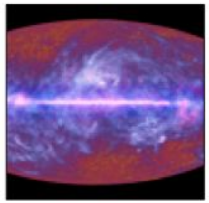
Data Science, Superfacility, & AI

Debbie Bard
Acting Data Department Head
Group Lead, Data Science Engagement

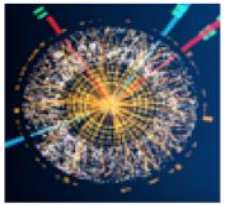
NERSC supports a large number of users and projects from DOE SC's experimental and observational facilities

Palomar Transient Factory
Supernova



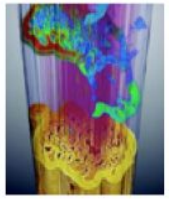
Planck Satellite
Cosmic Microwave Background
Radiation

Star
Particle Physics



Atlas
Large Hadron Collide

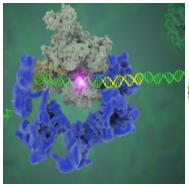
Dayabay
Neutrinos



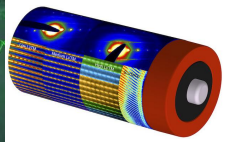
ALS
Light Source

LCLS
Light Source

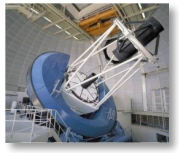
Joint Genome Institute
Bioinformatics



Cryo-EM

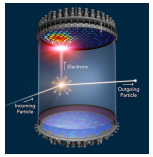


NCEM

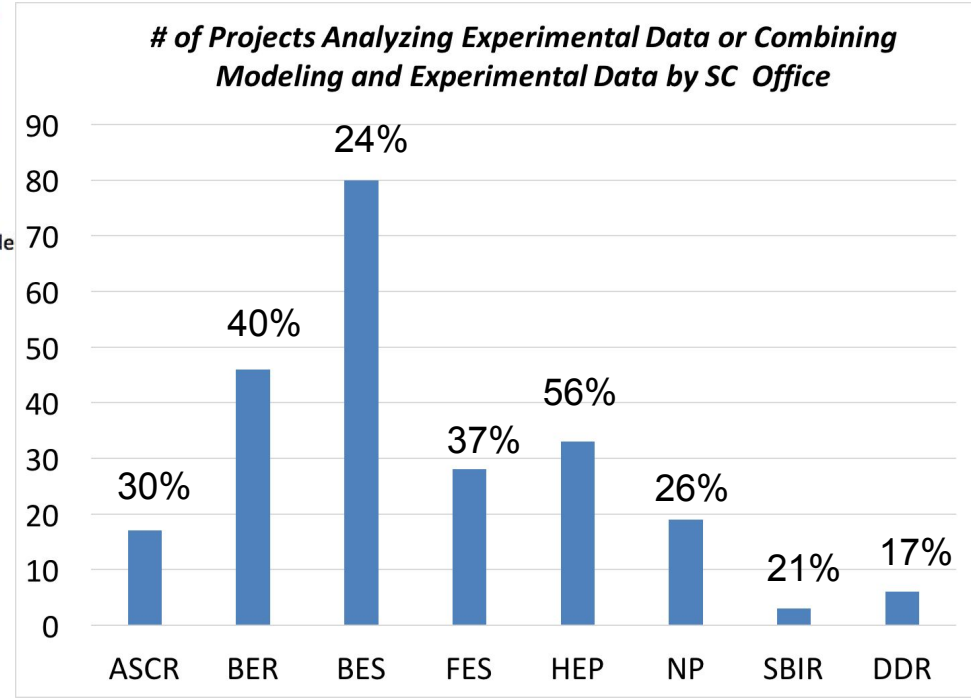


DESI

LSST-DESC



LZ

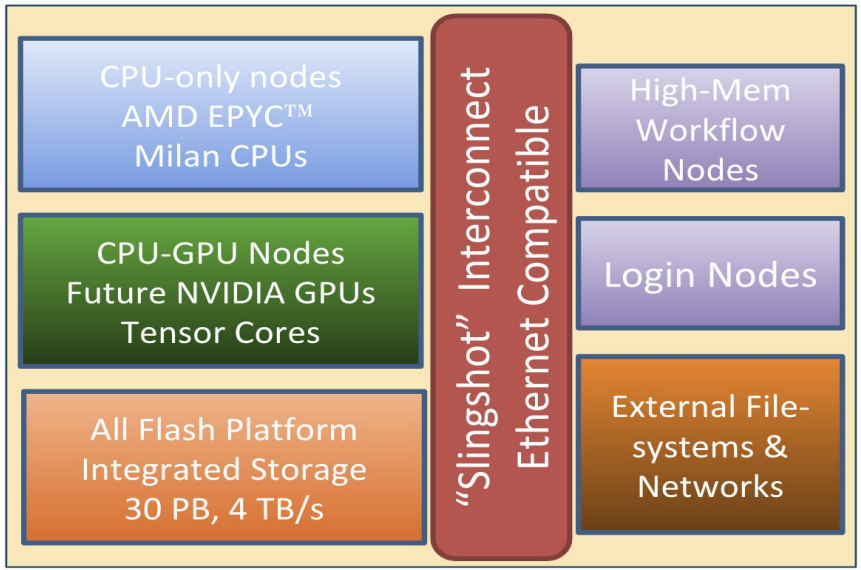


~35% (235) of ERCAP projects self identified as confirming the primary role of the project is to 1) analyze experimental data or; 2) create tools for experimental data analysis or; 3) combine experimental data with simulations and modeling

NERSC-9 has a great configuration for data science

NERSC-9 Capabilities

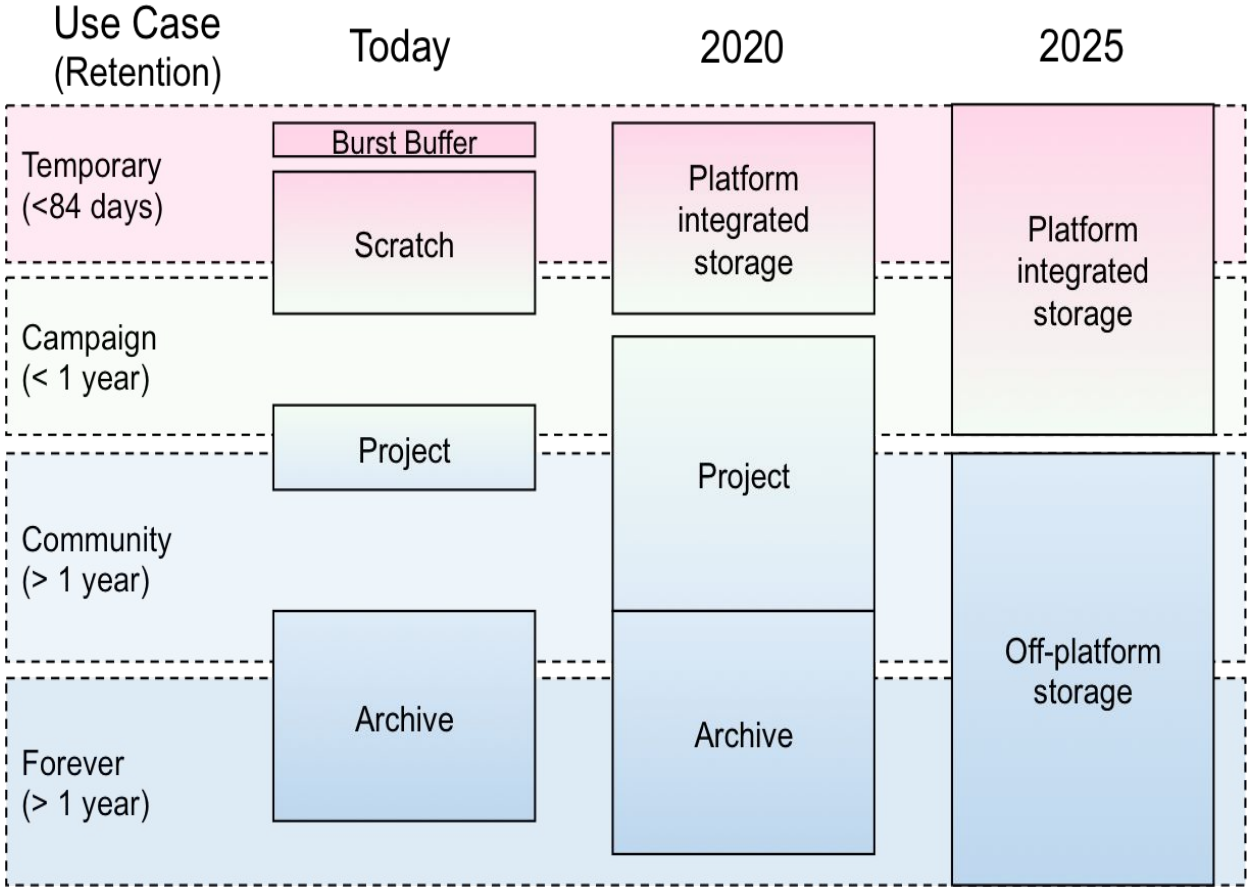
- Powerful platform for large-scale machine learning
- High-performance, configurable networking
- All-flash storage system will benefit complex pipelines
- Flexible system allows tight coupling to dedicated workflow resources within & outside system



NERSC

NERSC storage roadmap

- **Target 2020**
  - Collapse burst buffer and scratch into all-flash scratch
  - Invest in large disk tier for capacity
  - Long-term investment in tape to minimize costs
- **Target 2025**
  - Use single namespace to manage tiers of SCM and flash for scratch
  - Use single namespace to manage tiers of disk and tape for long-term storage

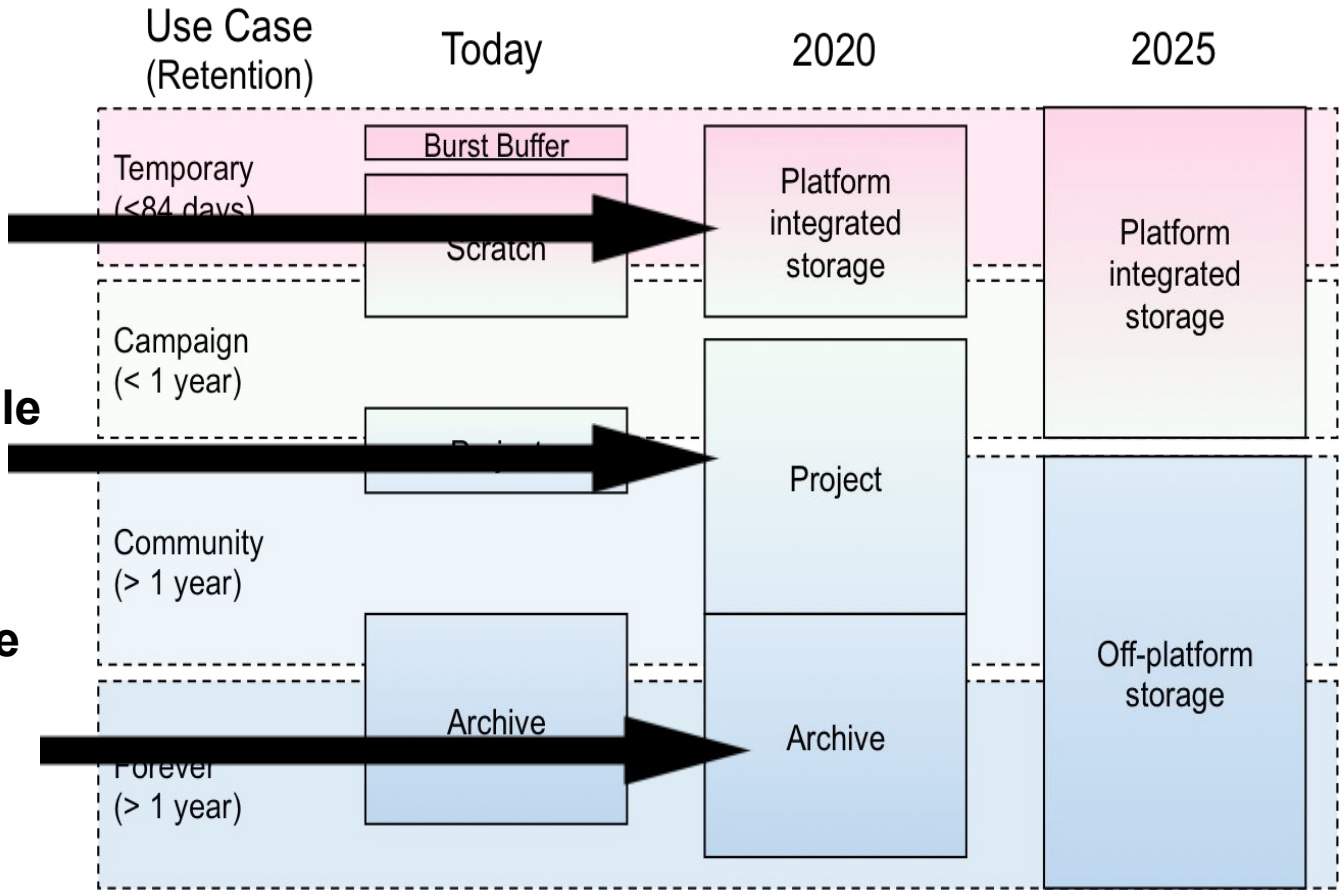


NERSC storage roadmap

All-flash parallel file system on NERSC-9

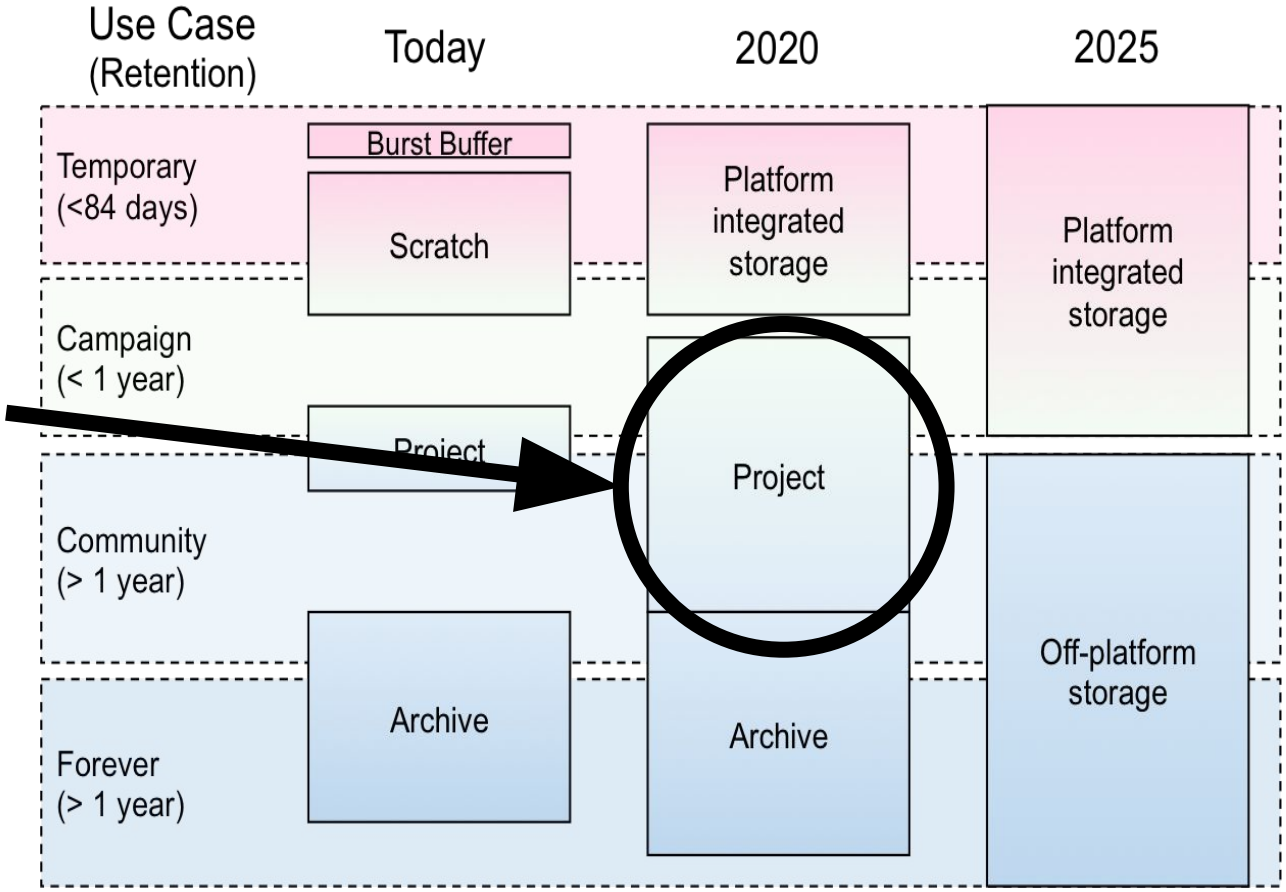
>150PB disk-based file system

>350PB HPSS archive w/ IBM TS4500 (w/ integrated cooling)



Community Filesystem under construction!

- In process of receiving delivery of ~80PB
- Vendor partner: IBM
- GPFS file system
- Targeting ~150-200PB deployment when NERSC-9 arrives



New Tape Libraries installed at CRT and pulling in data

- Although NERSC staff and majority of NERSC systems moved to Wang Hall in 2015, tape libraries did not.
- Energy efficient design of Wang Hall resulted in humidity fluctuations that could be damaging to tape
- Considered building an enclosed room for tapes, and an offsite location.
- NERSC partnered with IBM to deploy environmentally contained tape libraries, saving money and time

CS Area Strategic Plan: Superfacility Initiative

User Engagement

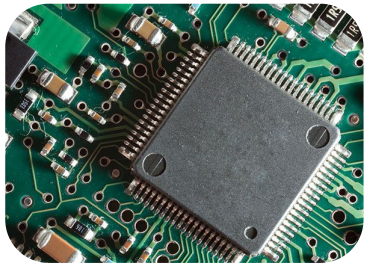
Engage with experimental, observational and distributed sensor user communities to deploy and optimize data pipelines for large-scale systems.

Data Lifecycle

Mange the generation, movement and analysis of data for scalability, efficiency and usability. Enable data reuse and search to increase the impact of experimental, observational and simulation data.

Automated Resource Allocation

Deliver a framework for seamless resource allocation, calendaring and management of compute, storage and network assets across administrative boundaries.



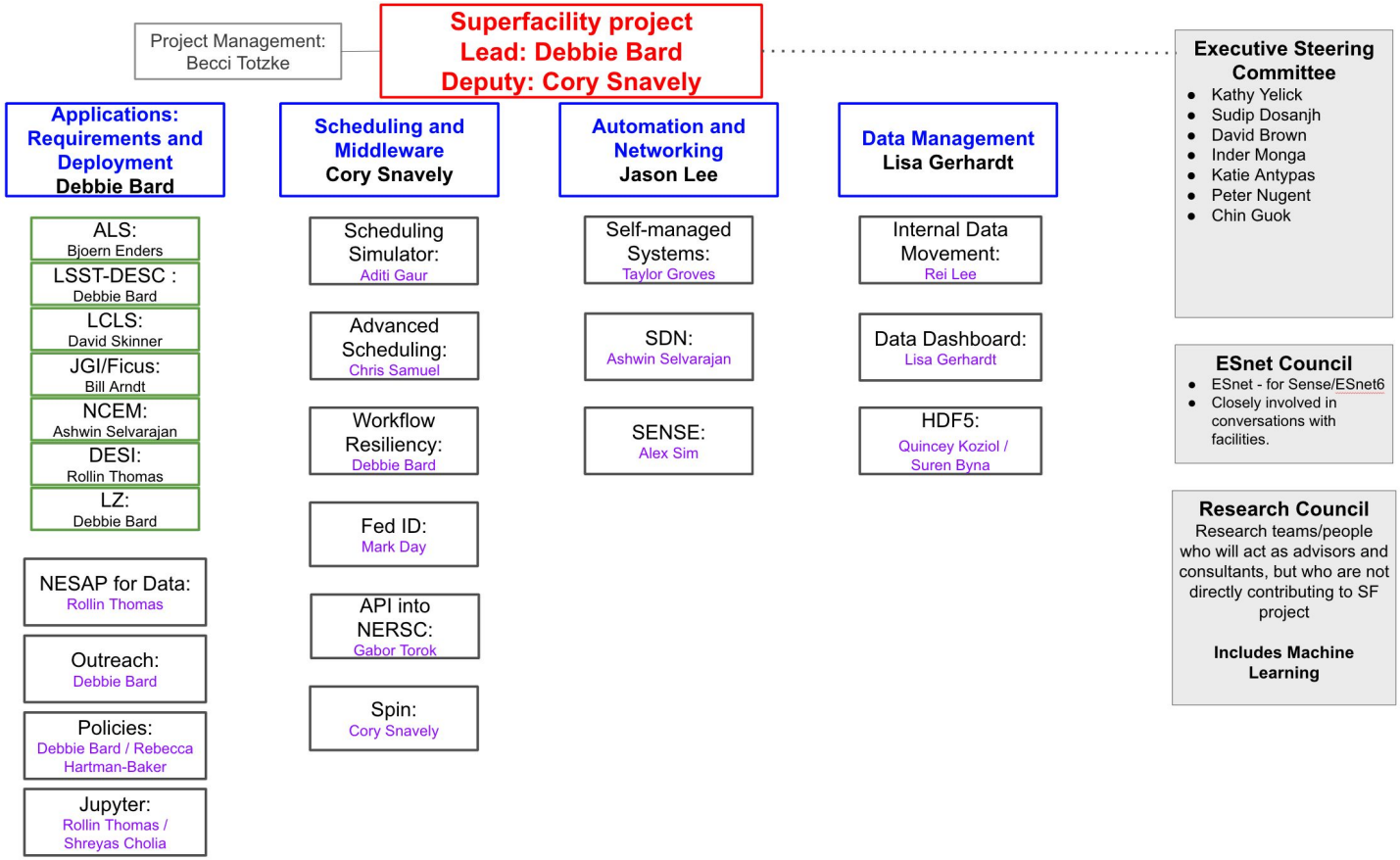
Computing at the Edge

Design and deploy specialised computing devices for real-time data handling and computation at experimental and computational facilities.

NERSC

https://cs.lbl.gov/strategy/

The Superfacility internal project will coordinate, plan and manage the technical work



The Superfacility Project will coordinate relevant work at NERSC/ESnet/CRD and increase communication and focus across the groups involved.

The Superfacility internal project will coordinate, plan and manage the technical work

Project Management: Becci Totzke

Superfacility project
Lead: Debbie Bard
Deputy: Cory Snavely

Applications: Requirements and Deployment
Debbie Bard

- ALS: Bjoern Enders
- LSST-DESC : Debbie Bard
- LCLS: David Skinner
- JGI/Ficus: Bill Arndt
- NCEM: Ashwin Selvarajan
- DESI: Rollin Thomas
- LZ: Debbie Bard
- NESAP for Data: Rollin Thomas
- Outreach: Debbie Bard
- Policies: Debbie Bard / Rebecca Hartman-Baker
- Jupyter: Rollin Thomas / Shreyas Cholia

Scheduling and Middleware
Cory Snavely

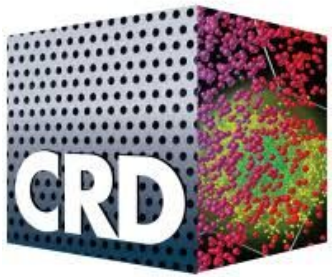
- Scheduling Simulator: Aditi Gaur
- Advanced Scheduling: Chris Samuel
- Workflow Resiliency: Debbie Bard
- Fed ID: Mark Day
- API into NERSC: Gabor Torok
- Spin: Cory Snavely

Automation and Networking
Jason Lee

- Self-managed Systems: Taylor Groves
- SDN: Ashwin Selvarajan
- SENSE: Alex Sim

Data Management
Lisa Gerhardt

- Internal Data Movement: Rei Lee
- Data Dashboard: Lisa Gerhardt
- HDF5: Quincey Koziol / Suren Byna



Executive Steering Committee

- Kathy Yelick
- Sudip Dosanjh
- David Brown
- Inder Monga
- Katie Antypas
- Peter Nugent
- Chin Guok

ESnet Council

- ESnet - for Sense/ESnet6
- Closely involved in conversations with facilities.

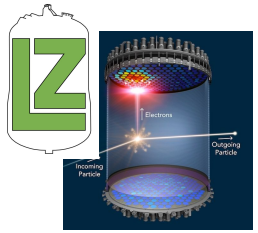
Research Council
Research teams/people who will act as advisors and consultants, but who are not directly contributing to SF project

Includes Machine Learning

NERSC-funded collaborations with CRD

- Roofline modeling for Perlmutter
- Accelerator exploration for NERSC-10
- Enabling Jupyter tools for the Superfacility model
- Tuning data precision for applications (PREMIX)
- Supporting complex workflows at NERSC

Science Engagements



Next-generation dark matter detection, continuously sending data to NERSC and UK

ALS

High-rate detectors use NERSC for real-time experimental feedback, data processing/management, and comparison to simulation

FICUS
Facilities Integrating Collaborations for User Science
JGI-EMSL Collaborative Science Call

Complex multi-stage workflow to analyse response of soil microbes to climate change

LSST DESC
Dark Energy Science Collaboration

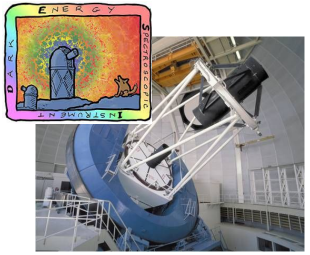
Processing streaming alerts (from NCSA) for detection of supernova and transient gravitational lensing events

MOLECULAR FOUNDRY

4D STEM data streamed to NERSC, used to design ML algorithm for future deployment on FPGAs close to detector

LCLS

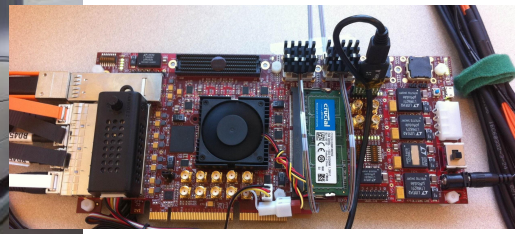
High-rate detectors use ESnet and NERSC for real-time experimental feedback and data processing



Nightly processing of galaxy spectra to inform next night's telescope targets

NCEM detector/DAQ development

- **Stream up to 400Gb/s directly to SSD (burst buffer) during test runs**
  - Use data to train AI algorithm to down-filter data stream.
  - Filtering algorithm will then be deployed on FPGAs close to instrument.
- **Transition to regular operations: burst to NERSC for high-intensity runs, real-time feedback on data quality.**

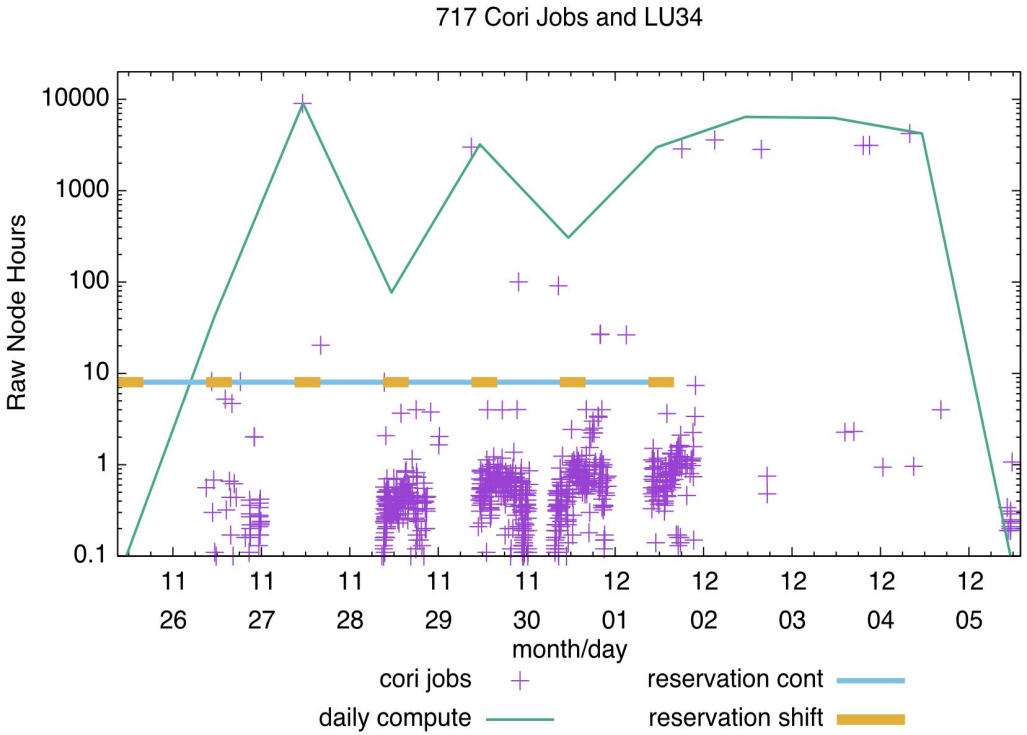


FPGA based readout system

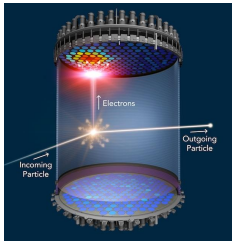
4 x 100 GbE

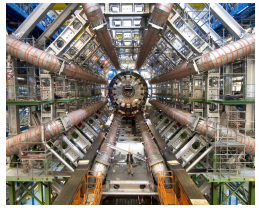
LCLS Experiments using NERSC in Production

- **Live analysis provided to beamline staff**
- Detector to Cori rate ~ 5GB/s
- Use compute reservation on Cori
- Feedback rate is ~ 20 images/sec using 8-16 Cori nodes
  - allows team to keep up with the experiment
- On LCLS cluster, only 10-20% of the data could be analyzed in realtime.



JGI and PDSF workloads Transitioned to Cori





JGI
JOINT GENOME INSTITUTE

NERSC

Mendel cluster retires in a week!

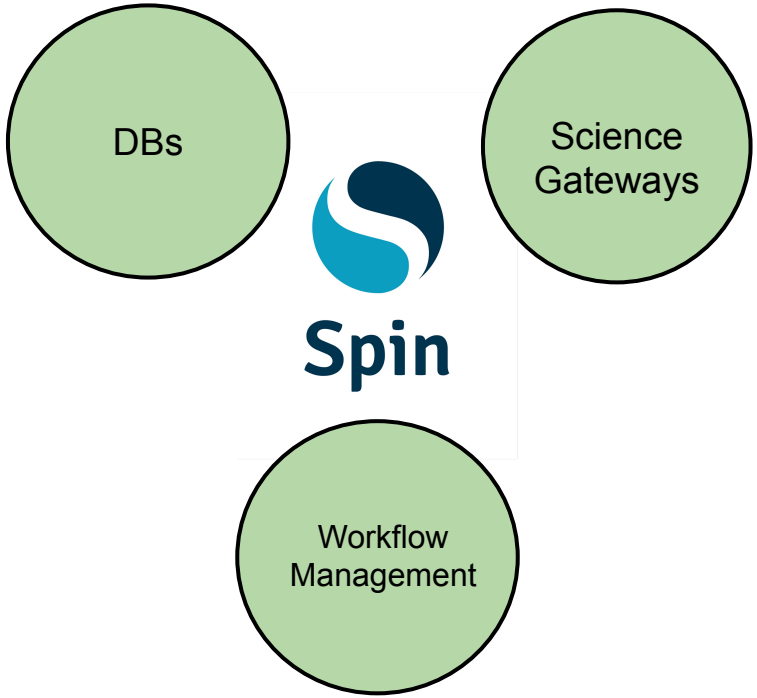
- PDSF was retired at the end of March
  - Longest continuously-running Linux cluster on the planet
  - Enabling of CVMFS and Shifter Images was key to transitioning Particle Physics workloads
- Denovo/Genepool retires at the end of July
  - Test “blackout” of Denovo hardware happened in April
  - All JGI pipelines now run on Cori (JGI partition and JGI allocation)
  - JGI Services transitioned to Spin/VMware

Spin: Edge Services for Complex Workflows

Workflows often require additional edge services (DBs, APIs, Portals) to achieve their science.

Spin: Container-based platform to easily and quickly create science gateways, workflow managers and other edge services, with limited assistance from staff

- Tightly coupled with HPC resources
- Scalable user defined services

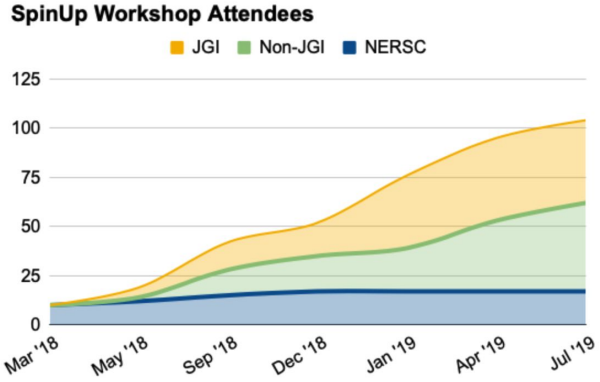


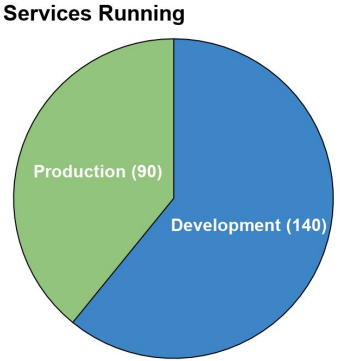
Spin in Production Pilot for Staff and Users

Since launch, **87 users** and **17 staff** have attended six workshops on Spin.

What's running in Spin?

- **ESS-DIVE** (data archive)
- **JupyterHub** (interactive notebooks)
- **Materials Project & Data Bank**
- **OpenChemistry Data Platform** (Kitware)
- **R Studio**
- **ScienceSearch** (ML-driven data index)
- various **JGI pipelines** and services
- *and more...*





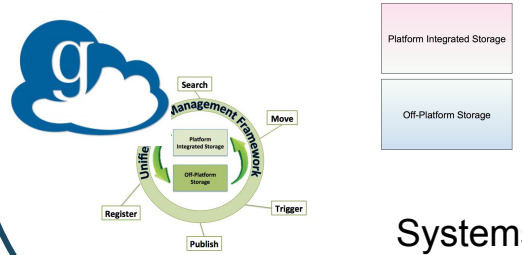
API for Experimental Facilities

Job management: submission, monitoring, retries

slurm workload manager

newt

Data Movement: Between layers, across facilities



Platform Integrated Storage

Off-Platform Storage

Systems section

Software section

Reservations: HPC, Storage, BW

SDN

slurm workload manager

Publish and Share Data

ESG OpenMSI newt ALS

Manage Identities

IAM Service

Ø

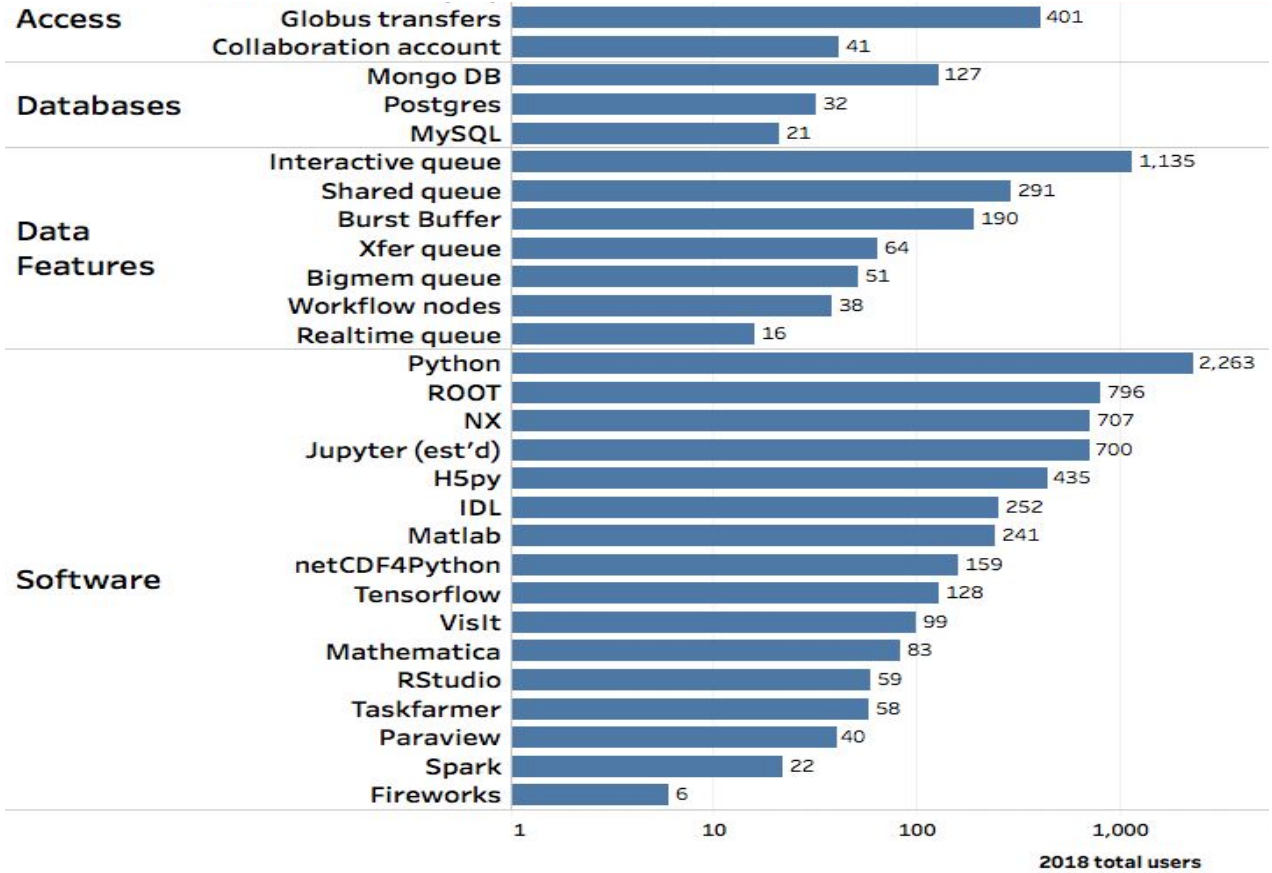
NERSC

Production Data Stack

Capabilities	Technologies
Data Transfer + Access	globus online, GridFTP, jupyter, !m, python django, newt
Workflows	FireWorks, TaskFarmer, 
Data Management	HDF, netCDF, ROOT Data Analysis Framework, mongoDB, MySQL, PostgreSQL
Data Analytics	python, Spark, julia, TensorFlow, Caffe, PYTORCH
Data Visualization	VisIt, ParaView

NERSC

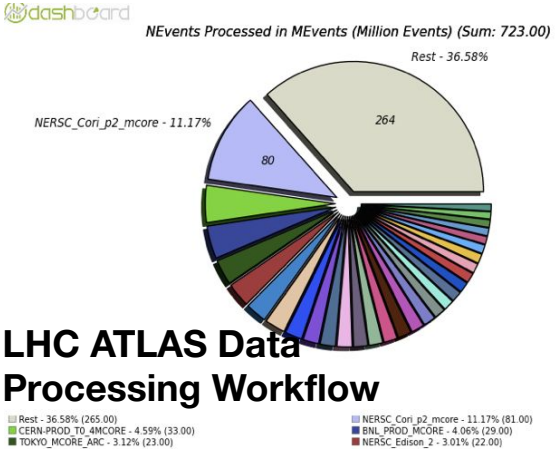
Strong Adoption of Data Stack



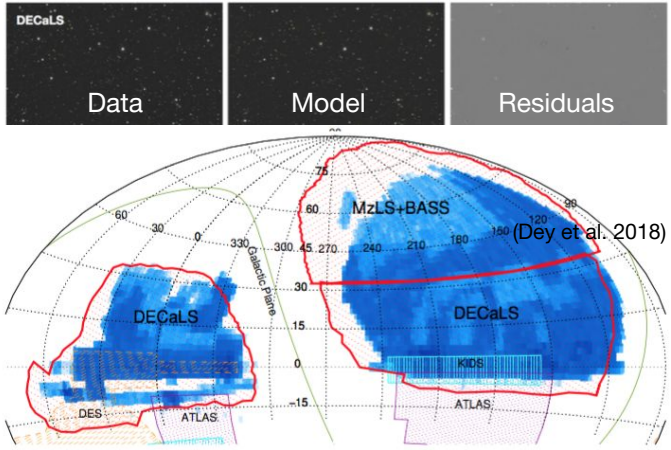
Science via Python@NERSC

The Materials Project

Powering Workflows to Understand Properties of Materials



LHC ATLAS Data Processing Workflow

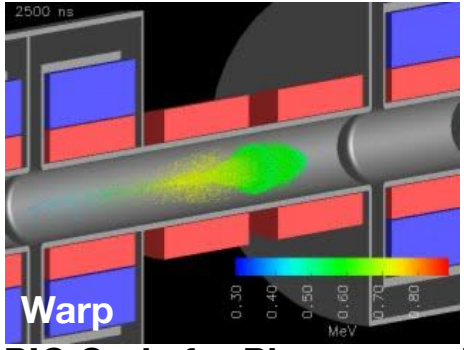


(Dey et al. 2018)

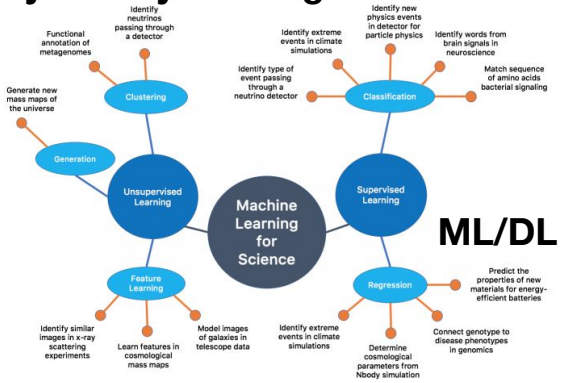
Sky Survey Catalogs for Cosmology

NBODYKIT

Modeling Dark Matter and Dark Energy



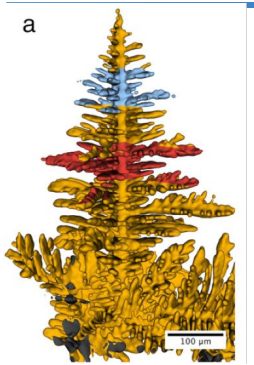
PIC Code for Plasmas and High Current Particle Beams

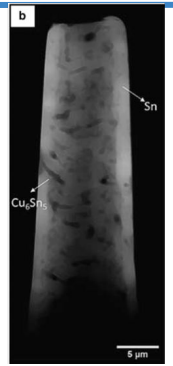


ML/DL

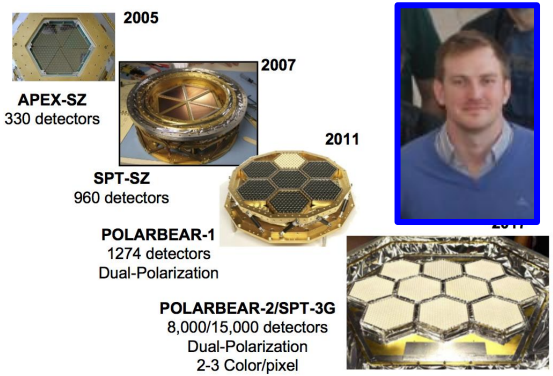
NERSC

NESAP for Data

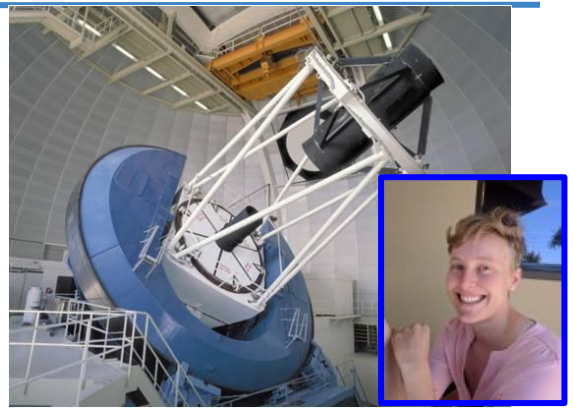




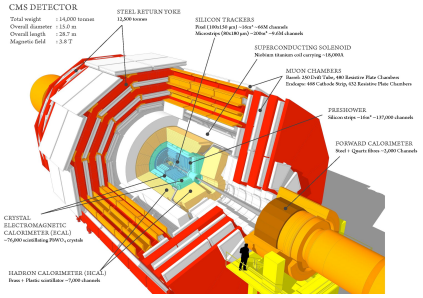
TomoPy (Postdoc: Z. Ronaghi)



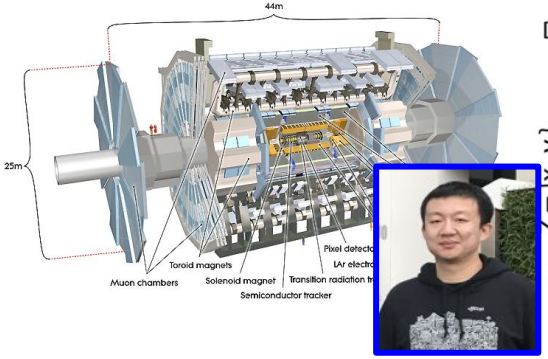
CMB (Postdoc: J. Madsen)



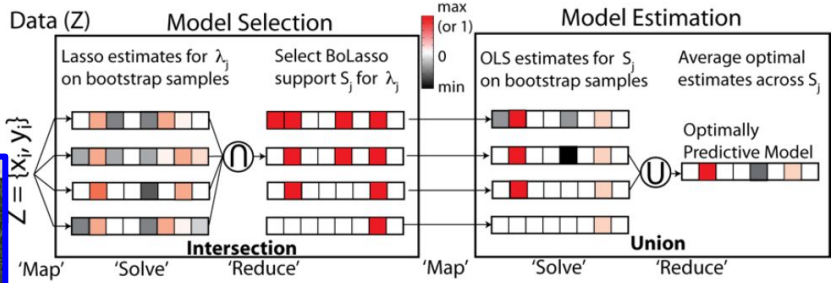
DESI (Postdoc: L. Stephey)



CMS



ATLAS (Postdoc: Y. Wang)



ML/Neuro

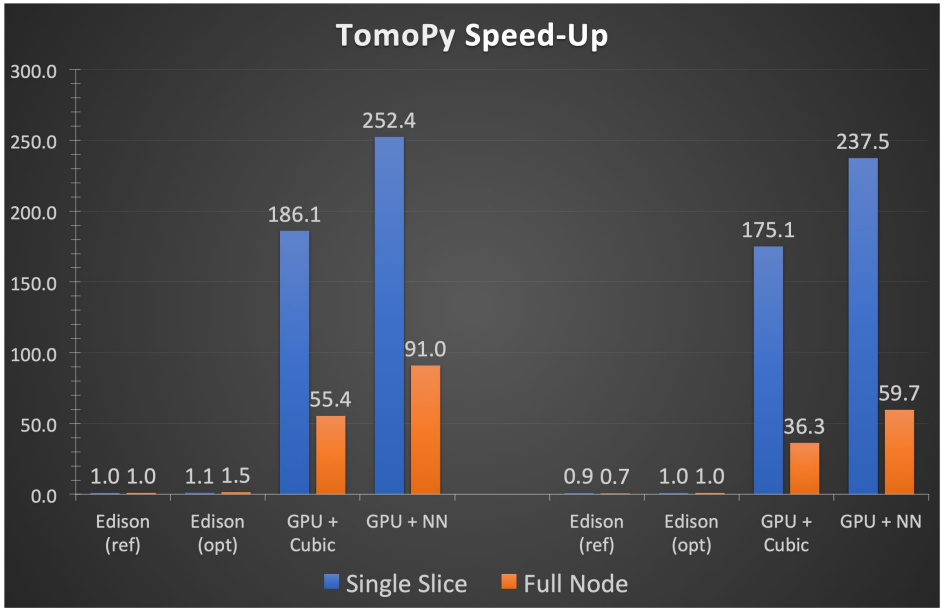
NERSC

NESAP for Data: now starting Round 2

Jonathan Madsen

TomoPy (APS, ALS, etc)

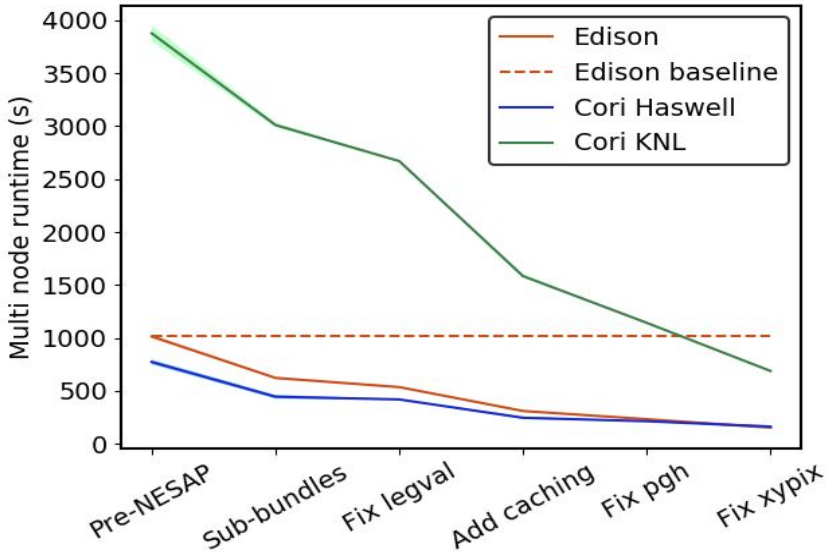
- GPU acceleration of iterative reconstruction algorithms
- New results from first NERSC-9 hack-a-thon w/NVIDIA, >200x speedup!



Laurie Stephey

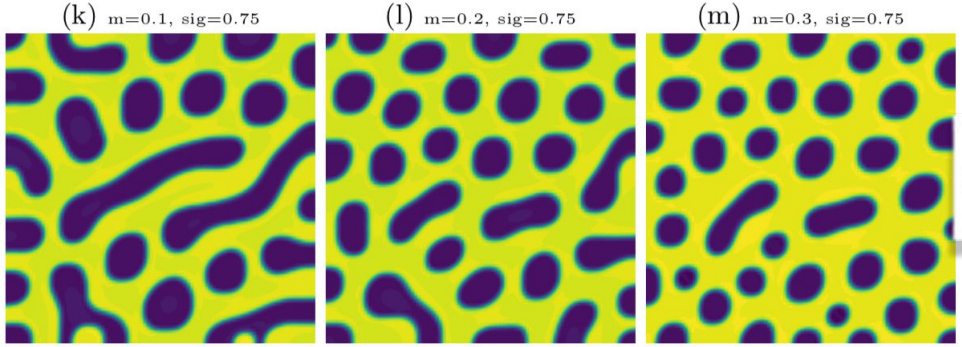
DESI Spectroscopic Extraction

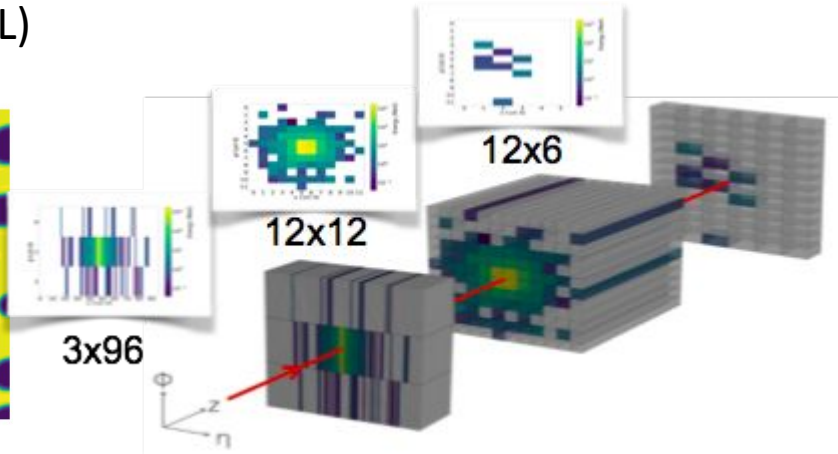
- Optimization of Python code on Cori KNL architecture
- Code is 4-7x faster depending on architecture and benchmark



NESAP for Learning: new for Perlmutter

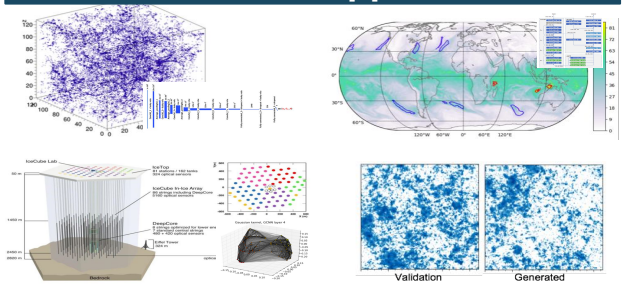
- FlowGAN (Marc Day, LBNL)
- Extreme Scale Spatio-Temporal Learning (Shinjae Yoo, BNL)
- Accelerating HEP Simulations with ML (Ben Nachman, LBNL, Jean-Roch Vlimant, Caltech)
- Deep Learning for Thermochemistry (Zachary Ulissi, CMU)
- RL for Light Sources (Christine Sweeney, LANL)





NERSC Learning strategy

Methods and Applications



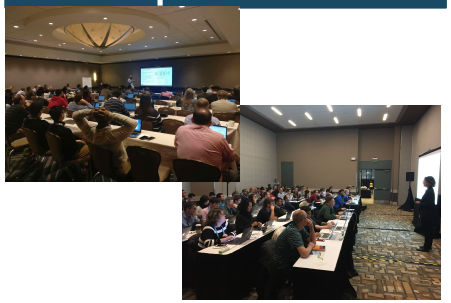
Deployment

Automation | Hubs | Notebooks

Software Frameworks and Libraries

Systems w/ Accelerators

Empowerment

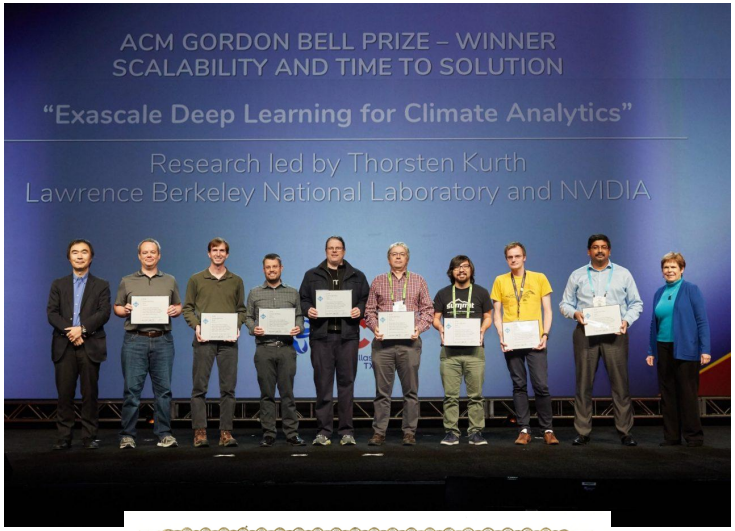


- **NERSC recognized growing ML importance for 5+ years:**
  - *Big Data Center* projects; Cori SW deployment; Requirements in Perlmutter design
- ***Apply ML* for science using cutting-edge methods**
  - In-depth engagements; NESAP for Learning; Leverage commonality via model hubs
- ***Deploy* optimized hardware and software systems**
  - Productive SW at HPC scale; Benchmarks; Vendor engagements; HW evaluation
- ***Empower* through seminars, workshops, training and schools**

NERSC

Leadership within Machine Learning

- SC'18 Gordon Bell Award
  - 1st Exascale Deep Learning application
- ICMLA '18 Best Paper Award
  - Graph NNs for IceCube Neutrino classification
- ISC'19 Hyperion Research HPC Innovation Excellence Award
  - Unsupervised Discovery of Coherent Fluid Flow Structures
- SC'19 Best Paper Finalist
  - etalumis: Combining Probabilistic Programming with DL
- SC'19 Gordon Bell submission
  - 1.2 EF PI-GAN implementation on Summit




17th IEEE International Conference on Machine Learning and Applications
IEEE December 17-20, 2018, Orlando, FL, USA AMLA

BEST PAPER AWARD

Graph Neural Networks for IceCube Signal Classification

Nicholas J Choma (New York University); Federico Monti (Universitá della Svizzera Italiana); Lisa Gerhardt (Lawrence Berkeley National Laboratory); Tomasz Palczewski (University of California, Berkeley); Zahra Ronaghi (Lawrence Berkeley National Laboratory); Prabhat (Lawrence Berkeley National Laboratory); Wahid Bhimji (Lawrence Berkeley National Lab); Michael Bronstein (Imperial College); Spencer R. Klein(University of California, Berkeley); Joan Bruna (Courant Institute of Mathematical Sciences, NYU, USA)

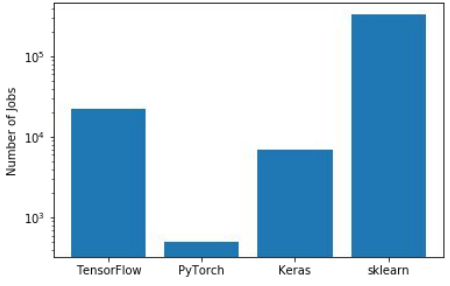
Mehmed Kantardzic, Conference Co-Chair

Moamar Sayed-Mouchaweh, Conference Co-Chair

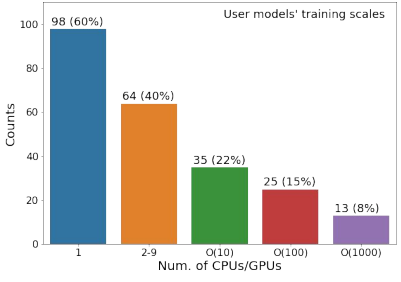
Production (Deployment)

- Software maintenance, documentation
- Workload analysis and survey
- Benchmarking
  - TensorFlow benchmarks
  - PyTorch benchmarks
  - Science benchmarks, *MLPerf HPC*
- Jupyter
  - Interactive *distributed* deep learning

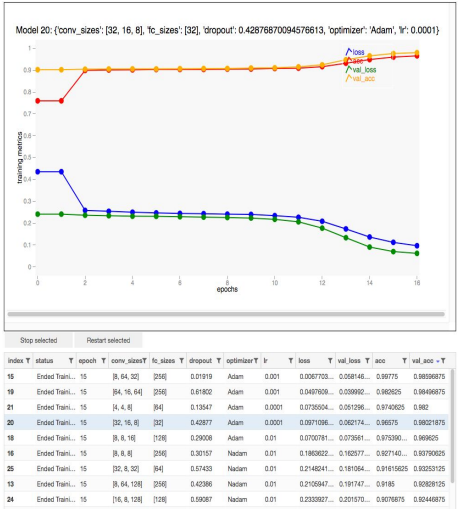
Current workload:

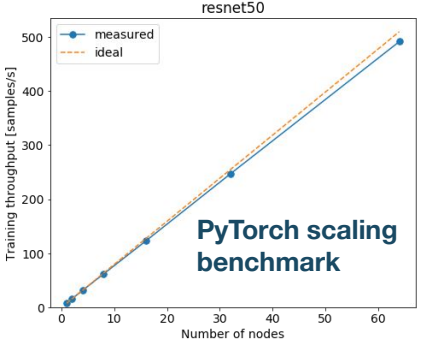


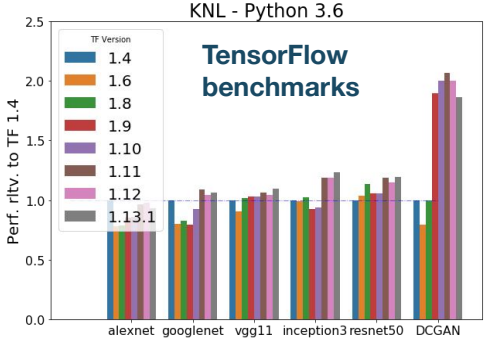
User surveys:



Jupyter deep learning:

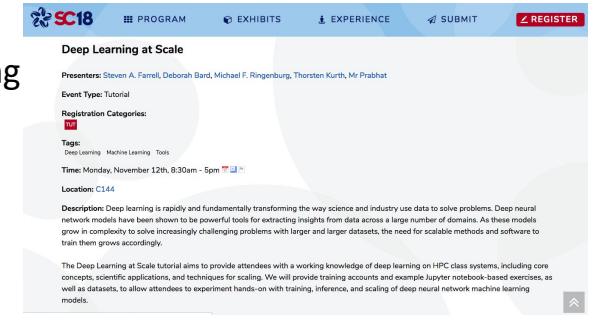






Outreach (Empowerment)

- **Workshops**
  - **Machine Learning for Science Workshop @LBL**
  - **Monterey Data Conference**
- **Training**
  - Data Day
  - ECP AHM
  - SC'18 Tutorial; SC'19 proposal accepted
  - CUG'19 Tutorial
  - SEA'19 Tutorial
  - ISC'19 Tutorial
  - **Deep Learning for Science Summer School**
- **Plenaries/Keynotes**
  - Intel HPCDevCon
  - CHEP
  - NCAR SEA
  - Google TensorFlow Dev Summit
  - IMSC tri-annual meeting
  - AGU, AMS



NERSC

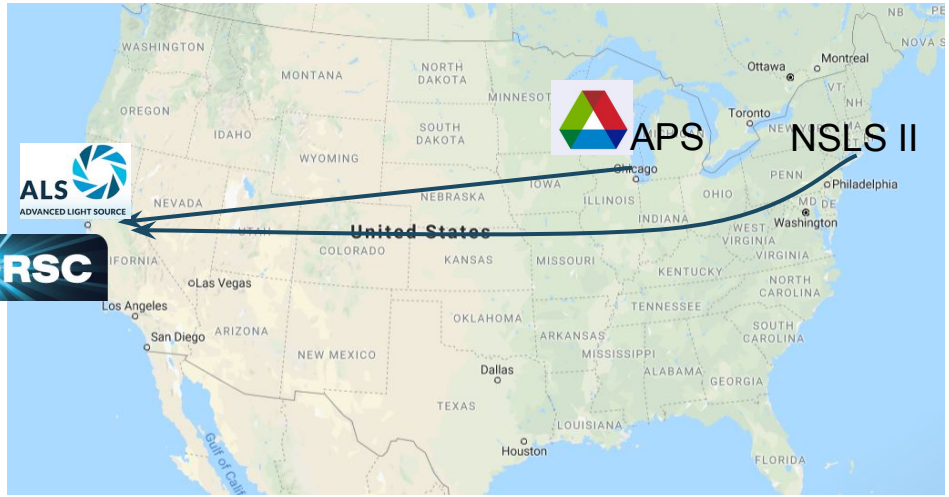
Summary

- We’re busy in the data department!
  - We’ve been focussing on the hardware we’re deploying, the services and technical solutions we’re developing, and the high performance scalable software we’re deploying.
- Deployment of Community File System is first step in realising the Storage2020 plan
- Superfacility project will ensure we deploy services, technical capabilities and policies for experimental science, useful across the NERSC userbase
- We continue to expand our program in ML and AI, focussing on methods, deployment and empowerment.

Needs from NERSC

Experiment	What runs at NERSC?	What runs elsewhere?
LCLS	5-10% of experiments that require >32PF compute in 2021 (~3% >128PF in 2027)	All other experiments at LCLS
ALS	2-3 beamlines with large computing requirements, i.e. tomography and ptychography (~200MB/s)	Other ~40 ALS beamlines
NCEM	Stream super high-rate (>400 Gb/s) detector data to NERSC for algorithm design	Low data-rate microscopes do not use NERSC
LSST-DESC	Large-scale cosmology and instrument simulations (NESAP team); Supernova alert processing draws on multiple PB-scale data sources	Small-scale analysis done at home institutions
DESI	Short-turnaround compute needs for rapid analysis, co-location of data and simulation	Small-scale analysis done at home institutions
LZ	Combination of large-scale simulations and relatively small data coming from the experiment	Mirror data processing in UK; small-scale analysis done at home institutions
JGI/FICUS	Complex multi-stage workflow with some large MPI components (FICUS); Large-scale assembly pipelines (hipmer)	JGI exploring appropriate compute options for some workloads

DOE synchrotron light sources collab



Big picture

Coherent or full-field experiments use high frame rate 2D detectors for their science.

Data volume and computation has become increasingly demanding to host on-site.

Value proposition

HPC enables live workflows for fast feedback (Appl.: Ptychography, Tomography)

Data analysis (HPC) and Sharing (Globus, Spin) and Archiving (HPSS)

Colocation of compute and data allows for custom, user-based post-processing and potential for reprocessing and virtual experiments

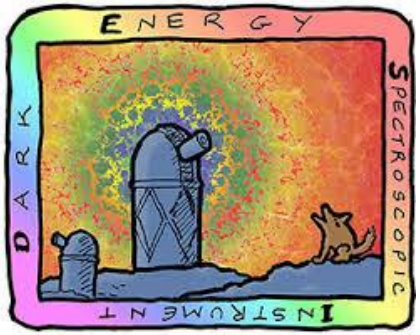
Example Ptychography@ALS

Needs from NERSC

- **SDN + Advanced Scheduling to stream data into readily available compute nodes.**
  - Data rates at 50-200MB/s per detector stream currently (2018)
- **Cori and/or Cori GPU for preprocessing and computation**
  - 50-150 CPU cores per detector
  - Computation requirements depend on size of dataset.
- **Data Mover API to archive and stage data**
  - 30-60 TB raw data per week per detector
- **SPIN for user gateway deployment or FEDid / user account**
- **An API to synchronize data deployment with NERSC uptime**

DESI: Dark Energy Spectroscopic Instrument

Science Purpose: Explaining the Physics of Dark Energy



- 3D map of the Universe over the past 10 billion years.
  Spectra from Kitt Peak of 10's of millions of galaxies and quasars
- Statistical properties of the 3D distribution:
  cosmological parameters ⇒ past/future evolution of Universe ⇒ fundamental physics

Importance of NERSC to DESI

- DESI needs to select targets.
- **Storage** needed to co-locate survey/sim data; **compute** needed to process, re-process, analyze the data and run target selection.
- Spectroscopic extraction pipeline is computationally intensive, needed work to optimize for future architectures ⇒ NESAP for Data.
- Spin enables DESI collaboration to monitor survey progress and share results.

DESI: Dark Energy Spectroscopic Instrument

2019 Superfacility Milestones for DESI

Data Management: Internal Data Movement, Data Dashboard

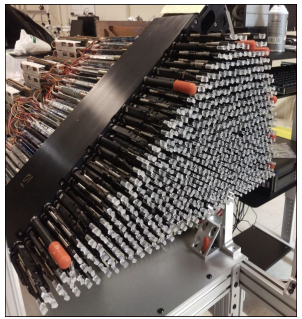
July 1 *Move data (preserving metadata) on large data sets, ingress data as project user, all without tickets; Globus sharing.*

Automation and Networking (or Scheduling and Middleware?)

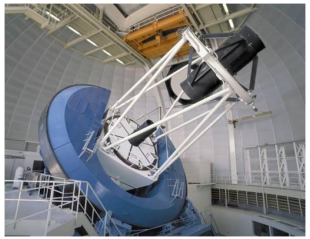
Sep 1 *"Cron on specific systems" (or equivalent without gap) with minimal MFA impact; ability to schedule or trigger actions throughout the center.*

Scheduling and Middleware: Advanced Scheduling

Sep 1 *Deadline scheduling; ability to resize jobs even if just give-backs.*



DESI Fiber Positioner Petal
1 Exposure = 30 Frames
= 15,000 Spectra



Mayall 4m Telescope, KPNO
DESI Commissioning: Sep 2019

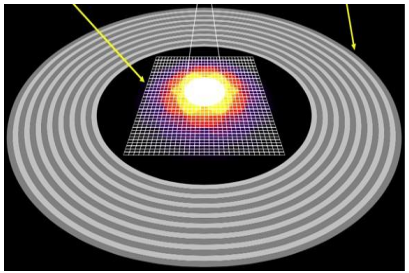
Projections for DESI through 2025: 7.6 PB & 182M MPP

- Up to 100 GB/night (raw), processed is 5-10x larger; need results next day
- 4.2 PB of spectroscopic data (raw+reduced)
  2.5 PB of simulation and analysis products
  ~1 PB of imaging survey data for target selection
- 182M MPP hours (pre-NESAP estimate, to be revised)

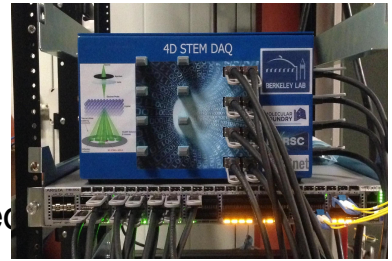
NCEM - 4D STEM Detector Development

Science Story: Development of a High Frame Rate 4D STEM detector

- NCEM is developing a high frame rate (100KHz) 4D detector system to enable fast real-time data analysis of scanning diffraction experiments in scanning transmission electron microscopy (STEM)
- High frame rate development aims to improve scanning diffraction experiments and will be installed on the Transmission Electron Aberration-corrected Microscope (TEAM)
- Team of scientists and engineers from NCEM, NERSC and the Engineering Division of Lawrence Berkeley National Laboratory co-designed detector system to collect, transport and analyze STEM data in real time



100 KHz frame rate



Custom FPGA development for data transfer

Value Proposition

- Direct High Speed Data Transfer: NERSC internal Network has been extended to NCEM building and the detector is hooked up at 400Gbps
- Data receiver batch jobs on Cori will receive complete raw image sets from the microscope and pass them on to processes that conduct online analysis and store the data

Aug 2018

NCEM - 4D STEM Detector Development

Needs from NERSC

- **SDN + Advanced Scheduling to stream data into readily available load-balanced compute nodes.**
  - Data rates up to 360 Gbps when complete raw-data is send to NERSC
  - Cori for preprocessing and computation
  - Compute needs depend on the Bursty traffic rate
  - Detector Images are buffered in Burst Buffer and distributed to analysis jobs

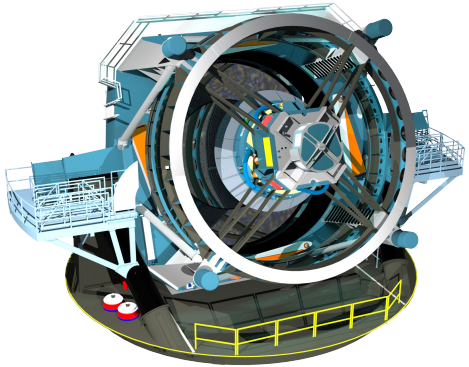
4D-STEM Milestones

- Spring 2019 - NCEM will start sending camera data to NERSC
- Summer 2019 - Computation on experiments selected for using 4D-STEM detector
- 2019 -2020 - Data compression to ~40Gbps using ML algorithms and/or other compression techniques developed on Cori

LSST Dark Energy Science Collaboration

Science story

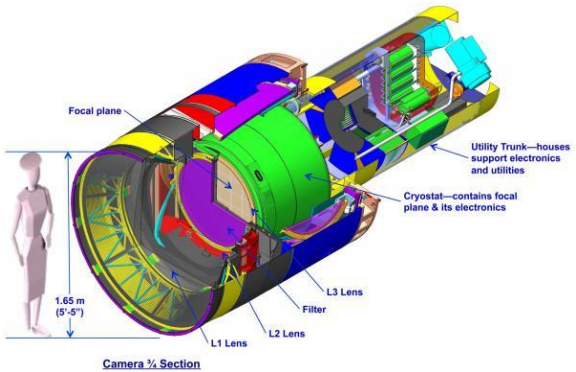
- Explain Dark Energy through multiple science probes: Galaxy catalogs, supernovae, lensing
- Survey covers the whole sky every few nights
- 3.2 Gpix camera built by DOE



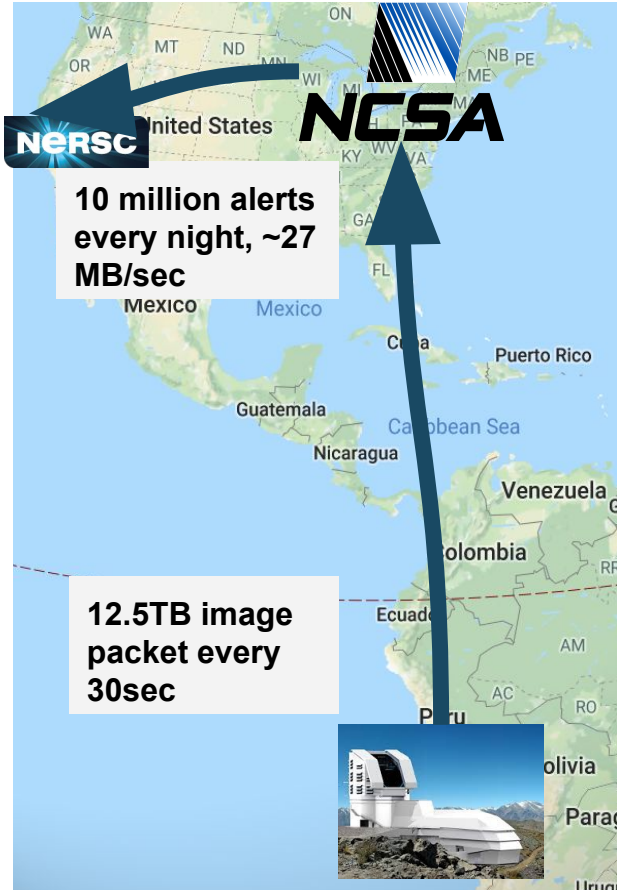
Value proposition

Ability to co-locate and combine data w/compute:

- Simulations: Cosmology, instrument, detector
- Non-LSST Data: Other surveys for context
- Data analysis (HPC) and Sharing (Globus, Spin)



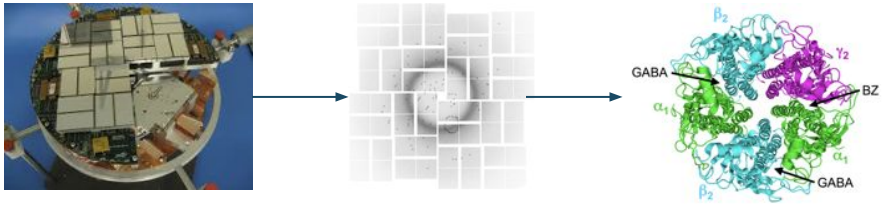
LSST Dark Energy Science Collaboration

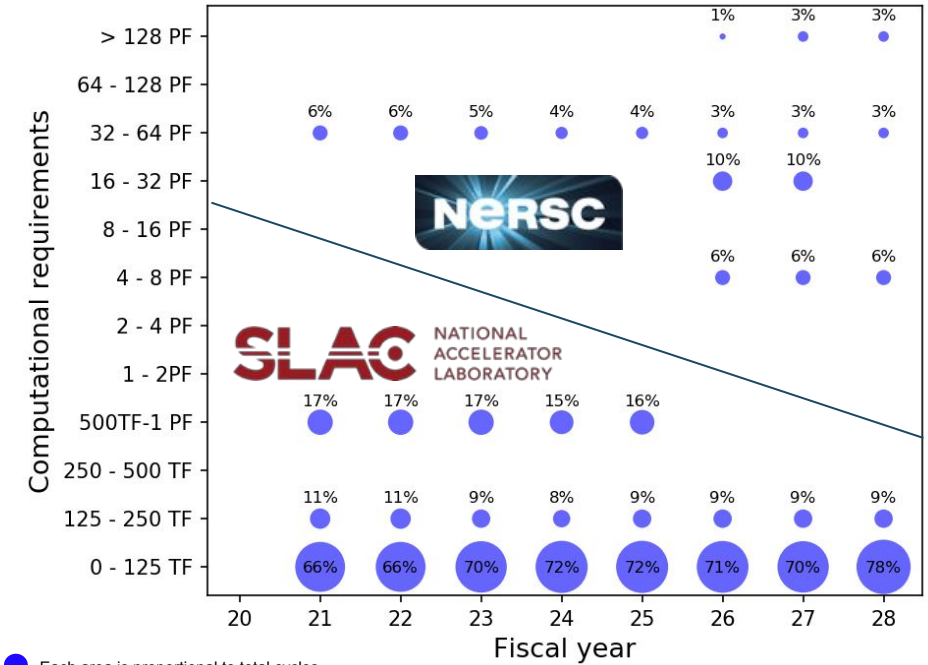


Needs from NERSC:

- **Spin for Supernova broker**
  - 10 million alerts/night
- **Cori for Simulations**
  - 138M MPP hours in 2019 (increasing annually), 3 month turnaround for sim campaign
  - NESAP support
  - 1.2PB project storage purchased, additional 1PB in FY19
- **Data Management to coordinate/share data**
  - Some simulations run at in2p3 hosted at NERSC
- **Jupyter for analysis**
  - Hundreds of scientists accessing notebooks

LCLS/ESnet/NERSC Collaboration





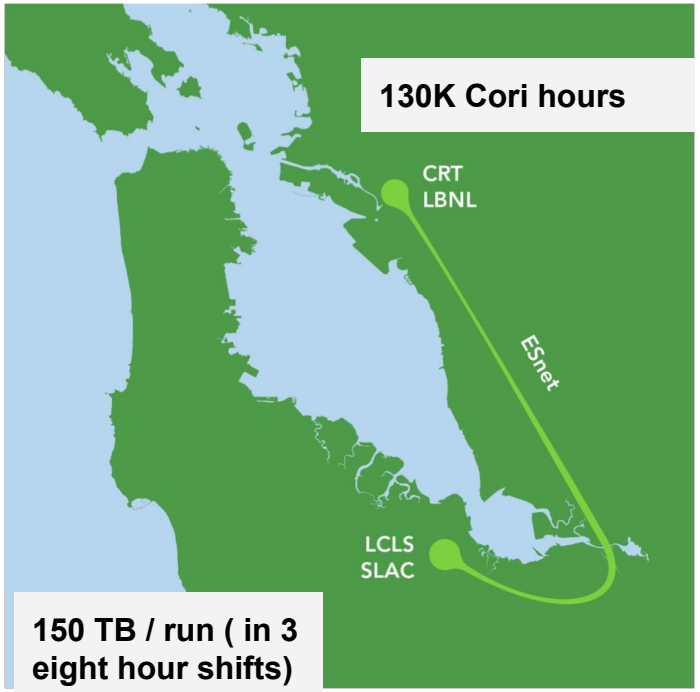
Science Summary:

- Streams of diffractive images reconstruct molecular structure and motion.
- Using HPC to speed data analysis allows on-shift understanding of collected data.

Value proposition:

- 20% of LCLS-II (2021-2028) experiments will require NERSC (dots above line)

LCLS/ESnet/NERSC Collaboration



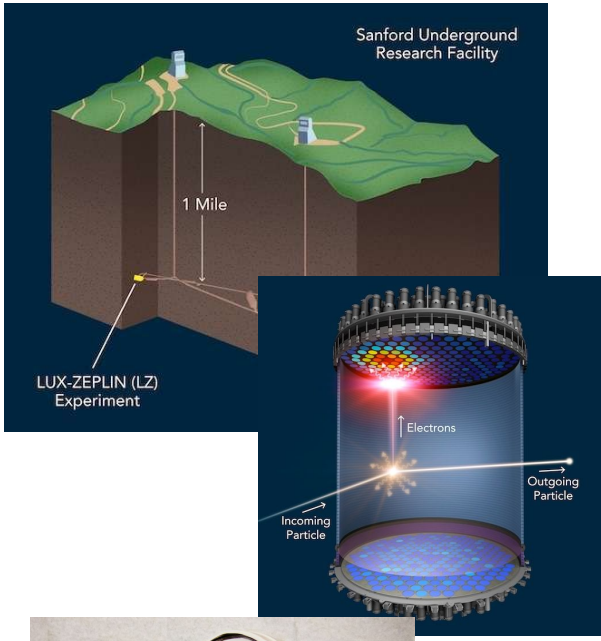
Needs from NERSC:

- **Spin for data transfer automation**
  - Reserve space and nodes for data
- **Cori for Data Analysis**
  - LCLS uses 130K hours per experiment
  - LCLS-II will 100x data rates
  - NERSC's ability to provide scheduled compute intensity is critical
- **GPUs for algorithm advancement**
- **WAN Bandwidth**
  - In cooperation with ESnet provide scheduled bandwidth to compute nodes.
  - Orchestrate NERSC and ESnet resources (SENSE, SDN, scheduling)

Detecting Dark Matter with LZ

Science Story

- Search for direct dark matter signals using a liquid Xenon target and dual-phase time projection chamber, located at the Sanford Underground Research Facility (SURF) in Lead, SD, USA.
- 3-year run will achieve a sensitivity close to the fundamental limits from cosmic ray neutrino background.

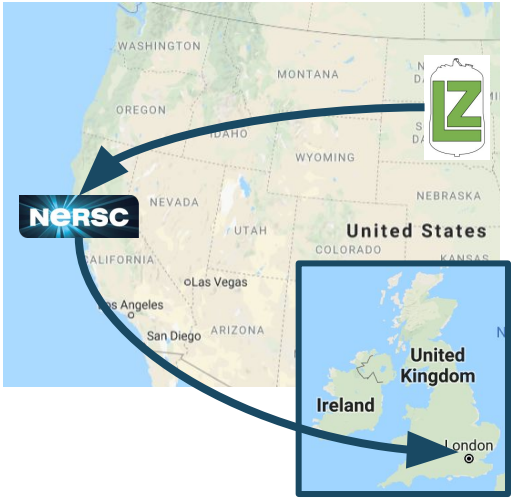


Value Proposition

- Real-time analysis of data (“prompt processing”) needed to flag potentially dangerous detector issues.
- Data archived at NERSC, regular re-processing required as analysis code is updated.
- All data and analysis mirrored at UK site.

U.S. DEPARTMENT OF ENERGY | Office of Science
BERKELEY LAB

Detecting Dark Matter with LZ



- *250 scientists, 36 institutions world-wide*
- *Operations start mid-2020*
- *Data moved from deep mine to surface before shipping to NERSC*

Needs from NERSC:

- **Advanced scheduling for real-time data analysis and experimental feedback**
  - Real-time queues ideal
- **Workflow resiliency for data flow**
  - When NERSC *storage* is down, data needs to stream to UK site for archiving, and synch back to NERSC afterwards.
- **Cori for simulations and data prep in 2019**
  - 13M MPP hours, 130TB required.
  - NESAP support to port code to KNL.
- **Data management**
  - Storing >PB/year data on tape, need seamless staging to project/scratch for processing.

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB