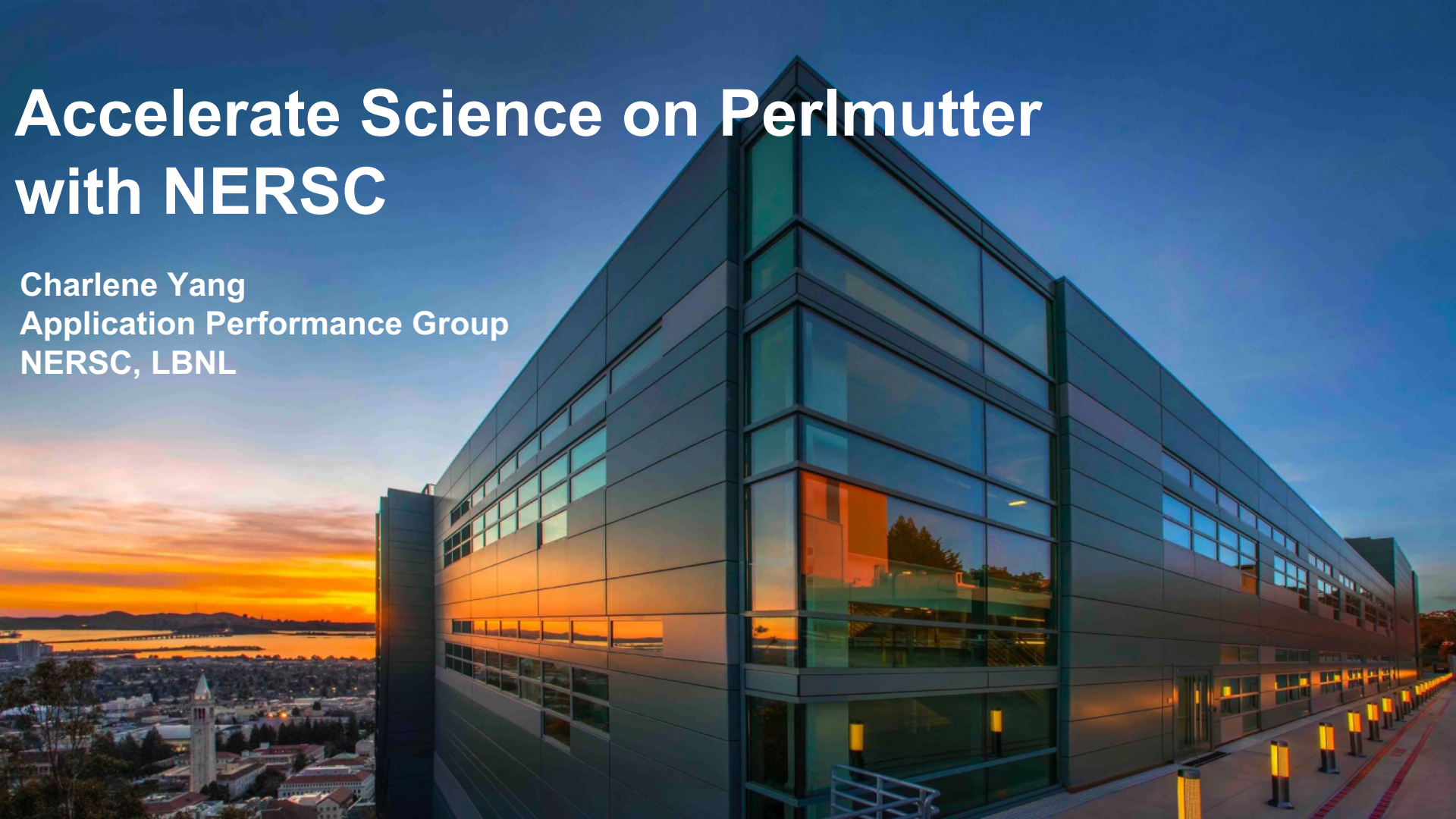
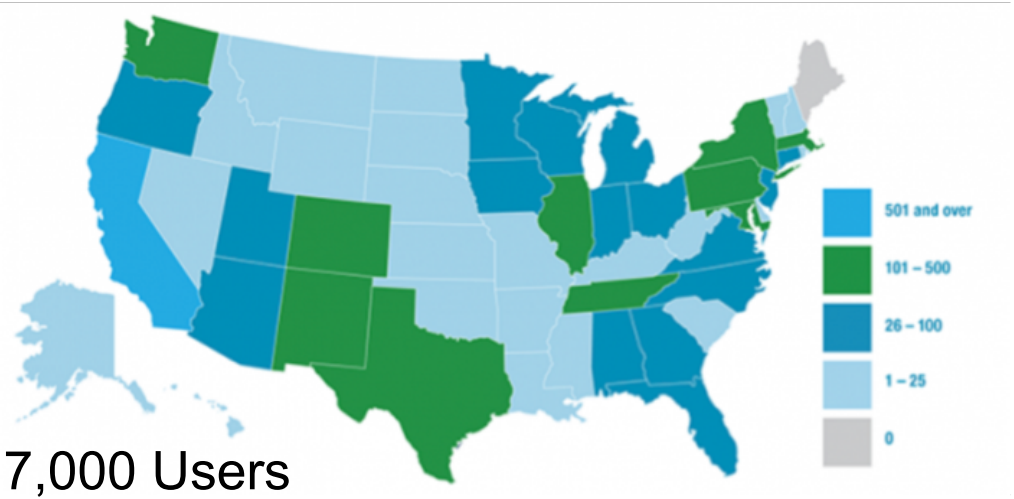
Accelerate Science on Perlmutter with NERSC

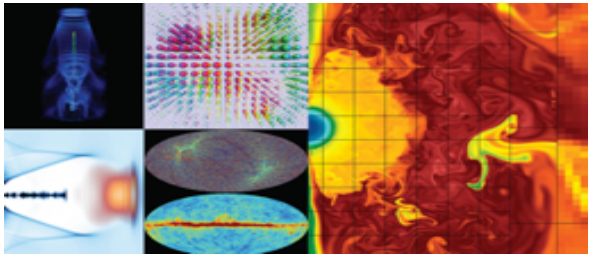
Charlene Yang
Application Performance Group
NERSC, LBNL



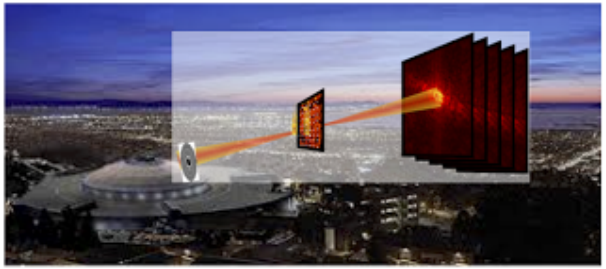
NERSC is the mission High Performance Computing facility for the DOE SC



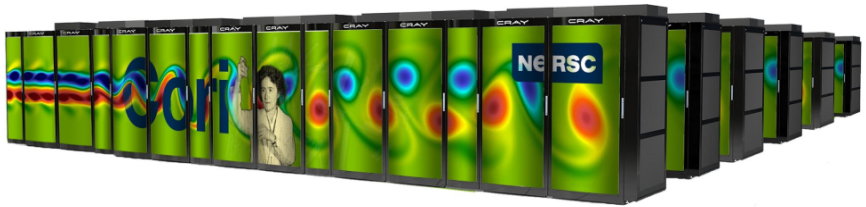
7,000 Users
800 Projects
700 Codes
2000 NERSC citations per year



Simulations at scale



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



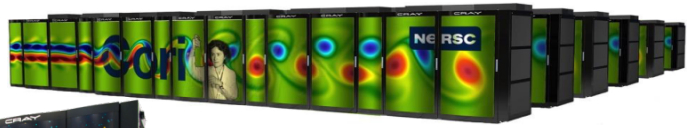
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

NERSC Systems Roadmap

NERSC-7: Edison
2.5 PFs
Multi-core CPU
3MW

2013

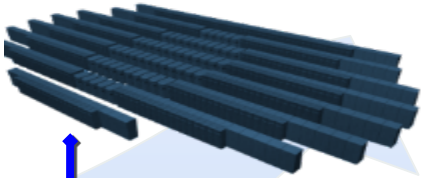


NERSC-8: Cori
30PFs
Manycore CPU
4MW

2016

NERSC-9: Perlmutter
3-4x Cori
CPU and GPU nodes
>5 MW

2020



NERSC-10
ExaSystem
~20MW

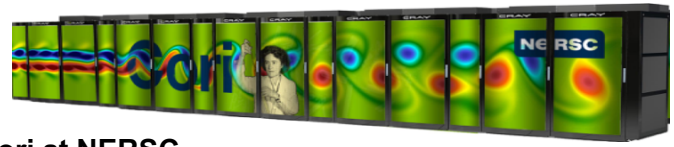
2024

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

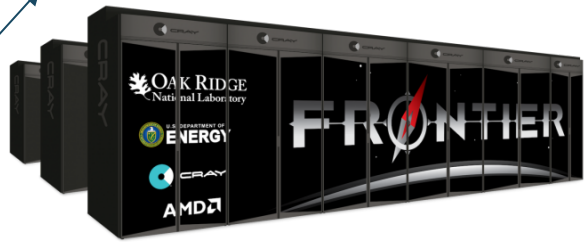
DOE HPC Roadmap

2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023



Cori at NERSC

Summit at OLCF (NVIDIA Volta)



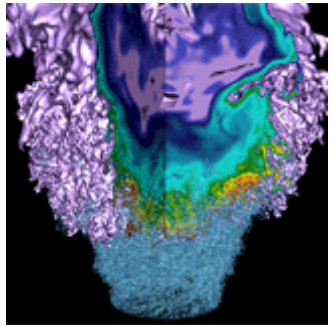
NVIDIA Volta GPUs

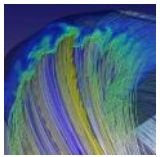
NVIDIA GPUs

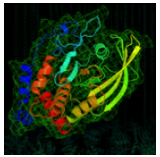
Intel GPUs

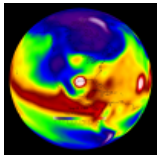
AMD GPUs

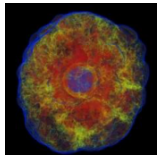
System Overview

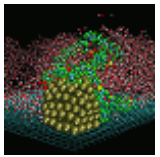












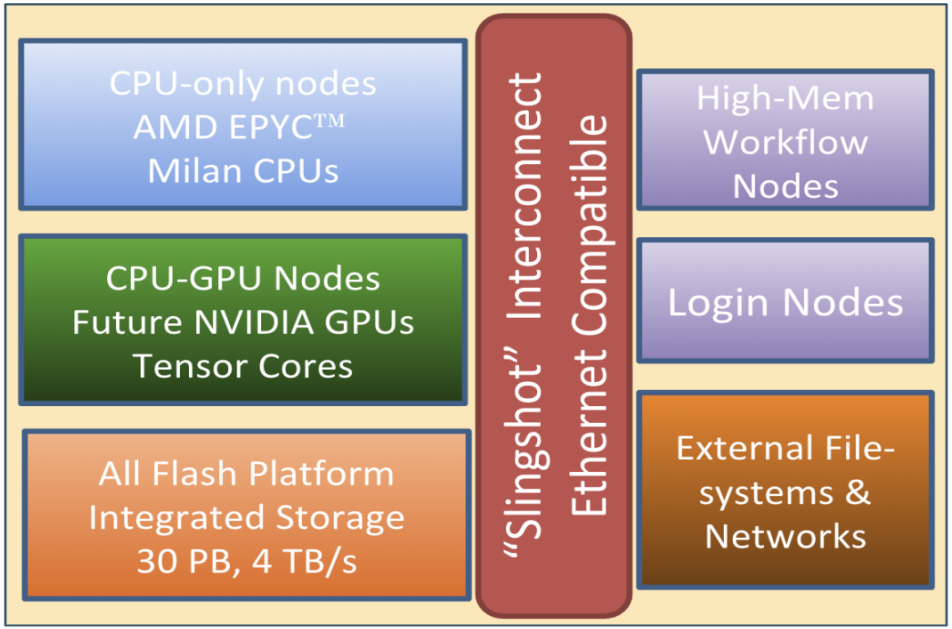
NERSC

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, 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

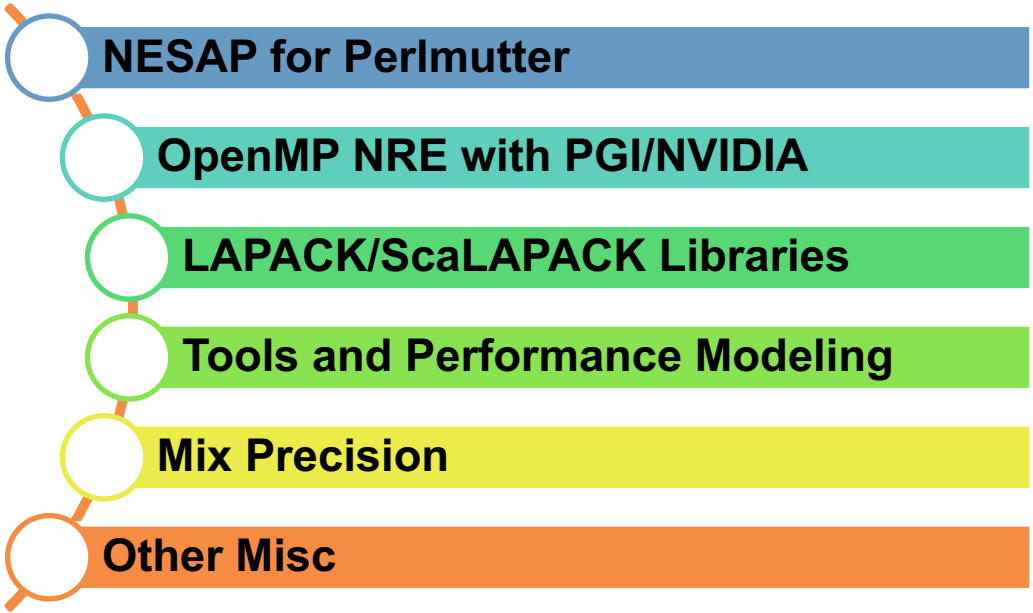
Exascale Requirements Reviews 2015-2018

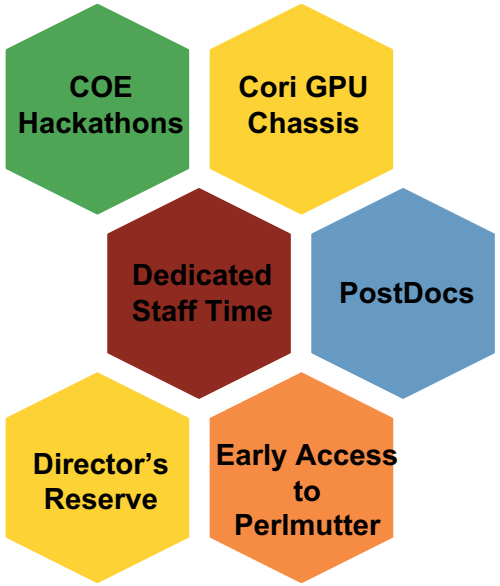
First time users from DOE experimental facilities broadly included

Our Grand Challenge

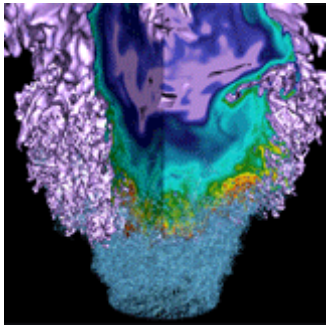
Enable a diverse community of ~7000 users and ~800 codes to run efficiently on advanced architectures such as Cori, **Perlmutter** and beyond

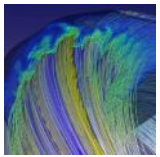
Our Solutions to it

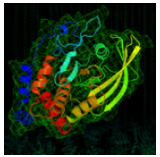


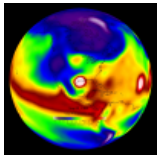


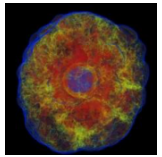
NESAP for Perlmutter

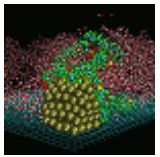












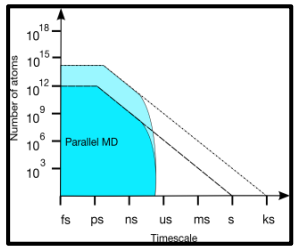
NERSC

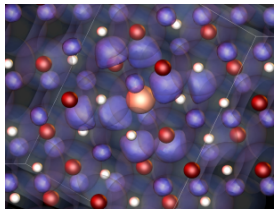
NESAP for Perlmutter

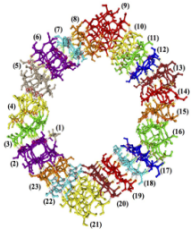
NESAP is NERSC's Application Readiness Program. Initiated with Cori; Continuing with Perlmutter.

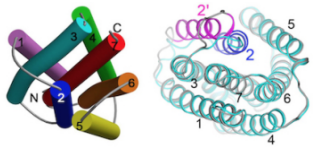
Strategy: Partner with app teams and vendors to optimize participating apps. Share lessons learned with with NERSC community via documentation and training.

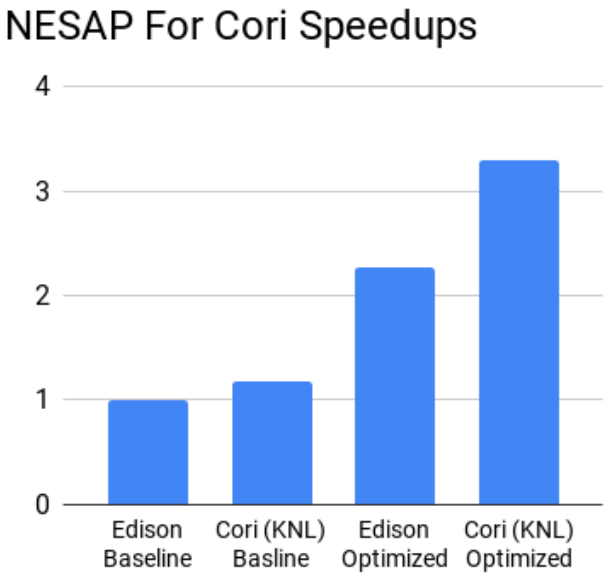
We are really excited about working with you to accelerate science discovery on Perlmutter!



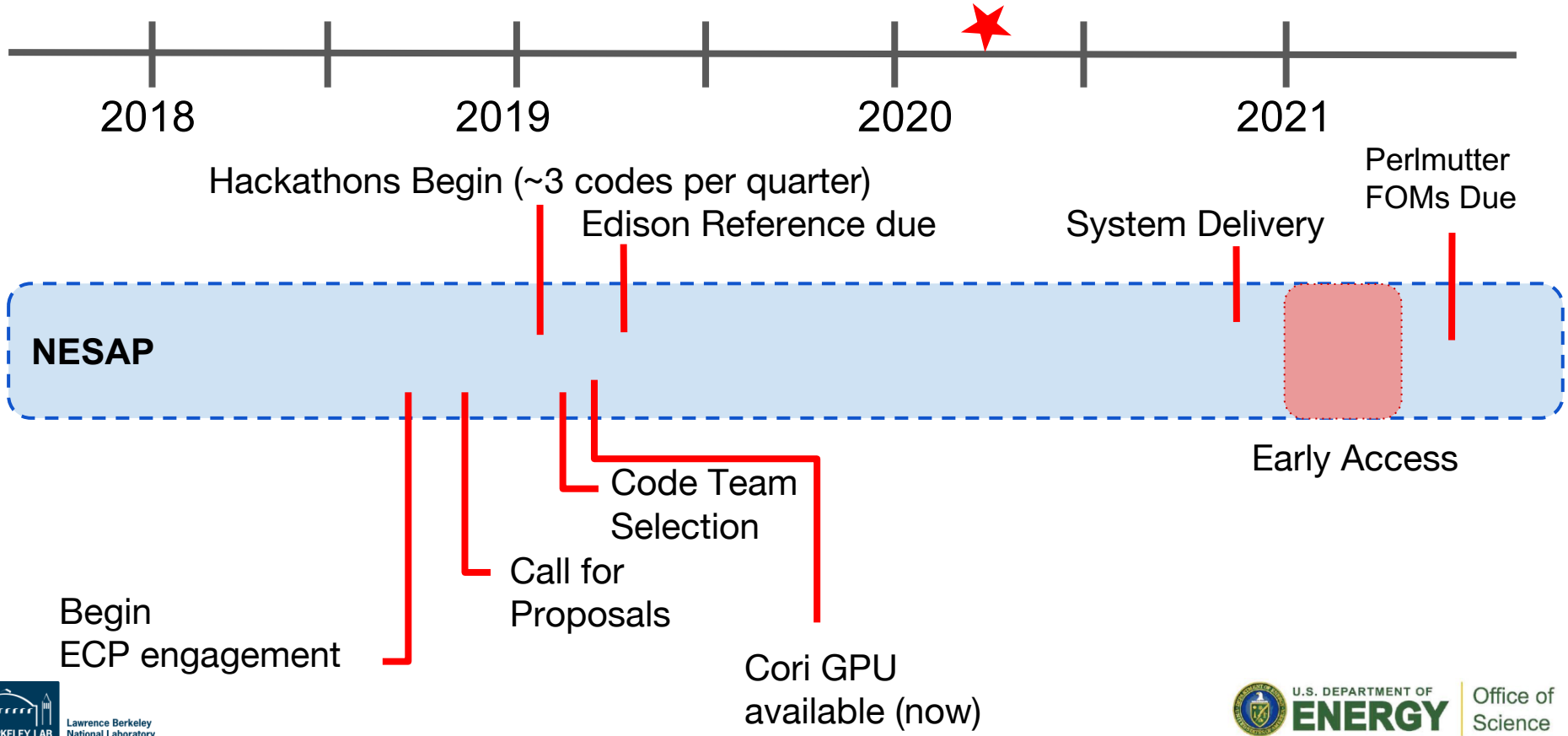








NESAP Timeline



Application Selection

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

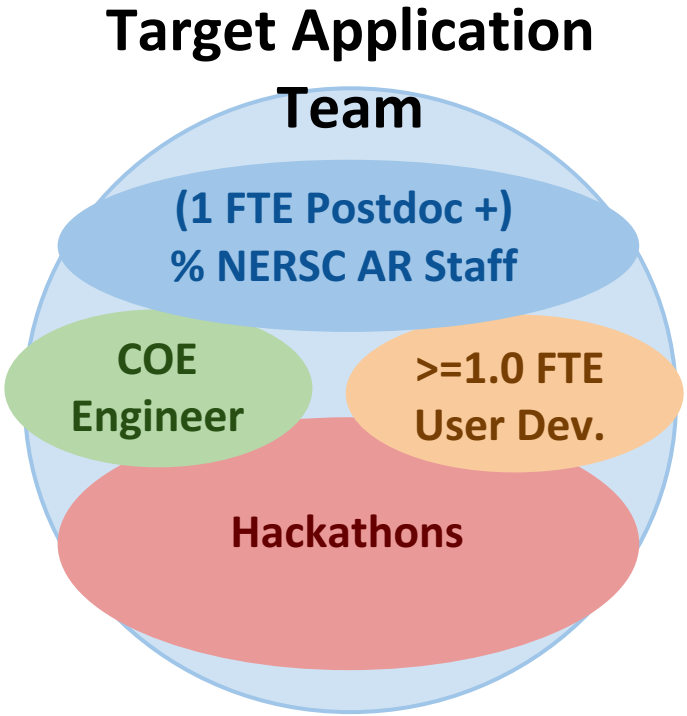
- 6 NESAP for Data apps continued
- 5 ECP Apps jointly selected (Participation funded by ECP)

- Open call for proposals
- Reviewed by a committee of NERSC staff, external reviewers and input from DOE PMs
- Multiple applications from each SC Office and algorithm area

- Beyond this 25 **Tier-1 apps**, additional applications selected for **Tier-2 NESAP**

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

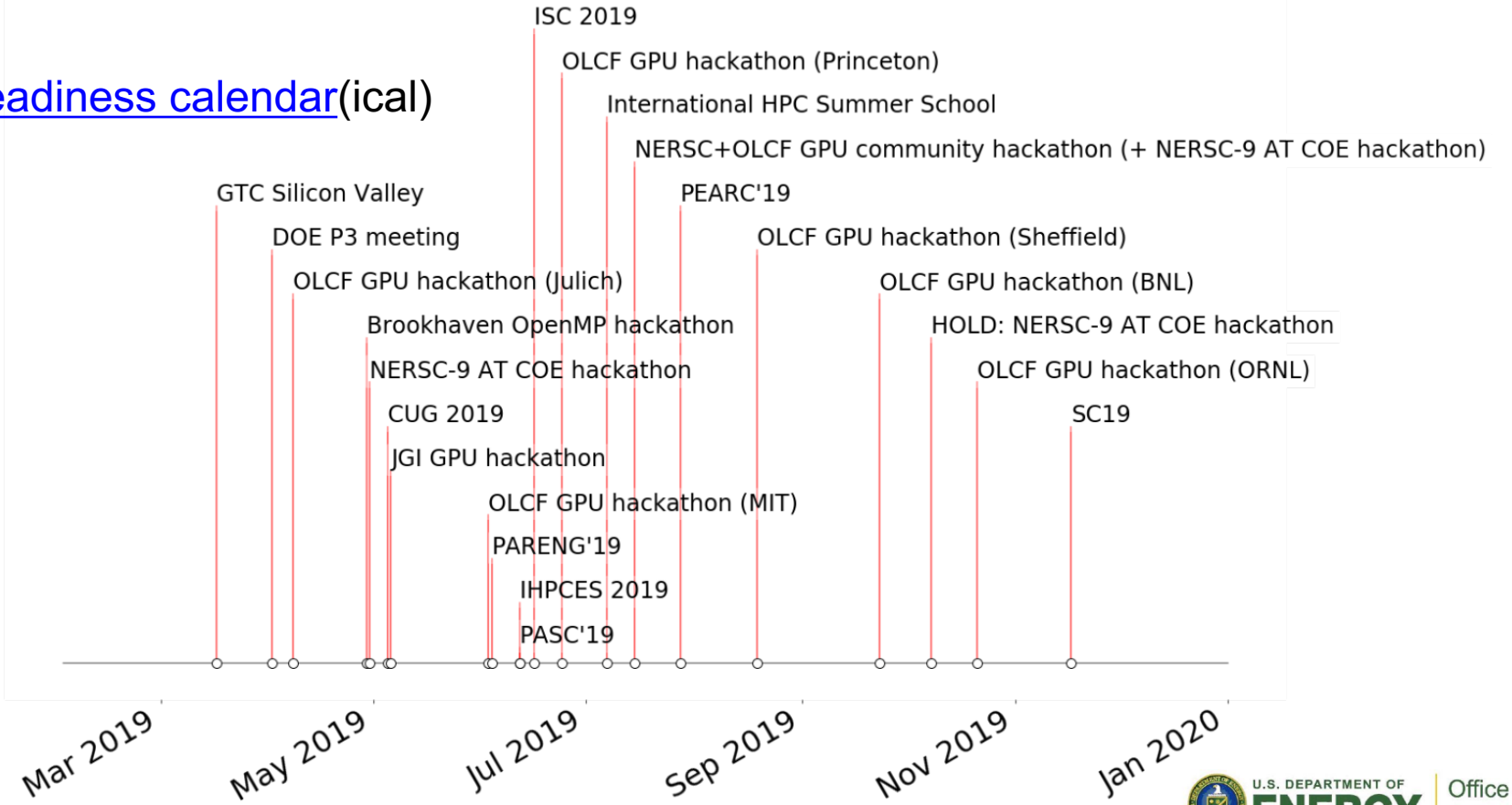


Hack-a-Thons

- Quarterly GPU hackathons from 2019-2021
- ~3 apps per hackathon
- 6-week prep with performance engineers, leading up to 1 week of hackathon
- Deep dives with experts from Cray, NVIDIA, NERSC
- Tutorials throughout the week on different topics
  - OpenMP/OpenACC, Kokkos, CUDA etc.
  - profiler techniques/advanced tips
  - GPU hardware characteristics, best known practices

Other Events

[App Readiness calendar](ical)



NESAP Postdocs

NERSC plans to hire a steady-state of between 10-15 PostDocs to work with NESAP teams towards Perlmutter readiness.

Positions are non-traditional from most academic PostDocs. Project is mission driven (to optimize applications for Perlmutter).

Projects with a mix of Science, Algorithms and Computer Science are often most compelling/successful. **Need to be well connected w/ team**.

PostDocs sit at NERSC and collaborate closely with other NESAP staff but available to regularly travel to team location.

Previous NESAP Postdocs

Mathieu Lobet (WARP)
La Maison de la Simulation (CEA)
(Career)

Brian Friesen (Boxlib/AMReX)
NERSC (Career)

Tareq Malas (EMGEO)
Intel (Career)

Andre Ovsyanikov (Chombo)
Intel (Career)

Taylor Barnes (Quantum ESPRESSO)
MOLSSI (Career)

Zahra Ronaghi (Tomopy)
NVIDIA (Career)

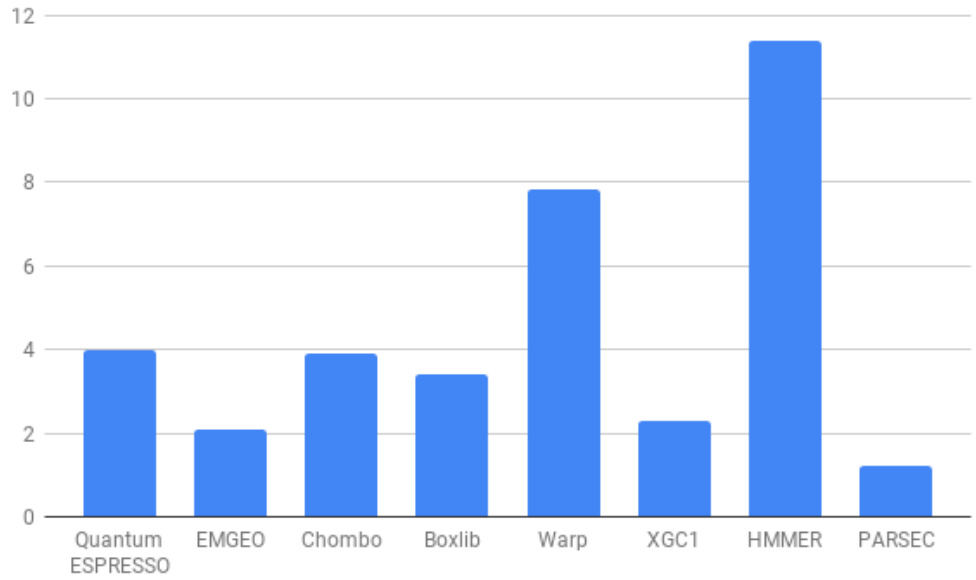
Rahul Gayatri (Perf. Port.)
ECP/NERSC (Term)

Tuomas Koskela (XGC1)
Helsinki (Term)

Bill Arndt (E3SM)
NERSC (Career)

Kevin Gott (PARSEC)
ECP/NERSC (Term)

Postdoc Speedups for Cori



PostDocs made average of 4.5X SpeedUp in NESAP for Cori

Published 20+ Papers Along with NESAP Teams and Staff

We Need Your Help!

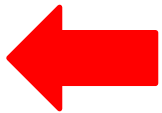
The best way to guarantee your project a PostDoc is to help us recruit one!

Encourage bright, qualified and eligible (must have less than 3 years existing PostDoc experience) candidates to apply (and email Jack Deslippe - jrdeslippe@lbl.gov)

We are interested in advertising in your domain mailing lists.

NESAP PostDoc Position:

http://m.rfer.us/LBLRJs1a1



NERSC Liaisons

NERSC has steadily built up a team of Application Performance experts who are excited to work with you.

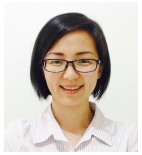
Jack Deslippe
Apps Performance Lead
NESAP LEAD

Brandon Cook
Simulation Area Lead

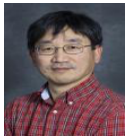
Thorsten Kurth
Learning Area Lead

Rollin Thomas
Data Area Lead

Brian Friesen
Cray/NVIDIA COE Coordinator

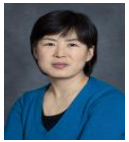


Charlene Yang
Tools/Libraries Lead



Woo-Sun Yang

Doug Doerfler



Zhengji Zhao

Helen He

Stephen Leak

Kevin Gott

Lisa Gerhardt

Jonathan Madsen

Rahul Gayatri

Chris Daley

Wahid Bhimji

Mustafa Mustafa

Steve Farrell

Mario Melara

Lawrence Berkeley National Laboratory

U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC Liaisons

What we can and can't help with:

Can:

- Help Facilitate Between Team and Vendors/NERSC
- Help Profile, Analyze Performance and Guide Optimization
- Get hands on with code, suggest patches for well contained regions
- Help guide PostDocs' progress and provide career advice

Can't (in most cases):

- Become Domain Experts in Your Field
- Redesign an application/algorithm from scratch
- Rewrite/Refactor large sections of your application
- Be the only point-of-contact a NESAP PostDoc has with team

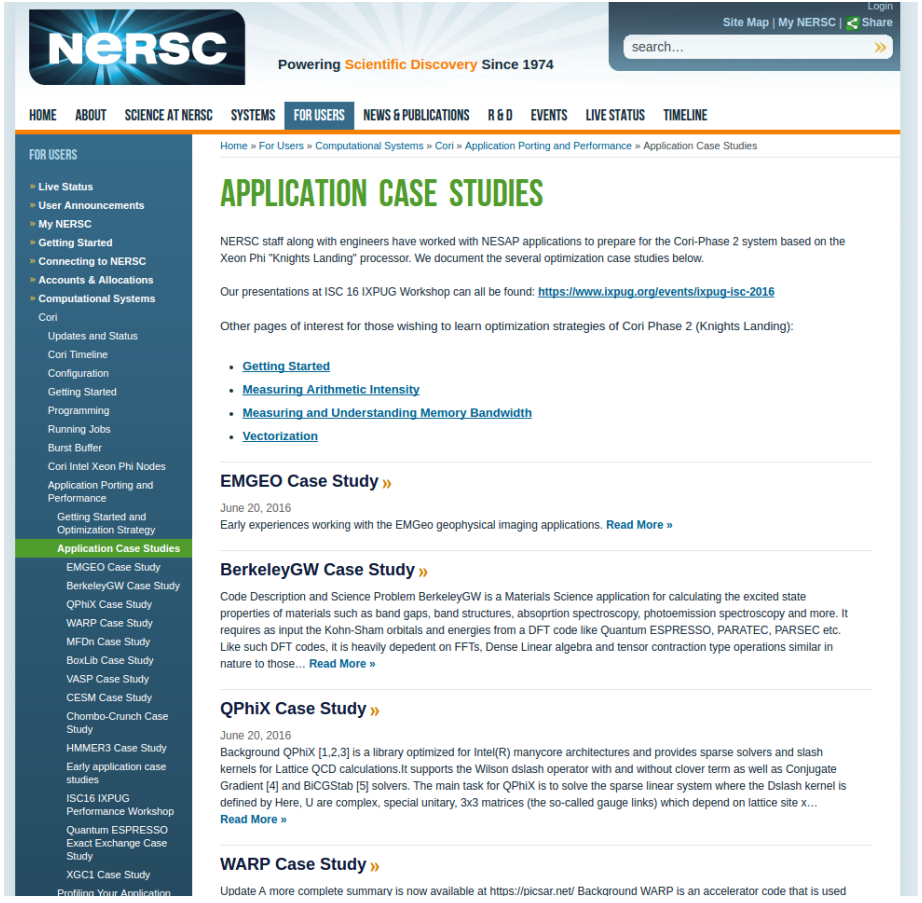
Cori GPU Access

- 18 nodes in total, each node has:
  - 2 sockets of 20-core Intel Xeon Skylake processor
  - 384 GB DDR4 memory
  - 930 GB on-node NVMe storage
  - 8 NVIDIA V100 Volta GPUs with 16 GB HBM2 memory
    - Connected with NVLink interconnect
- CUDA, OpenMP, OpenACC support
- MPI support
- Access for NESAP Teams by request
  - Request form link will be sent to NESAP mailing list

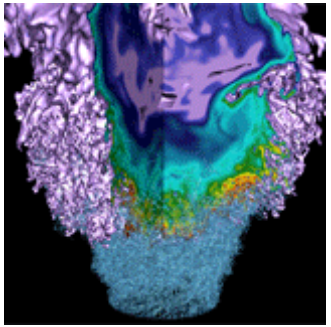
Training, Case Studies and Documentation

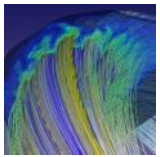
- For those teams **NOT** in NESAP, there will be a robust training program

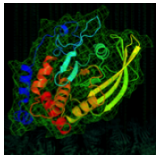
- Lessons learned from deep dives from NESAP teams will be shared through case studies and documentation

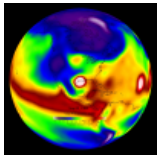


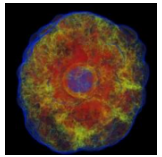
OpenMP NRE

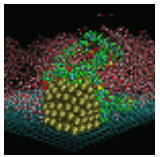












NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

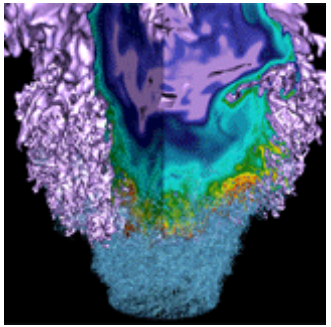
BERKELEY LAB

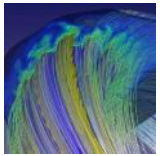
OpenMP NRE

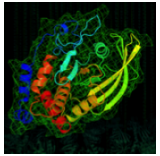
- Add OpenMP GPU-offload support to PGI C, C++, Fortran compilers
  - Performance-focused subset of OpenMP-5.0 for GPUs
  - Compiler will be optimized for NESAP applications

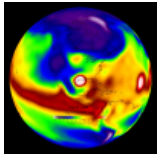
- Early and continual collaboration will help us improve the compiler for you. Please
  - Strongly consider using OpenMP GPU-offload in your NESAP applications
    - Let us help you to use OpenMP GPU-offload
  - Share representative mini-apps and kernels with us
    - Experiment with the GPU-enabled OpenMP compiler stacks on Cori-GPU (LLVM/Clang, Cray, GNU)
  - Contact Chris Daley (csdaley@lbl.gov) and/or your NESAP project POC

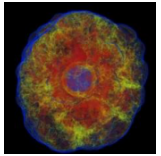
(Sca)LAPACK Libraries

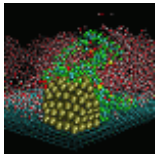












NERSC

Lack of (Sca)LAPACK on GPUs

	Library	Support for NVIDIA GPUs
Single GPU	cuSolver	Incomplete LAPACK (cuSolverDN, cuSolverSP, cuSolverRF)
	MAGMA	Incomplete LAPACK
	Cray LibSci_ACC	Incomplete LAPACK and not promised/planned for Perlmutter
	PETSc	Certain subclasses ported using Thrust and CUSP
	Trilinos	Certain packages implemented using Kokkos
	SLATE	Ongoing ECP work, due to finish in 2021
Multiple GPUs (Distributed)	ELPA	Only support eigensolvers
	???	???

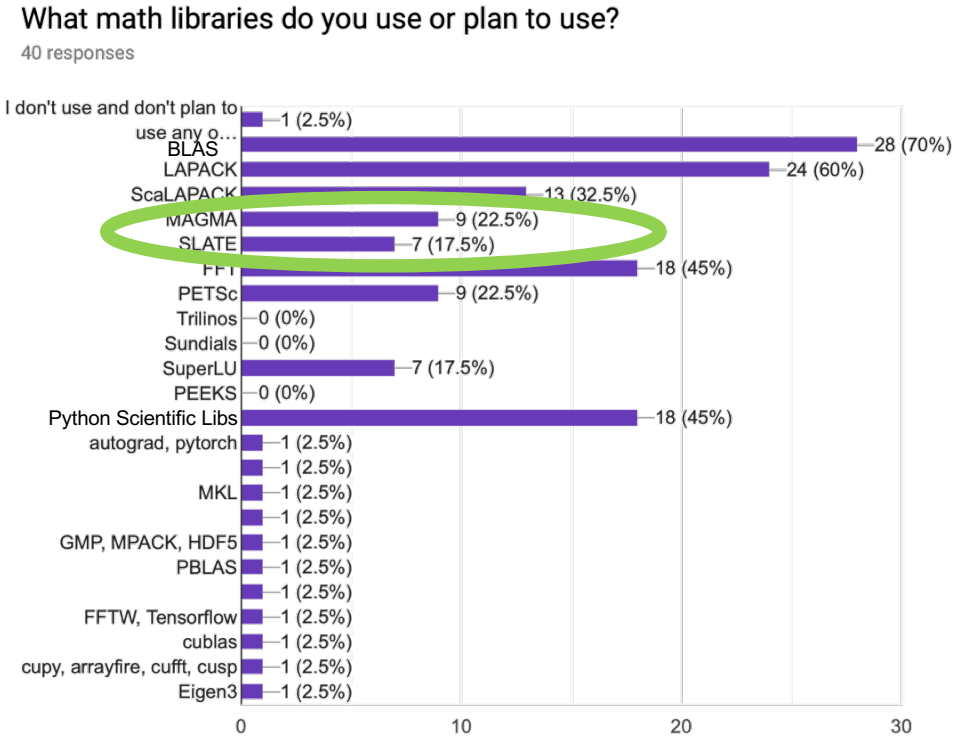
NESAP Survey

- April 10-30; 40 responses

- What libraries do you use?
- What routines in LAPACK?
- What routines in ScaLAPACK?

  (% of runtime, matrix size, sdcz)

- More details at [Results]



Importance to NERSC

ScaLAPACK required by

- VASP
- Quantum Espresso
- NAMD
- CP2K
- BerkeleyGW
- NWChemEx
- WEST
- Qbox
- DFT-FE
- ExaSGD
- PARSEC
- M3DC1
- MFDn
- WDMApp

Even more for LAPACK...

$$\epsilon_{\mathbf{GG'}}(\mathbf{q};0) = \delta_{\mathbf{GG'}} - v(\mathbf{q}+\mathbf{G})\chi_{\mathbf{GG'}}(\mathbf{q};0)$$

$$W_{\mathbf{GG'}}(\mathbf{q};0) = \epsilon^{-1}_{\mathbf{GG'}}(\mathbf{q};0)v(\mathbf{q}+\mathbf{G'})$$

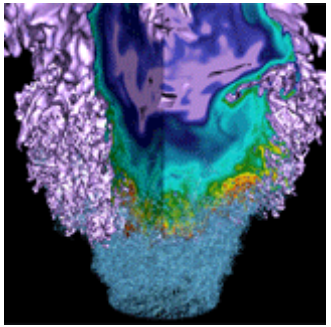
$$\left(E^{\mathrm{QP}}_{c\mathbf{k}} - E^{\mathrm{QP}}_{v\mathbf{k}}\right)A^S_{vc\mathbf{k}} + \sum_{v'c'\mathbf{k}'}\left\langle vc\mathbf{k}\middle|K^{\mathrm{eh}}\middle|v'c'\mathbf{k}'\right\rangle = \Omega^S A^S_{vc\mathbf{k}}$$

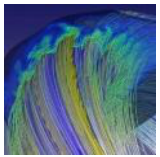
Diagonalization and inversion of large matrices, e.g. 200k x 200k

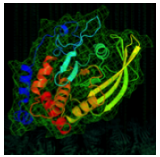
Collaboration with PGI/NVIDIA

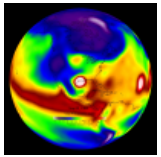
- A drop-in replacement for LAPACK/ScaLAPACK
- Support distributed memory systems with NVIDIA GPUs
- Possibly leverage SLATE and ELPA efforts

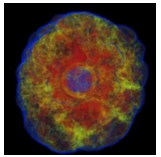
Tools and Performance Models

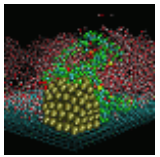












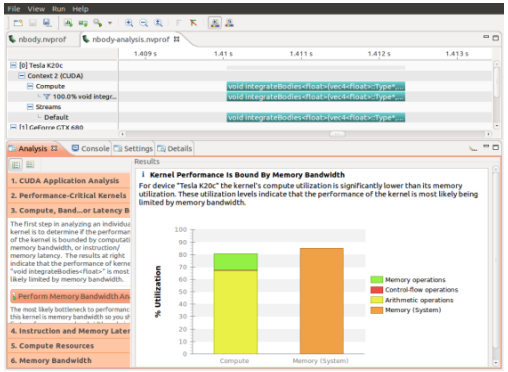
NERSC

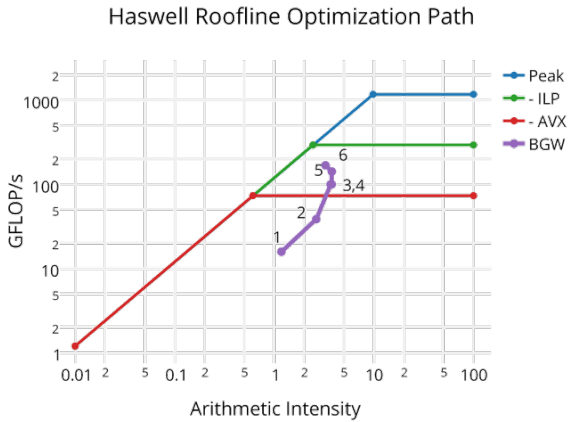
Tools and Roofline

Profiling tools
- provide a rich set of features
- nvprof/nvvp, Nsight Systems, Nsight Compute
- TAU, HPC Toolkit

Roofline Performance Model:
- offers a holistic view of the application
- captures effects of bandwidth/latency, memory coalescing, instruction mix, thread divergence, *etc*

We are actively working with NVIDIA towards GPU Roofline analysis using nvprof/Nsight Compute.





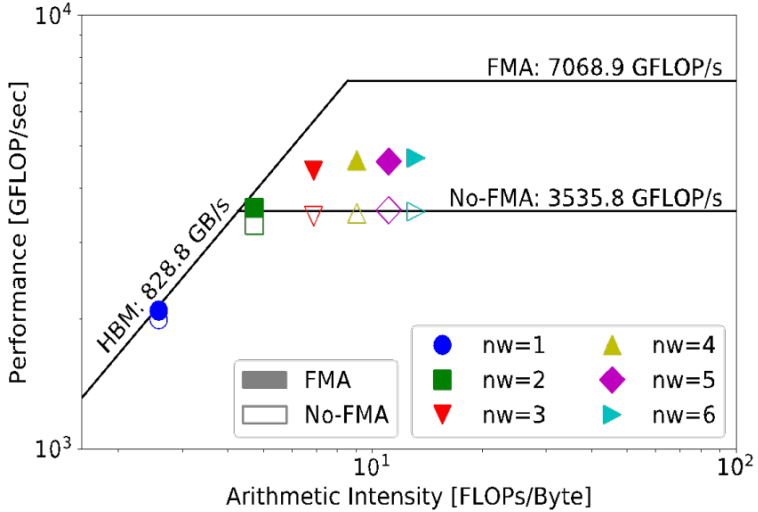
Roofline on GPUs

So far, we've been able to construct a hierarchical Roofline on NVIDIA GPUs

- nvprof metrics for runtime, FLOPs, and bytes
- memory hierarchy: L1/shared, L2, DRAM, *etc*.

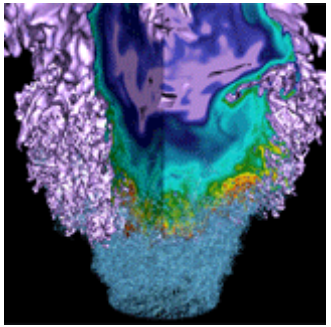
WorkFlow:

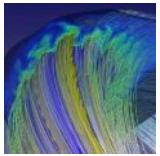
1. Use nvprof to collect application data (FLOPs, bytes, runtime)
2. Calculate Arithmetic Intensity (FLOPs/byte) and application performance (GFLOP/s)
3. Plot Roofline

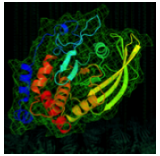


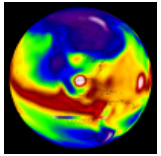
GPP on V100

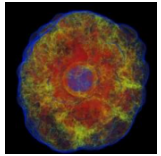
Mixed Precision

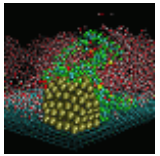












NERSC

Mixed Precision

Benefits of reduced/mixed precision:

- From FP64 to FP32
  - 2x speedup due to bandwidth savings or compute unit availability
  - similar savings in network communication
- More modern architectures support efficient FP16 operations
  - speedup of about 15x possible compared to FP64 for certain operations
- Similar speedups are possible if most operations are done in lower precision

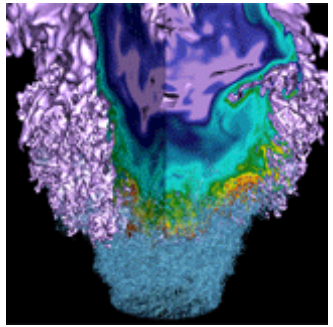
NESAP collaboration with CRD (Costin Iancu) and NVIDIA (Chris Newburn):

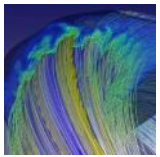
- Investigate the applicability of mixed precision arithmetic
- Extract general guidelines and rules of when it works when it doesn't
- Apply findings to some NESAP applications to improve performance

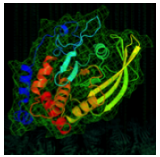
How can I get involved?

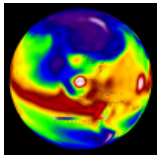
- Follow opportunities to follow on the NESAP mailing list

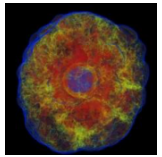
Other Work

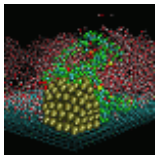










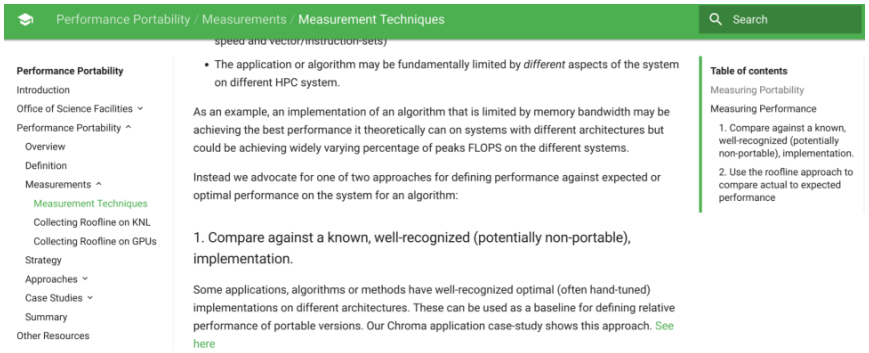


NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Performance Portability



OpenACC
Directives for Accelerators

NERSC now a member.

NERSC leading development of performanceportability.org

NERSC hosted 2016 C++ Summit and ISO C++ meeting on HPC.

NERSC leading 2019 DOE COE Perf. Port. Meeting

NERSC

Thank You