Exploring Emerging Technologies in the HPC Co-Design Space

Jeffrey S. Vetter

Presented to
AsHES Workshop, IPDPS
Phoenix
19 May 2014

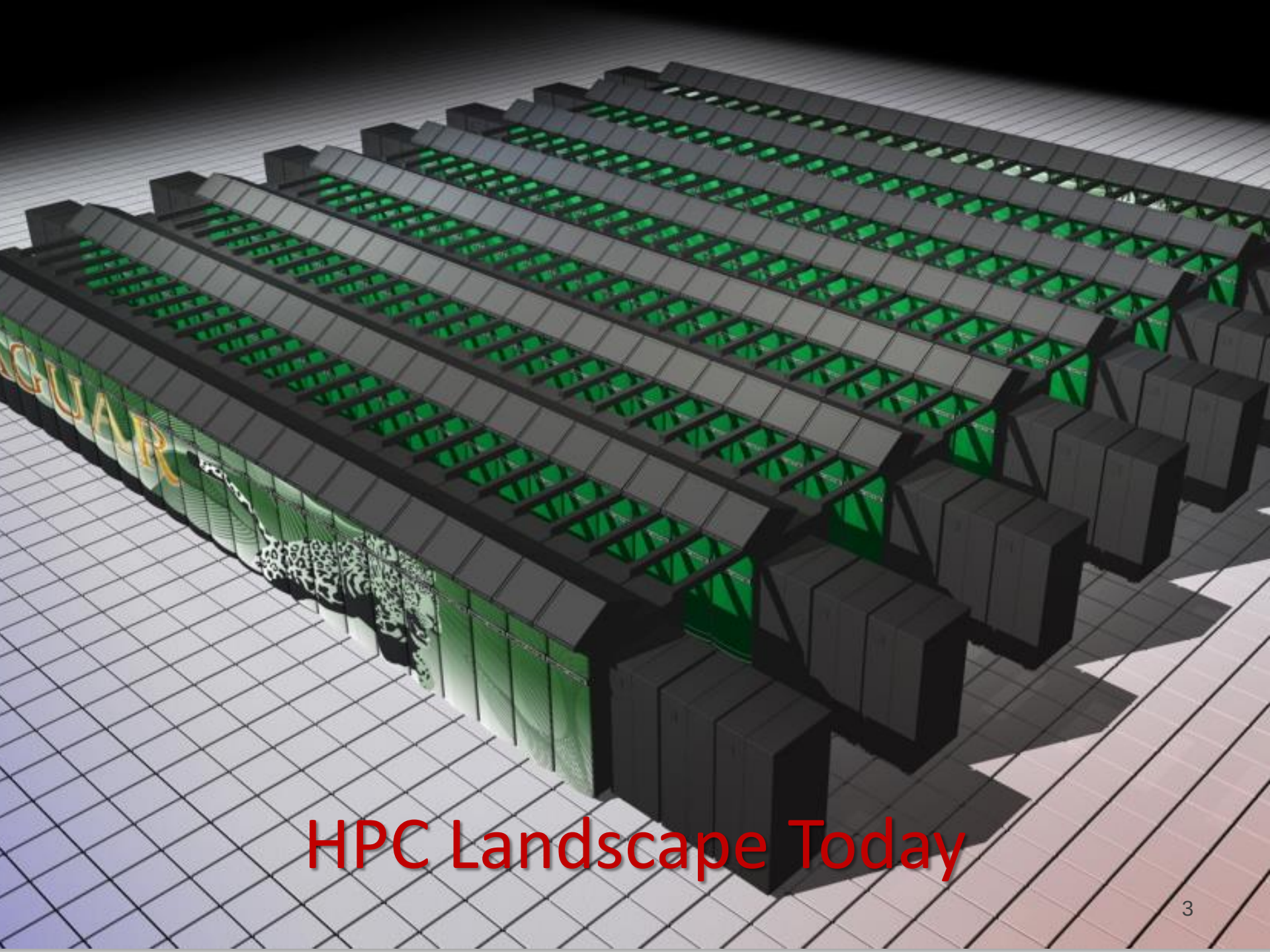
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Georgia Tech College of Computing
Computational Science and Engineering

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

Presentation in a nutshell

- **Our community expects major challenges in HPC as we move to extreme scale**
  - Power, Performance, Resilience, Productivity
  - Major shifts in architectures, software, applications
    - Most uncertainty in two decades
- **Applications will have to change in response to design of processors, memory systems, interconnects, storage**
  - DOE has initiated Codesign Centers that bring together all stakeholders to develop integrated solutions
- **Technologies particularly pertinent to addressing some of these challenges**
  - Heterogeneous computing
  - Nonvolatile memory
- **We need to reexamine software solutions to make this period of uncertainty palpable for computational science**
  - OpenARC
  - Memory allocation strategies



HPC Landscape Today

Notional Exascale Architecture Targets

(From Exascale Arch Report 2009)

System attributes	2001	2010	"2015"		"2018"	
System peak	10 Tera	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	~0.8 MW	6 MW	15 MW		20 MW	
System memory	0.006 PB	0.3 PB	5 PB		32-64 PB	
Node performance	0.024 TF	0.125 TF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW		25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	16	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	416	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW		1.5 GB/s	150 GB/sec	1 TB/sec	250 GB/sec	2 TB/sec
MTTI		day	O(1 day)		O(1 day)	

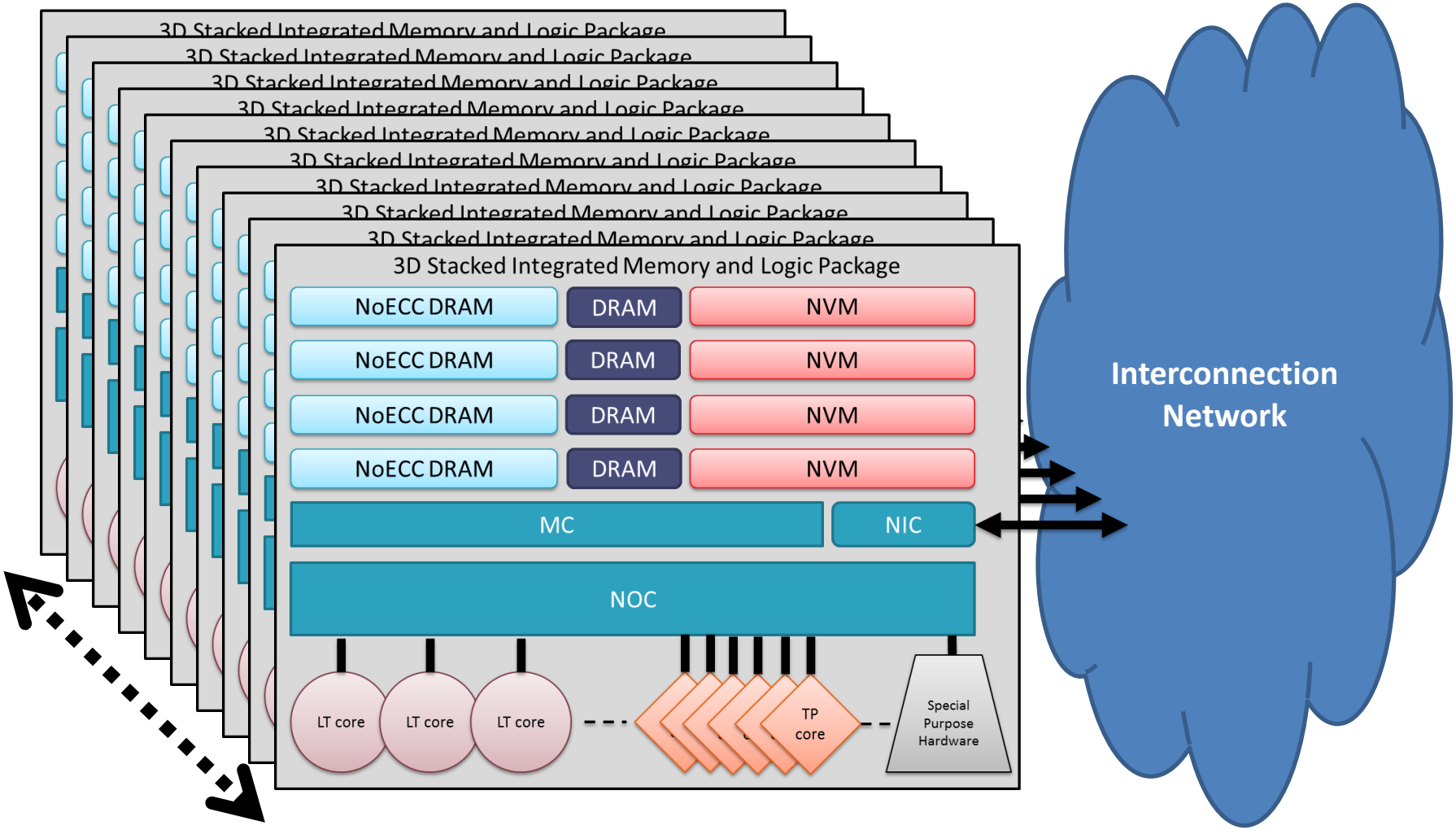
http://science.energy.gov/ascr/news-and-resources/workshops-and-conferences/grand-challenges/

Contemporary HPC Architectures

Date	System	Location	Comp	Comm	Peak (PF)	Power (MW)
2009	Jaguar; Cray XT5	ORNL	AMD 6c	Seastar2	2.3	7.0
2010	Tianhe-1A	NSC Tianjin	Intel + NVIDIA	Proprietary	4.7	4.0
2010	Nebulae	NSCS Shenzhen	Intel + NVIDIA	IB	2.9	2.6
2010	Tsubame 2	TiTech	Intel + NVIDIA	IB	2.4	1.4
2011	K Computer	RIKEN/Kobe	SPARC64 VIIIfx	Tofu	10.5	12.7
2012	Titan; Cray XK6	ORNL	AMD + NVIDIA	Gemini	27	9
2012	Mira; BlueGeneQ	ANL	SoC	Proprietary	10	3.9
2012	Sequoia; BlueGeneQ	LLNL	SoC	Proprietary	20	7.9
2012	Blue Waters; Cray	NCSA/UIUC	AMD + (partial) NVIDIA	Gemini	11.6	
2013	Stampede	TACC	Intel + MIC	IB	9.5	5
2013	Tianhe-2	NSCC-GZ (Guangzhou)	Intel + MIC	Proprietary	54	~20

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

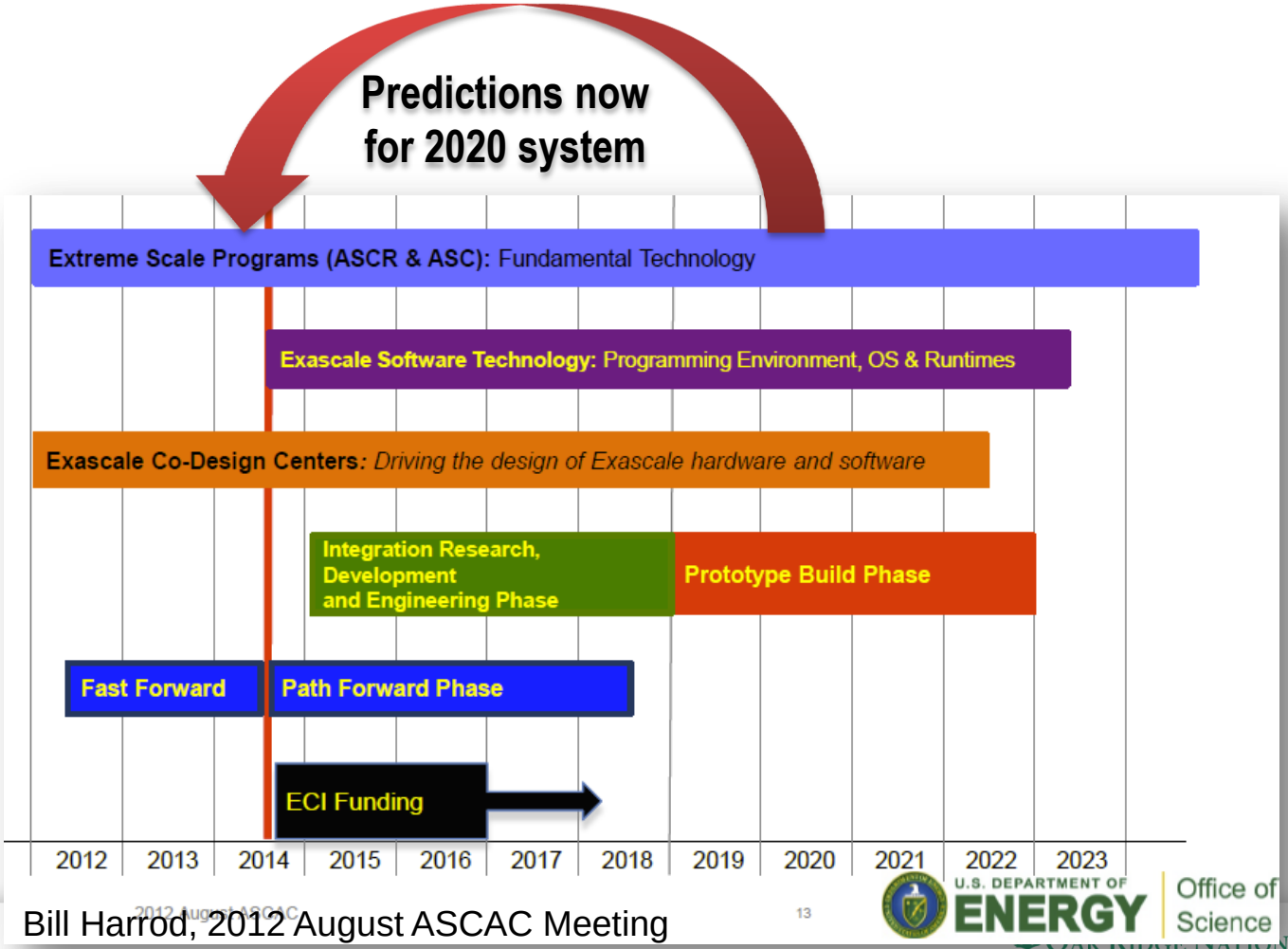
Notional Future Architecture



Co-designing Future Extreme Scale Systems

Designing for the future

- Empirical measurement is necessary but we must investigate future applications on future architectures using future software stacks



U.S. DEPARTMENT OF ENERGY | Office of Science
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Holistic View of HPC

Performance, Resilience, Power, Programmability

Applications

- Materials
- Climate
- Fusion
- National Security
- Combustion
- Nuclear Energy
- Cybersecurity
- Biology
- High Energy Physics
- Energy Storage
- Photovoltaics
- National Competitiveness

- Usage Scenarios
  - Ensembles
  - UQ
  - Visualization
  - Analytics

Programming Environment

- Domain specific
  - Libraries
  - Frameworks
  - Templates
  - Domain specific languages
  - Patterns
  - Autotuners

- Platform specific
  - Languages
  - Compilers
  - Interpreters/Scripting
  - Performance and Correctness Tools
  - Source code control

System Software

- Resource Allocation
- Scheduling
- Security
- Communication
- Synchronization
- Filesystems
- Instrumentation
- Virtualization

Architectures

- Processors
  - Multicore
  - Graphics Processors
  - Vector processors
  - FPGA
  - DSP
- Memory and Storage
  - Shared (cc, scratchpad)
  - Distributed
  - RAM
  - Storage Class Memory
  - Disk
  - Archival
- Interconnects
  - Infiniband
  - IBM Torrent
  - Cray Gemini, Aires
  - BGL/P/Q
  - 1/10/100 GigE

Oak Ridge National Laboratory
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Holistic View of HPC – Going Forward
Large design space –> uncertainty!

Performance, Resilience, Power, Programmability

Applications

- Materials
- Climate
- Fusion
- National Security
- Combustion
- Nuclear Energy
- Cybersecurity
- Biology
- High Energy Physics
- Energy Storage
- Photovoltaics
- National Competitiveness

- Usage Scenarios
  - Ensembles
  - UQ
  - Visualization
  - Analytics

Programming Environment

- Domain specific
  - Libraries
  - Frameworks
  - Templates
  - Domain specific languages
  - Patterns
  - Automata
- Platf
  - Langu
  - Compilers
  - Int
  - Pe Correc
  - Source

System Software

- Resource Allocation
- Scheduling
- Security
- Comm nication
- Sy nization
- ems
- ent

Architectures

- Processors
  - Multicore
  - Graphics Processors
  - Vector processors
  - FPGA
  - DSP
- Memory and Storage
  - Shared (cc, scratchpad)
  - Distributed
  - RAM
  - Storage Class Memory
  - Disk
  - Archival
- Interconnects
  - Infiniband
  - IBM Torrent
  - Cray Gemini, Aires
  - BGL/P/Q
  - 1/10/100 GigE

Large design space is challenging for apps, software, and architecture scientists.

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Slide courtesy of Karen Pao, DOE

U.S. DEPARTMENT OF ENERGY

Three Exascale Co-Design Centers selected after intense competition

Exascale Co-Design Center for Materials in Extreme Environments (ExMatEx)

Director: Timothy Germann (LANL)

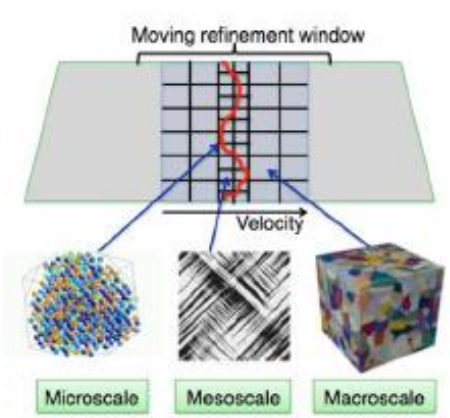
Center for Exascale Simulation of Advanced Reactors (CESAR)

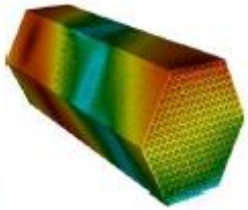
Director: Andrew Siegel (ANL)

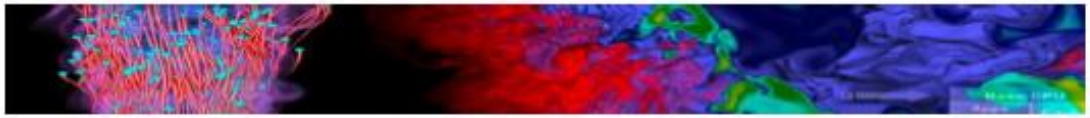
Center for Exascale Simulation of Combustion in Turbulence (EXaCT)

Director: Jacqueline Chen (SNL)

	ExMatEx (Germann)	CESAR (Rosner)	EXaCT (Chen)
National Labs	LANL	ANL	SNL
	LLNL	PNNL	LBNL
	SNL	LANL	LANL
	ORNL	ORNL	ORNL
		LLNL	LLNL
			NREL
University & Industry Partners	Stanford	MIT	Stanford
	CalTech	TAMU	GA Tech
		Rice	Rutgers
		U Chicago	UT Austin
		IBM	Utah
		TerraPower	
		General Atomic	
		Areva	





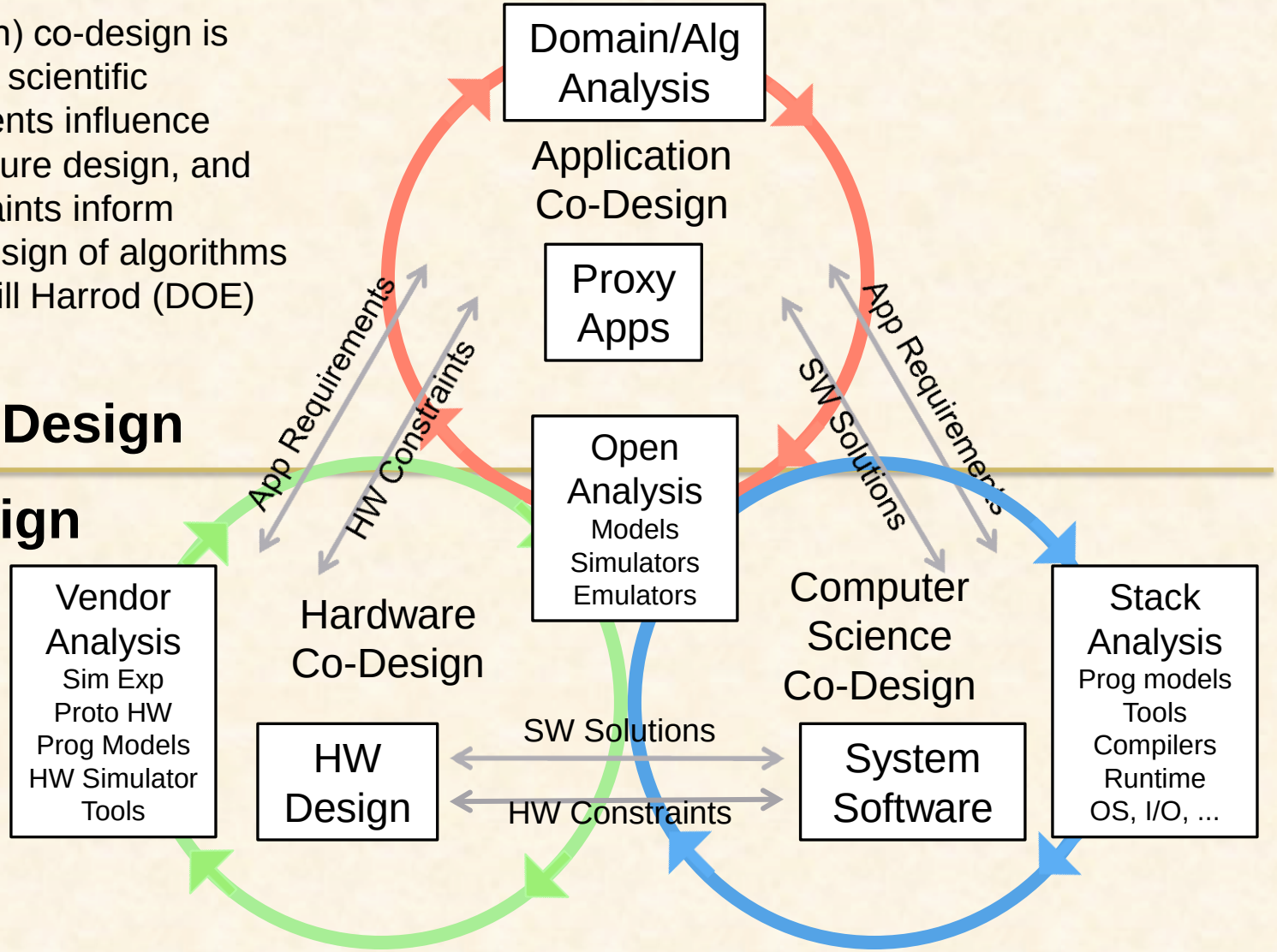


Each project is $4M/yr for 5 years, subject to satisfactory progress as gauged by frequent reviews

Slide courtesy of ExMatEx Co-design team.

Workflow within the Exascale Ecosystem

"(Application driven) co-design is the process where scientific problem requirements influence computer architecture design, and technology constraints inform formulation and design of algorithms and software." – Bill Harrod (DOE)

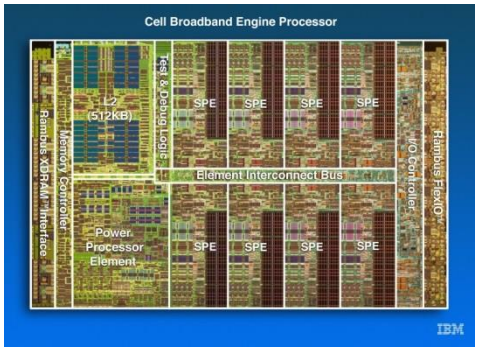


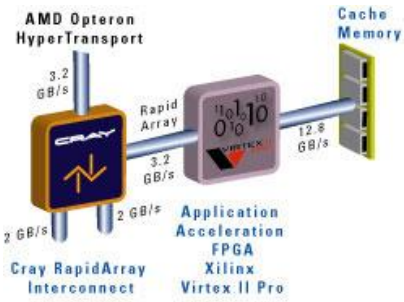
Emerging Architectures

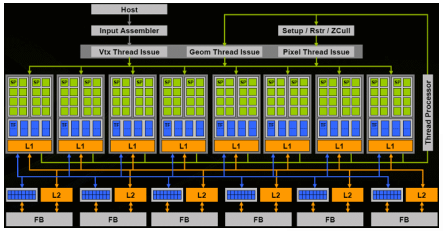
Earlier Experimental Computing Systems

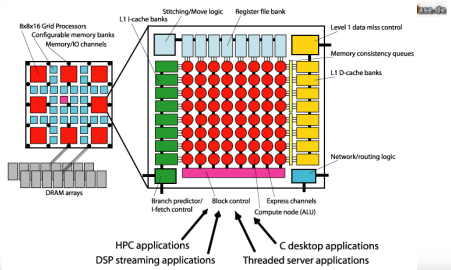
- The past decade has started the trend away from traditional 'simple' architectures
- Mainly driven by facilities costs and successful (sometimes heroic) application examples
- Examples
  - Cell, GPUs, FPGAs, SoCs, etc
- Many open questions
  - Understand technology challenges
  - Evaluate and prepare applications
  - Recognize, prepare, enhance programming models

Popular architectures since ~2004





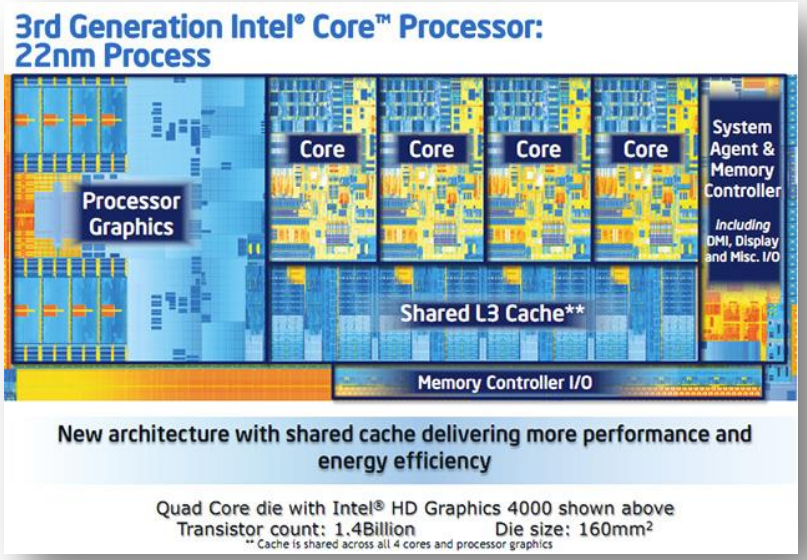


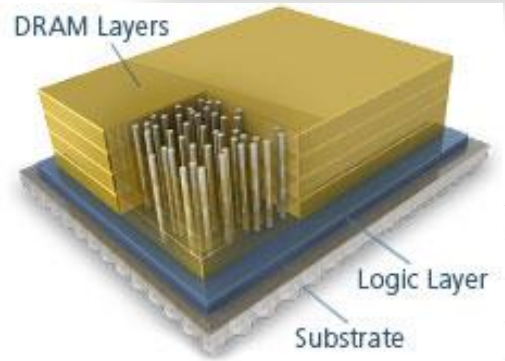


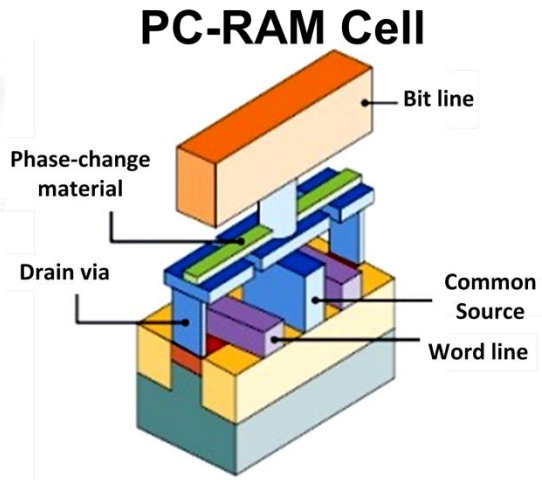
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Emerging Computing Architectures – Future

- Heterogeneous processing
  - Latency tolerant cores
  - Throughput cores
  - Special purpose hardware (e.g., AES, MPEG, RND)
  - Fused, configurable memory
- Memory
  - 2.5D and 3D Stacking
  - HMC, HBM, WIDEIO2, LPDDR4, etc
  - New devices (PCRAM, ReRAM)
- Interconnects
  - Collective offload
  - Scalable topologies
- Storage
  - Active storage
  - Non-traditional storage architectures (key-value stores)
- Improving performance and programmability in face of increasing complexity
  - Power, resilience





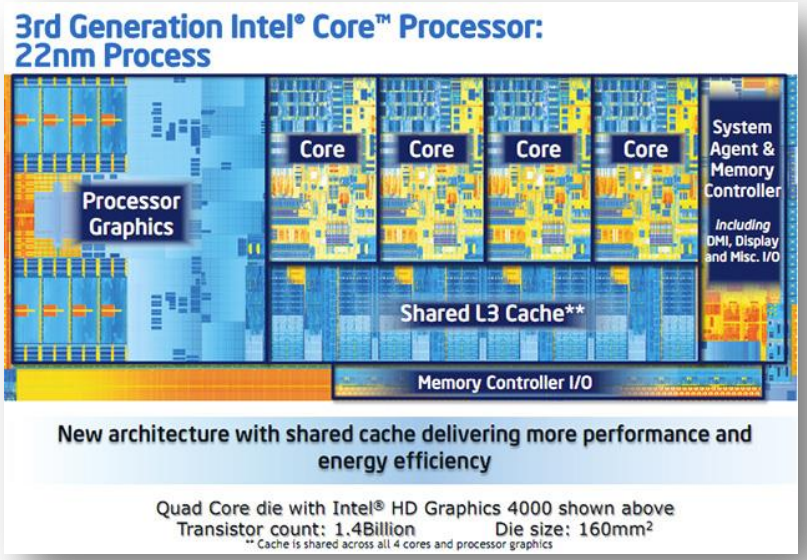


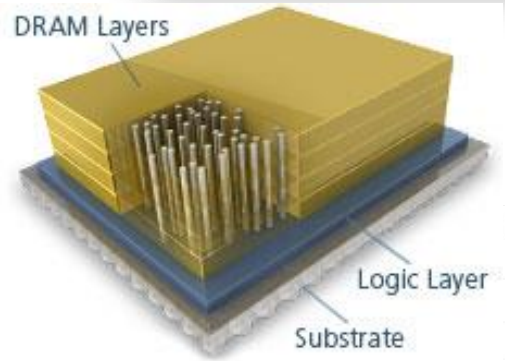
HPC (mobile, enterprise, embedded) computer design is more fluid now than in the past two decades.

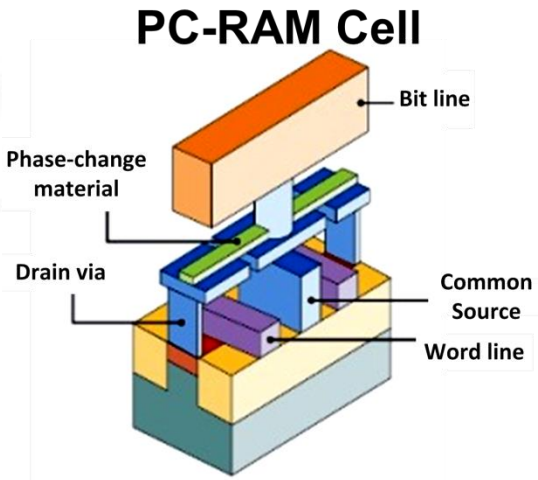
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Emerging Computing Architectures – Future

- Heterogeneous processing
  - Latency tolerant cores
  - Throughput cores
  - Special purpose hardware (e.g., AES, MPEG, RND)
  - Fused, configurable memory
- Memory
  - 2.5D and 3D Stacking
  - HMC, HBM, WIDEIO2, LPDDR4, etc
  - New devices (PCRAM, ReRAM)
- Interconnects
  - Collective offload
  - Scalable topologies
- Storage
  - Active storage
  - Non-traditional storage architectures (key-value stores)
- Improving performance and programmability in face of increasing complexity
  - Power, resilience







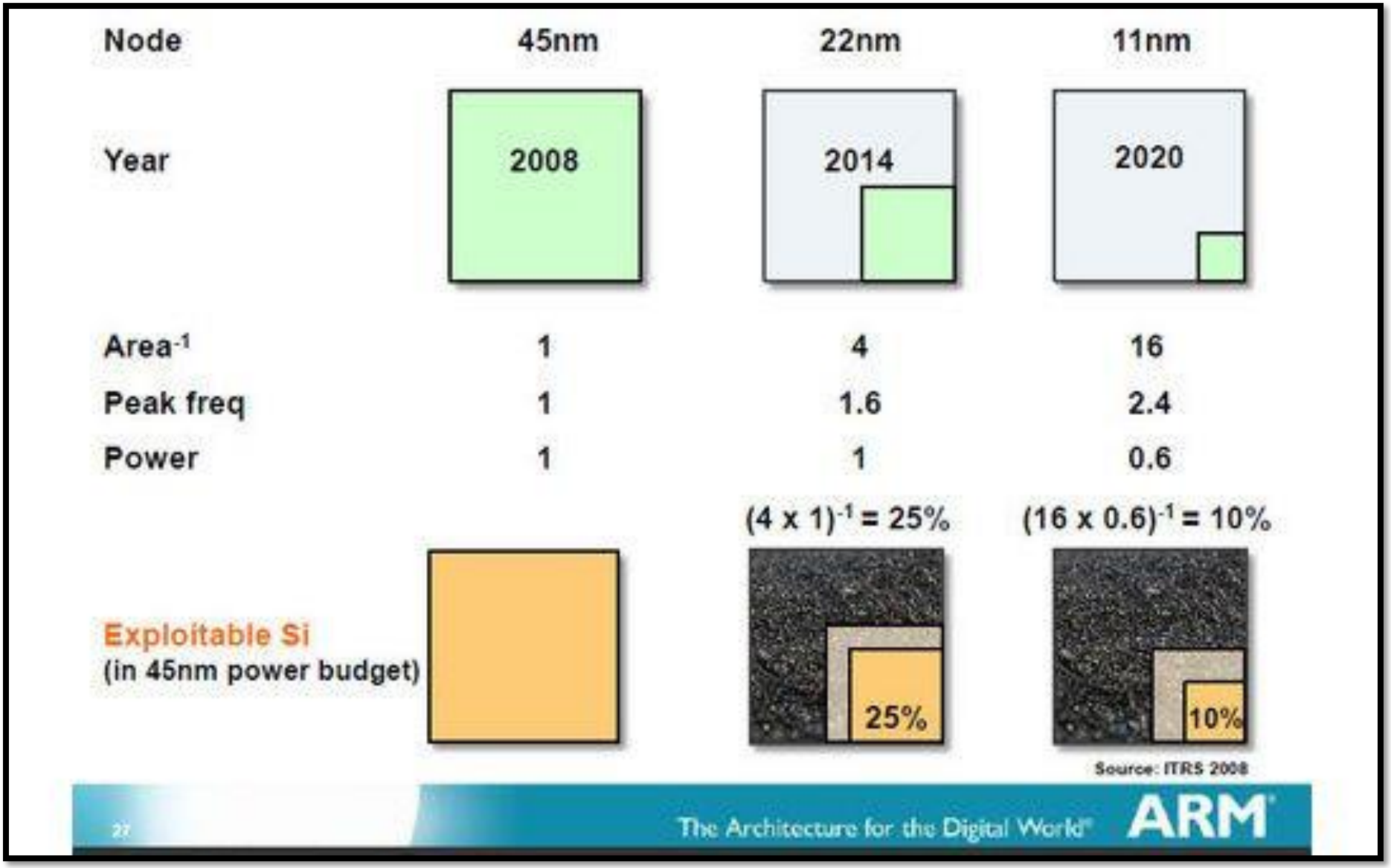
HPC (mobile, enterprise, embedded) computer design is more fluid now than in the past two decades.

MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Heterogeneous Computing

You could not step twice into the same river. -- Heraclitus

Dark Silicon Will Make Heterogeneity and Specialization More Relevant



Source: ARM

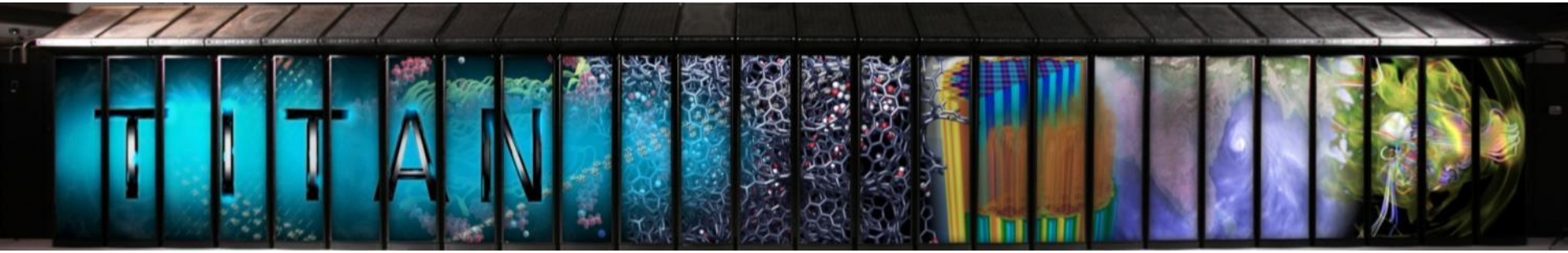
TH-2 System

- 54 Pflop/s Peak!
- Compute Nodes have 3.432 Tflop/s per node
  - 16,000 nodes
  - 32000 Intel Xeon cpus
  - 48000 Intel Xeon phis (57c/phi)
- Operations Nodes
  - 4096 FT CPUs as operations nodes
- Proprietary interconnect TH2 express
- 1PB memory (host memory only)
- Global shared parallel storage is 12.4 PB
- Cabinets: 125+13+24 = 162 compute/communication/storage cabinets
  - ~750 m2
- NUDT and Inspur

TH-2 (w/ Dr. Yutong Lu)

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

DOE's "Titan" Hybrid System: Cray XK7 with AMD Opteron and NVIDIA Tesla processors



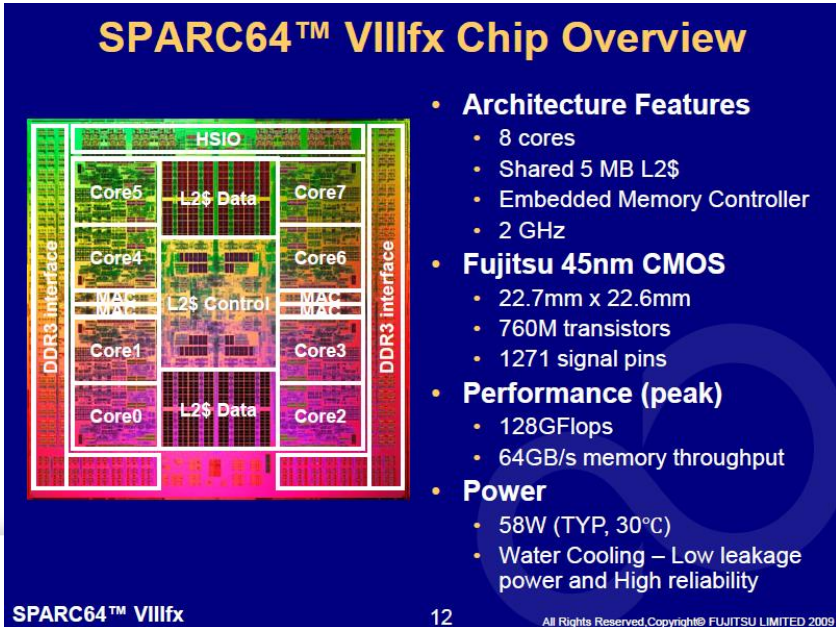
4,352 ft^2

SYSTEM SPECIFICATIONS:

- Peak performance of 27.1 PF
  - 24.5 GPU + 2.6 CPU
- 18,688 Compute Nodes each with:
  - 16-Core AMD Opteron CPU
  - NVIDIA Tesla "K20x" GPU
  - 32 + 6 GB memory
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
- 8.9 MW peak power

And many others

- BlueGene/Q
  - QPX vectorization
  - SMT
  - 16 cores per chip
  - L2 with memory speculation and atomic updates
  - List and stream prefetch
- K - Vector system
  - SPARC64 VIIIfx
  - Tofu interconnect
- Standard clusters
  - Tightly integrated GPUs
  - Wide AVX – 256b
  - Voltage and frequency islands
  - Transactional memory
  - PCIe G3

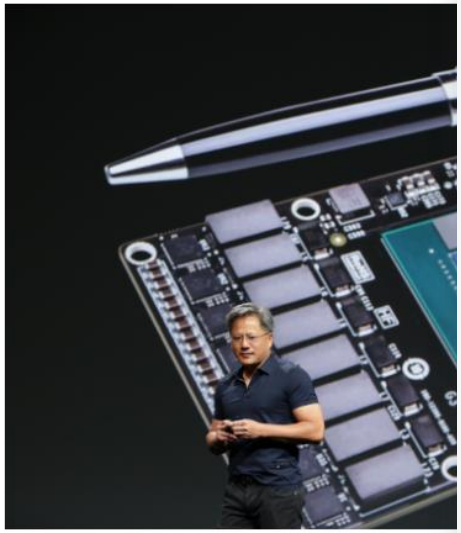


Integration is continuing …

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



Nvidia CEO Jen-Hsun Huang introduces Pascal at today's GPU Tec

Nvidia

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

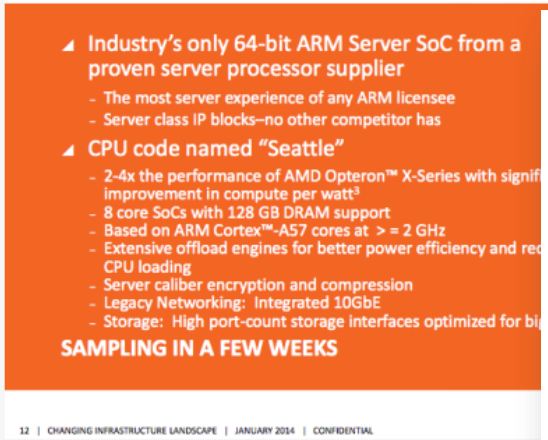
It Begins: AMD Announces Its First ARM Based Server SoC, 64-bit/8-core Opteron A1100

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



Around 15 months ago, AMD announced that it would be build
2014. Less than a month into 2014, AMD made good on its pro
A1100: a 64-bit ARM Cortex A57 based SoC.

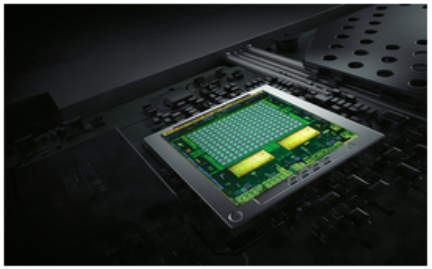
The Opteron A1100 features either 4 or 8 AMD Cortex A57 cor
talking about harvested die to make up the quad-core configur
away entirely, but since we're at very early stages of talking ab
bets going on. Each core will run at a frequency somewhere no
process at Global Foundries.

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.

Despite Nvidia having announced on its blog that it is "now shipping", the development kit that is powered by a Tegra K1 chip won't actually ship until 15 May.

Claiming to be "the world's first mobile supercomputer", the Jetson TK1 kit is built for embedded systems to aid the development of computers attempting to simulate human recognition of physical objects, such as robots and self-driving cars.

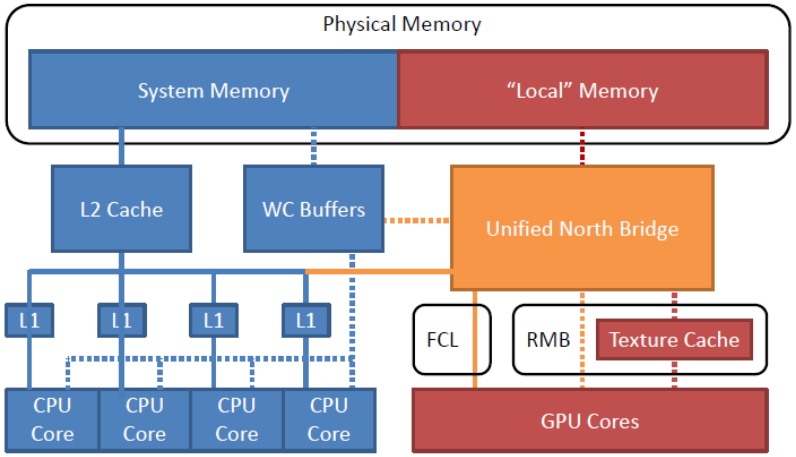
Speaking at the GPU Technology Conference (GTC) in March, Nvidia co-founder and CEO Jen Hsun Huang described it as capable of running anything the Geforce GTX Titan Z graphics card can run, but at a slower pace.

With a total performance of 326 GFLOPS, the Jetson TK1 should be more powerful than the Raspberry Pi board, which delivers just 24 GFLOPS, but will retail for much more, costing $192 in the US - a number that matches the number of cores in the Tegra K1 processor that Nvidia launched at CES in Las Vegas in January.

"The Jetson TK1 also comes with this new SDK called Vision Works. Stacked onto CUDA, it comes with a whole bunch of primitives whether it's recognising corners or detecting edges, or it could be classifying objects.

Parameters are loaded into this Vision Works primitives system and all of a sudden it recognises objects," Huang said on stage during the Jetson TK1 launch.

Fused memory hierarchy: AMD Llano



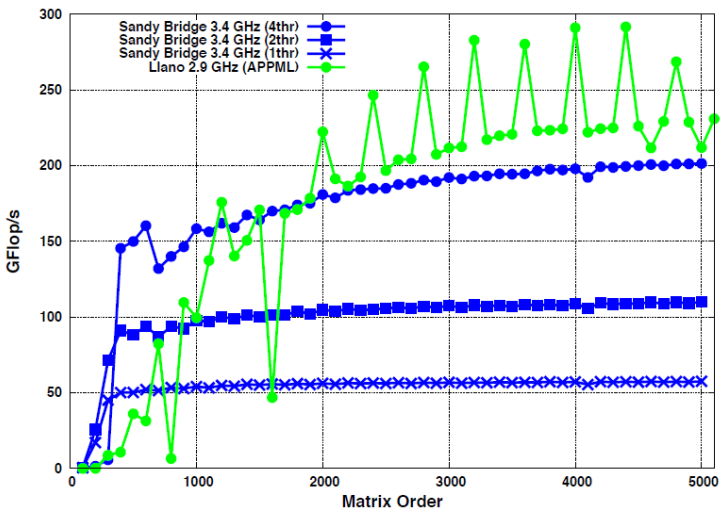
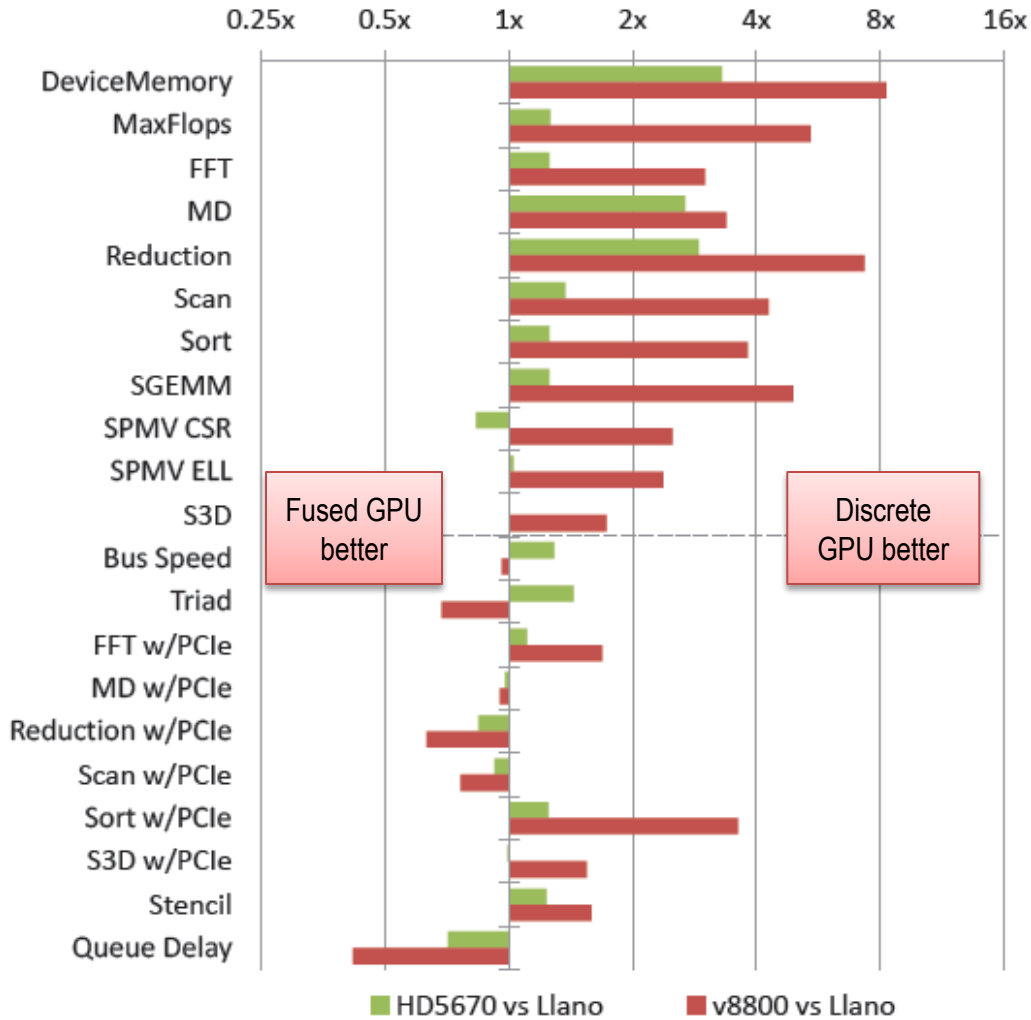


Figure 3: SGEMM Performance (one, two, and four CPU threads for Sandy Bridge and the OpenCL-based AMD APPML for Llano's fGPU)

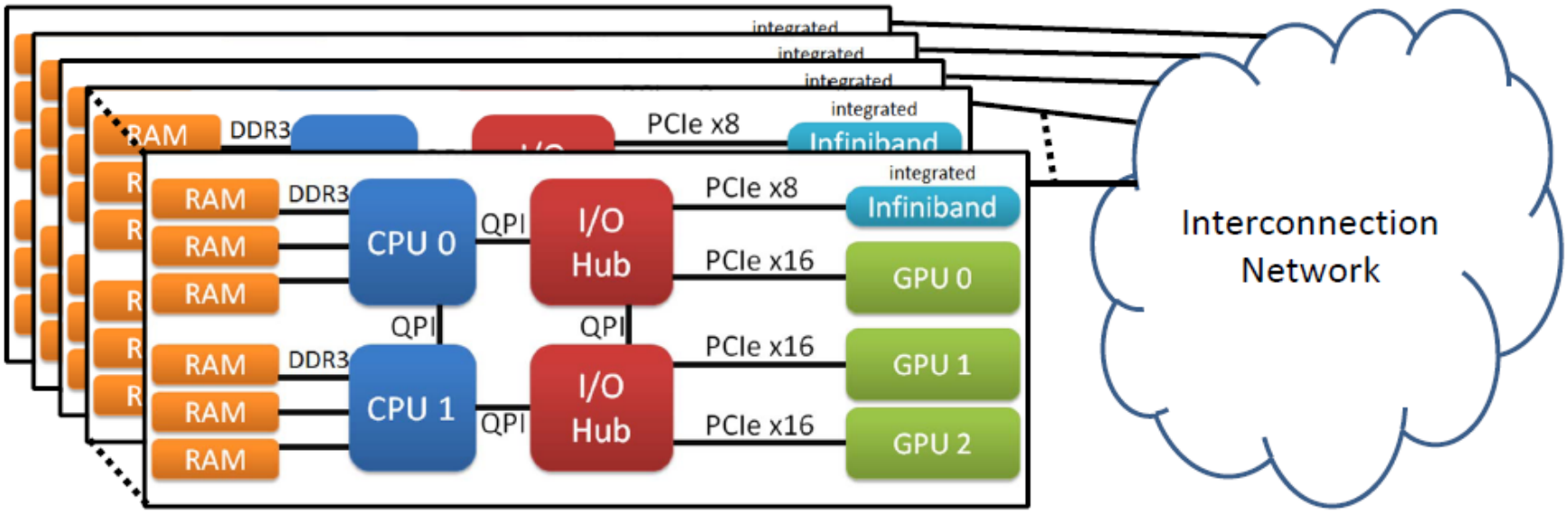


K. Spafford, J.S. Meredith, S. Lee, D. Li, P.C. Roth, and J.S. Vetter, "The Tradeoffs of Fused Memory Hierarchies in Heterogeneous Architectures," in ACM Computing Frontiers (CF). Cagliari, Italy: ACM, 2012. Note: Both SB and Llano are consumer, not server, parts.

MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Programming Heterogeneous Systems Productively

Applications must use a mix of programming models for these architectures



MPI

- Low overhead
- Resource contention
- Locality

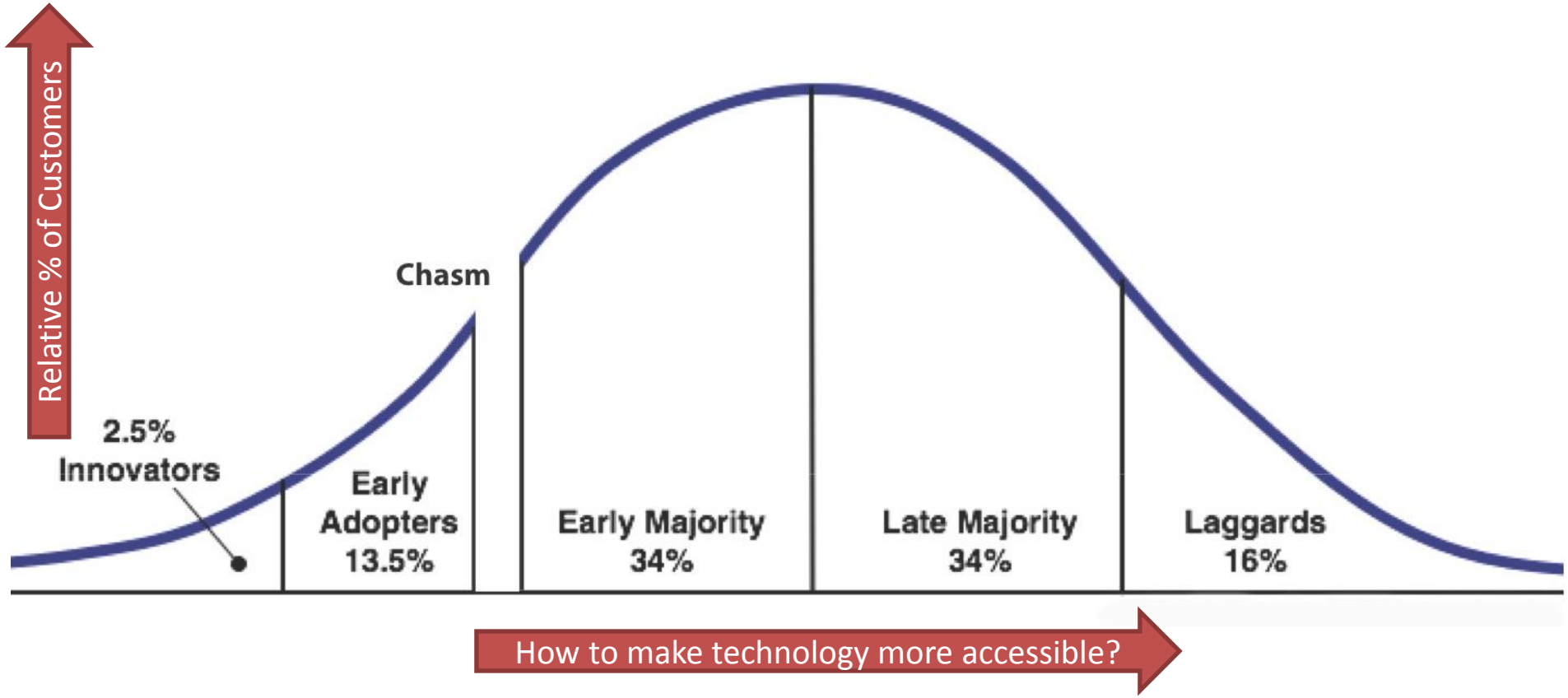
OpenMP, Pthreads

- SIMD
- NUMA

OpenACC, CUDA, OpenCL, OpenMP4, …

- Memory use, coalescing
- Data orchestration
- Fine grained parallelism
- Hardware features

Technology Adoption Lifecycle



Crossing the Chasm, Geoffrey A. Moore

Realizing performance portability across contemporary heterogeneous architectures

- Can we develop a 'write once, run anywhere efficiently' application with advanced compilers, runtime systems, and autotuners?

Table 1: Comparison of Heterogeneous Architectures

Property	CUDA	GCN	MIC
Programming models	CUDA, OpenCL	OpenCL, C++ AMP	OpenCL, Cilk, TBB, LEO OpenMP
Thread Scheduling	Hardware	Hardware	Software
User Managed Cache	Yes	Yes	No
Global Synchronization	No	No	Yes
L2 Cache Type	Shared	Private per core	Private per core
L2 Total Size	upto 1.5MB	upto 0.5MB	25MB
L2 Line-size	128	64	64
L1 Data Cache	Read-only + Read-write	Read-only	Read-write
Native Mode	No	No	Yes

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

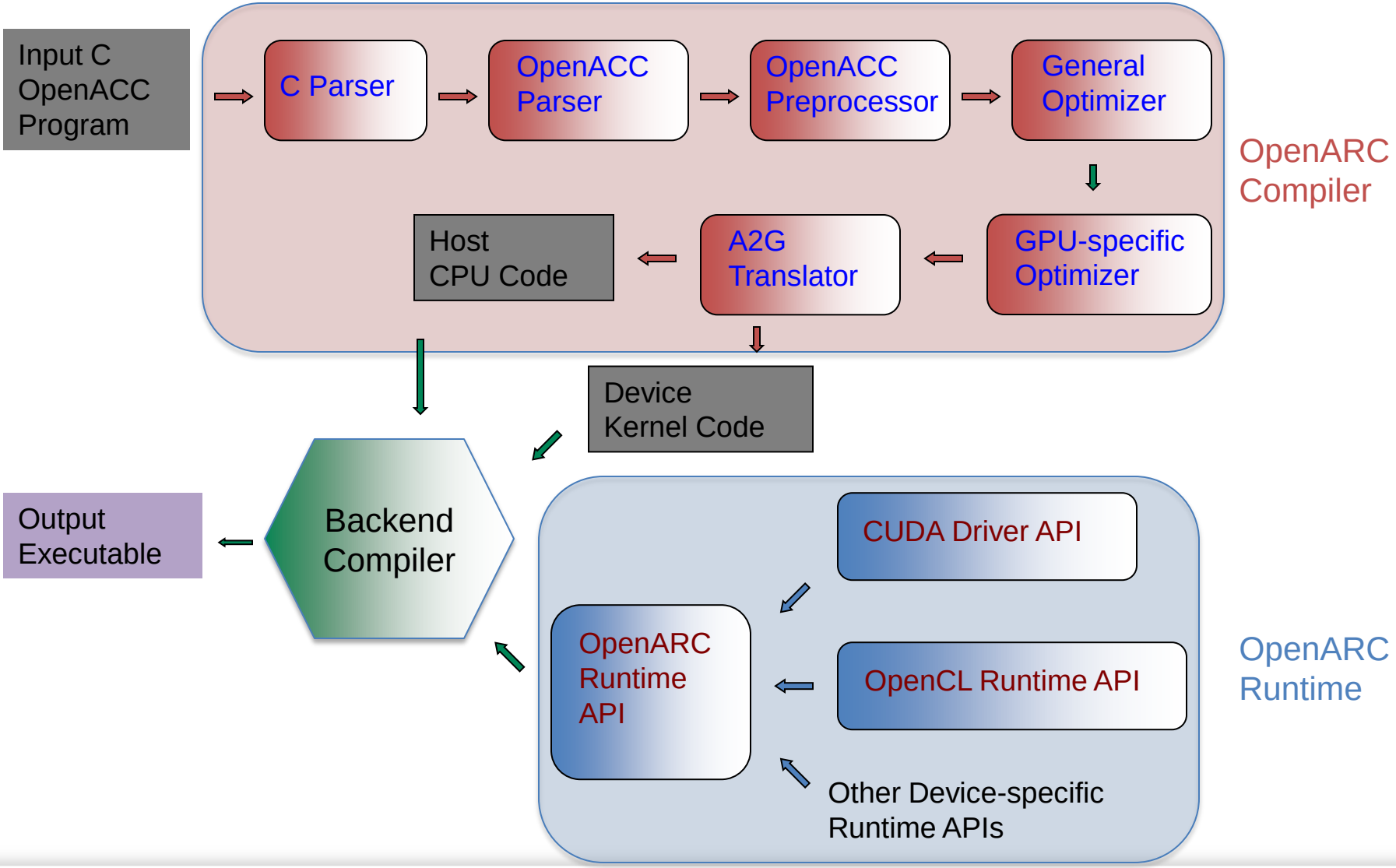
"Write one program and run efficiently anywhere"

- 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
      - Supports OpenMP3
  - 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.
  - Building common high level IR that includes constructs for parallelism, data movement, etc

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*

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

OpenARC System Architecture



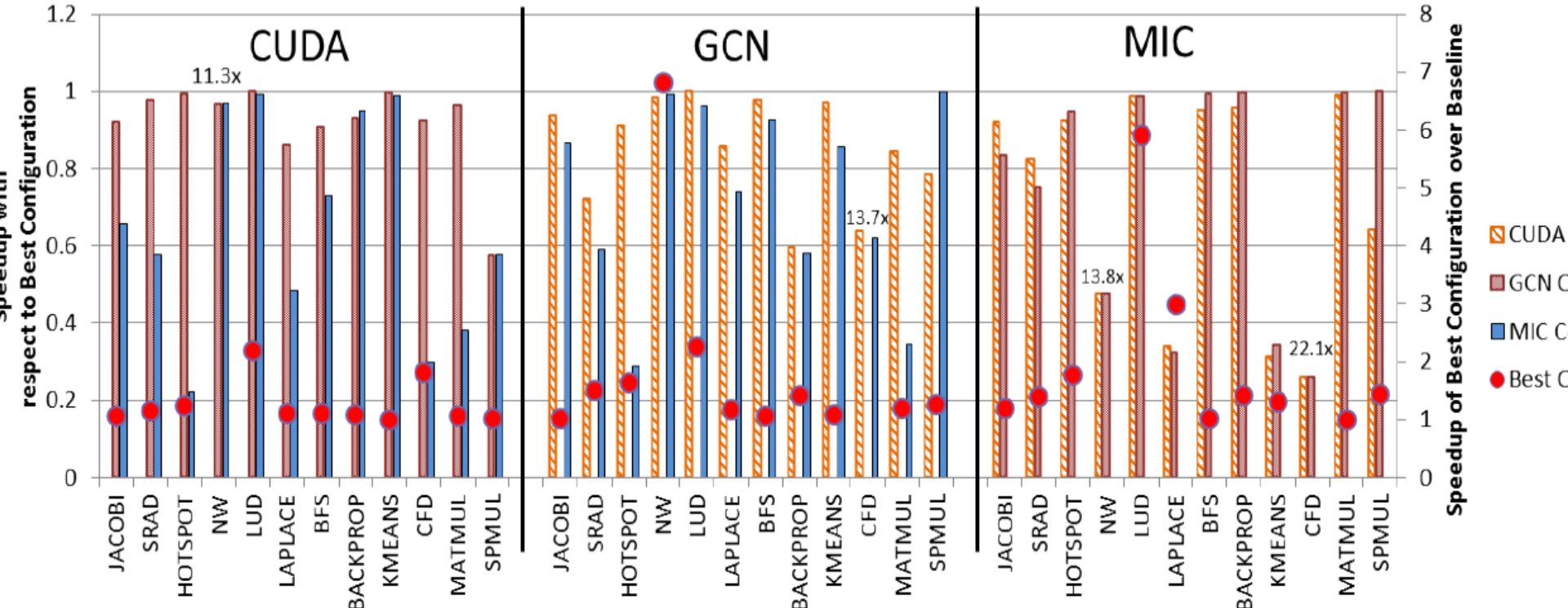
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

39

Performance Portability is critical and challenging

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

- One 'best configuration' on other architectures
- Major differences
  - Parallelism arrangement
  - Device-specific memory
  - Other arch optimizations



Automating selection of optimizations based on machine model

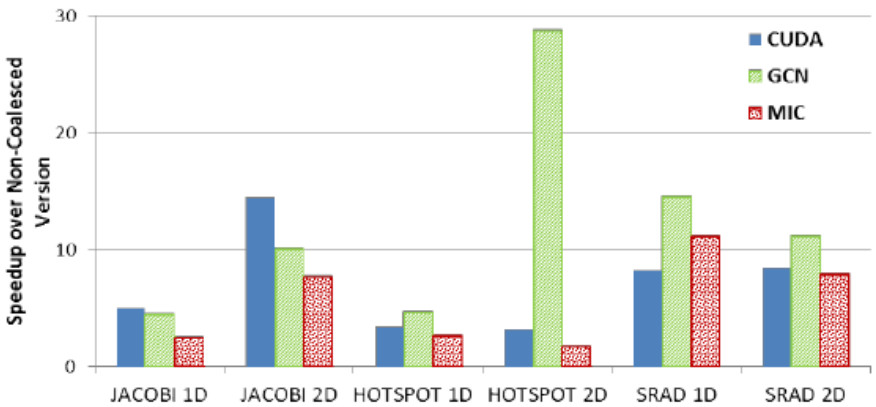


Figure 5: Memory Coalescing Benefits on Different Architectures : MIC is impacted the least by the non-coalesced accesses

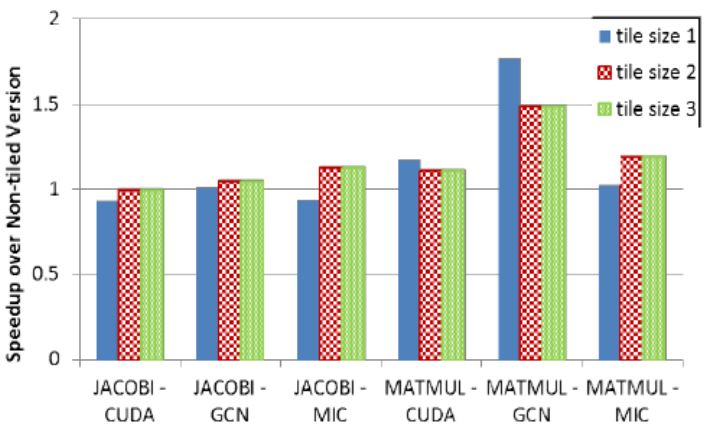


Figure 7: Impact of Tiling Transformation : *MATMUL* shows higher benefits than *JACOBI* owing to more contiguous accesses

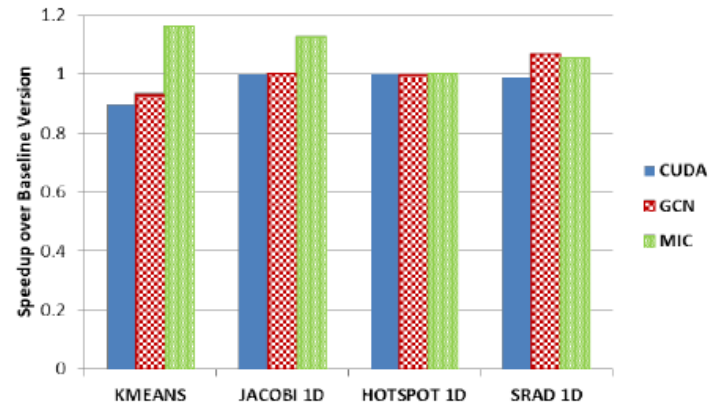


Figure 9: Effects of Loop Unrolling - MIC shows benefits on unrolling

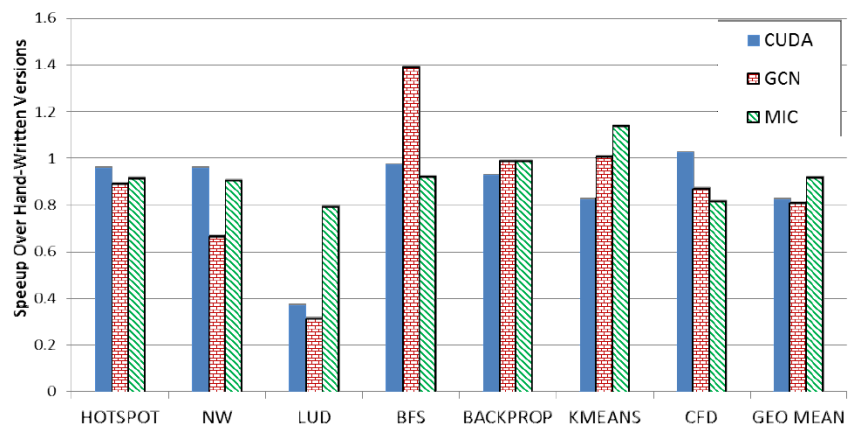


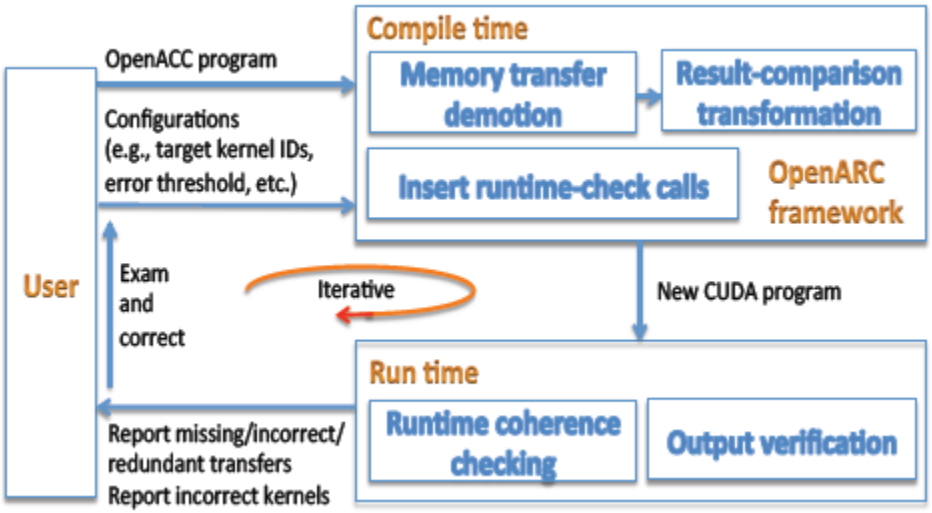
Fig. 11: Comparison of hand-written CUDA/OpenCL programs against auto-tuned OpenARC code versions : Tuned OpenACC programs perform reasonably well against hand-written codes

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Optimization and Interactive Program Verification with OpenARC

Talk Tues

- Problem
  - *Too much abstraction* in directive-based GPU programming!
    - Debuggability
      - Difficult to diagnose logic errors and performance problems at the directive level
    - Performance Optimization
      - Difficult to find where and how to optimize



- Solution
  - Directive-based, interactive GPU program verification and optimization
    - OpenARC compiler:
      - Generates runtime codes necessary for *GPU-kernel verification* and *memory-transfer verification and optimization*.
    - Runtime
      - Locate trouble-making kernels by comparing execution results at kernel granularity.
      - Trace the runtime status of CPU-GPU coherence to detect incorrect/missing/redundant memory transfers.
    - Users
      - Iteratively fix/optimize incorrect kernels/memory transfers based on the runtime feedback and apply to input program.

S. Lee, D. Li, and J.S. Vetter, "Interactive Program Debugging and Optimization for Directive-Based, Efficient GPU Computing," in *IEEE International Parallel and Distributed Processing Symposium (IPDPS). Phoenix: IEEE, 2014*

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Example Optimization: Identify and Optimize Data Transfers

- By adding additional instrumentation, OpenARC can help identify redundant and incorrect data transfers
- User can optimize by adding pragmas

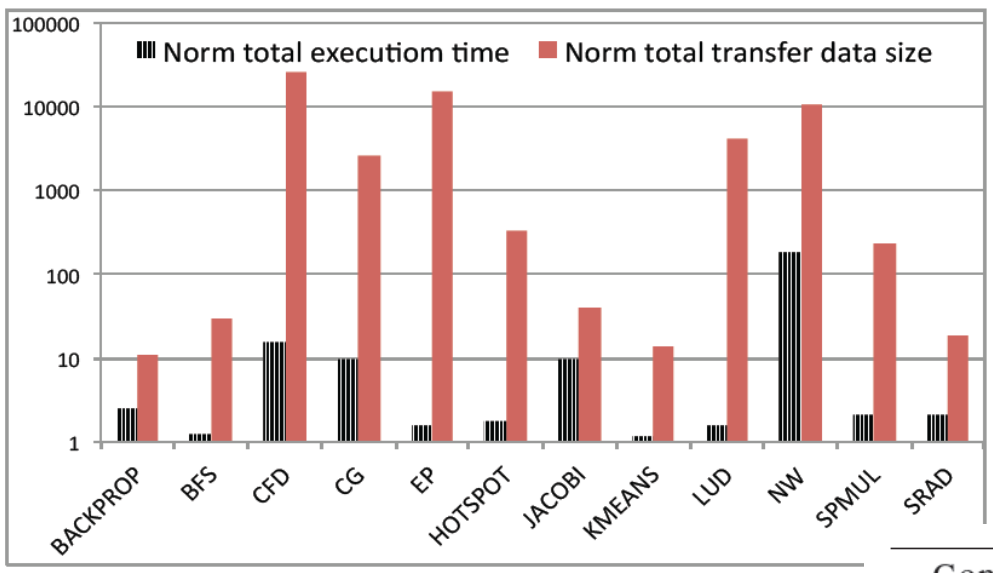


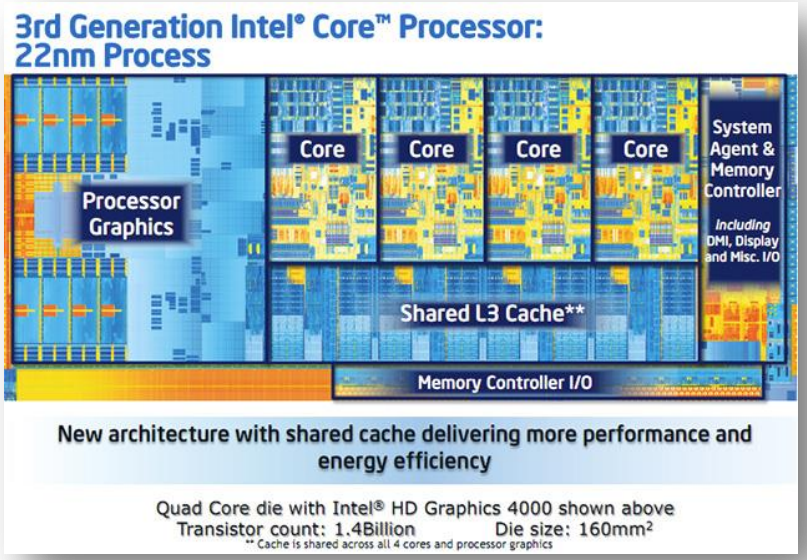
Figure 1: Total data transferred by unopt memory management scheme and its executio normalized to fully optimal scheme

```
– Copying b from device to host in update0
  (enclosing loop index = 1) is redundant.
– Copying b from device to host in update0
  (enclosing loop index = 2) is redundant.
...//repeated until the last k–loop iteration
```

Listing 4: Sample debugging messages for JACOBI kernel in Listing 3. *update0* refers to *memcpyout(b)* in line 8 in Listing 3.

Future Directions in Heterogeneous Computing

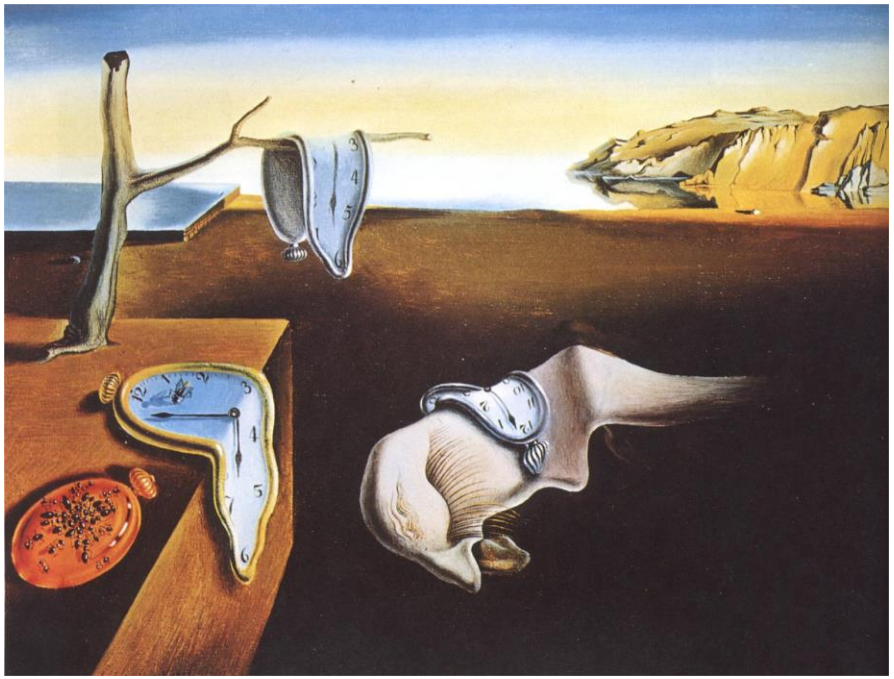
- Over the next decade: Heterogeneous computing will continue to increase in importance
  - Embedding and mobile community have already experienced this trend
- Manycore
  - Integrated GPUs, special purpose HW
- Hardware features
  - Transactional memory
  - Random Number Generators
    - MC caveat
  - Scatter/Gather
  - Wider SIMD/AVX
  - AES, Compression, etc
- Synergies with BIGDATA, mobile markets, graphics
- Top 10 list of features to include *from application perspective. Now is the time!*



- The future is about new productive programming models
- Inform applications teams to new features and gather their requirements

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Memory Systems



http://www.wikipaintings.org/en/salvador-dali/the-persistence-of-memory-1931

The Persistence of Memory

Notional Exascale Architecture Targets

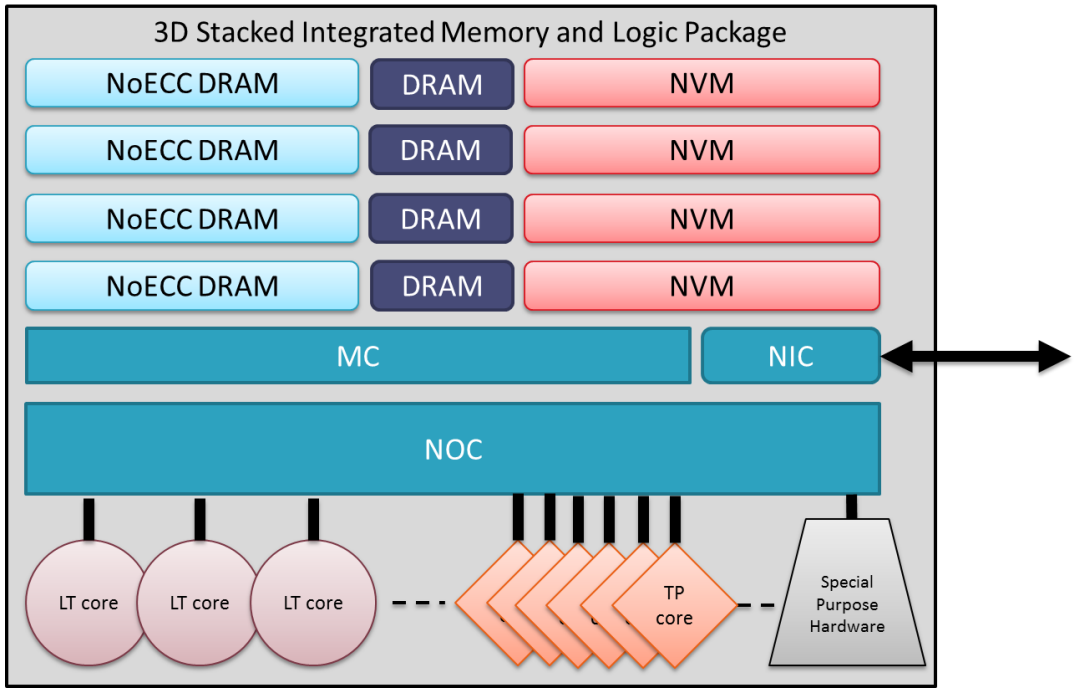
(From Exascale Arch Report 2009)

System attributes	2001	2010	"2015"		"2018"	
System peak	10 Tera	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	~0.8 MW	6 MW	15 MW		20 MW	
System memory	0.006 PB	0.3 PB	5 PB		32-64 PB	
Node performance	0.024 TF	0.125 TF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW		25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	16	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	416	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW		1.5 GB/s	150 GB/sec	1 TB/sec	250 GB/sec	2 TB/sec
MTTI		day	O(1 day)		O(1 day)	

http://science.energy.gov/ascr/news-and-resources/workshops-and-conferences/grand-challenges/

Notional Future Node Architecture

- NVM to increase memory capacity
- Mix of cores to provide different capabilities
- Integrated network interface
- Very high bandwidth, low latency to on-package locales



Blackcomb: Comparison of emerging memory technologies

Jeffrey Vetter, ORNL
Robert Schreiber, HP Labs
Trevor Mudge, University of Michigan
Yuan Xie, Penn State University

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

	SRAM	DRAM	eDRAM	NAND Flash	PCRAM	STTRAM	ReRAM (1T1R)	ReRAM (Xpoint)
Data Retention	N	N	N	Y	Y	Y	Y	Y
Cell Size (F^2)	50-200	4-6	19-26	2-5	4-10	8-40	6-20	1- 4
Read Time (ns)	< 1	30	5	10^4	10-50	10	5-10	50
Write Time (ns)	< 1	50	5	10^5	100-300	5-20	5-10	10-100
Number of Rewrites	10^{16}	10^{16}	10^{16}	10^4-10^5	10^8-10^{12}	10^{15}	10^8-10^{12}	10^6-10^{10}
Read Power	Low	Low	Low	High	Low	Low	Low	Medium
Write Power	Low	Low	Low	High	High	Medium	Medium	Medium
Power (other than R/W)	Leakage	Refresh	Refresh	None	None	None	None	Sneak

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

NVRAM Technology Continues to Improve – Driven by Market Forces

designlines MEMORY

News & Analysis

3D NAND Production Starts at Samsung

Peter Clarke
8/6/2013 08:05 AM EDT
16 comments

NO RATI
1 saves
LOGIN T

Like 17 Tweet 7 Share 10 g+1 3

LONDON — Samsung Electronics Co. Ltd. has begun n production of a 128 Gbit NAND flash memory that is inte multiple layers, and claims that it is the first company to

The memory is based on a charge-trap cell rather than conventional floating gate non-volatile cell used in 2D N In the vertical arrangement this charge-trap cell shows i reliability between a factor of 2 and a factor of 10 over conventional floating-gate NAND flash memory, Samsur in a press release.

The technology is capable of stacking up 24 layers, but did not disclose how many layers it had used in its 128 vertical NAND, nor whether the memory cells are multile whether it had relaxed the design geometry from the lea in 2D memory, which stands at about 19 or 16 nm.

The company did say that the memory would provide improvements in performance and area ratio, and a V-N is suitable for a wide range of consumer and commercial applications including embedded NAND storage and solid-state drives.

The V-NAND component has the same memory capacity as a 128

designlines MEMORY

News & Analysis

3D NAND Transition: 15nm Process Technology Takes Shape

Gary Hilson
5/13/2014 08:15 AM EDT
5 comments

Like 15 Tweet 6 Share 6

TORONTO – With 3D NAND unlikely to make at least 2015, SanDisk and its flash foundry recently announced 15nm process technolog flash.

SanDisk's 1Z-nm technology will be applied t and 3-bit-per-cell NAND flash memory archite production ramp to begin in the second half technology scales chips along both axes, an a broad range of SanDisk offerings, from ren enterprise SSDs.

Toshiba's new process replaces its 19nm pro is aimed at providing a transitional step to 3D Nelson, senior VP of Toshiba America Electro memory business unit. Toshiba's 15nm proce conjunction with improved peripheral circuitry chips that achieve the same write speed as c second generation 19nm process technolog transfer rate to 533 megabits a second -- 1.3 employing a high-speed interface.

Nelson said there is room to advance floating

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

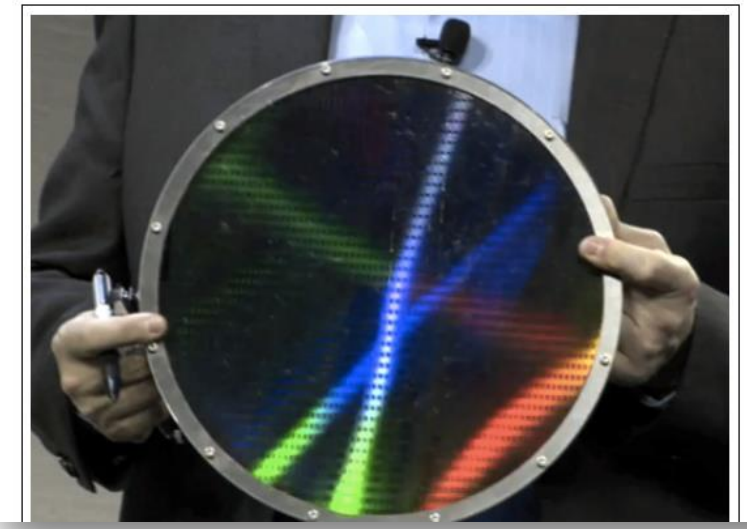
By Chris Mellor

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

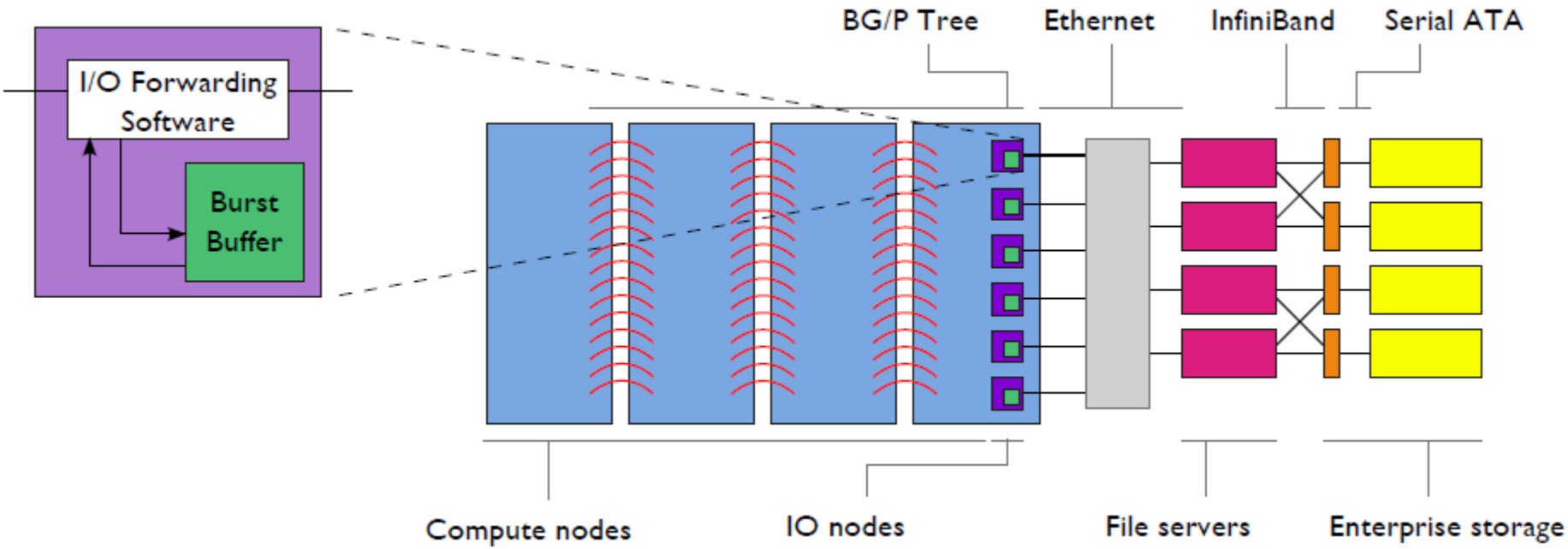
Blocks and Files HP has warned *El Reg* not to get its hopes up too high after the tech titan's CTO Martin Fink suggested StoreServ arrays could be packed with 100TB Memristor drives come 2018.

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.

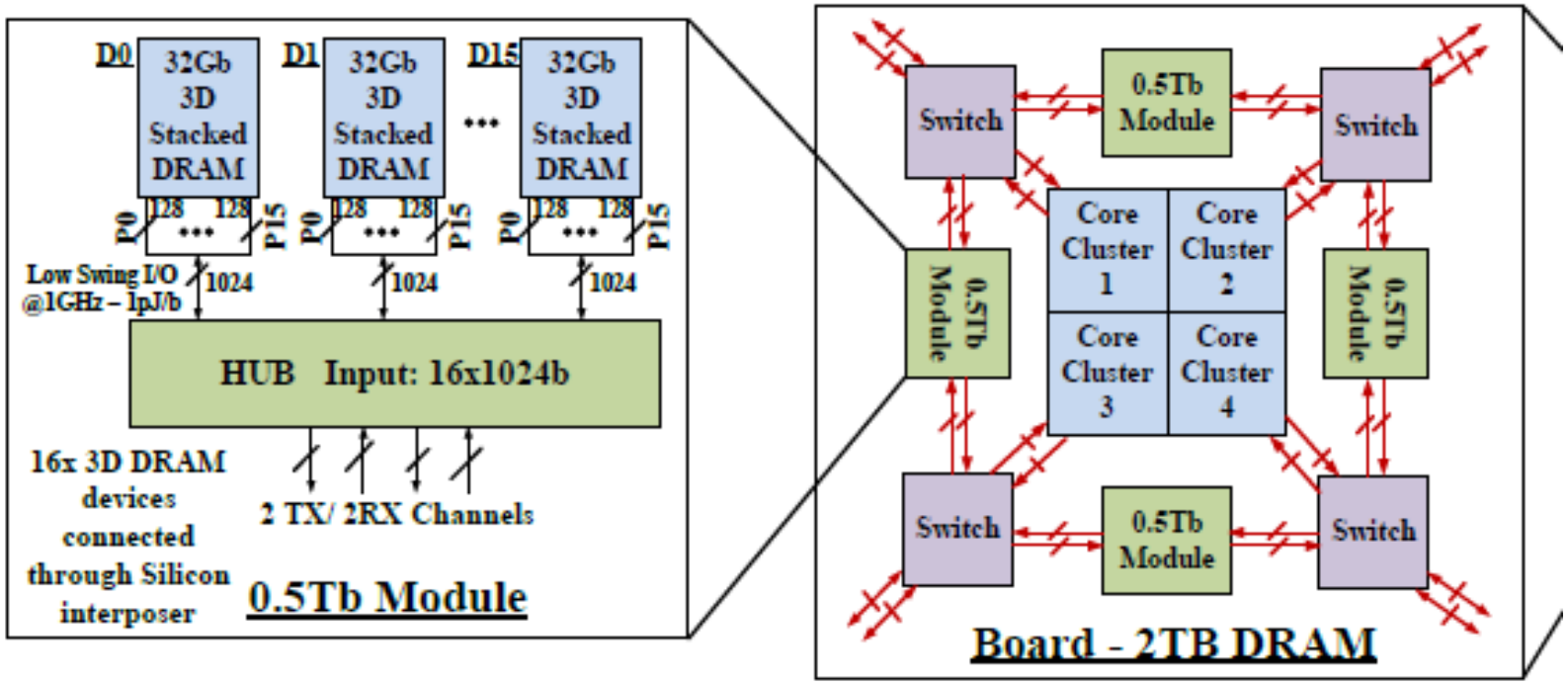


Early Uses of NVRAM: Burst Buffers



N. Liu, J. Cope, P. Carns, C. Carothers, R. Ross, G. Grider, A. Crume, and C. Maltzahn, "On the role of burst buffers in leadership-class storage systems," Proc. IEEE 28th Symposium on Mass Storage Systems and Technologies (MSST), 2012, pp. 1-11,

Tradeoffs in Exascale Memory Architectures



- Understanding the tradeoffs
  - ECC type, row buffers, DRAM physical page size, bitline length, etc

"Optimizing DRAM Architectures for Energy-Efficient, Resilient Exascale Memories," *SC13, 2013*

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Programming Interfaces Example: NV-HEAPS

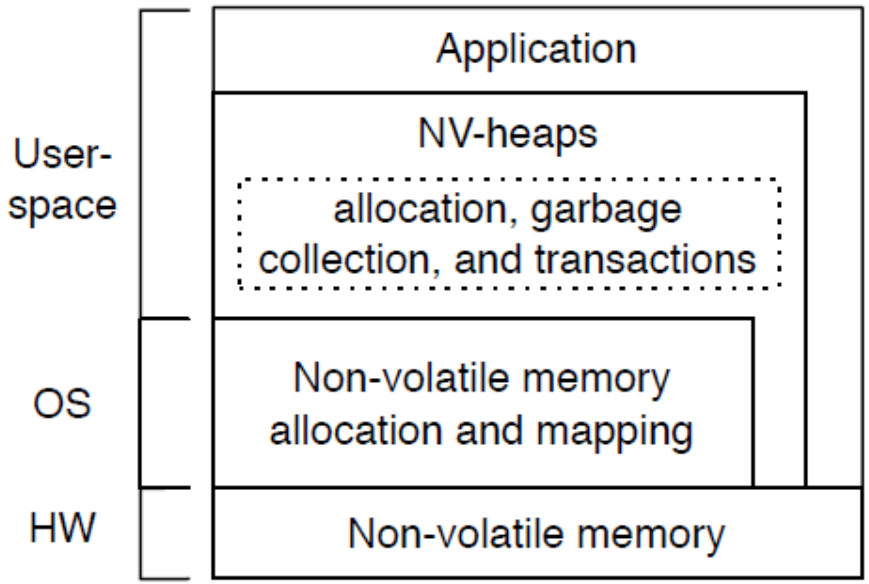


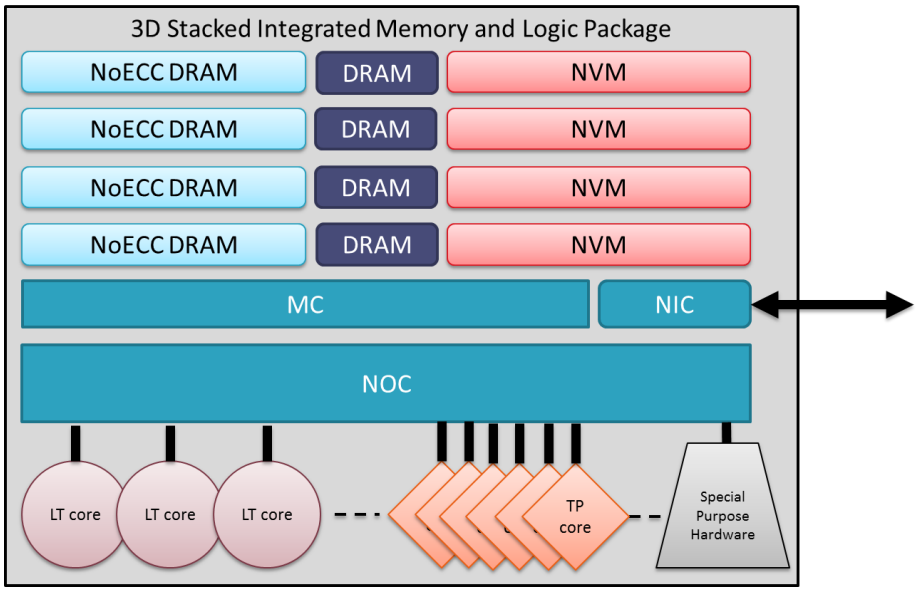
Figure 1. The NV-heap system stack This organization allows read and write operations to bypass the operating system entirely.

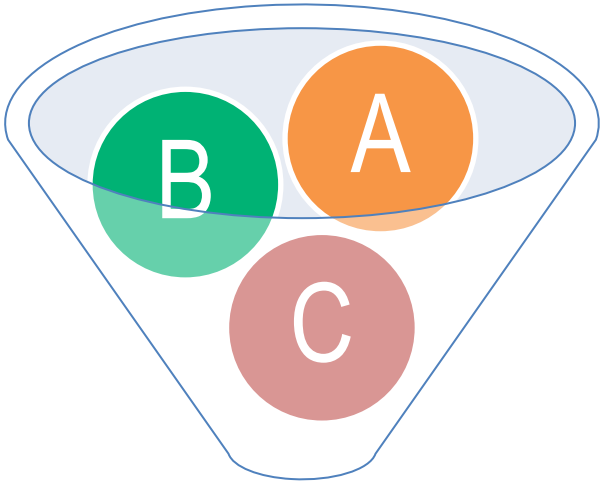
```
class NVList : public NVObject {
  DECLARE_POINTER_TYPES(NVList);
public:
  DECLARE_MEMBER(int, value);
  DECLARE_PTR_MEMBER(NVList::NVPtr, next);
};

void remove(int k)
{
  NVHeap * nv = NVHOpen("foo.nvheap");
  NVList::VPtr a =
            nv->GetRoot<NVList::NVPtr>();
  AtomicBegin {
    while(a->get_next() != NULL) {
      if (a->get_next()->get_value() == k) {
        a->set_next(a->get_next()->get_next());
      }
      a = a->get_next();
    }
  } AtomicEnd;
}
```

Figure 2. NV-heap example A simple NV-heap function that atomically removes all links with value k from a non-volatile linked list.

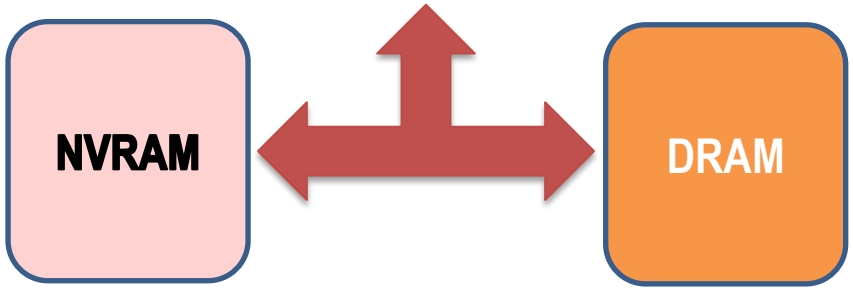
J. Coburn, A.M. Caulfield *et al.*, *"NV-Heaps: making persistent objects fast and safe with next-generation, non-volatile memories," in Proceedings of the sixteenth international conference on Architectural support for programming languages and operating systems. Newport Beach, California, USA: ACM, 2011, pp. 105-18, 10.1145/1950365.1950380.*

New hybrid memory architectures: What is the ideal organizations for our applications?





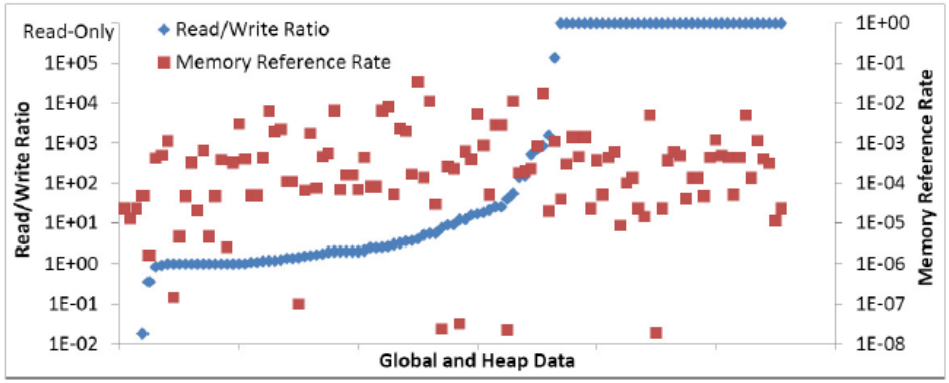
Natural separation of applications objects?



D. Li, J.S. Vetter, G. Marin, C. McCurdy, C. Cira, Z. Liu, and W. Yu, "Identifying Opportunities for Byte-Addressable Non-Volatile Memory in Extreme-Scale Scientific Applications," in *IEEE International Parallel & Distributed Processing Symposium (IPDPS). Shanghai: IEEEE, 2012*

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Measurement Results



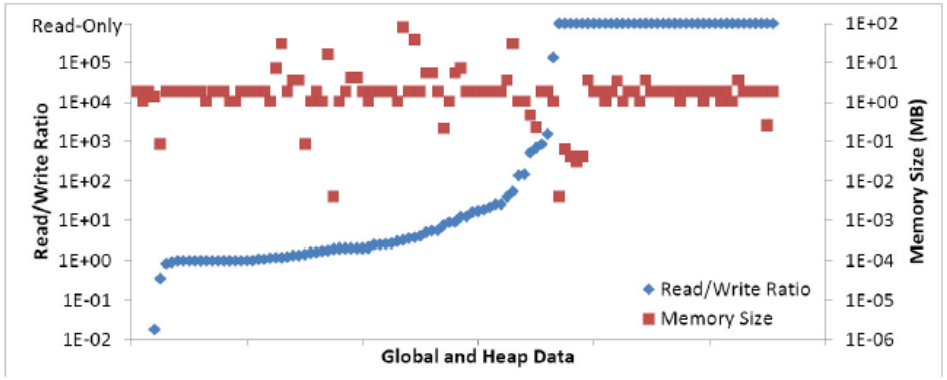
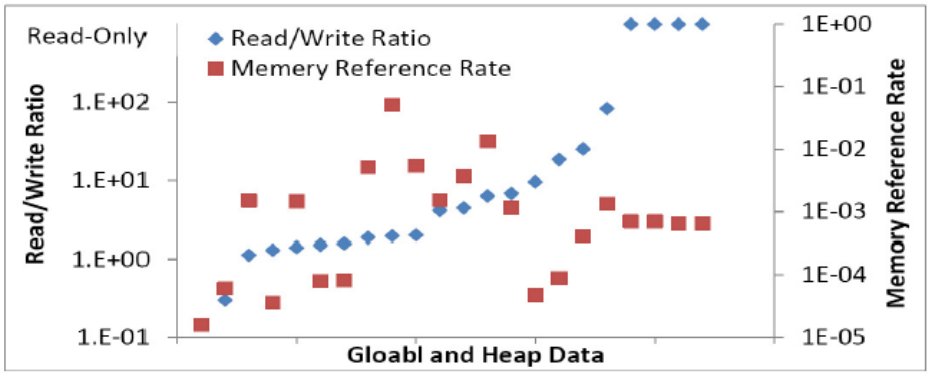


Figure 3: Read/write ratios, memory reference rates and memory object sizes for memory objects in Nek5000



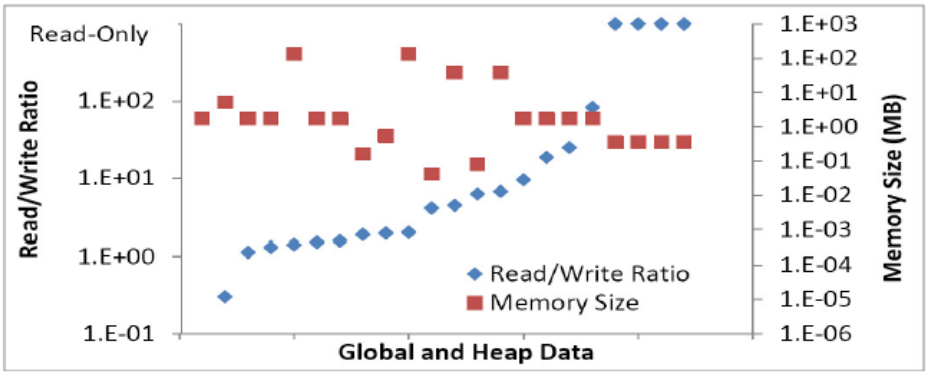


Figure 6: Read/write ratios, memory reference rates and memory object sizes for memory objects in S3D

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

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

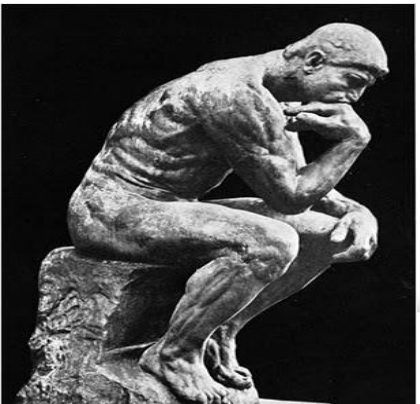
- Many lookup, index, and permutation tables
- Inverted and ‘element-lagged’ mass matrices
- Geometry arrays for grids
- Thermal conductivity for soils
- Strain and conductivity rates
- Boundary condition data
- Constants for transforms, interpolation
- …

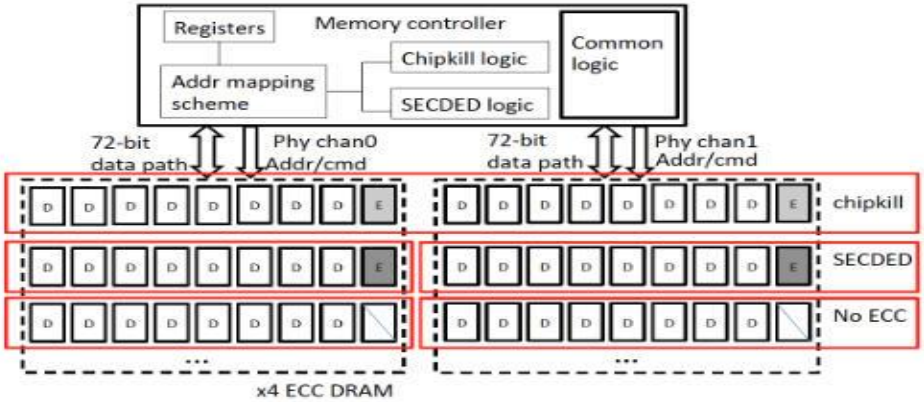
Redesigning algorithms for multi-mode memory systems

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Rethinking Algorithm-Based Fault Tolerance

- Algorithm-based fault tolerance (ABFT) has many attractive characteristics
  - Can reduce or even eliminate the expensive periodic checkpoint/rollback
  - Can bring negligible performance loss when deployed in large scale
  - No modifications from architecture and system software
- However
  - ABFT is completely opaque to any underlying hardware resilience mechanisms
  - These hardware resilience mechanisms are also unaware of ABFT
  - Some data structures are over-protected by ABFT and hardware





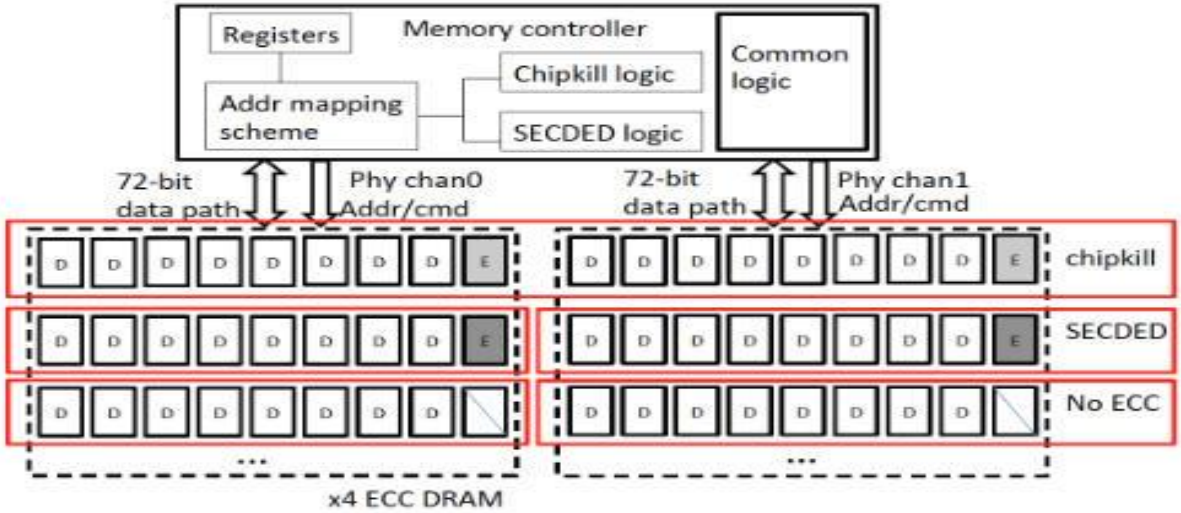
D. Li, C. Zizhong, W. Panruo, and S. Vetter Jeffrey, "Rethinking Algorithm-Based Fault Tolerance with a Cooperative Software-Hardware Approach," Proc. International Conference for High Performance Computing, Networking, Storage and Analysis (SC13), 2013,

We consider ABFT using a holistic view from both software and hardware

- We investigate how to integrate ABFT and hardware-based ECC for main memory
- ECC brings energy, performance and storage overhead
- The current ECC mechanisms cannot work
  - There is a significant semantic gap for error detection and location between ECC protection and ABFT
- We propose an explicitly-managed ECC by ABFT
  - A cooperative software-hardware approach
  - We propose customization of memory resilience mechanisms based on algorithm requirements.

System Designs

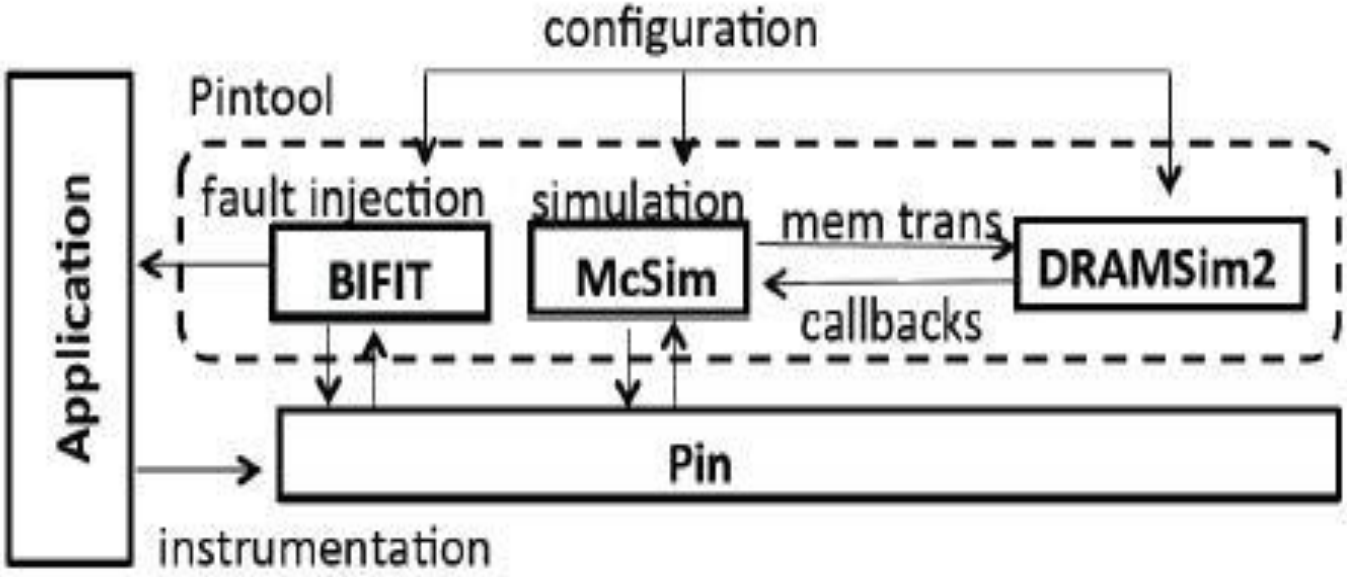
- Architecture
  - Enable co-existence of multiple ECC
  - Introduce a set of ECC registers into the memory controller (MC)
  - MC is in charge of detecting, locating, and reporting errors
- Software
  - The users control which data structures should be protected by which relaxed ECC scheme by ECC control APIs.
  - ABFT can simplify its verification phase, because hardware and OS can explicitly locate corrupted data



ATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Evaluation

- We use four ABFT (FT-DGEMM, FT-Cholesky, FT-CG and FT-HPL)



- We save up to 25% for system energy (and up to 40% for dynamic memory energy) with up to 18% performance improvement

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Future Directions in Next Generation Memory

- Next decade will be exciting for memory technology
- New devices
  - Flash, ReRam, STTRAM will challenge DRAM
  - Commercial markets already driving transition
- New configurations
  - 2.5D, 3D stacking removes recent JEDEC constraints
  - Storage paradigms (e.g., key-value)
  - Opportunities to rethink memory organization
- Logic/memory integration
  - Move compute to data
  - Programming models

	SRAM	DRAM	eDRAM	NAND Flash	PCRAM	STTRAM	ReRAM (1T1R)	ReRAM (Xpoint)
Data Retention	N	N	N	Y	Y	Y	Y	Y
Cell Size (F^2)	50-200	4-6	19-26	2-5	4-10	8-40	6-20	1- 4
Read Time (ns)	< 1	30	5	10^4	10-50	10	5-10	50
Write Time (ns)	< 1	50	5	10^5	100-300	5-20	5-10	10-100
Number of Rewrites	10^{16}	10^{16}	10^{16}	10^4-10^5	10^8-10^{12}	10^{15}	10^8-10^{12}	10^6-10^{10}
Read Power	Low	Low	Low	High	Low	Low	Low	Medium
Write Power	Low	Low	Low	High	High	Medium	Medium	Medium
Power (other than R/W)	Leakage	Refresh	Refresh	None	None	None	None	Sneak

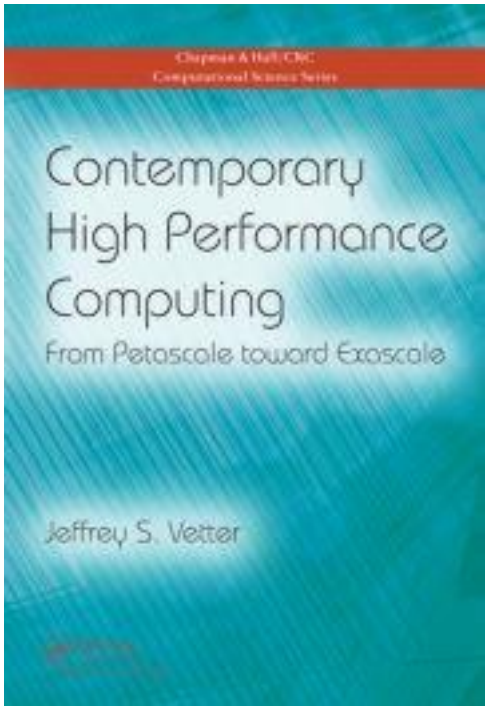
- Refactor our applications to make use of this new technology
- Add HPC programming support for these new technologies
- Explore opportunities for improved resilience, power, performance

OAK RIDGE National Laboratory

Summary

- **Our community expects major challenges in HPC as we move to extreme scale**
  - Power, Performance, Resilience, Productivity
  - Major shifts and uncertainty in architectures, software, applications
- **Applications will have to change in response to design of processors, memory systems, interconnects, storage**
  - DOE has initiated Codesign Centers that bring together all stakeholders to develop integrated solutions
- **Technologies particularly pertinent to addressing some of these challenges**
  - Heterogeneous computing
  - Nonvolatile memory
- **We need to reexamine software solutions to make this period of uncertainty palpable for computational science**
  - OpenARC
  - Memory use and allocation strategies

- **New book surveys the international landscape of HPC**
- **24 chapters with many of today's top systems/facilities: Titan, Tsubame2, BlueWaters, Tianhe-1A**



http://j.mp/YhLiQP

Q & A

More info: vetter@computer.org

Recent Publications from FTG (2012-3)

[1] F. Ahmad, S. Lee, M. Thottethodi, and T.N. VijayKumar, "MapReduce with Communication Overlap (MaRCO)," *Journal of Parallel and Distributed Computing, 2012, http://dx.doi.org/10.1016/j.jpdc.2012.12.012.*

[2] C. Chen, Y. Chen, and P.C. Roth, "DOSAS: Mitigating the Resource Contention in Active Storage Systems," in *IEEE Cluster 2012, 2012, 10.1109/cluster.2012.66.*

[3] A. Danalis, P. Luszczek, J. Dongarra, G. Marin, and J.S. Vetter, "BlackjackBench: Portable Hardware Characterization," *SIGMETRICS Performance Evaluation ReviewSIGMETRICS Performance Evaluation Review, 40, 2012,*

[4] A. Danalis, C. McCurdy, and J.S. Vetter, "Efficient Quality Threshold Clustering for Parallel Architectures," in *IEEE International Parallel & Distributed Processing Symposium (IPDPS). Shanghai: IEEEE, 2012, http://dx.doi.org/10.1109/IPDPS.2012.99.*

[5] J.M. Dennis, J. Edwards, K.J. Evans, O. Guba, P.H. Lauritzen, A.A. Mirin, A. St-Cyr, M.A. Taylor, and P.H. Worley, "CAM-SE: A scalable spectral element dynamical core for the Community Atmosphere Model," *International Journal of High Performance Computing Applications, 26:74–89, 2012, 10.1177/1094342011428142.*

[6] J.M. Dennis, M. Vertenstein, P.H. Worley, A.A. Mirin, A.P. Craig, R. Jacob, and S.A. Mickelson, "Computational Performance of Ultra-High-Resolution Capability in the Community Earth System Model," *International Journal of High Performance Computing Applications, 26:5–16, 2012, 10.1177/1094342012436965.*

[7] K.J. Evans, A.G. Salinger, P.H. Worley, S.F. Price, W.H. Lipscomb, J. Nichols, J.B.W. III, M. Perego, J. Edwards, M. Vertenstein, and J.-F. Lemieux, "A modern solver framework to manage solution algorithm in the Community Earth System Model," *International Journal of High Performance Computing Applications, 26:54–62, 2012, 10.1177/1094342011435159.*

[8] S. Lee and R. Eigenmann, "OpenMPC: Extended OpenMP for Efficient Programming and Tuning on GPUs," *International Journal of Computational Science and Engineering, 8(1), 2013,*

[9] S. Lee and J.S. Vetter, "Early Evaluation of Directive-Based GPU Programming Models for Productive Exascale Computing," in *SC12: ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis. Salt Lake City, Utah, USA: IEEE press, 2012, http://dl.acm.org/citation.cfm?id=2388996.2389028, http://dx.doi.org/10.1109/SC.2012.51.*

[10] D. Li, B.R. de Supinski, M. Schulz, D.S. Nikolopoulos, and K.W. Cameron, "Strategies for Energy Efficient Resource Management of Hybrid Programming Models," *IEEE Transaction on Parallel and Distributed SystemsIEEE Transaction on Parallel and Distributed Systems, 2013, http://dl.acm.org/citation.cfm?id=2420628.2420808,*

[11] D. Li, D.S. Nikolopoulos, and K.W. Cameron, "Modeling and Algorithms for Scalable and Energy Efficient Execution on Multicore Systems," in *Scalable Computing: Theory and Practice, U.K. Samee, W. Lizhe et al., Eds.: Wiley & Sons, 2012,*

[12] D. Li, D.S. Nikolopoulos, K.W. Cameron, B.R. de Supinski, E.A. Leon, and C.-Y. Su, "Model-Based, Memory-Centric Performance and Power Optimization on NUMA Multiprocessors," in *International Symposium on Workload Characterization. San Diego, 2012, http://www.computer.org/csdl/proceedings/iiswc/2012/4531/00/06402921-abs.html*

[13] D. Li, J.S. Vetter, G. Marin, C. McCurdy, C. Cira, Z. Liu, and W. Yu, "Identifying Opportunities for Byte-Addressable Non-Volatile Memory in Extreme-Scale Scientific Applications," in *IEEE International Parallel & Distributed Processing Symposium (IPDPS). Shanghai: IEEEE, 2012, http://dl.acm.org/citation.cfm?id=2358563, http://dx.doi.org/10.1109/IPDPS.2012.89.*

Recent Publications from FTG (2012-3)

[14] D. Li, J.S. Vetter, and W. Yu, "Classifying Soft Error Vulnerabilities in Extreme-Scale Scientific Applications Using a Binary Instrumentation Tool," in *SC12: ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis. Salt Lake City, 2012, http://dl.acm.org/citation.cfm?id=2388996.2389074, http://dx.doi.org/10.1109/SC.2012.29.*

[15] Z. Liu, B. Wang, P. Carpenter, D. Li, J.S. Vetter, and W. Yu, "PCM-Based Durable Write Cache for Fast Disk I/O," in *IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS). Arlington, Virginia, 2012, http://www.computer.org/csdl/proceedings/mascots/2012/4793/00/4793a451-abs.html*

[16] G. Marin, C. McCurdy, and J.S. Vetter, "Diagnosis and Optimization of Application Prefetching Performance," in *ACM International Conference on Supercomputing (ICS). Euguene, OR: ACM, 2013*

[17] J.S. Meredith, S. Ahern, D. Pugmire, and R. Sisneros, "EAVL: The Extreme-scale Analysis and Visualization Library," in *Proceedings of the Eurographics Symposium on Parallel Graphics and Visualization (EGPGV), 2012*

[18] J.S. Meredith, R. Sisneros, D. Pugmire, and S. Ahern, "A Distributed Data-Parallel Framework for Analysis and Visualization Algorithm Development," in *Proceedings of the 5th Annual Workshop on General Purpose Processing with Graphics Processing Units. New York, NY, USA: ACM, 2012, pp. 11–9, http://doi.acm.org/10.1145/2159430.2159432, 10.1145/2159430.2159432.*

[19] A.A. Mirin and P.H. Worley, "Improving the Performance Scalability of the Community Atmosphere Model," *International Journal of High Performance Computing Applications, 26:17–30, 2012, 10.1177/1094342011412630.*

[20] P.C. Roth, "The Effect of Emerging Architectures on Data Science (and other thoughts)," in *2012 CScADS Workshop on Scientific Data and Analytics for Extreme-scale Computing. Snowbird, UT, 2012, http://cscads.rice.edu/workshops/summer-2012/data-analytics*

[21] K. Spafford, J.S. Meredith, S. Lee, D. Li, P.C. Roth, and J.S. Vetter, "The Tradeoffs of Fused Memory Hierarchies in Heterogeneous Architectures," in *ACM Computing Frontiers (CF). Cagliari, Italy: ACM, 2012, http://dl.acm.org/citation.cfm?id=2212924, http://dx.doi.org/10.1145/2212908.2212924.*

[22] K. Spafford and J.S. Vetter, "Aspen: A Domain Specific Language for Performance Modeling," in *SC12: ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis, 2012, http://dl.acm.org/citation.cfm?id=2388996.2389110, http://dx.doi.org/10.1109/SC.2012.20.*

[23] C.-Y. Su, D. Li, D.S. Nikolopoulos, M. Grove, K.W. Cameron, and B.R. de Supinski, "Critical Path-Based Thread Placement for NUMA Systems," *ACM SIGMETRICS Performance Evaluation ReviewACM SIGMETRICS Performance Evaluation Review, 40, 2012, http://dl.acm.org/citation.cfm?id=2381056.2381079,*

[24] V. Tipparaju and J.S. Vetter, "GA-GPU: Extending a Library-based Global Address Space Programming Model for Scalable Heterogeneous Computing Systems," in *ACM Computing Frontiers (CF), 2012, http://dx.doi.org/10.1145/2212908.2212918.*

[25] J.S. Vetter, *Contemporary High Performance Computing: From Petascale Toward Exascale, vol. 1, 1 ed. Boca Raton: Taylor and Francis, 2013, http://j.mp/RrBdPZ,*

[26] J.S. Vetter, R. Glassbrook, K. Schwan, S. Yalamanchili, M. Horton, A. Gavrilovska, M. Slawinska, J. Dongarra, J. Meredith, P.C. Roth, K. Spafford, S. Tomov, and J. Wynkoop, "Keeneland: Computational Science using Heterogeneous GPU Computing," in *Contemporary High Performance Computing: From Petascale Toward Exascale, vol. 1, CRC Computational Science Series, J.S. Vetter, Ed., 1 ed. Boca Raton: Taylor and Francis, 2013, pp. 900,*

[27] W. Yu, X. Que, V. Tipparaju, and J.S. Vetter, "HiCOO: Hierarchical cooperation for scalable communication in Global Address Space programming models on Cray XT systems," *Journal of Parallel and Distributed ComputingJournal of Parallel and Distributed Computing, 2012.*

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY