RC23586 (W0504-054) April 12, 2005
Electrical Engineering

IBM Research Report

De-convolving Variability in Technology/Circuit Co-Design

Ruth A. Wang, Paul Friedberg
Department of Electrical Engineering and Computer Sciences
University of California at Berkeley
Berkeley, CA 94720

Kerry Bernstein, Dale Pearson, Mark B. Ketchen, Wilfried Haensch
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. Ithas been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distributionoutside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g, payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at http://domino.watson.ibm.com/library/CyberDig.nsf/home

De-convolving Variability in Technology/Circuit Co-Design

Ruth A. Wang[1], Paul Friedberg
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94720 USA
{ruthwang, pfriedbe}@eecs.berkeley.edu

Kerry Bernstein, Dale Pearson
Mark B. Ketchen, Wilfried Haensch
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598 USA
{kbernste, dale_pearson, mbketchen, whaensch}@us.ibm.com

Abstract—A SPICE-based Monte Carlo simulation methodology is described and used to analyze performance variability in 90nm PD-SOI circuits. Process and operating parameters of NAND chains and 16-bit adders are subjected to simulated variations in manufacturing process and operating conditions. Overall variability levels in delay and active power are compared across logic evaluation style, circuit complexity, and architecture. Individual parameter contributions to total variation levels are de-convolved; the most variation-sensitive parameters and designs are identified.

I. Introduction

The phenomenal success of CMOS technology is rooted in its scalability. Transistor counts traditionally double every 18 months, which greatly increases the functionality and productivity of each successive design. However, with the steadily decreasing feature sizes come new challenges that must be met. At the forefront of these is the trade-off between performance and power, because total power dissipation is limited by thermal constraints. In current leading edge technologies, passive power (leakage) has emerged as a significant fraction of the total chip power. Furthermore, the limited tolerances of the manufacturing process aggravate these tradeoffs, posing a major challenge to designers. For this reason, device solutions with improved performance at fixed leakage levels will likely propel future technology scaling. One such solution is the partially depleted silicon-on-insulator (PD-SOI) device, in which the Si device body sits atop an isolating oxide layer [1]. The body thickness is chosen such that the junctions of the device abut the isolating oxide layer, resulting in minimized junction capacitances and thus improved AC performance. However, the improvements gained from device and circuit design solutions are limited; manufacturing tolerances will inevitably set the ultimate operable circuit range.

In this work, we explore the use of Monte Carlo simulation techniques as a means of objectively assessing the robustness of logic circuit topologies in the face of increased scaling-induced variability. Variation in delay and active power is evaluated for a set of representative (canonical) circuits in 90nm PD-SOI, whose parameters are subjected to manufacturing process and operating variations. Further parameter contributions to total variability levels are de-convolved in order to quantify the sensitivity of each circuit to each varying parameter.

[1] Now with Intel Corporation.

II. Experimental Setup

A. Monte Carlo Simulation Framework

Monte Carlo analysis is a well-established technique [2] for exploring circuit performance sensitivities over a range of process and operating conditions. A generalized simulation flow used in this study is diagrammed in Figure 1. Each Monte Carlo simulation comprises a batch of SPICE simulations of a given circuit. For each SPICE simulation, parameter values are drawn randomly from their respective distributions to define a particular circuit instance, which is then simulated to measure its active power dissipation and delay. This process is then repeated, with the next circuit created by choosing and applying a new set of randomly selected parameter values.

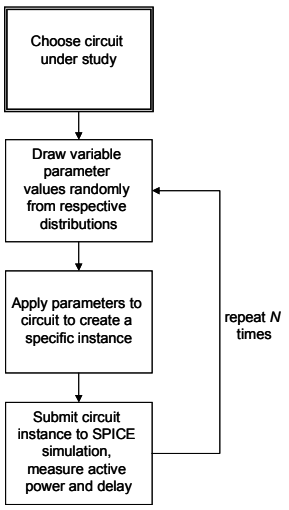

Fig. 1. Block diagram of SPICE-based Monte Carlo simulation flow.

Simulation time constraints limit either the number, *N*, of SPICE simulations per batch or circuit complexity. To achieve reasonable sampling of the model parameter distributions, this technique is limited to small scale circuits (i.e. on the order of hundreds of devices) and simulation batch sizes ranging from 200 – 1000 simulations. Distributions for

various process parameters (device width W, length L, threshold voltage V_{th} and SiO_2 thickness t_{ox}) are specified by the BSIM SOI model [3], with limits consistent with those predicted by the ITRS [4], and with those observed experimentally. The operating supply voltage V_{dd} is varied over a normal distribution with a nominal value of 1V and 3σ value of 50mV. The spatial correlation coefficient ρ is set to one (i.e. parameter values are the same for all devices in a given circuit instance). The operating temperature is held at the worst-case value of 85°C.

Each circuit instance is subjected to two sets of Monte Carlo simulations. In the first simulation, all five parameters are varied simultaneously to capture the full statistical distribution in overall delay and active power. In the second simulation, each parameter is isolated and varied individually, while all others are held at their nominal values. All interdependencies between parameters (e.g. V_{th} dependence on L, W, t_{ox}) are reconciled within the simulator, as asserted by BSIM models.

B. Circuits under Study

The circuits under study represent basic microprocessor datapath elements. Two circuit functions are chosen: a six stage chain of NAND gates and a family of 16-bit adders. Furthermore, each of these circuit types is designed using multiple logic evaluation styles, and for the adder family, different circuit architectures.

1) NAND Chains

The canonical NAND chain consists of six three-input gates, with all non-switching inputs tied to V_{dd}. In total, four NAND chains are designed based upon three logic evaluation styles: static CMOS, pulsed-static CMOS (PS-CMOS) [5], dynamic domino [6], and passgate (LEAP) [7]. Each NAND chain is submitted to $N = 1000$ SPICE simulations.

The output of the NAND chain is loaded with a static capacitor of value C_L = 10fF, consistent with the input capacitance of a typical stage. This static load is modeled as an ideal capacitor in the SPICE simulation; its value remains constant throughout each of the simulations and is unaffected by the random parameter selection process. In order to compare this scheme with one that models the fluctuating input capacitance of an active successive stage, a second loading condition is designed using fanout-of-three (FO3) loading [8]. Because the FO3 load contains active devices, each of its transistors is subjected to the same process parameter variations as the other gates that form the chain.

2) 16-bit Adders

In total, eleven 16-bit adders, which span a range of circuit architectures and logic evaluation styles, are designed and submitted to both sets of Monte Carlo simulations. The three basic architectures are: ripple carry adder with a passgate-based Manchester carry chain (static and dynamic) [7], logarithmic carry-select (static, dynamic, and passgate) [7] and carry lookahead (Kogge-Stone radix 2 and radix 4) [9], Han-Carlson [10], and Brent-Kung [11]. A fanout-of-four (FO4) static inverter loads the critical paths for all adder designs. Due to the increase in both logic complexity and transistor count as compared to the NAND chains, a reduced number of simulations (N = 200) is run for the adders. Although the sample size for adder simulations is thus smaller than for the NAND chains, it is sufficiently wide to reveal clear insights into the performance variability of various implementations.

3) Optimization of Transistor Sizes.

In order to conduct an unbiased comparison of the effects of process variability on designs within each circuit type, transistor sizes are objectively optimized for delay, given a fixed set of area and timing constraints. The specifics of these constraints differ for the NAND chains and the adders, as dictated by differences in circuit complexity; however the goal of objective sizing remains consistent for both types. All circuits presented in this study are optimized using an in-house software routine implementing a genetic algorithm.

III. Results and Analysis

The variability levels of all circuit performance metrics are calculated as the standard deviation (sigma) of the simulated value normalized to its corresponding mean value (σ/μ). 95% confidence intervals in this normalized measure of variability are denoted on each results plot in the form of error bars.

A. Delay and Power Variability

Figure 2 plots normalized coefficients of variability for the delay of NAND chains with static and FO3 loads. In the static capacitive loading case, static CMOS displays the most well-controlled delay variation levels, with a normalized variability of 6.4%, while LEAP suffers significantly greater variability, at 8.7%. The dynamic and pulsed static styles remain comparable to the static case with 6.7% and 6.8% variability, respectively. Similar results may be seen for FO3 loads.

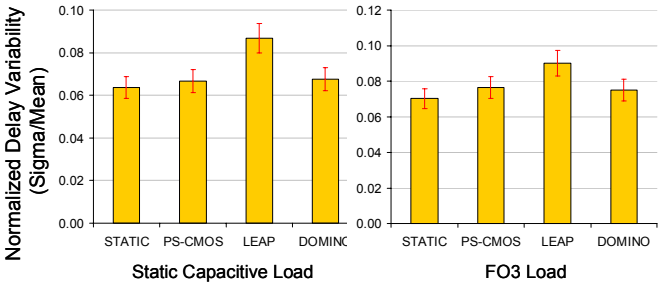

Fig. 2. Normalized delay variability of NAND chain.

The relative robustness of static CMOS is shown in Figure 3, which plots normalized power variation levels for the NAND chains. While the static CMOS implementation displays a normalized power variability of 4.3%, the LEAP style suffers the highest amount, at 5.7%. Variability of the dynamic and pulsed static styles remains lower than LEAP, at 4.6% and 5.1%.

Simulation results for the family of 16-bit adders indicate that the static implementation of the carry-select adder is the most resistant to delay variation (5.4%), as shown in Figure 4. Furthermore, while variability levels for most other static and dynamic designs fall within 20% of the static carry-select,

passgate families clearly suffer from the least amount of variation control. The three designs with the highest relative delay variabilities are the static ripple carry adder with passgate-based Manchester carry chain (7.1%), the passgate implementation of the carry-select (8.2%), and passgate-based radix 2 Kogge Stone (9.1%).

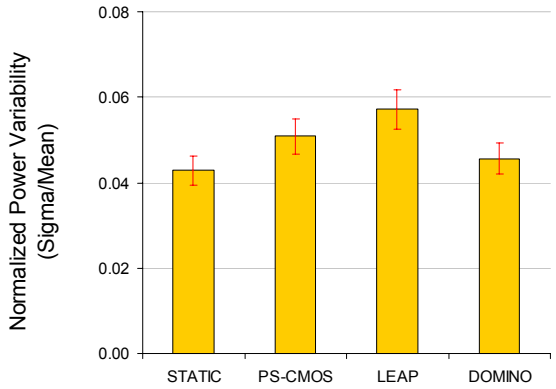

Fig. 3. Normalized power variability of NAND chain with static capacitive loading.

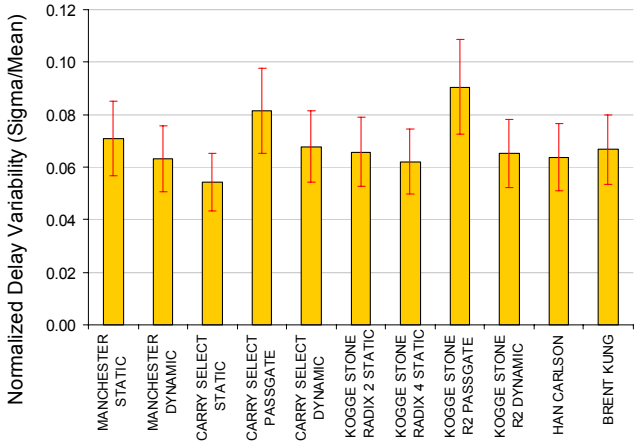

Fig. 4. Normalized delay variability of 16-bit adders.

Trends in adder power variability are shown in Figure 5. The static ripple carry adder using the Manchester carry chain displays the most predictable power values (3.8% variability), while the variation in other designs range between 22% and 137% higher. The two least-robust designs from a power perspective are the static, radix 2 Brent Kung (7.9%) and static, radix 4 Kogge Stone (9.1%) adders, each with spreads over 100% larger. This result may be attributed to the higher relative complexities of these designs, each having large intermediate capacitances along critical path nodes.

The Brent-Kung topology has widely varying internal fan-outs at each node, characteristic of its irregular tree structure, while the radix 4, Kogge Stone architecture has the tallest transistor stack height of all designs (four each of PMOS and NMOS). These loads are composed of internal capacitances of active transistors, which fluctuate according to variations in process parameters. During adder operation, the active power drawn to continuously charge and discharge these varying capacitances fluctuates correspondingly, resulting in the

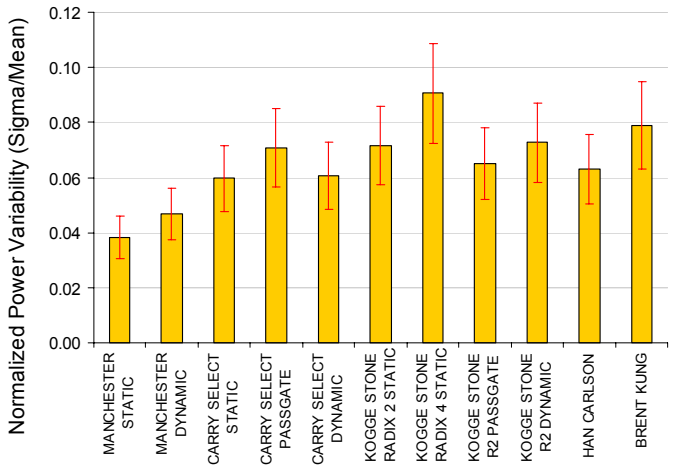

Fig. 5. Normalized power variability of 16-bit adders.

higher relative power variability for these complex architectures. In comparison, the more regular adder architectures display less performance spread.

Power-delay-products (PDPs) of all adders are compared in Figure 6, with raw values plotted with ±3σ error bars. According to these results, the adder implementation with the smallest mean PDP value is the dynamic ripple carry adder using a Manchester carry chain. In terms of normalized variability, this style displays the least-varying power and delay values. However, the ripple-carry architecture is least optimal from a speed perspective; the Han-Carlson implementation emerges as a much higher performance design with comparably high power and delay predictability.

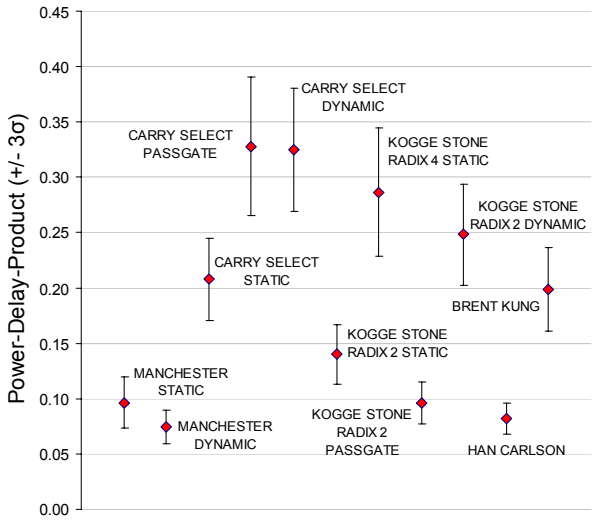

Fig. 6. Power-delay-product (PDP) of 16-bit adders.

B. Individual Parameter Contributions

Figure 7 and 8 show the normalized individual parameter contributions to delay variability for the NAND chains with FO3 loads and adders, respectively. The threshold voltage is found to be the most significant parameter in both cases, with an average contribution of 4.3% for the NANDs and 3.7% for the adders. Furthermore, the designs that are most sensitive to

variations in threshold voltage are the passgate-based styles: the LEAP implementation of the NAND chain and the four passgate adders (the static and dynamic Manchester chain implementations for the ripple adder and the passgate carry-select and radix 2 Kogge-Stone architectures) display the highest sensitivities to V_{th} variation.

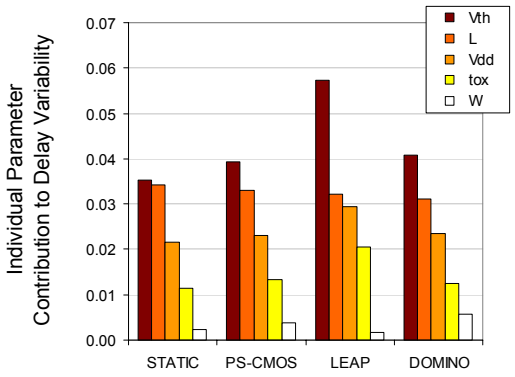

Fig. 7. Individual parameter contributions to delay variability of NAND chain with FO3 loading.

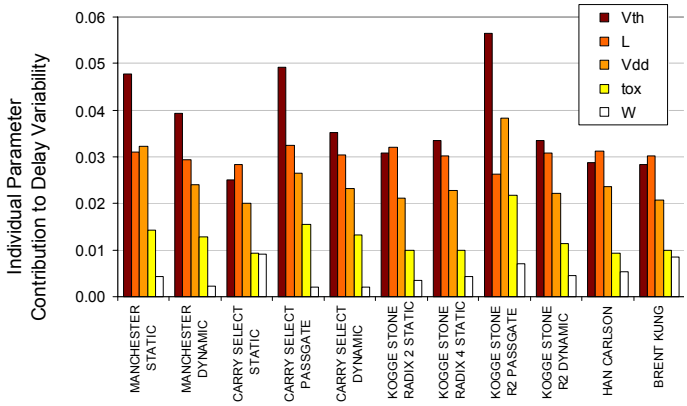

Fig. 8. Individual parameter contributions to delay variability of 16-bit adders.

Variation in gate length L is nearly as significant as V_{th} contributions, accounting for an average of 3% of the overall variability in both cases. Furthermore, supply voltage variations account for average contributions of 2.4% (NAND chains) and 3% (adders). The process parameters t_{ox} and W are typically very well-controlled, and contribute on average 1.4% and 0.3% for the NAND chains, and 1.2% and 0.5% for the adders. These results quantify the high sensitivity of delay to fluctuations in V_{th}, V_{dd} and L, consistent for NAND chains and the family of adders, across all logic evaluation styles. Clearly, efforts to impose tighter control over these three parameters during manufacturing and design processes would significantly improve the ability to control the range of transistor gate delays.

Variability in active power dissipation is affected by supply tolerance: Figure 9 shows average V_{dd} contributions of 4.7% for the adders. Fluctuations in V_{th} are also significant, accounting for 3.2% of the power spreads. Techniques for controlling V_{th} during manufacturing and for reducing V_{dd} noise during circuit operation improve the predictability of power dissipation.

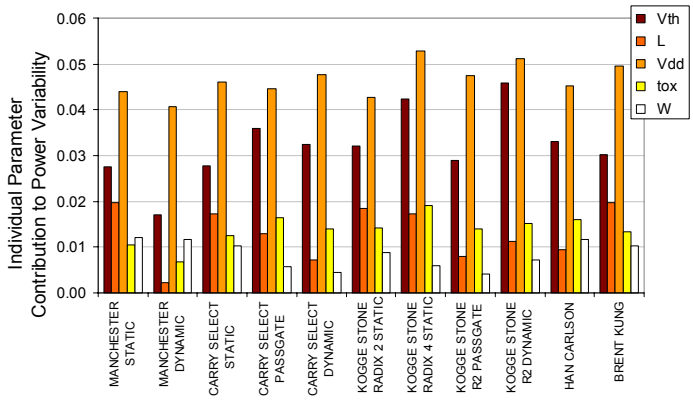

Fig. 9. Individual parameter contributions to power variability of 16-bit adders.

IV. Discussion

For the circuit topologies studied in this work, static CMOS circuits are most tolerant of parameter variation. Meanwhile, passgate-based circuits suffer 36% to 69% more delay spreads than corresponding static implementations, consistent with the observed significant dependence of delay variability on V_{th} variations. Power variation trends of adder circuits indicate dependence upon intermediate node fanout; designs with larger fluctuating capacitances on internal nodes generally yield the least predictable power, while designs with both fewer transistors and more balanced internal signal fanouts display the least amount of power fluctuation.

Total variability is approximated by the sum of five individual parameter contributions. The most significant contributors are identified as V_{dd}, V_{th} and L, accounting for an average total of 10% toward both delay and power variability, for all circuits studied. Among these three factors, V_{th} emerges as the most significant physical parameter affecting both delay and power, with contributions from L nearly as significant.

References

[1] K. Bernstein and N. Rohrer, *SOI Circuit Design Concepts*. Kluwer Academic Publishers, 2000.

[2] A. Hall, "On an experimental determination of PI," October, 1873.

[3] BSIM SOI Device Model; http://www-device.eecs.berkeley.edu/~bsimsoi/

[4] International Technology Roadmap for Semiconductors; http://public.itrs.net/

[5] C.-L. Chen and G. S. Ditlow, "Pulse Static CMOS Circuit," U.S. Patent no. 05495188, Feb. 1996.

[6] K. Bernstein et al., *High Speed CMOS Design Styles*, 1st ed. Kluwer Academic Publishers, 1998.

[7] J. Rabaey, A. Chandrakasan, B. Nikolic, *Digital Integrated Circuits*, 2nd ed. Prentice Hall, 2003.

[8] M. Horowitz, "VLSI Scaling for Architects," Presentation slides, Computer Systems Laboratory, Stanford University, 2000.

[9] P. Kogge and H. Stone, "A parallel algorithm for the efficient solution of a general class of recurrence equations," *IEEE Transactions on Computers*, 1973.

[10] T. Han and D. Carlson, "Fast area-efficient VLSI adders," in *8th Annual Symposium on Computer Arithmetic*. Como Italy, March 1982.

[11] R. Brent and H. Kung, "A regular layout for parallel adders," *IEEE Transactions on Computers*, 1982.