Technology-Driven Architecture Innovations: Opportunities and Challenges
Past, Present, and Future

Yuan Xie
University of California, Santa Barbara
Technology and Architecture Interaction

- Technology or Architecture: Contribution
  - Contribution to computer performance growth roughly equally between technology and architecture, with architecture credited with $\approx 80 \times$ improvement since 1985*

  *Danowitz, et al., “CPU DB: Recording Microprocessor History”, CACM 04/2012

- Technology and Architecture: Evolving Interaction
  - New technologies affect decision making by architects
  - Development in architecture impacts the viability of technologies

Computer Technology and Architecture: An Evolving Interaction

IEEE Computer, 09/1991

John L. Hennessy, Stanford University
Norman P. Jouppi, Digital Equipment Corporation

- Two technology trends:
  - Transistor scaling
  - Increasing memory density

- Two architecture trends:
  - Processor Architecture: Pipelining/ILP
  - Memory Architecture: Caching
The Next 15 Years in Computer Architecture Research?

“The best way to predict the future is to study the past”
- Robert Kiyosaki

- After Hennessy & Jouppi’s Summary in 1991, what was the trend since then?
- We studied the topics of each ISCA papers from 1992-2016

Computer Technology and Architecture: An Evolving Interaction

IEEE Computer, 09/1991

John L. Hennessy, Stanford University
Norman P. Jouppi, Digital Equipment Corporation
Topics in Components

- Memory architecture gains more importance since 2005
- Interconnect architecture since 2002 (NoC)
- GPU architecture since 2008
- Accelerator architecture since 2008

![Graph showing the development of processor, memory, accelerators, and interconnect from 1992 to 2016.](image)
Topics on Optimization Goals:

- ISCA 2000:
  - Wattch (Princeton) and SimplePower (PennState) (2000)
  - Transient fault detection via simultaneous multithreading (2000)
- Power/Reliability became major topics for architecture research since 2000
CMOS Technology May Not Scale Anymore

Street Dates for Intel’s Lead Generation Products

- 14nm slips by 2 quarters
- 10nm slips by 5-6 quarters
- 7nm by end 2020?

Moore’s Law Slope
(2x every 18mos)

Moore’s Crawl?
(2x every 36mos)

Technology Node (nm)

Courtesy David Brooks @ Harvard
Emerging Technologies

- Emerging Technologies other than traditional CMOS scaling may provide new opportunities for new architecture innovations
  - 3D die-stacking
  - Non-volatile memory
  - Nanophotonics
  - Quantum
Technology-Driven Architecture

- Technology and Architecture: Evolving Interaction
  - New technologies affect decision making by architects
  - Development in architecture impacts the viability of technologies

A Case Study on 3D Die-Stacking Architecture
Design Space Exploration of 3D Architectures

YUAN XIE
Pennsylvania State University

GABRIEL H. LOH
Georgia Institute of Technology

BRYAN BLACK
Intel Corporation

and

KERRY BERNSTEIN
IBM Corporation

As technology scales, interconnects have become a major source of power consumption for microprocessors. Increasingly, alternate ways of building modern microprocessors have been developed, where a stack of multiple device layers with direct vertical interconnects on the same chip. As fabrication of 3D integrated circuit and architectural techniques is imperative to explore this area, we give a brief introduction to 3D integration technology that can enable the adoption of 3D ICs, and present techniques for implementing single-core and multicore 3D processors.

Three-dimensional integration is an emerging fabrication technology that vertically stacks multiple integrated chips. The benefits include an increase in device density; much greater flexibility in routing signals, power, and clock; the ability to integrate disparate technologies; and the potential for new 3D circuit and microarchitecture organizations. This article provides a technical introduction to the technology and its impact on processor design. Although our discussions here primarily focus on high-performance processor design, most of the observations and conclusions apply to other microprocessor market segments.

3D integration technology overview

Although there are several candidate variants on 3D integration technology, at the heart of all of them is the vertical stacking of two or more individual integrated chips. (This article doesn't cover processes that "grow" multiple layers of devices such as wafer bonding. (See the "stack" sidebar for an example of multiple whole silicon wafers bonded with 3D integrated chips.)
Intel® 3D Pentium® 4 (ICCD 2004)

Source: Intel

Source: B.Black (Intel)
Design and Management of 3D Chip Multiprocessors Using Network-in-Memory

Feihui Li, Chrysostomos Nicopoulos, Thomas Richardson, Yuan Xie, Vijaykrishnan Narayanan, Mahmut Kandemir
Dept. of CSE, The Pennsylvania State University
University Park, PA 16802, USA
{feli,nicopoul,trichard,yuanxie,vijay,kandemir}@cse.psu.edu

Figure 1. A typical NoC mesh.

Figure 4. Proposed 3D Network-in-Memory architecture
Intel’s 3D +NOC Prototyping (2007)

80 Cores

SRAM

20MB 3D local memory for TFLOP performance
BW 12GB/s/tile @ full core clock (3GHz)
~1TB/s for TFLOP

heat sink
heat spreader
Polaris die
Freya die
LGA substrate
top metal
top metal
TSVs

Courtesy: T. Karnik (Intel)
Gabe Loh, Yuan Xie. “3D Stacked Microprocessor: Are We There Yet?”  
IEEE Micro, Volume 30 Issue 3, pp. 60-64, May. 2010
More and more transistors can be integrated into a single package

About 100MB-1GB on-package DRAM would be available

How to use these transistors efficiently?
- Multi-core, and many-core?
- Larger cache size or deeper cache hierarchy?
- On-package main memory?

X. Dong et al. “Simple but Effective Heterogeneous Main Memory with On-Chip Memory Controller Support” (SC 2010)
In-package 3D Memory with GPU

Conventional GDDRs, off-chip

Wide-bus routing on silicon interposer

3D-Stacked DRAMs
SMs and MCs
Silicon Interposer

Package Substrate

Top View

Side View

Die-Stacking is Happening

AMD Announcement on June 16, 2015

- The Fiji GPU Packaging is 50x50mm
- The interposer size is 26x32mm
- The GPU is about 20x24mm
- There are four 1GB HBM stacks for a total of 4GB of memory
Nvidia Pascal (3/2016)
Technology-Driven Architecture Innovation

- New technologies affect decision making by architects
- Development in architecture impacts the viability of technologies

Technology development:

- **1990-2004**: Single-core Fine-granularity uP (ISCA06, MICRO06)

- **2004**: Single-core Fine-granularity uP

- **2006**: Multi-core coarse-granularity uP (ISCA06, MICRO06)

- **2007**: Multi-core coarse-granularity uP (ISCA06, MICRO06)

- **2010**: 3D->2.5D (SC10)

- **2011**: 2.5D GPU (ISLPED12)

- **2012**: HMC

- **2015**: 3D GPU AMD Fury X
Emerging Non-volatile Memories

- Magnetic RAM (MRAM)
  - EverSpin (130nm, up to 16Mb)

- Spin-Torque-Transfer RAM (STTRAM)
  - Grandis (54nm, acquired by Samsung)

- Phase-Change RAM (PCRAM)
  - Samsung (20nm, diode, up to 8Gb)

- Resistive RAM (ReRAM)
  - Micron (16Gb, 27nm, ISSCC14)

- Intel 3D Xpoints (2016)
Architecture Opportunities with NVM

Opportunities:

- Leveraging NVM as LLC/Memory/Storage (HPCA 09-10, ISCA 11, 12, 16)
  - Performance and energy improvement
- Leveraging Nonvolatility for instant power-on/power-off (HPCA 15)
- Leveraging Nonvolatility for persistency Support (MICRO 13, MICRO 14)

Challenges:

- Wear-out, write-overhead, asymmetric read/write
Emerging Application Domains

- Mobile/embedded
- Data center
- AI/ML Application
The (Re)Rising of AI Applications

- Supercomputers: Drug design
- Data Centers: Automatic translation
- Smartphones: Audio recognition
- Embedded Devices: Business analytics, Robotics, Consumer electronics

Source: Yuji Chen (ICT)
Emerging Application + Emerging Technology

- When Emerging Application meets Emerging Technology -> Emerging Architecture

Welcome to Session 1A-3 (11:40am – 12:00pm)!

PRIME: A Novel Processing-in-memory Architecture for Neural Network Computation in ReRAM-based Main Memory

Ping Chi*, Shuangchn Li*, Tao Zhang†, Cong Xu‡, Jishen Zhaoδ, Yu Wang#, Yongpan Liu#, Yuan Xie*
Architecture-driven innovations are still the dominant themes in ISCA

Since 2000, there were increasing interests in

- **Technology-driven** architectural innovations: 3D, NVM, optical, Quantum etc.
- **Application-driven** architectural innovations: Datacenter, mobile, NN etc.
Neuromorphic → Neural Network → Large Graph Processing → VR/AR

Application-Driven

Heterogeneous Computing

GPU → ASIC → FPGA

Emerging Technology

3D Stacking

STT-RAM/ReRAM → PCM/Memristor

Intel HARP → Cambricon → TESLA P100

Heterogeneous Computing Emerging Technology

Heat Sink

Off-chip Optical Interconnects

TSVs

Processor/L1 Die

SRAM/DRAM memory

PCM/ReRAM memory

Optical Die

Laser

Logic Die 4 DRAMs

ullump interface for 2.5D or 3D
8 x 128b Channels

Architecture 2030 Workshop. 28 June 18, 2016

HP labs, 2012