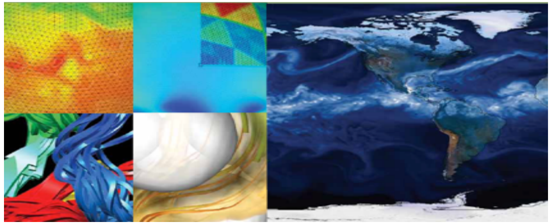
NERSC Director's Perspective

Sudip Dosanjh
Director, NERSC

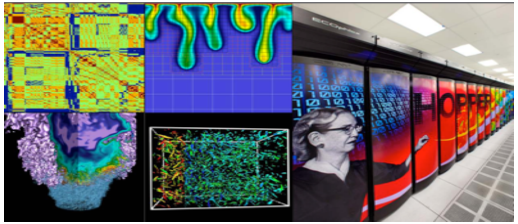
NERSC: the Mission HPC Facility for DOE Office of Science Research

U.S. DEPARTMENT OF ENERGY | Office of Science

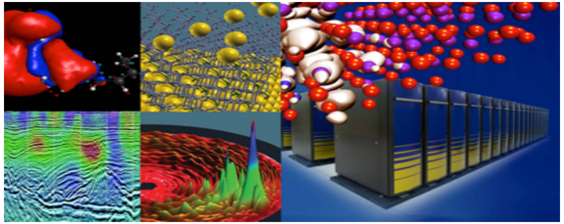
Largest funder of physical sciences research in the U.S.



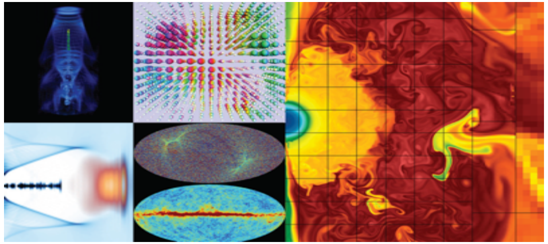
Bio Energy, Environment



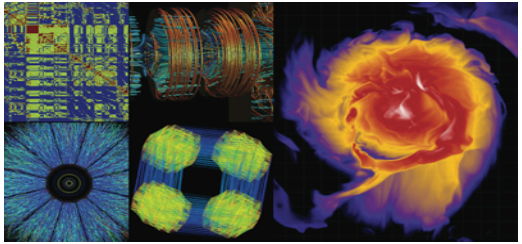
Computing



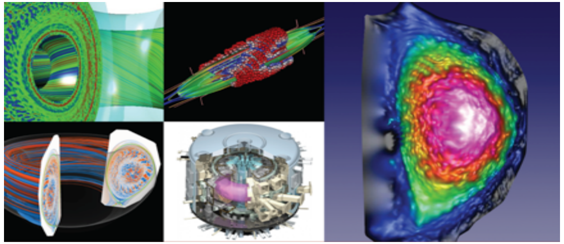
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics

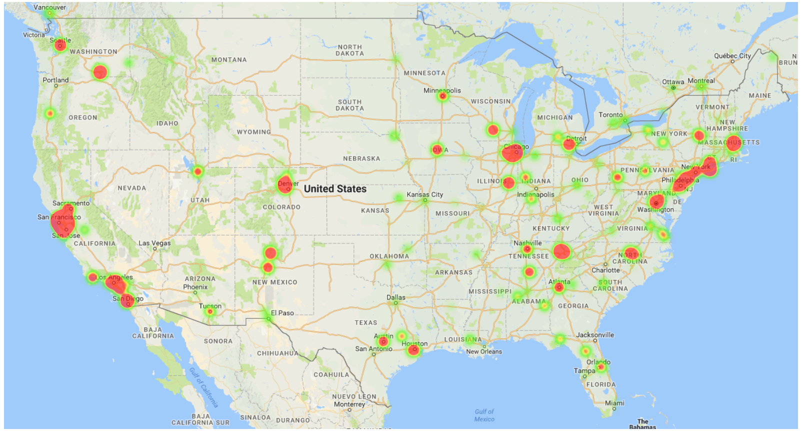


Nuclear Physics

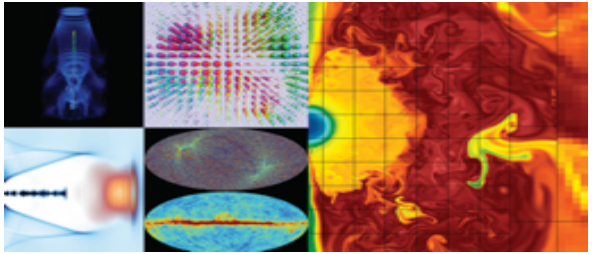


Fusion Energy, Plasma Physics

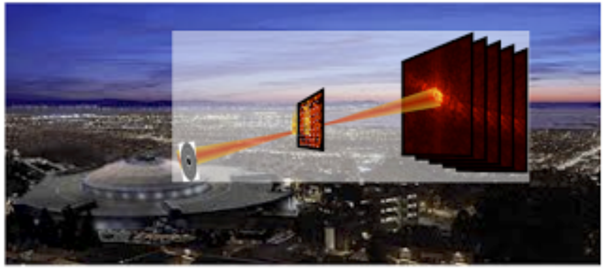
NERSC supports a broad user base



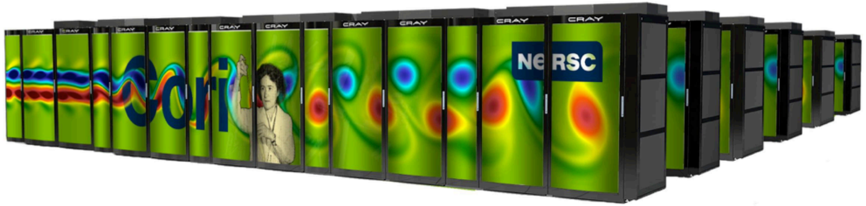
7,000 Users
800 Projects
700 Codes
~2000 publications per year



Simulations at scale



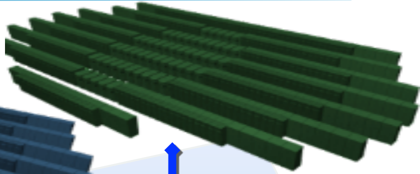
Data analysis support for DOE's experimental and observational facilities
Photo Credit: CAMERA

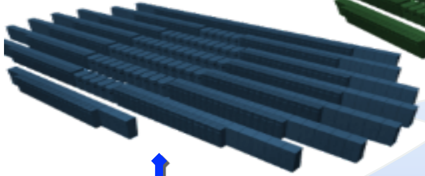


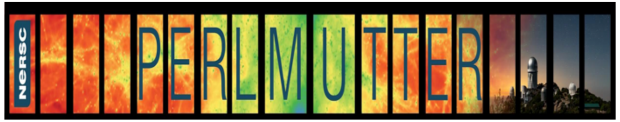
U.S. DEPARTMENT OF ENERGY

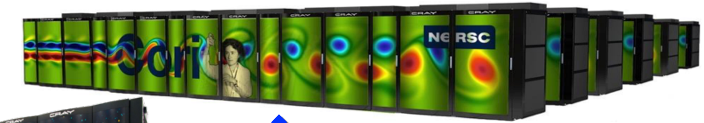
BERKELEY LAB

NERSC Systems Roadmap









NERSC-7:
Edison
Multicore CPU

2013

NERSC-8: Cori
Manycore CPU
NESAP Launched:
transition applications to advanced architectures

2016

NERSC-9:
CPU and GPU nodes
Continued transition of applications and support for complex workflows

2020

NERSC-10:
Exa system

2024

NERSC-11:
Beyond Moore

2028

What's changing?

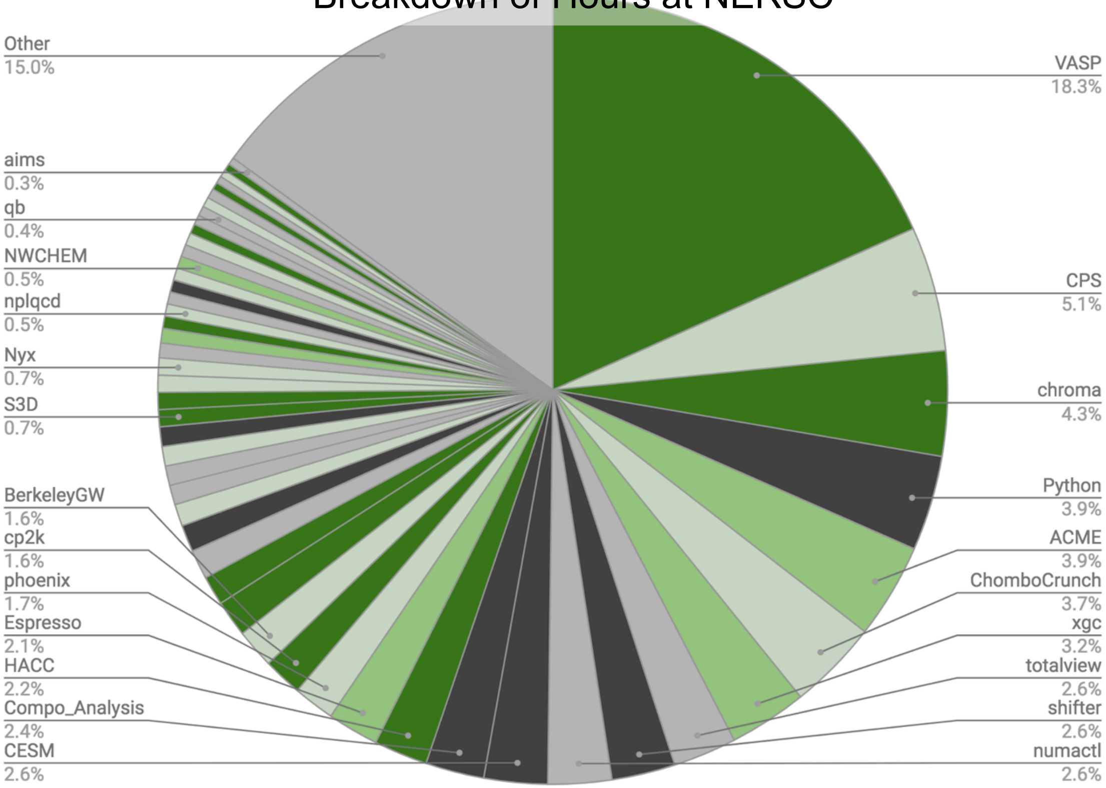
- Applications need to adapt to new systems
- Increasing engagement from Experimental and Observational Facilities
- More emphasis on data analysis, AI and Deep Learning

NERSC-9: A System Optimized for Science

- Cray Shasta System providing 3-4x capability of Cori system
- First NERSC system designed to meet needs of both large scale simulation and data analysis from experimental facilities
  - Includes both NVIDIA GPU-accelerated and AMD CPU-only nodes
  - Cray Slingshot high-performance network will support Terabit rate connections to system
  - Optimized data software stack enabling analytics and ML at scale
  - All-Flash filesystem for I/O acceleration
- Robust readiness program for simulation, data and learning applications and complex workflows
- Delivery in late 2020

GPU Readiness Among NERSC Codes

Breakdown of Hours at NERSC



GPU Status & Description	Fraction
Enabled: Most features are ported and performant	32%
Kernels: Ports of some kernels have been documented.	10%
Proxy: Kernels in related codes have been ported	19%
Unlikely: A GPU port would require major effort.	14%
Unknown: GPU readiness cannot be assessed at this time.	25%

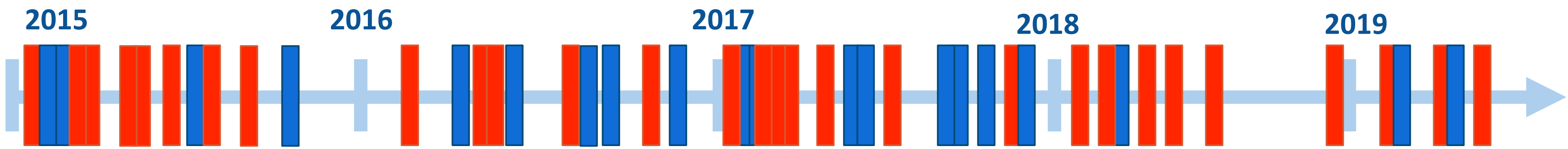
A number of applications in NERSC workload are GPU enabled already.

We will leverage existing GPU codes from CAAR + Community

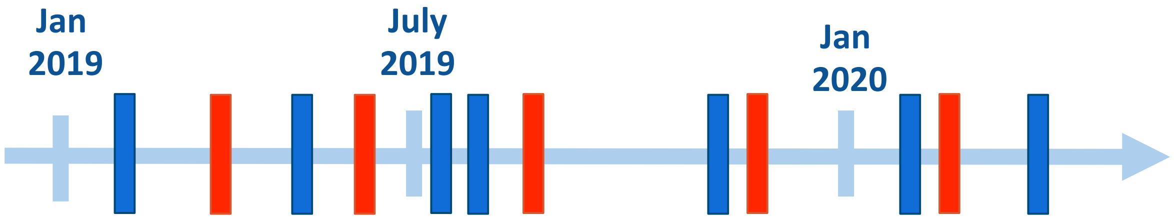
Hack-A-Thons and App Perf. Training

Community Training
Hack-a-Thon

KNL Targeted Events



GPU Targeted Events



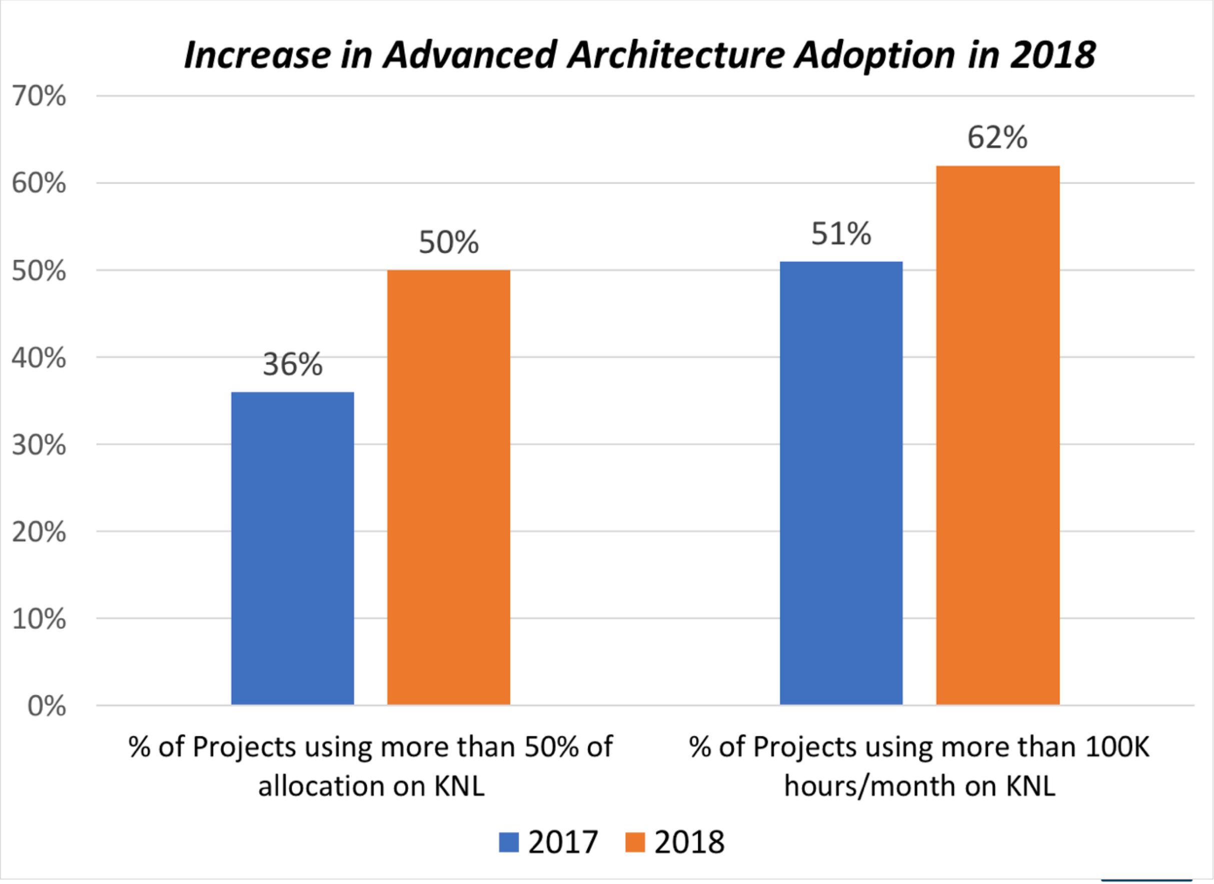
NESAP for Perlmutter

Simulation 12 Apps	Data Analysis 8 Apps	Learning 5 Apps

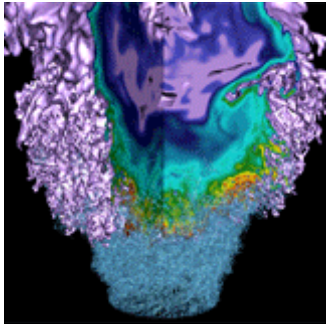
- 5 ECP Apps Jointly Selected (Participation Funded by ECP)
- 20 additional teams selected through Open call for proposals.
  - https://www.nersc.gov/users/application-performance/nesap/nesap-projects/
- Access to Cori GPU rack for application readiness efforts.

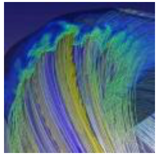
Transition of the entire NERSC workload to advanced architectures

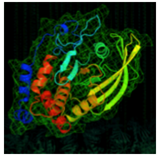
To effectively use Cori KNL, users must exploit parallelism, manage data locality and utilize longer vector units. All features that will be present on exascale era systems

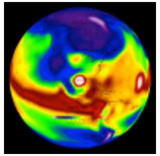


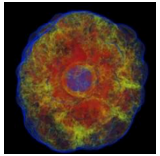
Engagement with Experiments

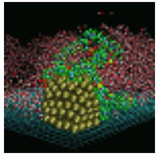








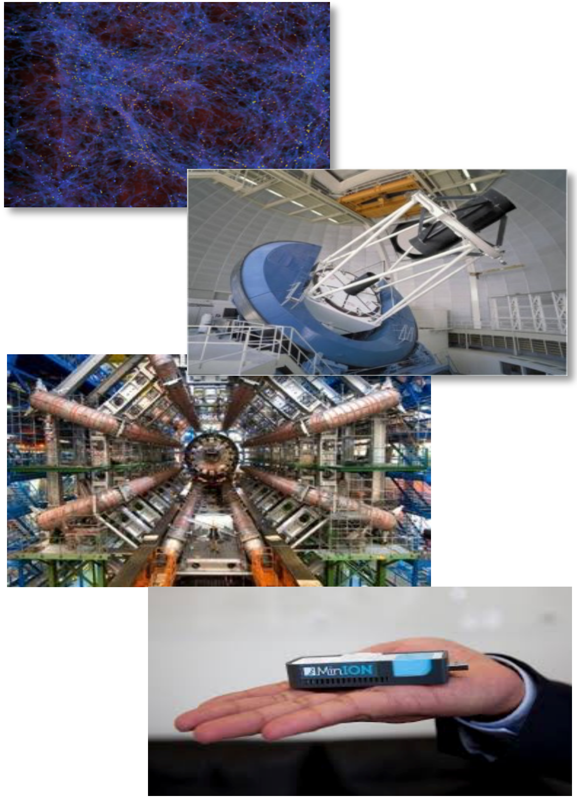




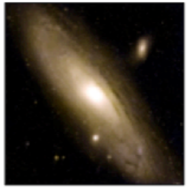
NERSC

What's different?

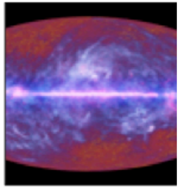
- **Proliferation of data from DOE user facilities**
- **Scientific workflows have become more complex**
  - Streaming data to HPC facilities
  - Real-time/Interactive access
  - Rich 'Data' stack
- **Important scientific problems are requiring both simulation and data analytics**
  - Advanced Machine Learning and Statistics methods + tools required



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

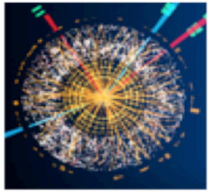


Palomar Transient Factory Supernova

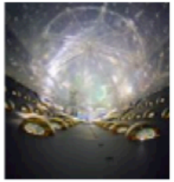


Planck Satellite Cosmic Microwave Background Radiation

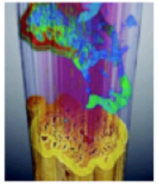
Star Particle Physics



Atlas Large Hadron Collide

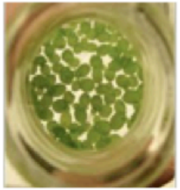


Dayabay Neutrinos

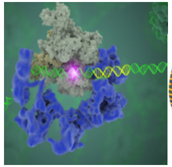


ALS Light Source

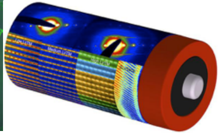
LCLS Light Source



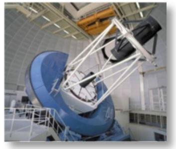
Joint Genome Institute Bioinformatics



Cryo-EM

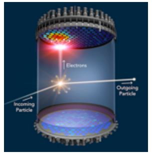


NCEM

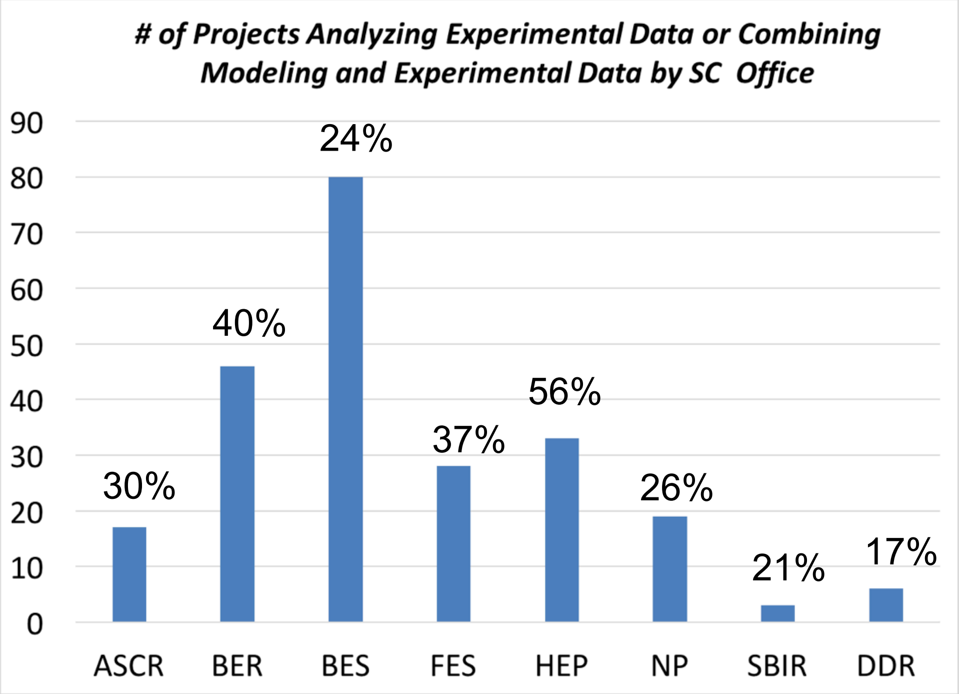


DESI

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

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Requirements reviews and users from experimental facilities describe numerous pain points

- **Workflows** require manual intervention and custom implementations
- Difficult to surge experimental pipelines at HPC facility in '**real-time**'
- I/O performance, storage space and access methods for **large datasets** remain a challenge
- Searching, publishing and sharing **data** are difficult
- **Analysis codes** need to be adapted to advanced architectures
- Lack of **scalable analytics software**

Research

- **Resilience strategy** needed for fast-turnaround analysis needs
  - including: coordinating maintenances, fault tolerant pipelines, rolling upgrades, alternative compute facilities...
- No **federated identity** between experimental facilities and NERSC
- Not all scientists want command-line access.

Policy

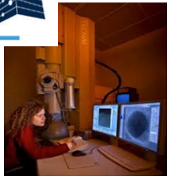
Science Engagements

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

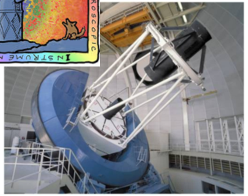
Processing streaming alerts (from NCSA) for detection of supernova and transient gravitational lensing events

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

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



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

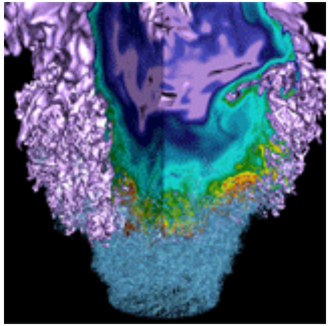


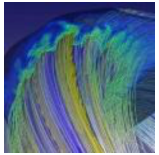
Nightly processing of galaxy spectra to inform next night's telescope targets

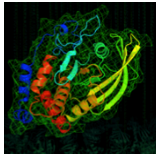
Needs from NERSC

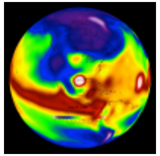
Experiment	What runs at NERSC?	What runs elsewhere?
LCLS	5-10% of experiments that require >32PF compute in 2021 (~3% >128PF in 2027)	All other experiments at LCLS
ALS	2-3 beamlines with large computing requirements, i.e. tomography and ptychography (~200MB/s)	Other ~40 ALS beamlines
NCEM	Stream super 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

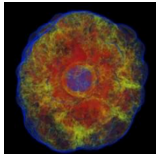
Data, AI and Deep Learning

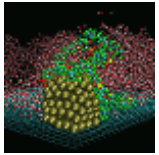








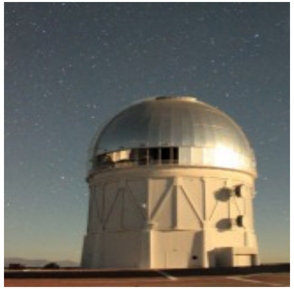




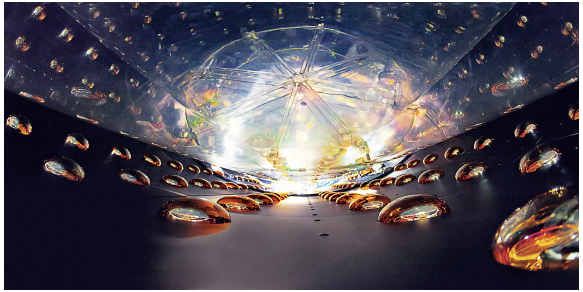
NERSC

	HEP			BER		BES		NP	FES
	Astronomy	Cosmology	Particle Physics	Climate	Genomics	Light Sources	Materials	Heavy Ion Colliders	Plasma Physics
Classification	✗		✗	✗	✗	✗	✗	✗	✗
Regression		✗			✗	✗	✗	✗	✗
Clustering		✗	✗	✗	✗	✗	✗	✗	✗
Dimensionality Reduction				✗				✗	
Surrogate Models	✗	✗	✗				✗	✗	✗
Design of Experiments		✗		✗			✗		✗
Feature Learning	✗	✗	✗	✗	✗	✗	✗	✗	✗
Anomaly Detection	✗		✗	✗		✗		✗	

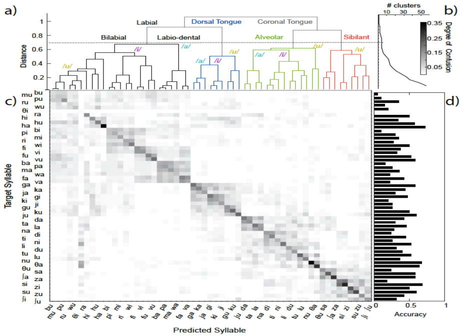
Deep Learning for Science

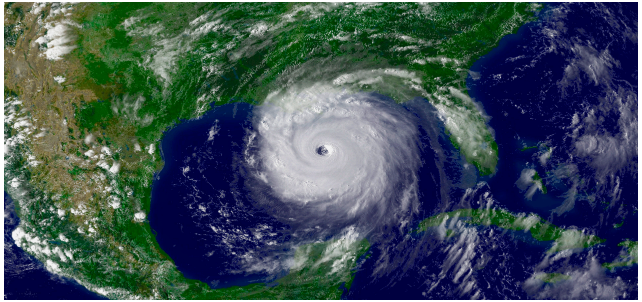
Modeling galaxy shapes



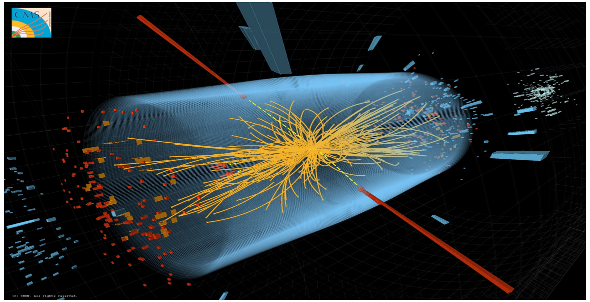
Clustering Daya Bay events



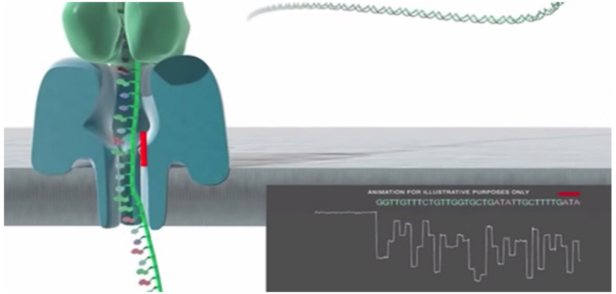
Decoding speech from ECoG



Detecting extreme weather

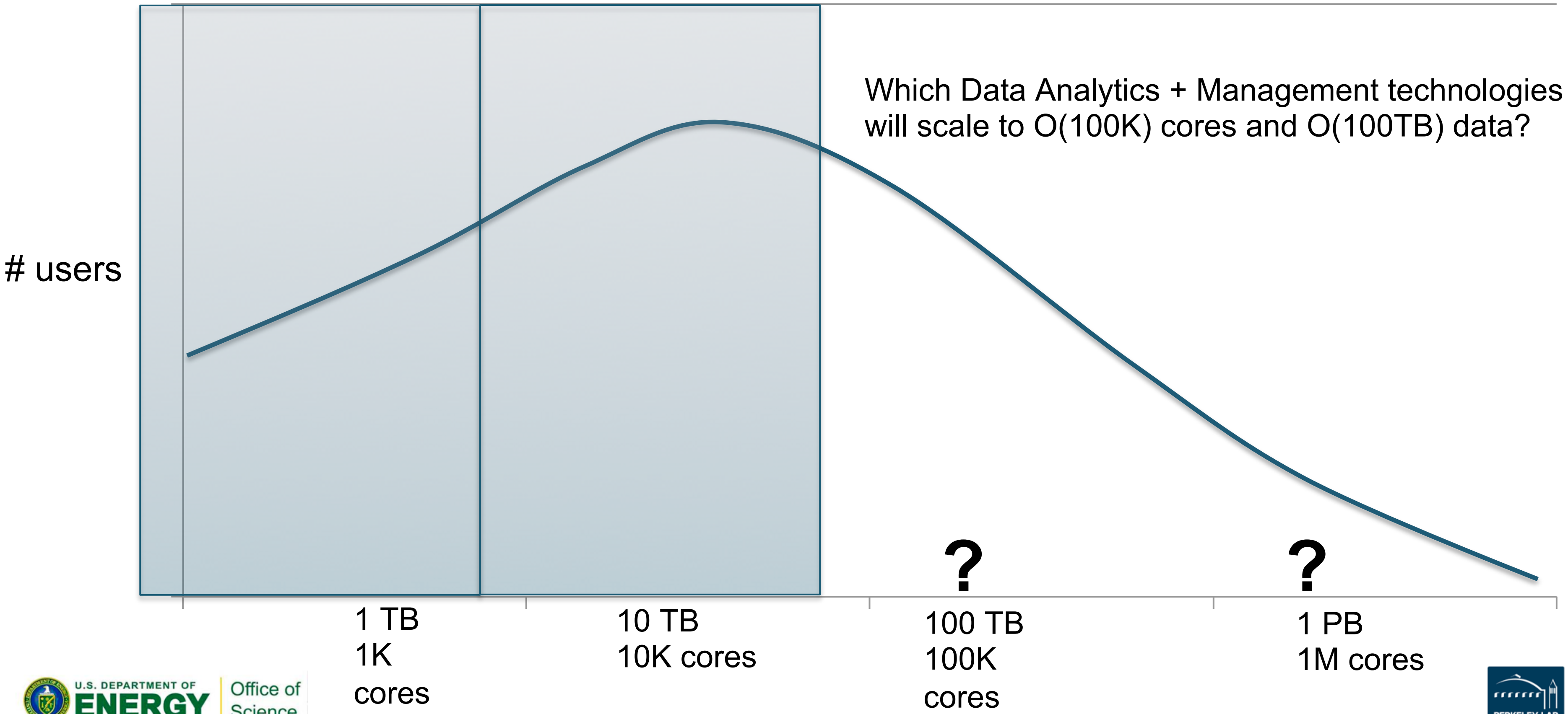


Classifying LHC events



Oxford Nanopore sequencing

Big Data Center collaboration with Cray and Intel



Enabling Precision Analytics for Climate Science

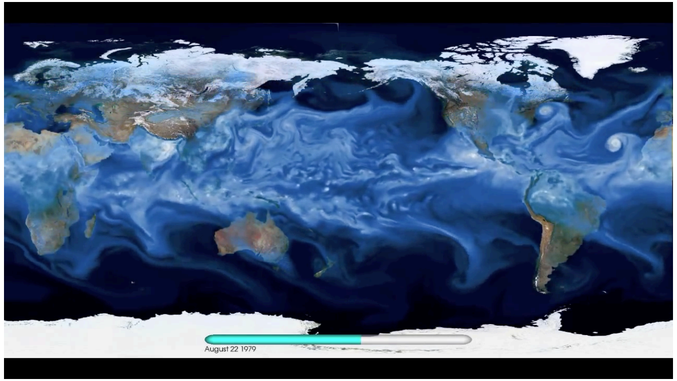
Climate projections employ coarse analytics on high fidelity model output

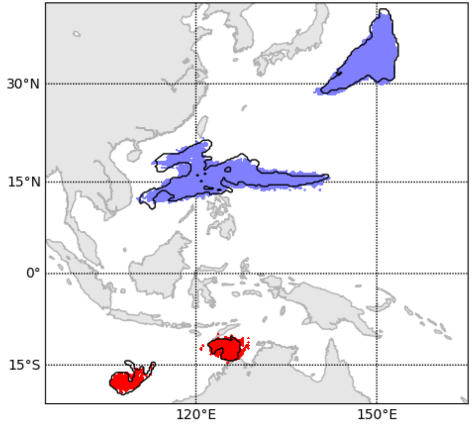
- Mean temperature increase, sea level rise
- Characterizing impact of extreme weather requires precision analytics

Our Goal

- Analyze complex datasets with high spatio-temporal resolution
- Extract high quality segmentation masks leveraging state-of-the-art Deep Learning models (Deeplabv3+)

Gordon Bell Prize for achieving over an Exaflop (FP16) on Summit

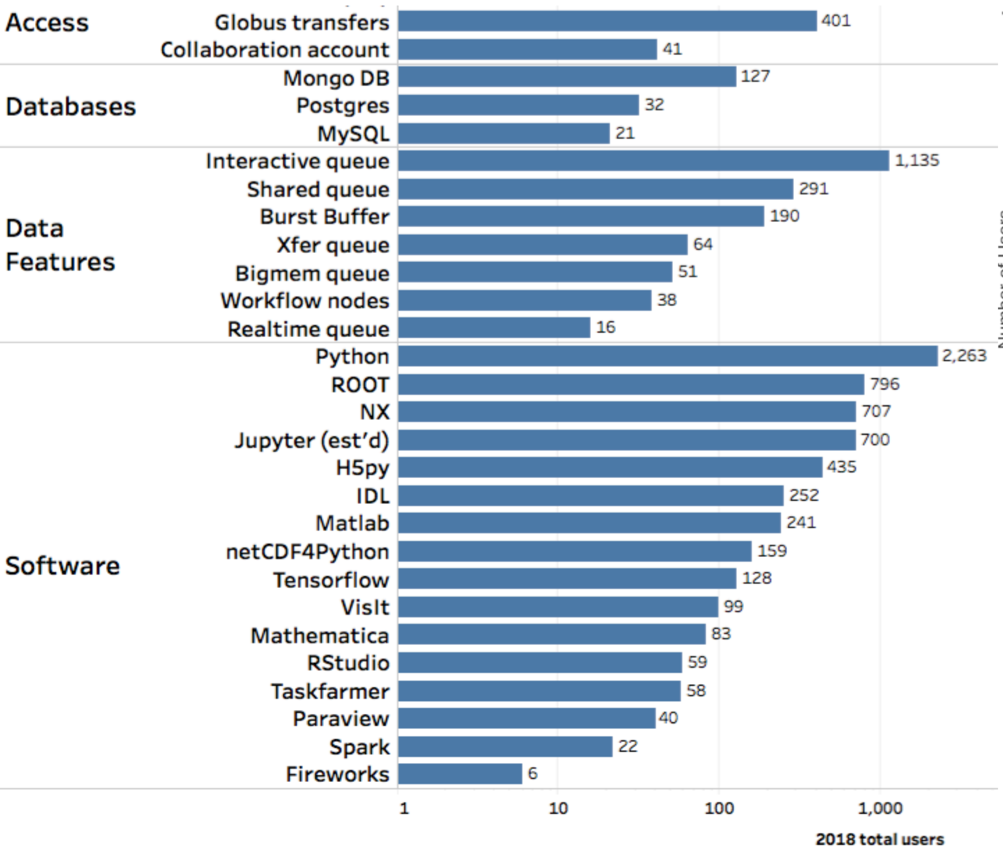


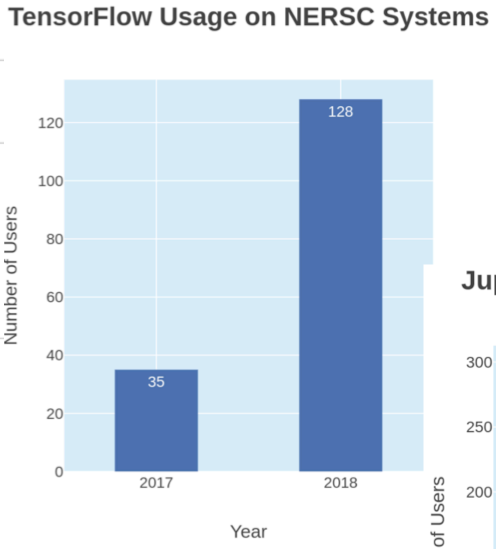


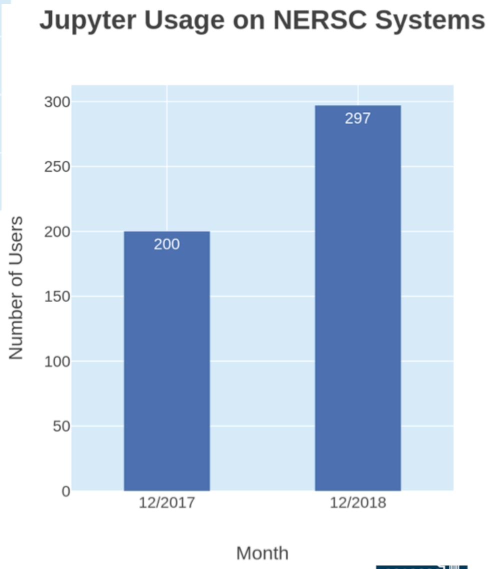
NERSC Big Data Stack

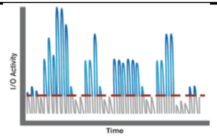
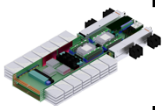
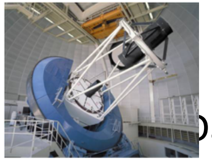
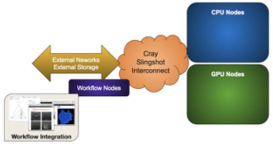
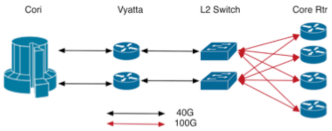
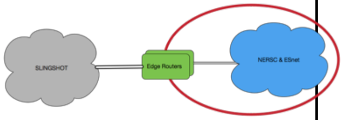
Capabilities	Technologies
Data Transfer + Access	globus online, GridFTP, jupyter, python django, newt
Workflows	FireWorks, taskfarmer
Data Management	HDF, netCDF, ROOT Data Analysis Framework, mongoDB, MySQL, PostgreSQL
Data Analytics	python, Spark, TensorFlow, PYTORCH, R, julia, Caffe, K
Data Visualization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

Strong Adoption of Data Software Stack







Data Features	Cori experience	N9 enhancements
I/O and Storage	Burst Buffer 	All-flash file system: performance with ease of data management 
Analytics - **Production stacks** - **Analytics libraries** - **Machine learning**	User defined images with Shifter NESAP for data New analytics and ML libraries	 Benchmark Production Analytics workflows. Data apps in NESAP at outset Optimised analytics libraries and deep learning application benchmarks
Workflow integration	SchedMD Real-time queues	 SLURM co-scheduling Workflow nodes integrated
Data transfer and streaming	SDN 	Slingshot ethernet-based converged fabric 

NERSC Community Engagements

Jupyter Community Workshop
June 11-13, 2019 • NERSC and Berkeley Institute for Data Science, Berkeley, CA

GPUs for Science Day
July 2-3, 2019 • NERSC, Berkeley, CA

Deep Learning for Science Summer School
July 15-19, 2019 • NERSC, Berkeley, CA

Monterey Data Conference
August 5-8, 2019 • Monterey Marriott, Monterey, CA

Tutorials Lead at SC18, ECP19, GTC19 and ISC19:

Parallel I/O in Practice (SC18)
Exascale I/O Technologies (ECP19)
Deep Learning at Scale (SC18, ECP19, ISC19)
Getting Started with Containers on HPC (ISC19)
OpenMP Common Core: a "Hands-On" Exploration (SC18)
Managing HPC Software Complexity with Spack (SC18, ECP19)
Container Computing for HPC and Scientific Workflows (SC18, ECP19)
Performance Tuning of Scientific Codes with the Roofline Model (SC18, ECP19, ISC19, GTC19)

Conclusions

- **We are very excited about Perlmutter**
  - Significant increase in capabilities over Cori
  - NESAP is a key to broad adoption of GPU technologies
  - 1st NERSC system designed with data in mind from the very beginning
    - All flash file system, new interconnect, big data stack
- **Demand from Experimental and Observational Facilities is increasing dramatically**
- **NERSC has made a significant investment in data, AI and deep learning**

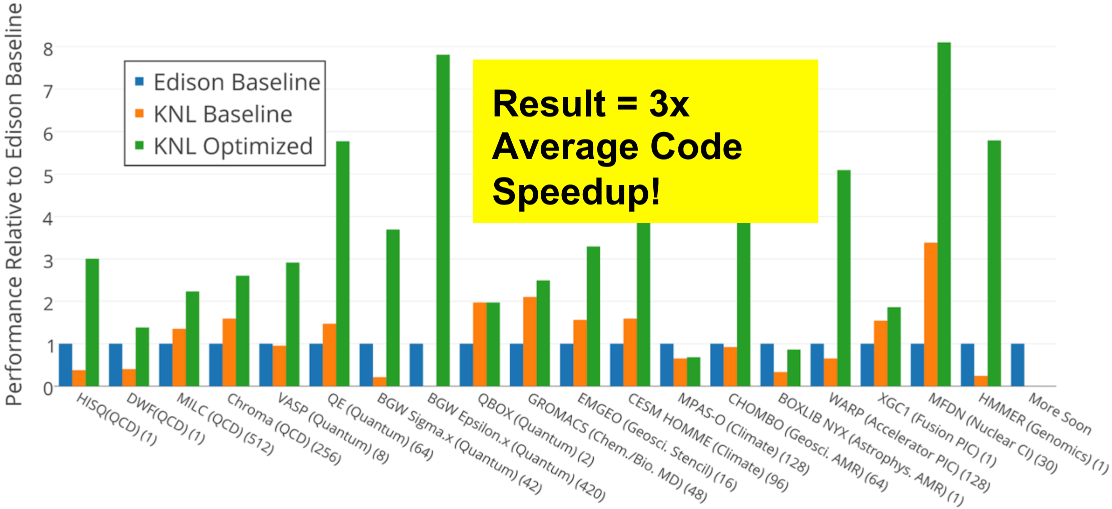
Questions?

Cori: Pre-Exascale System for DOE Science

- **Cray XC System - heterogeneous compute architecture**
- **9600 Intel KNL compute nodes**
  - 68 cores, 4 HW treads/core, AVX-512, 16GB HBM, 96GB DRAM
- **>2000 Intel Haswell nodes**
- **Cray Aries Interconnect**
- **NVRAM Burst Buffer, 1.6PB and 1.7TB/sec**
- **Lustre file system 28 PB of disk, >700 GB/sec**
- **Investments to support large scale data analysıs**
  - High bandwidth external connectivity to experimental facilities from compute nodes
  - Virtualization capabilities (Shifter/Docker)
  - More login nodes for managing advanced workflows
  - Support for real time and high-throughput queues

NERSC Exascale Scientific Application Program (NESAP)

- Prepare DOE SC users for advanced architectures like Cori
- Partner closely with ~20 application teams and apply lessons learned to broad NERSC user community.
- >½ of projects have used > ½ their time on Cori KNL



NESAP ECP Engagement

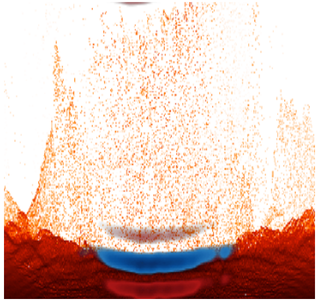
NESAP For Cori Included the Following ECP Apps

LatticeQCD	***WDMApp***	E3SM
NWChemEX	***WarpX***	Urban
GAMESS	Exastar	ExaSGD
EXAALT	Exasky	CANDLE
ExaAm	EQSIM	***ExaFEL***
QMCPACK	SubSurface	ExaSMR
Combustion	***Exabiome***	MFIX

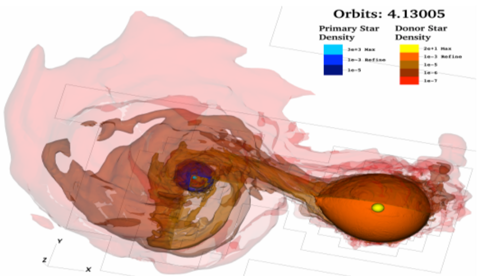
ECP NESAP for Cori Codes
ECP Codes Used at NERSC

EXAALT, WDMApp, WaprX, Exabiome, ExaFEL are Pre-Selected for NESAP for Perlmutter. More apps to be announced Feb. 2019.

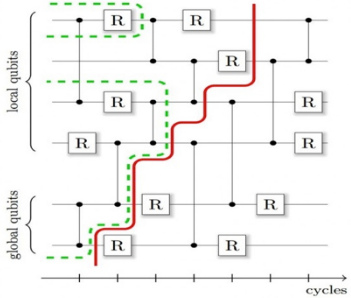
Users Demonstrate Groundbreaking Science Capability



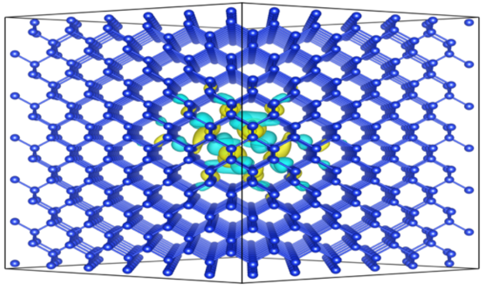
Large Scale Particle in Cell Plasma Simulations



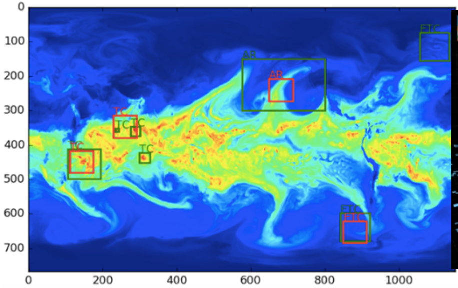
Stellar Merger Simulations with Task Based Programming

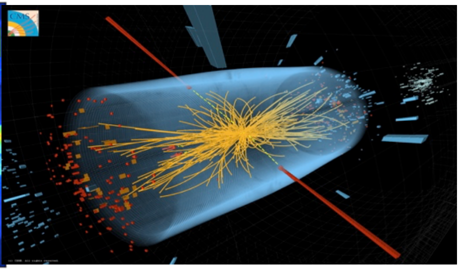


Largest Ever Quantum Circuit Simulation



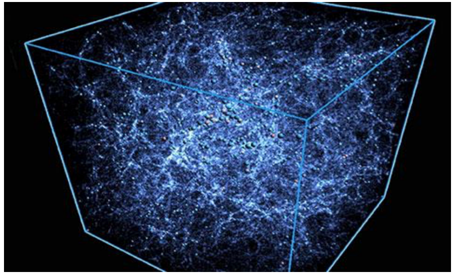
Largest Ever Defect Calculation from Many Body Perturbation Theory > 10PF





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

U.S. DEPARTMENT OF ENERGY | Office of Science

608 M hours provided to ECP teams in 2018

BERKELEY LAB

Compute Node Details

- **CPU only nodes**
  - Next Generation AMD CPUs
  - CPU only cabinets will provide approximately same capability as ***full*** Cori system (~8B hours) > 4000 nodes
  - Efforts to optimize codes for KNL will translate to NERSC-9 CPU only nodes
- **CPU + GPU nodes**
  - Next Generation NVIDIA GPUs with Tensor cores, high bandwidth memory and NVLINK-3
  - Unified Virtual Memory for improved programmability
  - 4 to 1 GPU to CPU ratio
  - (> 16B hours)

Slingshot Network

Commodity based

- Ethernet physical layer
  - Native support for generic Linux Sockets
- Traffic Class (QoS levels)
- Converged fabric

Optimized for HPC

- Low latency
- MPI and collective offload
- Congestion control
- Adaptive routing

Commodity + HPC

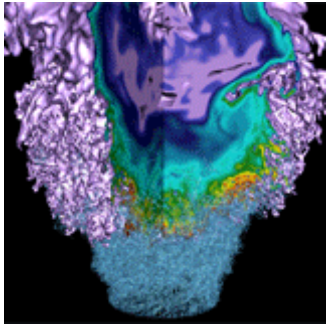
- 64-port Rosetta Switches
- 25GBps Cray NIC (200Gbps)

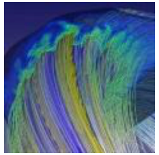
NESAP for Perlmutter

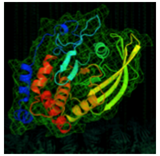
Simulation ~12 Apps	Data Analysis ~8 Apps	Learning ~5 Apps

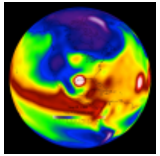
- 5 ECP Apps Jointly Selected (Participation Funded by ECP)
- Open call for proposals.
  - **App selection will contain multiple applications from each SC Office and algorithm area**
  - **Additional applications (beyond 25) will be selected for second tier NESAP with access to vendor/training resources and early access**

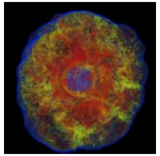
Workflows and Data Analytics

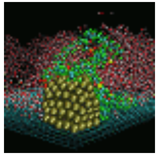












NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

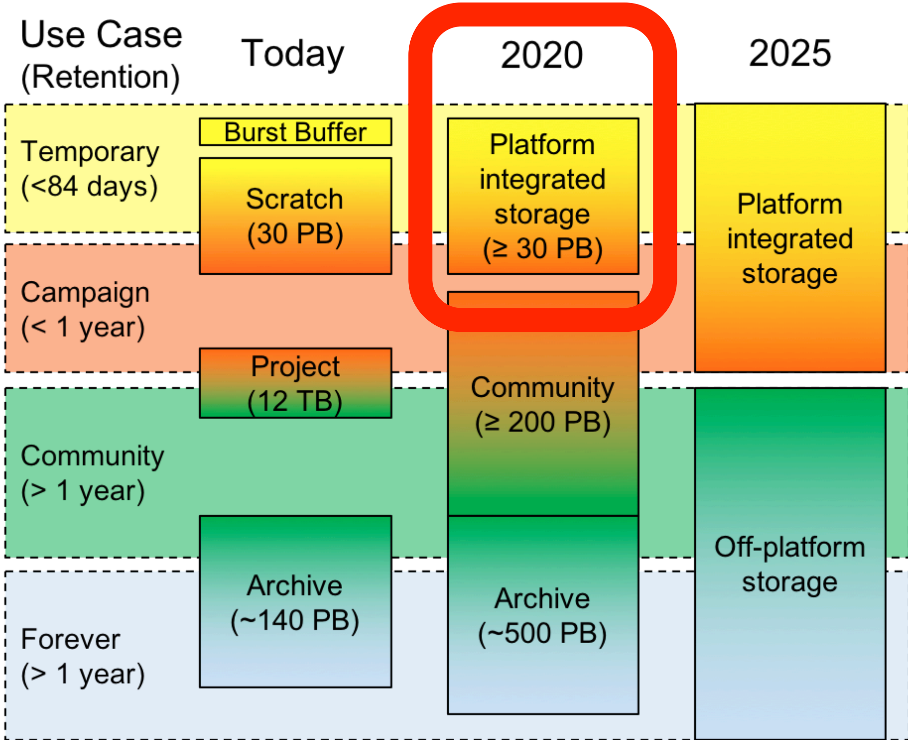
BERKELEY LAB

NERSC-9: A System Optimized for Science

- Cray Shasta System providing 3-4x capability of Cori system
- First NERSC system designed to meet needs of both large scale simulation and data analysis from experimental facilities
  - Includes both NVIDIA GPU-accelerated and AMD CPU-only nodes
  - Cray Slingshot high-performance network will support Terabit rate connections to system
  - Optimized data software stack enabling analytics and ML at scale
  - All-Flash filesystem for I/O acceleration
- Robust readiness program for simulation, data and learning applications and complex workflows
- Delivery in late 2020

Platform Storage System Design Goals

- **Meet the needs of users**
  - Support high IOPS & metadata rates for data analysis
  - Retain benefits of burst buffer
  - Collapse performance tiers
- **Prepare for the future**
  - Prove Lustre optimizations for all-flash performance world



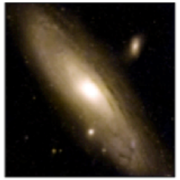
"Storage 2020: A Vision for the Future of HPC Storage," Berkeley, CA, 2017. Available online: https://escholarship.org/uc/item/744479dp

NESAP for Perlmutter

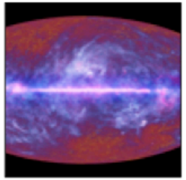
Simulation ~12 Apps	Data Analysis ~8 Apps	Learning ~5 Apps

- 5 ECP Apps Jointly Selected (Participation Funded by ECP)
- Open call for proposals.
  - **App selection will contain multiple applications from each SC Office and algorithm area**
  - **Additional applications (beyond 25) will be selected for second tier NESAP with access to vendor/training resources and early access**

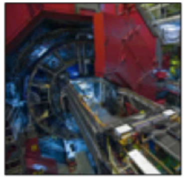
NERSC already supports a large number of users and projects from DOE SC's experimental and observational facilities



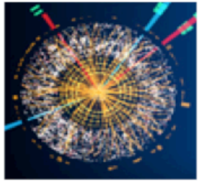
Palomar Transient Factory Supernova



Planck Satellite Cosmic Microwave Background Radiation

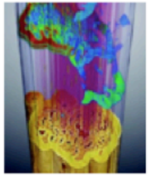


Alice Large Hadron Collider

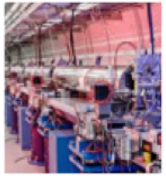


Atlas Large Hadron Collider

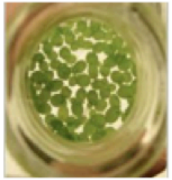
Dayabay Neutrinos



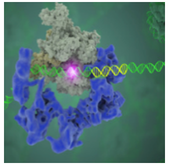
ALS Light Source



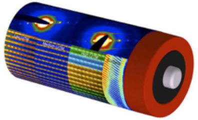
LCLS Light Source



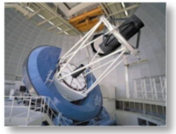
Joint Genome Institute Bioinformatics



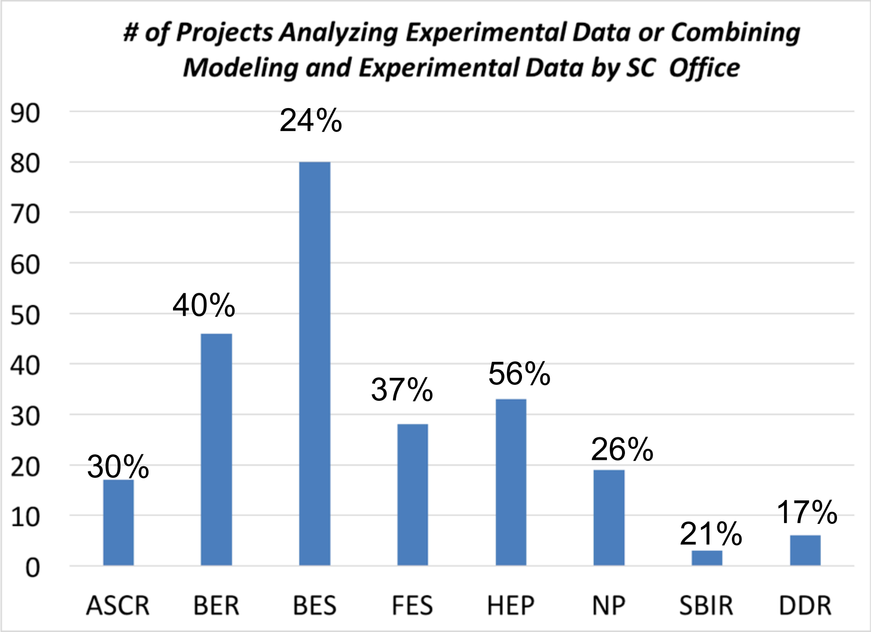
Cryo-EM



NCEM



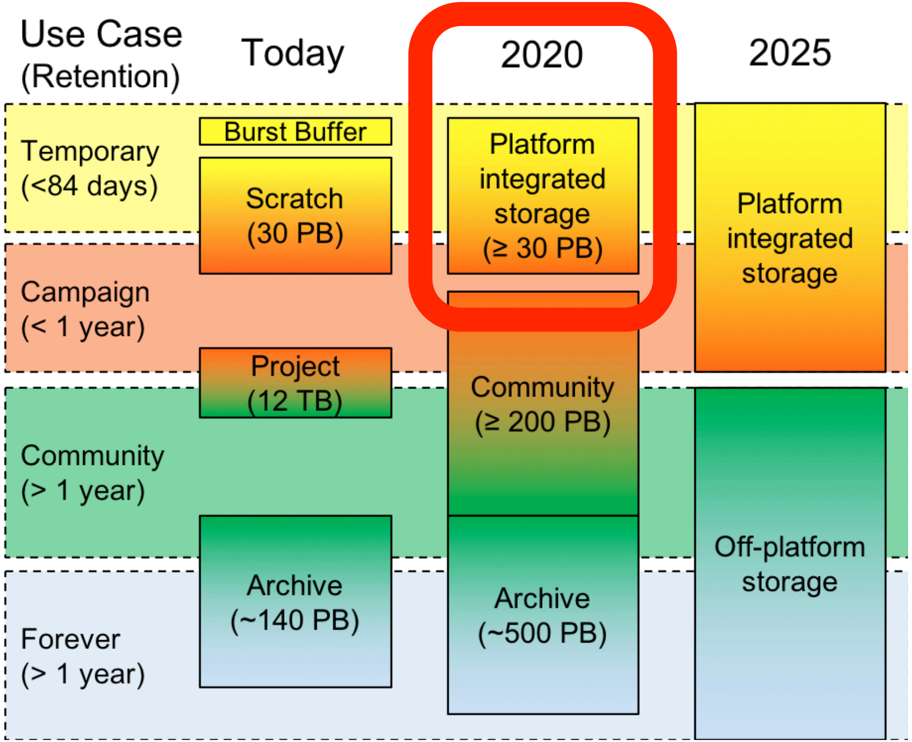
DESI



~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

Platform Storage System Design Goals

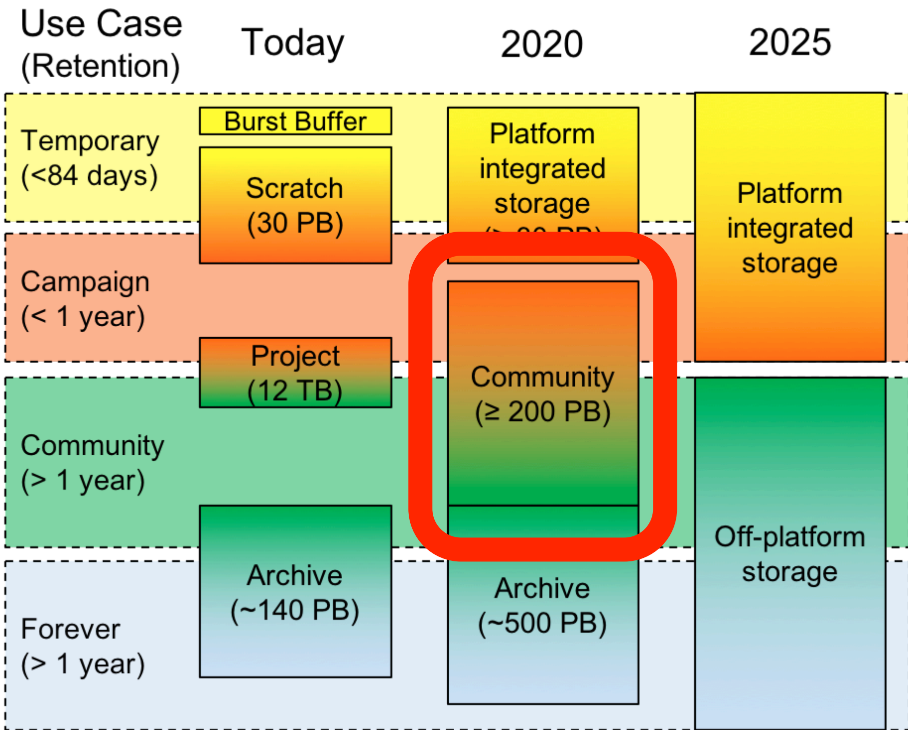
- **Meet the needs of users**
  - Support high IOPS & metadata rates for data analysis
  - Retain benefits of burst buffer
  - Collapse performance tiers
- **Prepare for the future**
  - Prove Lustre optimizations for all-flash performance world



"Storage 2020: A Vision for the Future of HPC Storage," Berkeley, CA, 2017. Available online: https://escholarship.org/uc/item/744479dp

Community File System: High-capacity storage

- **NERSC-9 scratch for hot, short-term data**
- **Community file system for cooler, longer-term data**
  - Replace project file system
  - ~50 PB initially
  - Grow to > 200 PB for N9
  - Emphasis on data management, access, and sharing



"Storage 2020: A Vision for the Future of HPC Storage," Berkeley, CA, 2017.
Available online: https://escholarship.org/uc/item/744479dp

NERSC Big Data Stack

NERSC

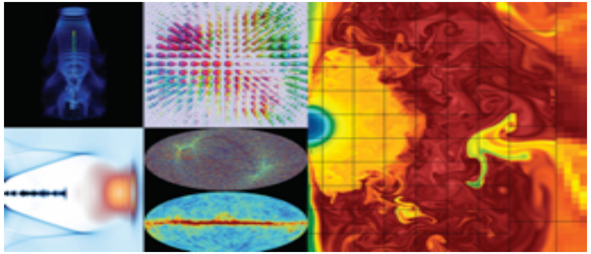
Capabilities	Technologies
Data Transfer + Access	globus online, GridFTP, jupyter, python django, newt
Workflows	FireWorks, taskfarmer
Data Management	HDF, netCDF, ROOT Data Analysis Framework, mongoDB, MySQL, PostgreSQL
Data Analytics	python, Spark, TensorFlow, PYTORCH, R, julia, Caffe, theano
Data Visualization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

NERSC-9 System Deployment Timeline

Milestone	Date
NESAP Call for Proposals Due	Dec. 2018
GPU Rack on Cori available for NESAP Users	Dec. 2018
NERSC-9 System Delivery	Oct. 2020
System Integration with NERSC Complete	Dec. 2020
Acceptance Testing Begins	Dec. 2020
NESAP Teams on NERSC-9 System	Jan. 2021
All users enabled on NERSC-9 System	Apr. 2021
System Acceptance	Aug. 2021

NERSC-9 System Optimized for Simulation and Data Users

In order to meet science requirements and demonstrate exascale-era technologies accelerator technology was essential



- **Significant fraction of NERSC workload can now use GPUs**
  - GPU programming has matured
  - Improved software environment
  - Increases in GPU memory capacity improve programmability
- **System contains large number of CPU-only nodes for applications that are not yet ready**

System designed to meet needs of data analysis from experimental facilities

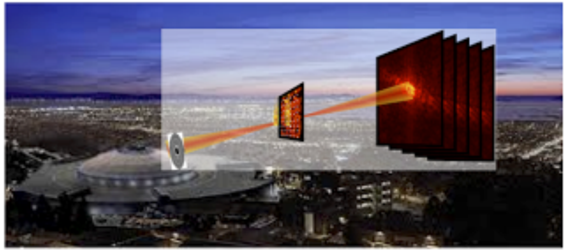
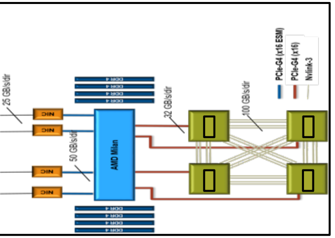


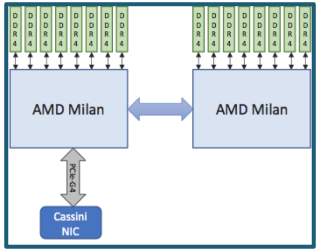
Photo Credit: CAMERA

- **System well balanced between network and FLOPS**
- **Optimized network for data ingest from experimental facilities**
- **Real-time scheduling capabilities**
- **Supported analytics stack including latest ML/DL software**
- **System software supporting rolling upgrades for improved resilience**

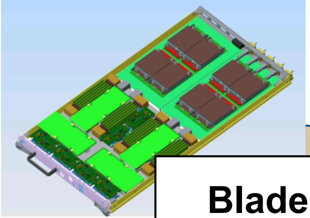
NERSC-9 at a glance

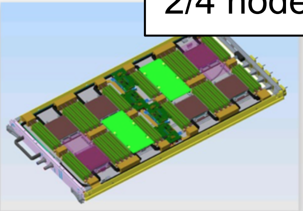


GPU Node
1 CPU: 4 GPU
>30 TF
>128 GiB DDR
>128 GiB HBM
4 NICs

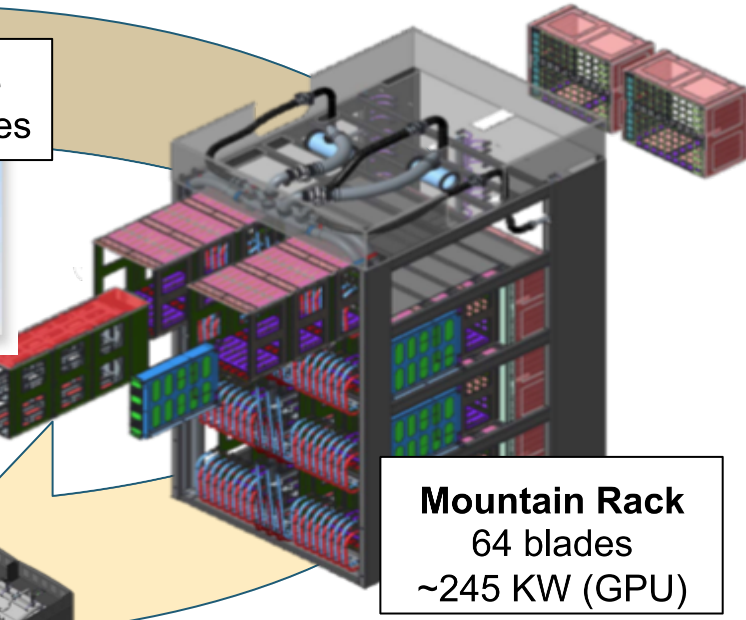


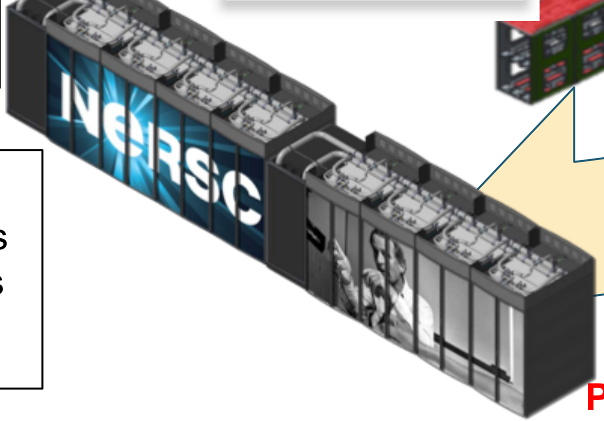
CPU Node
2 CPU
~4 TF
>256 GiB DDR
1 NIC





Blade
2/4 nodes





Mountain Rack
64 blades
~245 KW (GPU)

System
X CPU racks
Y GPU racks
3-4x Cori

External Network Connectivity

A new approach for System-to-Center integration

- Ethernet compatible Slingshot allows for seamless connectivity
  - blurs the line of what is “inside” or “outside” the system
- Slingshot on Rosetta Switches to Ethernet on Edge Routers
- Using a pair of Edge Routers with 400Gb ports
  - 1.9 Tb/s sustained bandwidth across a converged fabric
  - ~4x increase in bandwidth from what is available to Cori
  - provides integration point for N9, the Center and the WAN

Transitioning From KNL to AMD Processors

Codes optimized on Xeon Phi (KNL) will run well on Perlmutter

Many KNL architecture features are present on Perlmutter CPUs

- Many-Core
- MPI+OpenMP Programming Model Will Continue

Easier Onramp to “Many-Core” with Perlmutter CPUs than with KNL

- More Traditional Cores
- Single Memory Technology

GPU Transition Path for CPU Apps

NESAP for Perlmutter will extend activities from NESAP for Cori

1. Identifying and exploiting on-node parallelism - threads + vector
2. Understanding and improving data-locality within the cache-memory hierarchy

What's New?

1. Heterogeneous compute elements
2. Identification and exploitation of even more parallelism
3. Emphasis on performance-portable programming approach:

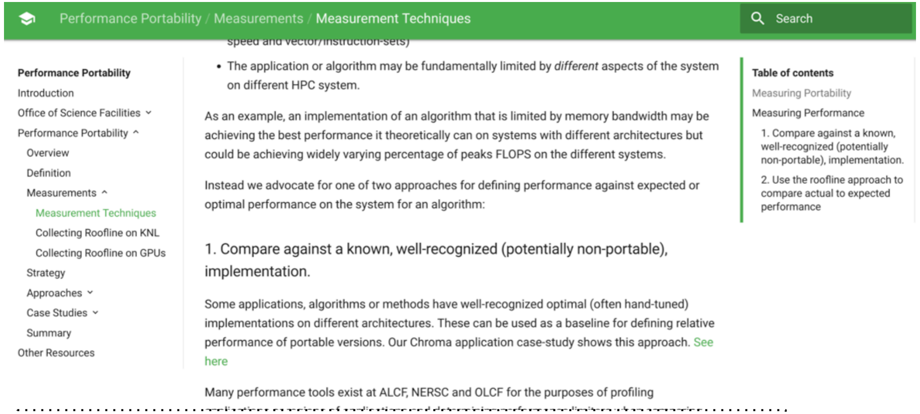
Programming Models Supported

CUDA, CUDA FORTRAN, OpenACC, Kokkos, Raja, OpenMP NRE with PGI/NVIDIA

Engaging around Performance Portability

NRE with PGI to enable OpenMP GPU acceleration

NERSC will pursue membership in OpenACC



Performance Portability / Measurements / Measurement Techniques

Search

speed and vector/instruction-sets)

Performance Portability
Introduction
Office of Science Facilities
Performance Portability
Overview
Definition
Measurements
Measurement Techniques
Collecting Roofline on KNL
Collecting Roofline on GPUs
Strategy
Approaches
Case Studies
Summary
Other Resources

- The application or algorithm may be fundamentally limited by *different* aspects of the system on different HPC system.

As an example, an implementation of an algorithm that is limited by memory bandwidth may be achieving the best performance it theoretically can on systems with different architectures but could be achieving widely varying percentage of peaks FLOPS on the different systems.

Instead we advocate for one of two approaches for defining performance against expected or optimal performance on the system for an algorithm:

1. Compare against a known, well-recognized (potentially non-portable), implementation.

Some applications, algorithms or methods have well-recognized optimal (often hand-tuned) implementations on different architectures. These can be used as a baseline for defining relative performance of portable versions. Our Chroma application case-study shows this approach. See here

Many performance tools exist at ALCF, NERSC and OLCF for the purposes of profiling

Table of contents
Measuring Portability
Measuring Performance
1. Compare against a known, well-recognized (potentially non-portable), implementation.
2. Use the roofline approach to compare actual to expected performance

NERSC leading development of performanceportability.org

Doug Doerfler leading performance portability workshop at SC18, and 2019 DOE COE perf. port. meeting

N9 will also have compiler support for Kokkos and RAJA