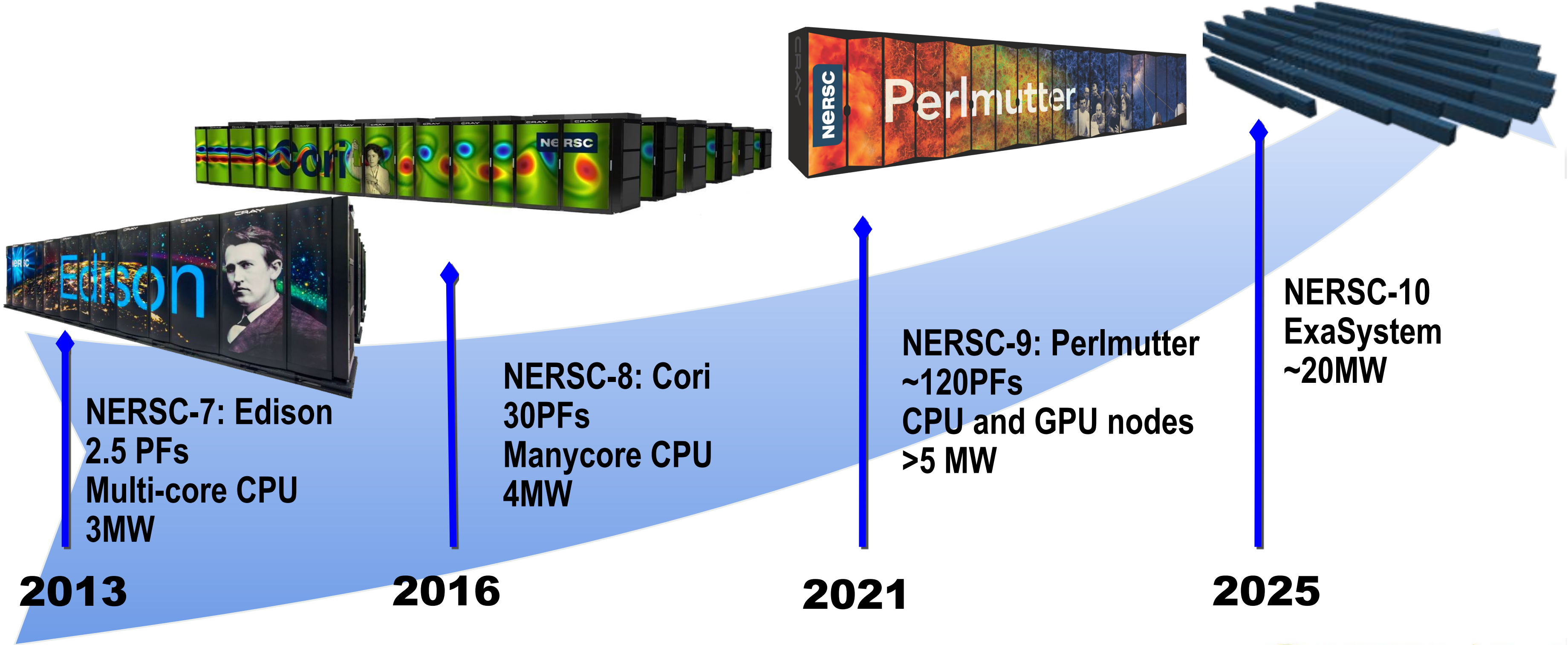
Early Science on Perlmutter

NERSC

Jack Deslippe
Application Performance Lead

September 2021

NERSC Systems Roadmap



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

NERSC-8: Cori
30PFs
Manycore CPU
4MW

NERSC-9: Perlmutter
~120PFs
CPU and GPU nodes
>5 MW

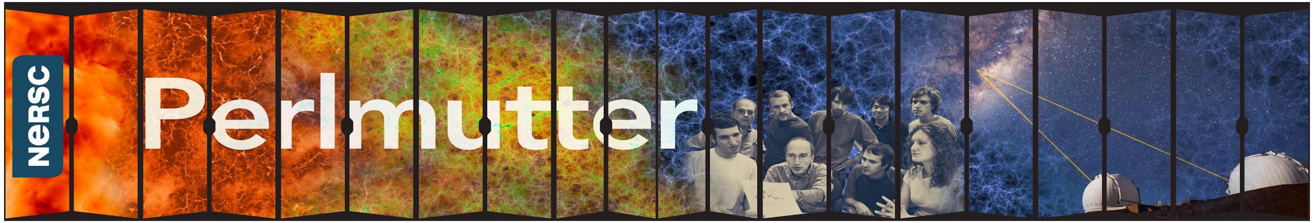
NERSC-10
ExaSystem
~20MW

2013

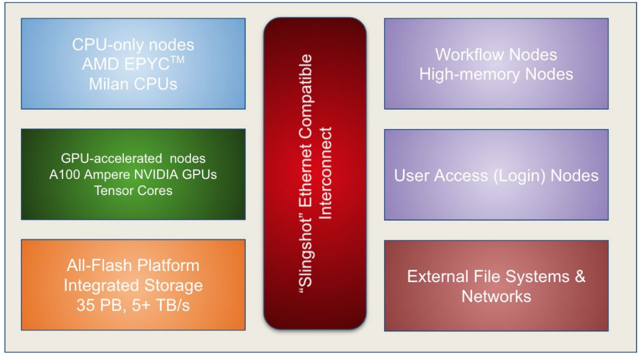
2016

2021

2025



- HPE Cray System with 4x capability of Cori
- GPU-accelerated (GPU/CPU) and CPU-only nodes
- HPE Cray Slingshot high-performance network
- All-Flash filesystem
- Application readiness program (NESAP)



CPU-only nodes AMD EPYC™ Milan CPUs

GPU-accelerated nodes A100 Ampere NVIDIA GPUs Tensor Cores

All-Flash Platform Integrated Storage 35 PB, 5+ TB/s

"Slingshot" Ethernet Compatible Interconnect

Workflow Nodes High-memory Nodes

User Access (Login) Nodes

External File Systems & Networks

Phase I: Arrived Spring 2021

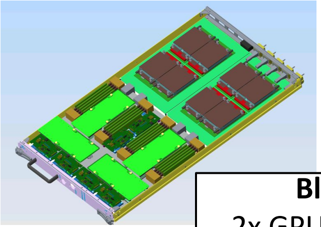
- 1,536 GPU-accelerated nodes
- 1 AMD "Milan" CPU + 4 NVIDIA A100 GPUs per node
- 256 GB CPU memory and 40 GB GPU high BW memory
- 35 PB FLASH scratch file system
- User access and system management nodes

Phase II Addition - arrives this Winter

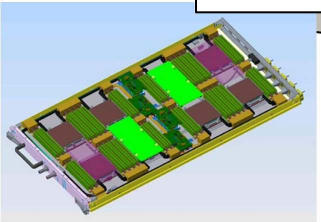
- 3,072 CPU only nodes
- 2 AMD "Milan" CPUs per node
- 512 GB memory per node
- Upgraded high speed network
- CPU partition will match or exceed performance of entire Cori system

Perlmutter at a glance

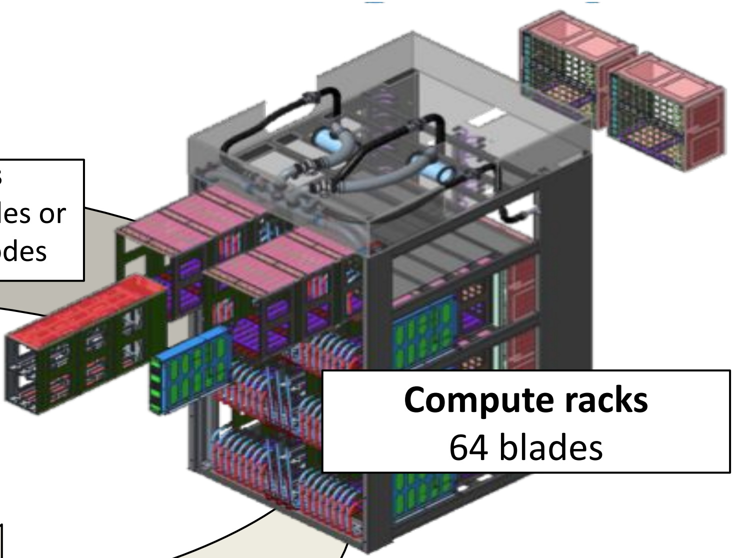
NVIDIA A100 GPU Nodes
4x GPU + 1x CPU
160 GiB HBM + DDR
4x 200G "Slingshot" NICs



AMD EPYC 7003 CPU Node
2x CPUs
> 256 GiB DDR4
1x 200G "Slingshot" NIC



Blades
2x GPU nodes or
4x CPU nodes



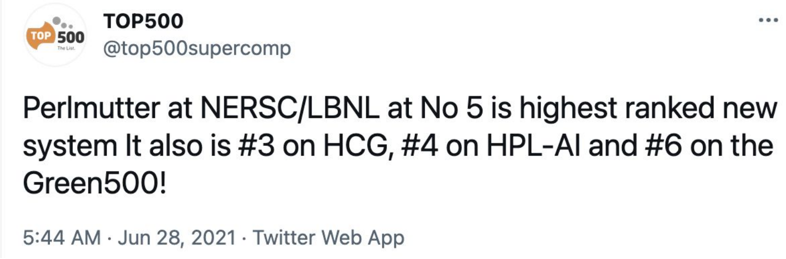
Compute racks
64 blades

Centers of Excellence
Network
Storage
App. Readiness
System SW

→

Perlmutter system
12 GPU racks
12 CPU racks
~6 MW

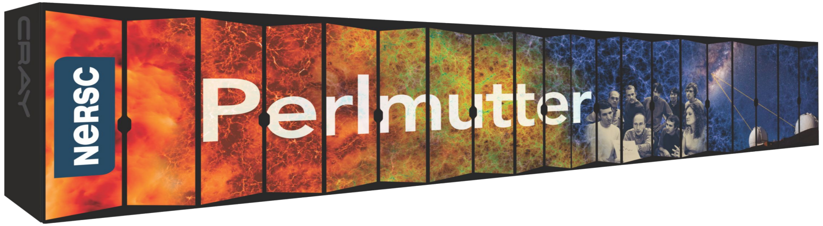
Perlmutter is #6 on Green500 and the most energy-efficient of the Top500 Top 10



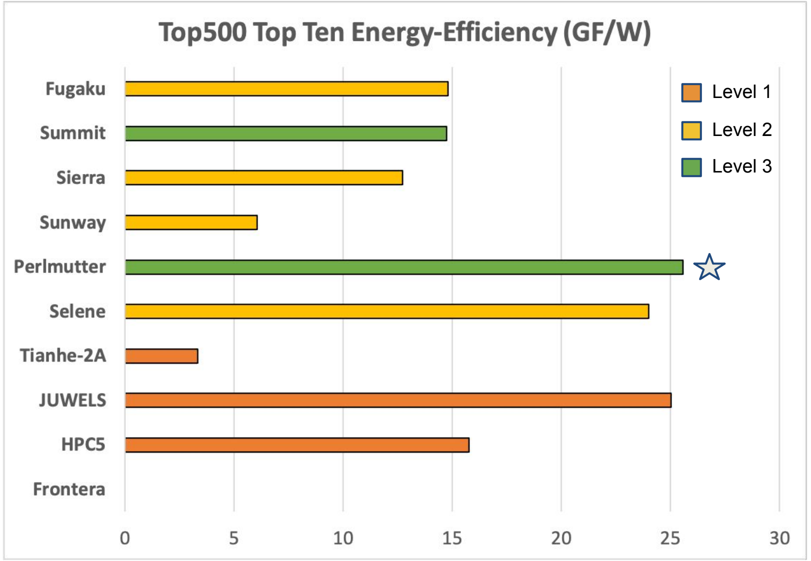
TOP500
@top500supercomp

Perlmutter at NERSC/LBNL at No 5 is highest ranked new system It also is #3 on HCG, #4 on HPL-AI and #6 on the Green500!

5:44 AM · Jun 28, 2021 · Twitter Web App



HPL Performance: 64.59 PF
Energy Efficiency: 25.55 GF/W (core phase)



Comparison of Perlmutter and Cori

Attribute	Cori (2016)	Perlmutter (2021)
Peak Performance	~30 PF	~120 PF
Peak Power	< 4MW	~6 MW
System Memory	~ 1PB (DDR4 + HBM)	> 2PB (DDR4 + HBM)
Node Performance	> 3 TF	> 70 TF
Node Processors	Intel KNL + Intel Haswell	AMD EPYC (Milan) + Nvidia A100 GPUs
# of Nodes	9300 KNL + 1900 Haswell	1536 GPU Accelerated + 3072 CPU only
Intra-Node Interconnect	N/A	NVLink across GPUs; PCIe
Inter-Node Interconnect	Aries	Slingshot
Filesystem	28 PB, 0.75 TB/s	35PB All-Flash; > 4TB/s

Transitioning a Broad Workload to GPUs

NERSC has the most broad/diverse workload in the DOE. Many users have little GPU experience.

NESAP is our application readiness program for preparing our workload for new systems.

Strategy: Partner with application development teams and vendors to port & optimize key applications of importance to the Office of Science. Share lessons learned with with NERSC community via documentation and training.

Resource that have been Available to Teams: NERSC Staff technical liaisons, performance postdocs, access to vendor application engineers, hackathons, early access to hardware (GPU nodes on Cori and Perlmutter)

NESAP Applications Cover the Broad Workload

Electronic Structure	
Quantum ESPRESSO	BES
NWChemEX	BES
VASP	BES
MFDn	NP
WEST	BES
BerkeleyGW	BES

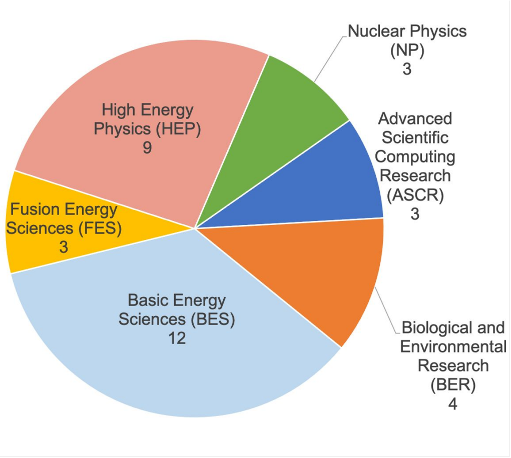
Molecular Dynamics	
EXAALT	FES, NP, BES
NAMD	BES, BER

Data	
DESI	HEP
TomoPy	BES
ATLAS	HEP
ExaFel	BES, ECP
CMS	HEP
ExaBiome	BER, ECP
TOAST	HEP
JGI WorkFlows	BER
LZ	HEP

Learning	
ExaRL	BES
HEP Accel ML	HEP
Catalyst ML	BES
Extreme Spatio-Temporal ML	ASCR
FlowGAN	ASCR

Particles & Grids	
ASGarD	FES, ASCR
WarpX	HEP, ECP
ImSim	HEP
ChomboCrunch	BES, ECP
E3SM	BER, ECP
WDMAPP	FES, ECP

LQCD	
LQCD Consortium	HEP, NP



Tier 1 NESAP Teams

+29 Tier 2 NESAP teams

58 Total NESAP Teams

Perlmutter Supports Every GPU Programming Model

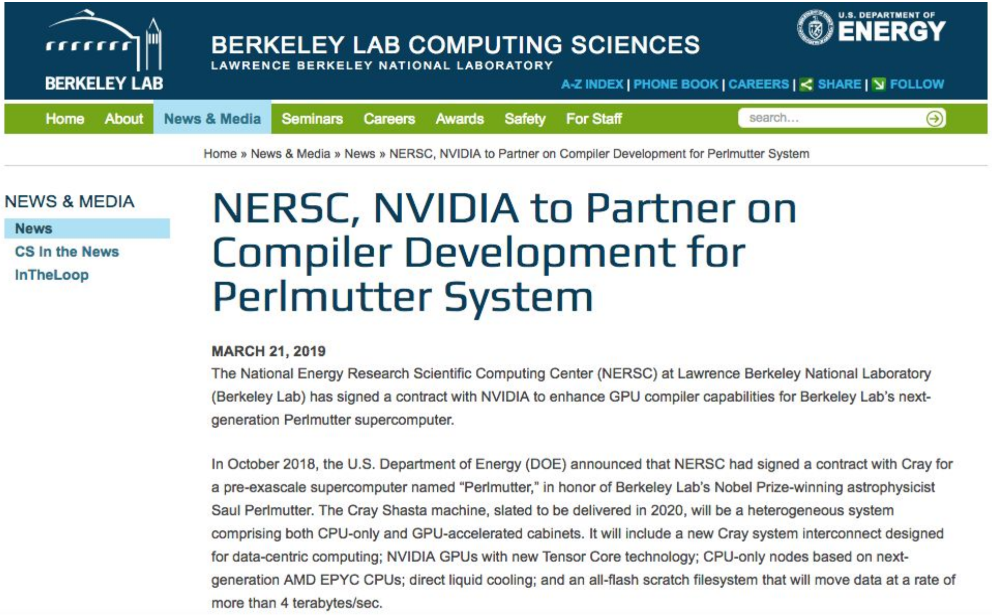
	Fortran/ C/C++	CUDA	OpenACC 2.x	OpenMP 5.x	CUDA Fortran	Kokkos / Raja	MPI	HIP	DPC++ / SYCL
NVIDIA	🟩	🟩	🟩	🟩	🟩	🟩	🟩		
CCE	🟩					🟩	🟩		
GNU	🟩	🟩	🟩	🟩		🟩	🟩		
LLVM	🟧	🟧		🟧		🟧	🟧	🟧	🟧

🟩 Vendor Supported

🟧 NERSC Supported

OpenMP NRE partnership with NVIDIA

- Agreed upon subset of OpenMP features to be included in the NVIDIA (was PGI) compiler
- OpenMP test suite created with micro-benchmarks, mini-apps, and the ECP SOLLVE V&V suite
- 5 NESAP application teams partnered with NVIDIA to add OpenMP target offload directives
- The production OpenMP offload compiler was released in April 2021.


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.

Hackathons

CRAY
a Hewlett Packard Enterprise company

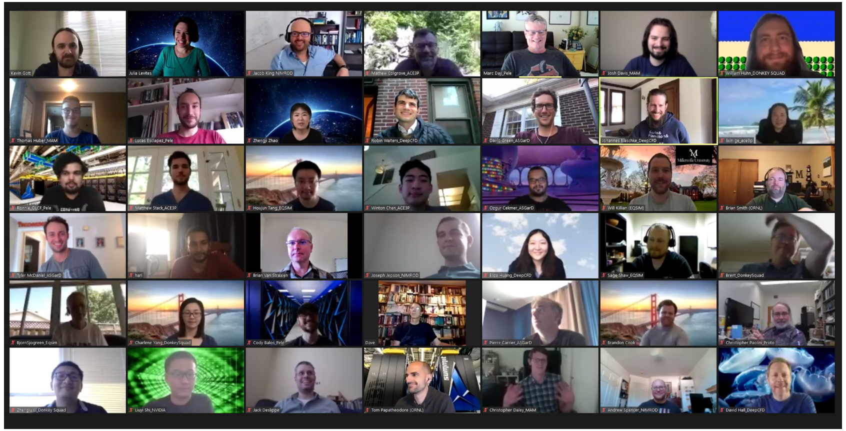
NVIDIA

"Hackathons" have proven to be a highly effective tool for preparing applications for new architectures.

1. **Private** COE Hackathons
   Quarterly with 2-3 NESAP teams + Cray and NVIDIA engineer support.

2. **Public** GPU Hackathons
   (https://www.gpuhackathons.org) **NERSC provided more team mentors than any other institution to worldwide events in 2020**.

   Allows us to reach NERSC teams all around the country and world



NERSC adapted the hackathon format for the COVID work-from-home environment. Instead of on-site, full-day sessions, we moved to a series of shorter sessions spread out over 6-8 weeks.

Features of this format were popular and effective and we plan to incorporate them into future hackathons.

Broad impact and enablement

Programming models and languages

kokkos

SYCL

ROCm

OpenMP

OpenACC
Directives for Accelerators

ortran

Community Codes

VASP

LAMMPS

CP2K

WEST!

QUANTUMESPRESSO

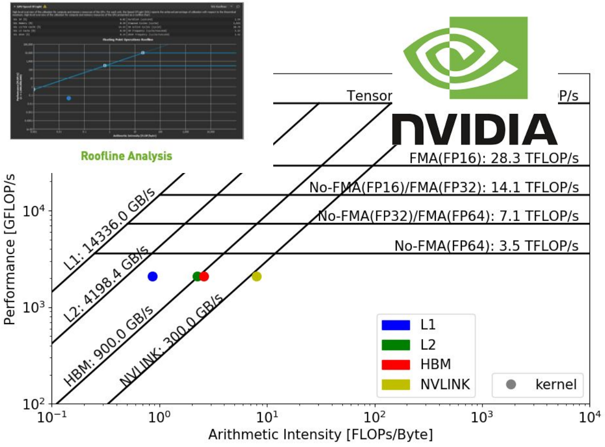
Community Resources

NERSC NERSC Documentation

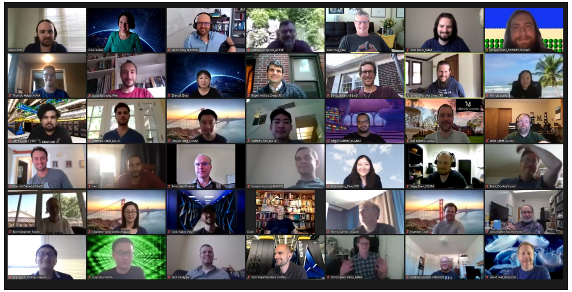
NERSC TRAINING EVENTS

SC20

Vendor tools



Community GPU hack-a-thons

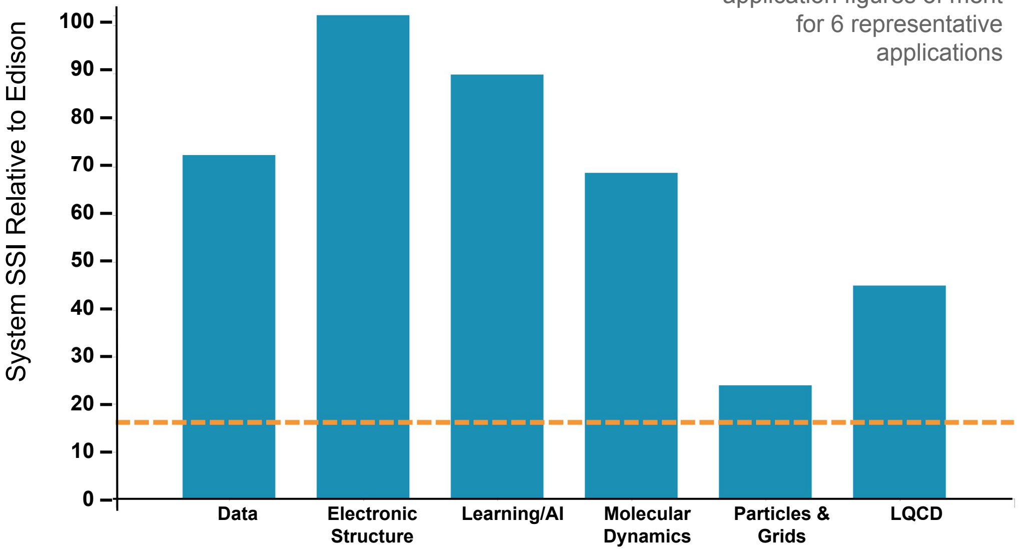


Projected Application Performance

- We use Perlmutter and Previous GPU performance measurements to estimate/extrapolate a system wide throughput speedup on Perlmutter vs. Edison (the NERSC-7 system).

- Applications from different science areas and algorithmic spaces are able to utilize Perlmutter GPUs

Based on measured application figures of merit for 6 representative applications

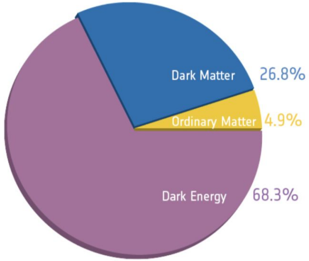


Perlmutter System-Wide Performance Performance

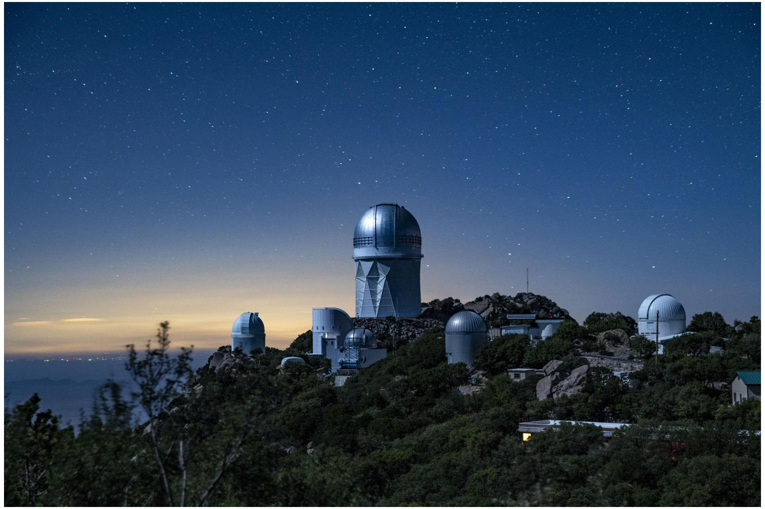
6 applications from different areas of the workload achieve 20X Systemwide speedup over Edison.

DESI Dark Energy Spectroscopic Instrument

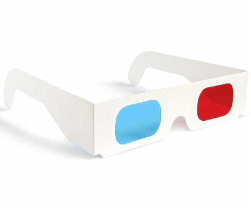
Science: Understand Dark Energy



Scientists believe about 70 percent of the universe is dark energy, although we don’t have a good understanding of what it is



The DESI instrument will send NERSC data every night for 5 years

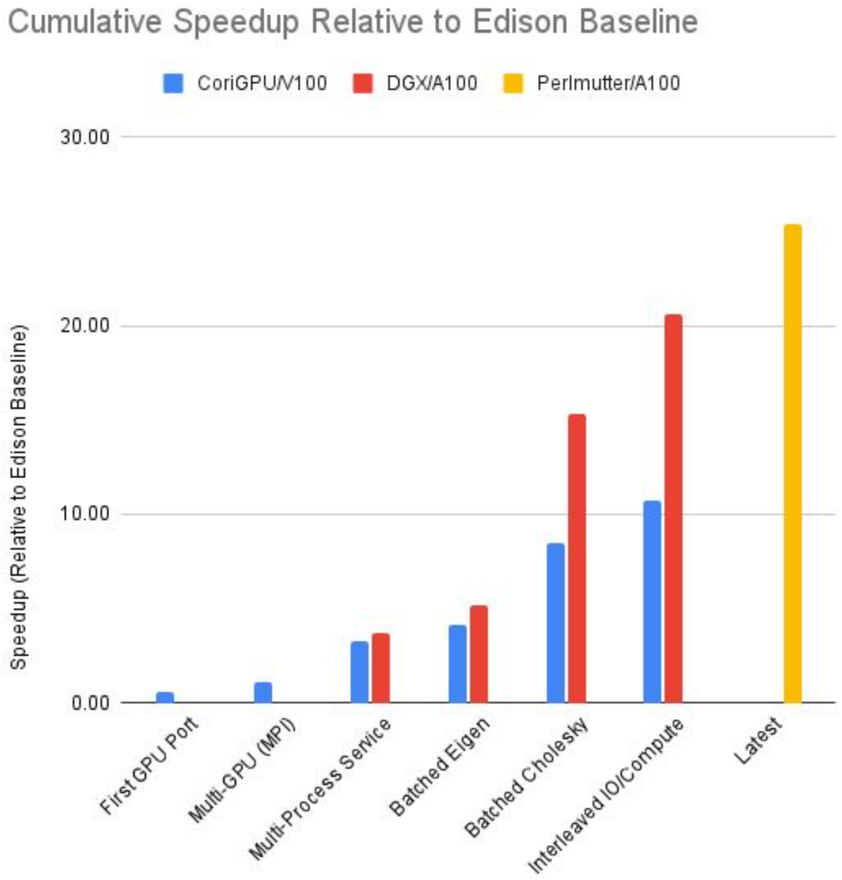


Data will be used to construct the most detailed 3D map of the universe to date and better understand the nature of dark energy

DESI

Dark **E**nergy **S**pectroscopic **I**nstrument

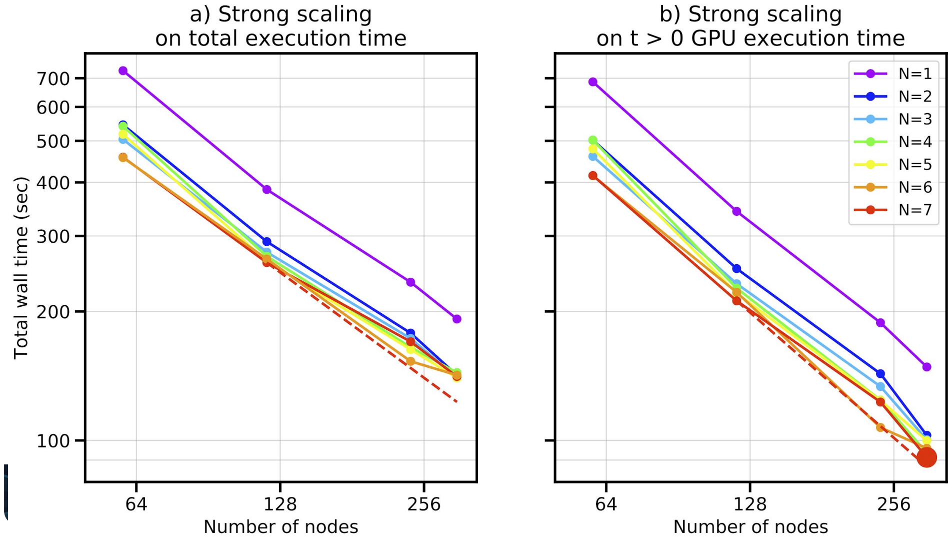
- DESI Spectral Extraction is an image processing code implemented in Python.
- Completed major refactor of optimized CPU code and initial GPU port in early 2020.
- Major optimization milestones include: saturating GPU utilization using MPI and CUDA Multi-Process Service, refactoring code to leverage batched linear algebra operations on GPU, and interleaving IO with computation.
- **25x** improvement in per-node throughput using Perlmutter compared to Edison baseline.



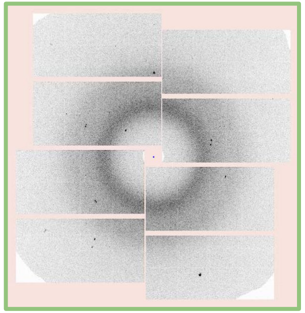
ExaFEL

XFEL requires **real-time data analysis** to make decisions **during ongoing experiments**. Data collection rates outpacing computational resources at the experimental sites, **requiring a Superfacility approach**.

In two years, NESAP has developed a highly scalable CUDA/GPU application. **CCTBX/nanoBragg w/ runtime improved from 12.3 hours on Edison, to 2 minutes on Summit**.

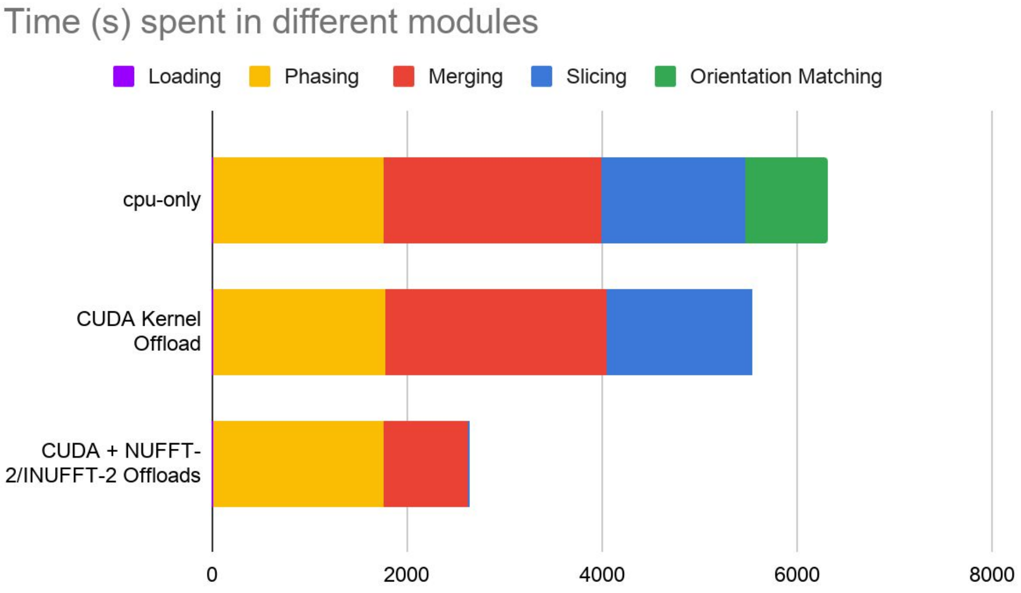


CCTBX/nanoBragg strong scaling on Summit. Colored lines show number of concurrent streams per GPU



ExaFEL

NESAP has been essential in **developing a scalable version of the MTIP algorithm** (figure, right). By offloading kernels to CUDA, **MTIP/Spinifel runtime was decreased by 2.4x over CPU-only code**.



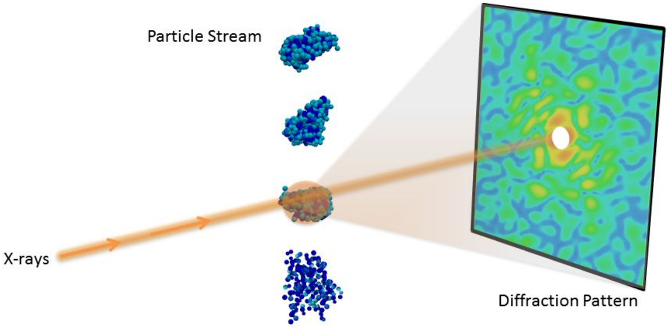
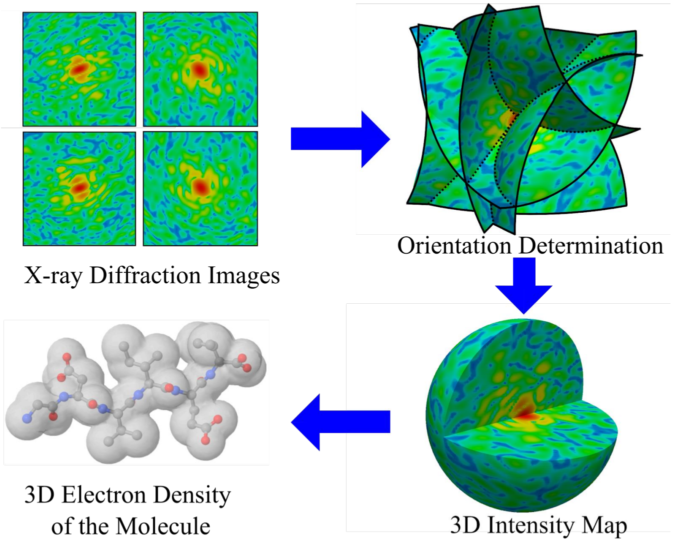
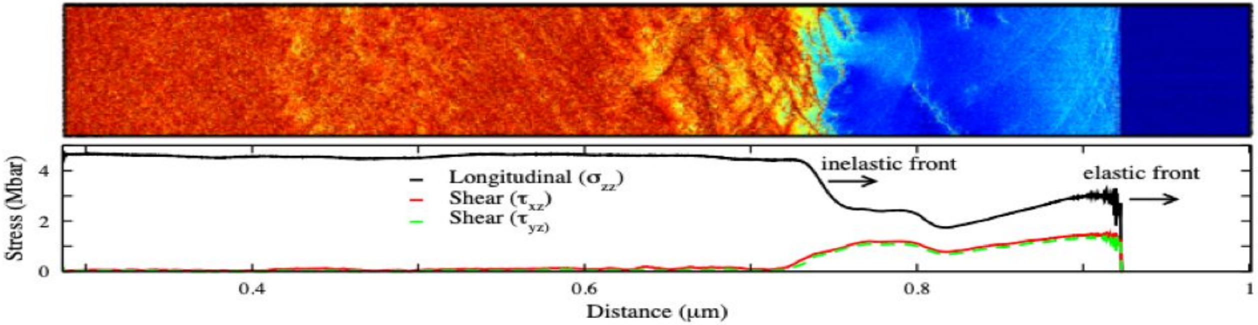


Illustration of **SPI technique**: the X-ray beam interacts with only a few molecules a time


X-ray Diffraction Images
Orientation Determination
3D Intensity Map
3D Electron Density of the Molecule

Record Scale MD With LAMMPs
Gordon Bell Finalists

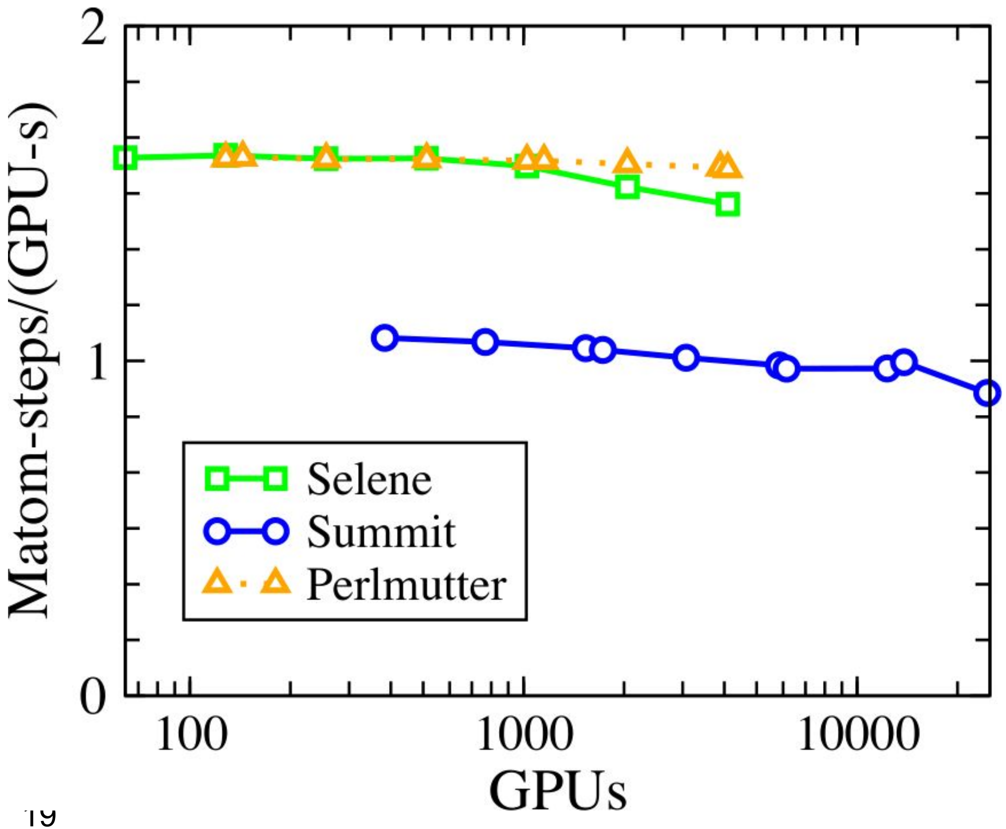
- Collaborative effort: University of South Florida, Sandia, NERSC and NVIDIA
- Billion atom molecular dynamics simulation (20B atoms)
  - SNAP quantum-accurate machine learned interatomic potential
  - Kokkos CUDA backend for NVIDIA GPUs
  - A run achieved 11.24 PFLOPS on Perlmutter on 1024 nodes (~ 2/3rd of the total machine)
- Simulation model shock compression of carbon at extreme pressures and temperatures.



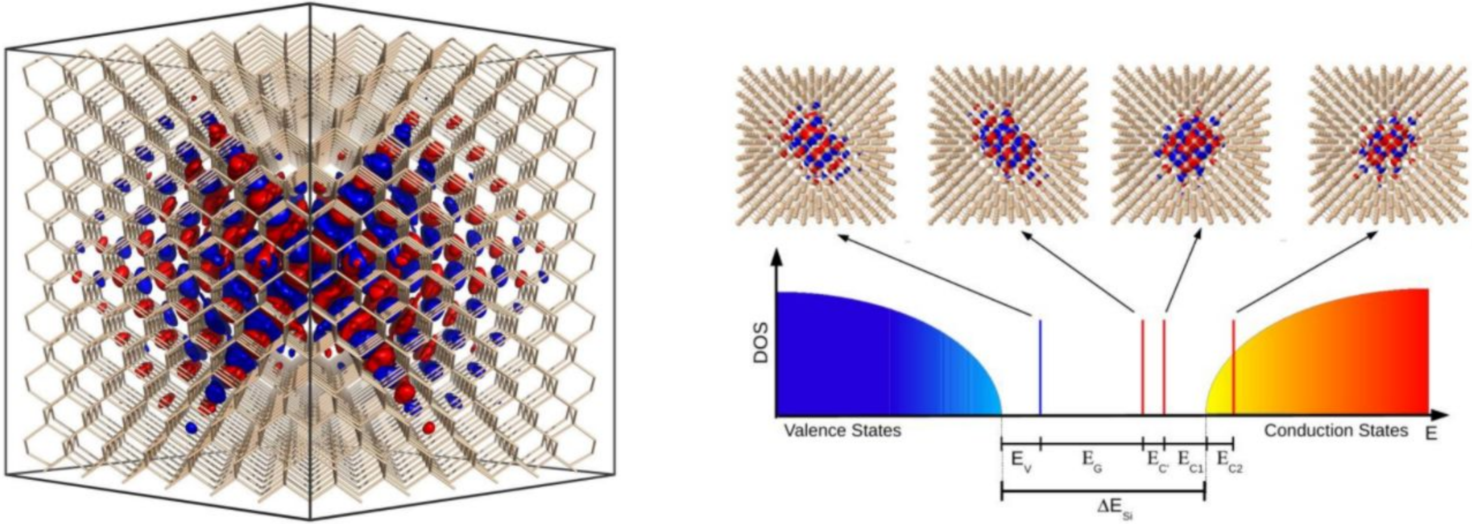
1.8 billion carbon atom simulation of split elastic-inelastic shock wave propagating in single crystal diamond (dark blue). The elastic precursor (light blue) is followed by an inelastic wave (red), which exhibits an unexpected stress relaxation mechanism

Record Scale MD With LAMMPs Gordon Bell Finalists

Strong scaling the amorphous carbon benchmark on Perlmutter and related systems.



Qubit Design With BerkeleyGW



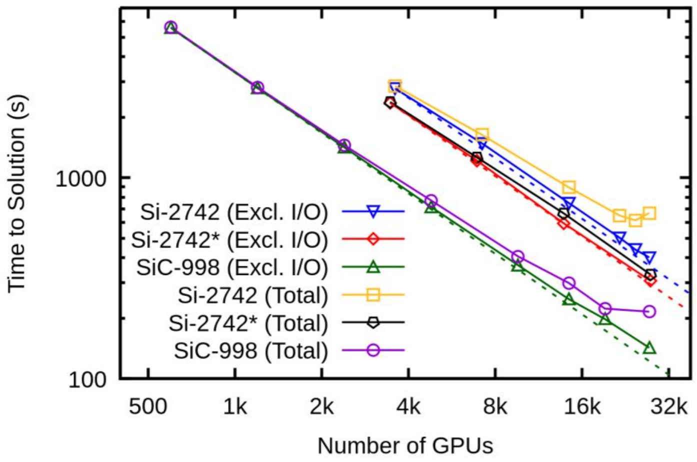
Example: Divacancy point defect in crystalline silicon, prototype of a solid-state Qubit

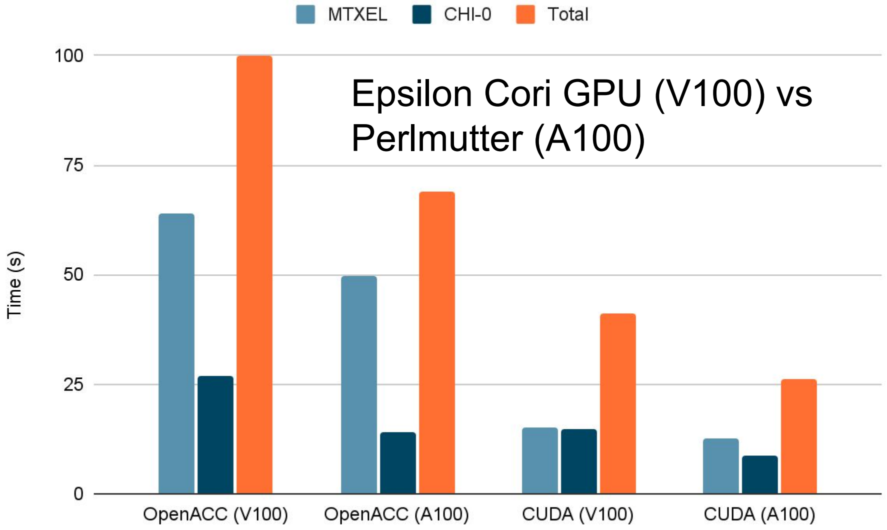
Accurate prediction requires:

- Accuracy beyond DFT: **GW and GW+BSE**
- Unprecedented simulation sizes: **1000's of atoms**

Qubit Design

The BerkeleyGW NESAP team was recognized as a Gordon Bell finalist in 2020.




Epsilon Cori GPU (V100) vs Perlmutter (A100)

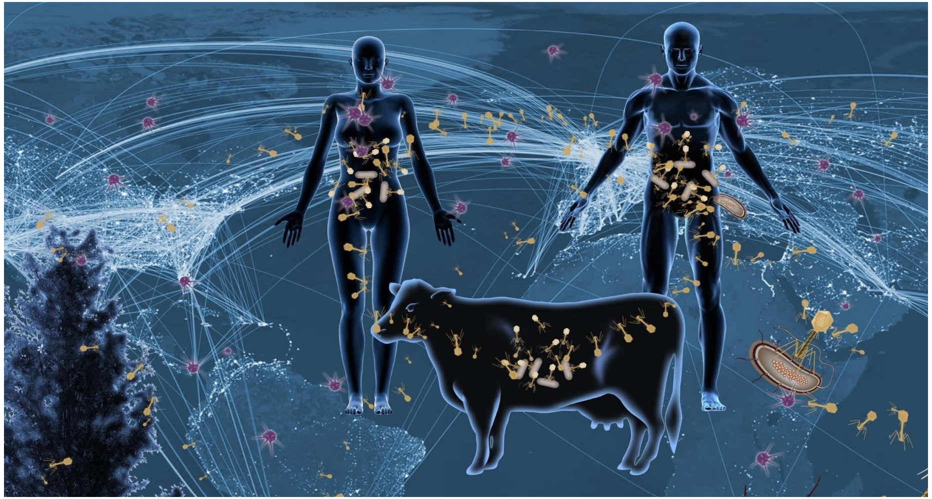
	MTXEL	CHI-0	Total
OpenACC (V100)	64	27	100
OpenACC (A100)	49.8	14.2	69
CUDA (V100)	15.2	14.7	41
CUDA (A100)	12.6	8.7	26.2

- Si-214 system (scaled: 4Ry CT ; 3000 bands). 8 GPUs each.

BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY
Office of Science

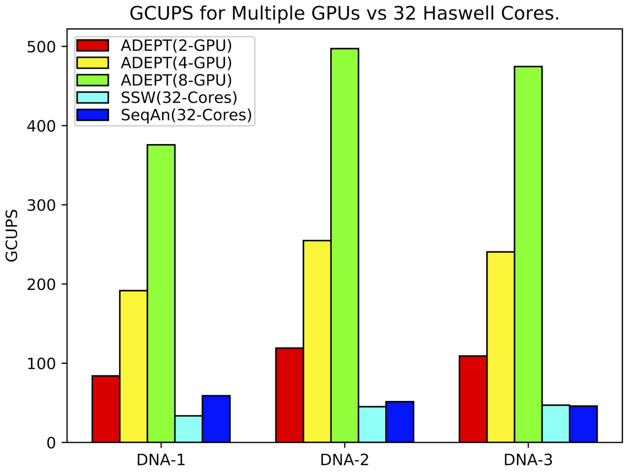
Exabiome (Meta-Genomics)

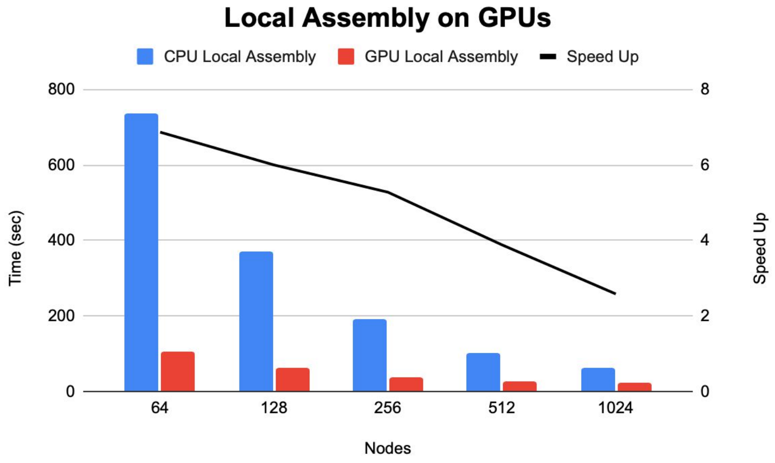


- **Microbes:** these are single cell organism, e.g. viruses, bacteria
- **Microbiomes:** communities of microbial species living in our environment.
- **Metagenomics:** genome sequencing of these communities.

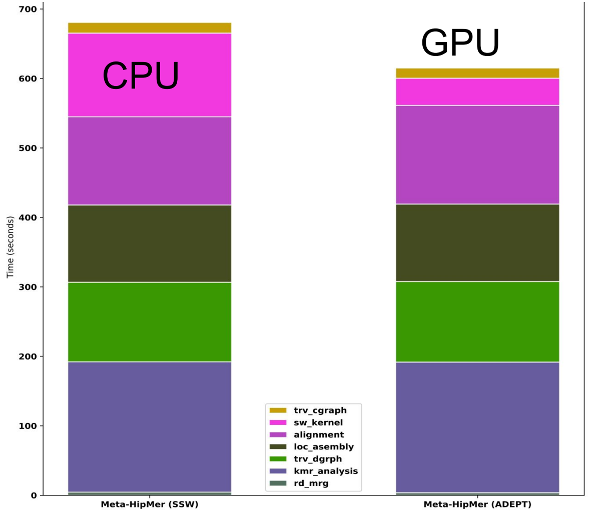
Exabiome (Meta-Genomics)

- A lot of progress has been made on GPU algorithms for meta-genomics.
- This NESAP team wrote the world's fastest GPU aligners using a lot of clever strategies, newly available GPU intrinsic instructions etc.
- With the help of warp level intrinsics, dynamic data structures were written for GPUs from scratch to re-write the Local Assembly stage.





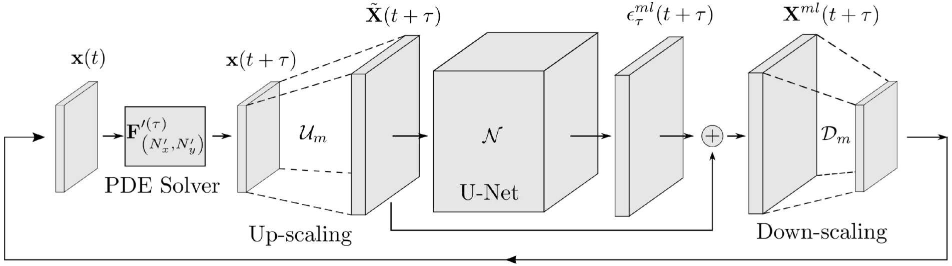
At large scales, sensitivity to communication latency dominates. Being addressed.



Accelerating CFD with GANs on Perlmutter

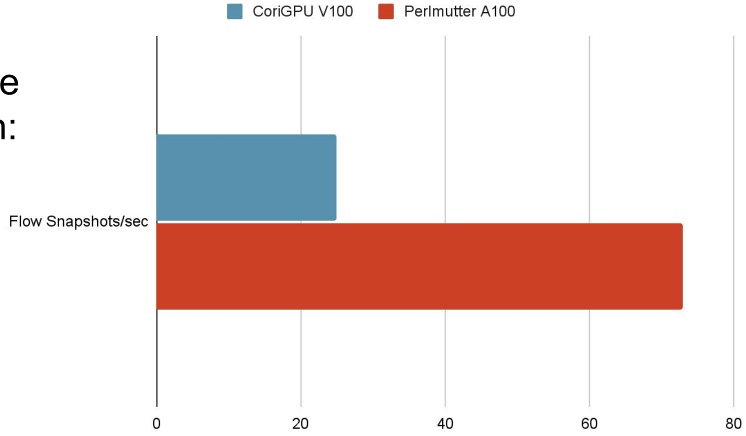
The FlowGAN project introduces a technique based on a deep neural network architecture to augment traditional numerical simulations of fluid flows. The ML model is used to correct the numerical errors induced by a coarse-grid simulation of turbulent flows at high-Reynolds numbers.

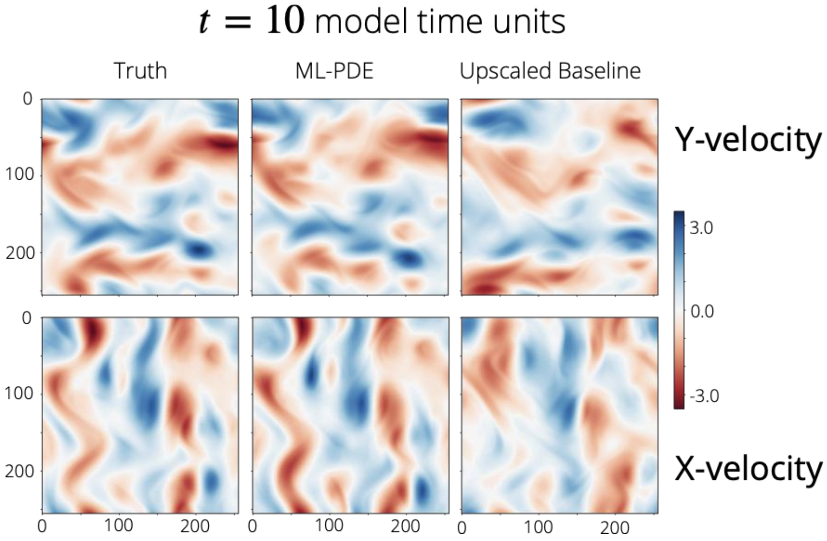
FlowGAN architecture:


$\mathbf{x}(t)$
$\mathbf{F}'^{(\tau)}_{(N'_x,N'_y)}$
PDE Solver
$\mathbf{x}(t+\tau)$
$\mathcal{U}_m$
Up-scaling
$\tilde{\mathbf{X}}(t+\tau)$
$\mathcal{N}$
U-Net
$\epsilon^{ml}_{\tau}(t+\tau)$
$\mathbf{X}^{ml}(t+\tau)$
$\mathcal{D}_m$
Down-scaling

Performance Comparison:

2.9x performance improvement over CoriGPU on ML workflow





BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY | Office of Science

Key Takeaways

- NERSC successful in preparing a significant number of key Office of Science applications for Perlmutter
  - Early engagement and access to GPU technologies
  - Embedded Postdocs
  - Focused Hackathons
- NERSC continuing to engage w/ broad NERSC community to enable use of Perlmutter productively
  - Encouraging community to join GPUHackathons.org events all over the country next year
- GPU optimizations (Increasing Parallelism, Understanding and Minimizing Code Movement) continue lesson learned from Cori
- OpenMP and C++ Frameworks (Kokkos etc.) are viable performance portable options.

Questions?

Large Scale Combustion Modeling w/ Pele

Combustion Fuel □ Methane

- DRM19 chemistry with 21 species
- ERK chemistry solver
- 2 AMR levels

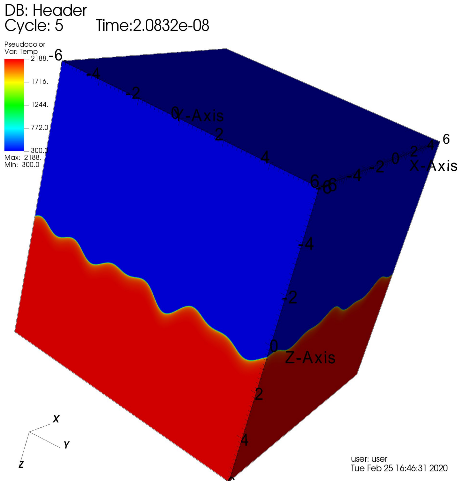
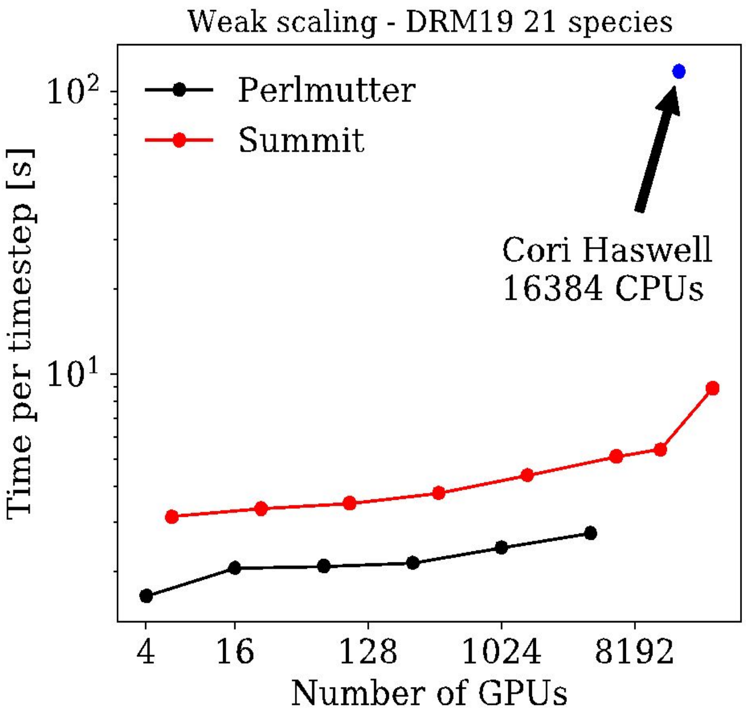


Figure above shows a statistically stationary flame. This flame configuration has been extensively used in DNS calculations and in this case it is used for scaling tests.

The configuration can be easily reproduced with different chemical mechanisms.

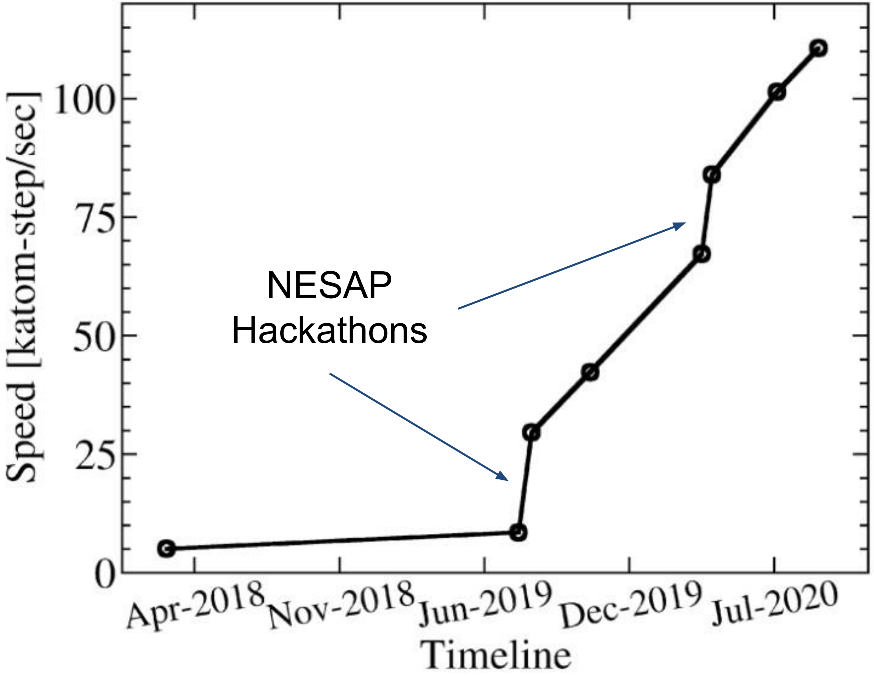


LAMMPs

- LAMMPS is a classical molecular dynamics code with a focus on materials modeling
- Production LAMMPS/Kokkos version was highly optimized over a serious of hackathons - Joint effort of NERSC/NESAP, ECP, NVIDIA and HPE
- Every kernel was rewritten and optimized individually, compared to baseline
- **22x** improvement in performance compared to baseline on NVIDIA V100 GPU (previous generation than on Perlmutter).
- SSI is the system-wide throughput increase over Edison in atom-steps/second.

SSI: 69

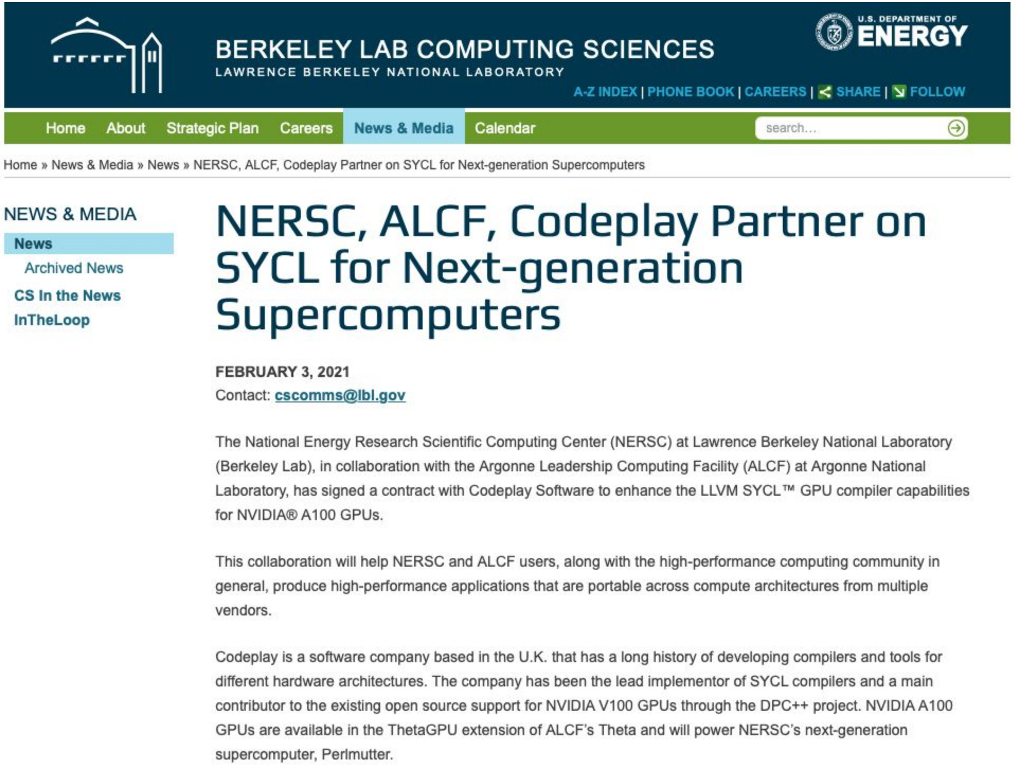
Node vs Node Speedup: 250x



NESAP Hackathons

NERSC, ALCF and Codeplay partnership on SYCL

- Target SYCL 2020 (latest specification) support on Ampere A100 GPUs
- Open LLVM based compiler
- Provides Portability for Apps Developed for Aurora
- Extensions for A100
  - Asynchronous Copy
  - Asynchronous Barrier
  - Tensor core types/ APIs


BERKELEY LAB COMPUTING SCIENCES
LAWRENCE BERKELEY NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
A-Z INDEX | PHONE BOOK | CAREERS | SHARE | FOLLOW
Home
About
Strategic Plan
Careers
News & Media
Calendar
search...
Home » News & Media » News » NERSC, ALCF, Codeplay Partner on SYCL for Next-generation Supercomputers
NEWS & MEDIA
News
Archived News
CS In the News
InTheLoop

NERSC, ALCF, Codeplay Partner on SYCL for Next-generation Supercomputers

FEBRUARY 3, 2021
Contact: cscomms@lbl.gov

The National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory (Berkeley Lab), in collaboration with the Argonne Leadership Computing Facility (ALCF) at Argonne National Laboratory, has signed a contract with Codeplay Software to enhance the LLVM SYCL™ GPU compiler capabilities for NVIDIA® A100 GPUs.

This collaboration will help NERSC and ALCF users, along with the high-performance computing community in general, produce high-performance applications that are portable across compute architectures from multiple vendors.

Codeplay is a software company based in the U.K. that has a long history of developing compilers and tools for different hardware architectures. The company has been the lead implementor of SYCL compilers and a main contributor to the existing open source support for NVIDIA V100 GPUs through the DPC++ project. NVIDIA A100 GPUs are available in the ThetaGPU extension of ALCF's Theta and will power NERSC's next-generation supercomputer, Perlmutter.