order to study gate failure rates on a typical die.

The general setup of the simulations is shown in Fig. 1. Each unloaded device under test (DUT) was connected to a separate voltage supply so that energy consumed could be calculated by integrating the supply current over the cycle time. The gates considered were:

- Inverter
- NOR2
- NAND2
- NOR3
- NAND3
- Latch (see Fig. 2)

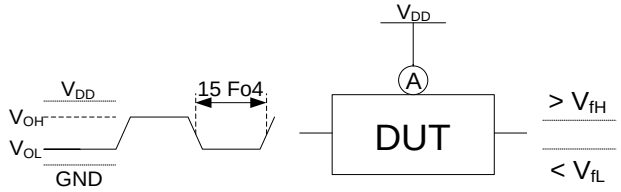

Fig. 1. Diagram of general simulation setup. The output of the DUT is compared with failure criteria. The input of the DUT is affected by noise and is held steady long enough to eliminate timing failures.

Our simulations show that increasing the pMOS width sometimes improved the minimum energy under a failure rate constraint but that minimum nMOS width was preferable for energy, as will be discussed further in Section III. Thus, we considered gates with minimum nMOS width and pMOS width ranging from 1-2× minimum.

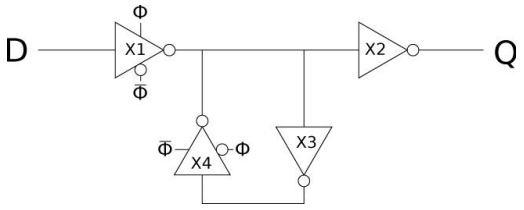

Fig. 2. The design of latch used in this study.

The results are sensitive to the period and the failure criteria selected. The period was obtained from a Monte Carlo

simulation of a chain of 15 fanout-of-4 (FO4) inverters and was set to the mean + 3 sigma to simulate the worst case performance of a logic pathway of moderate length [9]. Each logic gate was supplied with a four input pattern to test the functionality. The latch was similarly tested in its ability to both write and hold both high and low values. The latch receives both a clock input for 15 FO4 high and 15 FO4 low during each cycle, for a cycle time of 30 FO4. A 15 FO4 delay was found to be enough to resolve all timing failures [9]. Therefore, to match the input timing for the latch, we chose to apply each input pattern for the combinational logic for 15 FO4 as well. The period is dependent on V_{DD} [10] and the pMOS width, and is shown in Fig. 3. The same pMOS width was used in the DUT and in the FO4 delay calculation to account for the fact that gates with wider pMOS can run faster and leak for a shorter period.

A gate will fail when its voltage transfer characteristic is so badly distorted by variation that the subsequent stage cannot recognize the output. There are many possible definitions of failure because different gates have different logic levels. Moreover, the amount of noise from crosstalk and other sources is unknown.

This study was based on inverter logic levels. As shown in Fig. 4, each inverter has input logic levels V_{IL} and V_{IH} corresponding to the unity gain points. At these levels, the output is V_{OH} or V_{OL}.

We assume each gate receives an input of V_{OL} or V_{OH} reflecting noise propagated from the previous stage. We set the output failure levels as V_{fL} and V_{fH}, where $V_{fL} = V_{IL}/2$ and $V_{fH} = (V_{IH}+V_{DD})/2$. This allocates half the next stage's input margin to the current gate's distorted output and half to other noise sources.

These margins conservatively allocate 50% of the input margins for other sources of noise such as capacitive coupling crosstalk. Changing the failure criteria definition to allow for more or less noise will respectively shift the results to show a higher or lower predicted yield for a given energy. Simulations showed that non-minimum sizing trends discussed in Section III are observed regardless of the failure criteria.

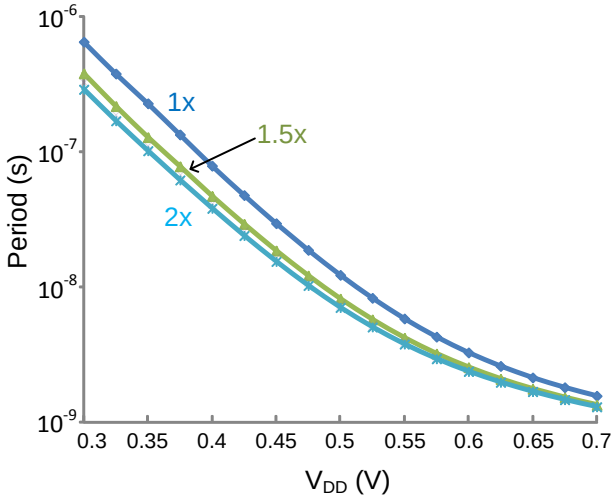

Fig. 3. 15 FO4 period as a function of supply voltage for multiple pMOS sizings.

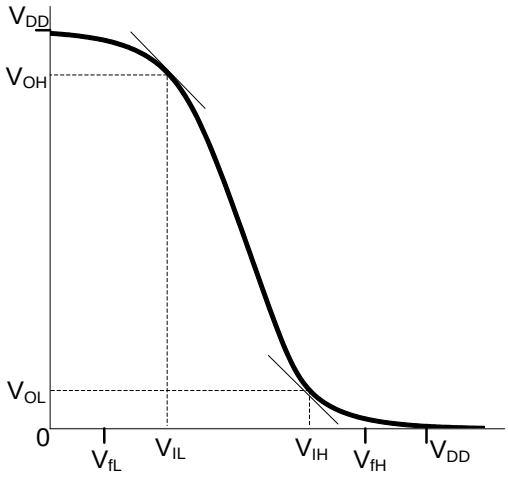

Fig. 4. Failure rate definition criteria based on static noise margins of an ideal inverter.

The gate failure rate is defined to be the fraction of gates in a Monte Carlo simulation whose outputs do not always conform to these levels for one or more input patterns.

III. Results

We observed that minimum energy is achieved by using minimum-width nMOS transistors and upsizing pMOS as necessary for a particular failure rate. Increasing the size of transistors decreases failure rate due to a lowering of σ_{V_t}, which has an inverse square root relationship to transistor area [9][11]. For some gates, increasing transistor width also skews the gate to have more favorable transfer characteristics in the absence of variation. Fig. 5 shows the impact of increasing transistor widths in a NAND2 gate by 50%. Widening either the nMOS or pMOS results in lower failure rates at a given voltage, but widening the pMOS has the greater impact.

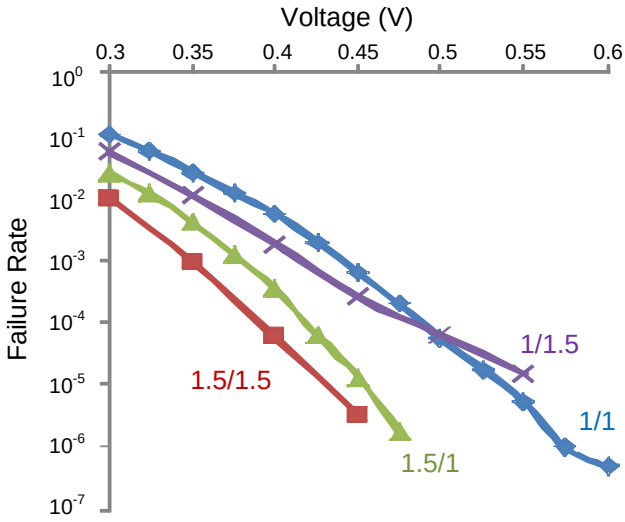

Fig. 5. Failure rate vs. supply voltage for a NAND2 gate for multiple nMOS and pMOS sizings for various P/N ratios (multiples of minimum size).

Fig. 6 plots energy vs. failure rate for the NAND2 gate. In each curve the voltage is varied, resulting in a tradeoff between energy and reliability. At high failure rates, minimum width gives lowest energy, consistent with past practice. But at failure rates lower than 5×10^{-5}, widening the pMOS reduces total energy. Widening the nMOS is never beneficial, even in a NAND2 which has series nMOS transistors. Increasing the ratio of pMOS to nMOS transistor sizes corrects asymmetric mobility and makes the DC transfer curve more symmetric which reduces the failure rate. Thus the remainder of this study focuses on pMOS sizing.

We simulated NAND2 gate failure rate and energy across voltage and sizing, then combined these results to study energy vs. failure rate. Fig. 7 shows that failure rate decreases exponentially with voltage and transistor width, as reported in [7]. Fig. 8 shows the relationship of average energy to voltage and sizing. At low voltage, leakage energy dominates because of the long periods. At high voltage, dynamic energy dominates. The minimum energy point, occurring at 475 mV for minimum-width devices, balances these two components. As the pMOS becomes wider, the dynamic energy increases because of the greater capacitance, and the leakage energy decreases because variability drops, shifting the curve up and to the left. Fig. 9 plots energy vs. failure rate for the NAND2. On the left side of the curve are voltages which have both high failure rates and high energy. These very low voltages are never a desirable operating point. The minimum energy point occurs at a failure rate of 2×10^{-4} using minimum sized devices and a 475 mV supply. Lower failure rates require a higher supply voltage. For failure rates between 10^{-4} and 10^{-5}, it becomes advantageous to widen the pMOS by 25%, and for even lower failure rates, by 50%.

Figs. 10-14 show energy vs. failure rate for other gates. The inverter always achieves lowest energy for a given failure rate when using minimum-width devices. The NOR2, NAND3, and NOR3 are similar to the NAND2, benefiting from wider transistors at low failure rates. The NOR3 and latch both show at least an order of magnitude higher failure rate at the minimum energy point and benefit from widening the pMOS by as much as 100%. Like SRAM [5], reliable latch operation comes at a substantial increase in energy.

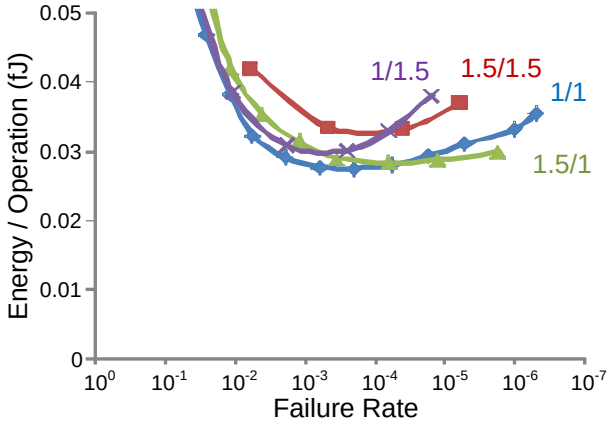

Fig. 6. Energy / operation vs. failure rate for a NAND2 gate for various P/N ratios (multiples of minimum size).

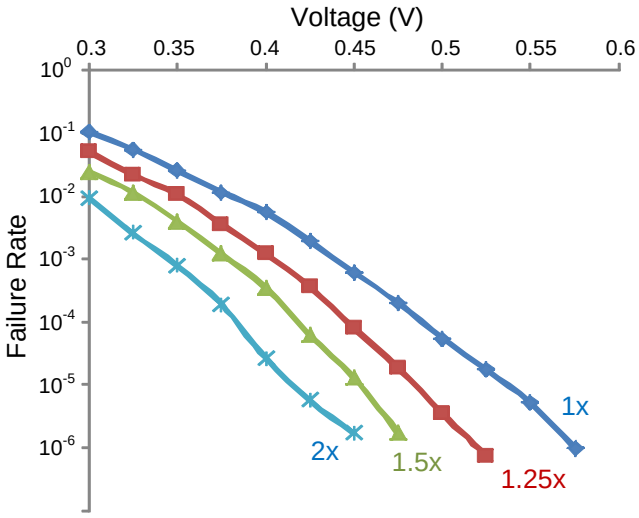

Fig. 7. Failure rate vs. supply voltage for a NAND2 for multiple pMOS sizings.

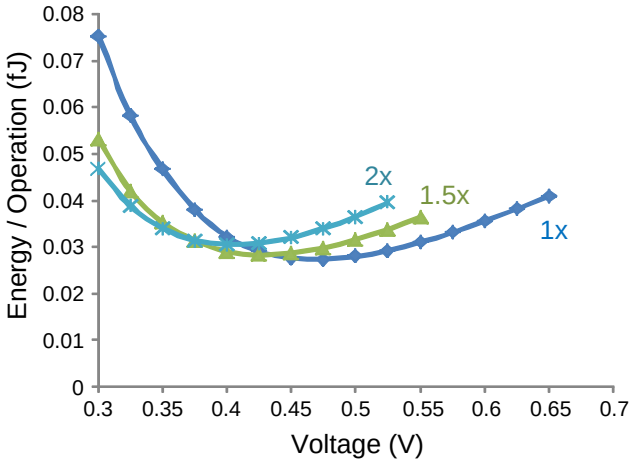

Fig. 8. Energy / operation vs. supply voltage for a NAND2 for multiple pMOS sizings.

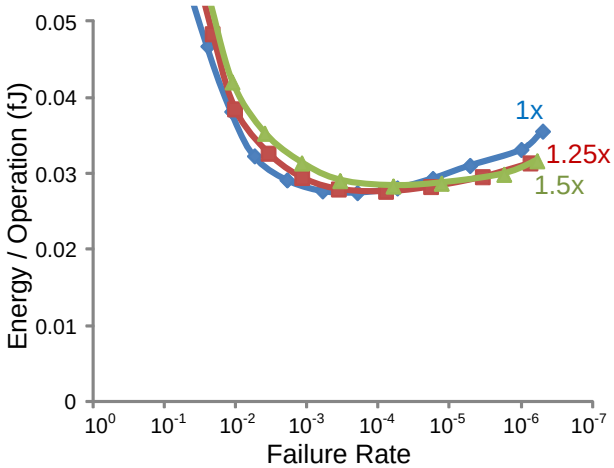

Fig. 9. Energy / operation vs. failure rate for a NAND2 for multiple pMOS sizings. Non-minimum sizing minimizes energy for failure rates below 10^{-4}.

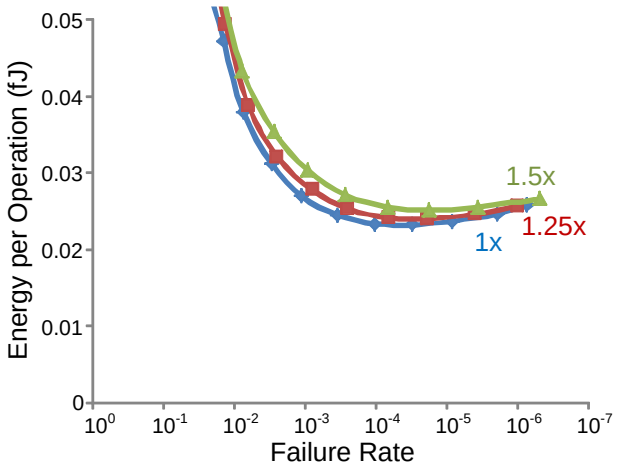

Fig. 10. Energy / operation vs. failure rate for an inverter for multiple pMOS sizings. Minimum sizing always minimizes energy for an inverter.

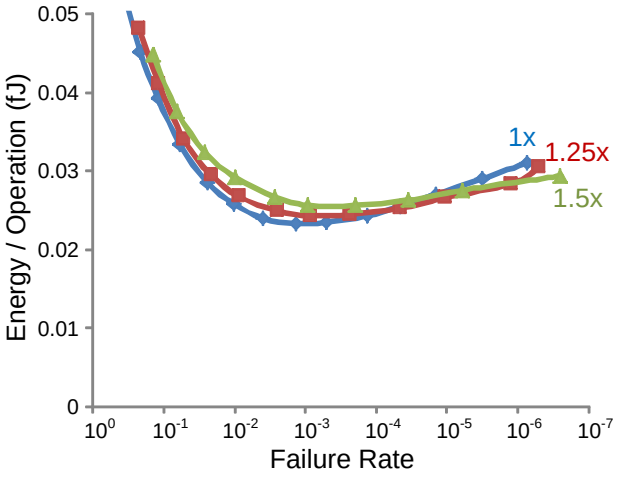

Fig. 11. Energy / operation vs. failure rate for a NOR2 for multiple pMOS sizings. Non-minimum sizing minimizes energy for failure rates below 5×10^{-5}.

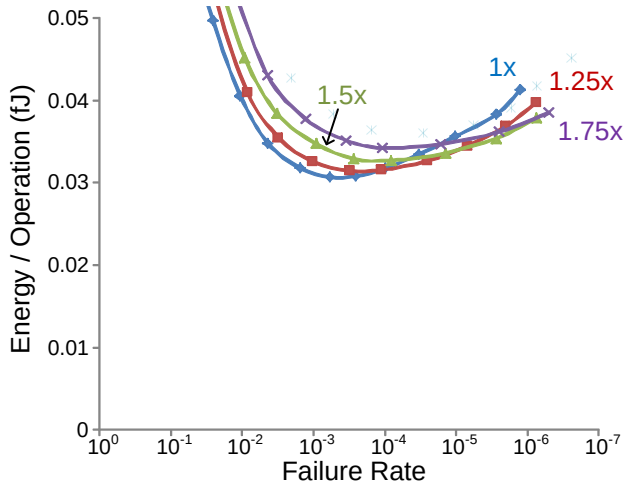

Fig. 12. Energy / operation vs. failure rate for a NAND3 for multiple pMOS sizings. Non-minimum sizing minimizes energy for failure rates below 10^{-4}.

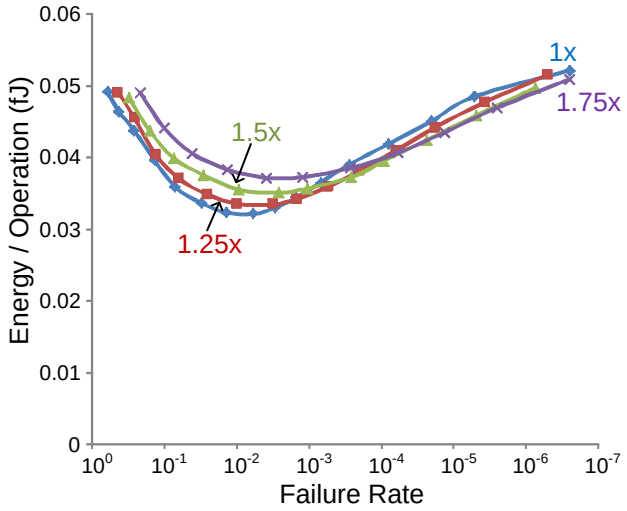

Fig. 13. Energy / operation vs. failure rate for a NOR3 for multiple pMOS sizings. Non-minimum sizing minimizes energy for failure rates below 2×10^{-3}.

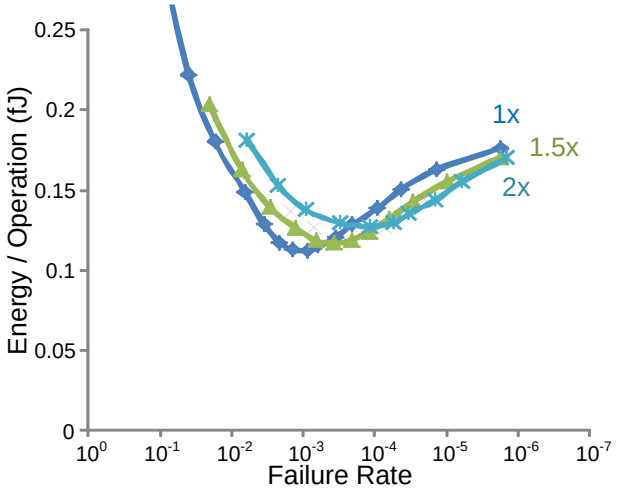

Fig. 14. Energy / operation vs. failure rate for a latch for multiple pMOS sizings. Non-minimum sizing minimizes energy for failure rates below 7×10^{-4}.

Some applications may require lower failure rates than are feasibly detectable using Monte Carlo simulations. Data for significantly lower failure rates can be extrapolated from the curves presented in this study or can be computed using importance sampling [5].

Using devices with different threshold voltages does not reduce the energy in this process. For a NAND2 gate at a failure rate of 10^{-6}, simulations showed a 12% and 8.5% increase in minimum energy using high-V_t and low-V_t transistors respectively.

IV. Conclusion

This study demonstrates that the minimum-energy operating point under a yield constraint occurs in the near-threshold region. Gates should use minimum-width nMOS transistors. However, to achieve gate failure rates better than 10^{-4} that are necessary to build large systems it is beneficial to widen the pMOS transistors of most types of gates rather than

to only increase the supply voltage. The effect is especially pronounced for latches and NOR3 gates whose failure rates at the minimum energy point are very high.

As variability gets worse with feature size scaling, we expect that wider transistors will show even more benefit. An area of future work would be an analytical model relating sizing, variability, and yield. Future work would also confirm that results are not process dependent.

ACKNOWLEDGMENT

We appreciate conversations with David Blaauw at the University of Michigan. We also acknowledge contributions by Sibilla Franca.

REFERENCES

[1] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and Practical Limits of Dynamic Voltage Scaling," in *Design Automation Conf.*, 2004, pp. 868-873.

[2] E. A. Vittoz, "Low-Power Design: Ways to Approach the Limits," in *Intl. Solid-State Circuits Conf.*, Feb. 1994, pp.14-18.

[3] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and Sizing for Minimum Energy Operation in Subthreshold Circuits," in *Intl. Solid-State Circuits Conf.*, Sept. 2005, pp. 1778-1786.

[4] A. Wang and A. Chandrakasan, "A 180-mV Subthreshold FFT Processor Using a Minimum Energy Design Methodology," in *Intl. Solid-State Circuits Conf.*, vol.40, no.1, Jan. 2005, pp. 310-319.

[5] G. Chen, D. Sylvester, D. Blaauw, and T. Mudge, "Yield-Driven Near-Threshold SRAM Design," in *IEEE Trans. Very Large Scale Integration (VLSI) Systems*, vol.18, no.11, Nov. 2010, pp. 1590-1598.

[6] S. Jain, S. Khare, S. Yada, A. V, P. Salihundam, S. Ramani, et al., "A 280mV-to-1.2V Wide-Operating-Range IA-32 Processor in 32nm CMOS," in *Intl. Solid-State Circuits Conf.*, 2012, pp. 66.

[7] J. Kwong and A. P. Chandrakasan, "Variation-Driven Device Sizing for Minimum Energy Sub-threshold Circuits," in *Intl. Symp. Low Power Electronics and Design*, Oct. 2006, pp. 8-13.

[8] N. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*, 4th ed. Boston: Addison-Wesley, 2011.

[9] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and Mitigation of Variability in Subthreshold Design," in *Intl. Symp. Low Power Electronics and Design*, Aug. 2005, pp. 20-25.

[10] S. Hanson, B. Zhai, D. Blaauw, D. Sylvester, A. Bryant, and X. Wang, "Energy Optimality and Variability in Subthreshold Design," in *Intl. Symp. Low Power Electronics and Design*, Oct. 2006, pp. 363-365.

[11] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching Properties of MOS Transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, Oct. 1989, pp. 1433-1440.