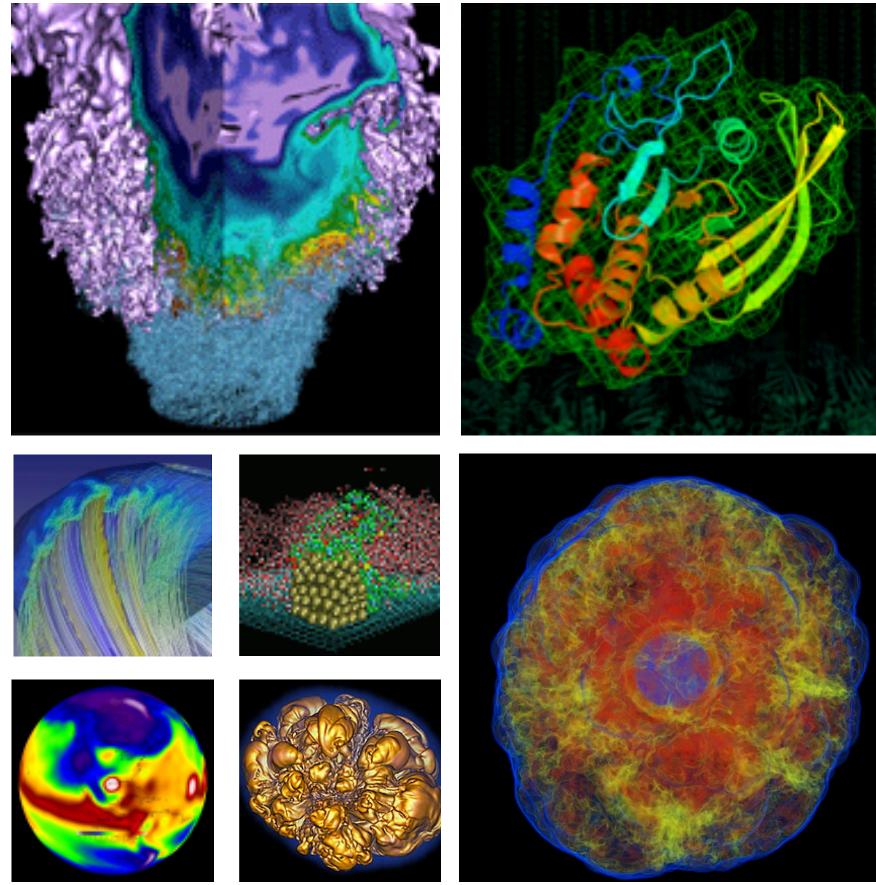
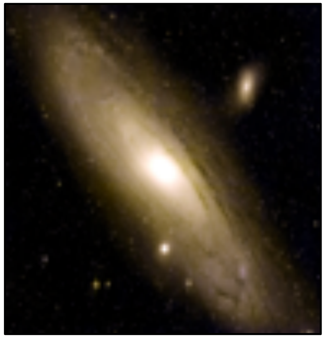
The Convergence of HPC and Data Science at NERSC

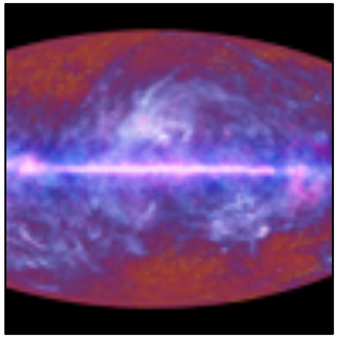


Katie Antypas
NERSC Deputy for Data Science

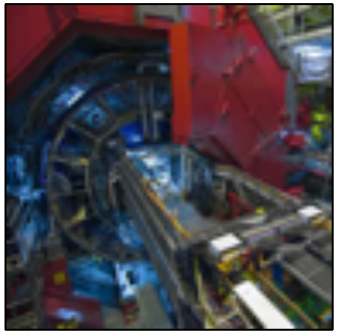
NERSC has been supporting data intensive science for a long time



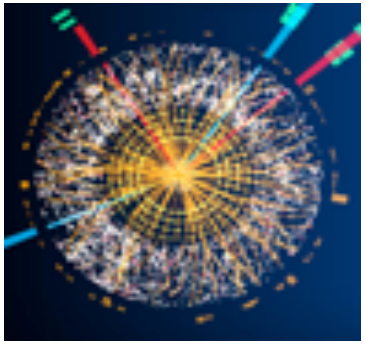
Palomar Transient Factory Supernova



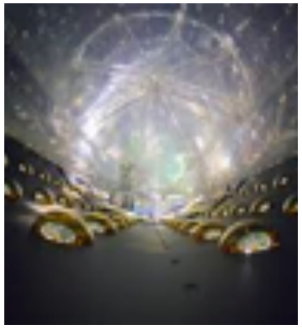
Planck Satellite Cosmic Microwave Background Radiation



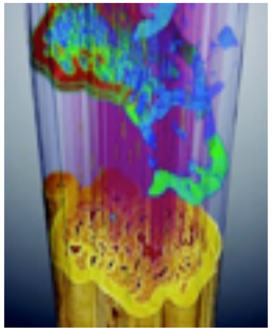
Alice
Large Hadron Collider



Atlas
Large Hadron Collider



Dayabay
Neutrinos



ALS
Light Source

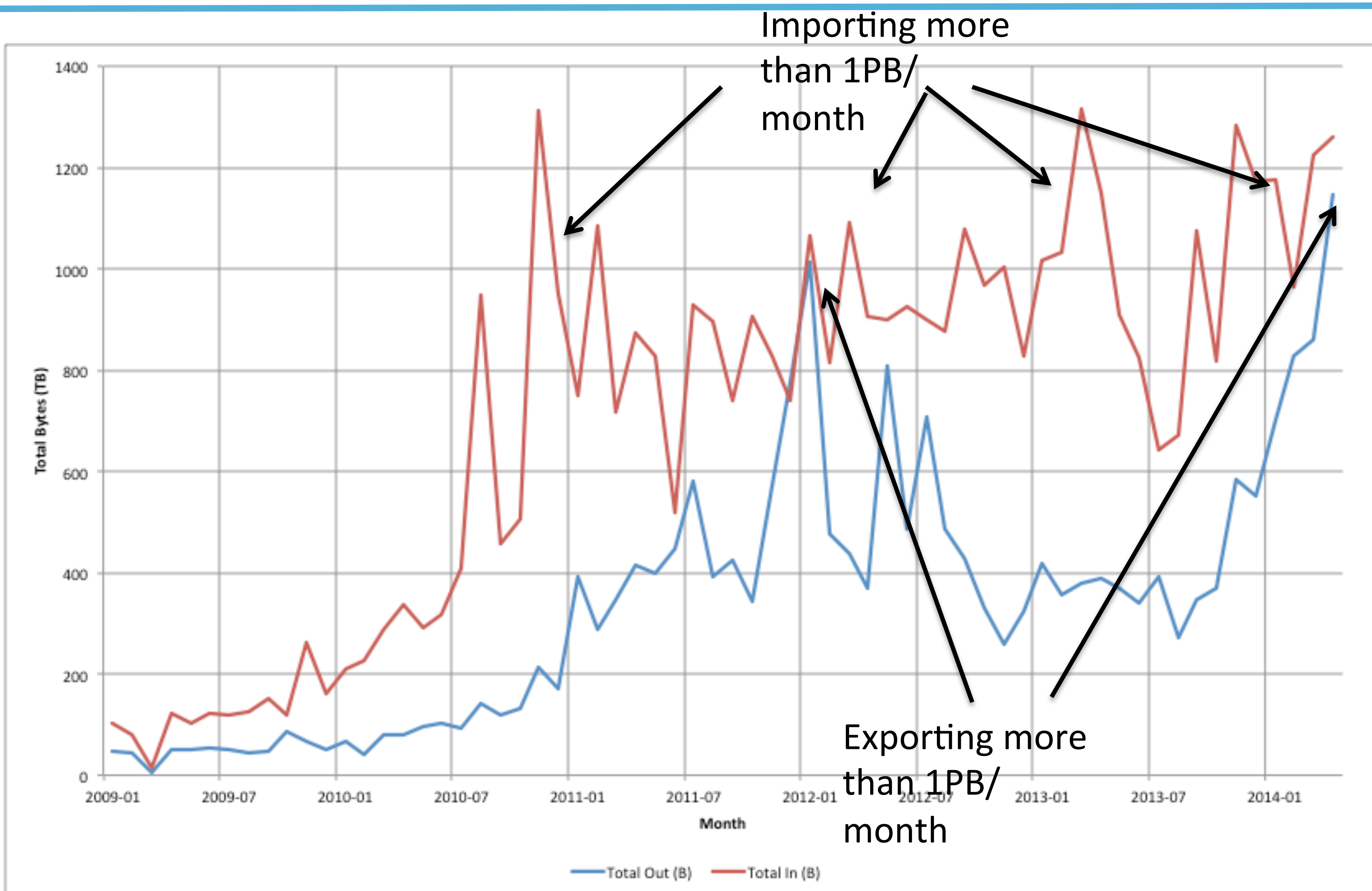
LCLS
Light Source

Joint Genome Institute Bioinformatics

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NERSC users import more data than they export!



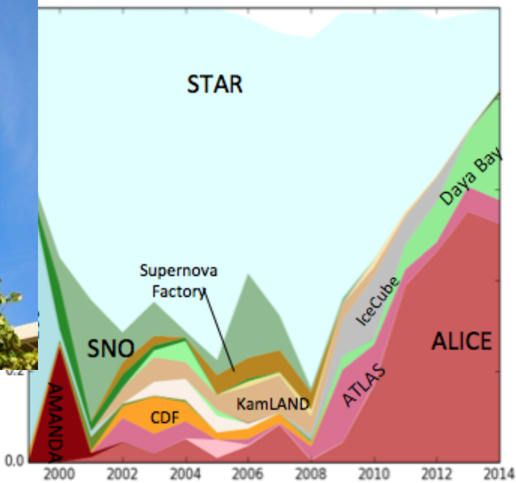
Historically NERSC has deployed separate Compute Intensive and Data Intensive Systems

Compute Intensive

Data Intensive

Carver

Genepool



PDSF

Need for Change

- Dramatically growing data sets require Petascale+ computing for analysis
- We increasingly need to couple large-scale simulations and data analysis

But how different really are the compute and data intensive platforms?

Policies

- Fast-turn around time. Jobs start shortly after submitted
- Can run large numbers of throughput jobs

Software/Configuration

- Support for complex workflows
- Communication and streaming data from external databases and data sources
- Easy to customize user environment

Hardware

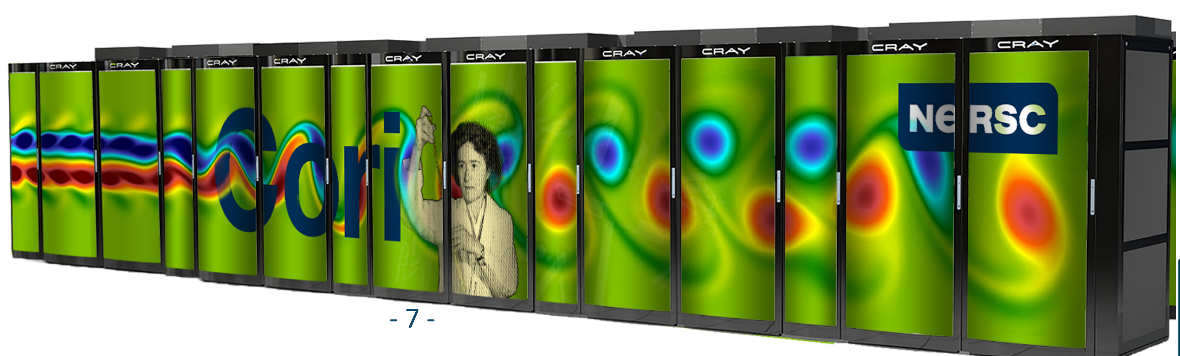
- Local disk for fast I/O
- Some systems (not all) have larger memory nodes
- Support for advanced workflows (DB, web, etc)

Differences are primarily software and policy issues with some hardware differences in the ratio of I/O, memory and compute

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

NERSC

- **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**



NERSC Organization

- **SUDIP DOSANJH** — Division Director
  - **JEFF BROUGHTON** — Division Deputy for Operations
  - **KATIE ANTYPAS** — Division Deputy for Data Science
  - **JOHN SHALF** — Chief Technology Officer

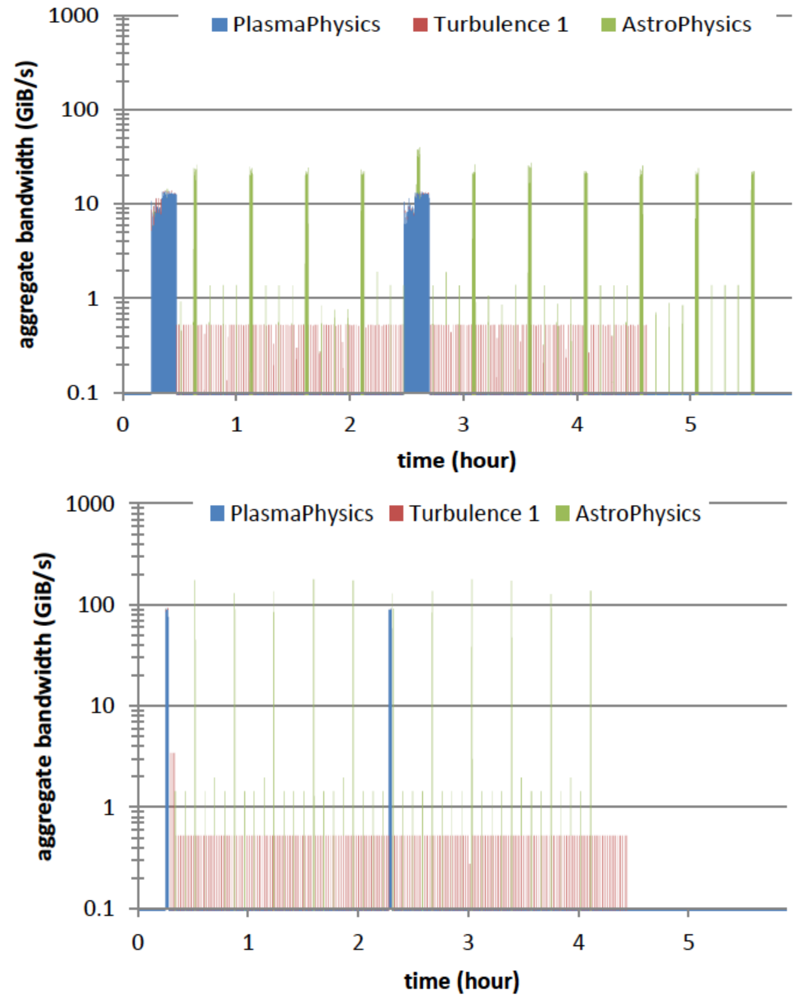
- HPC DEPARTMENT — **RICHARD GERBER**, Acting Department Head
  - ADVANCED TECHNOLOGIES — **NICK WRIGHT**, Group Leader
  - APPLICATION PERFORMANCE — **JACK DESLIPPE**, Acting Group Leader
  - COMPUTATIONAL SYSTEMS — **JAY SRINIVASAN**, Group Leader
  - USER ENGAGEMENT — **REBECCA HARTMAN-BAKER**, Acting Group Leader

- SCIENTIFIC COMPUTING & DATA SERVICES 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**, Acting Group Leader

- SYSTEMS DEPARTMENT — **JEFF BROUGHTON**, Department Head
  - NETWORK & SECURITY — **BRENT DRANEY**, Group Leader
  - OPERATIONS TECHNOLOGY GROUP — **ELIZABETH BAUTISTA**, Group Leader

Burst Buffer Motivation

- **Flash storage is significantly more cost effective at providing bandwidth than disk (up to 6x)**
- **Flash storage has better random access characteristics than disk, which help many SC workloads**
- **Users' biggest request (complaint) after wanting more cycles, is for better I/O performance**

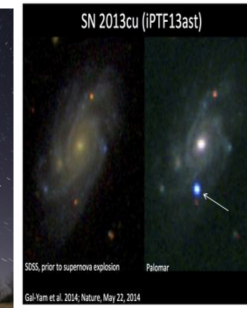


Application perceived I/O rates, with no burst buffer (top), burst buffer (bottom).

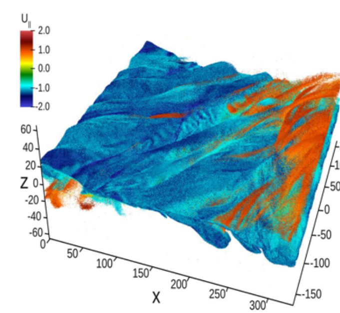
Analysis from Chris Carothers (RPI) and Rob Ross (ANL)

NERSC is exploring Burst Buffer Use Cases beyond checkpoint-restart

- **Accelerate I/O**
  - Checkpoint/restart or other high bandwidth reads/writes
  - Apps with high IOP/s e.g. non-sequential table lookup
  - Out-of-core applications
  - Fast reads for image analysis
- **Advanced Workflows**
  - Coupling applications, using the Burst Buffer as interim storage
  - Streaming data from experimental facilities
- **Analysis and Visualization**
  - In-situ/ in-transit
  - Interactive visualization

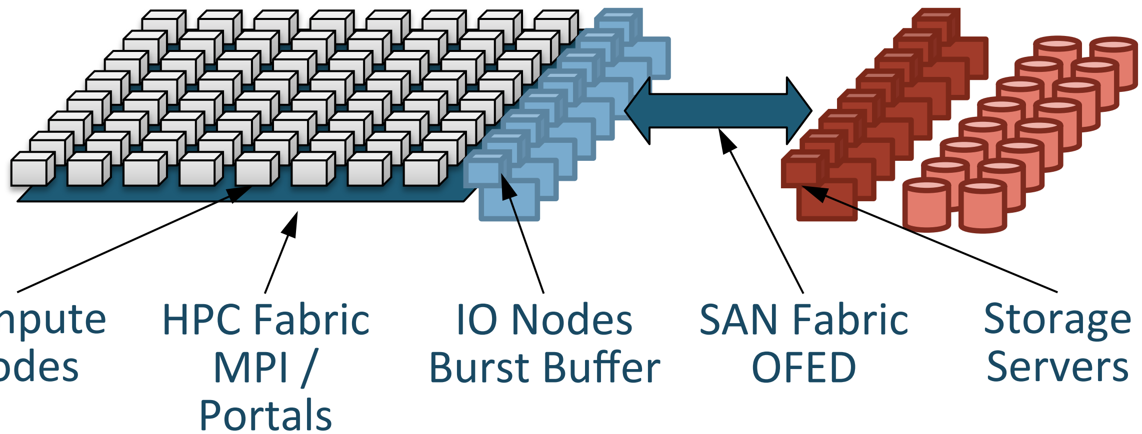


Palomar Transient Factory Pipeline:
Use Burst Buffer as cache for fast reads



VPIC – in situ visualization of a trillion particles

Burst Buffer Software Development Efforts



Create Software to enhance usability and to meet the needs of all NERSC users

- Scheduler enhancements
  - Automatic migration of data to/from flash
  - Dedicated provisioning of flash resources
  - Persistent reservations of flash storage
- Caching mode – data transparently captured by the BB nodes
  - Transparent to user -> no code modifications required
- Enable In-transit analysis
  - Data processing or filtering on the BB nodes – model for exascale

Burst Buffer Early User Program

- **Aug 10th: solicited proposals for BB Early Users program.**
  - Award of exclusive early use of BB on Cori P1, plus help of NERSC experts to optimise application for BB.
- **Selection criteria include:**
  - Scientific merit.
  - Computational challenges.
  - Cover range of BB data features.
  - Cover range of DoE Science Offices.
- **Great interest from the community, 29 proposals received. Good distribution across offices...**

NERSC supported projects

Project	DoE office	BB data features
Nyx/Boxlib cosmology simulations (*Ann Almgren, LBNL*)	HEP	I/O bandwidth with BB; checkpointing; workflow application coupling; in-situ analysis.
Phoenix: 3D atmosphere simulator for supernovae (*Eddie Baron, U. Oklahoma*)	HEP	I/O bandwidth with BB; staging intermediate files; workflow application coupling; checkpointing.
Chombo-Crunch + Visit for carbon sequestration (*David Trebotich, LBNL*)	BES	I/O bandwidth with BB; in-situ analysis/ visualization using BB; workflow application coupling.
Sigma/UniFam/Sipros Bioinformatics codes (*Chongle Pan, ORNL*)	BER	Staging intermediate files; high IOPs; checkpointing; fast reads.
XGC1 for plasma simulation (*Scott Klasky, ORNL*)	Fusion	I/O bandwidth with BB; intermediate file I/O; checkpointing.
PSANA for LCLS (*Amadeo Perazzo, SLAC*)	BES/BER	Staging data with BB; workflow management; in-transit analysis.

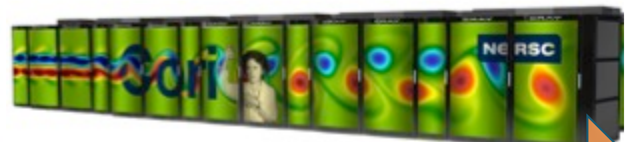
NERSC supported projects

Project	DoE office	BB data features
ALICE data analysis (*Jeff Porter, LBNL*)	NP	I/O bandwidth with BB; read-intensive I/O.
Tractor: cosmological data analysis (DESI) (*Peter Nugent, LBNL*)	HEP	Intermediate file I/O using BB; high IOPs.
VPIC-IO performance (*Suren Byna, LBNL*)	HEP/ACSR	I/O bandwidth with BB; in-situ data analysis; BB to stage data.
YODA: Geant4 sims for ATLAS detector (*Vakhtang Tsulaia, LBNL*)	HEP	BB for high IOPs; stage small intermediate files.
ALS SPOT Suite (*Craig Tull, LBNL*)	BES/BER	BB as fast cache; workflow management; visualization.
TomoPy for ALS image reconstruction (*Craig Tull, LBNL*)	BES/BER	I/O throughput with BB; workflow management; read-intensive I/O.
kitware: VPIC/Catalyst/ParaView (*Berk Geveci, kitware*)	ASCR	in-situ analysis/visualization with BB; multi-stage workflow.

A variety of use cases are represented by the Burst Buffer Early Users

Application	I/O bandwidth : reads	I/O bandwidth: writes (checkpointing)	High IOPs	Workflow coupling	In-situ / in-transit analysis and visualization	Staging inter-mediate files/ pre-loading data
Nyx/Boxlib		X		X	X	
Phoenix 3D		X		X		X
Chomo/Crunch + Visit		X		X	X	
Sigma/UniFam/Sipros	X	X	X			X
XGC1	X	X				X
PSANA				X	X	X
ALICE	X					
Tractor			X	X		X
VPIC/IO					X	X
YODA			X			X
ALS SPOT/TomoPy	X			X	X	X
kitware				X	X	

Upgrading Cori's External Connectivity



Enable 100Gb+ Instrument to Cori

- Streaming data to the supercomputer allows for analytics on data in motion
- Cori network upgrade provides SDN (software defined networking) interface to ESnet. 8 x 40Gb/s bandwidth.
- Integration of data transfer and compute enables workflow automation

<u>Cori Network Upgrade Use Case:</u>

- X-ray data sets stream from detector directly to Cori compute nodes, removing need to stage data for analysis.
- Software Defined Networking allows planning bandwidth around experiment run-time schedules
- 150TB bursts now, LCLS-II has 100x data rates

Realtime access to HPC systems

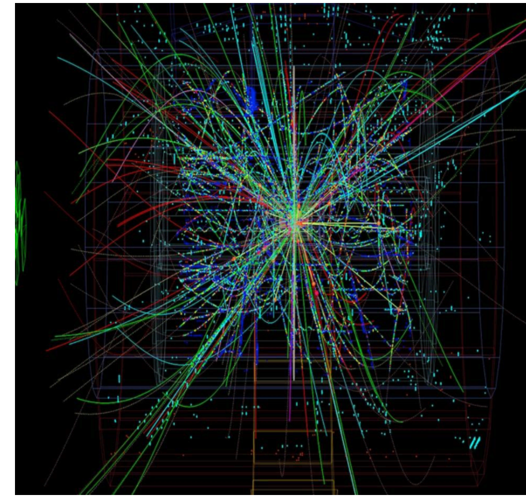
- We've heard from a number of users that lack of 'realtime' access to the system is a barrier to scientific productivity
- We added a question to ERCAP about realtime needs to assess demand and size realtime resources.
- And received 19 responses from 5 out of 6 Offices
- With NERSC's new batch scheduler, SLURM, we have implemented a 'real-time' queue on Cori Phase 1
- With approval from program managers we approved 15 projects for the real-time queue and ~4 are running already

Transition to SLURM to better support data intensive science

- **NERSC made the switch from the Torque/Moab scheduler to SLURM on both Edison and Cori**
- **Open source, NERSC can contribute to development**
- **Enables a number of features for data intensive science**
  - Real time queues
  - High throughput queues

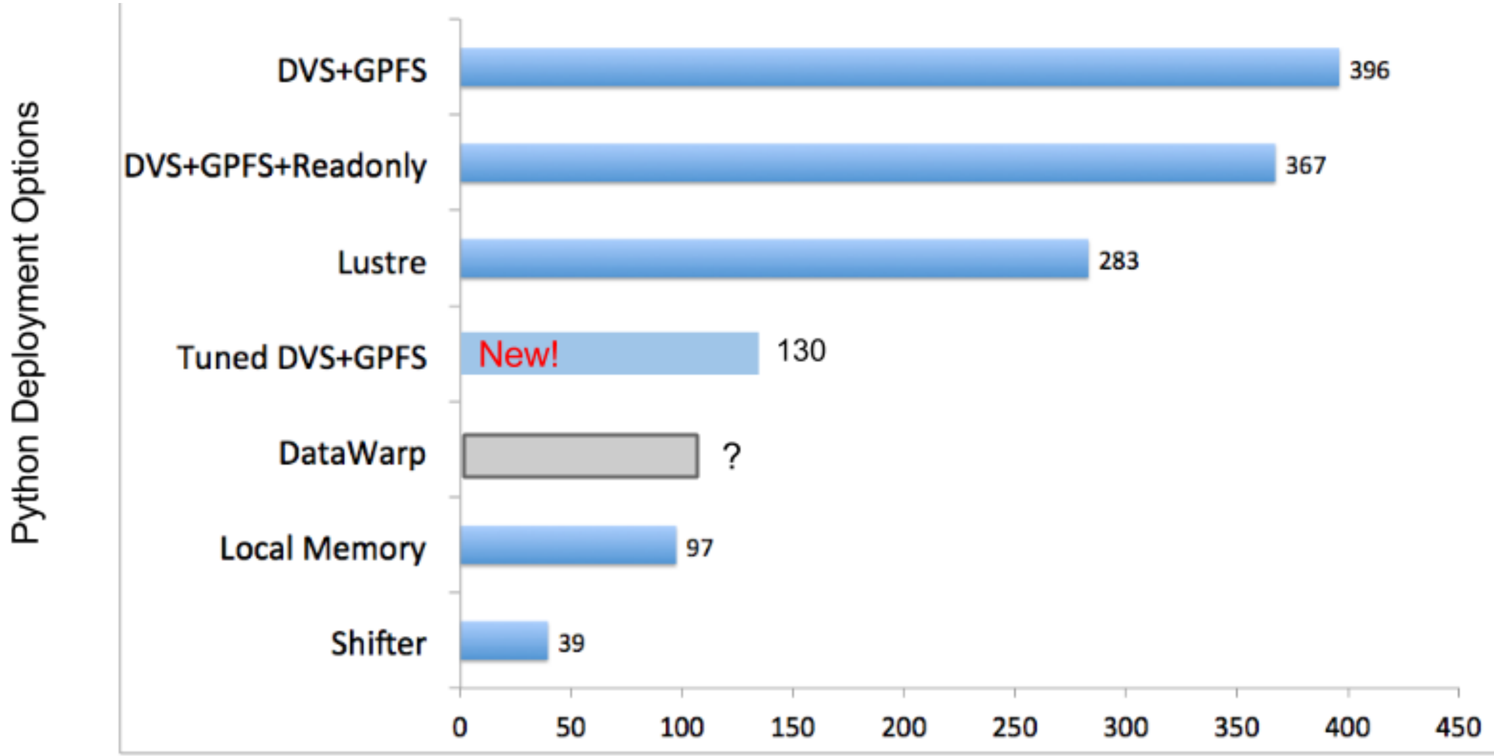
Shifter brings user defined images to supercomputers

- **Shifter, a container for HPC, allows users to bring a customized OS environment and software stack to an HPC system.**
- **Use cases**
  - High energy physics collaborations that require validated software stacks
  - Cosmology and bioinformatics applications with many 3rd party dependencies
  - Light source applications that with complicated software stacks that need to run at multiple sites



Improving Python Performance

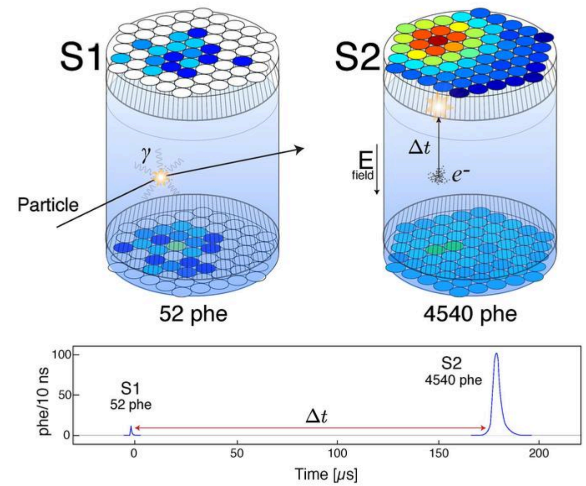
- Python is a critical tool for many data intensive applications
- We have tried various ways to improve python performance on Cray systems



Time (seconds) to complete Python Performance (Test on 300 nodes, 24 cores per node)

New database capabilities: Improving the LUX experiment's workflow with SciDB

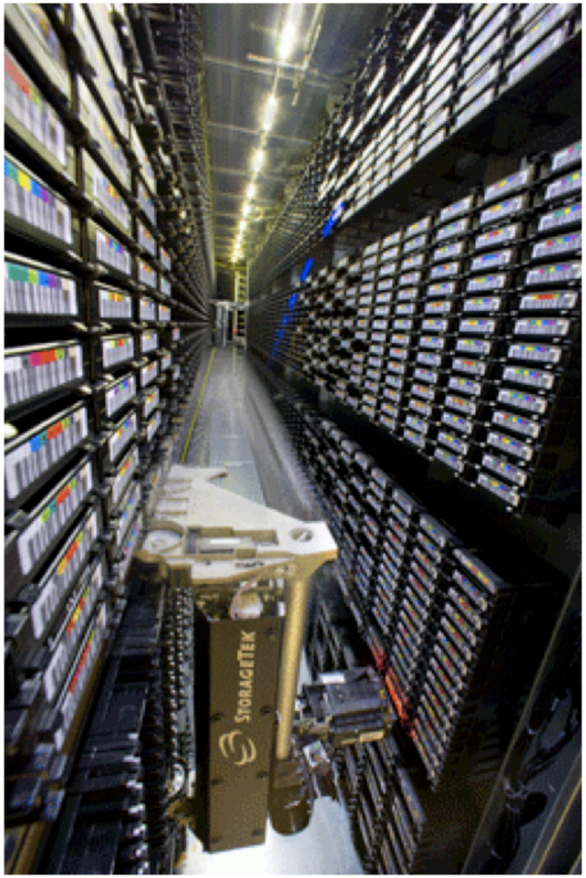
- **Large Underground Xenon (LUX) dark matter experiment operates a mile underground at the Sanford Underground Research Facility in South Dakota with data rates of 250 TB/year**
- **NERSC staff worked with LUX team to load raw data from inaugural run into NERSC's SciDB testbed, open source database for large array-structured data**
- **Using SciDB analysis that would have taken a day, can now be done in 1.5 minutes**



Example LUX event, showing initial flash of light from interaction of WIMP with detector, and subsequent signal from cloud of electrons created in the interaction. This double-pulse is the signature of a dark matter interaction.

Increased Archive Disk Cache improves HPSS performance for users

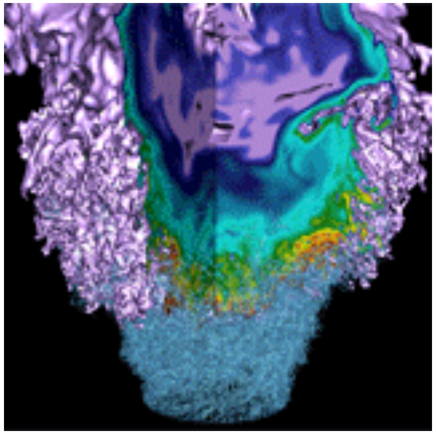
- Disk cache increased by 10x from 200TB to over 2PB
- Before the increase, files stayed on disk cache for 2.5 days and now stay on 24.5 days (10x improvement)
- Impact for users is enormous, latency to tape is 90 seconds while disk cache is < 1 sec
- Of the files read, 75% are read within 30 days days of writing – disk cache close to optimal capacity

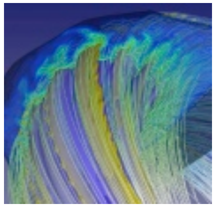


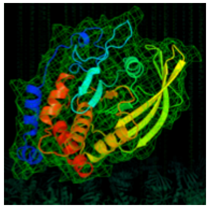
NERSC-9 Will Provide Capabilities for DOE Data-Intensive Users in 2020

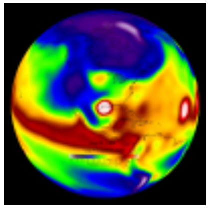
- **NERSC-9 will build upon the successes of the data different components of Cori**
- **End to end workflow requirements and performance are critical for the design and optimization of the system**
- **Overall goal is to enable seamless data motion with dynamic allocation and scheduling of resources**
  - Enable first steps towards exascale-era storage system
  - Vendor community excited about engagement and collaboration opportunities

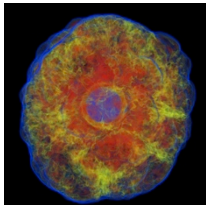
Superfacility Concept

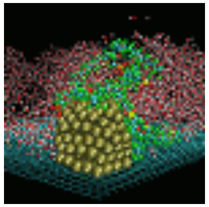




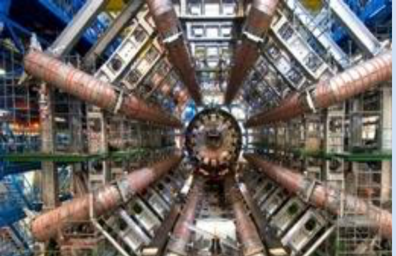
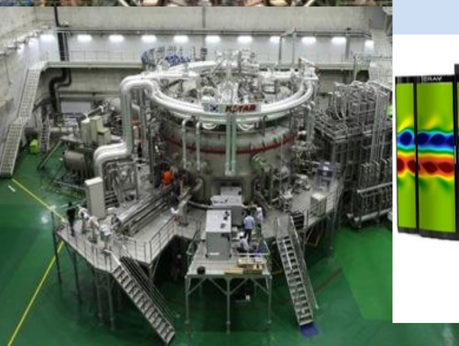
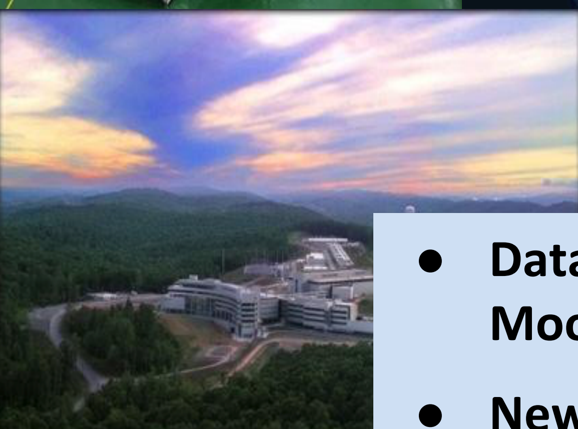


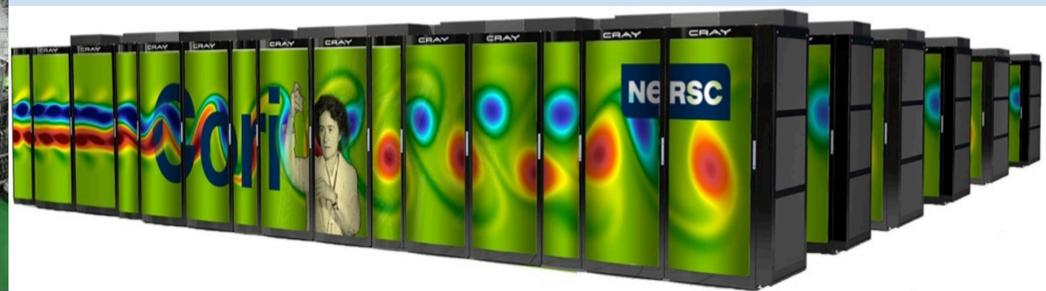






Experimental and observational science is at crossroads

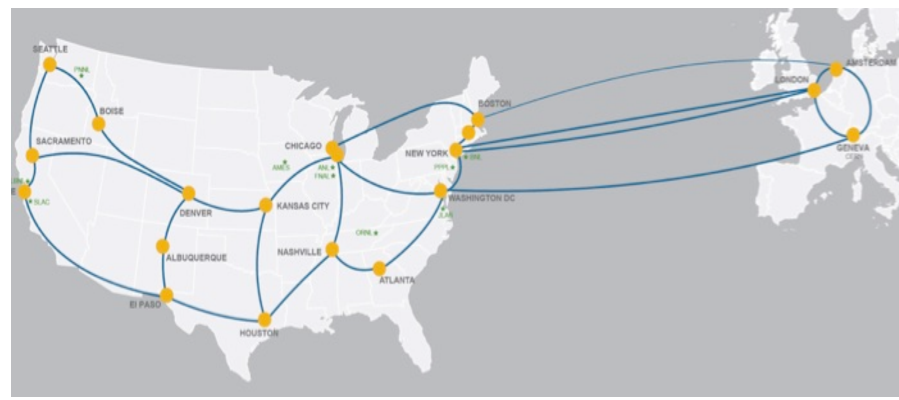




- Data volumes are increasing faster than Moore’s Law
- New algorithms and methods for analyzing data
- Infeasible to put a supercomputing center at every experimental facility

ESnet: Instrument for Broad Impact

- **ESnet's unique role:**
  - –Growing 2x commercial nets
  - –50% of traffic is from "big data"
  - –100Gigabits/sec cross continent
  - –80% starts or ends outside

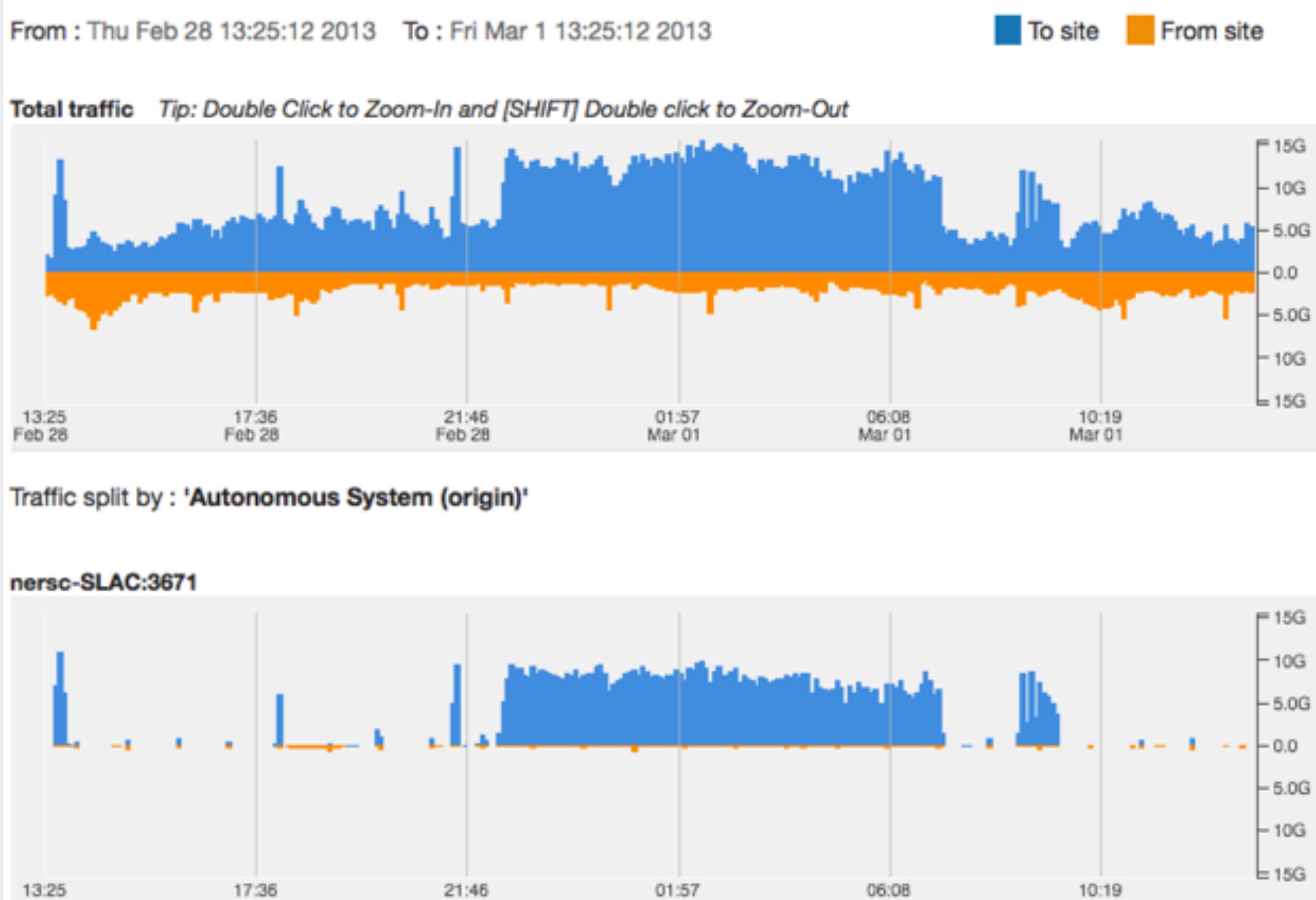


  - –New extension to Europe

Cross Bay Data Transfer

All NERSC Traffic

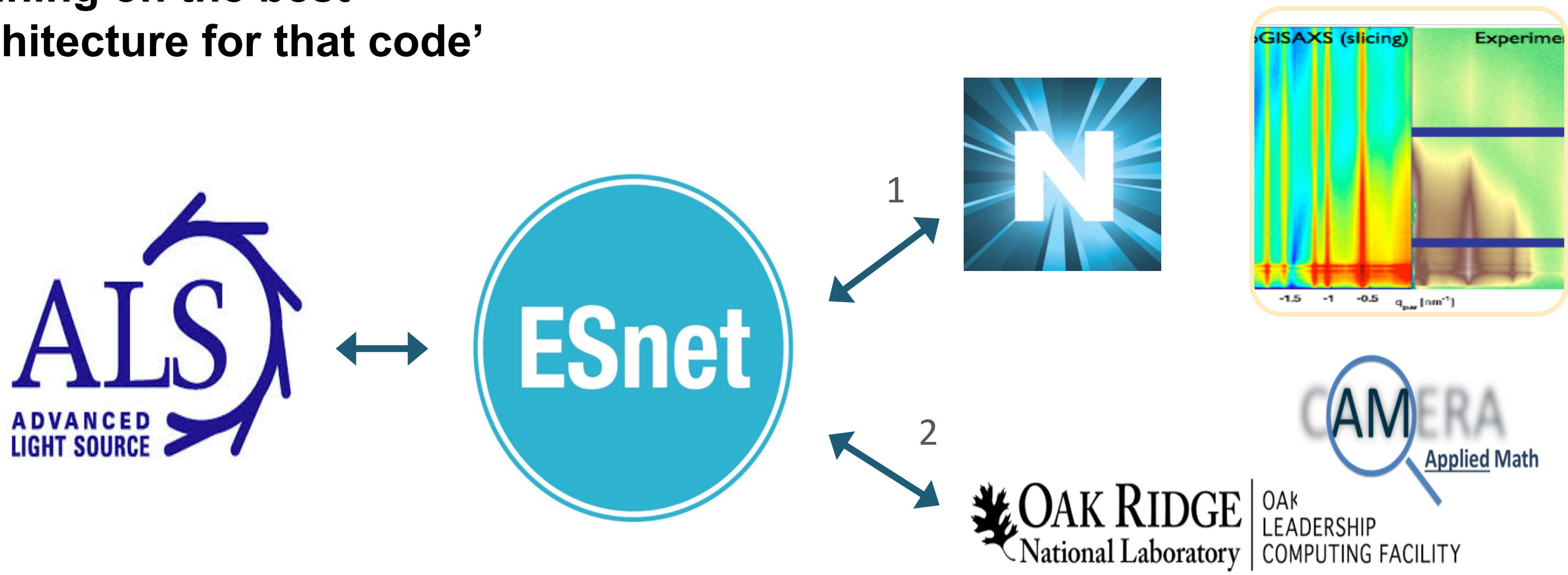
Photosystem II X-Ray Study



Superfacility Prototype and Use Case : Process of science transformed

‘Eliminate boundaries between the Scientist and the world’s best algorithms running on the best architecture for that code’

Real-time analysis of ‘slot-die’ technique for printing organic photovoltaics, using ALS + NERSC (SPOT Suite for reduction, remeshing, analysis) + OLCF (HipGISAXS running on Titan w/ 8000 GPUs).



http://www.es.net/news-and-publications/esnet-news/2015/esnet-paves-way-for-hpc-superfacility-real-time-beamline-experiments/ Results presented at March 2015 meeting of American Physical Society by Alex Hexemer. Additional DOE contributions: **GLOBUS** (ANL), **CAMERA** (Berkeley Lab)

Conclusions

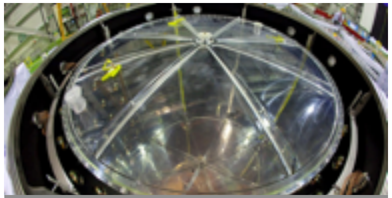
- Meeting the challenges of performance growth will impact all scales of computing and big data (in science)
- We need to develop an infrastructure to support computing and data science because both are becoming increasingly coupled

Solving the Puzzle of the Neutrino

- **HPC and ESnet vital in the measurement of the important "θ_{13}" neutrino parameter.**
  - Last and most elusive piece of a longstanding puzzle: why neutrinos appear to vanish as they travel
  - The result affords new understanding of fundamental physics; may eventually help solve the riddle of matter-antimatter asymmetry in the universe.
- **HPC for simulation / analysis; HPSS and data transfer capabilities; NGF and Science Gateways for distributing results**
  - All the raw, simulated, and derived data are analyzed and archived at a single site
  - => Investment in experimental physics requires investment in HPC.
- **One of Science Magazine's Top-Ten Breakthroughs of 2012**

The Daya Bay experiment counts antineutrinos at three detectors (shown in yellow) near the nuclear reactors and calculates how many would reach the detectors if there were no oscillation. transformation.

NERSC's PDSF cluster



Daya Bay detectors

PI: Kam-Biu Luk (LBNL)