UNIVERSITATEA LUCIAN BLAGA DIN SIBIU

Doctoral School of Engineering Sciences and Mathematics
Doctoral field: Computers and Information Technology

PhD Thesis

Anticipative and Predictive Techniques in Multicore Microprocessors

Author:

Claudiu-Raul Buduleci, MSc

PhD Supervisor:

Prof. Remus Brad, PhD

SIBIU 2024

ABSTRACT

This thesis provides a comprehensive review of existing research and advancements in various aspects of microprocessor design and optimization, covering key challenges and innovative solutions. It introduces the issue bottleneck, a fundamental problem in microprocessor architectures, which refers to the limitations in the number of instructions that can be issued per clock cycle. It reviews different strategies and mechanisms proposed in the literature to overcome these limitations and improve overall processor throughput. In addition, it addresses the critical issues of security and data consistency in microprocessor systems with speculative execution capabilities. Further, it delves into various aspects of dynamic instruction reuse (DIR), value prediction (VP), and dynamic voltage and frequency scaling (DVFS), including theoretical foundations and practical implementations. Additionally, it covers state-of-the-art multi/many-core microprocessor simulators and parallel benchmarks, which are used for evaluating and optimizing microprocessor designs.

It starts by presenting the augmentation of the Sniper simulator, with access to the operand values of a specific group of instructions. The motivation for accessing the operand values is given because computer programs, in particular graphic and multimedia applications, are characterized by a high degree of redundancy which can be exploited by DIR techniques. This work falls into the category of "open architectures" starting from the open-source concept, as it provides researchers with a methodology for reading the values of instruction operands. The most important added value is a technical one, comprising details about the simulator's architecture and modifications. Furthermore, it is proposing a simulation methodology to study the potential reusability of specific types of dynamic instructions. The experimental study is done on the Splash-2 benchmark suite using the modified simulator, by varying the following parameters: the number of cores, the compiler optimization method, and the operand history length (depth of different operand values stored for each instruction). The results are indicating a promising potential of reusability, on average varying from 84 % to 87 %, on selected types of dynamic instructions and lead to the idea of implementing a 4-way associative reuse buffer (RB) in a multi/many-core system. The level of optimization of the compiler influences the degree of reusability and overall performance. A rough estimation of potential gain in performance (speedup) is also calculated, reaching a maximum of 17.5 % and 3.6 % on average.

Another focus point is an original contribution that augments the Intel Nehalem multicore architecture with a selective high-latency arithmetic set-associative RB. The architecture is simulated using Sniper, which was further adapted to estimate the power consumption, area of integration, and chip temperature, including latency modifications for the newly added unit. The implementation of a set-associative RB is a new approach, along with its applicability in a multicore microprocessor, applied to long-latency arithmetical instructions targeting dataflow bottleneck and increasing CPU performance. Additionally, a manual design space exploration process was done for the enhanced

microarchitecture, by varying the associativity and the size of the proposed RB unit and evaluating the impact on the interested metrics. The simulations on the Splash-2 benchmarks, revealed an average reuse rate of up to 33.27 % allowing a maximum speedup of 6.56 %. While the energy consumption remains stable, it can be seen an average chip temperature reduction of 2.8 °C along with an increase in associativity.

Another highlight of this work consists in the implementation and evaluation of a selective VP in a multicore environment, with a focus on long latency arithmetical instructions, having the goal to break the dataflow bottleneck of each core, thus increasing the overall performance. The Sniper simulator was used to augment the Intel Nehalem architecture with a VP and to estimate the computing performance, area of integration, power consumption, energy efficiency, and chip temperature for the enhanced architecture. Multiple simulation scenarios were performed, where multiple parameters of the VP unit were varied: the number of entries, associativity, and the number of values that are used for prediction for each instruction. By increasing the history length, it was measured, on average, more than 3 % increase in performance (core speed-up), a reduction in chip temperature from 57.8 °C to 56.17 °C, and lower energy consumption in most cases, compared with the baseline configuration. A comparison between the VP and DIR techniques in an equitable condition (to exploit the same value locality) was done, out of which the advantages and disadvantages of each technique in the given context are highlighted. The comparison was done in variation with the number of cores in the system.

Moreover, an empirical analysis of the consecrated Splash-2 benchmark suite vs. the latest version of Splash-4 was performed. It was shown that on a 64-core configuration, half of the simulated benchmarks reach temperatures well beyond the critical threshold of 105 °C, emphasizing the necessity of a multi-objective evaluation from at least the following perspectives: energy consumption, performance, chip temperature, and integration area. During the analysis, it was observed that the cores spend a large amount of time in the idle state, around 45 % on average in some configurations. This can be exploited by implementing a predictive DVFS technique called the Simple Core State Predictor (SCSP) to enhance the Intel Nehalem architecture and to simulate it using Sniper. The aim was to decrease the overall energy consumption by reducing power consumption at core-level while maintaining the same performance. More than that, the SCSP technique, which operates with core-level abstract information, was applied in parallel with VP or a DIR technique, which rely on instruction-level information. Using the SCSP alone, a 9.95 % reduction in power consumption and an energy reduction of 10.54 % were achieved, maintaining the performance. By combining the SCSP with the VP technique, a performance increase of 8.87 % was obtained while reducing power and energy consumption by 3.13 % and 8.48 %, respectively.

List of Publications

The publications done during the development of this thesis is the following:

- **C. Buduleci**, A. Gellert, A. Florea, R. Chis, and R. Brad, "Multi-Objective Optimization of Speculative and Anticipative Multi-Core Architectures," in *Advanced Computer Architecture and Compilation for High-performance Embedded Systems*, Fiuggi, Italy: HiPEAC, 2020, pp. 11–14.
- **C. Buduleci**, A. Gellert, A. Florea, and A. Matei, "Extending Sniper with Support to Access Operand Values: A Case Study on Reusability Measurement," in *2022 23rd International Carpathian Control Conference (ICCC)*, Sinaia, Romania: IEEE, May 2022, pp. 70–75. doi: 10.1109/ICCC54292.2022.9805869.
- **C. Buduleci,** A. Gellert, and A. Florea, "Selective High-Latency Arithmetic Instruction Reuse in Multicore Processors," in *2023 27th International Conference on System Theory, Control and Computing (ICSTCC)*, Timisoara, Romania: IEEE, Oct. 2023, pp. 410–415. doi: 10.1109/ICSTCC59206.2023.10308483. **(Best Paper Award)**
- **C. Buduleci**, A. Gellert, A. Florea, and R. Brad, "Architectural and Technological Approaches for Efficient Energy Management in Multicore Processors," *Computers*, vol. 13, no. 4, p. 84, Mar. 2024, doi: 10.3390/computers13040084.
- **C. Buduleci**, A. Gellert, A. Florea, and R. Brad, "Improving Multicore Architectures by Selective Value Prediction of High-Latency Arithmetic Instructions," *Adv. Electr. Comp. Eng.*, vol. 24, no. 2, pp. 61–72, 2024, doi: 10.4316/AECE.2024.02007.

Previously published papers of the author that are strongly connected to this work:

- A. Florea, **C. Buduleci**, R. Chis, A. Gellert, and L. Vintan, "Enhancing the Sniper simulator with thermal measurement," in 2014 18th International Conference on System Theory, Control and Computing (ICSTCC), Sinaia: IEEE, Oct. 2014, pp. 31–36. doi: 10.1109/ICSTCC.2014.6982386.
- R. Chis, A. Florea, **C. Buduleci**, and L. Vintan, "Multi-Objective Optimization for an Enhanced Multi-Core SNIPER Simulator," Proceedings of The Romanian Academy, Series A: Mathematics, Physics, Technical Sciences, Information Science, vol. 19, no. 1, pp. 85–93, Mar. 2018.

The success of this thesis is largely due to the remarkable efforts of many people who gave inspiration, expertise, and resources.

I would like to start by expressing my deepest gratitude to my advisors, Prof. Dr. Ing. Remus Brad, Prof. Dr. Ing. Adrian Florea, and Conf. Dr. Ing. Arpad Gellert, for their invaluable mentorship, support, and guidance throughout this journey. I am also grateful to Conf. Dr. Ing. Daniel Morariu for his valuable insights and for being a member of the committee.

On a personal note, I am incredibly grateful to my family and friends for their unwavering love, encouragement, and patience throughout this challenging process. A special thanks to my wife, Mihaela, who has been a constant source of happiness and supported me in achieving this goal. I am also deeply grateful to my close friend Bogdan for his unwavering support and inspiration over the years.

With profound gratitude, I would like to dedicate this work to the memory of Prof. Dr. Ing. Lucian Vinţan, who believed in my potential and inspired me to embark on this challenging journey.

Finally, I would like to thank everyone who has directly or indirectly contributed to this work.

TABLE OF CONTENTS