Blackcomb

Jeffrey Vetter

Presented to

SOS16, Santa Barbara

Oak Ridge National Laboratory

Georgia Tech College of Computing
Computational Science and Engineering

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

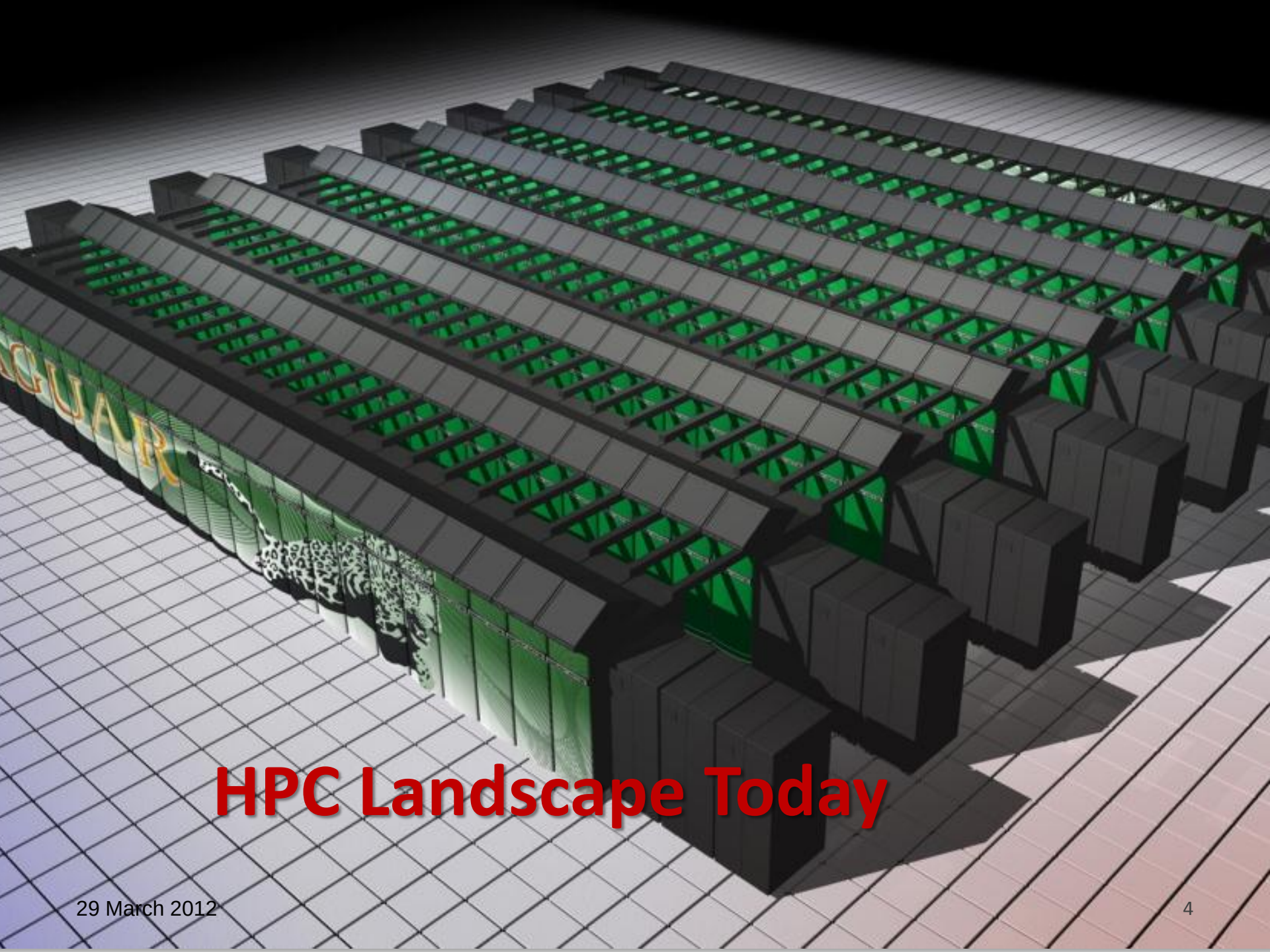
Acknowledgements

- **Contributors**
  - Future Technologies Group: http://ft.ornl.gov
  - NSF Keeneland Project: http://keeneland.gatech.edu
  - DOE
    - 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
  - International Exascale Software Project: http://www.exascale.org/iesp/Main_Page
  - DARPA NVIDIA Echelon

- **Sponsors**
  - NVIDIA
    - CUDA Center of Excellence
  - US National Science Foundation
  - US Department of Energy Office of Science
  - US DARPA

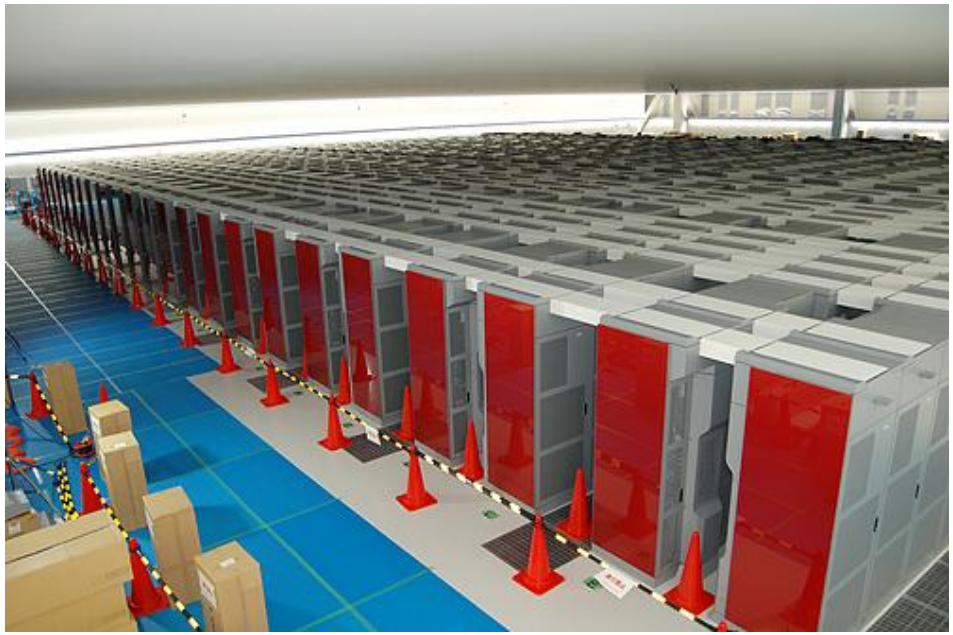
Highlights

- **Memory capacity and bandwidth will be a major challenge for Exascale**
- **New NVRAM technologies may offer a solution**
  - Many open questions, however
- **Blackcomb: how does NVRAM fit into Exascale?**
- **Initial architecture and applications results are promising**



HPC Landscape Today

RIKEN/Fujitsu K: #1 in November 2011



- **10.5 PF (93% of peak)**
  - 12.7 MW
- **705,024 cores**
- **1.4 PB memory**

Source: Riken, Fujitsu

ORNL's "Titan" System

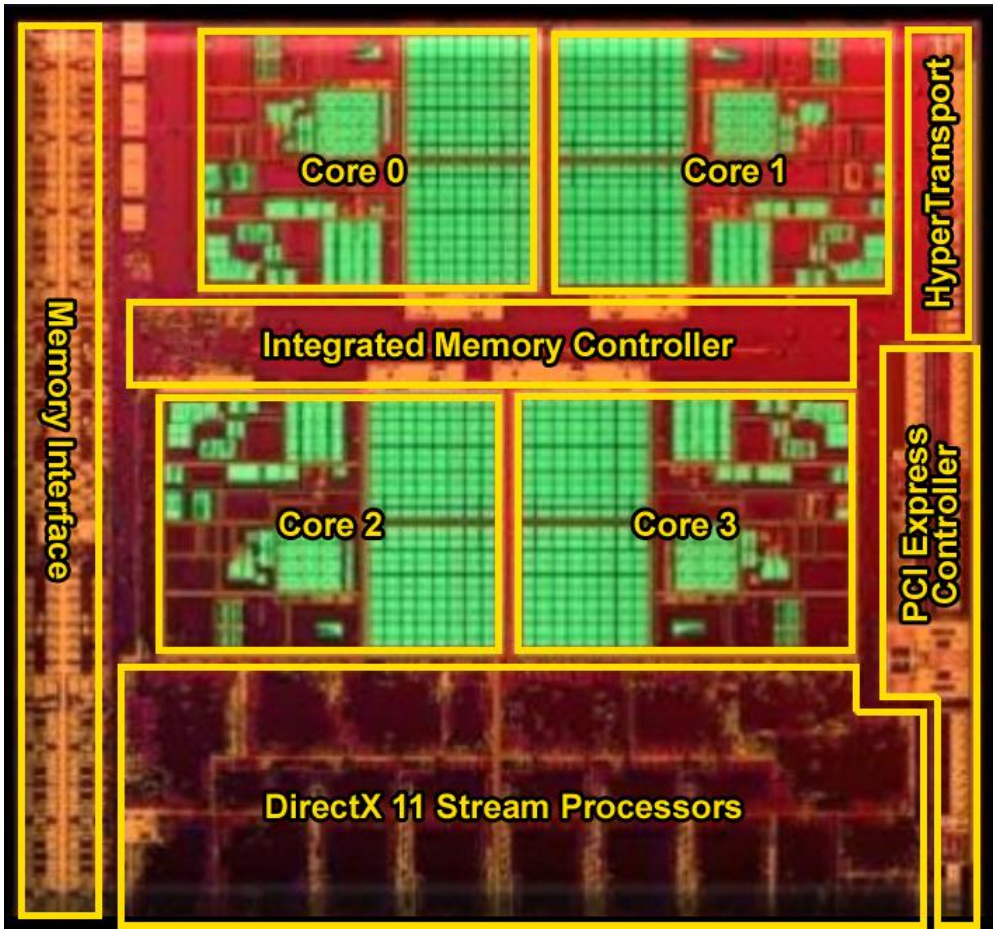
- Upgrade of existing Jaguar Cray XT5
- Cray Linux Environment operating system
- Gemini interconnect
  - 3-D Torus
  - Globally addressable memory
  - Advanced synchronization features
- AMD Opteron 6200 processor (Interlagos)
- New accelerated node design using NVIDIA multi-core accelerators
  - 2011: 960 NVIDIA M2090 "Fermi" GPUs
  - 2012: 10-20 PF NVIDIA "Kepler" GPUs
- 10-20 PFlops peak performance
  - Performance based on available funds
- 600 TB DDR3 memory (2x that of Jaguar)

Titan Specs	
Compute Nodes	18,688
Login & I/O Nodes	512
Memory per node	32 GB + 6 GB
NVIDIA "Fermi" (2011)	665 GFlops
# of Fermi chips	960
NVIDIA "Kepler" (2012)	>1 TFlops
Opteron	2.2 GHz
Opteron performance	141 GFlops
Total Opteron Flops	2.6 PFlops
Disk Bandwidth	~ 1 TB/s

Source: OLCF

AMD's Llano: A-Series APU

- **Combines**
  - 4 x86 cores
  - Array of Radeon cores
  - Multimedia accelerators
  - Dual channel DDR3
- **32nm**
- **Up to 29 GB/s memory bandwidth**
- **Up to 500 Gflops SP**
- **45W TDP**



Source: AMD

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/

Challenges to ~~Exascale~~

Performance Growth

1) **System power** is the primary constraint
2) **Memory** bandwidth and capacity are not keeping pace
3) **Concurrency** (1000x today)
4) **Processor** architecture is an open question
5) **Programming model** heroic compilers will not hide this
6) **Algorithms** need to minimize data movement, not flops
7) **I/O bandwidth** unlikely to keep pace with machine speed
8) **Reliability and resiliency** will be critical at this scale
9) **Bisection bandwidth** limited by cost and energy

Unlike the last 20 years most of these (1-7) are equally important across scales, e.g., 100 10-PF machines

Source: Hitchcock, Exascale Research Kickoff Meeting

Memory Bandwidth and Capacity

Critical Concern : Memory Capacity

	2010	2018	Factor Change
System peak	2 Pf/s	1 Ef/s	500
Power	6 MW	20 MW	3
System Memory	0.3 PB	10 PB	33
Node Performance	0.125 Tf/s	10 Tf/s	80
Node Memory BW	25 GB/s	400 GB/s	16
Node Concurrency	12 CPUs	1,000 CPUs	83
Interconnect BW	1.5 GB/s	50 GB/s	33
System Size (nodes)	20 K nodes	1 M nodes	50
Total Concurrency	225 K	1 B	4,444
Storage	15 PB	300 PB	20
Input/Output bandwidth	0.2 TB/s	20 TB/s	100

Table 1: Potential Exascale Computer Design for 2018 and its relationship to current HPC designs.

- **Small memory capacity has profound impact on other features**

- **Feeding the core(s)**
- **Poor efficiencies**
- **Small messages, I/O**

New Technologies May Offer a Solution

Device Type	HDD	DRAM	NAND Flash	FRAM	MRAM	STTRAM	PCRAM	NRAM
Maturity	Product	Product	Product	Product	Product	Prototype	Product	Prototype
Present Density	400Gb/in^2 [7]	8Gb/chip [9]	64Gb/chip [10]	128Mb/chip	32Mb/chip	2Mb/chip	512Mb/chip	NA
Cell Size (SLC)	(2/3)F^2	6F^2	4F^2	6F^2	20F^2	4F^2	5F^2	5F^2
MLC Capability	No	No	4bits/cell	No	2bits/cell	4bits/cell	4bits/cell	No
Program Energy/bit	NA	2pJ	10nJ	2pJ	120pJ	0.02pJ	100pJ	10pJ [11]
Access Time (W/R)	9.5/8.5ms [8]	10/10ns	200/25us	50/75ns	12/12ns	10/10ns	100/20ns	10/10ns [11]
Endurance/Retention	NA	10^{16}/64ms	10^5/10yr	10^{15}/10yr	10^{16}/10yr	10^{16}/10yr	10^5/10yr	10^{16}/10yr

Device Type	RRAM	CBRAM	SEM	Polymer	Molecular	Racetrack	Holographic	Probe
Maturity	Research	Prototype	Prototype	Research	Research	Research	Product	Prototype
Present Density	64Kb/chip	2Mb/chip	128Mb/chip	128b/chip	160Kb/chip	NA	515Gb/in^2	1Tb/in^2
Cell Size	6F^2	6F^2	4F^2	6F^2	6F^2	N/A	N/A	N/A
MLC Capability	2bits/cell	2bits/cell	No	2bits/cell	No	12bits/cell	N/A	N/A
Program Energy/bit	2pJ	2pJ	13pJ	NA	NA	2pJ	N/A	100pJ [12]
Access Time (W/R)	10/20ns	50/50ns	100/20ns	30/30ns	20/20ns	10/10ns	3.1/5.4ms	10/10us
Endurance/Retention	10^8/10yr	10^8/Months	10^9/days	10^4/Months	10^5/Months	10^{16}/10yr	10^5/50yr	10^5/NA

M.H. Kryder et al., “After Hard Drives,” IEEE Trans. Mag., 45(10):3406-13, 2009.

To More than Exascale HPC …

Several Open Questions

- **Which technologies will pan out?**
  - Manufacturability, economics, system integration, timeline, device endurance
- **How should NVRAM be integrated into an Exascale system?**
  - Disk replacement
  - Hybrid DRAM-NVRAM main memory
  - Something else?
- **How can applications make use of NVRAM**
  - With no/minor changes to the application?
  - With major changes to the application?
- **How can the system software and programming environment support this capability and device characteristics?**

Blackcomb Overview

Blackcomb Project Overview

Applications
- ORNL

Software and Programming Model
- HP, ORNL

System Architecture
- HP, Michigan, ORNL

Memory Architecture
- PSU, HP, Michigan

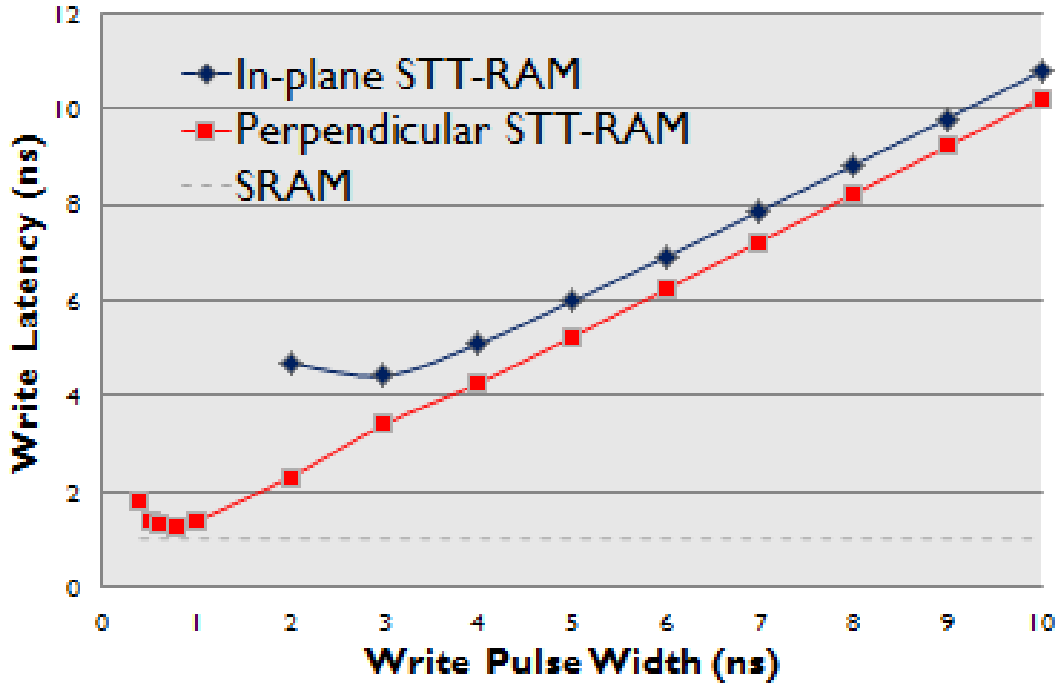
Device Technology
- PSU

- **Jeffrey Vetter, ORNL**
- **Robert Schreiber, HP Labs**
- **Trevor Mudge, Michigan**
- **Yuan Xie, PSU**

Device-Architecture

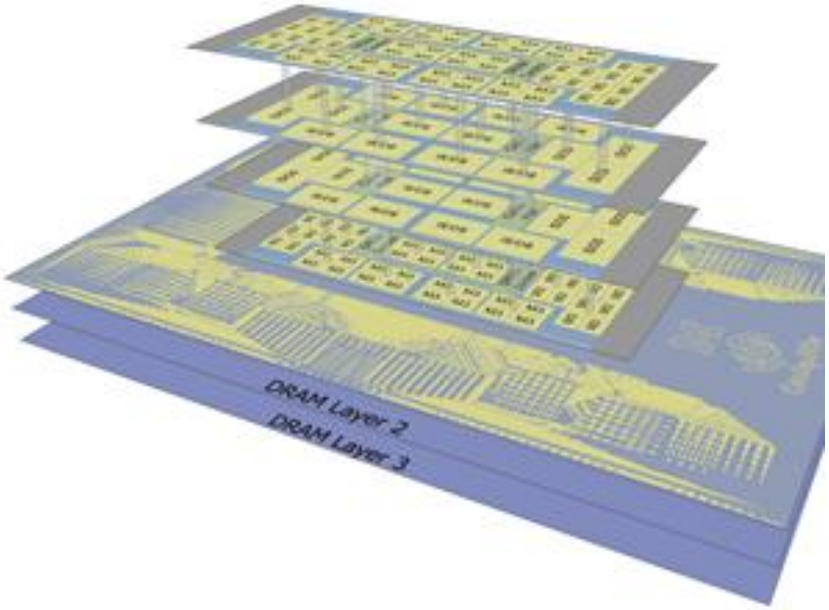
Write Latency

- 2MB STT-RAM macros and SRAM macro

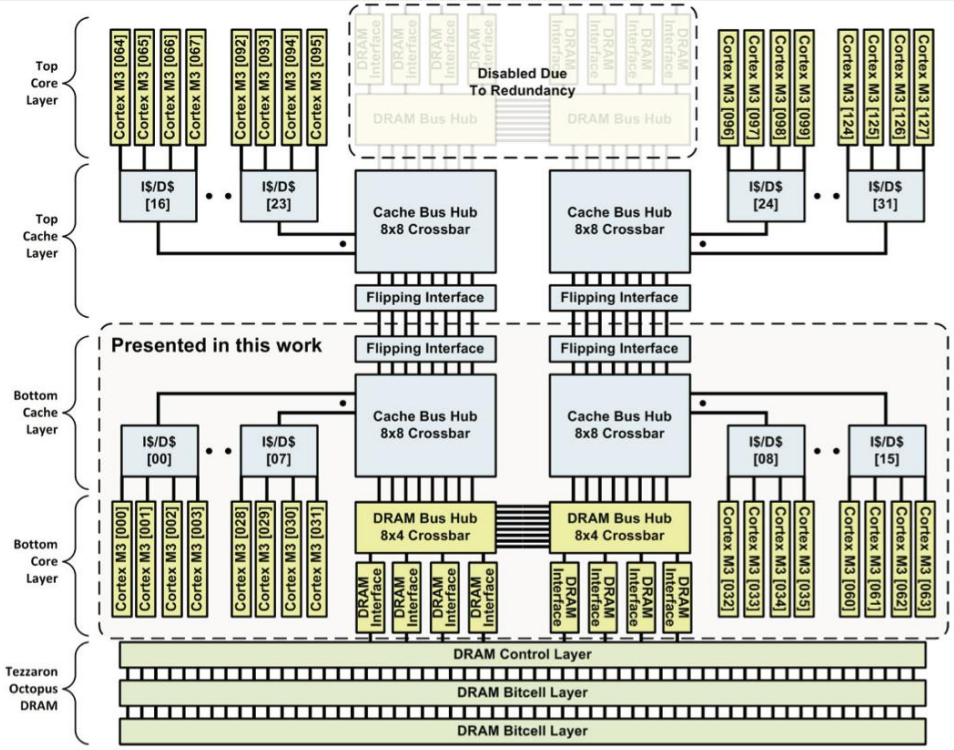


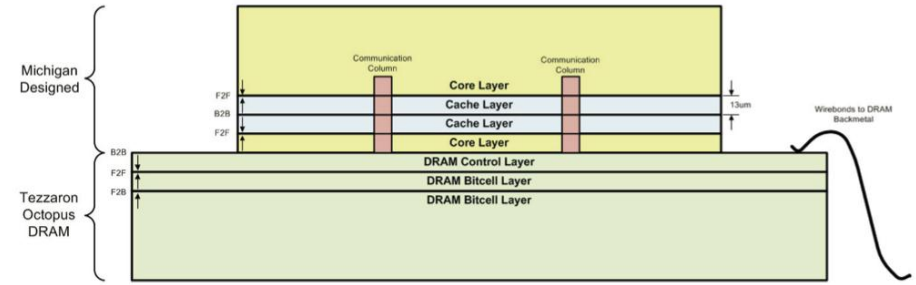
C. Xu et al., "Device-Architecture Co-Optimization of STT-RAM Based Memory Design," **ICCAD 2011 .**

Centip3De - 3D NTC Architecture



† near threshold computing





Courtesy: Mudge, Michigan from D. Fick, R. Dreslinski et al., "Centip3De: A 3930DMIPS/W Configurable Near-Threshold 3D Stacked System with 64 ARM Cortex-M3 Cores," in IEEE International Solid-State Circuits Conference (ISSCC). San Francisco: IEEE, 2012, http://www.isscc.org (Sponsored by ARM and NSF)

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

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

How can applications make use of NVRAM?

How could/should NVRAM be integrated into an Exascale system?

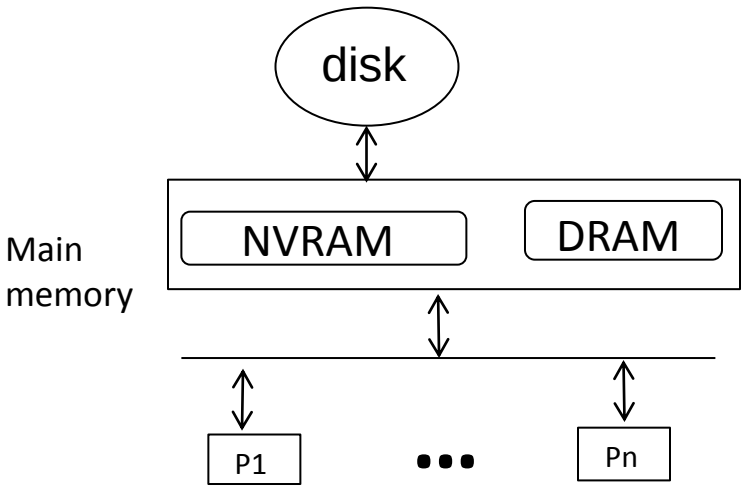
- **Traditional I/O**
  - Local disk replacement
  - I/O Buffer
  - Checkpoint Buffer
  - In-situ analysis
- **Most of the capability and characteristics of NVRAM are hidden by system abstractions**
- **Integrate with memory hierarchy**
- **Where?**
  - Cache
  - DRAM
  - New level of memory backing DRAM?

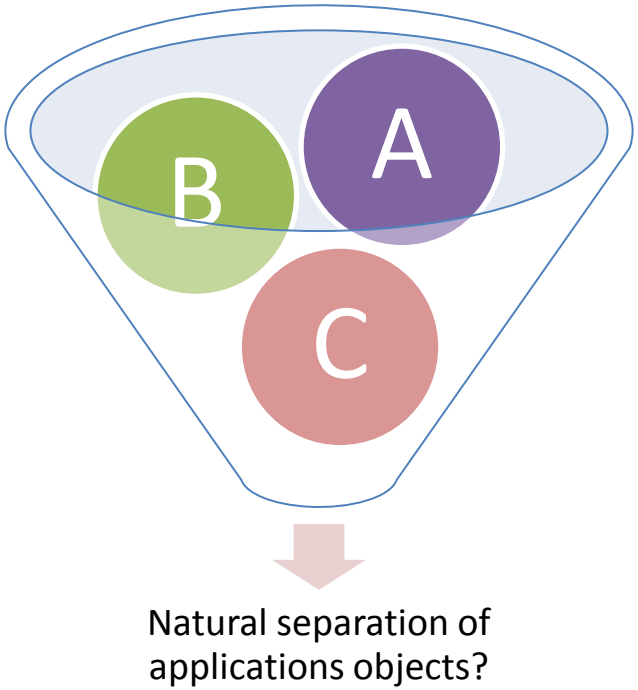
Setup

- **Assume that main memory has two partitions: DRAM and NVRAM**
- **Byte addressable**
- **Similar memory infrastructure otherwise**
- **In general, NVRAM devices have**
  - Near zero standby power
  - Higher latencies and energy for writes

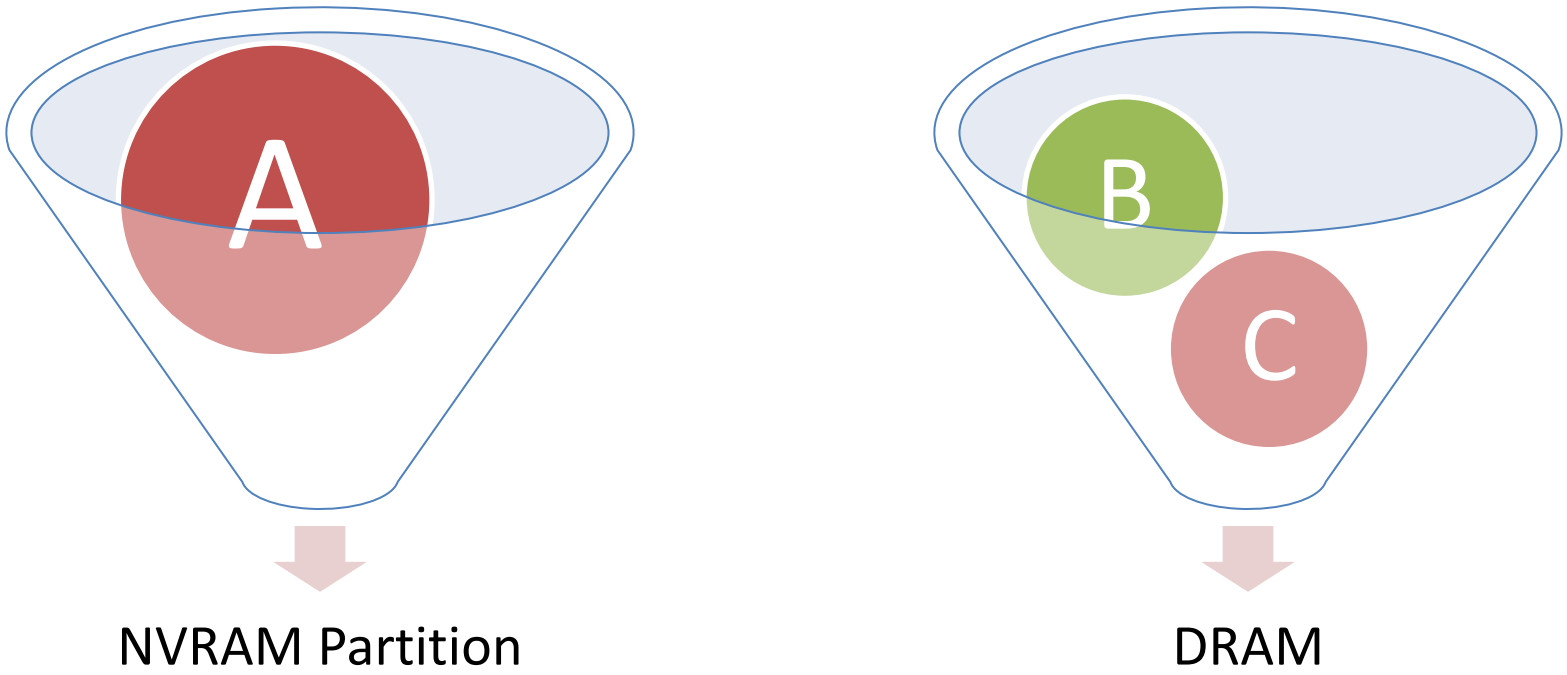
D. Li, J.S. Vetter, G. Marin, C. McCurdy, C. Cira, Z. Liu, and W. Yu, "Identifying Opportunities for Byte-Addressable Non-Volatile Memory in Extreme-Scale Scientific Applications," in *IEEE International Parallel & Distributed Processing Symposium (IPDPS). Shanghai: IEEEE, 2012*

Natural separation of applications objects for a hybrid DRAM-NVRAM configuration?





Can we identify this separation with some basic metrics?

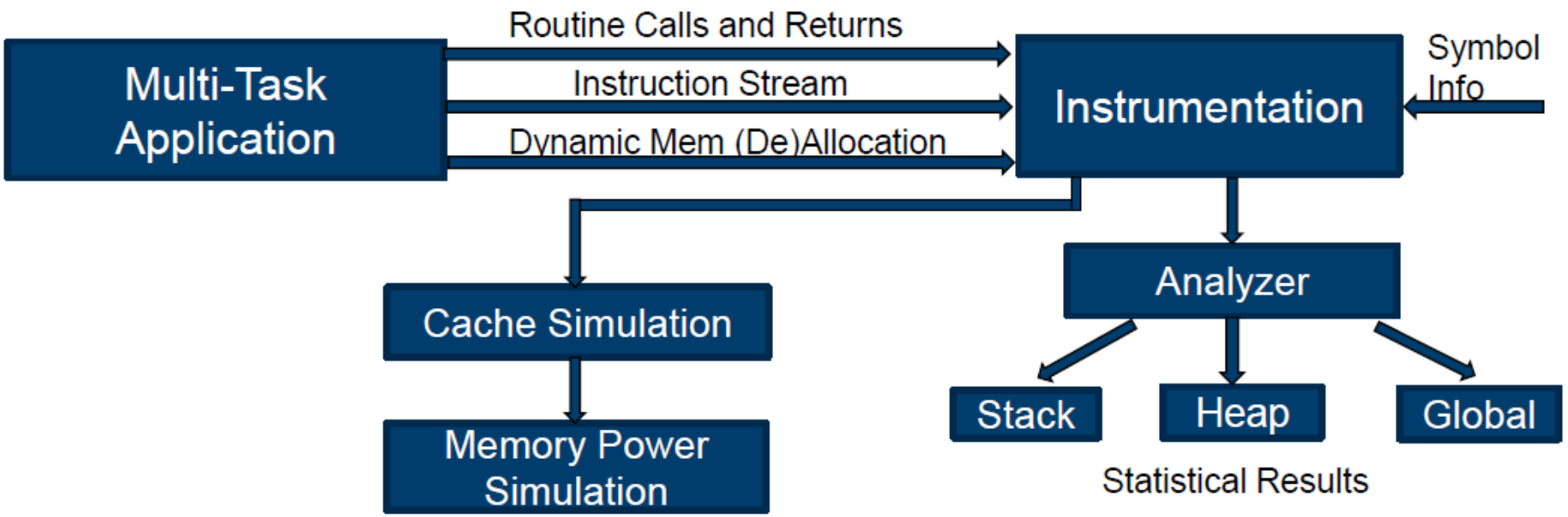


Our Metrics and Rationale

- **Empirical**
  - Object size – size matters for power
  - Reference rate – will it be cached
  - Read/write ratio – writes are expensive
- **Simulated**
  - Performance – needs to be competitive
  - Power – reduce it (prime directive)

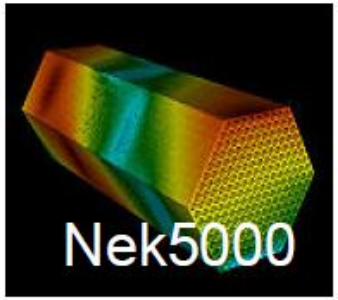
Methodology: NV-Scavenger

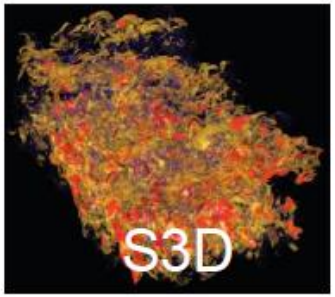
- Empirically measure and simulate memory behavior of real applications
- Gather statistics per app object for locality, read/write ratios, etc

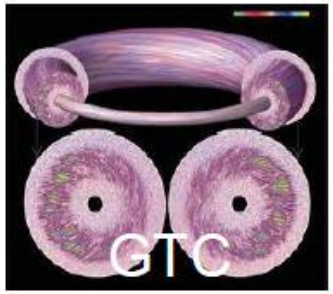


Focus on DOE workload

- Co-design center and other DOE applications
- Unmodified applications
- Runs in natural environment
- Initialization and finalization elided







Measurement Results

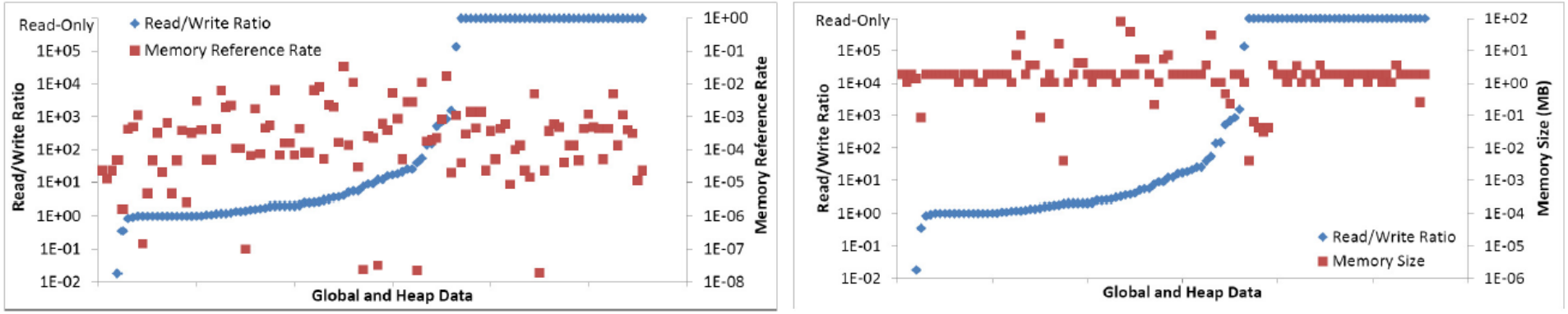


Figure 3: Read/write ratios, memory reference rates and memory object sizes for memory objects in Nek5000

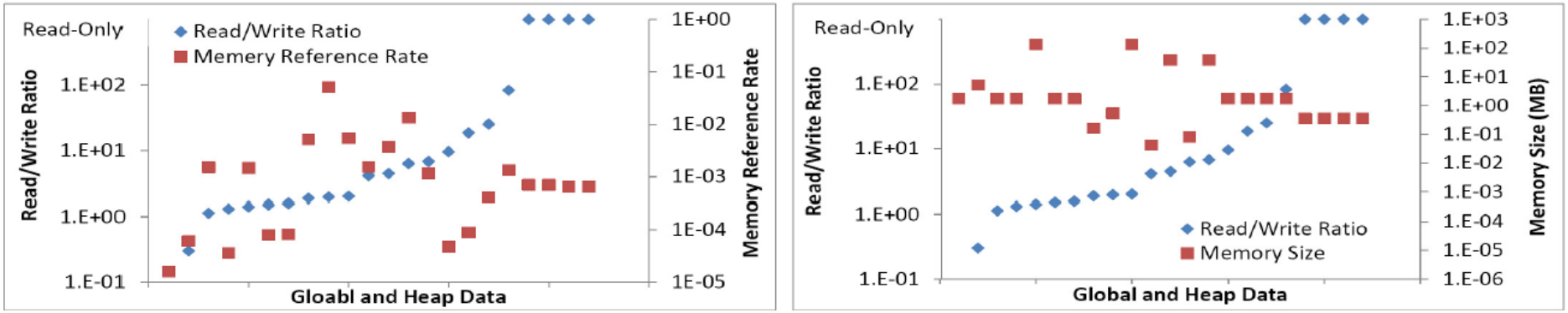


Figure 6: Read/write ratios, memory reference rates and memory object sizes for memory objects in S3D

Observations: Numerous characteristics of applications are a 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

Based on this evidence …

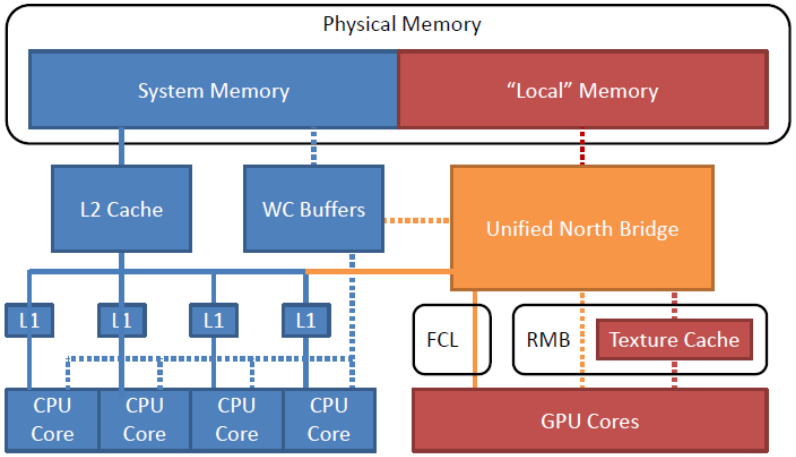
- **It appears that most scientific applications have a set of objects that could easily be mapped into NVRAM**
  - Without modification
  - With no/little performance impact
  - With lower overall power
- **Application co-design may increase this ratio significantly**

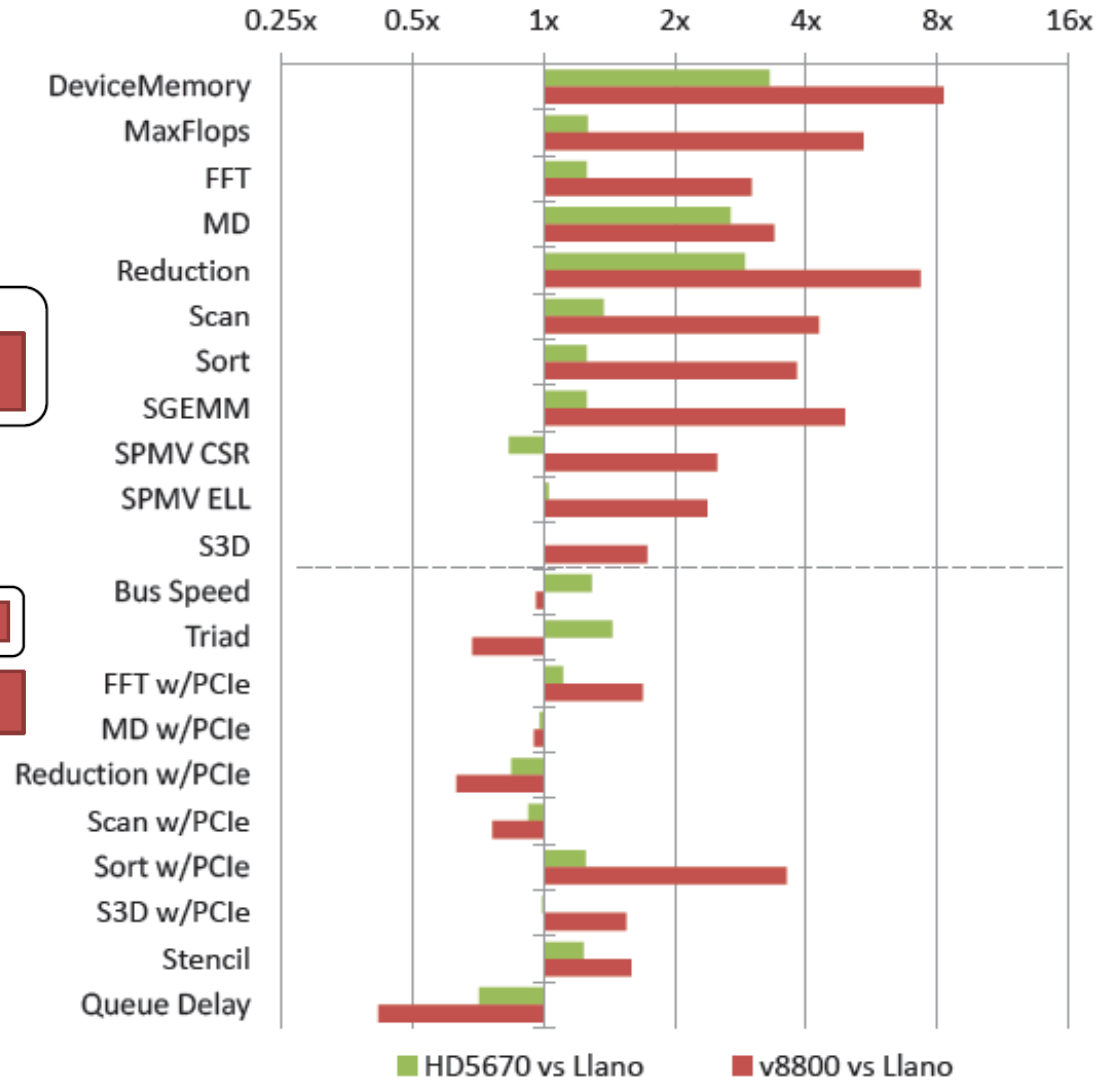
Current Status

Current Applications and Software Tasks

- **Continue working on software support for exploiting byte-addressable NVRAM**
  - Programming models
  - Compiler support
- **Understand future memory hierarchies and NVRAM insertion points**
- **Understand how apps might change in response to algorithms and other technology constraints**

Memory integration forces important decisions: Evaluation of Llano v. other options



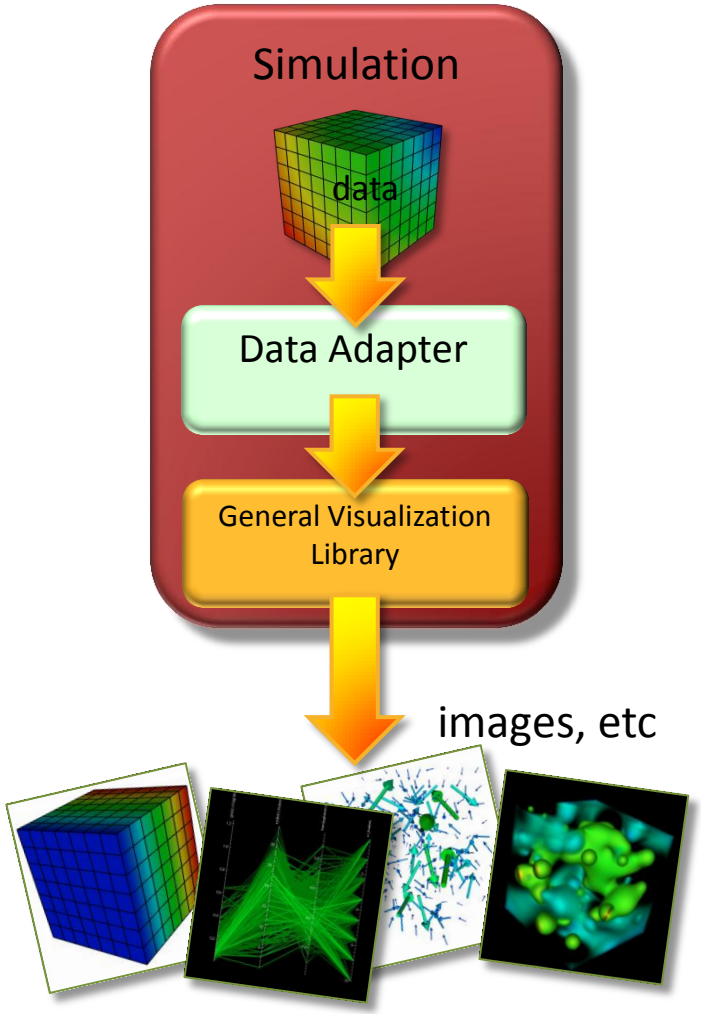


Note: Llano is a consumer, not server, part.

To appear: 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

Tightly Coupled General In Situ Processing

- **Simulation uses data adapter layer to make data suitable for general purpose visualization library**
- **Rich feature set can be called by the simulation**
- **Operate directly on the simulation's data arrays when possible**
- ***Write once, use many times***



B. Whitlock, J. Favre, and J.S. Meredith, "Parallel In Situ Coupling of a Simulation with a Fully Featured Visualization System," in *Eurographics Symposium on Parallel Graphics and Visualization (EGPGV) in association with Eurographics, 2011*

Bonus Slides

FAQ

Blackcomb: Hardware-Software Co-design for Non-Volatile Memory in Exascale Systems

Jeffrey Vetter, ORNL
Robert Schreiber, HP Labs
Trevor Mudge, U Michigan
Yuan Xie, PSU

A comparison of various memory technologies

	SRAM	DRAM	NAND Flash	PC-RAM	STT-RAM	R-RAM
Data Retention	N	N	Y	Y	Y	Y
Memory Cell Factor (F^2)	50-120	6-10	2-5	6-12	4-20	<1
Read Time (ns)	1	30	50	20-50	2-20	<50
Write / Erase Time (ns)	1	50	10^6-10^5	50-120	2-20	<100
Number of Rewrites	10^{16}	10^{16}	10^5	10^{10}	10^{15}	10^{15}
Power Read/Write	Low	Low	High	Low	Low	Low
Power (Other than R/W)	Leakage Current	Refresh Power	None	None	None	None

Novel Ideas

- **New resilience-aware designs for non-volatile memory applications**
  - Mechanical-disk-based data-stores are completely replaced with energy-efficient non-volatile memories (NVM).
  - Most levels of the hierarchy, including DRAM and last levels of SRAM cache, are completely eliminated.
- **New energy-aware systems/applications for non-volatile memories (nanostores)**
  - Compute capacity, comprised of balanced low-power simple cores, is co-located with the data store.

Impact and Champions

- **Reliance on NVM addresses device scalability, energy efficiency and reliability concerns associated with DRAM**
  - **More memory** – NVM scalability and density permits significantly more memory/core than projected by current Exascale estimates.
  - **Less power** – NVMs require zero stand-by power.
  - **More reliable** – alleviates increasing DRAM soft-error rate problem.
- **Node architecture with persistent storage near processing elements enables new computation paradigms**
  - Low-cost checkpoints, easing checkpoint frequency concerns.
  - Inter-process data sharing, easing in-situ analysis (UQ, Visualization)

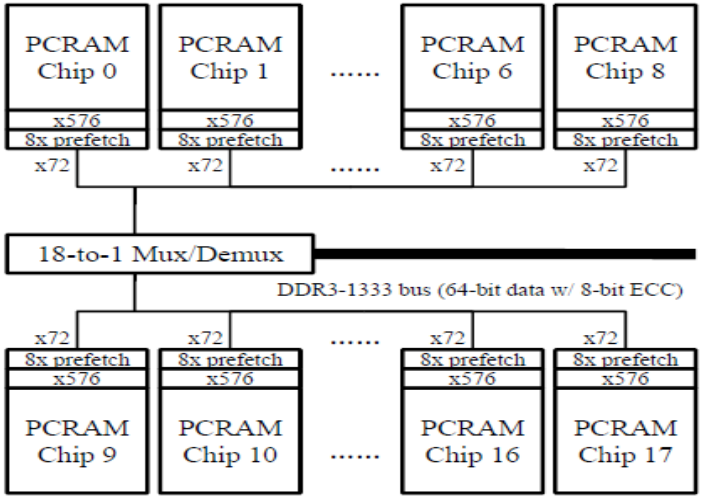
Milestones

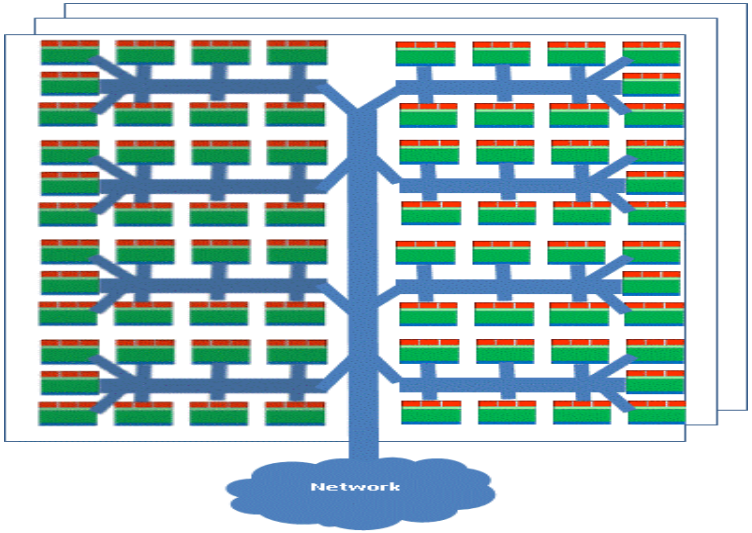
- Identify and evaluate the most promising non-volatile memory (NVM) device technologies.
- Explore assembly of NVM technologies into a storage and memory stack
- Build the abstractions and interfaces that allow software to exploit NVM to its best advantage
- Propose an exascale HPC system architecture that builds on our new memory architecture
- Characterize key DOE applications and investigate how they can benefit from these new technologies

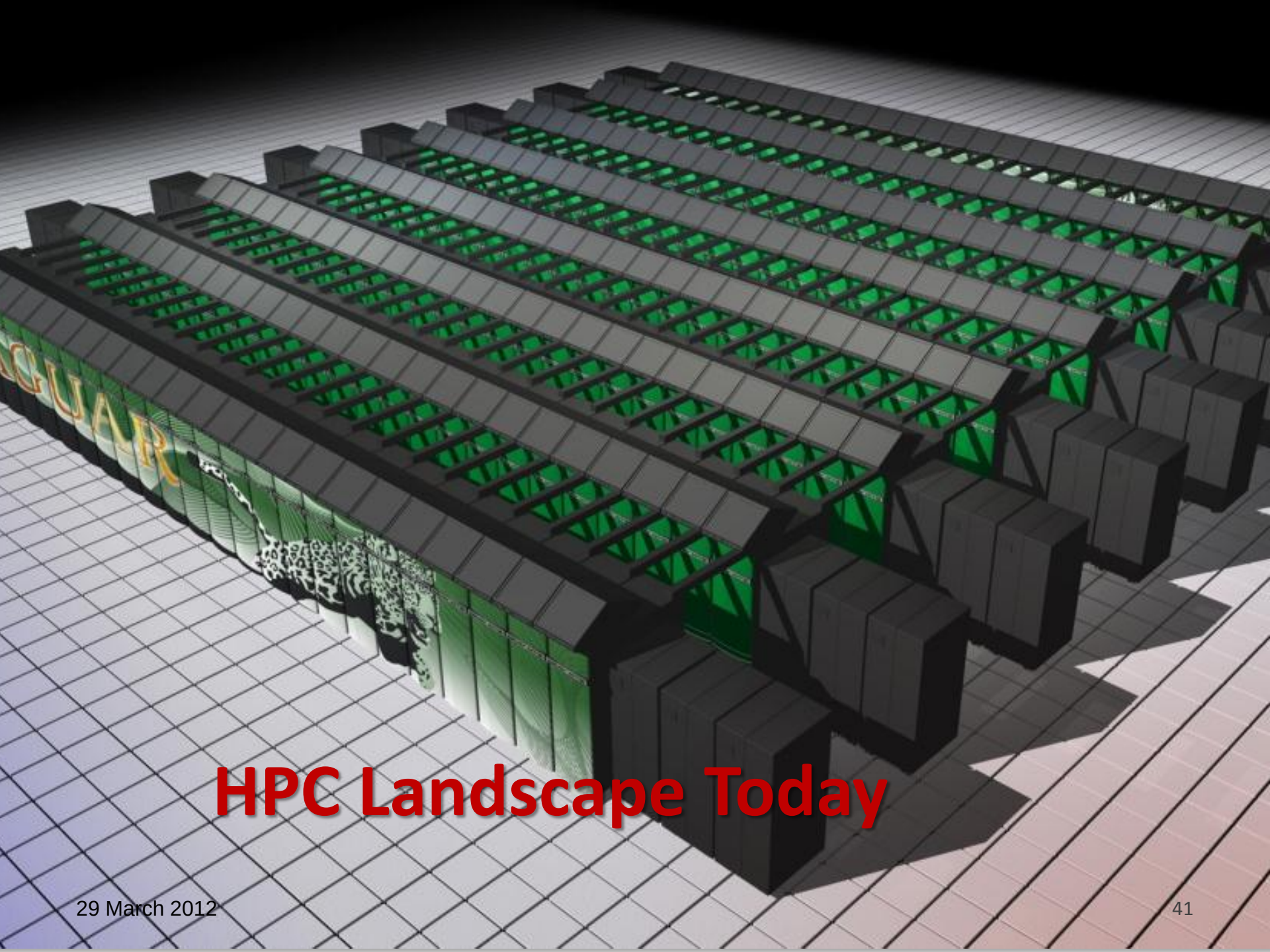
4/20/2011

Opportunities go far beyond a plugin replacement for disk drives…

- **New distributed computer architectures that address exascale resilience, energy, and performance requirements**
  - replace mechanical-disk-based data-stores with energy-efficient non-volatile memories
  - explore opportunities for NVM memory, from plug-compatible replacement (like the NV DIMM, below) to radical, new data-centric compute hierarchy (nanostores)
  - place low power compute cores close to the data store
  - reduce number of levels in the memory hierarchy
- **Adapt existing software systems to exploit this new capabilities**

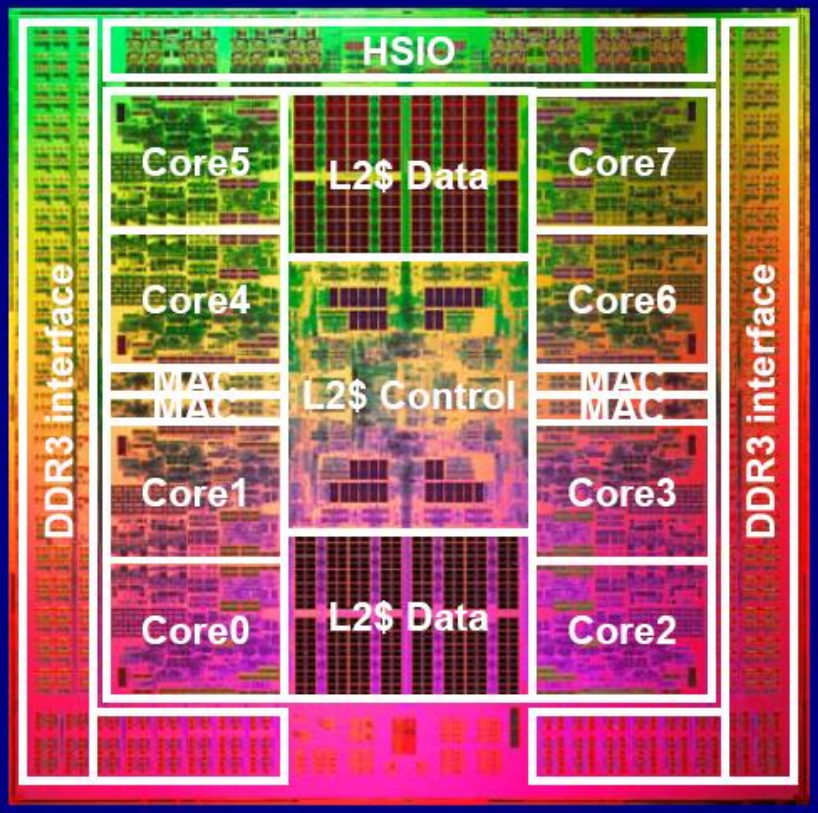






HPC Landscape Today

SPARC64™ VIIIfx Chip Overview



- **Architecture Features**
  - 8 cores
  - Shared 5 MB L2$
  - Embedded Memory Controller
  - 2 GHz
- **Fujitsu 45nm CMOS**
  - 22.7mm x 22.6mm
  - 760M transistors
  - 1271 signal pins
- **Performance (peak)**
  - 128GFlops
  - 64GB/s memory throughput
- **Power**
  - 58W (TYP, 30℃)
  - Water Cooling – Low leakage power and High reliability

#2: Tianhe-1A uses 7,000 NVIDIA GPUs

- **Tianhe-1A uses**
  - 7,168 NVIDIA Tesla M2050 GPUs
  - 14,336 Intel Westmeres
- **Performance**
  - 4.7 PF peak
  - 2.5 PF sustained on HPL
- **4.04 MW**
  - If Tesla GPU's were not used in the system, the whole machine could have needed 12 megawatts of energy to run with the same performance, which is equivalent to 5000 homes
- **Custom fat-tree interconnect**
  - 2x bandwidth of Infiniband QDR

EDITION: INTERNATIONAL | U.S. | MÉXICO | ARABIC

Set edition preference

CNN

Home | Video | World | U.S. | Africa | Asia | Europe | Latin America | Middle East | Business | W

World's fastest supercomputer belongs to China

Mashable

By **Stan Schroeder**, Mashable

October 28, 2010 -- Updated 1406 GMT (2206 HKT) | Filed under: Innovation

COURTESY NVIDIA

The supercomputer was unveiled yesterday at the Annual Meeting of National High Performance Computing.

STORY HIGHLIGHTS

- Tianhe-1A unveiled Wednesday at HPC China 2010 in Beijing
- Supercomputer has a performance record of 2.507 petaflops
- Tianhe-1A designed by the National University of Defense Technology
- System cost $88 million and its 103 cabinets weigh 155 tons

RELATED TOPICS

China

Computer Technology

(Mashable) -- The United States no longer owns the world's fastest supercomputer.

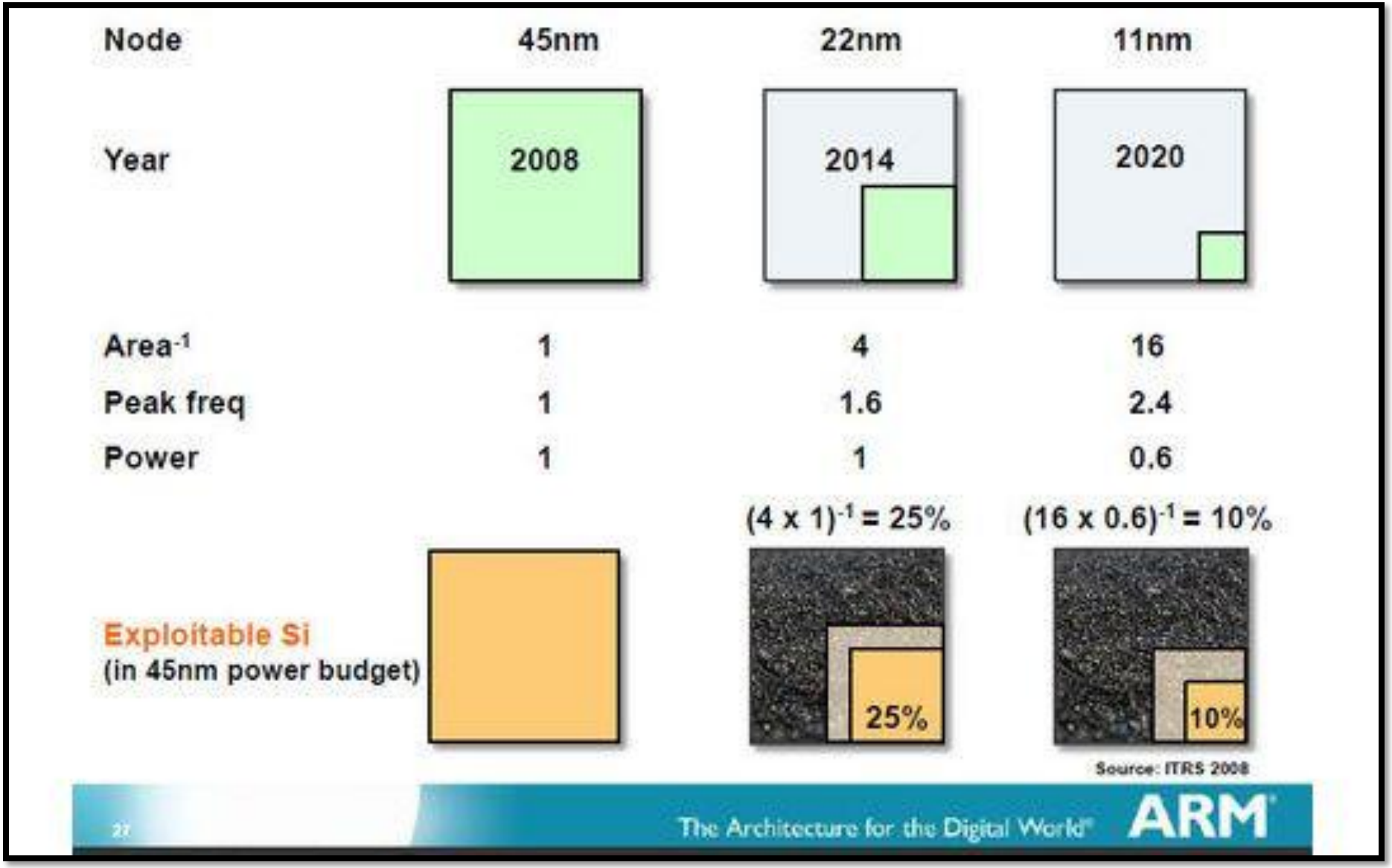
A computer called Tianhe-1A, unveiled on Wednesday at a conference in Beijing, China, can run calculations faster than the previous speed leader, a computer at a U.S. lab in Tennessee.

The new computer set a performance record by crunching 2.507 petaflops of data at once. The previous leader, a computer called Cray XT5 Jaguar and located at the Oak Ridge National Laboratory, completed 1.75 petaflop calculations.

Analysts say the new record underscores China's place as a global tech leader.

Trend #1: Facilities and Power

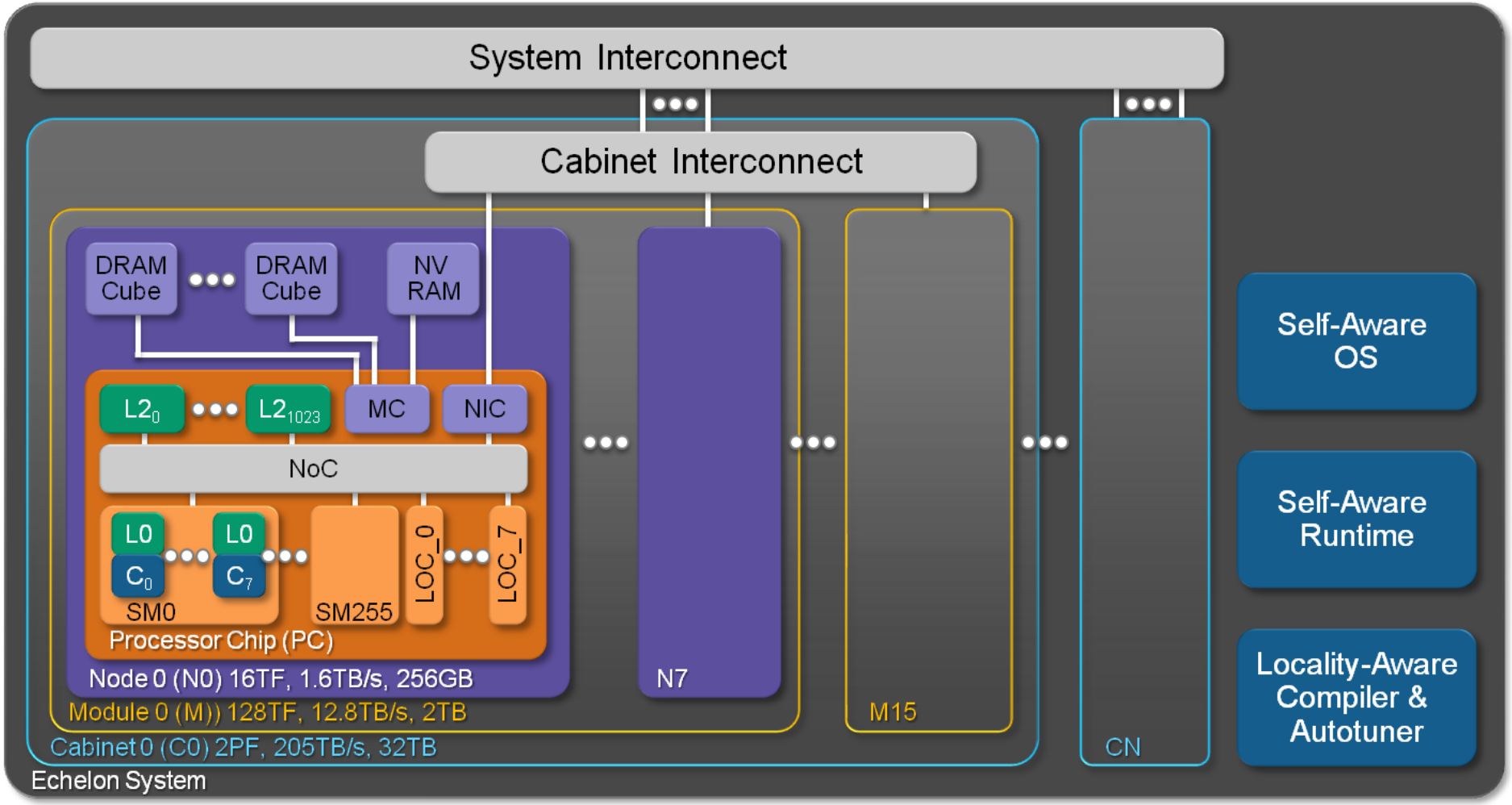
Trend #2: Dark Silicon Will Make Heterogeneity and Specialization More Relevant



Source: ARM

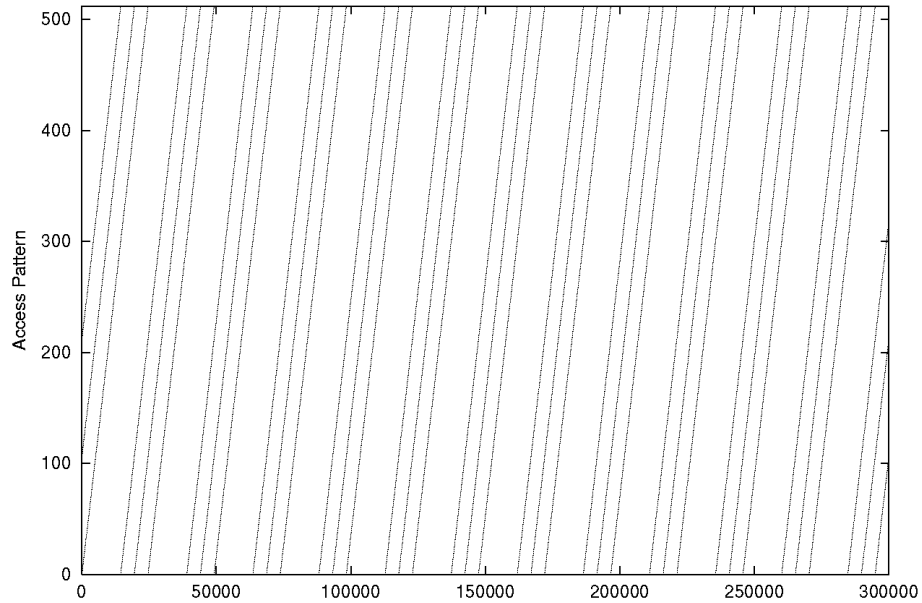
Looking Forward to Exascale

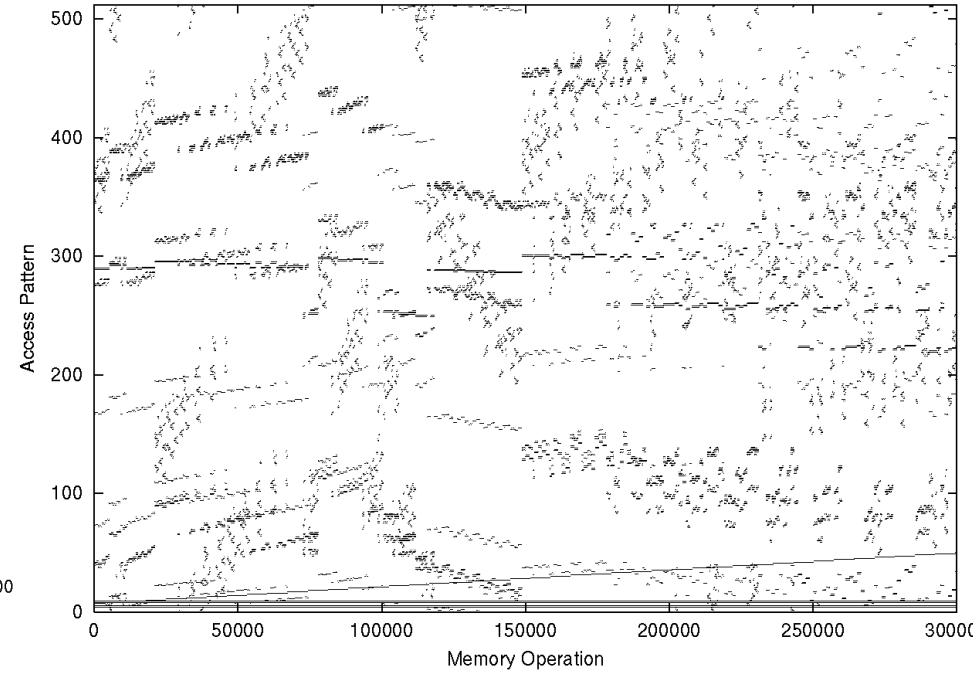
NVIDIA Echelon System Sketch

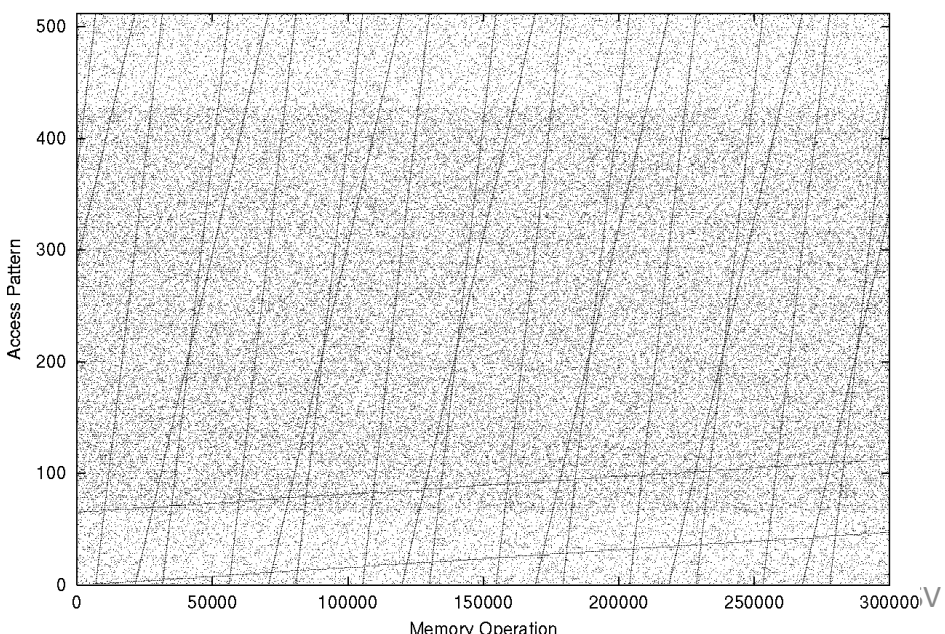


DARPA Echelon team: NVIDIA, ORNL, Micron, Cray, Georgia Tech, Stanford, UC-Berkeley, U Penn, Utah, Tennessee, Lockheed Martin

Application Memory Access Patterns



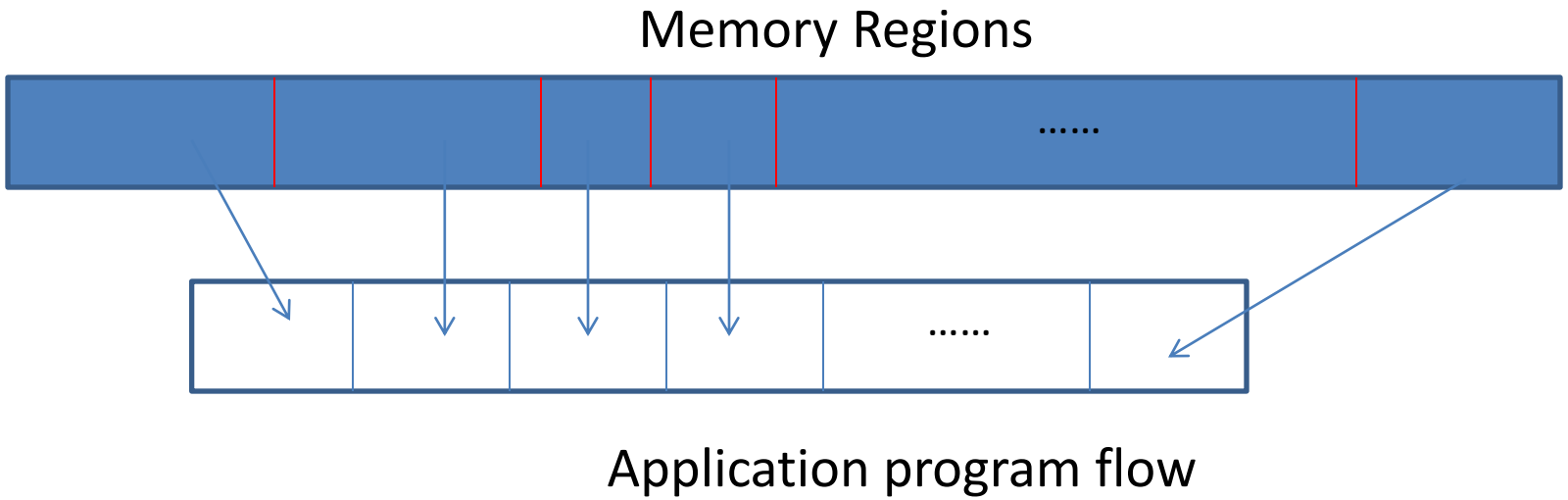




- **What features do we need?**
  - Scatter/Gather
  - Streaming load/store
  - Atomics, Transactions

NVRAM-Friendly Access Patterns

- **Not frequently written data**
- **Have good memory locality**
  - Memory references hit caches
- **"Streaming processing" pattern**



RWT

- **Target: capture NVRAM-friendly access patterns**
- **Read-Write-Total (RWT)**
  - Read/write ratio matters
  - Absolute number of memory references matters too

MR_R: the number of read references to a memory region

MR_w: the number of write references to a memory region

$T_{R/W}$: the total number of read and write references to all mem regions

Absolute number matters too!

Ratio matters!

$$RWT = \left(\frac{MR_R + MR_w}{T_{R/W}}\right) log \frac{MR_R}{MR_W}$$

Make RWT independent of app

Read major, or write major?

CMCER

- **RWT: an architecture-independent metric**
- **Last level cache miss + cache eviction ratio (CMCER)**
  - An architecture-dependent metric
  - Reflect the number of accesses to a memory region
  - Identify which memory region accesses the main memory most

MR_{CM}: the number of access to a memory region due to the last level cache miss

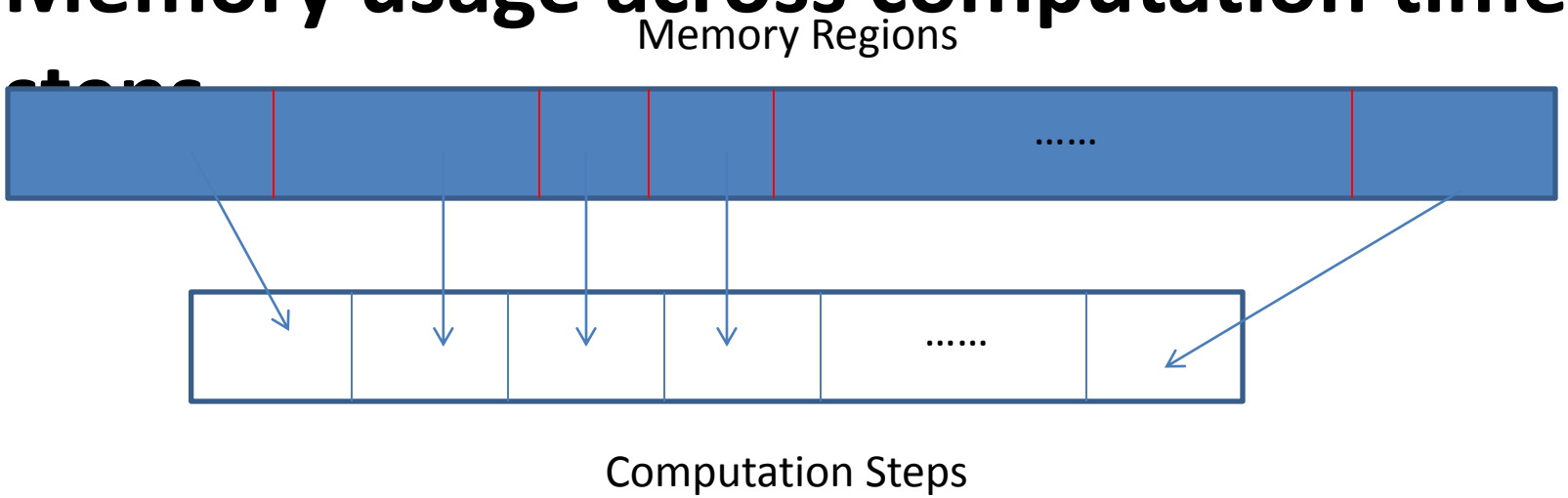
MR_{CE}: the number of access to a memory region due to the cache eviction

$T_{R/W}$: the total number of read and write references to all memory regions

$$CMCER = \frac{MR_{CM} + MR_{CE}}{T_{R/W}}$$

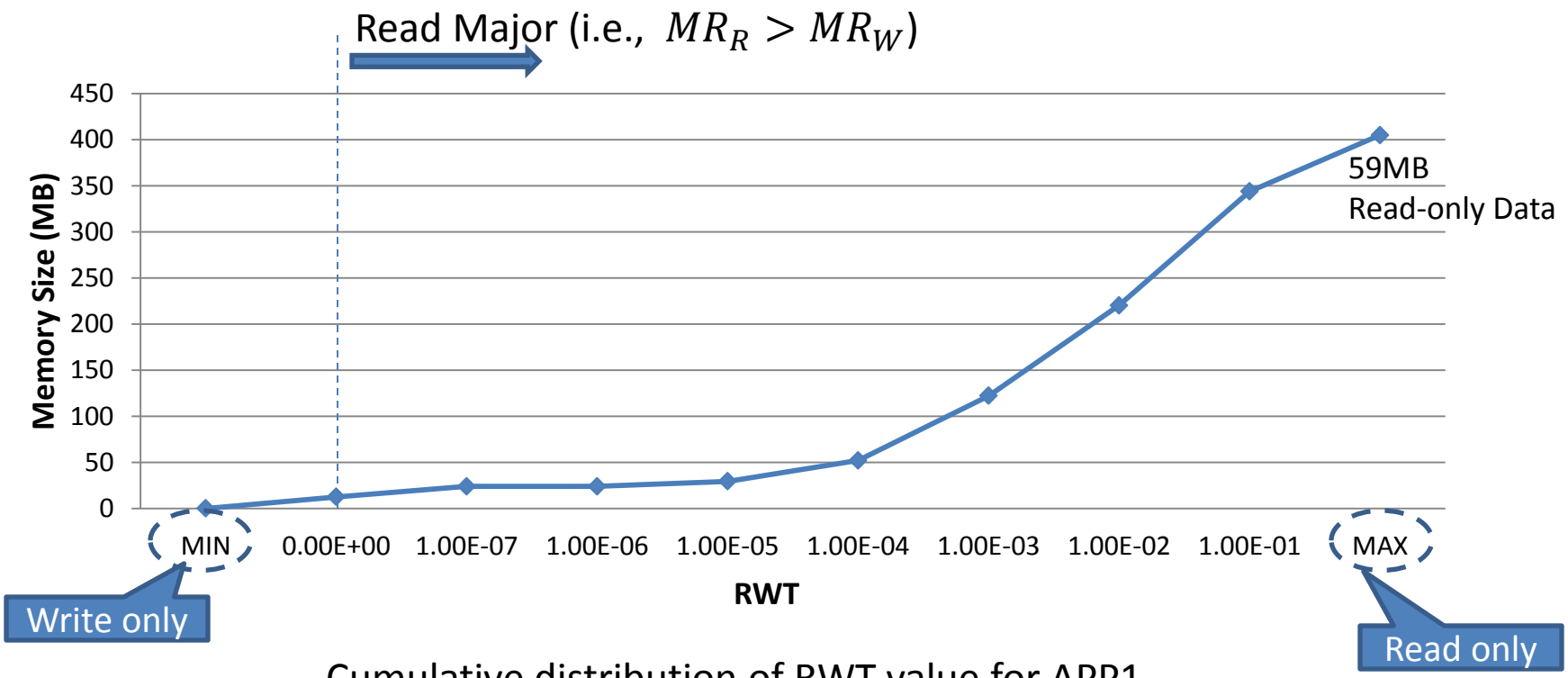
Rate

- **Statistical data for all level cache misses**
  - Bypass the cache hierarchy to save energy
  - Can we remove the cache for NVRAM? (explore new architecture designs)
- **Memory usage across computation time steps**



APP1

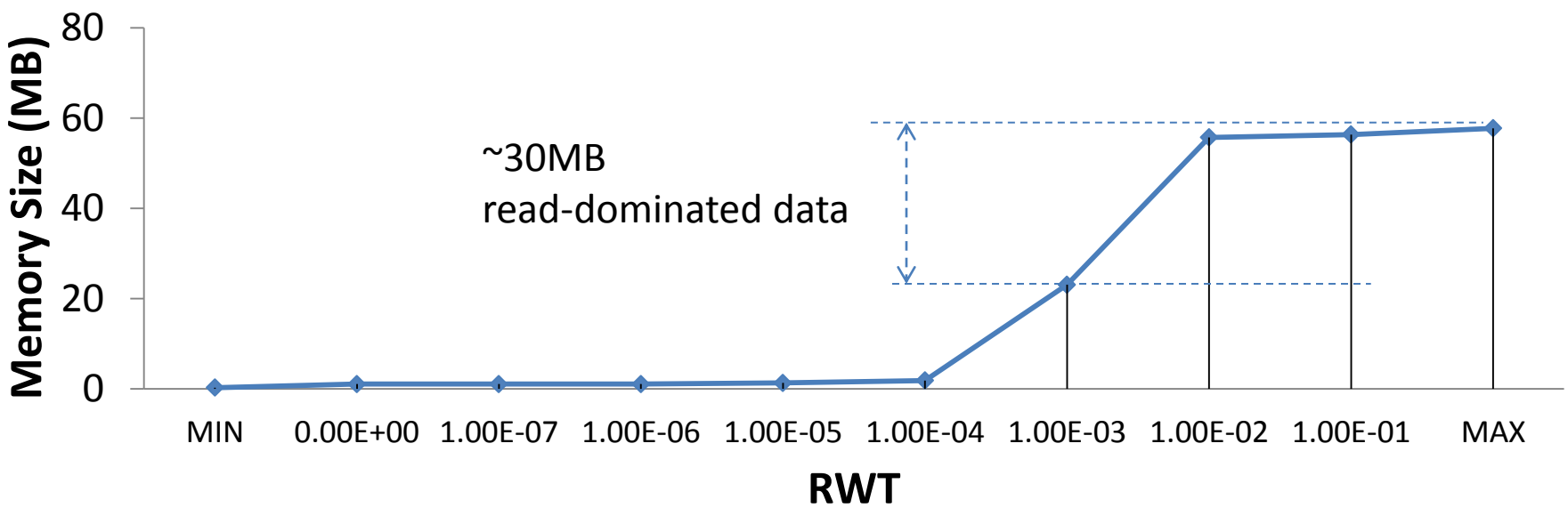
- **A computational fluid dynamic solver**
  - cover a broad range of applications
  - We instrument the eddy problem, a 2D problem
- **RWT**



Cumulative distribution of RWT value for APP1

APP2

- A turbulent reacting flow solver
- A high-order accurate, non-dissipative numerical scheme solved on a three-dimensional structured Cartesian mesh
- RWT

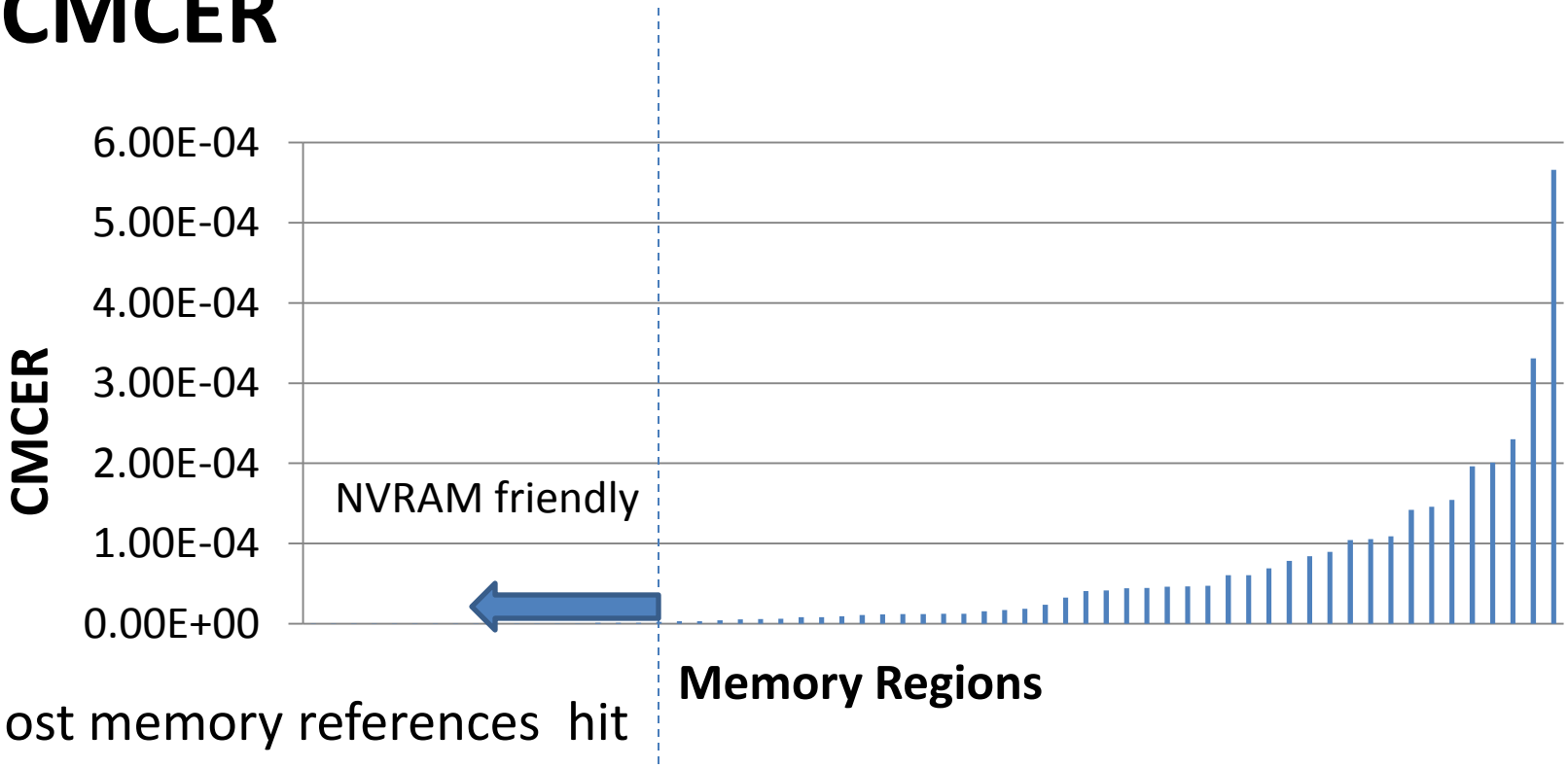


Cumulative distribution of RWT value for APP1

Memory size will be further increased if we increase the number of grid points

APP1

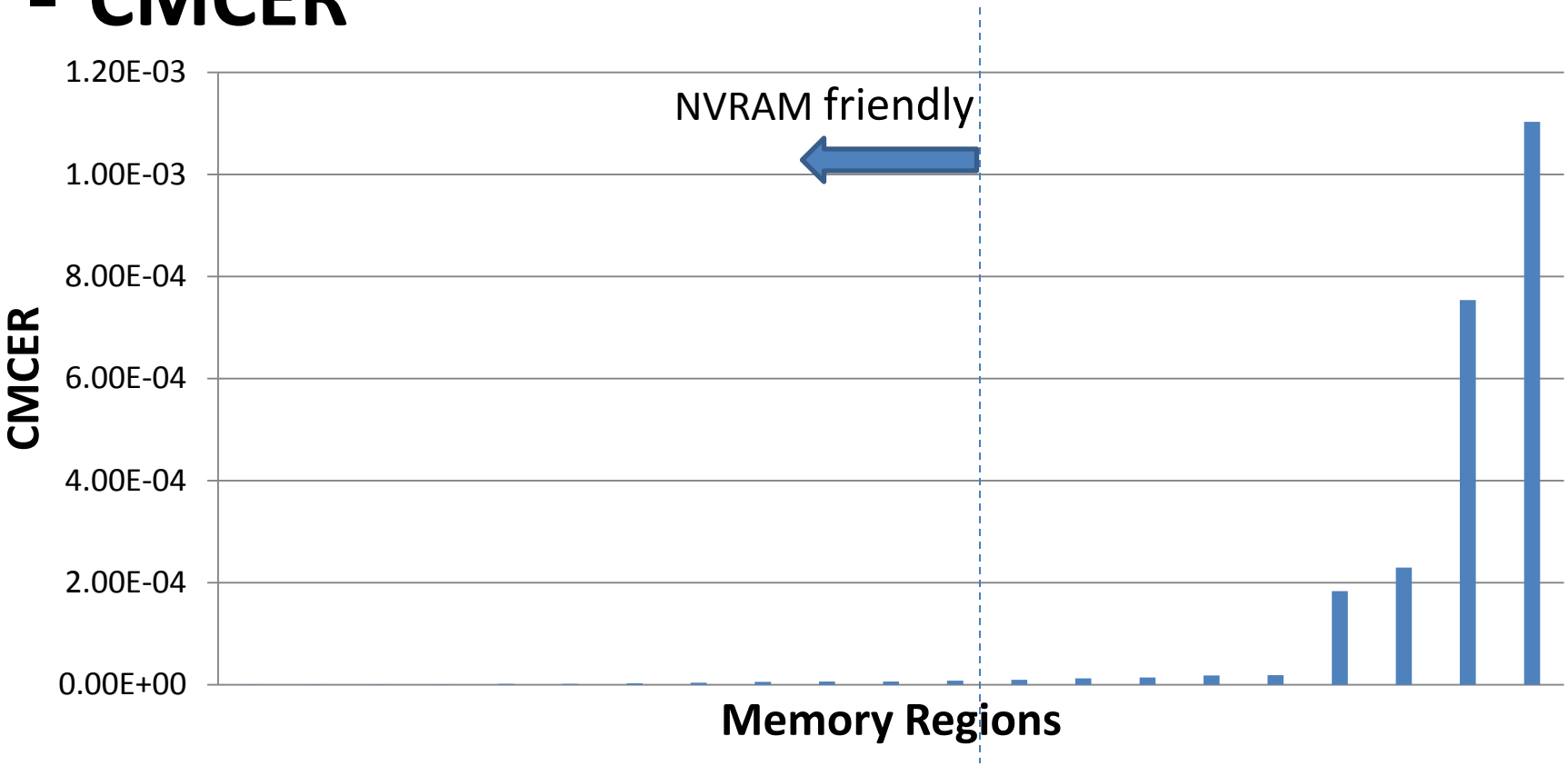
- **CMCER**



Most memory references hit caches

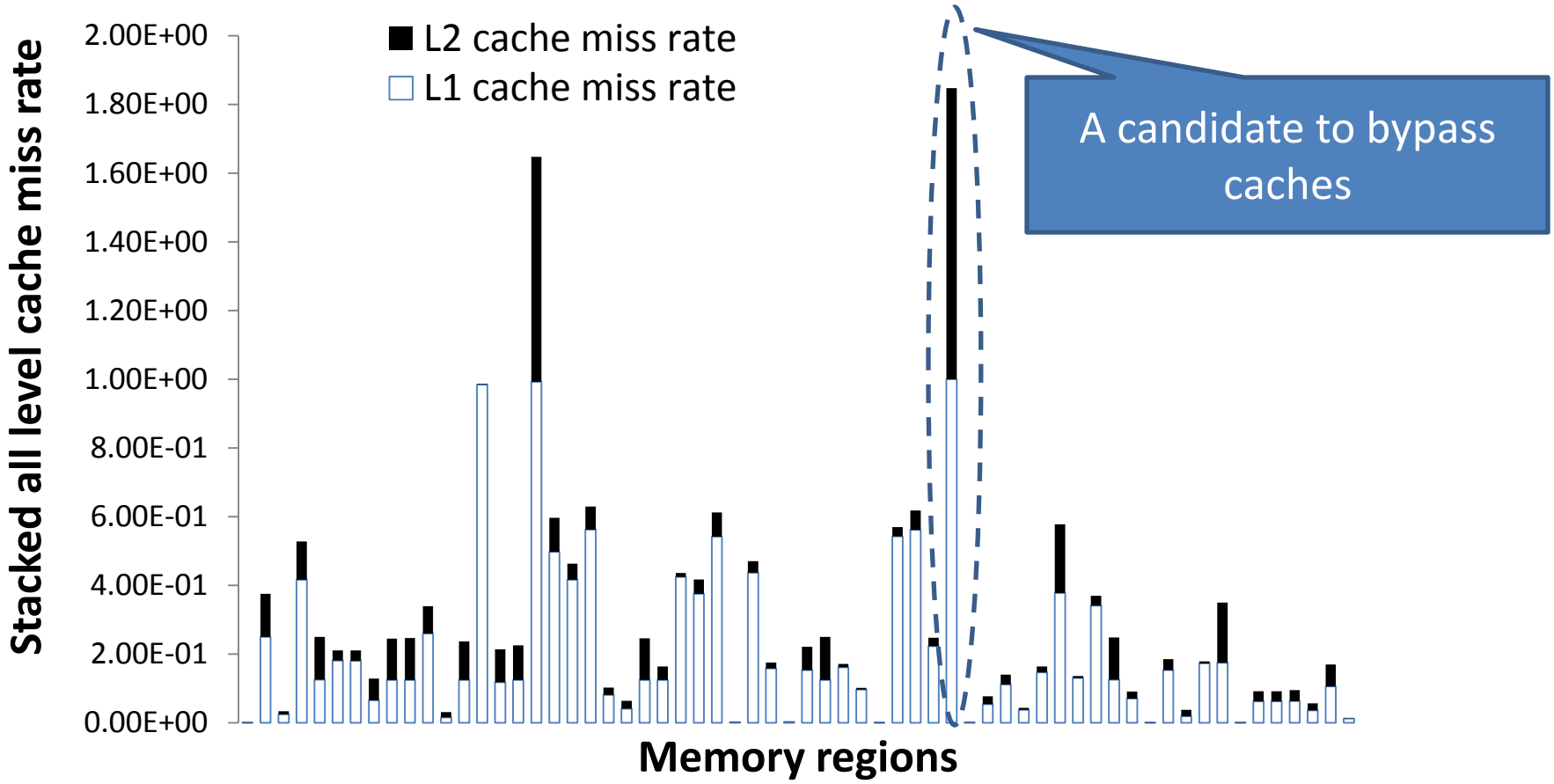
APP2

- **CMCER**


1.20E-03
1.00E-03
8.00E-04
6.00E-04
4.00E-04
2.00E-04
0.00E+00
CMCER
NVRAM friendly
Memory Regions

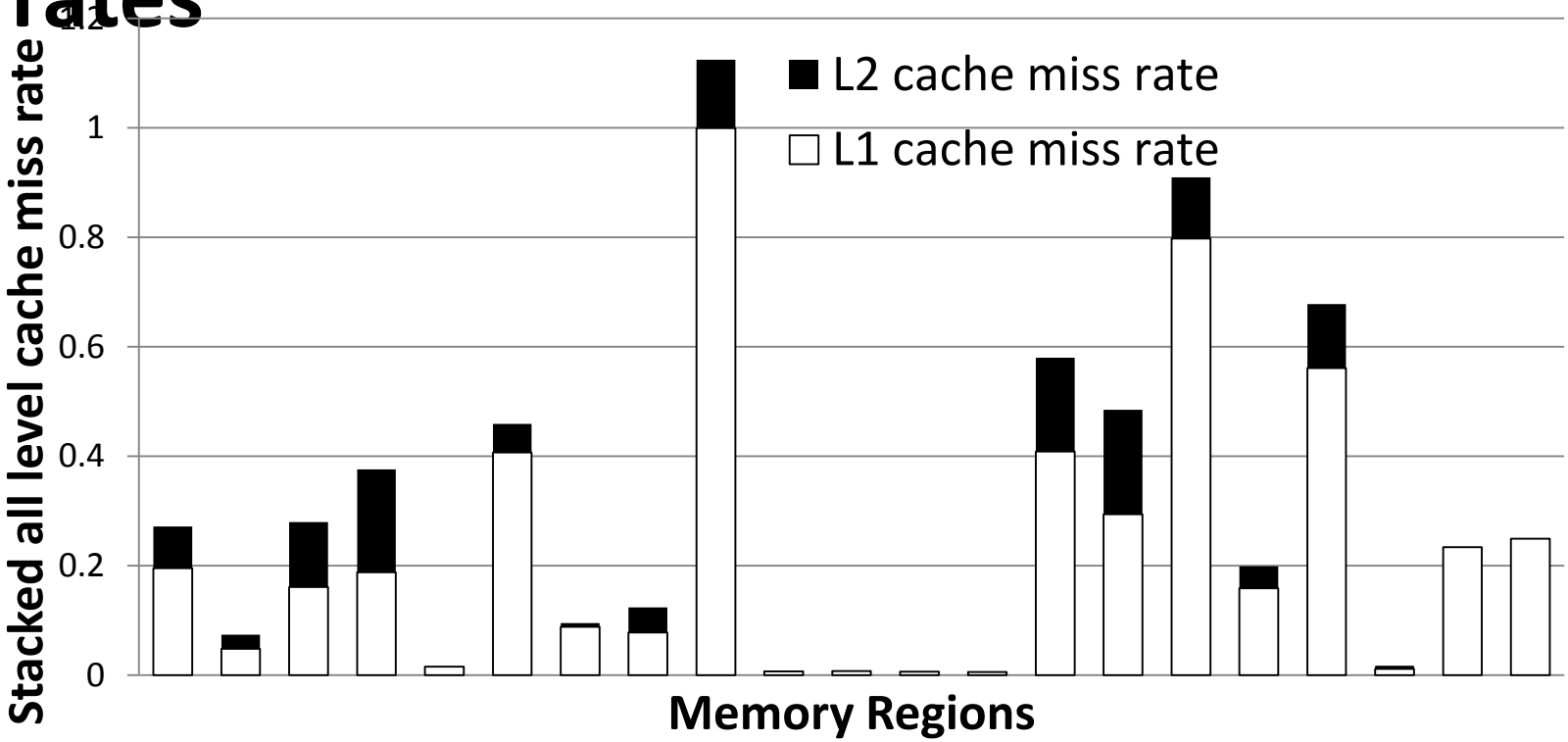
APP1

- **Statistical data for all level cache miss rates**



APP2

- **Statistical data for all level cache miss rates**



Most references are satisfied by the caches

In-situ Analysis

in si·tu

[in **sahy**-too, -tyoo, **see**-; *Lat.* in **sit**-oo] Show IPA

–***noun***

1. situated in the original, natural, or existing place or position: *The archaeologists were able to date the vase because it was found in situ.*

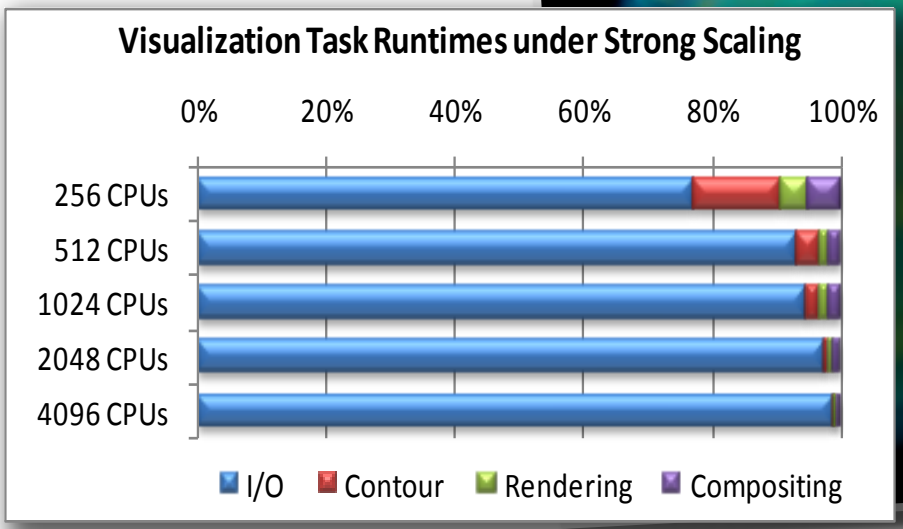
2. *Medicine/Medical* .

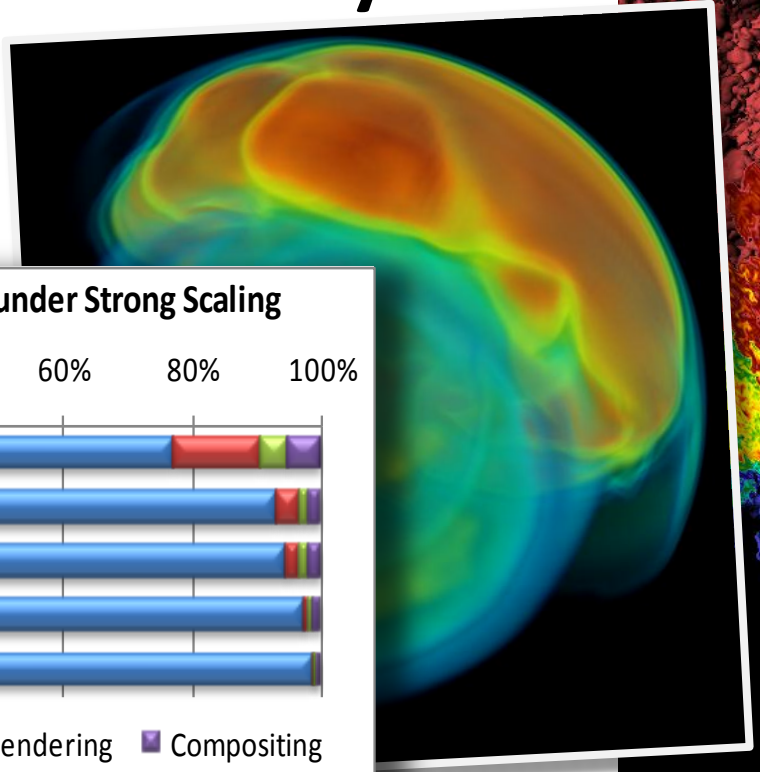
a. in place or position; undisturbed.

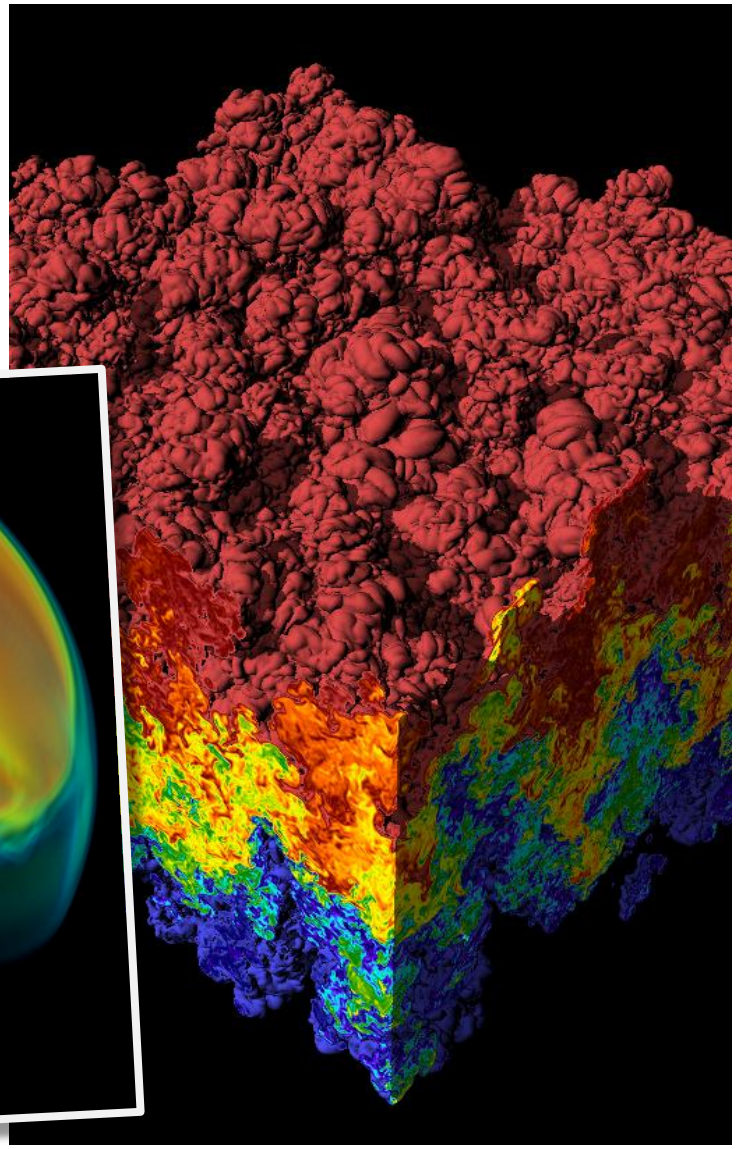
b. in a localized state or condition: *carcinoma in situ.*

Visualization and Analysis Depends on I/O

- Scaling studies with VisIt report 90% time in I/O
- Worse as concurrency increases







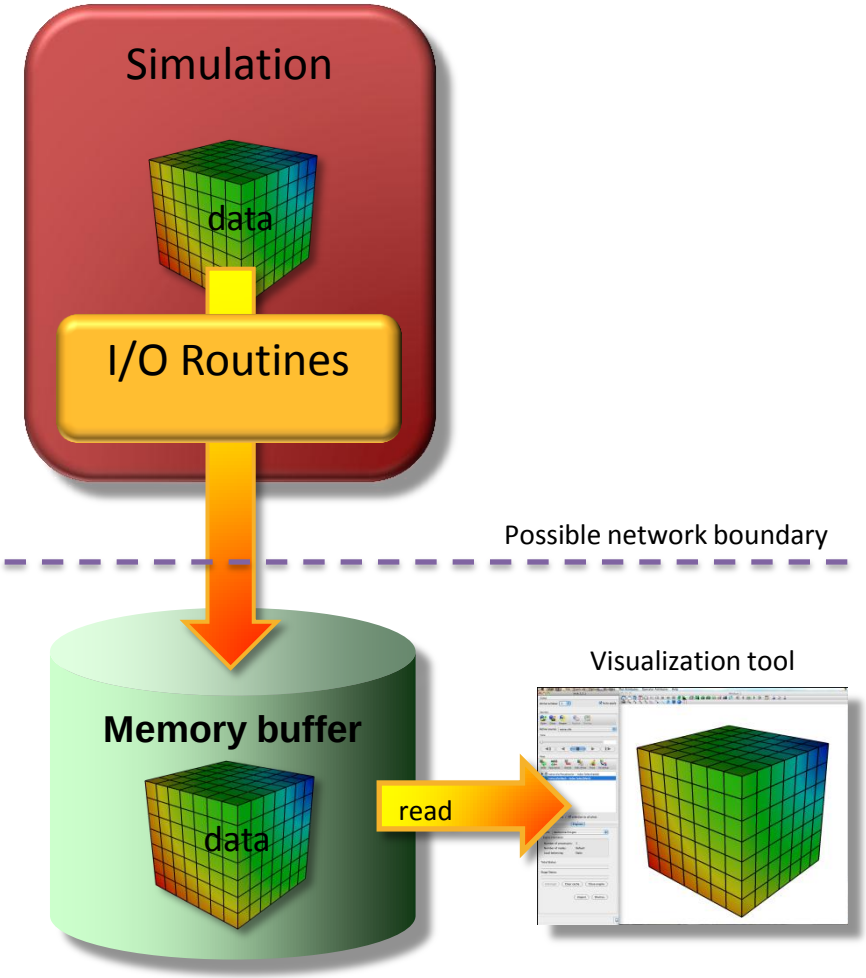
In Situ Analysis Avoids I/O

- **Many types of analysis are amenable to calculation as the data is generated, i.e. *in situ***
- **Three general types of in situ processing:**

Strategy	Description	Caveats
Tightly-coupled	Visualization/analysis have direct access to memory of simulation code	• Very memory constrained • Potential performance, stability costs
Loosely-coupled	Visualization/analysis run on concurrent resources and access data over network	• Data movement costs • Requires separate resources
Hybrid	Data is reduced in a tightly coupled setting and sent to a concurrent resource	• Most complex • Shares caveats of the other strategies

Loosely Coupled In Situ Processing

- I/O routines stage data into secondary memory buffers, possibly on other compute nodes
- Visualization applications access the buffers and obtain data
- Separates visualization processing from simulation processing
- Copies and moves data



Tightly Coupled Custom In Situ Processing

- **Custom visualization routines are developed specifically for the simulation and are called as subroutines**
  - Optimized for data layout
  - Create best visual representation
- **Tendency to concentrate on very specific visualization scenarios**
- ***Write once, use once***