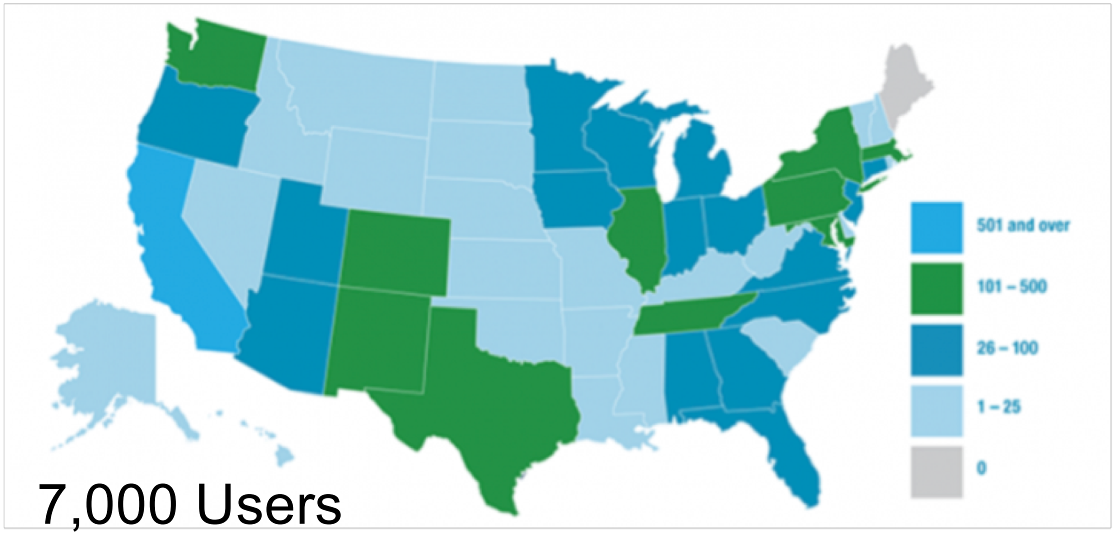
Perlmutter - A 2020 Pre-Exascale GPU-accelerated System for NERSC - Architecture and Application Performance Optimization

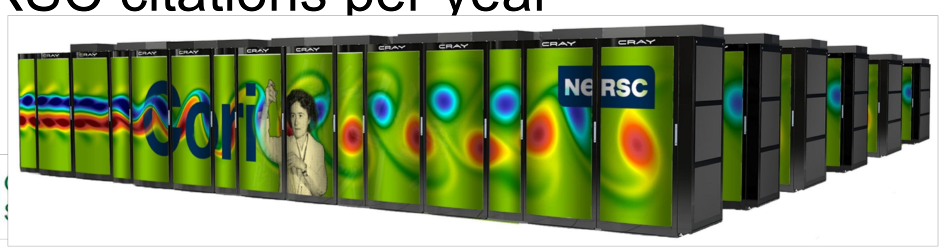
Nicholas J. Wright
Perlmutter Chief Architect

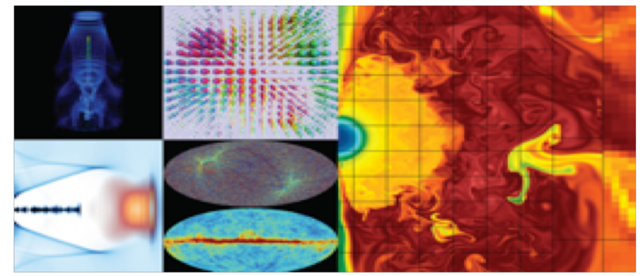
GPU Technology Conference
San Jose
March 21 2019

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

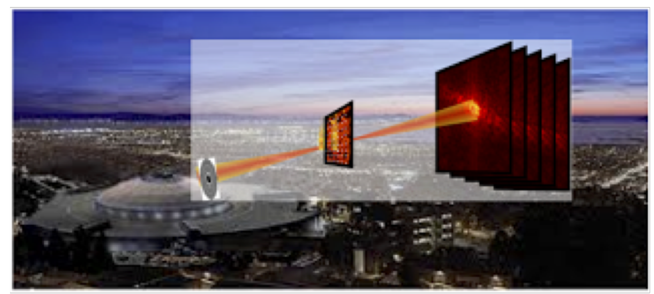


7,000 Users
800 Projects
700 Codes
2000 NERSC citations per year





Simulations at scale



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

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

NERSC

Perlmutter is a Pre-Exascale System

Pre-Exascale Systems

2013 2016 2018 2020

Argonne
IBM BG/Q

ORNL
Cray/NVidia K20

LLNL
IBM BG/Q

Argonne
Intel/Cray KNL

LBNL
Cray/Intel Xeon/KNL

LANL/SNL
Cray/Intel Xeon/KNL

ORNL
IBM/NVIDIA
P9/Volta

LLNL
IBM/NVIDIA
P9/Volta

PERLMUTTER

LBNL
Cray/NVIDIA/AMD

Exascale Systems

2021-2023

Argonne
Intel/Cray

Frontier

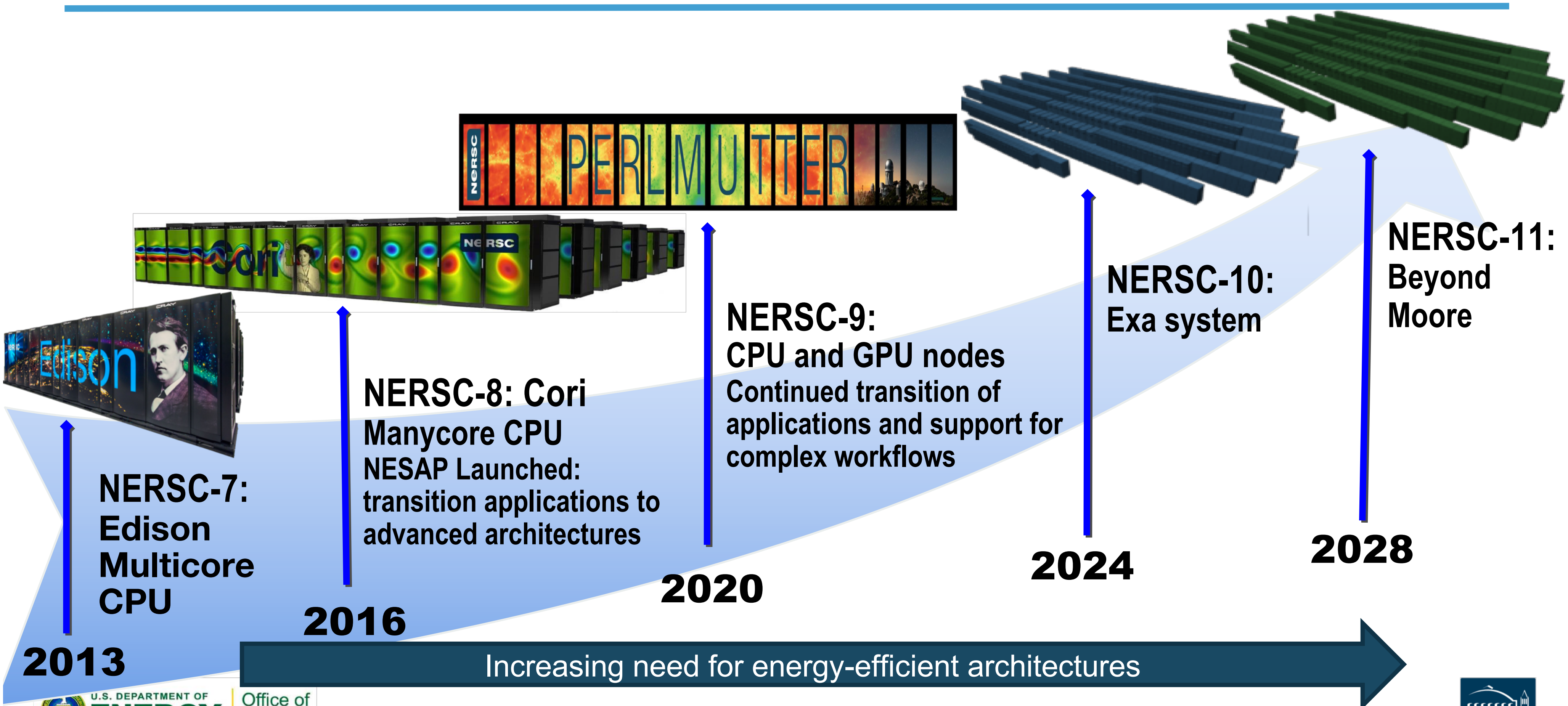
ORNL
TBD

Crossroads

LANL/SNL
TBD

LLNL
TBD

NERSC Systems Roadmap



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

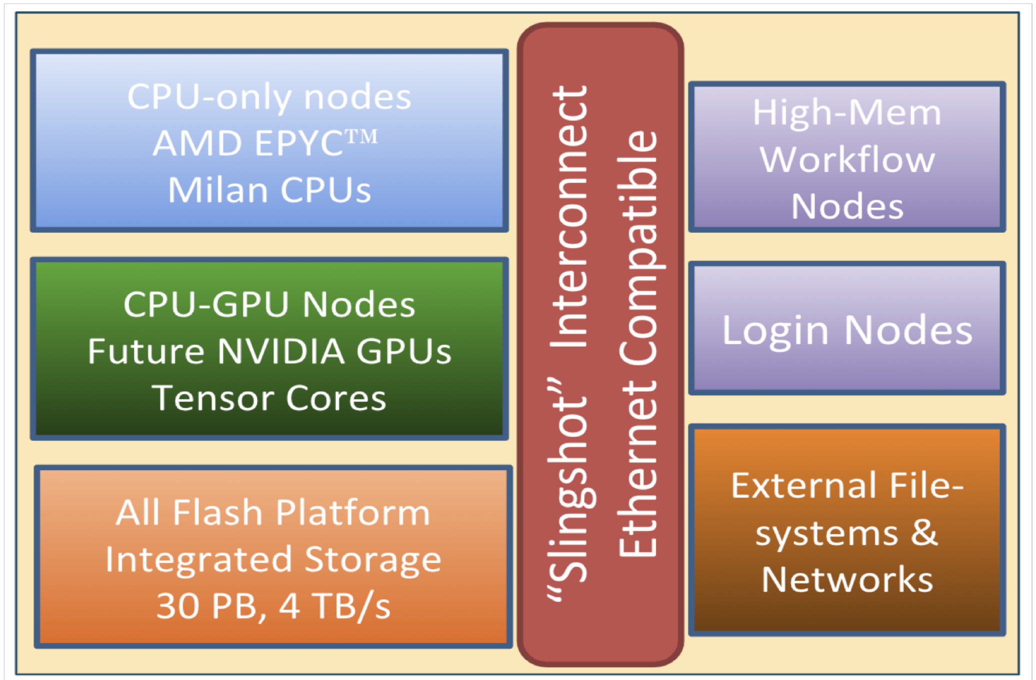
NVRAM Burst Buffer 1.5 PB, 1.5 TB/sec

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: 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



GPU nodes

4x NVIDIA "Volta-next" GPU

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

Volta specs

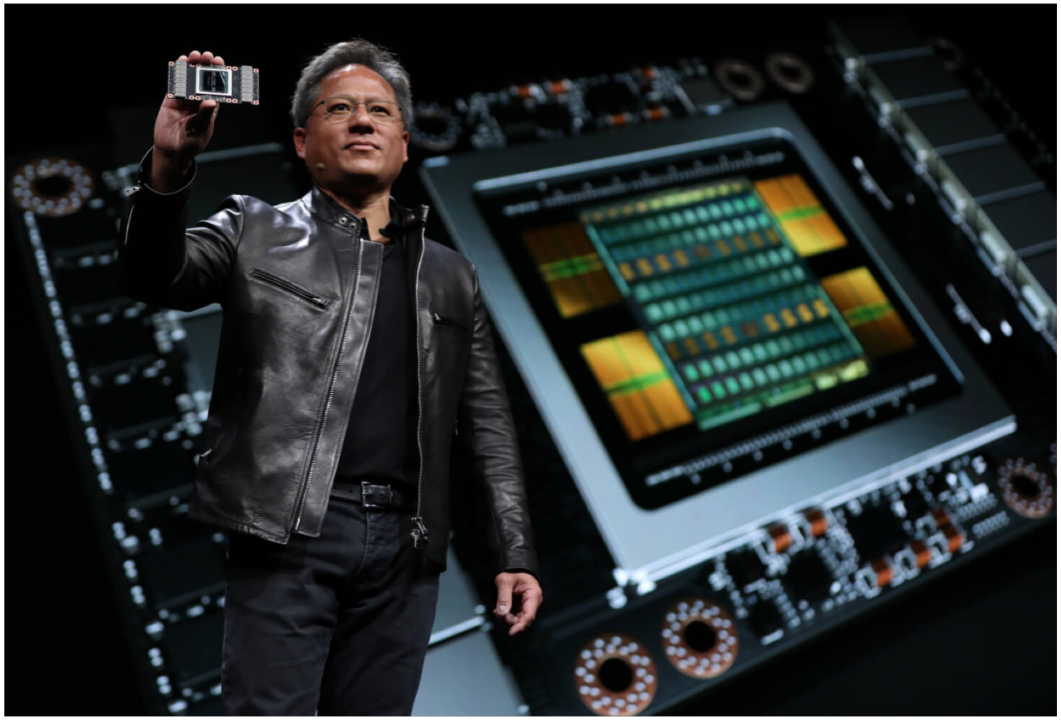
1x AMD CPU

4 Slingshot connections

- 4x25 GB/s

GPU direct, Unified Virtual Memory (UVM)

2-3x Cori



AMD CPU nodes

AMD “Milan” CPU

- ~64 cores
- “ZEN 3” cores - 7nm+
- AVX2 SIMD (256 bit)

Rome specs

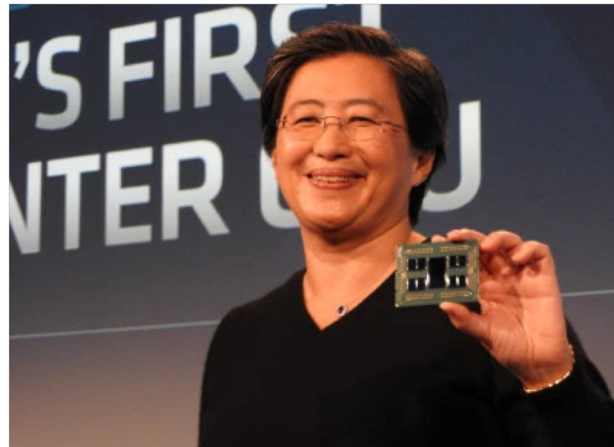
8 channels DDR memory

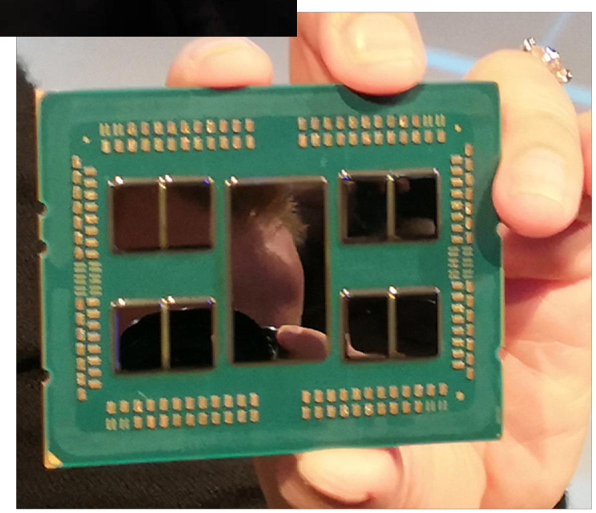
- >= 256 GiB total per node

1 Slingshot connection

- 1x25 GB/s

~ 1x Cori





U.S. DEPARTMENT OF ENERGY | Office of Science

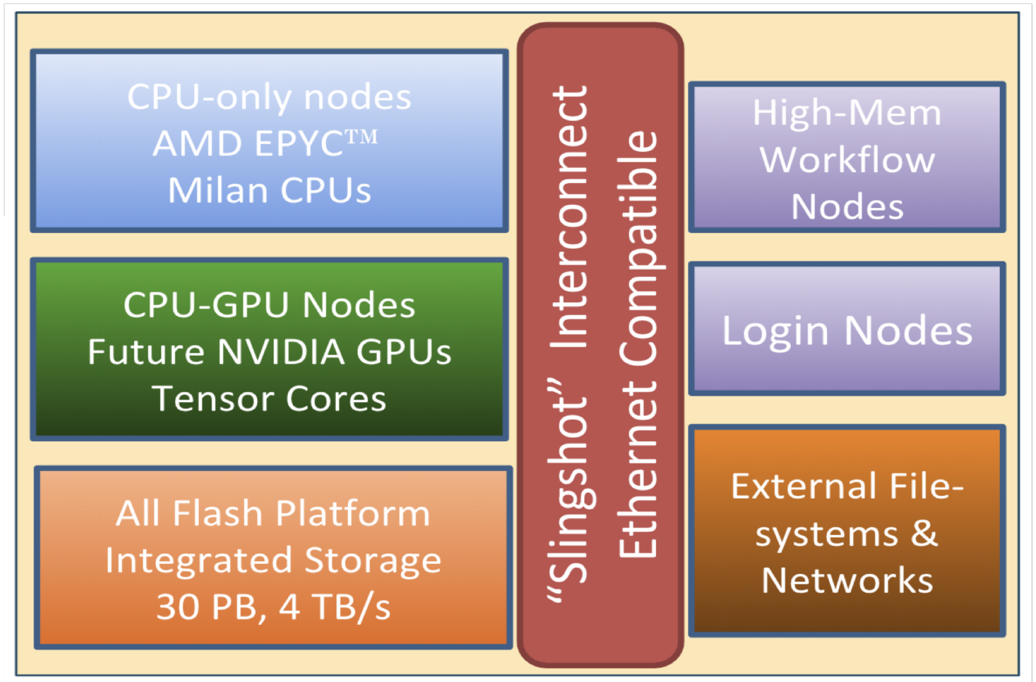
BERKELEY LAB

Perlmutter: A System Optimized for Science

NERSC

- **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

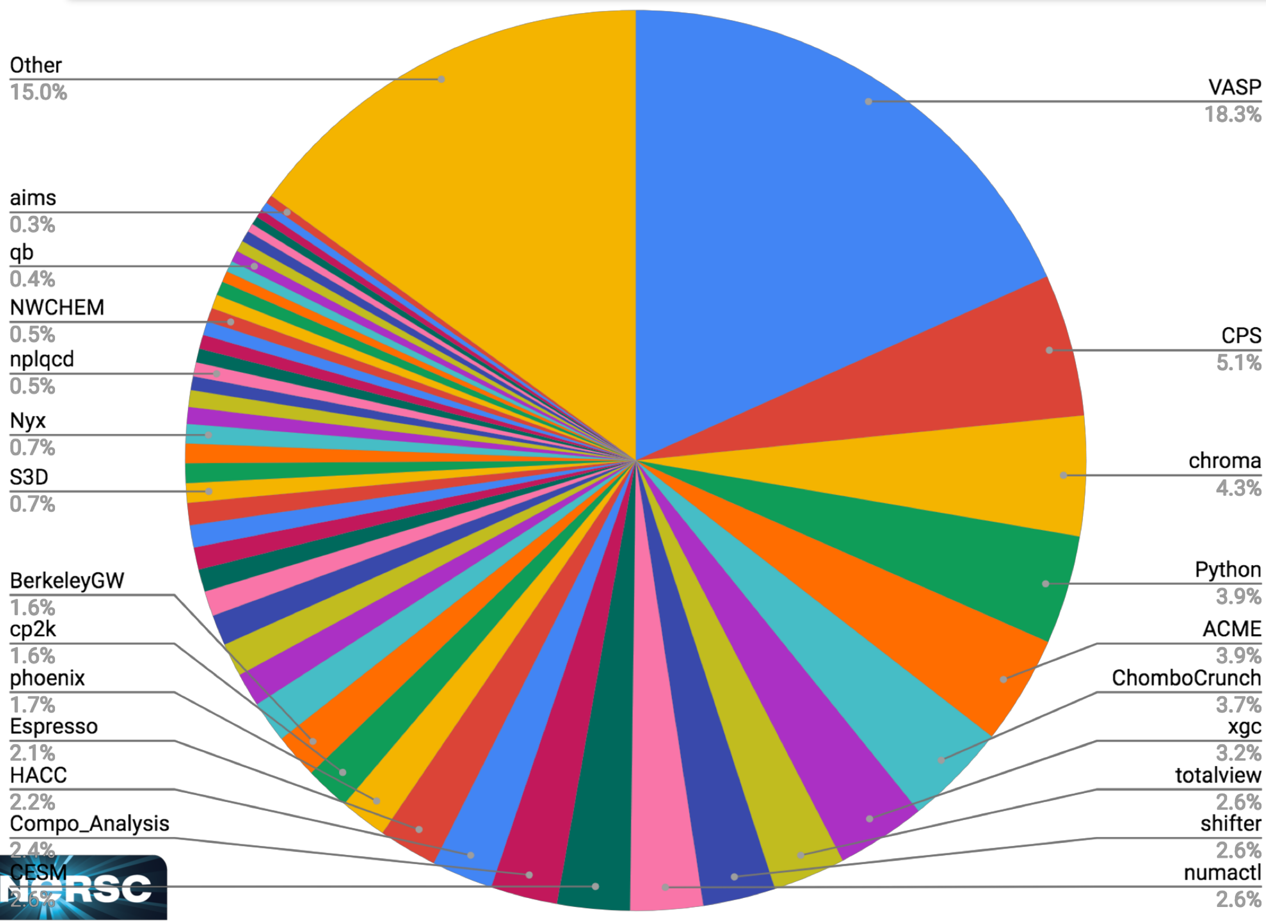
How do we optimize the size of each partition?



U.S. DEPARTMENT OF ENERGY | Office of Science

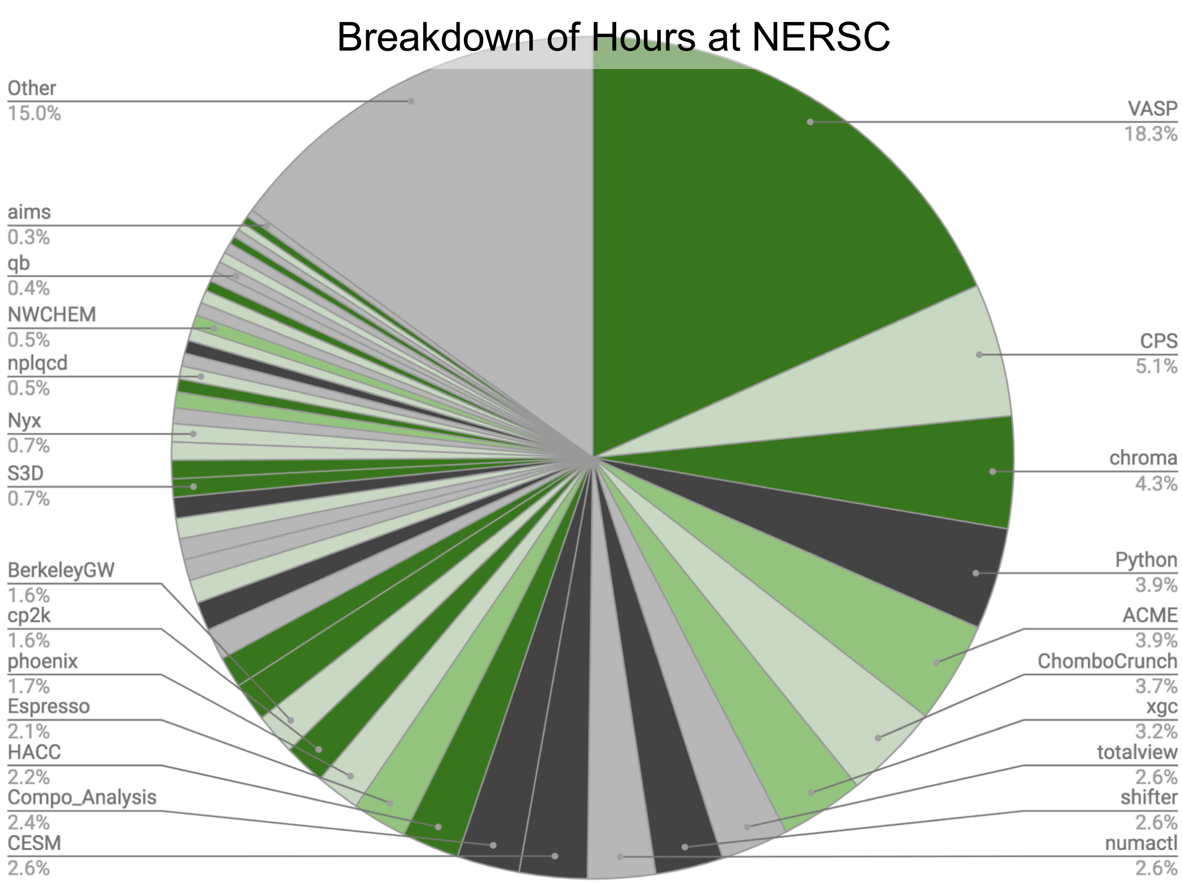
BERKELEY LAB

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 be 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

$$SSI = \left\langle \frac{\#Nodes \times Jobsize \times Perf\_per\_node}{\#Nodes_{Ref} \times Jobsize_{Ref} \times Perf\_per\_node_{Ref}} \right\rangle$$

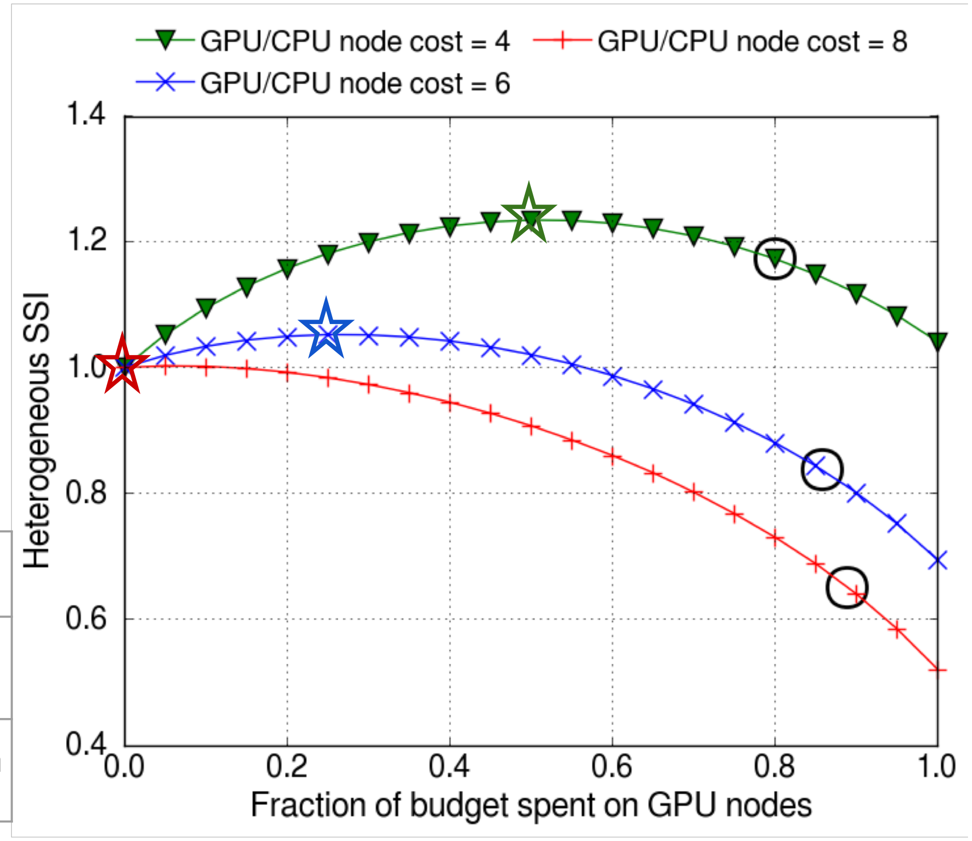
Application	Description
Quantum Espresso	Materials code using DFT
MILC	QCD code using staggered quarks
StarLord	Compressible radiation hydrodynamics
DeepCAM	Weather/Community Atmospheric Model 5
GTC	Fusion PIC code
“CPU Only” (3 Total)	Representative of applications that cannot be ported to GPUs

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.

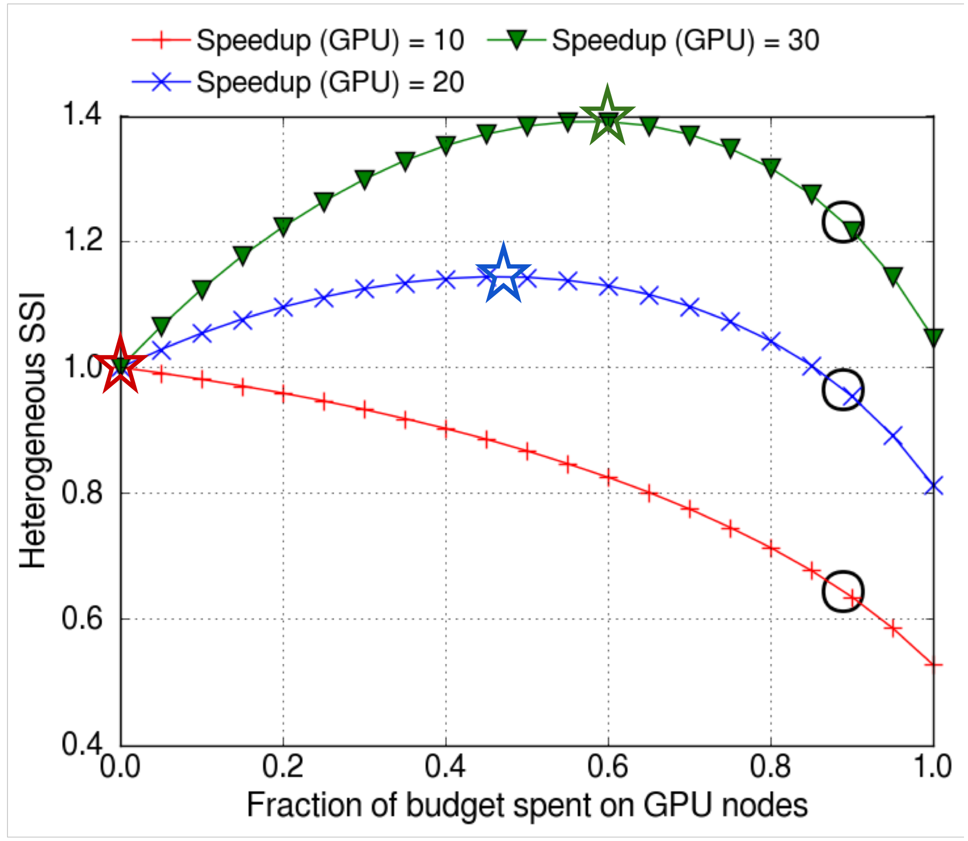
B. Austin, C. Daley, D. Doerfler, J. Deslippe, B. Cook, B. Friesen, T. Kurth, C. Yang, N. J. Wright, "A Metric for Evaluating Supercomputer Performance in the Era of Extreme Heterogeneity", *9th IEEE International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS18),* November 12, 2018,

Application readiness 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, C. Daley, D. Doerfler, J. Deslippe, B. Cook, B. Friesen, T. Kurth, C. Yang, N. J. Wright, "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,

Application Readiness Strategy for Perlmutter

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?

- 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**

NERSC

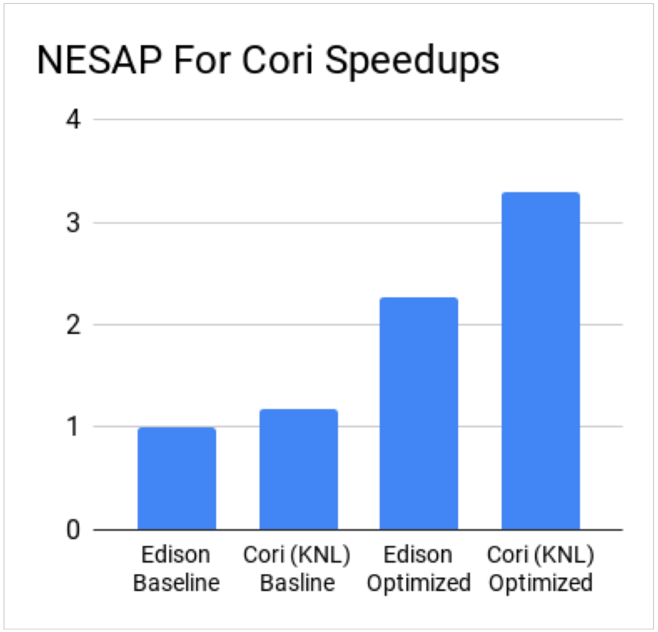
GPU Transition Path for Apps

NESAP for Perlmutter will extend activities from NESAP

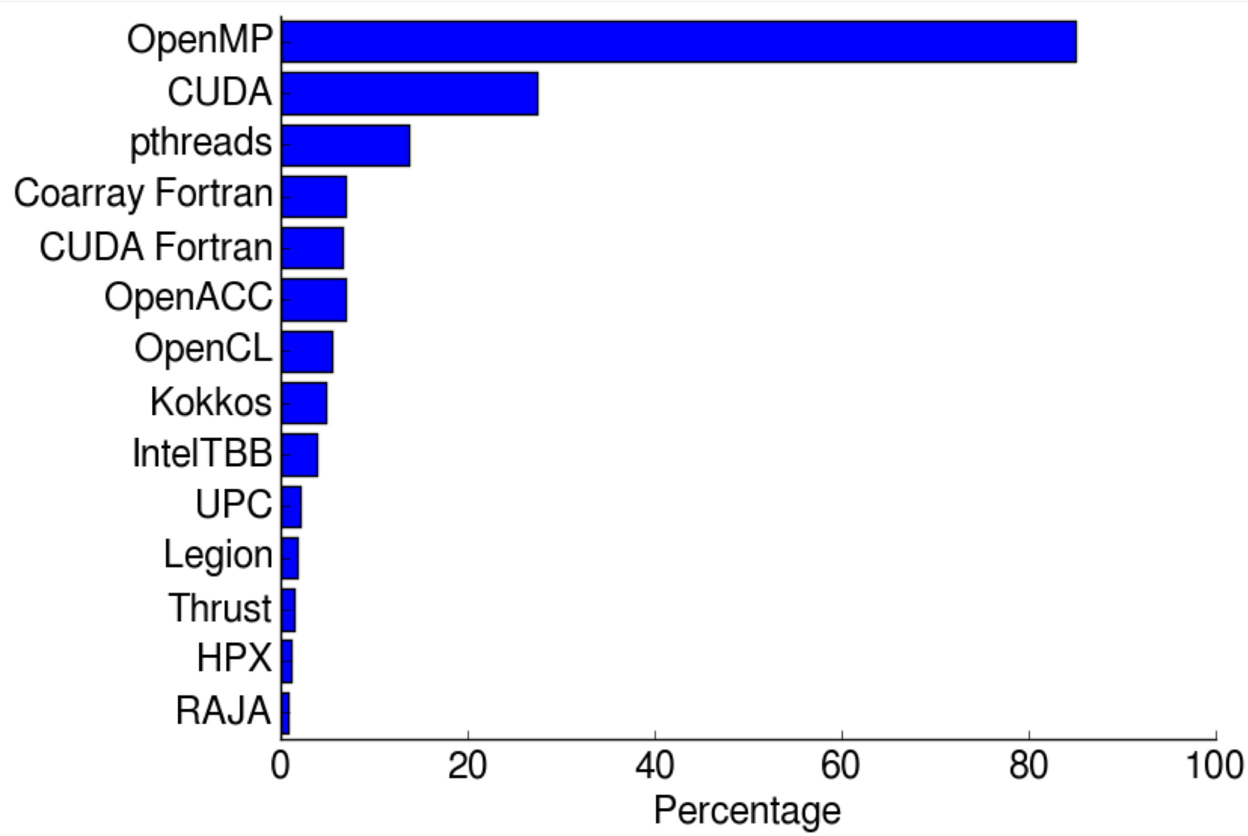
1. Identifying and exploiting on-node parallelism
2. Understanding and improving data-locality within the memory hierarchy

What's New for NERSC Users?

1. Heterogeneous compute elements
2. Identification and exploitation of even more parallelism
3. Emphasis on performance-portable programming approach:
   – Continuity from Cori through future NERSC systems



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**

OpenMP meets the needs of the NERSC workload

- **Supports C, C++ and Fortran**
  - The NERSC workload consists of ~700 applications with a relatively equal mix of C, C++ and Fortran
- **Provides portability to different architectures at other DOE labs**
- **Works well with MPI: hybrid MPI+OpenMP approach successfully used in many NERSC apps**
- **Recent release of OpenMP 5.0 specification – the third version providing features for accelerators**
  - Many refinements over this five year period

Ensuring OpenMP is ready for Perlmutter CPU+GPU nodes

- **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

Breaking News !

NERSC

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.

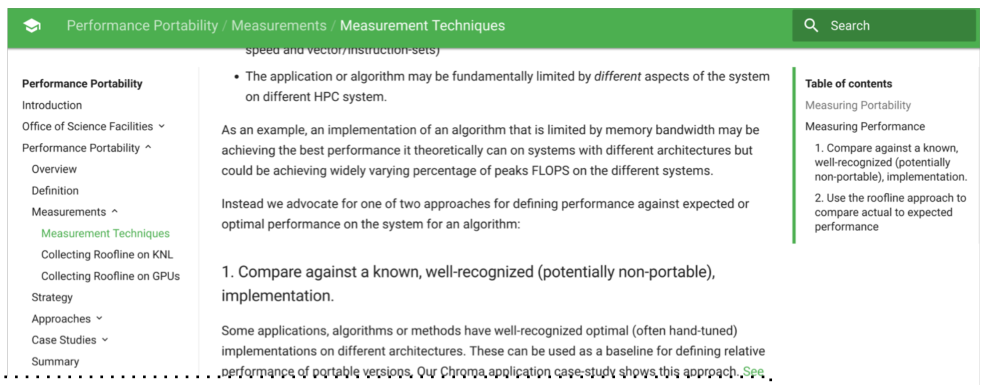
In October 2018, the U.S. Department of Energy (DOE) announced that NERSC had signed a contract with Cray for a pre-exascale supercomputer named "Perlmutter," in honor of Berkeley Lab's Nobel Prize-winning astrophysicist Saul Perlmutter. The Cray Shasta machine, slated to be delivered in 2020, will be a heterogeneous system

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

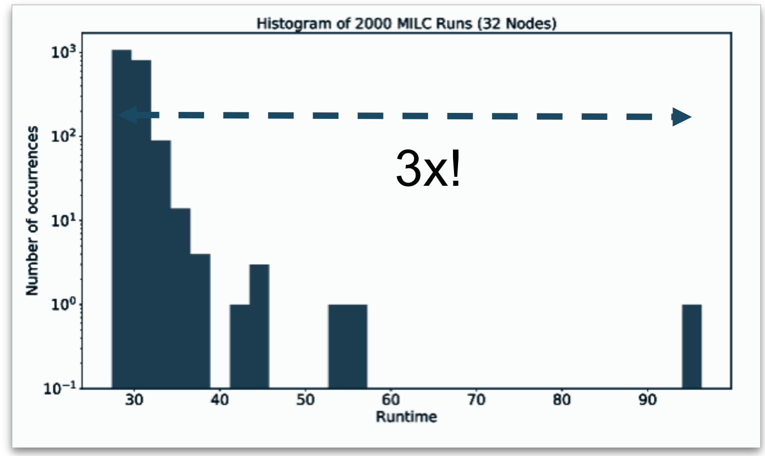


NERSC is leading development of performanceportability.org

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

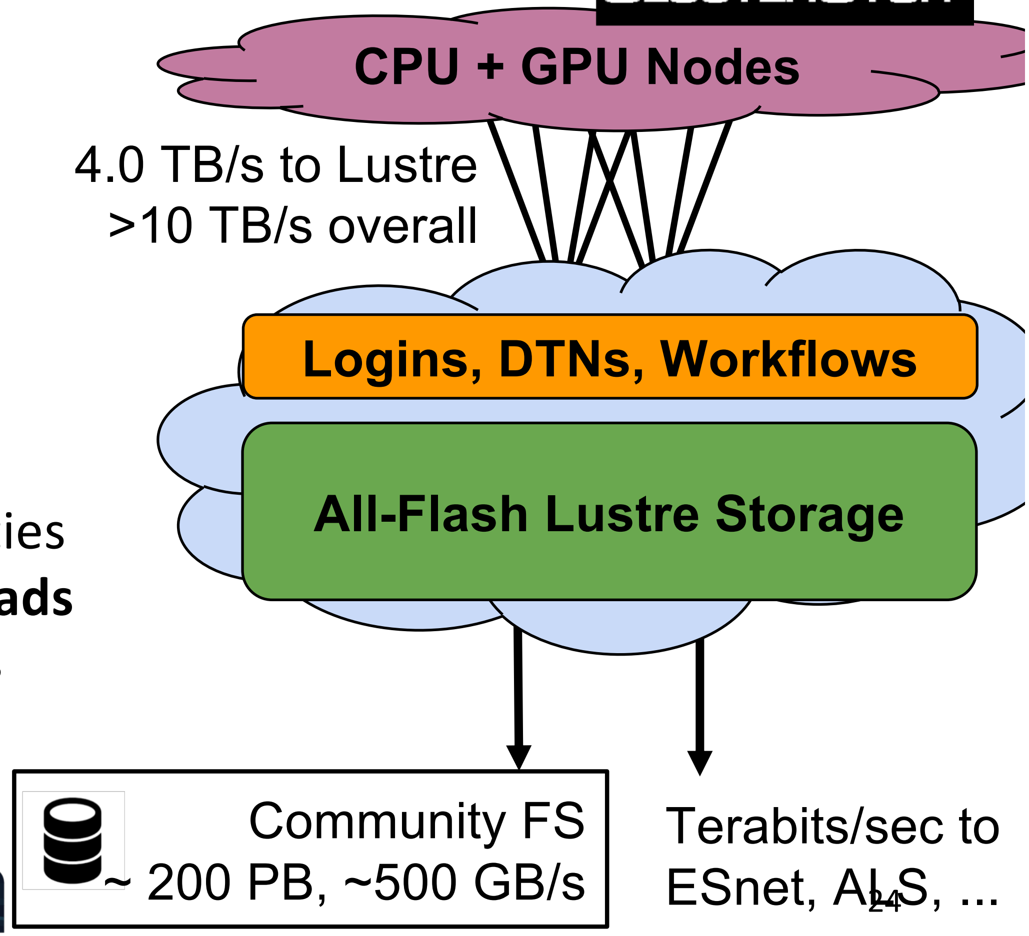
Slingshot Network

- **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 a All-Flash Filesystem

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



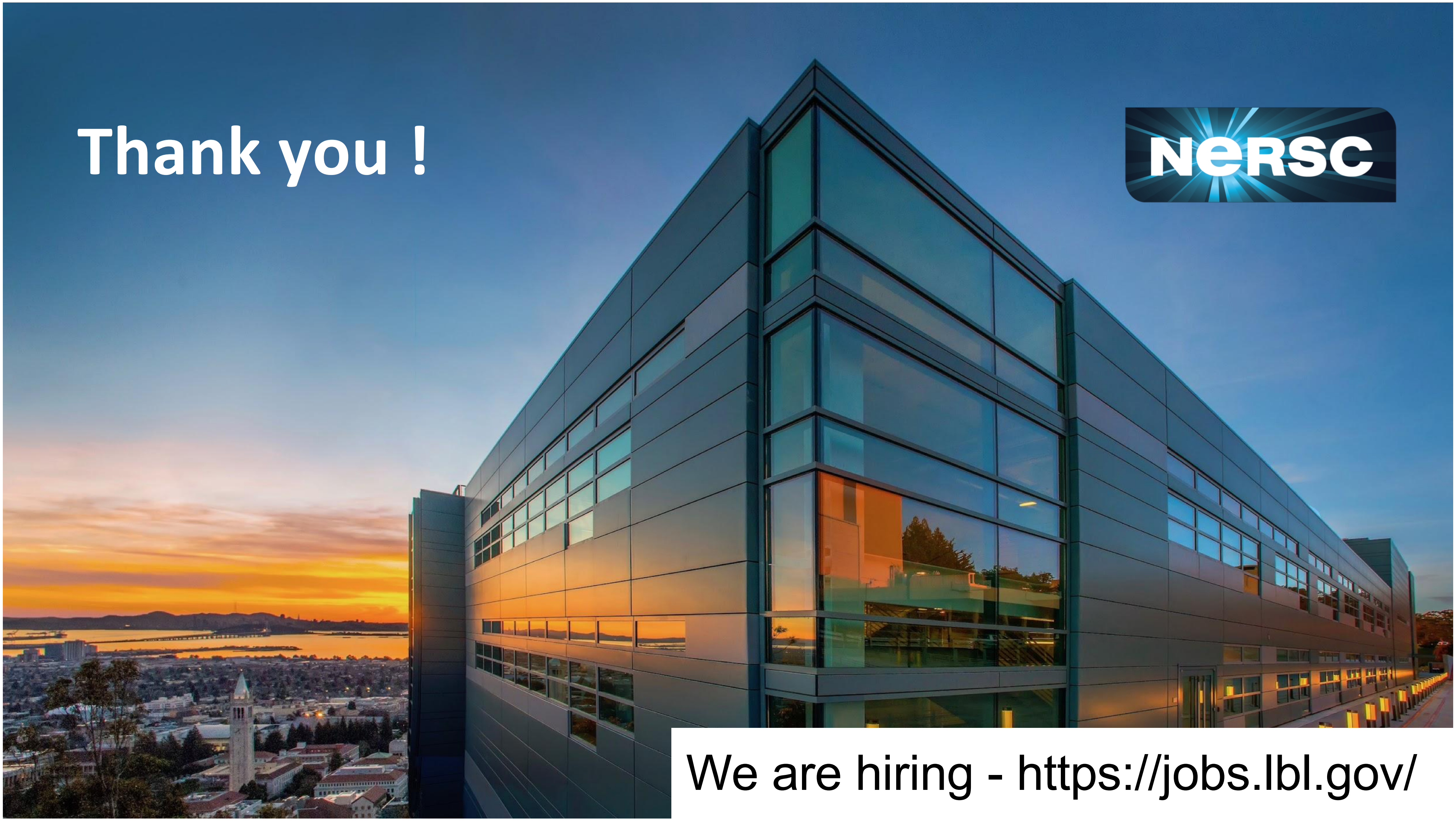
NERSC-9 will be named after Saul Perlmutter

- 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

- **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 - https://jobs.lbl.gov/