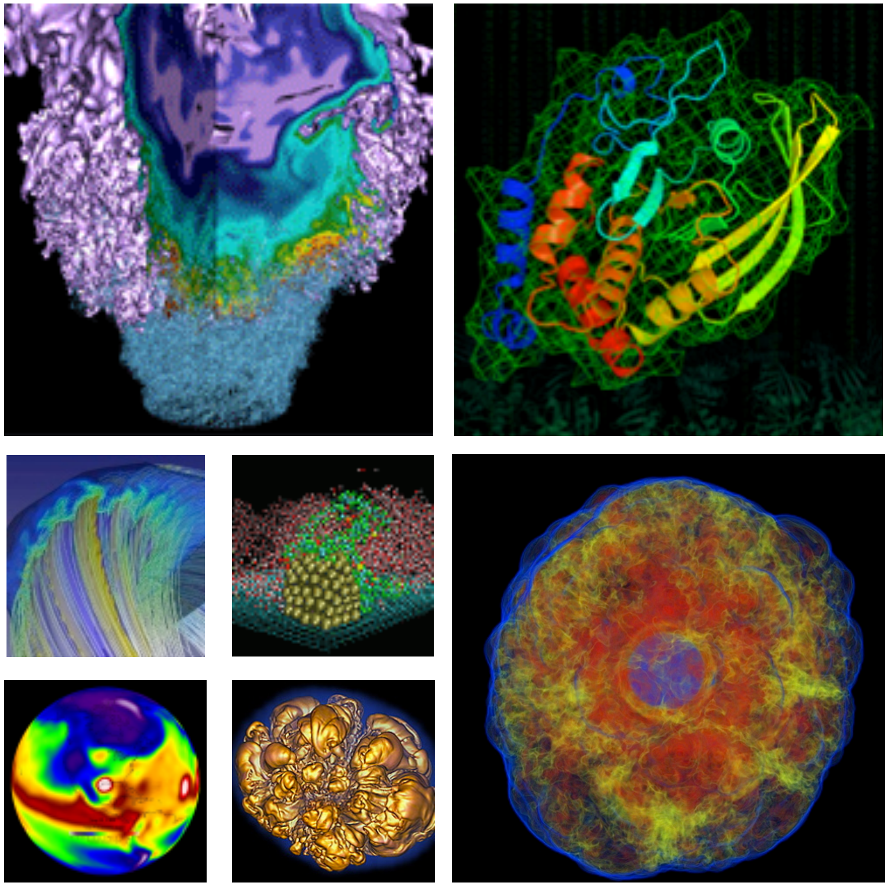
Supporting Data Intensive Workloads at NERSC



May 4, 2017
Katie Antypas
Data Department Head

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

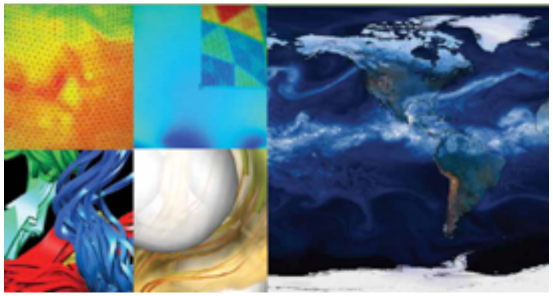
- NERSC deploys advanced HPC and data systems for the broad Office of Science community

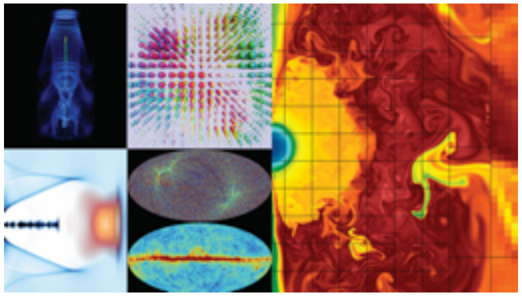
- NERSC staff provide advanced application and system performance expertise to users

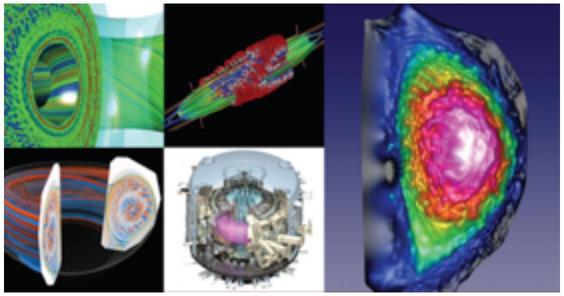
- Approximately 6000 users and 750 projects

- 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.***

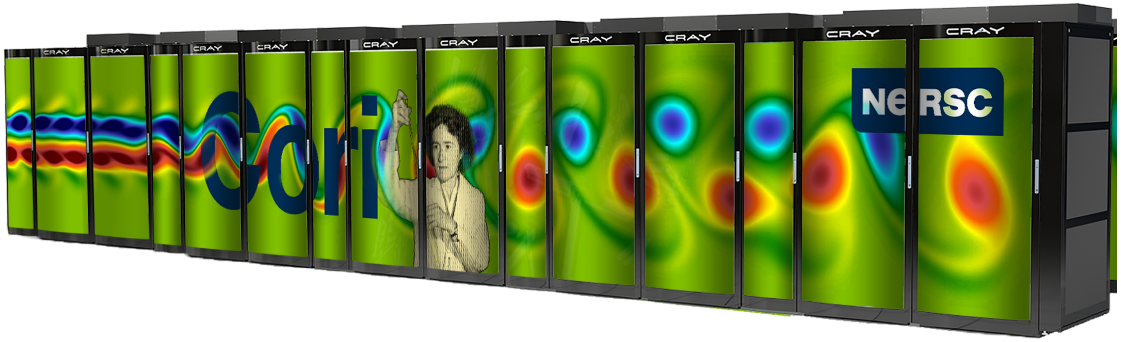




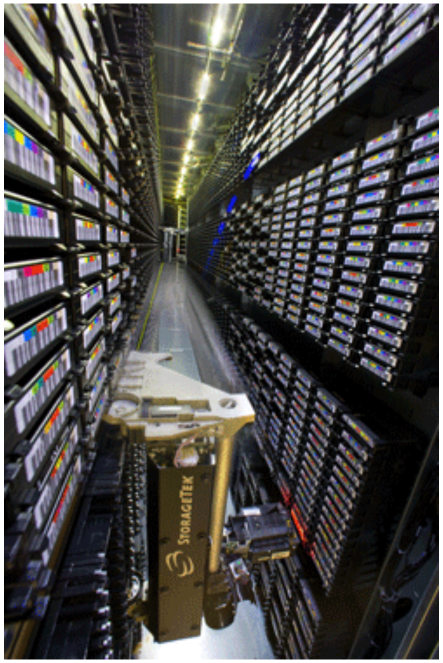


NERSC Resources at a Glance

Cori: 30PFs, 30PB disk



Edison: ~3PFs, 8PB disk

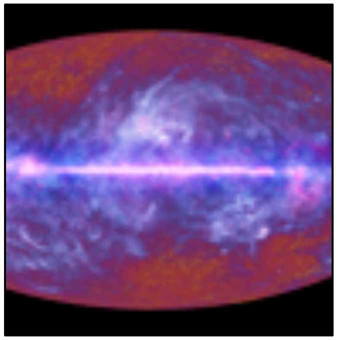


HPSS Archive: ~100 PBs

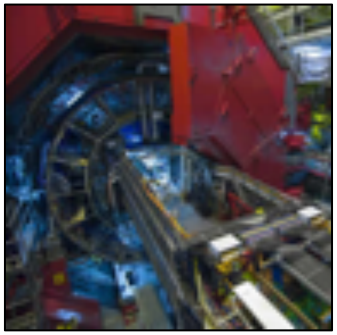
NGF: 40TB/project and buy-in model

NERSC has been supporting data intensive science for a long time

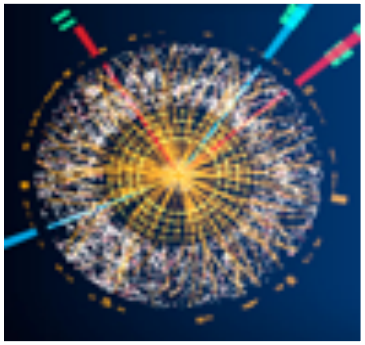
Palomar Transient
Factory
Supernova



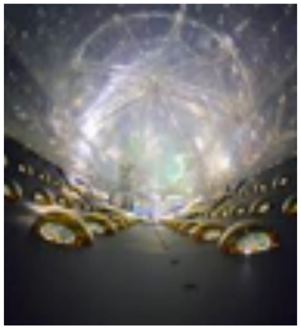
Planck Satellite
Cosmic Microwave
Background
Radiation



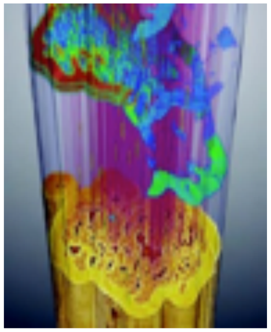
Alice
Large Hadron
Collider



Atlas
Large Hadron
Collider



Dayabay
Neutrinos

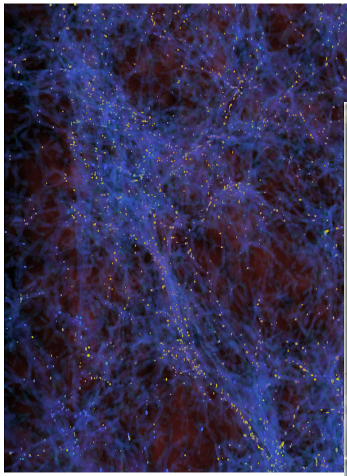


ALS
Light Source

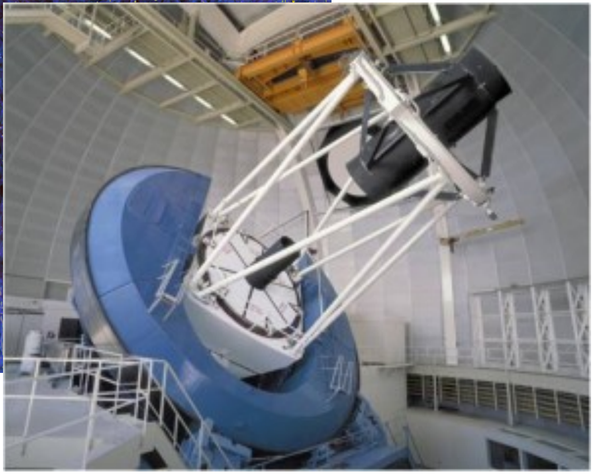
LCLS
Light Source

Joint Genome
Institute
Bioinformatics

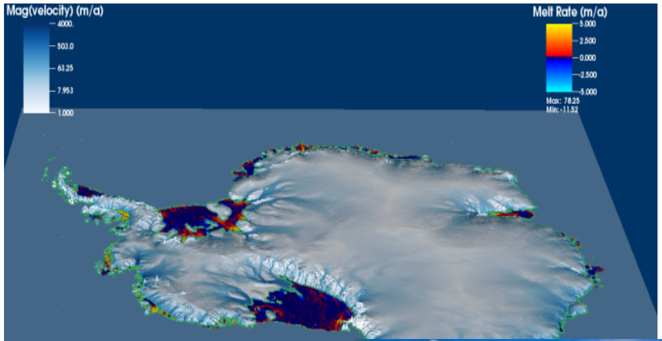
What has changed? Coupling of experiments with large scale simulations



Nyx simulation of Lyman alpha forest



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

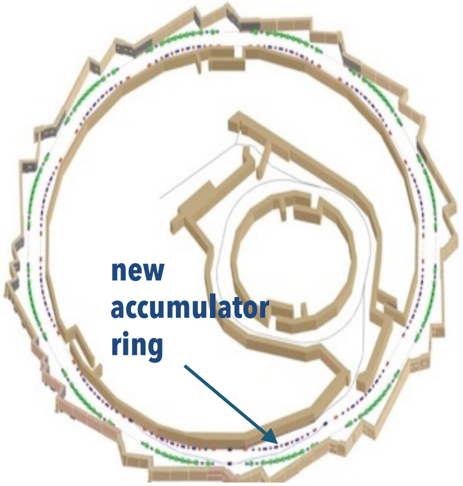


New climate modeling methods, produce new understanding of ice

Genomes to watersheds

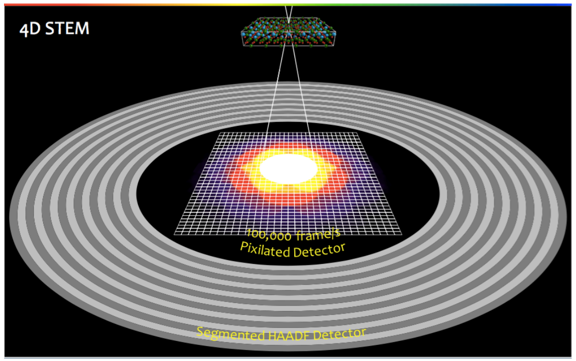
What has changed? Increased data rates and new sensing capabilities

LCLS
Light Source



Advanced Lightsource Upgrade

Environmental sensors



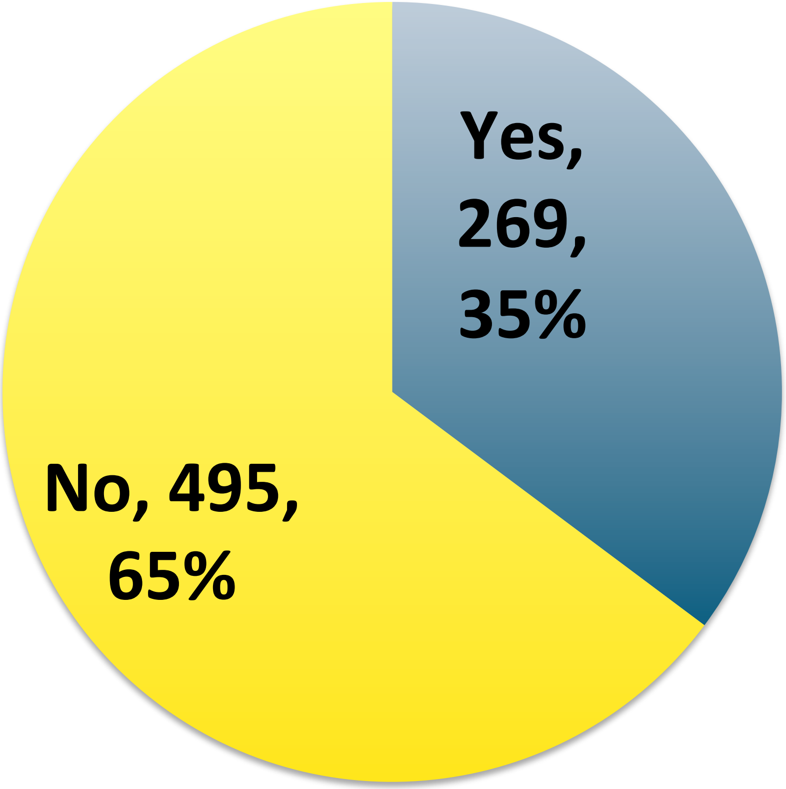
Next generation electron microscope

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 ERCAP Question This Year

- **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?

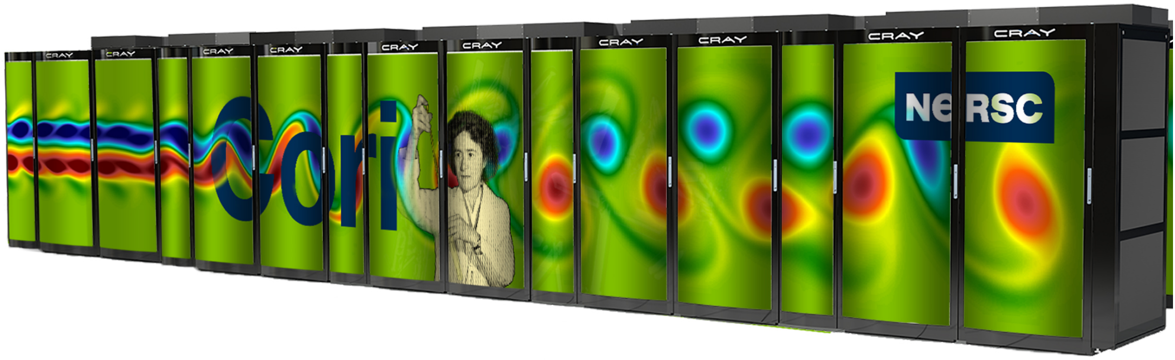


764 projects

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

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

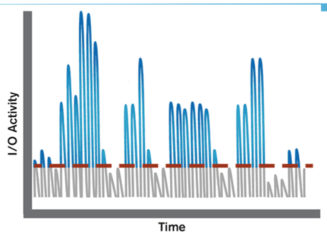
- **High bandwidth external connectivity to experimental facilities from compute nodes (Software Defined Networking)**
- **NVRAM Flash Burst Buffer as I/O accelerator**
  - 1.5PB, 1.5 TB/sec
  - User can request I/O bandwidth and capacity at job launch time
  - Use cases include, out-of-core simulations, 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

NERSC

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

SHIFTER

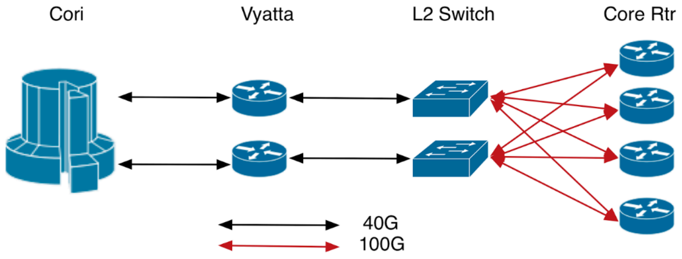
I need real-time feedback for my workflow

SchedMD

Real-time queues

Internal network limits how I can import data to supercomputer

SDN



There is limited software for analytics on HPC systems

New analytics and ML libraries

Caffe

scikit learn

Spark

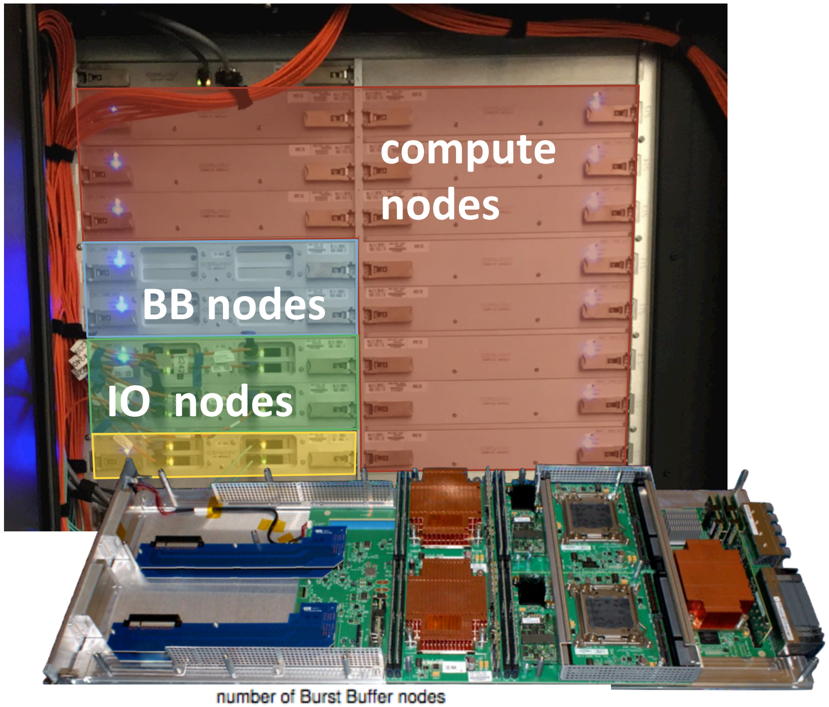
K

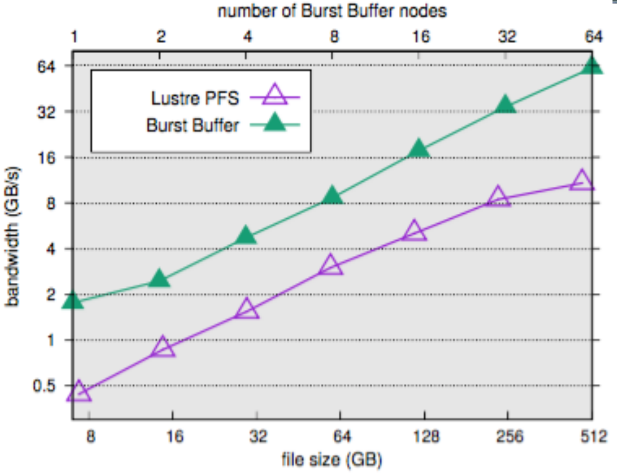
TensorFlow

U.S. DEPARTMENT OF ENERGY | Office of Science

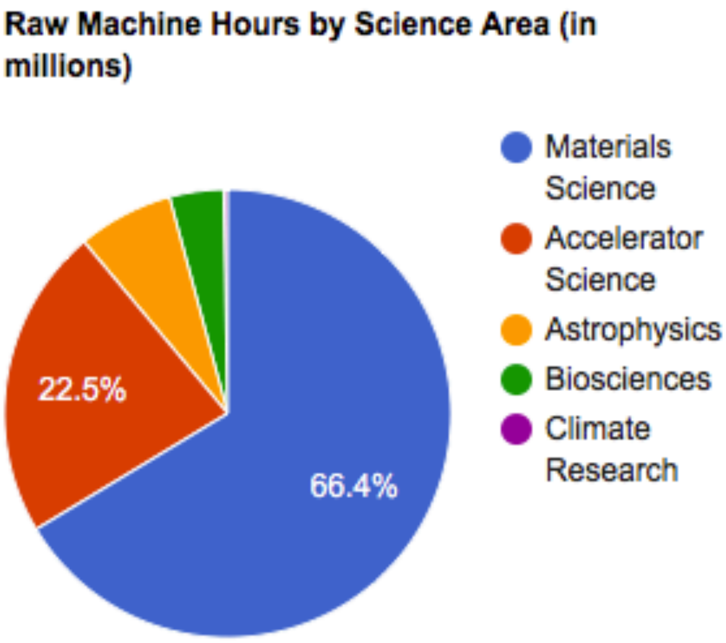
Burst Buffer is gaining momentum

- 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





Real-time queue makes inroads at NERSC

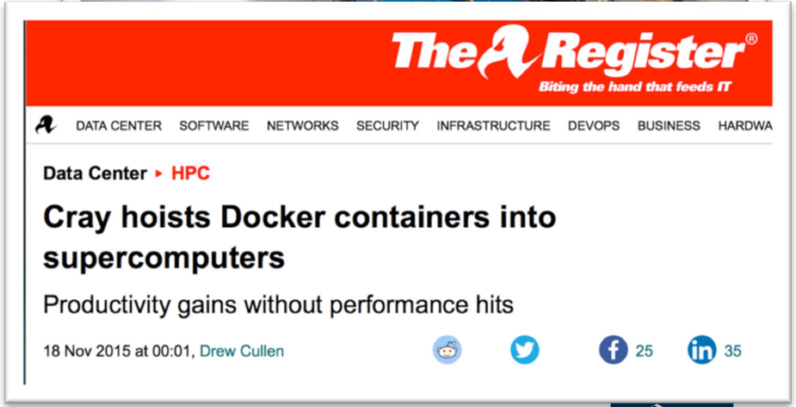


Raw Machine Hours by Science Area (in millions)

- 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

Shifter: Containers for HPC

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 HARDWA

Data Center ▸ HPC

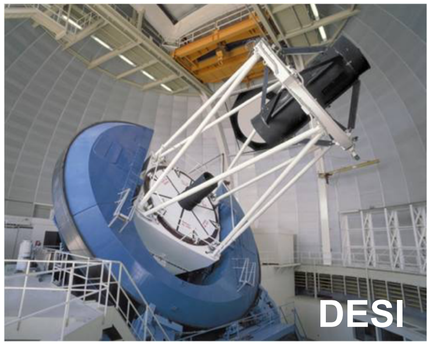
Cray hoists Docker containers into supercomputers

Productivity gains without performance hits

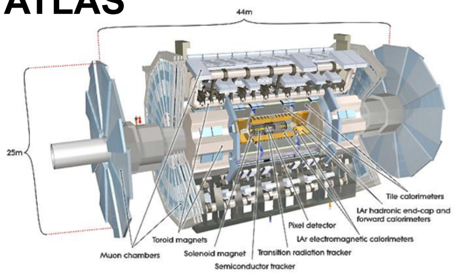
18 Nov 2015 at 00:01, Drew Cullen

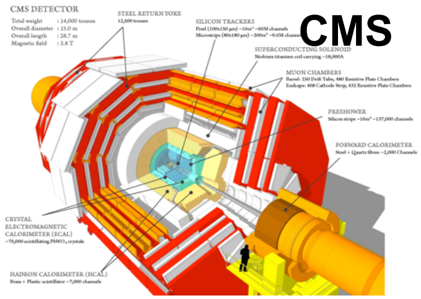
NERSC Exascale Science Application Program (NESAP) for Data

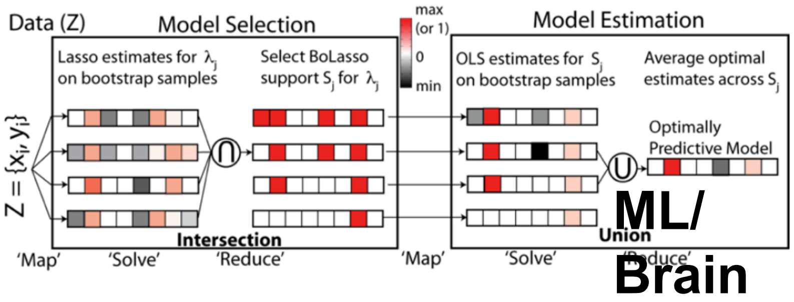
- **Applications that analyze data from experiments and instrumentation also need help preparing for exascale**
- **Teams get access to vendor expertise and NERSC liaison.**
- **Proposal process for code teams:**
  - Call for proposals in October.
  - 6 selections in December (pictured)
- **NESAP postdocs:**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



ATLAS


44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

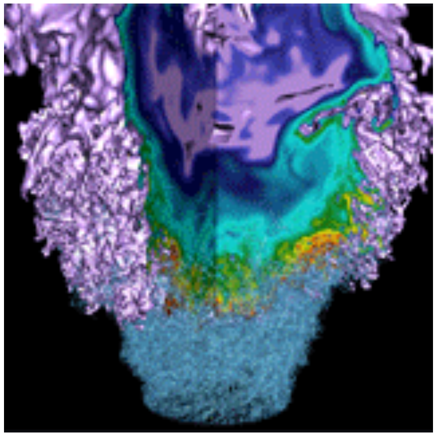

CMS
CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

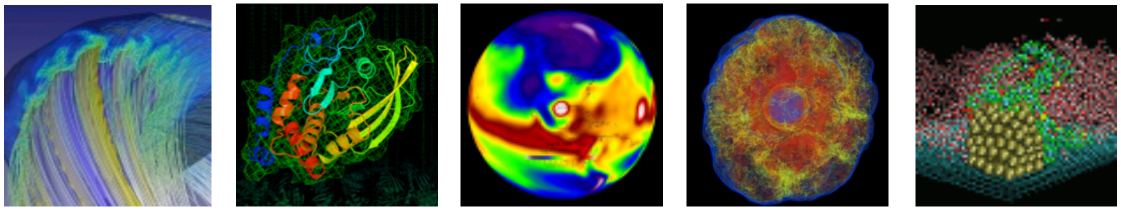

Data (Z)
Model Selection
max
(or 1)
0
min
Model Estimation
Lasso estimates for λj on bootstrap samples
Select BoLasso support Sj for λj
OLS estimates for Sj on bootstrap samples
Average optimal estimates across Sj
Optimally Predictive Model
Z = {xi, yi}
∩
Intersection
∪
Union
'Map'
'Solve'
'Reduce'
'Map'
'Solve'
'Reduce'
ML/
Brain

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**
- **Rolling upgrades and system outages**
  - Considering redundancy between sites
- **Interactivity and queue turn around times for experimental facilities**
- **Supporting diverse workflows, few common tools**

Looking towards the future





New Data Requirements from Users

- Informal feedback from users, the annual NERSC user survey, and more formal DOE requirements reviews describe similar data requirements.
- While there is some variation across SC Offices, the requirements are surprisingly similar.
- From NP report in executive summary:

New hardware and software tools are needed to analyze, track, and manage data produced by experiments and in simulations, including developments in databases, and to move data effciently between sites for appropriate analysis.

Requirements Reviews: Machine Learning and Analytics Software

- **Improved tools needed for machine learning and deep learning, which are now are a part of analysis (pattern recognition, anomaly detection, (BES)**
- **Community would benefit from development of better algorithms (such as Machine Learning methods) and data-processing tools for lossless real-time data reduction near the beam line. (NP)**
- **New techniques for data analysis are urgently needed to address overwhelming data volumes and streams from both experiments and simulations (HEP)**
- **New approaches to interpreting large data sets are needed and may include neural networks, image segmentation and other ML approaches. (BER)**

Requirements Reviews: High Bandwidth Networking

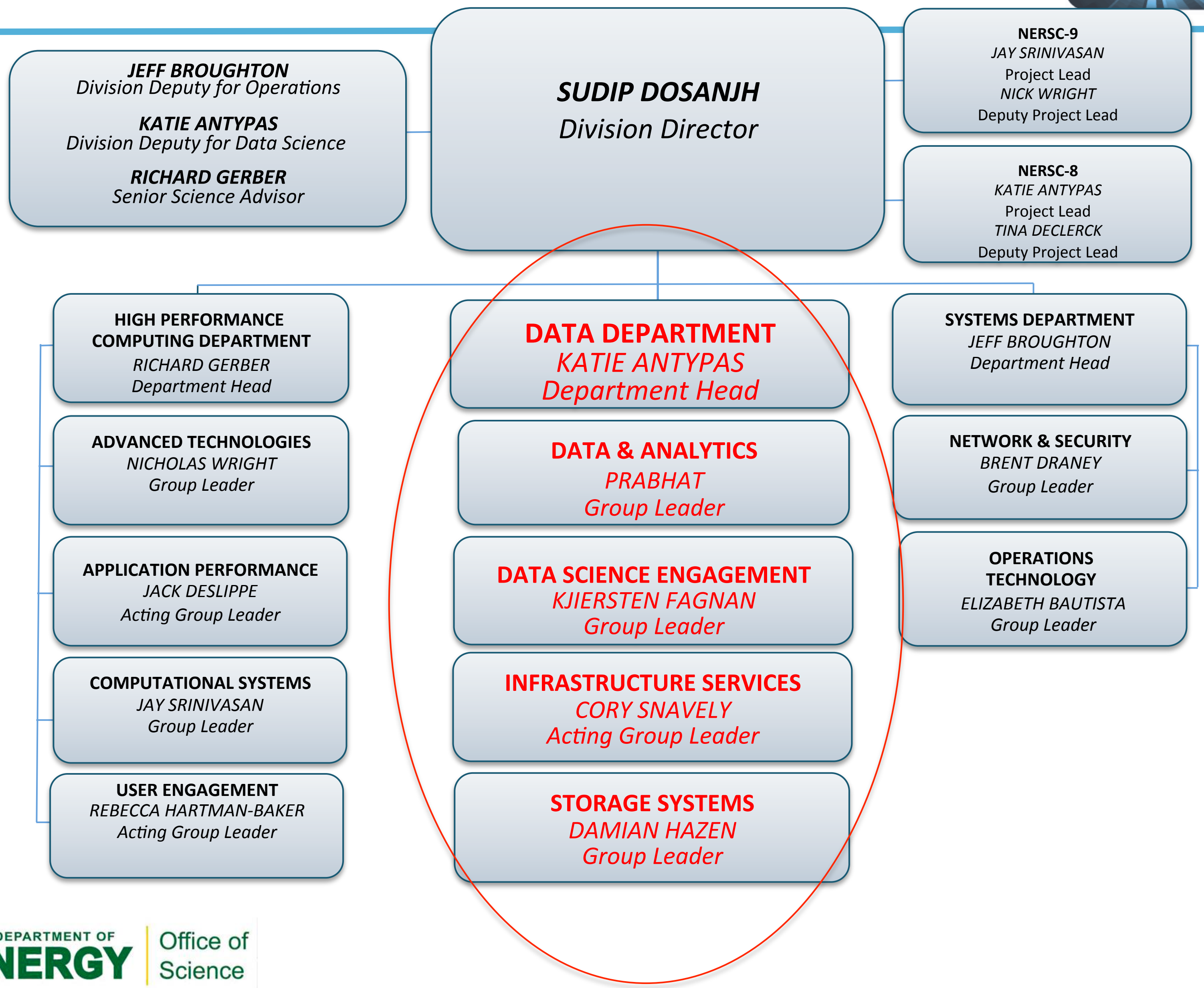
- Having access to high I/O bandwidth to stream data into an HPC system from some external measurement device or the local storage system will also be essential (BER)
- On-demand, high-performance networking will be required to enable this inter-facility operation. (BES)
- Treating networks as a resource that needs to be managed and planned for is an important area of future ASCR and HEP interaction. (HEP)
- A streaming readout system requires a combination of HPC and storage coupled to the detector by a low-latency, high-bandwidth network. (NP)

Requirements Reviews: 'Real-time' and fast turnaround computing

- Efficient and effective use of BES facilities requires real-time access to ASCR HPC facility-class resources to support streaming analysis and visualization to guide experimental decisions. (BES)
- Software development and performance tuning can be highly interactive processes, incorporating rapid prototyping; policies that enable rapid evaluation and ability to rapidly acquire interactive resources can significantly improve productivity in these communities (ASCR)
- The experimental program would benefit from real-time access to the advanced computing capabilities of ASCR and NSF (NP)
- Increasingly, BER community is depending on facilities that generate huge amounts of data, sometimes continuously and in real time. (BER)
- Scheduling tools and policies for optimized usage of computers… providing queues that enable sufficiently quick turn-around time for model development and test purposes, will improve researcher efficiency. (BER)
- Energy Frontier applications require real-time remote access to resources while the associated jobs are running. (HEP)

National Energy Research Scientific Computing Center (NERSC)

NERSC



SUDIP DOSANJH
Division Director

JEFF BROUGHTON
Division Deputy for Operations

KATIE ANTYPAS
Division Deputy for Data Science

RICHARD GERBER
Senior Science Advisor

NERSC-9
JAY SRINIVASAN
Project Lead
NICK WRIGHT
Deputy Project Lead

NERSC-8
KATIE ANTYPAS
Project Lead
TINA DECLERCK
Deputy Project Lead

HIGH PERFORMANCE COMPUTING DEPARTMENT
RICHARD GERBER
Department Head

ADVANCED TECHNOLOGIES
NICHOLAS WRIGHT
Group Leader

APPLICATION PERFORMANCE
JACK DESLIPPE
Acting Group Leader

COMPUTATIONAL SYSTEMS
JAY SRINIVASAN
Group Leader

USER ENGAGEMENT
REBECCA HARTMAN-BAKER
Acting Group Leader

DATA DEPARTMENT
KATIE ANTYPAS
Department Head

DATA & ANALYTICS
PRABHAT
Group Leader

DATA SCIENCE ENGAGEMENT
KJIERSTEN FAGNAN
Group Leader

INFRASTRUCTURE SERVICES
CORY SNAVELY
Acting Group Leader

STORAGE SYSTEMS
DAMIAN HAZEN
Group Leader

SYSTEMS DEPARTMENT
JEFF BROUGHTON
Department Head

NETWORK & SECURITY
BRENT DRANEY
Group Leader

OPERATIONS TECHNOLOGY
ELIZABETH BAUTISTA
Group Leader

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

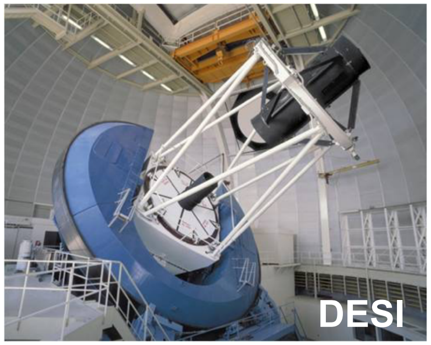
07.20.16

National Energy Research Scientific Computing Center

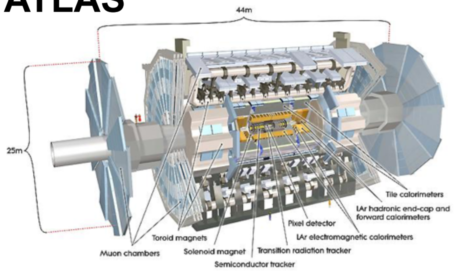
NESAP for Data

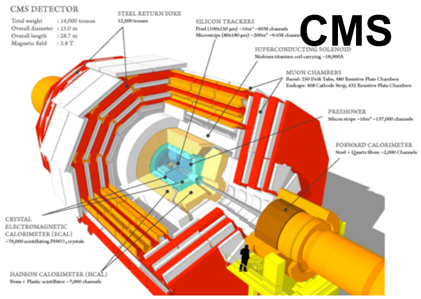
- **Applications that analyze data from experiments and instrumentation also need help preparing for exascale**
- **Teams get access to vendor expertise and NERSC liaison.**
- **Proposal process for code teams:**
  - Call for proposals in October.
  - 6 selections in December (pictured)
- **NESAP postdocs:**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

ALS

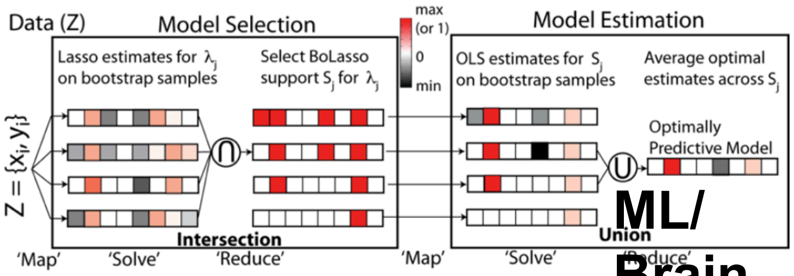

DESI

ATLAS


44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

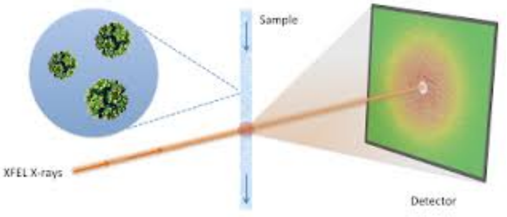

CMS
CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

CMB

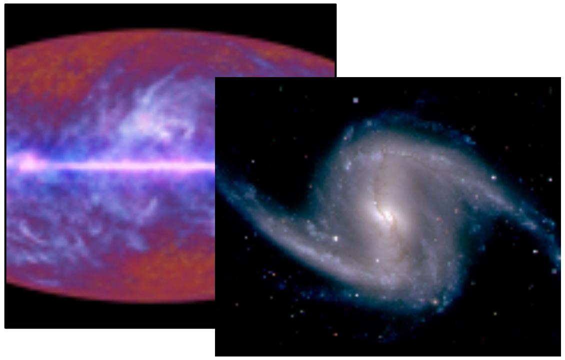

Data (Z)
Model Selection
Model Estimation
max (or 1)
0
min
Lasso estimates for λj on bootstrap samples
Select BoLasso support Sj for λj
OLS estimates for Sj on bootstrap samples
Average optimal estimates across Sj
Optimally Predictive Model
Z = {xi, yi}
Intersection
Union
'Map'
'Solve'
'Reduce'
'Map'
'Solve'
'Reduce'

ML/ Brain

Some exemplars

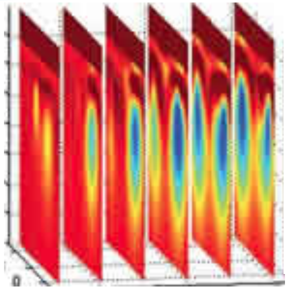


ASCR: Algorithms for next generation light sources
PI: Sethian



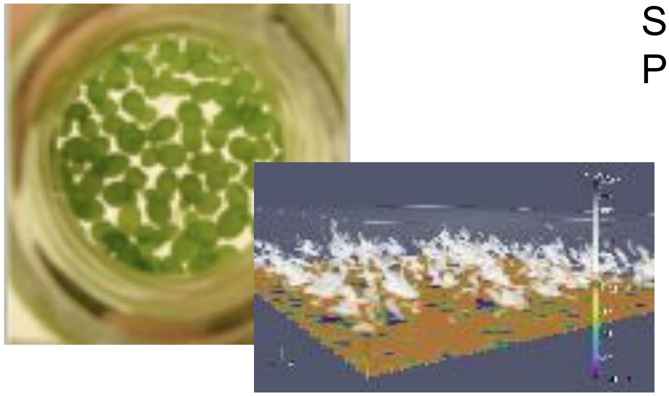
HEP: CMB Data Analysis for Planck Satellite
PI: Borrill

HEP: Dark Energy Survey
PI: Habib



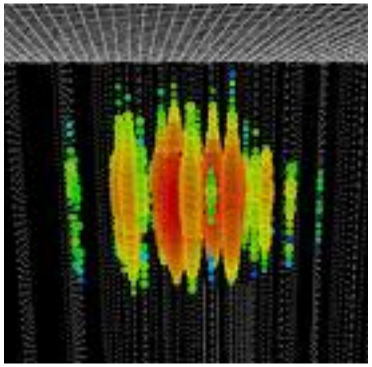
BES: Large Scale 3D Geophysical Inversion & Imaging
PI: Newman

BES: Advanced Light Source
PI: Banda

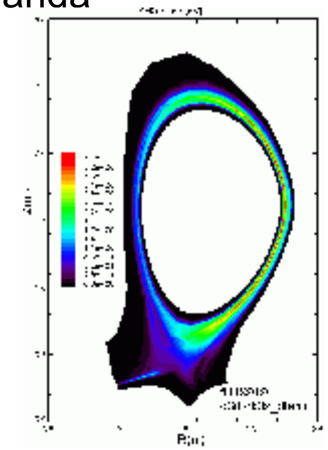


BER: Joint Genome Institute, Production Sequencing
PI: Ruben/Acting

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



NP: Simulations and Analysis for IceCube
PI: Palczewski



FES: LLNL MFE Supercomputing
PI: Maxim

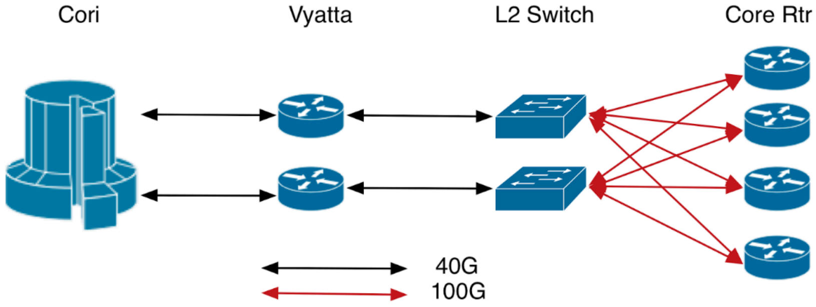
Enhanced Cori WAN Networking

Progress

- HW and SW installed and configured
- Phase 1 (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