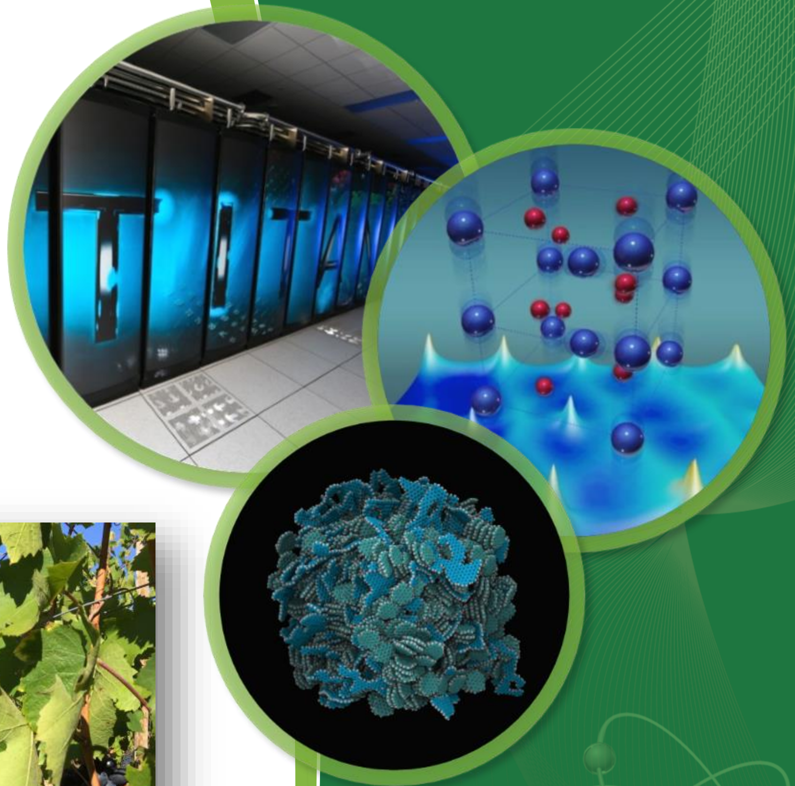
Performance Portability for Extreme Scale High Performance Computing

Jeffrey S. Vetter

and many collaborators



Presented to
Workshop on Clusters, Clouds, and Data for Scientific Computing (CCDSC 2016)

Châteauform'
La Maison des Contes
427 Chemin de Chanzé, France

5 Oct 2016

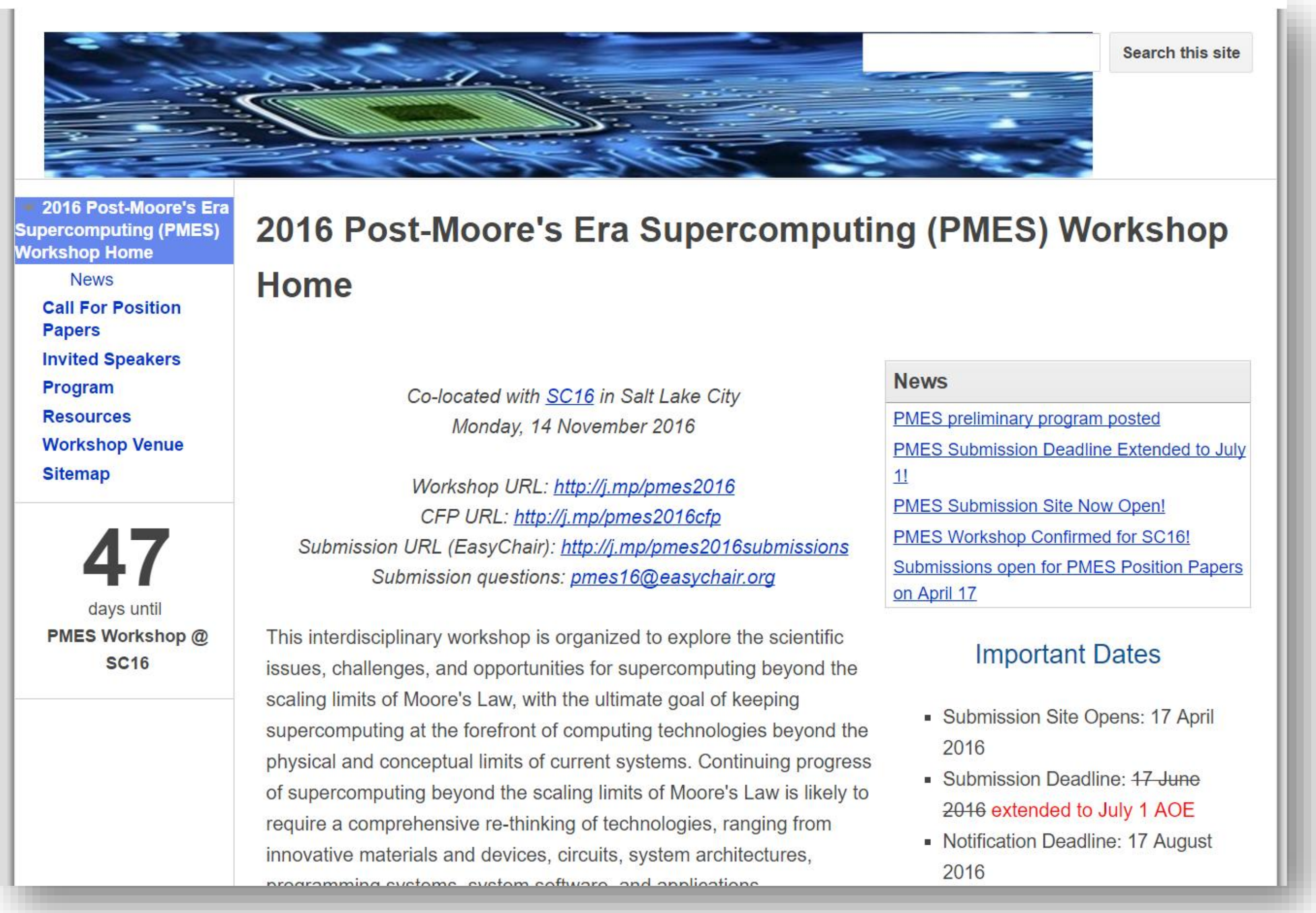
Future Technologies Group

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

OAK RIDGE National Laboratory

https://j.mp/pmes2016

2016 Post-Moores Era Supercomputing Workshop @ SC16 (Nov 14)



2016 Post-Moore's Era Supercomputing (PMES) Workshop Home

News

Call For Position Papers

Invited Speakers

Program

Resources

Workshop Venue

Sitemap

47

days until

PMES Workshop @ SC16

2016 Post-Moore's Era Supercomputing (PMES) Workshop Home

Co-located with SC16 in Salt Lake City
Monday, 14 November 2016

Workshop URL: http://j.mp/pmes2016
CFP URL: http://j.mp/pmes2016cfp
Submission URL (EasyChair): http://j.mp/pmes2016submissions
Submission questions: pmes16@easychair.org

This interdisciplinary workshop is organized to explore the scientific issues, challenges, and opportunities for supercomputing beyond the scaling limits of Moore's Law, with the ultimate goal of keeping supercomputing at the forefront of computing technologies beyond the physical and conceptual limits of current systems. Continuing progress of supercomputing beyond the scaling limits of Moore's Law is likely to require a comprehensive re-thinking of technologies, ranging from innovative materials and devices, circuits, system architectures, programming systems, system software, and applications

News

PMES preliminary program posted

PMES Submission Deadline Extended to July 1!

PMES Submission Site Now Open!

PMES Workshop Confirmed for SC16!

Submissions open for PMES Position Papers on April 17

Important Dates

- Submission Site Opens: 17 April 2016
- Submission Deadline: ~~17 June 2016~~ extended to July 1 AOE
- Notification Deadline: 17 August 2016

- Accepted papers include
  - Quantum computing (Dwave)
  - Neuromorphic computing
  - Probabilistic
  - Approximate computing, numerics
  - Reconfigurable
  - Photonics
  - Software
  - Performance modeling

Overview

- 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 performance portable programming models now more than ever!
- Programming systems must provide performance portability (in addition to functional portability)!!
  - New memory hierarchies with NVM everywhere
  - Heterogeneous systems

Trends toward Exascale

Exascale architecture targets circa 2009

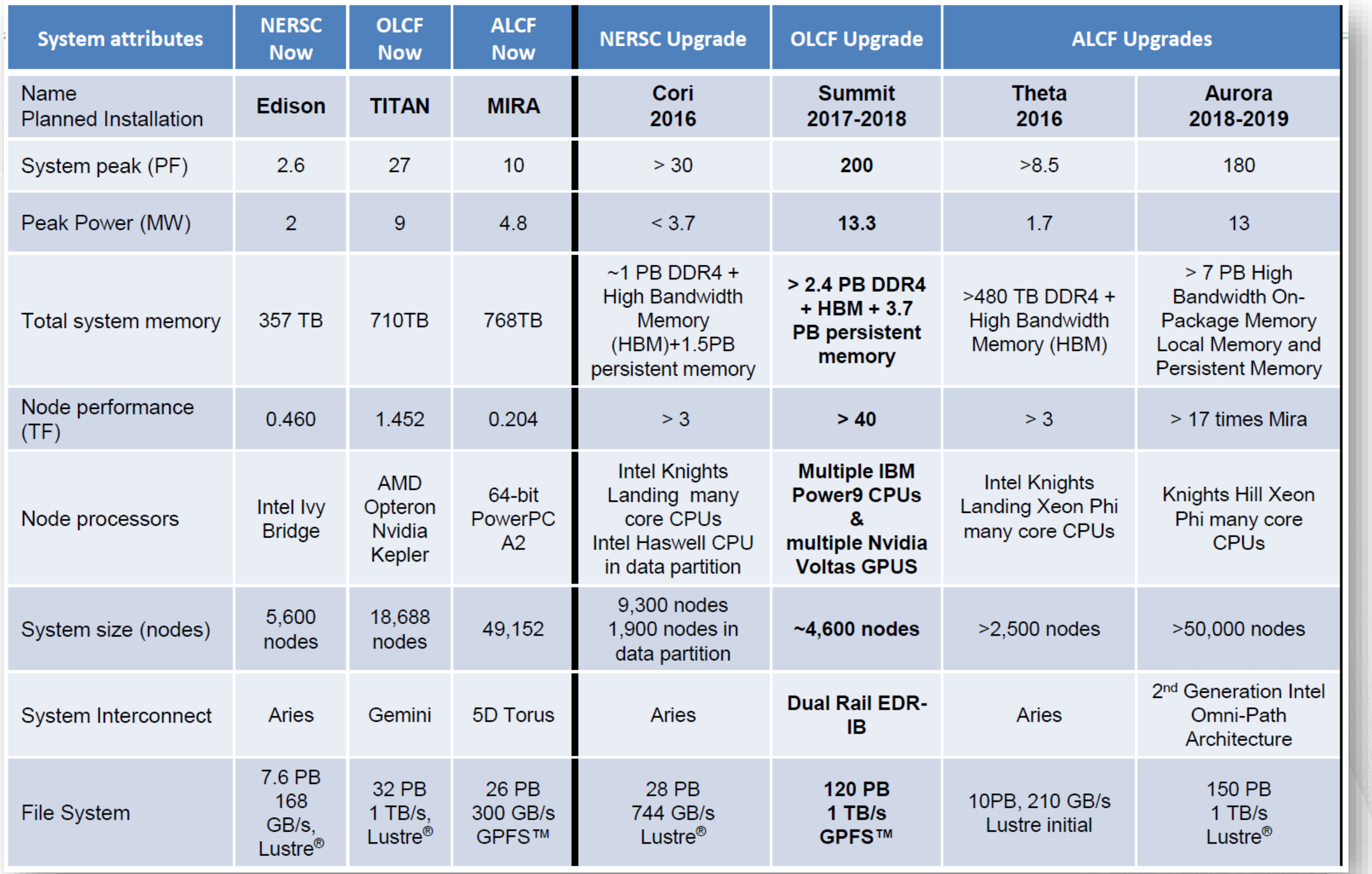
2009 Exascale Challenges Workshop in San Diego

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)	

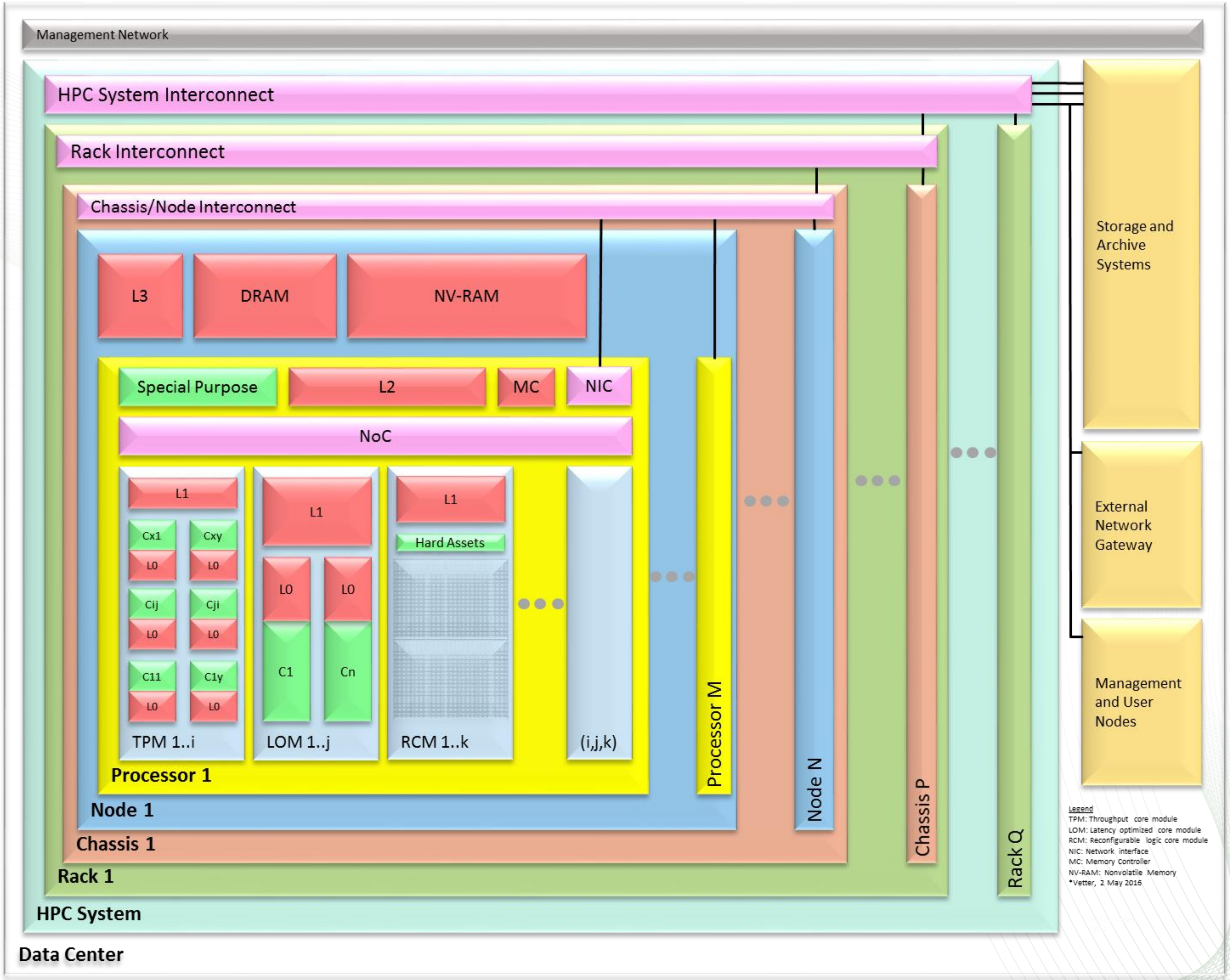
Contemporary ASCR Computing At a Glance



System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrades	
Name Planned Installation	**Edison**	**TITAN**	**MIRA**	**Cori** **2016**	**Summit** **2017-2018**	**Theta** **2016**	**Aurora** **2018-2019**
System peak (PF)	2.6	27	10	> 30	**200**	>8.5	180
Peak Power (MW)	2	9	4.8	< 3.7	**13.3**	1.7	13
Total system memory	357 TB	710TB	768TB	~1 PB DDR4 + High Bandwidth Memory (HBM)+1.5PB persistent memory	**> 2.4 PB DDR4 + HBM + 3.7 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	**~4,600 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®

Binkley, ASCAC, April 2016

Complexity α T



Management Network

HPC System Interconnect

Rack Interconnect

Chassis/Node Interconnect

L3

DRAM

NV-RAM

Special Purpose

L2

MC

NIC

NoC

L1

Cx1

Cxy

L0

L0

Cij

Cji

L0

L0

C11

C1y

L0

L0

TPM 1..i

L1

L0

L0

C1

Cn

LOM 1..j

L1

Hard Assets

RCM 1..k

(i,j,k)

Processor 1

Processor M

Node 1

Node N

Chassis 1

Chassis P

Rack 1

Rack Q

HPC System

Data Center

Storage and Archive Systems

External Network Gateway

Management and User Nodes

Legend

TPM: Throughput core module

LOM: Latency optimized core module

RCM: Reconfigurable logic core module

NIC: Network interface

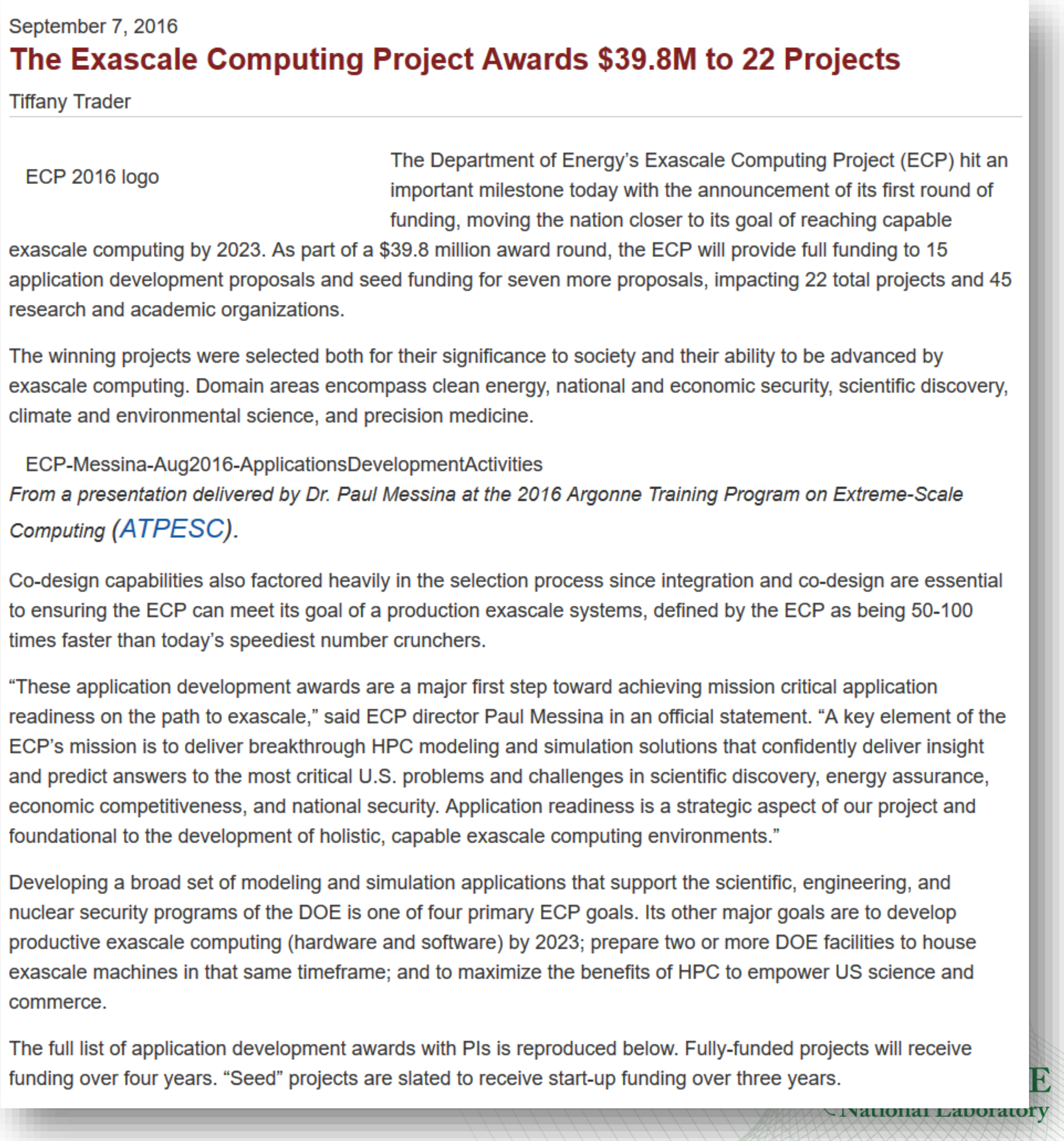
MC: Memory Controller

NV-RAM: Nonvolatile Memory

*Vetter, 2 May 2016

Complexity is the next major challenge!

- Time of rapid change in computer architectures
  - Heterogeneous cores
  - Deep, multimode memory systems
  - I/O architectures
  - Reliability
  - Changing system balance
- Uncertainty, Ambiguity among current and future architectures
  - Managing complexity is our main challenge!
    - Complex systems → Fewer apps → Smaller HPC
- Critical questions
  - How do we design future systems so that they are faster than current systems on mission applications?
    - Entirely possible that the new system will be slower than the old system!
  - How do we design applications for some level of performance portability?



September 7, 2016

The Exascale Computing Project Awards $39.8M to 22 Projects

Tiffany Trader

ECP 2016 logo

The Department of Energy's Exascale Computing Project (ECP) hit an important milestone today with the announcement of its first round of funding, moving the nation closer to its goal of reaching capable exascale computing by 2023. As part of a $39.8 million award round, the ECP will provide full funding to 15 application development proposals and seed funding for seven more proposals, impacting 22 total projects and 45 research and academic organizations.

The winning projects were selected both for their significance to society and their ability to be advanced by exascale computing. Domain areas encompass clean energy, national and economic security, scientific discovery, climate and environmental science, and precision medicine.

ECP-Messina-Aug2016-ApplicationsDevelopmentActivities

From a presentation delivered by Dr. Paul Messina at the 2016 Argonne Training Program on Extreme-Scale Computing (ATPESC).

Co-design capabilities also factored heavily in the selection process since integration and co-design are essential to ensuring the ECP can meet its goal of a production exascale systems, defined by the ECP as being 50-100 times faster than today's speediest number crunchers.

"These application development awards are a major first step toward achieving mission critical application readiness on the path to exascale," said ECP director Paul Messina in an official statement. "A key element of the ECP's mission is to deliver breakthrough HPC modeling and simulation solutions that confidently deliver insight and predict answers to the most critical U.S. problems and challenges in scientific discovery, energy assurance, economic competitiveness, and national security. Application readiness is a strategic aspect of our project and foundational to the development of holistic, capable exascale computing environments."

Developing a broad set of modeling and simulation applications that support the scientific, engineering, and nuclear security programs of the DOE is one of four primary ECP goals. Its other major goals are to develop productive exascale computing (hardware and software) by 2023; prepare two or more DOE facilities to house exascale machines in that same timeframe; and to maximize the benefits of HPC to empower US science and commerce.

The full list of application development awards with PIs is reproduced below. Fully-funded projects will receive funding over four years. "Seed" projects are slated to receive start-up funding over three years.

Performance Portability : what is it?

- Effectively from application perspective, “write once, run anywhere efficiently”
- Performance portability is not a new topic
  - Kuck, 1996
- For two decades, expectations were set by ‘(Curse of) Moore’s Law’ with exception for MPI for scaling parallelism
  - Recompile and relink
- More important then ever
  - Becoming difficult to hide complexity for even functional portability
- Efficiently use resource of interest

		Application				
		A	B	C	D	
Architecture	W					?
	X					?
	Y					?
	Z					?
		?	?	?	?	

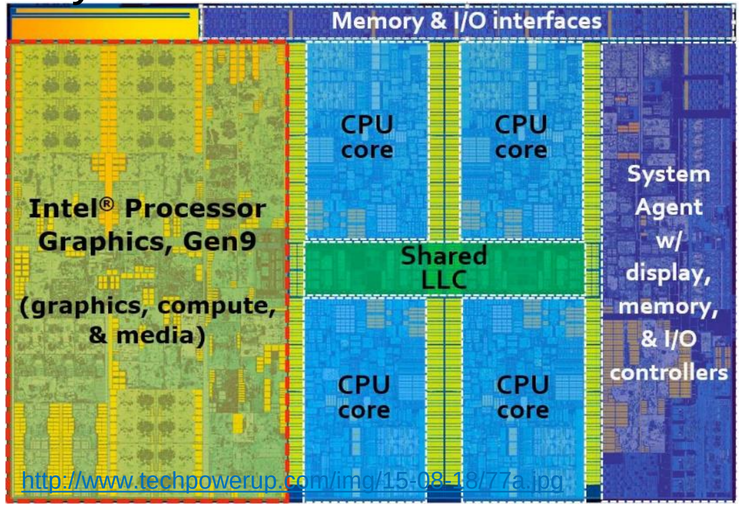
D.J. Kuck, *High performance computing: challenges for future systems. New York: Oxford University Press, 1996.*

Motivating Heterogeneous Systems

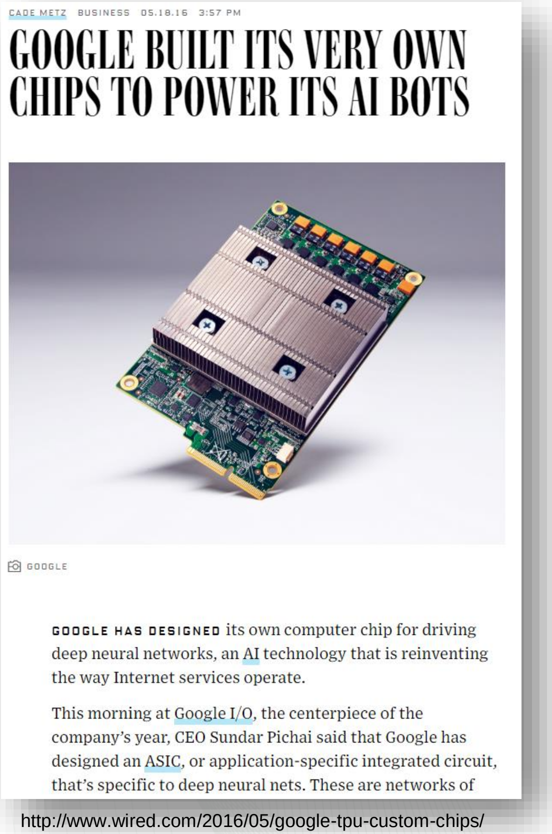
Specialization is here to stay: Core, Processor Architectures

- Cores
  - CPU
  - GPUs (discrete, integrated)
  - FPGAs
  - Special purpose engines
    - RNGs
    - AES, video engines
    - Transactional memory
    - Virtualization support
- SIMD/short vector
- SMT, threading models
- DVFS (incl Turboboost)
- etc

Skylake


Memory & I/O interfaces
Intel® Processor Graphics, Gen9
(graphics, compute, & media)
CPU core
CPU core
Shared LLC
CPU core
CPU core
System Agent w/ display, memory, & I/O controllers

http://www.techpowerup.com/img/15-08-18/77a.jpg



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

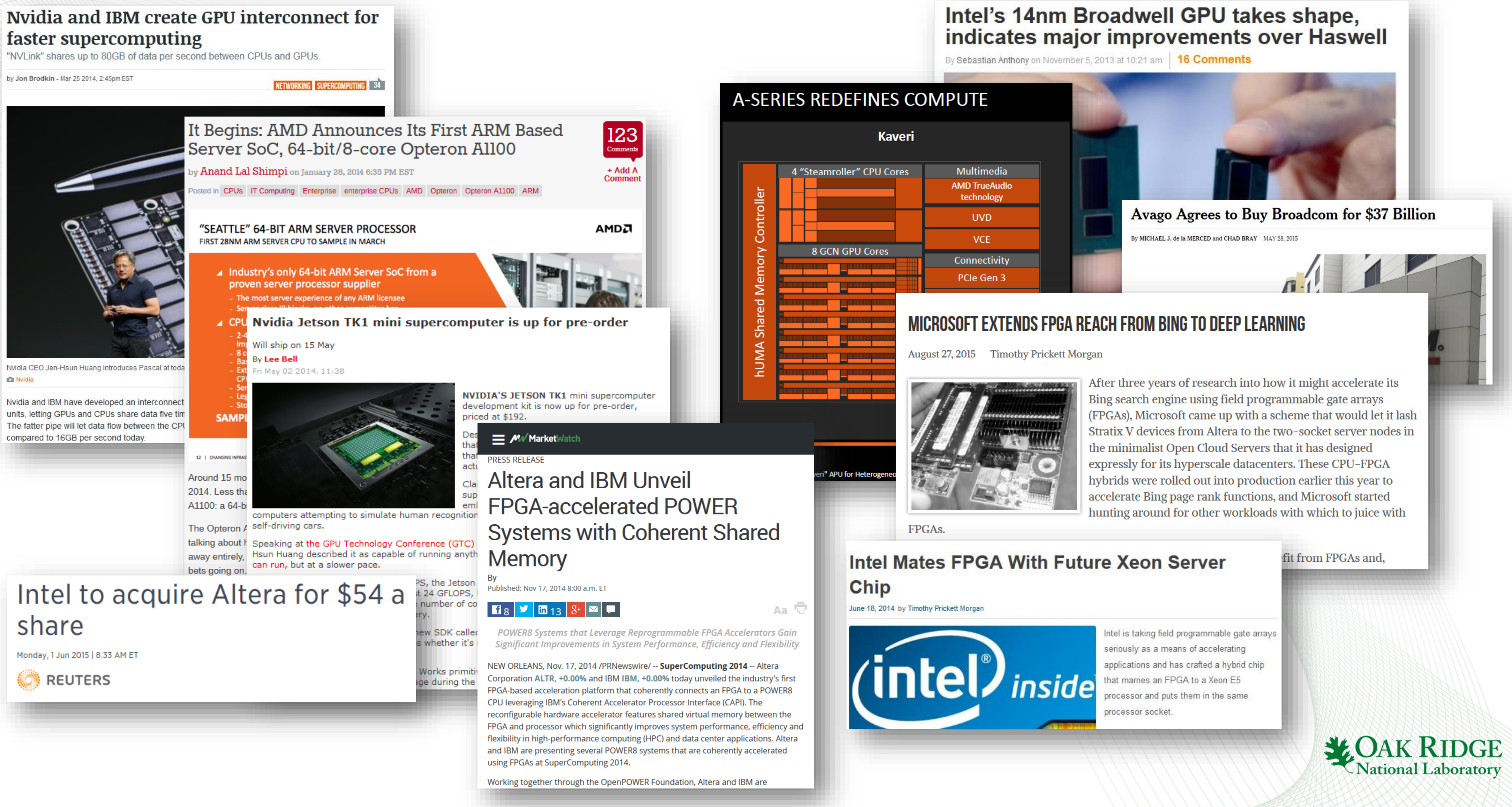
http://www.wired.com/2016/05/google-tpu-custom-chips/

Pascal

http://cdn.wccftech.com/wp-content/uploads/2014/03/NVIDIA-Pascal-GPU-Chip-Module.jpg

D.E. Shaw, M.M. Deneroff, R.O. Dror *et al., "Anton, a special-purpose machine for molecular dynamics*

In the news



Nvidia and IBM create GPU interconnect for faster supercomputing

"NVLink" shares up to 80GB of data per second between CPUs and GPUs.

by Jon Brodkin - Mar 25 2014, 2:45pm EST

NETWORKING SUPERCOMPUTING 34

Nvidia CEO Jen-Hsun Huang introduces Pascal at toda

Nvidia

Nvidia and IBM have developed an interconnect units, letting GPUs and CPUs share data five tim The fatter pipe will let data flow between the CPU compared to 16GB per second today.

It Begins: AMD Announces Its First ARM Based Server SoC, 64-bit/8-core Opteron A1100

123 Comments

+ Add A Comment

by Anand Lal Shimpi on January 28, 2014 6:35 PM EST

Posted in CPUs IT Computing Enterprise enterprise CPUs AMD Opteron Opteron A1100 ARM

"SEATTLE" 64-BIT ARM SERVER PROCESSOR

FIRST 28NM ARM SERVER CPU TO SAMPLE IN MARCH

AMD

- Industry's only 64-bit ARM Server SoC from a proven server processor supplier
  - The most server experience of any ARM licensee
  - Ser
- CPU
  - 2-4 imp
  - 8 c
  - Bas
  - Ext CPU
  - Ser
  - Leg
  - Sto

SAMPL

12 | CHANGING INFRAS

Around 15 mo 2014. Less tha A1100: a 64-bi

The Opteron A talking about h away entirely, bets going on.

Nvidia Jetson TK1 mini supercomputer is up for pre-order

Will ship on 15 May

By Lee Bell

Fri May 02 2014, 11:38

NVIDIA'S JETSON TK1 mini supercomputer development kit is now up for pre-order, priced at $192.

Des that that actu

Cla sup eml computers attempting to simulate human recognition self-driving cars.

Speaking at the GPU Technology Conference (GTC) Hsun Huang described it as capable of running anyth can run, but at a slower pace.

PS, the Jetson t 24 GFLOPS, number of co ry.

ew SDK called s whether it's

Works primiti ge during the

Intel to acquire Altera for $54 a share

Monday, 1 Jun 2015 | 8:33 AM ET

REUTERS

MarketWatch

PRESS RELEASE

Altera and IBM Unveil FPGA-accelerated POWER Systems with Coherent Shared Memory

By

Published: Nov 17, 2014 8:00 a.m. ET

8 13 Aa

POWER8 Systems that Leverage Reprogrammable FPGA Accelerators Gain Significant Improvements in System Performance, Efficiency and Flexibility

NEW ORLEANS, Nov. 17, 2014 /PRNewswire/ -- **SuperComputing 2014** -- Altera Corporation ALTR, +0.00% and IBM IBM, +0.00% today unveiled the industry's first FPGA-based acceleration platform that coherently connects an FPGA to a POWER8 CPU leveraging IBM's Coherent Accelerator Processor Interface (CAPI). The reconfigurable hardware accelerator features shared virtual memory between the FPGA and processor which significantly improves system performance, efficiency and flexibility in high-performance computing (HPC) and data center applications. Altera and IBM are presenting several POWER8 systems that are coherently accelerated using FPGAs at SuperComputing 2014.

Working together through the OpenPOWER Foundation, Altera and IBM are

A-SERIES REDEFINES COMPUTE

Kaveri

hUMA Shared Memory Controller

4 "Steamroller" CPU Cores

8 GCN GPU Cores

Multimedia

AMD TrueAudio technology

UVD

VCE

Connectivity

PCIe Gen 3

veri" APU for Heterogeneo

Intel's 14nm Broadwell GPU takes shape, indicates major improvements over Haswell

By Sebastian Anthony on November 5, 2013 at 10:21 am | 16 Comments

Avago Agrees to Buy Broadcom for $37 Billion

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

MICROSOFT EXTENDS FPGA REACH FROM BING TO DEEP LEARNING

August 27, 2015 Timothy Prickett Morgan

After three years of research into how it might accelerate its Bing search engine using field programmable gate arrays (FPGAs), Microsoft came up with a scheme that would let it lash Stratix V devices from Altera to the two-socket server nodes in the minimalist Open Cloud Servers that it has designed expressly for its hyperscale datacenters. These CPU-FPGA hybrids were rolled out into production earlier this year to accelerate Bing page rank functions, and Microsoft started hunting around for other workloads with which to juice with FPGAs.

fit from FPGAs and,

Intel Mates FPGA With Future Xeon Server Chip

June 18, 2014 by Timothy Prickett Morgan

Intel is taking field programmable gate arrays seriously as a means of accelerating applications and has crafted a hybrid chip that marries an FPGA to a Xeon E5 processor and puts them in the same processor socket.

Current ASCR Computing At a Glance

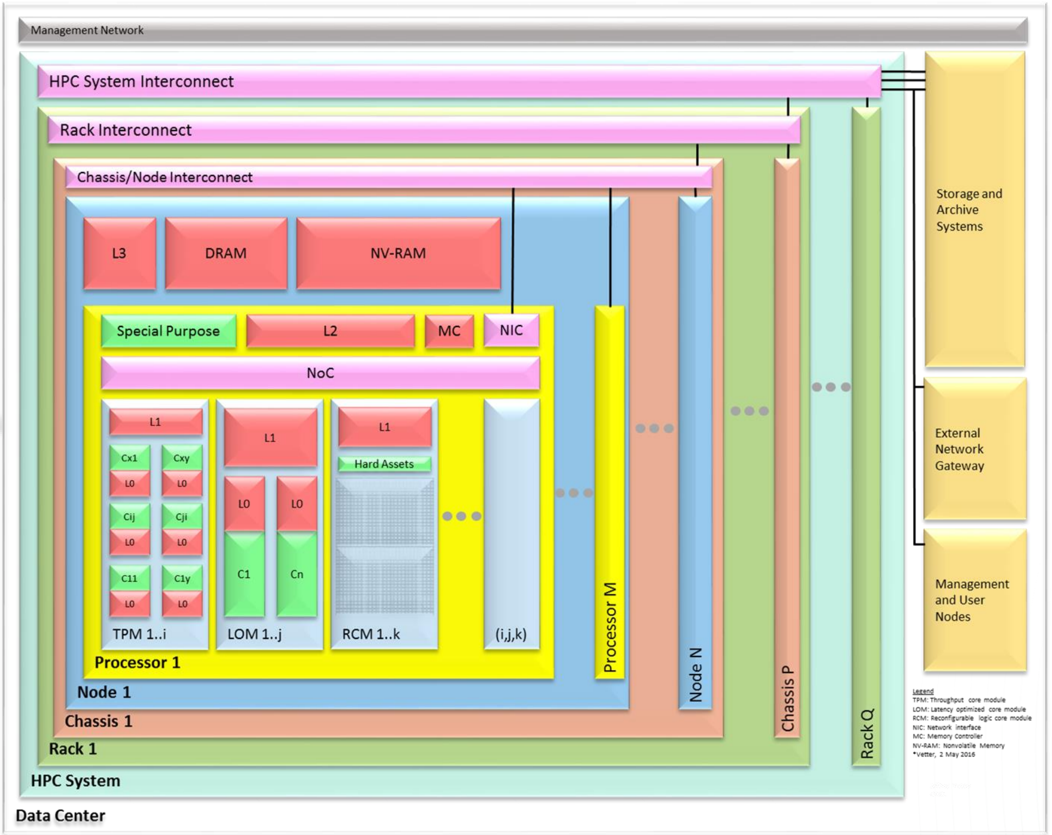
System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrades	
Planned Installation	**Edison**	**TITAN**	**MIRA**	**Cori** **2016**	**Summit** **2017-2018**	**Theta** **2016**	**Aurora** **2018-2019**
System peak (PF)	2.6	27	10	> 30	150	>8.5	180
Peak Power (MW)	2	9	4.8	< 3.7	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 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®

Binkley, ASCAC, April 2016

Complexity α T

Programming Heterogeneous Systems

…Yields Complex Programming Models



Legend
TPM: Throughput core module
LOM: Latency optimized core module
RCM: Reconfigurable logic core module
NIC: Network interface
MC: Memory Controller
NV-RAM: Nonvolatile Memory
*Vetter, 2 May 2016

- This approach is not scalable, affordable, robust, elegant, etc.
- Not performance portable

System: MPI, Legion, HPX, Charm++, etc

Low overhead

Resource contention

Locality

Node: OpenMP, Pthreads, U-threads, etc

SIMD

NUMA, HBM

Cores: OpenACC, CUDA, OpenCL, OpenMP4, …

Memory use, coalescing

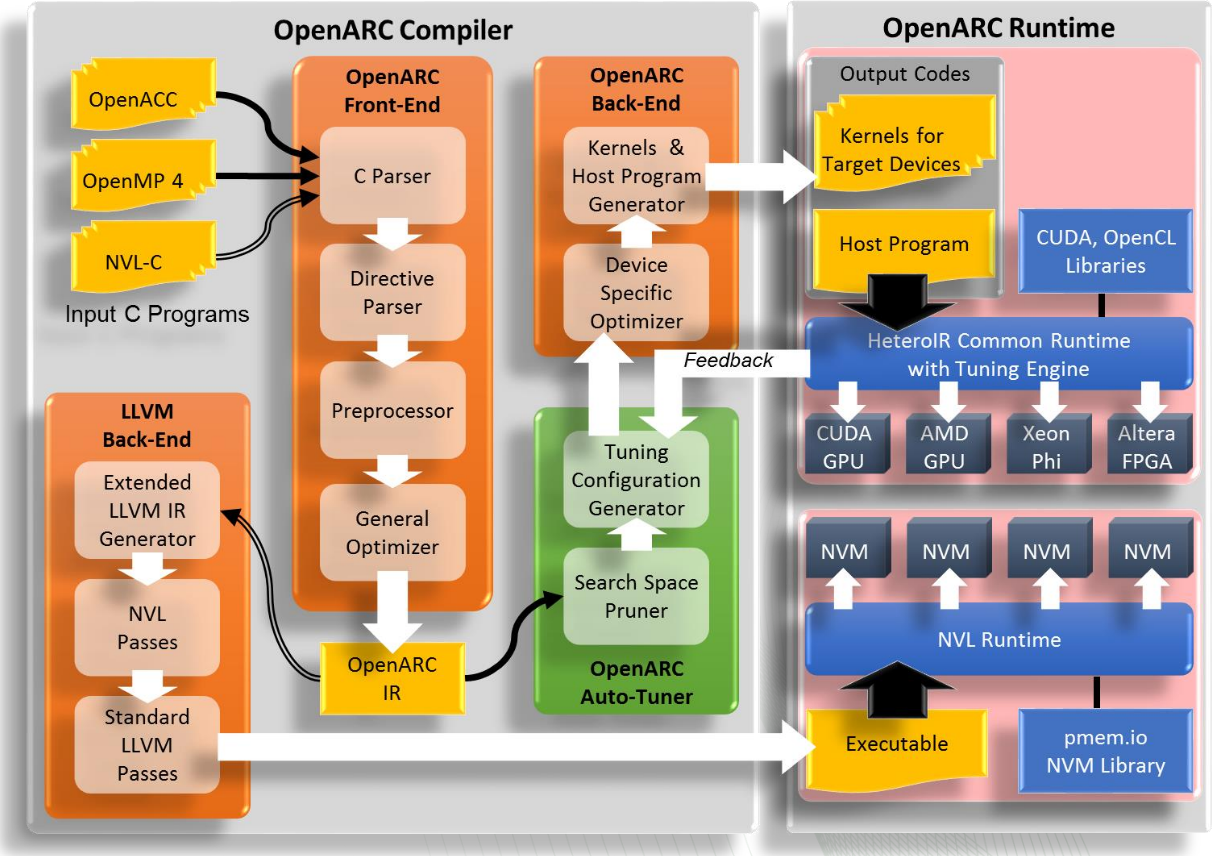
Data orchestration

Fine grained parallelism

Hardware features

OpenARC: Open Accelerator Research Compiler

- Open-Sourced, High-Level Intermediate Representation (HIR)-Based, Extensible Compiler Framework.
  - Perform source-to-source translation from OpenACC C to target accelerator models.
    - Support full features of OpenACC V1.0 (+ array reductions and function calls)
    - Support both CUDA and OpenCL as target accelerator models
  - Provide common runtime APIs for various back-ends
  - Can be used as a research framework for various study on directive-based accelerator computing.
    - Built on top of Cetus compiler framework, equipped with various advanced analysis/transformation passes and built-in tuning tools.
    - OpenARC's IR provides an AST-like syntactic view of the source program, easy to understand, access, and transform the input program.

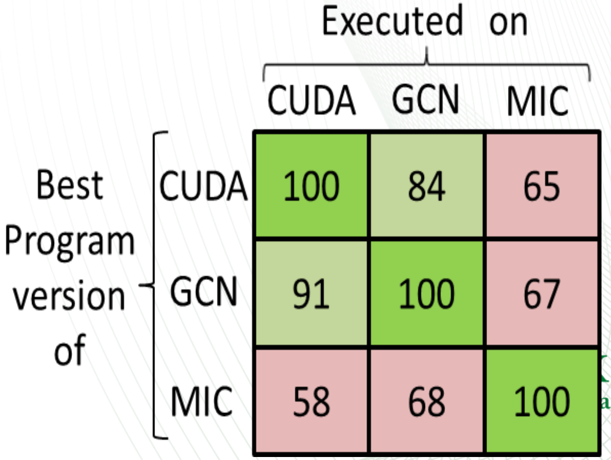


S. Lee and J.S. Vetter, "OpenARC: Open Accelerator Research Compiler for Directive-Based, Efficient Heterogeneous Computing," in *ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC). Vancouver: ACM, 2014*

Understanding Performance Portability of High-level Programming Models for Heterogeneous Systems

- Problem
  - Directive-based, high-level accelerator programming models such as OpenACC provide code portability.
    - How does it fare on performance portability?
    - And what architectural features/compiler optimizations affect the performance portability? And how much?
- Solution
  - Proposed a high-level, architecture-independent intermediate language (HeteroIR) to map high-level programming models (e.g., OpenACC) to diverse heterogeneous devices while maintaining portability.
  - Using HeteroIR, port and measure the performance portability of various OpenACC applications on diverse architectures.

- Results
  - Using HeteroIR, OpenARC ported 12 OpenACC applications to diverse architectures (NVIDIA CUDA, AMD GCN, and Intel MIC), and measured the performance portability achieved across all applications.
  - HeteroIR abstracts out the common architecture functionalities, which makes it easy for OpenARC (and other compilers) to support diverse heterogeneous architectures.
  - HeteroIR, combined with rich OpenARC directives and built-in tuning tools, allows OpenARC to be used for various tuning studies on diverse architectures.



		Executed on		
		CUDA	GCN	MIC
Best Program version of	CUDA	100	84	65
	GCN	91	100	67
	MIC	58	68	100

Amit Sabne, Putt Sakdhnagool, Seyong Lee, and Jeffrey S. Vetter. Understanding Portability of a High-level Programming Model on Contemporary Heterogeneous Architectures, IEEE Micro Volume 35, Issue 4 (DOI: 10.1109/MM.2015.73), 2015.

Intelligent selection of optimizations based on target architecture

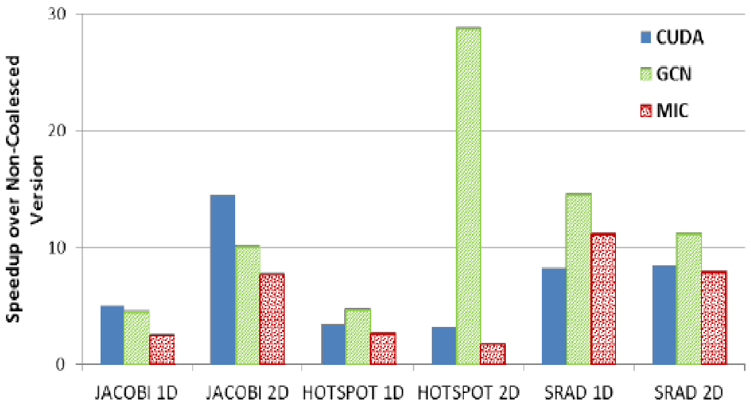


Figure 5: Memory Coalescing Benefits on Different Architectures : MIC is impacted the least by the non-coalesced accesses

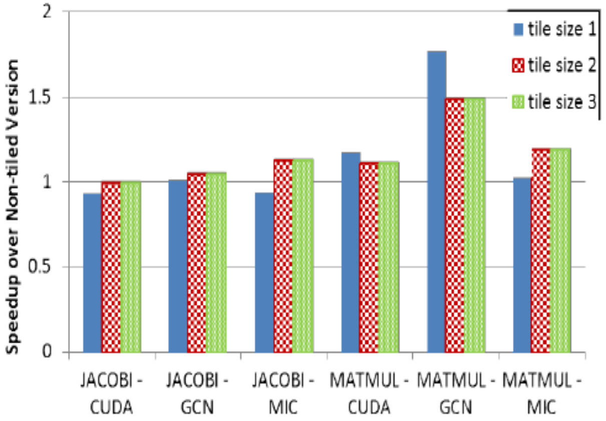


Figure 7: Impact of Tiling Transformation : *MATMUL* shows higher benefits than *JACOBI* owing to more contiguous accesses

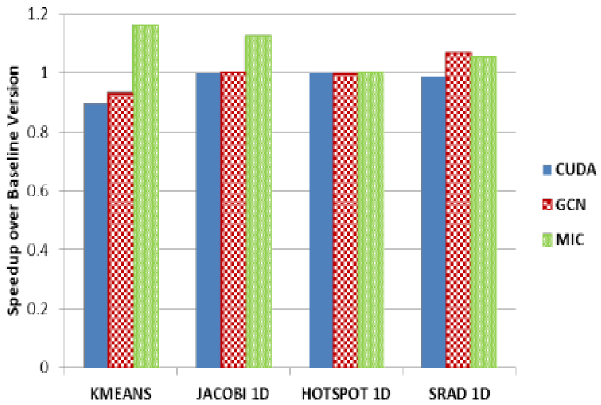


Figure 9: Effects of Loop Unrolling - MIC shows benefits on unrolling

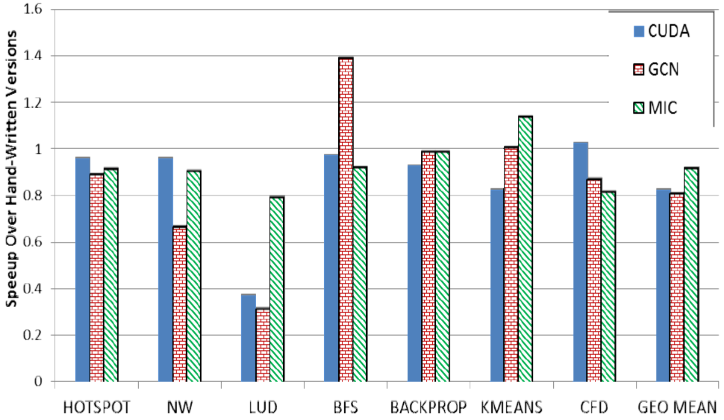


Fig. 11: Comparison of hand-written CUDA/OpenCL programs against auto-tuned OpenARC code versions : Tuned OpenACC programs perform reasonably well against hand-written codes

OpenACC to FPGA: A Framework for Directive-Based High-Performance Reconfigurable Computing

- OpenACC-to-FPGA translation framework
  - source-to-source translation of the input OpenACC program into an output OpenCL code,
  - further compiled to an FPGA program by the underlying backend Altera OpenCL compiler.
  - Prototyped new OpenACC directives to support pipelining of kernels
- Recent Results
  - Proposed several FPGA-specific OpenACC compiler optimizations and pragma extensions to achieve higher throughput.
  - Evaluated the framework using eight OpenACC benchmarks, and measured performance variations on diverse architectures (Altera FPGA, NVIDIA/AMD GPUs, and Intel Xeon Phi).

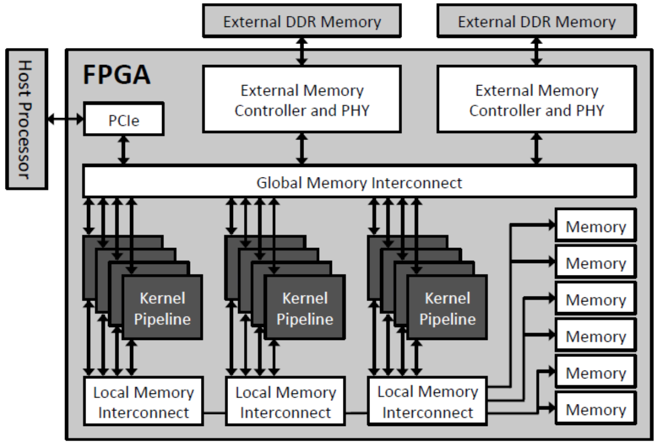
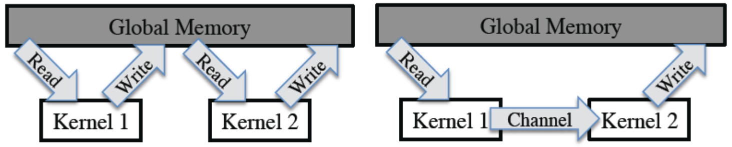
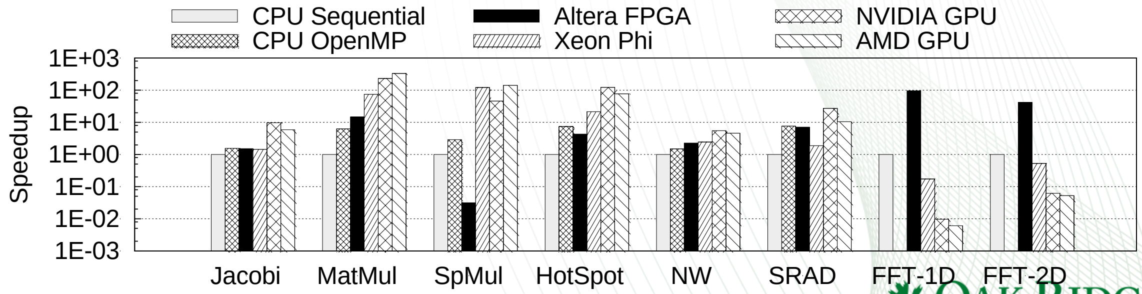

External DDR Memory
External DDR Memory
Host Processor
FPGA
PCIe
External Memory Controller and PHY
External Memory Controller and PHY
Global Memory Interconnect
Kernel Pipeline
Kernel Pipeline
Kernel Pipeline
Memory
Memory
Memory
Memory
Memory
Memory
Local Memory Interconnect
Local Memory Interconnect
Local Memory Interconnect

Figure 2: FPGA OpenCL Architecture



(a) Global Memory Access Without Channels (b) Global Memory Access With Channels

Figure 3: Difference in Global Memory Access Pattern as a Result of Channels Implementation



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.

Emerging Non-volatile Memory Systems

Exascale architecture targets circa 2009

2009 Exascale Challenges Workshop in San Diego

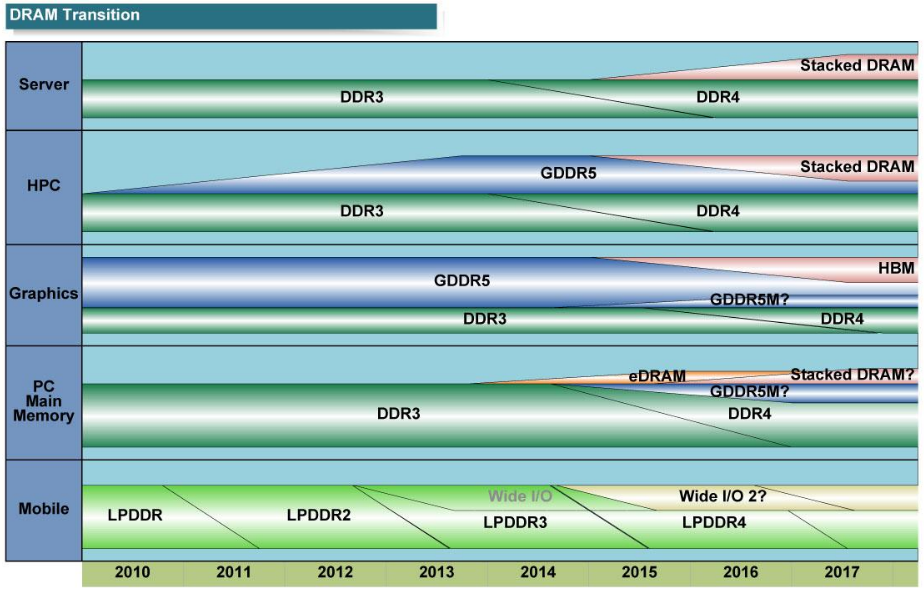
Attendees envisioned two possible architectural swim lanes:

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

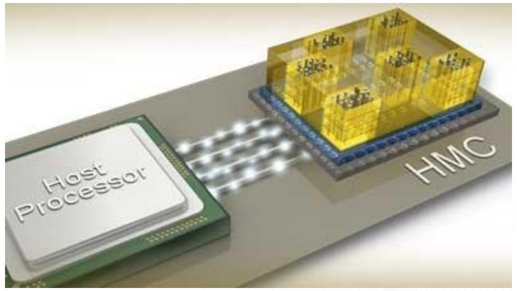
System attributes	2009	"Pre-Exascale"		"Exascale"	
System peak	2 PF	100-200 PF/s		1 Exaflop/s	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32–64 PB	
Storage	15 PB	150 PB		500 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/s	1 TB/s	0.4 TB/s	4 TB/s
Node concurrency	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)	

Memory Systems are Diversifying

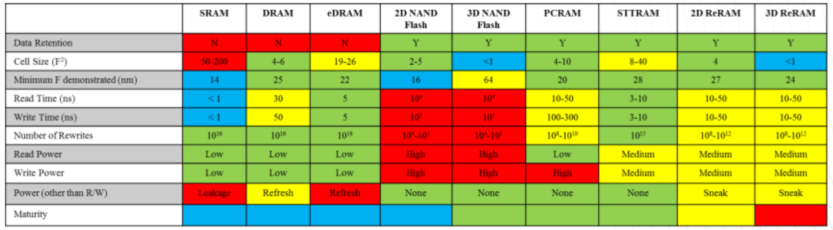
- HMC, HBM/2/3, LPDDR4, GDDR5X, WIDEIO2, etc
- Configuration diversity
  – Fused, shared memory
  – Scratchpads
  – Write through, write back, etc
  – Consistency and coherence protocols
  – Virtual v. Physical, paging strategies
- 2.5D, 3D Stacking
- New devices (ReRAM, PCRAM, STT-MRAM, Xpoint)



Copyright (c) 2014 Hiroshige Goto All rights reserved. http://gigglehd.com/zbxe/files/attach/images/1404665/988/406/011/788d3ba1967e2db3817d259



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



	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	10^4	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	🟦	🟦	🟦	🟦	🟩	🟩	🟩	🟨	🟥

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

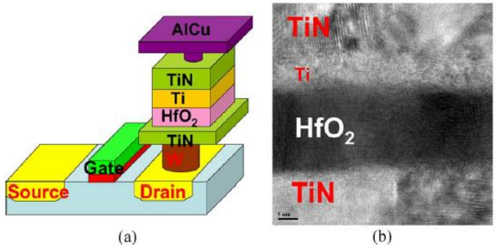
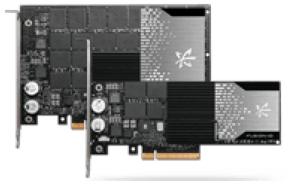
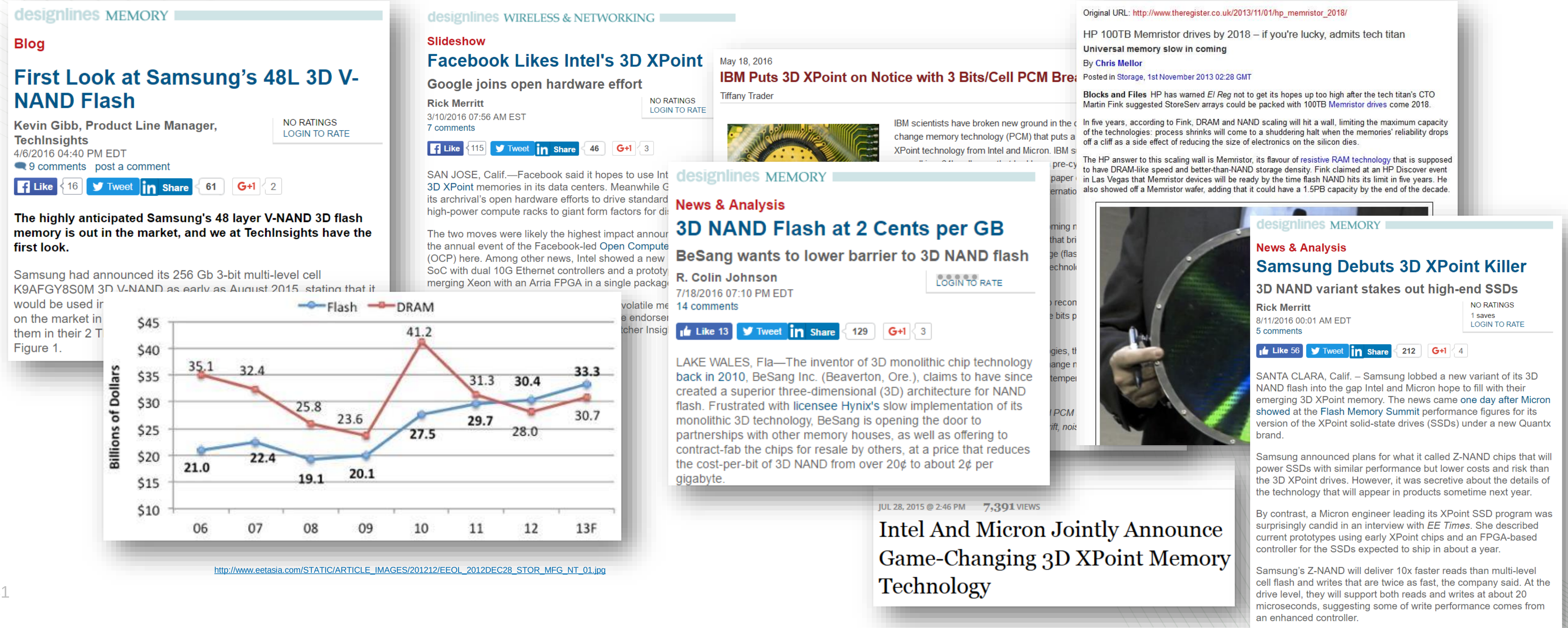


Fig. 4. *(a) A typical 1T1R structure of RRAM with HfO_x; (b) HR-TEM image of the TiN/Ti/HfO_x/TiN stacked layer; the thickness of the HfO_2 is 20 nm.*

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

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
9 comments post a comment

NO RATINGS
LOGIN TO RATE

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 ir
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
volatile me
e endorser
tcher Insig

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 c
change memory technology (PCM) that puts a
XPoint technology from Intel and Micron. IBM s
pre-cy
paper
ernatio
ming n
that bri
ge (flas
echnol
recon
e bits p
gies, th
ange m
temper
PCM
ift, nois

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

Current ASCR Computing At a Glance

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 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®

Binkley, ASCAC, April 2016

Complexity α T

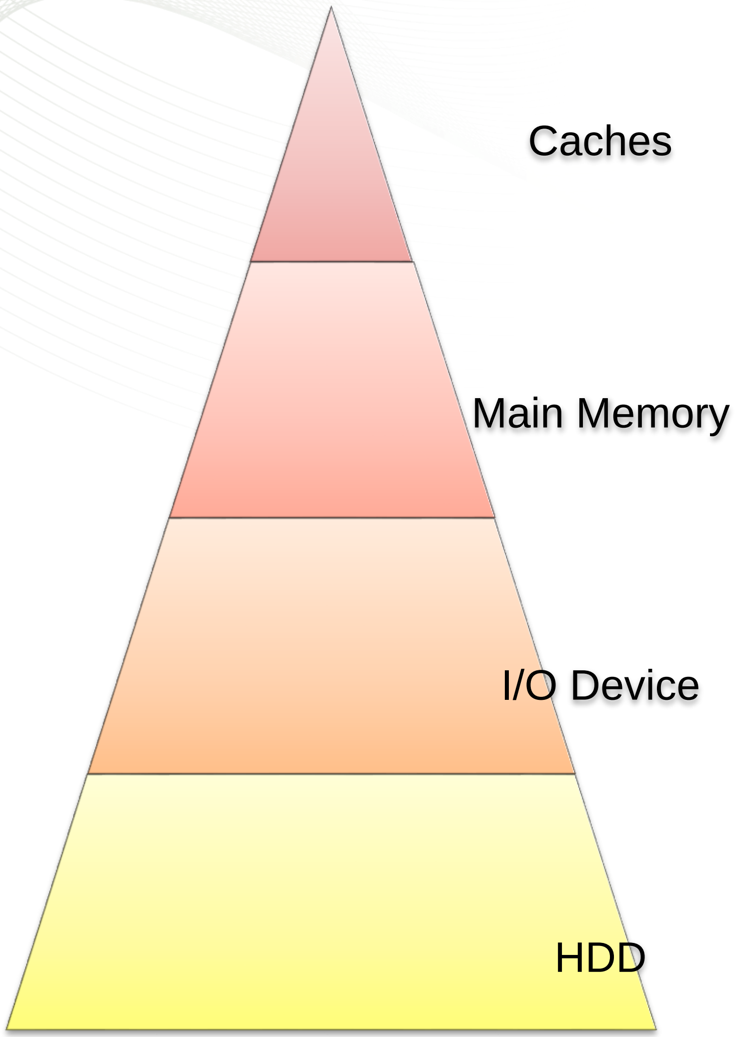
Comparison of Emerging Memory Technologies

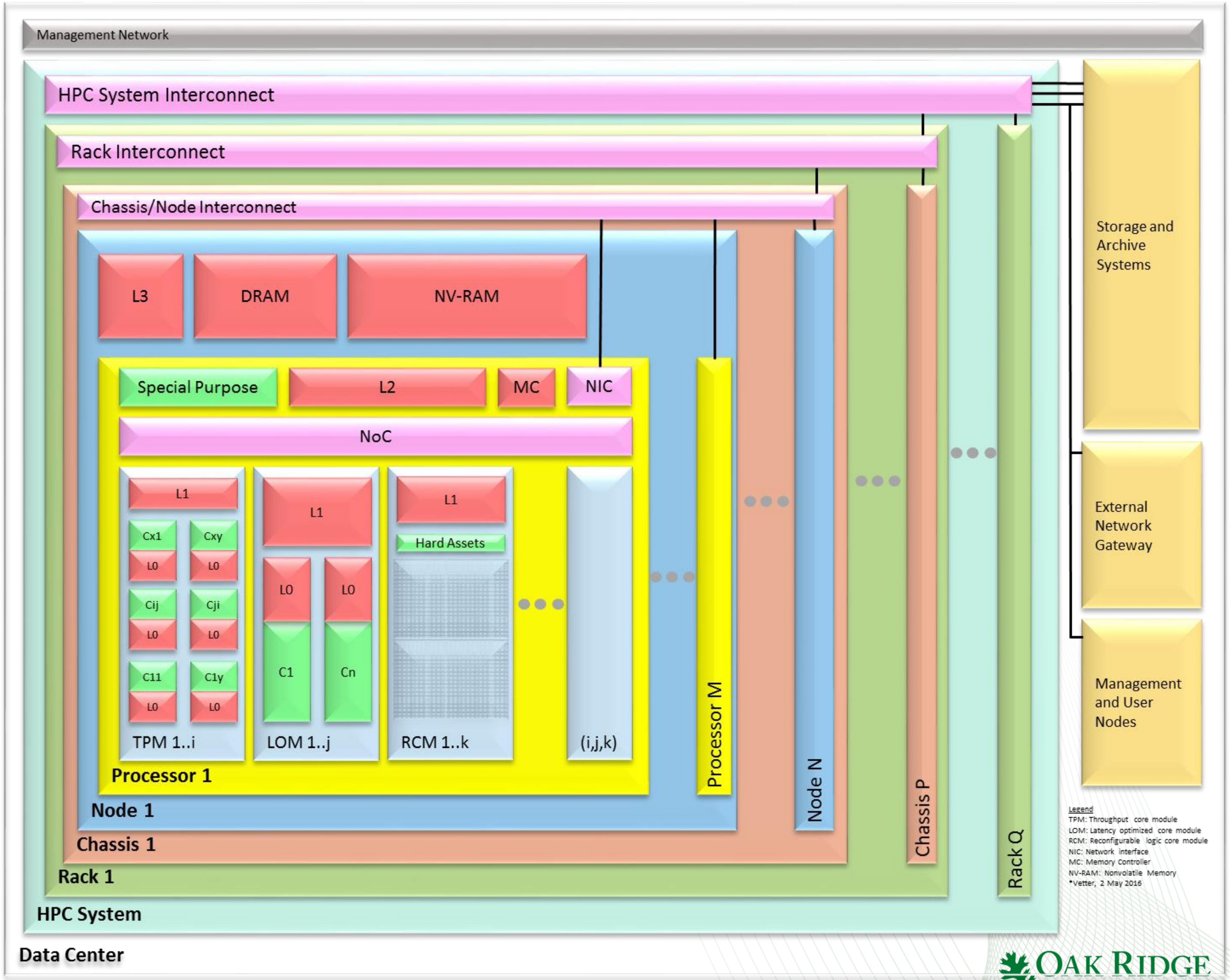
	Deployed					Experimental			
	SRAM	DRAM	eDRAM	2D NAND Flash	3D NAND Flash	PCRAM	STTRAM	2D ReRAM	3D ReRAM
Data Retention	N	N	N	Y	Y	Y	Y	Y	Y
Cell Size (F^2)	50-200	4-6	19-26	2-5	<1	4-10	8-40	4	<1
Minimum F demonstrated (nm)	14	25	22	16	64	20	28	27	24
Read Time (ns)	< 1	30	5	10^4	10^4	10-50	3-10	10-50	10-50
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

Migration up the hierarchy





Legend
TPM: Throughput core module
LOM: Latency optimized core module
RCM: Reconfigurable logic core module
NIC: Network interface
MC: Memory Controller
NV-RAM: Nonvolatile Memory
*Vetter, 2 May 2016

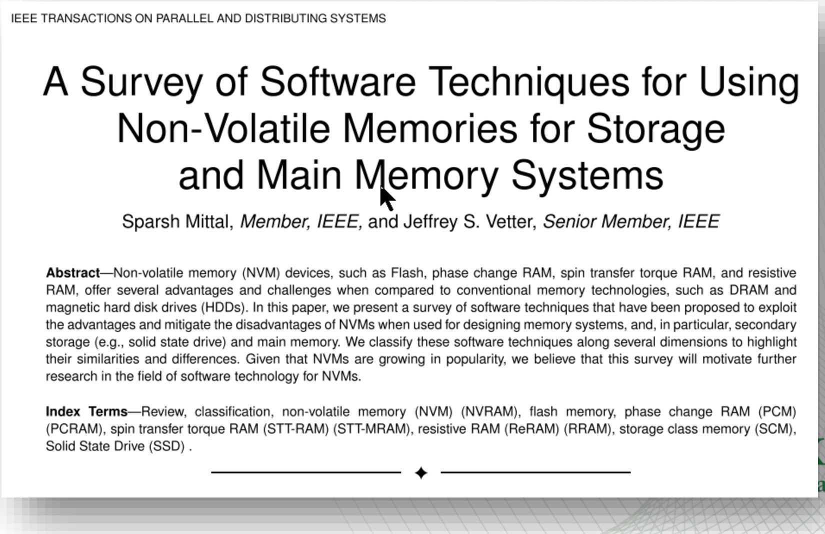
Programming NVM Systems

NVM Programming Systems : Goals

- Architectures will vary dramatically
  - How should we design the node?
  - Portable across various NVM architectures
  - MPI and OpenMP do not solve this problem.
- Two modes of operation
  - Drop in replacement for DRAM
  - Exploit persistence
- Active area of research
- Performance for HPC scenarios
  - Allow user or compiler/runtime/os to exploit NVM
  - Asymmetric R/W
  - Remote/Local
- Assume lower power costs under normal usage
- Security

- Correctness and durability
  - A crash or erroneous program could corrupt the NVM data structures
  - Programming system needs to provide support for this model
  - Enhanced ECC for NVM devices
- ACID
  - Atomicity: A transaction is “all or nothing”
  - Consistency: Takes data from one consistent state to another
  - Isolation: Concurrent transactions appears to be one after another
  - Durability: Changes to data will remain across system boots

http://j.mp/nvm-sw-survey



IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTING SYSTEMS

A Survey of Software Techniques for Using Non-Volatile Memories for Storage and Main Memory Systems

Sparsh Mittal, *Member, IEEE,* and Jeffrey S. Vetter, *Senior Member, IEEE*

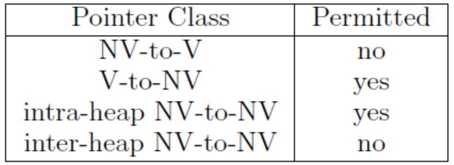
Abstract—Non-volatile memory (NVM) devices, such as Flash, phase change RAM, spin transfer torque RAM, and resistive RAM, offer several advantages and challenges when compared to conventional memory technologies, such as DRAM and magnetic hard disk drives (HDDs). In this paper, we present a survey of software techniques that have been proposed to exploit the advantages and mitigate the disadvantages of NVMs when used for designing memory systems, and, in particular, secondary storage (e.g., solid state drive) and main memory. We classify these software techniques along several dimensions to highlight their similarities and differences. Given that NVMs are growing in popularity, we believe that this survey will motivate further research in the field of software technology for NVMs.

Index Terms—Review, classification, non-volatile memory (NVM) (NVRAM), flash memory, phase change RAM (PCM) (PCRAM), spin transfer torque RAM (STT-RAM) (STT-MRAM), resistive RAM (ReRAM) (RRAM), storage class memory (SCM), Solid State Drive (SSD) .

NVL-C: Portable Programming for NVMM

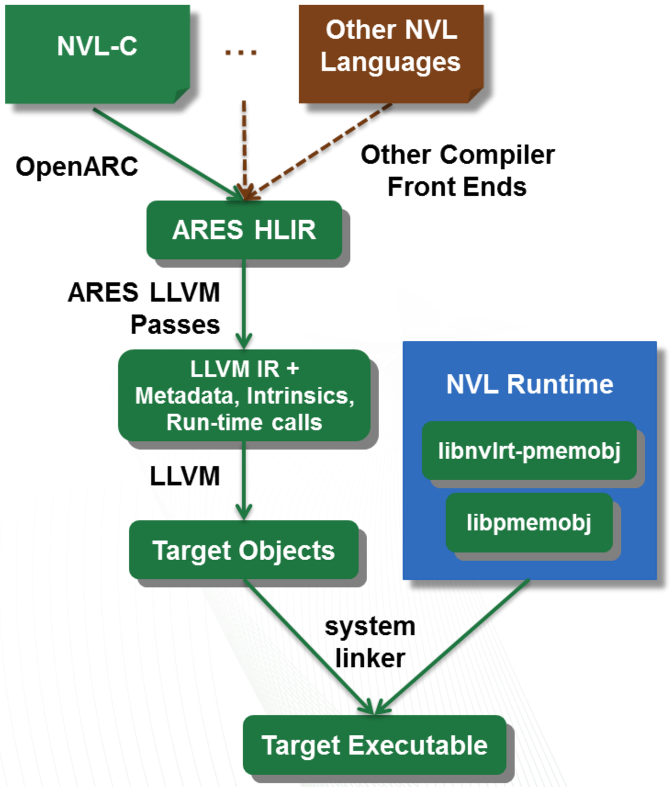
- Minimal, familiar, programming interface:
  - Minimal C language extensions.
  - App can still use DRAM.
- Pointer safety:
  - Persistence creates new categories of pointer bugs.
  - Best to enforce pointer safety constraints at compile time rather than run time.
- Transactions:
  - Prevent corruption of persistent memory in case of application or system failure.
- Language extensions enable:
  - Compile-time safety constraints.
  - NVM-related compiler analyses and optimizations.
- LLVM-based:
  - Core of compiler can be reused for other front ends and languages.
  - Can take advantage of LLVM ecosystem.

```
#include <nvl.h>
struct list {
  int value;
  nvl struct list *next;
};
void remove(int k) {
  nvl_heap_t *heap
    = nvl_open("foo.nvl");
  nvl struct list *a
    = nvl_get_root(heap, struct list);
  #pragma nvl atomic
  while (a->next != NULL) {
    if (a->next->value == k)
      a->next = a->next->next;
    else
      a = a->next;
  }
  nvl_close(heap);
}
```



Pointer Class	Permitted
NV-to-V	no
V-to-NV	yes
intra-heap NV-to-NV	yes
inter-heap NV-to-NV	no

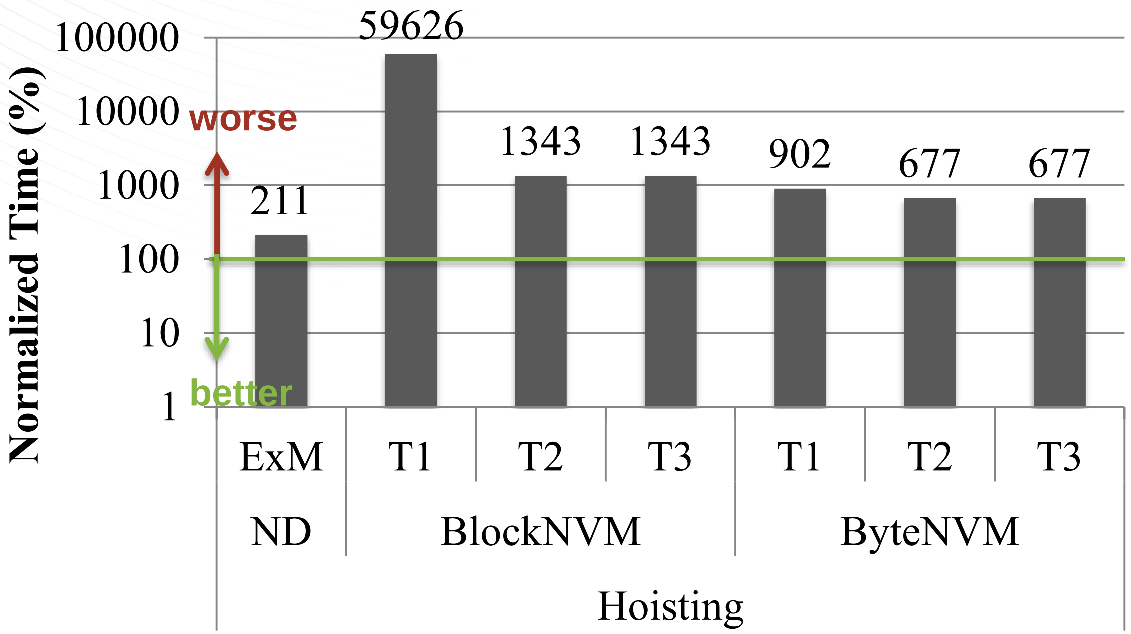
Table 1: Pointer Classes



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*

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

Summary

- Recent trends in extreme-scale HPC paint an uncertain 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 systems must provide performance portability (beyond functional portability)!!
  - Heterogeneous systems
  - New memory hierarchies

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