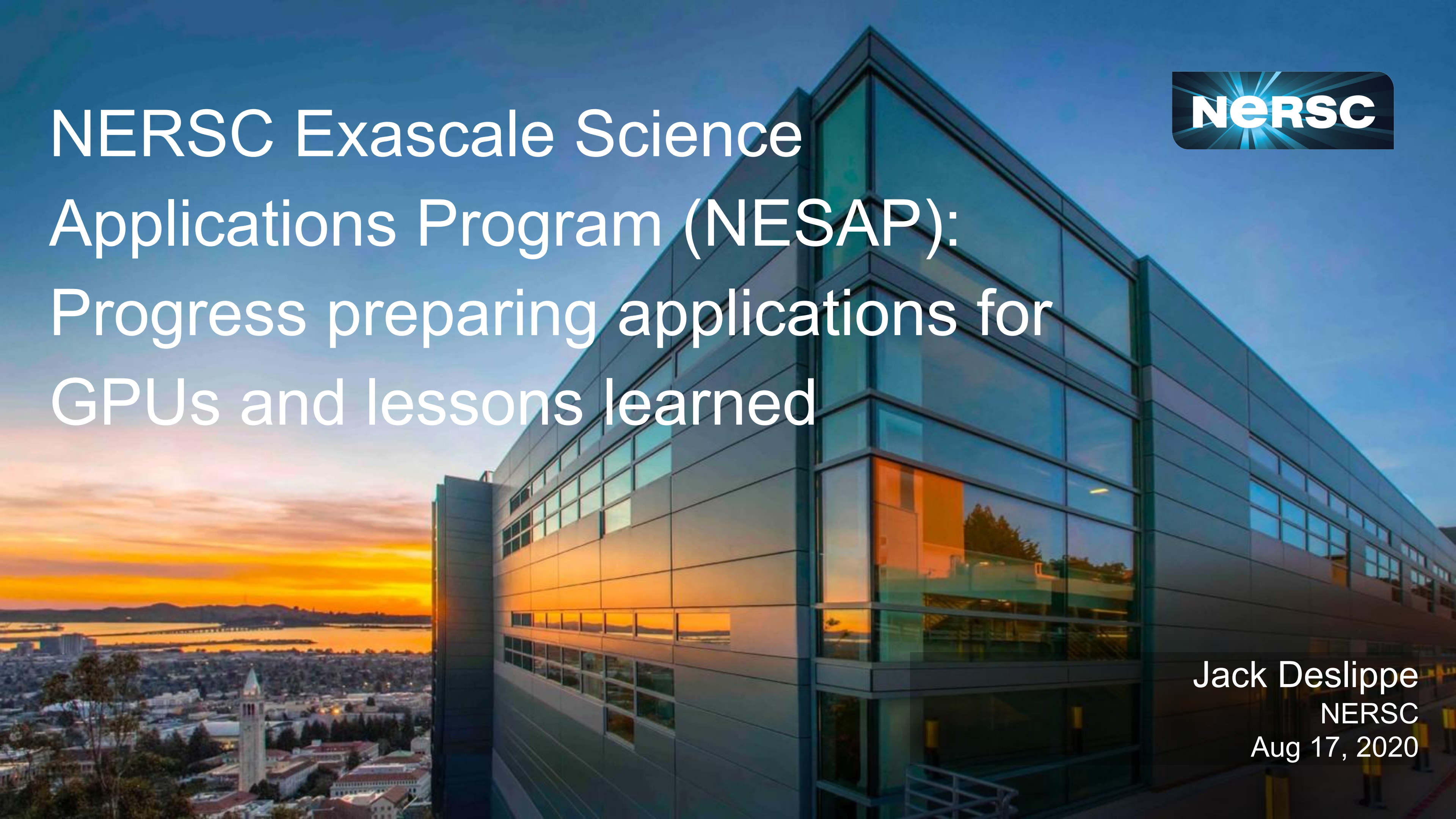
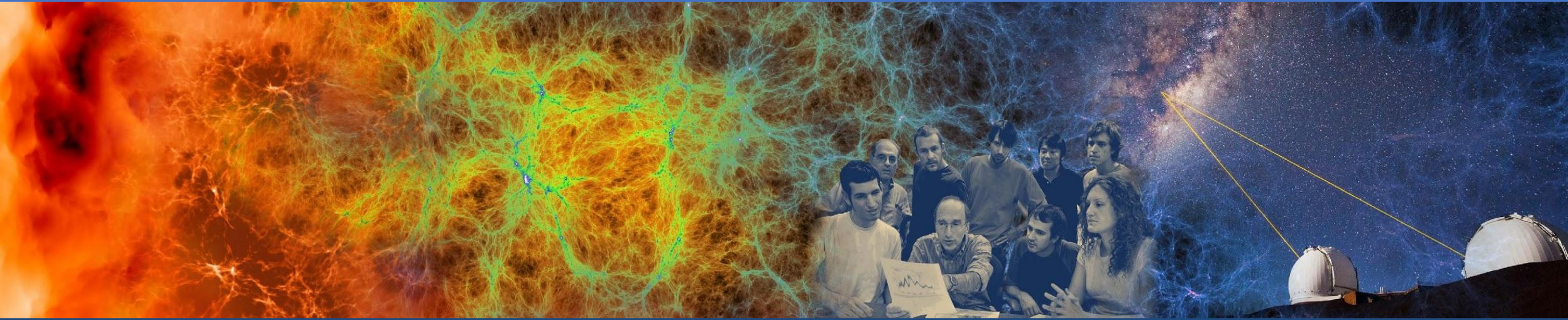


NERSC Exascale Science Applications Program (NESAP): Progress preparing applications for GPUs and lessons learned

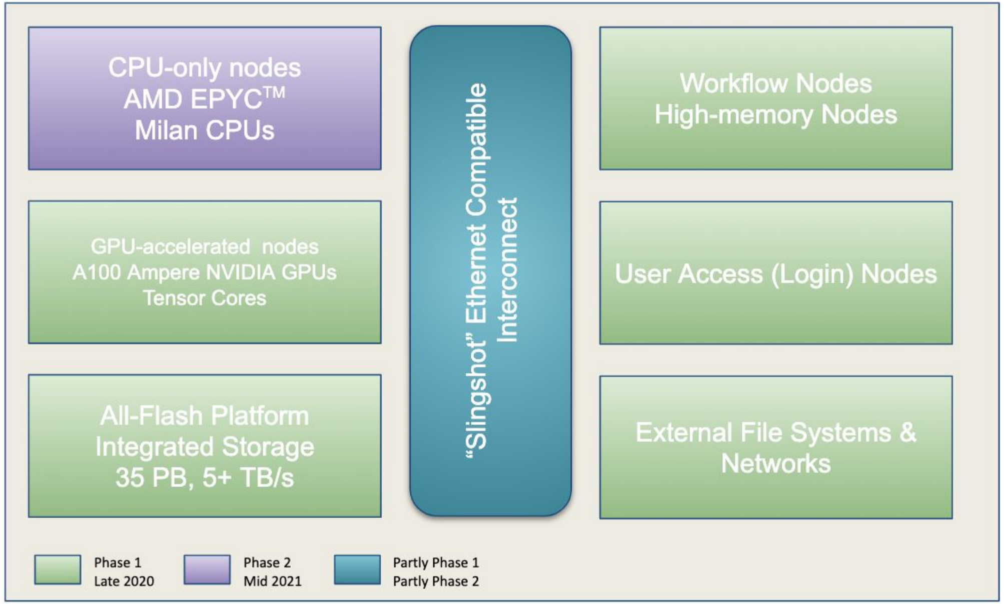
Jack Deslippe
NERSC
Aug 17, 2020

Perlmutter and NERSC Roadmap



Perlmutter: a System Optimized for Science

- NVIDIA A100-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



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 1 Late 2020

🟪 Phase 2 Mid 2021

🟦 Partly Phase 1 Partly Phase 2

CPU Nodes

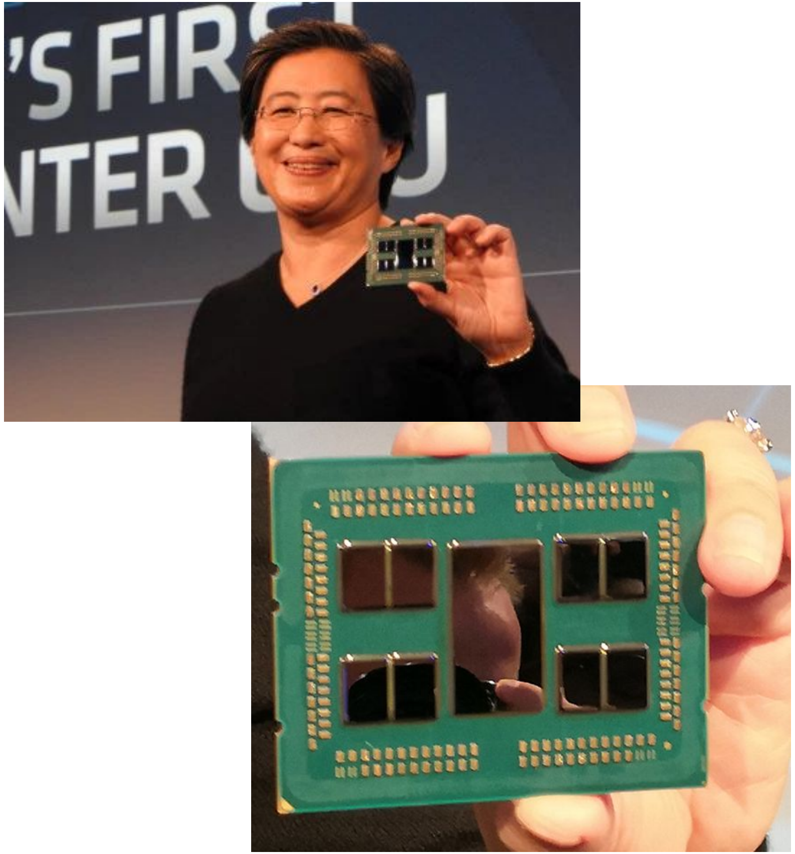
AMD “Milan” CPU

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

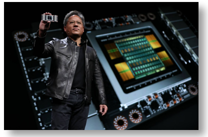
>=Rome specs

8 channels DDR memory

~ 1x Cori



GPU Nodes

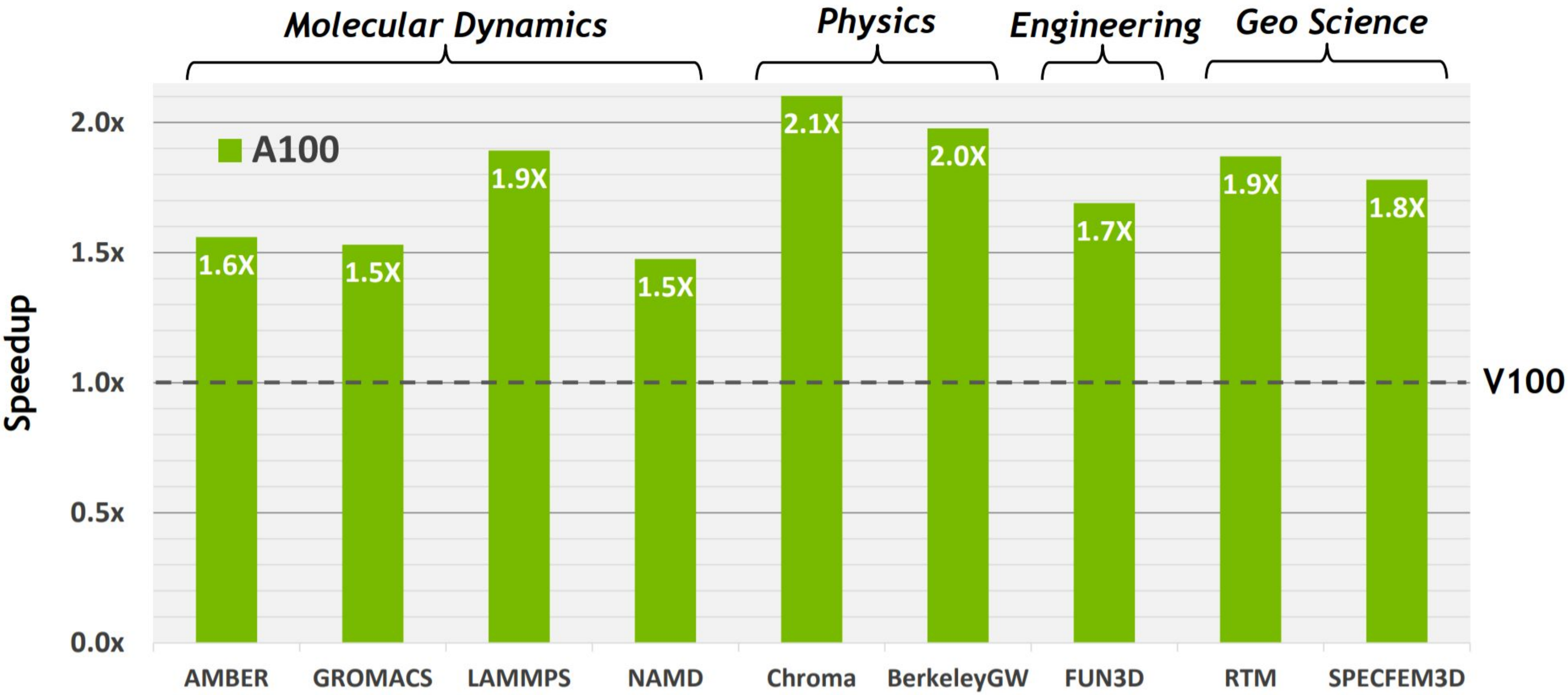


- 4x NVIDIA Ampere (A100) GPUs
- 1 AMD Milan CPU

- NVLINK-3 (Between 4 GPUs)
- FP16, TF32, FP64 Tensor Cores
- GPU direct
- Multi-Instance GPU (MIG)

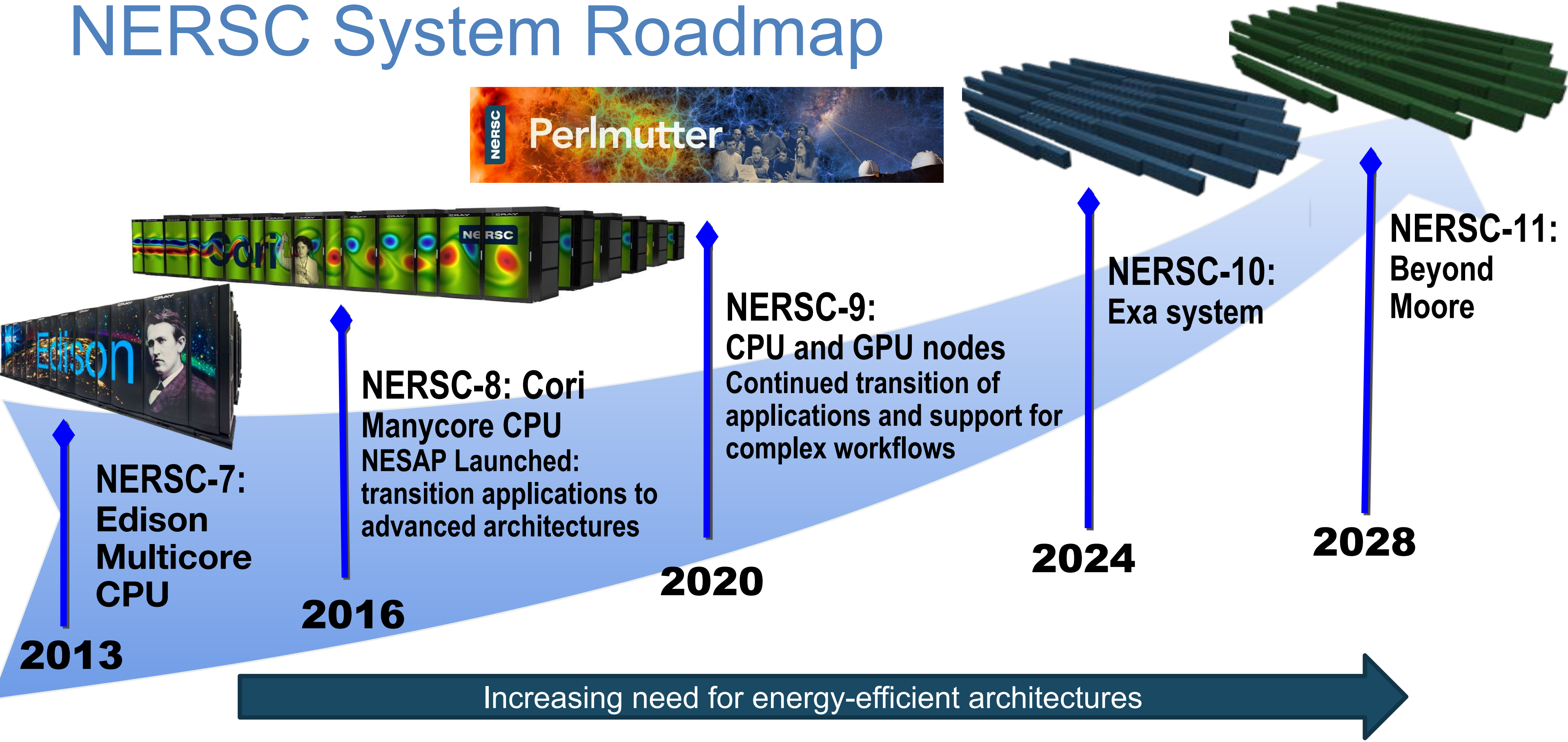
	V100	A100
FP64 Peak	7.5 TF FMA	19.5 TF TC (9.7 TF FMA)
FP16 Peak	125 TF TC	312 TF TC
SMs	80	108
Memory BW	900 GB/s	1555 GB/s
Memory Size	16 GB	40 GB
L2 Cache	6 MB	40 MB
Shared Mem. / SM	96 KB	164 KB

A100 vs V100



Molecular Dynamics
Physics
Engineering
Geo Science
🟩 A100
Speedup
2.0x
1.5x
1.0x
0.5x
0.0x
V100
1.6X
1.5X
1.9X
1.5X
2.1X
2.0X
1.7X
1.9X
1.8X
AMBER
GROMACS
LAMMPS
NAMD
Chroma
BerkeleyGW
FUN3D
RTM
SPECFEM3D

NERSC System Roadmap



2013

NERSC-7: Edison Multicore CPU

2016

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

2020

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

2024

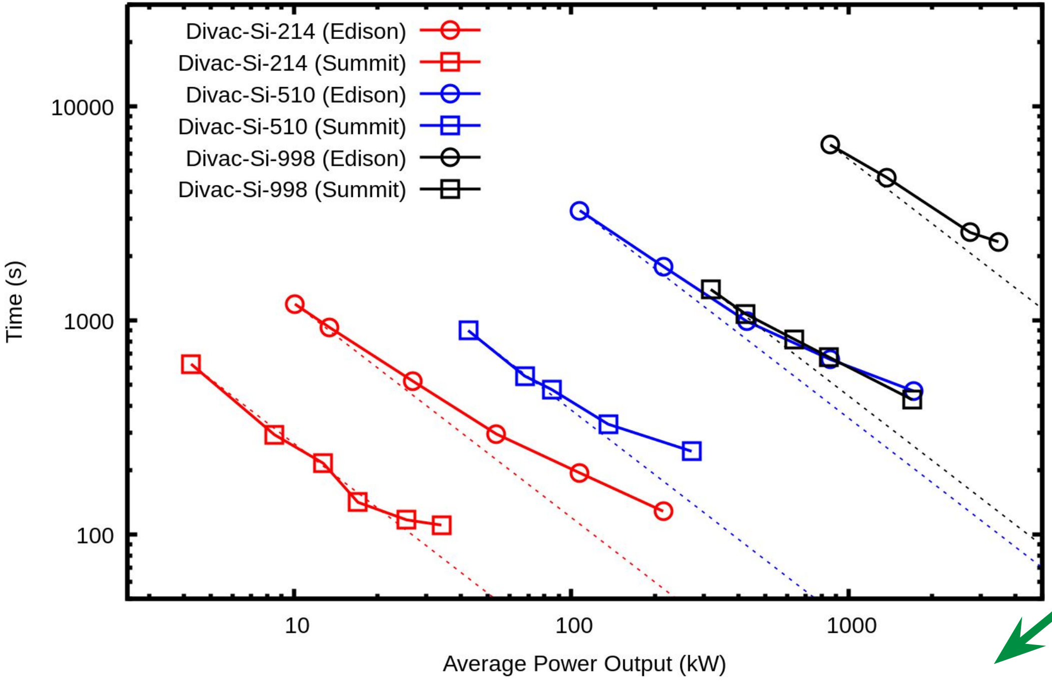
NERSC-10: Exa system

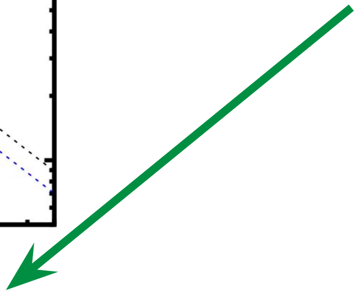
2028

NERSC-11: Beyond Moore

Increasing need for energy-efficient architectures

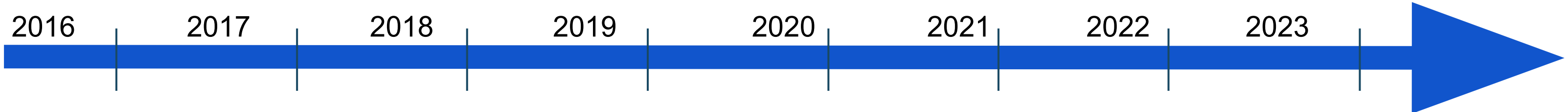
Why GPUs

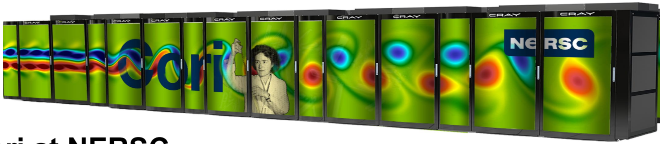




Improving Energy Efficiency

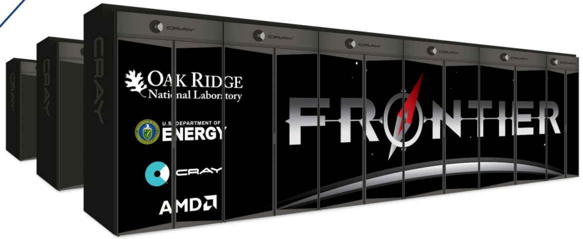
DOE HPC Roadmap - GPUs





Cori at NERSC

Summit at OLCF (NVIDIA Volta)



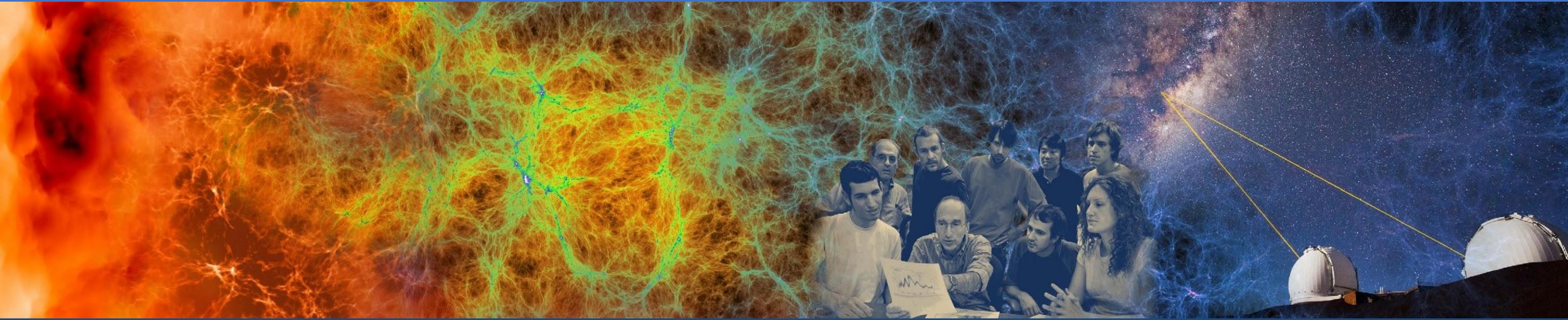
NVIDIA Ampere

Intel GPUs

NVIDIA Volta GPUs

AMD GPUs

Application Readiness for Perlmutter Overview



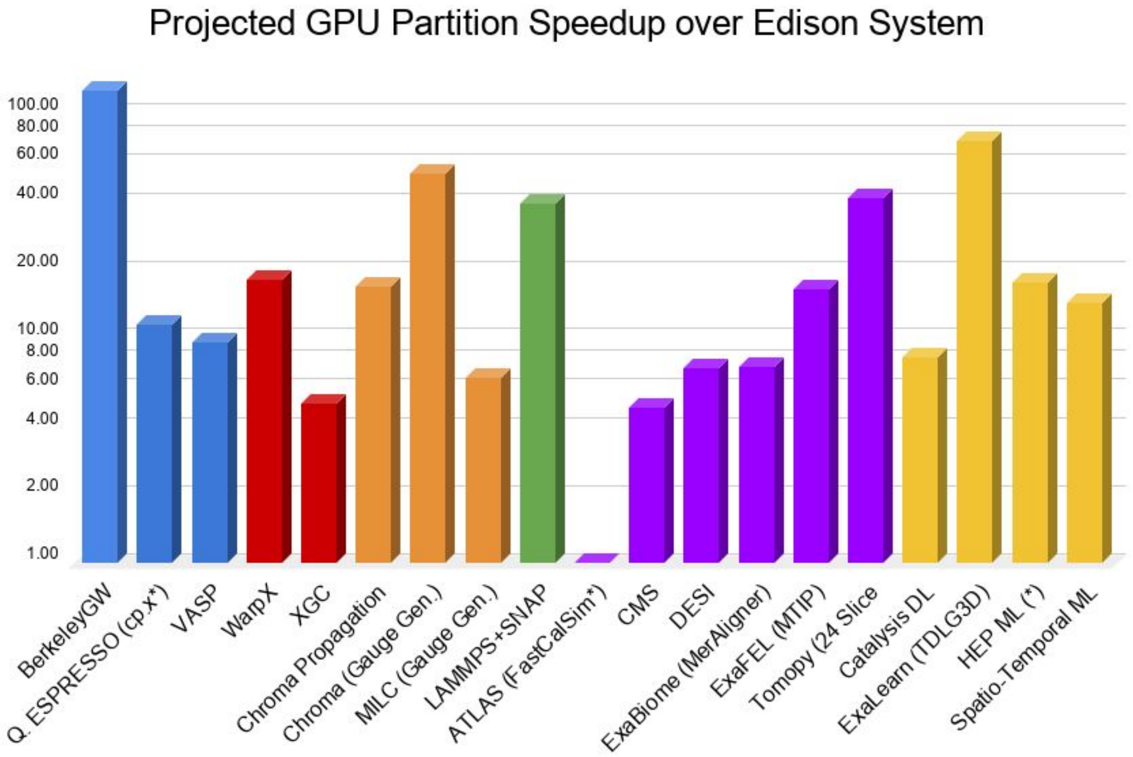
Our Common Challenge

How to enable NERSC’s diverse community of 7,000 users, 800 projects, and 700 codes to run on advanced architectures like Perlmutter and beyond?

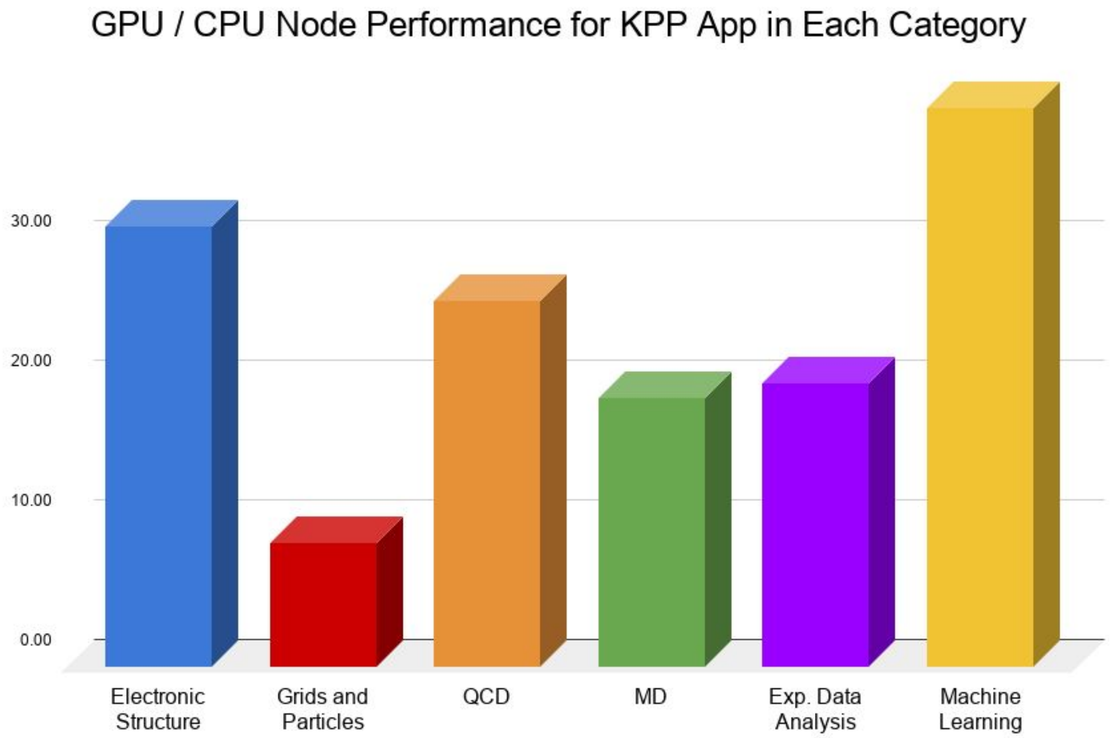
Application Readiness Strategy for Perlmutter

- Vendor engagements
  - hack-a-thons with HPE/Cray, NVIDIA
  - NRE investment (OpenMP)
- Partnership with key code teams (NESAP)
  - ~25 projects spanning science domains
- Postdoctoral program
  - ~15 fellows focused on performance
- **Community engagement**
  - training events, tutorials, public hack-a-thons
  - publications in journals and conferences
  - **Cori GPU Node -** https://docs-dev.nersc.gov/cgpu/

Early NESAP Progress



Early NESAP Progress



Hackathons

NESAP Cray COE Hackathons

- 4 Per Year. ~3 NESAP
- 1-2 Cray, NERSC, NVIDIA mentors per team.
- 12 Week (½ day per week) Virtual Working group.

Community Hackathons:

- https://gpuhackathons.org/
- Open to applications from anyone
- 2-3 NVIDIA, NERSC, ORNL, Community mentors per team.
- 1 + 3 Day Virtual Events during Pandemic

Roofline for Performance Analysis

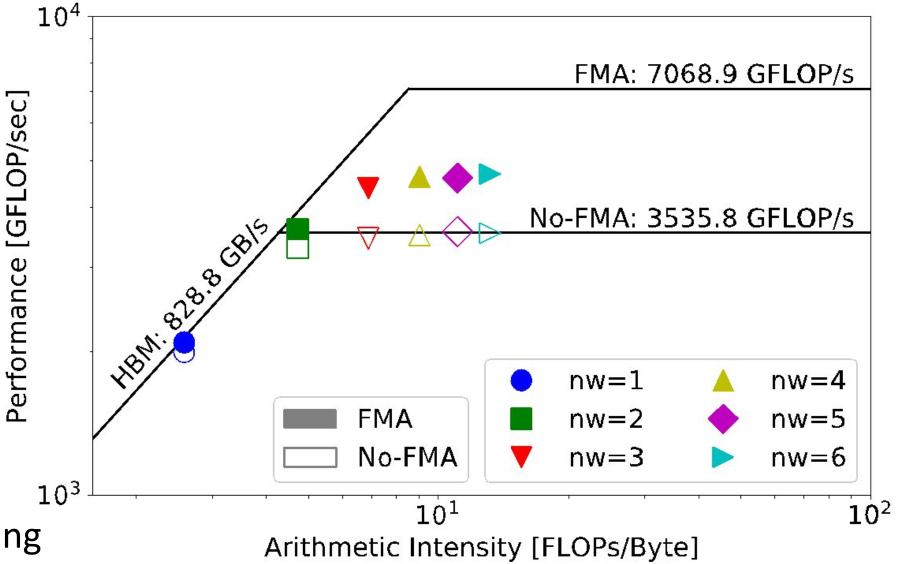
Users Want to Know:

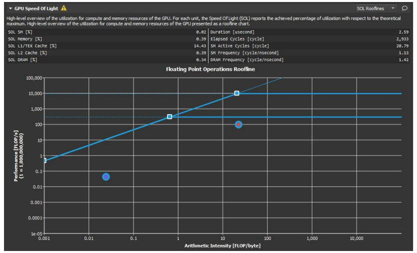
- What part of my code should I move to GPU?
- How do you know what HW features to target: HBM, Latency Hiding, Shared Mem, Packed Warps...
- How do you know how your code performs in an absolute sense and when to stop?

Progress Towards Roofline on GPUs:

Worked with NVIDIA to ensure NVProf/NSight can collect all required metrics including data motion from multiple levels:
L1/Shared, L2, DRAM, Host DRAM *etc*.

Roofline is now Integrated in NVIDIAs NSIGHT tool

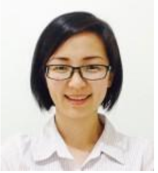




High-level overview of the utilization for compute and memory resources of the GPU. For each unit, the Speed Of Light (SOL) reports the achieved percentage of utilization with respect to the theoretical maximum. High-level overview of the utilization for compute and memory resources of the GPU presented as a roofline chart.

SOL SM [%]	0.02	Duration [usecond]	2.59
SOL Memory [%]	0.39	Elapsed Cycles [cycle]	2,933
SOL L1/TEX Cache [%]	14.43	SM Active Cycles [cycle]	20.79
SOL L2 Cache [%]	0.39	SM Frequency [cycle/nsecond]	1.13
SOL DRAM [%]	0.34	DRAM Frequency [cycle/nsecond]	1.42

Floating Point Operations Roofline
100,000
10,000
1,000
100
10
1
0.1
0.01
0.001
0.0001
1e-05
Performance [FLOP/s] (1 = 1,000,000,000)
0.001
0.01
0.1
1
10
100
1,000
10,000
Arithmetic Intensity [FLOP/byte]



Charlene Yang
Leading

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 ...

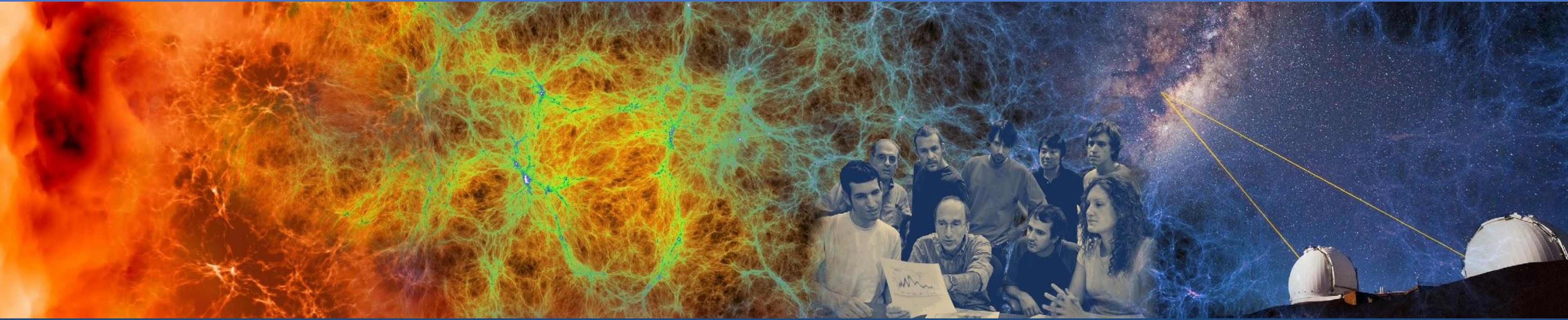
Raja: SW4

OpenMP NRE

Goal: Enable Directives Porting strategy from Cori to Perlmutter

- **Agreed on the subset of OpenMP target offload features to be included in the PGI compiler**
- **Created an OpenMP test suite containing micro-benchmarks, mini-apps, and the ECP SOLLVE V&V suite to evaluate correctness and performance**
- **Selected 5 NESAP application teams to partner with NVIDIA/PGI to add OpenMP target offload directives to the applications**

NESAP Success Stories

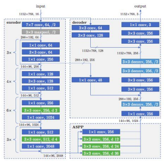


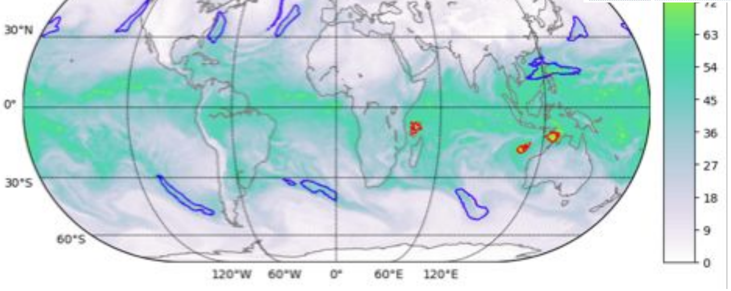
Applications at Scale w/ NESAP Expertise

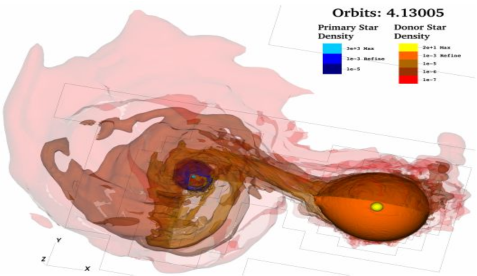
Exascale DL for Climate

Segmentation

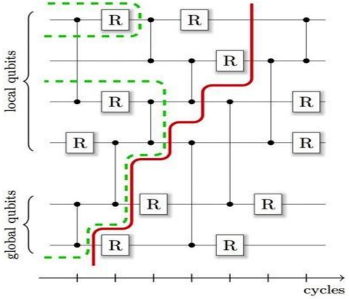
SC18 Gordon Bell Prize: arXiv:1810.01993



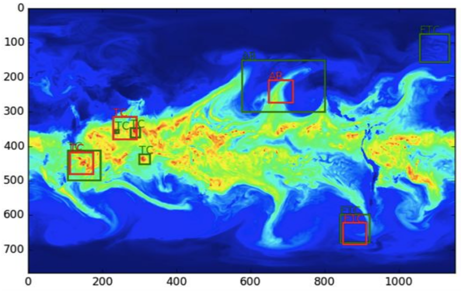


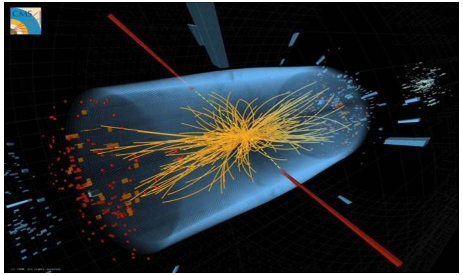


Stellar Merger Simulations with Task Based Programming



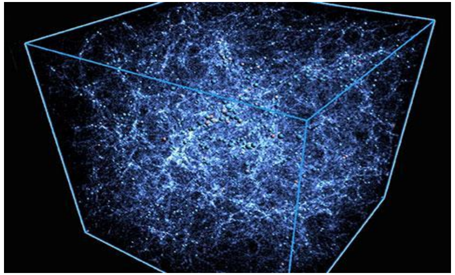
Largest Ever Quantum Circuit Simulation





Deep Learning at 15PF (SP) for Climate and HEP on Cori

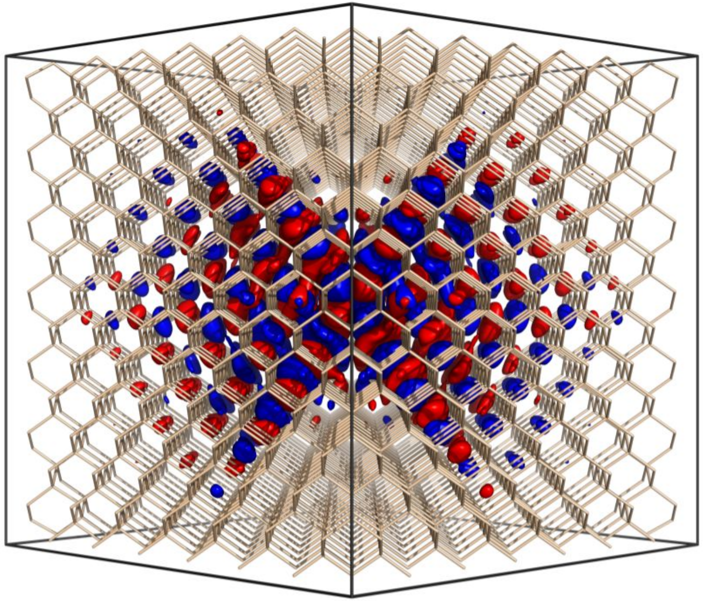
Celeste: 1st Julia app to achieve 1 PF



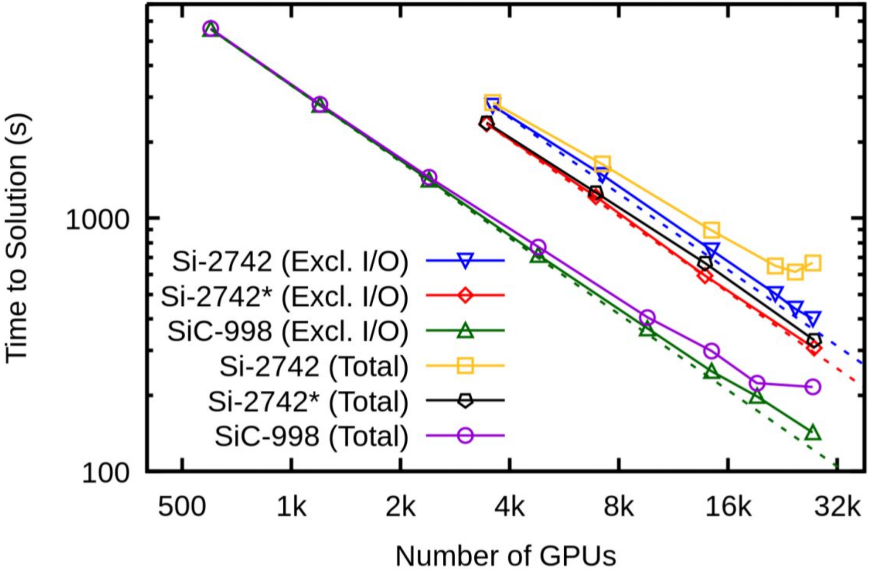
Galactos: Solved 3-pt correlation analysis for Cosmology @9.8PF

A NESAP App. is 2020 Gordon Bell Finalist

The BerkeleyGW NESAP team completed the largest ever excited state calculations using ~30k GPUs, achieving over 100 PFLOPs of peak performance on Summit. Excited to use Perlmutter's A100 GPUs.



Localized Defect state in a Semiconductor of Relevance to Qubits.



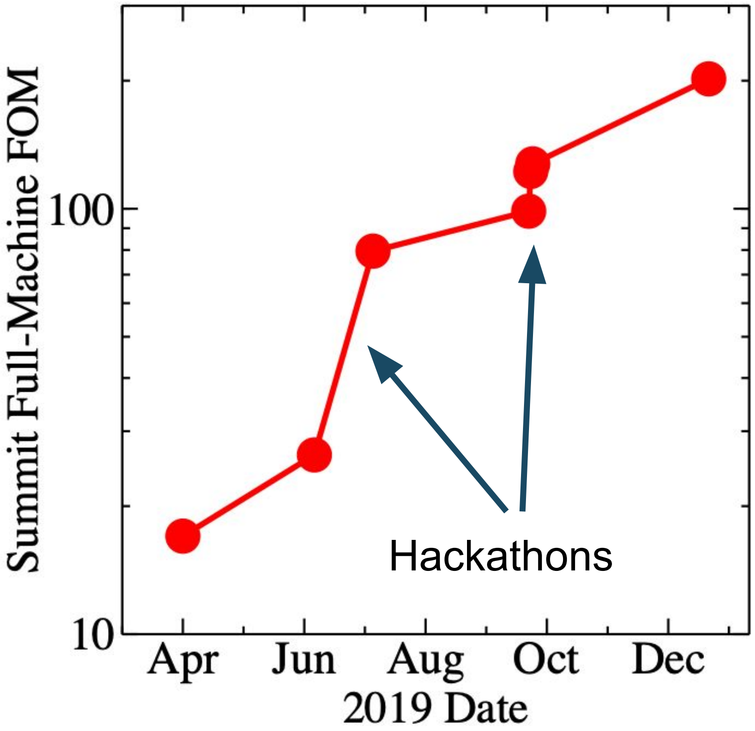
LAMMPs NESAP Effort

- LAMMPs is part of the ECP EXAALT project. Working with NERSC on acceleration for Perlmutter as a Waypoint for Exascale systems.

- Bottleneck is calculation of Forces/Potentials on atoms.

- Team made a tremendous amount of progress by developing a Mini-App - TestSNAP, for use at hackathons.

- Team has both a Kokkos and OpenMP 4.5 implementation of TestSNAP. Kokkos is used in production.



Lead by NESAP
Staff Rahul Gayatri

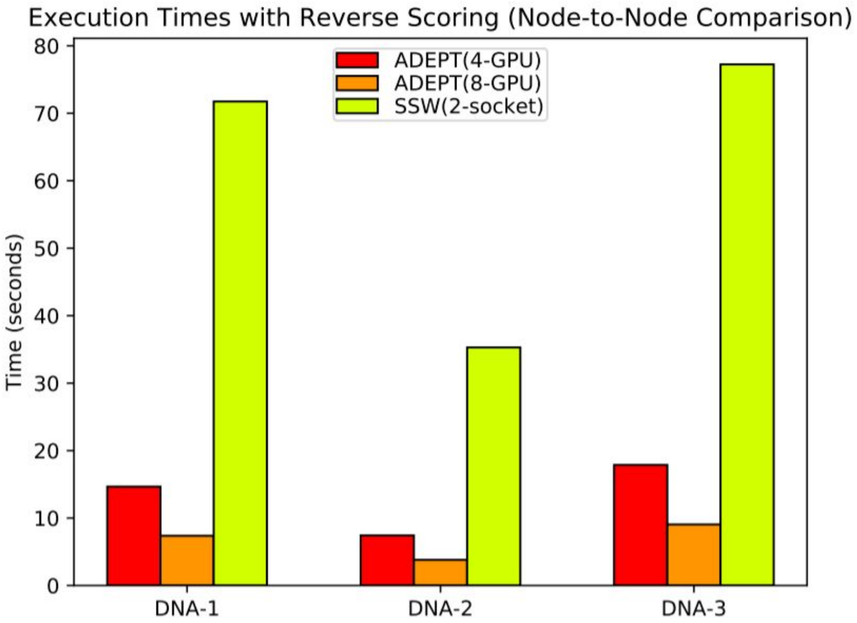
Smith-Waterman Alignment on GPUs

- Bio-informatics can sometimes be a challenging space for GPU performance.

- NESAP team developed novel Smith-Waterman alignment algorithm for multiple GPUs.

- Fastest ever GPU node implementation for DNA and Protein alignment.

- Performance
  4 V100 Node > 5x Cori HSW Node



Lead by NESAP Staff Muaaz Awan

We are Looking Forward to Seeing You at a Future Event!

NERSC

nersc.gov/users/training
gpuhackathons.org

Tomopy

Benchmark problem is a SIRT Tomographic reconstruction with 100 iterations. Each 2D slice was 2048 x 2048 pixels and the number of projection angles was 1501.

Required Porting: New Algorithm targeting GPUs

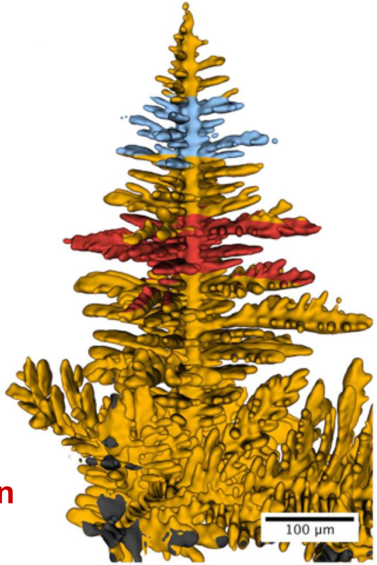
Baseline 24 slice reconstruction time (**Edison**)

walltime	28252.003

GPU 24 slice reconstruction time (**4 V100s**)

walltime	278.872

Optimization by NESAP PostDOC Jonathan Madsen



Implementation of New GPU Algorithm