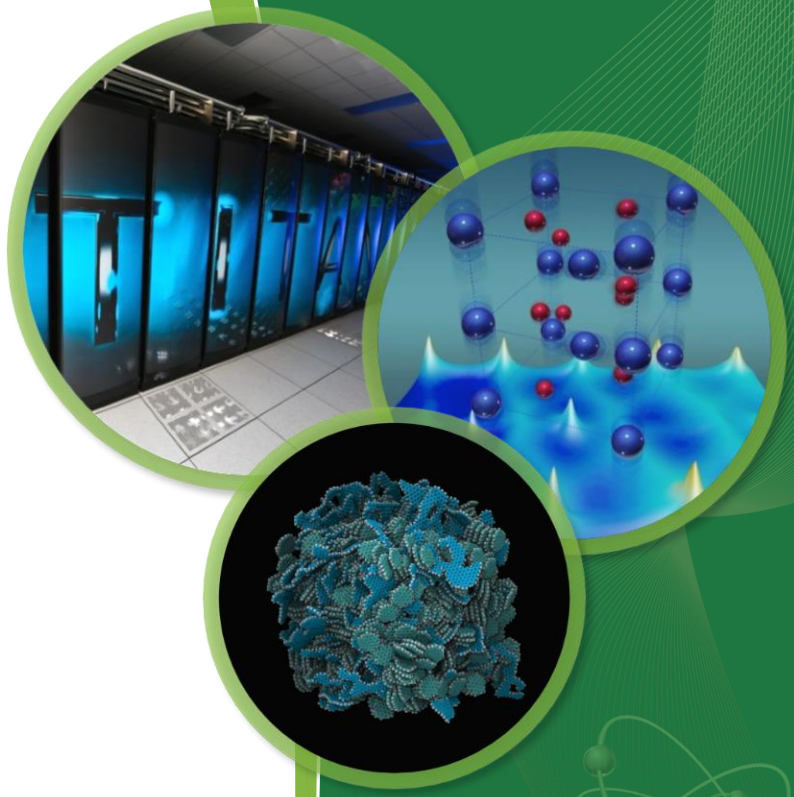
Exploring Emerging Memory Technologies in Extreme Scale High Performance Computing

Jeffrey S. Vetter



Presented to
ISC Third International Workshop on Communication Architectures for HPC, Big Data, Deep Learning and Clouds at Extreme Scale

Frankfurt

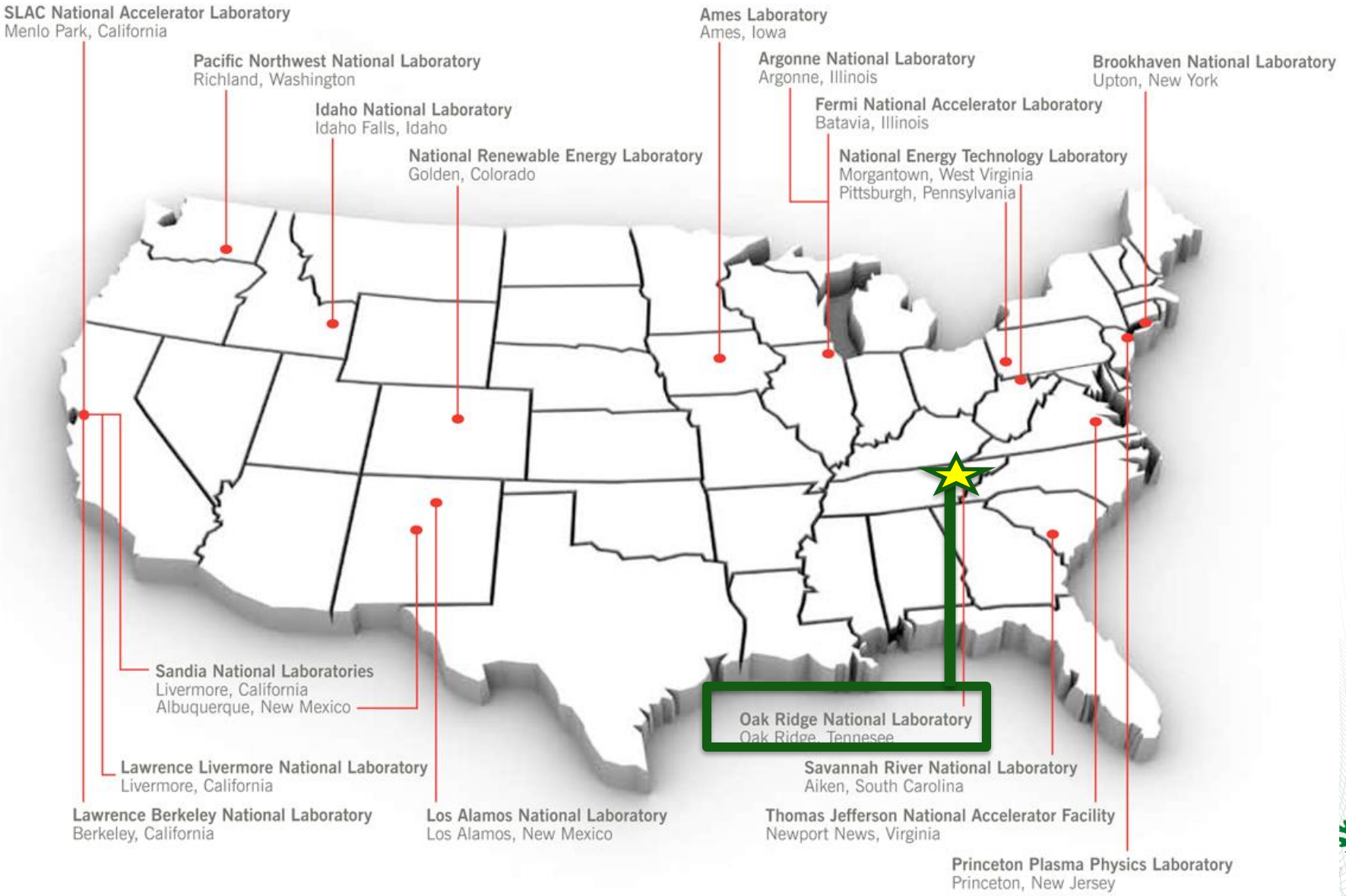
22 Jun 2017

Future Technologies Group ornl

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

OAK RIDGE National Laboratory

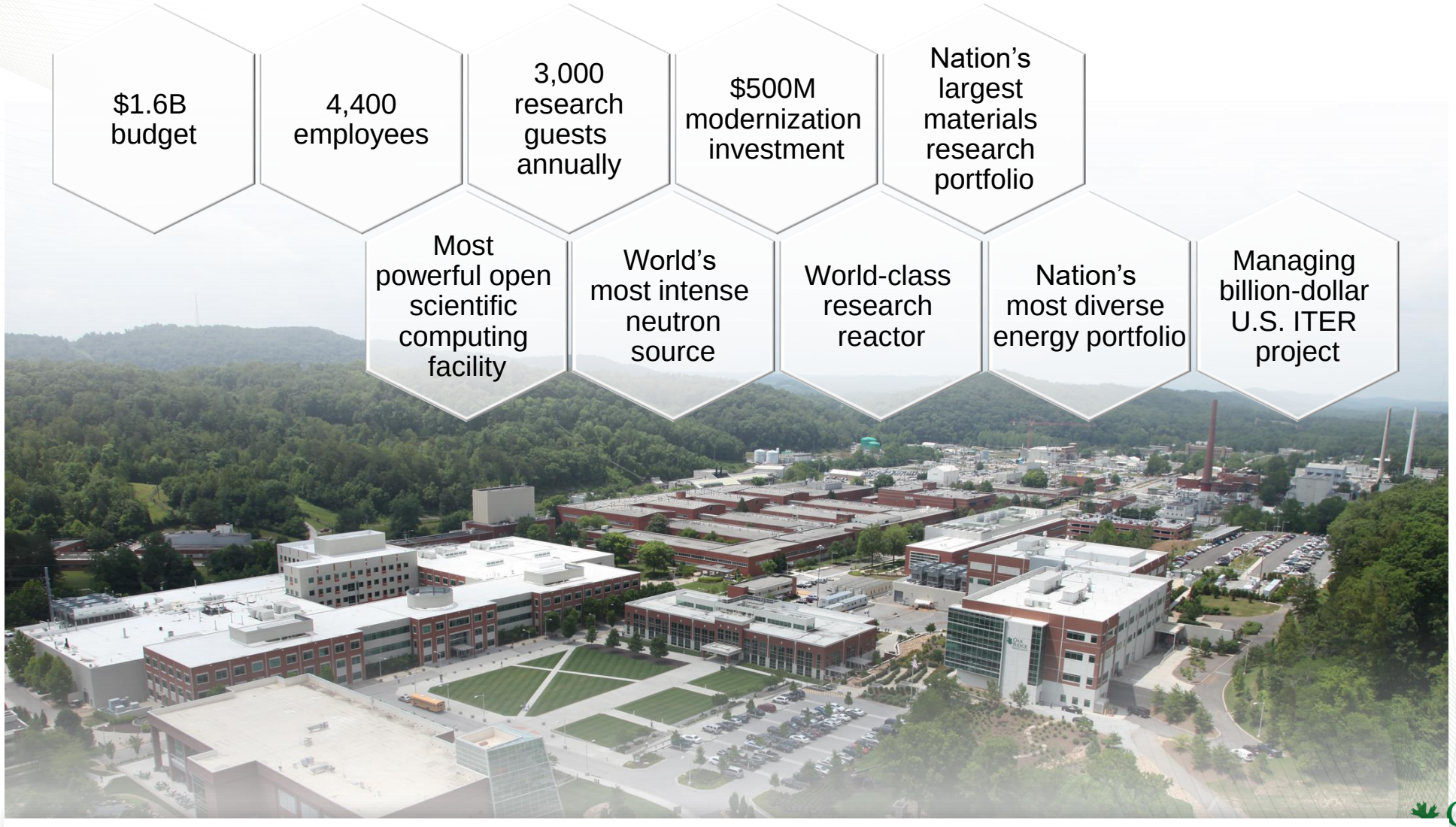
Oak Ridge National Laboratory is the DOE Office of Science's Largest Lab



7

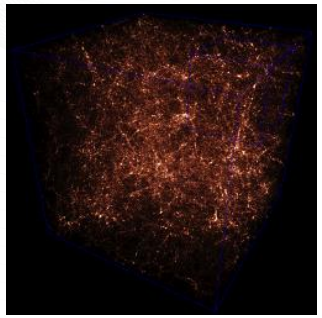
OAK RIDGE National Laboratory

ORNL is Characterized by Diverse Scientific Portfolio



What Does 1000x Increase Provide? Mission Impact

Cosmology

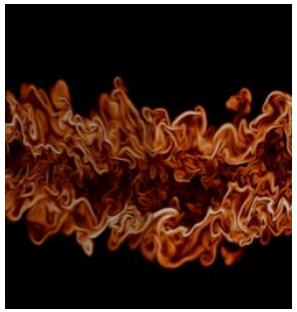


Salman Habib
Argonne National Laboratory

Habib and collaborators used its HACC Code on Titan's CPU–GPU system to conduct today's largest cosmological structure simulation at resolutions needed for modern-day galactic surveys.

K. Heitmann, 2014. *arXiv.org,* 1411.3396

Combustion

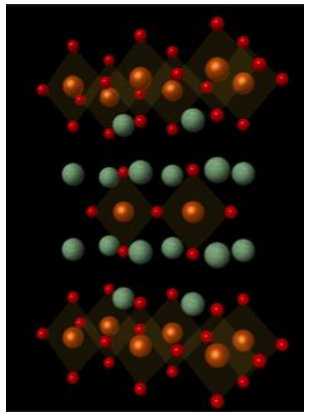


Jacqueline Chen
Sandia National Laboratory

Chen and collaborators for the first time performed direct numerical simulation of a jet flame burning dimethyl ether (DME) at new turbulence scales over space and time.

A. Bhagatwala, et al. 2014. *Proc. Combust. Inst.* **35**.

Superconducting Materials



Paul Kent
ORNL

Paul Kent and collaborators performed the first ab initio simulation of a cuprate. They were also the first team to validate quantum Monte Carlo simulations for high-temperature superconductor simulations.

K. Foyevtsova, et al. 2014. *Phys. Rev. X* **4**

Molecular Science

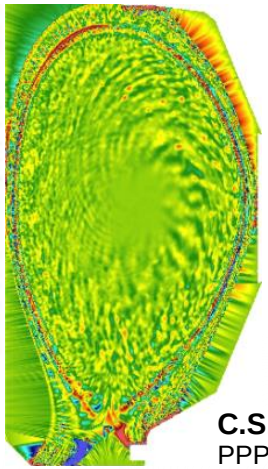
Michael Klein
Temple University

Researchers at Procter & Gamble (P&G) and Temple University delivered a comprehensive picture in full atomistic detail of the molecular properties that drive skin barrier disruption.

M. Paloncyova, et al. 2014. *Langmuir* **30**

C. M. MacDermaid, et al. 2014. *J. Chem. Phys.* **141**

Fusion



C.S. Chang
PPPL

Chang and collaborators used the XGC1 code on Titan to obtain fundamental understanding of the divertor heat-load width physics and its dependence on the plasma current in present-day tokamak devices.

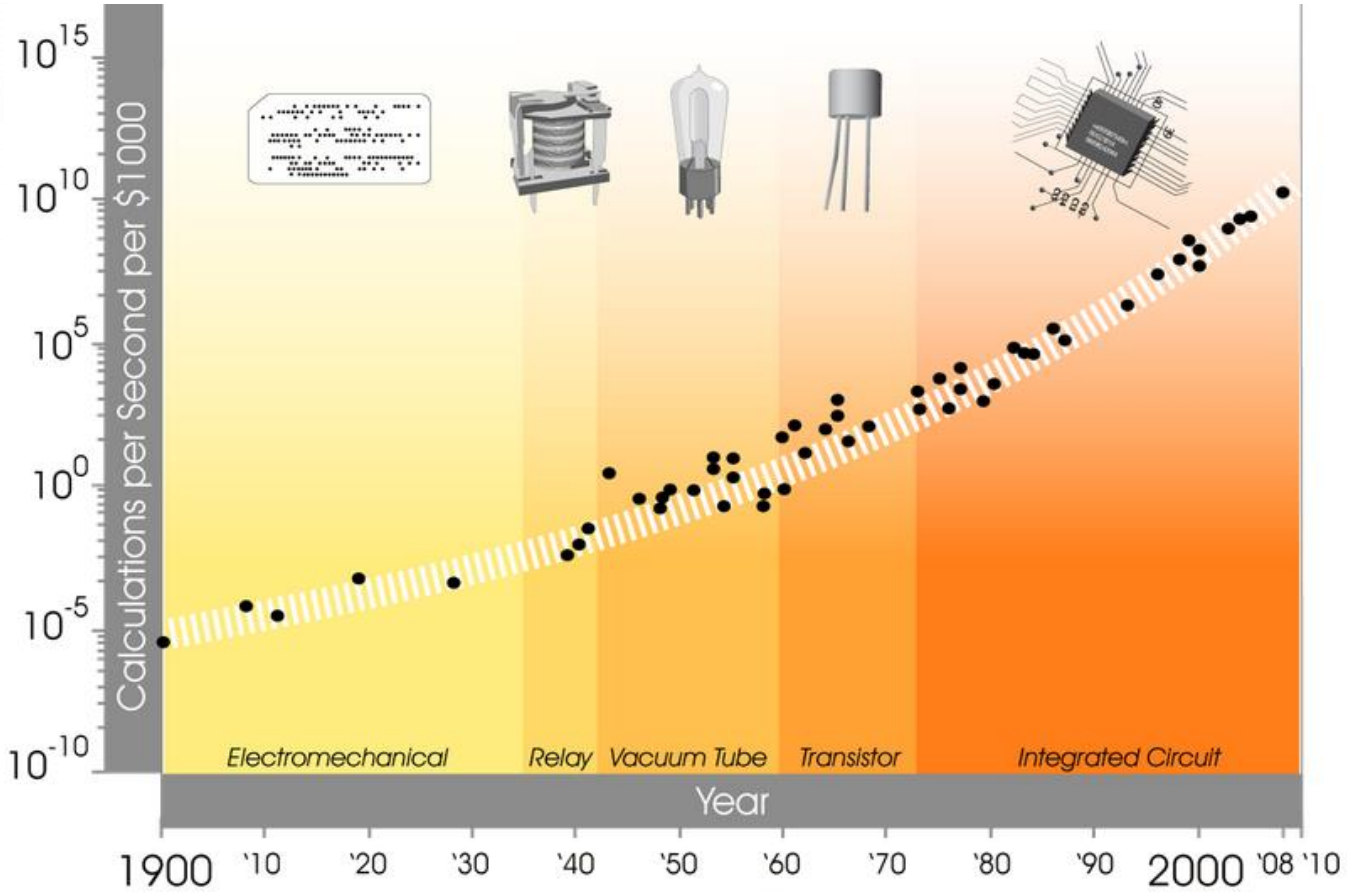
C. S. Chang, et al. 2014. *Proceedings of the 25th Fusion Energy Conference, IAEA*, October 13–18, 2014.

Highlights

- Recent trends in extreme-scale HPC paint an ambiguous future
  - 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
- Memory systems are changing now!
  - New devices
  - New integration
  - New configurations
  - Vast (local) capacities
- Programming systems must provide performance portability (in addition to functional portability)!!
  - We need new programming systems to effectively use these architectures
  - NVL-C
  - Papyrus(KV)
- Changes in memory systems will alter communication and storage requirements dramatically

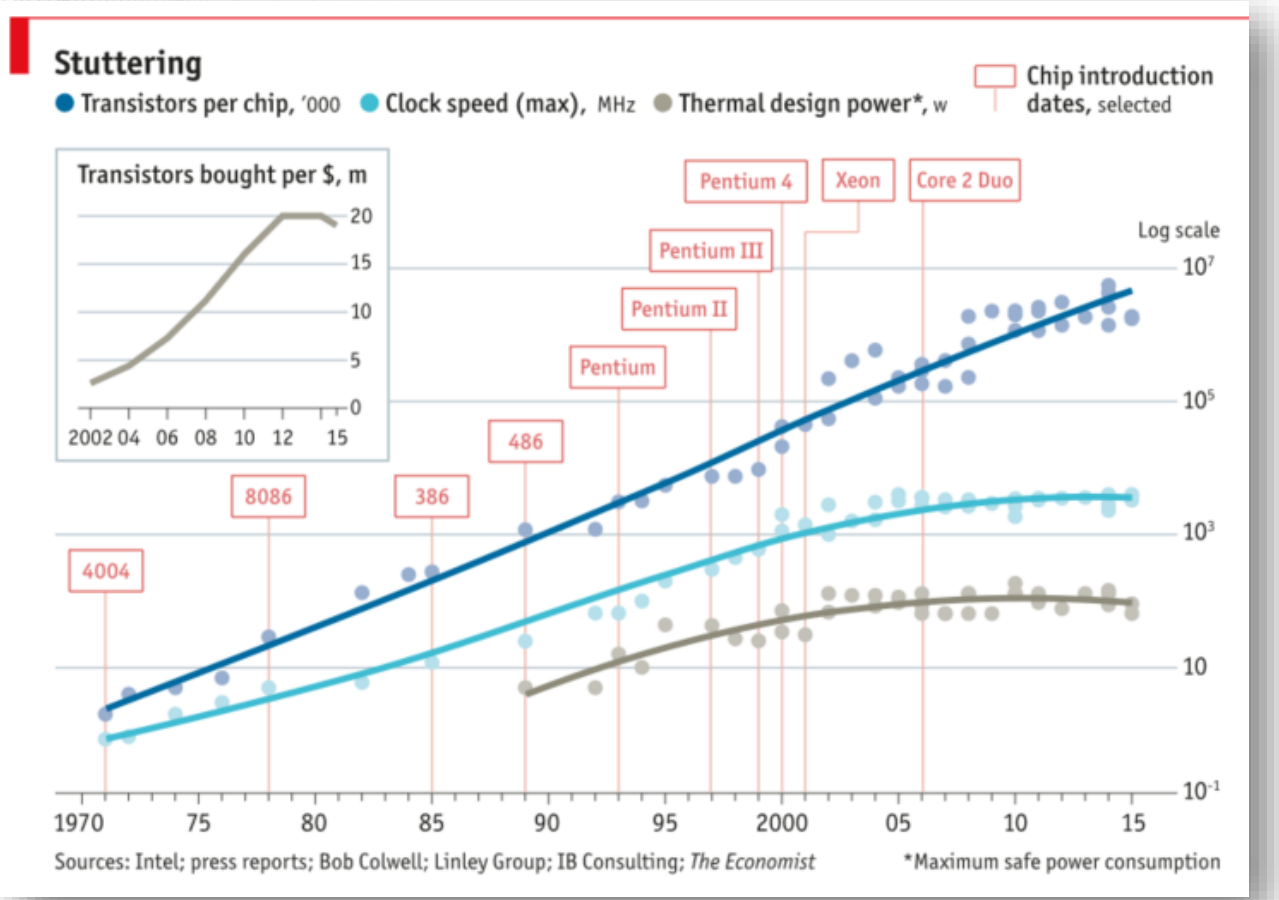
Major Trends in Computing

Sixth Wave of Computing



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

Contemporary devices are approaching fundamental limits



Economist, Mar 2016

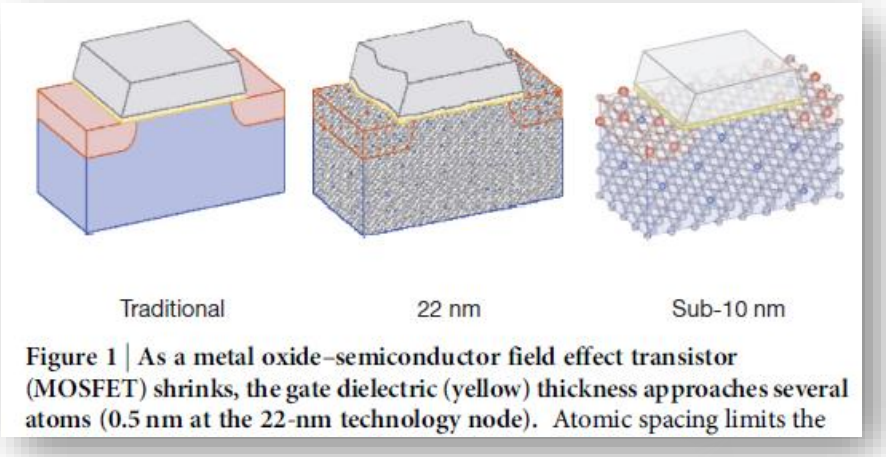


Figure 1 | As a metal oxide–semiconductor field effect transistor (MOSFET) shrinks, the gate dielectric (yellow) thickness approaches several atoms (0.5 nm at the 22-nm technology node). Atomic spacing limits the

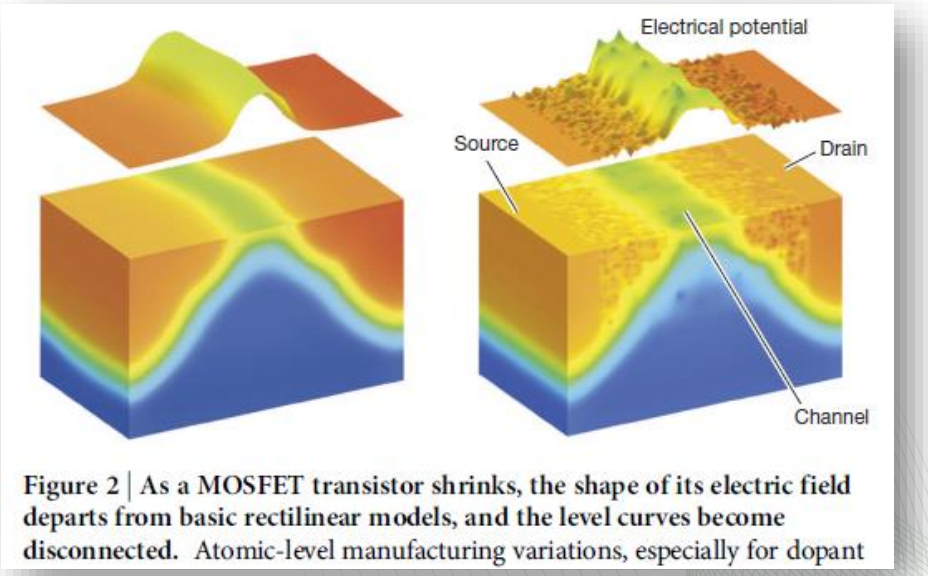


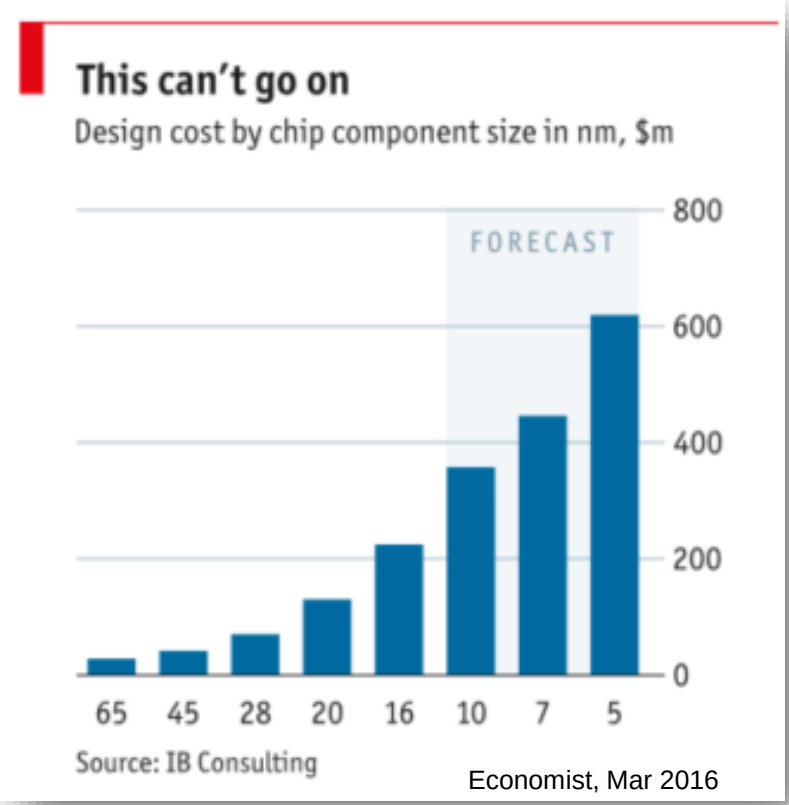
Figure 2 | As a MOSFET transistor shrinks, the shape of its electric field departs from basic rectilinear models, and the level curves become disconnected. Atomic-level manufacturing variations, especially for dopant

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.*

Semiconductors are taking longer and cost more to design and produce



Economist, Mar 2016

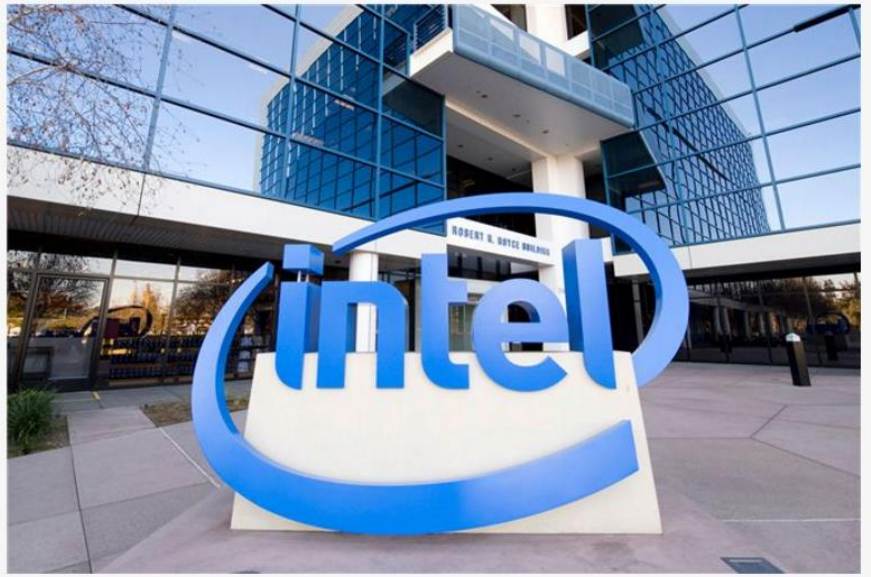
Intel's 'Tick-Tock' Seemingly Dead, Becomes 'Process-Architecture-Optimization'

98 Comments

+ Add A Comment

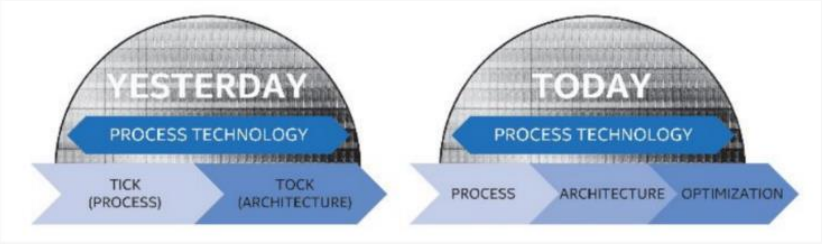
by Ian Cutress on March 22, 2016 6:45 PM EST

Posted in CPUs | Intel | 14nm | 10nm | EUV | Lithography | Tick-Tock | Process-Architecture-Optimization



As reported at The Motley Fool, Intel's latest 10-K / annual report filing would seem to suggest that the 'Tick-Tock' strategy of introducing a new lithographic process note in one product cycle (a 'tick') and then an upgraded microarchitecture the next product cycle (a 'tock') is going to fall by the wayside for the next two lithographic nodes at a minimum, to be replaced with a three element cycle known as 'Process-Architecture-Optimization'.

Intel's Tick-Tock strategy has been the bedrock of their microprocessor dominance of the last decade. Throughout the tenure, every other year Intel would upgrade their fabrication plants to be able to produce processors with a smaller feature set, improving die area, power consumption, and slight optimizations of the microarchitecture, and in the years between the upgrades would launch a new set of processors based on a wholly new (sometimes paradigm shifting) microarchitecture for large performance upgrades. However, due to the difficulty of implementing a 'tick', the ever decreasing process node size and complexity therein, as reported previously with 14nm and the introduction of Kaby Lake, Intel's latest filing would suggest that 10nm will follow a similar pattern as 14nm by introducing a third stage to the cadence.



http://www.anandtech.com/show/10183/intels-tick-tock-seemingly-dead-becomes-process-architecture-optimization

RIDGE Laboratory

Semiconductor business is highly process-oriented, optimized, and growing extremely capital intensive

designlines INDUSTRIAL CONTROL

News & Analysis

Semi industry fab costs limit industry growth

Nicolas Mokhoff
10/3/2012 03:00 PM EDT
Post a comment

NO RATINGS
LOGIN TO RATE

Like 4 Tweet Share G+1 0

MANHASSET, N.Y. -- The fundamental economics of the semiconductor industry may start changing sooner rather than later, according to market research firm Gartner Inc.

The costs of staying at the leading edge in semiconductor manufacturing are rising. Semiconductor manufacturers need to plan on equipment costs increasing at about 15 percent for each new node, according to Gartner (Stamford, Conn.).

It's possible that 450-mm manufacturing will achieve the goal of 30 percent cost reduction. But that equates to only three or four years of increasing equipment costs, and consequently, delays the inevitable, Gartner said. It is also possible that new technologies will emerge that will slow the rate of cost increases, according to the firm.

According to Gartner, the costs of manufacturing equipment needed for leading-edge semiconductor manufacturing are increasing at a rate between 7 percent and 10 percent per year, depending on the basic process.

By 2020, current cost trends will lead to an average cost of between $15 billion and $20 billion for a leading-edge fab, according to the report. By 2016, the minimum capital expenditure budget needed to justify the building of a new fab will range from $8 billion to $10 billion for logic, $3.5 billion to $4.5 billion for DRAM and $6 billion to $7 billion for NAND flash, according to the report.

The Gartner report predicts that at current spending rates, only eight companies could afford to build fabs in the next few years.

EE Times

By 2020, current cost trends will lead to an average cost of between $15 billion and $20 billion for a leading-edge fab, according to the report. By 2016, the minimum capital expenditure budget needed to justify the building of a new fab will range from $8 billion to $10 billion for logic, $3.5 billion to $4.5 billion for DRAM and $6 billion to $7 billion for NAND flash, according to the report.

Context: Intel Reports Full-Year Revenue of $55.4 Billion, Net Income of $11.4 Billion (Intel SEC Filing for FY2015)

Major 2013 IC Foundries (Pure-Play and IDM)

2013 Rank	2012 Rank	Company	Foundry Type	Location	2011 Sales ($M)	2012 Sales ($M)	2012/2011 Change (%)	2013 Sales ($M)	2013/2012 Change (%)
1	1	TSMC	Pure-Play	Taiwan	14,299	16,951	19%	19,850	17%
2	2	GlobalFoundries	Pure-Play	U.S.	3,195	4,013	26%	4,261	6%
3	3	UMC	Pure-Play	Taiwan	3,760	3,730	-1%	3,959	6%
4	4	Samsung	IDM	South Korea	2,192	3,439	57%	3,950	15%
5	5	SMIC*	Pure-Play	China	1,320	1,542	17%	1,973	28%
6	8	Powerchip**	Pure-Play	Taiwan	374	625	67%	1,175	88%
7	9	Vanguard	Pure-Play	Taiwan	520	582	12%	713	23%
8	6	Huahong Grace***	Pure-Play	China	619	677	9%	710	5%
9	10	Dongbu	Pure-Play	South Korea	500	540	8%	570	6%
10	7	TowerJazz	Pure-Play	Israel	611	639	5%	509	-20%
11	11	IBM	IDM	U.S.	420	432	3%	485	12%
12	12	MagnaChip	IDM	South Korea	350	400	14%	411	3%
13	13	WIN	Pure-Play	Taiwan	304	381	25%	354	-7%
—	—	Top 13 Total	—	—	28,464	33,951	19%	38,920	15%
—	—	Top 13 Share	—	—	89%	90%	—	91%	—
—	—	Other Foundry	—	—	3,446	3,669	6%	3,920	7%
—	—	Total Foundry	—	—	31,910	37,620	18%	42,840	14%

*Does not include Wuhan Xinxin (now XMC) for 2012 or 2013.
**Powerchip transitioned from an IDM foundry to a pure-play foundry in 2013.
***Hua Hong NEC and Grace merged in 2012 (excludes Shanghai Huali).

Source: IC Insights, company reports

http://www.icinsights.com/

OAK RIDGE National Laboratory

Business climate reflects this uncertainty, cost, complexity, consolidation

designlines WIRELESS & NETWORKING

Blog

IC Merger Mania Hits Fever Pitch

Dylan McGrath, Contributing Editor
12/2/2015 10:13 AM EST
NO RATINGS
LOGIN TO RATE
1 comments post a c
Like 10 Twee

With the announcem
PMC-Sierra, the total
acquisitions announ

The wave of consolid

designlines AUTOMOTIVE

News & Analysis

Foundries' Sales Show Hard Times Continuing

er Clarke
2016 09:33 PM EDT
mments
Like 6 Tweet

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 provide an alternative.

DECEMBER 19TH, 2016 - BY: MARK LAPEDUS

As several chipmakers ramp up their 10nm finFET processes, with 7nm just around the corner, R&D has begun for 5nm and beyond. In fact, some are already moving full speed ahead in the arena.

TSMC recently announced plans to build a new fab in Taiwan at a cost of $15.7 billion. The proposed fab is targeted to manufacture TSMC's 5nm and 3nm processes, which are due out in 2020 and 2022, respectively. Other chipmakers, including GlobalFoundries, Intel and Samsung, also are looking at technologies for 5nm and beyond.

Both 5nm and 3nm present a multitude of unknowns and challenges. For one thing, the specs of these technologies are murky, if not confusing. And not all of the technologies are alike.

Regardless, based on the roadmaps from various chipmakers, Moore's Law continues to slow as process complexities and costs escalate at each node. The roadmaps, of course, could change. But for now, Intel plans to ramp up 10nm in the second half of 2017, with 7nm slated for production in early to mid-2020, according to industry sources. Intel's 5nm production is targeted for early 2023, sources said, meaning its traditional 2-year process cadence is extending to roughly 2.5 to 3 years.

Intel to acquire Altera for $54 a share

Monday, 1 Jun 2015 | 8:33
REUTERS

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 23, 2014

ACQUISITION TO BOOST SANDISK'S ENTERPRISE GROWTH

MILPITAS, Calif., July 23, 2014 - SanDisk Corporation (NASDAQ: SNDK), a global leader in flash storage solutions, today announce
hardware

"I am deli
the Fusio
and chief
the indus

to-market talent of
ehrotra, president
se flash solutions in

Western Digital Now A Storage Powerhouse With SanDisk Acquisition

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

Britain's largest tech firm,
deal including UK jobs gu

Britain's most successful techno
in a £24bn deal by Japan's SoftBank

Britain's biggest technolo
ARM Holdings accepted a

Qualcomm to Acquire NXP Semiconductors for $38.5 Billion

By CHAD BRAY and QUENTIN HARDY OCT. 27, 2016

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

© FT montage

33

MARCH 8, 2017 by: Arash Massoudi and George Parker in London

Japan's SoftBank is set to sell a roughly $8bn stake in Arm, the UK chip designer
purchased only six months ago, placing 25 per cent of Britain's largest technolog
company into a new, Saudi-backed $100bn investment fund.

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

In April-June 2016, Toshiba had a 20.4% share in global NAND flash memory market.

The Dollar Business Bureau

Japanese conglomerate Toshiba Corporation is in discussion to sell a minority stake in its well-performing flash memory division to the US maker Western Digital Corporation in a bid to raise funds, according to a source.

Earlier, in a statement, Toshiba said that it had been looking for several options for its flash memory business, even a spin-off, but nothing firm had been decided.

The laptops-to-engineering multinational, still coping with an accounting scandal of $1.3 billion, surprised its investors last month by declaring cost overruns at the US nuclear business it acquired in 2015, which could at present mean a charge against profit reaching $4 billion.

Broadcom acquires Brocade in $5.9 billion deal

Posted 1 hour ago by Ron Miller (@ron_miller)

Next Story

t Brocade Communications Systems today for $5.9 billion, giving the
ering in networking storage business.

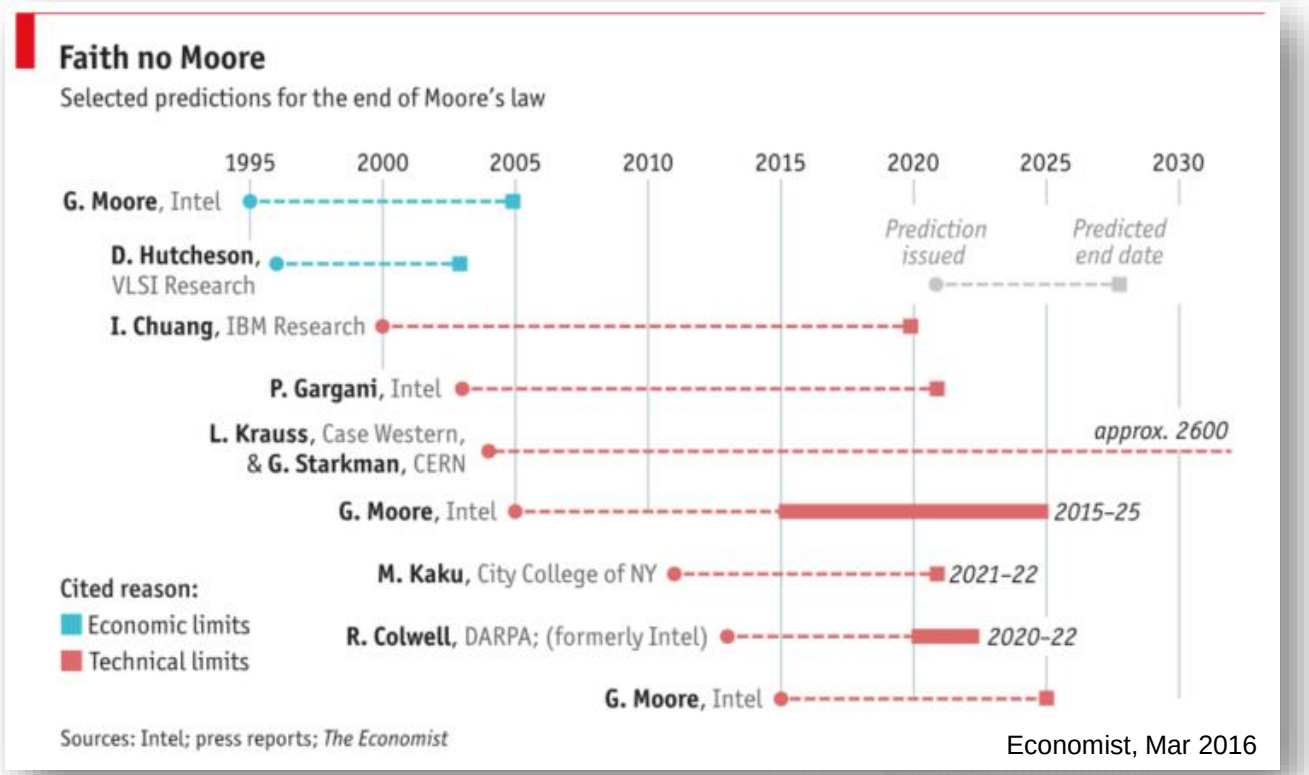
or Brocade shareholders, who gain a 47 percent premium over the $8.69
ay night (October 28th). Under the terms of the deal, Broadcom will pay

lion transaction with Broadcom also assuming $400 million of net debt for

OAK RIDGE National Laboratory

End of Moore's Law

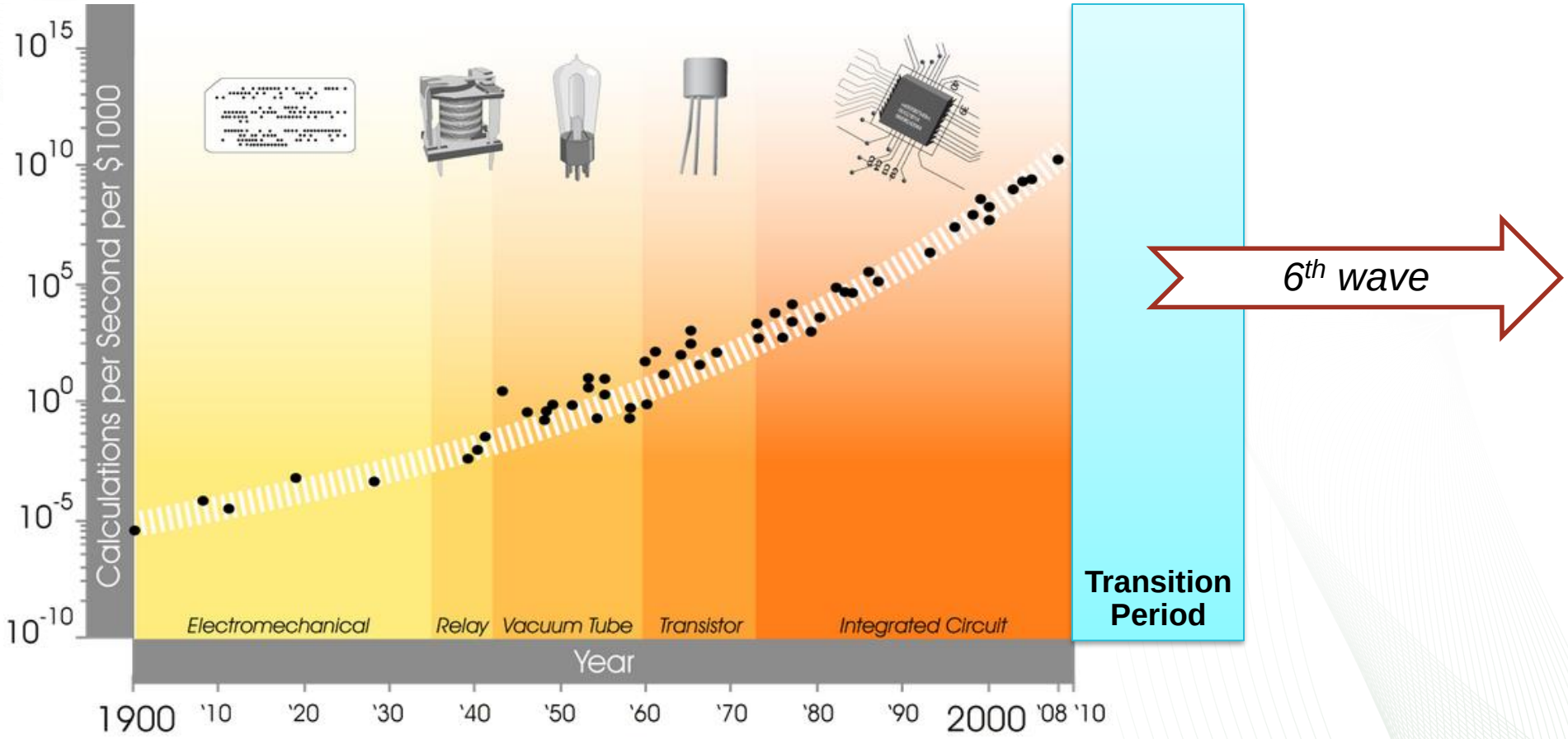
- Device level physics will prevent much smaller feature size of current transistor technologies
- Business trends indicate asymptotic limits of both manufacturing capability and economics
- What now?



Economist, Mar 2016

"The number of people predicting the death of Moore's Law doubles every two years." – Peter Lee, Microsoft

Sixth Wave of Computing



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

Our Transition Period Predictions

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 efficiently for our workloads
- Integrate components to boost performance and eliminate inefficiencies

Emerging Technologies

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

Our Transition Period Predictions

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

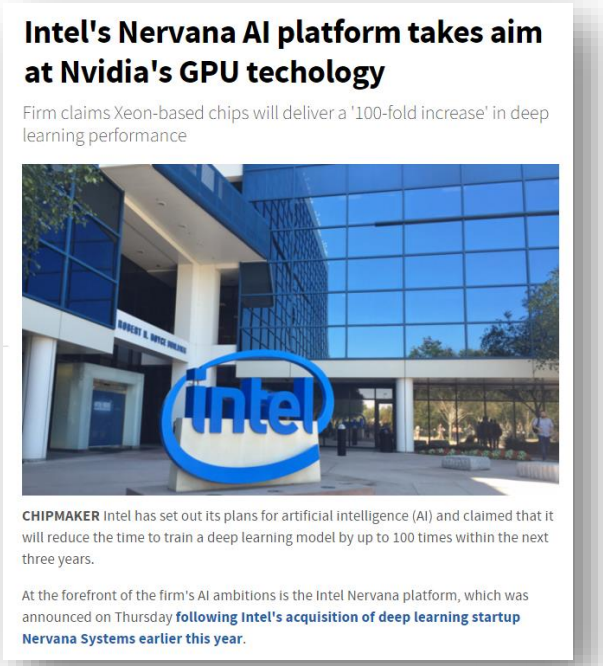
- Use CMOS more efficiently for our workloads
- Integrate components to boost performance and eliminate inefficiencies

Emerging Technologies

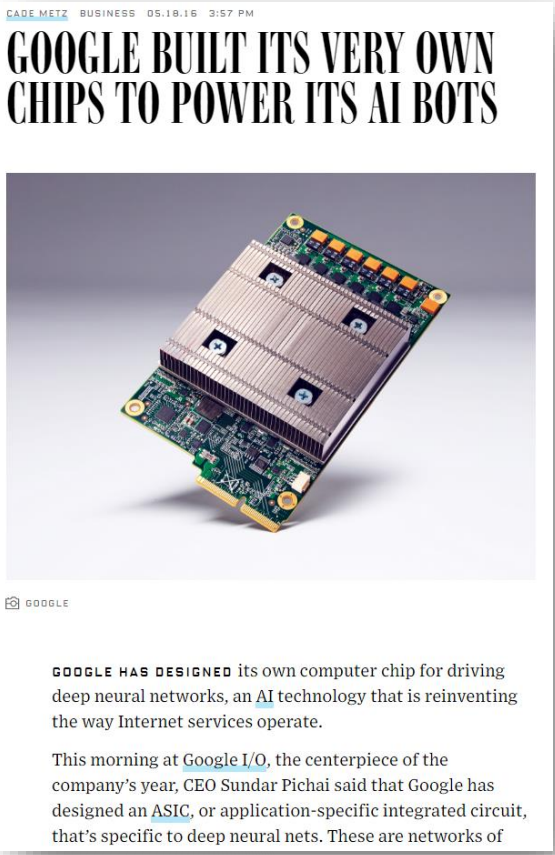
- Investigate new computational paradigms
  - Quantum
  - Neuromorphic
  - Advanced Digital
  - Emerging Memory Devices

Architectural specialization will accelerate

- Vendors, lacking Moore's Law, will need to continue to differentiate products (to stay in business)
- Grant that advantage of better CMOS process stalls
- 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



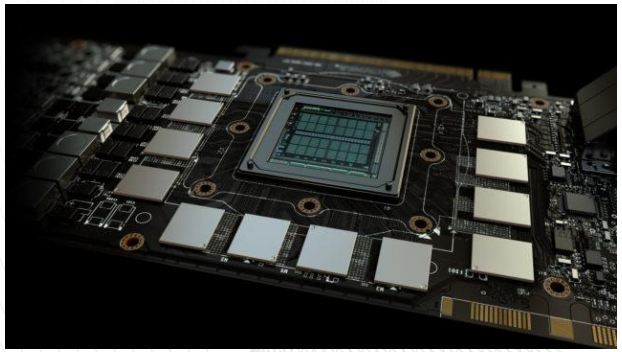
http://www.theinquirer.net/inquirer/news/2477796/intels-nervana-ai-platform-takes-aim-at-nvidias-gpu-techology



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

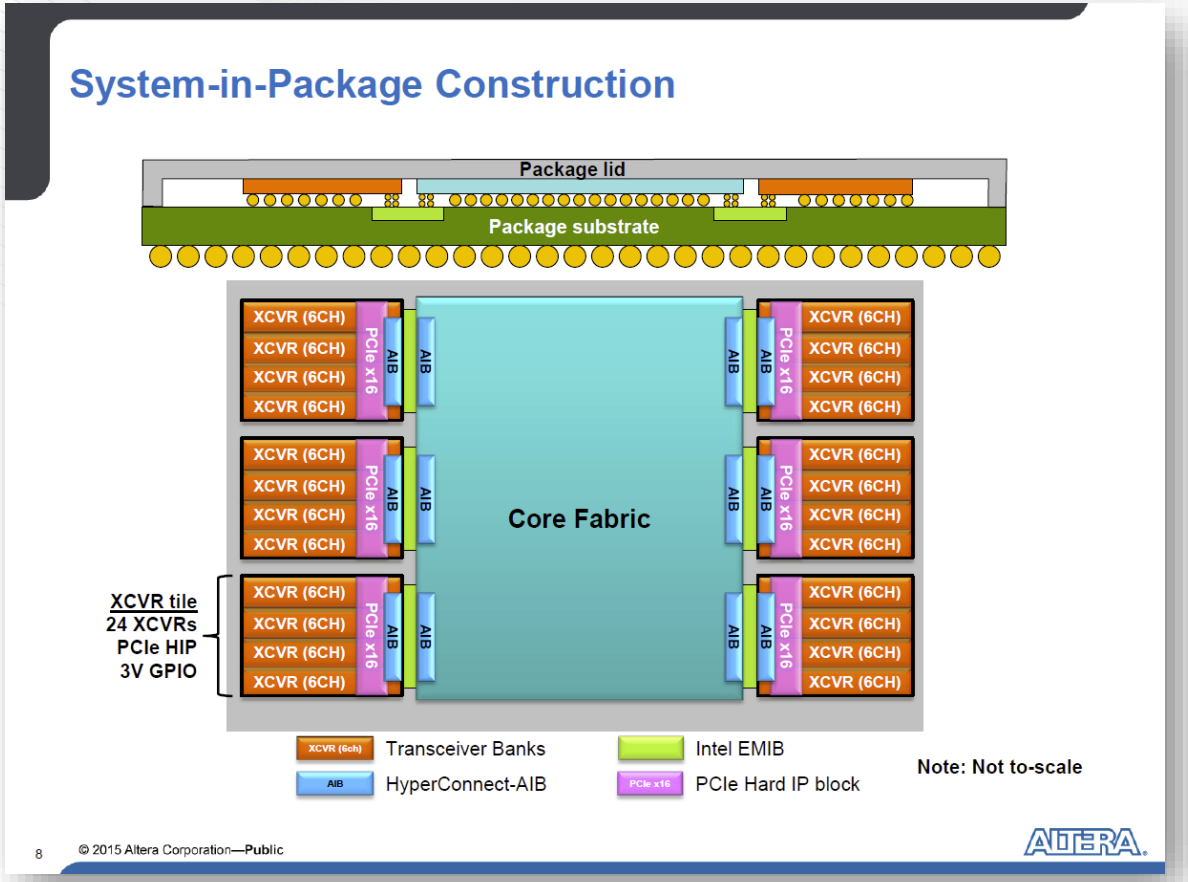
https://www.thebroadcastbridge.com/content/entry/1094/altera-announces-arria-10-2666mbps-ddr4-memory-fpga-interface

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.*

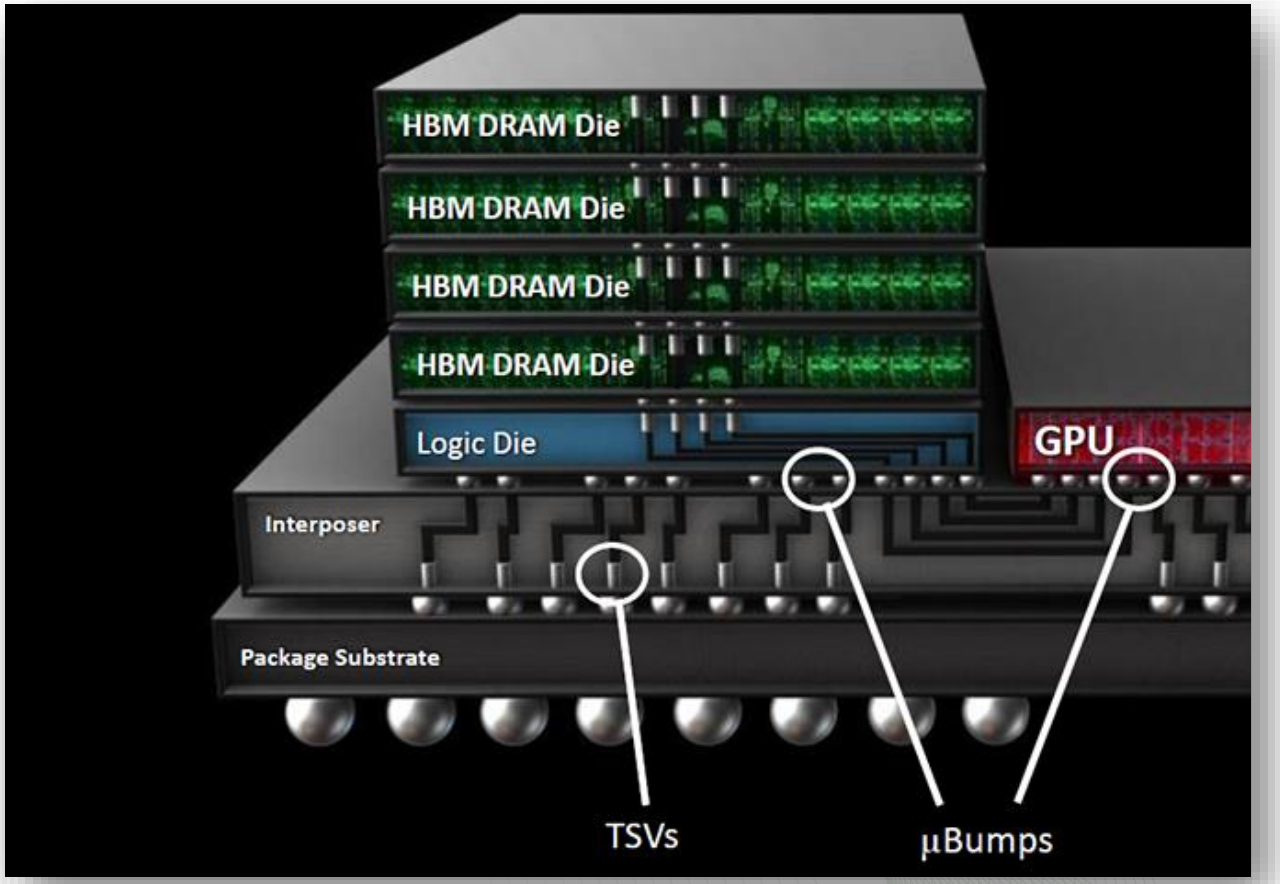


https://fossbytes.com/nvidia-volta-gddr6-2018/

Tighter integration and manufacturing of components will provide some benefits: components with different processes, functionality; local bandwidth



HotChips 27, Altera Stratix 10



http://www.anandtech.com/show/9969/jedec-publishes-hbm2-specification

OAK RIDGE National Laboratory

Our Transition Period Predictions

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

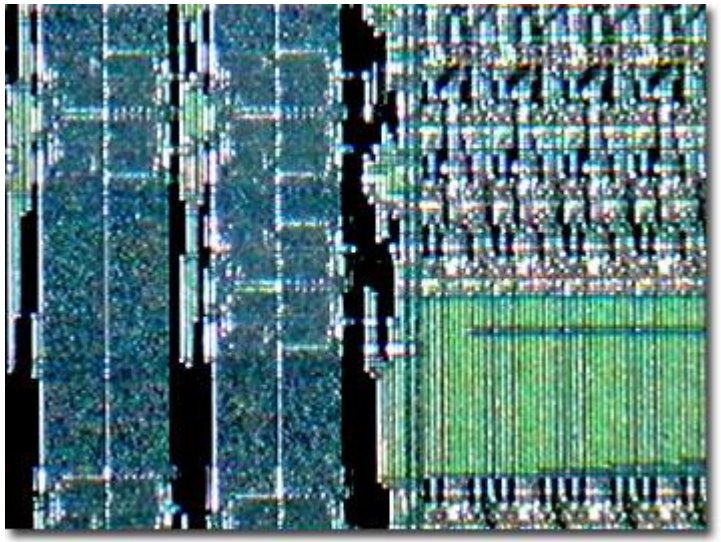
- Use CMOS more efficiently for our workloads
- Integrate components to boost performance and eliminate inefficiencies

Emerging Technologies

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

Exploration of emerging technologies is different this time – I promise

- Three decades of alternative technologies have fallen victim to 'curse of Moore's law': general CPU performance improvements without any software changes
  - Weitek Floating Point accelerator (circa 1988)
  - Piles of other types of processors: clearspeed,
  - FPGAs
- Some of these technologies found a specific market to serve
  - But most failed
- Now, the context and parameters have changed!



https://micro.magnet.fsu.edu/optics/olympusmicd/galleries/chips/weitekmathmedium.html

http://www.clearspeed.com

Transition Period will be Disruptive

- New devices and architectures may not be hidden in traditional levels of abstraction
  - 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
- Solutions need a co-design framework to evaluate and mature specific technologies

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

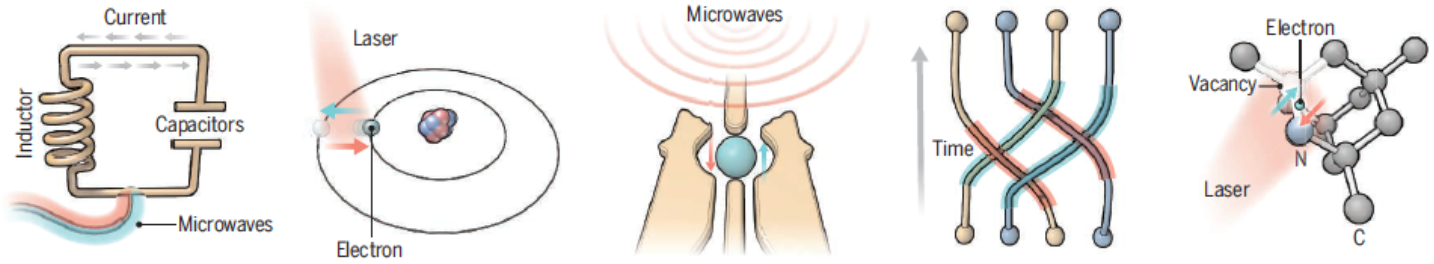
Adapted from IEEE Rebooting Computing Chart

Quantum computing: Qubit design and fabrication have made recent progress

Science 354, 1091 (2016) – 2 December

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 super-position 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 semi-conductor 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.

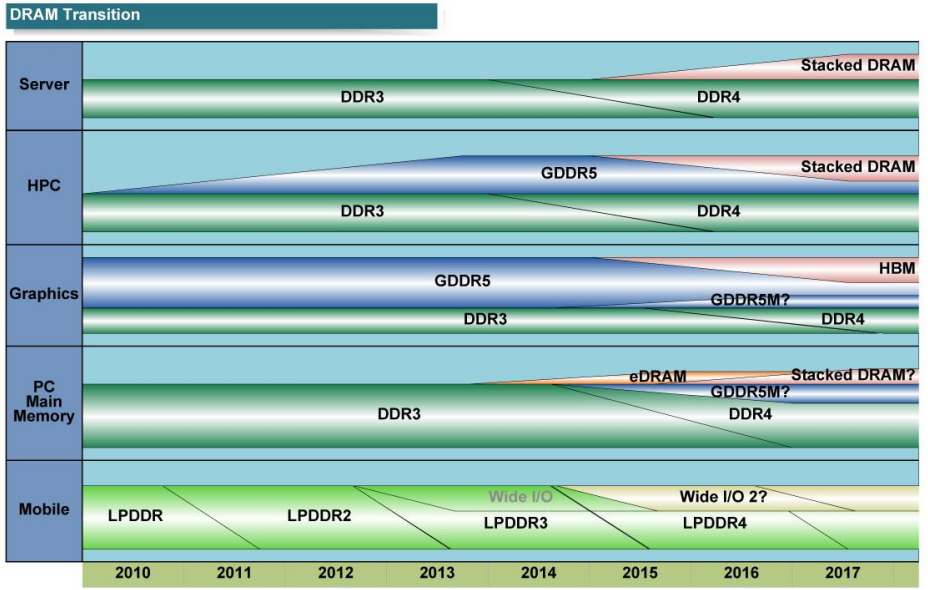
- Technological progress
  - Demonstrated qubits
- IBM, Google, Microsoft, Intel, D-Wave, etc investing in quantum
  - IBM has 17-qubit quantum computer on cloud
    - https://www.research.ibm.com/ibm-q/
- DOE has released to RFPs for quantum computing

New Memory Devices and Systems

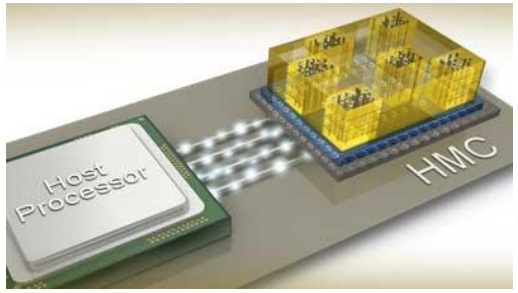
OAK RIDGE National Laboratory

New Memory Devices and Systems

- HMC, HBM/2/3, LPDDR4, GDDR5X, WIDEIO2, etc
- 2.5D, 3D Stacking
- New devices (ReRAM, PCRAM, STT-MRAM, Xpoint)
- Configuration diversity
  - Fused, shared memory
  - Scratchpads
  - Write through, write back, etc
  - Consistency and coherence protocols
  - Virtual v. Physical, paging strategies



http://gigglehd.com/zbxe/files/attach/images/1404665/988/406/011/788d3ba1967e2db3817d259d2e83c88e_1.jpg



https://www.micron.com/~/media/track-2-images/content-images/content_image_hmc.jpg?la=en

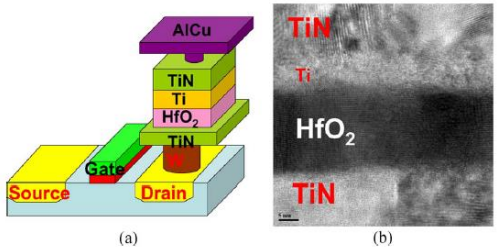


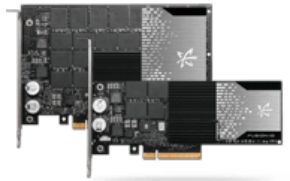
Fig. 4. *(a) A typical 1T1R structure of RRAM with HfO_x; (b) HR-TEM image of the TiN/Ti/HfO_x/TiN stacked layer; the thickness of the HfO_2 is 20 nm.*

H.S.P. Wong, H.Y. Lee, S. Yu *et al., "Metal-oxide RRAM," Proceedings of the IEEE, 100(6):1951-70, 2012.*

	SRAM	DRAM	eDRAM	2D NAND Flash	3D NAND Flash	PCRAM	STTRAM	2D ReRAM	3D ReRAM
Data Retention	N	N	N	Y	Y	Y	Y	Y	Y
Cell Size (F^2)	50-200	4-6	19-26	2-5	<1	4-10	8-40	4	<1
Minimum F demonstrated (nm)	14	25	22	16	64	20	28	27	24
Read Time (ns)	< 1	30	5	10^6	10^5	10-50	3-10	10-50	10-50
Write Time (ns)	< 1	50	5	10^5	10^5	100-300	3-10	10-50	10-50
Number of Rewrites	10^{16}	10^{16}	10^{16}	10^4-10^5	10^4-10^5	10^8-10^{10}	10^{15}	10^8-10^{12}	10^8-10^{12}
Read Power	Low	Low	Low	High	High	Low	Medium	Medium	Medium
Write Power	Low	Low	Low	High	High	High	Medium	Medium	Medium
Power (other than R/W)	Leakage	Refresh	Refresh	None	None	None	None	Sneak	Sneak
Maturity									

J.S. Vetter and S. Mittal, "Opportunities for Nonvolatile Memory Systems in Extreme-Scale High Performance Computing," CiSE, *17(2):73-82, 2015.*

NVRAM Technology Continues to Improve – Driven by Market Forces



designlines MEMORY

Blog

First Look at Samsung's 48L 3D V-NAND Flash

Kevin Gibb, Product Line Manager, TechInsights

4/6/2016 04:40 PM EDT

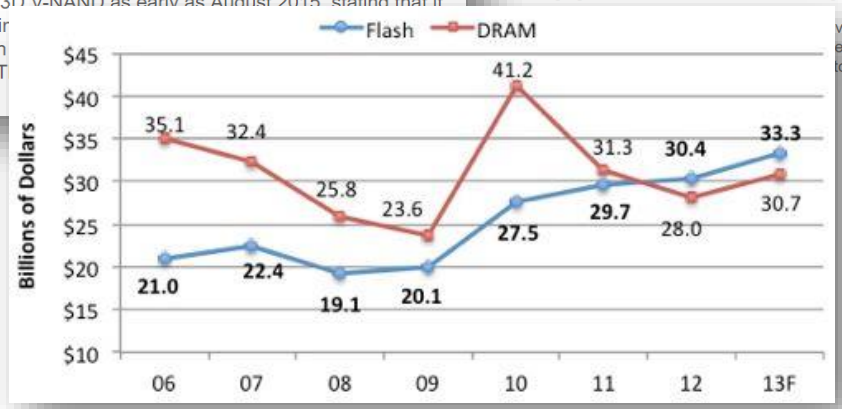
NO RATINGS
LOGIN TO RATE

9 comments post a comment

Like 16 Tweet Share 61 G+1 2

The highly anticipated Samsung's 48 layer V-NAND 3D flash memory is out in the market, and we at TechInsights have the first look.

Samsung had announced its 256 Gb 3-bit multi-level cell K9AFGY8S0M 3D V-NAND as early as August 2015, stating that it would be used in ... on the market in ... them in their 2 T... Figure 1.



http://www.eetasia.com/STATIC/ARTICLE_IMAGES/201212/EEOL_2012DEC28_STOR_MFG_NT_01.jpg

designlines WIRELESS & NETWORKING

Slideshow

Facebook Likes Intel's 3D XPoint

Google joins open hardware effort

Rick Merritt

3/10/2016 07:56 AM EST

7 comments

NO RATINGS
LOGIN TO RATE

Like 115 Tweet Share 46 G+1 3

SAN JOSE, Calif.—Facebook said it hopes to use Int... 3D XPoint memories in its data centers. Meanwhile G... its archrival's open hardware efforts to drive standard... high-power compute racks to giant form factors for di...

The two moves were likely the highest impact annour... the annual event of the Facebook-led Open Compute... (OCP) here. Among other news, Intel showed a new... SoC with dual 10G Ethernet controllers and a prototy... merging Xeon with an Arria FPGA in a single packag...

May 18, 2016

IBM Puts 3D XPoint on Notice with 3 Bits/Cell PCM Brea...

Tiffany Trader

IBM scientists have broken new ground in the ... change memory technology (PCM) that puts a... XPoint technology from Intel and Micron. IBM s...

designlines MEMORY

News & Analysis

3D NAND Flash at 2 Cents per GB

BeSang wants to lower barrier to 3D NAND flash

R. Colin Johnson

7/18/2016 07:10 PM EDT

14 comments

LOGIN TO RATE

Like 13 Tweet Share 129 G+1 3

LAKE WALES, Fla—The inventor of 3D monolithic chip technology back in 2010, BeSang Inc. (Beaverton, Ore.), claims to have since created a superior three-dimensional (3D) architecture for NAND flash. Frustrated with licensee Hynix's slow implementation of its monolithic 3D technology, BeSang is opening the door to partnerships with other memory houses, as well as offering to contract-fab the chips for resale by others, at a price that reduces the cost-per-bit of 3D NAND from over 20¢ to about 2¢ per gigabyte.

Original URL: http://www.theregister.co.uk/2013/11/01/hp_memristor_2018/

HP 100TB Memristor drives by 2018 – if you're lucky, admits tech titan

Universal memory slow in coming

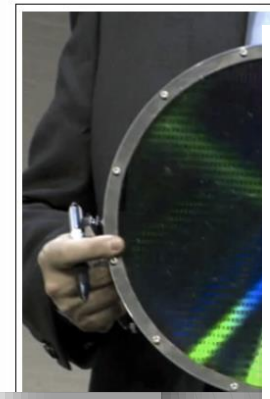
By Chris Mellor

Posted in Storage, 1st November 2013 02:28 GMT

Blocks and Files HP has warned *El Reg* not to get its hopes up too high after the tech titan's CTO Martin Fink suggested StoreServ arrays could be packed with 100TB Memristor drives come 2018.

In five years, according to Fink, DRAM and NAND scaling will hit a wall, limiting the maximum capacity of the technologies: process shrinks will come to a shuddering halt when the memories' reliability drops off a cliff as a side effect of reducing the size of electronics on the silicon dies.

The HP answer to this scaling wall is Memristor, its flavour of resistive RAM technology that is supposed to have DRAM-like speed and better-than-NAND storage density. Fink claimed at an HP Discover event in Las Vegas that Memristor devices will be ready by the time flash NAND hits its limit in five years. He also showed off a Memristor wafer, adding that it could have a 1.5PB capacity by the end of the decade.



designlines MEMORY

News & Analysis

Samsung Debuts 3D XPoint Killer

3D NAND variant stakes out high-end SSDs

Rick Merritt

8/11/2016 00:01 AM EDT

5 comments

NO RATINGS
1 saves
LOGIN TO RATE

Like 56 Tweet Share 212 G+1 4

SANTA CLARA, Calif. – Samsung lobbed a new variant of its 3D NAND flash into the gap Intel and Micron hope to fill with their emerging 3D XPoint memory. The news came one day after Micron showed at the Flash Memory Summit performance figures for its version of the XPoint solid-state drives (SSDs) under a new Quantx brand.

Samsung announced plans for what it called Z-NAND chips that will power SSDs with similar performance but lower costs and risk than the 3D XPoint drives. However, it was secretive about the details of the technology that will appear in products sometime next year.

By contrast, a Micron engineer leading its XPoint SSD program was surprisingly candid in an interview with *EE Times*. She described current prototypes using early XPoint chips and an FPGA-based controller for the SSDs expected to ship in about a year.

Samsung's Z-NAND will deliver 10x faster reads than multi-level cell flash and writes that are twice as fast, the company said. At the drive level, they will support both reads and writes at about 20 microseconds, suggesting some of write performance comes from an enhanced controller.

JUL 28, 2015 @ 2:46 PM 7,391 VIEWS

Intel And Micron Jointly Announce Game-Changing 3D XPoint Memory Technology

Comparison of Emerging Memory Technologies

	SRAM	DRAM	eDRAM	2D NAND Flash	3D NAND Flash	PCRAM	STTRAM	2D ReRAM	3D ReRAM
	Deployed					Experimental			
Data Retention	N	N	N	Y	Y	Y	Y	Y	Y
Cell Size (F^2)	50-200	4-6	19-26	2-5	<1	4-10	8-40	4	<1
Minimum F demonstrated (nm)	14	25	22	16	64	20	28	27	24
Read Time (ns)	< 1	30	5	10^4	10^4	10-50	3-10	10-50	10-50
Write Time (ns)	< 1	50	5	10^5	10^5	100-300	3-10	10-50	10-50
Number of Rewrites	10^{16}	10^{16}	10^{16}	10^4-10^5	10^4-10^5	10^8-10^{10}	10^{15}	10^8-10^{12}	10^8-10^{12}
Read Power	Low	Low	Low	High	High	Low	Medium	Medium	Medium
Write Power	Low	Low	Low	High	High	High	Medium	Medium	Medium
Power (other than R/W)	Leakage	Refresh	Refresh	None	None	None	None	Sneak	Sneak
Maturity									

Intel/Micron Xpoint?
Samsung Z-NAND?

http://ft.ornl.gov/trac/blackcomb

Aggressively-pursued NVM Technologies Continue to Improve

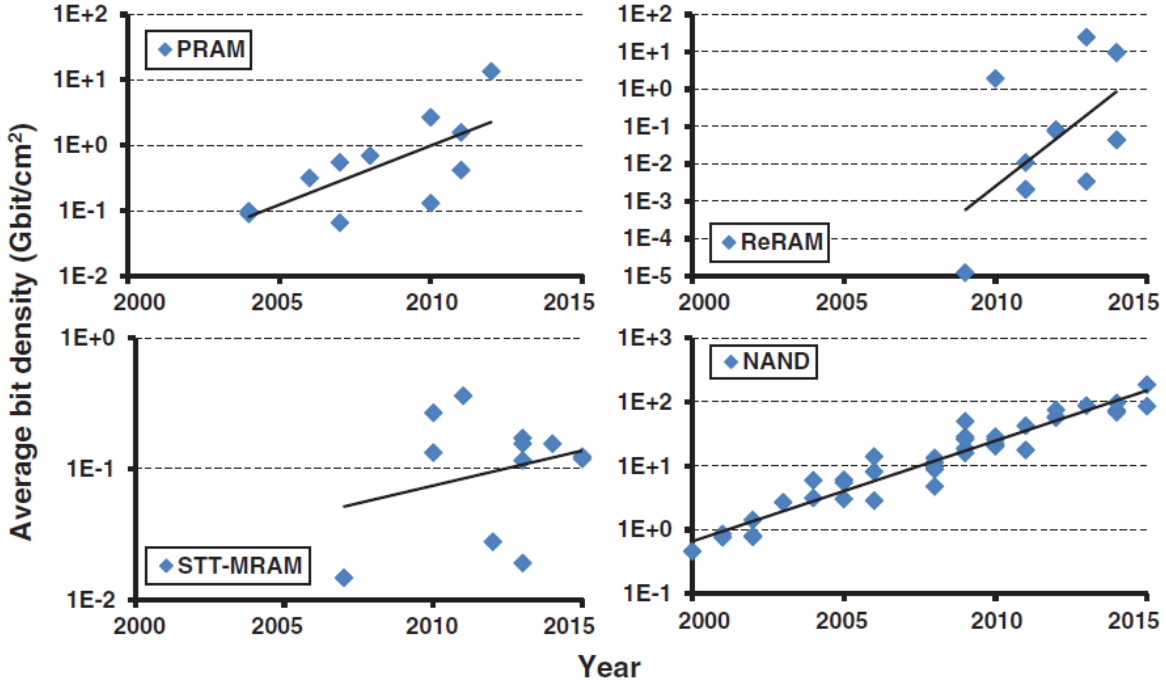


Figure 3: Average bit density

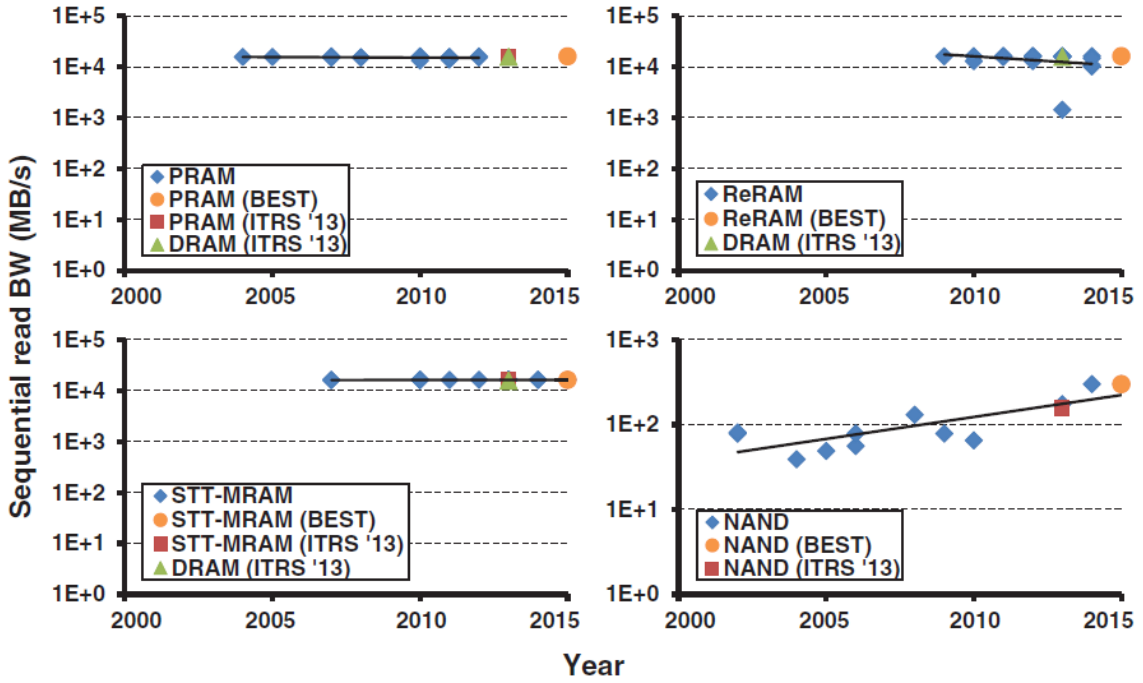


Figure 11: Sequential read BW

Source: Suzuki, 2015

Microcosm of DOE HPC Architectures are Reflecting these Trends

Projections: Exascale architecture targets circa 2009

2009 Exascale Challenges Workshop in San Diego

Attendees envisioned two possible architectural swim lanes:

1. Homogeneous many-core thin-node system
2. Heterogeneous (accelerator + CPU) fat-node system

System attributes	2009	"Pre-Exascale"		"Exascale"	
System peak	2 PF	100-200 PF/s		1 Exaflop/s	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32–64 PB	
Storage	15 PB	150 PB		500 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/s	1 TB/s	0.4 TB/s	4 TB/s
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	500,000	50,000	1,000,000	100,000
Node interconnect BW	1.5 GB/s	150 GB/s	1 TB/s	250 GB/s	2 TB/s
IO Bandwidth	0.2 TB/s	10 TB/s		30-60 TB/s	
MTTI	day	O(1 day)		O(0.1 day)	

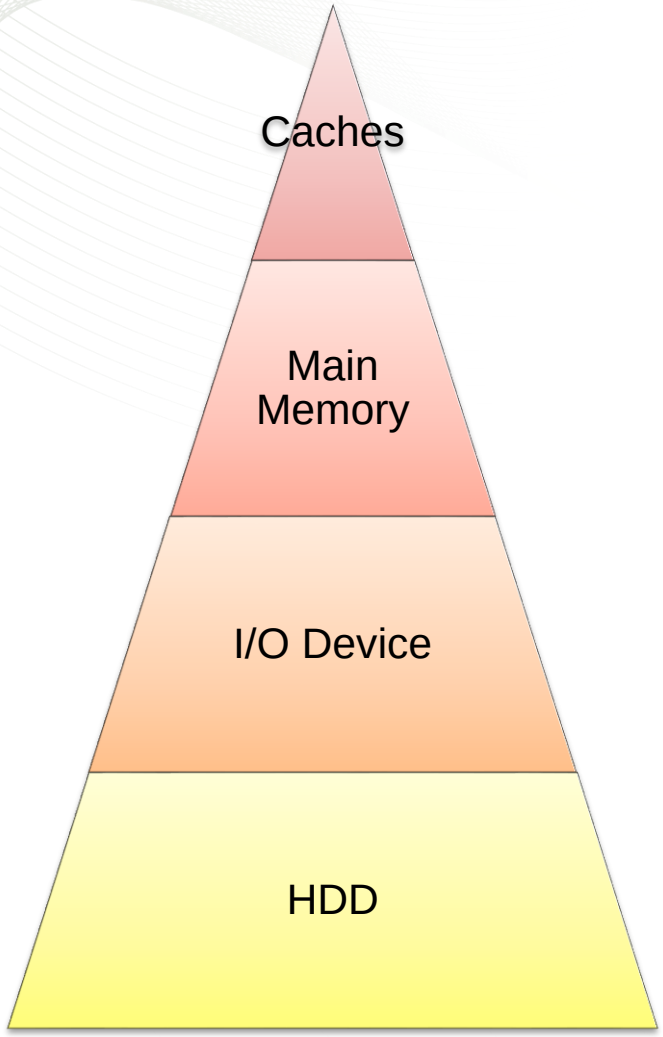
Architectural specialization and integration will create very complex platforms – no two systems alike

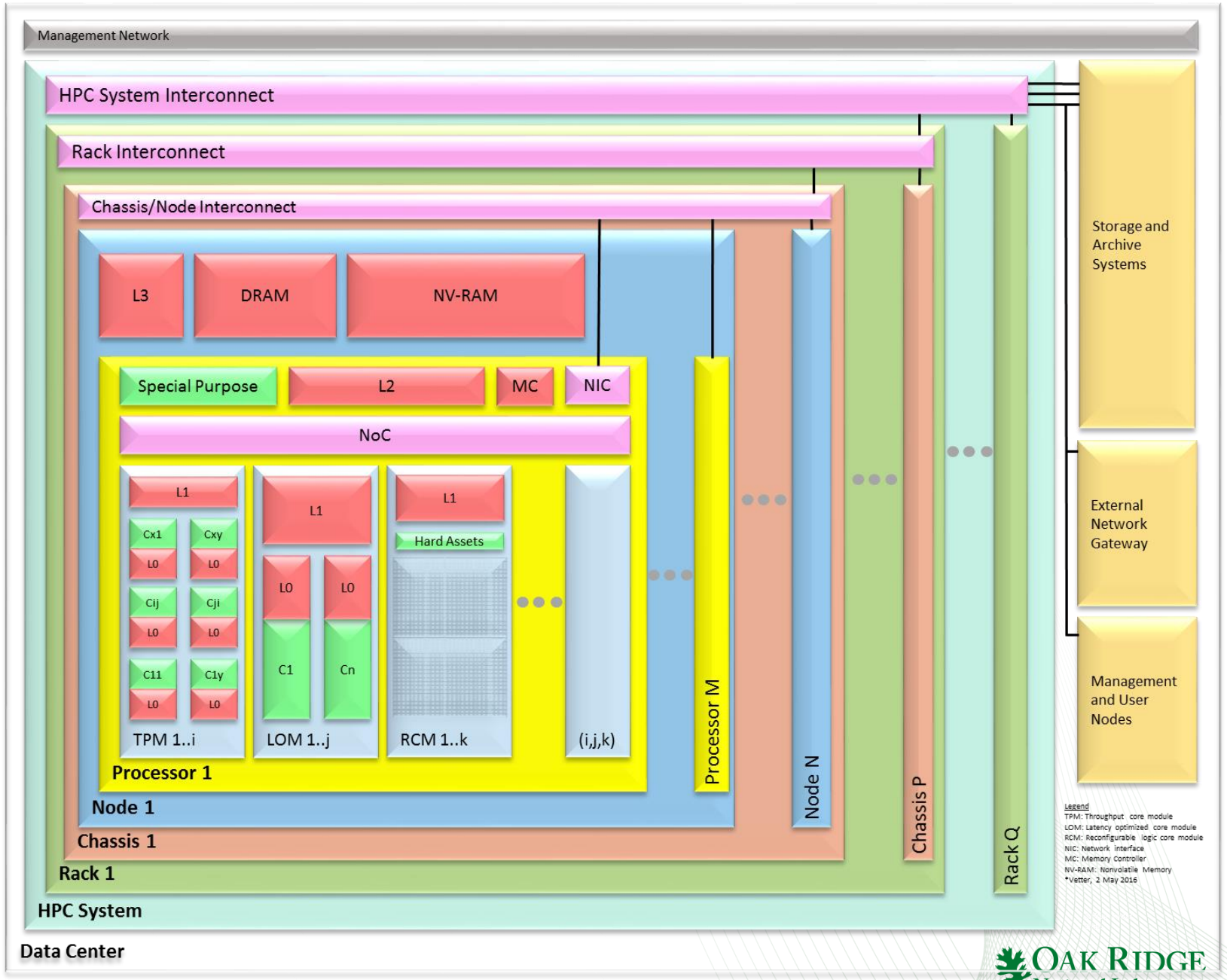
- Architectural specialization and integration will create very complex platforms
  - Heterogeneous computing
  - Deep memory hierarchies incl NVM
  - Plateauing I/O forces app redesign
- Implications
  - DOE and companies will need to understand design tradeoffs
  - Programming systems will need to be portable
  - Performance portability will be a stretch goal

System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrades	
Planned Installation	**Edison**	**TITAN**	**MIRA**	**Cori 2016**	**Summit 2017-2018**	**Theta 2016**	**Aurora 2018-2019**
System peak (PF)	2.6	27	10	> 30	150	>8.5	180
Peak Power (MW)	2	9	4.8	< 3.7	10	1.7	13
Total system memory	357 TB	710TB	768TB	~1 PB DDR4 + High Bandwidth Memory (HBM)+1.5PB persistent memory	> 1.74 PB DDR4 + HBM + ~~2.8~~ ~~3.7~~ ~7 PB persistent memory	>480 TB DDR4 + High Bandwidth Memory (HBM)	> 7 PB High Bandwidth On-Package Memory Local Memory and Persistent Memory
Node performance (TF)	0.460	1.452	0.204	> 3	> 40	> 3	> 17 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	Multiple IBM Power9 CPUs & multiple Nvidia Voltas GPUS	Intel Knights Landing Xeon Phi many core CPUs	Knights Hill Xeon Phi many core CPUs
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	~3,500 nodes	>2,500 nodes	>50,000 nodes
System Interconnect	Aries	Gemini	5D Torus	Aries	Dual Rail EDR-IB	Aries	2nd Generation Intel Omni-Path Architecture
File System	7.6 PB 168 GB/s, Lustre®	32 PB 1 TB/s, Lustre®	26 PB 300 GB/s GPFS™	28 PB 744 GB/s Lustre®	120 PB 1 TB/s GPFS™	10PB, 210 GB/s Lustre initial	150 PB 1 TB/s Lustre®

Complexity α T

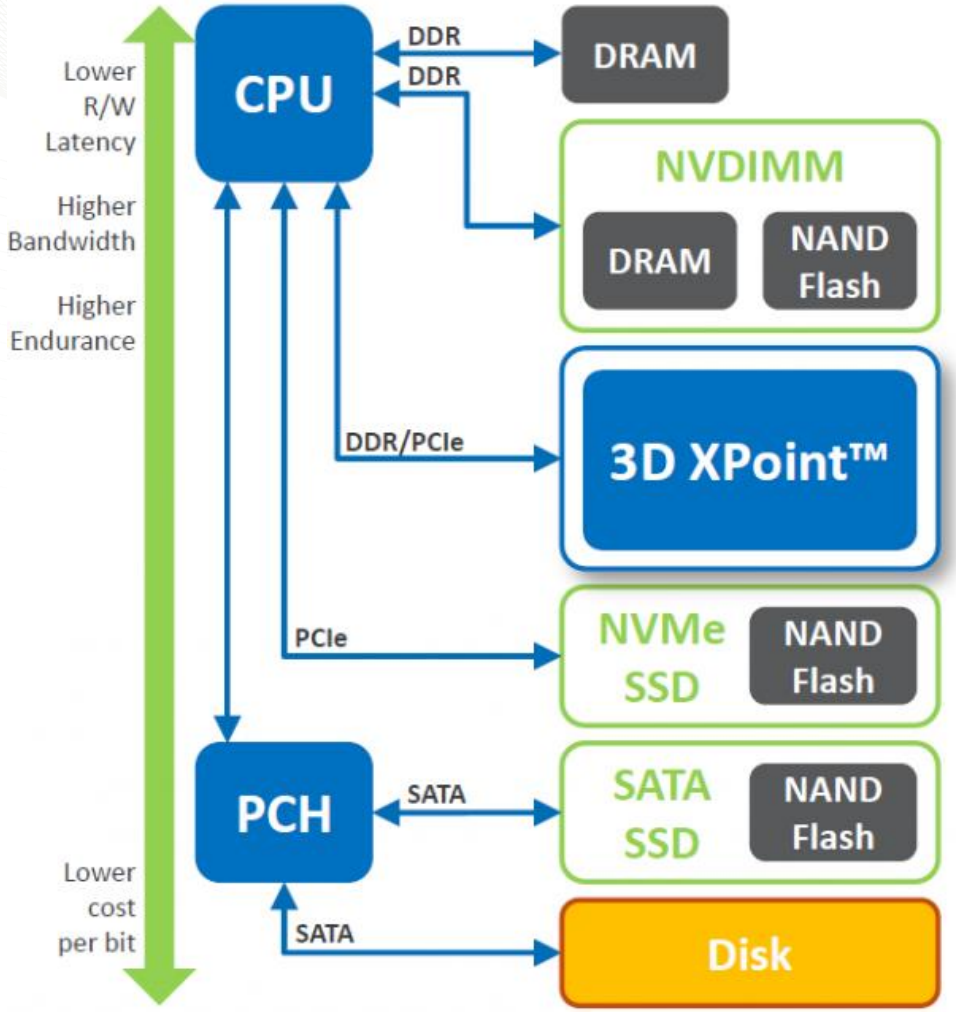
Migration up the hierarchy

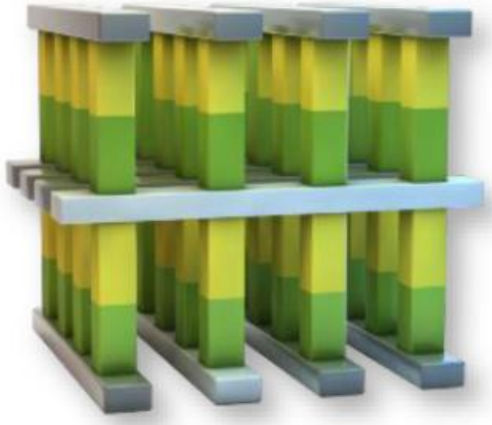




Programming NVM on a Node

NVM Architectural Assumptions



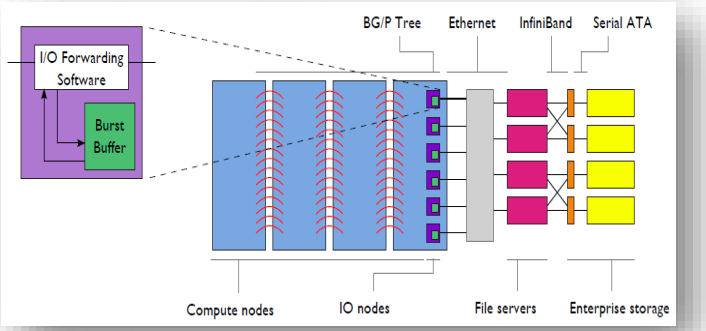


- 3D XPoint technology provides the benefit in the middle
- It is considerably faster than NAND Flash
- Performance can be realized on PCIe or DDR buses
- Lower cost per bit than DRAM while being considerably more dense

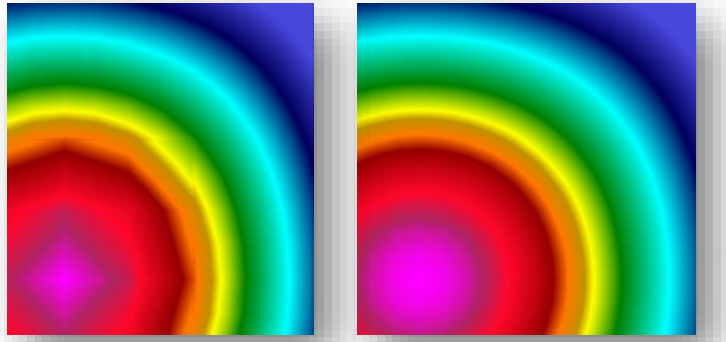
https://www.kitguru.net/components/memory/anton-shilov/intel-first-3d-xpoint-ssds-will-feature-up-to-6gbs-of-bandwidth/

HPC Application Scenarios for NVM

- Burst Buffers, C/R [Liu, et al., MSST 2012]



- In situ visualization



http://ft.ornl.gov/eavl

- In-mem tables

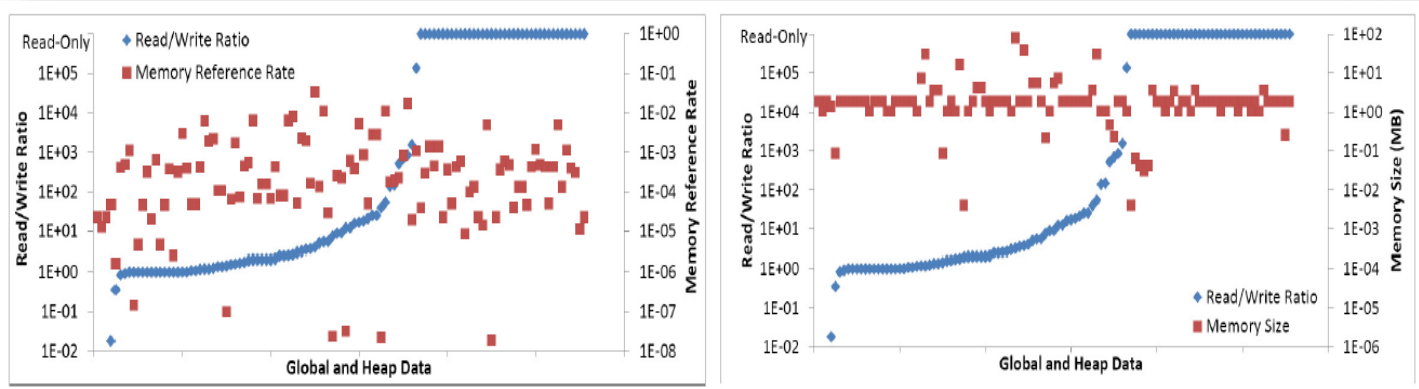


Figure 3: Read/write ratios, memory reference rates and memory object sizes for memory objects in Nek5000

OAK RIDGE National Laboratory

J.S. Vetter and S. Mittal, "Opportunities for Nonvolatile Memory Systems in Extreme-Scale High-Performance Computing," *Computing in Science & Engineering, 17(2):73-82, 2015.*

Observations: Numerous characteristics of applications are a good match for byte-addressable NVRAM

Empirical results show many reasons…

- Lookup, index, and permutation tables
- Inverted and ‘element-lagged’ mass matrices
- Geometry arrays for grids
- Thermal conductivity for soils
- Strain and conductivity rates
- Boundary condition data
- Constants for transforms, interpolation
- MC Tally tables, cross-section materials tables…

large scale NVM into applications

- Transactional protocol to preserve local and global consistency and redundancy
- Name indexing
- Object attributes allow local caching

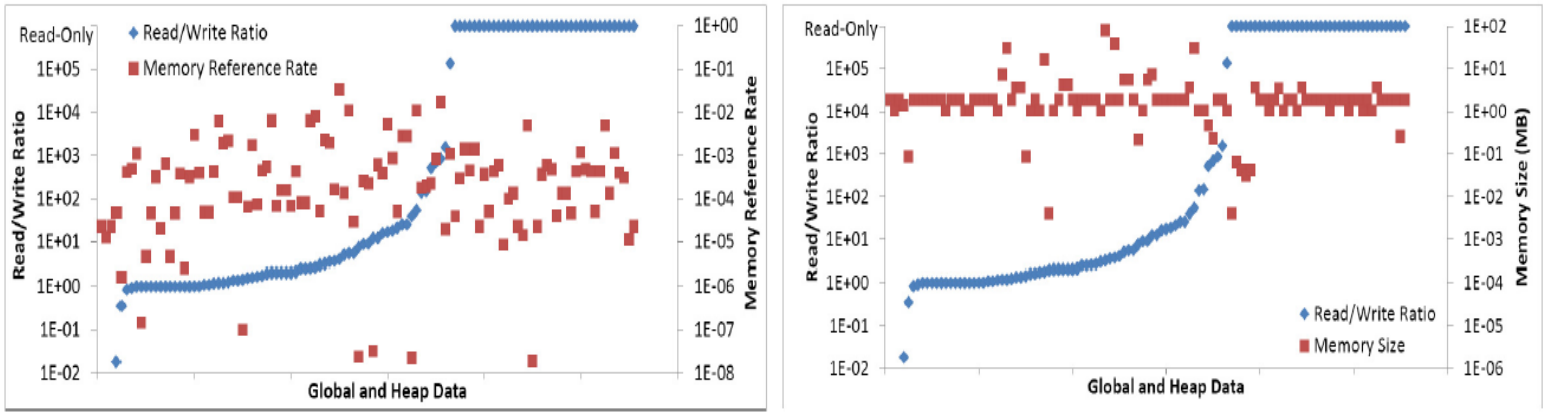


Figure 3: Read/write ratios, memory reference rates and memory object sizes for memory objects in Nek5000

NVL-C: extending C to support NVM

J. Denny, S. Lee, and J.S. Vetter, “NVL-C: Static Analysis Techniques for Efficient, Correct Programming of Non-Volatile Main Memory Systems,” in *ACM High Performance Distributed Computing (HPDC). Kyoto: ACM, 2016*

OAK RIDGE National Laboratory

Design Goals: Familiar programming interface

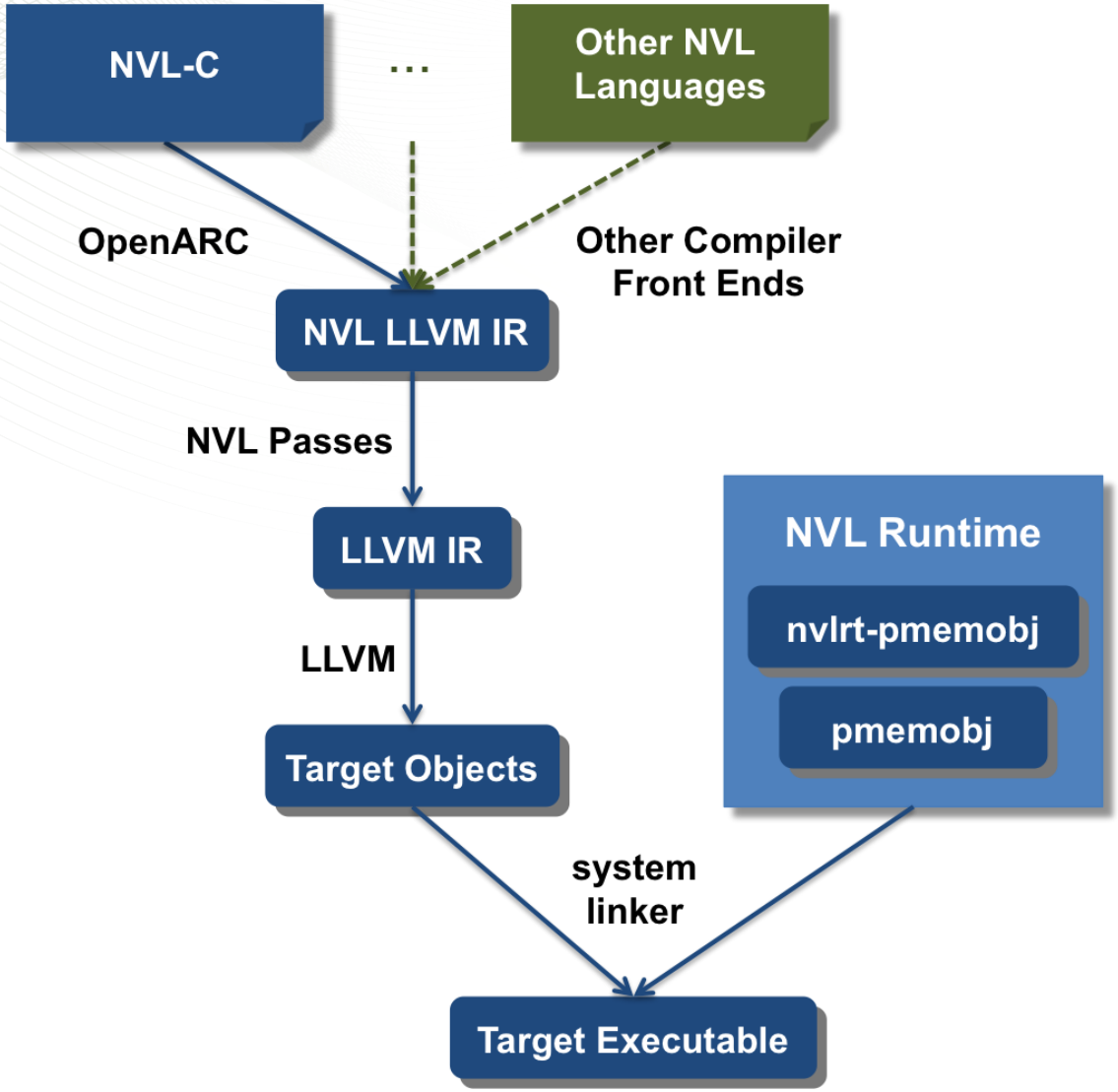
```
#include <nvl.h>
struct list {
  int value;
  nvl struct list *next;
};
void add(int k, nvl struct list *after) {
  nvl struct list *node
    = nvl_alloc_nv(heap, 1, struct list);
  node->value = k;
  node->next  = after->next;
  after->next = node;
}
```

- Small set of C language extensions:
  - Header file
  - Type qualifiers
  - Library API
  - Pragmas
- Existing memory interfaces remain:
  - NVL-C is a superset of C
  - Unqualified types as specified by C
  - Local/global variables stored in volatile memory (DRAM or registers)
  - Use existing C standard libraries for HDD

Design Goals: Avoiding persistent data corruption

- New categories of pointer bugs:
  - Caused by multiple memory types:
    - E.g., pointer from NVM to volatile memory will become dangling pointer
  - Prevented at compile time or run time
- Automatic reference counting:
  - No need to manually free
  - Avoids leaks and dangling pointers
- Transactions:
  - Avoids persistent data corruption across software and hardware failures
- High performance:
  - Performance penalty from memory management, pointer safety, and transactions
  - Compiler-based optimizations
  - Programmer-specified hints

Design Goals: Modular implementation



- Core is common compiler middle-end
- Multiple complier front ends for multiple high-level languages:
  - For now, just OpenARC for NVL-C
- Multiple runtime implementations:
  - For now, just Intel's pmem (pmemobj)

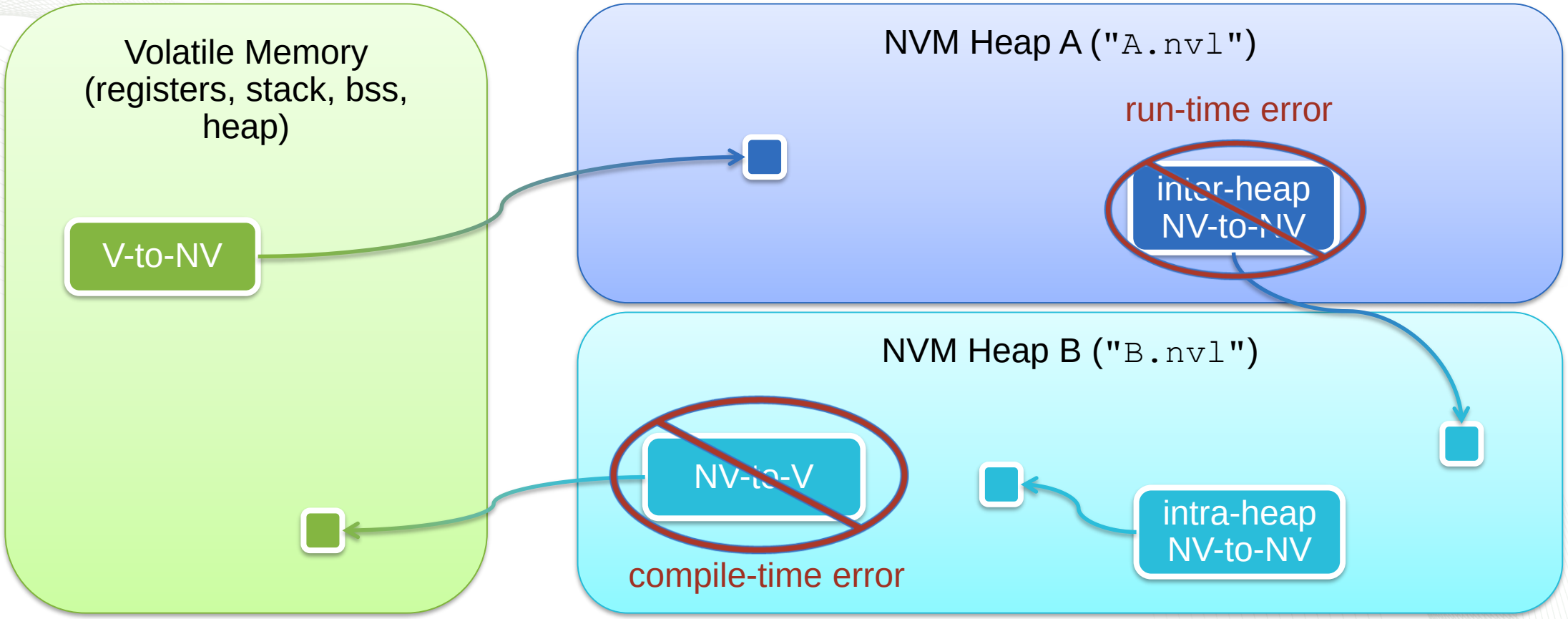
Programming Model: NVM Pointers

```
#include <nvl.h>
struct list {
  int value;
  nvl struct list *next;
};
void add(int k, nvl struct list *after) {
  struct list *node
    = malloc(sizeof(struct list));
  node->value = k;
  node->next  = after->next;
  after->next = node;
}
```

compile-time error
explicit cast won't help

- **nvl** type qualifier:
  - Indicates NVM storage
  - On target type, declares NVM pointer
  - No NVM-stored local or global variable
- Stricter type safety for NVM pointers:
  - Does not affect other C types
  - Avoids persistent data corruption
  - Facilitates compiler analysis
  - Needed for automatic reference counting
  - E.g., pointer conversions involving NVM pointers are strictly prohibited

Programming Model: Pointer types (like Coburn et al.)



avoids dangling pointers when memory segments close

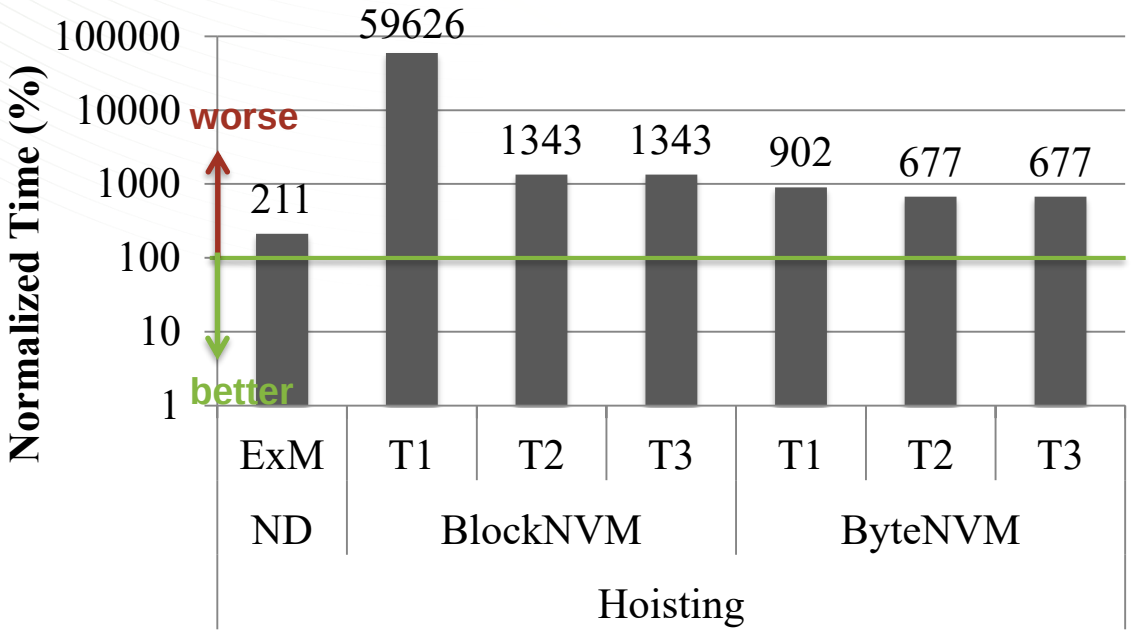
Programming Model: Transactions: Undo logs

```
#include <nvl.h>
void matmul(nvl float a[I][J],
            nvl float b[I][K],
            nvl float c[K][J],
            nvl int *i)
{
  while (*i<I) {
    #pragma nvl atomic heap(heap)
    {
      for (int j=0; j<J; ++j) {
        float sum = 0.0;
        for (int k=0; k<K; ++k)
         sum += b[*i][k] * c[k][j];
        a[*i][j] = sum;
      }
      ++*i;
    }
  }
}
```

- Before every NVM store, transaction creates undo log to back up old data
- Undo log contains metadata plus old data being overwritten
- Problem: large overhead because an undo log is created for every element of `a` (every iteration of `j` loop)

Evaluation: LULESH

- `backup` is important for performance
- `clobber` cannot be applied because old data is needed



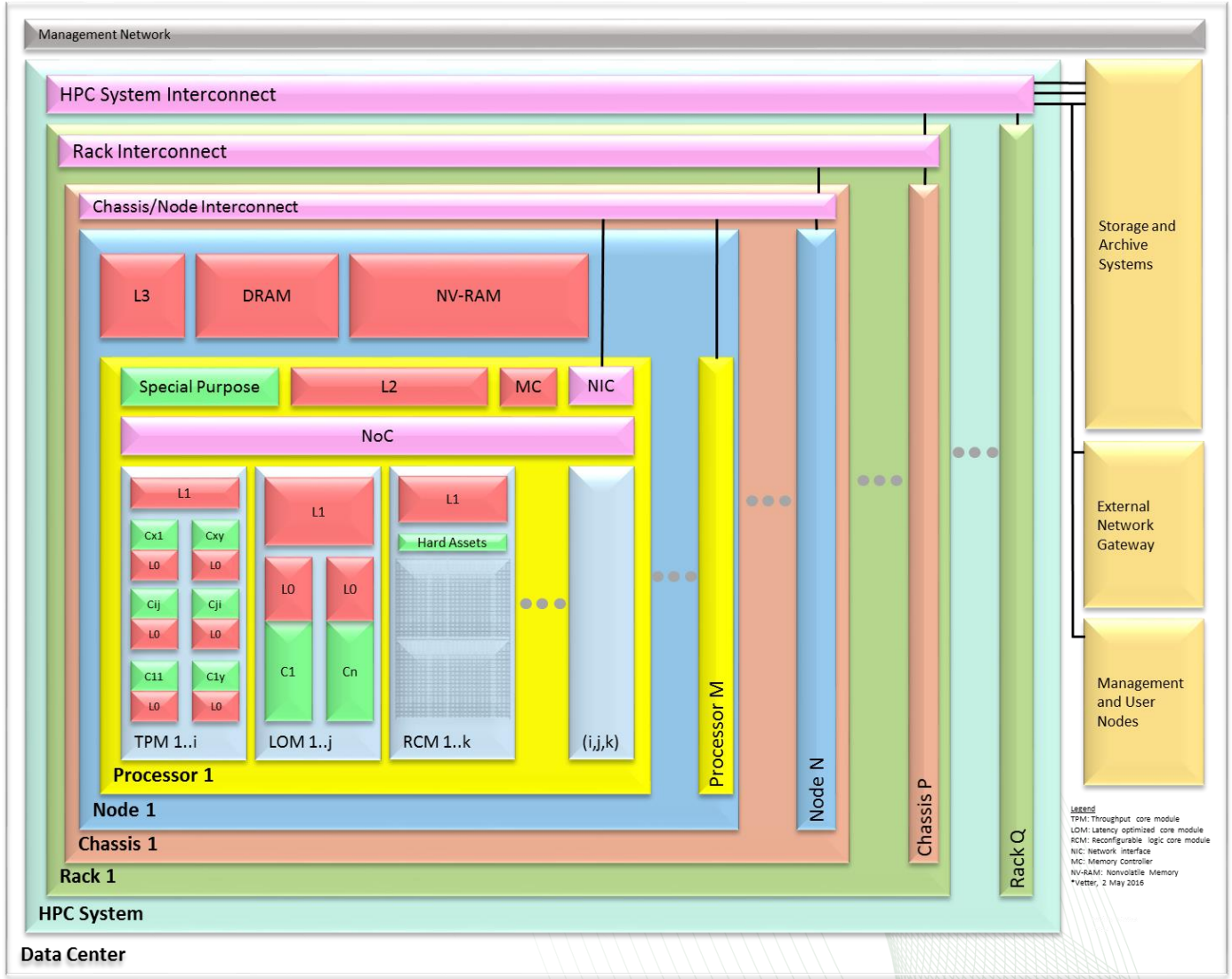
- ExM = use SSD as extended DRAM
- T1 = BSR + transactions
- T2 = T1 + `backup` clauses
- T3 = T1 + `clobber` clauses
- BlockNVM = `msync` included
- ByteNVM = `msync` suppressed

Programming Scalable NVM with Papyrus(KV)

OAK RIDGE National Laboratory

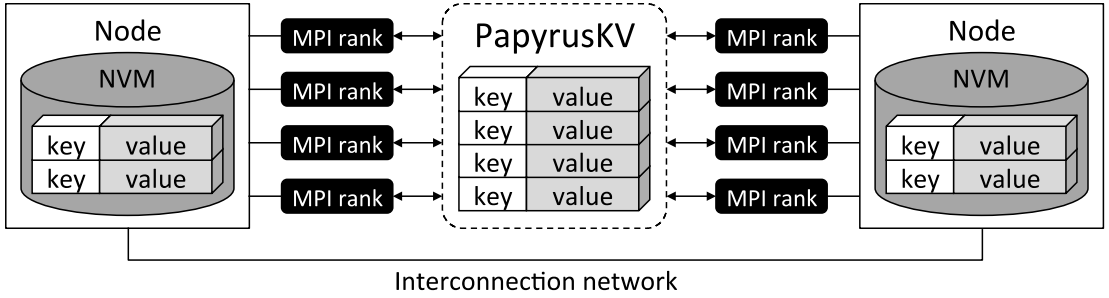
Scalable NVM Architectural Assumptions

- NVM Architectures will vary dramatically -> Portability
  - Exploit persistence?
  - Where in the hierarchy?
    - Already in external storage system
    - Rack mounted appliance (Cori)
    - Chassis shared?
    - Node shared? (Summit)
- Design Goals
  - Performance
    - Must provide performance better than or equal to alternatives for providing large shared storage for HPC apps
  - Scalability (to millions of threads)
  - Interoperability with existing HPC programming models
    - Can be incrementally introduced
    - Leverage features of other programming models
  - Application customizability
    - Usage scenarios vary dramatically
    - Tunable consistency model, protection attributes (e.g., RO)

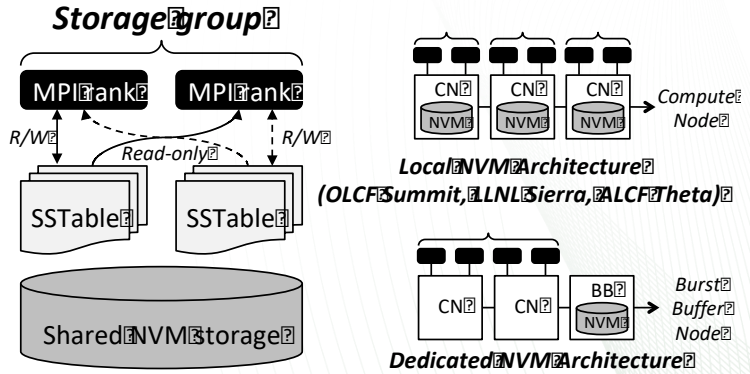


PapyrusKV: A High-Performance Parallel Key-Value Store for Distributed NVM Architectures

- Leverage emerging NVM technologies
  - High performance
  - High capacity
  - Persistence property
- Designed for the next-generation DOE/NNSA systems
  - Portable across local NVM and dedicated NVM architectures
  - An embedded key-value store (no system-level daemons and servers)
- Designed for HPC applications
  - MPI/UPC-interoperable
  - Application customizability
    - Memory consistency models (sequential and relaxed)
    - Protection attributes (read-only, write-only, read-write)
    - Load balancing
  - Zero-copy workflow, asynchronous checkpoint/restart



PapyrusKV stores keys and values in arbitrary byte arrays across multiple NVM devices in a distribute system



PapyrusKV is portable across local NVM and dedicated NVM architectures

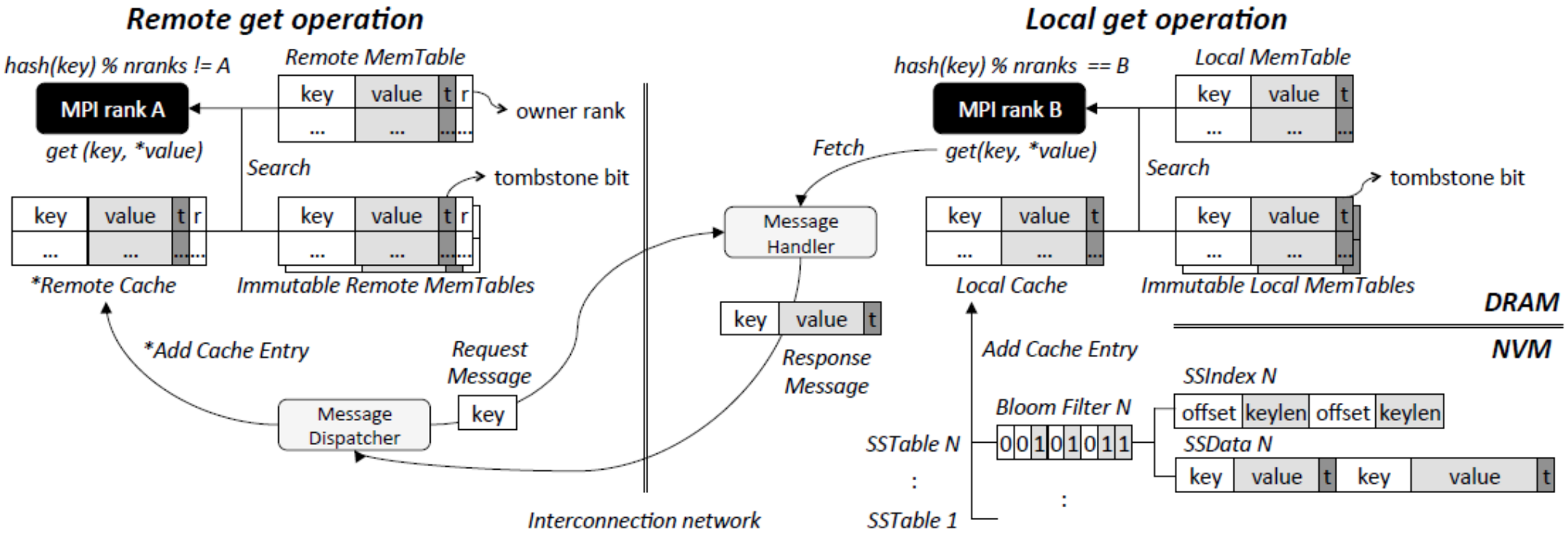
J. Kim, S. Lee, and J. S. Vetter, "PapyrusKV: A High-Performance Parallel Key-Value Store for Distributed NVM Architectures," *In Proc. of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2017 (accepted)

PapyrusKV Application API

Table 1: The PKV API.

API Function	Description	Collective
(a) Environment		
pkv_init(int argc, char*** argv, const char* repository)*	Initialize execution environment using *repository* path	×
pkv_finalize()	Terminate execution environment	×
(b) Basic		
pkv_open(const char name, int flags, pkv_option_t* opt, pkv_db_t* db)*	Open or create *db* with *name*	×
pkv_close(pkv_db_t db)	Close *db*	×
pkv_put(pkv_db_t db, const char key, size_t keylen, const char* value, size_t valuelen)*	Insert or update a *key-value* pair to *db*	
pkv_get(pkv_db_t db, const char key, size_t keylen, char** value, size_t* valuelen)*	Retrieve *value* for a given *key* from *db*. If *value* is not allocated in memory, PKV allocates a new heap region from the PKV memory pool. Otherwise, data is copied to *value* directly	
pkv_delete(pkv_db_t db, const char key, size_t keylen)*	Delete a key-value pair for a given *key* from *db*	
pkv_free(pkv_db_t db, char val)*	Release a heap memory region allocated from the PKV memory pool	
(c) Consistency		
pkv_signal_notify(int signum, int ranks, int count)*	Send signals to *ranks*	
pkv_signal_wait(int signum, int ranks, int count)*	Wait for signals from *ranks*	
pkv_fence(pkv_db_t db)	Migrate the remote MemTable and immutable MemTables to the owner ranks immediately	
pkv_barrier(pkv_db_t db, int level)	Collective memory fence with a flushing *level* (`PKV_MEMTABLE` or `PKV_SSTABLE`). With `PKV_SSTABLE` level, the whole *db* data are flushed to SSTables.	×
pkv_consistency(pkv_db_t db, int mode)	Set memory consistency mode on *db* to *mode* (`PKV_SEQUENTIAL` or `PKV_RELAXED`)	×
pkv_protect(pkv_db_t db, int prot)	Set protection attribute on *db* to *prot* (`PKV_RDWR`, `PKV_WRONLY`, or `PKV_RDONLY`)	×
(d) Persistence		
pkv_checkpoint(pkv_db_t db, const char path, pkv_event_t* event)*	Generate a snapshot of *db* into *path*. It runs asynchronously if *event* is not NULL	×
pkv_restart(const char path, const char* name, int prot, pkv_db_t* db, pkv_event_t* event)*	Revert *db* with *name* from a snapshot stored in *path*. It runs asynchronously if *event* is not NULL	×
pkv_destroy(pkv_db_t db, pkv_event_t event)*	Remove *db* and all its data from NVM. It runs asynchronously if *event* is not NULL	×
pkv_wait(pkv_db_t db, pkv_event_t event)	Wait for *event* to complete	×

PapyrusKV Example Get operations



Present design allows remote cache only for RO data.

Evaluation

- Evaluation results on OLCF's SummitDev, TACC's Stampede (KNL), and NERSC's Cori

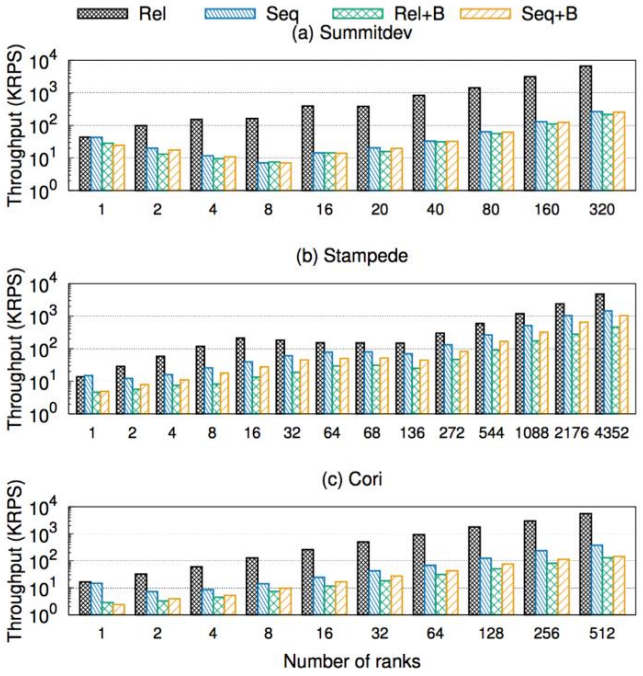


Figure 7: Put operation performance in relaxed (Rel) and sequential (Seq) consistency modes. B refers to Barrier.

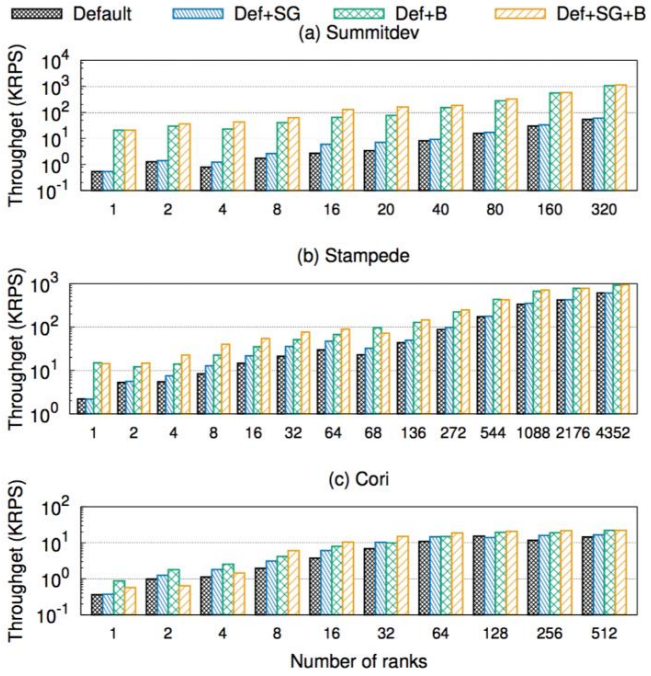


Figure 8: Get operation performance. SG and B refer to Storage Group and SSTable Binary search, respectively.

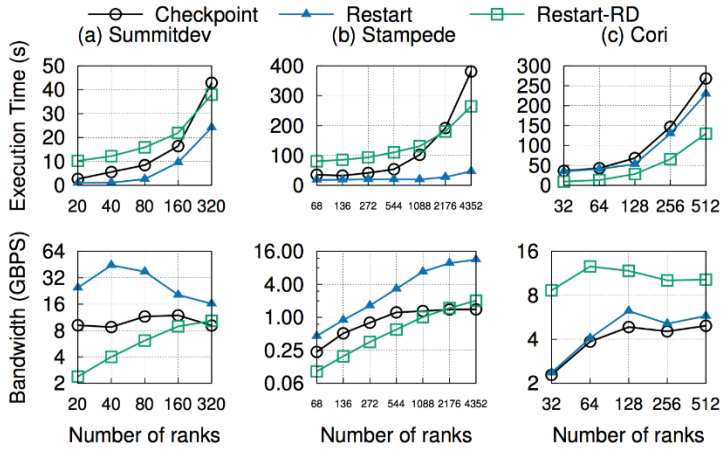


Figure 10: Checkpoint, restart, and restart with redistribution (RD) performance.

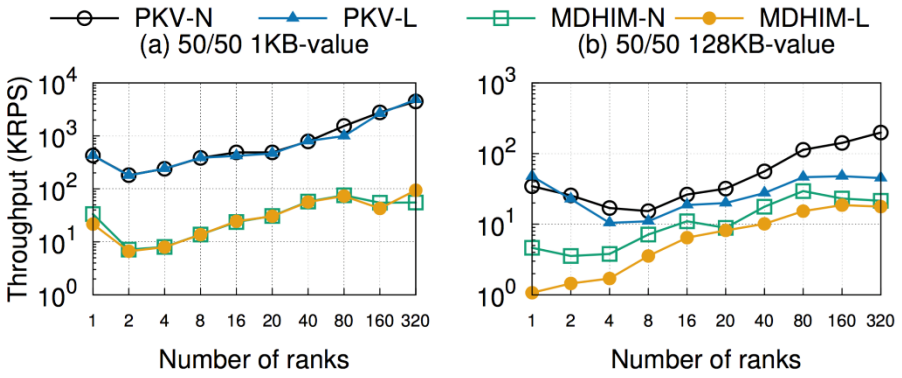
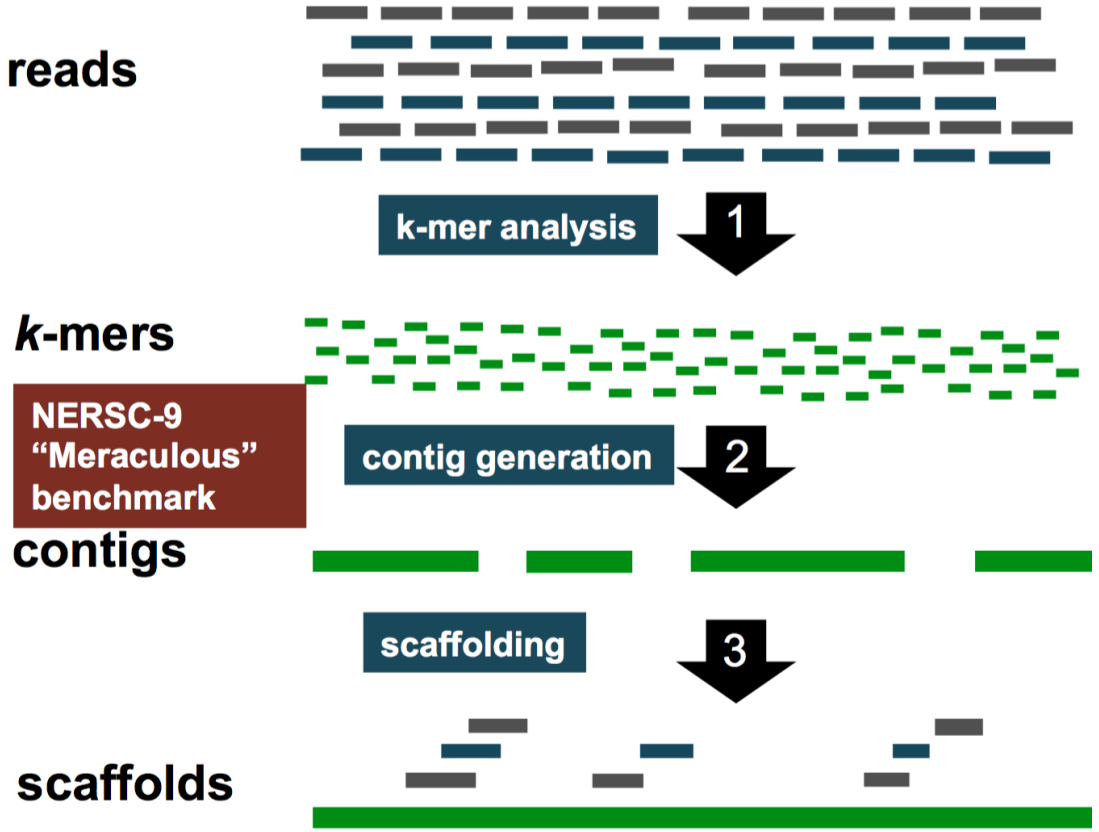


Figure 11: Performance comparisons with MDHIM on Summitdev. NVMe (N) and Lustre (L) are used for their data storages.

ECP Application Case Study 1: Meraculous (UPC)

- A parallel De Bruijin graph construction and traversal for De Novo genome assembly
  - *ExaBiome, Exascale Solutions for Microbiome Analysis, LBNL*



Graphic from ExaBiome: Exascale Solutions to Microbiome Analysis (LBNL, LANL, JGI), 2017

Table 1: Source lines of code.

Source file	UPC	UPC+PapyrusKV
meraculous.c	469	475 (+6)
buildUFXhashBinary.h	315	173 (-143)
kmer_hash.h	457	129 (-328)
UU_traversal_final.h	1754	1724 (-30)
Modified Total	2995	2501 (-494)
Grand Total	5971	5477 (-494)

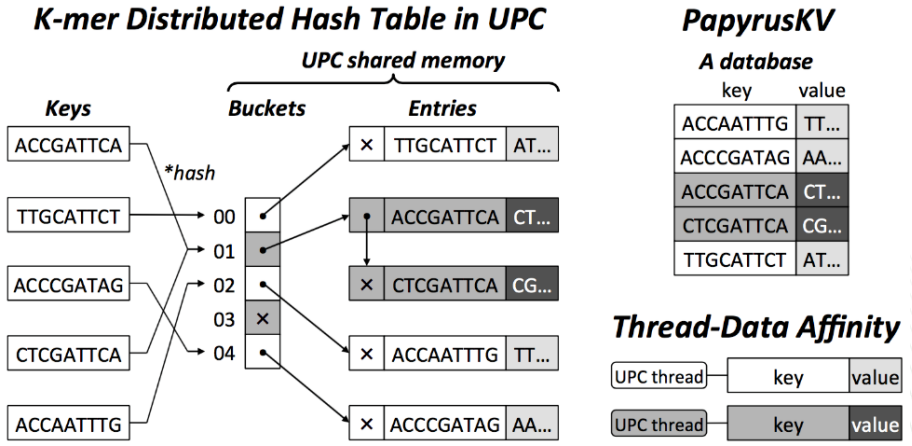
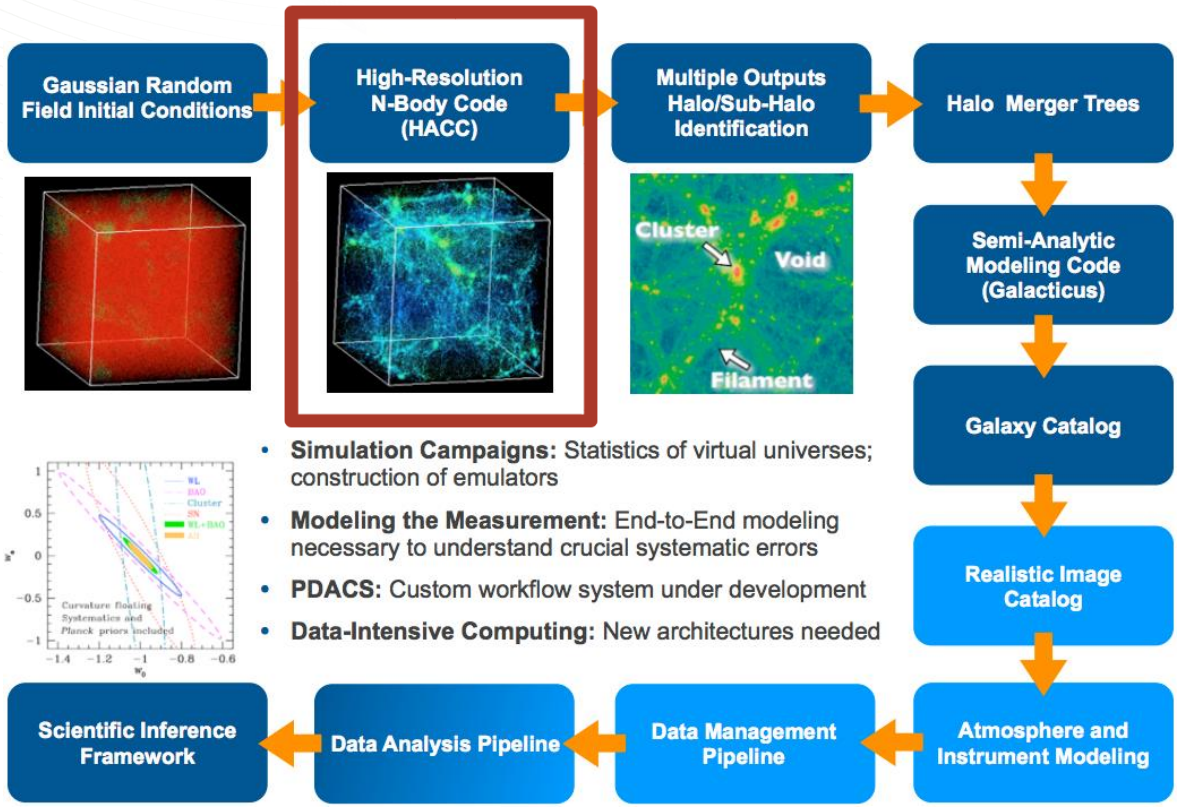


Figure 5: Distributed hash table implementations in UPC and PapyrusKV. *The same user *hash* function in the UPC application can be used in PapyrusKV.

ECP Application Case Study 2: HACC (MPI)

- An N-body cosmology code framework
  - *ExaSky, Computing the Sky at Extreme Scales, ANL*



Graphic from HACCing the Universe on the BG/Q (ANL), 2014

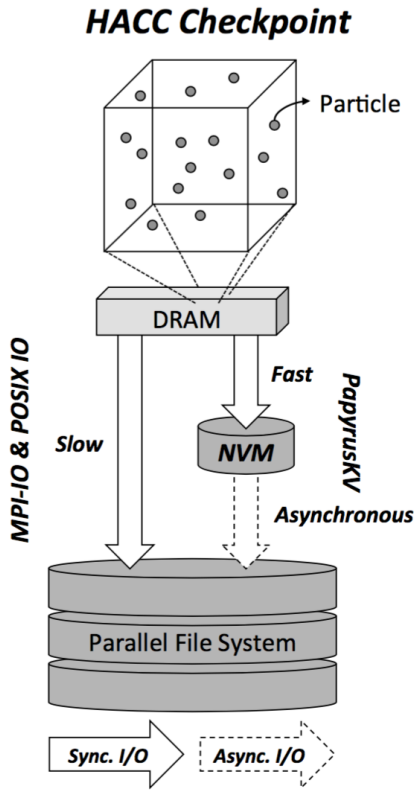


Figure 7: Two-phases checkpointing. PapyrusKV reduces the I/O overhead with help from fast access of NVM. Asynchronous checkpoint hides the I/O overhead between NVM and parallel file system from the application.

Implications for Interconnects and Storage

Predictions for Interconnects and Storage

1. Device and architecture trends will have major impacts on HPC in coming decade
    1. NVM in HPC systems is real!
2. Performance trends of system components will create new opportunities
3. Sea of NVM allows applications to operate differently
    1. Sea of NVM will permit applications to run for weeks without doing I/O to external storage system
    2. Applications will simply access local/remote NVM
    3. Longer term productive I/O will be ‘occasionally’ written to Lustre, GPFS
    4. Checkpointing (as we know it) will disappear
4. Requirements for interconnection networks will change
    1. Increase in byte-addressable memory-like message sizes and frequencies
    2. Reduced traditional IO demands
    3. KV traffic could have considerable impact – need more applications evidence

Summary

- Recent trends in extreme-scale HPC paint an ambiguous future
  - 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
- Memory systems are changing now!
  - New devices
  - New integration
  - New configurations
  - Vast (local) capacities
- Programming systems must provide performance portability (in addition to functional portability)!!
  - We need new programming systems to effectively use these architectures
  - NVL-C
  - Papyrus(KV)
- Changes in memory systems will alter communication and storage requirements dramatically

Acknowledgements

- Contributors and Sponsors
  - Future Technologies Group: http://ft.ornl.gov
  - US Department of Energy Office of Science
    - DOE Vancouver Project: https://ft.ornl.gov/trac/vancouver
    - DOE Blackcomb Project: https://ft.ornl.gov/trac/blackcomb
    - DOE ExMatEx Codesign Center: http://codesign.lanl.gov
    - DOE Cesar Codesign Center: http://cesar.mcs.anl.gov/
    - DOE Exascale Efforts: http://science.energy.gov/ascr/research/computer-science/
  - Scalable Heterogeneous Computing Benchmark team: http://bit.ly/shocmarx
  - US National Science Foundation Keeneland Project: http://keeneland.gatech.edu
  - US DARPA
  - NVIDIA CUDA Center of Excellence

Bonus Material

OAK RIDGE National Laboratory