A Primer on Compression in the Memory Hierarchy

Synthesis Lectures on Computer Architecture

Synthesis Lectures on Computer Architecture publishes 50- to 100-page publications on topics pertaining to the science and art of designing, analyzing, selecting and interconnecting hardware components to create computers that meet functional, performance and cost goals. The scope will largely follow the purview of premier computer architecture conferences, such as ISCA, HPCA, MICRO, and ASPLOS.

Power-Efficient Computer Architectures: Recent Advances
Magnus Själander, Margaret Martonosi, and Stefanos Kaxiras
2014

FPGA-Accelerated Simulation of Computer Systems
Hari Angepat, Derek Chiou, Eric S. Chung, and James C. Hoe
2014

A Primer on Hardware Prefetching
Babak Falsafi and Thomas F. Wenisch
2014

On-Chip Photonic Interconnects: A Computer Architect's Perspective
Christopher J. Nitta, Matthew K. Farrens, and Venkatesh Akella
2013

Optimization and Mathematical Modeling in Computer Architecture
Tony Nowatzki, Michael Ferris, Karthikeyan Sankaralingam, Cristian Estan, Nilay Vaish, and David Wood
2013

Security Basics for Computer Architects
Ruby B. Lee
2013

The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Second edition
Luiz André Barroso, Jimmy Clidaras, and Urs Hölzle
2013

Shared-Memory Synchronization
Michael L. Scott
2013

Resilient Architecture Design for Voltage Variation
Vijay Janapa Reddi and Meeta Sharma Gupta
2013

Multithreading Architecture
Mario Nemirovsky and Dean M. Tullsen
2013

Performance Analysis and Tuning for General Purpose Graphics Processing Units (GPGPU)
Hyesoon Kim, Richard Vuduc, Sara Baghsorkhi, Jee Choi, and Wen-mei Hwu
2012

Automatic Parallelization: An Overview of Fundamental Compiler Techniques
Samuel P. Midkiff
2012

Phase Change Memory: From Devices to Systems
Moinuddin K. Qureshi, Sudhanva Gurumurthi, and Bipin Rajendran
2011

Multi-Core Cache Hierarchies
Rajeev Balasubramonian, Norman P. Jouppi, and Naveen Muralimanohar
2011

A Primer on Memory Consistency and Cache Coherence
Daniel J. Sorin, Mark D. Hill, and David A. Wood
2011

Dynamic Binary Modification: Tools, Techniques, and Applications
Kim Hazelwood
2011

Quantum Computing for Computer Architects, Second Edition
Tzvetan S. Metodi, Arvin I. Faruque, and Frederic T. Chong
2011

High Performance Datacenter Networks: Architectures, Algorithms, and Opportunities
Dennis Abts and John Kim
2011

Processor Microarchitecture: An Implementation Perspective
Antonio González, Fernando Latorre, and Grigorios Magklis
2010

Transactional Memory, 2nd edition
Tim Harris, James Larus, and Ravi Rajwar
2010

Computer Architecture Performance Evaluation Methods
Lieven Eeckhout
2010

Introduction to Reconfigurable Supercomputing
Marco Lanzagorta, Stephen Bique, and Robert Rosenberg
2009

On-Chip Networks
Natalie Enright Jerger and Li-Shiuan Peh
2009

The Memory System: You Can't Avoid It, You Can't Ignore It, You Can't Fake It
Bruce Jacob
2009

Fault Tolerant Computer Architecture
Daniel J. Sorin
2009

The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines
Luiz André Barroso and Urs Hölzle
2009

Computer Architecture Techniques for Power-Efficiency
Stefanos Kaxiras and Margaret Martonosi
2008

Chip Multiprocessor Architecture: Techniques to Improve Throughput and Latency
Kunle Olukotun, Lance Hammond, and James Laudon
2007

Transactional Memory
James R. Larus and Ravi Rajwar
2006

Quantum Computing for Computer Architects
Tzvetan S. Metodi and Frederic T. Chong
2006

A Primer on Compression in the Memory Hierarchy

Somayeh Sardashti, Angelos Arelakis, Per Stenström, and David A. Wood

ISBN: 978-3-031-00623-4 paperback
ISBN: 978-3-031-01751-3 ebook

DOI 10.1007/978-3-031-01751-3

A Publication in the Springer series
SYNTHESIS LECTURES ON COMPUTER ARCHITECTURE

Lecture #36
Series Editor: Margaret Martonosi, *Princeton University*
Founding Editor Emeritus: Mark D. Hill, *University of Wisconsin, Madison*
Series ISSN
Print 1935-3235 Electronic 1935-3243

A Primer on Compression in the Memory Hierarchy

Somayeh Sardashti
University of Wisconsin, Madison

Angelos Arelakis
Chalmers University of Technology

Per Stenström
Chalmers University of Technology

David A. Wood
University of Wisconsin, Madison

SYNTHESIS LECTURES ON COMPUTER ARCHITECTURE #36

ABSTRACT

This synthesis lecture presents the current state-of-the-art in applying low-latency, lossless hardware compression algorithms to cache, memory, and the memory/cache link. There are many non-trivial challenges that must be addressed to make data compression work well in this context. First, since compressed data must be decompressed before it can be accessed, decompression latency ends up on the critical memory access path. This imposes a significant constraint on the choice of compression algorithms. Second, while conventional memory systems store fixed-size entities like data types, cache blocks, and memory pages, these entities will suddenly vary in size in a memory system that employs compression. Dealing with variable size entities in a memory system using compression has a significant impact on the way caches are organized and how to manage the resources in main memory. We systematically discuss solutions in the open literature to these problems.

Chapter 2 provides the foundations of data compression by first introducing the fundamental concept of value locality. We then introduce a taxonomy of compression algorithms and show how previously proposed algorithms fit within that logical framework. Chapter 3 discusses the different ways that cache memory systems can employ compression, focusing on the trade-offs between latency, capacity, and complexity of alternative ways to compact compressed cache blocks. Chapter 4 discusses issues in applying data compression to main memory and Chapter 5 covers techniques for compressing data on the cache-to-memory links. This book should help a skilled memory system designer understand the fundamental challenges in applying compression to the memory hierarchy and introduce him/her to the state-of-the-art techniques in addressing them.

KEYWORDS

cache design, memory system, memory hierarchy, data compression, performance, energy efficiency

Contents

List of Figures

List of Tables

Preface

This primer is intended for readers who are interested in learning about the different ways that data compression can be applied to the computer memory hierarchy, including caches, main memory, and the links that connect them. This audience includes computing industry professionals and graduate students. We expect our readers to be familiar with the basics of computer architecture. Knowing the details of out-of-order execution is unnecessary, but readers should be comfortable with the basics of cache and memory hierarchy design.

This primer's primary goal is to provide readers with a basic understanding of the key challenges and opportunities in applying data compression to the memory hierarchy. We address the complementary issues of data compression and data compaction, presenting high-level concepts and how they manifest in previously proposed designs. We introduce a taxonomy to help understand which compression algorithms are most applicable to different levels of the memory hierarchy. We use another taxonomy to classify different ways of compacting variable-size compressed blocks into memory structures that are classically designed to hold fixed-size blocks.

A secondary goal of this primer is to make readers aware of the opportunity that data compression has to improve the performance, energy efficiency, and cost of future computer systems. We believe that technology trends are converging in such a way as to make data compression a compelling solution and hope this primer may help expedite its widespread adoption throughout the memory system. It is not a goal of this primer to cover all topics in depth, but rather to cover the basics and help readers identify which topics they may wish to pursue in greater depth.

Somayeh Sardashti, Angelos Arelakis, Per Stenström, and David A. Wood
December 2015

Acknowledgments

We owe many thanks for the help and support we have received during the development of this primer. We would like to thank Alaa Alameldeen, Martin Burtscher, Magnus Ekman, Aamer Jaleel, and Martin Thuresson for their helpful feedback on this work. While our reviewers provided great feedback, they may or may not agree with all of the final content of this primer.

We would also like to thank our former co-authors Alaa Alameldeen, Chloe Alverti, Fredrik Dahlgren, Magnus Ekman, Andre Seznec, and Martin Thuresson for their role in helping us understand the issues in applying data compression to the memory hierarchy. Without their contributions, this work would not have been possible.

This work was supported in part by the National Science Foundation (CCF-1218323, CNS-1302260, CCF-1438992, CCF-1533885). Professor Wood has a significant financial interest in AMD and Google. The views expressed herein are not necessarily those of the National Science Foundation, nor anyone other than the authors. This work was also supported in part by the CHAMPP project funded by the Swedish Research Council, the FP7 EUROSERVER project funded by the European Commission, and the SCHEME project funded by the Swedish Foundation for Strategic Research.

Somayeh thanks her colleagues and coauthors for sharing their expertise, her husband, Dr. Hamid Reza Ghasemi, and parents, Aghdas Zeinali and Khosro Sardashti, for all their love, support, and encouragement.

Angelos thanks his coauthors of this book for the pleasant collaboration in this inspiring work, his parents Dimitris and Mary, and his sister Stella for encouraging him to always follow his dreams, and his wife Christina for always being on his side and making his life joyful and for her love.

Per thanks all his collaborators in the past and foremost the great collaboration with the coauthors of this book for an inspiring and very enjoyable mission. Most importantly, he thanks his wife Carina and their daughter Sofia for all the love and support they provide.

David thanks his coauthors for putting up with his deadline-challenged work style, his parents Roger and Ann Wood for inspiring him to be a second-generation Computer Sciences professor, and Jane, Alex, and Zach for helping him remember what life is all about.

Somayeh Sardashti, Angelos Arelakis, Per Stenström, and David A. Wood
December 2015