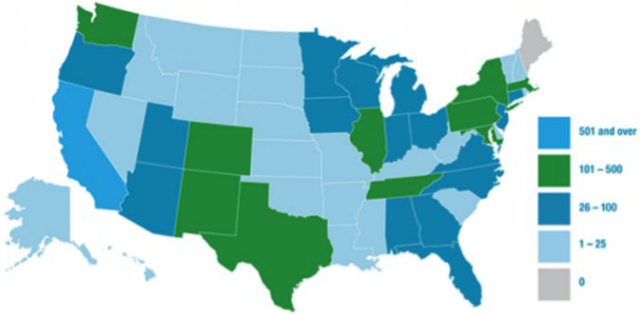
Why GPUs?

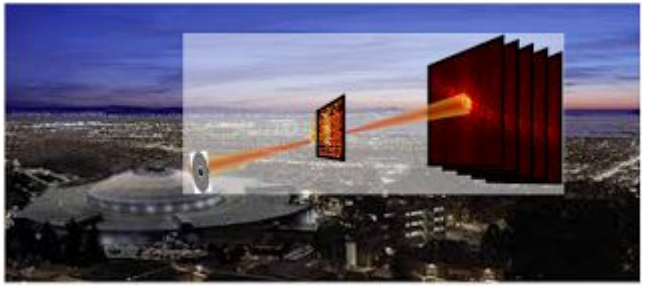
Jack Deslippe

Application Performance Lead
NERSC

NERSC has a dual mission to advance science and the state-of-the-art in supercomputing

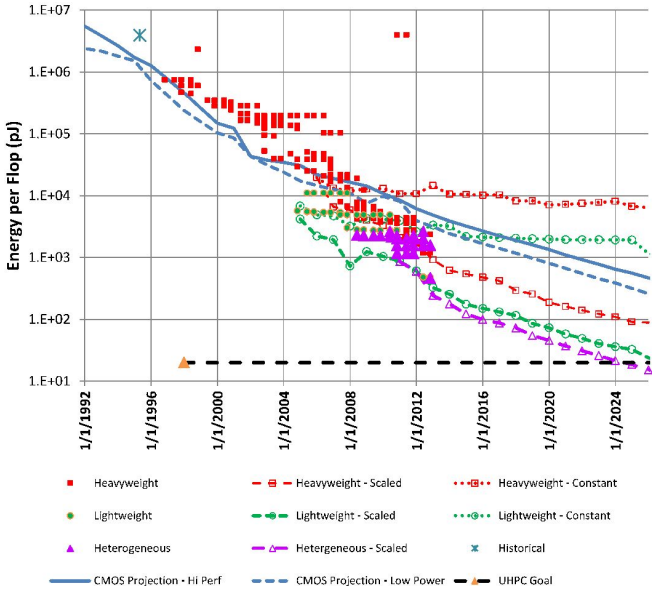
- 7,000 Users; 800 Projects; 700 Codes; 2000 NERSC citations per year
- We collaborate with computer companies years before a system's delivery to deploy advanced systems with new capabilities at large scale
- We provide a highly customized software and programming environment for science applications
- We are tightly coupled with the workflows of DOE's experimental and observational facilities
- Our staff provide advanced application and system performance expertise to users

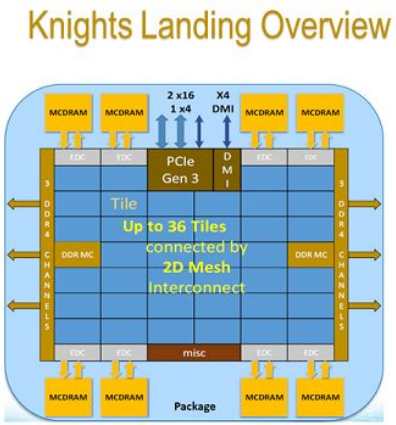




Change Has Arrived

Driven by power consumption and heat dissipation toward lightweight cores

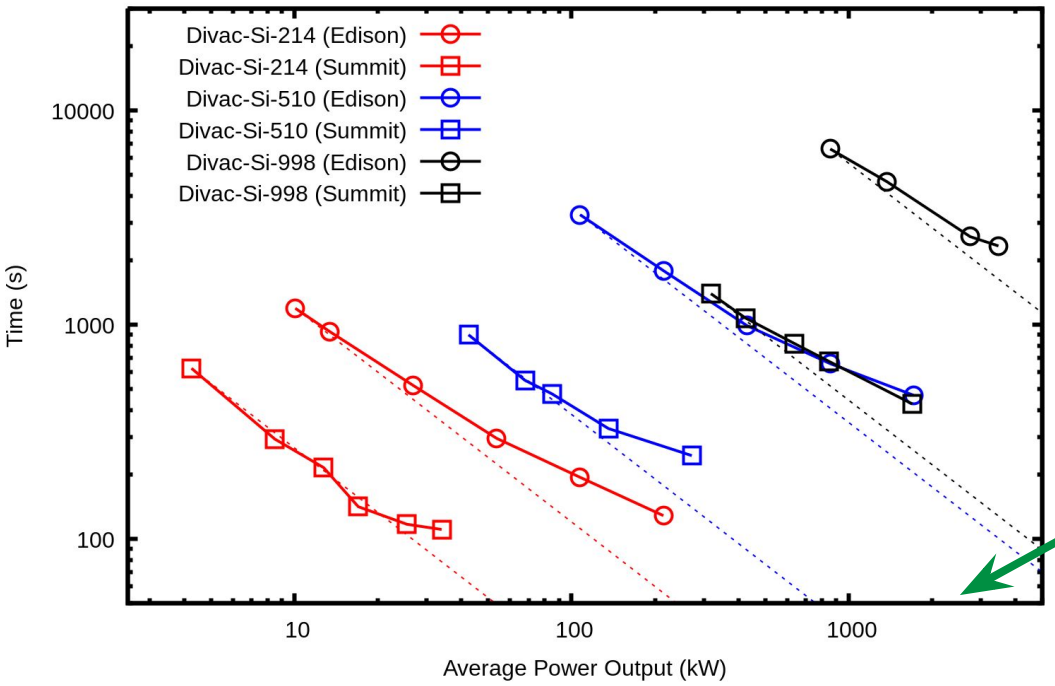




KNL: 215-230 W
2-socket Haswell: 270 W

Cori is a boon to science in the U.S. because of new capabilities, but the Intel Xeon Phi many-core architecture requires a code modernization effort to use efficiently.

Energy Efficiency Across Architectures



Improving Energy Efficiency

Circles: EDISON@NERSC CPU only

Squares: SUMMIT@OLCF CPU+GPU

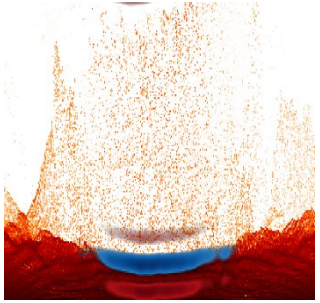
Where does increased performance come from?

On, KNL / GPUs getting peak performance comes from effectively using:

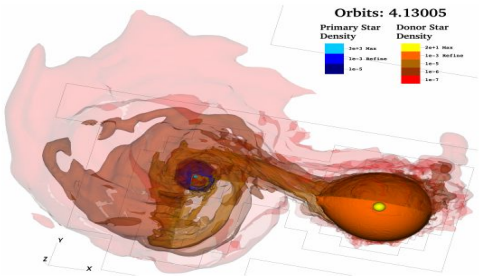
- ~~Increase Clock Speed~~
- O(100) Cores/SMs per chip with Many Hyperthreads/Warps to hide any latency
- 8-32 Double Precision wide Vectors
- Multiple FLOPS per vector lane using FMA, Tensors etc.

And... Cache, HBM, Memory Hierarchy that Allows Enough Bandwidth to Feed Compute Unites

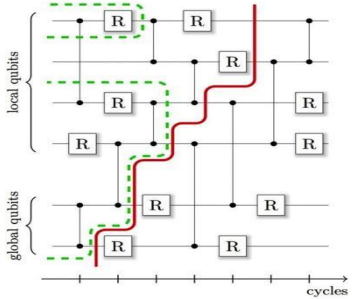
NERSC's Users Demonstrating Groundbreaking Science Capability on KNL



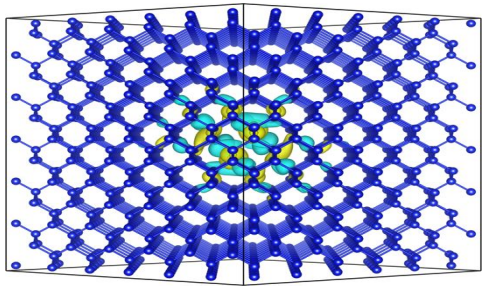
Large Scale Particle in Cell Plasma Simulations



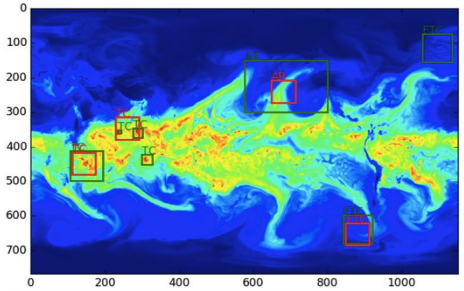
Stellar Merger Simulations with Task Based Programming

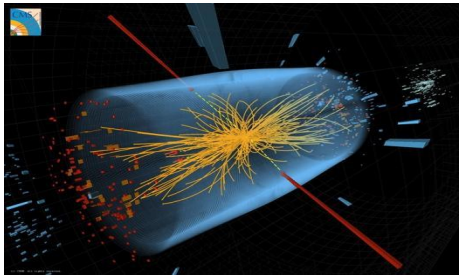


Largest Ever Quantum Circuit Simulation



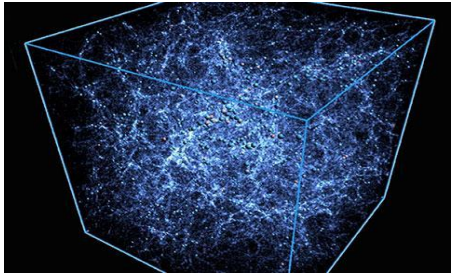
Largest Ever Defect Calculation from Many Body Perturbation Theory > 10PF





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

Celeste: 1st Julia app to achieve 1 PF



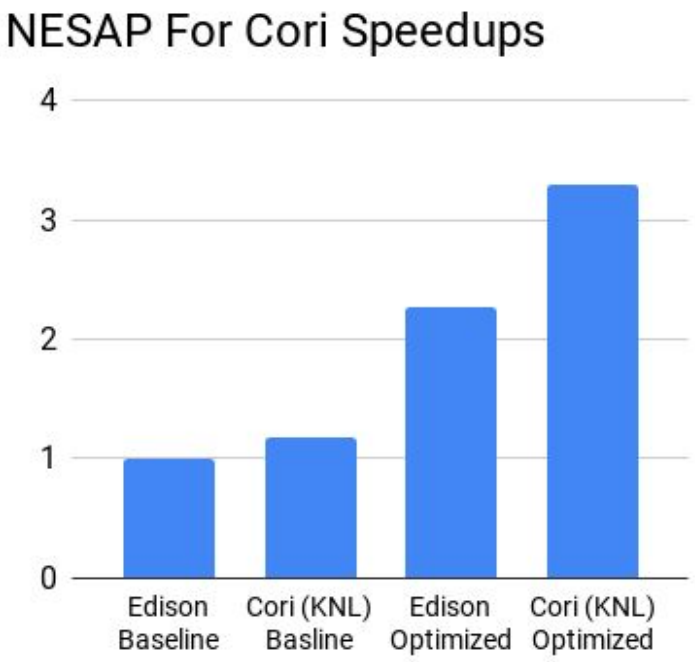
Galactos: Solved 3-pt correlation analysis for Cosmology @9.8PF

Modernizing Codes is Possible

NESAP Program Shows Investments in Codes Pays Off!

- ~25 Projects selected from competitive application process with reviews
- ~15 postdoctoral fellows
- Leverage vendor expertise and hack-a-thons
- **Optimize codes with improvements relevant to multiple architectures**

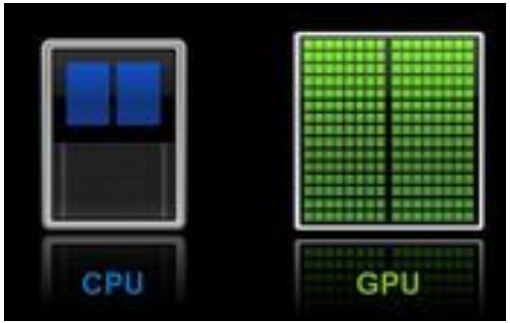
- **Lessons from NESAP Available to All In Venues like this!**



CPUs to GPUs

CPU (Haswell)

- 64 cores
- 2 threads each
- 2x256-bit vectors
- pipelined instructions
- double precision
  - ~2000 way parallelism (64*4*8)



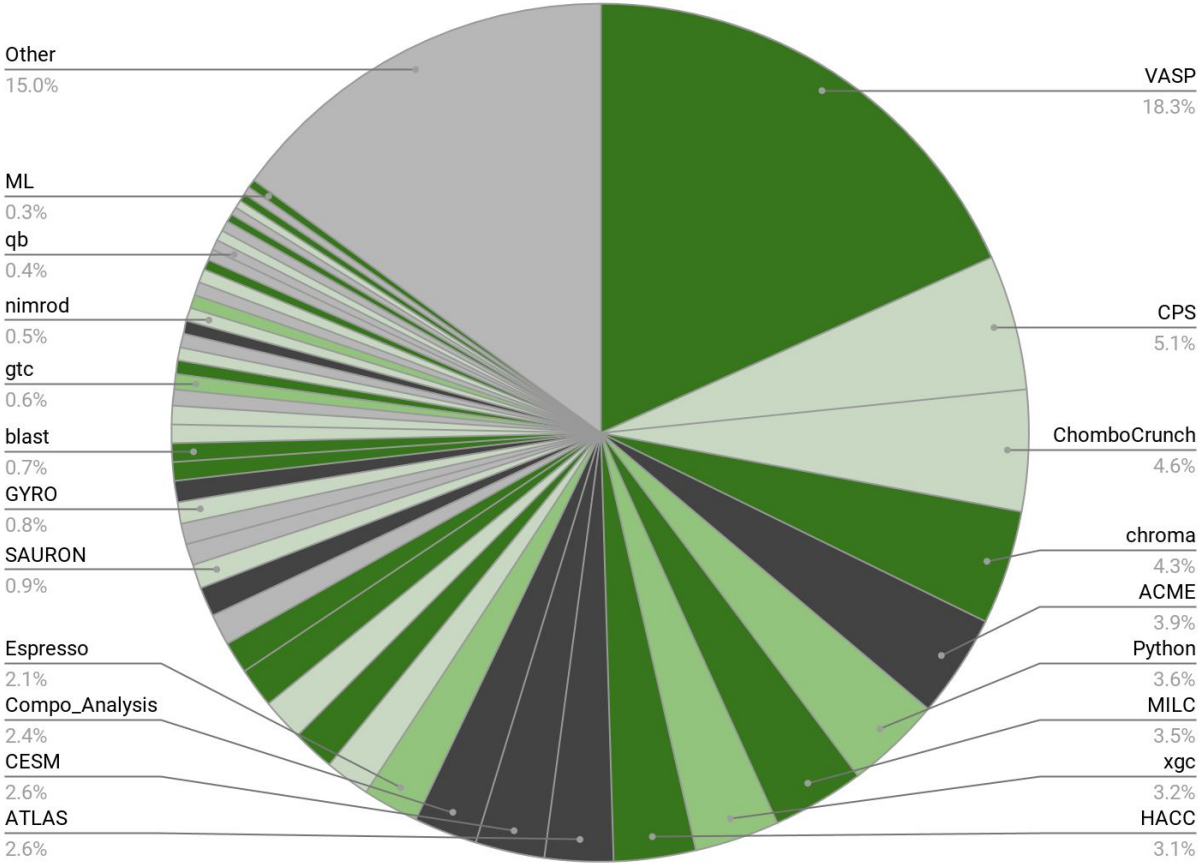
GPU (V100)

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

CPU - Speed

GPU - Throughput

GPU Readiness Among NERSC Codes (Aug'17 - Jul'18)



GPU Status & Description	Fraction
Enabled: Most features are ported and performant	37%
Kernels: Ports of some kernels have been documented.	10%
Proxy: Kernels in related codes have been ported	20%
Unlikely: A GPU port would require major effort.	13%
Unknown: GPU readiness cannot be assessed at this time.	20%

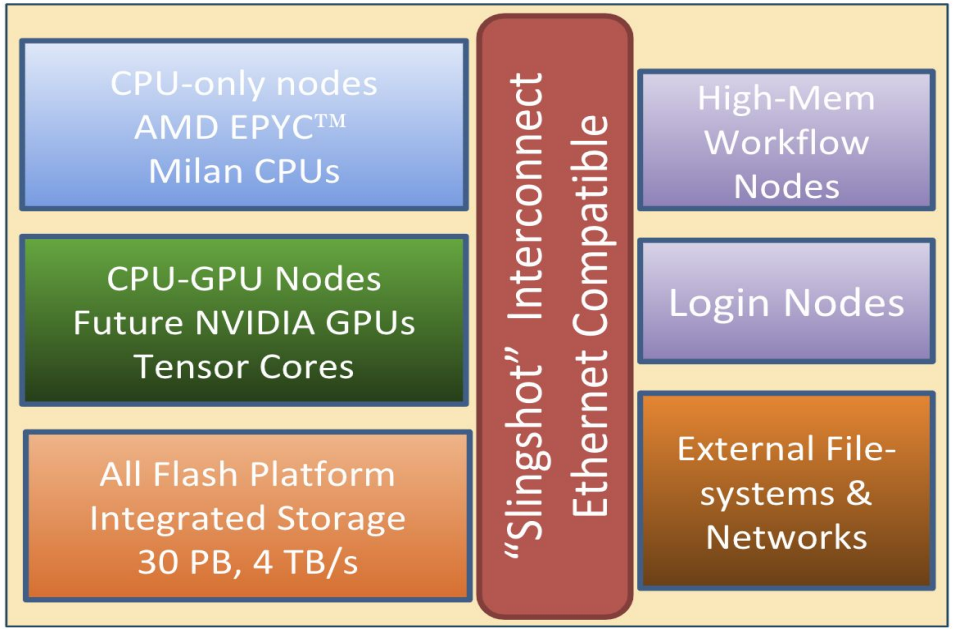
A number of applications in NERSC workload are GPU enabled already.

NERSC-9 will be named after Saul Perlmutter

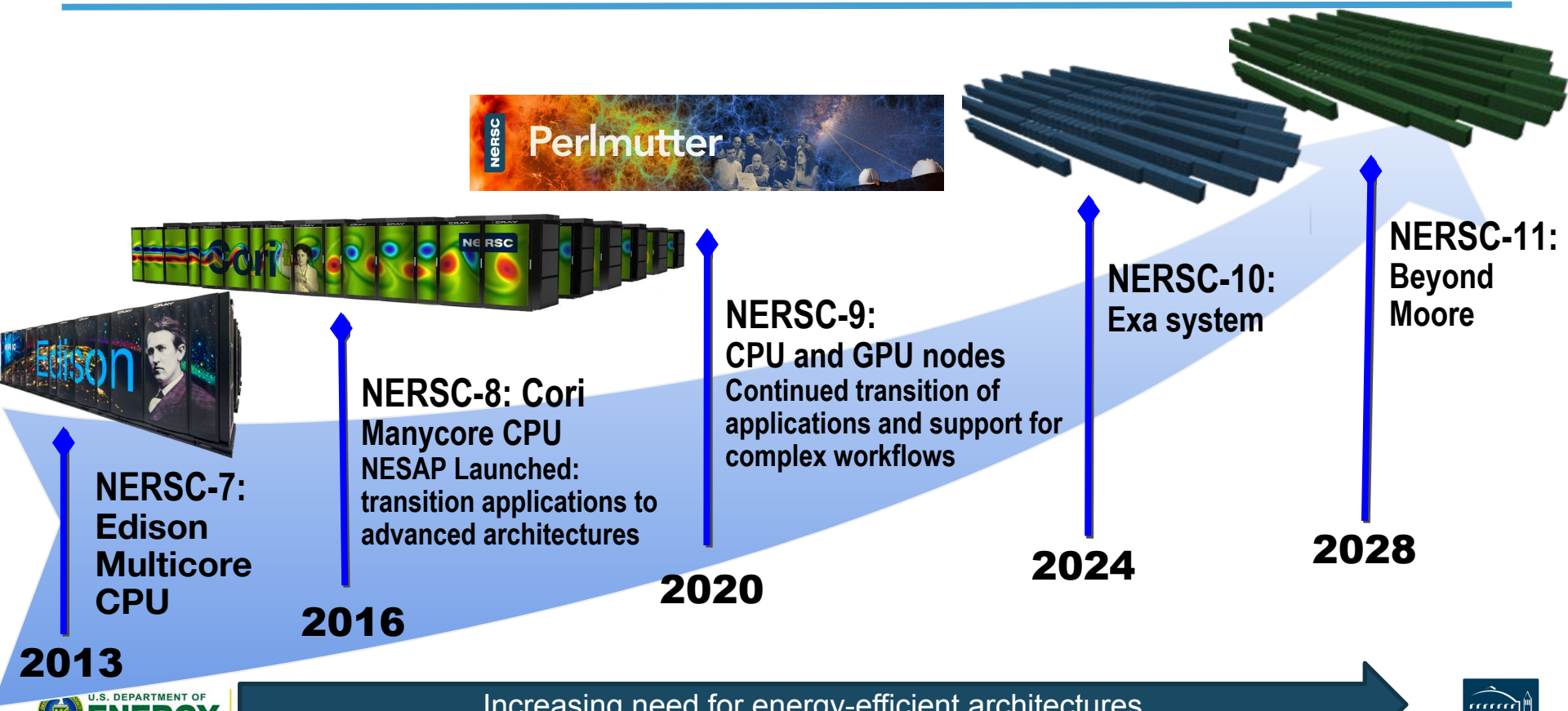
- Winner of 2011 Nobel Prize in Physics for discovery of the accelerating expansion of the universe.
- Supernova Cosmology Project, lead by Perlmutter, was a pioneer in using NERSC supercomputers combine large scale simulations with experimental data analysis
- Login “saul.nersc.gov”

Perlmutter: A System Optimized for Science

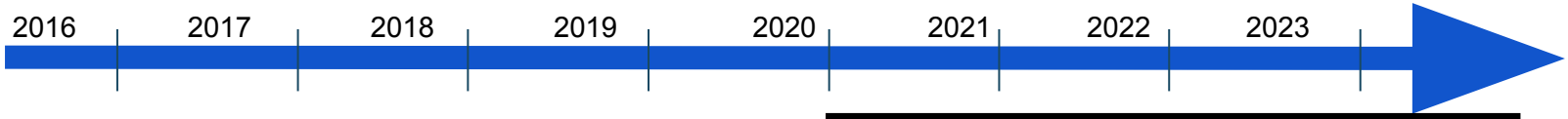
- GPU-accelerated and CPU-only nodes meet the needs of large scale simulation and data analysis from experimental facilities
- Cray "Slingshot" - High-performance, scalable, low-latency Ethernet-compatible network
- Single-tier All-Flash Lustre based HPC file system, >6x Cori's bandwidth
- Delivery in early FY21

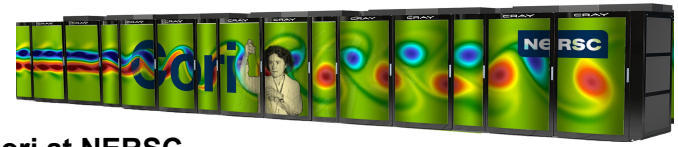


NERSC Systems Roadmap



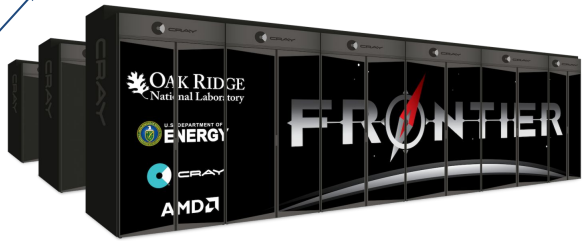
DOE HPC Roadmap





Cori at NERSC

Summit at OLCF (NVidia Volta)



NVIDIA Volta GPUs

NVIDIA GPUs

Intel GPUs

AMD GPUs

AMD CPU nodes

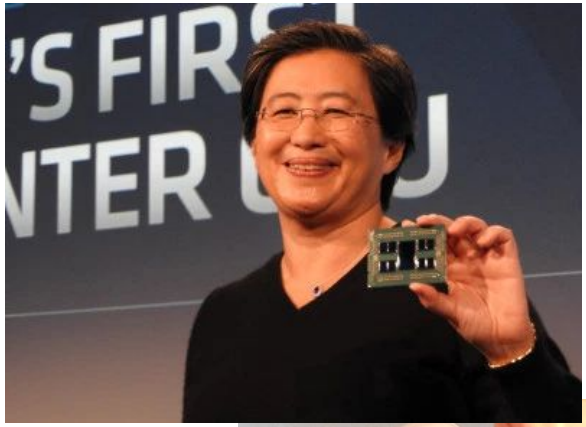
AMD “Milan” CPU

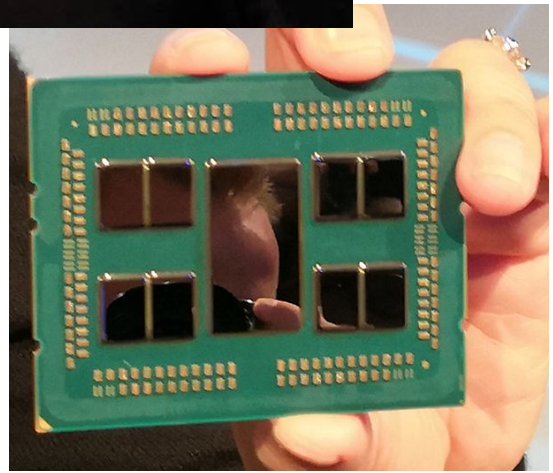
- ~64 cores
- “ZEN 3” cores - 7nm+
- AVX2 SIMD (256 bit)

>=Rome specs

8 channels DDR memory

~ 1x Cori





GPU nodes

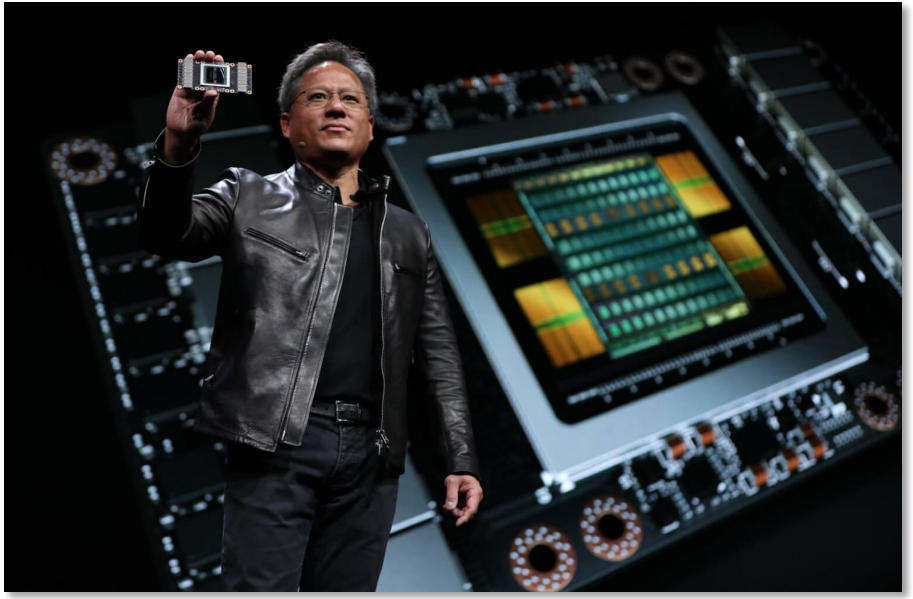
4x NVIDIA “Volta-next” GPU

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

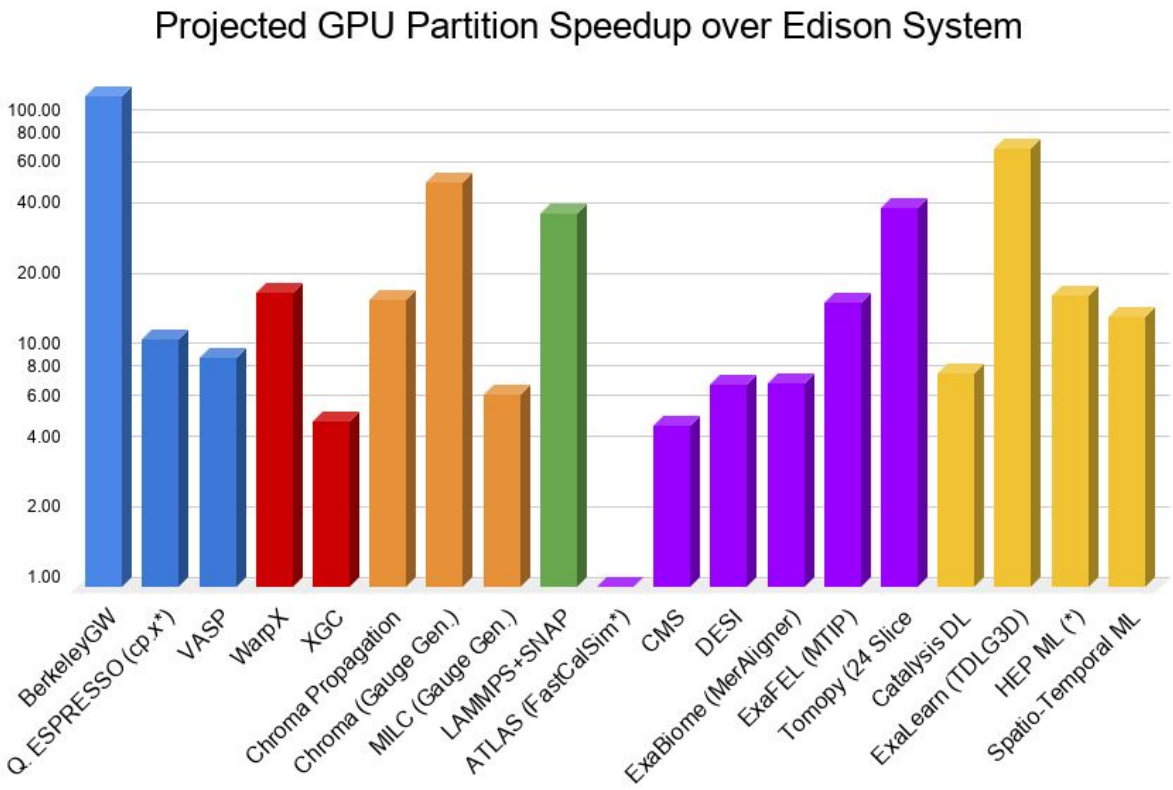
Volta specs

1x AMD CPU

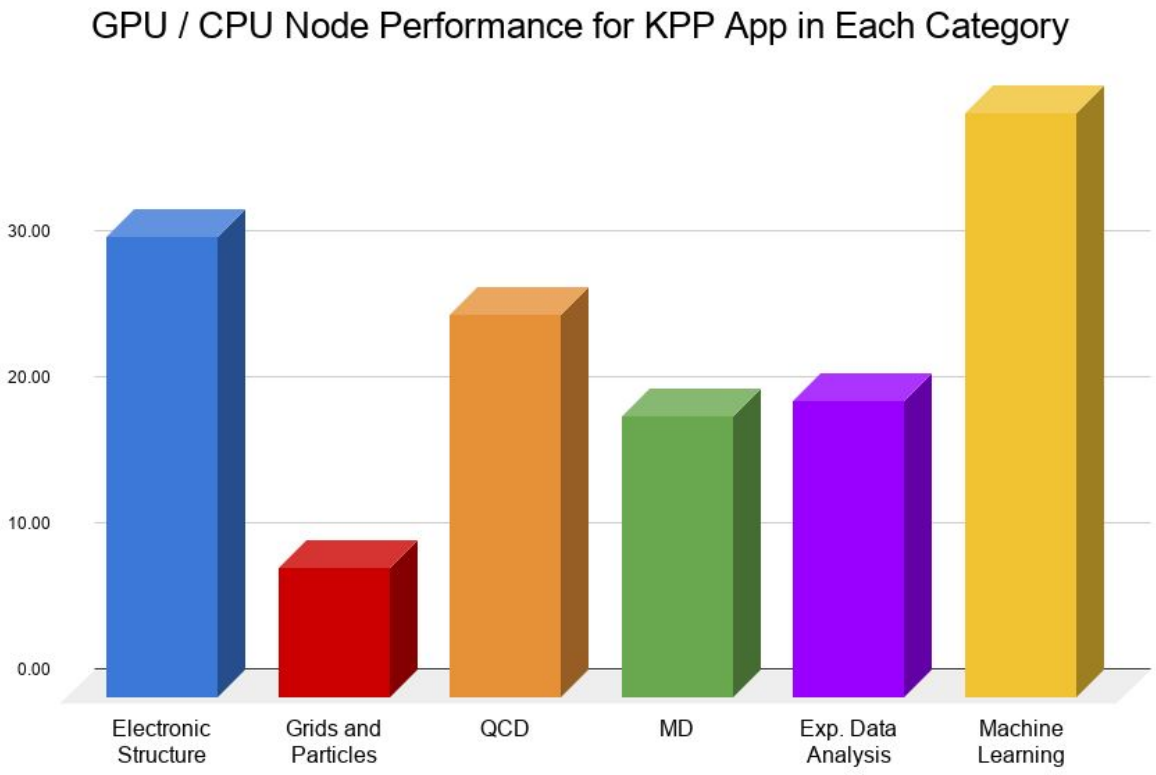
GPU direct



Early NESAP Progress



Early NESAP Progress



Example Early Progress (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.

FOM = 1 / (WallTime Per 2D Slice)

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

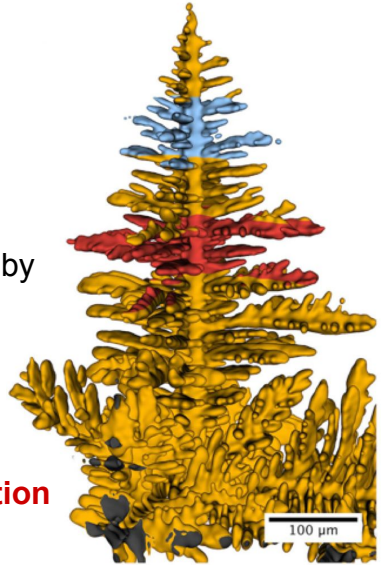
walltime	28252.003

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

walltime	278.872

Optimization by NESAP PostDOC Jonathan Madsen

Implementation of New GPU Algorithm



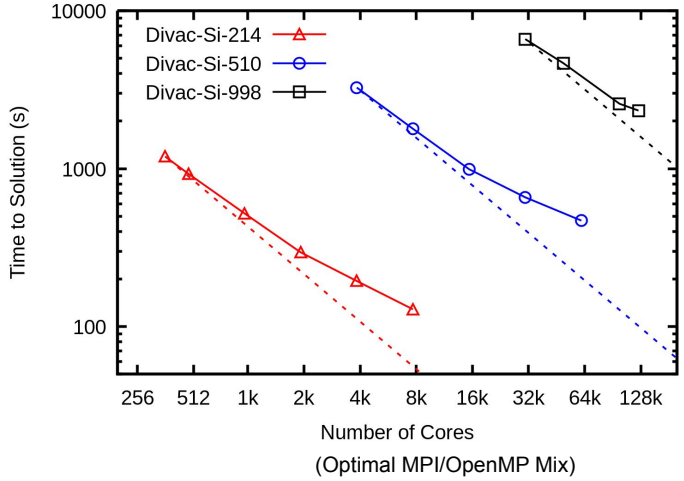
Example Early Progress (BGW Epsilon)

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.

FOM= 1 / (WallTime)

Port of Bottlenecks to CUDA Kernels and Math Libraries

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		

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 – Status & Future Plans

Items completed

- 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

Next Items

- Evaluate upcoming compiler releases

Optimization Strategy: Roofline on GPUs

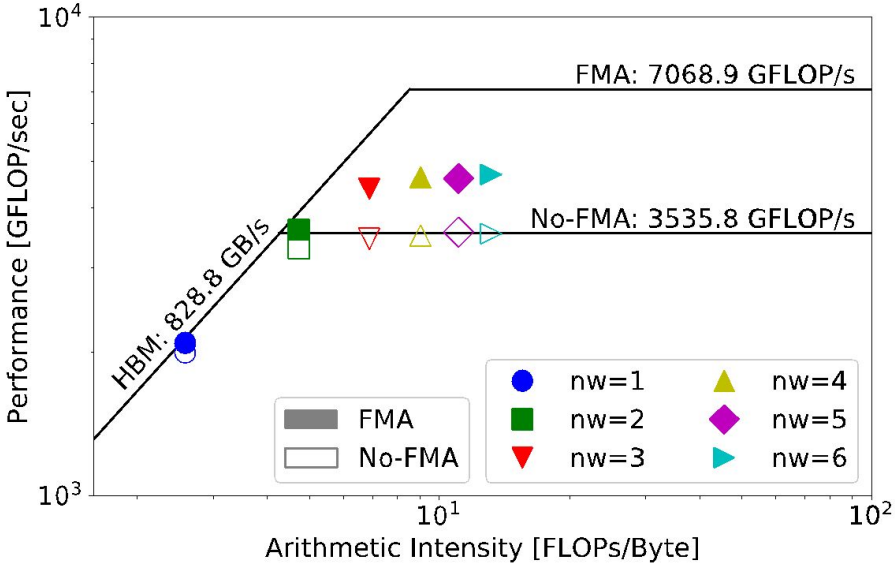
Users Want to Know:

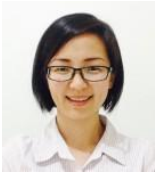
- What part of my code should I move to GPU?
- How do you know what HW features to target: HBM, Latency Hiding, Shared Mem, Packed Warps...
- How do you know how your code performs in an absolute sense and when to stop?

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*.

Hoping to automate roofline presentation



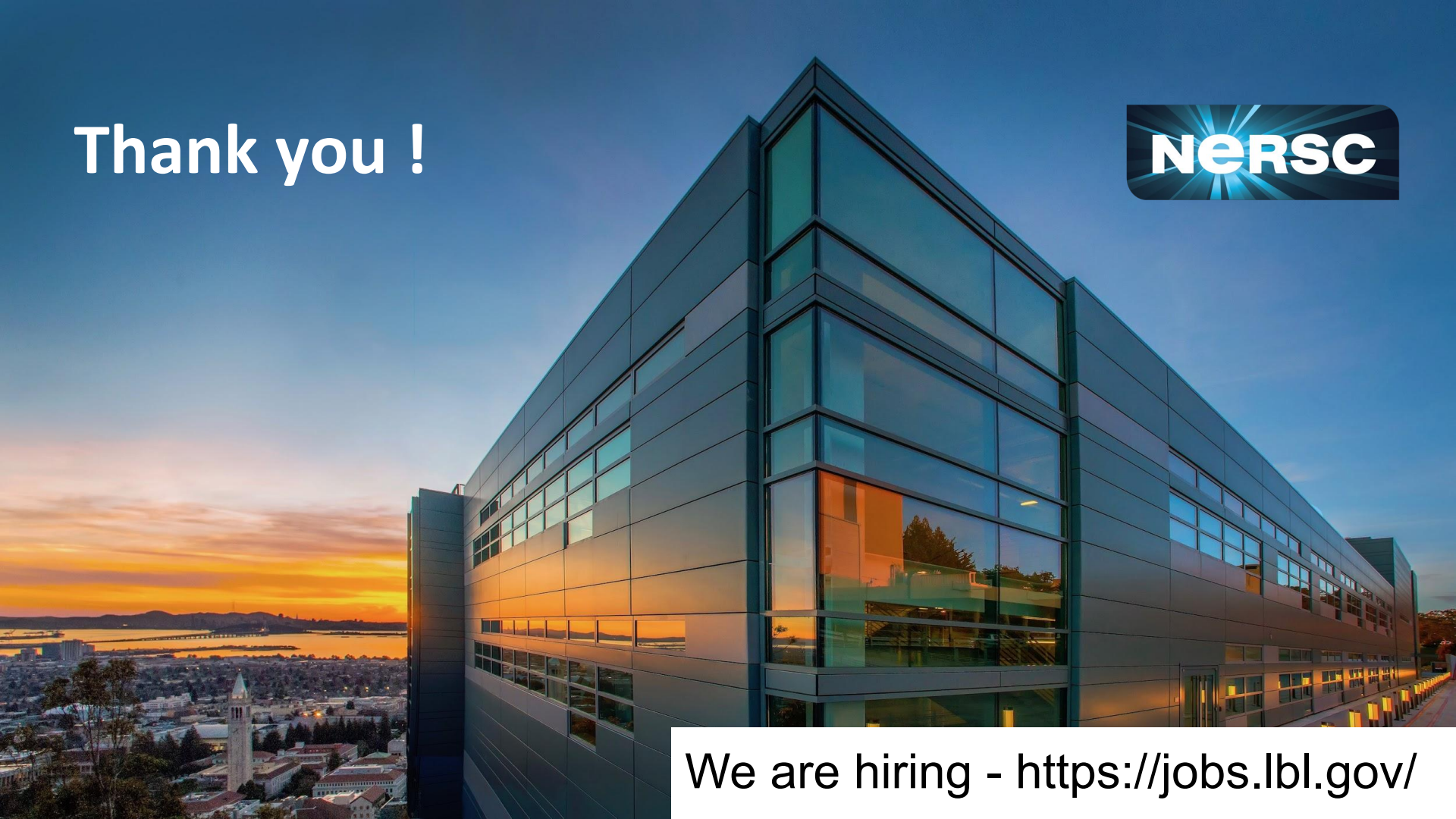


Charlene Yang
Leading

Perlmutter: A System Optimized for Science

- Cray Shasta System providing 3-4x capability of Cori system
- First NERSC system designed to meet needs of both large scale simulation and data analysis from experimental facilities
  - Includes both NVIDIA GPU-accelerated and AMD CPU-only nodes
  - Cray Slingshot high-performance network will support Terabit rate connections to system
  - Optimized data software stack enabling analytics and ML at scale
  - All-Flash filesystem for I/O acceleration
- Robust readiness program for simulation, data and learning applications and complex workflows
- Delivery in early FY 2021

Thank you !



We are hiring - https://jobs.lbl.gov/