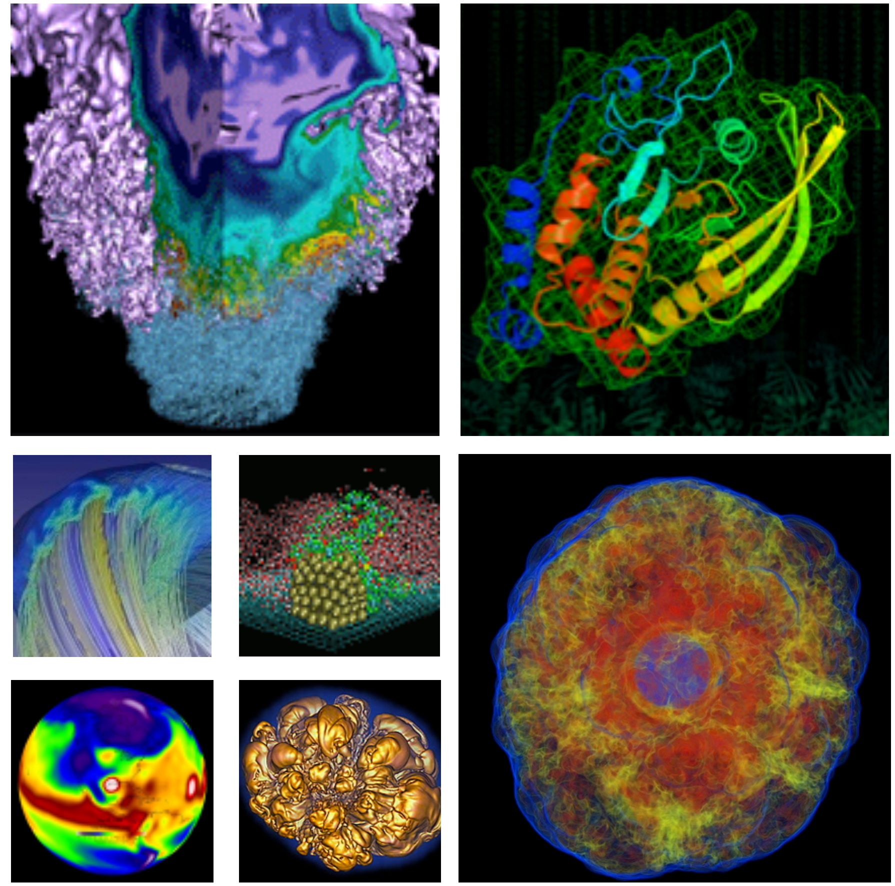
ASCR Facility Plans

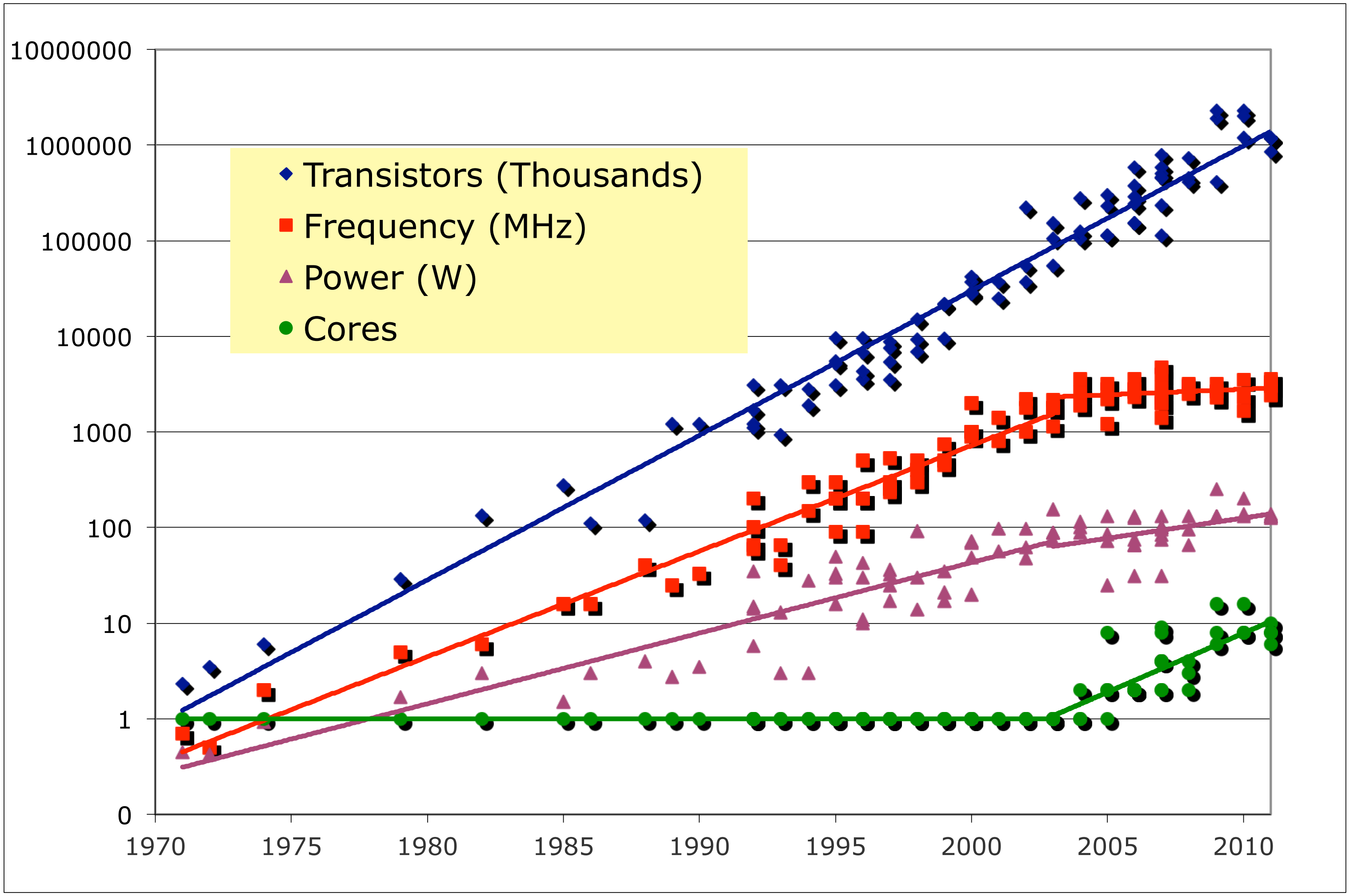


NERSC 40 YEARS at the FOREFRONT 1974-2014

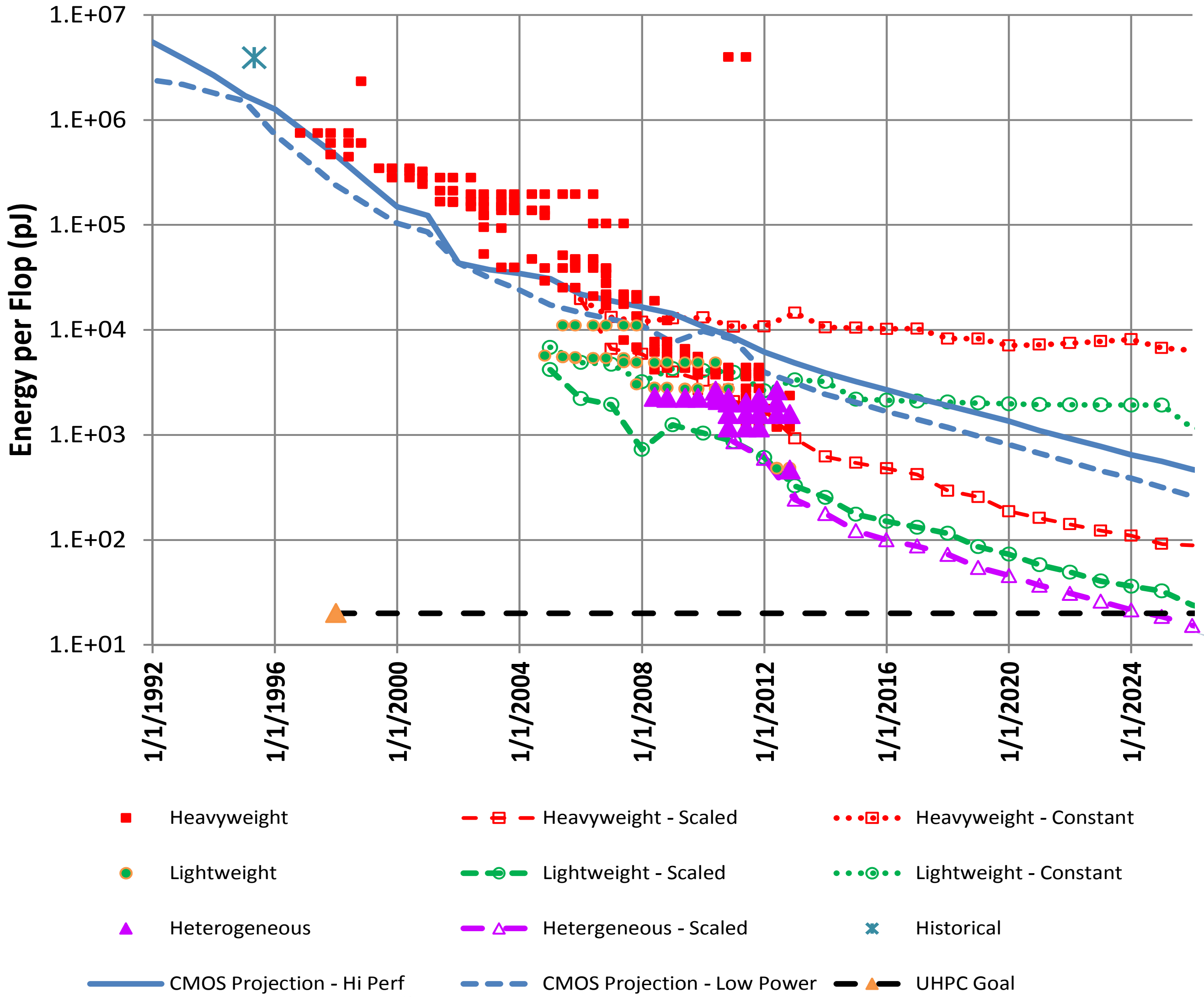
Sudip Dosanjh
Director

June 10, 2015

Projections on Moore's Law and other trends

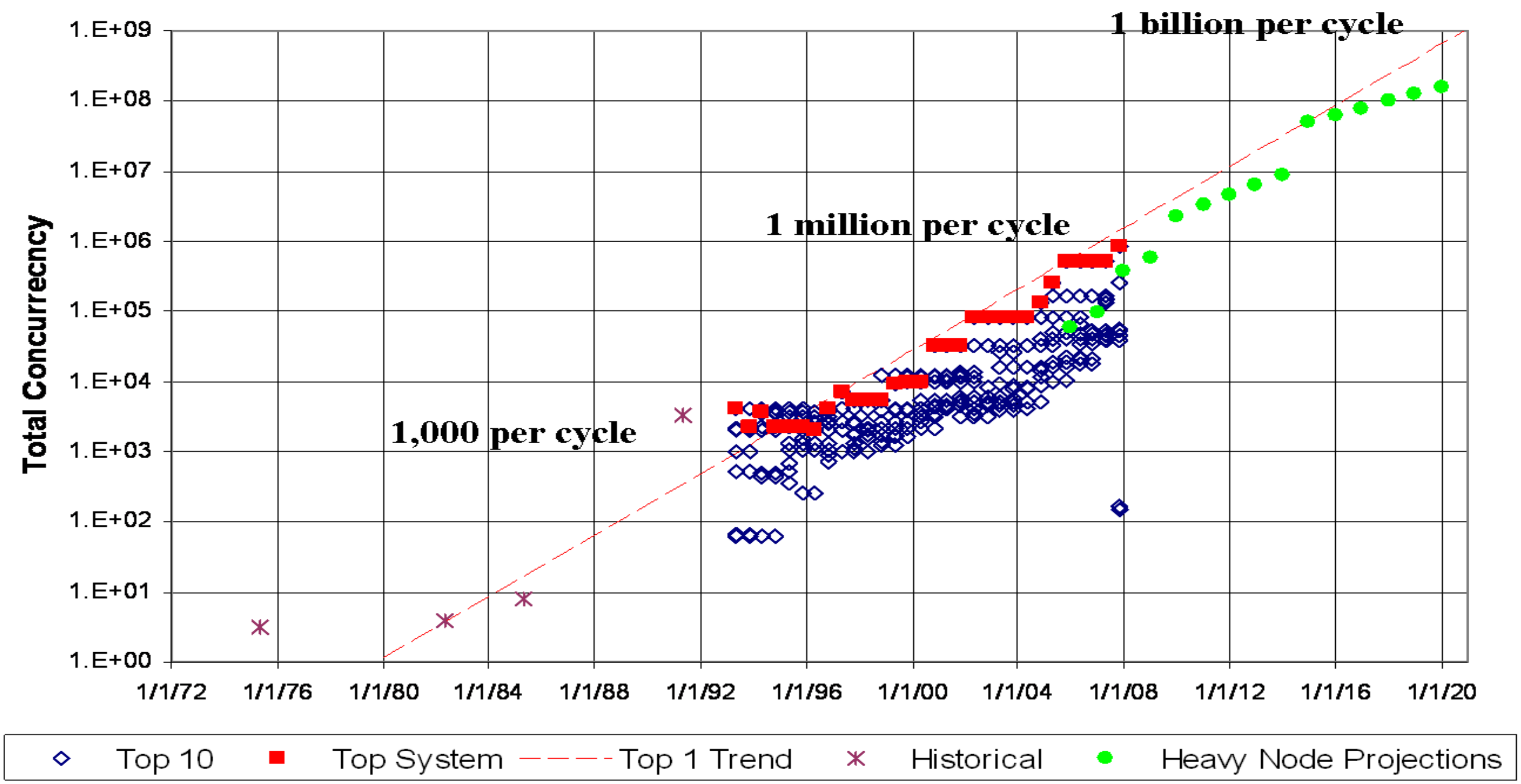


We need to transition to energy efficient architectures



Manycore or Hybrid is the only approach that crosses the exascale finish line

Projected Parallelism for Exascale

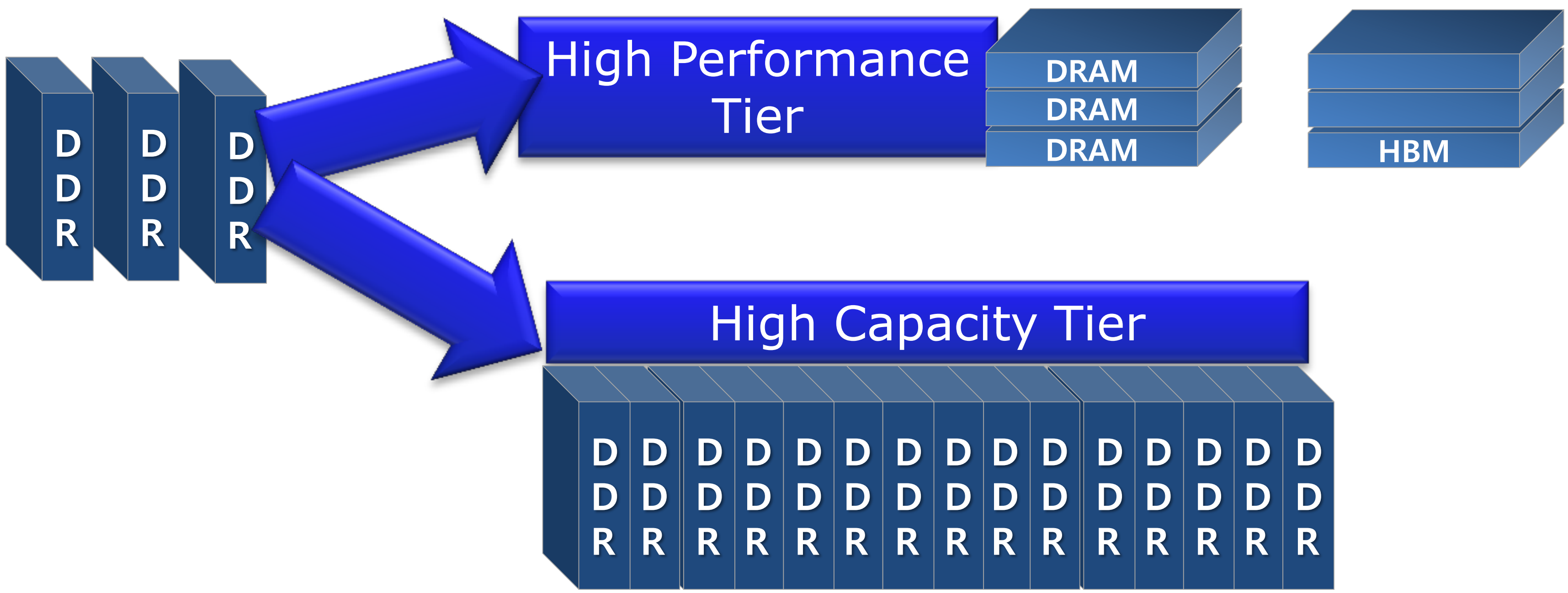


Can Get Capacity *OR* Bandwidth
But Cannot Get Both in the Same Technology

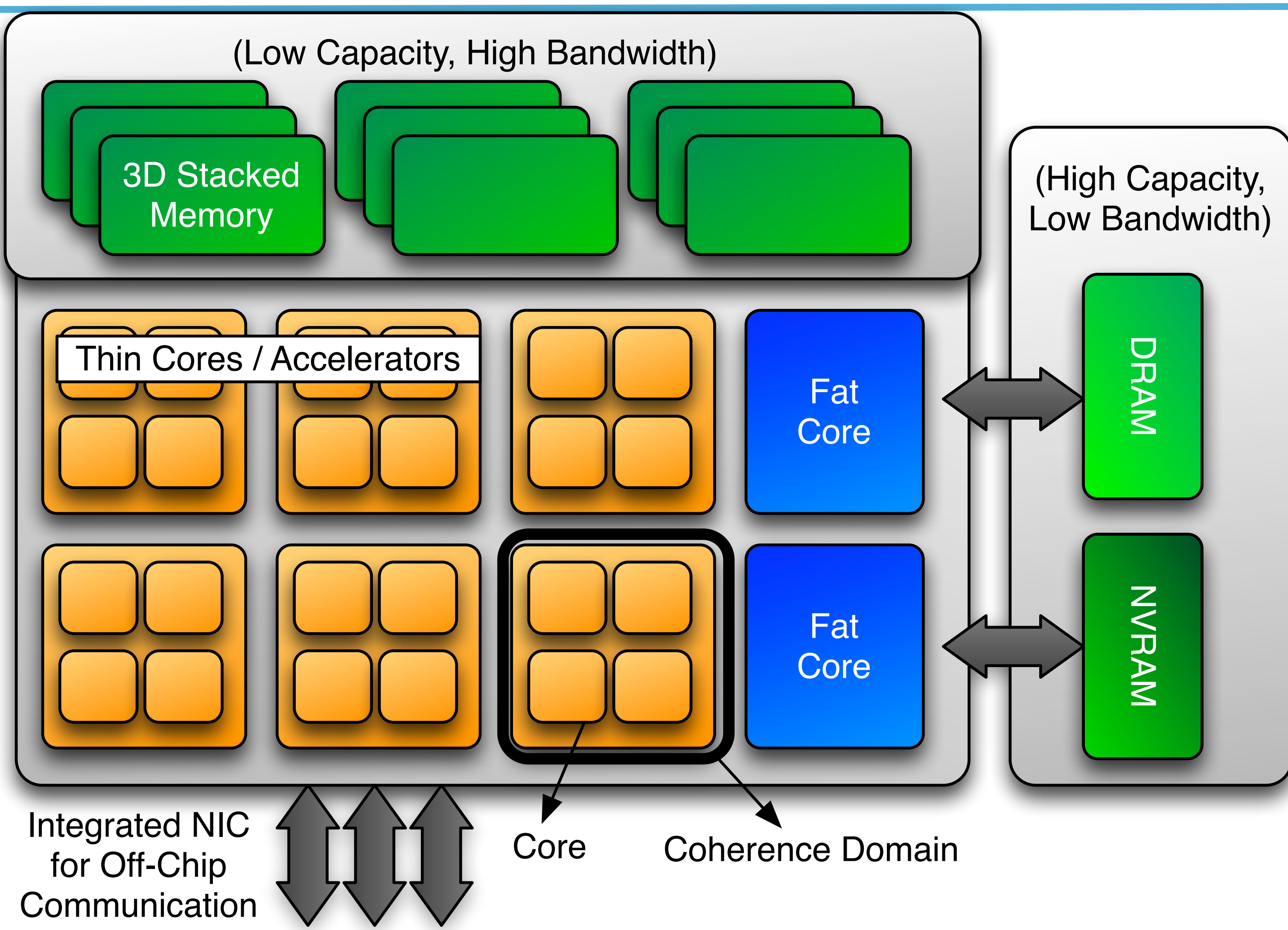
Cost (increases for higher capacity and cost/bit increases with bandwidth)

Power

Bandwidth\Capacity	16 GB	32 GB	64 GB	128 GB	256 GB	512 GB	1 TB
4 TB/s							
2 TB/s	Stack/PNM						
1 TB/s			Interposer				
512 GB/s				HMC organic			
256 GB/s					DIMM		
128 GB/s							NVRAM

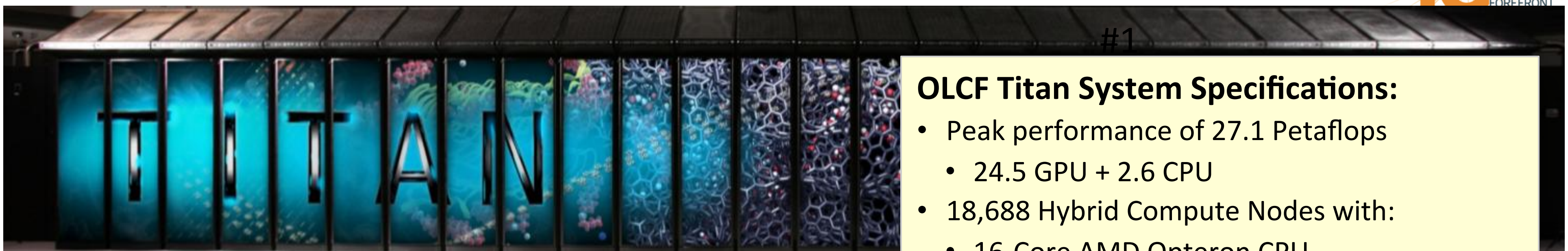


Abstract Machine Model



Leadership Computing for Scientific Discovery

NERSC 40 YEARS at the FOREFRONT



#1

OLCF Titan System Specifications:
- Peak performance of 27.1 Petaflops
  - 24.5 GPU + 2.6 CPU
- 18,688 Hybrid Compute Nodes with:
  - 16-Core AMD Opteron CPU
  - NVIDIA Tesla “K20x” GPU
  - 32 + 6 GB memory
- 200 Cabinets; 710 TB total system memory; 8.9 MW peak power

- Peer reviewed projects are chosen to advance science, promote innovation, and strengthen industrial competitiveness.
- Demand for these machines has grown each year, requiring recent upgrades of both.

ALCF Mira System Specifications:
- Peak performance of 10 Petaflops
- 49,152 Compute Nodes each with:
  - 16-Core Power PC A2 CPU with 64 Hardware Threads and 16 Quad FPUs
  - 16 GB memory
- 56 Cabinets; 786 TB total system memory; 4.8 MW peak power

FY 2013 research projects include; advancing materials for lithium air batteries, solar cells, and superconductors ; improving combustion in fuel-efficient, near-zero-emissions systems; understanding how turbulence affects the efficiency of aircraft and other transportation systems; designing next-generation nuclear reactors and fuels ; developing fusion energy systems

NERSC: 40 years of High Performance Computing for DOE

System Specifications:

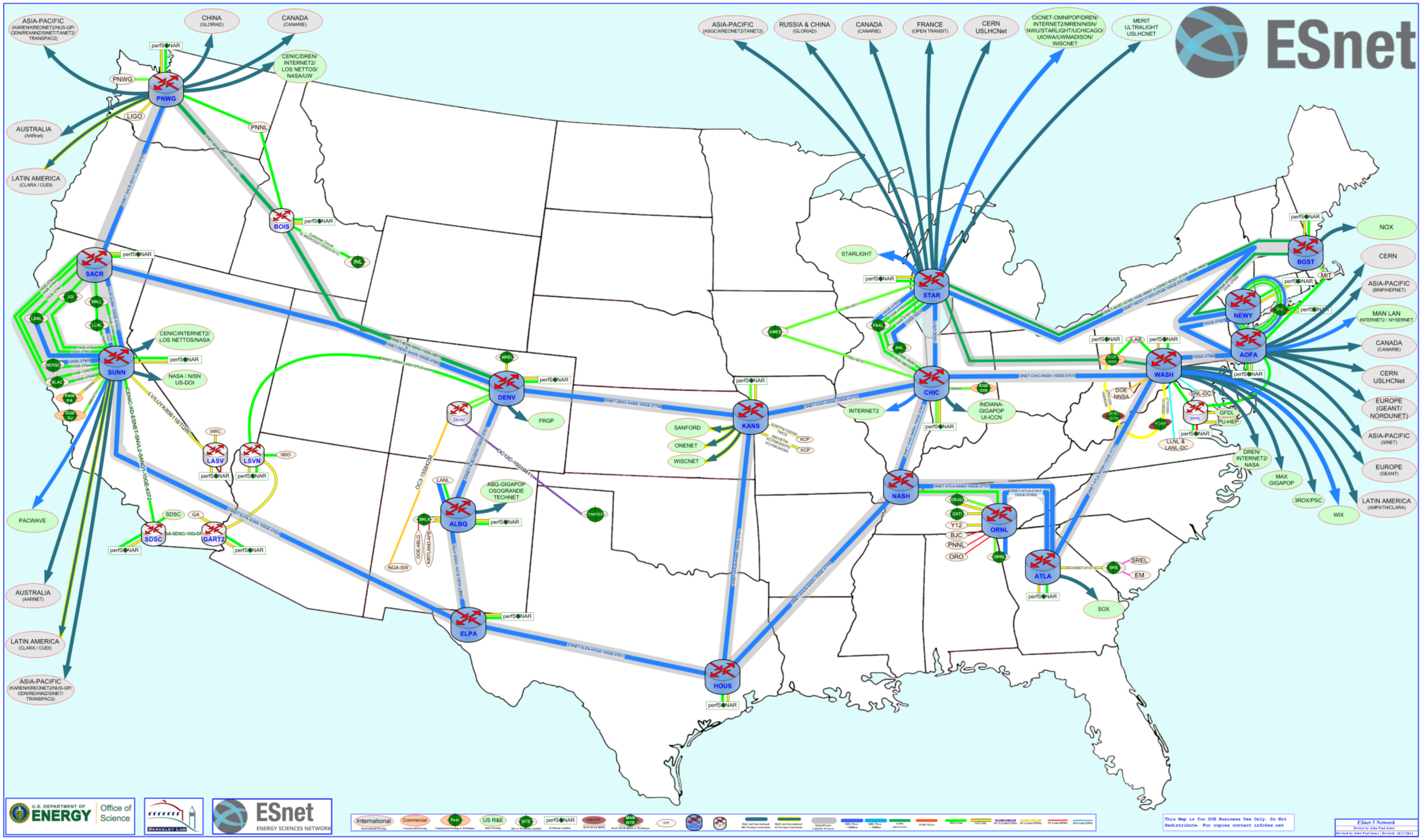
- *Hopper XT5 (2010)*
  - 1.3PF, 212TB, 2.9 MW peak power
- *Edison XC30 (in acc*
  - Based on DARPA/DOE HPCS system
  - 2.4PF, 333TB, 2.1 MW peak power
- *400TF mixed use clusters*
  - NERSC, JGI, HEP/NP, Materials, Kbase

Computational Research and Theory Building will provide 12 MW power and cooling for future NERSC computing resources

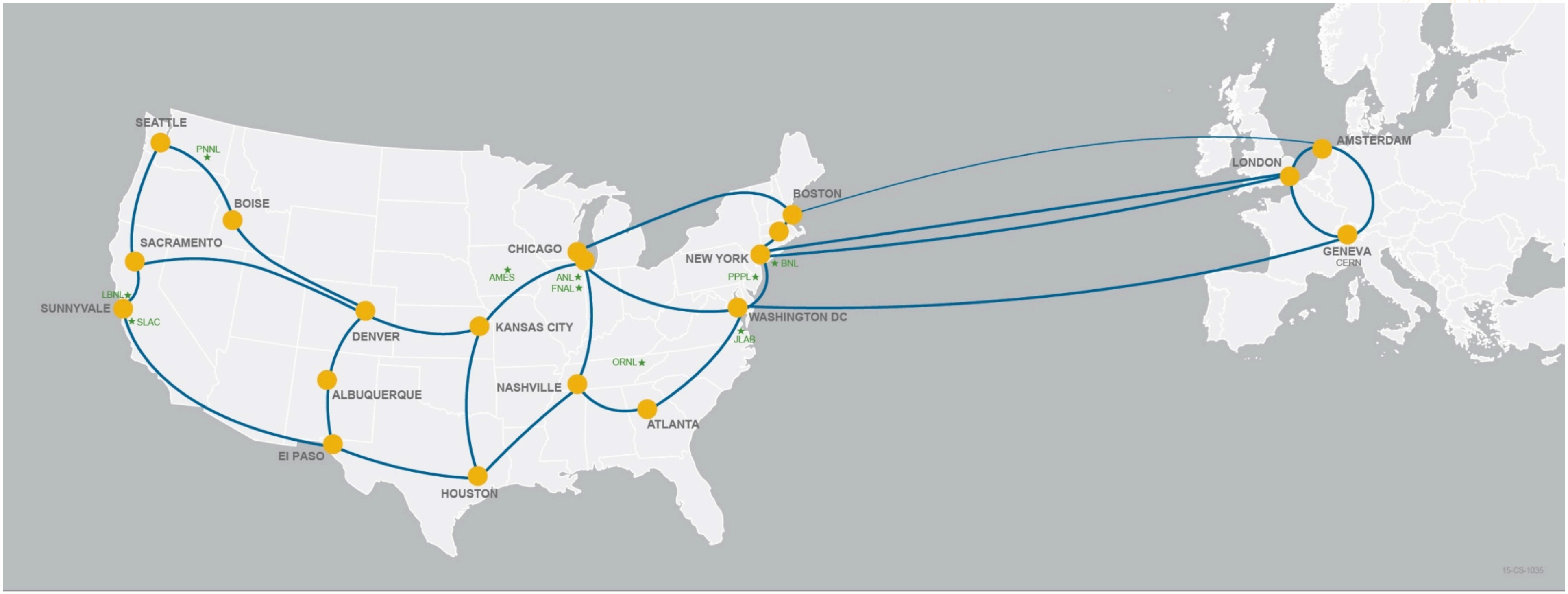
ASCR Computing At a Glance

now ← → future

System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrades	
Name Planned Installation	**Edison**	**TITAN**	**MIRA**	**Cori** **2016**	**Summit** **2017-2018**	**Theta**	**Aurora** **2018-2019**
System peak (PF)	2.6	27	10	> 30	150	>8.5	180
Peak Power (MW)	2	9	4.8	< 3.7	10	1.7	13
Total system memory	357 TB	710TB	768TB	~1 PB DDR4 + High Bandwidth Memory (HBM) +1.5PB persistent memory	> 1.74 PB DDR4 + HBM + 2.8 PB persistent memory	>480 TB DDR4 + High Bandwidth Memory (HBM)	> 7 PB High Bandwidth On-Package Memory Local Memory and Persistent Memory
Node performance (TF)	0.460	1.452	0.204	> 3	> 40	> 3	> 17 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	Multiple IBM Power9 CPUs & multiple Nvidia Voltas GPUS	2nd gen Intel Xeon Phi processor (code name Knights Landing)	3rd gen Intel Xeon Phi processor (code name Knights Hill)
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	~3,500 nodes	>2,500 nodes	>50,000 nodes
System Interconnect	Aries	Gemini	5D Torus	Aries	Dual Rail EDR-IB	Aries	2nd Generation Intel Omni-Path Architecture
File System	7.6 PB 168 GB/s, Lustre®	32 PB 1 TB/s, Lustre®	26 PB 300 GB/s GPFS™	28 PB 744 GB/s Lustre®	120 PB 1 TB/s GPFS™	10PB, 210 GB/s Lustre initial	150 PB 1 TB/s Lustre®

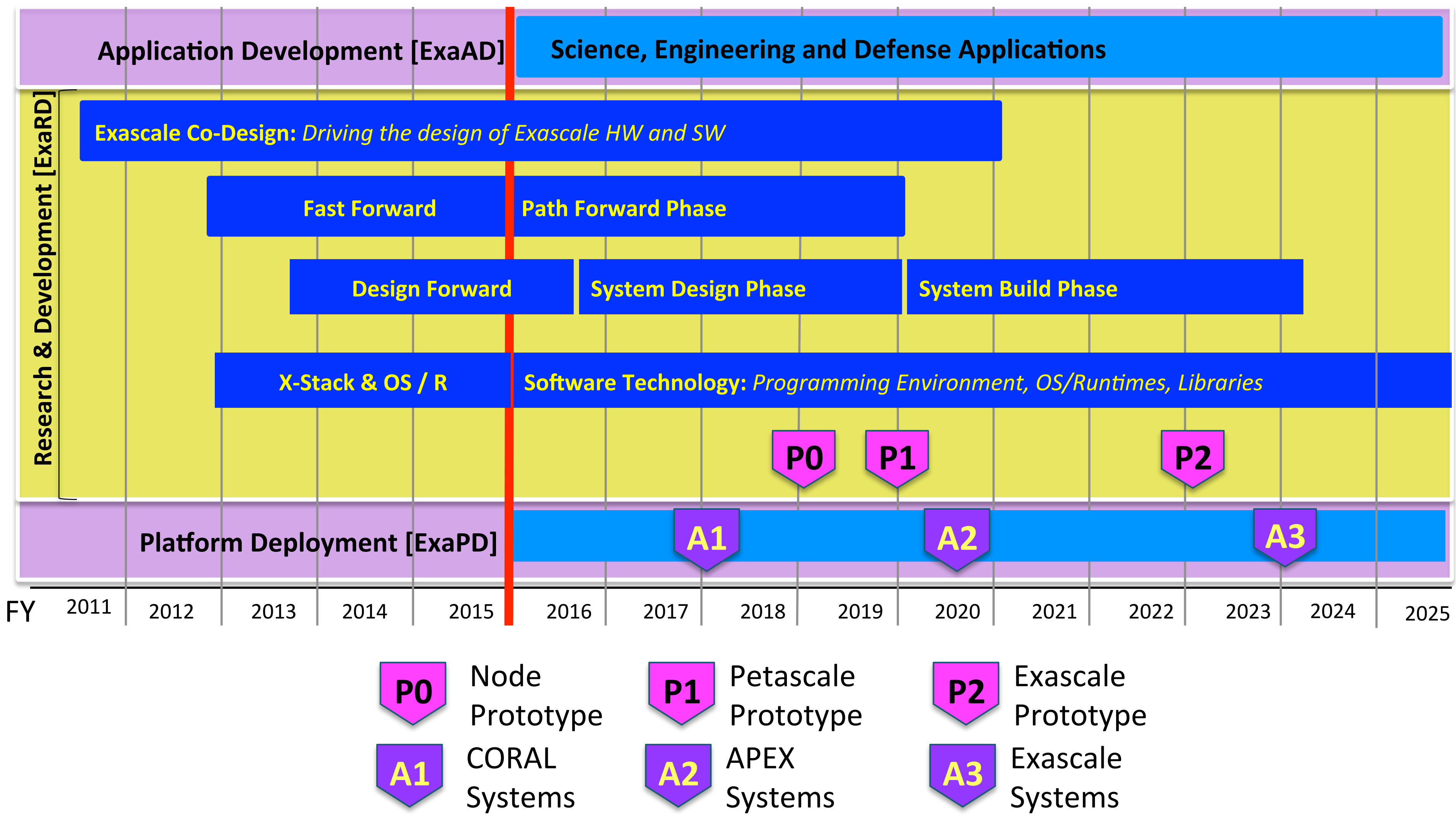

ESnet
ASIA-PACIFIC
CHINA
CANADA
AUSTRALIA
LATIN AMERICA
ASIA-PACIFIC
RUSSIA & CHINA
CANADA
FRANCE
CERN
USLHCNet
PNWG
SACR
SUNN
BOIS
LASV
LSVN
SDSC
GART2
DENV
ALBQ
ELPA
HOUS
KANS
STAR
CHIC
NASH
ORNL
ATLA
WASH
AOFA
NEWY
BOST
PACWAVE
STARLIGHT
INTERNET2
SANFORD
ONENET
WISCNET
FRGP
SOX
NOX
CERN
MAN LAN
EUROPE
WIX
MAX GIGAPOP
3ROX/PSC
perfSONAR

ESnet Goes Global: Extension to Europe

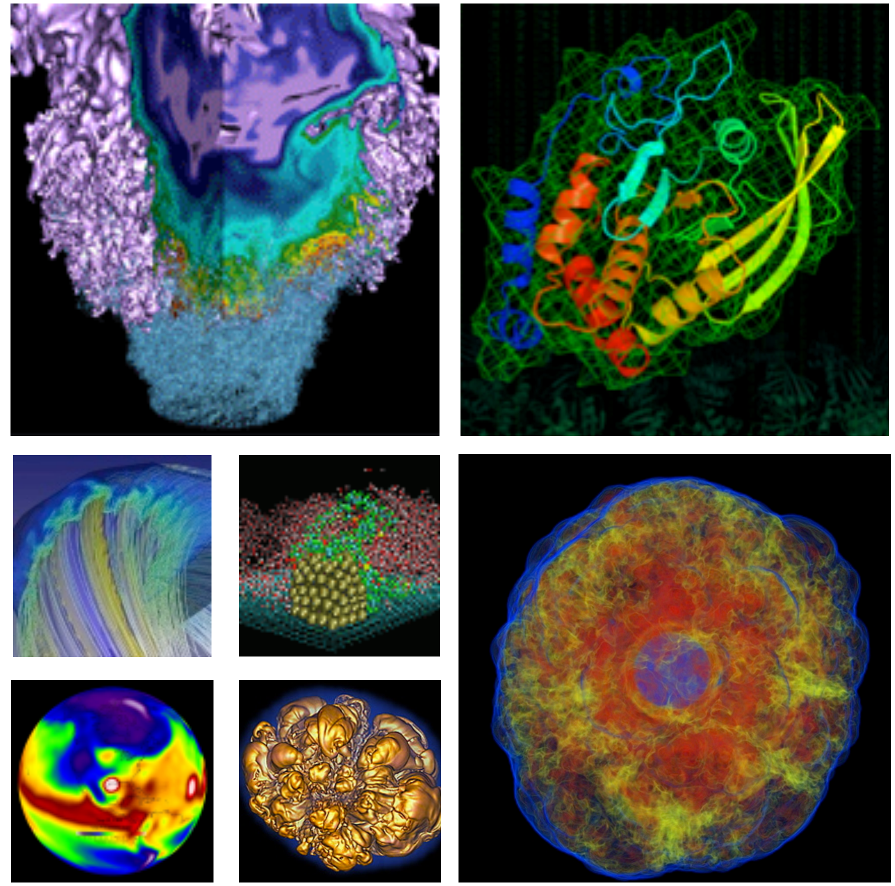


- 25% of all ESnet traffic goes to/from Europe
- 3x100+ Gbps across the Atlantic with redundant paths to serve all DOE missions
- Ready by March 2015 to support LHC Run 2 (was operational in Jan 2015)
- Will support 10x increase in transatlantic traffic from Large Hadron Collider

Exascale Computing Initiative
Timeline



Recent Developments at NERSC



NERSC 40 YEARS at the FOREFRONT 1974-2014

Sudip Dosanjh
Director

June 10, 2015

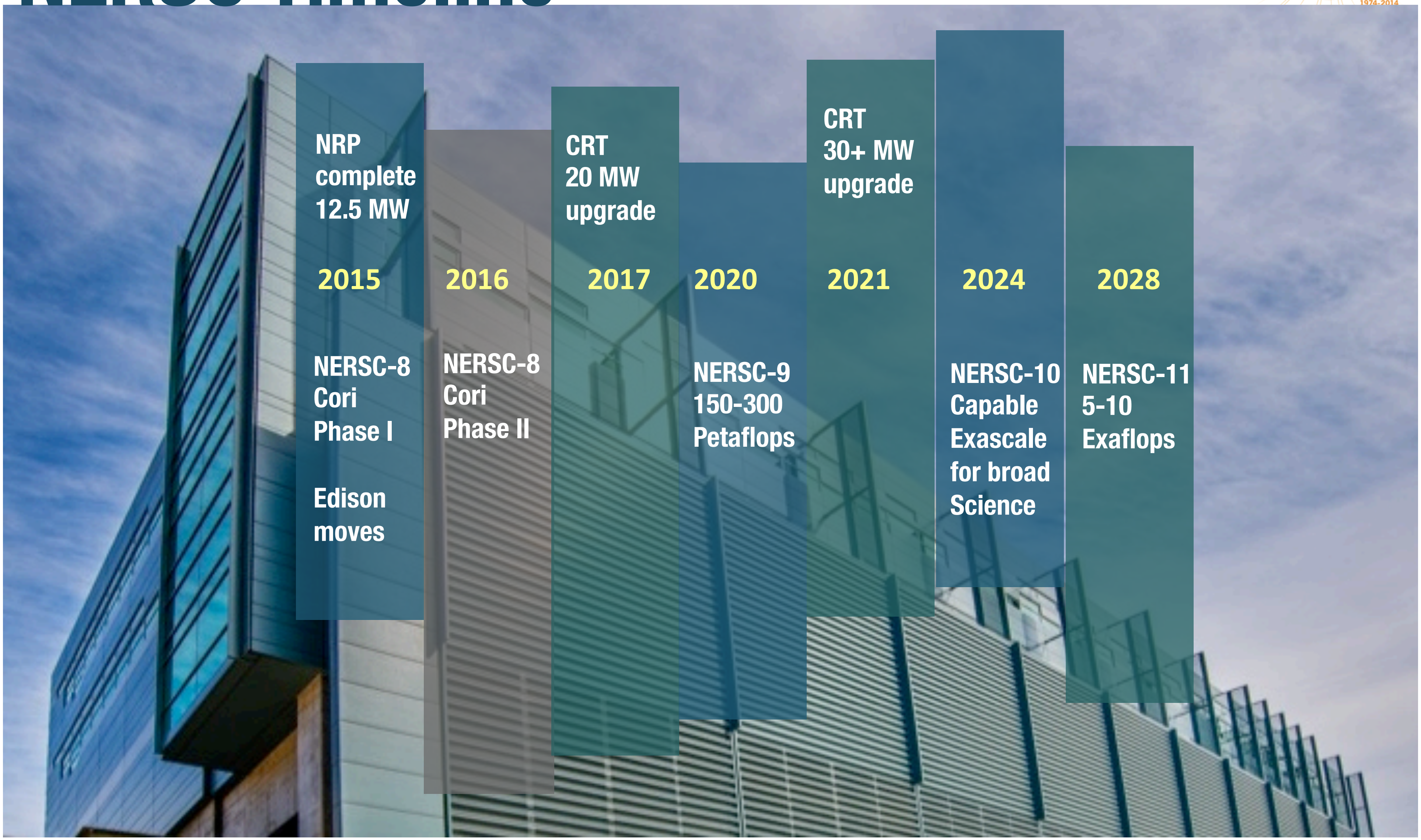
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Cori will be installed in the Computational Research and Theory (CRT) Facility

- **Four story, 140,000 GSF**
  - 300 offices on two floors
  - 20K -> 29Ksf HPC floor
  - 12.5MW -> 40 MW to building
- **Located for collaboration**
  - CRD and ESnet
  - UC Berkeley
- **Exceptional energy efficiency**
  - Natural air and water cooling
  - Heat recovery
  - PUE < 1.1
  - LEED gold design
- **Initial occupancy 2015**

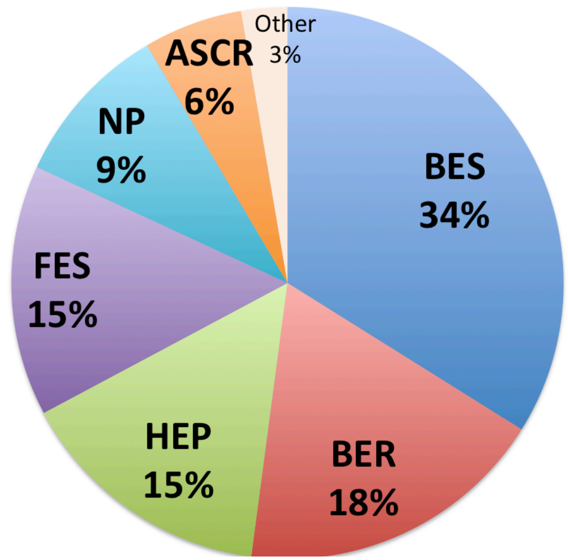
NERSC Timeline

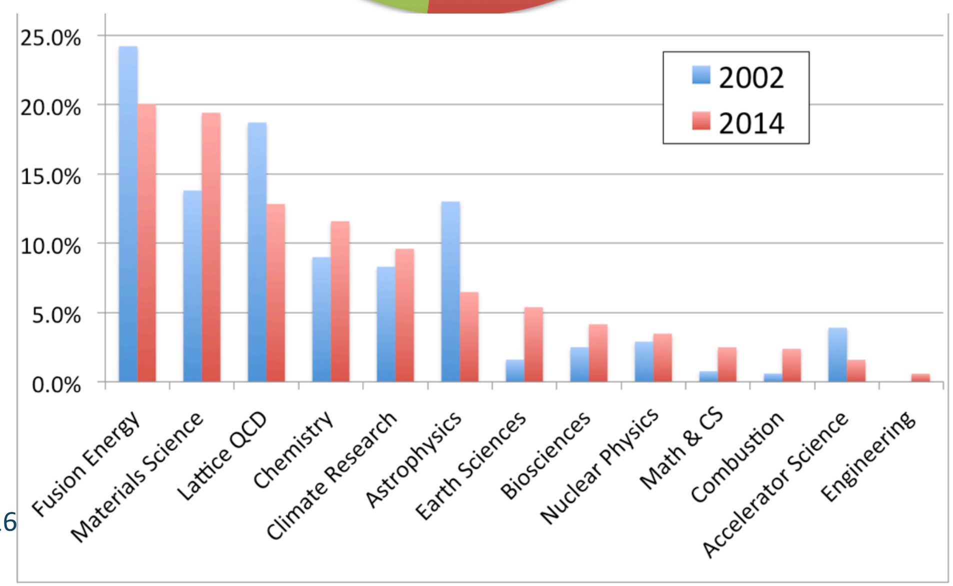


We directly support DOE's science mission

- **We are the primary computing facility for DOE Office of Science**
- **DOE SC allocates the vast majority of the computing and storage resources at NERSC**
  - Six program offices allocate their base allocations and they submit proposals for overtargets
  - Deputy Director of Science prioritizes overtarget requests
- **Usage shifts as DOE priorities change**

2014 Allocation Breakdown

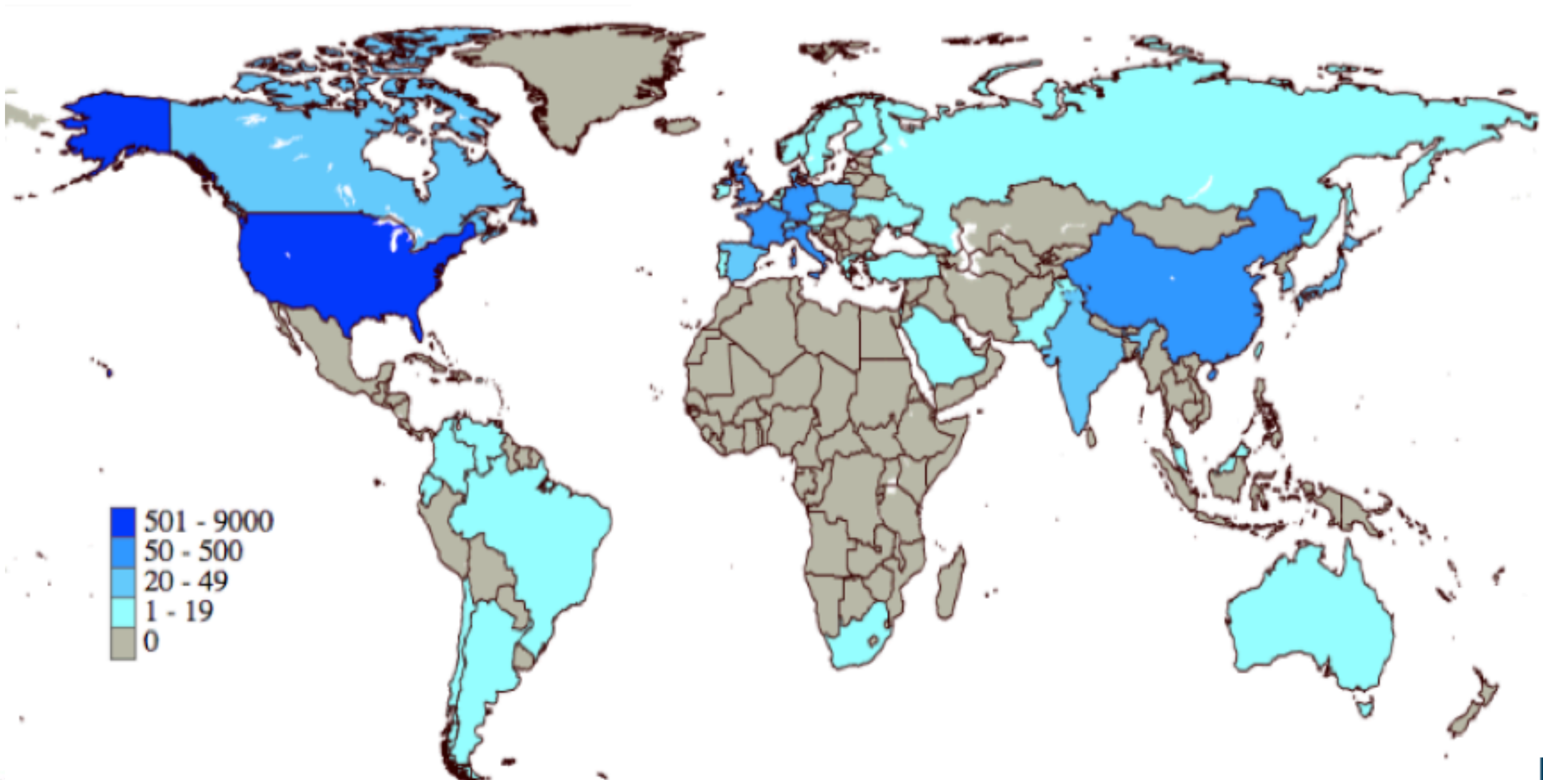




U.S. DEPARTMENT OF ENERGY | Office of Science

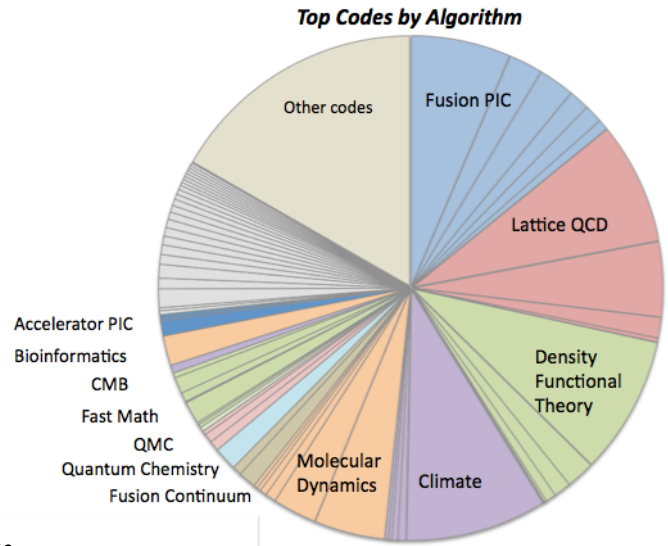
We support a broad user base

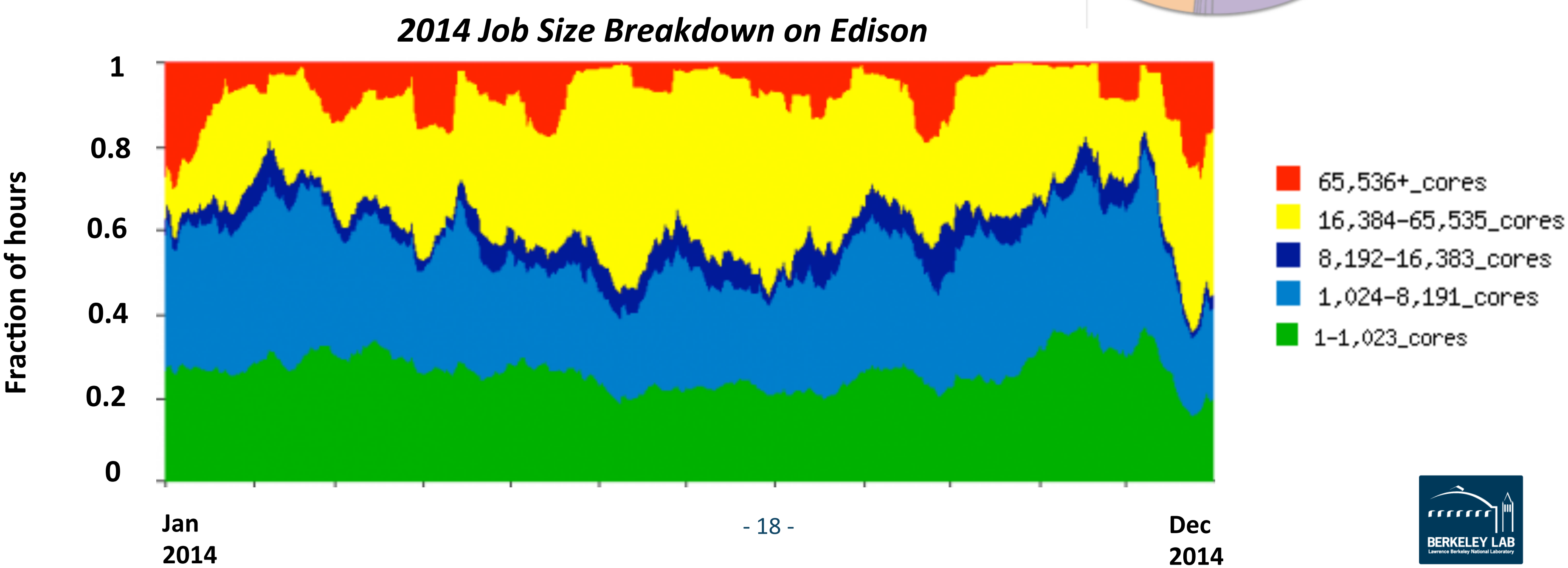
- **~6000 users, and we typically add 300-500 per year**
- **Geographically distributed: 48 states as well as multinational projects**



We support a diverse workload

- **Many codes (600+) and algorithms**
- **Computing at scale and at high volume**





NERSC collaborates with computer companies to deploy advanced HPC and data resources

- **Hopper (N6) and Cielo (ACES) were the first Cray petascale systems with a Gemini interconnect**
- **Architected and deployed data platforms including the largest DOE system focused on genomics**
- **Edison (N7) is the first Cray petascale system with Intel processors, Aries interconnect and Dragonfly topology (serial #1)**
- **Cori (N8) will be one of the first large Intel KNL systems and will have unique data capabilities**

The NERSC-8 System: Cori

- **Cori will support the broad Office of Science research community and begin to transition the workload to more energy efficient architectures**
- **Cray XC system with over 9300 Intel Knights Landing compute nodes –mid 2016**
  - Self-hosted, (not an accelerator) manycore processor with over 60 cores per node
  - On-package high-bandwidth memory
- **Data Intensive Science Support**
  - 10 Haswell processor cabinets to support data intensive applications – Summer 2015
  - NVRAM Burst Buffer to accelerate data intensive applications
  - 28 PB of disk, >700 GB/sec I/O bandwidth
- **Robust Application Readiness Plan**
  - Outreach and training for user community
  - Application deep dives with Intel and Cray
  - 8 post-docs integrated with key application teams

Image source: Wikipedia

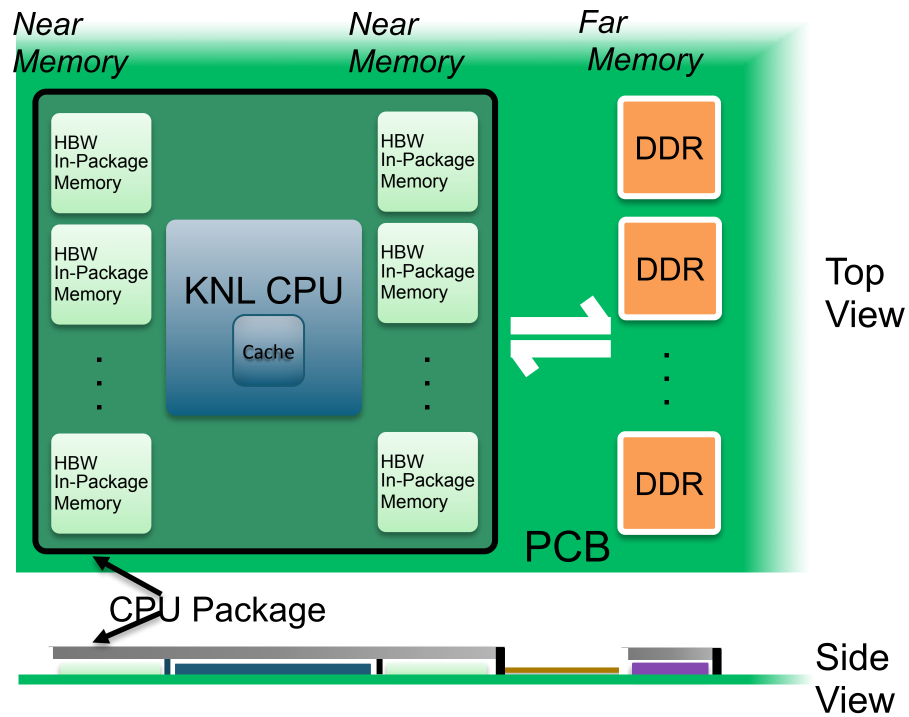
System named after Gerty Cori, Biochemist and first American woman to receive the Nobel prize in science.

Intel "Knights Landing" Processor

- Next generation Xeon-Phi, >3TF peak
- Single socket processor - Self-hosted, not a co-processor, not an accelerator
- Greater than 60 cores per processor with support for four hardware threads each; more cores than current generation Intel Xeon Phi™
- Intel® "Silvermont" architecture enhanced for high performance computing
- 512b vector units (32 flops/clock – AVX 512)
- 3X single-thread performance over current generation Xeon-Phi co-processor
- High bandwidth on-package memory, up to 16GB capacity with bandwidth projected to be 5X that of DDR4 DRAM memory
- Higher performance per watt

Knights Landing Integrated On-Package Memory

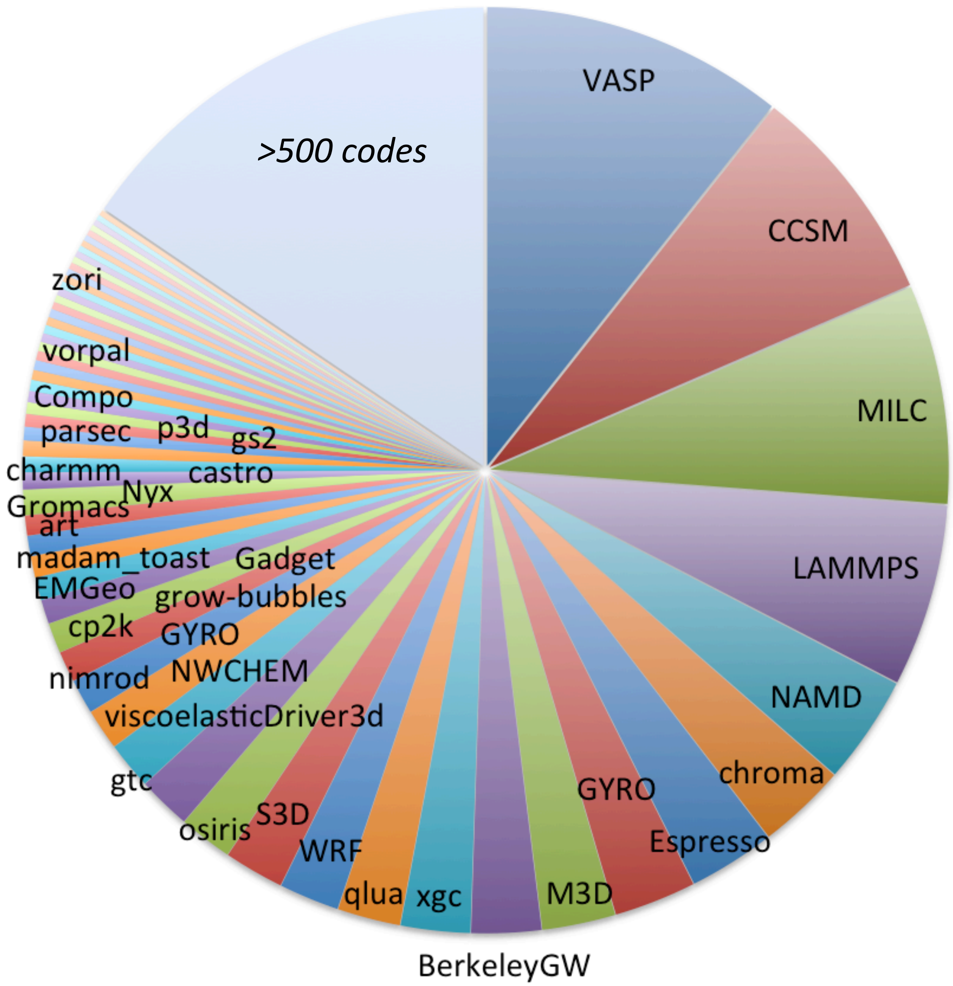
Cache Model	Let the hardware automatically manage the integrated on-package memory as an "L3" cache between KNL CPU and external DDR
Flat Model	Manually manage how your application uses the integrated on-package memory and external DDR for peak performance
Hybrid Model	Harness the benefits of both cache and flat models by segmenting the integrated on-package memory



Maximum performance through higher memory bandwidth and flexibility

We will initially focus on 20 codes

Breakdown of Application Hours on Hopper and Edison 2013



- **10 codes make up 50% of the workload**
- **25 codes make up 66% of the workload**
- **Edison will be available until 2019/2020**
- **Training and lessons learned will be made available to all application teams**

20 NESAP Tier-1 and Tier-2 codes

ASCR (2)

Almgren (LBNL) – **BoxLib AMR Framework** used in combustion, astrophysics

Trebotich (LBNL) – **Chombo-crunch** for subsurface flow

BES (5)

Kent (ORNL) – **Quantum Espresso**
Deslippe (NERSC) – **BerkeleyGW**
Chelikowsky (UT) – **PARSEC** for excited state materials
Bylaska (PNNL) – **NWChem**
Newman (LBNL) – **EMGeo** for geophysical modeling of Earth

BER (5)

Smith (ORNL) – **Gromacs** Molecular Dynamics
Yelick (LBNL) – **Meraculous** genomics
Ringler (LANL) – **MPAS-O** global ocean modeling
Johansen (LBNL) – **ACME** global climate
Dennis (NCAR) – **CESM**

HEP (3)

Vay (LBNL) – **WARP & Synergia** accelerator modeling
Toussaint (U Arizona) – **MILC** Lattice QCD
Habib (ANL) – **HACC** for *n*-Body cosmology

NP (3)

Maris (U. Iowa) – **MFDn** *ab initio* nuclear structure
Joo (JLAB) – **Chroma** Lattice QCD
Christ/Karsch (Columbia/ BNL) – **DWF/HISQ** Lattice QCD

FES (2)

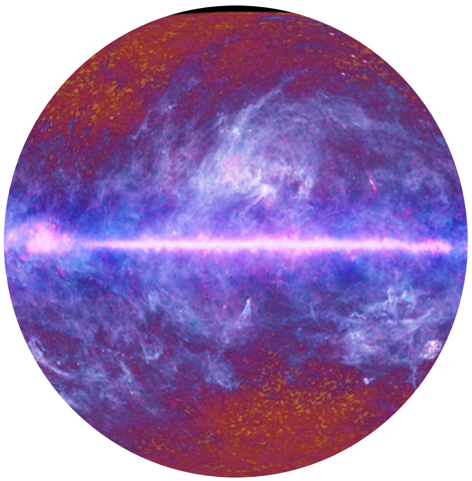
Jardin (PPPL) – **M3D** continuum plasma physics
Chang (PPPL) – **XGC1** PIC plasma

We are launching the NERSC Exascale Science Applications Program (NESAP)

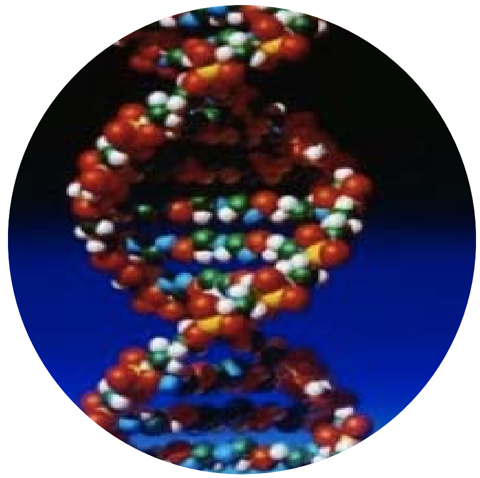
- **NESAP components:**

- Strong support from vendors
- Developer Workshops for 3rd-Party SW
- Early engagement with code teams
- Leverage existing community efforts
- Postdoc Program
- NERSC training and online modules
- Early access to N8 technology

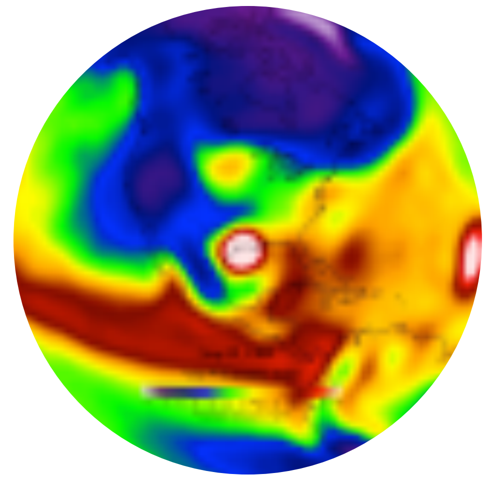
DOE Facilities are Facing a Data Deluge



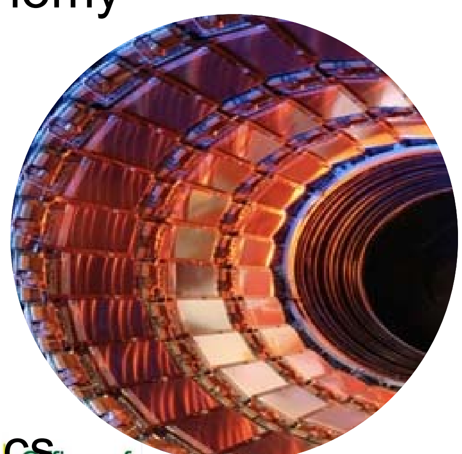
Astronomy



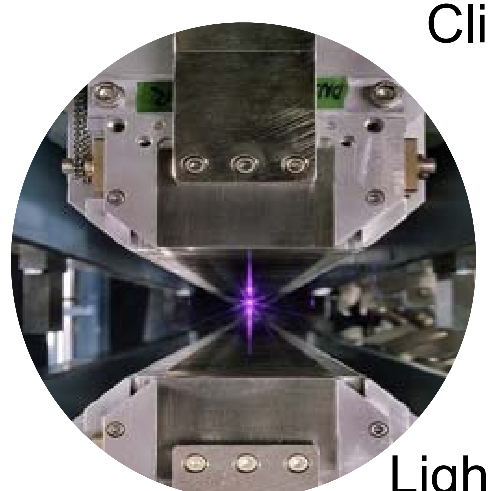
Genomics



Climate

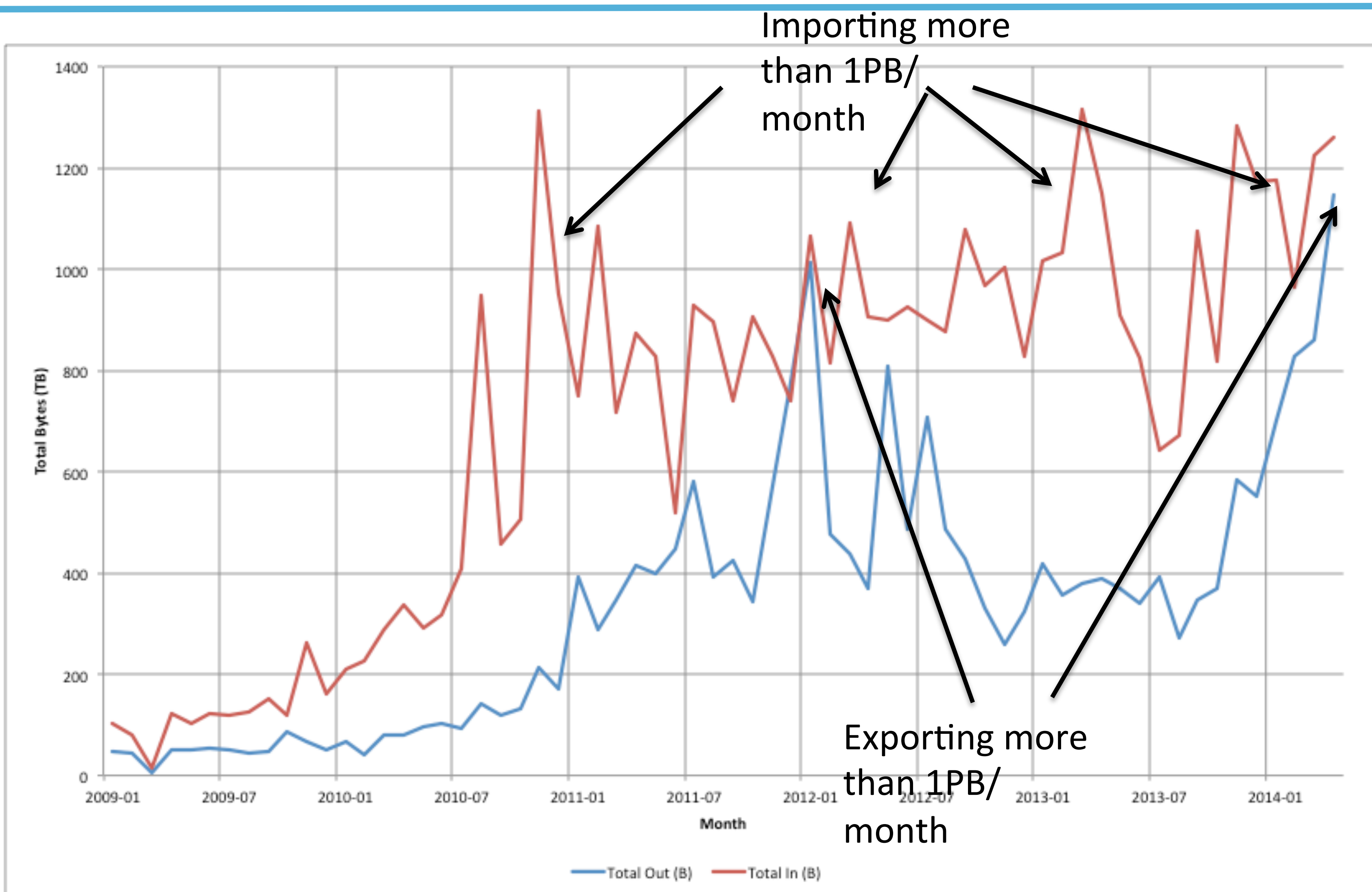


Physics



Light Sources

NERSC users import more data than they export!



Cori Data Enhancements

- **Data partition with large memory nodes, software to enable data workflows (including user defined images)**
- **IO enhancements-- NVRAM nodes on the interconnect fabric for caching, software defined networking**
- **Larger disk system**

Goals are to enable the analysis of large experimental data sets and in-situ analysis coupled to Petascale simulations

NERSC, LANL and Sandia formed a partnership for next-generation supercomputers

- **Alliance for application Performance at Extreme-scale (APEX)**
  - Visible collaboration between ASCR and ASC
  - Strengthen impact on industry
  - Address challenges transitioning applications to advanced manycore architectures with a broader coalition
  - Risk mitigation on technical challenges
- **Successive deployments**
  - Informal partnership on Hopper and Cielo (2010)
  - Cori and Trinity in 2015-2016
  - NERSC-9 and Crossroads in 2020