OAK RIDGE
National Laboratory

Preparing for Extremely Heterogeneous Systems

Jeffrey S. Vetter

With many contributions from FTG Group and Colleagues

PPAM 2019
Bialystok, Poland
11 Sep 2019

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

Future Technologies Group

http://ft.ornl.gov vetter@computer.org

Highlights

- Recent trends in extreme-scale HPC paint an ambiguous future
  - Contemporary systems provide evidence that power constraints are driving architectures to change rapidly
  - Multiple architectural dimensions are being (dramatically) redesigned: Processors, node design, memory systems, I/O
  - Complexity is our main challenge
- Applications and software systems are all reaching a state of crisis
  - Applications will not be functionally or performance portable across architectures
  - Programming and operating systems need major redesign to address these architectural changes
  - Procurements, acceptance testing, and operations of today’s new platforms depend on performance prediction and benchmarking.
- We need portable programming models and performance prediction now more than ever!
  - Heterogeneous processing
    - OpenACC->FGPAs
    - Intelligent runtime system (IRIS) (not covered today)
    - Clacc – OpenACC support in LLVM (not covered today)
  - Emerging memory hierarchies (NVM)
    - DRAGON – transparent NVM access from GPUs (not covered today)
    - NVL-C – user management of nonvolatile memory in C (not covered today)
    - Papyrus – parallel aggregate persistent storage (not covered today)
- Performance prediction is critical for design and optimization (not covered today)

OAK RIDGE
National Laboratory

Time for a short poll…

Q: Think back 10 years. How many of you would have predicted that many of our top HPC systems would be GPU-based architectures?

Yes

No

Revisionist ☺

OAK RIDGE
National Laboratory

Q: Think forward 10 years. How many of you predict that most of our top HPC systems will have the following architectural features?

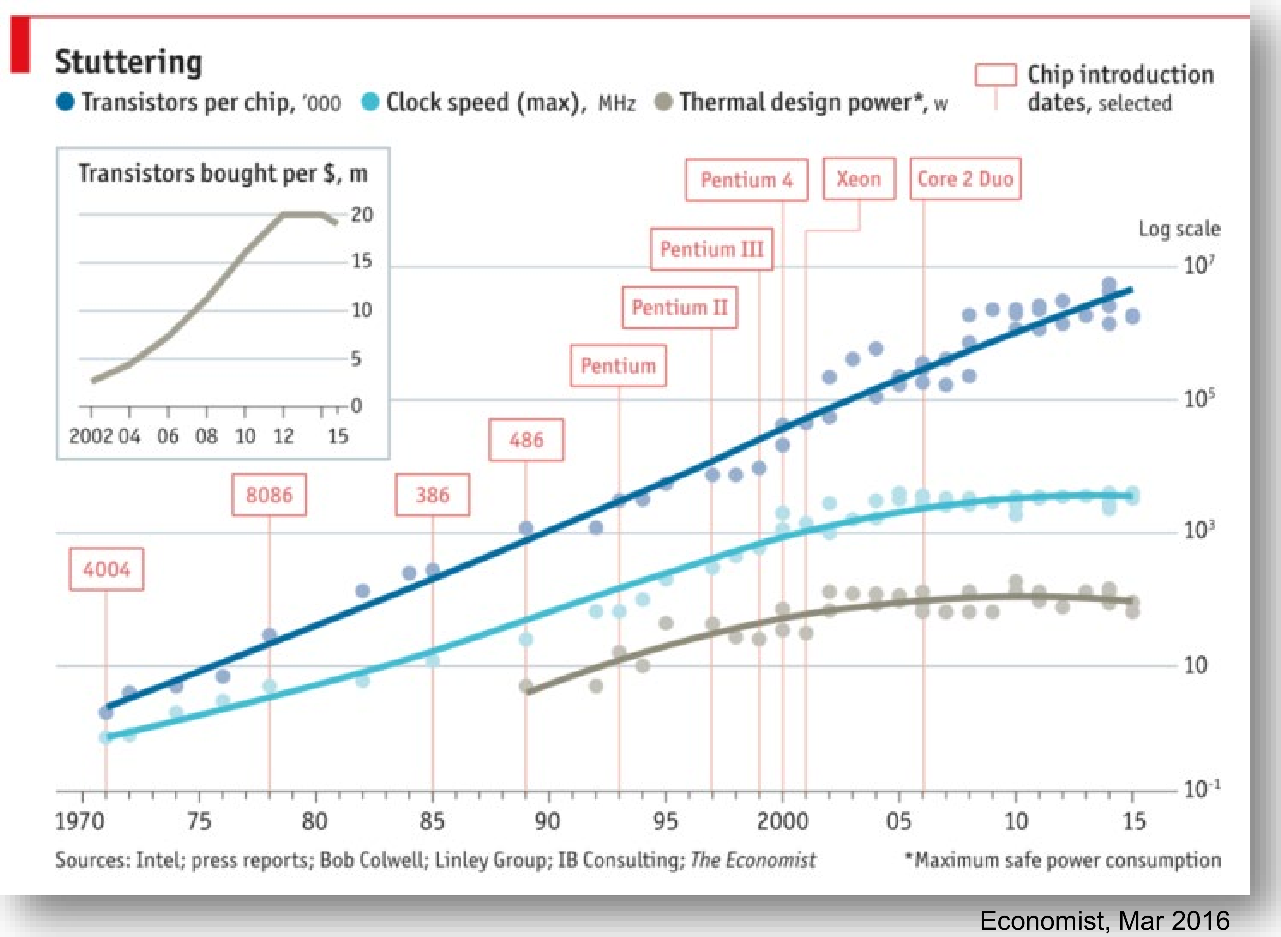
- X86 multicore CPU
- GPU
- FPGA/Reconfigurable processor
- Neuromorphic processor
- Deep learning processor
- Quantum processor
- RISC-V processor
- Some new unknown processor
- All/some of the above in one SoC

Q: Now imagine you are building a new application with ~3M LOC and 20 team members over the next 10 years. What on-node programming model/system do you use?

- C, C++, Fortran
- C++ templates, policies, etc (e.g., AMP, Kokkos, RAJA,)
- CUDA, cu***, HIP
- OpenCL, SYCL
- OpenMP or OpenACC
- R, Python, Matlab, etc
- A Domain Specific Language (e.g., Claw, PySL)
- A Domain Specific Framework (e.g., PetSc)
- Some new unknown programming approach
- All/some of the above

Motivating Trends

Contemporary devices are approaching fundamental limits



Sources: Intel; press reports; Bob Colwell; Linley Group; IB Consulting; *The Economist*

*Maximum safe power consumption

Economist, Mar 2016

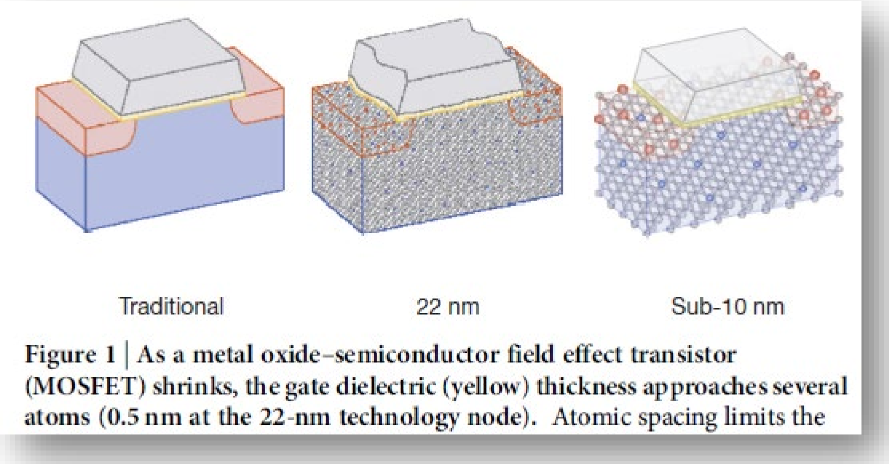


Figure 1 | As a metal oxide–semiconductor field effect transistor (MOSFET) shrinks, the gate dielectric (yellow) thickness approaches several atoms (0.5 nm at the 22-nm technology node). Atomic spacing limits the

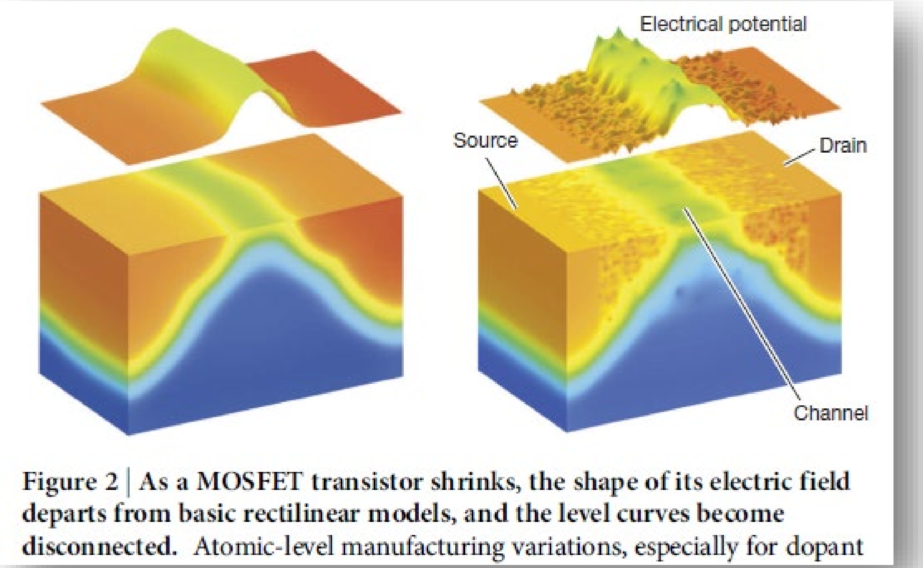
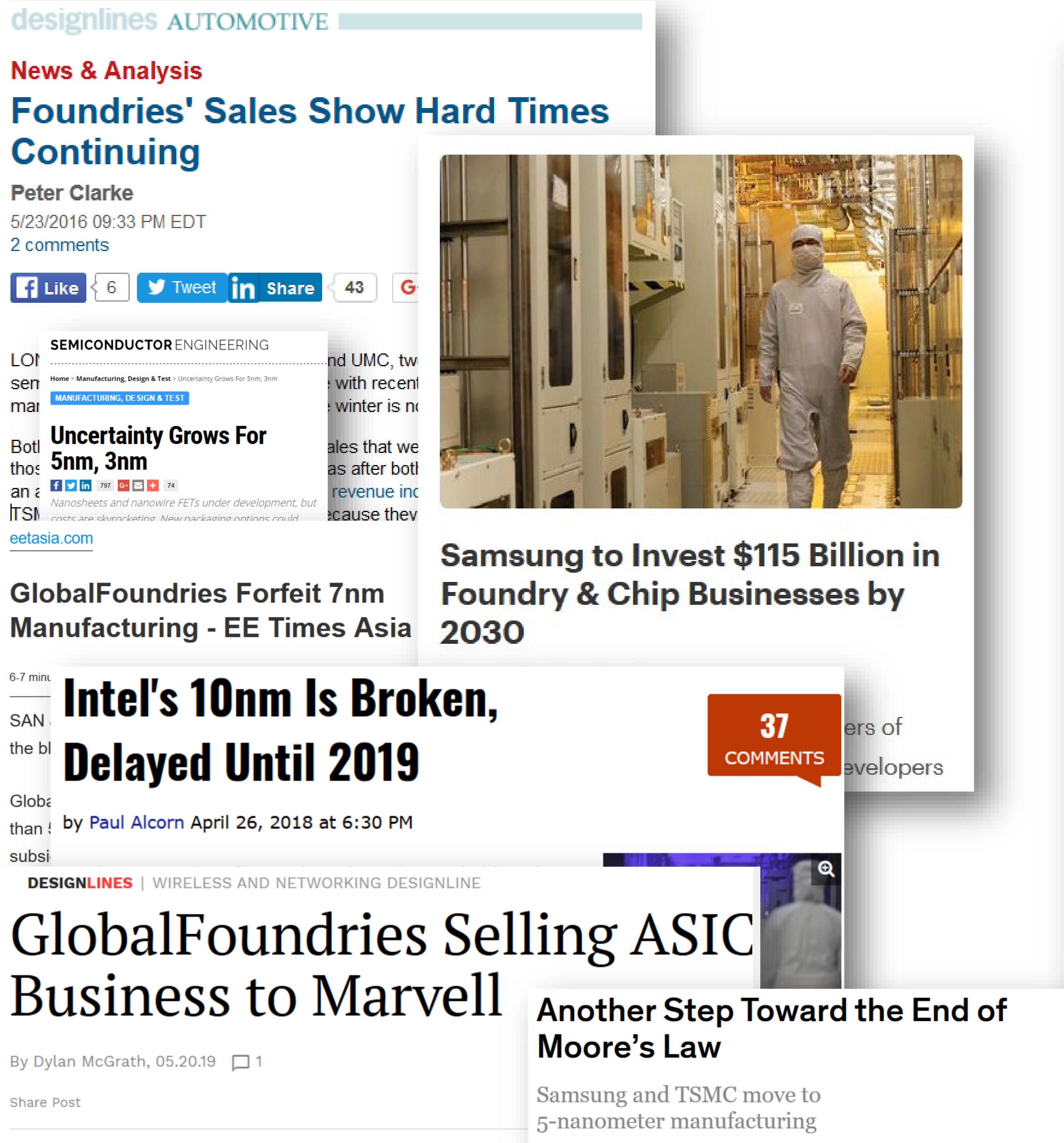


Figure 2 | As a MOSFET transistor shrinks, the shape of its electric field departs from basic rectilinear models, and the level curves become disconnected. Atomic-level manufacturing variations, especially for dopant

Dennard scaling has already ended. Dennard observed that voltage and current should be proportional to the linear dimensions of a transistor: 2x transistor count implies 40% faster and 50% more efficient.

R.H. Dennard, F.H. Gaensslen, V.L. Rideout, E. Bassous, and A.R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits, 9(5):256-68, 1974,*

I.L. Markov, "Limits on fundamental limits to computation," *Nature, 512(7513):147-54, 2014, doi:10.1038/nature13570.*



designlines AUTOMOTIVE

News & Analysis

Foundries' Sales Show Hard Times Continuing

Peter Clarke
5/23/2016 09:33 PM EDT
2 comments

Like 6 Tweet in Share 43 G

LON nd UMC, tw
sem with recent
mar winter is no
Both ales that we
thos as after both
an a revenue inc
TSM ecause they

SEMICONDUCTOR ENGINEERING

Home > Manufacturing, Design & Test > Uncertainty Grows For 5nm, 3nm

MANUFACTURING, DESIGN & TEST

Uncertainty Grows For 5nm, 3nm

797 74

Nanosheets and nanowire FETs under development, but costs are skyrocketing. New packaging options could

eetasia.com

GlobalFoundries Forfeit 7nm Manufacturing - EE Times Asia

6-7 minu

SAN
the bl

Globa
than
subsi

Samsung to Invest $115 Billion in Foundry & Chip Businesses by 2030

ers of
evelopers

Intel's 10nm Is Broken, Delayed Until 2019

37 COMMENTS

by Paul Alcorn April 26, 2018 at 6:30 PM

DESIGNLINES | WIRELESS AND NETWORKING DESIGNLINE

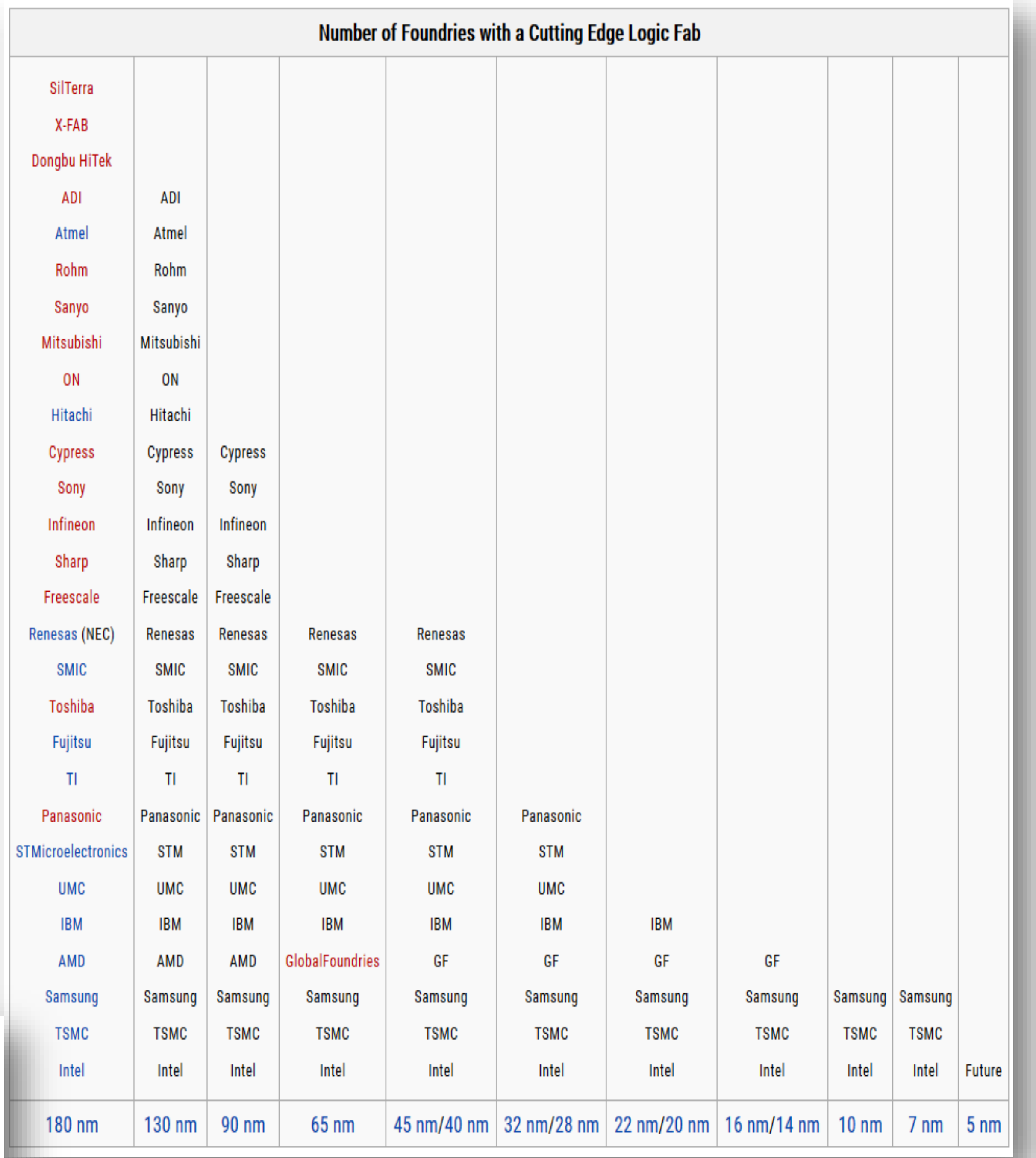
GlobalFoundries Selling ASIC Business to Marvell

By Dylan McGrath, 05.20.19 1

Share Post

Another Step Toward the End of Moore's Law

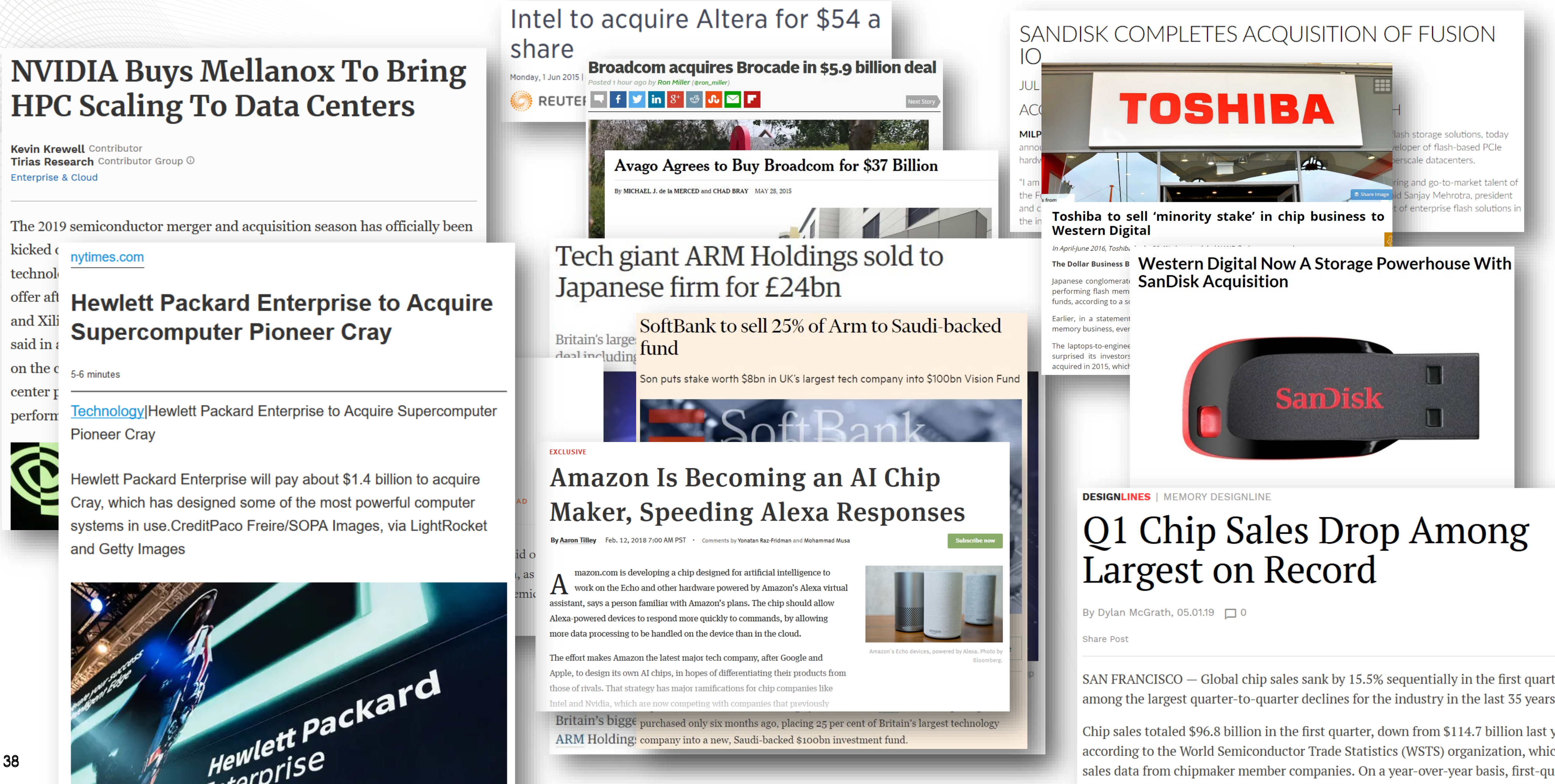
Samsung and TSMC move to 5-nanometer manufacturing



Number of Foundries with a Cutting Edge Logic Fab

SilTerra										
X-FAB										
Dongbu HiTek										
ADI	ADI									
Atmel	Atmel									
Rohm	Rohm									
Sanyo	Sanyo									
Mitsubishi	Mitsubishi									
ON	ON									
Hitachi	Hitachi									
Cypress	Cypress	Cypress								
Sony	Sony	Sony								
Infineon	Infineon	Infineon								
Sharp	Sharp	Sharp								
Freescale	Freescale	Freescale								
Renesas (NEC)	Renesas	Renesas	Renesas	Renesas						
SMIC	SMIC	SMIC	SMIC	SMIC						
Toshiba	Toshiba	Toshiba	Toshiba	Toshiba						
Fujitsu	Fujitsu	Fujitsu	Fujitsu	Fujitsu						
TI	TI	TI	TI	TI						
Panasonic	Panasonic	Panasonic	Panasonic	Panasonic	Panasonic					
STMicroelectronics	STM	STM	STM	STM	STM					
UMC	UMC	UMC	UMC	UMC	UMC					
IBM	IBM	IBM	IBM	IBM	IBM	IBM				
AMD	AMD	AMD	GlobalFoundries	GF	GF	GF	GF			
Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung	
TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	TSMC	
Intel	Intel	Intel	Intel	Intel	Intel	Intel	Intel	Intel	Intel	Future
180 nm	130 nm	90 nm	65 nm	45 nm/40 nm	32 nm/28 nm	22 nm/20 nm	16 nm/14 nm	10 nm	7 nm	5 nm

Business climate reflects this uncertainty, cost, complexity, consolidation



NVIDIA Buys Mellanox To Bring HPC Scaling To Data Centers

Kevin Krewell Contributor
Tirias Research Contributor Group ⓘ
Enterprise & Cloud

The 2019 semiconductor merger and acquisition season has officially been kicked technol offer aft and Xili said in on the center perform

Intel to acquire Altera for $54 a share

Monday, 1 Jun 2015 |

REUTER

Broadcom acquires Brocade in $5.9 billion deal

Posted 1 hour ago by Ron Miller (@ron_miller)

Next Story

Avago Agrees to Buy Broadcom for $37 Billion

By MICHAEL J. de la MERCED and CHAD BRAY MAY 28, 2015

SANDISK COMPLETES ACQUISITION OF FUSION IO

JUL

AC H

MILP annou hardw flash storage solutions, today veloper of flash-based PCIe perscale datacenters.

"I am the F and c the in ring and go-to-market talent of id Sanjay Mehrotra, president t of enterprise flash solutions in

Share Image

s from

Toshiba to sell 'minority stake' in chip business to Western Digital

In April-June 2016, Toshiba

The Dollar Business B

Japanese conglomerate performing flash mem funds, according to a s

Earlier, in a statement memory business, ever

The laptops-to-enginee surprised its investors acquired in 2015, which

Western Digital Now A Storage Powerhouse With SanDisk Acquisition

nytimes.com

Hewlett Packard Enterprise to Acquire Supercomputer Pioneer Cray

5-6 minutes

Technology|Hewlett Packard Enterprise to Acquire Supercomputer Pioneer Cray

Hewlett Packard Enterprise will pay about $1.4 billion to acquire Cray, which has designed some of the most powerful computer systems in use.CreditPaco Freire/SOPA Images, via LightRocket and Getty Images

Tech giant ARM Holdings sold to Japanese firm for £24bn

Britain's large deal including

SoftBank to sell 25% of Arm to Saudi-backed fund

Son puts stake worth $8bn in UK's largest tech company into $100bn Vision Fund

purchased only six months ago, placing 25 per cent of Britain's largest technology company into a new, Saudi-backed $100bn investment fund.

Britain's bigge ARM Holding

EXCLUSIVE

Amazon Is Becoming an AI Chip Maker, Speeding Alexa Responses

By Aaron Tilley Feb. 12, 2018 7:00 AM PST · Comments by Yonatan Raz-Fridman and Mohammad Musa

Subscribe now

Amazon.com is developing a chip designed for artificial intelligence to work on the Echo and other hardware powered by Amazon's Alexa virtual assistant, says a person familiar with Amazon's plans. The chip should allow Alexa-powered devices to respond more quickly to commands, by allowing more data processing to be handled on the device than in the cloud.

The effort makes Amazon the latest major tech company, after Google and Apple, to design its own AI chips, in hopes of differentiating their products from those of rivals. That strategy has major ramifications for chip companies like Intel and Nvidia, which are now competing with companies that previously

Amazon's Echo devices, powered by Alexa. Photo by Bloomberg.

AD

id o , as emic

DESIGNLINES | MEMORY DESIGNLINE

Q1 Chip Sales Drop Among Largest on Record

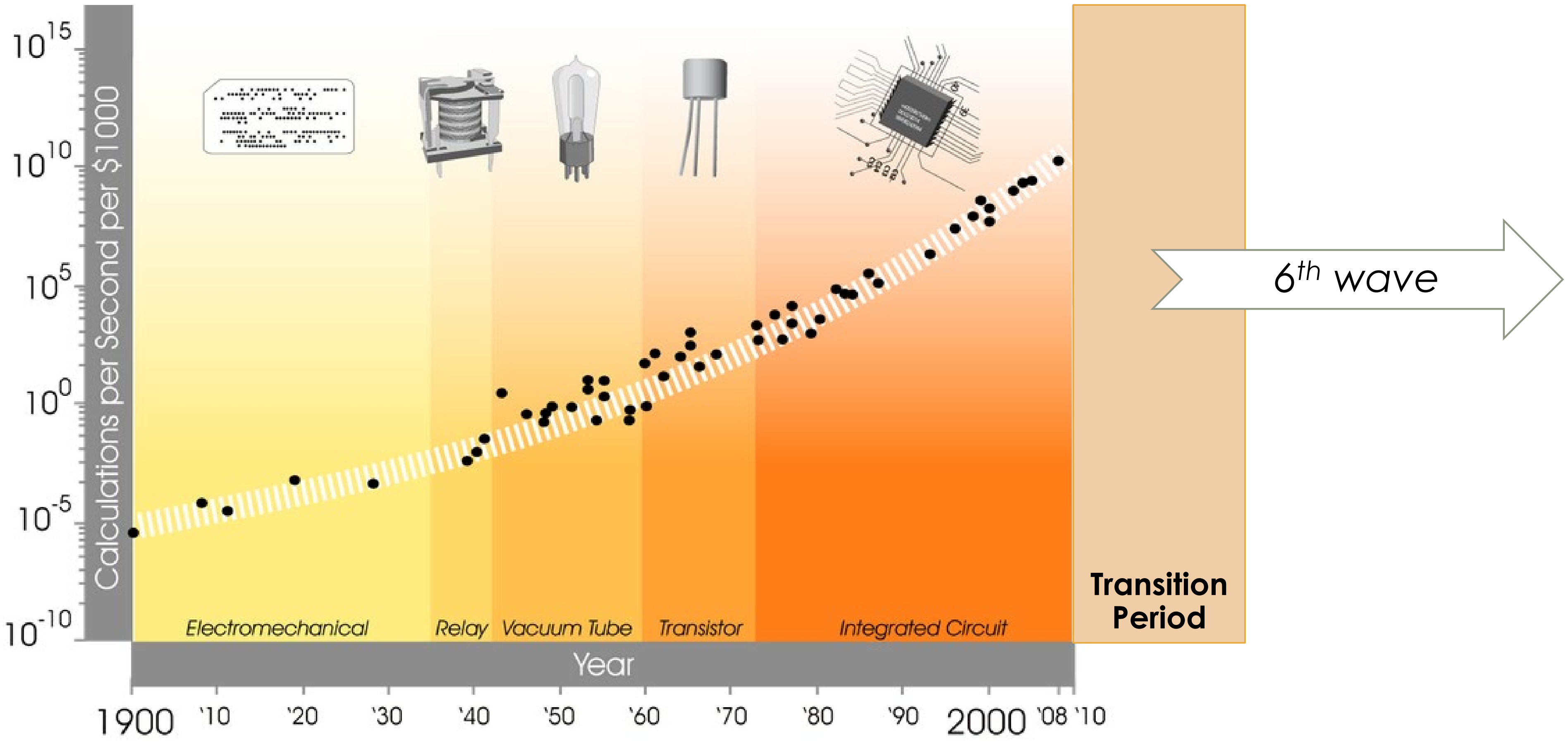
By Dylan McGrath, 05.01.19 🗨 0

Share Post

SAN FRANCISCO — Global chip sales sank by 15.5% sequentially in the first quart among the largest quarter-to-quarter declines for the industry in the last 35 years

Chip sales totaled $96.8 billion in the first quarter, down from $114.7 billion last y according to the World Semiconductor Trade Statistics (WSTS) organization, whic sales data from chipmaker member companies. On a year-over-year basis, first-qu

Sixth Wave of Computing



Transition Period

6th wave

http://www.kurzweilai.net/exponential-growth-of-computing

Predictions for Transition Period

Optimize Software and Expose New Hierarchical Parallelism

- Redesign software to boost performance on upcoming architectures
- Exploit new levels of parallelism and efficient data movement

Architectural Specialization and Integration

- Use CMOS more effectively for specific workloads
- Integrate components to boost performance and eliminate inefficiencies
- Workload specific memory+storage system design

Emerging Technologies

- Investigate new computational paradigms
  - Quantum
  - Neuromorphic
  - Advanced Digital
  - Emerging Memory Devices

Predictions for Transition Period

Optimize Software and Expose New Hierarchical Parallelism

- Redesign software to boost performance on upcoming architectures
- Exploit new levels of parallelism and efficient data movement

Architectural Specialization and Integration

- Use CMOS more effectively for specific workloads
- Integrate components to boost performance and eliminate inefficiencies
- Workload specific memory+storage system design

Emerging Technologies

- Investigate new computational paradigms
  - Quantum
  - Neuromorphic
  - Advanced Digital
  - Emerging Memory Devices

Predictions for Transition Period

Optimize Software and Expose New Hierarchical Parallelism

- Redesign software to boost performance on upcoming architectures
- Exploit new levels of parallelism and efficient data movement

Architectural Specialization and Integration

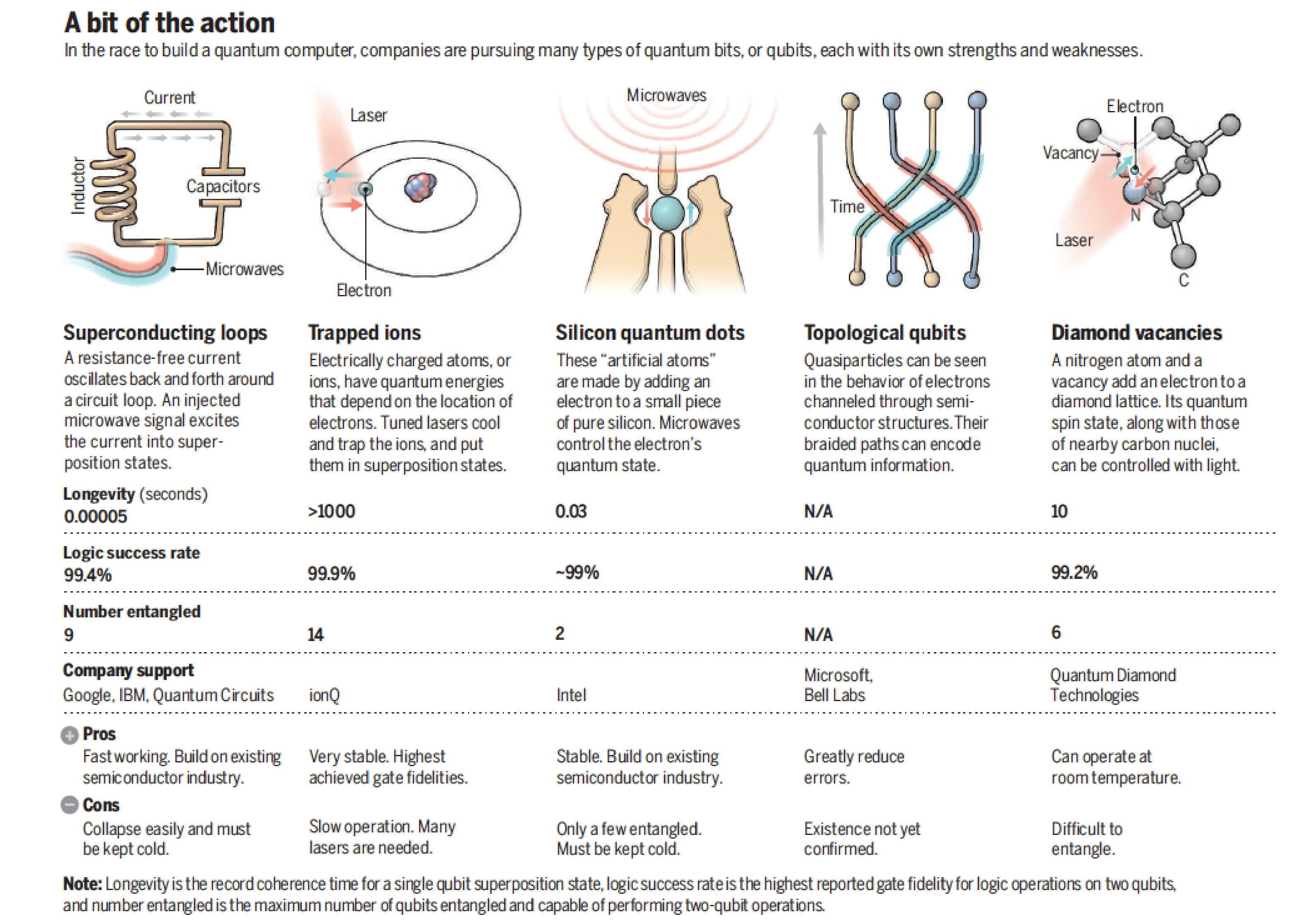
- Use CMOS more effectively for specific workloads
- Integrate components to boost performance and eliminate inefficiencies
- Workload specific memory+storage system design

Emerging Technologies

- Investigate new computational paradigms
  - Quantum
  - Neuromorphic
  - Advanced Digital
  - Emerging Memory Devices

Quantum computing: Qubit design and fabrication have made recent progress but still face challenges

Science 354, 1091 (2016) – 2 December



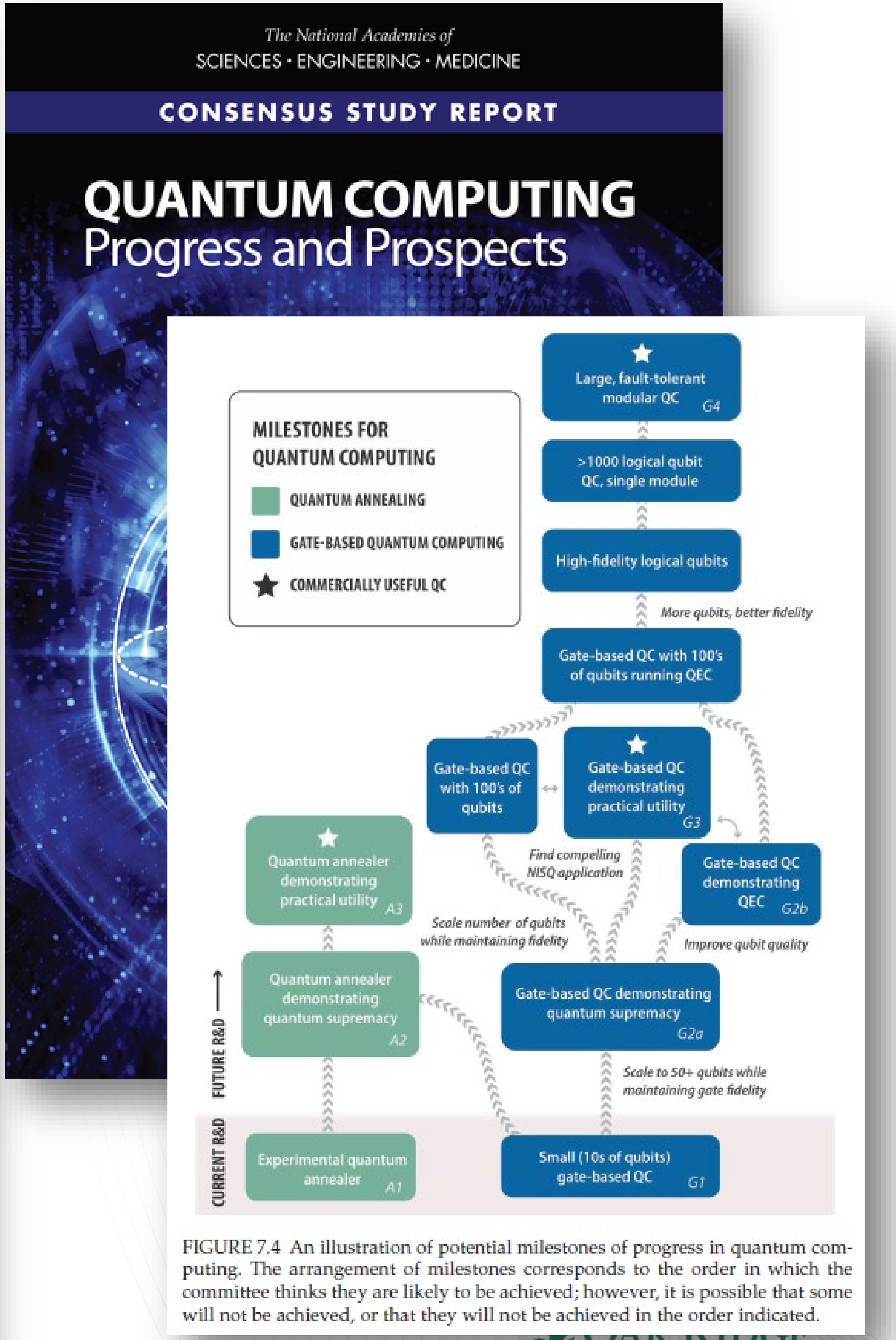
Current
Inductor
Capacitors
Microwaves
Laser
Electron
Microwaves
Time
Electron
Vacancy
Laser
N
C

A bit of the action

In the race to build a quantum computer, companies are pursuing many types of quantum bits, or qubits, each with its own strengths and weaknesses.

Superconducting loops	Trapped ions	Silicon quantum dots	Topological qubits	Diamond vacancies
A resistance-free current oscillates back and forth around a circuit loop. An injected microwave signal excites the current into superposition states.	Electrically charged atoms, or ions, have quantum energies that depend on the location of electrons. Tuned lasers cool and trap the ions, and put them in superposition states.	These "artificial atoms" are made by adding an electron to a small piece of pure silicon. Microwaves control the electron's quantum state.	Quasiparticles can be seen in the behavior of electrons channeled through semiconductor structures. Their braided paths can encode quantum information.	A nitrogen atom and a vacancy add an electron to a diamond lattice. Its quantum spin state, along with those of nearby carbon nuclei, can be controlled with light.
Longevity (seconds) 0.00005	>1000	0.03	N/A	10
Logic success rate 99.4%	99.9%	~99%	N/A	99.2%
Number entangled 9	14	2	N/A	6
Company support Google, IBM, Quantum Circuits	ionQ	Intel	Microsoft, Bell Labs	Quantum Diamond Technologies
⊕ **Pros** Fast working. Build on existing semiconductor industry.	Very stable. Highest achieved gate fidelities.	Stable. Build on existing semiconductor industry.	Greatly reduce errors.	Can operate at room temperature.
⊖ **Cons** Collapse easily and must be kept cold.	Slow operation. Many lasers are needed.	Only a few entangled. Must be kept cold.	Existence not yet confirmed.	Difficult to entangle.

Note: Longevity is the record coherence time for a single qubit superposition state, logic success rate is the highest reported gate fidelity for logic operations on two qubits, and number entangled is the maximum number of qubits entangled and capable of performing two-qubit operations.



The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

QUANTUM COMPUTING
Progress and Prospects

MILESTONES FOR QUANTUM COMPUTING

🟩 QUANTUM ANNEALING

🟦 GATE-BASED QUANTUM COMPUTING

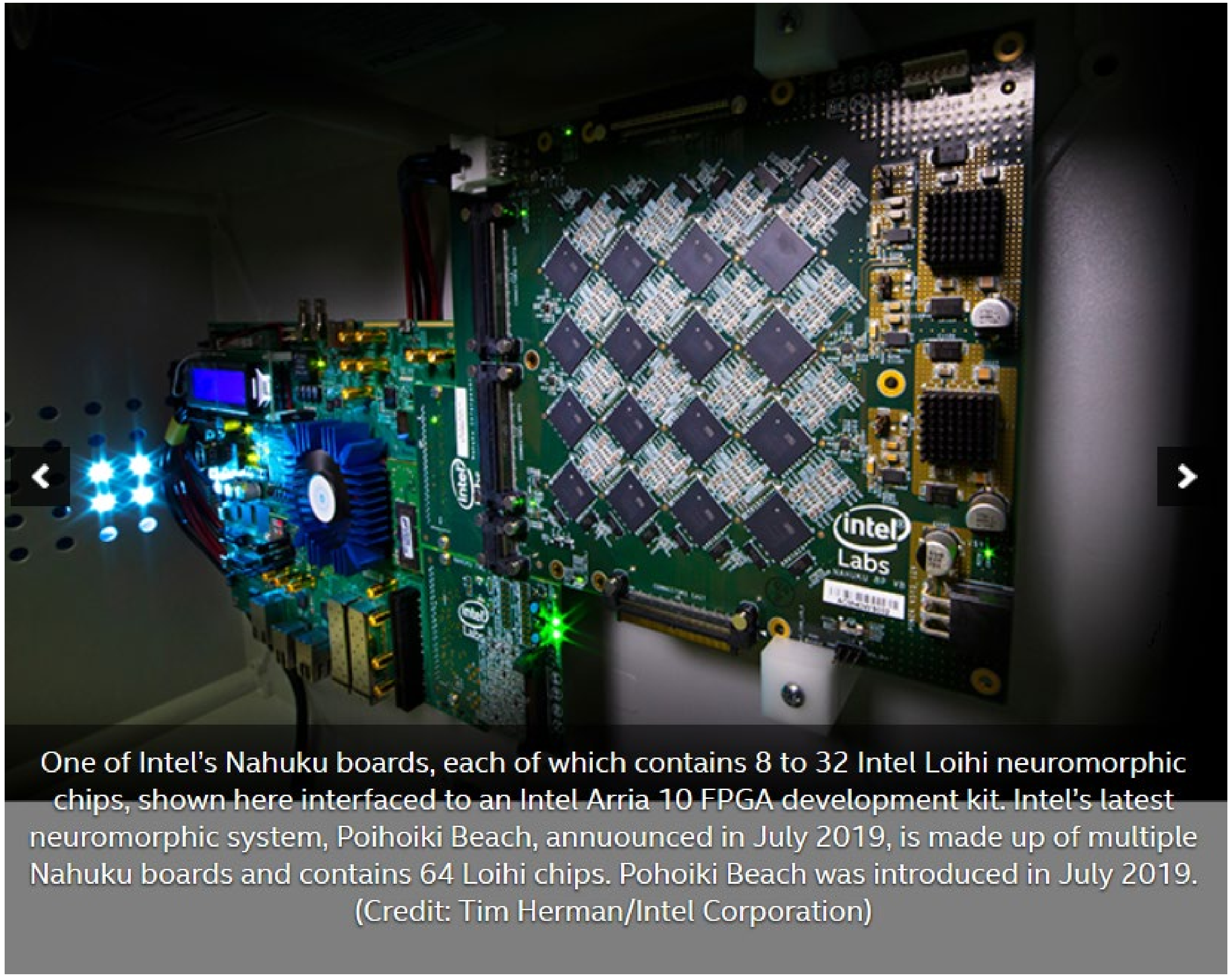
★ COMMERCIALLY USEFUL QC

★ Large, fault-tolerant modular QC G4
>1000 logical qubit QC, single module
High-fidelity logical qubits
More qubits, better fidelity
Gate-based QC with 100's of qubits running QEC
Gate-based QC with 100's of qubits
★ Gate-based QC demonstrating practical utility G3
Find compelling NISQ application
Gate-based QC demonstrating QEC G2b
★ Quantum annealer demonstrating practical utility A3
Scale number of qubits while maintaining fidelity
Improve qubit quality
Quantum annealer demonstrating quantum supremacy A2
Gate-based QC demonstrating quantum supremacy G2a
Scale to 50+ qubits while maintaining gate fidelity
FUTURE R&D
CURRENT R&D
Experimental quantum annealer A1
Small (10s of qubits) gate-based QC G1

FIGURE 7.4 An illustration of potential milestones of progress in quantum computing. The arrangement of milestones corresponds to the order in which the committee thinks they are likely to be achieved; however, it is possible that some will not be achieved, or that they will not be achieved in the order indicated.

http://nap.edu/25196

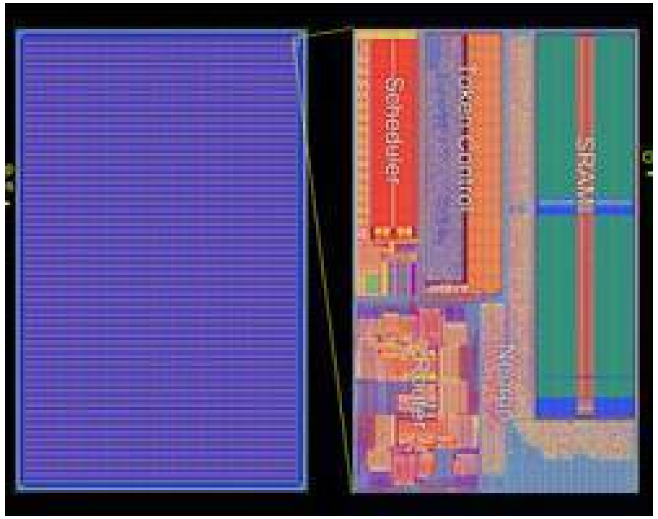
Neuromorphic (Brain Inspired) Computing



One of Intel's Nahuku boards, each of which contains 8 to 32 Intel Loihi neuromorphic chips, shown here interfaced to an Intel Arria 10 FPGA development kit. Intel's latest neuromorphic system, Poihoiki Beach, annuounced in July 2019, is made up of multiple Nahuku boards and contains 64 Loihi chips. Pohoiki Beach was introduced in July 2019. (Credit: Tim Herman/Intel Corporation)

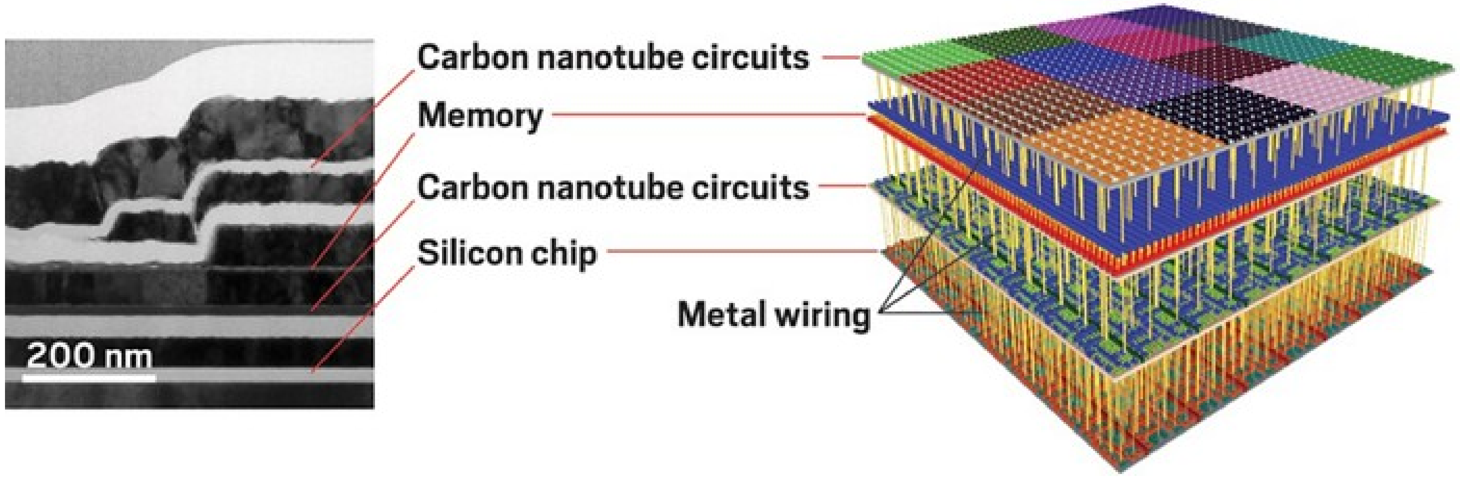
https://newsroom.intel.com/news/intels-pohoiki-beach-64-chip-neuromorphic-system-delivers-breakthrough-results-research-tests/

- SpiNNaker
- IBM True North
- BrainScaleS
- DANNA
- Others…

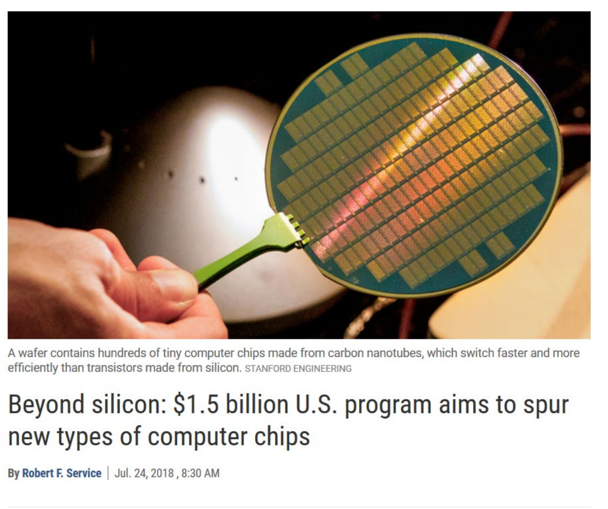


https://m-cacm.acm.org/news/201072-the-future-of-microchips

New devices: Carbon Nanotube Transistors and Circuits



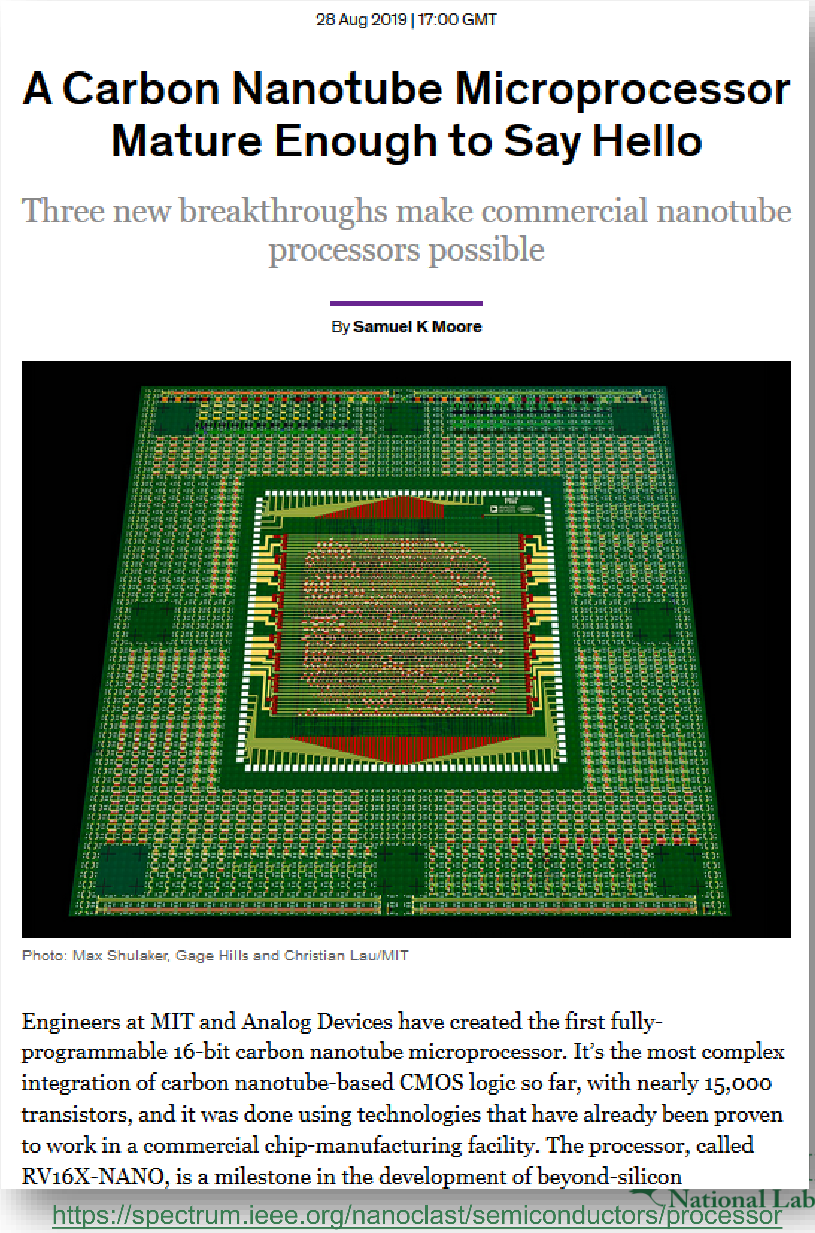
https://cen.acs.org/materials/electronic-materials/Carbon-nanotube-computers-face-makebreak/97/i8



A wafer contains hundreds of tiny computer chips made from carbon nanotubes, which switch faster and more efficiently than transistors made from silicon. STANFORD ENGINEERING

Beyond silicon: $1.5 billion U.S. program aims to spur new types of computer chips

By Robert F. Service | Jul. 24, 2018 , 8:30 AM



28 Aug 2019 | 17:00 GMT

A Carbon Nanotube Microprocessor Mature Enough to Say Hello

Three new breakthroughs make commercial nanotube processors possible

By Samuel K Moore

Photo: Max Shulaker, Gage Hills and Christian Lau/MIT

Engineers at MIT and Analog Devices have created the first fully-programmable 16-bit carbon nanotube microprocessor. It's the most complex integration of carbon nanotube-based CMOS logic so far, with nearly 15,000 transistors, and it was done using technologies that have already been proven to work in a commercial chip-manufacturing facility. The processor, called RV16X-NANO, is a milestone in the development of beyond-silicon

https://spectrum.ieee.org/nanoclast/semiconductors/processors/modern-microprocessor-built-using-carbon-nanotubes

Emerging Memory Devices

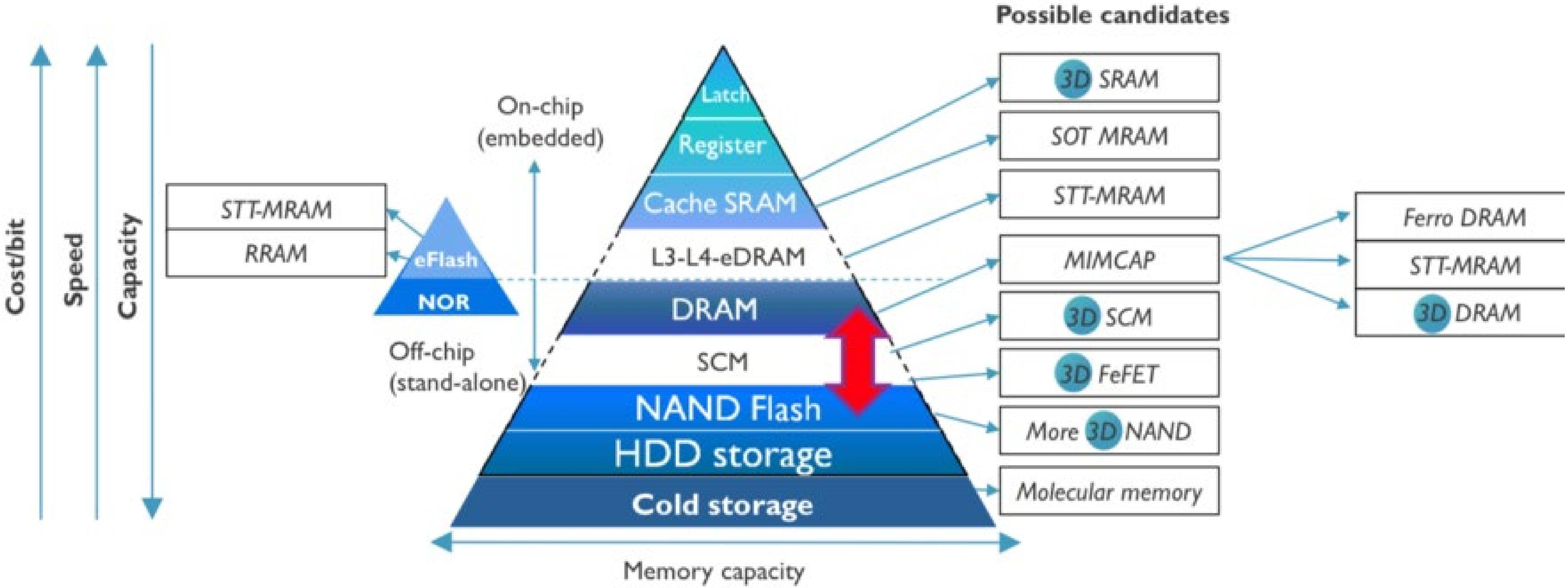


Image Source: IMEC

Predictions for Transition Period

Optimize Software and Expose New Hierarchical Parallelism

- Redesign software to boost performance on upcoming architectures
- Exploit new levels of parallelism and efficient data movement

Architectural Specialization and Integration

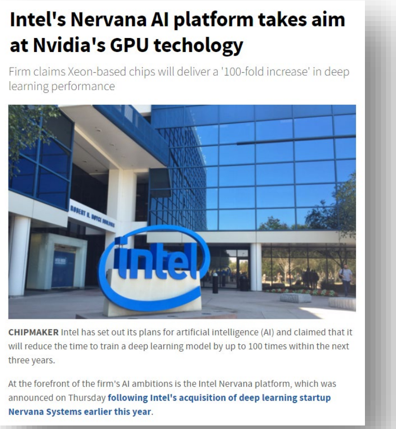
- Use CMOS more effectively for specific workloads
- Integrate components to boost performance and eliminate inefficiencies
- Workload specific memory+storage system design

Emerging Technologies

- Investigate new computational paradigms
  - Quantum
  - Neuromorphic
  - Advanced Digital
  - Emerging Memory Devices

Pace of Architectural Specialization is Quickening

- Industry, lacking Moore's Law, will need to continue to differentiate products (to stay in business)
  - Use the same transistors differently to enhance performance
- Architectural design will become extremely important, critical
  - Dark Silicon
  - Address new parameters for benefits/curse of Moore's Law
- 50+ new companies focusing on hardware for Machine Learning



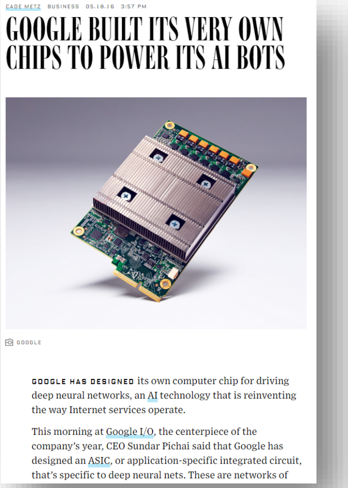
Intel's Nervana AI platform takes aim at Nvidia's GPU techology

Firm claims Xeon-based chips will deliver a '100-fold increase' in deep learning performance

CHIPMAKER Intel has set out its plans for artificial intelligence (AI) and claimed that it will reduce the time to train a deep learning model by up to 100 times within the next three years.

At the forefront of the firm's AI ambitions is the Intel Nervana platform, which was announced on Thursday **following Intel's acquisition of deep learning startup Nervana Systems earlier this year**.

http://www.theinquirer.net/inquirer/news/2477796/intels-nervana-ai-platform-takes-aim-at-nvidias-gpu-techology



CADE METZ BUSINESS 05.18.16 3:57 PM

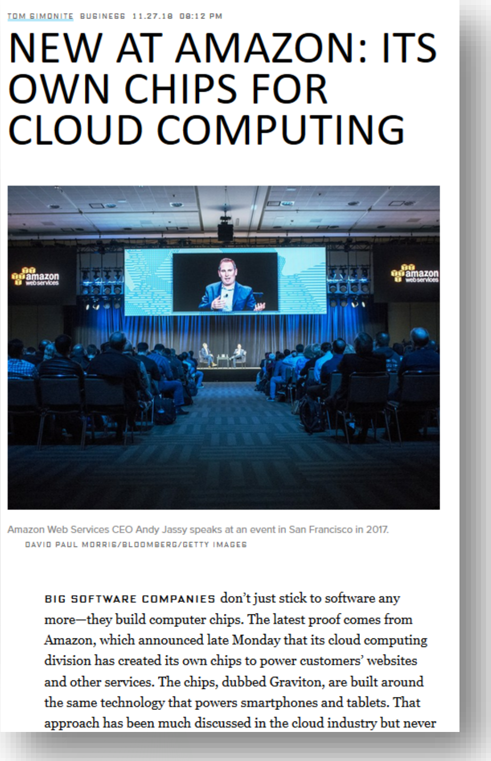
GOOGLE BUILT ITS VERY OWN CHIPS TO POWER ITS AI BOTS

GOOGLE

GOOGLE HAS DESIGNED its own computer chip for driving deep neural networks, an AI technology that is reinventing the way Internet services operate.

This morning at Google I/O, the centerpiece of the company's year, CEO Sundar Pichai said that Google has designed an ASIC, or application-specific integrated circuit, that's specific to deep neural nets. These are networks of

http://www.wired.com/2016/05/google-tpu-custom-chips/



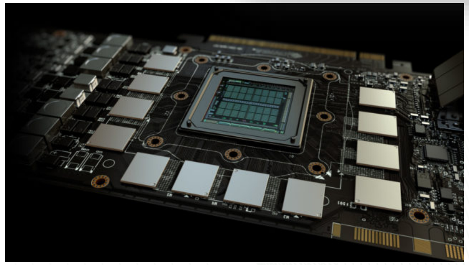
TOM SIMONITE BUSINESS 11.27.18 08:12 PM

NEW AT AMAZON: ITS OWN CHIPS FOR CLOUD COMPUTING

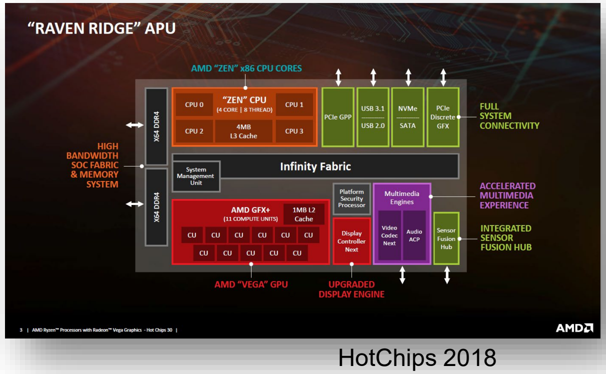
Amazon Web Services CEO Andy Jassy speaks at an event in San Francisco in 2017.
DAVID PAUL MORRIS/BLOOMBERG/GETTY IMAGES

BIG SOFTWARE COMPANIES don't just stick to software any more—they build computer chips. The latest proof comes from Amazon, which announced late Monday that its cloud computing division has created its own chips to power customers' websites and other services. The chips, dubbed Graviton, are built around the same technology that powers smartphones and tablets. That approach has been much discussed in the cloud industry but never

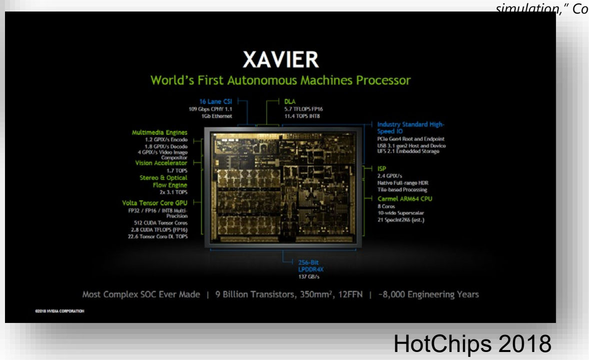
D.E. Shaw, M.M. Deneroff, R.O. Dror *et al., "Anton, a special-purpose machine for molecular dynamics simulation," Communications of the ACM, 51(7):91-7, 2008.*



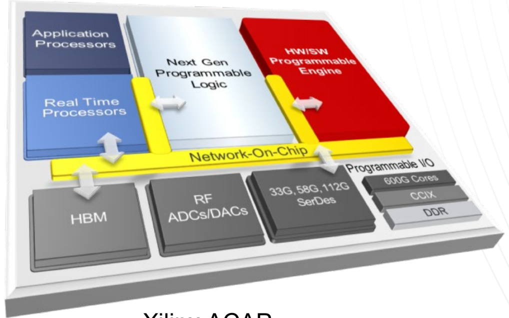
https://fossbytes.com/nvidia-volta-gddr6-2018/



HotChips 2018



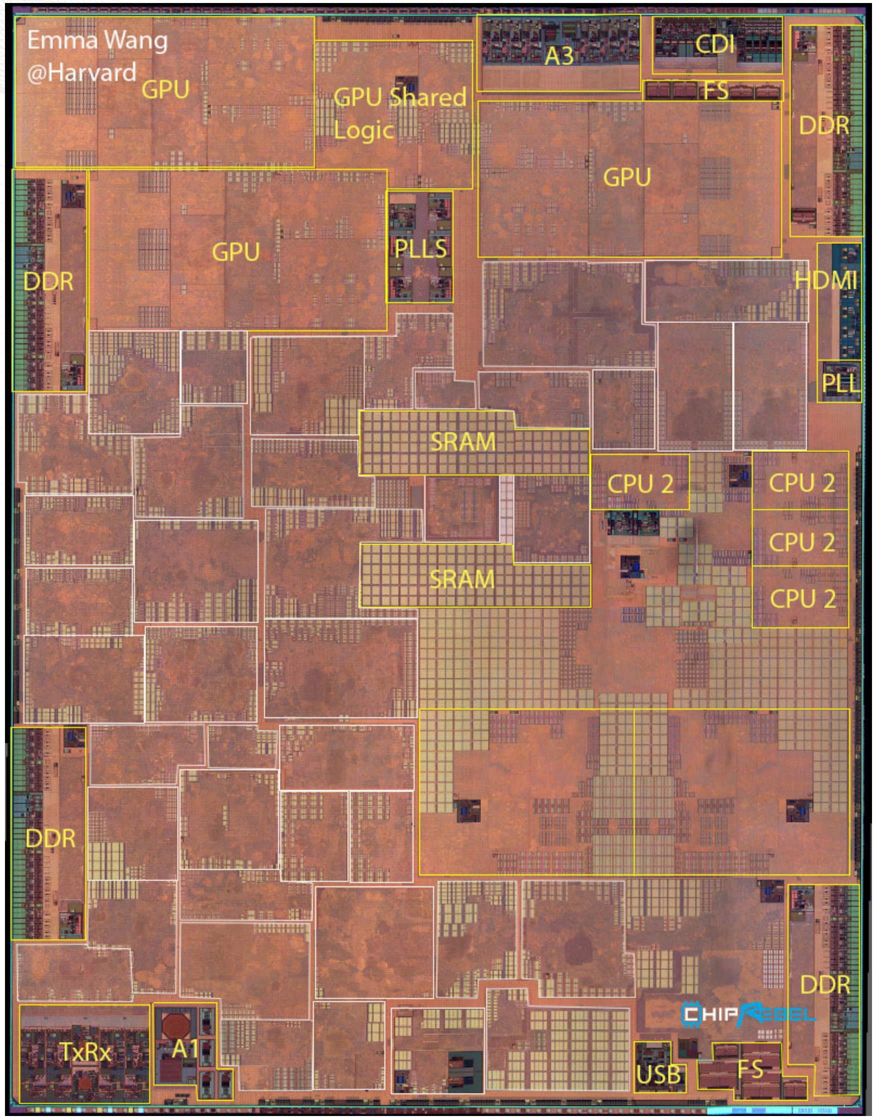
HotChips 2018

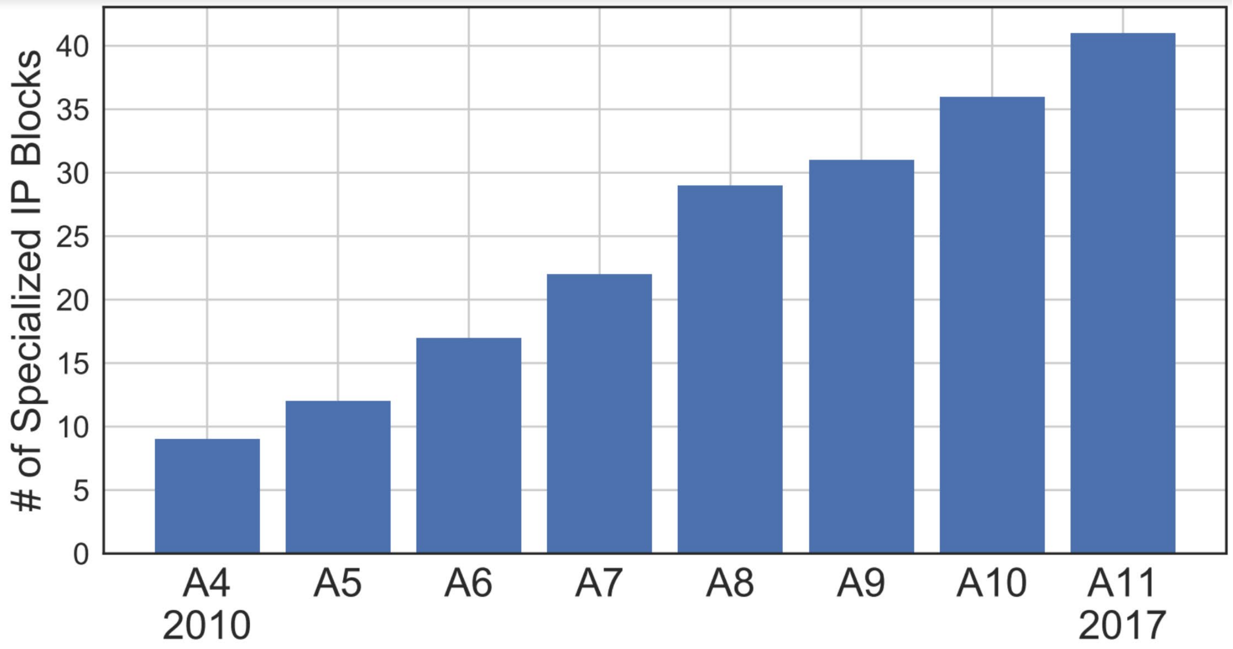


Xilinx ACAP

https://www.thebroadcastbridge.com/content/entry/1094/altera-announces-arria-10-2666mbps-ddr4-memory-fpga-interface

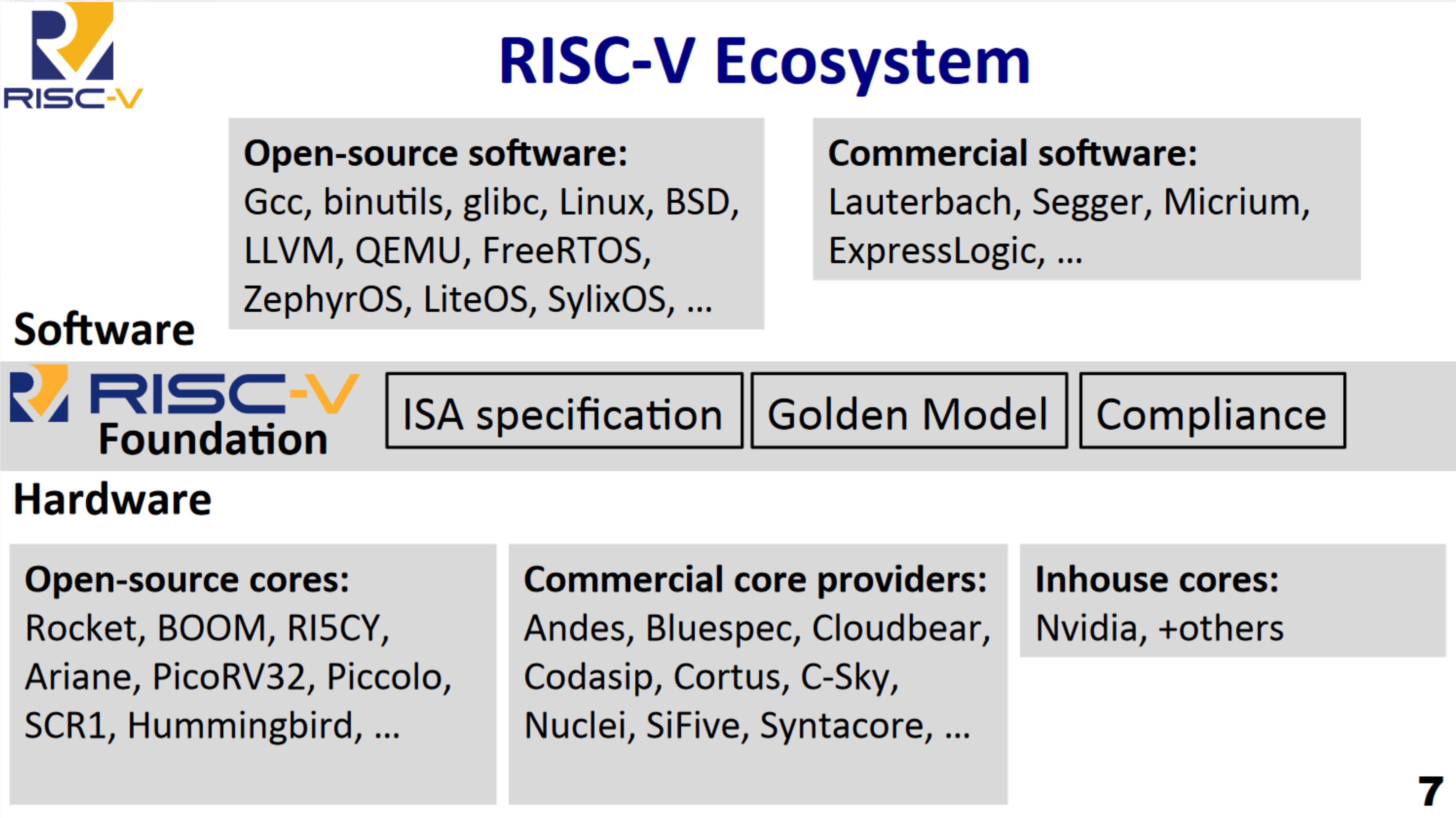
Analysis of Apple A-* SoCs





http://vlsiarch.eecs.harvard.edu/accelerators/die-photo-analysis

Growing Open Source Hardware Movement Enables Rapid Chip Design



RISC-V Ecosystem

Software

Open-source software:
Gcc, binutils, glibc, Linux, BSD, LLVM, QEMU, FreeRTOS, ZephyrOS, LiteOS, SylixOS, …

Commercial software:
Lauterbach, Segger, Micrium, ExpressLogic, …

RISC-V Foundation

ISA specification | Golden Model | Compliance

Hardware

Open-source cores:
Rocket, BOOM, RI5CY, Ariane, PicoRV32, Piccolo, SCR1, Hummingbird, …

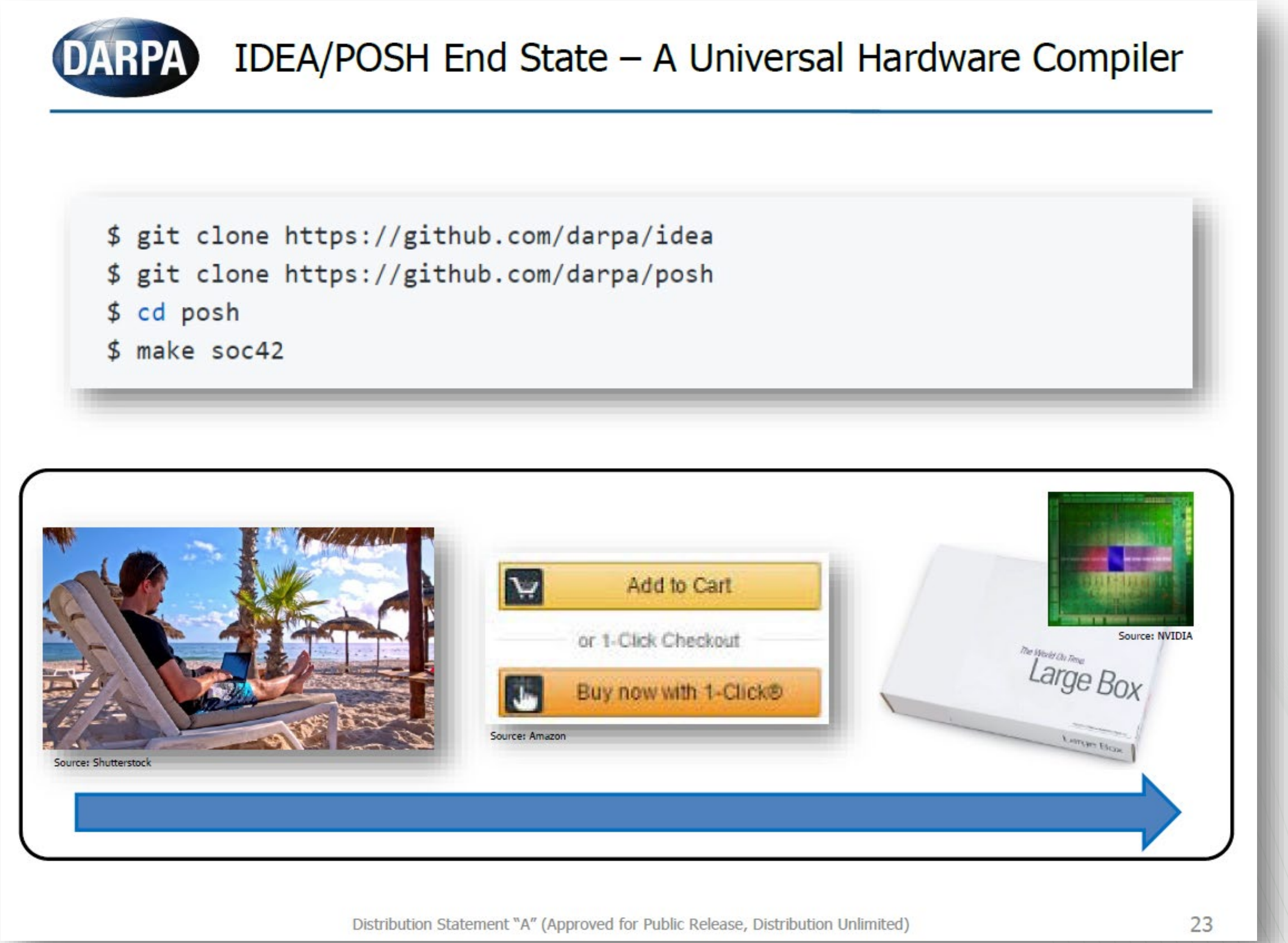
Commercial core providers:
Andes, Bluespec, Cloudbear, Codasip, Cortus, C-Sky, Nuclei, SiFive, Syntacore, …

Inhouse cores:
Nvidia, +others

7

RISC-V Summit, 2018

DARPA ERI Programs Aiming for Agile (and Frequent) Chip Creation



IDEA/POSH End State – A Universal Hardware Compiler

```
$ git clone https://github.com/darpa/idea
$ git clone https://github.com/darpa/posh
$ cd posh
$ make soc42
```

Source: Shutterstock

Source: Amazon

Source: NVIDIA

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

23

A. Olofsson, 2018

Oak Ridge National Laboratory

Summary:
Transition Period will be Disruptive – Opportunities and Pitfalls Abound

- New devices and architectures may not be hidden in traditional levels of abstraction

- Examples
  - A new type of CNT transistor may be completely hidden from higher levels
  - A new paradigm like quantum may require new architectures, programming models, and algorithmic approaches

Layer	Switch, 3D	NVM	Approximate	Neuro	Quantum
Application	1	1	2	2	3
Algorithm	1	1	2	3	3
Language	1	2	2	3	3
API	1	2	2	3	3
Arch	1	2	2	3	3
ISA	1	2	2	3	3
Microarch	2	3	2	3	3
FU	2	3	2	3	3
Logic	3	3	2	3	3
Device	3	3	2	3	3

Adapted from IEEE Rebooting Computing Chart

Department of Energy (DOE) Roadmap to Exascale Systems

An impressive, productive lineup of *accelerated node* systems supporting DOE’s mission

Pre-Exascale Systems [Aggregate Linpack (Rmax) = 323 PF!]

First U.S. Exascale Systems

2012	2016	2018	2020	2021-2023

ORNL
Cray/AMD/NVIDIA

ANL
IBM BG/Q

LLNL
IBM BG/Q

ANL
Cray/Intel KNL

LBNL
Cray/Intel Xeon/KNL

LANL/SNL
Cray/Intel Xeon/KNL

ORNL
IBM/NVIDIA

LLNL
IBM/NVIDIA

LBNL
Cray/AMD/NVIDIA

CROSSROADS

LANL/SNL
TBD

Heterogeneous Cores

Deep Memory incl NVM

Plateauing I/O Performance

FRONTIER

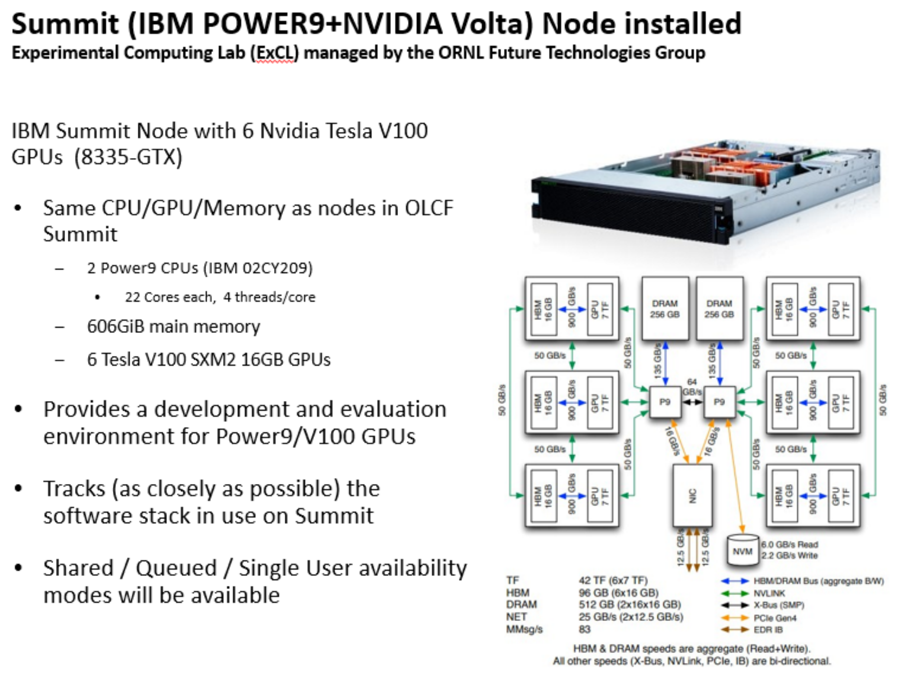
ORNL
AMD/Cray

ANL
Intel/Cray

LLNL
TBD

OAK RIDGE
National Laboratory

Jan 2018



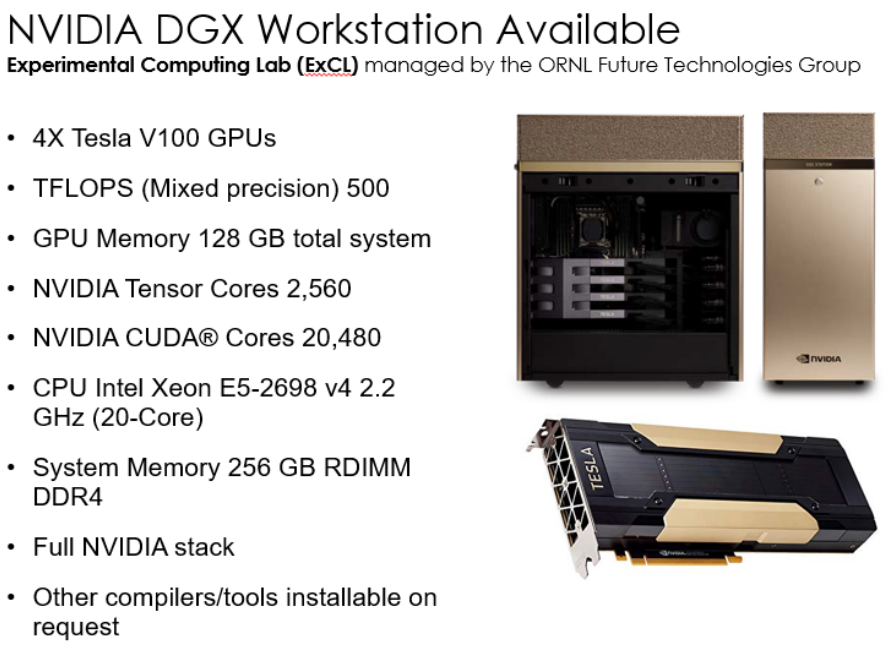
Summit (IBM POWER9+NVIDIA Volta) Node installed

Experimental Computing Lab (ExCL) managed by the ORNL Future Technologies Group

IBM Summit Node with 6 Nvidia Tesla V100 GPUs (8335-GTX)

- Same CPU/GPU/Memory as nodes in OLCF Summit
  - 2 Power9 CPUs (IBM 02CY209)
    - 22 Cores each, 4 threads/core
  - 606GiB main memory
  - 6 Tesla V100 SXM2 16GB GPUs
- Provides a development and evaluation environment for Power9/V100 GPUs
- Tracks (as closely as possible) the software stack in use on Summit
- Shared / Queued / Single User availability modes will be available

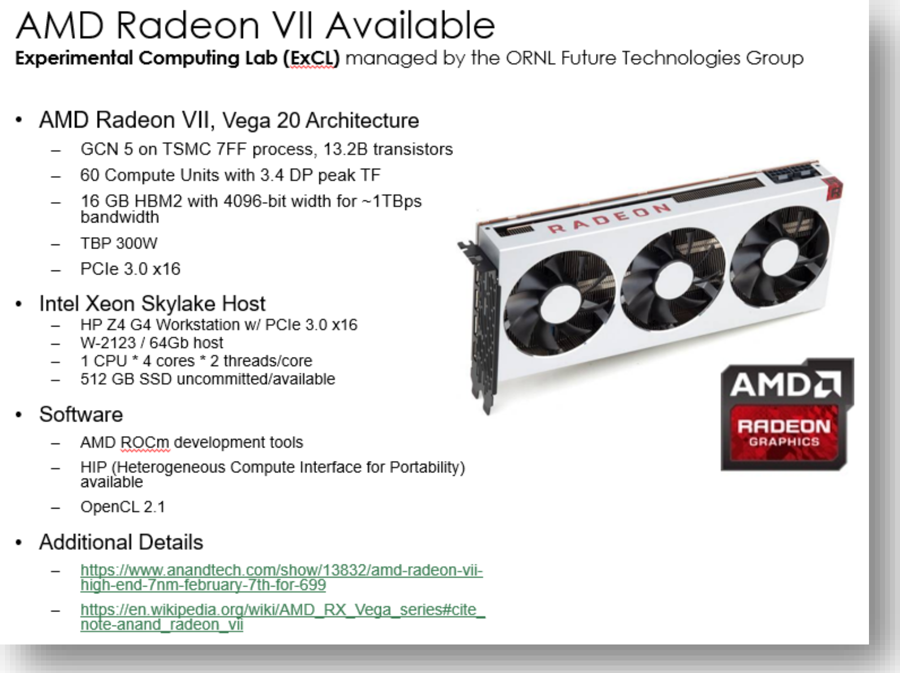
HBM & DRAM speeds are aggregate (Read+Write).
All other speeds (X-Bus, NVLink, PCIe, IB) are bi-directional.



NVIDIA DGX Workstation Available

Experimental Computing Lab (ExCL) managed by the ORNL Future Technologies Group

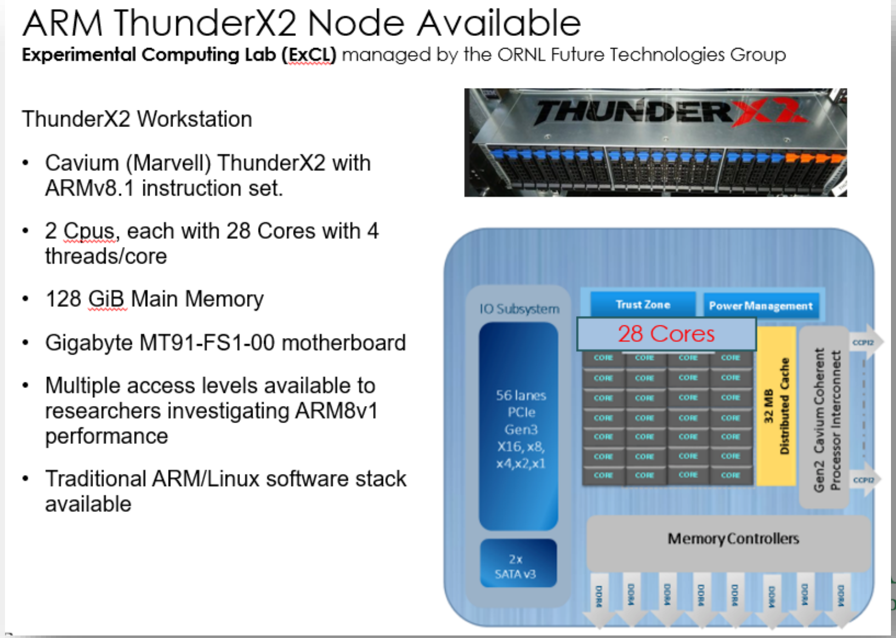
- 4X Tesla V100 GPUs
- TFLOPS (Mixed precision) 500
- GPU Memory 128 GB total system
- NVIDIA Tensor Cores 2,560
- NVIDIA CUDA® Cores 20,480
- CPU Intel Xeon E5-2698 v4 2.2 GHz (20-Core)
- System Memory 256 GB RDIMM DDR4
- Full NVIDIA stack
- Other compilers/tools installable on request



AMD Radeon VII Available

Experimental Computing Lab (ExCL) managed by the ORNL Future Technologies Group

- AMD Radeon VII, Vega 20 Architecture
  - GCN 5 on TSMC 7FF process, 13.2B transistors
  - 60 Compute Units with 3.4 DP peak TF
  - 16 GB HBM2 with 4096-bit width for ~1TBps bandwidth
  - TBP 300W
  - PCIe 3.0 x16
- Intel Xeon Skylake Host
  - HP Z4 G4 Workstation w/ PCIe 3.0 x16
  - W-2123 / 64Gb host
  - 1 CPU * 4 cores * 2 threads/core
  - 512 GB SSD uncommitted/available
- Software
  - AMD ROCm development tools
  - HIP (Heterogeneous Compute Interface for Portability) available
  - OpenCL 2.1
- Additional Details
  - https://www.anandtech.com/show/13832/amd-radeon-vii-high-end-7nm-february-7th-for-699
  - https://en.wikipedia.org/wiki/AMD_RX_Vega_series#cite_note-anand_radeon_vii



ARM ThunderX2 Node Available

Experimental Computing Lab (ExCL) managed by the ORNL Future Technologies Group

ThunderX2 Workstation

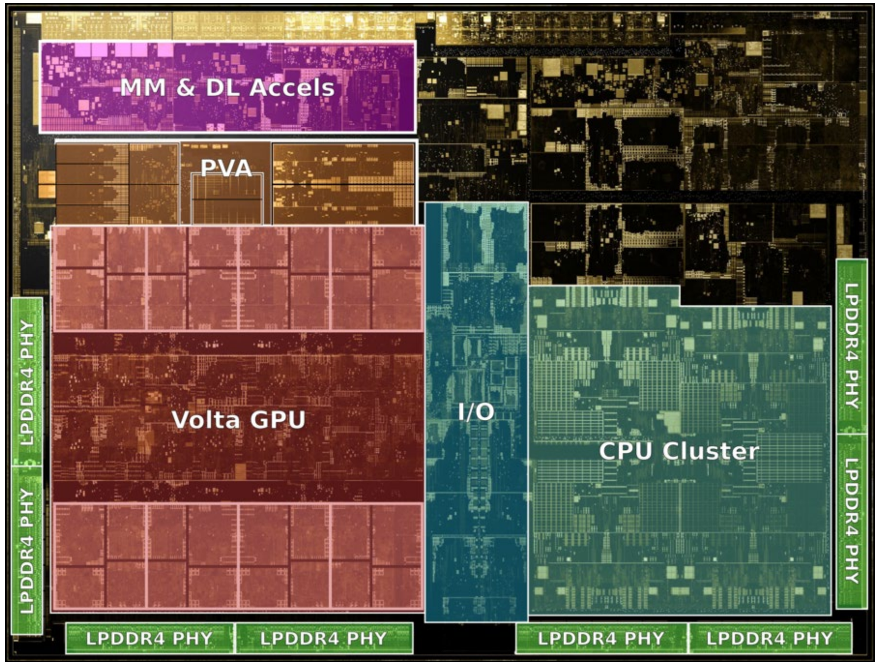
- Cavium (Marvell) ThunderX2 with ARMv8.1 instruction set.
- 2 Cpus, each with 28 Cores with 4 threads/core
- 128 GiB Main Memory
- Gigabyte MT91-FS1-00 motherboard
- Multiple access levels available to researchers investigating ARM8v1 performance
- Traditional ARM/Linux software stack available

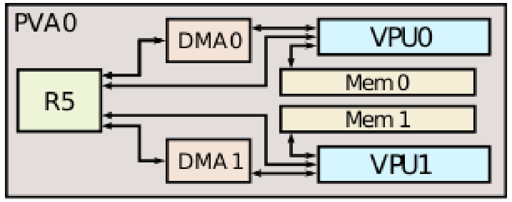
NVIDIA Jetson AGX Xavier SoC available

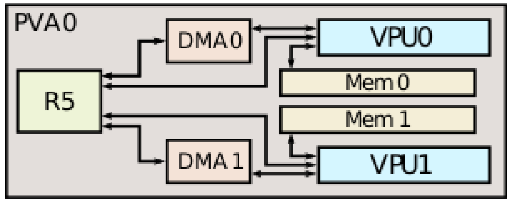
Experimental Computing Lab (ExCL) managed by the ORNL Future Technologies Group

NVIDIA Jetson AGX Xavier:

- High-performance system on a chip for autonomous machines
- Heterogeneous SoC contains:
  - Eight-core 64-bit ARMv8.2 CPU cluster (Carmel)
  - 1.4 CUDA TFLOPS (FP32) GPU with additional inference optimizations (Volta)
  - 11.4 DL TOPS (INT8) Deep learning accelerator (NVDLA)
  - 1.7 CV TOPS (INT8) 7-slot VLIW dual-processor Vision accelerator (PVA)
  - A set of multimedia accelerators (stereo, LDC, optical flow)
- Provides researchers access to advanced high-performance SOC environment



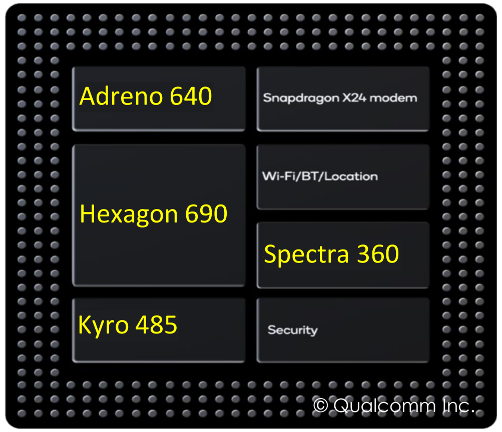




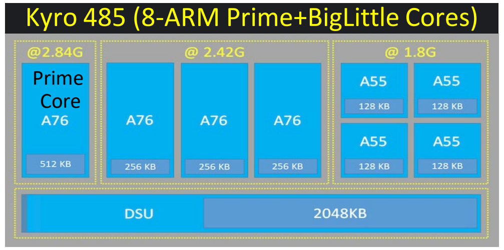
For more information or to apply for an account, visit https://excl.ornl.gov/

Qualcomm 855 SoC (SM8510P)

Experimental Computing Lab (ExCL) managed by the ORNL Future Technologies Group



Kyro 485 (8-ARM Prime+BigLittle Cores)

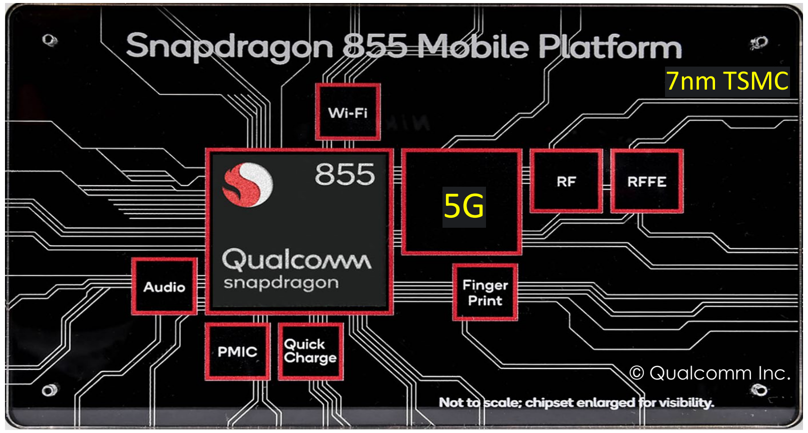


Hexagon 690 (DSP + AI)

- Quad threaded Scalar Core
- DSP + 4 Hexagon Vector Xccelerators
- New Tensor Xccelerator for AI
- Apps: AI, Voice Assistance, AV codecs

Adreno 640

- Vulkan, OpenCL, OpenGL ES 3.1
- Apps: HDR10+, HEVC, Dolby, etc
- Enables 8k-360º VR video playback
- 20% faster compared to Adreno 630



Connectivity (5G)

- Snapdragon X24 LTE (855 built-in) modem LTE Category 20
- Snapdragon X50 5G (external) modem (for 5G devices)
- Qualcomm Wi-Fi 6-ready mobile platform: (802.11ax-ready, 802.11ac Wave 2, 802.11ay, 802.11ad)
- Qualcomm 60 GHz Wi-Fi mobile platform: (802.11ay, 802.11ad)
- Bluetooth Version: 5.0
- Bluetooth Speed: 2 Mbps
- High accuracy location with dual-frequency GNSS.

Spectra 360 ISP

- New dedicated Image Signal Processor (ISP)
- Dual 14-bit CV-ISPs; 48MP @ 30fps single camera
- Hardware CV for object detection, tracking, streo depth process
- 6DoF XR Body tracking, H265, 4K60 HDR video capture, etc.

Qualcomm Development Board connected to (mcmurdo) HPZ820

- Connected Qualcomm board to HPZ820 through USB
- Development Environment: Android SDK/NDK
- Login to mcmurdo machine
  ```
  $ ssh –Y mcmurdo
  ```
- Setup Android platform tools and development environment
  ```
  $ source /home/nqx/setup_android.source
  ```
- Run Hello-world on ARM cores
  ```
  $ git clone https://code.ornl.gov/nqx/helloworld-android
  $ make  compile  push  run
  ```
- Run OpenCL example on GPU
  ```
  $ git clone https://code.ornl.gov/nqx/opencl-img-processing
  ```
  - Run Sobel edge detection
    ```
    $ make  compile  push  run  fetch
    ```
- Login to Qualcomm development board shell
  ```
  $ adb shell
  $ cd /data/local/tmp
  ```

OAK RIDGE National Laboratory

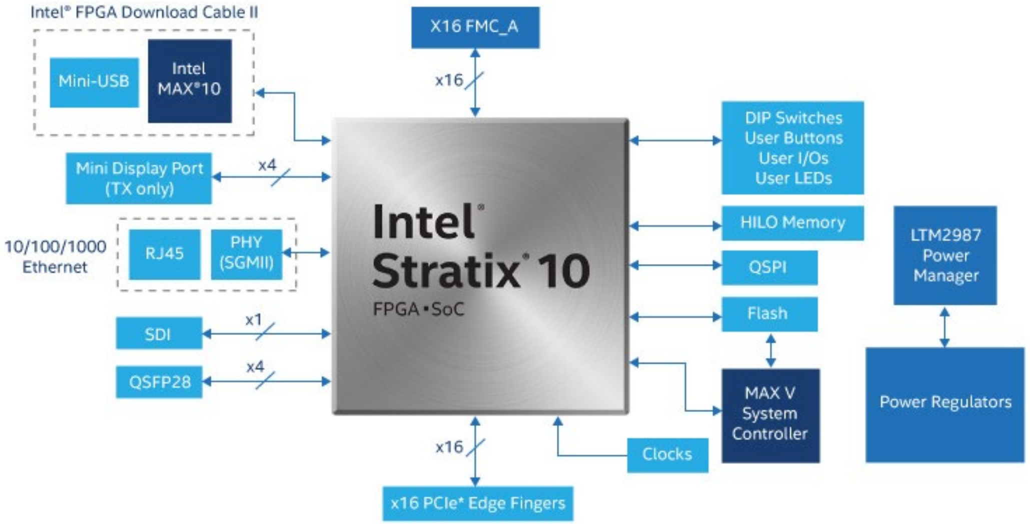
For more information or to apply for an account, visit https://excl.ornl.gov/

Created by Narasinga Rao Miniskar, Steve Moulton

Intel Stratix 10 FPGA available

Experimental Computing Lab (ExCL) managed by the ORNL Future Technologies Group

- Intel Stratix 10 FPGA and four banks of DDR4 external memory
  - Board configuration: Nallatech 520 Network Acceleration Card
- Up to 10 TFLOPS of peak single precision performance
- 25MBytes of L1 cache @ up to 94 TBytes/s peak bandwidth
- 2X Core performance gains over Arria® 10
- Quartus and OpenCL software (Intel SDK v18.1) for using FPGA
- Provide researcher access to advanced FPGA/SOC environment



For more information or to apply for an account, visit https://excl.ornl.gov/

Directive-based Solutions for FPGA Computing

[1] J. Lambert, S. Lee *et al., "Directive-Based, High-Level Programming and Optimizations for High-Performance Computing with FPGAs," in Proceedings of the 2018 International Conference on Supercomputing. Beijing, China: ACM, 2018, pp. 160-71, 10.1145/3205289.3205324.*

[2] S. Lee, J. Kim, and J.S. Vetter, "OpenACC to FPGA: A Framework for Directive-based High-Performance Reconfigurable Computing," Proc. IEEE International Parallel & Distributed Processing Symposium (IPDPS), 2016, 10.1109/IPDPS.2016.28.

Challenges in FPGA Computing

- Programmability and Portability Issues
  - Best performance for FPGAs requires writing Hardware Description Languages (HDLs) such as VHDL and Verilog; too complex and low-level
    - HDL requires substantial knowledge on hardware (digital circuits).
    - Programmers must think in terms of a state machine.
    - HDL programming is a kind of digital circuit design.
  - High-Level Synthesis (HLS) to provide better FPGA programmability
    - SRC platforms, Handel-C, Impulse C-to-FPGA compiler, Xilinx Vivado (AutoPilot), FCUDA, etc.
    - None of these use a portable, open standard.

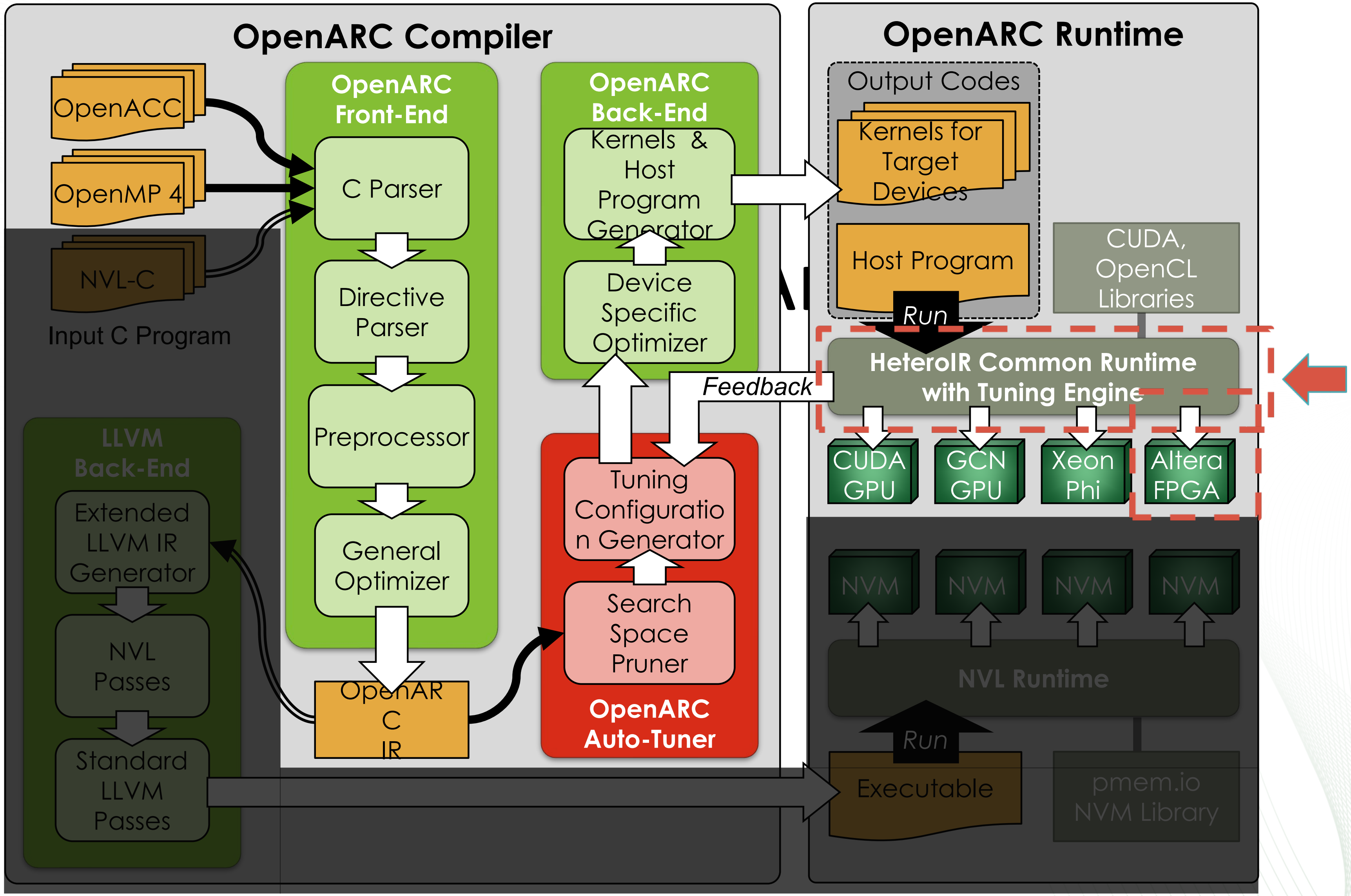
Standard, Portable Programming Models for Heterogeneous Computing

- OpenCL
  - Open standard portable across diverse heterogeneous platforms (e.g., CPUs, GPUs, DSPs, Xeon Phis, FPGAs, etc.)
  - Much higher than HDL, but still complex for typical programmers.
- Directive-based accelerator programming models
  - OpenACC, OpenMP4, etc.
  - Provide higher abstraction than OpenCL.
  - Most of existing OpenACC/OpenMP4 compilers target only specific architectures; none supports FPGAs.

FPGAs| Approach

- Design and implement an OpenACC-to-FPGA translation framework, which is the first work to use a standard and portable directive-based, high-level programming system for FPGAs.
- Propose FPGA-specific optimizations and novel pragma extensions to improve performance.
- Evaluate the functional and performance portability of the framework across diverse architectures (Altera FPGA, NVIDIA GPU, AMD GPU, and Intel Xeon Phi).

S. Lee, J. Kim, and J.S. Vetter, "OpenACC to FPGA: A Framework for Directive-based High-Performance Reconfigurable Computing," Proc. IEEE International Parallel & Distributed Processing Symposium (IPDPS), 2016, 10.1109/IPDPS.2016.28.



OpenARC Compiler

OpenACC

OpenMP 4

NVL-C

Input C Program

OpenARC Front-End

C Parser

Directive Parser

Preprocessor

General Optimizer

OpenARC IR

LLVM Back-End

Extended LLVM IR Generator

NVL Passes

Standard LLVM Passes

OpenARC Back-End

Kernels & Host Program Generator

Device Specific Optimizer

Feedback

Tuning Configuration Generator

Search Space Pruner

OpenARC Auto-Tuner

OpenARC Runtime

Output Codes

Kernels for Target Devices

Host Program

Run

CUDA, OpenCL Libraries

HeteroIR Common Runtime with Tuning Engine

CUDA GPU

GCN GPU

Xeon Phi

Altera FPGA

NVM

NVM

NVM

NVM

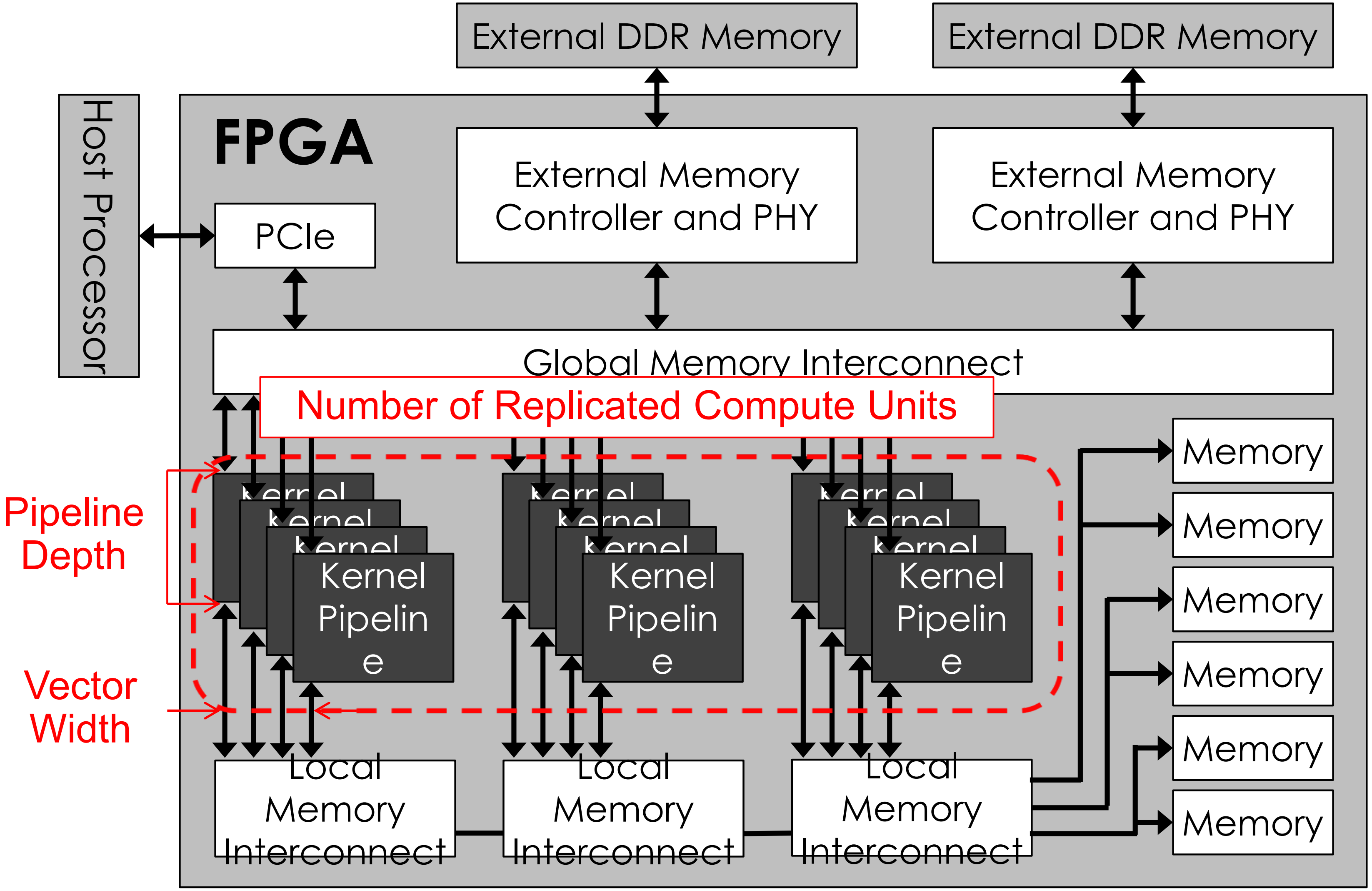
NVL Runtime

Run

Executable

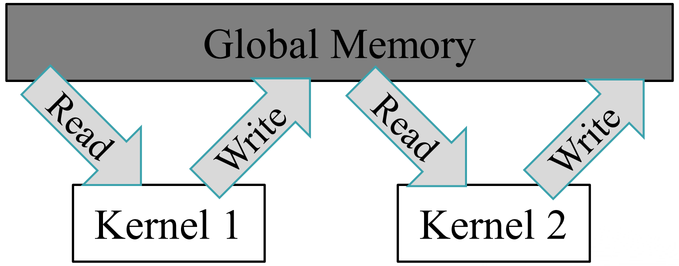
pmem.io NVM Library

FPGA OpenCL Architecture

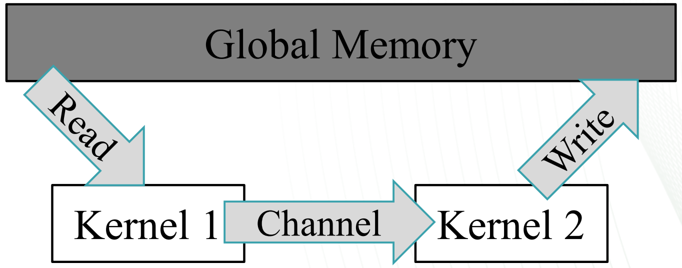


Kernel-Pipelining Transformation Optimization

- Kernel execution model in OpenACC
  - Device kernels can communicate with each other only through the device global memory.
  - Synchronizations between kernels are at the granularity of a kernel execution.
- Altera OpenCL channels
  - Allows passing data between kernels and synchronizing kernels with high efficiency and low latency



Kernel communications through global memory in OpenACC

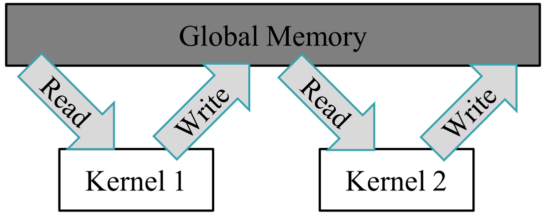


Kernel communications with Altera channels

Kernel-Pipelining Transformation Optimization (2)

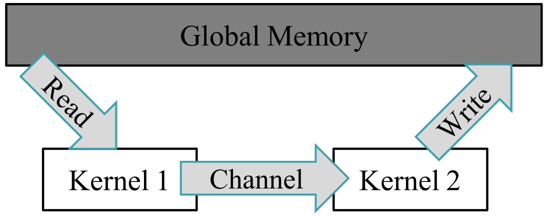
(a) Input OpenACC code

```
#pragma acc data copyin (a) create (b) copyout (c)
{
    #pragma acc kernels loop gang worker present (a, b)
    for(i=0; i<N; i++) { b[i] = a[i]*a[i]; }
    #pragma acc kernels loop gang worker present (b, c)
    for(i=0; i<N; i++) {c[i] = b[i]; }
}
```



(b) Altera OpenCL code with channels

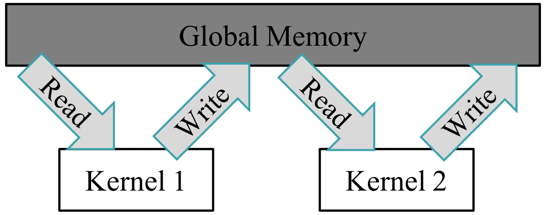
```
channel float pipe_b;
__kernel void kernel1(__global float* a) {
    int i = get_global_id(0);
    write_channel_altera(pipe_b, a[i]*a[i]);
}
__kernel void kernel2(__global float* c) {
    int i = get_global_id(0);
    c[i] = read_channel_altera(pipe_b);
}
```



Kernel-Pipelining Transformation Optimization (3)

(a) Input OpenACC code

```
#pragma acc data copyin (a) create (b) copyout (c)
{
    #pragma acc kernels loop gang worker present (a, b)
    for(i=0; i<N; i++) { b[i] = a[i]*a[i]; }
    #pragma acc kernels loop gang worker present (b, c)
    for(i=0; i<N; i++) {c[i] = b[i]; }
}
```

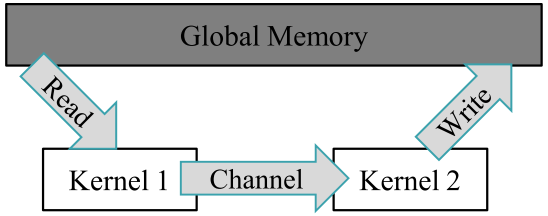


Kernel-pipelining transformation

Valid under specific conditions

(c) Modified OpenACC code for kernel-pipelining

```
#pragma acc data copyin (a) pipe (b) copyout (c)
{
    #pragma acc kernels loop gang worker pipeout (b) present (a)
    For(i=0; i<N; i++) { b[i] = a[i]*a[i]; }
    #pragma acc kernels loop gang worker pipein (b) present (c)
    For(i=0; i<N; i++) {c[i] = b[i];}
}
```

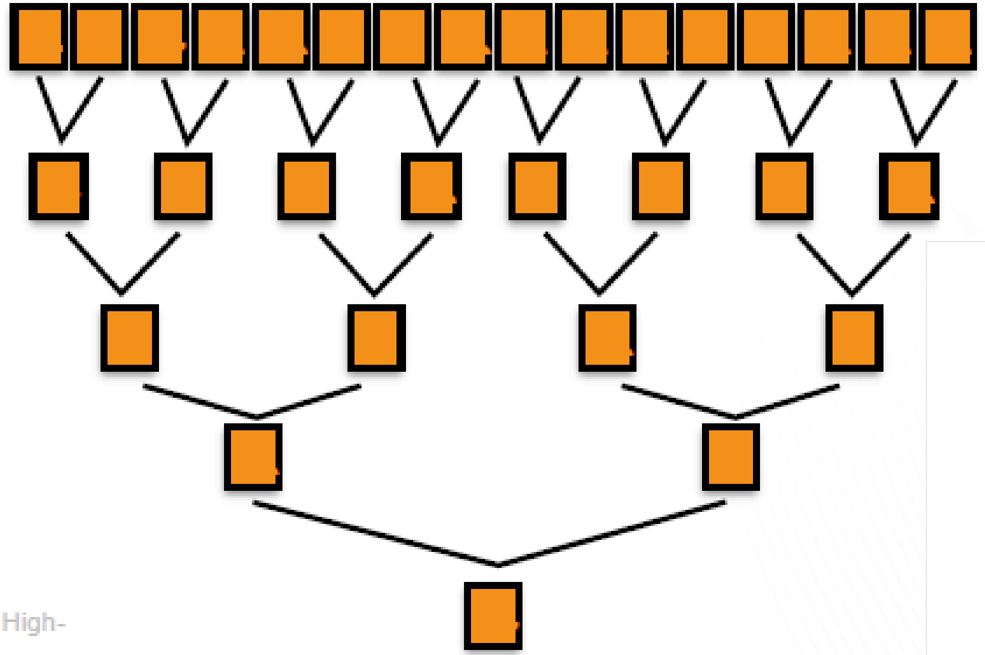


FPGA-specific Optimizations

- Single work-item
- Collapse
- Reduction
- Sliding window
- (Branch-variant code motion)
- (Custom unrolling)

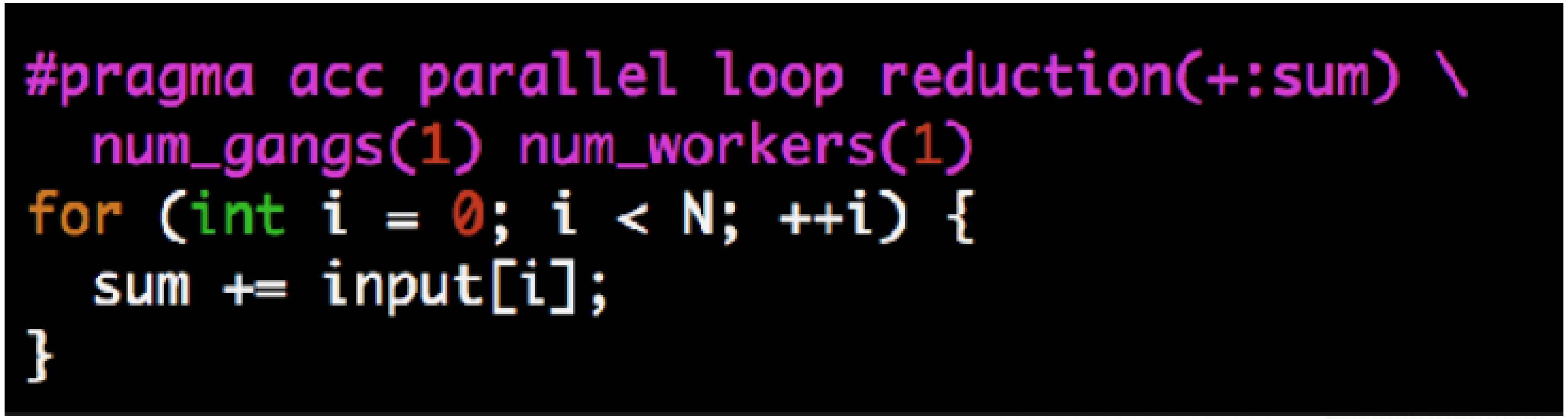
Reduction Optimization

- Scalar reduction
- Combine an array of values into a single scalar value
- Associative Operations: Sum, Multiply, Max, Min, AND, OR
- OpenMP and OpenACC support a *reduction* clause to direct the compiler to implement parallel reductions.



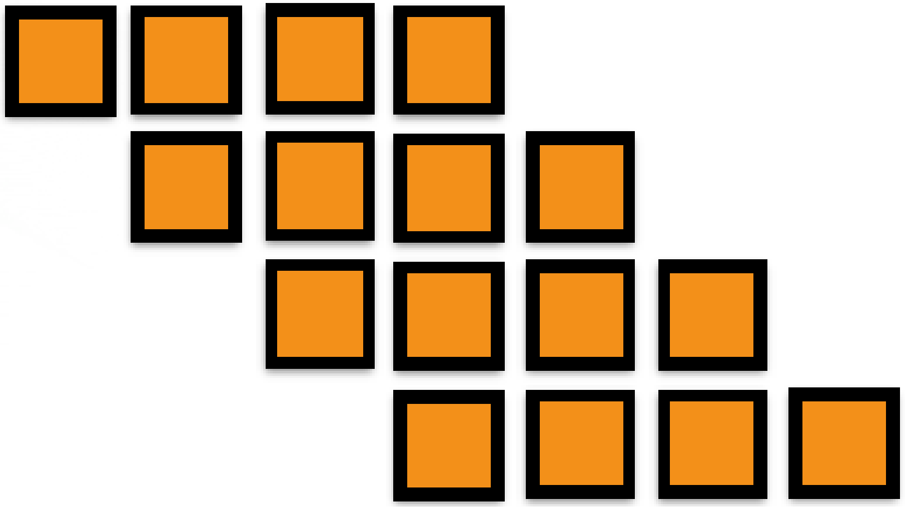
Pipeline Parallel Reduction

- We can do multi-threaded parallel tree-based reduction on FPGA
- Can we also do pipeline parallel reduction?
- Scalar reductions have a loop-carried dependency on the reduction variable

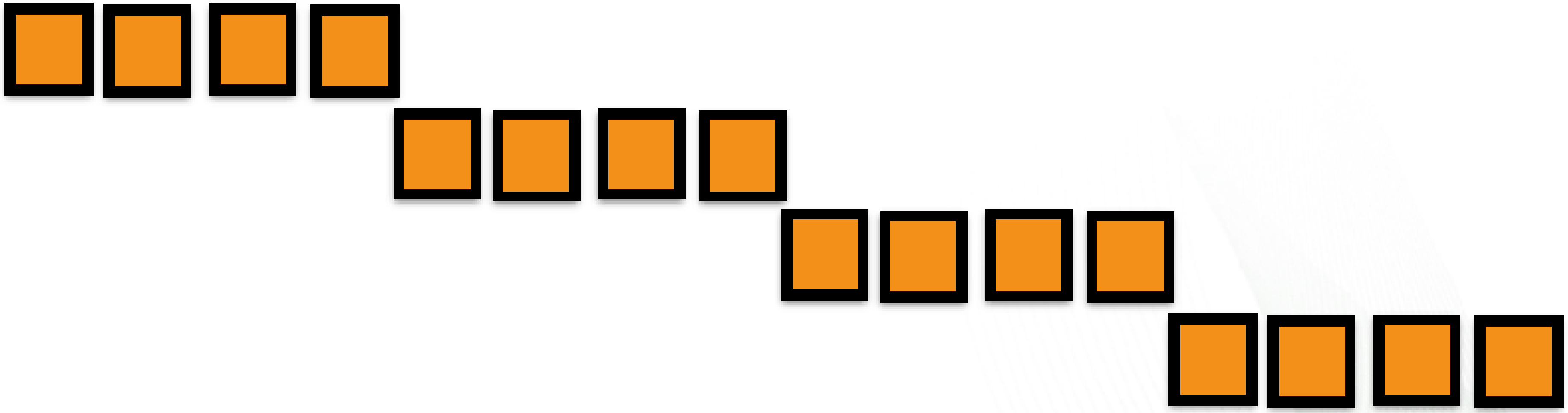


```
#pragma acc parallel loop reduction(+:sum) \
  num_gangs(1) num_workers(1)
for (int i = 0; i < N; ++i) {
  sum += input[i];
}
```

What we want

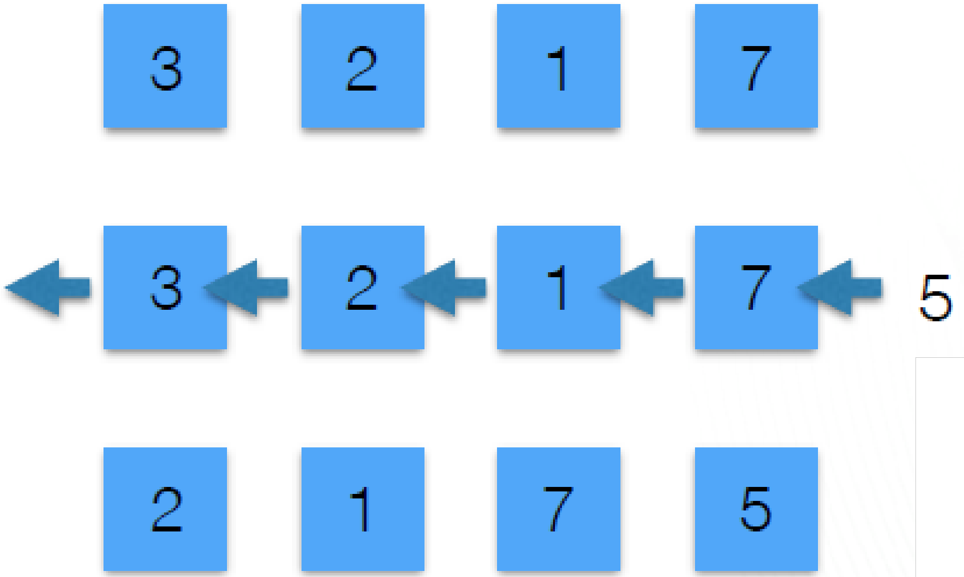


What we have



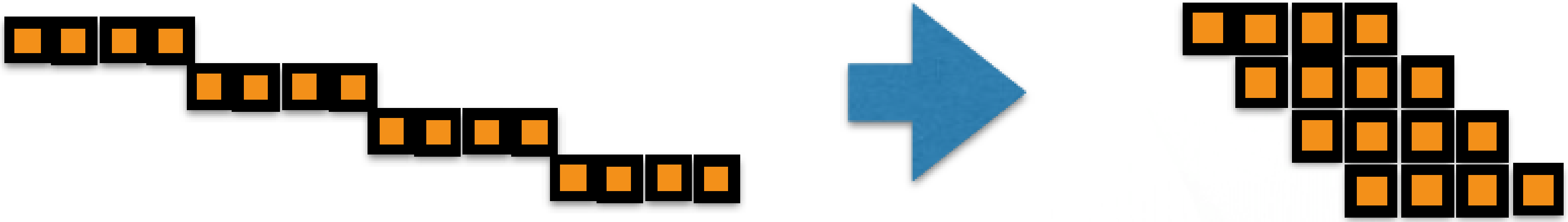
Shift Registers

- “A cascade of flip flops sharing the same clock in which the output of each flip-flop is connected to the 'data' input of the next flip-flop in the chain”
- An array where each value can be shifted left or right efficiently (in one cycle).



Shift-register Reduction

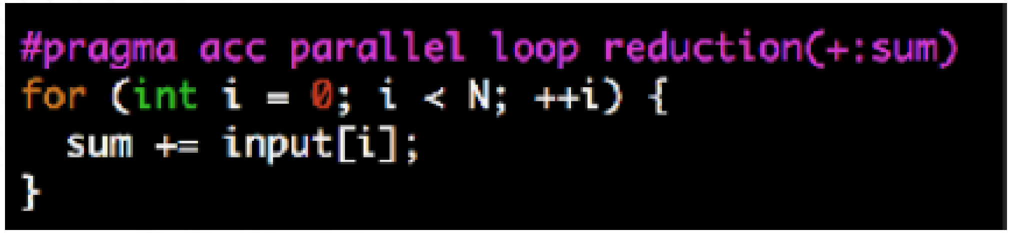
- Accumulate partial sums in a shift registers, and perform a final reduction over the partial sums
- By using shift registers we can increase the dependance distance, improving pipeline efficiency



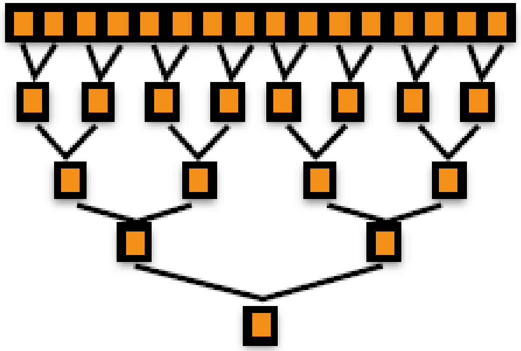
- The size of the shift register needed is directly proportional to the cost of the reduction operation.

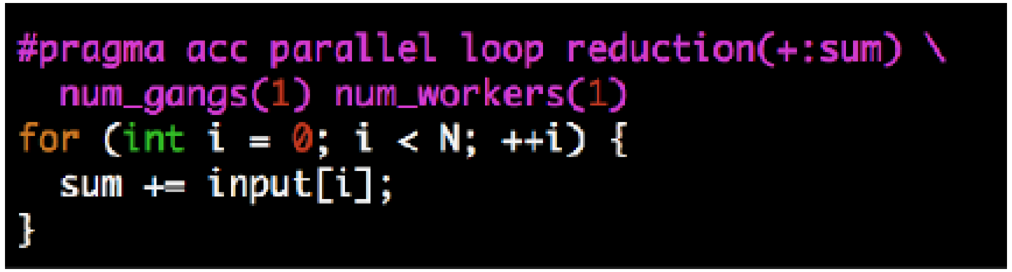
OpenACC input

OpenCL output (Generated by OpenARC)

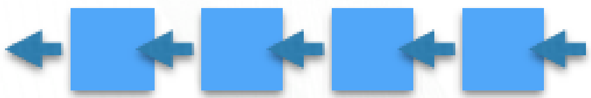


```
#pragma acc parallel loop reduction(+:sum)
for (int i = 0; i < N; ++i) {
  sum += input[i];
}
```

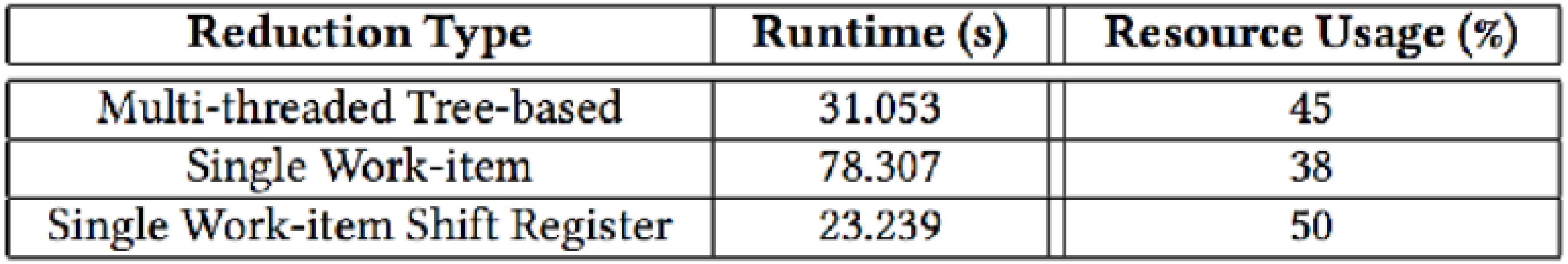




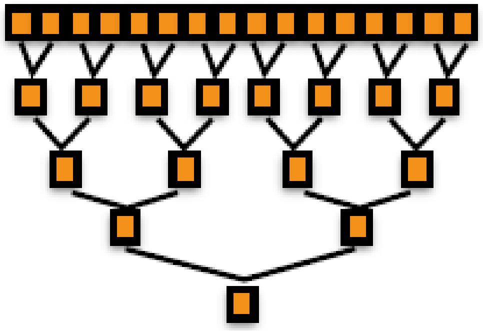
```
#pragma acc parallel loop reduction(+:sum) \
  num_gangs(1) num_workers(1)
for (int i = 0; i < N; ++i) {
  sum += input[i];
}
```

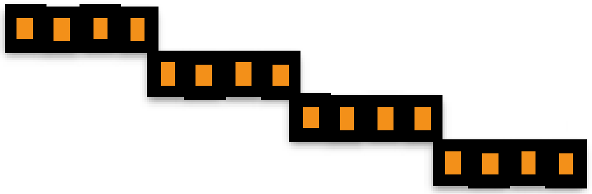


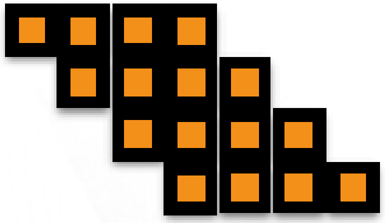
Reduction Optimization (SRAD Benchmark)



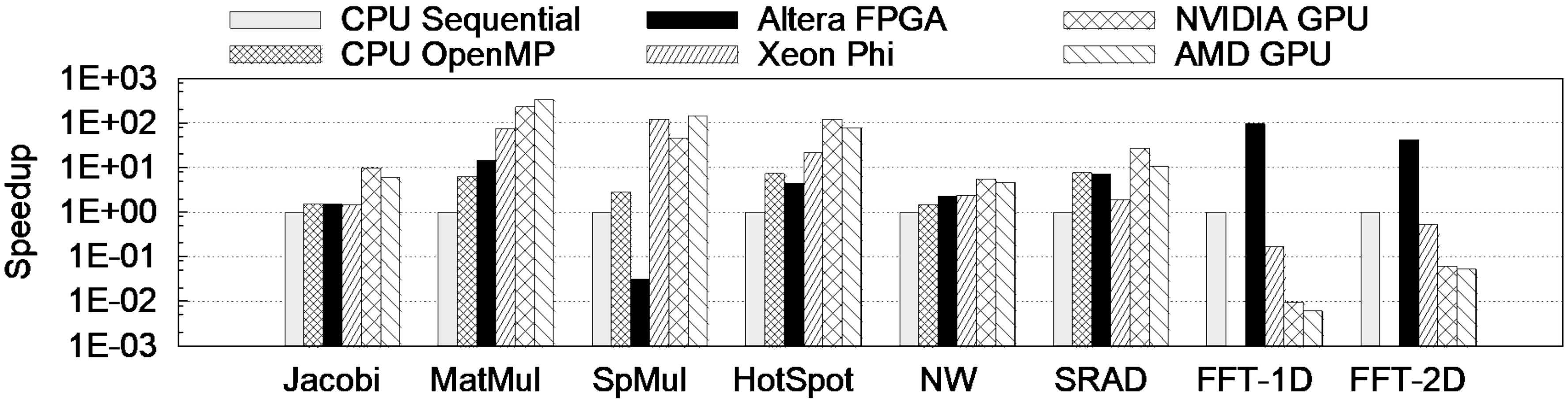
Reduction Type	Runtime (s)	Resource Usage (%)
Multi-threaded Tree-based	31.053	45
Single Work-item	78.307	38
Single Work-item Shift Register	23.239	50







Overall Performance



FPGAs prefer applications with deep execution pipelines (e.g., FFT-1D and FFT-2D), performing much higher than other accelerators.

For traditional HPC applications with abundant parallel floating-point operations, it seems to be difficult for FPGAs to beat the performance of other accelerators, even though FPGAs can be much more power-efficient.

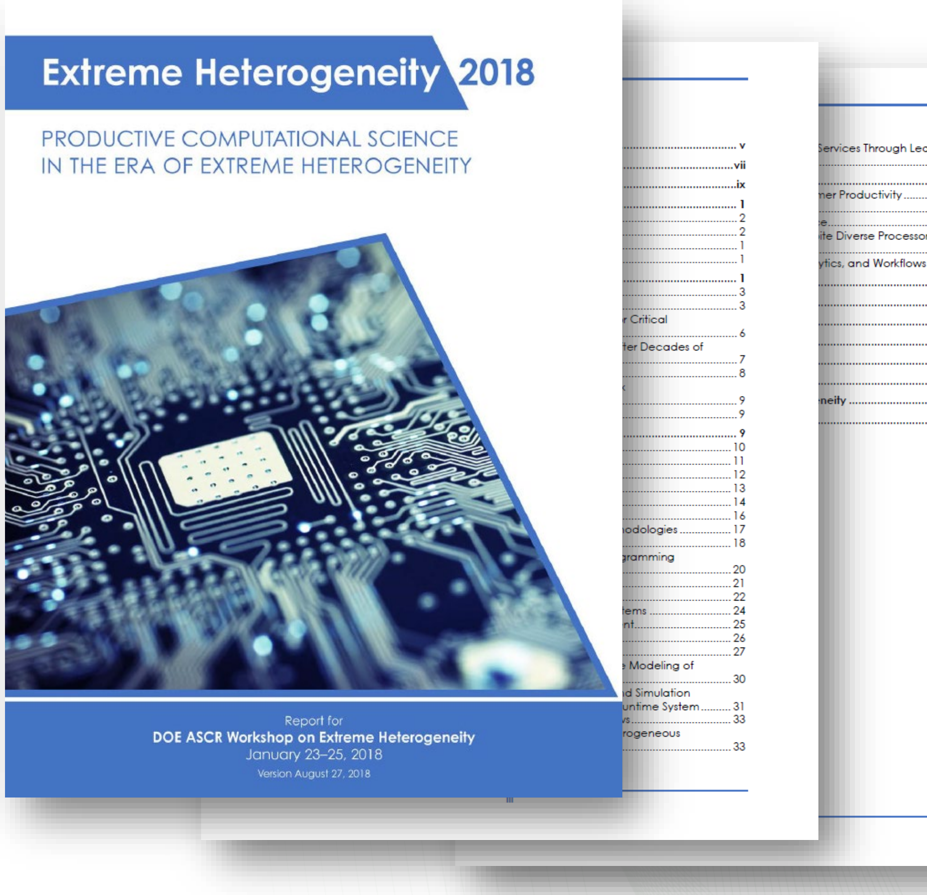
- Tested FPGA does not contain dedicated, embedded floating-point cores, while others have fully-optimized floating-point computation units.

Current and upcoming high-end FPGAs are equipped with hardened floating-point operators, whose performance will be comparable to other accelerators, while remaining power-efficient.

Final Report on Workshop on Extreme Heterogeneity

1. Maintaining and improving programmer productivity
   - Flexible, expressive, programming models and languages
   - Intelligent, domain-aware compilers and tools
   - Composition of disparate software components

- Managing resources intelligently
  - Automated methods using introspection and machine learning
  - Optimize for performance, energy efficiency, and availability
- Modeling & predicting performance
  - Evaluate impact of potential system designs and application mappings
  - Model-automated optimization of applications
- Enabling reproducible science despite non-determinism & asynchrony
  - Methods for validation on non-deterministic architectures
  - Detection and mitigation of pervasive faults and errors
- Facilitating Data Management, Analytics, and Workflows
  - Mapping of science workflows to heterogeneous hardware and software services
  - Adapting workflows and services to meet facility-level objectives through learning approaches



Extreme Heterogeneity 2018

PRODUCTIVE COMPUTATIONAL SCIENCE
IN THE ERA OF EXTREME HETEROGENEITY

Report for
DOE ASCR Workshop on Extreme Heterogeneity
January 23–25, 2018
Version August 27, 2018

https://orau.gov/exheterogeneity2018/

https://doi.org/10.2172/1473756

Recap

- Recent trends in computing paint an ambiguous future (for HPC and broader community)
  - Contemporary systems provide evidence that power constraints are driving architectures to change rapidly (e.g., Dennard, Moore)
  - Multiple architectural dimensions are being (dramatically) redesigned: Processors, node design, memory systems, I/O
- Major transition point for computing
  - New devices
  - New architectures
  - New programming systems
- Complexity and uncertainty are ubiquitous
- Programming systems must provide performance portability (in addition to functional portability)!!
- In near term, rate of change will accelerate and grow more diverse

- Visit us
  - We host interns and other visitors year round
    - Faculty, grad, undergrad, high school, industry
- Jobs in FTG
  - Postdoctoral Research Associate in Computer Science
  - Software Engineer
  - Computer Scientist
  - Visit https://jobs.ornl.gov
- Contact me vetter@ornl.gov

OAK RIDGE
National Laboratory