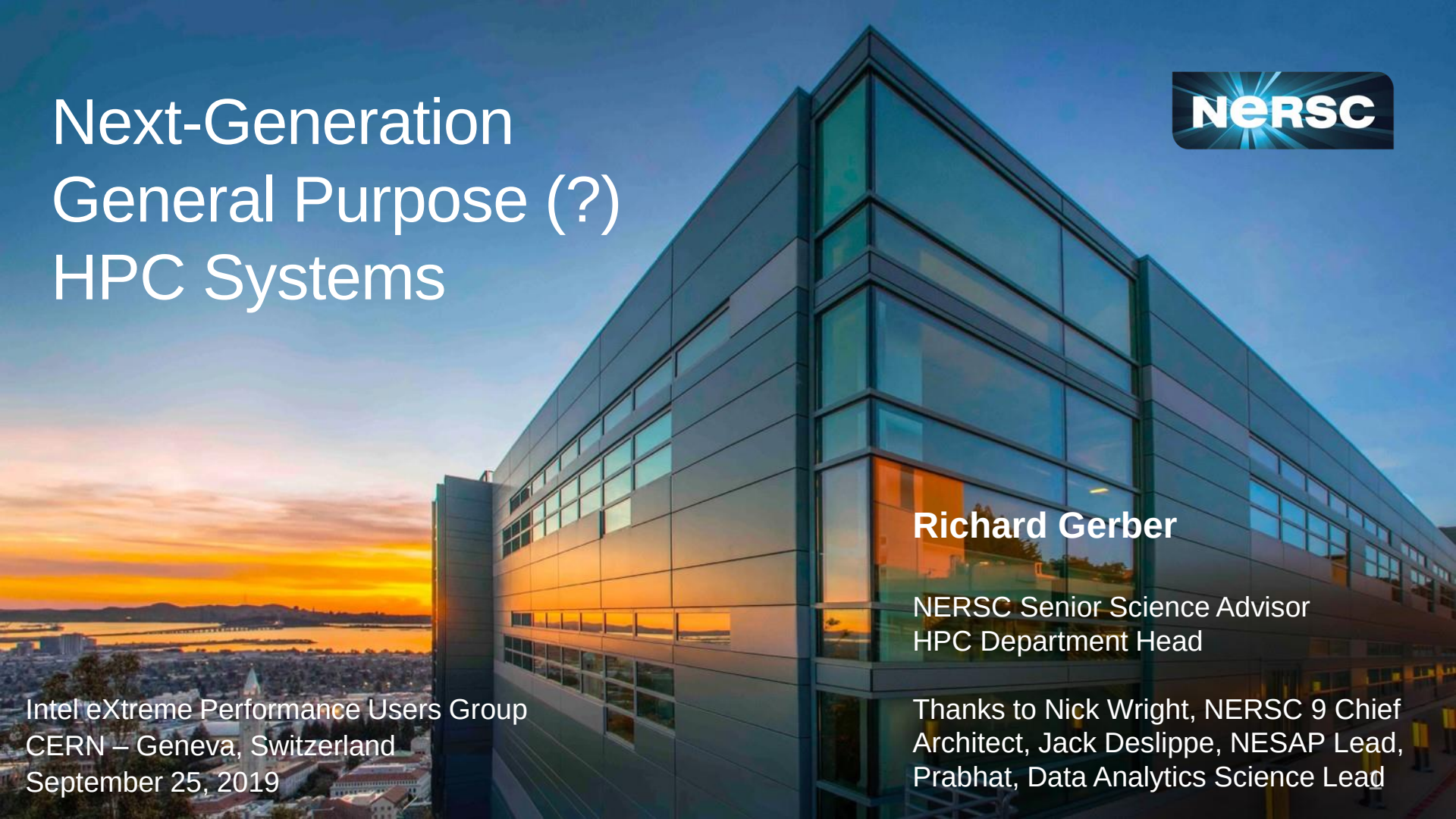
Next-Generation General Purpose (?) HPC Systems



Richard Gerber

NERSC Senior Science Advisor
HPC Department Head

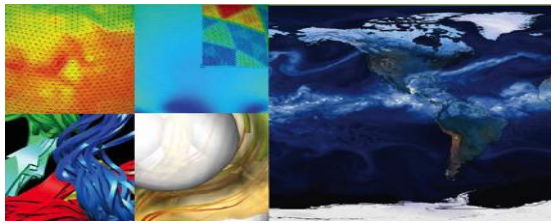
Intel eXtreme Performance Users Group
CERN – Geneva, Switzerland
September 25, 2019

Thanks to Nick Wright, NERSC 9 Chief Architect, Jack Deslippe, NESAP Lead, Prabhat, Data Analytics Science Lead

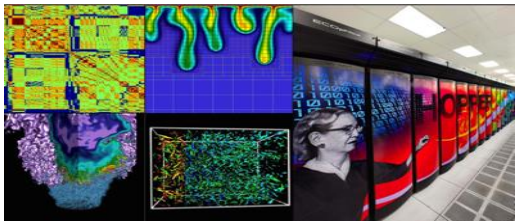
NERSC: Mission HPC for DOE Office of Science Research

U.S. DEPARTMENT OF ENERGY | Office of Science

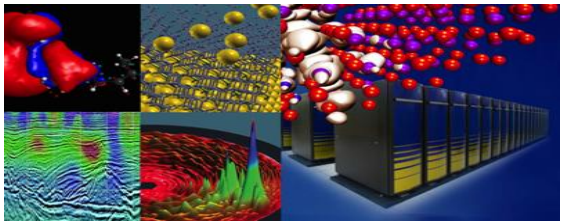
Largest funder of physical science research in the U.S. - $6.5B budget



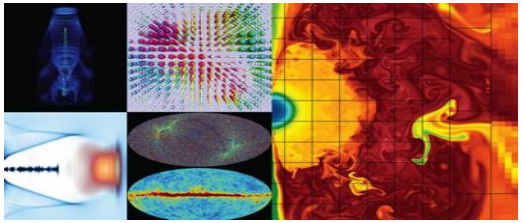
Bio Energy, Environment



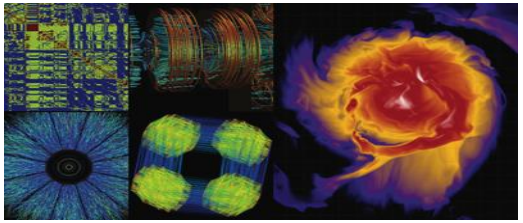
Computing



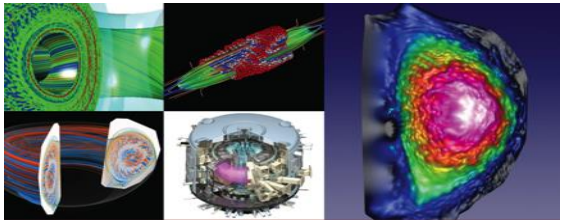
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



Nuclear Physics



Fusion Energy, Plasma Physics

7,000 users, 800 projects, 700 codes, 50 states, 40 countries, universities & national labs

Focus on Science

NERSC supports the broad mission needs of the six DOE Office of Science program offices

8,000 users and 900 projects

Simulation, data, learning

NERSC science engagement team provides outreach and POCs

6 Nobel Prize Winners

> 2,500 refereed publications in 2018

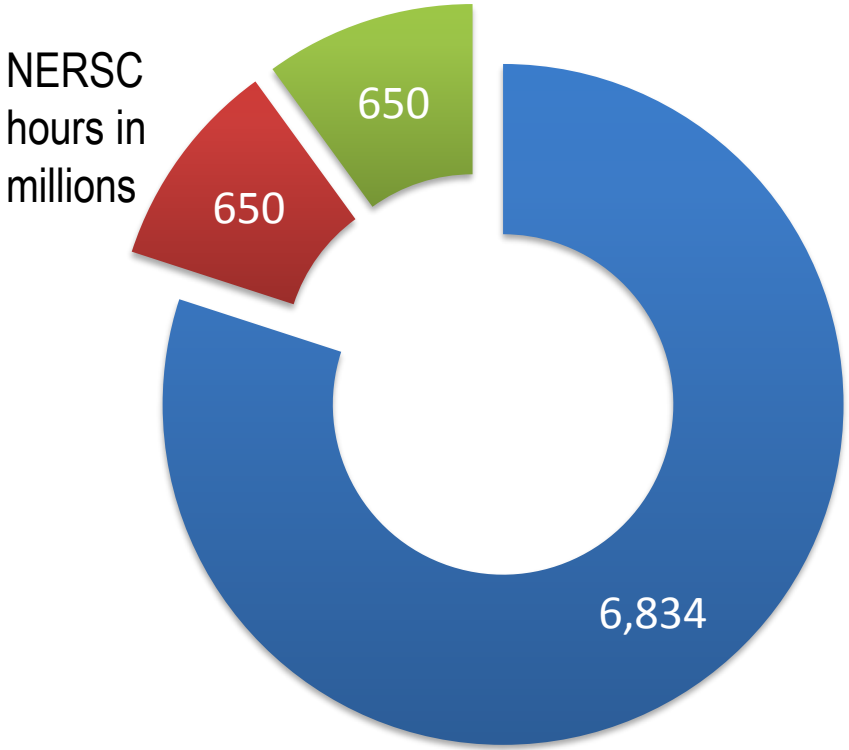
120 in Nature Journals

9 in Science

300 in Physical Review <X>

15 in PNAS

Allocation of Computing Time 2019



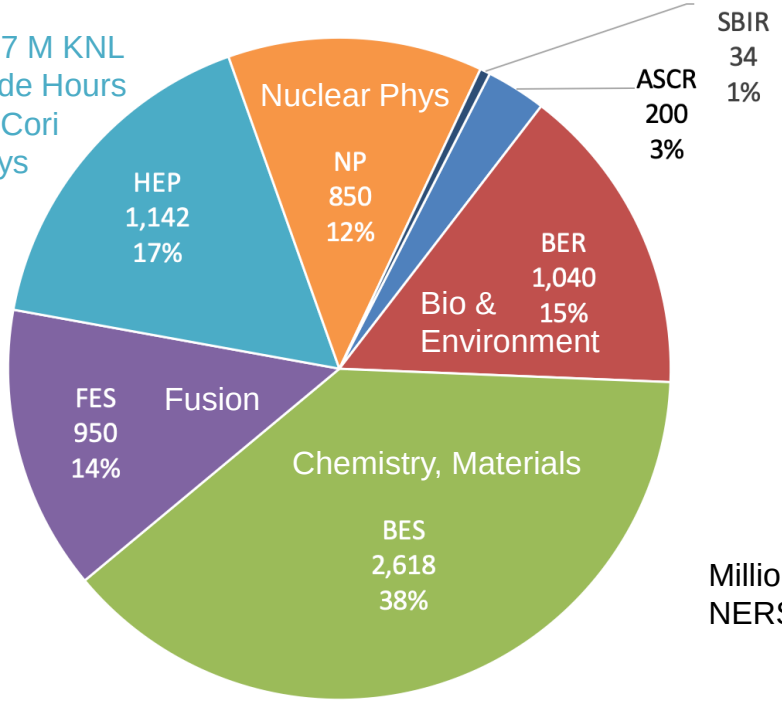
- **DOE Mission Science 80%**
  Distributed by DOE Office of Science program managers
- **ALCC 10%**
  Competitive awards run by DOE Advanced Scientific Computing Research Office
- **Directors Discretionary 10%**
  Strategic awards from NERSC

NERSC supports important mission science, not just projects that have optimized codes

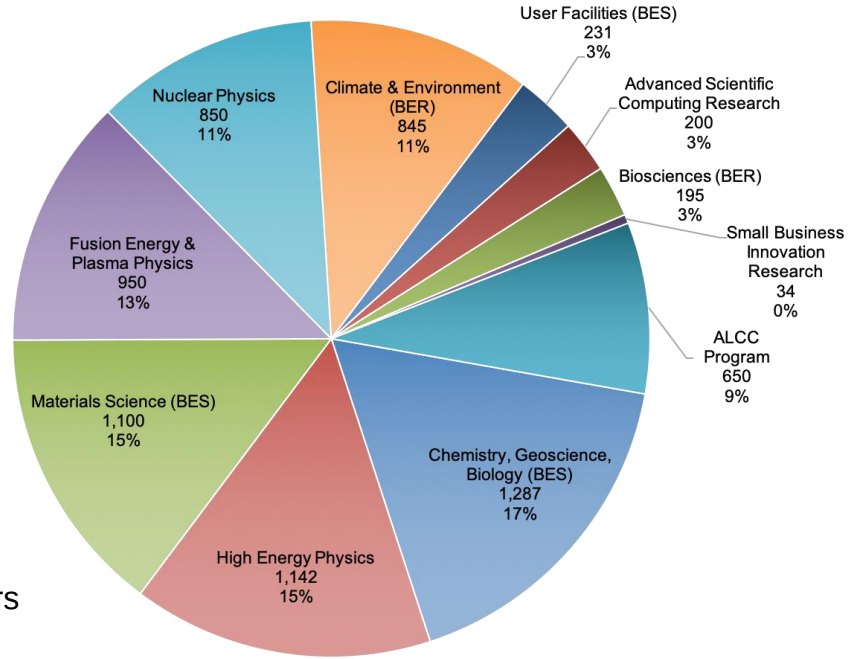
Allocation Breakdown

By Program Office

12.7 M KNL Node Hours
55 Cori Days

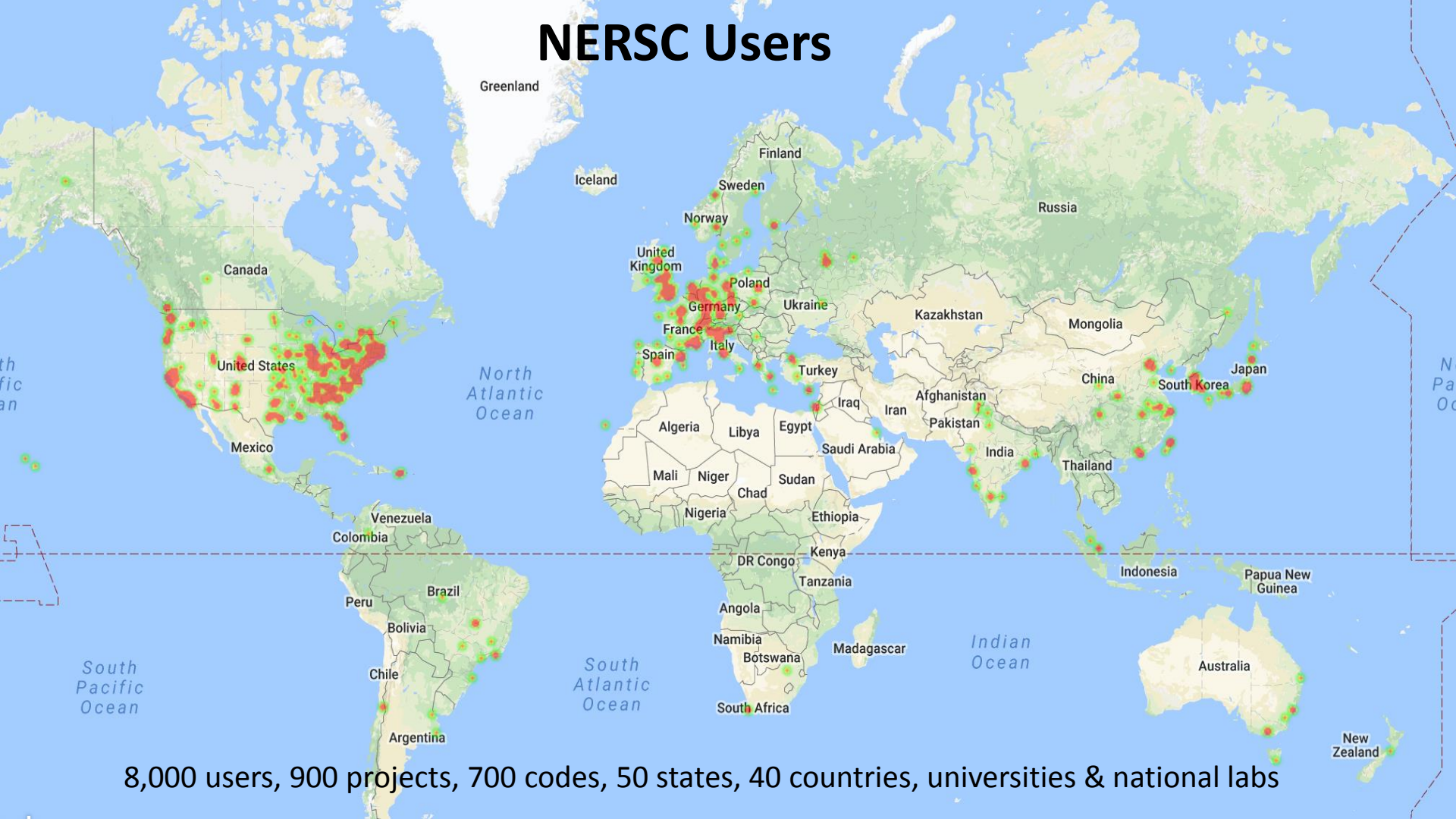


By Subprograms



Millions of NERSC Hours

NERSC Users



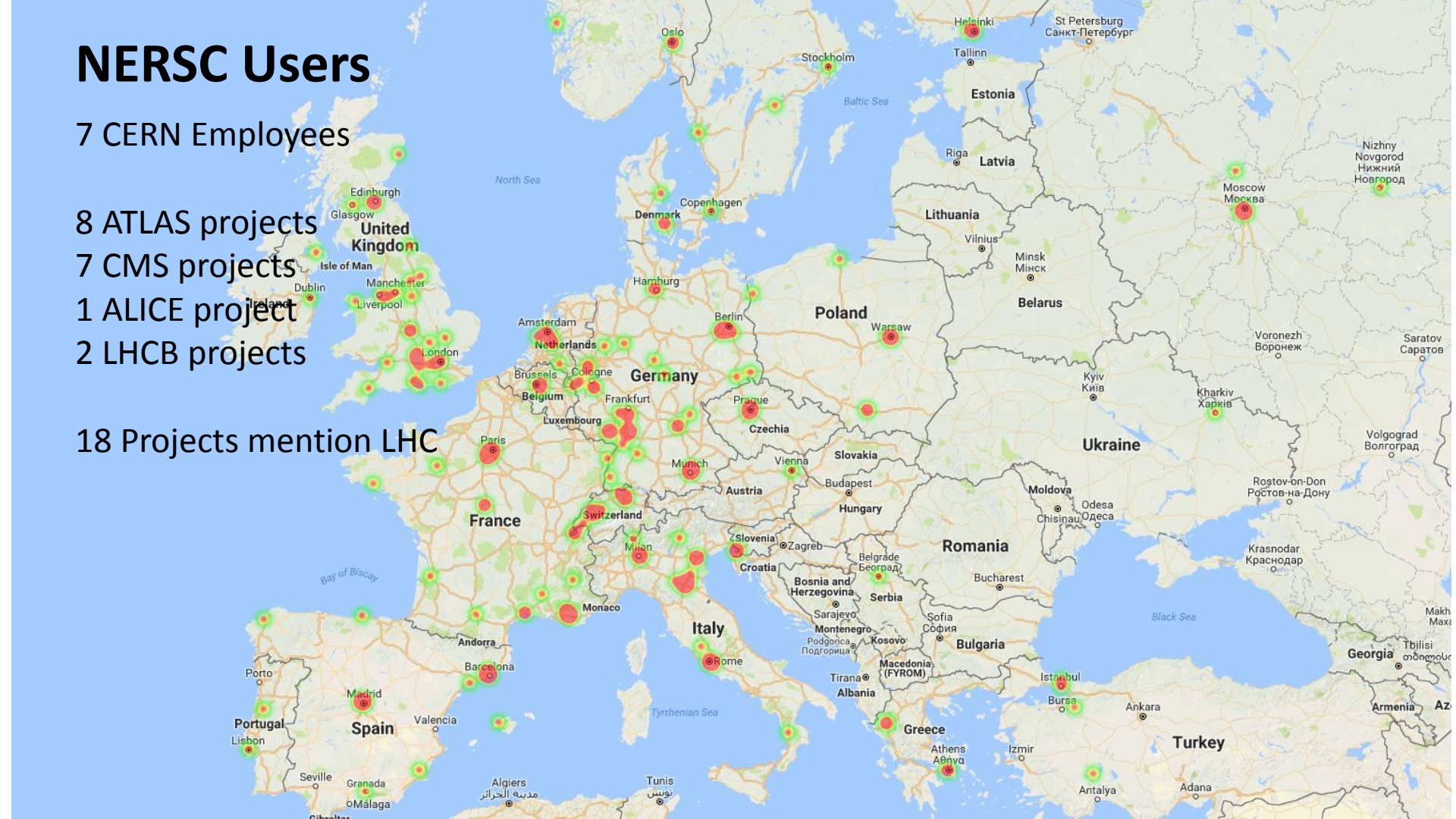
8,000 users, 900 projects, 700 codes, 50 states, 40 countries, universities & national labs

NERSC Users

7 CERN Employees

8 ATLAS projects
7 CMS projects
1 ALICE project
2 LHCB projects

18 Projects mention LHC

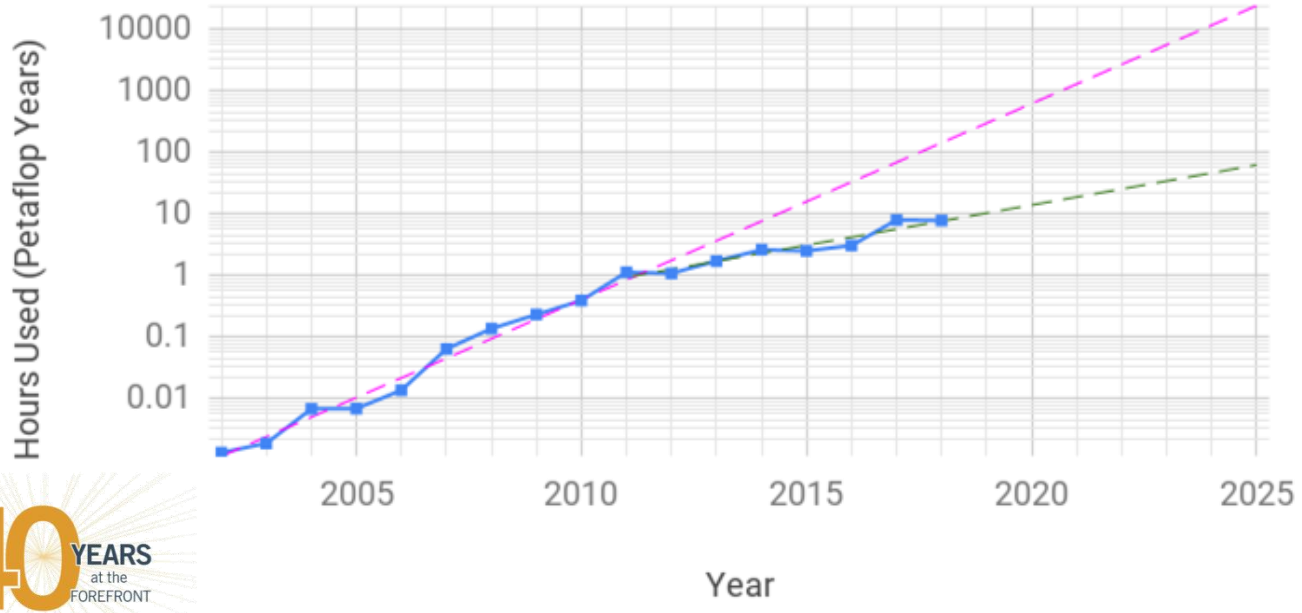


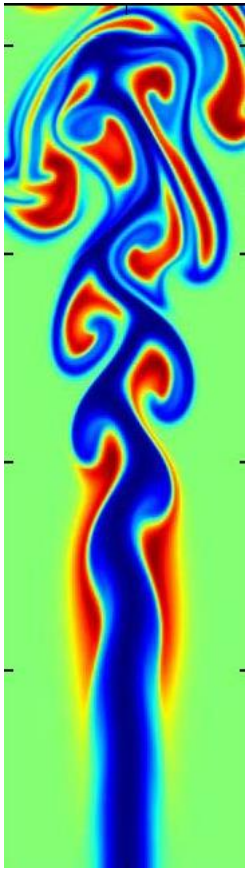
Top LHC Projects

PI	Org	Title	Hours
P. Calafiura	LBNL	Detector Simulation of the ATLAS Detector on NERSC HPCs	122 M
O. Gutsche, …	Fermilab	Enabling HEP Frontier Science through HEPCloud	147 M
S. Hoeche	SLAC	Precision QCD calculations and event simulation for the Large Hadron Collider	4.5 M
Z. Marshall	LBNL	Simulation and analysis of ATLAS Physics	1.0 M
J. Porter	LBNL	Data analysis and simulations for the ALICE experiment at the LHC	10 M
X. Ju	LBNL	Machine Learning for HEP Tracking at Extreme Scales	2.5 M
Z. Ligeti	LBNL	Theoretical Particle Physics Simulations for LHC Processes	1.8 M
N. Nachman	LBNL	Deep Learning for Jet Physics	1.0 M

NERSC Delivers HPC for Science

NERSC has provided exponentially increasing Top-10 HPC systems for science, largely used by simulation.





NERSC 2014 40 YEARS at the FOREFRONT

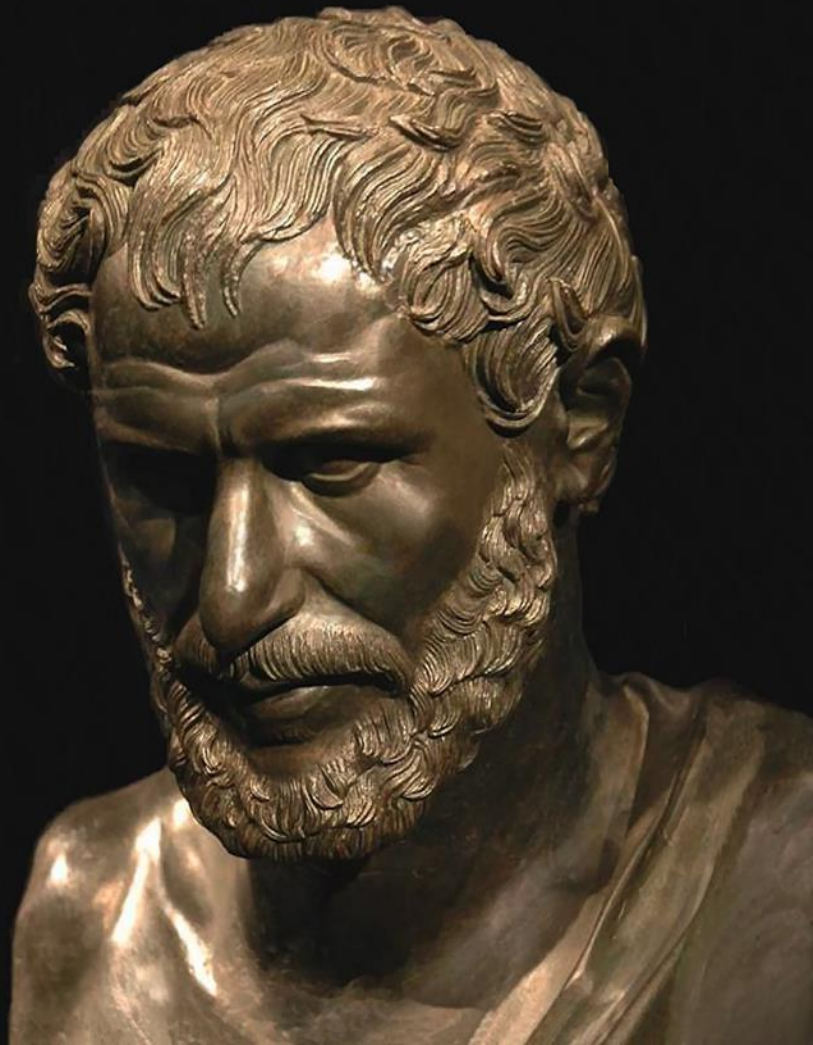
NERSC Has Been Supporting Data for a Long Time

- Required modest computing resources, e.g. PDSF (~2,000 CPUs)
- Leveraged NERSC's large data facilities & high speed network
- Detector simulation & data analysis
- Data intensive, high throughput workflows
- Grid Support

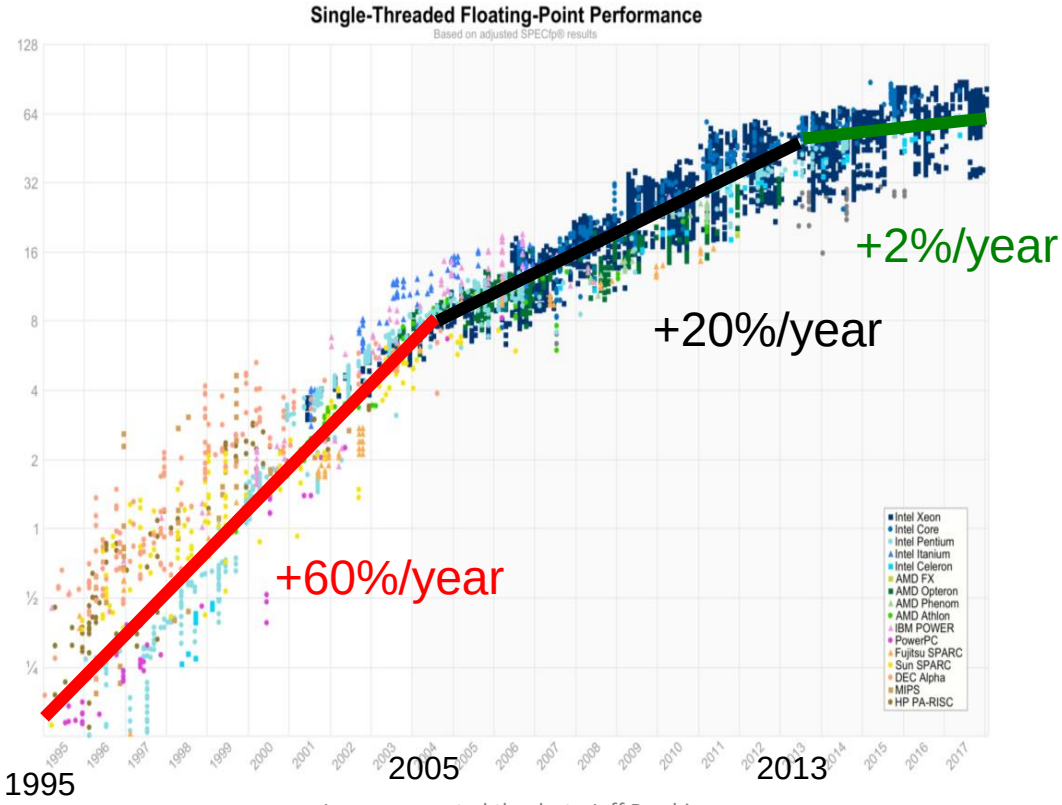
Today ~ 200 PB on tape, building a new >250 PB disk-based capacity system.

“The only thing constant is change”

–Heraclitus of Ephesus



Change #1: Processor Technology

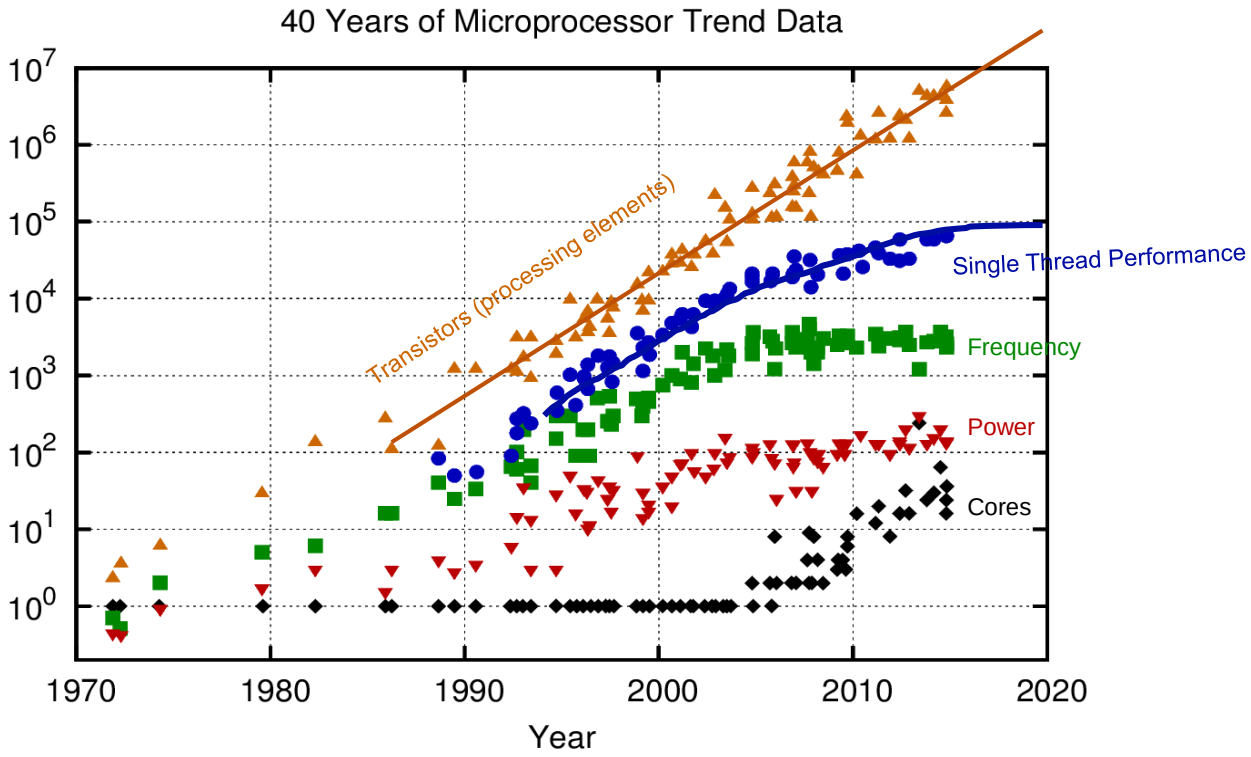


1995 2005 2013

- Single-thread (single-core) performance is flat, driven by power and heat dissipation constraints.
- Performance gains are being provided by combining light-weight, low-power cores on a single chip.
- Examples: Intel Xeon Phi “Knight’s Landing” and Graphical Processing Units (GPUs)
- Requires programming for fine-grained parallelism on many-core and/or GPU processors.

Image generated thanks to Jeff Preshing (https://github.com/preshing/analyze-spec-benchmarks)

Many-Core KNL Shook Things Up



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

MPI Parallelism

FLOP Optimization

Data Locality | Threading | Vectorization

Threading

Data Locality

Vectors

FLOPS | MPI

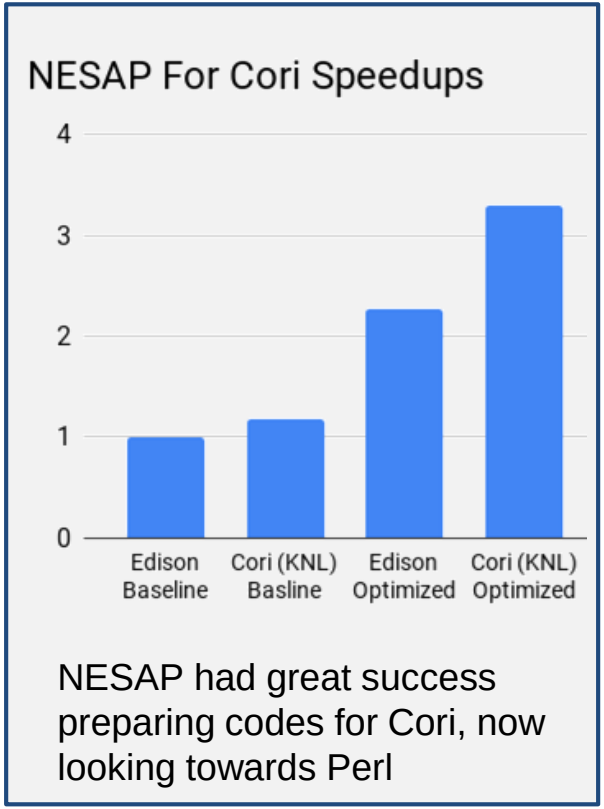
NESAP: Application Readiness

NESAP is NERSC's Application Readiness Program for new Systems.

Strategy: Partner with application teams representing ~50% of workload and vendors to optimize participating applications. Share lessons learned with with NERSC community via documentation and training.

Resource Available to Teams: NERSC Staff development effort and expertise, performance post-docs, access to vendor application engineers, hack-a-thons, early access to hardware (GPU nodes on Cori and Perlmutter)

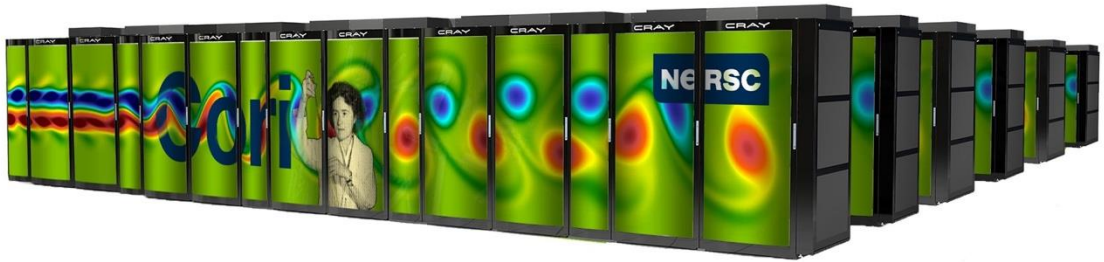
IXPUG
INTEL EXTREME PERFORMANCE USERS GROUP



NESAP had great success preparing codes for Cori, now looking towards Perl

NERSC's Cori System

9,300 Intel Xeon Phi "KNL" nodes
2,000 Intel Xeon "Haswell" nodes
700,000 processor cores
1.2 PB memory
Cray XC40 / Aries Dragonfly intercon
30 PB Lustre Cray Sonexion scratch
1.5 PB Burst Buffer @ 1.9 TB/sec



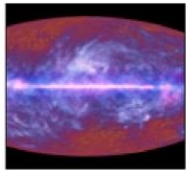
NESAP successfully enabled much of the workload on KNL

KNL node utilization ~100%

However, there remain a significant number of users who use Haswell only, many of them "data users"

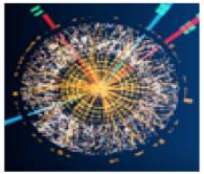
Change #2: Exploding Demand from Experimental Facilities

Palomar Transient Factory
Supernova

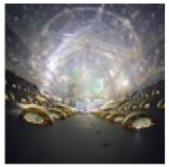


Planck Satellite
Cosmic Microwave Background
Radiation

Star
Particle Physics



Atlas
Large Hadron Collide

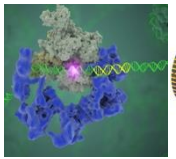


Dayabay
Neutrinos

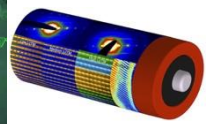
ALS
Light Source

LCLS
Light Source

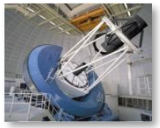
Joint Genome Institute
Bioinformatics



Cryo-EM

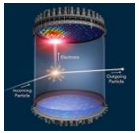


NCEM

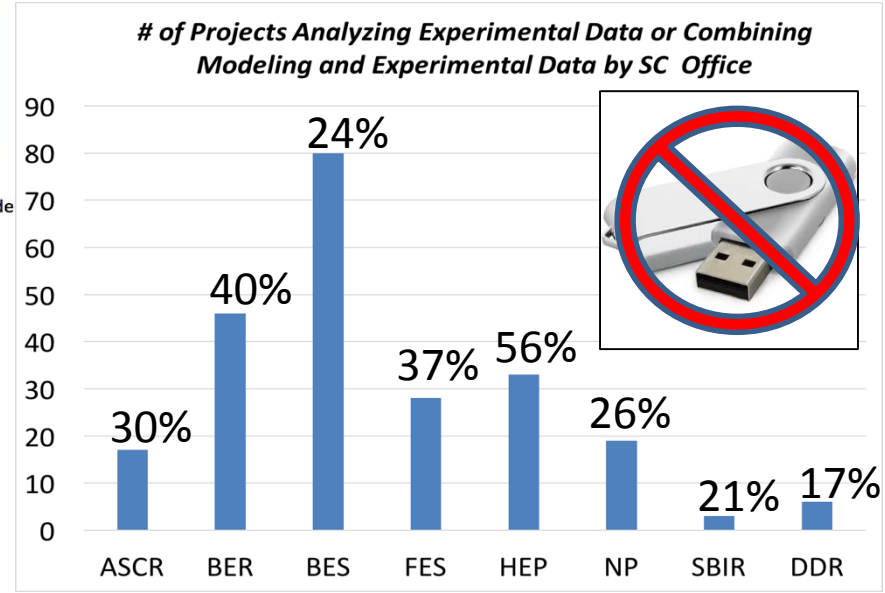


DESI

LSST-DESC



LZ



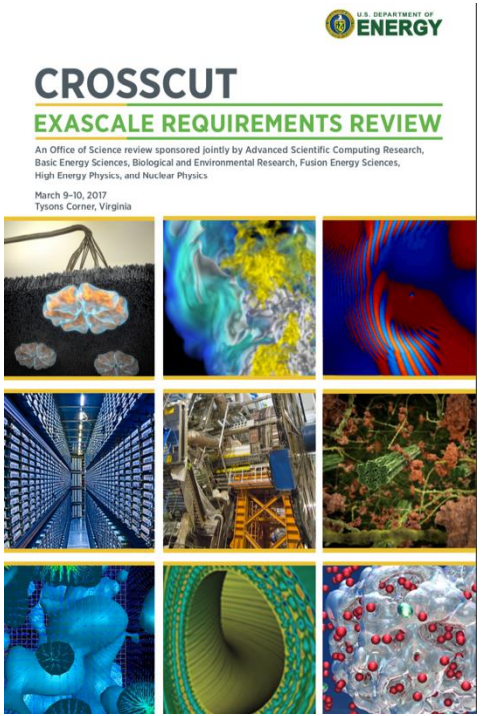
~35% (235) of ERCAP projects self identified as confirming the primary role of the project is to 1) analyze experimental data or; 2) create tools for experimental data analysis or; 3) combine experimental data with simulations and modeling

Exascale Requirements Reviews Key Data Findings

The volumes, the rapidity, and the complexity of data … require data management, archiving, and curation capabilities that will not be satisfied by the best practices employed today.

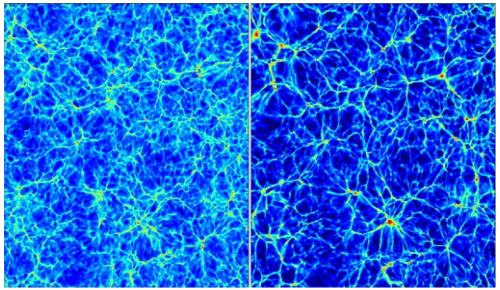
As workflows … become more complex, ASCR and other user facilities will be challenged to work out co-scheduling and data management approaches that will optimize the productivity of all facility resources.

The I/O capabilities of large HPC systems need to … grow faster. [Codes] cannot spend excessive amounts of time blocking on I/O, and data read/write rates can[not] limit performance of data analysis pipelines.

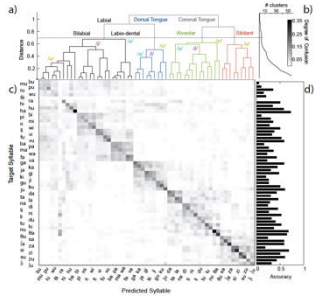


Change #3: Deep Learning (& AI) for Science

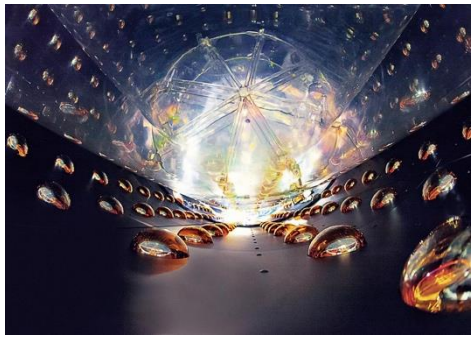
Modeling galaxy shapes



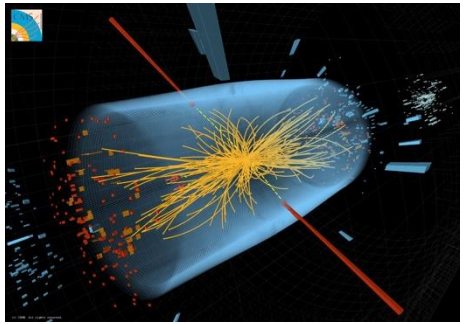
Generating cosmology mass maps



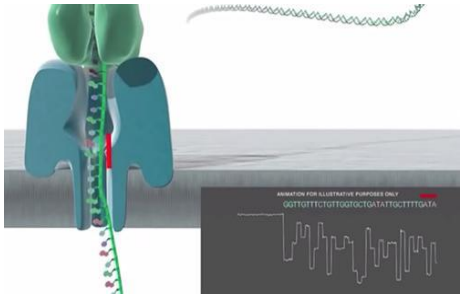
Decoding speech from ECoG



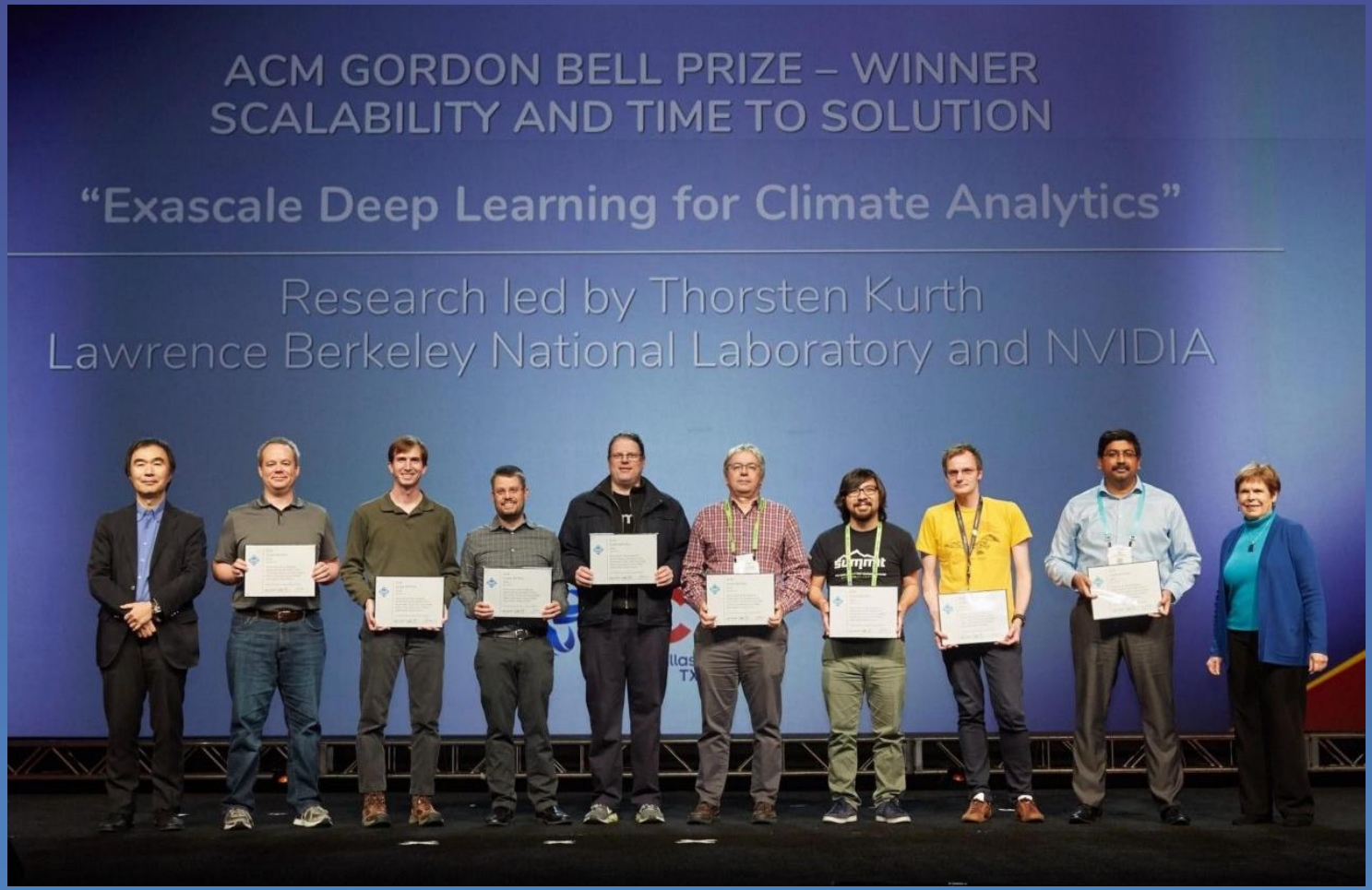
Clustering Daya Bay events



LHC Signal/Background classification



Oxford Nanopore sequencing

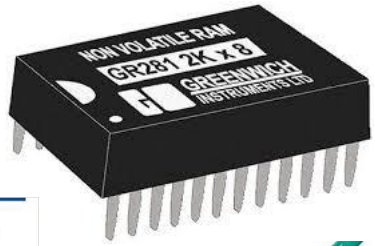


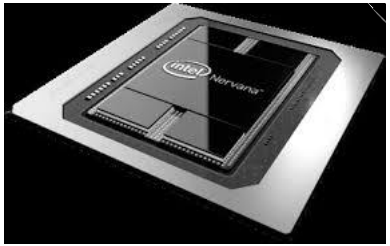
Contributors:
Thorsten Kurth,
Sean Treichler,
Joshua Romero,
Mayur Mudigonda,
Nathan Luehr,
Everett Phillips,
Ankur Mahesh,
Michael Matheson,
Jack Deslippe,
Massimiliano Fatica,
Michael Houston.
NERSC, NVIDIA,
UC Berkeley, OLCF

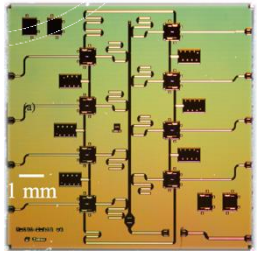
Change #4: Emerging Technologies

FireWorks

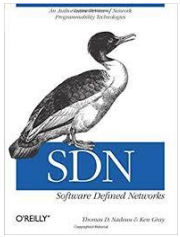
slurm
workload manager

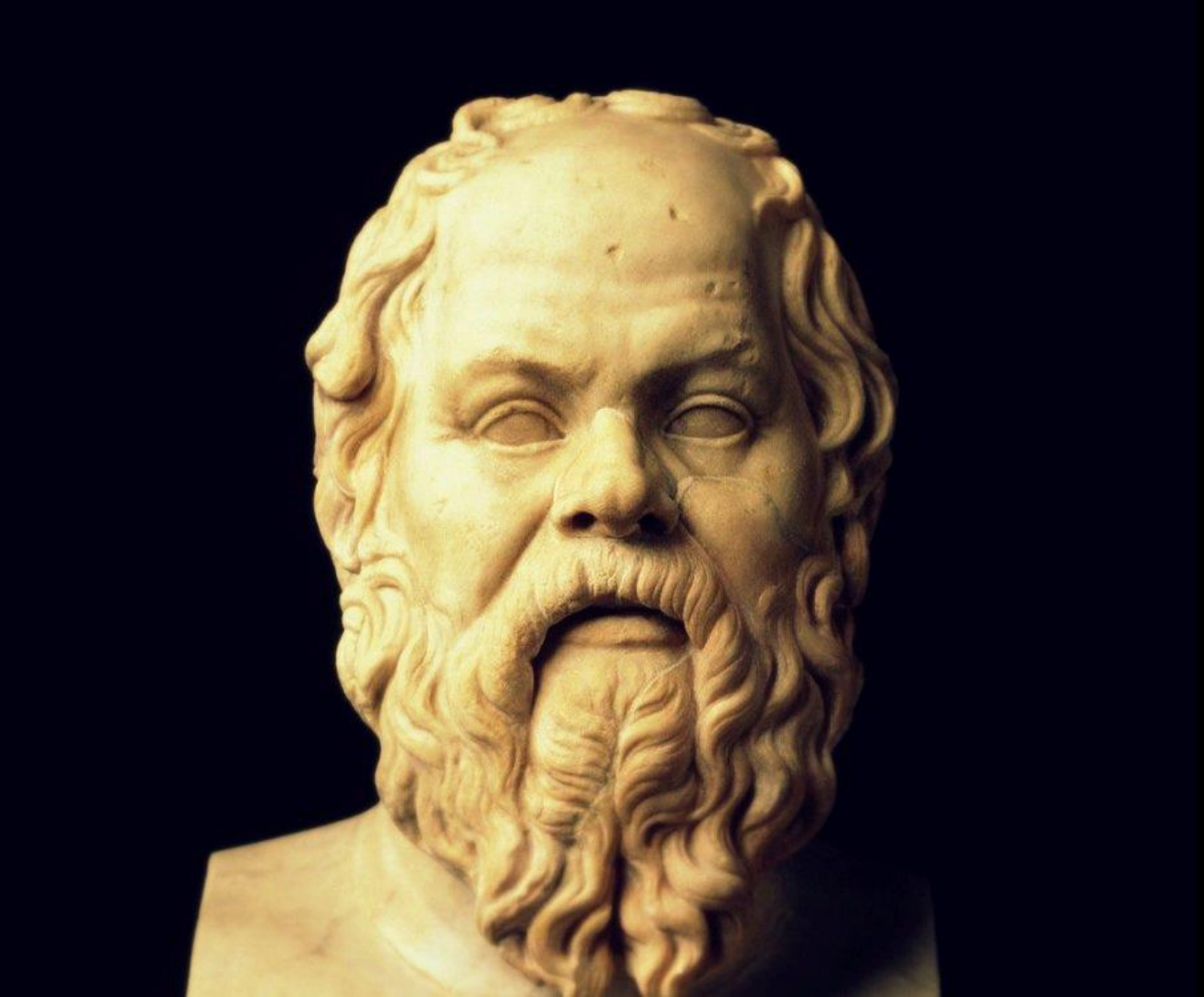






kubernetes

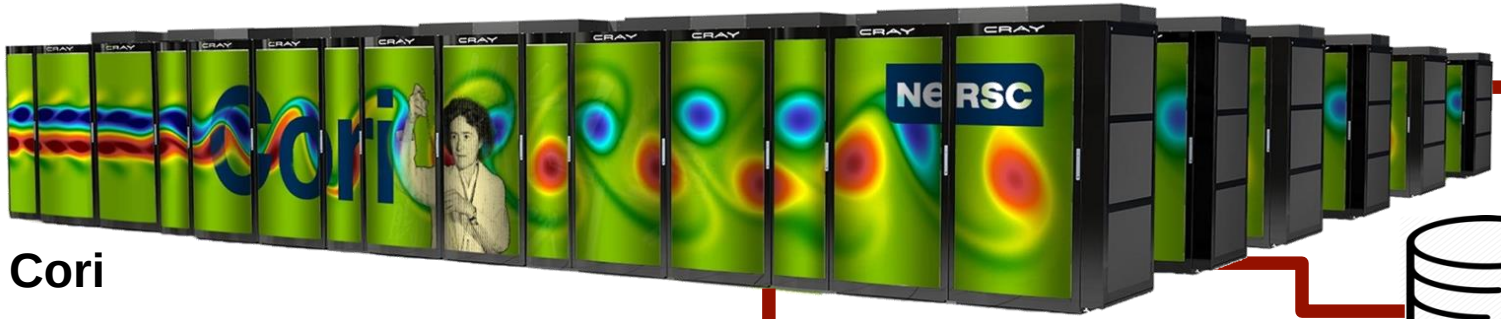




“The secret of change is to focus all of your energy, not on fighting the old, but on building the new.”

– Socrates

NERSC Starts New Class of Systems for Science



Cori

9,600 Intel Xeon Phi “KNL” manycore nodes
2,000 Intel Xeon “Haswell” nodes
700,000 processor cores, 1.2 PB memory
Cray XC40 / Aries Dragonfly interconnect

700 GB/s

1.5+ TB/s

2 PB
Burst Buffer

28 PB
Scratch

50 GB/s

150 GB/s

60-300 PB
/cfs

5 GB/s

275 TB
/home

DTNs, Spin, Gateways

ESnet
ENERGY SCIENCES NETWORK
2 x 10 Gb/s
2 x 100 Gb/s
SDN

Ethernet & IB Fabric
Science Friendly Security
Production Monitoring
Power Efficiency
WAN

Production Software Stack for Data & Learning

NERSC Data Department Formed

Data Analytics Services

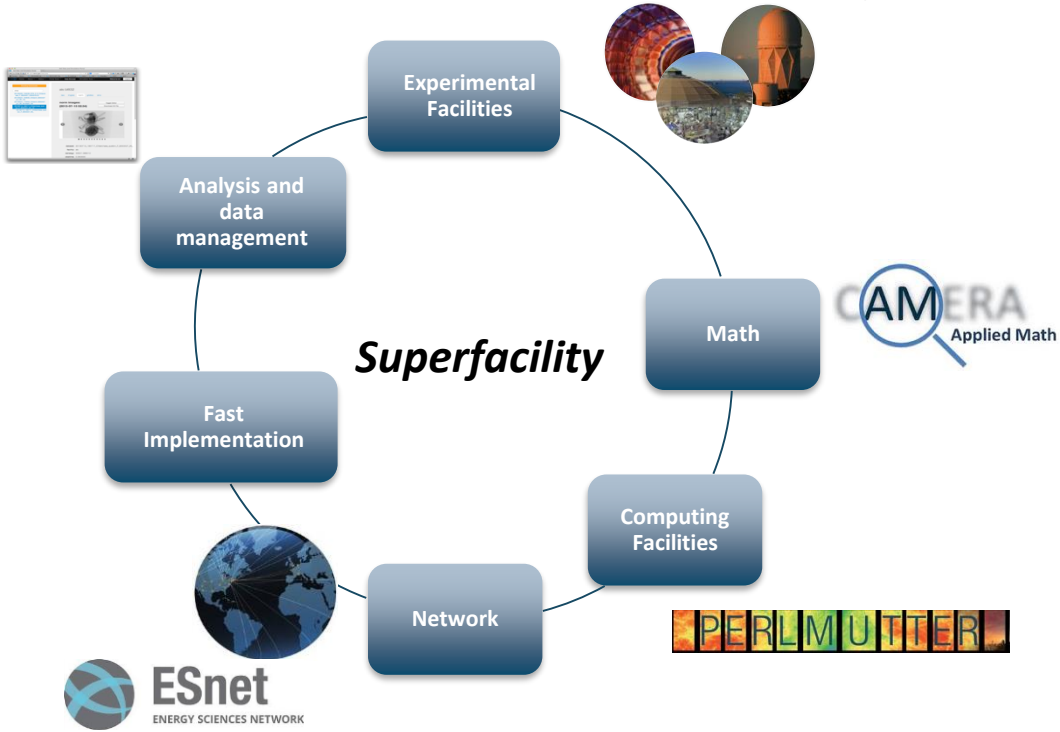
Data Engagement Group

NESAP for Data

Capabilities	Technologies
Data Transfer + Access	globus online, GridFTP, jupyter, IM, python django, newt
Workflows	FireWorks, TaskFarmer
Data Management	HDF, netCDF, ROOT Data Analysis Framework, mongoDB, MySQL, PostgreSQL
Data Analytics	python, Spark, TensorFlow, K, R, julia, caffe, PYTORCH
Data Visualization	VisIt, ParaView

Superfacility: A network of connected scientific facilities, software and expertise to enable new modes of discovery

- Large scale computing and storage resources
- Expertise optimizing pipelines & workflows
- Reusable building blocks & APIs
- Advanced scheduling and resource allocation
- Scalable infrastructure to launch services
- Edge services and external connectivity



NERSC's Next System: Optimized for Science

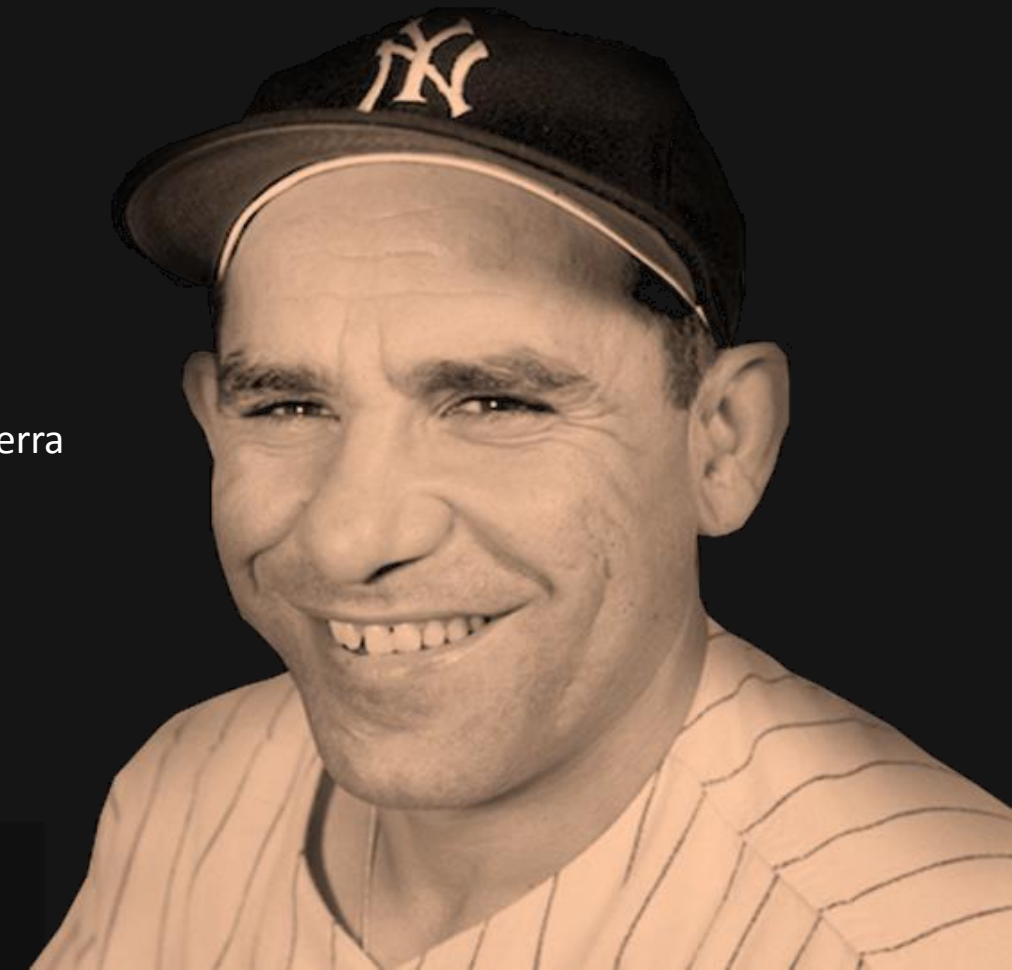
- Cray Shasta System providing 3-4x capability of Cori
- First NERSC system designed to meet needs of both large scale simulation and data analysis from experimental facilities
  - Includes both NVIDIA GPU-accelerated and AMD CPU-only nodes
  - Cray Slingshot ethernet-compatible network will support TB connectivity
  - Optimized data software stack enabling analytics and ML at scale
  - All-Flash scratch filesystem for I/O acceleration
- Robust readiness program for simulation, data and learning applications and complex workflows
- Delivery in late 2020

What will be next?

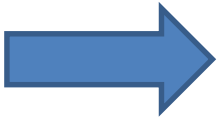
- Requires supporting traditional diverse NERSC workload while delivering an increase computing capability
- Must support Superfacility and data analysis needs
- Needs to support emerging AI / Learning paradigms

“It's tough to make predictions, especially about the future”

–Yogi Berra



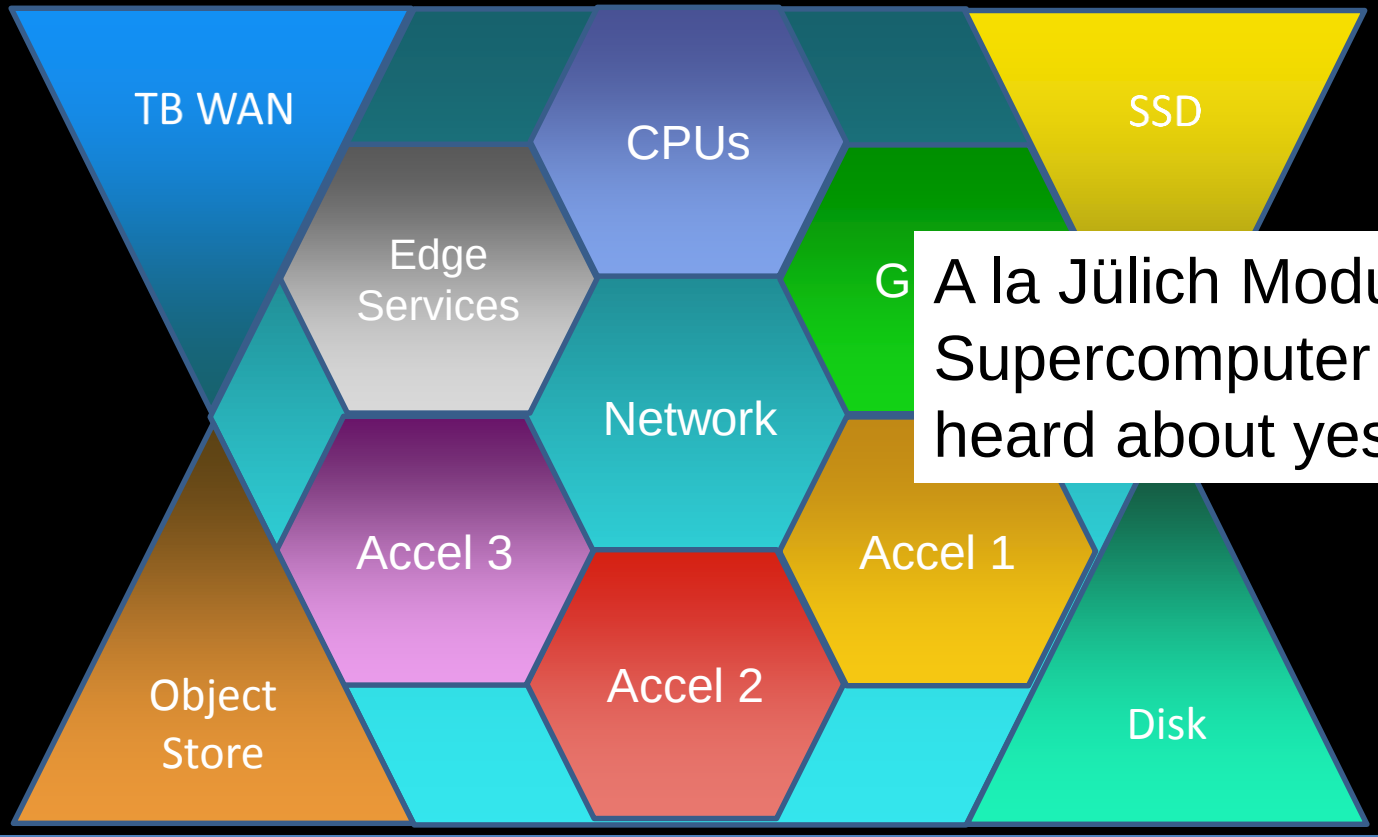

HERO
Find your inner hero.
RUN



Next-Generation System (~2025)

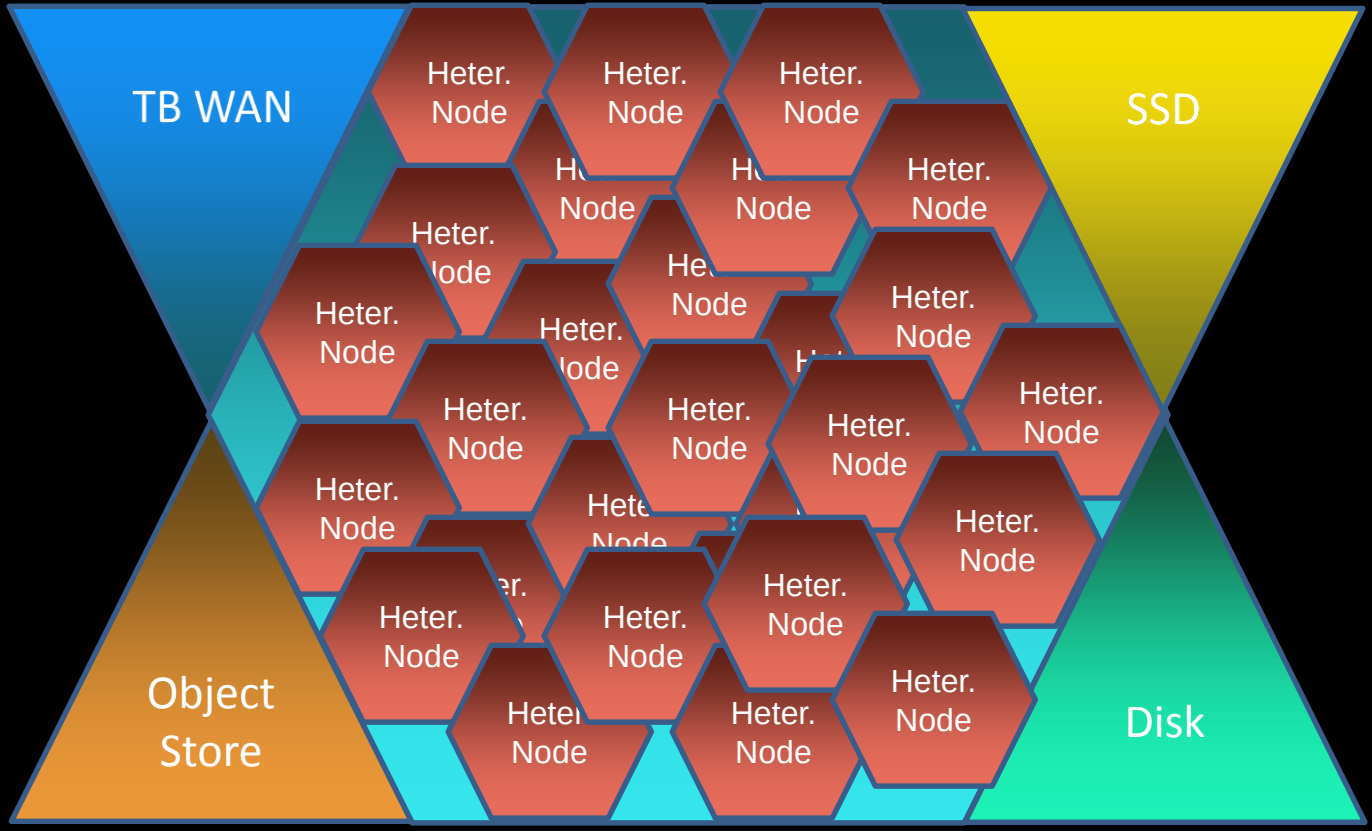
- Specialized energy-efficient heterogeneity in processing units
- Improved memory hierarchy (CPU to permanent storage)
- Advanced data movement to/from to compute engines
  - Real-time feedback for co-scheduled experiments
  - Fast analysis for followup
- Advanced resource scheduling
  - Co-scheduling of individual heterogenous resources
  - Quality of service for components
- Containers or similar for portability and reproducibility
- Productive software stack with workflow and framework support

Heterogenous System of Homogenous Parts

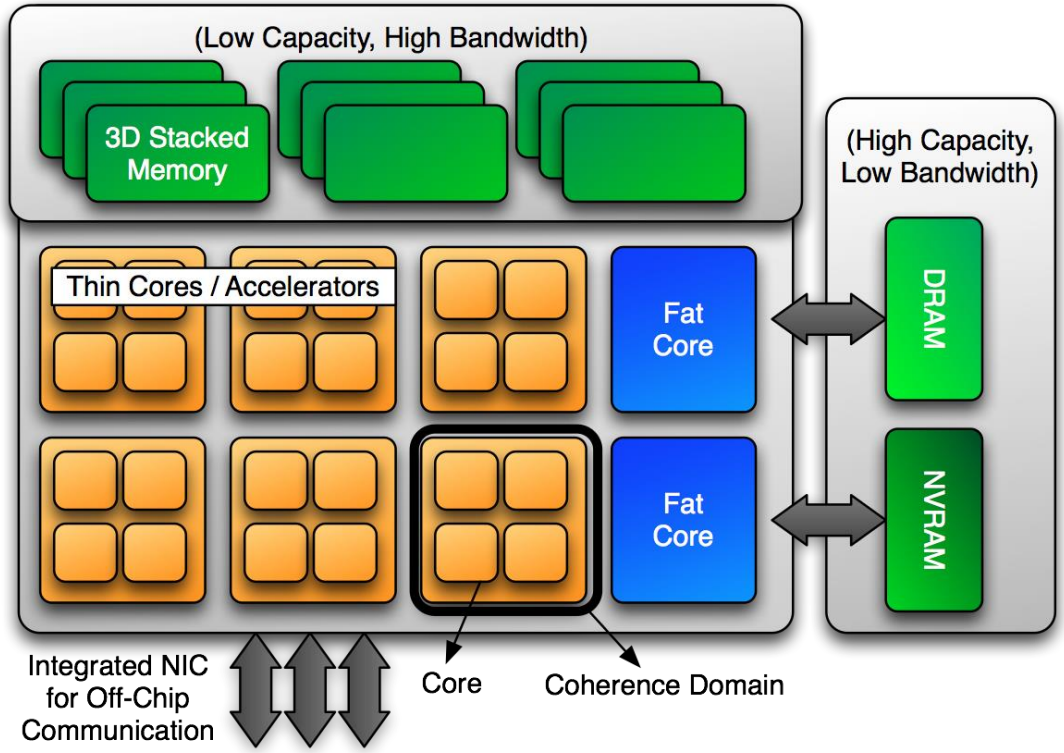


A la Jülich Modular Supercomputer we heard about yesterday

Homogeneous System of Heterogeneous Nodes



On-Node Heterogeneity



Can this be a general-purpose processing node?

Portable Performance Programming

I have no idea what's going to prevail!

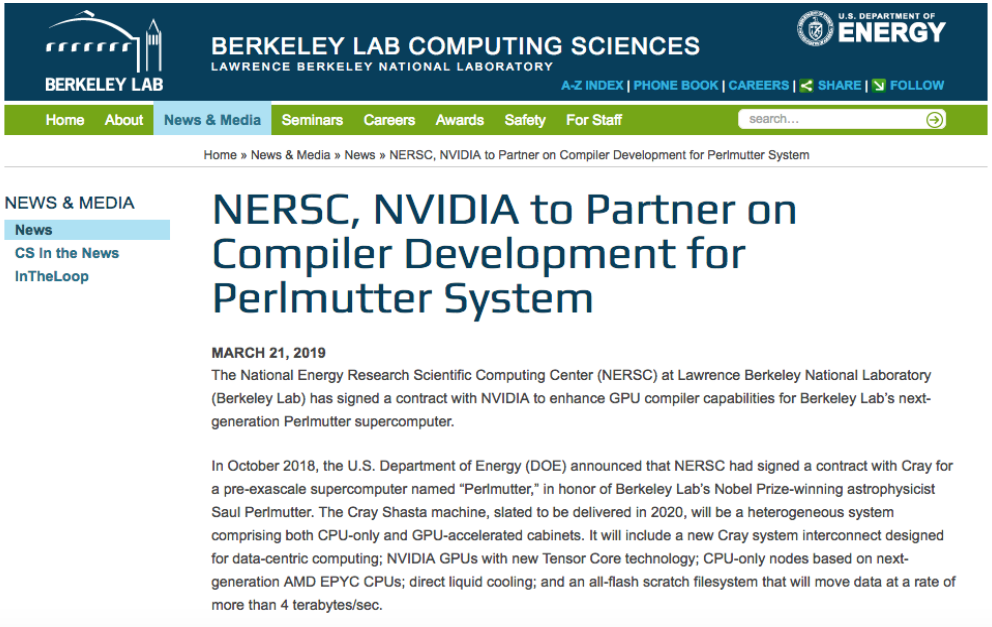
But we're working for you

OpenMP for Productivity

- 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 codes that run at NERSC
- Recent release of OpenMP 5.0 specification – the third version providing features for accelerators
  - Many refinements over this five year period
- But will OpenMP work with future accelerators?

NRE partnership with PGI/NVIDIA

- Agreed upon subset of OpenMP features to be included in the PGI compiler
- OpenMP test suite created with micro-benchmarks, mini-apps, and the ECP SOLLVE V&V suite
- 5 NESAP application teams will partner with PGI to add OpenMP target offload directives
- Alpha compiler is due soon - Limited access. Closed Beta - Apr 2020 and Open Beta – Oct 2020 - Greater access
- We want to hear from the larger community. Tell us your experience, including what OpenMP techniques worked / failed on the GPU.



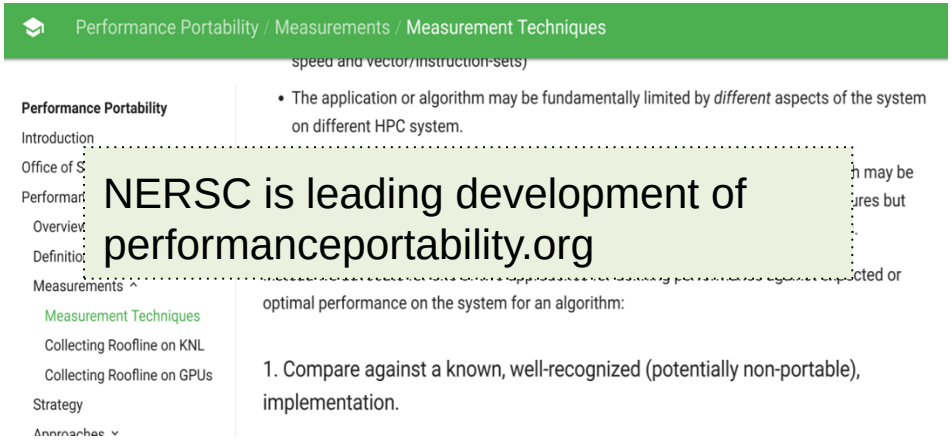
Engaging around Performance Portability

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

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

NERSC Hosted Fortran Standards Meeting

NERSC recently joined as a Member



NERSC is leading development of performanceportability.org

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

NERSC Systems Roadmap

NERSC-7:
Edison
Multicore
CPU

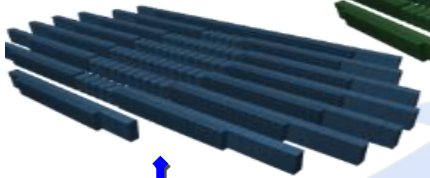
2013

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

2016

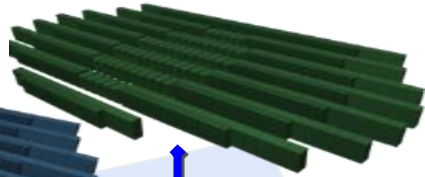
NERSC-9:
CPU and GPU nodes
Continued transition of
applications and support for
complex workflows

2020



NERSC-10:
Exa system

2024



NERSC-11:
Beyond
Moore

2028

Future

- Large monolithic simulation gives way to enabling end-to-end complex workflows on heterogenous systems
- Can the “General Purpose” HPC center continue to exist?
- Yes! – HPC centers retain their unique roles by providing integrated systems designed for science, providing capabilities not available anywhere else