Improving Performance, Power Efficiency, Yield, and Reliability Using Programmable Power-gating Techniques

By

Abhishek Arvind Sinkar

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Electrical Engineering)

at the

UNIVERSITY OF WISCONSIN – MADISON

2012

Date of final oral examination: 08/27/2012

The dissertation is approved by the following members of the Final Oral Committee:

Nam Sung Kim, Assistant Professor, Electrical and Computer Engineering

Katherine Compton, Associate Professor, Electrical and Computer Engineering

Karthikeyan Sankaralingam, Assistant Professor, Computer Sciences

Mikko H. Lipasti, Philip Dunham Reed Professor, Electrical and Computer Engineering

Kewal K. Saluja, Professor, Electrical and Computer Engineering

To my family ...with gratitude for their love and encouragement.

Acknowledgments

Only when I started writing these acknowledgments, I realized how difficult it can be to remember and include everyone who contributed to the success of this work in some way. The list is very long and I would like to apologize in advance if I leave out any deserving individual.

First, I would like to thank my advisor, Prof. Nam Sung Kim, for giving me an opportunity to work with him. Nam came in my life at a point when I had been in graduate school for nearly four years and when my PhD dream was faltering. Slowly and patiently, he guided me till I reached this day. During the course of this research, I gained from him not only technical knowledge but also a careful research attitude and ethics.

I will always be indebted to my family in India and the United States for their unconditional love and encouragement, especially to my parents for overlooking my negligence of my responsibilities towards them in the last few years.

I also would like to thank my PhD committee members, Professor Katherine Compton, Professor Kewal K. Saluja, Professor Karu Sankaralingam, and Professor Mikko H. Lipasti for their precious time in reviewing my thesis and their valuable advice to improve my work.

I would like to thank Brian Busse and family for providing me with a "home" during my years in Madison. Thanks to them, the pain of missing my family was reduced.

A long journey becomes pleasant and bearable in the company of kind and friendly fellow-travelers. I was fortunate to meet some very nice and kind people in the form of my

colleagues in the lab and friends I made at UW-Madison. I will not mention names to avoid the embarrassment of leaving anyone out but I will never forget the help granted by each one of them in my times of need.

Finally, I would like to sincerely thank the Center for Quick Response Manufacturing at UW-Madison for giving me an opportunity to work with them during my initial years in Madison.

Abstract

As technology scales, power consumption, variability, and reliability have become serious concerns in processor design. Increasing transistor densities and subthreshold leakage have increased chip power dramatically. High power consumption makes the use of affordable packaging and cooling solutions difficult in high-performance processor platforms while it degrades battery life of mobile processors. As a result, manufacturers now focus on maximizing performance/Watt rather than raw processing performance. Further, smaller transistors are more susceptible to process variations and aging effects like bias temperature instability (BTI). Leakage, variability, and aging reduce parametric yield, i.e., the number of functional dies which meet the frequency and power constraints. For a long time, designers have addressed variability and aging by incorporating design margins. However, to maximize power-efficiency, it is necessary to reduce design margins by using dynamic techniques. The shift to multi- and many-core architectures presents new challenges for implementing such dynamic power, variability, and reliability management techniques, as it is necessary to keep their design, verification, and test cost low in the multi- and many-core environment.

Power-gating (PG) is commonly used to reduce standby leakage power in multi-core processors. The PG device incurs an overhead in terms of chip area and it is impacted by BTI and process variations, which affects the frequency and power of the power-gated circuit. On the other hand, it provides a knob to control the voltage (hence frequency

and power) of the connected processor core. This thesis presents low-cost, static as well as dynamic techniques to optimize the PG device for improving performance, power efficiency, yield, and reliability of multi-core processors.

First, a method of post-manufacturing tuning of PG device to improve yield and performance of multi-core processors with a power constraint, in the presence of die-to-die (D2D) and within-die (WID) variations, is presented. To improve yield in presence of D2D variations, the strength of the PG device is adjusted in fast-but-leaky dies such that they can operate in an acceptable power and frequency region. Simulations with ISCAS benchmark circuits demonstrate that ∼88% and 98% of discarded leaky dies can be recovered by the proposed method in fixed frequency and variable frequency designs, respectively. In processors with a shared voltage domain but individual PG domains, when each core shows different frequency due to WID variations, the strength of a PG device of each core is adjusted to make their maximum operating frequencies even. This allows faster cores to consume less active leakage power, reducing the total power consumption well below a power constraint in a globally-clocked design. Subsequently, global supply voltage is increased for higher overall frequency until the power constraint is satisfied. The PG tuning improves the performance by 3%-21% on average for 2-, 4-, 8-, and 16-cores in presence of WID variations.

Second, a circuit technique is proposed to track the BTI aging of a PG device and chip temperature variation, and adjust the strength of PG device dynamically. This dynamic technique eliminates the design margins used to account for aging and high temperature. As a result, leakage power is reduced by ∼10% and dynamic power is reduced by ∼4% in early chip lifetime and low temperatures. The proposed tracking technique also reduces the gate oxide failure rate by 5.1%, 3.8%, and 4.1% for fast, nominal, and slow process corners over a period of 7.5 years.

Third, the impact of PG size on frequency and power of multi-core processors with

process variations is analyzed. It is shown that increasing the size of a PG device beyond a certain value does not improve the frequency of the power-gated circuit while it increases the power consumption significantly. A PG device is often sized to minimize the voltage drop (thus the frequency degradation), requiring considerable die area. Meanwhile, adaptive voltage scaling (AVS) has been used to improve yield of power-constrained processors exhibiting a large spread of maximum frequency and total power due to process variations. Based on the above observation, a methodology to optimize both the size of PG devices and the degree of AVS jointly, is proposed such that the PG size is minimized while maximizing performance and power efficiency of power-constrained processors. The PG sizing methodology is applied to multi-core processors adopting global and frequency-island clocking schemes. Simulation results demonstrate that the joint optimization considering both D2D and WID variations reduces the size of power-gating devices by more than 50% with $\sim$3% performance improvement for power-constrained multi-core processors.

Fourth, the use of per-core PG device as on-chip, linear voltage regulator (VR) is proposed as a cost-effective way to provide per-core voltage domains in multi-core processors. Experimental results show that when core-to-core voltage variations in a multi-core processor are relatively small, the $MIPS^3/W$ of a processor using the proposed VRs is slightly higher than that of a processor using switching VRs. In addition, a VR using PG device requires significantly less area compared to a switching VR and can be implemented with low cost as it shares its principal component with the PG device and does not require bulky inductors which are challenging to manufacture.

Finally, in processors with per-core voltage/frequency domains and power and thermal constraints, a method is proposed to optimize the voltage/frequency of cores running highly compute-bound, multi-threaded workload. In the presence of WID variations, fast cores are often subject to thermal throttling which reduces the overall performance of the processor. Although, techniques such as thread migration are used to solve the thermal throttling

problem, there is not much opportunity for thread migration when all cores are running highly compute-bound applications most of the time. In such a case, the proposed method can balance the core temperatures leading to less frequent thermal throttling of fast cores and thus higher maximum performance under power and thermal constraints.

Contents

Acknowledgments **ii**

Abstract **iv**

List of Figures **xii**

List of Tables **xv**

1 Introduction **1**

1.1 Motivations 1

1.1.1 Power Management 4

1.1.2 Process Variations 5

1.1.3 Device Aging Effects 6

1.2 Thesis Contributions 7

1.2.1 Yield and performance improvement 7

1.2.2 Active leakage reduction and gate oxide reliability improvement . . . 8

1.2.3 AVS-aware PG sizing algorithm 8

1.2.4 Low cost voltage regulation using PG device 9

1.2.5 Performance improvement of power/thermal constrained processors with per-core V/F domains 9

1.3 Organization 10

2 Background **11**

2.1 Power Gating 11

2.1.1 Physical Implementation 13

2.1.2 Power Gating Challenges 14

2.2 Power Delivery 15

2.2.1 Voltage Regulators 17

2.3 Process Variations 19

2.4 Circuit Reliability 21

3 Frequency and Yield Optimization Using Programmable Power-gating **23**

3.1 Programmable-width Power Gating 24

3.2 Impact on Delay, Leakage, and VV_{DD} 25

3.3 Die-to-die Variation-aware Yield Improvement with PPG 27

3.3.1 Simulation Methodology for Yield Experiments 28

3.3.2 Designs with Frequency Target: Fixed P_{LEAK} Constraint 28

3.3.3 Designs with Frequency Binning: Variable P_{LEAK} Constraint 30

3.3.4 Impact of Limiting VV_{DD} Drop 34

3.4 Frequency Improvement with Multiple PPG Domains 36

3.4.1 Example of F_{MAX} Improvement in a Quad-core Processor 37

3.4.2 F_{MAX} Optimization Problem Formulation 38

3.4.3 Simulation Methodology for F_{MAX} Optimization 39

3.4.4 F_{MAX} Optimization Results 40

3.5 Related Work 42

3.6 Chapter Summary 43

4 Active Leakage Reduction and Gate Oxide Reliability Improvement of Power-gated Circuits **45**

4.1 Impact of Temperature on VV_{DD} and Active Leakage Power 46

4.2 Impact of NBTI on VV_{DD} and Active Leakage Power 48

4.3 Auxiliary PG Device for Clamping VV_{DD} 50

4.4 NBTI Tracking Scheme 52

4.4.1 Simulation Setup 54

4.4.2 Active Leakage Reduction with NBTI Tracking 57

4.5 Temperature Tracking Scheme 58

4.6 Tracking with Within-die Spatial Process and Temperature Variations . . . 60

4.6.1 Simulation Setup for Tracking with WID Variations 61

4.6.2 Tracking circuit performance with WID variations 62

4.7 Impact of VV_{DD} Clamping on Gate oxide Reliability 63

4.8 Related Work 65

4.9 Chapter Summary 67

5 AVS-aware Power-gate Sizing for Multi-core Processors **69**

5.1 Impact of D2D Variations and PG Size on VV_{DD}, F_{MAX}, and P_{TOT} 70

5.2 AVS-aware PG Size Optimization for Performance and Power Efficiency . . 73

5.2.1 PG Size Optimization Problem Formulation 74

5.3 AVS-aware PG Sizing with WID Variations 77

5.3.1 WID Variations with Global Clocking 78

5.3.2 WID Variations with Frequency-Island Clocking 81

5.4 Simulation Methodology 84

5.5 Chapter Summary 86

6 Low Cost Per-core Voltage Regulation Using Power-gating Device **88**

6.1 Per-core Voltage Domains 89
6.1.1 Motivation for Per-core Voltage Domains 89
6.1.2 Challenges for Supporting Per-core Voltage Domains 91
6.2 LDO VRs Exploiting C2C Voltage Variations and PCPG device 93
6.2.1 C2C Voltage Variations 93
6.2.2 PCPG Based LDO VRs 94
6.2.3 LDO vs. Switching VR Efficiency Comparison 97
6.3 Evaluation of DVFS with LDO VR 100
6.3.1 DVFS Algorithms for Efficiency Comparison 100
6.3.2 Architectural Simulation Environment 102
6.3.3 Core Frequency and Power Modeling 103
6.3.4 MIPS^3/W Comparison 105
6.4 Related Work 111
6.5 Chapter Summary 112

7 Throughput Optimization under Power and Thermal Constraints 114
7.1 C2C Frequency, Power, and Temperature Variations 115
7.2 Throughput model for Per-core Clocking 117
7.3 Maximizing Performance Under Power Constraint 120
7.4 Maximizing Performance Under Power and Thermal Constraints 124
7.5 Implementation Cost 126
7.6 Related Work 127
7.7 Chapter Summary 127

8 Conclusions 129

List of Figures

1.1 ITRS projection for maximum allowable power 3

2.1 Logical structure of a PG device 12

2.2 Physical layout styles of PG switch 14

2.3 Circuit schematic of buck voltage regulator 17

2.4 Circuit schematic of linear voltage regulator 18

2.5 Systematic WID variation map for a 16-core processor 19

2.6 Percentage degradation in V_{TH} of PMOS due to NBTI 21

3.1 Programmable width PG device 25

3.2 Leakage, Delay, VV_{DD}, and P_{PG} 26

3.3 F_{MAX} and P_{LEAK} distribution of ISCAS85 C432 after PG optimization . . 29

3.4 V_{TH} variation map and initial F_{MAX} and P_{LEAK} of quad-core processor . . 37

3.5 F_{MAX} and P_{LEAK} of quad-core processor after F_{MAX} equalization with PPG 38

3.6 Simulation setup to measure dynamic and leakage power in PG circuit . . 40

3.7 Avg. F_{MAX} improvement with PPG and V_{DD} tuning 41

4.1 VV_{DD} vs. total current through PG device at high and low temperature . . 47

4.2 Impact of temperature on VV_{DD} and leakage current across process corners 48

4.3 V_{TH} and VV_{DD} degradation due to NBTI 49

4.4 Impact of PG upsizing on VV_{DD} and I_{LEAK} in early chip life 50

4.5 Auxiliary PG device for NBTI and Temperature compensation 51

4.6 A scheme for tracking NBTI degradation of PG device 53
4.7 Simulation setup for VV_{DD} clamping experiment with 55
4.8 VV_{DD} and P_{LEAK} versus time with nbti tracking 57
4.9 A scheme for tracking temperature variations 59
4.10 VV_{DD} and P_{LEAK} versus temperature with temperature tracking 60
4.11 Within-die process and temperature variation map 62
4.12 VV_{DD} tracking with WID process and temperature variations 63
4.13 Time to breakdown versus Voltage for 1nm oxide 64
4.14 Oxide failure rate reduction over time by clamping VV_{DD} 66
5.1 Impact of D2D process variations on VV_{DD} 71
5.2 VV_{DD} across process corners and versus PG size 72
5.3 F_{MAX} and P_{TOT} as function of PG size 73
5.4 P^3/W vs. PG size and optimized PG size, F_{MAX}, and V_{DD} 75
5.5 Systematic WID variation map for a 16-core processor 78
5.6 p^3/w and optimum PG size and V_{DD} with global clk. and WID variations . 80
5.7 Speedup with frequency island clocking relative to global clocking 82
5.8 V_{DDOPT}, s_{PGOPT} versus no. of cores 84
6.1 Comparison of $MIPS^3$/W with Chip-wide and per-core DVFS 90
6.2 Impact of splitting the voltage domain on the overall VR capacity 92
6.3 Maximum voltage difference between cores with individual V/F domains. . 94
6.4 Low dropout VR architecture 96
6.5 Efficiency comparison between switching and LDO VRs 98
6.6 $MIPS^3$/W comparison of 8-core processors supported by LDO and switching VRs (VR power loss not included) 106
6.7 $MIPS^3$/W comparison of 8-core processors supported by LDO and switching VRs (power loss in on- and off-chip VRs included) 107

6.8 MIPS^3/W comparison of 8-core processors supported by LDO and switching VRs in multi-program environment 109
6.9 MIPS^3/W comparison of 8-core processors supported by LDO and switching VRs in multi-program environment (no core voltage constraint) 111
7.1 Systematic WID Variations and Normalized F_{MAX} and P_{LEAK} in 16-core Processor 115
7.2 Temperature profile of a 16-core processor with per-core clocking 117
7.3 Throughput versus average F_{MAX} values of 16-core processors 118
7.4 F_{MAX} vs. P_{TOT} of fast and slow cores in a die 121
7.5 $F_{MAX,i}$ & $P_{TOT,i}$ of 16-core die sample before and after optimization 122
7.6 Distribution of avg. F_{MAX} improvement 123
7.7 Temperature profile of 16-core die after optimization 125
7.8 Distribution of avg. F_{MAX} improvement under power and thermal constraints 126

List of Tables

3.1 Yield loss recovery for fixed P_{LEAK} constraint 31
3.2 Yield loss recovery for variable P_{LEAK} constraint 33
3.3 Yield loss recovery wiyh PG drop limited to 100mV and 150mV 35
4.1 NBTI model parameters used to calculate ΔV_{TH} 56
5.1 AVS-aware optimum PG size, V_{DD}, and die F_{MAX} across process corners. . 77
5.2 Key hardware parameters for GPGPUSim simulations 85
6.1 Summary of VR design parameters 97
6.2 Summary of DVFS Algorithms. 101
6.3 Processor Simulation Parameters. 102
6.4 Frequency and Power consumption of each core as function of $V_O[i]$. 105
7.1 Per-core Clocking Throughput Model Parameters 119

Chapter 1

Introduction

1.1 Motivations

Since the invention of the transistor in 1947, the computing industry has grown tremendously while positively impacting every area of human life. Due to the advances in semiconductor manufacturing technology, processor manufacturers have been able to produce faster and more powerful processors every two years, for the last six decades. Shrinking process technology allows more and faster transistors to be integrated in a given die area, thereby allowing a continuous improvement in performance. Processor performance is given by the Iron Law equation as:

$$Performance = \frac{1}{No.\ of\ Instructions} \times \frac{Instructions}{Cycles} \times Clk\ freq. \tag{1.1}$$

For years, manufacturers improved performance by increasing clock frequency and instructions per cycle (IPC) by evolving microarchitecture to exploit instruction-level parallelism (ILP) in programs. As a result, high-performance computing systems are capable of achieving performance of the order of petaflops (10^{15} floating point operations per second). This tremendous increase in performance is driven by the simulation needs of models

from the disciplines of astrophysics, biology, climate and weather forecasting, and national security, to name but a few. The US Department of Energy's target for high-performance computing systems is to reach exaflops of performance with a total power consumption limited to 20MW by 2018 [1]. This represents a 3× increase in the power consumption and 500× increase in performance of present systems which deliver 2 petaflops at 6MW of power. These performance and power targets will require a power management solution for an exaflop system which is over 150 times more efficient than current technology. Improving performance and power efficiency simultaneously is a significant challenge for processor manufacturers. Both increasing clock frequency and microarchitectural innovations to improve performance increase power consumption.

The power dissipated in a CMOS circuit can be divided into two types; the power due to switching capacitance of the transistors, also called dynamic power (P_{DYN}), and the power dissipated in idle transistors due to leakage current, (P_{LEAK}). P_{DYN} is given by:

$$P_{DYN} = \alpha \cdot C \cdot V^2 \cdot f \tag{1.2}$$

where α is the switching activity factor ($0 < \alpha < 1$) of a circuit node, C is the node capacitance, V is the node voltage change during the switching transition and f is the switching frequency. Although the per-transistor P_{DYN} decreases with technology scaling, chip-wide P_{DYN} has been increasing due to increase in the effective switching capacitance per unit chip area and die size. The effective switching capacitance per unit chip area increases by ∼40% with each technology generation (capacitance of a transistor changes by 0.7× while area changes by 0.5×). Die size increases by ∼25% every technology generation due to increase in the amount of last level cache needed and improvements in fabrication technology which enable higher yield for large reticle sizes. P_{LEAK} depends on the applied voltage and is proportional to the number of off transistors in a given chip area. P_{LEAK} increases with technology scaling due to increased transistor density and lower threshold

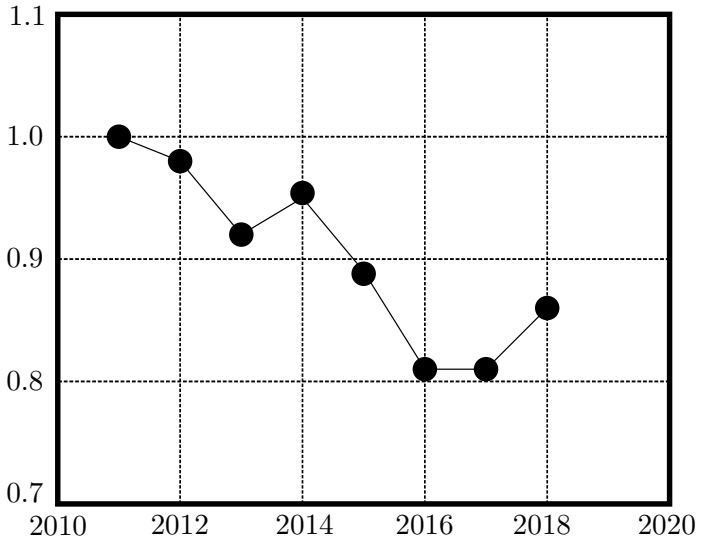


Figure 1.1: ITRS projection for maximum allowable power for high-performance processors. (Normalized to 2011 value [2].)

voltage. For a given process technology, microarchitectural methods to exploit ILP often increase the transistor count which also increases P_{LEAK}. Thus, both P_{DYN} and P_{LEAK} have increased with technology scaling, resulting in the total power reaching levels which make the use of affordable packaging and cooling solutions practically impossible. Further, higher power consumption necessitates the use of bigger voltage regulators (VRs), increasing platform cost and size. In the area of mobile and embedded processors, increasing power consumption leads to reduced battery life. Figure 1.1 shows the *international technology roadmap for semiconductors* (ITRS) [2] projection for maximum allowable chip power for high-performance and cost-performance processors normalized to the 2011 value. Modern high-performance processor platforms have stabilized maximum power dissipation at approximately 120W due to package cost, reliability, and cooling cost issues. As a result, performance gains in future processors will have to be achieved while staying within the power envelope which requires use of low power design methodology at the circuit, architecture, and software levels. Partly as a result of the power wall, processor designers have adopted new design methodologies in the recent years. Clock frequency has flattened circa

2005 and the trend is towards integrating more processing units or cores in a die. Such cores operate at low power consumption while delivering required performance with the aid of multi-threading support and efficient on-chip interconnect network. At the same time, high-performance computing, traditionally confined to scientific laboratories, is now making a move to the mainstream. Multi-core processors are also being widely adopted in the arena of mobile devices because they allow complex applications to be written and executed concurrently for devices like smart-phones and tablets which can provide the user with more visually rich graphics and multimedia experiences.

1.1.1 Power Management

The shift to multi-core processors was considered to be beneficial to reducing power consumption due to the notion that high throughput can be achieved by dividing a task into several parallel portions, each running on a core at a lower voltage/frequency. However, single-threaded or sequential performance is still required in many applications due to inherently limited parallelism and yet distant mainstream adoption of parallel programming. Further, increase in the number of cores causes increase in the uncore (i.e., shared on-chip cache, interconnect, etc) power dissipation and area to maintain scalability of performance. A major component ($\sim$ 40%) of total chip power dissipation in today's high-performance processors is sub-threshold leakage power which increases with technology scaling [3]. As a result, the power problem continues to exist even in the multi-core era as the power per core does not scale down at the same rate as the increase in number of cores.

The root cause of the power problem lies in the packaging and cooling solutions' ability to remove heat from the processor die. Technological advancements in packaging and cooling solutions have traditionally lagged behind those in processor technology. Also, platform cost increases when high-end packaging and cooling solutions are employed. As a result, today's processors are designed to operate with a limited power budget, also known as

thermal design power (TDP), which is the maximum average power that the cooling solution must dissipate. Given the limits on how much power consumption can be reduced through fundamental improvements in semiconductor and cooling technology, the only way in which today's designs can work is by wisely utilizing the power budget that is available.

A commonly employed technique is to distinguish between active and idle (standby) states for a system component. System components (e.g cores, caches, translation lookaside buffer arrays) which are not being utilized can be put into a standby state by powering them down with power-gating (PG) devices. Components which store data related to the system state can be put into a data retention low power state by lowering their supply voltage. During a component's active state, power reduction can be obtained by adjusting the performance of the component depending on the needs of the workload. Performance tuning is achieved by varying the voltage/frequency, also known as dynamic voltage and frequency scaling (DVFS). In multi-core processors, fine-grained power/performance trade-off can be achieved by individual voltage/frequency scaling of the cores. Such fine-grained control requires VRs per core or per group of cores. As the number of cores increases, however, providing per-core off-chip VR increases platform size and cost prohibitively. There has been an effort in the research community and industry to integrate VRs on the processor die. However, on-chip VRs suffer from low efficiency due to low quality integrated inductors in CMOS technology and increase the design time and effort. Hence, a low cost and complexity voltage regulation scheme with efficiency comparable to off-chip switching VRs is needed.

1.1.2 Process Variations

Technology scaling worsens shifts in transistor parameters such as threshold voltage (V_{TH}), channel length (L_{EFF}), gate oxide thickness (T_{OX}), and channel doping. Parameter variations broadly fall in two categories: inter-die (die-to-die (D2D)) and intra-die (within-

die (WID)). Inter-die variations include variations that arise between different dies in the same wafer or across different wafers or wafer lots, while intra-die variations account for variations that arise within the same die or more generally within a reticle field. These device level variations are manifested as variations of circuit delay and leakage power among manufactured dies and also among cores in a die. Dies that do not meet the frequency target or power constraint have to be discarded which reduces the yield. Variations are addressed by incorporating design margins and binning of manufactured parts. Margins are added to device sizes, process parameters such as doping, and supply voltages such that manufactured dies which are slow meet the target frequency. In high-performance processors, post manufacturing frequency binning along with adaptive voltage scaling (AVS) and adaptive body biasing (ABB) is used to improve yield by sorting dies into several frequency bins. Leaky dies are accepted by placing them in low frequency bins with reduced voltage or reverse body bias which reduces profit. Within-die variations such as core-to-core(C2C) frequency and leakage power variations can provide new opportunities for power management and performance improvement in multi- and many-core processors. The speed of fast cores can be traded with useful power headroom by doing static (one-time) optimization techniques such as post manufacturing tuning of circuits as well as dynamic optimization techniques such as voltage/frequency scaling and thread migration. The benefits and overheads of such optimization techniques warrant thorough analysis considering the design and operating space involved.

1.1.3 Device Aging Effects

CMOS circuits undergo aging and even irreversible failure due to the continuous stress applied to the devices during circuit operation. Aging effects such as bias temperature instability (BTI) and oxide degradation create long term shifts in device parameters such as threshold voltage and gate tunneling current which translate into circuit frequency degra-

dation and power increase with time. Further, circuit reliability is impacted by process variations. BTI depends on threshold voltage while the time to breakdown of oxide is a function of oxide thickness. Both these parameters show a statistical distribution within a chip and across chips. Design margins are commonly used to counter the impact of aging. However, margins can lead to deterioration of other circuit metrics such as leakage power or area overhead in the absence of aging (early chip life). As a result, dynamic techniques, which compensate for aging based on usage conditions, are desirable.

1.2 Thesis Contributions

This thesis is a multi-faceted work which proposes novel methods to improve performance, power efficiency, yield, and reliability of multi-core processors. To achieve these objectives, our methods optimize the structure and size of PG device with post-manufacturing as well as dynamic runtime tuning. The original contributions of this thesis are as follows:

1.2.1 Yield and performance improvement

For multi-core processor with D2D F_{MAX} and P_{LEAK} variations, a post-silicon method is shown to optimize the strength of the PG device to improve the parametric yield. The PG devices of leaky dies can be weakened during a post-manufacturing stage to reduce their virtual rail voltage (VV_{DD}). This reduces the P_{LEAK} (hence P_{TOT}) of these leaky dies. Dies which would be discarded due to violation of P_{TOT} constraint are recovered with small overhead of tuning.

The post-silicon PG tuning can be applied on a per-core basis to multi-core processors to increase the operating frequency. The PG devices of fast (and leaky) cores can be weakened during a post manufacturing stage to reduce their VV_{DD} and P_{LEAK}. The resulted power headroom can be used to increase the global supply voltage and clock frequency F_{MAX} of the processor. The F_{MAX} increase is limited by the power constraint of the die. The

method relies on characterization of F_{MAX} and P_{LEAK} of the cores. Modern processors with dynamic voltage/frequency scaling (DVFS) feature are subject to such characterization during manufacturing testing for purpose of speed binning. Hence, the PG tuning method can be incorporated in existing test flow with small additional overhead.

1.2.2 Active leakage reduction and gate oxide reliability improvement

PG devices are often upsized during design to account for negative bias temperature instability (NBTI) (or PBTI in NMOS) aging which leads to undesirable, high active leakage power in early chip life. Leakage power dissipated in circuit blocks which are in use (i.e., active) is refered to as active leakage. This thesis proposes a NBTI tracking mechanism and PG device architecture which adjusts its strength based on usage (aging) to clamp the VV_{DD} to a fixed value. Our simulation results show that our clamping technique reduces the active leakage power by ∼10% and dynamic power by ∼4% in early chip life. Further, the analysis presented shows that active leakage of power-gated circuits does not decrease as much as expected when chip temperature falls. At low temperatures VV_{DD} of a power-gated circuit increases resulting in higher active leakage. Another tracking circuit and PG device to reduce active leakage in presence of temperature variations is also demonstrated. Both aging and temperature tracking methods are evaluated in the presence of WID variations.

The impact of the VV_{DD} clamping technique on gate oxide reliability is analyzed. With the proposed clamping method, gate-oxide failure rate decreases by up to 5%, 3.8%, and 4.1% for fast, nominal, and slow process corners over a period of 7 years.

1.2.3 AVS-aware PG sizing algorithm

First, we analyze the impact of PG size on the VV_{DD} and F_{MAX} across process corners. PG devices are often sized such that the connected circuit meets timing specifications under the worst-case current, process corner, and aging. This causes PG devices to occupy a

significant die area. The analysis shows that as the PG size is increased beyond a certain range, the F_{MAX} gain diminishes quickly while power increases at a faster rate than F_{MAX}. Based on the above analysis, an algorithm to optimize the size of the PG device is proposed. The algorithm utilizes the AVS technique commonly used to compensate the impact of process variations in manufactured processor dies. Across process corners, the sizing technique reduces the PG device size by up to 43%. In the presence of WID variations the sizing algorithm reduces the PG size by 58%. The proposed sizing has negligible impact on the F_{MAX} and leads to F_{MAX} improvement when WID variations are considered.

1.2.4 Low cost voltage regulation using PG device

H. Gashemi *et al.*[4] analyzed the voltage demand of cores in an eight-core server class processor running a variety of benchmarks and showed that the maximum voltage difference between cores with per-core DVFS is less than or equal to 100mV for at least 90% of the execution intervals in most applications. Based on this observation, a cost-effective and efficient method for providing per-core voltage domains in a multi-core processor is proposed. The proposed technique uses per-core PG devices to regulate the core voltages with a small area overhead. The area cost and efficiency of the proposed regulator are analyzed and compared with a switching regulator. It is shown that the proposed linear regulator can achieve higher MIPS^3/W on a multi-core processor than a switching counterpart while occupying a significantly lower die area.

1.2.5 Performance improvement of power/thermal constrained processors with per-core V/F domains

An optimization technique is shown to maximize the average F_{MAX} (i.e., throughput) of power- and thermal-constrained processors considering the WID C2C variations. In a multi-core processor, WID C2C variations result in different power and frequency trade-offs

between fast and slow cores. The throughput is proportional to the average frequency of cores in a processor when the processor allows each core at its own frequency (i.e., per-core clocking). Further, fast cores, which consume more power due to higher frequency and leakage, experience more thermal throttling than slow cores. The proposed optimization technique searches the most performance/power-efficient voltage/frequency point for each core during a post-silicon tuning process and results in less thermal throttling of leaky cores. Hence, it can provide higher average F_{MAX} for the processor under a power constraint. Simulation results show that the maximum throughput of 16-core processors with high C2C power variance can be improved by nearly up to 10% with the proposed optimization.

1.3 Organization

This thesis is organized as follows. Chapter 2 provides background material on topics of power gating, power delivery and VRs, process variations and yield and device aging. Chapter 3 presents the frequency and yield improvement method using programmable PG. Chapter 4 demonstrates leakage power reduction and gate oxide reliability improvement in the presence of BTI aging and temperature variations using a tracking circuit and auxiliary PG device. Chapter 5 presents the optimum PG sizing methodology while maximizing performance and power efficiency of multi-core processors. Chapter 6 investigates the feasibility of using per-core PG as cost-effective on-chip linear regulators for efficient power delivery in multi-core processors. Chapter 7 presents per-core voltage/frequency tuning in presence of within-die variations and thermal and power constraints. Chapter 8 summarizes the research and presents concluding remarks.

Chapter 2

Background

This chapter presents an overview of state-of-the-art in power gating (PG) and power delivery in multi-core processors including voltage regulators (VR). A brief background on process variations and circuit reliability phenomenon of bias temperature instability is presented which forms the basis for understanding the research contributions in the following chapters.

2.1 Power Gating

A PG device, also known as a sleep transistor, is an integrated on-off switch placed between the supply (or ground) rail and a digital circuit. Conceptually, such a PG device is illustrated in Figure 2.1. Both PMOS (header) and NMOS (footer) based implementations are possible. A PMOS device offers greater leakage reduction in the sleep state compared to an NMOS device for the same size. Circuits gated with footer are more sensitive to ground noise due to higher leakage of NMOS compared to those using headers [5]. However, in the on-state, a larger PMOS device is required compared to an NMOS for attaining a certain voltage drop due to the lower current drive of the PMOS. During the standby mode of the circuit, the PG device is turned off, disconnecting the circuit from the power supply

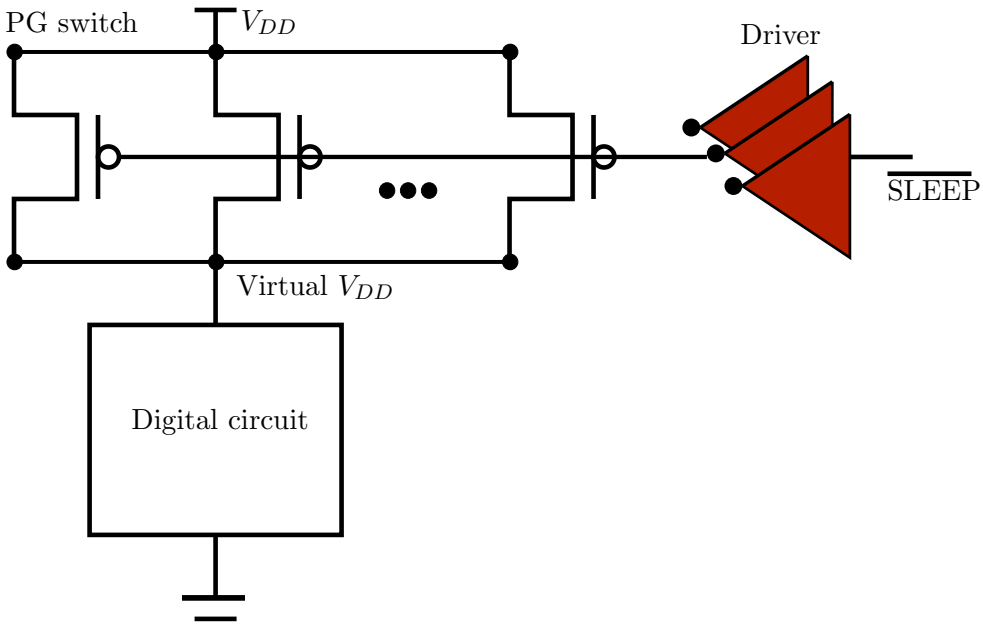


Figure 2.1: Logical structure of a power gating device.

which eliminates any standby leakage current flowing through the circuit. During the active mode, the PG device is turned on, connecting the circuit to the supply. In the on-state, the finite resistance of the PG device leads to a voltage drop across it which reduces the voltage applied to the circuit. In a PG circuit, the delay and power characteristics of the circuit are governed by the value of the virtual rail voltage (VV_{DD}) rather than V_{DD}. The voltage drop across the PG device is proportional to the total current (I_{TOT}) (i.e., dynamic + active leakage) drawn by the circuit and the effective width of the PG switch. Thus, the PG device is often sized to guarantee a minimum VV_{DD} level at the maximum I_{TOT} of the circuit at which the timing analysis for a multi-core processor is often performed. A common industry practice is to size the PG device for a voltage drop of 50mV$\sim$100mV at the worst case current drawn by the core [5]. Buffered drivers are typically needed to drive the large capacitance of the PG device which further adds to the area overhead of PG.

2.1.1 Physical Implementation

A PG device consists of a large number of transistors (fingers) in parallel between the supply rail and the circuit to be gated. One implementation approach is to embed a PG transistor in each standard cell [6]. This is advantageous from the point of view of design tools since the standard cell containing the PG switch can be characterized and its characteristics can be used by conventional computer aided design (CAD) tools for timing analysis, physical synthesis and so on. However, this approach consumes unnecessary chip area. A relatively smaller PG switch can be shared between cells considering the fact that all cells do not draw maximum current at the same time. When the PG switches are shared between cells, the standard cell is designed to include power, ground, and virtual rail. The cell library contains special power supply cells which consist of the PG header or footer.

These power supply cells are placed in each row as shown in Figure 2.2-(a). A single switch often consists of multiple fingers placed adjacent to each other depending on the required switch strength. Many industrial designs adopt the distributed PG network in Figure 2.2-(a) in one form or other. Another physical layout style for PG is the ring style in which the PG transistors are placed on the periphery of the circuit block and form a ring connecting the main and virtual rails as shown in Fig 2.2-(b). The ring style is less complicated to place and route than the distributed mesh since the PG switches and part of the power network are separated from the logic cells. Cells that require the permanent power supply such as isolation cells can be placed around the power domain areas without having to be distributed. With the advent of multi- and many-core designs and as core sizes become smaller with each technology generation, per-core PG device is becoming increasingly common where a PG device is used to shut down/turn on an entire core. Examples of recent multi-core processors using per-core PG are Intel's Core^{TM} $i7$ [7] and AMD's Llano [8].

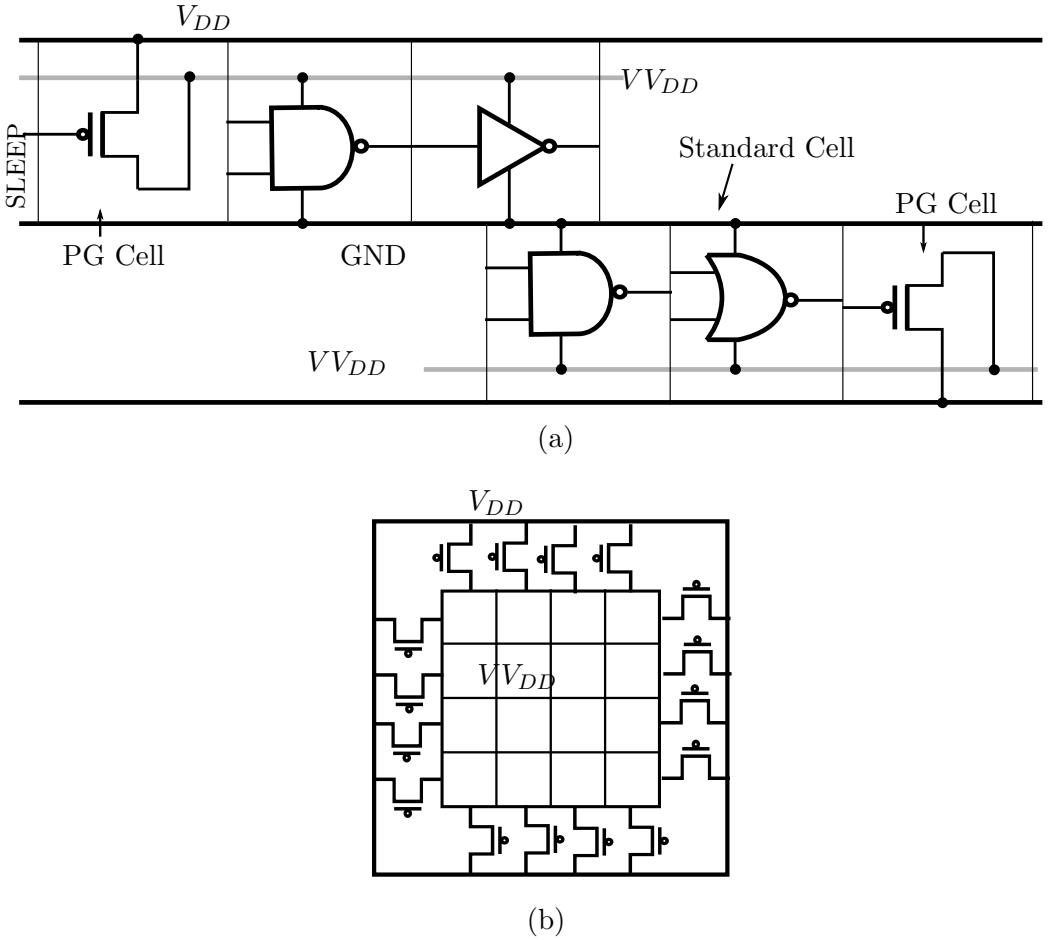


Figure 2.2: (a) Power rails and distributed PG switches in a standard cell environment (b) Ring style PG network.

2.1.2 Power Gating Challenges

Although the work proposed in this thesis does not attempt to solve the challenges associated with implementing power gating, we provide a brief overview in this section. Two most significant design challenges of power gating are sizing and in-rush current and associated supply/ground bounce during core wakeup. The size of a PG device impacts the circuit delay (i.e., F_{MAX}) in the active mode, the leakage current in the sleep mode, and area. To minimize the delay impact due to the drop across the PG, a large PG device

is preferable, but it increases both off-state leakage and area. Most sizing works focus on accurately estimating the current through the PG device. Once the current is known, the minimum PG size can be determined for a given voltage drop, e.g. 2% of V_{DD}. One sizing strategy is to estimate the average current drawn by the power gated circuit. The PG voltage drop is independent of the exact switching pattern of the gates in the circuit and different designs which have same average current tend to have the same voltage drop. Hence, the average current can be used to find the minimum required size of the PG device for a given performance penalty [9][10]. Other studies have demonstrated algorithms for accurate estimation of the maximum current through the sleep transistor [11][12].

The processor core connected to a PG device represents a large capacitance which is in discharged state when the PG device is off. When the PG is turned on, the large capacitance causes a large inrush current to flow which depends on the resistance of the PG device. This large current spike causes resistive and Ldi/dt voltage drops on the power distribution grid which result in ground bounce (for footer) or V_{DD} spike (for header). This ground bounce noise from one PG domain is transmitted to the adjacent domains which are already active through the shared power and ground networks. The current spike and ground bounce noise can be controlled by gradually turning on the PG device over multiple clock cycles. The wakeup phenomenon has been extensively studied in literature and various wakeup strategies have been published. Kim *et al.* [13] showed a method of turning on the PG transistors in which the gate-to-source voltage (V_{GS}) of the PG switch or some fingers of the switch are activated in a step wise non-linear fashion. The authors claimed to reduce both the magnitude of the voltage spike and the wakeup time by this method.

2.2 Power Delivery

The power delivery network of a modern processor consists of multiple voltage domains for cores, shared memory, IO, and other blocks constituting the platform. Power is supplied

to the individual voltage domains by voltage regulators (VRs) which are DC-DC step-down converters supplying an accurate and regulated voltage to the power rail. A typical high performance server platform design thus consists of several VRs of various power ratings (i.e., the total maximum power that the VR can supply). In most commercial processors, the VRs are located on the circuit board and consist of discrete power MOSFETs, inductors, capacitors, and VR controller IC. Power is carried through copper traces to the processor die.The on-chip power network consists of power and ground tracks routed in several metal layers on the die and connected with inter-layer vias. The current drawn by the processor leads to voltage drops across the resistance (IR) and inductance (Ldi/dt) of the power delivery network. A significant amount of on-chip and off-chip decoupling capacitor is added to the power network to reduce the voltage transient noise. The VR design and decoupling requirements depend on the platform specifications. The voltage scaling time of a VR increases proportionally to the decoupling capacitance added on the power delivery path. Depending on platform design, the voltage scaling time of an off-chip VR can be of the order of tens of microseconds [14].

Providing a separate voltage domain for each core or group of cores can lead to fine-grained (per-domain) voltage scaling of the cores resulting in greater power savings than single voltage domain [15]. However, as the number of cores increases, per-core or multiple voltage domains can increase platform cost significantly. A VR is typically designed to provide peak power to the cores in a voltage domain. In case of multiple voltage domains, each VR needs to be designed for higher power drawn by a single domain/core while the other domains are shut down [16][17]. Further, multiple voltage domains are usually associated with multiple clock domains and separate power distribution networks which increases verification complexity and time.

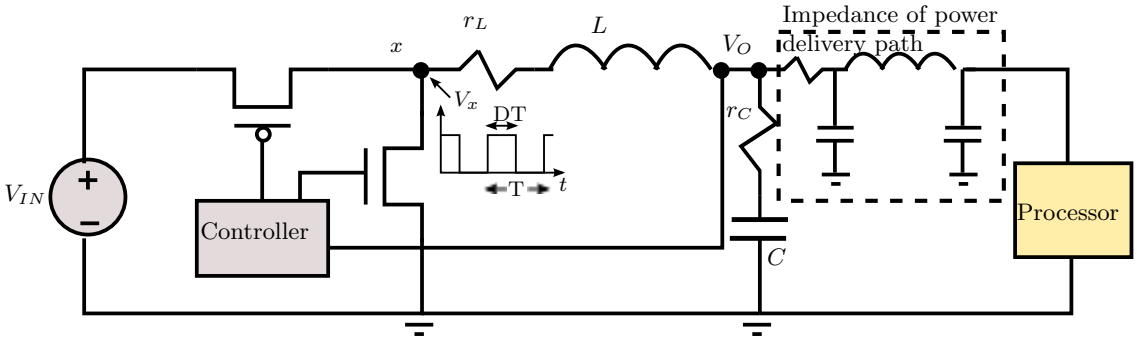


Figure 2.3: Circuit schematic of buck voltage regulator. The output voltage V_O can be controlled by adjusting the duty ratio D of the PMOS switch ($V_O = D \cdot V_{IN}$).

2.2.1 Voltage Regulators

VRs are required to provide a stable voltage to a processor and dynamically scale the voltage to implement power management algorithms. VRs used to power processors can be broadly divided into two categories: switching and linear VRs. Switching VRs are designed using power MOSFETs and typically involve power transfer using the magnetic field of an inductor. Linear VRs, on the other hand, use a MOSFET or bipolar device in linear mode of operation to reduce the voltage. Switching VRs have been preferred for powering processors due to their higher power conversion efficiencies over a wide output voltage range compared to linear VRs. Figure 2.3 shows the circuit of a VR topology commonly use for processors, also known as buck converter. The PMOS and NMOS transistors are switched alternately to create a pulse waveform at the node x from the input voltage V_{IN}. The LC filter is used to filter the switching frequency component and produce a DC voltage V_O at the output. A feedback control loop is used to sense and regulate the output voltage. The average value of V_O is given by

$$V_O = \frac{t_{on}}{T} \times V_{IN} \tag{2.1}$$

where t_{on} is the on-time of the PMOS and T is the switching period. Multiple phases of the converter can be connected in parallel to supply a high current to a processor die. The

power efficiency of a switching VR is a function of its components' technology and sizes. Off-chip switching VRs built using discrete components can attain conversion efficiencies of 90% [18].

A linear VR, also known as low dropout (LDO) regulator consists of a series pass transistor operating in the linear region between the input and output voltage. Figure 2.4 shows the circuit for an LDO. The output voltage is regulated by controlling the gate drive of the pass device with a feedback loop. The efficiency of an LDO is a function of the difference $V_{IN} - V_O$, also called the dropout voltage of the regulator. For small dropout voltages, the efficiency is high and decreases linearly as dropout voltage increases.

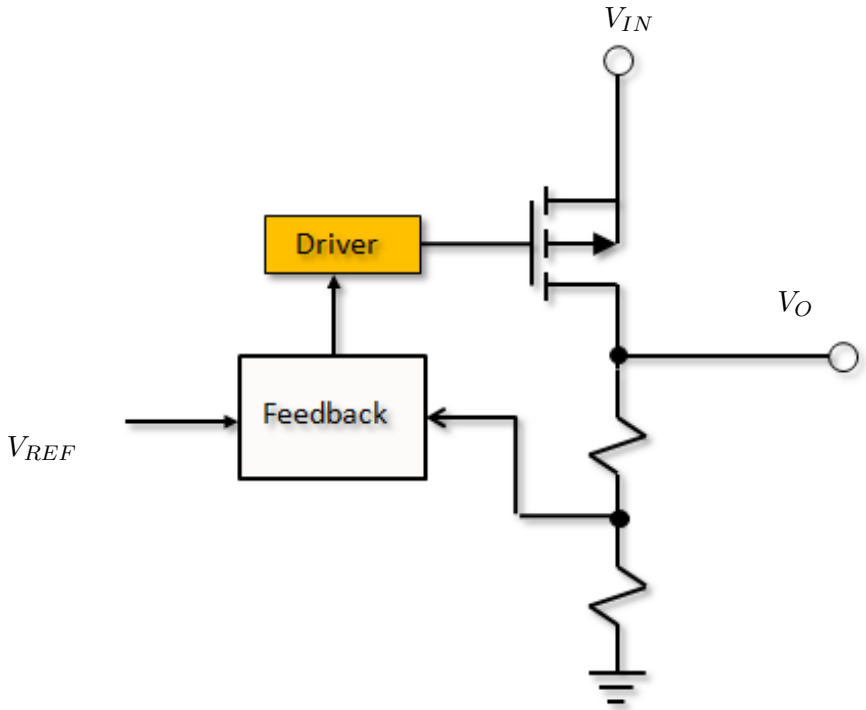
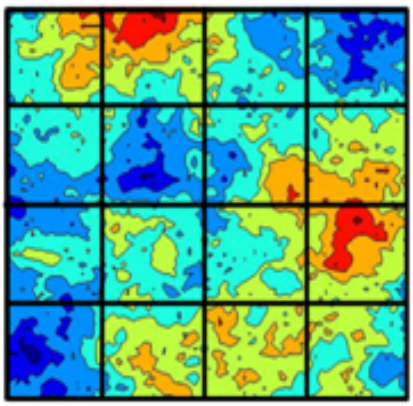


Figure 2.4: Circuit schematic of linear voltage regulator.

2.3 Process Variations

Non-idealities in the semiconductor manufacturing process introduce variations in the key electrical parameters of devices and interconnect. Die-to-die (D2D) variations include variations that arise between different dies in the same wafer or across different wafers, while intra-die, or within-die (WID) variations account for variations that arise within the same die or more generally within a reticle field. WID variations are caused, for example, by lens aberrations and lithographic hot spots, while D2D variations are caused by wafer-level physical phenomena as chemical-mechanical polishing and photoresist coating mechanisms [19]. From a circuit design point of view, key parameters, which are impacted by process variations, are V_{TH} and L_{EFF} of transistors. WID variations have random and systematic components. The random component of parameter variation changes randomly from device



(a)

1.07, 1.75	1.00, 1.21	1.19, 2.89	1.43, 6.43
1.36, 4.10	1.37, 5.84	1.12, 2.38	1.12, 1.25
1.34, 4.54	1.24, 2.20	1.17, 2.02	1.02, 1.00
1.34, 8.78	1.17, 1.57	1.13, 1.22	1.20, 2.06

(b)

Figure 2.5: (a) Systematic WID V_{TH} variation map for a 16-core processor (b) Corresponding F_{MAX} and P_{LEAK}.

to device, i.e., without any correlation between devices. The systematic component follows a governing law, where the correlation between devices is based on the distance between those devices. This spatial correlation is locally layout-dependent and circuit-specific, and globally location dependent. As a result of these variations, significant variation in F_{MAX}

and P_{LEAK} is observed among dies and among cores within a die. As more cores are integrated in a die with technology scaling, spatially correlated WID variations lead to more considerable core-to-core (C2C) F_{MAX} and P_{LEAK} variation. Dighe *et al.* reported 28% F_{MAX} variation and 1.75× P_{LEAK} variation between slowest and fastest core in a 80 core processor fabricated in 65nm technology [20]. In this work, we use the model of within die variations developed by Sarangi *et al.* [21]. According to this model, the systematic WID variation in L_{EFF} is related to the V_{TH} variation by the following equation:

$$L_{EFF} = L_{EFF,0} \times \left(1 + \frac{V_{TH} - V_{TH,0}}{2 \times V_{TH,0}}\right) \tag{2.2}$$

D2D variation is modeled as normally-distributed with zero mean and $\sigma_{V_{TH}}$= 5.0%. Figure 2.5-(a) shows a systematic V_{TH} variation map for a die containing 16 cores. Each rectangle represents a core, and each pair of numbers in a rectangle in Figure 2.5-(b) corresponds to F_{MAX} and P_{LEAK} of each core at V_{DD}, normalized to the F_{MAX} of the slowest core and the P_{LEAK} of the least leaky core, respectively. Note that core 4 is 1.43× faster than the slowest core while core 13 is to 8.7× leakier than the least leaky core in the die; faster cores often consume more P_{LEAK}. As more cores are integrated in a die (i.e., die size becomes larger), relative C2C F_{MAX} and P_{LEAK} variations across a die increase [22].

Various post-manufacturing tuning techniques are used by chip manufacturers to mitigate the impact of process variations on yield. A basic method is to allocate a die to one of several frequency bins depending on its F_{MAX} and marketing it at a price based on the bin frequency. This process is known as frequency binning. Adaptive voltage scaling (AVS) adjusts the voltage of a die based on its process corner to maximize its frequency under a power constraint. Dies which are fast but leaky are allocated lower than nominal voltage to reduce their power consumption while slow dies are allocated higher voltage to increase their F_{MAX}. AVS is more beneficial than simple frequency binning because it reduces both standby leakage and dynamic power, which allows dies to be accepted in a higher

frequency bin than with frequency binning alone. However, employing AVS using multiple voltage domains to mitigate WID variations in a multi-core processor can be challenging because it increases design and verification time and requires multiple voltage regulators and decoupling capacitors. Adaptive body biasing (ABB) is another technique to tighten the delay and leakage spread of dies. By increasing (decreasing) the V_{TH} of fast (slow) dies, it possible to decrease the F_{MAX} spread of a distribution of dies. As technology scales, the effectiveness of ABB reduces due to increased short channel effects (V_{TH} roll off and drain induced barrier lowering) both of which reduce sensitivity of V_{TH} to body potential [23]. Additionally, implementing ABB at the WID level requires extra routing resources.

2.4 Circuit Reliability

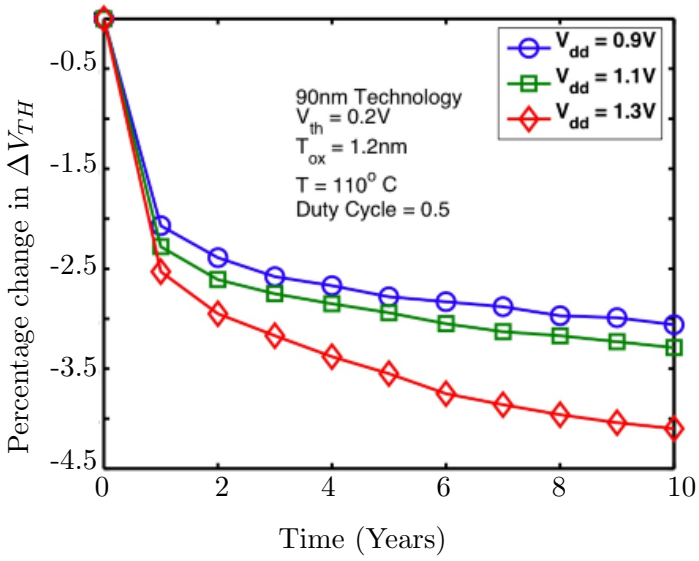


Figure 2.6: Percentage degradation in V_{TH} of PMOS due to NBTI [24].

The reliability of CMOS circuits decreases with usage over time due to device aging effects. Aging is caused by factors such as applied voltage, operating temperature, and growth of defects inherent in silicon. One the most important reliability concern for future

CMOS technology is bias temperature instability (BTI), which increases the threshold voltage (V_{TH}) of transistors. This in turn reduces the current drive of the transistors and circuit speed. In conventional CMOS devices, the junction of Si and SiO_2 is made of dangling bonds or traps. Traditionally, these traps are passivated by introduction of hydrogen which forms Si-H bonds at the interface. However, recent scaling trends such as slow reduction in supply voltages compared to aggressive gate-oxide scaling and process modifications such as the introduction of nitrided oxides to prevent boron penetration from the poly-gate increase the dissociation of the Si-H bonds over time during device operation. The increased trap density at the Si/SiO_2 interface leads to reduction in channel charge carriers, thus effectively increasing the V_{TH} of the device. Figure 2.6 shows the V_{TH} degradation with time for various stress duty ratios calculated using the NBTI aging model in [24]. As technology scales, BTI is becoming an important concern for chip manufacturers due to aggressive oxide thickness scaling, reduced rate of supply voltage scaling and increasing average chip temperatures.

Chapter 3

Frequency and Yield Optimization Using Programmable Power-gating

As core size continues to shrink with each technology generation, process variations lead to substantial die-to-die (D2D) and core-to-core (C2C) variations of maximum operating frequency (F_{MAX}), active leakage power (P_{LEAK}), and thus total power (P_{TOT}) in multi-core processors. This in turn may reduce frequency and/or yield of power-constrained designs. Yield is impacted because many dies which are fast enough to satisfy the F_{MAX} constraint, can be discarded due to excessive active P_{LEAK}. This becomes worse as the spread of P_{LEAK} and its proportion in total power consumption increases with technology scaling. For dies that meet the P_{LEAK} constraint, F_{MAX} is limited by the slowest core when chip-wide voltage/frequency domain is used while fast cores consume considerably more P_{LEAK}. Under a power constraint, the fast-but-leaky cores limit the F_{MAX} of the die.

To minimize standby power, a power-gating (PG) device, placed between a core and its power supply rails, is commonly used in multi-core processors [25]. In this chapter, we propose two optimization methods to improve the yield and F_{MAX} of dies using PG devices

that already exist in many power-constrained designs. The first method attempts to increase the number of dies that can operate at a given target frequency (i.e., yield) by reducing the strength of the PG device (hence P_{LEAK}) of leaky dies during a post-manufacturing step. The second optimization method assigns the most performance/power efficient voltage/frequency point for each core by tuning its PG device during a post-manufacturing step. As a result of the optimization, the power consumption of fast and slow cores can be balanced, leading to higher maximum throughput (i.e., F_{MAX}) under a power constraint.

3.1 Programmable-width Power Gating

Figure 3.1 illustrates a concept of a programmable-width power gating (PPG) device [26]. PMOS header switches are connected to the SLEEP signal through NAND gates; the other input of each NAND gate is connected to a configuration bit. Hence, the header switches with a configuration bit set to "0" are always turned off while those with a configuration bit set to "1" can be turned on (when SLEEP is "0") and off (when SLEEP is "1"). As a result, a less number of switches will be turned on during active mode (when SLEEP is "1") when we have less configuration bits set to "1". This increases the series resistance and the voltage drop across the switches (decreases VV_{DD} linearly), which reduces the leakage power (exponentially) and the speed (close-to-linearly with a small amount of VV_{DD} drop) of the circuit. Therefore, we can control P_{LEAK} and thus F_{MAX} of a particular die depending on how we program the configuration bits after manufacturing.

Configuration bits can be programmed by one-time-programmable (OTP) e-fuses after manufacturing characterization of F_{MAX} and P_{LEAK}. Note that characterizing and programming can be done like any other variation compensation technique (e.g., ABB or AVS). To reduce the number of configuration bits while providing various sizes of header switches that designers can choose from, the header switches are arranged in exponentially increasing widths as shown in Figure 3.1. Each switch, in turn, may consist of several smaller switches

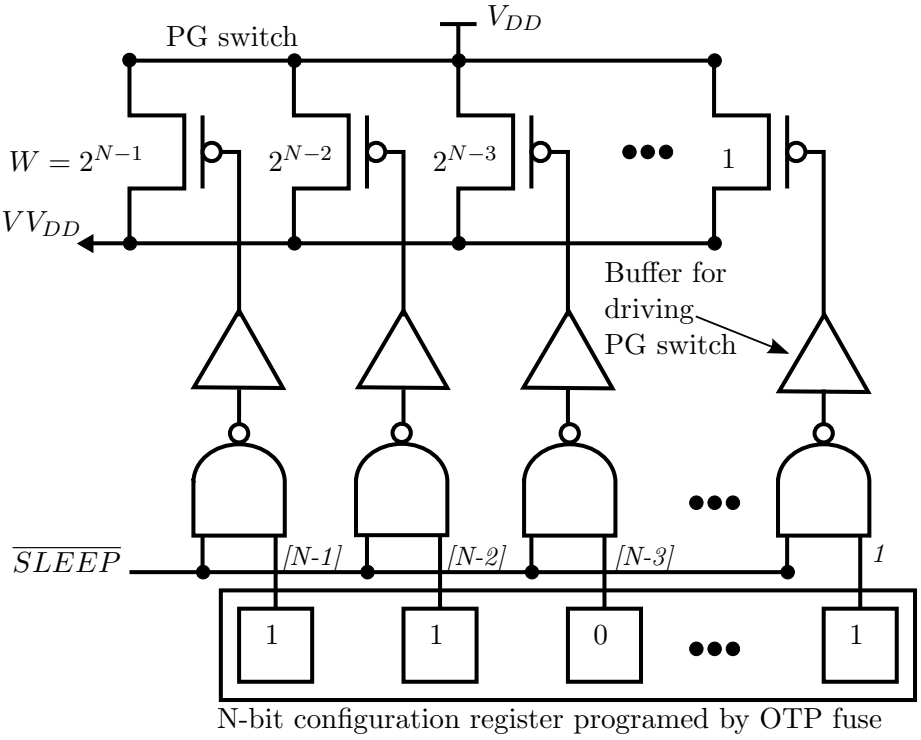


Figure 3.1: Implementation of programmable power gating device.

connected in parallel. This allows the sum of header widths proportional to the binary value of configuration bits. For example, only 8 configuration bits are enough to provide 4mV granularity for adjusting VV_{DD}. The header switches and buffers in Figure 3.1 represent the required resource to implement a conventional PG. Therefore, the additional area by the NAND gates, which is an overhead of PPG, is small since the large PG switches are driven by existing buffers.

3.2 Impact on Delay, Leakage, and VV_{DD}

To analyze the impact of tuning PG strength on F_{MAX}, P_{LEAK}, and VV_{DD} in a 32nm technology node, a PPG device is initially sized to provide 25 mV drop from V_{DD}, which is assumed to be 0.9 V at the nominal process corner and 100 C. Under such a condition, we

assume that the ratio between dynamic power and active leakage power of an IC connected to the PPG device is 7:3 in P_{TOT} [3]. We model I_{DYN} and I_{LEAK} with a dummy circuit as illustrated in Section 3.4.3. P_{TOT} is directly measured at the V_{DD} node to include the

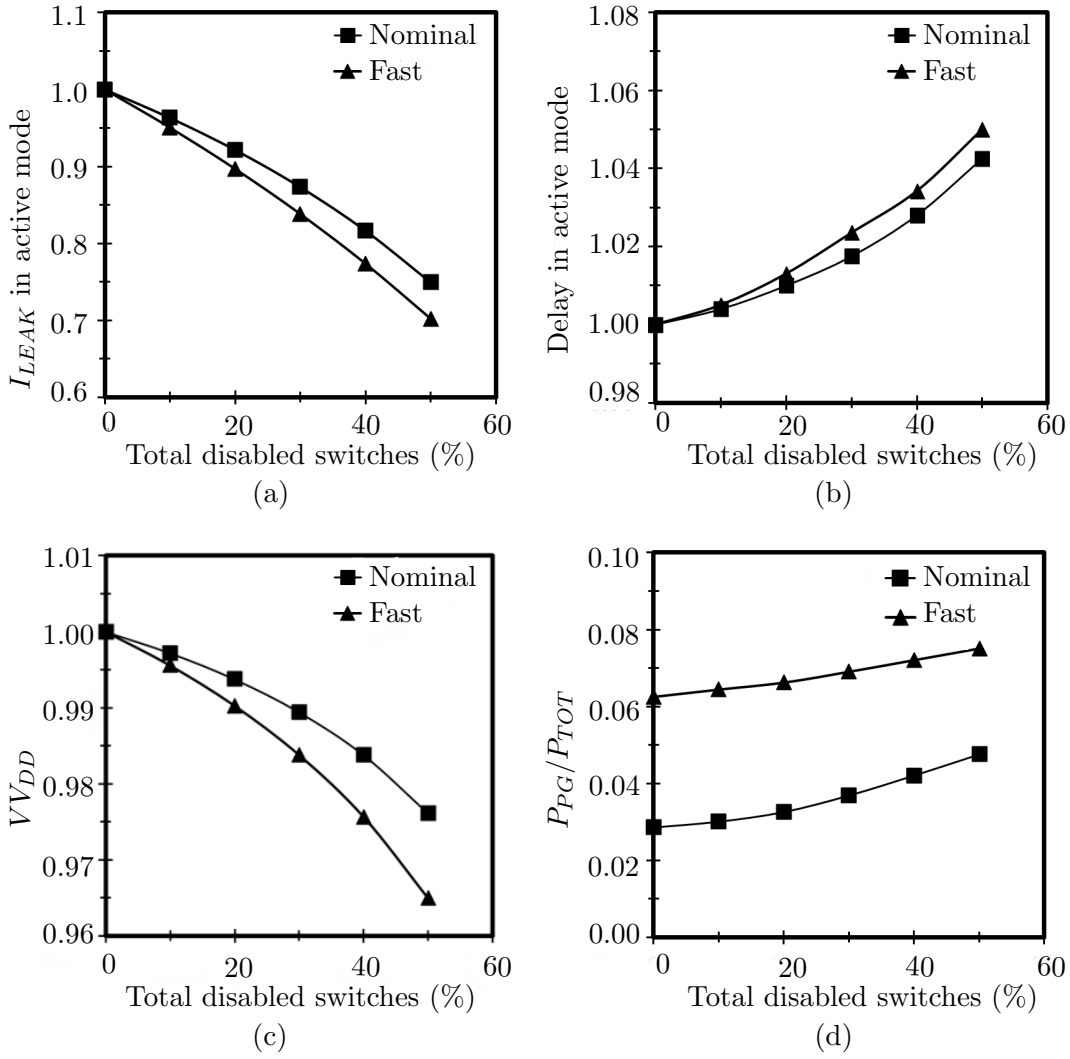


Figure 3.2: Normalized (a) leakage current, (b) delay, (c) VV_{DD}, and (d) P_{PG} /P_{TOT} versus the fraction of off PPG switches.

power consumption of the PG device itself. Figure 3.2-(a) and (b) show active I_{LEAK} and delay while the total width of disabled PG switches is varied; the figures are normalized

to those of an IC in which all switches are enabled. As the fraction of disabled switches increases to 10%, 20%, and 30% at the nominal process corner, active I_{LEAK} decreases by 3.7%, 7.9%, and 12.7%, while the delay increases by 0.5%, 1.0%, and 1.8%, respectively. It is the fast corner in which the proposed technique will be mainly applied to dies, because the dies are often unnecessarily fast thereby consuming too much P_{LEAK}.

Figure 3.2-(c) and (d) show the VV_{DD} normalized to that of an IC in which all switches are enabled, as well as the proportion of power consumption of the PPG device (PPG) in P_{TOT}. As the fraction of disabled switches increases to 10%, 20%, and 30%, the resistance across the PPG switches increases, thereby reducing VV_{DD} by 10%, 22%, and 37% (relative to the initial 25 mV drop), respectively. This also reduces P_{DYN} of the connected IC. As the resistance increases, on the other hand, more power is consumed by the PPG switches. However, the portion of PPG in P_{TOT} is very small as shown in Figure 3.2-(d); it is only 8% at the fast corner although 50% of the PPG switches are disabled.

3.3 Die-to-die Variation-aware Yield Improvement with PPG

Dies in the fast corner exhibit significantly higher P_{LEAK} than those in nominal corner due to shorter L_{EFF} and lower V_{TH} for their transistors. As a result, they violate the $P_{TOT,MAX}$ constraint and have to be discarded, thereby reducing the yield. This section presents a method to improve the yield of power-constrained designs using PPG devices for two scenarios: designs with: 1) fixed and 2) variable F_{MAX} and P_{LEAK} constraints, respectively. Discarded fast-but-leaky dies can be recovered by adjusting (weakening) the strength of power gates to reduce the P_{LEAK} until each die can satisfy the $P_{TOT,MAX}$ constraint (but within the F_{MAX} constraint).

3.3.1 Simulation Methodology for Yield Experiments

The yield improvement technique was verified with a simulation setup consisting of ISCAS85 and OpenCore circuits. Monte-Carlo simulations were performed using SPICE with a predictive 45-nm technology model [27][28] to apply D2D and WID process variations to each die sample of the selected circuits; we apply 3σ variations (i.e., 0.4 V and 10 nm) to the nominal V_{TH} and L_{EFF} of NMOS and PMOS, respectively. Initially, PPG switches at the nominal corner are sized such that the maximum voltage drop across the switches does not exceed 50mV for the peak current consumption ($I_{DD,MAX}$) of each circuit; The header switches in Figure 3.1 are all connected (i.e., all the configuration bits are initially set to "1"). $I_{DD,MAX}$ is estimated by applying 10000 vectors at 100 °C and selecting a pair of vectors causing the worst-case current consumption.

3.3.2 Designs with Frequency Target: Fixed P_{LEAK} Constraint

In typical ASIC designs, F_{MAX} and P_{TOT} of each die have to satisfy frequency target (F) and power target (P), respectively, i.e.

$$F_{MAX}(VV_{DD}) \geq F \tag{3.1}$$

$$P_{TOT} = P_{DYN}(VV_{DD}) + P_{LEAK}(VV_{DD}) \leq P \tag{3.2}$$

In this section, F_{MAX} is not an operating frequency but a maximum frequency that can be achieved by a particular die; unless F exceeds F_{MAX}, dies will operate at F. Since all the components (F_{MAX}, P_{DYN}, and P_{LEAK}) in Eqs. (3.1) and (3.2) are functions of VV_{DD}, some fast-but-leaky dies that satisfy Eq. (3.1) may be forced back to satisfy both Eq. (3.1) and Eq. (3.2) after adjusting VV_{DD} (or the strength of a programmable width PG). In Eq. (3.2), P_{DYN} can be considered to be constant if: 1) operating frequency is fixed at F and 2) after adjusting the PPG, VV_{DD} takes a value that is not significantly different from

the initial VV_{DD}. This allows us to consider the power constraint in Eq. (3.2) simply as a leakage target:

$$P_{LEAK}(VV_{DD}) < P' \tag{3.3}$$

where P' is the difference between P and P_{DYN} (i.e., P_{LEAK} budget). In fact, Eq. (3.3) is conservative since P_{DYN} becomes smaller after adjusting VV_{DD}. In other words, Eq. (3.2) is always satisfied if Eq. (3.3) is satisfied. For each die sample of a benchmark circuit, P_{LEAK} and F_{MAX} are measured using SPICE as the configuration bits are varied until the die satisfies Eq. (3.3). P in Eq. (3.3) is arbitrarily set to 3× of the nominal P_{LEAK} i.e. $P_{LEAK,NOM}$ (leakage power of a die without process variations); F is set to -3σ of the nominal F_{MAX}, $F_{MAX,NOM}$ where σ is standard deviation of F_{MAX} over 1,000 dies. The P_{LEAK} and the F_{MAX} of 1000 dies of ISCAS85 benchmark C432 obtained through

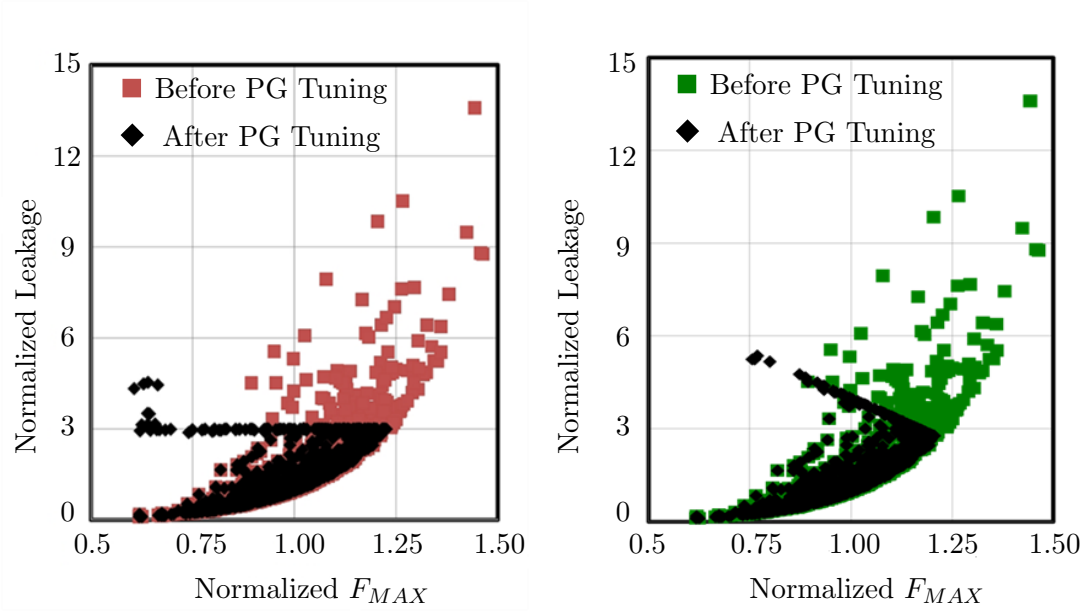


Figure 3.3: Normalized P_{LEAK} and F_{MAX} distribution before and after applying the optimization method to maximize the yield under (a) fixed and (b) variable P_{LEAK} constraints for ISCAS85 C432.

SPICE simulation are shown as a scatter plot (square boxes marked "Before PG Tuning") in Figure 3.3-(a). 116 dies are rejected in the example circuit, due to the leakage constraint, i.e. about 88% of yield. For each of the rejected dies due to excessive P_{LEAK}, we tried to change its configuration bits (by setting some of them to 0) so that it was moved under the $3\times P_{LEAK}$ boundary. The results are shown as another scatter plot (diamond boxes marked "After PG Tuning") in Figure 3.3-(a). Only 12 dies are rejected after the tuning method is applied, and the yield is improved by recovering 104 dies (99%) out of 116 rejected ones. Table 3.1 summarizes the yield loss due to violating P_{LEAK} constraints (exceeding $3\times$ and $4\times$ of the nominal P_{LEAK}) and the recovery, before and after the optimization is applied, respectively. For the dies violating the P_{LEAK} constraints, the strength of PG switches is reduced until the die satisfies the P_{LEAK} constraint; meanwhile it must not violate the target frequency constraint F which is -3σ frequency of the nominal F_{MAX} among 1000 samples per each circuit. We assume that the optimization process failed and recovery was unsuccessful, if the F_{MAX} of a die becomes slower than the target frequency constraint. On average, we recovered 90% and 92% of discarded fast-but-leaky dies when the P_{LEAK} constraints are $3\times P_{LEAK,NOM}$ and $4\times P_{LEAK,NOM}$, respectively. Relaxing the P_{LEAK} constraint gives fewer violations before applying the optimization, but it also provides more opportunity to recover the discarded dies from the violations, resulting in a similar or higher percentage of yield improvement within a certain range of P_{LEAK} constraints.

3.3.3 Designs with Frequency Binning: Variable P_{LEAK} Constraint

In high-performance processor designs, a list of frequency targets, $F_1 < F_2 < \ldots < F_N$, are provided. The F_{MAX} of a die is compared to the targets and then it is put into an

Table 3.1: Yield loss recovery for fixed P_{LEAK} constraint

circuit	$\boldsymbol{P_{LEAK}}$ **constraint**	**no. of violations**		**yield loss recovery**
		before PG optimization	**after PG optimization**	
C432	$3\times P_{LEAK,NOM}$	116	12	90%
	$4\times P_{LEAK,NOM}$	56	6	89%
C499	$3\times P_{LEAK,NOM}$	118	13	89%
	$4\times P_{LEAK,NOM}$	60	4	93%
C880	$3\times P_{LEAK,NOM}$	101	12	88%
	$4\times P_{LEAK,NOM}$	45	4	91%
C1355	$3\times P_{LEAK,NOM}$	106	6	94%
	$4\times P_{LEAK,NOM}$	43	2	95%
C1908	$3\times P_{LEAK,NOM}$	121	12	90%
	$4\times P_{LEAK,NOM}$	67	7	90%
C2670	$3\times P_{LEAK,NOM}$	119	9	92%
	$4\times P_{LEAK,NOM}$	60	2	97%
C3540	$3\times P_{LEAK,NOM}$	118	18	85%
	$4\times P_{LEAK,NOM}$	65	5	92%
average	$3\times P_{LEAK,NOM}$	114	12	90%
	$4\times P_{LEAK,NOM}$	57	4	92%

appropriate bin, i.e,

$$f(F_{MAX}) = \begin{cases} F_i, & F_i \leq F_{MAX} < F_{i+1}, i = 1, 2, ..., N \\ F_N & \text{otherwise} \end{cases} \tag{3.4}$$

where f is a function that assigns an operating frequency; this process is called frequency binning. As a result, dies have different leakage power constraints depending on the bins where they are placed, since P_{DYN}, which is proportional to operating frequency, is different for different bins. Meanwhile, the sum of P_{DYN} and P_{LEAK} has to be no greater than a fixed power constraint. PPG tuning can be applied in this case with the consideration of a variable P_{LEAK} constraint. Since the bins of higher frequencies are preferred, the optimization objective is to maximize the operating frequency of each die, i.e.

$$maximize f(F_{MAX}(VV_{DD})) \tag{3.5}$$

subject to the constraint:

$$P_{TOT} = P_{DYN}(f(F_{MAX}), VV_{DD}) + P_{LEAK}(VV_{DD}) \leq P \tag{3.6}$$

Figure 3.3-(b) shows the results of applying the PG tuning optimization to ISCAS C432 samples with variable P_{LEAK} constraint. In this figure, P is equal to $P_{DYN,NOM}$ + $4\times P_{LEAK,NOM}$, where $P_{DYN,NOM}$ takes about 60% of P. Before the proposed method is applied to maximize the yield, the dies above the diagonal line are discarded because their P_{TOT} exceeds P. As explained earlier, the dies exhibiting higher F_{MAX} have less P_{LEAK} budget. Initially 106 dies are rejected from C432. However, after the method is applied (diamond boxes marked "After PG Tuning"), only 1 die is rejected, recovering almost all the dies in this particular example. Table 3.2 summarizes the yield loss due to

Table 3.2: Yield loss recovery for variable P_{LEAK} constraint

circuit	P_{LEAK} **at** $F_{MAX,NOM}$	**no. of violations**		**yield loss recovery**
		before PG optimization	**after PG optimization**	
C432	$0.3\times P$	204	2	99%
	$0.4\times P$	106	1	99%
C499	$0.3\times P$	208	4	98%
	$0.4\times P$	110	2	98%
C880	$0.3\times P$	190	0	100%
	$0.4\times P$	104	0	100%
C1355	$0.3\times P$	193	3	98%
	$0.4\times P$	94	1	99%
C1908	$0.3\times P$	214	2	99%
	$0.4\times P$	120	0	100%
C2670	$0.3\times P$	202	0	100%
	$0.4\times P$	108	0	100%
C3540	$0.3\times P$	203	5	98%
	$0.4\times P$	117	3	97%
average	$0.3\times P$	202	2	98%
	$0.4\times P$	108	1	99%

violating P constraint and the recovery before and after the optimization is applied, respectively. For the dies violating the target power constraint, we adjust the strength of power-gating switches until we satisfy the target power constraint, while maximizing F_{MAX} of each die. As F_{MAX} of a die becomes slower by adjusting the strength of power gating switches, more P_{LEAK} will be allowed since F_{MAX} decrease reduces P_{DYN} of the die, which allows more power budget for P_{LEAK}. To set the target power constraint P, we assume that $P_{TOT,MAX}$ is $P_{DYN,NOM} + 4 \times P_{LEAK,NOM}$ at the nominal F_{MAX}, and that the ratios of $P_{DYN,NOM}$ to $4 \times P_{LEAK,NOM}$ at the nominal F_{MAX} are 1) 0.6 and 0.4 and 2) 0.7 and 0.3 to see the sensitivity on the percentage of P_{LEAK}, in P, since the percentage of P_{LEAK} in most recent digital designs like microprocessors has been between 30∼40%. On average, we recovered 98% of the discarded dies, when the P_{LEAK} constraints at the $F_{MAX,NOM}$ point are $0.3 \times P$ and $0.4 \times P$, respectively. Note that less P_{LEAK} budget (e.g., $P_{LEAK} = 0.3 \times P$ at the nominal F_{MAX}) incurs more violations than more P_{LEAK} budget ($P_{LEAK} = 0.4 \times P$ at the nominal F_{MAX}) before applying the optimization due to less P_{LEAK} headroom for P_{LEAK} variations at higher F_{MAX}.

3.3.4 Impact of Limiting VV_{DD} Drop

The tuning of the PG device to limit the P_{LEAK} of fast dies may increase the voltage drop across the power-gating device since some switches are disabled, thereby increasing its resistance in active mode. However, an excessive voltage drop can degrade the noise margin of the designs. Table 3.3 shows the yield recovery by applying the PG tuning to ISCAS85 benchmarks with the PG drop limited to 150mV and 100mV. When the voltage drop limited to 150mV (100mV) for the fixed P_{LEAK} constraints = $3 \times P_{LEAK,NOM}$ and $4 \times P_{LEAK,NOM}$, we recovered 52% (30%) and 53% (32%) of discarded fast-but-leaky dies on average. For the variable P_{LEAK} constraints, i.e., P_{LEAK} at $F_{MAX,NOM} = 0.3 \times P$ and $0.4 \times P$, we recovered 56% (30%) and 61% (33%) of discarded fast-but-leaky dies with the PG voltage drop limited

to 150mV (100mV). The circuits with low activity factors lead to very high P_{LEAK} in P_{TOT} while D2D variations can increase P_{LEAK} by orders of magnitude. Limiting the voltage drop across the PG devices also limits the maximum P_{LEAK} decrease for fast-but-leaky dies, impacting the percentage of recoverable dies. However, for a synthetic circuit whose

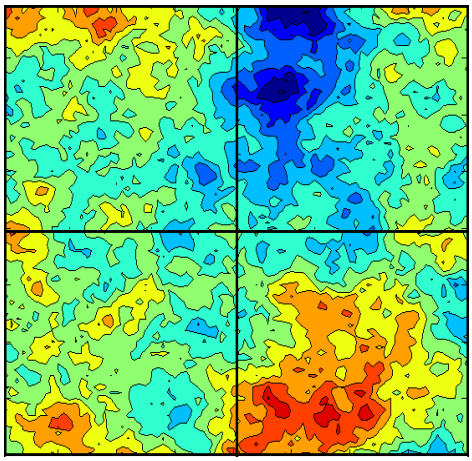
Table 3.3: Yield loss recovery wiyh PG drop limited to 100mV and 150mV

circuit	**fixed P_{LEAK} constraint**	**Yield loss recovery(%)**		**var. P_{LEAK} constraint**	**Yield loss recovery(%)**	
		150mV	**100mV**		**150mV**	**100mV**
C432	$3\times P_{LEAK,NOM}$	53	31	$0.3\times P$	40	31
	$4\times P_{LEAK,NOM}$	54	30	$0.4\times P$	57	32
C499	$3\times P_{LEAK,NOM}$	51	31	$0.3\times P$	69	40
	$4\times P_{LEAK,NOM}$	53	38	$0.4\times P$	69	40
C880	$3\times P_{LEAK,NOM}$	56	26	$0.3\times P$	69	39
	$4\times P_{LEAK,NOM}$	49	27	$0.4\times P$	64	43
C1355	$3\times P_{LEAK,NOM}$	58	36	$0.3\times P$	56	30
	$4\times P_{LEAK,NOM}$	47	23	$0.4\times P$	65	33
C1908	$3\times P_{LEAK,NOM}$	45	27	$0.3\times P$	62	38
	$4\times P_{LEAK,NOM}$	64	43	$0.4\times P$	60	33
C2670	$3\times P_{LEAK,NOM}$	55	32	$0.3\times P$	52	6
	$4\times P_{LEAK,NOM}$	50	28	$0.4\times P$	51	16
C3540	$3\times P_{LEAK,NOM}$	46	24	$0.3\times P$	46	29
	$4\times P_{LEAK,NOM}$	52	35	$0.4\times P$	64	32
average	$3\times P_{LEAK,NOM}$	52	30	$0.3\times P$	56	30
	$4\times P_{LEAK,NOM}$	53	32	$0.4\times P$	61	33

nominal P_{LEAK} fraction in P_{TOT} is similar to commercial processors, $88\% - 93\%$ of dies can be recovered although the voltage drop is limited to 100mV.

3.4 Frequency Improvement with Multiple PPG Domains

A large design often consists of several cores or groups of cores where each has its own PG device to minimize standby leakage power. Depending on the workload demand, some of the cores can be disabled using the associated PG devices. In a multi-core processor employing a global-clocking scheme, the operating frequency, when all or some of the cores are active, is set by the F_{MAX} of the slowest core (unless each core is clocked at its own frequency using frequency islands). Note that many commercial SoCs and multi-core processors use a global-clocking scheme to avoid clock-domain crossing that complicates design, verification, and test. A power constraint is often imposed when all cores are running simultaneously at a maximum sustainable performance point. This limits the increase of V_{DD} and F_{MAX}. In the following discussion V_{DDTDP} and $F_{MAX,TDP}$ denote the V_{DD} and F_{MAX} which the processor can reach under the given power constraint ($P_{TOT,MAX}$ (usually thermal design power)). In a multi-core processor with per core PPG, we can set the PG configuration bits of each core such that the F_{MAX} of each core can be set as even as possible to that of the slowest core. This will lead to a significant amount of each core's P_{LEAK} reduction without impacting the F_{MAX} of the processor, which, in turn, lets the total power consumption (P_{TOT}) of the processor become much smaller than the power constraint. Consequently, we can increase the global V_{DD} (thus F_{MAX}) of the processor until power and other constraints such as maximum junction temperature (T_j), maximum V_{DD} ($V_{DD,MAX}$), etc. are not violated so that we can put the die in a higher frequency bin. This F_{MAX} optimization procedure, as applied to a quad-core processor die sample, is shown below followed by a formulation of the optimization problem and simulation methodology adopted in this work.



(a)

W **1.19** **1.18**	**X** **1.37** **2.79**
Y **1.16** **1.25**	**Z** **1.00** **1.00**

(b)

Figure 3.4: (a) Systematic V_{TH} variation map for a quad-core processor. (b) The initial F_{MAX} and P_{LEAK} of each core, normalized to the F_{MAX} of the slowest core and P_{LEAK} of the least leaky core respectively.

3.4.1 Example of F_{MAX} Improvement in a Quad-core Processor

Figure 3.4-(a) shows a V_{TH} variation map of a quad-core processor where each rectangle represents a core. A pair of numbers within each core shown in Figure 3.4-(b) corresponds to the F_{MAX} and the P_{LEAK} of each core; they are each normalized to the smallest values from all four cores. Since the F_{MAX} of a multi-core processor employing a global-clocking scheme is limited by the slowest core Z, we take cores W, X, and Y and change the PG configuration bits of each of them (i.e. decrease their VV_{DD}) until their F_{MAX}s become as close as possible to the F_{MAX} of core Z. Since the VV_{DD} of faster cores is reduced, the P_{LEAK} of core W, X, and Y become smaller as well, as shown in Figure 3.5-(a) where all the cores now have the same F_{MAX}. After the above F_{MAX} equalization, the sum of P_{LEAK} becomes 4.1 while it was 6.22 in Figure 3.4-(b), which is about 34% P_{LEAK} reduction. As a result, the P_{TOT} is reduced by 10.75% in Figure 3.5-(a) assuming that sum of P_{LEAK} before we program the PG configuration bits of each core is 40% of $P_{TOT,MAX}$ at V_{DDTDP}. Next the global V_{DD} (thus the F_{MAX}) is increased until the P_{TOT} becomes the same as

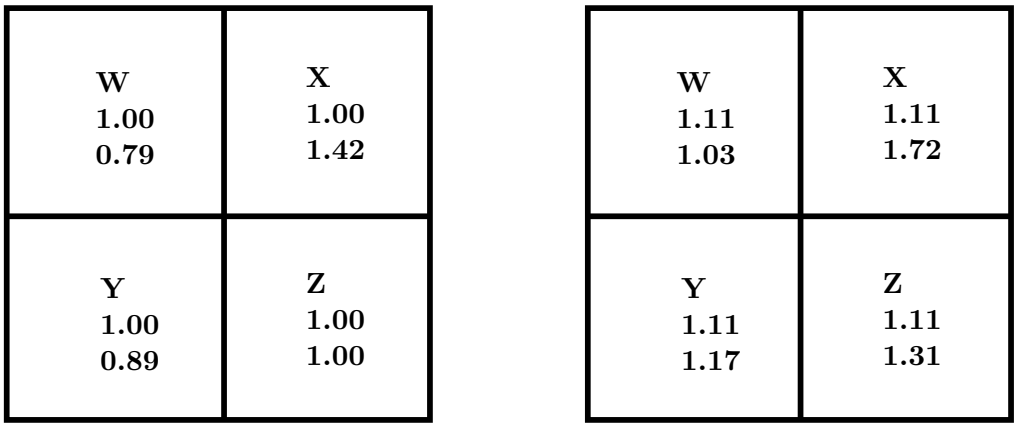


Figure 3.5: Normalized P_{LEAK} and F_{MAX} of cores: (a)after applying core-by-core programmable power-gating and (b) increasing global V_{DD} until power constraint is satisfied again.

before (i.e. the P_{TOT} of Figure 3.4-(b)) as long as $T_{j,MAX}$ and $V_{DD,MAX}$ constraints are not violated; Figure 3.5-(b) shows the result, where we can see F_{MAX} increase of 11%.

3.4.2 F_{MAX} Optimization Problem Formulation

Let VV_{DD} of PPG domain i, $VV_{DD,i}$ be represented by $VV_{DD,i} = v_i(W_i).VV_{DD}$, where W_i is the total effective width of the PPG device in domain i; v_i is a function that returns the VV_{DD} scaling factor of domain i for given W_i. VV_{DD} is a strong function of W_i and its value is always lower than $VV_{DD,i}$ in PPG. However, if the global V_{DD} is increased, VV_{DD} can become higher than its initial V_{DD}. Modulating the strength of a PPG device affects the VV_{DD} of the corresponding domain alone while scaling the global V_{DD} affects the VV_{DD} of all the domains. The objective of the problem is to maximize the F_{MAX} of a given die while $P_{TOT,MAX}$ and $V_{DD,MAX}$ are satisfied. Objective:

$$maximize(F_{MAX}(V_{DD}, VV_{DD,1}, VV_{DD,2}, ..., VV_{DD,N})) \tag{3.7}$$

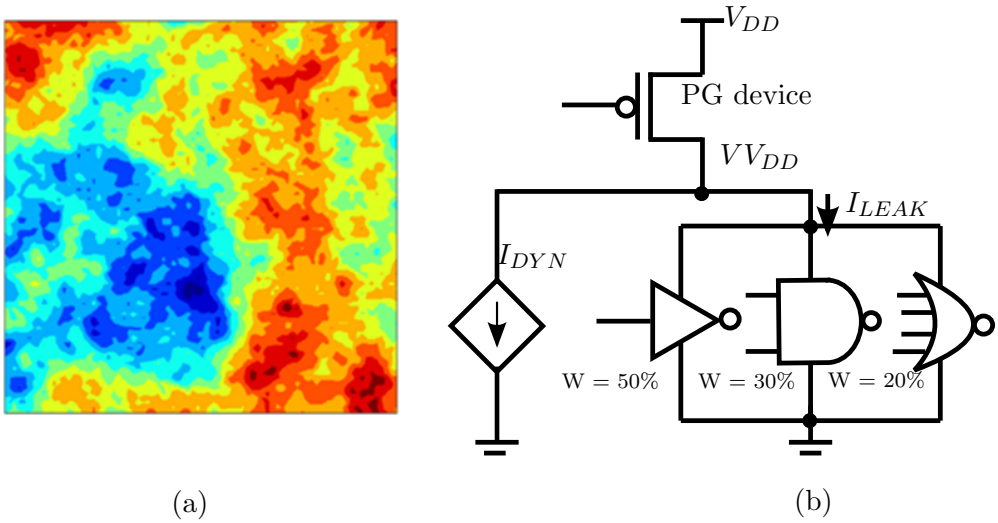
Constraints:

$$P_{TOT} = \sum_{i=1}^{N} P_{TOT,i}(VV_{DD,i}, F_{MAX}) \leq P_{TOT,MAX}, V_{DD} < V_{DD,MAX} \tag{3.8}$$

where $F_{MAX,i}$ and $VV_{DD,i}$ are F_{MAX} and VV_{DD} of the circuitry in PPG domain i, respectively; F_{MAX} is $minF_{MAX,1}(VV_{DD,1}), ..., F_{MAX,N}(VV_{DD,N})$; N is the number of PPG domains; and $P_{TOT,i}$ corresponds to the total power consumption that includes dynamic and static components in PPG domain i.

3.4.3 Simulation Methodology for F_{MAX} Optimization

A multi-core processor die with WID variations is generated with an 80 $\times$ 80 grid where each grid point is assigned a distinct V_{TH} and L_{EFF} combination. The die area is assumed to be 35mm^2. WID correlation distance coefficient ϕ (0.5), WID V_{TH} variation σ^{sys} (6.4%), and D2D variation σ^{D2D} (5.0%) were used to model WID and D2D V_{TH} and L_{EFF} variation. Figure 3.6-(a) shows systematic, correlated L_{EFF} and V_{TH} variations across a die in a 32nm technology [27][28]. An F_{MAX} and P_{LEAK} combination is obtained for each grid point with the associated V_{TH} and L_{EFF}, by simulating a 16-stage FO4 chain for F_{MAX} and a dummy circuit as illustrated in Figure 3.6-(b) consisting a large number of INV (50%), NAND (30%), and NOR (20%) gates for P_{LEAK}. V_{TH} and L_{EFF} value pairs corresponding to the grid points are applied to a 32nm technology model [27][28] to obtain F_{MAX} and P_{LEAK} (and F_{MAX} and P_{LEAK} scaling factors relative to the F_{MAX} and P_{LEAK} at $V_{DD,TDP}$) as functions of V_{DD} using SPICE and a curve fitting tool; each gate excluding INVs has a various number of inputs (2 $\sim$ 4) and randomly selected input states are applied to measure P_{LEAK}. The F_{MAX} scaling factor of a core (i.e. PG domain) is decided by the slowest grid point in the core, while P_{LEAK} is obtained from the summation of the leakage current from all the grid points in the core. In the F_{MAX} simulations P_{TDP} at $V_{DD,TDP}$ is assumed to be 120W, which is typical for a server class multi-core processor,



(a) (b)

Figure 3.6: (a) Systematic WID $V_{TH}(L_{EFF})$ variation map, (b) dummy circuit to model leakage power consumption.

and the P_{LEAK} percentage in P_{TDP} at $V_{DD,TDP}$ is varied between 10% and 40%. With the assumed P_{TDP} and the P_{LEAK} percentage in P_{TDP} at $V_{DD,TDP}$, we estimate P_{DYN} and P_{LEAK} using the generated F_{MAX} and P_{LEAK} scaling factors at any given VV_{DD}. With the calculated $P_{TOT,i}$, we compute the T_j of each domain in a die sample using HotSpot [29]. We used 0.3K/W for the convection resistance [30] and the given die size (35mm^2) assuming that the $T_{j,max}$ is 100 °C; 120W power consumption across a 35mm^2 die results in $T_j =\sim 100$ °C with the provided convection resistance.

3.4.4 F_{MAX} Optimization Results

The F_{MAX} optimization method was applied to 100 die samples of 2,4,8, and 16 core processors with D2D and WID variations. Figure 3.7-(a) shows the average active P_{LEAK} reduction after the first PPG tuning step. As the number of cores per die increases, there is more relative P_{LEAK} and F_{MAX} spread between the cores. This provides more opportunity for reducing active P_{LEAK} of fast cores and improving the overall F_{MAX} of a die as a result. As shown in Figure 3.7-(a), P_{LEAK} reduces by 9%-41% for 2-, 4-, 8-, and 16-core

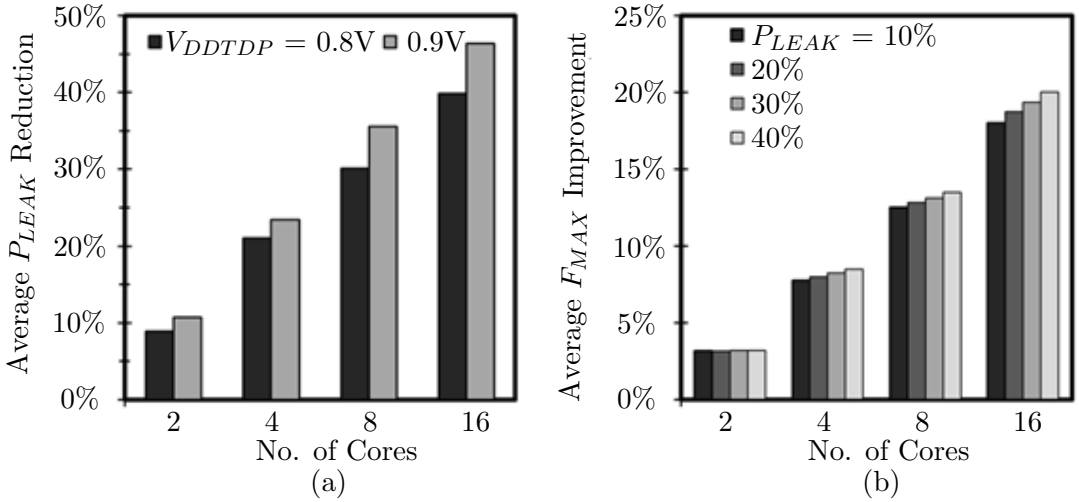


Figure 3.7: (a) Avg. P_{LEAK} reduction after the first PPG tuning step. (b) Avg. F_{MAX} improvement after the second PPG + V_{DD} tuning step. In (a) and (b), P_{LEAK} is responsible for 30% of $P_{TOT,MAX}$ and V_{DDTDP} is 0.8V respectively.

processors after the first tuning step. Here it is assumed that: 1) a processor consumes $P_{TOT} = P_{TOT,MAX}$ at V_{DD} = 0.8 or 0.9V (i.e.,V_{DDTDP}) and 2) P_{LEAK} is responsible for 30% of $P_{TOT,MAX}$ before the first PPG tuning step. Since P_{LEAK} scales more substantially at higher voltage, processors with higher V_{DDTDP} can provide more P_{LEAK} reduction opportunities; V_{DDTDP} = 0.9 V offers 2%-6% more P_{LEAK} reduction, potentially leading to more improvement in F_{MAX}. When V_{DDTDP} is 0.8V and P_{LEAK} is 40% of $P_{TOT,MAX}$ at V_{DDTDP}, F_{MAX} can be improved by 3%-21% on average for 2-, 4-, 8-, and 16-core processors as shown in Figure 3.7-(b). The percentage of P_{LEAK} in $P_{TOT,MAX}$ should also impact the F_{MAX} improvements since P_{LEAK} can change more dramatically than P_{DYN} for adjusting the PPG device and V_{DD}. However, as illustrated in Figure 3.7-(b), increasing the percentage of P_{LEAK} in P_{TOT} from 20% to 40% results in only 1%-5% difference in F_{MAX} improvement for 2, 4, 8, and 16 cores. This is because P_{LEAK} scales at a similar rate as P_{DYN} when V_{DD} is around the V_{DDTDP} region. For a given $P_{TOT,MAX}$ constraint, the F_{MAX} improvement can be affected by V_{DDTDP} in two ways: 1) A higher V_{DDTDP} can re-

sult in more P_{LEAK} scaling as shown in Figure 3.7-(a), but 2) there is less power headroom to improve F_{MAX} because a design with lower V_{DDTDP} is assumed to have higher power than one with higher V_{DDTDP} at the same V_{DD}. Hence, the difference of V_{DDTDP} should not affect the average F_{MAX} improvement significantly; $V_{DDTDP} = 0.9$V provides less than 1% difference in F_{MAX} improvement for 2-, 4-, 8-, and 16-core processors even when P_{LEAK} is 40% of $P_{TOT,MAX}$. Finally, in the above experiments, limiting the voltage drop across the PPG devices to account for noise issues and reliability problems at low voltage, leads to 0%-3% less F_{MAX} improvement depending on the number of cores per chip or the initial fraction of P_{LEAK} in P_{TOT}.

3.5 Related Work

To mitigate yield loss due to process variations, adaptive body biasing (ABB) has been adopted as an effective technique since it can either reduce P_{LEAK} or improve F_{MAX}. J. Tschanz *et al.* [31] showed the impact of ABB on yield with a test chip containing processor critical path replica circuits. Impact of ABB on $F_{MAX}$and P_{LEAK}, applied through separate on-chip power distribution networks, was measured for both D2D and WID variation scenarios. The authors reported recovery of 100% test dies with an area overhead of 2%∼3%. Several problems, however, have been identified in the use of ABB: reverse body biasing (RBB) increases the amount of threshold voltage variations [23]; ABB in dual-V_{TH} design is very difficult due to different body effect coefficient of high- and low-V_{TH} devices [32]; ABB to both NMOS and PMOS devices requires a triple-well process. Adaptive voltage scaling (AVS) after manufacturing testing is another option to mitigate yield loss. T. Chen *et al.* [33] showed significant yield improvement by applying AVS to test circuits in a 0.1μm technology on a per-die basis. However, employing multiple voltage domains on a chip to mitigate WID variations (such as core-to-core F_{MAX} and P_{LEAK} variations in multi-core processors) is very challenging in practice due to the increased cost of 1) design,

verification, and testing time [34] as well as 2) voltage regulators and decoupling capacitors [15]. A programmable (or variable-width) sleep transistor to control the F_{MAX} and P_{LEAK} spread of dies utilizing footer switches, that can be configured to have different widths, was shown by Deogun *et al.* [26]. Their work used variable gate voltage and variable width sleep transistors to reduce the spread of F_{MAX} and P_{LEAK} in 32-bit carry look-ahead adder under the impact of D2D and WID variations. The work shown in this thesis applies the PPG technique to multi-core processors with per-core PG domains where power headroom of fast cores can be redistributed to the slow cores to improve F_{MAX} of the die.

3.6 Chapter Summary

Two methods are proposed to improve the maximum operating frequency and yield of power-constrained multi-core designs. These methods use programmable power gating devices which can be tuned during post-manufacturing characterization process depending on the speed and leakage characteristics of the die or individual cores. The first method recovers discarded dies due to excessive active leakage power; a necessary amount of active leakage power is reduced by decreasing the strength of power-gating devices until the dies are brought back into the acceptable operating region. The effectiveness of the method is demonstrated for two different design scenarios: 1) ASIC type designs with fixed leakage power and 2) processor-type designs with variable leakage power constraints. Our experiments demonstrated that about 88% and 98% of discarded dies could be recovered by the proposed methods in designs with fixed and variable frequency targets, respectively.

The second method improves the maximum operating frequency of multi-core designs implemented with multiple power-gating domains by adjusting the strength of power-gating devices, domain by domain, which is followed by scaling global supply voltage for higher operating frequency. Our experimental results showed that the proposed method improved the maximum operating frequency by 3%-21% for 2-, 4-, 8-, and 16-core processors. More

frequency improvement can be obtained as the fraction of leakage power in total power increases.

Chapter 4

Active Leakage Reduction and Gate Oxide Reliability Improvement of Power-gated Circuits

Power gating (PG) and reverse body biasing (RBB) are commonly used techniques to reduce leakage power during the standby mode of a processor core. However, even when the processor cores are active, each logic block spends a significant amount of time in idle (non-switching) condition. Since full voltage is applied to the circuits in active mode, leakage power is dissipated in the idle blocks in the active mode. This power is further aggravated by the fact that die areas which are active have high temperatures. Device aging due to bias temperature instability (BTI) is a serious concern in scaled technologies since it slows down circuit speed over time. Due to negative bias temperature instability (NBTI) aging, the voltage drop across a PG device utilizing PMOS transistors increases which reduces the virtual rail voltage (VV_{DD}) in the active mode and reduces circuit speed. To

account for the aging effect, the PG device is usually upsized such that required maximum voltage drop is maintained over chip lifetime. Moreover, the PG device is also sized for the worst-case voltage drop partly resulted by a large amount of active leakage current at high temperature. This leads to higher VV_{DD} and active leakage power in early chip life and at low temperature. To minimize active leakage power increase due to this effect, we propose two techniques that adjust strength of a PG device based on its usage and circuit's temperature at runtime. The technique is applied to an experimental setup modeling total current consumption of a processor in 32nm technology and its efficacy is demonstrated in the presence of within-die (WID) spatial process and temperature variations. The proposed techniques also improve the gate oxide reliability of the power-gated circuit due to clamping of the virtual rail voltage.

4.1 Impact of Temperature on VV_{DD} and Active Leakage Power

The voltage drop across a PG device, which is placed between the V_{DD} rail and the VV_{DD} node of an integrated circuit (IC), should be very small (e.g., less than 100mV) to minimize IC's performance degradation. In such a condition, the PG device operates in a linear region, modeled as small resistance. As a result, the voltage drop across the PG device is approximately proportional to total (dynamic + leakage) current of the IC. To prevent any timing failure, the PG device is often sized such that it guarantees a minimum voltage level ($V_{DD,MIN}$) when the maximum current (I_{MAX}) is consumed as shown in Figure 4.1-(a). This figure shows the variation of VV_{DD} as a function of the total current through the PG device. The voltage drop across the PG ($V_{DD}-VV_{DD}$) comprises of two components: a drop due to leakage current (between two dashed lines in Figure 4.1-(a)) and that due to dynamic current (between lower dashed line and $I(VV_{DD})$ axis). Since the amount of leakage current

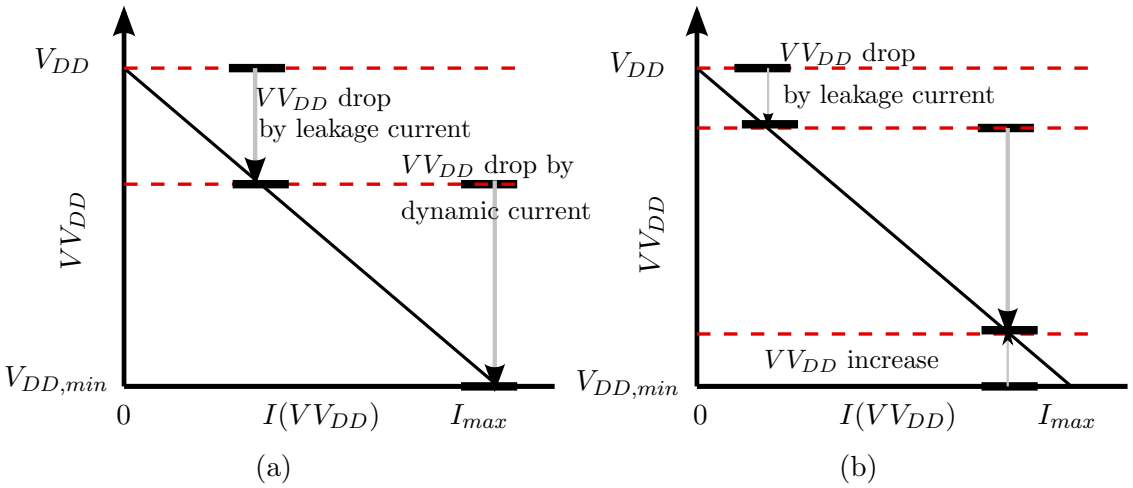


Figure 4.1: VV_{DD} versus total current flowing through a PG device at (a) the worst-case and (b) lower temperatures.

is significantly smaller at low temperature, total current decreases, increasing VV_{DD} as illustrated in Figure 4.1-(b). Note that PG voltage drop due to leakage current decreases while the drop due to the dynamic portion of the current remains unaltered. The rise in VV_{DD} causes higher than necessary (i.e., at $VV_{DD} = V_{DDmin}$) leakage power consumption at the low temperature. Note that VV_{DD} and leakage current will impact each other cyclically until an equilibrium state is reached. Figure 4.2 shows a temperature effect on VV_{DD} and leakage current of an IC connected to a PG device versus process corners modeling D2D process variations. Each bar in Figure 4.2-(a) represents VV_{DD} at given temperature. At each process corner (slow, nominal, and fast), VV_{DD} increases due to less leakage (thus total) current as temperature decreases from 100 to 60 °C. Each solid-line in Figure 4.2-(b) illustrates leakage current versus temperature. As temperature decreases, VV_{DD} increases. As a result, leakage power does not reduce as much as expected at lower temperature. Each dashed line in Figure 4.2-(b) shows leakage current when VV_{DD} is clamped to a target level, i.e., 0.947, 0.975, and 0.978V for fast, nominal, and slow corners at 100 °C. Note that there is substantial leakage current difference between the two cases as

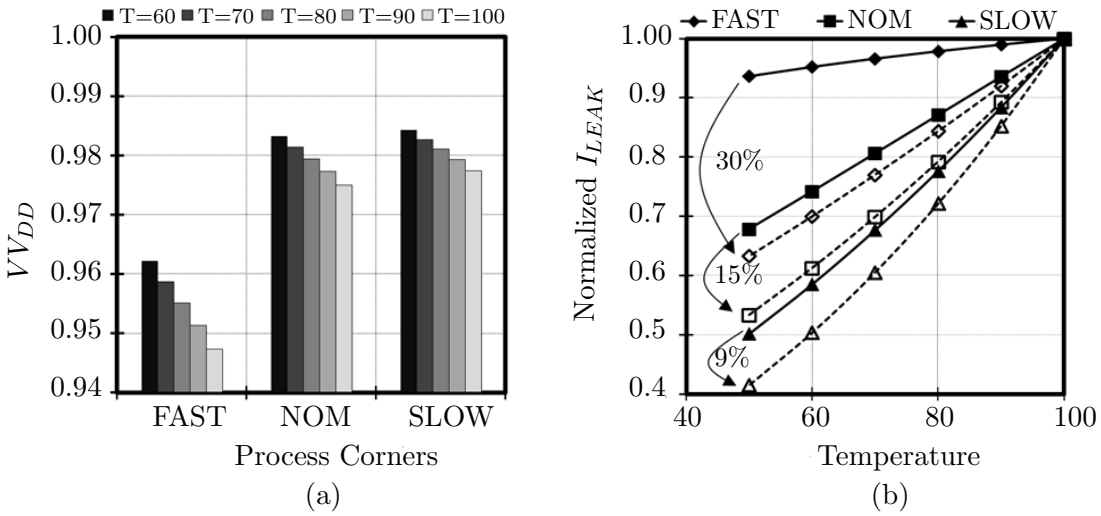


Figure 4.2: (a) VV_{DD} versus process corners for different temperatures. (b) Normalized leakage current versus temperature for different process corners.

temperature decreases; the VV_{DD} increase of the IC connected to the PG device contributes to 7∼30% (fast), 4∼15% (nominal), and 3∼9% (slow) more leakage current at the given temperature range (60∼ 90 °C). This suggests that preventing VV_{DD} increase at lower temperature can reduce a considerable amount of leakage power.

4.2 Impact of NBTI on VV_{DD} and Active Leakage Power

NBTI degrades V_{TH} i.e., threshold voltage of PMOS PG devices over time. This can lead to potential timing failures since the weakened PG devices (thus higher resistance between the V_{DD} rail and the VV_{DD} node) yield lower VV_{DD} than expected when I_{MAX} is consumed. Furthermore, NBTI degradation is strongly dependent on device usage time, temperature, and switching frequency of signals applied to devices [24]. As a result, PG devices exhibit relatively higher V_{TH} degradation than logic ones that experience much faster signal switching frequency with far lower average usage time per gate. Hence, to prevent any potential timing failure in late chip lifetime due to decreasing VV_{DD} over time,

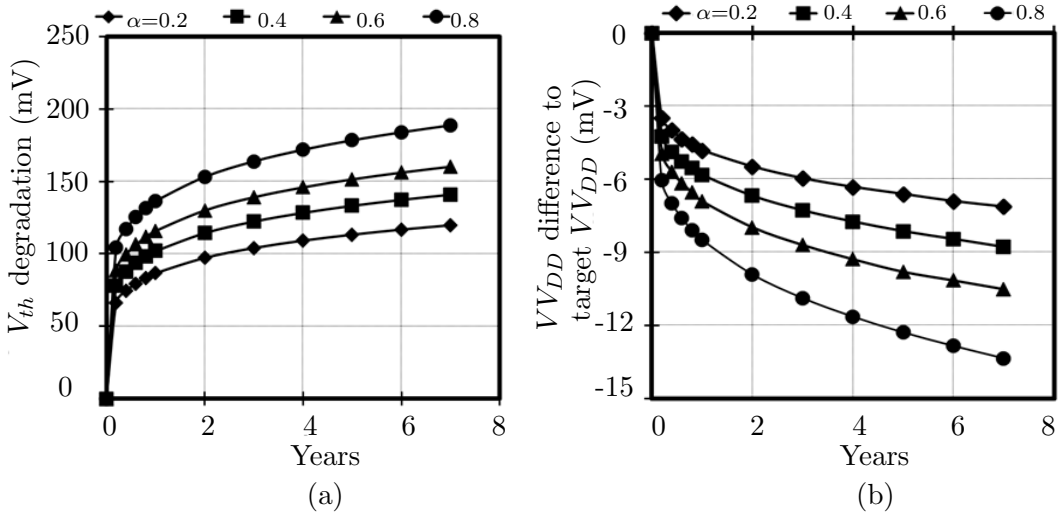


Figure 4.3: Degradation of (a) V_{TH} and (b) VV_{DD} versus time for different duty cycle ratios of the PG wakeup/sleep signal. The target VV_{DD} drop is 25mV.

the PG devices must be upsized considering their expected usage time and signal switching frequency. However, upsizing PG devices to compensate NBTI degradation leads to higher VV_{DD} (thus more leakage power) since the PG devices are stronger (thus less resistance between the V_{DD} rail and the VV_{DD} node) in early chip lifetime.

Figure 4.3-(a) plots the amount of V_{TH} degradation versus stress time of a PG device for various average duty ratios of the PG wakeup/sleep signal (α). We calculate the amount of V_{TH} degradation for different ratios of PG usage time assuming that the average PG wakeup/sleep signal switching period is 1ms. For instance, the PG device is turned on for 0.2ms in every 1ms when α is 0.2. This is a conservative assumption since typical power-management algorithms show a longer wakeup/sleep period than 1ms, leading to more NBTI degradation. Depending on α, PG device's V_{TH} can degrade by 100∼200mV. This, in turn, decreases VV_{DD} as illustrated in Figure 4.3-(b) since the PG device weakens significantly after 7-year degradation. Figure 4.4 shows the impact of upsizing a PG device on VV_{DD} and leakage current in early chip lifetime at different process corners. The PG device is upsized to provide a target voltage drop (25mV) after 7-year NBTI degradation at

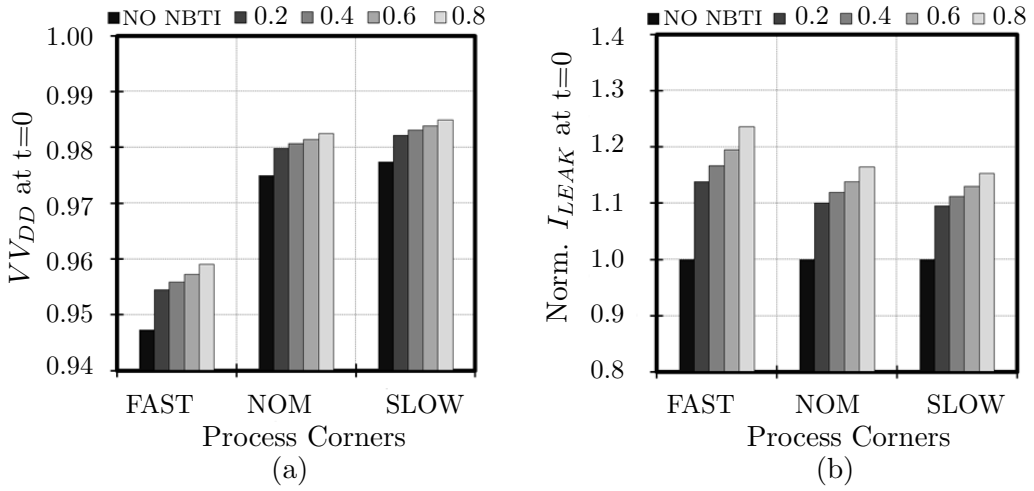


Figure 4.4: (a)VV_{DD} and (b) normalized leakage at t=0 versus process corners for different usage fraction (α).

the nominal corner. This leads to 10∼23% higher leakage current (depending on a process corner and α due to higher VV_{DD} than the required level in early chip lifetime.

4.3 Auxiliary PG Device for Clamping VV_{DD}

Figure 4.5 illustrates two auxiliary (simply AUX) PG devices to compensate the strength of a main PG device for temporal temperature variation of an IC connected to the main PG device and its NBTI degradation. Logically both are separate structures from the main PG device, which will be sized for 60 °C die temperature without considering NBTI degradation in this study, but physically transistors (or fingers) in the AUX PG devices are interleaved with those of the main PG device with a certain ratio. All the fingers in the AUX PG device for adapting temporal temperature variation are turned on at 100 °C, but more of them are gradually turned off as temperature decreases. Finally, all the fingers in the AUX PG device will be completely turned off at 60 °C in this study. This prevents VV_{DD} (thus leakage) increase at low temperature.

All the fingers in the AUX PG device for compensating NBTI degradation are off at

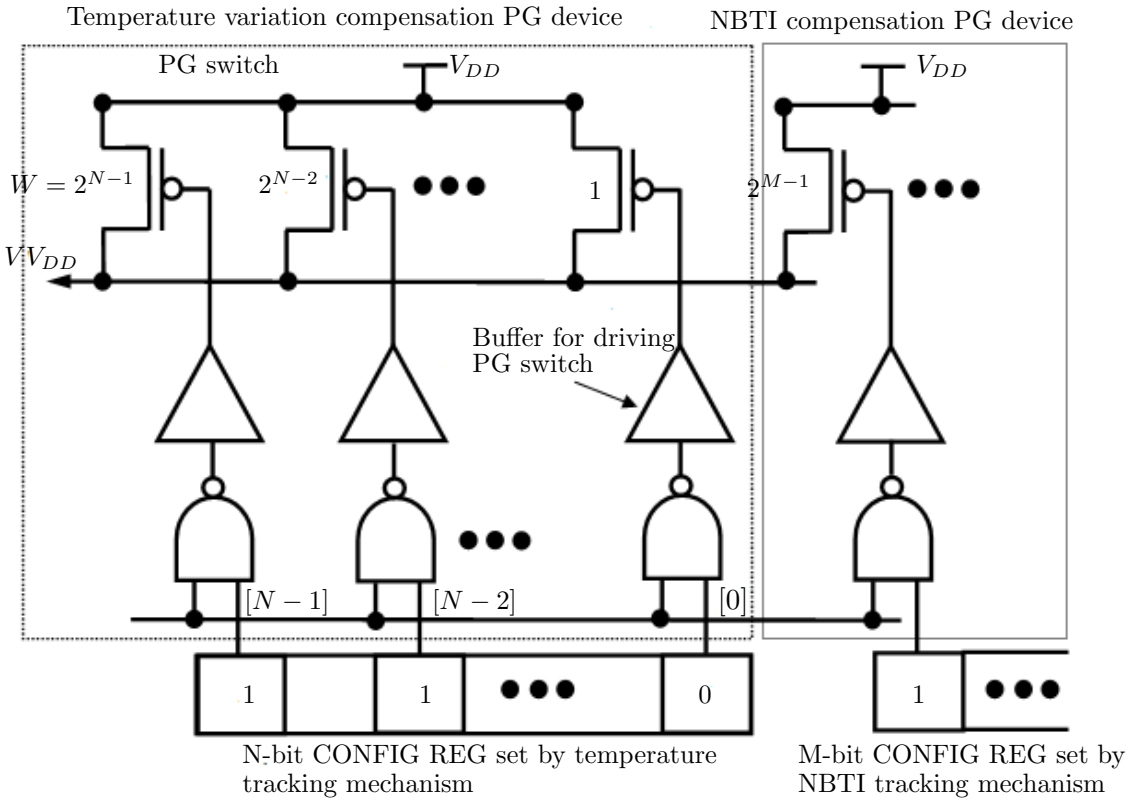


Figure 4.5: Auxiliary PG device for NBTI and Temperature compensation.

time zero, but they are gradually turned on as the main PG device becomes weaker over time. This minimizes leakage increase by suppressing VV_{DD} increase caused by a stronger PG device in early chip lifetime. Meanwhile, it maintains VV_{DD} above a target level throughout the chip lifetime to prevent any timing failures. Each group of PG device fingers is progressively sized such that the strength of two AUX PG devices is proportional to the M- and N-bit binary values from the configuration registers shown in Figure 4.5.The registers are set by the temperature and NBTI tracking mechanisms that will be presented in the following section. The AUX PG devices in this study compensate the strength of the main PG devices such that they can maintain VV_{DD} close to a target level regardless of temperature variation and NBTI degradation. This will allow us to minimize unnecessary leakage power at low temperature and/or in early chip lifetime. The illustrated AUX PG devices allow us to vary their effective width in the field, modulating their overall strength.

4.4 NBTI Tracking Scheme

Figure 4.6 illustrates the proposed scheme to track NBTI degradation of a main PG device over time using reference and stressed ring oscillators (ROSCs). To track the actual PG device stress time (thus the corresponding NBTI degradation), the stressed ROSC is active only when the PG device is on. Meanwhile, the reference ROSC is active only when a periodic AGING_TEST signal is on. Although NBTI is a very slow process, the AGING_TEST signal will be asserted only once in a while. This will minimize the usage time of the reference ROSC, preventing the degradation of the reference ROSC. As PG devices strength degrades proportional to its usage time, temperature, and voltage, the frequency of the stressed ROSC becomes slower and the value of the counter, clocked by the stressed ROSC, decreases accordingly. When the value of the counter, clocked by the reference ROSC, reaches at a target saturation value (X), the other counter, clocked by the stressed

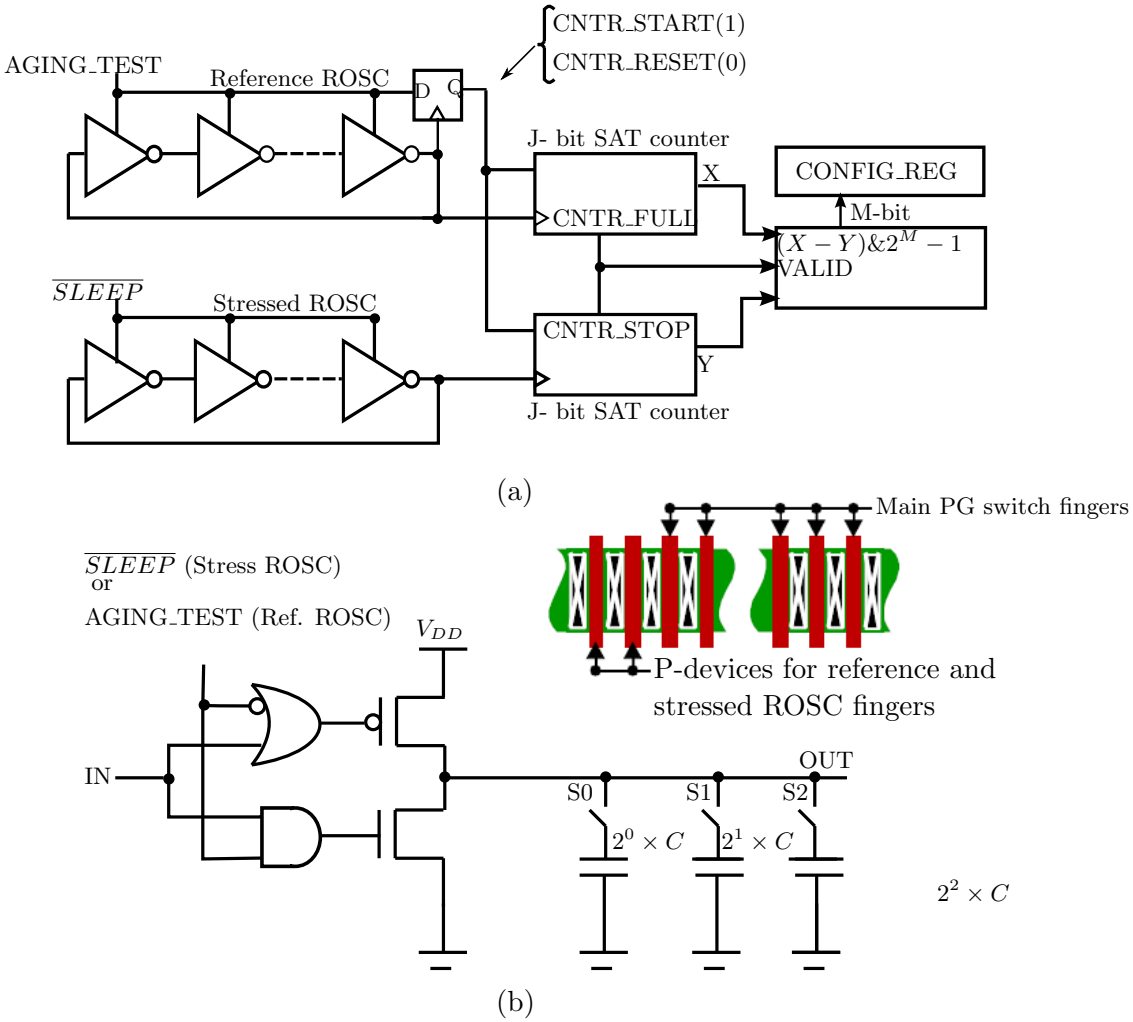


Figure 4.6: (a) A scheme for tracking the NBTI degradation of a main PG device and configuring its AUX PG device accordingly. (b) A circuit for reference and stressed ring oscillators.

ROSC, stops and provides a counted value (Y) during the NBTI characterization period. Since the lower M-bit value of X-Y stored in the configuration register is proportional to the main PG device degradation; it is used to adjust the strength of the AUX PG device, compensating that of the main PG device. To turn on all the fingers in the AUX PG device after 7 years, X should satisfy:

$$X = (2^M - 1)/\left(1 - \frac{f_{ROSC,7\_ROSC\_stressed}}{f_{ROSC,reference}}\right) \tag{4.1}$$

where $f_{ROSC,reference}$ is the frequency of the reference ROSC, and $f_{ROSC,7_{years}tressed}$ is the expected frequency of the stressed ROSC for a process corner after 7 years degradation. A different X value should be programmed into the counter during a manufacturing process using one-time-programmable (OTP) fuses since the rate of degradation is also dependent on a process corner even for the same usage time. The circuit of each ROSC element is also shown in Figure 4.6-(b). When SLEEP or AGING_TEST is off, both NMOS and PMOS are off to prevent any NBTI degradation. The PMOS devices for the ROSCs have the same size as each PG device finger since device degradation is also dependent on the device size. At t = 0, both ROSCs should have the same frequency. The PMOS devices for both ROSCs are placed next to each other to track WID spatial variation and its effect on the frequency of the ROSCs. A large number of ROSC stages is used to cancel out the effect of random variations on the ROSC frequency; 105 stage ROSC is used in this study. A frequency tuning mechanism consisting of capacitors programmed by OTP fuses (S0∼S2), is provided to make the frequency of both oscillators equal or similar at t=0.

4.4.1 Simulation Setup

Figure 4.7 shows the circuit schematic for NBTI tracking experiments. A PG device is first sized to provide 25mV voltage drop with a supply voltage V_{DD} = 1V at the nominal process corner. In such a condition, the dynamic to leakage current ratio of an IC connected

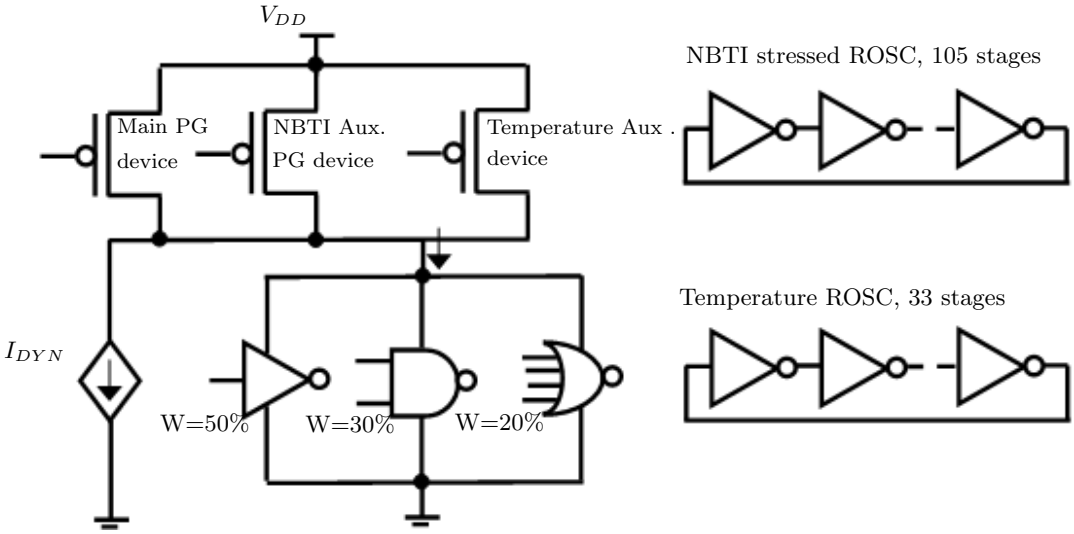


Figure 4.7: Simulation setup for clamping VV_{DD} using auxiliary PG device and NBTI and temperature tracking ring oscillators..

to the PG device is assumed to be 7:3 [3]. The dynamic current is modeled with a current source (7A) modeling the worst-case maximum dynamic current and the leakage current (3A) is modeled with a dummy circuit at 100 °C. The dummy circuit to model the leakage current consists of a large number of INV, NAND, and NOR gates that are comprised of 50%, 30%, and 20% of its total effective channel widths, respectively. Also, randomly selected input states are applied to each gate that has 1∼4 inputs to measure the leakage current. Finally, the fast and slow process corners are modeled by introducing $\pm 3\sigma$ variations of 7.5% and 15% in L_{EFF} and V_{TH}, respectively. The threshold voltage of the PMOS transistors comprising the PG device is increased by an amount equivalent to V_{TH} shift after 7 years of NBTI degradation. The NBTI aging model from [24] is used to calculate the V_{TH} increase for each process corner as follows:

$$\Delta V_{TH} = \left[\frac{\sqrt{k_v^2 \times \alpha \times T_{cycle}}}{1 - \beta^{1/(2n)}} \right]^2 n \tag{4.2}$$

In Eq. (4.2) the parameters k_v and β are given by

$$k_v = \left(\frac{q \cdot t_{ox}}{\epsilon_{ox}}\right) \cdot K_1^2 \cdot C_{ox} \cdot (V_{gs} - Vth) \cdot \sqrt{C} \cdot e^{\frac{2E_{ox}}{E_{o1}}} \tag{4.3}$$

$$\beta = 1 - \frac{2 \cdot \xi_1 \cdot t_e + \sqrt{\xi_1 \cdot C \cdot (1-\alpha) \cdot T_{cycle}}}{2 \cdot T_{ox} + \sqrt{C} \cdot t} \tag{4.4}$$

where T_{cycle} is the average stress cycle time; t is the total aging duration in sec; α is the average signal duty ratio; $E_{ox} = (V_{gs}Vth)/t_{ox}$; $C_{ox} = \epsilon_{ox}/t_{ox}$;$C = (exp(-E_a/kT))/T_0$. Table 4.1 shows the values of the parameters used in the simulations. ΔV_{TH} calculated using Eqn. (4.2) is applied to the PMOS based PG device. Next, the strength of the auxiliary PG device for NBTI is increased until the PG drop is restored to 25mV. This gives the total required size of the auxiliary PG device that must be allocated to account for 7 years of NBTI aging. ΔV_{TH} is applied to the PMOS devices of the NBTI ring oscillator and the frequency reduction at each time step is mapped to a number of auxiliary PG device fingers which are turned on.

Table 4.1: NBTI model parameters used to calculate ΔV_{TH}

T_{cycle}	1ms	C_{ox}	3.5×10^{-20} F/nm
α	0.2∼0.8	C	23.952 nm^2/s
E_{ox}	0.7 V/nm	T_0	10^{-8}
n	0.16(0.25) for H(H_2) diffusion	ξ_1	0.9
ξ_2	0.5	K_1	$8 \times 10^{-4} s^{-0.25} C^{-0.5} nm^{-2}$
E_{01}	0.335 V/nm		

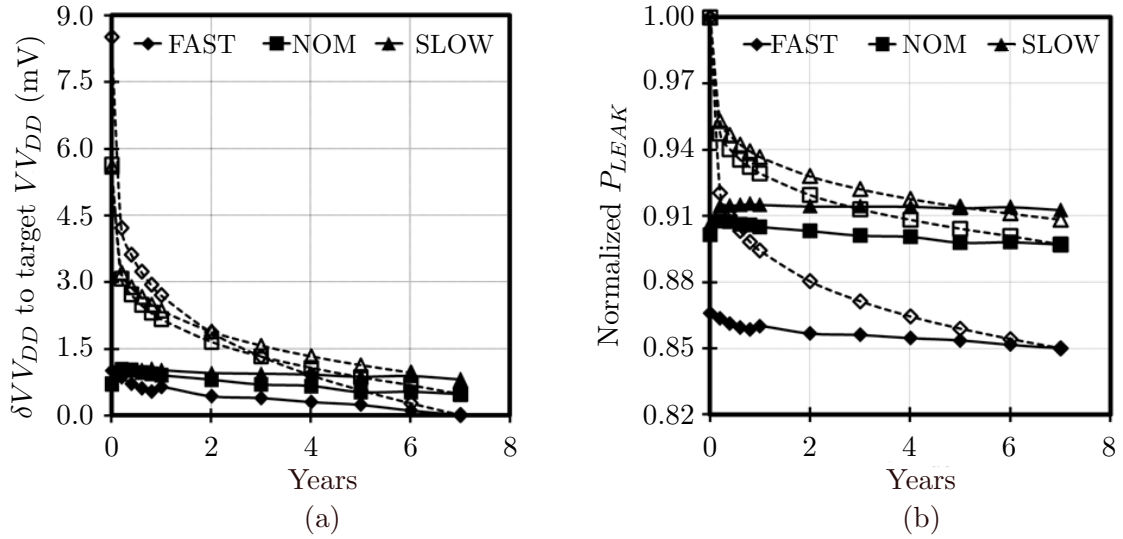


Figure 4.8: (a) VV_{DD} relative to a target level and (b) normalized leakage versus time for different process corners with and without tracking NBTI.

4.4.2 Active Leakage Reduction with NBTI Tracking

Figure 4.8 shows VV_{DD} relative to a target level after 7- year NBTI degradation and leakage power versus usage time for different process corners. The solid- and dashed-lines represent the results with and without tracking and compensating NBTI degradation. The results shown in Figure assume 0.2 usage time ratio and 10% tracking guardband. In other words, to account for various inaccuracy issues in the tracking mechanism, 10% of the auxiliary PG device fingers for compensating NBTI degradation are always on if the main PG device is on. Meanwhile 90% of the auxiliary PG device fingers are controlled by the NBTI tracking mechanism. J and M are assumed to be 9 and 6 respectively for this experiment. As shown in Figure 4.8-(a), VV_{DD} stays at a level very close to its target for each process corner throughout 7 years. This reduces leakage power by 9∼14% in early chip lifetime even with a very conservative PG device usage time as shown in Figure 4.8-(b). In this figure, the leakage power values are normalized to the value at t=0 for each process corner. Note that more leakage will be reduced for the PG device upsized for a higher usage

time ratio. Leakage power saving by clamping VV_{DD} is greatest for circuits at the fast corner as ΔVV_{DD} is highest and leakage power reduces more significantly with VV_{DD} than slow and nominal corners.

4.5 Temperature Tracking Scheme

As discussed in Section 4.1, at low temperatures leakage power of a power-gated circuit does not decrease as expected due to rise in the VV_{DD} level. The following work proposes a method to track temperature change of an IC and clamp VV_{DD} to a target level regardless of die temperature. Figure 4.9 illustrates the proposed scheme to track temporal temperature change of an IC. Current-starved inverters (INVs) are used in the ROSC to make them less sensitive to V_{DD} fluctuation due to noise. We also assume that quiet analog V_{DD} used for PLLs is applied to the ROSC since it often exhibits only Gaussian white-noise. In Figure 4.9, the target saturation value of the K-bit counter, clocked by SYS_CLK clock from a PLL, should be set to provide the time to count Y - the value of the L-bit saturation counter, clocked by the ROSC, at 100 °C as follows:

$$Y = (2^N - 1) / \left(\frac{f_{ROSC,60\,^\circ\mathrm{C}}}{f_{ROSC,100\,^\circ\mathrm{C}}} - 1 \right) \tag{4.5}$$

where $f_{ROSC,60\,^\circ\mathrm{C}}$ and $f_{ROSC,100\,^\circ\mathrm{C}}$ are the frequencies of the ROSC at 60 and 100 °C, respectively. The value of X - Y is inversely proportional to the AUX PG device strength and the lower N bits of the complemented value of X - Y are used to decrease the strength of the AUX PG device at low temperature. Also, the L-bit counter must be saturated at Y + (2N - 1) to prevent an overflow effect. In other words, X - Y must be less than or equal to 2N - 1. A 33-stage current-starved INV chain is used to implement the ROSC that tracks die temperature.

Figure 4.10 shows VV_{DD} relative to a target level and leakage power versus tempera-

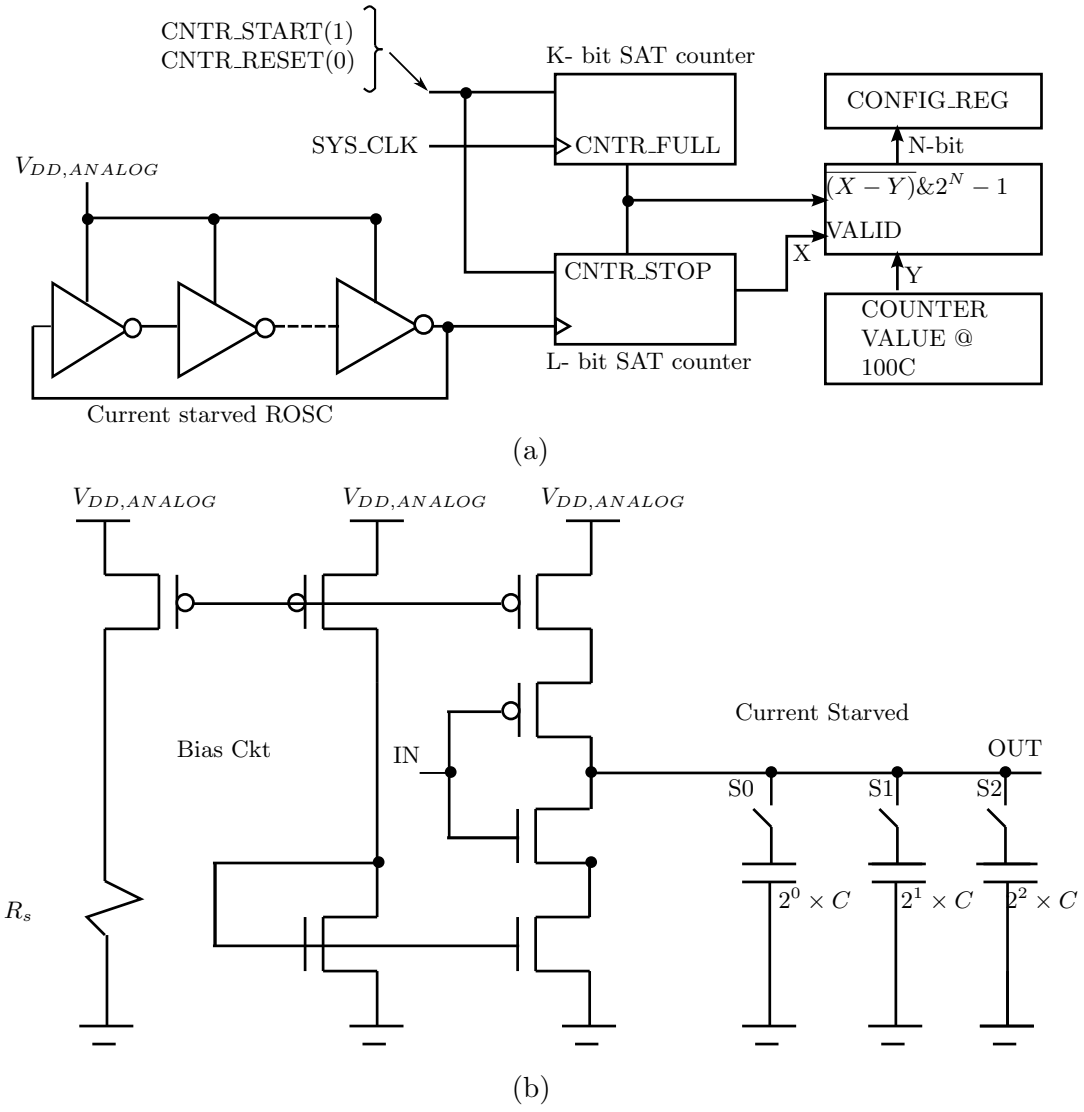


Figure 4.9: (a) A scheme for tracking temporal temperature change of an IC and configuring its AUX PG device accordingly. (b) A ring oscillator circuit implemented with current-starved inverters.

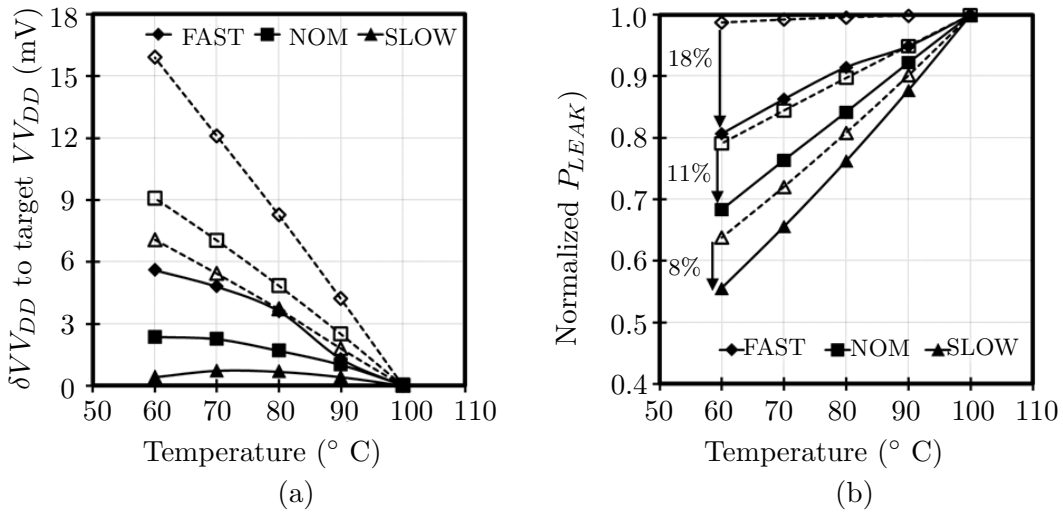


Figure 4.10: (a) VV_{DD} relative to a target level and (b) normalized leakage versus temperature for different process corners with and without tracking die temperature.

ture for different process corners. The solid- and dashed-lines represent the results with and without temperature tracking, respectively. Similar to the NBTI tracking scheme, 5% tracking guardband is used to account for various inaccuracy issues in the tracking mechanism. We also use 9 and 6 for L and N. As shown in Figure 4.10-(a), VV_{DD} is clamped to the value very close to the target level for each process corner for the given range of temperatures. As a result, leakage power is reduced by 5∼18% (fast), 3∼11% (nominal), and 2∼8% (slow) at lower temperature than 100 °C; seen in Figure 4.10-(b). Note that the higher VV_{DD} values than the target ones at 60 °C are partly resulted from the 5% gaurdband applied to the AUX PG device in Figure 4.10-(a).

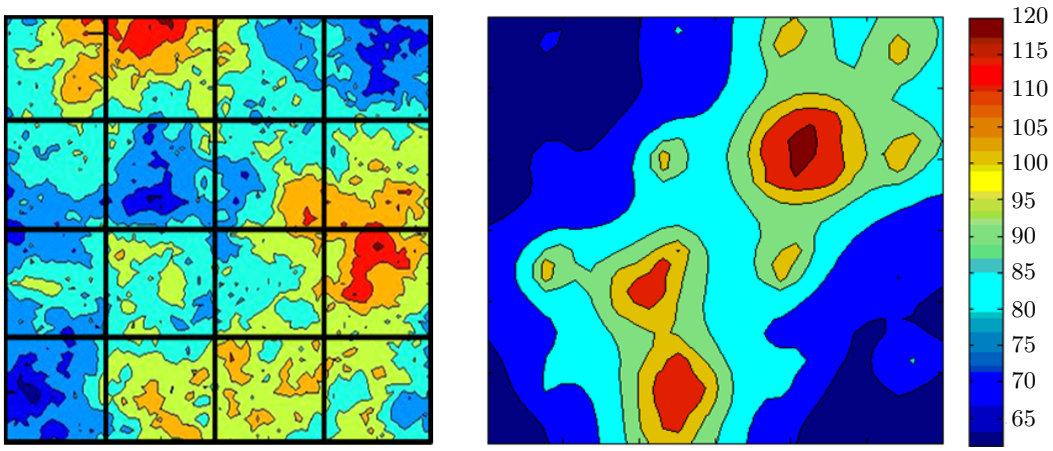
4.6 Tracking with Within-die Spatial Process and Temperature Variations

The ROSCs can track D2D variations fairly accurately for NBTI degradation and temperature variation as shown in Sections 4.4 and 4.5. However, substantial within-die (WID)

spatial process and temperature variations have been observed in ICs manufactured with nanoscale technology. As a result, these WID variations can significantly impact the frequency of the ROSCs depending on their locations on a die. In this section, we demonstrate the efficacy of the proposed tracking circuits in the presence of WID spatial process and temperature variations.

4.6.1 Simulation Setup for Tracking with WID Variations

A WID variation map is generated by mapping an 80×80 grid on to a 9mm×9mm die size. Each point in the grid is assigned a pair of V_{TH} and L_{EFF} values. To model the variation in V_{TH} and L_{EFF} we use the following parameters: WID correlation distance coefficient $\phi = 0.5$, WID systematic V_{TH} variation $\sigma^{sys}_{V_{TH}}$=6.4% and D2D variation $\sigma^{sys}_{V_{TH}}$=5% as presented in [22]. The die is divided into 16 sections (each 20×20 grid points) as shown in Figure 4.11-(a). Each section contains a 0.7A current source and a dummy circuit consuming 0.3A leakage current at the nominal corner connected to a PG device to model dynamic and leakage current, respectively. To track the WID variations, we distribute a chain of INVs across the die. For example, we divide a die into 16 sections and place the $1/16^{th}$ of ROSC INV chains in each section. WID spatial temperature variations are applied by selecting a chip temperature map shown in Figure 4.11-(b). Temperature maps with three different average die temperatures of 65 °C, 74 °C, and 83 °C are used in this study. We apply each corresponding pair of V_{TH} and L_{EFF} values and ΔV_{TH} generated using the 32nm predictive technology [27][28] and NBTI models to simulate the described tracking circuits. SPICE temperature parameter DTEMP is used to apply the temperature variation to individual ROSC inverters.



(a) (b)

Figure 4.11: (a) A WID systematic L_{EFF} and V_{TH} variation map (b) A temperature variation map of the corresponding die.

4.6.2 Tracking circuit performance with WID variations

Figure 4.12 plots VV_{DD} and leakage power versus time for a die sample that models WID process and temperature variations. We pick a die sample that lies between fast and slow corners in terms of its average process parameter characteristics. Also, we use three different temperature maps with the average die temperatures of 83, 74, and 65 °C. As shown in Figure 4.12-(a), the proposed NBTI and temperature tracking schemes track WID spatial process and temperature variations, maintaining nearly constant VV_{DD} regardless of die temperature or PG device usage. As a result, in early chip life, we could reduce leakage power by 10, 12, and 13% for the three different average die temperatures, respectively. Note that the leakage power values are normalized to the value at 83 °C at time zero in Figure 4.12-(b). On average of 100 die samples, in early chip lifetime, we can reduce the leakage power by 8, 9, and 10% for 83, 74, and 65 °C die temperatures respectively.

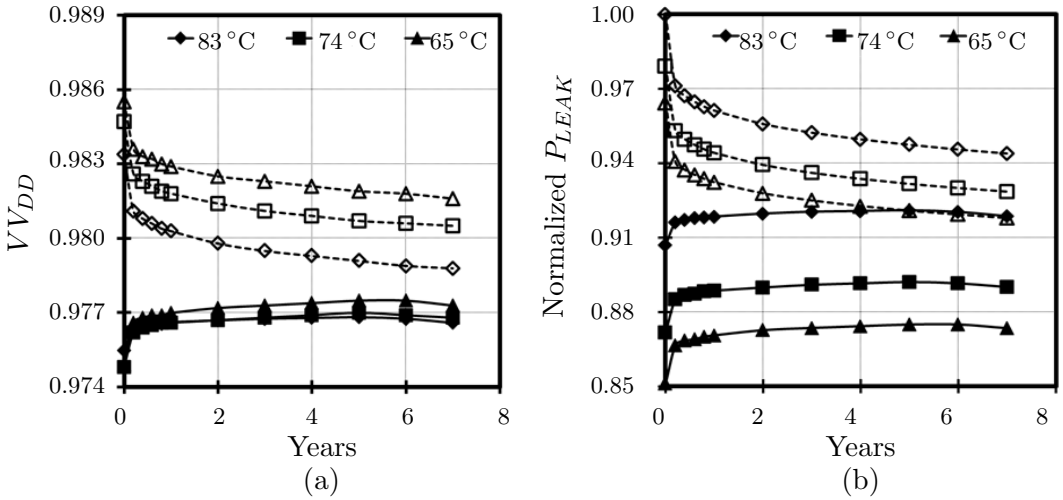


Figure 4.12: (a) VV_{DD} and (b) normalized leakage power versus time for different die average temperatures with and without tracking the NBTI and temperature for a die sample.

4.7 Impact of VV_{DD} Clamping on Gate oxide Reliability

As shown in Figure 4.4-(a), upsizing the PG device to account for 7 years of NBTI aging increased VV_{DD} at t=0 by 8.5mV, 5.63mV, and 5.59mV at the fast, nominal and slow process corners, respectively. This increase of VV_{DD} in early chip life has a substantial deteriorating effect on the reliability of the gate oxide of devices. With technology scaling, the oxide thickness of transistors has been reduced aggressively while the operating voltage has not scaled with the same pace. As a result, transistors are subjected to high electric field stress leading to rapid degradation of oxide reliability and ultimately oxide breakdown. The breakdown process is attributed to defect formation in the gate oxide which accumulate over the device lifetime and form a conducting path through the oxide [35].

Since defect generation in the oxide is a non-deterministic process, oxide time to breakdown (TBD) is modeled as a random variable and its distribution is typically described by fitting experimental breakdown data to Weibull probability distribution function which is

given by [35]

$$F(T_{BD}) = 1 - \exp\left[-\left(\frac{T_{BD}}{\alpha}\right)^{\beta}\right] \tag{4.6}$$

where F is the cumulative distribution function or failure rate of the time to breakdown, α is the characteristic lifetime or the time for 63.2% of the device samples to breakdown and β is the slope parameter of the Weibull distribution. The dependence of α on voltage, temperature and device area was shown with experimental breakdown data in [36]. The slope parameter β varies linearly with the oxide thickness and can be assumed to be independent of voltage and temperature [37][38]. Figure 4.13 shows α as a function of voltage for four temperatures and oxide area of $0.02048cm^2$ (corresponding to 1 billion devices with W = 64nm and L = 32nm) obtained by extrapolating the results shown in [36].

Due to the power-law dependence of oxide breakdown time on operating voltage, clamping VV_{DD} can improve the chip oxide reliability compared to the higher VV_{DD} resulting from conventional PG device sizing. In the case of time varying voltage and temperature,

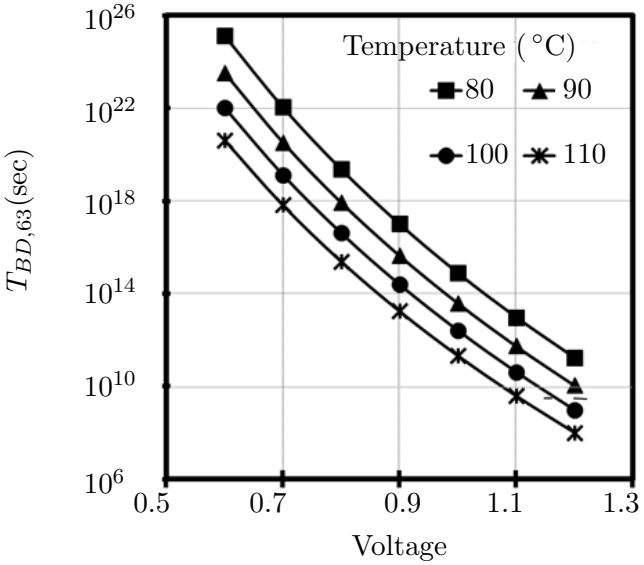


Figure 4.13: Time to breakdown ($T_{BD,63\%}$) vs. Voltage for different temperatures for tox = 1nm based on experimental data in [36].

the failure probability in Eq. (4.6) can be computed by using the dynamic reliability model in [39]. In this section this model is used to calculate the failure rates for a chip containing 1 billion devices with oxide thickness of 1nm. The oxide reliability at any time 't' is obtained by dividing the entire time period into k time frames, each of duration Δt during which the voltage is assumed constant [39]. The failure rate at time t, $F(k\Delta t)$ can be found by computing the probability that the system does not fail with any single time interval from 0 to $k\Delta t$ as follows:

$$F(k\Delta t) = 1 - \prod_{i=1}^{k} [1 - (F(i\Delta t) - F((i-1)\Delta t))] \tag{4.7}$$

where $F(i\Delta t) - F((i-1)\Delta t)$ denotes the probability of device failure in the i^{th} interval calculated using updated values of voltage at the beginning of the interval. The model parameters (α and β) were obtained from [36] after applying the area and thickness scaling laws. Figure 4.14-(a), (b) and (c) show the failure rate (F) and percentage reduction at 100 °C over a period of 7.5 years for the fast, nominal and slow process corners, respectively. At t = 2, 5 and 7.5 years, respectively, the failure rate is reduced by 10.4%, 6.7%, and 5.1% for the fast corner, 8%, 5.1%, and 3.8% for the nominal corner, 8.3%, 5.4%, and 4.1% for the slow corner.

4.8 Related Work

Vattikonda *et al.* [40] and Bhardwaj *et al.* [24] developed predictive models for threshold shift due to NBTI considering static and dynamic stresses, and explored the design space of basic circuits using the models. Prior work on NBTI tracking and measurement has been done in the context of achieving high resolution, unbiased and uninterrupted measurement with minimal and inexpensive test hardware. Fernndez *et al.*[41] demonstrated on-chip NBTI measurement with individual PMOS test device and inverter, capable of

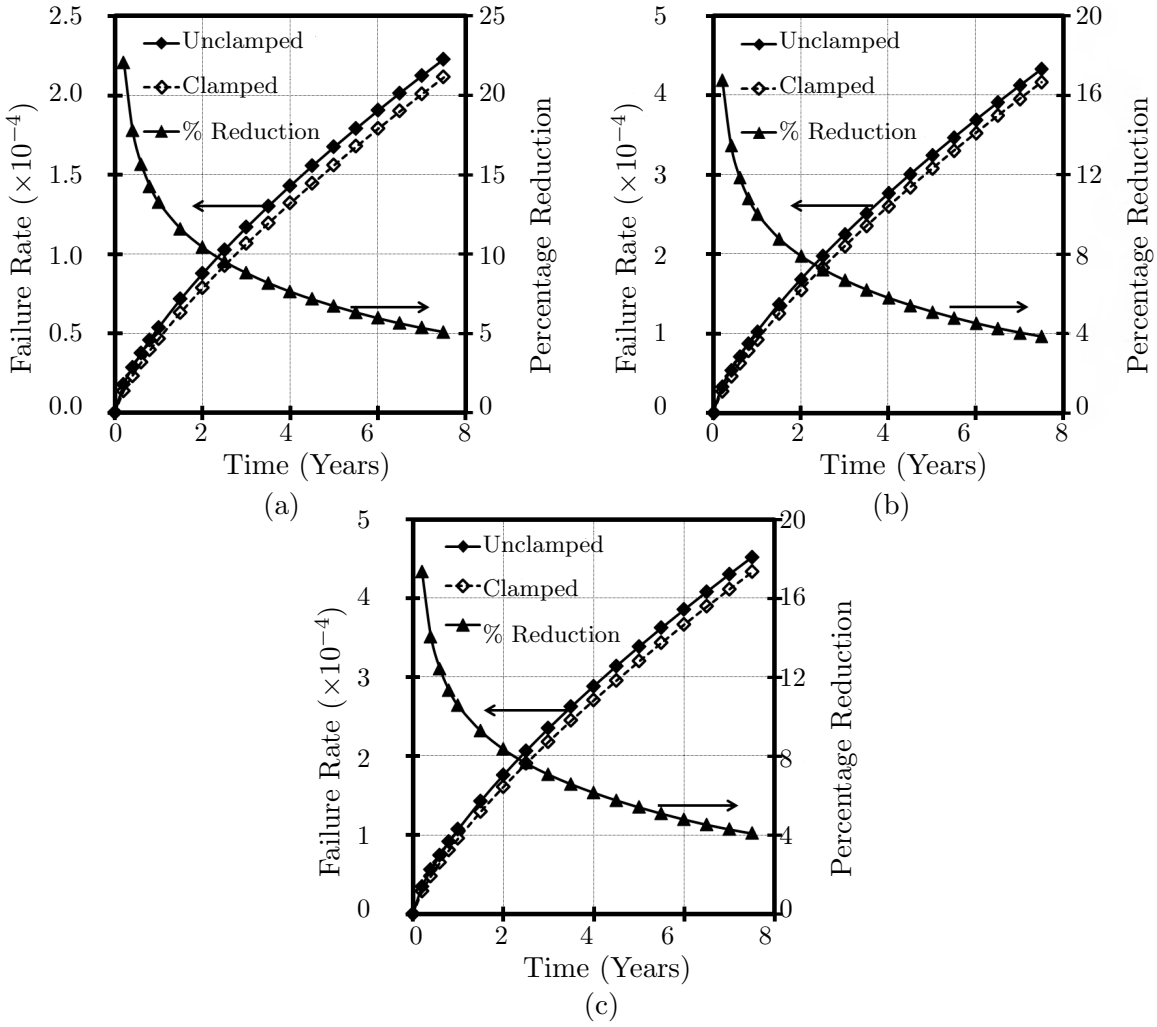


Figure 4.14: Oxide failure rate vs. time with unclamped (solid) and clamped(dashed) VV_{DD}, T=100 °C for (a) fast (b) nominal, and (c) slow process corner.

being stressed with high-frequency AC and DC stress in both interrupted and on-the-fly fashion. An on-die ring oscillator with frequency divider is used to generate the AC stress signal. They measured the ΔV_{TH} due to both AC and DC stress and claimed that the V_{TH} shift due to AC stress is independent of the frequency of the stress signal over a 1Hz-2GHz range. Kim *et al.* [42] demonstrated an NBTI measuring circuit consisting of stressed and unstressed ring oscillators with high resolution and immunity to temperature and voltage fluctuations.

Dependence of oxide breakdown time on voltage and temperature was experimentally studied by Wu *et al.* [37]. In thin oxides, breakdown time follows power law dependence with respect to voltage ($TBD\ (V) - n$) where the power law exponent, n, is independent of oxide thickness to a first order. The time dependent failure rate model presented in Section 4.7 is adopted from the dynamic reliability framework presented by Zhuo *et al.* [39]. This framework was used in conjunction with a DVFS processor to balance reliability degradation with performance.

4.9 Chapter Summary

In power-gated ICs, low temperature does not reduce leakage power at the expected rate due to increased VV_{DD}. An upsized PG device to compensate NBTI degradation in later chip lifetime increases leakage power in early chip lifetime due to the increased VV_{DD}. Two techniques are shown that track NBTI degradation and temperature variation, and adjust the strength of PG devices accordingly. They clamp VV_{DD} close to a target level at runtime in spite of any given NBTI degradation and/or temperature variation within the specified ranges. As a result, leakage power is reduced by 8∼10% for the given range of average die temperatures in early chip lifetime. We demonstrated that they maintain VV_{DD} close to a target level even in the presence of WID spatial process and temperature variations.

With technology scaling, oxide thickness has been reduced aggressively while the oper-

ating voltage has not scaled with the same pace. As a result devices are subjected to high electric field stress leading to rapid degradation of oxide reliability and ultimately oxide breakdown. Oversizing the PG device to account for NBTI degradation leads to higher than necessary VV_{DD} and thus increased oxide stress. The VV_{DD} clamping method proposed in this work can reduce the oxide stress and improve device reliability. Over a period of 7.5 years, the oxide failure rate is reduced by 5.1%, 3.8%, and 4.1% for fast, nominal, and slow corners by VV_{DD} clamping.

Chapter 5

AVS-aware Power-gate Sizing for Multi-core Processors

In a multi-core processor with per-core power gating (PCPG), the voltage drop across a PG device, for given size, is proportional to each core's total current (I_{TOT}), i.e., dynamic + leakage current ($I_{DYN} + I_{LEAK}$). This voltage drop affects maximum core frequency (F_{MAX}) negatively. PG devices are often sized large to minimize the voltage drop (thus frequency degradation) which takes up considerable die area. A PG device and its associated interfacing circuitry were reported to occupy ∼10% of the core area in [43][44]. This overhead can be expected to worsen with technology scaling because the fraction of leakage power in total power increases as the channel length of devices decreases while the transistor I_{ON}/I_{OFF} ratio does not improve proportionally [45]. Hence, larger PG devices will be necessary in future technologies.

Due to die-to-die (D2D) process variations, dies in the fast corner often violate the P_{TOT} constraint due to excessive P_{LEAK} (and thus P_{TOT}) while ones in the slow corner suffer from lower maximum operating frequency (F_{MAX}) than ones in the nominal corner although they consume much less P_{TOT}. In such cases, adaptive voltage scaling (AVS) can

be applied in a post-manufacturing step to dies to satisfy the power constraint (and thus improve yield) in both low-power and high-performance processors.

In this chapter, we propose a methodology to optimize both the size of PG device and the degree of AVS jointly to minimize the PG device size while maximizing performance and power efficiency of power-constrained processors. First, we analyze the impact of PG device size on both F_{MAX} and P_{TOT} of processors in the presence of process variations. Second, we apply the proposed optimization methodology to multi-core processor at fast, nominal, and slow process corners such that we minimize the device size while maximizing performance and power efficiency under a power constraint. Finally, we extend our analysis and optimization for multi-core processors under the impact of within-die (WID) process variations. Impact of the proposed sizing methodology on processors with global and frequency-island (FI) clocking is examined.

5.1 Impact of D2D Variations and PG Size on VV_{DD}, F_{MAX}, and P_{TOT}

The voltage drop across a PG device is a function of the total current drawn by the circuit connected to the device and the width of the device. The voltage drop increases as I_{TOT} increases, lowering the virtual rail voltage (VV_{DD}) which is given by:

$$VV_{DD} = V_{DD} - R_{PG}(V_{DD}, VV_{DD}) \times I_{TOT}(VV_{DD}) \tag{5.1}$$

where R_{PG} is the on-resistance of the PG device as a function of both V_{DD} and VV_{DD}. Figure 5.1 shows a graphical illustration of Eq. (5.1) where VV_{DD} is plotted as a linear function of the total PG current (I_{TOT}). A fixed size PG device also exhibits a different amount of R_{PG} depending on its process corner as illustrated in Figure 5.1; channel length (L_{EFF}) and threshold voltage (V_{TH}) variations affect PG device's strength, i.e., R_{PG}. For

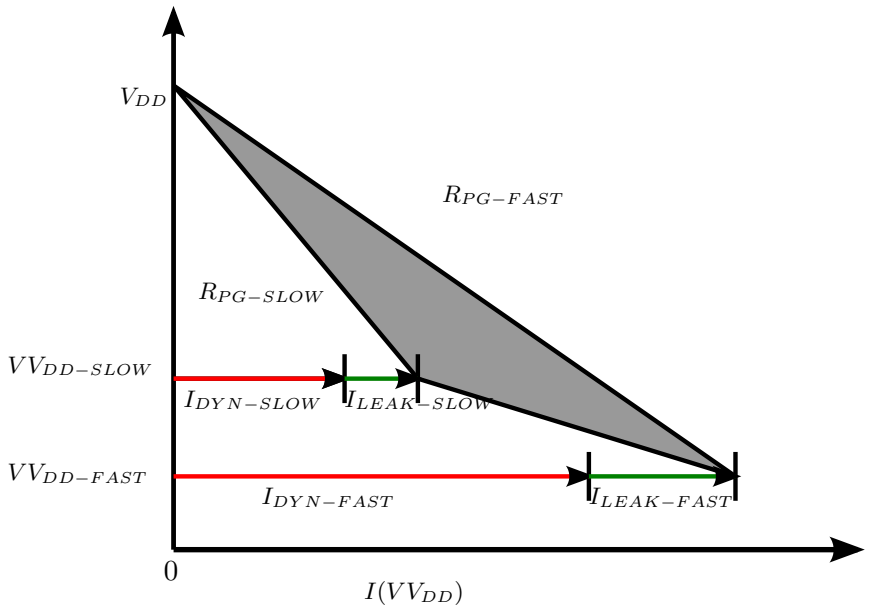


Figure 5.1: Impact of D2D process variations on VV_{DD}.

example, R_{PG} of PG devices in the slow corner, $R_{PG-SLOW}$ is larger than that of PG devices in the fast corner, $R_{PG-FAST}$. Note that R_{PG} is constant for the illustration purpose in Figure 5.1, although it is also a function of VV_{DD} and I_{TOT}. Meanwhile, processors in the slow process corner consume much less I_{DYN} and I_{LEAK} than ones in the fast corner, resulting in a higher VV_{DD} value, $VV_{DD-SLOW}$. On the other hand, processors in the fast corner operate at faster F_{MAX} than ones in the slow corner, consuming more I_{DYN} and I_{LEAK}. Figure 5.2-(a) and (b) present VV_{DD} versus normalized PG device size (s_{PG}) at the slow, nominal, and fast corners; the PG device is sized to provide 25mV voltage drop, i.e., $VV_{DD} = 0.875$V for total 90W power consumption at the nominal corner and 100 °C and I_{LEAK} is 30% of I_{TOT} at the nominal corner. As the PG device size increases, R_{PG} decreases. This increases VV_{DD}, i.e., decreases the voltage drop across the PG device. Note that in a PG circuit both VV_{DD} and R_{PG} impact each other. However, the VV_{DD} increase diminishes rapidly as the PG size increases while the main incentive to implement large PG device size is to minimize the F_{MAX} degradation incurred by the voltage drop.

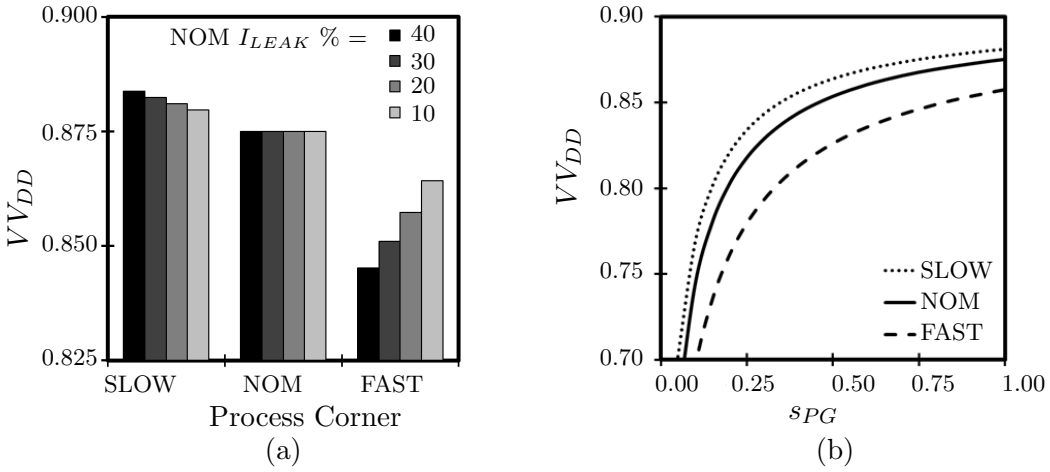


Figure 5.2: VV_{DD} versus process corner for 4 different leakage fractions in P_{TOT}. (b) VV_{DD} versus PG size for 3 different process corners.

Figure 5.3-(a) and (b) show F_{MAX} and P_{TOT} versus PG-device size at the slow, nominal, and fast corners. Both F_{MAX} and P_{TOT} are normalized to the F_{MAX} and P_{TOT} values at the nominal corner and $s_{PG} = 1$. As shown in Figure 5.3-(a), the increase of F_{MAX} diminishes quickly with a larger PG device. For a quantitative comparison purpose, we define the rate of F_{MAX} increase as a function of $s_{PG}(ROIF_{MAX})$ as follows:

$$ROIF_{MAX}(s_{PG}) = \frac{\Delta f_{MAX}(s_{PG})}{\Delta s_{PG}} \tag{5.2}$$

where s_{PG} is the normalized PG-device size, $f_{MAX}(s_{PG})$ is the normalized F_{MAX} as a function of s_{PG}. Note that $ROIF_{MAX}$ becomes less than 0.1 once s_{PG} is larger than 0.43, 0.53, and 0.72 at the slow, nominal, and fast corner, respectively at the points marked by "▲" in Figure 5.3-(a). Similar to $ROIF_{MAX}$, the rate of P_{TOT} increase as a function of s_{PG}, i.e.$(ROIP_{TOT})$ is defined as follows:

$$ROIP_{TOT}(s_{PG}) = \frac{\Delta p_{TOT}(s_{PG})}{\Delta s_{PG}} \tag{5.3}$$

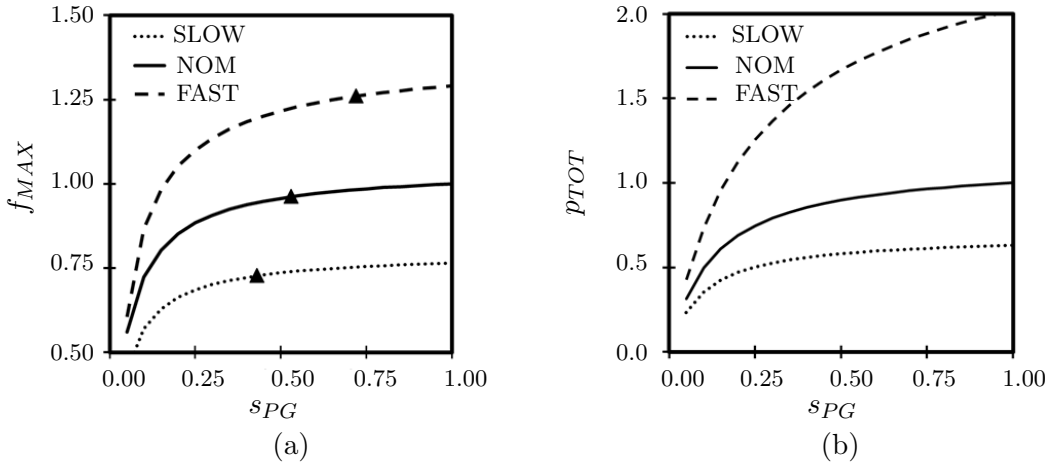


Figure 5.3: (a)f_{MAX} versus s_{PG} and (b) p_{TOT} versus s_{PG} at the slow, nominal, and fast corners.

where $p_{TOT}(s_{PG})$ is the normalized P_{TOT} as a function of s_{PG}. As shown in Figure 5.3, $ROIP_{TOT}$ is more notable than $ROIF_{MAX}$. For instance, increasing the PG size from 0.5 to 1 only improves F_{MAX} by 3% while growing P_{TOT} by 10% in the nominal corner; the $ROIP_{TOT}$ values at the slow, nominal, and fast corners are 2, 3, and 7× higher at $ROIF_{MAX} = 0.1$. This suggests that there must be a performance- and power-optimal PG device size since increasing its size too much does not improve F_{MAX} while increasing P_{TOT} as well as die area associated with the PG device.

5.2 AVS-aware PG Size Optimization for Performance and Power Efficiency

AVS is applied to dies in a post-manufacturing step to maximize F_{MAX} while satisfying the P_{TOT} constraint. A lower value of V_{DD} is programmed for fast dies to limit their power consumption to thermal design power (P_{TDP}) while slow dies are assigned higher V_{DD} to maximize their F_{MAX}. Increasing V_{DD} through AVS can reduce PG device size. In this section, we present a design methodology that optimizes PG device size anticipating

that AVS will be applied to a die to satisfy F_{MAX} and P_{TDP} constraint based on the die's process corner and target computing segment. The proposed PG size optimization algorithm is motivated by the nature of F_{MAX} and P_{TOT} variation with PG size as shown in Figure 5.3. First, the optimization is applied to dies with D2D variations at the fast, nominal, and slow process corners. Next, processors with C2C variations are considered.

5.2.1 PG Size Optimization Problem Formulation

To select the best PG size and V_{DD} combination while maximizing performance and power efficiency, we define a metric performance-cube-per-Watt (P^3/W), i.e., the inverse of energy-delay-square-product (ED^2P). Both performance-square-per-Watt (P^2/W), i.e., the inverse of energy-delay-product (EDP) and P^3/W have been widely used to compare the performance and power efficiency of processors. However, P^3/W emphasizing performance more is more commonly used for high-performance processors while P^2/W is used for mobile processors [46]. In this study, it is assumed that the performance of processors running future multi- and many-core applications, e.g., recognition, mining, synthesis (RMS) applications, is linearly proportional to F_{MAX} within a reasonable range [47]. In such a case, we define the normalized P^3/W (p^3/w) and the rate of P^3/W increase ($ROIP^3/W$) as functions of s_{PG} as follows:

$$p^3/w(s_{PG}) = \frac{(f_{MAX}(s_{PG}))^3}{p_{TOT}(s_{PG})} \tag{5.4}$$

$$ROIP^3/W(s_{PG}) = \frac{\Delta(p^3/w(s_{PG}))}{\Delta s_{PG}} \tag{5.5}$$

The performance-cube-per-Watt was measured for the PG circuit at the nominal process corner while varying PG size and V_{DD} (detailed experimental methodology in Section 5.4). Figure 5.4-(a) shows p^3/w versus s_{PG} for three different V_{DD} values, 0.9, 0.925, and 0.950V at the nominal corner. The $ROIP^3/W$ begins to be less than 0.1 over the point marked with "▲" at each V_{DD} point, implying that P^3/W does not improve effectively with larger

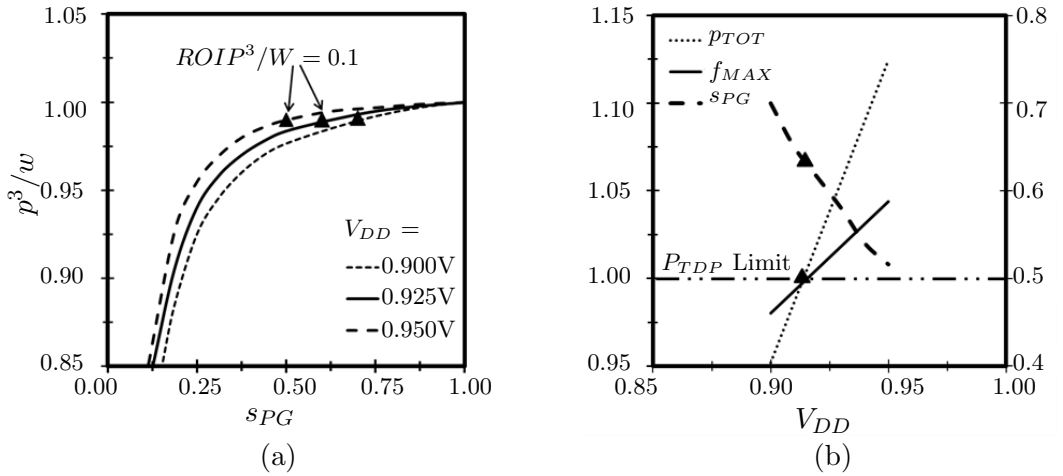


Figure 5.4: (a) p^3/w versus s_{PG} for 3 different V_{DD} and (b) p_{TOT}, f_{MAX} and s_{PG} versus V_{DD} at the nominal corner.

PG device size. Note that the PG device larger than 0.7, 0.6, and 0.5 improves P^3/W only negligibly at V_{DD} = 0.900, 0.925, and 0.950V, respectively.

Note that a smaller PG device can be as P^3/W effective as larger one as V_{DD} increases. Based on this observation, the proposed optimization algorithm constrains $ROIP^3/W$ to be less than 0.1 while minimizing PG size. The optimization problem can be expressed as:

Objective: Minimize(s_{PG})

Constraints: $V_{DD} < V_{DD,MAX}$, $P_{TOT} \leq P_{TDP}$, $ROIP^3/W \leq 0.1$

$ROIP^3/W(s_{PG}) < 0.1$ prevents choosing any arbitrary small s_{PG} value that can lead to very low F_{MAX} satisfying $V_{DD} < V_{DD,MAX}$ and $P_{TOT} \leq P_{TDP}$ constraints. The corresponding pseudo-code for the joint optimization of V_{DD} and s_{PG} is illustrated in Algorithm 1. First, an s_{PG} value is chosen at a diminishing point where $ROIP^3/W$ at each V_{DD} point begins to be less than a target value, e.g., 0.1. A different diminishing point can be chosen for the optimization depending on how much die cost a manufacturer is willing to pay for additional but marginal P^3/W improvement in this case. Second, if P_{TOT} for the given pair of s_{PG} and V_{DD} is less than or equal to P_{TDP}, then s_{PGOPT} is updated with the smallest one so far. Finally, the first and second steps with an incremented V_{DD} value are

Input: Mapping of F_{MAX} and P_{LEAK} to voltage
Output: Optimal PG size and V_{DD}
$V_{DD} = V_{DD,MIN}$;
$s_{PG,OPT} = s_{PG,MAX}$;
$V_{DD,OPT} = V_{DD,MIN}$;
while $V_{DD} \leq V_{DD,MAX}$ **do**
    $s_{PG} = s_{PG,MIN}$;
    while $ROIP^3/W > 0.1$ **do**
        Calculate $ROIP^3/W(V_{DD},s_{PG})$;
        $s_{PG} = s_{PG} + s_{PGstep}$;
    end
    if $P_{TOT}(V_{DD},s_{PG}) \leq P_{TDP}$ *And* $s_{PG} \leq s_{PG,OPT}$ **then**
        $s_{PG,OPT} = s_{PG}$;
        $V_{DD,OPT} = V_{DD}$;
    end
    $V_{DD} = V_{DD} + V_{DDstep}$;
end

Algorithm 1: Algorithm pseudo-code for AVS-aware PG size optimization.

repeated until P_{TOT} becomes equal to P_{TDP} and the most updated s_{PGOPT} and V_{DDOPT} values are returned in the end.

Figure 5.4-(b) shows p_{TOT}, f_{MAX}, and s_{PG} versus V_{DD} at the nominal corner using Alorithm 1. In Figure 5.4-(b), 1) P_{TDP} is 90W, and 2) P_{TOT} and F_{MAX} are normalized to the values at $s_{PG} = 1$ and $V_{DD} = 0.9$V, respectively. This result demonstrates that there is virtually no F_{MAX} degradation with nearly 40% smaller PG device size at $V_{DD} = 0.915$V while satisfying the same 90W P_{TDP} constraint. The corresponding p_{TOT}, f_{MAX}, and s_{PGOPT} points are marked with "▲" at the $V_{DD} = 0.915$V. Table 5.1 summarizes the target P_{TDP} for each process corner; V_{DD} satisfying P_{TDP} with $s_{PG} = 1$; s_{PGOPT} and V_{DDOPT} at the same P_{TDP}; and the resulted f_{MAX} using the optimization algorithm. As noted, faster/slower dies consuming more/less I_{LEAK} are often used for higher-/lower-end computing platforms that are equipped with more/less power-delivery and cooling capacity. Therefore, they usually have higher/lower P_{TDP} budget. The results exhibit that slightly higher V_{DD} leads to much smaller PG-device size at the same P_{TDP} and F_{MAX}; s_{PG} is

Table 5.1: AVS-aware optimum PG size, V_{DD}, and die F_{MAX} across process corners.

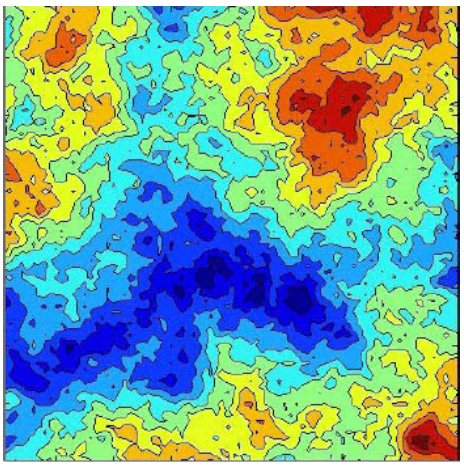
Proc. Corners	**Slow**		**Nom**		**Fast**	
P_{TDP}	65W	70W	90W	100W	120W	130W
V_{DD} @ $s_{PG} = 1$	0.945	0.955	0.900	0.923	0.825	0.840
s_{PGOPT}	0.755	0.715	0.640	0.515	0.568	0.455
V_{DDOPT}	0.955	0.975	0.915	0.948	0.845	0.875
f_{MAX}	0.999	0.999	0.999	~ 1.000	~ 1.000	~ 1.000

reduced by 24∼43% at the slow, nominal, and fast corners.

5.3 AVS-aware PG Sizing with WID Variations

As technology scaling enables more cores to be integrated in a die, core size is becoming smaller relative to die size. In this section, first, we assume that the transistors constituting a PG device are also impacted by WID variations in addition to the core. Then we analyze the impact of WID variations and the number of cores per die on the optimal PG device size and V_{DD}. In particular, both the global clocking (GC) and frequency-island (FI) schemes will be examined since they impact performance and power efficiency of multi-core processors differently in the presence of WID variations.

For the analysis presented in this section, we generated a WID correlated V_{TH} (and L_{EFF}) variation profile for a 16-core die as shown in Figure 5.5-(a). Detailed methodology for generating the map is described in Section 5.4. Figure 5.5-(b) shows the normalized F_{MAX} and P_{LEAK} values for each core normalized to the slowest and least leaky core respectively. In a die sample with process parameters close to the nominal corner, we observe that certain cores are up to 12% faster and 82% leakier than the slowest cores according to Figure 5.5-(b). Such a large amount of I_{LEAK} (thus I_{TOT}) variations affects



1.03 1.15	1.03 1.15	1.03 1.15	1.03 1.15
LLC		LLC	
1.03 1.15	1.03 1.15	1.03 1.15	1.03 1.15
1.03 1.15	1.03 1.15	1.03 1.15	1.03 1.15
LLC		LLC	
1.03 1.15	1.03 1.15	1.03 1.15	1.03 1.15

(a) (b)

Figure 5.5: (a) Systematic L_{EFF} variations across a die.(b) Normalized F_{MAX} and P_{LEAK} of each core in a 16-core processor.

VV_{DD}, F_{MAX}, and P_{TOT} of individual cores. Further, as the number of cores per die increases, the relative F_{MAX} and I_{LEAK} variations among the cores changes more notably.

5.3.1 WID Variations with Global Clocking

When the GC scheme is adopted, the F_{MAX} of a multi-core processor is often determined by that of the slowest core. Meanwhile, faster cores do not contribute to improving F_{MAX} although they consume much higher I_{LEAK} (thus P_{TOT}). Considering WID variations that result in C2C I_{LEAK} (thus VV_{DD}) variations, the P_{TOT} of a die can be expressed as follows:

$$P_{TOT} = \sum_{i}^{N} (I_{DYN,i}(VV_{DD,i}) + I_{LEAK,i}(VV_{DD,i})) \cdot VV_{DD,i} \tag{5.6}$$

where $I_{DYN,i}$ is the I_{DYN} of i^{th} core and N is the number of cores in a die. Since all the cores run at one frequency, $I_{DYN,i}$ can be modeled as follows:

$$I_{DYN,i}(VV_{DD,i}) = F_{MAX,j}(VV_{DD,j}) \cdot C_{EFF} \cdot VV_{DD,i} \tag{5.7}$$

where C_{EFF} is the effective switching capacitance of each core and $F_{MAX,j}(VV_{DD,j})$ is F_{MAX} of the slowest core (j) in a die; experimental methodology to obtain C_{EFF} is described in Section 5.4. Figure 5.6-(a) shows p^3/w versus s_{PG} of a 16-core die sample for three different V_{DD} values, i.e., 0.950, 0.965, and 0.980V at the nominal corner considering WID process variations. For this simulation, the F_{MAX} of the slowest core and total power, P_{TOT}, given by Eqn. (5.6), is measured to calculate p^3/w for each PG size. First, p^3/w reaches the peak values, i.e., ROIP^3/W = 0, with s_{PG} = 0.43, 0.41, and 0.38 at V_{DD} = 0.950, 0.965, and 0.980V before it starts to decrease with increasing s_{PG}. Thus, an s_{PG} value providing ROIP^3/W = 0 can be regarded as the maximum performance and power efficient point; p^3/w for the cases modeling only D2D variations kept increasing although it was very negligible as demonstrated in Section 5.2. Second, the s_{PGOPT} values at the same V_{DD} is smaller than the ones only considering D2D variations (0.37 versus 0.55 at V_{DD} = 0.95V and ROIP^3/W = 0.1). Both the first and second observations can be explained as follows: 1) The faster and leakier cores with lower V_{TH} and L_{EFF} due to WID variations increase I_{LEAK} dramatically as VV_{DD}, (i.e., PG size), increases. 2) However, it does not contribute to increasing F_{MAX} that is determined by the slowest core with higher V_{TH} and L_{EFF} when the GC scheme is used for multi-core processors. 3) As a consequence, this begins to reduce P^3/W as the increase of I_{LEAK} becomes more substantial than that of F_{MAX} with the increase of PG device size. Figure 5.6-(b) presents p_{TOT}, f_{MAX}, and s_{PG} versus V_{DD} of a 16-core processor. f_{MAX} at V_{DDOPT} = 0.960V and s_{PGOPT} =0.41 is 3% higher than V_{DDOPT} = 0.9145V and s_{PGOPT} = 1.00 at the same P_{TDP} = 90W. First, V_{DDOPT} considering WID variations together is higher than the V_{DDOPT} considering only

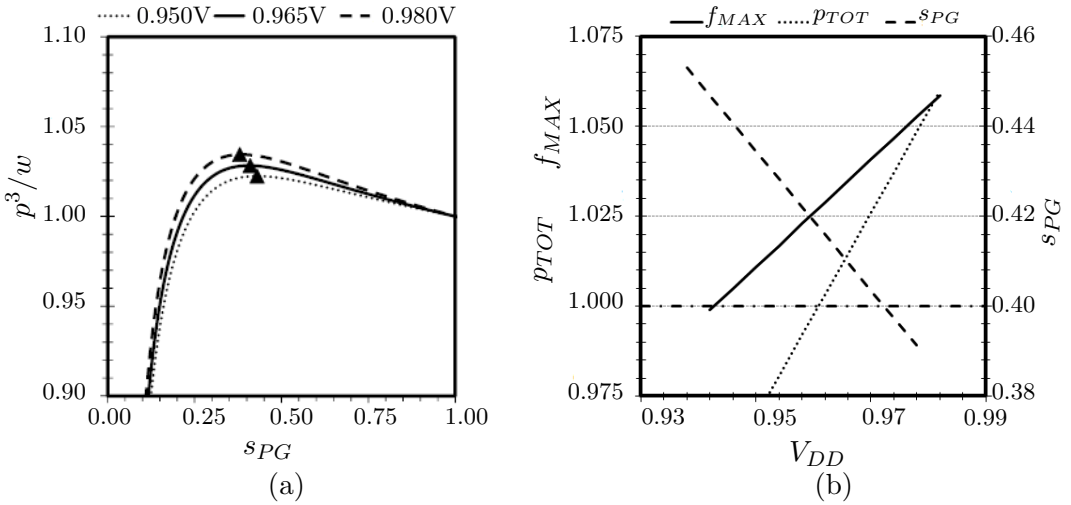


Figure 5.6: (a) p^3/w versus s_{PG} (b) p_{TOT}, f_{MAX} and s_{PG} versus V_{DD} of a 16-core processor die sample.

D2D variations at $s_{PGOPT} = 1.00$ since I_{DYN} with WID variations is lower due to lower F_{MAX} at the same V_{DD}. This provides the headroom to increase V_{DD} to satisfy the P_{TDP} constraint. Second, larger PG devices increase the I_{LEAK} of faster cores more considerably while they do not contribute to the increase of processor F_{MAX}. Meanwhile, smaller PG devices exhibiting larger R_{PG} reduce VV_{DD} of faster cores more considerably than that of slower cores. This decreases the P_{TOT} more effectively with far less F_{MAX} impact; the die in the fast (slow) corner showed much lower (higher) VV_{DD} than one in the nominal corner according to Figure 5.2. However, F_{MAX} begins to decrease too dramatically with much higher power consumption by the PG device itself if the PG device size becomes too small, leading to very low P^3/W. Finally, there is no notable difference in terms of V_{DDOPT} and s_{PGOPT} when 1-, 4-, and 16-core processors are examined with the same WID variation profiles. However, the observed trend is that more cores per die lead to slower F_{MAX} and more P_{TOT} at the same V_{DD} (thus lower P^3/W and smaller s_{PGOPT}).

In summary, it is expected that certain cores will consume even higher I_{LEAK} without contributing to the F_{MAX} due to increasing WID C2C F_{MAX} and I_{LEAK} relative variations

with continuing technology scaling. Thus, relatively smaller PG devices that can suppress the I_{LEAK} of such cores more effectively with the minimal F_{MAX} impact on the slowest core in the die will be favored for the performance and power efficiency. Finally, the optimal s_{PG} and V_{DD} choices reduces the size of PG devices by 59% while improving F_{MAX} by 3% for the power-constrained processor in the presence of WID variations.

5.3.2 WID Variations with Frequency-Island Clocking

The FI scheme allows each core to run at its own F_{MAX}. Thus, it should be more performance- and power-efficient than the GC scheme. The data synchronizations among the cores are very infrequent in RMS applications running on future multi-core processors. Thus, if workloads with a sufficient number of threads are compute-bound, the FI scheme can boost the performance proportional to the average F_{MAX} (F_{MAXAVG}) of all the cores in a die. Then the theoretical upper-bound of performance improvement approaches:

$$F_{MAXAVG} = \left[\sum_{i}^{N} F_{MAX,i}\right] / N \tag{5.8}$$

where $F_{MAX,i}$ is the F_{MAX} of i^{th} core in a die. Figure 5.7-(a) presents the speedup of 4-, 8-, and 16-core processors with FI relative to ones with GC. The F_{MAX} of each core is obtained using the die shown in Figure 5.5 at 0.9V and the detailed properties of all the benchmark programs including the acronyms are illustrated in [48]; Section 5.4 describes the detailed simulation methodology. As the number of cores per die increases, the relative C2C F_{MAX} variations increase, leading to higher F_{MAXAVG}. For instance, the die shown in Figure 5.5 provides 2.6, 4.1, and 6.3% higher F_{MAXAVG} for 4-, 8-, and 16-core processors than the F_{MAX} of a single-core processor. This results in higher speedup for processors with FI relative to ones with GC as the number of cores per die increases in Figure 5.7-(a). Note that certain workloads like AES have higher speedup than the theoretical upper-bound value,

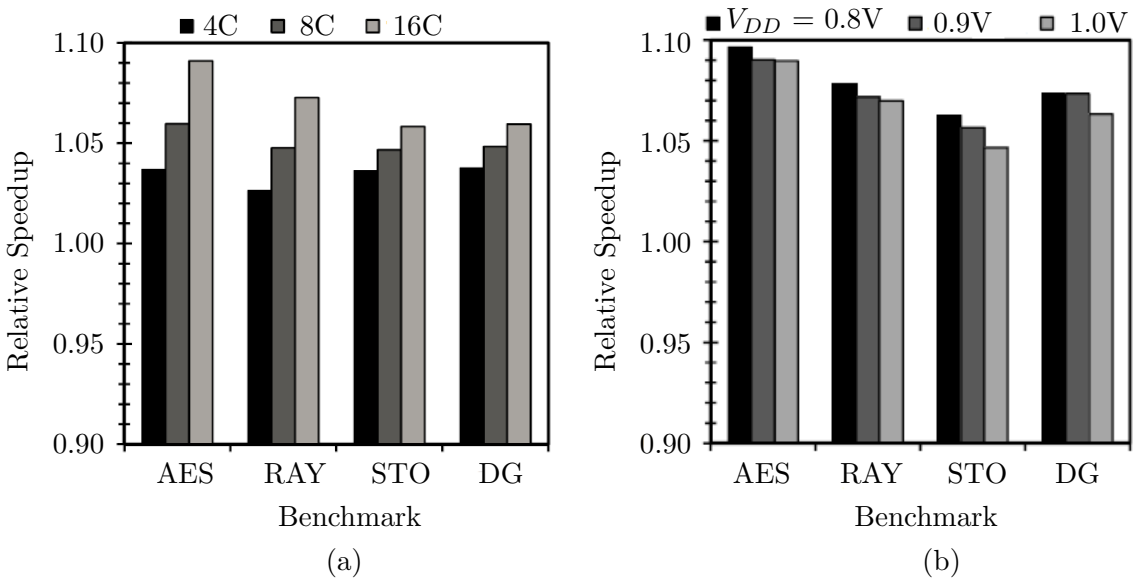


Figure 5.7: Speedup of 4-, 8-, and 16-core processors with FI relative to ones with GC. (b) Speedup versus F_{MAXAVG} at V_{DD} = 0.8, 0.9, and 1.0V.

which can be explained as follows. The workloads run many identical threads containing load/store instructions. They often cause bursty memory accesses since all the cores run at the same F_{MAX} and multiple threads request the accesses at the same time. This results in memory resource conflicts, degrading performance. However, the memory accesses are distributed when all the cores run at different F_{MAX} values. This results in less memory resource conflicts and higher performance. Figure 5.7-(b) presents relative speedup of the 16-core processor die shown in Figure 5.5 at V_{DD} = 0.8, 0.9, and 1.0V compared to a global clocking scheme. This shows that the speedup degrades as V_{DD} (thus F_{MAXAVG}) increases due to the limited main memory resource, but it still scales proportional to F_{MAXAVG}. Note that the compute-bound workloads with a sufficient number of parallel threads are chosen for the simulation to demonstrate that the performance improves proportional to F_{MAXAVG}. Although, either memory-bound workloads do not show the improvement with FI, we assume that the processors are designed for providing the maximum performance for the compute-bound workloads. Then we replace F_{MAX} in Eqn. (5.4) with F_{MAXAVG}

to obtain P^3/W. Since the FI scheme allows an independent F_{MAX} for each core, we can still use Eqn. (5.6) for P_{TOT}, but the $I_{DYN,i}$ can be expressed as follows:

$$I_{DYN,i}(VV_{DD,i}) = F_{MAX,i}(VV_{DD,i}) \cdot C_{EFF} \cdot VV_{DD,i} \tag{5.9}$$

Figure 5.8-(a) shows V_{DDOPT} and s_{PGOPT} at ROIP^3/W = 0 versus the number of cores per die for the same die sample as used in Figure 5.5. As the number of cores per die increases from 1 to 4 and 16, V_{DDOPT} decreases from 0.960 to 0.955 and 0.950V while s_{PGOPT} increases from 0.41 to 0.42 and 0.43, respectively. In terms of p^3/w versus s_{PG} of 1-, 4-, and 16-core processors, they exhibit the same trend with the peak P^3/W values at the V_{DDOPT} and s_{PGOPT} point as Figure 5.6-(a). However, V_{DDOPT} and s_{PGOPT} are shifted depending on the number of cores per die. Figure 5.8-(b) presents p_{TOT}, f_{MAX}, and s_{PG} versus V_{DD} of a 16-core processor for the same die sample used in Figure 5.5. s_{PGOPT} = 0.43 and V_{DDOPT} =0.950V provides F_{MAXAVG} ~3% faster than s_{PG} = 1 and V_{DD} = 0.9145V at P_{TDP} = 90W.

In summary, faster (and leakier) cores also contribute to increasing F_{MAXAVG} when the FI scheme is used for multi-core processors. Thus, s_{PGOPT} increases while V_{DDOPT} decreases as the number of cores per die increases unlike the GC case. In other words, larger PG devices help to improve P^3/W. Meanwhile, when larger PG devices are used, V_{DDOPT} decreases to satisfy the P_{TDP} constraint due to the increased I_{LEAK} in the faster cores. The optimal s_{PG} and V_{DD} choices reduces the size of PG devices by 57~59% while improving F_{MAX} by ~3% for the power-constrained processor in the presence of WID variations.

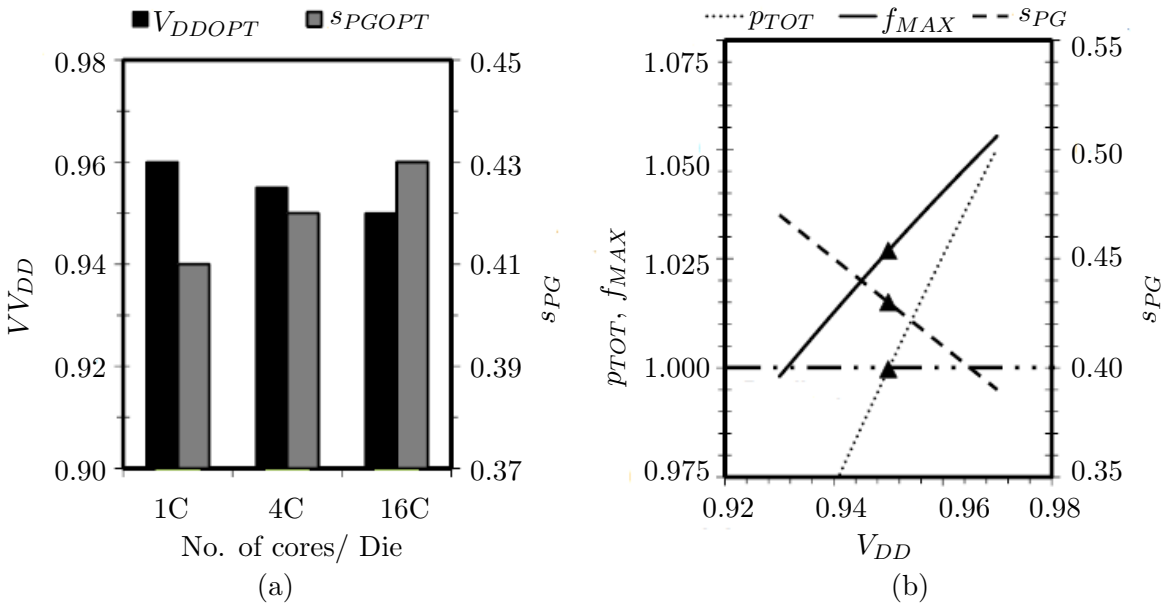


Figure 5.8: (a) V_{DDOPT} and s_{PG} versus the number of cores per die and (b) p_{TOT}, f_{MAX} and s_{PG} versus V_{DD} of a 16-core processor for the same die sample.

5.4 Simulation Methodology

The PG device is sized to provide 25mV drop from the nominal V_{DD} (0.9V) at the nominal process corner and 100 °C in our simulations using the 32nm predictive technology model [27][28]. Under such a condition, the ratio between I_{DYN} and I_{LEAK} of a core connected to the PG device is assumed to be 7:3 [3] for $P_{TOT} = P_{TDP} = 90$W. We model 1) the I_{DYN} with a voltage dependent current source modeling the worst-case peak I_{DYN} and 2) the I_{LEAK} with a dummy circuit as shown in Figure 3.6. To model the voltage dependence of I_{DYN}, first, we obtain F_{MAX} at $VV_{DD} = 0.875$V and its scaling factor, f_{MAX} as a function of VV_{DD} using a 24-stage FO4 inverter chain for a range of VV_{DD} values (0.6∼1.0V). Second, we approximate C_{EFF} based on the following relationship:

$$I_{DYN}(VV_{DD}) = F_{MAX}(VV_{DD}) \cdot C_{EFF} \cdot VV_{DD} \tag{5.10}$$

Table 5.2: Key hardware parameters for GPGPUSim simulations

# of SM Cores	4/8/16	Shared Mem/ SM	16KB
SIMD Width / SM	1/4/8	# of Mem Ch.	4
# of Threads / SM	1024	BW/Mem Ch.	8B/Cycle
3 of CTAs / SM	8	DRAM Rq. Queue	16
# of Registers / SM	16384	Mem Controller	FR-FCFS
Constant & Texture Cache Sizes	8KB, 2-Way, 64B Line	GDDR3 Mem. $t_{CL}/t_{RP}/t_{RC}/t_{RAS}$	10/10/35/25

and the known I_{DYN} value at 0.875V, i.e. (0.7×90W)/0.875V. The dummy circuit for I_{LEAK} modeling consists of a large number of gates (INV: 50%, NAND: 30% and NOR: 20% effective widths) where randomly selected input states are applied to each gate with 1∼4 inputs to measure the leakage power. Third, P_{TOT} is directly measured at the V_{DD} node to include the power consumption of the PG device itself. Finally, f_{MAX} and p_{TOT} as functions of s_{PG} at each V_{DD} point are obtained using a curve fitting tool, Origin 8.1 [49] with the following fitting function: y0 + A1.e-x/t1 + A2.e-x/t2 + A3.e-x/t3. As we increase the number of cores per die, C_{EFF} and I_{LEAK} based on 90W per die is divided by the number of cores, and so are the size of the PG device and the number of gates in the dummy circuit.

For analysis in Section 5.3, a multi-core processor die with WID variations is generated with an 80×80 grid where each grid point is assigned a distinct V_{TH} and L_{EFF} combination. The die area is assumed to be 35mm^2. WID correlation distance coefficient $\Phi(0.5)$, and WID V_{TH} variation σ^{sys} (6.4%) with mean V_{TH} and L_{EFF} values at the nominal corner were used to model WID variation [22]. Note that F_{MAX} is decided by the slowest grid point in each core [21], while I_{LEAK} is obtained from the leakage current of all the grid points in each core.

Finally, we analyze performance scaling trends of various data intensive multi-core workloads using GPGPU-Sim [48]. GPGPU-Sim was modified to support the FI scheme. The number of SM cores, i.e., 4∼16 used in this study aims at the entry-level to mid-range GPGPUs. Table 5.2 summarizes the key hardware configuration parameters; other parameters are identical to ones shown in [48]. A brief description of benchmarks used for throughput experiment is as follows:

- **AES**: Advanced Encryption Standard algorithm in CUDA for encrypting and decrypting files.
- **RAY**: Rendering an image by ray tracing; each pixel corresponds to a scalar thread in CUDA.
- **STO**: StoreGPU is a library that accelerates hashing-based primitives designed for middleware.
- **DG**: gpuDG is a discontinuous Galerkin time-domain solver, used in electromagnetic field computations.

5.5 Chapter Summary

In this chapter, first, the impact of PG-device size on both F_{MAX} and P_{TOT} of processors in the presence of process variation was analyzed. The experimental results showed that the F_{MAX} increase diminishes rapidly while the P_{TOT} grows more notably than the F_{MAX} with a larger PG device than a smaller one. Second, based on above observation, a design methodology is proposed that optimizes both the size of PG devices and the degree of AVS jointly such that the PG device size is minimized while maximizing performance and power efficiency of power-constrained processors. The experimental results demonstrate that the joint optimization considering D2D and WID variations can reduce the size of PG devices

by 59% while increasing F_{MAX} by 3% of power-constrained processors using GC. When the optimization is applied to the multi-core processors using FI, the size of PG devices was reduced by 58 and 57% while improving F_{MAX} by ∼3% for 4- and 16-core processors; the optimal size of PG devices increases gradually while the optimal V_{DD} for AVS decreases as the number of cores per die increases.

Chapter 6

Low Cost Per-core Voltage Regulation Using Power-gating Device

Users run different applications on a processor's cores. These applications differ significantly from each other in their hardware usage patterns and hence instructions per cycle (IPC). In such a diverse core utilization scenario, per-core dynamic voltage/frequency scaling (DVFS) can provide the best power savings while maximizing the performance of the processor. Per-core DVFS requires the implementation of per-core voltage domains on a die with the ability to efficiently scale the voltage of each domain. Most multi-core processors have a single voltage domain which is shared by all cores and powered from a single off-chip voltage regulator (VR). In such a scheme, the VR regulates the voltage of the shared domain to the maximum value demanded by a core. This reduces the performance/power benefits offered by DVFS. Transferring the VR to the die can reduce the amount of decoupling capacitor required on the power delivery network; thus reducing platform cost and enabling faster voltage scaling. On-chip VRs also help to reduce power distribution loss and voltage

drops.

Traditionally, switching VRs have been used invariably as the regulator of choice for powering different types of computing platforms due to their high power conversion efficiency over a wide dynamic voltage range. However, providing per-core voltage domains and integrating switching VRs on the processor die, both present several challenges making the shift to monolithic per-core VRs difficult for manufacturers. In this chapter, we first outline some of the challenges associated with implementing per-core voltage domains and integrating VRs. Next, we propose a method of using the per-core power-gating (PCPG) devices for on-chip voltage regulation and compare them with switching VRs. Using $MIPS^3/W$ as a metric , we analyze and compare the performance and power of the proposed low-cost linear VRs with switching VRs for a set of DVFS workloads.

6.1 Per-core Voltage Domains

6.1.1 Motivation for Per-core Voltage Domains

Most commercial processors support DVFS, yet they only have a single chip-wide voltage domain due to the high cost of supporting multiple voltage domains. As more cores share the chip-wide voltage domain, the performance/power benefit of DVFS diminishes. This is because a single voltage domain cannot allow a multi-core processor to effectively exploit runtime performance (i.e., IPC) variations across cores running multiple threads or applications for a given DVFS interval. For example, some cores running threads in memory-intensive phases can operate at lower V/F without impacting the performance while other cores executing threads in compute-intensive phases must operate at higher V/F to maximize performance. Consequently, many researchers have investigated various DVFS algorithms to exploit multiple voltage domains effectively [15] [50].

Figure 6.1 compares the *million instructions per second cubed per Watt* ($MIPS^3/W$) of

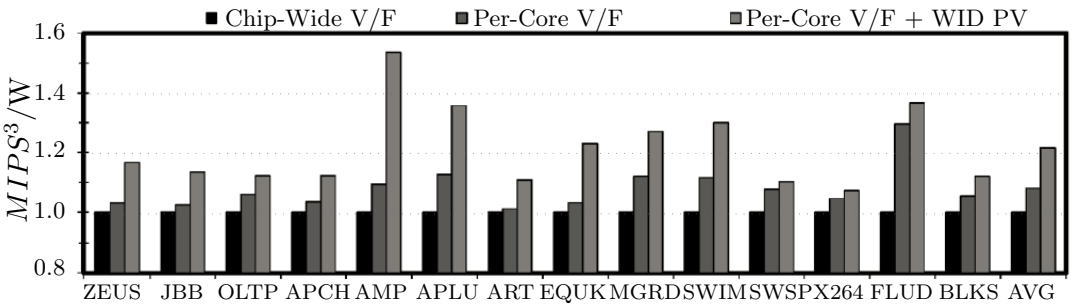


Figure 6.1: MIPS^3/W comparison of 8-core processors supporting per-chip V/F domain, per-core V/F domains, and per-core V/F domains exploiting WID process variations [4]. An oracular approach [15] is applied to each runtime DVFS interval. Each interval is comprised of 10-million executed instructions, which is equivalent to a few hundred microseconds depending on IPC values. A total of 1-billion instructions are executed after 100-million instructions are executed for warming up on-chip caches.

8-core processors that are supported by a chip-wide V/F domain, per-core V/F domains, and per-core V/F domains exploiting within-die (WID) process variations (PV), which leads to C2C frequency and power variations. All the MIPS^3/W results are normalized to that of the 8-core processor with a chip-wide V/F domain without considering the power efficiency of VRs. We use four commercial workloads (Apache, JBB, OLTP, and Zeus denoted by APCH, JBB, OLTP, and ZEUS) [51], six SPEC OMP V3.2 benchmarks (Ammp, Applu, Art, Equake, Mgrid, and Swim denoted by AMP, APLU, ART, EQUK, MGRD, and SWIM), and four PARSEC benchmarks (Swaptions, X264, Fluid, and BlackScholes denoted by SWSP, X264, FLUD, and BLKS) [52] running on a modified GEMS multi-core simulator [53]. An oracle DVFS algorithm [15] is modified to maximize MIPS^3/W for a given maximum power constraint. MIPS^3/W is chosen as a metric to emphasize the performance aspect of server-class multi-core processors under a power constraint more than MIPS/W which is used for mobile processors [54]. See Section 6.3 for the detailed processor configuration and the methodology modeling C2C frequency and power variations due to WID process variations.

The per-core DVFS that does not exploit C2C frequency and power variations increase a

geometric mean of $MIPS^3/W$ by 8%, while the one that exploits the C2C variations increase a geometric mean of $MIPS^3/W$ by 22% over the chip-wide DVFS. Further, as the number of cores per processor is increased, we observe that per-core DVFS achieves even more $MIPS^3/W$ improvement than chip-wide DVFS. For example, the $MIPS^3/W$ improvement of 12- and 16-core processors using per-core DVFS over chip-wide DVFS is nearly 3× and 4× higher than that of a 8-core processor. The increase becomes even larger when the C2C process variations are exploited. This signifies the growing importance of supporting per-core DVFS to maximize performance and power efficiency under a power constraint.

6.1.2 Challenges for Supporting Per-core Voltage Domains

Supporting per-core voltage domains can allow multi-core processors to exploit C2C frequency and power variations, and hence significantly increase performance and power efficiency. However, most commercial processors have only one chip-wide V/F domain. This is because splitting the voltage domain and providing multiple off-chip VRs incurs a high cost for the platform and package designs. Figure 6.2 illustrates the impact of splitting the voltage domain to provide per-core DVFS on the overall VR capacity required. Assume that the maximum power consumption of the processor is limited to 120W and there are four cores. When all four cores are running, each core can consume up to 30W; thus, it seems that each per-core VR only needs to support up to 30W. However, for example, when only two out of four cores are active due to limited parallelism, the two active cores can run at higher V/F (e.g., Intel Turbo Boost TechnologyTM [55]) without violating their thermal and maximum supply voltage constraints for reliability. If the two active cores consume 40W at such an operating voltage/frequency, the capacity of each VR needs to be increased to 40W and the total combined capacity of all the VRs becomes 160W.

When the voltage domain is shared, however, a 120W VR is still sufficient for such a case; the total power consumption of two cores running at the turbo mode is a total of 80W,

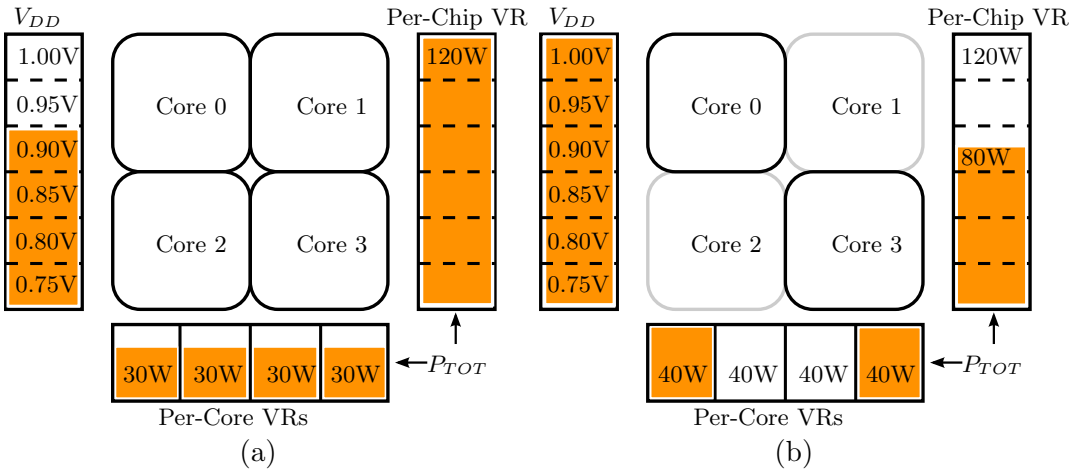


Figure 6.2: Impact of splitting the voltage domain on the overall VR capacity of a quad-core processor to provide per-core voltage domains. All cores are active and consume a total of 120W in (a) and only two cores are active in (b).

which is below the maximum capacity of the VR. Although it is feasible for only a subset of cores to run in turbo mode, platform designers cannot increase the VR capacity for only a subset of the cores. This is because cores are put into turbo mode in a round-robin fashion to prevent excessive aging of a specific core or subset of cores, requiring designers to provide the capacity for turbo mode for all the cores. Finally, increasing WID process variations leads to substantial C2C frequency and (leakage) power variations [20]. In other words, some cores consume notably more power than others due to a high fraction of leakage power in total power (e.g., ∼30% [3]) and a large variation of the leakage power across cores. Thus, the per-core VR capacity is determined by such cores, increasing the overall VR capacity even further.

The increased total power capacity requires larger components for VRs and more package pins for power delivery. Form-factor is critical even for server platforms to maintain high integration density in data centers, and VRs are the second largest components next to DRAM modules; VRs occupy 63% more platform area than the CPU, the third largest components [56]. Furthermore, many commercial chips are heavily constrained by the

available pins; nearly a half of all pins are already dedicated for power delivery and increased overall VR capacity requires more pins. Although the platform and package cost associated with multiple off-chip VRs can be lowered by using on-chip switching VRs [57], integrating cores and high-quality inductors for the VRs on the same chip has also been a major technical challenge for manufacturers, potentially impacting both the efficiency of the VRs and the yield of dies.

6.2 LDO VRs Exploiting C2C Voltage Variations and PCPG device

6.2.1 C2C Voltage Variations

In the experiment for Figure 6.1, the maximum voltage difference between cores in a processor supported by per-core voltage domains is not large at each DVFS interval. Figure 6.3 shows that the maximum voltage difference between cores for the "Per-Core V/F" case in Figure 6.1 is less than or equal to 100mV for at least 90% of the execution intervals in most applications. Similar statistics are observed for the "Per-Core V/F + WID PV" case. In such a case, the power loss in low dropout (LDO) VRs can be lower than that in switching VRs when a proper V_{IN} value for LDO VRs is selected to minimize the difference between the V_{IN} value and the V_O values across cores. Furthermore, an LDO VR can be implemented very cost-effectively since (i) it does not require inductors or large capacitors [58] and (ii) it can share its largest component (i.e., the output device) with a PCPG device.

In the next section, we show that, under such conditions, the power loss of LDO VRs can be lower than the power loss of switching VRs.

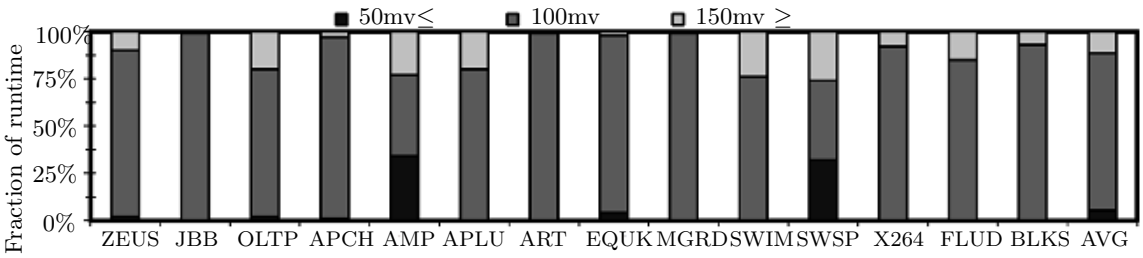


Figure 6.3: Percentage of intervals exhibiting various maximum voltage differences between cores for an 8-core processor supported by per-core V/F domains [4]. An oracle approach, similar to one used in [15], is used to determine V/F of each core at each runtime DVFS interval. Each interval is comprised of approximately 10-million executed instructions, which is equivalent to a few hundred microseconds depending on IPC values. A total of 1-billion instructions are executed after 100-million instructions are executed to warm up on-chip caches.

6.2.2 PCPG Based LDO VRs

PCPG devices are typically provided for commercial multi-core processors to reduce standby leakage power of idle cores [59]. In active state, a PG device incurs a small voltage drop across it (i.e., between the supply voltage and the actual voltage applied to the core). The voltage drop is inversely proportional to the size (i.e., total transistor width) of the PG device for a given amount of total current (dynamic + leakage) drawn by the core. In fact, the voltage applied to the core can be modulated by controlling the effective width (i.e., resistance) of the PG device [60].

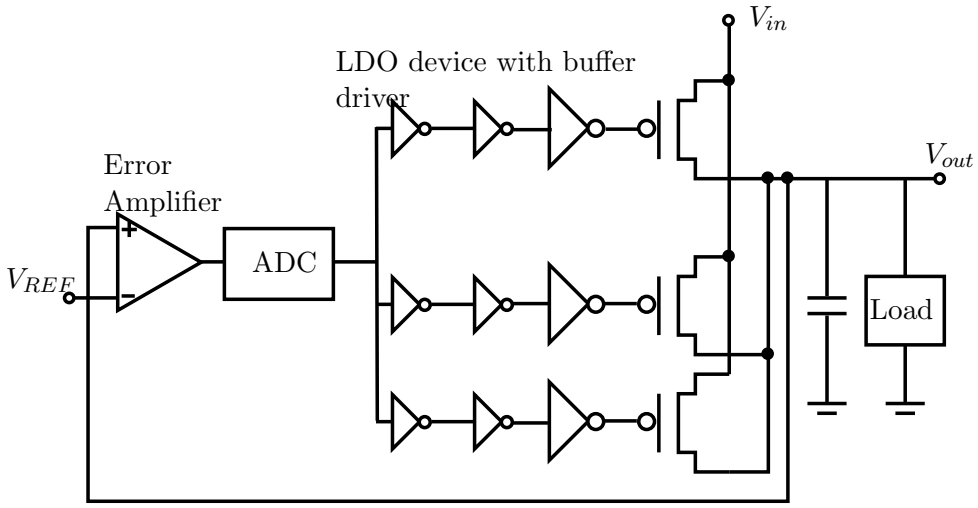
A PG device, which is implemented with many parallel transistors and on/off signal buffers, is similar to the largest component (i.e., the pass device between V_{IN} and V_O) of a typical LDO VR, as illustrated in Figure 6.4-(a). In other words, an LDO VR can be implemented by augmenting a PG device with feedback control circuitry comprised of an error amplifier, an analog-to-digital converter, and a reference voltage generator. In [58], it was reported that the output device and its buffers, both of which can be shared with a PG device, accounted for 83% of the total LDO VR area. Since a PG device consumes 5%∼10% of a core's area [61], we estimate that the extra overhead due to the feedback

control circuitry to implement LDO VR is 0.85%∼1.7% of the core's area. By contrast, on-chip switching VRs require large inductors and capacitors for their implementations. As a result, a switching VR has nearly four times larger chip area than a comparable LDO VR [62]. Furthermore, LDO VRs can provide faster transient responses than switching VRs [63] and, unlike switching VRs, they do not inject switching noise in the substrate. This is desirable for the operation of highly sensitive mixed signal circuits.

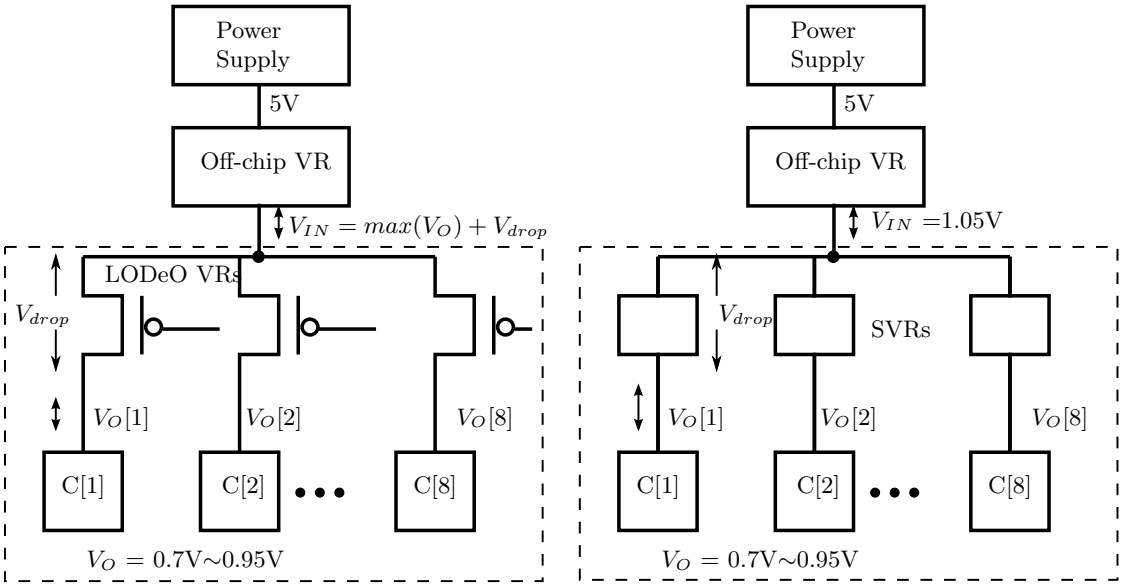
Figure 6.4-(b) shows two different approaches to distribute supply voltages to an 8-core processor with per-core V/F domains. Both approaches use a first stage off-chip VR to convert 5V to an intermediate voltage level, V_{IN} of on-chip per-core VRs. We cannot supply 5V for on-chip switching VRs directly considering the oxide reliability of nanoscale transistors implementing both VRs and cores. This voltage is further down-converted using on-chip per-core VRs to the voltage (V_O[i]) required by core i. The arrangement on the left uses LDO VRs (i.e., PCPG devices augmented with the control and reference circuitry to implement LDO VRs). The efficiency of an LDO VR is a function of its V_O/V_{IN} ratio.

When the voltages demanded by individual cores are restricted to a limited range (e.g., within 100mV of one another as shown in Figure 6.3), a high V_O/V_{IN} ratio can be achieved for all the cores by adjusting the V_O of the first stage (i.e., V_{IN} of the second stage) such that it is sufficient to provide the highest V_O demanded by any of cores. Thus, a processor adopting per-core LDO VRs can be tuned to achieve high efficiency by jointly optimizing both their V_{IN} and V_O. The arrangement on the right uses per-core on-chip switching VRs to provide the necessary core voltage. A switching VR uses two active devices, inductors and capacitors to provide high voltage conversion efficiency across a wide range of V_O. This efficiency is primarily determined by the switching losses in the active devices and the conduction losses in active devices and inductor. The V_{IN} value for switching VRs is fixed to 1.05V in this example.

In summary, an LDO VR can be implemented very cost-effectively since (i) it does not



(a)



(b)

Figure 6.4: (a) A typical LDO VR architecture; the cartoon is reproduced from [58]. (b) An example of V_{IN} and V_O ranges of LDO VRs in the left and switching VRs denoted by SVRs in the right for supporting per-core voltage domains;"C[i]" in (b) denotes core i.

require inductors and (ii) it can share its largest component (i.e., the output device) with a PCPG device. Furthermore, its efficiency can be very high when cores running multiple threads or applications demand similar voltage values. In the next section, we investigate the power efficiencies of the two different VR types.

6.2.3 LDO vs. Switching VR Efficiency Comparison

Figure 6.5 compares the efficiency of a switching VR with that of an LDO VR (the on-chip second stage only in (a) and both the off- and on-chip stages in (b), respectively). The efficiency of LDO VRs is higher than switching VRs when V_{IN} - V_O is small (or V_O/V_{IN} is high), but it becomes lower as $V_{IN} - V_O$ increases (or V_O/V_{IN} decreases). If $V_{IN} - V_O$ is more than 100mV, the efficiency of LDO VRs usually is lower than that of switching VRs as shown in Figure 6.5-(a). We model the efficiencies of both switching and LDO VRs assuming each core consumes the maximum allowed current for each operating voltage. To measure the maximum efficiency of the switching VR at each operating point (i.e.,voltage/current), we search for and activate the optimal number of phases out of eight available phases for a given voltage/current. Table 6.1 summarizes the key design parameters of various VR stages described in this study. The off-chip switching VR efficiency computation is based on [64]

Table 6.1: Summary of VR design parameters

	Off-chip Switching VR	On-chip Switching VR	On-chip LDO VR
V_{IN}/V_O	5V/1.05V to 5V/0.85V	1.05V/0.95V to 1.05V/0.7V	0.95V/0.7V to 0.7V/0.95V
Tech.	Discrete devices	32nm	32nm
f_{sw}	300KHz	100MHz	N/A
L/phase	360nH (r_L=0.5mΩ)	63.5nH (Q=20@100MHz)	N/A
# of phases	6	8	N/A

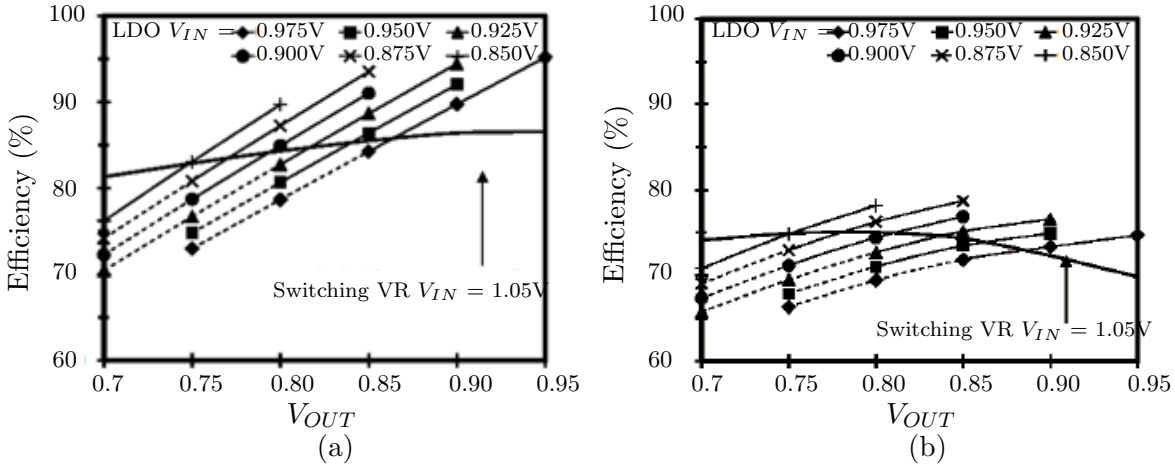


Figure 6.5: Efficiency Comparison between switching and LDO VRs. (a)The on-chip (second stage) only and (b) both the off- and on-chip (first and second stages).

with V_{IN} fixed to 5V. Off-chip switching VR designs built with off-the-shelf components typically have very high efficiencies ($> 90\%$) due to low loss inductors and capacitors. Their efficiency reaches a maximum for a certain load current and then drops with further increase in current due to an increase in conduction losses. Consequently, as the off-chip regulator output for LDO VRs decreases, the efficiency degrades, and thus the overall efficiency of LDO VRs becomes slightly lower than that of switching VRs, as plotted in Figure 6.5-(b). The efficiency of an LDO VR can be calculated as:

$$\eta_{LDO} = \frac{V_O \cdot I_O}{V_O \cdot I_O + (V_{IN} - V_O) \cdot I_O + V_{bias} \cdot I_Q} \tag{6.1}$$

where I_Q is the quiescent current of the LDO and V_{bias} is the biasing voltage for the reference and feedback control circuitry. A steady analog $V_{bias} = 0.9$V generated on chip from the variable V_{IN} is assumed in this work. The current efficiency of a typical LDO VR is defined by:

$$\eta_I = \frac{I_O}{I_O + I_Q} \tag{6.2}$$

η_I is a measure of the power loss in the control and biasing circuitry of the LDO. On-chip LDO designs with current efficiencies in the range of 95% to 99% have been reported. The LDO efficiency is computed assuming a current efficiency of 97% [58] at I_O corresponding to 120W/0.9V.

The efficiency of switching VRs with integrated inductors was modeled for different CMOS technology generations in [65]. The efficiency is a strong function of the inductor Q factor. Inductors in CMOS processes are made from the available metal layers and attain low Q values for realistic dimensions due to the substrate losses and frequency dependent conduction losses. It was shown in [65] that a fully monolithic switching VR achieves ~62% efficiency with on-die inductors in 90nm CMOS, which is not acceptable considering the performance benefit that can be brought by per-core voltage domains under a power constraint. The efficiency can be improved by using alternate inductor technologies with high Q. This may include inductors with magnetic materials compatible with a CMOS process or inductors mounted externally on the package while only the active devices are integrated on die [57]. A switching VR with 80%-87% efficiency with integrated active devices and on-package inductors ($Q = 20$) was demonstrated in [57].

The efficiency analysis presented in this work assumes a 32nm CMOS process with inductors ($Q = 20$ @ 100MHz) similar to [57], since switching VR with on-die inductors exhibit poor efficiency. On-package inductors incurs packaging design and integration issues, but are being adopted in industry recently [66]. The design is optimized to achieve a conversion ratio of 1.05V/0.9V at a load current of 16.67A per core (corresponding to total 120W for 8 cores at 0.9V) with an efficiency of 88%. An 8-phase topology is used with 63.5nH inductance per phase. As V_O and load current are reduced, the efficiency of switching VRs decreases monotonically. This is because the switching loss constitutes a higher percentage of the output power as the V_O value reduces. The efficiency is strongly dependent on the operating point at which the switching VR design is optimized. For a

design optimized for a higher V_O, the efficiency at low output voltage drops more rapidly compared to a design optimized for a lower V_O [65]. In Figure 6.5-(a) the on-chip switching VRs are optimally designed for $V_O = 0.8$V.

6.3 Evaluation of DVFS with LDO VR

6.3.1 DVFS Algorithms for Efficiency Comparison

The key objective of this section is to evaluate the effectiveness of the LDO VRs derived from PCPG devices. Thus, we can use various per-core DVFS algorithms optimized for high-performance multi-core processors including the algorithms exploiting C2C frequency and power variations along with thread migrations (TMs) [50]. For the evaluation, we adopt an integer linear programming (ILP) method for the DVFS algorithms. The ILP formulation is similar to one used in [15], which attempts to minimize the power consumption of a multi-core processor for a given performance constraint. We modify the formulation such that we search for the optimal V_O for each core to maximize MIPS^3/W under a power constraint at each DVFS interval as follows:

Objective:

$$maximize \left(\sum_{i=1}^{N} MIPS_i = \sum_{i=1}^{N} \sum_{j=1}^{M} IPC_i \cdot F_{ij} \cdot x_{ij} \right) \tag{6.3}$$

Constraints:

$$\begin{aligned} \sum_{i=1}^{N} \sum_{j=1}^{M} P_{ij} \cdot x_{ij} &\leq P_{TOT,MAX} \\ \sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij} &\leq N \end{aligned} \tag{6.4}$$

where N is the number of cores; M is the number of V_O steps supported by the DVFS

Table 6.2: Summary of DVFS Algorithms.

Algorithms	Voltage Domain	Frequency Domain	Process Var. Aware	Thread Migration	Off-chip VR V_O	On-chip VR	Constraint
ShV/F	Shared	Shared	No	No	Varying	N/A	$V_{O1} = V_{O2} = ... = V_{ON}$
SeV/F	Separate	Separate	No	No	Fixed	SVR	
SeV/F(PV)			Yes	No			
SeV/F(PV/TM)			Yes	Yes			
LDOSeV/F	Virtually Separate	Separate	No	No	Varying	LDO VR	$V_{IN} - V_O[i] \leq 100$mV
LDOSeV/F(PV)			Yes	No			
LDOSeV/F(PV/TM)			Yes	Yes			

"Sh," "Se," "PV," and "TM" indicate "Shared," "Seperate," "Process Variation," and "Thread Migration" respectively.

algorithm; $MIPS_i$ and IPC_i are the MIPS and IPC of core i; F_{ij} is the frequency of core i at voltage level j; x_{ij} corresponds to one bit of an M-bit binary variable for core i that is guaranteed to assign core i to only one of M possible V/F states (i.e., $\forall i : \sum_{j=1}^{M} x_{ij} = 1$). P_{ij} is the power consumption of core i, which is a function of $V_O[i]$; $P_{TOT,MAX}$ is the allowed total power consumption of the processor; and Eq. (6.4) is the constraint, respectively. In Eq. (6.4), the second constraint is to enforce one V_O selection for each core. The V_{IN} for all LDO VRs is determined by taking the maximum value among $V_O[1]$, $V_O[2]$,..., and $V_O[N]$.

As discussed in [50], this algorithm requires manufacturers to store per-core frequency and power values at each voltage level for DVFS algorithms to exploit C2C frequency and power variations. These values can be characterized by the manufacturer and stored, along with many other processor tuning parameters, in a non-volatile memory of the processor. Like other DVFS algorithms, we also need to predict workload characteristics like the IPC of each thread to assign a proper V/F to each core for the next DVFS interval. Although various methods can be used to predict the IPC of the next interval based on the current IPC, we assume that the IPC value of each thread at every interval is known in advance (as an oracle method). This is to isolate the impact of the IPC prediction from the $MIPS^3/W$

Table 6.3: Processor Simulation Parameters.

Fetch/Issue/Retire	4/4/4	# of Cores	8
IL1	32KB/4-way/64B 3 cycles	Branch Pred./BTB/RAS	YAGS/1K/32
L2	512KB/8-way/64B 10 cycles	DL1	32KB/4-Way/64B 3cycles
Cache Coherency Protocol	Directory-based MESI	Main Memory size/block/page/latency	DDR3-1.6GHz 4GB/64B/4KB/7-7-7-/ 20ns
# of MSHRs	8	Write-buffer entries	16

results so that we can fairly compare the efficacy of the two different VR schemes. Finally, we adopt a simple scheme for the TM technique; we assign threads to cores one-to-one in the order of IPC and frequency values. For example, the thread with the highest IPC is assigned to the core with the highest frequency at a given voltage (i.e., the fastest core considering C2C frequency variations). Table 6.2 summarizes the DVFS algorithms explored in this study and constraints for specific algorithms. Our baseline processor has a single, chip-wide V/F domain using an off-chip VR (i.e., ShV/F).

6.3.2 Architectural Simulation Environment

The processor configuration contains eight cores. Each core is four wide with 32KB private L1 cache and a shared 512KB L2 cache. The cores are connected to each other using crossbar switches. We evaluate different DVFS algorithms using a full-system cycle-accurate simulator, GEMS [53], which is modified to support per-core frequency domains and TM requiring L1 cache flushing. In addition to four commercial workloads (Apache, JBB, OLTP, and Zeus)[51], six SPEC OMP V3.2 benchmarks (ammp, applu, art, equake, mgrid, and swim), and four PARSEC benchmarks (Swaptions, X264, Fluid, and Black Scholes)[52], five mixes of compute- and memory-bound SPEC2006 benchmarks (eight copies of Bzip2, six copies of Bzip2 and two copies of Libquantum, four copies of Bzip2 and four copies

of Libquantum, two copies of Bzip2 and six copies of Libquantum, and eight copies of Libquantum denoted by 8B0L, 6B2L, 4B4L, 2B6L, and 0B8L,respectively) are used. The processor simulation parameters are summarized in Table 6.3.

6.3.3 Core Frequency and Power Modeling

Typically, an operating system (OS) determines V/F of cores based on a given power management algorithm, but both the OS and VRs cannot track and respond to instantaneous changes of power consumption. Thus, the OS must conservatively assume the power consumption of cores at each given operating V/F and guarantee that the entire chip does not exceed its maximum power consumption, if it aims to optimize performance without violating a power constraint at any given moment. To model the maximum power consumption of cores, we assume that (i) the total maximum power consumption of 8 cores is 120W and (ii) 30% of the total power is active leakage at 0.9V. Each core has its own shared L2 cache that shares the V/F domain with the core. Thus, we assume that the L2 power scales with the core power consumption. The power consumption of I/O and other peripheral components including on-chip interconnects, which are tied to other separate fixed voltage/frequency domains, is not included in the analysis since it can be regarded as a fixed power cost for all the cases explored in this work; I/O and on-chip interconnects are responsible for 15% of the total power in Niagara 2 [67].

Due to WID C2C frequency and leakage power variations, the power consumption of each core differs. To analyze the impact of WID process variations on the frequency and leakage power consumption of each core, we first generate 100 variation maps for threshold voltage (V_{TH}) and effective channel length (L_{EFF}) of transistors in a die and characterize frequency and power consumption by following the methodology presented in [21] and described in Chapter 3 : WID correlation distance coefficient $\Phi = 0.5$ and WID V_{TH} and L_{EFF} variations $\sigma_{sys} = 6.4\%$ and 3.2% of the nominal V_{TH} and L_{EFF} values respectively. We apply the

V_{TH} and L_{EFF} values of each grid point to a FO4 inverter chain and a dummy circuit, which is comprised of 50% inverters, 30% NAND gates, and 20% NOR gates, to obtain the frequency and leakage power scaling factors of each core, respectively; NAND and NOR gates in a dummy circuit can have up to 4 inputs and their inputs are assigned randomly with either 1 or 0.

The frequency and leakage scaling factors of each grid point are measured over a voltage range of 0.95V to 0.7V using a 32nm technology model [27] [28] and SPICE. We assume that the frequency of each core is determined by the slowest grid point in the core [21] and the frequency of the slowest core is 3.2GHz at 0.9V. Each core's maximum dynamic power consumption at 0.9V is $\left(F_i / \sum_{j=1}^{N} F_j\right) \times 120W \times 0.7$ where F_i and F_j are the frequency of core i and j, and N is the number of cores. With the known frequency, voltage, and dynamic power values, the maximum core switching capacitance (i.e., C_{dyn}) can be obtained. This allows us to calculate the dynamic power at any given voltage. The leakage power of each grid point is scaled such that the sum of the leakage power from all grid points in a die is equal to 30% of 120W at 0.9V. The sum of the scaled leakage power from all the grid points belonging to a particular core gives the core's leakage power.

Finally, in our experiments, some cores are allowed to run at V/F higher than 0.9V/3.2GHz as long as the total power constraint is satisfied; this is possible when other cores run at V/F lower than 0.9V/3.2GHz. Since all cores in our baseline processor, which uses a per-chip single VR, run at the same frequency, the dynamic power consumption of the processor is lower than when other processors use per-core V/F domains. Thus, we increase the V/F of the processor until 120W is fully used (i.e., 0.9125V and 3.3GHz). The C2C frequency and power variations change across different dies. However, for analyses presented in this work, a typical die map is chosen from the 100 generated maps. Thus, the $MIPS^3/W$ results, which exploit C2C frequency and leakage power variations, represent the value close to the median value of the 100 die maps. Table 6.4 tabulates the frequency and power consumption

Table 6.4: Frequency and Power consumption of each core as function of $V_O[i]$.

$V_O[i]$	Core1		Core 2		Core 3		Core 4		Core 5		Core 6		Core 7		Core 8	
0.95V	3.6	15.9	3.8	17.9	4.4	18.2	3.9	16.9	3.8	17.2	4.1	19.3	4.4	29.1	4.1	19.7
0.90V	3.2	12.5	3.4	14.0	4.0	14.5	3.5	13.3	3.4	13.5	3.7	15.1	4.0	22.0	3.7	15.3
0.85V	2.8	9.6	3.0	10.8	3.5	11.3	3.0	10.2	3.0	10.4	3.3	11.7	3.5	16.6	3.3	11.8
0.80V	2.4	7.2	2.5	8.1	3.1	8.6	2.6	7.7	2.4	7.8	2.8	8.8	3.1	12.3	2.8	8.9
0.75V	2.0	5.3	2.1	6.0	2.6	6.4	2.2	5.7	2.1	5.7	2.3	6.5	2.6	9.0	2.4	6.6
0.70V	1.6	3.7	1.7	4.2	2.1	4.6	1.7	4.0	1.7	4.1	1.9	4.6	2.1	6.4	1.9	4.7

of each core a s function of $V_O[i]$. For each core the frequency (GHz) and power (Watts) are given in the left and right columns, respectively.

6.3.4 MIPS^3/W Comparison

Impact of Limiting V_{IN}-V_O range

Figure 6.6 compares MIPS^3/W of 8-core processors using LDO and switching VRs. The DVFS algorithms begining with "LDOSeV/F" use LDO VRs while ones with "SeV/F" use switching VRs. Although WID C2C process variations are not exploited and the TM technique is not applied, the MIPS^3/W difference between LDOSeV/F and SeV/F is around 2% (3% versus 5% improvement over ShV/F) on average (i.e.,geometric mean). However, when WID C2C process variations are exploited, the MIPS^3/W difference between the processors using LDO and switching VRs becomes 1% (14% versus 15% improvement over ShV/F) on average. Finally, the MIPS^3/W difference between the two schemes leads to a 3% difference (22% versus 25% improvement over ShV/F) on average when both the TM technique is applied and WID C2C process variations are incorporated with the DVFS algorithms. Finally, exploiting C2C frequency/power variations and TMs can mitigate the potential limitation of LDO VRs and its relative benefit is higher for processors using LDO VRs.

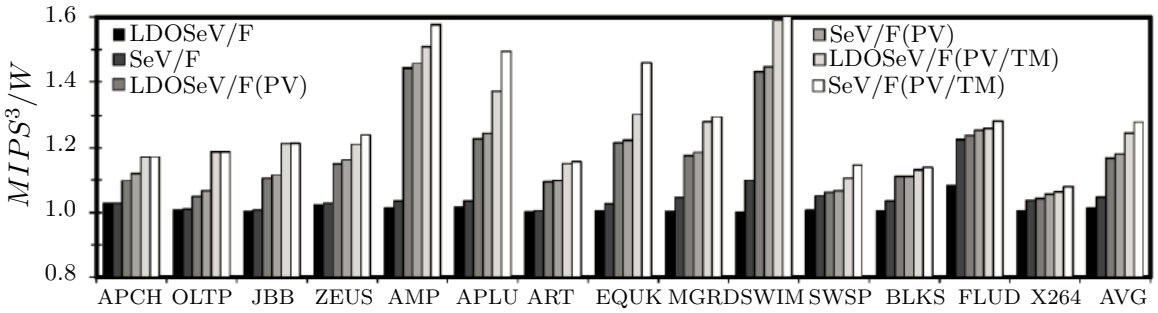


Figure 6.6: MIPS^3/W comparison of 8-core processors supported by LDO (algorithms beginning with the LDOSeV/F prefix) and switching VRs (algorithms beginning with SeV/F). All results are normalized to a processor with ShV/F and do not include the power loss by the VRs. Each interval is comprised of 10-million executed instructions.

Impact of VR Efficiency on MIPS^3/W

MIPS^3/W result shown in Figure 6.6 does not include the power consumption (i.e., power loss) by both on- and off-chip VRs to see the impact of constraining the V_O range for LDO VRs, where it is observed that the MIPS^3/W difference between processors using LDO and switching VRs is very small. However, since C2C voltage differences are limited to 100mV at each DVFS interval, the differences between the V_{IN} value and the V_O values must be small. In other words, LDO VRs must exhibit higher efficiency than switching VRs for most DVFS intervals, as shown in Figure 6.5-(b). Consequently, as the efficiencies of both on- and off-chip VRs are considered, a processor using LDO VRs can provide higher MIPS^3/W than a processor using switching VRs as shown in Figure 6.7. First, the processors using LDOSeV/F and SeV/F, which do not exploit WID process variations, result in worse MIPS^3/W than the processor using ShV/F, which uses only an off-chip VR. This is because the power loss by the on-chip VRs completely negates the benefit of supporting per-core V/F domains for multi-threaded applications. Second, when WID process variations are exploited, VSeV/F(PV) and SeV/F(PV) can provide 6% and 4% higher MIPS^3/W than ShV/F on average. The processor using LDO VRs exhibits higher MIPS^3/W than the one using switching VR. This is the opposite of the trend shown in

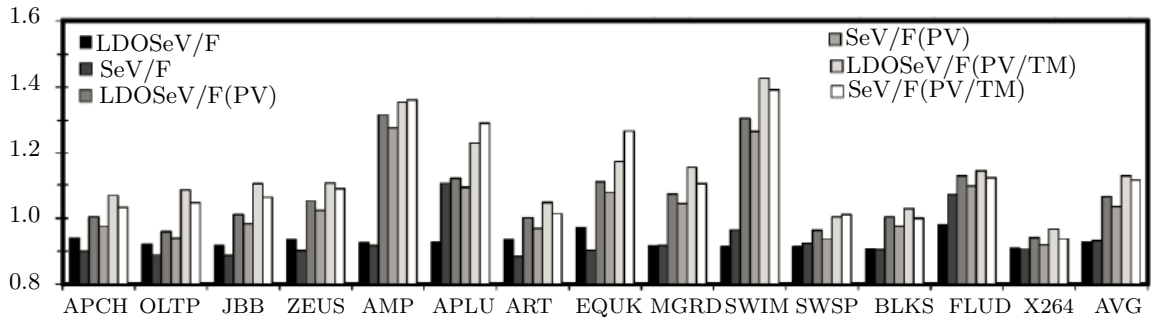


Figure 6.7: MIPS^3/W comparison of 8-core processors supported by LDO (algorithms beginning with the LDOSeV/F prefix) and switching VRs (algorithms beginning with SeV/F) including the power loss by both on- and off-chip VRs. All results are normalized to a processor with ShV/F and include the power loss by the off-chip VR. Each interval is comprised of 10-million executed instructions.

Fig 6.6 where the power loss by VRs was not considered in computing MIPS^3/W and the MIPS^3/W of the processor using LDO VRs was lower than one using switching VRs. This is mainly due to small C2C voltage variations in multi-threaded applications, which allows LDO VRs to provide voltages with higher efficiency than switching VRs as shown in Figure 6.5-(b). Finally, when the TM technique is also applied, LDOSeV/F(PV/TM) and SeV/F(PV/TM) can provide 13% and 12% higher MIPS^3/W than ShV/F on average.

Impact of DVFS interval on MIPS^3/W

The interval period for applying DVFS algorithms also impacts the benefit of DVFS. In theory, shorter DVFS intervals can capture more C2C IPC variations and thus lead to higher performance/power efficiency. Thus, we reduce the DVFS interval to every 5-million instructions while keeping the TM interval at 10-million instructions. As expected, MIPS^3/W for both LDOSeV/F(PV/TM) and SeV/F(PV/TM) increases, but the relative difference between them remains almost the same. The DVFS interval is often determined by considering both (i) the computational overhead of the DVFS algorithm and (ii) the VR efficiency degradation during V changing periods [15]; (ii) prohibits a very short interval for

a simple threshold-based DVFS algorithm even though the current state-of-the-art off-chip switching VRs can support much faster voltages changes. For example, Microsoft Windows Vista uses 20ms for the default value while it could support a more aggressive interval value (e.g., 1ms) for DVFS [68].

Multi-program Environment

A processor executing multiple applications can exhibit more substantial C2C IPC variations than when it is running multi-threaded applications, depending on the mix and characteristics of applications. Consequently, supporting a wider range of V_O values using switching VRs may lead to higher MIPS^3/W than using LDO VRs under a specified power constraint. Figure 6.8 shows the MIPS^3/W comparison between two processors using switching and LDO VRs when running five mixes of memory- and compute-bound applications; the mixes of applications were run using the multi-core simulator (not in isolation) to accurately model the interaction between applications as well.

First, when the power loss by VRs is not considered, a processor with per-core voltage domains using either switching or LDO VRs has much higher MIPS^3/W than one using a single chip-wide voltage domain for 6B2L, 4B4L,and 2B6L in Figure 6.8-(a). This is due to these applications mixes having much higher C2C IPC variations than the multi-threaded applications. For example, LDOSeV/F(PV/TM) and SeV/F(PV/TM) can provide 34% and 55% higher MIPS^3/W than ShV/F on average. The processor using LDO VRs provides substantially higher MIPS^3/W than one using a single chip-wide VR, but 16% lower MIPS^3/W than one using switching VRs. Second, when the power loss by the VRs is considered, as shown in Figure 6.8-(b), LDOSeV/F(PV/TM) and SeV/F(PV/TM) can yield 24% and 39% higher MIPS^3/W than ShV/F on average. Unlike the multi-thread applications, LDOSeV/F(PV/TM) results in lower MIPS^3/W than SeV/F(PV/TM), yet the difference between LDOSeV/F(PV/TM) and SeV/F(PV/TM) is reduced to 12%. This is because the

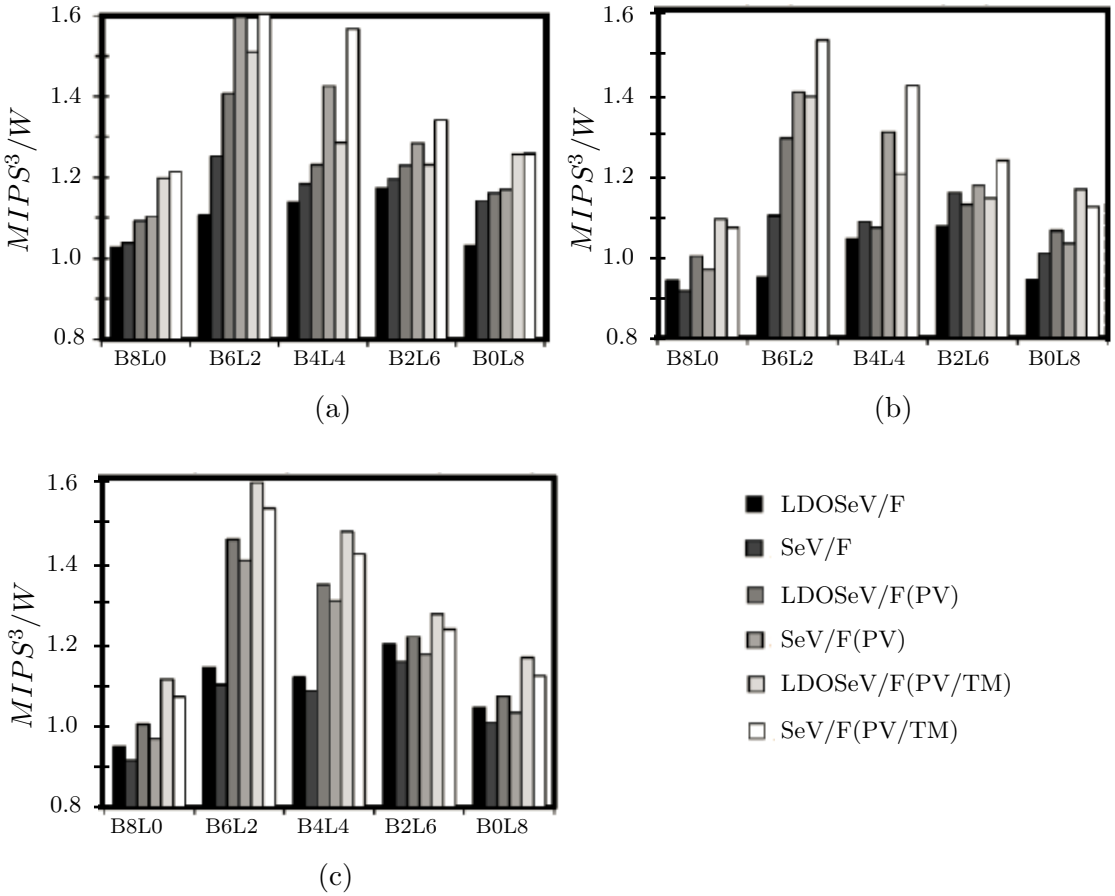


Figure 6.8: MIPS^3/W comparison of 8-core processors supported by LDO (algorithms beginning with the LDOSeV/F prefix) and switching VRs (algorithms beginning with SeV/F). The power loss of VRs is not included in (a) and is included in (b). The V_{IN} - V_O constraint is removed in (c) for LDO VRs. Each interval is comprised of 10-million executed instructions.

power loss by LoDeO VRs is still lower than switching VRs in many DVFS intervals.

In the previous experiments, the difference V_{IN} - V_O was limited to 100mV. This is because the efficiency of LDO VRs becomes lower than switching VRs once the voltage difference becomes larger than 100mV. On the other hand, forcing such a constraint misses potential power reduction opportunities that can be achieved by operating cores at lower V/F. In other words, the benefit of reducing V/F of cores more can outweigh lower power efficiency of LDO VRs operating at V_{IN} - V_O larger than 100mV. Thus, to evaluate the impact of such a constraint, we remove the V_{IN} - V_O constraint for LDO VRs in Figure 6.8-(c). When V_{IN} - V_O is larger than 100mV, the power loss by LDO VRs is higher than that of switching VR. However, the power loss of LDO VRs becomes lower than that of switching VRs for the DVFS intervals exhibiting V_{IN} - V_O less than 100mV. Consequently, as long as there exist more DVFS intervals with V_{IN} - V_O less than 100mV, the processor using LDO VRs can lead to higher MIPS^3/W than the one using switching VRs. Figure 6.8-(c) shows that LDOSeV/F(PV/TM) has 4% higher MIPS^3/W than SeV/F(PV/TM). To validate this result, we analyzed the fraction of DVFS intervals exhibiting V_{IN} - V_O more than 100mV. For B4L4, we measured the fraction of DVFS intervals in which LDO VRs have lower efficiency than switching VRs after the V/F and core power consumption profiles obtained from SeV/F(PV/TM) are applied to both efficiency functions of LDO and switching VRs. This reveals that LDO VRs show higher efficiency than switching VRs for close to 60% of the total DVFS intervals that are experienced by individual cores.

We re-evaluate the MIPS^3/W for multi-treaded workloads in Figure 6.9 after the voltage range constraint is removed. Although WID C2C process variations are not exploited and the TM technique is not applied, LDOSeV/F provides 4% higher MIPS^3/W than SeV/F on average. When both WID C2C process variations and TM are exploited, LDOSeV/F and SeV/F provide 16% and 12% higher MIPS^3/W than ShV/F on average; LDOSeV/F leads to 4% higher MIPS^3/W than SeV/F whether or not WID C2C process variations

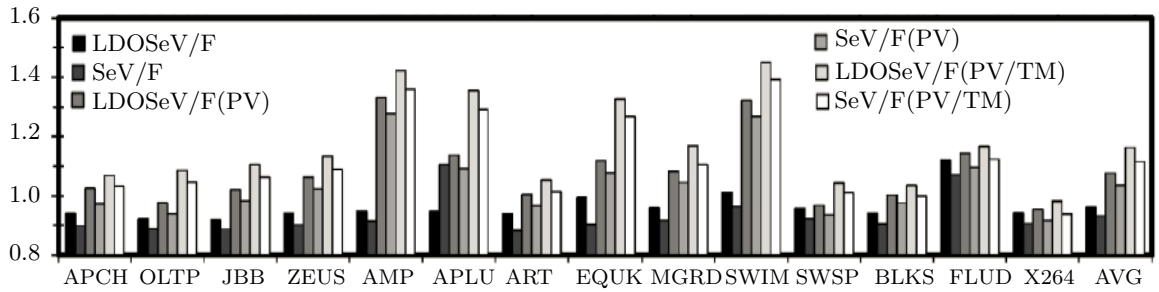


Figure 6.9: MIPS^3/W comparison of 8-core processors supported by LDO (algorithms beginning with the LDOSeV/F prefix) and switching VRs (algorithms beginning with SeV/F) including the power loss from both on- and off-chip VRs and without limiting V_{IN} - V_O. All results are normalized to a processor with ShV/F and include the power loss by the off-chip VR. Each interval is comprised of 10-million executed instructions.

and/or TM are exploited. This is because LDO VRs exhibit higher efficiency than switching VRs for most DVFS intervals. This is mainly due to small C2C voltage variations in multithreaded applications, which allows LDO VRs to provide voltages with higher efficiency than switching VRs.

6.4 Related Work

Several prior studies have investigated the benefits of DVFS for multi-core processor. Li *et al.* analyzed the performance of DVFS combined with dynamic core scaling for parallel workloads on multi-core processors[69]. They exploit the observation that parallel workloads with limited problem size do not use all cores efficiently. Thus, they jointly adjust the number of active cores along with performing DVFS to maximize the performance under a power constraint. Kim *et al.* [15] demonstrated the potential benefit of per-core DVFS using on-chip switching VRs for embedded processors, and provided detailed background on switching VRs using air-core inductors and an analysis of their efficiency. Recently, S. Eyerman *et al.* also evaluated the benefit of fine-grain applications of DVFS and proposed a fine-grain microarchitecture-driven DVFS mechanism [70]. Many researchers also studied

the impact of WID process variations, which lead to C2C frequency and power variations [71] [20], on the performance of multi-core processors, and proposed a DVFS algorithm that can exploit the C2C frequency and power variations. Teodorescu *et al.* investigated a DVFS algorithm based on linear programming to maximize the performance of multi-core processors under a power constraint [50]. Their DVFS algorithm also exploits WID C2C frequency and power variations for workload scheduling, as well as power management. Rangan *et al.* proposed a thread migration technique to minimize the cost of the transition time for the VR output voltage [72]. They introduced voltage domains in which each core operates at a fixed but different voltage level. If threads require different V/F levels for power-efficient operations, they migrate to the cores that can provide an appropriate performance level, instead of changing the voltage/frequency of cores.

6.5 Chapter Summary

Splitting voltage domains for per-core V/F control increases platform cost and validation complexity. Integrated switching VRs are hindered by the unavailability of high quality monolithic inductors. In this chapter, we demonstrated that PCPG devices augmented with small circuitry can operate as low-cost LDO VRs. Unlike on-chip switching VRs, LDO VRs do not require on-chip inductors and do not inject switching noise in the substrate. The efficiency of LDO VRs reduces as their output voltage applied to cores drops (i.e., large difference between input and output voltage of the LDO VRs). Consequently, per-core DVFS using LDO VRs may lead to lower performance/power efficiency than using switching VRs. However, experiments in [4] show that C2C voltages variations are relatively small when the voltages are optimized to maximize performance under a power constraint. After modeling the power efficiency of both LDO and switching VRs using a 32nm technology, we show that the $MIPS^3/W$ of an 8-core processor using LDO VRs is slightly higher than that of a processor using switching VRs. This is because the efficiency of LDO VRs is higher

than that of switching VRs for small C2C voltage variations in each DVFS interval, which was observed through the experiments.

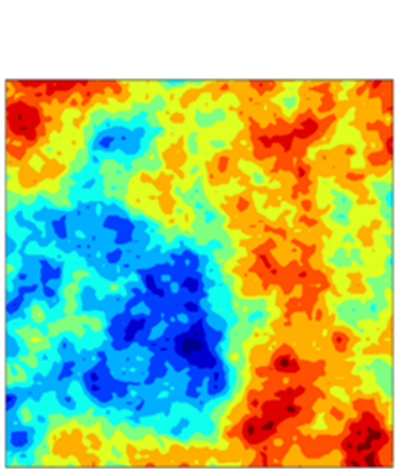
Future processors will have more cores (i.e., many-core processors) and will potentially exhibit higher C2C IPC variations depending on the mix of single and multi-threaded applications. This will result in wider V_{IN}-V_O range and thus worse MIPS^3/W due to poor power conversion efficiency for LDO VRs. On the other hand, it will be impractical to provide a large number of switching VRs for many-core processors due to the cost; integrating high quality on-chip inductor becomes more challenging with technology scaling while integrating on-package inductors will not be a scalable solution for a larger number of cores due to the physical constraint. In such a case, a hierarchical VR scheme where a switching VR provides a shared voltage domain for a sub-set (or a cluster) of cores and LDO VRs provide per-core voltage domains within each cluster may be used. Such a hierarchical power delivery architecture can support cost-effective per-core voltage domains, which can minimize both the number of switching VRs while maximizing the power efficiency of individual LDO VRs.

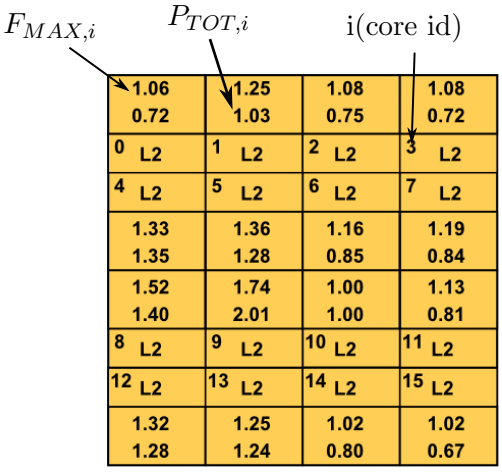
Chapter 7

Throughput Optimization under Power and Thermal Constraints

Core-to-core (C2C) variations in frequency and leakage power reduce the power efficiency of processors using a single chip-wide voltage/frequency (V/F) domain. This is because the maximum operating frequency of the die, (F_{MAX}), is decided by the slowest core while fast cores consume considerably more leakage power (P_{LEAK}) at the unnecessary high voltage. To maximize the power efficiency of processors in the presence of C2C F_{MAX} and P_{LEAK} variations, the effectiveness of per core V/F domains has been explored [15][20][73]. Individual V/F domains achieve high performance/Watt by operating each core or group of cores with optimal voltage and frequency depending on the workload demand.

Chapter 6 presented a cost-effective way of implementing per-core V/F domains in a multi-core processor. In this chapter, we show how a post-silicon F_{MAX} tuning process can be applied to multi-core processors with per-core V/F domains and power/thermal constraint. Fast cores often have higher P_{LEAK} than slow cores due to their shorter channel length (L_{EFF}) and lower threshold voltage (V_{TH}). This can result in considerable C2C temperature (and thus performance) variations under a thermal constraint as cores with





(a) (b)

Figure 7.1: (a) Systematic L_{EFF} and V_{TH} variations across a die. (b) Normalized F_{MAX} and P_{TOT} of each core in a 16-core processor. The blue region in (a) has shorter L_{EFF} and lower V_{TH} resulting in faster and leakier cores.

higher P_{LEAK} create local hotspots and experience more frequent thermal throttling. The proposed optimization method can balance the power consumption and hotspot temperature of fast and slow cores which in turn leads to less frequent thermal throttling of fast cores and thus higher maximum performance under power and thermal constraints.

7.1 C2C Frequency, Power, and Temperature Variations

With technology scaling, spatially correlated within-die (WID) variations in transistor L_{EFF} and V_{TH} manifest as C2C F_{MAX} and P_{LEAK} variations in multi- and many-core processors [21]. For the results described in this chapter, a multi-core processor die with WID variations is generated with a 100×100 grid where each grid point is assigned a distinct V_{TH} and L_{EFF} combination (detailed methodology for generating variation map is described in Chapter 3). Figure 7.1-(a) shows systematic, correlated L_{EFF} and V_{TH} variations across a die in a 32nm technology. An F_{MAX} and P_{LEAK} combination is obtained for each grid point by simulating a 16-stage FO4 chain and a dummy circuit with the associated V_{TH} and

L_{EFF}. The F_{MAX} of a core is decided by the slowest grid point in the core, while P_{LEAK} is obtained from the summation of the leakage current from all the grid points in the core.

As shown in Figure 7.1-(b) certain cores are up to 74% faster (and consume 100% more P_{TOT}) than the slowest core in a 16-core processor die sample. [20] reported the fastest core in an Intel®'s 80-core processor to be 28% faster than the slowest core in a 65nm technology. As the number of cores per die increases with technology scaling, the relative F_{MAX} and P_{TOT} variation among the cores changes more notably. When all the cores in a processor operate at the same F_{MAX} based on the F_{MAX} of the slowest core (i.e., chip-wide clocking), they consume the same dynamic power (P_{DYN}) for the same activity. However, cores which can potentially operate at faster F_{MAX} consume more P_{LEAK} than the slowest core due to their shorter L_{EFF} and lower V_{TH}. This can result in considerable C2C temperature (and thus performance variations) under a thermal constraint. This is because the cores consuming more P_{LEAK} (and thus experiencing higher hotspot temperature) can experience more thermal throttling, which usually reduces the F_{MAX} of the core until the temperature is reduced. In other words, frequent thermal throttling of cores reduces the average F_{MAX} (i.e., performance) of the cores over time. Finally, the faster cores in a processor adopting per-core clocking can experience even more frequent thermal throttling due to higher P_{DYN} due to higher F_{MAX}. Figure 7.2 shows the C2C temperature variations running 16 instances of gcc, the hottest of the SPEC CPU2000 benchmarks, on the die shown in Figure 7.1. It is assumed here that each core runs at its own F_{MAX} (per-core clocking) and the die temperature profile was obtained using Hotspot [29], the gcc power trace, and the core floorplan similar to an Alpha 21264 processor that are included in the Hotspot package. Initially, the gcc dynamic power trace in the Hotspot package is generated by running Wattch [74] with an architecture configuration similar to Alpha 21264. The P_{DYN} numbers in the power trace are scaled such that the peak power consumption in the trace for a single core corresponds to 70% of 5.94W (i.e., 70% of per-

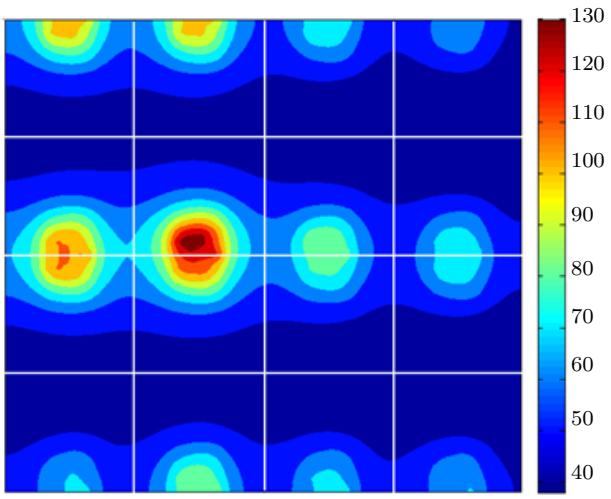


Figure 7.2: Temperature profile of a per-core clocking 16-core processor running 16 gcc instances. The hotspot temperature of core 4, 5, 8, and 9 is well over 100 °C. The low temperature area is occupied with L2 caches.

core power budget) assuming that the total power budget of a processor with the nominal process parameters is 95W. The remaining 30% of 5.94W is P_{LEAK} of the core. The Hotspot configuration parameters are adjusted such that the hotspots of all cores with the nominal process parameters reach 100 °C; all cores run at the same F_{MAX} and consume the same P_{TOT} for the initial setup. In Figure 7.2, faster cores are much hotter due to their higher F_{MAX} (and thus P_{DYN}) and more P_{LEAK}. For example, the fastest core that is 74% faster shows ∼60 °C higher temperature than the slowest core. This can potentially incur more frequent thermal throttling for the fast cores since they consume more P_{TOT}, and results in slower average F_{MAX} than their peak F_{MAX} over long runtime periods.

7.2 Throughput model for Per-core Clocking

In this section, the impact of F_{MAX} and P_{LEAK} variations on throughput of many-core processors under power and thermal constraints is evaluated. The throughput model from

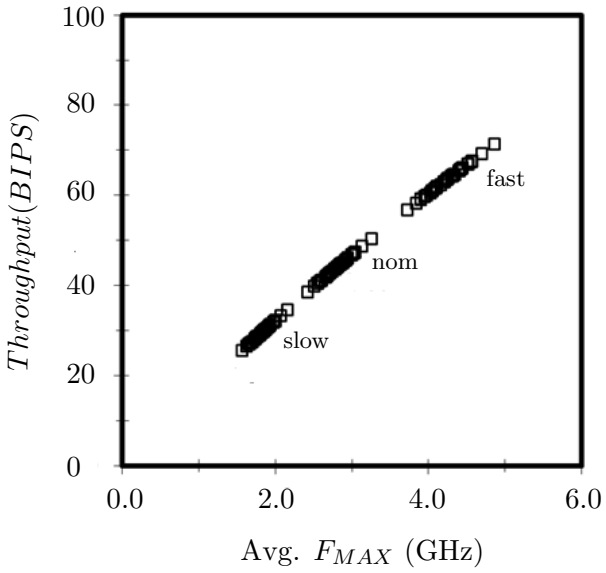


Figure 7.3: Throughput versus average F_{MAX} values of 16-core processors.An average F_{MAX} value of a processor is the arithmetic mean of individual core F_{MAX} values. Each point represents a die sample.

[22][75], used for the analysis is as follows

$$TP_i = IPC_i \times F_{MAX,i} = \frac{1}{\frac{CPI_{com}}{F_{MAX,i}} + M_{rate}(S_{L2(1)}) \times t_{miss}} \tag{7.1}$$

$$t_{miss} = \frac{L_{mem}}{N_{pr}} + L_s \times \left(1 + \frac{U}{2(1-U)}\right) \tag{7.2}$$

$$\sum_{i=0}^{N-1} TP_i = \frac{U}{M_{rate}(S_{L2(1)}) \times L_s} \tag{7.3}$$

where TP_i and IPC_i are the throughput and instructions per cycle of given core i; CPI_{com} is the core CPI with a perfect L2 cache; $S_{L2(1)}$ is the size of L2 cache per core; $M_{rate}(S_{L2(1)})$ is the number of misses per instruction for a cache size of $S_{L2(1)}$, which can be estimated using $M_{rate}(1MB)/\sqrt{S_{L2(1)}/S_{1MB}}$; L_{miss}, L_{mem}, and L_s are the number of cycles for handling L_2 cache miss, fetching data from the DRAM array, and service latencies per L2 cache miss relative to $F_{MAX,i}$; N_{pr} is the number of parallel memory requests that can be serviced per

Table 7.1: Per-core Clocking Throughput Model Parameters

N_{pr}	1	CPI_{com}	0.92
$S_{L2(1)}$	512KB	$M_{rate(1MB}$	1.8×10^{-3}
L_{mem}	100ns$\times F_{MAX,i}$	L_s	48GB/s$\times 128B \times F_{MAX,i}$

core; and U is the number of memory requests per cycle. Since U and IPC are dependent on each other, IPC in Eq. (7.1) reduces to a quadratic equation, where its roots will produce an explicit IPC expression. In this analysis, a medium-size in-order core was assumed with N_{pr}, $S_{L2(1)}$, $M_{rate}(1MB)$, L_{mem}, L_s, and CPI_{com} values that are tabulated in Table 7.1 for Eq. (7.1) - (7.3); all parameters except for L_{mem} and L_s are from [75]; L_{mem} and L_s are from DDR3-1600; the main memory bandwidth is from a Sun Microsystems' 16-core Rock processor running at 2.6GHz [76]; and $F_{MAX,i}$ of a die is obtained using the methodology described in Section 7.1.

Figure 7.3 plots the throughput versus the average F_{MAX} values of 16-core processors. An average F_{MAX} value of a processor is the arithmetic mean of core F_{MAX} values in the processor. 112 die samples from three different process corners (i.e., slow, nominal, and fast) are explored after WID L_{eff}/V_{th} variations are applied to each corner. The result from the plot shows that the overall throughput of a many-core processor is proportional to its average F_{MAX} value of cores, when each core is allowed to run at its own F_{MAX}. This implies that maximizing the average F_{MAX} of cores in a many-core processor results in the maximum throughput as long as the power and thermal constraints are satisfied.

7.3 Maximizing Performance Under Power Constraint

In this section, the problem of maximizing the throughput of a multi-core processor with per-core V/F domains under a maximum total power constraint is analyzed. Technological advances in on-chip voltage regulators indicate the feasibility of such per-core V/F domains in near future. The F_{MAX} of each core can be different due to WID variations although the voltage of all the cores is initially set to the same value. In other words, some cores are faster and consume more power than others, and vice-versa. Reducing the voltage of the fast cores often decreases their P_{LEAK} dramatically while it does not decrease F_{MAX} significantly as illustrated in Figure 7.4. P_{LEAK} of circuits with shorter L_{EFF} and lower V_{TH}, which enable higher F_{MAX} for the faster cores, respond to a voltage change more sensitively due to the stronger drain induced barrier lowering (DIBL) effect. On the other hand, increasing the voltage of slow cores leads to higher F_{MAX} for them while it does not increase P_{LEAK} as much as faster cores do as shown in Figure 7.4. As shown in Section 7.2, the throughput of a many-core processor supporting per-core clocking is proportional to the average F_{MAX} value of all the cores in the processor. When the power budget for a processor is fixed, optimizing the voltage of each core (i.e., decreasing voltage for fast cores while increasing voltage for slow cores) can lead to a higher average F_{MAX} value (and thus higher throughput) for the processor.

To maximize the throughput of a many-core processor for a given power constraint, the optimization problem consists of searching the most performance/power efficient operating point for each core for a given power constraint as follows:

Objective:

$$Maximize \left(\sum_{i=0}^{N-1} F_{MAX,i}(V_{DD,i}) \right) \tag{7.4}$$

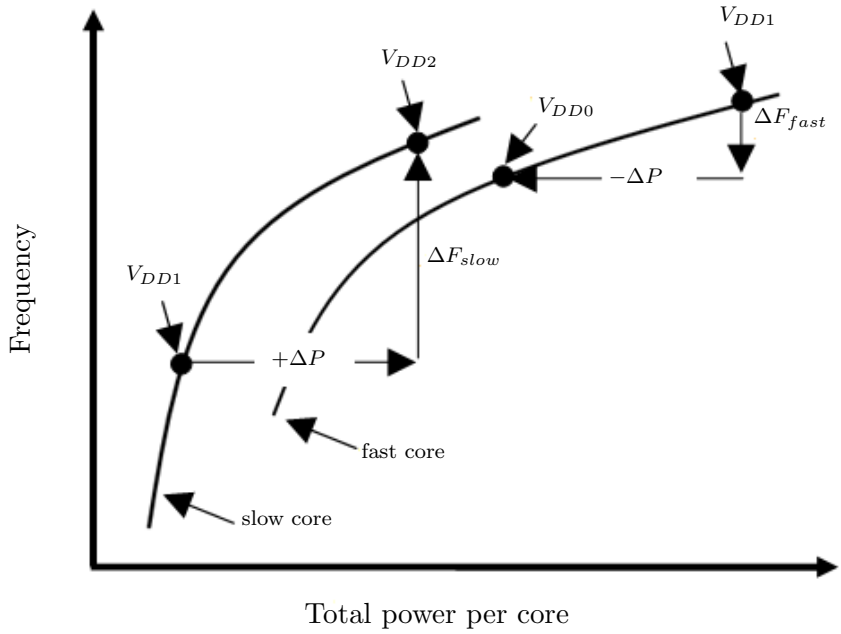


Figure 7.4: F_{MAX} versus P_{TOT} of fast and slow cores in a die.($V_{DD0} < V_{DD1} < V_{DD2}$).

Constraints:

$$\sum_{i=0}^{N-1} P_{TOT,i}(V_{DD,i}, F_{MAX,i}) \leq P_{TDP}$$
$$\forall i : V_{DD,MIN} \leq V_{DD,i} \leq V_{DD,MAX} \tag{7.5}$$

where $P_{TOT,i}$ is the total power consumption of core i as a function of its voltage, $V_{DD,i}$ and $F_{MAX,i}$ are voltage and frequency of core i, modeled with the methodology illustrated in [60]; P_{TDP} is the maximum P_{TOT} constraint which is the thermal design power ; $V_{DD,MIN}$ is the minimum voltage limited by on-chip memory element failures; and $V_{DD,MAX}$ is the maximum voltage limited by the reliability of transistors' gate-oxide. The optimization algorithm is implemented using the characterized $F_{MAX,i}$ and $P_{TOT,i}$ for each core in a die and the Matlab implementation of the active-set non-linear optimization algorithm. The $F_{MAX,i}$ and $P_{TOT,i}$ of each core are characterized as a function of $V_{DD,i}$ with step value =

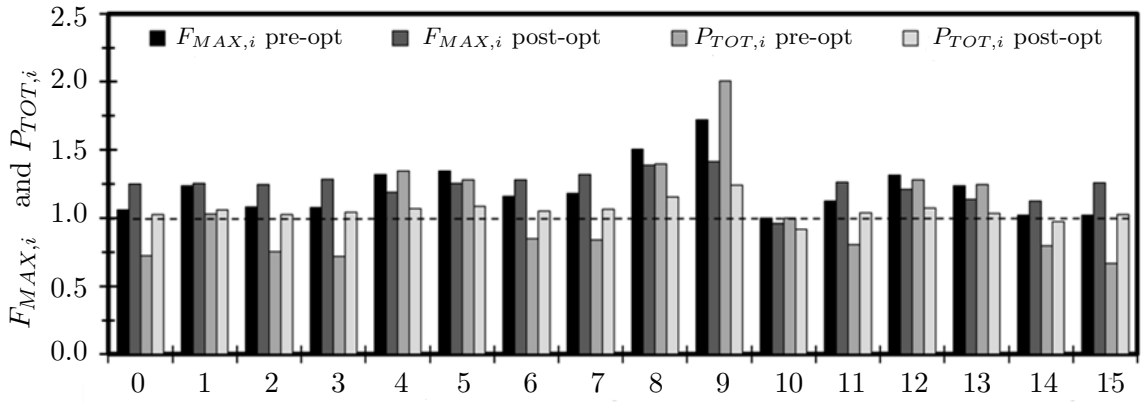


Figure 7.5: $F_{MAX,i}$ and $P_{TOT,i}$ of 16-core die sample before and after the optimization.

6.25mV. Further, the optimization algorithm is applied to each processor die using $V_{DD,i}$ step value = 6.25mV, which is the minimum resolution for voltage scaling using state-of-the-art voltage regulator for commercial processors [77].

Figure 7.5 shows the F_{MAX} and P_{TOT} of each core ($F_{MAX,i}$ and $P_{TOT,i}$) in a 16-core processor die before and after the optimization is applied to the die sample used for generating Figure 7.1. The F_{MAX} and P_{TOT} of each core are normalized to those of the slowest core in the die before the optimization (i.e., core 10 in Figure 7.1-(b)). After the optimization, the F_{MAX} and P_{TOT} of core 0, 2, 3, 6, 7, 11, 14, and 15, which are relatively slower and consume less power than the other cores before the optimization, are increased. Meanwhile, those of core 4, 5, 8, 9, 12, and 13, which are relatively faster and consume more power, are decreased after the optimization. The F_{MAX} and P_{TOT} of core 1 and 10 are barely changed. Overall, for this particular die sample, the average F_{MAX} is improved by 2.3% while the P_{TOT} remains nearly constant. The F_{MAX} improvement is relatively small for the die sample considered. However, the proposed optimization technique balances power consumption (and thus temperature) between fast and slow cores. This can lead to even higher average F_{MAX} improvement when (i) the C2C P_{TOT} variance of a die is high and (ii) a thermal constraint and the impact of thermal throttling are considered as shown in

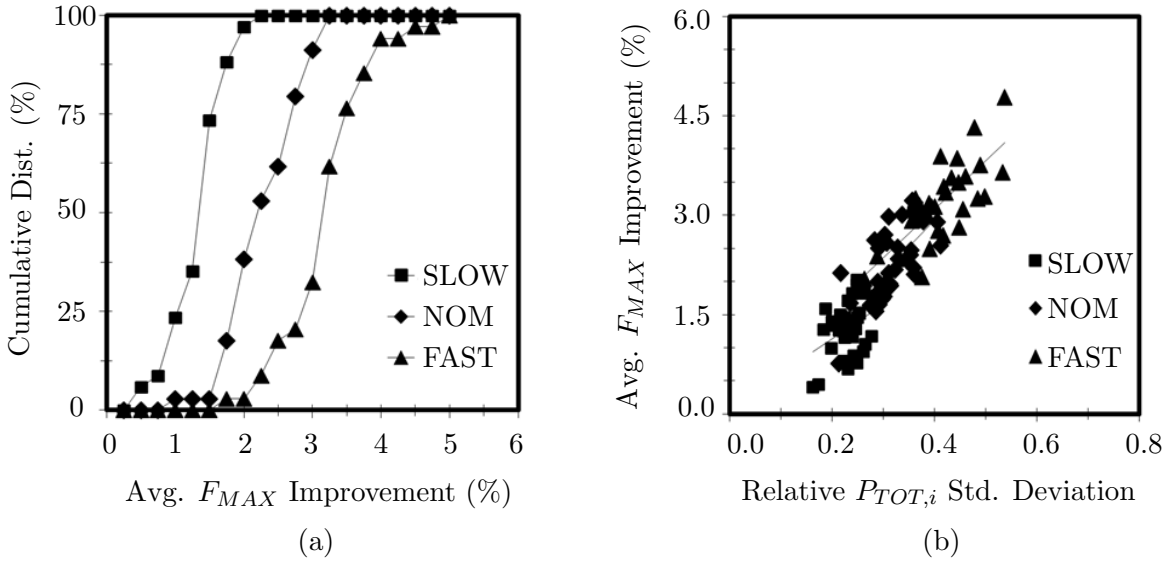


Figure 7.6: (a) Cumulative distribution of the average F_{MAX} improvement of 16-core processors under a power constraint after optimization. (b) The corresponding average F_{MAX} improvement versus C2C relative $P_{TOT,i}$ std. deviation.

the next section.

Figure 7.6-(a) plots the cumulative distributions of the average F_{MAX} improvement of 16-core processors under a power constraint after the optimization is applied. For slow-, nominal-, and fast-corner dies, P_{TDP} is set to 50W [78], 95W [79], and 185W [80], respectively; dies from slow, nominal, and fast corners are often used for different computing segments (i.e., mobile, desk-top, and servers) that have different power constraints due to their very different speed and power consumption characteristics. The proposed optimization for 16-core processors can improve the average F_{MAX} by up to 2.0%, 3.2%, and 4.8% for the slow, nominal, and fast corners, respectively. Figure 7.6-(b) plots the average F_{MAX} improvement versus C2C $P_{TOT,i}$ standard deviation. The $P_{TOT,i}$ standard deviation values are normalized to P_{TDP} (= 50W, 95W, and 185W for slow, nominal, and fast corners)/16. Clearly, the F_{MAX} improvement is greater for dies with more WID C2C P_{TOT} variations statistically. Note that the relative C2C F_{MAX} and P_{LEAK} variations also increase as more cores are integrated per die. In turn, this can increase the benefit of the proposed

optimization techniques.

7.4 Maximizing Performance Under Power and Thermal Constraints

As described in Section 7.1, in the presence of C2C P_{LEAK} variations, even a many-core processor adopting chip-wide clocking leads to very high temperature variations (and thus performance variations) across cores. This is because fast cores with higher P_{LEAK} will experience more thermal throttling although they run at the same F_{MAX} as the slowest one. This problem worsens when each core is allowed to operate at its own fastest F_{MAX} (i.e., per-core clocking). The fast cores will experience even more thermal throttling due to higher F_{MAX} (i.e. P_{DYN}), resulting in a lower average F_{MAX} value than their peak F_{MAX} value over long runtime periods. Potentially, this may diminish the benefit of per-core clocking. For example, as shown in Figure 7.2, where no thermal constraint is enforced, the steady-state hotspot temperature of fast cores (i.e., core 4, 5, 8, 9, 12, and 13 in Figure 7.5) is much higher than that of slow cores due to their very high power consumption. This can lead to more frequent thermal throttling for the faster cores under a thermal constraint, degrading the throughput of the faster cores as a consequence. For example, the average F_{MAX} of the 16-core processor die shown in Figure 7.1 decreases by 8.2% when thermal throttling is engaged for cores whose hotpsot temperature exceeds 100 °C. To model the impact of thermal throttling on the average F_{MAX} over a long runtime period, in this work an oracle approach is assumed for the dynamic thermal management (DTM) scheme. In other words, we reduce the voltage/frequency of the cores exceeding 100 °C such that their hotspot temperature is just below 100 °C. We observed that a realistic DTM scheme will result in more average F_{MAX} degradation since it searches the operating point adaptively at runtime by trial and errors through a feedback mechanism. Furthermore, a thread migration

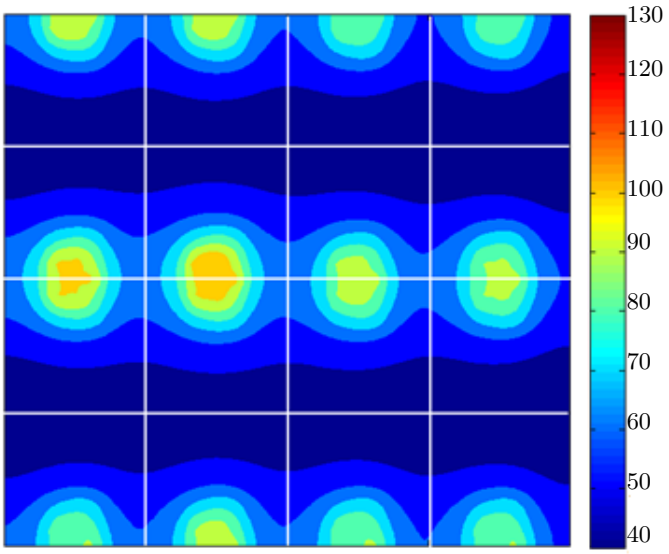


Figure 7.7: Temperature profile of a per-core clocking 16-core processor running 16 gcc instances after the optimization technique is applied.

technique is not applicable since all the cores are running homogeneous threads; a thread running on a hot core should migrate to another cool core which is also running a thread.

On the other hand, the proposed optimization technique can positively impact the throughput of thermal-constrained many-core processors. It reduces the power consumption of the fast cores, and re-allocates the reduced power consumption to the slow cores that originally consumed much less power. Figure 7.7, which plots the temperature profile of the 16-core processor after the optimization technique applied, shows more uniform hotspot temperatures across cores than before optimization shown in Figure 7.2 for the same die. This results in much less average F_{MAX} degradation when the thermal throttling is engaged for the cores whose hotspot temperature reaches at 100 °C at runtime; the average F_{MAX} after the optimization decreases by only 1.7%. Overall, the optimization technique improves the average F_{MAX} by 7.1%.

Figure 7.8 plots the cumulative distributions of the average F_{MAX} of 16-core processors under both power and thermal constraints after the optimization and thermal throttling are

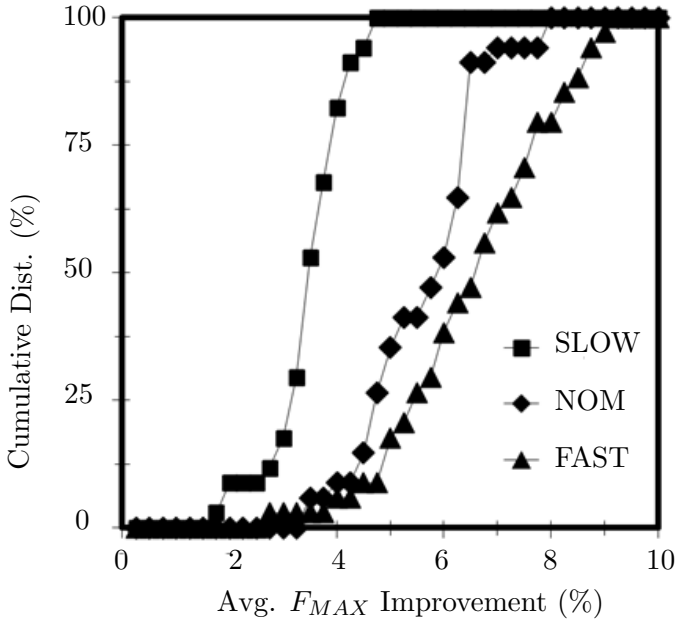


Figure 7.8: The cumulative distributions of the average F_{MAX} improvement of 16-core processors under both power and thermal constraints after the optimization is applied.

applied. The proposed optimization for 16-core processors can improve the average F_{MAX} by up to 4.7%, 7.9%, and 9.2% for the slow, nominal, and fast corners. In this experiment, we also confirm the strong correlation between the average F_{MAX} improvement and the relative C2C $P_{TOT,i}$ standard deviation, as shown in Figure 7.6-(b), even when both power and thermal constraints are considered.

7.5 Implementation Cost

The proposed optimization algorithm requires that the full F_{MAX} and P_{TOT} characteristics of each core in a processor are known for the post-silicon tuning process. F_{MAX} characterization is often done at a few voltage points which are interpolated by power management firmware to implement P-state transitions [81]. To minimize the cost of characterization, we can estimate F_{MAX} by interpolating the few F_{MAX} values available from

the P-state characterization process. Similar methodology can be adopted for P_{TOT} characterization.

7.6 Related Work

K. Bowman *et al.* and J. Tschanz *et al.* evaluated the impact of D2D and WID variations on the maximum operating frequency distribution of single core processors [82][31]. K. Bowman *et al.* also presented analytical throughput models for globally-clocked multi-core processors and the sensitivity of several processor designs' throughputs to process variations [75]. E. Humenay *et al.* modeled the impact of WID variations on C2C frequency and power variations in multi-core processors [83]. They also analyzed the impact of applying per-core adaptive body-biasing (ABB) and adaptive voltage scaling (AVS) on performance symmetry considering only a thermal constraint. S. Herbert *et al.* presented analytical models for the throughput of multi-core processors using frequency island (FI) clocking (per-core clocking) and quantified performance benefit of FI clocking across a range of multi-core processor designs [22]. R. Rao *et al.* provided analytical thermal and power models, and analyzed the impact of thermal constraints on performance of multi-core processors [84]. J. Donald *et al.* presented power-performance trade-off analysis in the presence of process variations, and they provided a method to predict an optimal cut-off point for turning-off extra cores in multi-core processors [85].

7.7 Chapter Summary

C2C frequency, power, and temperature variations limit the performance of power-/thermal-constrained many-core processors. To maximize the throughput of such processors, an optimization algorithm is proposed that maximizes the average frequency of processors by balancing the power consumption between fast and slow cores in processors

supporting per-core voltage/frequency domains. The proposed optimization algorithm is based on the following observations: (1) the throughput of a many-core processor supporting per-core clocking is proportional to the average frequency of cores; (2) each core in a processor exhibits considerably different trade-offs between power consumption and frequency due to WID process variations and the DIBL effect; and (3) C2C power consumption and frequency variations also lead to substantial C2C temperature variations, causing more thermal throttling for fast cores that consume considerably more power than slow cores. Experimental results demonstrate the proposed optimization algorithm can improve the average frequency of 16-core processors by up to 10% for dies exhibiting high C2C power variances. Considering that future technology will exhibit higher C2C frequency and power variations, it is expected that the improvement using the proposed technique will be higher.

Chapter 8

Conclusions

As feature sizes shrink with every generation of CMOS technology, processors are increasingly impacted by parameter variations and reliability issues such as bias temperature instability (BTI). In addition, as the transistor density on a chip increases, power consumption has emerged as an important design constraint, requiring processor manufacturers to focus on maximizing performance/Watt rather than pure performance in high-performance as well as mobile processors. Achieving high performance/Watt (i.e., power efficiency) while maintaining low cost requires efficient and cost-effective power delivery methods for today's processors. Power-gating (PG) devices are commonly used in multi-core processors to reduce standby leakage power. A PG device is impacted by process variations as well as aging effects. In this thesis, we proposed novel methods for improving the performance, yield, power efficiency, and reliability of multi-core processors using PG devices.

First, we demonstrated the use of programmable PG devices to improve the yield of manufactured dies in the presence of die-to-die (D2D) variations. The strength of a programmable PG device can be tuned in a post-manufacturing step to reduce the leakage power (P_{LEAK}) of leaky dies, thereby increasing the number of acceptable dies which satisfy the total power constraint (i.e., yield). The effectiveness of the method is demonstrated

for two different design scenarios: 1) ASIC type designs with fixed leakage power and 2) processor-type designs with variable leakage power constraints. Simulation results demonstrated that about 88% and 98% of discarded dies could be recovered by the proposed methods in two design scenarios, respectively. Further, the programmable PG device tuning can be applied at a per-core level to improve the F_{MAX} of multi-core processors with global clocking scheme. By tuning the PG device of each core, the proposed optimization method assigns the most performance/power efficient voltage/frequency point for each core during a post-manufacturing step. As a result of the optimization, the power consumption of fast and slow cores can be balanced, leading to higher maximum throughput (i.e., F_{MAX}) under a power constraint. Our results showed an F_{MAX} improvement of 3%-21% on average for 2-, 4-, 8-, and 16-core processors.

Second, the impact of BTI aging and chip temperature variations on the virtual rail voltage (VV_{DD}) of power-gated circuits is analyzed. Our simulations showed that in power-gated circuits, low temperature does not reduce leakage power at the expected rate due to increased VV_{DD}. Also, upsizing the PG device to compensate BTI degradation in later chip lifetime increases leakage power in early chip lifetime due to the increased VV_{DD}. We proposed a circuit technique that track NBTI degradation of a PMOS PG device and chip temperature variation, and adjusts the strength of PG devices accordingly. The proposed method clamps VV_{DD} close to a target level at runtime in spite of any given NBTI degradation and/or temperature variation within the specified ranges. As a result, leakage power is reduced by $\sim$10% and dynamic power is reduced by $\sim$4% for the given range of average die temperatures in early chip lifetime. We demonstrated that the proposed tracking method maintains VV_{DD} close to a target level even in the presence of within-die (WID) spatial process and temperature variations. Oversizing the PG device to account for BTI degradation leads to higher than necessary VV_{DD} and thus increased gate oxide stress. The VV_{DD} clamping method proposed in this thesis can reduce the oxide stress and improve

device reliability. Over a period of 7.5 years, the gate oxide failure rate reduced by 5.1%, 3.8%, and 4.1% for fast, nominal, and slow corners by VV_{DD} clamping.

Third, we showed that increasing the size of a PG device beyond a certain value does not improve the performance (i.e., F_{MAX}) of the power-gated circuit while increasing the power consumption significantly. Hence, sizing the PG device purely to minimize F_{MAX} degradation can result in a large PG area while reducing the power efficiency of the processor. We proposed a PG sizing algorithm to optimize the size of a PG device and the supply voltage in processors with adaptive voltage scaling (AVS) while maximizing performance and power efficiency. Our experimental results demonstrate that the joint optimization considering D2D and WID variations can reduce the size of PG devices by 59% while increasing F_{MAX} by 3% of power constrained processors using GC. The performance gain is achieved due to the power headroom created by reducing the PG size of the fast cores which reduces their P_{LEAK} significantly without impacting F_{MAX}. When the optimization was applied to the multi-core processors using FI, the size of PG devices is reduced by 58% and 57% while improving F_{MAX} by $\sim$3% for 4- and 16-core processors.

Fourth, we investigated the use of per-core PG (PCPG) in multi-core processors as on-chip linear voltage regulator (VR). Off-chip switching VRs are invariably used to supply power to the different components of a high-performance processor platform. However, providing multiple off-chip switching VRs for individual voltage/frequency control of cores in multi-core processors increases platform cost while on-chip inductors result in low efficiency on-chip VRs. We demonstrated that PCPG devices augmented with small circuitry can operate as low-cost low dropout (LDO) VRs. Unlike on-chip switching VRs, LDO VRs do not require on-chip inductors and do not inject switching noise in the substrate. Our investigation showed that a switching VR requires nearly four times larger chip area than a comparable LDO VR. The efficiency of LDO VRs reduces as their output voltage drops (i.e., large difference between input and output voltage of the LDO VRs). However, our

experiments showed that C2C voltage variations are relatively small when the voltages are optimized to maximize performance under a power constraint. After modeling the power efficiency of both LDO and switching VRs using a 32nm technology, we showed that the $MIPS^3/W$ of an 8-core processor using LDO VRs is slightly higher than that of a processor using switching VRs.

Finally, we proposed an optimization algorithm that maximizes the performance of multi-core processors with per-core clocking and power/thermal constraint. The proposed algorithm balances the power consumption and hotspot temperatures of fast and slow cores in processors supporting per-core voltage/frequency domains by selecting the most performance/power optimal voltage and frequency for each core during post-manufacturing testing. Core-to-core (C2C) power and frequency variations, which arise due to WID process variations, lead to substantial C2C temperature variations, causing more thermal throttling for fast cores that consume considerably more power than slow cores. This reduces the average F_{MAX} (i.e., performance) of the processor. The proposed optimization leads to less frequent thermal throttling of fast cores and thus higher maximum throughput under power and thermal constraints. Experimental results demonstrate the proposed optimization algorithm can improve the average frequency of 16-core processors by up to 10% for dies exhibiting high C2C power variances.

References

[1] K. Alvin, B. Barrett, R. Brightwell, S. S. Dosanjh, A. Geist, K. S. Hemmert, M. A. Heroux, D. Kothe, R. C. Murphy, J. Nichols, R. Oldfield, A. Rodrigues, and J. S. Vetter, "On the path to exascale," *IJDST*, vol. 1, no. 2, pp. 1–22, 2010.

[2] ITRS, "International technology roadmap for semiconductors," International Technology Roadmap for Semiconductors, Tech. Rep., 2011. [Online]. Available: http://www.itrs.net/Links/2011ITRS/Home2011.htm

[3] K. Aygün, M. J. Hill, K. Eilert, K. Radhakrishnan, and A. Levin, "Power delivery for high-performance microprocessors," *Intel Technology Journal*, vol. 09, pp. 273–283, 2005.

[4] H. R. Ghasemi, A. A. Sinkar, M. J. Schulte, and N. S. Kim, "Cost-effective power delivery to support per-core voltage domains for power-constrained processors," in *49th Annual Design Automation Conference*, ser. DAC '12. New York, NY, USA: ACM, 2012, pp. 56–61. [Online]. Available: http://doi.acm.org/10.1145/2228360.2228372

[5] K. Shi, Z. Lin, and Y.-M. Jiang, "A power network synthesis method for industrial power gating designs," in *8th International Symposium on Quality Electronic Design, (ISQED '07)*, march 2007, pp. 362 –367.

[6] K. Usami, N. Kawabe, M. Koizumi, K. Seta, and T. Furusawa, "Automated selective

multi-threshold design for ultra-low standby applications," in *International Symposium on Low Power Electronics and Design, (ISLPED '02)*, 2002, pp. 202 – 206.

[7] Intel Corp. Intel Core i7 Processor. [Online]. Available: http://www.intel.com/content/www/us/en/processors/core/core-i7-processor.html

[8] AMD. AMD A-Series Processor-in-a-Box. [Online]. Available: http://www.amd.com/us/products/desktop/processors/a-series/Pages/a-series-pib.aspx

[9] S. Mutoh, S. Shigematsu, Y. Gotoh, and S. Konaka, "Design method of mtcmos power switch for low-voltage high-speed lsis," in *Asia and South Pacific Design Automation Conference, (ASP-DAC '99)*, jan 1999, pp. 113 –116 vol.1.

[10] H.-S. Won, K.-S. Kim, K.-O. Jeong, K.-T. Park, K.-M. Choi, and J.-T. Kong, "An mtcmos design methodology and its application to mobile computing," in *International Symposium on Low Power Electronics and Design, (ISLPED '03)*, aug. 2003, pp. 110 – 115.

[11] J. Kao, S. Narendra, and A. Chandrakasan, "Mtcmos hierarchical sizing based on mutual exclusive discharge patterns," in *Design Automation Conference, (DAC'98)*, june 1998, pp. 495 –500.

[12] D.-S. Chiou, D.-C. Juan, Y.-T. Chen, and S.-C. Chang, "Fine-grained sleep transistor sizing algorithm for leakage power minimization," in *44th ACM/IEEE Design Automation Conference, (DAC '07)*, june 2007, pp. 81–86.

[13] S. Kim, S. Kosonocky, and D. Knebel, "Understanding and minimizing ground bounce during mode transition of power gating structures," in *International Symposium on Low Power Electronics and Design, (ISLPED '03)*, aug. 2003, pp. 22 – 25.

[14] W. T. Ng and O. Trescases, "Power management for modern vlsi loads using dynamic

voltage scaling," in *7th International Conference on Solid-State and Integrated Circuits Technology*, vol. 2, oct. 2004, pp. 1412 – 1415.

[15] W. Kim, M. S. Gupta, G.-Y. Wei, and D. Brooks, "System level analysis of fast, per-core dvfs using on-chip switching regulators," in *High Performance Computer Architecture, (HPCA'08)*. IEEE Computer Society, 2008, pp. 123–134.

[16] E. Rotem, R. Ginosar, A. Mendelson, and U. Weiser, "Multiple clock and voltage domains for chip multi processors," in *42nd Annual IEEE/ACM International Symposium on Microarchitecture, (MICRO-42)*, dec. 2009, pp. 459–468.

[17] Intel Corp. (2008, November) Intel Turbo Boost Technology in Intel Core Microarchitecture (Nehalem) based Processors. [Online]. Available: http://download.intel.com/design/processor/applnots/320354.pdf

[18] W. Huang, G. Schuellein, and D. Clavette, "A scalable multiphase buck converter with average current share bus," in *18th Annual IEEE Applied Power Electronics Conference and Exposition, (APEC '03)*, vol. 1, feb. 2003, pp. 438 –443 vol.1.

[19] S. Reda and S. R. Nassif, "Analyzing the impact of process variations on parametric measurements: Novel models and applications," in *Design, Automation, and Test in Europe, (DATE'09)*. IEEE, 2009, pp. 375–380.

[20] S. Dighe, S. R. Vangal, P. A. Aseron, S. Kumar, T. Jacob, K. A. Bowman, J. Howard, J. Tschanz, V. Erraguntla, N. Borkar, V. K. De, and S. Borkar, "Within-die variation-aware dynamic-voltage-frequency-scaling with optimal core allocation and thread hopping for the 80-core teraflops processor," *Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 184–193, 2011.

[21] S. R. Sarangi, B. Greskamp, R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas,

"Varius: A model of process variation and resulting timing errors for microarchitects," in *IEEE Transactions on Semiconductor Manufacturing*, 2008.

[22] S. Herbert and D. Marculescu, "Characterizing chip-multiprocessor variability-tolerance," in *45th ACM/IEEE Design Automation Conference, (DAC'08)*, june 2008, pp. 313 –318.

[23] S. Narendra, D. Antoniadis, and V. De, "Impact of using adaptive body bias to compensate die-to-die vt variation on within-die vt variation," in *International Symposium on Low Power Electronics and Design, (ISLPED'99)*, aug. 1999, pp. 229–232.

[24] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, and S. Vrudhula, "Predictive modeling of the nbti effect for reliable design," in *IEEE Custom Integrated Circuits Conference, (CICC '06)*, sept. 2006, pp. 189–192.

[25] K. Shi and D. Howard, "Challenges in sleep transistor design and implementation in low-power designs," in *43rd ACM/IEEE Design Automation Conference, (DAC'06)*, 0-0 2006, pp. 113–116.

[26] H. Singh Deogun, D. Sylvester, R. Rao, and K. Nowka, "Adaptive mtcmos for dynamic leakage and frequency control using variable footer strength," in *IEEE International SOC Conference*, sept. 2005, pp. 147 – 150.

[27] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm design exploration," in *7th International Symposium on Quality Electronic Design, (ISQED '06)*, march 2006, pp. –590.

[28] Predictive technology model (ptm). [Online]. Available: http://ptm.asu.edu/

[29] Hotspot 5.0 temperature modeling tool. [Online]. Available: http://lava.cs.virginia.edu/HotSpot/index.htm

[30] R. Rao, S. Vrudhula, and C. Chakrabarti, "Throughput of multi-core processors under thermal constraints," in *International Symposium on Low Power Electronics and Design, (ISLPED'07*, 2007.

[31] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," in *2002 IEEE International Solid-State Circuits Conference, (ISSCC'02)*, vol. 1, 2002, pp. 422–478 vol.1.

[32] Y. Yasuda, N. Kimizuka, Y. Akiyama, Y. Yamagata, Y. Goto, and K. Imai, "System lsi multi-vth transistors design methodology for maximizing efficiency of body-biasing control to reduce vth variation and power consumption," in *IEEE International Electron Devices Meeting, (IEDM'05)*, dec. 2005, pp. 68 –71.

[33] T. Chen and S. Naffziger, "Comparison of adaptive body bias (abb) and adaptive supply voltage (asv) for improving delay and leakage under the presence of process variation," *IEEE Transactions on Very Large Scale Integration Systems (TVLSI)*, vol. 11, no. 5, pp. 888 –899, oct. 2003.

[34] D. Lackey, P. Zuchowski, T. Bednar, D. Stout, S. Gould, and J. Cohn, "Managing power and performance for system-on-chip designs using voltage islands," in *IEEE/ACM International Conference on Computer Aided Design, (ICCAD'02)*, nov. 2002, pp. 195 – 202.

[35] J. Stathis, "Physical and predictive models of ultrathin oxide reliability in cmos devices and circuits," *IEEE Transactions on Device and Materials Reliability (TDMR)*, vol. 1, no. 1, pp. 43 –59, mar 2001.

[36] E. Y. Wu and J. Su, "Power-law voltage acceleration: A key element for ultra-thin gate oxide reliability," *Microelectronics Reliability*, pp. 1809–1834, 2005.

[37] E. Wu, J. Su, W. Lai, E. Nowak, J. McKenna, A. Vayshenker, and D. Harmon, "Interplay of voltage and temperature acceleration of oxide breakdown for ultra-thin gate oxides," *Journal of Solid-State Electronics*, vol. 46, no. 11, pp. 1787 – 1798, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S003811010200151X

[38] E. Wu, J. Sune, and W. Lai, "On the weibull shape factor of intrinsic breakdown of dielectric films and its accurate experimental determination. part ii: experimental results and the effects of stress conditions," *IEEE Transactions on Electron Devices*, vol. 49, no. 12, pp. 2141 – 2150, dec 2002.

[39] C. Zhuo, D. Sylvester, and D. Blaauw, "Process variation and temperature-aware reliability management," in *Design, Automation Test in Europe Conference Exhibition (DATE'10)*, march 2010, pp. 580–585.

[40] R. Vattikonda, W. Wang, and Y. Cao, "Modeling and minimization of pmos nbti effect for robust nanometer design," in *43rd ACM/IEEE Design Automation Conference, (DAC'06)*, 0-0 2006, pp. 1047–1052.

[41] R. Fernandez, B. Kaczer, A. Nackaerts, S. Demuynck, R. Rodriguez, M. Nafria, and G. Groeseneken, "Ac nbti studied in the 1 hz – 2 ghz range on dedicated on-chip cmos circuits," in *International Electron Devices Meeting, (IEDM '06)*, dec. 2006, pp. 1 –4.

[42] T.-H. Kim, R. Persaud, and C. Kim, "Silicon odometer: An on-chip reliability monitor for measuring frequency degradation of digital circuits," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 43, no. 4, pp. 874–880, april 2008.

[43] J. Tschanz, S. Narendra, Y. Ye, B. Bloechel, S. Borkar, and V. De, "Dynamic sleep transistor and body bias for active leakage power control of microprocessors," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 38, no. 11, pp. 1838 – 1845, nov. 2003.

[44] D. Chinnery and K. Keutzer, *Closing the Power Gap Between ASIC & Custom Tools and Techniques for Low Power Design.* Springer, 2007, pp. 251–280.

[45] B. Chatterjee, M. Sachdev, S. Hsu, R. Krishnamurthy, and S. Borkar, "Effectiveness and scaling trends of leakage control techniques for sub-130 nm cmos technologies," in *International Symposium on Low Power Electronics and Design, (ISLPED '03)*, aug. 2003, pp. 122 – 127.

[46] P. Bose, M. Martonosi, and D. Brooks, "Modeling and analyzing cpu power and performance: Metrics, methods, and abstractions," in *Tutorial, ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems, Cambridge, MA*, 2001.

[47] K. Bowman, A. Alameldeen, S. Srinivasan, and C. Wilkerson, "Impact of die-to-die and within-die parameter variations on the throughput distribution of multi-core processors," in *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED'07)*, aug. 2007, pp. 50–55.

[48] A. Bakhoda, G. L. Yuan, W. W. L. Fung, H. Wong, and T. M. Aamodt, "Analyzing cuda workloads using a detailed gpu simulator," in *IEEE International Symposium on Performance Analysis of Systems and Software, (ISPASS'09).* IEEE, 2009, pp. 163–174.

[49] Originlab - origin and originpro - data analysis and graphing software. [Online]. Available: http://www.originlab.com

[50] R. Teodorescu and J. Torrellas, "Variation-aware application scheduling and power management for chip multiprocessors," in *ISCA.* IEEE, 2008, pp. 363–374.

[51] A. R. Alameldeen, C. J. Mauer, M. Xu, P. J. Harper, M. M. K. Martin, D. J. Sorin, M. D. Hill, and D. A. Wood, "Evaluating non-deterministic multi-threaded commercial

workloads," in *Fifth Workshop on Computer Architecture Evaluation Using Commercial Workloads*, 2002, pp. 30–38.

[52] C. Bienia, "Benchmarking modern multiprocessors," Ph.D. dissertation, Princeton University, January 2011.

[53] M. M. K. Martin, D. J. Sorin, B. M. Beckmann, M. R. Marty, M. Xu, A. R. Alameldeen, K. E. Moore, M. D. Hill, and D. A. Wood, "Multifacet's general execution-driven multiprocessor simulator (gems) toolset," *SIGARCH Computer Architecture News*, vol. 33, no. 4, pp. 92–99, 2005.

[54] D. M. Brooks, P. Bose, S. E. Schuster, H. Jacobson, P. N. Kudva, A. Buyuktosunoglu, J.-D. Wellman, V. Zyuban, M. Gupta, and P. W. Cook, "Power-aware microarchitecture: Design and modeling challenges for next-generation microprocessors," *IEEE Micro*, vol. 20, pp. 26–44, 2000.

[55] Intel Corp. (2009, Jun.) Intel turbo boost technology 2.0. [Online]. Available: http://www.intel.com/technology/turboboost/index.htm

[56] ——. (2006, oct) Intel workstation board s975xbx2 technical product specification. [Online]. Available: http://download.intel.com/support/motherboards/server/s975xbx2/sb/s975xbx2_tps_rev_11.pdf

[57] P. Hazucha, G. Schrom, J. Hahn, B. Bloechel, P. Hack, G. Dermer, S. Narendra, D. Gardner, T. Karnik, V. De, and S. Borkar, "A 233-mhz 80%-87% efficient four-phase dc-dc converter utilizing air-core inductors on package," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 40, no. 4, pp. 838–845, april 2005.

[58] P. Hazucha, S. T. Moon, G. Schrom, F. Paillet, D. Gardner, S. Rajapandian, and T. Karnik, "High voltage tolerant linear regulator with fast digital control for biasing

of integrated dc-dc converters," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 42, no. 1, pp. 66–73, jan. 2007.

[59] S. Rusu, S. Tam, H. Muljono, J. Stinson, D. Ayers, J. Chang, R. Varada, M. Ratta, S. Kottapalli, and S. Vora, "A 45 nm 8-core enterprise xeon; processor," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 45, no. 1, pp. 7–14, jan 2010.

[60] N. S. Kim, J. Seomun, A. A. Sinkar, J. Lee, T. H. Han, K. Choi, and Y. Shin, "Frequency and yield optimization using power gates in power-constrained designs," in *ISLPED*, J. Henkel, A. Keshavarzi, N. Chang, and T. Ghani, Eds. ACM, 2009, pp. 121–126.

[61] Y. Hoskote, S. R. Vangal, A. Singh, N. Borkar, and S. Borkar, "A 5-ghz mesh interconnect for a teraflops processor." *IEEE Micro*, vol. 27, no. 5, pp. 51–61, 2007. [Online]. Available: http://dblp.uni-trier.de/db/journals/micro/micro27.html#HoskoteVSBB07

[62] W. Fu and A. Fayed, "A feasibility study of high-frequency buck regulators in nanometer cmos technologies," in *IEEE Dallas Circuits and Systems Workshop,(DCAS'09)*, oct. 2009, pp. 1–4.

[63] P. Hazucha, T. Karnik, B. Bloechel, C. Parsons, D. Finan, and S. Borkar, "Area-efficient linear regulator with ultra-fast load regulation," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 40, no. 4, pp. 933 – 940, april 2005.

[64] J. Klein. (2006) Fairchild semiconductors. [Online]. Available: http://www.fairchildsemi.com/an/AN/AN-6005.pdf

[65] J. Lee, G. Hatcher, L. Vandenberghe, and C.-K. K. Yang, "Evaluation of fully-integrated switching regulators for cmos process technologies," *IEEE Transactions on*

Very Large Scale Integration Systems (TVLSI), vol. 15, no. 9, pp. 1017 –1027, sept. 2007.

[66] D. Gardner, G. Schrom, F. Paillet, and S. Chickamenahalli, "Inductors for integrated circuit packages," US Patent US 7 911 313, 03 22, 2011. [Online]. Available: http://www.patentlens.net/patentlens/patent/US_7911313/en/

[67] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "Mcpat: an integrated power, area, and timing modeling framework for multicore and manycore architectures," in *MICRO*, D. H. Albonesi, M. Martonosi, D. I. August, and J. F. Martínez, Eds. ACM, 2009, pp. 469–480.

[68] Microsoft. Ppm in windows vista and windows server 2008. [Online]. Available: http://msdn.microsoft.com/en-us/windows/hardware/gg463252.aspx

[69] J. Li and J. F. Martnez, "Dynamic power-performance adaptation of parallel computation on chip multiprocessors," in *High Performance Computer Architecture (HPCA'06)*, 2006, pp. 77–87.

[70] S. Eyerman and L. Eeckhout, "Fine-grained dvfs using on-chip regulators," *TACO*, pp. 1–1, 2011.

[71] S. Herbert and D. Marculescu, "Variation-aware dynamic voltage/frequency scaling," in *High Performance Computer Architecture (HPCA'09)*, 2009, pp. 301–312.

[72] K. K. Rangan, G.-Y. Wei, and D. Brooks, "Thread motion: fine-grained power management for multi-core systems," in *IEEE International Symposium on Computer Architecture (ISCA'09)*, 2009, pp. 302–313.

[73] J. Howard, S. Dighe, Y. Hoskote, S. Vangal, D. Finan, G. Ruhl, D. Jenkins, H. Wilson, N. Borkar, G. Schrom, F. Pailet, S. Jain, T. Jacob, S. Yada, S. Marella, P. Salihundam,

V. Erraguntla, M. Konow, M. Riepen, G. Droege, J. Lindemann, M. Gries, T. Apel, K. Henriss, T. Lund-Larsen, S. Steibl, S. Borkar, V. De, R. Van Der Wijngaart, and T. Mattson, "A 48-core ia-32 message-passing processor with dvfs in 45nm cmos," in *IEEE International Solid-State Circuits Conference (ISSCC'10)*, feb. 2010, pp. 108 –109.

[74] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: a framework for architectural-level power analysis and optimizations," in *27th International Symposium on Computer Architecture (ISCA'00)*, june 2000, pp. 83–94.

[75] K. Bowman, A. Alameldeen, S. Srinivasan, and C. Wilkerson, "Impact of die-to-die and within-die parameter variations on the clock frequency and throughput of multi-core processors," *IEEE Transactions on Very Large Scale Integration Systems (TVLSI)*, vol. 17, no. 12, pp. 1679 –1690, dec. 2009.

[76] S. Chaudhry, R. Cypher, M. Ekman, M. Karlsson, A. Landin, S. Yip, H. Zeffer, and M. Tremblay, "Rock: A high-performance sparc cmt processor," *IEEE Micro*, vol. 29, pp. 6–16, 2009.

[77] Intel Corp. (2009, sept) Vrm and evrd 11.1 design guidelines. [Online]. Available: www.intel.com/assets/PDF/designguide/321736.pdf

[78] ——. Intel Xeon Processor L5320(8M Cache, 1.86GHz, 1066MHz FSB). [Online]. Available: http://ark.intel.com/Product.aspx?id=29767

[79] ——. Intel Xeon Processor 5050(4M Cache, 3.00 GHz, 667 MHz FSB). [Online]. Available: http://ark.intel.com/Product.aspx?id=27210

[80] ——. Intel Itanium Processor 9350(24M Cache, 1.73 GHz, 4.80GT/s Intel QPI). [Online]. Available: http://ark.intel.com/Product.aspx?id=43410

[81] V. J. Zimmer, M. A. Rothman, and D. C. Estrada, "Methods and apparatus to perform power management in processor systems," US Patent Application US 2007/0 234 075 A1, 10 04, 2007. [Online]. Available: http://www.patentlens.net/patentlens/patent/US_2007_0234075_A1/en/

[82] K. Bowman, S. Duvall, and J. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 37, no. 2, pp. 183 –190, feb 2002.

[83] E. Humenay, D. Tarjan, and K. Skadron, "Impact of process variations on multicore performance symmetry," in *Design, Automation Test in Europe Conference Exhibition (DATE '07)*, april 2007, pp. 1 –6.

[84] R. Rao, S. Vrudhula, and C. Chakrabarti, "Throughput of multi-core processors under thermal constraints," in *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED'07)*, aug. 2007, pp. 201 –206.

[85] J. Donald and M. Martonosi, "Power efficiency for variation-tolerant multicore processors," in *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED'06)*, oct. 2006, pp. 304 –309.