Leadership AI at the OLCF

Jack C. Wells

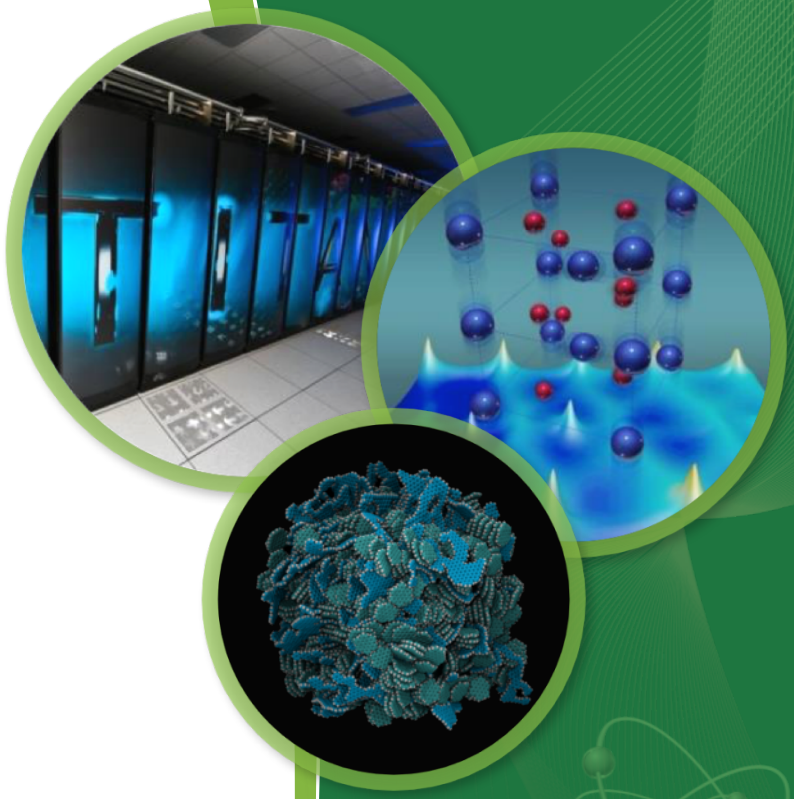
Director of Science
Oak Ridge Leadership Computing Facility
Oak Ridge National Laboratory

Workshop on Machine Learning in HPC Environments

@SC17

13 November 2017

Denver



This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. Some of the work presented here is from the TOTAL and Oak Ridge National Laboratory collaboration which is done under the CRADA agreement NFE-14-05227. Some of the experiments were supported by an allocation of advanced computing resources provided by the National Science Foundation. The computations were performed on Nautilus at the National Institute for Computational Sciences.

Machine Learning*

- Machine learning is a type of ***artificial intelligence*** (***AI***) that provides computers with the ability to learn without being explicitly programmed
  - Machine learning focuses on the development of computer programs that can change when exposed to new data
- The process of machine learning is similar to that of data mining
  - Both systems search through data to look for patterns, but instead of extracting data for human comprehension -- as is the case in data mining applications -- machine learning uses that data to detect patterns in data and adjust program actions accordingly
- Machine learning algorithms are often categorized as being supervised or unsupervised
  - Supervised algorithms can apply what has been learned in the past to new data. Unsupervised algorithms can draw inferences from datasets.

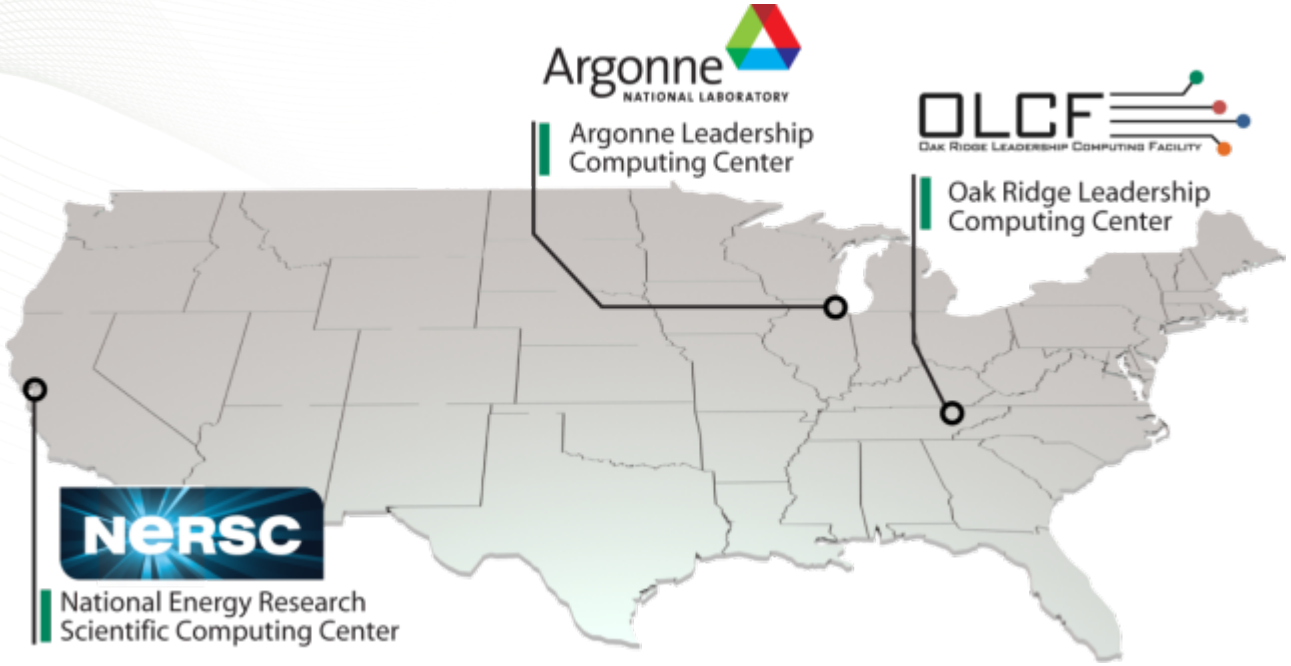
*http://whatis.techtarget.com/definition/machine-learning

AI (Artificial Intelligence) is Rapidly Advancing in Today's World

Domain	Domain Examples	Sample Techniques	DOE Science Overlap
Voice recognition and textual assistants	Siri, Alexa, Wolfram	Deep Learning, Semi-supervised Learning	Moderate
Games and search	Alpha-Go, Rubiks-Robot, Watson	Large-scale classification and semi-supervised and reinforcement learning	Moderate
Recommender systems	Amazon, Netflix	Supervised Learning	Low
Weather and climate prediction, Earthquake predictions	Weather.com, NHC	Large-Scale Classification and Prediction, and Extremes Prediction	High
Driverless cars	Gcar, Tesla	Image processing, classification, multi modal fusion	Moderate to High
Large-Graph analysis, Linked structures	Omics, Multivariates	Graph Traversal, Clustering, Tensors	High
Quantum Physics and Chemistry	Many Body Problem, Monte-Carlo methods	Clustering, Deep Neural Networks	High

1. 1000 years ago: **experimental science** - description of natural phenomena
2. Last few hundred years: **theoretical science** - Newton's Laws, Maxwell's Equations
3. Last few decades: **computational science** - simulation of complex phenomena
4. Present: **data-intensive science** - move from data to information to knowledge

DOE's Office of Science, Advanced Scientific Computing Research (ASCR) Computation User Facilities



NERSC
Edison is 2.57 PF

ALCF
Mira is 10 PF

OLCF
Titan is 27 PF

- DOE is a leader in open High-Performance Computing
- Provide the world's most powerful computational tools for open science
- Access is free to researchers who publish
- Boost US competitiveness
- Attract the best and brightest researchers

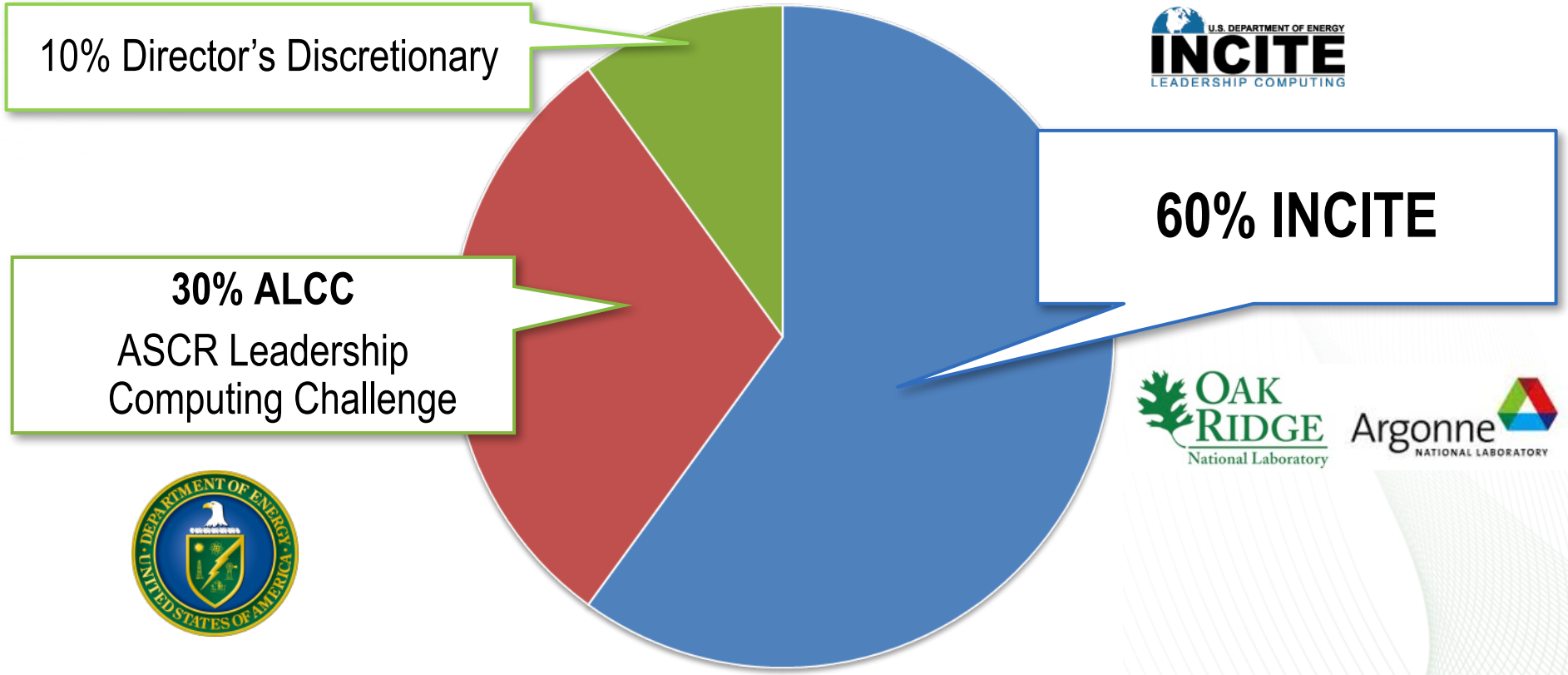
What is the Leadership Computing Facility (LCF)?

- Collaborative DOE Office of Science user-facility program at ORNL and ANL
- Mission: Provide the computational and data resources required to solve the most challenging problems.
- 2-centers/2-architectures to address diverse and growing computational needs of the scientific community
- Highly competitive user allocation programs (INCITE, ALCC).
- Projects receive 10x to 100x more resource than at other generally available centers.
- LCF centers partner with users to enable science & engineering breakthroughs (Liaisons, Catalysts).

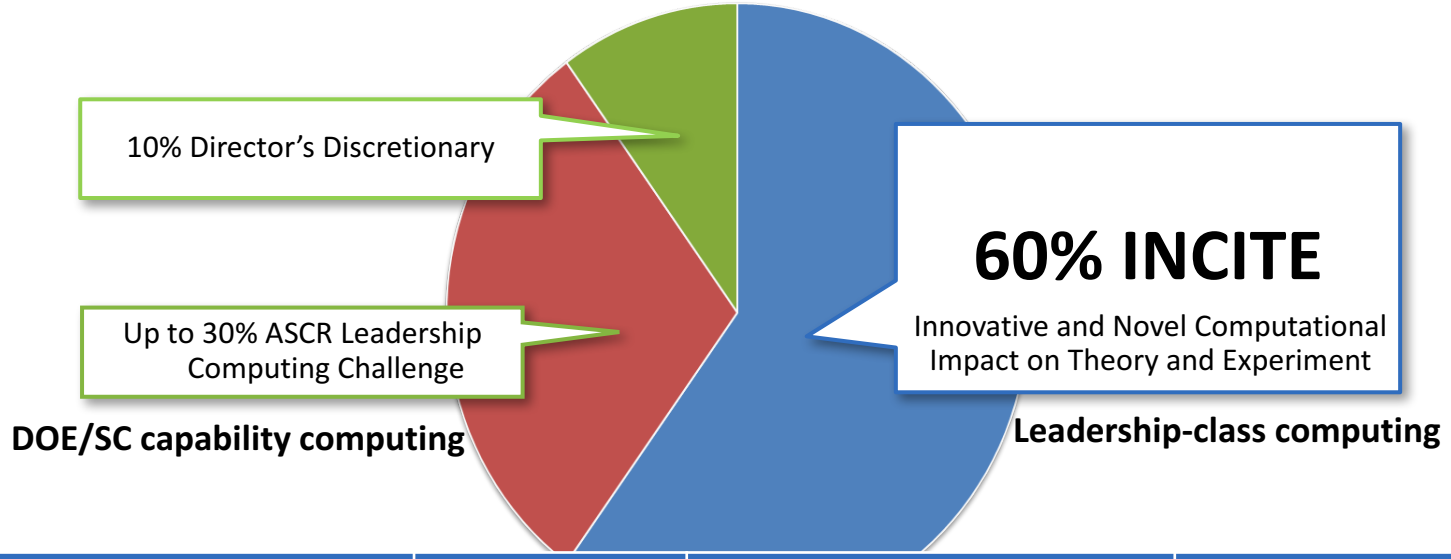
OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

Three primary user programs for access to LCF

Distribution of allocable hours



User Programs & Characteristics

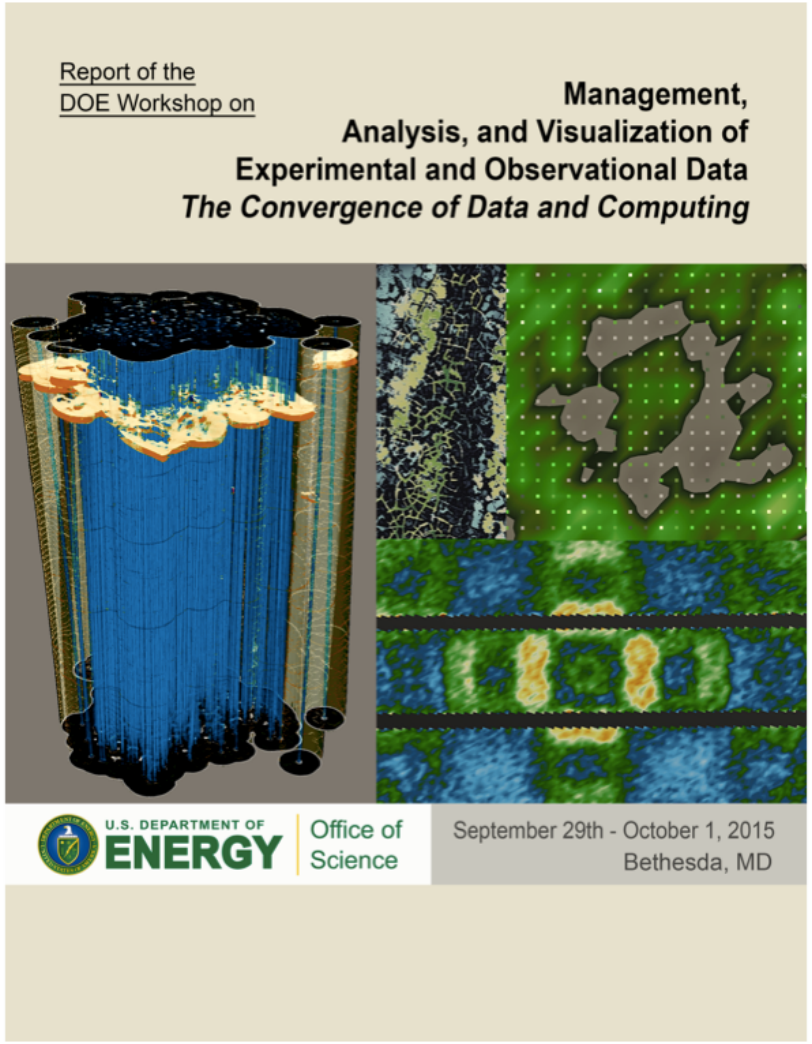


	Target	Review Process	Allocation Range for Titan	Proposal Submission Deadline	Allocation Period
INCITE	Open Science	1. Independent Peer Review 2. Computational Readiness	50-150 Mch	New: June Renewal: July	January – December Up to three years
ALCC	DOE Mission Science	DOE ASCR: Merit Review & Program Priority	10-250 Mch	February	July-June One year
DD	1. HPC Preparation 2. User Base Extension 3. ORNL Agenda	OLCF Internal Review	<5 Mch	None	Anytime, < one year

Outline

- Quick introduction to machine learning & leadership computing
- Science requirements gathering: highlight DOE/SC/ASCR workshop reports.
- Case Studies:
  - Multimodal characterization of materials
  - Predict plasma disruptions in tokamak fusion reactor
  - Vertex reconstruction in neutrino physics experiments
  - Big health data analytics: massive-scale-analysis of pathology reports
- Discussion of Policy & Technical Challenges
- Highlight OLCF's Summit Project.

Experimental and Observational Science Data is Exploding

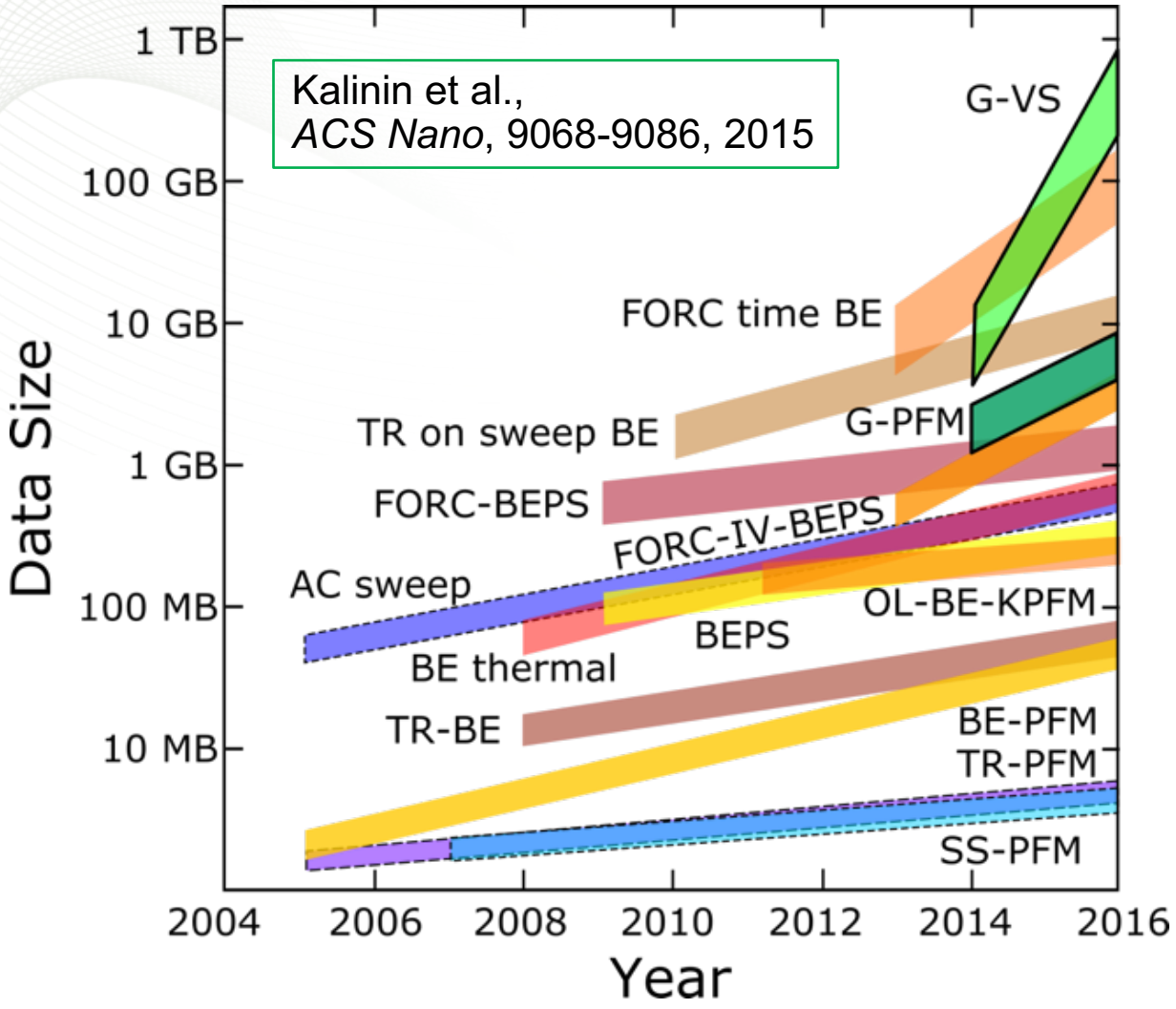


DOE Workshop: Data Management, Analysis and Visualization for Experimental and Observational Data (EOD) Workshop, (2015)

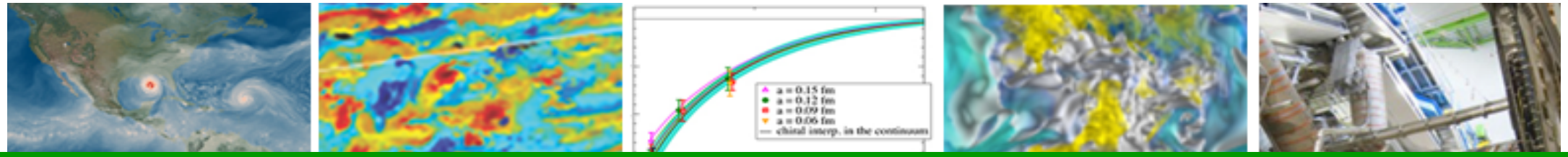
Program Leader: Lucy Nowell, DOE/ASCR

https://extremescaleresearch.labworks.org/events/data-management-visualization-and-analysis-experimental-and-observational-data-eod-workshop

Experimental and Observational Science Data is Exploding



- **Growing data sizes & complexity**
  - Cannot use desktop computers for analysis
  - ➢ **Need HPC!**
- **Multiple file formats**
  - Multiple data structures
  - Incompatible for correlation
  - ➢ **Need universal, scalable, format**
- **Disjoint and unorganized communities**
  - Similar analysis but reinventing the wheel
  - Norm: emailing each other code, data
  - ➢ **Need centralized repositories**
- **Proprietary, expensive software**
  - ➢ **Need robust, open, free software**



DOE-ASCR Exascale Requirements Reviews

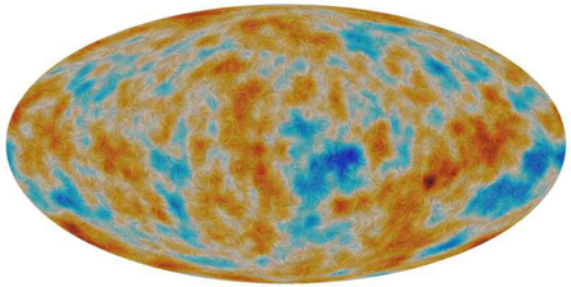
ASCR facilities conducted six exascale requirements reviews in partnership with DOE Science Programs

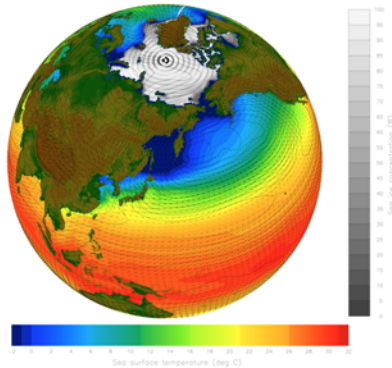
- Goals included:
  - Identify mission science objectives that require advanced scientific computing, storage and networking in exascale timeframe
  - Determine future requirements for a computing ecosystem including data, software, libraries/tools, etc.

Schedule

June 10–12, 2015	HEP
November 3–5, 2015	BES
January 27–29, 2016	FES
March 29–31, 2016	BER
June 15–17, 2016	NP
Sept 27–29, 2016	ASCR
March 9–10, 2017	XCut

Common Themes Across DOE Science Offices
Data: Large-scale data storage and analysis





BIG DATA Analytics

Experimental and simulated data set volumes are growing exponentially.
Examples: High luminosity LHC, light sources, climate, cosmology data sets ~ 100s of PBs.
Current capability is lacking.

Methods and workflows of data analytics are different than those in traditional HPC.
Machine learning is revolutionizing field.
Established analysis programs must be accommodated.

DOE/SC Requirements Crosscut Report:

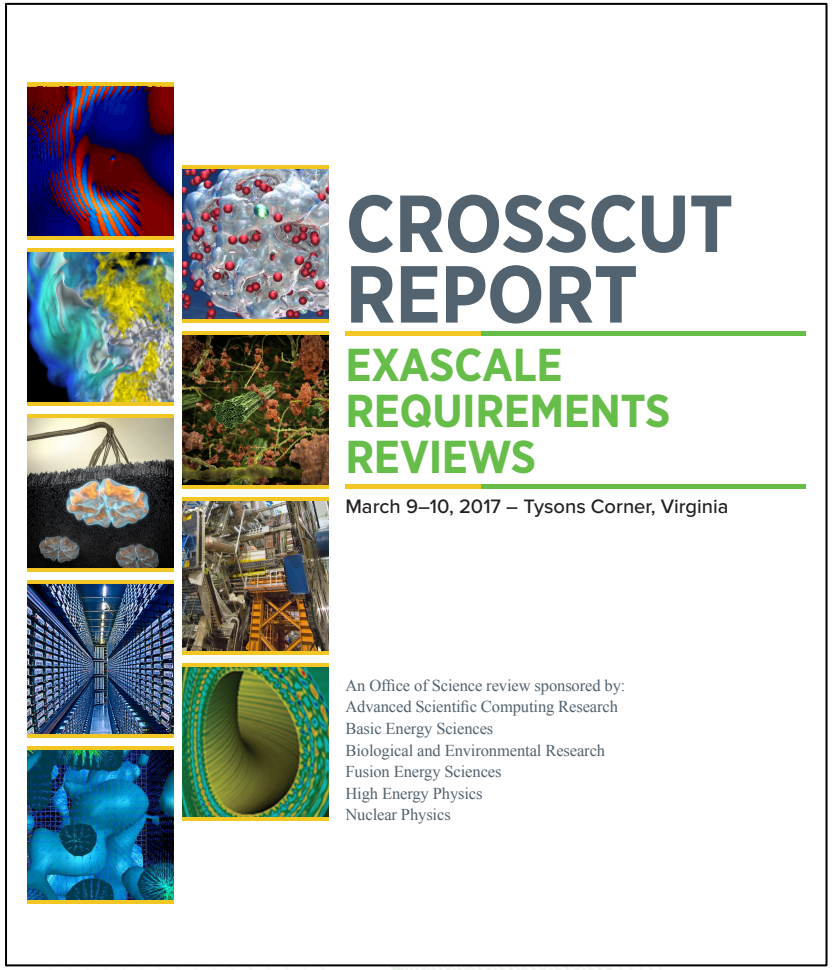
Executive summary findings support machine-learning needs

Data:

- "[…] performing analyses of big datasets and drawing inferences based on these data are revolutionizing many fields. **New approaches are needed for analyzing large datasets including advanced statistics and machine learning.**"

Software and Application Development:

- **Scalable data processing, data analysis, machine learning, discrete algorithms**, and multiscale/ multiphysics simulations are crucial for reducing and **understanding the large-scale data that will be produced by exascale systems.**



All 6 workshop reports are available online, X-Cut online soon: http://exascaleage.org/

Example: Multimodal characterization of materials

Novel microscopic and spectroscopic techniques allow characterization of the different aspects of the materials at the nanosacle.

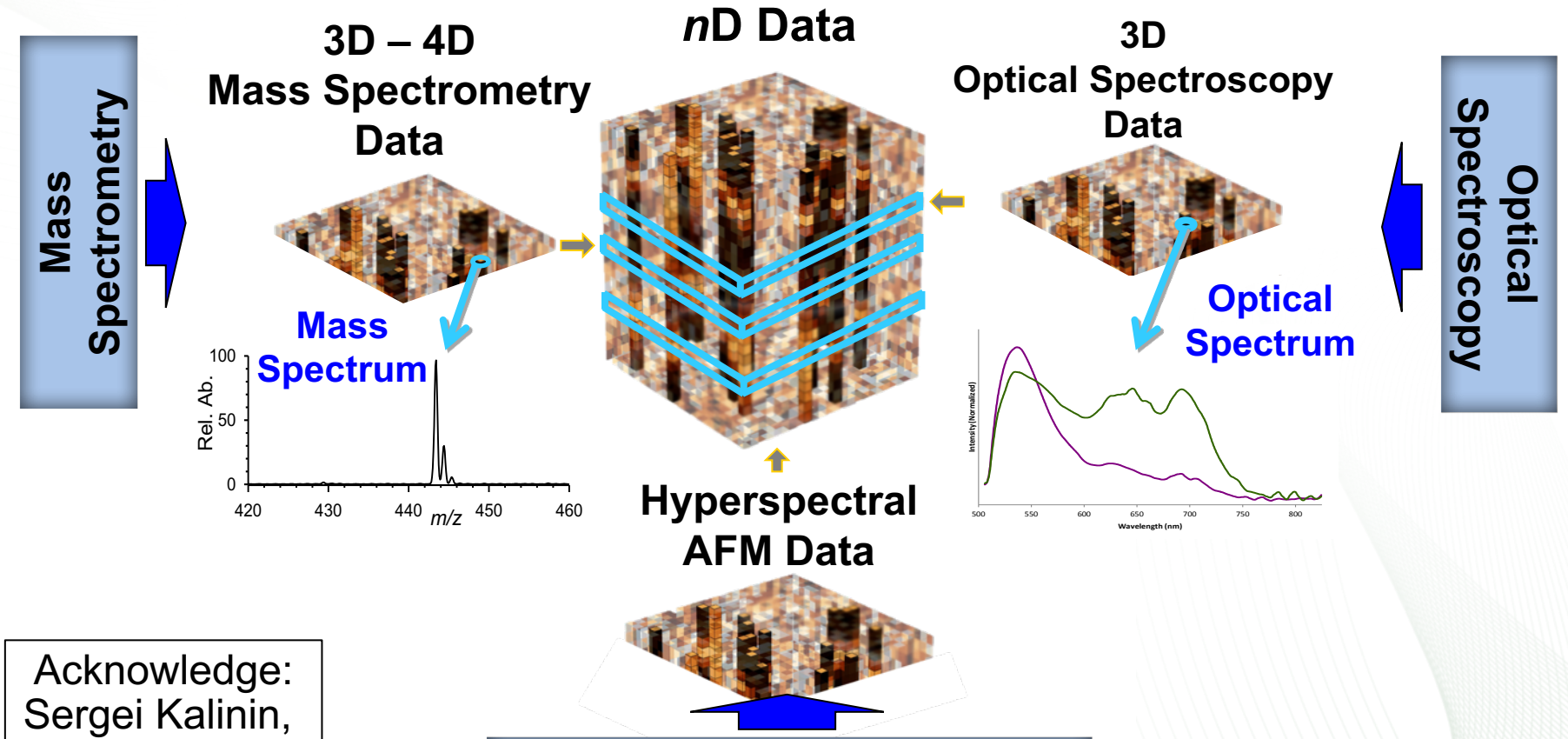
- Associated data analysis is difficult: Outcome data is big (up to Tb) and multidimensional

Combination of the microscopic and spectroscopic techniques (“multimodal”) is required for comprehensive characterization of materials.

- Example, Secondary Ion Mass Spectrometry (SIMS) can be combined with Atomic force microscopy (AFM), enabling correlated characterization of functional response at the nanoscale (AFM) with chemical composition (SIMS).
- Optical spectrometry can be introduced for characterization of the optical properties and studied sample crystallography, increasing the dimensionality
- Analysis of multimodal data is even more complicated because of dimensionality
- **This requires mathematical methods for dimensionality reduction, which would enable automated or semi-automated data processing and analysis.**

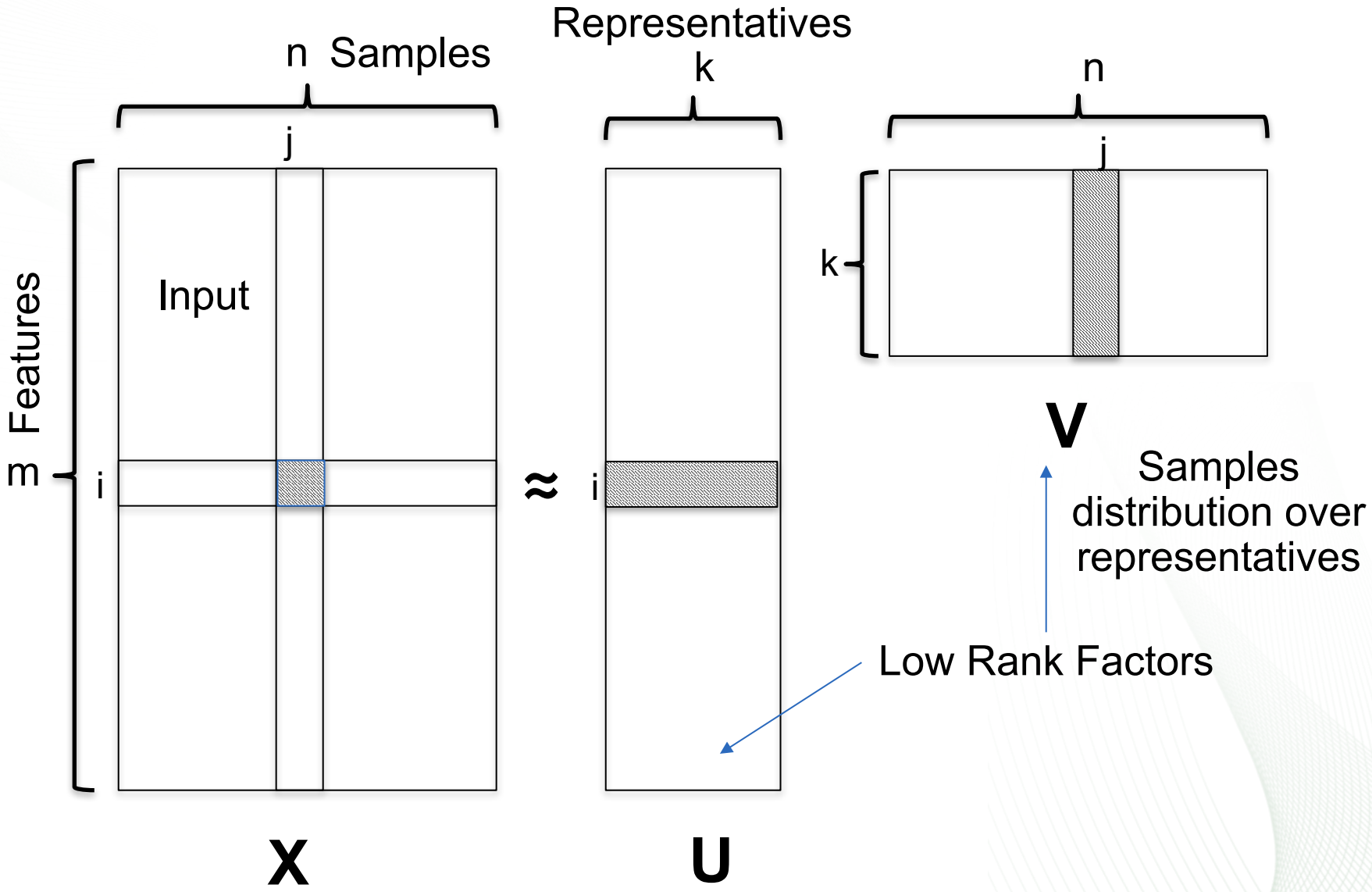
Tensor Factorizations at Scale for Scientific Data

- Multimodal characterization of materials – *comprehensive characterization from chemical composition to functional properties on the nanoscale*



Acknowledge:
Sergei Kalinin,
Anton Ievlev
ORNL/CNMS

Non-negative Matrix Factorization (NMF)



Non-negative matrix factorization

- A group of algorithms in multivariate analysis where a matrix X is factorized into two matrices U and V, with the property that all three matrices have no negative elements.
- NMF has an inherent clustering property, it automatically clusters the columns of input data X.
- Since NMF is not exactly solvable in general, it is commonly approximated numerically by minimizing the error function with the constraint that U and V be non-negative.
- NMF finds applications in such fields as computer vision, document clustering, chemometrics, chemical sensing, audio signal processing, and recommender systems.

https://en.wikipedia.org/wiki/Non-negative_matrix_factorization

MPI-FAUN: An MPI-Based Framework for Alternating-Updating Nonnegative Matrix Factorization

A new, high-performance parallel computational framework for a broad class of NMF algorithms that iteratively solves alternating non-negative least squares (NLS) subproblems for W and H.

- Distributed Communication avoiding NMF Algorithms
- https://github.com/ramkikannan/nmflibrary
- https://arxiv.org/abs/1609.09154, accepted at IEEE Trans. Knowledge and Data Eng.
- Miniapp benchmarked on OLCF Platforms

Rhea, 100 nodes, 1600 cores, Low Rank 50,

Dataset	Type	Matrix size	NMF Time
Video	Dense	1 Million x 13,824	5.73 seconds
Stack Exchange	Sparse	627,047 x 12 Million	67 seconds
Webbase-2001	Sparse	118 Million x 118 Million	25 minutes

Titan – Dense Matrix, Low Rank 50, 100 Iterations, 12,650 Nodes, 202500 Cores,

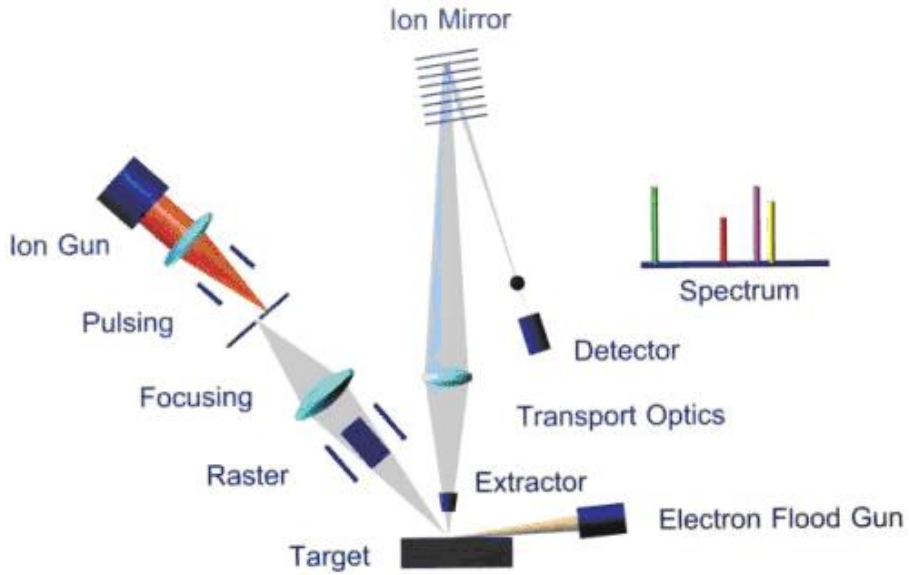
Matrix Size	Algos	NMF Time (in Secs)
2.7million x 2.7 million	MU	554
	HALS	197.75
	ANLS/BPP	219.8
3.03 million x 3.03 million	MU	554
	HALS	197.75
	ANLS/BPP	219.8

Acknowledge: Ramki Kannan, ORNL

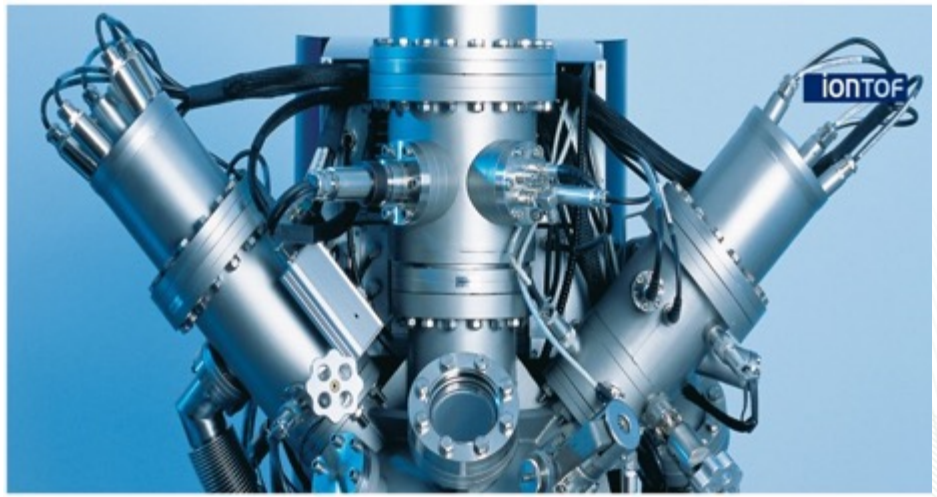
Case Study: NMF of ToF SIMS Data

- Time of Flight Secondary Ion Mass Spectrometry
  - Local investigations of the sample chemical composition
  - Ionization by Bi^+ ions
  - Sputtering by Cs^+ ions for investigations in the bulk
  - Time of flight of secondary ions is proportional to charge/mass ratio

ToF SIMS scheme



ION-TOF TOF. SIMS[5] (4100 C151)

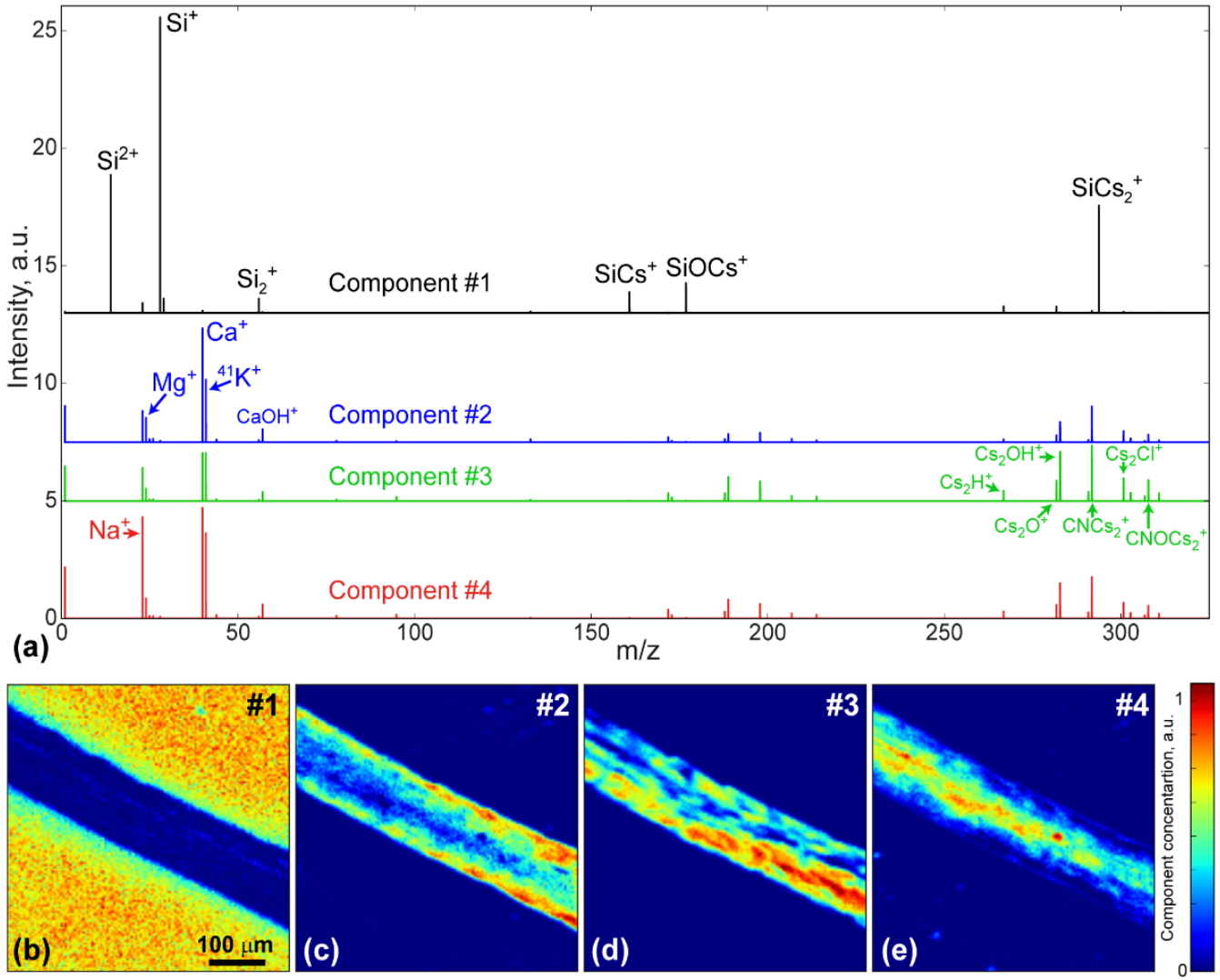


Acknowledge: Anton Ievlev ORNL/CNMS
https://www.ornl.gov/facility/cnms/output/chemical-imaging

Case Study: NMF of ToF SIMS Data

- TOF SIMS Data comes as a dense matrix
- Direct analysis of this matrix is difficult
- Task: Find a U matrix that represents the end members and a V matrix for the abundance map (mixture among these end members)
- For this study,
  - m = 128*128 (image size, “features”),
  - n = 1200 (signal length, “samples”),
  - k, the number of clusters, is between 2-6.

Case Study: NMF of ToF SIMS Data



Component 1 : Si substrate peaks
Component 2 : Inorganics, e.g., Mg, Ca, K cations
Component 3 : Cesium complexes
Component 4 : Higher Na cation concentrations
Component 5 : Appears to be noise

NMF (and PCA) approaches are currently insufficient to the task of multimodal data analysis

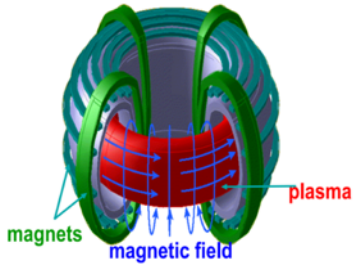
Acknowledge: Anton Ievlev ORNL; Ramki Kannan, ORNL, private communication

Case Study: Deep learning used to predict disruptions in a tokamak fusion reactor

PPPL PRINCETON PLASMA PHYSICS LABORATORY

Most critical problem for Fusion Energy: *avoid/mitigate large-scale major disruptions.*

Approach: Use of big-data-driven statistical/machine-learning (ML) predictions for the occurrence of disruptions in "Joint European Torus (JET)"



Princeton Team Goals include:

- improve physics fidelity via development of new *ML **multi-D**, time-dependent software including better classifiers;*
- develop ***"portable"*** (cross-machine) predictive software beyond JET to other devices and eventually ITER; and
- enhance execution speed of disruption analysis for very large datasets

*Development & deployment of advanced ML software via **Deep Learning Recurrent Neural Networks***

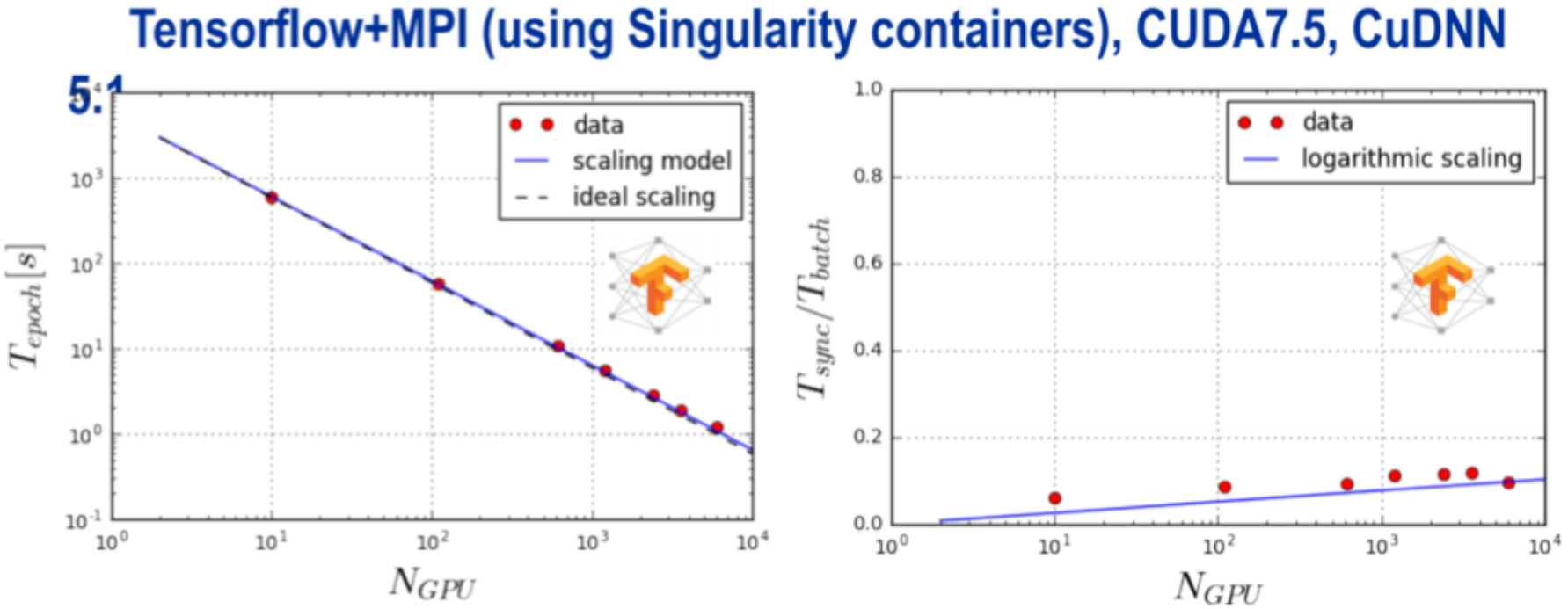
Bill Tang, "Accelerated Deep Learning Advances in HPC" Invited Talk #DC-7243, GTC-DC-2017

Alexey Svyatkovskiy, "Training Distributed Deep Recurrent Neural Networks with Mixed Precision on GPU Clusters", Machine Learning in HPC Environments (this workshop), SC17..

Case Study: Deep learning used to predict disruptions in a tokamak fusion reactor

PPPL PRINCETON PLASMA PHYSICS LABORATORY

- Deep Learning executing on ~6000 GPUs with TensorFlow+MPI.



Ack; Mike Matheson, ORNL, Titan implementation and scaling

Bill Tang, “Accelerated Deep Learning Advances in HPC” Invited Talk #DC-7243, GTC-DC-2017

Alexey Svyatkovskiy, “Training Distributed Deep Recurrent Neural Networks with Mixed Precision on GPU Clusters”, Machine Learning in HPC Environments (this workshop), SC17..

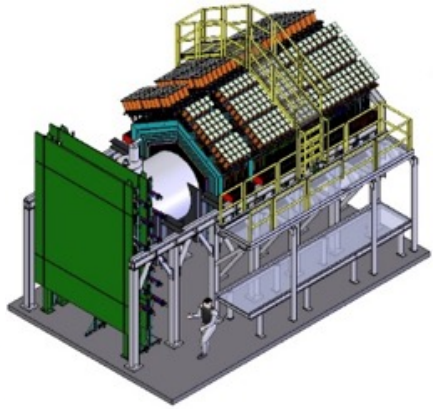
Pilot Study: Collision vertex reconstruction in HEP

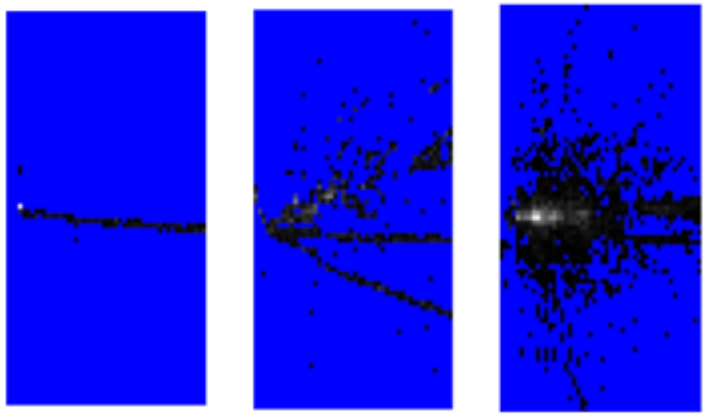
Fermilab

FNAL (MINERvA) and ORNL Computational Data Analytics Group (CDA) collaborated to improve their ML networks for vertex reconstruction.

A. Terwilliger, G. Perdue, D. Isele, R. M. Patton, and S.R. Young. “Vertex Reconstruction of Neutrino Interactions using Deep Learning.” In *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN),* pp. 2275-2281. IEEE Press, 2017. doi:10.1109/IJCNN.2017.7966131.

Neutrino Detection for MINERvA





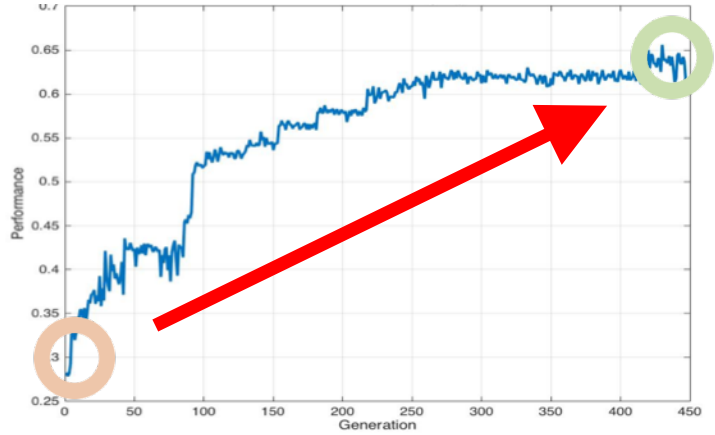
Analytics of Deep Learning Hyper-parameter search running at scale on Titan.

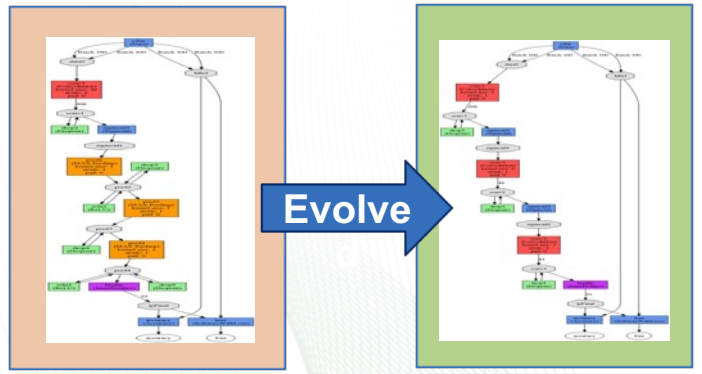
PI: R. M. Patton (ORNL), G. Perdue (FNAL)
Sponsor: ORNL LDRD, FNAL

Multi-node Evolutionary Neural Networks for Deep Learning (MENNDL)

Premise: For every data set, there exists a corresponding neural network that performs optimally with that data

- ORNL Data Analytics Group used Titan to develop an evolutionary algorithm to search for optimal hyper-parameters and topologies for ML networks.
- **Demonstrated on 18,000 nodes of Titan using high energy physics data through collaboration with Fermilab**
- Evaluated against multiple datasets
  - Standard computer vision datasets
  - Neutrino detector vertex finding dataset
  - SNS Small angle scattering model fitting dataset
- Currently exploring additional datasets and evaluating performance on Summit-Dev



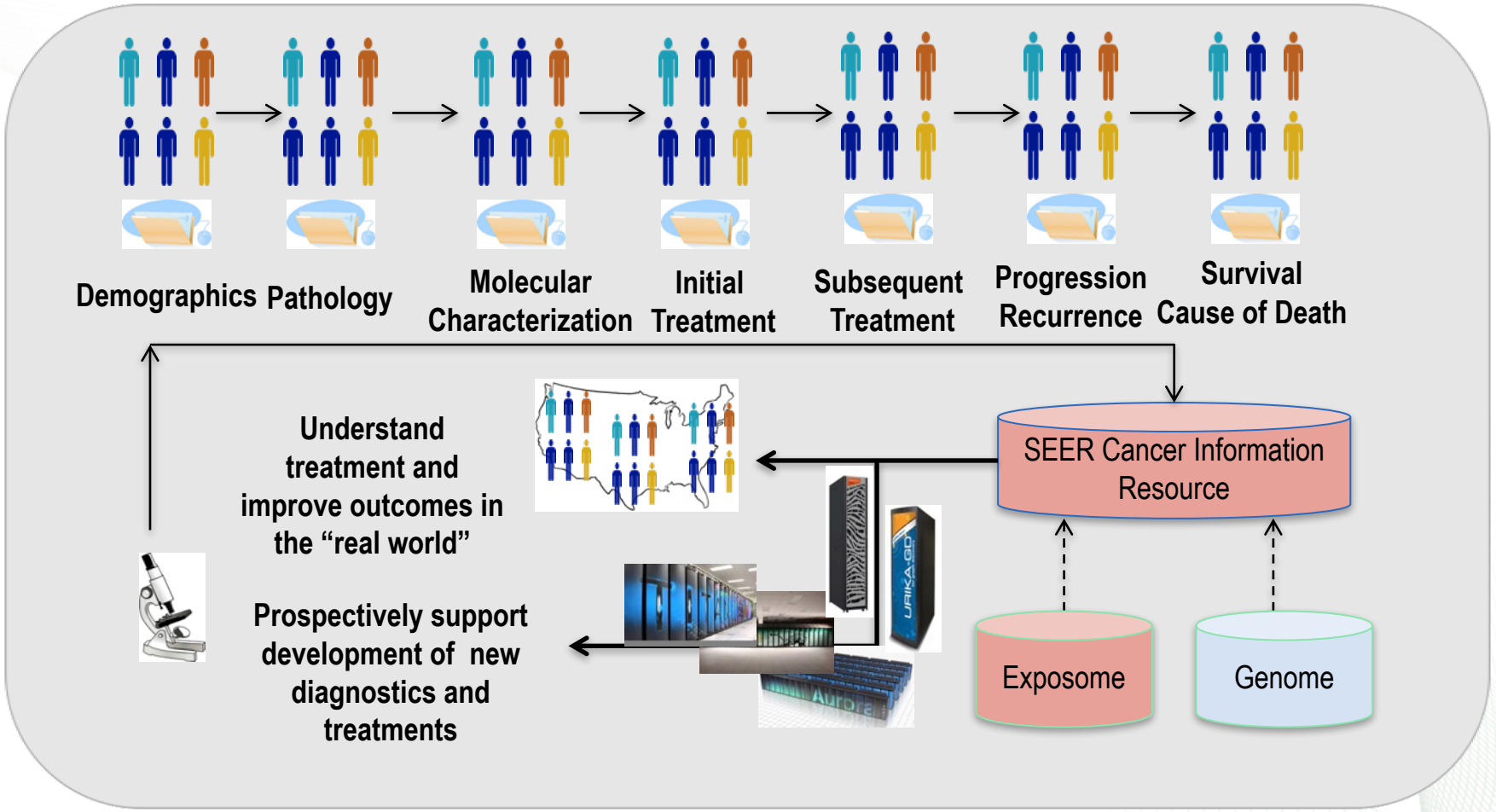


The “big data” revolution in health care is well underway

- **Advances in machine learning coupled with the explosion of health data is showing promise for**
  - accelerated biomedical research discovery
  - clinical decision support
  - guiding personalized treatments and disease management
  - helping uncover better preventive practices
  - improving workflow and streamlining communication and coordination
  - offering new ways to handle waste, fraud, and abuse
  - …

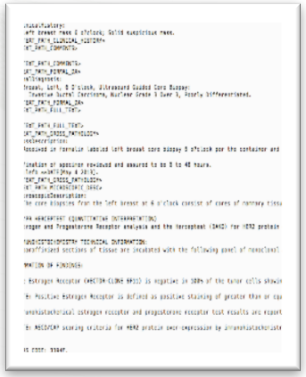
Ack: Georgia Tourassi, (ORNL), Arjun Shankar (ORNL)

AI to support national cancer surveillance



Ack: Georgia Tourassi, (ORNL)

A scalable framework for AI-assisted information extraction from pathology reports (text understanding)



Certified Tumor Registrar

CTR at a cancer registry reviews complete patient medical record + path report

Regional cancer registries collect case information and aggregate for NCI SEER database

NLP practice

R - research work:
set a goal →
devise an algorithm →
train the algorithm →
test its accuracy

D - development work:
implement the algorithm as an **API** with sufficient **performance** and **scaling** characteristics

Challenge: Scale deep learning – based natural language processing across *cancer patients*, *registrars*, and *information abstraction tasks*

Ack: Georgia Tourassi, (ORNL)

Experimental Pipeline

DATA PRE-PROCESSING
- Duplicate records
- Non-contradicting labels
- Incorrect organ annotations
- Small sample sizes
- Corpus curation

FEATURE REPRESENTATION
- TF-IDF
- Bag-of-graphs
- RAKE
- CHUNK
- GLOVE

TRAINING PROTOCOLS
With over sampling
Without over sampling

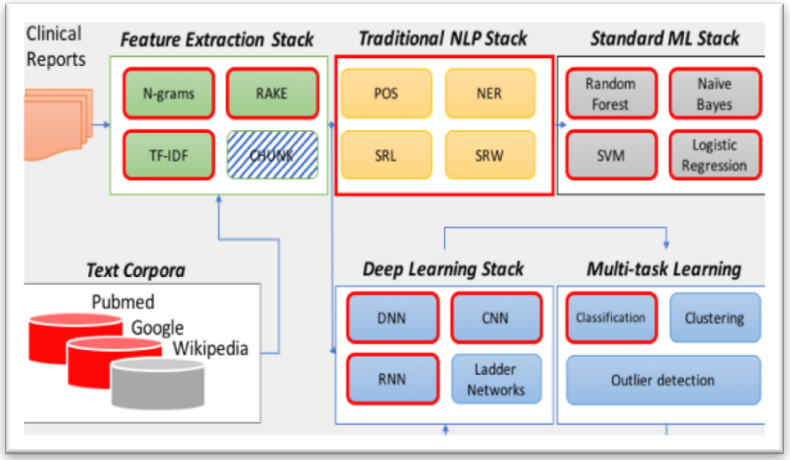
RULE-BASED SYSTEMS (RL)
- Contextualize (keywords for topics of interest)
- Term identification
- Classification

MACHINE LEARNING (ML)
- Naïve Bayes (NB)
- Logistic Regression (LR)
- Random Forest (RF)
- Support Vector Machines (SVM)
- Extreme gradient boosting tree (Xgboost)

DEEP LEARNING (DL)
- Convolutional neural nets (CNN)
- Hierarchical Attention nets (HAN)
- Multi-task Deep neural net (MT-DNN)

PERFORMANCE METRICS
- Precision (positive predictive value) / Recall (sensitivity) / F1 per class
- Macro / Micro scores (aggregate performance over all)

VALIDATION STRATEGIES
- K-fold cross validation (K-fold)
- Leave-one-registry out (LORO)
- Leave-one-case-out per registry (LOO_R)



Best performing algorithm is a Hierarchical Attention Neural Network

Primary Cancer Subsite: micro-F1 score=0.81±0.03

Histological Grade: micro-F1=0.90±0.02

H.-J. Yoon, A. Ramanathan, G.D. Tourassi, Multi-Task deep neural networks for automated extraction of primary site and laterality information from cancer pathology reports. *2016 INNS Conference on Big Data*, October 23-25, Greece

J.X. Qiu, H.-Y. Yoon, P.A. Fearn, G.D. Tourassi, "Deep Learning for Automated Extraction of Primary Sites from Cancer Pathology Reports," to appear in *IEEE Journal of Biomedical and Health Informatics* (2017).

S. Gao, M.T. Young, J.X. Qiu, J.B. Christian, P.A. Fearn, G.D. Tourassi, A. Ramanathan, "Hierarchical Attention Networks for Information Extraction from Cancer Pathology Reports," submitted to the *Journal of American Medical Informatics Association* (05/2017).

https://github.com/ECP-CANDLE/Benchmarks/tree/master/Pilot3/P3B1 (Multi-task deep neural network)

https://github.com/ECP-CANDLE/Benchmarks/tree/master/Pilot3/P3B2 (Generative models using recurrent neural networks)

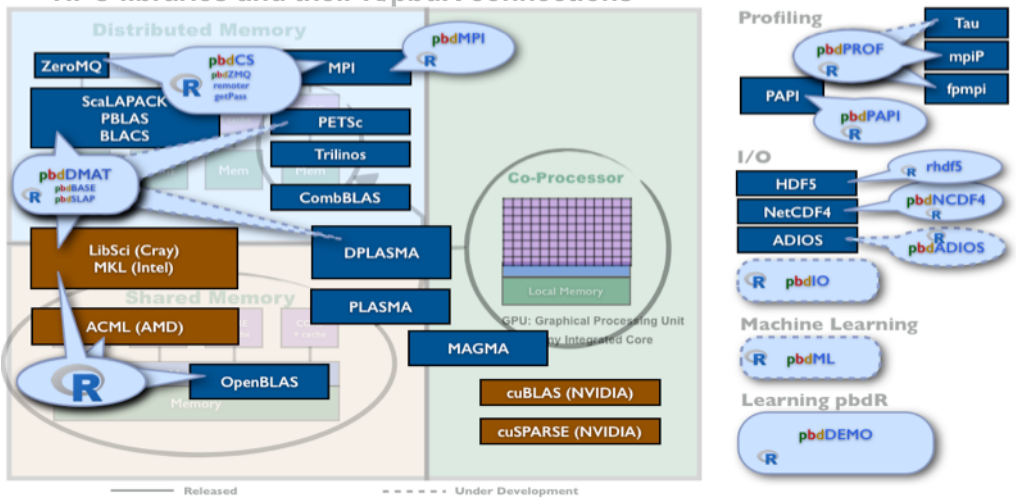
Scalable R Development Platform for Big Data Analytics

pbdR
Programming with Big Data in R

http://pbdr.org

- Engage parallel math libraries at scale
- R language unchanged
- New distributed concepts
- New profiling capabilities
- New interactive SPMD parallel
- In-situ distributed capability
- In-situ staging capability via ADIOS

HPC libraries and their R/pbdR connections



HPCwire July 6, 2016

"OLCF Researchers Scale R to Tackle Big Science Data Sets"

"***for situations where one needs interactive near-real-time analysis, the pbdR approach is much better*** *[than Apache Spark–like frameworks]."*

PCA of a 134 GB matrix: "*hours on . . . Apache Spark, . . . less than a minute using R.*"

"ORNL Researchers Bridge the Gap Between R, HPC Communities"

"…"untapped [R] domains" represent an enormous potential user base for world-class computers ."

Schmidt, Chen, Matheson, and Ostrouchov (2016). Programming with BIG Data in R: Scaling Analytics from One to Thousands of Nodes, ***Big Data Research,*** *in print online*.

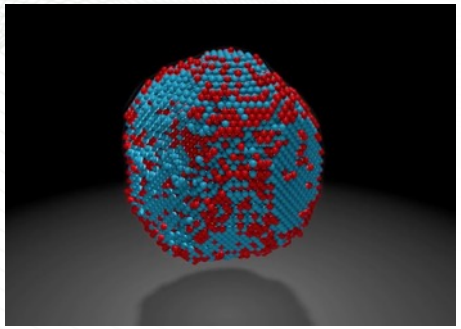
Schmidt, Chen, and Ostrouchov (2016). Introducing a New Client/Server Framework for Big Data Analytics with the R Language. ***XSEDE16 Conference on Diversity, Big Data, and Science at Scale***, Article No. 38.

Emerging Science Activities:
Selected Machine Learning Projects on Titan: 2016-2017

Program	PI	PI Employer	Project Name	Allocation (Titan core-hrs)
ALCC	Robert Patton	ORNL	Discovering Optimal Deep Learning and Neuromorphic Network Structures using Evolutionary Approaches on High Performance Computers	75,000,000
ALCC	Gabriel Perdue	FNAL	Large scale deep neural network optimization for neutrino physics	58,000,000
ALCC	Gregory Laskowski	GE	High-Fidelity Simulations of Gas Turbine Stages for Model Development using Machine Learning	30,000,000
ALCC	Efthimions Kaxiras	Harvard U.	High-Throughput Screening and Machine Learning for Predicting Catalyst Structure and Designing Effective Catalysts	17,500,000
ALCC	Georgia Tourassi	ORNL	CANDLE Treatment Strategy Challenge for Deep Learning Enabled Cancer Surveillance	10,000,000
DD	Abhinav Vishnu	PNNL	Machine Learning on Extreme Scale GPU systems	3,500,000
DD	J. Travis Johnston	ORNL	Surrogate Based Modeling for Deep Learning Hyper-parameter Optimization	3,500,000
DD	Robert Patton	ORNL	Scalable Deep Learning Systems for Exascale Data Analysis	6,500,000
DD	William M. Tang	PPPL	Big Data Machine Learning for Fusion Energy Applications	3,000,000
DD	Catherine Schuman	ORNL	Scalable Neuromorphic Simulators: High and Low Level	5,000,000
DD	Boram Yoon	LANL	Artificial Intelligence for Collider Physics	2,000,000
DD	Jean-Roch Vlimant	Caltech	HEP DeepLearning	2,000,000
DD	Arvind Ramanathan	ORNL	ECP Cancer Distributed Learning Environment	1,500,000
DD	John Cavazos	U. Delaware	Large-Scale Distributed and Deep Learning of Structured Graph Data for Real-Time Program Analysis	1,000,000
DD	Abhinav Vishnu	PNNL	Machine Learning on Extreme Scale GPU systems	1,000,000
DD	Gabriel Perdue	FNAL	MACHINE Learning for MINERvA	1,000,000
		TOTAL		**220,500,000**

OLCF Strategic Scientific Accomplishments: 2016

Nanoscience



Markus Eisenbach
ORNL

Eisenbach and team modeled the properties of strongly magnetic regions of an FePt nanoparticle. The researchers used the LSMS code on Titan to further determine the magnetic anisotropy of more than 1,300 atoms from regions of the nanoparticle.

Y. Yang, et al. 2017. Nature. **542**.

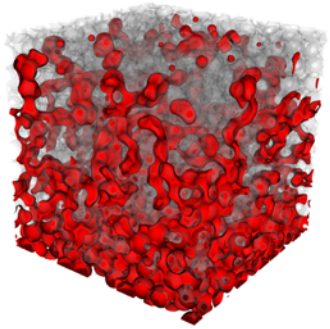
Engineering

Peter Vincent
Imperial College

Vincent's team is tackling unsteady airflow patterns in jet engines and providing engineers with an unprecedented tool to solve long-standing design problems.

P. Vincent, et al. 2016 Proc. of the Int'l. Conf. for HPC, Net., Storage and Analysis.

Geosciences

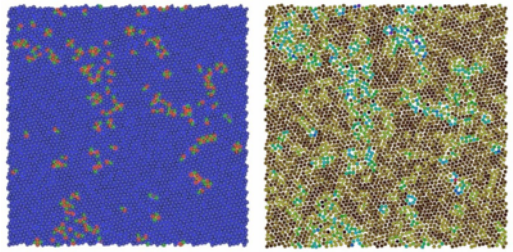


James McClure
Virginia Tech

McClure's team created a computational framework to study complex subsurface interactions, incorporating micro-CT imaging data to directly visualize the movement of fluids in underground reservoir rocks and other geologic materials.

R. T. Armstrong, et al. 2016. Phys. Rev. E. **94**.

Materials Science

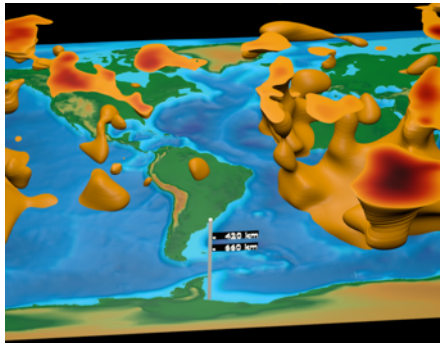


Sharon Glotzer
University of Michigan

Glotzer's team ran a series of hard particle simulations to study melting in 2-D systems, exploring how particle shape affects the physics of a 2-D solid-to-fluid melting transition.

J. A. Anderson, et al. 2016. Computer Physics Comm. **204.**

Geosciences



Jeoren Tromp
Princeton University

Tromp and his team modeled Earth's interior using Titan. This 3-D map shows shear wavespeed perturbations calculated using data from 253 earthquakes.

E. Bozdağ, et al. 2016. Geophysical J. Int'l. **207**.

All from 2016 work conducted on Titan through OLCF's INCITE program.

Future Policy and Technical Challenges/Opportunities: Convergence of HPC, Data Analytics, and AI Workflows

- Resource Management Systems (job schedulers) & Queue Policies
  - Too many jobs?, jobs too long? (relative to HPC jobs mix)
  - Interactive and/or real-time access & fine grain control
  - Data reuse on compute partition (e.g., NVRAM), spanning jobs, users, projects?
- Resource Specialization & Diversity (multi-level heterogeneity)
  - Subset of nodes with special operating conditions/requirements
- Resource Interoperability (middleware services)
  - Containers: interactions with diverse HPC system software
- HPC Center Support for Rapidly Evolving ML Frameworks?
- Resource & Data Access (authentication, authorization and data access)
  - Science applications are already spanning multiple data centers with varying security policies.
  - Data silos within and across organizations

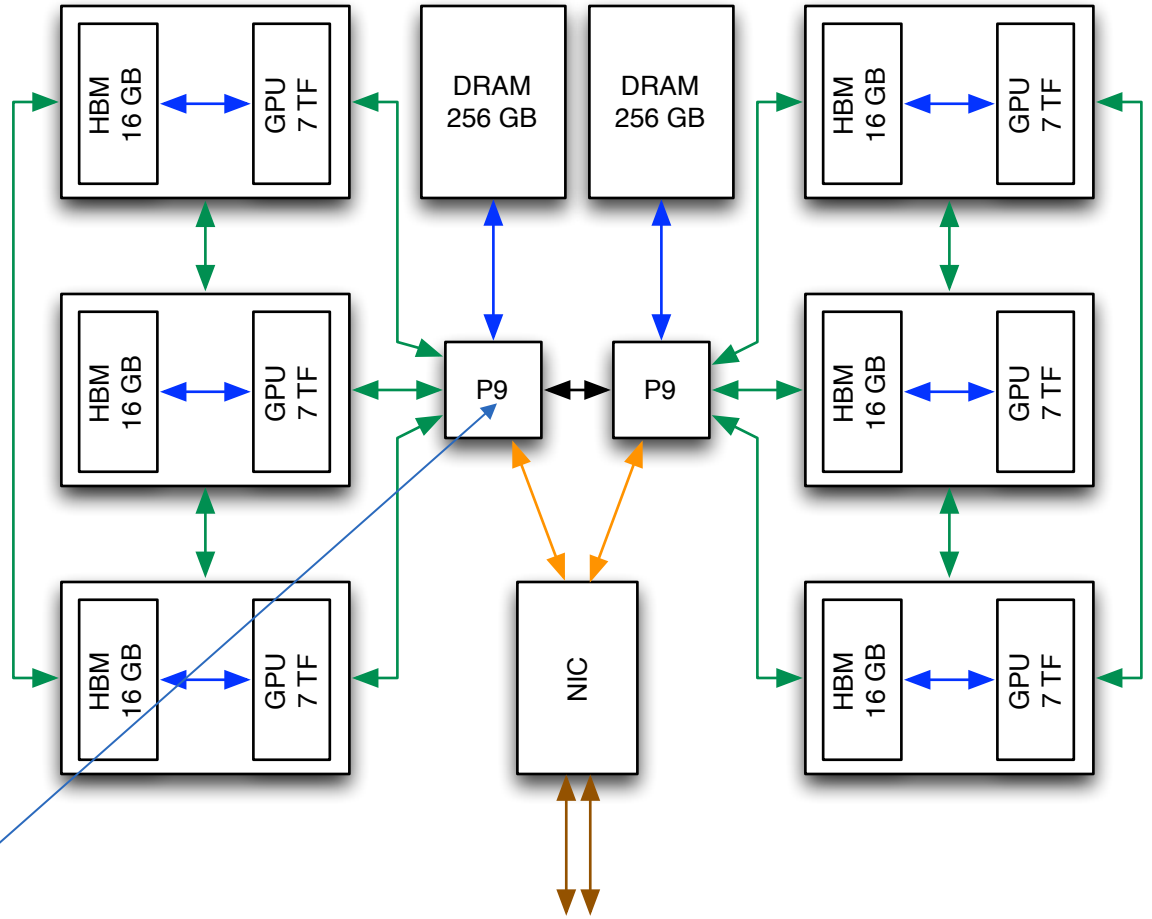
Ack: Sadaf Alam, (CSCS) for inspiration for organization of these topics.

Summit will replace Titan as the OLCF's leadership supercomputer

- Many fewer nodes
- Much more powerful nodes
- Much more memory per node and total system memory
- Faster interconnect
- Much higher bandwidth between CPUs and GPUs
- Much larger and faster file system

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 + 96 GB HBM2
NV memory per Node	0	1600 GB
Total System Memory	710 TB	>10 PB DDR4 + HBM2 + Non-volatile
System Interconnect (node injection bandwidth)	Gemini (6.4 GB/s)	Dual Rail EDR-IB (25 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

Summit Node Overview



1.6 TB NVMe SSD (800 GB per job)
6 GB/s Read + 2.1 GB/s Write

TF	42 TF (6x7 TF)
HBMH	96 GB (6x16 GB)
DRAM	512 GB (2x16x16 GB)
NET	23 GB/s (2x11.4 GB/s)
MMsg/s	83

- HBM/DRAM Bus
- NVLINK
- X-Bus (SMP)
- PCIe Gen4
- EDR IB

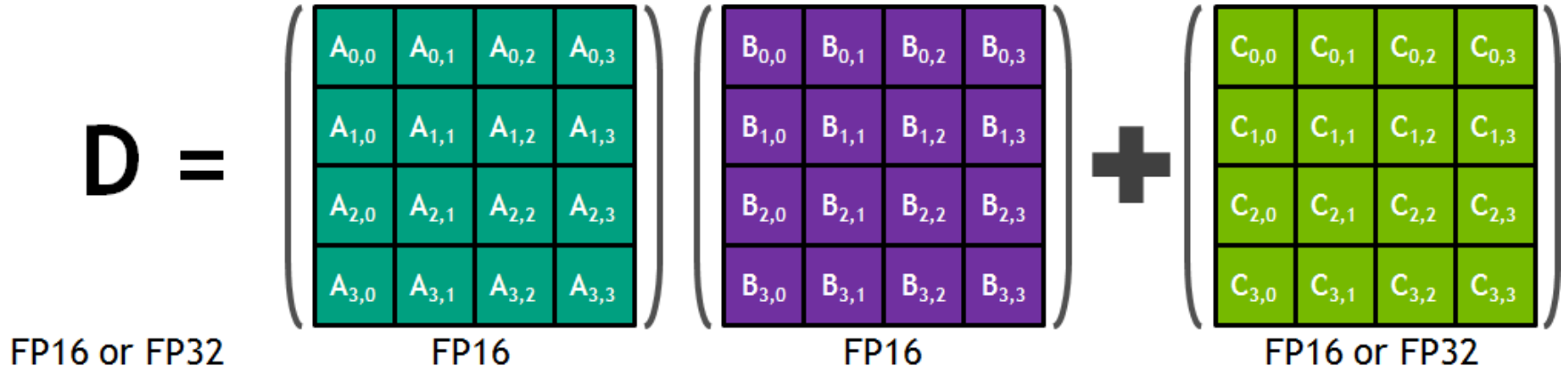
Volta Details

	Tesla V100 for NVLink	Tesla V100 for PCIe
PERFORMANCE with NVIDIA GPU Boost™	DOUBLE-PRECISION 7.8 TeraFLOPS	DOUBLE-PRECISION 7 TeraFLOPS
	SINGLE-PRECISION 15.7 TeraFLOPS	SINGLE-PRECISION 14 TeraFLOPS
	DEEP LEARNING 125 TeraFLOPS	DEEP LEARNING 112 TeraFLOPS
INTERCONNECT BANDWIDTH Bi-Directional	NVLINK 300 GB/s	PCIE 32 GB/s
MEMORY CoWoS Stacked HBM2	CAPACITY 16 GB HBM2	
	BANDWIDTH 900 GB/s	

TensorCores™
Mixed Precision
(16b MutliplyAdd and
32b Accumulate)

Note: The performance numbers are peak and not representative of Summit's Volta

Tesla V100 Tensor Cores



- Tesla V100's Tensor Cores are programmable matrix-multiply-and-accumulate units, delivering up to 125 Tensor TFLOPS for training and inference applications.
  - Each Tensor Core provides a 4x4x4 matrix processing array which performs the operation **D** = **A** * **B** + **C**, where **A, B**, **C** and **D** are 4×4 matrices.
  - The Tesla V100 GPU contains 640 Tensor Cores: 8 per SM.
  - The matrix multiply inputs **A** and **B** are FP16 matrices, while matrices **C** and **D** may be FP16 or FP32 matrices.
- Each Tensor Core performs 64 floating point FMA mixed-precision operations per clock and 8 Tensor Cores in an SM perform a total of 1024 floating point operations per clock.
- This is a 8X increase in throughput for deep learning applications per SM compared to Pascal GP100 using standard FP32 operations, resulting in a total 12X increase in throughput for the Volta V100 GPU compared to the Pascal P100 GPU.

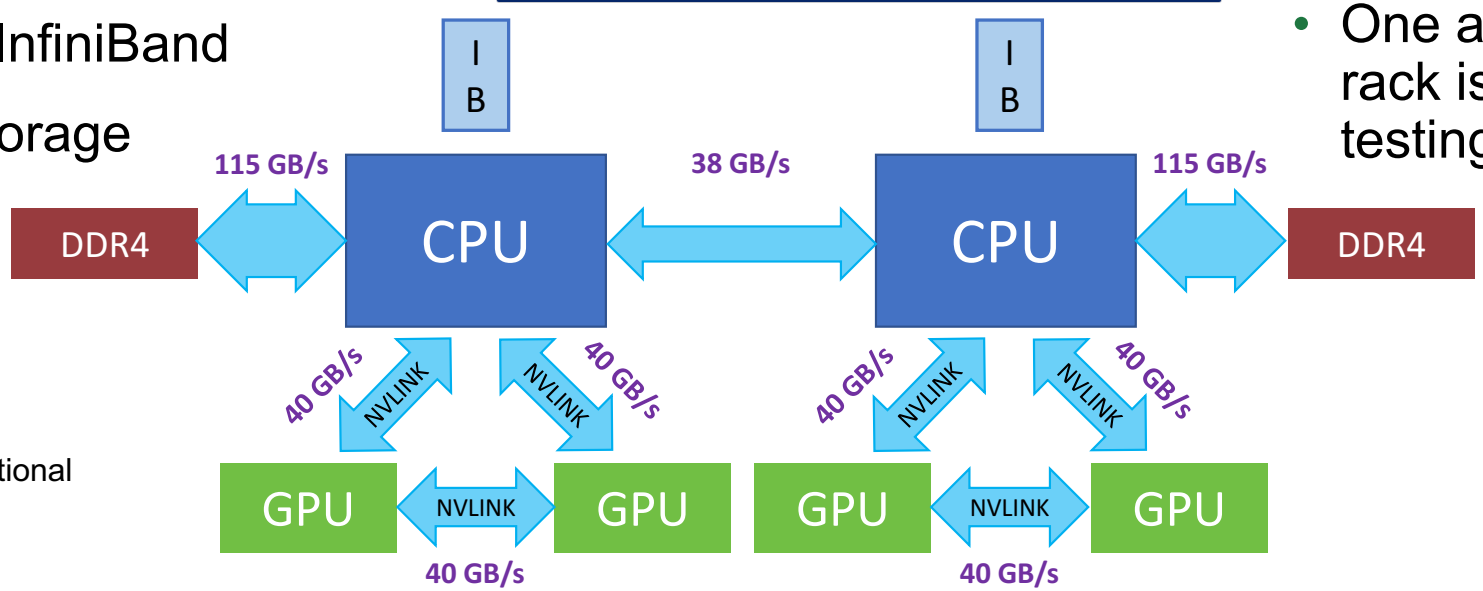
https://devblogs.nvidia.com/parallelforall/programming-tensor-cores-cuda-9/

Summit Early Evaluation System

Each IBM S822LC node has:

- 2x IBM POWER8+ CPUs
  - 32x 8GB DDR4 memory (256 GB)
  - 10 cores per POWER8+, each core with 8 HW threads
- 4x NVIDIA Tesla P100 GPUs
  - NVLink 1.0 connects GPUs at 80 GB/s
  - 16 GB HBM2 memory per GPU
- 2x Mellanox EDR InfiniBand
- 1600 GB NVMe storage

Mellanox EDR InfiniBand Fabric



DDR4 BW is Read+Write
X-Bus and NVLINK are bi-directional

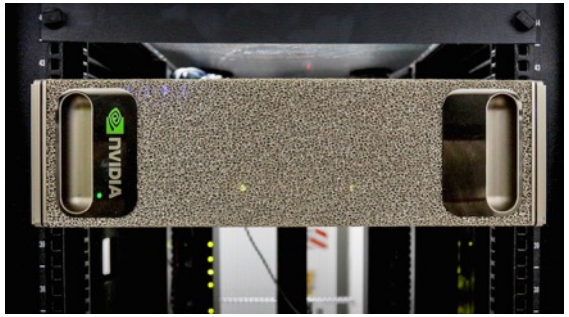
Summit EA System:

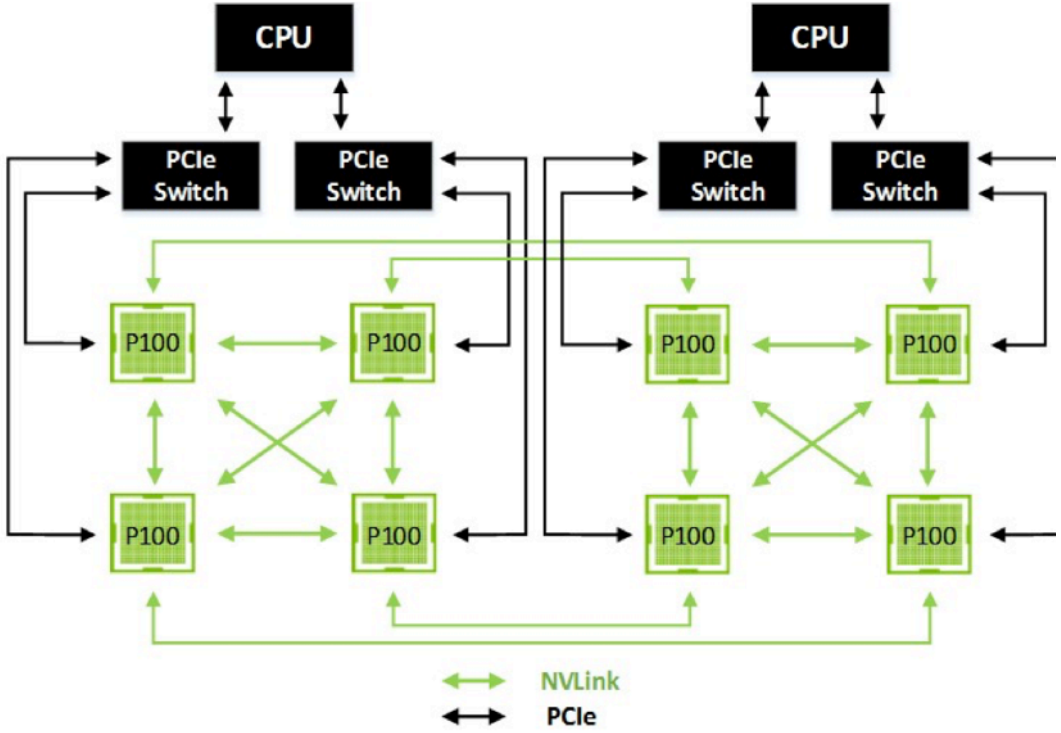
- 3 racks for development, each with 18 nodes
- One rack of login and support servers
- Nodes connected in a full fat-tree via EDR InfiniBand
- Liquid cooled w/ heat exchanger rack
- One additional 18-node rack is for system software testing

Preparation for Summit

- **DGX-1 architecture and use of NVLink similar to Summit-dev**
  - Ease the preparation for DNN training at larger scales

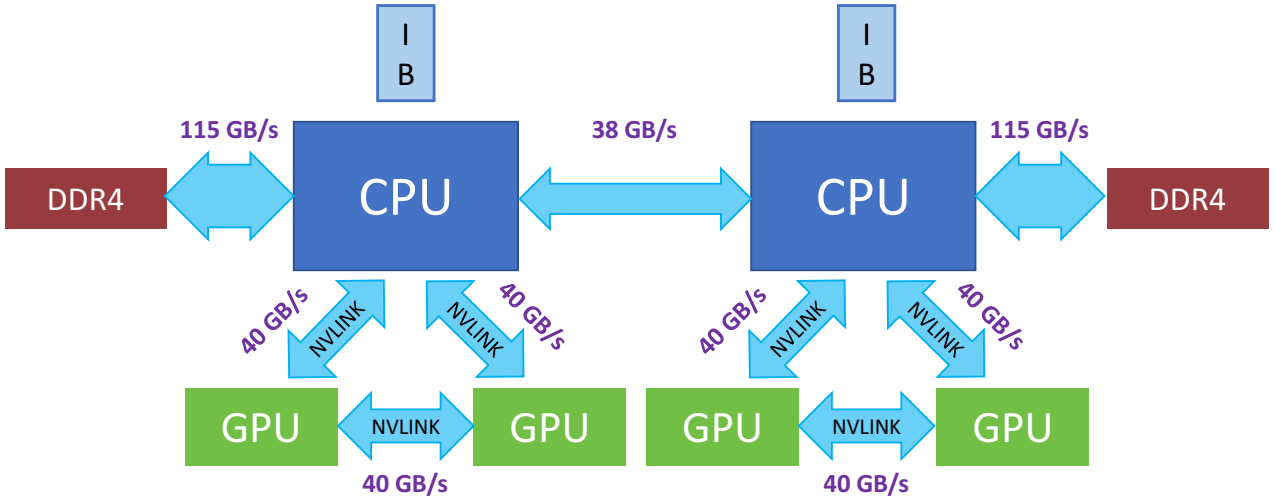
DGX-1





Summit-dev

(IBM's S822LC)



Thank You!
Acknowledgements

- Arjun Shankar (ORNL) and Robert Patton (ORNL)
- ASCR Facilities Division Requirements Workshop Team
  - Richard Coffey (ANL), Katherine Riley (ANL), Andre Manning (ANL), Richard Gerber (LBNL), Deborah Bard (LBNL), Eli Dart (LBNL), Katie Antypas (LBNL), Tjerk Straatsma (ORNL), Jim Hack (ORNL)
- Case Studies
  - Anton Levlev (ORNL), Ramki Kannan (ORNL), Bill Tang (Princeton U.), Gabriel Perdue (FNAL), Steven Young (ORNL),
- Vendor Partners
  - Cray, NVIDIA, IBM,

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. Funding for this work comes from DOE and ORNL's LDRD program. 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.*