INVESTIGATING THE EFFECTS OF DECOUPLING CACHE AND CORE SPEED ON POWER, THROUGHPUT, AND ENERGY CONSUMPTION

DAVID DANIEL PRUITT

Doctoral Program in Computer Science

APPROVED:

Eric Freudenthal, Ph.D., Chair

Shirley Moore, Ph.D.

Art Duval, Ph.D.

Luc Longpre, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

Dedication

To the friends that made this process possible

“It is possible to commit no errors and still lose. That is not a weakness. That is life.”

INVESTIGATING THE EFFECTS OF DECOUPLING CACHE AND CORE SPEED ON POWER, THROUGHPUT, AND ENERGY CONSUMPTION OF CPUS

by

David Daniel Pruitt BS CS MS CS

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

August 2022

Acknowledgements

I would like to thank everyone that assisted during my long and arduous journey in completing this project. I'm sure I'll missing some names here. I would like to thank the various members of the Robust Autonomic systems group for their support over the years. Gabe Arellano for sharing writing tips. Edward Dragone and Edward Hudgins for their help in various technical areas. Adrian Veliz for forging the path and advising of the landmines that appear. Daniel Cervantes for allowing me to bounce ideas off him and his support when I was trying to teach. I would also like to thank the unofficial members, David Reyes and Oscar Veliz for answering the stupid questions I had with the appropriate vigor.

I would also like to thank Eric Freudenthal and Salamah Salamah for providing assistance, useful advice, and direction when needed that made this journey possible. I want to thank Dr Shirly Moore for working through ideas, spotting problems, and providing helpful suggestions. I want to thank Dr Art Duval and Luc Longpre for taking the time out of their busy schedules to read over drafts and provide feedback.

"You ever figure procrastination is your brain's way of stopping you from making a terrible mistake? Yeah... Me too." Cayde-6

Abstract

A variety of computer systems from HPC to mobile systems are power limited and performance sensitive. These systems use very similar components at different scales. Dynamic Voltage and Frequency Scaling (DVFS) features enable modulation of CPU performance and efficiency characteristics to power, energy and timing requirements.

Programs have a variety of computational characteristics. If a CPU subsystem substantially limits a particular program's execution progress, that program's throughput will vary proportionally with the subsystem's clock frequency. In contrast, if a CPU subsystem does not substantially limit throughput, the impact of a change in its clock frequency will result in a diminimus change in a program's execution time.

Dynamic Voltage and Frequency Scaling (DVFS) power domains commonly encompass entire cores and their associated caches. This work indicates that moderate energy efficiency gains may be attainable for some programs if limiting and non-limiting subsystems' (D)VFS domains are decoupled. This decoupling enables tuning of their relative performance to application characteristics. Widely used simulation and modeling tools were extended to support this exploratory research.

Table of Contents

List of Tables

List of Figures

List of Equations

Chapter 1: Introduction

1.1 MOTIVATION

Modern computer systems from HPC to mobile systems are power limited and performance sensitive. These systems use very similar components at different scales [1].

Furthermore, application-specific CPUs may be optimized for energy efficiency. For example, CPUs and their subsystems can be selected and clocked to provide required performance. Even instruction sets are optimized to satisfy embedded systems' power, performance, and energy requirements.

(D)VFS features enable modulation of CPU performance and efficiency characteristics to better match system power and energy limitations, and application constraints. Higher frequency provides greater throughput at the cost of increased power consumption. When sufficient parallelism is exposed by the application and available from a system, optimization of efficiency via DVFS can also maximize throughput. Energy consumed by computing components is classified as either active or background. Active energy consumption is due to gate state changes on the computational data or control path. Background energy is consumed by leakage current and activity required to keep the system "alive."

The throughput of a CPU-limited computation is limited by the rate of gate transitions. The energy required for a gate transition activity is monotonic with voltage. Voltage is monotonic with frequency. Therefore, energy required to actively compute a result is monotonic with frequency. And therefore, if all energy went to active gate transitions, a computation would be most efficient at lowest available frequency.

Background energy consumption is the integral of background power consumption over the time required to complete a computation. Background power varies only 20% over the range