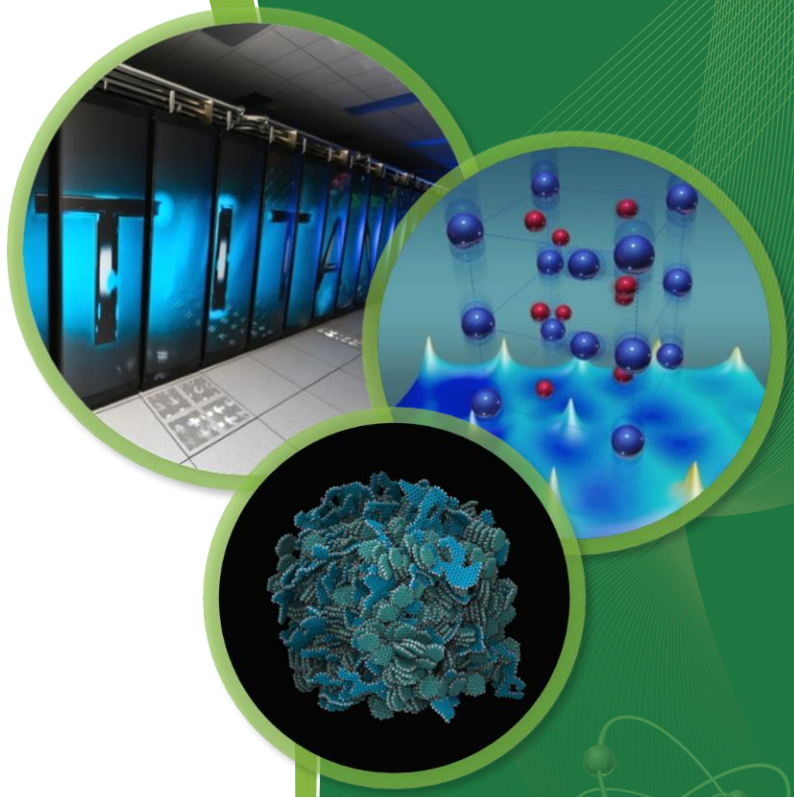
Exploring Emerging Memory Technologies in Extreme Scale High Performance Computing

Jeffrey S. Vetter

Many contributions from FTG Group and Friends



Presented to
PPAM 2017: 12th International Conference on Parallel Processing and Applied Mathematics

Lublin, Poland

17 Sep 2017

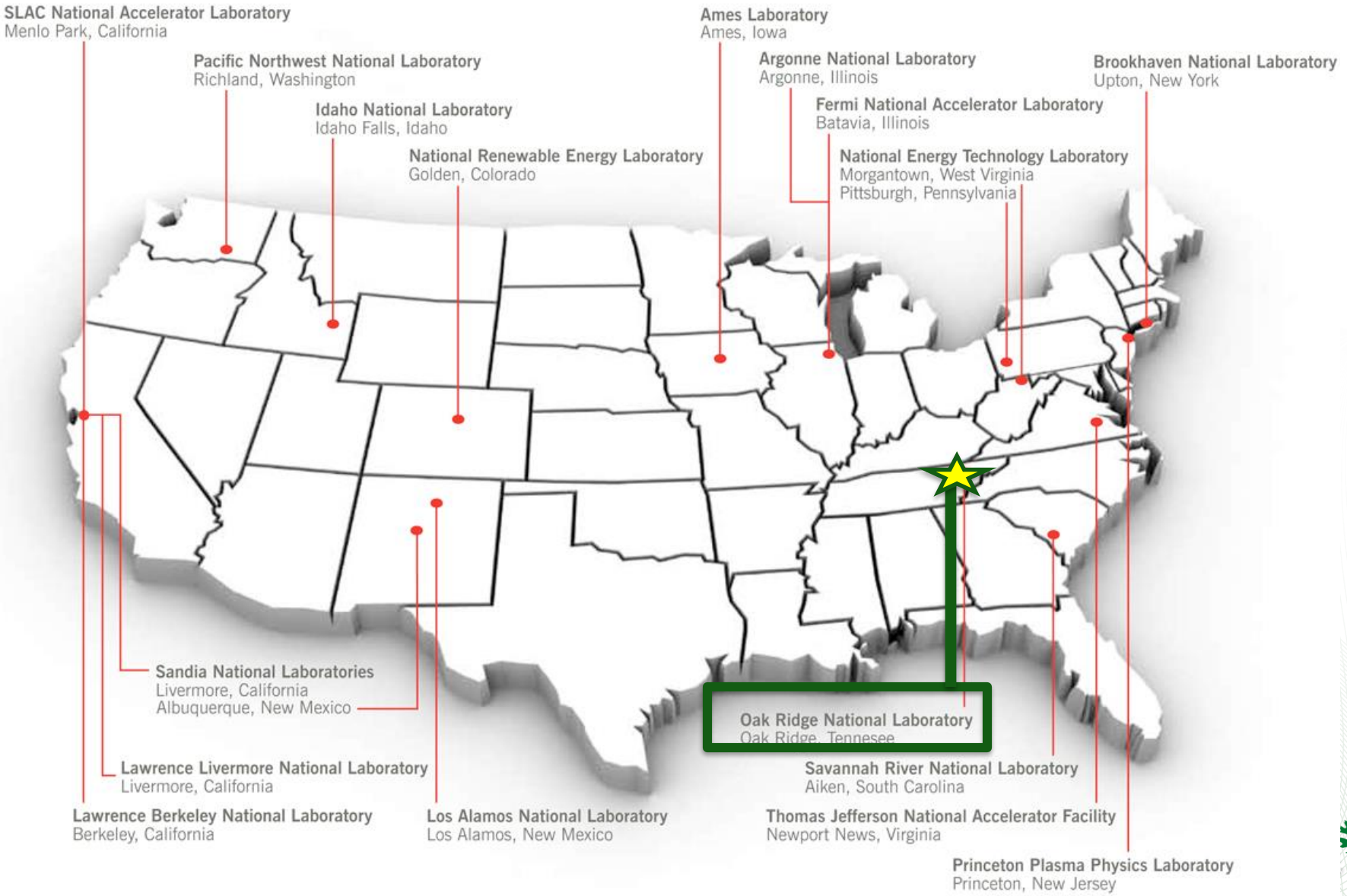
http://ft.ornl.gov vetter@computer.org

OAK RIDGE National Laboratory

Highlights

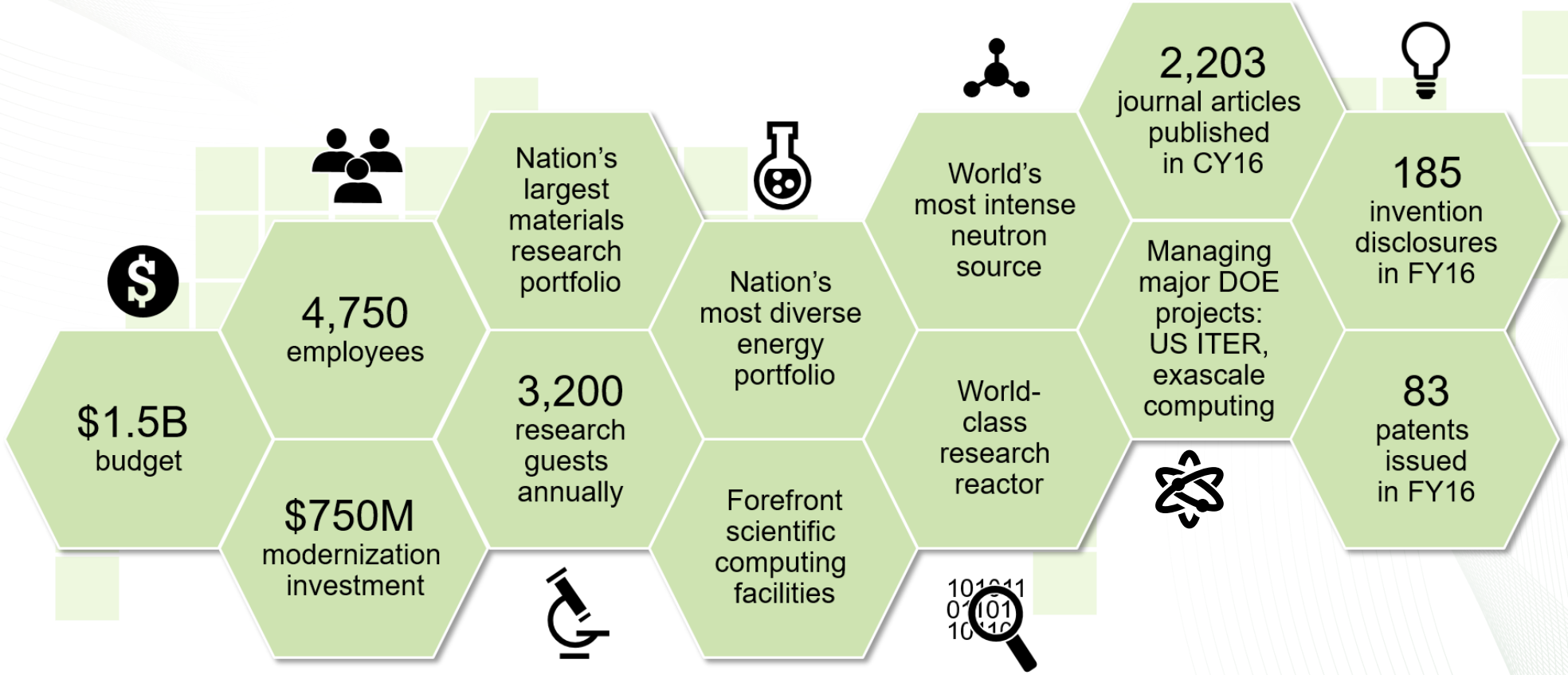
- Recent trends in extreme-scale HPC paint an ambiguous future
  - Contemporary systems provide evidence that power constraints are driving architectures to change rapidly (e.g., Dennard, Moore)
  - Markets and business strategies impact our goals
  - Multiple architectural dimensions are being (dramatically) redesigned: Processors, node design, memory systems, I/O
- Memory systems are leading the charge!
  - New devices and materials
  - New system organizations
  - New configurations
  - Vast (local) capacities
- Programming systems must support these new memory systems (and portability)!!
  - We need new programming systems to effectively use these architectures
  - Dragon: transparent access from GPUs to vast amounts of NVM
  - NVL-C: programming a hybrid DRAM-NVM main memory
  - Papyrus: aggregating NVM to provide distributed data structures
- These changes in underlying memory system technologies will have substantial impact on both architecture and application design

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



13

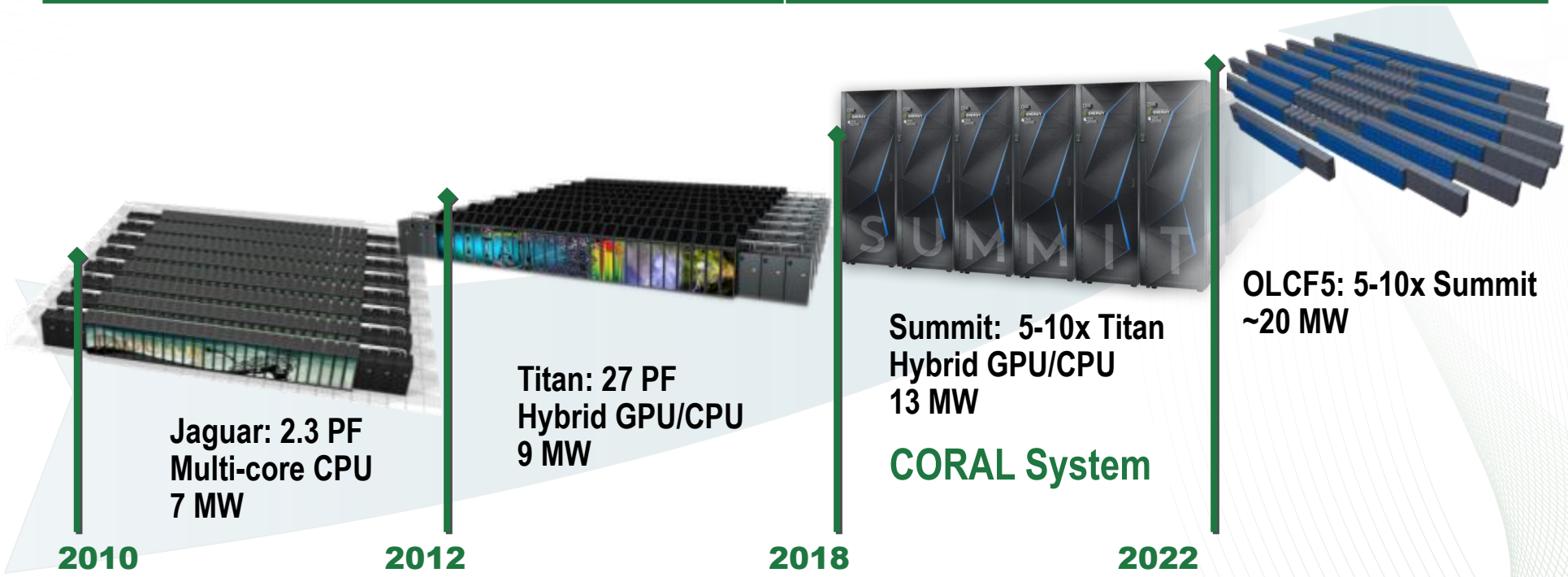
Today, ORNL is a leading science and energy laboratory



OAK RIDGE National Laboratory

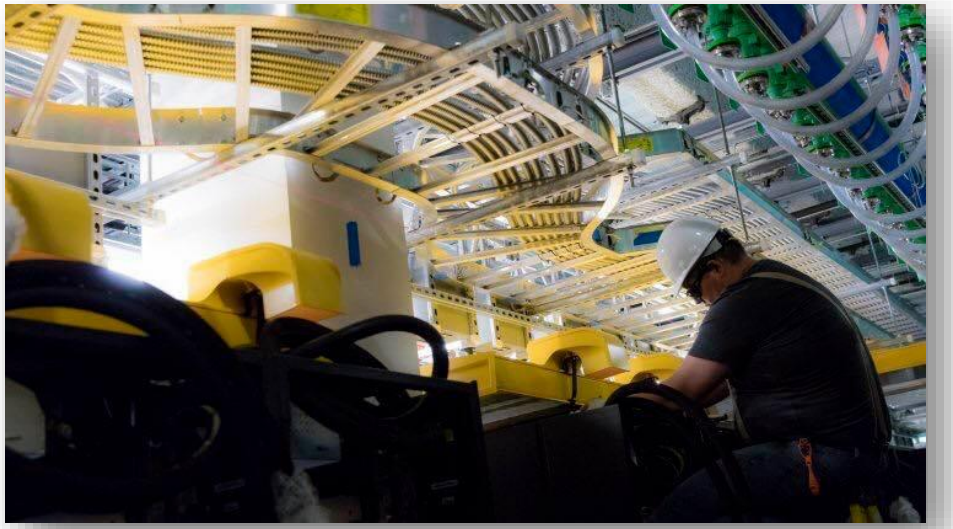
Our Science requires that we continue to advance our computational capability over the next decade on the roadmap to exascale.

Since clock-rate scaling ended in 2003, HPC performance has been achieved through increased parallelism. Jaguar scaled to 300,000 cores.	Titan and beyond deliver hierarchical parallelism with very powerful nodes. MPI plus thread level parallelism through OpenACC or OpenMP plus vectors



2018 OLCF leadership system Hybrid CPU/GPU architecture

summit

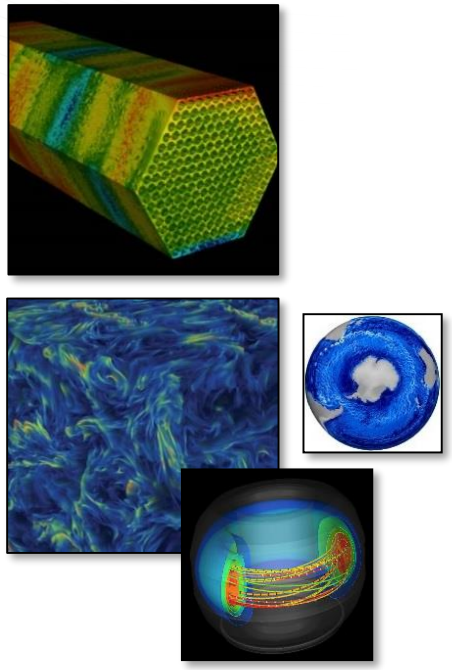


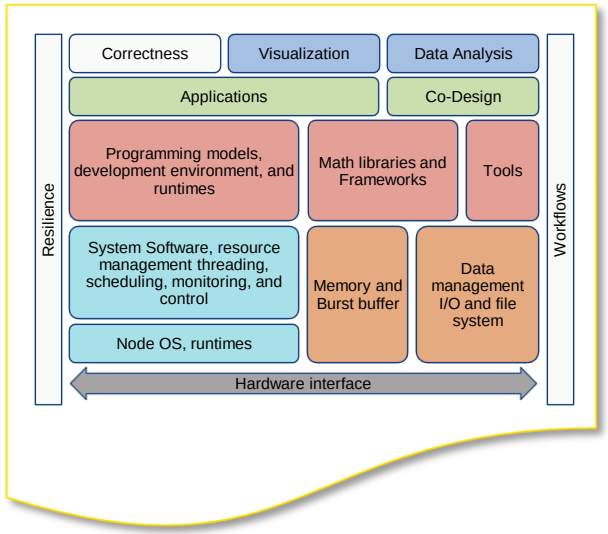
Vendor: IBM (Prime) / NVIDIA™ / Mellanox Technologies®

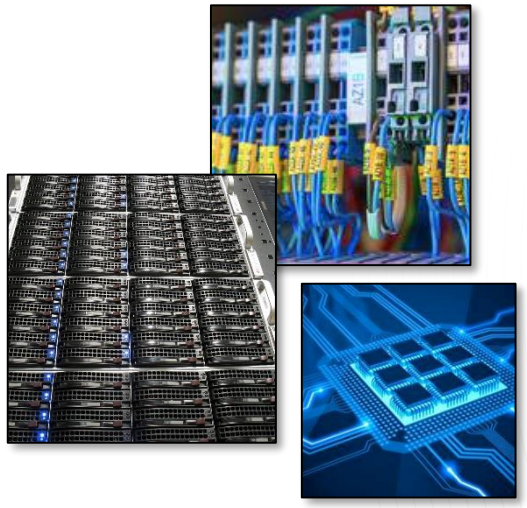
FEATURE	TITAN	SUMMIT
Application Performance	Baseline	5-10x Titan
Number of Nodes	18,688	~4,600
Node performance	1.4 TF	> 40 TF
Memory per Node	32 GB DDR3 + 6 GB GDDR5	512 GB DDR4 + HBM
NV memory per Node	0	1600 GB
Total System Memory	710 TB	>10 PB DDR4 + HBM + Non-volatile
System Interconnect (node injection bandwidth)	Gemini (6.4 GB/s)	Dual Rail EDR-IB (23 GB/s)
Interconnect Topology	3d Torus	Non-blocking Fat Tree
Processors	1 AMD Opteron™ 1 NVIDIA Kepler™	2 IBM POWER9™ 6 NVIDIA Volta™
File System	32 PB, 1 TB/s, Lustre©	250 PB, 2.5 TB/s, GPFS™
Peak power consumption	9 MW	15 MW

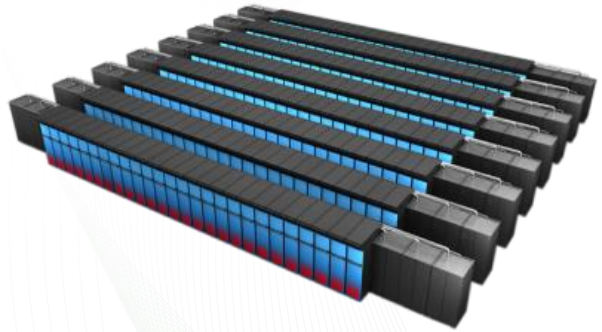
ECP has formulated a holistic approach that uses co-design and integration to achieve capable exascale

Application Development	Software Technology	Hardware Technology	Exascale Systems
Science and mission applications	Scalable software stack	Hardware technology elements	Integrated exascale supercomputers









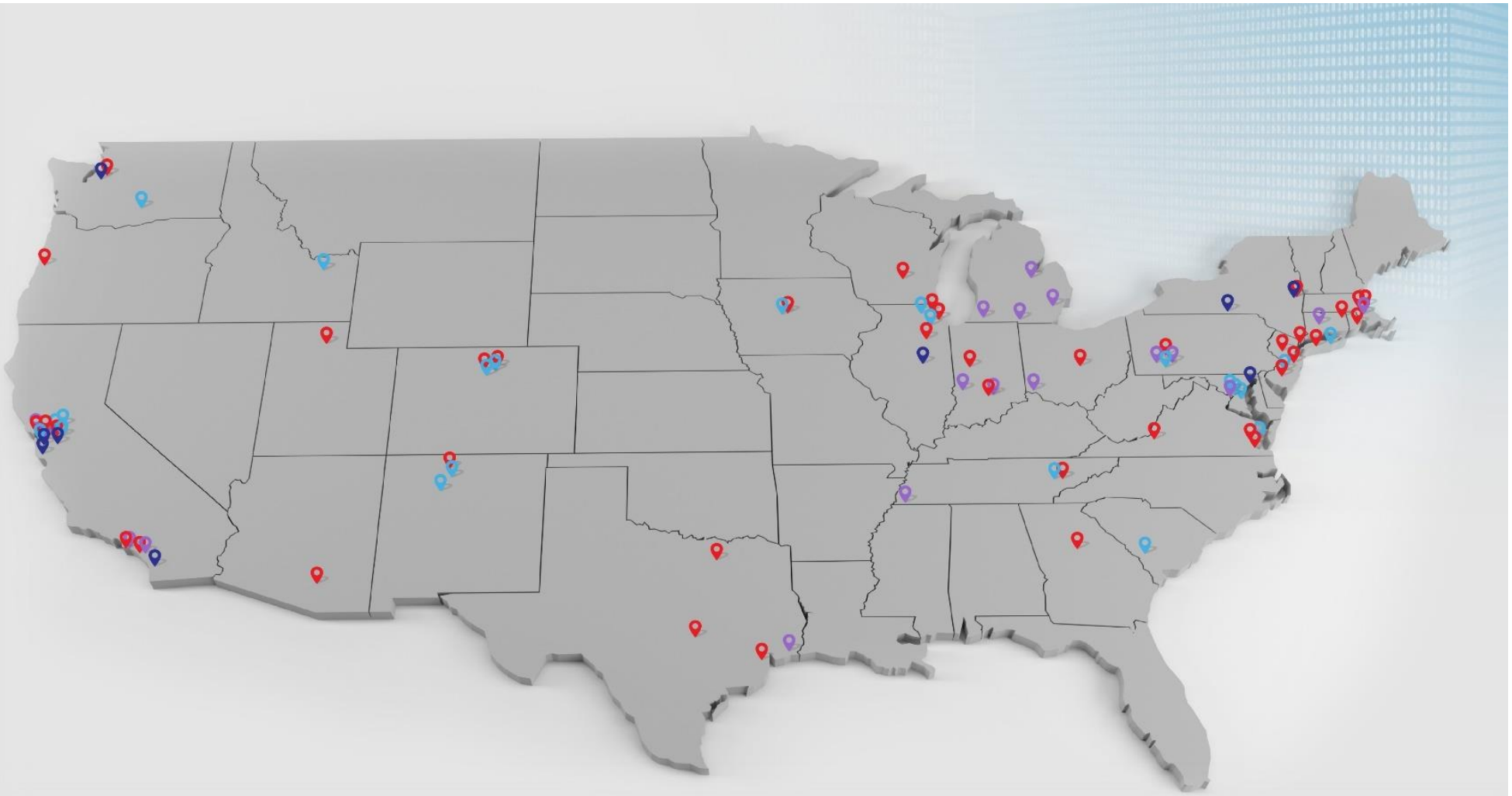
EXASCALE COMPUTING PROJECT

DOE LABORATORIES & AGENCY PARTNERS
22

PRIVATE SECTOR PARTNERS
9

UNIVERSITY RESEARCH PARTNERS
39

INDUSTRY COUNCIL MEMBERS
18



THE ECP ECOSYSTEM

- 800 Researchers
- 26 Application Development Projects
- 66 Software Development Projects
- 5 Co-Design Centers

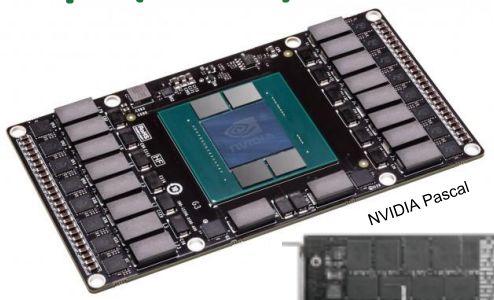
U.S. DEPARTMENT OF ENERGY | Office of Science

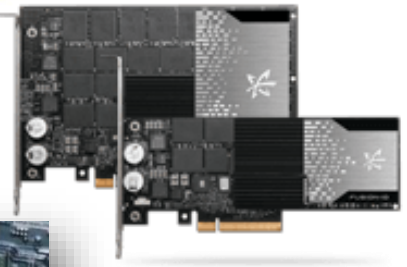
National Nuclear Security Administration

Future Technologies Group (FTG)

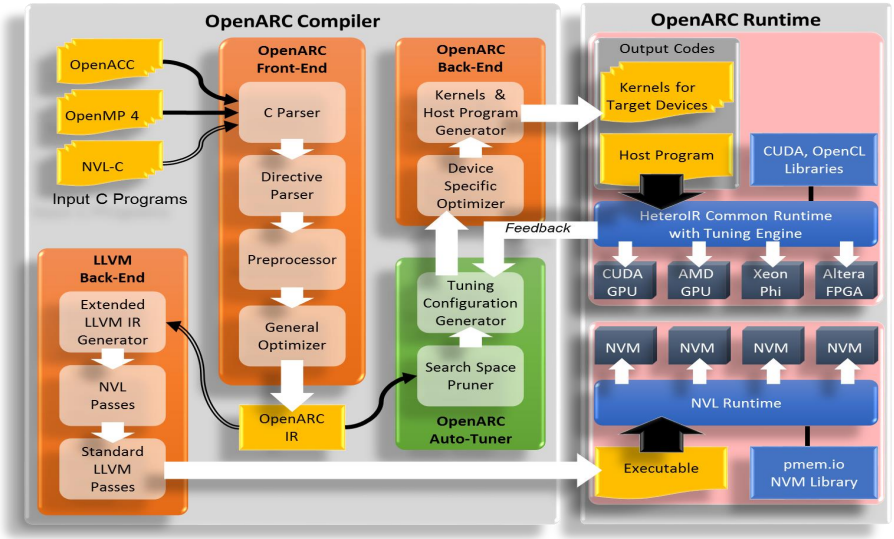
Jeffrey S. Vetter, Group Leader

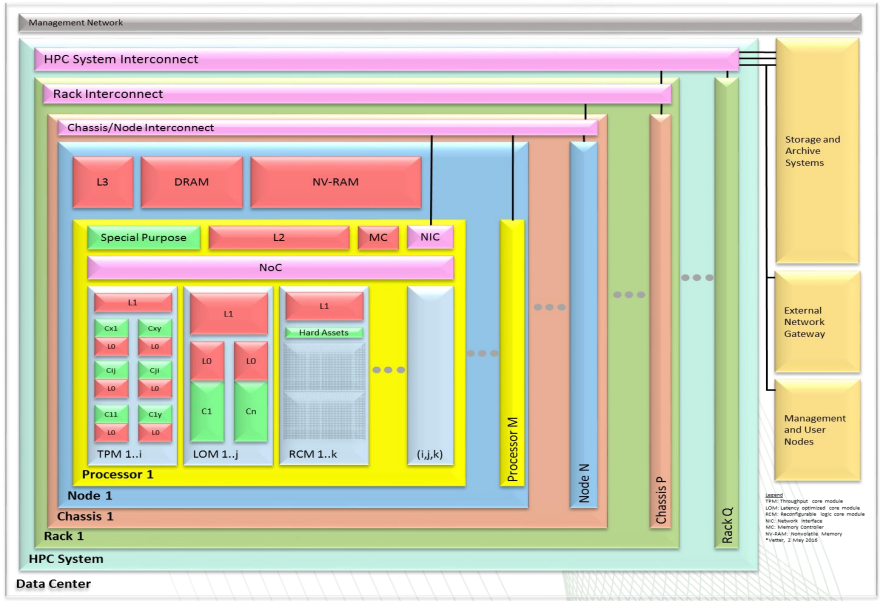
The Future Technologies Group performs research in core technologies for future generations of high-end computing architectures, including prototype computer architectures and experimental software systems. We investigate these technologies with the goal of improving the performance, energy efficiency, reliability, and productivity of these architectures for our sponsors and applications teams. See http://ft.ornl.gov.





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





Key Technical Areas

- Heterogeneous architectures
- Deep memory hierarchies including non-volatile memory
- Performance measurement, analysis, simulation, and modeling of emerging architectures.
- Programming systems to address emerging architectures
- Beyond Moore's Computing

Software Artifacts

- Scalable Heterogeneous Computing Benchmarks (SHOC)
- mpiP
- DESTINY
- Aspen
- OpenARC
- Papyrus
- NVL-C
- Oxbow
- LLVM Parallel IR
- NV-Scavenger

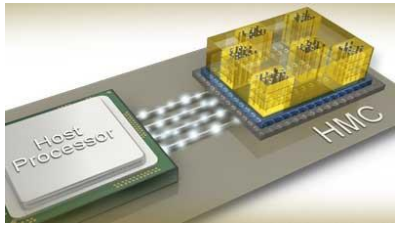
Impact

- Publications in SC, ICS, HPDC, TPDS, DATE, PLDI, IPDPS, Trans VLSI, etc.
- Two Gordon Bell awards
- NSF Keeneland
- DOE Titan
- IEEE TCHPC Early Career
- IEEE Fellow
- ~60 interns
- ~120 FTG seminars

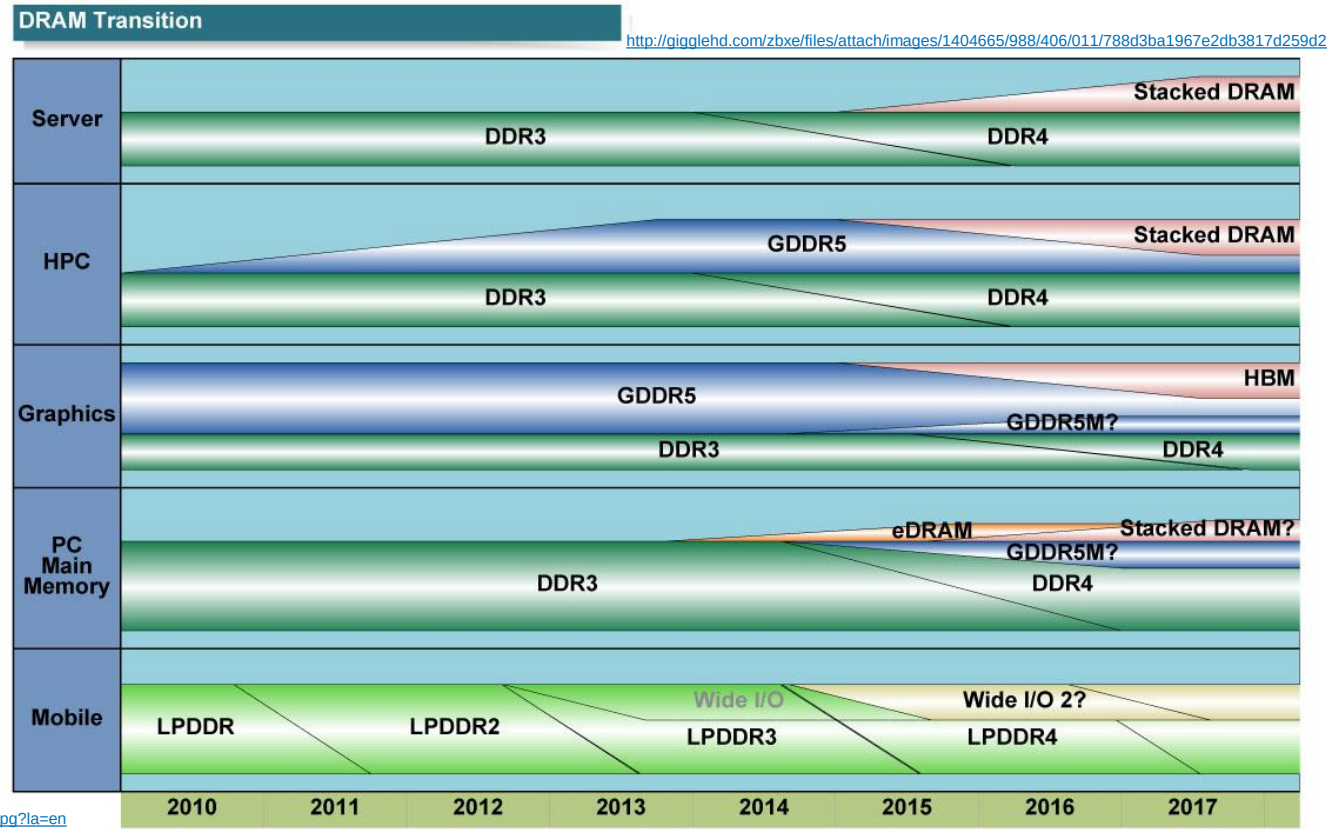
Emerging Memory Systems

Memory Systems Started Diversifying Several Years Ago

- **Architectures**
  - HMC, HBM/2/3, LPDDR4, GDDR5X, WIDEIO2, etc
  - 2.5D, 3D Stacking
- **Configurations**
  - Unified memory
  - Scratchpads
  - Write through, write back, etc
  - Consistency and coherence protocols
  - Virtual v. Physical, paging strategies
- **New devices**
  - ReRAM, PCRAM, STT-MRAM, Xpoint



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



http://gigglehd.com/zbxe/files/attach/images/1404665/988/406/011/788d3ba1967e2db3817d259d2e

	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.*

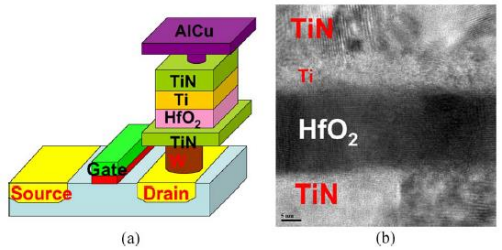


Fig. 4. *(a) A typical 1T1R structure of RRAM with HfO_x; (b) HR-TEM image of the TiN/Ti/HfO_x/TiN stacked layer; the thickness of the HfO_2 is 20 nm.*

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

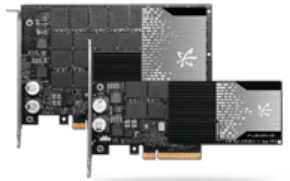
Current ASCR Computing At a Glance

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

NVRAM Technology Continues to Improve – Driven by Broad Market Forces



designlines MEMORY

Blog

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

Kevin Gibb, Product Line Mar
TechInsights
4/6/2016 04:40 PM EDT
9 comments post a comment

Like 16 Tweet Shar

The highly anticipated Samsu memory is out in the market, first look.

Samsung had announced its 25 K9AFGY8S0M 3D V-NAND as would be used in a variety of sol on the market in early 2016. True to their word, we managed to find them in their 2 TB capacity, mSATA, T3 portable SSD shown in Figure 1.

tom's HARDWARE PRODUCT REVIEWS NEWS DEALS FORUM

Samsung's 10-Year Plan Starts With 128TB QLC SSD, 960 Successor

22 COMMENTS

by Chris Ramseyer August 8, 2017 at 12:30 PM

designlines WIRELESS & NETWORKING

Slideshow

Facebook Likes Intel's 3D XPoint

Google joins open hardware effort

Rick Merritt

NO RATINGS
LOGIN TO RATE

3

use Int
while G
tandard
rs for di
annour
ompute
a new
prototy
packag

Facebook's support for the 3D XPoint non-volatile me
co-developed by Intel and Micron is "a huge endorse
Nathan Brookwood, principal of market watcher 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

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

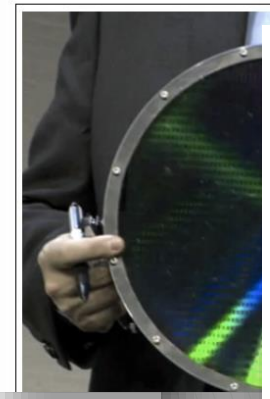
By Chris Mellor

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

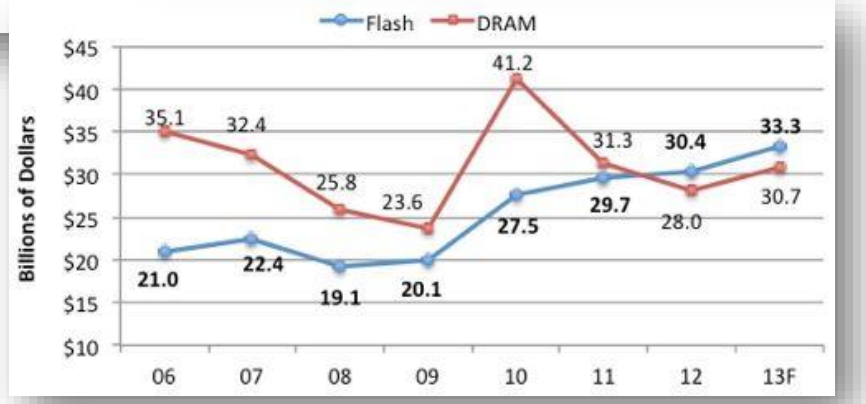
Like 56 Tweet Share 212 G+1 4

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.



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

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

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

Comparison of Emerging Memory Technologies

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

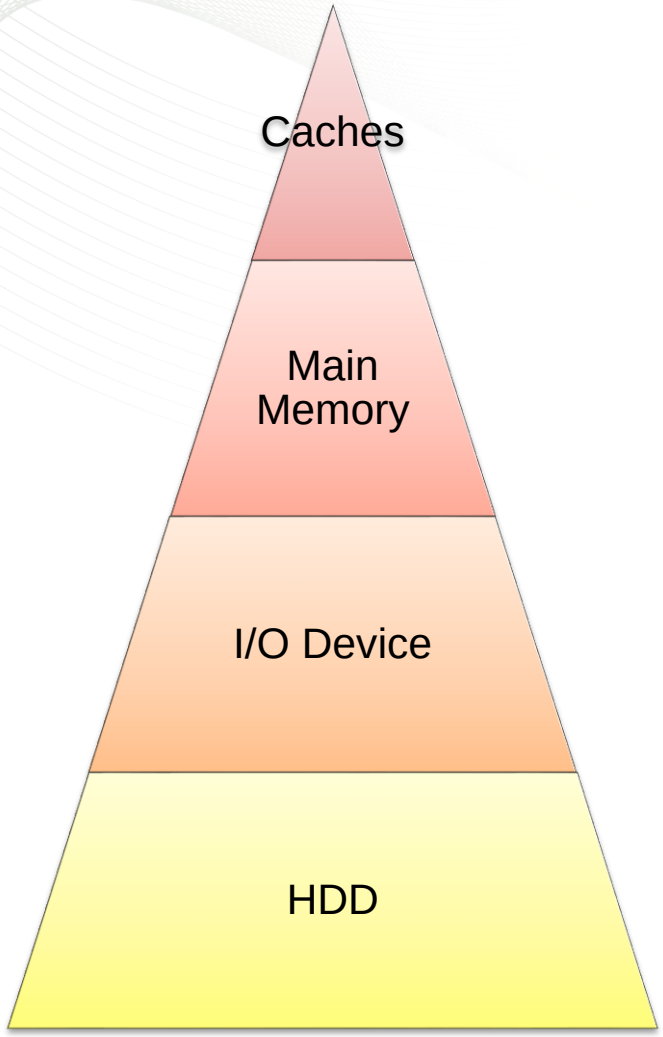
Intel/Micron Xpoint?
Samsung Z-NAND?

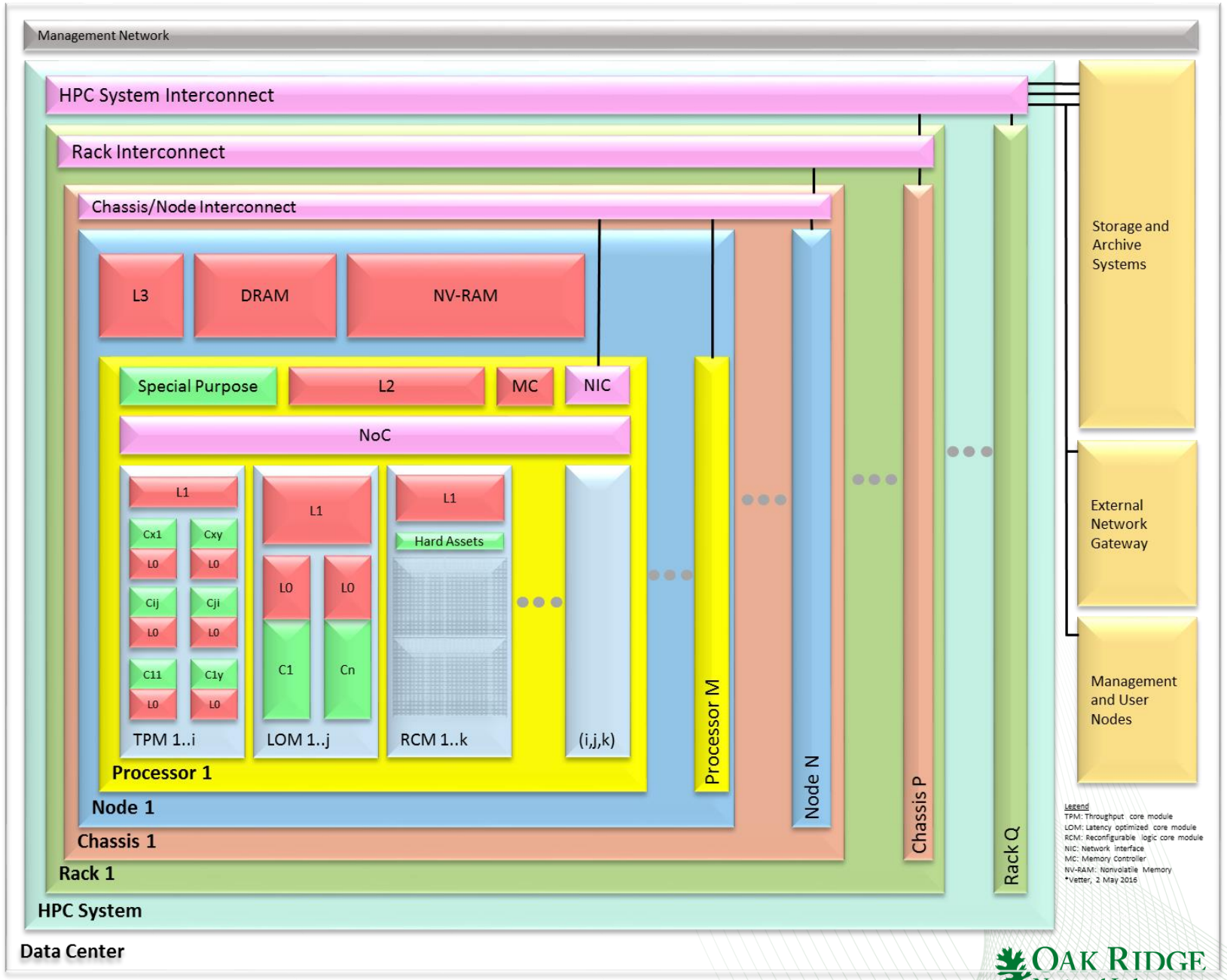
http://ft.ornl.gov/trac/blackcomb

Investigating Solutions to Memory and Storage Challenges

Applications	• Model and quantify application workloads with architecture independent metrics; analyze and optimize application performance • Aspen, NV-Scavenger, COMPASS
Programming Systems and Software	• Develop portable software to support many different architectures and application scenarios • DRAGON, NVL-C, Papyrus
System Architecture	• Architect system components for relevant workloads and scenarios • NoC alternatives, Cache strategies, Integration point, interfaces, KV support
Memory Architectures	• Organize new devices for performance, reliability, power efficiency • Stacking, compute-in-memory, error correction
Memory Devices and Materials	• Improve power efficiency to 2 pj/bit • JJ memory cell, DESTINY simulator

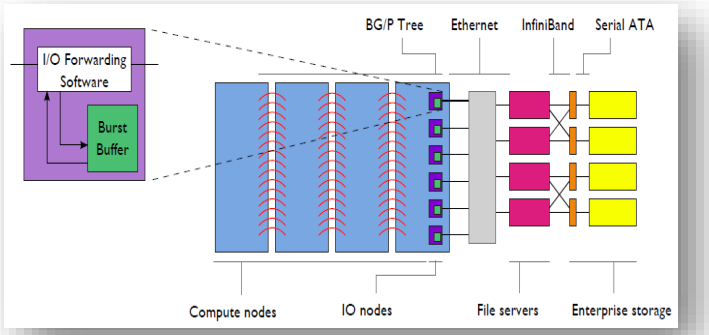
Migration up the hierarchy



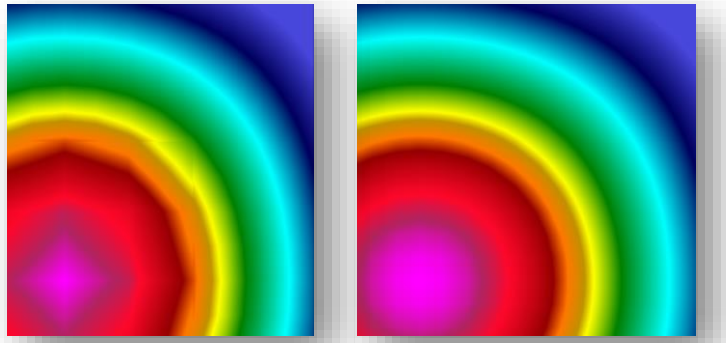


HPC Application Scenarios for NVM

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



- In situ visualization



http://ft.ornl.gov/eavl

- In-mem tables

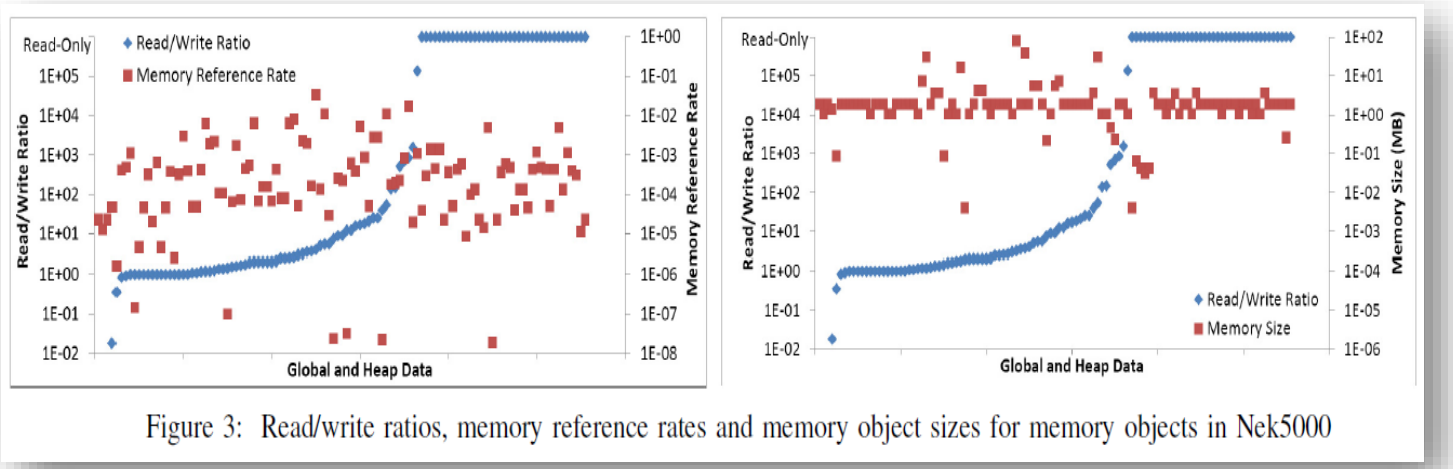


Figure 3: Read/write ratios, memory reference rates and memory object sizes for memory objects in Nek5000

Empirical results show many reasons…

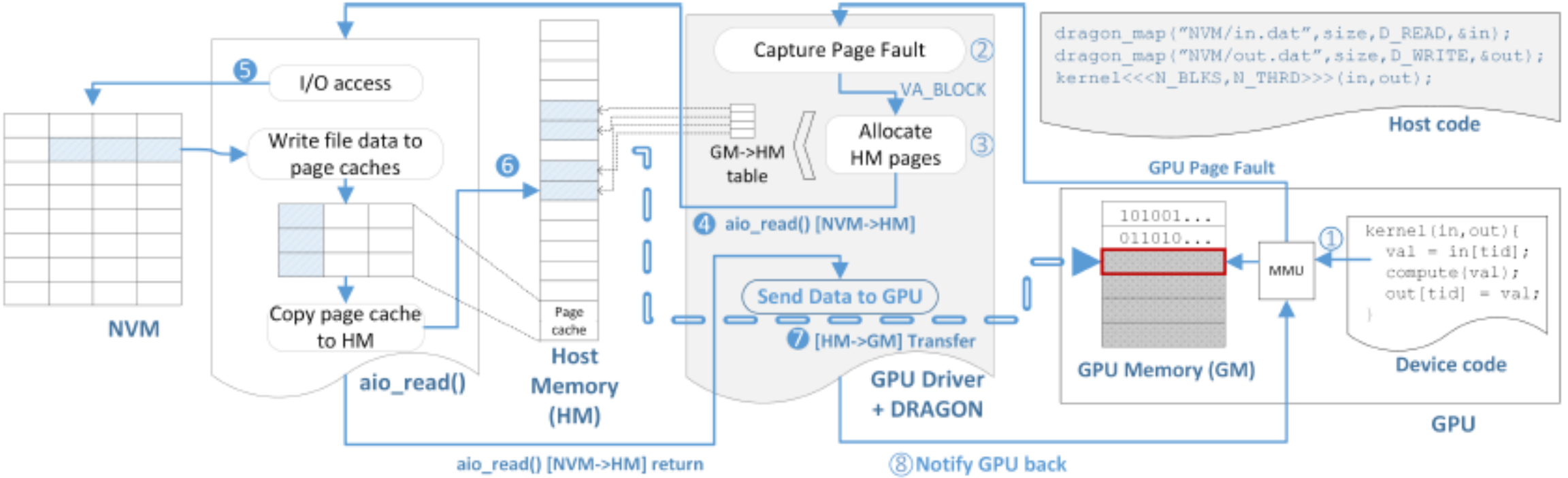
- Lookup, index, and permutation tables
- Inverted and 'element-lagged' mass matrices
- Geometry arrays for grids
- Thermal conductivity for soils
- Strain and conductivity rates
- Boundary condition data
- Constants for transforms, interpolation
- MC Tally tables, cross-section materials tables…

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

Runtime support for NVM

DRAGON provides NVM transparently to GPU through OS, drivers

- Provide vast NVM (FusionIO 1-5TB) to GPU (Pascal) transparently



Markthub, Belviranli, et al. *DRAGON: Direct Resource Access for GPUs over NVM*, submitted

Results with Caffe

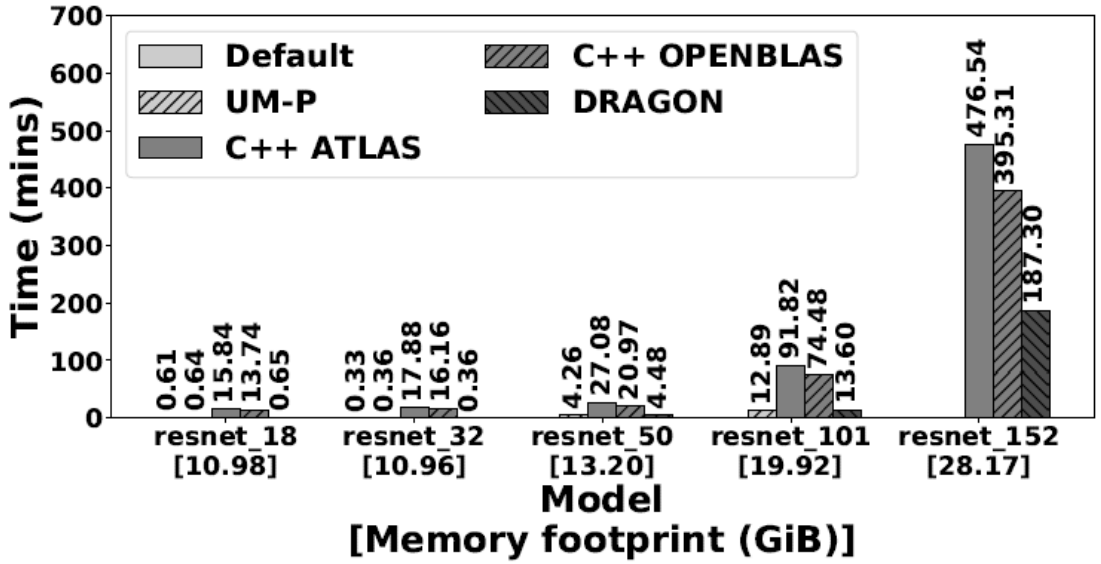


Figure 6: Comparison of ResNet execution times on Caffe.

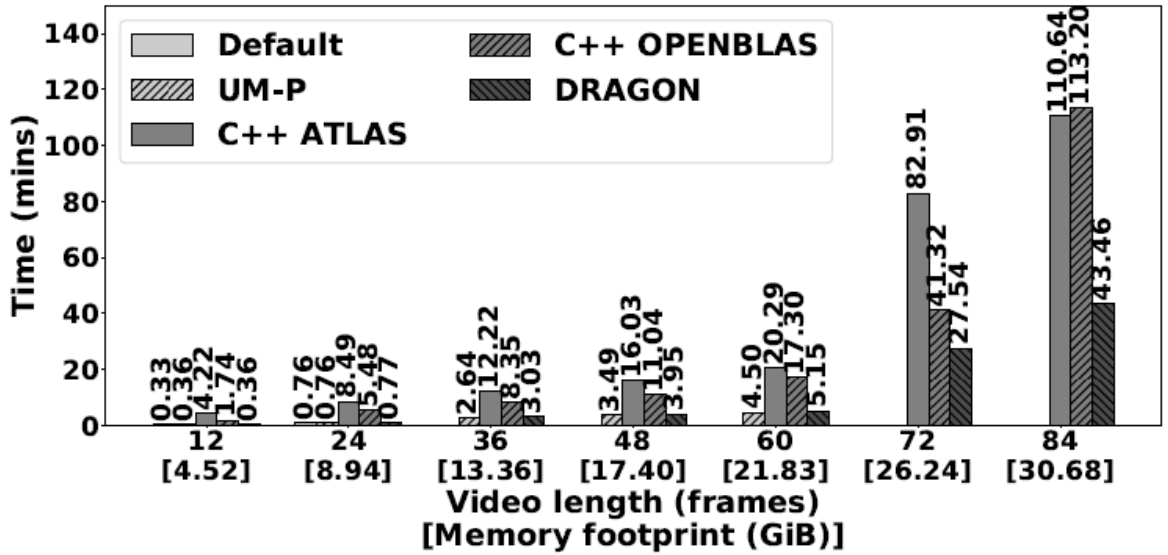


Figure 7: Comparison of C3D the execution times on Caffe.

Language support for NVM: NVL-C - extending C to support NVM

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

OAK RIDGE National Laboratory

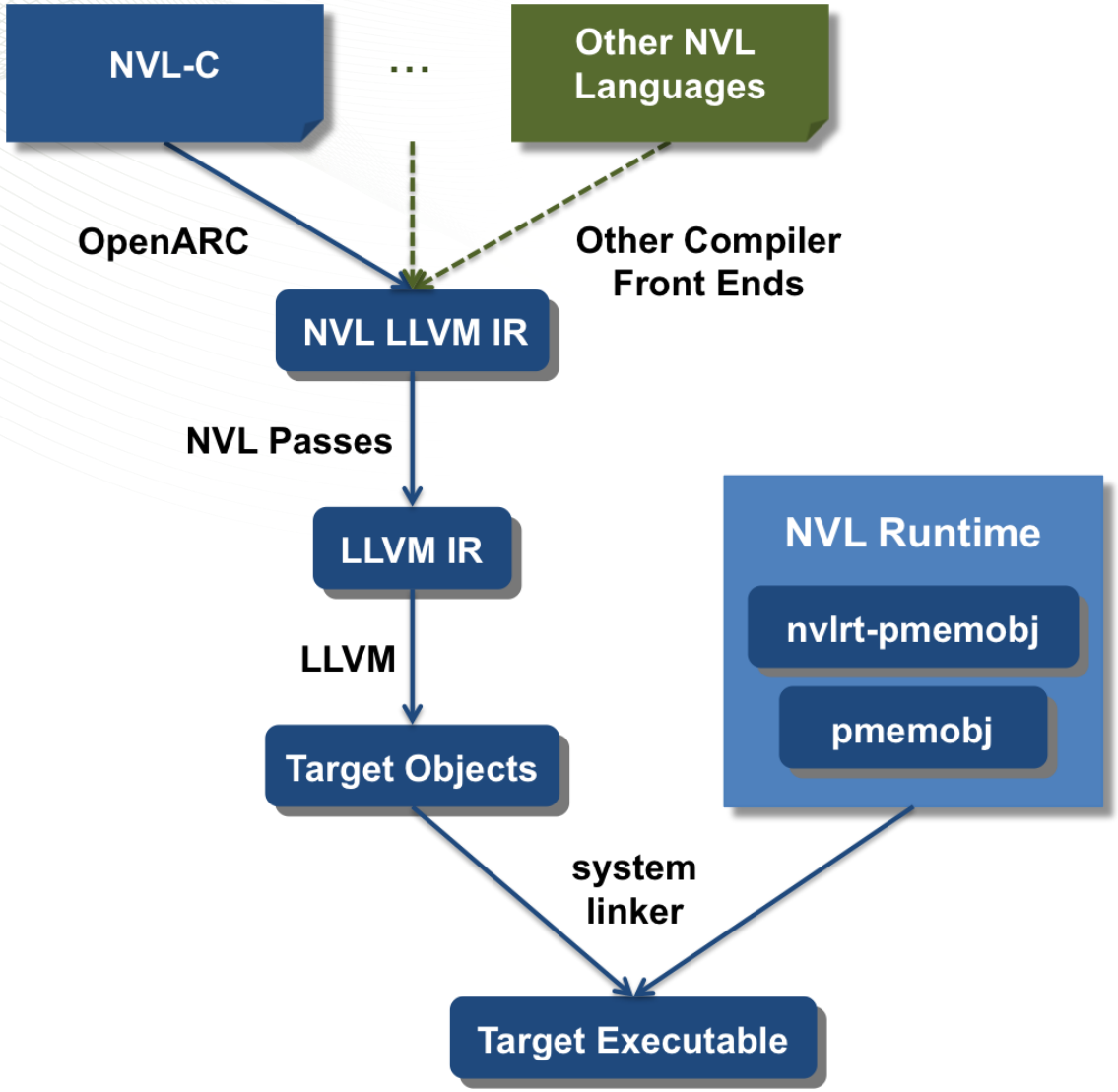
NVL-C: Programming Model

- 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);
}
```

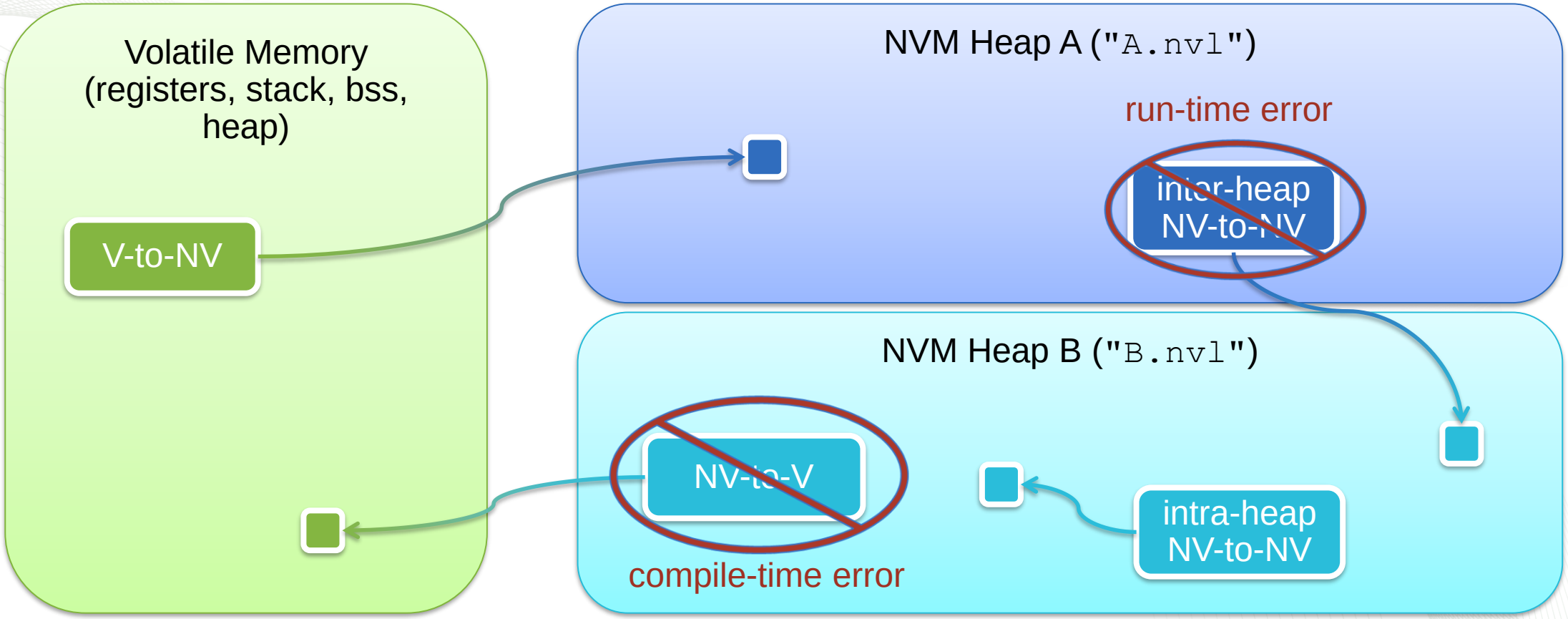

Denny, J.E., Lee, S., and Vetter, J.S.: 'NVL-C: Static Analysis Techniques for Efficient, Correct Programming of Non-Volatile Main Memory Systems'. Proc. Proceedings of the 25th ACM International Symposium on High-Performance Parallel and Distributed Computing, Kyoto, Japan2016 pp. Pages

Design Goals: Modular implementation



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

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



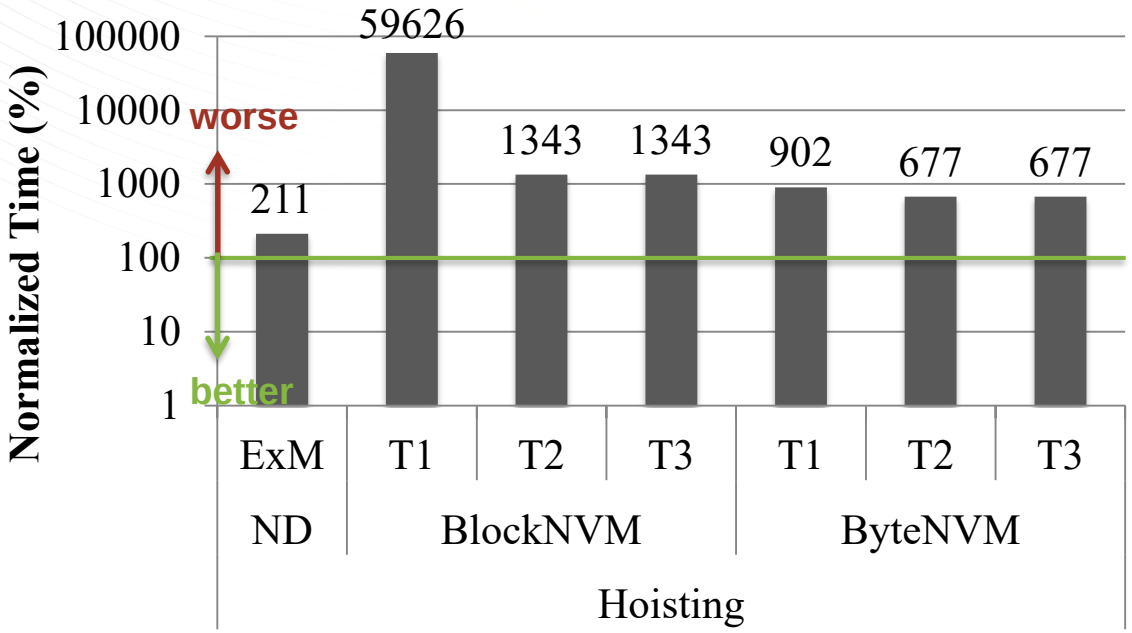
avoids dangling pointers when memory segments close

Programming Model: Transactions: Undo logs

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

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

Evaluation: LULESH



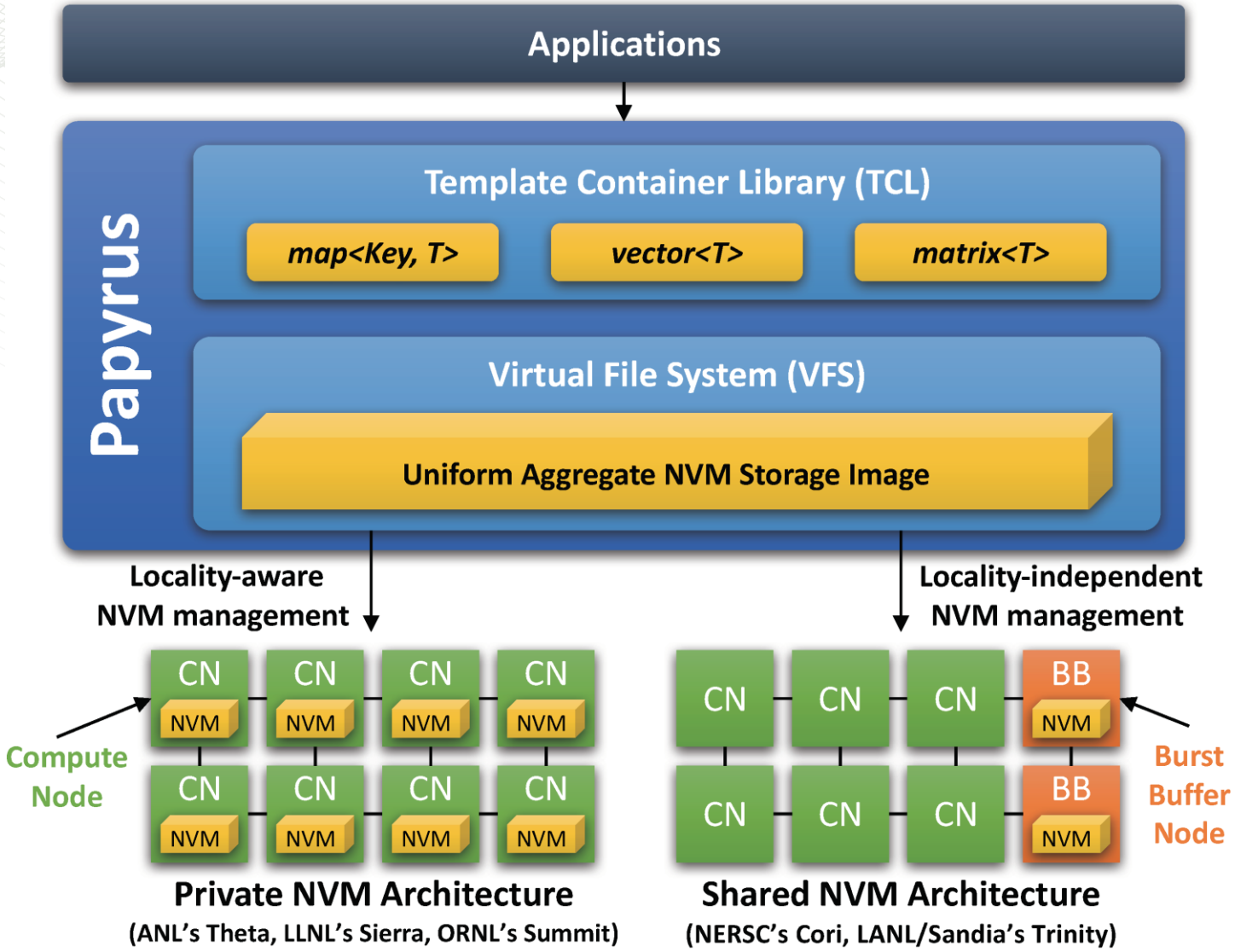
- ExM = use SSD as extended DRAM
- T1 = BSR + transactions
- T2 = T1 + `backup` clauses
- T3 = T1 + `clobber` clauses
- BlockNVM = `msync` included
- ByteNVM = `msync` suppressed

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

Programming Scalable NVM with Papyrus

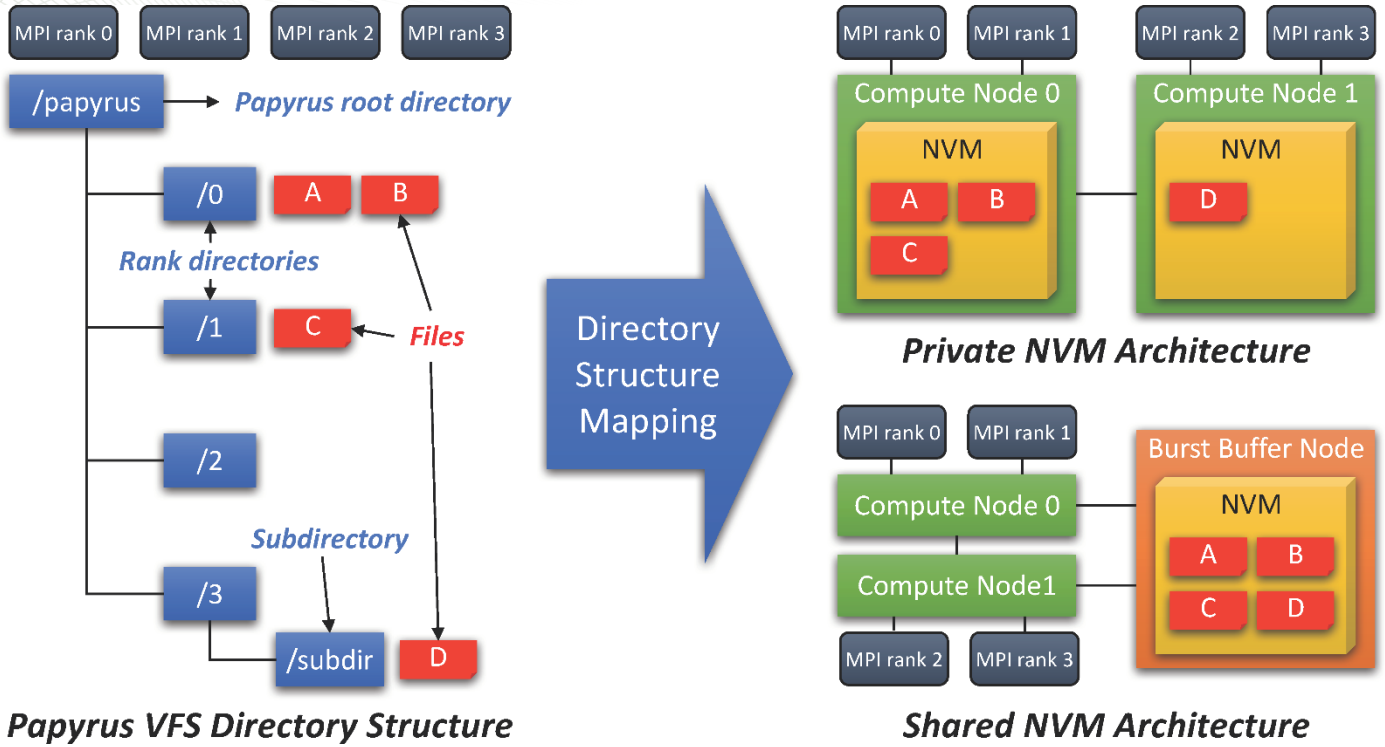
OAK RIDGE National Laboratory

Papyrus Overview



- Papyrus
  - A user-level library using MPI
    - MPI-interoperable
    - No daemon, no server
- Virtual File System (VFS)
  - Uniform aggregate NVM storage image
- Template Container Library (TCL)
  - High-level programming interface built on top of VFS
    - Data elements are distributed to multiple NVM nodes

Papyrus VFS Directory Structure



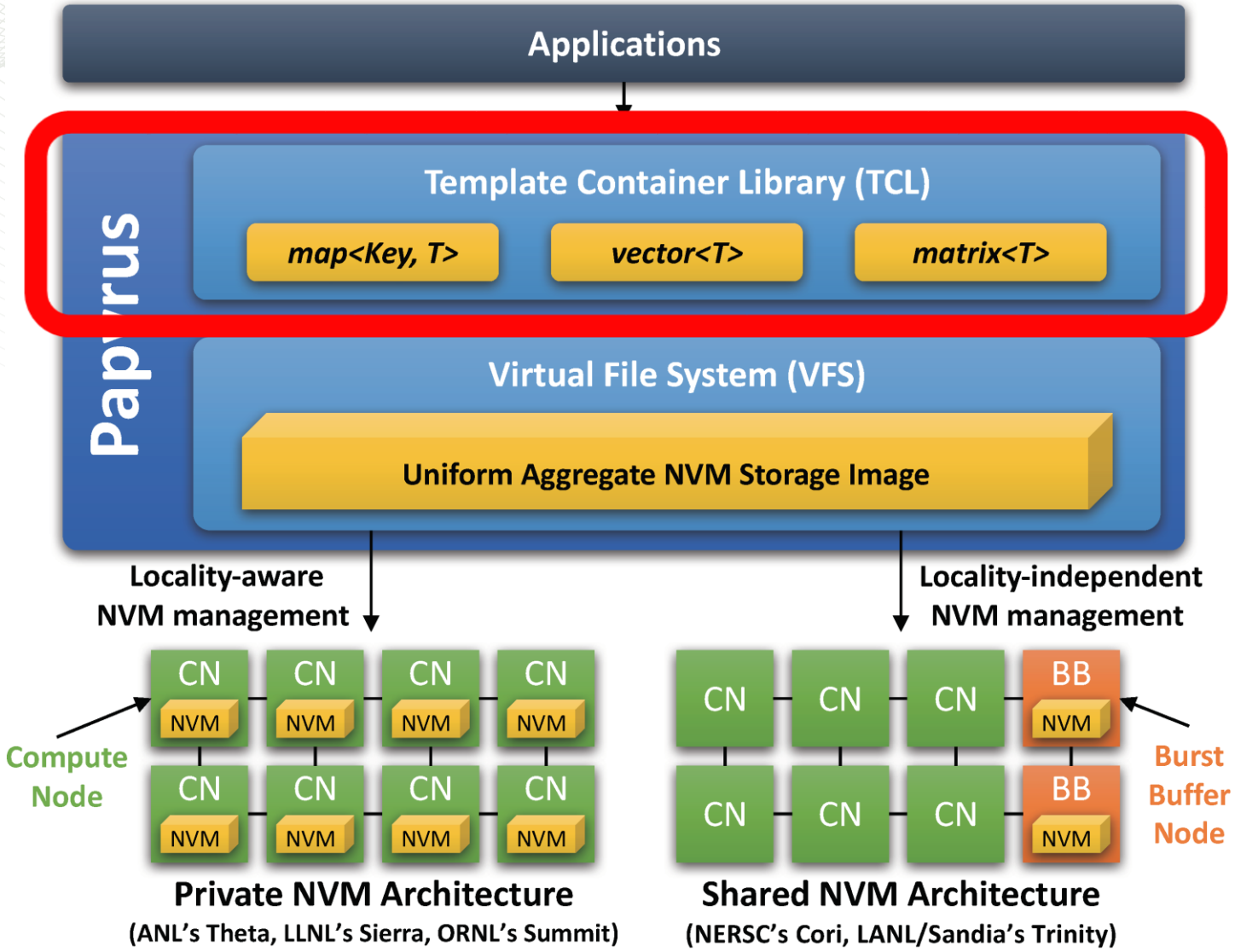
Papyrus VFS Directory Structure

- Uniform aggregate file directory structure across private and shared NVM architectures
- Papyrus root directory
  - Entry point to the aggregate NVM storage image
- Rank directories
  - Same number of ***rank directories*** as the number of the running MPI ranks

- A file on a rank directory ***N*** will be stored on

Private NVM Architecture	Shared NVM Architecture
An NVM in the node that runs MPI rank ***N*** ***(Locality-aware)***	A single NVM or striped over multiple NVMs on burst buffer ***(Locality-independent)***

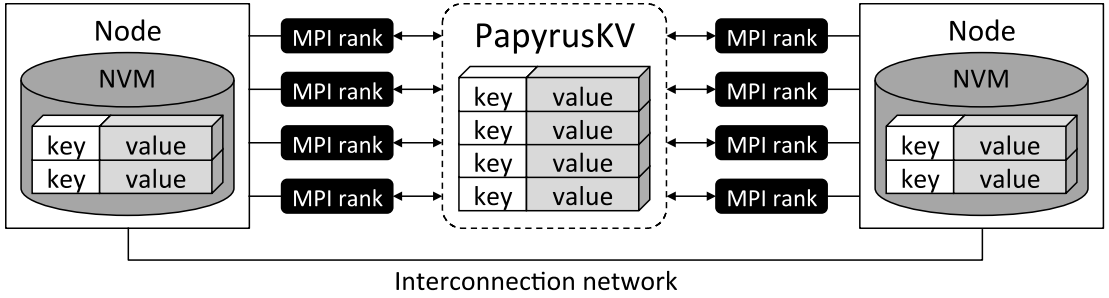
Papyrus Template Container Library (TCL)



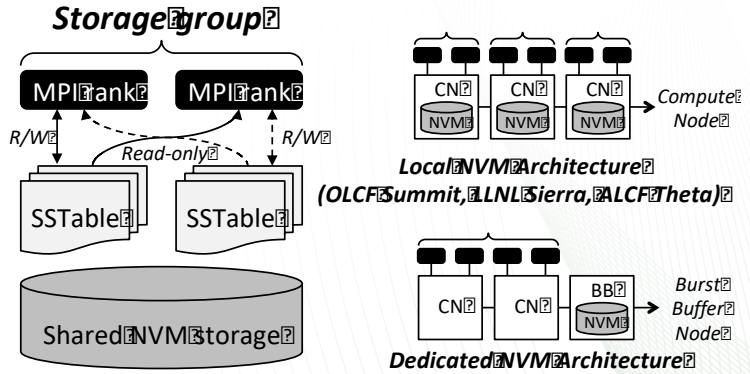
- A high-level programming interface on top of VFS
- Three C++ template containers
  - ***papyrus::map<Key, T>***
    - hashmap
  - ***papyrus::vector<T>***
    - mutable 1D array
  - ***papyrus::matrix<T>***
    - mutable 2D array
- Data elements are
  - Distributed to multiple NVM nodes
  - Globally accessed by all MPI ranks

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

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



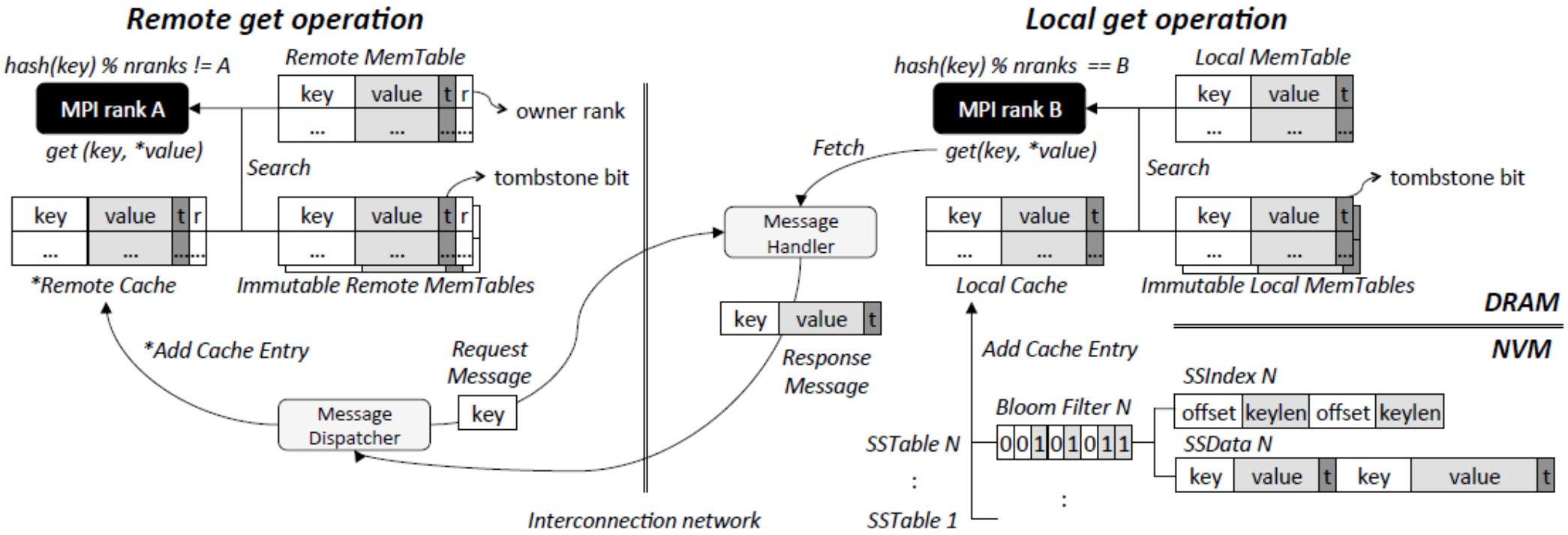
PapyrusKV stores keys and values in arbitrary byte arrays across multiple NVM devices in a distribute system



PapyrusKV is portable across local NVM and dedicated NVM architectures

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

PapyrusKV Example Get operations



Present design allows remote cache only for RO data.

Evaluation

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

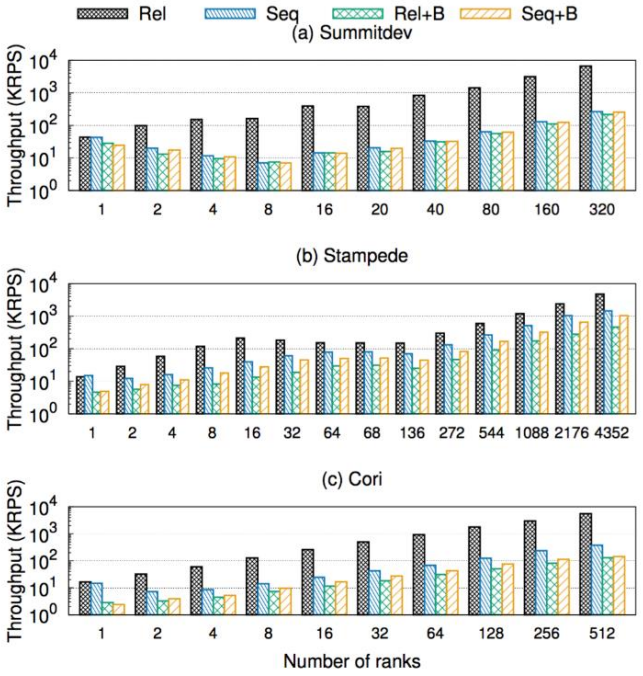


Figure 7: Put operation performance in relaxed (Rel) and sequential (Seq) consistency modes. B refers to Barrier.

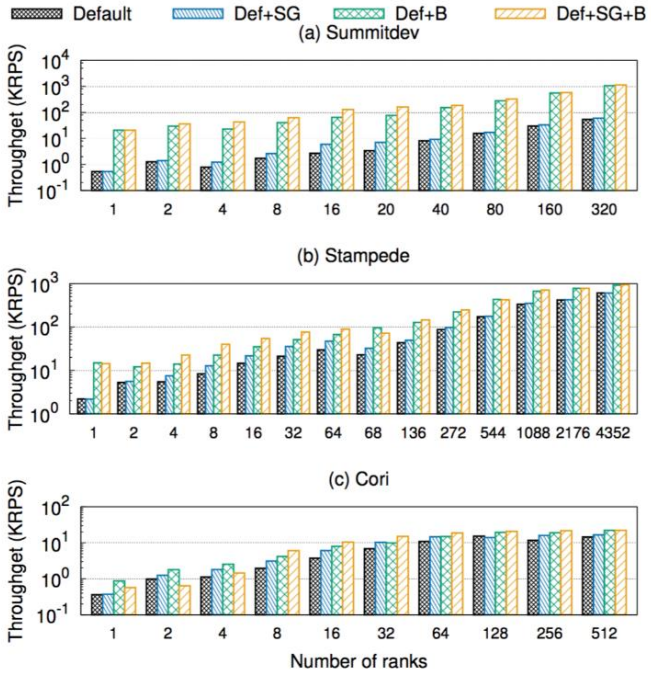


Figure 8: Get operation performance. SG and B refer to Storage Group and SSTable Binary search, respectively.

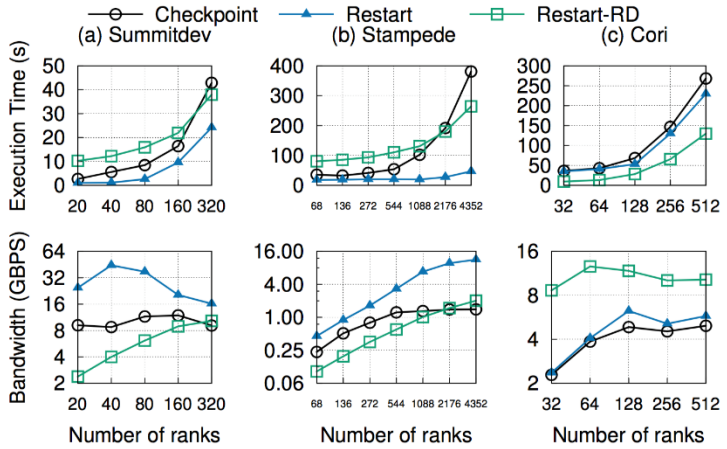


Figure 10: Checkpoint, restart, and restart with redistribution (RD) performance.

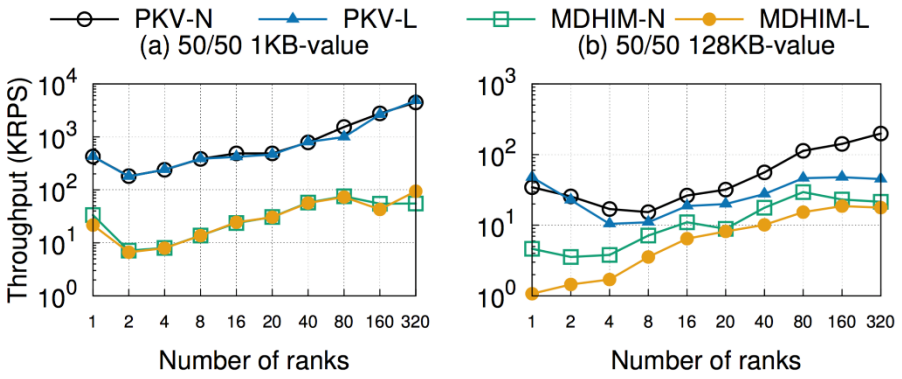
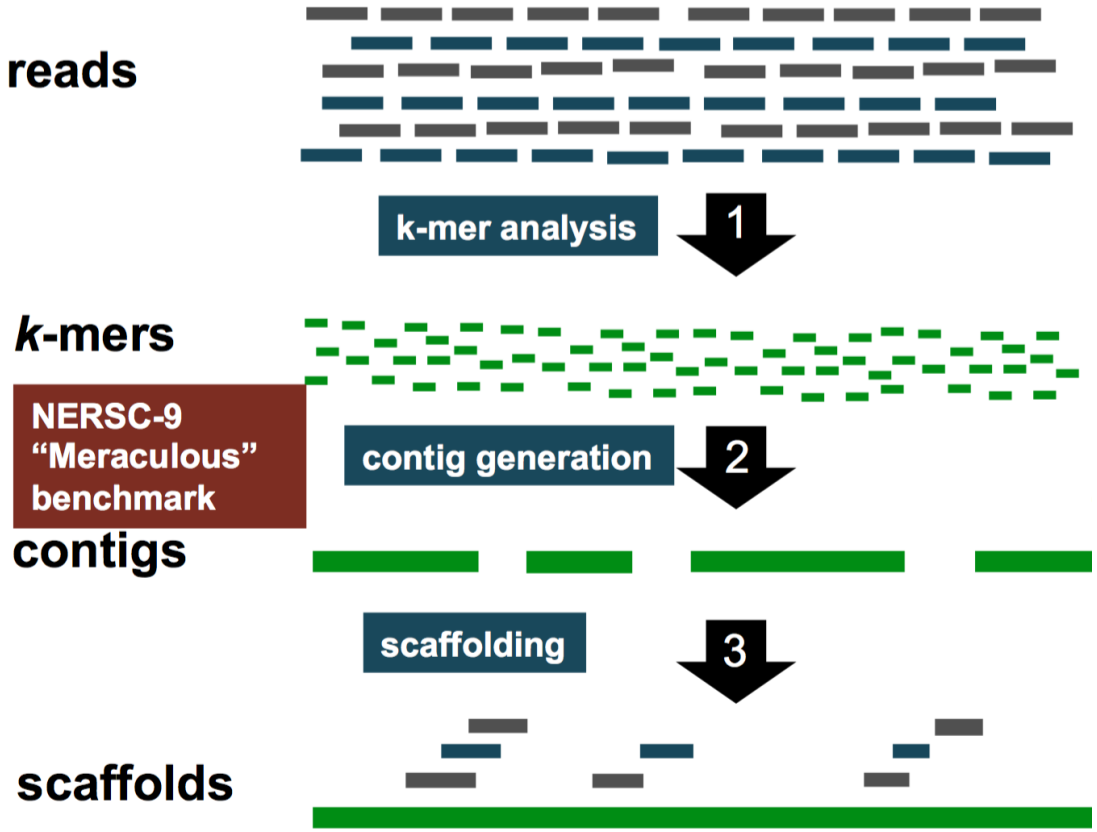


Figure 11: Performance comparisons with MDHIM on Summitdev. NVMe (N) and Lustre (L) are used for their data storages.

ECP Application Case Study 1 Meraculous (UPC)

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



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

Table 1: Source lines of code.

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

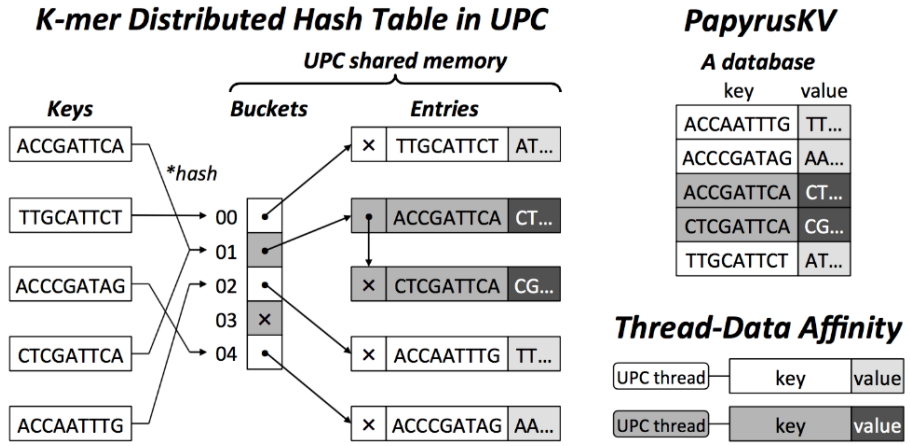


Figure 5: Distributed hash table implementations in UPC and PapyrusKV. *The same user *hash* function in the UPC application can be used in PapyrusKV.

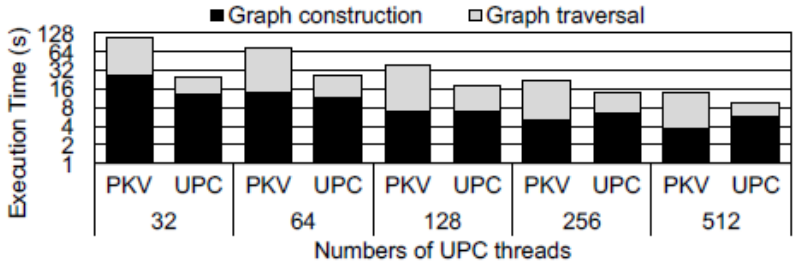
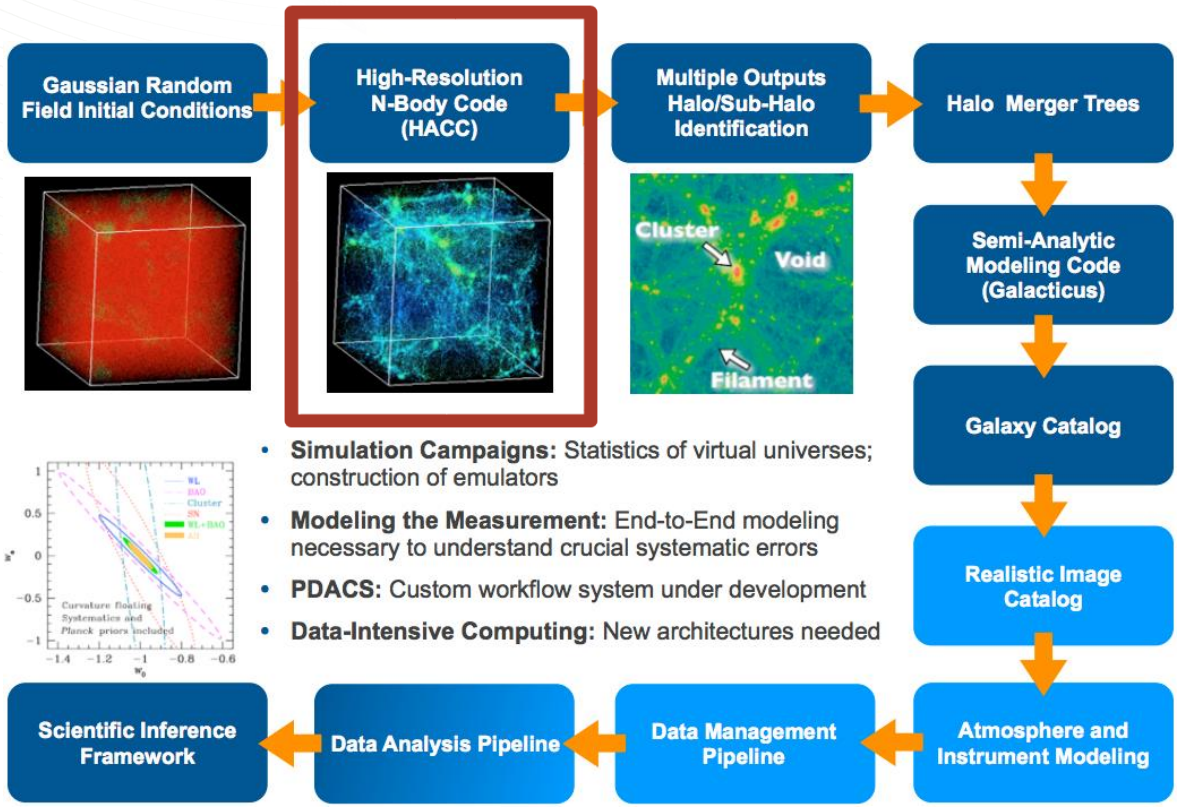


Figure 13: Meraculous performance comparison between PapyrusKV (PKV) and UPC on Cori.

ECP Application Case Study 2: HACC (MPI-IO)

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



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

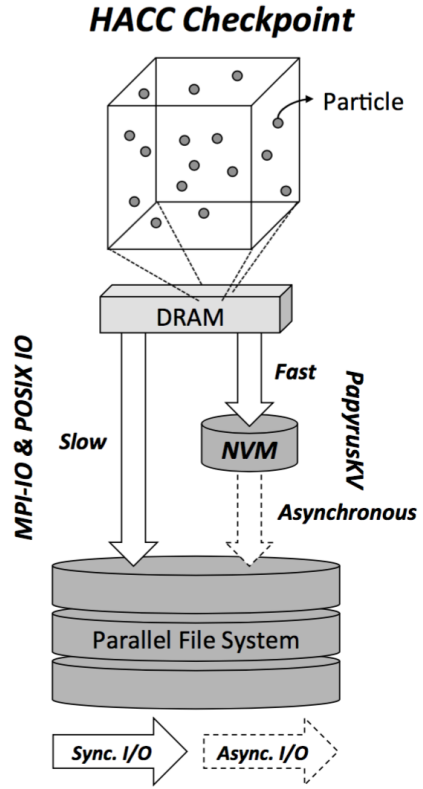


Figure 7: Two-phases checkpointing. PapyrusKV reduces the I/O overhead with help from fast access of NVM. Asynchronous checkpoint hides the I/O overhead between NVM and parallel file system from the application.

WIP: Initial results show about a 10% performance improvement in application performance.

Implications

OAK RIDGE National Laboratory

Implications

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

Summary

- Recent trends in extreme-scale HPC paint an ambiguous future
  - Contemporary systems provide evidence that power constraints are driving architectures to change rapidly (e.g., Dennard, Moore)
  - Multiple architectural dimensions are being (dramatically) redesigned: Processors, node design, memory systems, I/O
- Memory systems are leading the charge in BMC now!
  - New devices
  - New integration
  - New configurations
  - Vast (local) capacities
- Programming systems must support these new memory systems (and portability)!!
  - We need new programming systems to effectively use these architectures
  - NVL-C
  - Papyrus
- Changes in memory systems will dramatically impact systems and applications

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