

Research Infrastructures for Hardware Accelerators

Synthesis Lectures on Computer Architecture

Editor

Margaret Martonosi, *Princeton University*

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.

Research Infrastructures for Hardware Accelerators

Yakun Sophia Shao and David Brooks
2015

Analyzing Analytics

Rajesh Bordawekar, Bob Blainey, and Ruchir Puri
2015

Customizable Computing

Yu-Ting Chen, Jason Cong, Michael Gill, Glenn Reinman, and Bingjun Xiao
2015

Die-stacking Architecture

Yuan Xie and Jishen Zhao
2015

Single-Instruction Multiple-Data Execution

Christopher J. Hughes and
2015

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

© Springer Nature Switzerland AG 2022
Reprint of original edition © Morgan & Claypool 2016

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Research Infrastructures for Hardware Accelerators
Yakun Sophia Shao and David Brooks

ISBN: 978-3-031-00622-7 paperback
ISBN: 978-3-031-01750-6 ebook

DOI 10.1007/978-3-031-01750-6

A Publication in the Springer series
SYNTHESIS LECTURES ON COMPUTER ARCHITECTURE

Lecture #34
Series Editor: Margaret Martonosi, *Princeton University*
Series ISSN
Print 1935-3235 Electronic 1935-3243

Research Infrastructures for Hardware Accelerators

Yakun Sophia Shao and David Brooks
Harvard University

SYNTHESIS LECTURES ON COMPUTER ARCHITECTURE #34

ABSTRACT

Hardware acceleration in the form of customized datapath and control circuitry tuned to specific applications has gained popularity for its promise to utilize transistors more efficiently. Historically, the computer architecture community has focused on general-purpose processors, and extensive research infrastructure has been developed to support research efforts in this domain. Envisioning future computing systems with a diverse set of general-purpose cores and accelerators, computer architects must add accelerator-related research infrastructures to their toolboxes to explore future heterogeneous systems. This book serves as a primer for the field, as an overview of the vast literature on accelerator architectures and their design flows, and as a resource guidebook for researchers working in related areas.

KEYWORDS

accelerators, specialized architecture, SoC, high-level synthesis, simulators, design space exploration, workload characterization, benchmarks

Contents

Preface	xi
Acknowledgments	xiii
1 Why Accelerators, Now?	1
1.1 What is an Accelerator?	1
1.2 A Tale of Two Scalings	2
1.2.1 Moore Scaling	2
1.2.2 Dennard Scaling	3
1.3 The Combination of Moore and Dennard Scaling	4
1.3.1 Moore + Dennard—Where We Were	5
1.3.2 Moore Scaling Only—Where We Are	6
1.3.3 Dennard Only—Where We Are Unlikely To Be	7
1.3.4 A Future without Scaling: “The Winter of Despair”	7
1.4 To Live Without Scaling: “A Spring of Hope”	9
1.4.1 Why Not Architectural Scaling?	9
1.4.2 Specialization Makes a Difference	11
1.4.3 A Call for Tools in the Era of Accelerators	11
2 A Taxonomy of Accelerators.	13
2.1 Not All Apples Are Alike	13
2.2 Accelerator Taxonomy	13
2.2.1 Accelerators that Are Part of the Pipeline.	15
2.2.2 Accelerators that Are Attached to Cache	19
2.2.3 Accelerators that Are Attached to the Memory Bus	22
2.2.4 Accelerators that Are Attached to the I/O Bus	26
3 Accelerator Design Flow 101.	27
3.1 Standard RTL Design Flow	27
3.2 High-Level Synthesis	28
3.2.1 Bluespec SystemVerilog	29
3.2.2 Genesis2	29

3.2.3	Xilinx Vivado	29
3.2.4	Delite	30
3.2.5	Lime	30
3.2.6	Chisel	31
3.2.7	Spiral	31
3.2.8	PyMTL	31
4	Accelerator Modeling	33
4.1	Limitations of the RTL-Based Design Flow	33
4.2	Pre-RTL Modeling—Aladdin	34
4.2.1	Optimization Phase	36
4.2.2	Realization Phase	38
4.2.3	Integration with Memory System	42
4.2.4	Limitations	43
4.2.5	Aladdin Validation	43
4.2.6	Algorithm-to-Solution Time	45
4.2.7	Case Study: GEMM Design Space	46
5	Workload Characterization for Accelerators	51
5.1	ISA-Independent Workload Characterization—WIICA	51
5.1.1	Why ISA-Independent?	52
5.1.2	Methodology and Background	55
5.1.3	Compute	57
5.1.4	Memory	60
5.1.5	Control	63
5.1.6	Putting it All Together	65
6	Accelerator Benchmarks	67
7	Future Directions	71
	Bibliography	73
	Authors' Biographies	85

Preface

Specialized architectures have been a growing topic in both academic research and commercial development for the past decade. As traditional technology scaling slows, specialization becomes a viable solution for computer architects to continue performance growth and energy efficiency improvements without relying on technological advances.

This book aims to present a high-level overview of the state-of-the-art accelerator research in both industry and academia, with a special emphasis on research infrastructure available for accelerator-related research. This book begins by describing the technology trends that have led accelerator research to prominence. In Chapter 2, we present a taxonomy of accelerator research and practice, with the goal of introducing the reader to the flavor of accelerator designs that have been proposed in recent years. Chapter 3 presents the standard accelerator design flow from RTL generation, simulation, and synthesis. Recent advances in high-level synthesis (HLS) tools provide a promising path for accelerator development in the future, and we describe the capabilities of commercial tools like Xilinx's Vivado HLS and their limitations. Chapter 4 discusses pre-RTL modeling approaches to facilitate the rapid exploration of the design space of accelerators as well as the interaction between accelerators and the rest of the system. Chapter 5 focuses on workload characterization approaches in the context of accelerators and Chapter 6 discusses benchmarking. We end this book with a discussion on the challenges and opportunities of accelerator architectures and design tools in Chapter 7.

Yakun Sophia Shao and David Brooks
October 2015

Acknowledgments

We would like to thank Margaret Martonosi for encouraging us to write this book, as well as the feedback and support she has provided throughout the project. We would also like to thank Michael Morgan for providing us this opportunity and keeping us on schedule during the whole process. Many thanks to Kelly Shaw, Luis Ceze, and Glenn Holloway for their detailed comments that were invaluable in improving this manuscript. We would especially like to thank our many collaborators over the years: Gu-Yeon Wei, Viji Srinivasan, Simone Campanoni, Michael Lyons, Brandon Reagen, Sam Xi, and Robert Adolf. Much of the content of the book is built on wonderful collaborations and insightful discussions with many of them.

Yakun Sophia Shao and David Brooks
October 2015