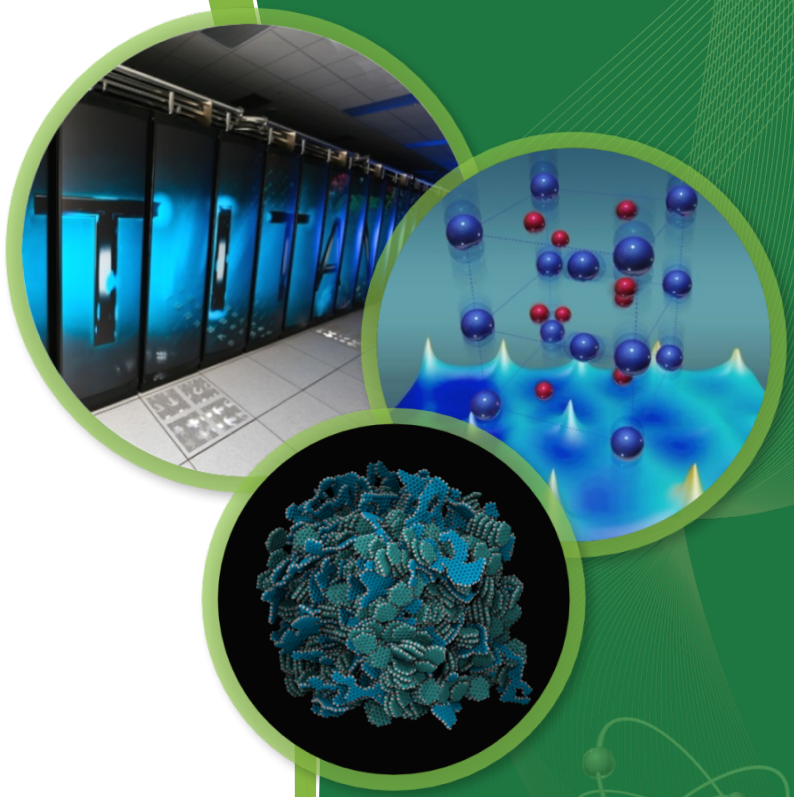
# Operating and Runtime Systems Challenges for HPC Systems




Barney Maccabe, Director
Computer Science and Mathematics Division

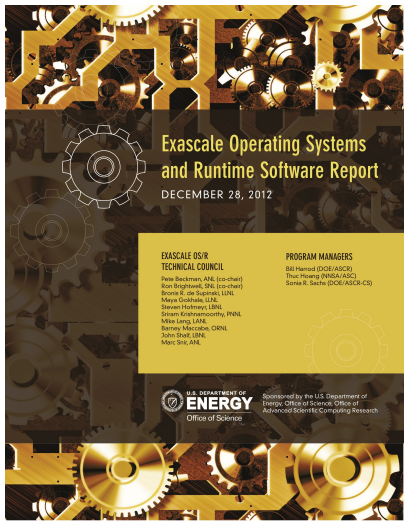
June 27, 2017 ROSS 2017 Washington, DC

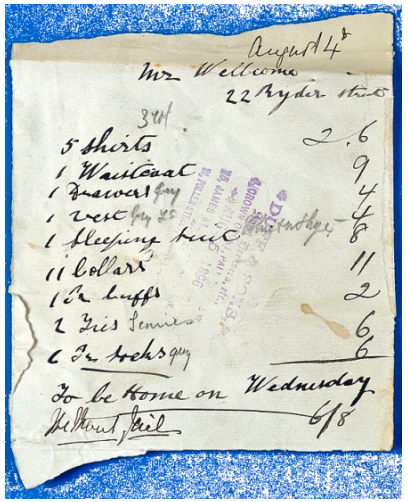
ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory

# Challenges Identified by OS/R Tech Council (Dec 2012)

- Technical Challenges
  - Resilience
    - Fault detection, notification, and management
    - OS survivability
  - Power
    - Hierarchical Management, Dynamics and Global optimization
  - Memory Hierarchy
    - NVRAM
    - Overhead Mgmt
    - Software-managed Memories
  - Parallelism
    - Scalable Synchronization, Scheduling, and Mgmt
    - Global Consistency, Coordination, and Control
  - Additional Hardware
    - Heterogeneity
    - Locality and Affinity Management
    - Hardware Event Handling

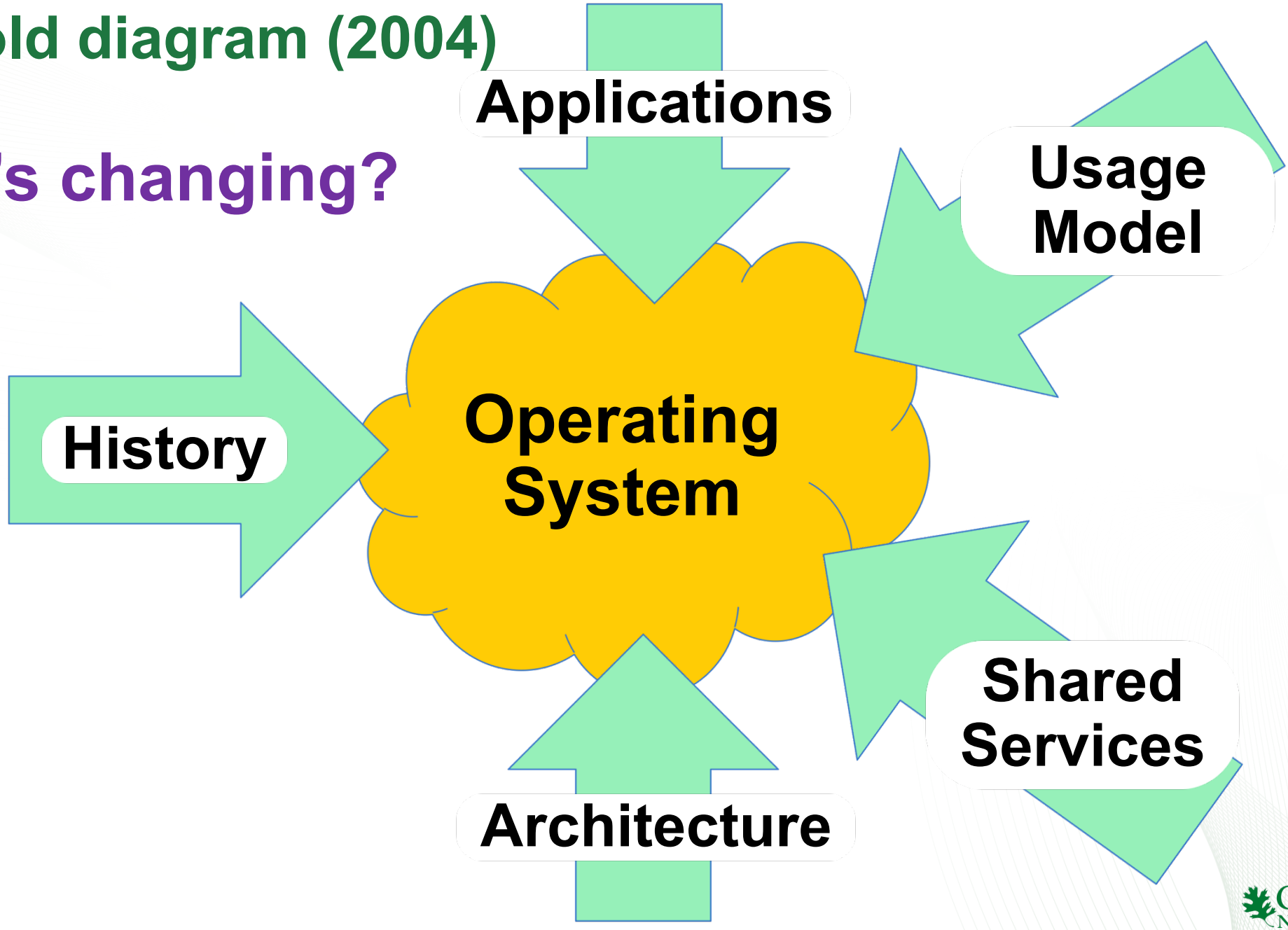


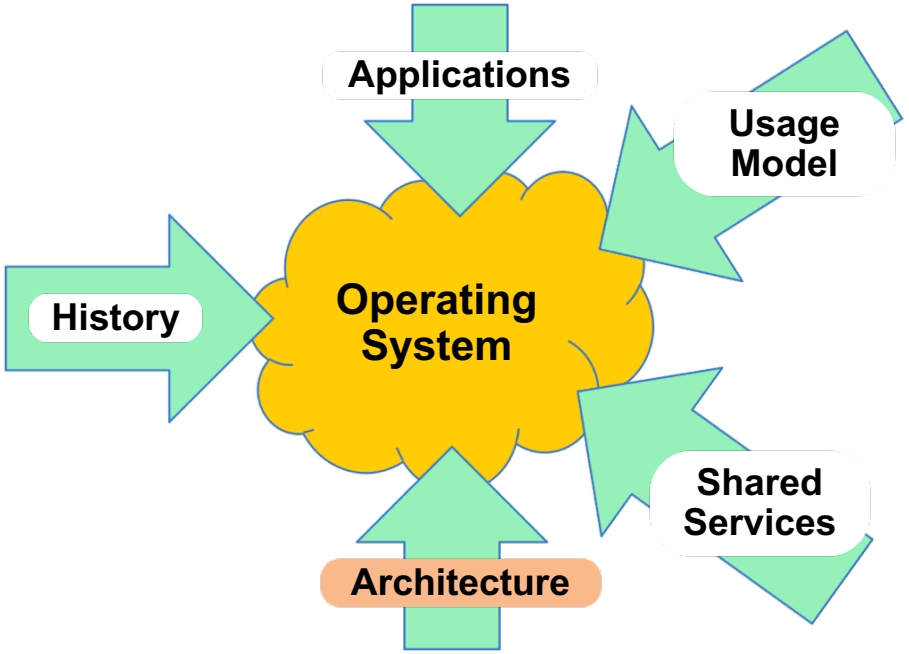

- Technical Challenges (continued)
  - OS/R Structural challenges
    - Misalignment of Requirements
    - User-space resource management
    - Parallel OS services
  - Legacy OS/R Issues
    - Changes in design assumptions
    - Internode parallelism
    - Enclaves
  - Application Structure
  - Dynamic Environments
- Business and Social Challenges
  - Existing code base
  - Lack of Vendor Transparency
  - Sustainability and Portability
  - OS Research not focused on HPC
  - Scaling down the Software Stack

# A very old diagram (2004)

## What's changing?



# Architecture: Compute, Memory, and Communication



# Hybrid Multicore Consortium (SC BOF 2010)

## Applications and Libraries

Define migration processes and libraries

*Application Communities*

## Programming Models

Programmer productivity and

Application performance portability

Co-Design

## Architecture and Metrics

Track and influence industrial development

## Performance and Analysis

Predictable application performance

Design feedback

OAK RIDGE National Laboratory

BERKELEY LAB

Los Alamos NATIONAL LABORATORY

# 2017 OLCF Leadership System
## Hybrid CPU/GPU architecture (like Titan)

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

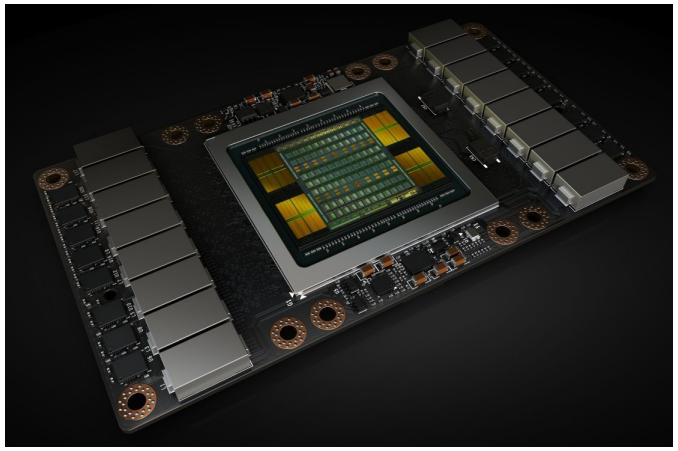
**At least 5X Titan's Application Performance**

**Approximately 3,400 nodes, each with:**

- Multiple IBM POWER9™ CPUs and multiple NVIDIA Volta® GPUs.
- CPUs and GPUs completely connected with high speed **NVLink**
- Large coherent memory: over 512 GB (**HBM** + DDR4)
  - all directly addressable from the CPUs and GPUs
- An additional 800 GB of NVRAM, which can be configured as either a **burst buffer** or as extended memory or both
- over 40 TF peak performance

**Dual-rail Mellanox® EDR-IB full, non-blocking fat-tree interconnect**

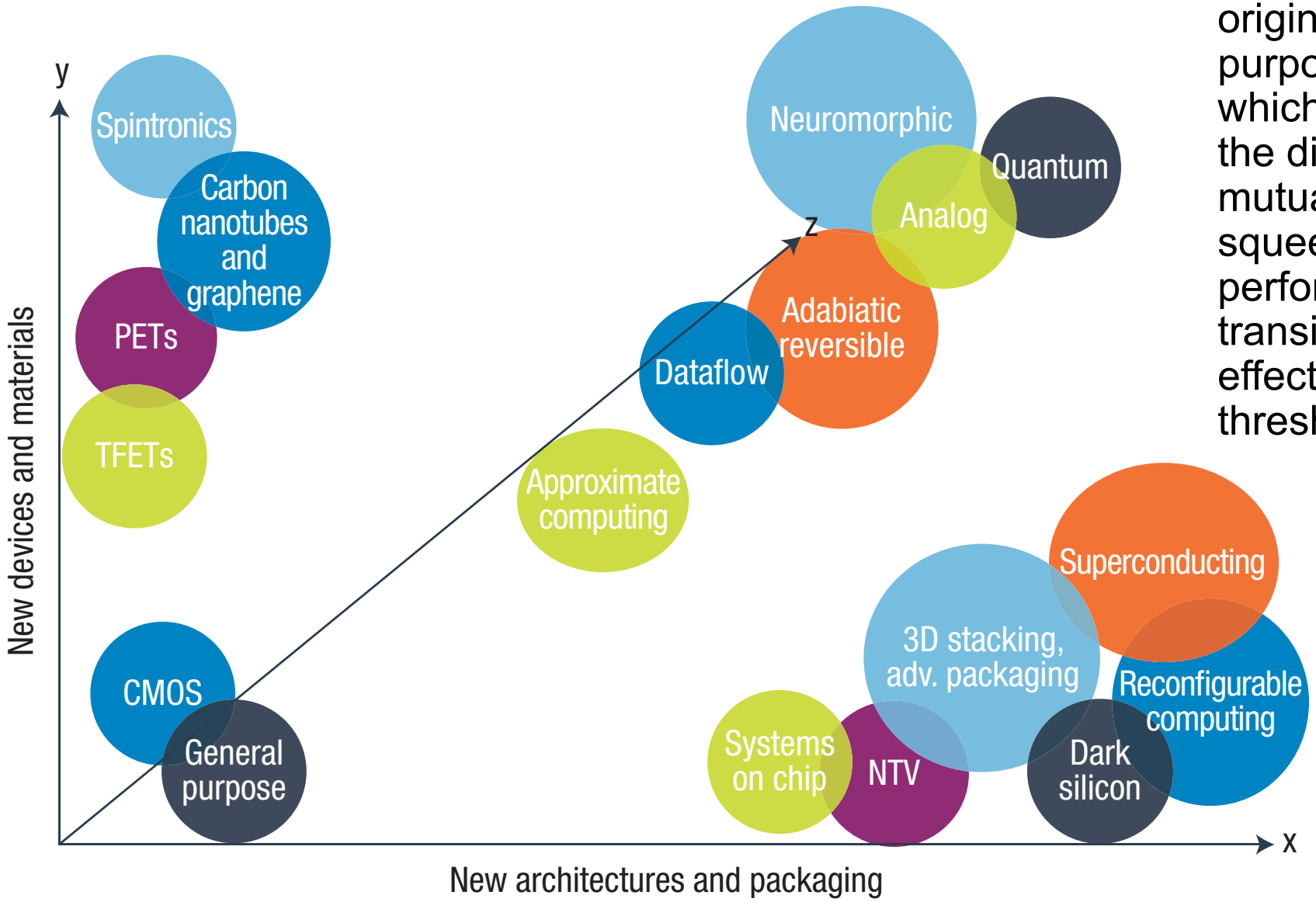
**IBM Elastic Storage (GPFS™) - 1TB/s I/O and 120 PB disk capacity.**



The Tesla V100 also includes 640 "tensor cores" designed to greatly accelerate machine learning, and a second-gen version of Nvidia's NVLink technology, which the company says surpasses PCIe transfer speeds by a whopping tenfold. None of that matters to PC gamers though.

PCWorld | MAY 11, 2017 3:00 AM PT

# Changing Compute Landscape




Technology scaling options along three dimensions. The graph's origin represents current general-purpose CMOS technology, from which scaling must continue. All the dimensions, which are not mutually exclusive, aim to squeeze out more computing performance. PETs: piezo-electric transistors, TFETs: tunneling eld-effect transistors; NTV: near-threshold voltage.

Credit: Shalf and Leland, IEEE Computer 2015

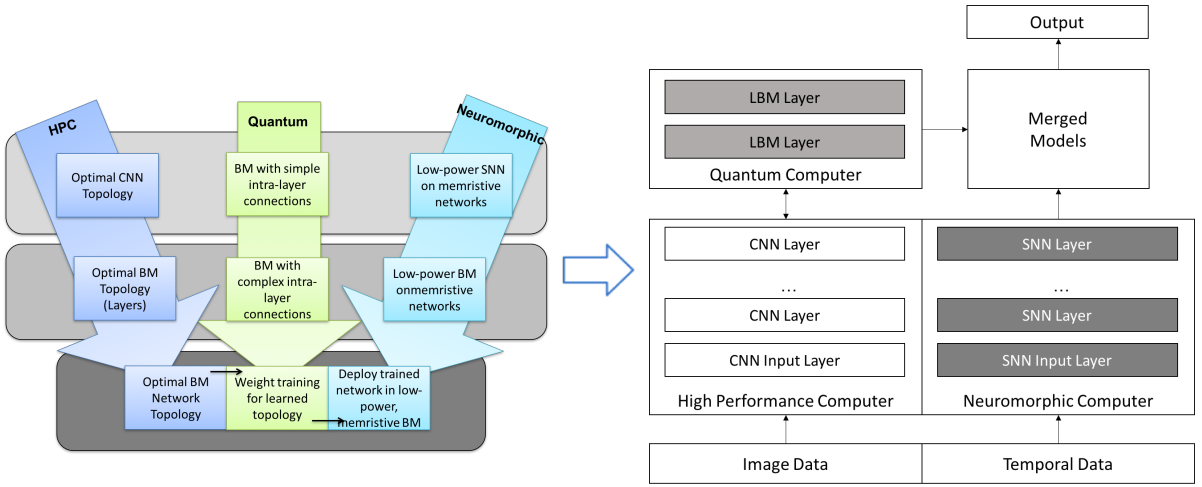
# Deep Learning on HPC-Quantum-Neuromorphic

## Objectives

- Evaluate deep learning networks on:
  - ORNL Titan to optimize hyper parameters
  - USC D-Wave to model complex topologies
  - UTK memristor-based neuromorphic low-power implementation
- Hybrid approach for analyzing scientific data

## Approach

- Find a common problem that can be evaluated on the architecture
  - Titan to explore hyper parameter optimization using a convolutional neural network
  - D-Wave to explore complex DL topologies using a Limited Boltzmann machine network
  - Neuromorphic hardware to explore implementations of spiking neural networks and Boltzmann machine networks

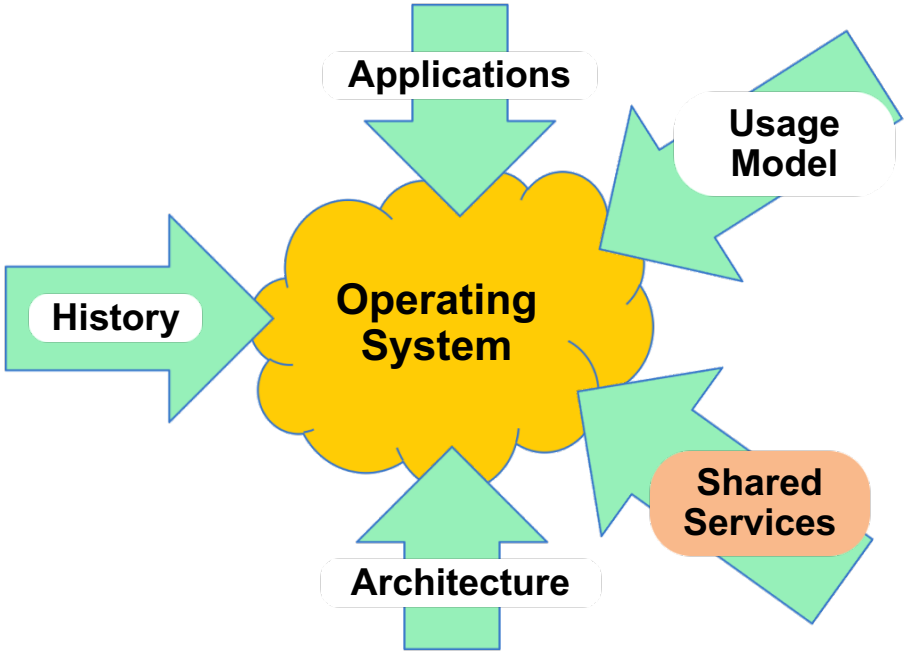



*Evaluating a deep learning network on a hybrid of three distinct architectures leads to proposed architecture*

## Impact

- Demonstrated MNIST Deep Learning solution on quantum, HPC, and neuromorphic computers
  - Trained a complex DL network on a quantum computer that is untrainable on conventional computers
  - Optimized the hyper-parameters on convolutional neural network using 16K models running on Titan
  - Simulated a spiking network running on memristive hardware
- New architecture for hybrid deep learning on scientific data

# Shared Services: Multitenant Storage and Advanced Networking

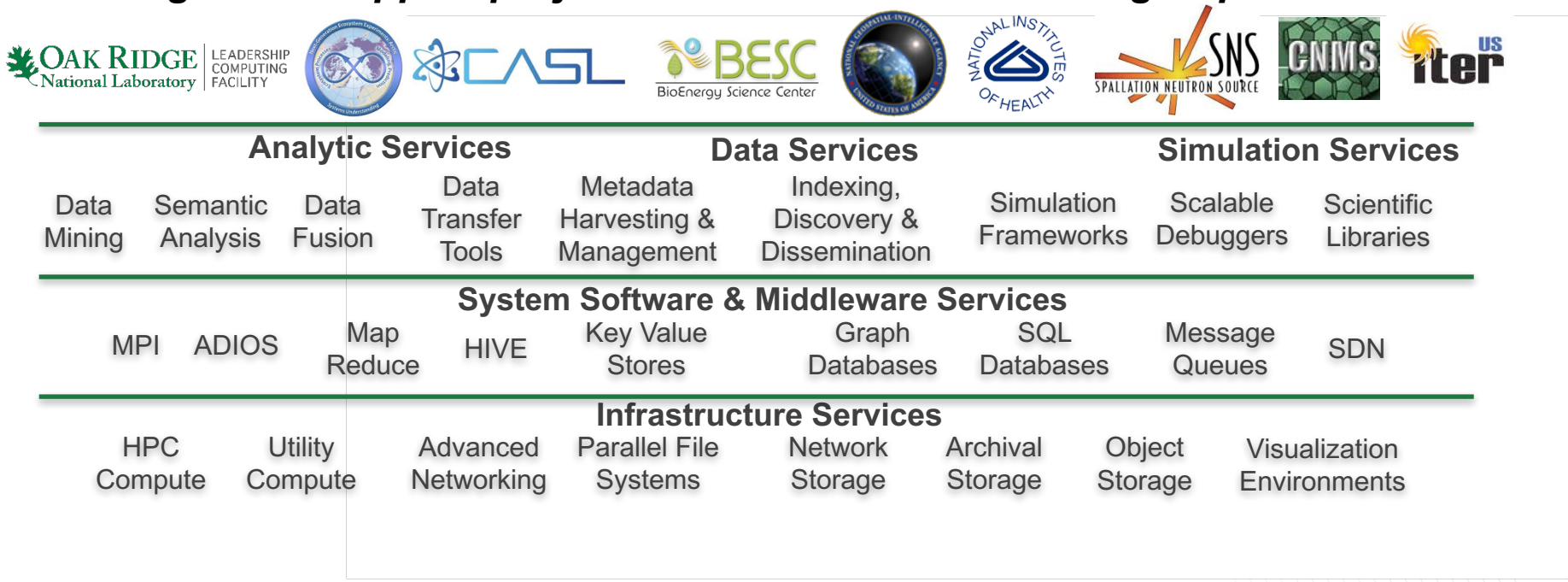


# CADES – Compute and Data Environment for Science

CADES is an **integrated compute and data science infrastructure** and **service portfolio** in support of ORNL Projects and Staff

- A diverse computing and data ecosystem
- Matrix staff with expertise in computing and data science
- Focused on the technical computing needs of the scientific and engineering R&D communities across ORNL
- Designed to deliver solutions to many projects

***Designed to support projects and staff with demanding requirements***

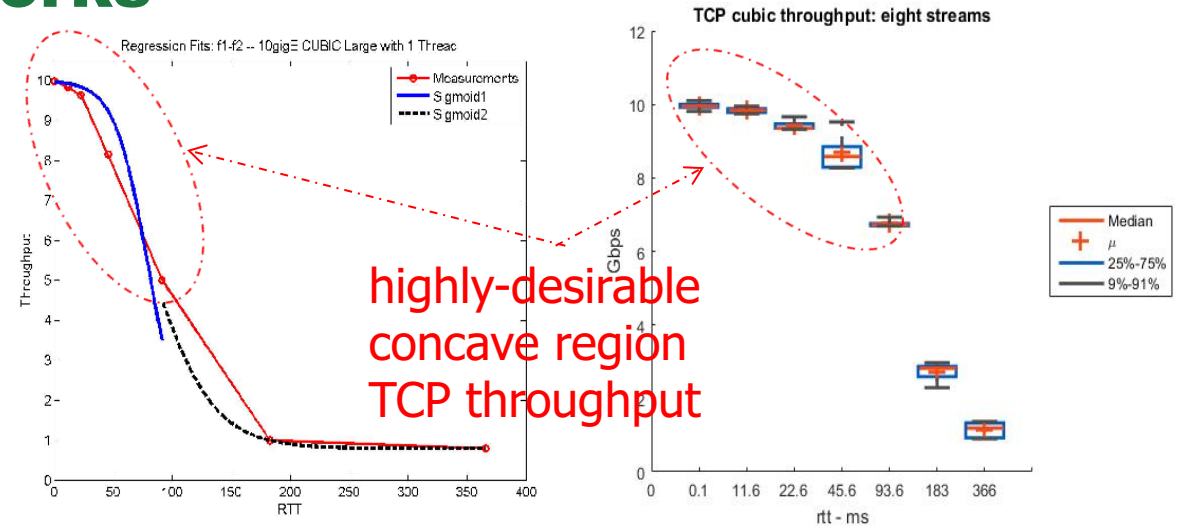


# Data Transfers Over Wide-Area Networks

## *Foundational Contributions:*

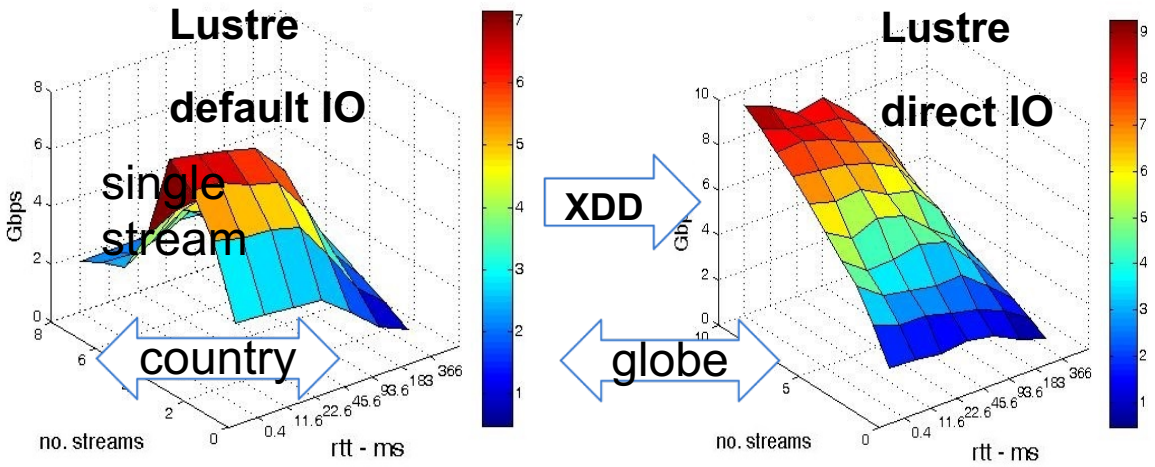
Overview: Develop, analyze and optimize data transfer solutions

- Developed Leading-Edge Solutions: 0-366ms rtt connections
  - optimized XDD file transfers
  - Lustre over Ethernet - LNet solutions
- Structured testbed experiments for performance assessment and tuning
- Developed Foundational Analytics: (i) concave-convex dual profiles, (ii) stability-throughput connection



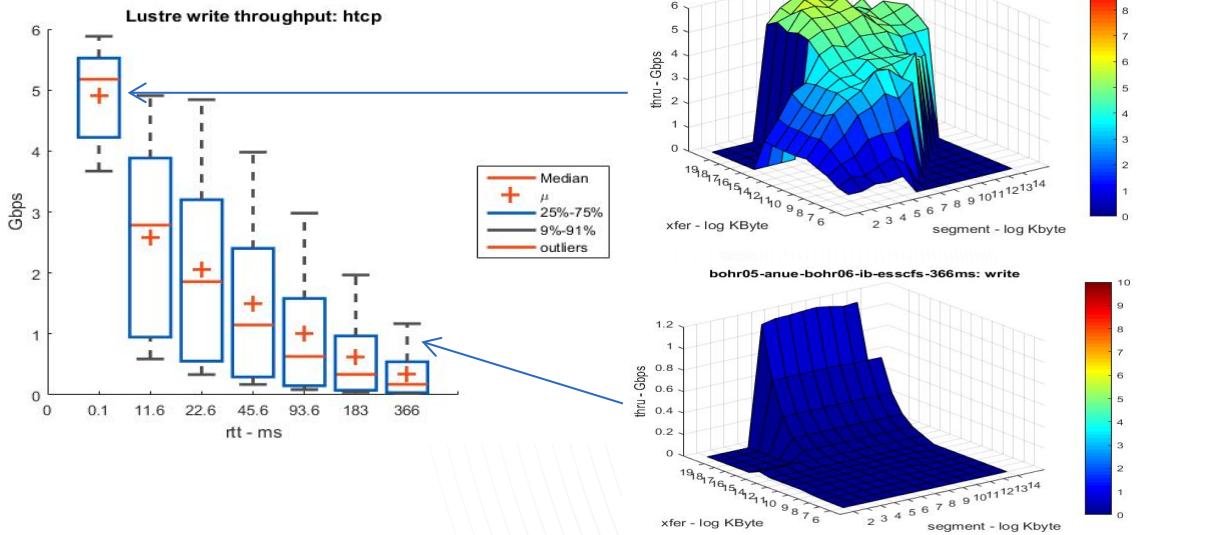

RTT: cross-country (0-100ms), cross-continents (100-200ms), across globe(366ms)

## *XDD Tuning: DirectIO for Improved Throughput*

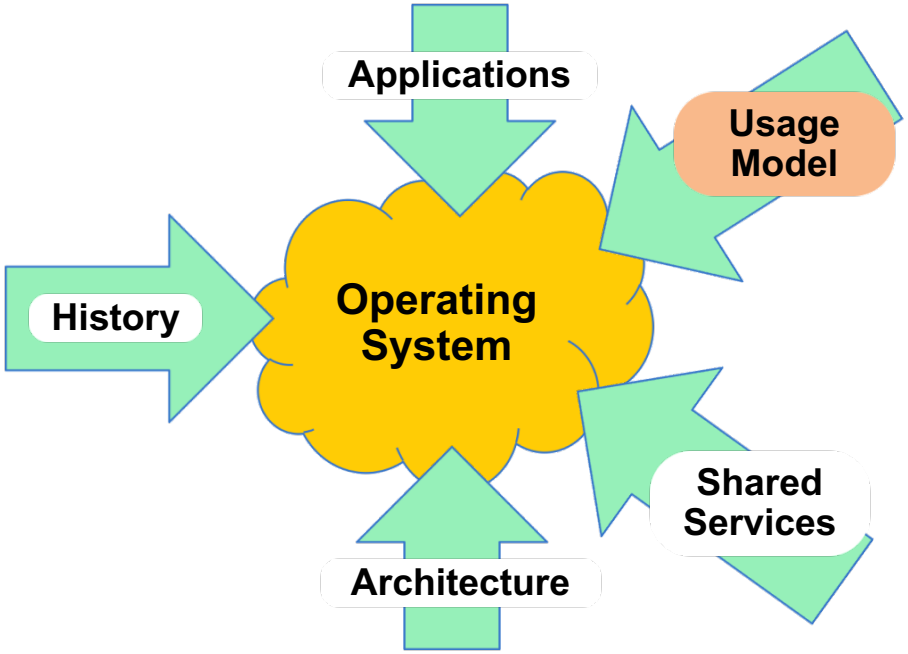



**Profiling leads to unimodality: – easier to optimize**

## *Proof-of-Principle: Lustre across globe*



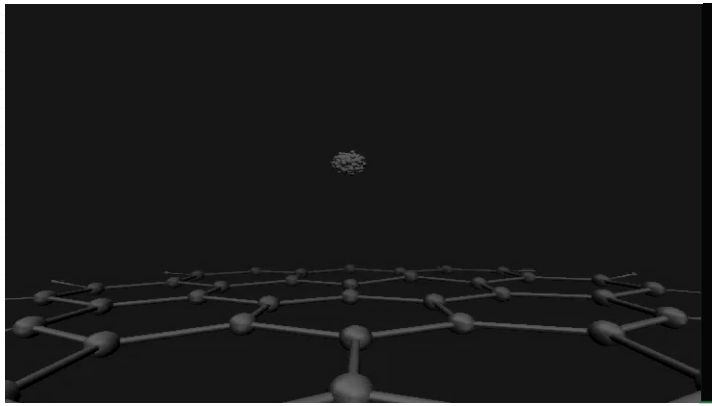
# Usage Models: Near Real Time and Backfill Access to HPC systems



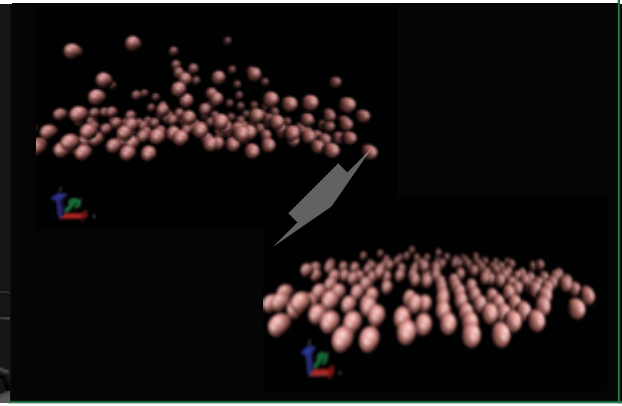
# Reaching atomic precision

We need to understand and control the *dynamic changes* in materials in *confined and non-equilibrium conditions* at the atomic level

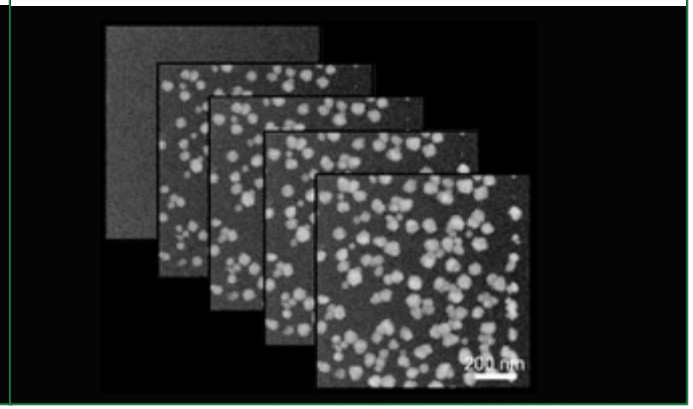
1. Understand how energy is imparted to matter: beam-matter interactions



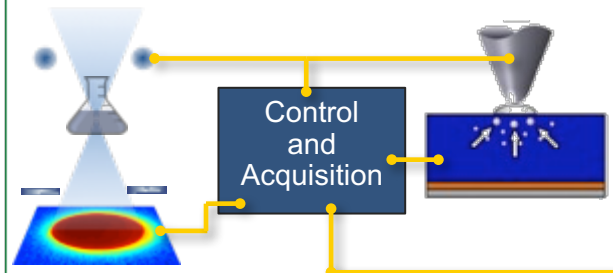
2. How energy and matter behave at the atomic level when confined



3. How reactions/diffusions progress as systems reach new equilibrium

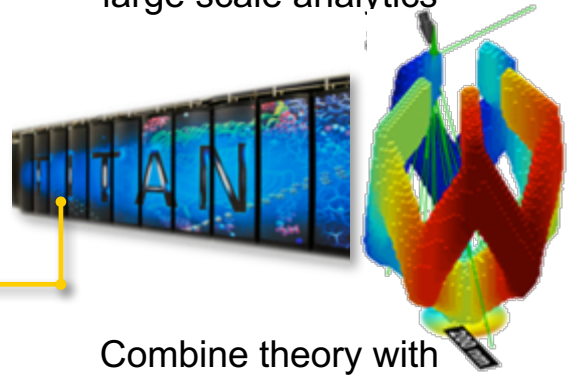


Use microscopes to not just image but transform matter



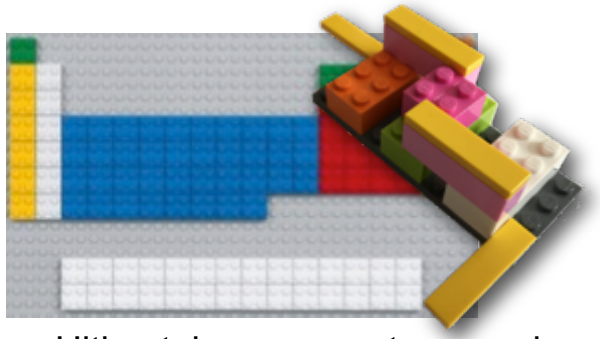

Feedback on dynamic changes to achieve fine control

Develop new theory, modelling, and large scale analytics



Combine theory with advanced analysis

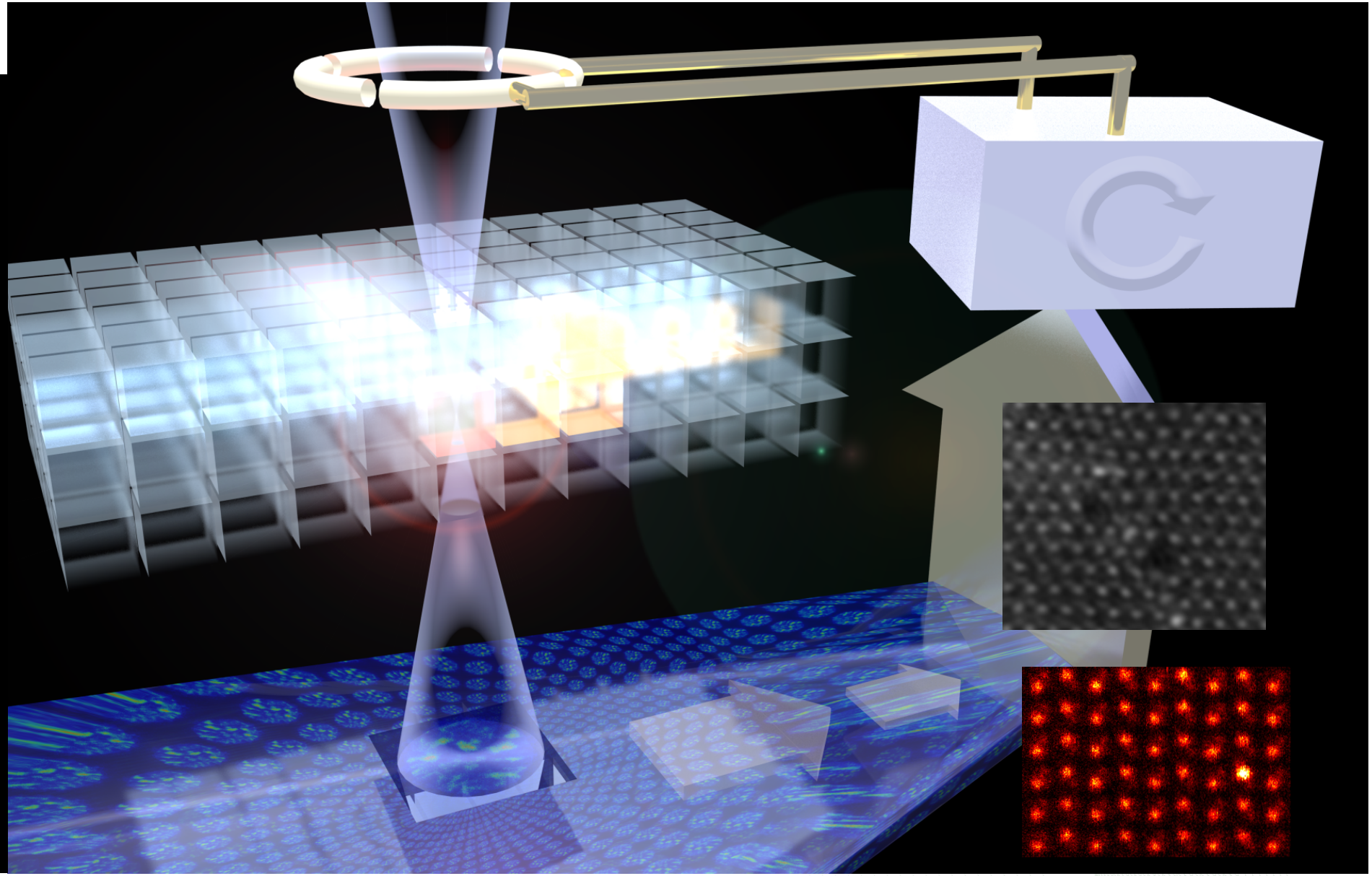
Observe and control dynamics



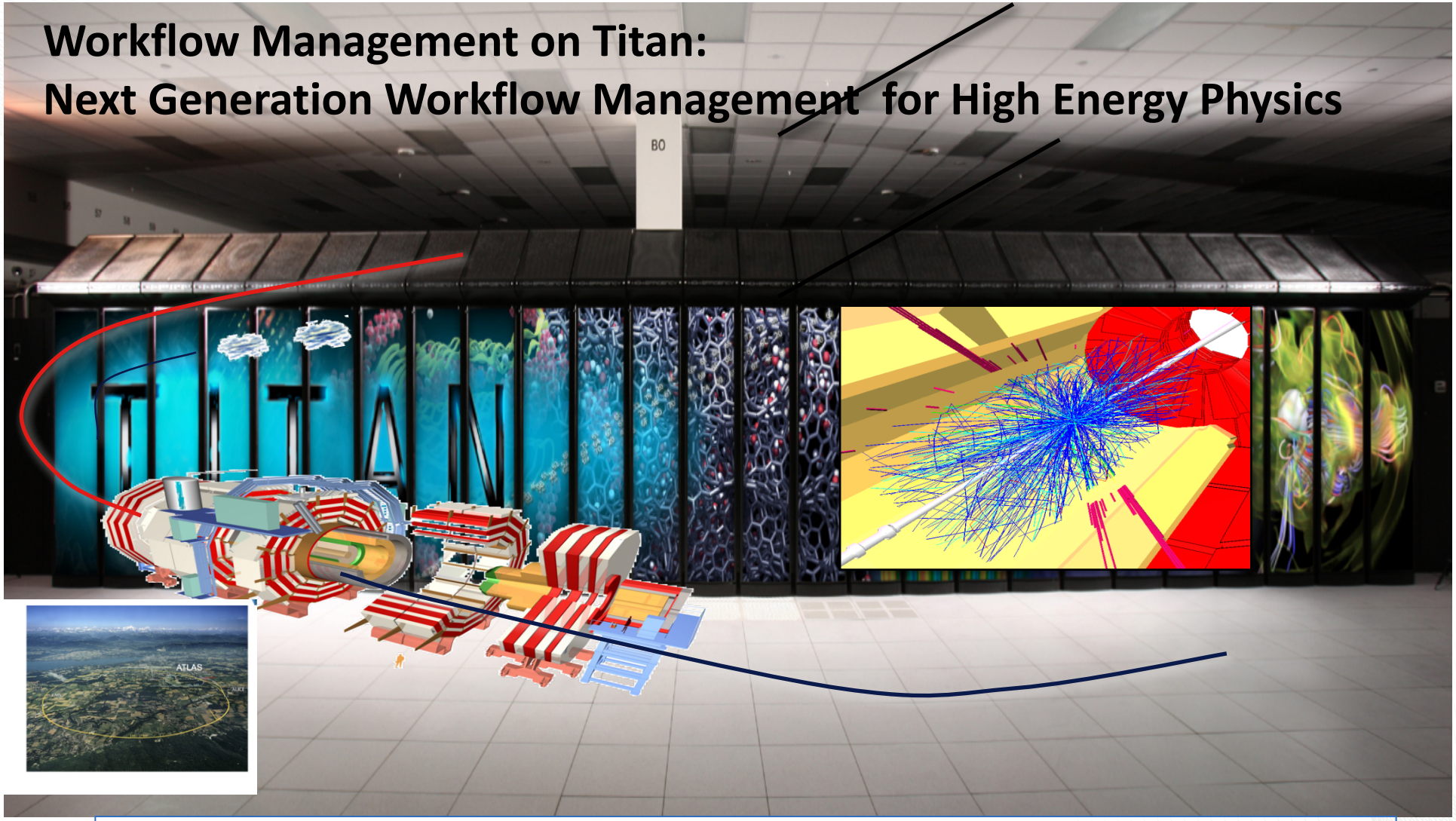
Ultimately arrange atoms and bonds to create new functionality

# Atomic Forge

- Uses ML to form beam (static control of 64 elements of corrector)
- Large volumes of data generation (movies, 4D sets)
- 100s-1000s platforms worldwide
- Need libraries of beam-matter interaction mechanisms
- Need AI control of beam position and intensity
- Will enable 3D atomic fabrication: quantum computing, spintronics, etc.

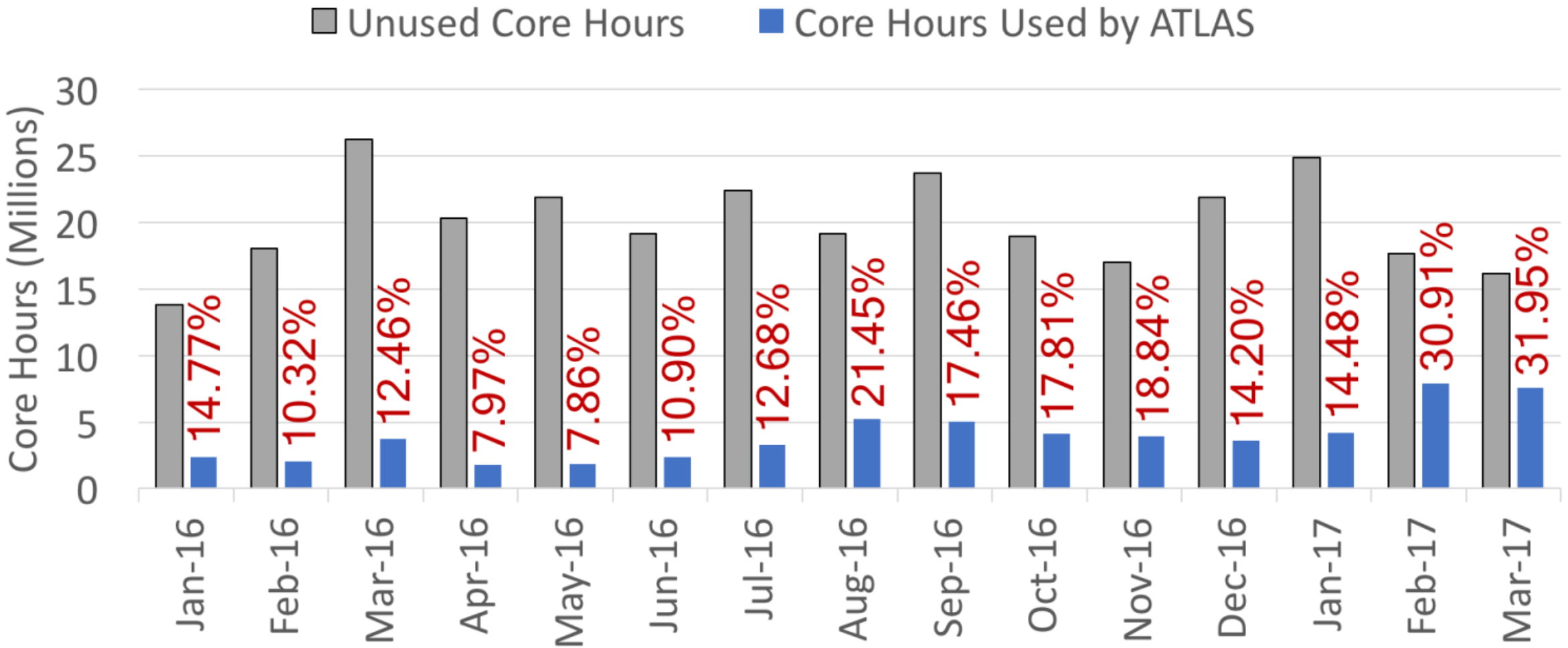


# LHC/ATLAS-OLCF Integration




"BigPanDA Workflow Management on Titan for High Energy and Nuclear Physics and for Future Extreme Scale Scientific Applications," DOE/SC/ASCR Next-Generation Networking for Science, Rich Carlson. PI: Alexei Klimentov (BNL); Co-Pis; K. De (U. Texas-Arlington), S. Jha (Rutgers U) J.C. Wells (ORNL)

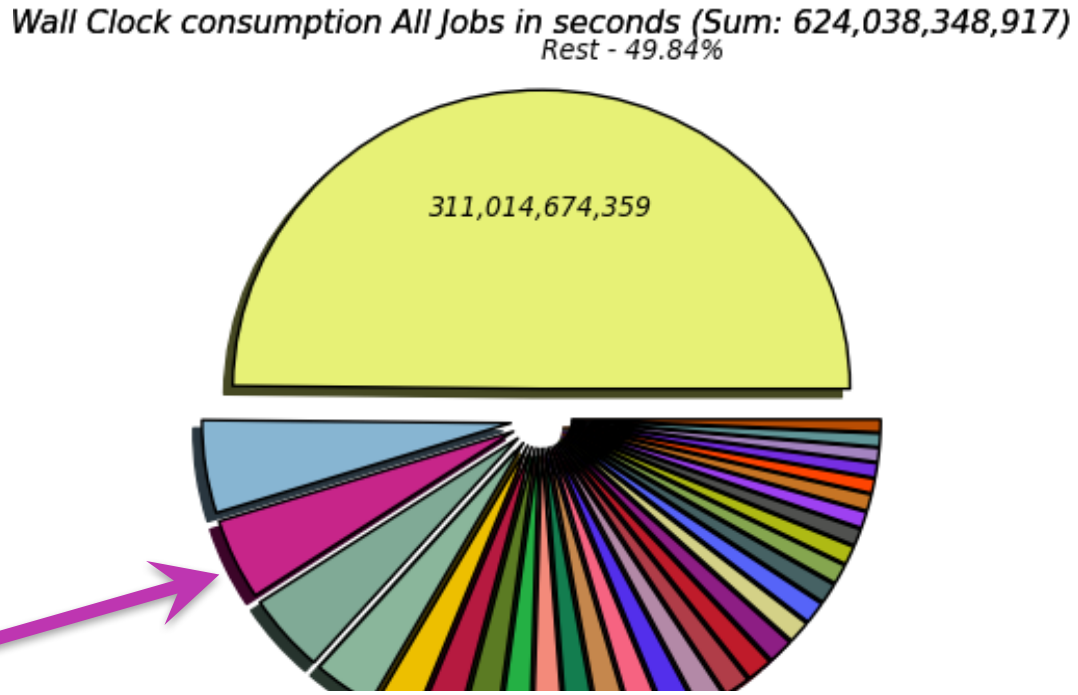
# Making Use of "Unusable Backfill"




Improved Titan utilization by 25 – equivalent of an INCITE project

# ATLAS simulation time worldwide: February 2017

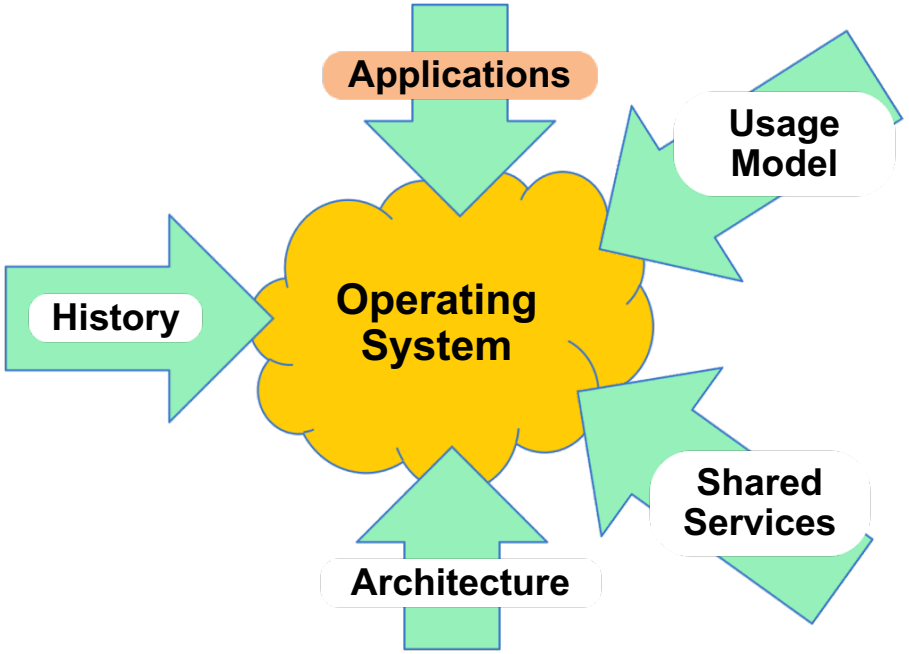
- ATLAS Detector Simulation integrated with Titan (OLCF)
- Titan has already contributed a large fraction of computing resources for MC simulations
  - Titan contributed 4.4% of total simulation time in February 2017.




Rest - 49.84% (311,014,674,359)
ORNL_Titan_MCORE - 4.39% (27,386,161,504)
CERN-P1_DYNAMIC_MCORE - 3.23% (20,185,049,416)
MWT2_MCORE - 2.07% (12,914,398,584)
BU_ATLAS_Tier2_MCORE - 1.66% (10,374,456,280)
CERN-PROD_T0_4MCORE - 1.65% (10,324,383,732)
MWT2_SL6 - 1.62% (10,127,987,628)
RAL-LCG2_MCORE - 1.43% (8,938,111,360)
DESY-HH_MCORE - 1.35% (8,399,710,720)
IAAS_MCORE - 1.19% (7,440,230,864)

CERN-P1_DYNAMIC_MCORE_LOWMEM - 5.02% (31,332,860,045)
BNL_PROD_MCORE - 4.26% (26,560,125,032)
AGLT2_MCORE - 2.16% (13,450,195,231)
BOINC_MCORE - 1.90% (11,842,949,450)
CERN-PROD_SHORT - 1.66% (10,365,047,898)
FZK-LCG2_MCORE - 1.65% (10,281,409,453)
IN2P3-CC_MCORE - 1.57% (9,785,324,136)
SLACXRD_MP8 - 1.36% (8,462,468,468)
TRIUMF_MCORE_LOMEM - 1.34% (8,333,559,388)
plus 12 more

# Applications: New kinds of Applications and New ways of building applications

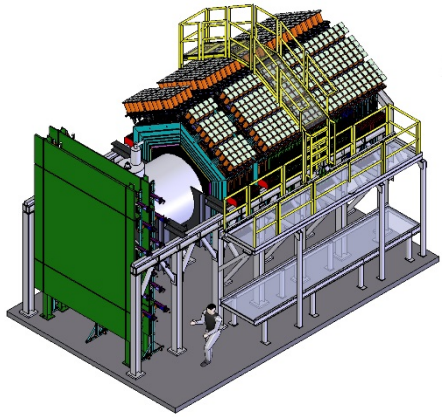


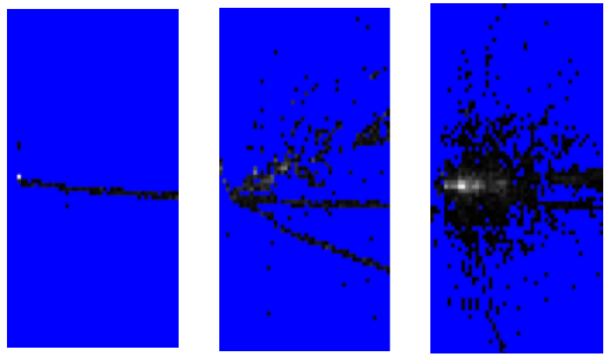
# Multi-node Evolutionary Neural Networks for Deep Learning (MENNDL)

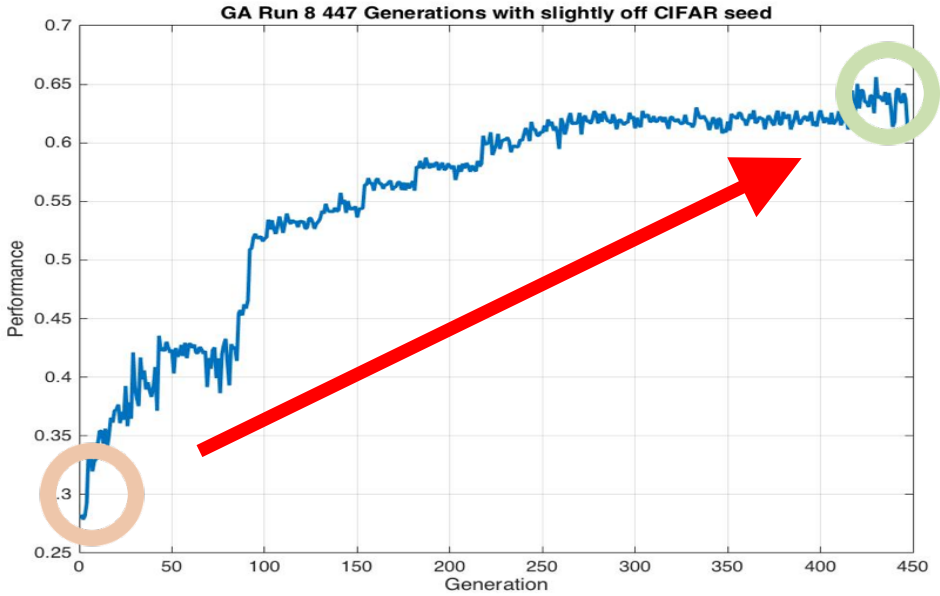
***Premise:*** For every data set, there exists a corresponding neural network that performs ideally with that data

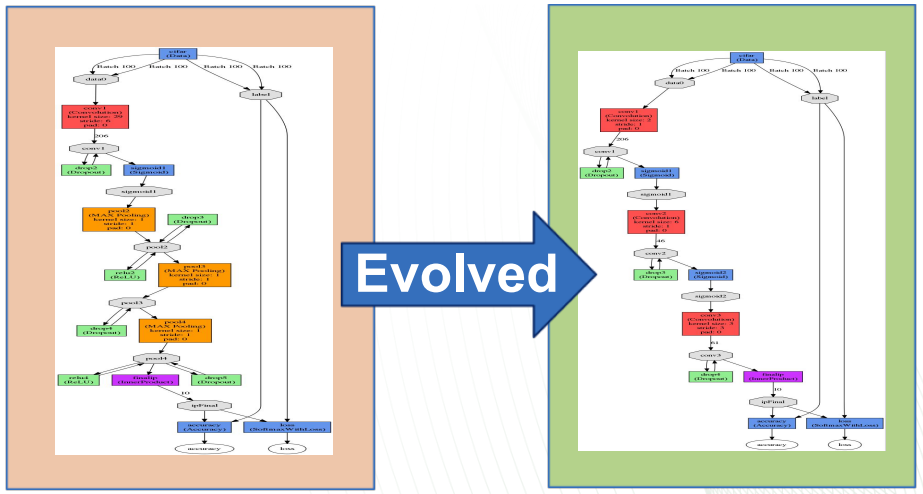
**Demonstrated on 15,000 nodes of Titan using High Energy Physics Data**

- Evaluated against multiple datasets
  - Standard computer vision datasets
  - Neutrino detector vertex finding datasets
- Currently exploring additional datasets and evaluating performance on Summit-Dev







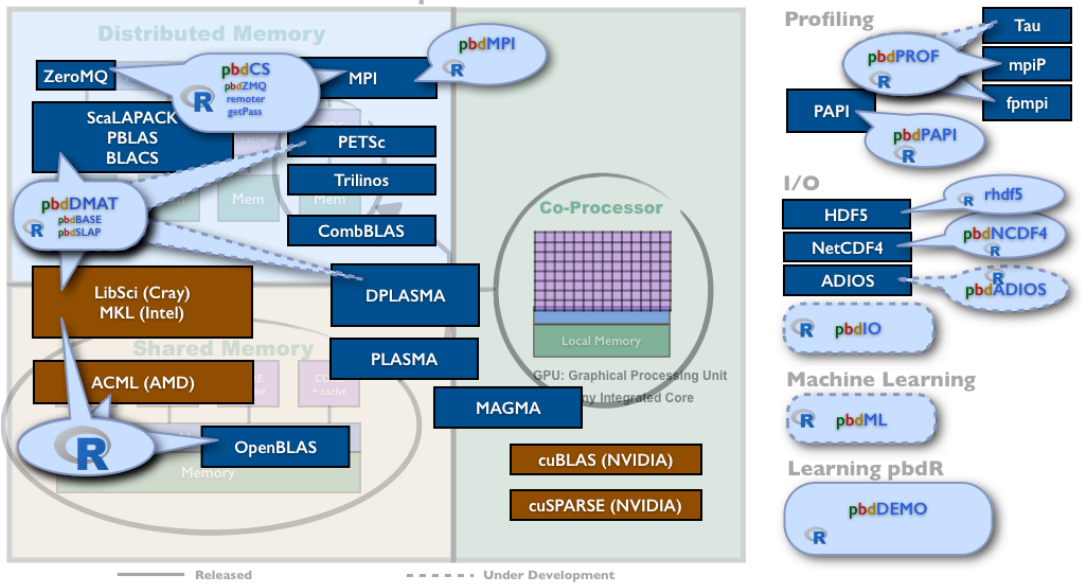
# **Highlight:** Highly Scalable Development Platform for Big Data Analytics

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




- **2016 ORNL Significant Event Award**
- Supported software on OLCF platforms (Rhea, Eos, Titan)
- Scalable algorithm engine for ORNL's BEAM workflow platform
- Scalable algorithm engine for iPLAR (Nanjing University)
- Supporting ORNL's bioinformatics random forest algorithms
- Supporting other DD allocations at OLCF

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."*

HPCwire April 20, 2017

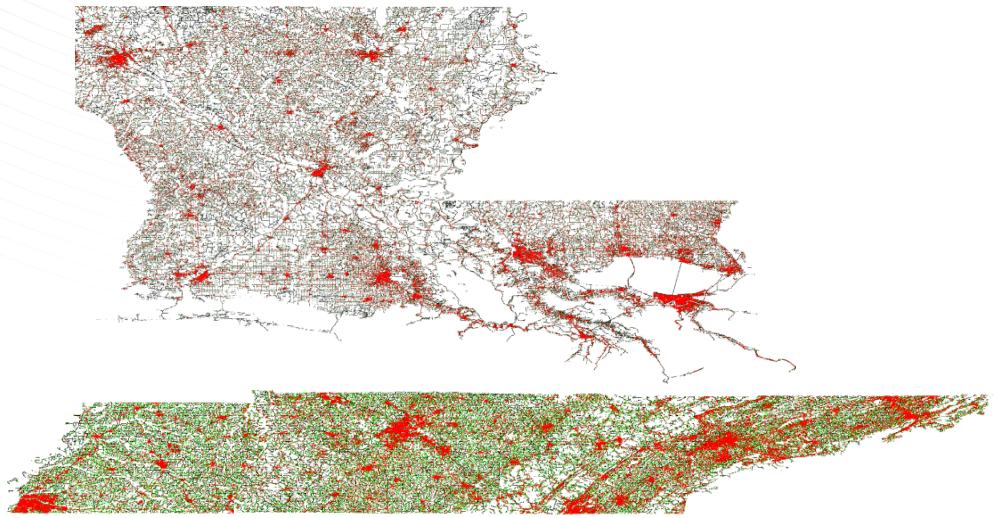
"**ORNL Researchers Bridge the Gap Between R, HPC Communities**"
"**…"untapped [R] domains" represent an enormous potential user base for world-class computers** ."

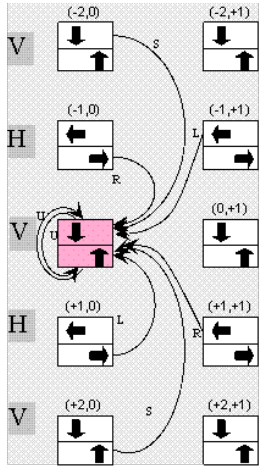
**BIG Data Analytics and Machine Learning = Modern statistical algorithms + pbdR infrastructure + HPC Libraries + HPC Hardware**

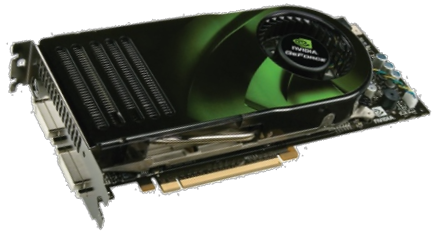
# Interactive, Massive (State or Regional-Scale) Vehicular Transportation Simulations on GPUs

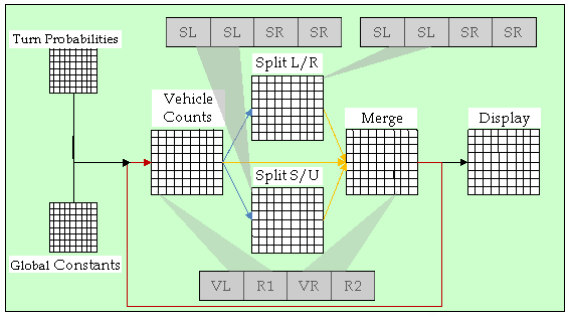
***PI: Kalyan Perumalla***

- Uniquely powerful capability to simulate **millions** of links and intersections
- Significantly faster than real-time even for largest network sizes (e.g., Texas, California)





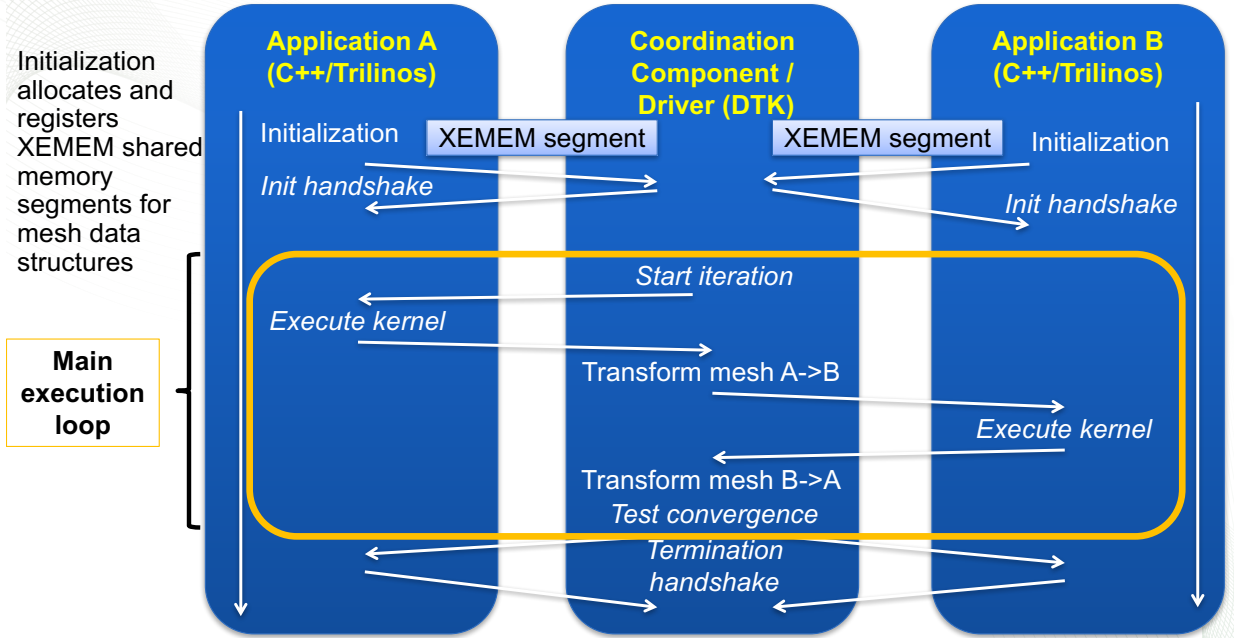



- "Towards Highly Interactive, GPU-based Evaluation of Evacuation Transport Scenarios at State-Scale," K. Perumalla, B. Aaby, S. Yoginath, and S. Seal, **SIMULATION**, Vol. 88(6), pp. 746-761 [*Journal Article*]
- "Towards Highly Interactive, GPU-based Evaluation of Evacuation Transport Scenarios at State-Scale," K. Perumalla, B. Aaby, S. Yoginath, and S. Seal, **National Evacuation Conference**, 2010
- "GPU-based Real-Time Execution of Vehicular Mobility Models in Large-Scale Road Network Scenarios," K. Perumalla, B. Aaby, S. Yoginath, and S. Seal, **IEEE/ACM/SCS PADS**, 2009

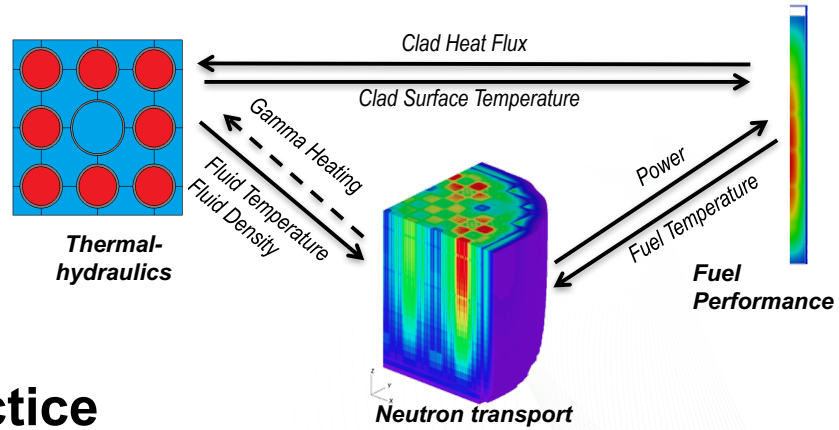
# System Support for Composition of Complex Applications

## Demonstrated application with minimal intrusion

- Used DTK for mesh exchanges
- Currently using artificial applications (but the physics makes sense)




## Example Composition

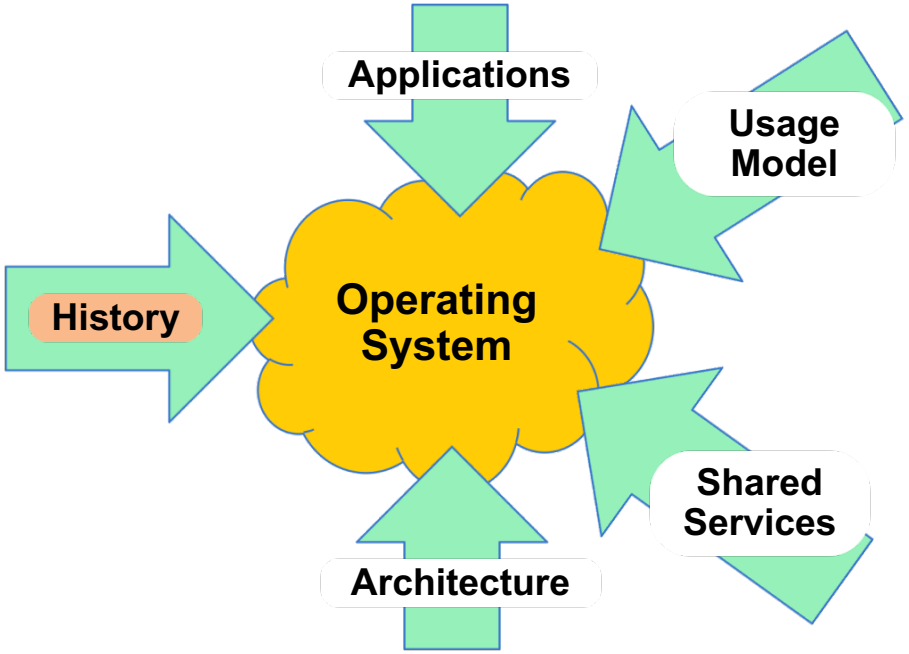



## Current Practice

- Transform components into libraries
- DTK becomes the main driver
- Costs:
  - More intrusive
  - Need to maintain a common build environment (4 FTEs in CASL)

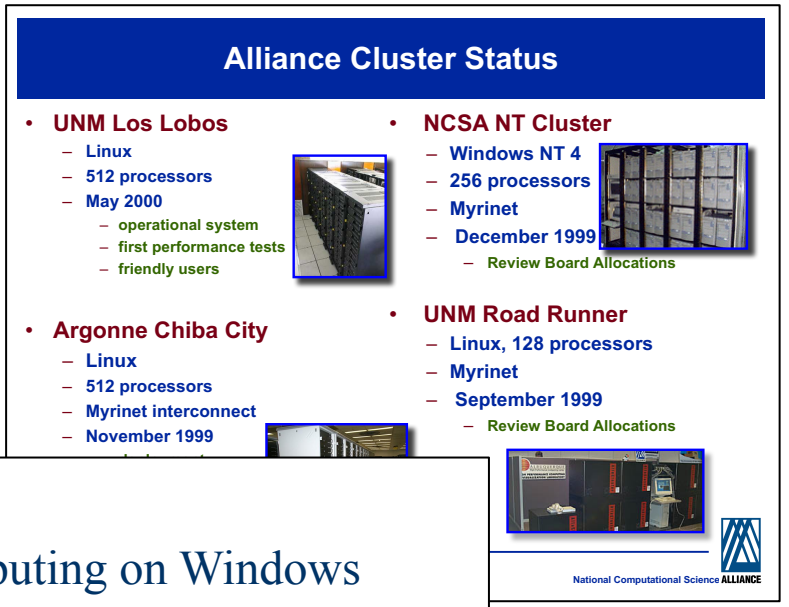
| Measurement | Baseline (µs) | Process (µs) | Container (µs) |
|---|---|---|---|
| Command queue | 48.3 ± 20.8 | 48.3 ± 20.8 | 37.1 ± 83.0 |
| appA -> appB mesh transformation | 14,080 ± 852 | 9,986 ± 692 | 9,400 ± 135 |
| appB -> appA mesh transformation | 17,298 ± 143 | 13,570 ± 111 | 13,153 ± 127 |
| Complete application iteration | 31,395 ± 660 | 25,132 ± 312 | 23,422 ± 469 |

# History: New kinds of users who don't have the common history

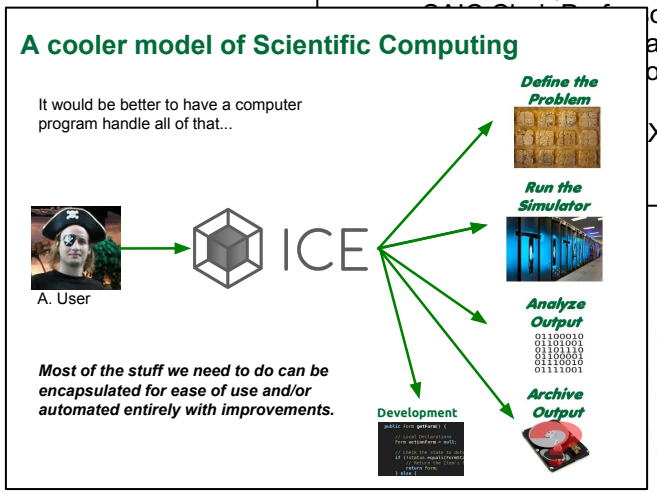


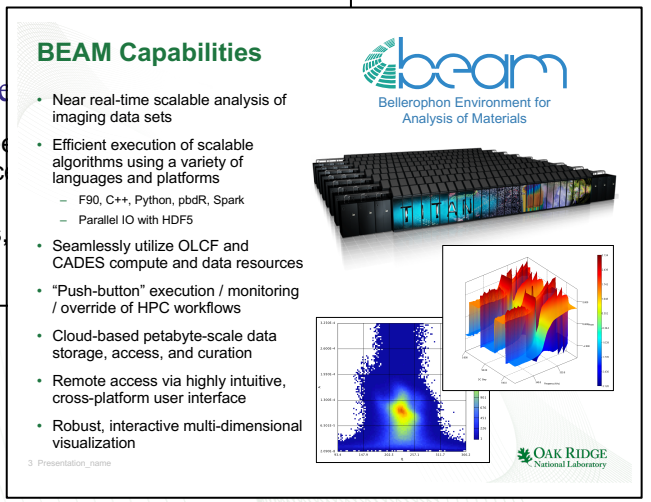
# History

- Scientific computing has been equivalent to Unix for several decades
  - Some instances of special purpose OS, but it's been Linux since 2000
  - Users understand the environment and have learned to work within its limitations
- Workflow systems are coming of age
  - New users will expect to be able to "mash up" applications and data sets
  - They won't be willing to make the investment needed to use current HPC systems

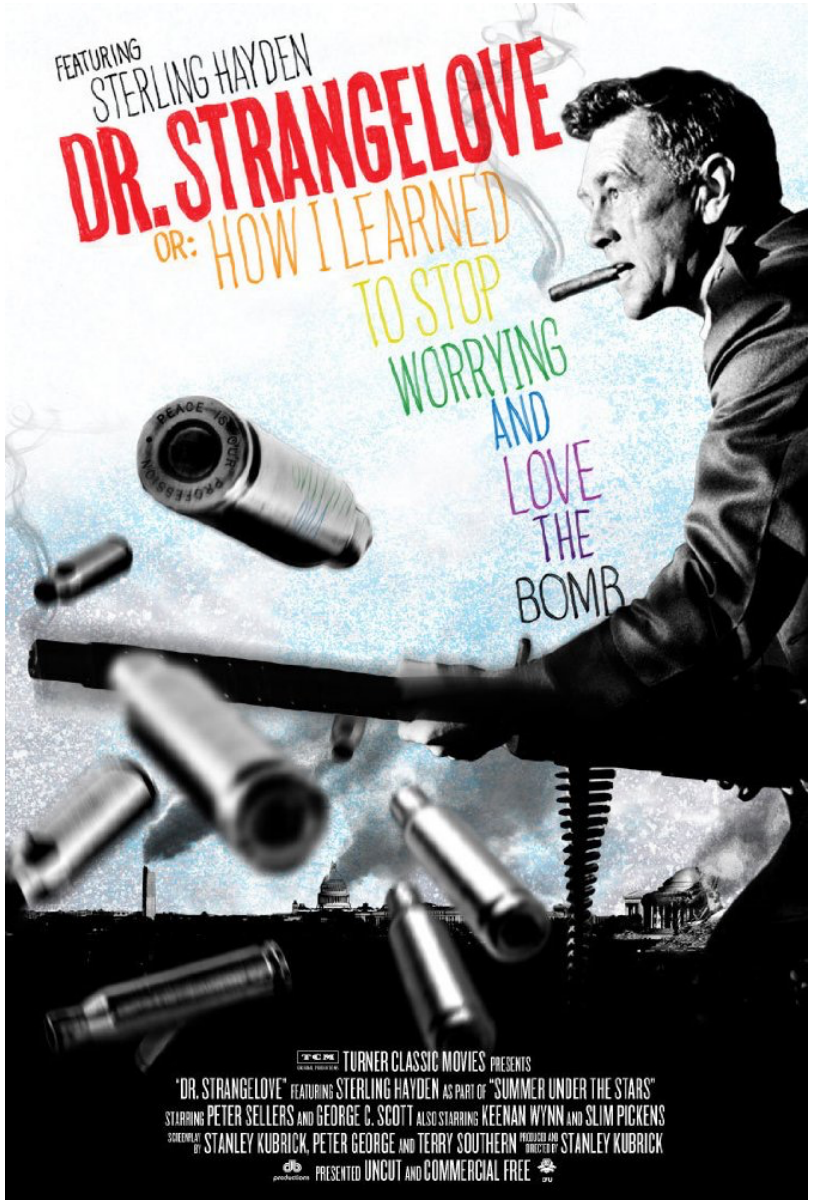






# Learning to live with Extreme Heterogeneity



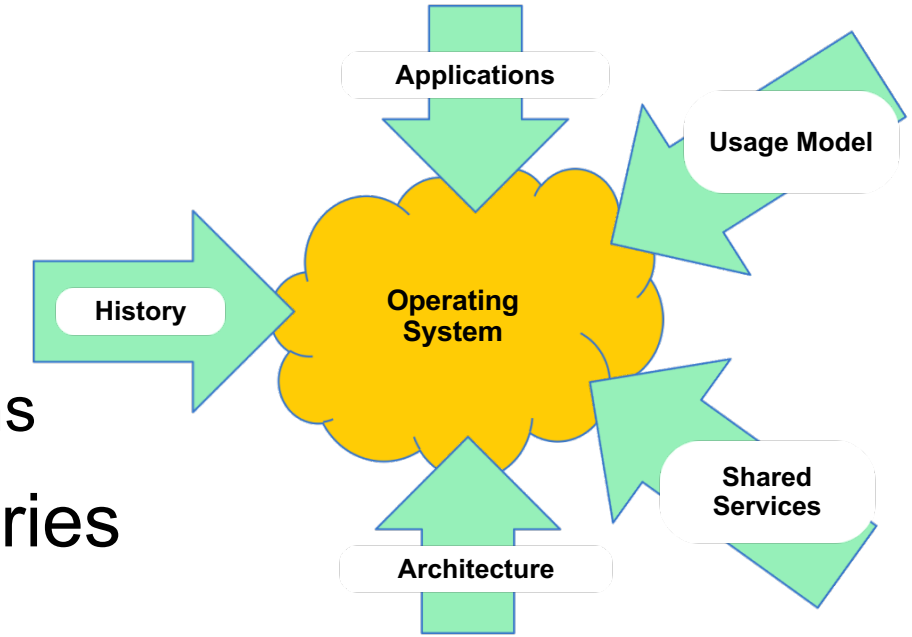

OAK RIDGE National Laboratory

# Extreme Heterogeneity (ASCR Summit)

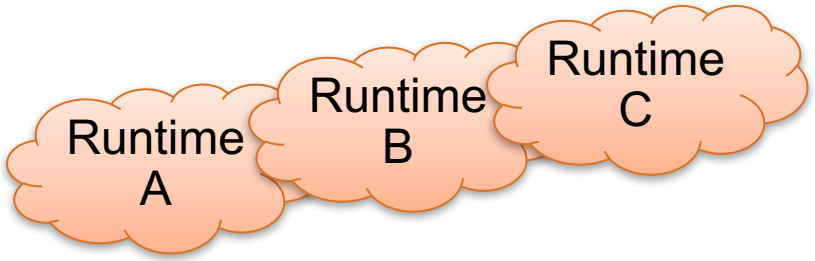
1. Dramatic increase in the diversity of processing and memory technologies available for HPC;
2. Increasingly diverse workflows for science and national security that bring increasing need to employ multiple heterogeneous technologies in concert to accomplish DOE mission goals;
3. An explosion in HPC use by DOE scientists across a wide array of new domains, as other programs in the Office of Science report that users of their Science User Facilities need to use supercomputing to accomplish their science, with demands such as near-real-time data analysis to support steering experiments, requiring the use of machine learning and artificial intelligence techniques; and
4. A broad range of programming expertise exhibited by our users, in part a consequence of (1) and (3).

# What is the role of an OS?

- Manage shared resources
  - Allocate resources to applications
  - Provide isolation between competing applications
- Interoperability of runtime systems and libraries
- Provide a abstractions for these resources
- Connect to other systems
- Runtime systems do the same thing at a “higher” level
  - Potentially have more understanding of how the resources will be used
  - May not have a strong notion of isolation; sharing may be more important
  - May not be concerned with other systems
  - The implementation of a programming model



# The critical challenge: co-design




- OS/R emphasized OS-R co-design
  - Enabling runtimes (and applications) to manage resources
  - Avoid switching between privilege levels
- The other co-design is harder
  - Must to be clear about what is needed
    - E.g., ability to virtualize and isolate virtual copies
    - May be able to entirely eliminate the OS!
  - Needs are never absolute
    - System software developers build careers on fixing problems with hardware: ***some accidental, some due to lack of foresight***
    - Examples
      - SUNMOS and broken NICs
      - Essential goal of Puma was to virtualize the network (and the local memory system)
- Critical need for testbeds

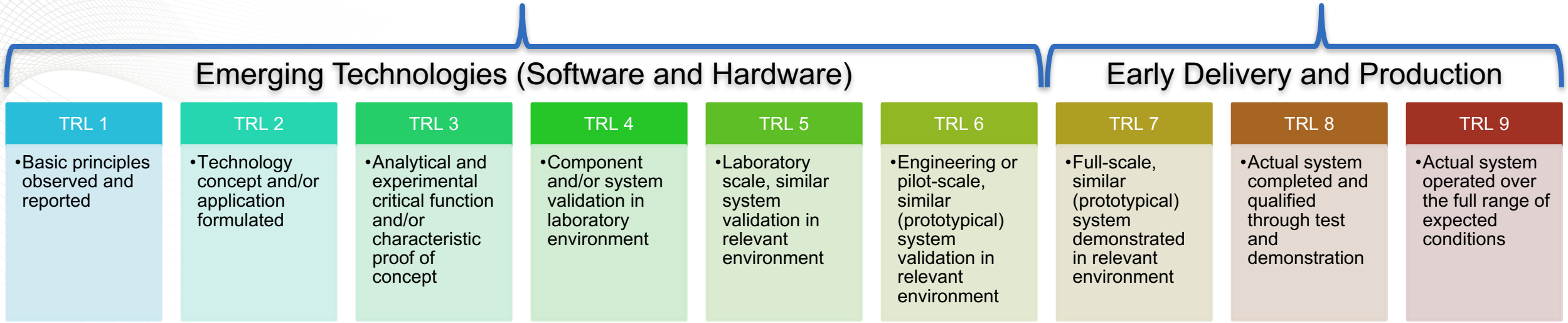
# Specific Challenge: Identify specific hardware features that might enhance functionality that you need

- I've made a career out of doing as little as possible: I highly recommend it ☺
  - Only do what is needed; then, try to make it someone else's problem
- First two papers: understanding OS Noise
  - Identify source of noise and then consider hardware solutions, e.g., offload "progress" to the NIC
- Next paper: Containers (admit it, multi-OS needs virtualization ☺)
  - Critical needs are the same as for virtualization: ability to give resources to a container, without constant OS attention
- Qthread scheduling
  - User level event (timer interrupt) handling would allow preemptive scheduling
- Asymmetric Performance: composing applications and/or sharing nodes
  - Studying the right problem: isolation;
- Telemetry Framework: getting the information needed for adaptive resource management
  - Application and system agnostic; the "hourglass" principle; essential for adoption; also essential to define better
- UNITY: Building new abstractions for memory and storage
  - Identify protocols to support multitenancy (e.g., LWFS ☺ ☺)

# ExCL: Engaging the Community

Testbed systems are essential:
1) explore the space
2) reproducibility
3) community

# Technology Readiness Levels (TRLs)




CNT, Quantum Computing | FPGAs, TrueNorth, D-Wave | Titan, Cori

CNMS, CRF, ***ESCRC***

LCFs, NERSC

ECP

- Consider dimensions of computing
  - Components (e.g., CPU, GPU, Memory)
  - Integration (e.g., CPU+GPU+Memory+NoC)
  - Scale (e.g., 1, 10, 100, 1000, 10000, 100000)
  - Applications (e.g., kernels, miniapps, mission)
  - Performance (e.g., measured, simulated, analytically modeled)
  - Timescales
  - Diversity

- Calibrate with full range
- Similar for software technology
- User Productivity

https://www.directives.doe.gov/directives-documents/400-series/0413.3-EGuide-04

# NSF Keeneland 2009 to 2015

- NSF Track 2D CSA: Largest project in history of GT College of Computing
- **Served 942 total users** of KIDS and KFS for research, development and educational purposes;
- **Contributed to at least 367** publications, presentations, and other software artifacts;
- **Provided over 50 million CPU hours** were used on KFS and over 10 million SUs were provided to XSEDE (equivalent to ~50 million SUs on a CPU-centric system);
- **Averaged 81.2% GPU utilization** over Keeneland's Full Scale 2-year production run (all GPUs, 24h, 7d);
- **Contributed to the development of many early software packages for GPU** heterogeneous computing: Scalable Heterogeneous Computing (SHOC) Benchmarks, MAGMA BLAS libraries, Ocelot emulator, and many others;

http://keeneland.gatech.edu

## Keeneland – Full Scale System

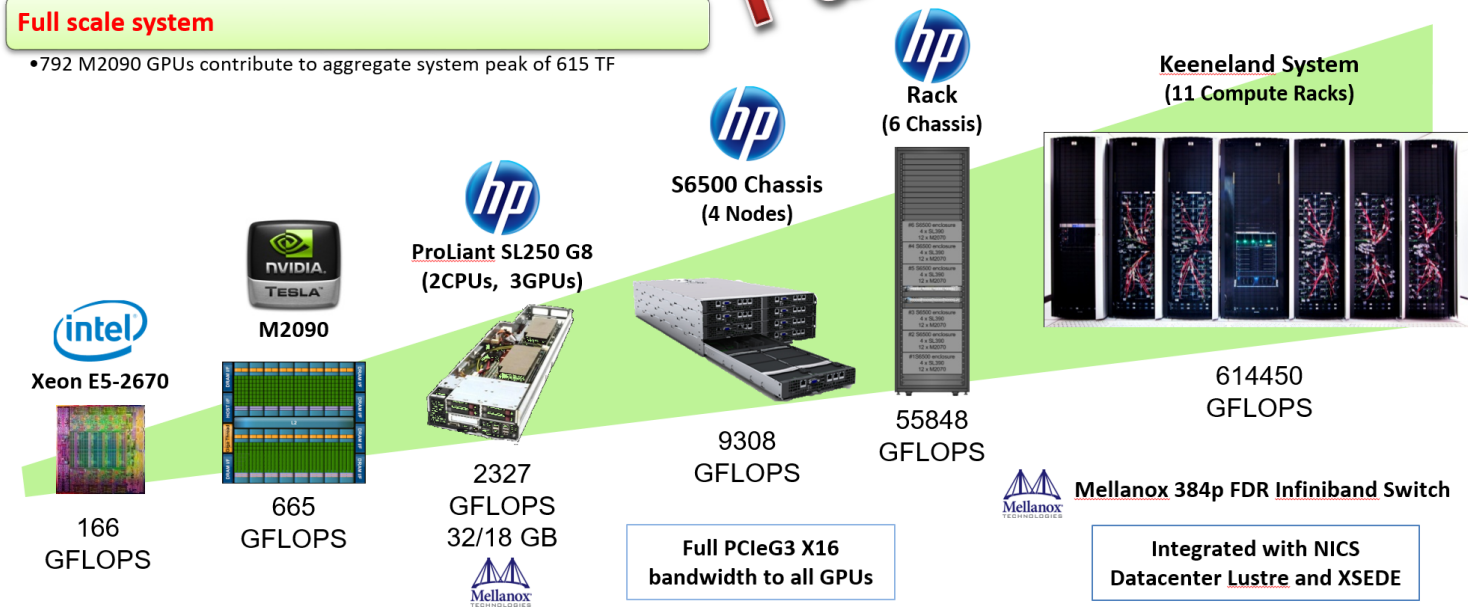
**Initial Delivery system** installed in Oct 2010

- 201 TFLOPS in 7 racks (90 sq ft incl service area)
- 902 MFLOPS per watt on HPL (#12 on Green500)
- Upgraded April 2012 to 255 TFLOPS
- Over 200 users, 100 projects using KID

**Full scale system**

- 792 M2090 GPUs contribute to aggregate system peak of 615 TF

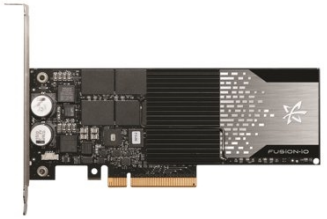
Full Scale

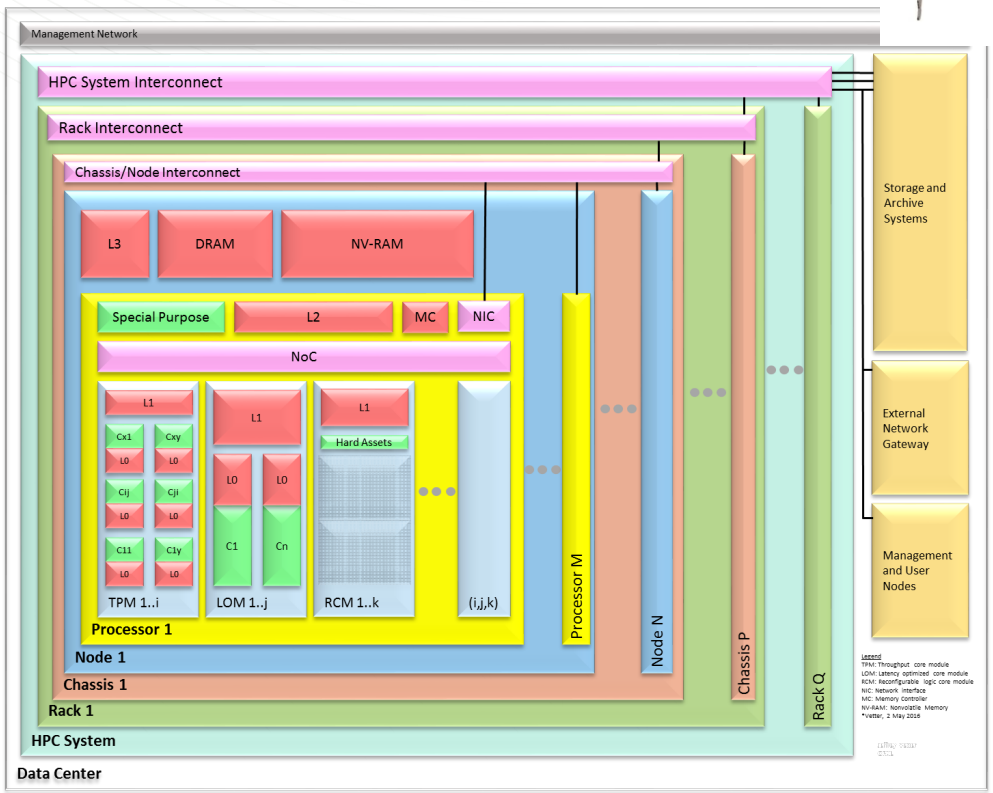



J.S. Vetter, R. Glassbrook *et al.*, *"Keeneland: Bringing heterogeneous GPU computing to the computational science community," IEEE Computing in Science and Engineering, 13(5):90-5, 2011, http://dx.doi.org/10.1109/MCSE.2011.83.*

# Reinvesting in ExCL: CCSD invested in two new platforms for ExCL

- Extreme Heterogeneous Cluster (4 nodes)
  - FusionIO card – 1TB
  - Pascal GPU
  - (2) Altera Arria 10 FPGAs







- Emu Systems (serial number 1 dev system)
  - Data analytics graph engine
  - Novel migrating thread architecture