Managing Variability for Ultimate Energy Efficiency

Borivoje Nikolić
Electrical Engineering and Computer Sciences, University of California, Berkeley, USA

***Abstract*— Technology scaling is in the era where the chip performance is constrained by its power dissipation. Although the power limits vary with the application domain, they dictate the choice of technology, architecture, and implementation techniques that trade off performance for power savings. Energy-efficient design is often achieved for designs that are sensitive to technology and design parameters. On the other hand, increased variability in semiconductor process technology and devices requires added margins in the design to guarantee the desired yield. Sources of variability in scaled technologies are reviewed, along with models and methods for their capture in design. Variability is characterized with respect to the distribution of its components, its spatial and temporal characteristics and its impact on specific circuit topologies. Methods of desensitizing the digital logic and SRAM to variability at low supply voltages are demonstrated.**

Keywords-CMOS, variability, SRAM.

I. INTRODUCTION

Increasing process variability is perceived as one of the major roadblocks for technology scaling [1]. Introduction of high-k dielectrics with metal gates and thin-body devices in advanced technology nodes has enabled better transistor on- and off characteristics and has suppressed variability. However, it is becoming increasingly difficult for the device tolerances to track the scaling rate of the minimum feature sizes. Sources of variability are in the transistors, interconnect, and in the operating environment (supply and temperature) [2]. Device parameters vary systematically because of deviations in nominal geometries, film thicknesses, dose of implants due to the manufacturing process, and changes in mechanical stress conditions [3]. Random device parameter fluctuations are associated with atomistic variations in device structure.

Concurrently, every IC design in scaled technologies is energy limited. Regardless of the type of the design, the maximum achievable performance depends on the efficiency of computation per unit of energy. Therefore, it is crucial to either minimize energy consumption subject to a throughput constraint, or maximize the amount of computation for a given energy budget. To enable better energy efficiency, a recent trend dictates the use of higher parallelism in each technology generation. In this scenario, optimization for energy requires further lowering of the supply voltages. To mitigate the impact of increased variability, appropriate design margins have to be added to every component of an integrated circuit. Impact of technology variability on performance and power is accentuated at low supply voltages. A certain amount of deviation from the nominal value of a technology parameter causes a larger relative change in circuit parameters at low voltages. Therefore, the requirements for robust operation at low voltages often contradict the needs for energy efficiency, and this is exacerbated by variability.

This paper reviews various classes of technology variability, analyzes their interactions and correlations, and presents methods for their realistic accounting in the design margins for SRAM and logic.

II. TECHNOLOGY VARIABILITY

CMOS process parameter variability is often classified into three categories: known systematic, known random and unknown [3]. Systematic process variations are deterministic shifts in space and time of process parameters, whereas random variations change the performance of any individual instance in the design in an arbitrary way. Systematic variations are, in general, spatially correlated, and affect all the devices in the same way. In practice, although many of the systematic variations have a deterministic source, they are either not known at the design time, or are too complex to model, and are thus treated as random. As a result, many of the sources of variability are not modeled in the design kits and have to be treated as random in the design process. The resulting 'random' variation component, depending on the way systematic variability is modeled, will often appear to have a varying degree of spatial correlation [4].

Variations reflect both the spatial as well as the temporal characteristics of the process and cause different dies and wafers to have different properties. Spatial variations are generally characterized as within-die (WID), die-to-die (D2D) and wafer-to-wafer (W2W) [2]. While the W2W variations dominated in the past, with scaling of the technology, WID and D2D variations can occupy a majority of the process spread. The performance of the manufacturing equipment, expressed through the dose, speed, vibration, focus, or temperature, varies within one die and from die to die. Those parameters that vary rapidly over distances smaller than the dimension of a die result in WID variations whereas variations that change gradually over the wafer will cause D2D variations. Similarly, even more parameters vary from wafer to wafer (W2W variations) and between different manufacturing runs (L2L variations).

Many sources of systematic spatial variability can be attributed to the different steps of the manufacturing process. The photolithography and etching contribute significantly to variations in nominal lengths and widths due to the complexity required to fabricate sublithographic lines that are much narrower than the wavelength of light used to print them. Significant contributors in this area include temperature non-uniformities in the critical post-exposure bake (PEB) and etch steps. Variation in film thicknesses (e.g., oxide thickness, gate stacks, wire and dielectric layer height) is due to the deposition and growth process, as well as the chemical-

978-1-4577-0618-9/11/$26.00 ©2011 IEEE

mechanical planarization (CMP) step. Additional electrical properties of CMOS devices are affected by variations in the dosage of implants, as well as the temperature of annealing steps. In recent technologies, overlay error, mask error, shift in wafer scan speed, rapid thermal anneal and the dependence of stress and proximity on layout have become notable sources of systematic variations.

Random device parameter fluctuations stem mainly from line-edge roughness (LER) [5], Si/SiO_2 and polysilicon (poly-Si) interface roughness [6] and random dopant fluctuations (RDF) [7].

Operating environment of the devices on a chip varies as well. Global variations in the supply voltage as well as variations in the local supply grid directly affect the CMOS gate delays, presenting sources of spatially-correlated variability. Operating temperature varies as well, both globally and locally, adding another spatially-correlated component of performance variability.

Device parameters also vary in time, during the design process or during the chip lifetime. Variations in time include intentional and random changes in the manufacturing process, time-dependent degradation in transistor parameters and changes in supply and temperature. Time-dependent degradation in transistor performance, particularly due to bias temperature instability (BTI), is a major concern. Negative BTI (NBTI) is caused by trapping of the carriers in the PMOS gate interfaces under high biases, which causes threshold increase and degraded current. BTI, caused by trapping of the carriers in the PMOS gate interfaces under high biases, which causes threshold increase and degraded current, used only to affect PMOS transistors in $Si-O_2$ gate stacks, now affects both NMOS and PMOS transistors in high-K metal gate devices [8]. Random telegraph signal (RTS) noise is also another time-dependent source of variability that is becoming a significant concern in design with highly scaled transistors, particularly in memory applications. It is estimated that V_{th} fluctuation due to RTS will exceed V_{th} variation due to RDF at 3 sigma levels at the 22nm technology node [9].

Chip yield is the probability that a chip is both functional and meets the parametric constraints, such as timing and power. A circuit with more design margin will have a higher yield, as it will be more immune to variability. The challenge is in finding the smallest margin necessary for the required yield so that performance is not overly constrained, resulting in large power overhead. The appropriate design margin depends on the type of design, circuit style, its function and use. In a typical VLSI design process, satisfying design corners is deemed necessary and assumed sufficient to validate a design. Process corners are the simplest and most commonly used methods of capturing variability in the design, where all devices on a chip are assumed to systematically vary their performance in accordance with each corner. Traditionally, this method has yielded sufficiently conservative designs that guaranteed desired yield. Nowadays, it is challenging to maintain design corners that will yield appropriate margins without an excessive power penalty.

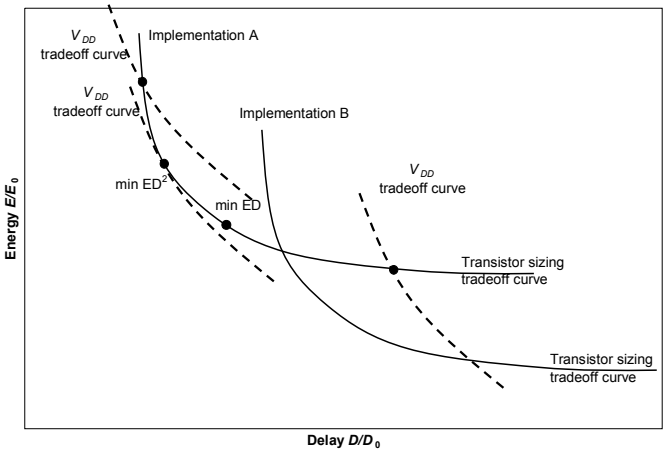

Figure 1: Illustration of energy-delay tradeoffs using architectural changes, supply voltages and transistor sizing.

III. Variability and Energy Efficiency

In the energy-constrained design regime, a system can be optimized to maximize the performance under energy constraints, or to minimize the energy under performance constraints. Either approach can be formulated as a constrained optimization problem. By using sensitivity based analysis or 'hardware intensities/ a tradeoff between energy and performance can be analyzed. Sensitivity is defined as the absolute gradient of energy to delay with respect to a change in some design variable.

There are usually several design variables that can be exploited to trade off energy for performance at various levels of design hierarchy. The tradeoff achieved by tuning some design variable x is given by the energy/delay sensitivity to variable x:

$$S_x(X) = \left.\frac{\partial E/\partial x}{\partial D/\partial x}\right|_{x=X}. \quad (1)$$

This quantity represents the amount of energy that can be traded for delay by tuning variable x, around the design point X. The energy-efficient design is achieved when the relative sensitivities to all the tuning variables are balanced , or when the limit value of a tuning variable is reached [10-11]. Example tuning variables are the supply voltage, V_{DD} or gate sizes, as illustrated in Figure 1 [12]. At high supply voltages, relative energy-delay sensitivity ($S(V_{DD})E/D$) of a design is approximately 2, meaning that 2% of energy savings can be traded off with 1% of delay increase. Relative sensitivity to V_{DD} decreases with the reduction in supply voltage and is practically limited with achieving a robust design at low supply voltages due to high delay variability due to process variations. Many highly parallel designs can achieve the optimal operating point at low supply voltages, near the transistor threshold. To enable this design point, it is necessary to provide a means for operating logic and memory at very low supply voltages, near the transistor threshold.

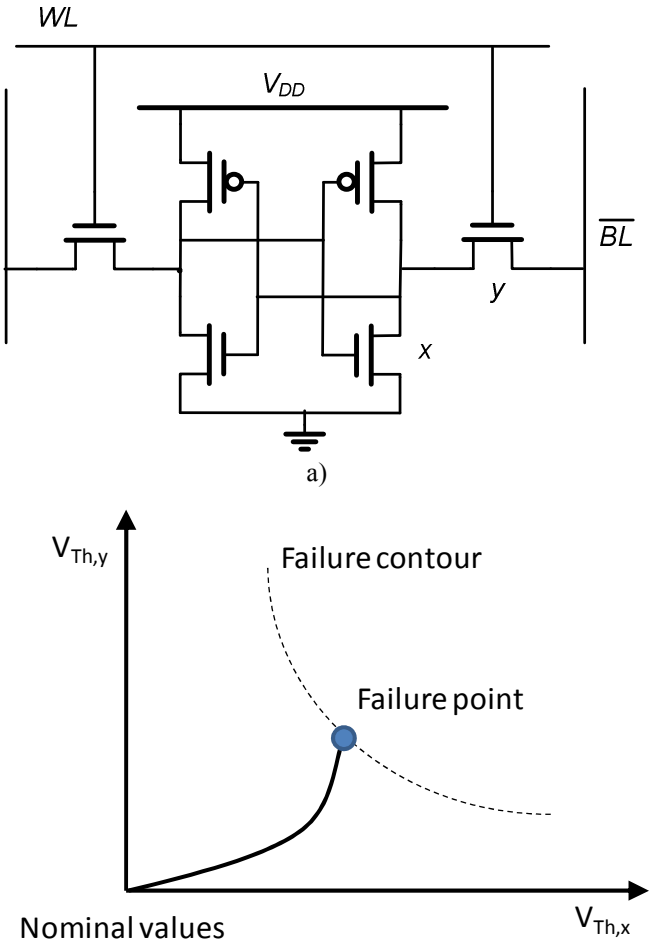

Figure 2: a) 6T SRAM cell; b) Illustration of the most likely failure point in SRAM: x and y illustrate a pair of devices on which the read failure depends the most.

IV. Memory and Logic In a Wide Supply Range

Maximizing energy efficiency for a wide range of applications with widely varying throughputs requires an ability to operate logic and embedded memory in a wide range of supply voltages.

SRAM. Guaranteeing yield for a large SRAM array is a challenging statistical optimization problem, even when the distributions of each transistor's parameters are Gaussian at relatively high supply voltages. Functionality of SRAM is usually assessed by evaluating its static and dynamic noise margins to determine the most likely point of failure. SRAM fails because of inability to write the data in target cycle time, to read it non-destructively or to retain it between access cycles. Sufficient margins against any of the failure modes are estimated through SPICE and TCAD simulations and array characterizations. When the process cannot guarantee sufficient yields, they are increased through the use of redundancy, error correction and assist techniques.

Regular layouts have allowed aggressive scaling of SRAM transistors compared to combinational logic. Simultaneously, these highly scaled devices with design rules relaxed compared to digital logic tend to exhibit higher sensitivities to systematic effects in addition to increased random variation. Systematic effects that affect SRAM have been attributed to temperature non-uniformities during annealing, STI-induced stress, and process-induced cell asymmetry [13]. Increased variability with technology scaling has a large negative impact on SRAM design. SRAM cells use the smallest transistors available, and therefore are susceptible to largest amounts of random variability, while the technology scaling enables integration of twice as many cells in each new process generation. As a result, it is becoming necessary to satisfy the design where the functionality of the cell is guaranteed more than 7 standard deviations away from the mean, while the SRAM yield is guaranteed through appropriate design margins against various failure modes. Different modes of operation of an SRAM cell stress different combinations of transistors. A particular combination of deviations of a set of transistor parameters, as illustrated by two thresholds in Figure 2, yields to the most likely failure point. The failure contour corresponds to a static or dynamic failure criterion.

To improve the read stability or writeability in SRAM, the average margin is increased by adjusting one of the terminal voltages. Lowering the column supply voltage or writing with bitline voltages less than 0V has been used to improve the writeability of the cell, meanwhile, lowering the wordline voltage has been demonstrated to improve the read stability while trading off writeability.

The key challenge in minimizing the SRAM operating voltage remains in the ability to accurately estimate the extreme values in the minimum operating voltage of its cells due to aggregate variations effects.

Logic. Digital logic typically utilizes larger devices than SRAM, which results in lower random variation per gate and a reduction in impact of some of the components of systematic variability. Furthermore, long critical paths in digital logic naturally average random, spatially uncorrelated variations. As a result, longer critical paths reduce the impact of random variability; the σ/μ of random variability roughly decreases with $\sqrt{N}$, where the N is the number of gates in the path. Longest paths in a circuit need to meet the setup time requirement for the receiving flip-flop, which need to be margined appropriately, as illustrated in Figure 3.a. Shortest paths need to be margined for avoiding the hold time violations. Hold margins are often dictated by the timing mismatches between individual gates and are not reduced through averaging. Systematic and spatially correlated variations are not averaged and σ/μ is independent of the logic depth. The hold time margin is essentially dictated by the mismatch in the delays of clock buffers and a Clk-Q path of the flip-flop, Figure 3.b.

Traditionally, correct functioning of digital logic is verified by using static timing analysis (STA), which checks if all timing paths meet their setup and hold requirements. To account for variability in STA, this verification has been

performed in multiple process corners. However, the closest point of failure does not necessarily correspond to one of the traditional corners; as a result, the number of process corners for design verification has been increasing. Furthermore, multi-corner STA may introduce artificially large margins in the design. An alternative approach for timing analysis is statistical STA (SSTA). In particular, block-based SSTA tries to recover linear run-time complexity, identification of a critical path and incremental nature of a traditional STA [14].

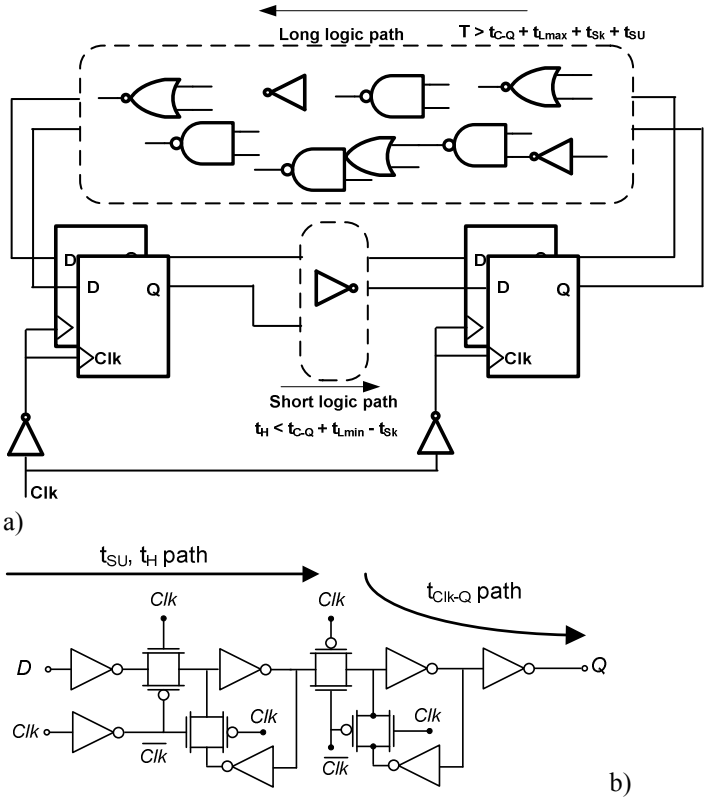

Figure 2: a) Illustration of long and short timing paths in a digital circuit; b) Clk-Q, setup and hold timing paths in a flip-flop.

To reduce the unnecessary margins in the design, while accounting for systematic and random variability the method of dynamic voltage scaling has been adjusting the operating supply and frequency by monitoring of critical path delay. To account for dependences in variability of different gate topologies on the supply voltage, an appropriate mix of gates should compose the set of critical and near-critical path replicas [15]. Still, mismatch between the actual critical paths and their replicas requires an added margin in the design. This margin can be reduced by monitoring the actual circuit failure caused by the inability of receiving latch to receive the data [16]. A challenge with this approach is to minimize the overhead of monitoring numerous critical paths for possible timing violations.

V. Conclusion

Variability limits the lowest operating voltage for a technology. This presents a challenge for continued scaling, where one of the major scenarios relies on continued improvements in energy efficiency of multicore processors through voltage scaling. To overcome voltage scaling barriers, variability characterization needs to be extended to enable compact, in-situ energy and performance monitoring of logic and memory blocks. Continued improvement in design techniques, which incorporate mitigation of the effects of variability, in addition to continuous performance monitoring would enable operation of high-volume products at near-threshold supplies.

References

[1] K.A. Bowman, S.G. Duvall, J.D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *IEEE Journal of Solid-State Circuits*, vol. 37, no.2, pp.183-190, Feb. 2002.

[2] K. Bernstein, *et al*, "High-performance CMOS variability in the 65-nm regime and beyond," *IBM Journal of Research and Development*, vol. 50, no.4-5, pp.433-449, July/Sept. 2006.

[3] S. Nassif, "Delay variability: sources, impacts and trends," *IEEE Int. Solid-State Circuits Conf., Dig. Of Tech. Papers*, San Francisco, CA, Feb. 2000, pp. 368 – 369.

[4] K. Qian, C. J. Spanos, "A comprehensive model of process variability for statistical timing optimization," in *Design for Manufacturability through Design-Process Integration II, V. K. Singh and M. L. Rieger, Eds., Proceedings of SPIE, Vol. 6925, Bellingham, WA: SPIE -- Society of Photo-Optical Instrumentation Engineers*, 2008, pp. 1G-1-11.

[5] A. B. Kahng, Y. C. Pati, "Subwavelength lithography and its potential impact on design and EDA," in *Proc. Design Automation Conference*, New Orleans, LA, USA, June 1999, pp. 799–804.

[6] P. Oldiges, *et al*, "Modeling line edge roughness effects in sub 100 nanometer gate length devices," *Proc. Int. Conf. on Simulation of Semiconductor Processes and Devicces, SISPAD 2000*. Seattle, WA, Sept. 6-8, 2000. pp. 131-134.

[7] A. Asenov, S. Kaya, J.H. Davies, "Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations," *IEEE Trans. on Electron Devices*, vol. 49, no. 1, pp. 112-119, Jan. 2002.

[8] J.C. Lin, A.S. Oates, C.H. Yu, "Time dependent Vcc_{min} degradation of SRAM fabricated with high-k gate dielectrics," *Proc.45th Annual IEEE International Reliability Physics Symposium,* Phoenix, AZ, Apr. 15-19, 2007, pp.439-444.

[9] N. Tega, *et al*, "Increasing threshold voltage variation due to random telegraph noise in FETs as gate lengths scale to 20nm," *2009 Symp. on VLSI Tech. Dig Tech. Papers*, Kyoto, Japan, June 2009, pp. 50-51.

[10] V. Zyuban, et al, "Integrated analysis of power and performance for pipelined microprocessors," *IEEE Transactions on Computers*, vol. 53, no. 8, pp. 1004 – 1016, Aug. 2004.

[11] D. Marković, V. Stojanović, B. Nikolić, M.A. Horowitz, R.W. Brodersen, "Methods for true energy-performance optimization," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 8, pp. 1282-1293, August 2004.

[12] B. Nikolić, "Design in the power-limited scaling regime," *IEEE Transactions on Electron Devices*, vol. 55, no. 1, pp. 71-83, January 2008.

[13] Z. Guo, A. Carlson, L.-T. Pang, K. Duong, T.-J. King Liu, B. Nikolić, "Large-scale SRAM variability characterization in 45nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 44, no.11, pp. 3174-3192, Nov. 2009.

[14] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, S. Narayan, "First-Order Incremental Statistical Timing Analysis", *Proc. Design Automation Conference, DAC'04*, June 2004.

[15] A. Drake, *et al*, "A distributed critical-path timing monitor for a 65nm high-performance microprocessor, *IEEE International Solid-State Circuits Conference, ISSCC'2007*, San Francisco, CA, Feb. 2007. pp. 398-399.

[16] D. Ernst, et al, "Razor: a low-power pipeline based on circuit-level timing speculation," *Proc. 36th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO-36*, 2003, pp. 7-18.