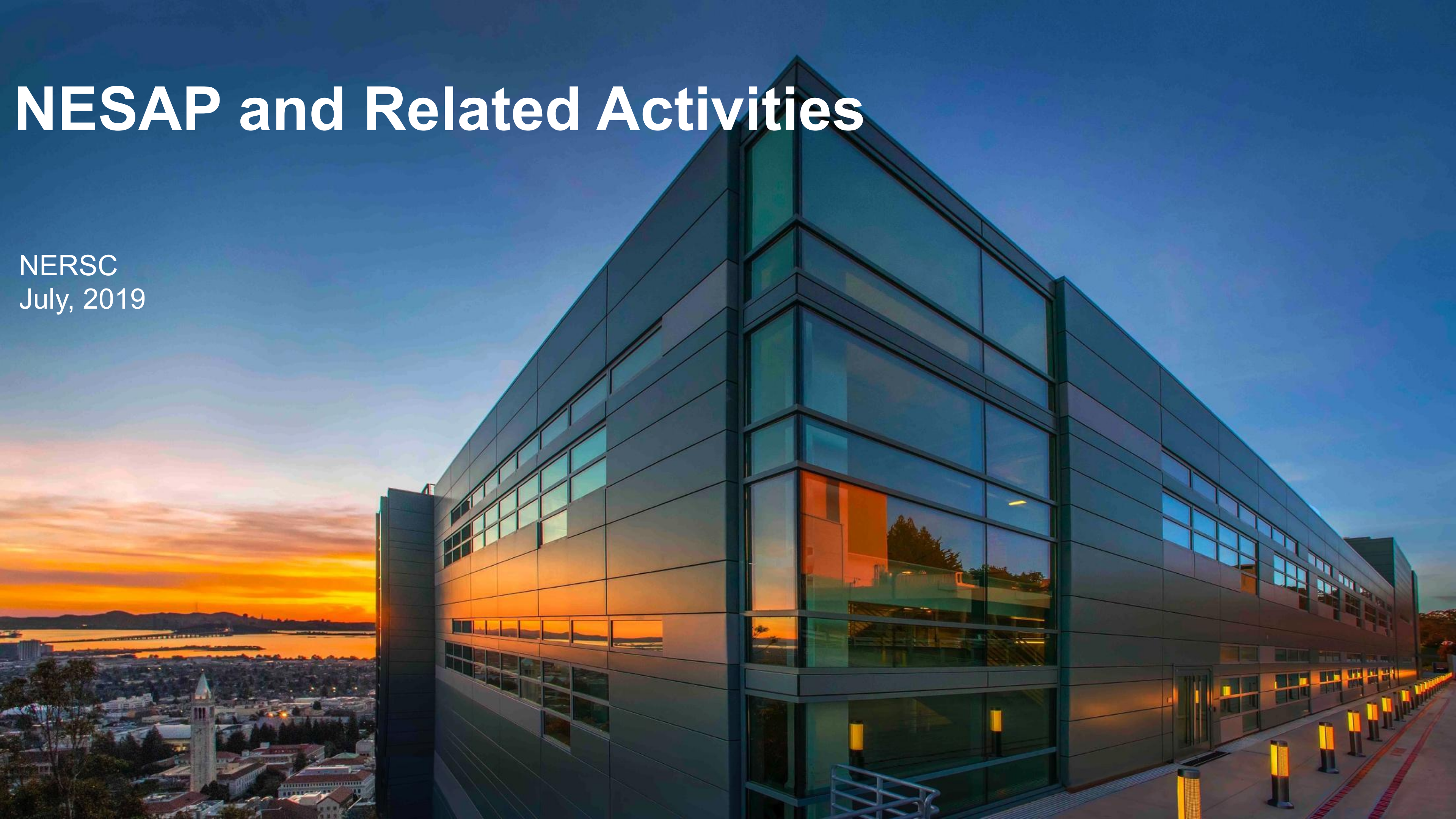
NESAP and Related Activities

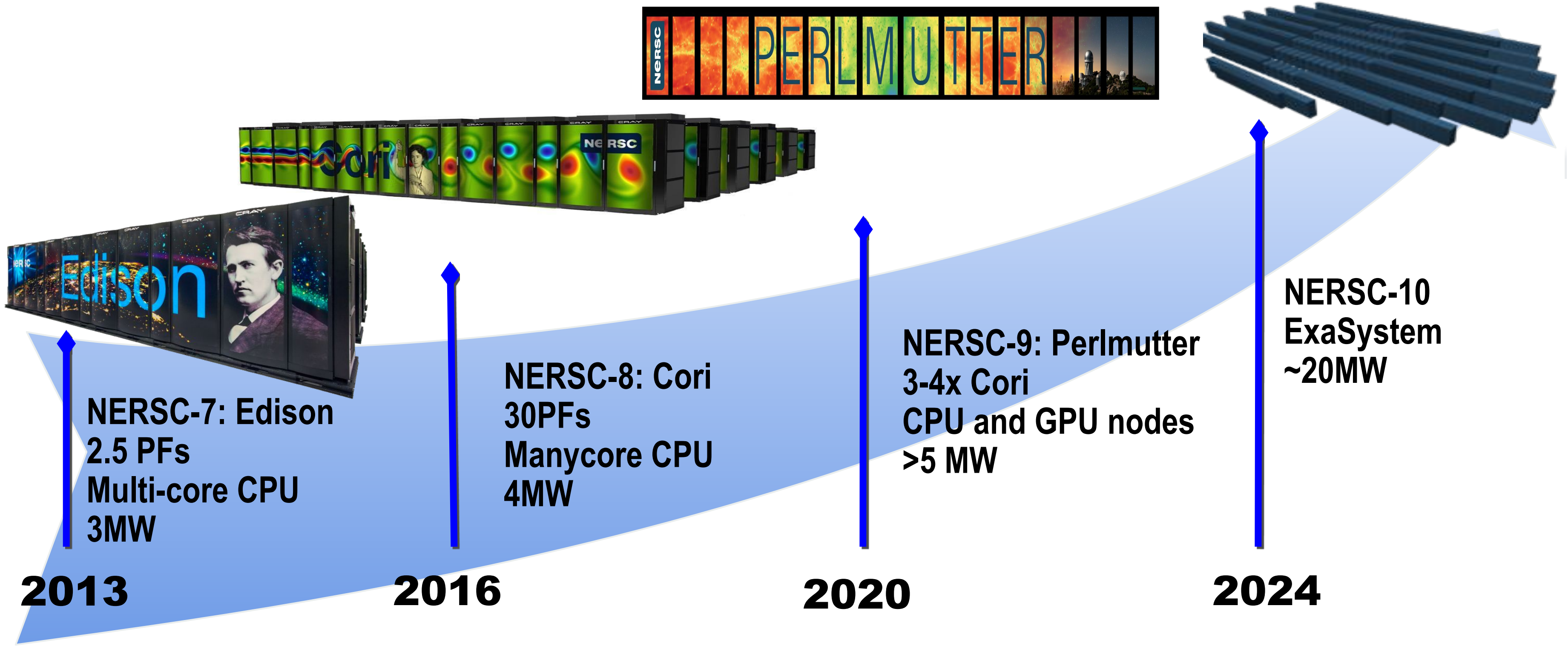
NERSC
July, 2019



Our Common Challenge

Enable a diverse community of scientific users and codes to run efficiently on advanced architectures like Cori, Perlmutter and beyond

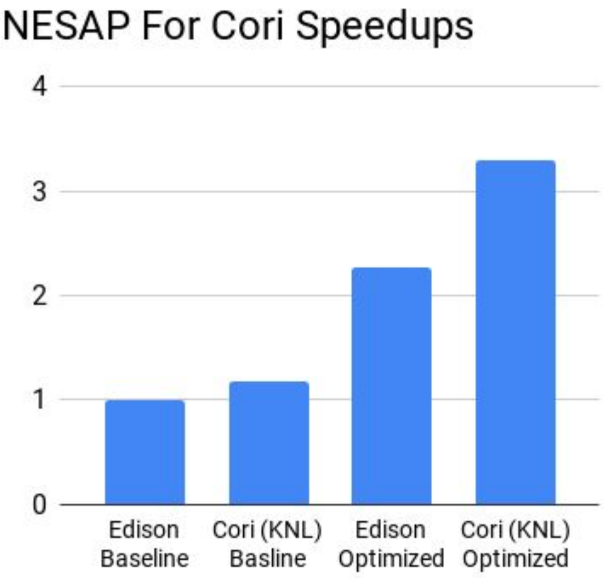
NERSC Systems Roadmap

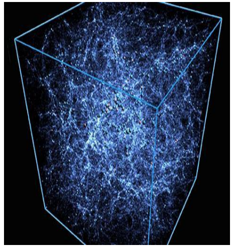


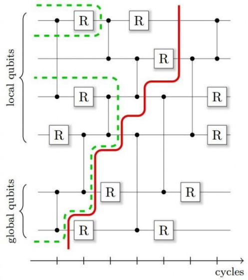
NESAP for Perlmutter

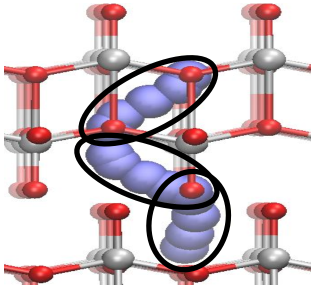
NESAP is NERSC’s Application Readiness Program. Initiated with Cori; Continuing with Perlmutter.

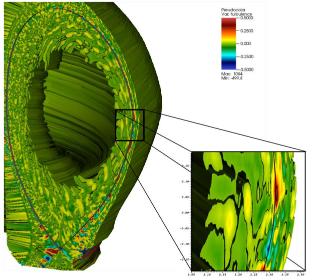
Strategy: Partner with app teams and vendors to optimize participating apps. Share lessons learned with with NERSC community via documentation and training.







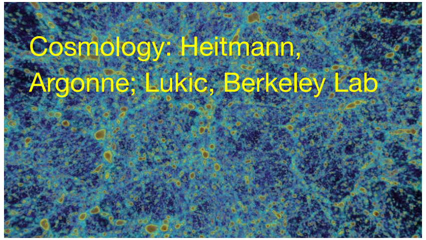




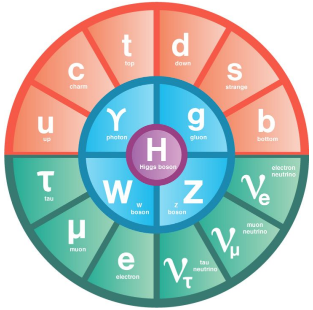
NESAP For Cori Key Takeaways

- NESAP can have a significant impact on application performance
- Data, learning apps can benefit from NESAP too
- Strong commitment from teams is required for success
- Hack-a-thons are keystone events
- Early access to relevant hardware is critical
- Engaging with vendors on tools and strategy is important

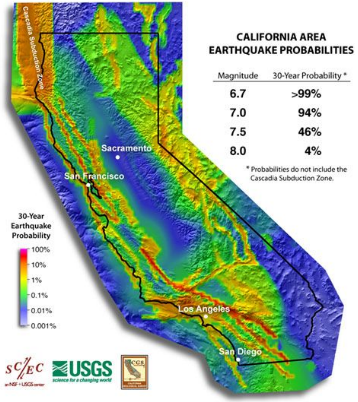
High Impact Science at Scale Projects

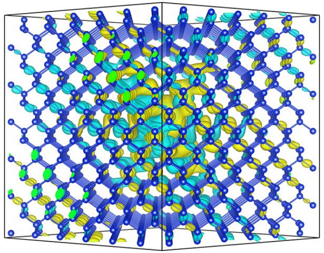


Strangeness and Electric Charge Fluctuations in Strongly Interacting Matter, Karsch, Brookhaven

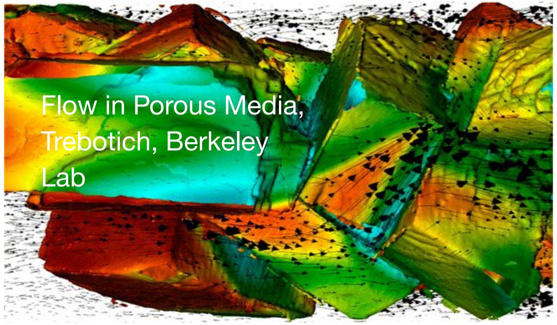


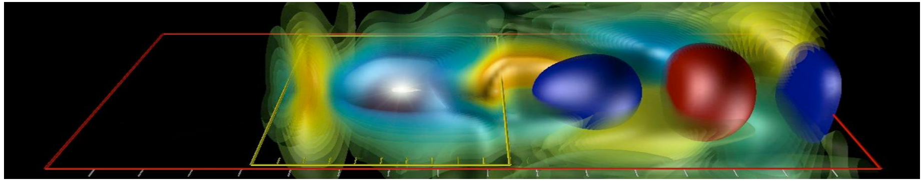
M8 Earthquake on the San Andreas Fault, Goulet, USC Earthquake Center





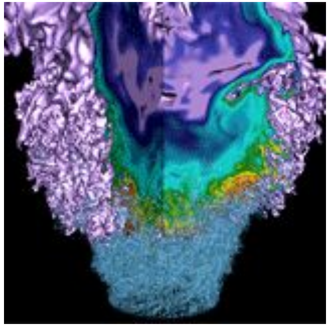
Properties of Complex Materials, Louie, UC Berkeley

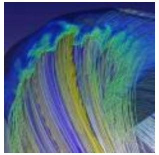


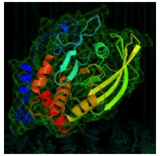


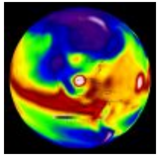
Asymmetric Effects in Plasma Accelerators, Vay, Berkeley Lab

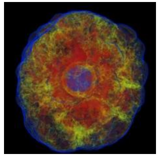
Perlmutter Overview

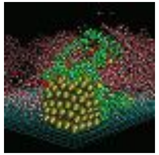






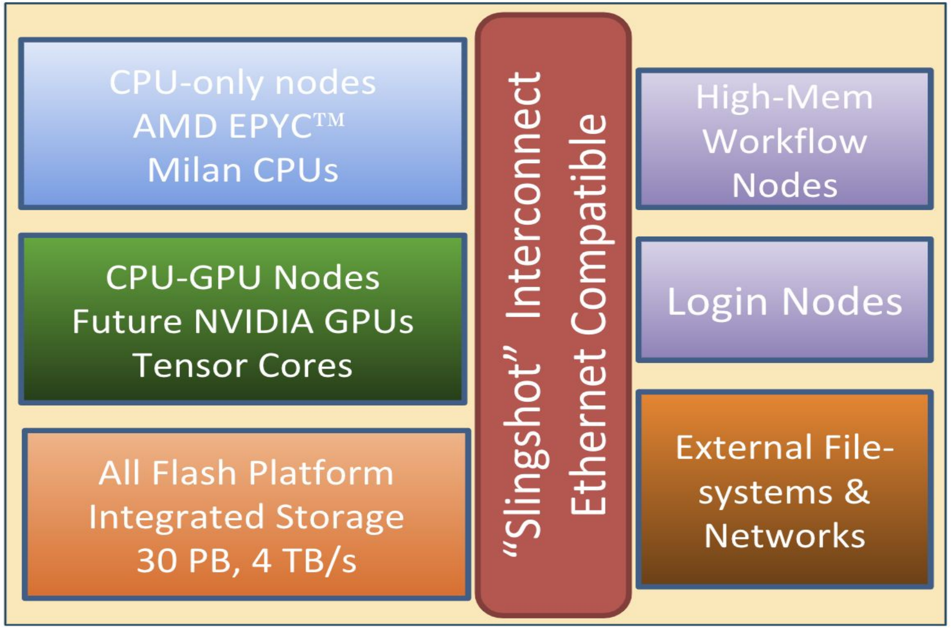






Perlmutter: A System Optimized for Science

- 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



AMD CPU nodes

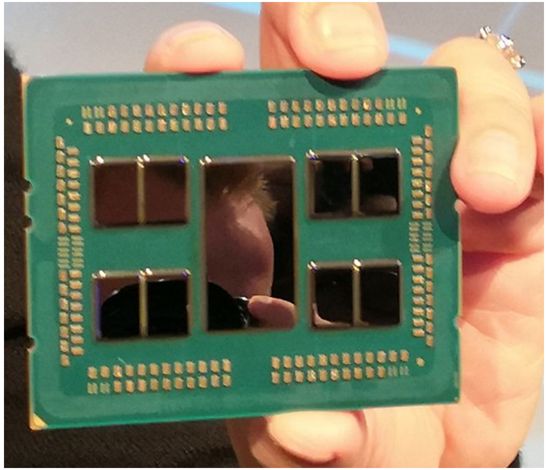
"Rome" specs

- ~64 cores
- AVX2 SIMD (256 bit)
  (Perlmutter will have Milan)

1 Slingshot connection

- 1x25 GB/s

~1 Cori **Optimizations for KNL expected to pay off on Milan**



GPU nodes

4x NVIDIA "Volta-next" GPU

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

Volta specs

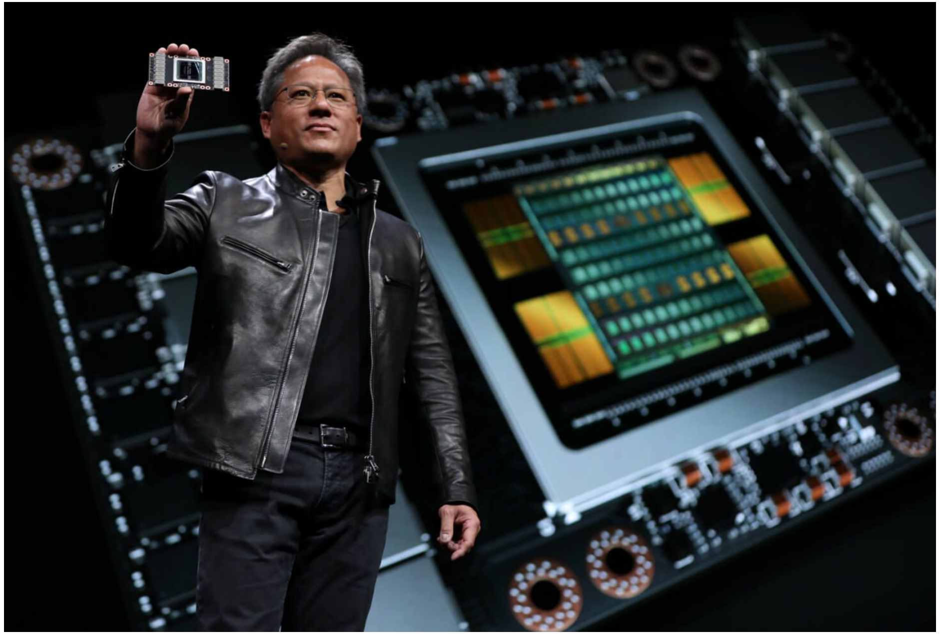
1x AMD Milan CPU

4 Slingshot connections

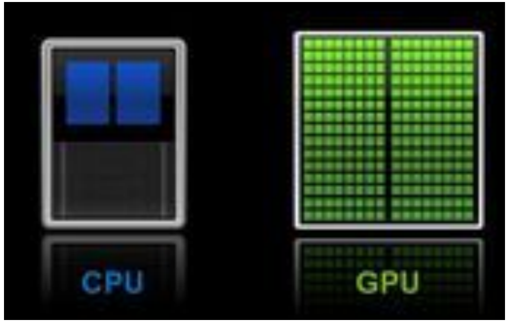
- 4x25 GB/s

GPU direct, Unified Virtual Memory (UVM)

2-3x Cori



CPU vs GPU



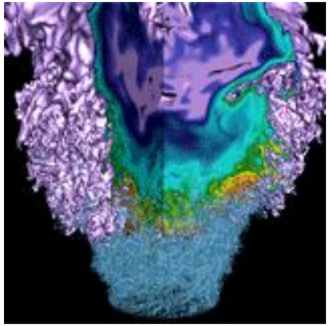
CPU (KNL)

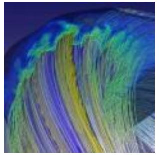
- 68 cores
- 4 threads each
- 512-bit vectors
- pipelined instructions
- double precision
  - ~2000 way parallelism (68*4*8)

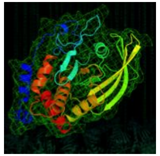
GPU (V100)

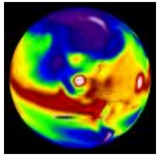
- 80 SM
- 64 warps per SM
- 32 threads per warp
- double precision
  - ~150,000+ way parallelism (80*64*32)

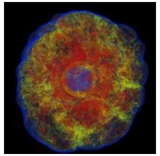
NESAP Teams

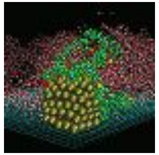












Two tiers of support for projects

With so many strong proposals we decided to support a large number of projects with 2 different levels of support

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
Number of applications in Tier	29 (18 new, 5 ECP and 6 NESAP for Data)	28

NESAP: Data

PI Name	Institution	Application name	Office
Maria Elena Monzani	SLAC	NextGen Software Libraries for LZ	HEP
Kjiersten Fagnan	JGI	JGI-NERSC-KBase FICUS Project	BER
Kathy Yelick	LBNL	Exabiome (ECP)	BER
Stephen Bailey	LBNL	DESI	HEP
Julian Borrill	LBNL	TOAST	HEP
Doga Gursoy	ANL	Tomopy	BES
Amedeo Perazzo	SLAC	ExaFEL (ECP)	BES
Paolo Calafiura	LBNL	Atlas	HEP
Dirk Hufnagel	LBNL	CMS	HEP

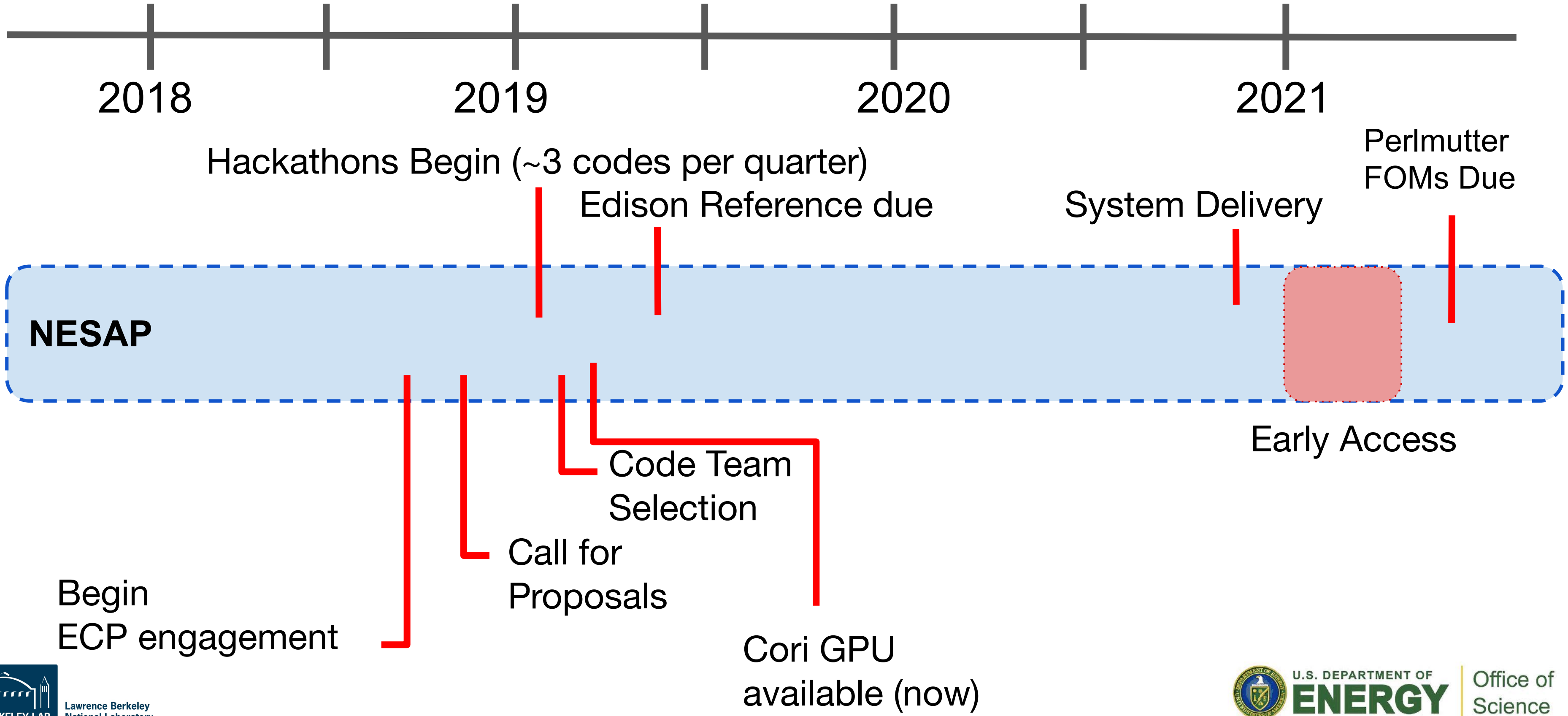
NESAP:Learning

PI Name	Institution	Application name	Office
Christine Sweeney	LANL	ExaLearn Light Source Application	BES
Marc Day	LBNL	FlowGAN	ASCR
Shinjae Yoo	BNL	Extreme Scale Spatio-Temporal Learning (LSTNet)	ASCR
Benjamin Nachman and Jean-Roch Vlimant	Caltech	Accelerating High Energy Physics Simulation with Machine Learning	HEP
Zachary Ulissi	CMU	Deep Learning Thermochemistry for Catalyst Composition Discovery/Optimization	BES

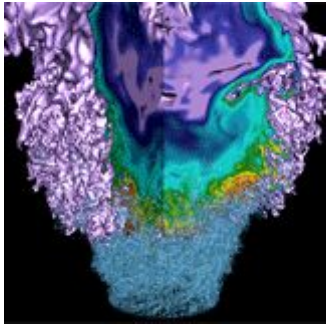
NESAP:Simulation

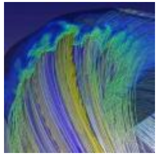
PI Name	Institution	Application name	Office
Josh Meyers	LLNL	ImSim	HEP
Carleton DeTar, Balint Joo	Utah; JLAB	USQCD (MILC, DWF, chroma, etc)	HEP;NP
Noel Keen, Mark Taylor	SNL; LBNL	E3SM	BER
David Green	ORNL	ASGarD (Adaptive Sparse Grid Discretization)	FES;ASCR
Mauro Del Ben	LBNL	BerkeleyGW	BES
Pieter Maris	Iowa State	Many-Fermions Dynamics for nuclear physics (MFDn)	NP
Hubertus van Dam	BNL	NWChemEx	BER;BES
David Trebotich	LBNL	Chombo-Crunch	BES
Marco Govoni	ANL	WEST	BES
Annabella Selloni, Robert DiStasio and Roberto Car	Princeton; Cornell	Quantum ESPRESSO	BES
Emad Tajkhorshid	UIUC	NAMD	BER;BES
CS Chang	PPL	WDMAPP	FES
Danny Perez	LANL	LAMMPS	BES;BER;FES
Ann Almgren / Jean-Luc Vay	LBNL	AMReX / WarpX	HEP

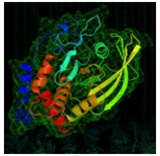
NESAP Timeline

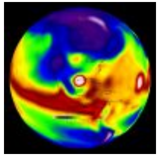


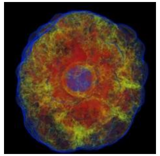
NESAP Staff

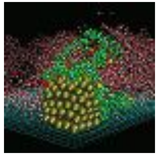












NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC Liaisons

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

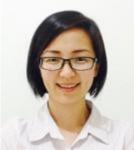
Jack Deslippe
Apps Performance Lead
NESAP LEAD

Brandon Cook
Simulation Area Lead

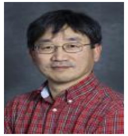
Thorsten Kurth
Learning Area Lead

Rollin Thomas
Data Area Lead

Brian Friesen
Cray/NVIDIA COE Coordinator

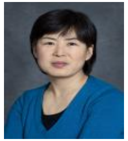


Charlene Yang
Tools/Libraries Lead



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

Postdocs

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

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

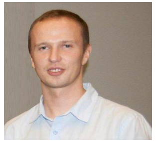
PostDocs sit at NERSC and collaborate closely with other NESAP staff but available to regularly travel to team location.

Where Have NESAP Postdocs Gone?

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)

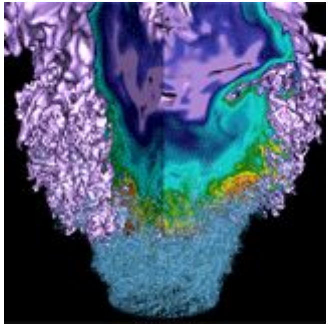
Rahul Gayatri (Perf. Port.)
ECP/NERSC (Term)

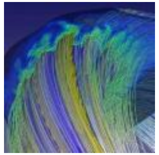
Tuomas Koskela (XGC1)
Helsinki (Term)

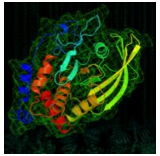
Bill Arndt (E3SM)
NERSC (Career)

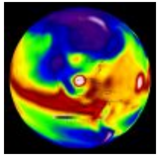
Kevin Gott (PARSEC)
ECP/NERSC (Term)

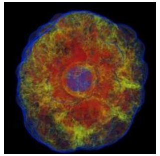
Hack-a-Thons

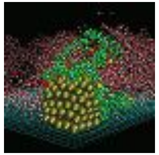












Hack-a-Thons

- Quarterly GPU hackathons from 2019-2021
- ~3 apps per hackathon
- 6-week prep with performance engineers, leading up to 1 week of hackathon

- Tutorials throughout the week on different topics
  - OpenMP/OpenACC, Kokkos, CUDA etc.
  - profiler techniques/advanced tips
  - GPU hardware characteristics, best known practices

CRAY

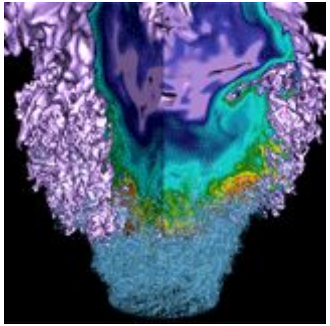
NVIDIA

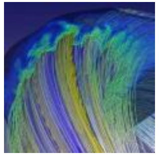
Recent Events

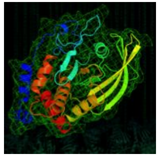
GPU For Science Days: July 2-3

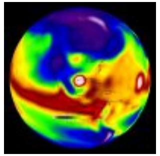
GPU Community Hackathon July 15-19

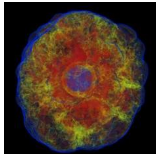
Early Examples

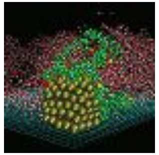












Example Tomopy

The figure of merit is the wall-clock time of reconstruction per each 2D slice and for a series of 2D slices (i.e. 3D dataset). The algorithm of choice was the SIRT algorithm with 100 iterations. Each 2D slice was 2048 x 2048 pixels and the number of projection angles was 1501.

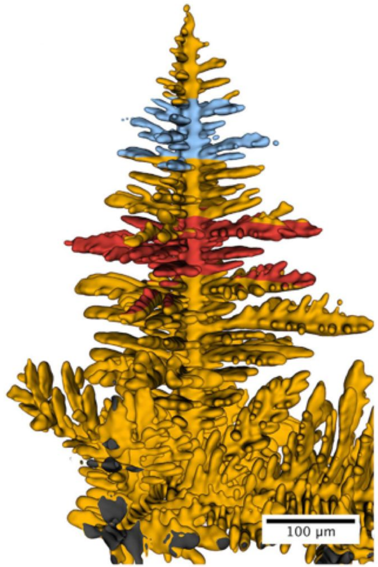
FOM= 1 / (< Fraction of System Used > * < WallTime Per Slice >)

Baseline 24 slice reconstruction time (Edison)

wall	28252.003
user	659962.850
system	5.680
cpu	659968.530

GPU 24 slice reconstruction time

wall	278.872
user	13475.050
system	7244.300
cpu	20719.350

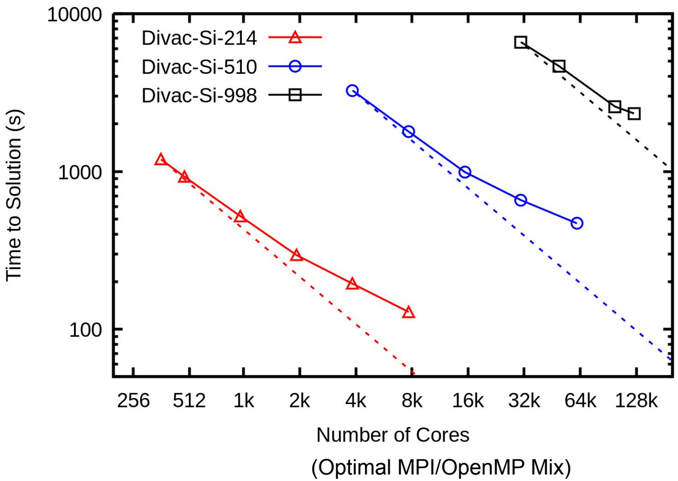


Rough Perlmutter System Performance Projection: ~20x

BerkeleyGW Example

The benchmark scientific problems chosen are three Si defect supercells of increasing size with 214, 510 and 998 atoms of a divacancy defect in Silicon. The label for these systems are Divac-Si-214, Divac-Si-510, and Divac-Si-998.

Edison Baseline Values:



GPU Numbers:

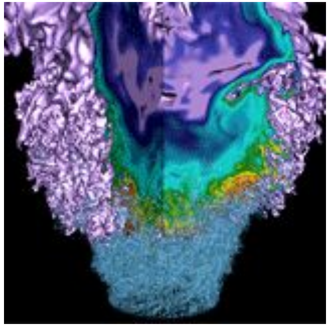
Edison		**GPU Nodes**	
Nodes	Time	Nodes	Time
1280	6618.316	150	1482.258
2048	4662.821	180	1295.988
5184	2333.096		

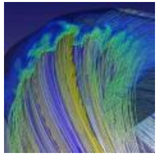
Rough Perlmutter System Performance Projection: ~ 10x

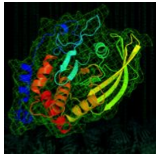
Path Forward

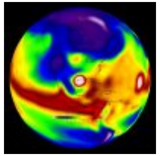
Volta-Next > Volta; GPU-ization of ELPA; Upcoming Hack-a-thon

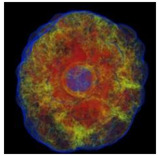
Programming Models and Development Activities

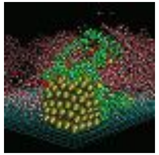












NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Supporting Existing GPU Apps

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

CUDA: MILC, Chroma, HACC ...

CUDA FORTRAN: Quantum ESPRESSO, StarLord (AMREX)

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

Kokkos: LAMMPS, PELE, Chroma ...

Raja: SW4

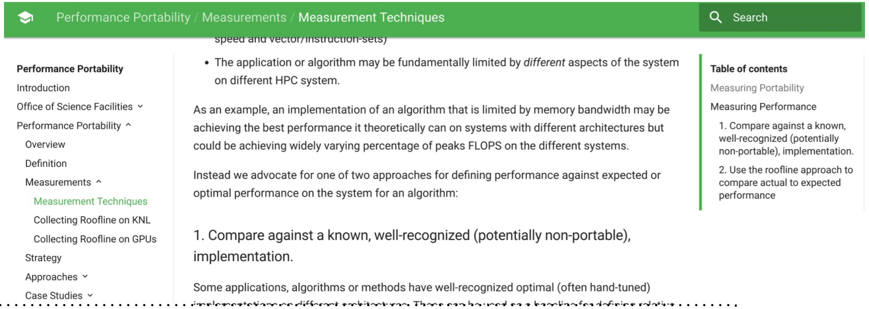
Engaging around Performance Portability

NERSC funding a PGI NRE effort to enable OpenMP GPU acceleration.

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

Directives For Accelerators

NERSC Now a member.


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

- 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)

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 leading development of performanceportability.org

NERSC Lead 2019 DOE COE Perf. Port. Meeting

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

Optimization Challenge and Strategy

Energy-Efficient Processors Have Multiple Hardware Features to Optimize Against:

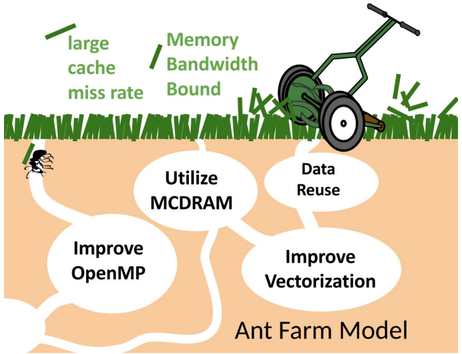
- Many (Heterogeneous) Cores
- Bigger Vectors
- New ISA
- Multiple Memory Tiers

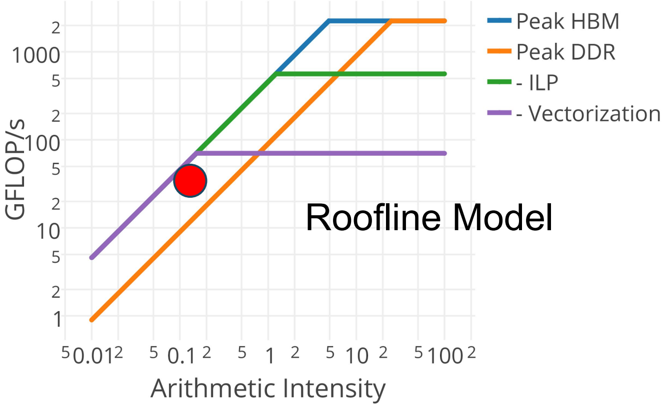
It is easy for users to get bogged down in the weeds:

- How do you know what KNL hardware feature to target?
- How do you know how your code performs in an absolute sense and when to stop?

NERSC has developed tools and strategy for users to answer these questions:

- Designed simple tests that demonstrate code limits
- Use roofline as an optimization guide
- Training and documentation hub targeting all users





Roofline on GPUs

Stay Tuned for Upcoming Training on Roofline Modeling on GPUs!

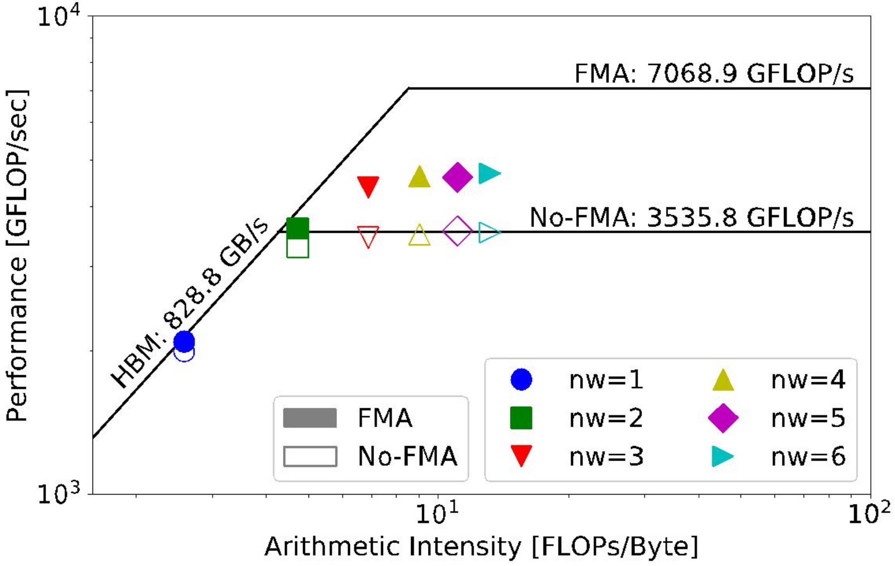
nvprof / Nsight can collect all required metrics including data motion from multiple levels of memory hierarchy: L1/Shared, L2, DRAM, *etc*.

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

Summary

NERSC has built up an application readiness team that is experienced and eager to work with the community to enable new science on Perlmutter.

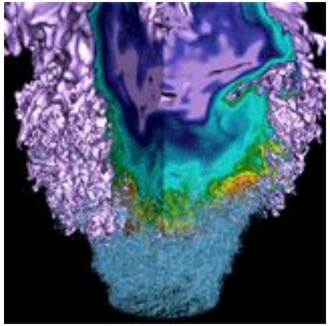
The excitement around GPUs for Science is inspiring.

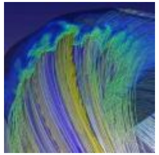
NESAP teams are hard at work.

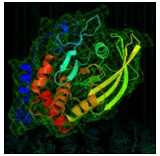
Plenty of opportunities for all NERSC users to participate:

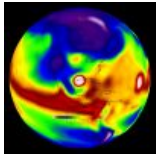
- NERSC staff are mentoring hackathon teams around the country/world
- Community events like GPU for Science Day and public trainings open to all
- NERSC Emphasizing Performance, Portability and Productivity

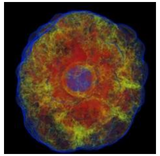
END

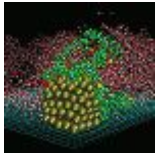




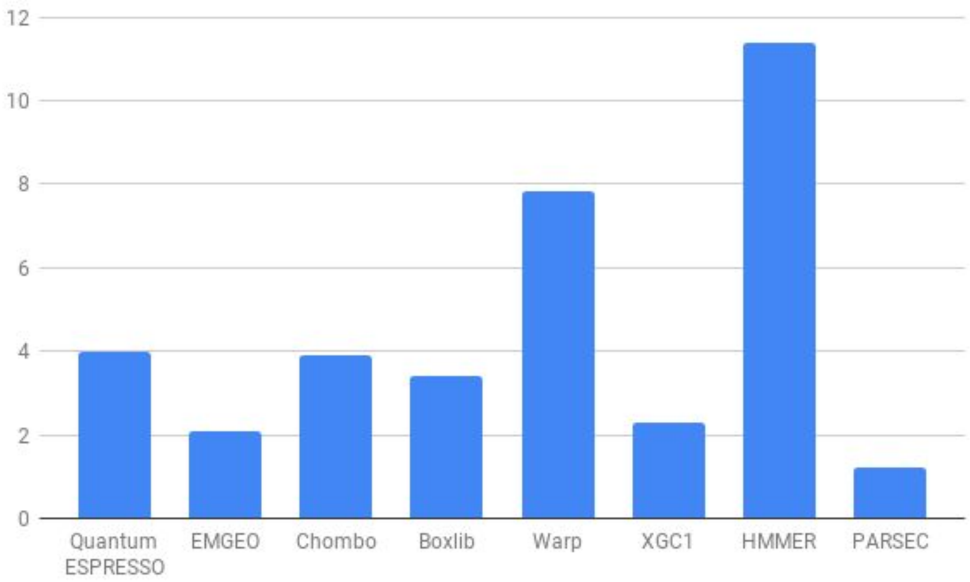








Postdoc Speedups



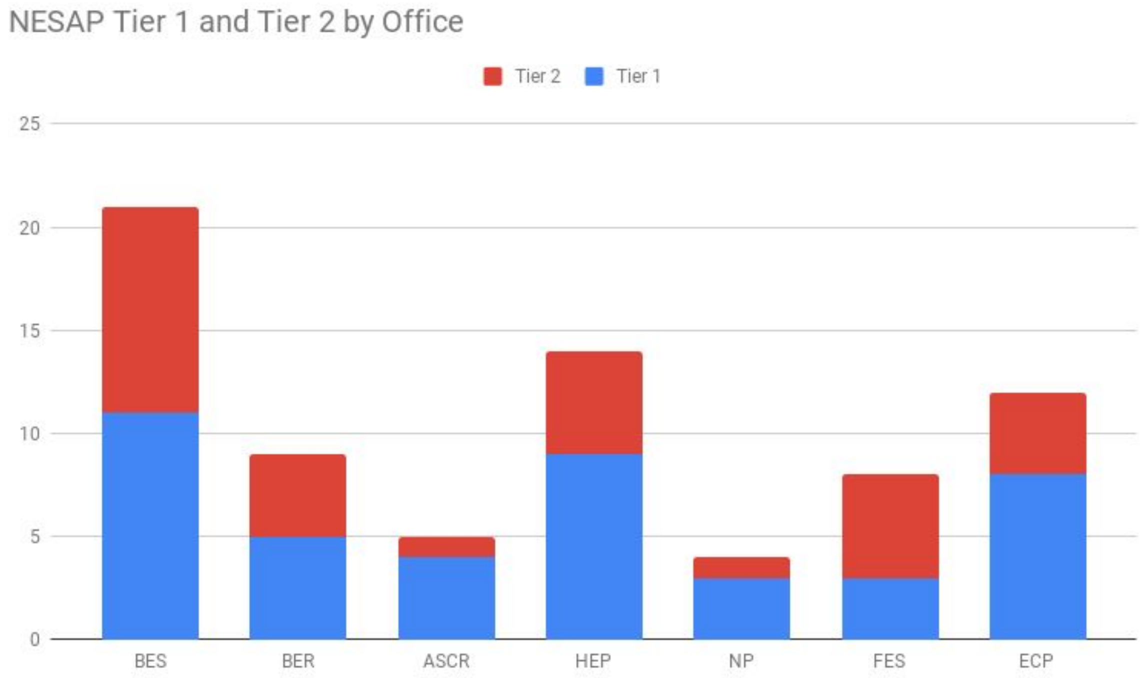
PostDocs made average of 4.5X SpeedUp in NESAP for Cori

Published 20+ Papers Along with NESAP Teams and Staff

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

NESAP By Office



*Some Applications (e.g. USQCD) selected multiple offices

Languages/Programming Models

	GPU Support	FORTRAN 2008	C11	C++17	OpenACC 2.x	OpenMP 5.x	PThreads
PGI							
CCE							
GNU						(Comm. Effort)	
LLVM						(Comm. Effort)	

	CUDA	CUDA FORTRAN	Kokkos	Raja	UPC	Cray MPI
PGI					(BerkeleyUPC)	
CCE						
GNU					(BerkeleyUPC)	
LLVM					(BerkeleyUPC)	

Vendor Supported

NERSC Supported

Data and Analytics Stack

Library	Vendor Supported	NERSC Supported	GPU Enabled	Implementations
Python 2				Cray/Anaconda
Python 3				Cray/Anaconda
Spark				Minerva
R				Minerva
TensorFlow				Minerva/NVIDIA
Keras				Minerva/NVIDIA
Caffe				Minerva/NVIDIA
PyTorch				Minerva/NVIDIA

Cray Will Provide its Minerva Data and Analytics SW Stack.

Vendor Supported

Common GPU Components Supported

NERSC

others available via pip/conda or docker: cuDF, cuGRAPH, ...

GPU Libraries

Library	Vendor Supported	NERSC Supported	Implementations
CrayMPICH			Cray
BLAS			cuBLAS, cuSPARSE, NVBLAS, PGI
RAND			cuRAND
LAPACK			PGI/NVBLAS
ScaLAPACK			PGI/NVBLAS
Magma			Open
FFT			CUFFT
Thrust			CUDA TK

EXASCALE COMPUTING PROJECT

NERSC has an engagement plan with ECP in place to Deploy ECP ST libraries and tools as they become available.

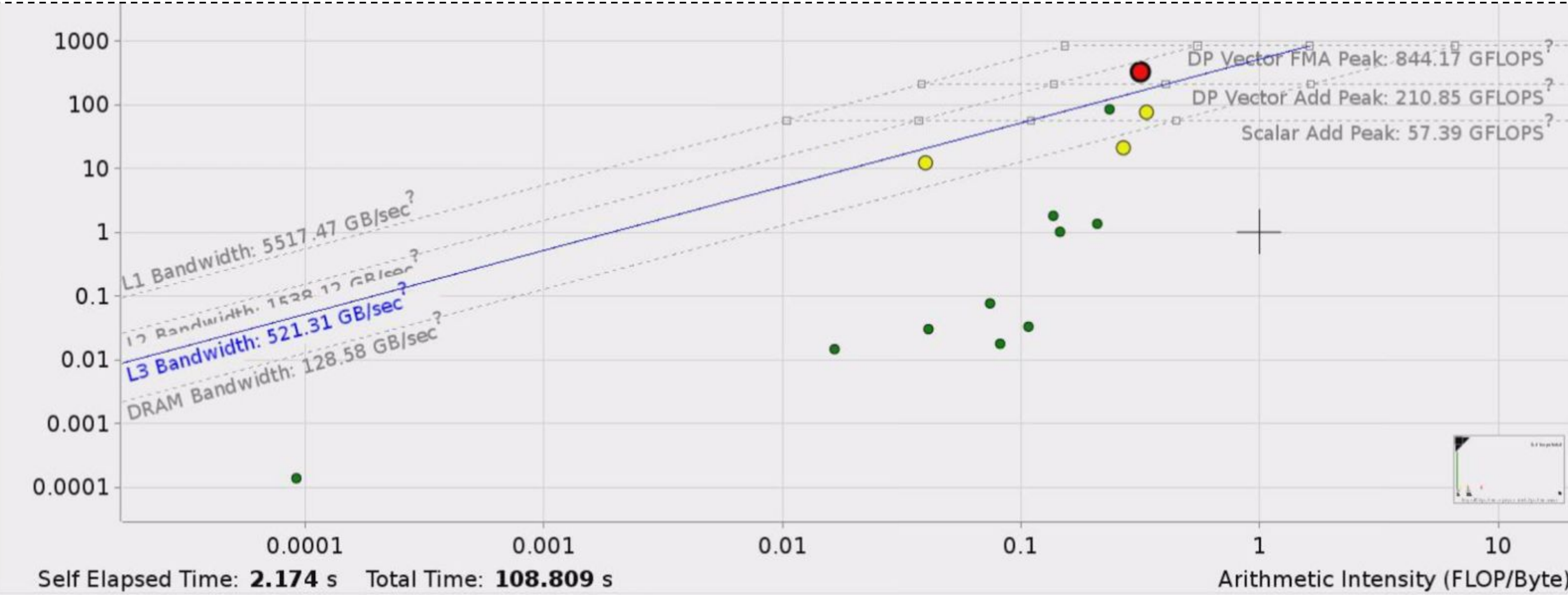
ECP Libraries include: PETSc, Trilinos, Sundials, SuperLU, PEEKS, SLATE ...

Tools

Tools	Vendor Supported	NERSC Supported	CPU	GPU
Totalview				
DDT				
PGI Debugger				
cu-memchk				
cu-gdb				
LGDB				
STAT				
ATP				

Tools	Vendor Supported	NERSC Supported	CPU	GPU
CrayPat				
Apprentice				
PAPI				
NVProf/ NSight				
PGProf				
Tau				
HPC Toolkit				

Tools CoDesign



Intel Vector-Advisor Co-Design - Collaboration between NERSC, LBNL Computational Research, Intel