Probabilistic & Statistical Design— the Wave of the Future

Shekhar Borkar

Intel Corp, 2111 NE 25th Ave, Hillsboro, OR 97124, USA
Shekhar.Y.Borkar@intel.com

Abstract. As technology scales, variability will continue to become worse. Random dopant fluctuations, sub-wavelength lithography, and dynamic variations due to circuit behavior will look like inherent unreliability in the design. Increasing soft errors will increase intermittent error rate by almost two orders of magnitude. As transistors scale even further, degradation due to aging will become worse. We discuss these effects and propose solutions in microarchitecture, design, and testing, for designing with billions of unreliable components to yield predictable and reliable systems, with probabilistic and statistical design methodology.

1 Introduction

VLSI system performance has increased by five orders of magnitude in the last three decades, made possible by continued technology scaling. This treadmill will continue, providing integration capacity of billions of transistors; however, power, energy, variability, and reliability will be the barriers to future scaling.

Die size, chip yields, and design productivity have so far limited transistor integration in a VLSI design. Now the focus has shifted to energy consumption, power dissipation and power delivery [1]. Transistor sub-threshold leakage continues to increase, and leakage avoidance, leakage tolerance, and leakage control techniques for circuits have been devised [2]. As technology scales further we will face new challenges, such as variability [3], single event upsets (soft errors), and device (transistor performance) degradation—these effects manifesting as inherent unreliability of the components, posing design and test challenges. We will discuss these effects and propose solutions in microarchitecture, circuit, and testing, for designing with many unreliable components (transistors) to yield reliable system designs in the future.

This problem is not new; even today we design systems to comprehend reliability issues. For example, error correcting codes are commonly used in memories to detect and correct soft errors. Careful design, and testing for frequency binning, copes with variability in transistor performance. What is new is that as technology scaling continues, the impact of these issues keeps increasing, and we need to devise techniques to deal with them effectively.

Please use the following format when citing this chapter:

Borkar, S., 2007, in IFIP International Federation for Information Processing, Volume 249, VLSI-SoC: Research Trends in VLSI and Systems on Chip, eds. De Micheli, G., Mir, S., Reis, R., (Boston: Springer), pp. 69–79

2 Sources of Variation

Primarily there are two types of variations: static and dynamic. Static variations are caused by variations induced during processing, and the behavior does not change over time. Dynamic variations, on the other hand, are caused by the behavior of the chip while it is functioning. An example of static variations is Vt mismatch between two adjacent transistors caused during fabrication. Supply voltage droop generated during operation of a chip causing circuit to slow down is an example of a dynamic variation.

The variations could be within a die, or between different dies; they could be systematic or random. Systematic variations are primarily induced by limitations in the processing, and random variations are caused by randomness in physical dimensions such as line edge roughness and discreteness of dopant atoms.

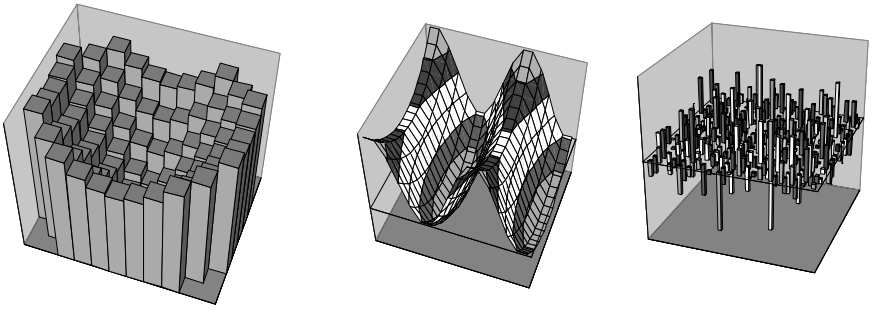

Fig. 1. Resist Thickness, Lens Aberrations, and Random placement of dopant atoms.

Figure 1 shows variation in photo resist thickness resulting in die to die variations, lens aberrations causing systematic variations, and random placement of dopant atoms causing random variations.

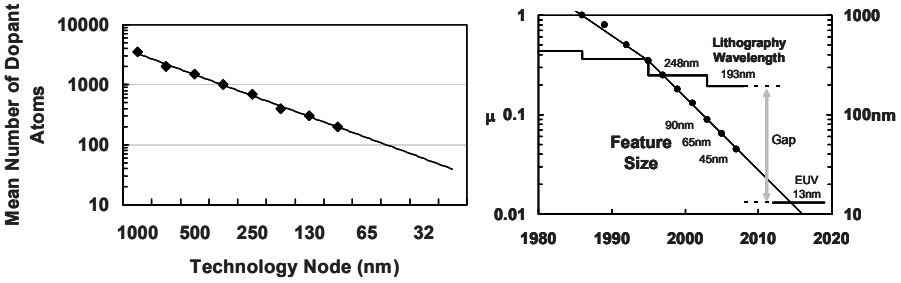

Fig. 2. (a) Mean number of dopant atoms in transistor channel decreases, and (b) sub-wavelength lithography until EUV.

Random Dopant Fluctuations results from discreteness of dopant atoms in the channel of a transistor [4]. Transistor channels are doped with dopant atoms to control their threshold voltage. Figure 2(a) shows dopant atoms in the channel of several generations of transistors. As a transistor scales in size each technology generation, its area

is reduced by half, and thus the number of dopant atoms in the channel reduces exponentially over generations. In one micron technology generation there were thousands of dopant atoms, whereas in 32 to 16 nm generation there will be only 10's of dopant atoms left in the channel, and the law of large numbers does not apply. Therefore, two transistors sitting side by side will have different electrical characteristics due to randomness in small number of dopant atoms, resulting in variability.

Another source of variability is due to sub-wavelength lithography shown in figure 2(b). Since 0.25µ technology generation, sub-wavelength lithography is being used for patterning transistors. For example, 248nm wavelength of light was used to pattern 0.25µ (250nm) and 0.18µ (180nm) transistors. The wavelength reduced to 193nm for 130nm technology, and since then it has remained constant until today for even 65nm transistors. There may be some additional breakthroughs to effectively reduce this wavelength (157nm light source or immersion technology) but the difference in the wavelength of light and the patterning width will continue to get wider until EUV (Extreme Ultra-violet, 13nm) technology becomes available. This sub-wavelength lithography is the primary reason for line edge roughness in transistors, and several other effects, resulting in variations.

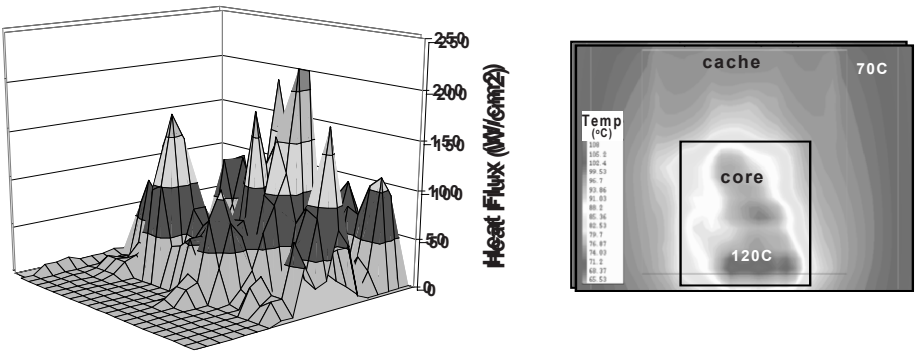

Fig. 3. Heat flux, and temperature variation across a microprocessor die.

Figure 3 shows heat flux (power density) across a microprocessor die, varying depending on the functionality of the circuit block. For example, the cache has less heat flux than an execution unit, and it also depends on the activity and compute load at any given time. Higher heat flux also puts more demand on the power distribution grid resulting in resistive and inductive voltage drops, creating time dependant, dynamic, supply voltage variations. Higher heat flux results in higher temperature, creating hot spots, and thus temperature variations across the die affecting circuit performance. This also results in higher sub-threshold leakage, variations in the leakage across the die, and dynamic variations in power delivery demand across the power distribution grid.

3 Impact of Variations on Products

Variations have been with us for a long time, and their impact on product was not profound in the past, but now it impacts performance, power, and yield. As technology scales, this impact will probably become worse [3].

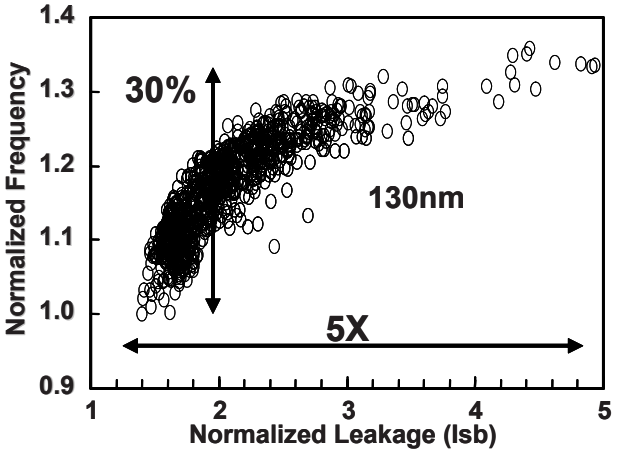

Fig. 4. Impact of variations on microprocessor frequency and leakage power.

Impact of variations on a product is shown in figure 4. It shows frequency and sub-threshold current measurements of about thousand microprocessors fabricated in 130nm technology, with 30% variation in the frequency distribution and about 5 to 10X spread in the sub-threshold leakage current. Since sub-threshold leakage power is a major portion (30-50%) of the total power consumption, 5-10X variation in the leakage power alone contributes to almost 50% variation in the total power. The behavior of the fabricated design in power and performance is different from what was intended, and hence the effect of variations looks like inherent unreliability in the design.

4 Variations in transistors

Variation in transistor length, width, and threshold voltage affect circuit performance, and thus overall performance of the VLSI chip. It also impacts yields and probability of meeting performance targets with yields.

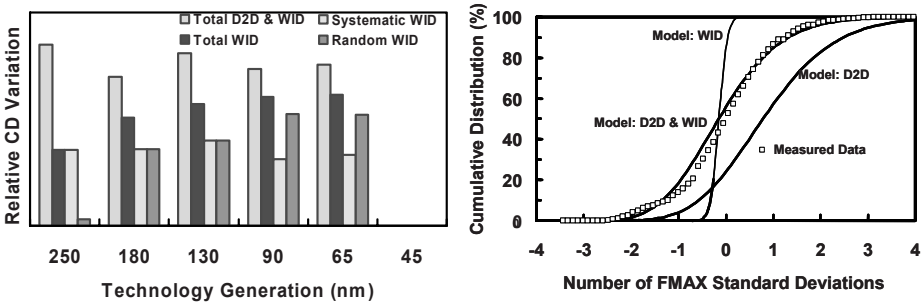

Fig. 5. (a) Gate length variation trend, and (b) within and die-to-die variation.

Figure 5(a) shows systematic and random, within-die and die-to-die gate length variation. The first bar shows the aggregate effect, and notice that it is almost a fixed percentage of the nominal gate length. However, random variations within die increase with gate length scaling. Figure 5(b) shows impact of die to die and within-die variations. Within-die variations impact Fmax mean, that is the mean of maximum frequency of operation, and die-to-die variations impact the variance of Fmax. So, loosely speaking, within-die variations reduce maximum frequency of operation, and die-to-die variations reduce the yield of the design at the maximum frequency.

5 Design Considerations

Design practice too impacts performance in the presence of variations as shown in figure 6.

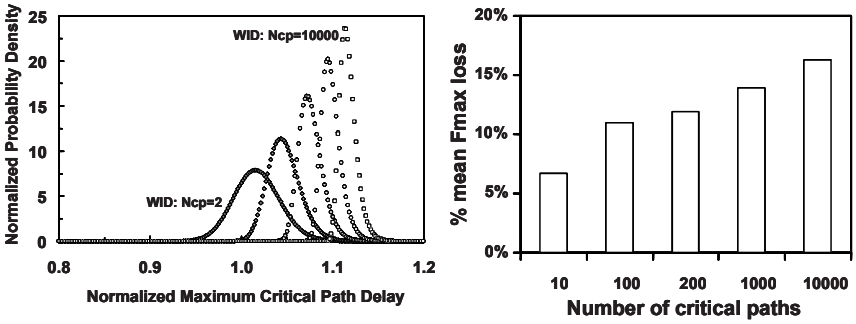

Fig. 6. Large number of time critical paths impact design performance.

Typically, the maximum frequency of operation of a design depends on a handful of time critical paths, and the number of such critical paths in the design plays important role in variation tolerant design. Notice that with a large number of critical paths in a design, the maximum critical path delay increases, thus lowering mean Fmax, but the variance also decreases. Reduction in the mean Fmax with the number of critical paths is logarithmic, that is, the reduction in Fmax slows down as the number of critical paths increase.

Microarchitecture also plays an important role in tolerating variations in a design. For example, microarchitecture with small number of gate stages in a clock period yields higher frequency of operation, compared to one with large number of gates. Both may provide equivalent logic throughput; however, with different variation tolerance. Large numbers of gates in a clock cycle tend to average out the effect of random variations, as shown in figure 7.

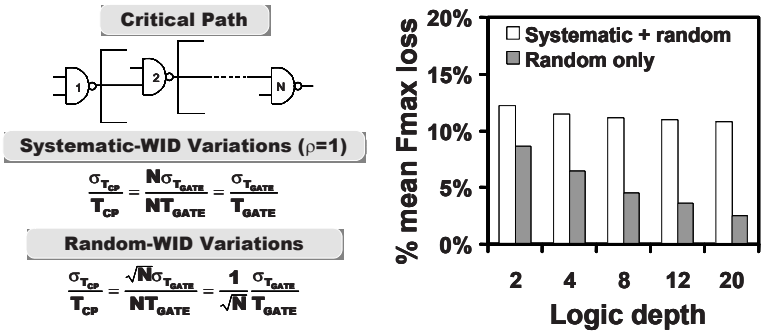

Fig. 7. Impact of random variations reduces with increasing logic depth.

6 Variation Tolerant Design

Numerous process technology, circuit, and architectural solutions have been proposed to deal with variations [3,5,6], which may require radical changes in our design methodology. For example, forward and reverse body bias can be used to tighten sub-threshold leakage and frequency distributions. Chips with higher leakage tend to be faster, hence reverse body bias may be applied to reduce the leakage and reduce the frequency. Similarly, slow chips can benefit from forward body bias to improve their speed at the expense of moderate increase in sub-threshold leakage power. Similarly, adaptive supply voltage too can be used in conjunction with body bias to tighten the distribution [6].

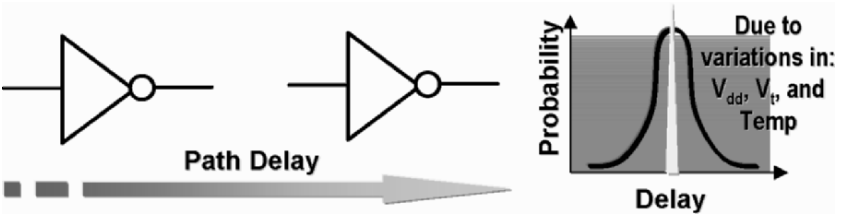

Fig. 8. Modeling and simulation of a critical path.

The chip frequency depends on the speed of the critical paths. Hence a critical path is typically modeled to have deterministic delay as predicted by a circuit simulator, as shown in figure 8; however, due to static and dynamic variations discussed before, the delay of the circuit is probabilistic. When the design is complete and conventional design methods are applied, transistors are typically down sized to save active power. As a result, transistors in the critical paths will also get down sized adequately. While every attempt is made during manufacturing to maintain the deterministic behavior, the increased variability, or not fully comprehended variability in transistor performance, can make these path delays probabilistic. Therefore, we need to deviate from conventional methodology of down-sizing transistors indiscriminately to reduce active power, because down-sizing indiscriminately makes many non-critical paths critical, and reduces probability of meeting the frequency goal.

Similarly, low threshold voltage transistor usage does not have to be minimal to reduce leakage power. With reduced low Vt usage across the design, the transistors near the critical path too may get replaced with high Vt transistors, and due to variations in the threshold voltage these paths could become slower, resulting in wider frequency distribution. Design tools and methodologies need to comprehend variations, and optimize the design not for frequency alone, but for active and leakage powers and their distribution.

When a micro-architecture is designed, the tendency is to improve frequency of operation by creating more critical paths, which reduces the probability of meeting the increased frequency goal. Furthermore, to meet higher frequency goals, micro-architecture tends to employ less number of gate-delays in a clock cycle. Since less number of gates in a clock cycle does a poor job of averaging and canceling the effects of variations, it results in reducing the probability of meeting the frequency goal. This once again, is contrary to conventional thinking and design methodology.

We need to evolve from today's deterministic design to probabilistic and statistical design for the future, comprehending variations, and optimizing for yield, performance, and power (figure 9).

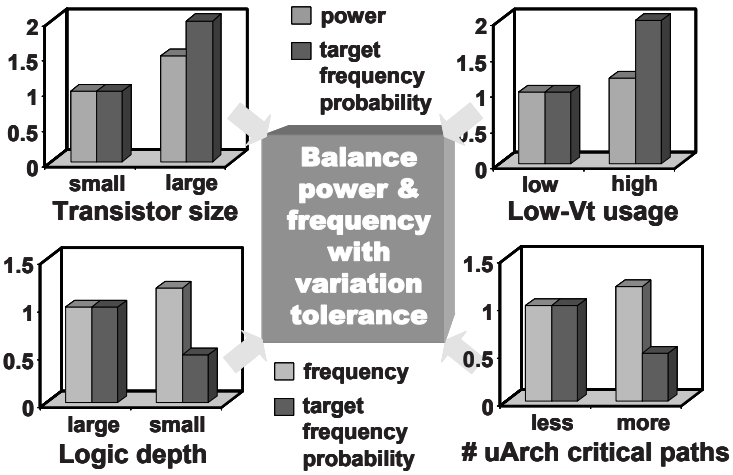

Fig. 9. Variation tolerant design methodology to optimize yield, performance, and power.

7 Longer Term Outlook

As technology continues to scale further, both static and dynamic variations will continue to get worse for the reasons discussed before, resulting in wider distribution of characteristics of the transistors. These variations in the transistors could be severe enough that it would be impossible to correct for them during design—they will have to be compensated somehow at the whole system level.

Single event upsets (soft-errors) are another source of concern. These errors are caused by alpha particles and more importantly cosmic rays (neutrons), hitting silicon chips, creating charge on the nodes to flip a memory cell or a logic latch. These errors are transient and random. These errors are relatively easy to detect and correct in memories by protecting the memory bits with parity, and employing error correcting codes. However, if such a single event upset occurs in a logic flip-flop then it is difficult to detect and correct.

We expect a modest increase in soft-error rate per logic state bit each technology generation [7]. Since the number of logic state bits on a chip double each technology generation (following Moore's Law), the aggregate effect on soft-error rate FIT (failure in time) could be almost 100X by the next few generations.

Aging has had significant impact on transistor performance. Studies have shown that transistor saturation current degrades over years due to oxide wear-out and hot carrier degradation effects. So far, the degradation is small enough such that it can be accounted for as an upfront design margin in the specification of a VLSI component. We expect this degradation to become worse as we continue to scale transistor

geometries beyond. It may become so bad that it would be impractical to absorb degradation effects upfront in a system design.

As gate dielectric scales, gate leakage will increase exponentially, and we fear that burn-in power will become prohibitive, making burn-in testing obsolete. Therefore, screening for defects and infant mortalities in VLSI chips will become increasingly difficult if not impossible. One-time factory testing will be insufficient, and what you need is the test hardware embedded in the design, to dynamically detect errors, isolate and confine the faults, reconfigure using spare hardware, and recover on the fly.

8 Paradigm Shift

There are several potential solutions in sight in all discipline to tackle most of the problems discussed before; however, all disciplines of VLSI will have to make concerted efforts to make this successful.

A shift from deterministic design to probabilistic and statistical design would ease impact of transistor variations on circuit performance. Today's design optimizations are performed with only one or two objectives, namely performance and power. This will have to change, with multi-variable design optimizations, comprehending performance, active & leakage power, reliability, yield, and bin-splits. Design tools to implement such optimizations, and statistical and probabilistic methodologies to go along with the tools need development.

In circuit design, replacing regular flip-flops by soft-error tolerant hardened flip-flops will improve soft-error rate tolerance by almost 10X. To catch dynamic errors, innovative techniques such as Razor [8] need serious consideration which will not only detect and correct errors, but will also allow the design to operate at optimum power and performance. This technique is power efficient because it replicates only those flip-flops that are critical and need to be checked for correctness, and is also capable of catching circuit marginalities arising from variations.

At the system architecture level, functional redundancy check may work; however, it may not be power and energy efficient, since it almost doubles the hardware and power consumption for the same performance. Any redundancy and checking hardware must be used judiciously to dynamically catch errors and take corrective action, and should not burden the system with excessive power consumption and complexity. An interesting microarchitecture is proposed in [9], where a traditional processor core is accompanied by a small, yet robust, core as a checker. The checker core is correct by construction, may be over designed to be variation tolerant and is made immune from any further errors—both static and dynamic. Since the checker core is small it consumes very little power, and can dynamically detect and correct any errors made by the large core, thus providing reliable operation of the system.

9 Multi—a Solution to Variability and Reliability

The key to the variability and reliability problems may be to exploit abundance of transistors using Moore's Law to your advantage. Instead of relying upon higher and higher frequency to deliver higher performance, a shift towards parallelism to deliver higher performance is in order, and thus Multi- may be the solution at all levels—from multiplicity of functional blocks in a design to multiple processor cores in a system [10].

Multiple functional blocks, operating at lower voltage and frequency, provide the same logic throughput, but at much reduced power, and can be used for redundancy and error checking. For example two ALUs (Arithmetic and Logic Units) could be used to provide higher throughput when needed, and can be used to check and correct results produced by each other (figure 10). Multiple cores in a system will provide similar performance and redundancy benefit with functional redundancy checking employed at a coarse level of granularity.

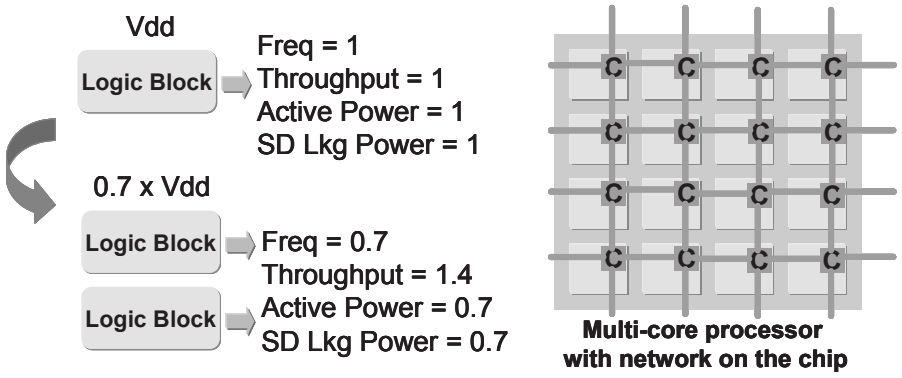

Fig. 10. (a) Multiple functional blocks, and (b) Multi-core processor for resiliency.

Test functionality can either be distributed as a part of the hardware, to dynamically detect errors, correct, and isolate aging & faulty hardware, or a sub-set of cores in the multi-core design can be used to do this task. This microarchitecture strategy, with Multi-cores to assist in redundancy, is called Resilient Microarchitecture, to continually detect errors, isolate faults, confine faults, reconfigure the hardware, and adapt. If such a strategy is made to work, then there is no need for one time factory testing or burn-in, since the system is capable of testing and reconfiguring itself to make itself work reliably throughout its lifetime.

All this is possible, but all disciplines from fabrication to software will have to cooperate and make the system reliable in spite of unreliable components. A lot of research and development needs to be done, however, to make this concept into a reality.

Acknowledgements

The author would like to thank Vivek De, Jim Tschanz, Ali Keshavarzi, Keith Bowman, Tanay Karnik, Peter Hazucha, and Jose Maiz, for their help and insightful discussions.

References

[1] Shekhar Borkar, "Design Challenges of Technology Scaling, IEEE Micro", July-August 1999.
[2] Shekhar Borkar, "Circuit Techniques for Subthreshold Leakage Avoidance, Control, and Tolerance", IEDM 2004.
[3] Shekhar Borkar et al., "Parameter Variations and Impact on Circuits and Microarchitecture", Proceedings of Design Automation Conference, 2003.
[4] Xinghai T et al., Intrinsic MOSFET parameter fluctuations due to random dopant placement, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Dec. 1997.
[5] Shekhar Borkar, "Probabilistic and Statistical Design—The wave of the future", VLSI-SOC, 2006.
[6] Jim Tschanz et al., "Effectiveness of adaptive supply voltage and body bias for reducing impact of parameter variations in low power and high performance microprocessors", IEEE Journal of Solid States Circuits, Volume 38, Issue 5, May 2003 Page(s):826-829
[7] Peter Hazucha et al., "Neutron Soft Error Rate Measurements in a 90-nm CMOS Process and Scaling Trends in SRAM from 0.25-μ to 90-nm Generation", IEDM 2003.
[8] Ernst D et al., Razor: a low-power pipeline based on circuit-level timing speculation, International Symposium on Microarchitecture, 2003. MICRO-36.
[9] Austin T, DIVA: a reliable substrate for deep submicron microarchitecture design, 32nd Annual International Symposium on Microarchitecture, 1999. MICRO-32.
[10] Shekhar Borkar, "Designing reliable systems from unreliable components: The challenges of transistor variability and degradation", IEEE Micro, Nov-Dec 2005.