

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 $\sim 80\times$ 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

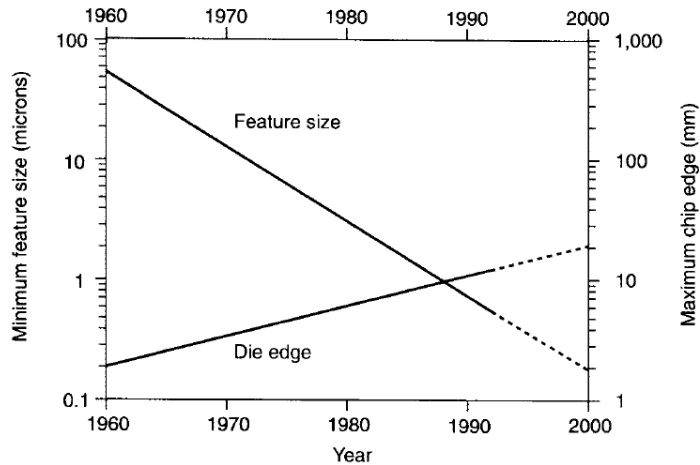
John L. Hennessy, Stanford University

Norman P. Jouppi, Digital Equipment Corporation

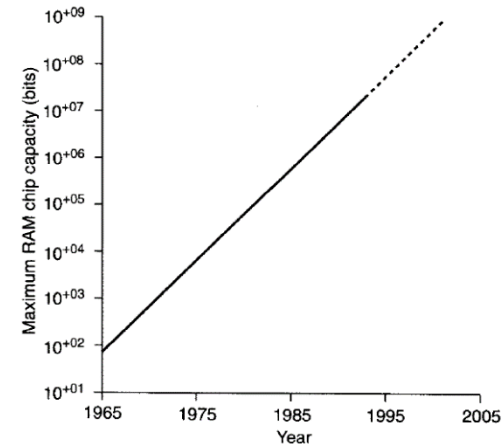
Technology and Architecture Interaction (1991)

Two technology trends:

Transistor scaling

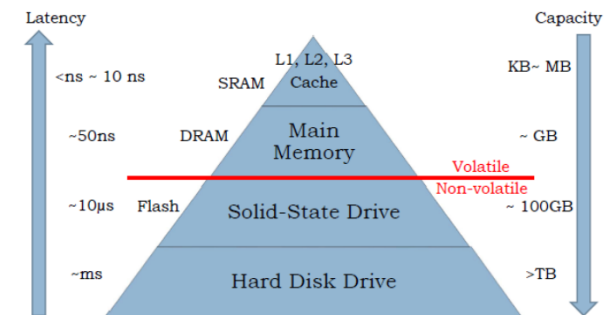
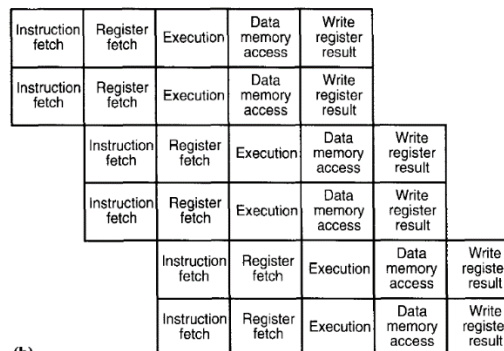
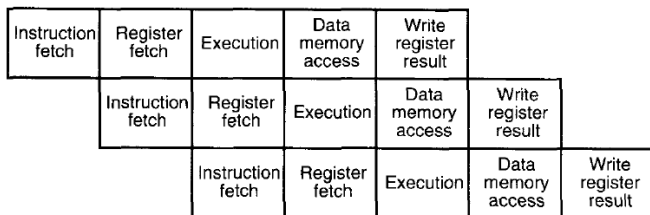


Increasing memory density



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

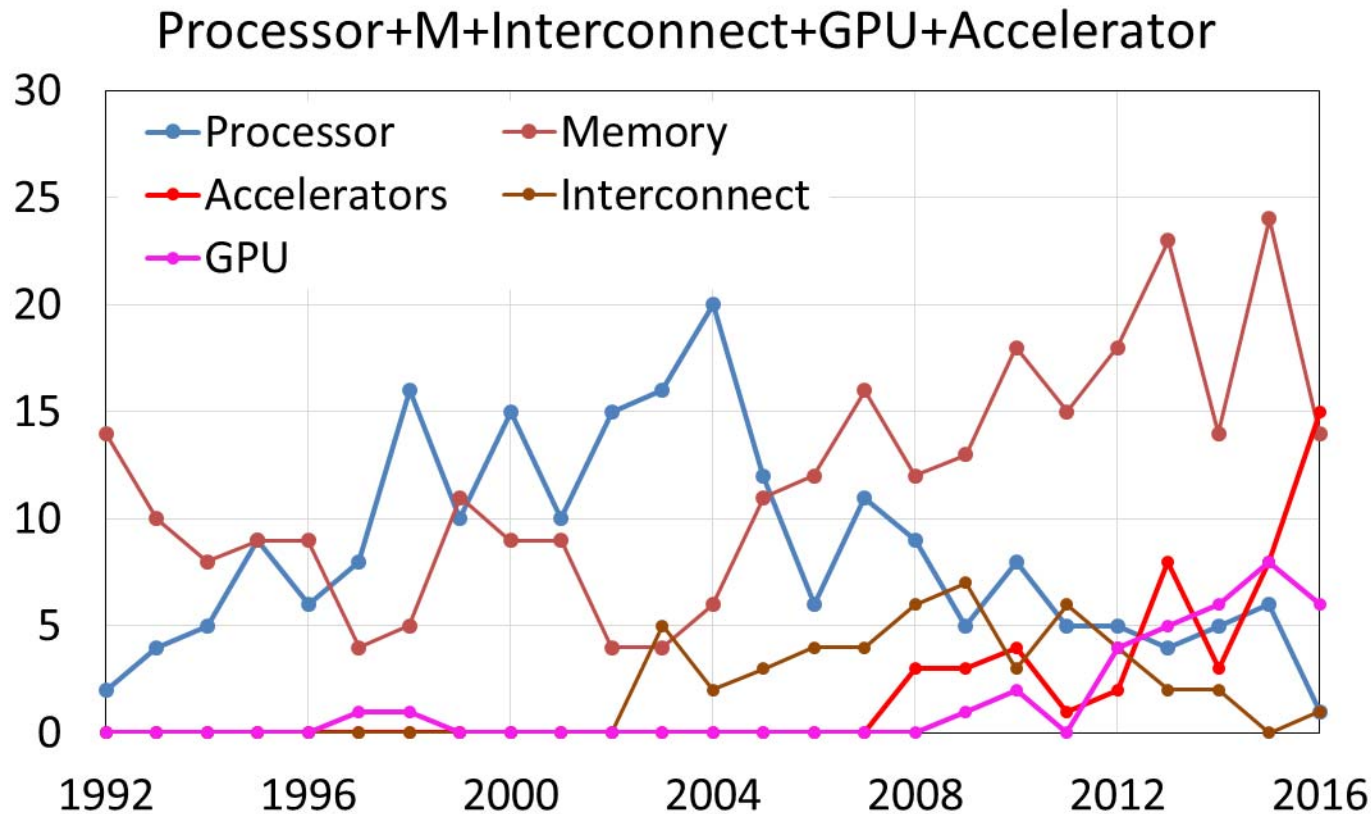
IEEE Computer, 09/1991

John L. Hennessy, Stanford University

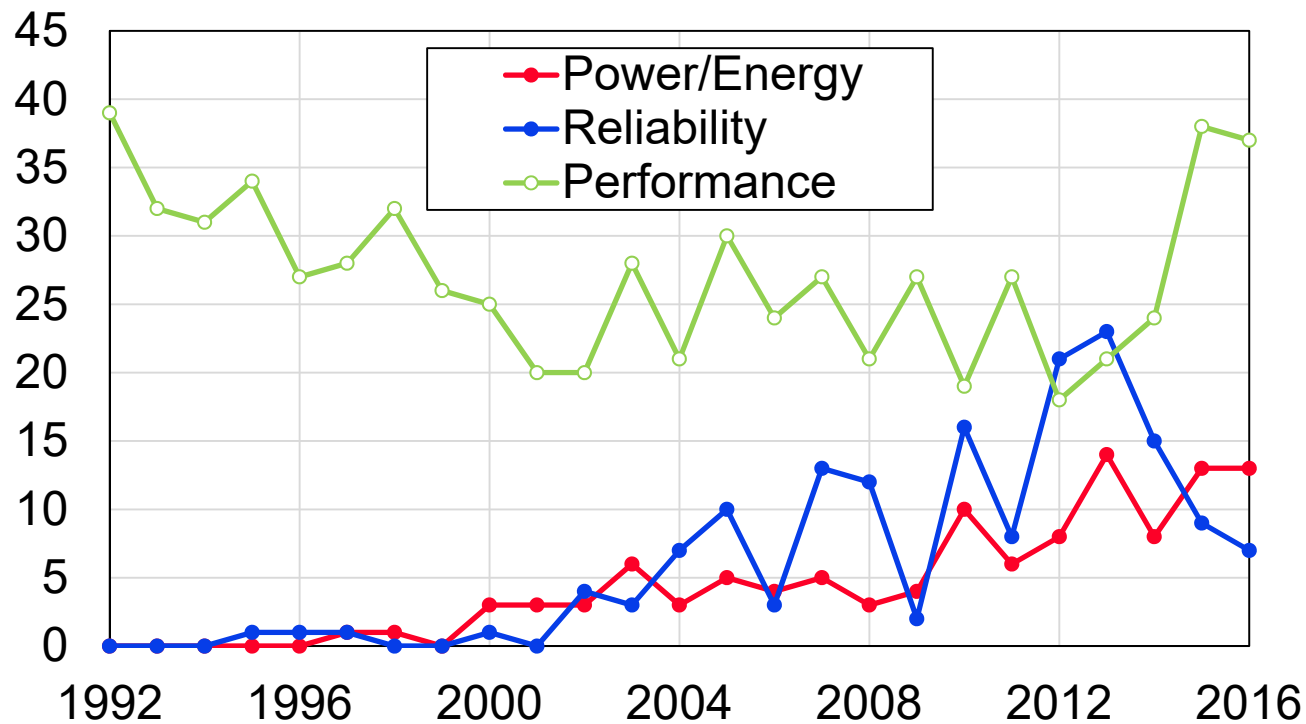
Norman P. Jouppi, Digital Equipment Corporation

Topics in Components

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



Topics on Optimization Goals:

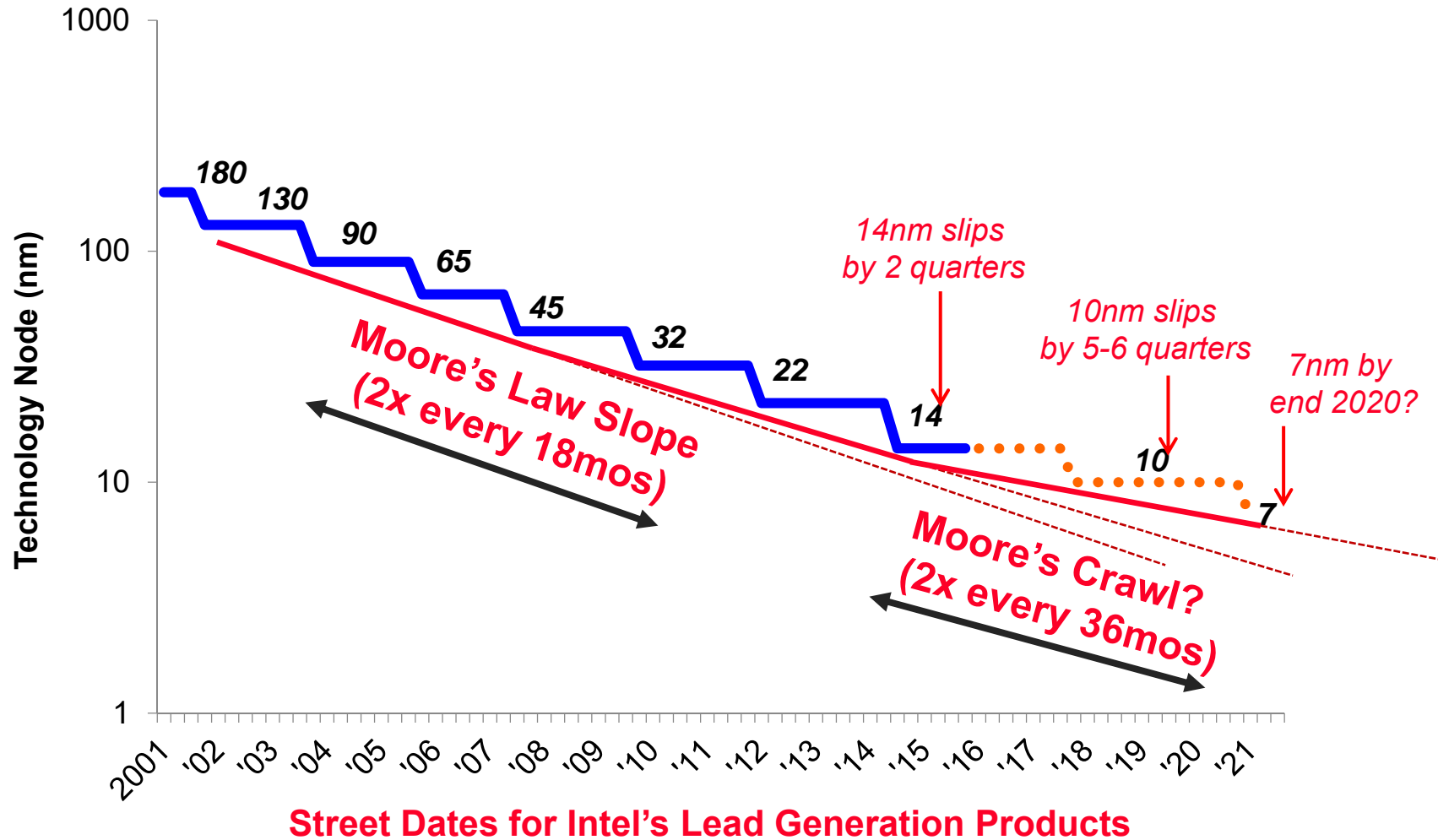


❑ ISCA 2000:

- Wattach (Princeton) and SimplePower (PennState) (2000)
- Transient fault detection via simultaneous multithreading (2000)

❑ Power/Reliability became major topics for architecture research since 2000

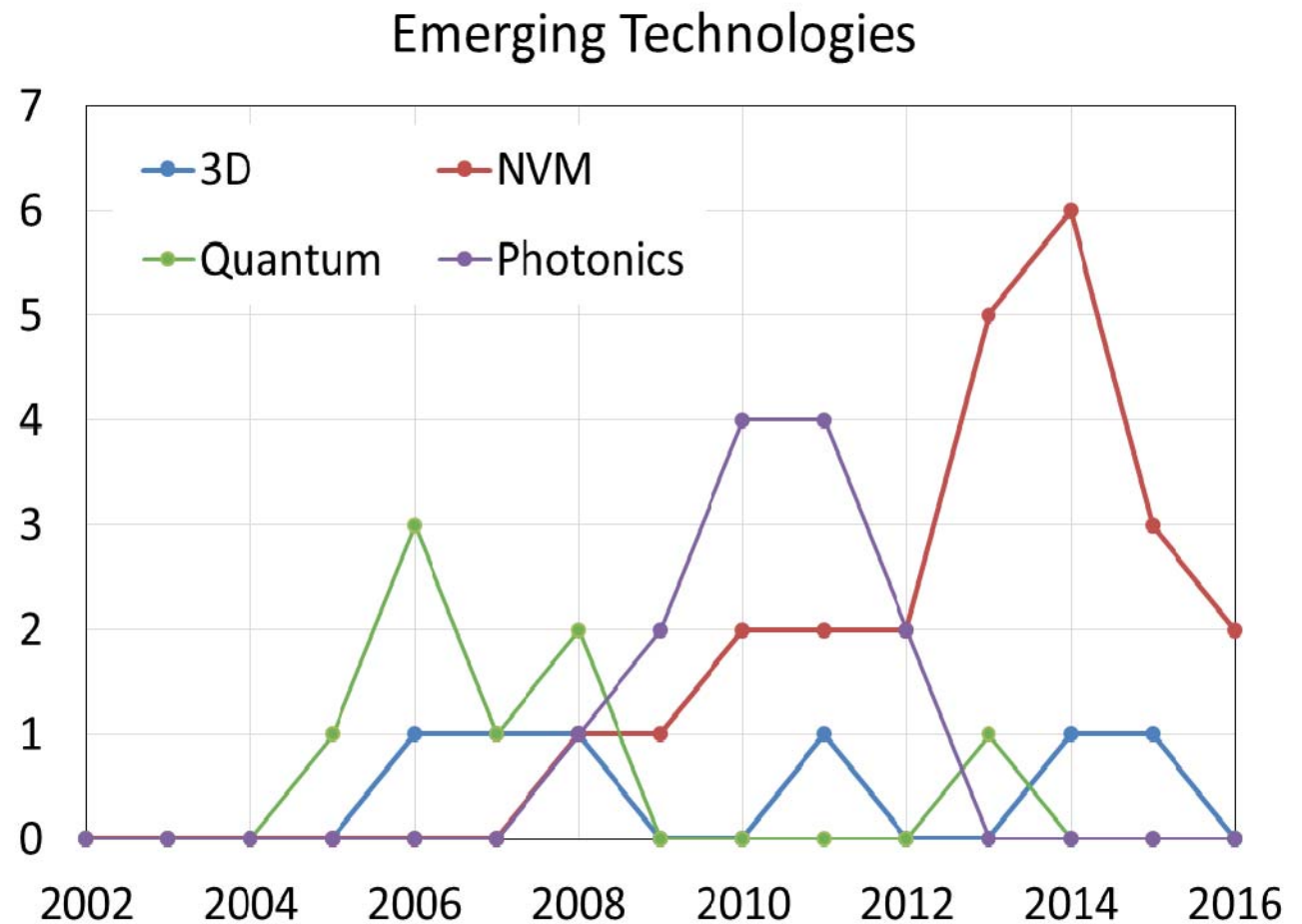
CMOS Technology May Not Scale Anymore



Courtesy David Brooks @ Harvard

Emerging Technologies

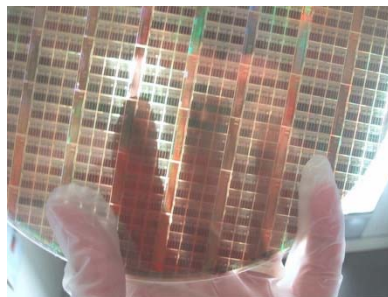
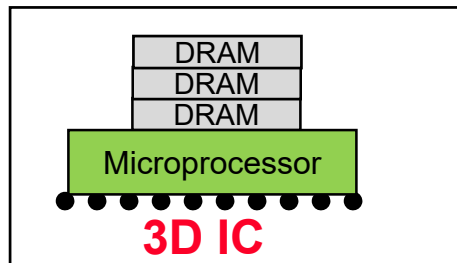
- ❑ Emerging Technologies other than traditional CMOS scaling may provide new opportunities for new architecture innovations
 - 3D die-stacking
 - Non-volatile memory
 - Nanophotonics
 - Quantum



Technology-Driven Architecture

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

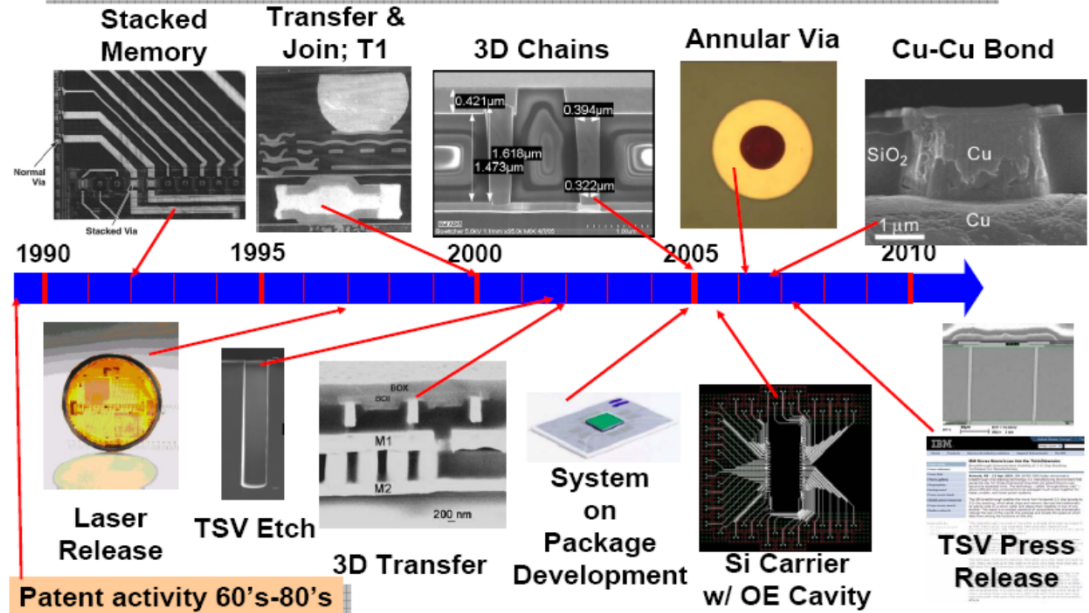
A Case Study on 3D Die-Stacking Architecture



2002/11/11

15 Years of IBM 3D Research

Many programs existed for extended periods before publication



Design Space Exploration 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 concern for power consumption for microprocessors. Increasingly, we consider alternate ways of building modern microprocessors where a stack of multiple device layers with direct vertical interconnects on the same chip. As fabrication of 3D integrated circuits and architectural techniques is imperative to explore this technology, in this article, we give a brief introduction to 3D integration technology that can enable the adoption of 3D ICs, and present three industrial case studies of design 3D microarchitectures.

PROCESSOR DESIGN IN 3D DIE-STACKING TECHNOLOGIES

THREE-DIMENSIONAL DIE-STACKING INTEGRATION STACKS MULTIPLE INTEGRATED CIRCUIT (IC) DIE ON A SINGLE SILICON WAFER. THIS TECHNOLOGY ENABLES THE DESIGN OF HIGH-DENSITY, LOW-LATENCY LAYERED ARCHITECTURES. AFTER PRESENTING A BRIEF BACKGROUND ON 3D DIE-STACKING TECHNOLOGY, THIS ARTICLE GIVES MULTIPLE CASE STUDIES ON DIFFERENT APPROACHES FOR IMPLEMENTING SINGLE-CORE AND MULTICORE 3D PROCESSORS AND DISCUSSES HOW TO DESIGN FUTURE MICROPROCESSORS GIVEN THIS EMERGING TECHNOLOGY.

..... 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 "stack" multiple layers of device such as

wafer bonding. (See the "stack" sidebar for an example of multiple whole silicon wafer 3D integrated chips.)

When considering the integration of two silicon dies, the topologies are face-to-face, where a die's "face" is metallized and its "back" is attached to the silicon substrate. The face-to-face bonding process builds the interconnects by depositing the copper on each die, and then being pressed together with a thermocompression process. A chemical-mechanical polishing process thins one die to reduce the distance for communication between the dies for external I/O and power

3D interconnects

From a processor design perspective, the most important interconnect

Gabriel H. Loh

Georgia Institute of

Technology

Yuan Xie

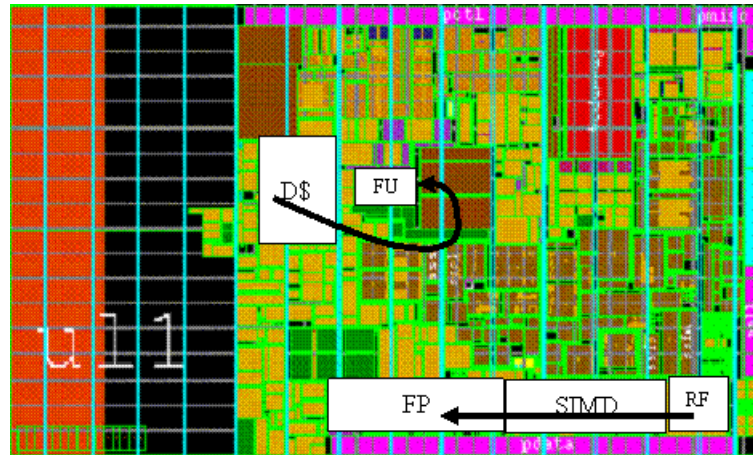
Pennsylvania State

University

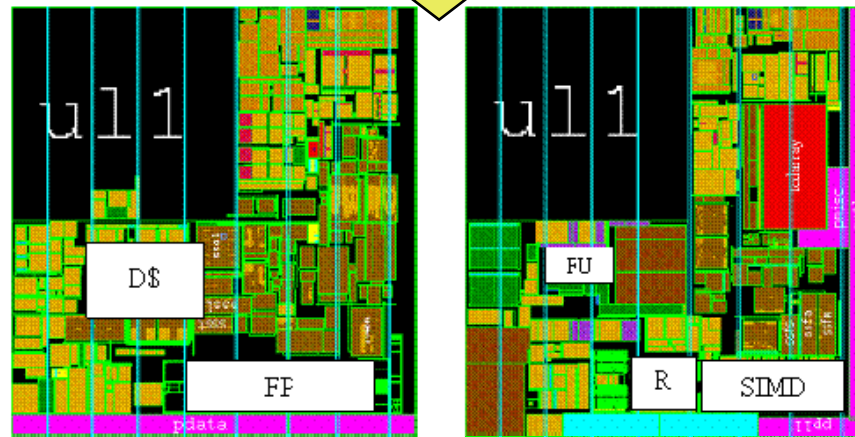
Bryan Black

Intel

Intel® 3D Pentium® 4 (ICCD 2004)



Source: Intel



Top

Bottom

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

ISCA 2006

{feli,nicopoul,trichard,yuanxie,vijay,kandemir}@cse.psu.edu

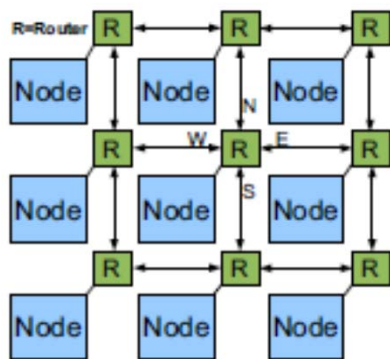


Figure 1. A typical NoC mesh.

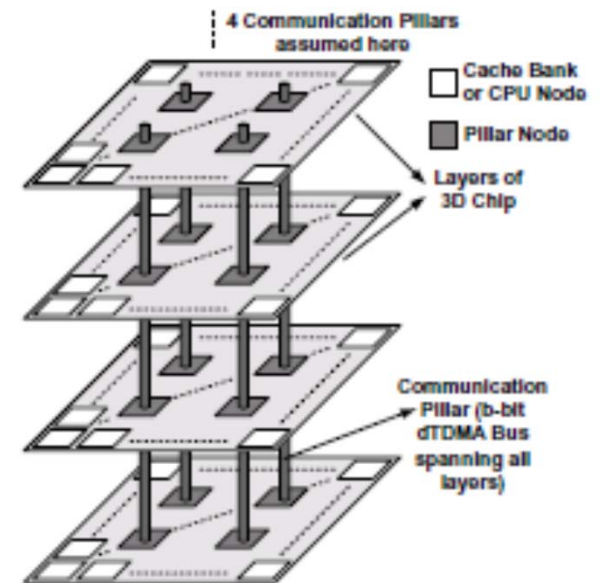
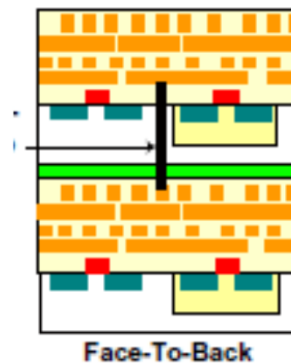
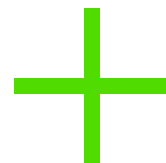



Figure 4. Proposed 3D Network-in-Memory architecture

Intel's 3D +NOC Prototyping (2007)



The image displays two micrographs on the left and a 3D schematic on the right. The first micrograph is labeled '80 Cores' and shows a dense grid of circuitry. The second is labeled 'SRAM' and shows a vertical strip of memory cells. The 3D schematic illustrates the stack: an LGA substrate at the base with TSVs (Through-Silicon Vias) connecting to a Freya die. Above the Freya die is a Polaris die, which is connected to a heat spreader and a heat sink. Labels for 'top metal' are also present, indicating the metal layers on the dies.

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



3D Stacked Microprocessor: Are We There Yet?

GABRIEL H. LOH

Georgia Institute of Technology

YUAN XIE

Pennsylvania State University

..... Three-dimensional integration has received considerable attention in the last several years from academic researchers and industry alike. This technology provides multiple layers of devices connected by a high-density, low-latency, layer-to-layer interface that can enable integrated circuits with more devices per unit area and allow the integration of different types of devices within the same 3D chip stack. Academic and

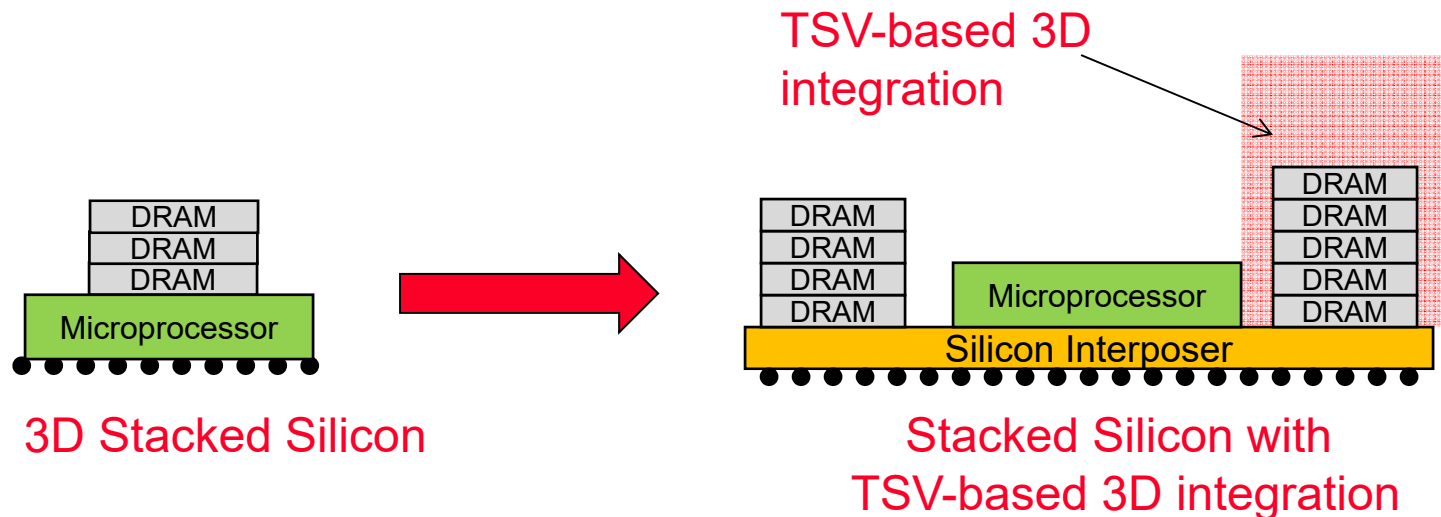
technological leaders from a range of institutions, including major semiconductor companies, government agencies, and industry consortia. (Most respondents answered our questions on condition of anonymity, and some chose not to reply at all due to concerns over confidentiality and exposure of proprietary information.) Their responses provide a view of where 3D integration technology for microprocessors currently stands,

Samsung, Tezzaron, and a few other companies have demonstrated, industry has reached the consensus that stacked memory will become mainstream. In this article, we focus on 3D stacking technology based on through-silicon-via (TSV) technology (see Figure 1b), which provides much faster and higher density inter-die connections than SiP or PoP.

The first question that many people are interested in is simply when TSV-

Gabe Loh, Yuan Xie. "3D Stacked Microprocessor: Are We There Yet?"
IEEE Micro, Volume 30 Issue 3, pp. 60-64, May. 2010

2D XPU+ 3D Memory = 2.5D Integration

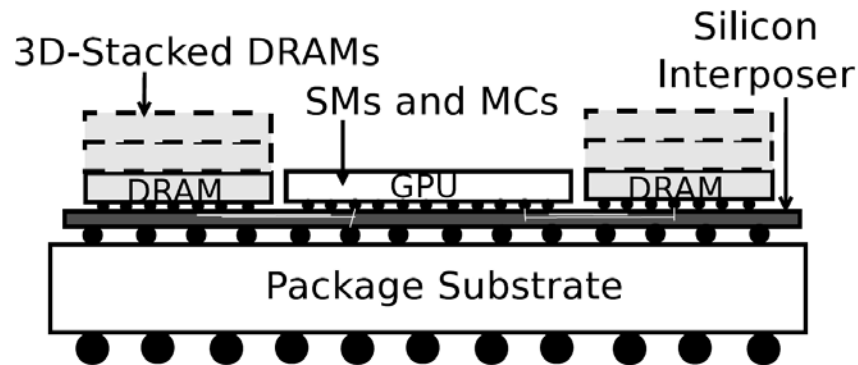
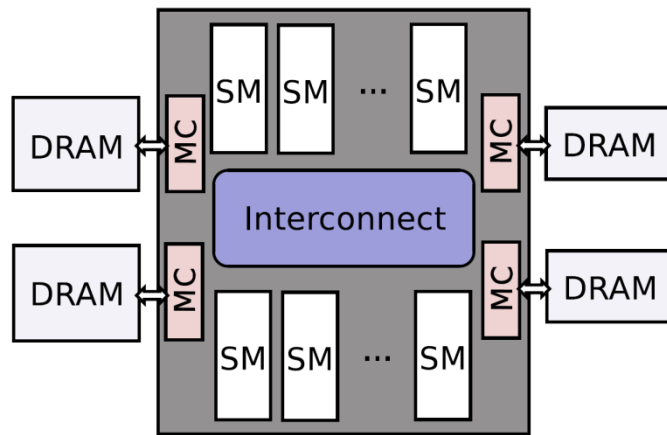


- ❑ More and more transistors can be integrated into a single package
- ❑ About 100MB-1GB on-package DRAM would be available
- ❑ How to use these transistors efficiently?
 - Multi-core, and many-core?
 - Larger cache size or deeper cache hierarchy?
 - On-package main memory?

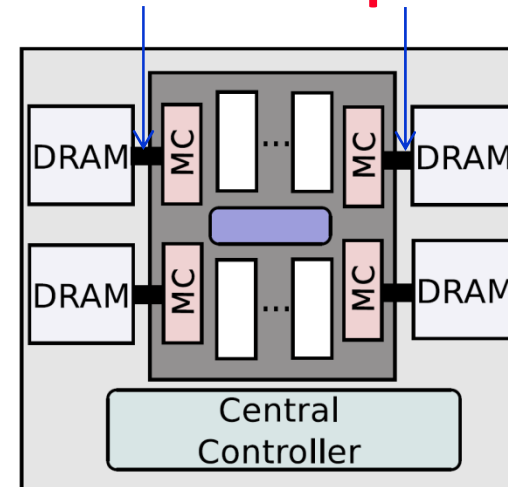
X. Dong et al. "Simple but Effective Heterogeneous Main Memory with On-Chip Memory Controller Support" (SC 2010)

In-package 3D Memory with GPU

Conventional GDDRs, off-chip



Wide-bus routing on silicon interposer



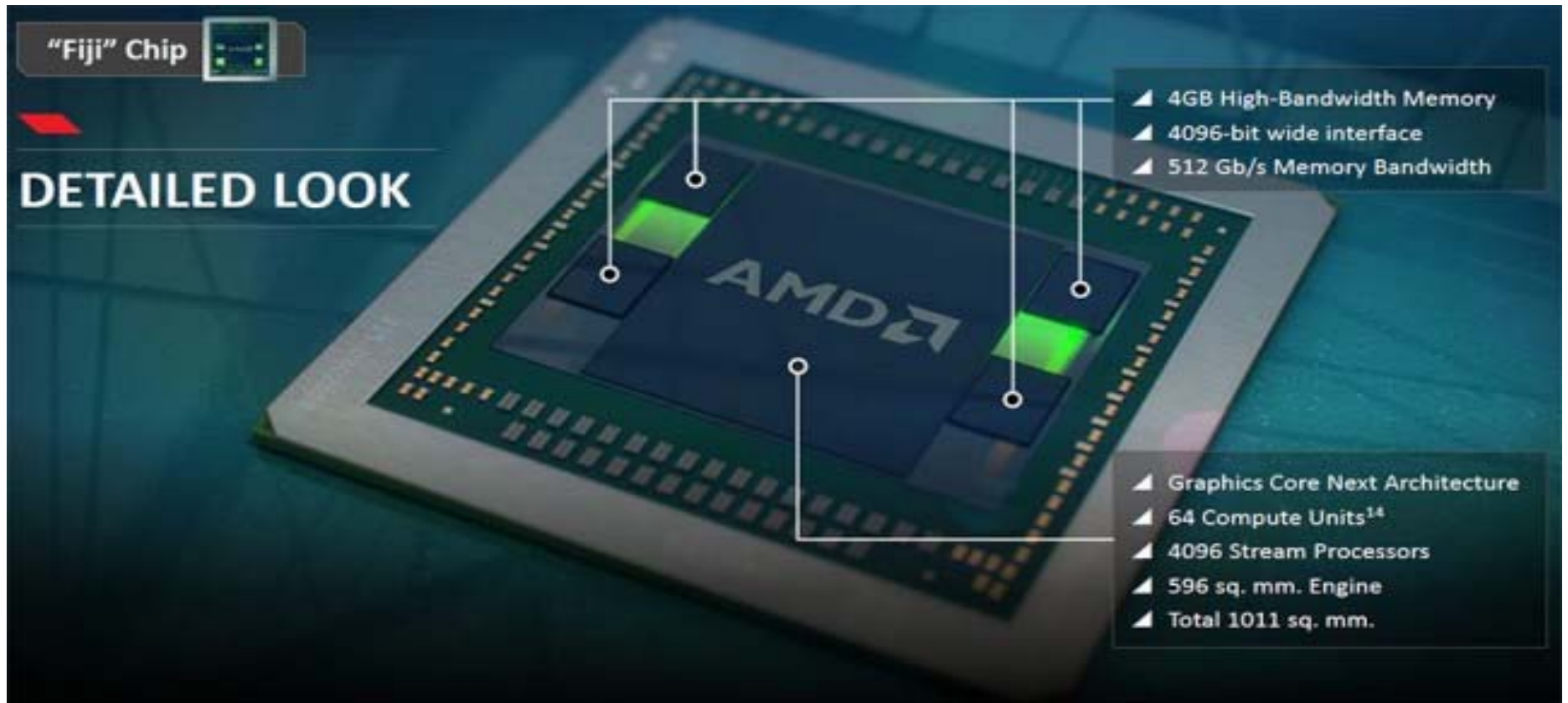
Top View

Side View

[Optimizing GPU Energy Efficiency with 3D Die-stacking Graphics Memory and Reconfigurable Memory Interface.](#) Jishen Zhao, Yuan Xie, Gabe Loh, *ISLPED 2012*.

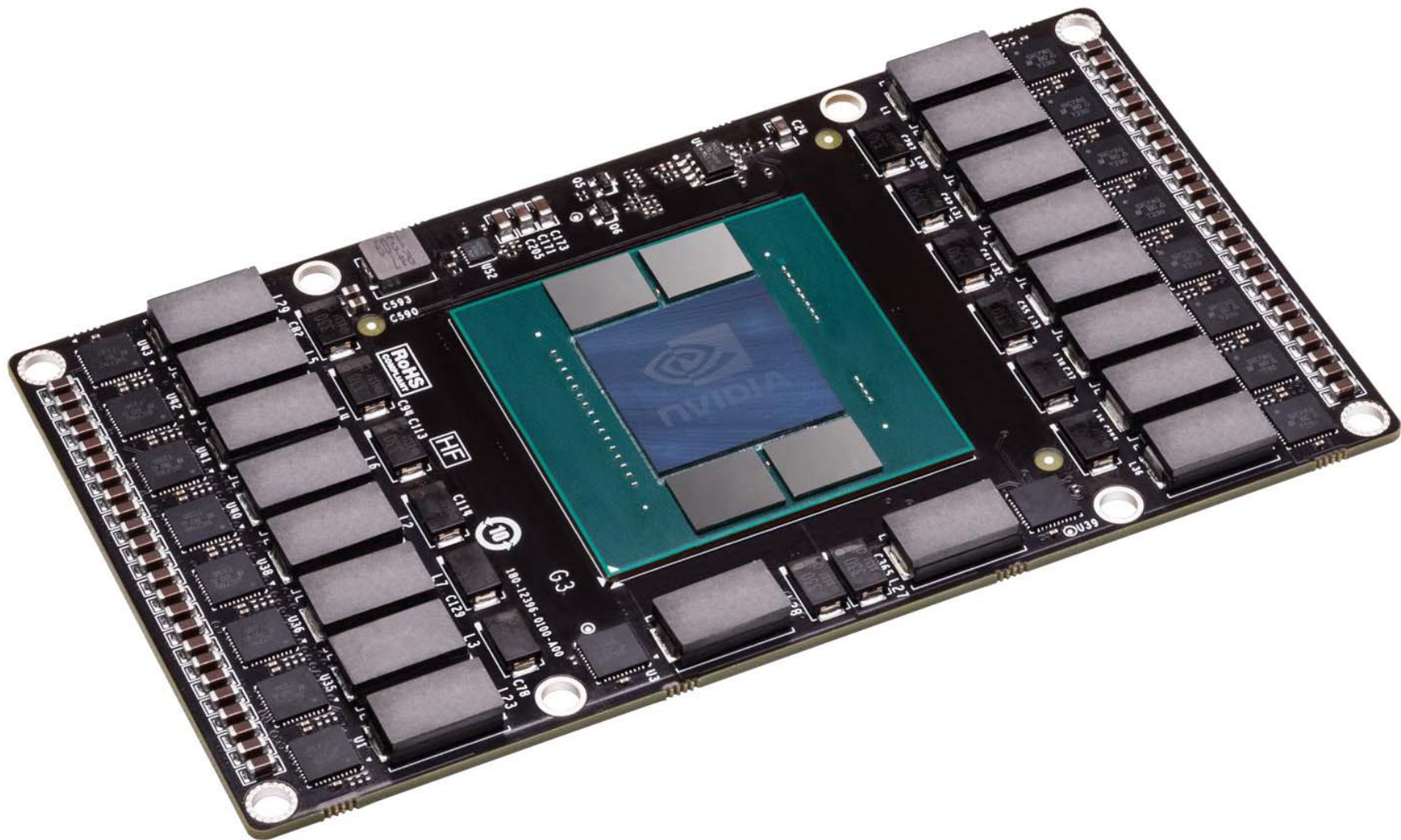
Die-Stacking is Happening

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)



Knights Landing

Holistic Approach to Real Application Breakthroughs



Platform Memory

NEW

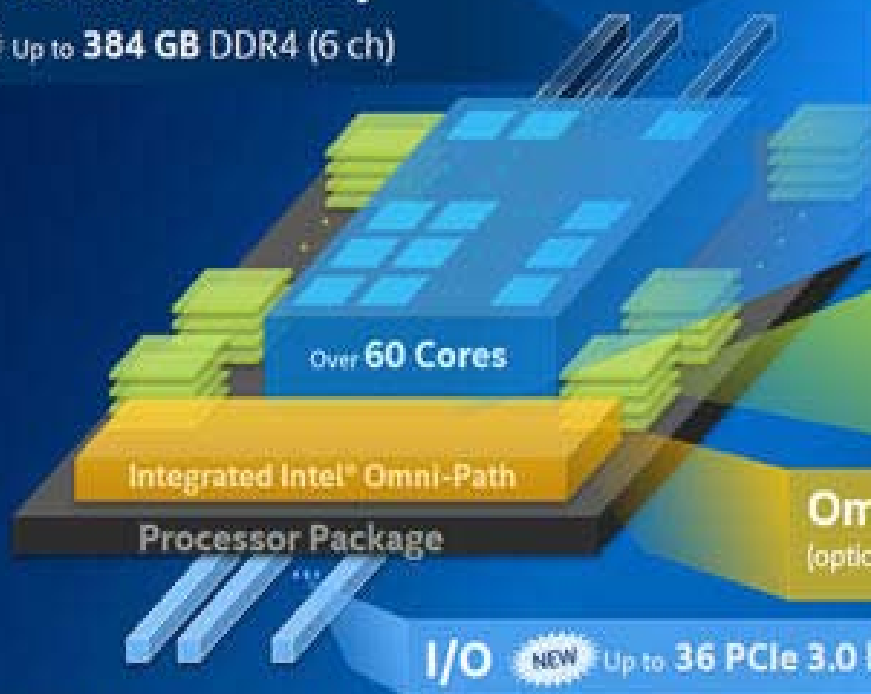
Up to **384 GB** DDR4 (6 ch)

Compute

- Intel® Xeon® Processor Binary-Compatible
- **3+ TFLOPS¹, 3X ST¹** (single-thread) perf. vs KNC
- **2D Mesh** Architecture
- **Out-of-Order** Cores

On-Package Memory

- Over **5x** STREAM vs. DDR4³
- Up to **16 GB** at launch



Over **60 Cores**

Integrated Intel® Omni-Path

Processor Package

Omni-Path

(optional)

- **1st** Intel processor to integrate

I/O

NEW

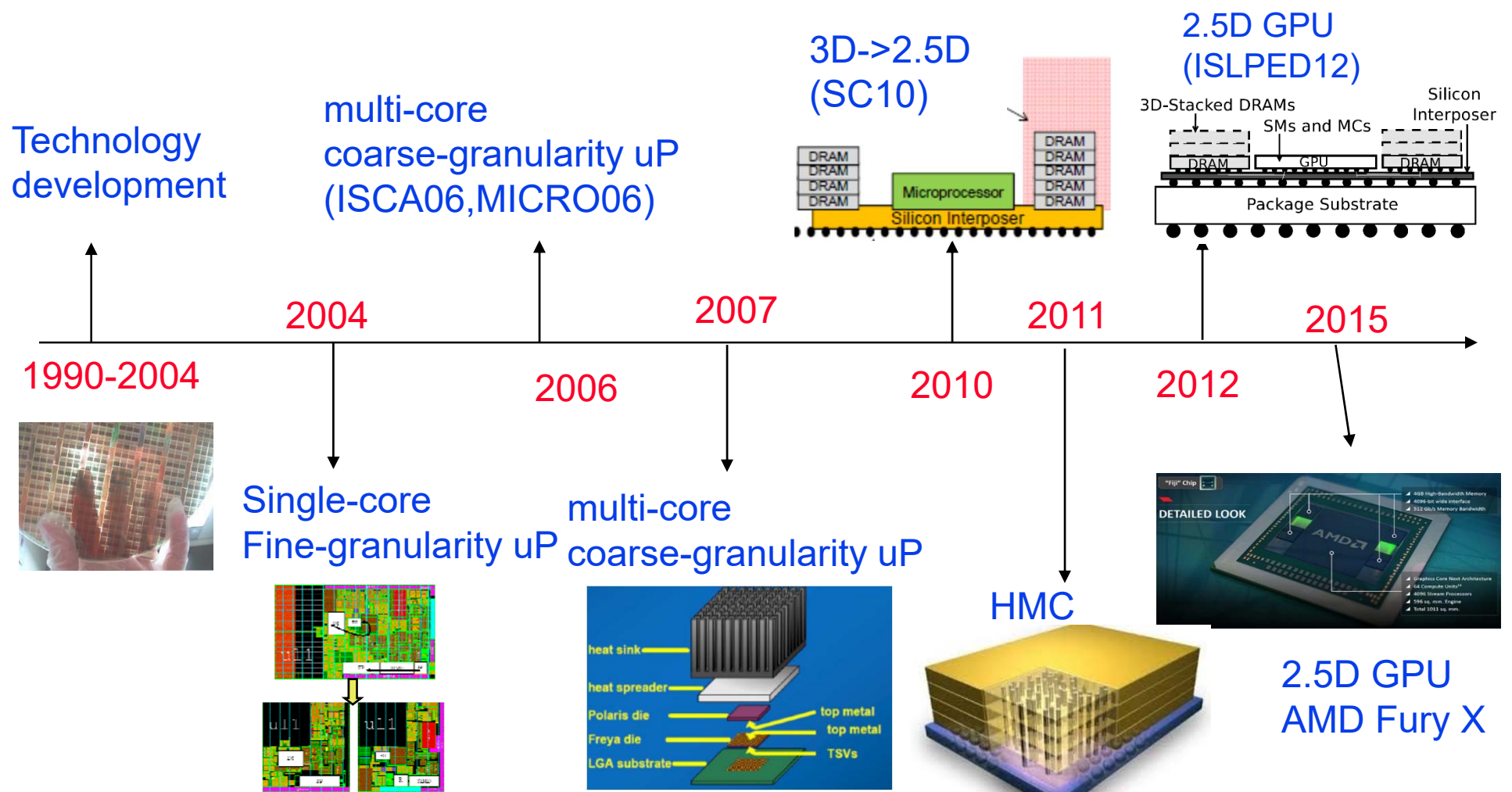
Up to **36 PCIe 3.0** lanes

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests are measured using specific computer systems, components, software, operations and functions. Any change to any of these factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchase, including the performance of that product when combined with other products. For more complete information visit <http://www.intel.com/performance>.



Technology-Driven Architecture Innovation

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



Emerging Non-volatile Memories

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

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

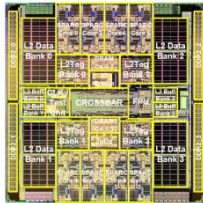
- ❑ Phase-Change RAM (PCRAM)
 - Samsung (20nm, diode, up to 8Gb)

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

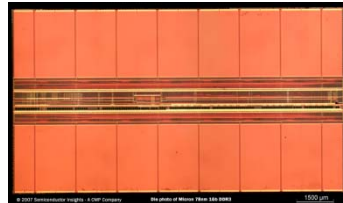
- ❑ Intel 3D Xpoints (2016)



Architecture Opportunities with NVM



On-chip memory
(SRAM, MRAM)



Off-chip memory
(DRAM, PRAM, ReRAM)



Solid State Disk
(Flash Memory, PRAM, ReRAM)



Secondary Storage
(HDD)

Opportunities:

- ❑ Leveraging NVM as LLC/Memory/Storage (HPCA 09-10, ISCA11,12,16)
 - Performance and energy improvement
- ❑ Leveraging Nonvolatility for instant power-on/power-off (HPCA15)
- ❑ Leveraging Nonvolatility for persistency Support (MICRO13, MICRO14)

Challenges:

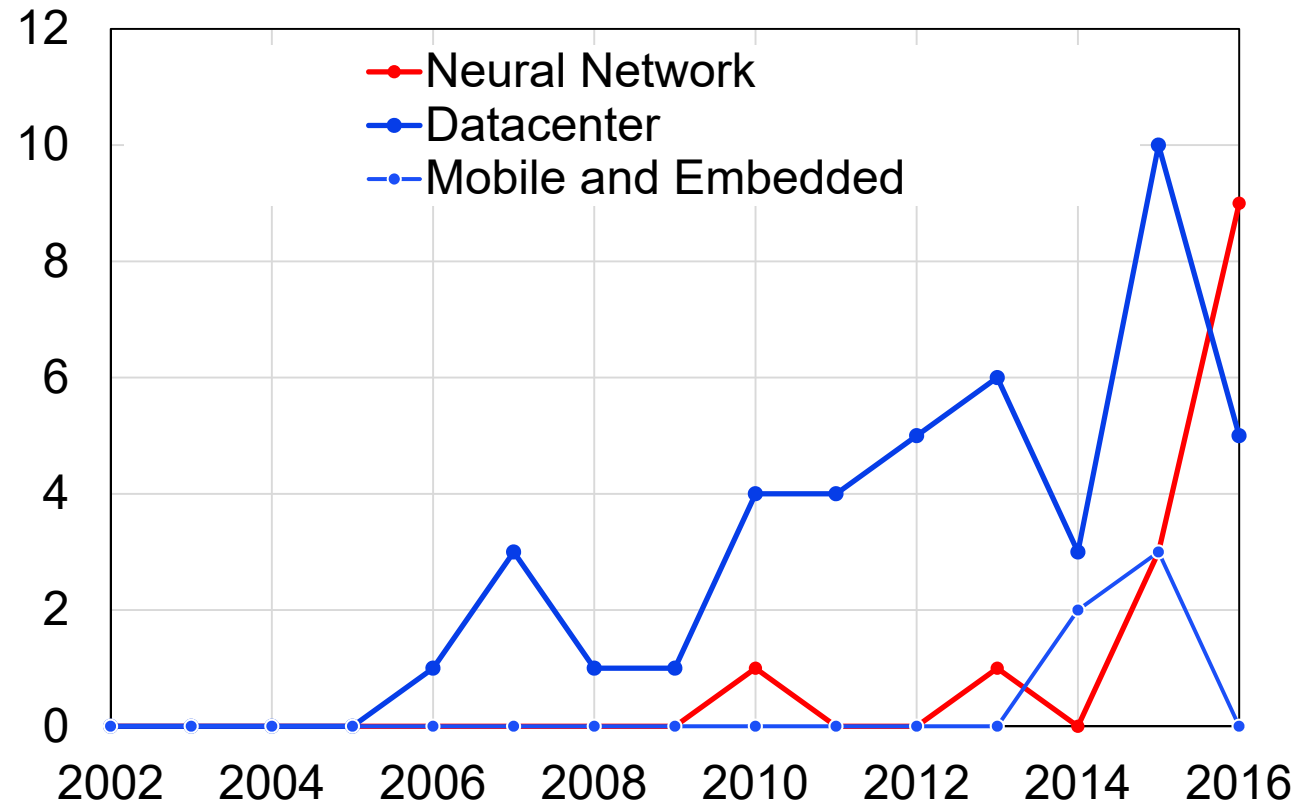
- ❑ Wear-out, write-overhead, asymmetric read/write

Emerging Application Domains

❑ Emerging application domains

- Mobile/embedded
- Data center
- AI/ML Application

Emerging Applications

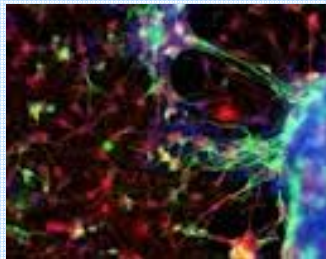


The (Re)Rising of AI Applications

Supercomputers



Business analytics



Drug design

Data Centers



Automatic translation



Smartphones



Audio recognition

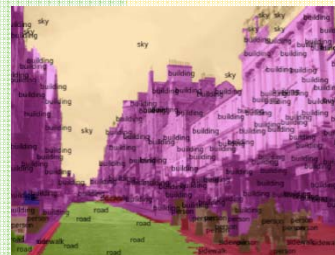
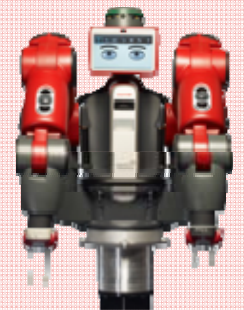


Image analysis

Embedded Devices



Robotics



Consumer electronics

Emerging Application + Emerging Technology

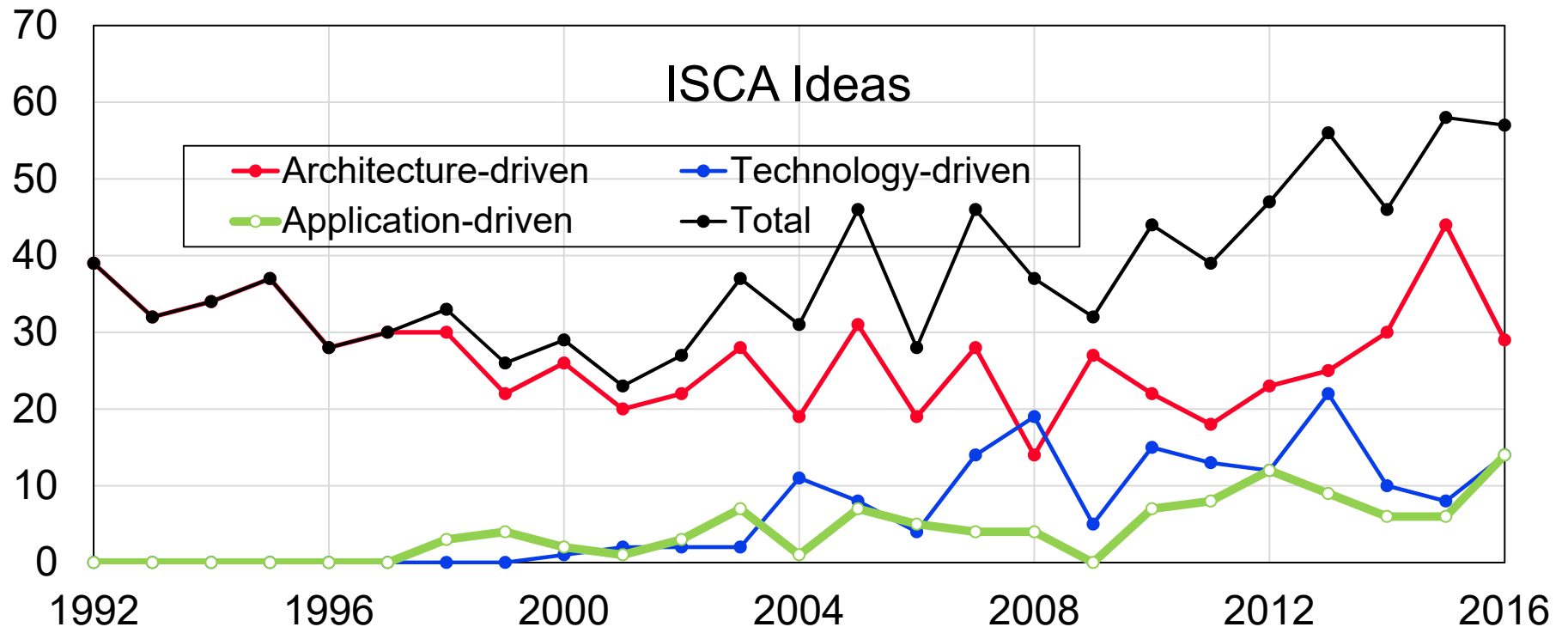
- **When Emerging Application meets Emerging Technology -> Emerging Architecture**

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 Zhao^δ, Yu Wang[#], Yongpan Liu[#], Yuan Xie^{*}

Overall Statistics for ISCA 1992-2016



- ❑ **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.



Data Center

AI

Mobile/Embedded

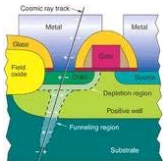
Application-Driven Innovations

Computer Architecture Innovations

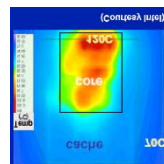
Technology scaling

Technology-Driven Innovations

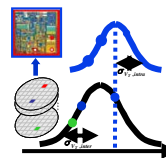
Emerging Technologies



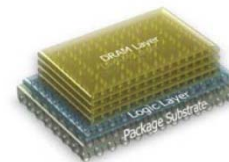
**Soft Errors
NBTI Aging**



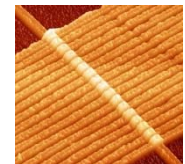
Power/Thermal



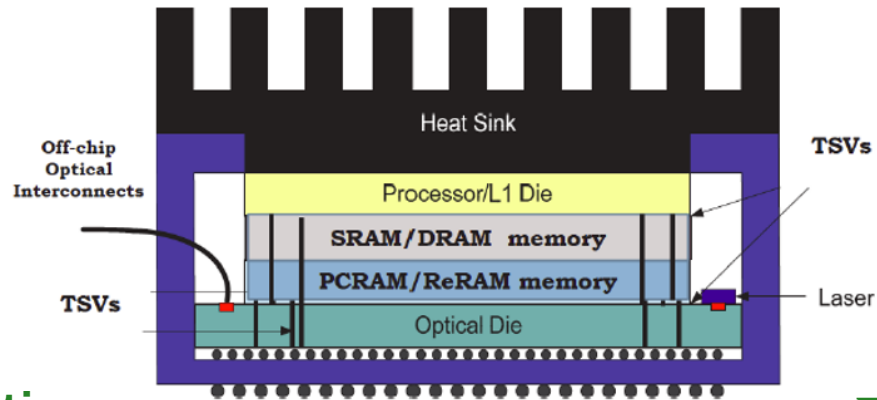
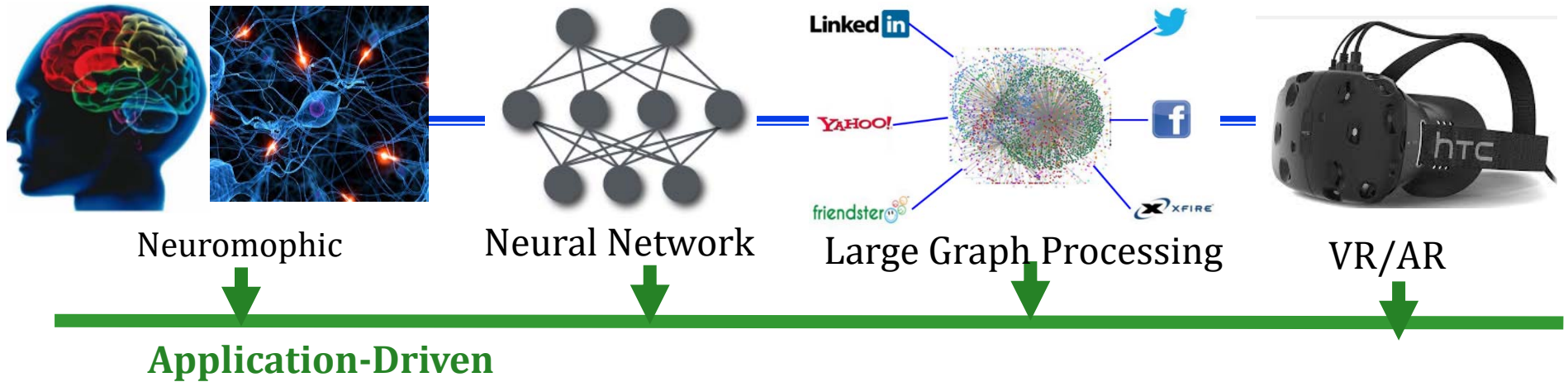
PVT Variations



3D Integration



Emerging NVM



Heterogeneous Computing

Emerging Technology

GPU

ASIC

FPGA

TESLA P100

Cambricon 寒武纪

Intel HARP

Architecture 2030 Workshop.28

3D Stacking

HP labs, 2012

STT-RAM/ReRAM

PCM/Memristor

Logic Die 4 DRAMs

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