# Design of High-Performance, Robust Datapaths with Delay Diagnostics for Scaled CMOS Technologies

by

Bhaskar P. Chatterjee

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2004

I hereby declare that I am the sole author of this thesis.

I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Bhaskar P. Chatterjee

I further authorize the University of Waterloo to reproduce this thesis by photocopying or other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Bhaskar P. Chatterjee

# Abstract

Over the past 30 years aggressive technology scaling and innovative design techniques have led to the design of high-performance microprocessors that operate at on-chip clock frequencies of more than 3GHz and have 100 million or more transistors. The projections from ITRS 2003 indicate that this trend will continue into the next decade resulting in the integration of over a billion transistors and on-chip clock frequency exceeding 10GHz by the year 2010. However, such aggressive technology scaling is not without its challenges. Some of the most important problems faced by high-performance logic design and test engineers are related to the high power demand, ensuring adequate noise margin and testability. In this thesis we address some of these issues in the context of bulk CMOS based logic and datapath designs.

During the course of this research work, a 32-bit, high performance ALU was designed with circuit level design modifications to ensure its low power operation. In particular, the critical and non-critical units of the ALU were identified and a dual supply design scheme was adopted in-order to minimize both switching and leakage power consumption during the active and standby modes of operation. In addition, a latch (flip-flop) scheme was developed that can support a reduced swing clocking scheme and interface signals between the different power supply domains without consuming additional static power. We also used a swing-restored CPL (SRCPL) based design approach for the non-critical logic and shifter units to lower the overall capacitance and data buffer sizes to reduce overall power (energy). Our results indicate that by using this strategy, it is possible to reduce the operating power by up to 24%.

As the technology is scaled, the transistor leakage current increases exponentially and causes noise margin degradation in digital circuits. Wide-OR domino logic circuits are used extensively in the design of ALU front-ends and register file (RFs). Such circuits are known to be especially susceptible to leakage induced logic upsets in scaled CMOS technologies. In this work we investigated several different circuit level schemes that have already been proposed and compared their effectiveness in improving circuit robustness. In particular, we considered schemes such as reverse body bias, channel length modulation, pseudo-static techniques, conditional keepers and forward body bias. We also proposed two circuit techniques that can be used to improve the wide-domino performance while maintaining iso-robustness scaling and reducing total energy consumption.

Finally in this research, we developed a design-for-testability (DFT) scheme for detecting delay faults in high performance datapaths. As technology is scaled, at-speed testing is becoming more difficult, while parametric faults are becoming more common. This is leading to both yield loss and long-term reliability problems. In addition, automatic test equipments (ATEs) are unable to keep up with the on-chip clock frequencies and the number of transistors/pin is increasing, making fault diagnostics a major challenge. The proposed DFT scheme enables us to detect delay faults with up to 60ps resolution and diagnose internal logic sub-units that cause such failures. In addition, our scheme allows for up to 5x lower test-mode clock frequency and converts the hard-to-detect delay faults into stuck-at faults at the primary outputs of the circuit under test (CUT).

The above design concepts were used to design a 32-bit, 180nm, 1.5GHz ALU with about 11.5k transistors. The semi-custom design measured $800\mu m$ x $600\mu m$ and the dif-

ferent design tradeoffs were quantified for the 180nm generation while the scaling trends for the 65nm technology were studied using the Berkeley Predictive Technology Models (BPTM). It is expected that this will help in the design of low power and reliable datapaths for scaled CMOS technologies.

**Acknowledgements**

I would like to take this opportunity to express my deep sense of gratitude and thanks to my supervisor Professor Manoj Sachdev. Without his constant help, suggestions and insights this work would not have been possible. What helped me most was his willingness to believe in a new idea and seldom, if ever at all, limiting the scope of my work.

I would also like to express my thanks and appreciation for the comments and inputs from all my PhD committee members. Their suggestions added immense value to my thesis and work. Also thanks to my mentors at Intel, Ram Krishnamurthy and Ali Keshavarzi whose suggestions and collaboration helped shape the course of this work. A special thanks to CMC and Phil Regier for all the tool support, equipments and technology access.

I would like to thank each member of our group and all my lab-mates. For all the jokes, late night chats, and afternoon coffee breaks! It was great getting to know all of you. I would especially like to thank Christine and Shahrzad for the research collaboration and their effort. These have been some of the most rewarding years of my student life and I am happy to know all my friends at the University of Waterloo.

I would like to thank all my friends and family members. I am especially indebted to the Hill and Dasgupta families during my stay here at Waterloo. Finally to my parents for all their kindness and love and my brother for always being there for me during trying times. Without their support this would not have been possible.

# Contents

**1 Introduction** **1**

1.1 VLSI Scaling and ITRS Projections . . . . . 2

1.2 Challenges for CMOS Scaling . . . . . 5

1.3 Reliability, Testing and Yield Issues . . . . . 11

1.4 Issues Addressed in this Work . . . . . 14

1.5 Summary and Thesis Outline . . . . . 17

**2 Building Blocks for High Performance Datapath Designs** **19**

2.1 Digital Logic Design and Circuit Families . . . . . 21

2.2 Footerless CDL and SRCPL Based Logic Design . . . . . 24

2.3 Pipelined Datapath Designs: Basic Concept . . . . . 28

2.4 Latches and Flip-flops for Pipelined Datapaths . . . . . 30

2.5 Multiplexors for High Performance Designs . . . . . 32

2.6 Summary . . . . . 35

**3 High Performance ALU Design and Low Power Operation** **36**

3.1 Different Power Components in CMOS Logic . . . . . 37
3.1.1 Switching Power Component . . . . . 38
3.1.2 Leakage Power Component . . . . . 40
3.1.3 Short Circuit Power Component . . . . . 41
3.2 Supply Scaling and MOSFET Current Components . . . . . 43
3.2.1 MOSFET OFF-Current Model . . . . . 45
3.2.2 MOSFET Gate Leakage Current . . . . . 47
3.3 Circuit Techniques for Low Power ALU Operation . . . . . 49
3.3.1 Latch and Flip-Flop Design for Dual Supply . . . . . 50
3.3.2 Swing Restored CPL Based Logic Unit . . . . . 54
3.4 ALU: Architecture and Circuit Design . . . . . 57
3.4.1 Design of Decoder and Logic-Shift Units . . . . . 59
3.4.2 ALU Critical Path and Adder Design . . . . . 60
3.4.3 Dual Supply Design and Assignment . . . . . 62
3.5 ALU Energy, Delay, Scaling Trends: Results . . . . . 64
3.5.1 ALU Performance for Sub-180nm Technologies . . . . . 65
3.5.2 ALU Energy for Sub-180nm Technologies . . . . . 65
3.5.3 ALU Standby Power and Current Demands . . . . . 67
3.6 Summary . . . . . 70

**4 Designing Robust Wide-Domino Logic for High Performance Datapaths 71**

4.1 RF Organization: A Simple Example . . . . . 73
4.2 RF and Wide-OR Domino MUX-es, Iso-Robustness Scaling . . . . . 75

4.2.1 RF Read Port and LBL, GBL Designs . . . . . 76
4.2.2 LBL and GBL: Design and Noise Margin Issues . . . . . 77
4.2.3 Robust Wide-Domino Schemes . . . . . 81
4.3 Leakage Control Schemes and RF Designs . . . . . 85
4.3.1 Dual Threshold LBL Design for 90nm Technology . . . . . 86
4.3.2 Transistor Level Leakage Control: Analysis . . . . . 89
4.4 Low Power, 65nm Wide-Domino Operation . . . . . 92
4.5 Precharge Current Control and 16-Wide Domino . . . . . 96
4.5.1 Precharge Current Control for RFs . . . . . 97
4.5.2 Designing Low Power 16-Wide Robust Dominos . . . . . 101
4.6 Implementation Issues . . . . . 104
4.7 Summary . . . . . 107

**5 Delay Fault Testability and Diagnostics for High Performance Datapaths 109**

5.1 High Performance Circuit Testing: Background . . . . . 110
5.2 Circuit Strategy for DSM Digital Testing . . . . . 112
5.2.1 DFT for Delay Testing in CDL gates . . . . . 113
5.2.2 Delay Fault Detection Range . . . . . 117
5.3 Design Overview of DFT Based 32-bit ALU . . . . . 119
5.3.1 Delay Testing Logic: Implementation . . . . . 123
5.3.2 Delay Testable ALU: Energy-Delay Tradeoffs . . . . . 129
5.4 ALU TEST Mode Operation . . . . . 130
5.4.1 DFT Implementation Issues . . . . . 134

5.5 Summary . . . . . 136

**6 Conclusion** **137**

6.1 Low Power ALU Design . . . . . 137

6.2 Robust Domino Designs for Datapath Circuits . . . . . 139

6.3 DFT Technique for High Performance Datapaths . . . . . 140

6.4 Future Work . . . . . 141

**7 Bibliography** **144**

**A Published Papers** **153**

**B Patents** **156**

# List of Tables


1.1 CMOS scaling: CVS and CFS basic trends . . . . . 5

3.1 Static CMOS vs. SRCPL bit-slice performance (ps) . . . . . 56

4.1 FBB characteristics for 65nm p-MOS transistors . . . . . 83

4.2 Data showing current ratio degradation . . . . . 91

4.3 Energy-delay data for low power 65nm design . . . . . 101

4.4 Low power RF designs, 65nm results . . . . . 103

5.1 Truth table for DFT logic and mode selection . . . . . 115

5.2 Defect resistance detection for 180nm technology . . . . . 120

5.3 Defect detection range for ALU DFT vs. non-DFT . . . . . 135

# List of Figures


1.1 Technology scaling basics . . . . . 4

1.2 ITRS 2003 technology and performance projections . . . . . 6

1.3 ITRS 2003 microprocessor power, current and levels of metallization . . . . 8

1.4 Pentium processor hotspot simulations: Source Intel Corp. . . . . 10

1.5 ITRS 2003 testing challenges and test cost . . . . . 12

2.1 Generic block diagram of a digital processor . . . . . 20

2.2 CMOS circuit families: static, pass transistor and dynamic logic . . . . . 22

2.3 Multi-stage footerless compound domino logic . . . . . 25

2.4 CPL vs. SRCPL comparisons . . . . . 27

2.5 Datapath pipelining: basic concepts . . . . . 29

2.6 Latch and filp-flop circuit schemes . . . . . 31

2.7 Multi-stage footerless compound domino logic . . . . . 34

3.1 Inverter during switching transient . . . . . 39

3.2 CMOS inverter leakage current components . . . . . 40

3.3 CMOS inverter short circuit current during switching transient . . . . . 42

3.4 Transistor OFF-state current scaling for sub-180nm CMOS technologies . . 45
3.5 Impact of supply scaling on transistor current . . . . . 49
3.6 TG latch with steady state current problem under dual supply clocking . . 51
3.7 Latch circuitry to support dual supply clocking . . . . . 52
3.8 Dual supply latch and flip-flop 180nm layouts . . . . . 52
3.9 Energy and delay plots for dual supply latch . . . . . 53
3.10 ALU bit-slice designed using SRCPL . . . . . 55
3.11 SRCPL based 180nm layout of ALU bitslice . . . . . 56
3.12 32-bit ALU organization block diagram . . . . . 58
3.13 Simplified architectural overview of adder . . . . . 62
3.14 32-bit ALU layout in 180nm CMOS technology . . . . . 64
3.15 32-bit ALU NORMAL mode performance and scaling trends . . . . . 66
3.16 32-bit ALU NORMAL mode total energy and scaling trends . . . . . 67
3.17 ALU standby mode power consumption . . . . . 68
3.18 ALU peak and average current demands . . . . . 69

4.1 Organization of 256-array, 64-bit RF . . . . . 74
4.2 Critical read path of a 64-array RF . . . . . 77
4.3 Waveforms for Wide-OR domino DC robustness . . . . . 79
4.4 Wide OR domino noise margin and scaling trends . . . . . 80
4.5 Organization of conditional keeper based wide-OR dominos . . . . . 84
4.6 Conditional keeper timing . . . . . 85
4.7 AC noise margin degradation of conditional keepers . . . . . 86

4.8 Pseudo-static technique based wide-OR domino . . . . . 87
4.9 Dual threshold based 90nm LBL organization . . . . . 88
4.10 Delay and noise margin comparisons for 90nm RFs . . . . . 89
4.11 Wide-domino organization with parasitic capacitances . . . . . 93
4.12 Leakage control schemes and their effectiveness . . . . . 95
4.13 RF read port energy-delay comparisons (65nm) . . . . . 96
4.14 RF read port precharge timing diagram . . . . . 98
4.15 RF with CLKB precharge transistors . . . . . 99
4.16 RF precharge current and voltage waveforms . . . . . 100
4.17 Cross-section for RBB and non-RBB n-MOS transistors . . . . . 106
4.18 LBL organization without RBB or precharge control . . . . . 106
4.19 16-wide LBL with RBB and precharge control . . . . . 107

5.1 CDL gates with DFT for delay testing . . . . . 114
5.2 Low frequency delay testing with DFT . . . . . 117
5.3 Resistive defects: typical location in CUT . . . . . 119
5.4 Defect resistance detection range DFT vs. non-DFT . . . . . 120
5.5 32-bit delay fault testable ALU architecture . . . . . 122
5.6 DFT logic for a delay fault testable ALU . . . . . 126
5.7 Alternate schemes for TEST mode clock generation . . . . . 127
5.8 DFT logic delay control using bias voltage (Scheme 2) . . . . . 128
5.9 DFT technique: delay impact, scaling trends . . . . . 129
5.10 TEST mode clock signals for ALU during delay testing . . . . . 131

5.11 Timing diagram showing ALU delay margins . . . . . 134

# Chapter 1

# Introduction

The past 30 years have seen unprecedented advancements in semiconductor technology and the emergence of the microelectronics industry. This has been made possible by several factors which include the discovery of the metal-oxide semiconductor field effect transistors (MOSFETs), the development of the complementary metal-oxide semiconductor (CMOS) process, aggressive technology scaling and improvements in the semiconductor manufacturing process. This has led to the integration of more and more transistors that offer more functionality, improved performance and enabled the design and development of modern high performance microprocessors. Higher functionality, larger die-sizes, lower cost/function have made the modern very large scale integration (VLSI) industry feasible and sustainable. However, as the semiconductor industry continues to scale to the deep submicron (DSM) regime and transistor feature size reaches below 180nm, designers and technologists are faced with several challenges. These stem from some of the fundamental parameters of the semiconductor material and MOSFET transistors that do not scale with

technology, resulting in non-linear shifts in transistor characteristics. This trend is making the scaling and integration of future CMOS technologies more difficult.

## 1.1 VLSI Scaling and ITRS Projections

Technology scaling proposed by [1] has been the main stay of the VLSI industry. This is based on shrinking of both the vertical and lateral transistor dimensions resulting in more efficient operation of the scaled designs. A direct consequence of technology scaling is the now famous Moore's Law [2] which predicts a doubling of performance of integrated circuits (IC) every 18 months. This has led to an exponential increase in circuit performance and throughput, with more functionality while reducing the cost/function for the scaled ICs. In this section we discuss the fundamentals of CMOS scaling and show some of the future projections from the International Technology Roadmap for Semiconductors (ITRS 2003) [3].

The basic idea of CMOS scaling is best explained with the help of Figure 1.1 which shows two transistors: one that has nominal features (Figure 1.1(a)), and the other that represents its scaled version (Figure 1.1(b)). It is clear that several transistor parameters need to be scaled in proportion: transistor channel length ($L$), thin-oxide ($T_{ox}$) and substrate doping concentration ($N_A$). The actual dimensions shown in Figure 1.1 are taken from the original paper that proposed CMOS scaling and correspond to a scaling factor of 5 (S=5). Traditionally, there have been two approaches for technology scaling [4]:

- Constant Voltage Scaling (CVS), and

- Constant Field Scaling (CFS).

The constant voltage scaling approach is where the transistor physical dimensions ($L$, $T_{ox}$), are scaled while the electrical parameters such as supply and threshold voltage are left unchanged. However, as the transistor's gate oxide becomes thinner in DSM technologies, the electric field stress starts to increase. This can cause thin-oxide breakdown and adversely affect the long-term reliability of CMOS circuits. Therefore in current circuits and VLSI implementations designers have moved away from the traditional constant voltage scaling technique (CVS) and use constant field scaling (CFS) instead. In this strategy, when the technology is scaled from one generation to the next, both the supply and threshold voltages are lowered in proportion along with the transistor's physical dimensions. This helps in maintaining the thin-oxide stress within acceptable limits and short-channel effects (SCE) under control.

The benefits of CMOS scaling result from the reductions in transistor parasitic capacitance ($C_{ox}$) associated with smaller device geometries, lower gate level average power ($P_{av}$), switching energy ($E_{sw}$) and improved propagation delay ($t_{delay}$). Table 1.1 shows the impact of technology scaling (CVS, CFS) on some of the transistor electrical parameters and important design metrics. Traditionally, a scaling factor of 0.7 has been used to shrink the feature size from one CMOS technology to the next. Based on the expressions shown in Table 1.1, it is clear that the capacitance ($C_{ox}$) reduces by 30% due to lower overall transistor area ($WL_{drawn}$). In addition, for the CFS approach, the inverter propagation delay reduces by 30%, while the average power ($P_{av}$) reduces by 50% and the switching energy ($E_{sw}$) by 65% for every technology generation. The power and energy savings ob-

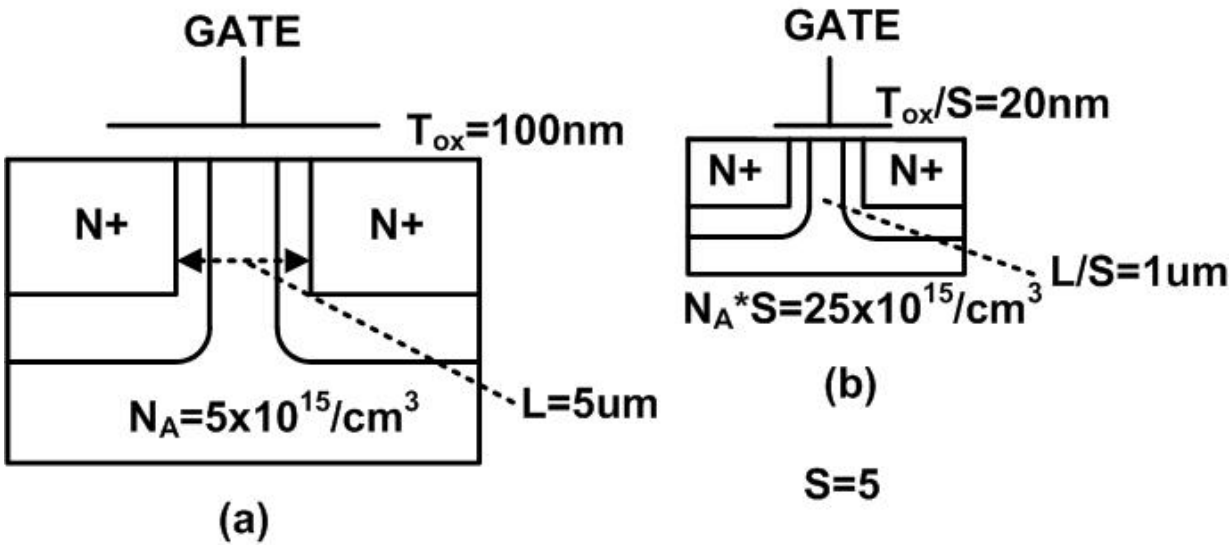


Figure 1.1: Technology scaling basics

tained from CMOS scaling is greater for CFS than for the CVS approach. This is because both the average power and switching energy are supply voltage dependent parameters and therefore scale more in the CFS approach as opposed to CVS where the power supply voltage is held constant.

The energy and delay improvements obtained as a direct consequence of CMOS scaling have led to unprecedented levels of integration and circuit performance. As a result, modern high performance microprocessors have achieved clock frequencies of more than 3 GHz and more than 100 million transistors on-die. It is expected that silicon based bulk CMOS technology will continue to be the mainstay of the microelectronics industry for the next decade [3]. The projected improvements in technology scaling and high-end microprocessor performance are indicated in Figure 1.2 based on data from the ITRS 2003 report. It is expected that by the year 2018, CMOS technology will reach the 18nm node and that the

Table 1.1: CMOS scaling: CVS and CFS basic trends

| Parameter | Relation | CVS | CFS |
|---|---|---|---|
| $W, L_{drawn}, T_{ox}$ | | $\frac{1}{S}$ | $\frac{1}{S}$ |
| $V_{DD}, V_{TH}$ | | 1 | $\frac{1}{S}$ |
| Area | $WL_{drawn}$ | $\frac{1}{S^2}$ | $\frac{1}{S^2}$ |
| Capacitance | $C_{ox}WL_{drawn}$ | $\frac{1}{S}$ | $\frac{1}{S}$ |
| Delay | $\frac{C_L V_{dd}}{I_{av}}$ | $\frac{1}{S^2}$ | $\frac{1}{S}$ |
| $P_{av}$ | $\frac{C_L V_{DD}^2}{t_{delay}}$ | $S$ | $\frac{1}{S^2}$ |
| Energy | $C_L V_{DD}^2$ | $\frac{1}{S}$ | $\frac{1}{S^3}$ |

on-die transistor count (including cache) will cross 1 billion by the year 2012, and reach up to 4.9 billion by 2018. In addition, the on-chip local clock frequency will surpass 10 GHz by 2008, and may equal 53 GHz by 2018.

These unprecedented levels of integration will lead to major design and testing challenges some of which will be addressed in this work.

## 1.2 Challenges for CMOS Scaling

Even though CMOS scaling has its advantages, it also poses certain technological and design challenges. Some of these problems can be listed as follows:

- transistor threshold voltage ($V_{TH}$) scaling,
- increase in subthreshold ($I_{OFF}$) and gate-leakage currents ($I_{GATE}$),

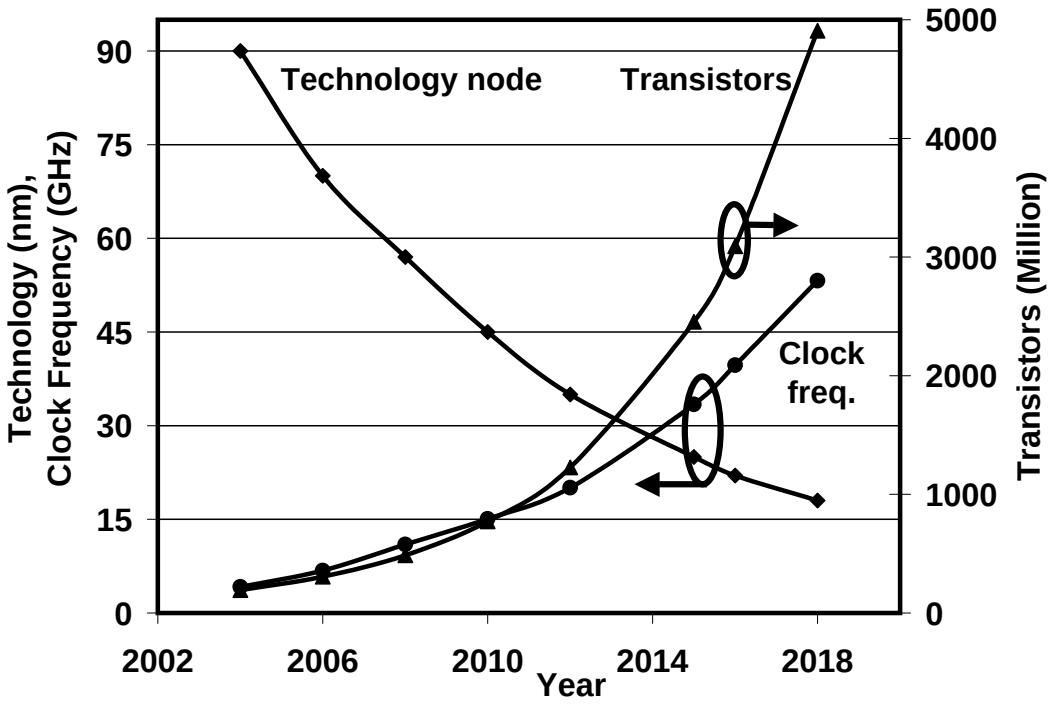


Figure 1.2: ITRS 2003 technology and performance projections

- increase in current density and total power demands,
- impact of interconnect delays on overall performance,
- thermal issues and impact of process variation, and
- degradation of transistor performance and reliability due to worsening short-channel effects (SCE).

These challenges in turn adversely affect the VLSI performance, manufacturability, yield and long-term reliability. Under the CFS regime, as the supply voltage is scaled with technology, the transistor threshold voltage needs to be reduced in proportion to maintain the gate-overdrive voltage ($V_{OV} = V_{DD} - V_{TH}$) and circuit performance. However, it is well known that this leads to an exponential increase in the transistor sub-threshold current [4], [5] . Transistor sub-threshold leakage is negligible compared to the switching current up to the 250nm technology. However, as the transistor $V_{TH}$ is aggressively scaled and more transistors are integrated on-die, the leakage component becomes more important. This can offset the reductions in switching energy obtained from technology scaling and lead to a higher system level total current and power. The increase in IC power is one of the most significant problems faced by designers that leads to many additional challenges. For example, higher system level power dissipation is associated with increases in both the average ($I_{av}$) and peak current ($I_{pk}$) demands. This in turn makes power delivery and the design of the power supply more challenging for scaled technologies. Some of the future trends for microprocessor current and power demands are shown in Figure 1.3. The $P_{av}$ for high-end microprocessors has already reached approximately 150W, and may exceed 300W by the year 2018 [3]. This increase in power will be largely due to increased leakage current, higher on-die clock frequency and higher levels of integration. It is expected that the power supply voltage will reach about 0.7V by 2018, resulting in an increase in $I_{av}$ from the present levels of about 130A to more than 400A by 2018. Such high currents can lead to excessive IR voltage drops and cause performance degradation as well as reduce the long term reliability of high end ICs.

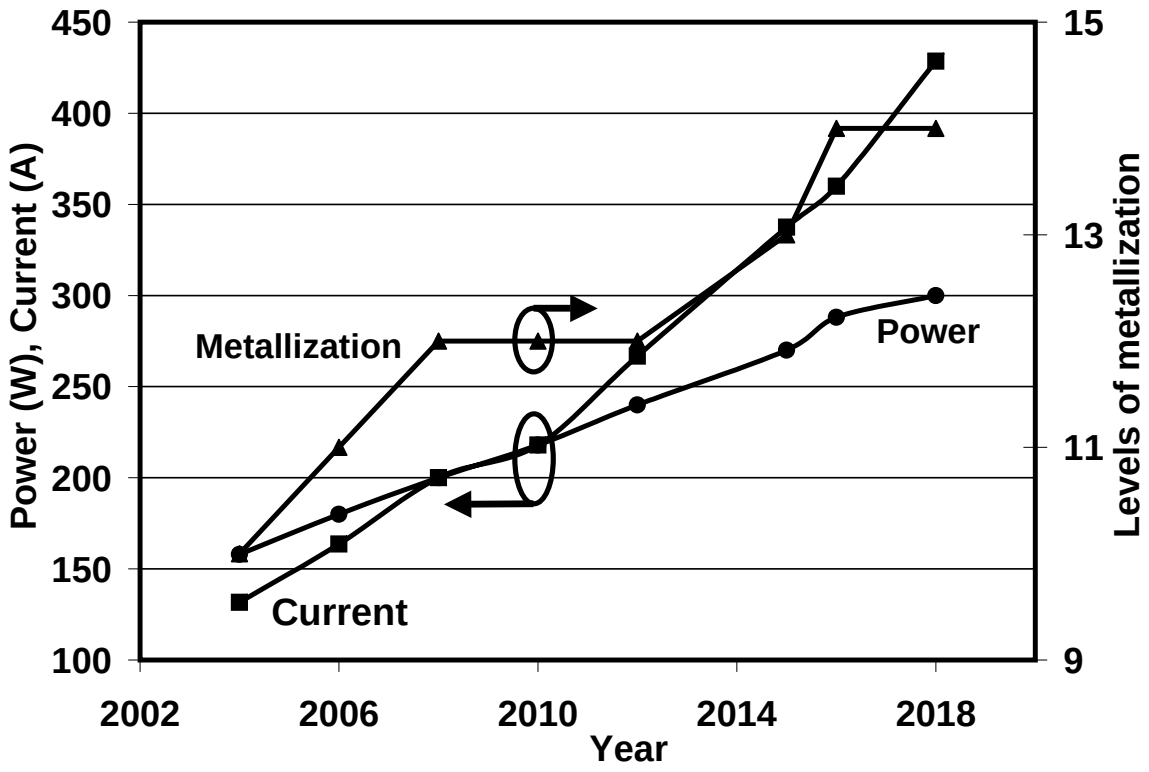


Figure 1.3: ITRS 2003 microprocessor power, current and levels of metallization

The increase in IC total power leads to thermal and reliability problems in DSM technologies. As the average and peak current demands increase, both the device and interconnect geometries are also scaled. This leads to higher current density and as a result, electro-migration related IC failures are becoming more common in high performance ICs [3], [6]. Another source of concern is the thermal issue and local hot-spots in high performance microprocessors. Large ICs typically have a relatively small portion of the die that is performance critical, which determines the system clock frequency and IC throughput.

These units operate at high switching frequency and switching activity thereby contributing significantly to the system power. Since a large portion of the IC total power is dissipated in a relatively small portion of the die, this results in high power density and localized thermal hot-spots. This is shown in Figure 1.4 which shows the simulation results for the on-die thermal map for a portion of the Pentium microprocessor. As shown in the figure, the performance critical address generation unit (AGU) and execution core have elevated temperatures of about $110^0C$. It is apparent that the hot-spot locations are concentrated in a relatively small area of the IC, while the rest of the die occupied by the cache operates at a relatively lower temperature of about $70^0C$. This problem is further compounded by the interdependence of transistor leakage current and operating temperature. As the on-die temperature rises, it leads to lower transistor threshold voltage ($V_{TH}$) and exponentially higher subthreshold current ($I_{OFF}$) [4], [5], [7]. In fact, in extreme cases this can lead to thermal run-away and destruction of the IC. This can occur during burn-in test during which an IC is subjected to both thermal and voltage stresses. Burn-in has been a well established test technique that is used to speed-up the failure of "weak" ICs and improve the long-term reliability of ICs shipped by VLSI vendors. However, the possibility of thermal hot-spots and runaway in scaled CMOS technologies is eroding the effectiveness of test techniques like burn-in and adversely impacting IC reliability. In addition to the thermal problem, high leakage currents also cause noise margin degradation in digital circuits. This problem is further compounded by the low power supply voltage in scaled CMOS technologies. This is especially true in the case of sub-90nm dynamic circuits and may pose a major challenge to high performance logic designers [8].

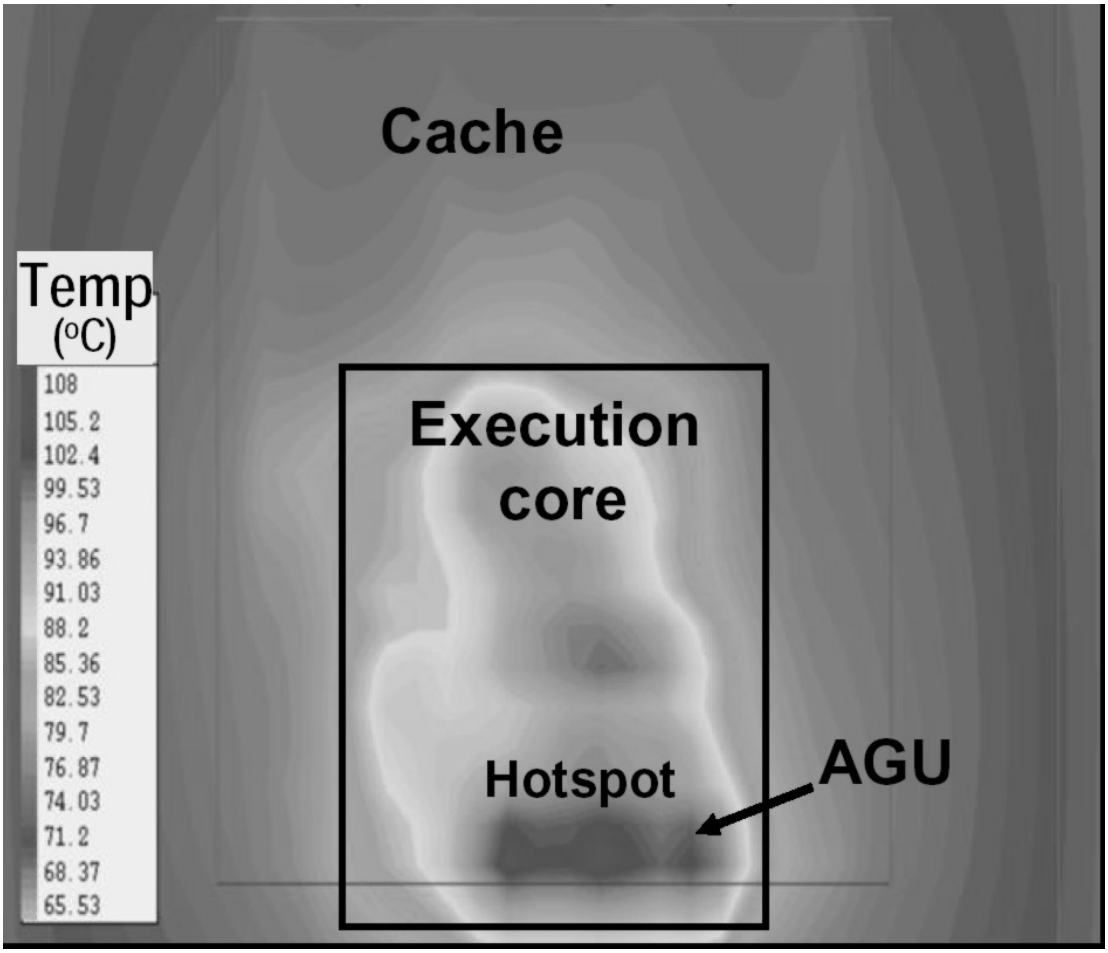


Figure 1.4: Pentium processor hotspot simulations: Source Intel Corp.

Process variation is also a major concern for sub-250nm CMOS technologies. As photolithography is stretched to its limits by CMOS scaling, there is increasing mismatch between on-die and die-die transistor parameters. This is further compounded by the increase in die size, number of mask levels and the introduction of dual $V_{TH}$ technologies. In particular, the variations in transistor drawn channel length ($L_{drawn}$), gate-oxide thickness ($T_{ox}$) and channel doping concentration can be significant ($N_{ch}$). These in turn lead to

variations in transistor threshold voltage ($V_{TH}$), saturation current ($I_{DSAT}$) and off-state current ($I_{OFF}$). In addition to variations in transistor parameters, the interconnect characteristics are also affected by process variation. The fluctuations in interconnect width and resistivity impact the overall IC performance. As more levels of metallization are used along with thicker inter-layer dielectrics (ILD) and thinner metal lines, resistive interconnects, vias, and contacts are becoming more common. As a result, sub-250nm ICs show delay spread and variability in both IC power consumption and circuit noise margins. Some of these process and manufacturing imperfections are spatial in nature (on die) while others are temporal (variations across time, die-to-die, batch-to-batch). Therefore, the impact of these variations on IC performance is becoming more difficult to model and track. This can cause logic failures and rejection of good ICs and is of special concern in scaled CMOS technologies.

## 1.3 Reliability, Testing and Yield Issues

The technological challenges enumerated earlier are posing significant problems for both circuit designers and the test community. Traditionally logic design and test strategies have evolved independently of one another. However, in DSM technologies both design and test are becoming more interdependent. For example, design choices that reduce system power may improve the effectiveness of test techniques like burn-in and $I_{DDQ}$. Also, test strategies that are well planned and integrated early into the design cycle can help improve IC yield. In the previous sections we discussed some of the design and power related challenges of scaled technologies. We now some of other issues related to IC testing, yield and reliability.

Testing of deep sub-micron ICs is becoming a major challenge primarily because of two reasons:

- effectiveness of existing test and IC screening techniques is being eroded by higher leakage currents and operating frequency ($f_{sw}$), and
- ICs are becoming more complex with more functionality and increasing transistors/pin.

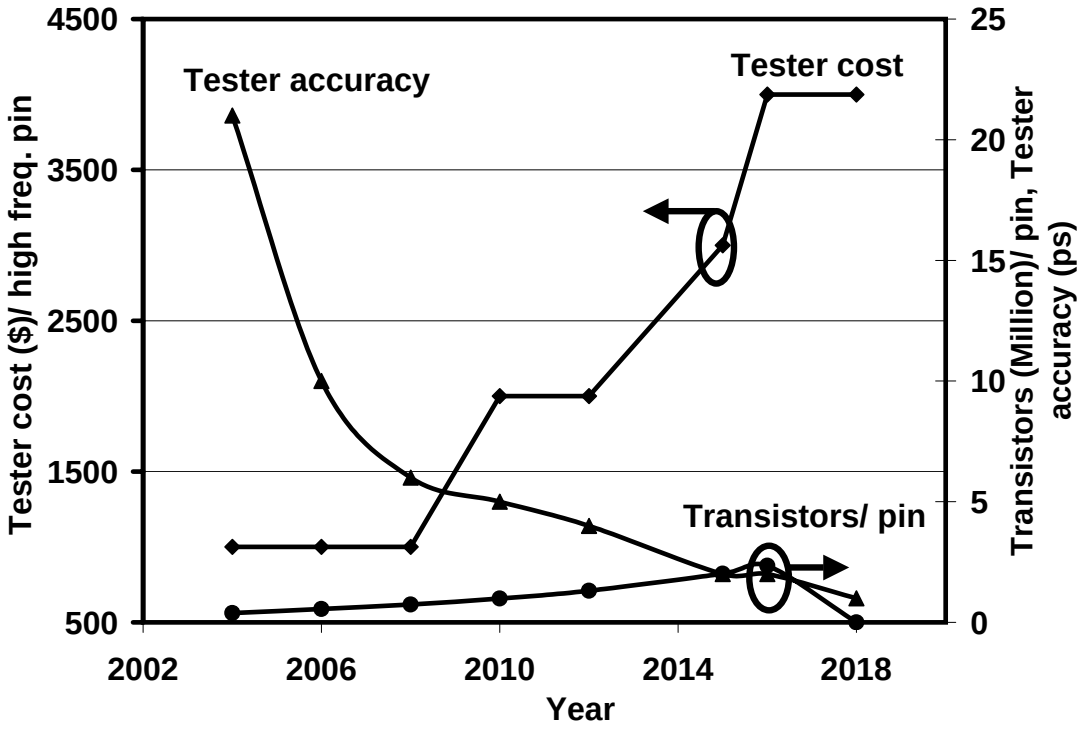


Figure 1.5: ITRS 2003 testing challenges and test cost

The increase in both the total IC current and power in scaled CMOS technologies is eroding the effectiveness of current based test techniques like burn-in and $I_{DDQ}$. As per ITRS 2003 projections the $I_{DDQ}$ quiescent current is expected to increase from the present level of 0.4mA to 20A by 2014. In addition, there is also expected to be a significant increase in the total variation in the IC quiescent current. This may render $I_{DDQ}$ less effective and require additional test strategies for screening high-end designs. Also as more transistors are integrated on-die, the IC pin count does not increase in proportion. Therefore the transistors/pin for complex ICs increases with scaling making their testing more difficult. This trend is shown in Figure 1.5 that indicates that the total transistors/pin count will increase from about 386k/pin to 2.3 million/pin by 2016. This is expected to be accompanied by an increase in tester cost as well. The automatic test equipment (ATE) frequency has improved by about 12% as opposed to the 30% increase of the device under test (DUT). This has eroded the speed advantage of the test equipment and is making at-speed testing more difficult. As a result, it is becoming more difficult to test ICs, causing test escapes, yield and revenue losses at the vendor's site. In fact, as indicated by the data in Figure 1.5, the ATE edge placement accuracy will have to within 1ps by the year 2018 in order to test the high performance microprocessors. The need for high performance testers with more pins and deeper memory is also adding to the overall test and debug cost and adding to the overall IC cost. For example, according to projections, it is expected that high-end ATEs may require a total of 2048 pins with each high frequency tester pin costing as much as 4000 dollars.

Therefore a holistic approach has to be adopted for the design and test of high performance ICs. As indicated in the ITRS 2003 report, both design as well as the testing of high-end ICs are becoming increasingly difficult. As the IC power density increases and electric stress across the thin-oxide reaches the breakdown limit the long-term reliability of high-performance circuits and microprocessors is compromised. As a result, there is degradation in the IC mean-time to failure (MTTF) causing in-situ parts to fail sooner than for earlier generations.

## 1.4 Issues Addressed in this Work

In this thesis the focus is on some of the design and test challenges associated with full-custom datapath units and their possible solutions. High performance logic units such as microprocessor arithmetic logic units (ALUs), adders and register files (RFs) operate with single cycle throughput and therefore consume a significant portion of the IC power. At the same time, since these units determine the system operating frequency, their design is timing critical. This is why such units are typically designed using the full-custom approach to optimize their performance for a given technology. This requires considerable design effort but also allows room for the integration of new circuit and architectural ideas. In this research the main focus is on three major areas:

- Low power circuits: reducing operating power of high performance logic units,
- Robust logic operation: ensuring robust, leakage tolerant circuit designs, and

- DFT for delay testing: develop a design for testability (DFT) scheme for delay fault detection and diagnostics.

*Low power circuits*: In this work a 32-bit ALU design with a high performance 32-bit binary unsigned adder unit and a logic-shifter unit is used as our reference design. The IC power consists of two components: switching power and leakage power. As explained earlier, the switching power ($P_{sw}$) and energy ($E_{sw}$) scale with technology whereas the leakage power component increases exponentially. In fact, beyond the 90nm technology it is expected that the IC power may be dominated by the leakage component [9],[10]. In this ALU design we incorporate several different low-power circuit techniques to reduce both the power and energy dissipation. For example, we use a dual supply strategy to lower both the switching and leakage power of the non-critical portions of the ALU. According to this design strategy the overall design is first partitioned into critical and non-critical units with the non-critical sections being allowed to operate from a second lower supply voltage. In addition, a latch and flip-flop circuitry that can support reduced swing clocking is developed and used to interface signals propagating between the different power supply domains. Also, a swing-restored complimentary pass transistor (SRCPL) based logic-shifter unit is employed in the ALU to reduce total switched capacitance and data buffer sizes. The details of this low power design along with the various datapath performance tradeoffs and their scaling trends are discussed in Chapter 3.

*Robust logic operation*: As the transistor off-state ($I_{OFF}$) current increases exponentially with scaling, the noise margin of certain digital circuits is being compromised. This is especially true for sub-130nm circuits implemented using domino logic. In fact, recent

literature shows that excessive leakage can cause leakage induced logic failure in DSM technologies [11], [12]. This is expected to render the use of domino logic in the design of high performance datapath circuits less effective. Especially susceptible to such leakage induced logic upsets are a special class of domino logic known as the wide-OR dominos that are used extensively in the design of high performance multiplexors (MUX-es) in ALUs and RFs. Many different circuit and leakage control strategies have been proposed that can improve the robustness of wide-OR domino circuits. In this research we compare the effectiveness of several different techniques for the 130nm to 65nm CMOS technologies and propose additional techniques that can improve the noise margin while ensuring their low power operation as well. The details of the work related to robust domino design is discussed in Chapter 4.

*DFT for datapaths*: Parametric faults are becoming more common in scaled CMOS technologies. Such defects are difficult to detect and their characteristics often vary over time, temperature and voltage cycles. High performance circuits and datapath units are especially susceptible to parametric faults that cause timing degradation. In this research we develop a design for testability (DFT) strategy that can detect delay faults in high-performance ALUs. We discuss the circuit level design and implementation of this scheme and demonstrate its ability to convert difficult-to-detect timing-failures into stuck at-faults at the ALU primary outputs. We develop a stage-to-stage testing strategy for the ALU using our DFT scheme that allows for built-in delay fault diagnostics. Furthermore, this DFT technique can detect delay faults when the clock frequency is lowered compared to the system clock frequency. This may allow allow for the usage of relatively cheaper testers

during the testing process. The details of this scheme are discussed in Chapter 5.

## 1.5 Summary and Thesis Outline

In this chapter we discussed some of the basics of CMOS scaling and its advantages. We then focussed on some of the problems and challenges associated with technology scaling. One of the main focus areas from a circuit and design perspective is the increase in system power of high performance datapath designs. Another area of concern is the possibility of leakage induced logic failure, noise margin degradation and parametric failures in digital circuits. As will be clear subsequently, in this work we focus on these aspects and design a 32-bit high-performance ALU to ensure low power and robust operation. Furthermore, we incorporate a DFT strategy to detect delay faults and parametric defects. We discuss the impact of these approaches on several design metrics including operating frequency, switching power (energy), leakage power (energy), noise margin, area penalty and present the details of the ALU physical design and its implementation.

The rest of this thesis is organized as follows: in Chapter 2 we discuss the basic concepts of digital logic design and some of the low power techniques proposed in literature. In Chapter 3 we discuss the details of the 32-bit ALU design and show the energy-delay scaling trends associated with our low power strategy. In Chapter 4 we present the concept of domino logic noise margin and its degradation with scaling. We also present several circuit techniques and compare their effectiveness for designing dominos scalable up to the 65nm CMOS technology. In this work all the results pertaining to the 180nm technology correspond to a TSMC process while those for the 130nm-65nm generations are for the

Berkeley Predictive Technology Models (BPTM) [13], [14]. The details of the DFT scheme for delay testability and diagnostics for the entire ALU are presented in Chapter 5. Finally, in Chapter 6 we present the conclusions and discuss possible future work.

# Chapter 2

# Building Blocks for High Performance Datapath Designs

The VLSI industry is being constantly pushed for better performance by the demands for computation intensive applications in the areas of scientific research, image processing, personal computing and network or database management and security. Modern microprocessors need to operate at high clock frequencies and maintain high data throughput. These requirements have led to several circuit and architectural innovations over the past years. Microprocessors normally have a central processing unit (CPU) which is responsible for data processing and complex computations associated with an application as well as a memory unit that stores the data operands (data memory) and the program instructions being executed (program memory). The datapath units include performance critical blocks like the arithmetic-logic unit (ALU), address generation unit (AGU), high performance adders and register files (RFs). On the other hand, the on-chip memory units

comprise of the various levels of cache hierarchy (L1, L2, L3) designed using static random access memory (SRAM), read only memory (ROM), shift registers and their control circuits. The simplified block diagram of a generic digital processor is shown in Figure 2.1 [4].

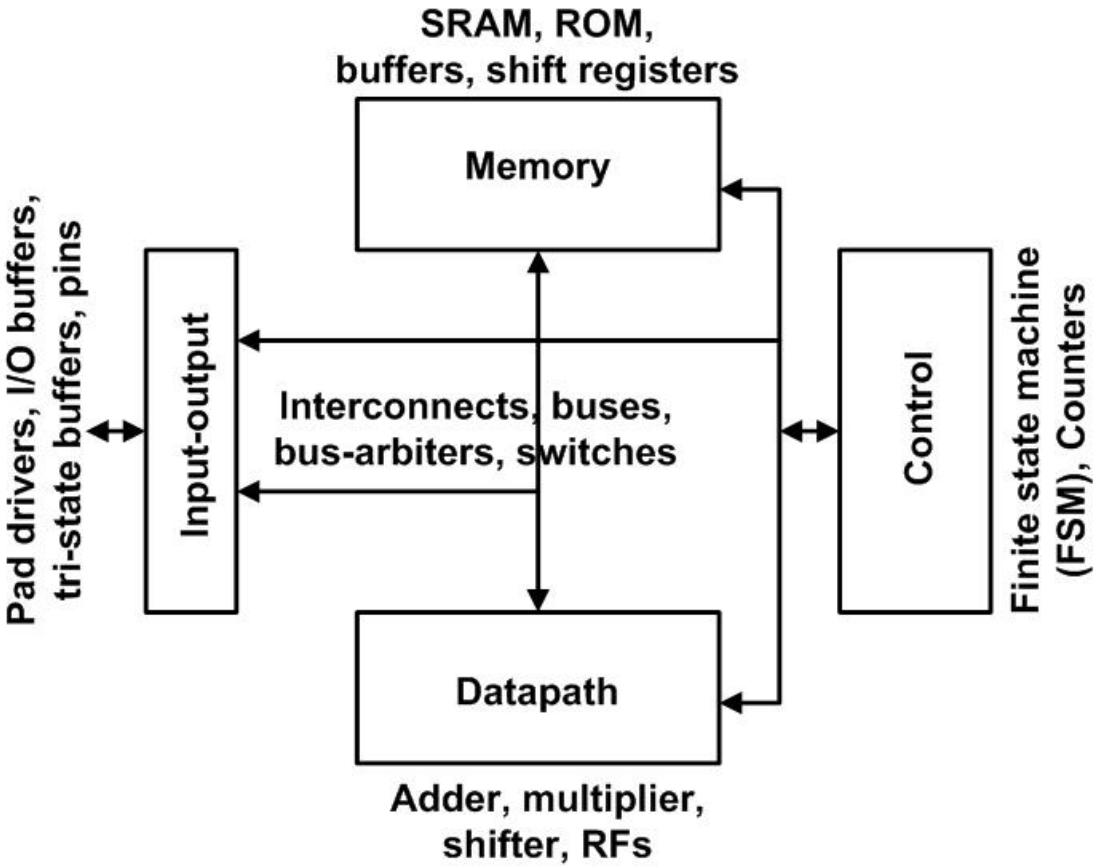


Figure 2.1: Generic block diagram of a digital processor

Digital VLSI design has traditionally been divided into two broad categories: logic design and memory design. This is because logic or datapath design and on-chip memory design have different design goals and objectives. For example, the primary goal for logic

and datapath units is the operating frequency and data throughput. As a result, logic design is associated with high frequency clocking, deep pipelining, parallel architectures, complex timing schemes allowing time-borrowing. On-chip memory is normally used to bridge the performance gap between the data processing unit and off-chip main memory. It acts as a buffer for storage of data and program and supplies operands to the processor's execution core. Therefore, the main design goals for on-chip memory is packing density, data stability, power consumption, cache coherency, improving cache hit algorithms and resolving data hazards.

In this research our focus is on high performance logic design. This chapter deals with some of the fundamentals of high performance circuit design and discusses the implementation of some of the basic building blocks. We present some of the most commonly used circuit families used in digital logic, discuss the concept of pipelining and circuit level operation of latches, flip-flops and multiplexors used in datapath designs.

## 2.1 Digital Logic Design and Circuit Families

Circuit designers have devised many different logic families in order to implement digital functions. Several different design metrics need to be considered when choosing a logic family for a particular design. These include factors like operating frequency, power (energy) consumption, area overhead, noise margin and transistor count. The different CMOS logic styles that are considered in this research are complementary static CMOS logic, dynamic logic and CMOS pass transistor logic (CPL). The basic topology and operation of each of these circuit families is briefly explained in Figure 2.2 with the help of 2-input NAND

gates [15].

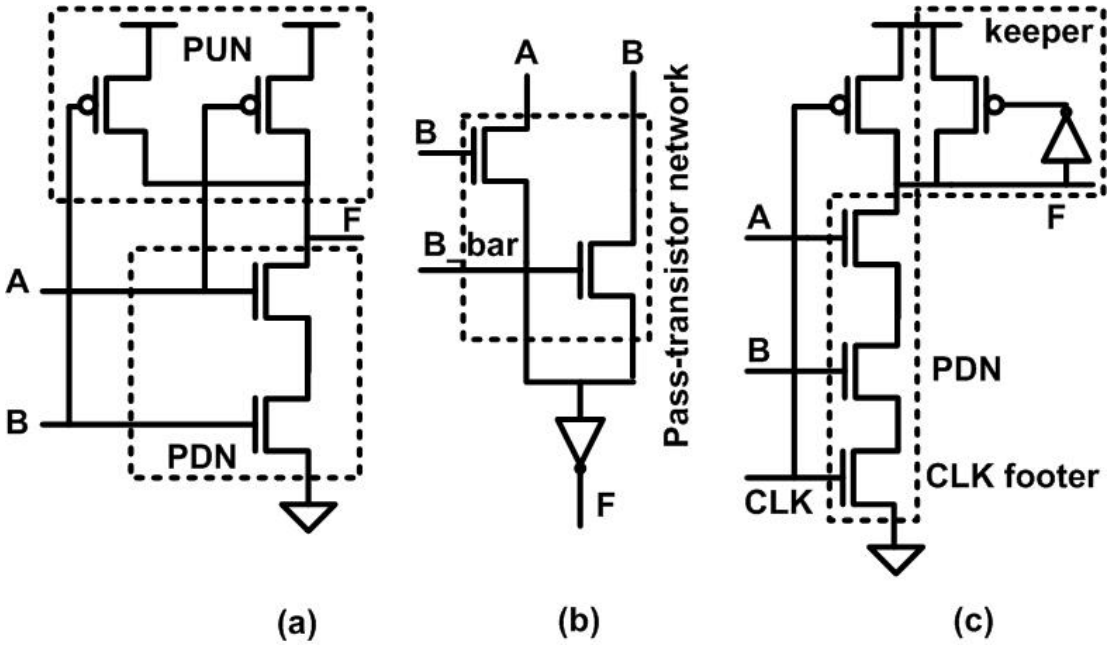


Figure 2.2: CMOS circuit families: static, pass transistor and dynamic logic

Complementary static CMOS logic gates (Figure 2.2(a)) are characterized by a p-MOS based pullup network (PUN) and an n-MOS based pulldown network (PDN). These two networks are dual of each other in that the transistors that are in series in one network appear in parallel in the other. Also, each of the primary input signals is connected to the gate terminal of at least two transistors. The p-MOS MOSFETs have carrier mobility 2-3 times lower than that of the n-MOS transistors. Therefore, in order to ensure equal rise/fall times and PUN/PDN conductivity, the p-MOS transistors are designed using widths that are proportionately larger than the n-MOS transistors. Even though the p-MOS PUN of complementary CMOS logic gates is area intensive, the gate output (F) is always connected

to either $V_{DD}$ or GND through a low impedance path and therefore has high noise margin. Complementary static CMOS logic gates are used for implementing non-critical functions and control logic and their noise margin/robustness scale well with technology.

CMOS pass transistor based designs (Figure 2.2(b)) use n-MOS pass transistors and wired-OR logic to implement logic functions. The p-MOS network of static CMOS gates is eliminated, allowing for lower switched capacitance, transistor count and smaller area. However, since the input signals are connected to the transistor drain terminal (A, B in Figure 2.2(b)), these gates do not have the driving capability of the static counterparts. In addition, the n-MOS transistors pass a "weak" logic 1. This degrades the signal rise and fall times and restricts the number of CPL logic gates that can be cascaded. Therefore designers often use inverters at the gate output to achieve full-rail signal swing and current drive capability as shown in Figure 2.2(b).

The dynamic logic family (Figure 2.2(c)) has been developed in order to achieve better performance compared to the static logic counterparts [4], [15]. It is characterized by the clocked (CLK) p-MOS precharge and n-MOS footer transistor. When CLK is logic low, the logic gate is in precharge phase and the dynamic output node F is connected to $V_{DD}$. However, when CLK is a logic high ($V_{DD}$), the pulldown n-MOS clocked transistor turns on and the gate can evaluate depending on the states of the input signals A and B. Unlike the static logic gates, the output node F can be floating and be in a high impedance state in the event that CLK=1 and any one or both A and B signals are logic low (GND). This can lead to several problems including charge-sharing, clock feed through and degraded noise margin [4], [5], [15]. To avoid such a scenario circuit designers typically employ a

weak p-MOS keeper as shown in Figure 2.2(c). It is clear that dynamic logic does away with the p-MOS PUN network of the static CMOS logic gates. This leads to significantly lower capacitance at the output node and improved propagation delay. However, the CLK signal has a high switching activity leading to higher power consumption, and these gates typically have poorer noise margins compared to the static logic gates. Circuit designers have developed many different types of dynamic logic styles which include:

- n-MOS domino (Figure 2.2(c)),
- p-MOS domino,
- n-p domino,
- zipper domino,
- multiple output domino logic (MODL), and
- compound domino logic (CDL).

A detailed discussion of the different domino logic techniques and the various design tradeoffs is presented in [15]. It should be noted that, for most high performance logic implementations like ALU, AGU, adder and RF, circuit designers prefer to use the compound domino logic (CDL) style.

## 2.2 Footerless CDL and SRCPL Based Logic Design

As mentioned earlier, complementary static CMOS logic offers improved noise margin and scalability while dynamic logic has better performance. Therefore, high performance logic

designers use CDL based logic for critical path designs that incorporate alternate stages of n-MOS dynamic and static CMOS logic gates to implement a given function. The inverting nature of static CMOS logic helps to interface these gates with the next stage of n-MOS dynamic logic gate and improve the overall noise margin. Figure 2.3 shows 4-stages of CDL logic gates implementing 2-input NAND-NOR logic.

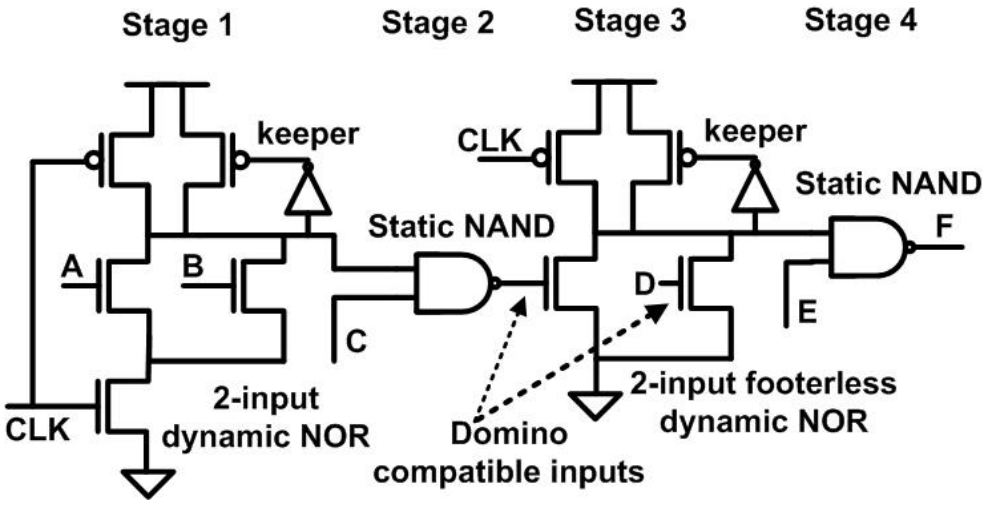


Figure 2.3: Multi-stage footerless compound domino logic

The first stage of logic gate (Stage 1, Figure 2.3) has primary inputs like A, B that can have arbitrary timing. Therefore the input stage of the domino logic gate has an n-MOS clocked footer to prevent any steady-state current. However, the internal signals (input D for Stage 3) are generated by the static logic gates and are therefore domino compatible (logic low during precharge). This eliminates the possibility of any steady-state current and the need for clocked footer transistor. This reduces the stack height, improves circuit performance and lowers power consumption. Therefore multi-stage logic

is normally designed using footerless CDL logic gates as shown in Figure 2.3. It should be noted that most critical path logic gates (static and dynamic) are restricted to 2 or 3 inputs. This ensures low output capacitance, limits the number of stacked transistors and body effect related performance degradation. This logic style is used in Chapter 3 to design the performance critical units of our 32-bit adder and ALU.

In addition to the CDL logic, our design also uses swing-restored complementary pass transistor logic (SRCPL). This logic is an extension of the basic CPL family shown in Figure 2.2(b). CPL logic based design has lower switched capacitance and occupies less area. In addition, the wired-OR based strategy makes it especially suitable for multiplexor (for instances on non-critical paths) and XOR designs. Therefore, it is well suited for implementing the logic and shifter units of the ALU. However, as mentioned earlier, CPL logic results in "weak" 1 and requires intermediate inverters in order to cascade multiple stages. Some of these problems can be overcome by using the SRCPL based design approach as shown in Figure 2.4.

The original CPL logic based 2-input NAND gate is shown in Figure Figure 2.4(a) while the SRCPL based implementation is given in Figure 2.4(b). When both inputs A and B are logic 1, the intermediate logic node of the CPL gate passes a "weak" logic 1 and reaches $V_{DD} - V_{TH}$. As a result, the p-MOS transistor of the subsequent inverter is not fully turned off causing large $I_{OFF}$ static current flow during steady-state operation. However, in the swing-restored CPL design (Figure 2.4(b)), a minimum sized p-MOS keeper is added which restores full-rail logic operation to the intermediate nodes as well. The swing-restorer circuitry adds capacitance to the internal node but also improves the signal rise

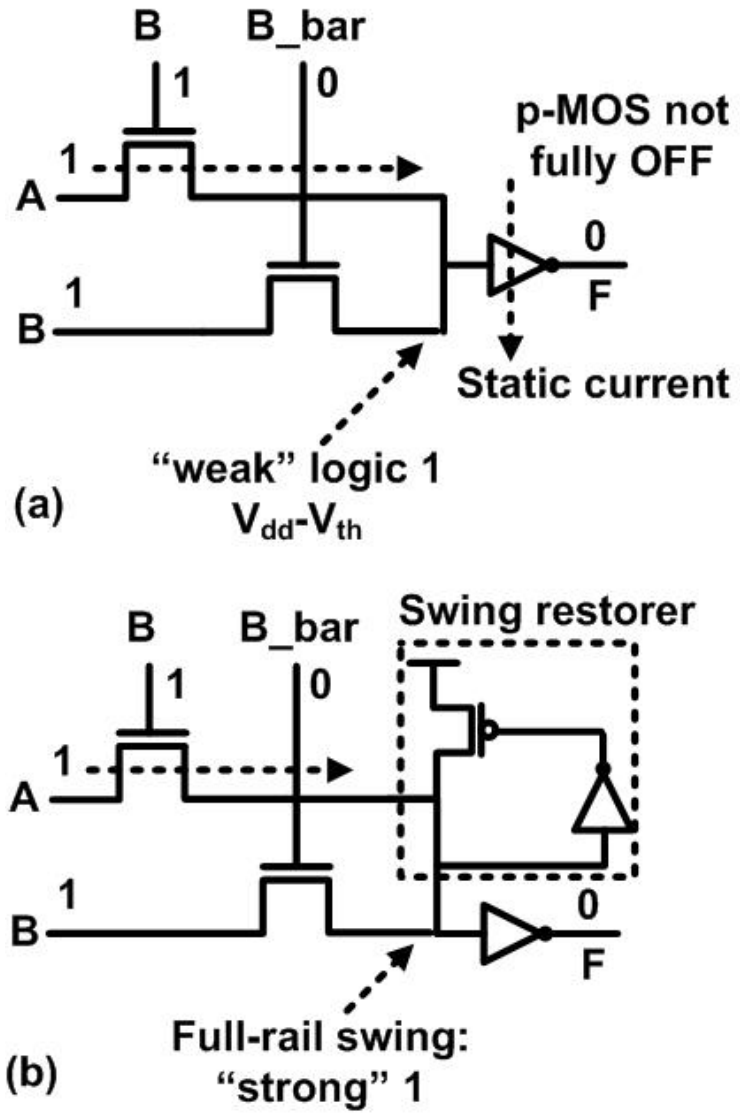


Figure 2.4: CPL vs. SRCPL comparisons

time during the $0 \rightarrow 1$ transition. In addition the SRCPL design ensures static power free logic operation and better noise margin.

## 2.3 Pipelined Datapath Designs: Basic Concept

One of the key features of modern high performance microprocessor datapaths is deep pipelining. The number of levels of pipelining is increased by reducing the amount of logic in each pipeline stage. This improves the processor clock speed and data throughput at the expense of increasing the processor latency and clock power (energy). The intermediate stages of a pipelined datapath normally use latches that allow time-borrowable logic operation. It is estimated that as much as 50% of the power in certain processors can be dissipated in the clock network [5], [7]. This is a direct consequence of the fact that modern microprocessors are heavily pipelined, the clock signal has high switching activity and the pipelined latches drive significant amounts of on-chip capacitance. The basic concept of datapath pipelining is shown in Figure 2.5. A non-pipelined datapath organization is shown in Figure 2.5(a) while a pipelined representation is given in Figure 2.5(b).

For the non-pipelined design, we show three cascaded logic blocks, with the first unit being an ALU and the subsequent blocks being represented by Fn. A and Fn. B respectively. The clock period ($T_{CLK}$) for such a design has to account for the worst case delay of the series connected logic blocks and is given by:

$$T_{CLK}^{non-pipelined} = T_{delay}^{reg} + T_{delay}^{ALU} + T_{delay}^{Fn.A} + T_{delay}^{Fn.B} + T_{setup}^{reg} \tag{2.1}$$

where $T_{CLK}^{non-pipelined}$ represents the clock period for the non-pipelined design, $T_{delay}^{reg}$ is the register (latch / flip-flop) propagation delay and $T_{setup}^{reg}$ is the register setup-time. Also $T_{delay}^{ALU}$, $T_{delay}^{Fn.A}$, $T_{delay}^{Fn.B}$ represent the propagation delays of the individual logic blocks. As seen in Figure 2.5(b), two additional stages of latches are inserted in the datapath that

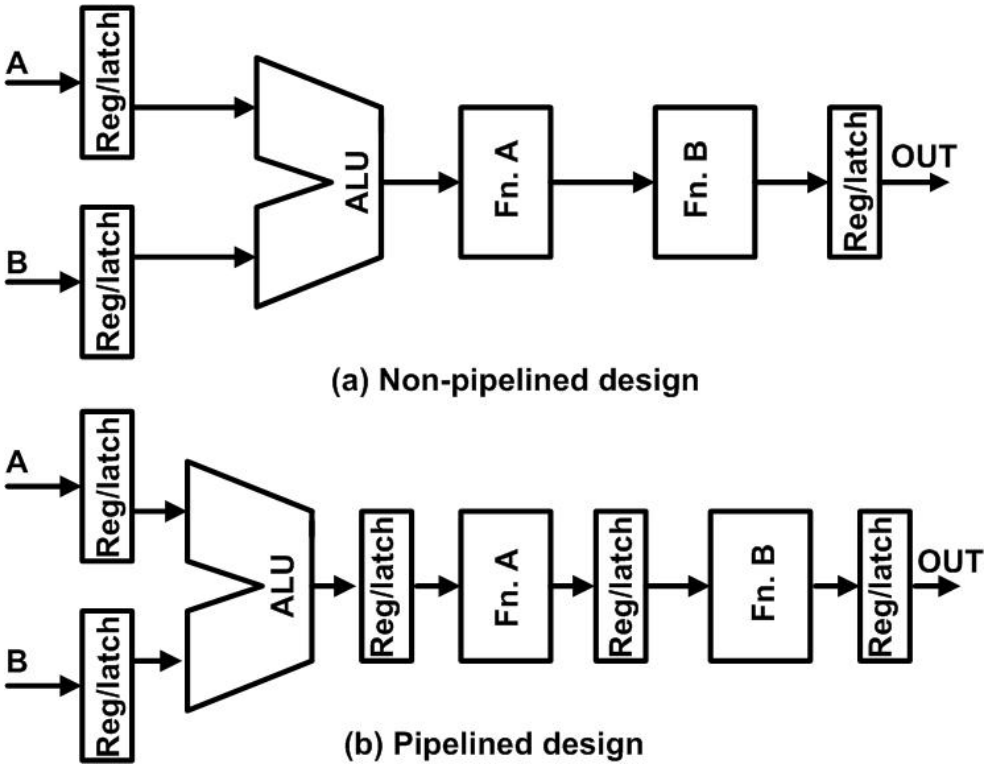


Figure 2.5: Datapath pipelining: basic concepts

break up the original datapath in to three pipelined stages. Therefore, in the case of the pipelined design the clock period has to accommodate the worst case delay of only one of the logic units and can be expressed as:

$$T_{CLK}^{pipelined} = T_{delay}^{reg} + max(T_{delay}^{ALU}, T_{delay}^{Fn.A}, T_{delay}^{Fn.B}) + T_{setup}^{reg} \tag{2.2}$$

It is clear that by inserting additional stages of registers, it is possible to reduce the total logic delay in an individual stage and thereby improve circuit performance. However, this also increases the depth of the pipeline and therefore the latency of the datapath and

overall clock and system power (energy). It will be explained subsequently that for our 32-bit ALU design, we use a two stage pipeline with one stage for the instruction decode operation and another stage for the opcode execution.

## 2.4 Latches and Flip-flops for Pipelined Datapaths

The design of high performance latch and flip-flop circuits is of importance for digital designs. Several different types of latch and flip-flop circuits have been proposed in the literature [5]. For example, some of the different types of flip-flops that have been discussed in different VLSI designs include D (data), SR (set-reset), T (toggle) and JK flip-flops with both static and dynamic logic based implementations. The prime concern for designers is the latch (flip-flop) performance, power consumption and data stability. VLSI datapath designers normally use static logic based D-type latches. This is because static logic based designs have better noise margins and scalability than their dynamic logic counterparts. Since latches and flip-flops are used extensively and are critical to data storage / retention, most designers prefer robust static logic based designs. For pipeline stages that require flip-flops, two cascaded D-type latches are used to design master-slave flip-flops.

The transmission gate (TG) based D-latch circuit and its timing diagram is shown in Figure 2.6. The input TG is off when CLK=1. During this time the feedback TG in on, which completes the feedback path of the latch retaining the latched data. At this time the latch data is not affected by input data changes and the latch is "opaque" (holds the stored data). However, when CLK=0, the input TG is on and the latch is "transparent". During this time the latch output Q tracks the input data but is inverted in phase. The

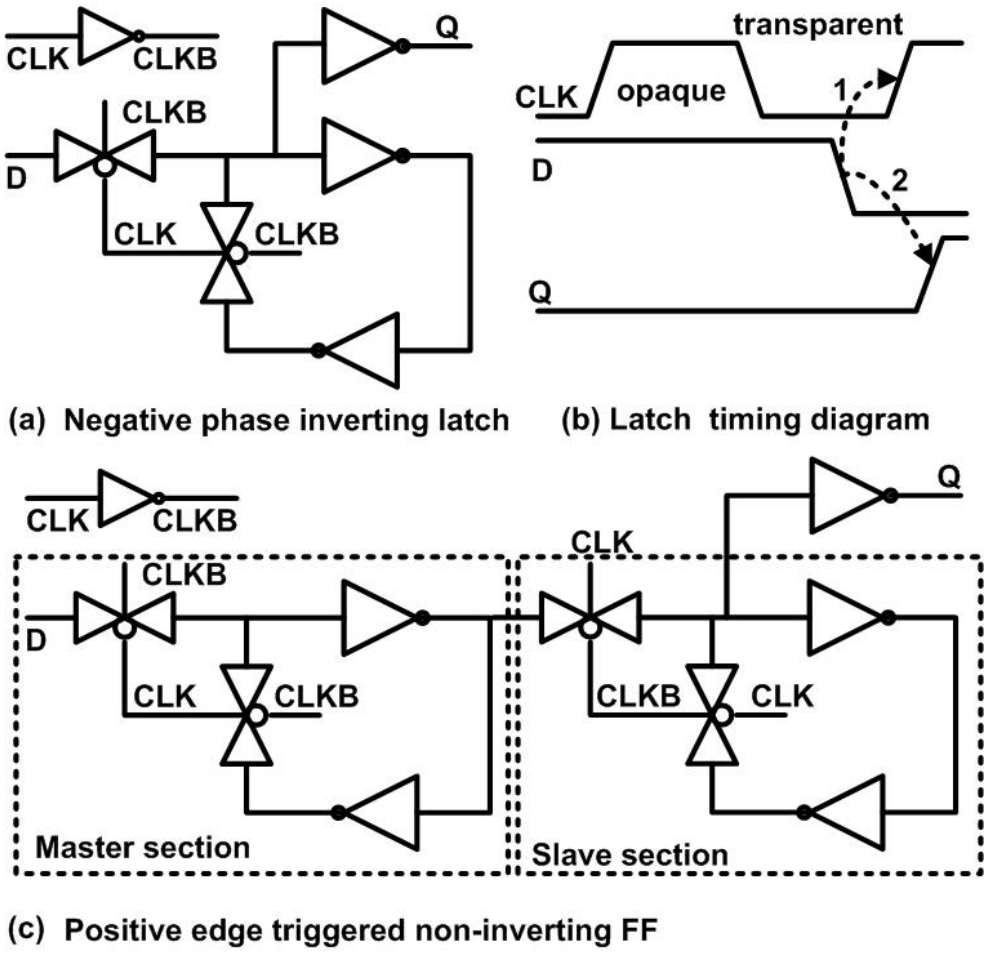


Figure 2.6: Latch and filp-flop circuit schemes

latch again becomes opaque after a $0 \rightarrow 1$ transition of the CLK signal. The data (D) can be pushed into the CLK=0 phase up to the point when the output Q does not any longer register the correct data. This time is indicated in Figure 2.6(b) by the label "1" and is referred to as the latch setup time ($T_{setup}^{latch}$). The performance of a latch is dependent on both its setup time and D $\rightarrow$ Q propagation delay. The latch propagation delay D $\rightarrow$ Q is

indicated in Figure 2.6(b) by the label 2.

The circuit schematic of a positive edge triggered D flip-flop is shown in Figure 2.6(c). The flip-flop is implemented using two back-to-back D-latches operating on opposite clock phases. For both the latch and flip-flop designs the CLKB (inverted clock) signal is generated from the system clock using a local inverter which helps to minimize the clock skew. A common design practice is to use minimum or close to minimum sized transistors for the feedback inverter to reduce contention and switching energy during writing of new data. Also, a driver-inverter is used at the latch and flip-flop output Q, to prevent directly driving signals and long interconnects using the storage node. This helps to improve the overall noise immunity of the design.

## 2.5 Multiplexors for High Performance Designs

In addition to the logic gates and sequential elements like latches and flip-flops, another important building block that is often used in digital circuit designs is the multiplexor (MUX). Multiplexors perform the logic operation of a many-to-one mapping and can be implemented in several different ways. MUX-es are used extensively to select the input data drivers at the ALU front-end and in the performance critical read circuitry of register files (RFs). The multiplexor operation selects the correct data driver amongst several possible inputs and drives the correct data on to a single output line (bus). The simplest multiplexor operation is that of a 2:1 MUX and can be expressed as shown:

$$F = S.(D1) + SBAR.(D0) \tag{2.3}$$

The signal S is the MUX control and helps select between input data D0 and D1. For example, when S=1, D1 is selected and is available at the MUX output F. On the other hand, D0 is selected when SBAR is 1 (S=0). This operation can be expanded to implement wide-MUXes (for example 4:1, 8:1 or 16:1) using wired-OR logic to merge several parallel logic units and drive a single bus. The operation of wide-MUXes will be explained in more detail in chapter 4 in the context of the 32-bit ALU front-end and high performance register file (RF) designs. There are several different circuit techniques for designing MUX-es. These include the transmission gate (TG) approach, $C^2MOS$ design and n-MOS domino logic based designs. The circuit level implementation of a 2:1 MUX using each of these circuit style is shown in Figure 2.7.

The different MUX implementations perform the same logic operation but each of them has certain design tradeoffs associated with it. For example, the static TG and $C^2MOS$ implementations (Figure 2.7(a) and (b), respectively) have larger propagation delay compared to the domino design (Figure 2.7(c)). However, the static MUX-es can be interfaced with other logic blocks designed using any circuit style (domino or static). This is not the case with the domino MUX which requires the select signals (S, SBAR) to be domino-compatible and be equal to logic low (GND) during the precharge phase. Therefore, the TG and $C^2MOS$ designs are normally used to multiplex non-critical logic blocks. In particular, in our design, we use the TG MUX-es with logic units that are implemented using complementary CMOS logic and $C^2MOS$ MUX-es at the output of SRCPL based logic units. This prevents the usage of series connected pass transistors and degraded signal rise/fall times. The dynamic MUX-es consume more switching power and

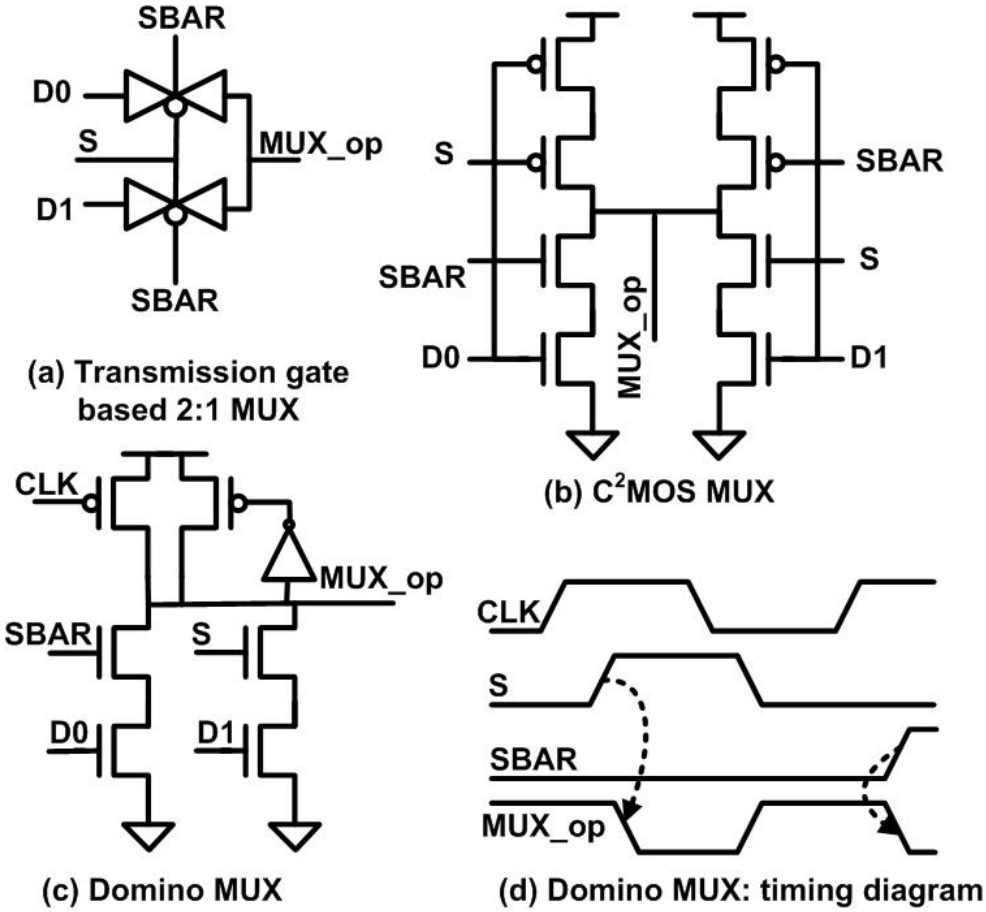


Figure 2.7: Multi-stage footerless compound domino logic

therefore in our design they are used exclusively for the 32-bit ALU front-end, ALU output mux-ing and RF read port. It should be noted that both the TG and $C^2MOS$ based MUX designs consume glitching power due to unbalanced path delays. However, this is not a concern in dynamic MUX designs and therefore static MUX-es require careful sizing to minimize this problem.

## 2.6 Summary

In this chapter we presented some of basic circuit styles used in high performance digital logic design. Some of the basic design tradeoffs associated with complementary static CMOS, pass transistor logic (CPL) and dynamic logic were discussed. The fundamental concepts associated with pipelining were enumerated and schemes for designing latch and flip-flop circuits were shown. Finally, we presented different MUX topologies and discussed their design tradeoffs. In the next chapter we will use these building blocks and circuit techniques to design a dual supply, SRCPL based, 32-bit ALU.

# Chapter 3

# High Performance ALU Design and Low Power Operation

In this chapter, we discuss the design details of a high performance 32-bit arithmetic and logical unit (ALU) and present the different circuit techniques used to ensure its low power (energy) operation. This design represents a fixed point unsigned ALU that can be used in the integer execution unit (IEU) of general purpose microprocessors. The ALU comprises a 32-bit high performance adder unit and a logic-shifter unit. In addition, there is a decoder unit to activate the different sub-blocks within the ALU depending on the instruction being executed. In this design, several modifications were made in order to reduce the ALU power consumption without degrading its performance. First, a dual power supply based clocking scheme was adopted to reduce both switching and leakage power (energy) consumption in the non-critical portions of the ALU. Second, the design was partitioned into critical and non-critical units to facilitate dual supply assignment and

routing. Third, a latch (flip-flop) circuitry was developed that can support dual supply clocking without resulting in static-power consumption. Finally, a swing-restored CPL (SRCPL) based design approach was adopted for the non-critical logic and shifter units to further reduce overall switching energy consumption. The 32-bit ALU design along with these modifications will be discussed in this chapter. In addition, the impact of the above techniques on different ALU performance metrics for the 180nm to the 65nm CMOS technologies will also be presented in this chapter.

## 3.1 Different Power Components in CMOS Logic

One of the key design issues in high performance logic is the total power (energy) consumption. As indicated in Chapter 1, the CMOS logic power (energy) is reduced with CFS scaling. With technology scaling, both the supply voltage and transistor threshold are scaled and leakage power becomes an increasingly important component of the total IC power. Consequently, CMOS IC power scaling deviates from the traditional scaling trends observed in earlier technology generations. Circuit designers now spend a considerable amount of effort in analyzing and reducing the total power (energy) consumption of high performance datapaths. The CMOS circuit power can be broken down into several different components [4]:

- Switching ($P_{sw}$),
- Leakage ($P_{leak}$), and
- Short-circuit ($P_{sc}$).

These different power components can be best explained in the context of a CMOS inverter. Some of the basic concepts are discussed in this section and can be extended to more complex logic gates and ICs. However, the problem of estimating the system level power consumption is dependent on many different factors that include the transistor sizes, interconnect parasitics, power supply voltage, input vectors and internal logic states. It is therefore difficult to develop an accurate analytical model for the overall power (energy) for a complex system and designers normally rely on circuit and system level simulators and EDA (electronic and design automation) tools to estimate the overall power consumption.

### 3.1.1 Switching Power Component

We now focus on some of components of CMOS circuit level power consumption and explain them with respect to an inverter. The detailed analysis of these power components is discussed extensively in the literature [4], [5], [7]. The switching power ($P_{sw}$) component is associated with the charging and discharging of the inverter parasitic and load capacitance whenever there is a change in its logic state. This is explained with the help of Figure 3.1 which shows a simple inverter with a lumped output capacitive load of $C_L$. When the input makes a logic $1 \rightarrow 0$ transition, the inverter p-MOS transistor turns on and the load capacitance is charged up from 0 to $V_{DD}$. During this time there is a transient charging current from the power supply ($V_{DD}$) indicated by $i_{DD}$.

Under steady state conditions, the inverter output reaches $V_{DD}$ and the current drops to zero. The same cycle is repeated when the inverter input makes a logic $0 \rightarrow 1$ transition, only that this time the capacitor discharges to ground. It can be shown [4] that this cycle

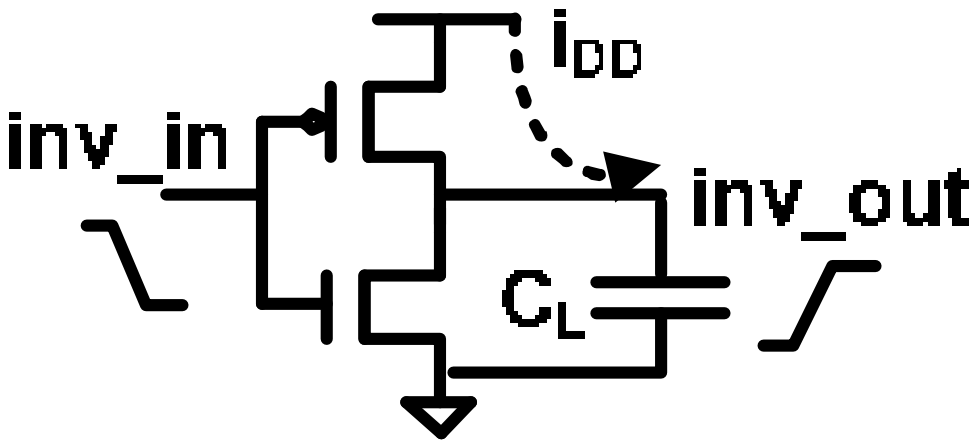


Figure 3.1: Inverter during switching transient

results in an overall power (energy) consumption per switching as shown below:

$$P_{sw} = C_L V_{DD}^2 f_{sw} \tag{3.1}$$

$$E_{sw} = C_L V_{DD}^2 \tag{3.2}$$

where $P_{sw}$ and $E_{sw}$ are the switching power and energy respectively, $C_L$ is load capacitance, $V_{DD}$ is the power supply voltage and $f_{sw}$ is the switching frequency. It should be noted that the dissipated power is converted into heat that causes localized hot spot problems making thermal management a major challenge in scaled technologies. In long channel transistors the dynamic power is the dominant component of the total IC power. However, this is not the case for DSM technologies and therefore we now discuss the leakage power

component in CMOS circuits.

### 3.1.2 Leakage Power Component

The leakage or static power component in CMOS circuits is typically negligible. This was one of the key advantages of CMOS circuits. Static (leakage) power is the amount of power consumed by the circuit when there is no switching activity and the logic is quiescent. The basic contributors to this power can be explained using the following Figure 3.2:

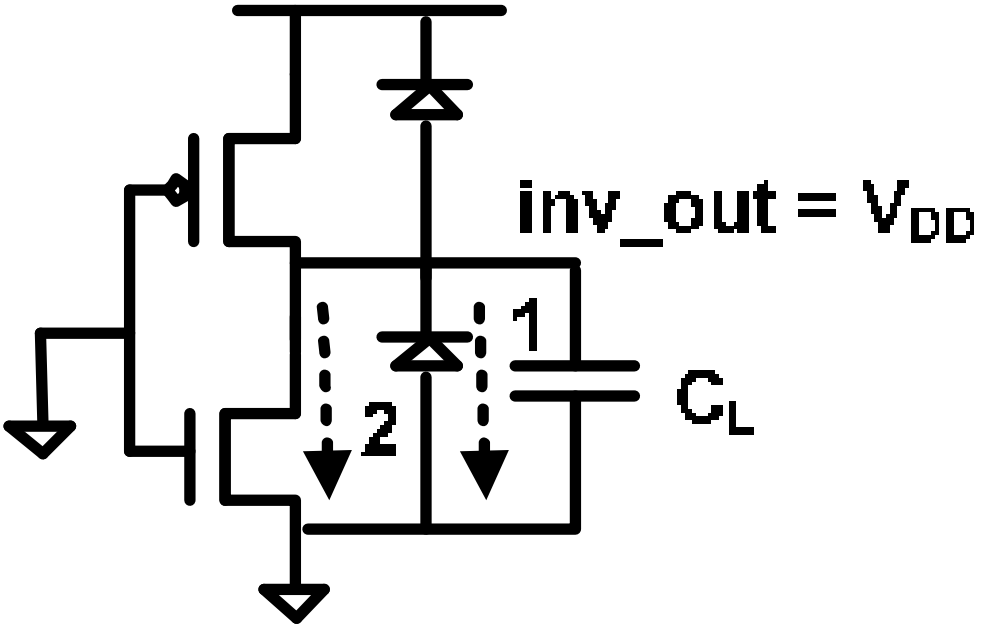


Figure 3.2: CMOS inverter leakage current components

There are two primary components of leakage power that are associated with CMOS

circuits and they are marked 1 and 2. The component 1 refers to the current (power) drawn by the reverse biased p-n junction diode that causes drain leakage. However, in most CMOS circuits this current component is negligible compared to the subthreshold current. This component is marked 2 in the figure and results in a drain to source leakage current even when the transistor is OFF ($V_{GS} < V_{TH}$). The subthreshold leakage current has an exponential dependence on the both the threshold voltage and operating temperature. Thus as the transistor $V_{TH}$ is lowered and operating frequency increases, this component is becoming increasingly important in scaled technologies. In fact, it is estimated that beyond the 90nm generation the total power of complex ICs maybe dominated by subthreshold leakage. The leakage power associated with CMOS circuits can be expressed as:

$$P_{leak} = I_{OFF} V_{DD} \tag{3.3}$$

There are many different physical and empirical models for the $I_{OFF}$ current [4], [5], [16] that have been reported in the literature and one of the models that is frequently used by digital circuit designers will be presented subsequently.

### 3.1.3 Short Circuit Power Component

During any transition, the CMOS logic consumes switching or dynamic power as mentioned earlier. However, the earlier discussion assumes that all the current during the transition goes from the power supply to charge / discharge the load capacitor. This is valid in the case of signals with zero rise and fall times. Under such a situation there is no short circuit path from $V_{DD}$ to ground. In more realistic situations, both the input and output signals

have finite rise and fall times. As a result, there is a short circuit direct path from $V_{DD}$ to ground. This situation is better explained with the help of the following Figure 3.3:

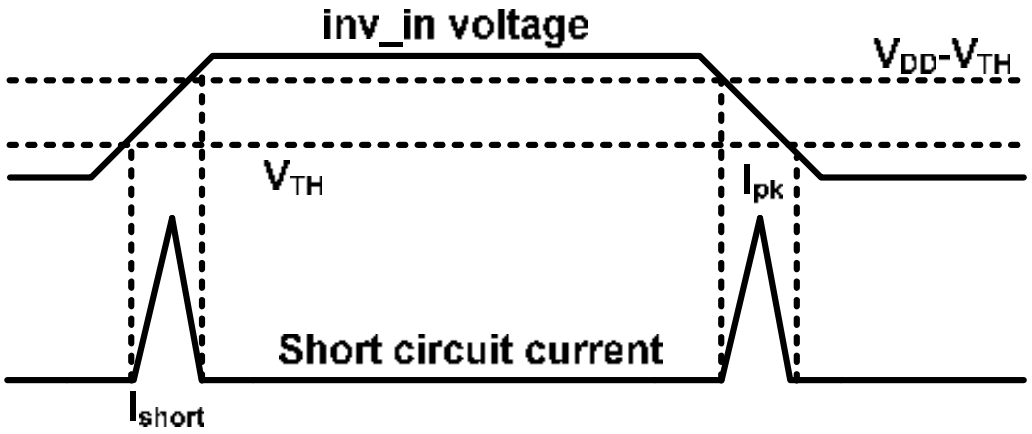


Figure 3.3: CMOS inverter short circuit current during switching transient

During the input signal transition time between voltage levels $V_{TH}$ and $V_{DD} - V_{TH}$, both the p-MOS and n-MOS transistor of the inverter are partially ON. Thus there is a direct path that results in the peak short circuit current as denoted by $I_{pk}$. The magnitude of this current is given by the actual transistor widths and the on-state saturation current ($I_{DSAT}$). The duration of the current flow depends on the signal rise and fall times and increases as the signal slopes degrade. The overall direct path energy can be obtained by integrating the current under the triangular waveforms during the duration of short circuit. A simplified expression for of the short circuit energy (power) is given in [4]:

$$E_{sc} = (\frac{t_r + t_f}{2})V_{DD}I_{pk} \tag{3.4}$$

$$P_{sc} = (\frac{t_r + t_f}{2}) V_{DD} I_{pk} f_{sw} \tag{3.5}$$

where $E_{sc}$ and $P_{sc}$ are the short circuit energy and power respectively, $t_r$ and $t_f$ are the signal rise and fall times while the rest of the symbols have their usual meanings. Normally high performance logic and datapath designs have sharp signal slopes (low $t_r$, $t_f$) that help to reduce the short circuit power. It is estimated that for well designed ICs, the short circuit component is between 10%-20% of the switching power.

Based on the above discussion is it possible to combine all of the different power components and obtain the total power for a CMOS inverter as shown below:

$$P_{total} = P_{sw} + P_{leak} + P_{sc} = C_L V_{DD}^2 f_{sw} + I_{OFF} V_{DD} + V_{DD} I_{pk} (\frac{t_r + t_f}{2}) f_{sw} \tag{3.6}$$

## 3.2 Supply Scaling and MOSFET Current Components

The above expression for power consumption shows that supply voltage scaling is one of the most effective ways to reduce IC power. In this section, we briefly discuss the different transistor level current components of a MOSFET and demonstrate the impact of supply scaling on each of them. This provides the motivation for a dual supply design for low power ALU operation and helps us understand the different design tradeoffs associated with such an approach. Several design techniques have been proposed that minimize transistor level leakage and system power consumption in high performance ICs [5], [9], [16], [17],

[18], [19], [20], [21], [22]. Some of these include transistor level leakage control techniques such as:

- dual $V_{th}$ techniques,
- multi-oxide or non-minimum channel length transistors,
- reverse body bias (RBB), and
- stack effect.

Other techniques such as dual supply designs [23], [24], [25], [26] low standby power (sleep mode) operation [5], and reduced swing logic have been proposed to tackle the IC power issue at a system level. However, each of the above techniques is associated with design overheads, which include performance degradation, the possible need to generate and route additional power supplies, area overhead or additional process steps.

One of the primary contributors to total IC power in scaled CMOS technologies is the transistor off-state current. The off-state current is increasingly exponentially while the on-state $I_{DSAT}$ current does not increase in proportion. This is resulting in a degradation in the $I_{ON}/I_{OFF}$ ratio which is one of the key metrics used by circuit designers. The exponential increase in the $I_{OFF}$ current and consequent degradation of the $I_{ON}/I_{OFF}$ ratio is shown in Figure 3.4.

These plots show the transistor $I_{ON}/I_{OFF}$ ratio and threshold voltages ($V_{TH}$) for low and high $V_{TH}$ n-MOS transistors for both the 130nm, 90nm and 65nm technologies using the BPTM models [13], [14]. This is resulting in excessive leakage currents for the sub-130nm generations and offsets the reduction in switching energy obtained from scaling.

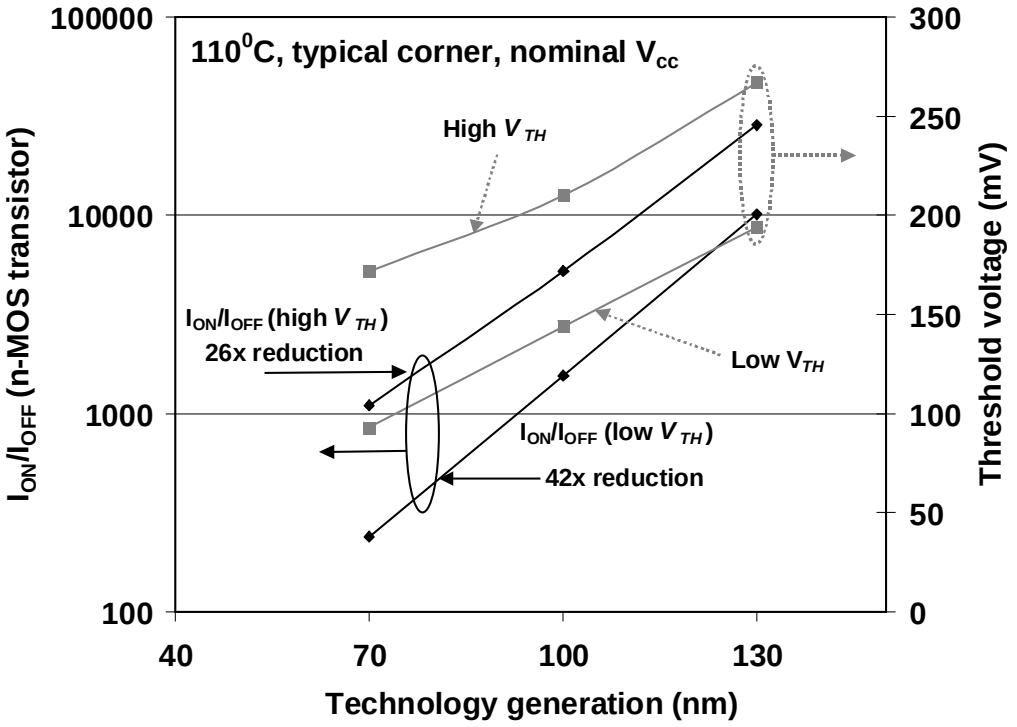


Figure 3.4: Transistor OFF-state current scaling for sub-180nm CMOS technologies

In addition, aggravated leakage current is causing thermal hot spots and thermal run away problems during burn-in and adversely affecting long-term reliability of high-end microprocessors [6], [27], [28].

### 3.2.1 MOSFET OFF-Current Model

The transistor $I_{OFF}$ comprises of several different components [5], [16], [29], of which the weak inversion and drain-induced barrier lowering (DIBL) currents are the most important. These two dominant leakage current components can be modelled as shown below:

$$I_{OFF} = Ae^{(\frac{V_{GS}-V_{TH0}-\gamma V_{SB}+\eta V_{DS}}{nv_T})}(1 - e^{\frac{-V_{DS}}{v_T}}) \tag{3.7}$$

where $A = \mu_0 C_{ox} \frac{W}{L_{eff}} v_T^2 e^{1.8}$, $\mu_0$ is the zero bias carrier mobility, $C_{ox}$ is the gate-oxide capacitance, $L_{eff}$ is the transistor effective channel length, W is the transistor width, $\eta$ is the DIBL coefficient, $\gamma$ is the linearized body effect coefficient, $n$ is the transistor sub-threshold swing coefficient and $v_T$ is the thermal voltage given by $\frac{kT}{q}$ ( 33mV at $110^0$C). In addition, $V_{TH0}$, $V_{GS}$, $V_{SB}$ and $V_{DS}$ denote the transistor zero-bias threshold voltage, gate-source, source-body, and drain-source voltages, respectively. We determine the worst-case transistor leakage by simulating the device $I_{OFF}$ at $110^0$C for $V_{GS} = 0$ and $V_{DS} = V_{DD}$. Based on Eq. (3.7), it is possible to establish closed form approximate expressions to model [30], [31] the reduction in leakage current due to supply scaling. We use the term $\frac{\Delta I_{OFF}}{I_{OFF}}$ as a figure of merit (F.O.M.) for $I_{OFF}$ reduction, where:

$$\frac{\Delta I_{OFF}}{I_{OFF}} = \frac{I_{OFF} - I_{OFF}^{final}}{I_{OFF}} = 1 - \frac{I_{OFF}^{final}}{I_{OFF}} \tag{3.8}$$

where $I_{OFF}$ and $I_{OFF}^{final}$ represent the transistor OFF-state current without and with $V_{DD}$ scaling, respectively. In order to simplify the modelling, we assume that $e^{\frac{-V_{DS}}{v_T}} \approx 0$. This approximation is justified since in this study, the ratio of $\frac{V_{DS}}{v_T} \geq 20$. Thus, the simplified transistor OFF-state current is given by:

$$I_{OFF} = Ae^{(\frac{V_{GS}-V_{TH0}-\gamma V_{SB}+\eta V_{DS}}{nv_T})} \tag{3.9}$$

Scaling the power supply lowers the transistor drain-source ($V_{DS}$) voltage, thereby reducing the DIBL current. In addition, since $V_{SB} = 0$, the corresponding term in Eq.

(3.7) is equal to zero. Thus, the savings in leakage current obtained by lowering of the supply voltage can be approximated as [10]:

$$\frac{\Delta I_{OFF}}{I_{OFF}}|^{V_{DD}} = 1 - e^{\frac{-\eta \Delta V_{DS}}{n v_T}} \tag{3.10}$$

### 3.2.2 MOSFET Gate Leakage Current

In addition to the transistor off-state current (sub-threshold leakage), it is expected that gate leakage will also be a significant component of total leakage for scaled CMOS technologies. This is especially true in the case of sub-90nm CMOS technologies where the transistor gate oxide thickness may be less than 2nm. Fowler-Nordheim (FN) and direct tunnelling are the two primary mechanisms that cause gate leakage current in deep submicron transistors. In the sub 1V regime, the voltage across the gate oxide is less than the barrier height for electrons in the conduction band ($V_{ox} < \phi_{ox}$). As a result, the $I_{GATE}$ component is determined primarily by the direct tunnelling current and can be approximated as [29], [32]:

$$J_{DT} = AE_{ox}^2 exp[-\frac{B(1-(1-\frac{V_{ox}}{\phi_{ox}})^{\frac{3}{2}})}{E_{ox}}] \tag{3.11}$$

where $J_{DT}$ is the direct tunnelling current density, and A and B are constants, $E_{ox}$ and $V_{ox}$ are the electric field and voltage across the gate oxide respectively, while $\phi_{ox}$ is the barrier height for electrons in the conduction band. The impact of supply voltage reduction $\Delta V_{DD}$ on the direct tunnelling current density $\frac{\Delta J_{DT}}{J_{DT}}$ can be expressed as follows:

$$\frac{\Delta J_{DT}}{J_{DT}} = \frac{J_{DT}^{V_{DD}} - J_{DT}^{V_{DD}-\Delta V_{DD}}}{J_{DT}^{V_{DD}}} = 1 - \frac{J_{DT}^{V_{DD}-\Delta V_{DD}}}{J_{DT}^{V_{DD}}} \tag{3.12}$$

By substituting $E_{ox} = \frac{V_{ox}}{T_{ox}}$ and dividing the current density expressions for the two different voltage conditions, we obtain:

$$\frac{\Delta J_{DT}}{J_{DT}} = 1 - (1 - \frac{\Delta V_{DD}}{V_{DD}})^2 exp(-BT_{ox})[\frac{(1 - \frac{V_{DD}-\Delta V_{DD}}{\phi_{ox}})^{\frac{3}{2}}}{V_{DD} - \Delta V_{DD}} - \frac{1 - (\frac{V_{DD}}{\phi_{ox}})^{\frac{3}{2}}}{V_{DD}}] \tag{3.13}$$

Since $V_{ox} < \phi_{ox}$, it is possible to expand and simplify the exponential terms in the above equation by neglecting the third and higher order terms. This allows us to obtain an approximate closed form expression for the direct tunnelling current density reduction due to supply scaling:

$$\frac{\Delta J_{DT}}{J_{DT}} \approx 1 - (1 - \frac{\Delta V_{DD}}{V_{DD}})^2 exp(\frac{3BT_{ox}\Delta V_{DD}}{8\phi_{ox}^2}) \tag{3.14}$$

For small changes in supply voltage, $\Delta V_{DD} \ll V_{DD}$ this can be further simplified and expressed as:

$$\frac{\Delta J_{DT}}{J_{DT}} \approx 1 - exp(\frac{3BT_{ox}\Delta V_{DD}}{8\phi_{ox}^2}) \tag{3.15}$$

Figure 3.5 presents simulation results for the 65nm technology (Berkeley PTM, level 54), demonstrating the impact of $V_{DD}$ scaling on the different transistor current componenets: $I_{DSAT}$, $I_{OFF}$ and $I_{GATE}$. This figure shows that a 30% reduction in supply voltage ($V_{DD}$ to $0.7V_{DD}$) results in up to 32% reduction in $I_{OFF}$ while lowering $I_{GATE}$ component by 84%. However, it also results in lower gate overdrive voltage and therefore reduces $I_{DSAT}$

by 48%. It is clear that using a lower power supply for performance critical circuits can result in unacceptable delay degradation. Therefore, a dual power supply arrangement is needed to minimize performance degradation while meeting the low power objectives.

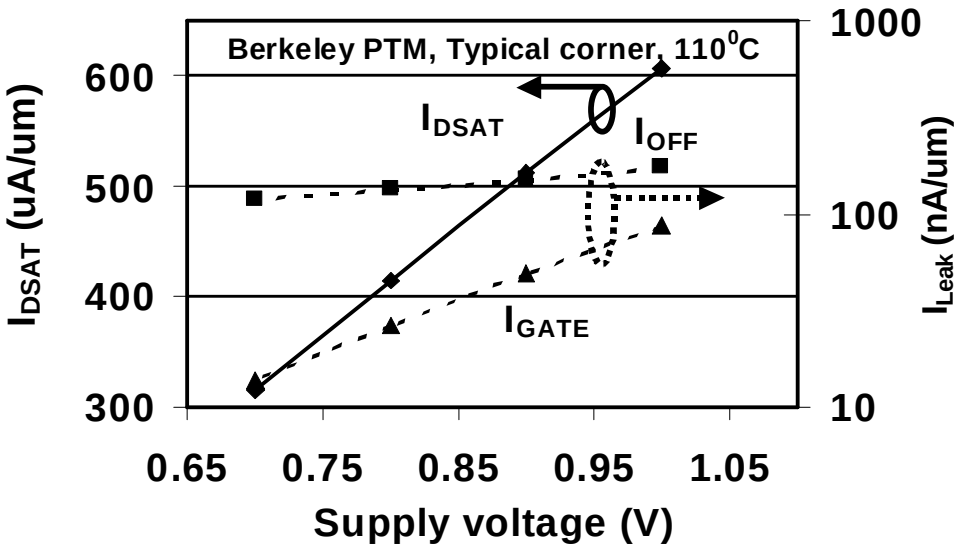


Figure 3.5: Impact of supply scaling on transistor current

## 3.3 Circuit Techniques for Low Power ALU Operation

In this section, we focus our attention on the circuit level strategies adopted in our design to achieve low power ALU operation. These techniques can be categorized as follows:

- Latches and flip-flops [23] that can support a reduced swing clocking scheme without

consuming static power, and

- Swing restored CPL (SRCPL) logic and $C^2$MOS MUX-es for designing the logic and shifter units to minimize switching capacitance and data driver sizes.

These techniques are used in conjunction with the dual supply clocking strategy to achieve low power ALU operation.

### 3.3.1 Latch and Flip-Flop Design for Dual Supply

High performance designs normally use static latches and master-slave FFs with transmission gate (TG) based circuits that have both clocked n-MOS and p-MOS transistors. The clock signal has a higher switching activity and is associated with high capacitance. Hence our goal was to reduce its swing ($0 \rightarrow V_{DDL}$) in order to lower clock related power consumption. On the other hand, the data circuitry of the entire ALU was maintained at a higher supply voltage ($0 \rightarrow V_{DDH}$) to minimize delay penalty. However, the TG based D latch and flip-flops cannot be used under such a dual supply clocking scheme. This is because the p-MOS transistors of the input TG do not fully turn OFF, resulting in static current (power) consumption. This condition can be better understood with the help of Figure 3.6.

In the event that the input data is different from the latched data, there is a full $V_{DDH}$ (high $V_{DD} = 1.8$V) voltage across the input transmission gate of the latch (or FF). However, when a dual supply clocking scheme is used with the clock transistors and buffers making a 0 to $V_{DDL}$ (low $V_{DD} = 1.3$V) transition, the p-MOS transistors of the input TG do not turn

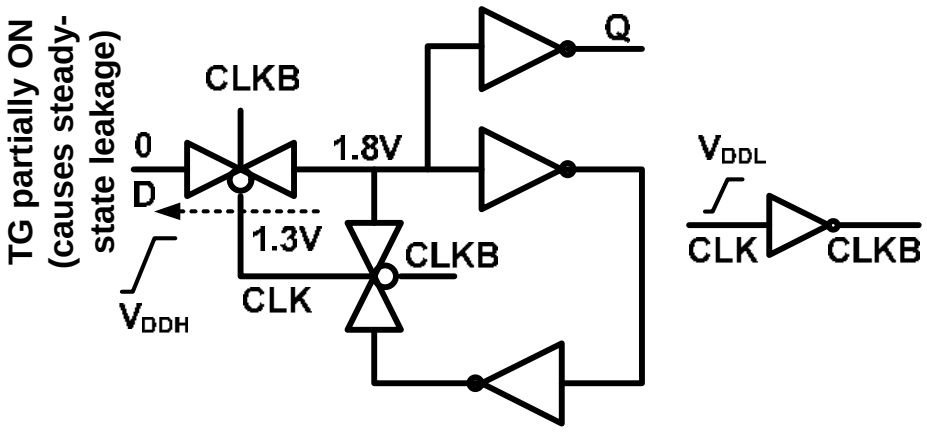


Figure 3.6: TG latch with steady state current problem under dual supply clocking

off fully. In fact, the gate-source voltage ($|V_{GSp}|$) across the p-MOS transistor is equal to the voltage difference between the clock and data network supply voltages ($V_{DDH}$-$V_{DDL}$). For the specific example shown in the figure this equals 0.5V for the 180nm technology and is approximately equal to the p-MOS transistor $V_{TH}$ resulting in exponentially higher steady state leakage current. This is a problem especially in high performance datapath designs that operate under conditions of high power density with local hot spots. This in turn leads to lower $V_{TH}$, and lends to higher $I_{OFF}$ and ALU power. The solution is to avoid clocked p-MOS transistors in the sequential circuits and replace the TG designs with the n-MOS only latch as shown in Figure 3.7:

Such a design supports low swing clocking, without consuming static power [23]. This circuit operates as an SRAM storage cell with single ended data. In this design one side of the storage cell is driven by an n-MOS pass transistor and is similar to a traditional

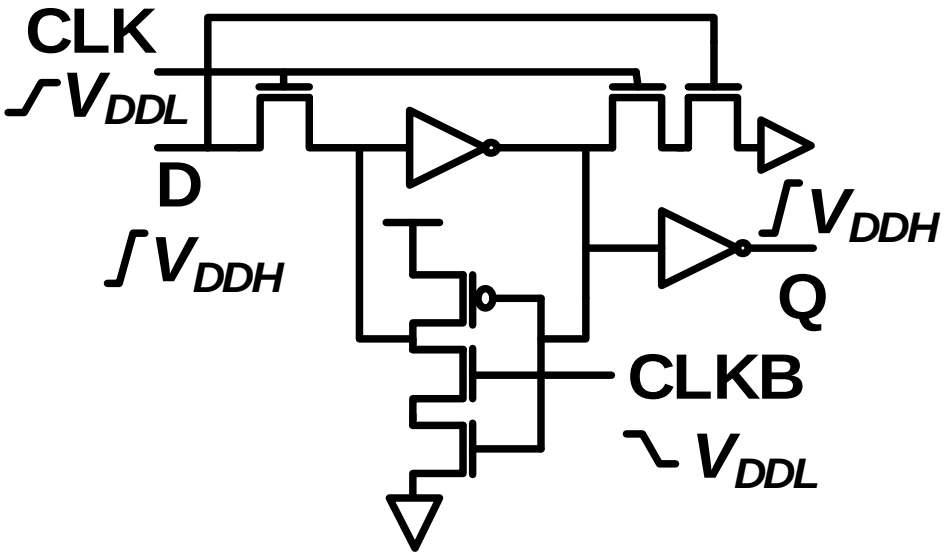


Figure 3.7: Latch circuitry to support dual supply clocking

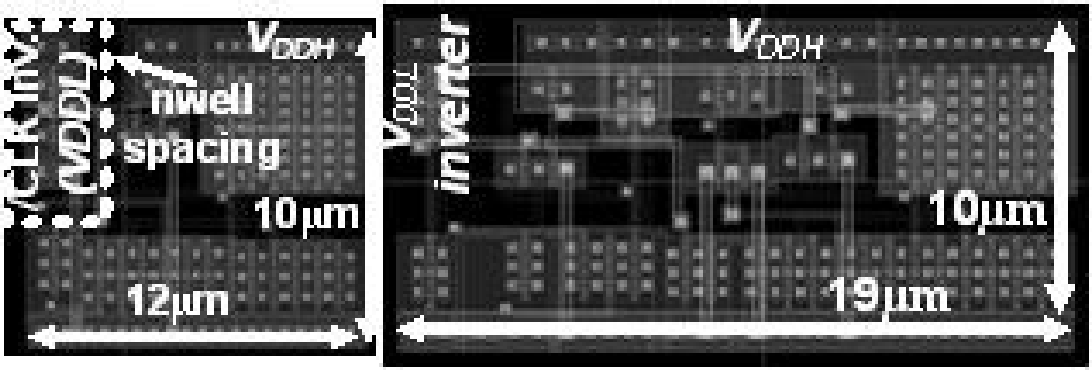


Figure 3.8: Dual supply latch and flip-flop 180nm layouts

SRAM cell, while the other end is connected to a 2-stack n-MOS pulldown.

The CLKB signal is generated locally using an inverter that operates from the $V_{DDL}$ supply. The master-slave flip-flops used in this ALU were designed by cascading 2 such

latches. Figure 3.8 shows the layout of the dual supply latch and flip-flop for the 180nm technology. We indicate the CLKB inverter operating from the $V_{DDL}$ supply and its separate n-well. Since these n-wells are not iso-potential, they are located further apart, resulting in 7% ( 3%) area overhead for the dual supply latch (flip-flop). It should be noted that the latch has 8 n-MOS and 3-p-MOS transistors and the x-dimension is determined by the total n-MOS width. This helps to limit the area penalty associated with the additional n-well required to implement the dual supply latch/FF design. The placement of the CLKB inverter is done so as to partition the cell and have easy access to $V_{DDH}$, $V_{DDL}$, and GND power rails when the bitslices are tiled for the 32-bit ALU.

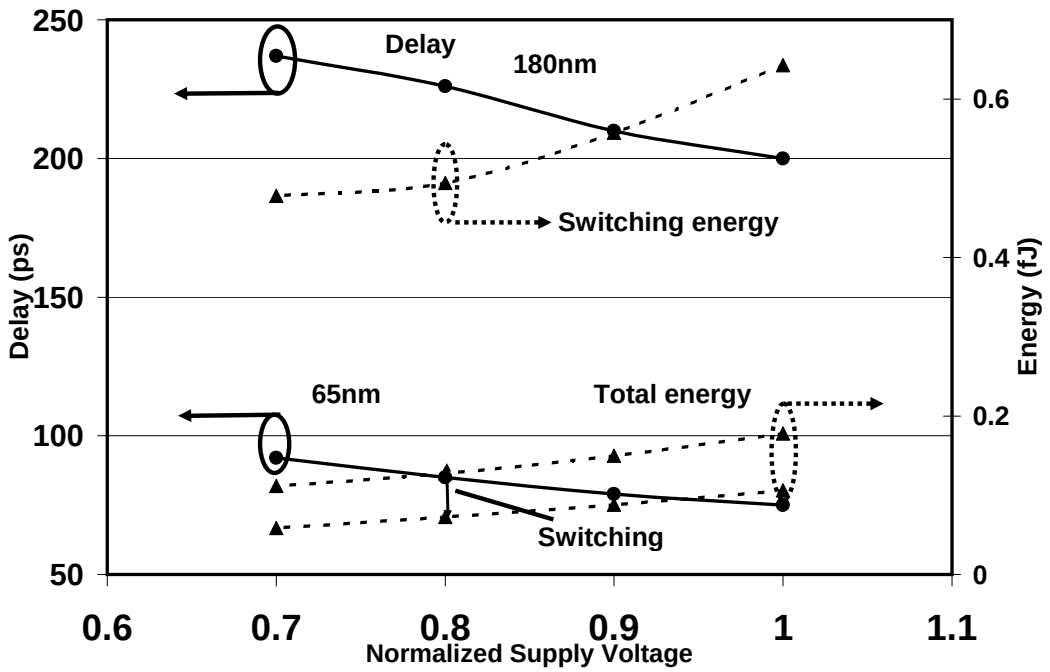


Figure 3.9: Energy and delay plots for dual supply latch

Figure 3.9 shows energy-delay plots for the dual supply latch for both 180nm and 65nm (Berkeley PTM) technologies. These data demonstrate that when the clock power supply is scaled ($V_{DDL}$), the latch total energy is reduced, while both the D→Q delay and $T_{setup}$ are degraded. There is a 26% (37%) reduction in total energy corresponding to a 30% supply scaling for the 180nm (65nm) technology. However, this is associated with a 19%-23% increase in the D→Q delay, while the $T_{setup}$ increased by ∼25ps for both technologies. In our ALU design we restrict the usage of such a dual-supply latch (flip-flop) scheme to only the non-critical units. This strategy allows us to absorb the additional delay penalty in the already existing timing slack.

### 3.3.2 Swing Restored CPL Based Logic Unit

The logic and shifter units constitute the non-critical blocks of the ALU. Therefore, we utilize the swing restored complementary pass transistor logic (SRCPL) to implement these units. SRCPL allows us to eliminate the p-MOS network required for a logic function when using the static CMOS style. This results in lower switching capacitance, smaller data buffer sizes and less area for the logic unit. A logic-shifter unit implemented using n-MOS pass transistors results in weak logic 1. Therefore in our design we used minimum sized output keepers to restore the CPL gate output signal to full swing (SRCPL). Figure 3.10 and Figure 3.11 show the circuit diagram and 180nm layout for a single bit-slice of the logic unit using SRCPL.

In our ALU we used SRCPL to implement the logic unit (INV, AND, OR, XOR) and a 5-bit shifter block. These were designed with minimum width pass transistors. Table 3.1

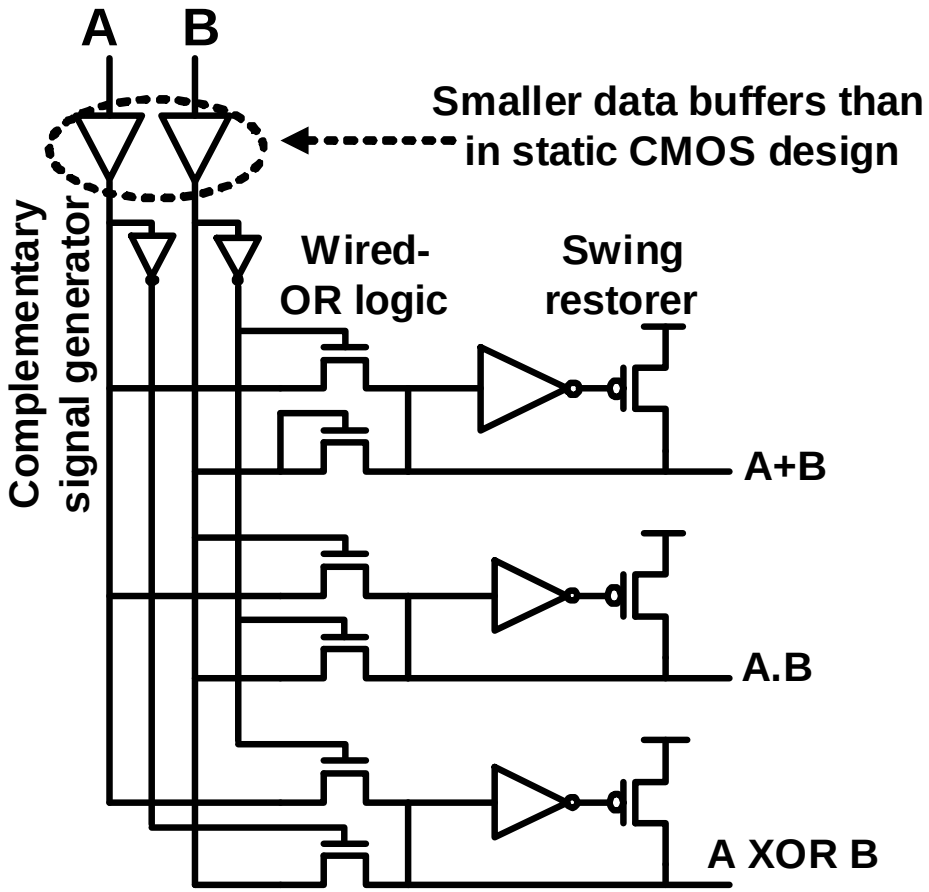


Figure 3.10: ALU bit-slice designed using SRCPL

compares the delay (in ps) for a bit-slice designed using SRCPL and static CMOS with identical input buffer and output load conditions. These results show that the SRCPL based design has a 12% lower delay compared to the worst-case performance (XOR) of the static CMOS implementation. In addition, our results show that the energy per bitslice

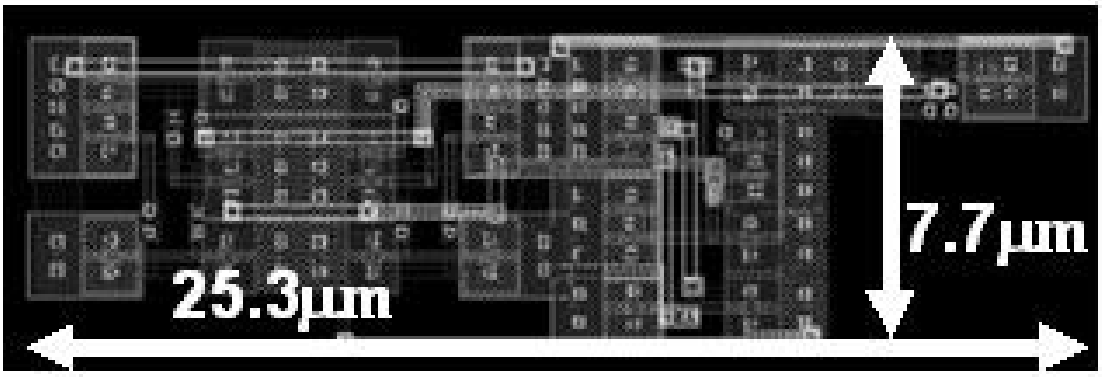


Figure 3.11: SRCPL based 180nm layout of ALU bitslice

for the SRCPL design is 1.26pJ as opposed to 1.79pJ for the static CMOS case. These results for both implementations correspond to an input vector pattern that considers all the different data combinations. They show that the SRCPL implementation results in 38% lower energy-delay product (EDP) and up to 30% lower switching energy compared to the static CMOS design without delay degradation.

Table 3.1: Static CMOS vs. SRCPL bit-slice performance (ps)

| Logic Function | Static CMOS | Swing Restored CPL |
|---|---|---|
| AND | 63 | 53 |
| OR | 64 | 84 |
| XOR | 110 | 97 |

## 3.4 ALU: Architecture and Circuit Design

In order to meet the constant demand for improved performance and lower power, several circuit and architectural design techniques have been proposed for high end microprocessors. As the transistor threshold voltage is scaled and number of pipelined stages are increased, the power dissipation in high performance microprocessors has roughly increased by three fold [25] every generation. Therefore, recent ALU designs that integrate a dual supply based approach are being investigated. There have been several recent designs that approach the issue of ALU power consumption and dual supply assignment [10], [23], [24], [25]. For example, the design reported in selectively uses a lower power supply voltage for p-MOS transistors using a shared n-well based design. In this case, the transistors that use the lower ($V_{DDL}$) and nominal ($V_{DDH}$) supplies both share the same n-well. This reduces the overall layout area overhead, but can also limit the voltage differential between the two supplies. A second design [24] uses a nominal supply voltage when the ALU operates using the 32-bit datapath, but in the 64-bit mode it uses an off-chip gated power supply to achieve low power operation. Both designs show lower than nominal power consumption (25%-33%) while resulting in 5%-25% delay degradation.

We now present the design overview of our proposed 32-bit ALU [33], [34] and demonstrate the impact of the different circuit techniques discussed in the earlier sections in ensuring low power operation. The basic ALU architecture is shown in Figure 3.12, and is similar to that reported in [24], [35]. In this design we use separate n-wells for the p-MOS transistors operating at different potentials. We also clearly partition the critical and non-critical units so that the dual supply operation can be carried out without performance

degradation. As will be shown subsequently, this approach is associated with a higher area penalty and less energy savings compared to the designs reported in [24], [25]. However, the energy savings obtained using our approach comes without any delay degradation and is therefore important in the context of high performance logic designs. The rest of this chapter discusses the design details, operation modes and scaling trends for important design metrics up to the 65nm CMOS technologies.

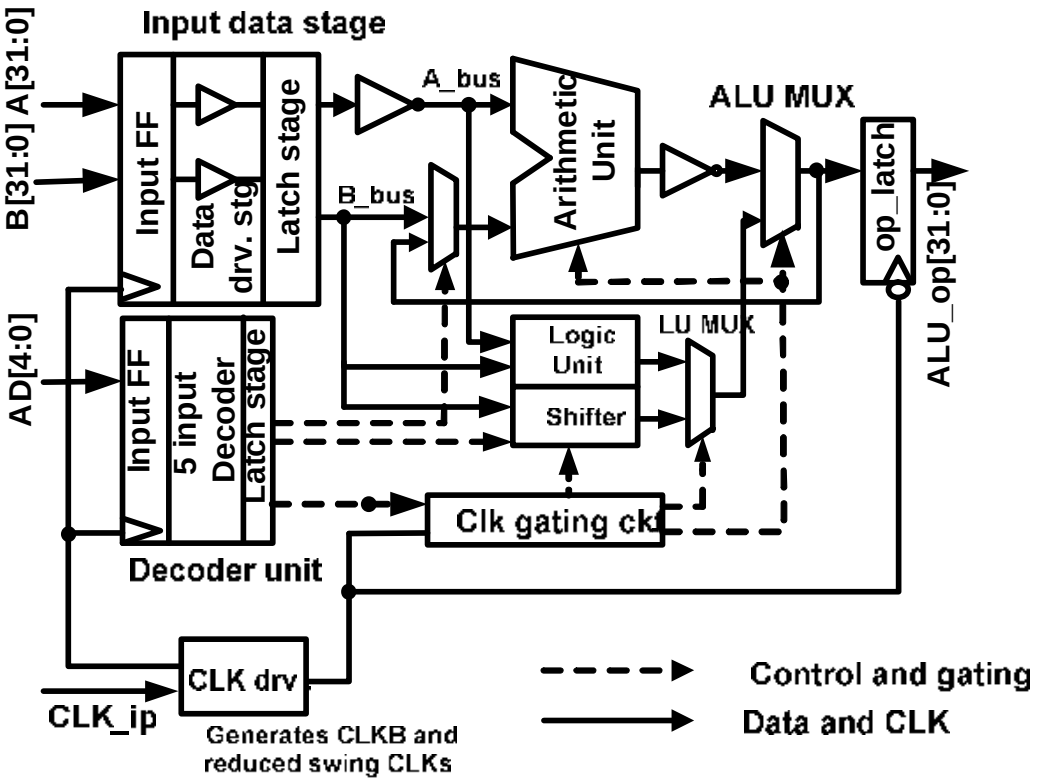


Figure 3.12: 32-bit ALU organization block diagram

### 3.4.1 Design of Decoder and Logic-Shift Units

As indicated in the block diagram shown in Figure 3.12, the ALU has several sub-units. The input stages for both the data buses are 32-bits wide and comprise data drivers and static master-slave flip-flops that support reduced swing clocking. The instruction decoder unit is implemented using 2-input static CMOS NAND-NOR logic gates and supports 15 different instructions. The design has two modes of operation: NORMAL and TEST that are selectable by setting the most significant bit of the decoder. In this chapter the focus is on the NORMAL mode of ALU operation. In the TEST mode a design for testability (DFT) scheme is used to detect and diagnose delay faults in the ALU. The details of the DFT scheme and TEST mode operation will be discussed in Chapter 5.

In the NORMAL mode, the ALU supports 32-bit unsigned addition, ALU loopback, logical operations such as invert, AND, OR, and XOR and shift of up to 5 bits. In addition to these instructions, the decoder also generates multiple internal clock gating signals by partially decoding the higher order input lines. This helps to stop clock toggling for deselected units and reduce overall power consumption. The decoder performance is non-critical, and its outputs use dual-supply latches to reduce switching energy. The A[31:0] and B[31:0] data buses drive the inputs of the performance critical 32-bit arithmetic unit as well as the non-critical logic and shifter units. We use bus-splitters for both A[31:0] and B[31:0] buses in order to reduce overall bus switching capacitance. The decoder unit generates the control signals for the bus-splitter based on the type of instruction being executed (arithmetic or logic-shift). The logic and shifter units are both implemented using SRCPL. The MUX-es at the output of the logic unit are realized using $C^2MOS$ logic

(instead of transmission gates) to avoid the usage of cascaded pass transistors that result in degraded signal rise/fall times and performance.

### 3.4.2 ALU Critical Path and Adder Design

The ALU critical path comprises the arithmetic unit (adder front-end MUX and 32-bit adder), and the ALU output MUX-es. In our design, these units were implemented using compound domino logic (CDL). The adder [36] is designed using a radix-2, Han-Carlson architecture with a sparse carry-merge tree (CMT). The adder output stage comprises parallel 4-bit static ripple carry adders (RCA). The carry inputs of the RCA are obtained from the CMT that generates every 1-in-4 carry signals ($C_3$, $C_7$, $C_{11}$, $C_{15}$, $C_{19}$, $C_{23}$, $C_{27}$, $C_{31}$). The adder input stage comprises a propagate-generate (PG) unit and uses footed clock transistors. However, subsequent stages of CDL gates are domino compatible and therefore do not require n-MOS clocked footer transistors.

The PG (propagate/generate) block and carry-merge-tree form the carry generate section while the Carry Select Adders (CSA) and the output MUX-es form the sum generate section of the adder. The PG block, carry-merge-tree and output MUX-es are the performance critical units and are implemented using CDL gates. The CSA adders operate in parallel with the carry generate section and are implemented using static CMOS gates. The PG block forms the propagate (P) and generate (G) terms based on the primary logic inputs A[31:0] and B[31:0] according to Eq. 3.16-3.17. The carry-merge-tree employs a binary merge algorithm that implements the recursive logic equation (Eq. 3.18), to produce the 1-in-4 carry signals ($C_3$, $C_7$, $C_{11}$, $C_{15}$ and so on) as shown in the simplified adder

architecture in Figure 3.13. The basic logic equations for the adder are given by:

$$P_i = A_i + B_i \tag{3.16}$$

$$G_i = A_i.B_i \tag{3.17}$$

$$C_i = G_i + P_i C_{i-1} \tag{3.18}$$

where $P_i$ and $G_i$ indicate the propagate and generate functions while $C_{i-1}$ represents the carry-merge operation. Also, $A_i$ and $B_i$ are the data for the $i^{th}$ bit position for the input A[31:0] and B[31:0] data buses. Figure 3.13 shows the simplified architecture of a 16-bit adder that can be extended to a 32-bit design. For each of the 4-bit adder blocks (Blocks A, B, C, D) shown in Figure 3.13, there are two parallel 4-bit CSA adders. One of these adders generates sum outputs assuming input carry to be logic 0 while the other assumes input carry to be logic 1. For example, the carry select adders pertaining to Block C, generate two sets of sum signals $S11:8^{(1)}$ and $S11:8^{(0)}$ using the corresponding block input carry $C_{in}^{c}$ to be logic 1 and logic 0 respectively. This happens in parallel with the carry-merge-tree and the CSA output signals become available at the inputs of the 2:1-MUX-es. When the $C_7$ signal becomes valid, the appropriate sum signals are selected and become available at the S11:8 primary outputs of the adder.

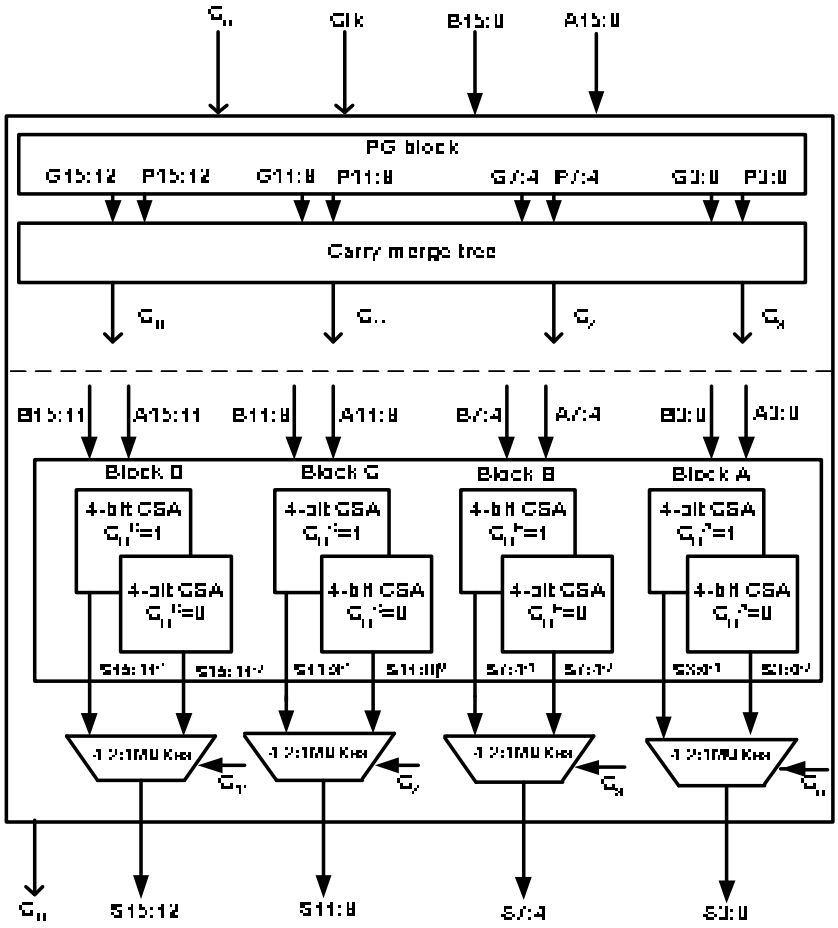


Figure 3.13: Simplified architectural overview of adder

### 3.4.3 Dual Supply Design and Assignment

In this design we partition the ALU into different power supply domains to reduce the performance overhead associated with the dual supply clock design scheme. The lower supply ($V_{DDL}$) is used for the clock signal of non-critical units only. This allows significant

energy reduction since clock is a high activity signal and is associated with large drivers and capacitance. The entire data network is maintained at the nominal ($V_{DDH}$) supply for the entire design, mitigating the problem of excessive delay degradation. In addition, performance critical units like the 32-bit adder and ALU output MUX-es are clocked using full-swing signals. As a result, the delay penalty associated with the dual supply clocking is limited to only the latches and FFs of the non-critical units and can be absorbed in the existing timing slack. Figure 3.14 shows the layout of our 32-bit ALU in 180nm bulk CMOS, 6 metal layer TSMC process and measures $800\mu m$ x $600\mu m$. The different logic and functional units are numbered as follows: 1 and 2 represent the scan chains for input and output data buses, 3 is the logic-shift unit, 4 is the input data stage for A[31:0] and B[31:0] buses along with the bus-splitters and drivers. The unit 5 represents the 32-bit adder core and ALU output MUX-es, 6 is for the decoder unit and the clock driver section. Finally 7 is for the DFT unit (discussed in detail in Chapter 5), the clock gating circuits and additional distributed clock drivers for the adder and ALU output MUX-es.

Our dual supply assignment (CLK: 0 to $V_{DDL}$; DATA: 0 to $V_{DDH}$) is restricted to units 3, 4, and 6 thereby indicating a clear partitioning between units using single and dual power supply regions. The low swing clock is derived from the full-rail input CLK signal (in unit 6) using a buffer stage operating from the $V_{DDL}$ power rail. The adoption of a dual supply clocking strategy increases the area of the input data stage by 3%, decoder unit by 2% and logic-shift unit by 3%. The DFT unit increases the ALU transistor count by 1.3% and along with the dual supply assignment and routing results in increased layout area. On the other hand, the SRCPL based logic-shifter unit helps reduce active area, thereby

limiting the overall area overhead associated with our design techniques to ∼4%.

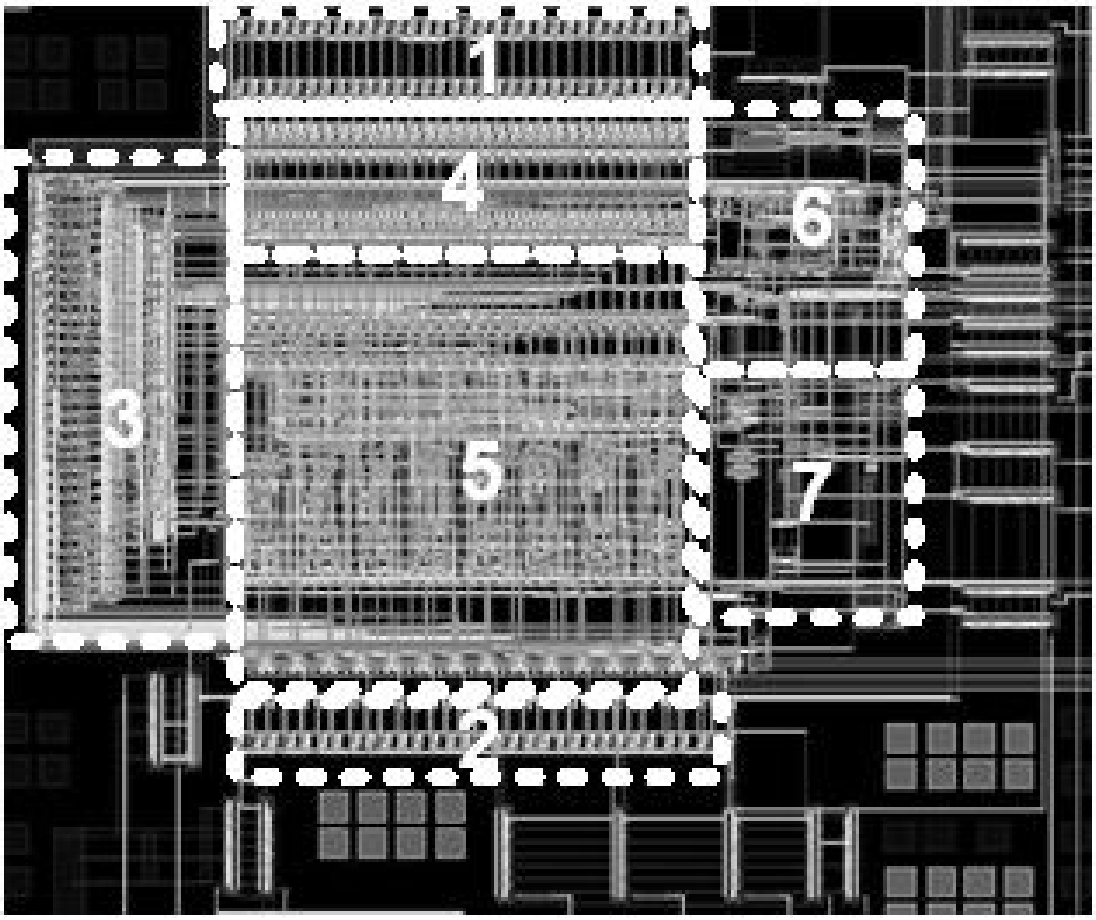


Figure 3.14: 32-bit ALU layout in 180nm CMOS technology

## 3.5 ALU Energy, Delay, Scaling Trends: Results

We now present the results for ALU performance and discuss its scaling trends for sub-180nm CMOS technologies. This section deals with some of the important ALU design

metrics such as worst-case delay, standby and active mode power, and peak and average current demand in the NORMAL mode of operation. We also show the scaling trends for these parameters for the 180nm-65nm CMOS technologies.

### 3.5.1 ALU Performance for Sub-180nm Technologies

The ALU worst-case delay corresponds to the maximum clock frequency at which it can operate. This is of special importance in high performance logic designs and determines the overall data throughput of the design. The worst case delay is obtained at an elevated temperature of $110^0C$ for the vector A[31:0]=FFFFFFFFH and B[31:0]=00000001H. The data points in Figure 3.15 for the 180nm technology correspond to a bulk CMOS TSMC process while the 130nm-65nm results were obtained using the Berkeley Predictive Technology Models (BPTM). Our results demonstrate that the 180nm 32-bit ALU operates at 1.5GHz and can be scaled to 4.2GHz for the 65nm CMOS technology.

### 3.5.2 ALU Energy for Sub-180nm Technologies

As the technology is scaled, we observe savings in switching energy due to reduction in the $CV^2$ product. However, there is also an exponential increase in the total leakage energy due to higher $I_{OFF}$ current. It is expected that this will result in an increase in both the total as well as the standby power (energy) consumption of DSM datapath designs. Figure 3.16 shows the normalized plots for the total energy for two different cases: Design 1 (Ref. ALU), and Design 2 (dual supply + SRCPL). Design 1 operates entirely at $V_{DDH}$, and uses static CMOS gates for its logic-shifter unit. Design 2, on the other

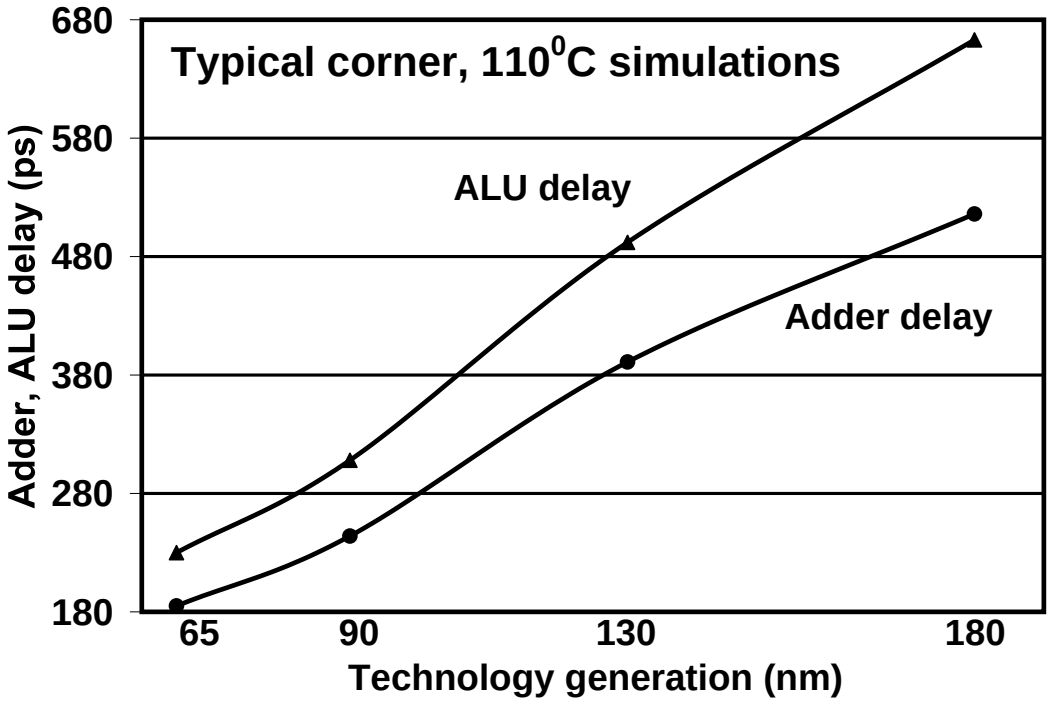


Figure 3.15: 32-bit ALU NORMAL mode performance and scaling trends

hand, demonstrates the energy savings obtained using SRCPL based logic-shifter units with reduced swing clocking. When the power supply voltage is lowered by ∼30% ($V_{DDL}$ = $0.7V_{DDH}$) there is 18%-24% savings in total energy for the 180nm-65nm technologies.

The energy plots in Figure 3.16 deviate from the traditional $CV^2$ scaling trends (especially for the 65nm generation) due to the dominance of subthreshold and gate leakage current components in DSM technologies. The results in Figure 3.16 were obtained for a data activity $\alpha$ of 0.1 in the NORMAL mode of ALU operation. It should be noted that the energy reduction for Design 2 was obtained without delay degradation. This was

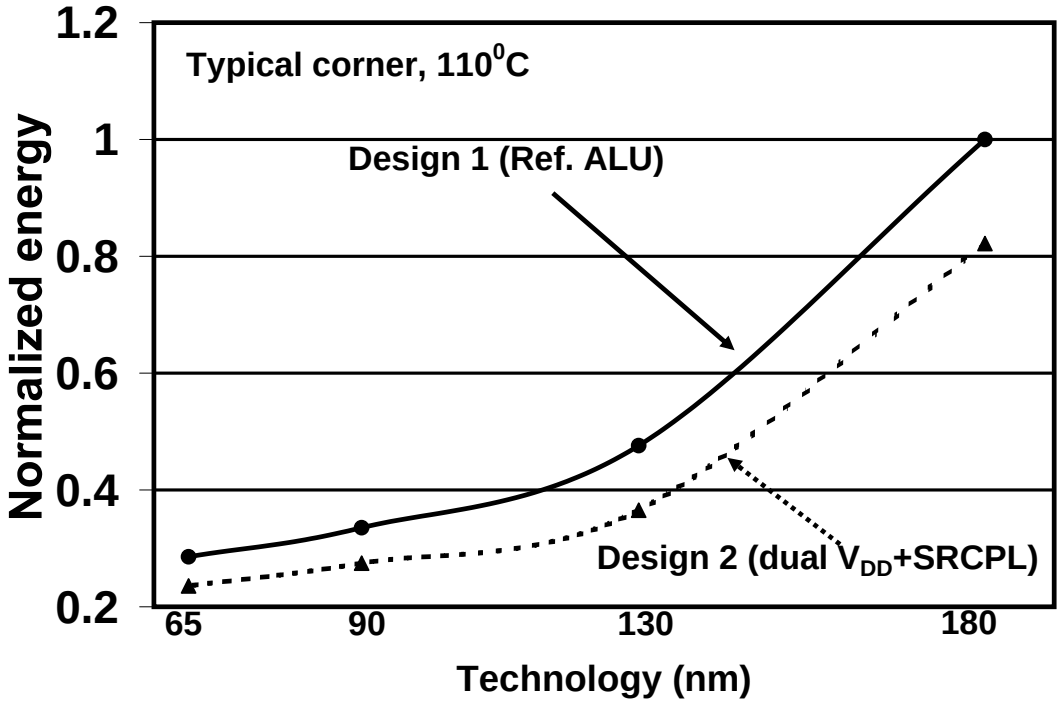


Figure 3.16: 32-bit ALU NORMAL mode total energy and scaling trends

possible because of the dual supply assignment strategy followed in this design, whereby all the critical unit data and clock signals were maintained at $V_{DDH}$.

### 3.5.3 ALU Standby Power and Current Demands

The adoption of a dual-supply clocking scheme helps us to reduce both active mode switching energy and standby power (no data or clock activity). In addition, the CPL based logic-shifter unit with smaller data drivers helps in reducing the overall switched capacitance. This translates into lower standby power for the 32-bit ALU design as can be seen

from the data in Figure 3.17.

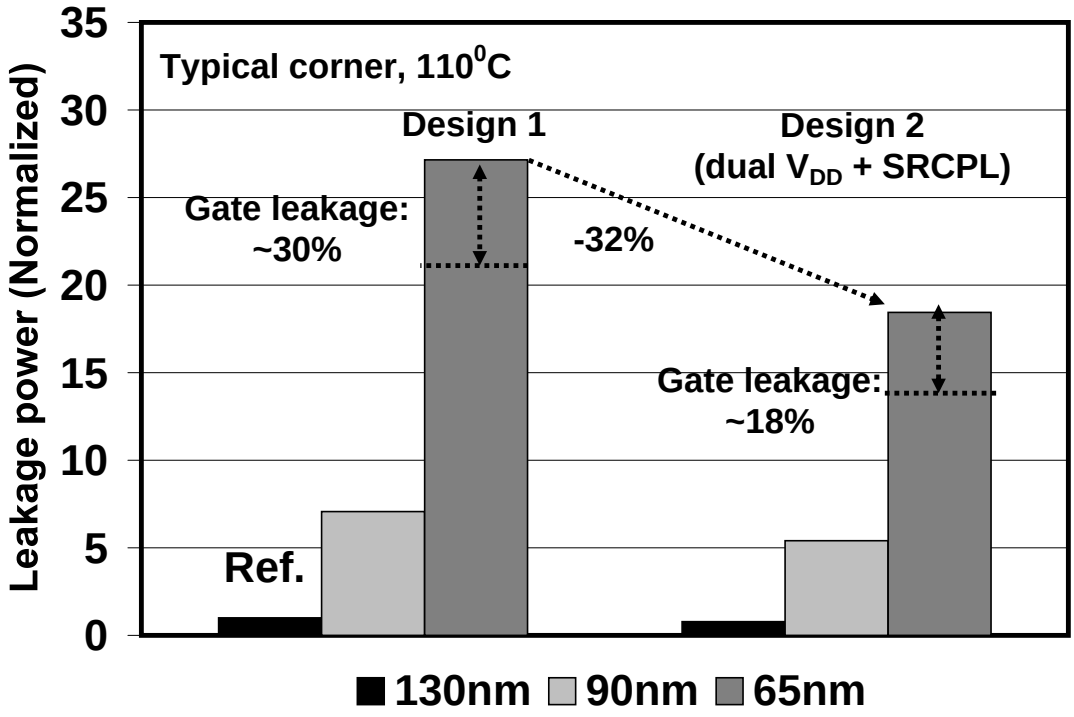


Figure 3.17: ALU standby mode power consumption

Our results show that the standby mode leakage power for the 180nm technology is negligible compared to that of the sub-130nm generations. Therefore, we use the 130nm ALU (Design 1) total standby power as our reference. These results demonstrate that when Design 1 is scaled from 130nm to 65nm technology, there is a 27x increase in the standby mode leakage power. Furthermore, it is estimated (HSPICE, level 54 simulations) that, for the 65nm technology node, gate leakage may account for ∼30% of the total leakage power. However, the total standby power for Design 2 (dual supply + SRCPL + scaled buffers) is 22% (32%) lower than the 130nm (65nm) generation. Furthermore, the gate leakage

component reduces significantly (∼40%) when the power supply is lowered for Design 2, and contributes to ∼18% of the total ALU leakage power for the 65nm generation.

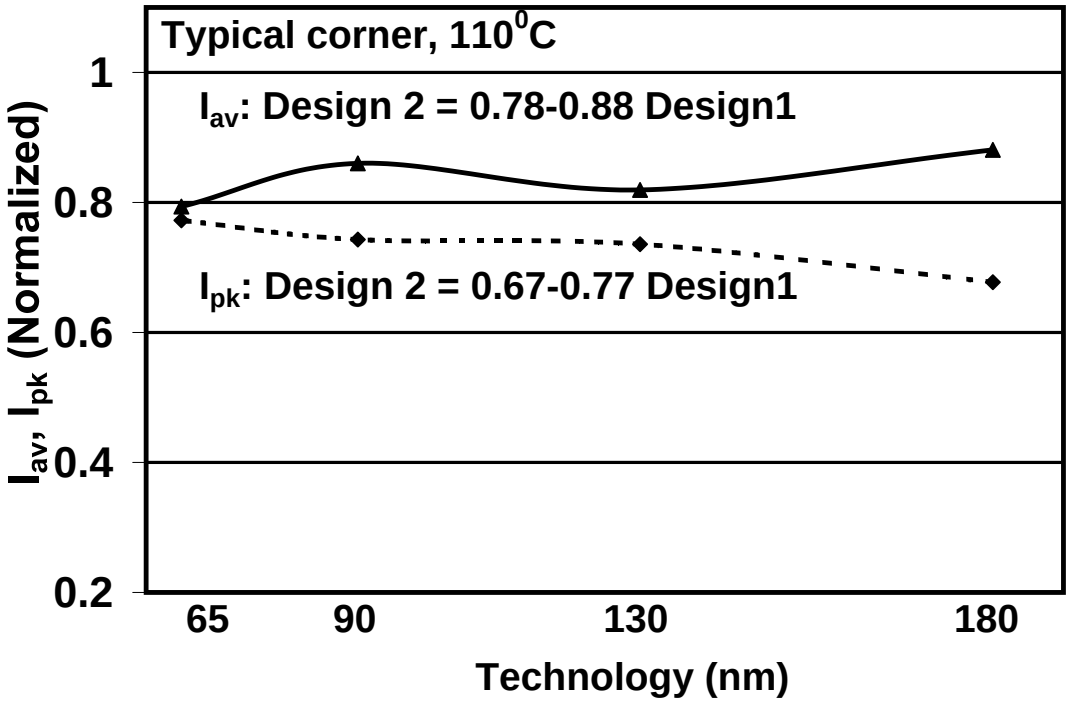


Figure 3.18: ALU peak and average current demands

Finally, we demonstrate the impact of the dual supply design on both peak and average current demands. The reductions in active and standby mode energy/power translate into reductions in the overall current demand. Figure 3.18 plots the Design 2 (dual supply) current demands normalized with respect to Design 1 for different technologies. It is clear that the low power circuit techniques result in 23%-33% reduction in $I_{pk}$, and 12%-21% lower $I_{av}$ demands. It is expected that this can help in lowering the possibility of localized Joule heating related hot spot occurrences (average current) and IR drop related

performance degradation (peak current) in DSM technologies.

## 3.6 Summary

In this chapter, we presented a high performance 32-bit ALU design and adopted a dual supply strategy to minimize total energy consumption. We discussed the partitioning of the performance critical and non-critical logic units as well as the usage of a latch/FF circuit that can support dual supply clocking. We demonstrated the scaling trends for the 180nm-65nm CMOS technologies showing reductions in ALU total energy (18%-24%), leakage power (22%-32%), and peak current (12%-21%) demands without delay penalty. In the subsequent Chapters 4 and 5 we will discuss circuit and DFT techniques that can be used in high performance ALUs and RFs to prevent leakage induced noise margin degradation and detect delay faults.

# Chapter 4

# Designing Robust Wide-Domino Logic for High Performance Datapaths

Domino logic has been the mainstay of high performance circuit and datapath designs. They are used in the implementation of microprocessor critical path units such as adders, ALUs and register files. In the previous chapter we presented the design of a low power, dual supply based 32-bit ALU. In this chapter we will discuss the design of leakage tolerant and robust wide domino logic gates for ALU front-ends and register files. Register files (RFs) are used extensively in high-end superscalar microprocessors [8], [37], [38]. They are high-speed, single-cycle, datapath units that provide the operands for the processor's integer and floating-point execution cores. The RF array depth and number of read-write ports increase with processor performance and number of on-die execution cores. The RF

read operation is timing critical during which an array is selected and full-rail data is read out by the processor [35], [39]. Therefore, the read circuitry is typically designed using multiple stages of cascaded wide-OR compound domino logic (CDL) gates that operate as high performance MUX-es. In addition to RF read ports, wide-dominos are also used to design the ALU front end MUX-es. This is because the data processing units can accept data from multiple sources that include the RF, cache, their own loopback bus, or local buffers depending on branch prediction results. Since they are on the processor critical path, these MUX-es are normally implemented using dynamic logic.

As CMOS technology is scaled, the domino logic noise margin is degraded substantially due to exponentially higher leakage current ($I_{OFF}$), lower power supply voltage and capacitance of the scaled devices [11], [40], [41]. This problem is further compounded by noise induced due to high switching frequencies and ground/$V_{DD}$ bounce. The wide-OR domino circuits are used for designing local and global bitlines (LBL, GBL) and are an integral part of the RF read port and ALU front-ends. However, these circuits are especially susceptible to leakage induced logic upsets in sub-130nm technologies because of the existence of multiple parallel pulldown paths. Several design strategies [12], [38], [41],[42] have been advanced in order to maintain iso-robustness scaling of high-performance RFs. Some of these include keeper-based techniques like:

- Upsized keepers,
- Conditional keepers, and
- Forward body-biased (FBB) keepers.

Other schemes that can suppress the n-MOS pulldown leakage and improve bitline noise margins [8], [18], [21], [43] include using:

- Pseudo-static schemes,
- Dual $V_{TH}$,
- Longer $L_{drawn}$, and
- Reserve body bias (RBB).

The challenge is to maintain the wide-OR domino/RF robustness, minimize the delay penalty and prevent further bitline fragmentation.

## 4.1 RF Organization: A Simple Example

Wide bit-width register files (RF) are performance-critical components of microprocessor integer/FPU execution cores and require single cycle read/write latency. In this section, we present the architectural overview of a 2-read, 1-write ported 256-entry 64-bit high performance RF whose organization is shown in Figure 4.1 [39]:

Each of the 256 RF entries is uniquely selected using an 8:256 decoder scheme that generates the read and write select (RS/WS) signals. In order to read/write from an entry, only one WL driver signal/port switches high (active) while the rest of the 255 drivers are inactive and are leaking. The read port word line (WL) drivers are on the RF critical path and are hence upsized to drive the 32 local bit-cells on each side of this partitioned RF. This results in increased leakage for the WL drivers of the 255 deselected entries. For

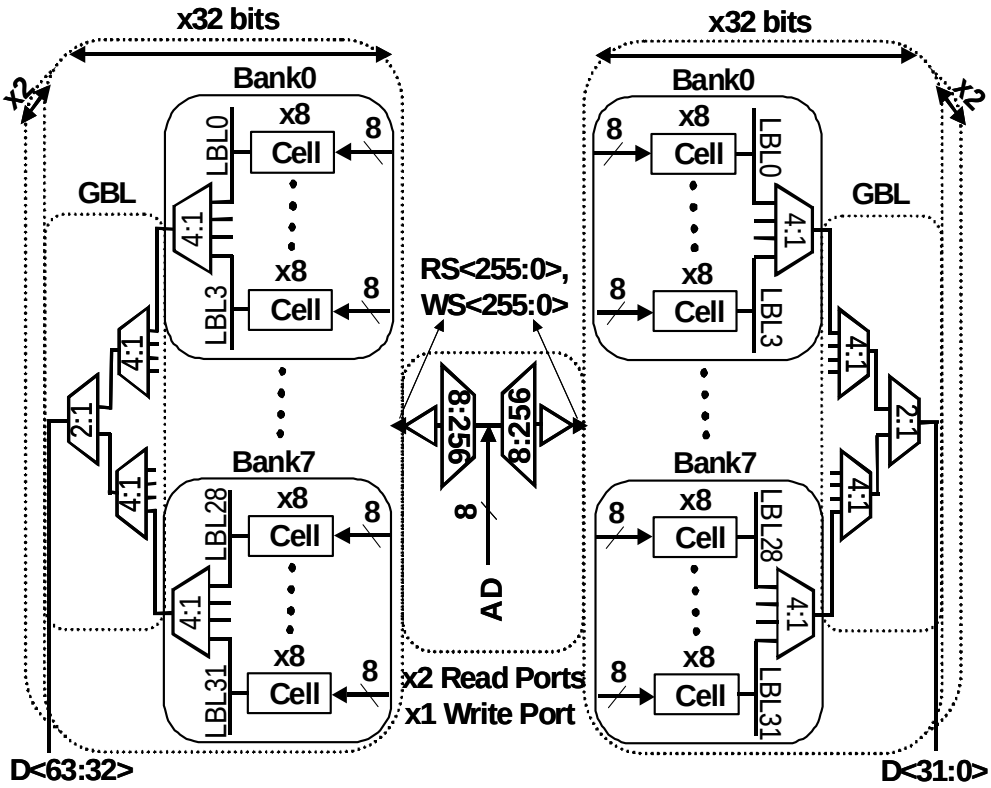


Figure 4.1: Organization of 256-array, 64-bit RF

high-performance RFs with large array depths and word sizes, the local and global bitlines (LBL, GBL) contribute significantly to the total energy. In addition, at any given time only one array is selected while the remaining arrays are deselected contributing to active leakage power. Therefore, in sub-130nm technologies leakage control schemes can help achieve both low power operation and robust (scalable) RF designs. In this chapter, we investigate the selective usage of the following techniques in high performance LBL and GBL designs [7], [18], [21]:

- Dual $V_{TH}$ technology,
- Longer $L_{drawn}$ transistors, and
- Reverse body bias (RBB).

We study the impact of the above techniques on the following design parameters: propagation delay, leakage and total energy, DC robustness, AC noise margin and overall area. Our results also show that a significant portion of the RF energy is consumed during the precharge phase. Therefore, in this work we present a circuit scheme to minimize the precharge contention and RF energy.

## 4.2 RF and Wide-OR Domino MUX-es, Iso-Robustness Scaling

We now discuss the organization of a 64-array, 32-bit wide RF and the implementation of wide-OR domino based LBL and GBL circuits. In addition, we introduce the concepts of AC and DC noise margins for wide dominos and demonstrate their degradation with scaling. This section also discusses some of the strategies that have already been proposed in the literature for designing leakage tolerant wide dominos. It should be noted that although the rest of the chapter deals with the design of robust LBL and GBL circuits for high performance RFs, the same concepts and tradeoffs are valid in the case of the domino wide-MUXes used in ALU front-end and output MUX-es.

### 4.2.1 RF Read Port and LBL, GBL Designs

Figure 4.2 shows the performance critical read port for a single bit-slice of a 64-array, 32-bit wide RF. This a scaled down version of the RF organization shown in the previous section but demonstrates the same noise margin and robustness related problems. We use this as our vehicle to study the different design tradeoffs involved in the design of robust domino logic gates. The design shown in Figure 4.2 comprises the read port decoder, word line (WL) driver unit, dynamic LBL and GBL and output buffer/latch section. The 6:64 decoder is used to uniquely select an array during the read operation. The decoder's timing is non-critical and therefore it is implemented using 2 input static NAND/NOR logic gates. The first stage of the WL driver section is implemented using a footed domino gate while the subsequent WL drivers are static inverters. When CLK=0, the RF intermediate nodes A, B, D (Figure 4.2) are precharged to $V_{DD}$ and all the read select signals ($RS_0$-$RS_{63}$) are 0. As a result, the LBL and GBL pulldown paths are cut-off and static power free operation is ensured.

During evaluation (CLK=1), the active high decoder output (Dec0) selects the RF array location that is to be accessed. This causes the word line (WL) driver section to evaluate and node A makes a $1 \rightarrow 0$ transition. Depending on the setup data ($D_0$) the LBL and GBL can evaluate resulting in nodes B and D making $1 \rightarrow 0$ transitions and the array data being read out. It should be noted that the keepers for the WL driver, LBL and GBL have been omitted in Figure 4.2 for the sake of clarity.

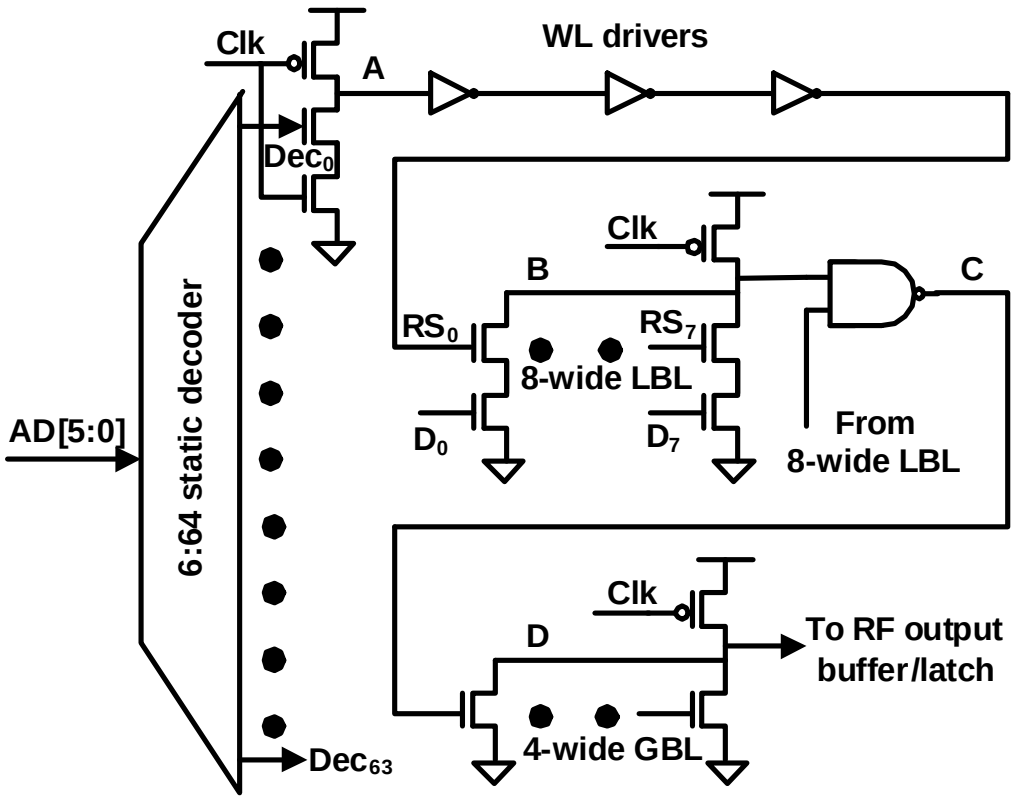


Figure 4.2: Critical read path of a 64-array RF

### 4.2.2 LBL and GBL: Design and Noise Margin Issues

Wide-OR dominos are used to design local and global bit lines (LBL, GBL) of high performance RFs. These LBL and GBL circuits shown in Figure 4.2 operate as high performance 8:1 and 4:1 domino MUX-es, respectively. The inputs to the pulldown network are domino compatible allowing for the removal of the clocked n-MOS footer transistor. This reduces the stack height, improves performance and lowers switching energy (LBLs are 2 n-MOS

stack, GBLs have single n-MOS pulldown). The domino logic gates typically use 3%-5% p-MOS keepers to hold the precharge value during evaluation phase in case the array is deselected or the input data = 0. Only one of the parallel pulldown paths is selected at a given time. As technology is scaled and the transistor $I_{OFF}$ current increases exponentially, the p-MOS keeper has to be upsized to ensure iso-robustness. However, this leads to increased contention and delay degradation requiring further bitline fragmentation or the incorporation of leakage control techniques into the LBL and GBL structures.

We use DC robustness and AC noise margin as metrics for comparing the effectiveness of different leakage tolerant domino designs. The DC robustness is defined with respect to node C (Figure 4.2) and is obtained under worst-case leakage conditions when the inputs $RS_0$-$RS_7$ are subjected to DC noise (simulated using a slow ramp signal). The voltage at which node C and read-select (RS) signals are equal is identified as the unity gain noise margin (UGNM) point. DC robustness for a given technology is defined as the normalized UGNM (UGNM/$V_{DD}$). The concept of DC robustness for wide-dominos is better explained with the help of the waveforms shown in Figure 4.3. This definition for DC robustness can be used for both LBLs and GBLs, and is well established in the context of leakage tolerant domino logic designs [8], [12], [40], [41].

In order to quantify the AC noise margin we subject the read selects ($RS_0$-$RS_7$) to a triangular waveform with equal rise and fall signal slopes of 10mV/ps. The peak of the triangular waveform is increased parametrically, until the subsequent dynamic gate (node D) evaluates, indicating a noise-induced logic failure. The AC noise margin accounts for effects such as switching transients ($CdV/dt$), coupling with neighboring metal lines

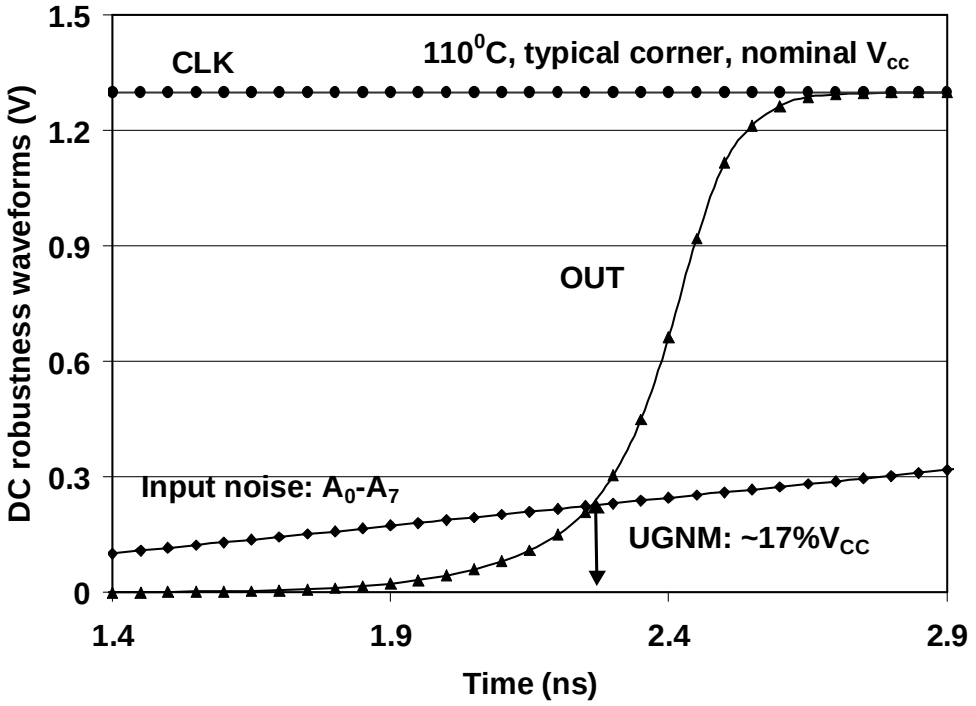


Figure 4.3: Waveforms for Wide-OR domino DC robustness

and impact of internal node capacitances that are neglected during the DC noise margin analysis.

We now show data for the LBL noise margin for sub-130nm technologies. Figure 4.4 shows results for an 8-wide, low $V_{TH}$ LBL with 5% p-MOS keeper. Simulations indicate that the 130nm design has a DC robustness of 17% and normalized AC noise margin of 36% (470mV). We use this as the target robustness for the designs in 90nm and 65nm technologies. This allows us to compare different design techniques and quantify their tradeoffs. It is possible to set the robustness threshold at a different absolute value, but the

energy-delay tradeoffs and scaling trends would remain unaffected. When the technology is scaled, the transistor $V_{TH}$ is lowered causing $I_{OFF}$ to increase by 3x-5x per generation. As a result, there is 35% (47%) degradation in DC robustness for the 90nm (65nm) technology (Figure 4.4). The AC noise margin also degrades by 11% and 17% for the corresponding technologies. It should be noted that the data for both the 130nm and 90nm technologies correspond to all low-$V_{TH}$ designs while that for the 65nm corresponds to that of a high-$V_{TH}$ pulldown. This is because a low $V_{TH}$ 65nm design does not have a UGNM crossover point and fails to operate due to excessive transistor leakage.

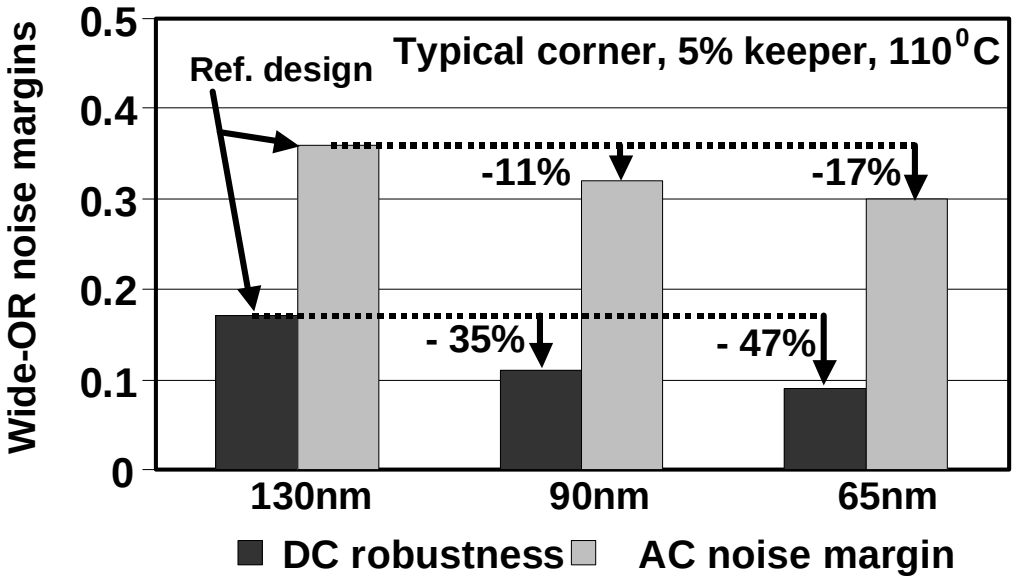


Figure 4.4: Wide OR domino noise margin and scaling trends

It is clear from these results that increased transistor $I_{OFF}$, lower capacitance of scaled devices, and higher switching frequency will significantly degrade the noise margin of high

performance domino logic gates. Therefore, we need to explore alternate design/leakage control techniques that improve circuit robustness while maintaining performance and low power RF operation.

### 4.2.3 Robust Wide-Domino Schemes

In this section, we focus on some of the techniques that have been reported in the literature [8], [11], [12], [38], [41], [42] for designing robust wide-OR dominos. These different schemes can be broadly categorized as:

- Keeper based, and
- Pulldown leakage reduction based approaches.

The keeper-based approaches strengthen the p-MOS pullup keeper and counter the n-MOS network leakage. Some of these techniques include keeper upsizing, forward body-biased keeper and conditional keepers. On the other hand, techniques that reduce the n-MOS network leakage include a pseudo-static scheme, dual $V_{TH}$, longer $L_{drawn}$ and reverse body-bias (RBB).

*Keeper upsizing*: When low $V_{TH}$ LBL and GBL designs are scaled from 130nm to 65nm the p-MOS keeper needs to be upsized 9x-10x to meet the AC and DC noise margin thresholds. For the 65nm technology, this amounts to a p-MOS keeper that is 45%-50% of the entire n-MOS pulldown width. This leads to excessive delay degradation (68%), contention power and can cause the bitlines to not have full-rail switching. Designers have resorted to further bitline fragmentation to avoid such a situation. However, this

adds to the number of cascaded logic stages and increases both the switching energy and overall area. Simple keeper upsizing is therefore not a suitable technique for designing high performance, robust, wide-OR dominos for sub-130nm technologies.

*Forward body-bias*: Another approach for improving wide-domino robustness is the forward body-bias (FBB) technique. In this technique, the n-well of the p-MOS keeper is connected to a potential lower than $V_{DD}$. This forward biases the keeper source-body junction and reduces its threshold voltage. This results in a higher ON current ($I_{keeper}$), thereby strengthening the domino node and its noise margin. However, the effectiveness of this method is limited by the allowable FBB voltage and sensitivity of $|V_{THp}|$ voltage to $V_{SB}$ (depends on body bias coefficient, $\gamma$). As the forward body bias voltage is increased, both the p-MOS keeper source and drain p-n junction diodes turn on resulting in static current. The simulation results in Table 4.1 for a 65nm p-MOS transistor show that it is possible to operate at a FBB voltage ($V_{SB}$) of 400mV. Beyond this, the $I_{ON}/I_{OFF}$ ratio degrades significantly, resulting in higher total and leakage energies for both the LBL and GBL designs.

*Conditional keeper technique*: Conditional keeper based design approaches have been proposed that can improve the DC robustness while minimizing the delay penalty and contention during evaluation. In this scheme, a parallel keeper is used in addition to the normally ON weak keeper. The extra keeper is turned on during the evaluation phase only if the LBL/GBL does not evaluate. The basic circuit scheme and timing diagrams for conditional keeper based wide-OR dominos are shown in Figure 4.5 and Figure 4.6.

Table 4.1: FBB characteristics for 65nm p-MOS transistors

| FBB voltage | $I_{ON}$ $\frac{\mu A}{\mu m}$ | $I_{OFF}$ $\frac{nA}{\mu m}$ | $I_{ON}/I_{OFF}$ |
|---|---|---|---|
| 0V (Ref) | 391 | 72 | 5430 |
| 100mV | 405 (1.04x) | 114 (1.58x) | 3552 |
| 200mV | 418 (1.07x) | 173 (2.4x) | 2416 |
| 300mV | 429 (1.1x) | 247 (3.43x) | 1736 |
| 400mV | 439 (1.12x) | 340 (4.72x) | 1291 |
| 500mV | 448 (1.15x) | 452 (6.28x) | 991 |

The condition generation circuitry requires additional logic gates and a delayed clock that is obtained from the system clock signal using a chain of static inverters. This increases the overall switching energy and domino node capacitance. Furthermore, the delay chain has to account for the WL driver delay and build in a safety margin so that the conditional keeper is turned on only after the worst-case LBL evaluation time. During this interval, a weak keeper holds the precharge value making the wide-OR domino design prone to logic upsets [44]. Our results show that the AC noise margin for a 65nm conditional keeper based design is only 19% indicating 47% degradation compared to the 130nm reference design as indicated by the waveforms shown in Figure 4.7.

*Pseudo-static technique*: The pseudo-static technique has been advanced as a means for designing robust wide-OR domino logic gates for deep submicron (DSM) technologies. The pseudo-static technique improves the wide-domino UGNM by introducing additional p-MOS transistors to force the LBL stack nodes to $V_{DD}$. Figure 4.8 shows the pseudo-static

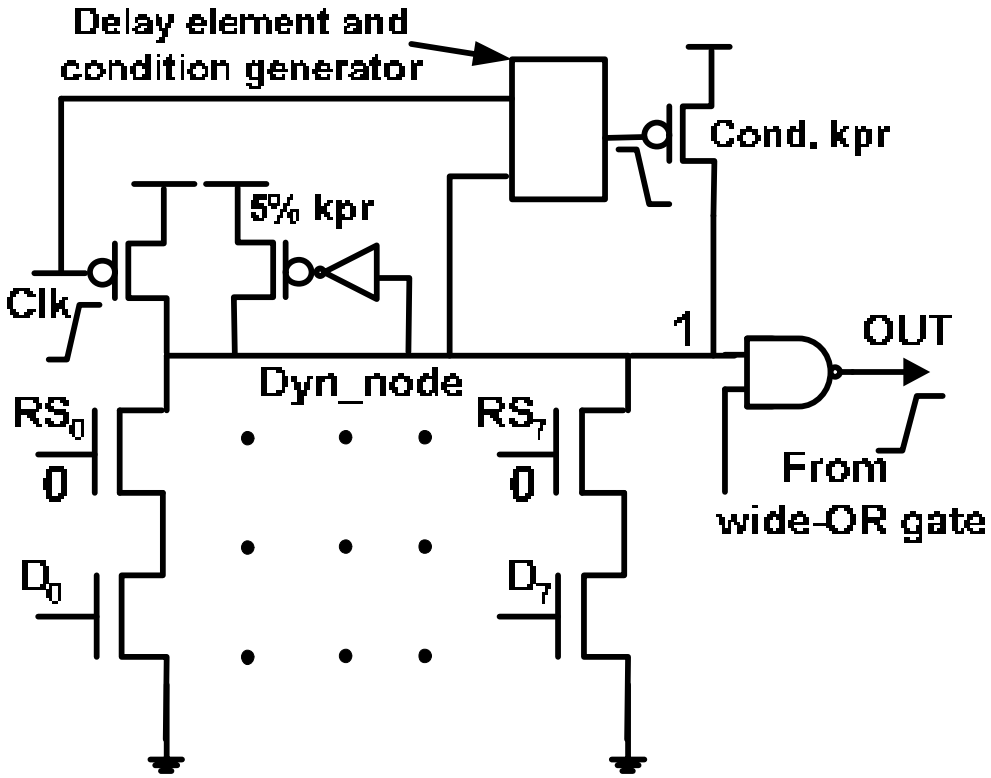


Figure 4.5: Organization of conditional keeper based wide-OR dominos

technique based wide-OR domino circuit scheme.

This scheme ensures that both transistors in the n-MOS stack are OFF, N2 has a higher "effective" threshold voltage (reverse body bias and reduced DIBL effect) and a negative $V_{GS}$ bias voltage. As a result, there is a significant reduction in leakage current though N2, resulting in improved UGNM. However, this technique also suffers from several

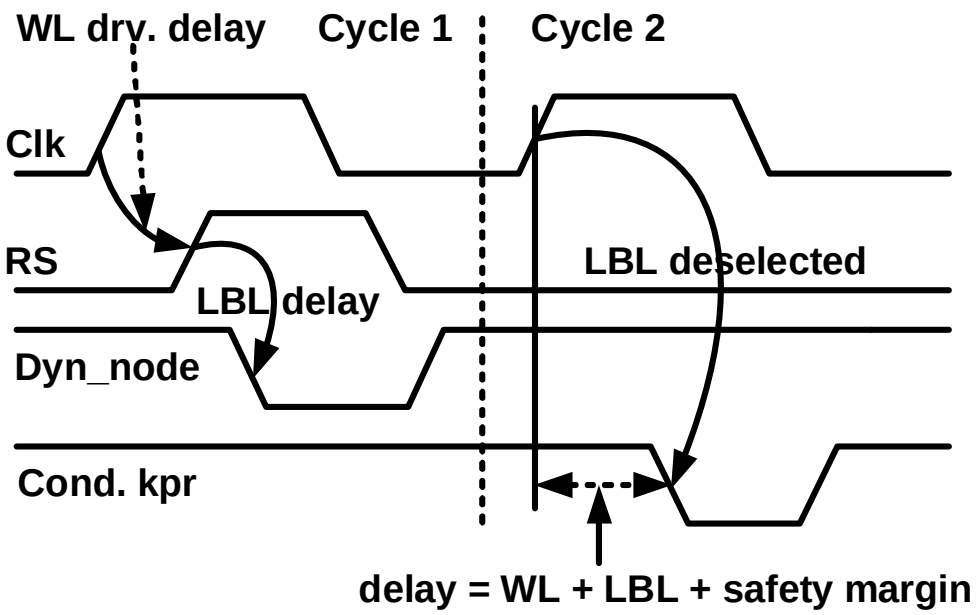


Figure 4.6: Conditional keeper timing

disadvantages and increases the overall RF delay due to LBL stack transistor re-ordering and an extra logic stage (2-input NOR) on the critical path. In addition, this strategy is not readily usable for GBLs that typically use single n-MOS pulldowns.

## 4.3 Leakage Control Schemes and RF Designs

In this section we discuss the selective usage of three different leakage control schemes: dual $V_{TH}$, non-minimum channel length (longer $L_{drawn}$) transistors and reverse body bias (RBB) for designing 8 and 16 wide-OR domino circuits. We also compare the transistor

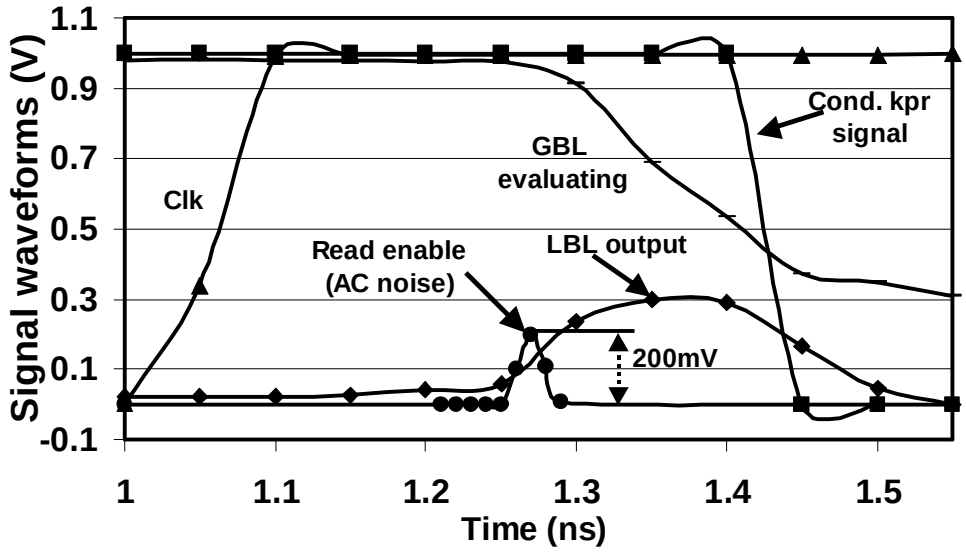


Figure 4.7: AC noise margin degradation of conditional keepers

$I_{ON}/I_{OFF}$ ratios to demonstrate the effectiveness of these techniques and show the design tradeoffs for the 90nm and 65nm technologies.

### 4.3.1 Dual Threshold LBL Design for 90nm Technology

In the dual-$V_{TH}$ LBL design approach we selectively use high threshold transistors, to minimize leakage current and limit delay degradation. Figure 4.9 shows the dual $V_{TH}$ assignment for an 8-wide LBL design. The 2-stack LBL dominos are organized such that the bottom transistors ($D_0$-$D_7$) are connected to the data from the RF local bitcells and are set up ahead of time. Consequently, the read select ($RS_0$-$RS_7$) transistors determine the domino gate's performance and its worst-case UGNM. In the dual-$V_{TH}$ scheme, we use

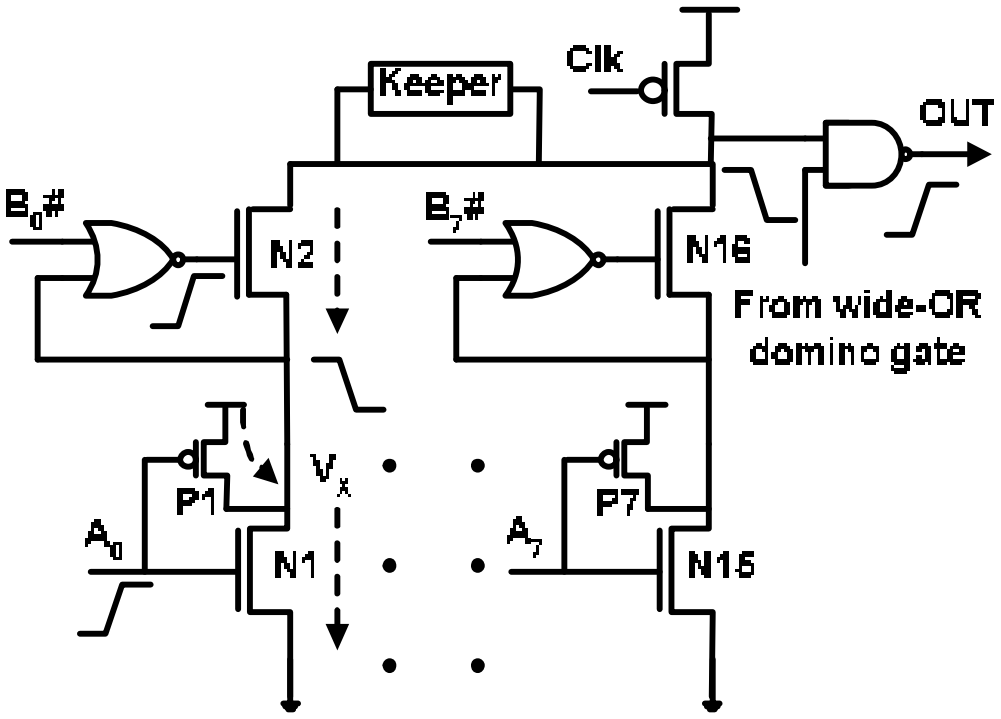


Figure 4.8: Pseudo-static technique based wide-OR domino

high $V_{TH}$ for the read select transistors, while low $V_{TH}$ transistors are used for $D_0$-$D_7$.

In this design, we implemented the static NAND gate using low-$V_{TH}$ transistors. This is because its p-MOS transistors determine the RF read cycle performance while a stronger n-MOS pulldown enhances the UGNM. Since the GBLs have single n-MOS pulldown they are implemented using high $V_{TH}$ n-MOS transistors to ensure robustness. The transistors for both the LBL and GBL keepers (p-MOS transistor and driver inverter) use low $V_{TH}$. We now show data for the dual $V_{TH}$ based 90nm RF read port design and compare them with a low $V_{TH}$ LBL implementation. The data for the two designs are compared under

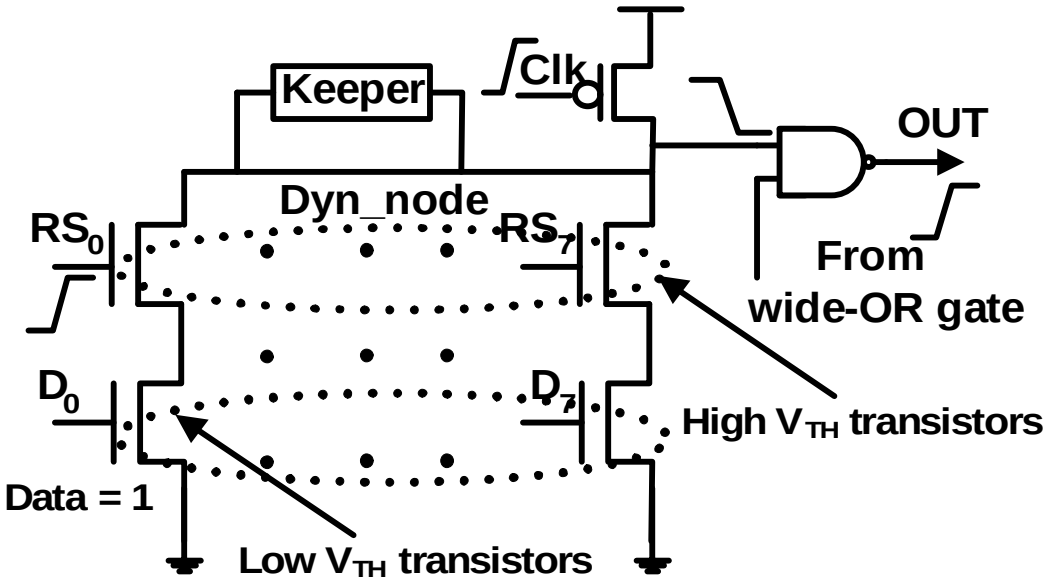


Figure 4.9: Dual threshold based 90nm LBL organization

conditions of equal UGNM. Our results show that the low $V_{TH}$ LBL and GBL designs require 16.5% p-MOS keeper to meet the 17% UGNM robustness threshold. However, the dual $V_{TH}$ designs meet the same robustness threshold with a 4.5% p-MOS keeper. Figure 4.10 compares the worst-case propagation delay and noise margins (AC and DC) for the entire 64-array, 32-bit 90nm RF read port whose organization was presented earlier. This data shows that direct scaling of the low-$V_{TH}$ RF with 5% LBL/GBL keepers to 90nm results in 35% DC and 11% AC noise margin degradations. It is clear that the dual $V_{TH}$ design shows better results compared to the low-$V_{TH}$ design with 12% less propagation delay. In addition, the dual $V_{TH}$ design has 24% lower total energy and 47% lower standby leakage. Clearly, the impact of lower current drive of high $V_{TH}$ transistors is offset by the

lower contention due to weaker p-MOS keepers in dual $V_{TH}$ LBL designs.

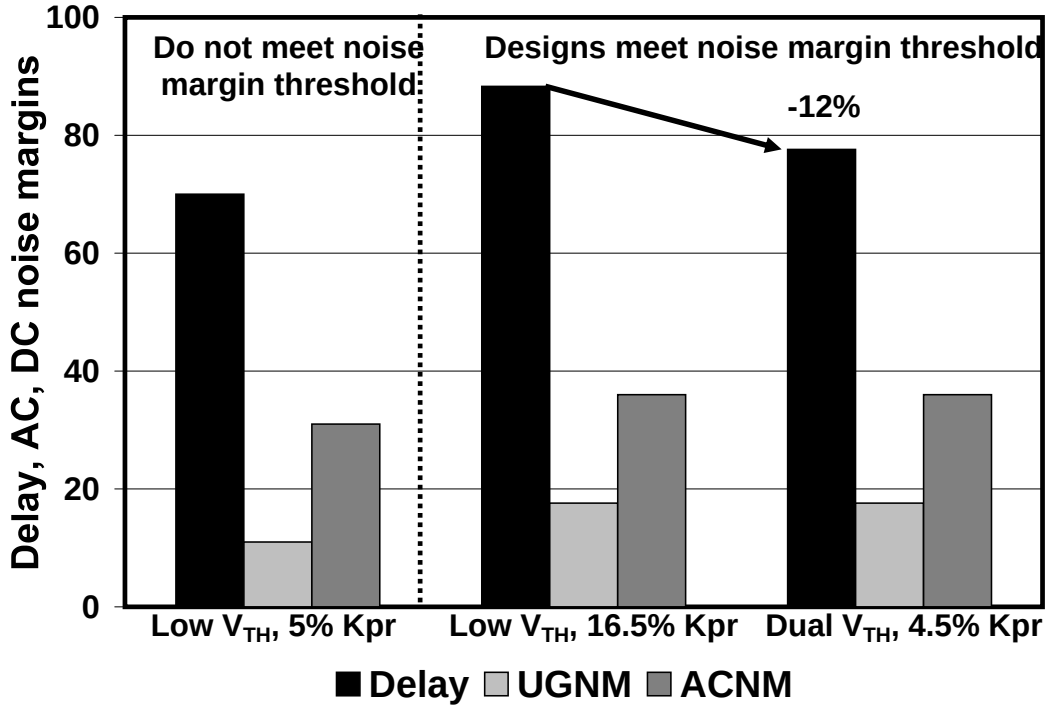


Figure 4.10: Delay and noise margin comparisons for 90nm RFs

### 4.3.2 Transistor Level Leakage Control: Analysis

As CMOS technology is scaled to the sub-90nm regime, the high $V_{TH}$ transistor $I_{OFF}$ also becomes significant. The $I_{OFF}/\mu m$ is about 25x that of the 130nm generation. As a result, LBL and GBL circuits for the 65nm generation designed using high $V_{TH}$ pulldowns and 5% keeper do not meet the noise margin thresholds. Thus, additional transistor level leakage control techniques are required to ensure their iso-robustness scaling. In this section, we compare the effectiveness of non-minimum channel length and reverse body bias (RBB)

techniques in reducing transistor $I_{OFF}$ and quantify their impact on $I_{ON}$ degradation. As discussed in Chapter 3, the transistor $I_{OFF}$ comprises of several components, of which the weak inversion and drain-induced barrier lowering (DIBL) currents are the most important. These two dominant leakage current components can be modelled using the equations as nentioned earlier. In this chapter we discuss the effect of two additional techniques in limiting transistor leakage current and use them to improve the robustness of wide-OR domino based datapath circuits.

*Impact of longer* $L_{drawn}$: Transistors with longer $L_{drawn}$ (non-minimum channel length) reduce the $I_{OFF}$ [10], [18] by increasing the zero-bias threshold voltage ($V_{TH0}$). For small increases in $L_{drawn}$ the threshold voltage increases almost linearly. The increase in the transistor zero bias threshold voltage can be approximated using $\Delta V_{TH0} = V_{TH0}(\frac{\Delta L_{drawn}}{L_{drawn}})$. In addition, the channel mobility remains approximately constant due to velocity saturation in DSM transistors. Therefore, the reduction in leakage current using non-minimum channel length transistors can be approximately modelled as [10]:

$$\frac{\Delta I_{OFF}}{I_{OFF}}|^{L_{drawn}} = 1 - \frac{1}{1 + \frac{\Delta L_{eff}}{L_{eff}}} e^{\frac{-\Delta V_{TH0}}{n v_T}} \tag{4.1}$$

where $\Delta L_{eff}$ is the change in effective channel length ($L_{eff}$) while all other terms have their usual meanings.

*Impact of reverse body bias*: Leakage current can also be suppressed by using the reverse body biasing (RBB) technique. The body of the n-MOS transistor is connected to a negative voltage with respect to the source terminal. The reduction in leakage current is proportional to the extent of the applied reverse bias voltage ($V_{SB}$). Recent research

indicates [10], [21] that beyond a certain optimal RBB voltage the transistor OFF-state current starts to increase due to increased gate induced drain lowering (GIDL), limiting the effectiveness of this technique. For the range of RBB voltages in the region of interest, Eq. (4.2) can be used to model the leakage current reduction as follows:

$$\frac{\Delta I_{OFF}}{I_{OFF}}|^{RBB} = 1 - e^{\frac{-\gamma V_{SB}}{n v_T}} \tag{4.2}$$

Both of the techniques discussed above result in an effective shift in transistor $V_{TH}$. Consequently, they are both associated with reduced gate overdrive voltages and lower $I_{ON}$ $[\propto (V_{DD} - V_{TH})^{\alpha}]$. An efficient leakage control technique is one that allows large reductions in $I_{OFF}$ with minimum $I_{ON}$ degradation. This helps to minimize their adverse impact on performance when used in high-end datapath circuits. For this purpose, we compare the degradation of the normalized $I_{OFF}/I_{ON}$ ratio as shown below:

$$\xi = \frac{\delta I_{OFF}}{\delta I_{ON}} / \frac{I_{OFF}}{I_{ON}} \tag{4.3}$$

Table 4.2 shows the value of the figure of merit (F.O.M.) $\xi$ and its scaling trends for both of the leakage control techniques.

Table 4.2: Data showing current ratio degradation

| $\xi = \frac{\delta I_{OFF}}{\delta I_{ON}} / \frac{I_{OFF}}{I_{ON}}$ | 130nm | 90nm | 65nm |
|---|---|---|---|
| Longer $L_{drawn}$ ($L_{drawn} + 30\%$) | 3.1 | 3.1 | 2.8 |
| RBB (30% $V_{DD}$ reverse bias) | 20.0 | 9.0 | 7.5 |

The data in Table 4.2 demonstrate that the effectiveness of both leakage control schemes

is reducing with scaling. Since RBB has a higher $\xi$ (F.O.M.), it is more efficient in controlling leakage current compared to channel length. However, the RBB approach requires a triple-well process with the generation and routing of an extra power supply voltage. The physical design of RBB and longer $L_{drawn}$ based LBL/GBL circuits and their area penalties will be discussed subsequently.

## 4.4 Low Power, 65nm Wide-Domino Operation

Multi-GHz RFs in high performance microprocessors can contribute significantly to the total IC power. Thus, it is important to reduce the LBL and GBL power while ensuring their robustness with minimum performance degradation. Both the supply voltage ($V_{DD}$) and transistor capacitance ($C_L$) reduce with technology scaling, resulting in lower switching energy/transition. It is expected that for the 65nm technology, a significant portion of the IC total energy will be due to the leakage component and this may reverse the energy scaling trends. This is especially true for large RF designs where only one array is selected at a given time while deselected entries are leaking.

When the transistor channel length is increased, the effective gate area ($WL_{drawn}$) increases proportionately. and contributes to higher gate capacitance ($C_{G0}$-$C_{G7}$) for the n-MOS pulldown transistors. The different parasitic capacitance components associated with the domino pulldown are shown in Figure 4.11. During precharge, (CLK=0) the read select signals ($RS_0 - RS_7$) are equal to 0 and the domino node is held at logic 1. There is no channel formation for transistors in cutoff region, resulting in the parasitic gate-drain capacitance component ($C_{GD}$) being equal to zero. During evaluation when an LBL is

selected (for example: $R_{S0}$ and Data=1), the n-MOS transistor's gate capacitance ($C_{G0}$) has to be charged up from 0 to $V_{DD}$. For LBL and GBL designs that use longer channel lengths, this can cause a near linear increase (8% higher $C_{G0}$ for 20nm longer $L_{drawn}$) in both switching energy and propagation delay.

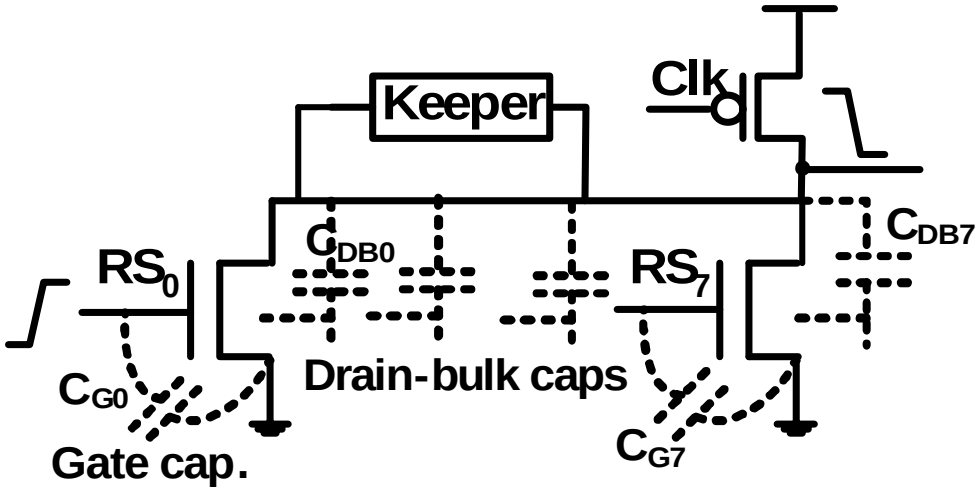


Figure 4.11: Wide-domino organization with parasitic capacitances

On the other hand, the RBB technique reverse biases the drain-bulk p-n junction of the n-MOS pulldown transistors. This lowers the diffusion capacitances ($C_{DB0}$ to $C_{DB7}$) of the entire network (Figure 4.11). Our results for the 65nm technology show that an RBB voltage of 200mV reduces the drain-bulk capacitance by 3.6%. The overall dynamic node capacitance of wide-OR gates is dominated by the diffusion capacitance [45]. Therefore, the RBB approach can result in 4% lower capacitance compared to longer $L_{drawn}$ transistors.

As discussed earlier, both these leakage control techniques result in $V_{TH}$ shifts thereby lowering the $I_{OFF}$ current and transistor $I_{ON}$. Therefore, the effectiveness of these techniques needs to be compared when they are used in high-end datapath designs. This is done in Figure 4.12 by plotting the 65nm transistor currents in the $I_{ON}$-$I_{OFF}$ plane. It is clear that a technique that has a steeper slope in the $I_{ON}$-$I_{OFF}$ plane is more efficient and results in less delay degradation when used in critical path designs. By comparing the data points A and B, we observe that there is 42% lower leakage current using RBB compared to longer $L_{drawn}$ transistors for the same $I_{ON}$. On the other hand, for points A and C, we observe an 18% higher $I_{ON}$ using RBB for the same leakage current. The above discussions demonstrate that RBB reduces the effective switched capacitance as well as offers a higher $I_{ON}/I_{OFF}$ ratio for the transistors.

We now compare the energy-delay plots for the 64-array, 32-bit RF read port implemented using LBL and GBL circuits that incorporate the above mentioned leakage control schemes. We compare the following three different designs:

- Upsized keeper (reference design),
- Longer $L_{drawn}$ (increased by up to 20nm), and
- RBB up to 200mV.

The LBL and GBL pulldown networks for these 65nm designs were implemented using high $V_{TH}$ n-MOS transistors. The high $V_{TH}$ leakage for 65nm technology is significant requiring 16% keeper upsizing to meet the UGNM threshold. On the other hand, when longer $L_{drawn}$ or RBB techniques are used, the UGNM threshold is met with 7.5%-8%

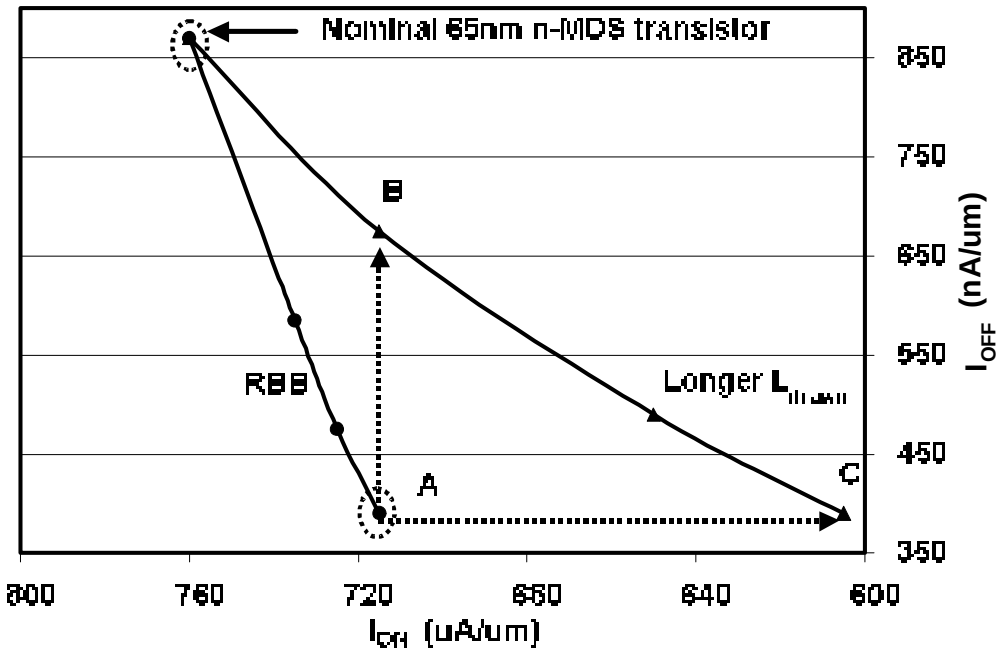


Figure 4.12: Leakage control schemes and their effectiveness

keepers. Figure 4.13 shows the energy and delay plots for the three different RF read port implementations. Our results show that using RBB (200mV) based LBL and GBL designs can improve the propagation delay by 10% while resulting in 16% lower total energy. These reductions result from lower switched capacitance, weaker p-MOS keeper and lower $I_{OFF}$ for the pulldown networks. The results for the longer $L_{drawn}$ (20nm longer) based design show similar reduction in energy but are associated with a 4% delay increase. This is because the longer $L_{drawn}$ transistors have lower $I_{ON}/I_{OFF}$ ratio compared to the RBB technique and higher overall switched capacitance. It should be noted that all the designs in Figure 4.13 have 17% DC robustness and 36% AC noise margin.

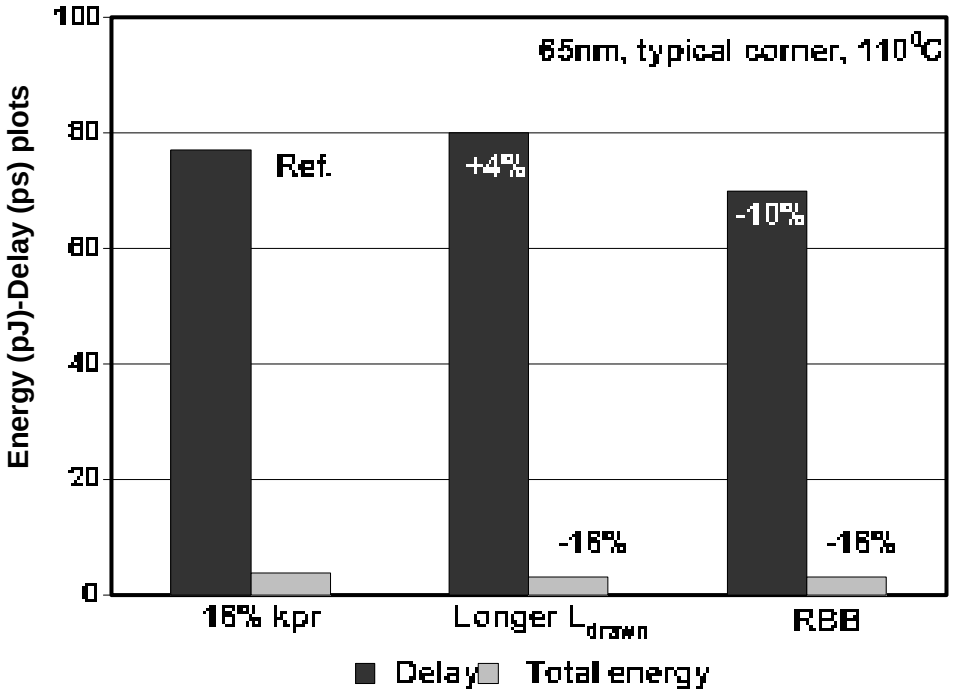


Figure 4.13: RF read port energy-delay comparisons (65nm)

## 4.5 Precharge Current Control and 16-Wide Domino

The RF energy/transition comprises energy consumed during both the evaluation and precharge phases. Our results show that a significant portion of RF energy is consumed during precharge operation due to short-circuit current ($I_{sc}$). The parallel paths of wide-OR dominos result in a self-loading effect and degrade its rise and fall times. In addition, large RFs use multiple stages of cascaded, footerless LBL and GBL circuits. In this section we discuss a circuit technique whereby CLKB (inverted clock) transistors are introduced at the outputs of the static gates to speed up precharge and minimize $I_{sc}$. We use this

scheme along with the leakage control techniques discussed earlier to design low power 16-wide LBL based RF read port designs.

### 4.5.1 Precharge Current Control for RFs

During precharge (CLK=0), the RF domino nodes (LBL, GBL) are held at logic 1 with the pulldown network read-select inputs equal to zero. This results in a short-circuit current free logic operation under steady-state conditions. However, during the switching transient from the evaluation to precharge phase, there can be significant short-circuit current. The short-circuit current during precharge phase is aggravated by the fact that the WL driver and LBL $0 \rightarrow 1$ transitions have to propagate through before the GBL node can fully precharge to $V_{DD}$ and also there are degraded rise times (self-loading) for LBL and GBL circuits. Figure 4.14 shows the timing diagram for a typical bit-line during the precharge phase.

In order to minimize this problem in this RF read port design, we use additional n-MOS CLKB transistors at the outputs of the static logic gates to speedup the precharge operation [46]. This limits the contention and reduces short-circuit current. The wide-OR domino organization along with the n-MOS precharge transistors and its impact on the precharge current is shown in Figure 4.15 and Figure 4.16. During evaluation, CLKB = 0, and the n-MOS precharge transistors are turned OFF. As a result, the RF can evaluate without any steady state current. However, when CLK=0, the CLKB transistors turn on pulling nodes $S_0$-$S_3$ to ground (Figure 4.15) and they cut off the GBL pulldown paths. The effect of these transistors is to provide additional parallel paths to speed up the precharge

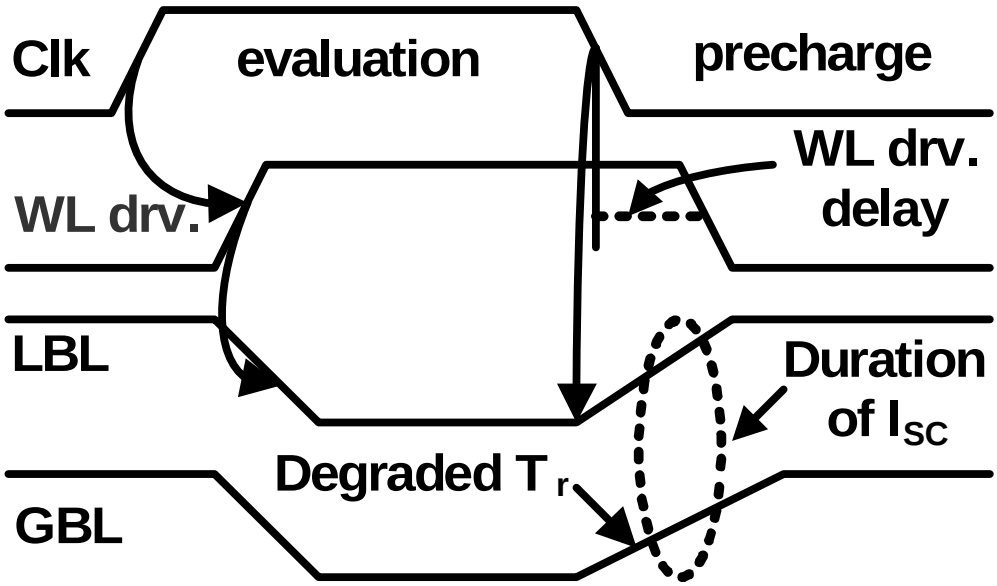


Figure 4.14: RF read port precharge timing diagram

process. Further short-circuit current reduction is possible by adding CLKB transistors at the WL driver outputs. However, our results show that this increases the clock load and overall switching energy. Therefore, for our 64-array RF design we restrict the usage of these precharge transistors to the outputs of the NAND gates ($S_0$-$S_3$ in Figure 4.15).

It is clear from the waveforms shown in (Figure 4.16) that the GBL signal $0 \rightarrow 1$ transition is significantly delayed for the conventional RF design. As a result, the $I_{sc}$ current flows for a longer duration adding to the overall RF energy consumption. However, as shown in Figure 4.16, when the CLKB transistors are added, the GBL reaches $V_{DD}$ sooner and the duration for short-circuit current flow is reduced. It can be seen from Figure 4.16

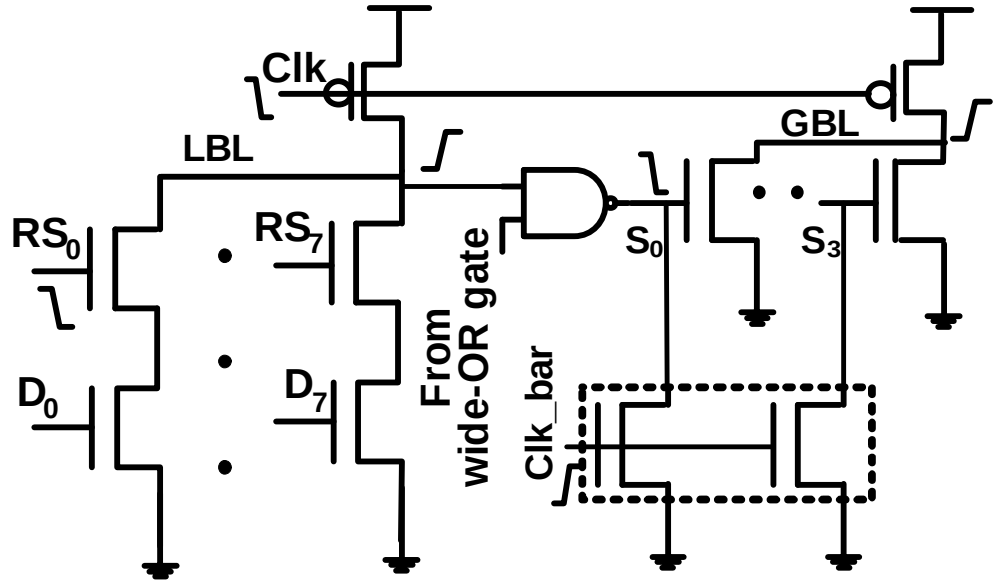


Figure 4.15: RF with CLKB precharge transistors

that the peak $I_{sc}$ for the proposed scheme is higher than that of the conventional design. If the CLKB transistors are too large, the increase in clock load offsets any savings in precharge short-circuit current. Therefore, we optimized the size of these transistors so that the overall EDP (energy-delay product) was minimized.

The optimized EDP data for different CLKB sizes are shown in Table 4.3. We use the 65nm design with high $V_{TH}$ pulldown as our reference. The data in the $2^{nd}$ column showing the RBB entry corresponds to 200mV reverse body bias voltage for the LBL and GBL designs. The subsequent columns correspond to data with progressively larger CLKB transistors. For example, RBB+0.1PC refers to the case when both the LBLs and

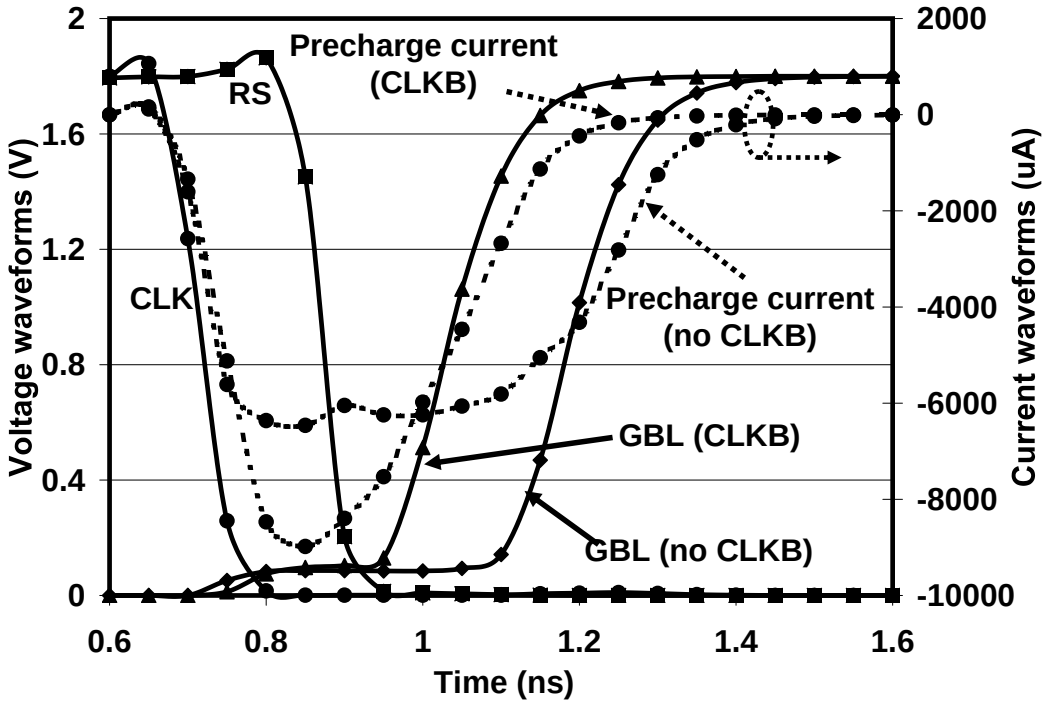


Figure 4.16: RF precharge current and voltage waveforms

GBLs use 200mV body bias and CLKB precharge (PC) transistors that are 10% of the GBL pulldown. The data in Table 4.3 show that as the precharge CLKB transistors are upsized (10%, 20%, 30% of GBL pulldown width), the energy savings reduce while the propagation delay increases proportionately. Our results show that the RBB+0.2PC data point corresponds to optimal operation with 30% lower EDP (25% from RBB, 5% from precharge transistors) than the reference design.

The data corresponding to longer channel length (nominal+20nm) LBL and GBL designs are shown in the $L_{drawn}$ section. These results show that the EDP for such a design is

Table 4.3: Energy-delay data for low power 65nm design

| | RBB | RBB+0.1PC | RBB+0.2PC | RBB+0.25PC |
|---|---|---|---|---|
| EDP (pJ.ps) | 214 | 207 | 201 | 202 |
| Energy | 0.82 | 0.78 | 0.75 | 0.75 |
| Delay | 0.91 | 0.92 | 0.93 | 0.94 |
| | $L_{drawn}$ | $L_{drawn}$+0.1PC | $L_{drawn}$+0.2PC | $L_{drawn}$+0.25PC |
| EDP (pJ.ps) | 248 | 237 | 232 | 233 |
| Energy | 0.83 | 0.78 | 0.75 | 0.75 |
| Delay | 1.04 | 1.06 | 1.07 | 1.08 |

approximately 16% higher than the RBB based implementation. This is due to the higher propagation delay associated with longer channel length transistor capacitances and the self-loading effect of wide-OR LBL and GBL designs. However, as can be seen from the results in Table 4.3, this design approach also reduces the EDP by ∼20% with respect to the original reference design.

### 4.5.2 Designing Low Power 16-Wide Robust Dominos

As technology is scaled, RF bitlines have to be fragmented to compensate for the increased transistor $I_{OFF}$ and maintain iso-robustness. The critical path of the 64-array RF read port considered in this work is organized as follows: 8-wide domino LBL, 2 input static NAND, and 4-wide domino GBL. Further bitline fragmentation (8-way LBL to 4-way LBL) increases the number of logic stages and overall delay. Fragmentation also increases

the total switched capacitance and number of clocked stages resulting in higher RF total energy. In this section, we show that leakage control techniques can be used to prevent further bitline fragmentation and design 16-wide, low power and robust RFs.

Leakage control techniques like RBB and longer $L_{drawn}$ transistors reduce the $I_{OFF}$ exponentially. The $I_{ON}$ degradation is proportional to the reduction in gate overdrive voltage [$\propto (V_{DD} - V_{TH})^{\alpha}$]. In the region of interest (RBB of up to 200mV, $L_{drawn}$ = 20nm longer), the effective $I_{ON}/I_{OFF}$ ratio increases by 2x-2.1x. This allows us to design a 16-wide domino with the same keeper strength as the original design and still meet the robustness threshold. The new read port critical path organization for the 64-array RF thus becomes: 16-wide domino LBL, 2-input static NAND, 2-wide domino GBL.

The new RF read port configuration increases the LBL capacitance by approximately 2 times while reducing the total number of parallel LBLs from 8 to 4. In addition, the GBL capacitance is also reduced by a factor of 2. Using a 16-wide LBL, 2-wide GBL configuration reduces the overall clock load and causes less short-circuit current in the GBL units during precharge. As a result, the RF total energy consumption is lowered. However, the larger LBL capacitance results in delay degradation. The contribution of LBL as a percentage to the total RF total read delay depends on several factors such as:

- Technology generation under consideration,
- Ratio between interconnect and gate delay,
- Whether leakage control techniques are employed, and
- Keeper strength used to ensure noise margin.

Our results show that the LBL delay accounts for 47%-63% of the read port delay. Therefore, the delay increase for the 16-wide LBL offsets the performance improvement obtained from using a 2-wide domino GBL. As a result, the new RF design shows an overall delay increase compared to the original design (8-wide LBL, 4-wide GBL). In this work we compare results for the following four designs:

- Design 1: Reference design with high $V_{TH}$ pulldown, 8-wide LBL and 16.5% upsized keepers,
- Design 2: 8-wide LBL, 4-wide GBL, 200mV RBB, and optimized CLKB transistors for precharge current reduction,
- Design 3: Extension of Design 2 with 16-wide LBL, 2-wide GBL and 200mV RBB, and
- Design 4: Longer $L_{drawn}$ (nominal+20nm) LBLs and GBLs, 16-wide LBL, 2-wide GBL, and optimized precharge transistors.

The delay, energy and EDP for these designs are compared in Table 4.4.

Table 4.4: Low power RF designs, 65nm results

| | Design 1 (Ref.) | Design 2 | Design 3 | Design 4 |
|---|---|---|---|---|
| EDP (pJ.ps) | 287 | 201 | 189 | 218 |
| Energy (norm.) | 1 | 0.75 | 0.59 | 0.6 |
| Delay (norm.) | 1 | 0.93 | 1.11 | 1.27 |

Our results show that Design 3 reduces the RF read port total energy by 41% and EDP by 34%. However, as explained earlier, the wider LBL design causes an 11% increase in the read delay. Design 4 shows similar energy savings as Design 3, but has larger delay degradation (27%). The higher delay for Design 4 (longer channel length) is due to lower $I_{ON}/I_{OFF}$ ratio and higher dynamic node capacitance.

## 4.6 Implementation Issues

In this section we discuss some of the design and implementation issues involved with the different leakage tolerant wide-OR domino schemes discussed in this research. It is clear that keeper upsizing is the easiest to implement, but shows unacceptable energy-delay tradeoffs for sub-130nm technologies. Our results show that dual $V_{TH}$ based RFs have acceptable energy-delay and noise margin characteristics for the 90nm technology. A dual $V_{TH}$ process requires additional process steps adding to the total manufacturing cost. However, such a design strategy may be adopted for robust 90nm designs since dual $V_{TH}$ transistors are already available with high performance logic processes.

It is expected that for 65nm technology, the leakage of even the high $V_{TH}$ transistor will be significant. Therefore, wide-OR dominos might require all high $V_{TH}$ pulldowns in order to meet DC robustness and AC noise margin thresholds. In our 65nm design, we selectively used RBB or longer $L_{drawn}$ transistors in addition to all high $V_{TH}$ n-MOS pulldowns to ensure proper noise margin. Using longer channel length transistors does not require additional masks but amounts to selective reverse scaling. Our results suggest that the drawn gate length has to be increased by 10%-20%. This would require a process with

optimized characteristics for both types of transistors (nominal and 20nm longer).

The RBB technique requires both design modifications and extra process steps. As shown in Figure 4.17, RBB requires a triple well process. The deep n-well region along with the n-diffusions on each side helps to isolate the n-MOS transistors with RBB from those formed in the substrate. Also, contacts are required to connect the n-well to $V_{DD}$ and reverse bias the parasitic p-n junctions. This creates a restriction on the minimum dimensions (A, B, C) for the n-well and the placement of neighboring n-MOS transistors in the substrate. In addition, the negative body bias voltage has to be generated and routed to the RF bitlines. This may require designing and integration of DC-DC converters or allocating additional pins at the package level depending on whether the power supply is generated on-die or off-chip. Thus, even though RBB shows improved performance, it has the maximum area and implementation overheads. The most significant area penalty for implementing RBB based RF bitlines is due to the n-well dimensions (A) and location of the neighboring n-MOS transistors (B). However, since the RF bitlines are regular structures and RBB is used selectively, we align the n-MOS transistors and use a shared n-well design to reduce overall area penalty.

Figure 4.18 and Figure 4.19 show the layout for two different 180nm RF bitline organizations. The first bitline implementation shown in Figure 4.18 represents the typical approach followed in RF designs. It shows two 8-wide LBLs merged into a 2-input static NAND gate to form a 16-wide bitline. The second design is shown in Figure 4.19 and it represents a 16-wide RBB based (shared well) LBL with n-MOS transistors for precharge current minimization.

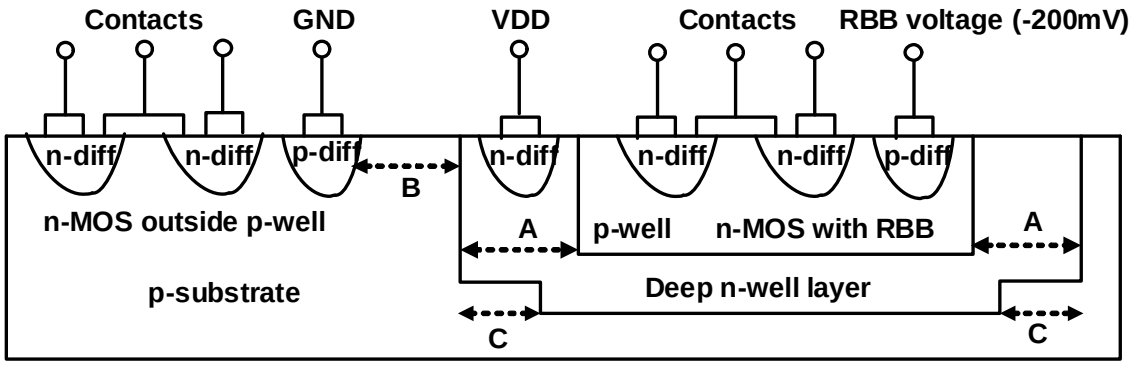


Figure 4.17: Cross-section for RBB and non-RBB n-MOS transistors

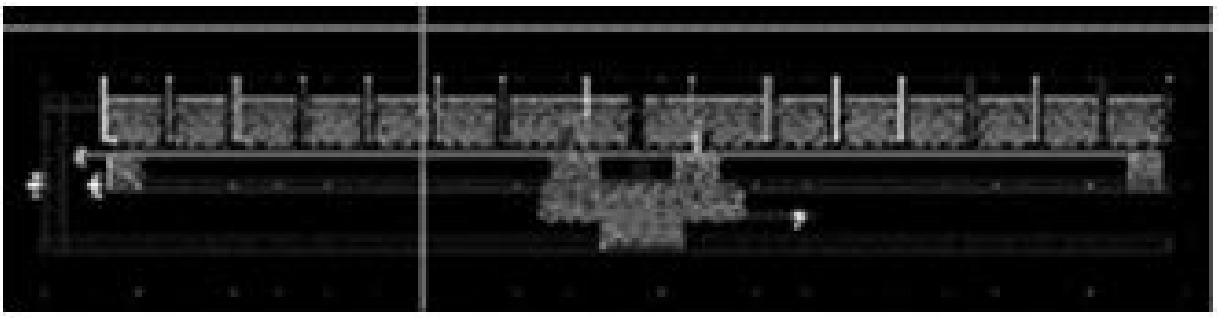

Figure 4.18: LBL organization without RBB or precharge control

The layout comparisons show that the RBB based bitline requires about 20% more area than the normal design. Most of the additional area requirement is due to the n-well with less than 1% area required for the n-MOS precharge transistors. The overhead is expected to be less for a larger RF since a significant portion of the area is occupied by the read-write decoders, multi-port design and input-output latches. On the other hand,

the area penalty for the 16-wide RF bitline design using longer channel length is less than 1%.

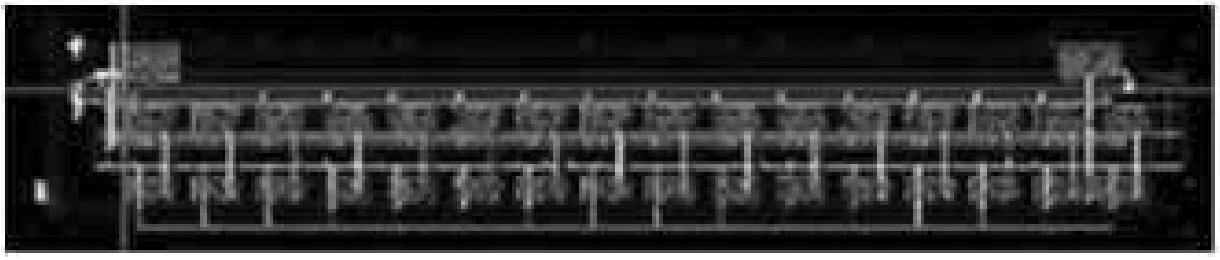

Figure 4.19: 16-wide LBL with RBB and precharge control

## 4.7 Summary

In this chapter we considered several leakage control and circuit techniques for designing robust, low power, wide-OR domino based RFs. We demonstrated that in addition to energy and delay comparisons, both AC and DC robustness need to be considered for sub-130nm RFs. Our results show that the dual-$V_{TH}$ technique is suitable for the 90nm technology and has acceptable energy-delay characteristics. However, additional circuit techniques along with all high-$V_{TH}$ pulldowns will be required for 65nm RF designs. Conditional keeper schemes may not provide adequate AC noise margins and can lead to higher switching energy, while FBB techniques may result in higher total power. Selective usage of longer channel length transistors (10nm-20nm longer) or RBB (200mV) based bitline designs can be used to design low power 65nm RFs. Our results show that the RBB based bitline has better performance but higher overall area and implementation overheads. We also integrated clocked n-MOS transistors with the RF bitline to minimize short-circuit

current during precharge. These techniques help us design RFs with wider LBLs and lower total energy by up to 40%. It is expected that such techniques will enable the designing of scalable LBL, GBL and MUX-es for RFs and ALU front ends. In the next chapter we present an on-chip DFT scheme that can detect delay faults and demonstrate how it can be integrated with high performance datapath units.

# Chapter 5

# Delay Fault Testability and Diagnostics for High Performance Datapaths

Modern microprocessors operate at clock frequencies more than 3GHz and have close to 100 million transistors on die. Digital IC performance has tracked Moore's Law and improved by 30% annually. However, the performance of the automatic test equipments (ATE) has improved by only 12% per year. In the 1980s, ATEs typically offered performance headroom of 5x or more over the device under test (DUTs). However, this advantage has now almost disappeared, and as the current trends continue, tester-timing errors are approaching the cycle time of the fastest devices [3]. As a result, at-speed testing is becoming more difficult. Thus, tester inaccuracy along with scaled geometry, and higher device speed, is expected to compromise IC yield and quality. Moreover, the higher number of DUT pins, demand

for higher ATE accuracy, and larger vector memories are expected to increase the cost of the state-of-the-art ATEs.

In order to maintain improved DUT performance and achieve higher levels of integration, supply voltage ($V_{DD}$), transistor threshold ($V_{TH}$) and oxide thickness ($T_{ox}$) are being scaled. This is resulting in a 3x-5x increase in the transistor $I_{OFF}/\mu m$ and IC background leakage every technology generation. Consequently, the total and peak current (power) demand of the circuit under test (CUT) is expected to increase. This is eroding the effectiveness of traditional test techniques such as $I_{DDQ}$ and stress testing (burn-in) [6], [18], [47]. As a result, parametric defects that cause timing-only failures as opposed to catastrophic logic failures are becoming more common in deep sub-micron (DSM) technologies [28], [48]. Such defects are difficult to detect and therefore result in increasing numbers of test escapes. This trend is posing a serious problem to the long-term reliability of future generation digital ICs.

In this chapter, we present a design for testability (DFT) technique that is geared towards the detection of such hard-to-detect defects in high performance digital ICs. Further more, we explore the possibility of using relatively low TEST mode clock frequency to detect such defects. This is expected to reduce the overall test cost, while improving the long-term reliability of high end digital ICs.

## 5.1 High Performance Circuit Testing: Background

VLSI defects are physical deformations caused by missing or extra material and manifest themselves in the form of shorts or opens. Depending on their impact, defects are typically

classified as:

- Global defects, and
- Local defects.

Global defects generally affect large areas on-die or even entire wafers and are normally easier to detect. On the other hand, local defects generally impact a smaller area on die. However, such defects are difficult to detect, and often require rigorous test practices for proper screening. Techniques used to detect IC defects can be broadly categorized as:

- Indirect (correlation based) methods, and
- Direct test methods.

An example of an indirect test technique is presented in [49], where $I_{DDQ}$ test results are correlated with the maximum operating frequency of a 32-bit microprocessor. The fact that shorter channel lengths lead to higher operating frequency and quiescent leakage current forms the basis of this technique. Another methodology, proposed in [50], is based on the Very Low Voltage Test (VLV) technique where ICs are performance tested at reduced $V_{DD}$. It was observed that delay faults were more noticeable at a lower $V_{DD}$ and hence easier to detect. However, the VLV technique affects only the transistor delay, while leaving the interconnect delay largely unchanged. In modern microprocessors, interconnects are responsible for an increasingly larger segment of the total delay. Hence, this method's suitability in DSM technologies is being eroded.

There has been an increased focus on direct test techniques which rely on:

- ATEs with improved capabilities/higher frequencies, and
- DFT and BIST (Built-In Self Test) for improved CUT testability.

Some of these methods [51], [52] are based on the incorporation of additional DFT structures and the creation of a low frequency TEST mode. The basic idea is to include an externally controlled, quantifiable delay to enable slow-speed testing. Such techniques are especially suited for combinational circuits bounded by flip-flops. However, these techniques can detect delay faults above a certain minimum value and require the routing of externally available, timing critical clock signals in the TEST mode. In addition, it is difficult to build-in diagnostics to locate a subset of logic gates causing the timing anomalies in large and complex CUTs.

In this chapter, we present a DFT technique that can detect delay faults with finer resolution and allows for the lowering of the TEST mode clock frequency. The work presented here demonstrates the applicability of this DFT technique for a 32-bit full custom ALU design. This is achieved without using any additional external timing critical signals (or pins) while maintaining the NORMAL mode energy-delay penalties within acceptable limits.

## 5.2 Circuit Strategy for DSM Digital Testing

Logic circuits implemented using the dynamic CMOS style offer higher performance over their static counterparts. Therefore, the performance critical microprocessor functional unit blocks (FUB) such as arithmetic logic units (ALU), and register files (RF), are of-

ten implemented using dynamic circuits. Some of these design details have already been discussed in Chapter 3. Such logic blocks normally have tight timing budgets and are therefore more prone to timing-only failures. In addition, the microprocessor operating frequency is closely tied to the performance of such FUBs and may be adversely affected by the presence of delay faults in such FUBs. Therefore, in this work we present a DFT strategy geared towards the detection of delay faults in performance critical FUBs that are designed using dynamic logic.

### 5.2.1 DFT for Delay Testing in CDL gates

Circuit designers have devised many different logic styles within the domino family in order to maintain high performance while ensuring scalability. In this research we focus on the testability of designs that use compound domino logic style (CDL) since it is used in the design of full-custom digital datapath designs [15], [36], [53]. CDL gates incorporate alternate stages of n-MOS domino and static CMOS logic gates thereby ensuring both improved performance and robustness. In particular, this logic style is used in the design of high performance MPU adders, ALUs, and register files. Figure 5.1 shows a chain of 7 CDL gates and is representative of the critical path of a 32-bit ALU. In addition, it also shows the proposed DFT structures [54], [55], [56], required to detect delay faults in such a circuit arrangement.

This circuit has 2 modes of operation:

- NORMAL, and
- TEST.

In the NORMAL mode of operation, the mode control signal T/N is set to logic 0. It is clear from Table 5.1 that this causes both the signals CTRL1 and CTRL2 to be treated as don't cares. During NORMAL mode operation, the 3 output signals of the DFT logic shown in Figure 5.1 are set to $V_{DD}$. As a result, the n-MOS footer transistors (N3, N5, N7) are always ON and allow the circuit to evaluate depending on the vectors applied at the primary logic inputs ($A_1 - A_M$).

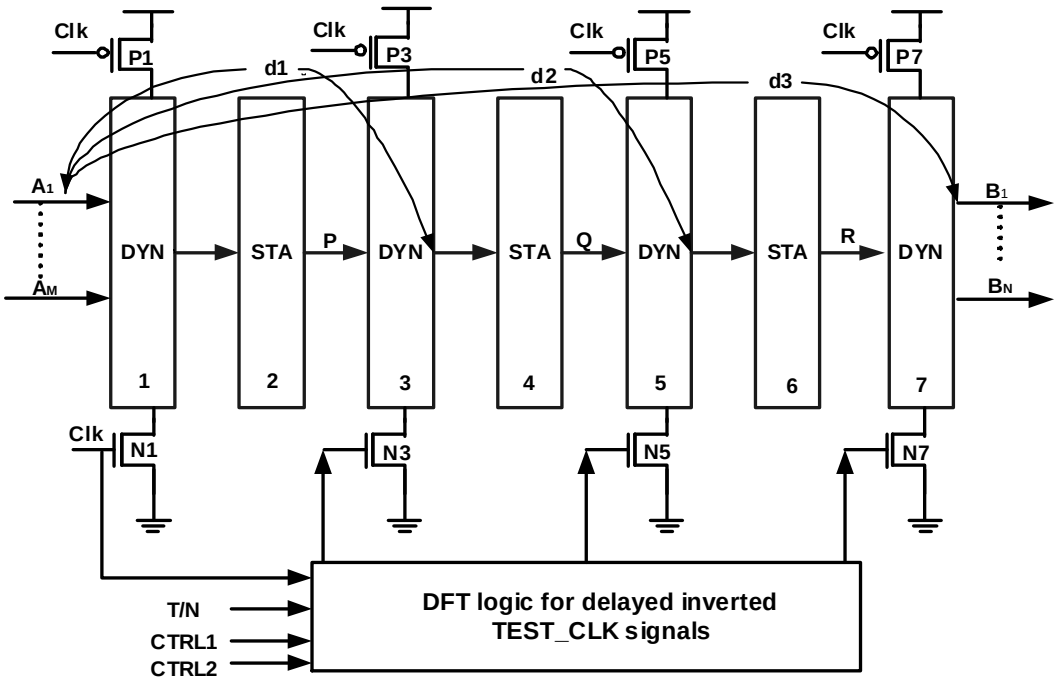


Figure 5.1: CDL gates with DFT for delay testing

In the TEST mode, we create an "evaluation window" for the circuit under test (CUT). This is achieved by applying the system clock signal (CLK) to the first stage of input logic gates and a delayed-inverted clock ($TESTCLK$) signal to subsequent logic stages. The

"window" duration is equal to the delay between the CLK and $TESTCLK$ signals. We build a safety margin into this "evaluation window" to account for delay variations caused by process, temperature and voltage fluctuations during testing in order to prevent the rejection of any good parts. The safety margin (∼60ps in 180nm) is a design parameter, and in this design, it was set to one inverter delay (F.O.=3).

Table 5.1: Truth table for DFT logic and mode selection

| T/N | CTRL1 | CTRL2 | Comment |
|---|---|---|---|
| 0 | x | x | NORMAL mode |
| 1 | 0 | 0 | TEST Section 1 (delay = d1) |
| 1 | 0 | 1 | TEST Section 2 (delay = d2) |
| 1 | 1 | 0 | TEST Section 3 (delay = d3) |
| 1 | 1 | 1 | Reserved (low power stress testing) |

For the case when the CUT is devoid of delay faults, the intermediate nodes (P, Q, R in Figure 5.1) can evaluate in the available "window". However, when a delay fault is present, circuit evaluation is delayed and signals get pushed out. In case the delay fault is excessive, the CUT fails to evaluate in the available evaluation time. Such a failure can then be detected at the primary outputs ($B_1 - B_N$) as a logic failure. Thus, our DFT technique helps convert delay faults internal to the combinational logic block into readily detectable stuck-at faults observable at the primary outputs. In addition, by setting the CTRL1 and CTRL2 signals appropriately (Table 5.1), it is possible to route the $TESTCLK$ signal to the selected n-MOS footer transistors (N3, N5, N7). This allows us to test a sub-section

of the CUT for delay faults using tight evaluation timing while the others are subjected to a more relaxed window. This allows us to trace a logic failure at the CUT primary outputs back to a set of internal gates and helps in creating built-in delay diagnostics. Another advantage of this DFT technique is the possibility of lowering the TEST mode clock frequency. The evaluation window used to detect delay faults has two edges:

- Opening edge, and
- Closing edge.

The system clock provides the opening edge, while the closing edge is obtained locally using the DFT logic. Thus, the detection of delay faults is dependent on the correct phase relationship between the CLK and $TESTCLK$ signals while being independent of their absolute signal frequencies. Hence, this DFT technique can enable delay fault testing at relatively low TEST mode clock frequency using cheaper ATEs. This concept is illustrated with the help of waveforms shown in Figure 5.2. We show the HSPICE simulations for 180nm CDL gates (with DFT) for a variable delay fault. The extent of the delay fault was controlled by introducing a variable resistance in series with the evaluation network of the logic gates. This has the impact of increasing the effective RC time constant and CUT delay [18], [55]. We use a TEST mode clock frequency that is 5x lower than the NORMAL mode of operation. It is clear that when DFT footer transistors are used, the CUT fails when the defect resistance is more than 1.25KOhms. However, in the absence of DFT, the same circuit fails to detect defect resistances of up to 3KOhms.

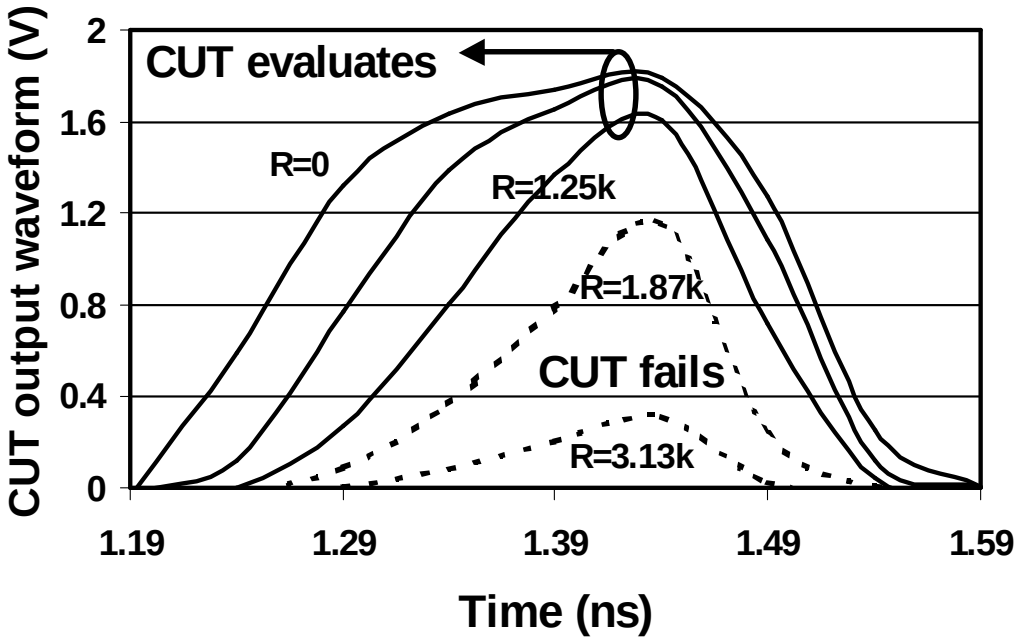


Figure 5.2: Low frequency delay testing with DFT

### 5.2.2 Delay Fault Detection Range

Our proposed DFT technique allows us to increase the range of detected defect resistance compared to a non-DFT circuit. This is crucial in high performance DSM circuits and FUBs, and can be better understood with the help of Figure 5.3 and Figure 5.4. In Figure 5.3, we show the location of some of the typical resistive defect in the domino logic gates (keeper omitted for clarity) under test. In this study, we considered both the cases when resistive defects were present on the transistor drain (R1) and source terminals (R2). In addition, we considered defects being present in pulldown paths that comprised single (R1, R2 in series with A) as well as multiple series (R3 in series with B, C) connected

n-MOS transistors.

For domino circuits with no DFT structures (N3, N5, N7 removed), the circuit has the entire duration when CLK=1 to evaluate. For our specific circuit example shown in Figure 5.4, this duration is about 200ps. As the defect resistance (R1) is increased, the CUT evaluation time gets pushed out and fails completely above a value of 3KOhms. However, when DFT is used, the circuit has a smaller evaluation window. Consequently, the CUT fails when the defect resistance is more than 1.25KOhms. However, it should be noted, that even with DFT, a certain range of defects (up to 1.25KOhms) still goes undetected. This is because the delay impact of such defects is within the safety margin, and an attempt to detect delay faults with finer resolutions can result in rejection of good parts and yield loss. It should also be noted that the defects in the high resistance range can however (above 3KOhms in this case) always be detected in our example.

Our results demonstrate that the CUT with DFT can consistently detect a larger range of defect resistance. This is clear from the simulation results for the resistances R1, R2, and R3 as shown in Table 5.2. We considered only one defect being present in the circuit at a given time and observed the defect resistance required for which the CUT begins to fail with and without DFT. Our results show that the extra range of resistance detected for R1, R2 and R3 using the DFT scheme are equal to 1.25K-3K, 3.13K-5K and 2.5K-3.1KOhms, respectively. It should be noted that the absolute values of detected defect resistances depend on the actual design, transistor width and ON-state resistance. However, the DFT scheme allows us to screen a larger range of DFT defects during testing.

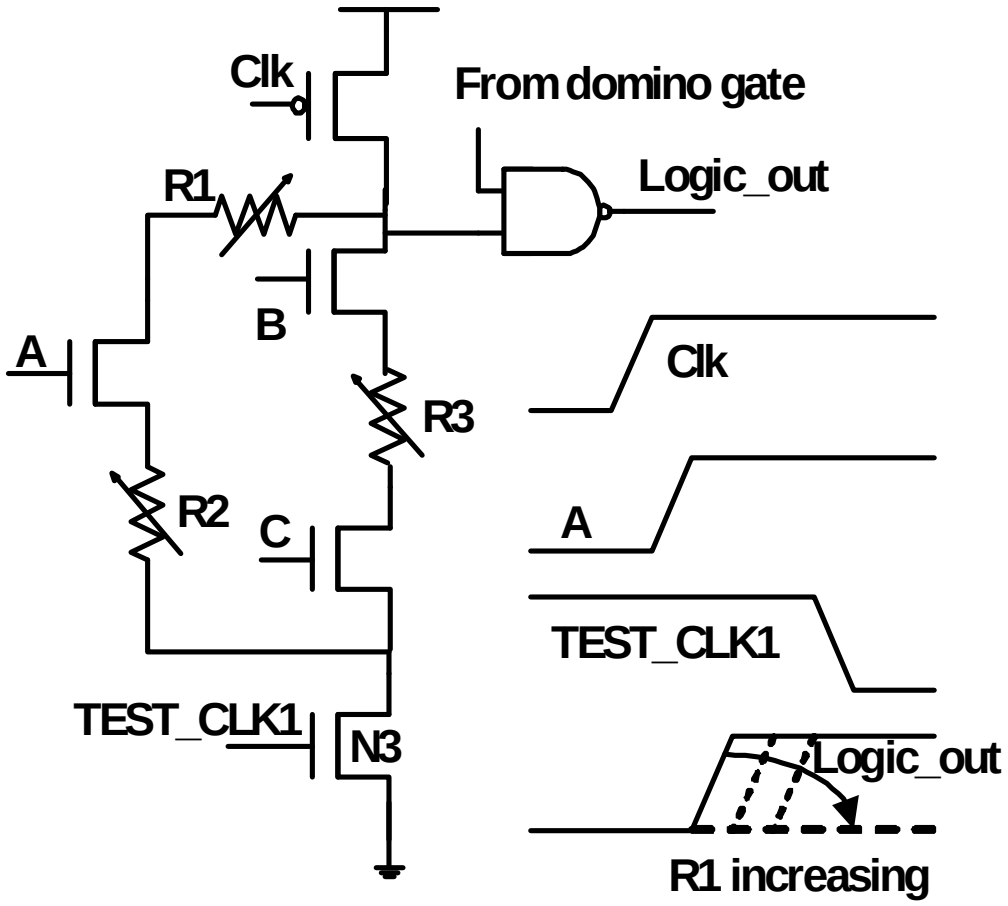


Figure 5.3: Resistive defects: typical location in CUT

## 5.3 Design Overview of DFT Based 32-bit ALU

In this section, we discuss the design of a delay fault testable, 32-bit high performance ALU. This is achieved by integrating our DFT technique with the low power ALU design discussed in Chapter 3. This design has two modes of operation, NORMAL and TEST. In

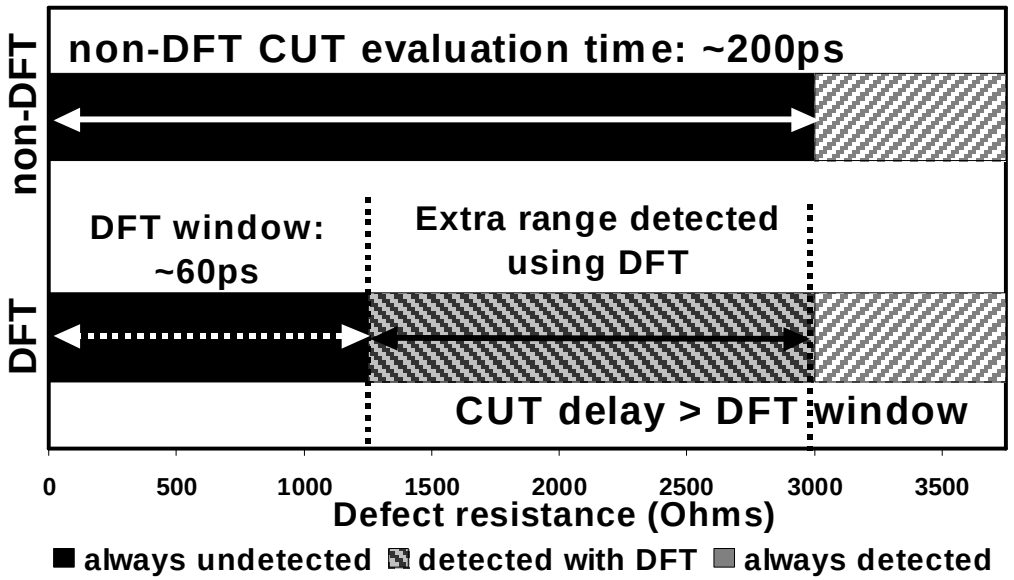


Figure 5.4: Defect resistance detection range DFT vs. non-DFT

Table 5.2: Defect resistance detection for 180nm technology

| | With DFT min. Ω | No DFT min. Ω | Extra Range |
|---|---|---|---|
| R1 | 1.25K | 3K | 1.25K-3K |
| R2 | 3.13K | 5K | 3.13K-5K |
| R3 | 2.5K | 3.1K | 2.5K-3.1K |

the NORMAL mode, the ALU performs arithmetic, logical and shift operations and can also support a low-power mode of operation using a dual-supply scheme. The ALU consists of approximately 11.5K transistors and operates at 1.5GHz for the 180nm technology. The ALU performance scales to 4.2GHz under worst-case conditions for the 65nm CMOS

technology. The ALU block diagram along with the on-chip DFT circuitry is shown in Figure 5.5 and its basic architecture is similar to that presented in [24], [36]. The block diagram indicates that the ALU comprises several sub-units. The input data stage comprise of master-slave static flip-flops and data drivers for the A[31:0] and B[31:0] buses. The decoder unit determines the actual instruction that is executed by the ALU (arithmetic, logical, or shift). Both the decoder and logic/shift units are non-critical in terms of performance and have relaxed timings. Therefore, the decoder is realized using static CMOS logic, while the logic unit and shifter are implemented using complementary pass transistor logic (CPL) to achieve low power operation. The ALU critical path comprises of the arithmetic unit (adder front-end MUX + 32-bit adder), output MUX-es, and output stage latches. In this design, these units were designed using CDL logic.

In the TEST mode of operation, the DFT logic can be used to perform delay testing on the performance critical units of the ALU. It should be borne in mind that the proposed DFT technique can be integrated with FUBs designed using dynamic logic and is independent of the ALU or its architecture. In this research, we used the ALU as a vehicle to demonstrate the effectiveness of our proposed test technique in detecting delay faults. The ALU on one hand is performance critical, while on the other, involves a reasonable degree of design complexity and a mix of different circuit design styles. This allows us to quantify the various energy-delay tradeoffs and scaling trends associated with our proposed technique.

The DFT logic unit shown in Figure 5.5 is implemented using static CMOS logic and $C^2$MOS MUX-es. When the input instruction to the ALU indicates that it is in the TEST

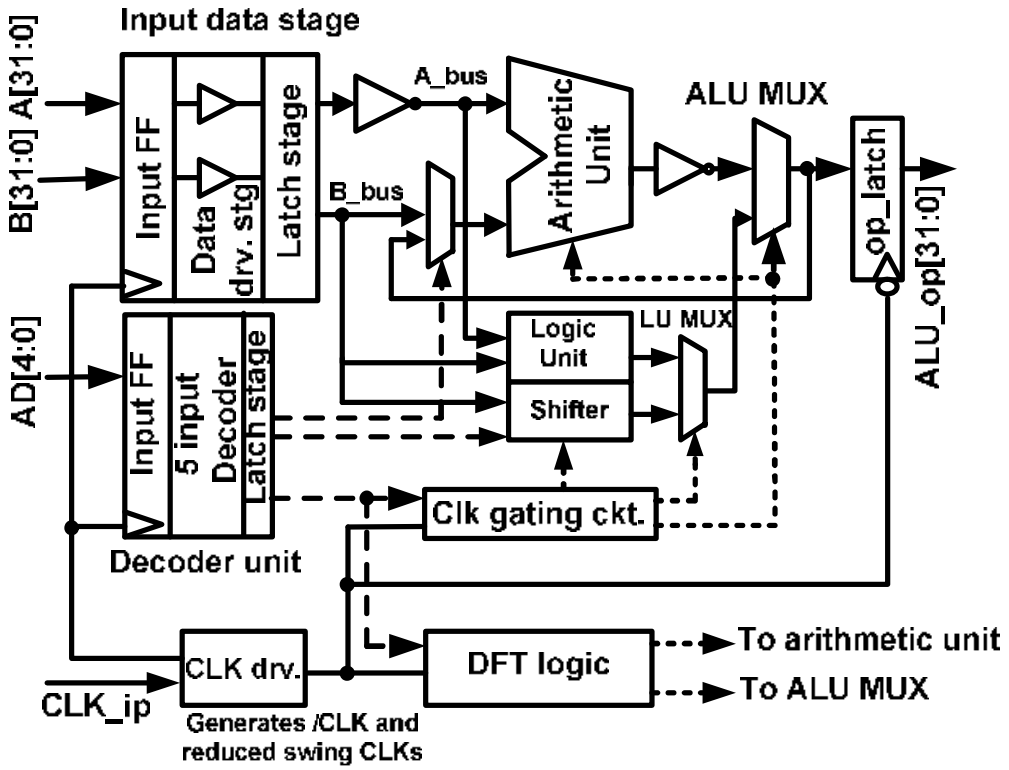


Figure 5.5: 32-bit delay fault testable ALU architecture

mode, the T/N signal is set to logic 1. As a result, the NORMAL mode control signals to the ALU are deactivated (logic 0). The decoder is designed such that it allows the arithmetic unit to operate during the TEST mode. In addition, it is possible to select the particular CDL stages within the ALU to be subjected to delay testing. This is indicated by the broken lines in Figure 5.5, from the output of the DFT logic to the arithmetic unit and the ALU output MUX-es.

### 5.3.1 Delay Testing Logic: Implementation

This section discusses the design details of the DFT logic unit that allows us to generate delayed-inverted $TESTCLK$ signals for ALU delay fault testing. The primary design objectives for the DFT logic are as follows:

- $TESTCLK$ signals should be generated on-chip and locally to the actual CUT logic to be tested,
- Eliminate the need for additional timing critical input signals to be supplied by the ATE,
- Minimize any additional clock load due to the DFT logic in NORMAL mode of ALU operation, and
- Minimize the transistor count, additional input pins and design complexity of the DFT logic.

The above considerations allow us to reduce the design overhead and make it easier to integrate the scheme with the overall logic design flow. Furthermore, it minimizes the additional clock load, and reduces the NORMAL mode switching energy penalty and clock skew. We explain the operation of the DFT scheme with the help of Figure 5.6. The DFT logic comprises two levels of MUX-es and a delay chain implemented using static CMOS inverters. The input stage MUX-es are connected to system signals, namely CLK, CLKB and the supply $V_{DD}$. The output MUX stage provides the gate control for the n-MOS footer transistors (N3, N5, N7) of the ALU. The transistors of the MUX and inverter chain

were sized appropriately in order to obtain the required evaluation window for each ALU section. It should be noted that there are several stages of inversion (act as gain-stages) between the input and output stages of MUX-es. This allows us to use minimum or close to minimum sized transistors for the input MUX stage and reduce the additional load on CLK and CLKB signals.

We used an odd number stages of inverters between the input and output MUX stages in order to obtain $TESTCLK$ signals that are inverted with respect to the input CLK, CLKB signals. It should be noted that we share a portion of the inverter delay chain between the $TESTCLK1$ and $TESTCLK2$ signals. This principle can be applied effectively in more complex designs to save transistor count and DFT logic area. The DFT MUX-es were implemented using $C^2MOS$ stages as opposed to transmission gate logic. This achieves better drive capability and sharp rise and fall time for the $TESTCLK$ signals.

It should be noted that the ALU logic operates on both the clock phases (CLK and CLKB). The input stages of the adder (PG unit) and the Carry Merge Tree evaluate when CLK = 1, while the ALU output MUX stage and output drivers evaluate using the negative phase when CLK = 0. Therefore, the DFT logic shown in Figure 5.6 generates 2 of the $TESTCLK$ signals ($TESTCLK1$, $TESTCLK2$) that are delayed-inverted with respect to the system clock (CLK) while the $TESTCLK3$ signal for the final stage was derived from CLKB. For the DFT unit design, we also ensure that the number of logic inversions on the delay chain equals that of the corresponding CUT section being tested. This helps us to match the delays of the DFT unit and CUT logic stages being tested. For our 180nm ALU design example, the $TESTCLK1$ and $TESTCLK2$ signals were delayed by 230ps,

and 390ps with respect to CLK, respectively. $TESTCLK3$ was delayed by 170ps with respect to CLKB. It should be noted that these delays also include the ∼60ps built-in safety margins.

In the NORMAL mode, the entire DFT logic is disconnected from the CLK grid via the input MUX that connects both node A and B (Figure 5.6) to $V_{DD}$. As a result, all the internal nodes of the DFT unit are actively connected to either $V_{DD}$ or ground. This eliminates the possibility of any intermediate node potentials within the DFT logic and excessive leakage currents during NORMAL operation. Furthermore, the output MUX-es connect the $TESTCLK$ signals to $V_{DD}$ thereby allowing NORMAL mode ALU operation.

In this study, we considered two alternative circuit level implementations for generating delayed-inverted $TESTCLK$ signals. These schemes are shown in Figure 5.7. Scheme 1 uses a chain of inverters followed by static CMOS NAND gate. In the TEST mode, the control signal from the decoder is set to logic 1, and the delayed clock signal turns the n-MOS footer transistor OFF after a predetermined duration. This scheme is different from that shown in Figure 5.6, in that it is not a MUX based design. As a result, it does not decouple the delay chain from the input clock signal in the NORMAL mode. This can result in additional clock skew and switching energy consumption.

Scheme 2 is based on the concept of current-starved inverters. The additional footer transistors on the inverters of the delay-chain are connected to $V_{DD}$ in the NORMAL mode and are fully ON. However, in the TEST mode, the gate voltage can be connected to an intermediate analog voltage (between $V_{DD}$ and 0V) through an external input pin. The input voltage ($V_{bias}$) allows us to control the gate-source overdrive voltage and control the

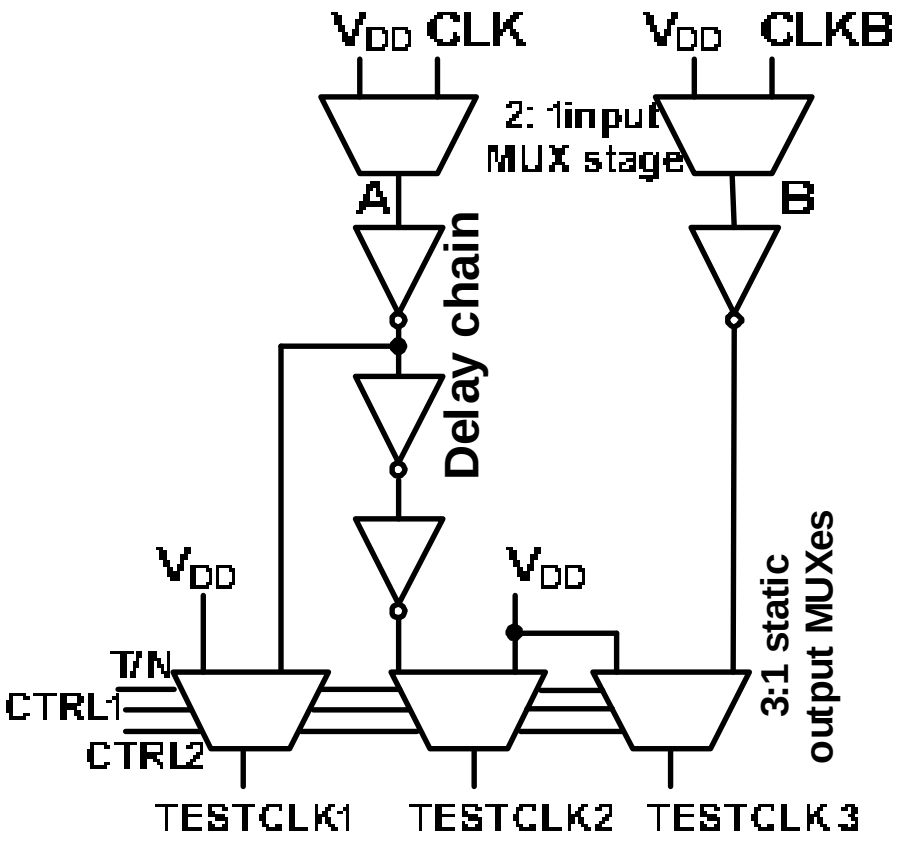


Figure 5.6: DFT logic for a delay fault testable ALU

CUT evaluation window. We show the impact of $V_{bias}$ control voltage on the delay chain in Figure 5.8. Our results indicate that as the $V_{bias}$ voltage is reduced, the control chain's delay increases and the signal rise/fall times (signal slopes) start to degrade. When the $V_{bias}$ voltage is in the range between $V_{DD} \rightarrow (V_{DD}\text{-}2V_{TH})$, the delay increases in small steps. Thus, this range of $V_{bias}$ can be used to fine-tune the CUT evaluation window. However, when the $V_{bias}$voltage is further lowered (less than $0.5V_{DD}$ for our 180nm technology), the

delay changes in much larger steps. When the $V_{bias}$ voltage is in this range, the DFT logic output has degraded rise and fall times.

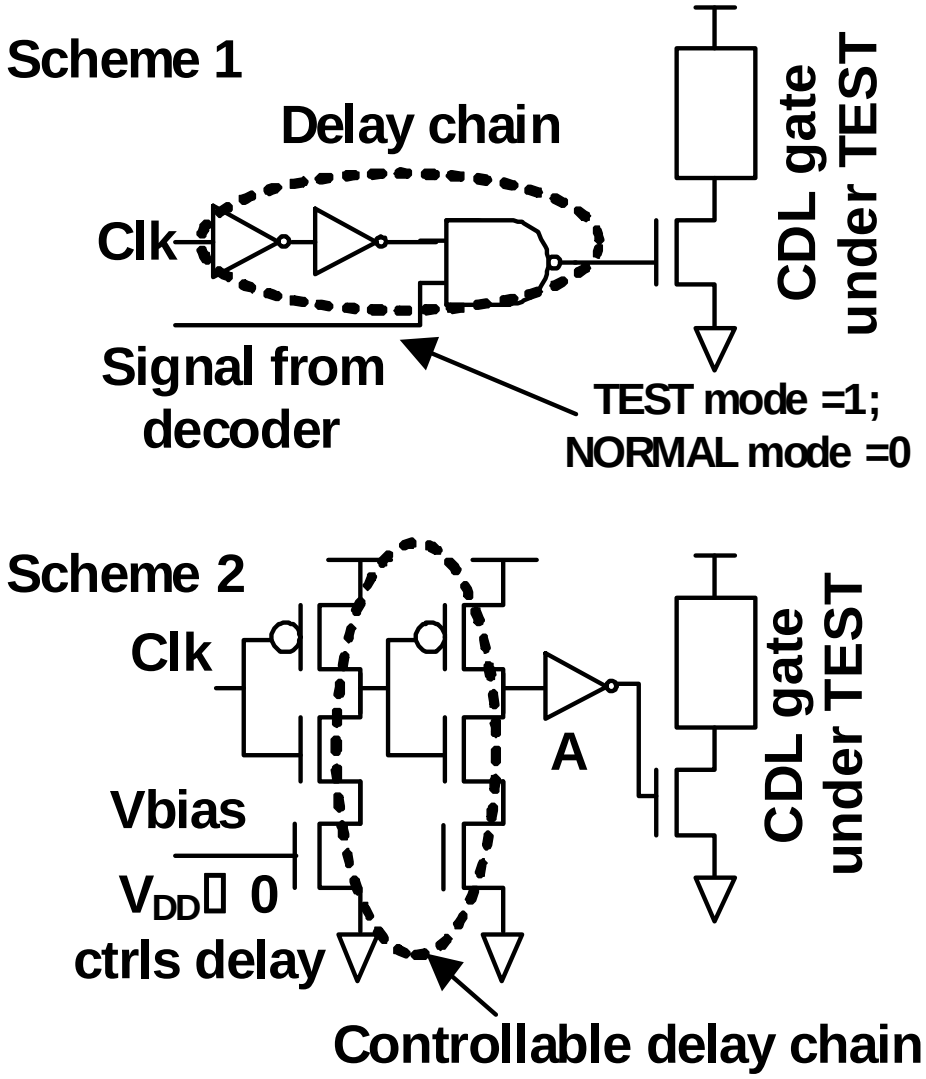


Figure 5.7: Alternate schemes for TEST mode clock generation

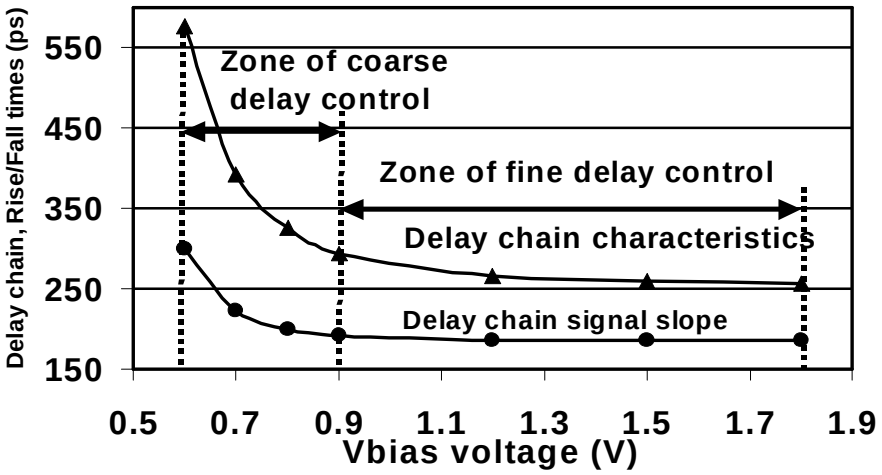


Figure 5.8: DFT logic delay control using bias voltage (Scheme 2)

Typically, the internal signals of high performance CUTs have sharp rise and fall signal slopes. Thus, it is not a good design practice to directly interface the DFT logic output signals (having degraded slopes) with the CUT. This can be mitigated, by allowing the degraded signal(s) to pass through a static CMOS inverter(s) that improves the final signal slope before interfacing with the CUT's footer transistors as shown in Figure 5.8 (inverter A in Scheme 2, Figure 5.7). This scheme can be used in designs that require more flexibility in the delay margins generated by the DFT logic. However, this design requires access to a controllable external analog voltage, an additional input pin and a precise mapping between the input signal voltage and DFT logic delay. Schemes 1 and 2 might be useful in certain applications but for our specific ALU design, we used the scheme enumerated in Figure 5.6.

### 5.3.2 Delay Testable ALU: Energy-Delay Tradeoffs

In this section, we present the simulation results showing the ALU performance and its scaling trends. We also discuss the energy-delay tradeoffs associated with the DFT technique. Our goal was to devise a DFT strategy for the high performance CUT, while minimizing the NORMAL mode delay and energy penalties. Figure 5.9 plots the worst-case delay of both the 32-bit adder and ALU, for the 180nm-65nm CMOS technologies. We plot results for both designs with and without DFT. This allows us to quantify the performance impact of the DFT technique on the NORMAL mode operation.

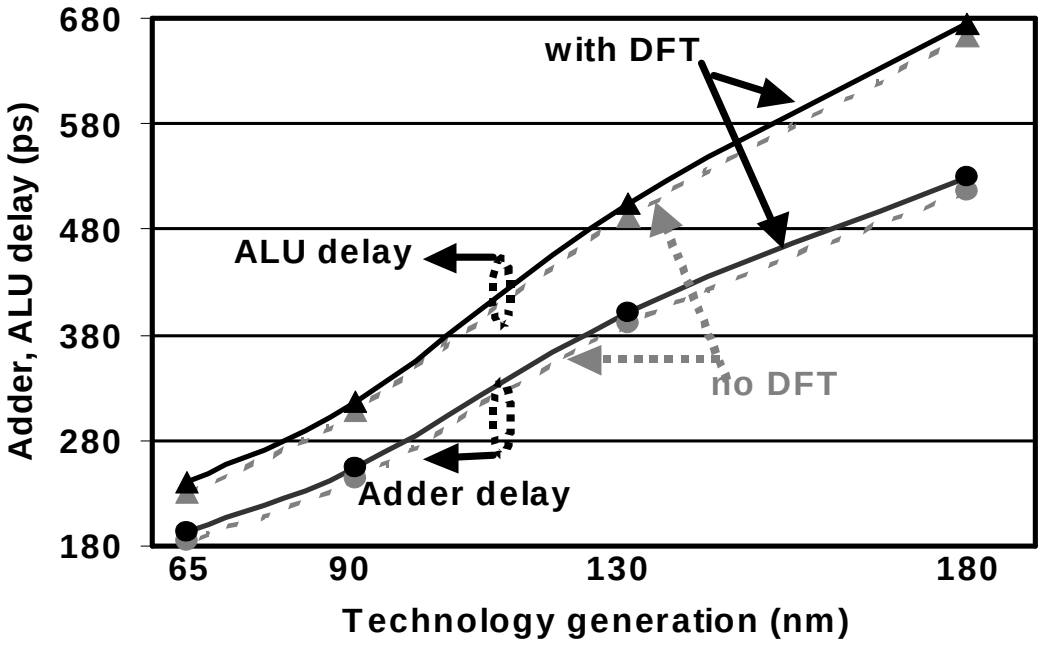


Figure 5.9: DFT technique: delay impact, scaling trends

The data points in Figure 5.9 for the 180nm technology correspond to a bulk CMOS

TSMC process while the 130nm-65nm results were obtained using the Berkeley Predictive Technology Models. Our results indicate that for both the adder and ALU, the DFT technique results in delay degradation. This is due to the additional n-MOS footer transistors (N3, N5, N7) inserted in the pulldown paths that increase the stack height and effective ON-state resistance of the evaluation path. However, the delay penalty can be maintained within acceptable limits by observing the following:

- The footer transistors are added to the dynamic logic gates only, with the alternate static gates left unchanged,
- In the NORMAL mode, these transistors are connected to $V_{DD}$ and are always ON, and
- Since the DFT transistors do not switch in the NORMAL mode, they can be upsized to minimize delay degradation without significantly increasing switching power.

Our results indicate that the DFT technique results in NORMAL mode delay degradation in the range of 2.7%-4.2% for the adder, and 1.8%-4.4% for the ALU for the 180nm-65nm technologies. In addition, the increase in the NORMAL mode switching energy is limited to less than 1% for the above technologies.

## 5.4 ALU TEST Mode Operation

This section deals with the TEST mode operation of the ALU and delay fault detection. In this study, we focused on delay defects existing in the performance critical arithmetic

unit and ALU output MUX-es that have tight timing budgets and are hence prone to parametric, timing-only failures. The other units such as the logic-shift unit and the decoder unit have significantly larger timing margins and hence any timing anomaly in them would be absorbed in the existing slack (unless they are catastrophic failures, which is not our focus).

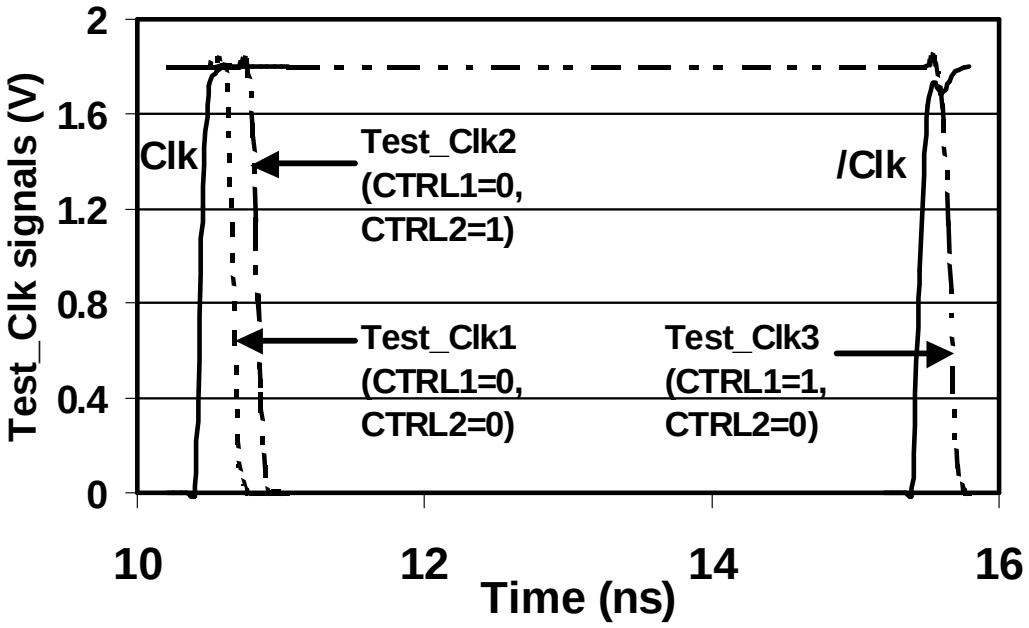


Figure 5.10: TEST mode clock signals for ALU during delay testing

We adopted a stage-to-stage delay testing strategy, where only a specific ALU stage was under TEST at a given time. This section had a tight evaluation window, while the rest of the ALU had relaxed timing. We carried out the delay testing at a $TESTCLK$ frequency 5x lower than NORMAL mode of operation. We show the $TESTCLK$ signals

in Figure 5.10 that were generated by the DFT logic for different CTRL1 and CTRL2 settings (Table 5.1). When the ALU is in the TEST mode, and both CTRL1 and CTRL 2 signals are equal to logic 0, section 1 is under TEST, and the N3 footer transistor is clocked with $TESTCLK1$ signal. This allows us to test the PG unit (propagate-generate) and the first stages of the Carry Merge Tree of the 32-bit adder unit. When $TESTCLK2$ is used to control the N5 footer transistor, the rest of the Carry Merge Tree is under test. Finally, when (TESTCLK3) is active, the adder output stage and ALU MUX-es are under test.

We now focus our attention on inserting resistive defects in the ALU and use our DFT test strategy to detect them. We introduced one delay defect at a time in the adder during the course of this study and the possibility of multiple defects being present simultaneously was not explored. The delay defects were introduced in the static gate p-MOS pullup network, and dynamic logic gate pulldown circuitry. The CDL logic precharge operation is non-critical (happens in parallel) and typically has more timing margin than the domino evaluation phase. As a result, parametric timing anomalies in the precharge network are not of concern in this study.

In order to conduct a representative study of the effectiveness of our DFT methodology, we introduced 11 unique delay faults in the 32-bit ALU. Table 5.3 shows the locations and nature of the defects and indicates that they were distributed evenly among the different logic stages. These resistive defects were introduced in the form of parametric resistances in series with the evaluation transistors. Normally for such defects, the delay impact is proportional to their resistance and they can be used to represent resistive metal lines, S-D

bridging defects, resistive vias and/or contacts.

It is clear from the results that the proposed DFT technique can detect a larger range of defect resistance compared to the non-DFT ALU. The resistances shown in Table 5.3 also map to equivalent circuit delay degradations. Our results indicate that the DFT ALU can detect faults of magnitude greater than the built-in safety margin (∼60ps for 180nm technology). For example, in the case fault 1 (F1 in table 5.3), the non-DFT design can detect resistive defects of magnitude greater than 3KOhms. This corresponds to approximately a delay fault of 330ps. However, with the built-in DFT scheme it is possible to detect defects of magnitude greater than 1.5KOhms. This corresponds to delay faults equal to the safety margin of 60ps. However, delay faults of smaller resolution go undetected. It should be noted that for the DFT design, it was possible to lower the TEST mode clock frequency to 200MHz, without compromising the fault detection range.

The results in Table 5.3 indicate that, for the non-DFT design, a larger range of defect resistance can go undetected. This range is determined by the timing margin between the CUT logic gate with the defect, and the CLK edge as shown in Figure 5.11. In this design, the CDL stages 1-3 evaluate before stages 4-6. As a result, they have more delay margin (Margin 1-3) and larger delay faults (resistance range) can go undetected. However, stages 4-6 evaluate closer to the closing edge of the evaluation phase (CLK $1 \rightarrow 0$ edge) and have a smaller timing margin (Margin 4-6). Thus, the deeper the logic level, the smaller is the timing margin (slack) for the non-DFT design. Consequently, for such gates, the detected resistance range is closer to that of the DFT design. In fact, the faults F7, F8 are in logic stages 5 and 6 that are closest to the CLK $1 \rightarrow 0$ edge resulting in the same defect

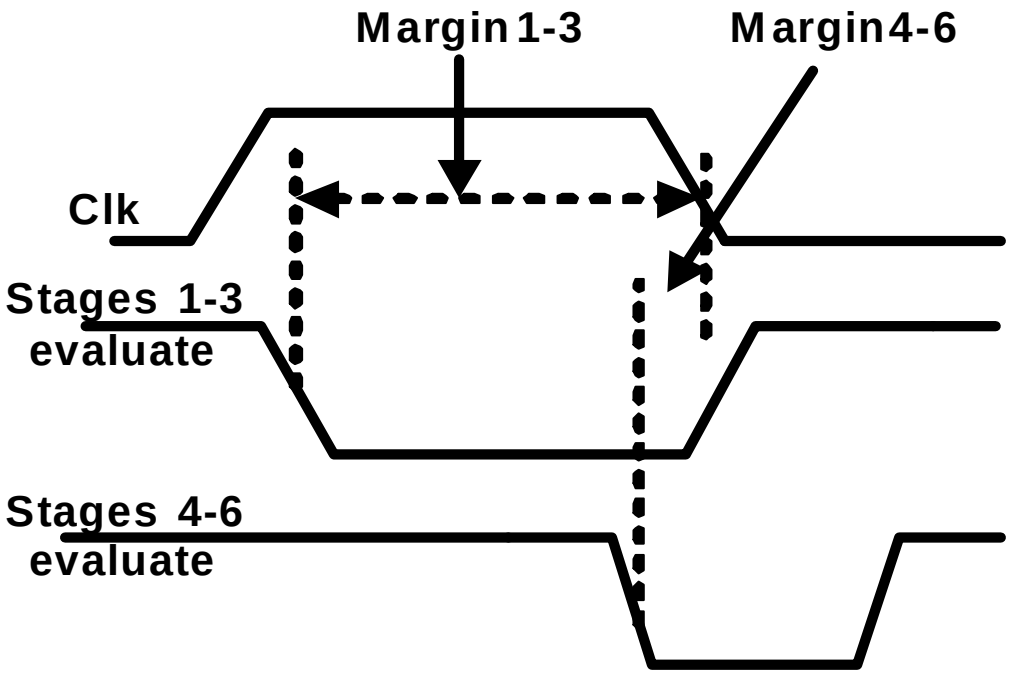


Figure 5.11: Timing diagram showing ALU delay margins

detection range as the DFT design. It should be noted that the ALU evaluates on both clock phases with the logic stages 1-6 evaluating when CLK =1, while stages 7-8 evaluate when CLK=0. This explains the trend in Table 5.3, where the additional range of detected resistance (delay fault) steadily decreases from stages 1-6 and again picks up for stages 7-8.

### 5.4.1 DFT Implementation Issues

Our proposed DFT technique has certain overheads associated with its design and implementation. This scheme requires the designing of a dedicated DFT unit to be activated

Table 5.3: Defect detection range for ALU DFT vs. non-DFT

| Fault No. | DFT (min. detected Ω)/delay fault | Non-DFT (min. detected Ω)/delay fault |
|---|---|---|
| F1 | 1.5K (60ps) | 3K(330ps) |
| F2 | 1.5K (60ps) | 3K(330ps) |
| F3 | 1.5K (60ps) | 3.5K (330ps) |
| F4 | 1K (60ps) | 2.5K (330ps) |
| F5 | 1K (55ps) | 1.5K (160ps) |
| F6 | 1K (55ps) | 2K (160ps) |
| F7 | 0.5K (55ps) | 0.5K (60ps) |
| F8 | 0.5K (55ps) | 0.5K (60ps) |
| F9 | 1.5K (60ps) | 3K (350ps) |
| F10 | 2K (60ps) | 3.5K (350ps) |
| F11 | 3.5K (60ps) | 6K (420ps) |

during the TEST mode. The layout of the DFT based 180nm, 32-bit ALU has an overall area of $800\mu m$ x $600\mu m$. The DFT unit results in a 1.3% increase in the ALU transistor count with an area of $200\mu m$ x $150\mu m$. This results in an area penalty of 4%. Our technique is geared towards performance critical datapath FUBs that are typically full-custom, hand crafted designs. It is therefore expected that the integration of this DFT technique with the logic design flow would not contribute significantly to additional turnaround time during layout. In addition, the proposed stage-to-stage testing methodology may result in longer test time or require additional test pattern generation. However, this is an issue

that remains the topic of future research and has not been addressed in this current study. Finally, we adopted a DFT unit design that results in the creation of a fixed, hard-coded evaluation window for the CUT. However, as has been mentioned in this work, it is possible to design for a delay margin with more flexibility at the expense of extra hardware.

## 5.5 Summary

In this chapter, we presented a DFT technique that can detect delay faults in a high performance 32-bit ALU design. We integrated this technique with the logic design flow and were able to detect a larger range of delay faults (∼60ps for 180nm technology) compared to the non-DFT design at a 5x lower test frequency. The delay (energy) penalty associated with this technique was shown to be between 2%-4% (1%) for the 180nm-65nm CMOS technologies. Further more, we demonstrated how this method can be used to convert delay faults into easy to detect stuck-at logic failures and build-in delay diagnostics using the stage-to-stage testing strategy. We also quantified the area and transistor count overhead of our scheme to be ∼4% and 1.3%, respectively, for the DUT under consideration. It is expected that this technique will help in improving delay fault detection and ensuring long term reliability of high-end digital ICs.

# Chapter 6

# Conclusion

In this thesis we first discussed some of the important challenges of deep submicron VLSI design. Our focus was the design and testability of high performance digital datapaths and logic units. Based on the ITRS projections presented in Chapter 1, we considered three major areas in this research. These include low power datapath operation, improved noise margins for logic circuits and delay testability of high performance designs. We now briefly summarize the main findings of our work and outline some of the possible future work in these area.

## 6.1 Low Power ALU Design

It is well known that as the technology is scaled leakage power is becoming a major problem leading to higher power dissipation and power density. This in turn can lead to power delivery problems, local thermal hot spots and degrade long-term reliability. In recent

years several different leakage control and design techniques have been advanced that can reduce the overall power consumption of performance critical microprocessor functional units. In this work we used a 32-bit high performance ALU design and used a dual supply design strategy along with a swing-restored CPL circuit technique to reduce power consumption. We first partitioned the design into critical and non-critical units and used a lower power supply clocking scheme for the non-critical blocks only. The datapath for the entire design was maintained at the higher supply voltage in order to minimize performance degradation. In addition, a latch (FF) scheme was developed that can interface between the different power supply domains without consuming additional static power. The logic and shifter units that are non-critical were designed using swing restored CPL to minimize the total switched capacitance and reduce the data driver sizes.

Based on these above design modifications we designed a 32-bit ALU using a 180nm, 6 metal layer TSMC process and studied the scaling trends up to the 65nm CMOS technology with the help of BPTM models. We quantified the impact of the design strategies on important design parameters such as total energy, delay, leakage power, peak and average current. The worst case performance of the ALU was 1.5GHz for the 180nm generation and scalable to 4.2GHz for the 65nm technology. It should be noted that the dual supply operation of the ALU can support this operating frequency without performance degradation due to the partitioning and dual supply assignment strategy adopted in this design. Our results show that the ALU total energy was reduced by 18%-24% for the 180nm-65nm technologies using our design strategy. In addition, the leakage power was reduced by 22%-32% when a 30% lower power supply voltage was used. The peak current that is

responsible for IR voltage drop and performance degradation during switching transients was also reduced by 12%-21%.

## 6.2 Robust Domino Designs for Datapath Circuits

The transistor off-state current is increasing by 3x-5x every generation due to threshold voltage scaling. This is not only increasing the IC total power but also has an adverse impact on the noise margin of digital logic. In particular, the wide-OR domino logic gates that are used as high performance MUX-es in the ALU front ends and read paths (LBL, GBL) of register files are most susceptible to leakage induced logic upset. Some of the basic means to improve their noise margin is to increase the domino keeper strength. However, this results in unacceptable delay and contention energy penalties. Therefore, the challenge is to improve the robustness while minimizing the overall delay and switching energy penalties. In this work we considered several different techniques already presented in the literature including keeper upsizing, conditional keepers, forward body biased and pseudo-static techniques. We demonstrated the effectiveness of additional leakage control techniques such as dual $V_{TH}$ dominos, selective channel length modulation and reverse body bias techniques in reducing transistor level leakage.

Based on our findings we designed 8-wide and 16-wide domino bitlines and demonstrated iso-robustness scaling trends for the 130nm-65nm CMOS technologies. Our results show that the dual-$V_{TH}$ technique is suitable for the 90nm technology and has acceptable energy-delay characteristics. However, additional circuit techniques along with all high-$V_{TH}$ pulldowns will be required for 65nm RF designs. Conditional keeper schemes may

not provide adequate AC noise margins and can lead to higher switching energy, while FBB techniques may result in higher total power. Selective usage of longer channel length transistors (10nm-20nm longer) or the RBB (200mV) based bitline designs can be used to design low power 65nm RFs. Our results show that RBB based bitline has better performance but higher overall area and implementation overhead. We also integrated clocked n-MOS transistors with the RF bitline to minimize short-circuit current during precharge. These techniques help us design RFs with wider LBLs and lower total energy by up to 40%.

## 6.3 DFT Technique for High Performance Datapaths

With scaling not only are power dissipation and transistor leakage major challenges, but testing of complex ICs is also becoming significantly more difficult. This is because of several different factors. For example, as the on-chip clock frequency increases by 30% annually due to scaling, the tester frequency is improving by about 12%. As a result, there is an erosion in the performance headroom that current testers have over the circuits under test. This is causing test escapes and posing challenges to the long term reliability of high-end parts. In addition, several new failure mechanisms manifest themselves only during at-speed testing. As ICs become more complex, the number of transistors/pin is increasing making it more difficult to access internal logic, generate test patterns and carry out effective diagnostics. Furthermore, more and more defects in modern ICs result in parametric timing failures that are difficult to detect and whose characteristics vary over time and operating conditions.

In this work we also developed a DFT technique that can detect hard-to-detect delay faults in high performance digital logic. We integrated this technique with the logic design flow and were able to detect a larger range of delay faults (∼60ps for 180nm technology) compared to the non-DFT design at a 5x lower test frequency. The delay (energy) penalty associated with this technique was shown to be between 2%-4% (1%) for the 180nm-65nm CMOS technologies. Furthermore, we demonstrated how this method can be used to convert delay faults into easy to detect stuck-at logic failures and build-in delay diagnostics using the stage-to-stage testing strategy. We also quantified the area and transistor count overhead of our scheme to be $\sim$ 4% and 1.3%, respectively, for the DUT under consideration. It is expected that this technique will help in improving delay fault detection and ensuring long term reliability of high end digital ICs.

## 6.4 Future Work

We now discuss some of the possible areas of possible future work that can be pursued based on the results presented in this research. Each of the three areas of investigation can be further investigated and expanded upon. For example, the concepts of low power datapath design can be extended to special purpose DSP architectures and design of multiplier units. Additional latch (FF) circuits that can support dual supply clocking with reduced overall clock load can also be investigated. Our design incorporated a dual supply approach to minimize IC power consumption. This has been an area of increased research effort in recent years. However, a detailed analysis of whether the additional power supply should be generated on-chip or off-chip is required. An on-chip second power supply will require

a DC-DC converter while an off-chip source may need additional pin assignments and package re-design. The efficiency of on-chip DC-DC converters should also be considered while determining the overall power (energy) savings of a dual supply scheme. It is expected that the overall system level power savings will be less than the trends shown here when the overheads associated with the on-chip regulator are included. Also the impact of low power datapath operation on on-chip hot spots, power (current) density and overall thermal map can be of significance and should be studied.

The study of leakage tolerant datapath designs can be extended so that the proposed schemes can be integrated with a large and more representative register file or ALU design. It is believed that the area penalties observed in our work will be much smaller in a more representative design. This is because the proposed leakage control techniques will be used only for the wide domino circuits, which constitute a relatively small fraction of the total design. In addition, the impact of switching transients and coupling with neighboring high frequency lines should be considered while evaluating the AC noise margin. In our research we considered the worst case noise margins by assuming that all of the input control lines may have DC (AC) noise. However, this is a pessimistic approach and results in excessive design margins. Therefore a more realistic approach should be developed based on the application, switching probabilities and layout information.

Finally, the DFT scheme for delay fault testability can be further extended to improve testability of other domino logic based units such as high performance multipliers and RFs. In addition, the impact of cycle-to-cycle jitter on the evaluation window can be investigated to quantify its impact on the defect detection range. Another factor that

merits investigation is the impact of such a scheme on the overall test pattern generation and test time. Furthermore, the use of the DFT transistors in the normal mode of operation as sleep transistors to reduce standby mode leakage power can be of interest to both the design and test communities.

# Bibliography


[1] R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE Journal of Solid State Circuits*, vol. 9, pp. 256–268, 1974.

[2] G. E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, pp. 82–85, 1965.

[3] S. I. A. (SIA). (2003) International Technology Roadmap for Semiconductors (ITRS). [Online]. Available: http://public.itrs.org

[4] J. Rabaey, A. Chandrakasen, and B. Nikolic, *Digital Integrated Circuits.* Upper Saddle River, N.J., U.S.A.: Prentice Hall, 2003.

[5] A. Chandrakasan, W. Bowhill, and F. Fox, *Design of High Performance Microprocessor Circuits.* Piscataway, N.J., U.S.A.: IEEE Press, 2000.

[6] V.D.Agarwal and M.L.Bushnell, *Essentials of Electronic Testing.* Boston, MA., U.S.A.: Kluwer Academic Publishers, 2002.

[7] A. Chandrakasan and R. W. Brodersen, *Low Power Digital CMOS Design.* Boston, MA., U.S.A.: Kluwer Academic Publishers, 1995.

[8] R. Krishnamurthy, A. Alvandpour, G. Balamurugan, N. Shanbag, K. Soumyanath, and S. Borkar, "A 130nm 6-GHz 256x32 bit Leakage-Tolerant Register File," *IEEE Journal of Solid-State Circuits*, vol. 37, pp. 624–632, 2002.

[9] T. Kuroda, "CMOS Design Challenges to Power Wall," in *Proc. of IEEE International Conference on Microprocessors and Nanotechnology*, 2001, pp. 6–7.

[10] B. Chatterjee, M. Sachdev, S. Hsu, R. Krishnamurthy, and S. Borkar, "Effectiveness and Scaling Trends of Leakage Control Techniques for Sub-130nm CMOS Technologies," in *Proc. of IEEE International Symposium on Low Power Electronic Design (ISLPED)*, 2003, pp. 122–127.

[11] S. O. Jung, K. W. Kim, and S. Kang, "Noise Constrained Power Optimization for Dual $V_T$ Domino Logic," in *Proc. of IEEE the International Symposium on Circuits and Systems (ISCAS)*, 2001, pp. 158–161.

[12] B. Chatterjee, M. Sachdev, and R. Krishnamurthy, "Leakage Control Techniques for Designing Robust Low-Power Wide-OR Domino Logic for sub-130nm CMOS Technologies," in *Proc. of IEEE International Symposium on Quality Electronic Design (ISQED)*, 2004, pp. 415–420.

[13] (2003) Berkeley Predictive Technology Models (BPTM). [Online]. Available: http://www-device.eecs.berkeley.edu

[14] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "New Paradigm of Predictive MOSFET and Interconnect Modeling for Early Circuit Design," in *Proc. of IEEE Custom Integrated Circuits Conference (CICC)*, 2000, pp. 201–204.

[15] K. Bernstein, K. Carrig, C. Durham, P. Hansen, D. Hogenmiller, E. Nowak, and N. Rohrer, *High Speed CMOS Design Styles.* Boston, MA., U.S.A.: Kluwer Academic Publishers, 1999.

[16] J.D.Meindl, "Low Power Microelectronics: Retrospect and Prospect," *Proc. of the IEEE*, vol. 83, pp. 6619–635, 1995.

[17] M. R. Stan, "Optimal Voltages and Sizing for Low Power," in *Proc. of IEEE International Conference on VLSI Design*, 1999, pp. 428–433.

[18] N. Sirisantana, L. Wei, and K. Roy, "High-Performance Low-Power CMOS Circuits Using Multiple Channel Length and Multiple Oxide thickness," in *Proc. of IEEE International Conference on Computer Design (ICCD)*, 2000, pp. 227–232.

[19] T. Sakurai, H. Kawaguchi, and T. Kuroda, "Low-Power CMOS Design through $V_{TH}$ Control and Low-Swing Circuits," in *Proc. of IEEE International Symposium on Low Power Electronic Design (ISLPED)*, 1997, pp. 1–6.

[20] T. Kuroda, T. Fujita, S. Mita, T. Nagamatsu, S. Yoshioka, K. Suzuki, F. Sano, M. Norishima, M. Murato, M. Kako, M. Kinugawa, M. Kakumu, and T. Sakurai, "A 0.9V, 150-MHz, 10-mW, 4mm$^2$, 2-D Discrete Cosine Transform Core Processor

with Variable Threshold-Voltage ($V_T$) Scheme," *IEEE Journal of Solid-State Circuits*, vol. 31, pp. 1770–1779, 1996.

[21] A. Keshavarzi, S. Ma, S. Nagendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar, and V. De, "Effectiveness of Reverse Body Bias for Leakage Control in Scaled Dual Vt CMOS ICs," in *Proc. of IEEE International Symposium on Low Power Electronic Design (ISLPED)*, 2001, pp. 207–212.

[22] M. Johnson, D. Somasekhar, and K. Roy, "Models and Algorithms for Bounds on Leakage in CMOS Circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 18, pp. 714–725, 1999.

[23] R. Krishnamurthy, S. Hsu, M. Anders, B. Bloechel, B. Chatterjee, M. Sachdev, and S. Borkar, "Dual Supply Voltage Clocking for 5GHz 130nm Integer Execution Core," in *Proc. of IEEE International Symposium on VLSI Circuits*, 2002, pp. 128–129.

[24] S. Matthew, M. Anders, B. Bloechel, T. Nguyen, R. Krishnamurthy, and S. Borkar, "A 4GHz, 300mW, 64-bit Integer Execution ALU with Dual Supply Voltages in 90nm CMOS," in *Proc. of IEEE the International Solid-State Circuits Conference (ISSCC)*, 2004, pp. 162–163.

[25] Y. Shimazaki, R. Zlatanovici, and B. Nikolic, "A Shared-Well Dual-Supply-Voltage 64-bit ALU," *IEEE Journal of Solid-State Circuits*, vol. 39, pp. 494–500, 2004.

[26] V. De and S. Borkar, "Technology and Design Challenges for Low Power and High Performance," in *Proc. of IEEE International Symposium on Low Power Electronic Design (ISLPED)*, 1999, pp. 163–168.

[27] Z. Cheng, L. Wei, and K. Roy, "On Effective $i_{DDQ}$ Testing of Low Voltage CMOS Circuits Using Leakage Control Techniques," in *Proc. of IEEE International Symposium on Quality Electronic Design ((ISQED))*, 2000, pp. 181–188.

[28] W. Needham, C. Prunty, and E. Yeoh, "High Volume Microprocessor Test Escapes, An Analysis of Defects Our Tests are Missing," in *Proc. of IEEE International Test Conference (ITC)*, 1998, pp. 25–34.

[29] K. Roy, S. Mukhopadhyay, and H. M. Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proc. of the IEEE*, vol. 91, pp. 305–327, 2003.

[30] R. Gu and M. I. Elmasry, "Power Dissipation Analysis and Optimization of Deep Submicron CMOS Digital Circuits," *IEEE Journal of Solid-State Circuits*, vol. 31, pp. 707–716, 1996.

[31] A. Srivastava, R. Bai, D. Blaauw, and D. Sylvester, "Modeling and Analysis of Leakage Power Considering Within-Die Process Variations," in *Proc. of IEEE International Symposium on Low Power Electronic Design (ISLPED)*, 2002, pp. 64–67.

[32] K. Schuegraf and C. Hu, "Hole Injection $SiO_2$ Breakdown Model for Very Low Voltage Lifetime Extrapolation," *IEEE Transactions on Electron Devices*, vol. 41, pp. 761–767, 1994.

[33] B. Chatterjee, M. Sachdev, and R. Krishnamurthy, "A CPL-Based Dual Supply ALU for Sub-180nm CMOS Technologies," in *Proc. of IEEE International Symposium on Low Power Electronic Design (ISLPED)*, 2004, pp. 248–251.

[34] B. Chatterjee and M. Sachdev, "Designing of a 1.7GHz, Low Power, Delay Fault Testable 32-bit ALU for Sub-180nm CMOS Technologies," *IEEE Transactions on VLSI Systems*, submitted Sept. 2004.

[35] S. Vangal, N. Borkar, E. Seligman, V. Govindarajulu, V. Erraguntala, H. Wilson, A. Panagal, V. Veeramachaneni, M. Anders, J. Tschanz, Y. Ye, D. Somasekhar, B. Bloechel, G. Dermer, R. Krishnamurthy, S. Nagendra, M. Stan, S. Thomspon, V. De, and S. Borkar, "A 2.5GHz 32 bit Integer-Execution Core in 130nm Dual Vt CMOS," in *Proc. of IEEE International Solid-State Circuits Conference (ISSCC)*, 2002, pp. 412–413.

[36] S. Matthew, R. Krishnamurthy, M. Anders, R. Rios, K. Mistry, and K. Soumyanath, "Sub-500ps 64-b ALUs in 180nm SOI/Bulk CMOS: Design and Scaling Trends," *IEEE Journal of Solid-State Circuits*, vol. 36, pp. 1636–1646, 2001.

[37] W. Hwang, R. Joshi, and W. Henkels, "A 500MHz, 32 Word x 64 bit, Eight-Port Self Resetting CMOS Register File," *IEEE Journal of Solid-State Circuits*, vol. 34, pp. 56–67, 1999.

[38] A. Alvandpour, R. Krishnamurthy, K. Soumyanath, and S. Borkar, "A sub-130nm Conditional Keeper Technique," *IEEE Journal of Solid State Circuits*, vol. 37, pp. 633–638, 2002.

[39] S. Hsu, B. Chatterjee, M. Sachdev, A. Alvandpour, R. Krishnamurthy, and S. Borkar, "A 90nm 6.5GHz 256x64b Dual Supply Register File with Split Decoder Scheme," in *Proc. of VLSI IEEE Circuits Symposium*, 2003, pp. 237–238.

[40] S. Thompson, I. Young, and M. Bohr, "Dual Threshold and Substrate Bias: Keys to High Performance, Low Power, 0.1$\mu$m Logic Designs," in *Proc. of IEEE Symposium on VLSI Technology*, 1999, pp. 69–70.

[41] M. Anis, M. Allam, and M. Elmasry, "Energy-Efficient, Noise-Tolerant Dynamic Styles for Scaled down CMOS and MTCMOS Technologies," *IEEE Transactions on VLSI Systems*, vol. 10, pp. 71–78, 2002.

[42] V. Kursun and E. G. Friedman, "Domino Logic with Variable Theshold Keeper," *IEEE Transactions in VLSI Systems*, vol. 11, pp. 1080–1093, 2003.

[43] J. P. Halter and F. N. Najm, "A Gate-Level Leakage Power Reduction Method for Ultra-Low Power CMOS Circuits," in *Proc. of IEEE Custom Integrated Circuits Conference (CICC)*, 1997, pp. 475–478.

[44] B. Chatterjee, M. Sachdev, and R. Krishnamurthy, "Designing Leakage Tolerant, Low Power Wide-OR Dominos for Sub-130nm CMOS Technologies," *Microelectronics Journal*, submitted July, 2004.

[45] C. Kwong, B. Chatterjee, and M. Sachdev, "Modeling and Designing of Energy-Delay Optimized Wide-OR Domino Circuits," in *Proc. of IEEE International Symposium on Circuits and Systems (ISCAS)*, 2004, pp. 921–924.

[46] B. Chatterjee and M. Sachdev, "Low Power Leakage Control Techniques for Scalable Register File Designs in Sub-180nm CMOS Technologies," *IEEE Transactions on VLSI Systems*, submitted Sept. 2004.

[47] M. Sachdev, "Current-based Testing of Deep-submicron VLSIs," *IEEE Design and Test of Computers*, vol. 18, pp. 76–84, 2001.

[48] P. Nigh, W. Needham, K. Butler, P. Maxwell, R. Aitken, and W. Maly, "So What is an Optimal Test Mix? A Discussion of the Sematech Methods Experiment," in *Proc. of IEEE International Test Conference (ITC)*, 1997, pp. 1037–1038.

[49] A. Keshavarzi, K. Roy, and C. Hawkins, "Intrinsic Leakage in Low Power Deep Submicron CMOS ICs," in *Proc. of IEEE International Test Conference (ITC)*, 1997, pp. 146–155.

[50] H. Hao and E. M. Cluskey, "Very-Low Voltage Testing for Weak CMOS Logic ICs," in *Proc. of IEEE International Test Conference (ITC)*, 1993, pp. 275–284.

[51] V. D. Agrawal and T. J. Chakraborty, "High-Performance Circuit Testing with Slow-Speed Tester," in *Proc. of IEEE International Test Conference (ITC)*, 1995, pp. 302–310.

[52] M. Shashaani and M. Sachdev, "A DFT Technique for High Performance Circuit Testing," in *Proc. of IEEE International Test Conference (ITC)*, 1999, pp. 267–285.

[53] J. Park, H. Ngo, J. Silberman, and S. Dhong, "470ps 64-bit Parallel Adder," in *Proc. of IEEE the Symposium on VLSI Circuits*, 2000, pp. 192–193.

[54] B. Chatterjee, M. Sachdev, and A. Keshavarzi, "A DFT Technique for Low Frequency Delay Fault Testing in High Performance Digital Circuits," in *Proc. of IEEE International Test Conference (ITC)*, 2002, pp. 1130–1139.

[55] ——, "DFT for Delay Fault Testing of High Performance Digital Circuits," *IEEE Design and Test of Computers*, vol. 21, pp. 248–258, 2004.

[56] ——, "A DFT Technique for Delay Fault Testability and Diagnostics in 32-bit High Performance CMOS ALUs," in *Proc. of IEEE International Test Conference (ITC)*, 2004, pp. 1108–1117.

# Appendix A

# Published Papers

- B. Chatterjee, M. Sachdev, and R. Krishnamurthy, "Designing of Low Power Robust Wide-OR Dominos for sub-130nm CMOS Technologies," (Special issue of Microelectronics Journal, submitted July 2004)

- B. Chatterjee, and M. Sachdev, "Dual Supply, CPL based 32-bit ALU design with DFT for Delay Testability and Diagnostics in sub-180nm Technologies," (submitted to Transactions on VLSI Systems, Sept. 2004)

- B. Chatterjee, and M. Sachdev, "Comparative Analysis and Design of Low Power, Robust, Wide Dominos for High Performance RFs," (submitted to Transactions on VLSI Systems, Sept. 2004)

- B. Chatterjee, M. Sachdev, A. Keshavarzi, "DFT for Delay Fault Testing of High Performance Digital Circuits," IEEE Design and Test of Computers (Special issue on Design for Yield and Reliability)", pp. 248-258, vol. 21, no. 3, May 2004

- B. Chatterjee, M Sachdev, R. Krishnamurthy, "A CPL Based Dual Supply 32-bit ALU Design for Sub-180nm CMOS Technologies," Proc. of IEEE ISLPED 2004, Newport Beach, pp. 248-251, Aug. 2004.

- B. Chatterjee, M Sachdev, A. Keshavarzi, "A DFT Technique for Delay Fault Testability and Diagnostics for 32-bit High Performance CMOS ALUs," Proc. of IEEE ITC 2004, Charlotte, pp. 1108-1117, Oct. 23-28, 2004.

- C. Kwong, B. Chatterjee, M. Sachdev, "Modelling and Designing Energy-Delay Optimized Wide Domino Circuits," Proc. of IEEE ISCAS 2004, Vancouver, pp. 921-924, May 23-26, 2004

- B. Chatterjee, M. Sachdev, R. Krishnamurthy, "Leakage Control Techniques for Designing Robust Low-Power Wide-OR Domino Logic for sub-130nm CMOS Technologies," Proc. of IEEE ISQED 2004, San Jose, pp. 415-420, March 22-24, 2004

- B. Chatterjee, M. Sachdev, S. Hsu, R. Krishnamurthy, S. Borkar, "Effectiveness and Scaling Trends of Leakage Control Techniques for sub-130nm CMOS Technologies," Proc. of IEEE ISLPED 2003, Seoul, pp. 122-127, August 25-27, 2003

- S. Hsu, B. Chatterjee, M. Sachdev, A. Alvandpour, R. Krishnamurthy, S. Borkar, "A 90nm 6.5GHz 256x64b Dual Supply Register File with Split Decoder Scheme," Proc. of IEEE VLSI Circuits Symposium 2003, Hawaii, pp. 237-238, June 12-14, 2003

- B. Chatterjee, M. Sachdev, A. Keshavarzi, "A DFT Technique for Low Frequency Delay Fault Testing in High Performance Digital Circuits," Proc. of IEEE ITC 2002,

Baltimore, pp. 1130-1139, October 7-10, 2002

- R. Krishnamurthy, S. Hsu, M. Anders, B. Bloechel, B. Chatterjee, M. Sachdev, S. Borkar, "Dual Supply Voltage Clocking for 5 GHz 130 nm Integer Execution Core," Proc. of IEEE VLSI Circuits Symposium 2002, Kyoto, pp. 128-129, June 13-15, 2002

- O. Semenov, B. Chatterjee and M. Sachdev, "Impact of Technology Scaling on Bridging Fault Modelling in CMOS Circuits," IEEE Intl. Workshop on Defect Based Testing (DBT), pp. 45-50, April 2001

# Appendix B

# Patents

- S. Hsu, B. Chatterjee, R. Krishnamurthy, "Level Converting Latch," US Patent No. 6563357, Granted May 13, 2003
- S. Hsu, B. Chatterjee, R. Krishnamurthy, "Low Clock Swing Latch for Dual-Supply Voltage Design," US Patent Application No. 20030117933, Filed June 26, 2003
- A. Keshavarzi, B. Chatterjee, R. Krishnamurthy, M. Sachdev, "Low Frequency Testing, Leakage Control, and Burn-In Control for High-Performance Digital Circuits," US Patent Application No. 20040051558, Filed March 18, 2003
- B. Chatterjee, S. Hsu, R. Krishnamurthy, "A Register File with a Selectable Keeper Circuit," Patent pending
- B. Chatterjee, S. Hsu, R. Krishnamurthy, "A Leakage Tolerant Register File," Patent pending