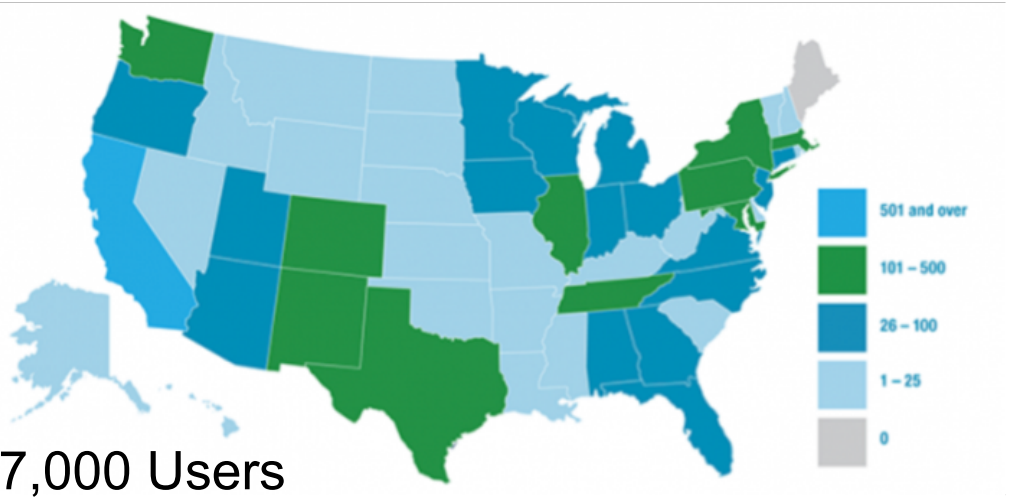
Preparing NERSC Applications for Perlmutter as an Exascale Waypoint

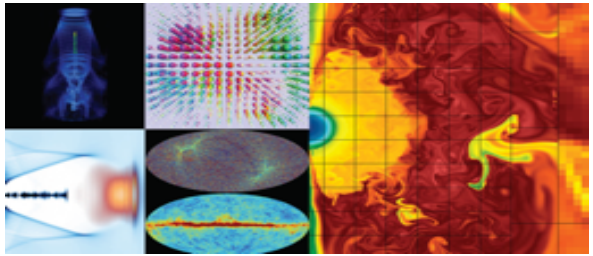
Charlene Yang
Application Performance Specialist
NERSC, LBNL

Pawsey Supercomputing Centre
Mar 26, 2019

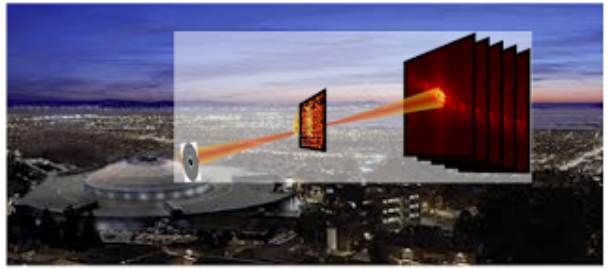
NERSC is the mission High Performance Computing facility for the DOE SC



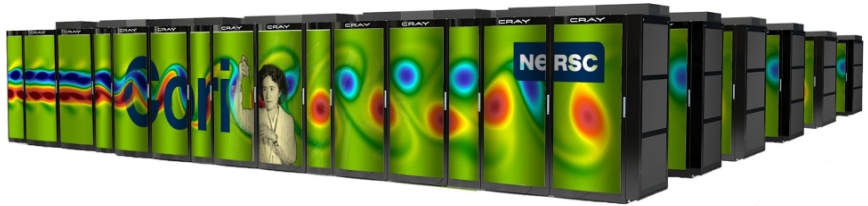
7,000 Users
800 Projects
700 Codes
2000 NERSC citations per year



Simulations at scale



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

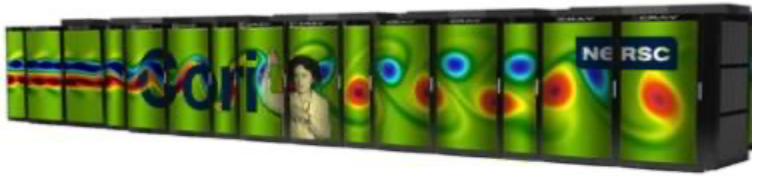


U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

NERSC has a dual mission to advance science and the state-of-the-art in supercomputing

- We collaborate with computer companies years before a system's delivery to deploy advanced systems with new capabilities at large scale
- We provide a highly customized software and programming environment for science applications
- We are tightly coupled with the workflows of DOE's experimental and observational facilities – ingesting tens of terabytes of data each day
- Our staff provide advanced application and system performance expertise to users



Perlmutter was announced 30 Oct 2018

NERSC

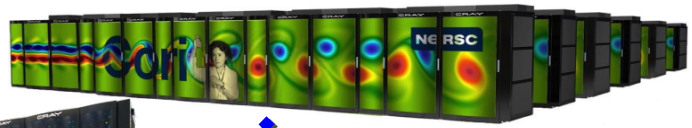
"Continued leadership in high performance computing is vital to America's competitiveness, prosperity, and national security," said U.S. Secretary of Energy Rick Perry. "This advanced new system, created in close partnership with U.S. industry, will give American scientists a powerful new tool of discovery and innovation and **will be an important milestone on the road to the coming era of exascale computing.**"

"We are very excited about the Perlmutter system," said NERSC Director Sudip Dosanjh. "It will provide a significant increase in capability for our users and **a platform to continue transitioning our very broad workload to energy efficient architectures.** The system is optimized for science, and we will collaborate with Cray, NVIDIA and AMD to ensure that Perlmutter meets the **computational and data needs of our users.** We are also launching a major power and cooling upgrade in Berkeley Lab's Shyh Wang Hall, home to NERSC, to prepare the facility for Perlmutter."

NERSC Systems Roadmap

NERSC-7:
Edison
Multicore
CPU

2013

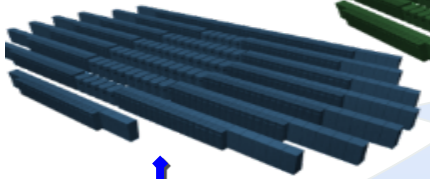


NERSC-8: Cori
Manycore CPU
NESAP Launched:
transition applications to
advanced architectures

2016

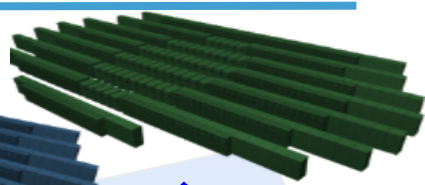
NERSC-9:
CPU and GPU nodes
Continued transition of
applications and support for
complex workflows

2020



NERSC-10:
Exa system

2024



NERSC-11:
Beyond
Moore

2028

Increasing need for energy-efficient architectures

Cori: A pre-exascale supercomputer for the Office of Science workload

Cray XC40 system with 9,600+ Intel Knights Landing compute nodes

68 cores / 96 GB DRAM / 16 GB HBM

Support the entire Office of Science research community

Begin to transition workload to energy efficient architectures

1,600 Haswell processor nodes

NVRAM Burst Buffer 1.5 PB, 1.5 TB/sec

30 PB of disk, >700 GB/sec I/O bandwidth

Integrated with Cori Haswell nodes on Aries network for data / simulation / analysis on one system

Perlmutter is a Pre-Exascale System

Pre-Exascale Systems (2013 – 2016 – 2018 – 2020)

Exascale Systems (2021-2023)

Argonne
IBM BG/Q

Argonne
Intel/Cray KNL

ORNL
IBM/NVIDIA
P9/Volta

LBNL
Cray/NVIDIA/AMD

Argonne
Intel/Cray

ORNL
Cray/NVidia K20

LBNL
Cray/Intel Xeon/KNL

Frontier

ORNL
TBD

LLNL
IBM BG/Q

LANL/SNL
Cray/Intel Xeon/KNL

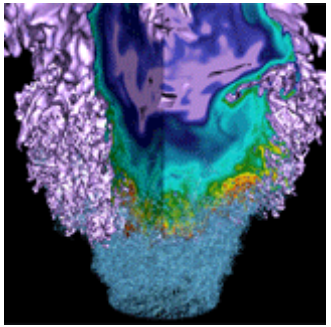
LLNL
IBM/NVIDIA
P9/Volta

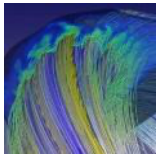
Crossroads

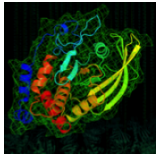
LANL/SNL
TBD

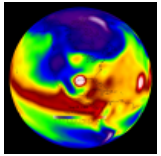
LLNL
TBD

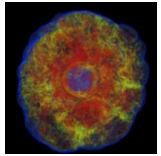
Perlmutter Overview

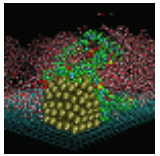












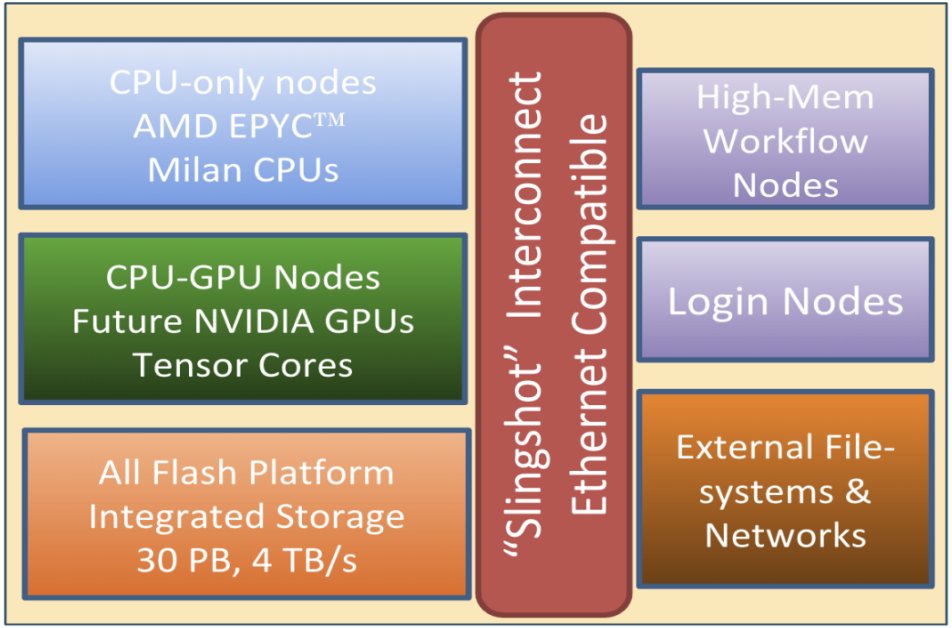
NERSC

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 combine 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



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

AMD CPU nodes

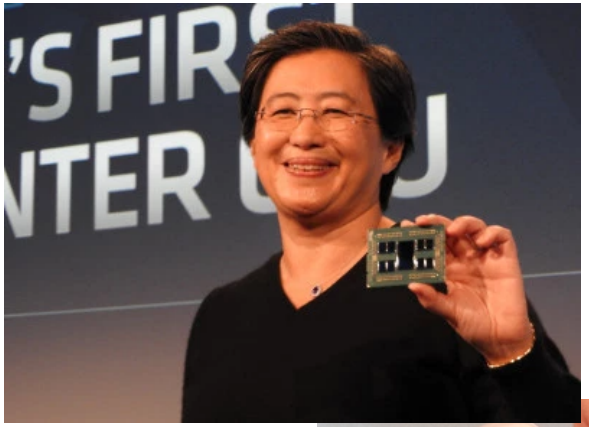
AMD "Milan" CPU
- ~64 cores
- "ZEN 3" cores - 7nm+
- AVX2 SIMD (256 bit)

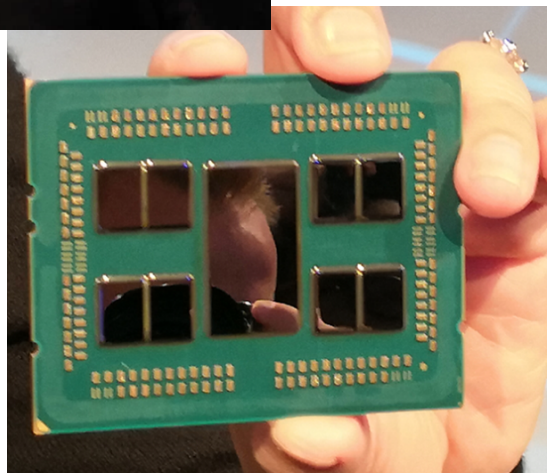
Rome specs

8 channels DDR memory
- >= 256 GiB total per node

1 Slingshot connection
- 1x25 GB/s

~ 1x Cori





GPU nodes

4x NVIDIA “Volta-next” GPU

- > 7 TF
- > 32 GiB, HBM-2
- NVLINK

Volta specs

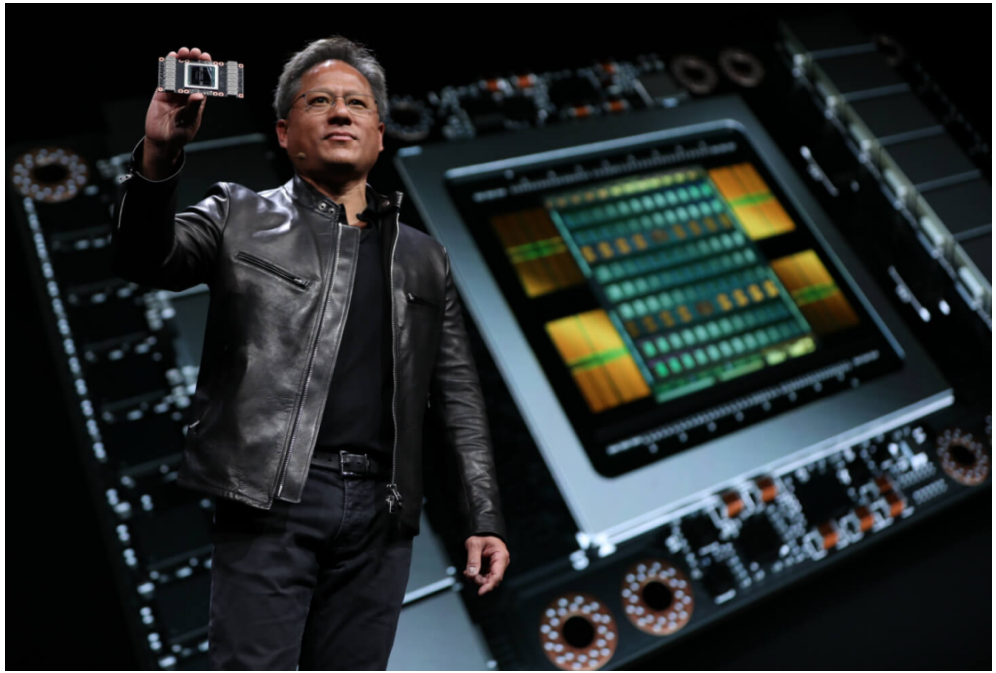
1x AMD CPU

4 Slingshot connections

- 4x25 GB/s

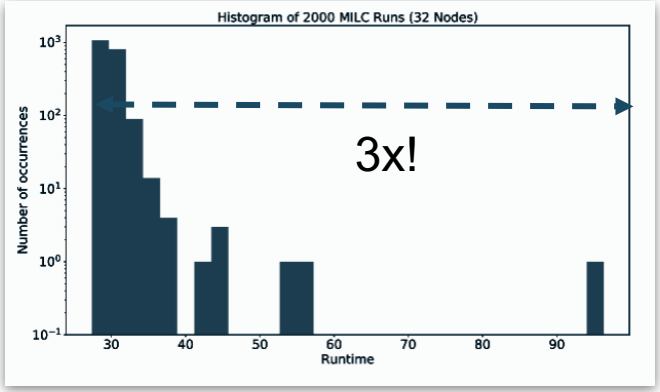
GPU direct, Unified Virtual Memory (UVM)

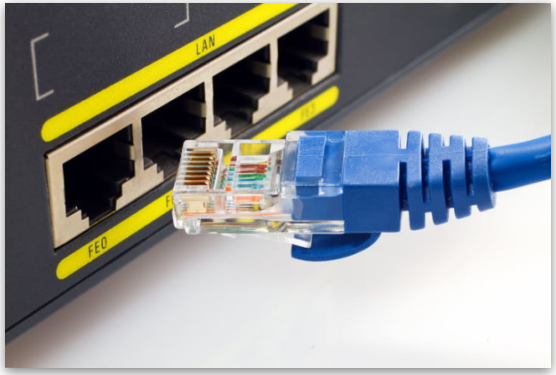
2-3x Cori



Slingshot Network

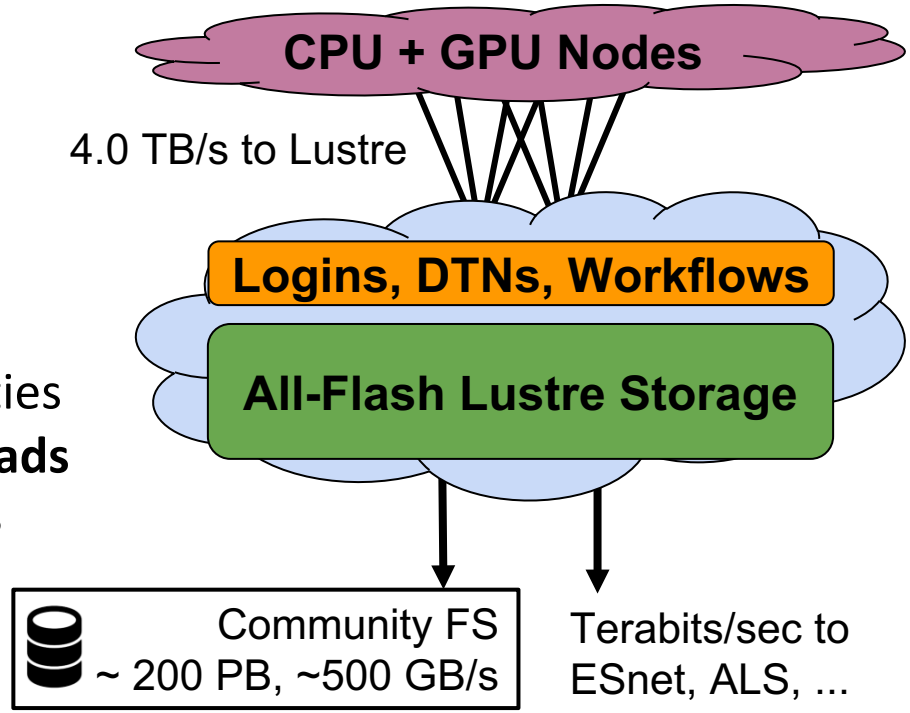
- **High Performance scalable interconnect**
  - Low latency, high-bandwidth, MPI performance enhancements
  - 3 hops between any pair of nodes
  - Sophisticated congestion control and adaptive routing to minimize tail latency
- **Ethernet compatible**
  - Blurs the line between the inside and the outside of the machine
  - Allow for seamless external communication
  - Direct interface to storage





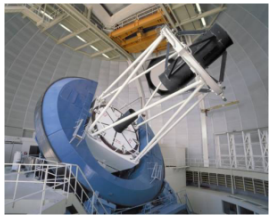
Perlmutter has an All-Flash Filesystem

- **Fast across many dimensions**
  - 4 TB/s sustained bandwidth
  - 7,000,000 IOPS
  - 3,200,000 file creates/sec
- **Usable for NERSC users**
  - 30 PB usable capacity
  - Familiar Lustre interfaces
  - New data movement capabilities
- **Optimized for NERSC data workloads**
  - NEW small-file I/O improvements
  - NEW features for high IOPS, non-sequential I/O

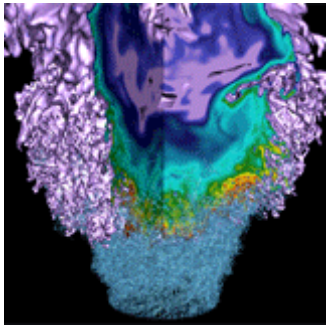


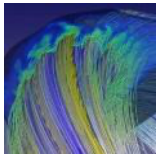
Analytics and Workflow Integration

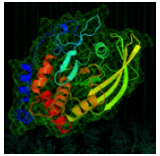
- **Software**
  - Optimized analytics libraries, includes Cray Analytics stack
  - Collaboration with NVIDIA for Python-based data analytics support
  - Support for containers
- **NERSC-9 will aid complex end-to-end workflows**
  - **Slurm co-scheduling of multiple resources and real-time/deadline scheduling**
  - **Workflow nodes: container-based services**
    - Connections to scalable, user workflow pool (via Spin) with network/scheduler access
  - **High-availability workflow architecture and system resiliency for real-time use-cases**

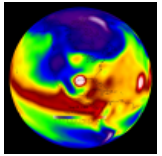


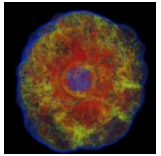
SSI-based Design

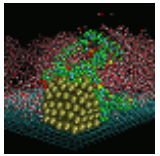












NERSC

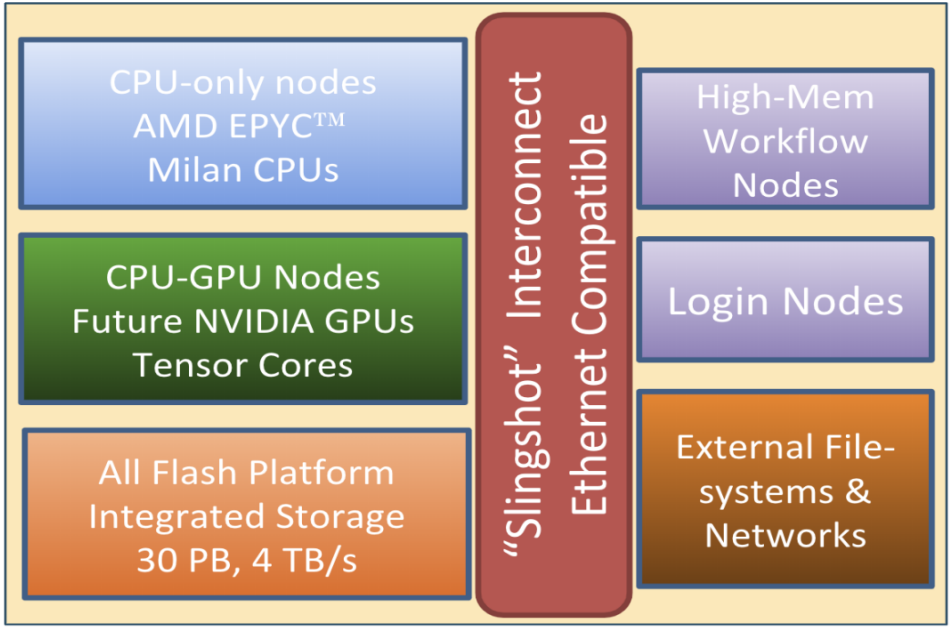
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

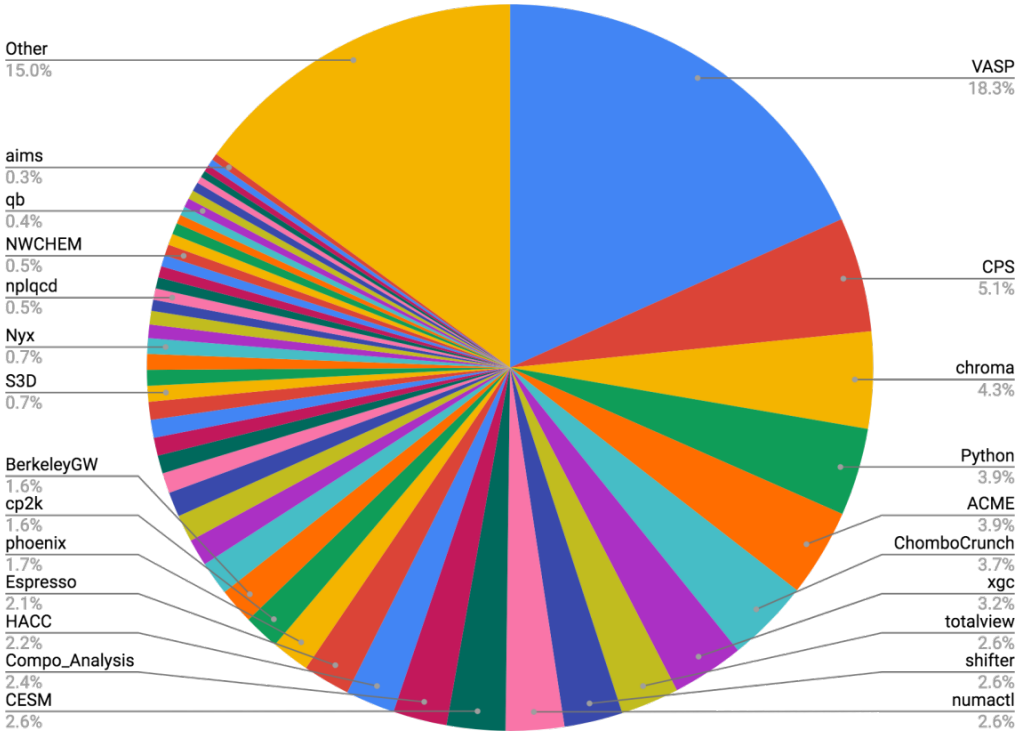
Perlmutter: A System Optimized for Science

- **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 [illegible]
- Dedicated login and high memory nodes to support complex workflows

How do we optimize the size of each partition?

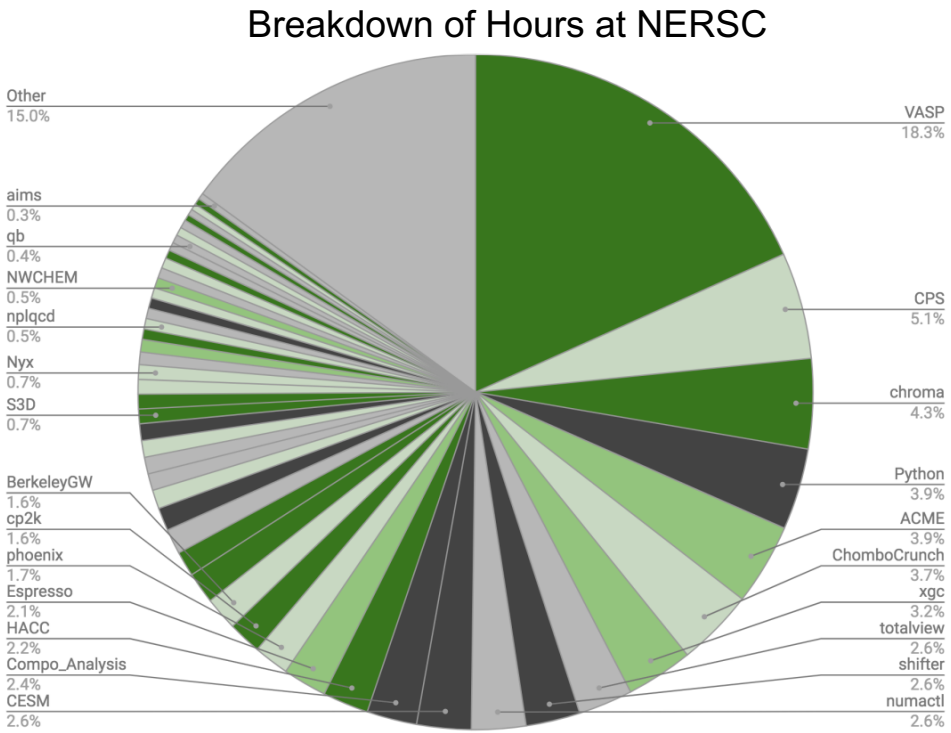


NERSC System Utilization (Aug'17 - Jul'18)



- 3 codes > 25% of the workload
- 10 codes > 50% of the workload
- 30 codes > 75% of the workload
- Over 600 codes comprise the remaining 25% of the workload.

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

How many GPU nodes to buy - Benchmark Suite Construction & Scalable System Improvement

Select codes to represent the anticipated workload

- Include key applications from the current workload.
- Add apps that are expected to contribute significantly to the future workload.

Scalable System Improvement

Measures aggregate performance of HPC machine

- How many more copies of the benchmark can be run relative to the reference machine
- Performance relative to reference machine

SSI - Sustained System Improvement

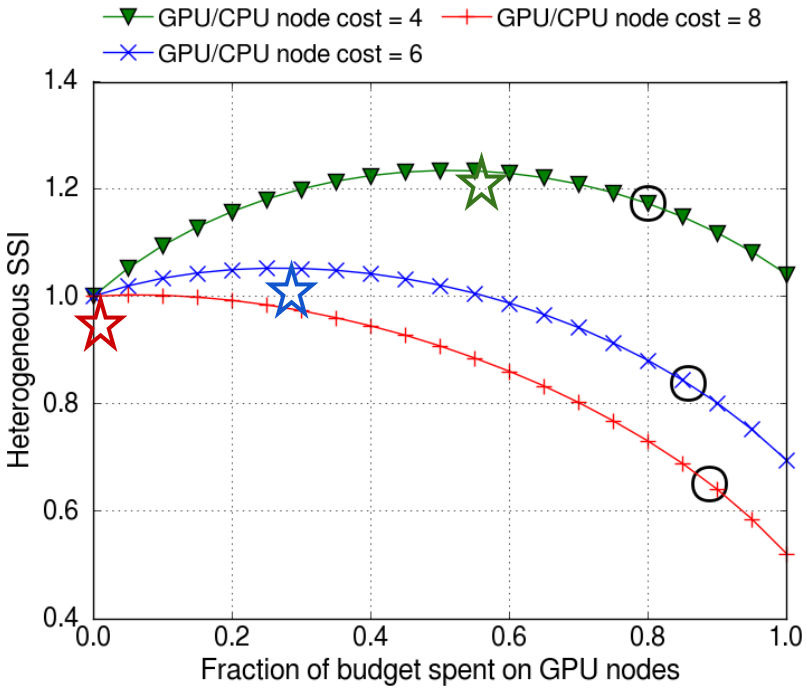
Application	Description
Quantum Espresso	Materials code using DFT
MILC	QCD code using staggered quarks
StarLord	Compressible radiation hydrodynamics
DeepCAM	Weather/Community Atmospheric Model 5
GTC	Fusion PIC code
"CPU Only" (3 Total)	Representative of applications that cannot be ported to GPUs

Hetero system design & price sensitivity: Budget for GPUs increases as GPU price drops

Chart explores an isocost design space

- Vary the budget allocated to GPUs
- Assume GPU enabled applications have performance advantage = 10x per node, 3 of 8 apps are still CPU only.
- Examine GPU/CPU node cost ratio

GPU/CPU $ per node	SSI increase vs. CPU-Only (@ budget %)	
8:1	None	No justification for GPUs
6:1	1.05x @ 25%	Slight justification for up to 50% of budget on GPUs
4:1	1.23x @ 50%	GPUs cost effective up to full system budget, but optimum at 50%



Circles: 50% CPU nodes + 50% GPU nodes
Stars: Optimal system configuration.

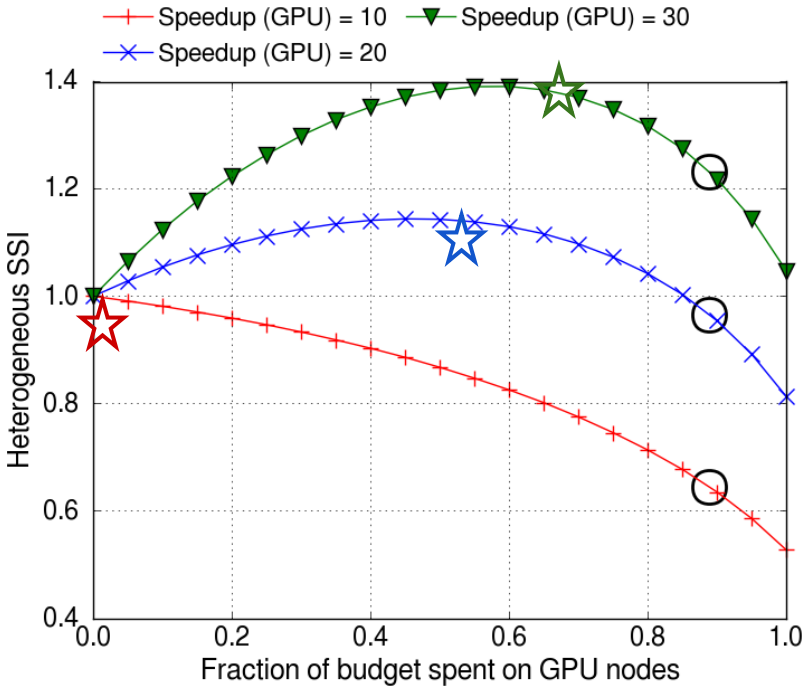
B. Austin, et al., "A Metric for Evaluating Supercomputer Performance in the Era of Extreme Heterogeneity", *9th IEEE International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS18)*, November 12, 2018

AR efforts justify larger GPU partitions

Explore an isocost design space

- Assume 8:1 GPU/CPU node cost ratio.
- Vary the budget allocated to GPUs
- Examine GPU / CPU *performance* gains such as those obtained by software optimization & tuning. 5 of 8 codes have 10x, 20x, 30x speedup.

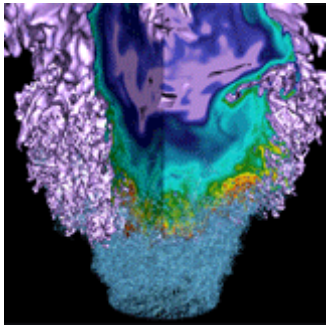
GPU/CPU perf. per node	SSI increase vs. CPU-Only (@ budget %)	
10x	None	No justification for GPUs
20x	1.15x @ 45%	Compare to 1.23x for 10x at 4:1 GPU/CPU cost ratio
30x	1.40x @ 60%	Compare to 3x from NESAP for KNL

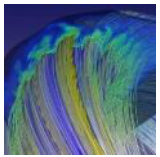


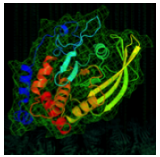
Circles: 50% CPU nodes + 50% GPU nodes
Stars: Optimal system configuration.

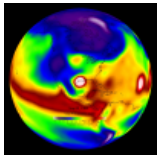
B. Austin, et al., "A Metric for Evaluating Supercomputer Performance in the Era of Extreme Heterogeneity", *9th IEEE International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS18)*, November 12, 2018

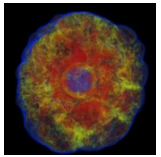
NESAP Overview

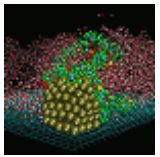












NERSC

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, 800 projects, and 700 codes to run on advanced architectures like Perlmutter and beyond?

Application Readiness Strategy for Perlmutter

NERSC Exascale Science Application Program (NESAP)

- http://nersc.gov/users/application-performance/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**

GPU Transition Path for Apps

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

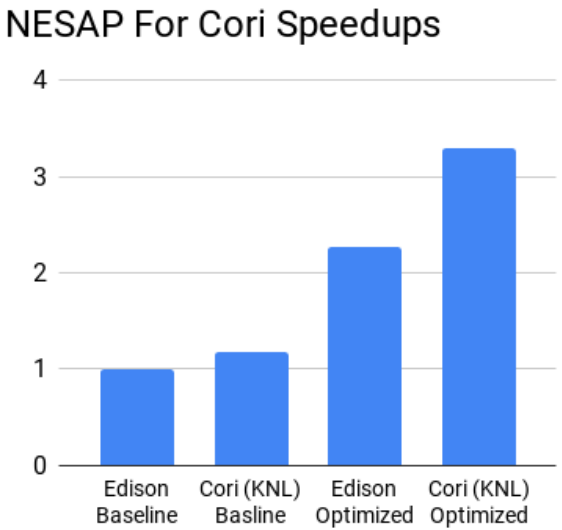
1. Identifying and exploiting on-node parallelism
2. Understanding and improving data-locality within the memory hierarchy

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

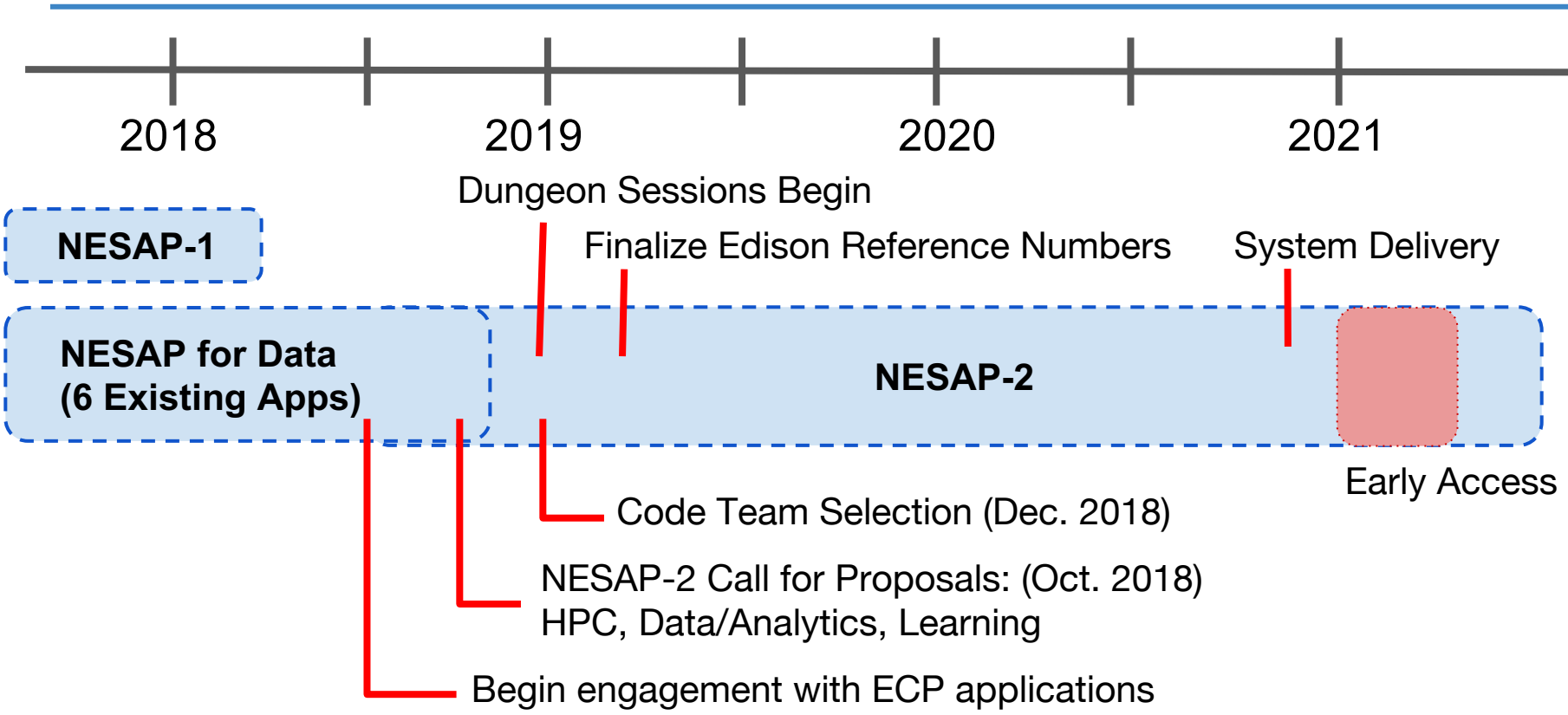


NESAP for Perlmutter

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)

Tier 1 apps have higher priority than Tier 2!

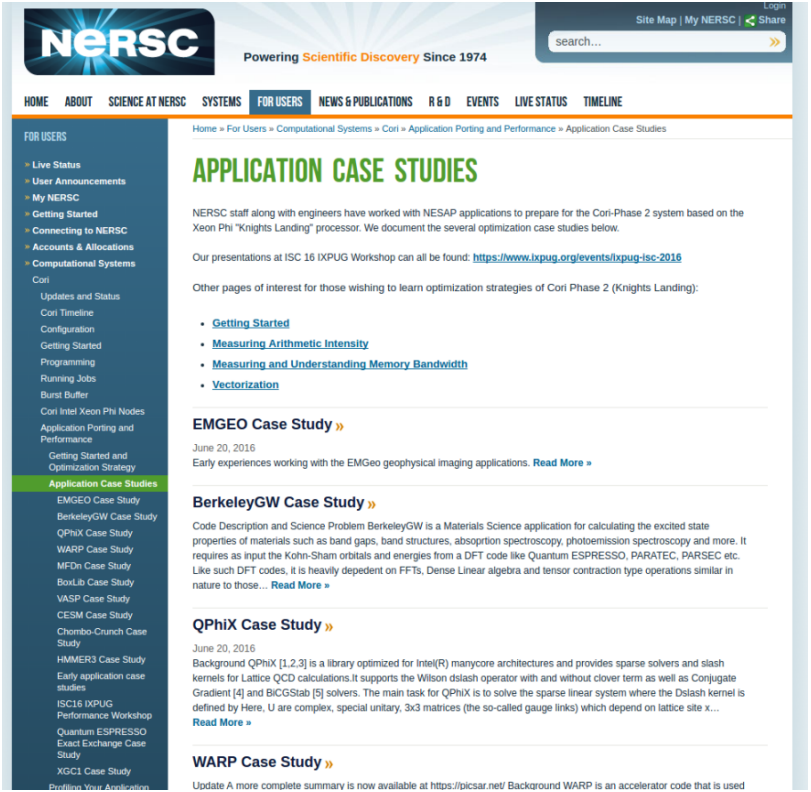
Support for NESAP Teams

Benefit	Tier 1	Tier 2
Early Access to Perlmutter	yes	eligible
Hack-a-thon with vendors	yes	eligible
Training resources	yes	yes
Additional NERSC hours from Director's Reserve	yes	eligible
NERSC funded postdoctoral fellow	eligible	no
Commitment of NERSC staff assistance	yes	no

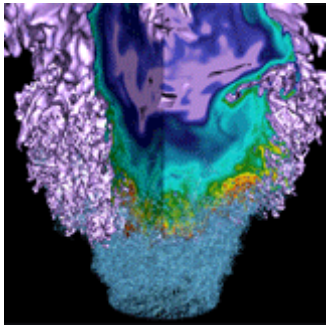
Training, Case Studies and Documentation

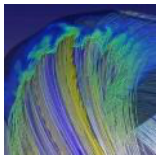
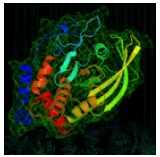
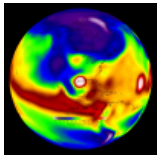
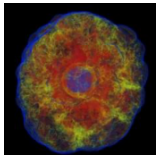
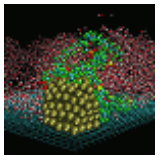
- For those teams not in NESAP, there will be a robust training program

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



Programming & Performance Portability



NERSC

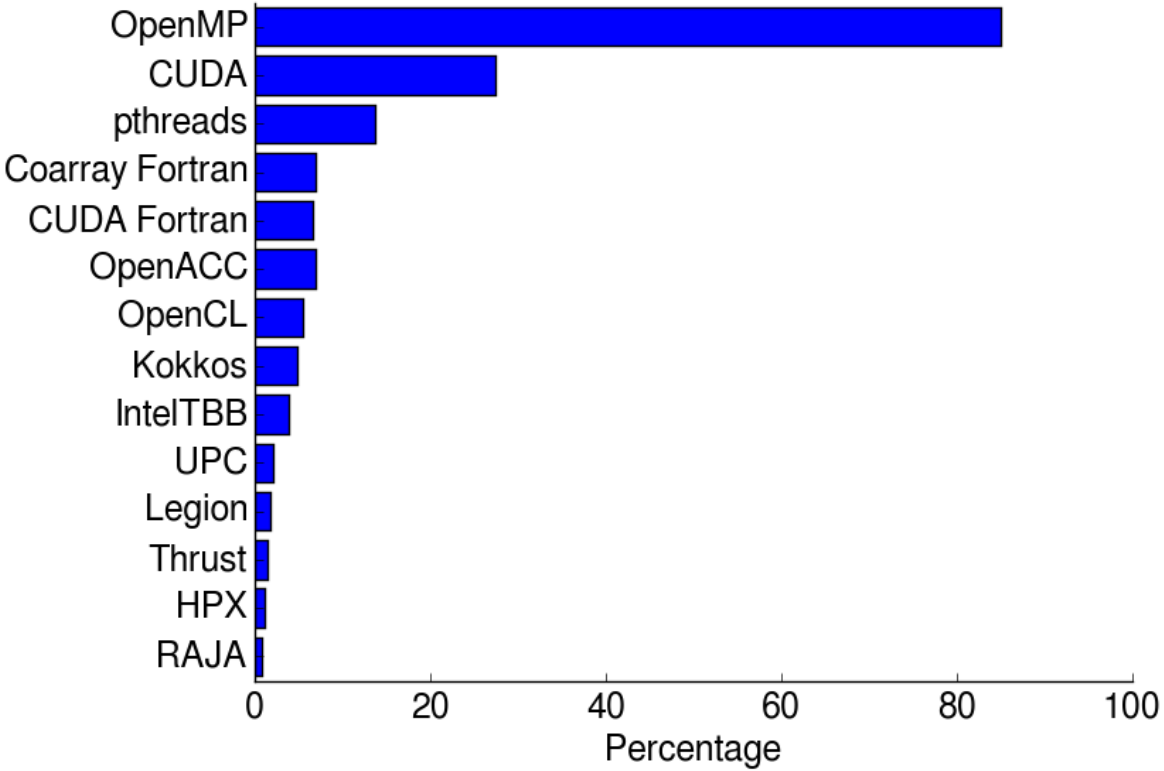
Performance Portability Strategies

- 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)

OpenMP is the most popular non-MPI parallel programming technique



- **Results from ERCAP 2017 user survey**
  - Question answered by 328 of 658 survey respondents
- **Total exceeds 100% because some applications use multiple techniques**

Breaking News !

BERKELEY LAB

BERKELEY LAB COMPUTING SCIENCES
LAWRENCE BERKELEY NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

A-Z INDEX | PHONE BOOK | CAREERS | SHARE | FOLLOW

Home | About | News & Media | Seminars | Careers | Awards | Safety | For Staff | search...

Home » News & Media » News » NERSC, NVIDIA to Partner on Compiler Development for Perlmutter System

NEWS & MEDIA

- News
- CS In the News
- InTheLoop

NERSC, NVIDIA to Partner on Compiler Development for Perlmutter System

MARCH 21, 2019

The National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory (Berkeley Lab) has signed a contract with NVIDIA to enhance GPU compiler capabilities for Berkeley Lab's next-generation Perlmutter supercomputer.

In October 2018, the U.S. Department of Energy (DOE) announced that NERSC had signed a contract with Cray for a pre-exascale supercomputer named "Perlmutter," in honor of Berkeley Lab's Nobel Prize-winning astrophysicist Saul Perlmutter. The Cray Shasta machine, slated to be delivered in 2020, will be a heterogeneous system comprising both CPU-only and GPU-accelerated cabinets. It will include a new Cray system interconnect designed for data-centric computing; NVIDIA GPUs with new Tensor Core technology; CPU-only nodes based on next-generation AMD EPYC CPUs; direct liquid cooling; and an all-flash scratch filesystem that will move data at a rate of more than 4 terabytes/sec.

Ensuring OpenMP is ready for Perlmutter

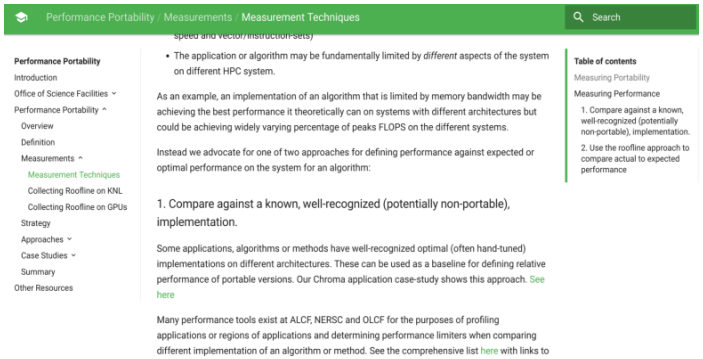
- **NERSC will collaborate with NVIDIA to enable OpenMP GPU acceleration with PGI compilers**
  - NERSC application requirements will help prioritize OpenMP and base language features on the GPU
  - Co-design of NESAP-2 applications to enable effective use of OpenMP on GPUs and guide PGI optimization effort
- **We want to hear from the larger community**
  - Tell us your experience, including what OpenMP techniques worked / failed on the GPU
  - Share your OpenMP applications targeting GPUs

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 is a Member



NERSC is leading development of performanceportability.org

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

Work with NVIDIA and SLATE to better support SCALAPACK

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 and preliminary work [1-2]

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

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 [3]

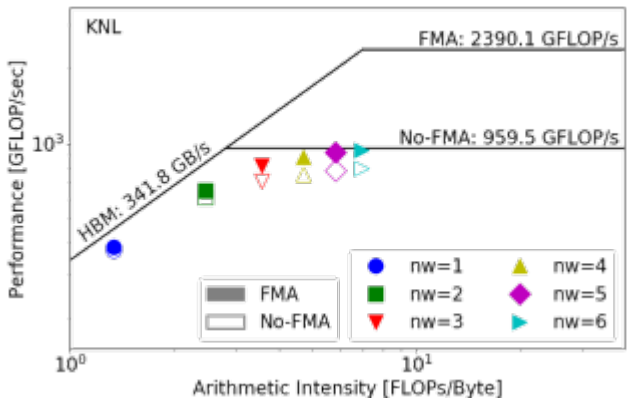
[1] S. J. Pennycook, et al, "A metric for performance portability," arXiv:1611.07409, 2016.
[2] Yang, Charlene, et al. "An empirical roofline methodology for quantitatively assessing performance portability." *P3HPC Workshop*. 2018.
[3] S. Williams, *et al*, "Roofline: An insightful visual performance model for multicore architectures," Communications of the ACM, vol. 52, no. 4, pp. 65–76, 2009.

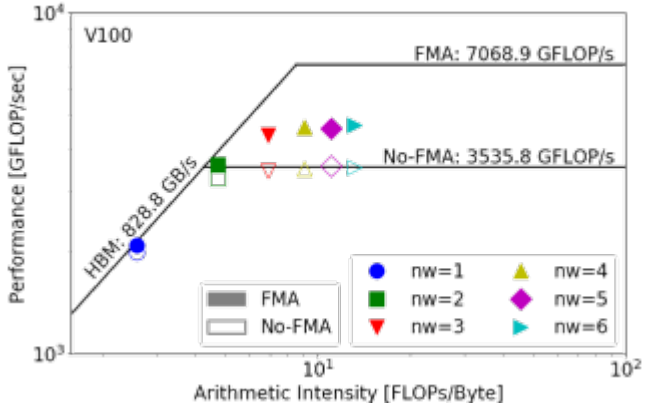
Methodology to Measure Perf. Port.

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.

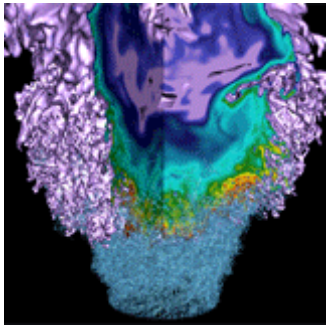
An example: GPP kernel from BerkeleyGW (Perf. Port. Scores)

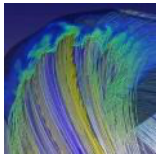
	Architectural Efficiency	nw=1	nw=2	nw=3	nw=4	nw=5	nw=6
FMA	KNL	84.98%	77.50%	66.77%	55.28%	46.56%	39.65%
	V100	97.36%	91.50%	76.70%	65.44%	65.07%	66.38%
	Performance Portability	**90.76%**	**83.92%**	**71.39%**	**59.93%**	**54.28%**	**49.65%**
No-FMA	KNL	82.06%	72.95%	73.74%	78.72%	81.28%	82.81%
	V100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

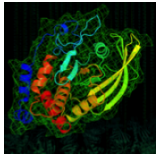
- Roofline captures application bottlenecks, machine balance, problem size, *etc*
- Perf. Port. metric captures performance changes across architectures

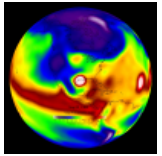
Yang, Charlene, et al. "An empirical roofline methodology for quantitatively assessing performance portability." *International Workshop on Performance, Portability and Productivity in HPC P3HPC*. 2018.

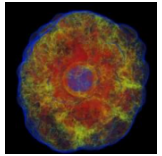
Summary

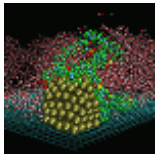












NERSC

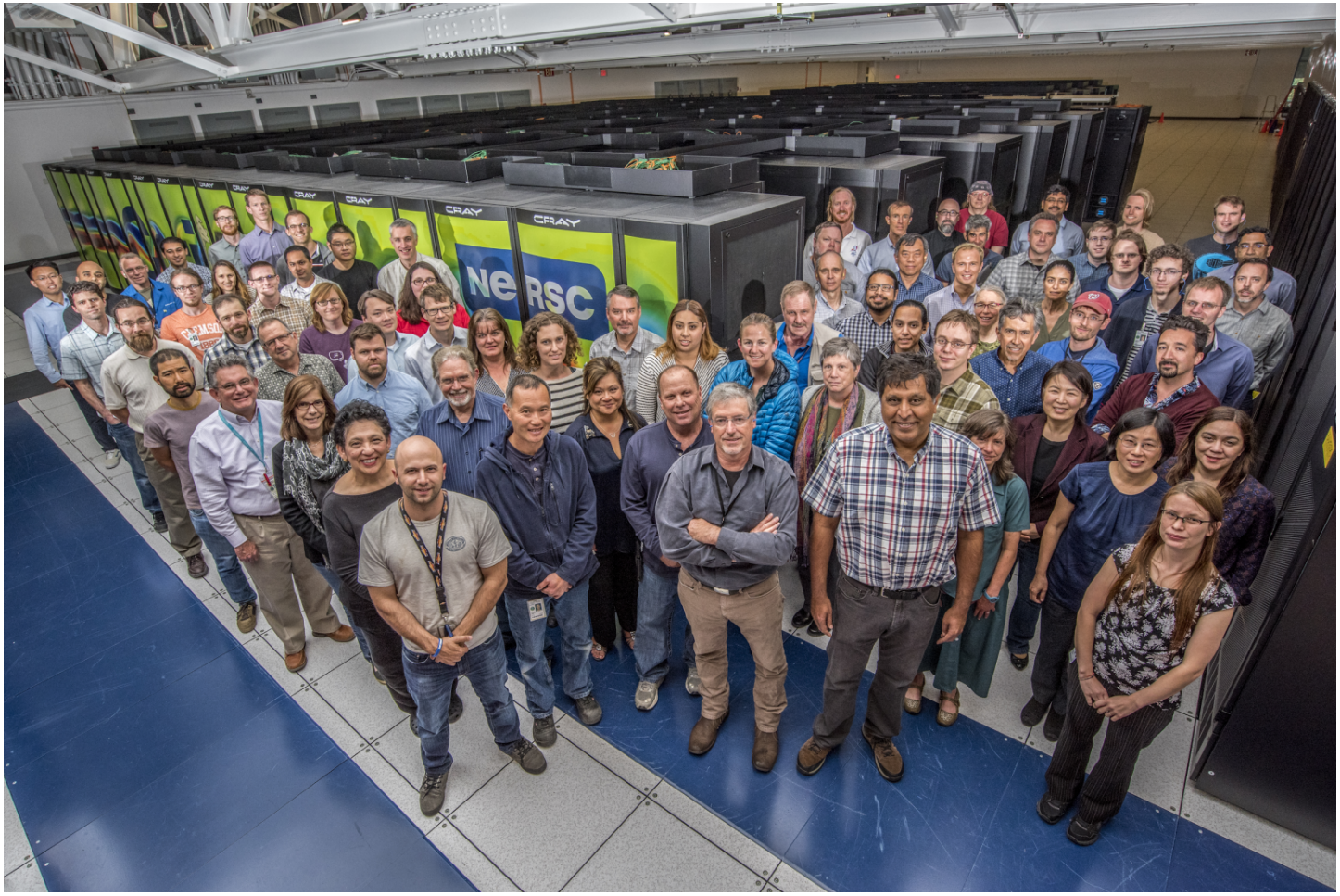
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Perlmutter: 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**

NERSC
ASK ME ABOUT
SAUL
CRAY


CRAY
NERSC