Design for Reliability and Low Power in Emerging Technologies

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte
Dissertation

von

M.Sc. Sami Alsalamin

Tag der mündlichen Prüfung: 28. July 2021
Erster Gutachter: Prof. Dr.-Ing. Jörg Henkel
Karlsruhe Institute of Technology (KIT)
Zweiter Gutachter: J.-Prof. Dr.-Ing. Hussam Amrouch
University of Stuttgart
Dritter Gutachter: Prof. Sergio Bampi
Federal University of Rio Grande do Sul

Acknowledgment

I would like to express my immeasurable appreciation and sincere gratitude to J. Professor Hussam Amrouch. This dissertation would not have been possible without his unreserved support. I am indebted and thankful for such a great supervisor. I am grateful for his precise guidance, countless enlightening discussions, and corrections during my research. His deep knowledge and enthusiasm for semiconductor physics and devices are one of the many motivating factors that kept me going during difficult times. His patience and encouragement indeed helped me overcome all obstacles during this work.

I would also like to extend my deepest gratitude to Prof. Sergio Bampi for accepting to be my co-examiner and providing erudite feedback. My sincere thankfulness goes to our collaborators Prof. Yogesh S. Chauhan, Prof. Andreas Gerstlauer, and Prof. Tulika Mitra for their great advice, ideas, and excellent guidance. I am deeply thankful for their help, which shaped my research.

I also want to thank all colleagues from the Chair for Embedded Systems for the good time. There are many colleagues who have helped, inspired and motivated me throughout these challenging years. Their support was an indispensable instrument to accomplish this research program. Therefore, I would like to express my thanks for all CES members. Special thanks go to my colleague and friend Victor van Santen. The in-depth technical discussions with him played an important role in improving the quality of this work. We together shared very great and difficult moments and he was a great assistant, supporter, and adviser in all aspects. I would also like to thank my colleague and kind friend Martin Rapp for the numerous technical discussions that helped me improve my knowledge in the topic area. I would also like to thank Dr. Georgios Zervakis for all his help. In particular, I am grateful for his eagerness to help and continuously advise.

This achievement would never be done without the continued support of my family. I would like to express my utmost heartfelt gratitude to my parents. Sadly, my father passed before I finished this work, who was the biggest motivation and encouraged me always. My mother, the inexhaustible source of love and tenderness, who has been continuously supporting me, taking care of me, and praying for me since ever. My brothers, Samer, Sameer, Anas, Ammar, and Ata, are the great support and source of proudness, this achievement would

have never be done without you. Finally, none of this would have been possible without the support of my wife, Maram, who was extraordinary patient during my Ph.D works. I would like to express my uttermost heartfelt gratitude to her and for my sons, Yousef and Malek, and my little queen Masa.

Abstract

The most crucial growth driver for the semiconductor industry is transistors downscaling. Consequently, for decades, circuits have become denser and more complex, following a continuous trend with every new technology node (further scaling). Previously, transistor scaling was always accompanied by supply voltage reduction to reduce power consumption, thus maintaining a constant power density. Entering the nanometer era has slowed down the scaling process due to many difficulties (e.g., physical limitations) and the non-ideality in voltage scaling, leading to increased power density. This has aggravated many reliability issues. Transistor aging phenomenon, excessive temperature, and self-heating effect are few examples of such issues. Conventionally, to mitigate these issues and sustain a reliable operation, timing guardbands have to be pessimistically considered, individually for every degradation effect, on top of the circuit's delay, to compensate for the induced delay degradations. This severely *degrades the overall performance*. However, mitigation can be alternatively achieved by applying substitutional techniques such as zero-temperature coefficient, approximate computing, etc. Even though these techniques can eliminate or greatly reduce the large timing guardbands, further consequences and trade-offs may be encountered.

Difficulties limiting the scaling of CMOS technology continue to challenge. These limitations and the corresponding technological challenges are currently dictating a shift in research from CMOS technology to that of emerging technologies. For instance, the Negative Capacitance Field-Effect Transistor (NCFET) is an emerging technology that has great potential to replace CMOS technology since NCFET exhibits considerable improvements in circuits' performance. Additionally, designers switched to complex models by employing parallel processing modules instead of higher frequencies. Such complex models necessitate advanced power management techniques at all design levels. These techniques must be revisited with new technology nodes, especially with emerging technologies, such as NCFET, where dependencies might change.

This dissertation presents novel approaches to solving these challenges on multiple design levels, providing techniques for analyzing and modeling circuit reliability and low-power design. Techniques are categorized into conventional ((a), (b), (c), and (d)), and unconventional techniques ((e), and (f)) as follows:

(a) Analysis performance gains accompanied with maximizing energy efficiency when operating in the near-threshold region, specifically at optimal energy point. Finding accurately such a point for multicore design is challenging as it changes following the optimization goals and workload as well.

(b) Revealing hidden interdependencies between transistor aging and voltage fluctuation caused by IR-drops. Hence, a novel technique is presented, avoiding under-/over-estimation of timing guardbands, considering these interdependencies towards the smallest, yet sufficient, guardband estimation.

(c) Towards containing transistor aging effects by employing graceful approximation technique, by making circuits faster only on-demand. Aging timing guardband is supplanted by employing approximate computing. The quantization technique is employed as a novel mechanism to maintain accuracy.

(d) Towards containing thermal-induced delay degradation through operating circuits near Zero-Temperature Coefficient (N-ZTC). Operating at N-ZTC minimizes thermal-induced variances in performance and power. Qualitative and quantitative comparisons are presented against traditional timing guardband.

(e) Modeling NCFET-aware power and energy management techniques for NCFET-based processors. NCFET technology has unique properties, that differ from CMOS technology, which makes traditional DVS and DVFS suboptimal. Hence, NCFET-aware power and energy management techniques are indispensably required, which are presented in this dissertation.

(f) Introducing a novel heterogeneous manycore design in NCFET. Such design employs only identical cores. Heterogeneity can be achieved by efficiently employing the optimal configurations. Extending Amdahl's law covering the execution of several new system-specific and application-specific parameters to quantify the benefits of the new design.

Evaluations of the proposed techniques are conducted through implementations and simulations at the circuit level (gate level) using the industrial chip design flow. Additionally, system-level simulators are used to implement and simulate manycore designs. The validation and quantification of the effectiveness of these techniques against state of the art are done through analytical, gate-level, and system-level simulations covering synthetic and real applications.

Zusammenfassung

Die fortlaufende Verkleinerung von Transistor-Strukturgrößen ist einer der wichtigsten Antreiber für das Wachstum in der Halbleitertechnologiebranche. Seit Jahrzehnten erhöhen sich sowohl Integrationsdichte als auch Komplexität von Schaltkreisen und zeigen damit einen fortlaufenden Trend, der sich über alle modernen Fertigungsgrößen erstreckt. Bislang ging das Verkleinern von Transistoren mit einer Verringerung der Versorgungsspannung einher, was zu einer Reduktion der Leistungsaufnahme führte und damit eine gleichbleibenden Leistungsdichte sicherstellte. Doch mit dem Beginn von Strukturgrößen im Nanometerbreich verlangsamte sich die fortlaufende Skalierung. Viele Schwierigkeiten, sowie das Erreichen von physikalischen Grenzen in der Fertigung und Nicht-Idealitäten beim Skalieren der Versorgungsspannung, führten zu einer Zunahme der Leistungsdichte und, damit einhergehend, zu erschwerten Problemen bei der Sicherstellung der Zuverlässigkeit. Dazu zählen, unter anderem, Alterungseffekte in Transistoren sowie übermäßige Hitzeentwicklung, nicht zuletzt durch stärkeres Auftreten von Selbsterhitzungseffekten innerhalb der Transistoren. Damit solche Probleme die Zuverlässigkeit eines Schaltkreises nicht gefährden, werden die internen Signallaufzeiten üblicherweise sehr pessimistisch kalkuliert. Durch den so entstandenen zeitlichen Sicherheitsabstand wird die korrekte Funktionalität des Schaltkreises sichergestellt, allerdings auf Kosten der Performance. Alternativ kann die Zuverlässigkeit des Schaltkreises auch durch andere Techniken erhöht werden, wie zum Beispiel durch Null-Temperatur-Koeffizienten oder Approximate Computing. Wenngleich diese Techniken einen Großteil des üblichen zeitlichen Sicherheitsabstandes einsparen können, bergen sie dennoch weitere Konsequenzen und Kompromisse.

Bleibende Herausforderungen bei der Skalierung von CMOS Technologien führen außerdem zu einem verstärkten Fokus auf vielversprechende Zukunftstechnologien. Ein Beispiel dafür ist der Negative Capacitance Field-Effect Transistor (NCFET), der eine beachtenswerte Leistungssteigerung gegenüber herkömmlichen FinFET Transistoren aufweist und diese in Zukunft ersetzen

könnte. Des Weiteren setzen Entwickler von Schaltkreisen vermehrt auf komplexe, parallele Strukturen statt auf höhere Taktfrequenzen. Diese komplexen Modelle benötigen moderne Power-Management Techniken in allen Aspekten des Designs. Mit dem Auftreten von neuartigen Transistortechnologien (wie zum Beispiel NCFET) müssen diese Power-Management Techniken neu bewertet werden, da sich Abhängigkeiten und Verhältnismäßigkeiten ändern.

Diese Arbeit präsentiert neue Herangehensweisen, sowohl zur Analyse als auch zur Modellierung der Zuverlässigkeit von Schaltkreisen, um zuvor genannte Herausforderungen auf mehreren Designebenen anzugehen. Diese Herangehensweisen unterteilen sich in konventionelle Techniken ((a), (b), (c) und (d)) und unkonventionelle Techniken ((e) und (f)), wie folgt:

(a) Analyse von Leistungszunahmen in Zusammenhang mit der Maximierung von Leistungseffizienz beim Betrieb nahe der Transistor Schwellspannung, insbesondere am optimalen Leistungspunkt. Das genaue Ermitteln eines solchen optimalen Leistungspunkts ist eine besondere Herausforderung bei Multicore Designs, da dieser sich mit den jeweiligen Optimierungszielsetzungen und der Arbeitsbelastung verschiebt.

(b) Aufzeigen versteckter Interdependenzen zwischen Alterungseffekten bei Transistoren und Schwankungen in der Versorgungsspannung durch „IR-drops“. Eine neuartige Technik wird vorgestellt, die sowohl Über- als auch Unterschätzungen bei der Ermittlung des zeitlichen Sicherheitsabstands vermeidet und folglich den kleinsten, dennoch ausreichenden Sicherheitsabstand ermittelt.

(c) Eindämmung von Alterungseffekten bei Transistoren durch „Graceful Approximation“, eine Technik zur Erhöhung der Taktfrequenz bei Bedarf. Der durch Alterungseffekte bedingte zeitlich Sicherheitsabstand wird durch Approximate Computing Techniken ersetzt. Des Weiteren wird Quantisierung verwendet um ausreichend Genauigkeit bei den Berechnungen zu gewährleisten.

(d) Eindämmung von temperaturabhängigen Verschlechterungen der Signallaufzeit durch den Betrieb nahe des Null-Temperatur Koeffizienten (N-ZTC). Der Betrieb bei N-ZTC minimiert temperaturbedingte Abweichungen der Performance und der Leistungsaufnahme. Qualitative und quantitative Vergleiche gegenüber dem traditionellen zeitlichen Sicherheitsabstand werden präsentiert.

(e) Modellierung von Power-Management Techniken für NCFET-basierte Prozessoren. Die NCFET Technologie hat einzigartige Eigenschaften, durch die herkömmliche Verfahren zur Spannungs- und Frequenzskalierungen zur Laufzeit (DVS/DVFS) suboptimale Ergebnisse erzielen. Dies erfordert NCFET-spezifische Power-Management Techniken, die in dieser Arbeit vorgestellt werden.

(f) Vorstellung eines neuartigen heterogenen Multicore Designs in NCFET Technologie. Das Design beinhaltet identische Kerne; Heterogenität entsteht durch die Anwendung der individuellen, optimalen Konfiguration der Kerne. Amdahls Gesetz wird erweitert, um neue system- und anwendungsspezifische Parameter abzudecken und die Vorzüge des neuen Designs aufzuzeigen.

Die Auswertungen der vorgestellten Techniken werden mithilfe von Implementierungen und Simulationen auf Schaltkreisebene (gate-level) durchgeführt. Des Weiteren werden Simulatoren auf Systemebene (system-level) verwendet, um Multicore Designs zu implementieren und zu simulieren. Zur Validierung und Bewertung der Effektivität gegenüber dem Stand der Technik werden analytische, gate-level und system-level Simulationen herangezogen, die sowohl synthetische als auch reale Anwendungen betrachten.

The Big Picture behind this Thesis

The Chair for Embedded Systems (CES) at the Karlsruhe Institute of Technology (KIT) has its core expertise in design and architectures for reliable and low-power embedded systems. Major research projects are dealing with Hardware/Software Co-Design aiming to incorporate the hardware and software technologies and exploit the synergy between the two to optimize and satisfy design constraints such as cost, performance, and power of the embedded systems (e.g., [CES1][CES2][CES3]), and developing low-power design of embedded systems based on power consumption and thermal managements (e.g., [CES4][CES5][CES6][CES13]). In addition, many works (e.g., [CES7][CES8][CES9]) focused on the Network on Chip (NOC) to improve the scalability and the power efficiency of systems-on-chip (SoC). Furthermore, several works (e.g., [CES10][CES11][CES12]) focused on approximate computing presenting a framework for approximate logic synthesis.

Furthermore, the CES played an essential role in the creation of the Priority Program (*Schwerpunktprogramm, SPP 1500*) and "*Invasive Computing*" (InvasIC).

DFG SPP 1500 "Dependable Embedded Systems" The key reason behind the significant focus on reliability at CES is the German national research program "Design and Architectures for Dependable Embedded Systems" (German Research Foundation, DFG SPP 1500) which mainly aims at finding new means to overcome the inherent challenges within the nano-CMOS era (e.g., [CES14][CES15][CES16]). Actually, since feature sizes of transistors began to approach atomic levels while voltage scaling is reaching its fundamental limit, technology scaling reached a point, where optimizing the on-chip systems for reliability is as important as optimizing them for performance, power, and cost.

The main goals of the DFG SPP 1500 Priority Program are [CES17]:

- Technology Abstraction.
- Dependable Hardware Architectures.
- Dependable Embedded Software.
- Operation, Observation, and Adaptation.
- Design Methodologies.

DFG Transregio TR89 – Invasive Computing "InvasIC": The main idea of InvasIC [CES19] is investigating and developing novel paradigm of invasive computing for designing and programming the future of parallel computing systems, where thousand of cores are expected to be integrated on a single chip. InvasIC is composed of different sub-projects to cover the different aspects such as resource-aware programming, hardware requirements to enable invasive computing, software requirements like compilers and operating systems (e.g., [CES18][CES19][CES20]).

Thesis Contributions in the scope of the DFG SPP 1500: As earlier mentioned in the Abstract, this thesis focuses on increasing/improving the reliability of embedded systems with respect to aging effects, IR-drop, and Self-Heating Effects (SHE). This thesis employ novel techniques to improve reliability of systems, such as approximate computing. In fact, these objectives are indeed an integral part of the key challenges that DFG SPP 1500 Priority Program addresses. For instance, the proposed aging and SHE techniques within this thesis contribute to the goal of "Dependable Hardware Architectures". Additionally, the proposed reliability mitigation using approximate computing contributes to the goal of "Technology Abstraction". Finally, the study of emerging technology (NCFET) within the scope of this thesis contributes to the goal of "Dependable Hardware Architectures" and "Design Methodologies".

[CES1] R. Ernst, J. Henkel and T. Benner, "Hardware-software cosynthesis for microcontrollers," in IEEE Design and Test of Computers, Dec. 1993, doi: 10.1109/54.245964.

[CES2] D. Herrmann, J. Henkel and R. Ernst, "An approach to the adaptation of estimated cost parameters in the COSYMA system," Third International Workshop on Hardware/Software Codesign, 1994, doi: 10.1109/HSC.1994.336718.

[CES3] J. Henkel and Yanbing Li, "Energy-conscious HW/SW-partitioning of embedded systems: a case study on an MPEG-2 encoder," Proceedings of the Sixth International Workshop on Hardware/Software Codesign. (CODES/CASHE'98), doi: 10.1109/HSC.1998.666233.

[CES4] Yanbing Li and J. Henkel, "A framework for estimating and minimizing energy dissipation of embedded HW/SW systems," Proceedings 1998 Design and Automation Conference. 35th DAC, 1998, doi: 10.1109/DAC.1998.724464.

[CES5] T. Ebi, M. A. Al Faruque and J. Henkel, "TAPE: Thermal-aware agent-based power econom multi/many-core architectures," 2009 IEEE/ACM International Conference on Computer-Aided Design - Digest of Technical Papers, 2009, doi: 10.1145/1687399.1687457.

[CES6] T. Ebi, D. Kramer, W. Karl and J. Henkel, "Economic learning for thermal-aware power budgeting in many-core architectures," 2011 Proceedings of the Ninth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), 2011, doi: 10.1145/2039370.2039401.

[CES7] J. Henkel, W. Wolf and S. Chakradhar, "On-chip networks: a scalable, communication-centric embedded system design paradigm," 17th International Conference on VLSI Design. Proceedings., 2004, doi: 10.1109/ICVD.2004.1261037.

[CES8] Mohammad Abdullah Al Faruque, T. Ebi and J. Henkel, "Run-time adaptive on-chip communication scheme," 2007 IEEE/ACM International Conference on Computer-Aided Design, 2007, doi: 10.1109/ICCAD.2007.4397239.

[CES9] M. A. Al Faruque, T. Ebi and J. Henkel, "Configurable links for runtime adaptive on-chip communication," 2009 Design, Automation and Test in Europe Conference and Exhibition, 2009, doi: 10.1109/DATE.2009.5090667.

[CES10] M. Shafique, W. Ahmad, R. Hafiz and J. Henkel, "A low latency generic accuracy configurable adder," 2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC), 2015, doi: 10.1145/2744769.2744778.

[CES11] K. Bhardwaj, P. S. Mane and J. Henkel, "Power- and area-efficient Approximate Wallace Tree Multiplier for error-resilient systems," Fifteenth International Symposium on Quality Electronic Design, 2014, doi: 10.1109/ISQED.2014.6783335.

[CES12] S. Salamin, G. Zervakis, O. Spantidi,I. Anagnostopoulos,J. Henkel, H. Amrouch, (2021). "Reliability-Aware Quantization for Anti-Aging NPUs", in IEEE/ACM 24th Design, Automation and Test in Europe Conference (DATE'21), 2021.

[CES13] H. Amrouch and J. Henkel, "Lucid infrared thermography of thermally-constrained processors," 2015 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), 2015, pp. 347-352, doi: 10.1109/ISLPED.2015.7273538.

[CES14] S. Rehman, M. Shafique, F. Kriebel and J. Henkel, "Reliable software for unreliable hardware: Embedded code generation aiming at reliability," 2011 Proceedings of the Ninth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), 2011, doi: 10.1145/2039370.2039408.

[CES15] J. Henkel, L. Bauer, H. Zhang, S. Rehman and M. Shafique, "Multi-layer dependability: From microarchitecture to application level," 2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC), 2014, doi: 10.1145/2593069.2596683.

[CES16] S. Tan, H. Amrouch, T. Kim, Z. Sun, C. Cook, J. Henkel. (2017). Recent advances in EM and BTI induced reliability modeling, analysis and optimization (invited). Integration, the VLSI Journal. 60. 10.1016/j.vlsi.2017.08.009.

[CES17] J. Henkel, L. Bauer, J. Becker, O. Bringmann, U. Brinkschulte, S. Chakraborty, M. Engel, R. Ernst, H. Hartig, L. Hedrich, A. Herkersdorf, R. Kapitza, D. Lohmann, P. Marwedel, M. Platzner, W. Rosenstiel, U. Schlichtmann, O. Spinczyk, M. Tahoori, J. Teich, N. When, H. Wunderlich "Design and architectures for dependable embedded systems," in IEEE International Conference on Hardware-Software Codesign and System Synthesis (CODES+ISSS), pp. 69-78, 2011.

[CES18] S. Kobbe, L. Bauer, D. Lohmann, W. Schröder-Preikschat and J. Henkel, "DistRM: Distributed resource management for on-chip many-core systems," 2011 Proceedings of the Ninth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), 2011, doi: 10.1145/2039370.2039392.

[CES19] J. Henkel, A. Herkersdorf, L. Bauer, T. Wild, M Hubner, R. Pujari, A. Grudnitsky, J. Heisswolf, M. Zaib, B. Vogel,V. Lari, S. Kobbe. (2012). Invasive Manycore Architectures. 10.1109/ASPDAC.2012.6164944.

[CES20] B. Oechslein, J. Schedel, J. Kleinöder, L. Bauer, J. Henkel, D. Lohmann, W. Schroeder-Preikschat. (2011). OctoPOS: A Parallel Operating System for Invasive Computing.

Contents

List of Figures

List of Tables

List of Publications in Journals

[J1] **S. Salamin**, G. Zervakis, Y. Chauhan, J. Henkel, and H. Amrouch, PROTON: Post-synthesis ferROelectric Thickness OptimizatioN for NCFET Circuits in The IEEE Transactions on Circuits and Systems (TCAS-I), submitted.

[J2] **S. Salamin**, G. Zervakis, F. Klemme, H. Kattan, Y. Chauhan, J. Henkel, and H. Amrouch Impact of NCFET Technology on Eliminating the Cooling Cost and Boosting the Efficiency of Google TPU in IEEE Transactions on Computers (TC'21), 2021, doi: 10.1109/TC.2021.3065454.

[J3] Georgios Zervakis, Iraklis Anagnostopoulos, **S. Salamin**, Yogesh S. Chauhan, Jörg Henkel, Hussam Amrouch, Impact of NCFET on Neural Network Accelerators in IEEE Access, 2021.

[J4] **S. Salamin**, V. M. van Santen, M. Rapp, J. Henkel and H. Amrouch, "Minimizing Excess Timing Guardbanding under Transistor Self-Heating though biasing at Zero-Temperature Coefficient," in IEEE Access, doi: 10.1109/ACCESS.2021.3057900.

[J5] **S. Salamin**, M. Rapp, J. Henkel, A. Gerstlauer and H. Amrouch, "Dynamic Power and Energy Management for NCFET-Based Processors," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 39, no. 11, pp. 3361-3372, Nov. 2020, doi: 10.1109/TCAD.2020.3012644.

[J6] **S. Salamin**, M. Rapp, A. Pathania, A. Maity, J. Henkel, T. Mitra and H. Amrouch, "Power-Efficient Heterogeneous Many-Core Design with NCFET Technology," in IEEE Transactions on Computers, doi: 10.1109/TC.2020.3013567.

[J7] H. Amrouch, G. Zervakis, **S. Salamin**, H. Kattan, I. Anagnostopoulos and J. Henkel, "NPU Thermal Management," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 39, no. 11, pp. 3842-3855, Nov. 2020, doi: 10.1109/TCAD.2020.3012753.

[J8] H. Amrouch, **S. Salamin**, G. Pahwa, A. D. Gaidhane, J. Henkel and Y. S. Chauhan, "Unveiling the Impact of IR-Drop on Performance Gain in NCFET-Based Processors," in IEEE Transactions on Electron Devices, vol. 66, no. 7, pp. 3215-3223, July 2019, doi: 10.1109/TED.2019.2916494.

[J9] **S. Salamin**, V. M. Van Santen, H. Amrouch, N. Parihar, S. Mahapatra and J. Henkel, "Modeling the Interdependences Between Voltage Fluctuation and BTI Aging," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 27, no. 7, pp. 1652-1665, July 2019, doi: 10.1109/TVLSI.2019.2899890.

List of Publications in Conferences

[C1] **S. Salamin**, G. Zervakis, O. Spantidi, I. Anagnostopoulos, J. Henkel and H. Amrouch "Reliability-Aware Quantization for Anti-Aging NPUs," 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE), Virtual Conference, Feb 1-5 2021.

[C2] **S. Salamin**, M. Rapp, H. Amrouch, A. Gerstlauer and J. Henkel, "Energy Optimization in NCFET-based Processors," 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, 2020, pp. 630-633, doi: 10.23919/DATE48585.2020.9116301.

[C3] J. Henkel, H. Amrouch, M. Rapp, **S. Salamin**, D. Reis, D. Gao, X. Yin, M. Niemier, C. Zhuo, X. S. Hu, H. -Y. Cheng, and C. -L. Yang, "The Impact of Emerging Technologies on Architectures and System-level Management: Invited Paper," 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), Westminster, CO, USA, 2019, pp. 1-6, doi: 10.1109/ICCAD45719.2019.8942102.

[C4] H. Amrouch, V. Santen, O. Prakash, H. Kattan, **S. Salamin**, S. Thomann, J. Henkel, "Reliability Challenges with Self-Heating and Aging in FinFET Technology," 2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS), Rhodes, Greece, 2019, pp. 68-71, doi: 10.1109/IOLTS.2019.8854405.

[C5] **S. Salamin**, M. Rapp, H. Amrouch, G. Pahwa, Y. Chauhan and J. Henkel, "NCFET-Aware Voltage Scaling," 2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), Lausanne, Switzerland, 2019, pp. 1-6, doi: 10.1109/ISLPED.2019.8824802.

[C6] M. Rapp, **S. Salamin**, H. Amrouch, G. Pahwa, Y. Chauhan and J. Henkel, "Performance, Power and Cooling Trade-Offs with NCFET-based Many-

Cores," 2019 56th ACM/IEEE Design Automation Conference (DAC), Las Vegas, NV, USA, 2019, pp. 1-6.

[C7] **S. Salamin**, H. Amrouch and J. Henkel, "Selecting the Optimal Energy Point in Near-Threshold Computing," 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE), Florence, Italy, 2019, pp. 1691-1696, doi: 10.23919/DATE.2019.8715211.

1. Introduction

Over decades, the cost of a single bit of a semiconductor memory has significantly dropped (more than 100 million times), and such a trend continues. Similarly, the cost of a logic gate has also dramatically dropped. Such a rapid price reduction has stimulated many new application domains where semiconductor devices have improved almost all human activities. Improvements of transistors are of great importance for driving innovations in an ever more digitalized and interconnected world. This, in turn, serves to promote the progress and benefits of human society. The rapidly increased functionality and the reduced cost of the semiconductors chips have brought many benefits to our life (e.g., health care system). Benefits can *only* be utilized if the newly developed technologies are able to perform a reliable operation (i.e., error-free results and failure-free operation).

With the invention of the Integrated Circuit (IC), and as the complementary metal-oxide-semiconductor (CMOS) became the leading technology in VLSI chips[1], chip manufactures have followed a continued trend of shrinking chips (i.e., scaling down area) by making them smaller and more dense following the what so-called *Moore's law*. Gordon Moore made an empirical observation, following Moore's perception, that the number of transistors on a single semiconductor chip doubles roughly every almost a fixed period. Moore's law is a description of the persistent periodic increase in the level of technology (i.e., transistor) shrinking [58][67]. Examples of technology generations are 0.18mm, 0.13mm, 90nm, 65nm, 45nm, and 7nm nodes. Numbers refer to the minimum metal line width of the transistor [164] (see Fig. 1.3). Each time the minimum line width is reduced, a new technology generation or technology node has been developed. With each new node, the feature sizes of the circuit layout are 70% of the previous technology node. Typically, a new technology node is developed every 2-3 years. Such a practice of periodic size reduction is called transistor scaling.

Transistor scaling is widely regarded as the best method to cope with the ever-growing demands for on-chip functionalities. For decades, technology almost follows fixed scaling factor $s \approx 0.7$x (i.e., -30%) of the transistors' dimensions, thus, reducing chips area by 50% (i.e., $0.7 \times 0.7 = 0.49$), every 2-3 years, as illustrated in Fig. 1.1. The figure shows the roadmap of the technology nodes following Moore's law. As a result, the number of transistors nearly doubles every new node due to the fixed scaling factor. Notably, besides transistor

[1] VLSI chip is a group of interconnected transistors that perform a particular function.

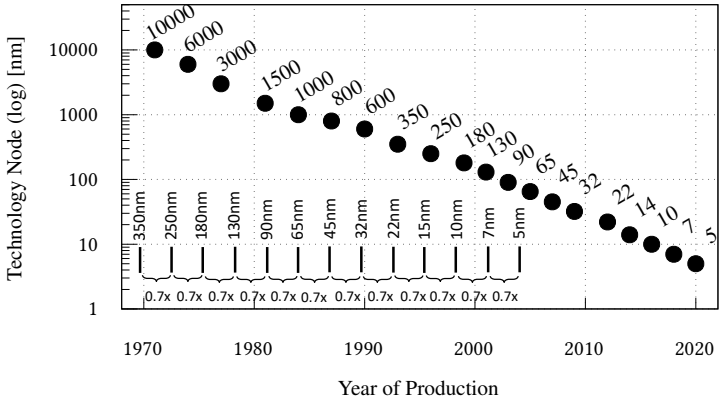


Figure 1.1.: Technology nodes over years where new technology is introduced every 2-3 years. Technology has been scaled almost by constant factor, $s \approx 0.7$.

scaling, CMOS technology has been continuously improved with respect to the geometric shape, structure, and the used materials.

Scaling transistors, for instance, from 65nm to the 45nm technology nodes enables billions of transistors to be fabricated on a single chip, as nearly twice as many transistors can be fabricated on the same silicon wafer (i.e., area) with the new technology node. Hence, the cost per chip is reduced significantly. Fig. 1.2 [146], for instance, shows a comparison of the most important metrics: Area, Delay, and Power of a single invertor over different technology nodes. For illustration purposes, the figure solely shows technologies from 180nm and below. For instance, as shown, the same area of a single invertor on the 180nm technology node can be used, theoretically, to fabricate up to 16, and 512 invertors using the 45nm, and 7nm technology nodes, respectively. Therefore, more logical cells can be manufactured within the same area, significantly increasing the chip's functionality.

However, besides the beneficial impact of the transistor scaling on area, transistors and hence logical cells become faster and more power-efficient, which, in turn, gathering space to revolutionize the advancements of the microprocessor industry. This is because that smaller transistors and shorter interconnects lead to smaller capacitances, and hence, switching delay and power drop [164] (see Eq. (1.1), Eq. (1.2) and Eq. (1.3)). Fig. 1.2 [146] shows the average power

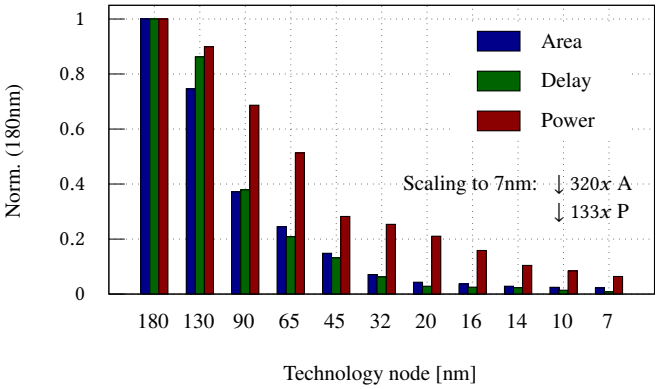


Figure 1.2.: The Area, Delay, and Power of a single invertor over different technology nodes using HSPICE simulator. All values are reduced towards the smaller nodes. Results are normalized to the 180nm technology node.

consumption and delay, using HSPICE simulator [153], of a single gate of an invertor over different technology nodes. Results are normalized to the largest node (180nm). As shown, as the area becomes smaller towards smaller technology nodes, the power and delay are decreased as well. Hence, the performance and functionality of the chip largely increase as the number of transistors that are fabricated on the chip increases, besides reducing manufacturing cost per chip. The rapidly increased functionality, cost reduction, and lower power consumption of the chips have brought many benefits to the end-users as well as to the semiconductor industry. Low manufacturing cost, increased computer processing power, and the ability to execute multiple simultaneous tasks are significant advantages for processor chips achieved by technology scaling.

Notably, area and power have been reduced for many technology nodes almost to the same degree following Dennard's scaling. As Dennard's scaling states, when a transistor is scaled down, the power density (i.e., power per area unit) must remain constant. Therefore, the transistor's power must remain in proportion to its area. This is achieved by scaling the voltage and the current in the ON state (I_{on}) of the transistor to the same degree as the area. Dennard observed that transistor dimensions (i.e., Width, Length, and the oxide layer (t_{ox}), see Fig. 1.3) could be scaled down by s≈0.7x (i.e., reduced by -30%) with every new technology node, and hence, reducing their area by 50%, which is in line with Moore's law. This would reduce transistor delay with the same factor, which, in turn, increases the switching frequency by around 40% (i.e.,

higher performance). However, to keep the power density constant, voltage is also reduced by 30%, which, in turn, reduces power consumption by 50% [24]. In summary, in every new technology node, while transistors density doubles (i.e., area halved or transistors number doubled), the transistor becomes 40% faster, and its power consumption becomes 50% less. As a final result, power density is not affected with scaling (i.e., remains constant). Chip sizes and their supply voltage have scaled with almost the same factor for technology nodes from 0.8μm (5V) through 0.5μm and 0.35μm (both 3.3V), 0.25μm (2.5V), 180nm (1.8V) to 120nm (1.2V) [164]. However, Dennard did not consider the physical limitations and roles of leakage current in smaller technology nodes [164] [24] where maintaining a constant power density could be impossible. Consequently, the increased power density has led to the discontinuation of Dennard's scaling, where the ideal supply voltage reduction could be no longer possible, in addition to many other limitations, as will be explained later.

Due to the discontinuation of Dennard's scaling, the nanometer era (i.e., a transistor in nanometer-scale, $< 1\mu$m) has brought new challenges for chip manufacturing. Even though most of these challenges are well-known phenomena, they become more significant and severe with smaller technology nodes, threatening the entire scaling process. For instance, as transistors become smaller, transistor aging, short-channel effects, Self-Heating effects, IR-Drop, and many other issues such as process variation start playing larger roles. While these challenges originate at the device level, their consequences propagate all the way up to the circuit and system levels. Chip design is essentially a cross-layer implementation requiring deep understanding and teamwork between device engineers, chip designers, system designers, and software engineers. Therefore, the nanometer CMOS designs have become a more complex process and need more focus on the physical design. Importantly, chip designers should account for the consequences of further scaling while overcoming the *scaling limitations* for reliable operation and maintaining the intended performance.

1.1. Scaling Limitations

There are many limitations and challenges behind slowing down or even stopping the scaling process. They can be categorized, within this dissertation,

into physical limitations, material limitations, economic limitations, and power and thermal limitations. These limitations are introduced here.

$$I_{on} \approx \frac{W}{L} C (V_{dd} - V_{th})^2 \tag{1.1}$$

$$f = \frac{1}{t} \propto I_{on} \; ; \; t \propto \frac{1}{(V_{dd} - V_{th})} \tag{1.2}$$

$$P_{dynamic} \approx C V_{dd}^2 f \tag{1.3}$$

$$P_{leakage} = V_{dd} I_{off} \; ; \; I_{off} \approx e^{-V_{th}} \tag{1.4}$$

Where I_{on} is the transistor current in the ON state, I_{off} is the transistor current in the OFF state (i.e., leakage current), W is the transistor width, L is transistor length, C is average switched capacitance, f is the switching frequency of the transistor, t is the delay of the transistor, V_{th} is the threshold voltage of transistor, $P_{dynamic}$ and $P_{leakage}$ are the dynamic and leakage power of the transistor, respectively, and V_{dd} is the supply voltage of transistor [23][60].

1.1.1. Physical Limitations

Physical limitations are due to the increment of leakage currents as the devices are becoming smaller and the shortage of interconnections to keep up with the increased current demands. Such limitations impact the performance and functionality of CMOS devices.

1.1.1.1. Supply and Threshold Voltages

Reducing the supply voltage of the transistor reduces the dynamic power consumption, as shown in Eq. (1.3). However, transistor's performance (switching speed) is proportional to its supply and threshold voltages, V_{dd} and V_{th} respectively (i.e., $\propto (V_{dd} - V_{th})$) as shown in Eq. (1.1) and Eq. (1.2). Hence, when the supply voltage is reduced, the transistor's performance follows. To maintain the performance, the threshold voltage must be reduced by the same degree following the supply voltage, as previously done in many technology nodes [24]. However, nowadays, threshold voltage cannot be reduced further (slightly changed) in order to sustain reliable operation and keep leakage current (I_{off})

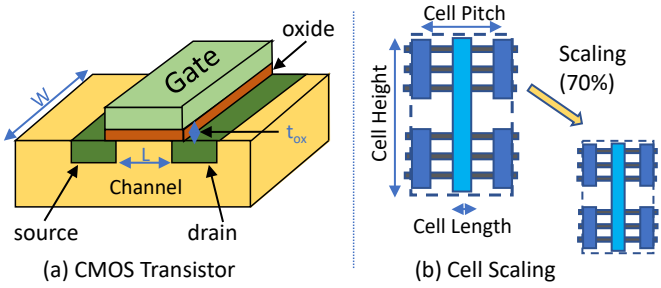


Figure 1.3.: (a) CMOS transistor structure. (b) Standard cell scaling.

at an acceptable level as $I_{off} \propto (V_{dd} - V_{th})$ [60], see Eq. (1.4). When the threshold voltage is reduced, the transistor cannot be completely turned off as it operates in a weak-inversion mode with an increase of the sub-threshold leakage current. For instance, reducing the threshold voltage of 85mV will increase the leakage current by 10× at 65nm technology node [96], which severely degrades the power and performance of the transistor. This has slowed down or even stopped the supply and threshold voltages reduction of transistors. For large chips, nowadays, leakage current could be similar to the switching current [46]. In summary, *the leakage current is an essential factor that limits the step of scaling as it is threshold voltage-dependent. With this limitation, both supply and threshold voltages cannot be reduced anymore.*

1.1.1.2. Gate Dielectric Thickness

The gate electrode with gate dielectric (i.e., the oxide layer t_{ox}, see Fig. 1.3(a)) control the switching operation of CMOS transistors. The voltage of the gate electrode controls the flow of electric current across the transistor. The gate dielectric should be designed as thin as possible in order to maximize the performance gain of the transistor. However, short-channel effects must remain under control when the transistor is turned on and reduces the sub-threshold leakage when a transistor is off. In order to maintain the electric field when the transistor is scaled, the gate dielectric thickness should also be scaled by the same ratio (i.e., $s = 0.7$). The thinner oxide layer raises the leakage current, which increases exponentially when the thickness scales down. With t_{ox} thickness below 2nm, quantum-mechanical tunneling of charge through

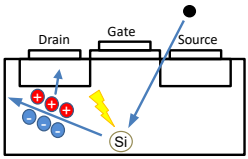
the gate oxide may occur, resulting in more leakage currents and reliability issues[164]. Currently, such a layer comprises only a few layers of atoms and is approaching the fundamental limits of around 1-1.5nm [64].

However, semiconductor technology is hitting the absolute limits to have a thin oxide layer while still acting as an insulator. In turn, scaling this layer further (i.e., thinner layer) with voltage reduction below 0.5V while keeping transistors deterministically behave as intended to be (i.e., as modeled) is becoming a challenge if not even impossible [164].

1.1.1.3. Short Channel Effect

The short channel enables a faster switching frequency of the transistor since a shorter time is required for the current to flow from source to drain in the ON state (i.e., I_{on}). However, a side-effect of the short channel is the high leakage current that flows from source to drain in the OFF state of the transistor (i.e., I_{off}), even if the gate voltage is below the threshold value. Therefore, scaling the channel, through scaling the transistors, would result in substantial increases in leakage current.

1.1.1.4. Soft Error



Soft errors (SER) are radiation-induced faults that happen due to a particle hit. When particles strike the silicon substrate within a transistor, they create hole-electron pairs that collect charges, which results in a transient current pulse. In combinational circuits, such a pulse results in a glitch that might be clocked in, and incorrect data is propagated. In storage elements (e.g., latch), if that pulse is large enough, it flips the stored value. Storage elements are mainly small node capacitances. As the technology node scales down, the node capacitances decrease, leading to an increase in soft error susceptibility. Moreover, the aggressive voltage scaling, to save dynamic power, results in lowering the pulse level, which is the reason behind increasing SER [34]. Therefore, the supply voltage should not be reduced below a certain level, especially for storage elements.

1.1.1.5. On-chip Metal Interconnections

While transistors are becoming smaller and closer to each other, besides the increasing number of transistors, on a single chip, narrower and longer metal interconnectors are then needed with every new technology node. Hence, this, in turn, increases the resistances and capacitances of the on-chip metal interconnectors [38]. The increasing resistances lead to higher voltage supply level drops in addition to the increased capacitance values, which cause more crosstalk noises. The combination of both leads to more considerable signal propagation delays, limiting the possible performance gain and might cause unreliable operation of the chip besides more heat dissipation. For instance, At 1GHz switching speed on 16nm technology node, signal rise and fall times must be less than 50ps within the available 1ns time frame. For such a case, even a line length of 3mm and above will become critical and may result in a large propagation delay [164]. Therefore, while transistors become faster through the transistor scaling, the on-chip interconnections result in a considerable increase in the propagation delays, dominating the overall performance gain [164]. For instance, Fig. 1.8 shows how the impact of the interconnection parasitics on circuit's delay (RC Delay) is becoming more significant with smaller technology nodes, which is mainly dominating the circuit's delay. Thus, metal interconnections start dominating the chip's performance, reliability, and signal integrity on smaller technology nodes [90].

1.1.2. Material Limitations

Material limitations are due to the inability of the materials, that are used for the dielectric and metal interconnections, to provide reliable isolation and conduction with further transistor scaling.

Designers always work to introduce new materials in order to keep up with transistors scaling to ensure the reliability of transistors as well as improve their performance [56]. Materials (e.g., silicon (Si), aluminum (Al), and copper (Cu)) are limited by their physical capabilities, such as carrier mobility and conductivity. When their physical limits are reached, transistors will not be improved (e.g., better performance or lower leakage current) further with scaling [141]. For instance, SiO_2 gate dielectric reliability degrades as it becomes thinner [49]. Therefore, high-k gate dielectric has been used in 45nm technology to replace SiO_2 [90]. The high-k materials enable better control of

the leakage current when the dielectric becomes thinner to support physical scaling [56]. On the other hand, Cu is less sensitive to electromigration than Al. Hence, Cu replaces Al as Al is more susceptible to defects when used as interconnect wires [141][56].

1.1.3. Economic Limitations

Economic limitations result from the increase of the fabrication (i.e., tapping out) and testing costs that may not be economically profitable for the semiconductor industry.

The rising cost in chip manufacturing is attributed to production and testing processes, which increases exponentially with transistor scaling. The adaptation of a new technological node is very costly. For instance, the average design costs (i.e., initial setup and test) for the 28nm chip may rise to 100–200M US$. Assuming the chip is for a consumer application with 1 US$ profit per device, then at least a volume of 100–200 million products is required to reach the break-even with respect to the development costs [164]. For smaller technology nodes, the total development costs will increase further. Smaller size circuit is more vulnerable to hard and soft defects (e.g., process variations). Hence, circuits should be carefully tested to guarantee the required performance and reliable operation. However, with scaling, more sophisticated test methods will incur additional testing steps and time and thus increasing test cost [56]. In short, in many application areas, the transition to the next technology node may no longer be economically profitable.

1.1.4. Power and Thermal Limitations

The power and thermal limitations are the results of the ever-increasing number of transistors in the area, which, in turn, demands higher energy consumption, and hence more heat is generated.

For instance, Fig. 1.2 shows that scaling an invertor from 180nm towards 7nm has led to 320x area reduction besides only 133x power reduction. Thus, power is not reduced with the same factor as area. Which, in turn, led to an increase in power density instead of constant density. In addition to many others, the reason above has led to the discontinuation of Dennard's scaling. Fig. 1.4 [137] shows how power densities increase with technology scaling, where an

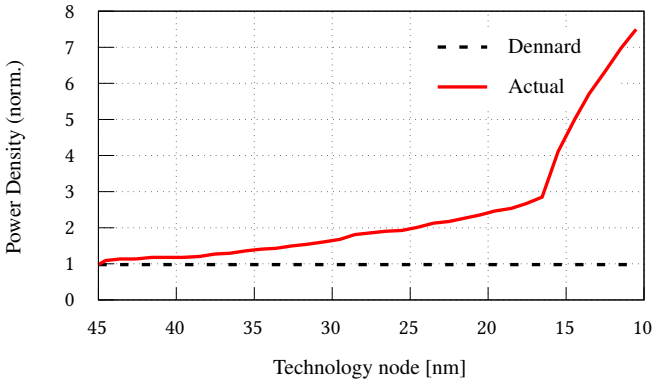


Figure 1.4.: Power densities for different technology nodes. Dennard's trend no longer holds where power density of the chip is continuously increasing.

inflection point appears at 45nm node, where Dennard's trend no longer holds. This was especially noticeable when the industry switched to using high-k dielectrics and metal gates with technology nodes 45nm and beyond [90].

However, as the generated heat of the CMOS chip is proportional to its power consumption, larger power consumption leads to rising the generated heat. Therefore, the increasing power consumption and power density, due to the transistor scaling, result in higher on-chip temperatures and presenting heat removal challenges. In short, although scaling has enabled billions of transistors to be integrated on a single chip, the growth in such integration rate contributes to the power and thermal problems. This negatively impacts the performance and reliability of CMOS transistors, and hence the chip, as the elevated power density and the elevated temperature are the main reasons behind aggravating many *reliability issues*.

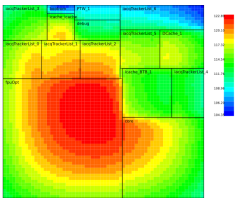
1.2. Reliability Issues

While transistor size scales faster than its power in addition to higher performance (higher frequency), with every new technology node, more power-hungry chips with increasingly on-chip power densities appear. This, in turn,

has aggravated many reliability issues. Reliability issues are stimulated directly or indirectly by excessive power densities. Nowadays, chips are constrained by their reliability and power to maintain error-free operation during their lifetime. In the following, the most critical reliability issues are introduced, which also are considered in this dissertation.

1.2.1. Excessive Temperature:

When the transistor switches ON, it consumes power and hence generates heat. Temperature affects two key parameters in transistor: threshold voltage (V_{th}) and carrier mobility (μ) [168]. Both parameters can be modeled as functions of temperature as shown in Eq. (1.5) and Eq. (1.6) [70]. As shown from equations, increasing the temperature will negatively affect the transistor's performance, and hence the chip, making it weaker and slower.



$$\mu(T) = \mu(T_{ambient})(\frac{T_{ambient}}{T})^m \tag{1.5}$$

$$V_{th}(T) = V_{th}(T_{ambient}) - k(T - T_{ambient}) \tag{1.6}$$

Where $T_{ambient}$ is the room temperature in Kelvin, m and k are positive constants, and T is transistor temperature.

The growing power density has elevated the on-chip temperature. Elevated on-chip temperatures vigorously agitate the chip's reliability as they affect the key characteristics of its transistors (e.g., switching speed and leakage current), which increases their susceptibility to timing violations, and hence runtime failures. For instance, high temperatures rapidly stimulate transistor aging phenomena with higher aging-induced reliability degradations [8]. Therefore, this increases the demand for chip-level cooling solutions to remove the excessive temperature and maintain a reliable operation.

1.2.2. BTI and HCI Transistor Aging:

Transistor aging: Are the results of electromigration and charge trapping within the transistor's channel, which manifests itself as a degradation in the electrical characteristics of the transistor (e.g., Threshold voltage (V_{th})) [127]. Such degradation makes transistors weaker and slower, increasing the susceptibility to timing violations, and hence runtime failures [13]. Aging in the form of Bias Temperature Instability (BTI) and HotCarrier Injection (HCI) are stimulated by higher temperatures, higher voltage, and continuous stress. The aging effect has become a critical issue as the size of transistors is scaling down while the supply voltage is not scaling.

Bias Temperature Instability (BTI): BTI is a two-phases mechanism; stress phase and partial recovery phase. Negative BTI (NBTI) occurs in the P-channel metal–oxide–semiconductor (PMOS) transistors, and Positive BTI (PBTI) occurs in the N-channel metal–oxide–semiconductor (NMOS) transistors. When the PMOS transistor switches ON, the stress voltage breaks the bound of Si-H at the interface. The separated hydrogen atoms combine into H_2 form, which diffuses towards the gate of the transistor. Broken Si-H bonds generate positively charged traps which increases the threshold voltage of the transistor (V_{th}). When PMOS switches OFF, where stress voltage is removed, the recovery phase starts where some of the traps are released, and some of the broken Si-H bonds heal. On the other hand, when the NMOS transistor is ON, electrons are trapped within the gate dielectric resulting in PBTI. During the recovery phase, these electrons are released and partially return to the channel, which leads to partially recover V_{th} [11, 13, 163].

Hot Carrier Injection (HCI): HCI mechanism is more prominent in NMOS transistors. The primary source of the hot carriers is the heating inside the channel of the transistor during operation. When the transistor switches ON, the high-energy carriers strike with other atoms and carriers in the transistor's channel. These energetic carriers can impact ionization within the substrate and the generated electrons or holes inside the channel. During this process, the injected carriers can generate interface or oxide defects, and as a result, the MOSFET characteristics are affected, e.g., threshold voltage [127, 163].

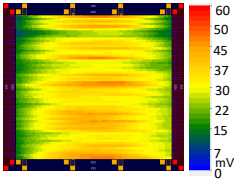
The primary sources of the aforementioned aging mechanisms in transistors is stress and high supply voltage. Aging rate increases in high temperatures. However, aging has been aggravated through transistor scaling, as can be seen

in Fig. 1.6. Further details regarding aging are provided in Section 2.2, and Chapter 4 and Chapter 5.

1.2.3. Voltage Drop (IR-drop):

In a semiconductor chip, the Power Delivery Network (PDN) is used to deliver supply voltage and ground from pads, which are physically connected to the external power supply, to all chip's components (e.g., logic cells) [53]. For instance, standard logic cells are placed in parallel rows such that power and ground lanes are distributed in pairs between cells. Metal interconnection layers are used for both PDNs and signals. Typically, PDN lanes are thicker and wider to have significantly lower resistance to ensure a good flow of current to the components [164]. The major challenge of chip design is obtaining a robust power grid by having sufficient power lanes to deliver the current demands efficiently. This is not only impossible due to the physical characteristics of the metal lanes (i.e., resistances and capacitances), but also this is very expensive and a waste of resources (e.g., area). Contrary, thinner lanes save resources, but PDN might not deliver sufficient currents to all cells [127].

Transistors scaling has increased the density of the transistors, which doubles with every new technology node. In turn, such an increase makes double the number of transistors to share the same PDN portions, which significantly increases the current demands on PDN. Static voltage drop can be analyzed relying on DC analysis of the PDN considering the average power of the chip to estimate a constant driving current in every cell within the chip. Static analysis guides designers early to further optimize the PDN (e.g., using wider power lanes, increase the number of lanes) towards minimizing static voltage drop as much as possible. Furthermore, the simultaneous switching activities in the chip causing enormous current demands, and current peaks occur. These currents cause undesirable dynamic voltage drops across PDN.

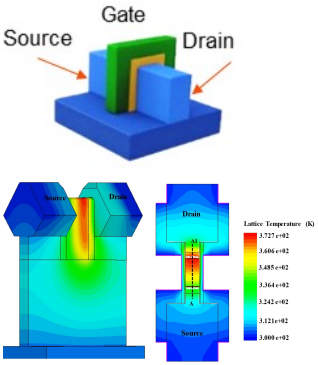


IR-drop: In power lanes is a fundamental property of any circuit due to the non-ideality in PDN, which originates from the fundamental parasitics (i.e., resistances and capacitances) of the power wires [127]. IR-drop results in sudden changes in the supply voltages (i.e., lower than nominal) that chip components receive. These changes are undesirable because some of the supplied energy is dissipated in addition to performance loss

resulted from lowering supply voltage (see Eq. (1.2)). With each clock, the simultaneous switching of cells, sharing the same power lanes, results in peaks of current demands. Due to the inability of PDN to deliver current demands, voltage drops at the terminals of cells leading to an increase in their delays [53]. However, with transistor scaling, the density of the transistors increases. Hence, the current demand on PDN increases with every new technology node. Further details are provided in Section 2.2 and Chapter 4.

1.2.4. Self-Heating Effect:

A fin field-effect transistor (FinFET) is a multigate metal-oxide-semiconductor field-effect transistor (MOSFET) built on a substrate. The gate wraps the channel on two, three, or four sides, forming a double gate structure. FinFET's invention represented a significant leap forward in the semiconductor industry, as it has brought many advantages to the semiconductor industry, especially in the nanometer era. The FinFET devices have significantly faster switching times and higher current density than the planar CMOS technology [149] (see Fig. 1.3).

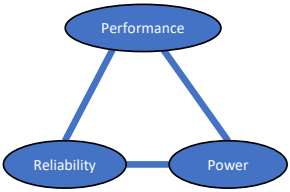


[135]

The switching to the FinFET technology has resulted in better control of the channel via the gate potential. Which, in turn, has helped to reduce the leakage current (I_{off}) as well as improve the performance of the transistors [149]. The channel of FinFET is encapsulated by the gate dielectric, which is also a thermal insulator. In turn, this has worsened some reliability phenomena. Self-Heating Effect (SHE) refers to elevated channel temperatures and their impact on the performance and leakage power of the transistor [12]. The channel temperature is elevated due to the Joule heating by the current flow through the channel [126]. Due to the thermal insulation of the gate, most of the heat generated within the transistor's channel remains inside the channel. Over time, the temperature is slowly dissipated to the body of the transistor. Elevated channel temperature reduces the drain current I_d in the ON state (i.e., slower operation) and increases the leakage current I_{off} in the OFF state. Additionally, elevated temperatures accelerate

other reliability phenomena, e.g., Aging. Further details are provided in Section 2.4 and Chapter 6.

1.3. Performance, Power, and Reliability



Performance, low power consumption, and reliable operation are the most critical design characteristics of any chip. Improving or sustaining their levels is the ultimate goal for designers with every new technology node. Such characteristics are wholly dependent where changing of one could influence the others. For instance, improving the chip's performance could increase the power consumption and worsen its reliability. Nowadays, we are witnessing an increasing demand for low-power computing with the raising of many new application domains, such as wearable and IoT devices. In such domains, devices are battery-powered, and hence power consumption could be more critical than the provided performance. Additionally, designers always work to reduce the power consumption in important computational nodes (e.g., supercomputers and data centers), where an immense amount of energy is required for both operating and cooling these nodes. Therefore, the ultimate design goal is maximizing performance at minimal power towards *power-efficient* devices. In many cases, performance and power have to be traded, as they influence each other.

Fig. 1.5 [164] shows the improvements in various performance parameters over the last decades. As shown, over the years, technology scaling is very beneficial for frequency as well as power efficiency (i.e., performance-per-unit of power). The figure also shows that after 2010, for technologies below 65nm, power efficiency and frequency improvements are very limited. Notably, the leakage power has increased, which, as discussed previously, is the main reason behind limiting the supply and threshold voltages scaling, leading to continuously increasing power density. These trends have essential impacts on chip design. For instance, when it comes to high-performance microprocessors, this has led to a substitutional design by switching from high-frequency operation towards complex multi-core architectures at lower frequencies. Such an intricate design helps to increase the chip's power efficiency by improving the computational

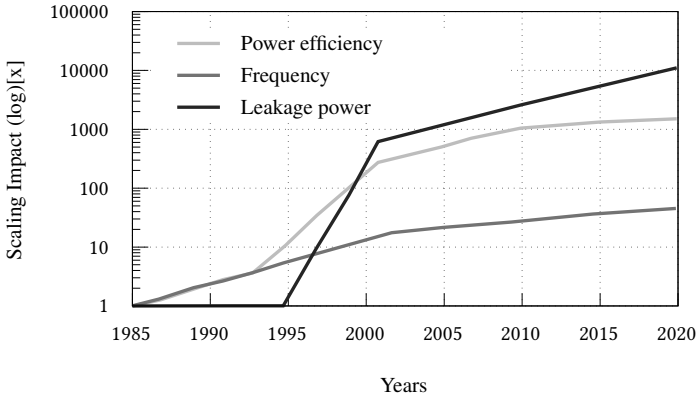


Figure 1.5.: Improvements in various chip's performance parameters over decades. Technology scaling is beneficial for frequency and power efficiency of the semiconductor chip, while leakage power is continuously increasing.

performance without the need to operate at higher frequencies. Moreover, to improve power efficiency, *power management* at all levels (i.e., device, circuit, and system levels) of the chip design has become indispensable for low-power computing devices.

On the other hand, Fig. 1.6 [164] shows various reliability parameters (issues) and how they have been aggravated with technology scaling. As the reliability issues inversely effect transistor's performance (i.e., switching speed), their contributions to the final chip's delay are continuously increasing. For instance, Fig. 1.7 shows the delay contributions of different characteristics over different technology nodes [150]. As shown, on the one hand, the delay contribution of cells (gates) is decreasing towards smaller technology nodes as transistors become faster. On the other hand, the performance loss is observed due to the increasing delay contributions of various reliability parameters (e.g., IR-Drop and interconnections) and the applied pessimistic mitigation techniques, where large timing guardbands are employed (see Section 1.4).

In summary, using the existing CMOS technology is inevitably approaching the limit of attainable power efficiency due to the fundamental limitations in scaling. This is due to the fact that the current technology is almost not scaling, the voltage remains practically unchanged, the performance and reliability are

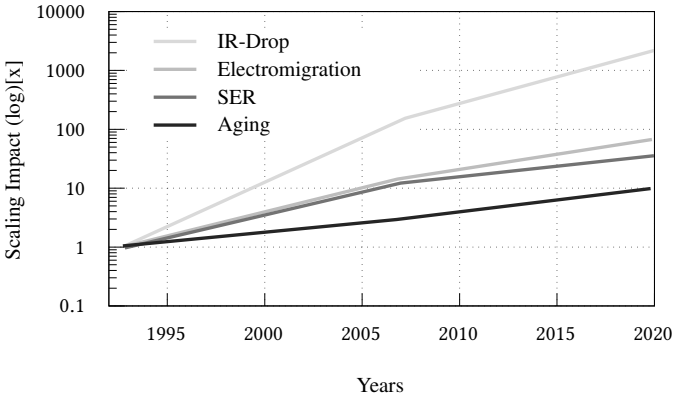


Figure 1.6.: Trends of different reliability parameters over the last decades, and how they are aggravated with technology scaling.

degrading, and the power density keeps increasing. Under these circumstances, designers started looking to optimize technology nodes instead of scaling. This indeed requires new techniques to sustain reliability, improve performance, and ultimately reduce power, which are mainly the goals of this dissertation. In addition, emerging technology, as supplant to CMOS, could also eliminate or solve the existing limitations.

1.4. Mitigation and Optimization Techniques

Various techniques have been proposed to mitigate the impacts of the reliability issues, minimize power consumption, and improve the performance of digital chips. Techniques employed and evaluated in this dissertation are classified into two main categories; conventional and unconventional techniques.

1.4.1. Conventional techniques

Conventional techniques are aiming at maximizing the utilization of the current technology node instead of further scaling. For the last decades, many tech-

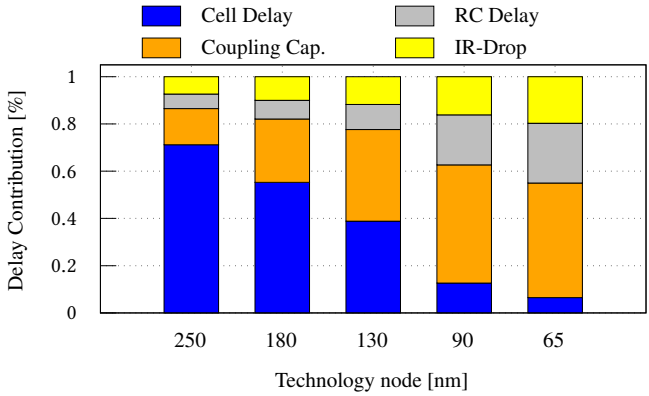


Figure 1.7.: Contribution of different components to chip's delay. Cell's delay decreases with technology scale, while other components become more significant with increasing contributions.

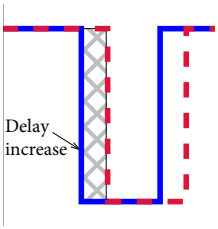
niques have been proposed to optimize operation under the current technology nodes and maximize the possible gains. Techniques vary with complexity and design level. This dissertation studied different techniques targeting CMOS technology to improve reliability, performance, and power consumption. It provides novel solutions improving the existing approaches for CMOS technology in terms of reliability and power consumption with respect to providing sufficiently accurate reliability estimation, proposing new techniques to improve reliability, avoiding pessimistic worst-case analyses, and minimizing the power consumption of CMOS circuits. Techniques as follows.

Near-Threshold Computing (NTC): Works by aggressively reducing the chip's supply voltage to approximately equal to the threshold voltage of its transistors. This operating region retains much of the power and energy savings. Total power is the sum of dynamic (i.e., switching) and leakage power. Dynamic power has quadratic proportional to voltage, while leakage is almost linearly proportional to voltage [123] (see Eq. (1.3) and Eq. (1.4)). With voltage reduction, significant power and energy savings can be obtained at the cost of performance, as performance is also proportional to voltage (see Eq. (1.2)). However, for many applications, where battery life is incredibly more important than performance, such as IoT devices, this is an acceptable trade-off. As the voltage reaches close towards the transistor threshold voltage (V_{th}), an inflection point is observed where leakage energy starts to increase (i.e., leakage increases stronger than dynamic decreases), dominating the total

energy consumption. At such a point, a trade-off is achieved where the optimal combination of leakage and dynamic power is found. This point is called the optimal energy point (OEP). For instance, Intel in 2012 demonstrated a complete x86 microprocessor operating in NTC that consumes around 2mW and could run a tiny solar panel [43]. Even though NTC was explored intensively for a long time, it was not employed in reality because of the significant performance loss, making the chip unusable. To recover such a performance loss, more parallelized functional units can be employed, which might in turn affect OEP. In parallel processors, finding OEP is challenging as it changes following optimization goals, the number of parallel processors and the workload being executed. This dissertation provides an analysis of the performance gains when employing parallelized NTC, showing how OEP could change following the optimization goals. Details in Section 2.1 and Chapter 3.

Efficient guardbanding: The delay of a chip is defined by the maximum (slowest) delay among all its timing paths. The main consequence of most reliability issues is altering the electrical characteristics of the transistors, which in turn degrades transistors' switching delay (i.e., delay increase). However, to protect chips against such degradation and thus sustain a reliable operation, a timing guardband (i.e., extra time slack) must be included on top of the chips delay to prevent unpredictable runtime timing violations [126].

However, such guardband could be significantly large, leading to a performance loss, as the chip will operate at a lower frequency than its full potential. Moreover, the contributions of timing guardbands to the chip's delay have increased with technology scaling, as illustrated in Fig. 1.7 [150], where guardbands can dominate the chip's performance, despite the improvement in cells' delay. Therefore, sufficiently accurate reliability estimation for *efficient guardbanding* while sustaining reliable operation is vitally important.



Overestimating guardband leads to extra performance loss, while underestimating leads to unreliable operation due to timing violations. For different reliability issues, designers consider them as independent issues where guardbands are individually estimated. The final guardband is, therefore, the magnitude of the sum of all guardbands. Even though such a technique might protect chips against timing violations, neglecting the interdependence between reliability

issues could be another issue. For instance, IR-drop could help in recovering, to some degree, aging degradation. Oppositely, aging worsens IR-drop. Therefore, the required timing guardband under both phenomena cannot be the magnitude of summing both guardbands. Hence, the smallest, yet sufficient, guardband needs to be constantly considered, and designers should not neglect similar dependencies. This dissertation presents a novel technique, considering correlations between many reliability issues towards the smallest, yet sufficient, guardband estimation, avoiding under-/over-estimation of timing guardbands. Further details in Section 2.2 and Chapter 4.

Approximate computing: Recently, eliminating timing guardband while maintaining reliable operation is of great interest for modern chips. Thus, this will help to sustain reliability without sacrificing performance. For instance, it will be more efficient to contain aging-induced timing degradation, rather than using a timing guardband, by gradually compensating for the increased delay. This can be achieved by making the circuit faster when needed. One of the most efficient techniques to achieve that is by applying approximate computing. Approximate computing can be employed to address aging in error-tolerant applications, which exploits the inherent error resilience of several applications, to trade-off computational accuracy with delay. Aging-aware approximate computing introduces directed approximations to improve a circuit's performance and mitigate the aging effects. This dissertation presents a technique to replace the aging timing guardband by approximate computing, employing quantization as a novel mechanism to maintain accuracy. Further details are presented in Section 2.3 and Chapter 5.

Zero-Temperature Coefficient: Thermal-induced timing degradation and the accompanying increase in leakage power can be efficiently mitigated by operating circuits at a lower voltage, specifically at Zero-Temperature Coefficient (ZTC) point. ZTC is a point where the temperature has no impact on the transistor's delay. Therefore, by operating at ZTC, the temperature does not affect the delay of the circuit. For instance, Self-Heating Effects (SHE) is a fundamental obstacle for current technologies and future transistor structures. It results in excessive temperatures across the transistor's channel, which severely degrading the switching speed and increasing leakage power. To sustain reliability and prevent timing errors, large timing guardbands are necessary, which leads to considerable performance losses. ZTC point is well-suited to minimize SHE impacts on the circuit's delay. However, lowering the supply voltage would result in an extra performance loss, and, therefore, a trade-off has to be found. This dissertation proposes to operate at N-ZTC to

suppress thermal-induced variances in performance and power. Further details are presented in Section 2.4 and Chapter 6.

1.4.2. Unconventional techniques

Unconventional techniques provide new approaches that consider reducing power consumption through supply voltage reduction, which trimming or eliminating, to some degree, reliability issues. Unlike conventional techniques, these techniques rely mainly on substituting the CMOS technology, as emerging technologies might overcome CMOS limitations.

Emerging technology: Designers work on inventing and designing new technologies that can overcome CMOS limitations. While many emerging technologies are proposed to supplant CMOS technology, none have proved to be technologically and economically a replacement to CMOS technology, yet. This mainly results from the technical difficulties of reconstructing the fabrication process to the new technology as well as the large cost required for adaptation. Negative Capacitance Field Effect Transistor (NCFET) technology is one of the best candidates to supplant CMOS technology. Recently, GlobalFoundries has demonstrated the compatibility of NCFET with the existing CMOS fabrication process. They fabricated the first NCFET-based circuit using their industrial 14nm FinFET technology [78]. NCFET shows the ability to rebound Dennard's scaling through supply voltage lowering while showing relatively high performance. Even NCFET might overcome existing limitations, it might bring newer challenges. For instance, voltage reduction in NCFET could increase the leakage power, rather than decrease, due to the negative Drain-Induced Barrier Lowering (DIBL) effect [105]. Hence, this could result in a novel trade-off between dynamic and leakage power in NCFET-based chips, which breaks the intuitive voltage selection (i.e., minimum voltage) when it comes to power and energy management techniques. This necessitates developing NCFET-aware power and energy management techniques instead of the conventional ones, which consider the optimization keys that NCFET brings. This dissertation presents the first NCFET-aware power and energy management techniques and the corresponding algorithms. Such techniques make the conventional DVS and DVFS aware of NCFET unique properties in order to maximize the intended performance under minimal power and energy consumption. In addition, it presents a novel NCFET-based heterogeneous

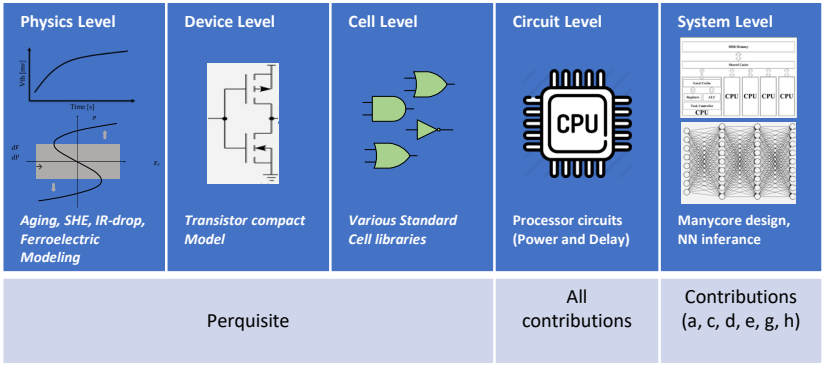


Figure 1.8.: The contributions provided in this dissertation are at different implementation layers.

manycore design, eliminating the associated overheads due to different microarchitectures in conventional heterogeneous manycore design. The dissertation extends Amdahl's law, covering the execution of several new system-specific and application-specific parameters of the new design. Further details in Section 2.5 and Part II.

1.5. Contributions of This Dissertation

The key objective of this dissertation is to investigate the challenges of providing sufficiently accurate reliability estimation along with proposing new techniques to improve the reliability of circuits, avoiding pessimistic worst-case estimations, with respect to transistor aging effects, IR-Drop, and Self-Heating effects as well as providing a comprehensive analysis on low power design, minimizing circuit's power, and ultimately increasing the power efficiency of circuits. Fig. 1.8 summarizes the contributions of this dissertation at different circuit implementation and design levels.

In particular, the novel contributions within this dissertation are as follows:

(a) Optimal Energy Point in parallelized Near-Threshold Computing: This dissertation provides analyses of the performance gains when employing parallelized Near-Threshold Computing (NTC), instead of single core, maximizing the energy efficiency of the parallelized NTC by operating at Optimal

Energy Point (OEP). It demonstrates how the optimization goals and running applications affect OEP in the scope of parallelized NTC, revealing that OEP is not necessarily within the near-threshold region. Details in Chapter 3, publication [123].

(b) Modeling Interdependencies between Voltage Fluctuation and Aging: This dissertation reveals the joint impact that transistor aging in conjunction with IR-drop has on the delay of circuits. It demonstrates the existing interdependencies between them and how they can amplify and mitigate each other, by employing a physics-based aging model, which is able to precisely consider fluctuations in supply voltage. It provides a novel approach to accurately estimate the required timing guardbands to sustain reliability under both phenomena for the projected lifetime of the circuit. Details in Chapter 4, publication [12, 127].

(c) Reliability-Aware Quantization for Anti-Aging: This dissertation presents a graceful approximation technique that, over time, suppresses transistor aging effects. Such technique is able to eliminate aging guardbands and associated performance loss with the smallest, yet acceptable, accuracy degradation. It employs the quantization technique as a novel mechanism to eliminate aging effects in circuits with tolerable accuracy. Details in Chapter 5, publications [130, 12].

(d) Minimizing Excessive Timing Guardband under Transistor Self-Heating: This dissertation presents a novel technique to analyze the impacts of Self-Heating Effects (SHE) on both delay (timing) and power of a full processor and multi-core design. It proposes a novel technique to mitigate SHE by operating circuits near zero-temperature coefficient (N-ZTC), minimizing SHE-induced variance in performance and power. The technique provides a methodology to approximate ZTC of circuits through the superposition of distinct ZTCs of all subcircuits within the circuit. Details in Chapter 6, publication [12, 126].

(e) Revealing Unique Properties of NCFET-based Circuit: This dissertation models NCFET-based processors, demonstrating that voltage scaling leads to a novel runtime trade-off between leakage and dynamic power, resulting in an optimal operating point at which total power is minimized. It shows how in NCFET-based processor, optimal voltage selection, required to minimize the total power consumption, is workload-dependent and follows the share of leakage power from total power. Details in Chapter 7, publications [128, 115].

(f) Power and Energy Management Techniques for NCFET-based Processors: This dissertation, based on the contribution mentioned above, presents comprehensive power and energy models for NCFET-based processors. These models enable the exploration of the simultaneous impacts of multiple optimization keys on power and energy of NCFET circuits. It presents novel NCFET-aware DVS and DVFS algorithms for power and energy minimization by selecting the optimal voltage/frequency pair at runtime while considering the characteristics of the running workloads. Details in Chapter 8, publications [128, 125, 124].

(g) Power-Efficient Heterogeneous Manycore Design for NCFET: This dissertation presents a novel NCFET-based heterogeneous manycore design, eliminating the associated overheads due to different microarchitectures in conventional heterogeneous manycore design, in order to maximize the power and energy efficiency. Such heterogeneity can be achieved by optimally selecting the correct configurations that make one core, at least, a super-core while the remaining are efficient cores. The dissertation extends Amdahl's law, providing comprehensive analysis to explore the simultaneous optimization corners at design-time, covering the execution of several new system-specific and application-specific parameters, quantifying the potential benefits of the new design. Details in Chapter 9, publications [128, 125, 124, 129].

1.6. Dissertation Outline

Before the contributions of this dissertation are presented in detail, Chapter 2 gives an overview of the preliminary backgrounds which are needed for a better understanding of dissertation contributions. These backgrounds cover Near Threshold Computing (NTC), transistor aging phenomena, IR-Drop, Approximate computing, Self-Heating Effects (SHE), and Zero-Temperature Coefficient (ZTC) as well as Negative capacitance Field Effect Transistor (NCFET). In addition, it discusses the state-of-the-art techniques.

This dissertation is in two parts. **Part (1) Reliability and Power under Conventional CMOS,** which consists of:

Chapter 3 presents investigations of how the performance loss can be compensated through parallelized computing when operating processors within the

Near-threshold region. Also, it shows how the Optimal Energy Point could be affected as the number of cores increases under different optimization goals.

Chapter 4 presents a technique for sustaining the reliability of processors under the interdependences between aging and IR-Drop phenomena. The technique considers the hidden correlations between these phenomena in order to effectively estimate the smallest, yet sufficient, timing guardband to protect the processor against timing violations at runtime. The effectiveness of the technique is evaluated along with comparisons to state-of-the-art techniques, illustrating how other techniques under- overestimate the guardbands.

Chapter 5 presents the reliability-aware quantization technique to mitigate aging effects in NPUs. The technique works by eliminating the aging timing guardbands. Such technique delivers a graceful accuracy degradation over time, while compensating for the aging-induced delay increase of the NPU. The effectiveness and the efficiency of the proposed technique are evaluated in sustaining a reliable operation for a set of state-of-the-art neural networks.

Chapter 6 presents a novel mitigation technique that protects the processor against Self-Heating Effect (SHE) while eliminating the accompanied large timing guardband. The technique exploits the potential in mitigating the thermal variations within the chip through operating the processor near Zero-Temperature Coefficient (N-ZTC) to contain the delay increases due to SHE.

Part (2) Low Power Computing: The Negative Capacitance Approach, which consists of:
Chapter 7 provides an implementation of NCFET-based processor chips. Following the standard chip design flow, it presents a cross-layer implementation, from transistor level all the way up to full chip. This also includes delay and power analysis of the final chip by applying timing and power signoffs.

Chapter 8 presents NCFET-aware power and energy management techniques based on Dynamic Voltage Scaling (DVS) and Dynamic Voltage and Frequency Scaling (DVFS). These techniques consider the unique properties of NCFET technology to ultimately minimize power/energy. It also provides mathematical optimization modeling for analytical design space exploration as well as gate-level simulations for accurate analyses. Analyses are presented along with the evaluation of the effectiveness of the proposed techniques in comparison with state-of-the-art techniques.

Chapter 9 discusses the novel heterogeneous manycore design based on NCFET. This chapter extends Amdhal's law covering the execution of sev-

eral new system-specific and application-specific parameters to quantify the potential benefits of the new design. Also, it provides a qualitative and quantitative comparison of the conventional heterogeneous design with NCFET-based design in terms of performance and power efficiency.

Finally, Chapter 10 concludes this dissertation and gives an outlook.

2. Preliminary Background and Related Works

The rapid technology scaling and evolution has brought many advantages in diverse aspects of IC manufacturing. However, the aggressive scaling has steadily increased the power density of ICs and increased their susceptibility to failures due to various kinds of reliability issues. In the following, preliminary background details are presented, which are required before presenting the contributions of this dissertation as well as discuss the state-of-the-art methodologies in estimating and evaluating each part.

2.1. Near-Threshold Computing (NTC)

The discontinuation of Dennard's scaling, where the supply voltage is almost unchanged, was the main reason behind not reducing the energy per operation of microprocessors. However, energy efficiency and low power consumption are essential to continue delivering logic throughput with much less energy consumption. The increasing demand to maximize the energy efficiency of devices has pushed Near-Threshold Computing (NTC) to the forefront of promising solutions because it enables circuits to operate close to their optimal energy point (OEP), through aggressive voltage down scaling, which in turn results in significant performance loss. Performance loss, due to the reduced clock frequency from the increased cells' delay when reducing supply voltage, can be recovered through parallelizing across cores [123]. This can be achieved by lowering the supply voltage to the near-threshold region and trading performance of single-core for parallelized many cores operating at a low voltage. Moreover, NTC has become a reality when Intel has developed a processor that operates in NTC [43].

Importantly, NTC has wide impacts on the whole design, starting from device-level up to system-level. Therefore, OEP must be accurately estimated for the parallelized design to maximize NTC's gain.

Propagation delay, power and voltage: Reducing the supply voltage (V_{dd}) of a circuit leads to a significant reduction in its total power due to the quadratic saving in dynamic power ($P_{dynamic}$) besides the exponential saving leakage power ($P_{leakage}$) (see Eq. (2.1) [23]). The consequence of reducing V_{dd} is the quadratic increase in the transistor's delay (see Eq. (2.2)). As a result, the maximum delay of the slowest path (i.e., critical path) of the circuit (t_{CP}) becomes larger, leading to a lower operating frequency ($freq$). Hence, a considerable performance loss is observed as V_{dd} approaches the NTC region.

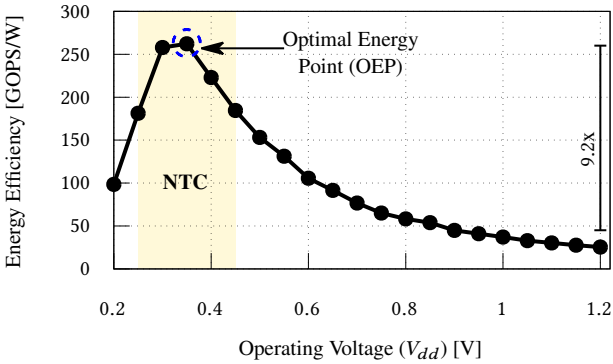


Figure 2.1.: The impact of reducing supply voltage on energy shows how the maximum energy efficiency exists within the near-threshold region. The Discrete Cosine Transform (DCT) circuit is considered in this analysis at the 45nm technology node.

$$P_{dynamic} \approx C_{eff} V_{dd}^2 f \ , P_{leakage} \approx V_{dd} K_1 e^{K_2 V_{dd}} \tag{2.1}$$

$$freq = \frac{1}{t_{CP}} \ ; t_{CP} = \sum_{T_i \in CP} t_{T_i} \ ; t_{T_i} \propto \frac{1}{(V_{dd} - V_{th})} \tag{2.2}$$

$$E = (P_{dynamic} + P_{leakage}) t_{CP} \tag{2.3}$$

Where C_{eff} is average switched capacitance, K_1 and K_2 are fitting parameters [23]. t_{T_i} is the delay of transistors that contribute to t_{CP} and V_{th} is the threshold voltage of the transistor.

Optimal Energy Point (OEP): it is important to determine the V_{dd} at which the energy efficiency is maximum because the energy of any circuit is the product of both total power and delay (see Eq. (2.3)). The total energy per cycle reduces with V_{dd} reduction until an inflection point at which it starts to exponentially increase as the leakage energy per cycle compensates for the reduction in dynamic energy per cycle [72]. The V_{dd} where such an inflection occurs typically represents the OEP, where the energy efficiency is maximized. For instance, Fig. 2.1 shows the energy efficiency of the Discrete Cosine Transform (DCT) circuit designed at 45nm technology. As shown, operating the circuit at 0.35V results in the maximum energy efficiency of 9.2x compared to the nominal voltage of 1.2V in the super-threshold region, for this particular

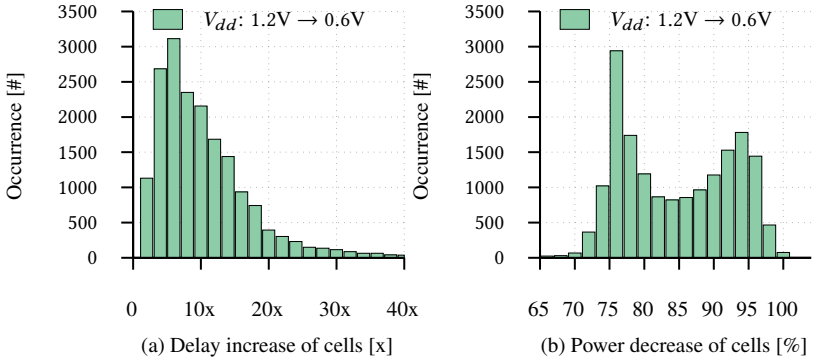


Figure 2.2.: (a) The same voltage reduction unevenly increases the delay of cells. (b) The same voltage reduction unevenly reduces the power of cells.

technology. However, for large circuits such as a processor, finding OEP is challenging. This is because standard cells are unevenly affected by voltage reduction with respect to delay and power.

Key challenges behind determining OEP: To understand the impact of voltage reduction on the processor's energy, the role of V_{dd} reduction on both delay and power has been investigated for the standard cells at the 45nm technology node. Fig. 2.2a presents how voltage reduction from 1.2V to 0.6V increases the delay of cells. As shown, the same V_{dd} reduction unevenly increases the delay of cells. While the delay of some cells is increased by only around 2x, the delay of other cells is increased by 10x and even up to 40x. Therefore, when studying the impact of voltage reduction on complex circuits like a full processor, it is difficult to accurately estimate the overall impact of voltage reduction on the processor's delay. This is because standard cells within the processor will be unevenly affected by the same V_{dd} reduction. Hence, every cell within the critical paths of the circuit will contribute its own delay increase to the over all delay increase. On the other hand, Fig. 2.2b presents the impact that voltage reduction from 1.2V to 0.6V has on the cells' power. As shown, voltage reduction also unevenly decreases the power of cells. While the power of some gates will be reduced by around 70%, the power of other cells may be reduced by more than 90%. All in all, the large variance in the power reduction of cells and in the delay increase of cells under the same voltage reduction makes estimating accurately how the energy of the circuit will profit from voltage scaling unclear, and hence, selecting where the OEP

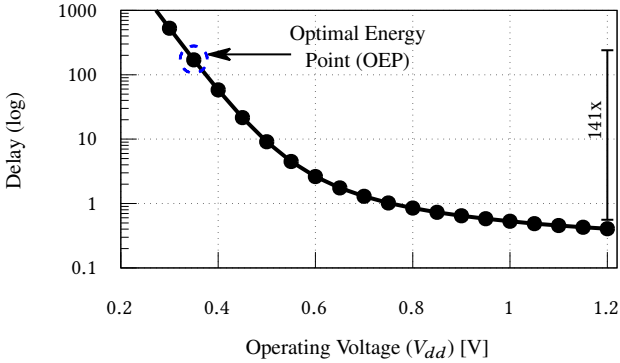


Figure 2.3.: Impact of voltage reduction on the delay of Discrete Cosine Transform (DCT) circuit, showing how the delay is reduced towards NTC.

exists is challenging. This even holds more when a complex circuit consisting of a numerous number of cells is analyzed where every cell within the netlist will be differently affected by the same V_{dd} reduction.

Parallelized Near-Threshold Computing: The key drawback of NTC is the significant performance loss. For instance, as shown in Fig. 2.3 for the DCT circuit, reducing V_{dd} from 1.2V to 0.35V, as its OEP, leads to 141x larger delay (i.e., lower frequency). Such a significant reduction considerably limits the possible services that processors may offer. For compensation, additional cores can be included where multiple tasks are parallelized. However, finding OEP then becomes more challenging. The main reason is that the sequential part of the executed applications akin to different limitations within the parallel execution. This plays a major role in defining the gain of parallelism and thus the required number of cores that are needed to fulfill the performance constraints. Additionally, despite the impact of using more cores on compensating the performance loss, a higher number of cores results directly in higher power consumption, and again OEP might be affected.

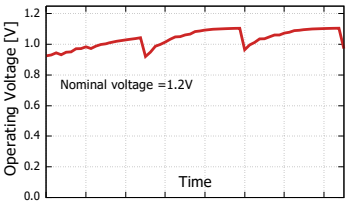
While finding accurately OEP is challenging for processors, finding OEP for parallelized processors is even more challenging. However, as finding OEP is crucial to maximizing energy efficiency, this dissertation proposes an accurate methodology to accurately find OEP for both single-core processors and parallelized processors, as will be presented in Chapter 3.

Prior works: Intel has presented in [43] its single-core processor based on the IA-32 architecture with the capability to operate at NTC. Intel has implemented multiple voltage domains to separate the supply voltage of logic from memory components. This is necessary to ensure that the minimum voltage V_{min} of the memory does not limit the V_{min} of the logic core. In [109], authors investigated the 7nm FinFET technology with respect to voltage reduction and compared it to planar MOSFET technologies. Results showed that FinFET provides 8.6x higher energy efficiency at NTC than the nominal operating voltage leading to a 4.8x gain compared to 20nm planar MOSFET. The analysis has been performed using a ring oscillator of thirty-one Inverters to investigate the impact of voltage reduction on delay and power at NTC towards analyzing the energy efficiency. [171] also studied FinFET technology with planar MOSFET w.r.t. energy for only a couple of different voltages (e.g., 0.3V, 0.45V, 0.8V, and 1.1V) by employing simple circuits such as Inverters chain, 16-bit adder, and 16-bit multiplier. The study showed that FinFET provides a better energy efficiency at NTC compared to planar CMOS technologies. [140] studied NTC with respect to approximate computing in error-tolerable applications. The work focuses on re-sizing channel length of nMOS transistors to avoid large transistors within the pull-up network.

As previously mentioned, the impact of voltage reduction on standard cells is non-uniform, and the variance is considerable. Hence, studying only a simple circuit with few cells (e.g., [140, 171]) or examining a chain of identical cells (e.g., [109, 108] is insufficient to investigate OEP. This holds even more when it comes to complex designs like a full processor.

Distinguishing from existing work: This dissertation's presented approach considers the non-uniform impact of voltage reduction on standard logic cells. This is achievable by studying delay and power for each cell individually over a range of operating voltage. This dissertation extends NTC analysis within the existence EDA tools by providing voltage-aware cell libraries, investigating the energy efficiency of parallelized NTC in manycore designs to accurately find OEP. Details in Chapter 3.

2.2. Interdependencies between Aging and IR-drop:



Voltage drop (IR-drop): voltage drops in metal lines is a fundamental property of any circuit due to the non-ideality in the Power Delivery Network (PDN), which originates from the parasitics (i.e., resistances and capacitances) of power lanes within PDN [20]. IR-drop results in a reduction in V_{DD} that reach the transistors within the circuit and hence a reduction in I_{ON} of transistors (see Eq. (2.4)). With each clock, the simultaneous switching of standard cells, sharing the same power lanes, results in peaks in the current demands. Due to the inability of PDN to deliver the current demands, supply voltage fluctuates at the terminals of standard cells leading to an increase in their delay.

BTI transistor aging: CMOS operates by switching the electric fields ON/OFF which stresses the materials within the transistor leading to the creation of defects [87]. Such defects are imperfections within the dielectric materials of transistors. Aging phenomena degrades the electrical characteristics of the transistor (e.g., V_{TH}, Carrier Mobility (μ), etc.). Bias Temperature Instability (BTI) is one of the key aging phenomena in the current technology nodes (i.e., technology <45nm) [114]. It is also predicted to remain a key degradation in the upcoming nodes [89]. The generated defects due to BTI, either at the Si-SiO_2 interface or deep within the dielectric, are undesired charges that interact with the applied electric field on the transistor [87]. This increases the threshold voltage, making the transistor weaker. ΔV_{th}, in turn, leads to a reduction in the I_{on} current (see Eq. (2.4)). Hence, it becomes slower.

While understanding the physical mechanisms of both aging and IR-Drop phenomena are not required at the system level, there is indeed still a substantial need to analyze their ability to induce degradations in order to accurately estimate reliability. For decades, both phenomena have been considered to be independent phenomena. Hence, interdependencies between aging and IR-drop remain hidden. However, the assumption mentioned above is invalid, as will be presented in Chapter 4. Therefore, estimating correctly the overall impact of both phenomena on the circuits necessitates investigating aging and

IR-Drop phenomena simultaneously, as the focus of this dissertation, and not individually as state-of-the-art techniques do.

Current technology: in this regard, the discontinuation of Dennard's scaling results in two key reliability problems at device and circuit levels:

1. Strong aging effects in transistors: elevated electric fields in transistors accelerate the underlying physical aging mechanisms (i.e., more defects). The generated defects manifest themselves as a shift in the key electrical characteristics of transistors, such as threshold voltage (V_{th}) [87]. In addition to ΔV_{th}, BTI can also alter other transistor parameters like carried mobility (μ), and sub-threshold slope (SS) [11]. However, ΔV_{th} is the dominant degradation caused by BTI [127]. In turn, V_{th} increase reduces the drain current of the transistor in the ON state, increasing the propagation delay of cells, as Eq. (2.4) and Eq. (2.5) demonstrate. Hence, timing violations (i.e., timing errors) occur in circuits due to unsustainable frequency if insufficient timing guardband (t_{GB}) was included on top of the critical path delay (see Eq. (2.5), Eq. (2.6)).
2. High IR-drops in circuits: elevated on-chip power densities increase the demands on the Power Delivery Network (PDN) of the circuit due to high current densities. In other words, due to technology scaling, the power lanes in circuits need to deliver higher currents to function more cells within the same area. The non-ideality in PDN manifests itself as IR-drops that lead to voltage fluctuations in which every cell within the circuit receives a different reduced V_{dd} from the ideal/nominal voltage. Hence, the propagation delay of cells increases due to the direct relation between V_{dd} and cell's delay and, additionally, due to the reduction in I_{on} (see Eq. (2.5) and Eq. (2.4)). Therefore, circuits under IR-drop become subject to sudden timing violations if insufficient guardband was included on top of the critical path delay (see Eq. (2.5) and Eq. (2.6)).

$$I_{on} \approx \frac{W}{2 \cdot L} \cdot C_{ox} \cdot \mu \cdot (V_{dd} - V_{th} - \Delta V_{th}) \tag{2.4}$$

$$t_{delay}(CP) = \sum_{m_i \in CP} \tau_{m_i}; \tau_{m_i} \approx \frac{C V_{dd}}{4} \left(\frac{1}{I_{onN}} + \frac{1}{I_{onP}} \right) \tag{2.5}$$

$$t_{clock} < t_{delay} + t_{GB} \Rightarrow \textit{timing errors} \; ! \tag{2.6}$$

$$t_{clock} > t_{delay} + t_{GB} \Rightarrow \textit{Perf. loss} \; ! \tag{2.7}$$

Here, m_i refers to transistors that form the critical path (CP) of the circuit. τ_{m_i} is the simplified propagation delay of an Inverter [60], as an example. C represents the capacitances connected to the cell. μ and C_{ox} are the carrier mobility and the oxide capacitance of the transistor, respectively.

Interdependencies between Aging and IR-drop: In the last decade, both aging and IR-drop degradations have been *independently* studied in detail. This is because aging has been typically considered as a long-term reliability degradation where its effects can be observed in hours and days, while IR-drop has been considered as a short-term reliability degradation where its effects can be observed within the microsecond regime (i.e., with every clock) [48]. Hence, interdependencies between aging and IR-drop remain hidden. Therefore, the required guardbands, for both impacts, are independently examined and the final guardband is the magnitude of summing of the two guardbands, which can result in performance loss when larger than the required guardband is considered (see Eq. (2.7)). However, advances in measurement equipment recently revealed that aging effects in current technologies could also be observed in a significantly shorter time regime [61, 134]. This is because technology scaling pushed transistors to an atomic level in which some types of defects can be generated and healed very fast [106, 52].

This part demonstrates that the existing interdependencies between aging and IR-drop do matter and cannot be ignored where aging and IR-drop effects influence each other, i.e., amplify or mitigate each other. Interdependencies are summarized in the following.
(a) Aging → IR-drop: aging-induced degradation (ΔV_{th}) amplifies the impact that IR-drop has on the delay of a circuit. This is observed from Eq. (2.4) and Eq. (2.5). The impact of V_{dd} reduction on I_{on} of transistor's delay becomes larger when aging-induced degradation occurs (i.e., $\Delta V_{th} > 0$).
(b) Aging ← IR-drop: IR-drop partially mitigates aging-induced ΔV_{th}. A

reduction in supply voltage allows some of the generated defects to heal. Therefore, IR-drop has the potential to reduce the induced ΔV_{th} and, hence, mitigate the deleterious impact of aging on the circuit's delay.
(c) Aging ↔ IR-drop: To find the efficient timing guardband under the joint impact of aging and IR-drop, the interdependencies between aging and IR-drop must be investigated. Neglecting the amplification impact leads to underestimating the required guardbands and thus including insufficient guardbands, which results in timing violation (see Eq. (2.6)). Neglecting the mitigation impact leads to overestimating the required guardbands and thus including inefficient (i.e., larger than what is actually needed) guardbands, leading to unnecessary performance losses (see Eq. (2.7)).

Prior work: Investigating the impact of aging or IR-drop alone on the delay of circuits and estimating the required timing guardband has been intensively explored in the past years, e.g., [9, 21]. Research w.r.t estimating the required timing guardbands under both phenomena *jointly* is still in its infancy. In [48], authors assumed that the rate of ΔV_{th} change due to BTI is significantly larger than the rate of V_{dd} change due to IR-drops. Hence, it was concluded that aging and IR-drop effects could be fully *independently* considered. Based on that, authors estimated the overall timing guardband under both phenomena as the magnitude summation of the guardbands required for each of them *individually*. Despite none of the existing studies analyzed how dynamics in V_{dd} (i.e., IR-drop) influences the underlying mechanisms of aging, some previous work aimed at estimating how aging-induced degradation (ΔVth) might have a different impact on the circuit's delay when V_{dd} becomes lower. [136] experimentally examined the delay of a ring oscillator under the effect of aging at various voltages showing that aging-induced ΔV_{th} becomes higher when switching from high to low V_{dd}. [2] studied the impact of ΔV_{th} due to BTI in SRAM cells under a constant V_{dd} reduction of 10%, mimicking the worst-case IR-drop scenario. Recently, [133] showed how voltage scaling for power management might narrow aging-induced ΔV_{th} due to the partial recovery which occurs when lowering V_{dd}.

Limitations in current EDA tools: Existing EDA tool for power signoff can extract the voltage waveforms due to IR-drop across the circuit. If multiple cell libraries at varied V_{dd} levels are provided to the timing signoff tool, the delay increase caused by IR-drop can be accurately estimated. Unlike IR-drop, estimating aging-induced delay increase is still not in the EDA tool.

Distinguishing from existing work: This work considers the interdependencies between aging and IR drop (i.e., aging ↔ IR-drop) to estimate the smallest, yet sufficient, timing guardband to protect circuits against both phenomena. The presented approach extends the available commercial EDA tool flows by integrating the proposed technique within the EDA tool for aging and IR-Drop analysis, employing the state-of-the-art physics-based BTI model, which can predict arbitrary voltage waveforms caused by the voltage fluctuations.

2.3. Aging-aware Approximate Computing:

Neural Networks (NNs) have increasing computational demands in many applications [69]. To achieve that and speed up the Deep NNs (DNNs), custom Application-specific integrated circuits (ASIC) for Neural Processing Units (NPUs) are becoming ubiquitous in general purpose and embedded computing [69]. NPU consists of thousands of multiply-accumulate (MAC) units [8], providing massive parallelism of the performed computations. For instance, Google TPU employs 64K MACs [69]. In NPUs, a large number of MAC units are compact and tightly packed within a small and dense area. Their inherent nature of performing massive parallelism makes MACs highly utilized and subject to elevated on-chip power densities that rapidly result in excessive on-chip temperatures [8].

Circuit aging: The very high utilization of MAC circuits within NPUs results in continuous stress with very small time for recovery. Hence, transistors age faster. Furthermore, excessive temperatures accelerate the mechanisms behind transistor aging [144]. Aging manifests itself as an increase in the threshold voltage (V_{th}) of the transistor, leading to a reduction of the drain current of a transistor in the ON state (I_{on}), and hence slower operation [127]. Therefore, circuits exhibit timing violations (i.e., errors) because the operating frequency becomes unsustainable over time (see Section 2.2). To protect circuits against timing violations, timing guardband (t_{GB}) must be included on top of the critical path delay for the projected lifetime. This results in considerable performance loss from the beginning until the end of the projected lifetime even though aging-induced delay degradation does not yet exist, or is very small, at the early phases of the chip's lifetime. Hence, the cost of guardbanding is paid from the very beginning, even when it is not yet needed.

Different approaches have been proposed [121, 73] to recover the associated performance losses. Such techniques reduce the aging impact. However, they induce area/power overhead [121].

Aging-aware approximation: Approximate computing has been employed to mitigate aging in error-tolerant applications [75]. It exploits the inherent error resilience of several applications to trade-off computational accuracy for delay [75]. Aging-aware works by improving the circuit's performance and hence mitigate aging effects. However, prior works examined very simple topologies, e.g., RCA adders and multipliers [75, 10].

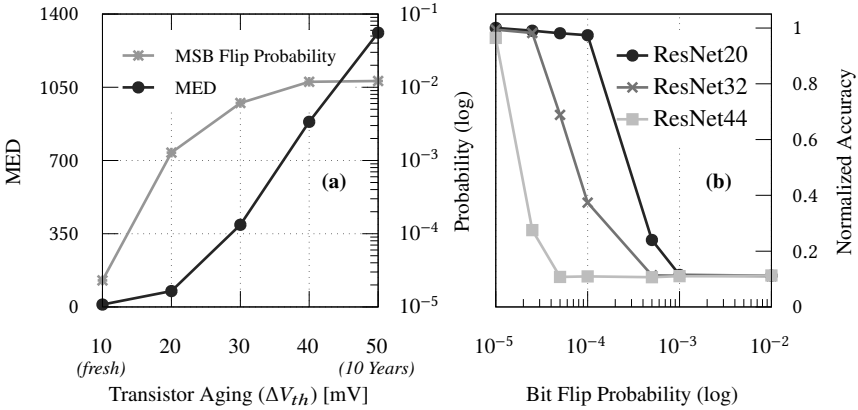


Figure 2.4.: (a) The error of an 8-bit multiplier under aging. (b) Accuracy of three NN (ResNets) when errors randomly injected in the 2 MSBs at the multiplications of the convolutional layers.

Aging-induced timing errors in NPU Induced timing errors result in unacceptable accuracy loss even after a short time[75]. In arithmetic circuits, errors occur mainly in the most significant bits (MSBs) [75]. Fig. 2.4 shows the aging-induced timing errors of an optimized 8-bit multiplier circuit that operates at the maximum frequency without an aging guardband [130]. Fig. 2.4(a) shows the Mean Error Distance (MED) for different aging levels (represented as an increasing level in threshold voltage (ΔV_{th})) in addition to the resulting probability of a bit flip in one of the two MSBs. Please note, ΔV_{th}=50mV represents the end of the projected lifetime (e.g., 10 years). As shown, more errors are produced over time (higher MED), and the probability of a bit-flip at the MSBs increases significantly.

Recent research demonstrates that the deeper NNs are, the more susceptible to errors generated in the multiplier units[158]. For instance, Fig. 2.4(b) shows an estimation of how aging-induced timing errors impact the accuracy of different NNs. The error is injected in the performed multiplications randomly by flipping one of the two MSBs with a given probability. Three ResNets are considered here, and the probability of the bit flip ranges between 10^{-5} and 10^{-2}. As shown, as the probability of bit flip increases with aging, the accuracy drops significantly and becomes unacceptable after a very small probability level of around $5{\times}10^{-4}$. Moreover, bit flip probability of 10^{-3}, which represents

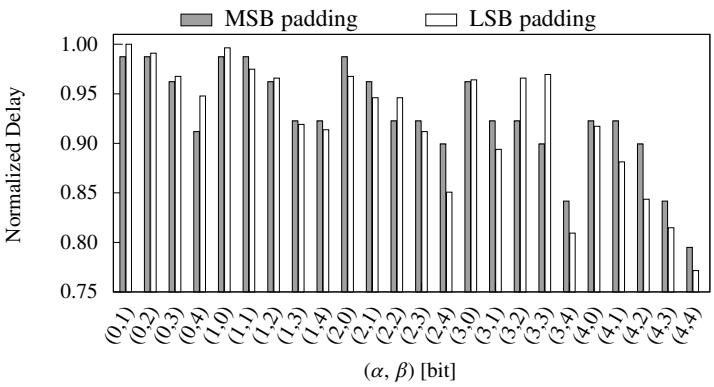


Figure 2.5.: Delay gain of the 8-bit MAC when applying (α, β) input compression. Both MSB and LSB paddings are evaluated.

1 year of aging, results in an unacceptable accuracy for all the examined NNs. In addition, for the deeper NNs, the accuracy drops much faster. In summary, aging-induced timing errors are critical for the NPU's accuracy, leading to poor quality even after a small degradation.

Input Compression: NPUs perform millions of MAC operations per inference. Hence the overall performance is defined by the performance of its MAC. This work considers a design similar to Google Edge TPU [32], consisting of 64×64 MAC array. MAC consists of an 8-bit unsigned multiplier and 22-bit unsigned adder to prevent accumulation overflow. Details in Chapter 5.

As a matter of fact, timing paths within the circuit feature different delays, and therefore, the active paths are input-dependent. For arithmetic circuits, lower bit-width numbers lead to faster operation[10]. Therefore, compressing the inputs of the MAC could result in faster operation. Instead of performing $A\times B+C$ for accurate MAC, circuit performs $A'\times B'+C'$. The compressed inputs A', B', and C' feature a bit-width of $8-\alpha$, $8-\beta$, and $22-(\alpha+\beta)$, respectively. Since the compressed inputs feature smaller bit-width, this work applies zero-padding to the remaining bits without altering the circuit's design.

Fig. 2.5 shows the evaluation of the delay of the MAC unit by compressing its inputs. Various compression values (α, β) for both padding options are examined. Input compression achieves up to 23% of delay gain. Some compression values are benefited by MSB padding while others by LSB padding.

Input Compression Through Quantization: Quantization is the process of mapping input values from a large set (i.e., large cardinality) to output values in a smaller set (i.e., low cardinality). Quantization forms the core of essentially all lossy compression algorithms. The mapping that realizes the quantization process is called a quantizer. Let X and Y denote the source and reproduction results with quantizer q. The quantization process is as follow:

$$q : X \rightarrow Y, \texttt{where } q(X) = \{q(x) : x \in X\} \tag{2.8}$$

In this dissertation, quantization forms an extra layer that maps the compressed input bits after applying the quantization model to MAC inputs. Quantization models are meant to reduce the output accuracy. However, many quantization models have been proposed for various applications and where performance and accuracy are the metric to differentiate between these models. For instance, for NN, quantization through quantizing the activations and the weights to lower bit width can be applied.

Over time, while the aging phenomena become stronger, a gradual increase of the compression level can be applied to increase the delay gain. Hence, by compressing the MAC inputs, the accuracy of NN will be degraded. To reduce such degradation, multiple low bit-width quantization techniques [77, 65, 19, 94] are employed in this work. Further details in Chapter 5.

In this work, a library of multiple low-bit width post-training quantization methods is created at design time based on the recently published approaches. Since some of them are optimized for specific NNs, or optimized for low precision, multiple methods must be integrated considering the diversity in NNs. However, some methods require off-line statistics and in some cases, the actual quantization is time-consuming due to the multiple optimizations. Importantly, all these methods do not require NN retraining and allow the utilization of different precision for weights and activations. More details about the employed quantization technique in [77, 65, 19, 94].

Prior work: In [92, 176, 158] approximate multipliers are employed and run-time reconfigurable approximate NN inference accelerators are implemented. [92, 176, 158] target solely the power consumption of circuit and not delay optimization. In [10], a fixed approximation through precision scaling is presented aiming on narrowing or removing aging guardbands. Nevertheless, fixed approximation leads to constant quality degradation that cannot be adapted over time. In [75], adaptive input cutting and masking techniques are

proposed to mitigate aging and achieve a graceful accuracy degradation of a DCT/IDCT accelerator. [75] was only applied to the very slow ripple-carry adder and array multiplier. Similarly, [25] shows the ability of adaptive approximation to control the chip temperature at run-time. Works in [75] and [25] require a control circuitry to set the run-time approximation. Moreover, [75] requires control circuitry to set the run-time approximation. [75, 10] reduce the computational precision of the accelerator itself. Hence, considering that errors due to approximate hardware are input-dependent, by just omitting some bits from the computations [75, 10], the quality loss for some inputs might be unacceptable [176].

Distinguishing from existing work: This work suppresses aging effects in NPUs by applying an adaptive approximation through input compression under reliability-aware quantization technique. With the smallest, yet negligible, inference accuracy loss, aging guardbands can be entirely eliminated for the entire projected lifetime by gracefully compressing the NPU inputs.

2.4. Self-Heating Effects:

A fin field-effect transistor (FinFET) is a multigate device based on a MOSFET (metal-oxide-semiconductor field-effect transistor) built on a substrate. FinFET devices are widely used due to their reduced leakage and excellent subthreshold slope compared to planar MOSFET[147]. FinFET advantages resulted from the new 3D structure of transistors with a vertical junction. Fig. 2.6b shows the 3D structure of FinFET consisting of a silicon channel (fin) and a gate that surrounds the channel on three sides. With the introduction of the 3D structure and due to the low thermal conductivity of the gate dielectric, the heat dissipation from a FinFET channel is very limited compared to planar MOSFETs. In MOSFETs, heat can be dissipated via the substrate layer by conducting heat, as shown in Fig. 2.6a. However, most of the heat generated within the FinFET transistor's channel remains within its channel as it slowly escapes to the body. The temperature spreads between drain and source where the hotspot appears at drain, as shown in Fig. 2.6b. This results in what so-called the Self-Heating Effect (SHE) [126].

SHE refers to the elevated channel temperatures (T_C) and their impact on the performance of transistors. The channel temperature is elevated due to Joule heating by current flow through the channel. This heat must then be conducted via the thermal resistance R_{th} between the channel and the substrate. However, the R_{th} between the channel and transistor's gate stack is very high. Therefore, only small amount of the heat over time can escape to the body and heat the chip (see Fig. 2.6b). Thus, the chip's temperature T_{chip} stays relatively cool ($T_C >> T_{chip}$). However, once the heat reaches the transistor's vias, slowly and steadily, it can escape and heats the entire chip [165].

Impacts on transistor: Elevated T_C affects the drain current I_D of the transistor and, thus, circuit performance [42]. This impact can be formulated as two opposing effects [168]: (1) the reduction of the threshold voltage (V_{th}) and (2) the reduction of channel carrier mobility (μ). Lower V_{th} results in higher drain currents I_D in the ON state (i.e., faster-switching speed). Reducing μ leads to lower I_D (i.e., slower switching speed and thus worse performance), opposing the beneficial impact of lower V_{th}. Therefore, the final performance of the transistor is the result of the compromise between V_{th} and μ effects [162]. In its simplest form, both parameters can be modeled as functions of temperature, according to [70], as follows:

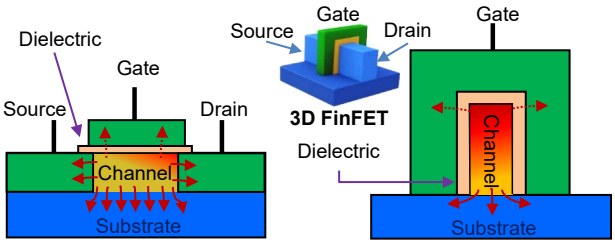


Figure 2.6.: (a) Planar MOSFET transistor: Heat dissipation from the channel is convenient due to conduction towards the substrate. (b) 3D FinFET (side view of the channel directly after drain showing the hotspot within the channel): Exhibits limited heat dissipation from its channel and thus, most of the heat is contained, exacerbating deleterious temperature effects.

$$\mu(T_C) = \mu(T_{ambient})(\frac{T_{ambient}}{T_C})^m \tag{2.9}$$

$$V_{th}(T_C) = V_{th}(T_{ambient}) - k(T_C - T_{ambient}) \tag{2.10}$$

Where $T_{ambient}$ is the room temperature in Kelvin, m and k are positive constants, and T_C is the channel temperature. These models show that V_{th} scales linearly with T_C increase, while μ scales with a power law. Please note, these models are only to simplify the relation between μ and V_{th}.

Thermal regions: Under nominal operation conditions (high V_{dd}), a temperature increase in T_C manifests itself as an I_D decrease and thus an increase in the transistor's delay. This means that the effect of lowering V_{th} is smaller than lowering μ: $\Delta I_D(V_{th}) < \Delta I_D(\mu)$. However, when looking across the range of all possible operating voltages, $\Delta I_D(V_{th})$ and $\Delta I_D(\mu)$ scale differently. At higher V_{dd}, $\Delta I_D(\mu)$ is larger, while at smaller V_{dd} $\Delta I_D(V_{th})$ is larger. Hence, three key regions emerge: Positive-Temperature Dependence (PTD) (i.e., increasing T_C reduces I_D), Zero-Temperature Coefficient (ZTC) (i.e., increasing T_C does not change I_D) and Inverse-Temperature Dependence (ITD) (i.e., increasing T_C increases I_D)[168]. The two dependencies compete and the stronger dependence determines how the transistor is affected by temperature.

The three thermal regions are summarized in Table 2.1. In these three regions, I_D falls, stays exactly the same or rises with increasing T_C, depending if $\Delta I_D(\mu)$ is larger or smaller than $\Delta I_D(V_{th})$. Following the proposed methodology in

Table 2.1.: ON-current dependencies on V_{th} and μ within the thermal regions

Thermal Region	I_D Dependencies	Rising T_C
PTD	$\Delta I_D(\mu) > \Delta I_D(V_{th})$	Lower overall I_D
ZTC	$\Delta I_D(\mu) \approx \Delta I_D(V_{th})$	Same overall I_D
ITD	$\Delta I_D(\mu) < \Delta I_D(V_{th})$	Higher overall I_D

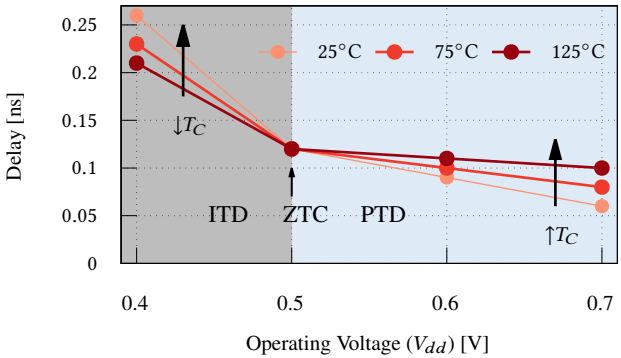


Figure 2.7.: Definition of the thermal regions when operating a Ring Oscillator (RO) circuit, consisting of 13 inverters designed at 7nm technology [39], at different voltages and three different temperatures where three regions emerge: Positive-Temperature Dependence (PTD), Zero-Temperature Coefficient (ZTC) and Inverse-Temperature Dependence (ITD).

the proposed analysis and evaluation (details in Chapter 6), a Ring Oscillator (RO) circuit is tested for ZTC. Fig. 2.7 shows the delay of the critical path $t_{delay}(CP)$ of RO, consisting of 13 inverters designed at 7nm technology [39], operating at three different T_Cs over a wide range of voltages. Delay values $t_{delay}(CP)$ start to converge in the PTD region with V_{dd} decreases. This trend remains until all $t_{delay}(CP)$ values meet at ZTC. Continuing over V_{dd} reduction, $t_{delay}(CP)$ values start to diverge again in the opposite direction in ITD. At ZTC (V_{ZTC}=0.5V in this particular example), $\Delta I_D(\mu)$=$\Delta I_D(V_{th})$ and thus, transistors, and thus the circuit, do not exhibit any thermal variance due to the compensation of beneficial ΔV_{th} with detrimental $\Delta\mu$. Please note, RO circuit is absolutely uniform ignoring local variation, i.e., all cells (subcircuits) are identical. Therefore, ZTC is identical for all subcircuits and no thermal variance is exhibited when operating at ZTC.

Temperature Modeling of Transistors: While for large transistors, Eq. (2.10) and Eq. (2.9) from 2001 [70] were fine. Nano-scale transistors have various additional dependencies, which must be considered. The temperature models $V_{th}(T_C)$ and $\mu(T_C)$, as well as the resulting $I_D(T_C)$, need to be more sophisticated to accurately predict transistor behavior and match reported experimental data. V_{th} temperature dependency is defined as follows:

$$V_{th} = V_{th0} + \Delta V_{th,all}{}^{2} \tag{2.11}$$

$$V_{th0} = \frac{kT}{q} \cdot ln\left[\frac{C_{ox}\frac{kT}{q} \cdot (C_{ox}\frac{kT}{q} + 2Q_{bulk} + 5C_{si}\frac{kT}{q})}{2q \cdot n_i \cdot \epsilon_{sub} \cdot \frac{kT}{q}}\right] + V_{fb} + \phi_B + \Delta V_{th,QM} + \frac{kT}{q} + q_{bs} \tag{2.12}$$

Where the following parameters are temperature dependent (i.e., feature the term) "$\frac{kT}{q}$": C_{ox} is the oxide capacitance, C_{si} is the body capacitance, Q_{bulk} is the fixed depletion charge, $\Delta V_{th,QM}$ is the surface potential considering quantum mechanical effect, k is Boltzmann constant, q is the electronic charge, n_i is the intrinsic carrier concentration, T is the temperature, ϵ_{sub} is the dielectric constant. V_{fb} is the flatband voltage, ϕ_B is the body-effect voltage parameter, q_{bs} is the body doping. Note the frequent occurrence of temperature terms "$\frac{kT}{q}$" highlights the actual complexity of taking elevated T_C into account.

$$\begin{aligned} t_{clk} &= t_{delay}(CP) + t_{GB} \\ t_{GB} &= \Delta t_{delay}(CP) \end{aligned} \tag{2.13}$$

Where $t_{delay}(CP)$ is the nominal propagation delay of the critical path in the circuit (see Eq. (2.5)), t_{GB} is the safety timing margin added to tolerate degradation in order to protect circuits against timing violations, and t_{clk} is the clock period. Larger $\Delta t_{delay}(CP)$ necessitates longer t_{GB} and thus longer t_{clk}, reducing f_{clk} and thus the circuit's performance. Therefore, t_{GB} must be minimized in order to keep performance as high as possible.

[2] Due to many phenomena that affect threshold voltage (e.g., roll-off, DIBL, reverse short channel effect, and temperature), accordingly, corrections have been added[26].

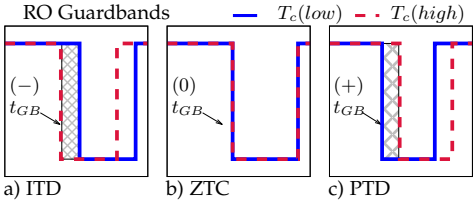


Figure 2.8.: The corresponding guardbands of the Ring Oscillator (RO), consists of 13 invertors, in the three thermal regions where t_{GB}=$t_{delay}(T_C(high))-t_{delay}(T_C(low))$.

Timing guardband: Traditionally, designers employ the worst-case timing scenario to overcome SHE-induced delay degradation (i.e., delay increases). Timing guard band (t_{GB}) must be added on top of the maximum delay of a circuit (i.e., critical path delay $t_{delay}(CP)$) to overcome delay degradations. This corresponds to a timing slack applied to the clock period shown in Eq. (2.13).

Nevertheless, t_{GB} is aimed to tolerate degradations regardless during high or low temperatures as illustrated in Fig. 2.8. The figure demonstrates that guardband can be positive, negative or zero following the thermal region that the circuit operates within. Note, there is no negative guardband, this just to highlight that the circuit becomes faster. Hence, it does not matter if t_{delay} starts to shift due to a high or low temperature from its nominal value. The guard band t_{GB} always follows worst-case timing (i.e., maximum delay). In ITD this means t_{delay} at low T_C, while in PTD this means t_{delay} at high T_C.

The SHE-induced timing guardband can be pretty significant, as T_C can be pretty high. ΔT_C due to SHE, and based on our analysis, could exceed 350°C in the worst case. Such a high T_C is not extraordinary as similar results (T_C>450°C) were previously reported in measurements in [98, 156, 16]. The timing guardband to protect the circuit against thermal variation can reach up to double the operating delay ($t_{GB} \approx 90\% \cdot t_{delay}$). Such large guardbands result in a severe performance loss.

Self-Heating is Unavoidable: SHE results from many design decisions in FinFET (3D structure, dielectric material, etc.) necessary for the good electrical characteristics of the FinFET. As long as these parameters are unchanged, SHE remains. Additionally, traditional temperature mitigation techniques are unable to reduce SHE. Therefore, SHE is unavoidable for the following reasons:

(a) Chip cooling is insufficient: Increasing the cooling capability of the chip does not help when it comes to SHE. SHE originated from the insulation of

the channel's surroundings, so cooling the surrounding area of a transistor has little impact on the heat contained inside the channel.

(b) DVFS is ineffective: Dynamic Voltage and Frequency Scaling (DVFS) (e.g., Intel Turbo boost [35]) oscillates between high-performance high-voltage operation and low-performance low-voltage phases to balance between performance and temperature. However, DVFS technique cannot handle SHE. Lowering the voltage by itself is efficient since the generated heat due to SHE is given by $I_D \cdot V_{DS}$ model and reducing the supply voltage V_{dd} reduces both V_{DS} and I_D (via V_{GS} reduction). Therefore, reducing the voltage permanently is effective against SHE. However, DVFS exploits the thermal capacitance of a chip. With low $R_{th}(chip)$ (good conduction via metal lines) and high $C_{th}(chip)$ (heat sink, heat spreader), the chip has a time constant τ_{chip}>100ms. Thus, the sub-microsecond oscillations (fully integrated or in-situ voltage regulators [28]) between V_{high} and V_{low} can sustain $T_{chip}<T_{critical}$. However, SHE does not heat the entire chip but instead acts locally and solely heats the transistor channel. As transistors are tiny, SHE has less material to heat (i.e., tiny C_{th}), resulting in time constants in the nano-second range, orders of magnitude smaller than the chip. Importantly, nano-second time constants are much smaller than ramp-up and ramp-down times of the voltage regulators in current chips [28]. DVFS should switch in the pico-second range in order to exploit the time constant of SHE and maintain $T_C<Tcritical$, which is entirely unfeasible with currently existing voltage regulator designs.

(c) Spatial thermal management is ineffective: Other thermal management techniques like task migration, thermal load balancing, etc. are equally ineffective. Spatial techniques either exploit cool neighboring computing cores, the thermal capacitance of the cores themselves (heat the core, then migrate away to cool) or want to prevent clustering of many hot cores together (formation of hotspots on chip). However, SHE does not affect or depend on the T_{chip} due to the insulation (high R_{th}) of the channel from its surroundings. Therefore, while these techniques successfully reduce T_{chip}, they have little impact on $\Delta T_C(SHE)$ and thus its induced thermal and delay variance.

(d) SHE requires new architecture-level mitigation: SHE is transistor-level problem, which currently cannot be solved at the transistor level. To continue geometry scaling, transistor designers had to improve the electrostatics of a transistor beyond MOSFETs to reduce leakage I_{OFF} and control the I_D. Therefore, 3D structures, like FinFETs, emerge as the solution, encompassing the channel to improve control over it. Yet, surrounding the channel in the gate

dielectric (FinFETs, Nanowires) or separating the channel from the substrate with a buried oxide (SOI MOSFET, SOI FinFET) increases R_{th} and thus results in SHE. The trend clearly tends towards channel insulation and as such, SHE started with SOI power MOSFET and worsens in each new generation [68].

In summary, SHE is inevitable and there is no way to bypass the SHE-induced degradations (i.e., power and delay) using the traditional power/temperature management techniques and designers can only try to reduce its impacts. To recover performance loss, the guardband must be lowered. ZTC operating point is well-suited to minimize SHE impacts on circuit's delay. By definition, it is a point (or region) where the temperature has little impact on the circuit's delay. Consequently, in this dissertation, we employ ZTC to minimize the impact of the unavoidable SHE. Details in Chapter 6.

Prior work: A large body of works studied only simple circuits to characterize ITD, from single transistors to small circuits. [70] studied the operation of transistors in different thermal regions. [1] presented an analysis of ZTC of a 32-bit CMOS adder on 65nm technology. [41] showed ITD impact on performance in 65nm CMOS ring oscillator in the sub-threshold regime. This quantitative study shows that ZTC occurs at V_{ZTC} = 0.9V. However, studying ITD and ZTC in a single transistor or simple circuits is insufficient because their ZTC is different, and thus a single RO is not representative for a chip. At the circuit level, Intel presented in [86] a 130nm test chip containing different types of ring oscillators. The study showed distinct V_{ZTC} for each cell type in the range between 0.783-0.866V.

SHE is well studied at the transistor level since it is well known for Silicon-On-Insulator (SOI) devices, e.g., [110]. Recently, transistor-level studies in FinFETs provide a good understanding of SHE in transistors in [4], [68], and [54]. However, these studies are limited to simple circuits, and the impact of SHE beyond ring oscillators and SRAM cells is not yet examined, which is covered in this dissertation.

Distinguishing from existing work: This dissertation presents an analysis of the impacts of Self-heating effects (SHE) on both the timing and power of large digital circuits by extending the existing Multi-Corner Multi-Mode (MCMM) approach used in EDA tool flow for SHE. It proposes operating the circuit at its Zero Temperature Coefficient (ZTC) to eliminate thermal variations due to SHE-induced variance with near-zero guardband, eliminating SHE-induced variances in performance and power. Further details in Chapter 6.

2.5. Negative Capacitance Field Effect Transistor (NCFET):

Negative Capacitance Field-Effect Transistor (NCFET) is one of the leading emerging technologies that considerably improves the transistor's efficiency by overcoming the fundamental limit of the sub-threshold swing (*SS*) that impedes all existing CMOS technologies including FinFET as well as futuristic nanowire and even nanosheet transistors [59, 122, 179, 83]. The fundamental limit of *SS* ($\approx$ 60 mV/dec) [179, 138] is the origin of non-ideal voltage scaling that resulted in the discontinuation of Dennard's scaling more than a decade ago [7]. There are two ways to form NCFETs. One is to add an external ferroelectric (FE) layer within the traditional FET (e.g., FinFET) gate stack, as shown in Fig. 2.9, with ferroelectric and high-κ as separate layers, where the Metal-Ferroelectric-Metal-Insulator-Semiconductor (MFMIS) structure is formed [104]. The other way is to replace the gate oxide (i.e., high-κ layer) of traditional FET with ferroelectric layer through doping the hafnium oxide (HO_2) material with zirconium to obtain ferroelectricity ($Hf_{0.5}Zr_{0.5}O_2$), where a Metal-Ferroelectric-Insulator-Semiconductor (MFIS) structure is formed [104]. However, with both strategies, NCFET does not come with an area overhead because the transistor's size (length and width) remains exactly the same. The increase in the thickness itself, due to the ferroelectric layer in MEMIS scenario, does not increase the area footprint that each FinFET transistor occupies. Hence, the area of the circuit is *almost* not affected.

The ferroelectric layer manifests itself as negative capacitance (NC) that leads to charge redistribution in which the gate switching occurs at lower applied potential than in traditional FET. Such a layer provides amplification of the vertical electric field that the transistor perceives, which can be attributed to the better gate control over the channel and the enhancements in the surface potential. In turn, this allows the transistor to overcome the fundamental limitation of the sub-threshold swing of 60mV at room temperature. The principle of NCFET was first proposed in 2008 by S. Salahuddin [59, 122].

Unique characteristics of NCFET:
In the following, the unique characteristics of NCFET over traditional FET are presented.
(a) Voltage amplification: The configuration of metal-ferroelectric-metal-insulator-semiconductor (MFMIS) in NCFET, which is considered in this dissertation, is done by integrating a ferroelectric layer within the gate stack of

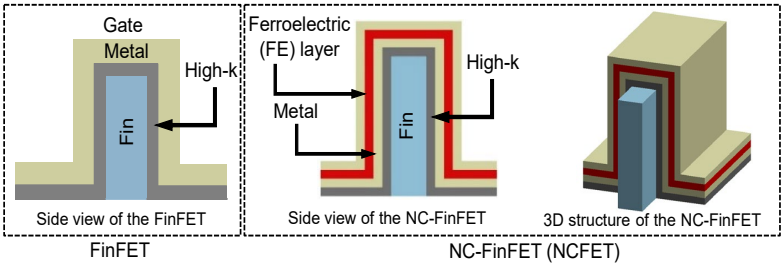


Figure 2.9.: The structure of both FinFET and NC-FinFET (NCFET) transistors. NCFET has a ferroelectric layer that is integrated within the gate stack of the transistor.

FinFET transistor, as illustrated in Fig. 2.9 in comparison with the traditional FET. The ferroelectric layer manifests itself as a negative capacitance (NC). Due to the NC, the total capacitance of the NCFET transistor is increased [7]. Fig. 2.10 shows the equivalent capacitances of the NCFET, which consists of the ferroelectric capacitance (C_{fe}) and the internal baseline FET capacitance (C_{int}). As shown, the two capacitances are connected in series. The well-known fact that when two capacitors are connected in series, the same amount of electrical charge is stored in each and, hence, a voltage drop occurs. Therefore, the total capacitance of the two capacitors is smaller than the smallest one. However, NC effects break the aforementioned fundamental law in physics where the total capacitance becomes larger instead of smaller [7]. This amplifies the internal voltage potential (V_{int}) that the internal gate perceives, as shown in Fig. 2.10. Therefore, the voltage amplification (A_V) at the internal gate can be then expressed as in Eq. (2.14).

$$A_V = \frac{\partial V_{int}}{\partial V_g} = \frac{|C_{fe}|}{|C_{fe}| - C_{int}} \tag{2.14}$$

For no hysteresis: $|C_{fe}| > C_{int} \Rightarrow A_V > 1$

$$V_{int} = A_{avg} V_g \text{ for a fixed } V_d;\ A_{avg} = \frac{1}{V_g} \int_0^{V_g} A_V \, dV_g \tag{2.15}$$

In turn, the provided voltage amplification by the ferroelectric layer will enable the NCFET to always reach a higher internal gate voltage V_{int}, as described in Eq. (2.15). Therefore, this resulted in a higher ON current (I_{ON}) at the same

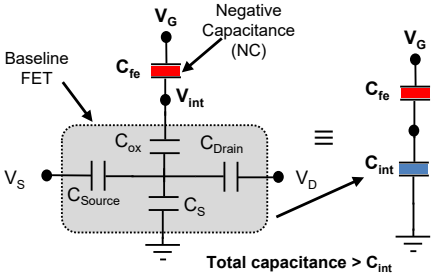


Figure 2.10.: The equivalent capacitance divider circuit and the equivalent circuit of the gate stack of NC-FinFET using the metal-ferroelectric-metal-insulator-semiconductor (MFMIS).

voltage. Therefore, as the maximum frequency of a circuit is governed by the I_{ON} of its transistors, NCFET-based circuits have, at the same voltage, higher frequency (i.e., performance) compared to traditional FET [7].

(b) Dynamic Power: The dynamic (switching) power (P_{dyn}) of a circuit is determined by the switching activity (α), total capacitance (C), operating voltage (V_{DD}), and frequency (f), as (Eq. (2.16)) shows.

$$P_{dyn} = \alpha \frac{V_{DD}}{T} \int_0^T I_{dd}(t)dt = \alpha C V_{DD}^2 f \tag{2.16}$$

However, while the C_{fe} exhibits a negative value while the condition $|C_{fe}| > C_{int}$ is always met within the range of operating voltages to ensure hysteresis-free operation [7], the total capacitance of NCFET, therefore, is always larger than the baseline FET (C_{int}), as Eq. (2.17) shows.

$$C_{NCFET} = \frac{C_{fe} \cdot C_{int}}{C_{fe} + C_{int}} > C_{int} \tag{2.17}$$

Therefore, compared to traditional FET, NCFET-based circuit will dissipate/-consume higher dynamic power as it will exhibit a larger total capacitance at the same V_{dd}. This is not only because of the larger total capacitance of the

transistor but also because the NCFET-based circuit can operate at a higher frequency (f). Hence, the total dynamic power is increased, see Eq. (2.16).

(c) CMOS Compatibility: NCFET technology is fully compatible with the existing CMOS fabrication process. In turn, this removes any cost overheads for the semiconductor industry to adopt. This became viable after the discovery of ferroelectricity in HfO_2 based materials, which are the standard materials in the current CMOS technology. Moreover, clear evidence was also provided after GlobalFoundries fabricated NCFET circuits using their mature commercial 14nm FinFET technology [78]. Importantly, NCFET transistors have the same footprint, as traditional FET transistors as explained previously. Hence, NCFET-based circuits do not incur any area overheads [129].

(d) Leakage and voltage dependency: In traditional FET technology, it is well known that the leakage current (I_{off}) decreases as the operating voltage V_{dd} decreases. Therefore, for instance, power and energy management techniques always aim at operating the circuits (e.g., processors) at the minimum possible V_{dd}, under performance constraints, to minimize power/energy. However, this is not always true when it comes to NCFET. Such a well-known voltage dependency (i.e., leakage current (I_{off}) decreases with V_{dd} decreases) becomes inverse with respect to leakage power in NCFET, under particular design corners, due to the negative DIBL effect, which is a typical characteristic of short-channel in NCFET [105]. Negative DIBL reduces the threshold voltage (V_{th}) of the transistor and thus increases I_{off} when the operating voltage decreases. In practice, when V_{dd} is increased in the OFF state, the gate charge reduces due to the electric field from drain to ferroelectric layer [102]. As the ferroelectric layer in NCFET is biased in a negative capacitance state, a decrease in charge means a corresponding increase in the voltage drop across the ferroelectric layer. Consequently, the voltage reaching the internal transistor gate decreases, which results in a rise in the energy barrier to the electrons coming from the source. As a result, I_{off} reduces with V_{dd} increases instead of increasing, as the case in traditional FET.

To illustrate such dependency, simulations were performed using the 7nm FinFET technology node [39] for both traditional FET (i.e., without a ferroelectric layer) and NCFET in which a ferroelectric layer with a 4nm thickness is employed. Results are extracted using the BSIM-CMG, the industrial standard compact model for FinFET technologies [26]. More details regarding the implementation are in Chapter 7. As demonstrated in Fig. 2.11, in traditional FET, reducing V_{dd} results in lower I_{off} and thus lower leakage power. However,

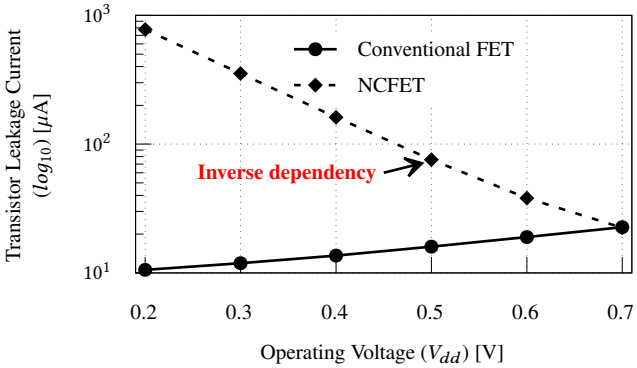


Figure 2.11.: Leakage current (I_{off}) of an NCFET transistor in comparison with traditional FET transistor, using the 7nm FinFET technology, over a wide range of voltages. NCFET exhibits an inverse dependency of leakage current with V_{dd}, unlike traditional FET.

unlike traditional FET, reducing V_{dd} in NCFET results in a noticeable increase in I_{off} and thus higher leakage power.

Importantly, in traditional FET, reducing the supply voltage of a processor reduces both dynamic and leakage power, and hence both total power and total energy are minimized. Therefore, power and energy can always be minimized by selecting the minimum voltage that sustains the required frequency (i.e., V/f pair) when operating in the nominal operating region (i.e., super threshold region). Therefore, traditional power/energy management techniques select V/f pair at design time, based on the optimization goal, minimizing power/energy. Because these dependencies might vary among different technologies, power and energy management techniques must be revisited and investigated when a new technology is introduced. This holds even more when it comes to emerging technologies in which the underlying physics fundamentally differs from traditional FET.

(e) Novel trade-off: NCFET results in inverse dependency of the leakage power over voltage at the device level as illustrated in Fig. 2.11. However, this also will affect the whole circuit and even propagate to the system level as well. For instance, examining a processor designed with traditional FET technology, and by scaling down the operating voltage (starting from nominal value), and by operating at the maximum possible frequency at that voltage, the total power always reduces toward low voltage, as shown in Fig. 2.12(a). In this

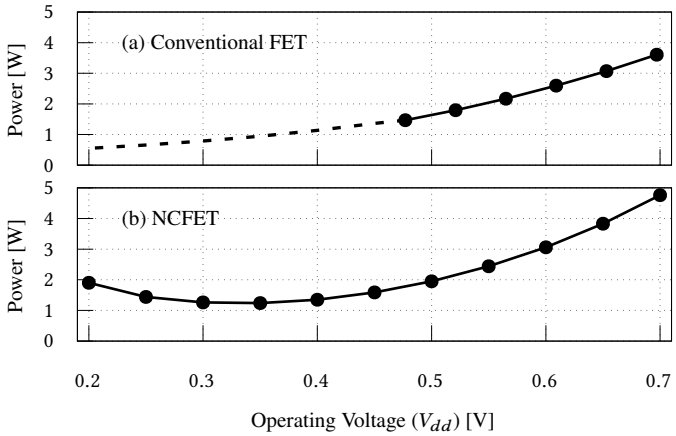


Figure 2.12.: Total power over a wide range of voltage for a processor designed with (a) traditional FET, (b) NCFET using 4nm thickness of the ferroelectric layer. The total power of the traditional FET-based processor decreases always with voltage. NCFET processor shows a novel trend where power decreases with voltage, until an inflection point appears, then power starts to increase again.

investigation, a full processor is examined within the normal operation region (i.e., super threshold) where voltage is only between 0.45V and 0.7V, because lower than that the processor would enter the NTC region (see Section 2.1). However, the expected power values are on the dashed line. On the other hand, repeating the same investigation for the same processor designed with NCFET using a 4nm ferroelectric layer shows a unique trend for power. Unlike traditional FET, NCFET-based processor shows a novel trend where power starts to decrease with voltage until reaching an inflection point, then it starts to increase again. This because the leakage power becomes dominant, and reducing the voltage further reducing the dynamic power but stronger increases the leakage power. Therefore, the minimum power can be achieved on a voltage that is higher than the minimum required. More details regarding this investigation and implementation are in Chapter 7 and Chapter 8.

This, in turn, results in a novel trade-off between power components (i.e., dynamic and leakage power), which has far-reaching consequences on the existing power/energy management techniques, as they always aim for minimum power/energy at minimum voltage [128]. Hence, Dynamic Voltage Scaling (DVS) and Dynamic Voltage/Frequency Scaling (DVS) techniques for pow-

er/energy minimization in NCFET must be aware of this property. Therefore, this dissertation presents the first NCFET-aware power/energy management techniques of this kind.

(f) NCFET-based circuit optimization: Due to previously mentioned characteristics of NCFET, this has two key implications when it comes to high-performance and low-power applications [128][129]: (1) compared to traditional FET, NCFET-based circuits can be operated at higher frequencies without the need to increase the voltage. (2) compared to traditional FET, NCFET-based circuits can be operated at the same frequency but at a much lower voltage. Under these characteristics, NCFET can be differently employed/configured based on the optimization goal.

(g) Manycore design with NCFET: Multicore and Manycore designs emerged to overcome the physical constraints of a single-core [148]. For instance, homogeneous manycore integrates cores with identical microarchitectures [27], and heterogeneous manycore uses cores with different microarchitectures [80]. Notably, the integration of cores with different microarchitectures comes with a high design cost [142].

NCFET technology is able to boost the performance of the already existing designs (i.e., circuits) by only altering the underlying transistors. Therefore, a heterogeneous NCFET-based manycore can be achieved by employing one or more cores in NCFET while the remaining cores are not affected. Even more, the ferroelectric layer can vary among cores, which creates another type of heterogeneity. This type of heterogeneity, which is presented in this dissertation (details in Chapter 9), comes *without changing the microarchitecture*. This allows designers to have benefits similar, if not even better, to conventional heterogeneous manycore without the associated overheads as all the cores share the same microarchitecture.

Prior works: Many techniques have been proposed to minimize the processor's power and energy. DVFS (dynamic voltage and frequency scaling) is used in almost all modern processors to optimize power/energy while meeting performance constraints[30]. Conventional DVFS selects the minimum frequency and voltage required to satisfy the performance constraint, and hence minimizes power/energy in FinFET technology [91, 45]. However, because NCFET could inverse the dependency of leakage power on voltage, existing DVFS techniques become sub-optimal, as has been discussed previously.

When it comes to the optimal energy point in FinFET technology, many studies (e.g., [72, 51]) showed that operating processors near-threshold voltage helps to achieve such a goal. Operating at such a low voltage leads to severe performance degradation, limiting the applicability of near-threshold computing.

Recently, several works have explored NCFET processor design and optimization. [115, 131] presented a comparison between FinFET and NCFET processors under different configurations (i.e., ferroelectric thicknesses). [131] showed an AES circuit design with NCFET and demonstrated 4× power-saving while operating at the same frequency as the baseline FinFET. [7] showed how NCFET could revive the prior trends in processor design for voltage and frequency studying a single-core processor. [115] illustrated how NCFETs impact the performance, power and temperature of a processor. Recently, a small number of works demonstrated voltage and frequency scaling for NCFETs [128, 125], which are part of this dissertation.

Many works have introduced heterogeneity into multi-/many-cores to achieve better performance and energy/power-efficiency. Device-level heterogeneity, such as threshold voltage, power gating, and voltage/frequency scaling, allows designers to achieve higher energy efficiency compared to a homogeneous processor. [161] proposed designing heterogeneous multi-/many-cores comprising steep slope devices to reach higher performance and better energy efficiency at different threshold voltages. Similarly, [63] proposed using different threshold voltages in addition to different supply voltages to reduce the leakage power for mobile applications. Power gating is a key design choice for better energy efficiency [174, 15].

Distinguishing from existing work: This dissertation demonstrates a comprehensive analysis of NCFET-based circuits, showing the far-reaching consequences of such an inverse dependency with respect to the existing power/energy management techniques. Moreover, it demonstrates that state-of-the-art Dynamic Voltage Scaling (DVS) and Dynamic Voltage/Frequency Scaling (DVS) techniques are sub-optimal with respect to NCFET. Therefore, it presents NCFET-aware DVS and DVFS techniques that can optimally minimize the power/energy of an NCFET-based processor. These techniques do not intuitively select the minimum operating V/f pair at design time as in traditional techniques. Instead, an advanced model is presented for runtime V/f pair selection following optimization goals and the dynamics of the application being executed. In addition, this dissertation also presents a novel heterogeneous manycore design with NCFET technology to improve the energy efficiency of

the conventional heterogeneous manycore design. Such novel design eliminates the heterogeneity at the microarchitecture level while achieving similar benefits, even better, of the conventional design.

Part I.

Reliability and Low Power Design under Conventional CMOS

3. Optimal Energy Point in Parallelized Near-Threshold Computing

Near-Threshold Computing (NTC) enables circuits to operate close to their optimal energy point (OEP) to maximize their energy efficiency through aggressive voltage scaling, resulting in a significant performance loss. When it comes to processor circuits, performance loss can be recovered through parallelizing across cores. This chapter presents the limitations to employ Near-Threshold Computing (NTC) w.r.t. energy efficiency and performance, and how parallelization can potentially overcome these limitations. In addition, it shows the impact that parallelized NTC has on shifting the Optimal Energy Point (OEP) out of the near-threshold region. However, relying on EDA tools, standard cell libraries are typically used to design digital circuits, such as processors. These libraries are not intended for NTC. Therefore, EDA tools must be extended for NTC first.

3.1. Voltage-Aware NTC Design

Standard cell libraries are typically developed and optimized to operate within the super-threshold region (i.e., nominal voltage) and not within the near-threshold region. Multi-Corner Multi-Mode (MCMM) are multiple executions of static timing analysis used to design digital chips across all modes and corners concurrently. MCMM covers few voltages and operating corners. MCMM analysis can be employed, in which the delay and power of circuits are estimated by the EDA tools through the interpolation between the available corners, while results suffer from high inaccuracies. However, for operating voltage (V_{dd}) out of the voltage range between the available corners, the tool cannot perform any delay/power estimations because extrapolation is impossible. Therefore, to accurately find the OEP of a processor, it is inevitable to create voltage-aware cell libraries in which the entire operating voltage range is fully covered. Importantly, as demonstrated in Section 2.1, operating voltage reduction has uneven impacts on both power and delay of the cells [123]. Therefore, the power and delay of every cell within the circuit should be examined over voltages. In the following, the creation of the voltage-aware cell libraries is presented and how they can be employed to perform delay and power analysis. Afterward, circuit design and the results of a full processor are presented.

Voltage-aware cell libraries: The presented analysis is targeting the 45nm technology node. However, the proposed methodology is not limited to any

specific technology, and it can be applied to any technology nodes. Firstly, the SPICE netlist is employed of different combinational and sequential cells from the 45nm open-cell library from Nangate [95]. To model the electrical characteristics of pMOS and nMOS transistors, Predictive Technology Modeling (PTM) is employed [111] for the high-performance 45nm technology. To model the dependencies of MOSFET's parameters on voltage, the industry-standard compact MOSFET model (BSIM model) is considered [36]. Finally, the HSPICE simulator is employed to accurately measure the delay and power of every standard cell under the impact of voltage reduction. To take the impact of the operating conditions into account, 7 input signal slews along with 7 output load capacitances are considered, which is typical in both industrial and academic cell libraries [151]. All the measured 49 values of every cell are then stored within a lookup table for timing and power information using the standard *liberty* format covering the voltage range from 1.2V to 0.2V with a 0.05V step. The cell libraries are compatible with standard tool flows for chip design, and hence, they can be employed directly within the design flow.

Voltage-aware timing and power analysis: By using the voltage-aware cell libraries, the mature algorithms of the existing EDA commercial tool can be leveraged in order to accurately estimate the energy consumption of any circuit (regardless of its complexity) under the entire range of V_{dd}. Therefore, first, the RTL design is synthesized at the nominal $V_{dd} = 1.2$V. Then, iteratively, voltage-aware delay and power analysis are performed, for the obtained netlist, at every voltage step. This allows an accurate estimation of the maximum delay of the circuit as well as its total power consumption. Hence, the consumed energy can be estimated at every voltage step to determine how the energy efficiency improves while voltage reduces towards precisely selecting where the OEP occurs.

Due to the uneven impact of voltage reduction on the delay of cells, the path that was initially critical at nominal voltage might not remain critical when V_{dd} is reduced, where another path might become critical at the lower voltage. Therefore, analyzing the entire circuit (i.e., netlist) is necessary, which is considered in the presented analysis.

3.2. Parallelized Near-Threshold Computing

Reducing the supply voltage leads to significant performance loss (see Section 2.1). To compensate for the significant performance loss at NTC, more parallel cores can be employed, maximizing the performance under a specific power budget while reducing the voltage towards NTC. However, the Amdahl serial factor (S_f) [5], which represents the obstacles that limit the gain of parallel execution (e.g., application sequence, communication latency, synchronization, I/O, etc.), should be considered. Therefore, finding OEP, in which the added cores provide the maximum energy efficiency, becomes more complicated due to the relation between the added cores, consumed power, and overall gained performance. However, this work focuses on exploring the design space at design time independent of the applications that might be executed. Therefore, in the presented analysis, varied S_f cases are covered that represent all possible applications instead of considering a specific application behavior. Later, when the exact S_f of the running application is known, designers can optimally select OEP from the already examined corners.

After that, the delay and power are examined using the voltage-aware analysis considering the minimum delay (i.e., maximum performance) of the studied design and its power at every voltage step. Then, at every voltage step, the maximum possible number of cores (n) is estimated within a given power budget. Finally, the achievable performance gain is evaluated under the impact of various S_f, as Eq. (3.1) clarifies. As shown from Eq. (3.1), the performance gain is defined as the ratio between the original performance of the single core at the nominal V_{dd} and the new performance obtained from parallelized NTC while considering the reduction in performance when reducing V_{dd}.

$$Performance\ Gain = \frac{S_c}{S_f + \frac{(1-S_f)}{n}} \tag{3.1}$$

Where S_c represents the core's performance at the low V_{dd} normalized to the core's performance before voltage scaling (i.e., nominal V_{dd}). Note that $0 \leq S_c \leq 1$.

Considering power saving in parallelized NTC: Besides its impact on the overall performance, S_f has an important impact on the total power consumption. The sequential portions of an application will always be executed on a single core. Thus, only one core out of n available cores will be active during

the sequential computation phase, whereas the remaining $(n-1)$ cores are idle. These cores consume only leakage power. On the other hand, during the phase of parallelism, all n cores are active. Hence, when the total consumed power is estimated to check against the predetermined budget, it is essential to consider the impact of S_f. The total power consumption can be conservatively estimated, assuming all cores are always active. The total power can then be calculated as Eq. (3.2) shows, based on [169].

$$W = \frac{W_c + (n-1)W_c k_c S_f}{S_f + \frac{1-S_f}{n}} \tag{3.2}$$

Where W is the total power consumption normalized to the consumed power at nominal voltage, W_c is the active core's power consumption relative to the power at nominal voltage, k_c is the fraction of the core's idle power normalized to the active core's power, S_f sequential fraction and n number of cores. Note that $0 \leq w_c \leq 1$ and $0 \leq k_c \leq 1$.

Impact of serial factor: S_f in parallelized NTC has two-fold impacts: (a) it leads to a loss in the performance gain, which is achieved from parallelized cores, (b) it reduces the total consumed power, and hence, the provided power budget is not fully utilized. Unlike the state-of-the-art approach, which neglects the impact of S_f on the total power consumption, this work accurately estimates the consumed power considering both sequential and parallel phases as Eq. (3.2) shows. The equation shows that during the sequential computation phase, one core is in the active state consumes W_c while all idle (n-1) cores consume $(n-1)W_c k_c S_f$. During the parallel computation phase, all cores are active, and they consume nW_c.

Performance per Watt: A trade-off is essential to find the OEP under the optimization conflicts (i.e., performance, power, and S_f). To accurately evaluate the overall gain in the parallelized NTC, it is necessary to evaluate the performance per watt (PW) as Eq. (3.3) shows. PW allows designers to figure out how the power budget is efficiently utilized and, thus, which V_{dd} maximizes the efficiency of the multi-core processor.

$$PW = \frac{S_c}{W_c + (n-1)w_c k_c S_f} \tag{3.3}$$

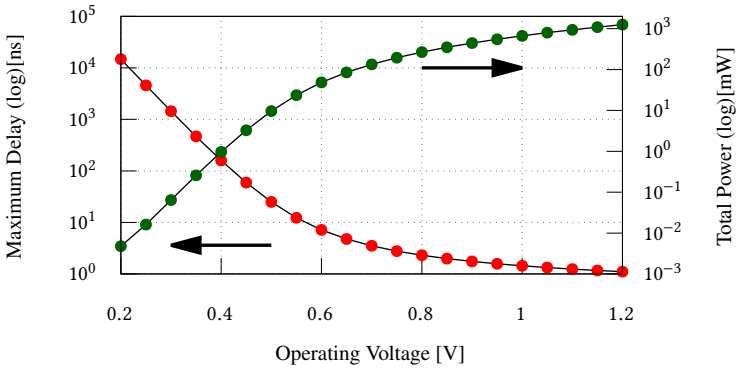


Figure 3.1.: Timing and power of the 64-bit Rocket Processor [159] over voltage.

Implementation: To automatically perform a design-space exploration, at design time, to find the OEP precisely, an automated framework is implemented. This framework considers the interactions between power and delay of the processor at every voltage together with the required power budget as well as the impact of S_f. At every voltage, the maximum number of cores within a given power budget is calculated to ultimately maximize the possible performance gain. Then, the optimal voltage is determined where the efficiency is maximized. The efficiency in the parallelized NTC is represented by Performance per Watt (PW).

3.3. Evaluation and Experimental Results

First, the experimental setup is presented. Then, the evaluation of the parallelized NTC and the resulted efficiency are discussed. The flow diagram of the implementation is presented in Appendix A.1.

Experimental setup: This work focuses on analyzing only the logical components (i.e., CPU). However, other components like caches, where errors could occur at low voltages (i.e., NTC), are assumed to operate at a higher V_{dd}, similar to the Intel implementation in [66]. The state-of-the-art 64-bit Rocket Processor [159] is considered for analysis, a RISC-V industry-competitive processor from Berkeley.

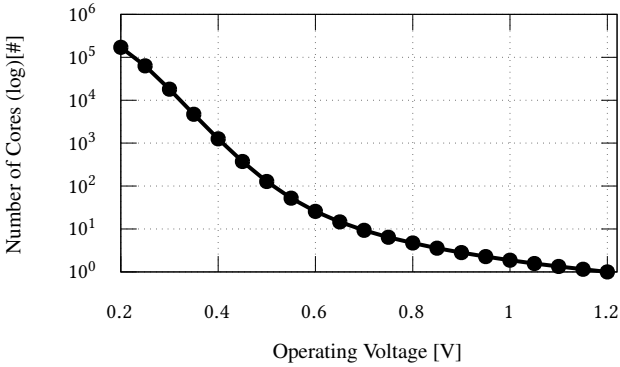


Figure 3.2.: Maximum number of cores within the power budget of the power at the nominal voltage of 1.2V.

OEP for single-core NTC: After generating the RTL description of the 64-bit Rocket processor, the gate-level netlist is generated by synthesizing the RTL using the original standard cell library at the nominal voltage (1.2V). Then, voltage-aware timing and power analysis are performed for the generated gate-level netlist over the whole voltage range. Fig. 3.1 summarizes the obtained power and timing results from nominal voltage 1.2V all the way down to 0.2V. The OEP of the processor appears around 0.35V.

OEP for parallelized NTC: S_f could change OEP based on the optimization goal (i.e., power or performance). However, different applications with different S_f result in different behavior. Therefore, for generalization, while considering the application's behavior is unknown in prior, a wide range of S_f values are examined in which varied runtime behaviors are reflected. In case S_f changes at runtime, designers could heuristically predict OEP from the expected range of S_f, and a conservative OEP is then selected. As explained earlier in Section 3.2, the key focus is obtaining V_{dd} where OEP occurs along with the maximum number of cores in which the performance is maximized. Concurrently, a certain power budget is fulfilled. For instance, considering the power budget at nominal voltage 1.2V, the maximum number of cores is shown in Fig. 3.2. The number of cores will change based on the power budget.

In Fig. 3.3(a and b), the achieved speedup from parallelized NTC is evaluated under different power budgets. Power budgets are examined as the total power consumption at different voltages and for two different cases of serial factors:

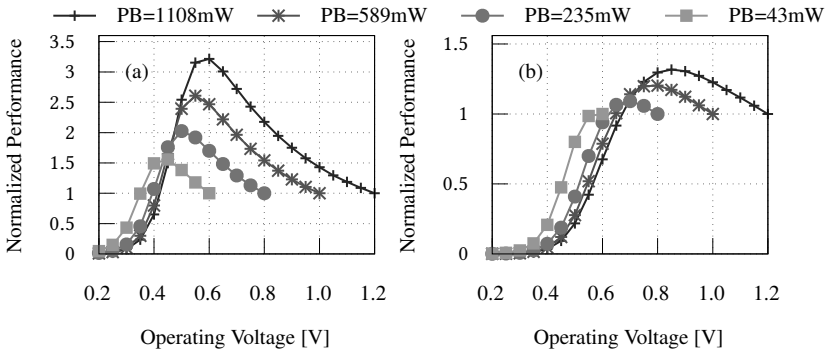


Figure 3.3.: Performance improvement at parallelized NTC under different power budgets (PB) for the case of S_f of (a) 0.01 and (b) 0.2. Smaller power budget pushes OEP towards the near-threshold region but when S_f is larger the OEP moves towards the super-threshold region. Note that the power budget estimated for every V_{dd}

.

$S_f = 0.01$ and $S_f = 0.2$. As shown in Fig. 3.3(a), providing a smaller power budget pushes OEP towards the left side (i.e., lower V_{dd}). However, higher S_f has a contradictory effect as it pushes OEP towards the right side (i.e., far from NTC), as shown in Fig. 3.3(b).

Fig. 3.4(a, b, and c) show the impact of various S_f on the gained performance as well as on shifting OEP. The evaluations have been done under one power budget, which is the total power at the nominal voltage of 1.2V. As expected, higher S_f leads to a lower speedup as Fig. 3.4(a) demonstrates. S_f plays a major role in determining where OEP occurs. Higher S_f shifts OEP towards the super-threshold region (i.e., higher voltages). For instance, a serial factor of 0.01 (i.e., 1% of the application is sequential while the rest are fully parallelized) leads to shifting OEP to 0.6V.

To explore further the impact of S_f, Fig. 3.4(b) shows the corresponding normalized total consumed power for different S_f values. As shown, when S_f becomes higher, the application spends more time within a single core and hence the rest of the cores are idle. Accordingly, the power budget is not fully exploited. In other words, higher S_f results in less utilization of the provided power budget. Thus, evaluating the overall efficiency must be represented by the Performance per Watt (details in Section 3.2) to determine where OEP precisely occurs. As Fig. 3.4(c) demonstrates, the impact of S_f becomes

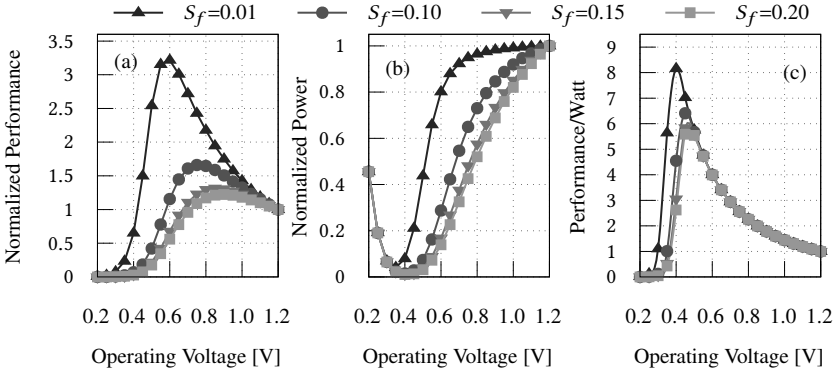


Figure 3.4.: The investigation of the impact of the serial factor on performance gain (a), power budget utilization normalized to the power at nominal voltage (b), and performance per watt (c). Higher S_f results in less performance gain and less power budget utilization (a) and (b). The maximum performance per watt occurs at (or close to) NTC, see (c).

marginal compared to the previously observed impact on both speedups (see Fig. 3.4(a), and Fig. 3.3(b)) and utilized power budget (see Fig. 3.4(b)). As shown, OEP under different S_f values approximately remains within or close to the near-threshold region. Hence, when optimizing for performance alone, OEP moves far from the near-threshold region (i.e., OEP occurs at $V_{dd} > 0.5$V). However, when optimizing for performance per watt, OEP remains close to the near-threshold region (i.e., OEP occurs at $V_{dd} < 0.5$V). Therefore, the optimization goal plays a major role in finding where the OEP occurs.

3.4. Summary and Conclusions

In summary, NTC is one of the proposed/available solutions to maximize the energy efficiency of the processors of the currently used technology. However, employing OEP leads to significant performance degradation. Employing more cores in parallel could recover the performance loss. Parallelized NTC plays a major role in shifting the OEP based on the selected optimization goal (performance or power optimization). Notably, at runtime, the application being executed will show different activities where the parallel utilization of the cores is governed by the serial execution factor S_f, which will affect OEP. This makes OEP inefficient to use, and preservative OEP should be selected.

4. Reliability-aware design under the interdependencies between voltage fluctuation and BTI aging

This chapter investigates the joint impact that aging in conjunction with IR-drop has on the circuit's delay. Both phenomena increase the circuit's delay. Traditionally, the required timing guardbands of both are individually estimated, and the final guardband is the magnitude of summing them up. Such pessimistic estimation leads to significant performance loss due to the largely employed timing guardband. However, using a physics-based aging model, which is able to precisely consider fluctuations in supply voltage (V_{dd}), the existing interdependencies between the two phenomena is demonstrated. Based on that, a novel technique is presented in this chapter to accurately estimate the smallest, yet sufficient, timing guardband that protects circuits against both phenomena and sustains reliability.

4.1. Impact of IR-drop and BTI Aging on Circuits Reliability

In the following, the origins behind degradations caused by IR-drops and aging are presented and how reliability is affected.

4.1.1. Voltage Fluctuation due to IR-drops

IR-drop is a fundamental property of any electrical circuit, which is mainly due to the radical parasitics (e.g., resistances and capacitances) inside the non-ideal Power Delivery Network (PDN) (i.e., power lanes)[20], which causes fluctuations in the supply voltage (V_{dd}). Two types of IR-drop are typically investigated when designing circuits [167, 145]:

1. **Static IR-drop analysis** aims to quickly obtain a rough estimation of the potential IR-drop in the circuit in the absence of any activities in the early stages of the chip design. Such analysis relies on DC analysis of the PDN considering the average power of the circuit to estimate a constant driving current for every cell within the circuit. Such analysis guides designers early to optimize the PDN further [127].
2. **Dynamic IR-drop analysis** aims to obtain more accurate and realistic estimations of the potential IR-drop across the chip under the impact of switching activities. The pattern of switching activities is driven by the

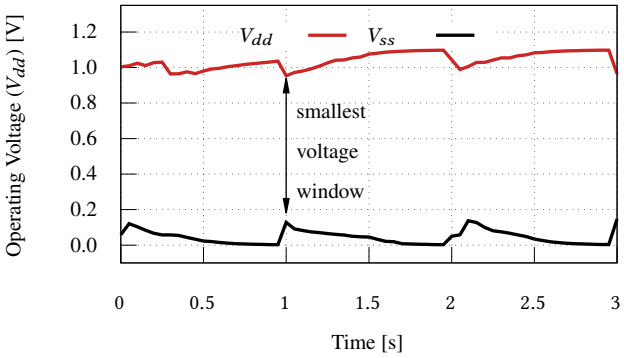


Figure 4.1.: V_{dd} and V_{ss} waveforms due to IR-drop extracted from 32-bit microprocessor implemented at the 45nm. As shown, the fluctuation in V_{ss} cannot be neglected and it should be considered. Therefore, a voltage window ($V_{dd} - V_{ss}$) needs to be considered when analyzing the impact of IR-drop on circuits' delay.

workloads being executed. The higher the activities, the larger the drop in V_{dd} due to the rapid increase in the current demands. With each rising edge of the clock, the simultaneous switching of standard cells results in peaks in driving current and thus high IR-drop[99, 145]. However, with the rapid technology scaling, IR-drop becomes a serious problem, and its consequence on circuit's delay is continuously increasing (see Chapter 1).

Voltage reduction, due to IR-drop, manifests itself as an increase in the circuit's delay [3]. This is due to the direct impact of V_{dd} on the propagation delay of cells and the indirect impact by lowering the I_{on} current of transistors (see Eq. (2.4) and Eq. (2.5)). Therefore, to suppress the deleterious impact of IR-drops, designers, in addition to optimizing the PDN, include a timing guard-band on top of the critical path delay of the circuit (see Eq. (2.6) and Eq. (2.13)) to always met the timing constraints and avoiding timing violations.

Voltage Window (V_{win}): When IR-drop occurs, both V_{dd} and V_{ss} are affected, in which V_{dd} drops below the nominal level, whereas V_{ss} increases above the nominal ground level (0V) [47]. This is because both V_{dd} and V_{ss} share similar power topologies over non-ideal power lanes within the PDN. Voltage, by definition, is an electric potential difference between two points, in this case between V_{dd} and V_{ss}. Therefore, considering V_{dd} fluctuations only leads to underestimating the overall impact of IR-drop on circuits. In order to

consider the impact of IR-drop on both V_{dd} and V_{ss}, a voltage window (see Eq. (4.1)) [47] is typically employed while analyzing the impact of IR-drop on the circuit's delay. The worst case occurs when the voltage window is minimum (i.e., maximum IR-drop). This leads to the highest impact on the circuit's delay. Fig. 4.1 shows an example of V_{dd} and V_{ss} waveforms for a microprocessor circuit designed at 45nm node. As shown, V_{dd} and V_{ss} fluctuate from nominal levels of 1.1V and 0V, respectively. The worst-case IR-drop occurs at the smallest voltage window.

$$V_{win}(t) = V_{dd}(t) - V_{ss}(t) \tag{4.1}$$

$$\text{IR-drop(Max)} = V_{dd}(nominal) - Min(V_{win}(t)) \tag{4.2}$$

4.1.2. Aging-Induced Degradation

Bias temperature instability (BTI) is one of the key aging phenomena in the current technology nodes [114], and it is predicted to remain a key degradation in the upcoming nodes [89]. Generated defects due to BTI are undesired charges that interact with the applied electric field on a transistor [87] which leads to an increase in its threshold voltage (V_{th}). As explained earlier (see Section 2.2), ΔV_{th}, in turn, leads to a reduction in the I_{on} current and thus an increase in the propagation delay of standard logic cells.

Impact of lowering V_{dd} on BTI: The underlying mechanisms of BTI are driven by the operating voltage (V_{dd}). Higher V_{dd} results in more vital electric fields, and therefore, more defects due to BTI are generated, leading to a larger ΔV_{th}[87]. However, when the V_{dd} is reduced, some of the generated defects are healed, leading to a reduction in ΔV_{th}[127]. Thus, lowering V_{dd} mitigates the BTI-induced ΔV_{th} due to the occurring partial recovery. Notably, according to the recent results in [52, 106], BTI is able to follow even the ultra-fast voltage changes in the current technology.

Notably, the presented measurements in [136] besides the simulation-based analysis in [133] show that the same BTI-induced ΔV_{th} results in a larger delay increase at lower V_{dd}. This is expected because I_{on} in the transistor is proportional to $(V_{dd} - V_{th} - \Delta V_{th})$ (see Eq. (2.4)). Hence, the impact of ΔV_{th} on the delay of a standard cell is *amplified* with V_{dd} reduction.

4.1.3. Bringing Voltage Fluctuation and BTI Together

Voltage fluctuations due to IR-drops influence BTI in two ways: (a) *amplify* the impact of induced ΔV_{th} on delay, (b) partially *mitigate* aging by reducing the induced ΔV_{th}. Because the underlying BTI mechanisms are able to follow ultra-fast V_{dd} changes in the current technology, as in [52, 106], it cannot be excluded that IR-drop and BTI-induced degradation influence each other.

4.2. Interdependencies between Voltage Fluctuations and BTI Aging

The underlying mechanism of BTI is explained here as well as how they react to V_{dd} changes along. Afterward, rates of defects generation/recovery due to BTI are described and how they can react to V_{dd} changes caused by IR-drops. Later, we describe how drops in V_{dd} can mitigate BTI-induced ΔV_{th} due to the partial recovery and how the BTI-induced degradation amplifies the impact that V_{dd} has on the delay.

4.2.1. Impact of Voltage on the Underlying Mechanisms of BTI

BTI is the generation of defects at the Si-SiO_2 interface and the capture/emission of carriers in defects inside the gate dielectric of a transistor. Both physical mechanisms depend on the strength of the electric field over the insulator. Therefore, the applied V_{dd} plays a major role in driving the underlying mechanisms of defect generation in BTI [87].

In order to capture the complex dependency of BTI on V_{dd}, physics-based BTI models must be employed because empirical BTI models (e.g., [85, 170]) can deal solely with a constant voltage. Hence, they cannot be employed to investigate the interdependencies between voltage fluctuation and BTI.

In this work, the state-of-the-art physics-based BTI model from [106] is employed (details are in Appendix A.3). Therefore, the BTI Analysis Tool (BAT) [127] is employed, which consists of various models to model the different underlying mechanisms of BTI accurately. BAT has been validated against

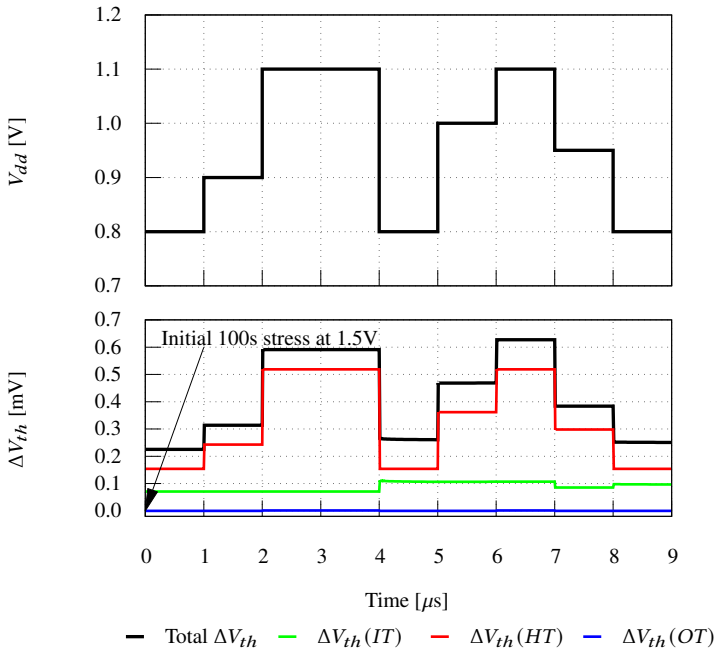


Figure 4.2.: Voltage dynamics governing the BTI-induced degradation (ΔV_{th}). When V_{dd} increases, ΔV_{th} increases due to BTI defects are being generated. When V_{dd} drops, BTI recovers, and hence ΔV_{th} decreases. To show the non-zero contribution of interface traps ($\Delta V_{th}(IT)$) and to demonstrate interface trap and hole trap interaction, the trace was prefaced by a 100s stress phase at 1.5V mimicking a month of operation at nominal voltage. HT reacts almost instantaneous to changes in voltage, while OT cannot contribute on this time-scale to the overall/total ΔV_{th} (further details in [127])

measurements, and it estimates ΔV_{th} induced by any arbitrary voltage waveforms. BAT covers the three defect types caused by BTI [127]: Interface Traps (IT), Oxide Traps (OT), and Hole Traps (HT) (see Appendix A.3) [127]. Interface Traps (IT) are broken *Si-H* bonds at the channel interface. Oxide Traps (OT) captures carriers in the newly generated defects in the interface layer (low-k gate dielectric). Hole Traps (HT) are capturing carriers in pre-existing defects due to imperfect manufacturing (see Chapter 2). More details on the defect generation mechanisms are available in [50]. Only by modeling the defect types mentioned above, BTI model can then model the voltage dependency of BTI correctly because each defect type has a different voltage acceleration factor [106]. The combination of the three defect types allows

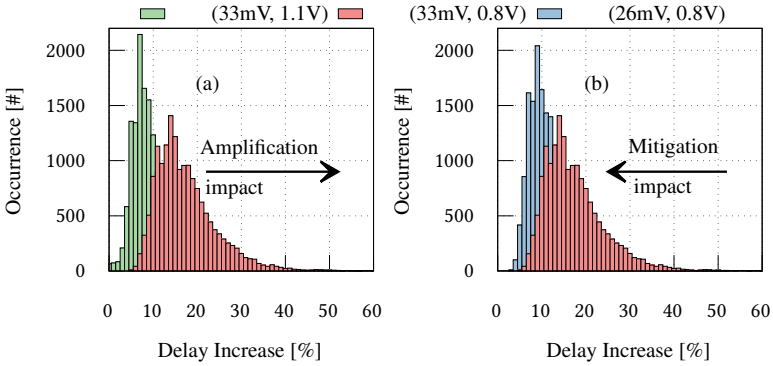


Figure 4.3.: (a) shows the amplification impact in which the same ΔV_{th} results in larger a delay increase in all standard cells within 45nm cell library. (b) shows how a reduction in ΔV_{th} due to BTI recovery results in mitigating the delay increases. Each pair refers to (ΔV_{th},V_{dd}) scenario.

considering how fluctuations in V_{dd} impact the resulting BTI-induced ΔV_{th}. To demonstrate that, Fig. 4.2 hows the resulting BTI-induced ΔV_{th} follows changes and dynamics in V_{dd}. As shown, when the V_{dd} drops (top figure), a partial recovery of the induced ΔV_{th} occurs, and when V_{dd} becomes high, the BTI degradation gets accelerated, and an increase in ΔV_{th} is observed [127].

4.2.2. Short- and long-Term BTI Degradations

BTI consists of two parts; long-term and short-term degradations [127]. Long-term BTI degradation is due to the accumulation of the generated defects over time. In practice, the longer the V_{dd} is applied to the transistor, the higher the BTI degradation becomes. Short-term BTI degradation occurs at a significantly smaller time scale. Traditionally, BTI was always considered a long-term phenomenon. However, scaling transistors to an atomic level beside the advent of ultra-fast sub-microsecond transistor measurements [116], revealed that BTI is a very fast phenomenon where its degradation can be observed after a couple of μs [116, 117, 52] (for more details, see [127]). The measurements are in good agreement with the predictions of recent physics-based BTI models, like the BAT model, which is employed in this work.

4.2.3. Joint Impact of IR-drop and BTI

Aging and IR-drop can mutually influence each other, creating interdependencies between them (aging ↔ IR-drop). On the one hand, BTI-induced ΔV_{th} amplifies (aging → IR-drop) the impact that IR-drop has on the circuit's delay. On the other hand, IR-drop mitigates, to some degree, (aging ← IR-drop) the BTI-induced ΔV_{th} due to the partial recovery as V_{dd} is dropped. In the following, the occurrence of the two dependencies occur is demonstrated.

(a) The Amplification Impact (Aging → IR-drop): When V_{dd} is dropped, the resiliency of the circuit against the increases in V_{th} decreases. For instance, [133] reported that the impact that the same ΔV_{th} of 10mV has on the delay of a circuit increases from 5% to around 10% when V_{dd} is reduced from 1.2V to 0.8V. To further investigate the amplification, the delay increase caused by ΔV_{th} of 33mV (i.e., typical one year of aging) is examined, using the physics-based aging model, across all standard logical cells within the 45nm standard library [95] for two different voltages; 1.1V and 0.8V. Fig. 4.3(a) shows how the distribution of the cell delay increase moves further towards the right side when V_{dd} drops from 1.1V to 0.8V. This demonstrates how the impact of the same ΔV_{th} is amplified when V_{dd} drops.

(b) The Mitigation Impact (Aging ← IR-drop): IR-drop can partially mitigate the BTI-induced ΔV_{th}. When the voltage is reduced, BTI-induced ΔV_{th} reduces as well, which results in BTI recovery. As voltage fluctuations occur in the μs scale, recovery follows as it is extremely fast [106, 52]. To demonstrate the impact of mitigating ΔV_{th} on reducing the delay increase of cells, Fig. 4.3(b) shows the delay increase distribution of cells and how it moves towards the left side when ΔV_{th} gets reduced by 7mV (i.e., ΔV_{th} is reduced from 33mV to 26mV).

For instance, Fig. 4.4 presents an example of the resulting ΔV_{th} due to IR-drop (details in Section 4.4). As shown, V_{dd} drops below the nominal voltage of 1.1V have led to a reduction of the induced BTI-induced ΔV_{th}.

Finally, knowing the joint impact of both degradation phenomena on cells' delay (i.e., aging amplifies the impact of IR-drop, and IR-drop partially mitigates the impact of aging degradation on standard cells' delay), estimating the smallest, yet sufficient, timing guardband to sustain reliability under both phenomena is crucial.

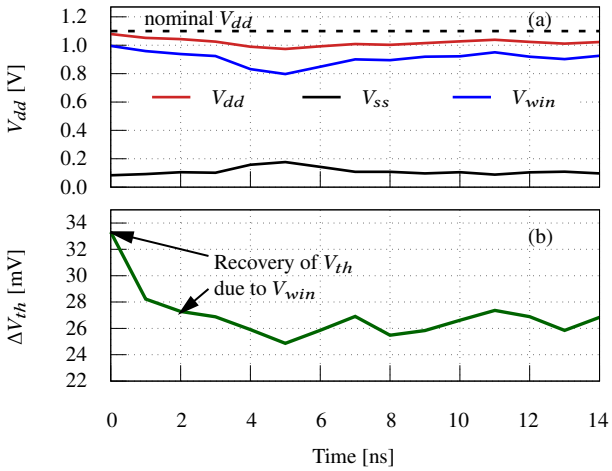


Figure 4.4.: (a) shows an example of V_{dd} and V_{ss} traces during voltage fluctuation caused by IR-drop as well as the corresponding voltage window (V_{win}) trace. (b) demonstrates the resulting ΔV_{th} waveform. A drop in V_{dd} results in BTI recovery. As shown, the resulting ΔV_{th} waveform follows the fluctuations present within the V_{win} trace.

4.3. Cross-layer Implementation Under Aging and IR-drop Effects

To estimate the smallest, yet sufficient, timing guardband that protects circuits against the joint impact of aging and IR-drop, the flow diagram shown in Fig. 4.5 is implemented. The implementation is a cross-layer implementation that links the physical and circuit levels under aging and IR-drop effects. The implementation is divided into the following steps:

(1) Physical design: Following the standard design flow of digital circuits from RTL to GDSII, the layout of the targeted circuit is created including an optimized PDN. Because starting from optimized PDN is essential to investigate the interdependencies between aging and IR-drop properly, PDN is iteratively optimized until the static IR drop becomes close to 0%, whereas dynamic IR-drop of both V_{dd} and V_{ss} becomes below 10%. This is similar to the reported state of the art in PDN optimization [127].

(2) IR-Drop Analysis: Using the power signoff tool, which is typically required before fabrication to perform the necessary electrical and physical

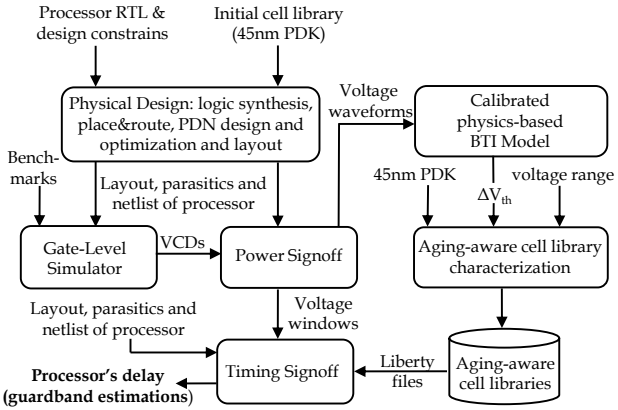


Figure 4.5.: General overview of our approach showing the phases of our implementation [127].

verifications, IR-drop is extracted for every cell within the circuit. The signoff tool estimates the power consumption over time of the circuit based on its switching activity. Then, it estimates the fluctuations in both V_{dd} and V_{ss} over time for every cell based on the existing parasitics (i.e., resistances and capacitances) of the PDN. However, as actual activities of the running workloads are unknown at design time, designers typically perform IR-drop analysis under different switching activities. Hence, this work considers switching activities ranging from 10% to 80%, covering a wide range of possible runtime activities. The voltage window waveform, using the extracted V_{dd} and V_{ss} waveforms, is employed to estimate the maximum IR-drop for every cell. Later, IR-drop is also analyzed under various benchmarks running on top of the circuit, considering workload-dependent analysis.

(3) Aging-IR-Aware Cell Libraries: When the voltage window waveform is applied to the gate-source voltage of the transistor, the physics-based BTI aging model [106, 127] model, which is able to consider the dependency of BTI on V_{dd}, calculates the generated defects over time and then estimates the maximum ΔV_{th}. The BTI model does consider any potential BTI recovery when voltage fluctuates below the nominal, i.e., IR-drop. Afterward, an aging-aware cell library is created where all standard cells are characterized in the scope of the estimated maximum ΔV_{th}. This is done using the commercial cell library characterization tool, which employs SPICE simulations to examine the delay and power of every cell under the effect that ΔV_{th} has at the targeted

voltage level. To this end, the BSIM compact modeling of high-k MOSFET is employed along with the Predictive Technology Model (PTM) card for nMOS and pMOS at 45nm [178]. However, to cover varied voltage levels which might occur due to runtime IR-drop, cell libraries for the entire range from nominal voltage 1.1V to 0.6V are created with a 50mV step.

(4) Timing Guardband Estimation: Finally, the timing signoff tool employs the created cell libraries along with the extracted IR-drop profile, containing the minimum voltage window for every cell, to calculate how the overall delay of circuit's paths will increase. Hence, the required timing guardband under the joint impact of aging and IR-drop can be accurately estimated.

Using the activity percentage of a circuit can quickly determine the needed guardband during design time, where actual activity is still unknown. For accurate estimation, a set of benchmarks is used to estimate the desired guardband for each application since each application has different activities.

4.4. Evaluation and Comparison

In the following, implementation details of the experimental setup are presented as well as present the evaluation demonstrating the individual and joint impacts of IR-drop and aging w.r.t timing guardbands in comparison with state-of-the-art approaches. This is important to demonstrate the necessity of considering the interdependencies between voltage fluctuation and aging.

4.4.1. Experimental Setup

In the following, a summary of the implementation layers is presented.
(1) EDA Tool Flows: Synopsys Design Compiler is used for logic synthesis and Cadence Innovus for layout and PDN design/optimization. For accurate IR-drop analysis and accurate guardband estimation, Cadence Voltus power signoff and Cadence Tempus timing signoff tools are employed.
(2) RTL Design: To consider a relatively complex circuit, this work targeted state-of-the-art PULPino processor [160], a 32-bit RISC-V.
(3) Cell Libraries Creation: This work employed the open-source 45nm Nangate library [95] to create IR-Aging-Aware Cell Libraries. Libraries were created under individual and joint impacts of IR-drop and aging effects in

Table 4.1.: Scenarios for comparison in the evaluation.

Scenario	Operation	IR	Aging	guardband
(a) IR-drop	IR only	Waveform	no	IR
(b) IR-Aging-1	Amplification	Waveform	ΔV_{th} = 33mV	Aging → IR-drop
(c) IR-Aging-2	Individual	V_{avg}	ΔV_{th} = 33mV	Aging + IR-drop
(d) IR-Aging-3	Individual	Waveform	Recovery	Aging ← IR-drop
(e) IR-Aging-4	Interdependencies	Waveform	Recovery	Aging ↔ IR-drop

order to enable a wide range of reliability analyses. For characterization, the Synopsys SiliconSmart tool has been used along with HSPICE.
(4) Circuit and Transistor Levels: The industrial MOSFET compact model (BSIM4.8) [26] is employed. The model is suitable for 45nm high-k MOSFET. As transistor modelcard, the 45nm high-performance PTM is employed [178]. For correct estimations of delay and power, post-layout SPICE netlists for standard cell libraries are considered. Every standard cell is characterized under (7×7) input signal slews and output load capacitances.
(5) Physical Level: As explained earlier, the state-of-the-art physics-based BTI model from [106] is employed. The model has been validated against semiconductor measurements and is able to precisely capture both short- and long- BTI degradations under arbitrary voltages.

4.4.2. Experimental Results and Comparisons

To cover various scenarios and for fair comparisons, the following case studies are implemented and analyzed.
(a) IR-drop alone: Timing guardbands are estimated under the impact of IR-drop alone, and hence, no aging (i.e., $\Delta V_{th} = 0$).
(b) IR-Aging-1 (amplification impact): Timing guardbands are estimated under the impact that IR-drop has together with the impact of BTI degradation. In this case, BTI aging recovery caused by IR drop is not considered (i.e., BTI stress without recovery). Hence, only the amplification impact is investigated.
(c) IR-Aging-2: Timing guardbands are estimated as the magnitude of summing both guardbands of aging and IR-drop individually, as in [48]. The average voltage (V_{avg}) due to IR-drop is considered, similar to [48, 100] (i.e., BTI stress without recovery).

(d) IR-Aging-3: Timing guardbands are estimated as the assumption of both individual guardbands of aging and IR-drop individually, similar to [48, 100]. Unlike the previous case, IR-drop waveforms are considered (instead of average voltage, i.e., BTI stress with recovery.) This gives fairer comparisons against our technique, which also considers the full IR-drop waveform.
(e) IR-Aging-4 (amplification and mitigation impacts): Timing guardbands are estimated under the joint impact of IR-drop and aging. Unlike case (b), BTI recovery due to IR-drop is additionally considered (i.e., BTI stress with recovery considering the entire voltage waveform under the effects of BTI recovery. Hence, both *amplification* and *mitigation* impacts are examined. All scenarios are summarized in Table 4.1 for better demonstrations.

In all the performed analyses, one year of lifetime (i.e., aging) is assumed, which results in ΔV_{th} = 33mV under BTI DC stress, in the absence of IR-drops. The switching activity that ranges from 10% to 80%, with a step of 10%, is analyzed to cover various IR-drop scenarios. Additionally, the actual IR-drop is also analyzed considering realistic switching activities caused by various workloads (benchmarks). The switching activities traces are extracted using the ModelSim tool as Value Change Dump (VCD) format.

(1) Timing guardband analysis: Fig. 4.6 summarizes the results of the required timing guardband under synthetic switching activities as well as the actual switching activities caused by different running benchmarks on top of the processor. Timing guardband is the delay increase which is caused by degradation. Fig. 4.6(a) demonstrates how neglecting the role of aging in amplifying the impact that IR-drop has on delay (i.e., aging $\rightarrow$ IR-drop) leads to underestimating the required guardband by up to 30% on average and up to 49% on the worst case. Fig. 4.6(b) shows that state-of-the-art techniques underestimate the required guardband by 12% on average and up to 20% (IR-Aging-3) and by 53% on average and up to 65% (IR-Aging-2). A lower guardband than required leads to unreliable operation during the projected lifetime as the processor becomes subject to timing violations due to the unsustainable frequency. Fig. 4.6(c) shows the impact of considering BTI recovery, caused by IR-drops, on narrowing the timing guardband. As shown, the mitigation impact reaches around 8%.

On the other hand, under realistic switching activity, similar observations and trends to the synthetic switching activities can still be observed. In this analysis, neglecting the role of aging on amplifying the impact that IR-drop has on delay leads to underestimating the required guardband by 44% as demonstrated

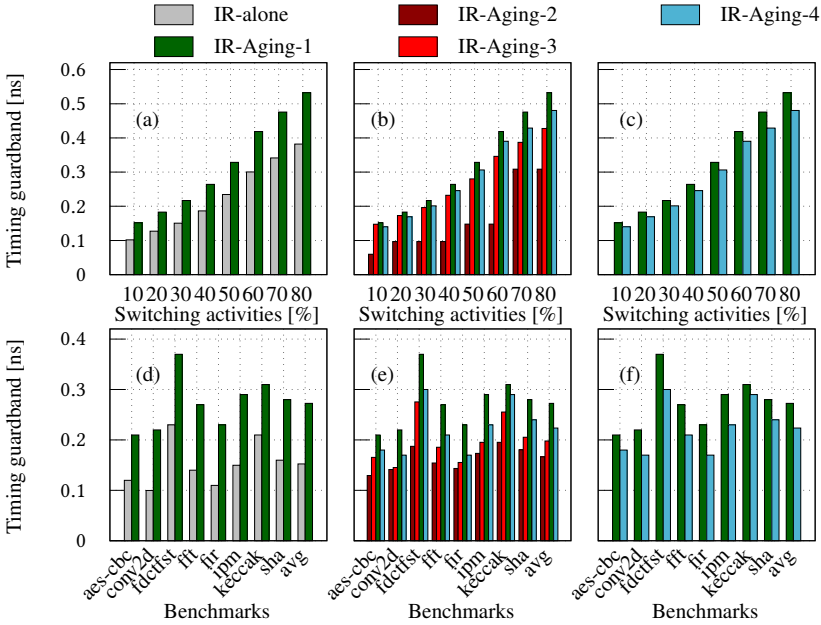


Figure 4.6.: Analysis under different switching activities (a, b, and c) as well as under different benchmarks executed on top of the processor (d, e, and f). (a+d) Evaluating the *amplification impact* due to aging, i.e., aging → IR-drop. (b+e) Comparison with state of the art demonstrating how neglecting the amplification impact leads to underestimating guardbands. (c+f) Evaluating the *mitigation impact* due to IR-drop, i.e., aging ← IR-drop, demonstrating how considering interdependencies, aging ↔ IR-drop, is necessary to estimate efficient guardbands.

in Fig. 4.6(d). In comparison, state-of-the-art techniques underestimate the required guardband by 27% and 39%, on average, for the case of IR-Aging-3 and IR-Aging-2, respectively as presented in Fig. 4.6(e). Finally, Fig. 4.6(e) shows that considering BTI recovery, caused by IR-drops, results in 18% narrower timing guardbands. *In summary, considering the interdependencies between both aging and IR-drop leads not only to estimate the correct timing guardbands but also to narrow them and reduce the performance loss.*

(2) Impact of IR-drop on mitigating aging: Fig. 4.7(a) demonstrates an example of the V_{dd} and V_{ss} traces for *conv2d* benchmark along with the resulting ΔV_{th} trace over time. As shown, IR-drops result in fluctuations in both V_{dd} and V_{ss}. The larger the IR-drop (i.e., smaller V_{win}), the larger the reductions in ΔV_{th} because more recovery for BTI defects will be possible.

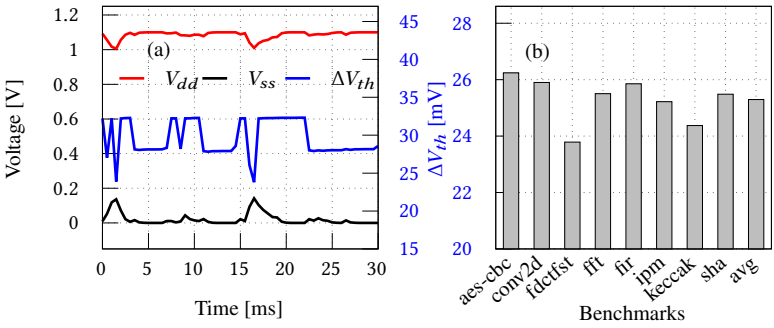


Figure 4.7.: (a) V_{dd} and V_{ss} fluctuations over time due to IR-drop, caused by "conv2d" benchmark, along with corresponding BTI-induced V_{th} trace. (b) Resulting BTI-induced ΔV_{th} due to the mitigation effect of IR-drop caused by varied benchmarks. IR-drop can lead to around 24% reduction in ΔV_{th} due to the recovery impact of voltage reductions. The baseline ΔV_{th} is 33mV, which is the impact of BTI after one year lifetime under the operation at 1.1V.

To clarify more the mitigation impact, Fig. 4.7(b) shows the resulting BTI-induced degradation (ΔV_{th}) for each benchmark. As previously mentioned, the baseline BTI degradation is ΔV_{th} of 33mV after one year lifetime under the operation at the nominal voltage of 1.1V. As shown, the average ΔV_{th} is around 25mV. Hence, BTI recovery caused by IR-drop can lead to around 24% mitigation for the induced ΔV_{th}.

4.5. Summary and Conclusions

Reliability-aware circuit design flows, in which the effects of voltage fluctuations and aging are *jointly* considered, do virtually not exist. In this chapter, a novel technique was presented that links physics all the way up to the circuit level to show the existing interdependencies between aging and voltage fluctuations. The impact of neglecting these interdependencies on timing guardband estimations has been presented under the proposed technique compared to state-of-the-art techniques. All in all, increasing the efficiency of circuits while sustaining reliability necessitates accurate estimations of the smallest, yet sufficient, timing guardbands that overcome the effects of aging as well as voltage fluctuations. *Without considering the existing interdependencies, achieving that goal would not be possible.*

5. Aging-aware Approximate Computing for NPU

As has been demonstrated previously, transistor aging profoundly degrades the reliability of circuits during their lifetime as transistors become slower, resulting in errors due to timing violations unless large guardbands are included, which leads to considerable performance losses. When it comes to Neural Processing Units (NPUs), where increasing the inference speed is the primary goal, such performance losses cannot be tolerated. In this chapter, a novel technique of reliability-aware quantization to eliminate aging effects in NPUs is presented. In NPUs, a large number of MAC units are tightly packed within a small footprint. Their inherent nature to perform massive parallelism makes MACs highly utilized, which results in excessive on-chip temperatures [8]. MAC units, in NPUs, are highly utilized with little time for relaxation, and hence excessive on-chip temperatures occur. This stimulates the aging mechanisms, which makes MACs very susceptible to aging-induced timing error. However, as demonstrated in Section 2.3, input compression results in a significant delay reduction of the MAC. Hence, applying input compression through quantization can mitigate aging-induced timing errors and eliminate the required guardband, which is presented here.

5.1. Input Compression Through Quantization:

By compressing the MAC inputs, the accuracy of the Neural network (NN) will be degraded due to the reduced bit width representation. To reduce such degradation, multiple low-bit-width quantization techniques can be employed [77, 65, 19, 94]. Particularly, the activations and the weights can be quantized to lower bit width $8 - \alpha$ and $8 - \beta$ bits, respectively, with the appropriate padding. This helps to gradually increase the compression (α and β values) over time in order to improve the delay gain when NPU ages. In this work, NNs are trained with 32-bit floating-point numbers, while the post-training quantization aims to reduce the number of bits for weights and activations (e.g., 8-bit integers). As a result, the model size is reduced while the accuracy remains close to the accuracy of the 32-bit floating-point model [77]. The reduced number of bits changes with the required demand.

A library of multiple low-bit width post-training quantization methods can be created based on state-of-the-art approaches. Since some of these approaches

are optimized for specific NNs, or optimized for low precision, multiple approaches must be considered. All these approaches do not require NN retraining and allow the utilization of different precision for weights and activations. Details regarding the employed quantization techniques in [77, 65, 19, 94].

For instance, considering 8-bit quantization, the activations and weights are quantized to the $[0, 2^8)$ segment, while the biases are quantized to $[0, 2^{16})$. Considering (α, β) compression, the activations are quantized to $[0, 2^{8-\alpha})$, weights to $[0, 2^{8-\beta})$, and biases to $[0, 2^{16-\alpha-\beta})$. When considering LSB padding, for example, the convolution inputs are shifted left, and thus, convolution operation equals:

$$\begin{aligned} F_{shifted} &= Bias \times 2^{(\alpha+\beta)} + \sum_{\forall j} \left((A_j \times 2^{\alpha}) \times (W_j \times 2^{\beta}) \right) \\ &= \left(Bias + \sum_{\forall j} \left(A_j \times W_j \right) \right) \times 2^{(\alpha+\beta)} = F \times 2^{(\alpha+\beta)} \end{aligned} \tag{5.1}$$

Hence, the output must be shifted to the right by $\alpha + \beta$ positions. However, this does not require additional circuity as it can be implemented at the software level. On the other hand, when MSB padding is used, no shift is required.

5.2. Implementation of Aging-Aware Quantization

Here, the implementation to perform the aging-aware quantization is presented. The implementation, in this work, starts from the device level up to the system level, where the NN inference accuracy is impacted.

5.2.1. Aging Modeling

Aging Model: In this work, the state-of-the-art physics-based aging model from [106] is considered, which can precisely capture the aging-induced degradation (ΔV_{th}) for the 14nm technology. The model has been validated against semiconductor measurements for various technologies.

Aging-Aware Cell Libraries: Following the approach in Section 4.3, the aging analysis within the commercial digital design tool flows is extended by creating the aging-aware cell libraries. First, the industrial transistor compact model (BSIM-CMG) model is calibrated to match Intel's 14nm FinFET technology measurements provided in [97]. Then, the state-of-the-art physics-based aging model is employed to estimate the corresponding ΔV_{th} over time, considering ten years as the typical projected lifetime. As the aging mechanism is affected by the operating conditions (e.g., utilization and temperature), ΔV_{th} is considered as an unbiased measure of the aging level. Therefore, aging is studied as a gradual increase in ΔV_{th} over time. Therefore, the examined aging period ranges from a fresh chip with $\Delta V_{th} = 0$ to 10 years lifetime with $\Delta V_{th} = 50$mV [106] with a 10mV step. Afterward, for each ΔV_{th} step, an aging-aware cell library is created by characterizing all standard cells.

Aging-induced Delay Analysis: For delay analysis at the circuit level, the same approach presented previously in Section 4.3 is used. The MAC design first is synthesized targeting maximum performance using $\Delta V_{th} = 0$ (i.e., no aging). Next, using Synopsys PrimeTime, static timing analysis (STA) is performed on the post-synthesis netlist. STA employs the aging-aware cell libraries to precisely capture the impact of aging on the circuit's delay. Notably, the worst-case analysis is considered where all transistors are equally degraded. Then, the delay is analyzed for both uncompressed and compressed inputs. For input compression, the respective input bits are padded with zeros (i.e., set to 0). This is important to precisely capture the circuit's delay w.r.t. the aging period and the paths that are activated with input compression.

5.2.2. Our Proposed Aging-Aware Quantization

As demonstrated in Section 2.3, input compression on the MAC unit delivers considerable delay gains (i.e., delay reduction), which potentially eliminating the aging-induced timing errors. Input compression in NPUs can be applied by employing low bit-width quantization in order to reduce accuracy loss. This work introduces an adaptive approximation approach by progressively increasing the required input compression over time. The full implementation of the proposed technique is described in Algorithm 5.1. First, the RTL description of the circuit is synthesized to obtain the post-synthesis netlist for the fresh circuit without aging. The MAC unit is considered as the driving circuit, as its accuracy and delay define the accuracy and speed of the NPU [65].

Algorithm 5.1 Aging-Aware Quantization [130]

Input: Synthesized Netlist, Aging Level (ΔV_{th}), Trained Model & Test Dataset
Output: Aging-Aware Quantized Model

1: List = []
2: **for all** $(\alpha, \beta) \in [0, 8]^2$:
3: Run STA with (corresponding aging library, compression (α, β))
4: **if** timing constraint is met**:** List $\leftarrow$ add (α, β)
5: $(\alpha, \beta) \leftarrow (\alpha, \beta)$ in List with $\min\left(\sqrt{\alpha^2 + \beta^2}\right)$
6: **for all** method **in** Quantization Library
7: Quantize Model using method and size (8-a, 8-b)
8: Capture accuracy on test dataset
9: **return** Quantized Model

Next, the signoff tool, along with the aging-aware libraries, is used to perform static timing analyses to identify all combinations of the compression values (α, β) that satisfy the timing constraint of the MAC unit (lines 2-4). This process is done for both MSB and LSB padding. Targeting minimum compression, the (α, β) that minimizes $\sqrt{\alpha^2 + \beta^2}$ is then selected (line 5). Finally, the obtained compression value (α, β) is used to quantize the NN model. The quantization size equals $8 - \alpha$ for the activations, $8 - \beta$ for the weights, and $16 - \alpha - \beta$ for the biases. Afterward, for the quantization procedure, all of the available methods in the created library are considered (see Section 5.1), and the inference accuracy on the test dataset by using the quantized model is examined (lines 6-8).

The (α, β) values are extracted at design time using the timing analysis to ensure that the timing constraint is met. Hence, no aging-induced timing errors occur, and accurate computations are always performed on the compressed inputs. Therefore, the inference accuracy is defined only by the accuracy delivered by quantization for the respective compression values. The inference accuracy can be captured at the software level without the need to perform post-synthesis timing simulations, which is infeasible for large datasets [158]. Note that the α and β values depend on the NPU microarchitecture (e.g., MAC size) and the aging period. Notably, the selected quantization method depends on both the (α, β) and NN.

In line 5 of Algorithm 5.1, by minimizing the employed compression, the algorithm selects the (α, β) that minimizes $\sqrt{\alpha^2 + \beta^2}$ using the Euclidean distance to the correct computation (i.e., no compression) case. For each quantization method and each NN, we quantize the NN using the respective method and (α, β) compression to capture the accuracy loss before compression. This procedure is repeated $\forall(\alpha, \beta) \in [0, 4]^2$, then ($\alpha$, β) are ranked based on

Table 5.1.: Achieved accuracy and selected quantization method for varying NNs at various aging levels (represented by ΔV_{th}) [130].

	Accuracy Loss (%) / Quantization Method Selected				
Neural Network	**10mV**	**20mV**	**30mV**	**40mV**	**50mV**
ResNet50	0.27 / M5*	0.36 / M5	0.97 / M3*	1.47 / M4*	2.37 / M4
ResNet101	0.26 / M5	0.36 / M5	1.28 / M5	0.97 / M4	1.84 / M4
ResNet152	0.28 / M5	0.34 / M5	1.08 / M5	1.12 / M4	2.10 / M4
VGG13	0.15 / M4	0.22 / M3	0.39 / M3	1.20 / M4	2.54 / M4
VGG16	0.05 / M5	0.14 / M5	0.29 / M3	0.73 / M4	1.09 / M4
VGG19	0.20 / M3	0.33 / M3	0.46 / M3	1.09 / M4	2.37 / M4
Alexnet	0.28 / M5	0.54 / M5	0.99 / M5	2.72 / M4	4.00 / M4
SqueezeNet 1.1	0.55 / M5	1.51 / M5	3.61 / M4	6.03 / M4	7.83 / M4
Wide ResNet50	0.14 / M5	0.24 / M5	0.67 / M5	1.27 / M4	2.49 / M4
Wide ResNet101	0.23 / M5	0.41 / M5	1.33 / M5	1.41 / M4	2.92 / M4

* M3: LAPQ [94], M4: ACIQ [19], M5: ACIQ w/o bias correction [19]

the computed accuracy loss. This process is done offline in order to avoid any extra runtime overhead.

5.3. Evaluation and Analysis

In order to evaluate the effectiveness of the proposed technique in eliminating the aging-induced timing errors in NPUs, the delay gain delivered by the technique is examined as well as the respective accuracy loss that has to be traded due to the applied input compression. For evaluation purposes, an architecture similar to the Edge TPU microarchitecture [32] is considered using MAC with an 8-bit multiplier and a 22-bit adder as the driving circuit. Ten NNs (listed in Table 5.1) with different characteristics are examined. For aging analysis, using the aging libraries, the delay is estimated at different aging levels for all possible input compression options. All the NNs are trained on the ImageNet dataset [44] and their implementation is based on official PyTorch repositories (Torchvision) [112]. Please note, in this work, the baseline design refers to the design where no compression is applied (i.e., α=β=0) as the MAC unit uses 8-bit quantization for activations and weights

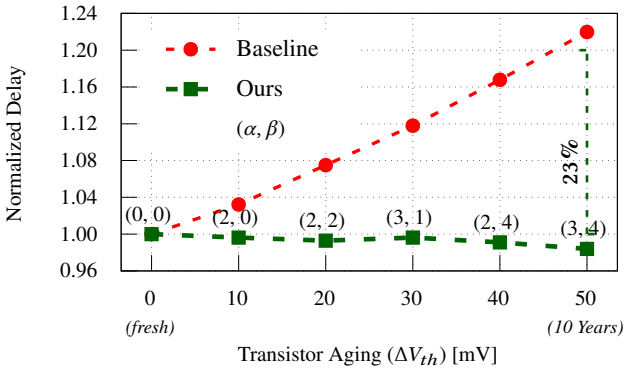


Figure 5.1.: The normalized delay, from the beginning until the end of the projected lifetime (10 years), of the baseline and our approach. The delay is normalized to the fresh baseline. In addition, figure shows the corresponding input compression (α, β) where zero padding is applied.

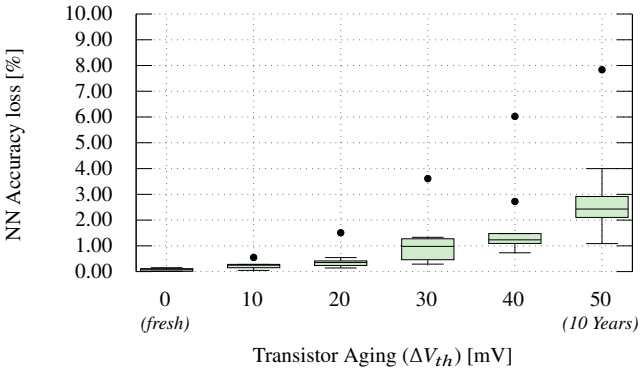


Figure 5.2.: Graceful accuracy degradation is delivered by the aging-aware quantization over time. Average accuracy degradation is presented by box plots for ten of the examined NNs.

[65]. Moreover, the accuracy loss is calculated with respect to the accuracy achieved with the original implementation (i.e., 32bit floating point inference). Therefore, even the baseline with no-aging will exhibit an insignificant, yet negligible, accuracy loss.

Fig. 5.1 shows the delay of the MAC employing the aging-aware quantization technique under input comparison as well as the delay of the baseline for the whole projected lifetime. The figure also shows the applied zero-padding bits

(α, β). The delay is normalized to the delay of the fresh baseline without timing guardbands. As shown, the baseline delay increases over time, resulting in a performance loss of 23% over the ten years of aging. This is expected behavior due to the aging phenomena where transistors become slower. Therefore, a timing guardband must be considered to protect the circuit against aging-induced timing errors for reliable operation. Contrarily, this is not the case when applying our technique. As shown, the proposed technique does not follow the baseline degradation trend because it adaptively compresses the inputs over time to gracefully compensate for the delay increase due to aging. The normalized delay is always less than or equal to 1, making the aging-aware quantization technique resilient against aging-induced delay degradation. Therefore, the technique effectively suppresses the aging effects, and no timing violations will occur while no guardband is required. Importantly, by eliminating the timing guardband, a 23% performance recovery is achieved compared to the baseline.

Nevertheless, the technique results in accuracy degradation due to lower width representation for the weights and activations. Fig. 5.2 shows the average accuracy degradation over aging for the examined NNs. Considering the worst-case delay analysis, the same compression is used for all the NNs. Fig. 5.2 shows the accuracy loss, presented by box plots over the examined NNs and aging periods. As shown, the technique delivers graceful accuracy degradation over time. For instance, the average accuracy loss is 0.24%, 0.45%, 1.11%, 1.80%, and 2.96% for aging (ΔV_{th}) 10mV, 20mV, 30mV, 40mV, and 50mV, respectively. The full accuracy results are summarized in Table 5.1 as well as the quantization method selected by the technique for each case.

Over time, the technique adaptively compresses the MAC inputs, which in turn reduces the switching activity, and thus, lowers power consumption. Fig. 5.3 shows the normalized energy of the proposed technique in comparison to the baseline. Please note again, the MAC unit operates at the maximum possible frequency, while the baseline requires a 23% timing guardband. The proposed technique delivers 46% energy reduction on average for the whole lifetime, ranging from 21% up to 67%.

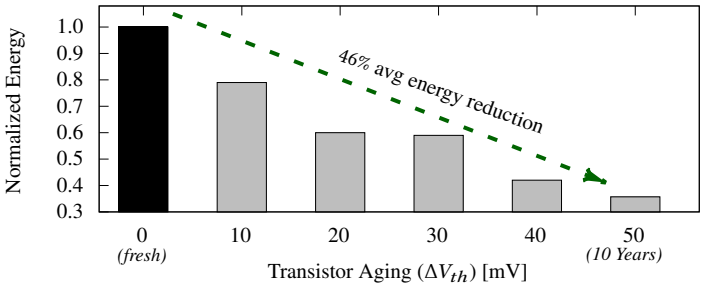


Figure 5.3.: Normalized energy consumption of the aging-aware quantization technique over the baseline for varying aging levels. Aging timing guardband is used for the baseline to prevent timing errors.

5.4. Summary and Conclusion

This chapter presents an adaptive aging-aware quantization technique over time to suppress aging impacts in NPUs. Such technique results in insignificant, yet negligible, accuracy degradation while eliminating the aging-induced timing violations without the need for aging timing guardbands. This results in boosting the NPU performance. As presented, the proposed technique significantly decreases energy consumption with no overhead.

6. Zero-Temperature Coefficient to contain Timing Guardband under Self-heating Effects

Self-Heating Effects (SHE) is a fundamental obstacle for the current FinFET devices and future transistor structures. It results in excessive temperatures (T_C) across the transistor's channel, which severely degrades the switching speed and increases the leakage power of processors. To sustain reliability and prevent timing violations, large timing guardbands are necessary, which lead to a considerable performance loss of the processor (see Section 2.4). At nominal supply voltages, temperature increases result in delay increases. However, at lower voltages, the dependence reverses.

The Zero-Temperature Coefficient (ZTC) operating point is well-suited to minimize SHE impacts on the delay of large circuits (e.g., processor). By definition, it is a point (or region) where the temperature has little impact on the circuit's delay. Consequently, ZTC can be employed to minimize the impact of the unavoidable SHE (details are Section 2.4). This removes the need for large guardbands and their negative impact on the performance of circuits at the cost of operating at a lower V_{dd}, which comes with its performance loss. Therefore, a trade-off has to be found between thermal timing guardband and performance loss.

To study the impact of SHE on large circuits using the standard EDA tools for chip design, these tools must become aware of SHE. Hence, the standard EDA tools must be enhanced first by bringing SHE analysis to the circuit level.

6.1. Self Heating modeling

Since SHE originally is analyzed at the transistor level, the analysis must start there. Transistor SPICE simulations are performed to determine $\Delta T_C(SHE)$ under different conditions (e.g., different V_{dd}, switching frequencies, number of fins, etc.). Later, SHE-aware standard cell libraries can be characterized, which necessitates analyzing standard cells under SHE by performing SPICE simulation for each cell under $\Delta T_C(SHE)$ to extract the cell's delay and power. Cell libraries are crucial for digital circuit design using EDA tools.

Modeling Self-Heating Effects: To model SHE, the transistor is abstracted using the duality between thermal and electrical modeling. SHE can be represented with an equivalent RC-thermal network, as shown in Fig. 6.1 [126]. The thermal resistance R_{th} represents the resistance to thermal flux between the channel and the rest of the chip. The thermal capacitance C_{th} is the amount

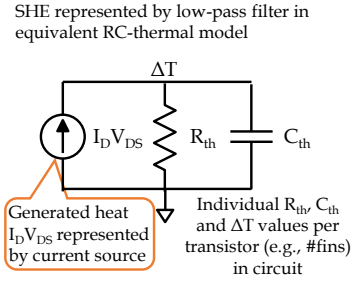


Figure 6.1.: Schematic diagram of a SHE model represented by a low-pass filter in an equivalent RC-thermal model [126].

of heat that is stored within the channel material. Both R_{th} and C_{th} depend on the semi-conductor technology (i.e., geometry, and materials of a transistor) and they are independent of voltage [76]. The generated heat inside a channel is the power loss due to joule heating within the channel, which is given by the product of the drain-current I_D and drain-source voltage V_{DS}.

This model is typically employed in SPICE during circuit simulations, as can be found in the industry-standard FinFET compact model BSIM-CMG [26]. With this model, $\Delta T_C(SHE)$ can be estimated by solving the voltage at node T (T_C). The temporal behavior of SHE is given by the time constant $\tau_{th} = C_{th} \cdot R_{th}$. A large time constant (e.g., τ_{th}=100ns) results in slow heating/cooling of the channel, while fast time constants result in rapid temperature changes. Currently, typical time constants are approximately 1ns [81].

Transistor SHE Simulations: To model the electrical characteristics of pFinFET and nFinFET transistors, the modelcard from the ASAP7 PDK is employed [39] inside the FinFET compact model BSIM-CMG[26]. Simulations for pFinFET and nFinFET are performed under a range of voltages for different numbers of fins. With increasing voltage, T_C rises and similarly with an increasing number of fins [68]. Then, BSIM-CMG is calibrated with 7nm FinFET SHE parameters from [68]. The simulation of a single transistor using typical operation conditions (i.e., 25°C V_{dd}=0.7V) with 3 fins shows $\Delta T_C(SHE)\approx$150°C. Multiple fins heat the substrate and thus each other. For instance, $\Delta T_C(SHE)$ significantly increased when the number of fins changed to 7. Consequently, increasing the number of fins to 7 results in a higher temperature of 350°C, as shown in Fig. 6.2. Such a high T_C occurs under the worst-case corner (i.e., continuous heating due to DC currents, high fin counts, and high voltage). Worst case means, in this context, the slowest delay always.

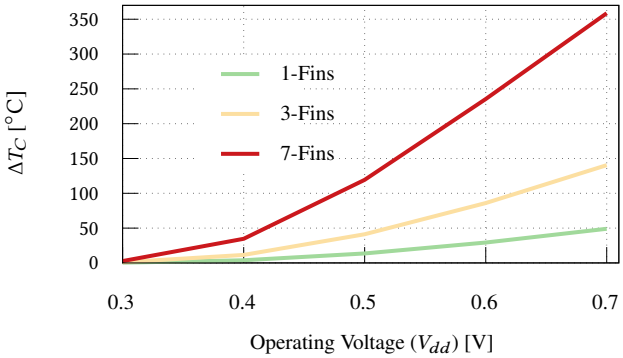


Figure 6.2.: Temperature increase within the transistor's channel (ΔT_C) due to SHE over a wide range of operating voltage V_{dd} [126]. Results are generated by employing the SHE model in Fig. 6.1 for transistors with 1,3 and 7-Fins at room temperature of 25°C.

However, in actual operation, the transistor, within digital circuits, will not observe such operating conditions and thus experience lower T_C. However, designers need always to account for the worst case [98, 156, 16]. Importantly, Fig. 6.2 shows how $\Delta T_C(SHE)$ decreases with V_{dd} decreases as it reaches ≈50°C and ≈120°C at 0.5V for 3 and 7 fins, respectively. Note that T_C only occurs within the transistor's channel while the chip temperature T_{chip} remains relatively cool and silicon copes with these temperatures [119].

6.2. Minimizing Thermal Dependence via ZTC operation in Large Circuits

This section shows the key challenge behind finding a single ZTC for large circuits exceeding 100K transistors. It illustrates the proposed technique in finding the point near ZTC with *minuscule* temperature-induced variance. Afterward, it features the implementation phases by creating cell libraries and the physical design. The physical design (i.e., chip layout) allows designers to accurately investigate SHE impacts on the chip *before fabrication* in terms of delay (i.e., performance) and then employing N-ZTC. The flow diagram of the proposed technique of finding N-ZTC of a chip is shown in Fig. 6.3.

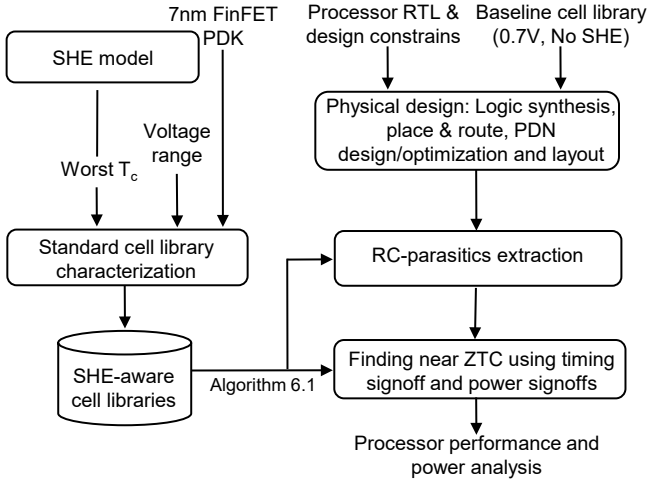


Figure 6.3.: The proposed technique, from the transistor model all the way to chip layout, aiming at mitigating SHE-induced delay degradation by employing N-ZTC. This necessitates finding the N-ZTC point of the chip following Algorithm 6.1 [126].

6.2.1. Finding the ZTC of standard cells

V_{ZTC}, in this context, refers to the operating voltage (V_{dd}) where ZTC is observed. Obtaining ZTC voltage for a large circuit, such as a processor, while considering SHE is challenging. A microprocessor features thousands of subcircuits. Each subcircuit contains many connected standard cells with a unique V_{ZTC} per cell type [86]. This is due to the different transistor types (e.g., more pFinFET than nFinFET) where each transistor type has its unique V_{ZTC}[70], different topology (e.g., transistors in series, transistors in parallel), and different transistor configurations (number of fins) per cell. Moreover, considering the various operating conditions of each cell creates a non-negligible variance in V_{ZTC}. To consider the impact of the operating conditions into account, 7 input signal slews (t_{slew}) along with 7 output load capacitances (C_{load}) are considered. These are typical values for industrial and academic cell library characterization[151]. Consequently, cell topology, t_{slew} and C_{load} result in various V_{ZTC} for different standard cells. The 7×7 propagation delay matrix for each standard cell is arranged as follow:

$$7 \times 7 = \begin{bmatrix} (t_{slew_1}, C_{load_1}) & \dots & (t_{slew_1}, C_{load_7}) \\ \vdots & \ddots & \vdots \\ (t_{slew_7}, C_{load_1}) & \dots & (t_{slew_7}, C_{load_7}) \end{bmatrix}$$

For instance, the 7×7 of V_{ZTC} matrix of NANDx2 (nand gate) cell experiments for the average rise delay shows various V_{ZTC} under SHE as follow:

$$cell = \begin{bmatrix} V_{ZTC_{(1,1)}} & \dots & V_{ZTC_{(1,7)}} \\ \vdots & \ddots & \vdots \\ V_{ZTC_{(7,1)}} & \dots & V_{ZTC_{(7,7)}} \end{bmatrix}$$

$$\text{NANDx2} = \begin{bmatrix} 0.53 & 0.53 & 0.52 & 0.52 & 0.51 & 0.50 & 0.49 \\ 0.53 & 0.53 & 0.53 & 0.52 & 0.51 & 0.50 & 0.49 \\ 0.53 & 0.53 & 0.53 & 0.53 & 0.51 & 0.51 & 0.50 \\ 0.54 & 0.54 & 0.53 & 0.53 & 0.53 & 0.51 & 0.50 \\ 0.54 & 0.54 & 0.54 & 0.53 & 0.53 & 0.53 & 0.53 \\ 0.54 & 0.54 & 0.54 & 0.54 & 0.53 & 0.53 & 0.53 \\ 0.55 & 0.54 & 0.54 & 0.54 & 0.54 & 0.54 & 0.53 \end{bmatrix}$$

NANDx2 exhibits V_{ZTC} that ranges between $0.55-0.49$V with a majority of ZTC at 0.53V. Still, there is a clear trend indicating a dependency on both t_{slew} and C_{load}. Selecting the operating condition is the responsibility of the synthesis tool that is unaware of SHE.

On the other hand, for example, INVx2 (Inverter) cell exhibits a different V_{ZTC} matrix than NANDx2, illustrating how each cell is different despite identical (t_{slew}, C_{load}) conditions. V_{ZTC} of the INVx2 cell exhibits significantly less variance with a smaller range of ranges between $0.54-0.50$V, with a stronger majority at 0.53V. This further highlights that variance in ZTC is circuit topology dependent.

$$\text{INVx2} = \begin{bmatrix} 0.53 & 0.53 & 0.53 & 0.53 & 0.52 & 0.52 & 0.50 \\ 0.53 & 0.53 & 0.53 & 0.53 & 0.53 & 0.52 & 0.50 \\ 0.53 & 0.53 & 0.53 & 0.53 & 0.53 & 0.52 & 0.51 \\ 0.53 & 0.53 & 0.53 & 0.53 & 0.53 & 0.53 & 0.51 \\ 0.54 & 0.53 & 0.53 & 0.53 & 0.53 & 0.53 & 0.53 \\ 0.54 & 0.54 & 0.54 & 0.53 & 0.53 & 0.53 & 0.53 \\ 0.54 & 0.54 & 0.54 & 0.54 & 0.53 & 0.53 & 0.53 \end{bmatrix}$$

Fig. 6.4 shows the histogram of all V_{ZTC} of all cells, which highlights the variances in all V_{ZTC}. Experiments cover all operating conditions for all cells, i.e., 101 standard cells $\times 7$ $t_{slew} \times 7$ C_{load} = 4949 simulation. The figure shows that the highest percentage of ZTC occurrence is at 0.54V, yet the span is quite large, from 0.49V to 0.55V. With such variance in V_{ZTC} within and across cells, it is *impossible* to operate every cell within the circuit *exactly* at ZTC.

However, a circuit consists of different subcircuits (i.e., cells) with different V_{ZTC}. Therefore, since each cell has a different matrix, finding the overall V_{ZTC} of the circuit is challenging, as it is must be the weighted average of V_{ZTC} of its subcircuits. To distinguish V_{ZTC} from the cells and the entire chip, we will refer to $V_{ZTC}(cell)$ and $V_{ZTC}(chip)$ from now on. $V_{ZTC}(chip)$ for the entire circuit is, thus, the weighted superposition of millions of $V_{ZTC}(cell)$ from all cell instances within the circuit. This would not entirely eliminate the thermal variance but will reduce it. However, the remaining variance is minuscule (i.e., negligible), as the circuit will operate close to the ZTC for most cells, as will be demonstrated later in Section 6.3.

Note, due to process variations; each transistor might have different characteristics. This results in a variation of the ZTC of transistors. The performed analysis shows that the variation of $V_{ZTC}(transistor)$ is small, and $V_{ZTC}(cell)$ is within the $V_{ZTC}(transistor)$ range (details in Appendix A.4.1).

6.2.2. ZTC for Large Circuits

Finding ZTC voltage of a large circuit is challenging due to the different $V_{ZTC}(cell)$. Cells within the circuit should be examined for delay under a set of conditions. With four dimensions t_{slew}, C_{load}, T_C and V_{dd} checking all conditions is unfeasible due to simulation time. Therefore, the proposed technique relies on static timing analysis tools (STA) to find and then employ

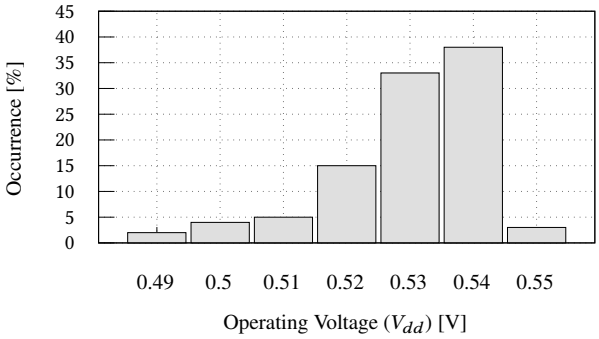


Figure 6.4.: The histogram of the experimental results of all V_{ZTC} of all cells. Experiments cover all operating conditions of all cells extracted by simulating every standard cell at high T_C (with SHE) and low T_C (without SHE) at a wide range of voltages.

$V_{ZTC}(chip)$. Consequently, circuits will operate at V_{dd} near ZTC (N-ZTC) of cells following Algorithm 6.1. The algorithm examines the circuit's delays at different T_Cs for a wide range of V_{dd}. The algorithm finds $V_{ZTC}(chip)$ by comparing the circuit's delays of a range of T_C, until a match is found (or within an acceptable delay variance ϵ). The full technique for employing N-ZTC of a circuit is summarized in Fig. 6.3 and Algorithm 6.1.

First, the circuit's layout is designed after synthesizing the RTL of the circuit. With the layout available, signoff tool creates best and worst-case corners for every voltage step based on the given $T_C(low)$ and $T_C(high)$ (i.e., the highest and lowest T_C). T_C follows the simulation results reported in Fig. 6.2, where $T_C = T_{chip} + \Delta T_C(\text{SHE})$. Note again that the worst case is always the highest delay, not the highest temperature. The signoff tool then examines the circuit's delay t_{delay} at these T_Cs. By applying the worst-case approach and as the actual T_C is within the examined range, this guarantees the error-free operation of the circuit. As V_{ZTC} is unknown in prior, the algorithm has to traverse all voltages within a suitable range (e.g., from $V_{ZTC}(pFinFET)$ to $V_{ZTC}(nFinFET)$) with the smallest possible step ($V_{step} = \alpha$). Iteratively, V_{dd} is reduced by a small step α=0.01V[1]. For each voltage, the analysis estimates at both high T_C ($T_C = T_{chip} + \Delta T_C(\text{SHE})$, see Fig. 6.2) and low T_C (without SHE, $T_C = T_{chip}$). After both simulations are complete, the algorithm examines the

Algorithm 6.1 Operating circuits Near their Zero-Temperature Coefficient (N-ZTC) aiming minuscule SHE-induced delay variance [126]

Require: *Voltage range, Voltage step α, channel ΔT_c list, SHE-aware libraries, chip layout, acceptable delay variance ϵ*
Ensure: N-ZTC at V_{ZTC}
1: Set V_{dd}=$V_{Nominal}$ ▷ Start from nominal=0.7V
2: **while** ZTC not found **do**
3: **for** Each ΔT_c in the list at V_{dd} (Fig. 6.2) **do**
4: $T_C = T_{chip} + \Delta T_C$ ▷ T_C(SHE)
5: **Create** Process corner at V_{dd} ▷ Using Voltus
6: **Set condition** set Temperature = T_c
7: **Parasitics extraction** ▷ Using Voltus
8: **STA** Chip's delay analysis ▷ Using Tempus
9: **Report** Delay $t_{delay}(T_c)$ ▷ Using Tempus
10: **end for**
11: $\Delta t_{delay} = t_{delay}(T_c(high)) - t_{delay}(T_c(low))$
12: **if** $\Delta t_{delay} \leq \epsilon$ **then** ▷ acceptable delay variance ϵ
13: ZTC found is True
14: **end if**
15: **Update** V_{dd}=$V_{dd}-\alpha$ ▷ update voltage
16: **Update** T_c at V_{dd} ▷ update T_c list
17: **end while**
18: Report Power ▷ Using Voltus at ZTC point for all temperatures

t_{delay} at every V_{dd} for both worst and best corners to find N-ZTC. The accepted delay variance is $\epsilon \leqslant 0.01$ ns (1% of total $t_{delay}(CP) \approx 1$ns).

6.2.3. SHE-Aware Standard Cell Libraries

Multi-Corner Multi-Mode (MCMM) are multiple executions of static timing analysis that are used in the design of digital chips across all modes and corners concurrently. A process corner is a cell library (delay and power tables for each cell) characterized by voltage, temperature, and manufacturing tolerance. Typically, a Process Design Kit (PDK) offers three corners, which are Slow-Slow (SS), Typical-Typical (TT), and Fast-Fast (FF) corners. The EDA tools can solely examine the circuit's delay and power at or between the available corners. For voltages or temperatures beyond the process corners (e.g., higher

[1] Smaller would not make sense as on-chip voltage regulators operate in 10mV intervals [74][71].

than highest temperature, lower than lowest voltage), the tools are unable to do any delay/power analysis since these tools do not support extrapolation. Importantly, available corners do not consider SHE. Hence, it is necessary to extend the available corners by creating *SHE-aware cell libraries*. In addition to higher temperatures, these cell libraries span a wide range of voltages to ensure ZTC is within the design space.

For this purpose, SHE-aware standard cell libraries are characterized by employing the SPICE netlists of combinational and sequential cells from the 7nm ASAP7 PDK[39]. The SHE-aware cell libraries are characterized considering the temperature used in the propagation delay simulations to the corresponding T_C under SHE. By using three fin configurations: 1, 3, and 7 fins, as shown in Fig. 6.2, this covers more than 90% of all transistors within the ASAP7 PDK. However, considering the worst-case operating condition (i.e., slowest delay), the highest fin count should be considered. Therefore, the peak T_C of the 7-fin transistor is used as the temperature during characterization. This temperature is then entered in the library characterization tool to determine the power and delay of the standard cells under various t_{slew} and C_{load} with the corresponding T_C=T_{chip}+ΔT_C(SHE) for a set of voltages V_{dd}. All the measured delays of every cell are then stored within a lookup table using the standard *liberty* format. In addition, for comparison purposes, the entire process is also done without SHE (T_C=T_{chip}).

6.3. Evaluation and comparison

This section evaluates the effectiveness of the proposed technique following Fig. 6.3. First, the physical chip design of the processor is presented. Then, $V_{ZTC}(cell)$ variance within the chip is demonstrated. Afterward, the technique is applied to determine N-ZTC of the chip ($V_{ZTC}(chip)$). Finally, the effectiveness of operating the designed chip at N-ZTC is presented in comparison with the traditional guardbanding technique w.r.t. performance and power.

6.3.1. Physical Chip Design

Large chip designs likely feature higher T_C variance, due to more combinations of f_{sw} (switching frequency), t_{slew}, C_{load} for a wider variety of standard cells.

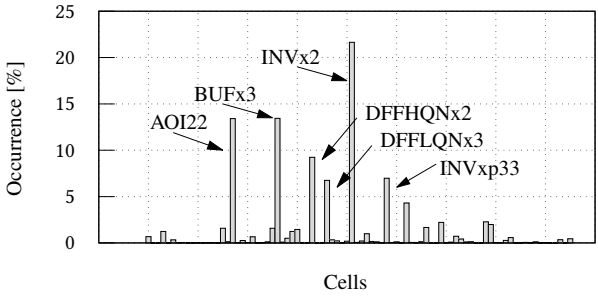


Figure 6.5.: The histogram of the employed cells within the OpenPiton chip layout as a percentage of occurrences of each cell to the total number of cells.

Therefore, a large circuit, such as a full processor, is considered to maximize T_C variance. Thus, a full computing tile of the state-of-the-art OpenPiton processor is employed. OpenPiton is an open-source processor based on the OpenSPARC T1 core [18]. A full tile consists of a CPU core, a floating-point unit (FPU), caches, and network-on-chip (NoC).

First, the RTL of the processor is synthesized using the baseline cell library from ASAP7 PDK [39], at nominal voltage 0.7V without SHE, using Synopsys DC compiler synthesis tool[152]. Then, the design is passed through place and route (i.e., layout design), including Power Delivery Network (PDN) design and optimization, using Cadence Innovus 7.1[29]. Then, N-ZTC is determined based on the post-layout simulations considering RC-parasitics and interconnects of the OpenPiton chip using the on-chip variation feature considering their impacts on delay and power.

6.3.2. ZTC Variance within The Processor

The designed chip consists of 448,668 different cells. The synthesis tool used 86 out of the available 101 standard cells in the PDK. Fig. 6.5 shows the histogram of the instantiated cells within the OpenPiton chip. For instance, selecting V_{dd}=0.54V, as the major occurring V_{ZTC} from Fig. 6.4 makes lots of cells operate exactly at their $V_{ZTC}(cell)$, some cells are in ITD and the remaining in PTD (see Section 2.4). Therefore, operating at $V_{ZTC}(chip)$ is a compromisation among all $V_{ZTC}(cell)$ as cells are distributed over all three thermal regions (see Section 2.4) while the majority in ZTC.

The variations in all V_{ZTC} can be examined using the standard deviation σ. Therefore σ of V_{ZTC}, defined in Eq. (6.1), is estimated for every operating condition (e.g., $V_{ZTC_{(1,1)}}$) across all the instantiated cells within the OpenPiton.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(V_{ZTCi} - \overline{V_{ZTC}})^2} \tag{6.1}$$

Where σ is the standard deviation, N is number of operating conditions, and $\overline{V_{ZTC}}$ is the arithmetic mean across all V_{ZTC} under the same t_{slew} and C_{load}.

$$\sigma = \begin{bmatrix} \sigma_{(1,1)} & \cdots & \sigma_{(1,7)} \\ \vdots & \ddots & \vdots \\ \sigma_{(7,1)} & \cdots & \sigma_{(7,7)} \end{bmatrix}$$

σ=0 indicates that only a single V_{ZTC}=$\overline{V_{ZTC}}$ does exist across the cells, and hence, all cells have identical V_{ZTC} under given t_{slew} and C_{load}. Oppositely, σ>0 indicates different V_{ZTC} from the mean $\overline{V_{ZTC}}$. Spanning the operating conditions reveals that the majority of cells operate at V_{ZTC} of $\overline{V_{ZTC}}\approx$0.53V contrary to the most occurring voltage of 0.54V with 38% occurrence (see Fig. 6.4). However, 0.53V is the second-most occurring voltage with 33% occurrence. This small difference results from the selection of cells and their surroundings stemming from the synthesis tool. Results of σ are summarized in the following matrix:

$$\sigma = \begin{bmatrix} 0 & 0 & 0.007 & 0.018 & 0.021 & 0.027 & 0.033 \\ 0 & 0 & 0 & 0.005 & 0.01 & 0.016 & 0.027 \\ 0 & 0.04 & 0 & 0 & 0.06 & 0.013 & 0.025 \\ 0.18 & 0.15 & 0.1 & 0 & 0 & 0.007 & 0.013 \\ 0.21 & 0.17 & 0.13 & 0.09 & 0 & 0 & 0 \\ 0.24 & 0.2 & 0.18 & 0.1 & 0.01 & 0 & 0 \\ 0.31 & 0.27 & 0.23 & 0.19 & 0.14 & 0.06 & 0 \end{bmatrix}$$

This highlights how under the same t_{slew} and c_{load}, different cells exhibit different V_{ZTC}. Hence, it is impossible to operate each cell *exactly* at its V_{ZTC}.

Hence, a compromise must be found. Instead of finding V_{ZTC} for every cell to find $V_{ZTC}(chip)$, $V_{ZTC}(chip)$ is to be estimated as discussed in Section 6.2.2.

6.3.3. Determining ZTC of The Processor

Algorithm 6.1 is employed to determine $V_{ZTC}(chip)$. For each iteration, V_{dd} is reduced by the smallest possible step α (0.01V), and then t_{delay} of the chip is examined with SHE ($T_C(high)$) and without SHE ($T_C(low)$) using Signoff tools based on the SHE-aware cell libraries. The chip's delay results (t_{delay}) of low and high T_C over the examined range of voltages converge towards $V_{ZTC}(chip)$. Since the examined range of voltages is large enough, t_{delay} must cross from the ITD region to the PTD region passing through ZTC. Hence, Algorithm 6.1 is deterministic, and it always terminates with $V_{ZTC}(chip)$.

Importantly, lowering the operating voltage reduces $\Delta T_C(SHE)$. Therefore, the proposed technique does not solely gain performance due to narrowing timing guardband, but also it *lowers* T_C. This is important because T_C stimulates other reliability phenomena like aging effects [4], and therefore, lowering T_C lowers aging impacts (i.e., reducing the guardband to protect against aging) [163, 134]. Therefore, delay and power estimations are based on the *variable* temperature with voltage changes.

Fig. 6.6 shows the delay (t_{delay}) of the OpenPiton chip with SHE ($T_c(high)$) and without SHE ($T_c(low)$) over V_{dd} where the thermal regions can be clearly identified (PTD, ZTC, and ITD). The delay is normalized to the nominal operating condition (V_{dd}=0.7V, 25°C). Guardband estimation follows the worst-case delay as shown with the gray curve in the same figure. Therefore, t_{clk} is always max($t_{delay}(T_C(low)), t_{delay}(T_C(high))$). The delay of both curves is expected to increase after a certain point, as shown on the dashed lines. This happens when $T_C(low)$=T_{chip}=$T_C(high)$ as at low voltages $\Delta T_C(SHE)$ tends to zero or when the thermal dependence of the chip's delay becomes weaker than the voltage dependence. However, as shown, $V_{ZTC}(chip)$ occurs near 0.53V (i.e., the cross point of the two delay curves). This $V_{ZTC}(chip)$ is closer to nominal V_{dd} than previously reported [41, 86] due to the smaller technology. This makes operating at ZTC more feasible, as the induced performance degradation, due to voltage lowering, is smaller if ΔV=$V_{nominal}$−V_{ZTC} is small. However, $V_{ZTC}(chip)$ is chip-specific and will vary from one chip to another.

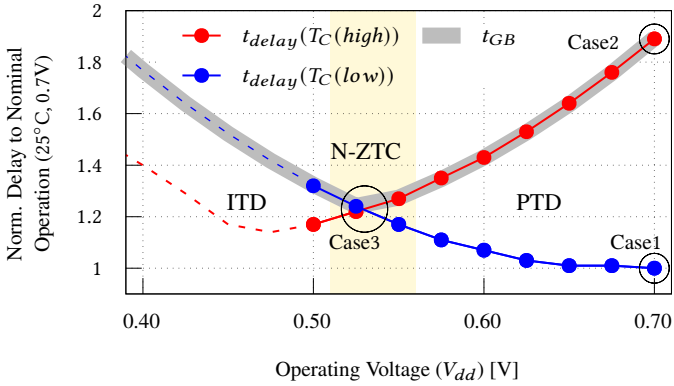


Figure 6.6.: The OpenPiton processor delay changes with T_C, due to SHE, normalized to the nominal operating condition (25°C, 0.7V). The delay of T_C(high) starts to decrease with voltage decreases considering the corresponding T_C (i.e., T_C decreases with voltage) with every voltage step. The delay of T_C(low) increases with voltage. Both delays are matched near V_{dd}≈0.53V. The delay of T_C(high) is expected to increase after dependencies changed as predicted in the dashed lines. Guardbands are estimated always at the worst-case delay, regardless if it occurs at high or low T_C. Hence, t_{clk} is always max($t_{delay}(T_C(high))$, $t_{delay}(T_C(high))$). The possible operating point cases are circled; Case-A: nominal without SHE, Case-B: traditional guardband, and Case-C: N-ZTC.

6.3.4. Traditional GuardBands for SHE Mitigation

The required guardband to mitigate SHE-induced delay degradation in the designed chip is illustrated in Fig. 6.7. SHE-induced delay degradation at nominal V_{dd} is very large where t_{GB}>90%·$t_{delay}(CP)$ and hence the circuit operates at a much higher delay (i.e., t_{clk} = $t_{delay}(CP)$+t_{GB}). The guardband t_{GB} reduces when V_{dd} reduces (starting in PTD with high V_{dd}) until reaching ZTC. Nevertheless, reducing voltage below $V_{ZTC}(chip)$ within the ITD region increases the worst delay *(now low T_C instead of high T_C)* again.

Operating the chip at $V_{ZTC}(chip)$ is a compromisation. The delay of the chip, in Fig. 6.6, is determined by the variances in the critical timing paths (i.e., slowest paths). The final delays of the critical paths still experience minuscule thermally induced delay variance (i.e., not totally eliminated but significantly reduced). By investigating such variances, the investigation shows a delay variance of <0.1% in the critical and near-critical timing paths. This is due to the tolerance factor ϵ (i.e., acceptable error) that Algorithm 6.1 employs due to having 10mV voltage steps as the algorithm might miss the perfect

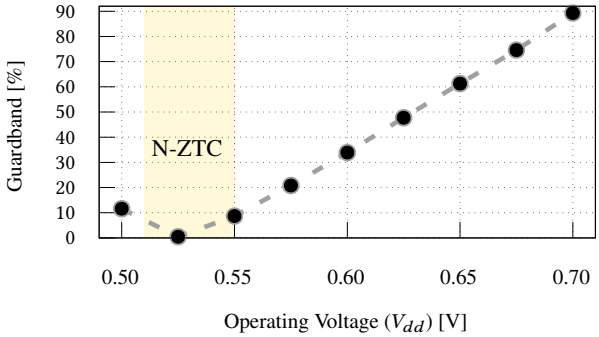
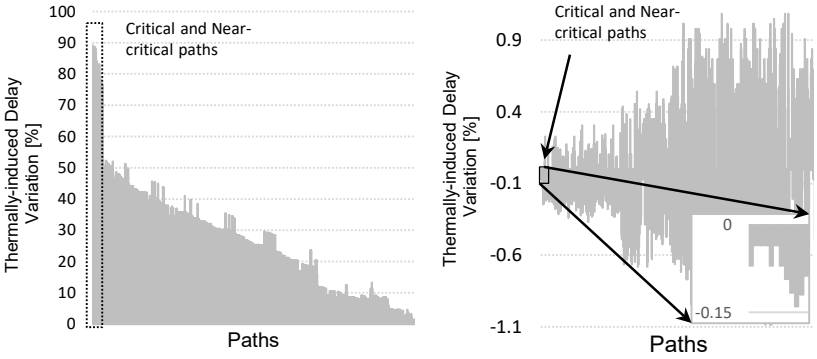


Figure 6.7.: Guardbands required to mitigate SHE-induced delay degradation over voltage in the OpenPiton chip. Guardband estimation follows the worst delay in Fig. 6.6.

V_{ZTC}. With this small delay variance, the required guardband is also small where t_{GB}<0.02ns, and hence *near-zero guardband* is required. However, the algorithm optimizes the critical path while non-critical paths exhibit larger thermally induced delay variance. Non-critical paths are, by definition, not critical (i.e., do not determine the timing of the entire chip). This is by design, where Algorithm 6.1 uses timing analysis of the entire chip to determine $V_{ZTC}(chip)$. The proposed technique considers near-critical paths becoming critical and always finds the path with the worst delay to determine $t_{delay}(chip)$. Thus, the paths with the worst delay always have near-zero thermal variances. However, all the other paths might still feature a negligible variance which has no impact on the overall chip delay.

To illustrate the delay variances within the timing paths, a sample set that covers a wide range of t_{delay} from timing paths within the designed chip is examined. Fig. 6.8a illustrates SHE-induced delay variances of the chip operating at a nominal voltage (V_{dd}=0.7V) where all paths are prolonged in their delay, as all cells operate in PTD and T_C is elevated. Fig. 6.8b illustrates SHE-induced delay variances of the chip employing N-ZTC (V_{dd}=$V_{ZTC}(chip)$=0.53V). As shown, the thermally induced delay variance is <0.1% in the critical paths. At the same time, delay variances in non-critical paths are larger (i.e., $\sigma(t_{delay})$<±1%). This is not an issue, as these paths will never become critical and thus cannot introduce timing violations. Nevertheless, the proposed technique minimized the delay variance as the original delay variance was $\sigma(t_{delay})$>90% and now became $\sigma(t_{delay})$<±1%.



(a) Thermally-induced delay variations due to SHE for a set of timing paths when operating the chip at 0.7V

(b) Thermally-induced delay variations due to SHE for a set of timing paths when employing N-ZTC

Figure 6.8.: Delay variance due to SHE-induced delay degradations of a set of paths within the designed chip. (a) variances due to SHE-induced delay degradation when operating at the nominal voltage (0.7V). (b) variances due to SHE-induced delay degradation employing N-ZTC (V_{dd}=0.53V).

Comparison between traditional guardbanding and N-ZTC:
The three possible operating points are:

- Case-A: Baseline at nominal voltage without SHE ($T_C(low)$).
- Case-B: Traditional guardband at nominal voltage with SHE ($T_C(high)$).
- Case-C: N-ZTC operation (lower V_{dd} and any T_C).

All cases are summarized in Table 6.1, and shown in Fig. 6.6 as well. Note that Case-A is a theoretical point, as it cannot be used to operate the circuit due to timing violations. This is expected, as Case-A would immediately exhibit timing violations when T_C increases above room temperature. Case-C (N-ZTC) does not reach the performance of Case-A. Instead, a delay degradation of 25% is still observed due to the lower V_{dd} when moving from nominal V_{dd} to $V_{ZTC}(chip)$. Notably, a 65% performance improvement can be observed of Case-C compared to Case-B due to the reduction of t_{GB}. In terms of power, N-ZTC reduces leakage power compared to Case-B and Case-A due to the reduced supply voltage, despite the elevated leakage from operating at high temperatures. The results are summarized in Table 6.1 in comparison with the theoretical baseline Case-A.

Table 6.1.: Comparison of the three possible operating points: Baseline, traditional guardband, and N-ZTC. Results are normalized to Case-A.

Case	$V_{dd}[V]$	GB	Delay inc.	Leakage	Freq. [GHz]	Reliable
A(Baseline)	0.7	No	0 [%]	100 [%]	1.77	No
B(Traditional)	0.7	Large	91 [%]	600 [%]	0.95	Yes
C(N-ZTC)	0.53	Near-zero	25 [%]	39 [%]	1.45	Yes

6.4. SHE Analysis on Multicore

This section evaluates the benefits of employing N-ZTC at the system level. Previously, delay analysis is linked directly to performance where minimizing guardband increases performance while reducing the voltage to $V_{ZTC}(chip)$ increases the delay and therefore reduces the performance. However, the system performance (e.g., the throughput of an application) differs from the circuit performance (e.g., cycles per second). Hence, the overall gain in the system performance must be evaluated. As shown previously, operating at lower V_{dd} reduces leakage power which reduces the total power. With the increase of the execution time and reducing power, the system's energy must be examined.

6.4.1. Experimental Setup

A multicore with four out-of-order cores modeling the *Gainestown* microarchitecture is configured and employed for evaluations. Each core is associated with private L1-I and L2-D caches with 32 KB each, and a private 256 KB L2 cache. The multicore contains an 8 MB shared L3 cache. The multicore is modeled to be implemented with the same 7nm PDK as employed previously for the OpenPiton design (see Section 6.3.1). Dynamic Voltage and Frequency Scaling sets frequencies from 0.95 GHz up to 1.77 GHz in multiples of 100 MHz. For simulation, Sniper [31] manycore simulator is employed, which allows multi-threaded simulation. McPAT[84] is used to estimate the power and energy consumption of the system. A set of applications are executed from the PARSEC benchmarks [22] with *simlarge* inputs. These applications cover compute-bound applications as well as memory-bound applications.

Because McPAT does not support 7nm FinFET, the power values are scaled of estimations performed with 45nm. To scale the power from 45nm to 7nm,

the OpenPiton processor is implemented using both the 45nm CMOS [95] and 7nm FinFET [39] to obtain scaling factors for dynamic and leakage power. These implementations follow the approach as described in Section 4.4.

6.4.2. Costs and Benefits from N-ZTC

Cases: The previous three cases summarized in Table 6.1 are examined here. Case-A is the baseline design without a guardband for SHE, which allows operating the multicore at a peak frequency of 1.77 GHz. This again is an unreliable point that features SHE-induced timing violations. This case acts as a baseline to see what theoretical performance would be achievable if SHE or thermal degradation would not be an issue. Case-B accounts for delay increases due to SHE and therefore employs timing guardband to its clock frequency, resulting in a lower frequency of 0.95 GHz. Case-C is employing N-ZTC by operating at $V_{ZTC}(chip)$=0.53V, with a near-zero timing guardband. However, lowering V_{dd} reduces the operating frequency from 1.77 to 1.45 GHz. As can be noticed, this is still faster than traditional guardbanding.

Four-threaded *PARSEC* applications are executed to fully utilize all cores operating at the voltage and frequency defined by each case. The benchmark execution time is recorded as a measure for system performance and the corresponding energy consumption.

Execution Time: Fig. 6.9a shows the execution time for different applications of the three cases. Results are normalized to the baseline Case-A. System performance of Case-A is faster than the reduced frequency in Case-B and slightly faster than Case-C. Importantly, applications are unequally affected by the reduced frequencies. For instance, while the performance of compute-bound applications like *blackscholes* scales almost linearly with the frequency, the performance of memory-bound applications like *canneal* depends strongly on the L3 and DRAM frequency, which is unchanged by operating at V_{ZTC}. In summary, N-ZTC shows better performance for all applications compared to the traditional SHE guardband technique.

Energy: Fig. 6.9b shows the energy consumption of the examined applications. Results are normalized to Case-B, as it uses the largest energy. Case-B uses the same voltage as Case-A but lower frequency due to the guardband. Hence, it takes the longest execution time and large power, which results in the highest energy consumption. Due to the timing violations when considering Case-A,

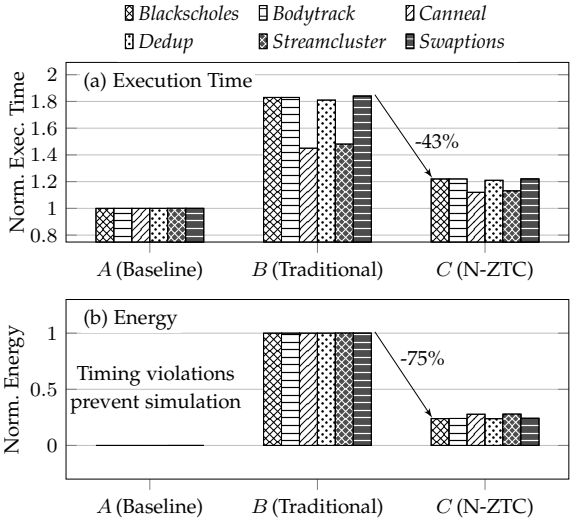


Figure 6.9.: Execution time and energy with the three studied cases. The baseline case does not employ SHE guardband and therefore it is unreliable case [126].

the energy results are not obtained. This is because Case-A is unrealistic since it causes timing violations. Thus its results are neglected.

Case-C operates at N-ZTC. Following the assumption that dynamic power is quadratic proportion to voltage [123], Case-C results in 58% lower dynamic power $(\frac{0.7-0.53V}{0.7})^2$ and these savings in the power consumption are larger than the loss in the execution time of 18% $(\frac{1.77-1.45GHz}{1.77GHz})$ and therefore energy should be reduced. The result of Case-C shows a reduction in energy consumption by 75% on average when employing N-ZTC where performance enhancement is up to 43% on average compared to Case-B, as shown in Fig. 6.9.

6.5. Summary and Conclusions

SHE-induced delay degradation, traditionally, can be mitigated by employing *large timing guardband* to guarantee error-free operation. Such guardband severely degrades the performance of the circuit. Operating near Zero-Temperature Coefficient (N-ZTC) can efficiently minimize the impact of SHE on the circuit's delay by eliminating the required guardbands. The presented

technique compromises all ZTCs within the chip, aiming to accurately locate the proper voltage closest to the ZTC of the chip with minuscule thermal variance. N-ZTC can entirely nullify the thermal-induced delay degradation with zero timing guardband for small circuits, such as ring oscillators. While for large circuits, such as a processor, near-zero guardband is still required. The simulation at the circuit level shows substantial improvements in performance and leakage power when employing N-ZTC compared to the traditional guardbanding technique. Similarly, the simulation at the system-level of multicore design shows significant enhancements w.r.t. performance and total energy when employing N-ZTC compared to traditional guardbanding technique.

Part II.

Low Power Computing: The Negative Capacitance Approach

7. NCFET-aware Modeling

7.1. NCFET: Physics and Device Modeling

Negative Capacitance Field Effect Transistor (NCFET) can be modeled (1) Metal-Ferroelectric-Metal-Insulator-Semiconductor (MFMIS) structure by adding an external ferroelectric (FE) layer within the conventional FET (e.g., FinFET) gate stack [104]. (2) Metal-Ferroelectric-Insulator-Semiconductor (MFIS) structure by replacing the gate oxide (i.e., high-κ layer) of conventional FET with a ferroelectric layer [104] (see Section 2.5).

In this work, NC-FinFET is modeled with a configuration of MFMIS, as shown in Fig. 2.9. The structure can be equivalently divided into the ferroelectric capacitor (C_{fe}) and the internal baseline FinFET (C_{int}) as shown in Fig. 2.10. The voltage amplification due to the ferroelectric layer is calculated as in Eq. (2.14). The baseline FinFET device is modeled using the industry-standard BSIM-CMG model [37, 26]. Then, the Negative Capacitance (NC) physics-based modeling for the NC effect is integrated within the industry-standard compact model of FinFET technology (BSIM-CMG). This is crucial for standard cell library characterization to create NCFET-aware libraries in order to be able to design and implement complex NCFET-based circuits, such as a full processor. The ferroelectric layer is modeled as shown in Eq. (7.1), which is based on the Taylor series as in the Landau-Khalatnikov (L-K) theory [82, 113]. Further details can be found in [53, 7].

$$V_{fe} = t_{fe}(2\alpha Q + 4\beta Q^3) \tag{7.1}$$

Where V_{fe} is the voltage across ferroelectric, Q is the gate terminal charge per unit area, t_{fe} is the thickness of ferroelectric. α and β are ferroelectric material-dependent parameters [82, 122].

The L-K equation is then solved in a self-consistent manner within the Verilog-A code of the BSIM-CMG model of FinFET [26] using a commercial SPICE simulator [153]. To model the FinFET parameters, the 7 nm Process Design Kit (PDK) is employed [39]. Regarding the ferroelectric layer, Al-doped HfO_2 material is considered due to its compatibility with CMOS material [93].

Fig. 7.1 summaries part of the device-level analysis [53]. To consider the role that the thickness of the ferroelectric layer play, various ferroelectric layer thicknesses are considered. The ferroelectric layer thickness is referred to as

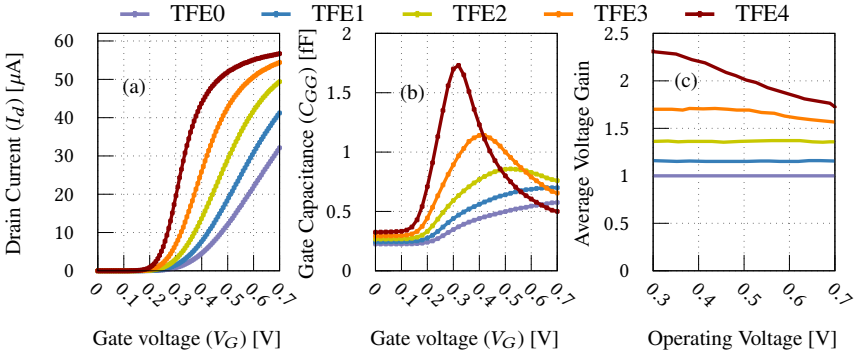


Figure 7.1.: Device-level analysis for the effects of different ferroelectric layers with varied thicknesses [53]. (a) shows the increase in the gate capacitance due to the negative capacitance. (b) presents the I_d current over V_G (at V_{DS} = 0.7V) demonstrating the increases in I_{ON} and the reductions in *SS*. In general, the thicker the ferroelectric layer, the higher the gain but, as a result, C_{GG} becomes much larger. (c) shows the average voltage gain (i.e., amplification).

TFEx, where x is the layer thickness in nanometers. TFE0, in this context, refers to the FinFET technology in which no ferroelectric layer is in use (more details in Section 7.2). As shown in Fig. 7.1(a), the existence of NC results in larger total gate capacitance (C_{GG}) of both NC-pFinFET and NC-nFinFET over baseline FinFET [103]. Moreover, as expected, the thicker the ferroelectric layer, the larger the NC is, and hence the C_{GG} becomes larger [103, 53]. In general, as shown in Fig. 7.1(a), for TFE4 as an example, the increase in C_{GG} is higher at lower V_G with a peak at around 0.35V. This, in turn, results in decreasing the sub-threshold swing and increasing the ON current, as shown in Fig. 7.1(b). Finally, Fig. 7.1(c) shows the average voltage amplification gain (A_{avg}), which is always greater than 1.

In summary, the thicker the ferroelectric layer, the higher the gain. As a result, the C_{GG} becomes much higher, which, together with an increased frequency, results in a higher dynamic power at the same operating voltage as explained earlier (see Section 2.5).

7.2. NCFET-aware cell library

As NCFET is an emerging technology and still in its infancy, real measurements of NCFET-based chip is unavailable where, for example, a processor chip has

not been fabricated yet. In all the analyses presented in this dissertation, full processor implementations are considered based on NCFET. Therefore, NCFET-based design, implementation, and analysis are performed through simulation based on a mature physics-based NCFET model. To reveal the trends of power and performance of a processor at different ferroelectric layers, a full processor's chip must be implemented first into gate-level then into a full layout (i.e., GDSII). Gate level using physical design (i.e., full circuit layout) is the standard and the most accurate method, considering circuit's parasitics and signal integrity, to examine the final chip using signoff tools before fabrication. To employ NCFET in circuit design, the standard chip design flow must be employed. Therefore, the NCFET-aware cell library is indispensable.

To model the electrical characteristics of pFinFET and nFinFET transistors, the modelcard from the ASAP7 PDK [39] is employed. The employed transistor model is BSIM-CMG [26]. The simulations, then, are performed for pFinFET and nFinFET under a wide range of voltages. To take the impact of the operating conditions into account, 7 input signal slews (t_{slew}) along with 7 output load capacitances (C_{load}) are considered. These are typical values for industrial and academic cell library characterization[151], which are also used for the original ASAP7 PDK [39].

On the other hand, the NC-FinFET model in Verilog-A code is employed within the commercial cell library characterization tool from Synopsys [154] to characterize the NCFET-aware cell libraries. The library contains the delay and power information of every standard sequential and combinational cells under the effects of NC in the standard *liberty* format. The post-layout cells' netlists, including the parasitics information, are obtained from the 7 nm PDK [39] for the characterization process to create the NCFET-aware cell libraries. Every cell is characterized under 7 input signal slews and 7 output load capacitances, as done previously for the baseline FinFET library.

The cell libraries are created for different voltages from the nominal voltage in the 7nm FinFET 0.7V down to 0.2V to accurately estimate how the power and frequency of NCFET-based circuit change as a function of voltage in comparison with FinFET. All the analyses in this work are applied at room temperature of $25°C$. To explore the impact of the ferroelectric layer thickness, four different thicknesses are considered from 1nm to 4nm in addition to the FinFET in which no FE layer is included. The ferroelectric layer thickness is referred to as TFEx, where x is the layer thickness in nanometers ($0 - 4$nm).

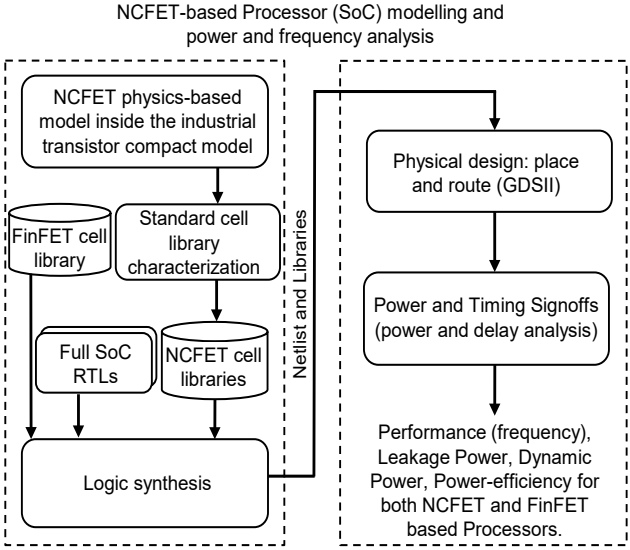


Figure 7.2.: Methodology of the NCFET-based processor modeling. Circuit-level modeling is a cross-layer implementation that links physics- and circuit-level aspects involved in designing an NCFET-based processor.

Note that TFE0 is FinFET. The thickness of the ferroelectric layer is limited to 4nm to ensure hysteresis-free operation [7].

7.3. NCFET-based Full-Chip Design

NCFET-aware cell libraries are fully compatible with the existing commercial EDA tool flows such as [152, 29]. Therefore, they can be deployed directly within the standard flow of chip design. Fig. 7.2 summarizes the implementation phases required for modeling and evaluating the NCFET-based processor. The evaluation covers the power and performance of the processor.

To reveal the trends of power and performance of a design (e.g., processor) at different ferroelectric layers, first, the Register-Transfer Level (RTL) of the circuit/design is synthesized into a gate-level implementation using a commercial logic synthesis tool [152] with NCFET cell libraries. During synthesis, the "*compile_ultra*" option within the Design Compiler is employed

to obtain well-optimized netlists under the highest optimization efforts. The gate-level design is then further realized using physical design processes (i.e., full circuit layout) as the standard method to examine the final chip before fabrication using the commercial layout tool flows [29]. This includes floor-planning, power delivery network (PDN), placement, routing, and optimization. Furthermore, the PDN is also optimized to gain the highest performance under the smallest voltage fluctuations considering the impacts of the circuit's RC-parasitic (resistances and capacitances) of the PDN on the voltage as changes in the circuit's delay during the analysis. This is done by using the industrial standard signoff tools by enabling the on-chip variation for signal and voltage integrity. By doing so, the tools simulate the impacts of PDN on the flow of the current as well as the reached voltage to the cells and hence the final delay and power of the chip.

The analysis at the gate level (i.e., chip implementation) focuses on studying the impact of NCFET on the power, energy, and delay of a full processor. Therefore, the physical implementation of memory structures, such as caches, is logic-based as neither SRAM modules nor memory compilers are available for NCFET yet. Flip-Flop (FF) based SRAM implementation is used in both FinFET (TFE0) and NCFET (TFE1-4) to enable fair comparisons. Moreover, the analysis shows that caches do not dominate the total power consumption with $< 21\%$ of the overall chip's power [129].

Finally, the processor's power and frequency (i.e., delay) are examined across the whole voltage range and for all ferroelectric layer thicknesses. The process is also done for FinFET. For accurate power and timing analysis, the complete RC-parasitics and interconnects of the entire chip are extracted and employed in Cadence Tempus Timing Signoff tool for delay analysis and Voltus IC Power Integrity signoff tool for power analysis. For accurate power estimations, Mentor QuestaSim is employed to extract the switching activity of the post-layout gate-level netlist. The extracted switching activity is then fed together with the final post-layout netlist and the RC-parasitics of the complete chip to the power signoff tool to estimate accurately the total power consumption of the chip. .

In this work, different types of processors have been studied, starting from a relatively small and low-power processor up to a high-performance one. As a summary, the following open-source single-core processors are implemented for analysis and comparisons: (1) A full tile of the state-of-the-art OpenPiton SoC, which is a general-purpose processor based on OpenSparc T1 architecture

Table 7.1.: Single-core processors comparison.

	BOOM	Rocket	OpenPiton
ISA	RISC-V	RISC-V	OpenSPARC T1
Word size	64bit	64bit	64bit
#Threads	2	1	1
Microarch.	Out-of-order	In-order	In-order
Branch-pred.	Yes	No	No
FPU	2	1	1
Pipeline stages	10	5	4
Reg. file WB	3w, 5r	2w, 2r	2w, 3r
Caches	all processors share the same cache configuration: 16K L1-I/D, 256K L2		

[18]. (2) The state-of-the-art 64-bit Rocket Processor [159], which is RISC-V processor. (3) Berkeley Out-of-Order Machine (BOOM) V2 [33] with medium configuration, which is an Out-of-Order 64-bit RISC-V Processor designed for high performance. Processors are configured to have relatively small caches. This is important to avoid caches from biasing results, as explained earlier. All the employed configurations in this work are summarized in Table 7.1. For instance, the resulting power and frequency of the OpenPiton SoC processor are presented in Fig. 7.3.

Fig. 7.3(a) shows how the maximum frequency (performance) of the OpenPiton increases with the ferroelectric layer. NC increases the total gate capacitance of FinFETs (see Section 7.1), which, together with the increased frequency, results in a higher dynamic power, as shown in Fig. 7.3(b). However, increasing the thickness of the ferroelectric layer inverses the dependency of leakage power on the operating voltage (V_{dd}) due to the negative drain-induced barrier lowering effect (DIBL) [105], as illustrated in Fig. 7.3(c) (see Section 2.5). Therefore, reducing V_{dd} at a high thickness increases the leakage power instead of decreasing it as in FinFETs technologies.

Importantly, all the examined processors show relatively similar trends in results with differences in the absolute values and the slop of the achieved values (e.g., BOOM shows the highest dynamic and leakage power).

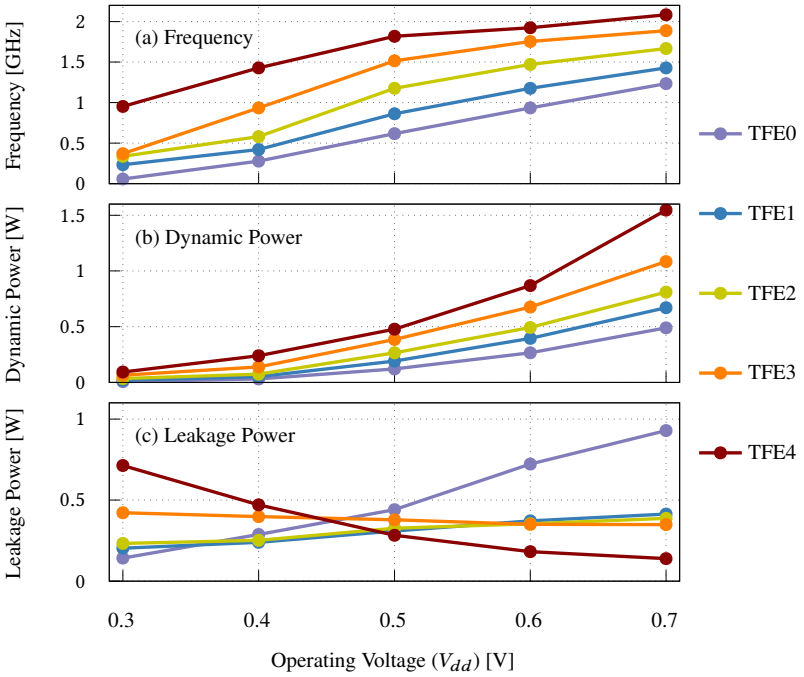


Figure 7.3.: Comparisons of frequency, dynamic, and leakage power of conventional FinFET technology and different NCFET technologies of OpenPiton processor. (a) NCFET boosts the maximum frequency of the processor at a given voltage. Gain increases with a thicker ferroelectric layer. (b) NCFET increases the dynamic power due to the increase in the frequency and total gate capacitance of the transistor. However, NCFET reduces the dynamic power at a fixed frequency as it allows to operate at a lower voltage. (c) NCFET with a thin ferroelectric layer weakens the dependency of leakage on voltage. At higher thicknesses, the dependency is reversed due to the negative DIBL.

8. NCFET-aware Voltage and Frequency Scaling

The power consumption of processor is proportional to the operating frequency (f) and operating voltage (V_{dd}). On the other hand, energy consumption measures the relationship between power consumption and the rate at which work is done (i.e., the integration of power consumption over execution time).

In conventional FET technology, both total power and total energy consumption are minimized by operating at the minimum voltage (V_{min}) that is required to sustain a frequency (f) under a given performance constraint [128, 125]. Reducing the supply voltage of a processor reduces both dynamic and leakage power. Similarly, due to the approximately linear relationship between frequency and voltage in the typical super-threshold region, despite the increase in execution time, both dynamic and leakage energy are minimized at reduced voltage, and frequency [51, 123, 125]. Traditional power/energy management techniques aim at operating the processor at the required frequency f_{min} and the corresponding minimum voltage (V_{min}) (i.e., V/f pair) to exploit these dependencies [88].

Because these dependencies vary among different technologies, power and energy management techniques must be revisited and investigated when a new technology is introduced. This holds even more when it comes to emerging technologies in which the underlying physics fundamentally differ from conventional FET technology.

NCFET is an emerging technology that has great potential to replace CMOS technology in the near future. As previously demonstrated, NCFET-based processors exhibit an observable performance enhancement compared to FinFETs in Section 2.5 and Fig. 7.3 by operating at the maximum possible frequency at a given V_{dd}. However, NCFET results in inverse dependency of the leakage power over voltage at the device and circuit level with higher thickness of the employed ferroelectric layer. This, in turn, could result in a novel trade-off between power components [128].

Hence, DVS and DVFS techniques for power/energy minimization in NCFET must be aware of this property. Therefore, this chapter presents the NCFET-aware power/energy management techniques.

8.1. Unique Properties of NCFET-based Processor

This section covers the importance and unique properties of NCFET-based processors.

Voltage selection for power minimization: Fig. 8.1(a) shows the total power consumption and its components (i.e., leakage and dynamic power) of a multicore system, designed in TFE4, running the master thread of PARSEC [22] *canneal* at 1 GHz. The minimum voltage (V_{min}) that is required to sustain this frequency is ≈0.2V. Starting from V_{min} and by increasing the operating voltage up to 0.6V under constant frequency, as shown in the figure, this increases the dynamic power but stronger decreases the leakage power. As a result, the total power consumption at first reduces towards higher voltage until an inflection point is reached, after which power starts to increase again as dynamic power becomes dominant. As shown, the power is minimized at a higher voltage (V_{opt}≈0.35V in this example) than the minimum voltage, which demonstrates that optimal voltage V_{opt} selection is a must in NCFET [128]. In turn, this has far-reaching consequences on power management techniques, which requires a non-intuitive voltage selection to be taken by the power management algorithm in order to optimally minimize the total power consumption [128].

Workload dependency: Different workloads have different runtime activities, and hence power components contribute differently to the total power. Accordingly, the leakage and dynamic power proportions to the total power change dynamically with workload behavior at runtime. This due to different characteristics of the running workloads. As V_{opt} depends on the trade-off between dynamic and leakage power, V_{opt} is expected to change with the workload being executed (i.e., different activities).

Fig. 8.1(b) shows the total power over voltage for two different workloads running on a multicore system designed in TFE4. Similarly, sweeping V_{dd} from V_{min} = 0.2V to 0.6V under constant frequency at 1GHz shows that V_{opt} differs for the two workloads. This demonstrates that the optimal voltage selection should follow the dynamics of workloads. Ignoring such attributes makes any power/energy management technique sub-optimal with respect to NCFETs technology.

Frequency and voltage selection for energy minimization: Operating a processor at a higher voltage allows the frequency to be increased accordingly.

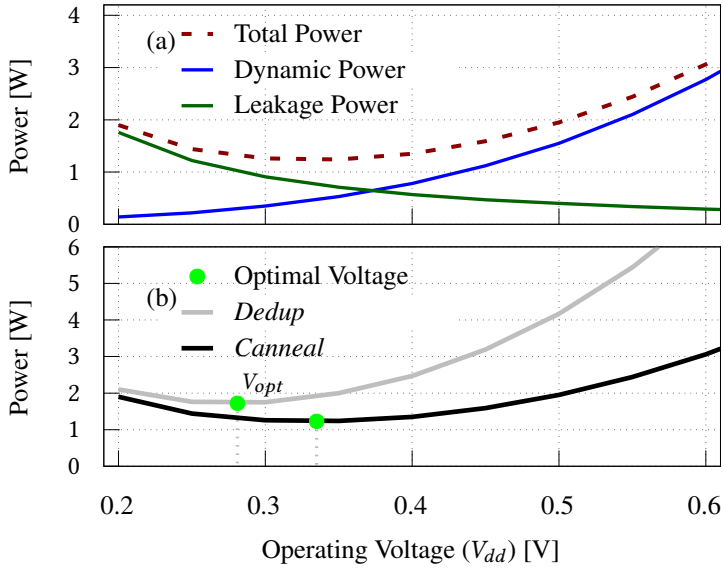


Figure 8.1.: (a) Total power and its components (i.e., leakage and dynamic) of the master thread of PARSEC *canneal* running on a multicore system designed in TFE4 at a fixed frequency (1 GHz) depend on the operating voltage. The total power decreases as voltage increases until it reaches an inflection point where it starts to increase again. (b) Total power consumption of two different workloads *canneal* and *dedup* running at the same frequency (1 GHz) over voltage. The total power of different workloads is minimized at different V_{opt}. Note that the ability of NCFET to operate at such low V (0.2 V) is due to the inherent voltage amplification provided by the integrated negative capacitance and does not result in near- or even sub-threshold computing.

Therefore, operating an NCFET-based processor at a higher frequency than f_{min} could result in lower energy. Minimal energy can be achieved by minimizing the power and operating at a proper frequency (e.g., fast execution with shorter leaking time) by trading off between power and execution time. In conventional FET technology, this is achievable by operating at the minimum required frequency f_{min} and minimum voltage V_{min} (see Chapter 3). In NCFET, this does not hold anymore where the minimal energy is achieved at a higher frequency than f_{min}.

On the one hand, in an NCFET-based processor, minimizing the total power consumption under fixed frequency always minimizes energy for all ferroelectric layer thicknesses. On the other hand, in some cases, increasing the operating frequency, under the corresponding voltage, increases the dynamic

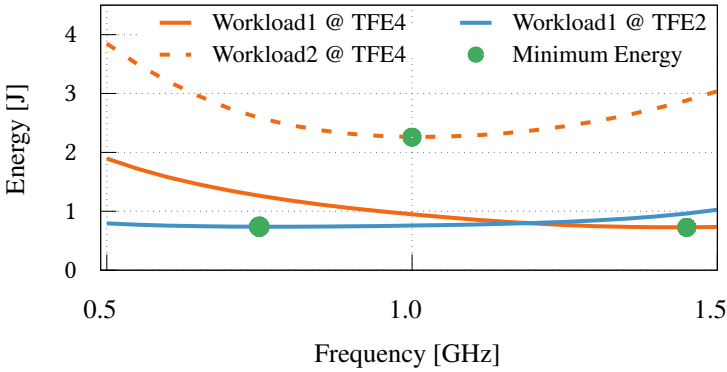


Figure 8.2.: Energy consumption over frequency, and the corresponding voltage, of two synthetic workloads with different dynamic power consumption running on a processor designed in TFE4. Energy is not minimized at the minimum operating frequency but at higher frequency. As the two workloads have different dynamic power consumption, their energy-minimizing frequencies differ. The figure also shows that the energy of the same workload can be optimized at different frequencies for two different FE layers (TFE2 and TFE4), showing the importance of selecting the optimal thickness at design time.

energy but more strongly decreases the leakage energy. Hence, the total energy decreases accordingly. This continues until an inflection point is reached where the dynamic energy becomes dominant and, therefore, increasing the frequency further starts to increase the total energy. To show such trends, Fig. 8.2 shows the energy of two synthetic workloads, with different dynamic power consumption, running on top of a TFE4 processor. As shown, the energy is minimized at a higher frequency than f_{min}=0.5GHz. Importantly, it also shows how the energy of two different workloads is minimized at different frequencies.

Thickness dependency: Fig. 8.2 presents the energy of a synthetic workload running on top of an NCFET-based processor using two different ferroelectric layer thicknesses. As shown, energy depends on the frequency and thickness of the ferroelectric layer. Different thicknesses are optimal (minimum energy) at different frequencies under the same workload, showing the importance of selecting the optimal thickness at design time.

In summary, (1) while a processor's power is always minimized at the lowest possible voltage that sustains the required frequency in FinFETs, this does not hold in NCFET at higher ferroelectric layer thicknesses [128]. Instead, operating at a higher voltage is more beneficial. (2) while a processor's energy is always minimized at the lowest voltage/frequency that satisfies the

performance constraint in FinFETs, this does not hold anymore in NCFET [124, 125]. Instead, operating at a higher voltage/frequency could be more beneficial. The optimal voltage/frequency is different for power or energy optimization. (3) thickness dependence besides workload characteristics must be considered when selecting voltage/frequency pairs for power and energy minimization. Therefore, developing new NCFET-aware power and energy optimization techniques is indispensable.

8.2. NCFET-aware Power and Energy Modeling

This section presents the application, frequency, power, and energy models that are needed for the *analytical design-space exploration* in NCFET processors. The results of this exploration are employed later to develop a runtime V/f pair selection algorithm that will be evaluated using detailed circuit- and system-level simulations. Power, energy, and frequency models are developed individually for each ferroelectric layer thickness. In this context, thickness is represented by (x).

8.2.1. Application Model

Two application models are covered here, as follows.

- NCFET-aware Dynamic Voltage Scaling (DVS) for power/energy minimization [128]. This model aims solely on selecting the optimal voltage V_{opt} at which the minimal power/energy is observed under constant frequency, satisfying given performance constraints. In this case, the required frequency is, therefore, the optimal frequency which is also the minimum frequency f_{min}. Note that this model implicitly reduces the energy by reducing the power at a constant frequency (i.e., fixed execution time). Hence, for this model, only the power modeling is thoroughly discussed. However, later, the energy results are presented.
- NCFET-aware Dynamic Voltage and Frequency Scaling (DVFS) for power/energy minimization [124, 125]. The required frequency is the

> minimum frequency (f_{min}). The optimal frequency (f_{opt}) is the operating frequency at which the processor's power/energy is minimized, which can be higher than f_{min}. Additionally, this model also aims on selecting the optimal voltage V_{opt} at f_{opt}.

$V_{min}(f_{opt})$ is the minimum voltage that is required to sustain f_{opt}. However, due to the negative DIBL that impacts NCFET transistors, the minimal power/energy could be achieved at a higher voltage, unlike conventional transistors, as previously mentioned. Therefore, $V_{opt}(f_{opt})$ is the optimal voltage for operating at f_{opt} [128].

The evaluation is done by assuming that the performance is linearly proportional to frequency to simplify the application model within the design-space exploration. This assumption significantly reduces the design-space corners. To explore the design-space, a synthetic workload is employed, which is defined as a *ratio of the dynamic to total power* running on a processor using the highest ferroelectric layer thickness running at the common highest frequency ($\hat{f}$) among all thicknesses (i.e., TFE4 at 1.2GHz, see Fig. 7.3). By sweeping this ratio, a large variety of workload domains can be covered from memory-bound to compute-bound applications. Later, a more realistic evaluation is presented based on the complex circuit- and system-level simulations under the execution of actual workloads.

8.2.2. Optimization Use Cases

Three optimization cases are covered, as follows:

(a) Power minimization under performance constraints with fixed work (W): Minimizes total power for executing a fixed amount of work W under the required frequency f_{min}. Thus, only the operating voltage is altered to minimize the average power consumption in this case. The execution time is not affected. The processor is assumed to be power-gated after finishing the execution W. Importantly, in the idle state, the processor can operate at a different voltage (V_{leak}) than in the active state to minimize leakage. Note, since a fixed frequency is considered, power and energy savings, as a percentage, are identical. Therefore, only the power is modeled here.

(b) Power minimization under performance constraints: Minimize average power for executing a given amount of work (W) under a fixed deadline (T) in a given period. The frequency can not be lower than a threshold (f_d) to fulfill T.

When the processor operates at a higher frequency, then its execution comprises two parts: First, the processor is active for T_{active}. Then, the processor is idle for $T - T_{active}$. During the idle phase, the processor is assumed to be clock-gated, i.e., its dynamic power is suppressed, but leakage continues. Importantly, in the idle state, the processor can operate at a different voltage (V_{leak}) than in the active state to minimize leakage.

(c) Energy minimization under fixed work (W): Minimizes total energy for executing a fixed amount of work (i.e., non-periodic work). Energy consumption is only incurred during the execution time of W and hence no idle time exists.

In all scenarios, a single thread is being executed on a single core. Furthermore, in the idle state, the processor operates at V_{leak} to minimize the leakage power. Based on the dependencies of leakage power over voltage, shown previously in Fig. 7.3(c), the highest voltage ($V_{leak} = 0.7$ V) will be used with TFE3-4 in idle state and the lowest voltage ($V_{leak} = 0.3$ V) with TFE0-2.

8.2.3. Power and Frequency Models

To develop the dynamic power, leakage power, and frequency models, their actual values, using the physical chip implementation of NCFET, must be obtained first. Follow the methodology in [115, 7, 128], the processor chip is designed and optimized using the standard chip design flow. Then, the dynamic and leakage power as well as the frequency of the processor are examined for the whole voltage range from 0.2V to 0.7 V, with 50mV steps. The process is done for all ferroelectric layer thicknesses individually (1-4 nm including FinFET), details in Section 7.3. Finally, the results are fitted into mathematical equations to use them within the analytical design space exploration.

The minimum voltage $V_{min}^{(x)}(f)$ at thickness x is required to sustain a frequency f, and inversely, the maximum sustainable frequency $f_{sus}^{(x)}(V)$ at thickness x and voltage V are given by:

$$V_{min}^{(x)}(f) = \left(\frac{\frac{1}{f} - c_{freq}^{(x)}}{a_{freq}^{(x)}} \right)^{\frac{1}{b_{freq}^{(x)}}} \tag{8.1}$$

$$f_{sus}^{(x)}(V) = \frac{1}{a_{freq}^{(x)} \cdot V^{b_{freq}^{(x)}} + c_{freq}^{(x)}}, \tag{8.2}$$

where $a_{freq}^{(x)}$, $b_{freq}^{(x)}$, $c_{freq}^{(x)}$ are constant fitting parameters.

Leakage and peak dynamic power when operating at $f_{sus}^{(x)}(V)$ are:

$$P_{leak}^{(x)}(V) = a_{leak}^{(x)} \cdot V^{b_{leak}^{(x)}} \tag{8.3}$$

$$P_{dyn,peak}^{(x)}(V) = a_{dyn}^{(x)} \cdot V^{b_{dyn}^{(x)}} + c_{dyn}^{(x)} \tag{8.4}$$

Here, $a_{dyn}^{(x)}$, $b_{dyn}^{(x)}$, $c_{dyn}^{(x)}$, $a_{leak}^{(x)}$, $b_{leak}^{(x)}$ are constant fitting parameters.

When operating at a frequency lower than $f_{sus}^{(x)}$ and the corresponding V, dynamic power is scaled linearly, and hence:

$$P_{dyn,max}^{(x)}(V, f) = \frac{f}{f_{sus}^{(x)}(V)} \cdot P_{dyn,peak}^{(x)}(V) \tag{8.5}$$

8.2.4. Workload-Dependence of Power and Energy:

Dynamic power consumption $P_{dyn}^{(x)}(V, f)$ is affected by the running workload, which induces some switching activity on the processor. The dynamic power consumption is scaled by a factor $r_{dyn} \geq 0$ from the peak dynamic power $P_{dyn,max}^{(x)}(V, f)$:

$$P_{dyn}^{(x)}(V, f) = r_{dyn} \cdot P_{dyn,max}^{(x)}(V, f) \tag{8.6}$$

The total power consumption $P^{(x)}_{total}(V, f)$ is the sum of dynamic and leakage power:

$$P^{(x)}_{total}(V, f) = P^{(x)}_{dyn}(V, f) + P^{(x)}_{leak}(V) \tag{8.7}$$

r_{dyn} is not constant since it represents the current workload activity that depends on the dynamic/total power ratio as a variable. The dynamic/total power ratio is defined, therefore, as r_{dyn} observed at the highest thickness $\hat{x}$=4 nm, at the highest common frequency $\hat{f}$ and the minimum required voltage $\hat{V} = V^{(4)}_{min}(\hat{f})$:

$$\begin{aligned} dyn/tot &= \frac{P^{(\hat{x})}_{dyn}(\hat{V}, \hat{f})}{P^{(\hat{x})}_{total}(\hat{V}, \hat{f})} \\ &= \frac{r_{dyn} \cdot P^{(\hat{x})}_{dyn,max}(\hat{V}, \hat{f})}{r_{dyn} \cdot P^{(\hat{x})}_{dyn,max}(\hat{V}, \hat{f}) + P^{(\hat{x})}_{leak}(\hat{V})} \end{aligned} \tag{8.8}$$

r_{dyn} can be calculated from a given dyn/tot as follows:

$$r_{dyn} = \frac{dyn/tot \cdot P^{(\hat{x})}_{leak}(\hat{V})}{P^{(\hat{x})}_{dyn,max}(\hat{V}, \hat{f}) \cdot (1 - dyn/tot)} \tag{8.9}$$

The average total power consumption under a performance constraint f_{min} with a fixed workload (i.e., optimization use case (a)) $P^{(x)}_{avg}(V, f)$ is the sum of power in active and idle reigns, where V_{leak} is selected at design time as described earlier:

$$P^{(x)}_{avg}(V, f) = P^{(x)}_{dyn}(V, f) + P^{(x)}_{leak}(V_{leak}) \tag{8.10}$$

The average total power consumption under a performance constraint f_{min} (i.e., optimization use case (b)) $P^{(x)}_{avg}(V, f)$ is the sum of power in active and idle reigns, where V_{leak} is selected at design time as described earlier:

$$P^{(x)}_{avg}(V, f) = (P^{(x)}_{dyn}(V, f) + P^{(x)}_{leak}(V)) \cdot \frac{W}{f} + P^{(x)}_{leak}(V_{leak}) \cdot (T - \frac{W}{f}) \tag{8.11}$$

(c) Total energy: Total energy for use case (c) is:

$$E^{(x)}_{total}(V, f) = (P^{(x)}_{dyn}(V, f) + P^{(x)}_{leak}(V)) \cdot \frac{W}{f} \tag{8.12}$$

8.2.5. Optimal Frequency and Voltage Selection

V_{opt} and f_{opt} that minimize total power can be obtained from the power model in the form of a minimization problem:

$$V_{opt}(f, r_{dyn}) = \underset{V^{(x)}_{min}(f) \le V \le V^{(x)}_{max}}{\arg\min} P^{(x)}_{avg}(V, f) \tag{8.13}$$

$$f_{opt}(r_{dyn}) = \underset{f^{(x)}_{min} \le f \le f^{(x)}_{max}}{\arg\min} P^{(x)}_{avg}(V_{opt}(f, r_{dyn}), f) \tag{8.14}$$

V_{opt} and f_{opt} that minimize total energy can be similarily obtained from the energy model:

$$V_{opt}(f, r_{dyn}) = \underset{V^{(x)}_{min}(f) \le V \le V^{(x)}_{max}}{\arg\min} E^{(x)}_{total}(V, f) \tag{8.15}$$

$$f_{opt}(r_{dyn}) = \underset{f^{(x)}_{min} \le f \le f^{(x)}_{max}}{\arg\min} E^{(x)}_{total}(V_{opt}(f, r_{dyn}), f) \tag{8.16}$$

Power/energy management (i.e., DVS/DVFS selection) is, therefore, an optimization problem that can be solved by exploring the design space of

$P^{(x)}_{total}(V,f)$, $P^{(x)}_{avg}(V,f)$ or $E^{(x)}_{total}(V,f)$ over all possible voltages and frequencies as well as ferroelectric layer thickness. Note again, in optimization case (a), the optimal frequency is f_{min} always.

8.2.6. Design Space Exploration Algorithm:

DVS and DVFS techniques operate under some constraints, e.g., to maximize performance given a power/energy budget or to minimize power/energy given a performance goal [88, 118]. The constraints are fulfilled by selecting the proper V/f pairs. V_{opt}/f_{opt} selection following Eq. (8.13), Eq. (8.14), and Eq. (8.16) (i.e., optimization use cases) are optimization problems that can be solved using a search algorithm. The algorithm to perform the required optimization is summarized in Algorithm 8.1. It performs a search by sweeping across all possible frequency and voltage steps. Since the number of discrete frequency/voltage settings is limited by the hardware and the analytical models, it can be evaluated quickly, and hence, the search is fast. In addition, it can be applied either online (i.e., executing Algorithm 8.1 at runtime) or offline (i.e., pre-characterizing a processor at design time and selecting predefined V/f pairs at runtime).

The offline technique works by characterizing the processor operating point as pairs of V_{opt}/f_{opt} as a function of workload ratios and performance at design time. Given the actual measured or derived workload characteristics and performance goals, operating points V_{opt}/f_{opt} can then be selected by the operating system or hardware at runtime.

8.3. Experimental and Evaluation Methodology

The experimental setup is summarized in Fig. 8.3. As shown, the experimental setup consists of three main parts:
(1) processor-level power and performance modeling using OpenPiton SoC processor [18]. Based on the previously obtained processor analysis in Fig. 7.3, the analytical power and performance models are created as described in Section 8.2.3. These models are integrated later within circuit- and system-level simulators for runtime selection.

Algorithm 8.1 NCFET-aware power/energy management algorithm for optimal operating point selection that minimizes power/energy.

Require: Power models for TFEx: $P_{avg}^{(x)}$, Energy models for TFEx: $E_{total}^{(x)}$, voltage range $[V_{min}, V_{max}]$, voltage step ϵ, frequencies (f_r), dyn/total power ratios, Work W, Time T, Deadline (required) frequency f_d
Ensure: Optimal frequency f_{opt} at optimal voltage V_{opt}

$f_{opt} \leftarrow f_{min}$
$rdyn$ at given $dyn/total$ ratio ▷ Eq. (8.9)
for each f in f_r **do**
    $V_{opt1} \leftarrow V_{min}^{(x)}(f)$ ▷ Eq. (8.1)
    repeat
        $power_a = P_{avg}^{(x)}$ for f=f_d=f_{min} ▷ Eq. (8.10)
        $power_b = P_{avg}^{(x)}$ for W, T, $f \geq f_d$ ▷ Eq. (8.11)
        $energy_c = E_{total}^{(x)}$ for W ▷ Eq. (8.12)
        if $Min.>$ {energy$_c$ OR power$_b$ OR power$_a$ } **then** ▷ Use case?
            $f_{opt} \leftarrow f$, $V_{opt} \leftarrow V_{opt1}$
        end if
        $V_{opt1} \leftarrow V_{opt1} + \epsilon$ ▷ iterative update
    until $V_{opt1} = V_{max}$ ▷ Termination criteria
end for
return f_{opt}, V_{opt}

(2) *System-level* simulation to evaluate the efficiency of a multicore system under the effects that conventional (i.e., NCFET-unaware) DVS and NCFET-aware DVS have.
(3) *Circuit-level* is a gate-level simulation to evaluate the efficiency of the processor under the effects that conventional (i.e., NCFET-unaware) DVFS and NCFET-aware DVFS have. Note that, as shown in Fig. 8.3, the NCFET-aware DVS has different implementation than DVFS. This is due to the complexity of the implementation and to avoid inaccuracy (i.e., approximation level for voltage and frequency implementation) at the system level. This because the scaling-based approach inherently could result in inaccuracies because it cannot accurately model differing device-level characteristics, and hence it is limited to DVS only. Therefore, the discussion of the experimental setup is presented here, and later the evaluation for each technique is individually presented. However, NCFET-aware DVS is also implemented at the circuit level for a fair comparison with NCFET-aware DVFS.

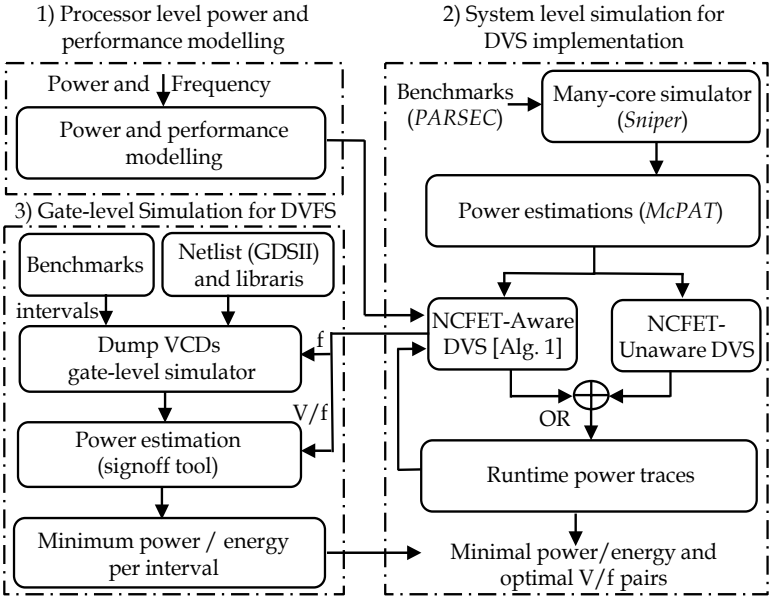


Figure 8.3.: Methodology for NCFET-based processor modeling besides analytical and experimental NCFET-aware power/energy management evaluation. (1) Analytical modeling of power and frequency of the OpenPiton processor. (2) System-level modeling and evaluation. (3) circuit-level (gate-level) modeling and evaluation.

8.3.1. NCFET-aware DVS Experimental Setup and Exploration

DVS analytical design-space exploration: Fig. 8.4 shows the design space with NCFET-aware DVS (V_{opt}) [128] and NCFET-unaware DVS (V_{min}) [88, 118, 166] by operating both cases at f_{min}. NCFET-unaware DVS sets the minimum voltage needed to sustain the required frequency, and therefore it does not consider workload behavior (i.e., characteristics). Contrarily, NCFET-aware DVS does consider the workload characteristics, using Eq. (8.10) and Algorithm 8.1, as it depends on the ratio of dynamic to total power. The explored design space in Fig. 8.4 reveals the following:

(a) Two trends can be observed: (1) the higher the required frequency, the higher V_{opt} is. This is consistent with NCFET-unaware DVS. (2) the lower the

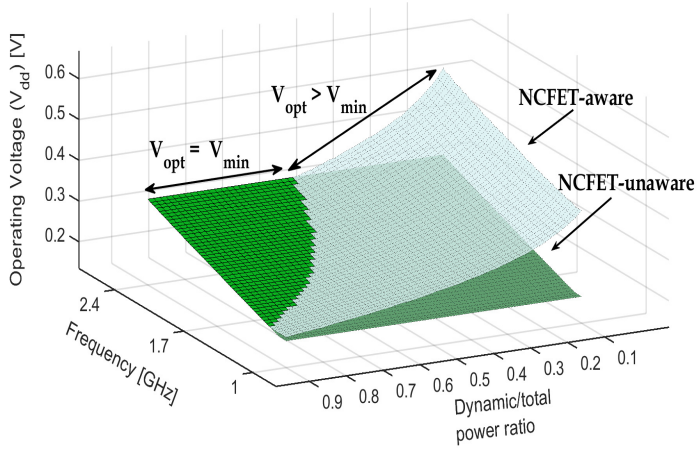


Figure 8.4.: Comparison of V_{dd} selected by NCFET-unaware (V_{min}) and NCFET-aware (V_{opt}) DVS based on frequency and leakage to total power ratio. NCFET-unaware DVS always selects V_{min} that sustains the required frequency. NCFET-aware DVS selects higher voltages (V_{opt}>V_{min}) for low frequencies or low dynamic/ total power ratio to minimize total power.

dynamic to total power ratio, the higher V_{opt} is. This is because leakage power gets prominent. Therefore, it should be reduced by selecting higher V_{dd}.

(b) Two distinct regions exist: (1) For high dynamic to total power ratio and for high frequencies, both techniques (NCFET-aware and conventional NCFET-unaware) select the same voltage (i.e., V_{opt}=V_{min}). (2) For low ratios of dynamic to total power or low frequencies, NCFET-aware DVS selects higher voltages than the minimum voltage to minimize the total power (V_{opt}>V_{min}).

System-level simulation: NCFET-aware DVS technique is evaluated at the system level using a multicore (2×2) system. Each core has private L1-I and L1-D caches with 32 KB. The per-core private L2 caches have a size of 256 KB, each. The 8 MB L3 cache is shared among all four cores. The *HotSniper* toolchain [107] is used to simulate the multicore system. It combines the *Sniper* multicore simulator[31] with a periodic invocation of *McPAT*[84] for runtime power estimation. Tasks are examined from the PARSEC benchmark suite [22], which is commonly used to evaluate multicore system. Since *McPAT* does not support the NCFET technology, power at 45nm using *McPAT* is examined first, and then scale dynamic and leakage power to 7nm NCFET. Therefore, the *OpenPiton* SoC is additionally implemented at the 45nm technology node

[95]. The frequency-dependent scaling factors are obtained by comparing the dynamic and leakage power consumption of both technology nodes based on [115]. The frequencies are set between 1.0 GHz and 2.4 GHz. The maximum frequency limit of 2.4 GHz comes from the employed *McPAT* at 45nm. V_{dd} is set between 0.2 V and 0.7 V. Notably, the low V_{dd} (i.e., 0.2V) does not result in sub-threshold computing due to the inherent voltage amplification provided by the negative capacitance.

For fair comparisons, both DVS cases are configured to have: the same frequencies, voltage range, and architecture, besides running the same benchmarks. Thus, only voltage selection differs based on the DVS decision.

8.3.2. NCFET-aware DVFS Experimental Setup and Exploration

DVFS analytical results exploration: Based on the power and frequency characterization of the processor (see Fig. 7.3), the power and frequency models are characterized (see Section 8.2.3), covering optimization use cases (b) and (c) (see Section 8.2.2). Afterward, Algorithm 8.1 is applied under given runtime constraints (e.g., T and W) to determine the optimal configurations for all possible workloads and ferroelectric layer thicknesses. To cover a wide range of workloads, dynamic/total power ratios are examined in the range of 0.1-0.9 for W=10^6 cycles, T=20ms (time to finish $10W = 10^7$ cycles at f_{min}=0.5GHz), and a performance constraint of f_d=0.8GHz to meet T. This exploration reveals the optimal operating pairs (V/f) under the given constraints and all possible configurations. The optimal pairs (V/f) reveal that TFE4 has the highest performance (i.e., highest frequency) among all thicknesses (more details in Appendix B.2).

Using the optimal pairs, the dependence of minimum power and energy on ferroelectric layer thickness is examined under the optimization use cases (b) and (c) (i.e., power and energy minimization scenarios) (see Section 8.2.2). Minimum power and energy results for different thicknesses and application characteristics are shown in Fig. 8.5. In Fig. 8.5(a), for power minimization, minimal power is at TFE4 for all workloads. Results also show that TFE3 has the highest power most of the time since it has the highest leakage power when operating at V_{leak} compared to others (see Fig. 7.3). By contrast, in Fig. 8.5(b), the minimal energy is again at TFE4 for most cases (i.e., for most workloads).

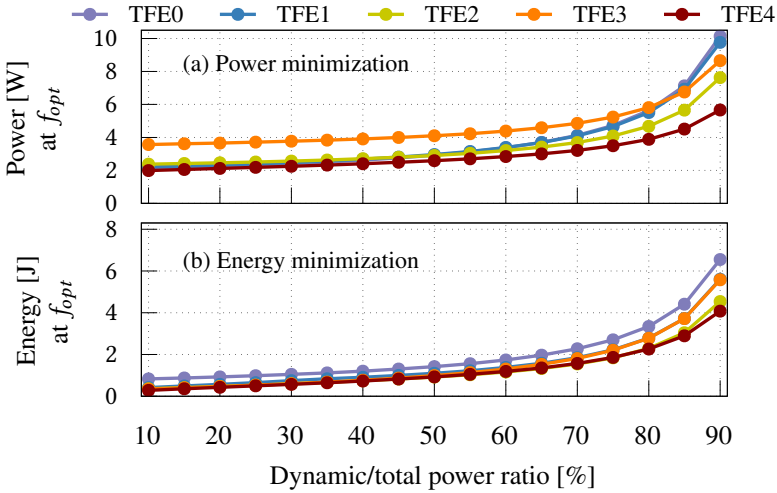


Figure 8.5.: Optimal power and energy over dynamic/total power ratios for different TFEx operating at optimal frequency f_{opt} using W=10^6 and T=$10W$. (a) Power minimization. TFE4 has the minimum power all the time. (b) Energy minimization. TFE4 has the minimum energy most of the time. For a small region 0.6-0.75, TFE2 has slightly better energy.

TFE2 and TFE4 are very close, where TFE2 is optimal for a small region between 0.6-0.75 with only small improvements. However, TFE4 achieves by far the best performance (see Appendix B.2). As a result, in both scenarios, *TFE*4 shows the minimal power and energy in addition to the best performance (i.e., higher f_{opt}) among all thicknesses. The preference, therefore, is for *TFE*4 as it minimizes energy and power at high performance. Therefore, DVFS is evaluated in TFE4 only.

DVFS at circuit level: The analytical design space exploration to expose the optimal configurations covers all possible operating corners under given constraints using synthetic workloads (i.e., varying the dynamic to total power ratio). Synthetic workloads are assumed to have certain properties that may differ from real applications. These properties were selected to allow a comprehensive analytical exploration. Real workloads may show execution phases with different power at runtime. Therefore, to minimize the power/energy of such a workload, it will be required to operate at various frequencies and voltages at runtime. Furthermore, the performance might not scale linearly with frequency as assumed in the analytical exploration. Therefore, V/f pair selection should follow dynamic workload characteristics. To consider the

dynamic characteristics of workloads, real applications must be evaluated using circuit-level simulation. However, as power/energy management techniques, such as DVFS, are complex algorithms, it is still unfeasible to simulate the full technique at the circuit level. This is due to the considerable simulation complexity and computational cost. Therefore, the DVFS technique can still be manually examined by simulating the runtime selections.

In practice, power/energy management always consists of two parts: (1) predicting future workload characteristics and (2) selecting (V/f) pairs based on the prediction. To make the simulation independent of any DVFS algorithm and provide an upper bound on the possible gains, an oracle approach is considered to assume workload characteristics are known in prior. This allows focusing on reporting the benefits from V/f selection in isolation while ignoring the impact of potentially wrong decisions made by heuristic or predictive techniques. NCFET-aware power and energy management can be employed for both pre-characterized and unknown applications. As the optimal V/f selection depends on the dynamic power consumption of the workload, it needs to be known or estimated when determining the V/f pair for the next execution phase. For pre-characterized workload, this is trivial. However, V/f selection needs to be combined with a heuristic or predictive technique for workload characterization for unknown applications. Note that this implementation is also done for *NCFET-aware DVS* as well for further comparisons.

Circuit-level simulation: Fig. 8.3(3) summarizes the circuit-level (i.e., gate-level) simulation used to evaluate V/f pair selection at runtime under realistic workloads. This simulation is complex, computationally expensive, and time-consuming. However, it is still important to provide an accurate evaluation that precisely considers the actual workload characteristics. Simulation is done by mimicking the functionality of the power/energy management technique using the optimal configurations obtained from the analytical exploration (i.e., V_{opt}/f_{opt} pairs). This is done by, first, splitting the execution time into equal intervals. Next, the power and energy are examined for every V_{opt}/f_{opt} pair for each interval. This process is a gate-level (i.e., post-layout) simulation, where VCD files are generated for every f_{opt}. Then, power is estimated using the power signoff tool for every interval by operating the circuit at V_{opt}. Finally, powers/energies are compared for similar intervals to find V_{opt}/f_{opt} with minimal power/energy. Again, the process is applied for *NCFET-aware DVS* as well.

8.4. Evaluation and Comparison

8.4.1. NCFET-aware DVS Evaluation and Analysis

This section evaluates the effectiveness of the NCFET-aware DVS [128] based on the optimization use case (a) in Section 8.2.2. First, it shows how NCFET-aware DVS saves power. Then it shows how total power saving varies at runtime. Later, to demonstrate the effectiveness of the NCFET-aware DVS technique, the energy savings are reported for different benchmarks compared to NCFET-unaware DVS [118, 166]. Note, as DVS minimizes power under fixed frequency constraints, power, and energy savings as a percentage would be identical (energy is the power consumption over time $E = P_{avg} * T$, $T = 1/f$).

When NCFET-aware DVS saves power: As explained previously, NCFET-aware DVS selects a higher V_{dd} than conventional DVS for low frequency or low ratio of dynamic to total power. This area is highlighted in Fig. 8.6 ($V_{opt} > V_{min}$). Fig. 8.6 also shows the ratio of leakage power to total power for a representative set of PARSEC benchmarks operating at different frequencies. Different workloads exhibit different ratios of dynamic to total power ratio. This ratio increases with increasing frequency because dynamic power consumption increases more strongly than leakage. For illustration purposes, the figure does not show all PARSEC benchmarks. As shown, almost for all scenarios, it necessitates selecting a higher operating voltage than V_{min} to minimize the total power. This experiment demonstrates that NCFET-aware DVS is required not only in some corner cases but in almost all execution scenarios of workloads. In the case where $V_{min} = V_{opt}$, conventional DVS already selects the optimal operating voltage, also selected by the NCFET-aware DVS.

How NCFET-aware DVS saves power (runtime example): Fig. 8.7 illustrates a runtime example of the master thread of PARSEC *canneal*, which shows distinct phases during execution phases. The total power consumption during phase-1 gradually decreases, as shown in Fig. 8.7b. The operating frequency is set at 1.7GHz. NCFET-unaware DVS sets V_{dd} to the minimum voltage (0.28V) required to sustain this frequency. In turn, dynamic power is minimized, but the leakage power is high. NCFET-aware DVS sets V_{dd} to a higher value (0.37V), which increases the dynamic power but stronger decreases the leakage power resulting in a lower total power compared to NCFET-unaware DVS. As can be noticed, V_{dd} is not constant. Instead, it varies

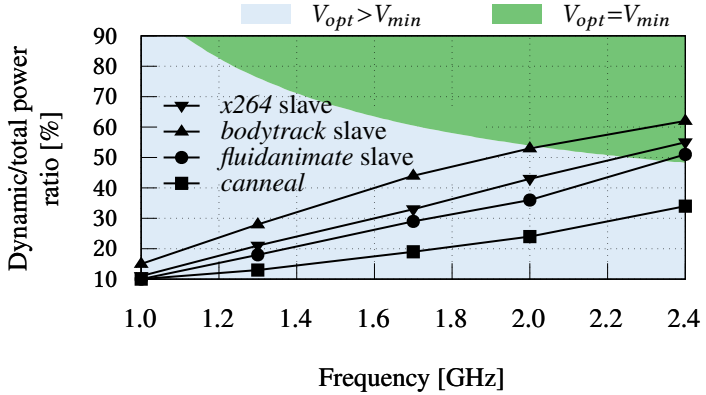


Figure 8.6.: Conventional DVS only selects the optimal V_{dd} (V_{opt}=V_{min}) for some PARSEC benchmarks when operated at very high frequency. In all other cases, total power is minimized at a higher operating voltage ($V_{opt}>V_{min}$) as in NCFET-aware DVS selection.

(i.e., increases slightly) over time. This is because dynamic power decreases, and therefore it is more beneficial to decrease the leakage power.

During phase-2, the master thread is become idle and awaits the termination of the slave threads. The frequency is, therefore, reduced to the minimum frequency (1.0GHz). In such a case, the leakage power dominates the total power. NCFET-unaware DVS would reduce V_{dd} down to 0.2V due to the low required frequency. Operating at such a low voltage strongly *increases* the leakage power in NCFET. Contrarily, NCFET-aware DVS increases the voltage to 0.53V to minimize the leakage power. In turn, the total power consumption during phase-2 is decreased by 67% compared to the NCFET-unaware DVS. Once the slaves are terminated, the master resumes operation in phase-3, and its frequency is increased again to 1.7GHz. However, CPU activity here is very low due to the frequent memory accesses. NCFET-aware DVS exploits this by using a higher V_{dd} than in phase-1, even though the same frequency is used in the two phases. The total energy consumption of all phases has been reduced by 17%. Importantly, NCFET-aware DVS does not statically increase V_{dd}, but in fact, it results in opposite behavior. NCFET-unaware decreases V_{dd} in phase-2, whereas V_{dd} needs to be increased to minimize the total power as shown at point B in Fig. 8.7a where V_{dd} is contradictorily selected. Furthermore, the V/f pairs model used in NCFET-unaware DVS does not hold anymore in NCFET. As shown in Fig. 8.7a for points A and C, the CPU operates at the same frequency but has different selected V_{dd}.

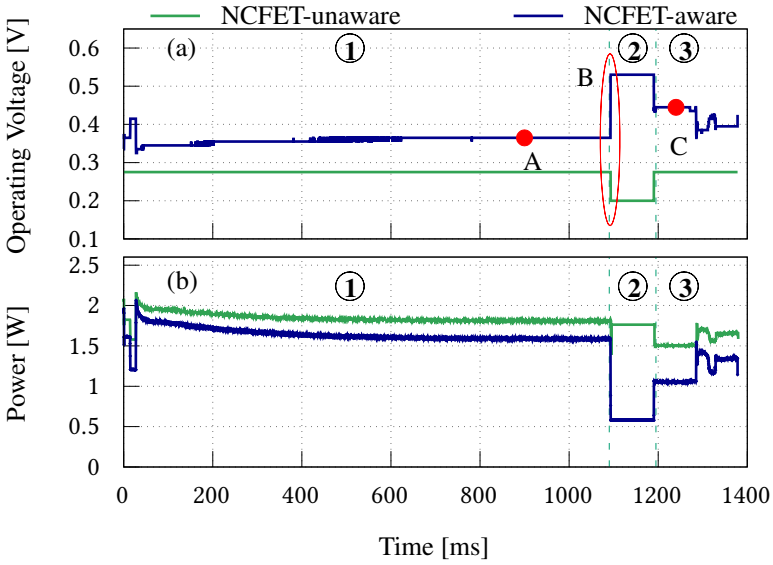


Figure 8.7.: (a) operating voltage V_{dd} and (b) total power consumption of the *canneal* master thread with NCFET-aware DVS and NCFET-unaware DVS. NCFET-aware DVS over scales the voltage based on the workload and thereby reduces the total power by up to 67% in phase-2 at the same CPU frequency and results in total energy savings of 17%. NCFET-unaware DVS fails when it comes to NCFET. As shown for points A and C, they have the same frequency but different voltages. At point B, voltage is *contradictorily* selected between the two DVS techniques.

Energy Savings: Fig. 8.8 summarizes the energy savings for different PARSEC benchmarks with *simsmall* inputs when active threads are operated at 1.7GHz, and idle cores are throttled to 1.0GHz. The DVS techniques do not affect performance since the frequency is the same with both techniques. Therefore, the only difference is the selected V_{dd}, which results in a different total power. Energy savings range from 14% and up to 27% for *blackscholes* and *dedup*, respectively.

Two factors affect the observed gains: (1) the CPU utilization that affects the dynamic power consumption of the active threads, and (2) the idle times of the threads due to the synchronization between threads. The higher the dynamic power consumption (i.e., dynamic to total power ratio) of active threads, the lower are the possible gains for these threads. This is the reason why *swaptions*, for instance, results in low gains. Long idle times of threads result in higher

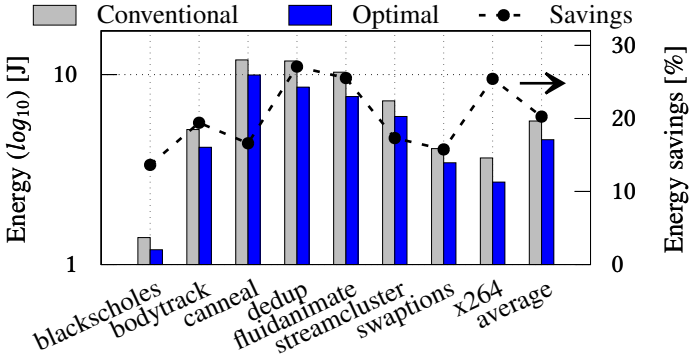


Figure 8.8.: Energy results and energy savings of different benchmarks running at 1.7GHz using NCFET-aware DVS in comparison with NCFET-unaware DVS. NCFET-aware DVS technique results in up to 27% energy savings (20% on average) while still providing the same performance.

gains since the total power consumption during idle times mainly consist of leakage which is reduced by NCFET-aware DVS technique. Overall, the average energy saving is 20%.

8.4.2. NCFET-aware DVFS Evaluation and Analysis

This section evaluates the effectiveness of the NCFET-aware DVFS [124, 125] based on optimization use cases (b) and (c) in Section 8.2.2. It also presents the achievable power and energy savings, using NCFET-aware power and energy management compared with NCFET-unaware techniques and the previously presented NCFET-aware DVS.

As demonstrated previously, TFE4 shows minimal power and energy over all thicknesses at f_{opt} (see Fig. 8.5). TFE4 also shows the highest frequency over all thicknesses (i.e., best performance). Therefore, evaluation covers power and energy only in TFE4 for three different scenarios:
(1) *NCFET-aware voltage and frequency selection (DVFS)*[124]: the processor operates at f_{opt} and $V_{opt}(f_{opt})$. (2) NCFET-aware voltage selection (DVS)[128] based on NCFET-aware DVS, the processor operates at the minimum frequency f_{min} that is required to meet the performance goal and the optimal voltage $V_{opt}(f_{min})$. This comparison is to show the impact of NCFET-aware frequency selection. (3) NCFET-unaware (conventional) technique [88, 166] where the processor operates at the minimum frequency f_{min} required to meet a

performance goal and the minimum voltage V_{min} required to sustain that frequency. All scenarios are summarized in Table 8.1.

NCFET-aware DVFS is evaluated at the circuit-level (gate-level) as described in Section 8.3.2 following the runtime scenarios that describe the different optimization use cases (see Section 8.2.2). Three representative micro-benchmarks are considered from [101] for evaluation: Matrix-Matrix multiplication (MM), Memory Test (MT), and Fast Fourier Transform (FFT).

The average dynamic/total power ratios for FFT, MM, and MT benchmarks at TFE4, $\hat{f} = 1.2$GHz and nominal voltage $\hat{V} = V_{min}(\hat{f})$ (matching the definition of dynamic/total power ratios) are 0.49, 0.61, and 0.68, respectively. Note that the examined workload ratios will depend on applied techniques to reduce leakage power/energy.

Workload executions are split into intervals to select V/f pair per interval. Note that optimal frequency and voltage selection does not depend on the interval size, but it is only a function of the average workload behavior in each interval. Due to limitations on the length of gate-level simulations, an interval size of $20-30$ instructions is selected.

Assuming an oracle approach that can predict the operating point before the interval start and employing V_{opt}/f_{opt} pairs that resulted from the analytical optimization, the power, and energy are estimated for every interval on all possible pairs. Then, the power/energy consumption is compared of all similar intervals to select V_{opt}/f_{opt} pairs that optimally minimize power/energy. Notably, the same used constraints in the analytical exploration are applied within the gate-level simulation, $f_{min} = f_d = 0.8$ GHz for power minimization and $f_{min} = 0.5$ GHz for energy minimization.

To highlight the fundamental differences for runtime frequency and voltage selection and for visualization, Fig. 8.9 plots samples of frequencies and voltages for three intervals when running the FFT benchmark. Fig. 8.9(a) and (b) show the selected frequencies and voltages by NCFET-aware DVFS

Table 8.1.: Scenarios for comparison in the DVFS evaluation.

Scenario	Freq.	Voltage
(1) NCFET-aware voltage and frequency selection (DVFS)	f_{opt}	V_{opt}
(2) NCFET-aware voltage selection (DVS)	f_{min}	V_{opt}
(3) NCFET-unaware technique (conventional)	f_{min}	V_{min}

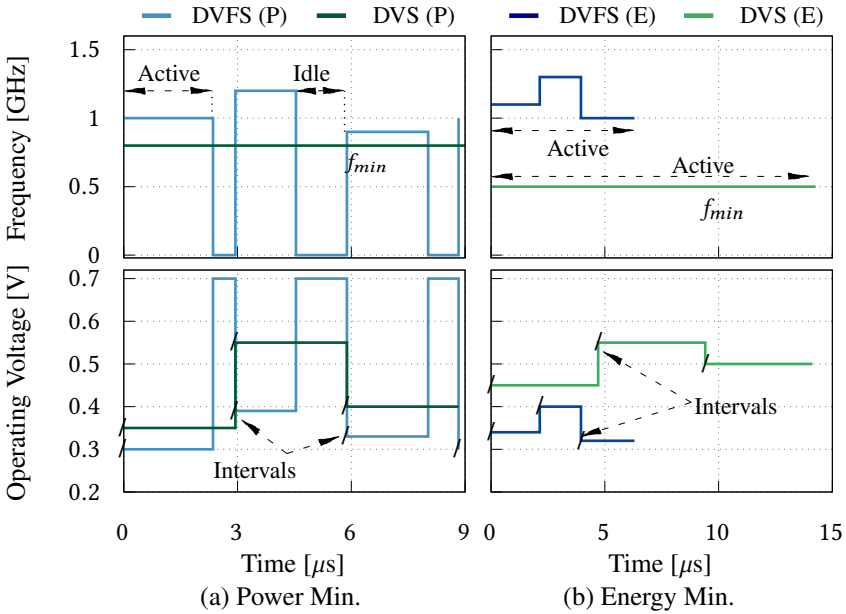


Figure 8.9.: Samples of three intervals while running FFT benchmark showing the selected frequencies and voltages by NCFET-aware DVFS and NCFET-aware DVS for (a) power minimization (P) and (b) energy minimization (E). NCFET-aware DVS selects f_{min} for both cases. f_{min}=f_d=0.8 GHz and f_{min}=0.5 GHz for power and energy minimization, respectively. By contrast, the NCFET-aware DVFS technique always selects higher frequencies, as shown in the top row for (a) and (b). For power minimization, NCFET-aware DVFS switches between idle and active regions. *Different V_{opt} are selected for the same interval based on the optimization goal.* For the power minimization case, two different voltages are selected in the active and idle region. The voltage is set to V_{leak}=0.7 V to reduce leakage power in the idle region.

compared to NCFET-aware DVS for power and energy minimization cases, respectively.

As shown for frequency selection, the NCFET-aware DVS technique uses f_{min} all the time. For power minimization, NCFET-aware DVS selects f_{min} = f_d = 0.8GHz, while for energy minimization, it selects f_{min} = 0.5GHz. On the other hand, NCFET-aware DVFS always selects higher frequencies than f_{min}. Frequencies, as expected, vary with intervals as the activity varies, which causes a varying dynamic power consumption. However, power minimization results in lower frequencies to be selected compared to the energy minimization case. This is due to the idle phase, where the processor has to wait for the

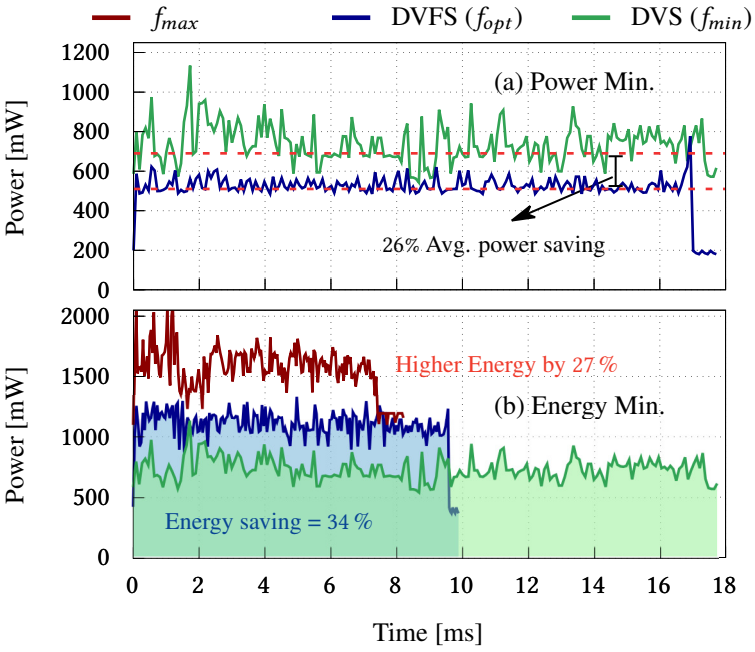


Figure 8.10.: (a) Power traces of FFT benchmark for power minimization case using NCFET-aware DVFS and DVS. Both scenarios finish at the same time. NCFET-aware DVFS results in lower average power. The average power is reduced by 26% in this particular example. (b) Power traces of FFT benchmark for energy minimization case using NCFET-aware DVFS and DVS. In NCFET-aware DVFS, the processor rushes to finish the execution early and shows 1.8x better performance in addition to lower energy (i.e., the highlighted area under the curve). Energy savings in this example are up to 34%. Additionally, a simple rush-to-completion strategy is examined at f_{max} and V_{opt} that provides fastest execution but shows 27% higher energy compared to the NCFET-aware DVFS technique.

deadline T (i.e., leaks more). Therefore, a slower execution is more beneficial to balance between powers and times of the active and idle regions.

As shown for voltage selection, for the two optimization cases, different V_{opt} are selected by NCFET-aware DVFS scenario for each interval. For the power minimization case, the voltage in the active region is relatively lower than the voltage in active region of the energy minimization case as a lower frequency is in use. In the idle region, the voltage is set to $V_{leak} = 0.7$V to reduce leakage power (see Section 8.2.3). Notably, the figure shows that different V_{opt}/f_{opt} are selected for the same interval based on the optimization goal.

Power and Energy Runtime Examples: The total runtime power/energy is the combination of all intervals operating at the optimal V/f pairs. Fig. 8.10(a) shows the power trace over time when running the FFT benchmark at the optimal V/f pairs for the power minimization case using NCFET-aware DVFS and DVS. As shown, DVFS results in lower average power compared to DVS. The power savings for this example are on average 26% compared to DVS. Fig. 8.10(b) shows the runtime power consumption for the energy minimization case when running the FFT benchmark at the optimal V/f pairs. As shown in the DVFS scenario, the processor rushes to finish earlier. Hence, the power consumption is higher than in the DVS scenario but with a shorter execution time. The overall energy (i.e., area under the curve) is, however, reduced. This is because the reduction in execution time is larger than the power increases (i.e., $E = P{\cdot}D$). The energy savings for this example are up to 34% compared to DVS. Moreover, performance is improved as $f_{opt} > f_{min}$. In this example, performance is enhanced by 1.8x. This again still comes with lower energy, as no energy-performance trade-off has to be made.

In addition, energy consumption is compared against a simple rush-to-completion strategy that runs at maximum frequency f_{max} and then goes to idle [55]. f_{max} is selected at the optimal voltage V_{opt} similar to the DVS voltage selection for a fair comparison. As shown in Fig. 8.10(b), such an approach results in 27% higher energy compared to the NCFET-aware DVFS technique. While operating at f_{max} achieves the fastest execution time, the associated increase in dynamic energy at higher voltages outweighs the gains from a larger idle time in TFE4.

Power and Energy Savings: The power and energy savings using NCFET-aware DVS in comparison with NCFET-aware DVS and an NCFET-unaware (conventional) technique are summarized in Fig. 8.11. The average power savings are up to 32% compared to the DVS scenario and up to 46% compared to the conventional scenario. The energy savings are up to 42% compared to DVS and up to 58% compared to a conventional scenario. Moreover, the performance improvements can reach up to 2.1x for the MT. Again, as shown before, NCFET-aware DVS results in higher savings than the conventional technique, as DVS is NCFET-aware, albeit for voltage selection only.

NCFET-aware DVFS vs. NCFET-aware DVS: Crucially, results for NCFET-aware DVFS show that, depending on the workload, minimal power/energy is achieved at a higher frequency than the performance constraint would re-

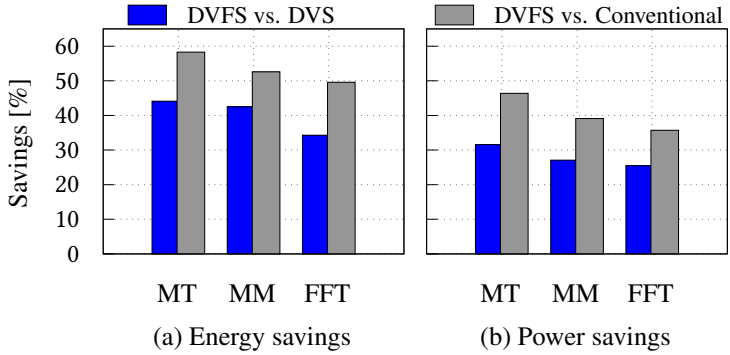


Figure 8.11.: Circuit-level simulation (a) energy savings, (b) power savings with NCFET-aware DVFS (DVFS) in comparison to NCFET-aware DVFS (DVS) and conventional techniques using three benchmarks. Savings are up to 58% and up to 49% compared to conventional and DVS scenarios, respectively, for the energy minimization case. For power minimization, savings are up to 46% and 32% compared to conventional and DVS scenarios, respectively.

quire. In other words, even optimal power management may necessitate more complex frequency optimization than NCFET-aware DVS alone.

8.5. Summary and Conclusions

NCFET is a promising emerging technology that provides outstanding performance in addition to better power optimization compared to FinFET technology. As conventional power and energy minimization techniques are unaware of the inverse dependency that leakage power exhibits in NCFET, they are suboptimal w.r.t. NCFET-based processors. This chapter presented NCFET-aware management techniques to optimize the power/energy of NCFET-based processors. The analysis demonstrated that the optimal frequency to achieve minimal power/energy is larger than the minimum frequency. The largest ferroelectric thickness provides both the best power/energy and performance. Simulations by applying the proposed techniques at the circuit level to characterize a real workload behavior demonstrated a considerable saving compared to state-of-the-art techniques.

9. NCFET-based Heterogeneous Manycore Design

In NCFET, the employed ferroelectric layer is the key design to optimize performance and power. However, in practice, the end goals of the overlying system determine the optimal ferroelectric layer thickness of an NCFET circuit under the optimization goal. Optimal thickness also depends on the operating voltage. For instance, a large thickness can be utilized to boost up the performance at any voltage. In contrast, low thickness can be employed to reduce power at a relatively high voltage.

The unique properties of NCFET open questions about their overall impacts on architecture and microarchitecture designs at both single-core and manycore levels. Manycore designs emerged to provide higher power efficiency and overcome the physical constraints of single-core [148, 129]. Homogeneous manycore (see Fig. 9.1(a)) integrates cores with identical microarchitectures[27] and heterogeneous manycore (see Fig. 9.1(b)) uses cores with different microarchitectures [80] to improve the efficiency and hence provides better power-performance trade-offs. Integration of cores with different microarchitectures comes with high design cost and runtime overheads [142].

However, a heterogeneous NCFET manycore can be achieved by changing the ferroelectric layer thickness among cores while keeping the microarchitecture untouched, as shown in Fig. 9.1(c). This allows designers to have benefits similar to conventional heterogeneous manycore without the associated overheads as all the cores share the same microarchitecture. However, the benefits of such a design are neither modeled nor quantified.

This chapter investigates employing cores with heterogeneous NCFET technology within the manycore design that can be the key to achieve a superior overall trade-off between power and performance (i.e., power efficiency). For instance, designers can increase the thickness of one or more core(s) at the cost of higher power to speed up the execution of critical modules that bottleneck the performance. Thus, the overall power efficiency is improved. This chapter explores, through analytical and quantitative modeling, system- and application-level benefits of NCFET-based heterogeneous manycore design in terms of performance and power-efficiency compared to the state-of-the-art FinFET-based designs.

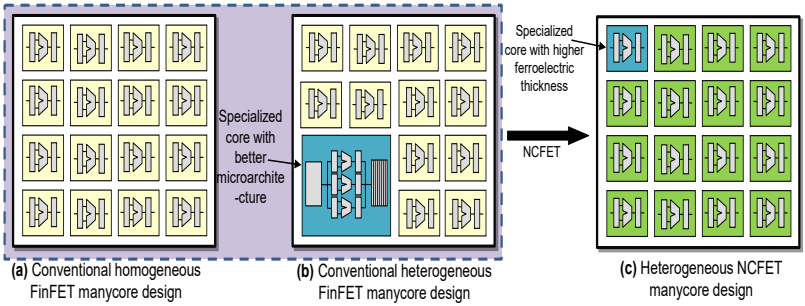


Figure 9.1.: An abstracted overview depicting (a) conventional homogeneous FinFET manycore, (b) conventional heterogeneous FinFET manycore combining different microarchitectures, and (c) the novel heterogeneous NCFET manycore combining different ferroelectric layer thicknesses with identical microarchitectures [129].

9.1. NCFET-based Heterogeneity Design

The novel heterogeneous NCFET manycore design is illustrated in Fig. 9.1, where the heterogeneity in NCFET is merely achieved by varying the ferroelectric layer thickness per core.

To investigate whether the heterogeneous NCFET manycore can provide better power efficiency than the conventional heterogeneous FinFET manycore, without any heterogeneity in core's microarchitecture, the following implementation phases are employed: (1) Amdahl's law is extended to model and quantify the power efficiency of heterogeneous NCFET manycore design. Using the extended Amdahl's law, a comprehensive analysis is presented of the design space associated with the power efficiency optimization of NCFET manycore to investigate the maximum theoretical power efficiency gains with NCFET manycore over the conventional FinFET manycore. To study the role of microarchitecture on the power efficiency of manycore designs, two different processor microarchitectures are modeled and analyzed. (2) The heterogeneity in NCFET is compared to the conventional heterogeneity in the microarchitecture. This is done by performing cycle-accurate single-core processor simulations at the RTL level to compare the energy and performance of in-order NCFET cores with out-of-order FinFET cores. (3) Then, performing system-level simulations of manycore processors to study the impact of the two sources of heterogeneity on manycore designs.

Amdahl's Law: Classical Amdahl's law [5] quantifies the maximum performance gains attainable using parallel processing in manycore systems. It quantifies the performance gain (speedup) of an application that has non-overlapping serial and parallel parts (phases) of size $(1-F)$ and F, respectively. The parallel part executes on all N processors, while the serial part executes only on one processor. Eq. (9.1) defines the performance of the parallel execution of an application under Amdahl's law.

$$Perf = \frac{Execution_{serial}}{Execution_{parallel}} = \frac{1}{(1-F) + F/N} \tag{9.1}$$

However, [148] extended Amdahl's law for manycore under different optimization goals, and [58] extended Amdahl's law to quantify the performance of conventional heterogeneous manycore under area constraints. Performance alone is an insufficient metric as high performance at the cost of unsustainably high power might not be acceptable. Power efficiency E measured in performance-per-watt is now the metric for evaluation of manycore designs.

It (i.e., E) is proportional to the inverse of the energy.

$$E = \frac{Perf}{Power} \propto \frac{1}{Energy}. \tag{9.2}$$

On the other hand, [169] extended Amdahl's law to quantify power efficiency in manycore designs. In general, Amdahl's law has been continuously extended to quantify the potential of new design techniques [17, 169, 148].

9.1.1. Implementation and RTL Simulation of Single-Cores

This work studies two open-source single-core processors for analyses and comparisons (details in Section 7.3). The NCFET heterogeneity is compared with conventional heterogeneity in the microarchitecture. To study the performance and energy of manycore with heterogeneous microarchitecture, the BOOM [33], and Rocket [14] processors are selected. The main reason is that these processors share the same ISA, caches configurations, and many other modules, but BOOM implements an out-of-order pipeline in contrast

to the in-order pipeline of Rocket. Table 7.1 presents further details of the microarchitecture of the processors.

Performance analysis (RTL simulation): To examine the performance/throughput difference between BOOM and Rocket, Synopsys VCS is employed for cycle-accurate RTL simulation by running a set of benchmarks on both processors and report the number of cycles. Such results are crucial to calibrate the system-level simulator into the correct configurations when mapping single-core parameters. Additionally, this helps to obtain the average energy consumption per benchmark of the single-core processor at different technologies and then finds how NCFET can improve performance and energy.

Power and frequency analysis: The Rocket chip is implemented with every technology (i.e., FinFET and NCFET with varying ferroelectric layer thickness) following the presented methodology in Section 7.3. On the other hand, BOOM is only implemented in FinFET.

9.1.2. Simulation of Multi-Threaded manycore

The *Sniper* simulator[107, 31] is employed to simulate the manycore processors with different microarchitectures (in-order and out-of-order) implemented in different technologies (FinFET and NCFET with varying ferroelectric layer thickness). However, to allow the execution of an application that utilizes all cores, in-order and out-of-order cores are selected to have the same ISA. BOOM and Rocket cores implement the same RISC-V ISA. *Sniper* already comes with configurations that are calibrated for BOOM. Additionally, the Rocket is modeled as an in-order version of BOOM.

McPAT [84] is used to estimate power consumption. Since *McPAT* is unaware of NCFET and following the same approach presented in Chapter 8, the power consumption at 45 nm is examined and then to scale the power consumption to 7nm NCFET. The scaling is done by comparing delay, dynamic, and leakage power signoffs of the processor at different voltages if implemented in 45nm Bulk CMOS, 7nm FinFET, or 7nm NCFET. Voltage- and frequency-aware scaling factors are extracted to scale the power traces.

A scaling-based approach inherently results in inaccuracies because it cannot accurately model differing device-level characteristics. However, relative power numbers of 7nm FinFET and 7nm NCFET still are comparable because

of the used voltage- and frequency-aware scaling factors that reflect the non-linear dependency of leakage and dynamic power on the voltage and frequency. *PARSEC* benchmarks [22] are used with simsmall inputs for evaluations and record their execution time and energy.

9.2. Analytical Modeling and Analysis of NCFET-based Manycore Designs

Based on the processor-level power and frequency signoffs of different processors at different voltages and ferroelectric thicknesses (see Chapter 7), the analytical models are derived for the NCFET-based manycore processors.

9.2.1. Power and Frequency of Single NCFET-based Core

Voltage and thickness affect the leakage and dynamic power of an NCFET core in different ways. Therefore, the leakage and dynamic power need to be separately examined. An NCFET core exhibits the lowest performance at the lowest ferroelectric thickness x (0 nm, i.e., FinFET) and voltage (0.3 V), where the performance is equivalent to the lowest performance of a FinFET core (i.e., no ferroelectric layer is in use). Hence, all results are normalized to this lowest performance configuration.

Let x within the range x_{min}=0 to x_{max}=4 represents the ferroelectric thickness of an NCFET core. The NCFET core using voltage scaling can operate at a voltage v, ranging from v_{min} to v_{max}. Section 7.3 restricts the values of v_{min} and v_{max} to 0.3 V and 0.7 V, respectively.

Let $\tilde{f}$, $\tilde{P_D}$, and $\tilde{P_L}$ be the operating frequency, dynamic power, and leakage power of a FinFET core (x_{min}) at v_{min}. These parameters depend on the implemented processor. The dynamic power consumption $\tilde{P_D}$ additionally depends on the application being executed. Let $f(x, v, \tilde{f})$, $P_D(x, v, \tilde{P_D})$, and $P_L(x, v, \tilde{P_L})$ be the operating frequency, dynamic power and leakage power of an NCFET core of a given x, v, and $\tilde{f}$. Then $\theta_f(x, v)$, $\theta_D(x, v)$, and $\theta_L(x, v)$ capture the scaling in the frequency, dynamic power, and leakage power of

the NCFET core with a change in x and v normalized to $\tilde{f}$, $\tilde{P_D}$, and $\tilde{P_L}$, respectively:

$$f(x, v, \tilde{f}) = \theta_f(x, v) \cdot \tilde{f} \tag{9.3}$$

$$P_D(x, v, \tilde{P_D}) = \theta_D(x, v) \cdot \tilde{P_D} \tag{9.4}$$

$$P_L(x, v, \tilde{P_L}) = \theta_L(x, v) \cdot \tilde{P_L} \tag{9.5}$$

The performance is normalized by assuming each FinFET core provides a unit performance at a v_{min} (i.e., $\tilde{f}$=1). This assumption further simplifies Eq. (9.3):

$$f(x, v) = \theta_f(x, v) \tag{9.6}$$

In order to obtain a continuous equation for $\theta_f(x, v)$, $\theta_D(x, v)$, and $\theta_L(x, v)$, a non-linear regression on empirical data is generated from the power and performance signoff tools of the core.

9.2.2. Application Execution Model

General-purpose processors are able to run a variety of applications with different characteristics. Traditionally, Amdahl's law uses the parallelization factor F to capture an application's scope of parallelization. A fraction F of the total work of the application parallelizes perfectly among all cores, while the remaining $(1 - F)$ fraction of the work executes only serially on a single core. The execution time of an application is assumed to be inversely proportional to the frequency of the underlying core.

Different applications consume different amounts of power on a FinFET core. Therefore, they will also consume different amounts of power when executing on an NCFET core [128]. Let $P_T(x, v, \tilde{P_D}, \tilde{P_L})$ represent the total power of an NCFET core for given x, v, $\tilde{P_L}$, and $\tilde{P_D}$.

$$P_T(x, v, \tilde{P_D}, \tilde{P_L}) = P_D(x, v, \tilde{P_D}) + P_L(x, v, \tilde{P_L}) \tag{9.7}$$

$$= \theta_D(x, v) \cdot \tilde{P_D} + \theta_L(x, v) \cdot \tilde{P_L} \tag{9.8}$$

Eq. (9.8) is obtained using Eq. (9.4) and Eq. (9.5). Let λ be the ratio of dynamic power to leakage power on a FinFET core at v_{min} for a given execution.

$$\lambda = \frac{\tilde{P_D}}{\tilde{P_L}} \tag{9.9}$$

$\tilde{P_D}$ depends on the execution, while $\tilde{P_L}$ is independent of the execution. Therefore, all power calculations can be normalized assuming $\tilde{P_L}$=1, which implies $\tilde{P_D}$=λ. After normalization Eq. (9.8) becomes

$$P_T(x, v, \lambda) = \theta_D(x, v)\lambda + \theta_L(x, v) \tag{9.10}$$

FinFET manycore simulations are executed using *PARSEC* benchmarks using the setup described in Section 9.1.2 to obtain a realistic range for λ. Results show the values of λ fall within the range of 0.01 and 0.2. The running application can show different behavior during the serial and parallel phases. Therefore, let $P_T^S(x, v)$ and $P_T^P(x, v)$ be the total power during the serial phase and parallel phase, respectively.

$$P_T^S(x, v, \lambda^S) = \theta_D(x, v)\lambda^S + \theta_L(x, v) \tag{9.11}$$

$$P_T^P(x, v, \lambda^P) = \theta_D(x, v)\lambda^P + \theta_L(x, v) \tag{9.12}$$

9.2.3. Manycore System Modeling

As previously mentioned, Amdahl's law has been continuously improved to cover new optimization keys. Hence, Amdahl's law must be improved again to capture the impact of NCFET on heterogeneous manycore. The models use N cores with the same microarchitecture and have the same area budget. The heterogeneity is achieved by varying the ferroelectric thickness of the cores. All cores operate at voltage v. NCFET, however, can exhibit different frequencies at different thicknesses, even at the same voltage. Note that unused idle cores are always power-gated to save power.

Heterogeneous NCFET Manycore: Let $M_{x,y}$ represent heterogeneous NCFET manycore with N NCFET cores. One core in $M_{x,y}$ has thickness x (i.e., specialized core), and all the remaining $N-1$ cores have a thickness y. The application's serial phase always uses the specialized core. In contrast, the application's parallel phase uses all the cores (including the specialized core).

Let $Perf_{M_{x,y}}$ represent the performance of an application executed on $M_{x,y}$. The application executed on $M_{x,y}$ compared to M experiences a speedup by order of $f(x,v)$ during its serial phase. It then experiences a performance speedup by order of $f(y,v)$ on $N-1$ parallel cores and an additional speedup by order of $f(x,v)$ on the specialized core.

$$Perf_{M_{x,y}} = \frac{1}{\frac{1-F}{f(x,v)} + \frac{F}{f(y,v)(N-1)+f(x,v)}} \tag{9.13}$$

Let $P_{M_{x,y}}$ represent the average power consumption of an application on $M_{x,y}$. Dividing total energy with total execution time gives $P_{M_{x,y}}$.

$$P_{M_{x,y}} = \frac{\frac{(1-F)P_T^S(x,v,\lambda^S)}{f(x,v)} + \frac{(N-1)FP_T^P(y,v,\lambda^P)+FP_T^P(x,v,\lambda^P)}{f(y,v)(N-1)+f(x,v)}}{\frac{1-F}{f(x,v)} + \frac{F}{f(y,v)(N-1)+f(x,v)}} \tag{9.14}$$

Let $E_{M_{x,y}}$ represent the power efficiency of an application on $M_{x,y}$. $E_{M_{x,y}}$ is achieved by combining Eq. (9.13) and Eq. (9.14).

$$E_{M_{x,y}} = \frac{1}{\frac{(1-F)P_T^S(x,v,\lambda^S)}{f(x,v)} + \frac{(N-1)FP_T^P(y,v,\lambda^P)+FP_T^P(x,v,\lambda^P)}{f(y,v)(N-1)+f(x,v)}} \tag{9.15}$$

9.2.4. Results of Analytical Analysis

The design space of the optimization process of the thicknesses of NCFET heterogeneous manycore is characterized for a given application, besides the gains in power efficiency. Gains are reported with NCFET over FinFET at the same voltage. NCFET always has a higher frequency than FinFET at the same voltage (see Fig. 7.3). Therefore, the power efficiency gains are accompanied always by a higher performance with NCFET over FinFET. Hence, efficiency gains do not come at the expense of performance.

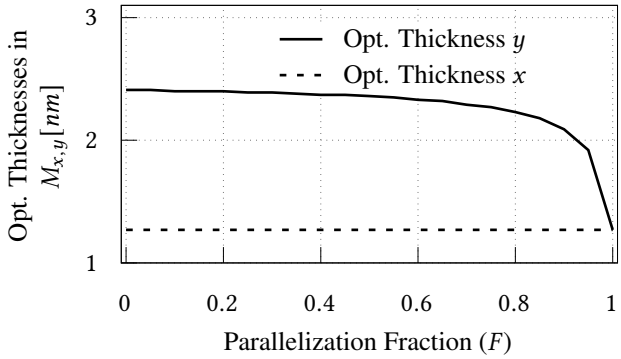


Figure 9.2.: The optimal ferroelectric thicknesses x and y in a heterogeneous NCFET manycore design. The optimal thickness of specialized cores x and remaining cores y in asymmetric NCFET $M_{x,y}$ for different values of parallel fraction F with other parameters fixed (v=0.7 V, λ_S=0.01, λ_P=0.2, N=64). The optimal thickness of specialized cores x moves from the value best suited for λ_S to the value best suited for λ_P as the value of F increases.

Serial and parallel parts have different values of λ in the real world. The design space, hence, becomes too large to visualize for all values of λ_S and λ_P. Therefore, λ_S and λ_P are set to two contrasting values of 0.01 and 0.2, respectively.

Fig. 9.2 shows how optimal thicknesses values of the specialized core x and the remaining efficient cores y change in the heterogeneous NCFET manycore $M_{x,y}$ with a change in the value of F for Rocket while other parameters are fixed. Thickness x gets optimized for λ_S when F is 0. The value of the thickness y is unimportant when F is 0 as non-specialized cores are always power-gated. Therefore, y could be of any value. Both x and y get optimized for λ_P when F is one as all cores work in parallel and no serial part exists. Both x and y are identical in this particular case. The value of y is always optimized for λ_P because it only executes the parallel fraction of the workload. The value of x for non-zero values of F gets optimized for both λ_S and λ_P because it executes both the serial and parallel fraction of the workload. However, the value of F determines the degree of optimization of x towards λ_S and λ_P.

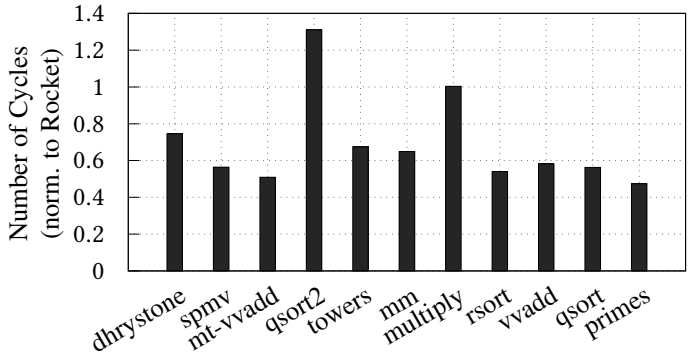


Figure 9.3.: Performance analysis of BOOM over Rocket processors by running a set of benchmarks. The BOOM outperforms the Rocket processor most of the time (i.e., < 1) due to the better microarchitecture design (i.e., out of order). Results do not consider that Rocket can operate at a higher frequency. Results are normalized to FinFET-based Rocket.

9.3. Quantitative Modeling of NCFET-based Manycore Designs

This section presents quantitative simulations of NCFET-based processors. First, the impact of microarchitectural heterogeneity is evaluated on the performance and energy of a single core because this is the current state-of-the-art approach when high performance and power efficiency are required and, therefore, serves as a baseline. These evaluations are performed using the RTL simulation setup described previously. Then, the simulations of multi-threaded applications are performed on manycore processors that ultimately show how NCFET-based designs improve power efficiency compared to heterogeneity in the microarchitecture.

9.3.1. Single-Core Exploration

Using the RTL simulator for BOOM and Rocket, several benchmarks taken from [120, 40] are examined in order to compare the performance and power-efficiency of the single-core BOOM and Rocket processors. Fig. 9.3 reports the number of cycles required to execute various benchmarks. Results are fully independent of technology as they do not yet capture the fact that different

microarchitectures can operate at different frequencies. Results show that BOOM needs fewer cycles for most benchmarks. The only exception is *qsort2*, which has few branches and has more data dependency between instructions, where Rocket needs fewer cycles due to the shorter pipeline. On average, BOOM requires >40% fewer cycles compared to Rocket. However, circuit-level analysis shows that this comes at the cost of around >70% more area and 110% more average-power for BOOM compared to Rocket when both are designed at FinFET.

The execution time not only depends on the number of cycles but also on the frequency. Circuit-level analyses of Rocket and BOOM at the nominal voltage (0.7V) are used to study the impact of technology on performance (i.e., execution time) and total energy consumption per application. The execution time is, therefore, the multiplication of delay per cycle by the number of cycles. The circuit-level analysis shows that, at FinFET, Rocket operates at a 32% higher frequency than BOOM due to lower hardware complexity. Despite the higher frequency of FinFET-based Rocket, BOOM still requires less execution time for most of the applications. Importantly, NCFET-based Rocket outperforms both FinFET-based Rocket and BOOM, as shown in Fig. 9.4(a). This performance enhancement comes at the cost of higher energy consumption when operating at the maximum frequency at the nominal voltage. The energy consumption is the average power per cycle multiplied by execution time. The energy consumption of NCFET Rocket is still lower than BOOM. Energy results per application are summarized in Fig. 9.4(b). Results in Fig. 9.3 and Fig. 9.4 are normalized to FinFET-based Rocket.

However, NCFET can result in less energy when operating at lower voltage while still achieving the same FinFET performance.

9.3.2. Manycore Exploration

This section investigates experimentally how the heterogeneous NCFET could yield similar, or better, performance and energy – without coping with different microarchitectures. First, several benchmarks are executed on the single in-order and out-of-order cores to verify the configurations of the system-level simulator and make sure it aligns with the RTL simulations. Fig. 9.5 shows the relative performance gain of FinFET BOOM over FinFET Rocket and NCFET Rocket over FinFET Rocket of two representative benchmarks. As shown, the

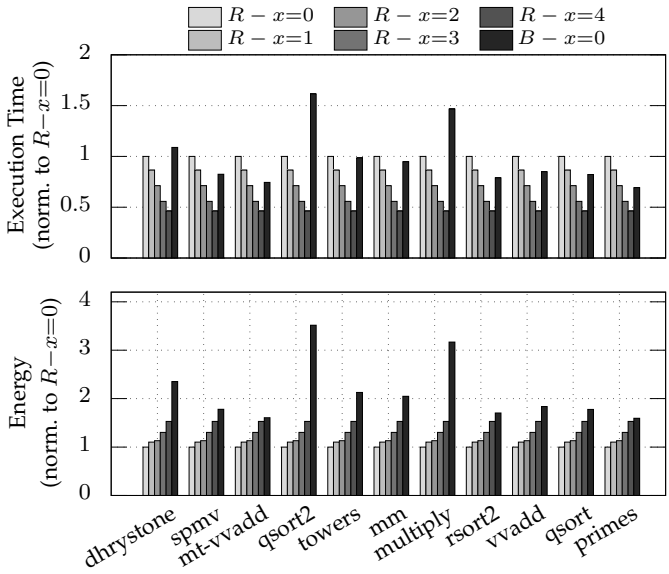


Figure 9.4.: (a) Execution time, and (b) energy consumption when running different benchmarks on Rocket (R) and BOOM (B) processors at different ferroelectric layer thicknesses x [129]. BOOM shows better performance than FinFET-based Rocket due to fewer execution cycles than Rocket for most applications, even though Rocket operates at a higher frequency. NCFET-based Rocket outperforms both FinFET-based Rocket and BOOM. This also comes with increases in energy considering the maximum possible frequency. Results are normalized to FinFET Rocket.

two simulators have a perfect matching as they report very similar performance gains. Hence, Sniper is correctly configured.

Two manycore processors are modeled with the same area. The first design implements the conventional heterogeneity at the level of microarchitecture. It combines one large out-of-order BOOM with four small in-order Rocket cores. The second does not employ microarchitectural heterogeneity and only uses small in-order Rocket cores. It uses six cores to occupy the same area as the first design. The difference with the conventional architectures is the implementation of NCFET. The ferroelectric layer thickness is a design-time choice that needs to be selected to suit many different applications at run-time. Analytical exploration reveals that the thickness ranges from 0.6nm to 2.4nm but converges towards 1nm for applications with high dynamic power consumption [129]. Therefore, x=1nm is selected.

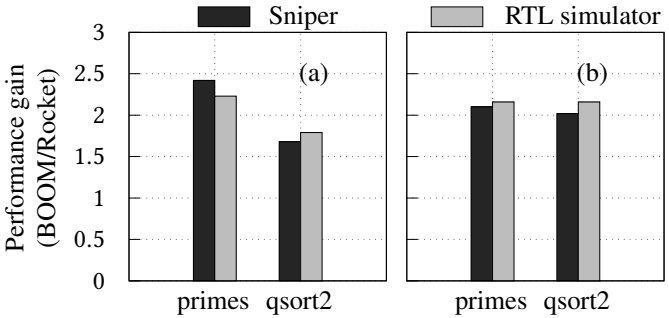


Figure 9.5.: The relative performance gain of (a) FinFET BOOM over FinFET Rocket and (b) NCFET Rocket ($R - x$=4) over FinFET Rocket of two benchmarks running on Sniper and RTL simulators. The two simulators show a perfect matching.

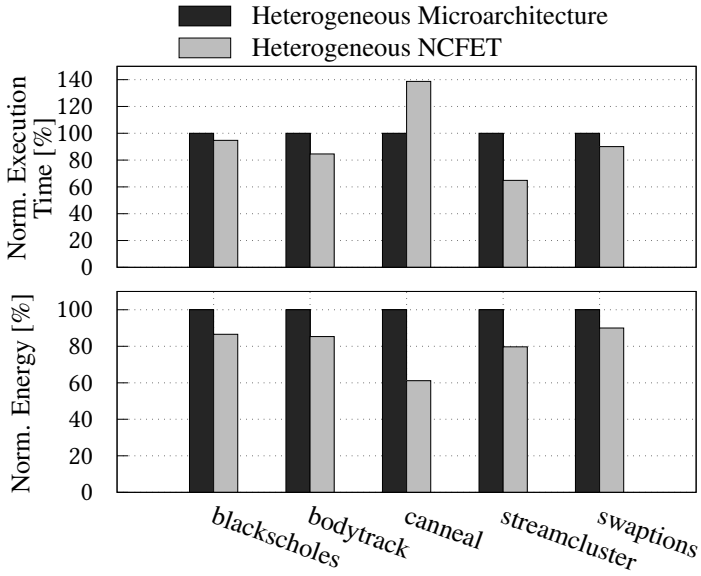


Figure 9.6.: Comparison of execution time (performance) and energy for different *PARSEC* benchmarks under different sources of heterogeneity. The proposed heterogeneity in the thickness in NCFET outperforms the conventional microarchitectural heterogeneity. It improves the performance on average by 8.3 % and energy by 20.2 %.

blackscholes, *bodytrack*, *canneal*, *streamcluster*, and *swaptions* are selected from *PARSEC* [22] to cover five and six threads. Idle cores are power-gated.

Fig. 9.6 shows the execution time (performance) and energy for different benchmarks on the two manycore processors. All benchmarks show higher performance with the NCFET-based manycore. The reasons are: (1) the slave threads dominate the performance of many benchmarks. (2) The NCFET-based manycore has one more core (six in total) than the manycore with heterogeneous microarchitecture (five in total). (3) NCFET cores run at a higher frequency. This, in turn, improves the overall performance of the benchmarks. The only exception is *canneal*. It has a low parallel fraction F and, therefore, its performance is dominated by the master thread. Furthermore, *canneal* is highly memory-bound. Consequently, it barely benefits from the increased frequency with NCFET. On average, the NCFET-based manycore increases performance by 8.3%.

Importantly, NCFET reduces the total energy of all benchmarks. This reduction is due to a combination of shorter execution time and avoiding the power-hungry out-of-order core and instead operating with power-efficient in-order cores. On average, the energy is reduced by 20.2% over conventional microarchitectural heterogeneity. This reduction demonstrates how NCFET increases the power efficiency of parallel execution of multi-threaded applications on manycore processors.

9.4. Summary and Conclusions

This chapter studied how NCFET affects the performance and power efficiency of the manycore designs through analytical and quantitative modeling. Manycore analytical modeling and comparative analysis are conducted through extending Amdahl's law to quantify the power efficiency gains in NCFET-based manycore designs. Analytical study based on the extended Amdahl's law demonstrated how the optimal ferroelectric layer's thickness(es) in NCFET depends on the employed microarchitecture and workload characteristics. This chapter introduced a novel heterogeneous NCFET manycore design that combines cores with different thicknesses of the ferroelectric layer. Unlike conventional heterogeneous FinFET manycore, the design improves the power efficiency without introducing microarchitecture heterogeneity.

10. Conclusions and discussion

10.1. Dissertation Conclusion

In the semiconductor industry, down-scaling transistor size is of great importance for driving innovations in an ever more digitalized and interconnected world. This will, in turn, serve to promote progress and the benefits of human society. Chip manufacturers have followed a continued trend of shrinking chips by making them smaller, denser, faster, more complex, and lower power per function as well as lower cost per circuit. These benefits, brought by smaller transistor sizes, can only be utilized if new technology nodes perform reliable computation (i.e., error-free results and failure-free operation). In recent decades, the rate of shrinking of chips has slowed down as many challenges accumulate. For instance, the supply and threshold voltages remain unchanged, performance and reliability are issues, and the power density keeps increasing are some of these challenges. Therefore, the current technology is hitting its fundamental limits where scaling further could be impossible, and many issues threaten the operation of modern semiconductor devices.

Designers are looking to optimize technology nodes instead of scaling. This indeed requires new techniques to sustain reliability, improve performance, and ultimately reduce power. This dissertation investigated the challenges of providing sufficiently accurate reliability estimation along with proposing new techniques to improve the reliability of circuits, avoiding pessimistic worst-case estimations (i.e., employing traditional timing guardband). Additionally, it provided comprehensive analyses on low power design under the Negative Capacitance Field-Effect Transistor (NCFET) as an emerging technology, making power and energy management techniques aware of its properties.

The presented works are in two parts. **Part-I** presented different techniques to sustain the reliability of embedded on-chip systems with respect to aging effects, self-heating effect, and IR-drop while significantly recovering most of the associated performance losses due to the traditional mitigation techniques. The presented techniques tackle different abstraction levels: from the device level to the circuit level and up to the system and multi-/many-core level. Evaluations of the proposed techniques were fully conducted through the implementations and simulations at the circuit level using the mature commercial chip design flow and system simulators as well. The contributions within this part are:
Chapter 3 explored the potential performance gains accompanied with maximizing energy efficiency when operating in Near-Threshold region, exactly at

Optimal Energy Point (OEP), by employing a multi-/many-core processor. It demonstrated how OEP is affected by optimization goals and running applications.
Chapter 4 revealed the hidden correlations between transistor aging and voltage fluctuation (IR-drop). It provided a novel approach to accurately estimate the required timing guardbands to sustain reliability under both phenomena for the whole lifetime of the circuit. Considering these correlations allows designers to estimate the smallest, yet sufficient, guardband under both phenomena.
Chapter 5 presented novel graceful-approximation technique that, over time, suppresses transistor aging effects. It provided a technique that is able to eliminate aging guardbands and the associated performance loss with the smallest, yet acceptable, accuracy degradation. It employed quantization technique to eliminate aging effects in circuits with tolerable accuracy.
Chapter 6 presented a novel technique to analyze the impacts of Self-Heating Effects (SHE) on both the timing and power of a full processor. It presented a novel technique to mitigate SHE by operating circuits near Zero-Temperature Coefficient (N-ZTC), showing that operating at N-ZTC minimizes SHE-induced variance in performance and power, showing how operating N-ZTC protects circuits against SHE without the need for timing guardbands.

Part-II provided comprehensive analyses on low power design under emerging technology. Negative Capacitance Field-Effect Transistor (NCFET) is an emerging technology that can go beyond voltage scaling limitations by operating at low voltages while featuring high performance. This part covered the unique properties of NCFET, made power and energy management techniques aware of its properties. Evaluations of the proposed techniques were fully conducted through the implementations and simulations at the circuit level using the mature commercial chip design flow and system simulators as well. The main contributions are as follows:
Chapter 7 modeled NCFET-based processors, demonstrating that voltage scaling leads to a novel runtime trade-off between leakage and dynamic power, resulting in an optimal point at which total power is minimized. It showed how in NCFET-based processor, the optimal voltage selection required to minimize the total power consumption is workload-dependent and follows the share of leakage power from total power.
Chapter 8 presented comprehensive power and energy models for NCFET-based processor. It showed how traditional DVS and DVFS techniques are suboptimal w.r.t. NCFET, as they are unaware of its properties. NCFET-aware DVS and DVFS models enable the exploration of the simultaneous impacts

of voltage, frequency, workload characteristics, and ferroelectric layer thickness on power and energy. It presented novel NCFET-aware DVS and DVFS algorithms for power and energy optimization by selecting the optimal voltage/frequency pair at runtime, considering the characteristics of the running workloads. It validated and quantified the effectiveness of these models against state-of-the-art, through analytical exploration, accurate circuit-level simulation, and system-level simulation considering synthetic and real workloads.
Chapter 9 presented a novel NCFET-based heterogeneous manycore design, in order to maximize the power and energy efficiency. Such design is able to eliminate the associated overheads due to different microarchitectures in conventional heterogeneous manycore design. To achieve heterogeneity, it proposed to optimally select the correct configurations by making one core as super-core while the remaining are power-efficient cores. It extended Amdhal's law covering the execution of several new system-specific and application-specific parameters to quantify the potential benefits of the new design.

10.2. Future Work

Many reliability issues, due to the technology scaling and fixed voltage, still need to be revisited in order to utilize the existing technology node. For instance, soft error and Random Telegraph Noise (RTN) are some of these issues. They are expected to be major issues in maintaining reliable operations in memory modules due to their strong ability to induce random errors.

Additionally, technology scaling has exceeded the tolerances of the semiconductor manufacturing process, resulting in ever-increasing process variations of transistors. Process variation becomes more critical for advanced technology nodes. As the industry continues to scale dimensions, its significance will continue to grow on performance and reliability. Hence, it must be considered in all circuit design and analyses covering all design levels.

On the other hand, this work implements NCFET-based memory, like caches, using flip-flop implementation. This is because of the absence of a memory compiler for NCFET. Important research can be based upon the memory compiler for NCFET and the accompanied reliability estimation. Reliability analysis could cover the scope of analyzing the SRAM cells reliability and other kinds of designs such as the typical 6-T SRAM and 8-T SRAM cells.

Bibliography

[1] A. Calimera and R. I. Bahar and E. Macii and M. Poncino. “Temperature-Insensitive Dual-V_{th} Synthesis for Nanometer CMOS Technologies Under Inverse Temperature Dependence”. In: *IEEE Transactions on Very Large Scale Integration Systems (TVLSI)* 18.11 (Nov. 2010), pp. 1608–1620. ISSN: 1063-8210. DOI: 10.1109/TVLSI.2009.2025884.

[2] I. Agbo et al. “Integral Impact of BTI, PVT Variation, and Workload on SRAM Sense Amplifier”. In: *Transactions on Very Large Scale Integration (TVLSI) Systems* 25.4 (Apr. 2017). ISSN: 1063-8210. DOI: 10.1109/TVLSI.2016.2643618.

[3] N. Ahmed, M. Tehranipoor, and V. Jayaram. “A Novel Framework for Faster-than-at-Speed Delay Test Considering IR-drop Effects”. In: *International Conference on Computer Aided Design (ICCAD)*. Nov. 2006. DOI: 10.1109/ICCAD.2006.320136.

[4] W. Ahn et al. “Integrated modeling of Self-heating of confined geometry (FinFET, NWFET, and NSHFET) transistors and its implications for the reliability of sub-20nm modern integrated circuits”. In: *Microelectronics Reliability (MR)* 81 (2018), pp. 262–273. ISSN: 0026-2714. DOI: 10.1016/j.microrel.2017.12.034.

[5] Gene M. Amdahl. “Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities”. In: *Spring Joint Computer Conference*. AFIPS ’67 (Spring) (1967), pp. 483–485. DOI: 10.1145/1465482.1465560.

[6] H. Amrouch et al. “Impact of Variability on Processor Performance in Negative Capacitance FinFET Technology”. In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 67.9 (2020), pp. 3127–3137. DOI: 10.1109/TCSI.2020.2990672.

[7] H. Amrouch et al. “Negative Capacitance Transistor to Address the Fundamental Limitations in Technology Scaling: Processor Performance”. In: *IEEE Access* 6 (2018), pp. 52754–52765. DOI: 10.1109/ACCESS.2018.2870916.

[8] H. Amrouch et al. “NPU Thermal Management”. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 39.11 (2020), pp. 3842–3855. DOI: 10.1109/TCAD.2020.3012753.

[9] H. Amrouch et al. “Reliability-aware design to suppress aging”. In: *Design Automation Conference (DAC)*. June 2016. DOI: 10.1145/2897937.2898082.

[10] H. Amrouch et al. “Towards aging-induced approximations”. In: *2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)*. 2017, pp. 1–6. DOI: 10.1145/3061639.3062331.

[11] Hussam Amrouch et al. “Impact of bti on dynamic and static power: From the physical to circuit level”. In: *Reliability Physics Symposium (IRPS), 2017 IEEE International*. IEEE. 2017, CR–3.

[12] Hussam Amrouch et al. “Reliability Challenges with Self-Heating and Aging in FinFET Technology”. In: July 2019, pp. 68–71. DOI: 10.1109/IOLTS.2019.8854405.

[13] Hussam Amrouch et al. “Reliability-aware design to suppress aging”. In: *Design Automation Conference (DAC), 2016 53nd ACM/EDAC/IEEE*. IEEE. 2016.

[14] Krste Asanović et al. *The Rocket Chip Generator*. Tech. rep. UCB/EECS-2016-17. EECS Department, University of California, Berkeley, Apr. 2016.

[15] Khaled Attia, Mostafa El-Hosseini, and Hesham Ali. “Dynamic power management techniques in multi-core architectures: A survey study”. In: *Ain Shams Engineering Journal* 8 (Oct. 2015). DOI: 10.1016/j.asej.2015.08.010.

[16] V. Axelrad et al. “Implementation of ESD Protection in SOI Technology: A Simulation Study”. In: *2005 International Conference On Simulation of Semiconductor Processes and Devices*. Sept. 2005, pp. 59–62. DOI: 10.1109/SISPAD.2005.201472.

[17] Bashayer M Al-Babtain et al. “A survey on Amdahl’s law extension in multicore architectures”. In: *International Journal of New Computer Architectures and their Applications (IJNCAA)* 3.3 (2013), pp. 30–46.

[18] Jonathan Balkind et al. “OpenPiton: An Open Source Manycore Research Framework”. In: *Int. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, 2016, pp. 217–232. ISBN: 978-1-4503-4091-5. DOI: 10.1145/2872362.2872414.

[19] Ron Banner, Yury Nahshan, and Daniel Soudry. “Post training 4-bit quantization of convolutional networks for rapid-deployment”. In: *NeurIPS*. 2019, pp. 7950–7958.

[20] Patrick Benediktsson, Jon A. Flandrin, and Chen Zheng. “Non Uniform On Chip Power Delivery Network Synthesis Methodology”. In: *Computing Research Repository (CoRR)* abs/1711.00425 (2017).

[21] B. Bhaskaran et al. “At-speed capture global noise reduction low-power memory test architecture”. In: *35th VLSI Test Symposium (VTS)*. Apr. 2017. DOI: 10.1109/VTS.2017.7928936.

[22] Christian Bienia et al. “The PARSEC Benchmark Suite: Characterization and Architectural Implications”. In: *Parallel Architectures and Compilation Techniques (PACT)*. 2008, pp. 72–81.

[23] David Blaauw et al. “Leakage current reduction in VLSI systems”. In: *Journal of Circuits, Systems, and Computers* 11 (Dec. 2002), pp. 621–636. DOI: 10.1142/S0218126602000665.

[24] M. Bohr. “A 30 Year Retrospective on Dennard’s MOSFET Scaling Paper”. In: *IEEE Solid-State Circuits Society Newsletter* 12.1 (2007), pp. 11–13. DOI: 10.1109/N-SSC.2007.4785534.

[25] B. Boroujerdian et al. “Trading Off Temperature Guardbands via Adaptive Approximations”. In: *2018 IEEE 36th International Conference on Computer Design (ICCD)*. 2018, pp. 202–209. DOI: 10.1109/ICCD.2018.00039.

[26] BSIM. *BSIM-CMG Technical Manual*. Oct. 2018. URL: %7Bhttp://www-device.eecs.berkeley.edu/bsim/?page=BSIMCMG%7D.

[27] Julian Bui, Chenguang Xu, and Sudhanva Gurumurthi. “Understanding Performance Issues on both Single Core and Multi-core Architecture ”. In: *Computer Organizatione*. MICRO 32. ACM, 2007, pp. 2–. DOI: 1-58113-000-0/00/0007.

[28] Edward Burton et al. “FIVR—Fully integrated voltage regulators on 4th generation Intel® Core™ SoCs”. In: *Applied Power Electronics Conference and Exposition (APEC)*. 2014. DOI: 10.1109/APEC.2014.6803344.

[29] *Cadence EDA tool flows*. https://www.cadence.com/. Oct. 2018.

[30] Enrico Calore et al. “Software and DVFS Tuning for Performance and Energy-Efficiency on Intel KNL Processors”. In: *Journal of Low Power Electronics and Applications* 8 (June 2018), p. 18. DOI: 10.3390/jlpea8020018.

[31] Trevor E Carlson, Wim Heirman, and Lieven Eeckhout. “Sniper: Exploring the Level of Abstraction for Scalable and Accurate Parallel Multi-Core Simulation”. In: *Int. Conf. for High Performance Computing, Networking, Storage and Analysis (SC)*. ACM. 2011, p. 52. DOI: 10.1145/2063384.2063454.

[32] Stephen Cass. “Taking AI to the edge: Google’s TPU now comes in a maker-friendly package”. In: *IEEE Spectrum* 56.5 (2019), pp. 16–17. DOI: 10.1109/MSPEC.2019.8701189.

[33] Christopher Celio et al. *BOOM v2: an open-source out-of-order RISC-V core*. Tech. rep. UCB/EECS-2017-157. EECS Department, University of California, Berkeley, Sept. 2017. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-157.html.

[34] V. Chandra and R. Aitken. “Impact of Technology and Voltage Scaling on the Soft Error Susceptibility in Nanoscale CMOS”. In: *2008 IEEE International Symposium on Defect and Fault Tolerance of VLSI Systems*. 2008, pp. 114–122. DOI: 10.1109/DFT.2008.50.

[35] Charles, J. and Jassi, P. and Ananth, N.S. and Sadat, A. and Fedorova, A. “Evaluation of the Intel Core i7 Turbo Boost feature”. In: *IEEE International Symposium on Workload Characterization (IISWC)*. 2009. DOI: 10.1109/IISWC.2009.5306782.

[36] Y.S. Chauhan et al. “BSIM - Industry standard compact MOSFET models”. In: *ESSCIRC*. 2012.

[37] Yogesh Singh Chauhan et al. “FinFET Modeling for IC Simulation and Design”. In: Oxford: Academic Press, 2015. ISBN: 978-0-12-420031-9. DOI: https://doi.org/10.1016/B978-0-12-420031-9.00012-9.

[38] I. Ciofi et al. “Impact of Wire Geometry on Interconnect RC and Circuit Delay”. In: *IEEE Transactions on Electron Devices* 63.6 (2016), pp. 2488–2496. DOI: 10.1109/TED.2016.2554561.

[39] Lawrence T. Clark et al. “ASAP7: A 7-nm finFET predictive process design kit”. In: *Microelectronics Journal (MJ)* 53 (2016). ISSN: 0026-2692. DOI: 10.1016/j.mejo.2016.04.006.

[40] Michael Clark and Bruce Hoult. “rv8: a high performance RISC-V to x86 binary translator”. In: (Oct. 2017). DOI: 10.13140/RG.2.2.30957.69601.

[41] D. Bol and C. Hocquet and D. Flandre and J. Legat. “The detrimental impact of negative Celsius temperature on ultra-low-voltage CMOS logic”. In: *European Solid-State Circuits Conference (ESSCIRC)*. Sept. 2010, pp. 522–525. DOI: 10.1109/ESSCIRC.2010.5619758.

[42] D. Jang and E. Bury and R. Ritzenthaler and M. G. Bardon and T. Chiarella and K. Miyaguchi and P. Raghavan and A. Mocuta and G. Groeseneken and A. Mercha and D. Verkest and A. Thean. “Self-heating on bulk FinFET from 14nm down to 7nm node”. In: *International Electron Devices Meeting (IEDM)*. Dec. 2015, pp. 11.6.1–11.6.4. DOI: 10.1109/IEDM.2015.7409678.

[43] V. De, S. Vangal, and R. Krishnamurthy. “Near Threshold Voltage (NTV) Computing: Computing in the Dark Silicon Era”. In: *IEEE Design Test* 34.2 (Apr. 2017), pp. 24–30.

[44] J. Deng et al. “ImageNet: A large-scale hierarchical image database”. In: *IEEE Conf. on Comp. Vis. and Pat. Recogn. (CVPR)*. June 2009, pp. 248–255.

[45] S. Dighe et al. “Within-Die Variation-Aware Dynamic-Voltage-Frequency-Scaling With Optimal Core Allocation and Thread Hopping for the 80-Core TeraFLOPS Processor”. In: *IEEE Journal of Solid-State Circuits* 46.1 (Jan. 2011), pp. 184–193. ISSN: 0018-9200. DOI: 10.1109/JSSC.2010.2080550.

[46] B Dilip, Surya Prasad, and R Bhavani. “LEAKAGE POWER REDUCTION IN CMOS CIRCUITS USING LEAKAGE CONTROL TRANSISTOR TECHNIQUE IN NANOSCALE TECHNOLOGY”. In: *International Journal of Electronics Signals and Systems (IJESS) ISSN: 2231- 5969, Vol-2 Iss-1, 2012* (Jan. 2012). DOI: 10.47893/IJESS.2013.1113.

[47] M. Eireiner et al. “Timing violations due to V_{DD}/V_{SS} bounce”. In: *Advances in Radio Science* 4 (Sept. 2006). DOI: https://doi.org/10.5194/ars-4-197-2006.

[48] F. Firouzi, S. Kiamehr, and M. B. Tahoori. “Statistical analysis of BTI in the presence of process-induced voltage and temperature variations”. In: *Asia and South Pacific Design Automation Conference (ASP-DAC)*. Jan. 2013. DOI: 10.1109/ASPDAC.2013.6509663.

[49] Andrea Ghetti. “Gate Oxide Reliability: Physical and Computational Models”. In: (Jan. 2004). DOI: 10.1007/978-3-662-09432-7_6.

[50] N. Goel et al. “A comprehensive modeling framework for gate stack process dependence of DC and AC NBTI in SiON and HKMG p-MOSFETs”. In: *Microelectronics Reliability (j.microrel)* (2014). DOI: 10.1016/j.microrel.2013.12.017.

[51] S. Goswami, B. Chowdhury, and M. Chanda. “Analytical modelling of power dissipation and voltage swing of CMOS logic circuit for near-threshold computing”. In: *Devices for Integrated Circuit (DevIC)*. Mar. 2017, pp. 658–663. DOI: 10.1109/DEVIC.2017.8074032.

[52] T. Grasser et al. “Advanced characterization of oxide traps: The dynamic time-dependent defect spectroscopy”. In: *International Reliability Physics Symposium (IRPS)*. Apr. 2013. DOI: 10.1109/IRPS.2013.6531957.

[53] H. Amrouch *et al.* “Unveiling the Impact of IR-Drop on Performance Gain in NCFET-Based Processors”. In: *IEEE Transactions on Electron Devices* 66.7 (2019), pp. 3215–3223. DOI: 10.1109/TED.2019.2916494.

[54] H. Jiang and S. Shin and X. Liu and X. Zhang and M. A. Alam. “The Impact of Self-Heating on HCI Reliability in High-Performance Digital Circuits”. In: *Electron Device Letters (EDL)* 38.4 (Apr. 2017), pp. 430–433. ISSN: 0741-3106. DOI: 10.1109/LED.2017.2674658.

[55] Jawad Haj-Yahya et al. *Energy Efficient High Performance Processors Recent Approaches for Designing Green High Performance Computing*. Apr. 2018. ISBN: 978-981-10-8554-3.

[56] Nor Zaidi Haron and Said Hamdioui. "Why is CMOS scaling coming to an END?" In: *2008 3rd International Design and Test Workshop*. 2008, pp. 98–103. DOI: 10.1109/IDT.2008.4802475.

[57] K. He et al. "Deep Residual Learning for Image Recognition". In: *IEEE Conf. on Comp. Vis. and Pat. Recogn. (CVPR)*. June 2016, pp. 770–778.

[58] M. D. Hill and M. R. Marty. "Amdahl's Law in the Multicore Era". In: *Computer* 41.7 (July 2008), pp. 33–38. ISSN: 0018-9162. DOI: 10.1109/MC.2008.209.

[59] Michael Hoffmann et al. "Unveiling the double-well energy landscape in a ferroelectric layer". In: *Nature* 565.7740 (2019), p. 464. DOI: 10.1038/s41586-018-0854-z.

[60] Chenming Hu. *Modern semiconductor devices for integrated circuits*. Prentice Hall, 2010. ISBN: 9780136085256.

[61] DS Huang et al. "Comprehensive device and product level reliability studies on advanced CMOS technologies featuring 7nm high-k metal gate FinFET transistors". In: *Reliability Physics Symposium (IRPS), 2018 IEEE International*. IEEE. 2018, 6F–7.

[62] Forrest N. Iandola et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size". In: *CoRR* abs/1602.07360 (2016). arXiv: 1602.07360. URL: http://arxiv.org/abs/1602.07360.

[63] Mitsuhiko Igarashi et al. "A 28 nm High-k/MG Heterogeneous Multi-Core Mobile Application Processor With 2 GHz Cores and Low-Power 1 GHz Cores". In: *JSSC* 50 (Jan. 2015), pp. 92–101. DOI: 10.1109/JSSC.2014.2347353.

[64] H. Iwai. "Future of CMOS technology". In: *2004 Semiconductor Manufacturing Technology Workshop Proceedings (IEEE Cat. No.04EX846)*. 2004, pp. 5–17. DOI: 10.1109/SMTW.2004.1393699.

[65] Benoit Jacob et al. "Quantization and training of neural networks for efficient integer-arithmetic-only inference". In: *CVPR*. 2018, pp. 2704–2713.

[66] S. Jain et al. “A 280mV-to-1.2V wide-operating-range IA-32 processor in 32nm CMOS”. In: *2012 IEEE International Solid-State Circuits Conference*. Feb. 2012, pp. 66–68.

[67] Raj Jammy. “Life beyond Si: More Moore or More than Moore?” In: *2010 IEEE International Integrated Reliability Workshop Final Report*. 2010, pp. ix–ix. DOI: 10.1109/IIRW.2010.5706469.

[68] Jang, Doyoung and Bury, Erik and Ritzenthaler, Romain and Bardon, M Garcia and Chiarella, Thomas and Miyaguchi, Kenichi and Raghavan, Praveen and Mocuta, Anda and Groeseneken, Guido and Mercha, Abdelkarim and others. “Self-heating on bulk FinFET from 14nm down to 7nm node”. In: *International Electron Devices Meeting (IEDM)*. IEEE. 2015, pp. 11–6. DOI: 10.1109/IEDM.2015.7409678.

[69] Norman Jouppi et al. “In-Datacenter Performance Analysis of a Tensor Processing Unit”. In: June 2017, pp. 1–12. DOI: 10.1145/3079856.3080246.

[70] K. Kanda and K. Nose and H. Kawaguchi and T. Sakurai. “Design impact of positive temperature dependence on drain current in sub-1-V CMOS VLSIs”. In: *IEEE Journal of Solid-State Circuits* 36.10 (Oct. 2001), pp. 1559–1564. ISSN: 0018-9200. DOI: 10.1109/4.953485.

[71] Z. Kamal, Q. Hassan, and Z. Mouhcine. “Full on chip capacitance pmos low dropout voltage regulator”. In: *2011 International Conference on Multimedia Computing and Systems*. Apr. 2011. DOI: 10.1109/ICMCS.2011.5945660.

[72] Himanshu Kaul et al. “A 320 mv 56 μw 411 gops/watt ultra-low voltage motion estimation accelerator in 65 nm cmos”. In: *IEEE Journal of Solid-State Circuits* 44.1 (2009), pp. 107–114.

[73] J. Keane and C. H. Kim. “Transistor aging”. In: *IEEE Spectr*. Vol. 48. Apr. 2011, pp. 28–33.

[74] B. Keller et al. “Sub-microsecond adaptive voltage scaling in a 28nm FD-SOI processor SoC”. In: *European Solid-State Circuits Conference (ESSCIRC)*. Sept. 2016, pp. 269–272. DOI: 10.1109/ESSCIRC.2016.7598294.

[75] H. Kim et al. “Aging Compensation With Dynamic Computation Approximation”. In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 67.4 (2020), pp. 1319–1332. DOI: 10.1109/TCSI.2020.2969462.

[76] R. Koh and T. Iizuka. “Parameter Extraction and Comparison of Self-Heating Models for Power MOSFETs Based on Transient Current Measurements”. In: *Transactions on Electron Devices (TED)* 60.2 (Feb. 2013), pp. 708–713. ISSN: 0018-9383. DOI: 10.1109/TED.2012.2226727.

[77] Raghuraman Krishnamoorthi. “Quantizing deep convolutional networks for efficient inference: A whitepaper”. In: *arXiv preprint arXiv:1806.08342* (2018).

[78] Zoran Krivokapic et al. “14nm Ferroelectric FinFET Technology with Steep Subthreshold Slope for Ultra Low Power Applications”. In: *IEEE Int. Electron Devices Meeting (IEDM)*. Dec. 2017, pp. 15.1.1–15.1.4. DOI: 10.1109/IEDM.2017.8268393.

[79] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Neural Information Processing Systems* 25 (Jan. 2012). DOI: 10.1145/3065386.

[80] Rakesh Kumar et al. “Single-ISA heterogeneous multi-core architectures: The potential for processor power reduction”. In: *Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture*. IEEE Computer Society. 2003, p. 81.

[81] Kumar, Ulayil Sajesh and Rao, Valipe Ramgopal. “A thermal-aware device design considerations for nanoscale SOI and bulk FinFETs”. In: *Transactions on Electron Devices (TED)* 63.1 (2016), pp. 280–287.

[82] L. D. Landau and I. M. Khalatnikov. “On the Anomalous Absorption of Sound near a Second Order Phase Transition Point”. In: *Dokladii Academii Nauk* 96 (1954), pp. 469–472.

[83] M. H. Lee et al. “Extremely Steep Switch of Negative-Capacitance Nanosheet GAA-FETs and FinFETs”. In: *2018 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2018, pp. 31–8.

[84] Sheng Li et al. “The McPAT Framework for Multicore and Manycore Architectures: Simultaneously Modeling Power, Area, and Timing”. In: *ACM Transactions on Architecture and Code Optimization (TACO)* 10 (Apr. 2013). DOI: 10.1145/2445572.2445577.

[85] D. Lorenz, M. Barke, and U. Schlichtmann. “Aging analysis at gate and macro cell level”. In: *International Conference on Computer-Aided Design (ICCAD)*. Nov. 2010. DOI: 10.1109/ICCAD.2010.5654309.

[86] M. Cho and M. Khellah and K. Chae and K. Ahmed and J. Tschanz and S. Mukhopadhyay. “Characterization of Inverse Temperature Dependence in logic circuits”. In: *Custom Integrated Circuits Conference (CICC)*. Sept. 2012, pp. 1–4. DOI: 10.1109/CICC.2012.6330659.

[87] S. Mahapatra et al. “A Comparative Study of Different Physics-Based NBTI Models”. In: *Transactions on Electron Devices (T-ED)* 60.3 (Mar. 2013). ISSN: 0018-9383. DOI: 10.1109/TED.2013.2238237.

[88] P. Meinerzhagen et al. “An energy-efficient graphics processor featuring fine-grain DVFS with integrated voltage regulators, execution-unit turbo, and retentive sleep in 14nm tri-gate CMOS”. In: *Solid - State Circuits Conference - (ISSCC)*. Feb. 2018, pp. 38–40. DOI: 10.1109/ISSCC.2018.8310172.

[89] S. Mishra et al. “Predictive TCAD for NBTI stress-recovery in various device architectures and channel materials”. In: *International Reliability Physics Symposium (IRPS)*. Apr. 2017. DOI: 10.1109/IRPS.2017.7936335.

[90] K. Mistry et al. “A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging”. In: *2007 IEEE International Electron Devices Meeting*. 2007, pp. 247–250. DOI: 10.1109/IEDM.2007.4418914.

[91] L. Mo, A. Kritikakou, and O. Sentieys. “Energy-Quality-Time Optimized Task Mapping on DVFS-Enabled Multicores”. In: vol. 37. 11. Nov. 2018. DOI: 10.1109/TCAD.2018.2857300.

[92] V. Mrazek et al. “ALWANN: Automatic Layer-Wise Approximation of Deep Neural Network Accelerators without Retraining”. In: *International Conference on Computer-Aided Design (ICCAD)*. Nov. 2019, pp. 1–8. DOI: 10.1109/ICCAD45719.2019.8942068.

[93] Stefan Mueller et al. “Incipient Ferroelectricity in Al-Doped HfO_2 Thin Films”. In: *Advanced Functional Materials* 22.11 (2012), pp. 2412–2417. ISSN: 1616-3028. DOI: 10.1002/adfm.201103119.

[94] Yury Nahshan et al. “Loss Aware Post-training Quantization”. In: *arXiv preprint arXiv:1911.07190* (2019).

[95] NanGate. *Open Cell Library*. https://www.silvaco.com/.

[96] Siva G. Narendra. “Challenges and Design Choices in Nanoscale CMOS”. In: *J. Emerg. Technol. Comput. Syst.* 1.1 (Mar. 2005), pp. 7–49. ISSN: 1550-4832. DOI: 10.1145/1063803.1063805. URL: https://doi.org/10.1145/1063803.1063805.

[97] S. Natarajan et al. “A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm2 SRAM cell size”. In: *2014 IEEE International Electron Devices Meeting*. 2014, pp. 3.7.1–3.7.3.

[98] Rajeev Pankaj Nelapati and Sivasankaran K. “Impact of self-heating effect on the performance of hybrid FinFET”. In: *Microelectronics Journal (MJ)* 76 (2018), pp. 63–68. ISSN: 0026-2692. DOI: 10.1016/j.mejo.2018.04.015.

[99] S. K. Nithin, G. Shanmugam, and S. Chandrasekar. “Dynamic voltage (IR) drop analysis and design closure: Issues and challenges”. In: *11th International Symposium on Quality Electronic Design (ISQED)*. Mar. 2010. DOI: 10.1109/ISQED.2010.5450515.

[100] Y. Ogasahara et al. “Validation of a Full-Chip Simulation Model for Supply Noise and Delay Dependence on Average Voltage Drop With On-Chip Delay Measurement”. In: *Transactions on Circuits and Systems II: Express Briefs (TCAS-II)* 54.10 (Oct. 2007). ISSN: 1549-7747. DOI: 10.1109/TCSII.2007.901574.

[101] *OpenSPARC T1*. Feb. 2019. URL: https://www.oracle.com/technetwork/systems/opensparc/opensparc-t1-page-1444609.html.

[102] G. Pahwa, A. Agarwal, and Y. S. Chauhan. “Numerical Investigation of Short-Channel Effects in Negative Capacitance MFIS and MFMIS Transistors: Subthreshold Behavior”. In: *IEEE Transactions on Electron Devices (TED)* 65.11 (Nov. 2018), pp. 5130–5136. ISSN: 0018-9383. DOI: 10.1109/TED.2018.2870519.

[103] G. Pahwa et al. “Analysis and Compact Modeling of Negative Capacitance Transistor with High *ON*-Current and Negative Output Differential Resistance – Part II: Model Validation”. In: *IEEE Transactions on Electron Devices* 63.12 (Dec. 2016), pp. 4986–4992. DOI: 10.1109/TED.2016.2614436.

[104] G. Pahwa et al. “Physical Insights on Negative Capacitance Transistors in Nonhysteresis and Hysteresis Regimes: MFMIS Versus MFIS Structures”. In: *IEEE Transactions on Electron Devices* 65.3 (2018), pp. 867–873. DOI: 10.1109/TED.2018.2794499.

[105] Girish Pahwa et al. “Designing Energy Efficient and Hysteresis Free Negative Capacitance FinFET with Negative DIBL and 3.5 XI ON Using Compact Modeling Approach”. In: *European Solid-State Circuits Conference (ESSCIRC)*. 2016, pp. 49–54.

[106] N. Parihar et al. “BTI Analysis Tool—Modeling of NBTI DC, AC Stress and Recovery Time Kinetics, Nitrogen Impact, and EOL Estimation”. In: *Transactions on Electron Devices (T-ED)* 65.2 (Feb. 2018). ISSN: 0018-9383. DOI: 10.1109/TED.2017.2780083.

[107] Anuj Pathania and Jörg Henkel. “HotSniper: Sniper-Based Toolchain for Many-Core Thermal Simulations in Open Systems”. In: *Embedded Systems Letters (ESL)* (2018).

[108] N. Pinckney et al. “Assessing the performance limits of parallelized near-threshold computing”. In: *Design Automation Conference (DAC), 2012*. June 2012, pp. 1143–1148.

[109] N. Pinckney et al. “Near-threshold computing in FinFET technologies: Opportunities for improved voltage scalability”. In: *Design Automation Conference (DAC), 2016*. June 2016, pp. 1–6.

[110] E. Pop, R. Dutton, and K. Goodson. “Thermal analysis of ultra-thin body device scaling [SOI and FinFet devices]”. In: *IEEE International Electron Devices Meeting 2003*. 2003, pp. 36.6.1–36.6.4. DOI: 10.1109/IEDM.2003.1269420.

[111] *Predictive Technology Model*. http://ptm.asu.edu/.

[112] Pytorch. *Torchvision models, https://pytorch.org/docs/stable/torchvision/models.html*.

[113] Karin M Rabe et al. “Modern physics of ferroelectrics: Essential background”. In: *Physics of Ferroelectrics*. Springer, 2007, pp. 1–30.

[114] A. Rahman et al. “Reliability studies of a 10nm high-performance and low-power CMOS technology featuring 3rd generation FinFET and 5th generation HK/MG”. In: *International Reliability Physics Symposium (IRPS)*. Mar. 2018. DOI: 10.1109/IRPS.2018.8353648.

[115] M. Rapp et al. “Performance, Power and Cooling Trade-Offs with NCFET-based Many-Cores”. In: *Design Automation Conference (DAC)* (2019).

[116] H. Reisinger et al. “A Comparison of Fast Methods for Measuring NBTI Degradation”. In: *Transactions on Device and Materials Reliability (TDMR)* 7.4 (Dec. 2007). ISSN: 1530-4388. DOI: 10.1109/TDMR.2007.911385.

[117] H. Reisinger et al. “Analysis of NBTI Degradation- and Recovery-Behavior Based on Ultra Fast VT-Measurements”. In: *International Reliability Physics Symposium (RELPHY)*. Mar. 2006. DOI: 10.1109/RELPHY.2006.251260.

[118] Chen Da-Ren, Chen Young-Long, and Chen You-Shyang. “Time and Energy Efficient DVS Scheduling for Real-Time Pinwheel Tasks”. In: *Journal of Applied Research and Technology* 12.6 (2014), pp. 1025–1039. ISSN: 1665-6423. DOI: https://doi.org/10.1016/S1665-6423(14)71663-3. URL: https://www.sciencedirect.com/science/article/pii/S1665642314716633.

[119] Eli Ringdalen and Merete Tangstad. “Softening and Melting of SiO2, an Important Parameter for Reactions with Quartz in Si Production”. In: *The 10th International Conference on Molten Slags, Fluxes and Salts 2016*. Jan. 2016. ISBN: 978-3-319-48625-3. DOI: 10.1007/978-3-319-48769-4_4.

[120] *RISC-V benchmark-tests*. 2019. URL: https://github.com/riscv/riscv-tests/tree/master/benchmarks.

[121] S. Roy et al. “OSFA: A New Paradigm of Aging Aware Gate-Sizing for Power/Performance Optimizations Under Multiple Operating Conditions”. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 35.10 (2016), pp. 1618–1629. DOI: 10.1109/TCAD.2016.2523439.

[122] Sayeef Salahuddin and Supriyo Datta. “Use of Negative Capacitance to Provide Voltage Amplification for Low Power Nanoscale Devices”. In: *Nano Letters* 8.2 (2008), pp. 405–410. DOI: 10.1021/nl071804g.

[123] S. Salamin, H. Amrouch, and J. Henkel. “Selecting the Optimal Energy Point in Near-Threshold Computing”. In: *Design, Automation Test in Europe Conference Exhibition (DATE)*. Mar. 2019, pp. 1691–1696. DOI: 10.23919/DATE.2019.8715211.

[124] S. Salamin et al. “Dynamic Power and Energy Management for NCFET-Based Processors”. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 39.11 (2020), pp. 3361–3372. DOI: 10.1109/TCAD.2020.3012644.

[125] S. Salamin et al. “Energy Optimization in NCFET-based Processors”. In: *2020 Design, Automation Test in Europe Conference Exhibition (DATE)*. Mar. 2020.

[126] S. Salamin et al. “Minimizing Excess Timing Guard Banding under Transistor Self-Heating through biasing at Zero-Temperature Coefficient”. In: *IEEE Access* 9 (2021), pp. 30687–30697. DOI: 10.1109/ACCESS.2021.3057900.

[127] S. Salamin et al. “Modeling the Interdependences Between Voltage Fluctuation and BTI Aging”. In: *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 27 (Mar. 2019), pp. 1–14. DOI: 10.1109/TVLSI.2019.2899890.

[128] S. Salamin et al. “NCFET-Aware Voltage Scaling”. In: *The International Symposium on Low Power Electronics and Design (ISLPED)* (2019).

[129] S. Salamin et al. “Power-Efficient Heterogeneous Many-Core Design with NCFET Technology”. In: *IEEE Transactions on Computers* (2020), pp. 1–1. DOI: 10.1109/TC.2020.3013567.

[130] S. Salamin et al. “Reliability-Aware Quantization for Anti-Aging NPUs”. In: *2021 Design, Automation Test in Europe Conference Exhibition (DATE)*. Mar. 2021.

[131] S. K. Samal et al. “Full chip power benefits with negative capacitance FETs”. In: *2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. July 2017. DOI: 10.1109/ISLPED.2017.8009170.

[132] M. Sandler et al. “MobileNetV2: Inverted Residuals and Linear Bottlenecks”. In: *IEEE/CVF Conf. on Comp. Vis. and Pat. Recogn. (CVPR)*. June 2018, pp. 4510–4520.

[133] V. M. van Santen et al. “Aging-aware voltage scaling”. In: *Design, Automation Test in Europe Conference Exhibition (DATE)*. Mar. 2016. DOI: 10.3850/9783981537079/0751.

[134] Victor Santen et al. “Reliability in Super- and Near-Threshold Computing: A Unified Model of RTN, BTI and PV”. In: *Circuits and Systems I: Regular Papers, IEEE Transactions on* PP (June 2017). DOI: 10.1109/TCSI.2017.2717790.

[135] Victor M. van Santen et al. “Impact of Self-Heating on Performance, Power and Reliability in FinFET Technology”. In: *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*. 2020, pp. 68–73. DOI: 10.1109/ASP-DAC47756.2020.9045582.

[136] S. Satapathy et al. “A revolving reference odometer circuit for BTI-induced frequency fluctuation measurements under fast DVFS transients”. In: *International Reliability Physics Symposium (IRPS)*. Apr. 2015. DOI: 10.1109/IRPS.2015.7112757.

[137] Zeinab Seifoori et al. “Introduction to Emerging SRAM-Based FPGA Architectures in Dark Silicon Era”. In: Jan. 2018. DOI: 10.1016/bs.adcom.2018.04.002.

[138] Mengwei Si et al. “Steep-Slope Hysteresis-Free Negative Capacitance MoS 2 Transistors”. In: *Nature Nanotechnology* 13.1 (2018), p. 24.

[139] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *arXiv 1409.1556* (Sept. 2014).

[140] Leonardo Bandeira Soares et al. “Near-threshold computing for very wide frequency scaling: Approximate adders to rescue performance”. In: *New Circuits and Systems Conference (NEWCAS), 2015 IEEE 13th International*. IEEE. 2015, pp. 1–4.

[141] P. M. Solomon. “Device Innovation and Material Challenges at the Limits of CMOS Technology”. In: *Annual Review of Materials Science* 30.1 (2000), pp. 681–697. DOI: 10.1146/annurev.matsci.30.1.681. eprint: https://doi.org/10.1146/annurev.matsci.30.1.681. URL: https://doi.org/10.1146/annurev.matsci.30.1.681.

[142] Thannirmalai Somu Muthukaruppan, Anuj Pathania, and Tulika Mitra. “Price theory based power management for heterogeneous multi-cores”. In: *ACM SIGPLAN Notices* 49.4 (2014), pp. 161–176.

[143] J. Song et al. “7.1 An 11.5TOPS/W 1024-MAC Butterfly Structure Dual-Core Sparsity-Aware Neural Processing Unit in 8nm Flagship Mobile SoC”. In: *Int. Solid-State Circuits Conf.* Feb. 2019, pp. 130–132.

[144] W. Sootkaneung, S. Howimanporn, and S. Chookaew. “Thermal Effect on Performance, Power, and BTI Aging in FinFET-Based Designs”. In: *DATE*. 2017, pp. 345–351. DOI: 10.1109/DSD.2017.35.

[145] *Static-and-dynamic-power-ir-analysis - Cadence forum*. https://community.cadence.com/cadence_technology_forums/f/digital-implementation/1061/static-and-dynamic-power-ir-analysis. Oct. 2020.

[146] Aaron Stillmaker and Bevan Baas. “Scaling equations for the accurate prediction of CMOS device performance from 180nm to 7nm”. In: *Integration* 58 (2017), pp. 74–81. ISSN: 0167-9260. DOI: https://doi.org/10.1016/j.vlsi.2017.02.002. URL: http://www.sciencedirect.com/science/article/pii/S0167926017300755.

[147] Vaidy Subramanian et al. “Planar bulk MOSFETs versus FinFETs: an analog/RF perspective”. In: *Transactions on Electron Devices (TED)* 53 (Jan. 2007), pp. 3071–3079. ISSN: 1557-9646. DOI: 10.1109/TED.2006.885649.

[148] Xian-He Sun and Yong Chen. “Reevaluating Amdahl’s Law in the Multicore Era”. In: *J. Parallel Distrib. Comput.* 70.2 (Feb. 2010), pp. 183–188. ISSN: 0743-7315. DOI: 10.1016/j.jpdc.2009.05.002.

[149] B. Swahn and Soha Hassoun. “Gate sizing: finFETs vs 32nm bulk MOSFETs”. In: *2006 43rd ACM/IEEE Design Automation Conference*. 2006, pp. 528–531. DOI: 10.1145/1146909.1147047.

[150] Synopsys. *Designing Advanced ASIC’s with Synopsys Design Tool Suite*. https://www.synopsys.com/apps/docs/pdfs/sps/rhet_rick_102407.pdf.

[151] Synopsys. *Synopsys Educational Design PDK*. Tech. rep. Oct. 2020. URL: http://web.engr.oregonstate.edu/~traylor/ece474/reading/SAED%5C_Cell%5C_Lib%5C_Rev1%5C_4%5C_20%5C_1.pdf.

[152] *Synopsys EDA Tool Flows*. https://www.synopsys.com/. Oct. 2021.

[153] *Synopsys HSPICE simulator*. Oct. 2018. URL: %7Bhttps://www.synopsys.com/verification/ams-verification/circuit-simulation/hspice.html%7D.

[154] *Synopsys SiliconSmart for cell library characterization*. Oct. 2018. URL: %7Bhttps://www.synopsys.com/implementation-and-signoff/signoff/siliconsmart.html%7D.

[155] C. Szegedy et al. "Going deeper with convolutions". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9.

[156] T. Takahashi et al. "Thermal-aware device design of nanoscale bulk/-SOI FinFETs: Suppression of operation temperature and its variability". In: *International Electron Devices Meeting (IEDM)*. Dec. 2011, pp. 34.6.1–34.6.4. DOI: 10.1109/IEDM.2011.6131672.

[157] M. Tan et al. "MnasNet: Platform-Aware Neural Architecture Search for Mobile". In: *IEEE/CVF Conf. on Comp. Vis. and Pat. Recogn. (CVPR)*. June 2019, pp. 2815–2823.

[158] Z. -G. Tasoulas et al. "Weight-Oriented Approximation for Energy-Efficient Neural Network Inference Accelerators". In: *IEEE Transactions on Circuits and Systems I: Regular Papers* 67.12 (2020), pp. 4670–4683. DOI: 10.1109/TCSI.2020.3019460.

[159] *The Rocket Chip Generator*. http://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-17.html.

[160] Andreas Traber et al. "PULPino: A small single-core RISC-V SoC". In: *3rd RISCV Workshop*. 2016. URL: %7Bhttp://iis-projects.ee.ethz.ch/index.php/PULP%7D.

[161] N. Tripathi, Amit Bhosle, and Ajit Pal. "Optimal assignment of high threshold voltage for synthesizing dualthreshold CMOS circuits". In: Feb. 2001, pp. 227–232. ISBN: 0-7695-0831-6. DOI: 10.1109/ICVD.2001.902665.

[162] Yannis Tsividis and Colin McAndrew. *Operation and modeling of the MOS transistor; 3rd ed.* Oxford series in electrical and computer engineering. New York, NY: Oxford Univ. Press, 2011. URL: https://cds.cern.ch/record/1546736.

[163] V. M. van Santen et al. “Aging-aware voltage scaling”. In: *Design, Automation Test in Europe Conference Exhibition (DATE)*. 2016, pp. 576–581. DOI: 10.3850/9783981537079_0751.

[164] Harry Veendrick. *Nanometer CMOS ICs*. Apr. 2017. ISBN: 978-3-319-47595-0. DOI: 10.1007/978-3-319-47597-4.

[165] S. Venkateswarlu et al. “Ambient Temperature-Induced Device Self-Heating Effects on Multi-Fin Si n-FinFET Performance”. In: *IEEE Transactions on Electron Devices* 65.7 (July 2018), pp. 2721–2728. ISSN: 1557-9646. DOI: 10.1109/TED.2018.2834979.

[166] U. Verner, A. Mendelson, and A. Schuster. “Extending Amdahl’s Law for Multicores with Turbo Boost”. In: *IEEE Computer Architecture Letters* 16.1 (2017), pp. 30–33. DOI: 10.1109/LCA.2015.2512982.

[167] Laung-Terng Wang, Yao-Wen Chang, and Kwang-Ting (Tim) Cheng, eds. *Electronic Design Automation: Synthesis, Verification, and Test*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2009. ISBN: 9780080922003.

[168] David Wolpert and Paul Ampadu. “Temperature Effects in Semiconductors”. In: *Managing Temperature Effects in Nanoscale Adaptive Systems*. Springer New York, 2012, pp. 15–33. ISBN: 978-1-4614-0748-5. DOI: 10.1007/978-1-4614-0748-5-2.

[169] D. H. Woo and H. S. Lee. “Extending Amdahl’s Law for Energy-Efficient Computing in the Many-Core Era”. In: *Computer* 41.12 (Dec. 2008). DOI:10.1109/MC.2008.494, pp. 24–31. ISSN: 0018-9162. DOI: 10.1109/MC.2008.494.

[170] K. Wu and D. Marculescu. “Aging-aware timing analysis and optimization considering path sensitization”. In: *Design, Automation Test in Europe (DATE)*. Mar. 2011. DOI: 10.1109/DATE.2011.5763249.

[171] Q. Xie et al. “Performance Comparisons Between 7-nm FinFET and Conventional Bulk CMOS Standard Cell Libraries”. In: *IEEE Transactions on Circuits and Systems II: Express Briefs* 62.8 (Aug. 2015), pp. 761–765.

[172] S. Xie et al. “Aggregated Residual Transformations for Deep Neural Networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5987–5995.

[173] J. Yang et al. “Quantization Networks”. In: *IEEE/CVF Conf. on Comp. Vis. and Pat. Recogn. (CVPR)*. June 2019, pp. 7300–7308.

[174] J. Yin et al. “Energy-Efficient Time-Division Multiplexed Hybrid-Switched NoC for Heterogeneous Multicore Systems”. In: *2014 IEEE 28th International Parallel and Distributed Processing Symposium*. 2014, pp. 293–303.

[175] Sergey Zagoruyko and Nikos Komodakis. “Wide Residual Networks”. In: *CoRR* abs/1605.07146 (2016). arXiv: 1605.07146. URL: http://arxiv.org/abs/1605.07146.

[176] G. Zervakis, H. Amrouch, and J. Henkel. “Design Automation of Approximate Circuits With Runtime Reconfigurable Accuracy”. In: *IEEE Access* 8 (2020), pp. 53522–53538. DOI: 10.1109/ACCESS.2020.2981395.

[177] X. Zhang et al. “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices”. In: *IEEE/CVF Conf. on Comp. Vis. and Pat. Recogn. (CVPR)*. June 2018, pp. 6848–6856.

[178] Wei Zhao and Yu Cao. “New Generation of Predictive Technology Model for Sub-45nm Design Exploration”. In: *International Symposium on Quality Electronic Design (ISQED)*. 2006, pp. 585–590. ISBN: 0-7695-2523-7. DOI: 10.1109/ISQED.2006.91.

[179] Victor Zhirnov and Ralph Cavin. “Nanoelectronics: Negative capacitance to the rescue?” In: *Nature nanotechnology* 3 (Feb. 2008), pp. 77–8. DOI: 10.1038/nnano.2008.18.

A. Appendix A

A.1. OEP under parallelized NTC

The general overview of the employed technique to analyze and evaluate the parallelized NTC approach is summarized in Fig. A.1.

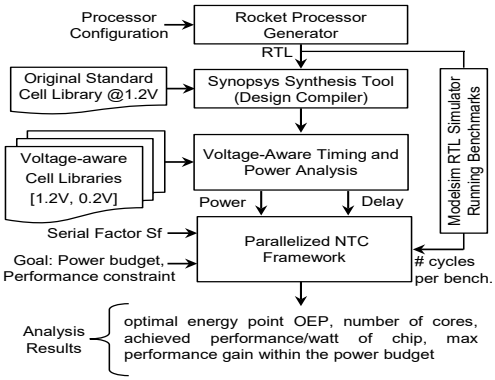


Figure A.1.: General overview of the presented voltage aware for parallelized NTC which is employed for the evaluation.

The number of cycles of the examined benchmarks is summarized in Table A.1.

Benchmarks	qsort	vvadd	towers	median
Cycles [#]	642240	162218	134267	157328
Benchmarks	dhrystone	multiply	mt-vvadd	spmv
Cycles [#]	490368	165593	604745	684101

Table A.1.: Summary of running benchmarks on the Rocket processor

A.2. Precision Scaling Modeling through NN Quantization

Precision scaling is a widely-used technique to reduce the power of circuits. By reducing the precision of the inputs, e.g., clock-gate some input bits, the circuit's switching activity reduces accordingly and consequently its dynamic power. Hence, the power density scales down accordingly, which therefore results in a less on-chip temperature. Similar to [8], we use precision scaling though quantization for the purpose of temperature management of NPUs. However, precision scaling reduces the numerical range of the inputs and consequently it leads to accuracy losses.

A.2.1. Quantization Modeling:

To enable precision scaling in NN inference, quantization is typically applied [173]. During quantization, weights and activations are converted into lower-precision numerical representations of the original 32-bit floating-point number. For example, it is demonstrated that "INT8" delivers almost the same accuracy as FLOAT32 and among others it delivers very high inference speedup since it avoids the high complexity of performing floating-point operations [143, 69]. In practice, considering that 8-bit is the baseline precision of the MAC array similar to Google TPU [69] and Samsung NPU[143], quantizing the NN to lower bit-widths reduces the precision of the inputs of the MAC array leading to lower power and temperature as explained above. Note that employing quantization instead of traditional precision scaling does result in a higher accuracy since the weights and activations will be scaled accordingly to the new precision range. To perform quantization, we employ the open-source machine learning Pytorch to train the NN using the default 32-bit float-point (FP) representation and then we apply n-bit quantization. We employ asymmetric min/max post-training quantization method using zero-point (ZP) in addition to the scale factor (S). By this method we map the min and max values of FP representation to the minimum and maximum range of the desired

precision level. Once the NN is trained, the tensor weight quantization to n-bit is performed (see Eq. A.1) and rounding is needed if result is not an integer.

$$\begin{aligned} x_{quant} &= S * x_{float} - ZP \\ S &= \frac{x_{quant}^{max} - x_{quant}^{min}}{x_{float}^{max} - x_{float}^{min}} \\ ZP &= x_{float}^{max} * S - x_{quant}^{max} \end{aligned} \tag{A.1}$$

A.2.2. NN Inference Accuracy Modeling:

All the examined NNs are developed using Pytorch and are trained with 32-bit FP precision. We capture their corresponding inference accuracy over the evaluation evaluation datasets [44]. We quantize (as described above) the weights, bias, and activations of the 32-bit FP representation to the baseline 8-bit precision and to the lower precision levels; 7-bit, 6-bit and 5-bit.

A.2.3. NN Accuracy Results

Precision scaling of NN inference accuracy is done by reducing number of bit using quantization model. To achieve this, we use PyTorch and we train 20 state-of-the-art NNs listed in Table A.2 for the ImageNet dataset [44]. Table A.2 summarizes the Top-5 accuracy of the examined NNs for 8-bit to 5-bit precision as well as for the reference 32-bit float point. As shown, different NNs exhibit different tolerance to precision. For example, ResNet-34 and ResNet-18 still feature $> 80\%$ accuracy even when performing inference at 5-bit.

A.3. BTI model

BTI model consists of a set of physics-based equations for each defect type (IT, HT, OT) [106]. Interface Traps (IT) are broken *Si-H* bonds at the channel interface or *H*-passivated defects at the high-k/low-k interface in the gate dielectric. Oxide Traps (OT) are the newly generated defects in the interface layer (low-k gate dielectric) in which capturing of carriers can occur. Hole

Table A.2.: Neural network accuracy for varying quantization

Neural Network	Accuracy (%)				
	32-float	**8-bit**	**7-bit**	**6-bit**	**5-bit**
ResNet34 [57]	91.24	91.31	91.27	90.32	82.87
ResNet18 [57]	89.08	88.97	88.84	87.87	80.13
VGG-19 [139]	90.88	90.79	90.56	89.39	77.24
VGG-16 [139]	90.38	90.3	90.18	88.92	74.76
VGG-13 [139]	89.25	89.18	89.06	87.43	71.67
VGG-11 [139]	88.63	88.57	87.98	84.5	57.82
SqueezeNet 1.1 [62]	80.62	80.44	79.42	75.52	44.06
SqueezeNet 1.0 [62]	80.42	80.35	79.22	70.06	30.8
ResNeXt50-32x4d [172]	93.7	93.63	93.39	91.92	27.37
ResNet50 [57]	92.86	92.81	92.63	89.93	7.60
ResNet101 [57]	93.55	93.45	93.29	90.46	0.88
ResNet152 [57]	94.05	94	93.85	90.93	0.72
Wide ResNet-50-2 [175]	94.09	94.03	93.82	88.31	0.52
ResNeXt101-32x8d [172]	94.53	94.48	94.37	70.94	0.46
AlexNet [79]	79.07	78.8	77.41	65.84	14.49
Wide ResNet-101-2 [175]	94.28	94.23	82.95	69.96	0.50
GoogLeNet [155]	89.53	88.85	86.96	24.48	0.48
MnasNet-1.0 [157]	91.51	87.18	74.84	14.77	0.51
ShuffleNet-v2 [177]	88.32	84.64	63.12	8.18	0.49
MobileNet-v2 [132]	90.29	83.87	66.67	16.03	0.63
Average	89.81	88.99	85.49	69.79	28.70

Traps (HT) is capturing carriers in pre-existing defects due to imperfect manufacturing. Details on the defect generation mechanisms are available in [50]. The model is calibrated and validated against semiconductor measurements under a wide range of operating conditions of voltage and temperature. To provide a quantitative insight of the used BTI model and demonstrate the voltage dependency, we present in the following a simplified version of the underlying equations. The equations describe IT stress (Eq. A.2), IT recovery (Eq. A.5), HT stress (Eq. A.3), HT recovery (Eq. A.6), OT stress (Eq. A.4) and OT recovery (Eq. A.7).

Stress:

$$\Delta V_{th}(IT) = A \cdot t^n \cdot e^{\Gamma E_{OX}} \cdot e^{\frac{-E_{AIT}}{kT}}$$

$$\Gamma = \Gamma_{\cap} + \frac{\alpha_{IT}}{kT} \tag{A.2}$$

$$\Delta V_{th}(HT) = B \cdot e^{\Gamma_{HT} E_{OX}} \cdot e^{\frac{-E_{AHT}}{kT} \cdot [1 - e^{-(\frac{t}{\tau_{HT}})^{\beta_{HT}}}]} \tag{A.3}$$

$$\Delta V_{th}(OT) = \frac{q}{C_{OX}} k_{FOT} \left[1 - e^{-(\frac{t}{m})^{\beta_S}}\right],$$

$$m = \eta \cdot (V_{OV})^{-\frac{\Gamma_{OT}}{\beta_S}} \frac{E_{AOT}}{e^{kT\beta_S}},$$

$$\Gamma_{OT} = \Gamma_{OT0} + \frac{\alpha_{OT}}{kT} \tag{A.4}$$

Recovery:

$$\Delta V_{th}(IT) = \Delta V_{IT}(EOS) \cdot e^{-(\frac{t}{\tau})^{\beta}} \tag{A.5}$$

$$\Delta V_{th}(HT) = \Delta V_{HT}(EOS) \cdot e^{-(\frac{t}{\tau_{DT}})^{\beta_{DT}}} \tag{A.6}$$

$$\Delta V_{th}(OT) = \Delta V_{OT}(EOS) \cdot e^{-(\frac{t}{\tau_R})^{\beta_R}} \tag{A.7}$$

A, B, K_{FOT}: pre-factors for bulk trap generation, n: time exponent, η: dispersion parameter, Γ: field acceleration factor, $\Gamma 0$: temperature independent field acceleration factor, Γ_{OT0}: voltage acceleration parameter, α_{IT}: bond polarization factor, α_{OT}: bond polarization factor, E_{AIT}: IT temperature activation, E_{AHT}: HT temperature activation, E_{AOT}: OT temperature activation, E_{OX}: oxide electric field, τ: Recovery time constant (depends on recovery bias), τ_{HT}: trapping time constant, T_{DT}: detrapping time constant, τ_{DT}: time constant for recovery, β: time constant dispersion parameter, β_{HT}: time constant dispersion parameter for trapping, β_{DT}: time constant dispersion parameter for detrapping, β_S: stretching parameter for stress, β_R: stretching parameter for recovery, $\Delta V_{IT}(EOS)$, $\Delta V_{HT}(EOS)$, $\Delta V_{OT}(EOS)$: induced degradation at the end of stress in the previous cycle.

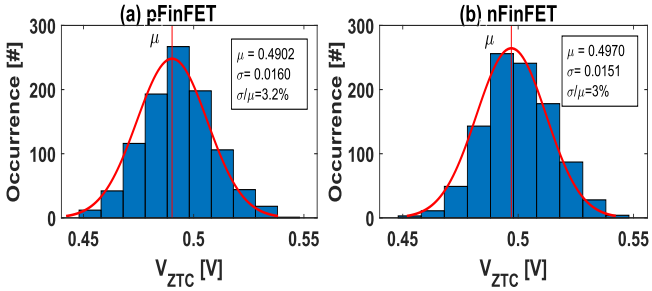


Figure A.2.: The histogram of ZTC of (a) pFinFET and (b)nFinFET transistors under process variations. V_{ZTC} values for both transistor types are distributed within a small range [0.45V - 0.55V].

A.4. Self-Heating Related Background

Here, we explain some background details which important for our work.

A.4.1. ZTC of transistors under process variations

Due to process variations, each transistor within the circuit could have different characteristics. This results in a variation of ZTC of transistors. To demonstrate such variation, we simulate 1000 different nFinFET and 1000 different pFinFET transistors (i.e., different length, width, etc) using HSPICE. The actual variability data are taken from [6, 97] for Intel 14nm FinFET technology. We study the variations for T_C high and low for a large range of voltages [0.2V-0.7V] with 10mV steps (see Algorithm 6.1). To determine ZTC of a transistor, we examine I_d of the transistor at high and low T_C. The voltage that shows no difference in I_d (because the propagation delay of the transistor is function of I_d) is therefore our ZTC. Results show that V_{ZTC} values for both transistor types are distributed within a small range [0.45V - 0.55V] as demonstrated in Fig. A.2. Importantly, by design, *V_{ZTC} of a chip must be located within this small range*.

B. Appendix B

B.1. NCFET Voltage Amplification

The ferroelectric (FE) layer manifests itself as a negative capacitance (NC) under certain conditions of capacitate matching. This, in turn, leads to charge redistribution in which the gate switching occurs at a lower applied potential than in conventional oxides. This can be attributed to the better gate control over the channel and hence the enhancements in the surface potential. The obtained gain from NC due to the better electrostatic integrity can be summarized in the following.

$$
\begin{aligned}
&E_{fe} = \frac{V_{fe}}{t_{fe}} \approx 2\alpha P + 4\beta P^3 + 6\gamma P^5 \; ; P \approx Q \Rightarrow \\
&V_{fe} = t_{fe}(2\alpha Q + 4\beta Q^3 + 6\gamma Q^5) \\
&\text{NC Effect: } Q_G > 0 \Rightarrow V_{fe} < 0 \Rightarrow V_{int} = V_G + |V_{fe}| \\
&C_{fe} \equiv \frac{\partial Q}{\partial V_{fe}} = \frac{1}{t_{fe}(2\alpha + 12\beta Q^2 + 30\gamma Q^4)} \\
&A_V = \frac{\partial V_{int}}{\partial V_G} = \frac{|C_{fe}|}{|C_{fe}| - C_{int}} \; . \\
&\text{To ensure hysteresis-free: } |C_{fe}| > C_{int} \Rightarrow A_V > 1
\end{aligned}
$$

Where α, β and γ are the material-specific Landau parameters required to define E_{fe}. V_{fe} is the voltage across the ferroelectric layer, which can be expressed as a function of its terminal charge Q following the phenomenological Landau-Devonshire (L-D) theory. C_{fe} and C_{int} are the ferroelectric layer capacitance and internal (i.e., underlying MOS) capacitance, respectively.

B.2. Optimal Frequency and Voltage Selection Point in NCFET-based Processor:

Power/energy management techniques (e.g., DVFS) operate under some constraints, e.g., to maximize performance given a power/energy budget or to minimize power/energy given a performance goal. The constraints are fulfilled by selecting the proper V/f pairs. V_{opt}/f_{opt} selection following Eq. (8.14) and Eq. (8.16) is an optimization problem that can be solved using a search algorithm. Our algorithm to perform the required optimization is summarized in Algorithm 8.1. It performs a search by sweeping across all possible frequency and voltage steps. Since the number of discrete frequency/voltage settings is limited by the hardware and closed-form analytical models can be evaluated quickly, the search is fast. In addition, it can be applied either online (i.e., executing Algorithm 8.1 at runtime) or offline (i.e., pre-characterizing a processor at design time and selecting predefined V/f pairs at runtime). The offline technique works by characterizing the processor operating point pairs V_{opt}/f_{opt} as a function of workload ratios and performance at design time. Given actual measured or derived workload characteristics and performance goals in the current power/energy management epoch, operating points V_{opt}/f_{opt} can then be selected by the operating system or hardware at runtime.

We further examine how the optimal frequency and the corresponding optimal voltage that minimize power or energy using our technique depends on possible workload characteristics. To cover a wide range of workloads, we examine dynamic/total power ratios in the range of 0.1-0.9 for W=10^6 cycles, T=20ms (time to finish $10W = 10^7$ cycles at f_{min}=0.5GHz), and a performance constraint of f_d=0.8GHz to meet T.

The optimal frequencies and voltages for power and energy minimization cases are shown in Fig. B.1(a) and (b), respectively. For power minimization, frequincies and voltages are shown for the active phase. The voltage is set to V_{leak} in the idle phase (see Section 8.2.2).

The figure shows that different frequencies and voltages are selected based on the optimization case (i.e., power or energy). Moreover, for all cases, TFE4 exhibits the best performance over all thicknesses.

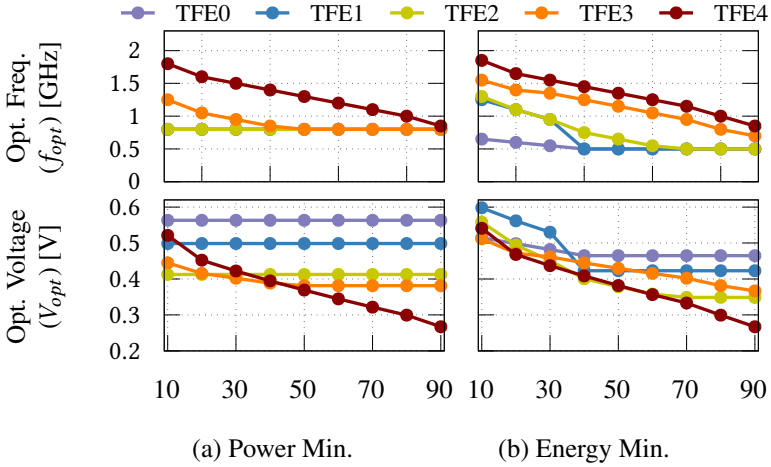


Figure B.1.: Optimal frequency and voltage selected by NCFET-aware power and energy management technique over dynamic/total ratio for thicknesses TFEx targeting (a) power minimization using W=10^6 and (b) energy minimization with f_d=0.8GHz, W=10^6 and T=20ms.