Mixed Swing Techniques for Low Energy/Operation Datapath Circuits

Ram Kumar Krishnamurthy

**A dissertation submitted to the
graduate school in partial fulfillment
of the requirements of the degree of**

**Doctor of Philosophy
in
Electrical and Computer Engineering**

**Carnegie Mellon University
Pittsburgh, Pennsylvania 15213**

December 1997

Abstract

The portable communications industry's vision of integrating a complete multimedia complex on a single die, coupled with the desktop computing industry's vision of integrating multimedia functionality into general-purpose microprocessors has transformed lowering the power dissipation of digital signal processing (DSP) datapath circuits into an increasingly important challenge in current and future fabrication processes. Fully-static CMOS logic accompanied with supply voltage scaling has enjoyed widespread usage in lowering datapath power dissipation over the last decade. However, fundamental limitations preclude device threshold voltage scaling under the constant drain-source field scaling paradigm in future deep-submicron processes, imposing limitations on voltage scaling. This has motivated a strong necessity for exploring new methodologies to lower the power dissipation of next-generation high-speed datapath circuits.

This thesis investigates *Mixed Swing* techniques for reducing the power dissipation of static CMOS datapath operators while retaining their high performance, or equivalently lowering their energy consumption per switching operation (energy/operation). Mixed swing techniques employ multiple operating voltages to implement standard datapath primitive functions by intermixing high- and low-voltage signal swings while driving interconnect and gate-fanout load capacitances at reduced voltage swings. Static and dynamic, single-ended and fully-differential mixed swing approaches are investigated to demonstrate the ability to voltage-scale more aggressively than static CMOS well into the deep-submicron regime.

Posynomial formulations for power and delay based on submicron MOS models are derived for mixed swing circuits to study and exploit the additional degrees of freedom available in their design space. On the basis of these models, optimization strategies for minimizing energy/operation are proposed and their efficiency is demonstrated on DSP datapath circuits. Worst-case process and temperature corner analyses are conducted to study low-voltage manufacturability and noise immunity challenges in mixed swing circuits. On-chip low-voltage series regulation approaches are developed to efficiently offset intra- and inter-die threshold variations, offering improved low-voltage manufacturability than full-swing static CMOS, while preserving high noise immunity. Further, on-chip series regulation eliminates the necessity for additional explicit off-chip supplies, transforming mixed swing techniques into a self-contained methodology which can replace full-swing static CMOS operating between a regular, high-voltage supply without warranting any technology or system-level modifications.

Experimental results showing substantial energy/operation savings are presented from (i) fabricated ICs and intensive circuit simulations on fixed-point DSP multiplier-accumulators over a range of operand bit-widths, power supply voltages, and commercial 0.5μm-0.16μm bulk-CMOS and fully-depleted SOI processes, and, (ii) data buses and multicast datapath nets of the floating-point units of two industrial next-generation multimedia-enriched microprocessors presently in design in a 0.16μm bulk-CMOS process.

Acknowledgments

I would like to express my sincerest gratitude to my advisor Prof. L. Richard Carley for his continuous inspiration, encouragement, and scholarly guidance throughout my education at Carnegie Mellon University. His erudite suggestions in helping me define and pursue my research have contributed invaluably towards my remaining on a productive path during my PhD study. I will forever cherish the innumerable discussions we have had envisioning future technological challenges in this rapidly evolving semiconductor industry.

I would also like to thank Prof. Rob Rutenbar, Dr. Herman Schmit, and Dr. Chris Nicol (Bell Laboratories, Holmdel, NJ) for taking an active part on my thesis committee and their invaluable inputs at critical junctures of this thesis. Rob's remarkable teaching and his presentation skills have been a constant source of inspiration to me. His sense of humor certainly made my PhD education an enjoyable one. Herman's remarkable patience listening to my ideas, the numerous suggestions for improvement, and his constructive critiques of my publications have gone a long way in helping me live up to the standards of industry-quality research. Chris's invaluable guidance in defining and solving

certain strategic problems as part of this thesis have tremendously helped me position my research better against competing work in the industry.

Interactions with the industry have played a vital role in this research. I would like to express my gratitude to Dr. Shekhar Borkar, Dr. Vivek De, and Dr. Soumya Krishnamurthy of Microprocessor Research Labs, Intel Corporation, for offering me the opportunity to investigate my ideas with them during Summer'97, and their invaluable inputs. I would also like to thank Prof. Andrzej Strojwas and Dr. Kimon Michaels (PDF Solutions. Inc.) for insightful discussions and commercial data on deep-submicron manufacturability, Dr. Balsha Stanisic (IBM Corporation) for commercial data on noise immunity, and Dr. Paul Davis (MIT Lincoln Labs) for help with fabricating our ideas.

Interactions with many colleagues of the SRC-CMU Center of Excellence for CAD, past and present, have made my PhD study an enriching experience. I would like to acknowledge the many insightful hallway discussions with Bulent Basaran, Chris Inacio, Pascal Meier, Tamal Mukherjee, Nitzan Weinberg, and Nick Zayed among many others. I would like to thank Cindy Meyers for her invaluable help with the layout of our datapath standard cell libraries.

Last but not least, I would like to thank my parents for being a constant source of moral support from continents across, and for firmly imbibing into me from a very young age that *perseverantia omnia vincit* - it is this perseverance that kept me going. This thesis is dedicated to them.

This work was supported by the Defense Advanced Research Projects Agency (Order A564), the National Science Foundation (Grant MIP9408457), and the Semiconductor Research Corporation (Contract 068.007). The U.S. government has certain rights to this material.

Table of Contents

List of Figures

1 Introduction

1.1 Motivation

There has been an accelerated consumer market demand over the last decade for portable communication devices with more and more multimedia functionality (e.g., bidirectional motion video, handwriting and voice recognition etc.) integrated onto them. Fueled by rapidly scaling feature sizes into the sub-0.25μm era, this has led to the vision of integrating a complete multimedia complex on a single die [Sasaki96], [Borel97]. With the major limitation to portability being battery space and weight, this has made lowering the power consumption of portable multimedia devices an increasingly important challenge in current and future technologies, in order to prolong battery life between successive charges. A majority of portable multimedia devices are essentially Digital Signal Processing (DSP) circuits interfacing with information from the real-world environment and/or human operators, and so there exists a strong motivation to minimize the power consumption of DSP circuits. In addition, DSP tasks, especially real-time applications, require maintaining a fixed rate of computation or throughput, and there exists no freedom to perform the computations at a slower rate (or motivation to perform them at a faster rate). This makes it imperative to achieve the low-power objectives without sacrificing performance.

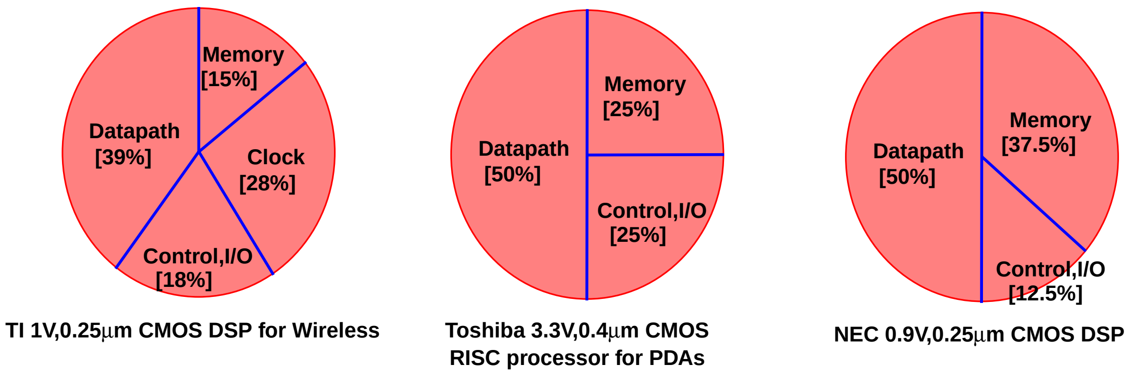
While the trend towards low-power has predominantly been driven by portability limitations, the desktop computing industry has also contributed to this trend. The growing integration of multimedia functionality onto general-purpose microprocessors coupled with rapidly increasing integration density has pushed integrated circuit (IC) power density (measured a IC power consumption per unit die area) to extreme limits making on-board heat dissipation a challenging and costly task. In addition, high power densities contribute to an increase in the junction and substrate temperatures which aggravates several high-temperature failure mechanisms such as thermal runaway, junction fatigue, and electromigration, causing an exponential degradation in the component's reliability with time [Chatterjee95]. These factors have made power reduction of multimedia-enriched microprocessors targeted for desktop markets as well a top priority in their traditional performance-area-reliability design space.

1.2 Thesis Focus

A majority of DSP circuits (e.g., Finite Impulse Response filters, convolution kernels etc.) are essentially signed, fixed-point datapath operators, specifically multiplications and/or accumulations. At the heart of a majority of DSP datapath is a multiplier-accumulator (MAC), typically short bit-width (8 - 24-bits), since this operand range dominates most DSP applications. The MAC lies directly on the critical circuit delay path and hence determines the operating clock frequency; many DSPs characterize their performance in terms of the number of MACs performed per second [Allen85], [Lapsley96]. Further, datapath operators display high switching activity due to both inherently high static transition probabilities and considerable amount of spurious transitions due to dynamic hazards. The high activity factors coupled with their high throughput requirements makes datapath power, essentially dominated by the MAC power, a substantial portion of total power of DSPs. Figure 1 illustrates this trend for three commercial CMOS DSPs and general-purpose RISC processors targeted for DSP applications: the datapath power component ranges from 39% [Wailee97a], [Wailee97b] up to 50% [Nagamatsu95], [Izumikawa97] of their respective total power. Therefore, there exists a strong necessity to focus attention on lowering the power consumption of DSP datapath circuits in general, and MAC circuits in particular.

The primary focus of this work is to investigate approaches to lower the *energy/operation* of datapath operators that are widespread in DSP applica-

FIGURE 1 Datapath circuit power dissipation trend of commercial DSP/RISC processors.



TI 1V,0.25µm CMOS DSP for Wireless

Toshiba 3.3V,0.4µm CMOS RISC processor for PDAs

NEC 0.9V,0.25µm CMOS DSP

tions. Energy/operation is defined as the energy consumed by a digital circuit per switching operation, or equivalently, the product of its power dissipation and operating clock period. The key challenge addressed in this thesis is to investigate approaches to minimize datapath circuit power dissipation while maintaining their high, target clock frequency specifications.

1.3 Research Overview

This thesis explores *Mixed Swing* techniques that enable more aggressive voltage scaling than fully static CMOS in order to reduce the energy/operation of datapath circuits in standard submicron bulk-CMOS and SOI fabrication processes. Mixed swing techniques employ multiple power supply voltages in order to expand the degrees of freedom available in the power-performance design space of static CMOS circuits. Standard digital logic gates are implemented in multiple stages by intermixing high- and low-voltage swing signals (hence the name *Mixed Swing* techniques), while driving interconnect and fanout load capacitances at low voltage swings. As we will show in Chapter 4, this allows the digital circuit designer to simultaneously exploit the best aspects of both static CMOS and voltage scaling, while preserving noise immunity and improving low-voltage manufacturability across worst-case process and temperature variations.

This thesis work is classified broadly into four focus areas. We now briefly discuss each of our focus areas and summarize their salient features.

1. **Mixed Swing Techniques - Gate Architectures:** Static CMOS-, Domino/ Pass-Transistor Logic-, and Cascode Voltage Switch Logic-based mixed

swing techniques are explored to construct standard datapath primitive gates. A fully static, single-ended, four-power-supply-rail methodology called *Mixed Swing QuadRail* presented here is shown to offer substantial energy/operation savings on datapath circuits with interconnect capacitance dominance, e.g., Wallace tree multipliers. A Domino/Pass-transistor Logic-based, single-phase clocked, single-ended methodology and a CVSL-based, fully static, fully-differential methodology presented here are shown to offer substantial energy/operation savings on datapath circuits with gate capacitance dominance, e.g., adders. The ability of these techniques to voltage-scale more efficiently than static CMOS well into the submicron regime, without warranting any specific technology modifications, is demonstrated through measurements on a test-chip and intensive HSPICE simulations. Further, in order to avoid explicit off-chip multiple power supplies, a series regulation technique for Mixed Swing QuadRail technique with sleep-mode control is developed. This approach efficiently generates on-chip Mixed Swing QuadRail's reduced swing power supply, making it a self-contained methodology. In addition, this is shown to significantly improve low-voltage manufacturability compared to full-swing static CMOS.

2. **Mixed Swing Techniques - Modeling and Optimization:** Mixed swing techniques perform multi-staged logic by employing multiple power supplies. Therefore, additional degrees of freedom are introduced into their power-delay optimization space. In order to explore this design space, posynomial power and delay formulations for Mixed Swing QuadRail are

developed using the n^{th}-Power Law submicron MOSFET model. The accuracy of these models are validated through HSPICE simulations. Based on our models, optimal voltage scaling and transistor sizing approaches are developed to minimize energy/operation of mixed swing circuits. The importance of employing these optimization approaches, particularly in future low-voltage technologies, is motivated through experimental results from a 16*16+36-bit Booth-recoded, Wallace-tree DSP multiplier-accumulator (MAC) in a commercial 3V, 0.5μm bulk-CMOS process.

3. **Mixed Swing Techniques - Low-voltage Challenges:** Two of the most critical low-voltage practicality challenges to mixed swing techniques are addressed - noise immunity and manufacturability:

- **Manufacturability:** Intra- and inter-die variations in device parameters across process and temperature corners cause substantial dispersions in power and delay of static CMOS circuits at reduced voltages. The variations are escalating at least linearly with scaling feature sizes contributing significantly to low-voltage parametric yield loss. Worst-case process and temperature corners are developed and a relative manufacturability analysis is performed on static CMOS and Mixed Swing QuadRail. The analysis is conducted in the 0.5μm process on the same 16-bit DSP MAC mentioned above, over a range of operating voltages. Improved dynamic control of intra- and inter-die threshold voltage variations is demonstrated by the series regulated Mixed Swing QuadRail approach at the cost of a small layout area penalty.

- **Noise immunity:** As feature sizes continue to scale rapidly, noise immunity of deep-submicron digital circuits, particularly at reduced power supply voltages, has become a metric of comparable importance as performance and power. This is particularly a concern in mixed swing techniques because of the reduced voltages across gate inputs, causing absolute noise margins to be lower than that of full-swing static CMOS circuits. However, at reduced voltages, primary sources of digital circuit noise are also scaled at least linearly. Worst-case process, temperature, and noise corners are developed and a relative low-voltage noise immunity analysis is performed on static CMOS and Mixed Swing QuadRail. The analysis is conducted in the 0.5µm process on the same 16-bit DSP MAC mentioned above. It is demonstrated that both methodologies possess adequately high noise immunity.

4. **Mixed Swing Techniques - Performance Analysis:** Two types of datapath circuits are studied to compare the power-delay space of mixed swing techniques with static CMOS:

- Fixed-point, signed (2’s complement), short bit-width DSP MACs are investigated to demonstrate the potential for energy/operation savings - because of the simultaneous power and performance bottleneck presented by MACs, they are a good vehicle to study both datapath-level and processor-level impact on DSP energy/operation. Power-delay comparisons between Static CMOS and Mixed Swing QuadRail approaches are performed through fabricated MACs and intensive HSPICE simulations. The analyses are conducted over a range of:

(i) MAC operand bit-widths that dominate industrial DSPs (8 - 24 bits) in order to study the energy/operation savings impact due to datapath width.

(ii) operating power supply voltages in order to study the energy/operation savings impact due to voltage scaling.

(iii) commercial submicron process generations: 0.5µm bulk-CMOS, 0.35µm bulk-CMOS, 0.25µm fully-depleted SOI, and 0.16µm bulk-CMOS processes, in order to study the energy/operation savings impact due to technology scaling.

- Static CMOS vs. mixed swing techniques power comparisons are performed on point-to-point data buses and multicast datapath nets within the floating-point units of two industrial next-generation microprocessors with extensive multimedia support, presently in design. The analyses are conducted in a commercial 0.16µm bulk-CMOS process using industrial circuit simulators over a range of operating power supply voltages and input data switching activities for target clock frequency specifications.

1.4 Thesis Organization

We now present the details of our approach to lowering the energy/operation of datapath circuits. The organization of this thesis is as follows.

Chapter 2 discusses the evolution of static CMOS as one of the most popular choices for high-speed/low-power DSP circuits. Previously published techniques for lowering the power consumption of static CMOS digital circuits are reviewed, with a technological update on the latest developments in this area.

Advantages and limitations of these approaches are examined with a special emphasis on their applicability in future deep-submicron processes.

Chapter 3 examines architectural choices for high-speed/low-power MAC circuits. A commonly used DSP MAC architecture is formulated that will be the focus of further study. Power-delay tradeoffs within MAC circuits are investigated, exploring opportunities for lowering their energy/operation.

Chapter 4 introduces the concept of multiple power supply voltage-based low-power digital circuit design. Previously published research on low-power multiple voltage techniques are discussed. The proposed mixed swing techniques are then described, motivating the usage of multiple voltages at the gate-level to construct standard datapath primitives. The ability of these techniques to voltage scale more effectively than static CMOS without requiring any process modifications is demonstrated. Advantages and limitations of these techniques are enumerated, and classes of datapath circuits that would best benefit from these techniques are proposed.

Chapter 5 explores the design space of mixed swing methodologies. Analytical power and delay models are derived, and power-delay tradeoffs are studied. Optimal voltage scaling and transistor sizing techniques are developed and experimental results are presented to demonstrate their effectiveness.

Chapter 6 investigates two of the most important low-voltage practicality challenges to mixed swing techniques viz., manufacturability and noise immunity. Rigorous worst-case manufacturability and noise immunity analyses are performed on DSP MACs relative to static CMOS. For improved low-voltage dynamic control of threshold voltage variations, a series regulation technique is developed for Mixed Swing QuadRail, demonstrating improved manufacturability over static CMOS.

Chapter 7 describes detailed power-delay space comparisons between static CMOS and mixed swing techniques on various DSP datapath circuits. Through fabricated datapath integrated circuits and intensive circuit simulations, the ability to achieve substantial energy/operation savings over a range of DSP operand bit-widths and operating voltages in current and future deep-submicron processes, without warranting any technology or system-level modifications, is convincingly demonstrated.

Finally, Chapter 8 summarizes the contributions of this thesis work. This is followed by a discussion of future directions to this research.

2 Background: Static CMOS Low-Voltage Design

In the design of low-power digital circuits, a key requirement is to avoid logic families that consume *extraneous* power, i.e., additional power dissipated than what is required to charge/discharge the capacitive load at the gate outputs to perform the logic function. This may be due to (i) a static totempole current path between the power and ground rails of every gate such as in ratioed logic families, or, (ii) the requirement of additional input-data-unrelated switching signals to perform the logic gate's function, such as in dynamic/clocked logic families [Bakoglu90]. Such techniques have traditionally been employed in high-speed digital circuits, where contrary to low-power design objectives, total power, much less extraneous power, is not a design issue.

The fully static CMOS methodology has evolved as one of the most popular techniques for lowering the power consumption of digital circuits in general, and datapath circuits in particular [Gray94], [Chandra95]. This is primarily because it demonstrates the lowest extraneous power dissipation among existing logic families. In addition, its superior low-voltage power-speed characteristics and high noise margins have been the driving factors towards its widespread usage. Unfortunately, simply employing static CMOS does not solve the problem of lowering the power consumption for datapath circuits - future technologies still demand more than an order of magnitude reduction in the power consumption of industrial DSPs and multimedia-enriched processors [Sasaki96], [Borel97]. Therefore, there exists a strong necessity to explore techniques for substantially lowering power dissipation of static CMOS datapath circuits.

We begin this chapter with a review of static CMOS power components and discuss previously reported techniques to lower its power consumption. Other proposed alternate circuit techniques for higher-speed and lower-power than static CMOS are also presented. Advantages and limitations of these approaches are identified, motivating the need for further exploration of methodologies to lower static CMOS power.

2.1 Static CMOS Power Components

In order to understand the evolution of static CMOS as one of the most popular low-power design approaches, we will first examine the sources of static CMOS power dissipation. The total power consumed by a static CMOS circuit consists of three components, given by the following expression:

$$P_{total} = P_{dynamic} + P_{short-circuit} + P_{static} \qquad \text{(EQ 1)}$$

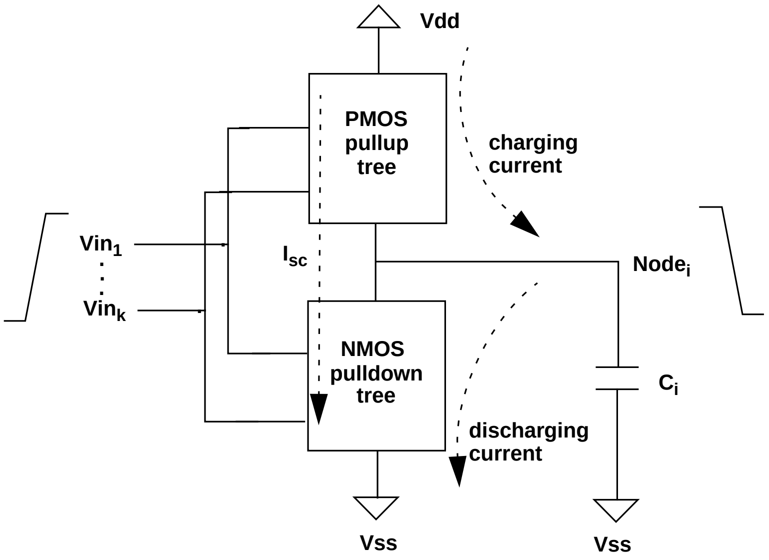
$P_{dynamic}$ represents the dynamic or switching power, i.e., the power dissipated in charging/discharging the physical load capacitance contributed by fanout gate loading, interconnect loading, and diffusion-substrate junctions at the CMOS gate outputs. C_i represents this capacitance at node i, lumped together as shown in Figure 2. For a static CMOS circuit with N switching nodes, operating at a clock frequency of f_{clk}, the dynamic power is given by [Chandra95]:

$$P_{dynamic} = \sum_{i=1}^{N} \alpha_i \cdot C_i \cdot V_{dd} \cdot V_{swing} \cdot f_{clk} \qquad \text{(EQ 2)}$$

where V_{dd} is the power supply voltage,V_{swing} is the voltage swing across the load capacitance which for a static CMOS gate is the same as V_{dd}, α_i is the switching activity at node *i* such that the product $\alpha_i.C_i$ is known as the effective switched capacitance per cycle at node *i*.

$\mathbf{P_{short\text{-}circuit}}$ represents the short-circuit power, i.e, the power consumed during switching because of a totempole current path between the power supply and ground, which exists for a short period of time during switching because of the finite input rise and fall times. Specifically, when the transition-

FIGURE 2 Static CMOS dynamic and short-circuit currents.



ing input voltage satisfies the condition $V_{tn} < V_{in} < V_{dd} - |V_{tp}|$ (V_{tn} and V_{tp} are the NMOS and PMOS device threshold voltages), there exists a conducting path between V_{dd} and ground as shown in Figure 2, during which both the NMOS and PMOS devices conduct simultaneously causing the short-circuit current I_{sc} to flow. This short-circuit power is given by [Sakurai90]:

$$P_{sc} = \alpha_i \cdot \frac{1}{n+1} \cdot \frac{1}{2^{n-1}} \cdot \frac{\beta}{2} \cdot (V_{dd} - (V_{tn} + |V_{tp}|))^{n+1} \cdot t_T \cdot f_{clk} \quad \text{(EQ 3)}$$

where, *n* is the velocity saturation index, typically between 1.0-1.5 in submicron processes, β is the transconductance gain factor of the pullup/pulldown transistor stack, and t_T is the input rise/fall time.

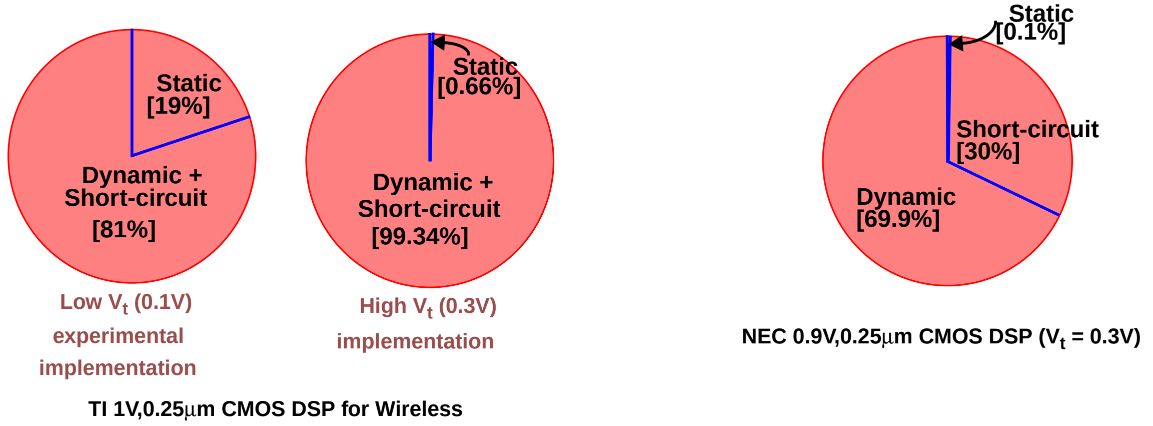
P_{static} represents the static power, i.e., the power dissipated even when there is no switching activity within the circuit. This is due to the leakage currents of the reverse-biased parasitic p-n junctions formed between the MOSFET drain and source diffusions to the substrate and well. These currents flow even when the devices are in cutoff region of operation, contributing to a constantly flowing static current between V_{dd} and ground. If I_s is the reverse saturation current of the source/drain p-n junctions, the static power is given by [Bakoglu90]:

$$P_{static} = I_{leakage} \cdot V_{dd} = I_S \cdot \left(e^{\frac{V_{rev}}{V_T}} - 1 \right) \cdot V_{dd} \quad \text{(EQ 4)}$$

where, V_{rev} is the reverse bias on the junction diodes and $V_T = KT/q$ is the thermal voltage.

Since several closely related parameters impact the three components of static CMOS power, depending on the specific circuit configuration, operating conditions, and fabrication process, any or all of these components may dominate total power. However, in a majority of static CMOS datapath circuits, dynamic power is the dominant component of total power, primarily because operating voltage has a full quadratic impact on it [Chandra95]. In addition, datapath operators display high switching activities due to their intrinsically high static transition probabilities and spurious/glitching transitions [Landman93], [Chandra95], [Favalli95], [Nagamatsu95], [Najm95]. This makes their effective switched capacitance per cycle substantial. These factors, coupled with their high-throughput demands, accounts for the dynamic power dominance. Short-circuit power also contributes significantly to total power, primarily because of the high switching activities and throughput requirements [Izumikawa97]. Since leakage currents are typically of the order of few nA/µm width of the transistors, the static or non-switching power is typically a few orders of magnitude smaller than dynamic power. Figure 3 demonstrates this

FIGURE 3 Dynamic, short-circuit, and static power dissipation trend of DSP processors.



trend on two commercial CMOS DSP processors [Wailee97a], [Izumikawa97], where the dynamic and short-circuit power components, in that order, dominate total power dissipation.

2.2 Voltage Scaling

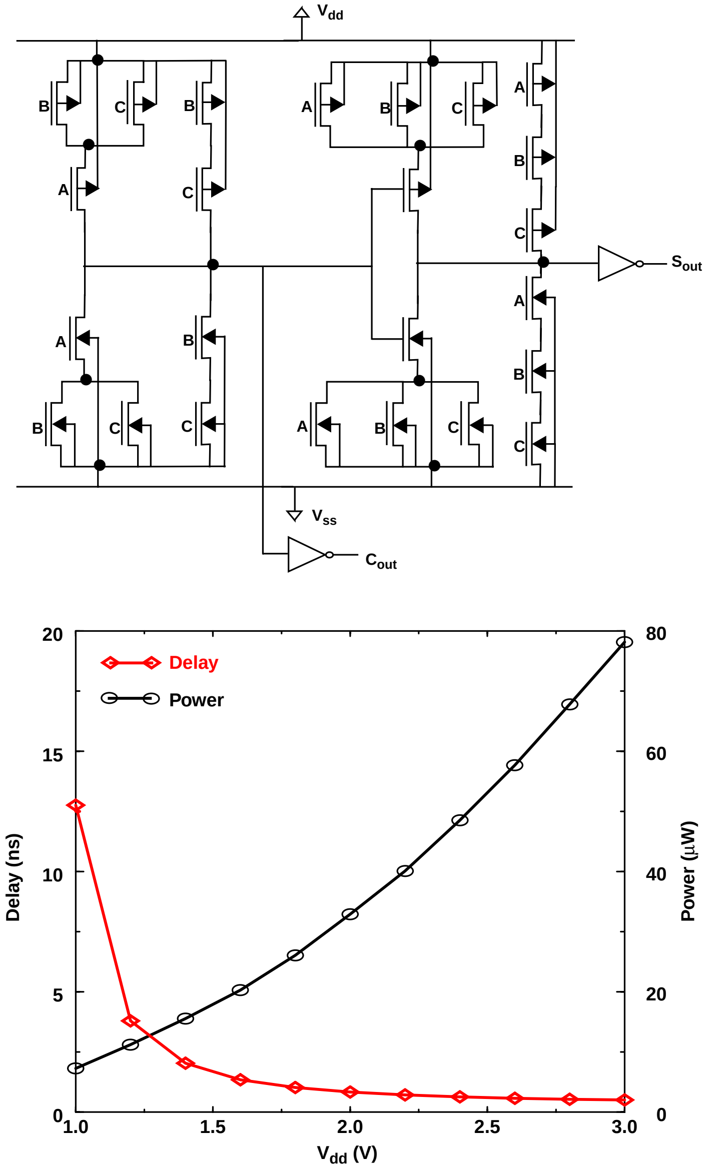
Voltage scaling, i.e., lowering the operating voltage below the maximum process-permitted voltage, has evolved as the most popular approach to lowering the power consumption of static CMOS circuits [Gray94], [Chandra95]. This, to some extent, is fairly obvious from Equation 2: lowering power supply voltage offers the largest factor of reduction (quadratic) achievable through lowering any parameter that impacts dynamic power. However, reduction in power supply voltage is accompanied with operating speed degradation due to reduced average transistor on-drive currents. Specifically, when voltages are scaled below the sum of the threshold voltages of the NMOS and PMOS devices, gate delays increase drastically, making them a substantial critical path delay contributor even in interconnect dominated circuits. Figure 4 demonstrates this effect for a static CMOS (3,2) Carry Save Adder (CSA), the basic building unit for a majority of datapath circuits, in a commercial 3V, 0.5μm bulk-CMOS process. The sum of the nominal NMOS and PMOS threshold voltages is approximately 1.6V. A nearly 9X improvement in total power is achieved through voltage scaling from 3V down to 1V; however, delay increases by nearly 28X simultaneously.

Two broad categories of solutions, (a) architectural and (b) technological, have been proposed to ease this bottleneck and compensate for the performance degradation, thereby extending the voltage scaling lower bound. We next examine these approaches and their advantages and limitations.

2.2.1 Architectural Speed Compensation Solutions

Architectural solutions are speed-enhancing modifications to the circuit architecture to compensate for the speed reduction due to voltage scaling

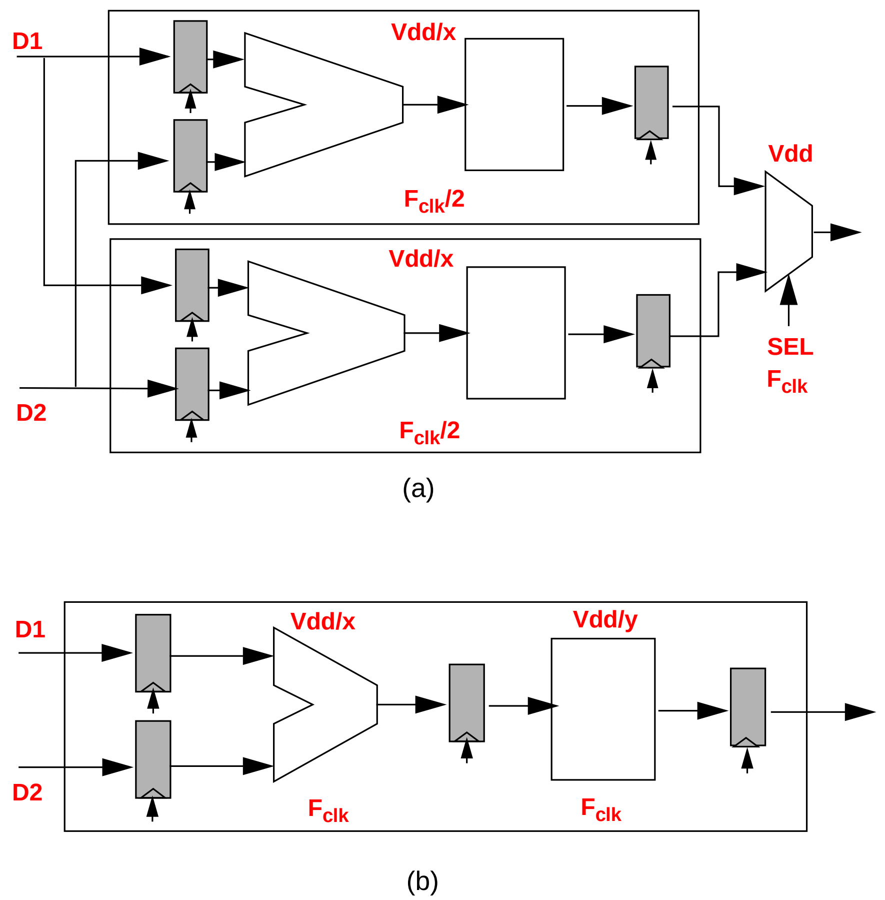
FIGURE 4 Static CMOS 3,2 CSA and its normalized power and delay in 0.5µm process.



[Chandra95]. Figure 5 shows the two popular architectural solutions, parallelism and pipelining, applied to an example datapath circuit:

1. **Parallelism** entails replicating the voltage scaled circuit, so that each circuit may operate at a lower clock frequency, while still retaining the desired throughput at the outputs. As an example, Figure 5(a) illustrates the example datapath circuit duplicated, with each circuit operating at a scaled voltage V_{dd}/x such that clock frequency is $F_{clk}/2$. The circuit outputs are

FIGURE 5 Architectural solutions for voltage scaling speed compensation.



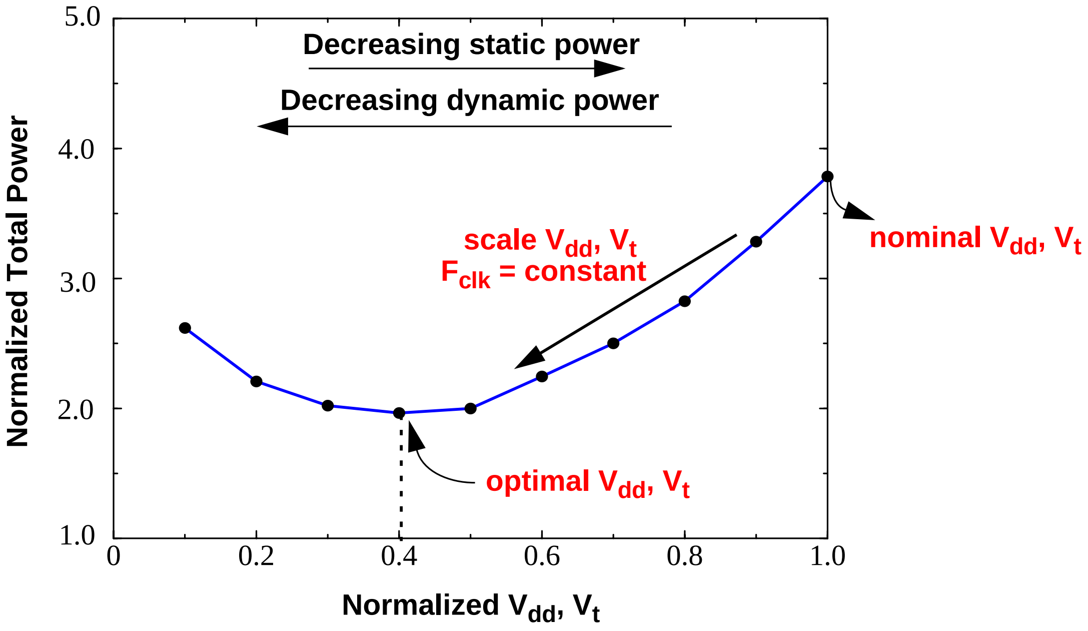
time-multiplexed at a clock frequency of F_{clk}, thereby retaining the desired external throughput. In general, the voltage may be scaled even lower by replicating the circuit N times, with each circuit clocking at F_{clk}/N and still retaining external throughput of F_{clk}. However, this approach requires a high layout area overhead and incurs the output multiplexor's delay penalty, both of which increase with N. Further, at low supply voltages, the power overhead due to parallelism offsets any power reduction achieved due to voltage scaling, essentially imposing a lower bound to voltage scaling.

2. **Pipelining** entails inserting register stages between the functional units within the circuit, so that each pipeline stage may operate at a lower voltage, while still retaining the desired external throughput. As an example, Figure 5(b) illustrates a register stage between the datapath circuit, with each pipeline stage operating at a lower supply voltage (V_{dd}/x and V_{dd}/y respectively), while still operating at a clock frequency of F_{clk}. This approach requires relatively lesser area penalty than parallelism, but increases the operation latency. Similar to parallelism, at low supply voltages, the additional register stages required to restore throughput contributes to increased clock power and area penalty, offsetting any power reduction achieved through voltage scaling. This essentially imposes a lower bound to voltage scaling as well.

2.2.2 Technological Speed Compensation Solutions

Technological solutions are fabrication process modifications that recommend simultaneous scaling of device threshold voltages and operating voltages to alleviate the speed penalty of voltage scaling [Liu93], [Burr94], [Gu96], [Frank97]. As shown in Figure 6, scaling threshold voltage and power supply simultaneously offers an exponential increase in static power and a quadratic reduction in dynamic power; since the latter typically dominates, an overall total power reduction is achieved. This continues until an optimum power supply and threshold voltage are reached when static and dynamic power are balanced, minimizing total power. Further threshold voltage or power supply scaling causes total power to increase due to static power domination. However, threshold voltage scalability is limited due to their intra- and inter-die variations caused by inevitable process and operating temperature fluctuations. The variations have been projected to increase at least linearly with decreasing feature sizes, becoming comparable to the threshold voltages themselves [Yan95], [Eisele95], [Strojwas96], [Tang96]. The threshold variations also cause increased delay and power dispersion [Sun94], [Davari96], [Frank97], with operating voltage scaling, degrading low-voltage manufacturability [Strojwas96]. As an example, measurements on a commercial 3V, 0.4μm bulk-CMOS process with nominal threshold voltages of 0.5V have demonstrated an exponential increase in clock frequency dispersion reaching up to 6X at V_{dd} = 1V due to threshold variations [Sun94]. Finally, threshold voltage scaling causes an exponential increase in leakage currents, typically by an order of magnitude for every 60-90mV of scaling in submicron processes [Bakoglu90]. This, from Equation 4, exponentially increases static power dissipation. In variable-load signal processing applications, where intermittent periods of computation (active operation mode) are separated by long periods of inactivity (sleep or standby mode), this high static power dissipation contributes to an unacceptably high off-state power [Chandra96]. The high leakage currents also prevent the effective usage of I_{DDQ} testing approaches [Acken83], commonly employed for detecting power-ground short-circuit/bridging faults [Shigematsu95]. These factors have made effective control of the threshold variations and the high leakage power with scaling threshold voltages prime challenges towards the applicability of technology-driven voltage scaling in the deep-submicron era. We next examine proposed approaches to tackle these challenges and evaluate their effectiveness in current and future fabrication processes.

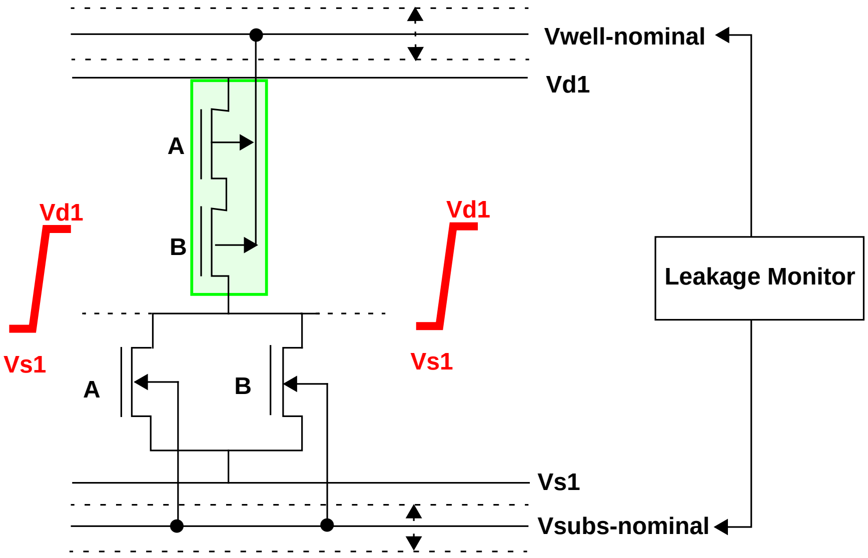
FIGURE 6 Technological solutions for voltage scaling speed compensation.



2.3 Self-Adjusting/Variable Threshold CMOS Approaches

Electronically controlling the threshold voltage variations by exploiting the body effect of MOS devices have been proposed [Kobayashi94], [Chen95], [Kuroda96]. Figure 7 illustrates the generic principle behind the Self-Adjusting/Variable Threshold Schemes (SATS/VTS), where the well and substrate connections are isolated as separate rails. Leakage current monitors in the proximity of the circuit being controlled sense threshold variations via variations in leakage currents (since leakage currents are strong functions of threshold voltage) and accordingly offset the substrate and well voltages to compensate the variations. Up to a 67% control in threshold variations has been demonstrated in a 0.7µm process with this approach [Kobayashi94]. An added bonus of this methodology is that during sleep-mode, the substrate/well rails are offset to their maximum voltages, maximally body-effecting the tran-

FIGURE 7 Self-Adjusting/Variable Threshold Scheme for electronic variations control.



sistors. This maximizes their threshold voltages, minimizing standby power dissipation. Up to four orders of magnitude reduction in leakage currents has been obtained in shifting from active to standby operation mode through this scheme [Kuroda96].

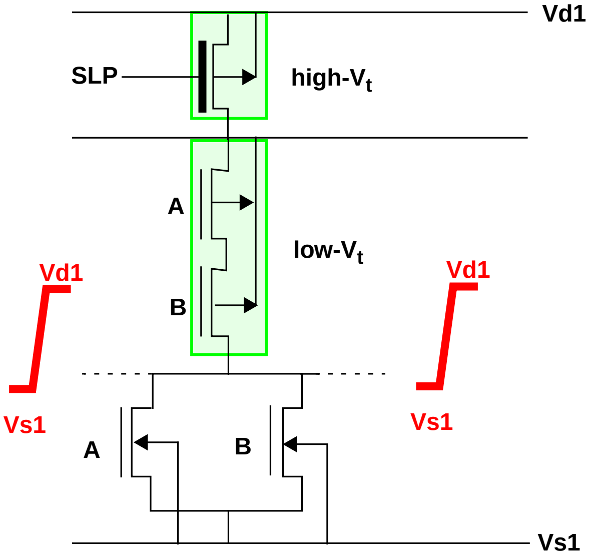
Unfortunately, the steeply increasing variations with process scaling may render these techniques ineffective at deep-submicron feature sizes, i.e., the bulk voltages required to compensate for the variations may substantially exceed the maximum process-permitted voltage. Furthermore, the absence of body effect in conventional partial- or fully-depleted SOI devices restricts their applicability in SOI processes. Although non-conventional body-tied SOI devices are being developed to overcome this restriction [Yang95], [Antoniadis97], [Douseki97], no commercial solutions have been reported to date.

2.4 Multiple Threshold CMOS Approaches

Multiple threshold voltage (multiple-well) approaches have been proposed to mitigate the aforementioned standby power problems due to high leakage currents [Shigematsu97]. These approaches entail the usage of dual threshold voltages (in principle extendable to any number of threshold voltages) by employing dual wells, one for each type of device, at an added fabrication cost due to modifying the process recipe.

Figure 8 illustrates the proposed usage of the two threshold voltages: the circuit implemented using the lower threshold voltage devices, and a PMOS “virtual power transistor” implemented with a higher threshold voltage. During active mode of operation, the virtual power transistor is enabled (SLP=Vs1) and delivers the circuit’s drive currents through it. During sleep-mode, the virtual device is disabled (SLP=Vd1), tristating the circuit. Since there exists no DC path between power supply and ground within the circuit, standby power is virtually eliminated, confined to the high threshold voltage PMOS device’s leakage power. Control circuits have been developed to transfer the data stored in the circuit’s registers to special latches before enabling sleep-mode in order to retain circuit state. The data is transferred back into the appropriate circuit registers to restore state when returning back into active mode. Although significant standby power savings can be achieved, these approaches incur substantial delay and dynamic power penalty in transferring state data. Particularly

FIGURE 8 Multiple Threshold Scheme for low standby power dissipation.



in large datapath circuits, the delay incurred in transferring back and forth the substantial state data may be prohibitive, i.e., a significant fraction of the sleep-mode period itself! Further, in variable-load signal processing applications [Chandra96], where significant transitions between active and sleep modes occur, the power penalty can be prohibitive as well, offsetting any standby power savings achieved. These factors confine the applicability of these techniques to small circuits, where the state transfer delay and power penalties are acceptable.

In summary, limitations to architecture- and technology-driven voltage scaling and the inability to effectively control intra- and inter-die threshold voltage variations, have motivated a strong quest for alternate low-power circuit methodologies in standard submicron CMOS and SOI processes, without mandating any technology modifications. In the next section, we examine the four broad categories of previously reported solutions in literature and evaluate their applicability in high-speed/low-power datapath circuits.

2.5 Alternate Low-Power Circuit Methodologies

Driven by the strong demand for high-speed and low-power digital circuits in general, and datapath circuits in particular, several alternate circuit families have been proposed, classified broadly into four categories: (a) Dynamic Logic-based techniques, (b) Pass-transistor Logic-based techniques, (c) Cascode Voltage/Current Switch Logic-based techniques, and (d) Adiabatic Logic-based techniques. While several variants have been developed under each category, we now review the most interesting approaches, with an emphasis on their applicability in future deep-submicron processes.

2.5.1 Dynamic Logic-based techniques

Domino CMOS [Krambeck82], [Goncalves83], Zipper CMOS [Lee86], and Clocked CMOS [Bakoglu90] approaches have been proposed for improved speed and lower power than static CMOS logic. Unfortunately, dynamic techniques require single- or multi-phase clock signals to perform their logic function. Since clock signals have unity switching activities, the precharge/evaluate transistors of every dynamic logic gate are charged/discharged each cycle, contributing to substantial additional power in large datapath circuits. Furthermore, since the output nodes are precharged and evaluated every cycle even when the input signals do not transition, dynamic techniques demonstrate significantly higher switching activities, thereby offsetting any dynamic power savings achieved due to their relatively lower input gate capacitance than static CMOS circuits [Wailee94], [Ng96]. Thus, dynamic methodologies have traditionally found usage only in high-speed digital circuits where power is not as much a concern as clock frequency.

2.5.2 Pass-transistor Logic-based techniques

Single-ended and fully-differential pass-transistor and transmission-gate logic techniques [Yano90], [Suzuki93], [Krishna95], [Param96], [Yano96], have been proposed as high-speed and/or low-power alternatives to the static CMOS methodology. However, since outputs of pass-transistors do not swing rail-to-rail, these approaches incorporate swing restoration circuitry to restore the logic gate outputs to full-swing (static CMOS) levels, degrading both speed and power. In addition, pass-transistor based techniques demonstrate rapid low-voltage speed degradation and relatively higher switching activities within the gates - even in single-ended implementations - offsetting any power reduction achieved due to their lower input gate capacitance. These factors make them power-inefficient compared to static CMOS in current and future low-voltage technologies [Yano96], [Zimmer97].

2.5.3 Cascode Voltage/Current Switch Logic-based techniques

Fully-differential Voltage-switch [Heller84] and Current-switch [Soma97] logic approaches have been proposed as high-speed/low-power alternatives to static CMOS. Unfortunately, they exhibit inherently higher switching activities due to being fully-differential, require routing both true *and* complimentary signals, necessitate single- or multi-phase clocks for operation (in some schemes), and display relatively lower noise immunity. Current-steering logic techniques [Ng97] have been developed which exhibit improved noise immunity, but high-speed is achieved at the cost of increased static DC bias currents; this contributes to high static power consumption. These factors have rendered them both power and speed inefficient except for large-fanin gate structures [Chu87], [Soma97].

2.5.4 Adiabatic Logic-based techniques

Fully-dynamic and quasi-static energy recovery logic approaches have been proposed to lower the power consumption of static CMOS circuits [De96], [Ye97], [Athas97]. However, adiabatic techniques require single- or multi-phase complimentary clocked power supplies and display significantly higher switching activities than static CMOS circuits. These factors, coupled with the necessity for efficient adiabatic power supply clock generators have confined the usage of these techniques to fairly low-speed (well below 100 MHz) applications, with their energy efficiency decaying exponentially with increasing clock frequency [Ye97]. Thus, adiabatic logic approaches have not been successfully attempted in literature for high-speed/low-power datapath circuits.

2.6 Summary

In this chapter, we described the evolution of static CMOS as one of the most popular choices of implementing high-speed/low-power datapath circuits. An overview of the components of power dissipation within static CMOS circuits was presented. Distributions of these power components from two commercial 0.25μm DSPs were shown to illustrate the increasing dominance of dynamic and short-circuit power components, in that order, over total power. Previously reported architectural and technological approaches to lower static CMOS power consumption were analyzed. Their advantages and limitations were outlined, emphasizing on their applicability in future deep-submicron processes. Advantages and limitations of previously published alternate logic families were also examined to evaluate their applicability in high-speed/low-power datapath circuits.

By investigating the limitations of existing approaches to lower static CMOS power consumption as well as existing logic families, we have re-emphasized the strong necessity for exploring alternate circuit methodologies for high-speed/low-power datapath circuits to achieve substantial energy/operation reduction over static CMOS in current and future deep-submicron processes. In the next chapter, we will examine power-delay trade-offs for a fully static CMOS implementation of a commonly employed DSP MAC architecture to explore opportunities for lowering energy/operation. In Chapter 4, we explore the potential for employing multiple power supply voltage-based techniques to exploit these opportunities and lower static CMOS energy/operation without warranting any technology modifications.

3 DSP MAC Circuits: Power-Delay Trade-offs

In this thesis, we focus our attention on large DSP datapath operators such as multipliers and MAC circuits, where lowering the energy/operation is of greatest research concern. These form the heart of a majority of commercial DSP processor datapath and therefore constitute a good vehicle to study both processor-level and system-level impact on DSP energy/operation [Allen85], [Lapsley96]. In this chapter, we examine architectural choices for signed (2's complement), fixed-point MAC circuits and formulate a commonly employed high-speed/low-power MAC architecture, that will be the focus of our further investigation. We then present a detailed exploration of power-delay trade-offs for this MAC architecture for a fully static CMOS implementation over a range of operand bit-widths, power supply voltages, and submicron fabrication processes. On the basis of this study, we determine opportunities to lower the energy/operation of MAC circuits, that will be exploited in future chapters.

3.1 MAC Architectural Choices

In this section we review radix multiplication and accumulation, and some commonly employed high-speed/low-power architectures for them. Radix multiplication consists of generating the product of two numbers called the *multiplicand* and the *multiplier.* While the multiplier and multiplicand may be of different bit-widths, this work specifically studies the most common case: when they are of the same bit-width. Multiplication can be conceptually seen as a sequence of shift-and-add operations. Accumulation, as the name implies, is the iterative addition of the multiplication results over all input vectors. The multiply-accumulate operation can be divided into three mutually exclusive parts, wherein architectural choices for each is independent of the others [Cavanagh84]:

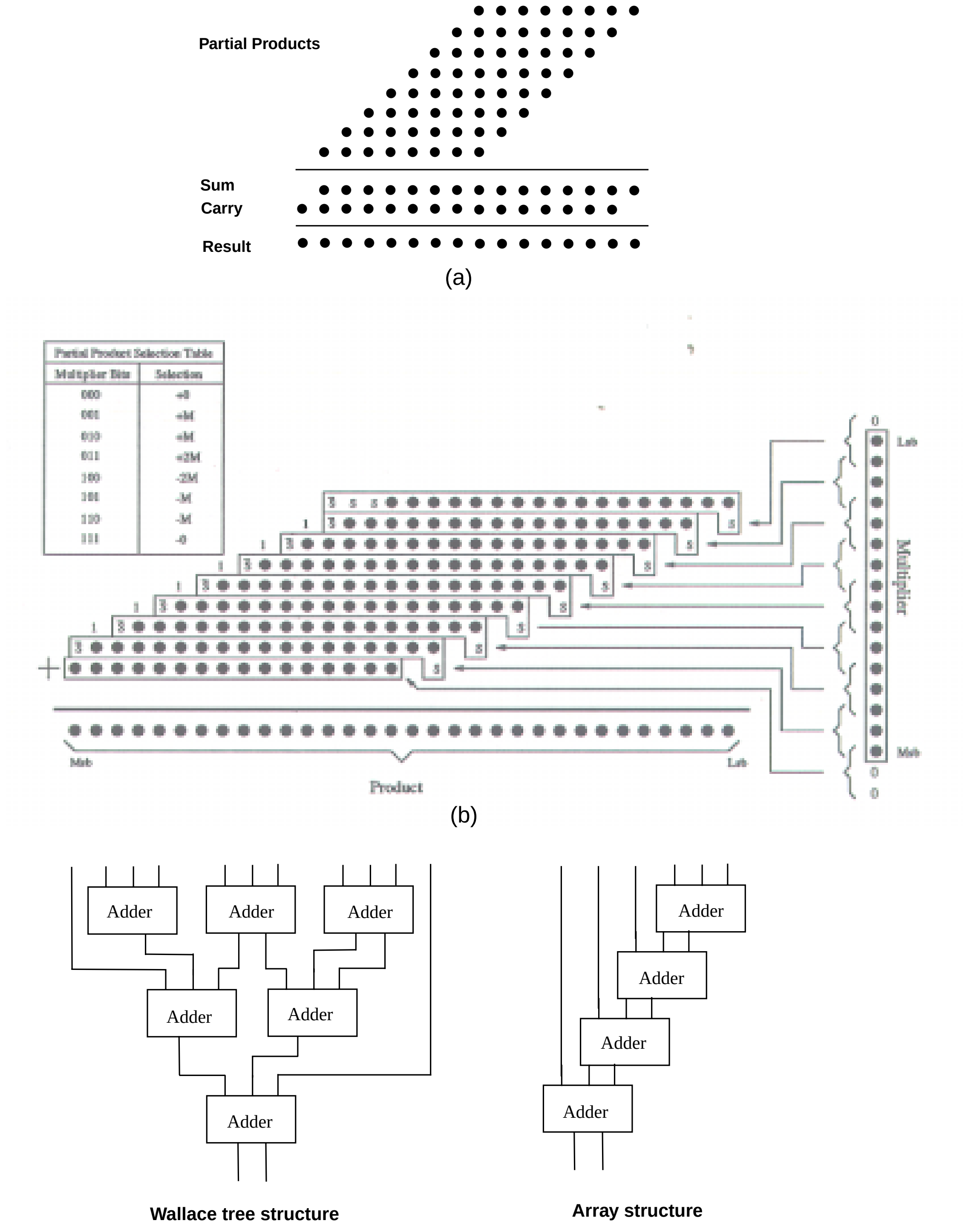
3.1.1 Partial Product Generation

The shifted multiplicand bits (called *summand*s) are generated here to form the *partial product array* [Cavanagh84], as shown in Figure 9(a). For a *n*n* multiplication, *n* partial product vectors, each of width *n* are produced. Booth recoding [Booth51] is a technique commonly used to reduce the number of partial product vectors, by recoding the *multiplier* bits into its multiples by examining consecutive bits of the *multiplicand.* Examining a larger number of multiplicand bits offers a proportionately larger reduction in the number of partial product vectors. Thus, Booth recoding-based partial product generation results in reduced hardware and subsequently power, at the cost of a slight penalty in encoding delay. However, this penalty is usually a small fraction of the total multiplication time reduction that this technique offers [Twaijry94]. The most commonly used Booth recoding approach is called Overlapped bit-pair recoding (or Modified Booth Algorithm) [Ardekani93], as shown in Figure 9(b). Here, the multiplier is recoded by examining every two successive bits of the multiplicand according to the table in Figure 9(b). This reduces the number of partial products generated by a factor of two (to *n/2*) compared to conventional AND-gate based partial product generation, offering substantial savings in area, power, and delay. Higher-order Booth recoding, i.e., recoding three bits and beyond, have been proposed; however, the recoding delay penalty incurred causes an overall increase in total multiplication time, even for wide bit-width multipliers [Twaijry94].

3.1.2 Partial Product Reduction

The *n*/2 partial products generated through bit-pair Booth recoding are added to produce two final *2n-1* bit vectors using Carry Save Adders (CSAs). Partial product reduction can be accomplished using either an array topology [Cavanagh84] or a (Wallace) tree topology [Wallace64], as illustrated in Figure 9(c). Array topologies have a logic depth of $O(n)$ and a regular structure, enabling easy layout. Wallace trees employ a parallel reduction scheme and have a logic depth of $O(log_{3/2}n)$, but an irregular structure making it difficult to layout. A majority of high-speed/low-power multipliers (over a wide range of bit-widths) have employed Wallace trees because of its shorter depth, fewer switching nodes, and lower switching activities (due to reduced spurious transitions) than array topologies [Goto92], [Ardekani93], [Twaijry94], [Wailee97b]. In addition, optimal layout topologies have been developed in order to overcome the irregular structure bottleneck of Wallace trees, demonstrating substantial area improvements [Twaijry96]. This has furthered the motivation to adopt Wallace tree-based partial product reduction in high-speed/low-power multipliers.

FIGURE 9 Multiplier Partial Product Generation and Reduction structures.



Partial Product Selection Table	
Multiplier Bits	Selection
000	+0
001	+M
010	+M
011	+2M
100	-2M
101	-M
110	-M
111	-0

A wide variety of CSA constructions have been used for partial product reduction. The most commonly used CSA construction is the (3,2) counter (Figure 4), which compresses 3 input bits to generate 2 output bits, although higher order CSAs (e.g., 5,3 and 7,3 counters) and/or an optimal combination of low- and high-order CSAs (e.g., 3,2 and 5,3 counters) may be employed to minimize delay or power [Twaijry96]. In this work, we specifically focus on the most general case of (3,2) CSA-based partial product reduction.

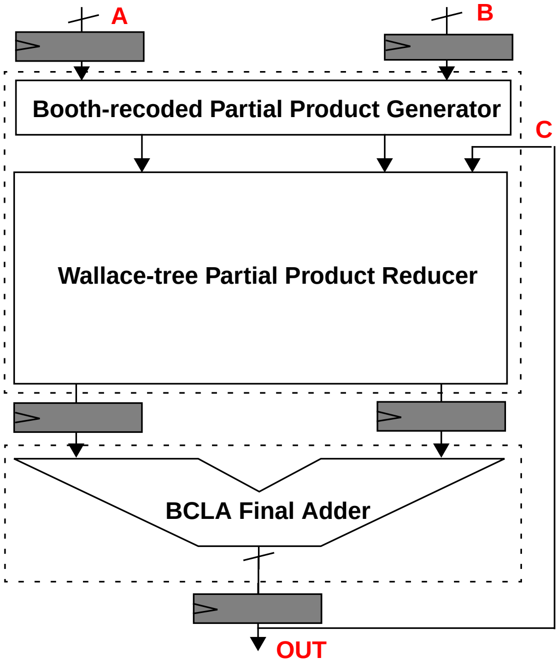
3.1.3 Final Addition

The two *2n-1*-bit reduced partial product vectors and current accumulator output are added to produce the next accumulator result. In high-speed MAC architectures, the current accumulator result is pushed into the Wallace tree partial product reduction stage in order to exploit the tree's logarithmic compression depth [Cavanagh84]. There exist many architectural choices for final addition which have been characterized on the power-delay space for different bit-widths [Nagendra94]. Block Carry Lookahead Adders [Cavanagh84], which use a parallel tree structure for rapid addition with a gate depth of $O(log_2n)$, were found to be among the least power-delay product architectures over a wide range of bit-widths, and is one of the most popular choices for high-speed/low-power final adders [Nagendra94].

In addition to these architecture choices, a degree of freedom available at the architectural level is the depth of pipelining within the MAC. In applications where a MAC operation is to be performed in one clock cycle, no pipelining is allowed. However, in high-throughput applications, a register stage is introduced between the multiplier and final adder [Lu93], [Nagamatsu95], [Jou95], [Nagendra96], [Murakami96], [Izumikawa97]. An added bonus due to the inserted pipeline stage is that it offers considerable reduction in spurious transitions, which depend quadratically on logic gate depth [Chandra95].

On the basis of this discussion, a high-performance and low-power MAC architecture commonly employed in DSP datapath is formulated, which will be the focus of further study. The architecture, shown in Figure 10, comprises a signed (2's complement), fixed-point, pipelined, Overlapped bit-pair Booth-recoded Partial Product Generator, (3,2) CSA-based Wallace tree Partial Product Reducer, and a Block Carry Lookahead Final Adder.

FIGURE 10 High-performance/Low-power DSP MAC architecture under study.


A
B
Booth-recoded Partial Product Generator
C
Wallace-tree Partial Product Reducer
BCLA Final Adder
OUT

3.2 Exploring MAC Power-Delay Trade-offs

We first examine power-delay space trade-offs within a fully-static CMOS implementation of our prototype MAC architecture over a range of operand bit-widths that dominate DSP processors, power supply voltages, and submicron process generations. This investigation will offer insights into exploring power minimization techniques while maintaining high speed.

A majority of DSP circuits are dominated by short bit-width datapath circuits, specifically over the range of 8 - 24-bit operands. Figure 11 shows the power distribution within the MAC’s building blocks for a 8*8+18-bit, 16*16+36-bit, and 24*24+56-bit static CMOS MAC employing the architecture in Figure 10, implemented in a commercial 3V, 0.5µm process. Figure 12 shows the power distribution within the MAC’s building blocks for the 16*16+36-bit MAC employing the same architecture, implemented over three additional submicron technology generations: 0.35µm bulk-CMOS, 0.25µm fully-depleted SOI, and 0.16µm bulk-CMOS. Power consumptions of each MAC implementation are obtained from measurements on fabricated ICs (0.5µm 16*16+36-bit MAC) and circuit simulations using BSIM models (other 0.5µm MACs and the 0.35µm, 0.25µm, and 0.16µm designs), across 500 pseudo-random input vectors. Some important conclusions can be drawn from Figure 11 and Figure 12:

- With increasing operand bit-widths, the percentage of total power dissipated in the multiplier circuit increases from nearly 50% in the 8*8+18 case up to 78% in the 24*24+56 case. This is primarily because the Booth encoders of the partial product generator and the CSAs within the Wallace tree drive substantial amounts of load capacitance at their outputs that is interconnect capacitance dominated. The registers and final adder drive comparatively lower output capacitances, that are gate capacitance dominated. Also, the multiplier displays significantly higher switching activities. Therefore, the multiplier's effective switched capacitance is much

FIGURE 11 Power distribution trend with operand size for MAC architecture under study.

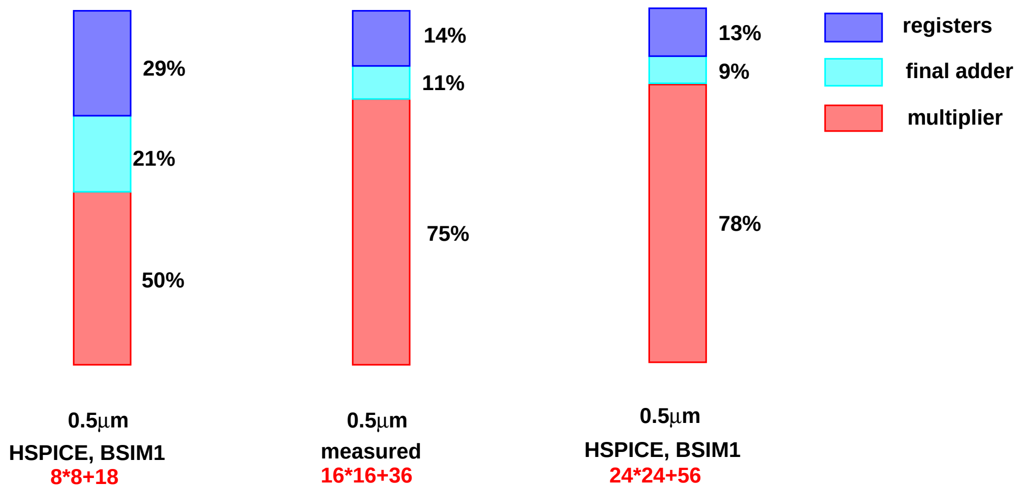
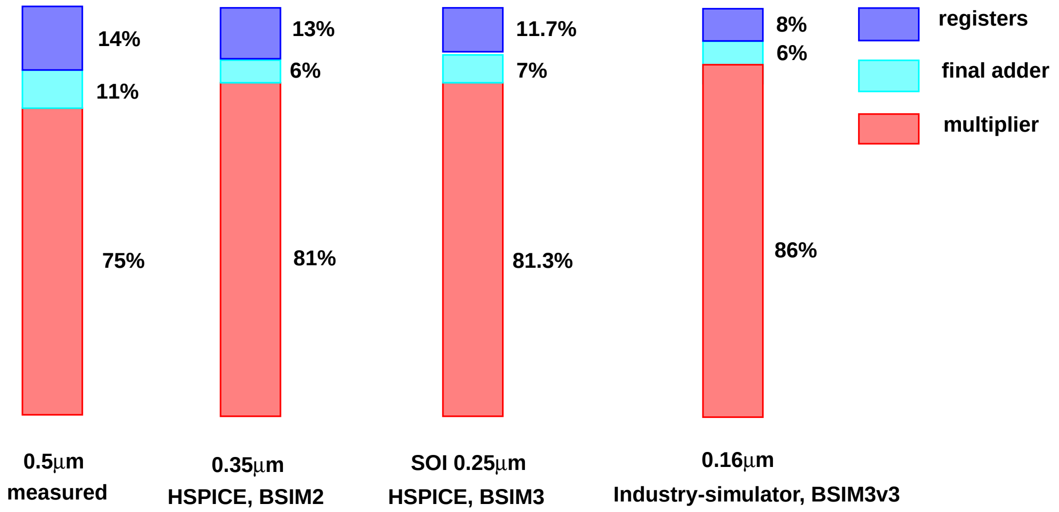


FIGURE 12 Power distribution trend with process scaling for MAC architecture under study.

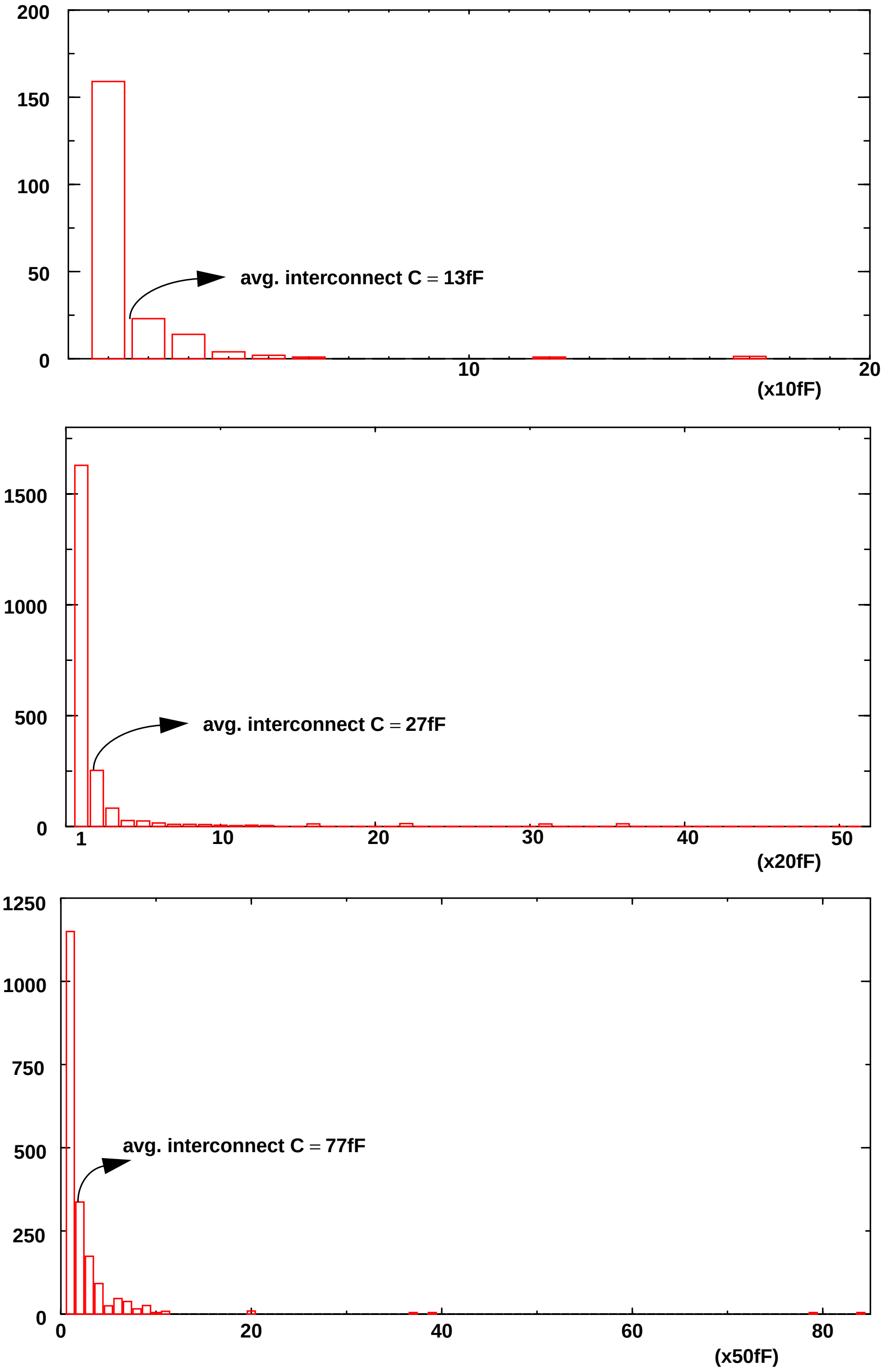


higher, making it the dominant power consumer. The dominance increases with increasing bitwidths. In order to explore this further, Figure 13 shows the interconnect capacitance distribution within the three multipliers in the 0.5µm process, extracted using Diva[1] from the fully placed-and-routed MAC layouts. It is observed that the average interconnect capacitance within the multiplier increases exponentially with bit-width, from approximately 13fF for the 8*8+18-bit MAC up to 77fF for the 24*24+56-bit MAC. Therefore, a strong necessity exists to focus attention on lowering the power consumption of the multiplier, particularly with increasing operand bit-widths.

- With scaling feature sizes, the percentage of total MAC power dissipated in the multiplier increases from 75% in the 0.5µm case up to 86% in the 0.16µm case for the 16*16+36-bit MAC, climbing up further in future deep-submicron processes. This trend is primarily because of the substantial interconnect capacitance driven by the Booth encoders and CSAs within the Wallace tree multiplier. Interconnect capacitance, dominated in deep-submicron processes by the fringing and coupling components, scales slower than gate capacitance with process scaling, making the multiplier a more and more dominant power consumer with scaling feature sizes. Figure 14 studies this trend in more detail: it shows the interconnect capacitance distribution within the 0.5µm and 0.16µm multipliers, extracted from the fully placed-and-routed MAC layouts. It is observed that the average interconnect capacitance within the 0.5µm multiplier is 27fF, about

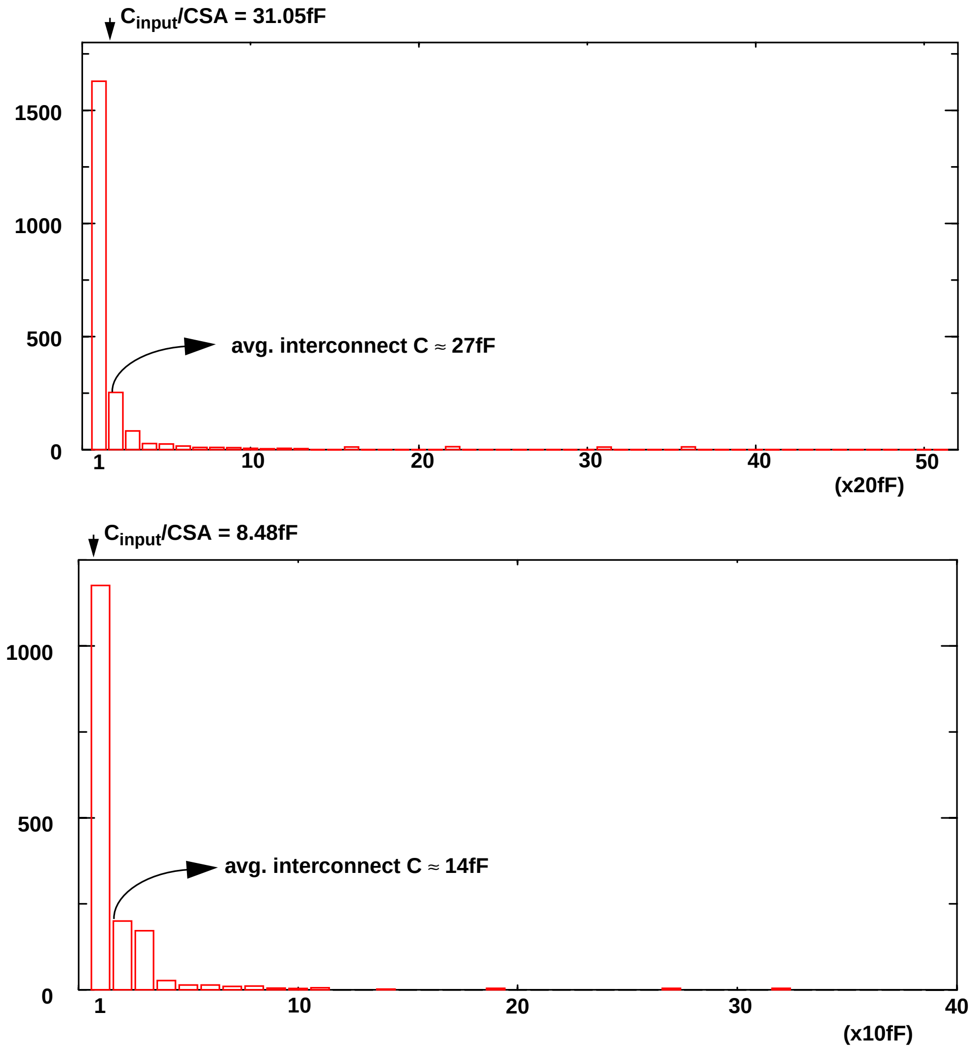
[1] Diva is a trademark of Cadence Design Systems, Inc.

FIGURE 13 8,16,24-bit multiplier interconnect distributions extracted from 0.5µm MAC layouts.



87% of the fanin gate capacitance per input of a CSA, which is 31.05fF. For the 0.16µm multiplier, the average interconnect capacitance is expectedly lower at 14fF, whereas the gate capacitance per input of a CSA drops much more rapidly to 8.48fF. The interconnect capacitance is now about 165% of the fanin gate capacitance, making the multiplier's power an even more dominant component. Therefore, there exists a strong necessity to

FIGURE 14 Multiplier interconnect distribution extracted from 0.5µm and 0.16µm MAC layouts.



focus attention on lowering the power consumption of the multiplier, particularly with scaling feature sizes.

Figure 15 shows the ratio of final adder to multiplier delays as a function of operating voltage scaling for the same 8*8+18-bit, 16*16+36-bit, and 24*24+56-bit MACs in the 0.5μm process. Figure 15 also shows this delay slack ratio for the 16*16+36-bit MAC in the 0.16μm process. Some important conclusions can be drawn from here:

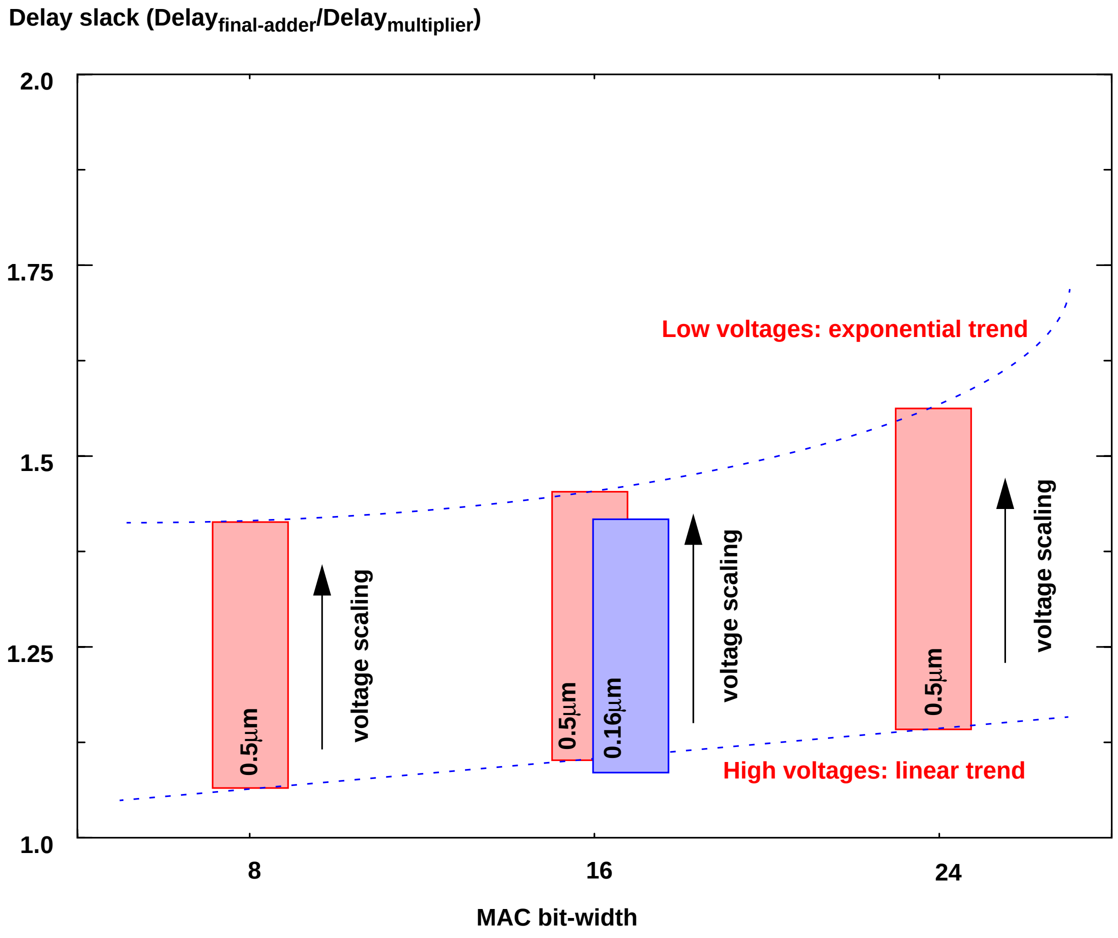
- The delay slack ratios are greater than unity over this range of MAC bit-widths and increasing with voltage scaling. The final adder determines the operable clock frequency of these MACs. This is due to its relatively higher logic gate depth than the multiplier. The adder’s delay dominance increases linearly with MAC bit-width at high voltages, whereas the increase becomes exponential at low voltages. This is because, with voltage scaling, transistors in the multiplier and adder are subjected to lower drain-source electric fields and hence lesser carrier velocity saturation [Sakurai90]. This causes the saturation-region drive currents to display a nearly quadratic relationship to operating voltage [Bakoglu90]. The higher logic gate depth of the adder therefore causes a steeper delay increase than the multiplier with voltage scaling, thereby increasing the delay slack.
- With scaling feature sizes, the delay slack between the final adder and multiplier decreases only slightly. This is due to the multiplier’s interconnect-dominated load capacitances which scale slower then the adder’s gate-dominated load capacitances. This causes a slightly steeper multiplier delay increase than the adder with voltage scaling. The delay slack contin-

ues to increase with voltage scaling, i.e., the final adder continues to determine the MAC's clock frequency.

3.3 Summary

The increasing dominance of interconnect capacitance over gate capacitance with process scaling makes the Wallace tree multiplier power a more and more dominant component of total MAC circuit power dissipation. Therefore, there

FIGURE 15 Final Adder:Multiplier delay slack trend with voltage scaling, process scaling, and operand bit-width for static CMOS MACs.



exists a strong necessity to focus attention on lowering multiplier power, more so in future technologies. The final adder determines the MAC's clock frequency over a range of operand bit-widths and operating voltages. Therefore, power-saving techniques that sacrifice speed are non-applicable to the final adder, particularly in fixed-throughput, real-time DSP circuits. Further, the increasing power criticality of the multiplier only makes the final adder less and less power critical with scaling feature sizes; applying power-reduction techniques, therefore, does not offer any tangible total power savings. However, the increasing final-adder-to-multiplier delay slack with voltage scaling in current and future submicron processes offers an opportunity to lower the multiplier power consumption without sacrificing performance. In the next chapter, we examine mixed swing techniques that exploit this opportunity by employing multiple operating voltages to achieve lower energy/operation.

4 Mixed Swing Techniques

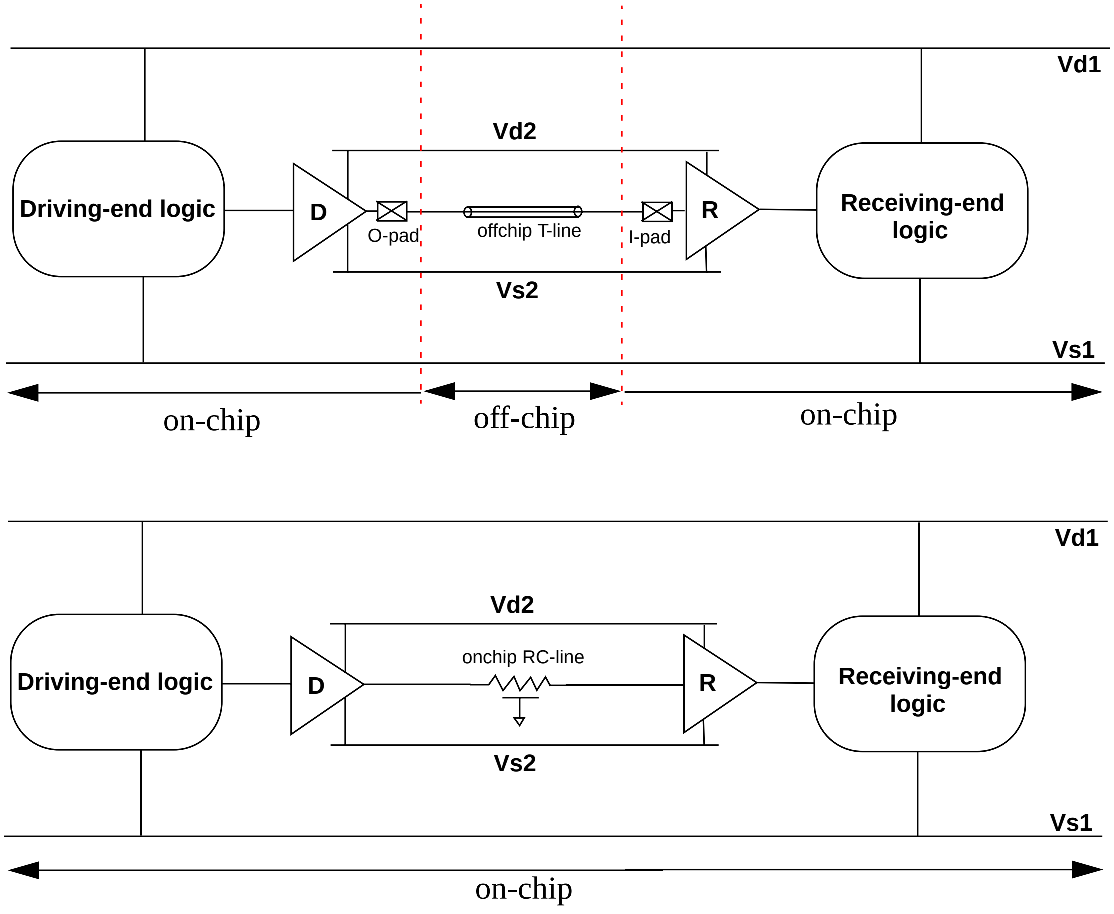
In this chapter, we introduce the concept of employing multiple power supply voltages to lower the energy consumption per switching operation of datapath circuits. We begin with a background on the evolution of multiple supply approaches to lower power dissipation, originally for off- and on-chip buses and then, more recently, for digital logic circuits. Advantages and limitations of these techniques are discussed. *Mixed Swing* techniques are then developed, which employ multiple supplies within a single gate to perform logic by intermixing high- and low-voltage signals. Static and dynamic, single-ended and fully-differential mixed swing techniques are investigated and their ability to voltage scale more effectively than fully static CMOS in standard submicron processes is demonstrated.

4.1 Background: Multiple Voltage Techniques

Multiple power supply-based techniques were originally developed to lower the power consumption of long off-chip [FutureBus83], [Knight88] and on-chip [Bakoglu85], [Shin89], [Nakagome93], [Sakurai97] buses. The motive behind these techniques is to drive the bus at a reduced voltage swing to lower the dynamic power dissipated in charging/discharging the large bus interconnect capacitance loads. Figure 16 illustrates the general principle behind these approaches, which essentially consist of two parts:

- A Driver circuit (represented as "D" in Figure 16) which interfaces the driving-end logic circuit operating between a regular, high-swinging pair of power supply rails (Vd1-Vs1) and the off-/on-chip bus being driven between a secondary, low-swinging pair of power supply rails (Vd2-VS2).

FIGURE 16 Previous off- and on-chip mixed voltage swing techniques general principle.



The Driver circuit receives the regular, high-swinging output signal from the driving-end logic circuit and drives the bus at the reduced swing.

- A Receiver circuit (represented as “R” in Figure 16) which interfaces the off-/on-chip bus being driven between the low-swinging pair of power supply rails (Vd2-Vs2) and the receiving-end logic circuit operating between the same regular, high-swinging pair of power supply rails (Vd1-Vs1). The Receiver circuit receives the low-swinging signal at the opposite end of the bus and amplifies it back to the regular, high-swing before feeding it into the receiving-end logic circuit.

Many approaches have been proposed in literature for implementing the driver and receiver circuits in order to maximize the power savings and minimize the delay penalty due to signal level conversion at the driving and receiving ends. The charging/discharging current requirements for long buses, even with their reduced swings, are substantial. So, there exists substantial simultaneous switching noise (power/ground bounce) on the driver/receiver power rails. Therefore, driver/receiver circuits demand highly noise-immune circuit topologies [Bakoglu85], [Knight88], [Bakoglu90], [Nakagome93]. A majority of driver/receiver techniques have employed the fully static CMOS methodology due to its high noise immunity.

The low-swing power rails may either be delivered from an explicit off-chip supply as in [Knight88] or locally generated through on-chip series regulation techniques as suggested by [Nakagome93]. On-chip regulation eliminates the necessity for an additional low-swing supply. However, since the low-swing stage’s drive currents are now sourced directly from the high-swing supply, there exists a DC series path between the high- and low-swing power rails. Therefore, from Equation 2, the dynamic power savings are now linear with the reduced swing. Employing an off-chip supply, on the other hand, offers a full quadratic reduction in bus dynamic power with the reduced swing.

4.2 Multiple Supply Digital Circuits

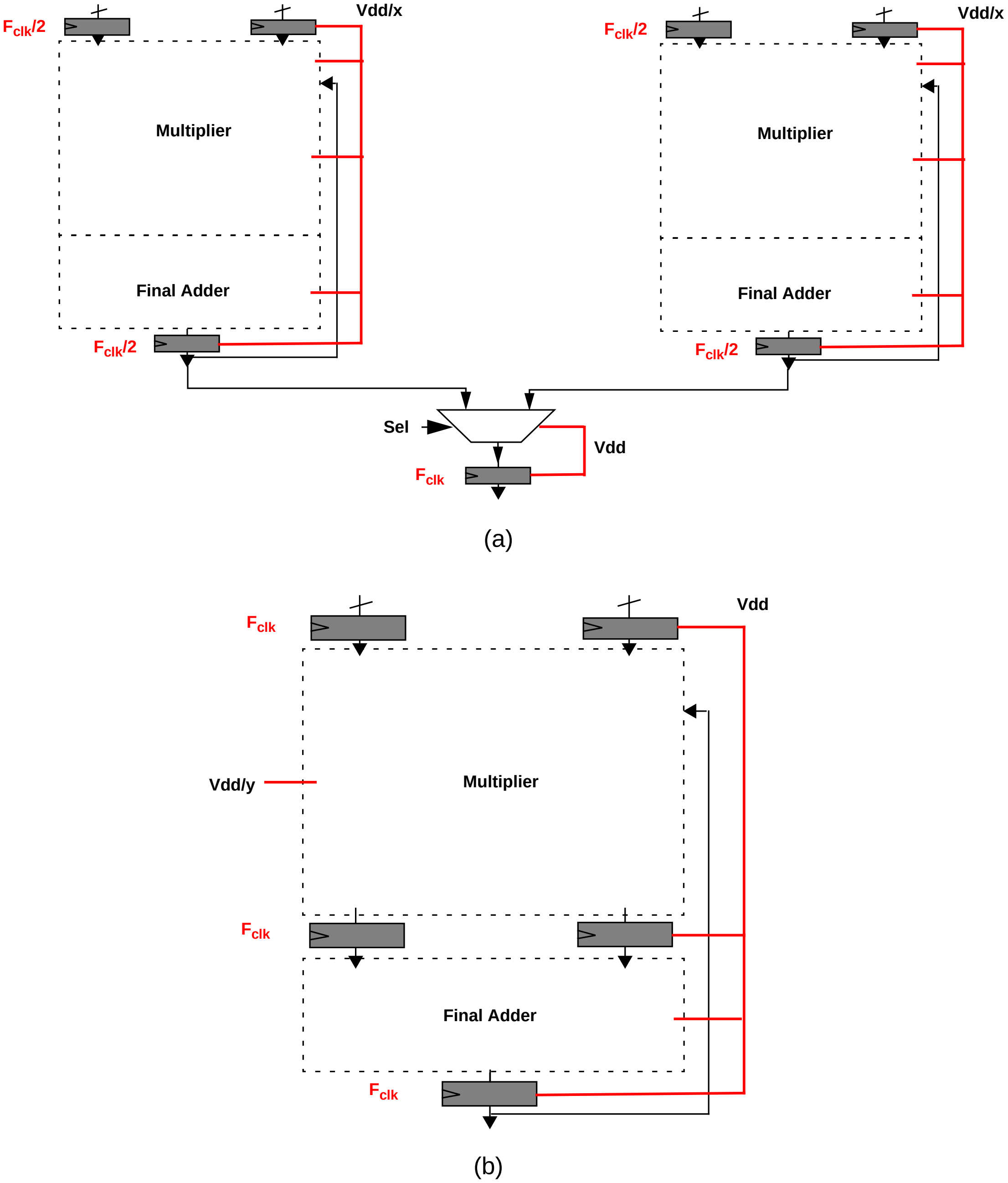
Limited work has been reported on employing multiple voltages to achieve the same power reduction goals *within* digital logic circuits. Two broad categories of multi-supply approaches have been proposed: (i) architecture-driven voltage scaling, and (ii) clustered voltage scaling. We next examine these approaches within the context of datapath circuits.

4.2.1 Architecture-driven Voltage Scaling

We have examined this class of multiple supply techniques previously in Chapter 2: parallelism and pipelining have been proposed as architectural solutions to compensate for the speed degradation of static CMOS circuits due to operating voltage scaling [Chandra95]. We now examine the effectiveness of these techniques in lowering the power consumption of datapath operators.

- **Parallelism:** Figure 17(a) illustrates the application of parallelism to an example MAC circuit. The MAC is replicated as shown, with each MAC operating at a voltage V_{dd}/x. The divisor x represents the extent of voltage scaling (and hence the extent of power reduction) this technique permits for an *internal* throughput of $F_{clk}/2$. The MAC outputs are time-multiplexed, i.e., the select signal *sel* is clocked at F_{clk} to extract an output from each MAC every cycle, thereby retaining target *external* throughput of F_{clk}. In general, the MAC may be replicated N times, each operating at a clock frequency F_{clk}/N, enabling further voltage scaling and hence power savings.

FIGURE 17 (a) Parallelism and (b) Pipelining applied to a typical DSP MAC architecture.



However, as pointed out in Chapter 2, parallelism approaches diminishing returns as N increases. Further, since parallelism mandates literal replication of hardware, the layout area penalty due to parallelizing becomes prohibitive for large datapath circuits such as MACs. Therefore, this approach has traditionally been confined to small, simple datapath circuits, e.g., short bit-width adders, subtractors etc.

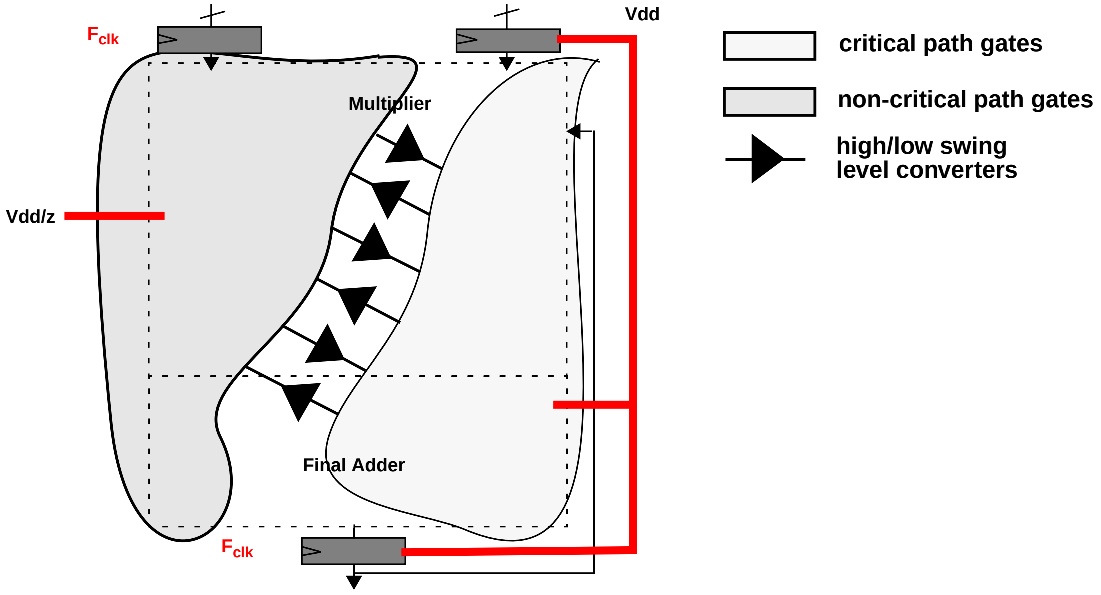
- **Pipelining:** Figure 17(b) illustrates the application of pipelining to the example MAC circuit. The MAC is pipelined between the multiplier and final adder as shown by inserting a single register stage in between them. The time-critical pipeline stage, assumed to be the final adder in Figure 17(b), operates at a regular, high voltage, V_{dd}. The non-time-critical pipeline stage, assumed to be the multiplier, operates at a lower voltage V_{dd}/y exploiting the delay slack between pipeline stages, while still retaining the target *external* throughput of F_{clk}. The divisor y represents the extent of voltage scaling this technique permits, and hence the extent of power reduction within the multiplier. I/O and pipeline registers operate at the high voltage to retain signal level compatibility with peripheral circuitry and level conversion circuits are inserted at the high/low voltage interfaces. In general, the MAC may be pipelined (more finely) N times. This enables further voltage scaling (and hence power savings), with each pipeline stage still operating at a clock frequency of F_{clk}, but at the cost of higher latency. However, as pointed out in Chapter 2, pipelining approaches diminishing returns with increasing N as well. Since pipelining only requires insertion of intermediate register stages, whose area penalty is significantly smaller than replicating hardware, it has evolved as a more feasible architectural solution to lowering power consumption of large datapath circuits than parallelism. An added bonus due to this approach is that spurious transitions, which are substantial within datapath circuits, decay quadratically with increased pipelining [Chandra95], further motivating its widespread usage.

4.2.2 Clustered Voltage Scaling

While pipelining exploits delay slack at the architectural level, clustered voltage scaling [Usami97] exploits it one level lower, at the circuit level. Multiple voltages are employed to exploit the delay slack between critical and non-critical paths *within* a digital circuit. Figure 18 shows the application of clustered voltage scaling to the same example MAC circuit. The critical and non-critical path gates are isolated into separate routing channels in the layout and tied to independent power supplies [Igarashi97]. The critical path gates operate at a regular, high voltage (V_{dd}) to meet the target throughput of F_{clk}. The non-critical path gates operate at a lower voltage V_{dd}/z exploiting the delay slack to equalize critical and non-critical path delays. Level convertion circuits are inserted at the high/low voltage interfaces. Power savings is achieved due to the reduced operating voltage of the non-critical gates. The divisor z represents the extent of voltage scaling this technique permits, and hence the extent of power reduction within the MAC.

The power reduction obtained through clustered voltage scaling is limited by the fraction of total gates that are non-critical and the available delay slack between critical and non-critical paths; higher the fraction of non-critical gates and delay slack, larger the power savings. Control path/random logic circuits typically have large delay slacks and a substantial fraction of non-critical gates, and hence are well-suited for this approach. As an example, application of clustered voltage scaling to seven random logic modules on the Mpact[1] media processor offered a 47% reduction in the power dissipated in those modules [Igarashi97]. The corresponding critical and non-critical gate voltages are 3.3V and 1.9V respectively. The savings were attributed to (i) 76% of the total gates being non-critical, enabling their voltage to be scaled, and (ii) substantial delay slacks, enabling the lower voltage to scale significantly (by 42%) below 3.3V to 1.9V.

FIGURE 18 Clustered voltage scaling applied to a typical DSP MAC architecture.



[1] Mpact is a trademark of Toshiba Corporation, Japan.

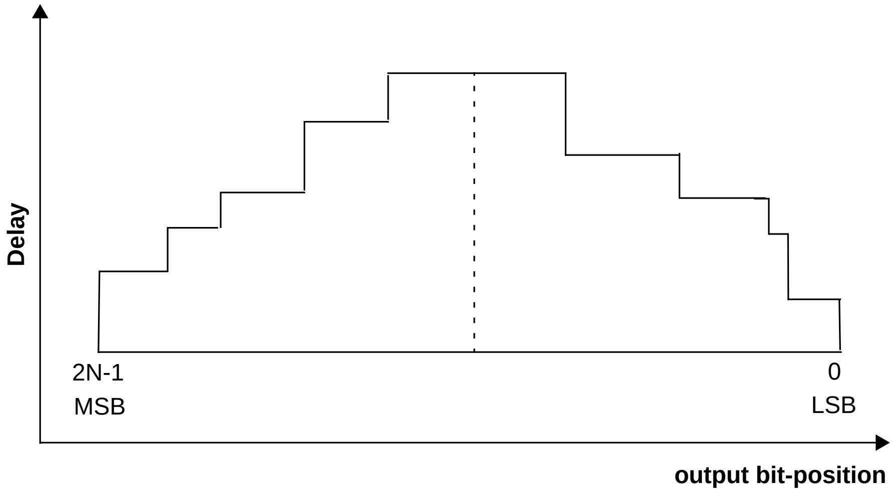
A majority of DSP datapath circuits typically have regular logic structures and hence do not possess high fractions of non-critical gates or large critical-to-non-critical path delay slacks. As an example, Figure 19 illustrates this behavior for a Booth-recoded Wallace tree multiplier's delay distribution as a function of its output bit positions [Ardekani93]. The semi-circular shape of this delay "wavefront" implies that only non-critical CSAs very close to the Least Significant Bit (LSB) and Most Significant Bit (MSB) positions will likely benefit from clustered voltage scaling. The extent to which the lower operating voltage can be scaled diminishes as we approach the central (critical path) bit positions. Since the majority of a Wallace tree's CSAs are concentrated in and around the central bit positions, the fraction of non-critical CSAs is significantly small, particularly for short bit-width multipliers which dominate DSPs. Therefore, (i) the maximum achievable dynamic power savings is limited, and (ii) any power savings obtained may be offset by the power penalty due to the slightly increased interconnect capacitances (because of the segregated high and low voltage channels in the layout [Igarashi97]) as well as the insertion of level converters at the high/low swing interfaces. These factors make clustered voltage scaling unattractive for most DSP datapath circuits.

4.3 The Mixed Swing QuadRail Methodology

The common motive behind both the architecture-driven voltage scaling and clustered voltage scaling approaches is to achieve dynamic power savings by employing multiple voltages, while still retaining their logic gate implementations unchanged. In this thesis, we investigate the usage of multiple voltages within a single gate to perform standard digital logic functions, specifically, datapath primitives. We demonstrate the ability to voltage scale more effectively than static CMOS well into the deep-submicron regime, offering substantial energy/operation reduction for static CMOS datapath circuits.

A multiple voltage circuit methodology called *Mixed Swing QuadRail* is investigated, which addresses maximum possible voltage scaling in standard submicron CMOS and SOI fabrication processes, without warranting any technology modifications. The described architecture requires four (as the name *QuadRail* suggests) power supply rails to be distributed, in order to expand the degrees of freedom available in the power-delay space of static CMOS circuits. Logic gates are implemented in multiple stages by intermixing high and low voltage signals (as the name *Mixed Swing* suggests) and substantial savings in dynamic power compared to static CMOS is obtained by driving capacitive

FIGURE 19 Booth-recoded Wallace tree delay distribution vs. output bit-position.



loads at the gate outputs at reduced voltage swings [Carley94], [Krishna96a], [Krishna96b].

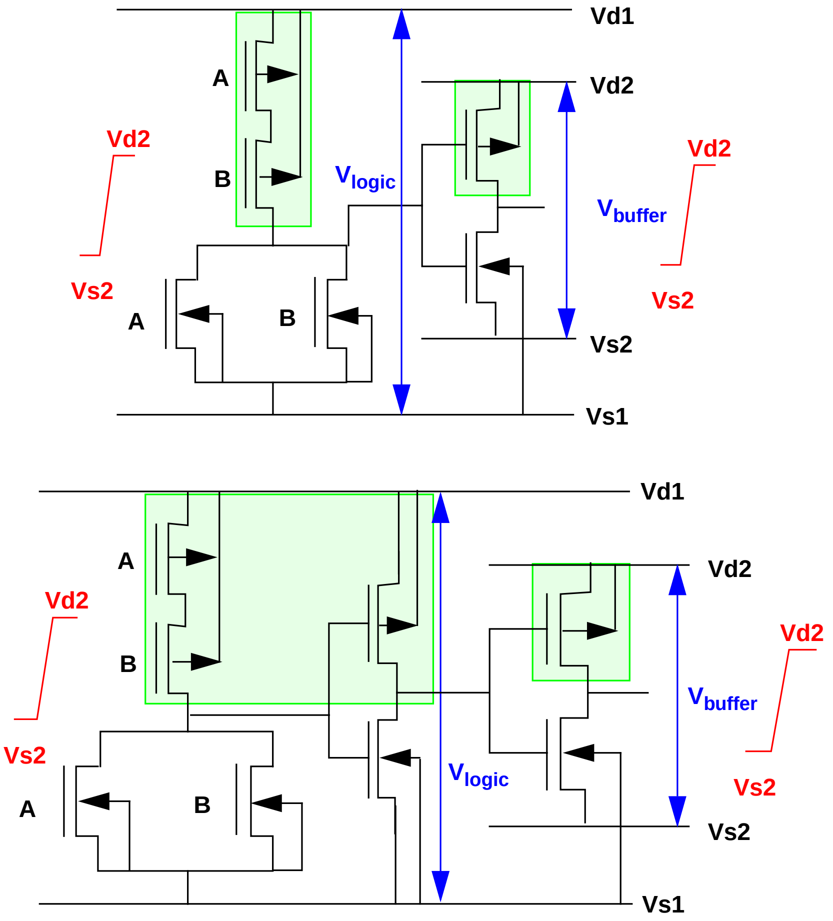
The essence of the Mixed Swing QuadRail methodology is that it allows exploitation of the best aspects of both voltage scaling and full swing static CMOS within a single logic gate. Figure 20 shows the Mixed Swing QuadRail gate architecture, consisting of a logic stage operating between the high-swinging power rails (i.e., Vd1-Vs1 = V_{logic}) and a driver/buffer stage operating between the low-swinging power rails (i.e., Vd2-Vs2 = V_{buffer}). The logic and buffer voltages are approximately centered to maximize noise margins and equalize rising and falling delays in either stage. The voltage swings are optimally selected to allow a small static current to flow in the logic stage, striking a balance between static power dissipation and performance. PMOS devices in both stages are ratioed wider than the NMOS devices to roughly equalize their respective drive capabilities. The buffer transistor widths are ratioed by a factor k (≥ 1) relative to that of logic stage transistors for improved buffer current over-drive. Each stage has its own n-well in order to minimize body effect on the PMOS devices, whereas the NMOS devices reside in the native p-substrate, staying compatible with conventional submicron n-well processes. Further, all devices in the logic and buffer stages are oriented in the same direction to minimize threshold voltage mismatches.

The buffer stage is essentially a static CMOS inverter, but with high-swinging inputs (V_{logic}) and low-swinging outputs (V_{buffer}). From Figure 20, the buffer stage gate-source on-drive voltage is approximately (V_{logic} + V_{buffer})/2 whereas the capacitive load voltage swing is only V_{buffer}. In submicron processes, for a given load capacitance and transistor aspect ratios, the buffer stage delay is related to load voltage swing and on-drive voltage as follows [Krishna97]:

$$Delay_{buffer-stage} \propto \frac{V_{buffer}}{\left(\frac{V_{logic}+V_{buffer}}{2} - V_t\right)^n} \quad \textbf{(EQ 5)}$$

where V_t is the threshold voltage and n is the velocity saturation index. n indicates the degree of carrier velocity saturation of the transistors, and is close to

FIGURE 20 Mixed Swing QuadRail (a) non-inverting and (b) inverting gate architectures.



1.0 in deep-submicron processes. Thus, on-drive currents are approximately linearly related to on-drive voltage, as opposed to the full quadratic dependence in long channel (>1μm feature size) devices [Bakoglu90]. Therefore, the ratio of load voltage swing to on-drive currents are *lower* than full-swing (V_{logic}) static CMOS, offering improved rise/fall delays at the output nodes. In addition, the reduced load voltage swing offers buffer stage dynamic and short-circuit power reduction, bounded by the ratio of V_{logic} to V_{buffer}. This enables V_{buffer} to be scaled well below the sum of the threshold voltages of the NMOS and PMOS devices while still retaining good switching performance compared to static CMOS.

The logic stage is identical to a CMOS inverting/non-inverting gate topology, except it has low-swinging inputs (V_{buffer}) and high-swinging outputs (V_{logic}), exploiting the fact that the transition region of a static CMOS gate is smaller than the complete input swing range. Similar to the buffer stage, for a given load capacitance and transistor aspect ratios, the logic stage delay is related to load voltage swing and on-drive voltage as follows [Krishna97]:

$$Delay_{\log ic-stage} \propto \frac{V_{\log ic}}{\left(\frac{V_{\log ic}+V_{buffer}}{2}-V_t\right)^n} \qquad \text{(EQ 6)}$$

Since the on-drive voltage is the same as that of the buffer stage, the relatively higher output swing causes the ratio of load voltage swing to on-drive currents to be *higher* than full-swing (V_{logic}) static CMOS, making rise/fall delays at the output nodes larger.

As feature sizes continue to shrink, both delay and power are becoming increasingly interconnect capacitance dominated rather than gate capacitance dominated. This is mainly because interconnect capacitance, dominated by coupling and fringing components, scales much slower than gate capacitance. This dominance causes the buffer stage's input gate capacitance to become less significant compared to the fraction of total load capacitance that is due to interconnect. Therefore, in current and future submicron processes, the buffer stage delay and power is becoming increasingly dominant over logic stage delay and power. This causes *overall* delay and power (i.e., sum of logic and buffer stage delay and power) to *improve* relative to full-swing static CMOS with process scaling. The delay and power savings increase with interconnect dominance and deep velocity saturation, both of which are inevitable in future deep-submicron processes. In addition, since the methodology is static and single-ended, the effective switched capacitance per cycle is identical to its equivalent full-swing static CMOS implementation. Therefore, the dynamic power savings achieved due to reduced output swing are not offset by an increase in any of the other parameters that impact dynamic power, unlike dynamic and/or fully-differential techniques. These factors make the Mixed Swing QuadRail approach best suited for large datapath circuits such as Wallace tree multipliers, where the buffer stage delay and power dominate due to the substantial interconnect capacitances at their gate outputs.

These advantages come with a modest layout area penalty that is incurred in bulk-CMOS processes because of the requirement for two n-wells within each gate as opposed to a single n-well required by its static CMOS counter-part. The area penalty is because of inter-well spacing design rules: wells maintained at different potentials require to be spaced far apart to avoid any possible encroachment caused by lateral diffusion of the implant atoms during ion-implantation of the wells [Sze83]. For a 16*16 Wallace tree multiplier in a 0.5µm bulk-CMOS process (implementation details to be described in Chapter 7), this results in an area penalty of nearly 10% over static CMOS. However, this penalty is non-existent in SOI processes due to the absence of wells. This is because the NMOS and PMOS devices are fabricated in local p-type and n-type "islands" respectively, grown epitaxially on an insulated substrate [Sze83]. This enables the sources of different devices *within* an "island" to be tied to different potentials while still satisfying only their inter-device spacing requirements.

4.4 Mixed Swing QuadRail Power-Delay Trend

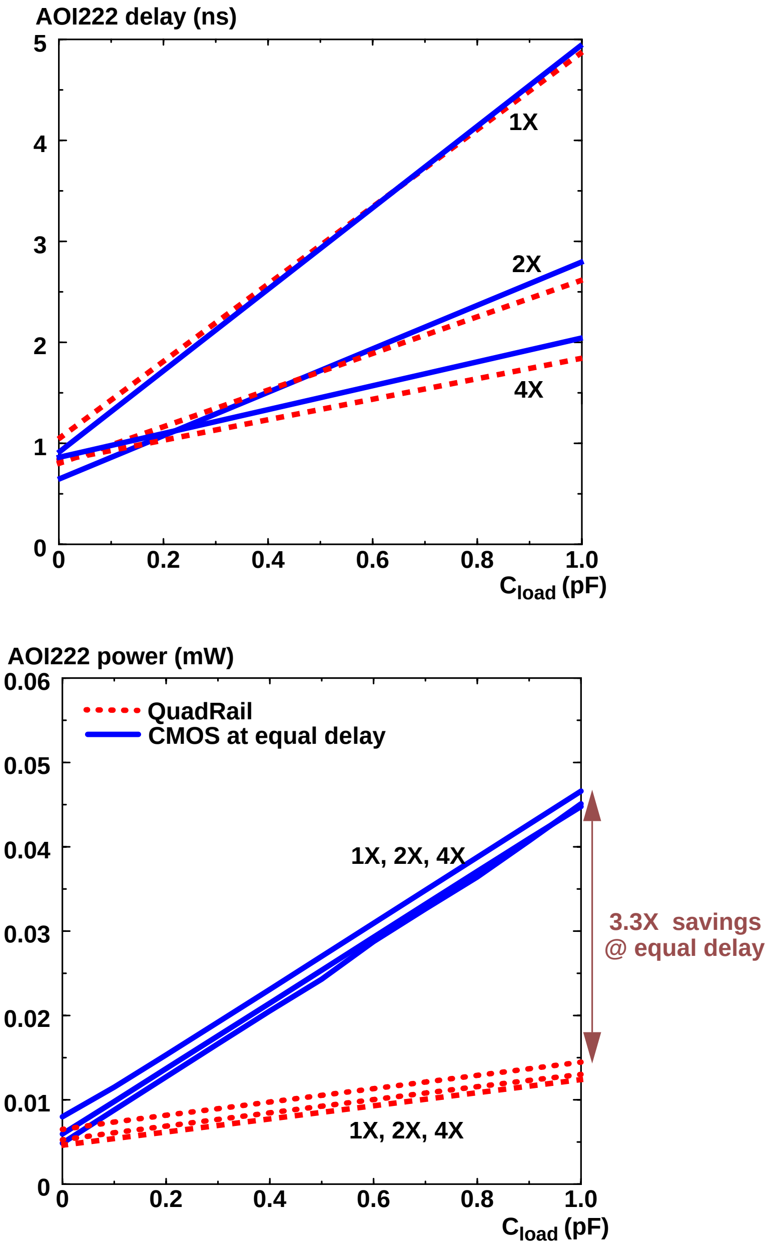
In order to illustrate the ability of the Mixed Swing QuadRail methodology to voltage scale more aggressively relative to static CMOS, Figure 21 shows the delay and power (at 100 MHz with α=1) of an example QuadRail and static CMOS AOI222 gate as a function of interconnect load capacitance in a commercial 3V,0.5µm bulk-CMOS process. 1-4X sized buffer transistors and interconnect capacitances in the range of 0-1pF are considered in both cases. The operating voltages are selected to approximately equalize their delays at any load capacitance. Delay and power are obtained through HSPICE simulations using Level13, BSIM1 models.

It is observed that with increasing interconnect load capacitance, both QuadRail and static CMOS delays increase with the same steepness, but QuadRail's rate of power increase is significantly lower than static CMOS due to the reduced load voltage swing. Thus, at C_{load} = 1pF, with equal delays, a 3.3X energy/operation reduction is obtained compared to static CMOS. The savings are even higher as interconnect capacitance increases beyond our range of analysis. At small loads (< 50fF), static CMOS and QuadRail power dissipation are almost equal at equal delays: this is due to QuadRail's logic stage static power, which becomes comparable to the buffer stage power. These observations are validated through experimental measurements (to within 10% of these HSPICE simulations) on a test-chip with chains of 17 AOI222 gates in static CMOS and QuadRail driving a range of interconnect loads (0.25mm, 0.5mm, 1.0mm and 2.0mm long, 1.2µm wide metal2 interconnects) fabricated in the 0.5µm process. Figure 22 shows the test-chip microphotograph, fabrication process characteristics, and sample measured input/output waveforms.

4.5 Multi-staged Mixed Swing QuadRail

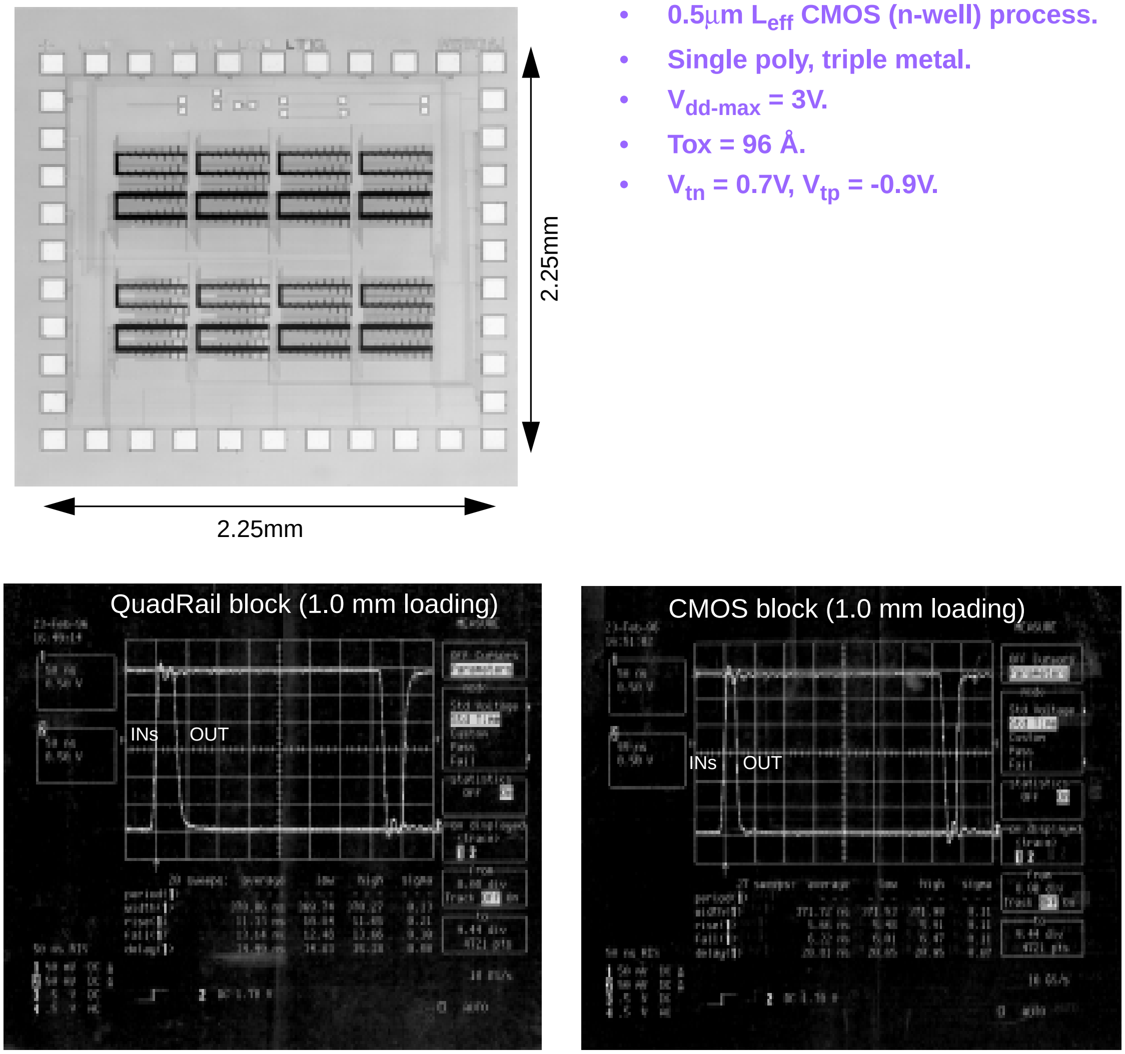
The Mixed Swing QuadRail methodology, in general, can be extended to three (or more) stages as shown in Figure 23 to allow larger voltage differences between the highest and lowest swing stages by using intermediate logic stages. The intermediate stages can be either tapered CMOS buffers/inverters or logic gates. Because the buffer's input swing is increased, the gate's output drive is greater for a given buffer transistor size. Any number of high voltage logic stages can be cascaded to form more complex functions, and followed by a buffer stage to deliver the output to the next gate. However, each additional stage requires its own independent pair of power rails which must be routed to all circuits sharing this methodology. Further, every additional voltage swing

FIGURE 21 QuadRail vs. static CMOS AOI222 delay and power vs. interconnect C_{load} trend.



requires either an explicit off-chip power supply or an on-chip series regulation mechanism. These factors make three- or higher-staged Mixed Swing Quad-Rail economically unattractive for most DSP datapath circuits. These approaches are best suited for constructing complex boolean functions (And/Nand-Or/Nor-Invert configurations) which are widely used in large control

FIGURE 22 AOI222 test-chip microphotograph, process characteristics, and sample measured waveforms.



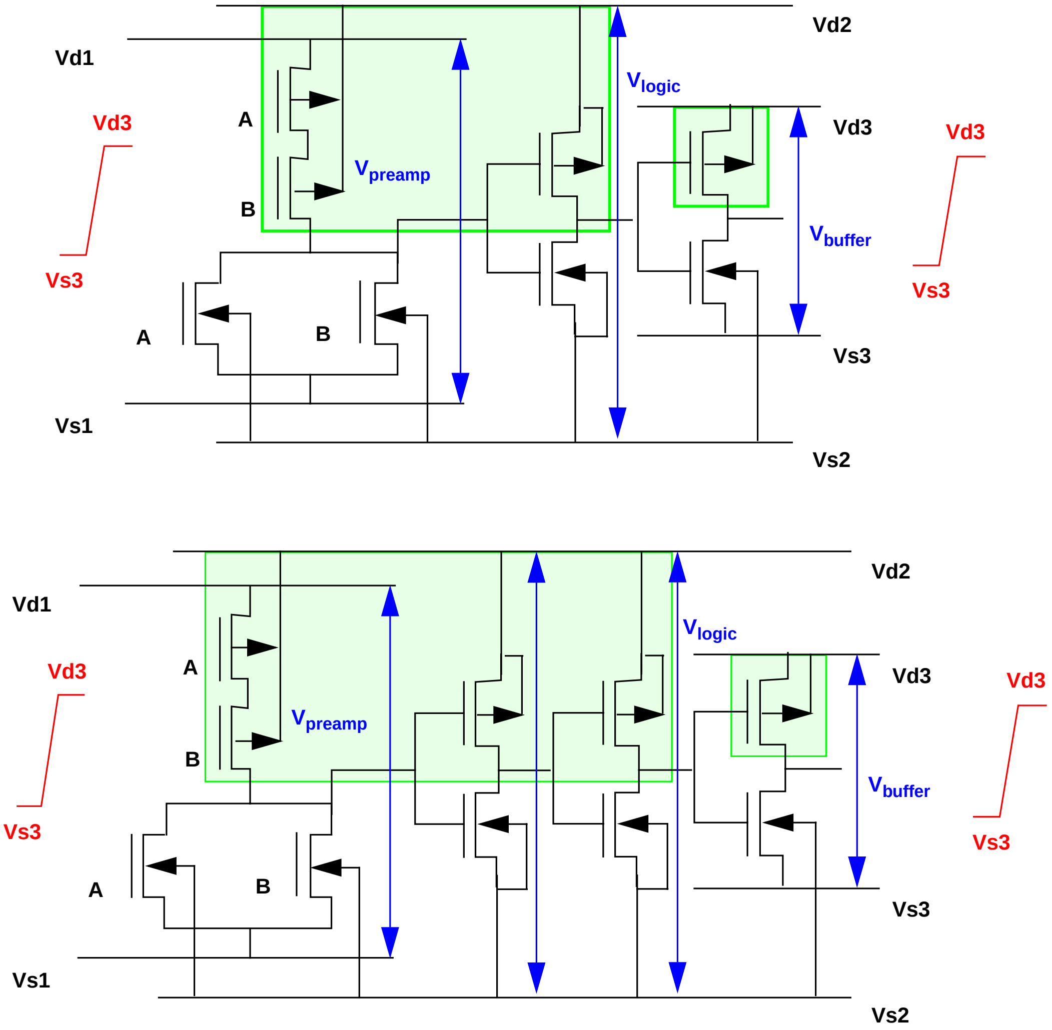
- $0.5\mu m$ L_{eff} CMOS (n-well) process.
- Single poly, triple metal.
- $V_{dd\text{-}max} = 3V$.
- Tox = 96 Å.
- $V_{tn} = 0.7V$, $V_{tp} = -0.9V$.

path/random logic circuits and which typically require tapered/buffered multi-staged gate implementations.

4.6 Alternate Mixed Swing Gate Architectures

The static, single-ended Mixed Swing QuadRail methodology described above renders itself well-suited for large datapath circuits such as multipliers and

FIGURE 23 Multi-stage QuadRail (a) inverting and (b) non-inverting gate architectures.



MACs, where interconnect capacitance dominates gate capacitance. However, there exists a whole gamut of small datapath operators such as adders and adder variants such as subtractors and comparators, where interconnect capacitance is significantly lower than gate capacitance. In order to reduce their energy/operation compared to static CMOS, Cascode Voltage Switch Logic (CVSL)-based and Domino/Pass-transistor Logic-based TriRail methodologies (employing three power supply rails) are investigated. The inherent speed advantages of CVSL [Heller84] and domino [Goncalves83] styles over static CMOS makes the proposed mixed swing counterparts best suited for small, short bit-width adders when energy/operation savings are sought compared to static CMOS.

4.6.1 Cascode Voltage Switch Logic-based TriRail

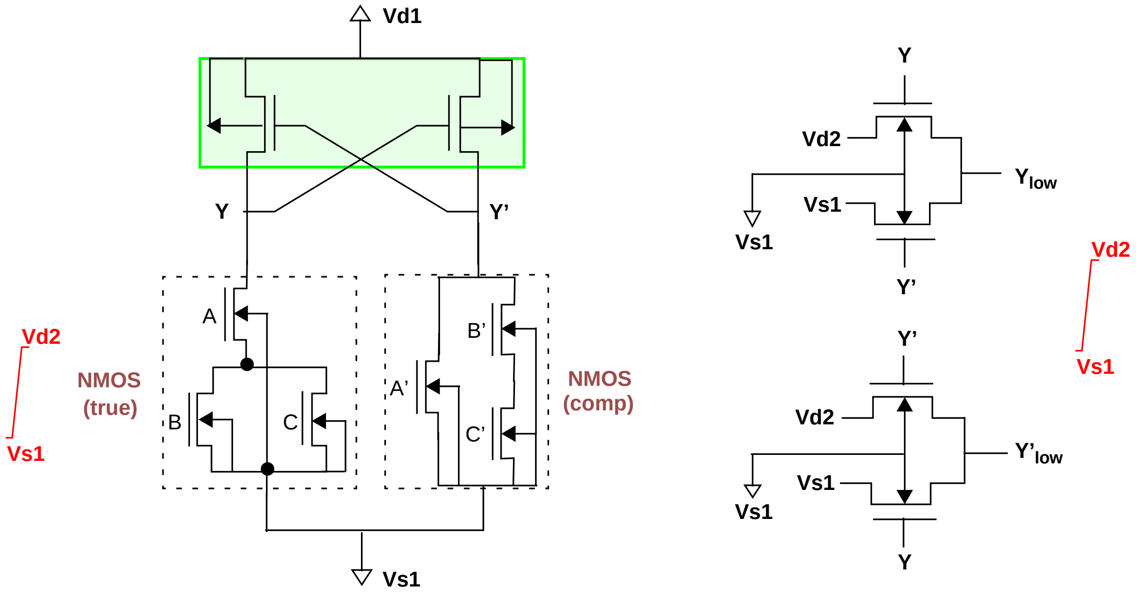
Figure 24 illustrates the static, fully-differential, CVSL-based TriRail gate architecture, where the logic stage is essentially a conventional CVSL implementation operating between a regular, high-swing supply (Vd1-Vs1) except that it has low-swinging true/complimentary inputs (Vd2-Vs1), making this a three-rail configuration. Vd2 is selected to be large enough to switch the CVSL tree for a given Vd1 and Vs1. The CVSL tree’s high-swinging outputs (Y and Y’ in Figure 24) form the control signal inputs to a pass-transistor-based buffer stage to regenerate the low swinging true/complementary outputs and drive the load capacitances at the reduced swing. The salient advantages of this approach are:

1. The inherently high-speed CVSL-based construction of the logic stage offers rapid low-to-high-swing level conversion while simultaneously performing the desired logic function.

2. The usage of NMOS devices only to implement the buffer stage offers improved transconductance gain factors per unit transistor width than equivalent static CMOS buffer stage, which requires both NMOS and PMOS devices. Thus, the buffer stage input gate capacitance driven by the high-swinging CVSL tree outputs is relatively lower, minimizing the logic stage power consumption.

3. Since the buffer stage is PMOS-free, a single n-well is adequate to accommodate the two PMOS devices of the CVSL tree, offering a layout density

FIGURE 24 CVSL-based Mixed Swing TriRail gate architecture.



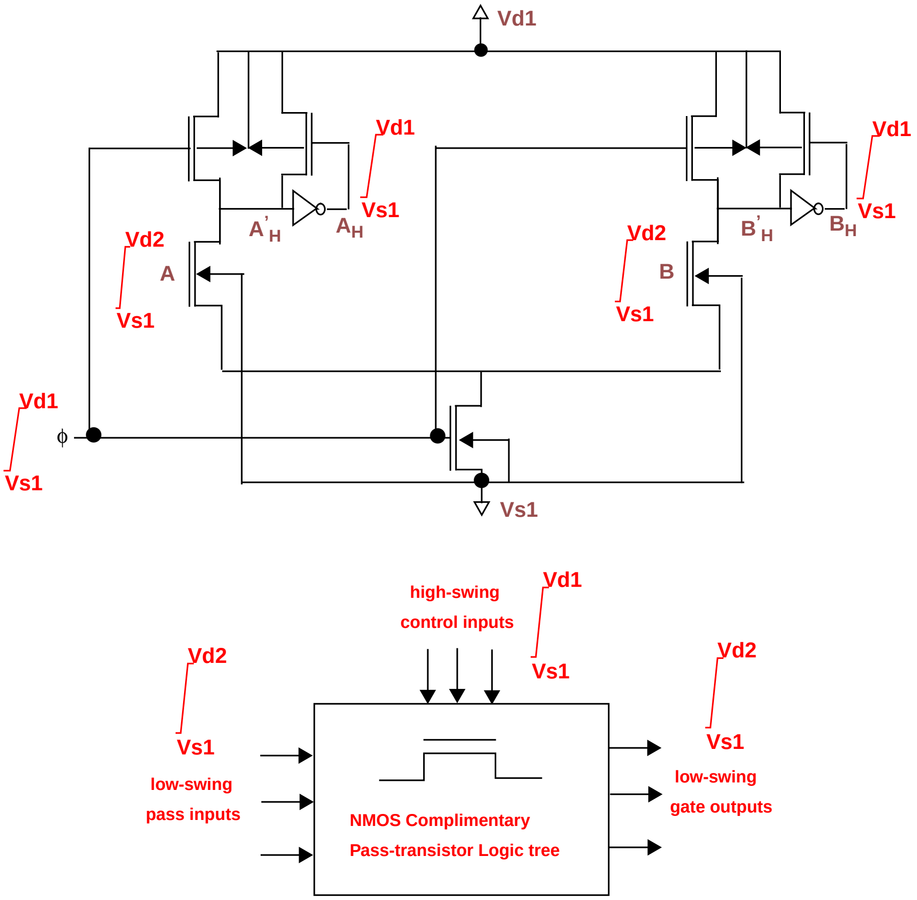
improvement compared to the twin-well-based Mixed Swing QuadRail approach.

However, since the CVSL architecture is fully-differential, true and complimentary NMOS logic trees are required. Further, true and complimentary signals require to be routed to every gate's inputs. These constitute a substantial layout area penalty, offsetting any area savings due to its single-well architecture. In addition, the fully-differential architecture implies that nearly twice the effective capacitance is switched every cycle, since each CVSL tree switches whenever its complementary tree switches. In fact, the effective switched capacitance in CVSL architectures is observed to be slightly greater than 2X due to the miller-coupling capacitance between the adjacently routed true and complimentary signals [Heller84], [Chu87], [Soma97], causing both a power and interconnect delay penalty. The increasing interconnect capacitance dominance in future deep-submicron processes further aggravates these penalties with process scaling. These factors make this approach unsuitable for large datapath, where the delay and power penalties due to fully-differential implementation offsets any delay or power savings achieved due to the reduced voltage swing. The CVSL-based mixed swing approach is well suited for small, short bit-width datapath such as adders, where the delay, power, and area penalties due to differential signalling and routing are minimal. In Section 4.6.3, we will demonstrate the energy/operation savings achieved by this approach over static CMOS for a 16-b Ripple Carry Adder over a range of operating voltages in the 0.5µm process.

4.6.2 Domino/Pass-transistor Logic-based TriRail

Figure 25 shows the single-phase (precharge/evaluate) clocked, single-ended domino/pass-transistor logic-based TriRail gate architecture. The domino preamplifier stage, operating between a regular, high-swinging supply (Vd1-Vs1), converts the single-ended low-swinging (Vd2-Vs1) inputs to high-swinging true/complimentary outputs (A_H,$A’_H$ and B_H,$B’_H$ in Figure 25). The static CMOS feedback “keeper” inverters are for improved preamplifier noise immunity against charge redistribution, and operate between the high-swing supply. The logic and buffer stages are integrated into a conventional pass-tran-

FIGURE 25 Domino/Pass-transistor Logic-based TriRail gate architecture.



sistor logic tree, except it has low-swinging pass signals and high-swinging control signals. The pass-transistor logic tree generates the low-swinging single-ended outputs and drives the load capacitances at the reduced swing. The salient advantages of this approach are:

1. The inherently high-speed domino-based preamplifier construction offers rapid low-to-high-swing level conversion to generate both true and complimentary outputs, essential for performing pass-transistor-based logic functions. Moreover, the relatively lower input gate capacitance and the absence of a pull-up/pull-down transistor stack contention current (such as in static CMOS) during switching offers further speed advantages.

2. Dynamic methodologies mandatorily precharge the output nodes every cycle and conditionally discharge during the evaluation phase. Therefore, there exists no spurious transitions within the preamplifier stage, although the pass-transistor logic/buffer stage demonstrates sneak current paths classical to pass-transistor-based logic families that may contribute to spurious transitions at the gate outputs [Izumikawa97], [Zimmer97].

3. The usage of NMOS devices only to implement the pass-transistor logic/ buffer stage offers improved transconductance gain factors per unit transistor width than equivalent static CMOS logic/buffer stages, which requires both NMOS and PMOS devices. Thus, the logic/buffer stage’s input gate capacitance driven by the high-swinging domino preamplifier stage outputs is relatively lower, minimizing the preamplifier power consumption.

4. Since the logic/buffer stage is PMOS-free, a single n-well is adequate to accommodate the preamplifier stage PMOS devices, offering a layout density improvement compared to the twin-well Mixed Swing QuadRail approach.

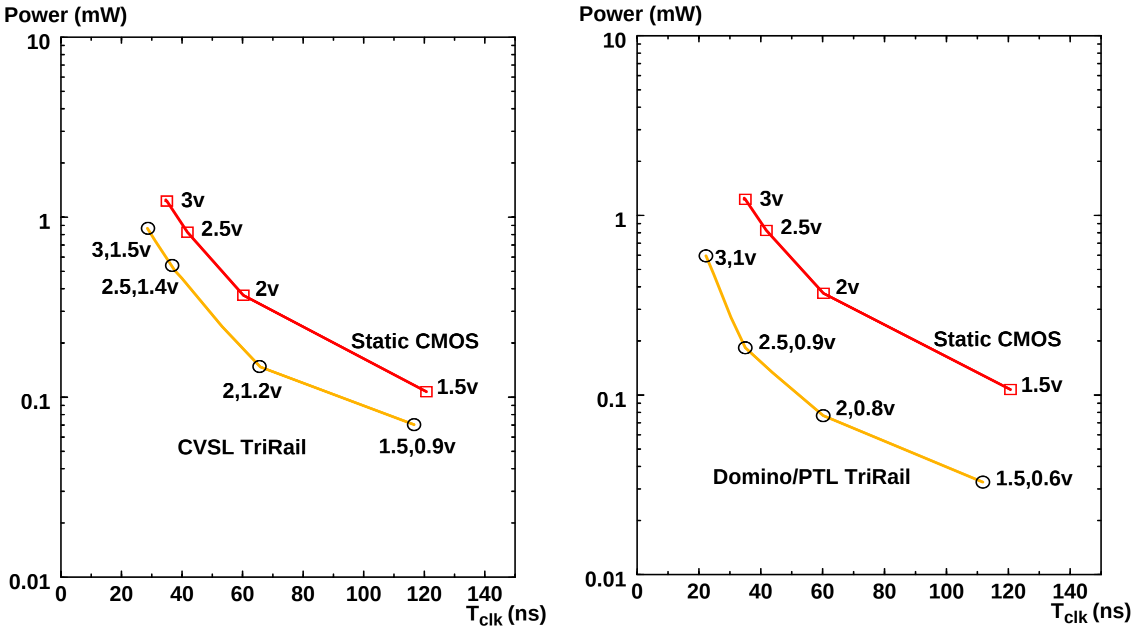
However, although this gate architecture is externally single-ended, it is internally fully-differential, since true *and* complimentary signals are required to construct pass-transistor logic trees. Therefore, the effective switched capacitance per cycle internally is nearly doubled. Further, domino approaches inherently demonstrate higher switching activities than their static counterparts, since their output nodes are precharged and evaluated every cycle, independent of input transition activity. Since the domino preamplifier's outputs are high-swinging, this constitutes a substantial dynamic power penalty. In addition, the domino preamplifier requires a high-swinging, single-phase clock (whose switching activity is unity) which is routed to every gate's precharge/evaluate devices. The interconnect capacitance due to routing this clock coupled with the precharge/evaluate device gate capacitances are charged/discharged every cycle. The increasing dominance of interconnect capacitance in future deep-submicron processes further aggravates the clock power penalty with process scaling. These factors may offset any power savings achieved due to the reduced load voltage swing in large datapath circuits. Therefore, the domino/pass-transistor logic-based TriRail approach is best-suited for small, short bit-width datapath such as adders, where the power penalties due to clock routing and internal differential signalling are minimal. In the next section, we study the energy/operation savings achieved by this approach over static

CMOS for a 16-b Ripple Carry Adder over a range of operating voltages in the 0.5µm process.

4.6.3 Adder Power-Delay Comparisons

In order to illustrate the ability of the CVSL- and domino/pass-transistor-based mixed swing methodologies to voltage scale more effectively relative to static CMOS, Figure 26 shows the power-delay comparisons between these techniques and static CMOS for a 16-bit Ripple Carry Adder in a commercial 3V, 0.5µm bulk-CMOS process. The comparisons are performed over a range of operating voltages, and across 500 pseudo-random input vectors. Delay and power are obtained through HSPICE simulations using Level13, BSIM1 models.

FIGURE 26 CVSL- and Domino/PTL-based TriRail vs. static CMOS power-delay comparisons.



It is observed that both approaches offer power as well as delay savings compared to static CMOS: while the power savings are predominantly due to the reduced output swing, the delay savings are due to both reduced output swing and the inherent speed advantage enjoyed by CVSL and domino logic families over static CMOS. The energy/operation savings for the CVSL-based approach ranges up to 1.62X. The domino/pass-transistor logic-based approach, because of domino's speed advantage over CVSL, allows increased voltage scaling than the CVSL approach at a given clock frequency. Therefore, the energy/operation savings are even higher, ranging up to 5.5X, i.e., nearly 3.5X better than the CVSL-based approach.

4.7 Summary

The usage of multiple power supply-based techniques for lowering the power consumption of static CMOS circuits was explored in this chapter. We presented earliest work on driver/receiver circuits employing multiple voltages for lowering the power consumption of off- and on-chip buses, essentially motivating the principle behind these techniques: reduced voltage swing across the load capacitance contributing to a nearly linear or quadratic dynamic power savings, depending on whether the low-swing voltage was locally generated on-chip or delivered from an explicit off-chip supply. This was followed by an examination of more recent work on employing multiple voltages *within* static CMOS circuits, specifically, the architecture-driven voltage scaling and clustered voltage scaling approaches. Advantages and limitations of these techniques were analyzed, with an emphasis on their applicability to large datapath circuits such as Wallace tree multipliers. Further, it was observed that these approaches employ multiple supplies within the circuit while still retaining the logic gate architecture unchanged.

Mixed swing techniques were then introduced, which motivate the usage of multiple voltages to construct standard digital logic gates, thereby exploiting the best aspects of both static CMOS and voltage scaling at the gate level. Static and dynamic, single-ended and fully-differential mixed swing techniques were investigated for lowering the energy/operation of datapath operators. The operating principle behind these techniques was illustrated: perform logic in multiple stages by intermixing high and low voltage signals while driving load capacitances at the gate outputs at reduced voltage swings.

A static, single-ended four power-supply-rail methodology called Mixed Swing QuadRail was investigated for lowering the power consumption of large, interconnect capacitance-dominated datapath operators such as Wallace tree multipliers. Advantages and shortcomings were outlined and the potential for high energy/operation savings relative to static CMOS, increasing with interconnect capacitance dominance, was demonstrated on a AOI222 test-chip fabricated in a 0.5µm bulk-CMOS process. CVSL-based and domino/pass-transistor logic-based TriRail approaches were also presented and their advantages and limitations were enumerated. Their ability to achieve substantial energy/operation savings over small, gate capacitance-dominated static CMOS datapath circuits was demonstrated on a 16-bit Ripple Carry Adder in the same 0.5µm process.

As mentioned in previous chapters, lowering the energy/operation of large, interconnect capacitance-dominated datapath circuits such as Wallace tree multipliers is the central focus of this thesis. Therefore, in the remaining chapters we focus our attention on the Mixed Swing QuadRail methodology and explore the extent to which we can lower energy/operation in current and future deep-submicron processes. In the next chapter, we develop optimization strategies to minimize QuadRail's energy/operation. In Chapter 6, we will investigate low-voltage challenges to QuadRail in order to demonstrate its practicality in future deep-submicron processes. Later, in Chapter 7, we perform power-delay comparisons between QuadRail and static CMOS on our prototype MAC architecture described in Chapter 3, among other datapath circuits, to demonstrate the ability to achieve substantial energy/operation savings.

5 Mixed Swing Circuits: Power-Delay Optimization

The Mixed Swing QuadRail methodology performs multi-staged logic within a single gate by employing multiple operating voltage swings. Therefore, additional degrees of freedom are introduced into its power-delay optimization space. Specifically, the logic and buffer stage transistor sizes and voltage swings are our additional degrees of freedom. While the transistor sizes are local to every QuadRail gate, the voltage swings are global across all QuadRail gates within a circuit. This thesis focuses on interconnect dominated datapath circuits, where both buffer stage delay and power significantly dominate over their logic stage counterparts. In such circuits, the logic stage transistors are typically sized minimum-width[1] in order to minimize gate capacitance loading on the fanin gates' buffer stages. The buffer stage transistors, on the other hand, require optimal sizing (i.e., wider than minimum-width) in order to minimize delay or power and to drive their large load capacitances with steep rise/fall times. However, as buffer transistor sizes increase, logic stage delay and power become comparable to the buffer stage delay and power. This is typi-

[1] Minimum-width for PMOS devices is typically 2-3X higher than the NMOS devices, since they are ratioed to approximately equalize high/low noise margins and rise/fall times.

cally addressed by (i) retaining the logic stage transistors as minimum-width and inserting tapered static CMOS inverters between the logic and buffer stages for improved buffer transistor current over-drive, and/or (ii) optimally sizing the logic stage transistors as well. However, as we will demonstrate later in Section 5.2.2, optimally sizing logic stage transistors in interconnect dominated datapath circuits does not offer any tangible improvements on QuadRail's power-delay space. On the other hand, optimally sizing buffer stage transistors is shown to offer substantial improvements on QuadRail's power-delay space. Therefore, we focus our attention on the additional degrees of freedom that have maximal potential impact: logic and buffer stage operating voltage swings and the buffer stage transistor sizes.

In this chapter we study the impact of these degrees of freedom on QuadRail's design space and explore opportunities to exploit them to minimize QuadRail circuit energy/operation. Analytical models for QuadRail power and delay are derived from submicron MOSFET I-V equations. These models are essential because they enable (i) rapidly studying QuadRail power-delay space trade-offs in current and future fabrication processes, and (ii) casting and solving a variety of QuadRail optimization problems, particularly for large circuits. The accuracy of these models is demonstrated through comparisons with HSPICE simulations using Level13, BSIM1 models. On the basis of these models, QuadRail's power-delay space is explored and optimal voltage scaling and buffer transistor sizing strategies are developed to minimize energy/operation [Krishna97]. The effectiveness of these strategies is demonstrated on a 16*16+36-bit MAC circuit in a commercial 3V, 0.5μm bulk-CMOS process.

5.1 Mixed Swing QuadRail Power, Delay Modeling

In this section, we develop QuadRail power and delay models from submicron MOSFET model I-V equations, compare their accuracy to HSPICE simulations, and evaluate power-delay trade-offs in QuadRail circuits. We propose to model both QuadRail power and delay as posynomial functions of buffer transistor size. A posynomial function P(k) of a positive variable k ∈ R is defined as [Ecker80]:

$$P(k) = \sum_{j} a_j \cdot \prod_{i=1}^{m} k_i^{b_{ij}} \qquad \text{(EQ 7)}$$

The coefficients a_j must be positive and b_{ij} must be real. Posynomial functions exhibit the distinct property that a local minimum of the function is a guaranteed global minimum. Posynomial models for power and delay are widely used for solving transistor sizing and gate sizing optimization problems for static CMOS circuits [Fishburn85], [Sapatnekar93].

One traditional approach employed in transistor-level optimization problems to model CMOS circuits is by modeling CMOS gates as RC-trees [Bakoglu90]. However, these models can deviate significantly from SPICE simulations, yielding suboptimal solutions [Hoppe90]. This is primarily due to not considering MOSFET short-channel effects which become significant at submicron feature sizes. On the other hand, developing accurate short-channel analytical models requires the usage of more precise MOSFET models, which are not only time-consuming but also require special device parameter extraction procedures. Shockley’s square-law MOSFET model [Bakoglu90] is widely used for simple analytical treatment of CMOS circuits but does not account for short-channel effects. The n^{th}-Power Law MOSFET model [Sakurai90] has been proposed as an extension to the square-law model and accounts for carrier velocity saturation and channel length modulation, both of which are dominant short-channel effects in submicron devices. Here, n is the velocity saturation index, a process-dependent parameter extracted from measured device I-V characteristics. n is approximately 1.0-1.5 for submicron processes and increases towards 2.0 with voltage scaling. This model has shown good agreement to measured I-V characteristics at least down to 0.25μm feature sizes.

We propose to employ the n^{th}-Power Law model I-V equations to develop our analytical formulations for QuadRail power and delay. Further, we take into consideration input waveform slope (approximated as a ramp signal), because of its significant contribution to delay and short circuit power [Heden87]. Our models are derived as functions of n, and hence they may be used to explore QuadRail's design space in various current and future submicron processes.

5.1.1 Analytical Delay Model

Defining Δ as the separation between rails[2], i.e., Vd1-Vd2 = Vs2-Vs1 from Figure 20, and λ as the channel length modulation factor, the differential equa-

[2] For simplicity, we assume a single Δ in our derivation. The resulting delay model can be modified for unequal NMOS and PMOS threshold voltages by substituting Δ with Δ_1 = Vd1-Vd2 for pull-up delay and Δ_2 = Vs2-Vs1 for pull-down delay, for both logic and buffer stages.

tion governing the logic stage’s output node charging/discharging is given by [Sakurai90]:

$$k \cdot C_{in} \cdot \frac{dV_{out}}{dt} = \frac{\beta_1}{2} \cdot \left(\Delta + V_{buffer} \cdot \frac{t}{t_T} - V_{t1}\right)^n \cdot (1 + \lambda \cdot V_{out}) \tag{EQ 8}$$

where, C_{in} is the input gate capacitance of a unit-sized buffer and k is the width of the buffer transistors relative to a unit-sized buffer, such that $k.C_{in}$ is the buffer stage’s input capacitance. Parasitic source/drain capacitances for the logic stage are accounted for in $k.C_{in}$. V_{out} is the time varying voltage across the buffer stage input capacitance, β_1 is the equivalent transconductance gain factor of the logic stage for short-channel devices [Sakurai91], t_T is the input rise/fall time, V_{t1} is the logic stage threshold voltage[3], and n is the velocity saturation index. Solving the above first order differential equation yields the expression for 50% rising/falling delay of the logic stage as follows:

[3] Similar to [Sakurai90], we assume NMOS and PMOS threshold voltages to be equal in our derivation. For unequal threshold voltages, V_{t1} in Equation 8 is appropriately replaced by V_{t1NMOS} or $|V_{t1PMOS}|$.

$$\begin{aligned} Delay_{\log ic} = {} & \frac{2 \cdot k \cdot C_{in}}{\beta_1 \cdot \lambda} \cdot \frac{1}{(\Delta + V_{buffer} - V_{t1})^n} \cdot \ln\left(\frac{V_{\log ic} + \frac{1}{\lambda}}{\frac{V_{\log ic}}{2} + \frac{1}{\lambda}}\right) + \\ & t_T - \\ & \left[\frac{t_T}{(n+1) \cdot V_{buffer}} \cdot \frac{1}{(\Delta + V_{buffer} - V_{t1})^n} \cdot \right. \\ & \left.\left((\Delta + V_{buffer} - V_{t1})^{n+1} - (\Delta - V_{t1})^{n+1}\right)\right] \end{aligned} \tag{EQ 9}$$

Similarly, buffer stage 50% rising/falling delay expression is derived from its governing charging/discharging first-order differential equation [Sakurai90]:

$$\begin{aligned} C_{load} \cdot \frac{dV_{out}}{dt} = {} & \\ & k \cdot \beta \cdot \left(\left\{(\Delta + V_{buffer} - V_{t2})^{n-1} \cdot V_{out}\right\} - \right. \\ & \left.\left\{\frac{V_{out}^2}{2} \cdot (\Delta + V_{buffer} - V_{t2})^{n-2}\right\}\right) \end{aligned} \tag{EQ 10}$$

where C_{load} is the QuadRail gate's load capacitance. Solving Equation 10 yields the buffer stage 50% rising/falling delay expression, given by:

$$\begin{aligned} Delay_{buffer} = {} & \\ & \left[\frac{C_{load}}{k \cdot \beta \cdot (\Delta + V_{buffer} - V_{t2})^{n-1}} \cdot \right. \\ & \left.\ln\left(\frac{4 \cdot (\Delta + V_{buffer} - V_{t2}) - V_{buffer}}{2 \cdot (\Delta + V_{buffer} - V_{t2}) - V_{buffer}}\right)\right] + \\ & m \cdot t_{1(r/f)} \end{aligned} \tag{EQ 11}$$

where, $t_{1(r/f)}$ is the logic stage output's 10% to 90% rise/fall time, given by:

$$
\begin{aligned}
t_{1(r/f)} = {} & \left[\frac{2 \cdot k \cdot C_{in}}{\beta_1 \cdot \lambda} \cdot \frac{1}{(\Delta + V_{buffer} - V_{t1})^n} \cdot \ln\left(\frac{0.9V_{logic} + \frac{1}{\lambda}}{\Delta + V_{buffer} - V_{t1} + \frac{1}{\lambda}}\right)\right] + \\
& t_T - \\
& \left[\frac{t_T}{(n+1) \cdot V_{buffer}} \cdot \frac{1}{(\Delta + V_{buffer} - V_{t1})^n} \cdot \left((\Delta + V_{buffer} - V_{t1})^{n+1} - (\Delta - V_{t1})^{n+1}\right)\right] + \\
& \left[\frac{k \cdot C_{in}}{\beta_1 \cdot (\Delta + V_{buffer} - V_{t2})^{n-1}} \cdot \ln\left(\frac{2 \cdot (\Delta + V_{buffer} - V_{t2}) - 0.1V_{logic}}{0.1V_{logic}}\right)\right]
\end{aligned}
\tag{EQ 12}
$$

where β is the transconductance gain factor of a unit-sized transistor, V_{t2} is the buffer stage threshold voltage[4], and m is an empirically fitted constant for a given set of voltage swings[5].

[4] Logic and buffer stage threshold voltages, i.e, V_{t1} and V_{t2} are different because opposite type devices are in conduction in either stage for any input combination that causes a transition at the output.

[5] Since only a portion of the logic stage output's slope affects the buffer stage delay, the input waveform slope's contribution is empirically fitted through HSPICE Level13, BSIM1 models in our analysis.

Increasing the buffer transistor size (k) leads to increased loading on the logic stage and hence logic stage delay. This, however, improves the buffer current drive, thereby decreasing buffer stage delay. Thus, *QuadRail delay is a posynomial function of buffer transistor size (k)* and there exists a *delay optimum* at which delay is minimized.

5.1.2 Analytical Power Model

The dynamic power dissipated by a QuadRail gate driving a load capacitance C_{load} can be expressed as the sum of the energies drawn by each stage from their respective supply rails over one clock cycle [Chandra95], i.e.,

$$\begin{aligned} P_{dyn} = \alpha \cdot k \cdot C_{in} \cdot (V_{\log ic})^2 \cdot f_{clk} + \\ \alpha \cdot C_{load} \cdot (V_{buffer})^2 \cdot f_{clk} \end{aligned} \qquad \text{(EQ 13)}$$

where, α is the switching activity and f_{clk} is the input signal frequency. Parasitic source/drain capacitances for the buffer stage are accounted for in C_{load}. The short-circuit power in the logic stage is given by [Sakurai90]:

$$P_{sc} = \alpha \cdot \frac{1}{n+1} \cdot \frac{1}{2^{n-1}} \cdot \frac{\beta_1}{2} \cdot (V_{drive} - 2V_{t1})^{n+1} \cdot t_T \cdot f_{clk} \qquad \text{(EQ 14)}$$

where, V_{drive} is the gate-source on-drive voltage, i.e., $(V_{logic} + V_{buffer})/2$. Equation 14 converges to the static CMOS short circuit power expression in [Sakurai90] when $V_{buffer} = V_{logic}$. Static power dissipation in the logic stage is given by:

$$P_{static} = I_{off} \cdot V_{\log ic} \qquad \text{(EQ 15)}$$

where, I_{off} is the logic stage off-current. If the logic stage gate-source off-drive voltage, i.e., $(V_{logic} - V_{buffer})/2$, is lesser (greater) than V_{t1}, the off devices are in subthreshold (strong inversion). Both short-circuit and static power dissipation are negligible for the buffer stage due to its reduced voltage swing and negative off-drive voltage respectively. As the buffer transistor size (k) increases, logic stage loading increases, increasing its dynamic power. This, however, decreases the buffer's output transition time and hence the input transition times for all fanout logic stages (t_T in Equation 14, which is a function of $1/k$), thereby reducing their short circuit power; the larger the number of fanouts, the more significant this reduction. Then, total QuadRail power consumption may be modeled as:

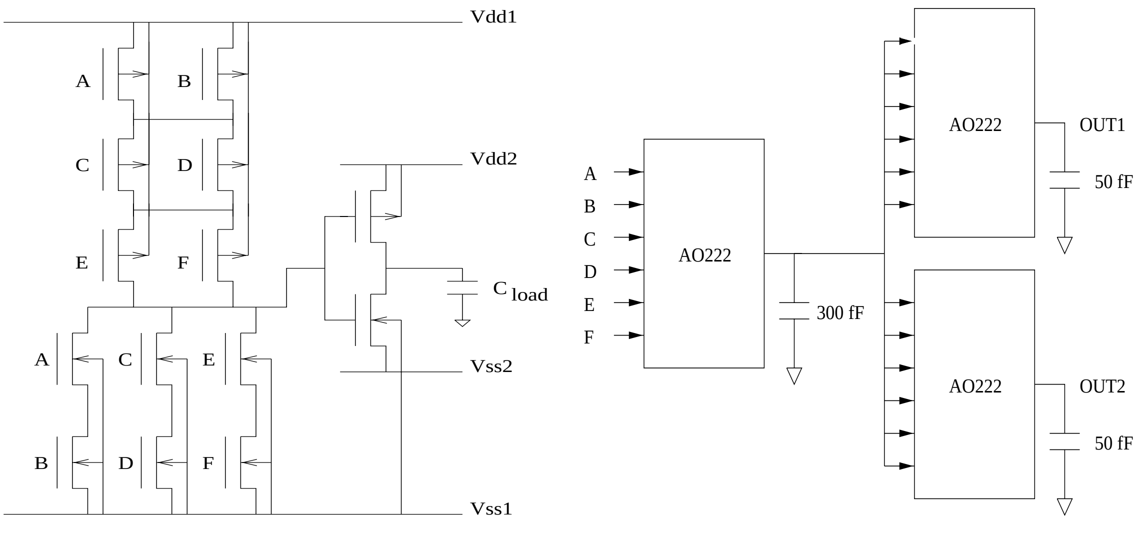
$$P_{total} = P_{dyn} + P_{static} + P_{sc} = A \cdot k + B + \frac{C}{k} \qquad \text{(EQ 16)}$$

From Equation 7 and Equation 16 we observe that *QuadRail power dissipation is also a posynomial function of buffer transistor size (k)* and there exists a global *power optimum* at which power is minimized.

5.1.3 Accuracy of Power, Delay models

In this section, we present comparisons of our models with HSPICE simulations using Level13, BSIM1 models in the 0.5μm process. Through measurements on the QuadRail test-chip described in Chapter 4 and HSPICE simulations using Level13, BSIM1 models, the value of n for this process was determined to be approximately 2.0 for voltages ≤ 3.0V. An experimental QuadRail circuit setup is considered for the comparisons as shown in Figure 27. The setup consists of a 6-input And-Or (AO222) gate cascade circuit. The driving gate drives all the fanout gates' inputs in addition to a capacitive load of 300fF (corresponding to approximately 2500μm of metal1 interconnect in the 0.5μm process). The fanout gates have unit-sized buffer transistors. Figure 28 shows the power (at 50MHz with $\alpha = 1$) and delay for this setup obtained at one operating point: $V_{logic} = 2.2V$ and $V_{buffer} = 0.8V$. The models show good agreement to HSPICE simulation results; the optimal buffer transistor sizes (*power optimum* and *delay optimum*) predicted by our models is within 2% of HSPICE results over a range of operating voltages (up to 3.0V) and capacitive loads studied. Note that both our models and HSPICE simulations correctly show a less steeper delay penalty for over-sizing than under-sizing as expected. This is due to the relative dominance of the logic and buffer stage delays in the total delay expression (Equation 9 and Equation 11 respectively).

FIGURE 27 QuadRail 6-input AND-OR (AO222) gate and AO222 experimental circuit setup.



5.1.4 Exploring QuadRail Power-Delay Space Trade-offs

In this section, we will employ our power and delay models to study the impact of our degrees of freedom on QuadRail's power-delay space and evaluate the power-delay trade-offs in the 0.5µm process. Figure 29 shows the delay and power for the same circuit setup as in Figure 27 obtained from our models with V_{buffer} = 0.8V, buffer transistor size *(k)* for the driving gate varying from 1X (minimum-width buffers) up to 10X, and V_{logic} varying from 1.5-3.0V. Also shown in Figure 29 are snapshots of the delay and power as a function of buffer transistor size at V_{logic} = 1.5V and V_{logic} = 3.0V. Some important conclusions can be drawn from these graphs:

- As V_{logic} approaches 3.0V, on-drive currents of both logic and buffer stages is increased, leading to reduced delays, despite an increase in the off-currents. Scaling V_{logic} towards 1.5V causes a hyperbolic delay increase in

FIGURE 28 QuadRail delay, power models compared to HSPICE Level13, BSIM1 simulations.

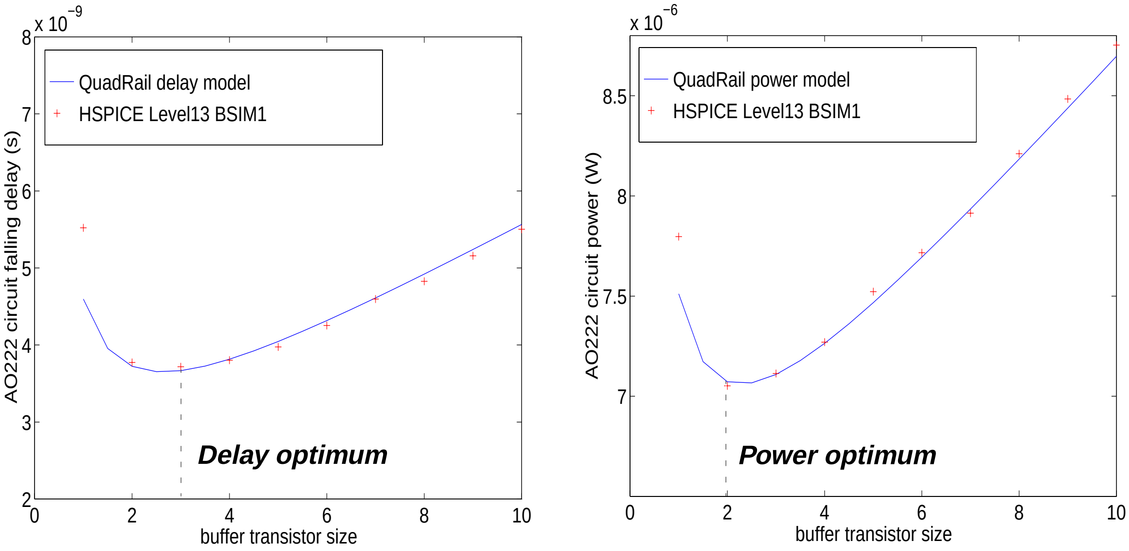
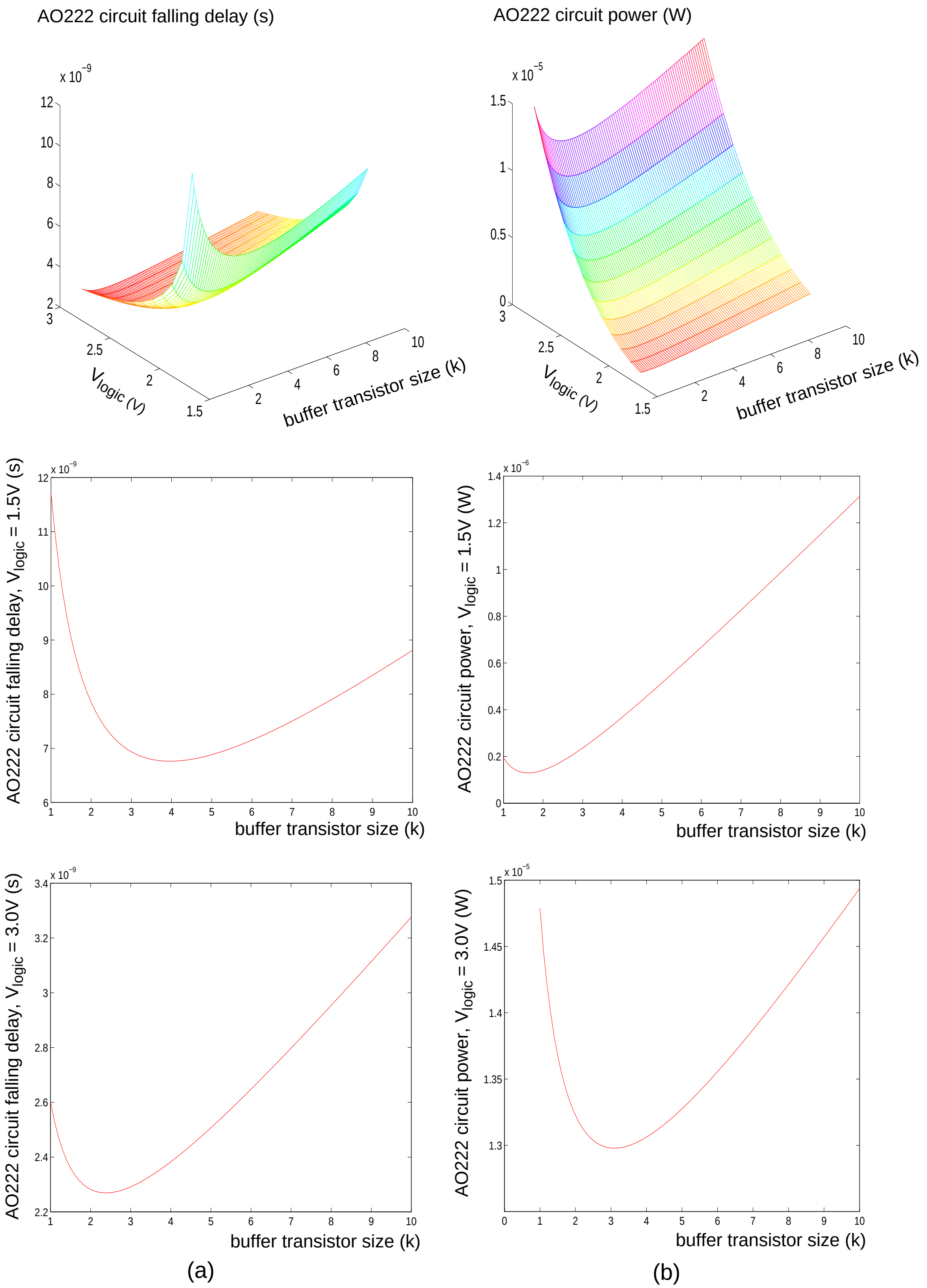


FIGURE 29 AO222 circuit delay and power vs. V_{logic} and buffer transistor size (k).



both logic and buffer stages, classical to static CMOS-based gate topologies [Bakoglu90].

- As V_{logic} approaches 3.0V, the increased buffer drive currents flatten the delay curve, i.e, the delay becomes less convex with increasing V_{logic}. Hence, although an optimal buffer transistor size exists at high logic stage voltage swings, the delay improvement obtained is not significant. Scaling V_{logic} towards 1.5V, i.e., tighter logic stage turn-off, causes steep delay penalties for non-optimal sizing, both for over- and under-sized buffers. The delay penalties for not sizing the buffer transistors at their *delay optimum* become more severe with even smaller buffer voltage swings (i.e., < 0.8V) or increased capacitive loads. Section 5.2.2 describes our approach for optimal buffer transistor sizing in QuadRail.
- As V_{logic} approaches 3.0V, short-circuit dissipation of the fanout gates is a significant component of total circuit power. This is particularly true with minimum-width buffers. When buffer transistor size is increased beyond minimum width, the driving gate’s output edge becomes steeper lowering the short-circuit power of the fanout gates and hence total power. When buffer size increases beyond the *power optimum*, dynamic power due to increased capacitive load dominates and total power starts increasing monotonically with buffer transistor size. Scaling V_{logic} towards 1.5V diminishes short-circuit power nearly cubically, and power penalty due to unit-sized buffers also diminishes. Thus, at reduced voltages, although there exists a *power optimum,* it is very close to minimum size.

- As V_{logic} approaches 3.0V, separation between logic and buffer stage swings is increased. Consequently, totempole off-currents in logic stage are substantially increased beyond nominal leakage currents. The increased static power may dominate total power. Moreover, the increased static currents reduce the steepness of the transfer characteristics and degrade noise margins. Scaling V_{logic} towards 1.5V causes improved turn-off lowering both static and dynamic power dissipation. Thus, selection of V_{logic} for a given V_{buffer} or vice versa involves careful consideration of static currents and noise margin degradation. Selection of V_{buffer} itself is determined by minimum noise margin requirements and target clock frequency constraints. Section 5.2.1 describes our approach for optimal voltage scaling in QuadRail.

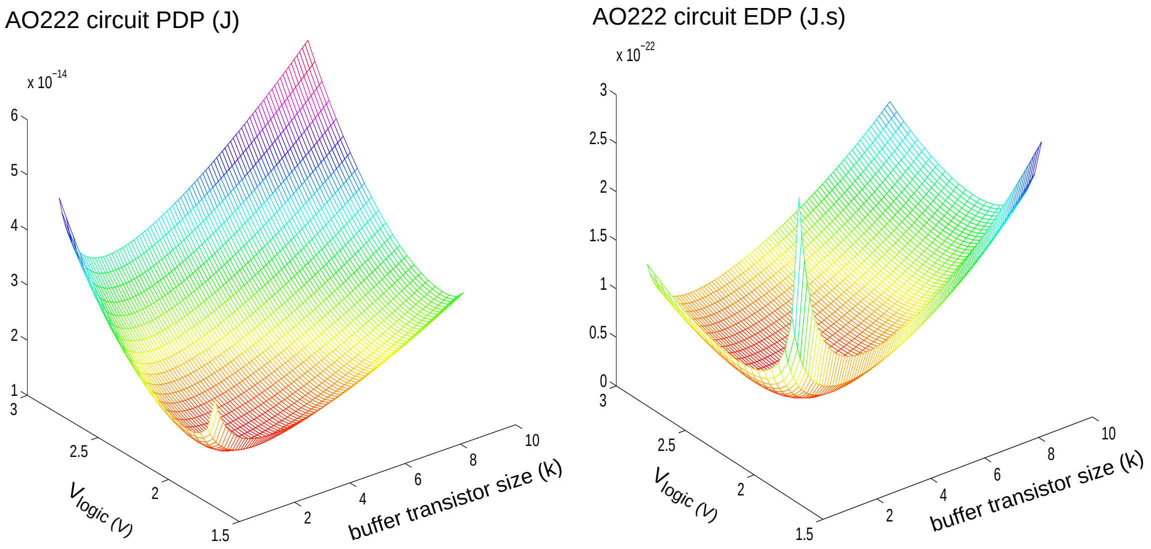
5.1.5 QuadRail Power-Delay Product, Energy-Delay Product Trade-offs

We now examine the effect of our degrees of freedom on QuadRail circuit power-delay product (PDP), i.e., power*delay, and energy-delay product (EDP), i.e., power*(delay)^2, two commonly employed metrics to compare power-delay trade-offs between circuit methodologies [Horowitz94], [Ko95]. Figure 30 shows the PDP and EDP for the same experimental setup as in Figure 27. Since V_{logic} has orthogonal effects on power and delay, and since both QuadRail power and delay are posynomial functions of buffer transistor size, QuadRail PDP and EDP are two-dimensional posynomial functions [Ecker80] of V_{logic} and buffer transistor size, i.e., there exists global optimal V_{logic} and *k* values at which PDP and EDP are minimized. Both non-optimal voltage scaling and buffer transistor sizing causes steep PDP/EDP penalties, emphasizing the importance of optimally selecting these quantities both from power and delay perspectives.

5.2 Mixed Swing QuadRail Optimization

For Mixed Swing QuadRail circuits, we assume the logic voltage swing to be the same as the power supply of peripheral static CMOS circuits to ensure I/O compatibility between QuadRail and the different static CMOS modules on-chip as well as off-chip. From a power savings point of view we would like to operate at the absolute smallest V_{buffer} and V_{logic} possible under noise margin constraints. Unfortunately, aggressive delay constraints may require a larger V_{buffer} and V_{logic} for increased buffer drive currents, forcing the designer to pay the quadratic dynamic power penalty. Given a global V_{logic} specification,

FIGURE 30 AO222 circuit PDP and EDP vs. V_{logic} and buffer transistor size (k).



we describe in this section, strategies to optimally select V_{buffer} and buffer transistor sizes. We then demonstrate their effectiveness in optimizing the energy/ operation of a 16*16+36-bit MAC circuit in the 0.5μm process, given various target clock frequencies. We do not place a constraint on total active area usage during optimization, but this feature can be introduced easily at the cost of obtaining sub-optimal solutions [Hoppe90].

5.2.1 Optimal Voltage Scaling

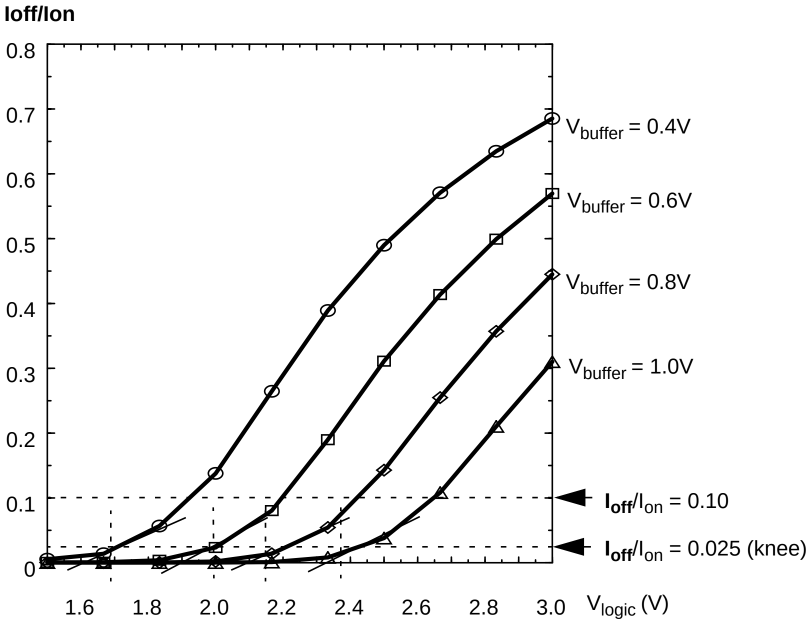
As mentioned in Section 5.1.4, selection of V_{logic} and V_{buffer} in QuadRail is critical for optimizing static power as well as noise margin degradation. In order to ensure adequately turned-off devices in the logic stage, we must restrict the off-currents to a small fraction of the average on-drive currents, striking a balance between static and dynamic power. Figure 31 shows the ratio of logic stage totempole off-current (I_{off}) to the worst-case on-drive current (I_{on}) for various V_{logic} and V_{buffer} values for the QuadRail gate in Figure 27 in the 0.5μm process, obtained through HSPICE simulations. It is observed that all graphs have two distinct regions - a steeply falling region, where I_{off} falls quadratically with V_{logic} due to strong inversion, and a flat region where I_{off} falls exponentially with V_{logic}, due to sub-threshold conduction. I_{on} falls linearly with V_{logic} in both regions. Selecting an I_{off}/I_{on} ratio defines unique buffer voltage swings at these logic voltage swings; the smaller this ratio, the better the turn-off.

If α is the circuit switching activity and N_d is the average logic gate depth per pipeline stage for a QuadRail circuit, the optimal I_{off}/I_{on} ratio to balance static and dynamic power, is given similar to [Burr91] as:

$$\left.\frac{I_{off}}{I_{on}}\right|_{optimal} = \frac{\alpha}{N_d} \tag{EQ 17}$$

As an example, I_{off}/I_{on} ratios of 0.025 (corresponding to the "knee" points) and 0.1 are chosen from Figure 31, corresponding to α = 0.025 and 0.1 respectively (since we are considering a single QuadRail gate in Figure 31, N_d = 1 for this case). The static currents are approximately 2.5% and 10% of the average on-drive currents. Figure 32 shows these example points on a V_{logic} vs. V_{buffer} plot. It is observed that the graphs are approximately linear, and each point on this line defines a unique pair of voltage swings satisfying the desired I_{off}/I_{on} ratio. In general, any QuadRail circuit with an activity factor α and an average

FIGURE 31 Off- to on-drive current ratios vs. logic stage voltage.



gate depth N_d is mapped onto the V_{logic} vs. V_{buffer} space as an approximate linear plot, having the form:

$$V_{logic} \approx V_{buffer} + \delta \cdot \left.\frac{I_{off}}{I_{on}}\right|_{optimal} \cdot 2V_{t1} \qquad \text{(EQ 18)}$$

where, δ is an empirically fitted constant and the optimal I_{off}/I_{on} ratio for that circuit is defined by Equation 17 and is the same at every point on the linear plot. Note that as the I_{off}/I_{on} ratio approaches zero, V_{logic} approaches V_{buffer}, i.e., fully static CMOS operation. Exactly which operating point (V_{buffer}, V_{logic}) is selected on this line depends on the designer's target clock frequency specifications; tighter delay constraints will force selection of higher voltage swings requiring higher power penalties. Thus, scaling down operating logic and buffer voltage swings along this line offers an efficient technique for simultaneous reduction of static and dynamic power, without degrading noise margins while ensuring adequately tight turn-off characteristics.

5.2.2 Optimal Buffer Transistor Sizing

From Equation 11 it is seen that for large load capacitances, typical along critical delay paths of digital circuits, minimum-width buffers have inadequate current drives and high delays. Since QuadRail delay is modeled as a posynomial function of buffer transistor size, there exists an optimal buffer size for which delay is minimized. This *delay optimum* is computed for every critical path gate as follows:

From Equation 9-Equation 12, total QuadRail gate delay can be expressed as:

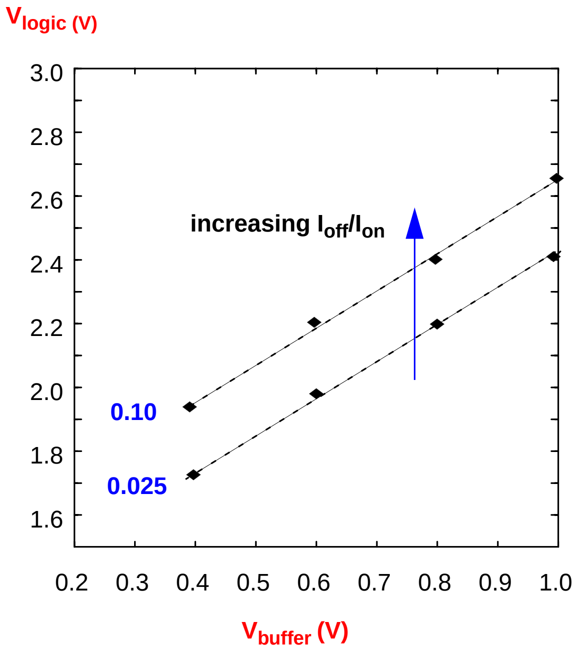
$$Delay_{total} = A \cdot k + B + C \cdot \frac{1}{k} \quad \text{(EQ 19)}$$

where, A, B, and C are the other design factors and process parameters independent of *k* from Equation 9-Equation 12. This posynomial expression has a global minimum, which is the *delay optimum*, given by:

$$k_{optimum} = \sqrt{\frac{C}{A}} \quad \text{(EQ 20)}$$

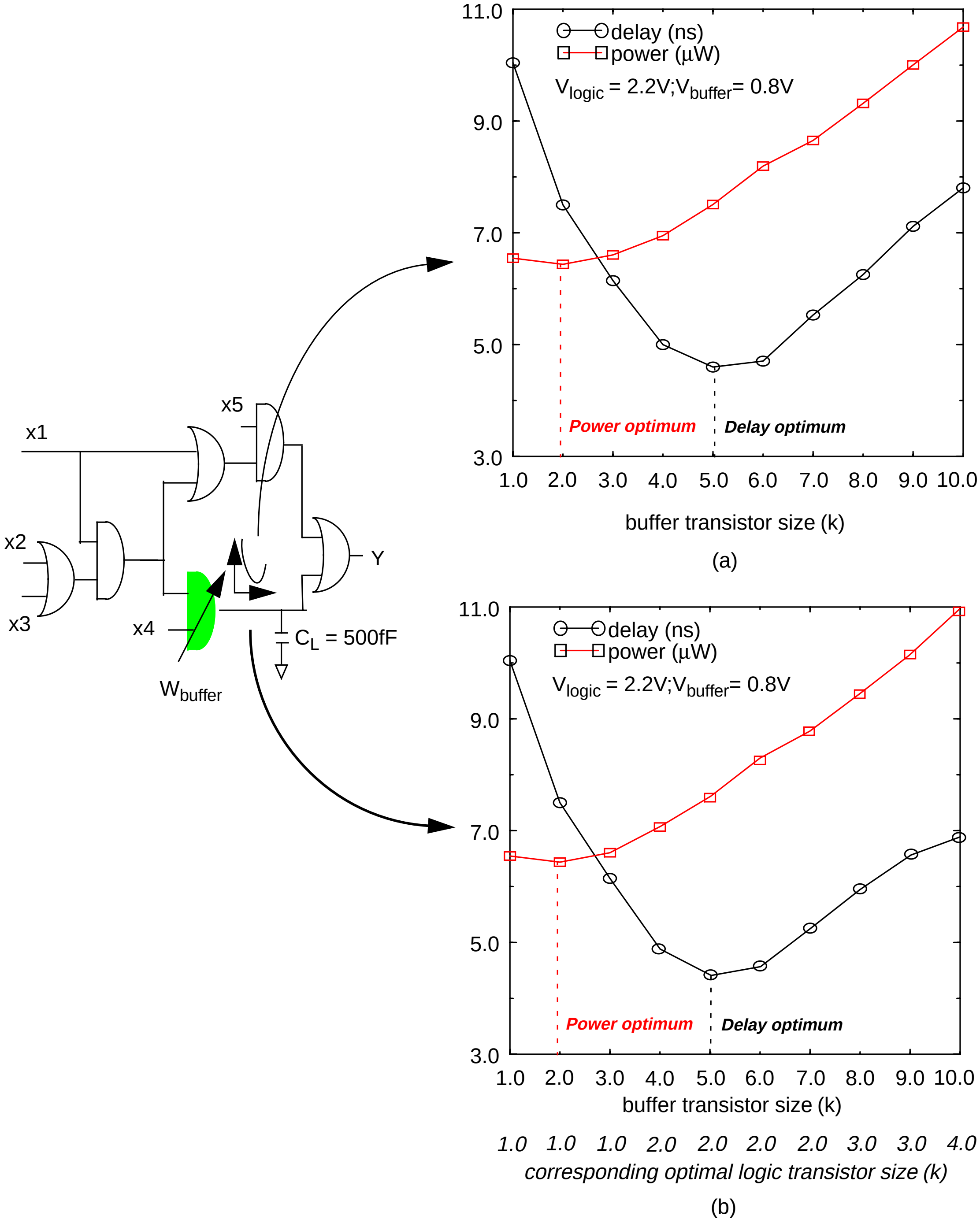
The optimal buffer transistor size depends on $\sqrt{C_{load}}$, $\sqrt{\beta_1}$, and is a non-linear function of the voltage swings. Since QuadRail power is also a posyno-

FIGURE 32 Logic vs. buffer stage voltage swing with I_{off}/I_{on} = 0.025 and 0.10.



mial function of buffer size, there exists a value of *k*, for which power is also minimized. In general, larger the fanout, larger the delay and power reduction obtained due to sizing the driving buffers at their *delay* and *power optima*. Thus, a QuadRail circuit with all transistors sized minimally is neither delay optimal nor power optimal, and increasing the buffer transistor size towards the *delay optimum* simultaneously offers a delay and power reduction. This continues until power starts to increase monotonically beyond the *power optimum*. Figure 33(a) illustrates this behavior for an example critical circuit delay path containing a 2-input AND gate driving a 500fF capacitive load in addition to a single fanout. Also shown are the *power* and *delay optima* for the AND gate for V_{logic} = 2.2V and V_{buffer} = 0.8V in the 0.5µm process. Increasing the AND gate's buffer transistor size beyond unit-size to its *power optimum* of 2X offers only a slight reduction (< 2%) in its contribution to total power. However, sizing the buffer transistors at their *delay optimum* of 5X offers a 2.2X reduction in its contribution to critical path delay. Increasing the AND gate's buffer transistor size beyond the *power optimum* to the *delay optimum* costs additional dynamic power in its logic stage; the power penalty due to delay optimal sizing is 15% higher than with minimum-sized buffers. Figure 33(b) illustrates the impact of optimally sizing both the logic stage and buffer stage transistors for the same experimental setup. For the same range of buffer transistor sizes (1X-10X), the corresponding optimal logic transistor sizes to minimize delay are determined through HSPICE simulations using Level 13, BSIM1 models. The optimal logic transistors sizes are shown under their respective buffer sizes in Figure 33(b). It is observed that since the delay and power are concentrated at the buffer stage, optimal logic stage sizing does not significantly impact delay

FIGURE 33 Optimal (a) buffer transistor sizing and (b) buffer and logic transistor sizing for an example critical circuit delay path.



or power. Thus, for buffer sizes in the range 1X-3X, the optimal logic transistor size continues to remain minimum-width (1X). For buffer sizes beyond 3X, logic stage delay's contribution becomes significant requiring it to be upsized beyond minimum-width to its optimum. However, the delay improvement achieved is only 1.047X (4.5%) at the buffer stage *delay optimum* of 5X. This is because of the continued buffer stage delay dominance. It is only beyond the buffer stage *delay optimum* of 5X that optimal logic transistor sizing offers any tangible delay savings, wherein logic stage delay is a significant portion of total delay. However, as mentioned earlier in this section, buffer sizes beyond the *delay optimum* result in both power and delay penalties and are therefore best avoided. Further, increasing the logic transistor sizes beyond minimum-width causes a monotonic power penalty, since it increases both the logic stage short-circuit power and the dynamic power of the fanin gate's buffer stages. Thus, optimal logic transistor sizing does not offer any significant improvements on the power-delay space beyond that offered by optimal buffer transistor sizing.

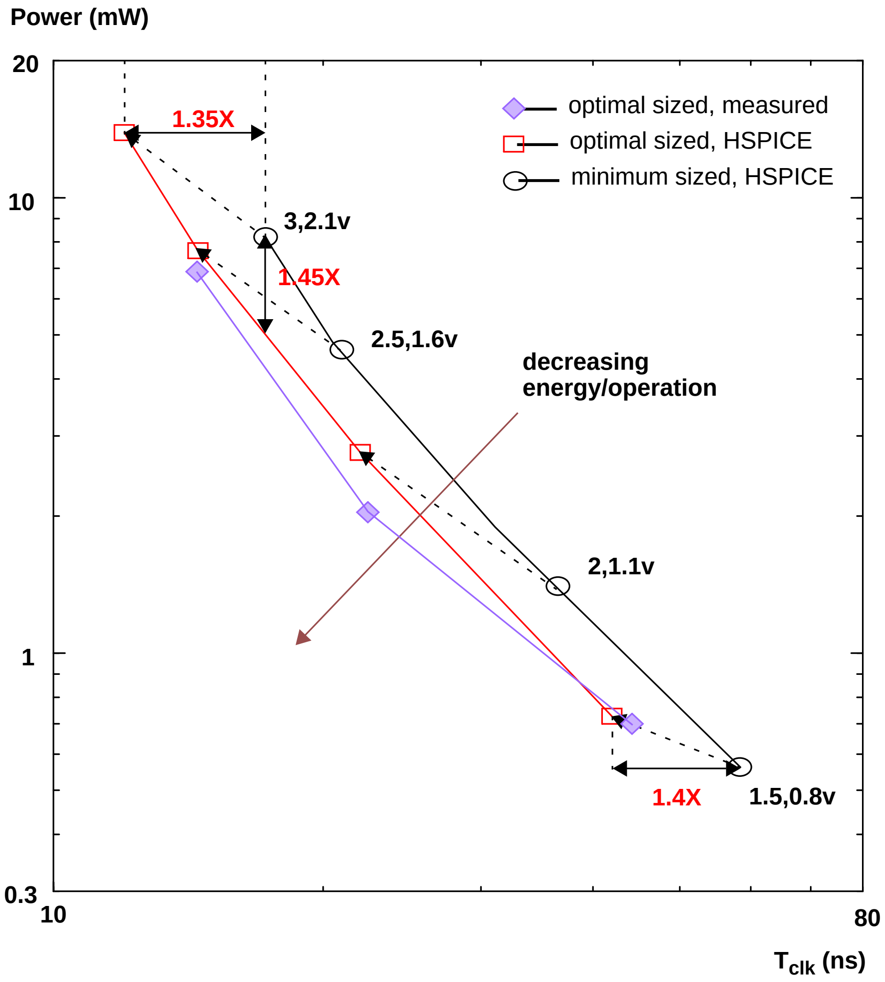
The effect of optimal voltage scaling and buffer transistor sizing on QuadRail's power-delay characteristics was first demonstrated on a 17-net ISCAS'85 combinational benchmark circuit (c17) [Brglez85] in the 0.5μm process, achieving up to 2.2X improvement in energy/operation [Krishna97]. Motivated by these results, we examine the effectiveness of these optimization techniques for a 16*16+36-bit QuadRail MAC implemented in our prototype architecture in the 0.5μm process.

5.2.3 16*16+36-bit MAC Optimization

The optimal voltage scaling and buffer transistor sizing techniques are applied to the QuadRail Wallace tree multiplier of a 16*16+36-bit MAC in the 0.5µm process. Implementation details will be described in Chapter 7. Optimal buffer transistor sizes are computed analytically for the Booth encoders, Booth multiplexors, and CSAs within the multiplier on the basis of Diva extracted parasitic capacitances at their outputs from the fully placed and routed MAC layout. A standard cell library of these primitives with multiple buffer sizes adopting a single cell footprint is created; thus, buffer resizing does not entail any layout modifications. A range of logic and buffer voltage swings is considered (V_{logic} = 1.5-3.0V and V_{buffer} = 0.8-2.1V), governed by the affine relationship V_{logic} = V_{buffer} + 0.9 for V_{logic} = 3.0V, 2.5V, and 2.0V and by the affine relationship V_{logic} = V_{buffer} + 0.7 at V_{logic} = 1.5V, corresponding to an optimal I_{off}/I_{on} ratio of 0.006667 (1/150). This is because with V_{logic} scaling, static power dominance increases relative to dynamic and short-circuit power, requiring a tighter turnoff at lower V_{logic} to maintain the same optimal I_{off}/I_{on} ratio.

Since the MAC was fabricated in the 0.5µm process, optimal sizing is performed at one operating point (V_{logic} = 3V, V_{buffer} = 2.1V) and then optimally voltage scaled. Figure 34 shows the multiplier power vs. T_{clk} characteristics for unit-sized buffer transistors (right), and with buffer transistors sized optimally (left), over our range of voltage swings. Power and delay are measured across 500 pseudo-random input vectors. Optimal scaling and sizing is observed to offer an essentially diagonal movement of the power-delay characteristics towards the origin, i.e., lower power for a target delay specification or improved speed for a target power budget. From Figure 34, we observe that despite optimal sizing at one set of voltages, our optimization techniques offer up to 1.45X reduction in energy/operation. For this range of voltages, up to 1.4X improvement in maximum operable speed is obtained. Further energy/operation improvements are achievable for a given clock frequency through *a priori* determination of the required operating voltages and then performing optimal sizing at those voltage swings.

FIGURE 34 Effect of optimization techniques on QuadRail power-delay characteristics.



5.3 Summary

In this chapter, we explored the design space of Mixed Swing QuadRail and outlined optimization strategies for minimizing QuadRail circuit energy/operation and hence maximizing the potential energy/operation savings against static CMOS. Analytical posynomial power and delay formulations were derived for QuadRail from the n^{th}-Power Law submicron MOSFET model I-V equations, that enabled studying the power-delay trade-offs in current and future fabrication processes. The accuracy of these models was demonstrated through comparisons with HSPICE simulations using Level13, BSIM1 models. The impact of QuadRail's degrees of freedom on the power, delay, power*delay product, and energy*delay product space of mixed swing circuits were investigated and optimal voltage scaling and buffer transistor sizing approaches to minimize QuadRail circuit energy/operation were developed. Their effectiveness was demonstrated on a 16*16+36-bit MAC circuit fabricated in a commercial 3V, 0.5µm bulk-CMOS process.

6 Mixed Swing Circuits: Low-Voltage Challenges

With feature sizes scaling well into the deep-submicron era, *manufacturability* of digital circuits has become an increasingly important design concern. This trend is particularly due to fluctuations in device and process parameters caused by inevitable disturbances in the fabrication process and variations in operating temperature. These fluctuations either result in the manufactured circuit not successfully performing the desired function (characterized as *functional* yield loss) or not meeting the target performance specifications (e.g., clock frequency, power dissipation) across worst-case process and temperature corners (characterized as *parametric* yield loss). With scaling feature sizes these fluctuations either remain non-scalable or worsen, making it an increasingly formidable research challenge to minimize the associated yield losses [Maly96], [Strojwas96].

Device and process parameter variations have raised yet another increasingly important design concern in deep-submicron processes: *noise immunity.* Degradation of digital circuit noise margins across worst-case process and temperature corners have contributed to a significant noise immunity loss with scaling feature sizes. The non-scalability or worsening of these fluctuations with process scaling has made designing for signal integrity an increasingly formidable research challenge as well [Shepard96].

Both manufacturability and noise immunity worsen with voltage scaling, due to the increased dispersion in circuit operating frequency, power dissipation, and noise margins across worst-case process and temperature corners at reduced voltages. This makes design for manufacturability and noise immunity all the more important in low-voltage deep-submicron circuits [Kakumu90], [Yan95], [Strojwas96].

In this chapter, we examine these two low-voltages challenges to study the practicality of mixed swing methodologies. Worst-case analysis is performed on a 16*16+36-bit MAC implemented in our prototype architecture, to study the manufacturability and noise immunity of Mixed Swing QuadRail relative to static CMOS in a 0.5µm bulk-CMOS process. For improved low-voltage manufacturability of QuadRail circuits in future deep-submicron processes, a series regulation technique is developed for local on-chip generation of QuadRail's low-swing power rails. This approach electronically offsets threshold voltage variations across the worst-case process/temperature corners. The series regulated approach, in essence, makes Mixed Swing QuadRail a self-contained methodology which can replace full-swing static CMOS operating between a regular, high-swing supply without warranting any fabrication process or system-level modifications.

6.1 Mixed Swing QuadRail Manufacturability

Of prime importance amongst all device and process parameter fluctuations are intra-die and inter-die MOSFET threshold voltage variations, since these worsen at least linearly with scaling feature sizes, becoming comparable to the threshold voltages themselves [Eisele95], [Yan95], [Strojwas96], [Tang96]. The increasing threshold variations results in substantial circuit delay and power dispersion across worst-case process and temperature corners [Sun94], [Davari96], [Frank97], only to be aggravated with voltage scaling due to the increased variations in transistor on-drive currents. The delay and power dispersions, therefore, contribute significantly to parametric yield degradation, particularly at low operating voltages, and more so with process scaling.

In this section, we quantify the power and delay dispersion for both static CMOS and Mixed Swing QuadRail across worst-case - Slow-NMOS-Slow-PMOS (SNSP) and Fast-NMOS-Fast-PMOS (FNFP) - process and temperature corners in a 0.5µm process. Table 1 shows the process and temperature corners for this process (Note that the FNFP and SNSP corners represent the worst-case power and delay scenarios respectively). The worst-case power/delay corners are formulated on the basis of FNFP and SNSP corner parameter variations data provided by PDF Solutions, Inc. [Michaels96]. Figure 36 shows the threshold voltage, transconductance gain factor, and saturation region on-drive current variations data, emphasizing the substantial parameter fluctuations. The static CMOS vs. QuadRail worst-case analysis is performed on the Wallace tree multiplier of a 16*16+36-bit MAC in the 0.5µm process, over a

FIGURE 35 NMOS vs. PMOS V_t, β, and I_{DS} variations.

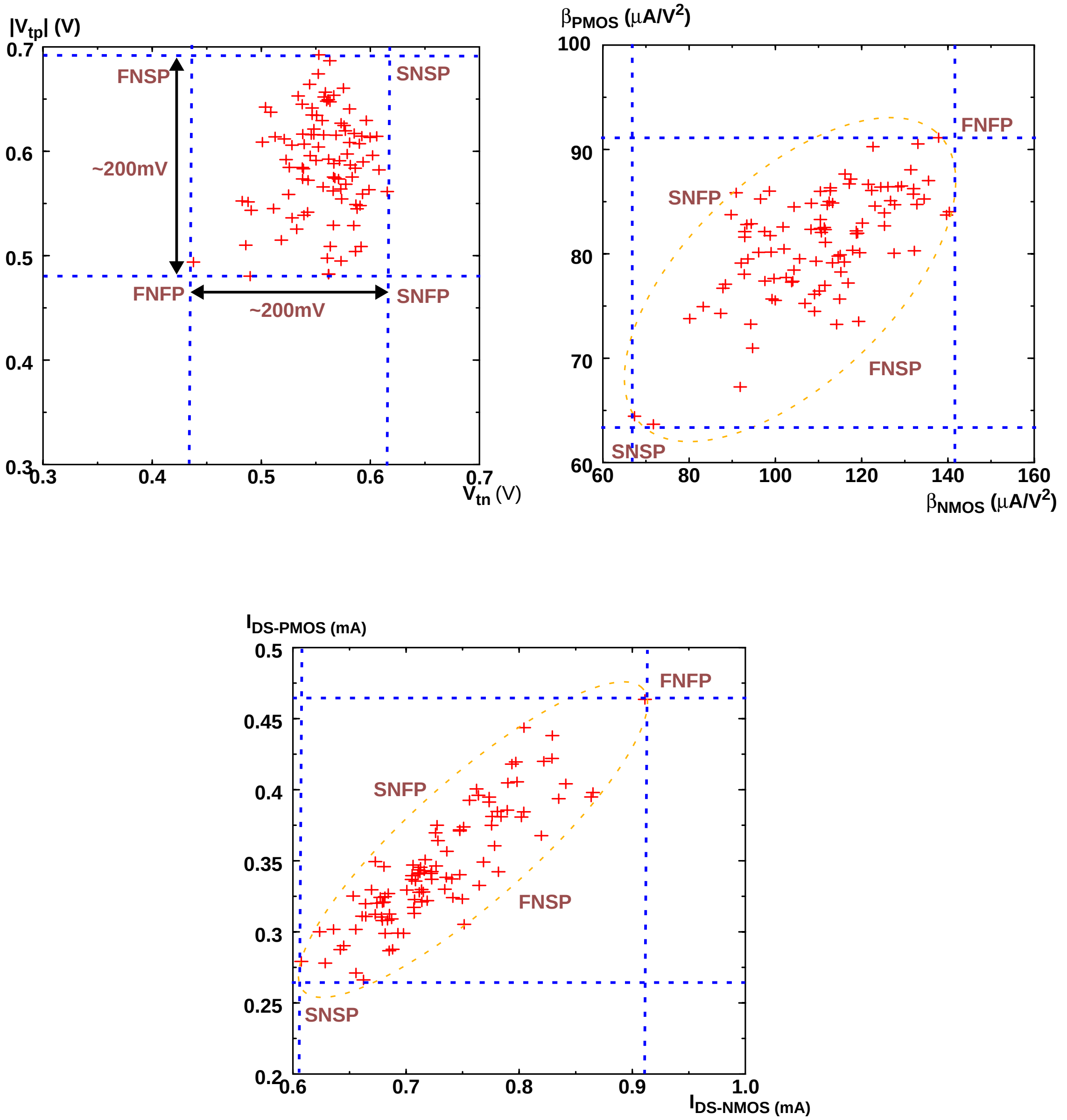
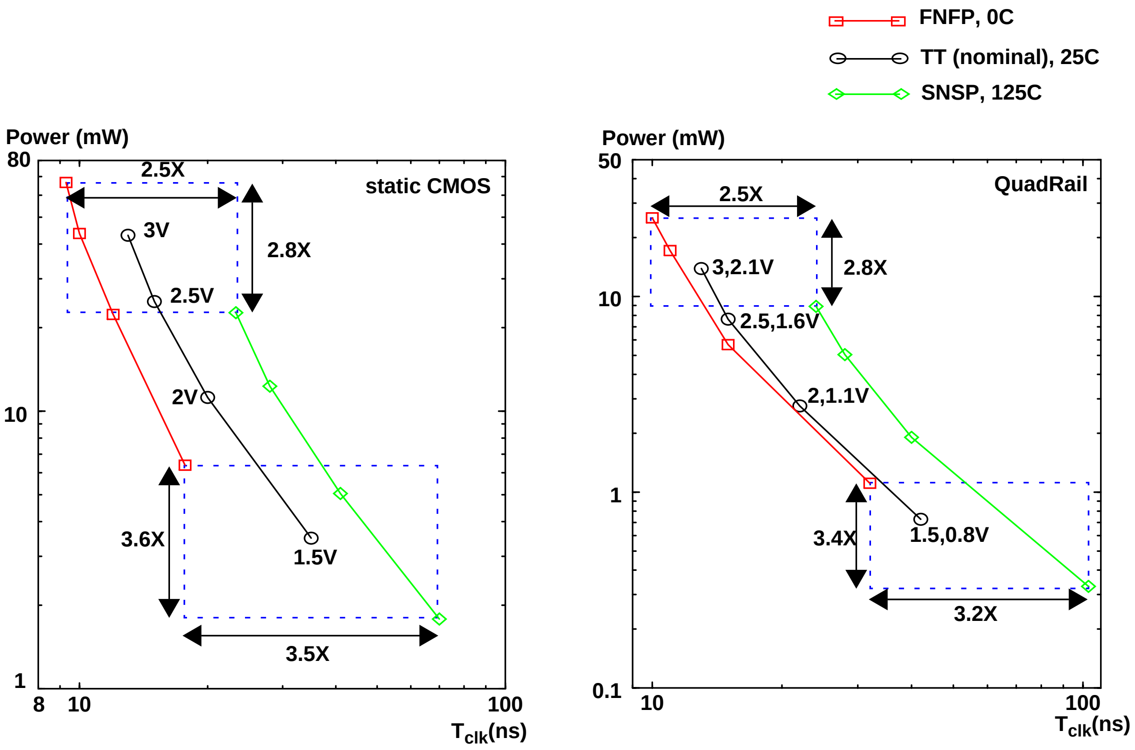


TABLE 1. Nominal and worst-case process and temperature corners in the 0.5µm CMOS process.

parameter	nominal	FNFP	SNSP
temperature (˙C)	25	0	125
T_{ox} (Å)	96	91	101
ΔL (µm)	0	-0.04	+0.04
ΔW (µm)	0	+0.06	-0.06
nMOS-V_t (V)	+0.70	+0.60	+0.80
pMOS-V_t (V)	-0.90	-0.80	-1.00

range of voltages. Implementation details will be described in Chapter 7. Figure 36 show the static CMOS and QuadRail power-delay dispersion obtained through HSPICE simulations using Level13, BSIM1 models across

FIGURE 36 Static CMOS vs. QuadRail worst-case analysis in 0.5µm process.



500 pseudo-random vectors. Both CMOS and QuadRail demonstrate similar delay and power dispersions at high voltage swings. However, at reduced swings, dispersions are slightly lower for QuadRail: at V_{logic}=1.5V, V_{buffer}=0.8V, we observe a power*delay dispersion of 10.88X for QuadRail as opposed to 12.6X for CMOS (corresponding V_{dd}=1.5V), i.e., about 1.2X better. This is primarily due to the reduced load voltage swings of QuadRail gates, causing the power and delay sensitivities to process and temperature corners to decrease approximately linearly with decreasing ratios of buffer to logic voltage swings. Thus, the Mixed Swing QuadRail approach demonstrates a modestly better low-voltage parametric yield than static CMOS. However, further containment of the delay and power dispersions will be essential in future low-voltage deep-submicron processes, because of the increasing threshold variations. This motivates the necessity for electronically offsetting the threshold variations in QuadRail, resulting in the development of an on-chip series regulated QuadRail methodology.

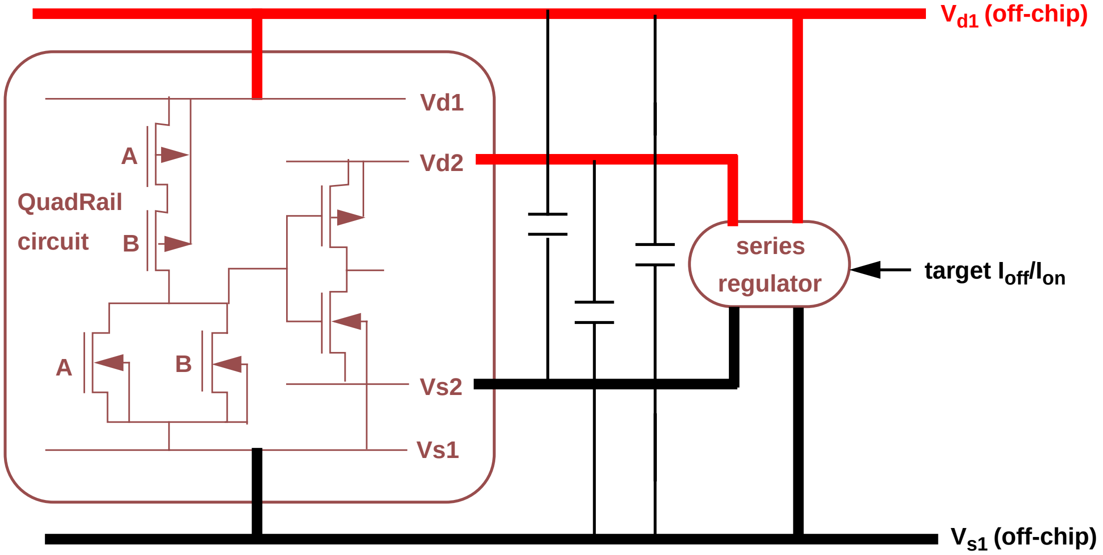
6.2 Series Regulated QuadRail Methodology

The Mixed Swing QuadRail methodology as described earlier employs explicit off-chip power supplies for the logic and buffer stages, which source their respective load capacitance charging/discharging currents. This approach offers a nearly quadratic reduction in buffer stage dynamic power since there exists no DC path between the high and low voltage supplies. However, this methodology has three limitations:

- Employing additional explicit off-chip supplies and its associated pin and pad requirements add to the total system cost and hence economically unattractive.
- When the buffer voltages are scaled well below the sum of the threshold voltages of NMOS and PMOS devices, the low-voltage off-chip supply is prone to significant inefficiencies, particularly if the drive-current requirements are high (e.g., if the buffer voltage supply delivers the drive-currents of many on-chip QuadRail circuits). This degrades overall system power efficiency.
- Due to the lack of any on-chip regulation (the separation between the supplies remains fixed), this methodology suffers from significant dispersions in delay and power at reduced operating voltages across worst-case process and temperature corners, contributing significantly to parametric yield degradation. Although the dispersions are modestly better than static CMOS, this is still a cause for concern in future deep-submicron processes.

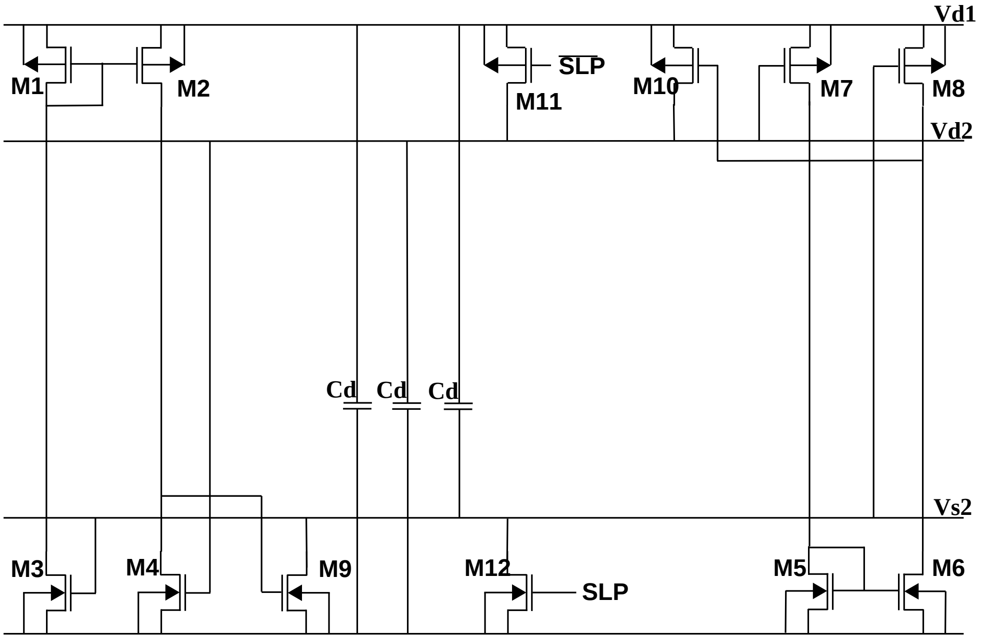
In order to overcome these limitations, an on-chip series regulation approach is developed for locally generating the buffer stage low-voltage supply for Mixed Swing QuadRail. Figure 37 shows the series regulated QuadRail methodology. Figure 38 shows the series regulator circuit. For a given high-swing voltage (Vd1-Vs1), the low-swing rail voltages (Vd2 and Vs2) are servoed to maintain a fixed ratio of off- to average on-drive current (I_{off}/I_{on}) within the QuadRail circuit, essentially implementing the optimal voltage scaling approach described in Chapter 5. The transistor pairs (M3:M4) and (M7:M8) are ratioed Nx:1x, where 1x is the minimum-width transistor and N is the desired I_{on}/I_{off} for the QuadRail circuit. By selecting the I_{off}/I_{on} ratio to be the ratio of switching activity to average gate depth of the QuadRail circuit, static and dynamic power are approximately balanced, minimizing the total circuit energy/operation. This maximizes the energy/operation savings compared to an equivalent static CMOS implementation operating between Vd1-Vs1 for a target clock frequency constraint. Further, this achieves the same goal of minimizing total energy/operation as the technological speed compensation solutions to voltage scaling [Liu93], [Burr94], [Gu96], [Frank97] described in Chapter 2, but without mandating any process recipe modifications. The current mirror devices (M1:M2) and (M5:M6) are ratioed 1:1. M9 and M10 provide the DC series path between the power rails and are sized to be able to source/sink the peak on-drive current requirement of the QuadRail

FIGURE 37 Series Regulated Mixed Swing QuadRail methodology.



circuit. All devices within the QuadRail circuit and the series regulator are oriented identically to minimize threshold voltage mismatches between them. This is critical because threshold mismatches between regulating and regulated circuits prevents effective electronic offset of threshold variations in any regulated circuit. Local inter-rail decoupling capacitors (C_d) are inserted to reduce rippling on the low-swing power rails due to simultaneous switching noise on the high- and low-swing power rails. M11 and M12 are sleep-mode enable devices that are disabled (SLP=Vs1) during normal operation. During power-down mode (SLP=Vd1), the low-swing power rails are shorted to the high-swing power rails, eliminating the DC path power consumption that exists during normal operating mode. This reduces QuadRail's sleep-mode power to that of full-swing static CMOS leakage power. Conventional static CMOS leakage

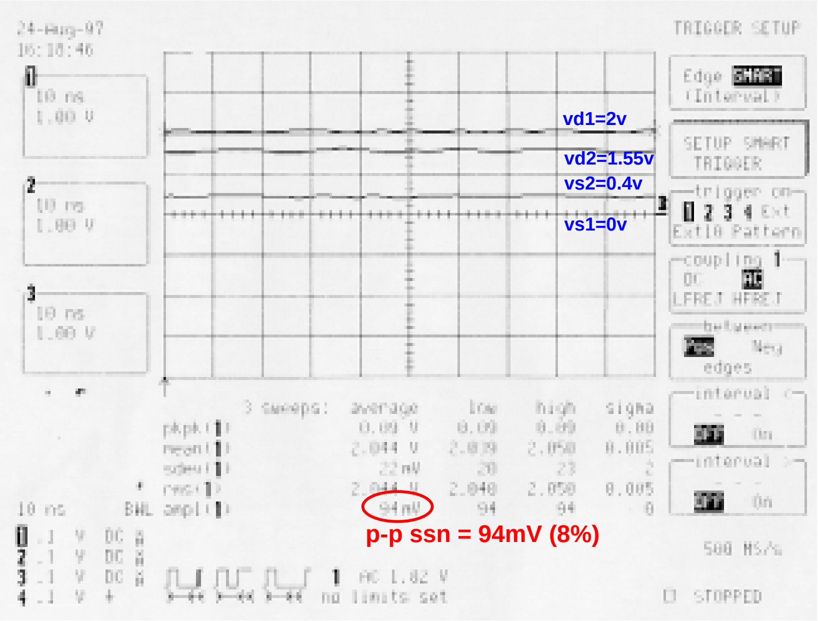
FIGURE 38 I_{off}/I_{on} ratio based Series Regulator circuit.



power reduction techniques can be adopted to further lower this sleep-mode power [Kuroda96], [Shigematsu97]. In order to demonstrate the series regulated QuadRail operation, Figure 39 shows sample waveforms from the off-chip high-voltage power rails and the on-chip series regulated low-voltage power rails, measured on the same 16*16+36-bit MAC fabricated in series regulated QuadRail in the 0.5μm process. Inter-rail MOS decoupling capacitors, 4pF each, are inserted to control the peak-peak simultaneous switching noise on the regulated power rails to within 8% of the rail-to-rail swing. Greater power/ground bounce suppression can be achieved at the cost of layout area through the insertion on larger inter-rail decoupling capacitors.

In order to study its impact on manufacturability relative to static CMOS and the off-chip regulated QuadRail approaches, worst-case process and tem-

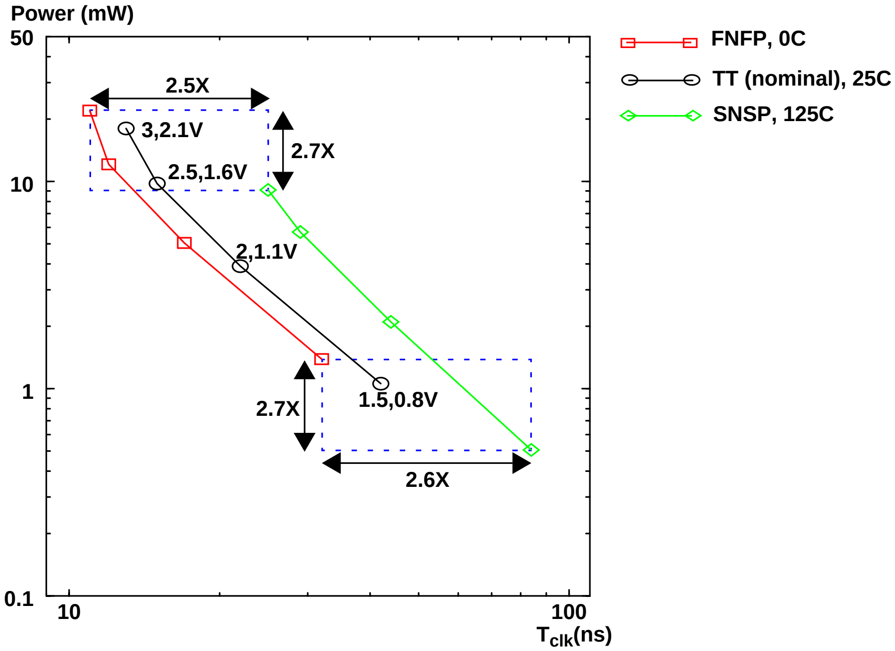
FIGURE 39 0.5μm 16*16+36-bit series regulated QuadRail MAC measured power-rail waveforms.



```
        3 sweeps:  average   low     high    sigma
pkpk(1)            0.09 V    0.09    0.09    0.00
mean(1)            2.044 V   2.039   2.050   0.005
sdev(1)            22 mV     20      23      2
rms(1)             2.044 V   2.040   2.050   0.005
10 ns   BWL ampl(1) 94 mV     94      94      0
1 .1 V DC
2 .1 V DC
3 .1 V DC
4 .1 V +
                        1 AC 1.82 V
                        no limits set
```

perature corner analysis is performed on the same Wallace tree multiplier of a 16*16+36-bit MAC in the 0.5µm process, but implemented with series regulation. The same process and temperature corners from Table 1 are employed here. Figure 40 shows the power-delay dispersion obtained through HSPICE simulations using Level13, BSIM1 models over the same 500 pseudo-random vectors. Series regulated QuadRail shows almost the same (1.04X lower) power*delay dispersion across corners compared to static CMOS and off-chip regulated QuadRail at V_{logic}=3V, V_{buffer}=2.1V. With voltage scaling, the dispersion remains well controlled because the series regulator adjusts the low-voltage power rails to effectively offset the threshold variations while maintaining the desired I_{off}/I_{on} ratio across process and temperature corners. At V_{logic}=1.5V, V_{buffer}=0.8V, the power*delay dispersion is 1.8X (1.55X) lower

FIGURE 40 Series Regulated QuadRail worst-case analysis in 0.5µm process.



than static CMOS (off-chip regulated QuadRail), demonstrating significantly improved low-voltage manufacturability.

6.3 Mixed Swing QuadRail Noise Immunity

We next address the other important low-voltage challenge to the practicality of mixed swing methodologies, viz., Noise Immunity. Figure 41 shows the QuadRail logic stage, buffer stage, and combined DC transfer characteristics of a CSA for V_{logic} = 1.5V, V_{buffer} = 0.8V in the 0.5μm process. Despite static current in the logic stage, the transfer characteristics are observed to be sharp, with fully restored outputs, due to multiple stages of gain. High and Low noise margins are almost equal and are approximately half of the buffer voltage swing (V_{buffer}/2). Therefore, the lower bound on V_{buffer} is set by the minimum permissible noise margin constraints [Kakumu90].

Although QuadRail's absolute noise margins are lower than that of an equivalent static CMOS gate operating at V_{logic} (which are approximately V_{logic}/2), primary sources of intrinsic digital circuit noise are also lower. In order to compare their relative noise immunity, a worst-case analysis is performed on a static CMOS and QuadRail CSA from within the Wallace tree multiplier of the 0.5μm 16*16+36-bit MAC. The goal is to study noise margin degradation of the static CMOS and QuadRail CSAs across Fast-NMOS-Slow-PMOS (FNSP) and Slow-NMOS-Fast-PMOS (SNFP) process and temperature

FIGURE 41 QuadRail logic stage, buffer stage, and combined DC transfer characteristics in 0.5µm process.

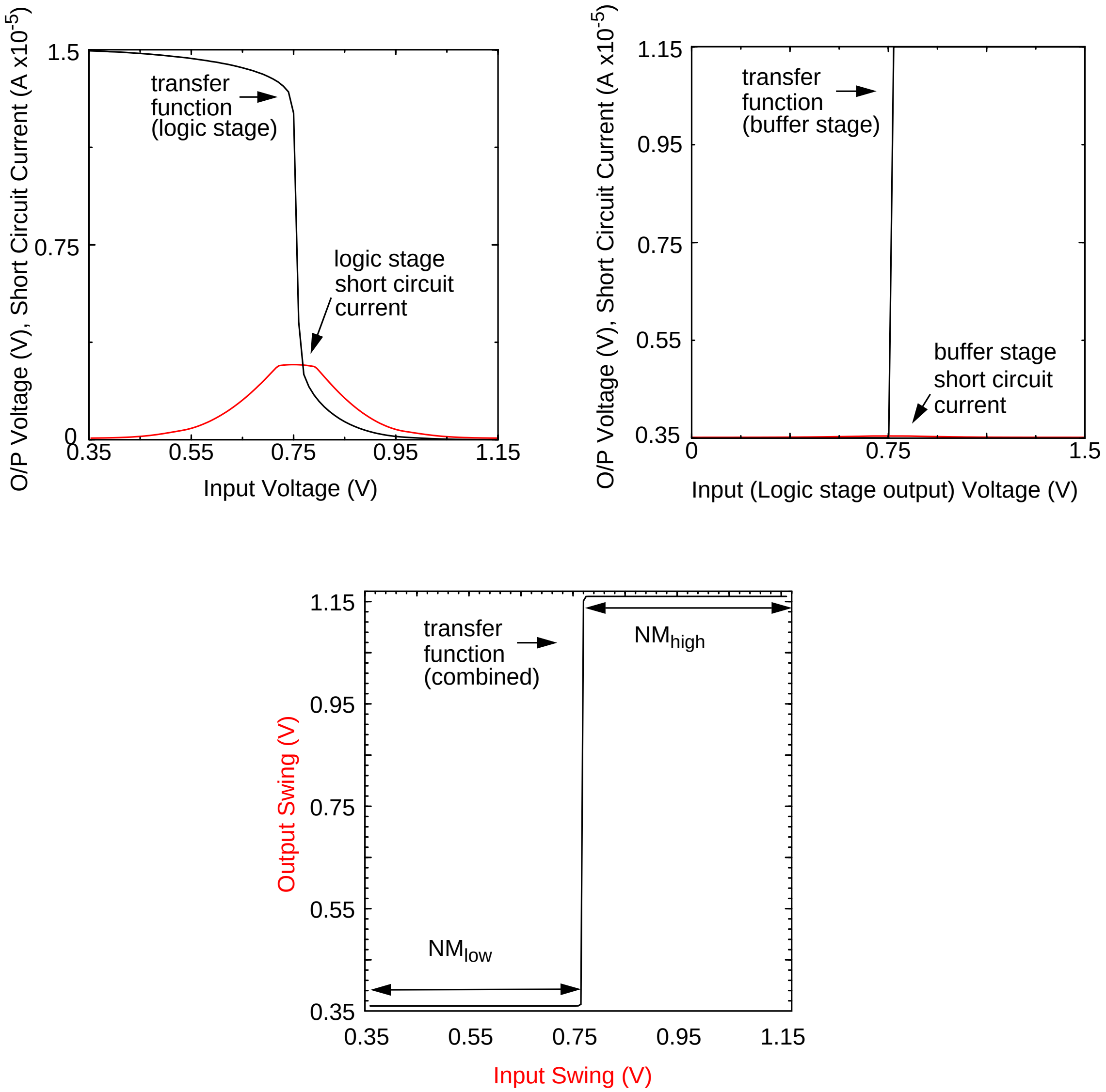


TABLE 2. Nominal and worst-case process and temperature noise corners in the 0.5µm CMOS process.

parameter	Typical	FNSP	SNFP
temperature (°C)	25	125	125
Tox (Å)	96	96	96
ΔL (µm)	0	0	0
ΔW (µm)	0	0	0
NMOS-Vt (V)	+0.70	+0.60	+0.8
PMOS-Vt (V)	-0.90	-1.00	-0.8

corners (Table 2), which represent the worst-case *noise* corners. The noise corners are formulated on the basis of FNSP and SNFP corner parameters data provided by PDF Solutions, Inc. [Michaels96]. Note that FNSP and SNFP corners assume no variations in gate-oxide thickness (T_{ox}), channel length (ΔL), and channel width (ΔW): this is due to the strong correlation between NMOS and PMOS devices in these parameter variations that precludes them from varying in opposite directions [Maly90]. On the other hand, variations in NMOS and PMOS threshold voltages do not exhibit a strong correlation and hence vary in opposite directions. Figure 35 explains this trend: NMOS and PMOS threshold voltage variations display a weak correlation, whereas their transconductance gain factors display a strong correlation due to several common process parameters affecting both [Bakoglu90]. The overall impact of the variations is a *strongly correlated* variation in the NMOS and PMOS saturation region on-drive currents.

Figure 42 shows the High and Low noise margin dispersions across the worst-case corners for the static CMOS and QuadRail CSAs, superimposed on their respective nominal DC transfer characteristics. The analysis is conducted at static CMOS V_{dd} = 1.5V and QuadRail V_{logic} = 1.5V, V_{buffer} = 0.8V in the 0.5µm process. The worst-case degradation in High and Low noise margins for static CMOS is observed to be 90mV and 95mV respectively, while the nominal noise margins are approximately 750mV. For QuadRail the corresponding High and Low noise margin degradations are 101mV and 103mV respectively, while the nominal noise margins are approximately 400mV.

It is observed that the fraction of nominal noise margins lost across worst-case corners is significantly higher for QuadRail. However, absolute *noise margins* across worst-case corners are not indicative of *noise immunity*, since primary sources of intrinsic digital circuit noise scale atleast linearly with reduced operating voltages [Bakoglu85], [Bakoglu90], [Shepard96]. In order to perform a realistic worst-case noise immunity analysis, we consider realistic worst-case intrinsic sources of noise within the static CMOS and QuadRail

FIGURE 42 Static CMOS vs. QuadRail noise margin dispersions across worst-case corners.

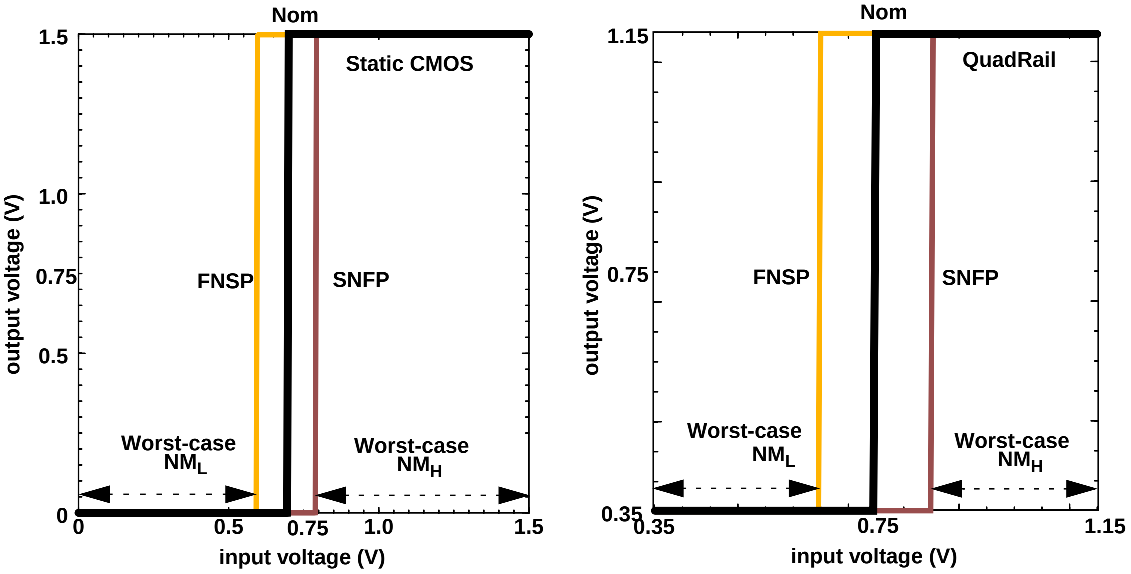


TABLE 3. Worst-case noise data for the 16*16+36-bit static CMOS and QuadRail MACs.

noise source	static CMOS V_{dd} = 1.5V	QuadRail V_{logic} = 1.5V, V_{buffer} = 0.8V
Power/ground bounce (across 500 pseudo-random vectors)	±60 mV	±31 mV
Signal crosstalk (1mm, minimum-spaced, metal2 interconnects switching antiphase)	±50 mV	±14 mV
Substrate coupling (±250 mV injected $V_{source\text{-}bulk}$)	±61 mV	±61 mV

16*16+36-bit MACs for the same operating conditions. The three primary sources of noise considered are (i) power/ground bounce, (ii) signal crosstalk, and (iii) substrate coupling [Bakoglu90]. On the basis of experimental measurements on the 16*16+36-bit MAC fabricated in the 0.5μm process and commercial low-voltage noise data [Stanisic97], [Nicol97], worst-case noise within the static CMOS and QuadRail MACs are computed, shown in Table 3. Figure 43 shows the leftover worst-case noise margins after allocating these noise values superimposed on the worst-case DC transfer characteristics from Figure 42. We define leftover worst-case noise margins as:

$$NM_{leftover} = \begin{bmatrix} NM_{nominal} - \\ max(NM_{L-deg}, NM_{H-deg}) + \\ Noise_{worst-case} \end{bmatrix} \qquad \text{(EQ 21)}$$

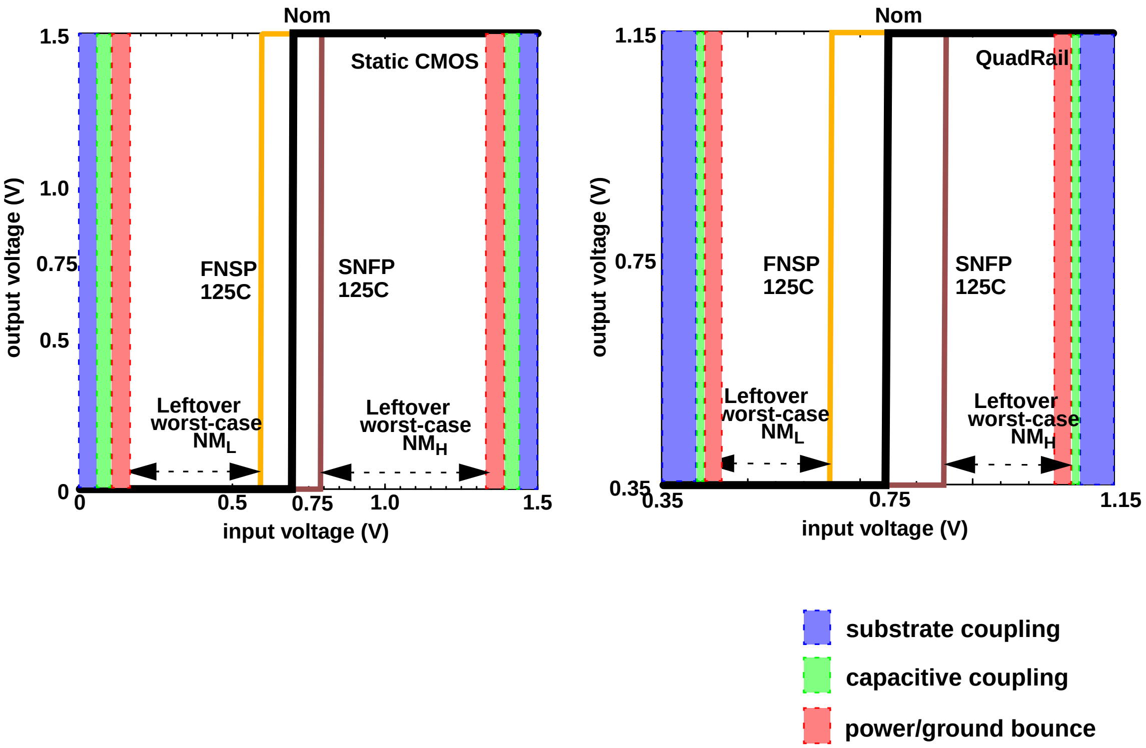
where, NM_{L-deg} and NM_{H-deg} are the worst-case Low and High noise margin degradations across corners respectively. Assuming the noise sources to be mutually exclusive and cumulative (catastrophic noise scenario), the leftover noise margins for static CMOS and QuadRail are computed as:

$$\begin{aligned} NM-CMOS_{leftover} &= [750-(95+171)]\ mV \\ &= 484\ mV \end{aligned} \quad \text{(EQ 22)}$$

$$\begin{aligned} NM-QuadRail_{leftover} &= [400-(103+105)]\ mV \\ &= 192\ mV \end{aligned} \quad \text{(EQ 23)}$$

From Equation 22 and Equation 23, it is observed that the leftover worst-case noise margins for static CMOS and QuadRail CSAs across process, temperature, and noise corners is 484mV and 192mV respectively. This corresponds to nearly 64% of the nominal noise margins for static CMOS and 48%

FIGURE 43 Leftover noise margins for the static CMOS and QuadRail CSAs.



for QuadRail, and is indicative of their respective noise immunity. Since a substantial fraction of the nominal noise margins are still leftover, both methodologies possess adequately high low-voltage noise immunity. However, to compensate for QuadRail's lower leftover noise margins, noise sources need to be controlled more tightly than in the equivalent static CMOS MAC: additional on-chip despiking capacitors, more effective shielding between the QuadRail and peripheral static CMOS circuits through extensive guard-banding and employing 'noise-aware' CAD tools that can assess noise-prone regions within QuadRail circuits and design to meet target worst-case noise margins [Bakoglu90], [Stanisic93], [Su93], [Shepard96]. This is required to enable further reliable voltage scaling of QuadRail circuits while maintaining the same fraction of leftover noise margins as static CMOS circuits.

6.4 Summary

Deep-submicron low-voltage practicality challenges to Mixed Swing QuadRail, specifically manufacturability and noise immunity, were investigated in this chapter. A worst-case analysis is performed on QuadRail as well as static CMOS to study their relative manufacturability and noise immunity in the 0.5µm process. The worst-case power, delay, and noise process and temperature corners were formulated on the basis of industrial parameter variations data.

A modestly lower power*delay dispersion is demonstrated for QuadRail over static CMOS at low voltages. However, further containment of the delay and power dispersions in future processes was noted to be essential for improved low-voltage manufacturability. Therefore, an on-chip series regulation approach with sleep-mode control was developed for Mixed Swing QuadRail for locally generating the buffer stage low-voltage supply. This technique electronically offsets threshold voltage variations by adjusting the low-voltage power rails, while maintaining a target I_{off}/I_{on} ratio across the worst-case corners. Up to a 1.8X better low-voltage manufacturability was achieved relative to static CMOS. Further, since the series regulated approach eliminates the necessity for an additional off-chip power supply, Mixed Swing QuadRail is transformed into a self-contained methodology which can replace full-swing static CMOS operating between a regular, high-voltage supply without warranting any technology or system-level modifications. Through the insertion of inter-rail on-chip decoupling capacitors within the series regulator, peak-peak power/ground bounce on the regulated low-voltage rails for a 16*16+36-bit MAC was measured to be within 8% of the rail-to-rail swing. In the next chapter, we will examine the energy/operation savings that the QuadRail methodology, both with and without series regulation, can achieve over static CMOS datapath circuits.

Worst-case high and low noise margin dispersions across the worst-case corners were also studied for QuadRail and static CMOS CSAs from within the Wallace tree multiplier of the 16*16+36-bit MAC. Through the superimposition of worst-case noise values on their worst-case DC transfer characteristics, both the methodologies were observed to display adequately high low-voltage noise immunity.

7 Mixed Swing Circuits: Performance Analysis

In this chapter, we perform power-delay comparisons between mixed swing and static CMOS methodologies to examine the achievable energy/operation savings on datapath circuits. We begin the comparisons on our prototype signed, fixed-point, DSP MAC architecture. By exploiting the increasing final-adder-to-multiplier delay slack with voltage scaling in current and future submicron processes, examined in Chapter 3, we demonstrate the ability to achieve substantial reduction in the energy/operation of Wallace tree multipliers over a range of operand bit-widths, process generations, and operating voltages.

The comparisons are next extended to single-layer point-to-point data buses and multi-layer multicast datapath nets within the floating-point units (FPUs) of two industrial next-generation microprocessors with extensive multimedia support, presently in design in a next-generation 0.16μm bulk-CMOS process. The ability to voltage scale more effectively than static CMOS over a wide range of input data switching activities for a target clock frequency is demonstrated.

7.1 DSP MAC Comparisons

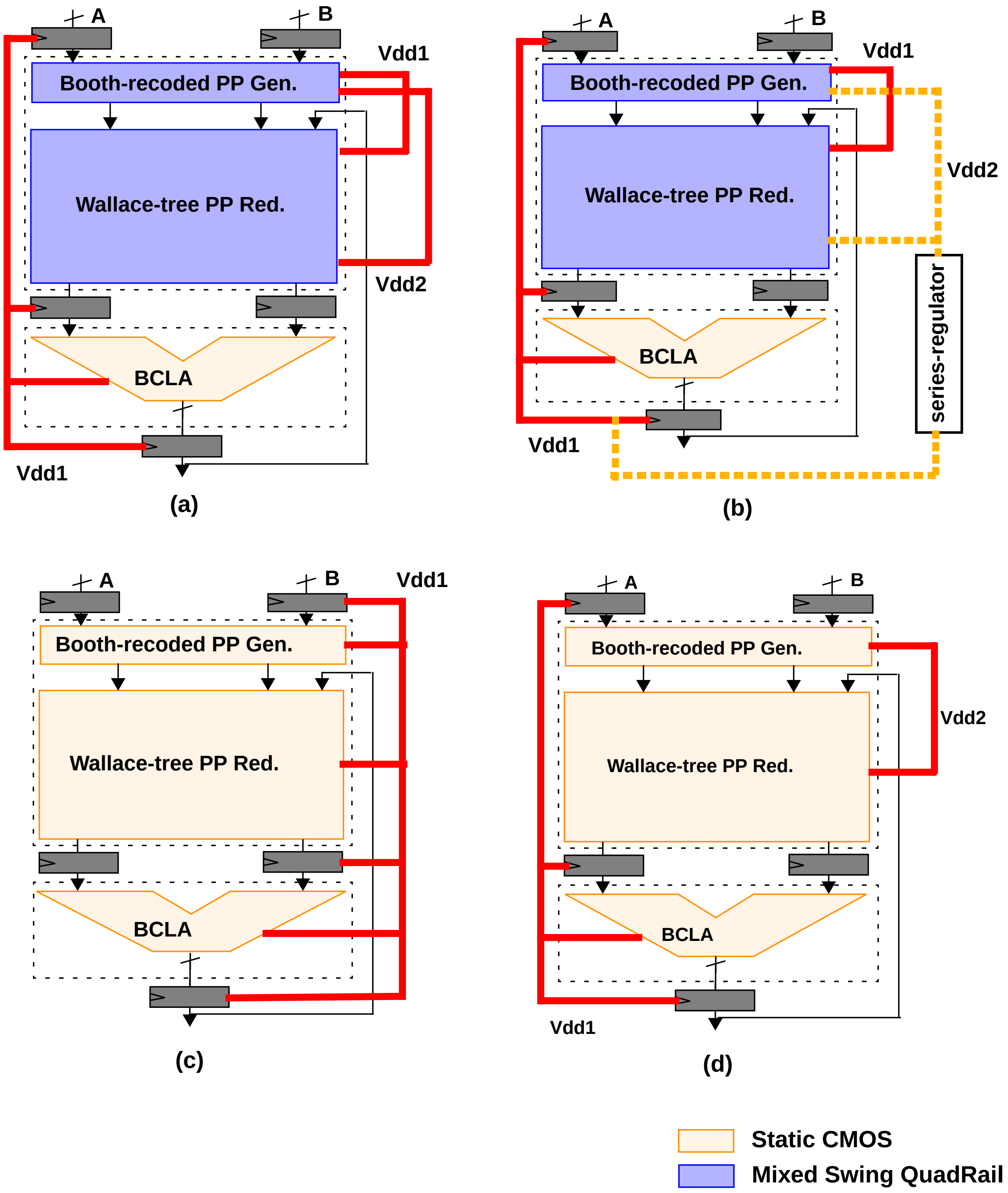
In this section, we perform power-delay space comparisons between the static CMOS and Mixed Swing QuadRail (both off-chip regulated and series regulated) MACs. The comparisons are performed over a range of (i) MAC bitwidths (8-24 bits), (ii) CMOS and fully-depleted SOI process generations (0.5μm - 0.16μm), and (iii) process-defined operating voltages.

As observed in Chapter 3, the increasing dominance of interconnect capacitance over gate capacitance with process scaling makes the Wallace tree multiplier power a more and more dominant component of total power within our prototype MAC architecture. Further, the final-adder-to-multiplier delay slack increases with voltage scaling in current and future submicron processes, as observed in Figure 15. This offers an opportunity to lower the multiplier power consumption while retaining target throughput by exploiting this delay slack. We exploit this by retaining the time-critical final adder as a fully static CMOS implementation. The power-critical Wallace tree multiplier is implemented in the off-chip regulated and series regulated Mixed Swing QuadRail approaches. The QuadRail MAC implementations, shown in Figure 45(a) and Figure 45(b), are compared against:

- a single-supply static CMOS MAC implementation operating on a single voltage, as shown in Figure 45(c).
- architecture-driven voltage scaling-based dual-supply static CMOS where the multiplier and final adder have separate power supplies to exploit the delay slack between the multiplier and final adder, as shown in Figure 45(d). We globally scale the final adder's operating voltage (as

above); the multiplier's power supply is scaled until it's delay equals the final adder's delay.

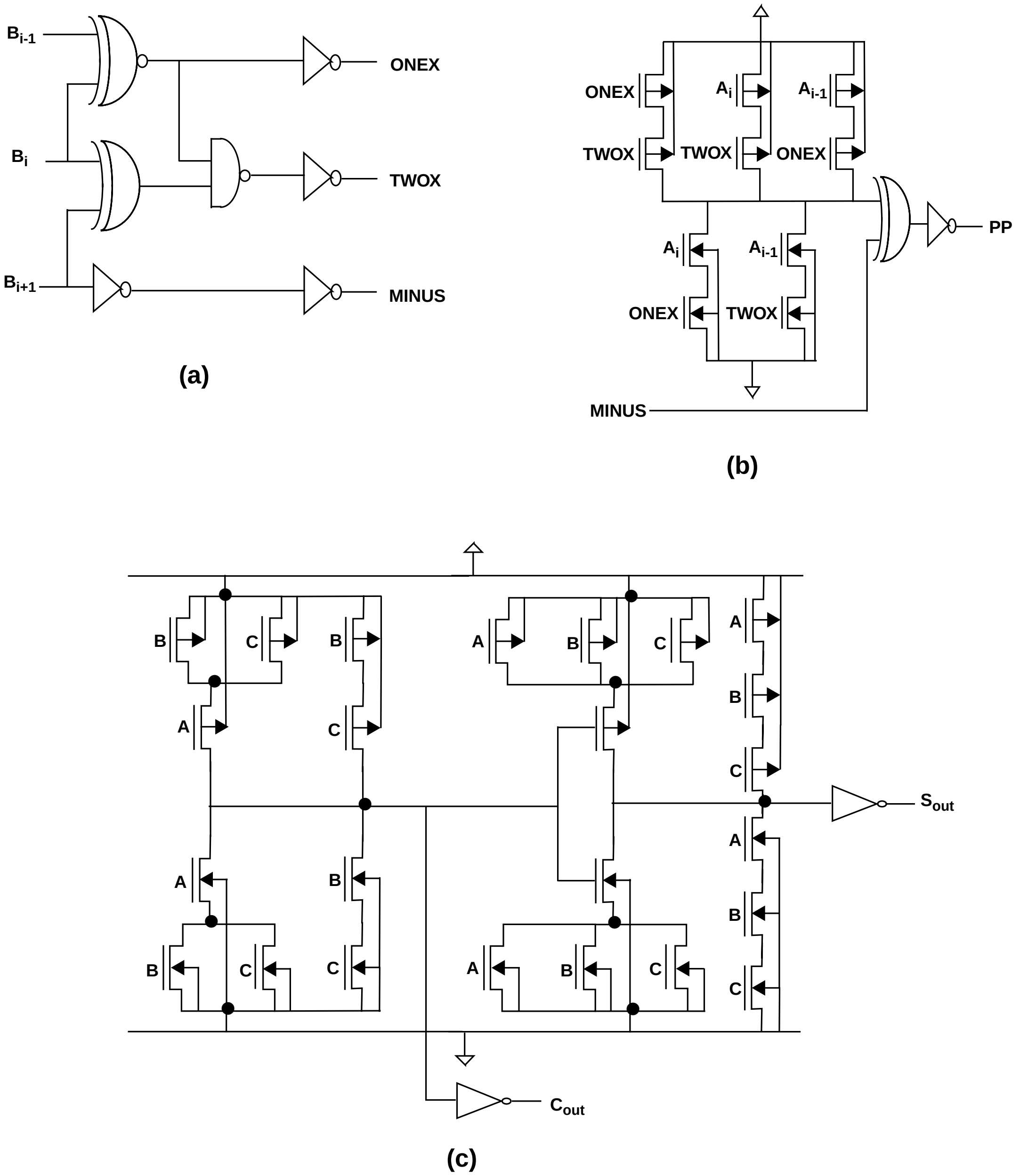
FIGURE 45 Static CMOS vs. Mixed Swing QuadRail power-delay comparison approaches.



7.1.1 Static CMOS and QuadRail MAC Implementation Details

Figure 46 and Figure 47 shows the static CMOS and Mixed Swing QuadRail implementations of the primitive building units of the multiplier: Booth

FIGURE 46 Static CMOS (a) Booth encoder, (b) multiplexor, and (c) CSA implementations.



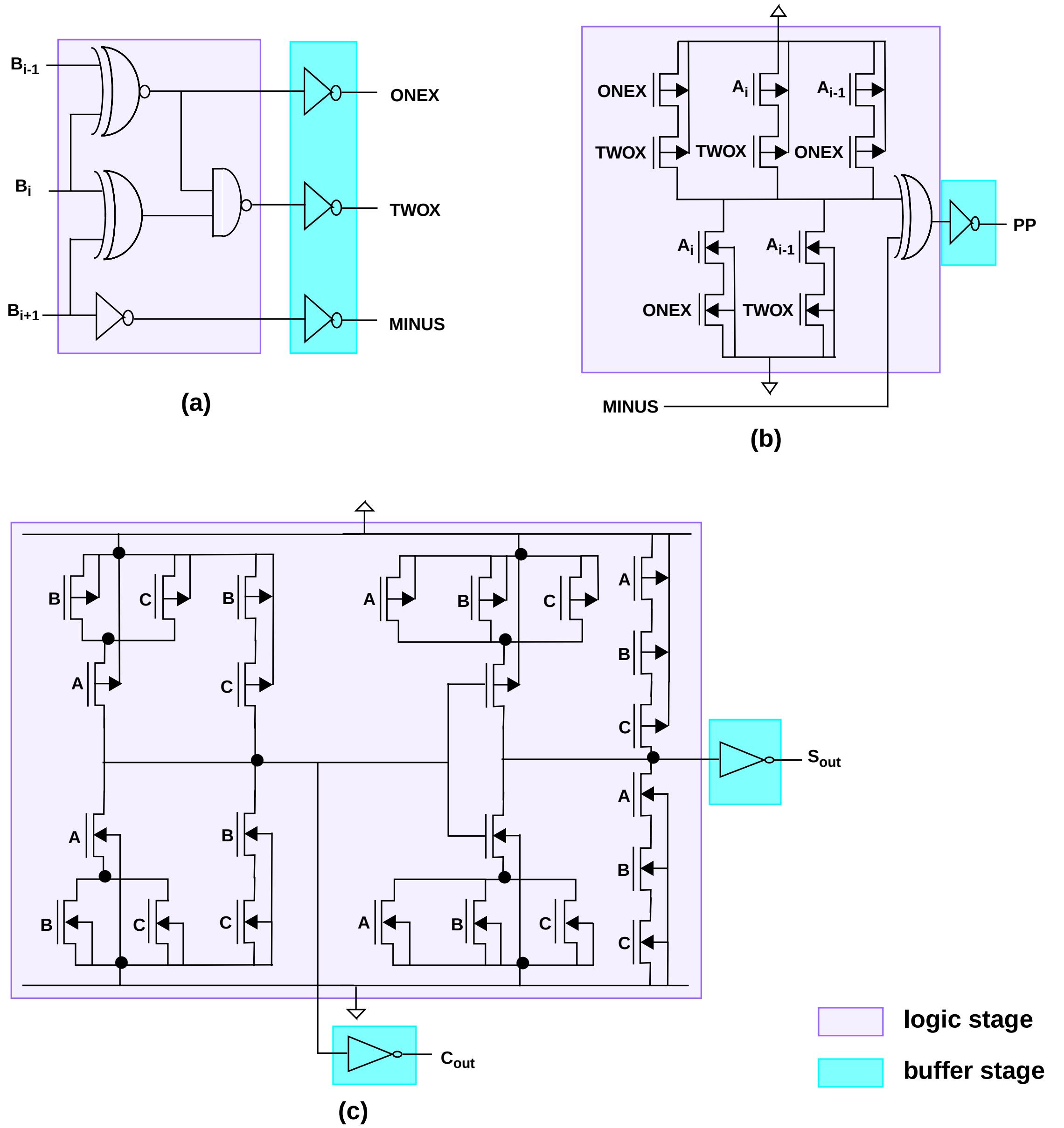
(a)

(b)

(c)

encoder, Booth multiplexor, and CSA. The Booth encoders and multiplexors are implemented such that their outputs are delivered through a static CMOS inverter, which is upsized to drive the desired load capacitance [Cavanagh84],

FIGURE 47 QuadRail (a) Booth encoder, (b) multiplexor, and (c) CSA implementations.



[Larsson96]. The CSA construction is one of the most delay-, power-, and area-efficient static CMOS implementations in literature [Montoye90], [Zimmer97].

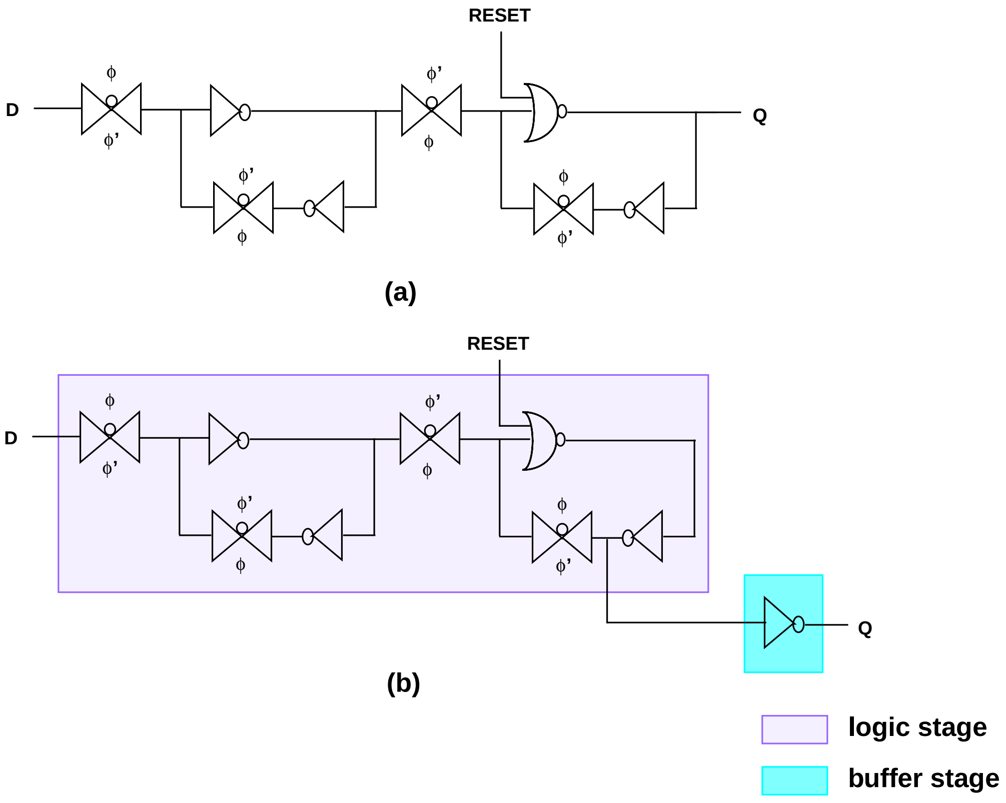
The input, output, and pipeline stage registers in both the static CMOS and QuadRail MACs are identical and operate at the full-swing static CMOS voltage. This is in order to have high-swinging (CMOS level) I/Os to enable interfacing with external static CMOS circuitry without level conversion. The registers are positive edge-triggered and constructed using the transmission-gate-based master-slave D-flip-flop implementation [Bakoglu90], as shown in Figure 48(a). Input registers for the QuadRail MAC alone have a low-swing output inverter as shown in Figure 48(b) in order to feed the QuadRail multiplier with low-swing inputs. Although a QuadRail gate can receive a high-swing input, this poses a signal crosstalk problem due to capacitive coupling between the high-swing input signals and neighboring low-swing routed signals within the multiplier layout. By lowering the signal voltage at the output of the input registers, we minimize the interaction between the high-swing inputs and the interior low-swing signals. Asynchronous RESET provision is introduced in both the static CMOS and QuadRail D-flip-flops, as shown in Figure 48, in order to enable clearing the MAC registers during testing.

The static CMOS MAC is automatically generated and optimally cell-sized by the commercial EPOCH[1] datapath compiler using its custom standard cell library and imported into the Cadence dfII[2] IC design environment. The Quad-

[1] EPOCH is a trademark of Cascade Design Automation, Inc.

Rail MAC layout is produced as follows: Structural Verilog-XL for the multiplier, extracted from the EPOCH datapath compiler, and the static CMOS final adder compiled by EPOCH are floorplanned by Preview, global-placed-and-routed by Block Ensemble, and detail-placed-and-routed by Cell3 Ensemble, all inside the Cadence dfII environment. A custom QuadRail standard cell library of the primitive building units for the multiplier (Figure 47 and Figure 48(b)) is constructed for this purpose. Each standard cell is created with a range of buffer transistor sizes (and hence drive strengths), but within a fixed

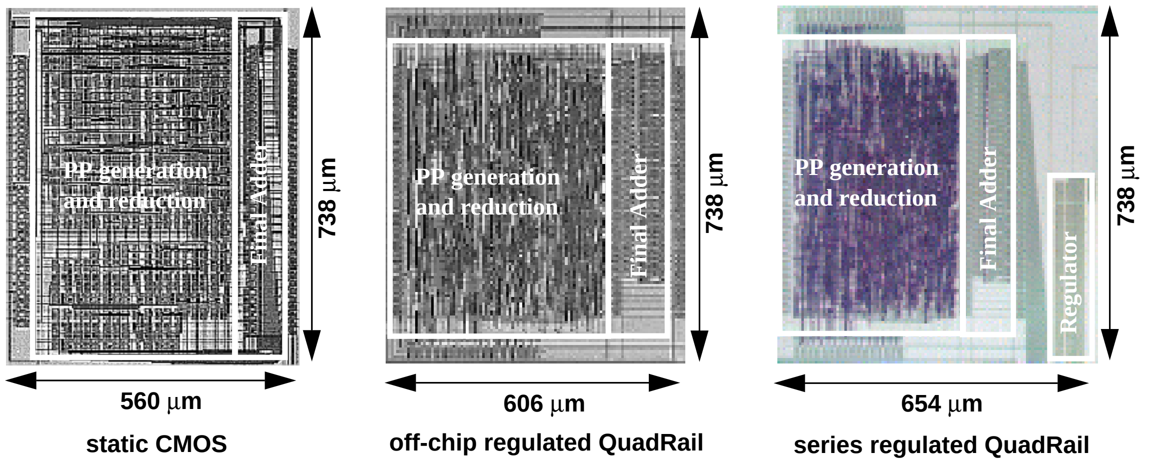
FIGURE 48 (a) Static CMOS/QuadRail and (b) QuadRail Input master-slave D-flip-flops.



[2] dfII, Verilog-XL, Preview, Block Ensemble, and Cell3 Ensemble are trademarks of Cadence Design Systems, Inc.

cell footprint in order to enable post-layout cell-sizing optimization without requiring any layout modifications. On the basis of post-layout study of the interconnect capacitance distribution within the QuadRail multiplier, extracted by Diva, buffer transistors of the Booth encoders, multiplexors, and CSAs are optimally cell-sized to minimize energy/operation, employing the approach described in Chapter 5. Both the static CMOS and QuadRail MACs are 100% over-the-cell routed, i.e., there exists no explicit routing channels in their layouts. Since the QuadRail cells internally perform high-swing logic, 100% over-the-cell routing involves careful layout considerations to minimize signal crosstalk due to coupling between the interior high-swing logic and the low-swing routed signals. Global metal1 obstruction regions over the high-swing logic regions within each QuadRail cell are defined that prevents Cell3 Ensemble from routing the low-swing signals over them, minimizing this coupling although not fully eliminating it. Figure 49 shows the 16*16+36-bit static CMOS and Mixed Swing QuadRail (off-chip regulated and series regulated)

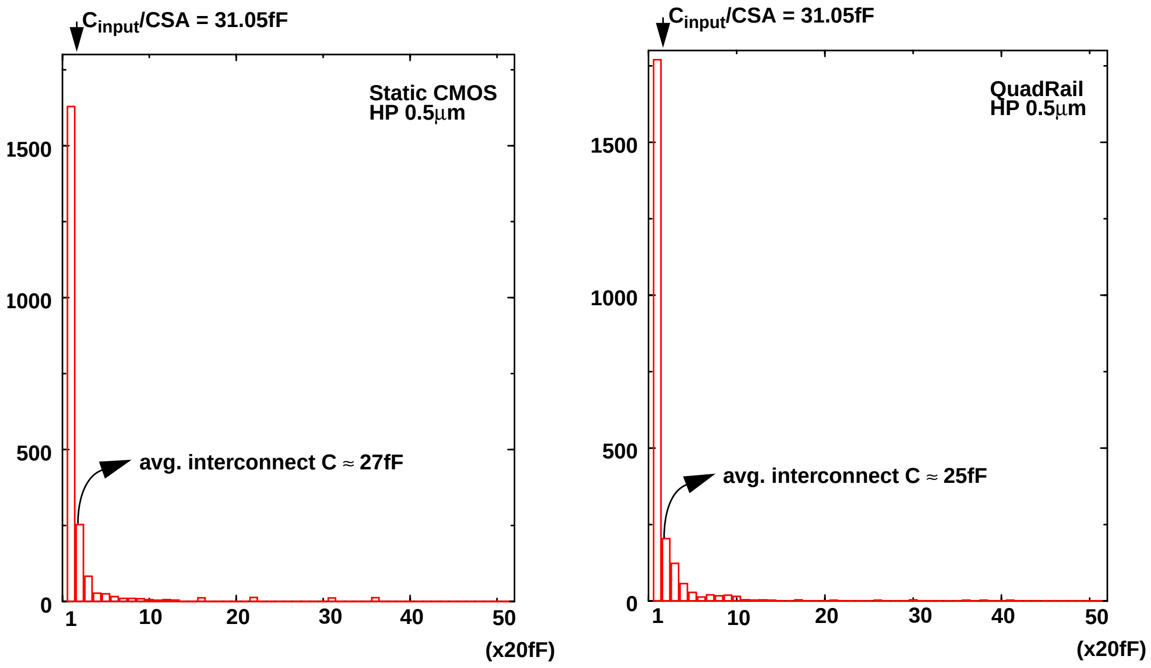
FIGURE 49 0.5µm 16*16+36-bit static CMOS and Mixed Swing QuadRail MAC layouts.


PP generation and reduction
Final Adder
738 µm
560 µm
static CMOS
PP generation and reduction
Final Adder
738 µm
606 µm
off-chip regulated QuadRail
PP generation and reduction
Final Adder
Regulator
738 µm
654 µm
series regulated QuadRail

MAC layouts in the 0.5µm process. Figure 50 shows the interconnect capacitance distributions within the static CMOS and QuadRail MACs extracted by Diva. The distributions are similar, ensuring that conclusions drawn from power-delay space comparisons between the two approaches are fair.

The Mixed Swing QuadRail MAC occupies approximately 10% higher layout area than static CMOS. This is primarily due to the slightly larger areas occupied by QuadRail's cells due to their inter-well spacing constraints. The series regulated QuadRail MAC occupies an additional 8% layout area because of the series regulator circuit: this, however, is dominated by three on-chip 4pF inter-rail MOS decoupling capacitors. Future deep-submicron processes, with

FIGURE 50 Static CMOS and QuadRail multiplier interconnect capacitance distributions.

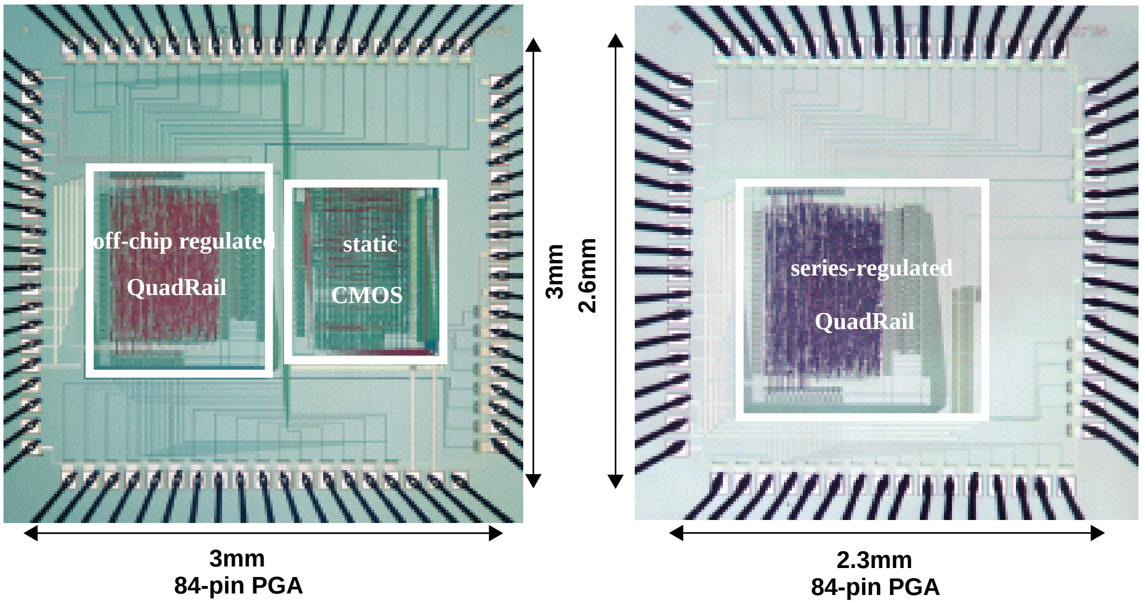


lower gate-oxide thicknesses, will offer higher MOS gate capacitance per unit area, mitigating this penalty significantly.

7.1.2 Static CMOS vs. Mixed Swing QuadRail Power-Delay Comparisons

We next present power-delay comparisons between the static CMOS and QuadRail approaches. We begin the comparisons with the 0.5µm process 16*16+36-bit MAC described in the previous section. Figure 51 shows the die microphotographs of the static CMOS and off-chip regulated and series regulated Mixed Swing QuadRail MACs fabricated in this process and the process characteristics. The comparisons are performed over a range of operating volt-

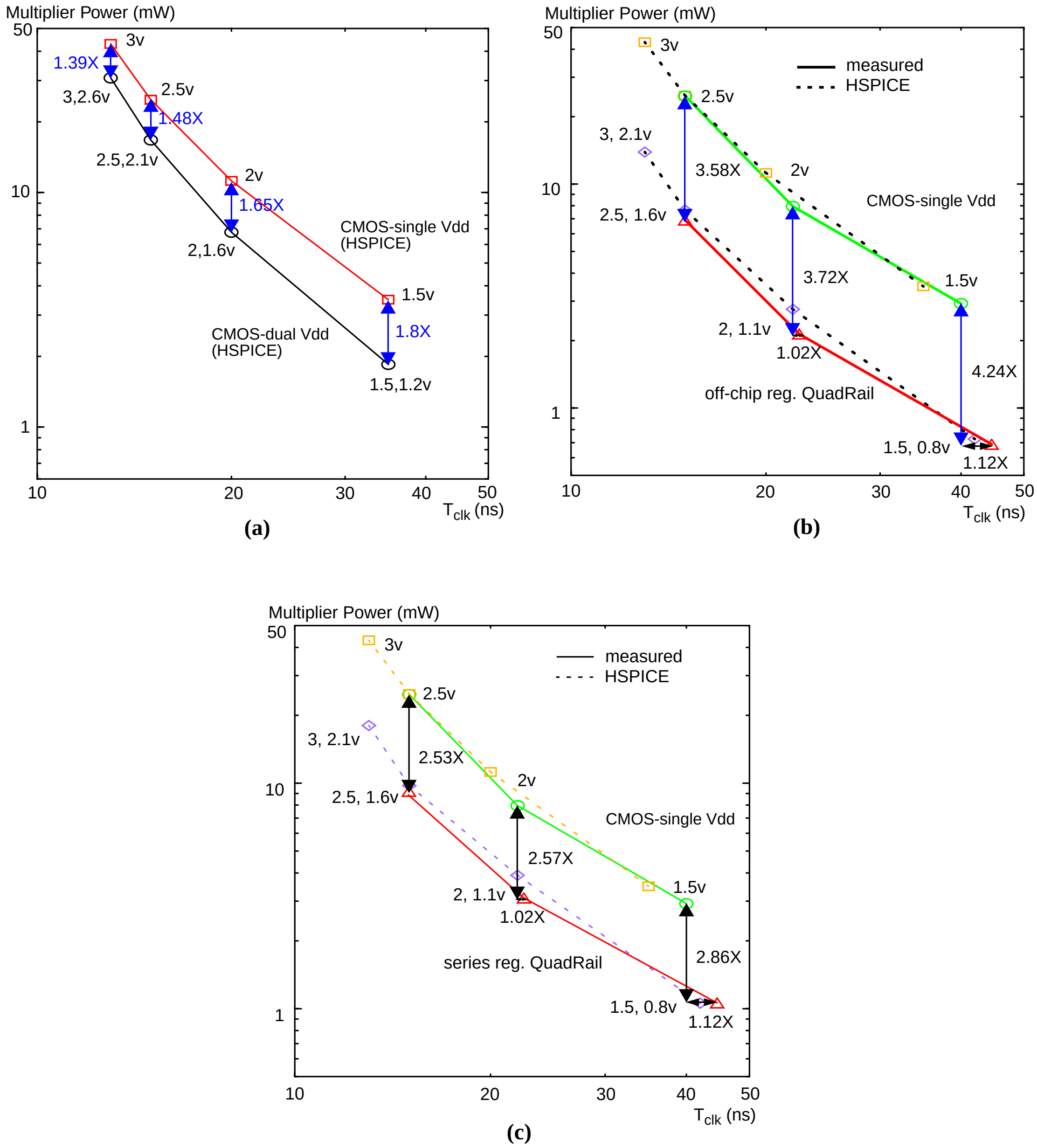
FIGURE 51 0.5µm 16*16+36-bit static CMOS and QuadRail MAC die microphotographs.



- 0.5µm L_{eff} CMOS (n-well) process.
- Single poly, triple metal.
- $V_{dd\text{-}max}$ = 3V.
- Tox = 96 Å.
- V_{tn} = 0.7V, V_{tp} = -0.9V.

ages to establish a power-delay comparison *space* (rather than perform the comparison at one operating voltage). Figure 52 shows the multiplier power vs. operating clock frequency comparisons for single-supply CMOS vs. dual-sup-

FIGURE 52 Multiplier power vs. T_{clk} comparisons for single-supply CMOS vs. dual-supply CMOS and QuadRail methodologies.



(a)

(b)

(c)

ply CMOS and the QuadRail methodologies. Power and delay are measured across 500 pseudo-random input vectors. Architecture-driven voltage scaling offers energy/operation savings ranging from 1.39X at CMOS V_{dd}=3V up to 1.8X at CMOS V_{dd}=1.5V. The final-adder-to-multiplier delay slack permits the multiplier's voltage to scale below the final adder's voltage offering this power savings. The off-chip regulated QuadRail methodology demonstrates even higher measured energy/operation savings ranging from 3.58X at V_{logic}=2.5V, V_{buffer}=1.6V (corresponding to the maximum measured clock frequency of 67 MHz) up to 3.79X at V_{logic}=1.5V, V_{buffer}=0.8V. This is because, as we observed in Chapter 4, QuadRail approach permits more effective voltage scaling than static CMOS, with the savings improving with even further voltage scaling. From the interconnect capacitance distributions of the static CMOS and QuadRail multipliers (Figure 50), the average interconnect capacitance within the multiplier is approximately 25fF. In addition, the average fanin gate capacitance 3,2 CSA in this process is 31.05fF, making the average load capacitance per switching-output node approximately 56fF. Analysis of the switching activity within the multiplier using unit-delay-model-based transition counting techniques reveals an activity factor of nearly 1.17, with up to 46% of the total transitions being spurious [Pursley97]. These factors make the effective switched capacitance per cycle substantial; the reduced voltage swing across this capacitance accounts for our energy/operation savings. In addition, HSPICE simulations show that nearly 28% of the buffer stage power is short-circuit power dissipation. The reduced buffer voltage swing, therefore, also offers a nearly cubic reduction in its short-circuit power, contributing to further energy/operation savings. The lower bounds for QuadRail voltage scaling (and hence maximum energy/operation savings) are limited by minimum noise margin constraints [Kakumu90].

At lower voltages, QuadRail demonstrates a small speed penalty. But the operating voltages still correspond to the lowest energy/operation, since that is our primary design goal. The delay penalty is because of the transition of the worst-case delay from the final adder to the QuadRail multiplier at low voltages. In the 0.5μm process, carrier velocity saturation is mitigated significantly at low voltages. Therefore, devices exhibit a nearly quadratic reduction in drive currents with voltage scaling. The lowered buffer voltage swing in the QuadRail multiplier thus incurs the delay penalty with voltage scaling. However, as we will shortly demonstrate, in future deep-submicron processes, drive current reduction with voltage scaling is linear due to carrier velocity saturation even at low voltages. This causes the final adder to remain the most time-critical MAC component even at low voltages, hence eliminating QuadRail's delay penalty.

The series regulated QuadRail methodology demonstrates comparatively lower measured energy/operation savings, ranging up to 2.55X (35% loss in savings compared to off-chip regulated QuadRail) at V_{logic}=1.5V, V_{buffer}=0.8V. This is due to the series regulator's static power which causes the buffer stage dynamic power savings to be linear rather than quadratic with voltage scaling. However, the nearly cubic short-circuit power reduction obtained is still retained despite series regulation, accounting for a measured savings slightly larger than linear. As we observed in Chapter 6, this methodology eliminates the necessity for an additional off-chip power supply while offering significantly improved low-voltage manufacturability. Further, because of its sleep-mode control, Series Regulated QuadRail's standby power at V_{logic}=1.5V,V_{buffer}=0.8V (152.5nW) is nearly three orders of magnitude lower than off-chip regulated QuadRail's standby power (143.8μW). This is because of the absence of a totempole current path in the logic stage during sleep mode. Figure 53 shows sample measured waveforms from the static CMOS and QuadRail MACs in the 0.5μm process.

To study the impact of process scaling on QuadRail, HSPICE simulated power-delay comparisons have been performed over three additional generations of commercial submicron processes: 3V,0.35μm bulk-CMOS; 2V,0.25μm fully-depleted SOI (FDSOI); and 0.16μm bulk-CMOS. Figure 54 shows the comparisons over a range of operating voltages in each process and the associated process characteristics. For proprietary reasons, the 0.16μm operating voltages and process details are not provided. Scaling feature sizes is accompanied with a modest increase in energy/operation savings at or near the

FIGURE 53 Measured waveforms from static CMOS and QuadRail 0.5μm 16*16+36-bit MACs.

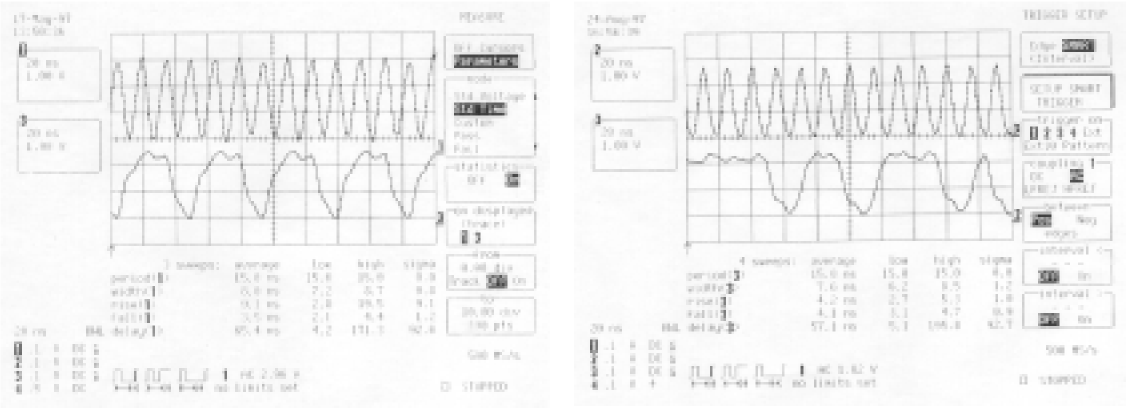
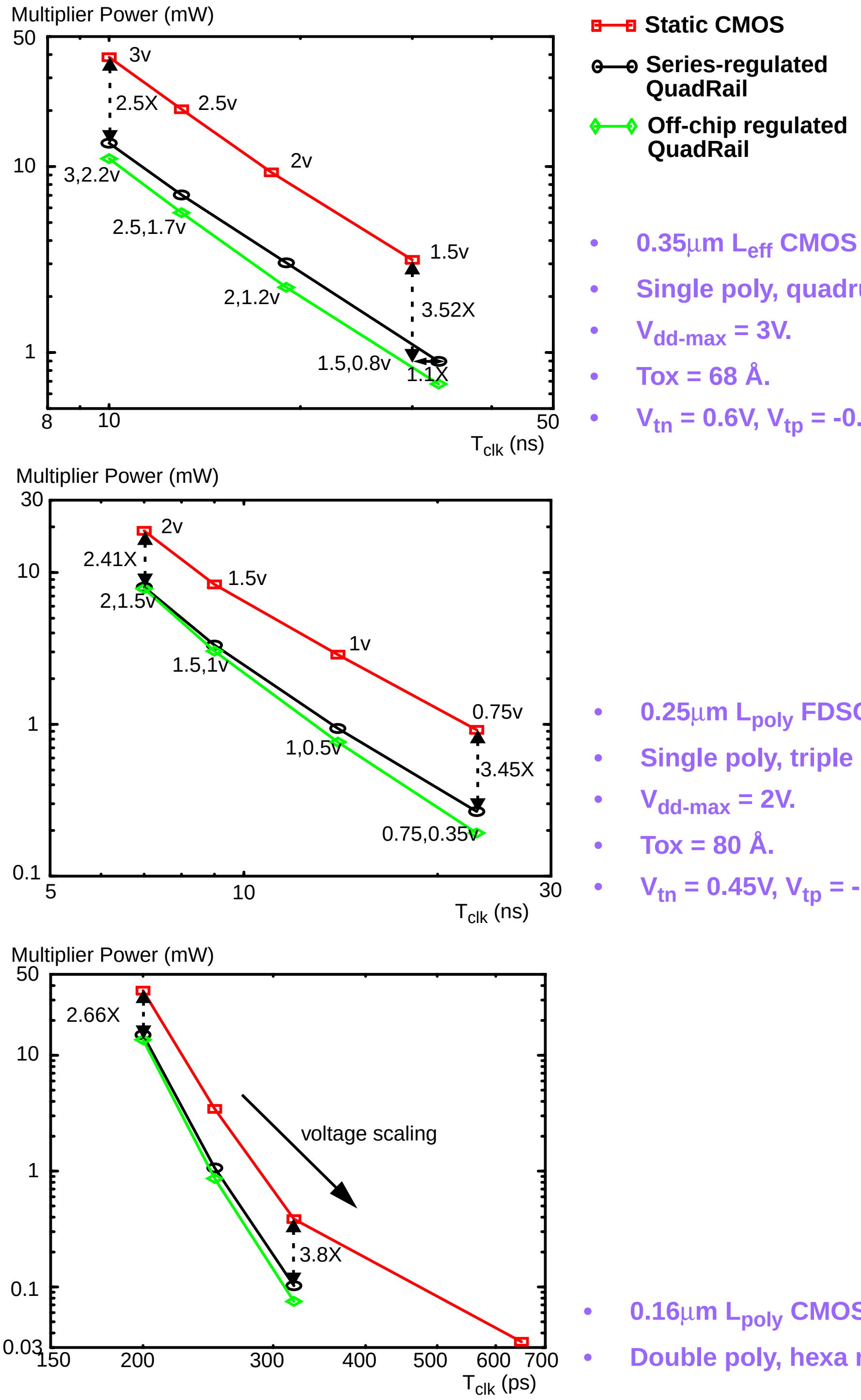


FIGURE 54 0.35µm, 0.25µm, and 0.16µm multiplier power vs. T_{clk} comparisons for single-supply CMOS vs. QuadRail methodologies.



- 0.35µm L_{eff} CMOS (n-well) process.
- Single poly, quadruple metal.
- $V_{dd\text{-}max}$ = 3V.
- Tox = 68 Å.
- V_{tn} = 0.6V, V_{tp} = -0.75V.

- 0.25µm L_{poly} FDSOI process.
- Single poly, triple metal.
- $V_{dd\text{-}max}$ = 2V.
- Tox = 80 Å.
- V_{tn} = 0.45V, V_{tp} = -0.45V.

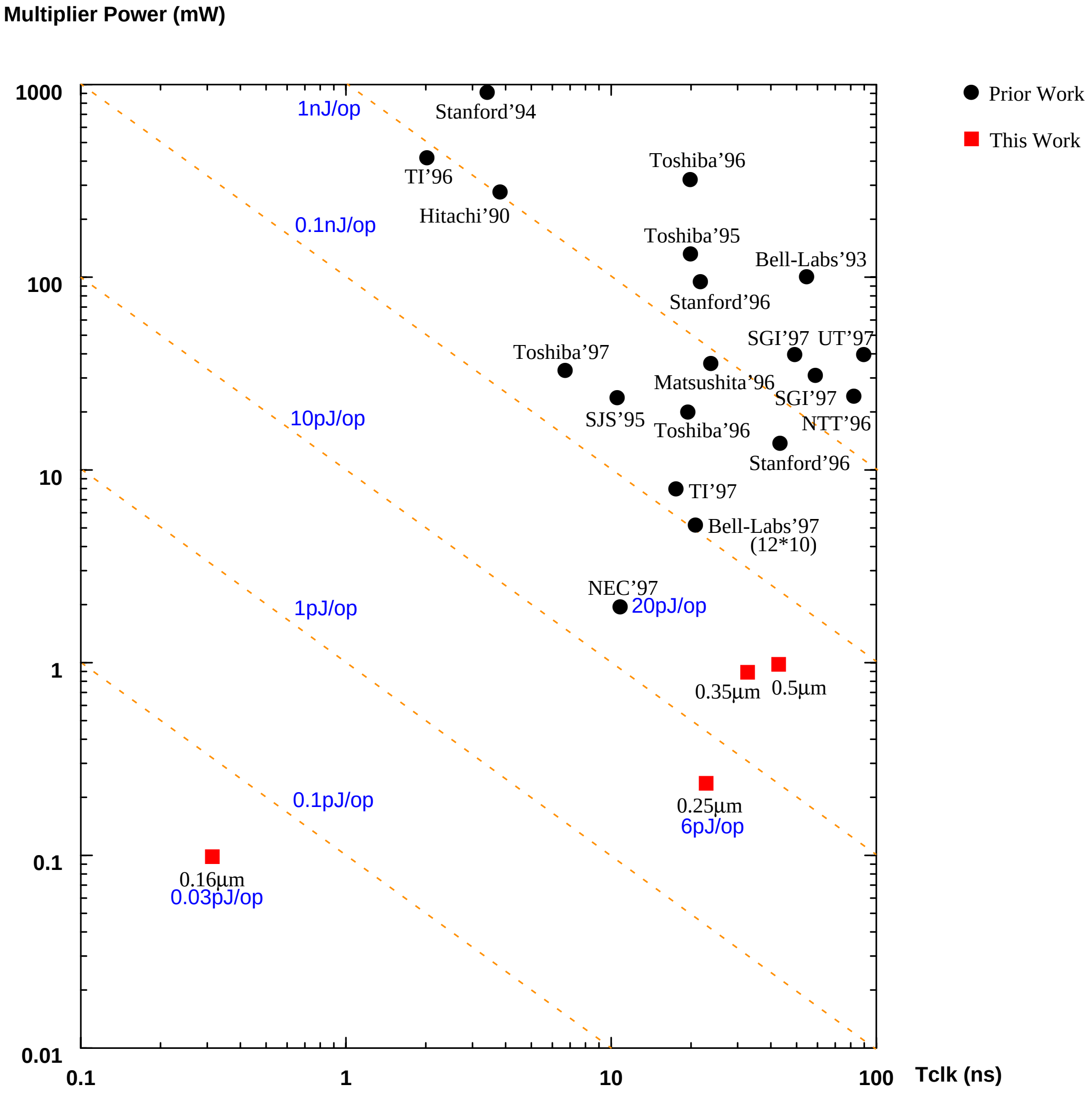
- 0.16µm L_{poly} CMOS (n-well) process.
- Double poly, hexa metal.

maximum process-permitted voltages. However, with voltage scaling, the energy/operation improvement is substantial: series regulated QuadRail's savings range up to 3.2X in 0.35μm, 3.45X in 0.25μm, and 3.8X in 0.16μm processes. This is attributed to the following deep-submicron QuadRail trends:

- Increasing ratios of logic to buffer voltage swings with voltage scaling. This leads to improved power savings with scaling feature sizes.
- Interconnect capacitance scaling slower than gate capacitance with process scaling. This results in improving energy/operation savings due to driving the load capacitances at reduced voltage swings.
- Carrier velocity saturation even at low voltages with scaling feature sizes. Therefore, lowering the buffer voltage swing continues to offer the power savings with process scaling, but with little or no impact on operating speed, significantly improving the energy/operation savings.
- lesser static power penalty due to series regulation due to lower multiplier on-drive current sourcing requirements with process scaling. Therefore, series regulated QuadRail's power-delay characteristics approach closer to off-chip regulated QuadRail's, making it more and more attractive in future deep-submicron processes.

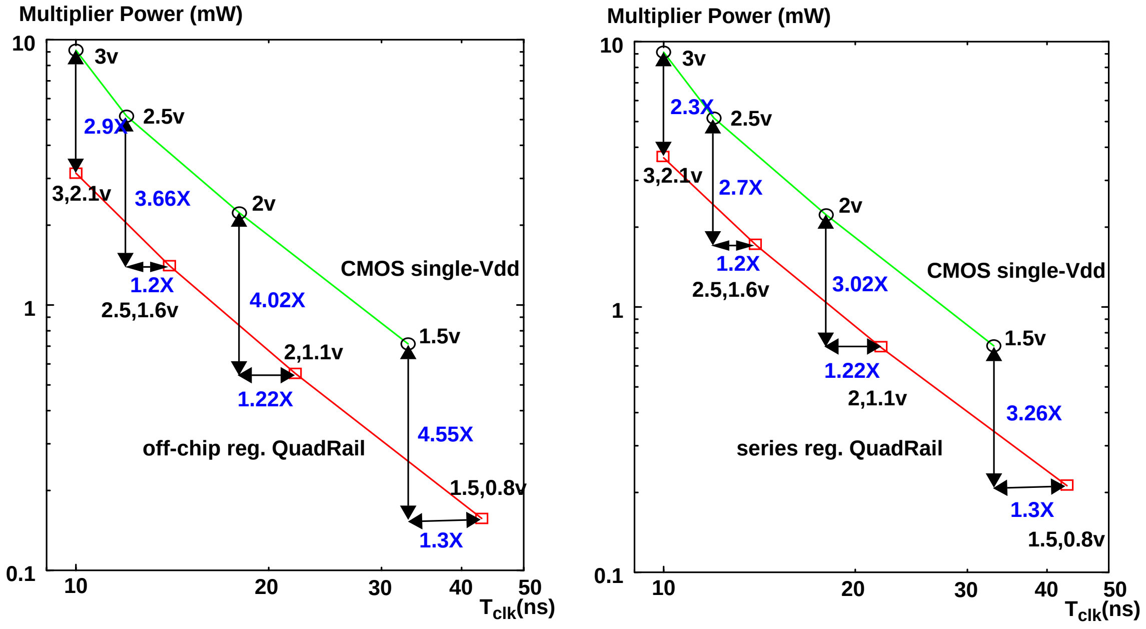
Figure 55 shows the series regulated QuadRail vs. previously published 16*16 multipliers energy/operation comparisons. Twenty of the lowest energy/operation multipliers that exist in literature to date to the best of our knowledge are displayed. These multipliers span a diverse spectrum of architectural, CAD toolflow, logic family, and bulk-CMOS/SOI fabrication process choices. The QuadRail approach offers a 3.3X lower energy/operation than the lowest energy/operation multiplier [Izumikawa97] in a comparable (0.25µm L_{poly}) process. To the best of our knowledge, the series regulated QuadRail 0.25µm and 0.16µm implementations are the first to cross below the 10pJ/operation barrier in standard submicron CMOS or SOI processes.

FIGURE 55 QuadRail vs. previous 16*16 multipliers energy/operation comparisons.



We next present static CMOS vs. QuadRail power-delay comparisons for the same MAC architecture, but across the range of bit-widths dominating commercial DSPs. Specifically, we consider 8*8+18-bit and 24*24+56-bit MACs in the 0.5µm process. Figure 56 and Figure 57 show the static CMOS vs. off-chip regulated and series regulated QuadRail multiplier power vs. T_{clk} comparisons for the 8*8+18-bit and 24*24+56-bit MACs respectively. Increasing operand bit-widths offers improved energy/operation savings due to the increasing effective switched capacitance per cycle, as we observed previously in Figure 13. In addition, increasing operand bit-widths causes an increase in the delay slack between the multiplier and final adder at low voltages, permitting further buffer voltage scaling and hence further energy/operation savings. Therefore, larger bit-width datapath circuits will benefit even further from the QuadRail methodologies.

FIGURE 56 Single-supply CMOS vs. QuadRail Power vs. T_{clk} comparisons for 8*8+18-bit MAC.

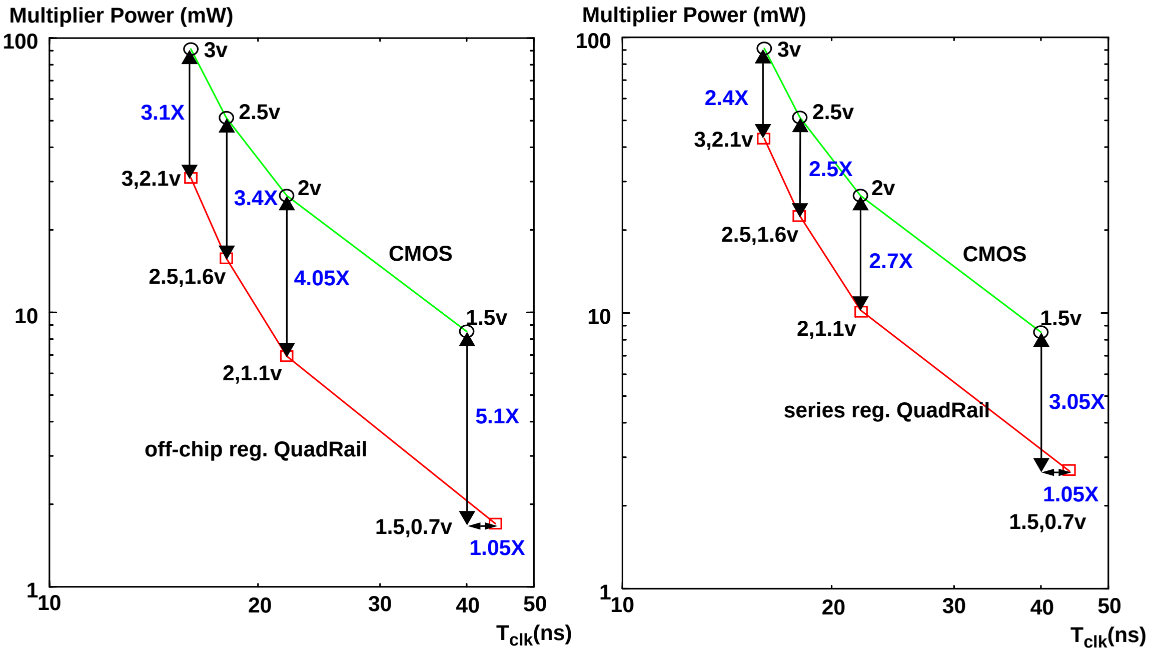


7.2 Microprocessor Floating-Point Units Comparisons

In this section, we describe the application of mixed voltage swing techniques to (i) single-layer point-to-point data buses and (ii) multi-layer multicast datapath nets within the floating-point units (FPUs) of two industrial next-generation microprocessors with extensive multimedia support in a 0.16µm bulk-CMOS process. The microprocessors are presently in design, and are expected to be announced in the 1999-2000 timeframe.

FPU data buses and multicast nets are becoming a substantial contributor to total power in next-generation general-purpose microprocessors. This is primarily because of the rapidly increasing integration of dedicated FPU-intensive multimedia instructions in modern processors [Ultrasparc95], [Pentium97].

FIGURE 57 Single-supply CMOS vs. QuadRail Power vs. T_{clk} comparisons for 24*24+56-bit MAC.

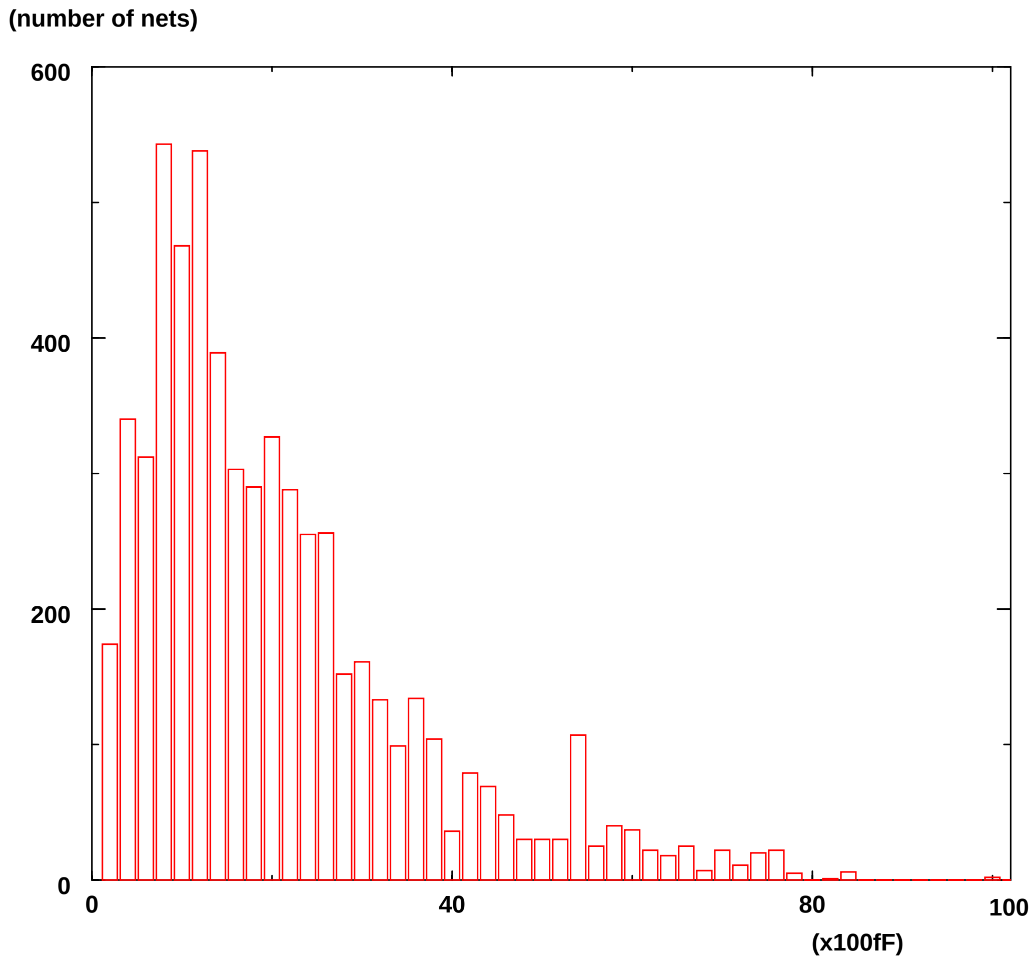


This has contributed to a significant increase in the physical capacitance charged/discharged within and between functional units in the FPU. In order to illustrate this, Figure 58 shows the FPU interconnect capacitance distribution within one of the two 0.16µm processors that this study focuses on. Interconnect capacitances are extracted through industrial in-house parasitic extractors from the fully placed-and-routed FPU layout. Further, the extensive multimedia support being incorporated has also contributed to a significant increase in the switching activities of FPUs, which were traditionally considered low activity. The substantial physical capacitance coupled with the high switching activities have made the effective switched capacitance per cycle, and hence dynamic power dissipation, within the FPUs a dominant bottleneck in next-generation microprocessors.

Figure 59 illustrates the generic experimental circuit setup, consisting of fully placed-and-routed buses and multicast nets between functional units within the FPU. The data buses are point-to-point and span a single metal layer (typically metal3 or metal5). The multicast datapath nets are inter-unit signals broadcasted to multiple receiving units and span across four metal layers (typically metal2 through metal5) in this six-metal-layer process. Full coverage of the signal interconnect on the top and bottom is considered to maximize coupling capacitance, thereby worst-casing performance and power dissipation. Further, the signal interconnect is shielded on both neighboring sides through grounded shield lines to minimize signal crosstalk. Signalling methodology is fully-differential to minimize common-mode noise coupling. The setup is illustrated in Figure 60 and Figure 61 for a 10000µm metal5 data bus and 10473µm multicast datapath net from the FPU respectively.

Static and dynamic fully-differential mixed swing methodologies are developed to implement driver and receiver circuits for the FPU data buses and multicast nets to lower their power consumption. The approaches are compared against static CMOS and proprietary dynamic mixed swing methodologies. The power comparisons are at a target clock frequency of 1GHz at the nomi-

FIGURE 58 FPU interconnect capacitance distribution.



nal-process, low-voltage, high-temperature corner. Specific case studies are conducted on 4000µm, 8000µm, and 10000µm data buses and a 10473µm multicast datapath net within the FPUs of the two processors. Optimal transistor sizing of the driver and receiver circuits, and optimal repeater insertion and wire sizing of the interconnects, are performed through in-house optimization

FIGURE 59 Experimental circuit setup: fully placed-and-routed buses and multicast nets.

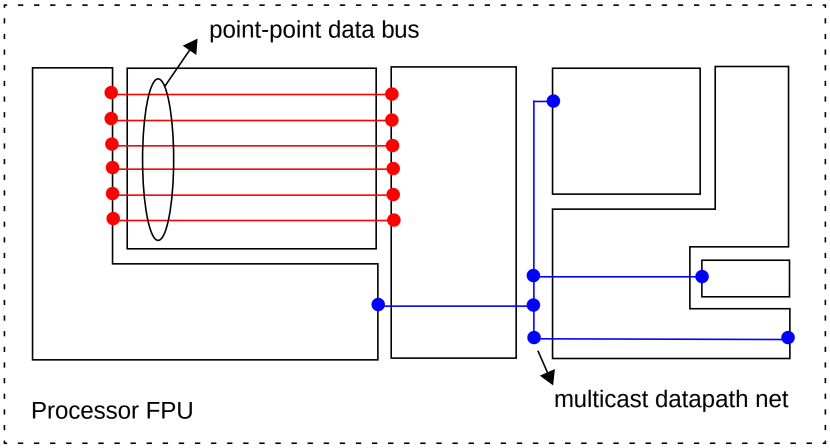
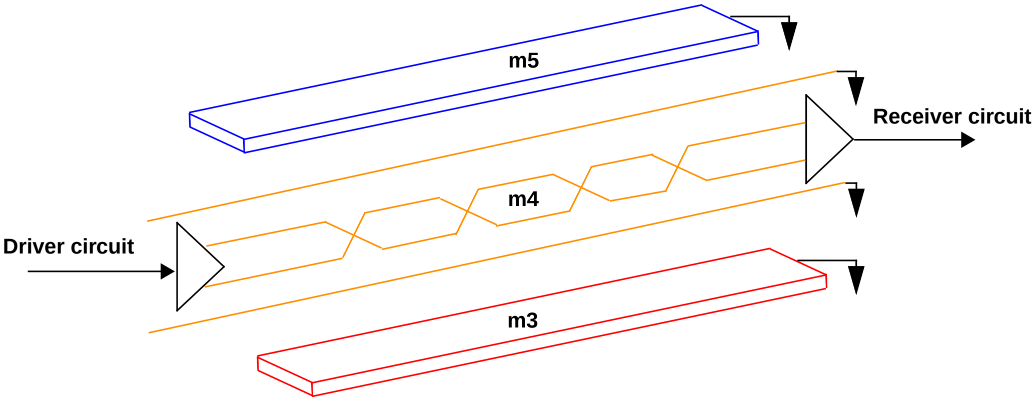


FIGURE 60 10000µm point-to-point FPU data bus experimental setup.



true/complimentary bus total capacitance = 2307fF.

toolsuites to minimize power consumption of each methodology while meeting the target clock frequency constraint.

Figure 62 through Figure 65 shows the power comparisons between static CMOS and the four mixed swing approaches developed in the 0.16µm process, obtained through in-house circuit simulations using customized BSIM3v3 models. The comparisons are performed over a range of input data switching activities (transitions per clock cycle) between 0.01 up to 1.0, and for two low-voltage swing specifications: 150mV and 500mV.

It is observed that the dynamic mixed swing approaches consume *higher* power than static CMOS at low input data activities. This is due to the inherently high switching activities of dynamic techniques, since output nodes are precharged and evaluated every clock cycle. Moreover, the high clock power required to drive the precharge/evaluate devices further penalizes their power

FIGURE 61 10473µm multicast FPU datapath net experimental setup.

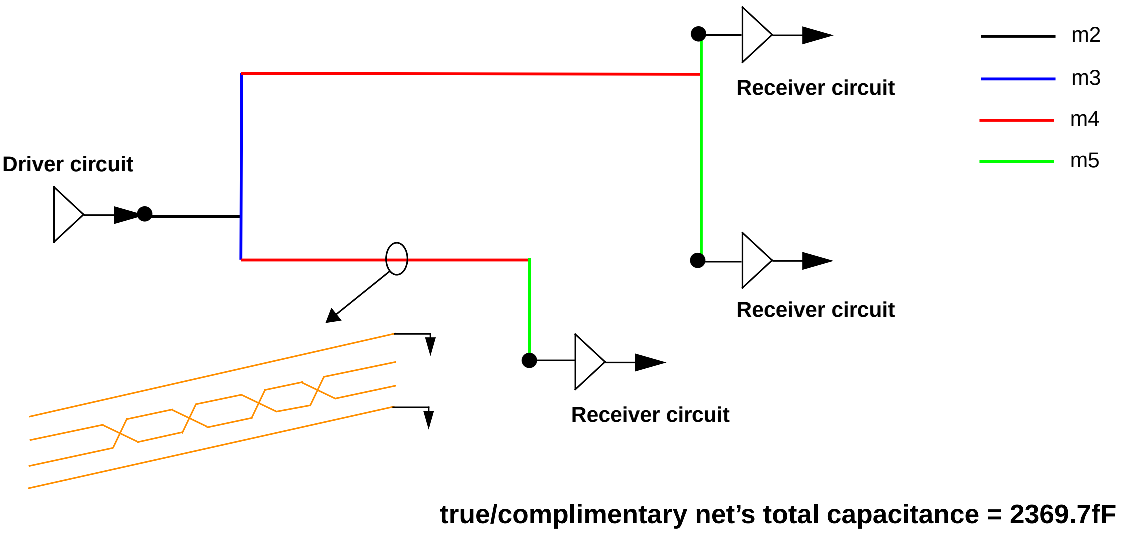


FIGURE 62 Power vs. input switching activity comparisons for 4000µm data bus.

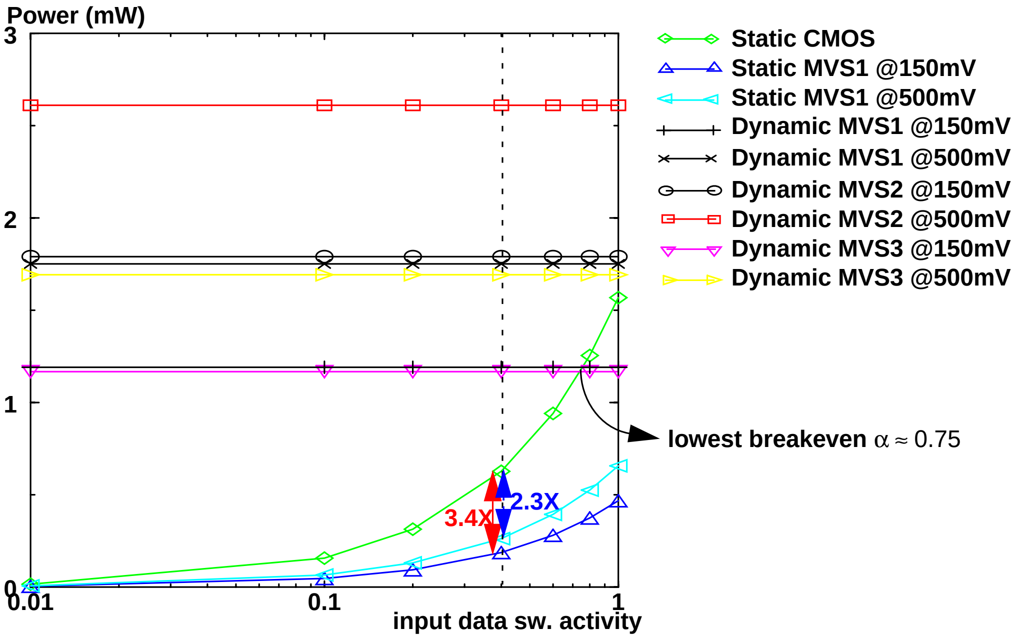


FIGURE 63 Power vs. input switching activity comparisons for 8000µm data bus.

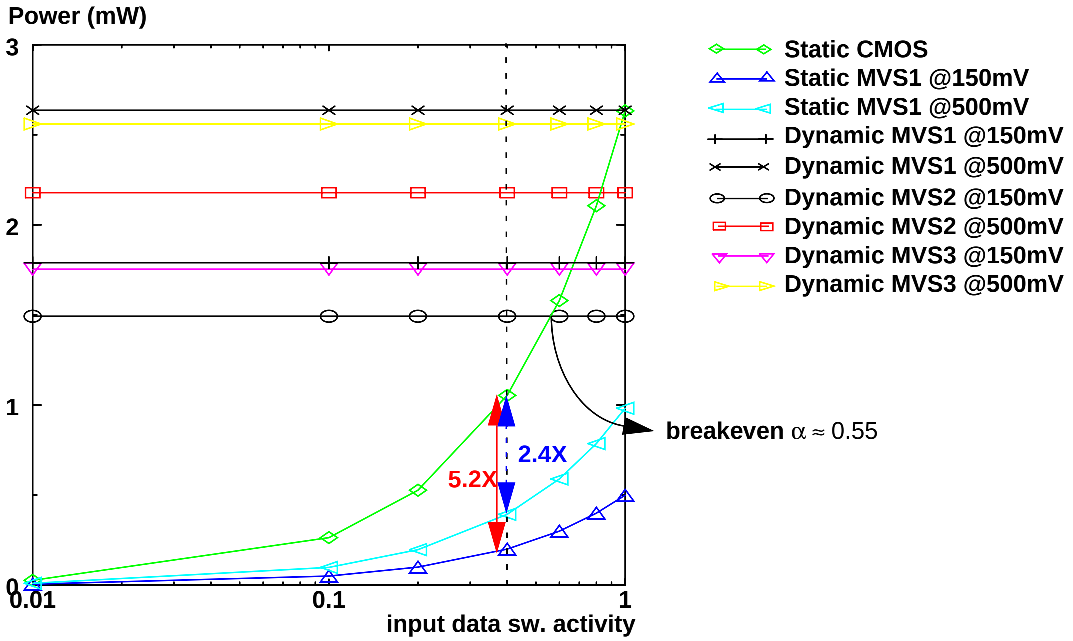


FIGURE 64 Power vs. input switching activity comparisons for 10000µm data bus.

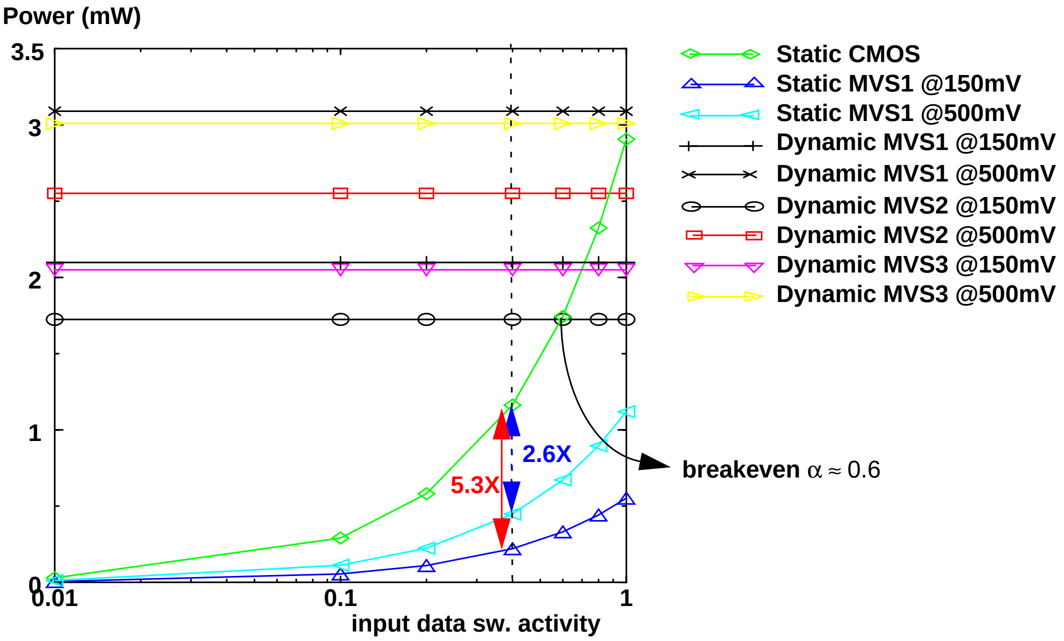
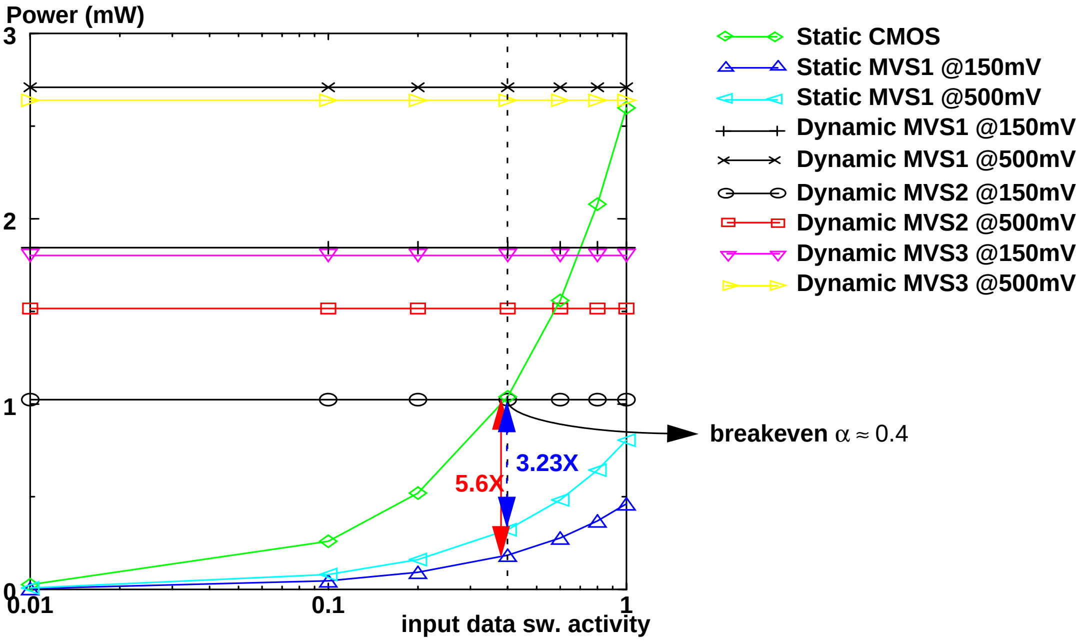


FIGURE 65 Power vs. input switching activity comparisons for 10473µm multicast net.



consumption, offsetting any savings achieved due to the lowered interconnect voltage swing. At high switching activities, dynamic techniques breakeven with static CMOS, and offer modest savings only at activities very close to unity. The lowest breakeven switching activity among the cases studied was 0.55. However, across several multimedia benchmarks the average switching activities of these FPUs was determined to be approximately 0.40. Therefore, dynamic mixed swing techniques consume higher power than full-swing static CMOS if employed in the FPUs of these processors.

Static mixed swing techniques have the potential to offer substantial power savings, because their switching activities are identical to that of static CMOS and do not require precharge/evaluate clock signals. This enables the dynamic power reduction achieved due to lower interconnect voltage swing to be maximally exploited, with the savings improving with increasing switching activities. At the average FPU switching activity of 0.40, the static mixed swing techniques demonstrate power savings up to 3.4X for the 4000μm data bus example and up to 5.6X for the 10473μm multicast net example, at the target operating clock frequency of 1 GHz.

7.3 Summary

In this chapter, we explored the potential of mixed swing approaches to achieve energy/operation savings over static CMOS datapath circuits. The studies were conducted on two types of datapath circuits: (i) signed, fixed-point DSP multiplier-accumulators over a range of operand bit-widths, power supply voltages, and commercial bulk-CMOS and fully-depleted SOI processes, and, (ii) data buses and multicast datapath nets of the floating-point units of two industrial next-generation multimedia-enriched microprocessors presently in design in a 0.16µm bulk-CMOS process.

By exploiting the final-adder-to-multiplier delay slack for our prototype MAC architecture, we investigated the ability of the off-chip regulated and series regulated QuadRail methodologies to lower the energy/operation savings of the power-critical Wallace tree multiplier over single-supply static CMOS and architecture-driven voltage scaled, dual-supply static CMOS approaches. The studies were performed on 8-24-bit MACs, since this range of bit-widths dominates commercial DSPs. Through measurements on fabricated MACs and intensive circuit simulations, substantial energy/operation savings were demonstrated with the savings increasing with operand bit-widths. The comparisons were conducted over a range of operating voltages to study the impact of our savings with voltage scaling. The increasing ratios of logic to buffer voltage swings was observed to offer improving energy/operation savings with voltage scaling. The comparisons were extended across four submicron process generations: 0.5µm bulk-CMOS, 0.35µm bulk-CMOS, 0.25µm FDSOI, and 0.16µm bulk-CMOS. In addition, the series regulated QuadRail 16*16 Wallace tree multiplier's energy/operation in these four processes were compared against twenty of the lowest energy/operation 16*16 multipliers published in literature. Increasing energy/operation savings with process scaling was demonstrated and deep-submicron trends that contribute to further improvements in QuadRail's savings in future processes were outlined.

Energy/operation comparisons were also performed on single-layer point-to-point data buses and multi-layer multicast datapath nets within the FPUs of two industrial next-generation multimedia-enriched microprocessors presently in design in a 0.16µm bulk-CMOS process. The comparisons were conducted over a range of input data switching activities at target clock frequency specifications. At an average switching activity of 0.40, up to 5.6X energy/operation savings over static CMOS was demonstrated.

8 Conclusions

8.1 Thesis Summary

The portable communications industry's vision of integrating a complete multimedia complex on a single die, coupled with the desktop computing industry's vision of integrating more and more multimedia functionality onto general purpose microprocessors has made lowering the power consumption of DSP datapath circuits an increasingly important priority in current and future fabrication processes. While fully-static CMOS techniques accompanied with supply voltage scaling have been popular low-power design techniques over the last decade, fundamental limitations impose a lower bound to their applicability in future deep-submicron processes, motivating a strong necessity for exploring alternate low-power datapath design methodologies.

This thesis has explored Mixed Swing techniques for lowering the energy/operation of static CMOS datapath circuits in standard submicron bulk-CMOS and SOI processes. Multiple power supply-based approaches were examined to implement standard datapath primitive functions by intermixing high- and low-voltage signal swings while driving interconnect and gate-fanout load capacitances at reduced voltage swings. We demonstrated that this approach allows exploiting the best aspects of both static CMOS and voltage scaling within a single gate. Static CMOS-, Domino/Pass-Transistor Logic-, and Cascode Voltage Switch Logic-based mixed swing techniques were investigated. A fully static, single-ended, four-power-supply-rail methodology called *Mixed Swing QuadRail* presented here was shown to offer substantial energy/operation savings on datapath circuits with interconnect capacitance dominance, e.g., Wallace tree multipliers. A Domino/Pass-transistor Logic-based, single-phase precharge/evaluate clocked, singe-ended methodology and a CVSL-based, fully static, fully-differential methodology developed here was shown to offer substantial energy/operation savings on datapath circuits with gate capacitance dominance, e.g., adders and adder variants.

In order to explore the design space of multi-supply approaches, posynomial power and delay formulations for Mixed Swing QuadRail were developed using the n^{th}-Power Law submicron MOSFET model and their accuracy validated through HSPICE simulations. Based on our models, optimal voltage scaling and transistor sizing approaches were developed to minimize energy/operation of mixed swing circuits. The importance of employing these optimization approaches, particularly in future low-voltage technologies, was motivated through experimental results from a 16*16+36-bit Booth-recoded, Wallace-tree DSP multiplier-accumulator (MAC) in a commercial 3V, 0.5µm bulk-CMOS process.

Two of the most critical low-voltage practicality challenges to mixed swing techniques - manufacturability and noise immunity - were addressed. Worst-case process and temperature corners were developed and a relative manufacturability and noise immunity analysis performed on static CMOS and Mixed Swing QuadRail. A modestly better low-voltage manufacturability and adequately high low-voltage noise immunity was demonstrated for QuadRail. For further improvement in manufacturability, a series regulation approach for Mixed Swing QuadRail was developed to effectively offset threshold voltage variations across worst-case corners. Up to a 1.8X better low-voltage manufacturability was achieved relative to static CMOS. Further, the series regulated approach eliminated the necessity for an additional explicit off-chip power supply, transforming Mixed Swing QuadRail into a self-contained methodology which can replace full-swing static CMOS operating between a regular, high-voltage supply without warranting any technology or system-level modifications.

Through fabricated datapath integrated circuits and intensive circuit simulations in commercial bulk-CMOS and SOI processes, we demonstrate the ability of off-chip regulated and on-chip series regulated mixed swing techniques to voltage-scale more aggressively than static CMOS well into the deep-submicron regime. Substantial energy/operation savings were achieved for Wallace tree multipliers of DSP MACs over a range of operand bit-widths, power supply voltages, and technology generations down until 0.16µm L_{poly} (0.12µm L_{eff}) feature sizes. Substantial power savings were also achieved over static CMOS on point-to-point data buses and multicast datapath nets within the floating-point units of two industrial 0.16µm next-generation microprocessors with extensive multimedia support, over a range of operating voltages and input data switching activities for target clock frequency specifications.

8.2 Future Directions

The research work explored in this thesis can be extended in a number of future directions:

- The feasibility of applying our mixed swing techniques beyond the domain of short bit-width datapath circuits should be investigated. Wider datapath operators, commonly employed in general-purpose processor integer and floating-point execution units, and control-path circuits with substantial interconnect capacitance are prime candidates for lowering energy/operation by exploiting mixed swing techniques. The increasing interconnect dominance within these circuits makes lowering their energy/operation all the more crucial in future deep-submicron processes.
- Alternate static, single-ended mixed swing methodologies to achieve even further energy/operation savings should be explored, particularly for variable throughput, data-driven signal processing datapath. The Mixed Swing QuadRail suffers from a modest static power penalty in the logic stage that is eliminated during sleep mode by the series regulator. In data-driven signal processing circuits, where throughput varies as a function of workload, sleep mode is not always enabled during inactivity, since very frequent transitions may occur between active and standby operation modes. In such applications, this static power penalty may be prohibitive. Some of the

ongoing research along this direction on investigating mixed swing pass-transistor logic-based techniques are outlined in [Carley97].

- While the ideas presented in this thesis have examined intermixing high- and low-voltage signals to perform datapath primitive logic functions, an improved form of clustered voltage scaling [Usami97] may be investigated to achieve further energy/operation savings, by intermixing static CMOS and mixed swing primitives within the same datapath. Due to the relatively lower absolute noise margins of mixed swing methodologies, this will involve careful ‘noise-aware’ layout of the datapath. Ongoing research along this direction addressing the associated physical CAD challenges are outlined in [Rutenbar97].

Bibliography

[Acken83] J.M. Acken, "Testing for Bridging Faults (Shorts) in CMOS Circuits", *Proc. IEEE/ACM Design Automation Conference,* June 1983, pp. 717-718.

[Allen85] J. Allen, "Computer Architecture for Digital Signal Processing", *Proc. of the IEEE,* Vol. 73, No. 5, May 1985.

[Antoniadis97] D. Antoniadis, "SOI CMOS as a Mainstream Low-Power Technology: A Critical Assessment", Digest of technical papers, *IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1997, pp. 295-300.

[Ardekani93] J.F. Ardekani, "MxN Booth Encoded Multiplier Generator Using Optimized Wallace Trees", *IEEE Trans. on VLSI Systems*, Vol. 1, June 1993, pp. 120-125.

[Athas97] W. Athas et al, "AC1: A Clock-Powered Microprocessor", Digest of technical papers, *IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1997, pp. 328-333.

[Bakoglu85] H.B. Bakoglu and J.D. Meindl, "New CMOS Driver and Receiver Circuits to Reduce Interconnection Propagation Delays", Digest of technical papers, *Symposium on VLSI Technology*, May 1985, pp. 54-55.

[Bakoglu90] H.B.Bakoglu, *Circuits, Interconnects, and Packaging for VLSI*, Addison-Wesley, 1990.

[Booth51] A.D. Booth, "A Signed Binary Multiplication Technique", *Quarterly Journal of Mathematics*, Vol. 4, 1951.

[Borel97] J. Borel, "Technologies for Multimedia Systems on a Chip", Digest of technical papers, *IEEE Intl. Solid State Circuits Conference*, February 1997, pp. 18-21.

[Brglez85] F. Brglez and H. Fujiwara, "A Neutral Netlist of 10 Combinational Benchmark Circuits and a Target Translator in FORTRAN", *Proc. IEEE Intl. Symposium on Circuits and Systems*, 1985, pp. 663-698.

[Burr91] J.B. Burr and A.M. Peterson, "Energy Considerations in Mutichip-module based Multiprocessors", Proc. *IEEE Intl. Conference on Computer Design*, 1991, pp. 593-600.

[Burr94] J.B. Burr and J. Shott, "A 200mV Self-Testing Encoder/Decoder using Stanford Ultra Low Power CMOS", Digest of technical papers, *IEEE Intl. Solid State Circuits Conference*, February 1994, pp. 84-85.

[Carley94] L.R. Carley, “QuadRail: A Design Methodology for Ultra Low Power Integrated Circuits”, *Proc. IEEE Intl. Workshop on Low Power Design*, April 1994.

[Carley97] L.R. Carley, "Design of Low Energy/Operation Digital Logic Circuits", *DARPA Review Meeting*, Dept. of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, August 1997.

[Cavanagh84] J.F. Cavanagh, *Digital Computer Arithmetic: Design and Implementation*, McGraw Hill, 1984.

[Chandra95] A.P. Chandrakasan and R.W. Broderson, *Low Power Digital CMOS Design*, Kluwer Academic, 1995.

[Chandra96] A.P. Chandrakasan et al, "Data-driven Signal Processing: An Approach for Energy Efficient Computing", Digest of technical papers, *IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1996, pp. 347-352.

[Chatterjee95] P.K. Chatterjee, "Trends for Deep-submicron VLSI and their Implications for Reliability", *Proc. Intl. Reliability Physics Symposium*, 1995, pp. 1-11.

[Chen95] Z. Chen et al, "Optimization of Quarter Micron MOSFETs for Low-Voltage/Low-Power Applications", Digest of technical papers, *IEEE Intl. Electron Devices Meeting*, December 1995, pp. 3.3.1-3.3.4.

[Chu87] K.M. Chu and D. Pulfrey, "Comparisons of CMOS Circuit Techniques: Differential CVSL Vs. Conventional Logic", *IEEE J. Solid-State Circuits*, Vol. 22, August 1987, pp. 528-532.

[Davari95] B. Davari, R. Dennard, and G. Shahidi, "CMOS Scaling for High Performance and Low Power - The Next Ten Years", *Proc. of the IEEE*, Vol. 83, April 1995, pp. 595-606.

[Davari96] B. Davari, "CMOS Technology Scaling 0.1mm and Beyond", *Proc. IEEE Intl. Electron Devices Meeting*, December 1996, pp. 21.1.1-21.1.4.

[De96] V.K. De and J.D. Meindl, "A Dynamic Energy Recycling Logic Family for Ultra Low-Power GSI", Digest of technical papers, *IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1996, pp. 371-375.

[Douseki97] T. Douseki et al, A 0.5V MTCMOS/SIMOX Logic Gate", *IEEE J. Solid-State Circuits*, Vol. 32, October 1997, pp. 1604-1609.

[Ecker80] J. Ecker, "Geometric Programming: methods, computations, and applications", *SIAM Review*, July 1980, pp. 338-362.

[Eisele95] M. Eisele et al, "Intra-Die Device Parameter Variations and their Impact on Digital CMOS gates at Low Supply Voltages", Digest of technical papers, *IEEE Intl. Electron Devices Meeting*, December 1995, pp. 3.4.1-3.4.4.

[Favalli95] M. Favalli and L. Benini, "Analysis of glitch power dissipation in CMOS ICs", *Proc. IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1995, pp. 123-128.

[Fishburn85] J.P. Fishburn and A.E. Dunlop, "TILOS: A Posynomial Programming Approach to Transistor Sizing", *Proc. IEEE Intl. Conference on Computer Aided Design*, November 1985, pp. 326-328.

[Frank97] D.J. Frank, P. Solomon, S. Reynolds, and J. Shin, "Aupply and Threshold Voltage Optimization for Low Power Design", *Proc. IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1997, pp. 317-322.

[FutureBus83] Project P896.1 - FutureBus Proposed Standard Specification, Draft 6.2, *IEEE Computer Society*, November 1983.

[Goncalves83] N.F. Goncalves and H. DeMan, "NORA: A Racefree Dynamic CMOS Technique for Pipelined Logic Structures", *IEEE J. Solid-State Circuits*, Vol. 18, June 1983, pp. 261-266.

[Goto92] G. Goto et al, "A 54x54 Regularly Structured Tree Multiplier", *IEEE J. Solid-State Circuits*, Vol. 27, September 1992, pp. 1229-1235.

[Gray94] P.R. Gray, H.S. Lee, J.M. Rabaey, C.G. Sodini, and B.A. Wooley, "Challenges and Opportunities in Low Power Integrated Circuit Design", *SRC Research Report S94019*, November 1994.

[Gu96] R.X. Gu and M.I. Elmasry, "Power Dissipation Analysis and Optimization of Deep-submicron CMOS Digital Circuits", *IEEE J. Solid-State Circuits*, Vol. 31, May 1996, pp. 707-713.

[Heden87] N. Hedenstierna and K.O. Jeppsen, "CMOS Circuit Speed and Buffer Optimization", *IEEE Trans. Computer Aided Design of IC's*, Vol. 6, March 1987, pp. 270-281.

[Heller84] L.G. Heller et al, "Cascode Voltage Switch Logic: A Differential CMOS Logic Family", Digest of technical papers, *IEEE Intl. Solid State Circuits Conference*, February 1984, pp. 16-17.

[Hoppe90] B. Hoppe, G. Neuendorf, D.S. Landsiedel, and W. Specks, "Optimization of High-Speed CMOS Logic Circuits with Analytical Models for Signal Delay, Chip Area, and Dynamic Power Dissipation", *IEEE Trans. Computer Aided Design of IC's*, Vol. 9, March 1990, pp. 236-247.

[Horowitz94] M. Horowitz, T. Indermaur, and R. Gonzalez, “Low-power Digital Design”, *Proc. IEEE Intl. Symposium on Low Power Electronics*, October 1994.

[Igarashi97] M. Igarashi et al, "A Low-Power Design Method using Multiple Supply Voltages", Digest of technical papers, *Proc. IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1997, pp. 36-41.

[Izumikawa97] M.Izumikawa et al., “A 0.25μm CMOS 0.9V 100MHz DSP Core”, *IEEE J. Solid-State Circuits*, Vol. 32, Jan. 1997, pp. 52-61.

[Jou95] S.J. Jou et al, "A Pipelined MAC using A High-Speed Low-Power Static and Dynamic Full Adder Design", *Proc. IEEE Custom Integrated Circuits Conference*, May 1995, pp. 593-596.

[Kakumu90] M. Kakumu and M. Kinugawa, "Power Supply Voltage Impact on Circuit Performance for Half and Lower Submicrometer CMOS LSI", *IEEE Trans. Electron Devices*, Vol. 37, August 1990, pp. 1902-1908.

[Khater96] I.S.A. Khater, A. Bellaouar, and M.I. Elmasry, "Circuit Techniques for CMOS Low-power High-Performance Multipliers", *IEEE J. Solid-State Circuits*, Vol. 31, October 1996, pp. 1535-1546.

[Knight88] T.F. Knight and A. Krymm, "A Self_terminating Low-Voltage Swing CMOS Output Driver", *IEEE J. Solid-State Circuits*, Vol. 23, April 1988, pp. 457-464.

[Ko95] U. Ko, P.T. Balsara, and W. Lee, "Low-power Design Techniques for High-Performance CMOS Adders", *IEEE Trans. on VLSI Systems*, Vol. 3, June 1995, pp. 327-333.

[Kobayashi94] T. Kobayashi and T.Sakurai, "Self-Adjusting Threshold-Voltage Scheme for Low-Voltage High-Speed Operation", *Proc. IEEE Custom Integrated Circuits Conference*, May 1994, pp. 271-274.

[Krambeck82] R.H. Krambeck et al, "High-speed Compact Circuits with CMOS" *IEEE J. Solid-State Circuits*, Vol. 17, June 1982, pp. 614-619.

[Krishna95] R.K. Krishnamurthy and R. Sridhar, "A CMOS Wave-pipelined Image Processor for Real-time Morphology", *Proc. IEEE Intl. Conference on Computer Design*, October 1995, pp. 638-643.

[Krishna96a] R.K. Krishnamurthy, I. Lys, and L.R. Carley, "Static Power-driven Voltage Scaling and Delay-driven Buffer Sizing in Mixed Swing QuadRail", *Proc. IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1996, pp. 381-386.

[Krishna96b] R.K. Krishnamurthy, I. Lys, and L.R. Carley, "Mixed Swing QuadRail: Exploring Multiple Voltage Swings for Low Energy/Operation of Digital Circuits", *SRC Research Report C96538*, November 1996.

[Krishna97] R.K. Krishnamurthy and L.R. Carley, "Exploring the Design Space of Mixed Swing QuadRail for Low Power Digital Circuits", *IEEE Trans. on VLSI Systems*, Vol. 5, December 1997, pp. 388-400.

[Kuroda96] T. Kuroda et al, "A 0.9V 150MHz 10mW 4mm2 2-D DCT Core Processor with Variable Threshold Voltage Scheme", Digest of technical papers, *IEEE Intl. Solid-State Circuits Conference*, February 1996, pp. 166-167.

[Landman93] P.E. Landman and J.M. Rabaey, "Power Estimation for High Level Synthesis", *Proc. European Design Automation Conference*, February 1993, pp. 361-366.

[Lapsley96] P. Lapsley, J. Bier, A. Shoham, and E. Lee, *DSP Processor Architectures and Features*, Berkeley Design Technology Inc., 1996.

[Larsson96] P. Larsson and C.J. Nicol, "Transition Reduction in Carry Save Adder Trees", *Proc. IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1996, pp. 85-88.

[Lee86] C.M. Lee and E.W. Szeto, "Zipper CMOS", *IEEE Circuits and Systems Magazine*, May 1986, pp. 10-16.

[Liu93] D. Liu and C. Svensson, "Trading Speed for Low Power by Choice of Supply and Threshold Voltages", *IEEE J. Solid-State Circuits*, Vol. 28, January 1993, pp. 10-17.

[Lu93] F. Lu and H. Samueli, "A 200 MHz CMOS Pipelined MAC Using Quasi-Domino Dynamic Full Adder Cell Design", *IEEE J. Solid-State Circuits*, Vol. 28, February 1993, pp. 123-132.

[Maly90] W. Maly, "Computer Aided Design for VLSI Circuit Manufacturability", *Proc. of the IEEE*, Vol. 78, February 1990, pp. 356-392.

[Maly96] W.Maly et al, "Design for Manufacturability in Submicron Domain", *Proc. IEEE/ACM Intl. Conference on Computer Aided Design*, Nov. 1996, pp. 690-697.

[Michaels96] K. Michaels, PDF Solutions Inc., San Jose, CA, *Private communications*, November 1996.

[Montoye90] R.K. Montoye et al, "An 18 ns 56-bit multiply-adder circuit", Digest of technical papers, *IEEE Intl. Solid State Circuits Conference*, February 1990, pp. 336-337.

[Murakami96] H. Murakami et al, "A MAC Macro for a 45 MIPS Embedded RISC Processor", *IEEE J. Solid-State Circuits*, Vol. 31, July 1996, pp. 1067-1071.

[Nagamatsu95] M. Nagamatsu et al, "A 150 MIPS/W CMOS RISC Processor for PDA Applications", Digest of technical papers, *IEEE Intl. Solid State Circuits Conference*, February 1995, pp. 114-115.

Nagendra94] C. Nagendra, R.M. Owens, and M.J. Irwin, "Low Power Tradeoffs in Signal Processing Hardware Primitives", *Proc. IEEE Workshop on VLSI Signal Processing,* October 1994, pp. 276-285.

[Nagendra96] C. Nagendra, R.M. Owens, and M.J. Irwin, "Design Tradeoffs in High Speed Multipliers and FIR Filters", *Proc. Ninth Intl. Conference on VLSI Design*, Jan. 1996, pp. 29-32.

[Najm95] F.N.Najm, "Power Estimation Techniques for Integrated Circuits", *Proc. IEEE/ACM Intl. Conference on Computer Aided Design,* Nov. 1995, pp. 492-499.

[Nakagome93] Y. Nakagome, K. Itoh, M. Isoda, K. Takeuchi, and M. Aoki, “Sub 1-V Swing Internal Bus Architecture for Future Low Power ULSIs”, *IEEE J. Solid-State Circuits,* Vol. 28, April 1993, pp. 414-419.

[[Ng96] P. Ng, P.T. Balsara, and D. Steiss, "Performance of CMOS Differential Circuits", *IEEE J. Solid-State Circuits,* Vol. 31, June 1996, pp. 841-846.

[Ng97] H.T. Ng and D.J. Allstot, "CMOS Current Steering Logic for Low-voltage Mixed-signal ICs", *IEEE Trans. on VLSI Systems*, Vol. 5, September 1997, pp. 301-308.

[Nicol97] C.J. Nicol, Bell Laboratories, Holmdel, NJ, *Private communications,* January 1997.

[Param96] A. Parameshwar, H. Hara, and T. Sakurai, "A Swing Restored Pass-transistor Logic MAC for Multimedia Applications", *IEEE J. Solid-State Circuits,* Vol. 31, June 1996, pp. 804-809.

[Pentium97] M.R. Choudhury and J.S. Miller, "A 300MHz CMOS Microprocessor with Multi-Media Technology", Digest of technical papers, *IEEE Intl. Solid State Circuits Conference*, February 1997, pp. 170-171.

[Pursley97] D.J. Pursley, “A Gate-Level Simulator for Power Consumption Analysis", *M.S. thesis*, Carnegie Mellon University, Pittsburgh, PA, 1997.

[Rutenbar97] R.A. Rutenbar, "Physical Design: Design of Low Energy/Operation Digital Logic Circuits", *DARPA Review Meeting,* Dept. of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, August 1997.

[Sakurai90] T. Sakurai and A.R. Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas", *IEEE J. Solid-State Circuits*, April 1990, pp. 584-594.

[Sakurai91] T. Sakurai and A.R. Newton, "Delay Analysis of Series Connected MOSFET Circuits", *IEEE J. Solid-State Circuits*, February 1991, pp. 122-131.

[Sakurai97] T. Sakurai, H. Kawaguchi, and T. Kuroda, "Low-power CMOS Design Through Vt Control and Low-Swing Circuits", Digest of technical papers, *IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1997, pp. 1-6.

[Sapatnekar93] S.S. Sapatnekar, V.B. Rao, P.M. Vaidya, and S.M. Kang, "An Exact Solution to the Transistor Sizing Problem for CMOS Circuits using Convex Optimization", *IEEE Trans. on Computer Aided Design of IC's*, Vol. 12, November 1993, pp. 1621-1634.

[Sasaki96] H. Sasaki, "Multimedia Complex on a Chip", Digest of technical papers, *IEEE Intl. Solid State Circuits Conference*, February 1996, pp. 16-19.

[Shepard96] K. Shepard and V. Narayanan, "Noise in Deep Submicron Digital Design", *Proc. IEEE/ACM Intl. Conference on Computer Aided Design*, November 1996, pp. 524-531.

[Shigematsu95] S. Shigematsu et al, "A 1-V High-speed MTCMOS Circuit Scheme for Power-down Applications", Digest of technical papers, *Symposium on VLSI Circuits*, June 1995, pp. 125-126.

[Shigematsu97] S. Shigematsu et al, "A 1-V High-speed MTCMOS Circuit Scheme for Power-down Applications", *IEEE J. Solid-State Circuits*, Vol. 32, June 1997, pp. 861-869.

[Shin89] H.J. Shin et al, "A 250 Mbits/s CMOS Crosspoint Switch", *IEEE J. Solid-State Circuits*, Vol. 24, April 1989, pp. 478-486.

[Soma97] D. Somasekhar and K. Roy, "LVDCSL: Low Voltage Differential Current Switch Logic", Digest of technical papers, *IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1997, pp. 18-23.

[Stanisic97] B. Stanisic, IBM Corp., Rochester, MN, *Private communications*, January 1997.

[Stanisic93] B. Stanisic, “Automatic Analog Power Distribution Synthesis in RAIL”, *PhD thesis*, Carnegie Mellon University, Pittsburgh, PA, 1993.

[Strojwas96] A.J. Strojwas et al., "Manufacturability of Low Power CMOS Technology Solutions", *Proc. IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1996, pp. 225-232.

[Su93] D.K. Su et al, "Experimental Results and Modeling Techniques for Substrate Noise in Mixed-Signal Integrated Circuits", *IEEE J. Solid-State Circuits*, Vol. 28, April 1993, pp. 420-430.

[Sun94] S.W. Sun and P.G.Y. Tsui, "Limitation of CMOS Supply Voltage Scaling by MOSFET Threshold Voltage Variation", *Proc. IEEE Custom Integrated Circuits Conference*, May 1994, pp. 267-270

[Suzuki93] M. Suzuki et al, "A 1.5ns 32b CMOS ALU in Double Pass-transistor Logic", Digest of technical papers, *IEEE Intl. Solid State Circuits Conference*, February 1993, pp. 90-91.

[Sze83] S.M. Sze, *VLSI Technology*, McGraw Hill, 1983.

[Tang96] X. Tang, V.K. De, and J.D. Meindl, "Effects of Random MOSFET Parameter Fluctuations on Total Power Consumption", Digest of technical papers, *IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1996, pp. 233-236.

[Twaijry94] H.A.Twaijry and M.J.Flynn, "Multipliers and Datapaths", *Technical Report CSL-TR-94-654*, Stanford University, CA, Dec. 1994.

[Twaijry96] H.A.Twaijry and M.J.Flynn, "Optimal Placement and Routing of Multiplier Partial Product Trees", *Technical Report CSL-TR-96-706*, Stanford University, CA, September 1996.

[Ultrasparc95] A. Chamas et al, "A 64b Microprocessor with Multimedia Support", Digest of technical papers, *IEEE Intl. Solid State Circuits Conference*, February 1995, pp. 178-179.

[Usami97] K. Usami et al, "Automated Low-Power Technique Exploiting Multiple Supply Voltages Applied to a Media Processor", *Proc. IEEE Custom Integrated Circuits Conference*, May 1997, pp. 131-134.

[Varhol97] P. Varhol, "Mainstream Processors gain DSP Features", *Portable Design*, September 1997, pp. 29-32.

[Wailee94] Wai-Lee, U. Ko, and P.T. Balsara, "A Comparative Study on CMOS Digital Circuit Families for Low-Power Applications", Digest of technical papers, *IEEE Intl. Workshop on Low Power Design*, August 1994, pp. 129-132.

[Wailee97a] Wai-Lee et al, "A 1V DSP for Wireless Communications", Digest of technical papers, *IEEE Intl. Solid State Circuits Conference*, February 1997, pp. 92-93.

[Wailee97b] Wai-Lee, "Low-Voltage Programmable DSP Processor Design", Tutorial, *IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1997.

[Wallace64] C.S. Wallace, "A Suggestion for a Fast Multiplier", *IEEE Trans. on Electron. Comp.*, Vol. 13, February 1964, pp. 14-17.

[Yan95] R.H. Yan et al, "Reducing Operating Voltages from 3,2, to 1V and Below: Challenges and Guidelines for Possible Solutions", Digest of tech. papers, *IEEE Intl. Electron Devices Meeting*, December 1995, pp. 3.1.1-3.1.4.

[Yang95] I.Y. Yang et al, "Back-gated CMOS on SOIAS for Dynamic Threshold Voltage Control", Digest of tech. papers, *IEEE Intl. Electron Devices Meeting*, December 1995, pp. 35.1.1-35.1.4.

[Yano90] K. Yano et al, "A 3.8ns CMOS 16*16 Multiplier Using Complimentary Pass-transistor Logic", *IEEE J. Solid-State Circuits*, Vol. 25, April 1990, pp. 388-395.

[Yano96] K. Yano et al, "Top-Down Pass-transistor Logic Design", *IEEE J. Solid-State Circuits*, Vol. 31, June 1996, pp. 792-803.

[Ye97] Y. Ye, K. Roy, and G. Stamoulis, "Quasi-Static Energy Recovery Logic and Supply Clock Generation Circuits", Digest of technical papers, *IEEE/ACM Intl. Symposium on Low Power Electronics and Design*, August 1997, pp. 96-103.

[Zimmer97] R. Zimmermann and W. Fichtner, "Low-power Logic Styles: CMOS Vs. Pass-Transistor Logic", *IEEE J. Solid-State Circuits*, Vol. 32, July 1997, pp. 1079-1090.