ENERGY-EFFICIENT MANAGEMENT OF RECONFIGURABLE COMPUTERS

by

Rathijit Sen

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2016

Date of final oral examination: 05/13/16

The dissertation is approved by the following members of the Final Oral Committee:
David A. Wood, Professor, Computer Sciences
Mark D. Hill, Professor, Computer Sciences
Mikko H. Lipasti, Professor, Computer Sciences
Gurindar S. Sohi, Professor, Computer Sciences
Michael M. Swift, Associate Professor, Computer Sciences

To my parents Meena Sen and Ranjit Kumar Sen

ACKNOWLEDGMENTS

I am honored to have Prof. David Wood as my advisor. He has patiently guided me and supported me over so many years. David is an erudite scholar and outstanding professor. I have learned so much from him about computer architecture and analytical modeling. He encouraged me to consider the theoretical underpinnings in addition to the practical feasibility of problem solutions. He helped me explore ideas, refine drafts of writeups, improve my presentation skills, and network with other researchers through attending conferences. Thank you, David, for your time, support, and help.

I would like to thank the other members of my committee, Profs. Mark Hill, Mikko Lipasti, Guri Sohi, and Mike Swift, for giving me valuable feedback on my work. I have learned a lot from their vast experience and deep insight. Mark and Mike gave me great inputs for improving my papers and posters. Mike pointed me to many interesting papers. I have learned a lot from Mark about queueing models and how to present my work better. I had insightful discussions with Mikko about computer architecture research. Guri gave me power meters that I used to measure server power for this work.

I am also grateful to the Multifacet research group for providing me with necessary infrastructure to do the work in this dissertation. The server clusters, Lapis full-system simulator, Ruby memory hierarchy simulator, and Wisconsin commercial workloads have been invaluable to me. Technical discussions at group meetings were highly instructive.

Dan Gibson helped me a lot with setting up of the simulation infrastructure. Somayeh Sardashti and Hamid Reza Ghasemi also helped with creating simulation checkpoints.

This work was supported in part by the National Science Foundation (CNS-0916725, CCF-1017650, CNS-1117280, CCF-1218323, CNS-1302260, CCF-1438992, CCF-1533885), Microsoft Corporation (MSN140822), Sandia National Labs/DOE (MSN123960/DOE890426), and a University of Wisconsin Vilas award.

I would like to thank Prof. Somesh Jha and Prof. Jignesh Patel for their collaboration and help on multiple projects in the areas of static program analysis and databases. Jignesh gave me an opportunity to contribute to the Quickstep project and generously supported me during my final summer at UW-Madison. I look forward to continuing our collaboration. I would also like to thank Prof. Tom Reps for supporting me during my first year at UW-Madison and for giving me an opportunity to work on projects involving static analysis of executables.

My current and former office mates—Jayneel Gandhi, Derek Hower, Lena Olson, and Somayeh Sardashti—have given me company and have always been ready to help me. I have enjoyed countless hours chatting with them. I will miss Lena's delicious home-baked muffins. Derek helped me understand many details of the Ruby simulator.

I have enjoyed working with Gagan Gupta on a long project involving static analysis and having many conversations with him about computer architecture, emerging technologies, reading groups, paper reviews and rebuttals. I also explored other projects with Arkaprava Basu, Asim Kadav, Nilay Vaish, Cong Wang, and Jianqiao Zhu. Discussions with Siddharth Barman helped me appreciate different perspectives on theory and systems research. Jason Power's elevator pitch workshop was useful and fun.

I have also benefited from interacting with many other excellent colleagues including Muhammad Shoaib Bin Altaf, Newsha Ardalani, Piramanayagam Arumuga Nainar, Raghuraman Balasubramanian, Emily Blem, Jayaram Bobba, Evan Driscoll, Polina Dudnik, Yasuko Eckert, Christopher Feilbach, Venkatraman Govindaraju, Swapnil Haria, Bill Harris, Joel Hestness, Nick Kidd, Marc De Kruijf, Akash Lal, Junghee Lim, Jaikrishnan Menon, Sanketh Nalli, Tony Nowatzki, Marc Orr, Sankaralingam Pannerselvam, Matt Sinclair, Srinath Sridharan, Swaminathan Sundararaman, Aditya Thakur, Aditya Venkataraman, Haris Volos, and Hongil Yoon.

A number of external researchers, including UW-Madison computer architecture alumni, gave useful feedback on my work. I would like to acknowledge Luiz Barroso, Edouard Bugnion, Lisa Hsu, Mike Marty, Kathryn McKinley, and Ravi Rajwar for helpful suggestions and comments. The Architecture Affiliates meetings helped to get useful feedback also from Alaa Almeldeen, Brad Beckmann, John Davis, Dan Gibson, Peter Hsu, Konrad Lai, Kevin Moore, Steve Reinhardt, Greg Thorson, Greg Wright, and others.

I learned about new applications for analytical models while working with Trishul Chilimbi, Wei Huang, Srilatha Manne, and Indrani Paul during internships at Microsoft and AMD. I also enjoyed interacting with Manish Arora, Joseph Greathouse, and fellow interns Lavanya Subramanian and Zhe Wang. Weekend trips around and near Seattle with Arkaprava Basu, Tushar Krishna, and Shekhar Srikantaiah were enjoyable.

Angela Thorp, our graduate program coordinator, always helped me with understanding and following graduate school and departmental procedures. The Computer Systems Lab staff, and in particular Tim Czerwonka, promptly resolved all issues with the computing infrastructure that I ran into while doing my work.

Prof. Y. N. Srikant (IISc) and Prof. Reinhard Wilhelm (Saarland University) have greatly encouraged me. I wish to also acknowledge the friendship of some wonderful people, among them being Prof. Jan Reineke (Saarland University), Oindrilla Gupta, Kuntal Dey, Mohamed Abdel Maksoud, S. V. N. Narayana Rao, and Arpan Sen.

My sister Rakhee, brother-in-law Sandip, and niece Swagnita have brought me joy with their kindness, encouragement, and arrangements for many memorable excursions.

It has been a long and difficult journey, one on which I could not have persevered without the steadfast support of my parents. They have been a pillar of strength for me. I dedicate this dissertation to them. Thank you for your boundless love, encouragement, and good wishes throughout the years.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Power and energy consumption are first-order constraints on the design and operation of computer systems today. Improving energy efficiency reduces the amount of energy needed to perform a given computation as well as enables more computation to be performed for the same amount of energy. This saves operational costs to use these systems as well as capital costs to provision for them.

Conventionally, energy proportionality (energy consumption in proportion to the work done, or equivalently, power consumption in proportion to utilization/performance/load served) as proposed by Barroso and Hölzle, has been the gold standard of an ideal system's energy efficiency. While this model is valid for fixed-resource systems, modern systems are reconfigurable in many aspects, allowing them to adapt to changing workload characteristics. Smart reconfigurability increases energy efficiency. However, we show that reconfigurability invalidates the conventional notions of ideal energy proportionality if the system starts to behave super-proportionally. Super-proportional systems provide more performance (or work) in proportion to the power (or energy) used. We propose a new ideal model, Energy Optimal Proportional (EOP), that subsumes the conventional model and improves upon it by also accounting for super-proportional systems.

EOP can guide system designers to improve the maximum efficiency attainable over the operating range and forms a basis for comparisons of energy efficiency across systems. Power-performance Pareto optimality, on the other hand, can guide system operators to manage load and configure resources appropriately to make the current system execute efficiently. We propose a new intellectual framework that interrelates these two complementary energy efficiency goals.

The rest of this dissertation focuses on energy-efficient management. We develop new reactive governors that coordinate processor frequency (and voltage) and hardware

prefetching to improve energy efficiency on a real (Haswell) server. We also propose a space-efficient hardware mechanism to estimate temporal locality (reuse) in cache accesses. The estimated distributions can be used by our new analytical models for cache performance to drive resizing decisions of the last-level cache.

Finally, we propose a new classification system for system reconfiguration capabilities. The classification is based on the semantics of what the reconfiguration affects—computation, communication, storage, scheduling, speculation. We hope that this classification will be insightful to future researchers while exploring the space of reconfigurable systems, in categorizing existing work and in identifying coordination options that have been less well explored.