Technology-Driven and Application-Driven Architecture Innovations: Past, Present, and Future

Yuan Xie

UCSB

Architecture 2030 Workshop @ ISCA'16

8:30	Intro remarks by Luis Ceze and Tom Wenisch
8:50	Mark Hill (Wisconsin) on "21st Century Computer Architecture"
9:10	Tom Conte (GeorgiaTech) on "IEEE Rebooting Computing Initiative & International Roadmap of Devices and Systems"
9:30	Devices Keynote: Philip Wong (Stanford) on "Device Technologies for the N3XT 1,000X Improvement in Computing Performance"
10:30	***Break***
11:00	Steve Keckler (nVidia/UT Austin) on "The Influence of Academic Research on Industry R&D"
11:25	Michael Taylor (UCSD) on "Open Source HW: Architecture's Only Hope for Survival"
11:45	Alvy Lebeck (Duke) on "Computing and Biomolecules"
12:05	Yuan Xie (UCSB) on "Technology-driven Architecture Innovation: Challenges and Opportunities"
12:30	***Lunch***
14:00	Applications Keynote: Kayvon Fatahalian (CMU) on "100 Quadrillion Live Pixels: The Challenge of Continuously Interpreting, Organizing, and Generating the World's Visual Information"

The Goal of the Workshop

- Community Efforts on the Vision for Computer Architecture research for the next 15 years

- Why now? A lot has changed in the last 5 years
  - Technology scaling is slowing down (Moore's law is dying)
  - Emerging technologies are getting mature.
  - Deep neural networks "caught us by surprise", machine learning now a key workload
  - Major platforms emerged (cloud, IoT, etc)
  - Vertical integration (systems companies)
  - Explosion of data (e.g., 1 trillion photos uploaded in 2015, genomics growing fast)

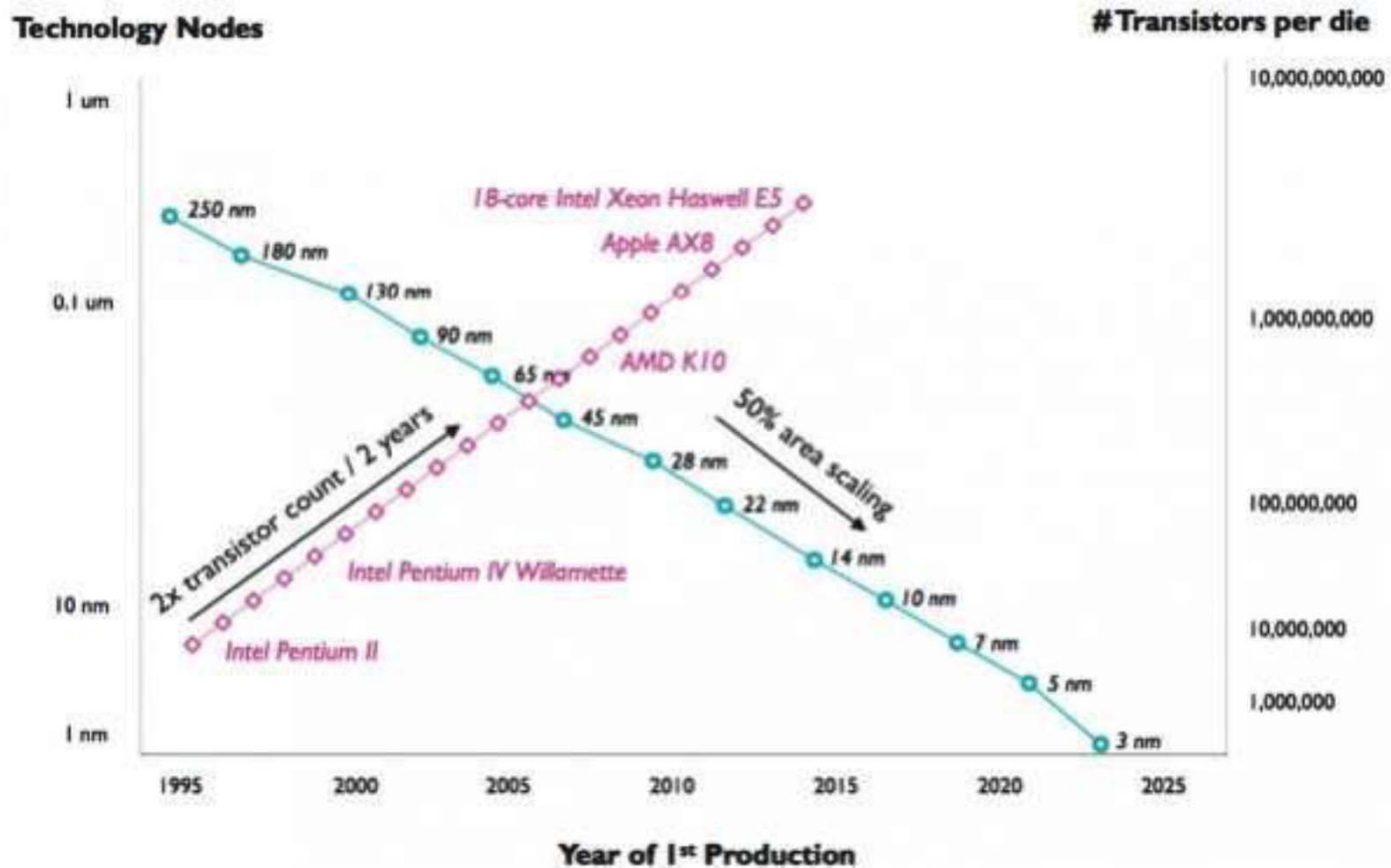
The Outcome of the Workshop

Arch2030: A Vision of Computer Architecture Research over the Next 15 Years

Summary ....
The Specialization Gap: Democratizing Hardware Design......
The Cloud as an Abstraction for Architecture Innovation......
Going Vertical ....
Architectures "Closer to Physics" ....
Machine Learning as a Key Workload ....
About this document ....

Technology Scaling: Key Contributor

- Technology scaling has been the key contributor to the performance improvement in microprocessor



Source: IMEC
An Steegen, 2015

Technology or Architecture Innovation?

- **Technology or Architecture**: Whose contribution is more significant for micprocessor performance improvement?
  - Contribution to computer performance growth roughly **equally** between technology and architecture *

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

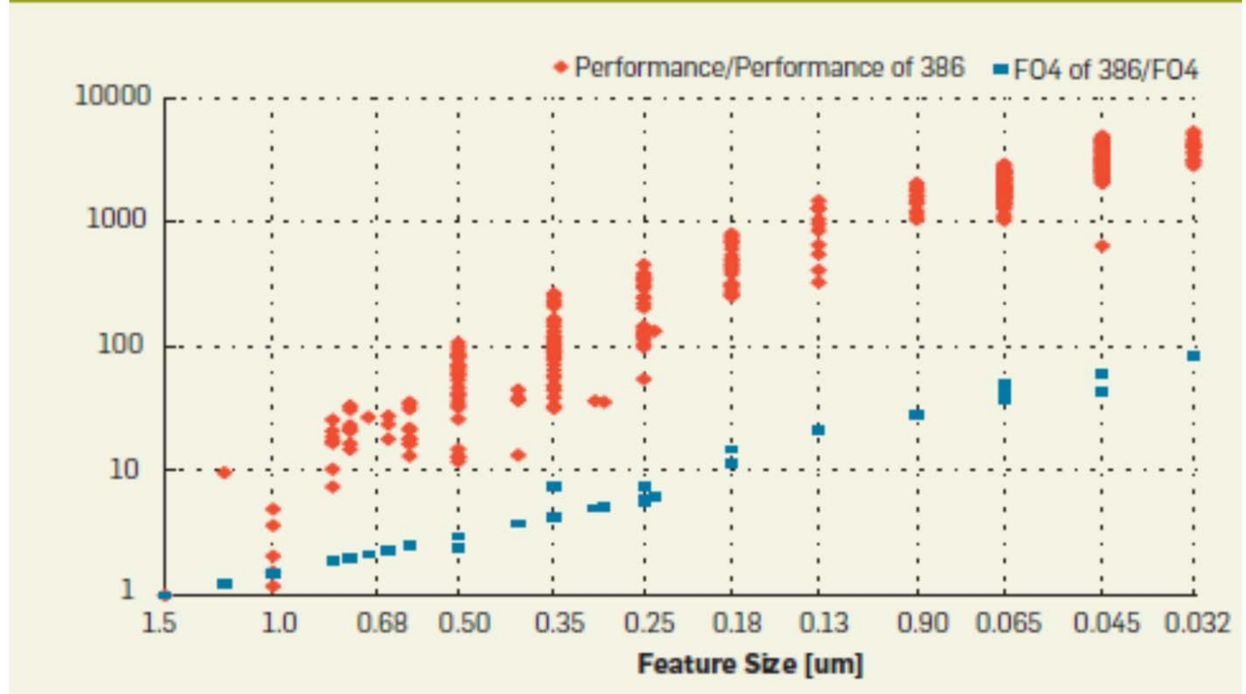
With this open database, you can mine microprocessor trends over the past 40 years.

BY ANDREW DANOWITZ, KYLE KELLEY, JAMES MAO, JOHN P. STEVENSON, AND MARK HOROWITZ

CPU DB: Recording Microprocessor History

IN NOVEMBER 1971, Intel introduced the world's first single-chip microprocessor, the Intel 4004. It had 2,300 transistors, ran at a clock speed of up to 740KHz, and delivered *60,000* instructions per second while dissipating 0.5 watts. The following four decades witnessed exponential growth in compute power,

Figure 1. The diamonds indicate how processor performance actually scaled with time, while the squares denote how much speedup came from improving the manufacturing process.



Technology or Architecture Innovation?

- **Technology or Architecture**: Whose contribution is more significant for micprocessor performance improvement?
  - Contribution to computer performance growth roughly **equally** between technology and architecture *

*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

John L. Hennessy, Stanford University

Norman P. Jouppi, Digital Equipment Corporation

IEEE Computer, 09/1991

3D Integration As a New Dimension of Scalability

Going Vertical

3D integration provides a new dimension of scalability.

A critical consequence of the end of Moore's Law is that chip designers can no longer scale the number of transistors in their designs "for free" every 18 months. Furthermore, over recent Silicon generations, driving global wires has grown increasingly expensive relative to computation, and hence interconnect accounts for an increasing fraction of the total chip power budget.

3D integration offers a new dimension of scalability in chip design, enabling the integration of more transistors in a single system despite an end of Moore's Law, shortening interconnects by routing in three dimensions, and facilitating the tight integration of heterogeneous manufacturing technologies. As a result, 3D integration enables greater energy efficiency, higher bandwidth, and lower latency between system components inside the 3D structure.

by 2021, to be replaced by 3D ...

e replaced by 3D ... 2015 when we noted the

PUs aren't expected until the 2021 ...

-by-2021-to-be... 2016-7-27

by 2021, to be replaced by 3D ...

e replaced by 3D integration. ... We discussed

ed for a Moore's law 3.0, ...

垂直探索

三维（3D）集成提供了一个新的可扩展维度。

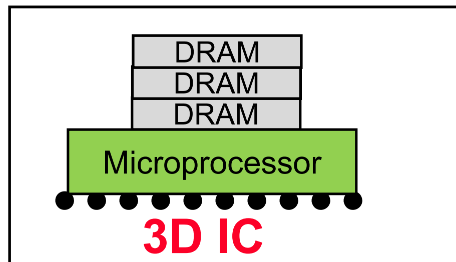
摩尔定律的终结带来的一个严重的后果是，芯片设计师再也不能“无偿地”每 18 个月将自己设计中的晶体管数量提高一倍。与此同时，近几代芯片中，驱动全局总线的开销已超过计算开销，并且不断快速增长，使得互连在芯片功耗预算中所占比例不断提升。

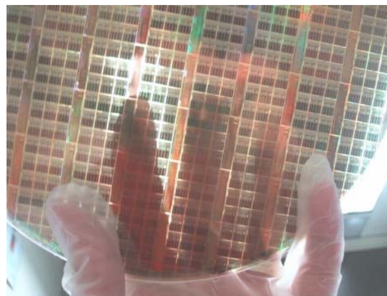
3D 集成为芯片设计提供了一个全新的维度，使得我们突破摩尔定律终结的阴霾，可以在单系统上继续集成更多的晶体管，从 3 个维度缩减互联开销，实现异构工艺技术紧密集成在一起。因此，3D 集成使系统组件在 3D 结构下能效更高、

Technology-Driven Architecture

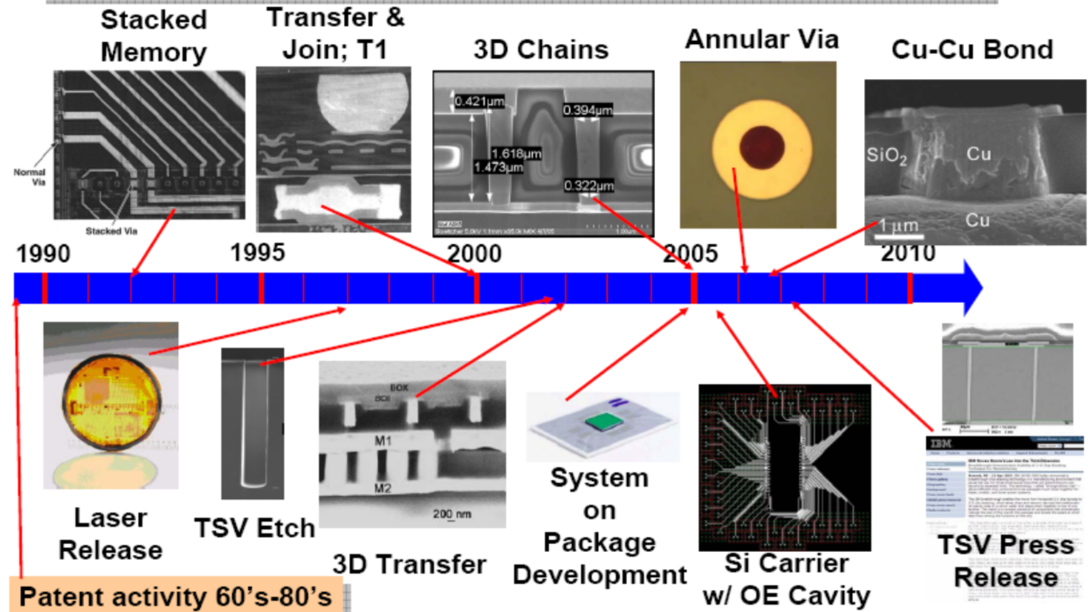
- Technology and Architecture：Evolving Interaction
  - **New technologies affect decision making by architects**
  - **Development in architecture impacts the viability of technologies**

A Case Study on 3D Die-Stacking Architecture





2002/11/11



Design Space Exploratio
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 majo
of power consumption for microprocessors. Increasing
sider alternate ways of building modern microprocesso
where a stack of multiple device layers with direct verti
on the same chip. As fabrication of 3D integrated circu
and architectural techniques is imperative to explore t
this article, we give a brief introduction to 3D integrati
that can enable the adoption of 3D ICs, and present th
components using 3D technology. An industrial case
design 3D microarchitectures.

PROCESSOR DESIGN IN 3D
DIE-STACKING TECHNOLOGI

THREE-DIMENSIONAL DIE-STACKING INTEGRATION STACKS MULTIPLE L
PROCESSED SILICON WITH A VERY HIGH-DENSITY, LOW-LATENCY LAYE
INTERCONNECT. AFTER PRESENTING A BRIEF BACKGROUND ON 3D DI
TECHNOLOGY, THIS ARTICLE GIVES MULTIPLE CASE STUDIES ON DIFFE
FOR IMPLEMENTING SINGLE-CORE AND MULTICORE 3D PROCESSORS A
HOW TO DESIGN FUTURE MICROPROCESSORS GIVEN THIS EMERGING T

Gabriel H. Loh
Georgia Institute of Technology

Yuan Xie
Pennsylvania State University

Bryan Black
Intel

•••••• 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

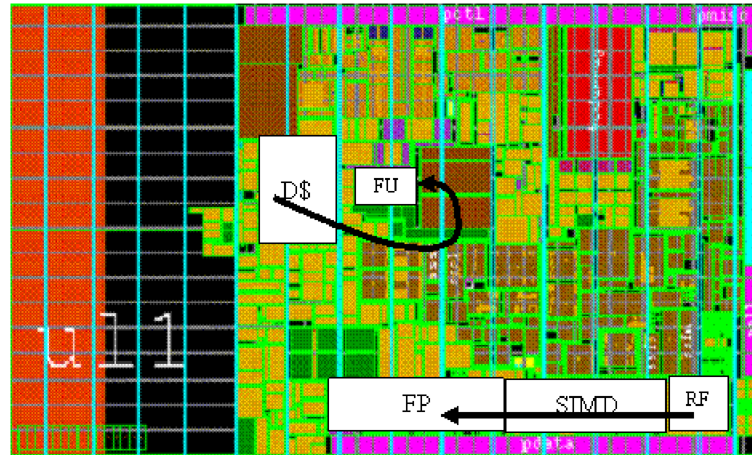
wafer bonding. (See the '
stack" sidebar for an e
multiple whole silicon wa
3D integrated chips.)

When considering th
of two silicon dies, the
topologies: face to face
where a die's "face" is
metallization and its "ba
the silicon substrate.
bonding process builds
tion, also called a *die-to-*
by depositing the coppe
on each die, and then bo
together with a thermoco
A chemical-mechanical
thins one die to reduc
communication between
for external I/O and pow

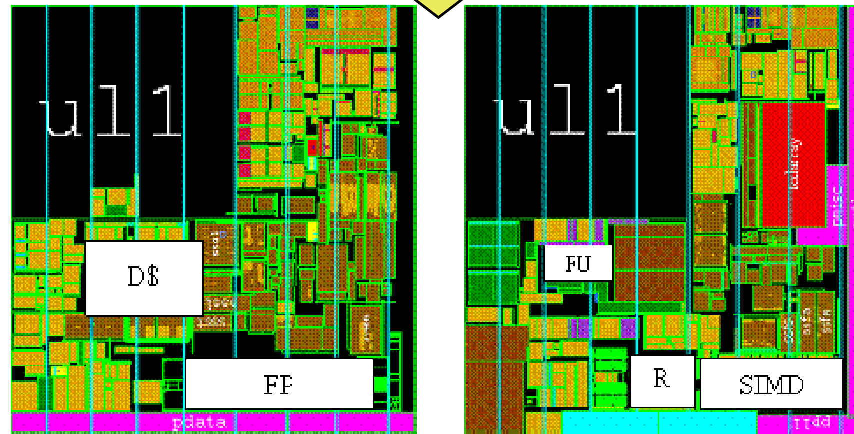
3D interconnects

From a processor des
the most important attri

Intel® 3D Pentium® 4 (ICCD 2004)



Source: Intel



Top **Bottom**

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

ISCA 2006

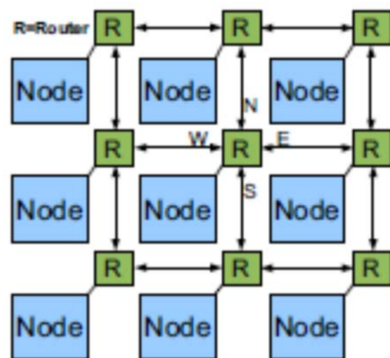
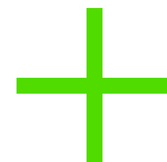
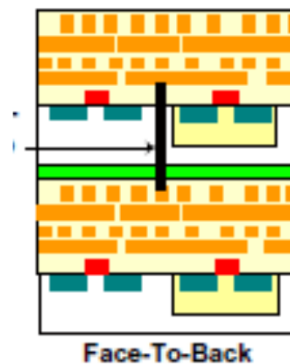


Figure 1. A typical NoC mesh.




Face-To-Back

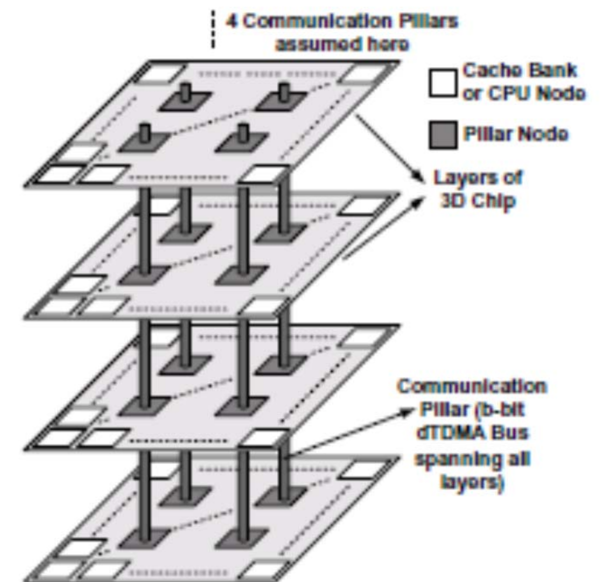
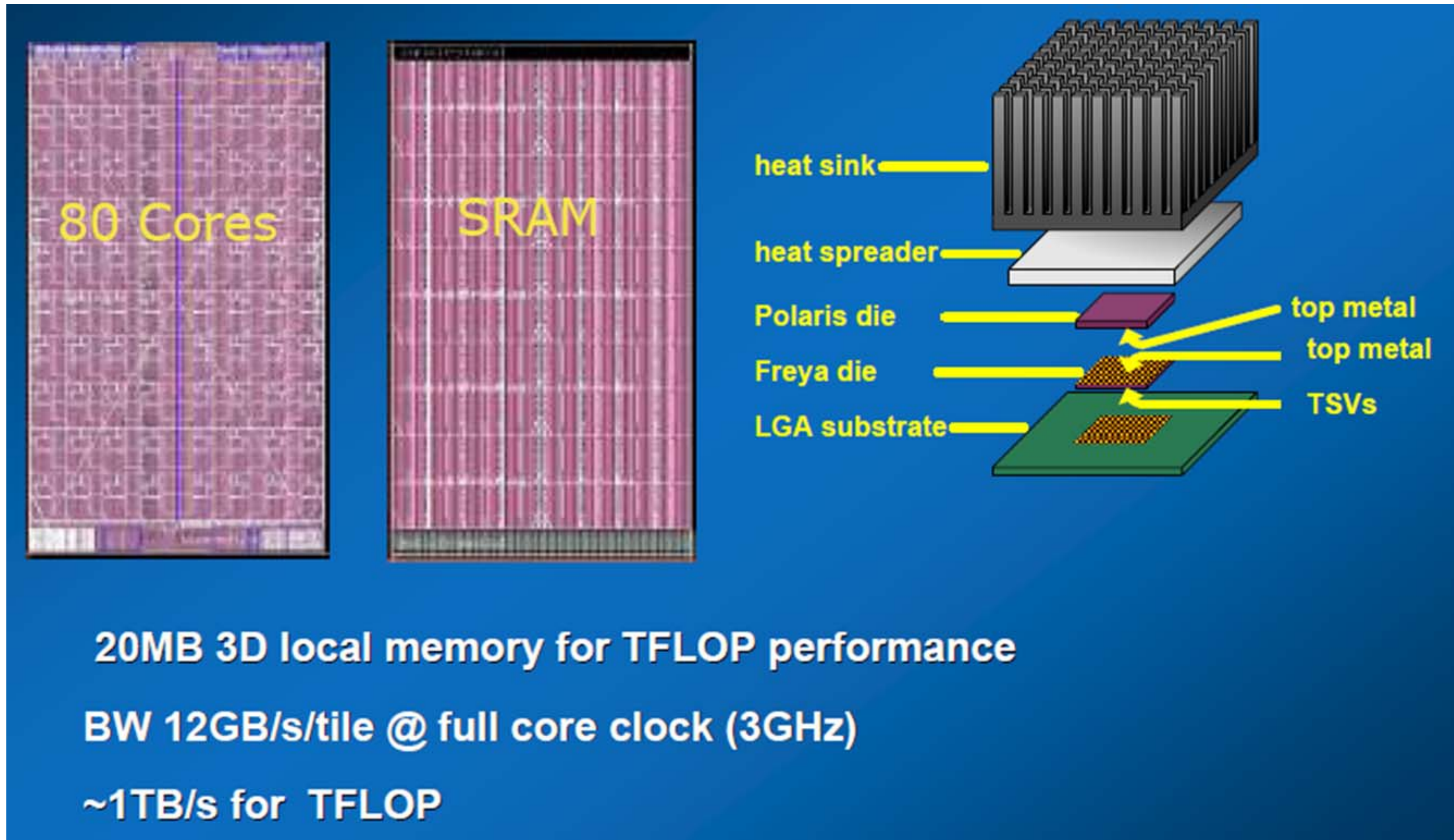


Figure 4. Proposed 3D Network-in-Memory architecture

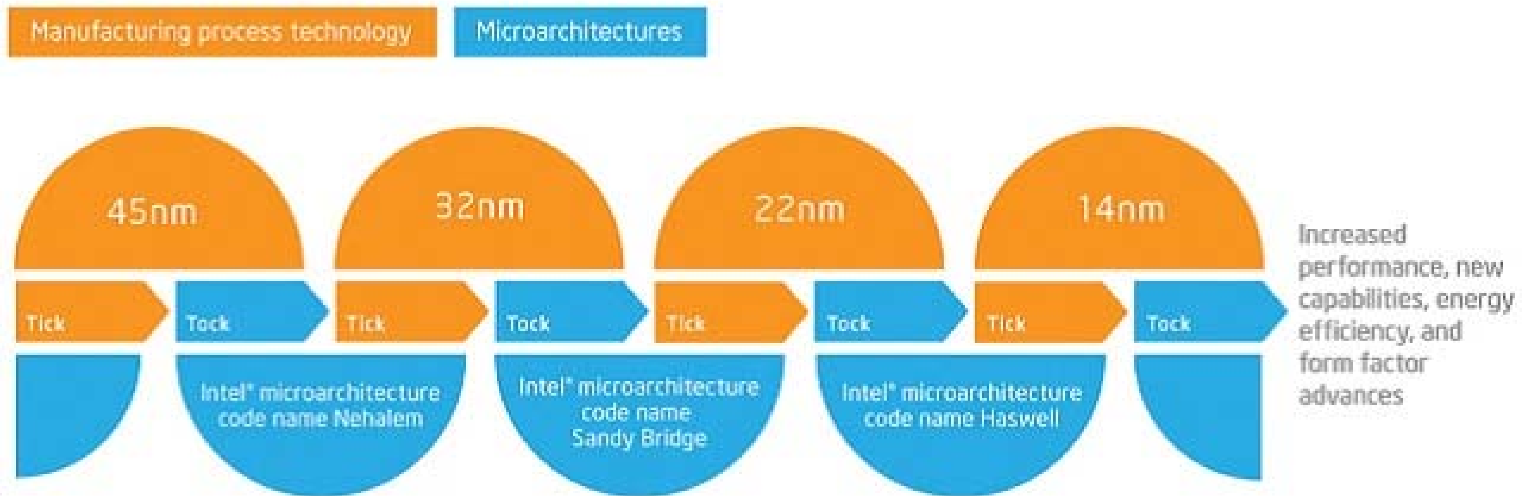
Intel's 3D +NOC Prototyping (2007)



Courtesy: T. Karnik (Intel)

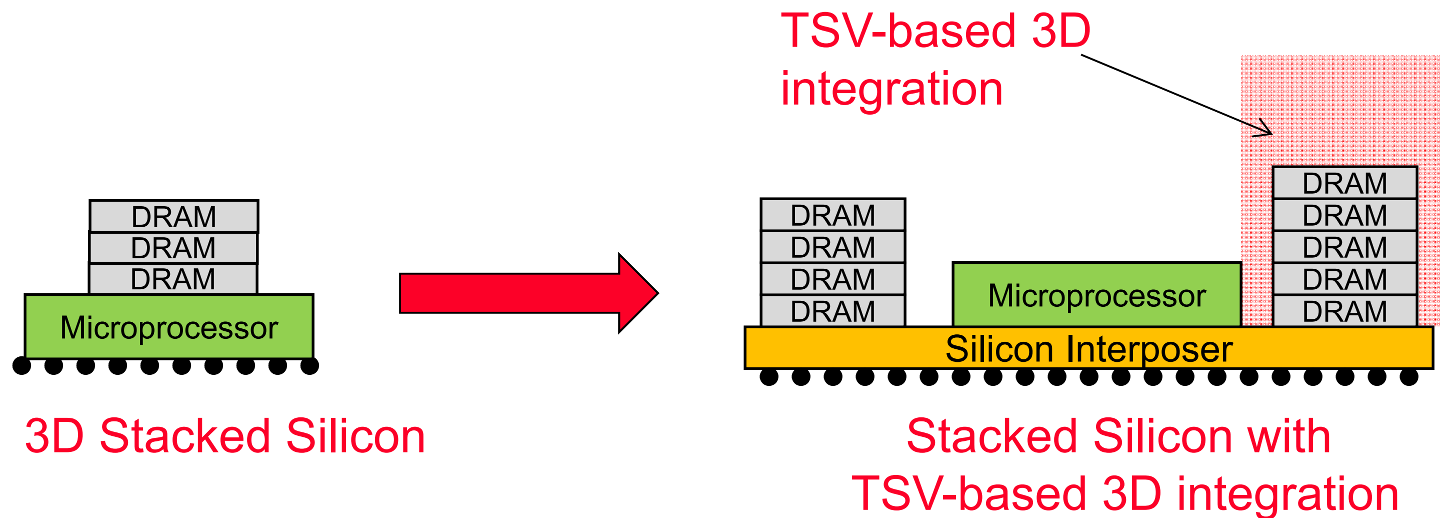
Intel's Tick-Tock Model

- Proposed in 2007
- Tick: technology change (e.g. 45nm->32nm)
- Tock: architecture change (e.g. Nehalem->SandyBridge)



2006	2007	2008	2009	2010	2011	2012	2013
65nm	65nm	45nm	45nm	32nm	32nm	22nm	22nm
NetBurst	Core	Core	Nehalem	Nehalem	Sandy bridge	Sandy bridge	Haswell
2014	2015	2016	2017	2018	2019	2020	2021
14nm	14nm	10nm	10nm	7nm	7nm	5nm	5nm
Haswell	Skylake	Skylake					

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)

Design Space Exploratio
3D Architectures

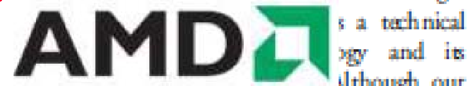
YUAN XIE
Pennsylvania State University
GABRIEL H. LOH
Georgia Institute of Technology
BRYAN BLACK
Intel Corporation
and
KERRY BERNSTEIN
IBM Corporation

As technology scales, interconnects have become a majo
of power consumption for microprocessors. Increasing
sider alternate ways of building modern microprocesso
where a stack of multiple device layers with direct verti
on the same chip. As fabrication of 3D integrated circu
and architectural techniques is imperative to explore t
this article, we give a brief introduction to 3D integrati
that can enable the adoption of 3D ICs, and present th
components using 3D technology. An industrial case
design 3D microarchitectures.

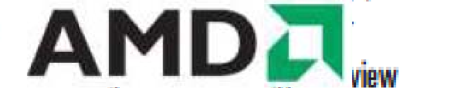
PROCESSOR DESIGN IN 3D DIE-STACKING TECHNOLOGIE

THREE-DIMENSIONAL DIE-STACKING INTEGRATION STACKS MULTIPLE L
PROCESSED SILICON WITH A VERY HIGH-DENSITY, LOW-LATENCY LAYE
INTERCONNECT. AFTER PRESENTING A BRIEF BACKGROUND ON 3D DI
TECHNOLOGY, THIS ARTICLE GIVES MULTIPLE CASE STUDIES ON DIFFE
FOR IMPLEMENTING SINGLE-CORE AND MULTICORE 3D PROCESSORS A
HOW TO DESIGN FUTURE MICROPROCESSORS GIVEN THIS EMERGING T

Gabriel H. Loh
Georgia Institute of Technology



Yuan Xie
Pennsylvania State University



Bryan Black
Intel

...... 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 orga- ... s a technical ... ogy and its ... Although our discussions here primarily focus on high-performance processor design, most of the observations and conclusions apply to other ...

... view

Although there are several candidate variants on 3D integration technology, at the heart of all of them is the vertical stacking ... ted chips. ... esses that ...

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

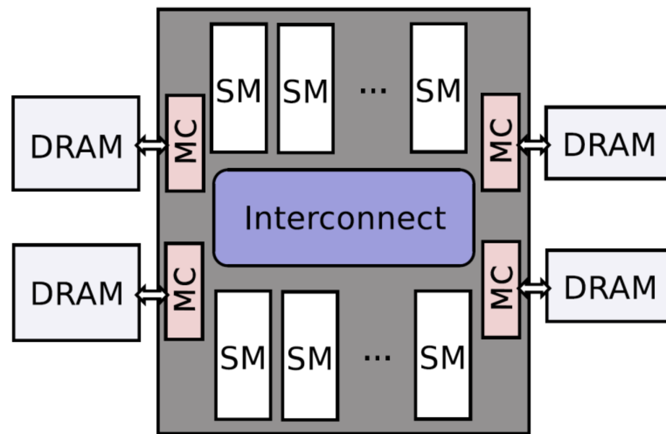
When considering th of two silicon dies, the topologies: face to face where a die's "face" is metallization and its "ba the silicon substrate. bonding process builds tion, also called a *die-to-* by depositing the coppe on each die, and then bo together with a thermoco A chemical-mechanical thins one die to redu communication between for external I/O and po

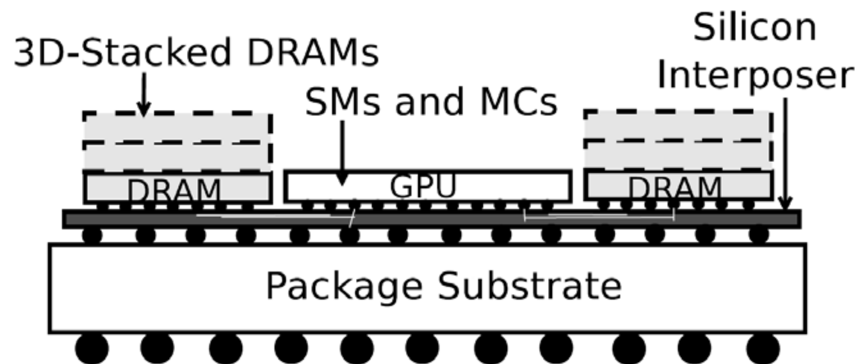
3D interconnects

From a processor des

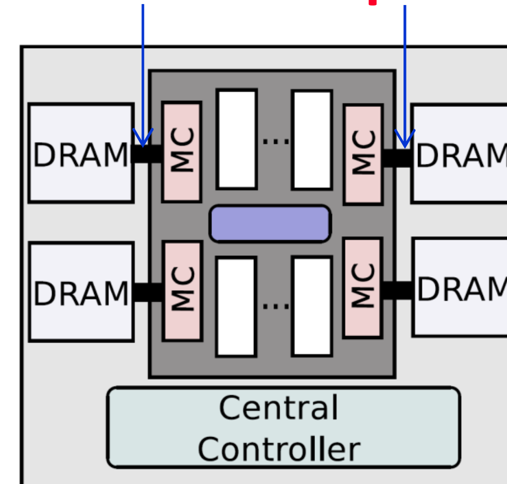
In-package 3D Memory with GPU

Conventional GDDRs, off-chip





Wide-bus routing on silicon interposer



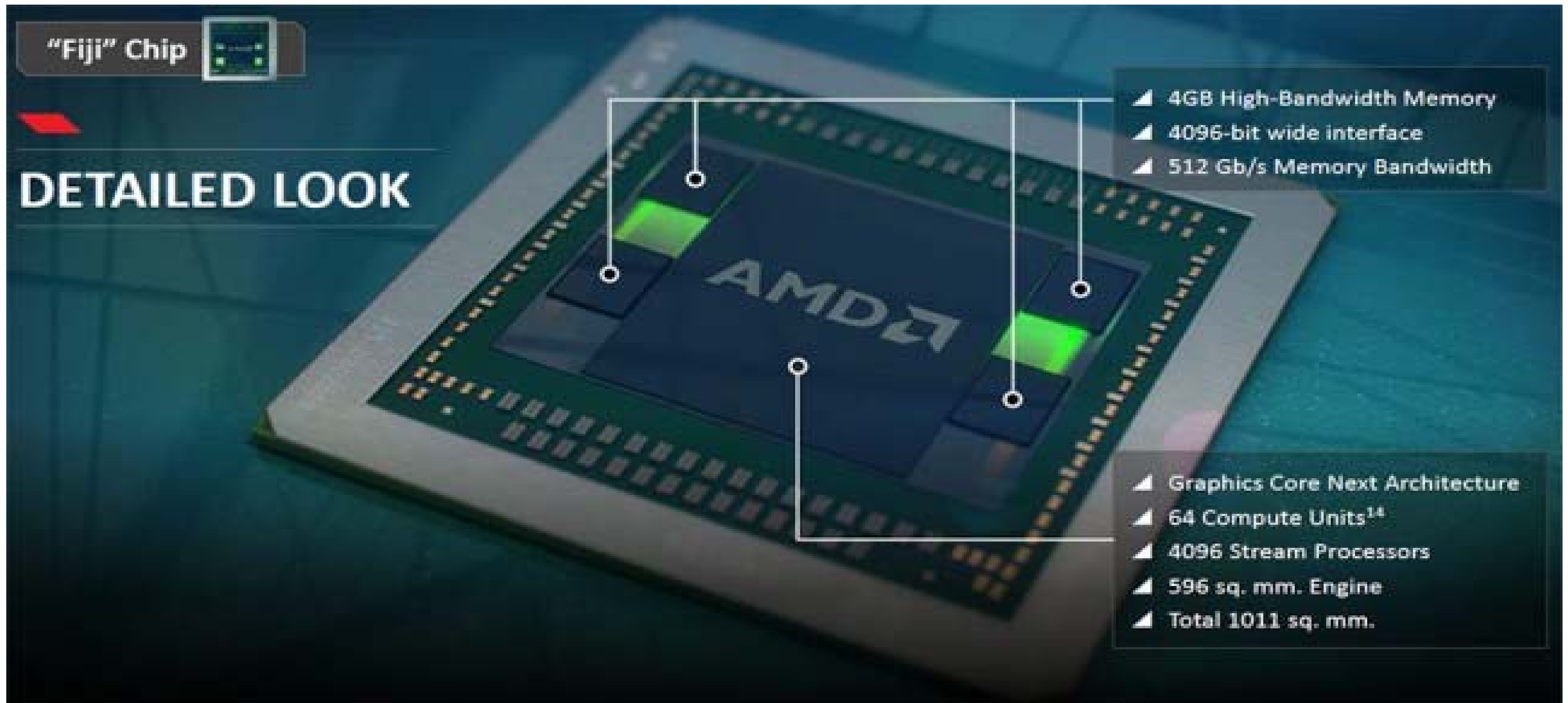
Top View

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 Happened!

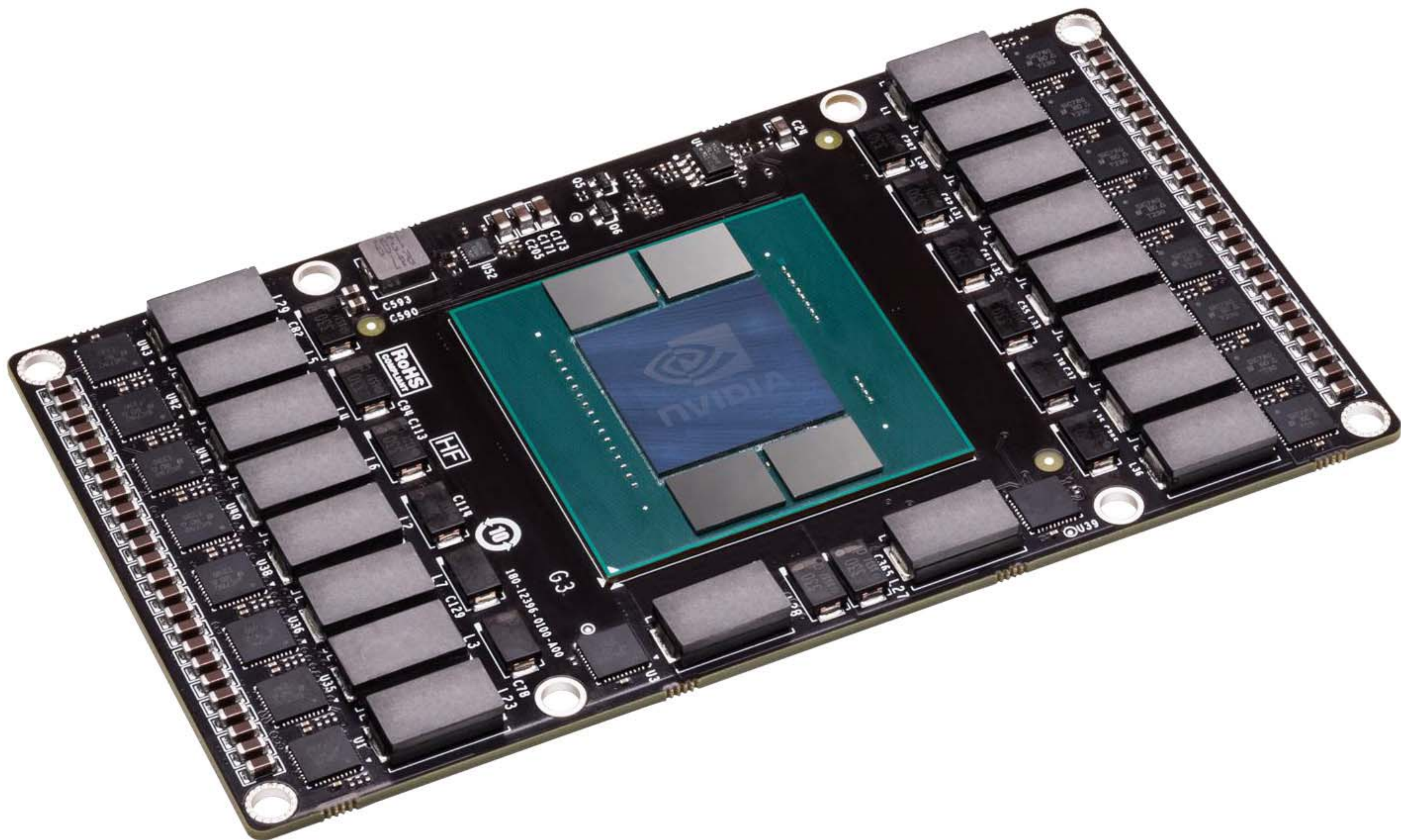
AMD Announcement on June 16, 2015



- 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

AMD

Nvidia Pascal (2016)

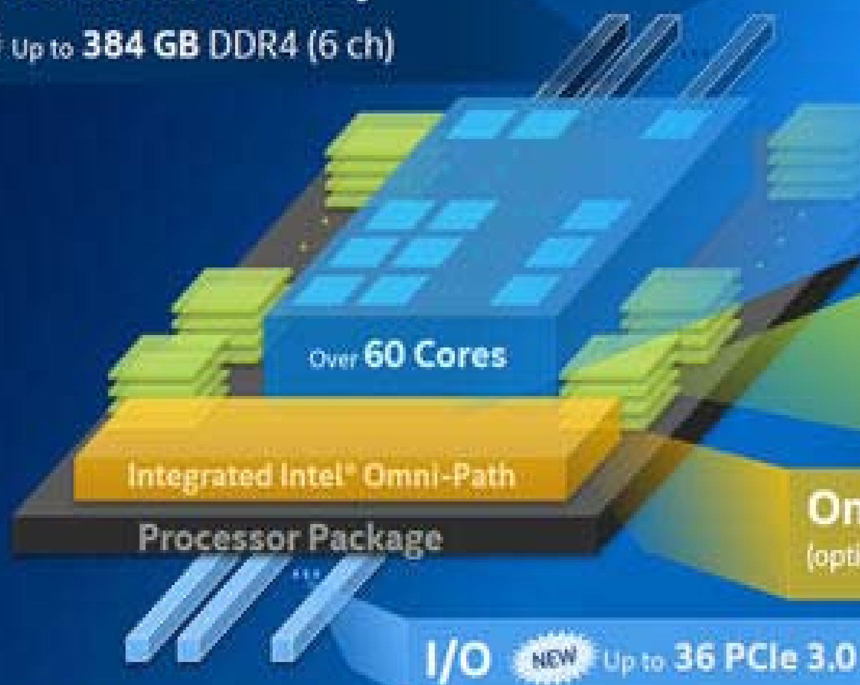


Knights Landing

Holistic Approach to Real Application Breakthroughs

Platform Memory

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



Compute

- Intel® Xeon® Processor Binary-Compatible
- **3+ TFLOPS**[1], **3X ST**[2] (single-thread) perf. vs KNC
- **2D Mesh** Architecture
- **Out-of-Order** Cores

On-Package Memory

- Over **5x** STREAM vs. DDR4[3]
- Up to **16 GB** at launch

Omni-Path

(optional)

- 1st Intel processor to integrate

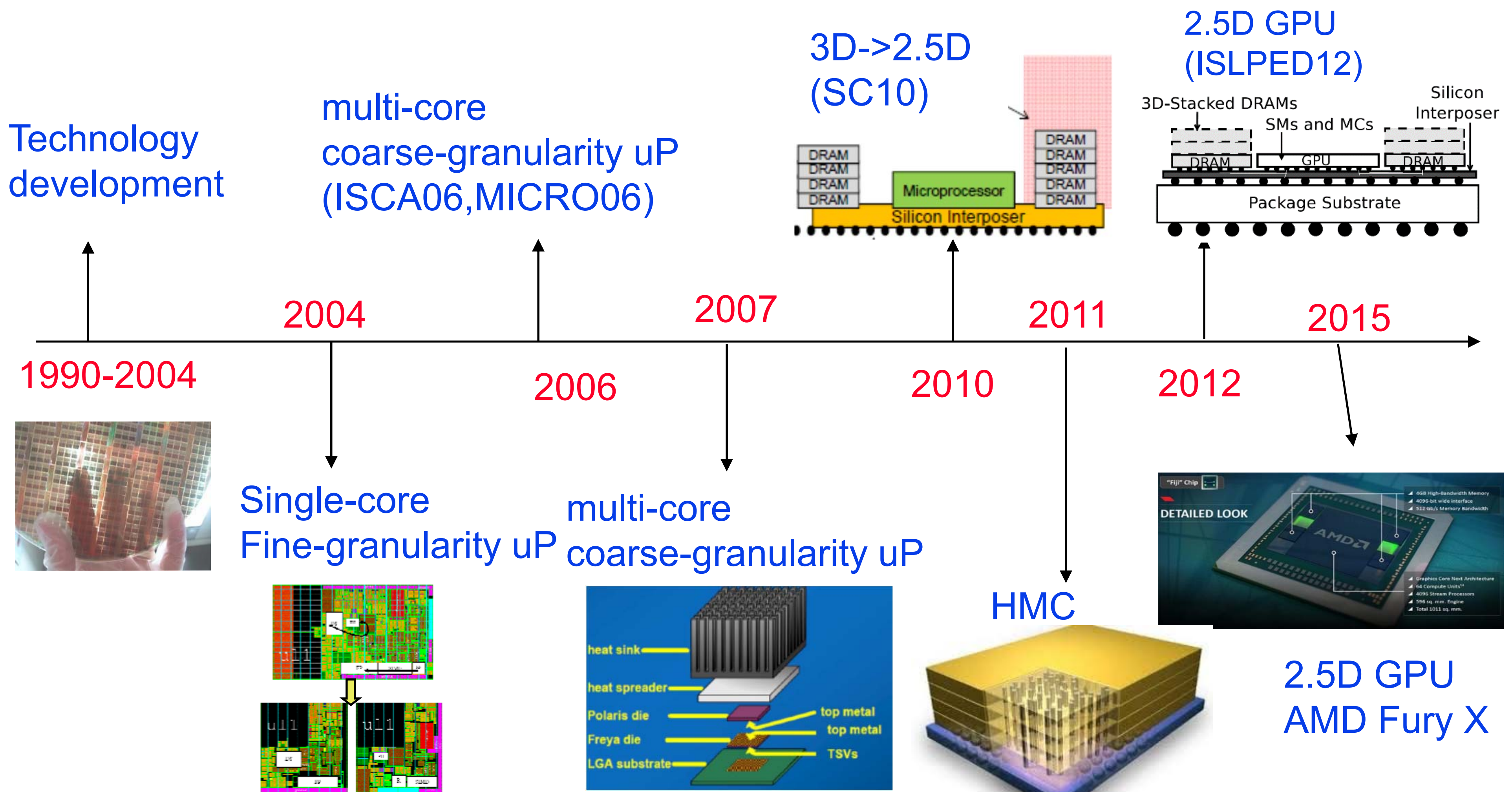
I/O

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 those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, 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



Technology-Driven and Application-Driven Architecture Innovations:

Emerging Application-Driven Architecture

Machine Learning as a Key Workload

Machine Learning is changing the way we implement applications. Hardware advancement make learning over big data possible.

Machine learning (ML) has made significant p over the last decade in producing application long been in the realm of science fiction, fron sought, practical voice-based interfaces to se cars. One can claim that this progress has be fueled by abundant data coupled with copiou power. Large-scale machine learning applicat motivated designs that range from storage s specialized hardware (GPUs, TPUs).

机器学习作为核心负载

机器学习正在改变我们实现应用程序的方式。硬件性能提升使得机器学习应用于大数据处理成为可能。

机器学习在过去十年中取得了长足的进步，生产了很多长期以来只存在于科幻小说里的应用，从长期追求实用的基于语音的接口到自动驾驶汽车。可以说，这一进步在很大程度上受益于丰富的数据和强大的计算能力。大规模机器学习应用也促进了包括存储系统和专用硬件（GPU，TPU）等的设计。

尽管目前的重点是支持云端的机器学习，但是在诸如智能手机和超低功耗传感器节点等低功耗设备中支持机器学习应用也存在很大潜力。幸运的是，许多机器学习内核具有相对规整的结构，可控的准确率和资源需求之间的权衡;因此，它们适用于专用硬件、重构和近似计算等技术，为体系结构的创新开启了新空间。

机器学习从业者花费相当长的时间用于模型训练。传闻有证据表明，即便使用超大规模的计算集群，花费一星期到一个月来训练一个模型也是普遍的。虽然这样的计算资源投资能够分摊到对模型多次调用，但模型较长的更新迭代周期可能会对用户体验产生负面影响。因此，对体系结构研究人员来说，设计更好地支持机器学习模型训练的系统是一个新的机遇。

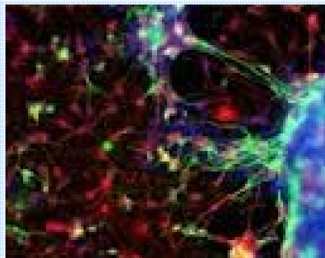
The (Re)Rising of AI Applications

Supercomputers | **Data Centers** | **Smartphones** | **Embedded Devices**

Business analytics

Audio recognition

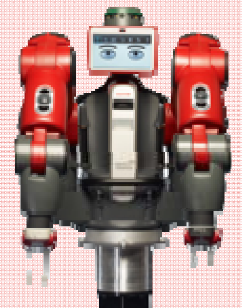
Robotics

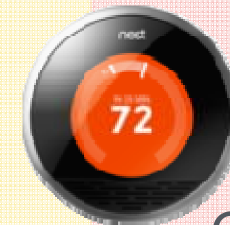


Drug design

Automatic translation

Image analysis





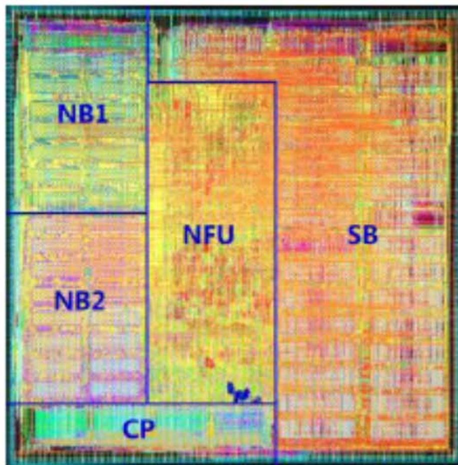
Consumer electronics

Courtesy: Yunji Chen, ICT

New Opportunities for Chinese Scholars

- Chinese researchers becomes the pioneers/leaders in the design of novel architecture for AI application.

寒武纪：开创深度学习处理器方向

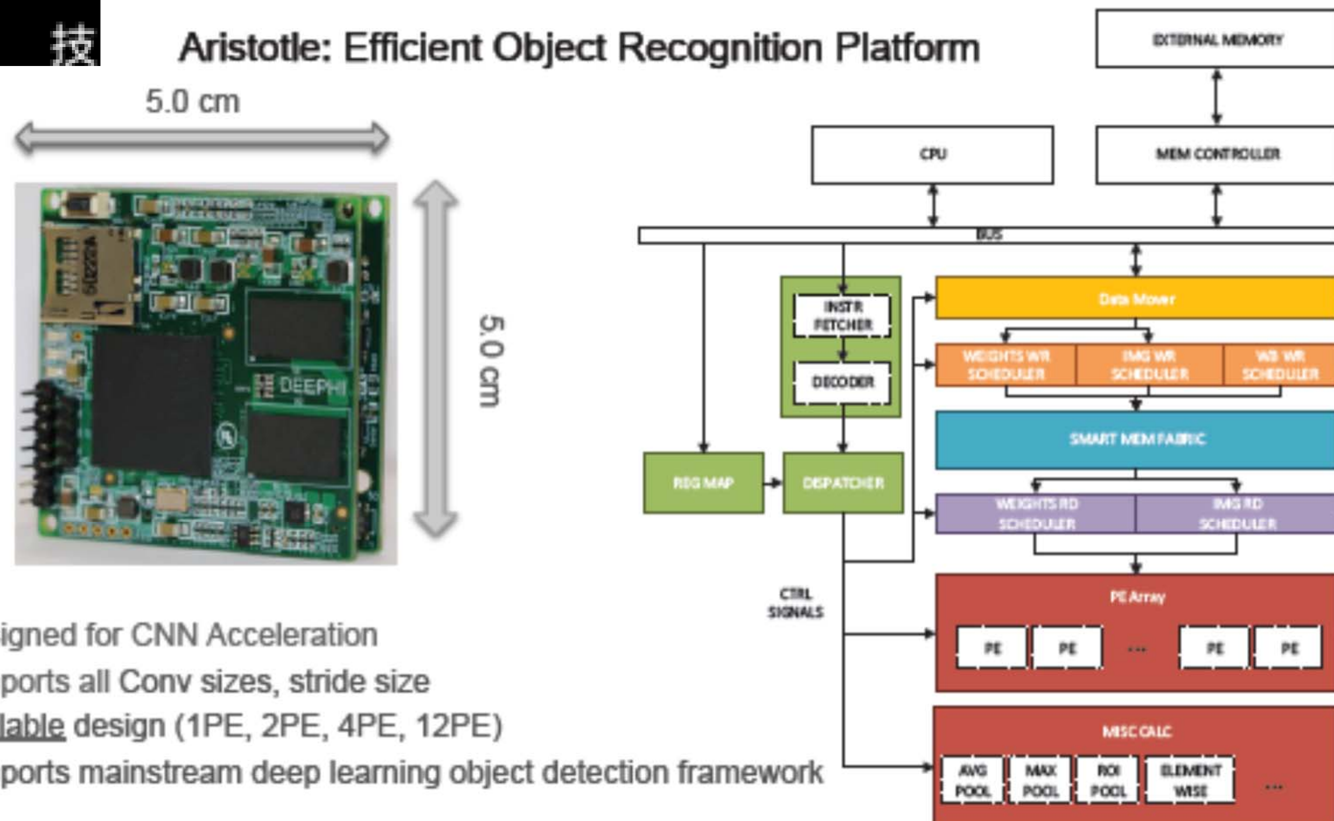


- 1GHz，0.485W @ 65nm，通用CPU 1/10的面积，100倍的性能

- **2013：国际首个深度学习处理器**
  - **DianNao: ASPLOS'14最佳论文**
  - **亚洲首获体系结构A类会议最佳论文**
- 2014：国际首个多核深度学习处理器
  - DaDianNao：MICRO'14最佳论文
- 2015：国际首个通用机器学习处理器
  - PuDianNao：ASPLOS'15
- 2015：摄像头上的智能识别IP
  - ShiDianNao：ISCA'15
- 2016：国际首个神经网络通用指令集
  - Cambricon：ISCA'16评审分数第一

New Opportunities for Chinese Scholars

- Chinese researchers becomes the pioneers/leaders in the design of novel architecture for AI application.



GPU's Important Role in AI HW Acceleartor

Building High-level Features
Using Large Scale Unsupervised Learning

Quoc V. Le QUOCLE@CS.STANFORD.EDU
Marc'Aurelio Ranzato RANZATO@GOOGLE.COM
Rajat Monga RAJATMONGA@GOOGLE.COM
Matthieu Devin MDEVIN@GOOGLE.COM
Kai Chen KAICHEN@GOOGLE.COM
Greg S. Corrado GCORRADO@GOOGLE.COM
Jeff Dean JEFF@GOOGLE.COM
Andrew Y. Ng ANG@CS.STANFORD.EDU

- 2012, training on a cluster with 1,000 machines (16,000 CPU cores) for three days.

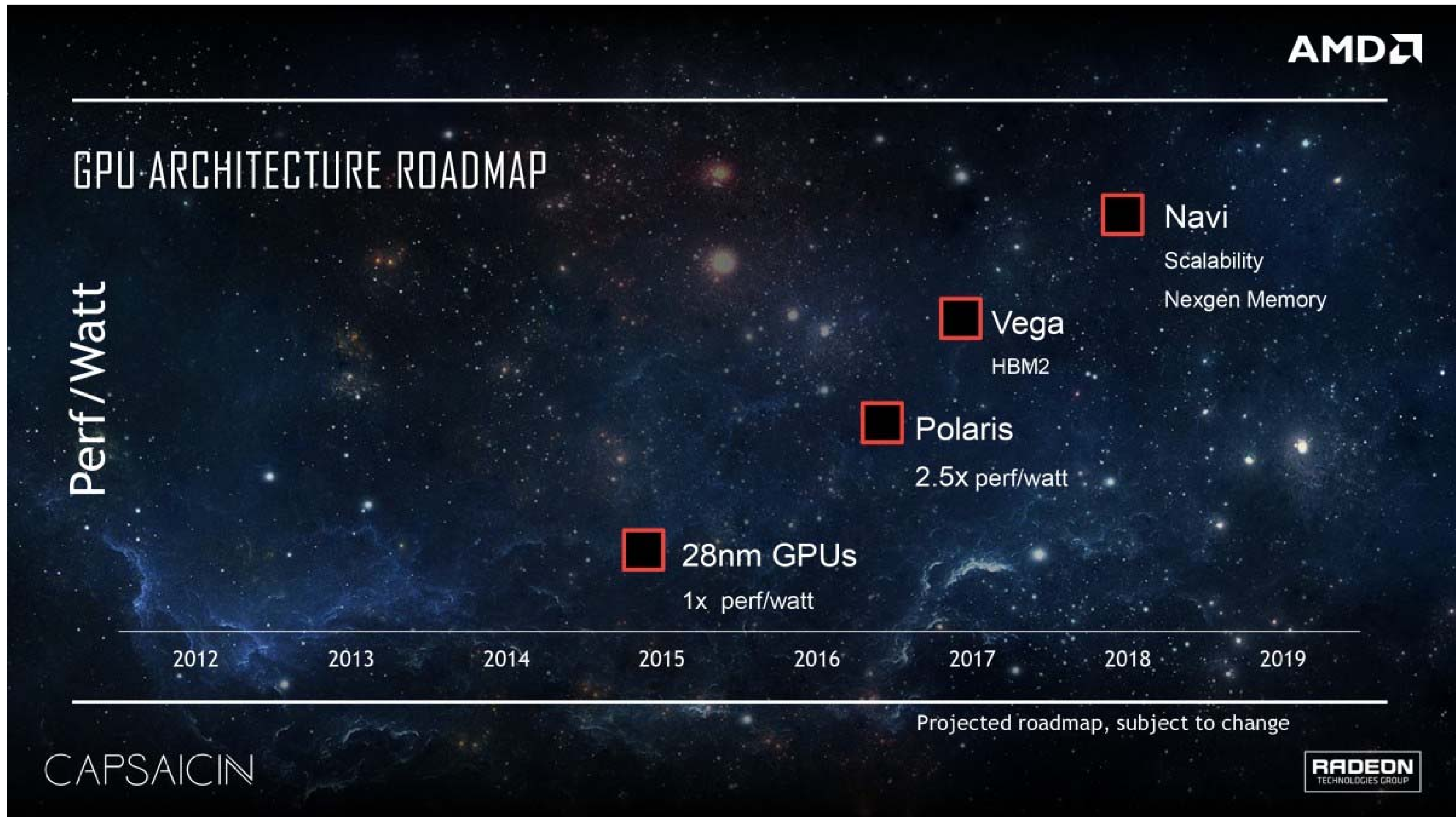
Deep learning with COTS HPC systems

Adam Coates ACOATES@CS.STANFORD.EDU
Brody Huval BRODYH@STANFORD.EDU
Tao Wang TWANGCAT@STANFORD.EDU
David J. Wu DWU4@CS.STANFORD.EDU
Andrew Y. Ng ANG@CS.STANFORD.EDU
Stanford University Computer Science Dept., 353 Serra Mall, Stanford, CA 94305 USA

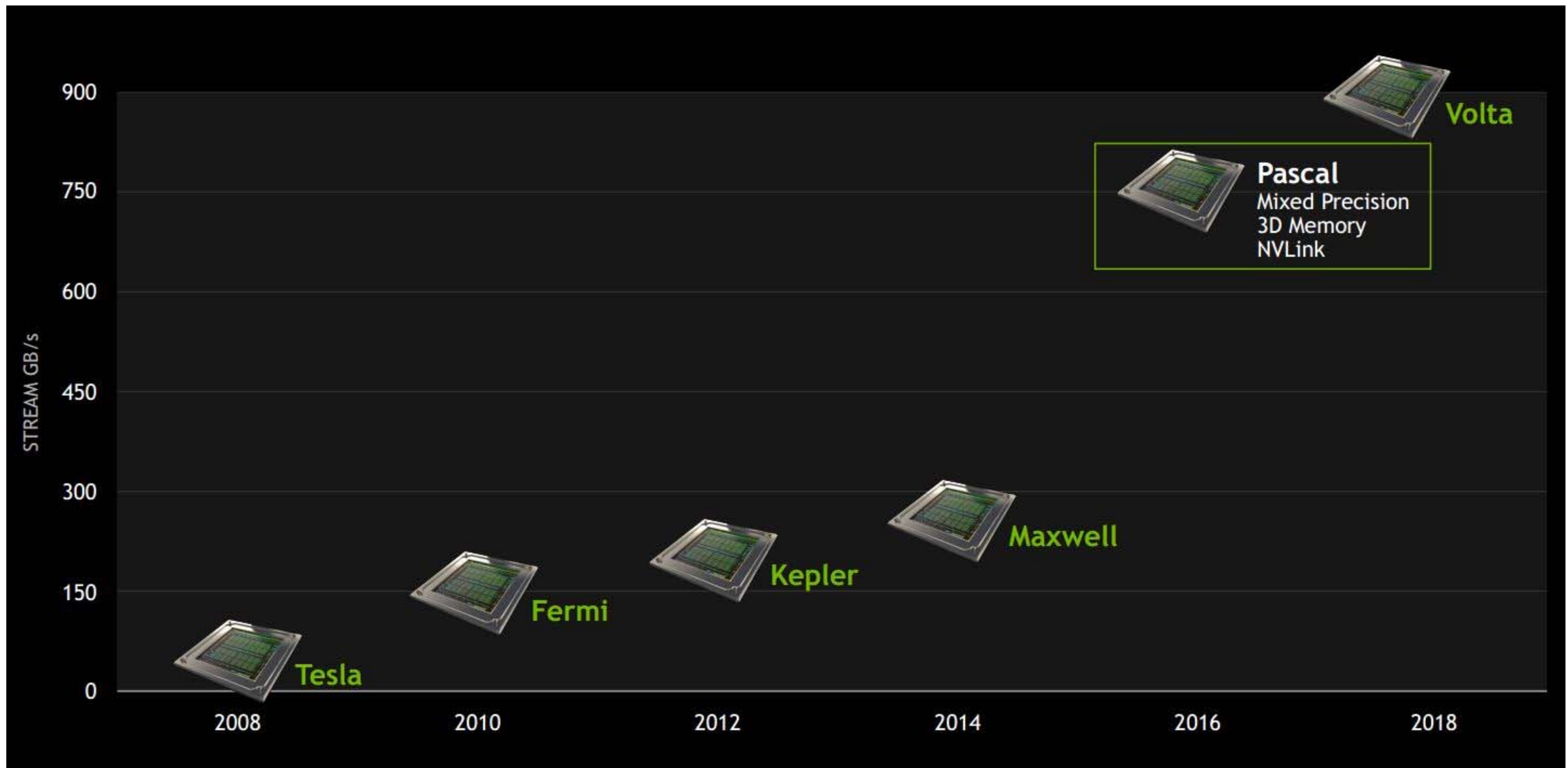
Bryan Catanzaro BCATANZARO@NVIDIA.COM
NVIDIA Corporation, 2701 San Tomas Expressway, Santa Clara, CA 95050

- 2013, training on a cluster of 3 GPU servers for 2 days

AMD Roadmap



Nvidia GPU Roadmap



AMD’s Catch to Nvidia’s on AI (2017)

Accelerators | RADEON INSTINCT

MI6

Passively Cooled Inference Accelerator

5.70 TFLOPS

224 GB/s Memory Bandwidth

<150W

MI8

Small Form Factor Accelerator

8.2 TFLOPS

512 GB/s Memory Bandwidth

<175W

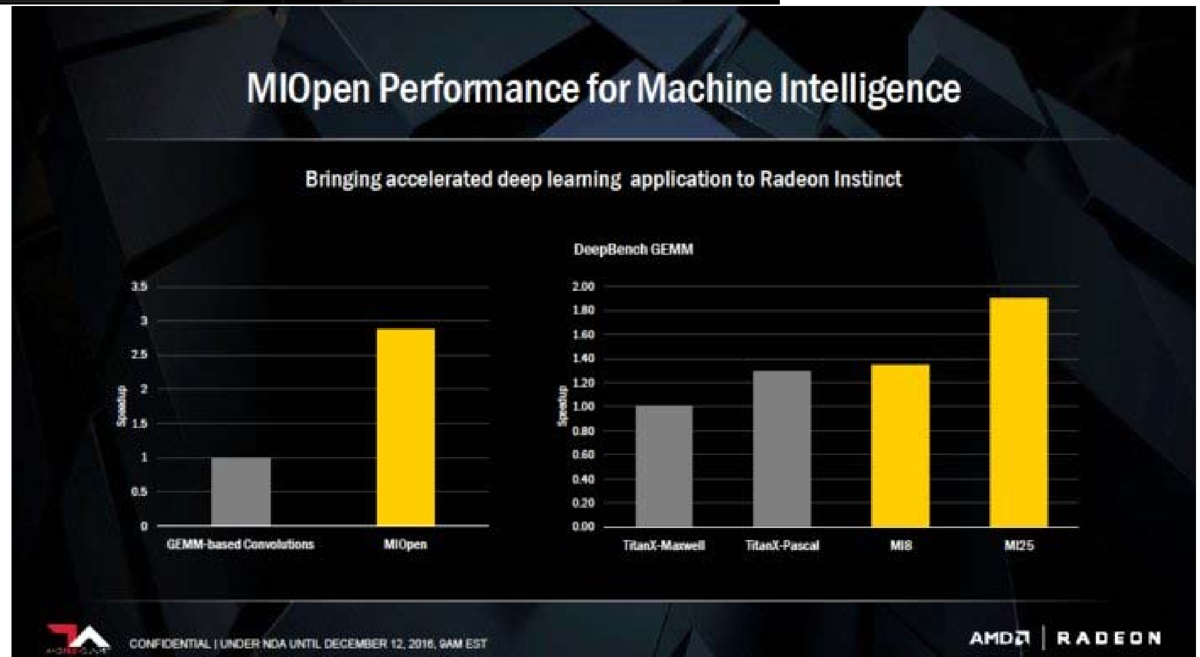
MI25 Vega with NCU

Passively cooled Training Accelerator

2X Packed Math

High Bandwidth Cache and Controller

<300W



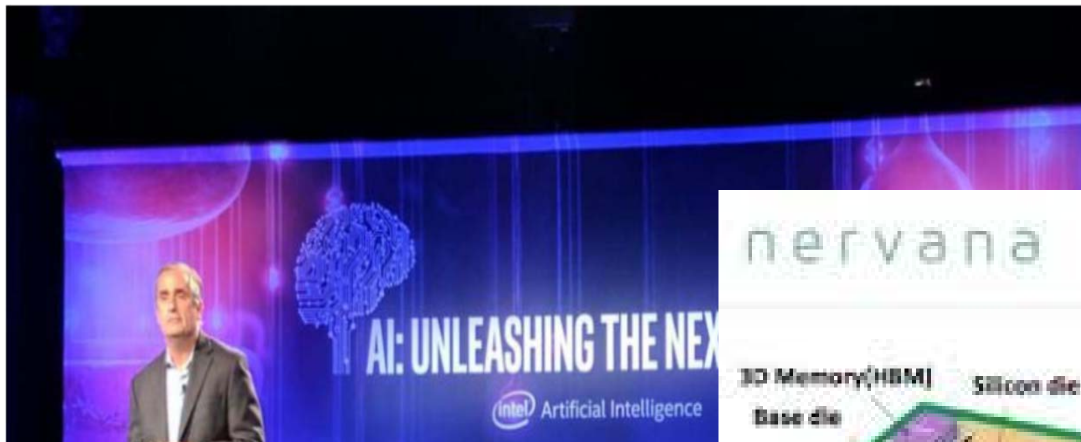
Intel/Nervana

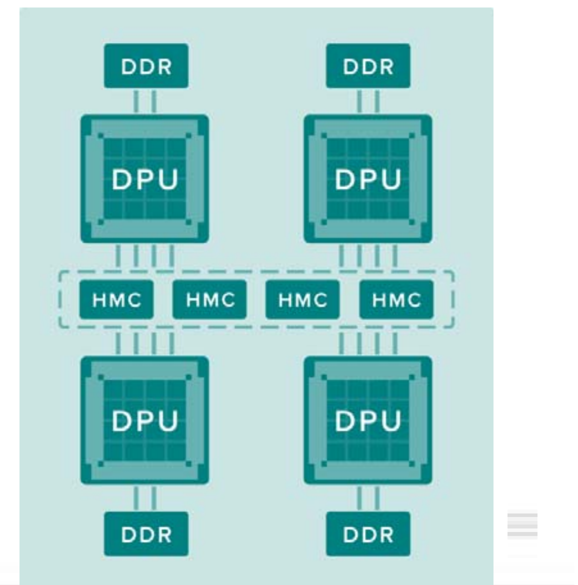
2016/8, Intel paid $350M to acquire Nervana to enter AI HW era

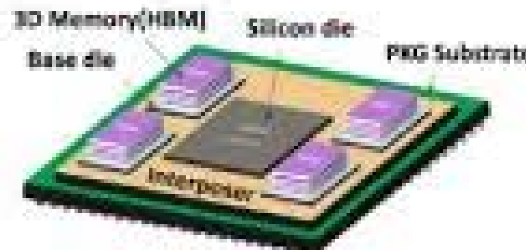
Intel will test Nervana's 'Lake Crest' silicon in first half of 2017, 'Knights Crest' also coming

JORDAN NOVET (HTTPS://VENTUREBEAT.COM/AUTHOR/JORDAN-NOVET)   @JORDANNOVET (HTTPS://TWITTER.COM/JORDANNOVET)

NOVEMBER 17, 2016 1:14 PM







Blazingly fast data access via high-bandwith memory

Training deep learning networks involves moving a lot of data, and current memory technologies are simply not up to the task. The Nervana Engine uses a new memory technology called High Bandwidth Memory that is both high-capacity and high-speed, providing 32 GB of on-chip storage and a blazingly fast 8 Tera-bits per second of memory access speed.

FPGA vs. GPU

Can FPGAs Beat GPUs in Accelerating Next-Generation Deep Neural Networks?

Eriko Nurvitadhi[1], Ganesh Venkatesh[1], Jaewoong Sim[1], Debbie Marr[1],
Randy Huang[2], Jason Gee Hock Ong[2], Yeong Tat Liew[2],
Krishnan Srivatsan[3], Duncan Moss[3], Suchit Subhaschandra[3], Guy Boudoukh[4]

[1]Accelerator Architecture Lab, [2]Programmable Solutions Group, [3]FPGA Product Team, [4]Computer Vision Group
Intel Corporation

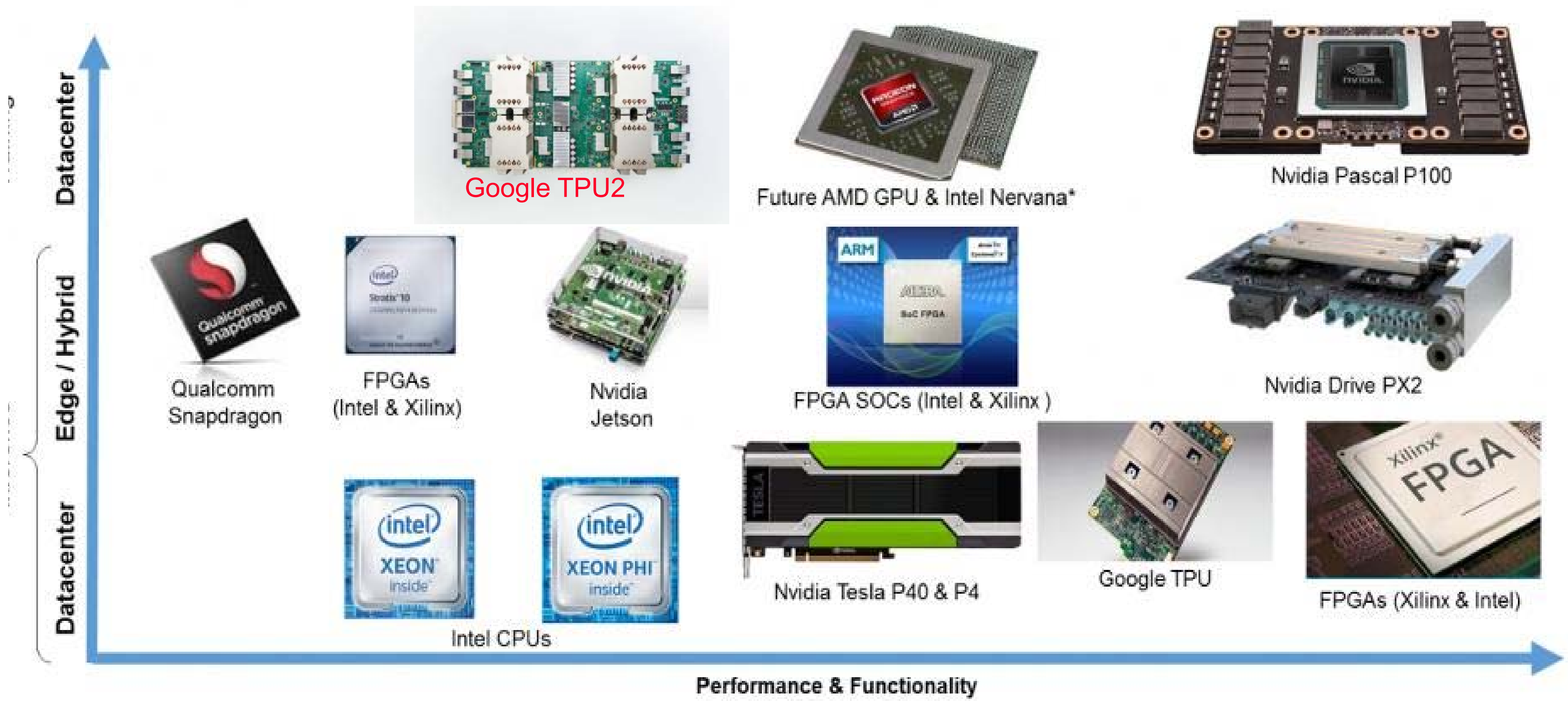
Microsoft has revealed that Altera FPGAs have been installed across every Azure cloud server, creating what the company is calling “the world’s first AI supercomputer.” The deployment spans 15 countries and represents an aggregate performance of more than one exa-op. -2016

GPU vs. FPGA for Deep Learning

GPU	FPGA	
Training in a cloud-based environment	- Excel at **inference**, where requires the most compute efficiency in terms of performance-per-watt.	
Large-scale **inference** workloads available on many public clouds	Possible training (see Intel's FPGA 17 paper)	
- Ease-of-use, - Well-established ecosystem, - Abundance of standardized libraries, frameworks, and support	- **Reconfigurable**, enabling leverage across a wide range of workloads and new evolving algorithms and neural networks (Compression, pruning, and variable / limited precision (8-bit to 1-bit layers in the same network) techniques	
- power consumption - lack of the ability to accommodate hardware changes as neural networks and algorithms evolve	- Difficulty to program	

No “one chip to rule them all” solutions

- ❑ Each has its advantages for a specific type of application, or data, that is being deployed and in a specific environment.
  - ❑ The data complexity and velocity determines how much processing is needed,
  - ❑ The environment determines the latency demands and the power budget.

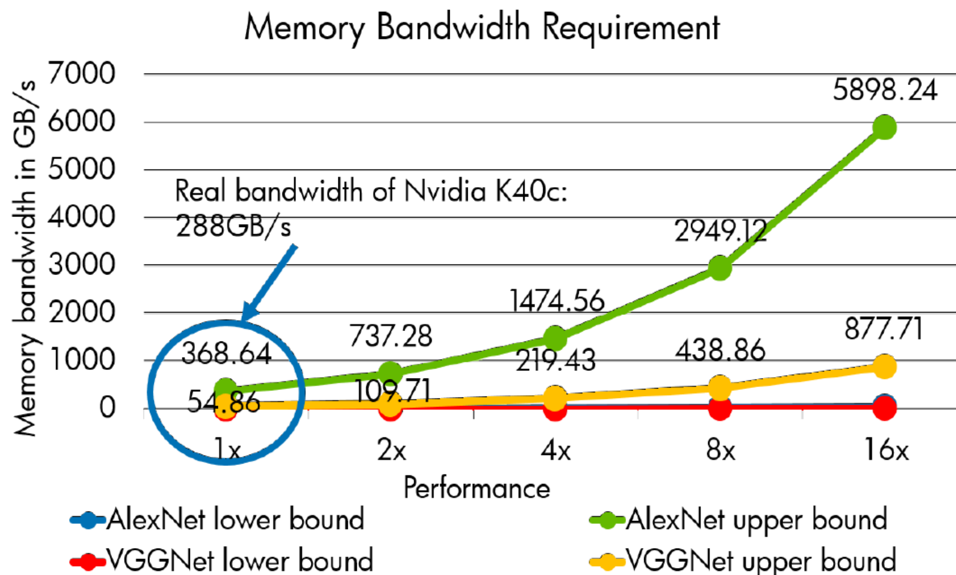


Copyright © 2017 Moor Insights & Strategy

Hardware across the Machine Learning landscape (Source: Moor Insights & Strategy)

Today's NN and DL Acceleration

- Neural network (NN) and deep learning (DL)
  - Provide solutions to various applications
  - Acceleration requires high memory bandwidth
    - Memory bandwidth becomes the bottleneck



Deng *et al*, "Reduced-Precision Memory Value Approximation for Deep Learning", HPL Report, 2015

- The size of NN increases
  - e.g., 1.32GB synaptic weights for Youtube video object recognition

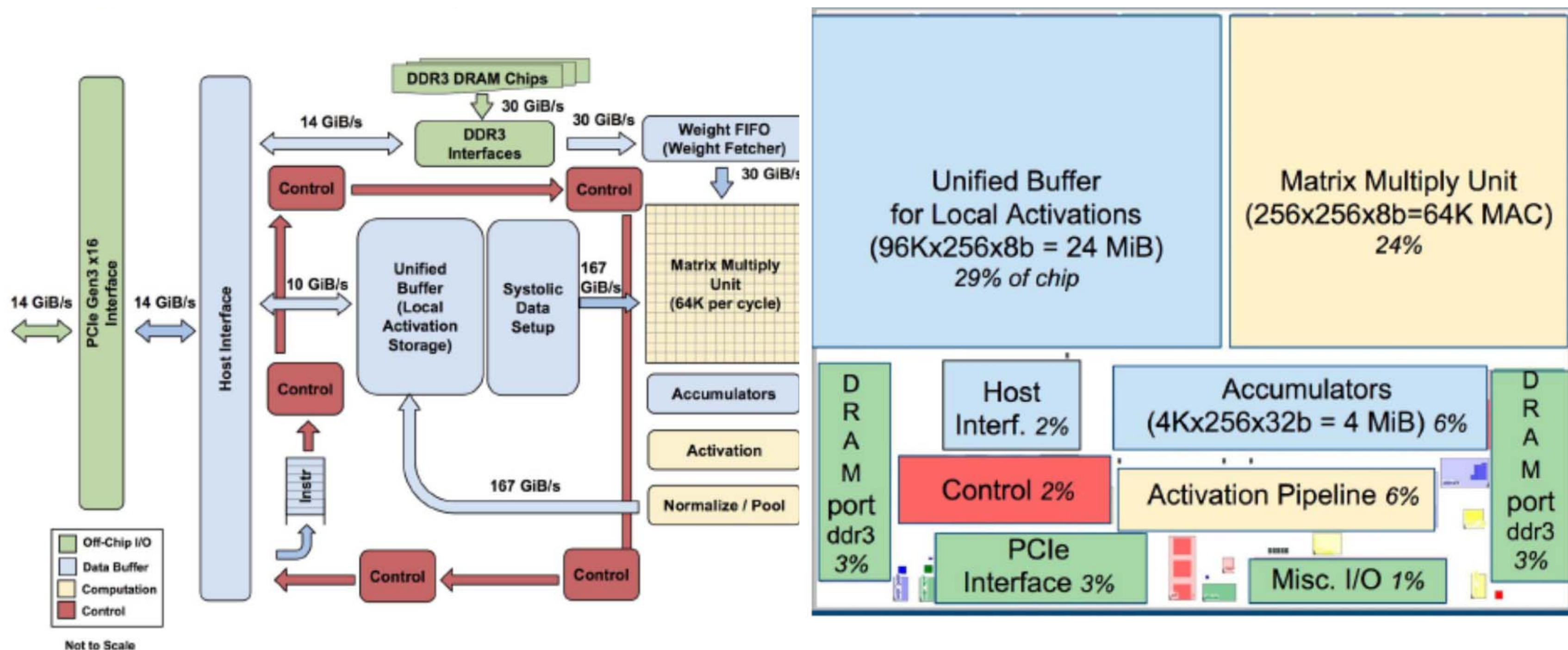
Google TPU Disclosed Last Month in ISCA

In-Datacenter Performance Analysis of a Tensor Processing Unit™

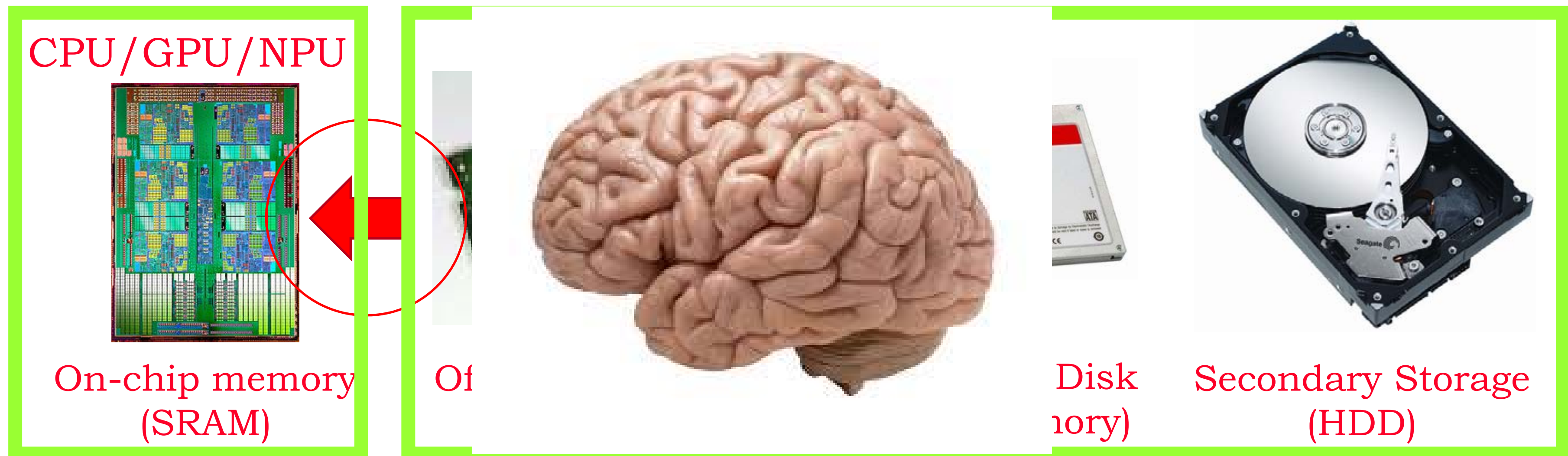
The TPU is about 15X - 30X faster at inference than the K80 GPU and the Haswell CPU.

n,

NN apps are **memory-bandwidth limited** on the TPU; if the TPU were revised to **have the same memory system as the K80 GPU**, it would be about 30X - 50X faster than the GPU and CPU.



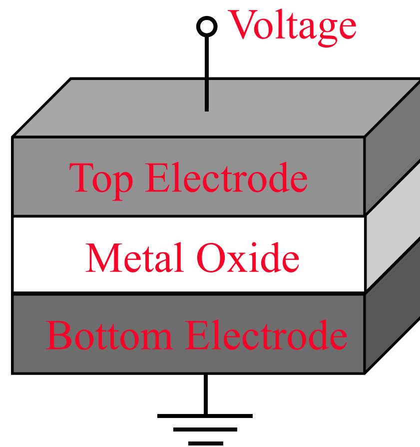
Today's Architecture

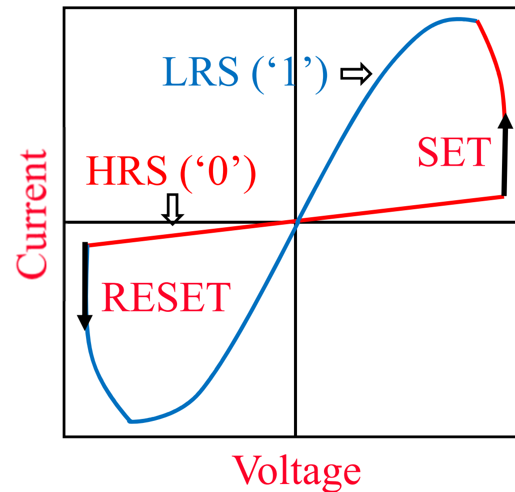


Our brain doesn't have a distinction of compute vs. memory

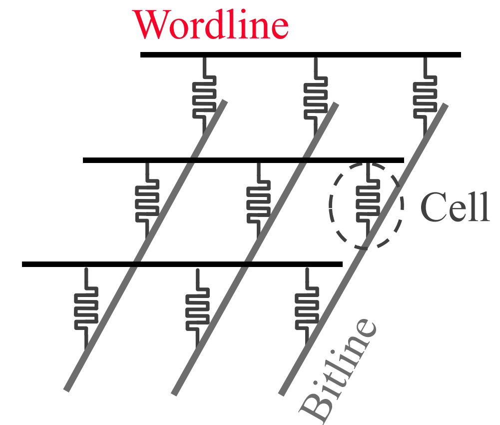
New Architecture:
In-Memory Computing/Near-Data Computing

ReRAM Based NN Computation

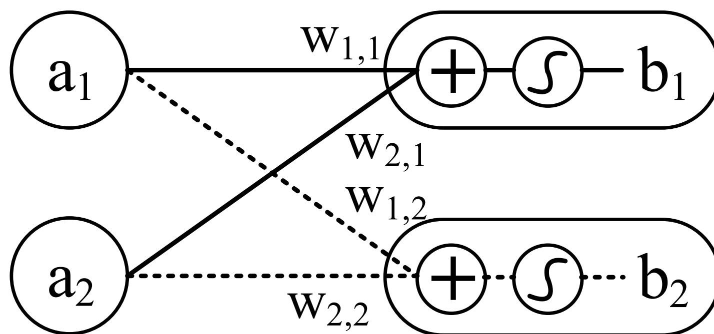




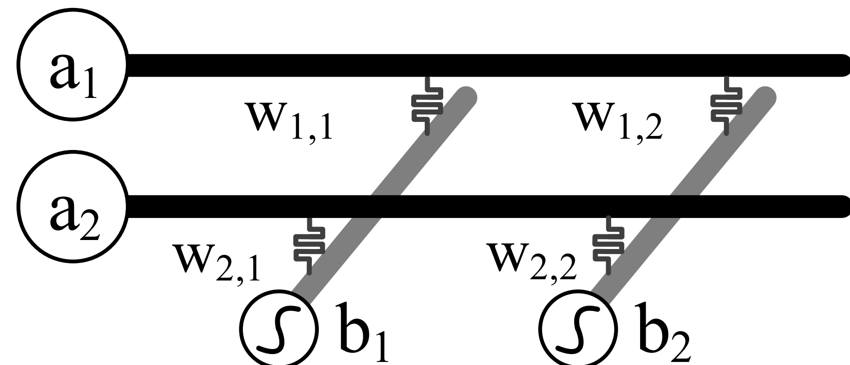
(b) I-V curve of bipolar switching



(c) schematic view of a crossbar architecture

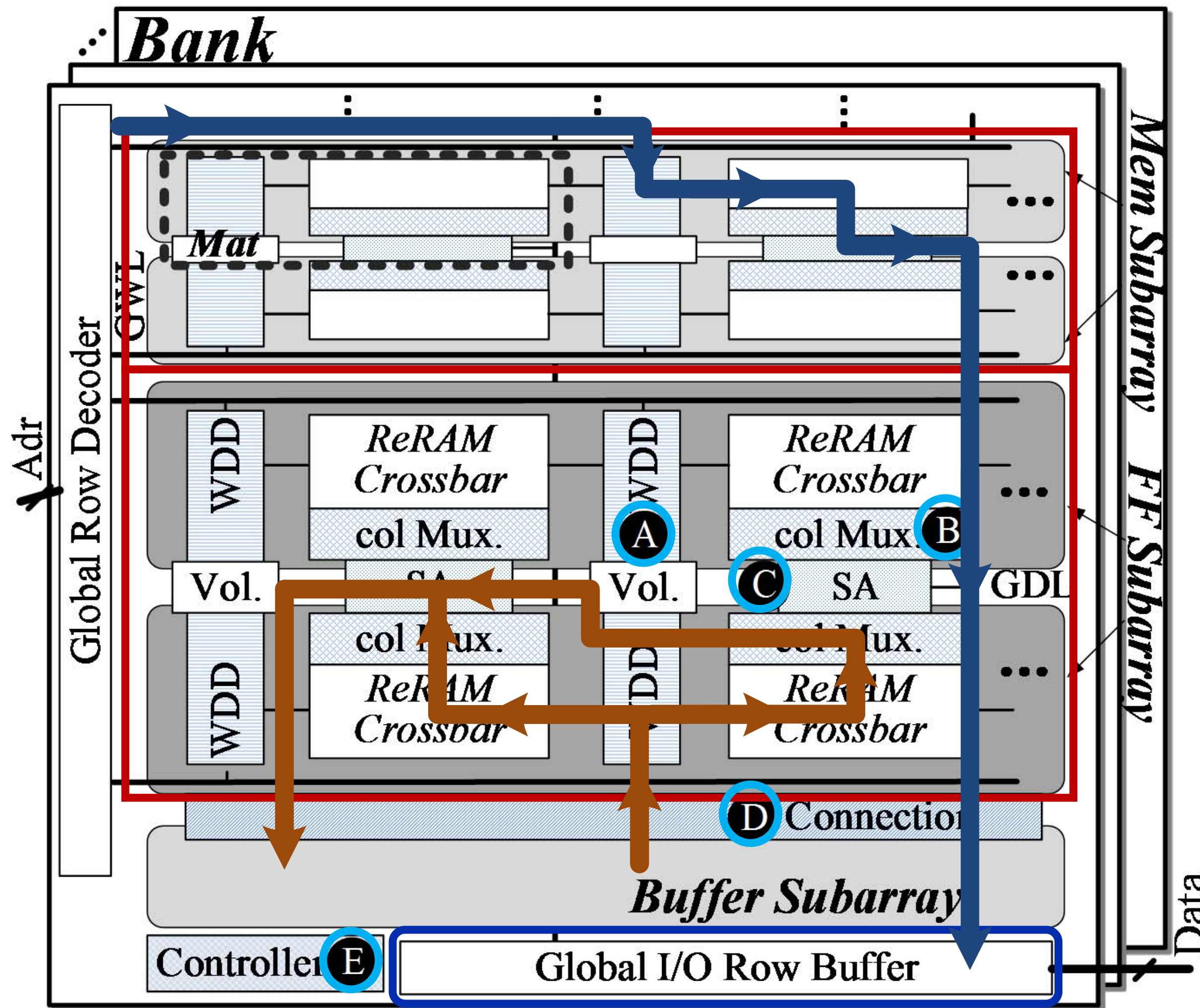


(a) An NN with one input and one output layer



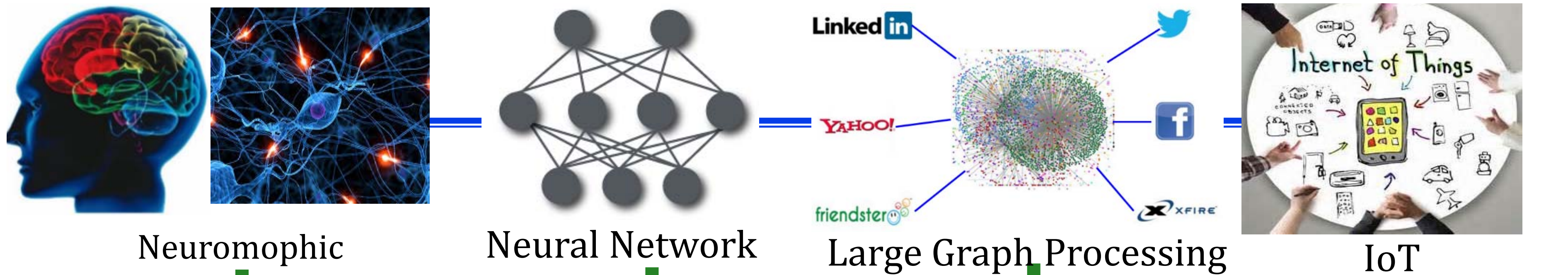
(b) using a ReRAM crossbar array for neural computation

PRIME Architecture Details



- (A) Wordline decoder and driver with multi-level voltage sources;
- (B) Column multiplexer with analog subtraction and sigmoid circuitry;
- (C) Reconfigurable SA with counters for multi-level outputs
- (D) Connection between the FF and Buffer subarrays;

Chi et al., "PRIME: A Novel Processing-in-memory Architecture for Neural Network Computation in ReRAM-based Main Memory", ISCA 2016



Application-Driven Innovations

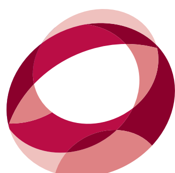
Computer Architecture Innovations

Technology-Driven Innovations

Heterogeneous Computing

Emerging Technology

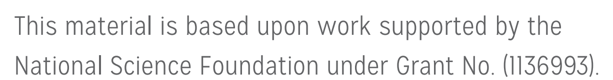
Arch2030: A Vision of Computer Architecture Research over the Next 15 Years



CCC

Computing Community Consortium

Catalyst



This material is based upon work supported by the National Science Foundation under Grant No. (1136993).

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Arch2030: A Vision of Computer Architecture Research over the Next 15 Years

Luis Ceze, Mark D. Hill, Thomas F. Wenisch

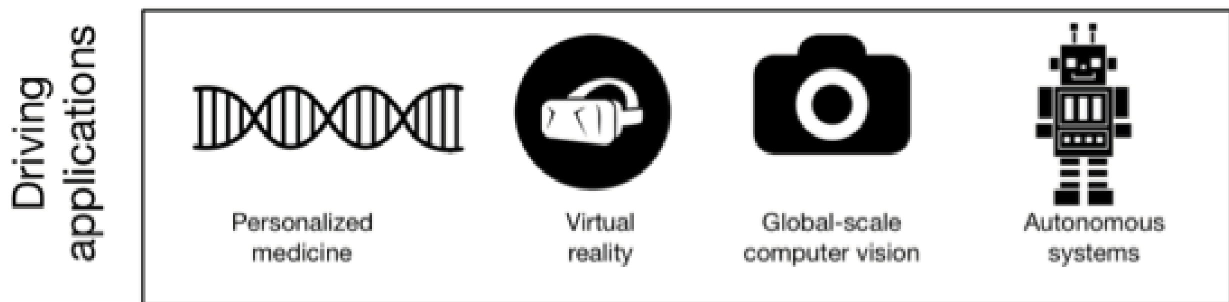
Sponsored by

Summary 1
The Specialization Gap: Democratizing Hardware Design 2
The Cloud as an Abstraction for Architecture Innovation 4
Going Vertical 5
Architectures “Closer to Physics” 5
Machine Learning as a Key Workload 6
About this document 7

Summary

Application trends, device technologies and the architecture of systems drive progress in information technologies. However, the former engines of such progress – Moore's Law and Dennard Scaling – are rapidly reaching the point of diminishing returns. The time has come for the computing community to boldly confront a new challenge: how to secure a foundational future for information technology's continued progress.

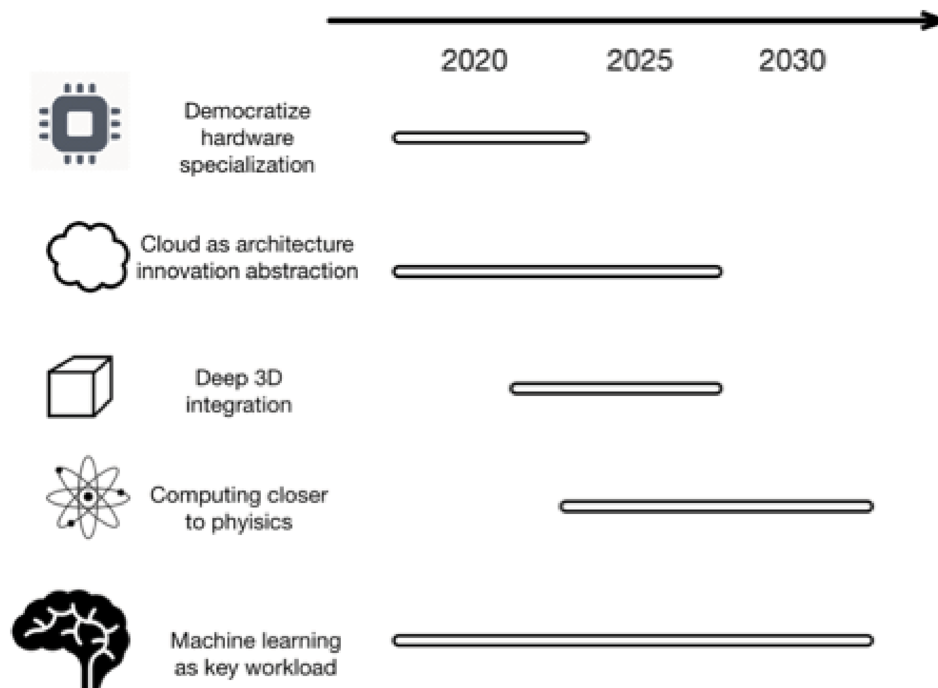
The computer architecture community engaged in several visioning exercises over the years. Five years ago, we released a white paper, *21st Century Computer Architecture*, which influenced funding programs in both academia and industry. More recently, the *IEEE Rebooting Computing Initiative* explored the future of computing systems in the architecture, device, and circuit domains.



This report stems from an effort to continue this dialogue, reach out to the applications and devices/circuits communities, and understand their trends and vision. We aim to identify opportunities where architecture research can bridge the gap between the application and device domains.

Why now? A lot has changed in just five years:

1. We now have a clear ***specialization gap*** — a gap between off-the-shelf hardware trends and application needs. Some applications, like virtual reality and autonomous systems, cannot be implemented without specialized hardware, yet hardware design remains expensive and difficult.
2. ***Cloud computing,*** now truly ubiquitous, provides a clear "innovation abstraction;" the Cloud creates economies of scale that make ingenious, cross-layer optimizations cost-effective, yet offers these innovations, often transparently, to even the smallest of new ventures and startups.
3. ***Going vertical*** with 3D integration, both with die stacking and monolithic fabrication, is enabling silicon substrates to grow vertically, significantly reducing latency, increasing bandwidth, and delivering efficiencies in energy consumption.
4. ***Getting closer to physics:*** device and circuit researchers are exploring the use of innovative materials that can provide more efficient switching, denser arrangements, or new computing models, e.g., mixed-signal, carbon nanotubes, quantum-mechanical phenomena, and biopolymers.
5. And finally, ***machine learning has emerged as a key workload;*** in many respects, machine learning techniques, such as deep learning, caught system designers "by surprise" as an enabler for diverse applications, such as user preference prediction, computer vision, or autonomous navigation.



We now describe each opportunity in greater detail.

The Specialization Gap: Democratizing Hardware Design

Developing hardware must become as easy, inexpensive, and agile as developing software to continue the virtuous history of computer industry innovation.

A widespread and emerging consensus maintains that classical CMOS technology scaling — the technical engine underlying Moore's Law that enables ever smaller transistors and denser integration — will come to an end in at most three more semiconductor technology generations (6-9 years)[1]. Further, Dennard scaling — the concomitant technical trend that enabled constant power per chip despite increasing CMOS integration density — ended in the mid-2000s[2][3], leading to a sea change in processor design: energy efficiency per operation has replaced area efficiency or peak switching speed as the most important design constraint limiting peak performance[4].

The effects of the imminent demise of classical scaling can be seen in recent industry announcements. Intel has abandoned its long-standing "tick-tock" model of releasing two major chip designs per technology generation, shifting instead to three designs; this extends the marketable lifetime of each generation as it drags the last gasps out of Moore's Law[5]. Further, the Semiconductor Industry Association has abandoned its biennial updates of the decades-old *International Technology Roadmap for Semiconductors*[6], a document that had been instrumental in coordinating technology, manufacturing, and system development across the industry. With no clear path to continued scaling, the value of the ITRS has ebbed.

[1] Chien and Karamcheti."Moore's Law: The First Ending and a New Beginning." *Computer* 46.12 (2013): 48-53.
[2] Fuller and Millett, "The Future of Computing Performance: Game Over or Next Level?," The National Academy Press, 2011 (http://books.nap.edu/openbook.php?record_id=12980&page=R1).
[3] Horowitz et al. "Scaling, power, and the future of CMOS." *IEEE International Electron Devices Meeting*, 2005.
[4] Mudge. "Power: A first-class architectural design constraint."*Computer* 34.4 (2001): 52-58.
[5] http://www.economist.com/technology-quarterly/2016-03-12/after-moores-law
[6] http://www.semiconductors.org/main/2015_international_technology_roadmap_for_semiconductors_itrs/

Yet, new applications continue to emerge that demand ever more computational capability. Foremost among these are the previously unimaginable applications enabled by large-scale machine learning, from image and speech recognition to self-driving cars to besting human experts at Go. Similar explosive growth can be seen in the need to process and understand visual data; some envisioned applications may demand the processing of gigapixels *per second for every human on earth*.

Past computing advances have been facilitated by the enormous investments in general-purpose computing designs enabled by classical scaling and made by only a handful of processor vendors. The large aggregate market of computing applications that benefited from these general-purpose designs amortized their substantial cost.

Given the twilight of classical scaling, continuing to meet emerging application performance demands by improving only a few general-purpose computing platforms is no longer feasible. Rather, over the past 5-10 years, a new strategy has emerged in some compute-intensive application domains: *specialized hardware design*. Specialized hardware (e.g., application-specific integrated circuits) can improve energy efficiency per operation by as much as 10,000 times over software running on a general-purpose chip[7]. The energy efficiency gains of specialization are critical to enable rich applications in the emerging Internet-of-Things. Specialization has been enormously successful in graphics rendering and video playback. Other initial evidence of commercial success is in machine learning applications. Indeed, the computer architecture research community has recognized and embraced specialization: of 175 papers in the 2016 computer architecture conferences (ISCA, HPCA, MICRO), 38 papers address specialization with GPUs or application-specific accelerators, while another 17 address specialized designs for machine learning.

However, commercial success of specialized designs, to date, has been demonstrated only for applications with enormous markets (e.g., video games, mobile video playback) that can justify investments of a scale similar to those made by general-purpose processor vendors. In terms of both time-to-market and dollars, the cost of designing and manufacturing specialized hardware is prohibitive for all but the few designs that can amortize it over such extensive markets.

To continue the virtuous innovation cycle, it is critical to reduce the barriers to application specific system design; to enable the energy efficiency advantages of specialization for all applications. Our vision is to "democratize" hardware design; that is, that hardware design become as agile, cheap, and open as software design. Software development teams can leverage a rich ecosystem of existing reusable components (often free and open source), use high-level languages to accelerate the capability of an individual developer, and rely on capable and automated program analysis, synthesis, testing, and debugging aids that help ensure high quality.

Despite decades of investment, computer-aided design has not delivered the same level of capability for hardware to a small development team. System designers require better tools to facilitate higher productivity in hardware description, more rapid performance evaluation, agile prototyping, and rigorous validation of hardware/software co-designed systems. Tool chains must mature to enable easy retargeting across multiple hardware substrates, from general purpose programmable cores to FPGAs, programmable accelerators, and ASICs. Better abstractions are needed for componentized/reusable hardware, possibly in the form of synthesizable intellectual property blocks or perhaps even physical chips/chiplets that can be integrated cheaply at manufacture. The architecture research community has an opportunity to lead in the effort to bridge the gap between general-purpose and specialized systems and deliver the tools and frameworks to make democratized hardware design a reality.

[7] Hameed et al. "Understanding sources of inefficiency in general-purpose chips." *International Symposium on Computer Architecture*, 2010.

The Cloud as an Abstraction for Architecture Innovation

By leveraging scale and virtualization, Cloud computing providers can offer hardware innovations transparently and at low cost to even the smallest of their customers.

The disruptive nature of Cloud computing to business-as-usual has been widely appreciated[8]. The Cloud lets new ventures scale far faster than traditional infrastructure investment. New products can grow from hundreds to millions of users in mere days, as evidenced by the meteoric launch of Pokemon Go in July 2016. However, the Cloud also disrupts traditional Fortune 500 business models since businesses that previously owned their own IT infrastructure realize the cost benefits derivable from leasing Cloud resources.

Less widely appreciated, however, is the Cloud computing model's ability to provide a powerful abstraction for cross-layer architectural innovation that was previously possible in only a very few, vertically integrated IT sectors (e.g., specialized high-performance supercomputers). The model provides two critical advantages: *scale* and *virtualization*.

Cloud computing providers can leverage scale not only for their own businesses, but for the benefit of their customers making investments in IT. As a result, these providers often find it cost effective to make enormous, non-recurring engineering investments, for example, to develop entirely new hardware and software systems in-house rather than relying on third-party vendor offerings.

We are beginning to see the emergence of specialized computer architectures enabling unprecedented performance in the Cloud. GPUs are becoming ubiquitous, not only in high-end supercomputers, but also in commercial Cloud offerings. Microsoft has publicly disclosed Catapult[9], its effort to integrate field-programmable gate arrays to facilitate compute specialization in its data centers. Cavium has released the ThunderX, a specialized architecture for Internet service applications. Google has disclosed the Tensor Processing Unit[10], a dedicated co-processor for machine learning applications. These projects demonstrate that the economic incentives are in place for Cloud providers to invest in computer architecture specialization.

For academic computer architecture researchers, now is the moment to seize this opportunity and present compelling visions for cross-layer specialization. For example, the ASIC Clouds effort presents a vision for how a large number of highly specialized processors can be deployed in concert to drastically accelerate critical applications[11]. The scale of the Cloud computing landscape has created a viable path for such academic proposals to demonstrate real, immediate impact. Another aspect of in-house specialization is the use of technologies that require special facilities, for example, atomic clocks for global time synchronization or superconducting logic that requires extremely low temperatures and makes sense only in a data-center environment.

The second critical advantage of the Cloud computing model is *virtualization*. By virtualization, we refer to a broad class of techniques that introduce new hardware and software innovations *transparently* to existing software systems. Virtualization lets a Cloud provider swap out processing, storage, and networking components for faster and cheaper technologies without requiring coordination with their customers. It also enables the oversubscription of resources — transparent sharing among customers with time-varying, fractional needs for a particular resource. Oversubscription is essential to the cost structure of Cloud computing: it lets Cloud providers offer IT resources at far lower prices than those individual customers would incur by purchasing dedicated resources.

---

[8] http://www.zdnet.com/article/eight-ways-that-cloud-computing-will-change-business/

[9] Putnam, et al. "A reconfigurable fabric for accelerating large-scale datacenter services." *ACM/IEEE 41st International Symposium on Computer Architecture*, 2014.

[10] https://cloudplatform.googleblog.com/2016/05/Google-supercharges-machine-learning-tasks-with-custom-chip.html

[11] Magaki et al. "ASIC Clouds: Specializing the Datacenter." *ACM/IEEE 43rd International Symposium on Computer Architecture*, 2016.

Academic computer architecture research has long been fundamental to enabling virtualization; indeed, VMWare, the most recognizable vendor of virtualization technology, was launched from a university research project. Academic architecture researchers must continue to play a key role in developing virtualization techniques that close the gap between virtualized and bare-metal performance. And, architecture researchers must develop new virtualization abstractions to enable transparent use and oversubscription of specialized hardware units, like the Catapult, TPU, or ASIC clouds.

Going Vertical

3D integration provides a new dimension of scalability.

A critical consequence of the end of Moore's Law is that chip designers can no longer scale the number of transistors in their designs "for free" every 18 months. Furthermore, over recent Silicon generations, driving global wires has grown increasingly expensive relative to computation, and hence interconnect accounts for an increasing fraction of the total chip power budget.

3D integration offers a new dimension of scalability in chip design, enabling the integration of more transistors in a single system despite an end of Moore's Law, shortening interconnects by routing in three dimensions, and facilitating the tight integration of heterogeneous manufacturing technologies. As a result, 3D integration enables greater energy efficiency, higher bandwidth, and lower latency between system components inside the 3D structure.

Architecturally, 3D integration also implies that computing must be near data for a balanced system. While 3D has long enabled capacity scaling in Flash and other memory devices, we are only now beginning to see integration of memory devices and high performance logic, for example, in Micron's Hybrid Memory Cube. 3D stacking has prompted a resurgence of academic research in "near-data computing" and "processing-in-memory" architectures, because it enables dense integration of fast logic and dense memory. Although this research topic was quite popular 20 years ago, processing-in-memory saw no commercial uptake in the 1990s due to manufacturability challenges. With the advent of practical die stacking and multi-technology vertical integration, such architectures now present a compelling path to scalability.

While 3D integration enables new capabilities, it also raises complex new challenges for achieving high reliability and yield that can be addressed with architecture support. For example, 3D-integrated memory calls to re-think traditional memory and storage hierarchies. 3D integration also poses novel problems for power and thermal management since traditional heat sink technology may be insufficient for the power density of high-performance integrated designs. Such problems and challenges open a new, rich field of architectural possibilities.

Architectures "Closer to Physics"

The end of classical scaling invites more radical changes to the computing substrate.

New device technologies and circuit design techniques have historically motivated new architectures. Going forward, several possibilities have significant architectural implications. These fall into two broad categories. The first is *better use of current materials and devices* by a more efficient encoding of information, one closer to analog. There has been a rebirth of interest in analog computing because of its good match to applications amenable to accuracy trade-offs. Further, analog information processing offers the promise of much lower power by denser mapping of information into signals and much more efficient functional units than their digital counterparts. However, such computing, more subject to noise, requires new approaches to error tolerance for it to make sense.

The second category of opportunities is *the use of "new" materials,* which can cover more efficient switching, denser arrangements, and unique computing models. Below we list a few prominent efforts worthy of the architecture community's attention.

New memory devices. For decades, data has been stored in DRAM, on Flash, or on rotating disk. However, we are now on the cusp of commercial availability

of new memory devices (e.g., Intel/Micron 3D XPoint memory[13]) that offer fundamentally different cost, density, latency, throughput, reliability, and endurance trade-offs than traditional memory/storage hierarchy components.

Carbon nanotubes. Electronics based on carbon nanotubes (CNTs) continues to make significant progress, with recent results showing simple microprocessors implemented entirely with CNTs[14]. CNTs promise greater density and lower power and can also be used in 3D substrates. This momentum makes CNTs a viable area for architects' consideration.

Quantum computing. Quantum computing uses quantum mechanics phenomena to store and manipulate information. Its key advantage is that the "superposition" quantum phenomenon effectively allows representation of 0 and 1 states simultaneously, which can be leveraged for exponential speed-ups compared to classical computing for select algorithms.

A sister effort of quantum computing is **superconducting logic.** Systems that use superconducting devices, such as Josephson junctions, offer "free" communication because they consume little energy to move a signal over a superconducting wire[12]. Operations on data, on the other hand, are more expensive than moving data. These trade-offs are the reverse of those in silicon CMOS, where most energy is dissipated in communication rather than operations on the data path.

Microsoft, Google, IBM and I-ARPA have publicized significant investments in quantum computing and superconducting logic. We conclude that the time is ripe for renewed academic interest in quantum computer architectures, with a likely path to practical impact within a decade.

Borrowing from biology. The use of biological substrates in computing has long been considered a possibility in several aspects of computer systems. DNA computing has demonstrated simple logic operations and more recent results show the potential of using DNA as a digital medium for archival storage and for self-assembly of nanoscale structure[15]. Progress in DNA manipulation[16] fueled by the biotech industry is making the use of biomaterials a more viable area for consideration among architecture researchers. Beyond DNA, there are other biomolecules that could be used for computing such as proteins, whose engineering advanced significantly in the past decade[17].

Machine Learning as a Key Workload

Machine Learning is changing the way we implement applications. Hardware advancement makes machine learning over big data possible.

Machine learning (ML) has made significant progress over the last decade in producing applications that have long been in the realm of science fiction, from long-sought, practical voice-based interfaces to self-driving cars. One can claim that this progress has been largely fueled by abundant data coupled with copious compute power. Large-scale machine learning applications have motivated designs that range from storage systems to specialized hardware (GPUs, TPUs).

While the current focus is on supporting ML in the Cloud, significant opportunities exist to support ML applications in low-power devices, such as smartphones or ultra-low power sensor nodes. Luckily, many ML kernels have relatively regular structures and are amenable to accuracy-resource trade-offs; hence, they lend themselves to hardware specialization, reconfiguration, and approximation techniques, opening up a significant space for architectural innovation.

---

[12] "Superconducting Computing and the IARPA C3 Program", http://beyondcmos.ornl.gov/documents/Session%203_talk1-Holmes.pdf

[13] http://www.intel.com/content/www/us/en/architecture-and-technology/non-volatile-memory.html

[14] https://www.technologyreview.com/s/519421/the-first-carbon-nanotube-computer/

[15] http://people.ee.duke.edu/~dwyer/pubs/TVLSI_dnaguided.pdf

[16] http://www.synthesis.cc/synthesis/2016/03/on_dna_and_transistors

[17] http://www.sciencemag.org/news/2016/07/protein-designer-aims-revolutionize-medicines-and-materials

Machine learning practitioners spend considerable time on computation to train their models. Anecdotal evidence suggests that week- to month-long training jobs are common, even when using warehouse-scale infrastructure. While such computational investments hopefully amortize over many invocations of the resulting model, the slow turnaround of new models can negatively affect the user experience. Consequently, architecture researchers have new opportunities to design systems that better support ML model training.

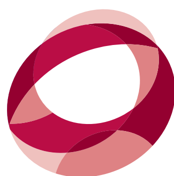
About this document

This document resulted from discussions held during the Arch2030 Workshop[18] at ISCA 2016, organized by Luis Ceze and Thomas Wenisch and shepherded by Mark Hill. The organizers also solicited input from the community in the form of an open survey as well as direct comments on this document. Ceze and Wenisch drafted the text and revised it based on community feedback. The contributors listed below provided feedback on drafts of this document and have granted their endorsement to its content.

Endorsers:

Luis Ceze, University of Washington
Thomas F. Wenisch, University of Michigan
Mark D. Hill, University of Wisconsin-Madison
Sarita Adve, University of Illinois at Urbana-Champaign
Alvin R. Lebeck Duke University
Michael Taylor, University of California San Diego
Josep Torrellas, University of Illinois at Urbana Champaign
Karin Strauss, Microsoft
Dan Sorin, Duke University
Doug Burger, Microsoft
Tom Conte, Georgia Institute of Technology, co-chair of the IEEE Rebooting Computing Initiative
Babak Falsafi, EPFL
Fred Chong, University of Chicago
Rakesh Kumar, University of Illinois at Urbana-Champaign
Todd Austin, University of Michigan
Christos Kozyrakis, Stanford University
Karu Sankaralingam, UW-Madison
James Tuck, NC State University
Trevor Mudge, University of Michigan
Martha Kim, Columbia University
Stephen W. Keckler, NVIDIA
Vikram Adve, University of Illinois at Urbana-Champaign
Timothy Sherwood, UC Santa Barbara
Kathryn S McKinley, Microsoft Research
Yuan Xie, UCSB
Lieven Eeckhout, Ghent University
Andrew Putnam, Microsoft
Nikos Hardavellas, Northwestern University
James Larus, EPFL IC
Simha Sethumadhavan, Columbia University
Andreas Moshovos, University of Toronto
David J. Lilja, University of Minnesota
Guri Sohi, University of Wisconsin-Madison
Antonio Gonzalez, UPC Barcelona
Jack Sampson, Penn State
Natalie Enright Jerger, University of Toronto
Mark Oskin, University of Washington
Ulya Karpuzcu, University of Minnesota
David Kaeli, Northeastern University

[18] The workshop was supported by the Computing Community Consortium.



1828 L Street, NW, Suite 800
Washington, DC 20036
P: 202 234 2111 F: 202 667 1066
www.cra.org cccinfo@cra.org