OAK RIDGE National Laboratory

Tiering in and Peering out
Approaches to Data-Intensive Science

SOS XXIII, 28th March 2019

Mallikarjun (Arjun) Shankar, Ph.D.

and

Sudharshan Vazkhudai, Ph.D.

Oak Ridge Leadership Computing Facility
National Center for Computational Science
ORNL

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

SOS 23 WORKSHOP

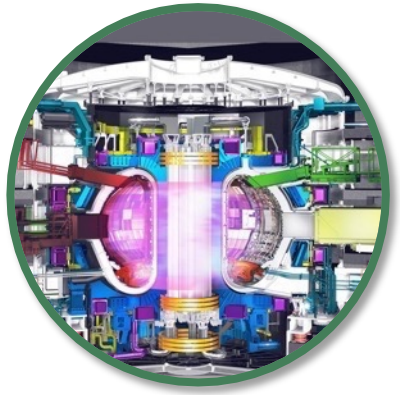
Asheville, NC | Exascale and Beyond

Science challenges for a smart supercomputer:

What can a smart supercomputer do?

Leadership science challenges for a smart supercomputer:

Identifying Next-generation Materials
By training AI algorithms to predict material properties from experimental data, longstanding questions about material behavior at atomic scales could be answered for better batteries, more resilient building materials, and more efficient semiconductors.



Predicting Fusion Energy
Predictive AI software is already helping scientists anticipate disruptions to the volatile plasmas inside experimental reactors. Summit's arrival allows researchers to take this work to the next level and further integrate AI with fusion technology.

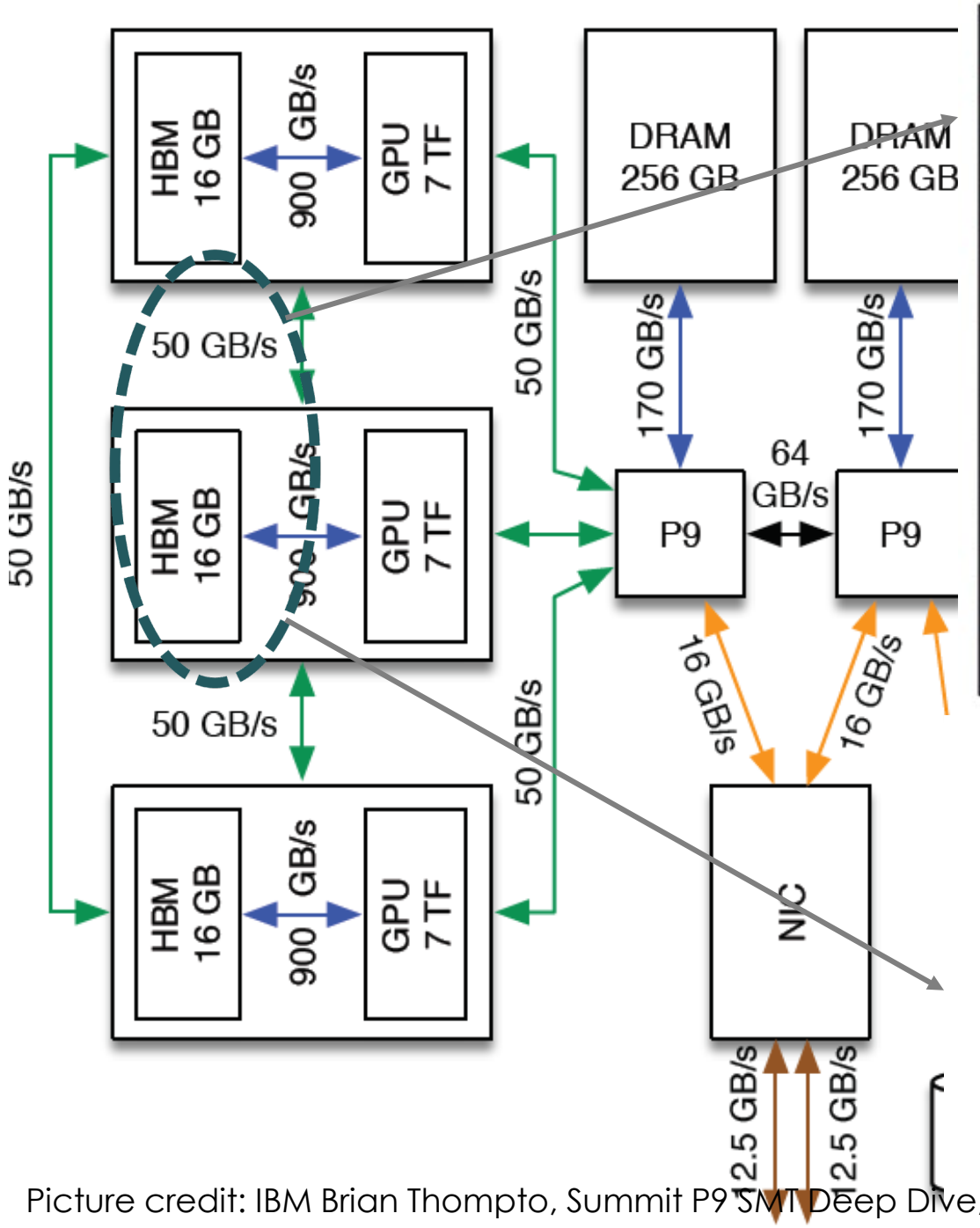
Deciphering High-energy Physics Data
With AI supercomputing, physicists can lean on machines to identify important pieces of information—data that's too massive for any single human to handle and that could change our understanding of the universe.

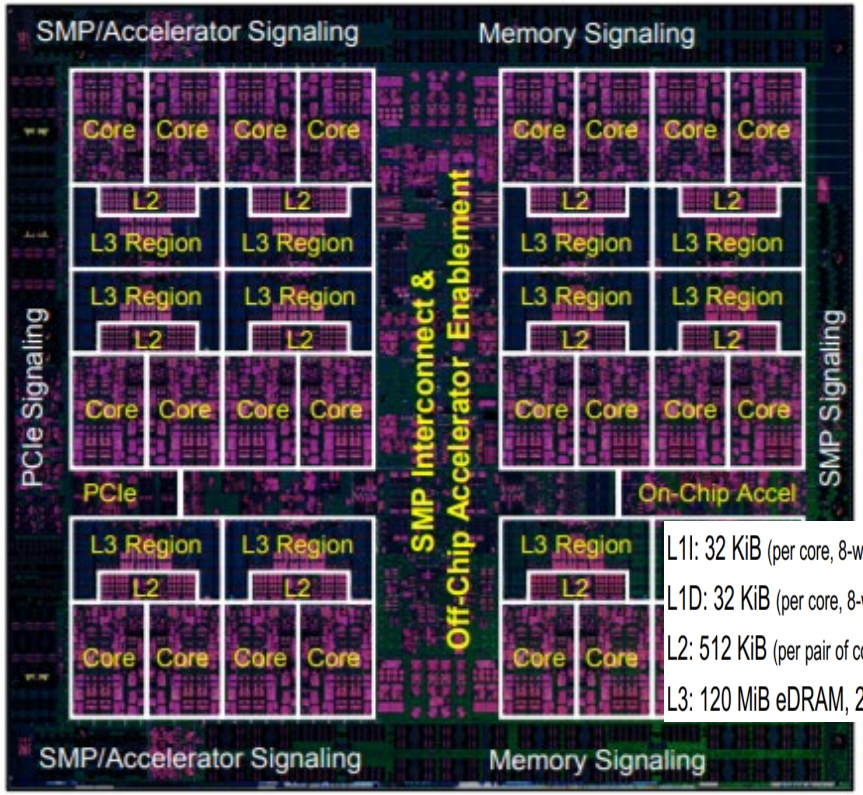
Combating Cancer
Through the development of scalable deep neural networks, scientists at the US Department of Energy and the National Cancer Institute are making strides in improving cancer diagnosis and treatment.

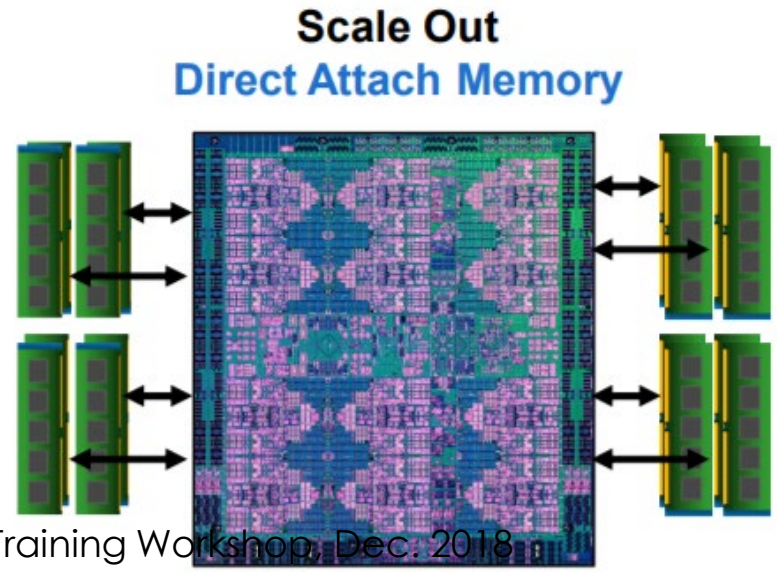
Our Tiered and Peered Data-Rich Landscape

- We are in "Tiered" territory
  - What is this doing for data science?
  - How does this inform our needs looking forward?
- "Peered" facilities for data-intensive science
  - Cross-facility data-intensive science
  - The gap to address and our needs going forward?

Summit P9 CPU & DRAM

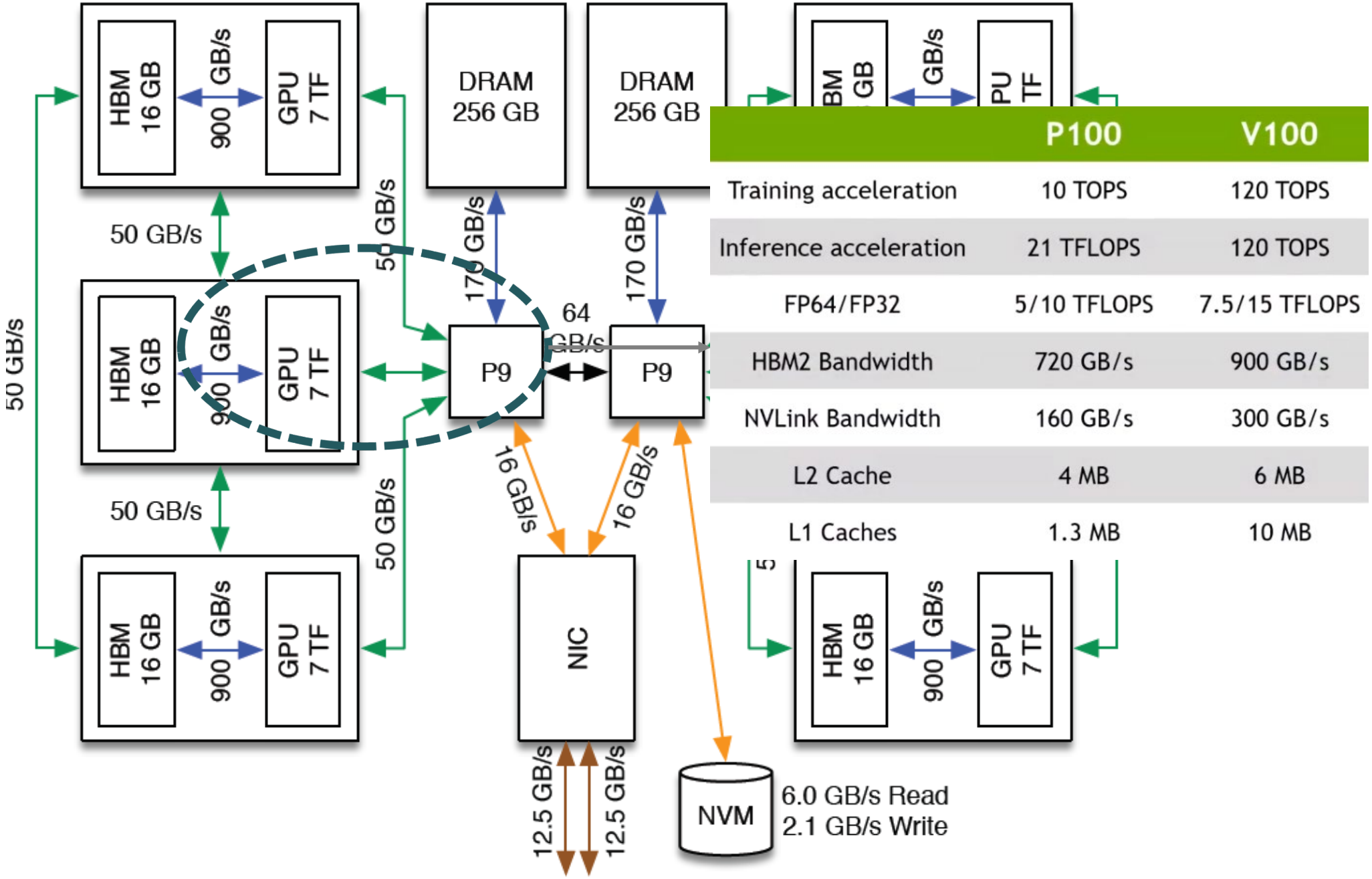






Picture credit: IBM Brian Thompto, Summit P9 SMT Deep Dive, Training Workshop, Dec. 2018

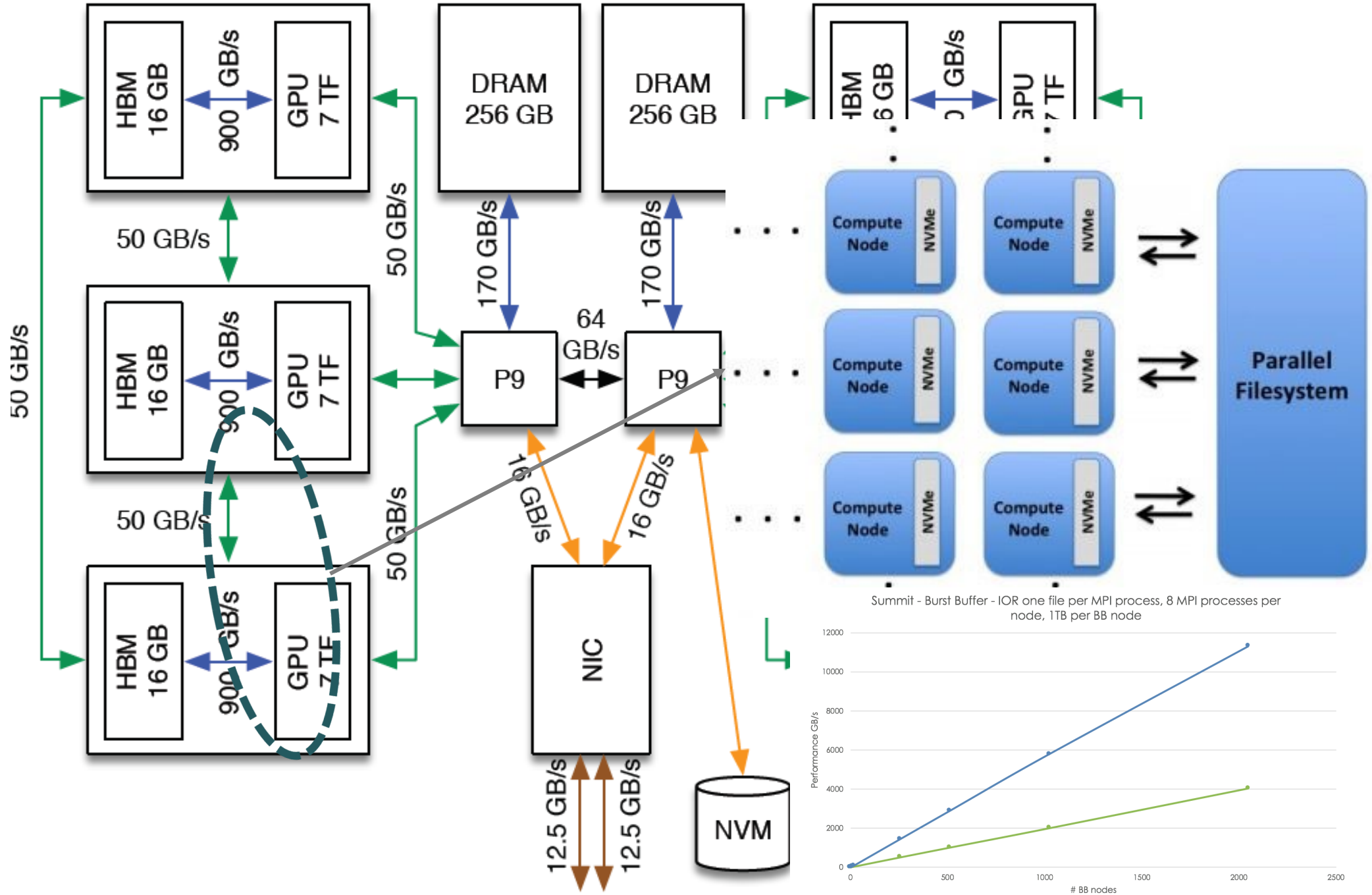
GPU HBM



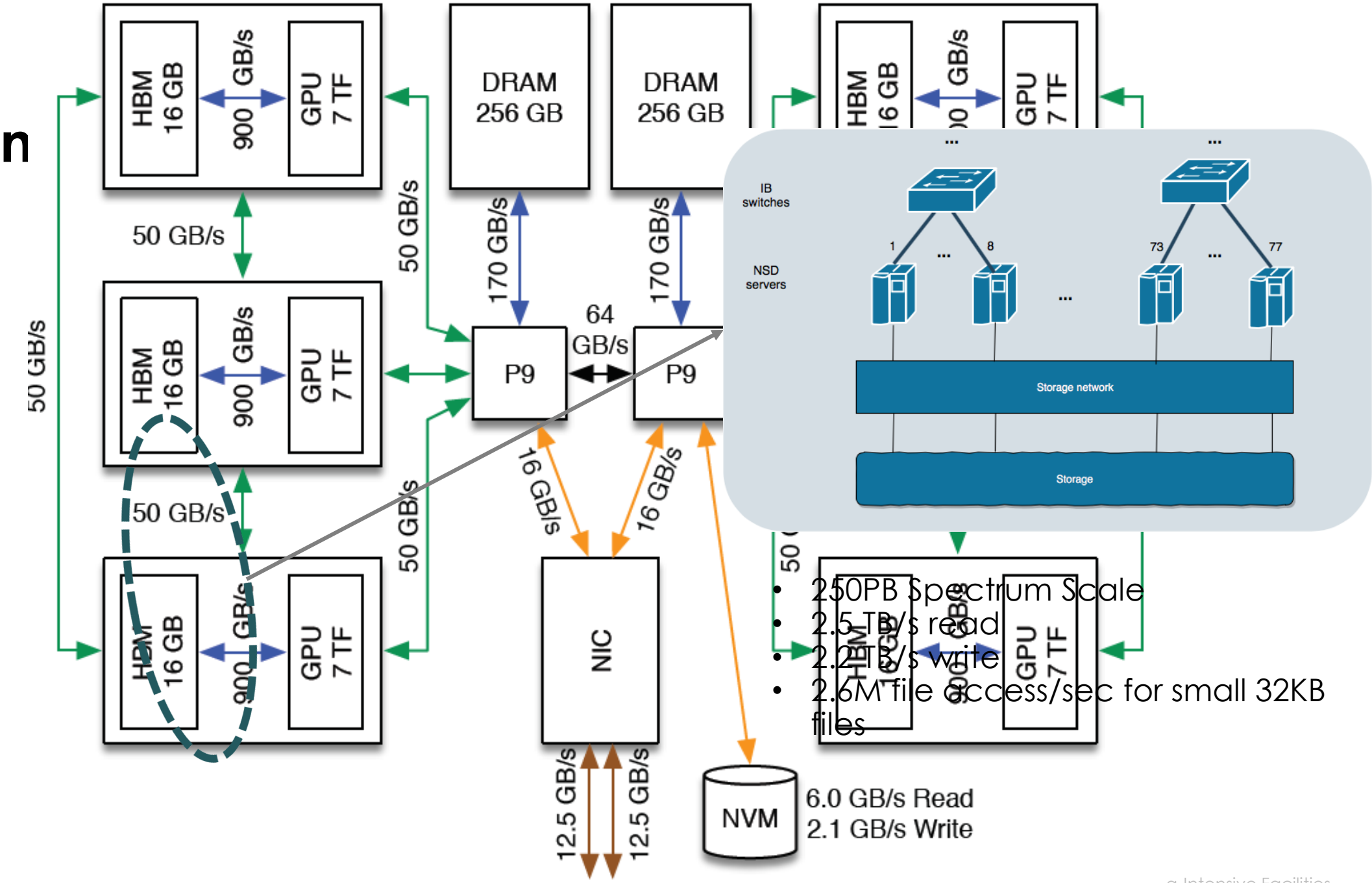
	P100	V100
Training acceleration	10 TOPS	120 TOPS
Inference acceleration	21 TFLOPS	120 TOPS
FP64/FP32	5/10 TFLOPS	7.5/15 TFLOPS
HBM2 Bandwidth	720 GB/s	900 GB/s
NVLink Bandwidth	160 GB/s	300 GB/s
L2 Cache	4 MB	6 MB
L1 Caches	1.3 MB	10 MB

Table credit: NVIDIA, Jeff Larkin, Summit Training Workshop, Dec. 2018

NVMe



Center-Wide Filesysten



- 250PB Spectrum Scale
- 2.5 TB/s read
- 2.2TB/s write
- 2.6M file access/sec for small 32KB files

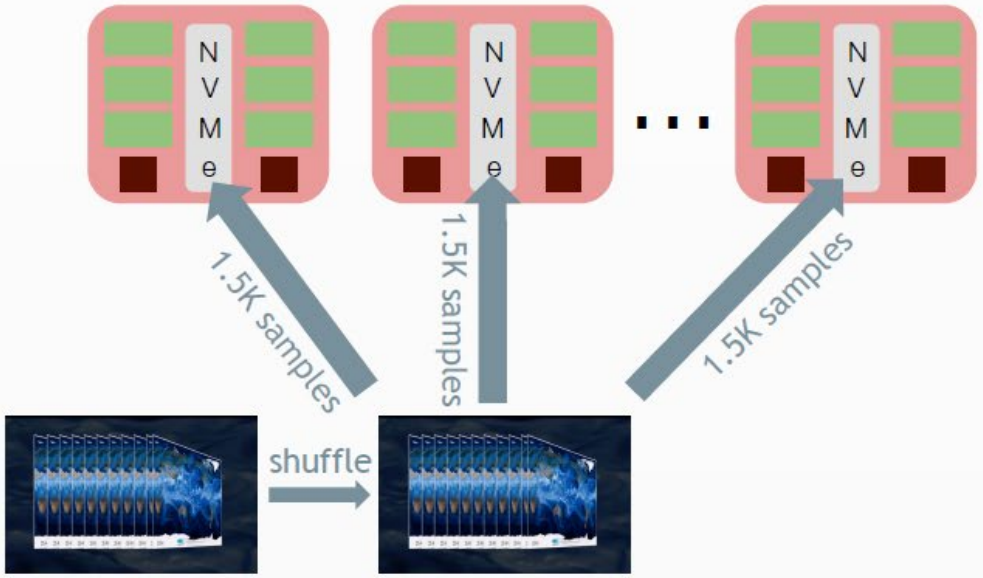
OAK RIDGE National Laboratory

Example 1: Exascale Climate Application

SC 18 Gordon Bell Award: Thorsten Kurth, Sean Treichler, Joshua Romero, Mayur Mudigonda, Nathan Luehr, Everett Phillips, Ankur Mahesh, Michael Matheson, Jack Deslippe, Massimiliano Fatica, Prabhat, Michael Houston; LBNL, NVIDIA, ORNL

Data Staging

Dataset Size	Required BW (27K GPUs)	GPFS/LUSTRE	BurstBuffer	NVM/e or DRAM
20 TB (~63K samples)	3.8 TB/s	~400 GB/s	~2 TB/s	~26 TB/s

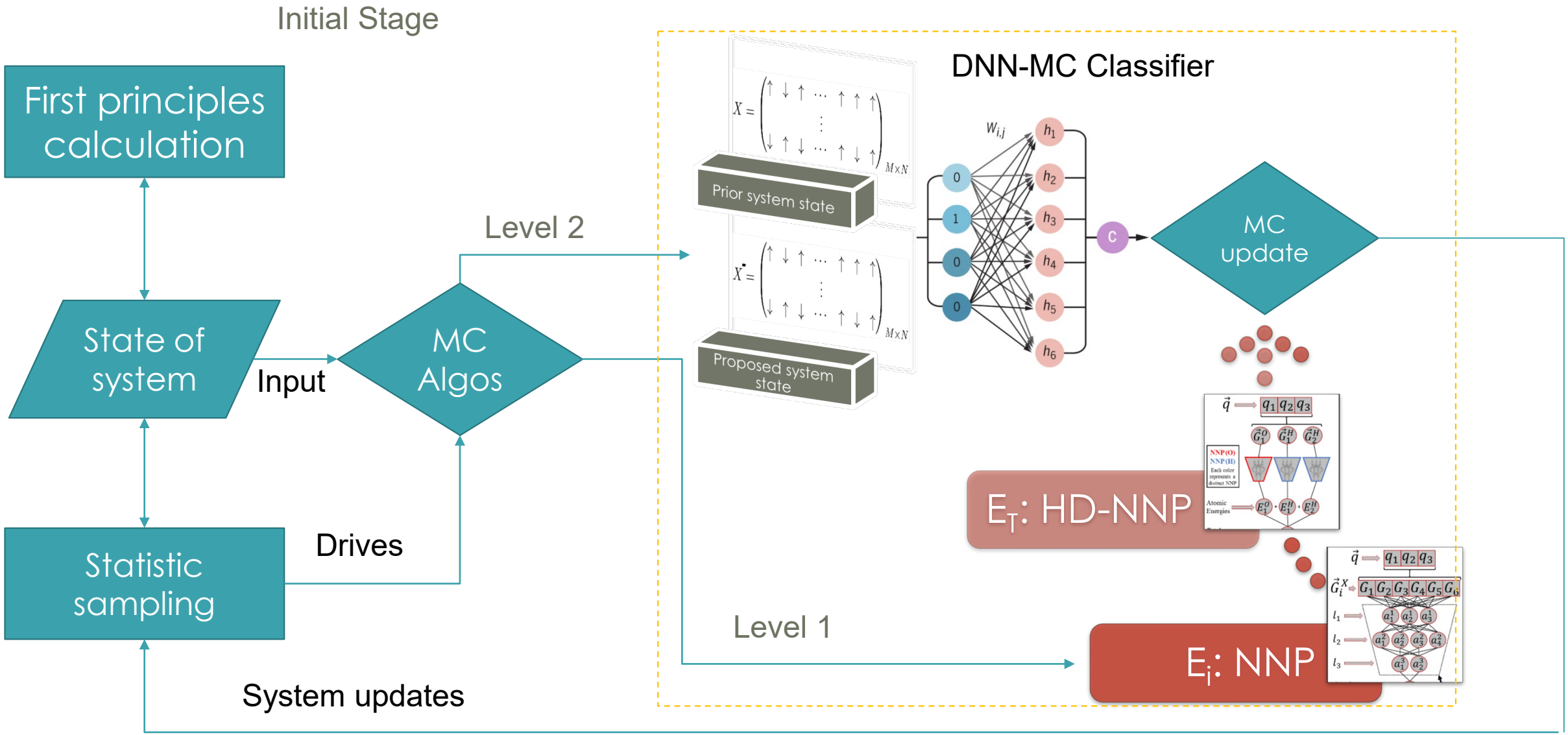


- 250 training samples/GPU (~15 GB), sample w/ replacement
- each file will be read at most once from FS
- files shared between nodes via MPI (`mpi4py`)

NERSC

Picture from Thorsten Kurth's Dec. 2018 ASCAC presentation.

Example 2: Simulation/ML-DL Interleaving
Materials Application: Surrogate modeling for MC



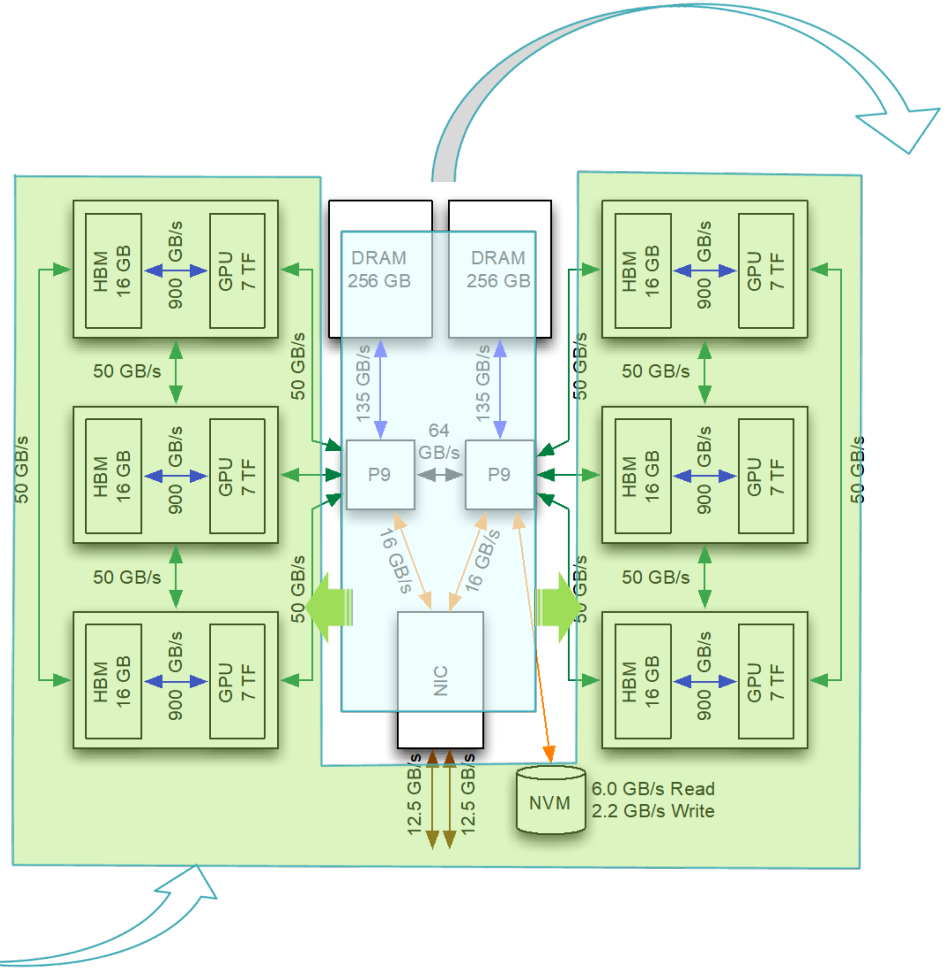
OAK RIDGE National Laboratory

M. Eisenbach, V. Cooper, J. Zhang, Z. Pei, Y.W. Li, M. Pasini, and J. Yin, *Machine Learning Surrogate Models to Accelerate Monte-Carlo Calculation,* Bulletin of the American Physical Society, March 2019

Data Science and Learning on the Memory Hierarchy

CORAL 2 Benchmark Suite	Description
Big Data Analytic Suite (BDAS)	PCA, K-Means, and SVM (based on pbdR)
Deep Learning Suite (DLS)	CANDLE, CNN, RNN, and ResNet-50 (distributed memory)

Traditional Node: PCA, K-Means, etc. excel due to the node's memory, CPU, and on-chip bandwidth



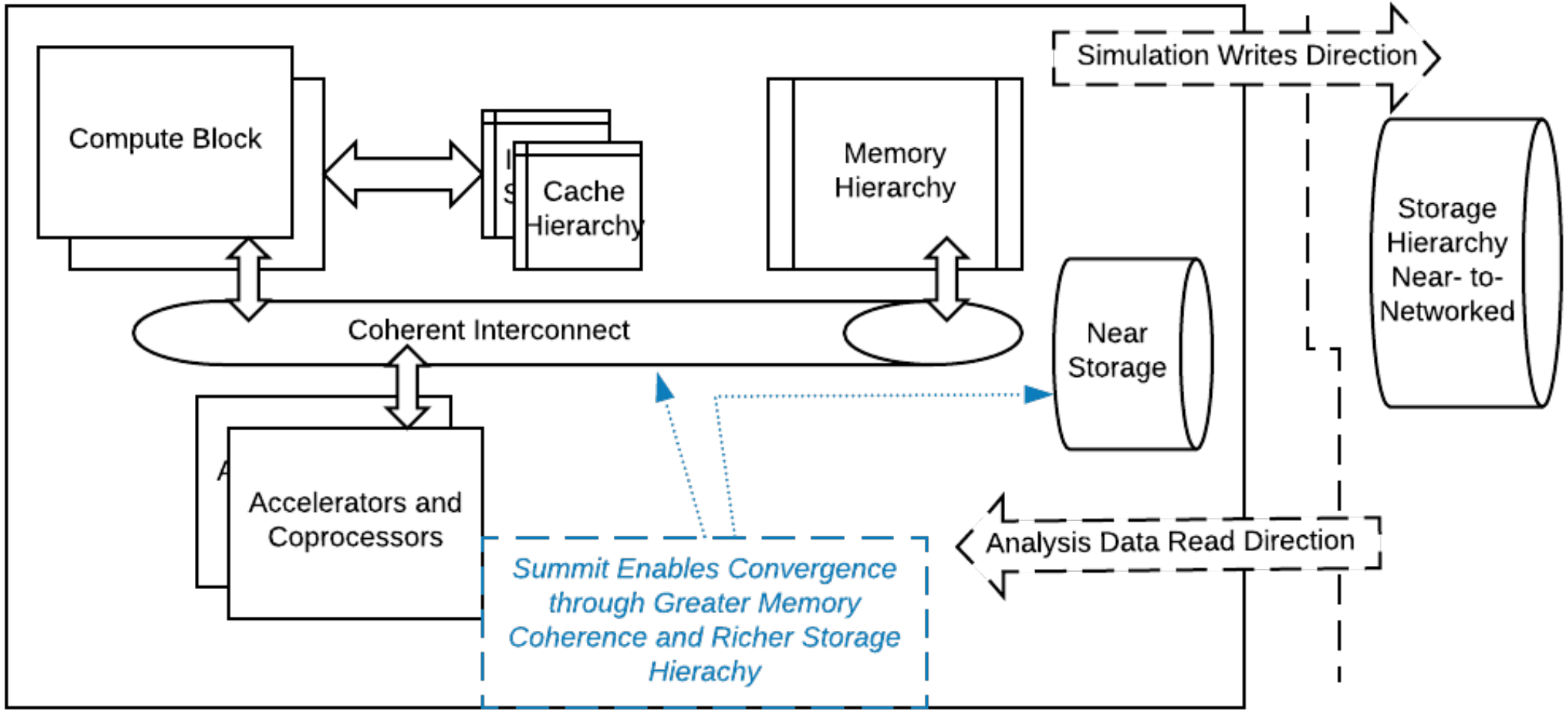
Deep Learning Codes (CNN; ResNet50; ..) excel here with NVM and GPUs enabling tensor operations.

Code suites are in the CORAL (Collaboration of Oak Ridge, Argonne, Livermore laboratories data benchmark suite: https://asc.llnl.gov/coral-2-benchmarks

Accommodating the Data-Science Landscape

Model-Driven and Simulation-Heavy Workloads	Data-Driven and Analytics (ML/DL) Workloads
Single Center-Wide Namespace	Read-Intensive Partitioned Namespace OK
Write Fraction (~15%) of HBM Rapidly (~5 Minutes)	High IOPs
Large Capacity Retention	Large Capacity Retention
N-N, N-1 write access patterns	Small, Random Reads

Tiering has moved us towards: “Convergence”



A rich tiered data hierarchy is a necessity given our **technology** (bandwidth, latency) and **cost** (per bit) landscape!

Tools to help the user to seamlessly process data across node-local storage, and node-local and center-wide storage are needed.

Shameless plug: ORNL is working on a couple - Spectral and Symphony FS.

HPC Cloud-Edge-**Fogging up the data-story…**

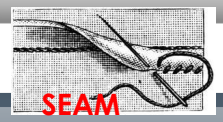
Scalable Cross-Facility Workflow and Ecosystem for Data Science

- *CADES, ...*
  - *Cloud, SHPC, DGX*
  - *Cross-Facility Data*
- *Globus*
- *PANDA, Fireworks, Pegasus, Swift, ...*

Data On-Ramps:
Observational Data
Scalable Computing

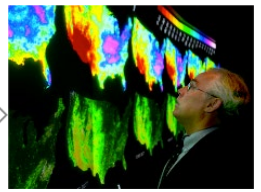
- *ADIOS*
- *Kubernetes*
  - *Slate/Marble*
  - *Slate/Onyx*
- *Singularity*

Workflows:
Staging and Cross-System and Cross-Facility



summit

- **Artifacts and Insights**



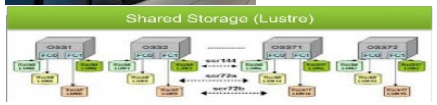
Capability Compute at OLCF:

Simulation &
Compute/Data Workflows at Scale

RHEA

EOS

CADES



Data Life-Cycle:
Dissemination, Sharing, Analytics Products

- *Constellation*
- *EDEN, Visit, Paraview, SIGHT*
- *CrossBOW*
- *Support: PDACS, Galaxy, Zeppelin, ...*

- *Well-known runtime and programming language extensions.*
- *Exploring new scheduling approaches.*

Analytics:
Analysis at Scale
Visual Analytics

- *pbdR*
- *Spark on Demand*
- *TensorFlow, PyTorch, Keras, Caffe, MxNet*

To Conclude..

1. Tiering: Need tools to support data processing across heterogeneous compute and memory hierarchy, node-local storage, and node-proximal and center-wide storage.

2. Peering: Need tools to support distributed data read/process/write abstractions across data-stores in data-intensive facilities.

#1 and #2 call for similar conceptual abstractions of programming primitives. The implementation will differ based on data affinity, currency, and latency awareness. Both could use an underlying data handling layer/API.

Thank You!

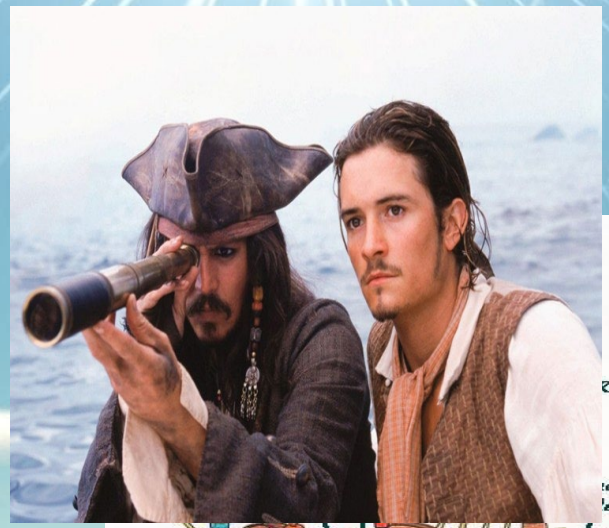
Acknowledgements:

OLCF/NCCS Colleagues (Junqi Yin, Jack Wells, Chris Zimmer); IBM; NVIDIA; Mellanox

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory (ORNL). *ORNL is operated by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.*

The United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. In addition, this work is supported by the Laboratory Director R&D Fund.

Peering Reinterpreted



Observational Facility

Computational Facility