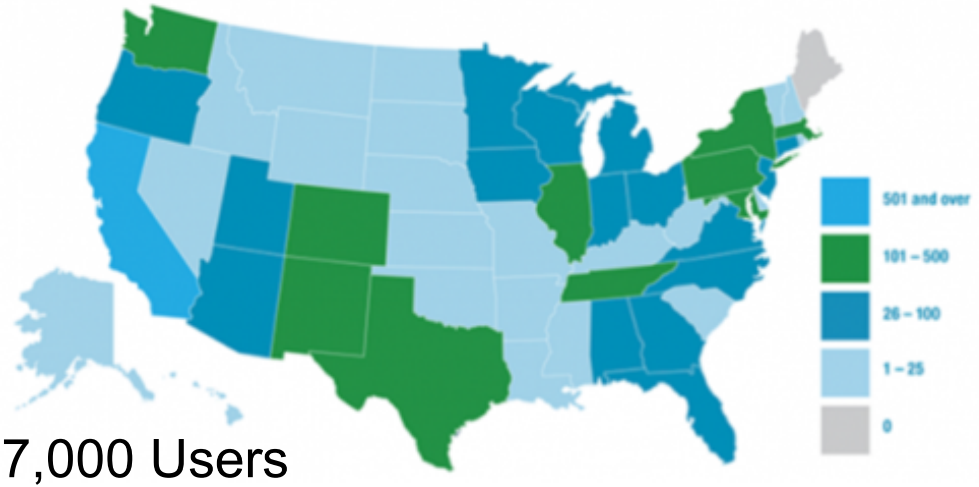
# Perlmutter - A 2020 Pre-Exascale GPU-accelerated System for NERSC - Architecture and Application Performance Optimization

Nicholas J. Wright
Perlmutter Chief Architect

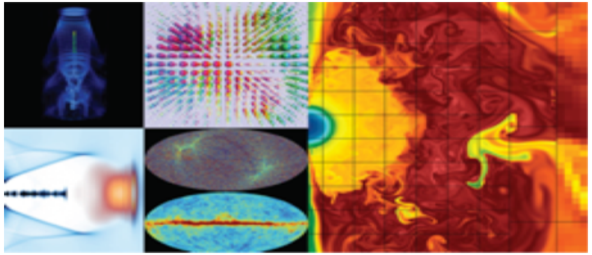
Sixth Workshop on Accelerator Programming Using Directives

18 November 2019

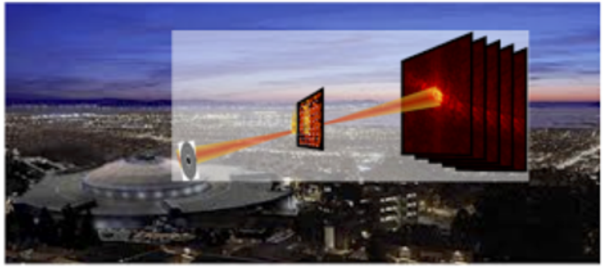
# NERSC is the mission High Performance Computing facility for the DOE SC



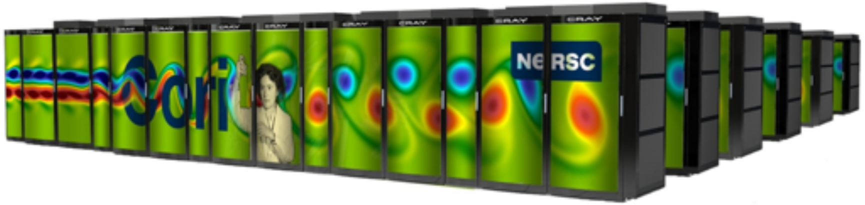
7,000 Users
800 Projects
700 Codes
2000 NERSC citations per year



Simulations at scale



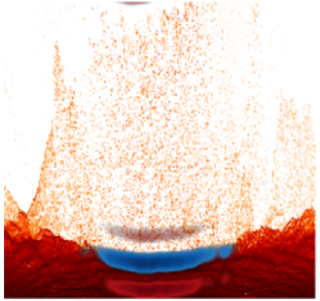
Data analysis support for DOE's experimental and observational facilities
Photo Credit: CAMERA



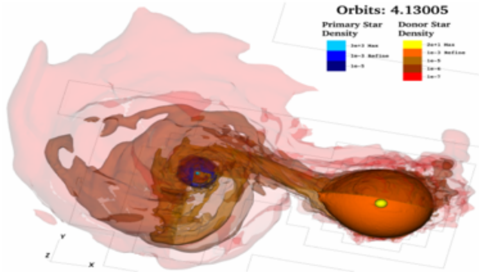
# NERSC has a dual mission to advance science and the state-of-the-art in supercomputing

- 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 – ingesting tens of terabytes of data each day
- Our staff provide advanced application and system performance expertise to users

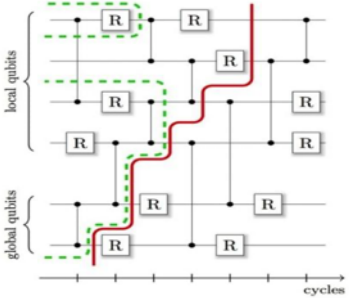
# NERSC's Users Demonstrate Groundbreaking Science Capability



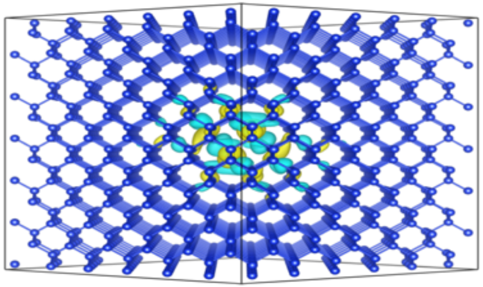
Large Scale Particle in Cell Plasma Simulations



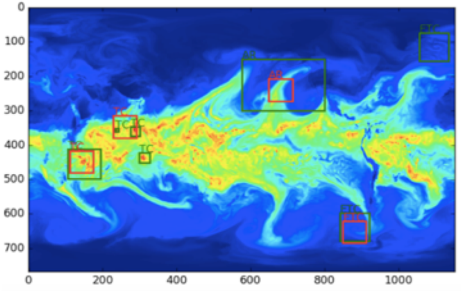
Stellar Merger Simulations with Task Based Programming

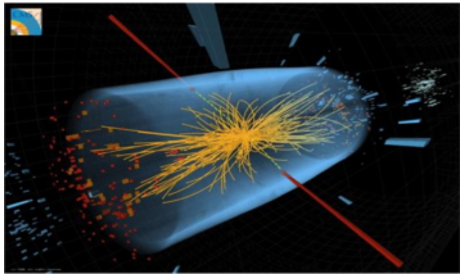


Largest Ever Quantum Circuit Simulation



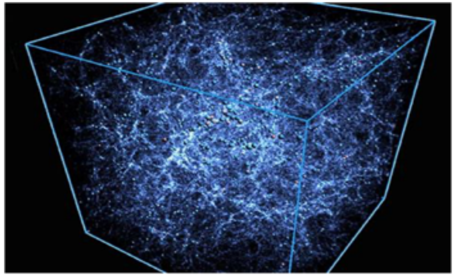
Largest Ever Defect Calculation from Many Body Perturbation Theory > 10PF





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

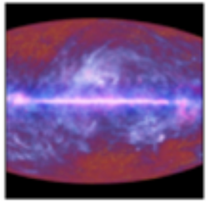
Celeste: 1st Julia app to achieve 1 PF



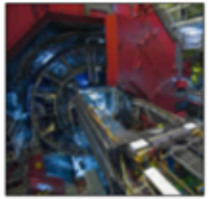
Galactos: Solved 3-pt correlation analysis for Cosmology @9.8PF

# NERSC also supports a large number of users and projects from DOE SC's experimental and observational facilities

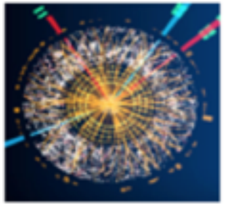
Palomar Transient Factory
Supernova



Planck Satellite
Cosmic Microwave Background Radiation

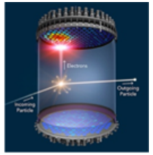


Alice
Large Hadron Collider

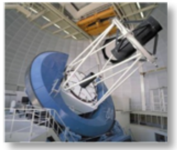


Atlas
Large Hadron Collider

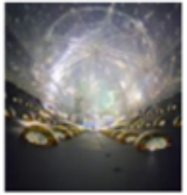
Star
Particle Physics



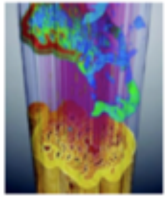
LZ



DESI



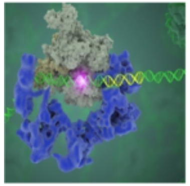
Dayabay
Neutrinos



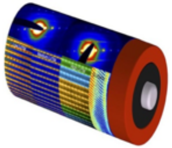
ALS
Light Source

LCLS
Light Source

Joint Genome Institute
Bioinformatics



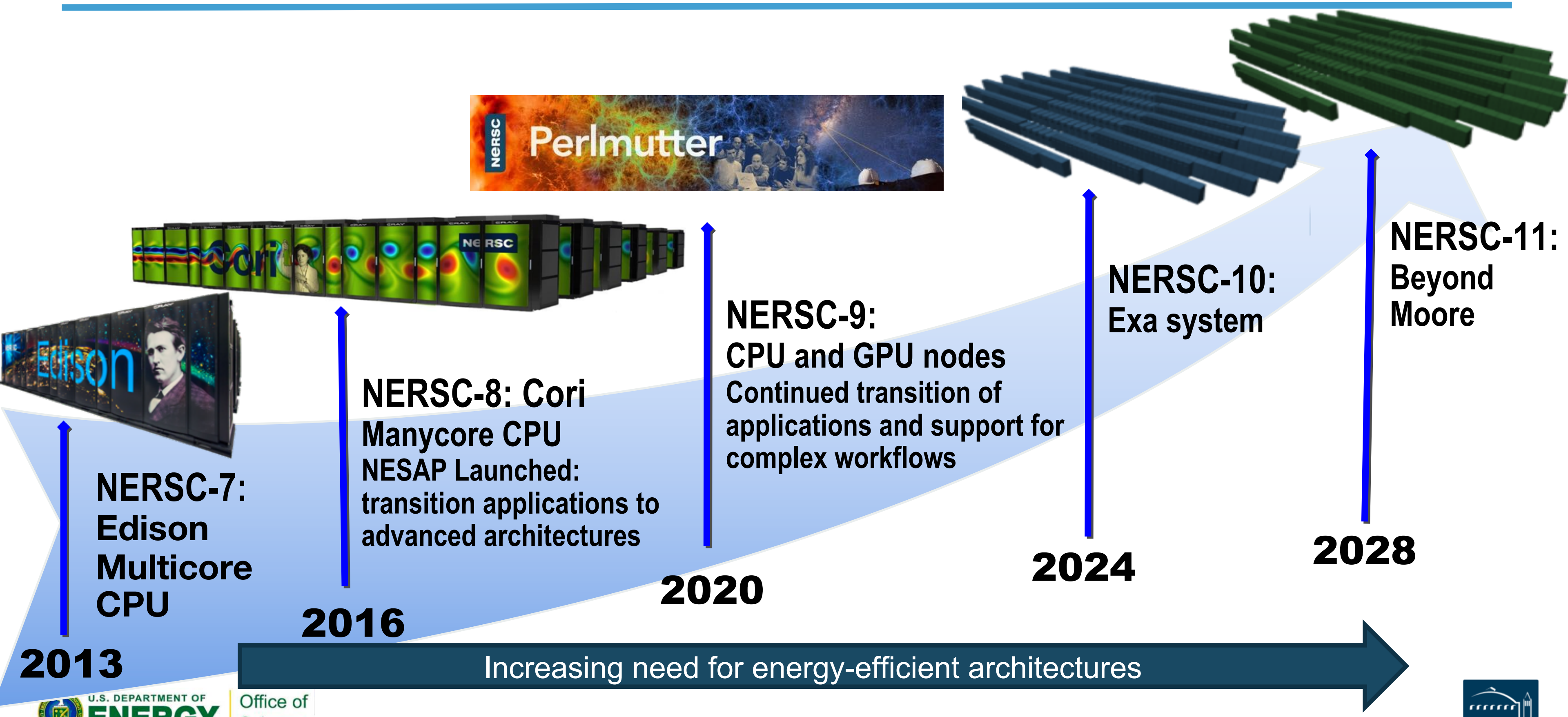
Cryo-EM



NCEM

LSST-DESC

# NERSC Systems 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

# Perlmutter is a Pre-Exascale System

## Pre-Exascale Systems

2013 | 2016 | 2018 | 2020

## Exascale Systems

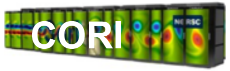
2021-2023

Argonne
IBM BG/Q

ORNL
Cray/NVidia K20

LLNL
IBM BG/Q

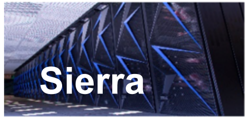
Argonne
Intel/Cray KNL



LBNL
Cray/Intel Xeon/KNL

LANL/SNL
Cray/Intel Xeon/KNL

ORNL
IBM/NVIDIA
P9/Volta



LLNL
IBM/NVIDIA
P9/Volta

LBNL
Cray/NVIDIA/AMD

CROSSROADS

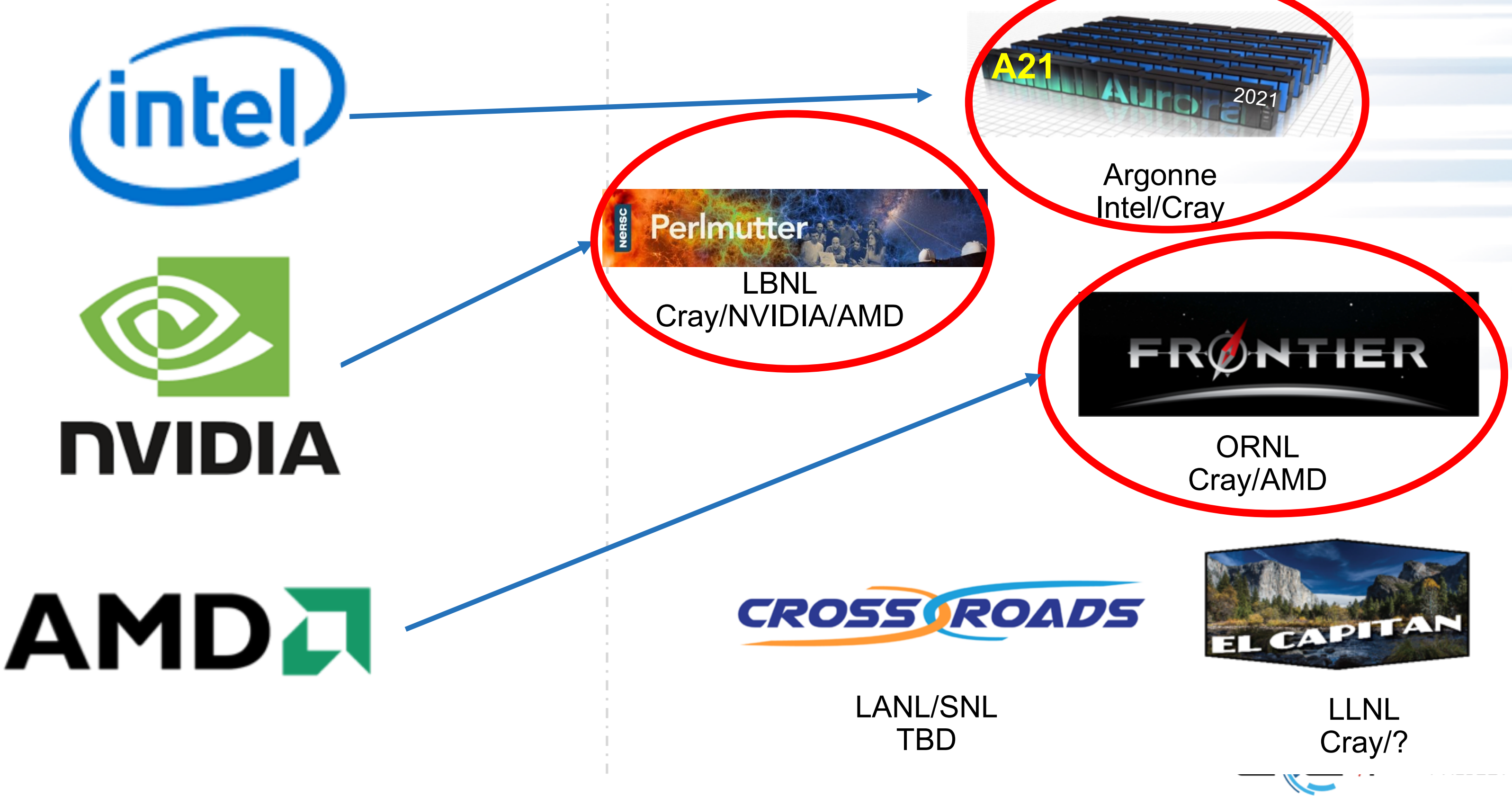
LANL/SNL
TBD

Argonne
Intel/Cray

ORNL
Cray/AMD

LLNL
Cray/?

| Pre-Exascale Systems | | | | Exascale Systems |
|---|---|---|---|---|
| 2013 | 2016 | 2018 | 2020 | 2021-2023 |



Argonne
Intel/Cray

LBNL
Cray/NVIDIA/AMD

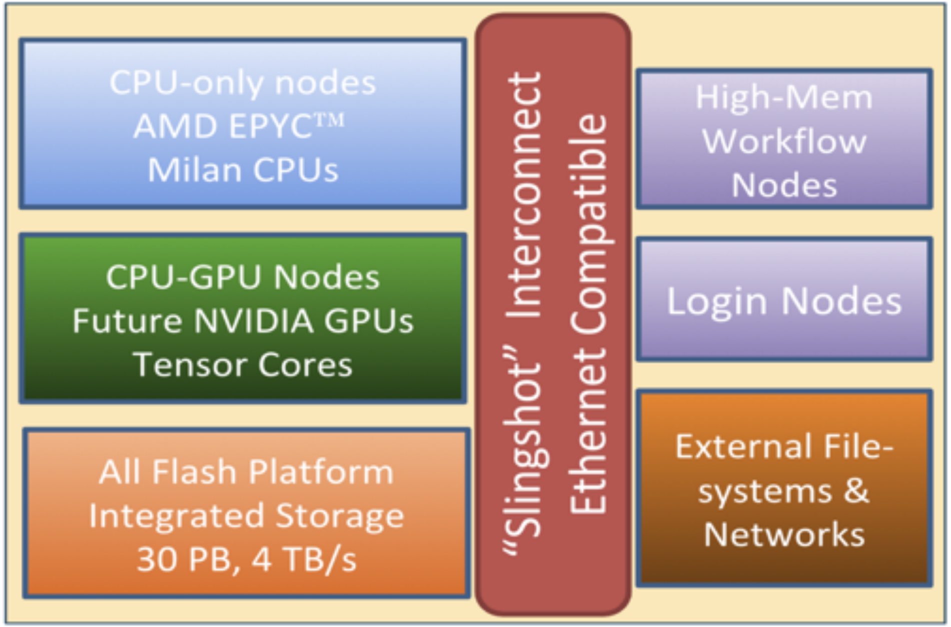
ORNL
Cray/AMD

LANL/SNL
TBD

LLNL
Cray/?

# 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
- Delivery in early FY21



# AMD CPU nodes

AMD "Milan" CPU

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

>=Rome specs

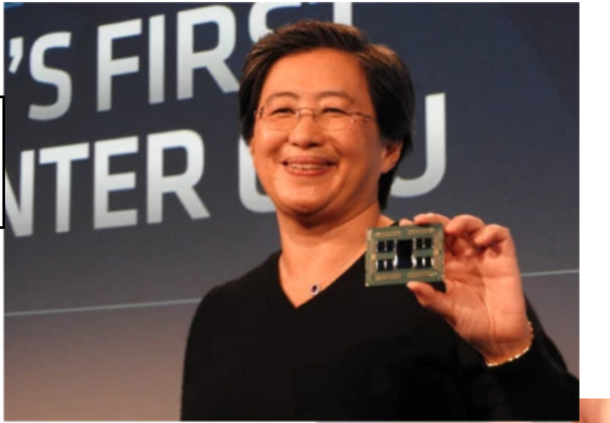
8 channels DDR memory

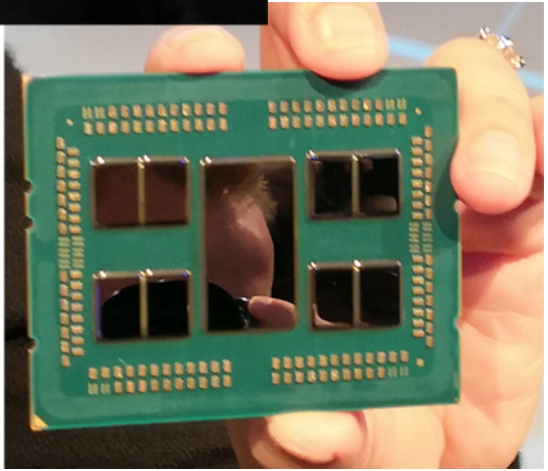
- >= 256 GiB total per node

1 Slingshot connection

- 1x25 GB/s

~ 1x Cori





# GPU nodes

4x NVIDIA “Volta-next” GPU

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

Volta specs

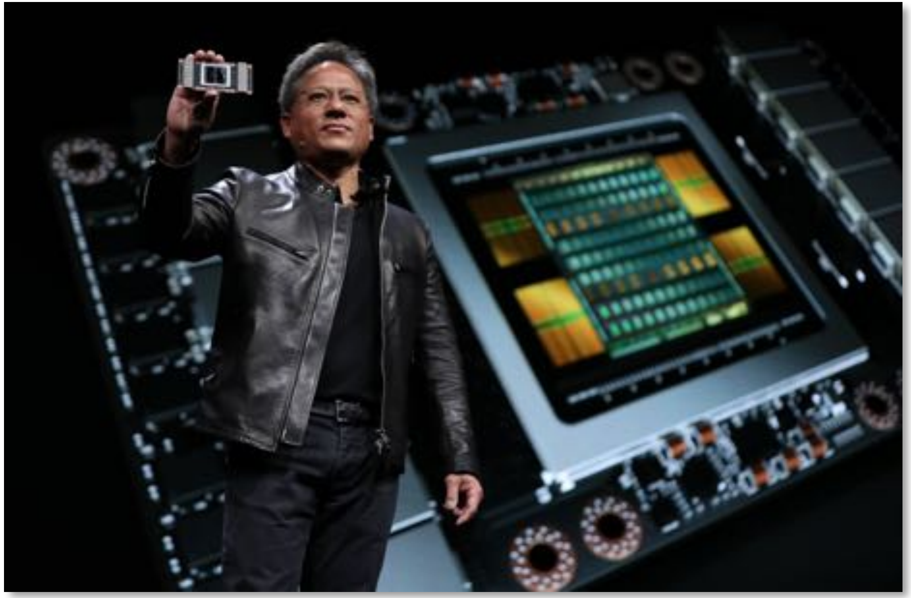
1x AMD CPU

4 Slingshot connections

- 4x25 GB/s

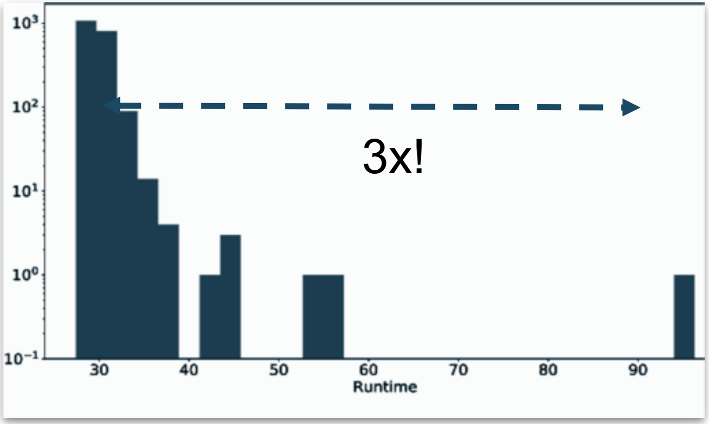
GPU direct, Unified Virtual Memory (UVM)

2-3x Cori



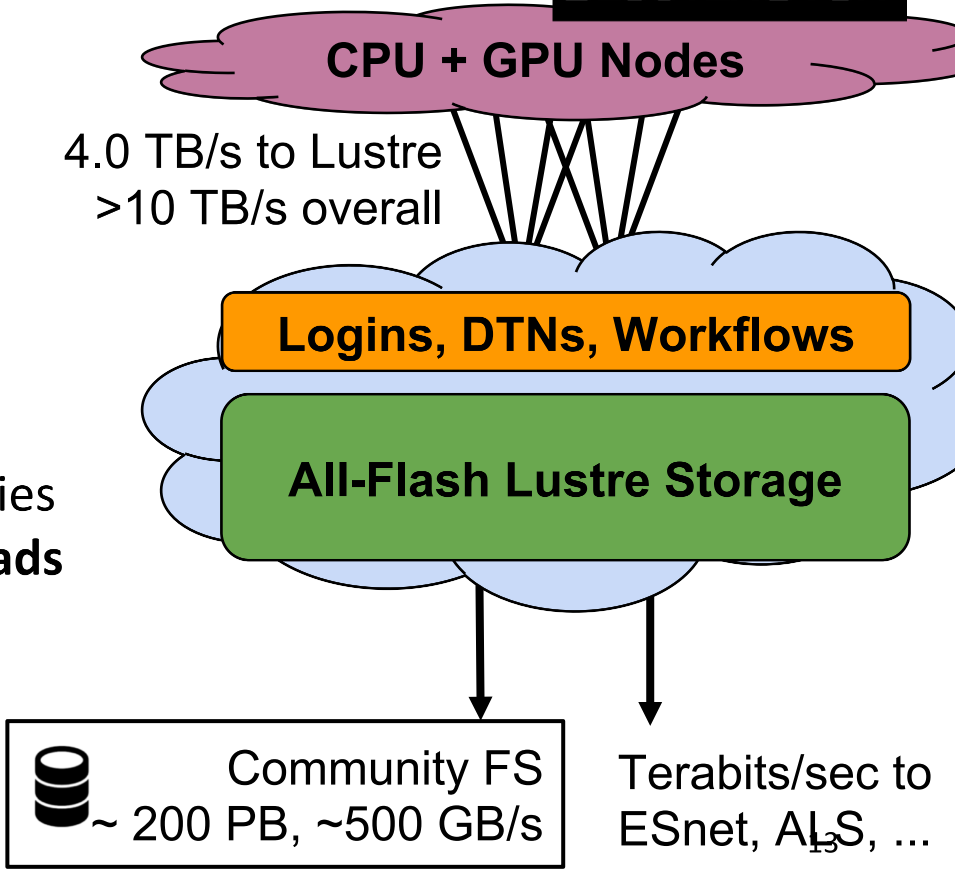
# Slingshot Network

- **High Performance scalable interconnect**
  - Low latency, high-bandwidth, MPI performance enhancements
  - 3 hops between any pair of nodes
  - Sophisticated congestion control and adaptive routing to minimize tail latency
- **Ethernet compatible**
  - Blurs the line between the inside and the outside of the machine
  - Allow for seamless external communication
  - Direct interface to storage



# Perlmutter has a All-Flash Filesystem

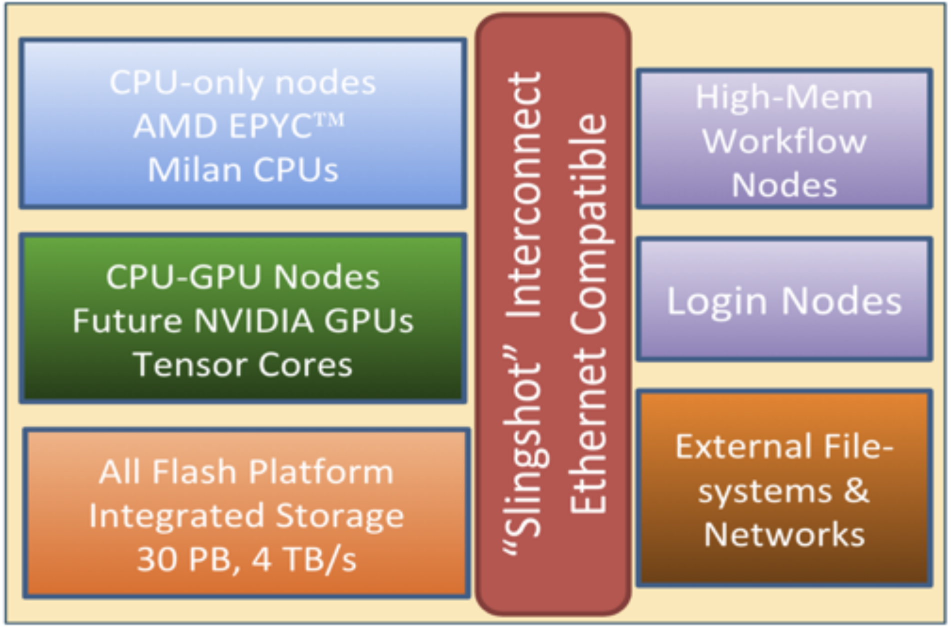
- **Fast across many dimensions**
  - 4 TB/s sustained bandwidth
  - 7,000,000 IOPS
  - 3,200,000 file creates/sec
- **Usable for NERSC users**
  - 30 PB usable capacity
  - Familiar Lustre interfaces
  - New data movement capabilities
- **Optimized for NERSC data workloads**
  - NEW small-file I/O improvements
  - NEW features for high IOPS, non-sequential I/O



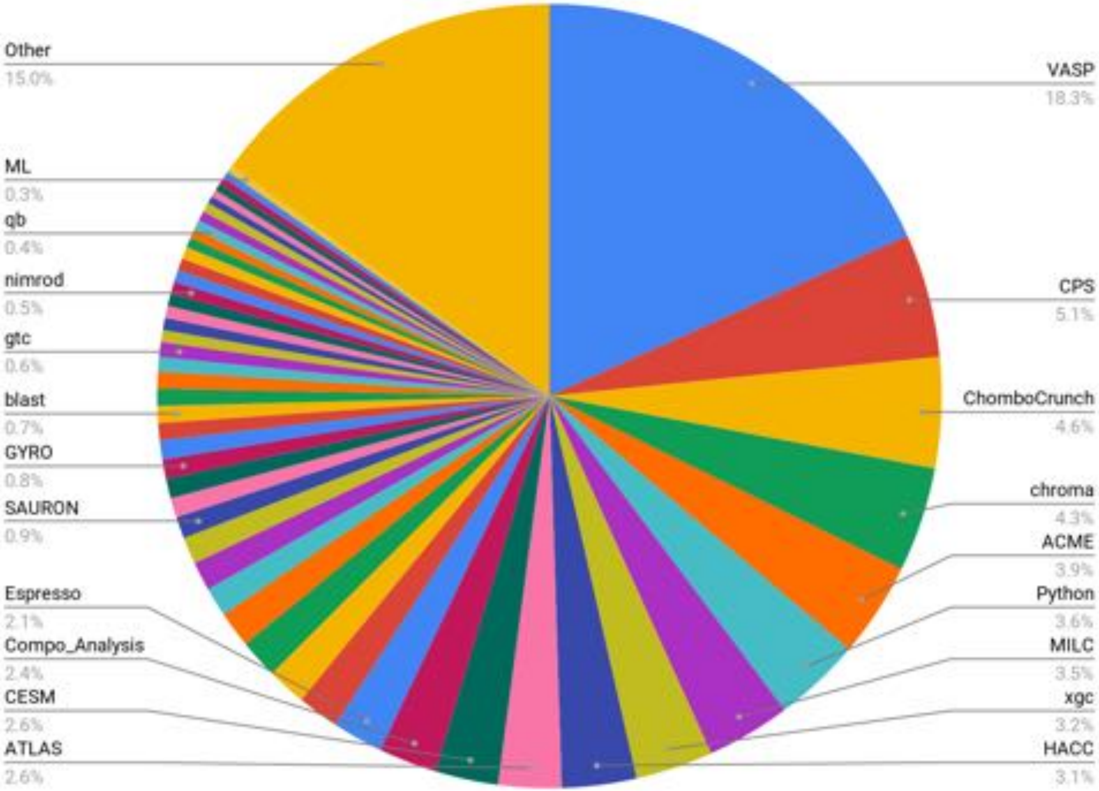
# Perlmutter: A System Optimized for Science

- **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

*How do we optimize the size of each partition?*

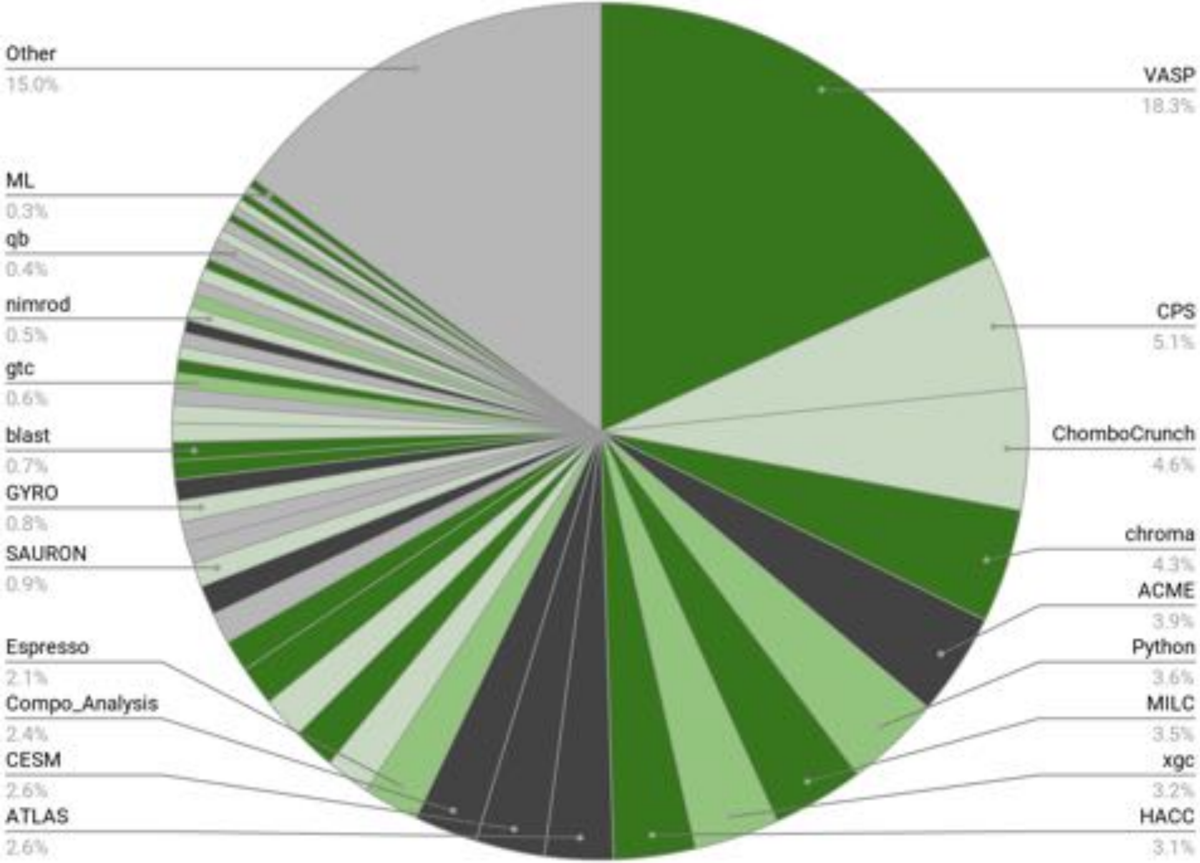


# NERSC System Utilization (Aug’17 - Jul’18)



- 3 codes > 25% of the workload
- 10 codes > 50% of the workload
- 35 codes > 75% of the workload
- Over 600 codes comprise the remaining 25% of the workload.

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



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

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

# How many GPU nodes to buy - Benchmark Suite Construction & Scalable System Improvement

Select codes to represent the anticipated workload

- Include key applications from the current workload.
- Add apps that are expected to be contribute significantly to the future workload.

**Scalable System Improvement**

Measures aggregate performance of HPC machine

- How many more copies of the benchmark can be run relative to the reference machine
- Performance relative to reference machine

$$SSI = \left\langle \frac{\#Nodes \times \ Jobsize \times Perf\_per\_node}{\#Nodes_{Ref} \times \ Jobsize_{Ref} \times Perf\_per\_node_{Ref}} \right\rangle$$

| Application | Description |
|---|---|
| **Quantum Espresso** | Materials code using DFT |
| **MILC** | QCD code using staggered quarks |
| **StarLord** | Compressible radiation hydrodynamics |
| **DeepCAM** | Weather/Community Atmospheric Model 5 |
| **GTC** | Fusion PIC code |
| **“CPU Only” (3 Total)** | Representative of applications that cannot be ported to GPUs |

NERSC

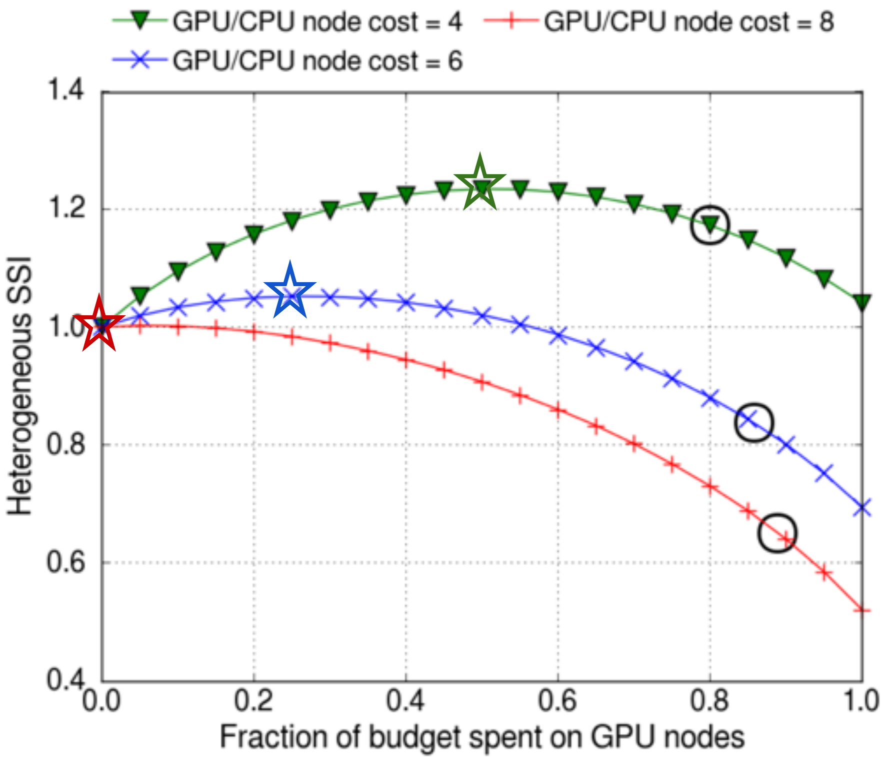
B. Austin, C. Daley, D. Doerfler, J. Deslippe, B. Cook, B. Friesen, T. Kurth, C. Yang, N. J. Wright, “A Metric for Evaluating Supercomputer Performance in the Era of Extreme Heterogeneity", *9th IEEE International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS18),* November 12, 2018,

# Hetero system design & price sensitivity: Budget for GPUs increases as GPU price drops

**Chart explores an isocost design space**

- **Vary the budget allocated to GPUs**
- **Assume GPU enabled applications have performance advantage = 10x per node, 3 of 8 apps are still CPU only.**
- **Examine GPU/CPU node cost ratio**

| **GPU / CPU $ per node** | **SSI increase vs. CPU-Only (@ budget %)** | |
|---|---|---|
| 8:1 | None | No justification for GPUs |
| 6:1 | 1.05x @ 25% | Slight justification for up to 50% of budget on GPUs |
| 4:1 | 1.23x @ 50% | GPUs cost effective up to full system budget, but optimum at 50% |



B. Austin, C. Daley, D. Doerfler, J. Deslippe, B. Cook, B. Friesen, T. Kurth, C. Yang, N. J. Wright, "A Metric for Evaluating Supercomputer Performance in the Era of Extreme Heterogeneity", *9th IEEE International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS18),* November 12, 2018,

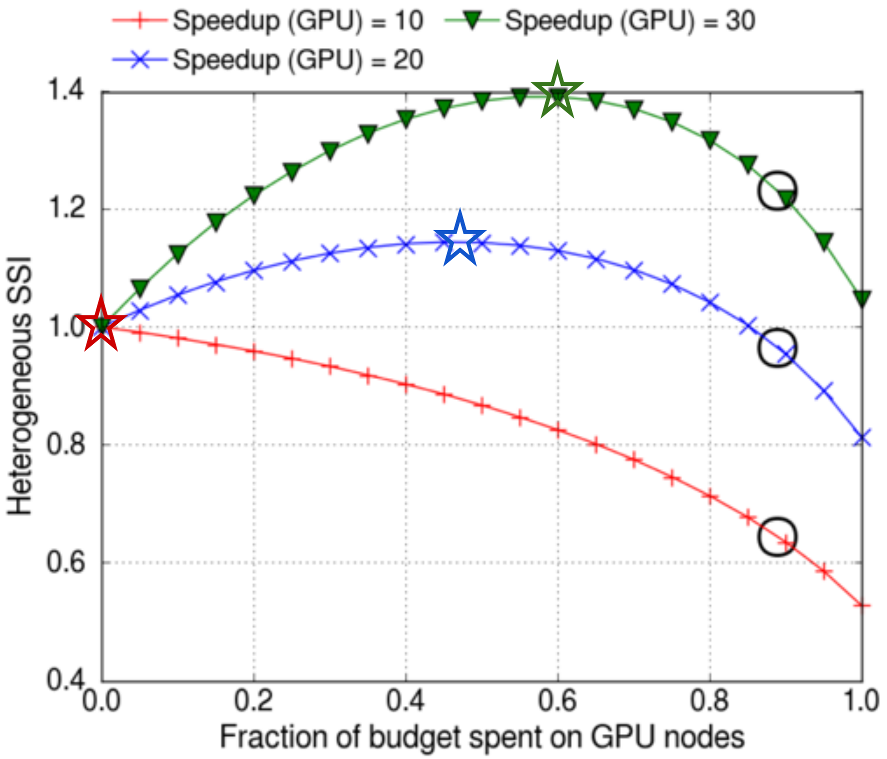
Circles: 50% CPU nodes + 50% GPU nodes
Stars: Optimal system configuration.

# Application readiness efforts justify larger GPU partitions.

**Explore an isocost design space**

- **Assume 8:1 GPU/CPU node cost ratio.**
- **Vary the budget allocated to GPUs**
- **Examine GPU / CPU *performance* gains such as those obtained by software optimization & tuning. 5 of 8 codes have 10x, 20x, or 30x speedup.**

| **GPU / CPU perf. per node** | **SSI increase vs. CPU-Only (@ budget %)** | |
|---|---|---|
| 10x | None | No justification for GPUs |
| 20x | 1.15x @ 45% | Compare to 1.23x<br>for 10x at 4:1 GPU/CPU cost ratio |
| 30x | 1.40x @ 60% | Compare to 3x<br>from NESAP for KNL |



Circles: 50% CPU nodes + 50% GPU nodes
Stars: Optimal system configuration

B. Austin, C. Daley, D. Doerfler, J. Deslippe, B. Cook, B. Friesen, T. Kurth, C. Yang, N. J. Wright, "A Metric for Evaluating Supercomputer Performance in the Era of Extreme Heterogeneity", *9th IEEE International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS18),* November 12, 2018,

# Application Readiness Strategy for Perlmutter

## How to Enable NERSC's diverse community of 7,000 users, 750 projects, and 700 codes to run on advanced architectures like Perlmutter and beyond?

- NERSC Exascale Science Application Program (NESAP)
- Engage ~25 Applications
- up to 17 postdoctoral fellows
- Deep partnerships with every SC Office area
- Leverage vendor expertise and community hack-a-thons
- Knowledge transfer through documentation and training for all users
- Optimize codes with improvements relevant to multiple architectures

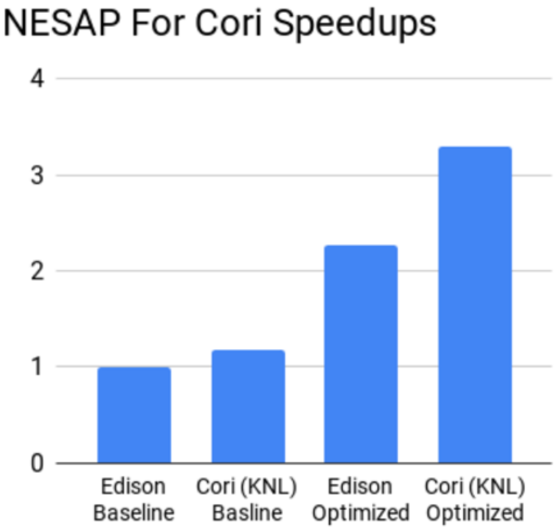
# GPU Transition Path for Apps

**NESAP for Perlmutter will extend activities from NESAP**

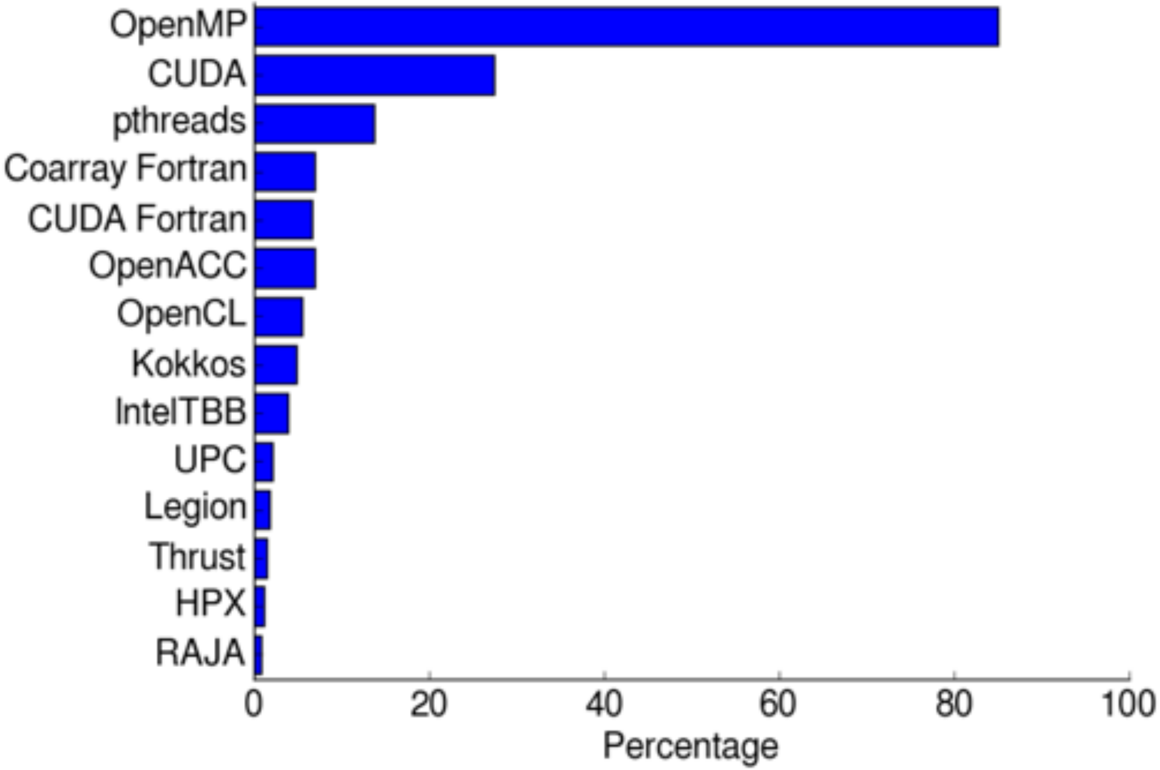
1. **Identifying and exploiting on-node parallelism**
2. **Understanding and improving data-locality within the memory hierarchy**

**What's New for NERSC Users?**

1. **Heterogeneous compute elements**
2. **Identification and exploitation of even more parallelism**
3. **Emphasis on performance-portable programming approach:**
   - Continuity from Cori through future NERSC systems and other DOE platforms



# OpenMP is the most popular non-MPI parallel programming technique

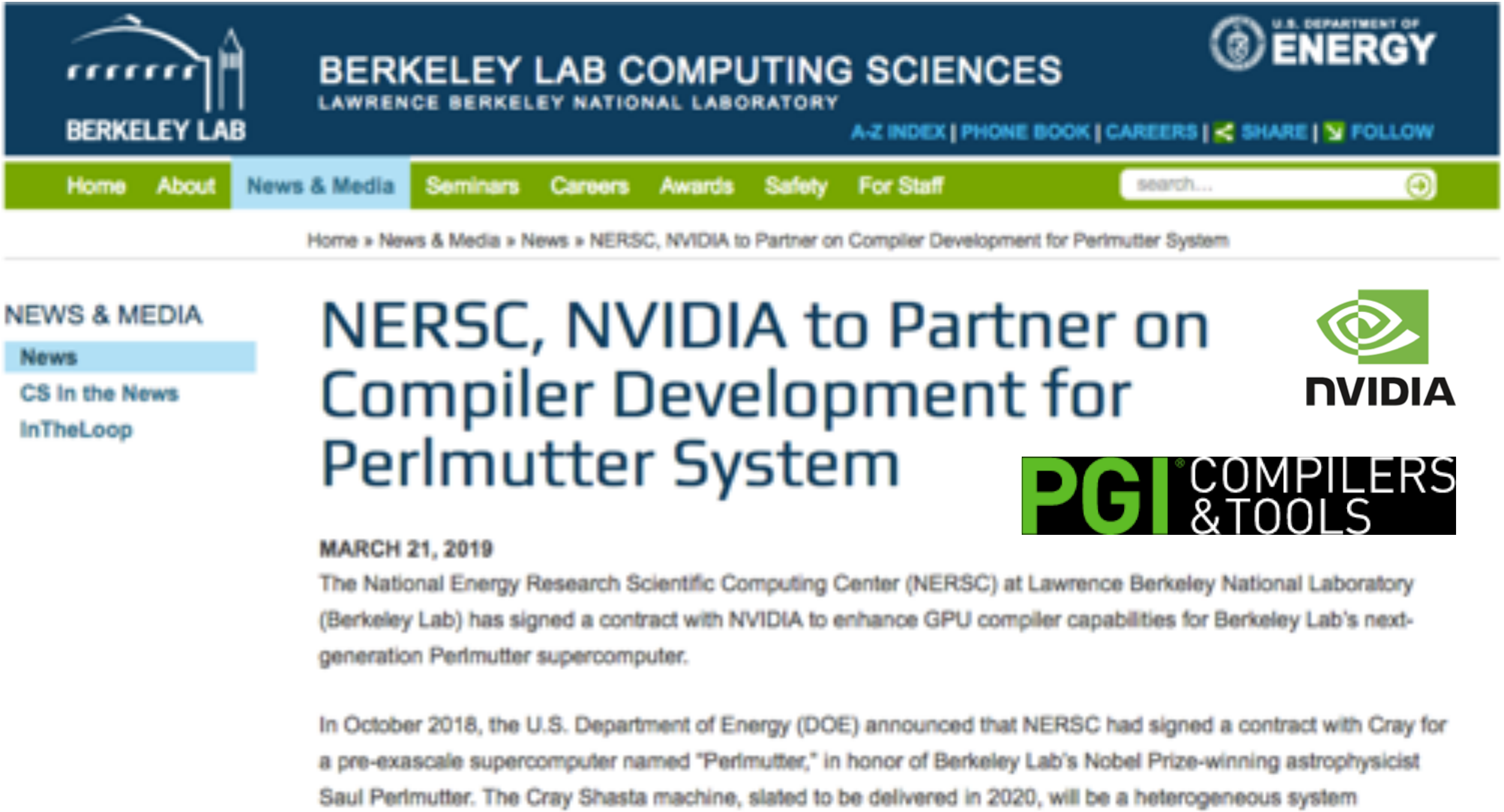


- **Results from ERCAP 2017 user survey**
  - Question answered by 328 of 658 survey respondents
- **Total exceeds 100% because some applications use multiple techniques**

# OpenMP meets the needs of the NERSC workload

- **Supports C, C++ and Fortran**
  - The NERSC workload consists of ~700 applications with a relatively equal mix of C, C++ and Fortran
- **Provides portability to different architectures at other DOE labs**
- **Works well with MPI: hybrid MPI+OpenMP approach successfully used in many NERSC apps**
- **Recent release of OpenMP 5.0 specification – the third version providing features for accelerators**
  - Many refinements over this five year period

# NRE partnership with PGI/NVIDIA


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
NVIDIA
PGI COMPILERS & TOOLS

## 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

# OpenMP NRE – Status & Future Plans

**Contract 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 contract items**

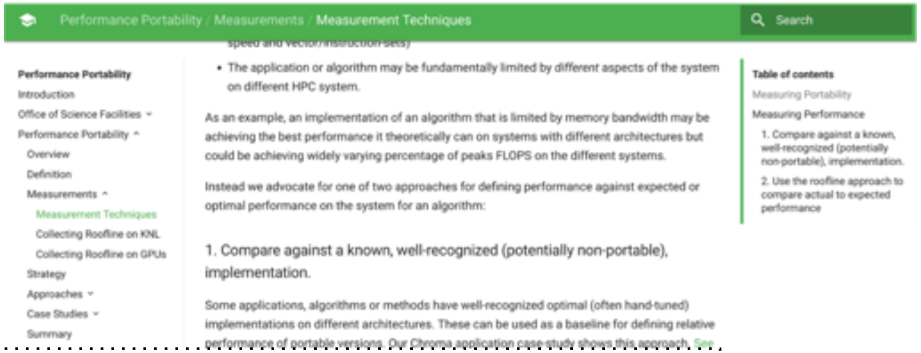
- Evaluate the Alpha compiler on Cori-GPU
- Evaluate upcoming compiler releases: Apr 2020 and Oct 2020
  - More NESAP and NERSC users will get access with each compiler release

# Engaging around Performance Portability

NERSC is working with PGI/NVIDIA to enable OpenMP GPU acceleration

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

NERSC is 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
Summary
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

speed and vector/instruction sets)

- 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) implementations on different architectures. These can be used as a baseline for defining relative performance of portable versions. Our Chroma application case-study shows this approach. See

NERSC is leading development of performanceportability.org

Doug Doerfler Lead Performance Portability Workshop at SC18. and 2019 DOE COE Perf. Port. Meeting

# NERSC-9 will be named after Saul Perlmutter

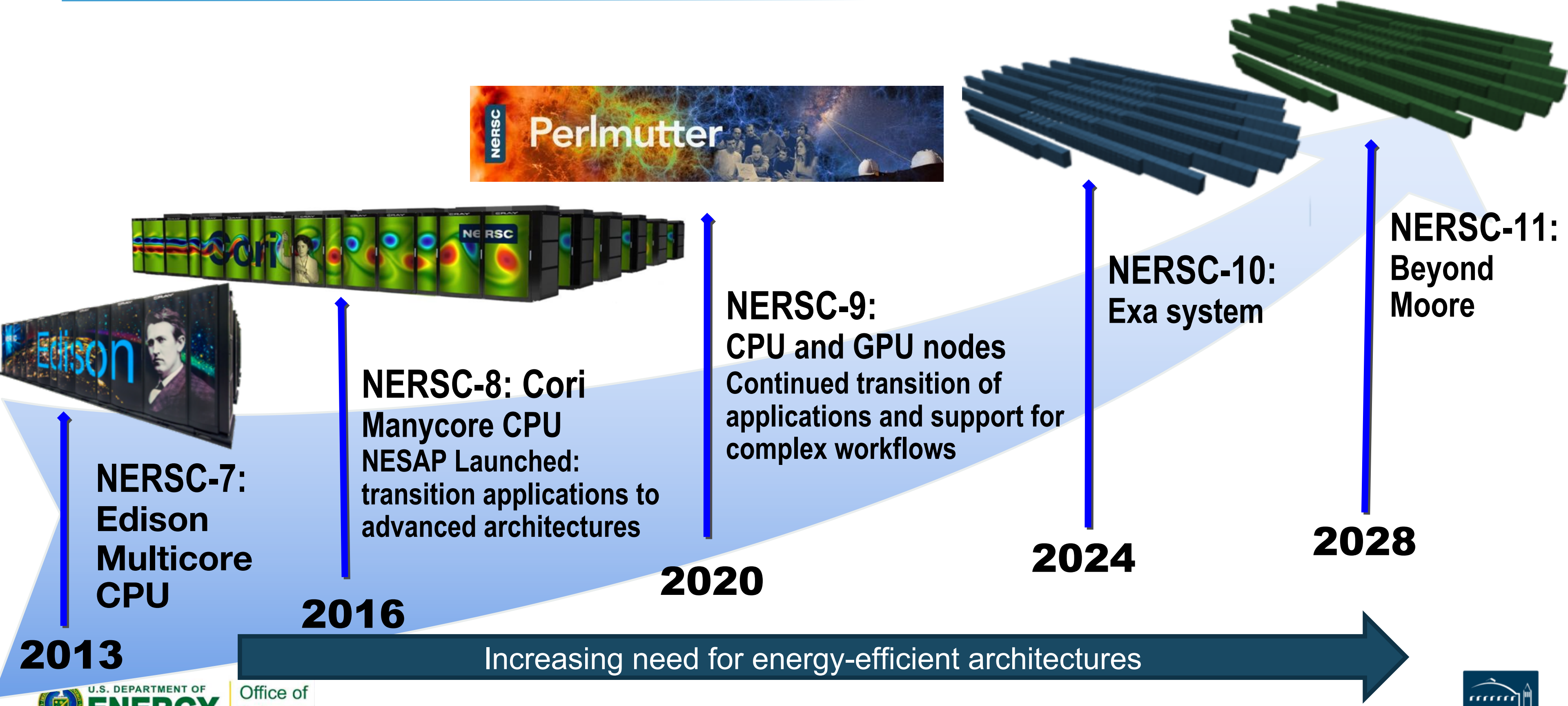
- Winner of 2011 Nobel Prize in Physics for discovery of the accelerating expansion of the universe.
- Supernova Cosmology Project, lead by Perlmutter, was a pioneer in using NERSC supercomputers combine large scale simulations with experimental data analysis
- Login “saul.nersc.gov”

NERSC

# 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

# NERSC Systems 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

# Will GPUs work for everybody?

- **Will 100% of the NERSC workload be able to utilize GPUs by 2024?**
  - Yes, they just need to modify their code
  - No, their algorithm needs changing
  - No, their physics is fundamentally not amenable to data parallelism
  - No, they just don’t have time or need too

# Technology Scaling Trends

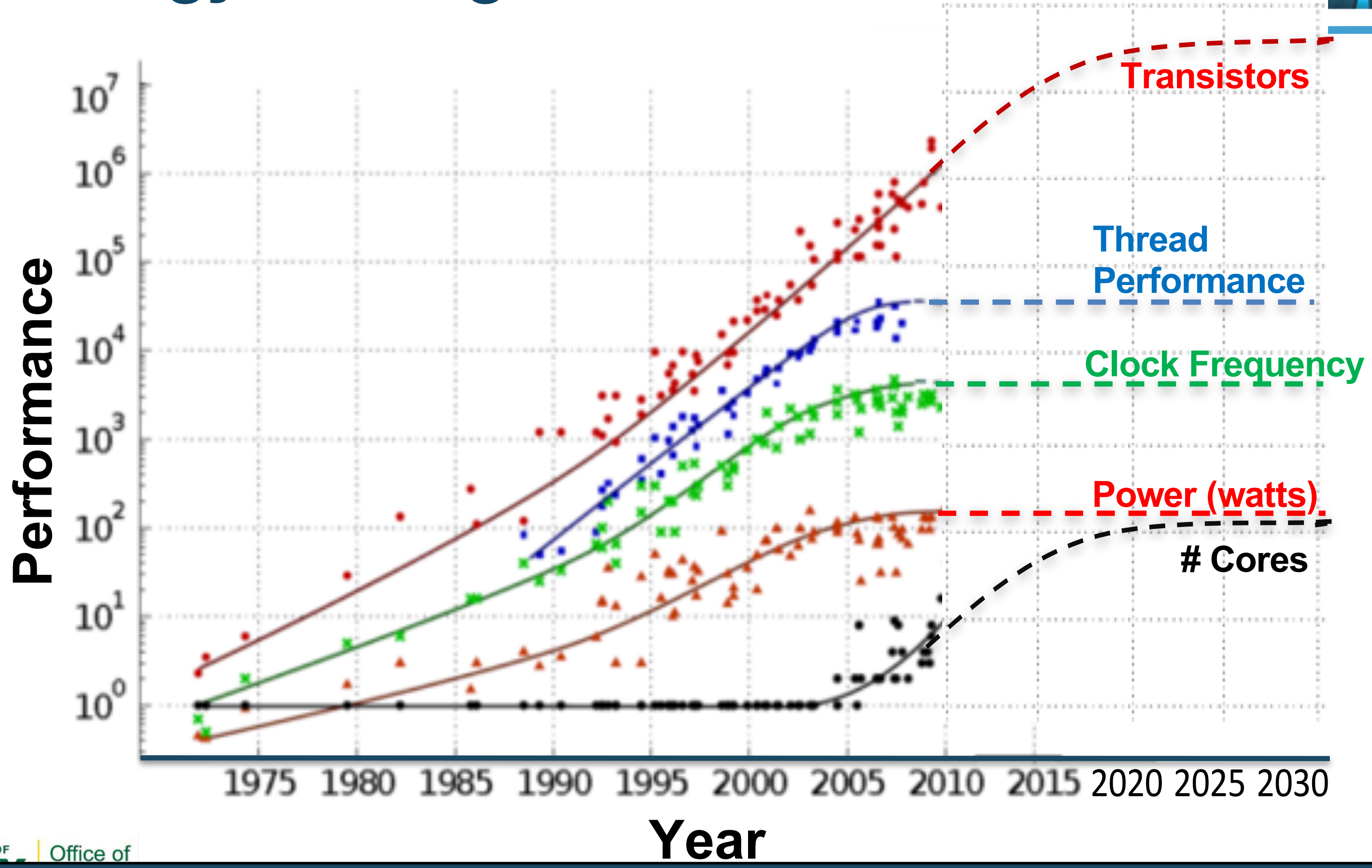
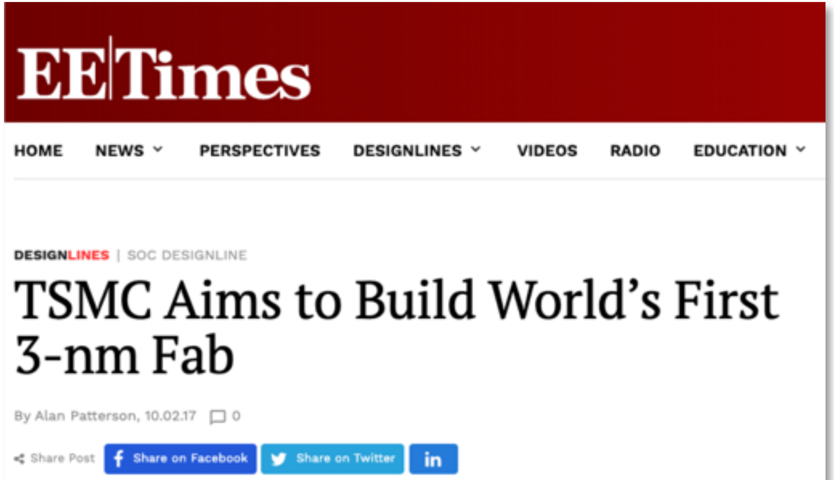


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

# Next-Next generation Process Nodes have been announced



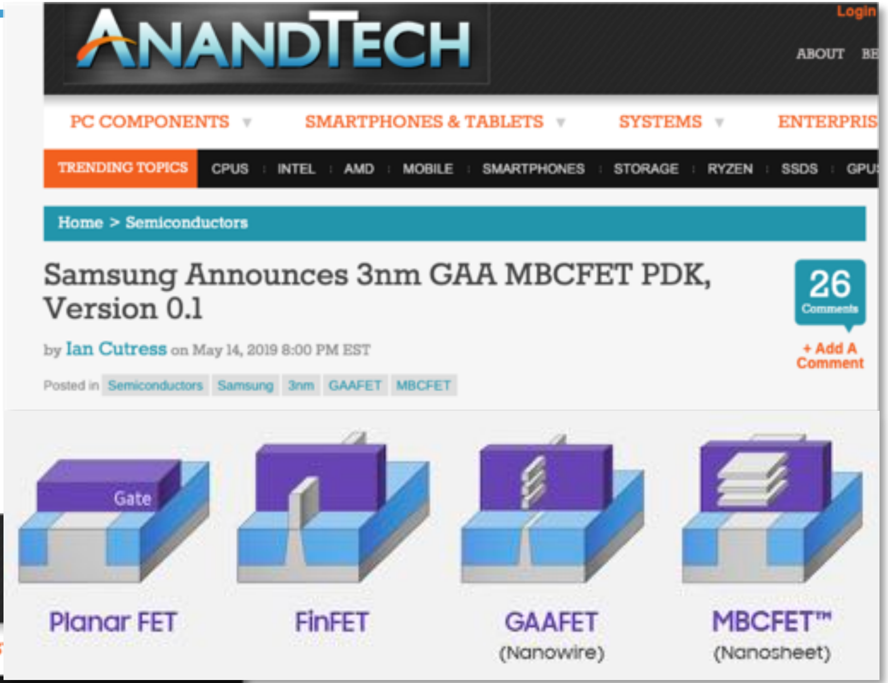
EETimes

DESIGNLINES | SOC DESIGNLINE

## TSMC Aims to Build World's First 3-nm Fab

By Alan Patterson, 10.02.17 0

TAIPEI — Taiwan Semiconductor Manufacturing (
in the Tainan Science Park in southern Taiwan, wl
manufacturing.



ANANDTECH

Home > Semiconductors

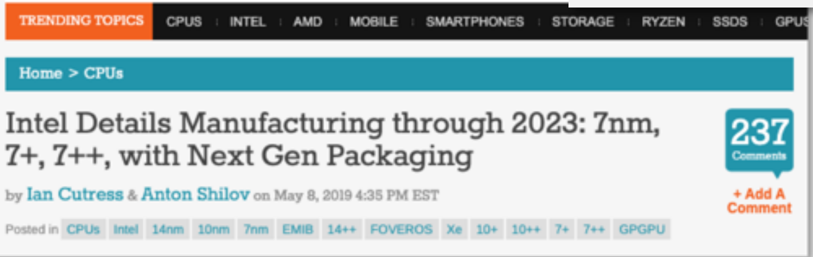
## Samsung Announces 3nm GAA MBCFET PDK, Version 0.1

26 Comments

+ Add A Comment

by Ian Cutress on May 14, 2019 8:00 PM EST

Posted in Semiconductors Samsung 3nm GAAFET MBCFET



ANANDTECH

Home > CPUs

## Intel Details Manufacturing through 2023: 7nm, 7+, 7++, with Next Gen Packaging

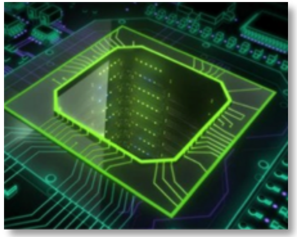
237 Comments

+ Add A Comment

by Ian Cutress & Anton Shilov on May 8, 2019 4:35 PM EST

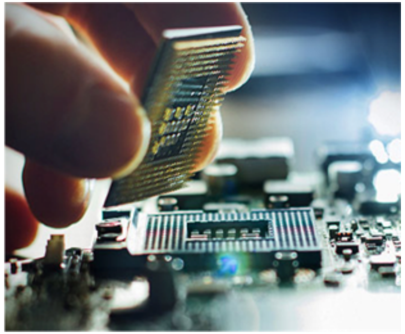
Posted in CPUs Intel 14nm 10nm 7nm EMIB 14++ FOVEROS Xe 10+ 10++ 7+ 7++ GPGPU

# Specialization: End Game for Moore’s Law



NVIDIA builds deep learning appliance with V100 Tesla’s

RISC-V is an open hardware platform

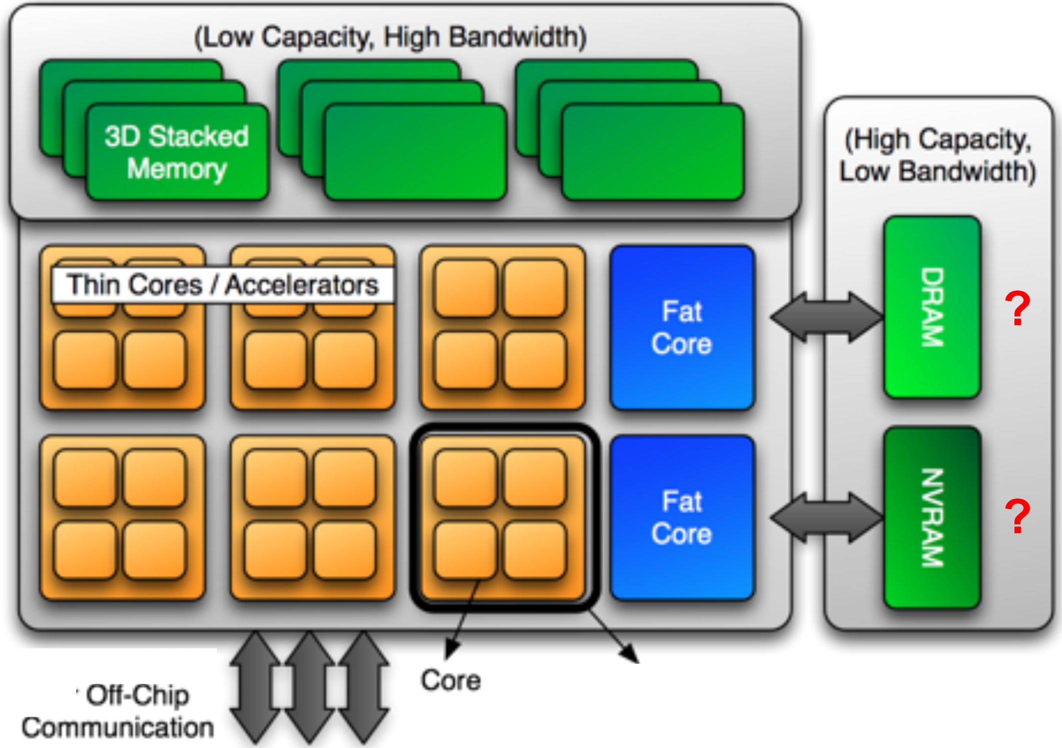


Intel buys deep learning startup, Nervana

FPGAs offer configurable specialization

Google designs its own Tensor Processing Unit (TPU)

# Potential 2024 Node


(Low Capacity, High Bandwidth)
3D Stacked Memory
(High Capacity, Low Bandwidth)
Thin Cores / Accelerators
Fat Core
Fat Core
DRAM
?
NVRAM
?
Off-Chip Communication
Core

- Vendors converging to a mixture of energy-efficient Thin Cores/Accelerators and Fat Cores

- Potentially with DRAM/NVRAM

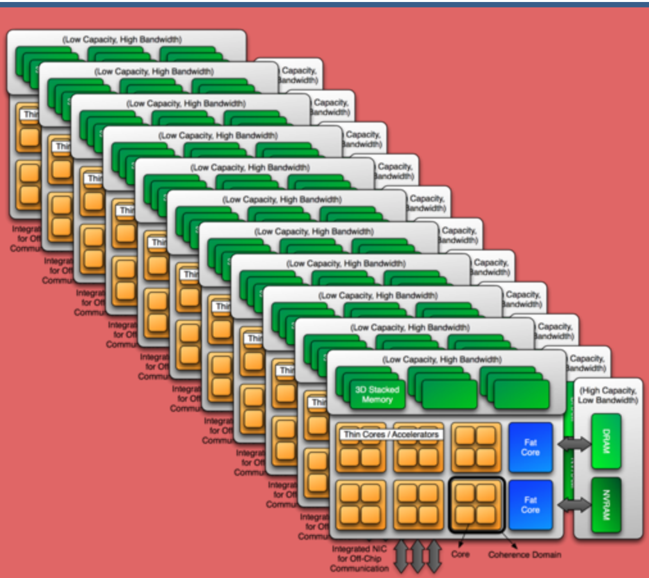
- (Hopefully) leads to less focus on data motion and more on identifying parallelism

J. A. Ang *et al*., "Abstract Machine Models and Proxy Architectures for Exascale Computing," *2014 Hardware-Software Co-Design for High Performance Computing*, New Orleans, LA, 2014, pp. 25-32.
doi: 10.1109/Co-HPC.2014.4 http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7017960&isnumber=7017953

# Potential NERSC-10 system #1

Edge Services

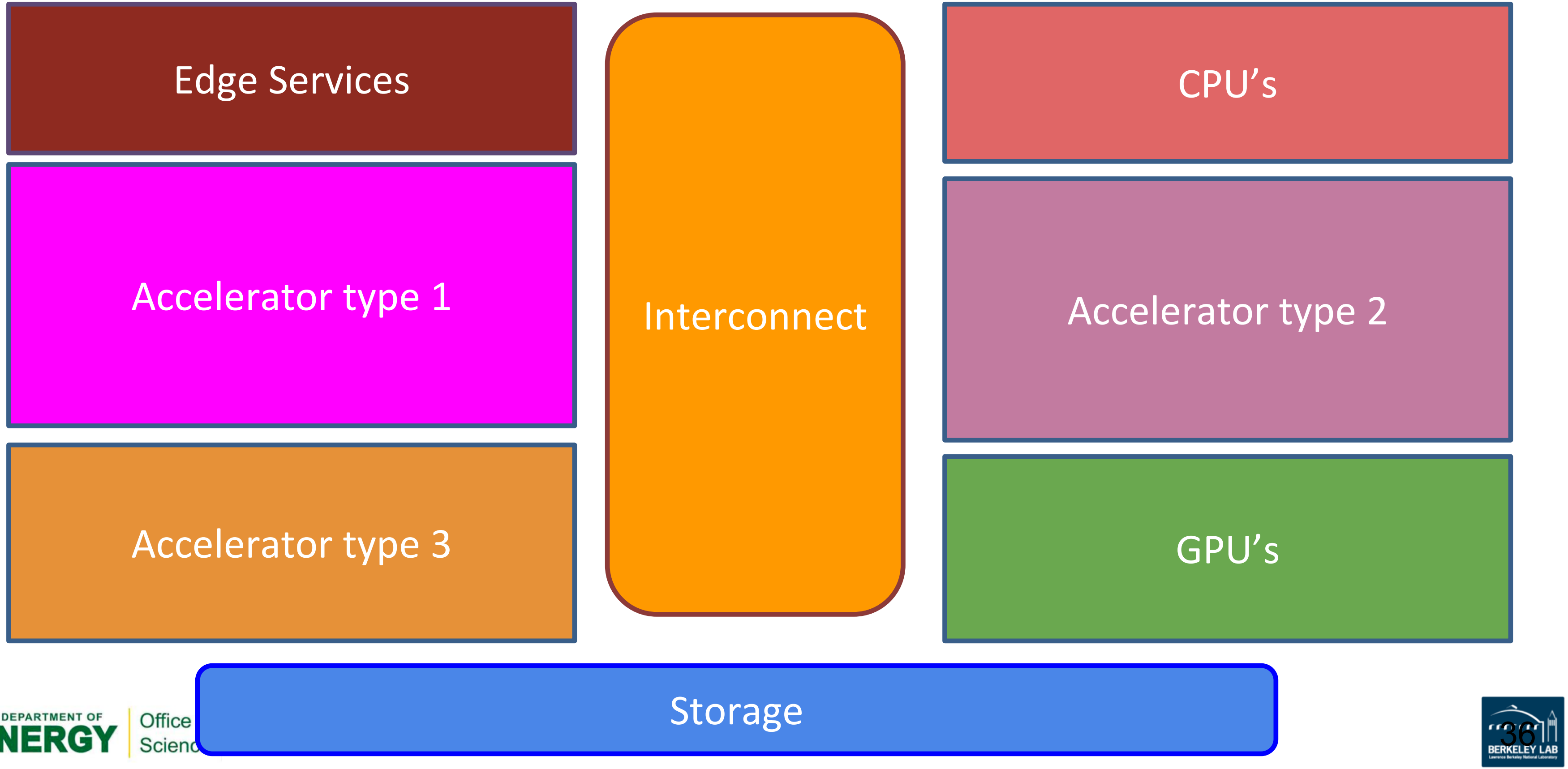
Interconnec
t


(Low Capacity, High Bandwidth)
3D Stacked Memory
Capacity, Bandwidth)
(High Capacity, Low Bandwidth)
Thin Cores / Accelerators
Fat Core
Fat Core
DRAM
NVRAM
Integrated NIC for Off-Chip Communication
Core
Coherence Domain

Heterogeneous
nodes

Storage

# Potential NERSC-10 system #2


Edge Services

Accelerator type 1

Accelerator type 3

Interconnect

CPU’s

Accelerator type 2

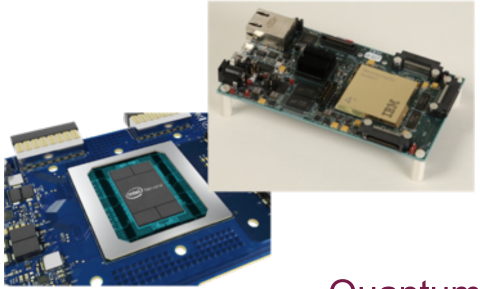
GPU’s

Storage

# Exploring Workflow Accelerators for SC Applications with NERSC-9 and Slingshot network

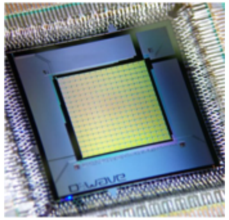
- *What accelerators map to existing SC workloads? And what SC challenges could be solved with emerging accelerators?*
- Key areas of investigation
  - Identify common algorithms, kernels, motifs that run well on emerging accelerators.
  - Determine feasibility of configurable processing technologies, e.g. FPGAs?
  - Analyze changing workload requirements, e.g. ML.

Neural Network Processors



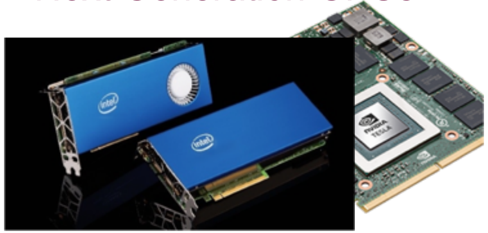
Emerging Technologies

Quantum Computing

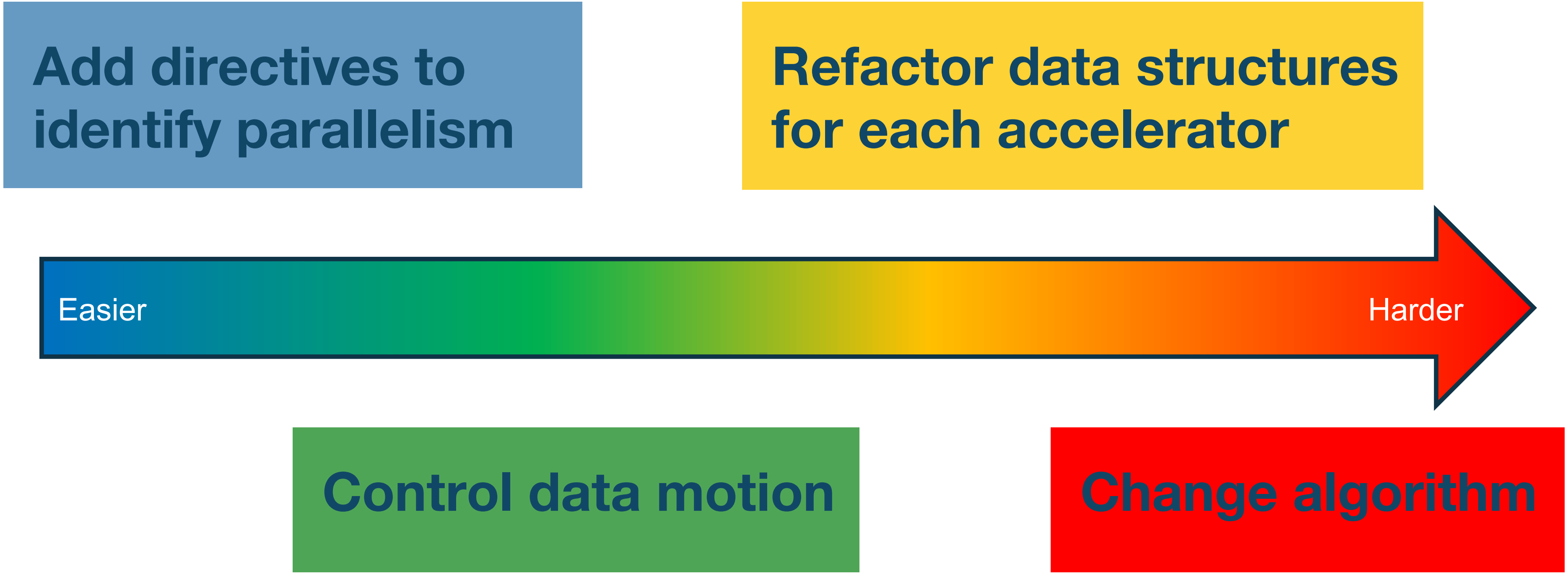


Programmable Arrays

Next Generation GPUs



# Spectrum of approaches to optimizing an application for accelerators - today



Add directives to identify parallelism

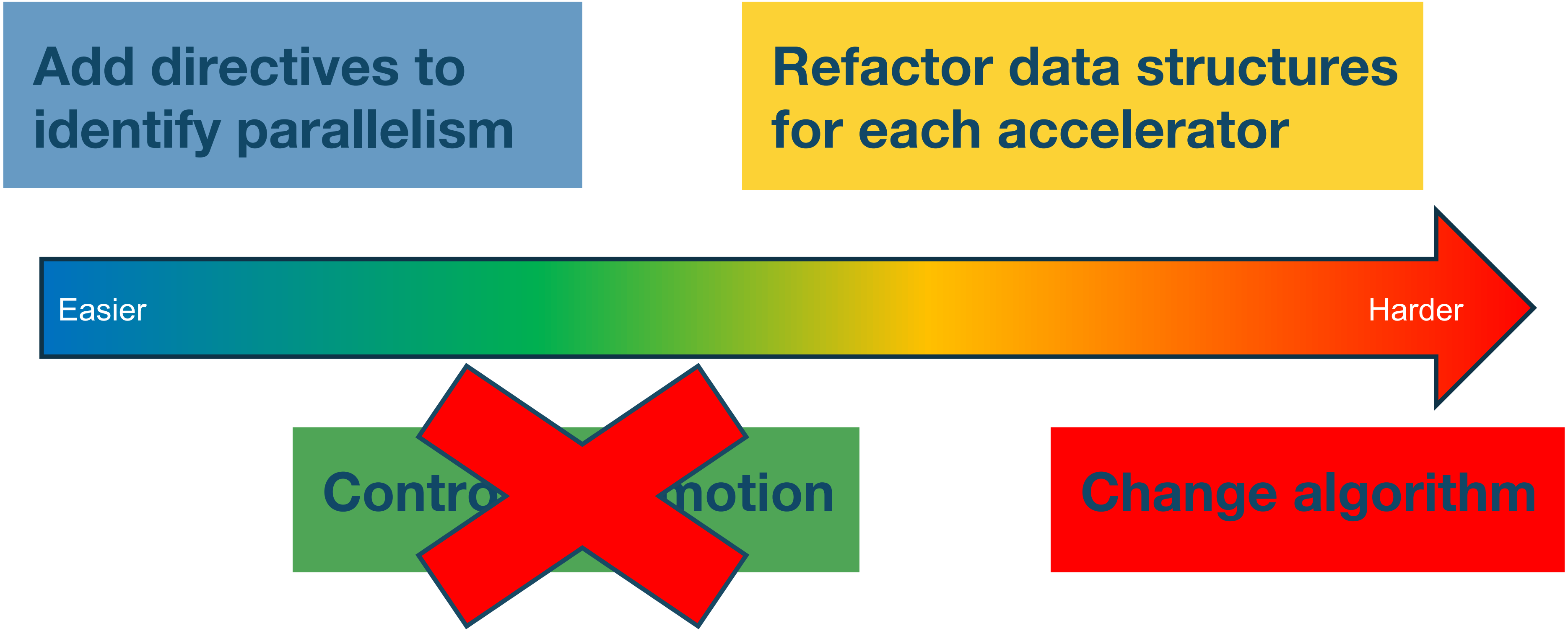
Refactor data structures for each accelerator

Control data motion

Change algorithm

Worse case is that a different algorithm is the optimal one for different architectures !

# Spectrum of approaches to optimizing an application for accelerators



Add directives to identify parallelism

Refactor data structures for each accelerator

Easier

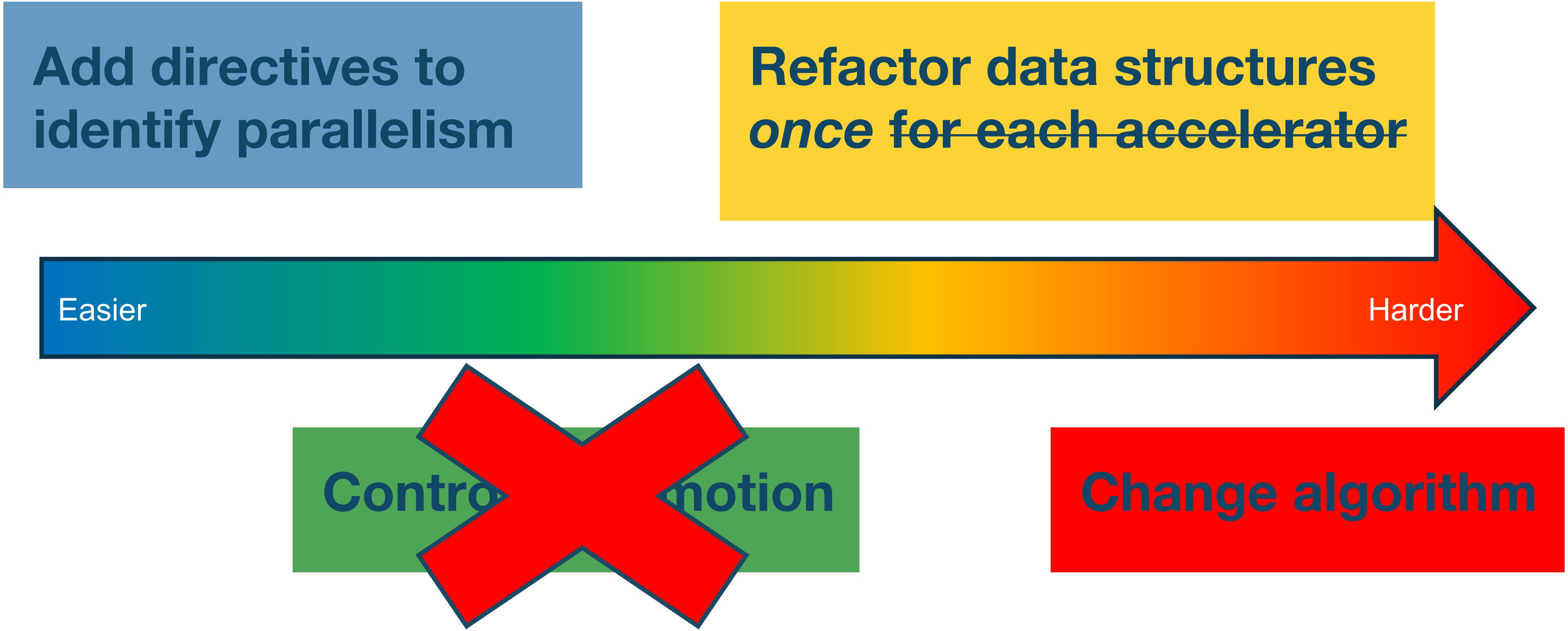
Harder

Contro motion

Change algorithm

Worse case is that a different algorithm is the optimal one for different architectures !

# Spectrum of approaches to optimizing an application for accelerators



Add directives to identify parallelism

Refactor data structures *once* ~~for each accelerator~~

~~Contro~~ ~~motion~~

Change algorithm

Worse case is that a different algorithm is the optimal one for different architectures !

# Summary

- **Hardware trends should reduce some of the burden on programmers today**
- **Software developments that separate or abstract away the details of the hardware should similarly help**
  - Programmer (or library expert) specializes the code for the hardware, e.g. Kokkos, Raja, OpenMP-5 declare variant and metadirective directives
  - Programmer specifies that parallel transformations are safe and allows compiler to specialize for the hardware, e.g. OpenMP-5 loop directive, Fortran do concurrent
- **Unrealistic to expect performance *and* portability while hardware has not converged**

# Question for the audience

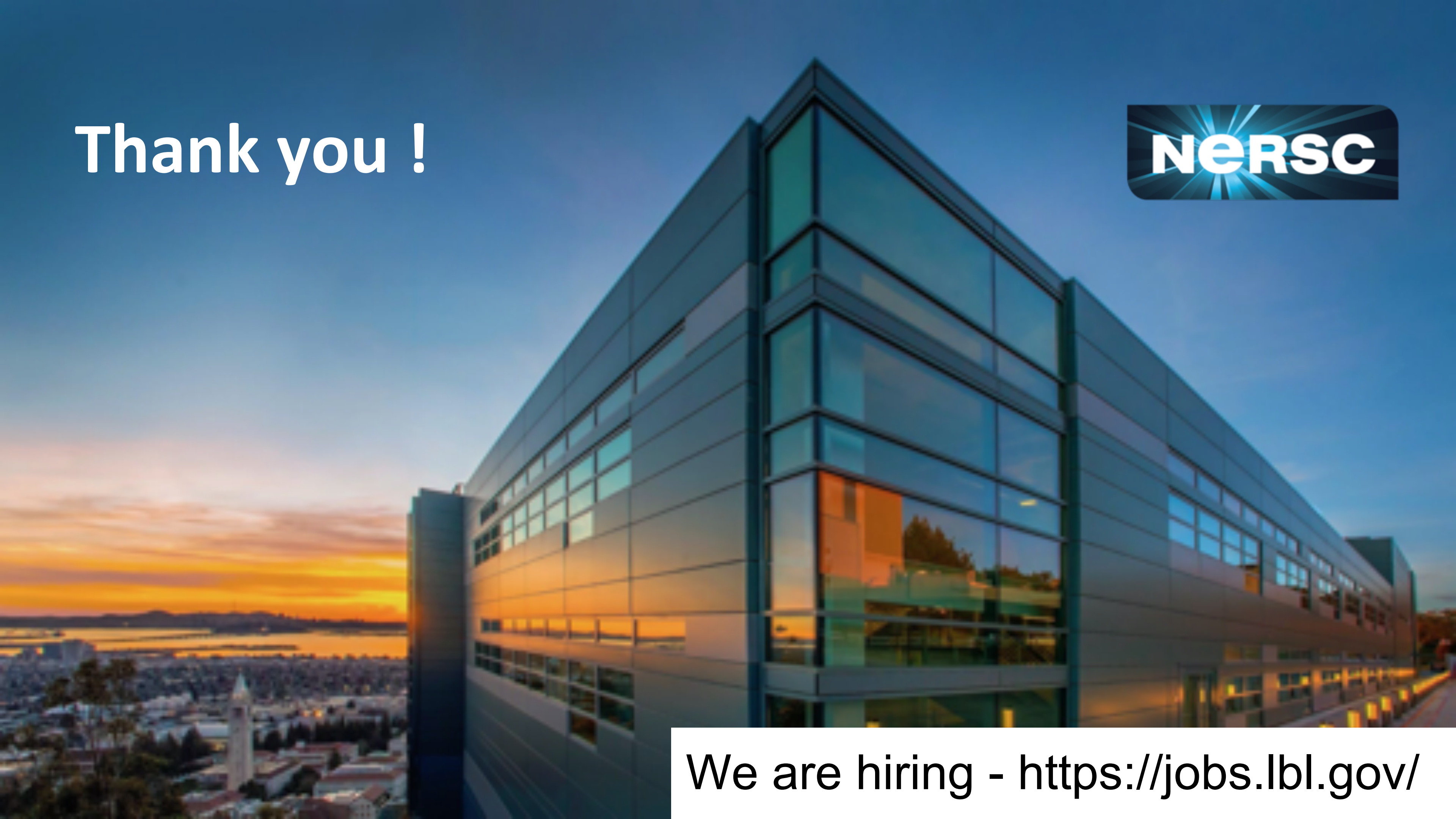
- **Will there be a Workshop on Accelerator Programming Using Directives (WACCPD) in 2024?**
  1. Yes
  2. No

- **Will we still need directives in 5 yrs? 10 yrs ?**
  1. Yes
  2. No

# Question for the audience

- **Will there be a Workshop on Accelerator Programming Using Directives (WACCPD) in 2024?**
  1. Yes
  2. **No – I hope the conversation will have moved on by then!**

- **Will we need directives in 5 yrs? 10 yrs?**
  1. **Yes – For some other reason….**
  2. No

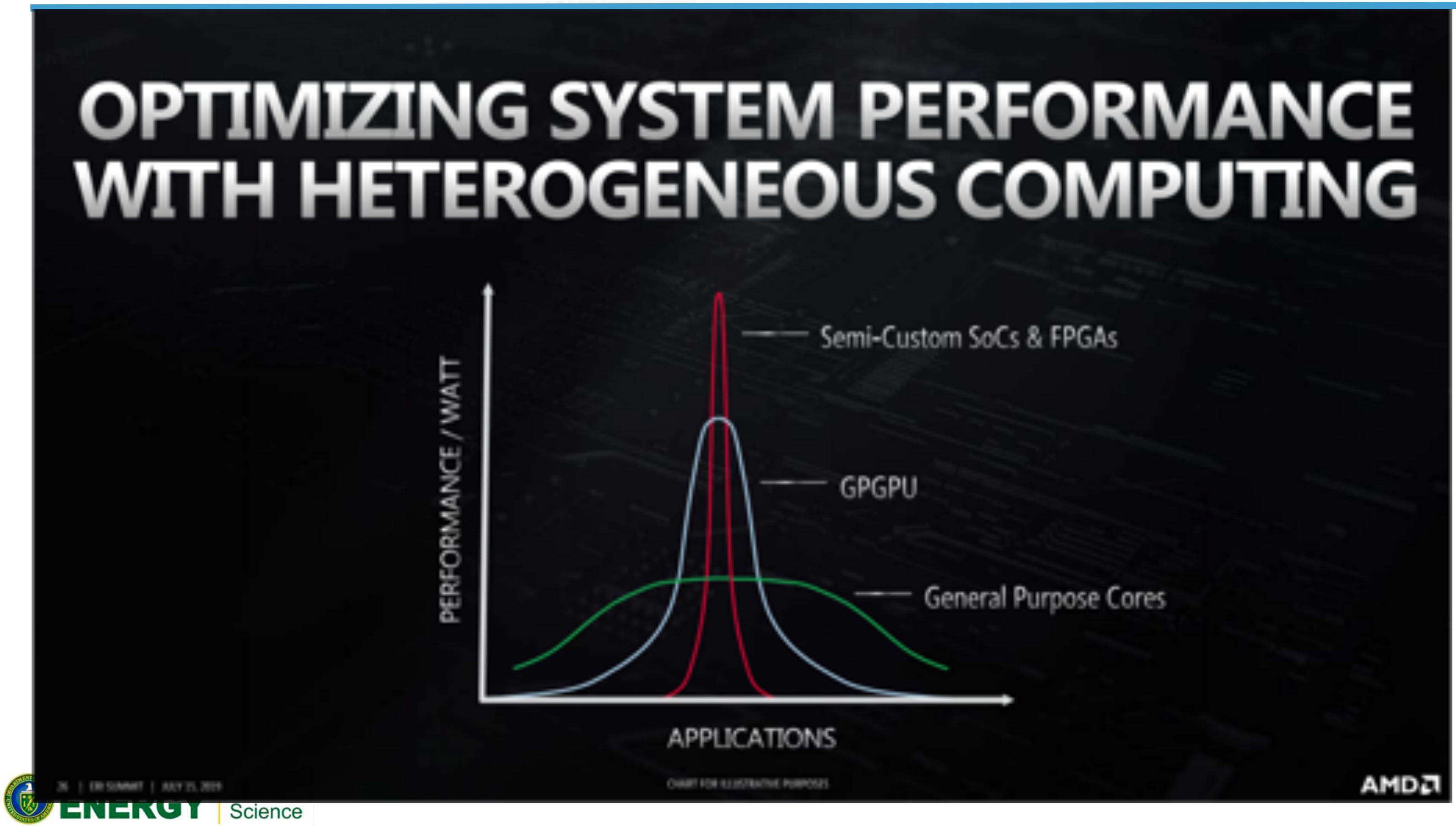


# Thank you !

NERSC

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

# View from AMD - can we exploit this to benefit NERSC users ?



## OPTIMIZING SYSTEM PERFORMANCE WITH HETEROGENEOUS COMPUTING

CHART FOR ILLUSTRATIVE PURPOSES