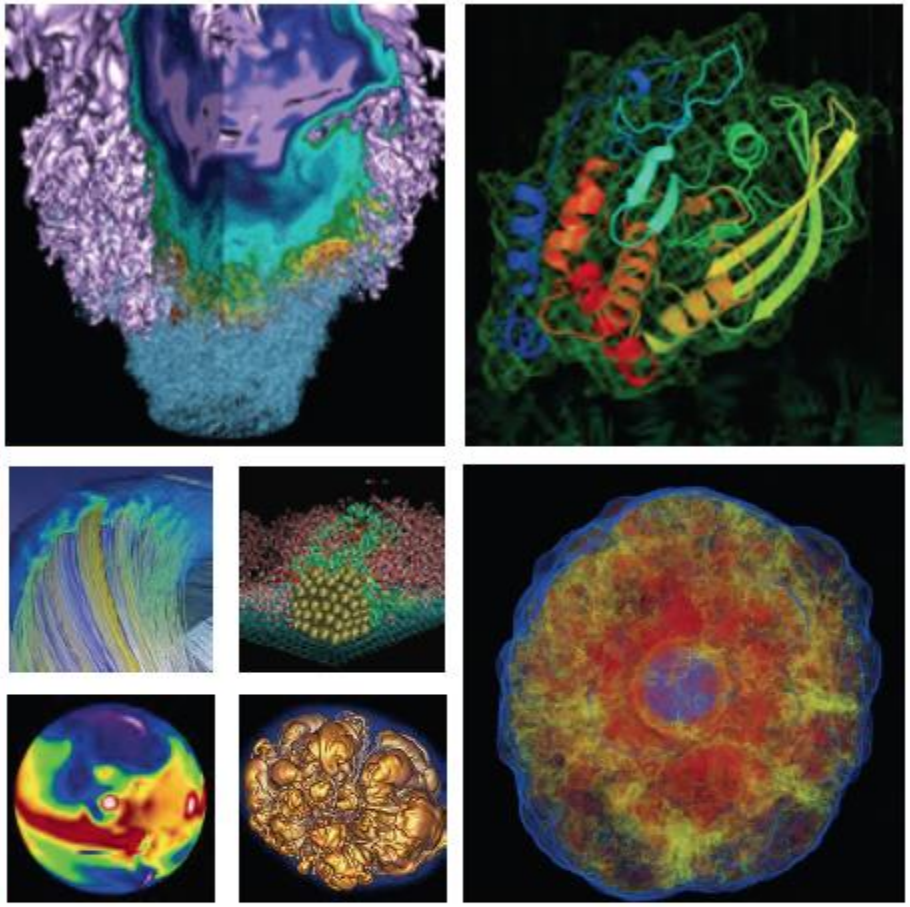
Supporting Data Intensive Workloads at NERSC



NERSC

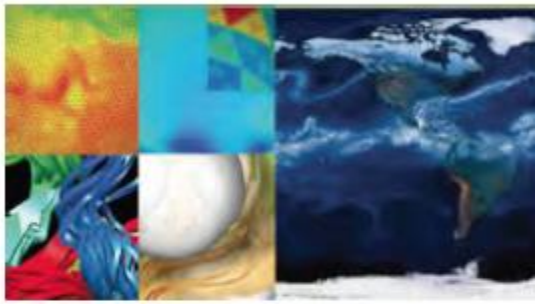
July 5, 2017
Katie Antypas
Data Department Head

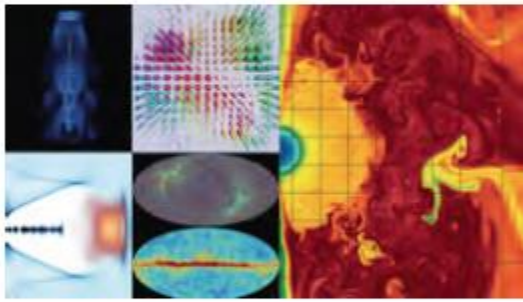
U.S. DEPARTMENT OF ENERGY | Office of Science

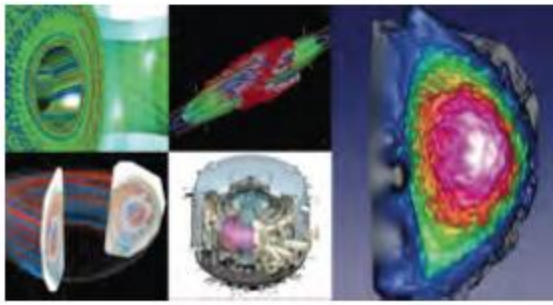
BERKELEY LAB

NERSC is the mission HPC computing center for the DOE Office of Science

- NERSC deploys advanced HPC and data systems for the broad Office of Science community
- NERSC staff provide advanced application and system performance expertise to users
- Approximately 6000 users and 750 projects
- Over 2000 publication resulting in NERSC resources per year
- New Data Initiative: ***Pioneer new capabilities to enable scientists to make large-scale data-intensive science discoveries.***

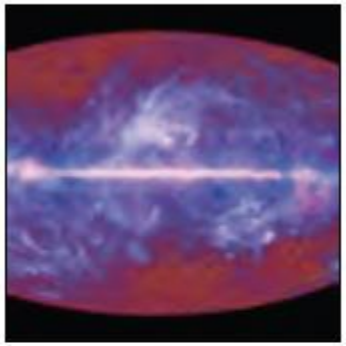






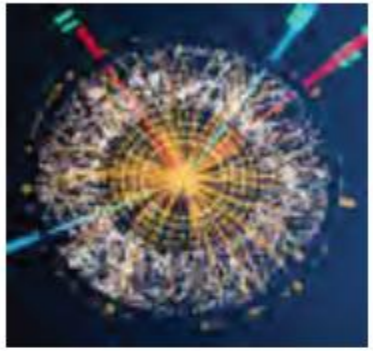
NERSC has been supporting data intensive science for a long time

Palomar Transient Factory
Supernova



Planck Satellite
Cosmic Microwave Background Radiation

Alice
Large Hadron Collider



Atlas
Large Hadron Collider

Dayabay
Neutrinos

ALS
Light Source

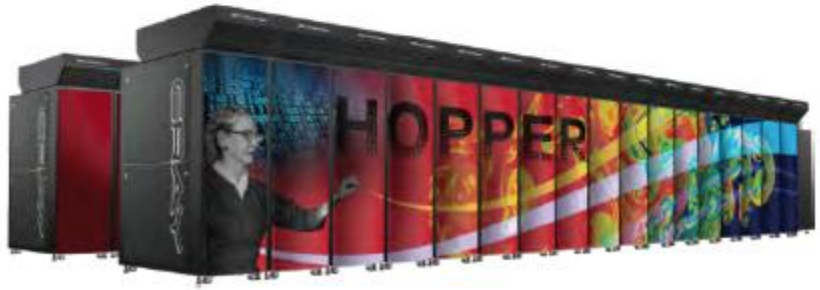
LCLS
Light Source

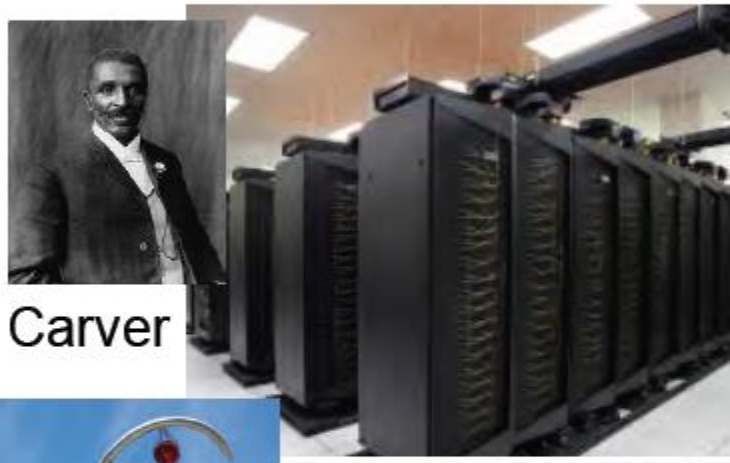
Joint Genome Institute
Bioinformatics

Historically NERSC has deployed separate Compute Intensive and Data Intensive Systems

NERSC

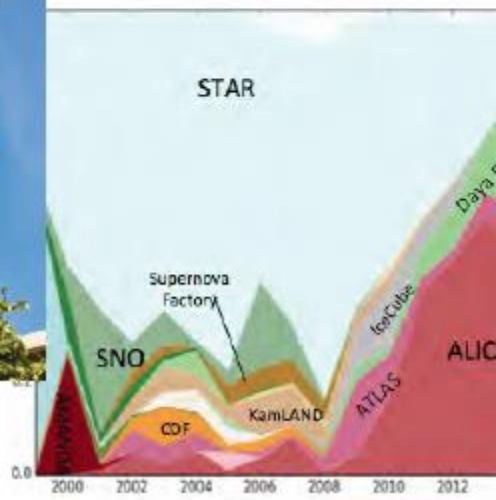
Compute Intensive ***2013*** ***Data Intensive***





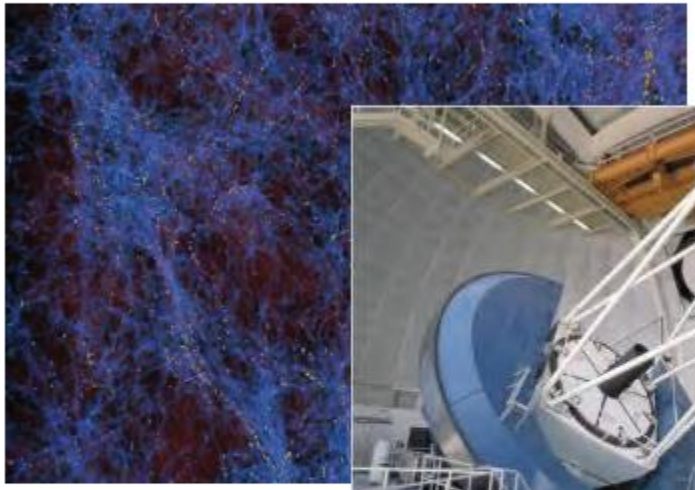
Carver

Genepool



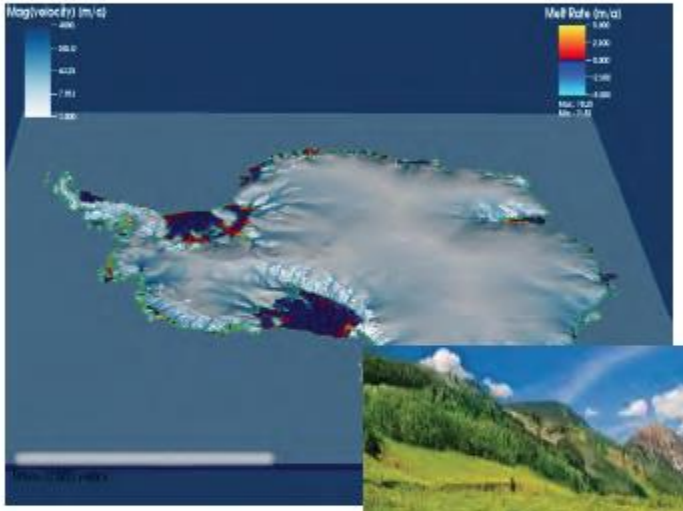
PDSF

What has changed? Coupling of experiments with large scale simulations



Nyx simulation of Lyman alpha forest

Kitt Peak National Observatory's Mayall 4-meter telescope, planned site of the DESI experiment



New climate modeling methods, produce new understanding of ice

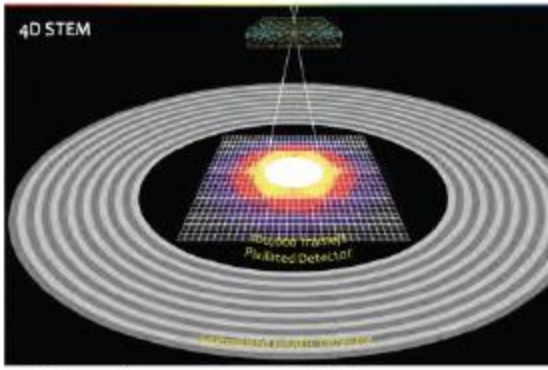
Genomes to watersheds

What has changed? Increased data rates and new sensing capabilities

LCLS
Light Source

Advanced Lightsource
Upgrade

Environmental
sensors



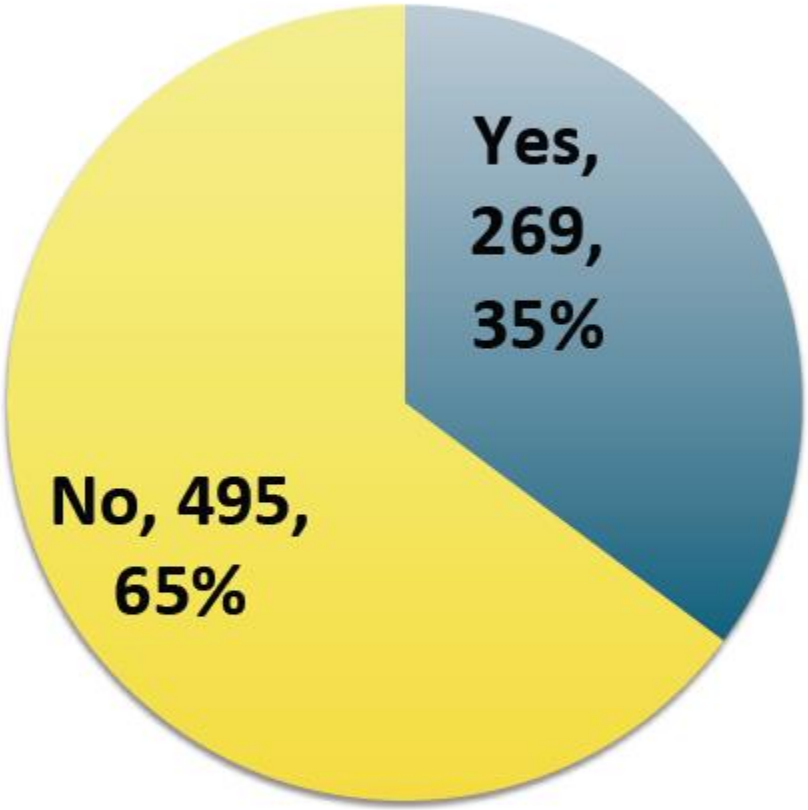
Next generation
electron microscope

Sequencers that fit into
the palm of your hand

- In the next 5 years, data rates will be approaching Tb/sec for many instruments
- Infeasible to put a supercomputer at the site of every data generator

New Data Question This Year on Propsals

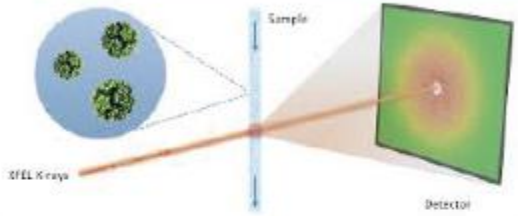
- **Is the primary role of this project to:**
  - Analyze data from experiments/ observational facilities; OR
  - Create tools and algorithms for analyzing exp/obs data; OR
  - Combining models and simulations with exp/obs data?



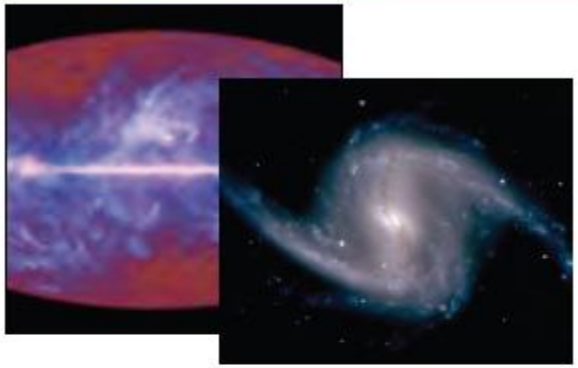
764 projects

The future of data intensive projects on NERSC systems, is now

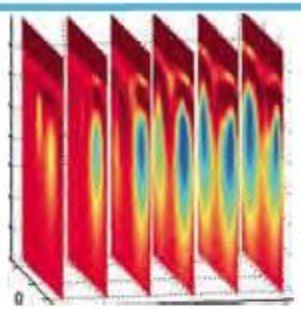
Some exemplars



ASCR: Algorithms for next generation light sources
PI: Sethian



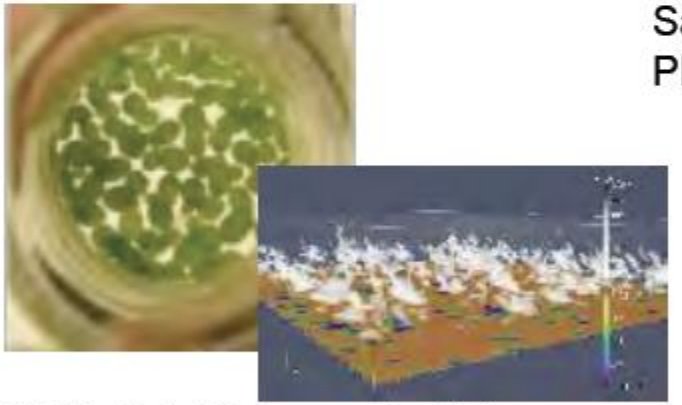
HEP: CMB Data Analysis for Planck Satellite
PI: Borrill



BES: Large Scale 3D Geophysical Inversion & Imaging
PI: Newman

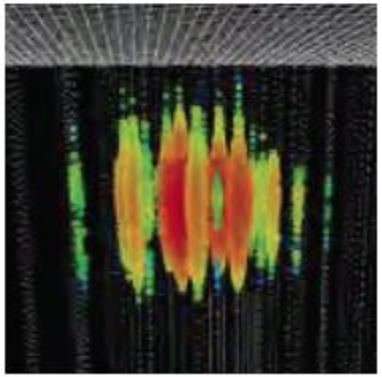
ALS

BES: Advanced Light Source
PI: Banda

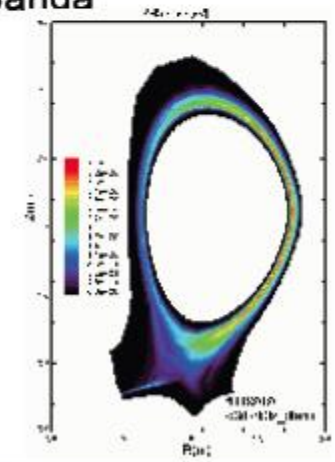


BER: Joint Genome Institute, Production Sequencing
PI: Ruben/Acting

HEP: Dark Energy Survey
PI: Habib



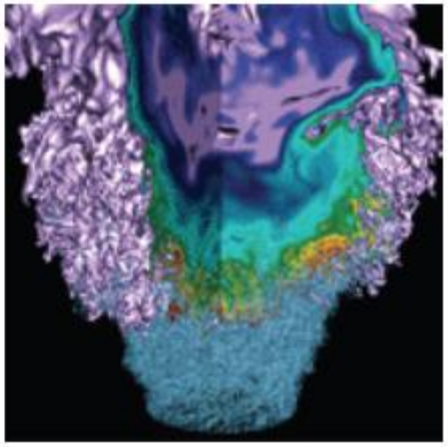
NP: Simulations and Analysis for IceCube
PI: Palczewski

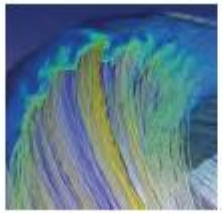


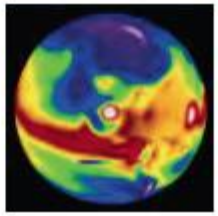
FES: LLNL MFE Supercomputing
PI: Maxim

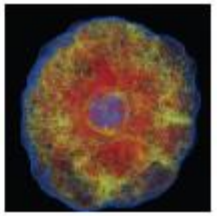
BER: Development of the LES ARM Symbiotic Simulation and Observation Workflow

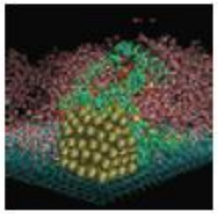
NERSC System Plans to Support Data Intensive Science





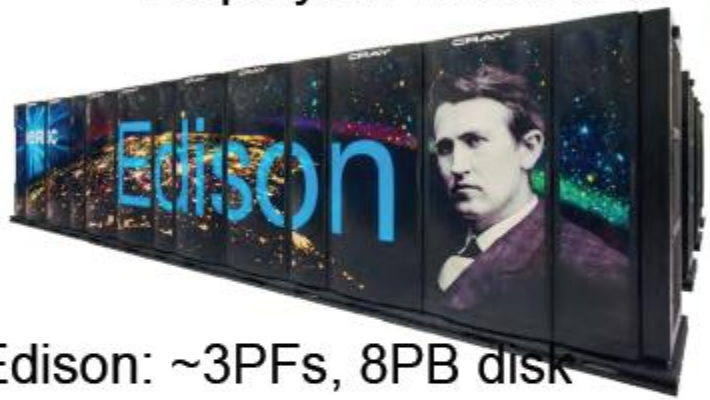






NERSC Resources at a Glance

Cori: 30PFs, 30PB disk
Deployed: 2015/2016



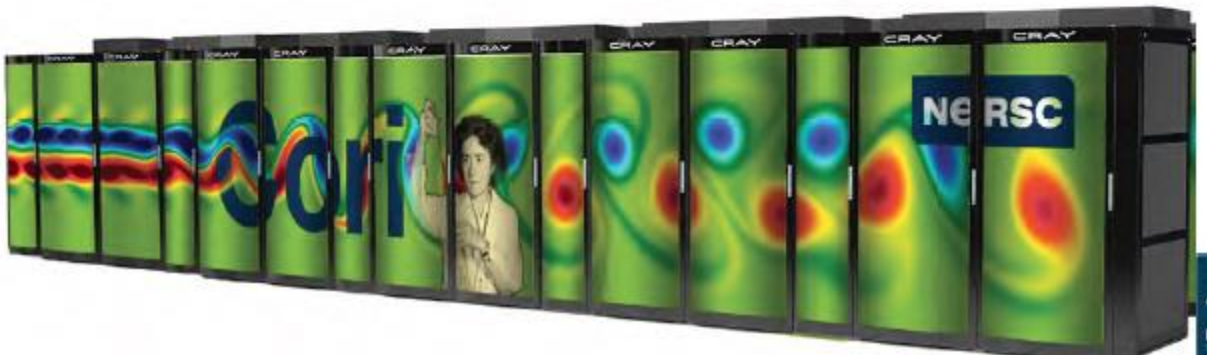
Edison: ~3PFs, 8PB disk
Deployed: 2013

NGF: 40TB/project and buy-in model

HPSS Archive: ~100 PBs

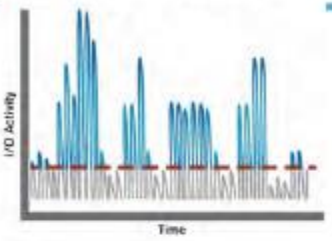
NERSC is making significant investments on Cori to support data intensive science

- **High bandwidth external connectivity to experimental facilities from compute nodes (Software Defined Networking)**
- **NVRAM Flash Burst Buffer as I/O accelerator**
  - 1.5PB, 1.5 TB/sec
  - User can request I/O bandwidth and capacity at job launch time
  - Use cases include, out-of-core simulations, image processing, shared library applications, heavy read/write I/O applications
- **Virtualization capabilities (Docker)**
- **More login nodes for managing advanced workflows**
- **Support for real time and high-throughput queues**



Data enhancements on Cori have addressed a number of user issues

I/O is too slow



Burst Buffer more than doubles available I/O bandwidth

It's difficult to bring complex software stacks to HPC systems

User defined images with Shifter

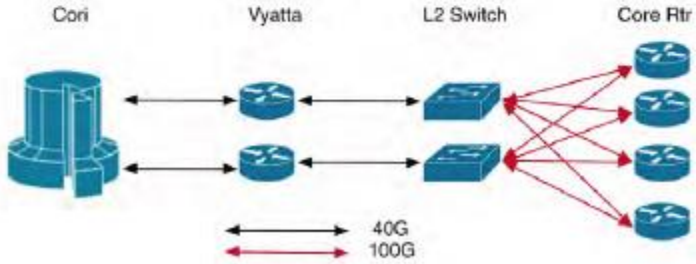
I need real-time feedback for my workflow

SchedMD

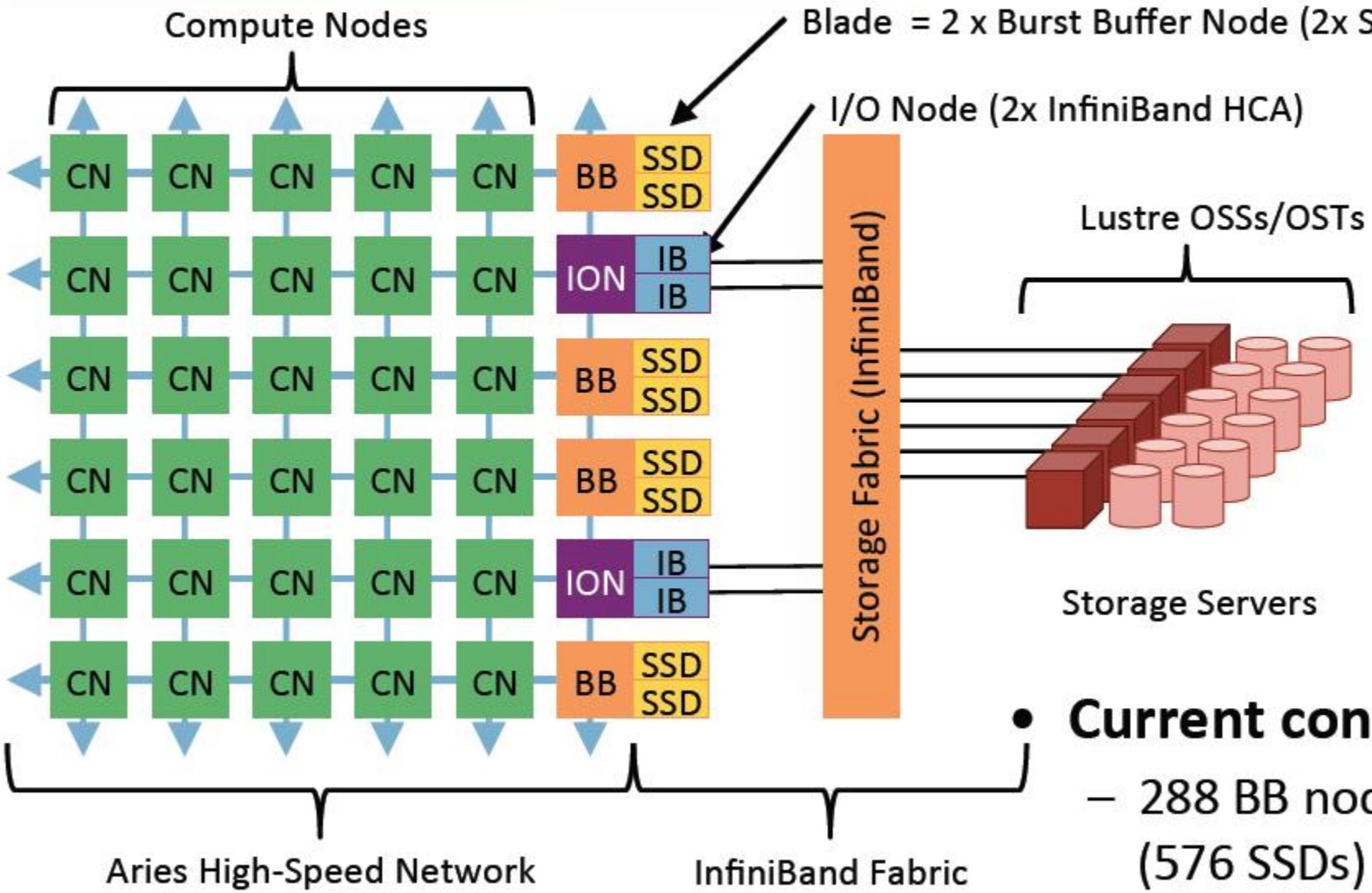
Real-time queues

Internal network limits how I can import data to supercomputer

SDN



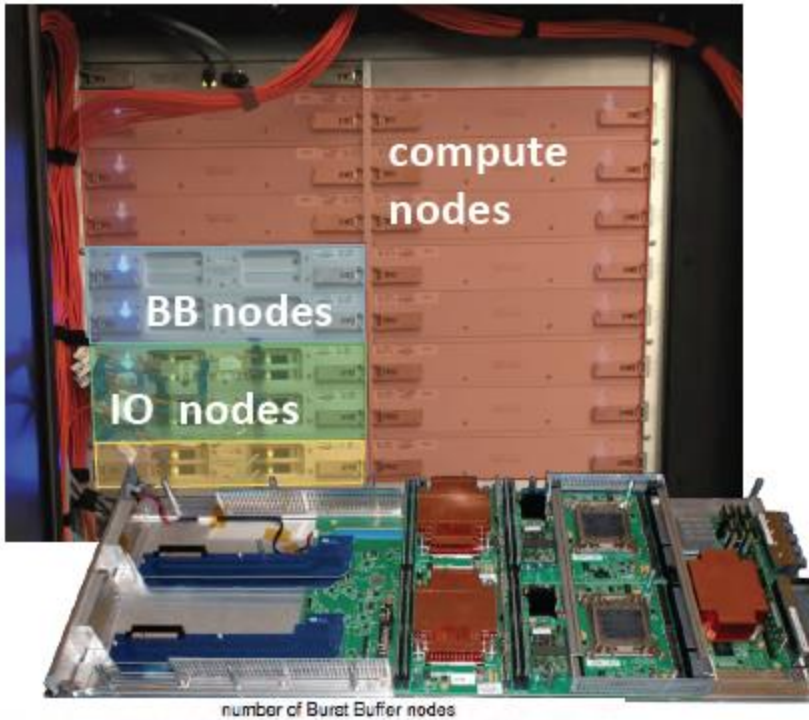
Burst Buffer Architecture

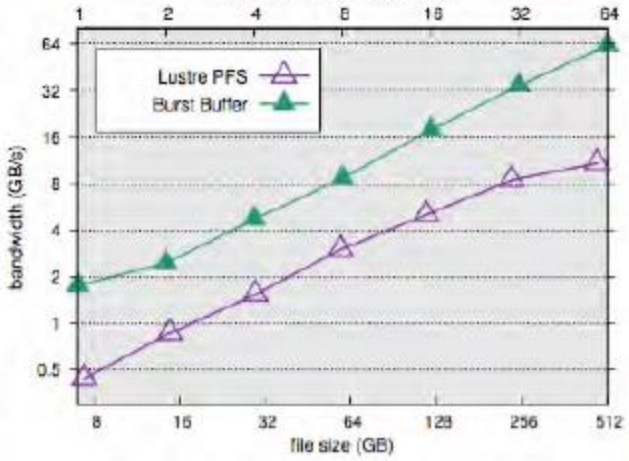


- **Current config**
  - 288 BB nodes (576 SSDs)

Burst Buffer is gaining momentum

- Many users are now seeing a 4-5x speed-up of their IO using the BB
- PHOENIX cosmology simulation code NESAP team: 5x speedup in entire code from BB.
- Initial tests of genomics reconstruction code sees 5-10x speedup in IO using the BB compared to Lustre
- Celeste Gordon Bell submission: using BB to stage 10M files (60TB) of astronomical image data for fast analysis





Shifter: Containers for HPC

NERSC

SHIFTER

docker

Enabling users to bring their own images to an HPC environment

The Register®
Biting the hand that feeds IT

DATA CENTER SOFTWARE NETWORKS SECURITY INFRASTRUCTURE DEVOPS BUSINESS HARDW

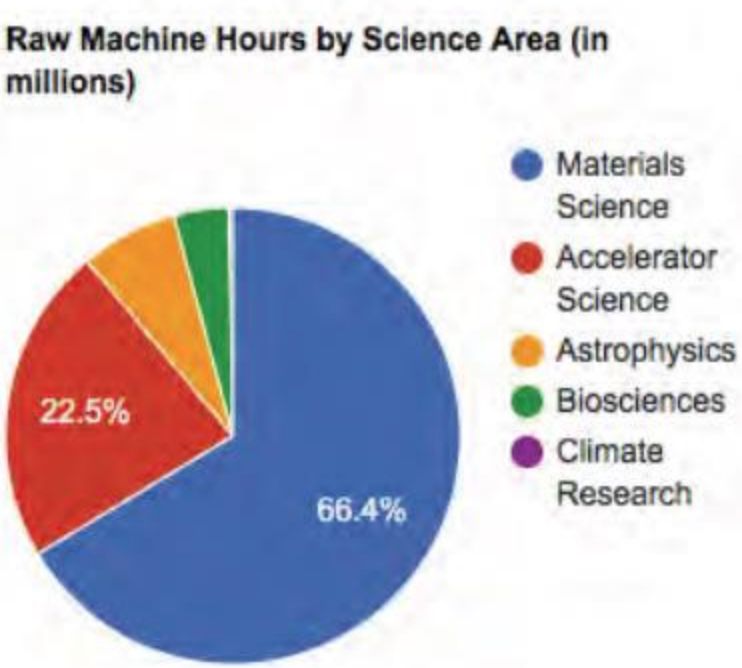
Data Center ▸ HPC

Cray hoists Docker containers into supercomputers

Productivity gains without performance hits

18 Nov 2015 at 00:01, Drew Cullen

Real-time queue makes inroads at NERSC



- Prototype queue used by a handful of projects at NERSC
- 32 nodes available for real-time queue
- Users apply to NERSC to get access
- Real-time queue accounts for <1% of time at NERSC
- NERSC is tracking usage and use cases closely

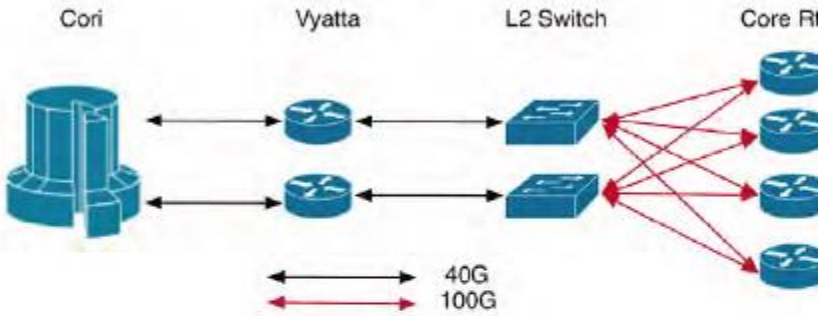
Enhanced Cori WAN Networking

Progress

- HW and SW installed and configured
- Simple outbound BW testing shows 4X improvement in bandwidth to compute nodes. RSIP 5.5 Gb/s, SDN 20Gb/s

Initial Science Uses Cases

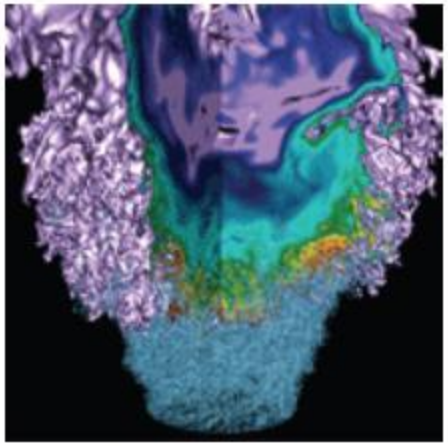
- General Atomics – 5x improvement talking to an external database used in a real-time workflow
- Globus-url-copy to CERN test point – 100x faster!
- LCLS to Cori BB now 100x faster!



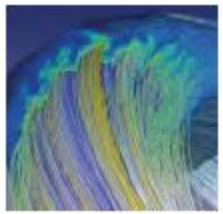
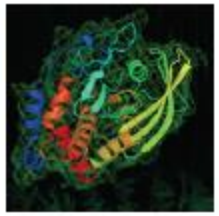
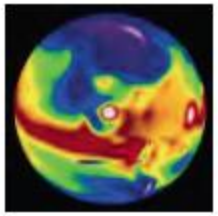
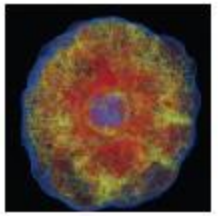
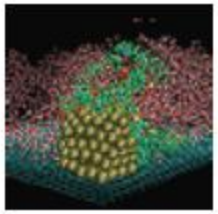
Next Steps

- Scale Testing 160 Nodes to 1 GW
- Multi-stream In-bound transfers
- Med Term: SLURM integration
- Long Term: OSCARS circuit testing and integration

User Support

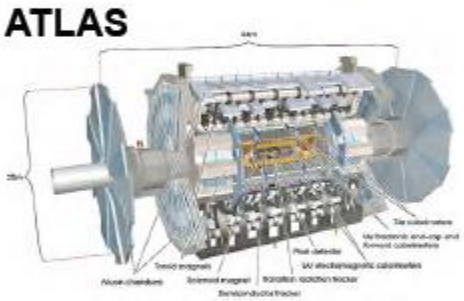


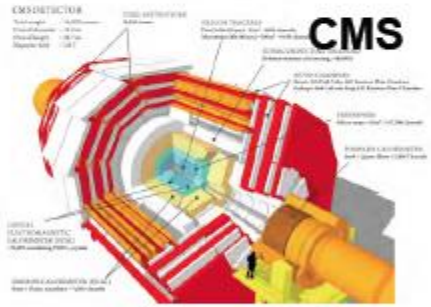
NERSC

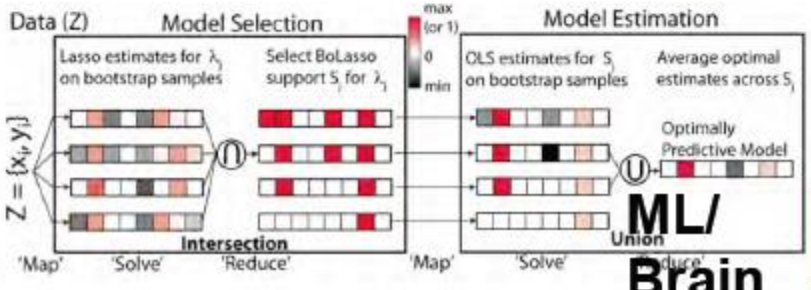






NERSC Exascale Science Application Program (NESAP) for Data

- **Applications that analyze data from experiments and instrumentation also need help preparing for exascale**
- **Teams get access to vendor expertise and NERSC liaison.**
- **Call for proposals resulted in 6 selected teams**
- **NESAP postdocs:**
  - 1 postdoc hired at NERSC.
  - Interviewing for 2 more now.
  - Code teams gathering initial performance data on KNL now.





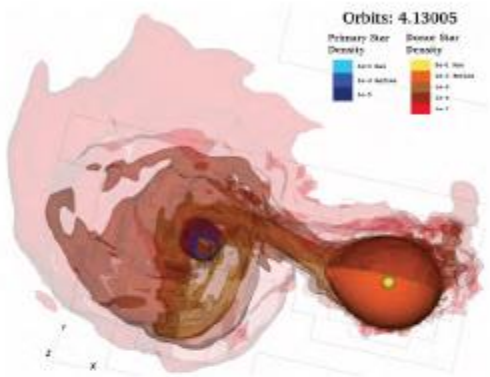


U.S. DEPARTMENT OF ENERGY | Office of Science

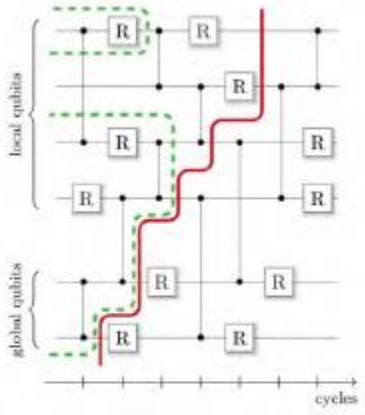
BERKELEY LAB

Large Scale SC submissions on Cori

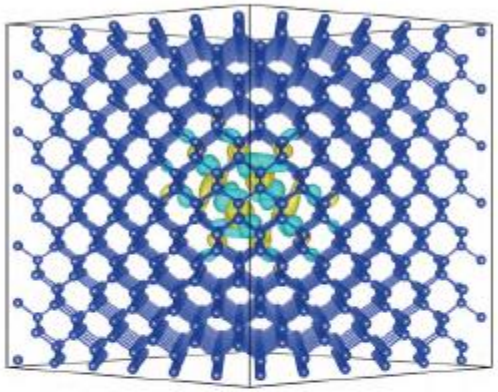
Large Scale Particle in Cell Plasma Simulations



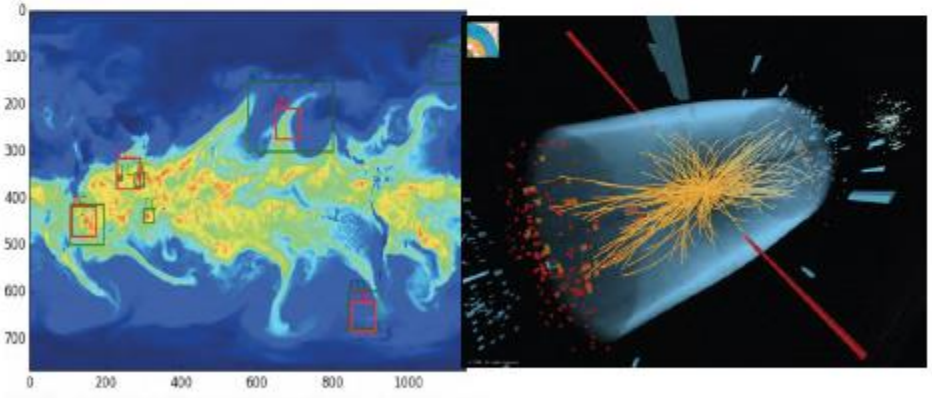
Stellar Merger Simulations with Task Based Programming



Largest Ever Quantum Circuit Simulation

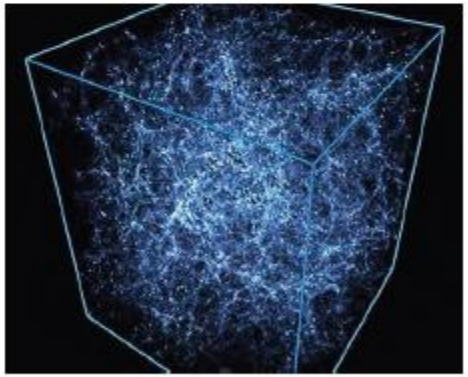


Largest Ever Defect Calculation from Many Body Perturbation Theory > 10PF



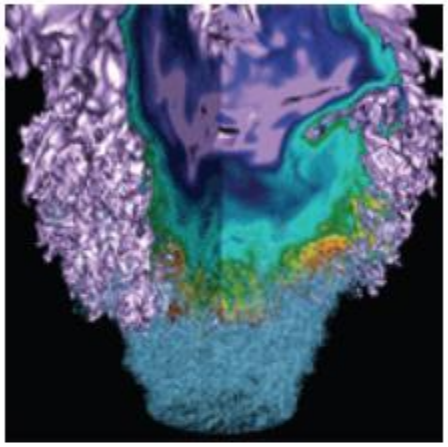
Deep Learning at 15PF (SP) for Climate and HEP

Celeste: 1st Julia application to achieve 1 PF

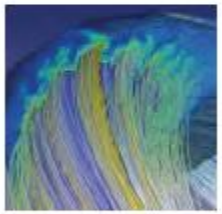


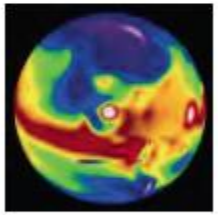
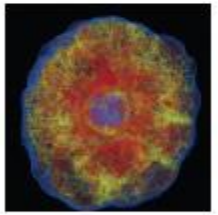
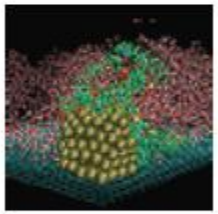
Galactos: Solved 3-pt correlation analysis for Cosmology @9.8PF

Deep Learning in Science?



NERSC



Can Deep Learning Work for Science?

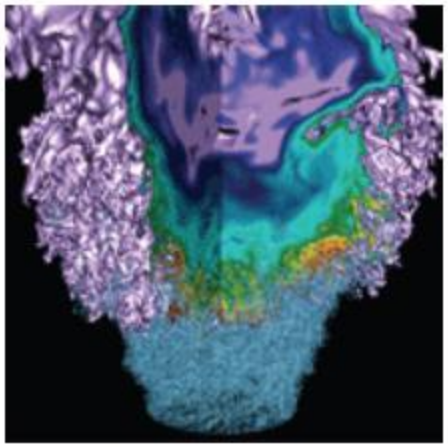
- **Similarities**
  - Tasks:
    - Pattern Classification
    - Regression
    - Clustering
    - Feature Learning
    - Anomaly Detection

- **Differences**
  - Unique attributes of Scientific Data
    - Multi-channel / Multi-variate
    - Double precision floating point
    - Noise and Artifacts
    - Statistics are likely different

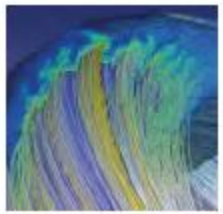
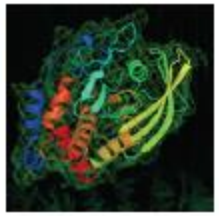
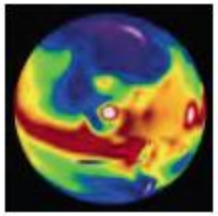
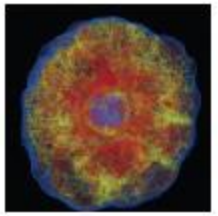
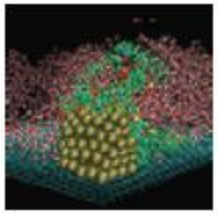
NERSC Supports Analytics and Machine Learning Libraries on Cori

- Machine learning libraries are installed and supported by NERSC staff
- Most packages were developed for single core or cloud environments and scalability on HPC systems is still a challenge
- NERSC is partnering with Intel and Cray to optimize and scale performance on HPC systems

Looking towards the future



NERSC

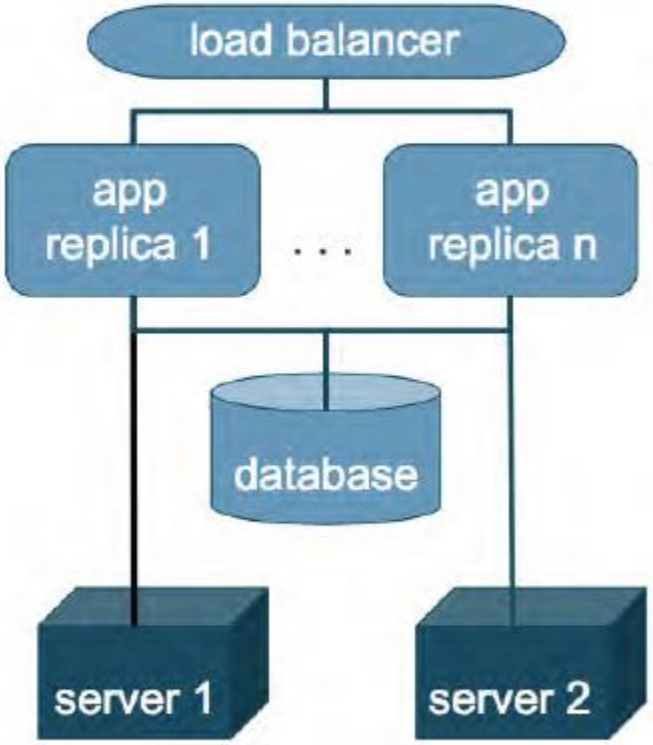
Future Data Requirements

- Scientists require support for analysis tools, many of which differ significantly from traditional simulation.
- Analyses and drawing inferences on big data sets are revolutionizing many fields. New approaches are needed for analyzing these datasets including advanced statistics and machine learning.
- Workflows in both simulation and analysis are becoming more complex, and need to be accommodated on HPC systems. Complexity often relates to data movement over wide area networks.
- Scientists at experimental facilities want to use HPC to help guide experiments in real time, which requires co-scheduling between ASCR facilities and facilities from other DOE offices.

SPIN Platform: Edge services to support data intensive science

Enable new approaches to scientific research with flexible, scalable, on-demand resources tightly integrated into the center resources

- Users require 'edge' services to support complex workflows, databases, science gateways, workflow managers
- Draw a clean line between user's role (developing a service) and center's role (providing infrastructure),
- Make it simpler to develop, deploy and scale services, but dropping in containers



Goals and Objectives for the NERSC-9 Project – Delivery in 2020

1. Provide a significant increase in computational capabilities over the Edison system on a set of representative DOE benchmarks.
2. Meet the needs of extreme computing and data users by accelerating workflow performance.
3. Provide a vehicle for the demonstration and development of exascale-era technologies
4. Delivery in the 2020 time frame

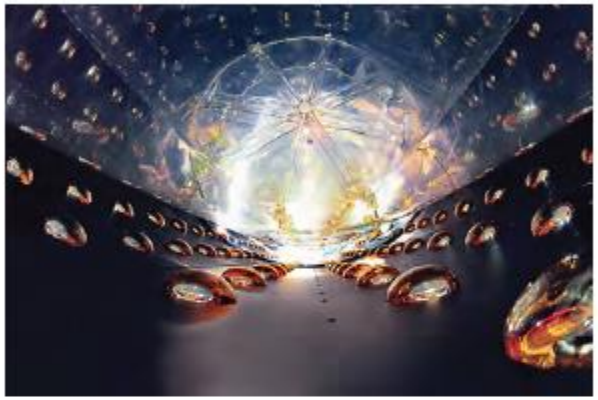
In Summary: We are making good progress but pain points remain

- **Authentication/trust/identity management between experimental facilities and NERSC**
- **Scalable analytics software**
- **Seamless data science workflows which include data transfer capabilities, supercomputer, databases, gateways and archiving**
- **Supporting diverse workflows, few common tools**
- **Rolling upgrades and system outages**
  - Considering redundancy between sites
- **Interactivity and queue turn around times for experimental facilities**

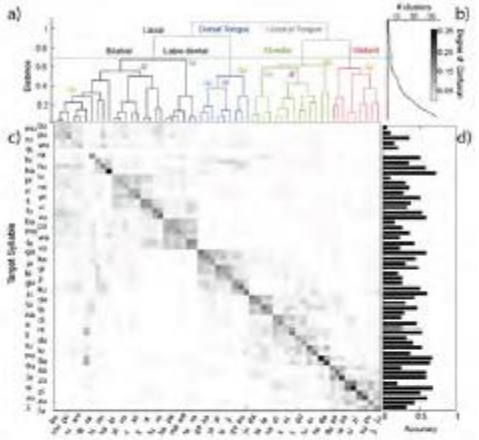
National Energy Research Scientific Computing Center

Deep Learning for Science

Modeling galaxy shapes

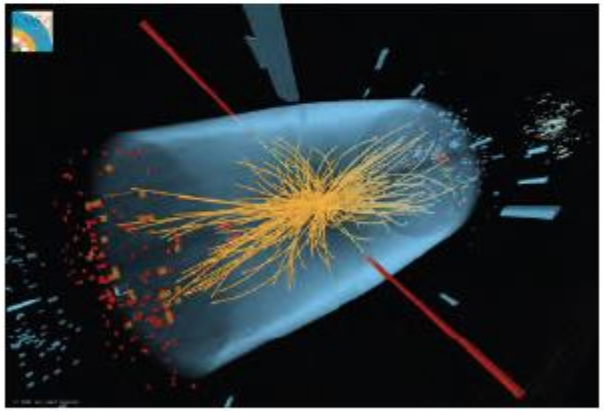


Clustering Daya Bay events

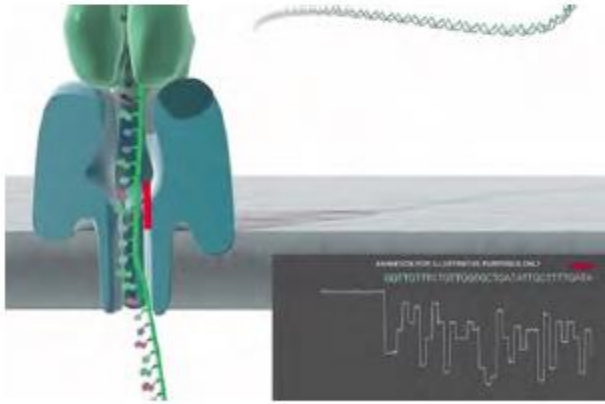


Decoding speech from ECoG

Detecting extreme weather



Classifying LHC events



Oxford Nanopore sequencing