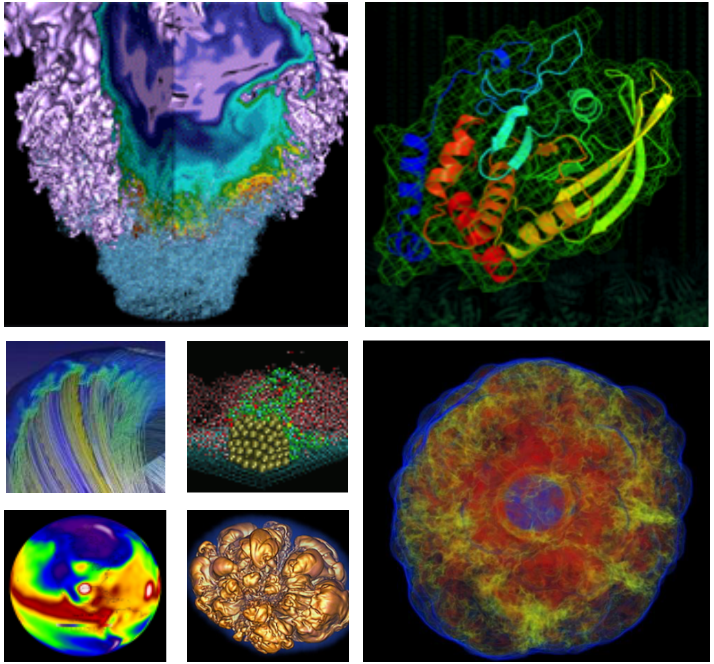
Preparing NERSC Applications for Perlmutter as an Exascale Waypoint



Charlene Yang

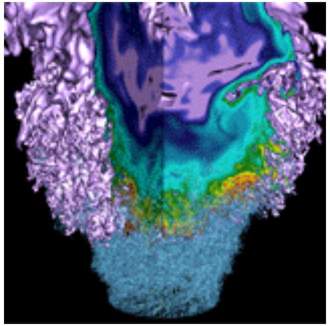
Application Performance Specialist

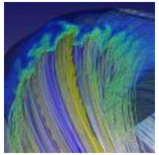
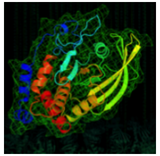
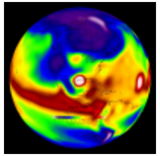
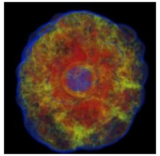
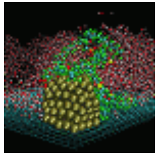
cjyang@lbl.gov

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Perlmutter Overview



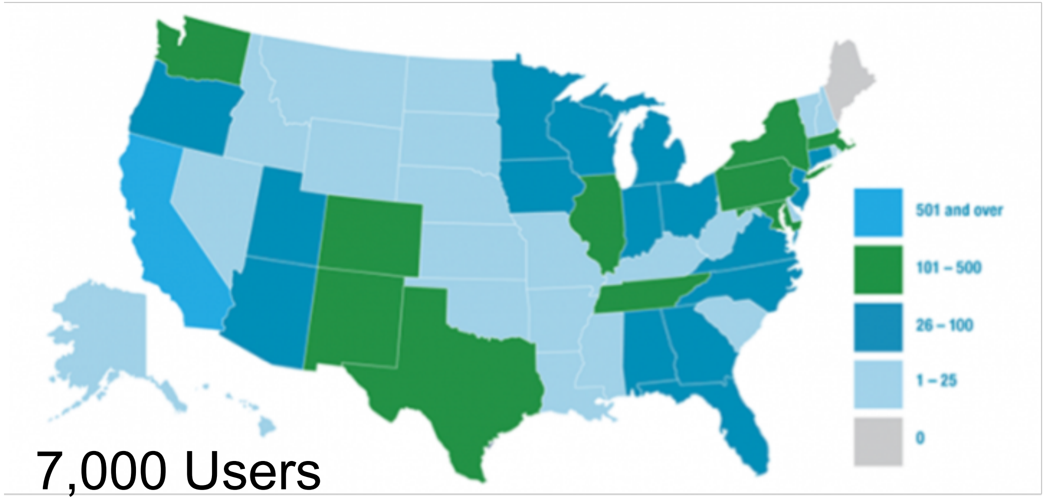






NERSC

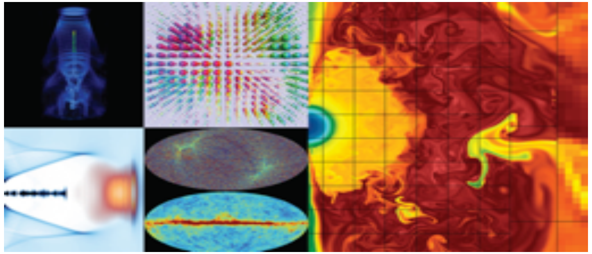
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

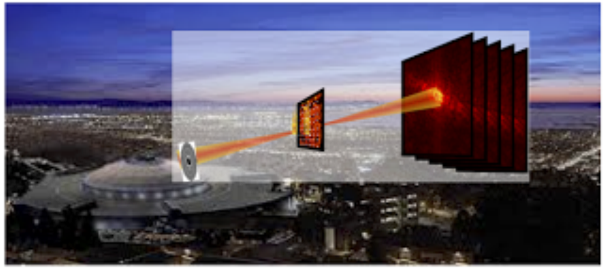
NERSC is the mission High Performance Computing facility for the DOE SC



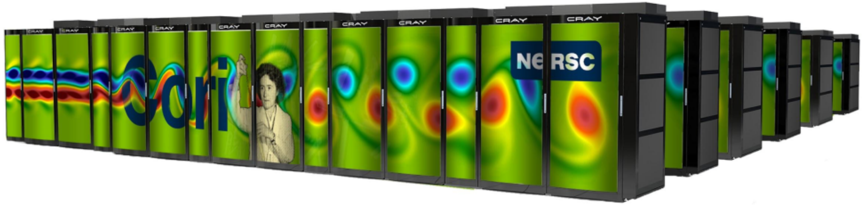
7,000 Users
800 Projects
700 Codes
2000 NERSC citations per year



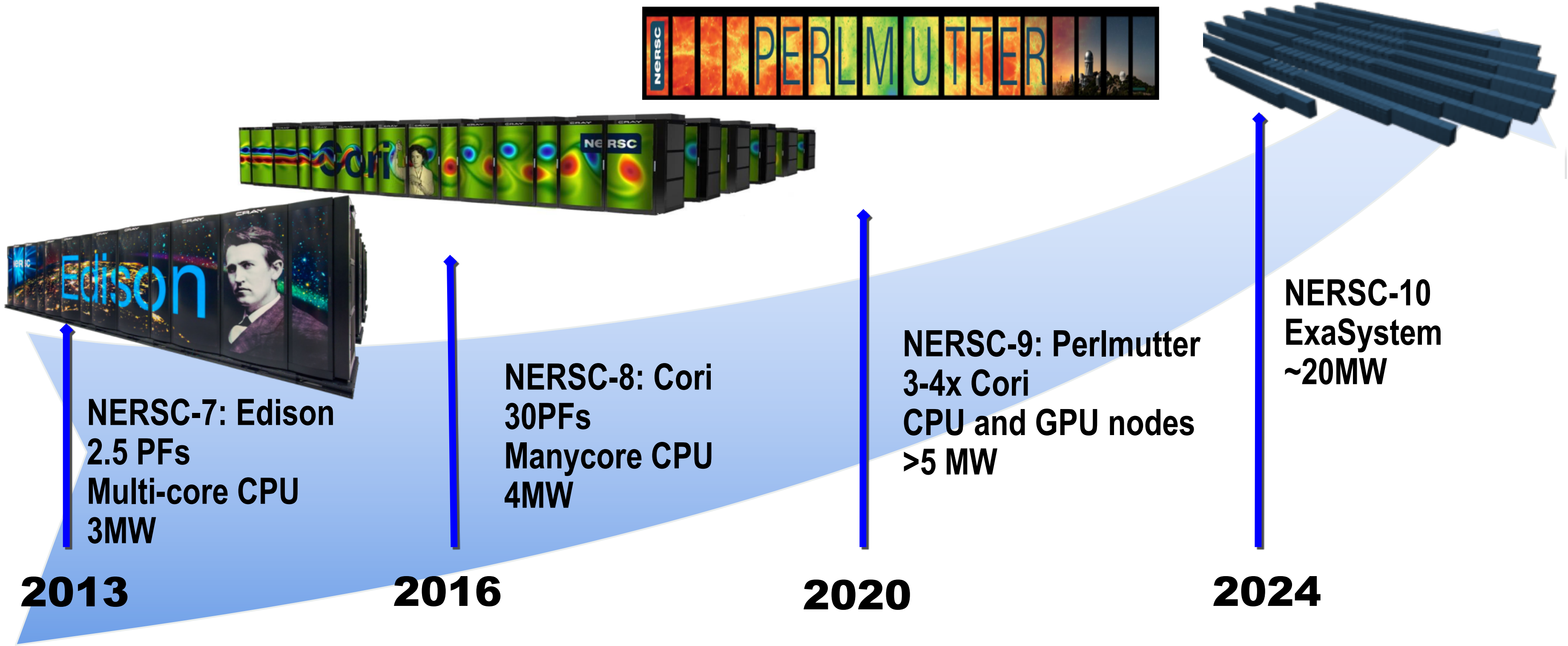
Simulations at scale



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



NERSC Systems Roadmap

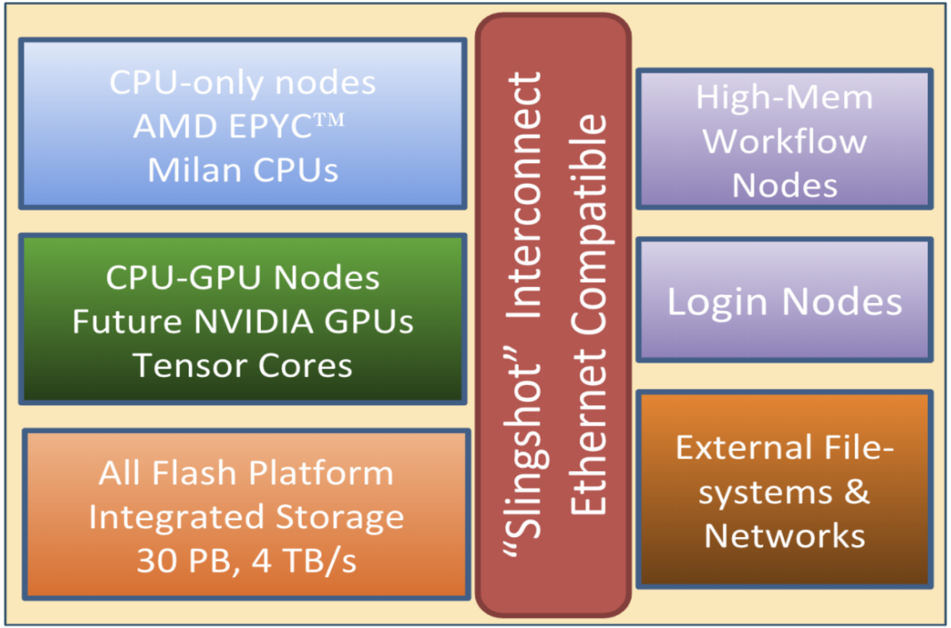


NERSC-9 will be named after Saul Perlmutter

- 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, combining large scale simulations with experimental data analysis
- Login “saul.nersc.gov”

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

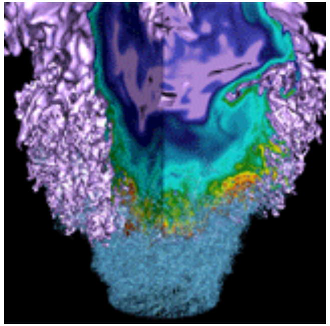
From the start NERSC-9 had requirements of simulation and data users in mind

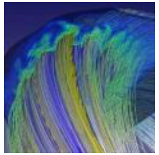
- All Flash file system for workflow acceleration
- Optimized network for data ingest from experimental facilities
- Dedicated workflow management and interactive nodes
- Real-time scheduling capabilities
- Supported analytics stack including latest ML/DL software
- System software supporting rolling upgrades for improved resilience

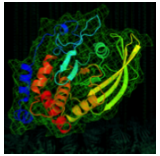
Exascale Requirements Reviews 2015-2018

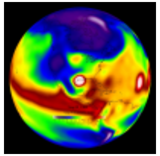
First time users from DOE experimental facilities broadly included

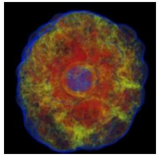
NESAP Overview

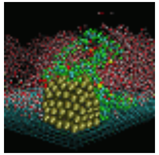












NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

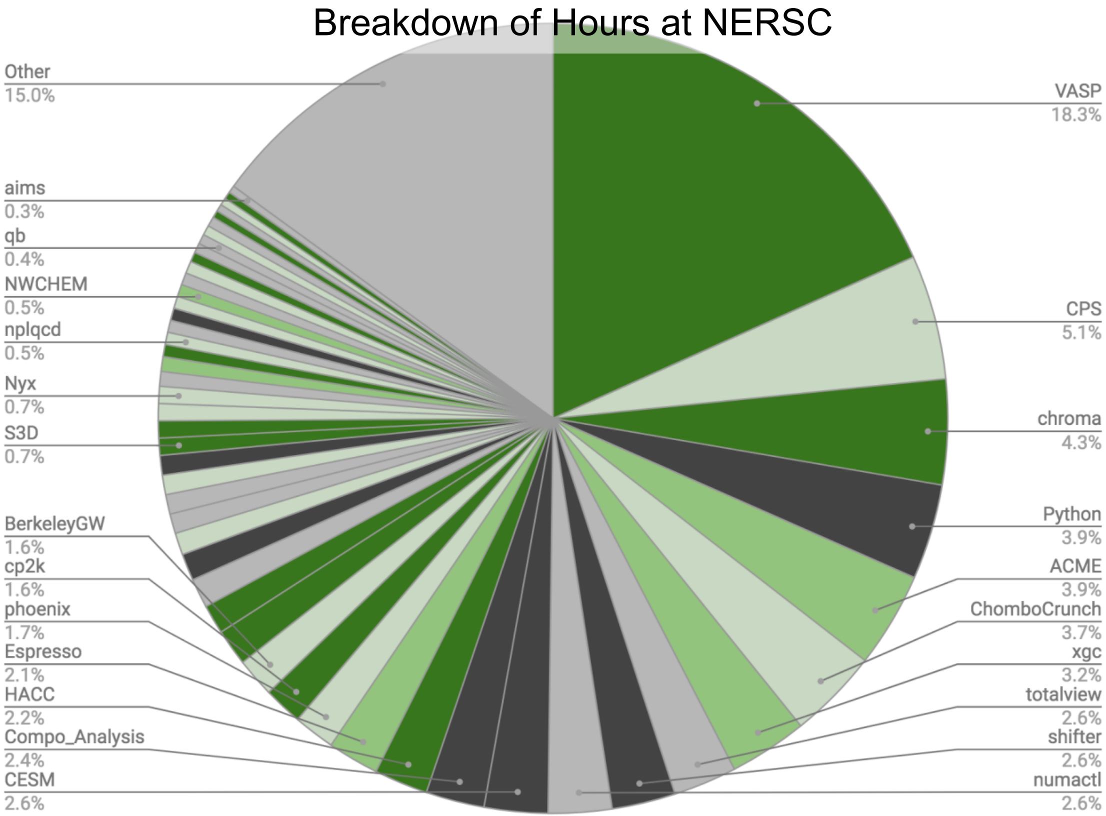
BERKELEY LAB

Application Readiness Strategy for Perlmutter

NERSC's Challenge

How to enable NERSC's diverse community of 7,000 users, 750 projects, and 700 codes to run on advanced architectures like Perlmutter and beyond?

GPU Readiness Among NERSC Codes (Aug'17 - Jul'18)



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

Application Readiness Strategy for Perlmutter

How to transition a workload with 700 Apps?

- **NERSC Exascale Science Application Program (NESAP)**
- Engage ~25 Applications
- up to 17 postdoctoral fellows
- Deep partnerships with every SC Office area
- Leverage vendor expertise and hack-a-thons
- Knowledge transfer through documentation and training for all users
- Optimize codes with improvements relevant to multiple architectures

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

GPU Transition Path for Apps

NESAP for Perlmutter will extend activities from **NESAP for Cori**

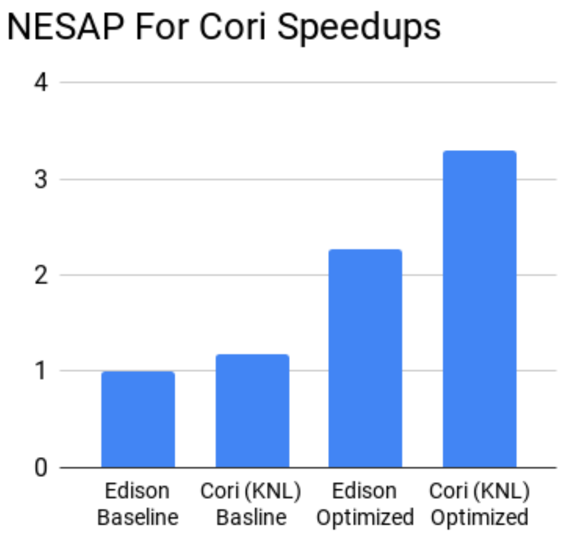
1. Identifying and exploiting on-node parallelism
2. Understanding and improving data-locality within the memory hierarchy

Knowledge and skills of multi/many-core optimization on HSW/KNL transferrable to AMD CPUs

What's New for NERSC Users?

1. Heterogeneous compute elements - NVIDIA GPUs
2. Identification and exploitation of even more parallelism
3. Data locality again, host/device

Emphasis on performance-portable programming approach: Continuity from Cori through future NERSC systems

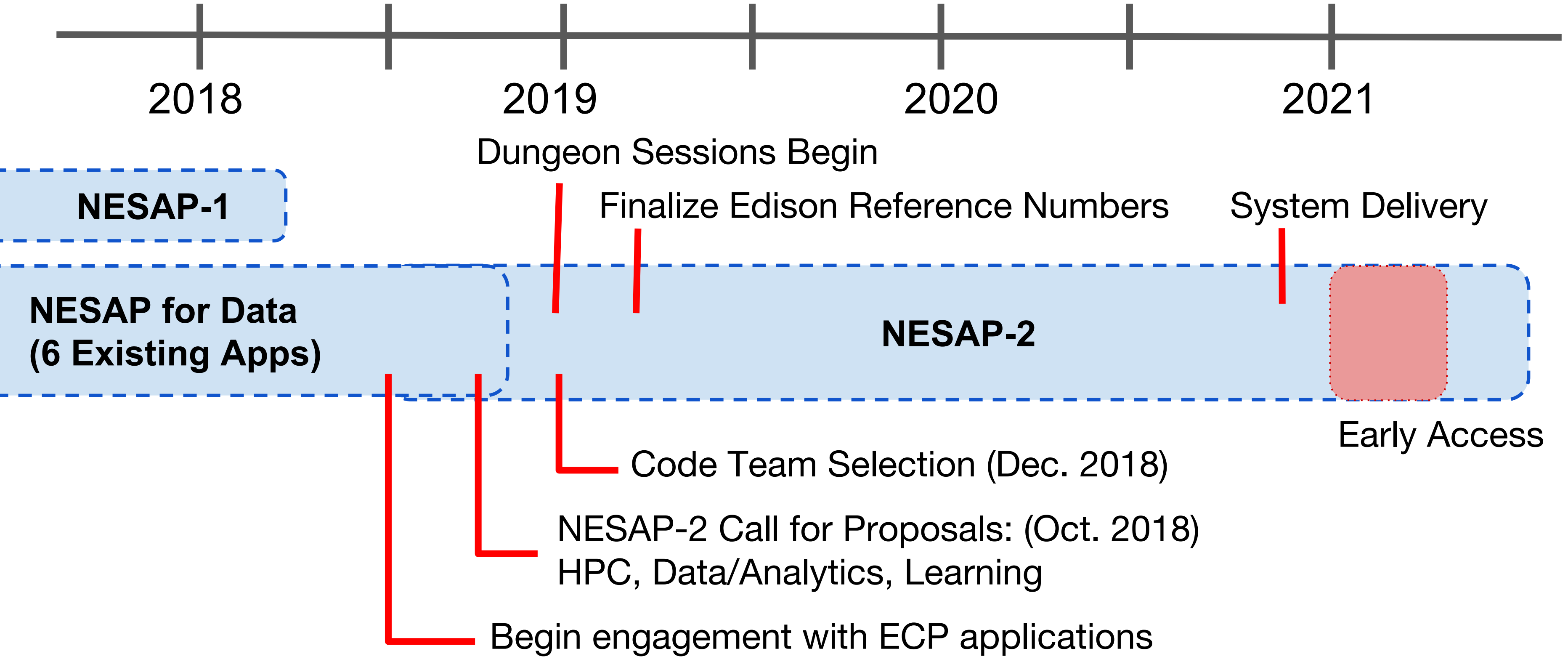


NESAP for Perlmutter

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

- 6 NESAP for Data apps will be continued. Additional apps focused on experimental facilities.
- 5 ECP Apps Jointly Selected (Participation Funded by ECP)
- Open call for proposals. Reviewed by a committee of NERSC staff, external reviewers and input from DOE PMs.
  - **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**

NESAP for Perlmutter Timeline



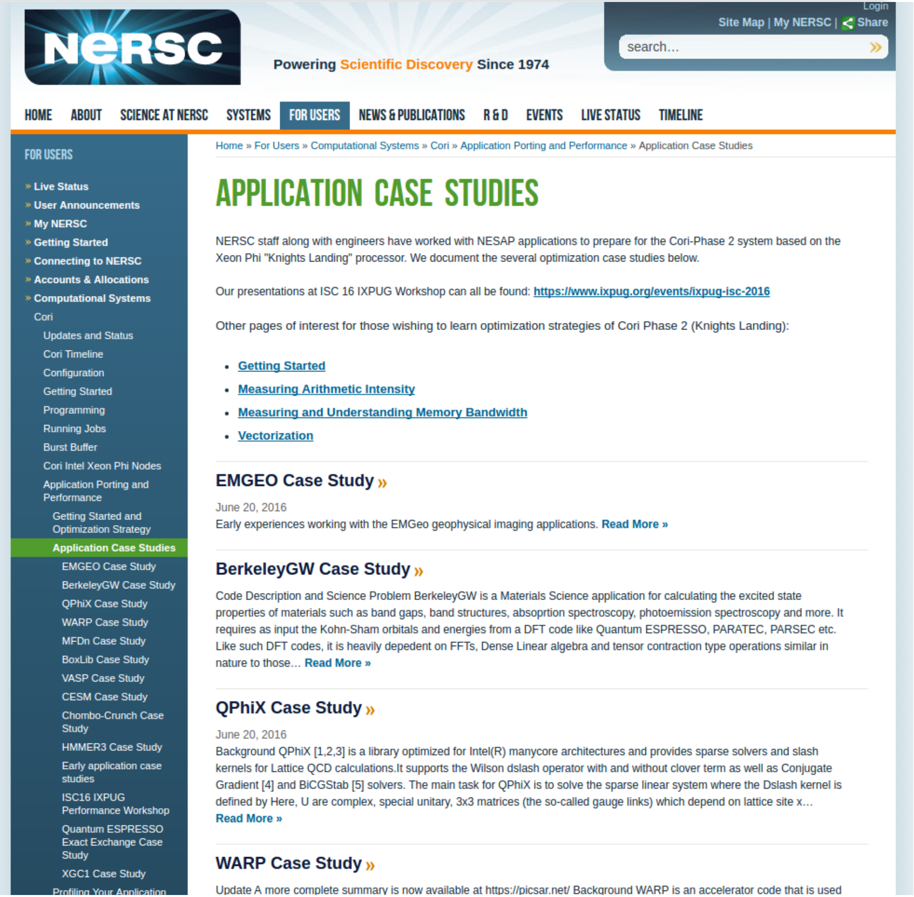
Resources Available to NESAP Awardees

- 1 hackathon session per quarter (Center of Excellence)
  - NERSC, Cray, NVIDIA engineer attendance
  - 3-4 code teams per hackathon
- Cray/NVIDIA engineer time before and after sessions
  - 6-week 'ramp-up' period with code teams and Cray/NVIDIA to ensure everyone is fully prepared to work on hackathon day 1
  - Tutorials/deep dives into GPU programming models, profiling tools, *etc*
- 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)

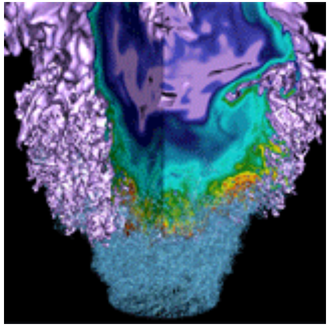
Training, Case Studies and Documentation

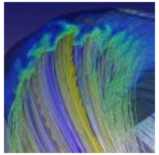
- For those teams not in NESAP, there will be a robust training program

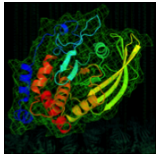
- Lessons learned from deep dives from NESAP teams will be shared through case studies and documentation

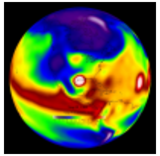


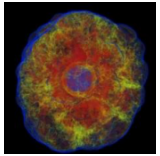
Programming & Performance Portability

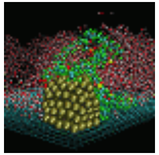












NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

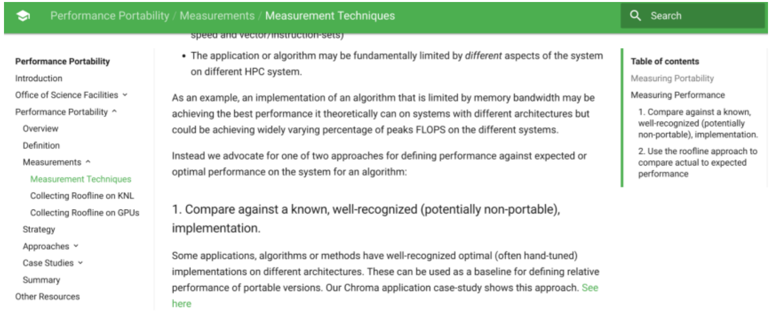
BERKELEY LAB

Engaging around Performance Portability

NERSC will work with PGI to enable OpenMP GPU acceleration with PGI compilers

- Ensures continuity of OpenMP added to NERSC apps for N8
- Co-design with PGI to prioritize OpenMP features for GPU
- Use lessons learned to influence future versions of OpenMP

NERSC has joined membership in OpenACC


Performance Portability / Measurements / Measurement Techniques

Search

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

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 collaboarting with OLCF and ALCF on development of performanceportability.org

- kokkos training in March 2019
- UPC++ available
- Are you part of an ECP ST project? Interested in contributing a NERSC hosted training?

Performance Portability Strategies

- Conditional compilation
- Directives: OpenMP, OpenACC
- Libraries: Use a library when possible
- Abstractions: Kokkos, Raja
- General-purpose high-level programming languages: UPC, Coarray Fortran
- DSLs: NMODL for neuroscience

Good coding practices

- Modularity, some high-level abstractions
- Data structures flexibly allocatable to different memory spaces
- Task level flexibility so work can be allocated to different compute elements (GPU & CPU)

Performance Portability

There is no consensus on the definition or metric for performance portability, but…

DOE SC Facility Definition (performanceportability.org)

An application is performance portable if it achieves a consistent ratio of the actual time to solution to either the best-known or the theoretical best time to solution on each platform with minimal platform specific code required.

Performance portability metric proposed by Pennycook et al. [1]

$$\boldsymbol{\Phi}(a,p,\boldsymbol{H}) = \begin{cases} \dfrac{|\boldsymbol{H}|}{\sum_{i\in \boldsymbol{H}} \frac{1}{e_i(a,p)}} & \text{if } i \text{ is supported}, \forall i \in \boldsymbol{H} \\ 0 & \text{otherwise} \end{cases}$$

Architectural Efficiency

Actual Application Performance

$$e_i(a,p) = \frac{P_i(a,p)}{\min(F_i,\ B_i \times I_i(a,p))}$$

Max Attainable Performance defined by Roofline [2]

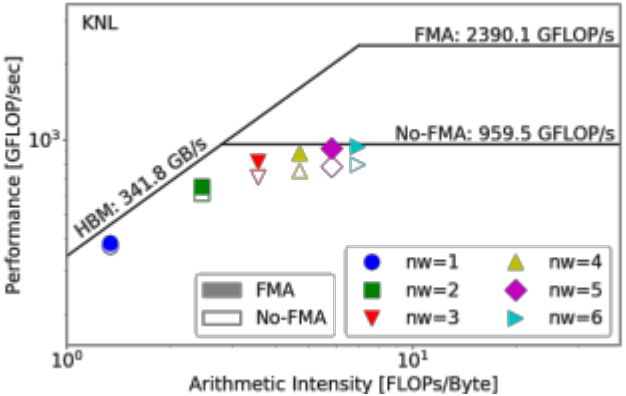
[1] S. J. Pennycook, J. D. Sewall, and V. Lee, "A metric for performance portability," arXiv:1611.07409, 2016.
[2] S. Williams, A. Waterman, and D. Patterson, "Roofline: An insightful visual performance model for multicore architectures," Communications of the ACM, vol. 52, no. 4, pp. 65–76, 2009.

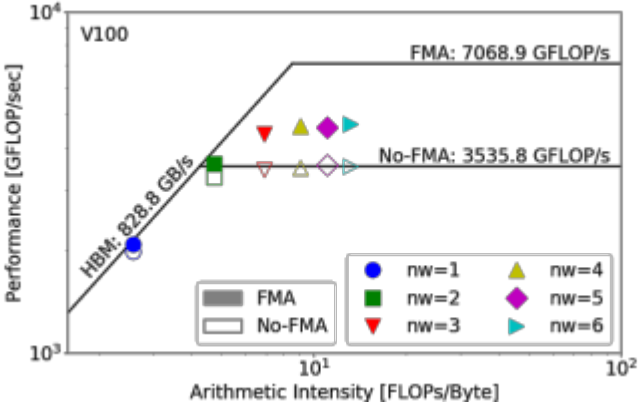
Methodology to Measure Perf. Port.

1. Measure empirical compute and bandwidth ceilings: ERT [3]
2. Measure application performance: SDE and LIKWID on KNL; NVPROF on V100

$$\text{Performance} = \frac{\text{SDE or } nvprof \text{ FLOPs}}{\text{Runtime}}, \quad \text{Arithmetic Intensity} = \frac{\text{SDE or } nvprof \text{ FLOPs}}{\text{LIKWID or } nvprof \text{ Data Movement}}$$

An example: GPP kernel from BerkeleyGW (Roofline)





[3] Empirical Roofline Toolkit. [Online]. Available: https://bitbucket.org/berkeleylab/cs-roofline-toolkit/src/master/

Methodology to Measure Perf. Port.

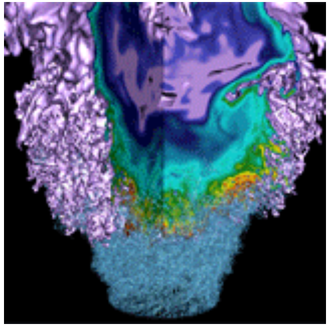
An example: GPP kernel from BerkeleyGW (Perf. Port. Scores)

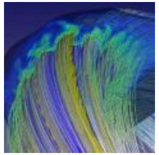
	Architectural Efficiency	nw=1	nw=2	nw=3	nw=4	nw=5	nw=6
FMA	KNL	84.98%	77.50%	66.77%	55.28%	46.56%	39.65%
	V100	97.36%	91.50%	76.70%	65.44%	65.07%	66.38%
	Performance Portability	**90.76%**	**83.92%**	**71.39%**	**59.93%**	**54.28%**	**49.65%**
No-FMA	KNL	82.06%	72.95%	73.74%	78.72%	81.28%	82.81%
	V100	92.88%	92.88%	97.43%	98.91%	1	99.73%
	Performance Portability	**87.14%**	**81.72%**	**83.95%**	**87.67%**	**89.93%**	**90.49%**

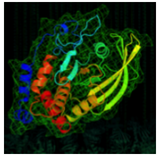
- Roofline captures application bottlenecks, machine balance, problem size, *etc*
- Perf. Port. metric captures performance changes across architectures

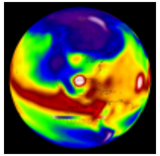
Yang, Charlene, et al. "An empirical roofline methodology for quantitatively assessing performance portability." *International Workshop on Performance, Portability and Productivity in HPC P3HPC*. 2018.

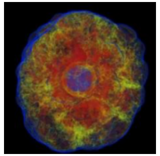
Summary

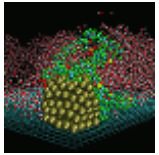










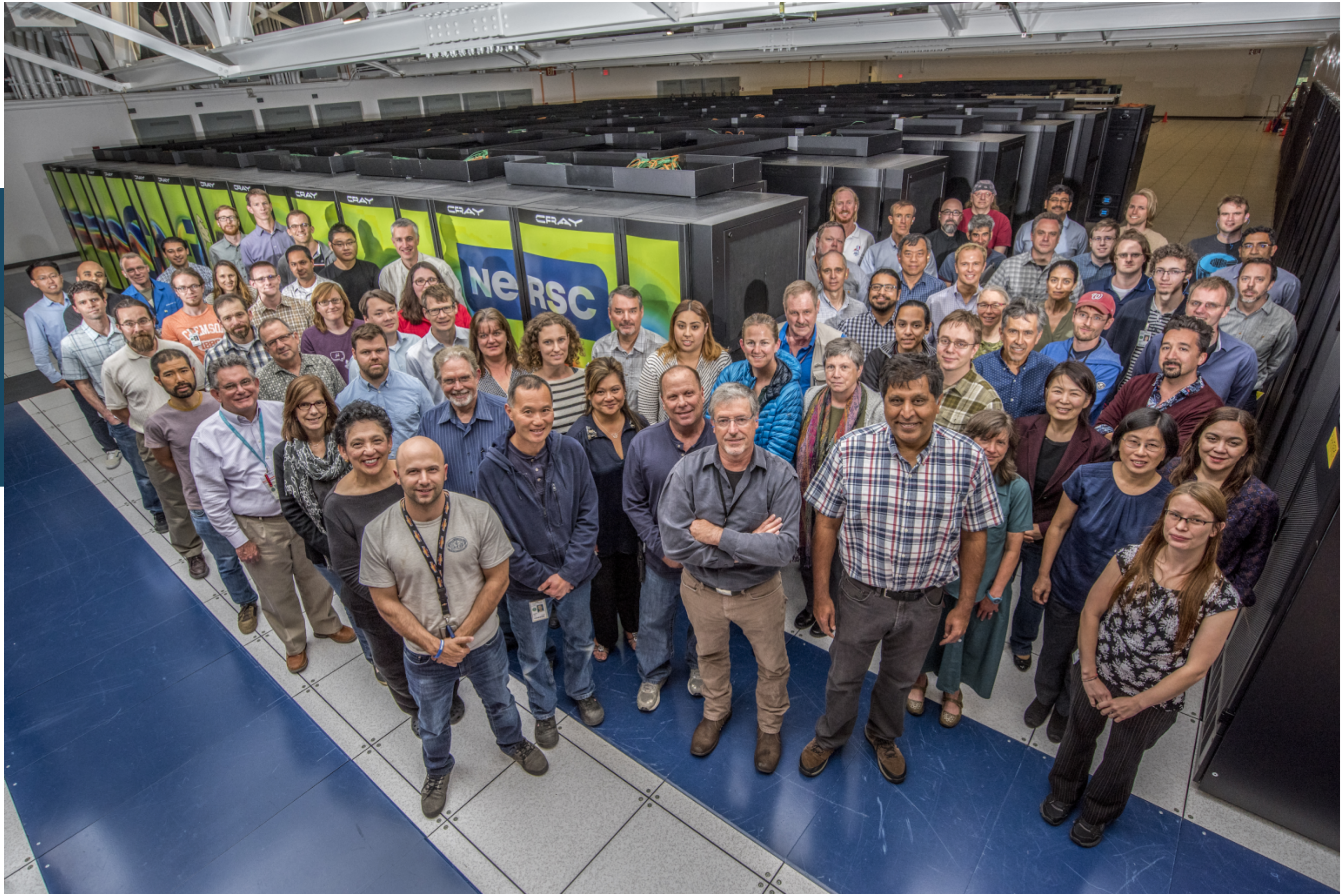


NERSC

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**

Thank You !



We are hiring!

- **Postdoctoral fellows**
  - including Grace Hopper fellowship
- **Application performance specialists**