Die-stacking Architecture

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.

Die-stacking Architecture
Yuan Xie and Jishen Zhao
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

Die-stacking Architecture

Yuan Xie and Jishen Zhao

ISBN: 978-3-031-00619-7 paperback
ISBN: 978-3-031-01747-6 ebook

DOI 10.1007/978-3-031-01747-6

A Publication in the Springer series
SYNTHESIS LECTURES ON COMPUTER ARCHITECTURE

Lecture #31
Series Editor: Margaret Martonosi, *Princeton University*
Series ISSN
Print 1935-3235 Electronic 1935-3243

Die-stacking Architecture

Yuan Xie
University of California, Santa Barbara

Jishen Zhao
University of California, Santa Cruz

SYNTHESIS LECTURES ON COMPUTER ARCHITECTURE #31

ABSTRACT

The emerging three-dimensional (3D) chip architectures, with their intrinsic capability of reducing the wire length, promise attractive solutions to reduce the delay of interconnects in future microprocessors. 3D memory stacking enables much higher memory bandwidth for future chip-multiprocessor design, mitigating the "memory wall" problem. In addition, heterogenous integration enabled by 3D technology can also result in innovative designs for future microprocessors. This book first provides a brief introduction to this emerging technology, and then presents a variety of approaches to designing future 3D microprocessor systems, by leveraging the benefits of low latency, high bandwidth, and heterogeneous integration capability which are offered by 3D technology.

KEYWORDS

emerging technology, die-stacking, 3D integrated circuits, memory architecture, heterogeneous integration

Contents

Preface

Three-dimensional (3D) integration is an emerging technology, where two or more layers of active devices (e.g., CMOS transistors) are integrated both vertically and horizontally in a single circuit. With continuous technology scaling, 3D integration is becoming an increasingly attractive technology in implementing microprocessor systems by offering much lower power consumption, lower interconnect latency, and higher interconnect bandwidth compared to traditional two-dimensional (2D) circuit integration.

In particular, 3D integration technologies promise at least four major benefits toward future microprocessor design.

- **Reduced interconnect wire length.** 3D integration can dramatically reduce interconnect wire length, especially by reducing global interconnects. This can directly lead to two benefits: improved circuit delay and reduced power consumption. Circuit delay reduction can be a straightforward effect of the reduced wire length; it can result in substantial system performance improvement. The power reduction is a result of reduced parasitic capacitance due to the shorter wire length and incorporating previously off-chip signals to be on-chip; it can lead to less heat generation and extended battery life.

- **Improved memory bandwidth.** 3D integration can improve memory bandwidth by an order of magnitude, because the number of interconnects between processor and the memory is no longer constrained by off-chip pin counts. In the era of big data and multithreading, applications adopt large memory working sets and multiple threads that simultaneously access the memory. Today, memory bandwidth is a fundamental performance bottleneck. 3D integration can be a critical technology in overcoming the memory bandwidth issue.

- **Enabling heterogeneous integration.** 3D integration enables heterogeneous integration, because circuit layers can be implemented with different and incompatible process technologies. Such heterogeneous integration can lead to novel architecture designs. For example, various incompatible memory technologies, such as SRAM, spin-transfer torque RAM (STT-RAM), and resistive RAM (ReRAM), can be integrated in a single processor chip to form a hybrid cache hierarchy [1].

- **Enabling smaller form factor.** 3D integration enables a much smaller form factor compared to traditional 2D integration technologies. Due to the addition of a third dimension to conventional 2D layout, it leads to a higher packing density and smaller footprint. This potentially leads to processor designs with lower cost.

Both academia and the semiconductor industry are actively pursuing this technology by developing efficient architectures in a variety of forms. From the industry prospective, 3D integrated memory is envisioned to become pervasive in the near future. Intel's Xeon Phi processors will deliver with 3D integrated DRAMs in 2016 [2]. NVIDIA announced that 3D integrated memory will be adopted in their new GPU products in 2016 [3]. AMD plans to ship high-bandwidth memory (HBM) with their GPU products and heterogeneous system architecture (HSA)-based CPUs in 2015 [4]. From the academia prospective, comprehensive studies have been performed across all aspects of microprocessor architecture design by employing 3D integration technologies, such as 3D stacked processor core and cache architectures, 3D integrated memory, and 3D network-on-chip. Furthermore, a large body of research has studied critical issues and opportunities raised by adopting 3D integration technologies, such as thermal issues which are imposed by dense integration of active electronic devices, cost issues which are incurred by extra process and increased die area, and the opportunity in designing cost-effective microprocessor architectures.

This book provides a detailed introduction to architecture design with 3D integration technologies. The book will start with presenting the background of 3D integration technologies (Chapter 1), followed by a detailed analysis of the benefits offered by these technologies including low latency, high bandwidth, heterogeneous integration capability, and cost efficiency (Chapter 2). Then, it will review various approaches to designing future 3D integrated microprocessors by leveraging the benefits of 3D integration (Chapter 3 through Chapter 6). These approaches cover all levels of microprocessor systems, including processor cores, caches, main memory, and on-chip network. Furthermore, this book discusses thermal issues raised by 3D integration and presents recently proposed thermal-aware architecture designs (Chapter 7). Finally, this book presents a comprehensive cost model which is built based on detailed cost analysis for fabricating 3D integrated microprocessors (Chapter 8). By utilizing the cost model, the book presents and compares cost-effective microprocessor design strategies.

While this book mostly focuses on designing high-performance processors, the concepts and techniques can also be applied to other market segments such as embedded processors and exascale high-performance computing (HPC) systems.

The target audiences for this book are students, researchers, and engineers in IC design and computer architecture, who are interested in leveraging the benefits of 3D integration for their designs and research.

Yuan Xie and Jishen Zhao
June 2015

Acknowledgments

Much of the work and ideas presented in this book have evolved over years in working with our colleagues and graduate students at Pennsylvania State University (in particular Professor Vijaykrishnan Narayanan, Professor Mary Jane Irwin, Professor Chita Das), and our industry collaborators including Dr. Gabriel Loh, Dr. Bryan Black, Dr. Norm Jouppi, and Mr. Kerry Bernstein.

We also thank Prof. Niraj Jha, Prof. Margaret Martonosi, and other reviewers for the comments and feedback to improve the draft.

Yuan Xie and Jishen Zhao
June 2015