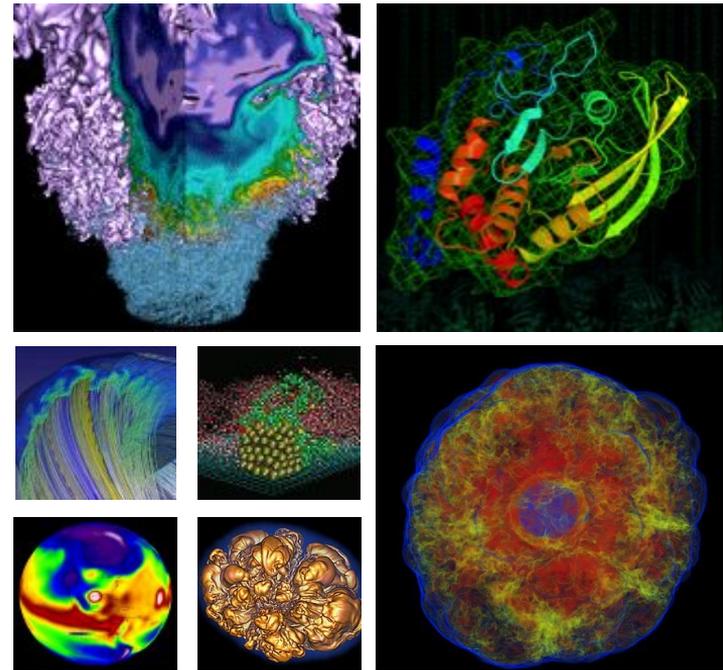


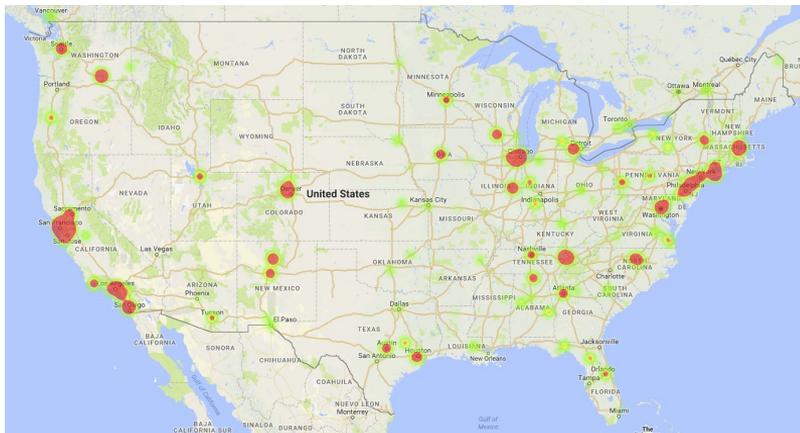
NERSC-9



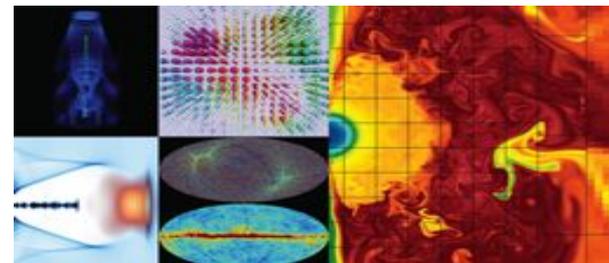
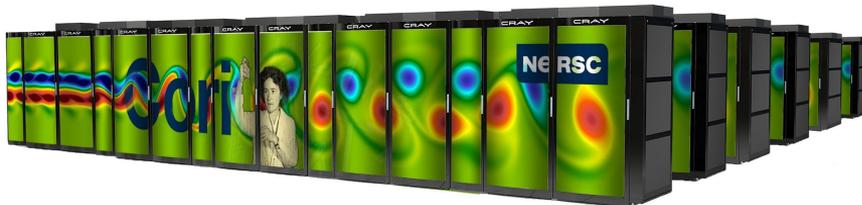
Sudip Dosanjh, NERSC Director
Katie Antypas NERSC-9 Project Director

ASCAC Meeting
Dec. 12, 2018

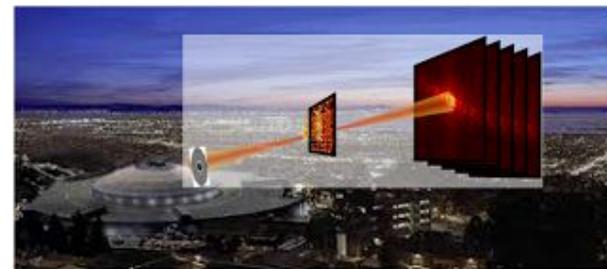
NERSC is the mission HPC facility for the DOE Office of Science



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



Simulations at scale

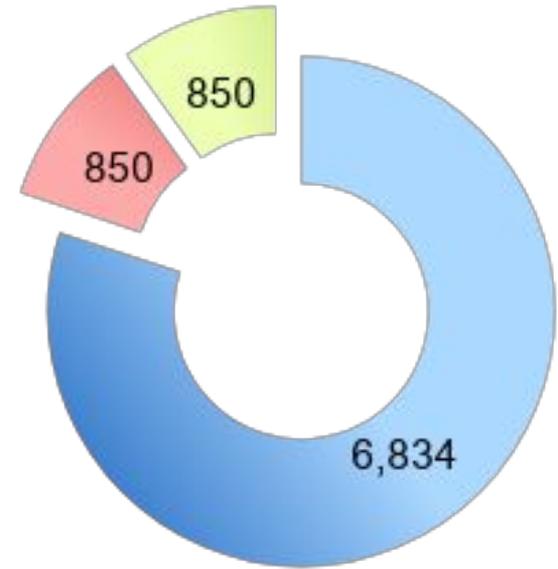
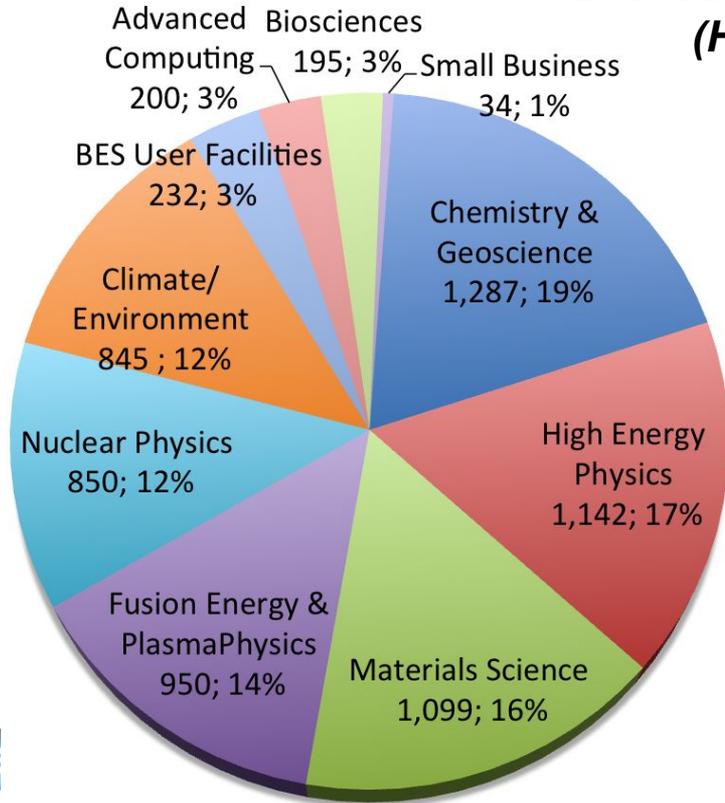


Data analysis support for
DOE's experimental and
observational facilities

Photo Credit: CAMERA

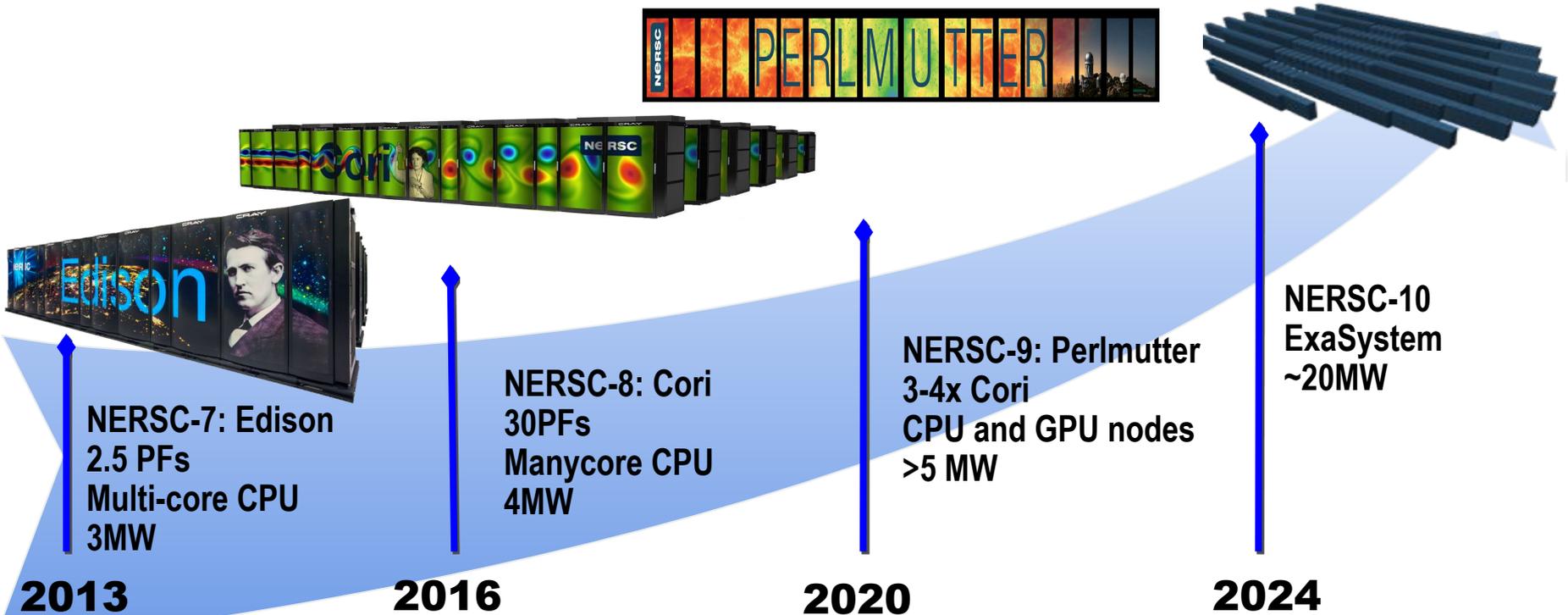
NERSC Directly Supports Office of Science Priorities

**2018 Allocation Breakdown
(Hours Millions)**



- DOE Mission Science
- ALCC
- Directors Discretionary

NERSC Systems Roadmap



NERSC-9 Project Major Scope Items

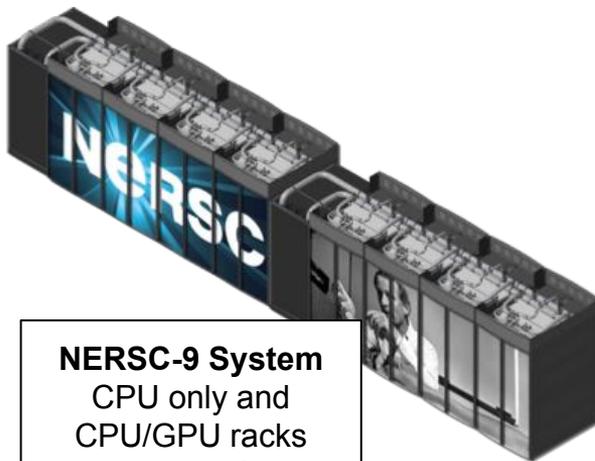


Facility Upgrade



- 12.5MW upgrade
- Cooling and electrical scope to support upgrade

System Deployment, Integration and Acceptance

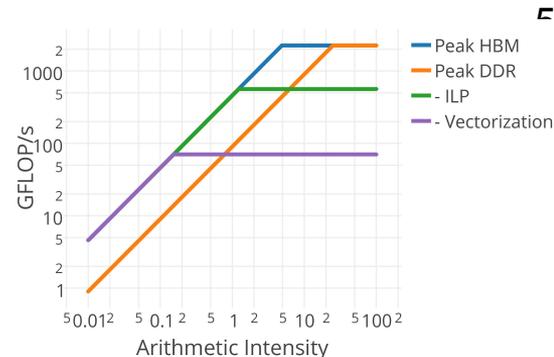
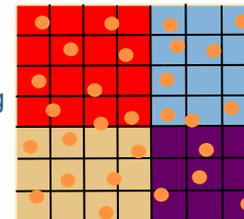


NERSC-9 System
CPU only and CPU/GPU racks
All Flash File System
Cray high speed network

Application Readiness

- Partnerships with ~25 apps teams and vendors for code optimization
- Training and documentation for broad community

Tiled layout for optimal caching (WARP code)

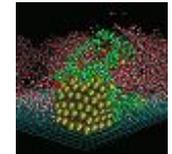
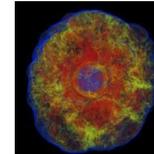
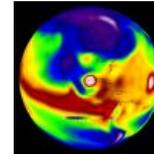
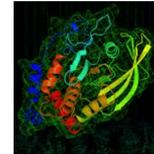
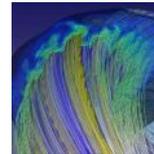
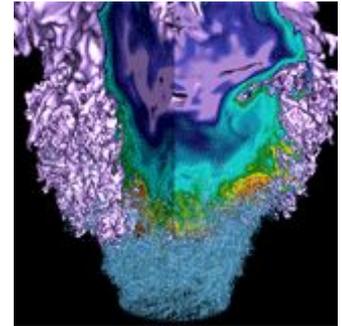


NERSC-9 will be named after Saul Perlmutter

- Winner of 2011 Nobel Prize in Physics for discovery of the accelerating expansion of the universe.
- Supernova Cosmology Project, lead by Perlmutter, was a pioneer in using NERSC supercomputers to combine large scale simulations with experimental data analysis
- Login “saul.nersc.gov”



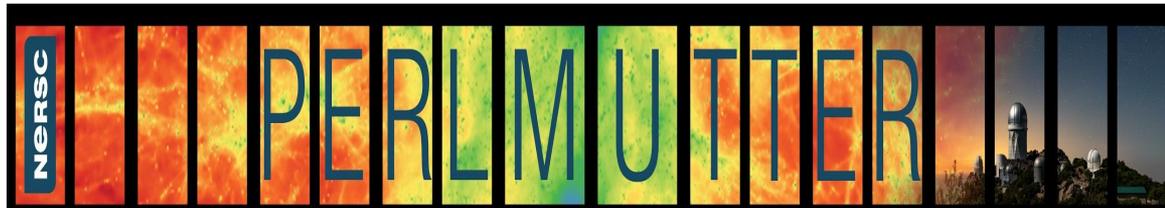
NERSC-9 Architecture



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**



From the start NERSC-9 had requirements of simulation and data users in mind

- All Flash file system for workflow acceleration
- 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
- Dedicated workflow management and interactive nodes

Exascale Requirements Reviews 2015-2018

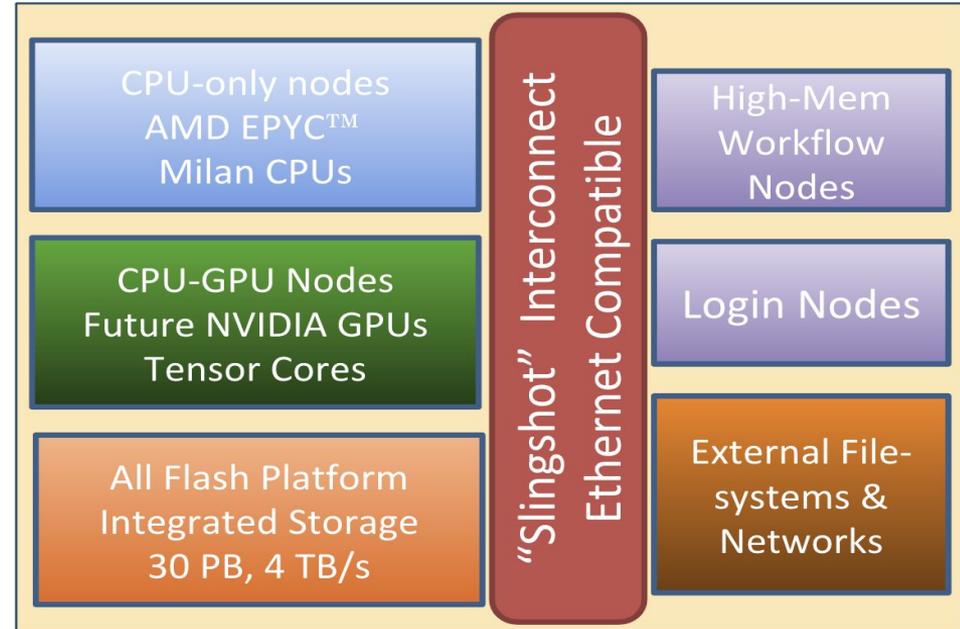
First time users from DOE experimental facilities broadly included



Perlmutter: A System Optimized for Science



- GPU-accelerated and CPU-only nodes meet the needs of large scale simulation and data analysis from experimental facilities
- Cray “Slingshot” - High-performance, scalable, low-latency Ethernet-compatible network
- Single-tier All-Flash Lustre based HPC file system, 6x Cori’s bandwidth
- Dedicated login and high memory nodes to support complex workflows

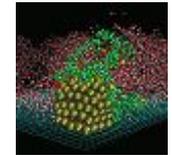
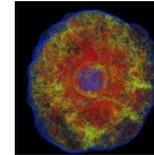
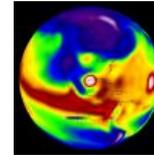
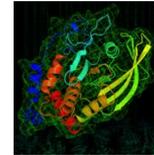
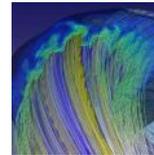
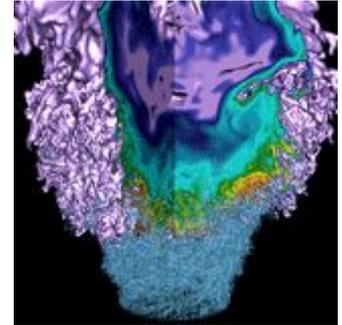


Compute Node Details

- **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)

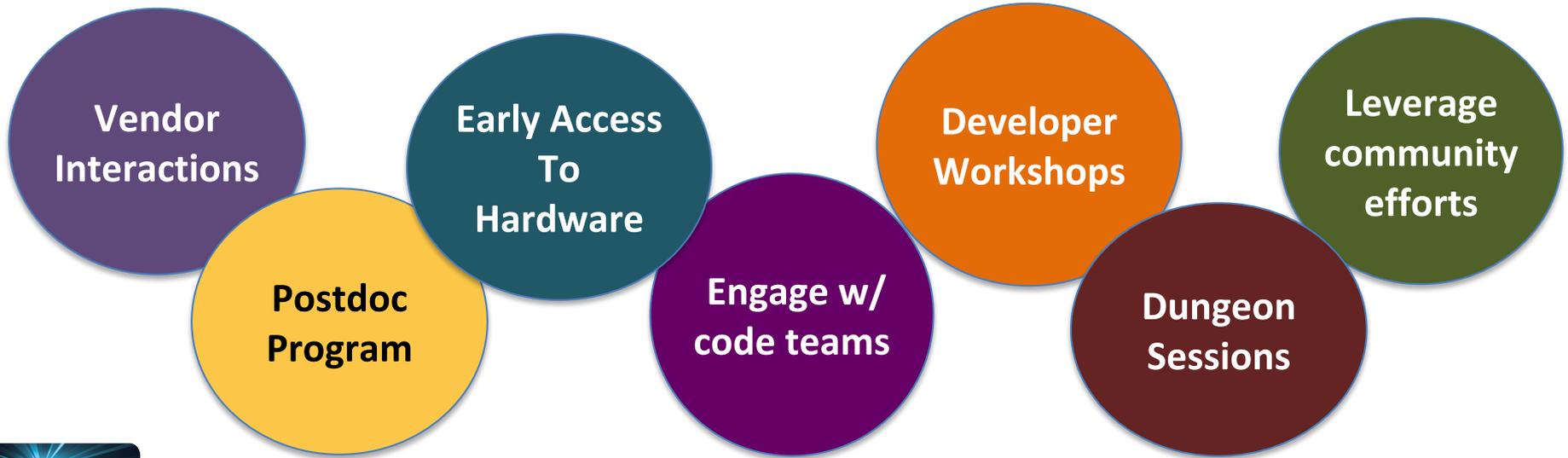


NERSC Exascale Science Application Program (NESAP)



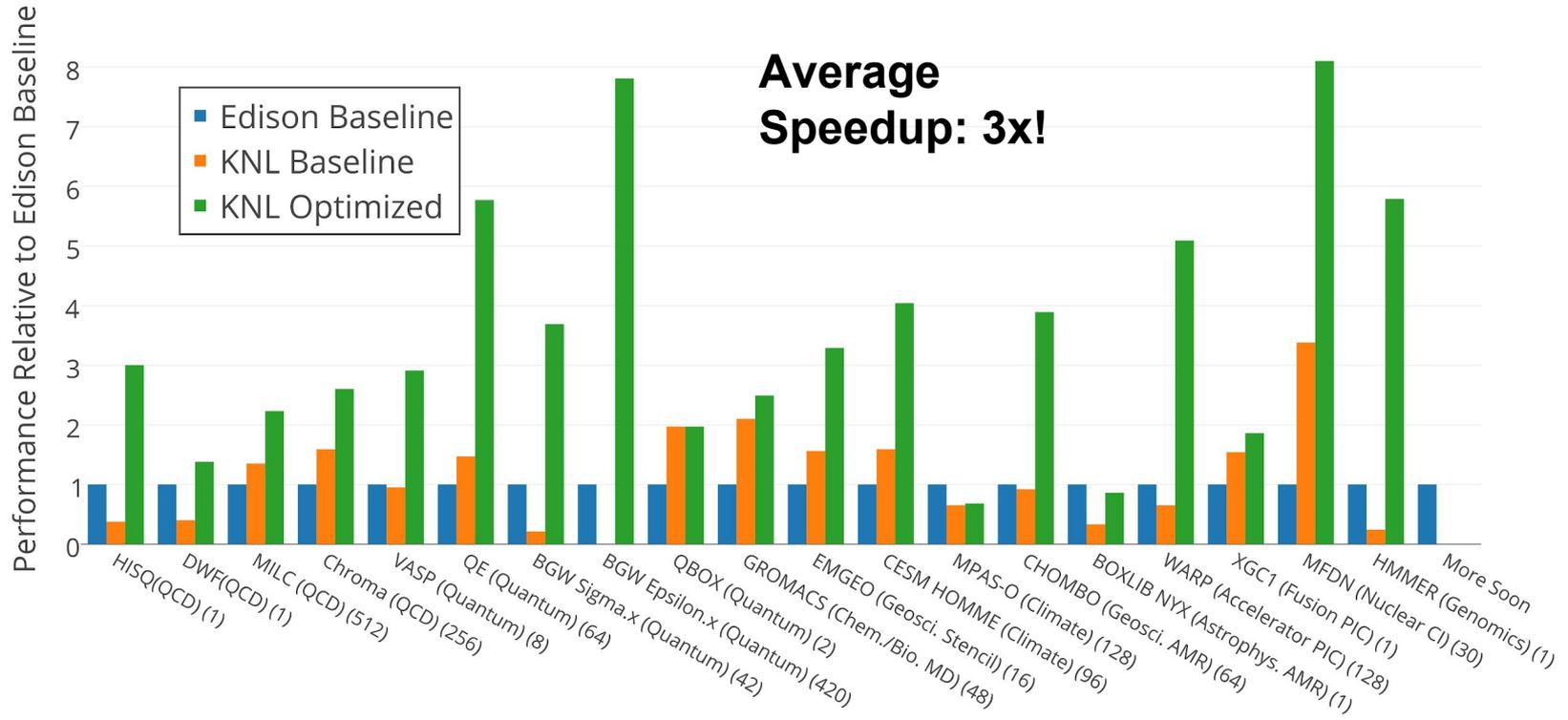
NERSC Exascale Scientific Application Program (NESAP)

- Prepare DOE SC users for advanced architectures like Cori
- Partner closely with ~20 application teams and apply lessons learned to broad NERSC user community.



Major Lessons From NESAP 1 on Cori

Code Engagement Leads to Significant Speedups



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**

<https://nersc.gov/users/application-performance/nesap/perlmutter/>

Open Now through December 18, 2018 12:00 noon PST

Selections announced in January 2019

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



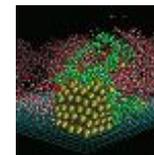
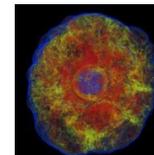
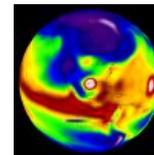
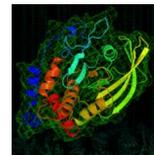
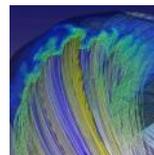
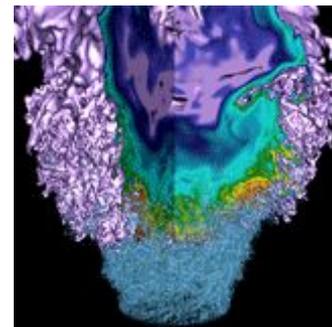
Training, Case Studies and Documentation

- For those teams not in NESAP, there will be a robust training program
- Lessons learned from deep dives from NESAP teams will be shared through case studies and documentation



The screenshot displays the NERSC website's 'APPLICATION CASE STUDIES' page. At the top, the NERSC logo is on the left, and the tagline 'Powering Scientific Discovery Since 1974' is on the right. A search bar is located in the top right corner. Below the header is a navigation menu with links for HOME, ABOUT, SCIENCE AT NERSC, SYSTEMS, FOR USERS (highlighted), NEWS & PUBLICATIONS, R & D, EVENTS, LIVE STATUS, and TIMELINE. The main content area is titled 'APPLICATION CASE STUDIES' and includes a breadcrumb trail: Home » For Users » Computational Systems » Cori » Application Porting and Performance » Application Case Studies. The text describes the work of NERSC staff with NESAP applications on the Cori-Phase 2 system. It lists presentations at the ISC 16 IXPUG Workshop and provides a link to the event page. A list of links for further reading includes 'Getting Started', 'Measuring Arithmetic Intensity', 'Measuring and Understanding Memory Bandwidth', and 'Vectorization'. Three case study entries are shown: 'EMGEO Case Study' (dated June 20, 2016), 'BerkeleyGW Case Study', and 'QPhIX Case Study' (dated June 20, 2016). A 'WARP Case Study' entry is partially visible at the bottom.

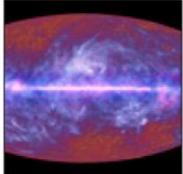
Workflows and Data Analytics



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



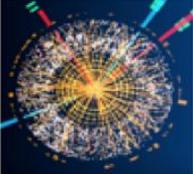
Palomar Transient Factory Supernova



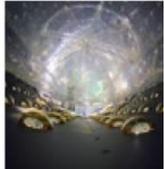
Planck Satellite Cosmic Microwave Background Radiation



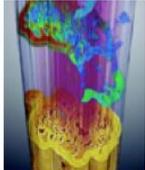
Alice Large Hadron Collider



Atlas Large Hadron Collider



Dayabay Neutrinos



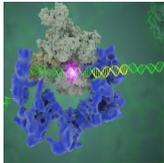
ALS Light Source



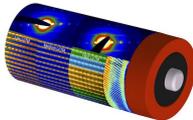
LCLS Light Source



Joint Genome Institute Bioinformatics



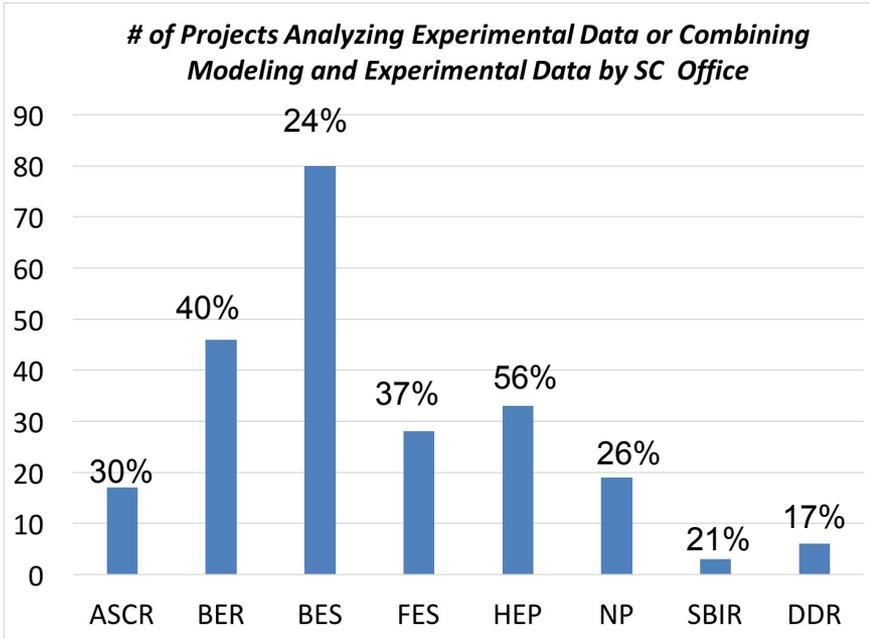
Cryo-EM



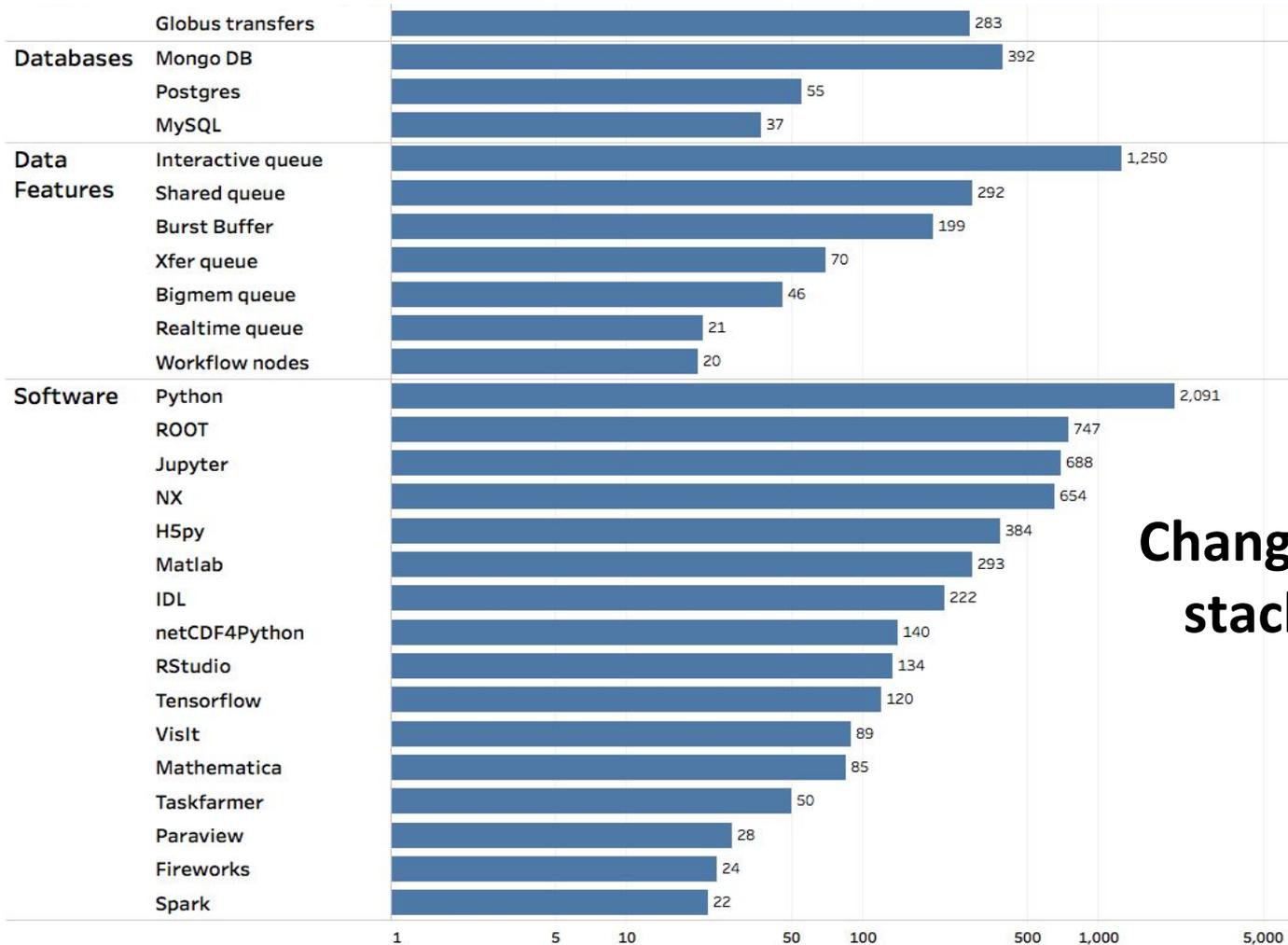
NCEM



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

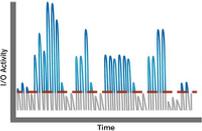


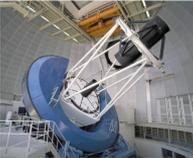
Changing software stack at NERSC

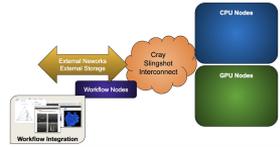
Number of Unique Users

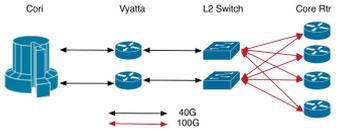
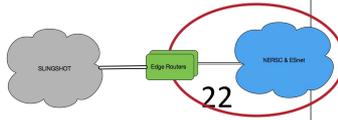


Data Features	Cori experience	N9 enhancements
---------------	-----------------	-----------------

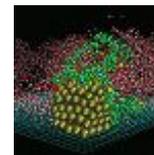
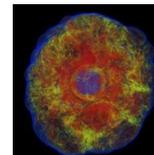
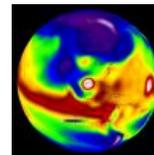
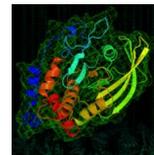
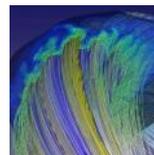
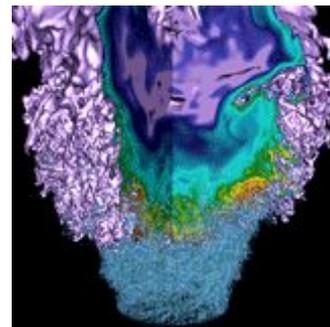
<h2>I/O and Storage</h2>	<p>Burst Buffer</p> 	<p>All-flash file system: performance with ease of data management</p> 
--------------------------	--	--

<h2>Analytics</h2> <ul style="list-style-type: none"> - Production stacks - Analytics libraries - Machine learning 	<p>User defined images with Shifter NESAP for data</p>  <p>New analytics and ML libraries</p> 	<p>Benchmark Production Analytics workflows. Data apps in NESAP at outset</p>  <p>Optimised analytics libraries and deep learning application benchmarks</p>
---	---	---

<h2>Workflow integration</h2>	<p>Real-time queues</p> 	<p>SLURM co-scheduling Workflow nodes integrated</p> 
-------------------------------	---	--

<h2>Data transfer and streaming</h2>	<p>SDN</p> 	<p>Slingshot ethernet-based converged fabric</p> 
--------------------------------------	--	---

Wrap up



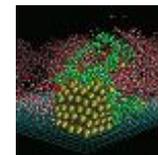
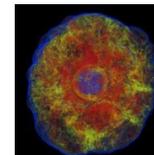
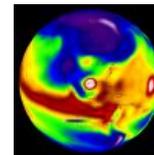
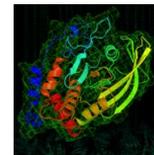
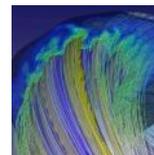
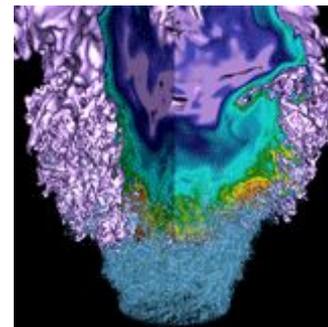
NERSC-9 System Deployment Timeline

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

Summary

- Planning for NERSC-9 started in 2015. We are thrilled to have the contract signed and start putting plans into action
- We look forward to working with the ASCR and SC community to make NERSC-9 a success!

Backup

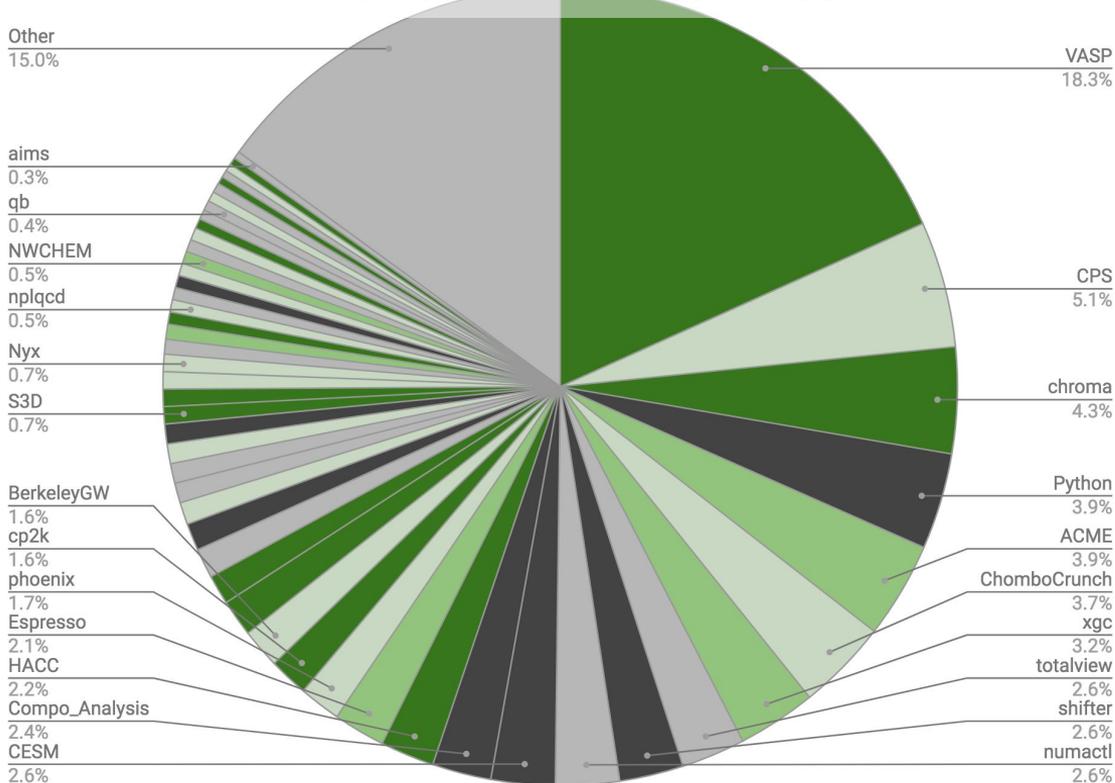


NERSC-9 System Optimized for Science

- In order to meet science requirements and mission need, accelerator technology was essential
- System contains large number of CPU-only nodes for applications that are not yet ready
- Significant fraction of NERSC workload can now use GPUs
 - GPU programming has matured
 - Improved software environment
 - Increases in GPU memory capacity improve programmability
- System well balanced between network and FLOPS
- All-flash filesystem for I/O acceleration

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

GPU Programming Models

We will support and engage our user community where their apps are:

CUDA: MILC, Chroma, HACC ...

CUDA FORTRAN: Quantum ESPRESSO, StarLord (AMREX)

OpenACC: VASP, E3SM, MPAS, GTC, XGC ...

Kokkos: LAMMPS, PELE, Chroma ...

Raja: SW4

Engaging around Performance Portability



NERSC is working with PGI/NVIDIA to enable OpenMP GPU acceleration



NERSC Hosted Past C++ Summit and ISO C++ meeting on HPC.



NERSC Will Pursue Membership in 2018

Performance Portability / Measurements / Measurement Techniques

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

- 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 Chrome application case:study.shows this approach. See

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

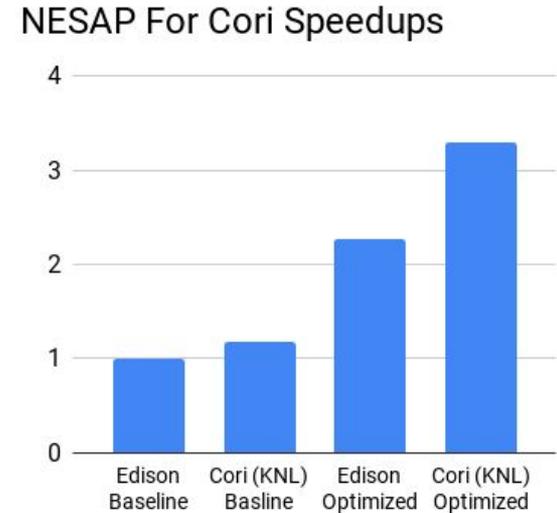
NERSC is leading development of performanceportability.org



Doug Doerfler Leading Accepted Performance Portability Workshop at SC18. and 2019 DOE COE Perf. Port. Meeting

NESAP call for proposals

- Resources available to awardees
 - 1 Hackathon Session Per Quarter
 - NERSC, Cray, NVIDIA Engineer Attendance
 - Cray/NVIDIA Engineer Time Before and After Sessions
 - NESAP PostDocs (NERSC will hire up to 17)
 - NERSC Application Performance Specialist Attention
 - General Programming, Performance and Tools Training
 - Early Access (Perlmutter and GPU testbed)



<https://nerSC.gov/users/application-performance/nesap/perlmutter/>

Open Now through December 18, 2018 12:00 noon PST

Selections announced in January 2019