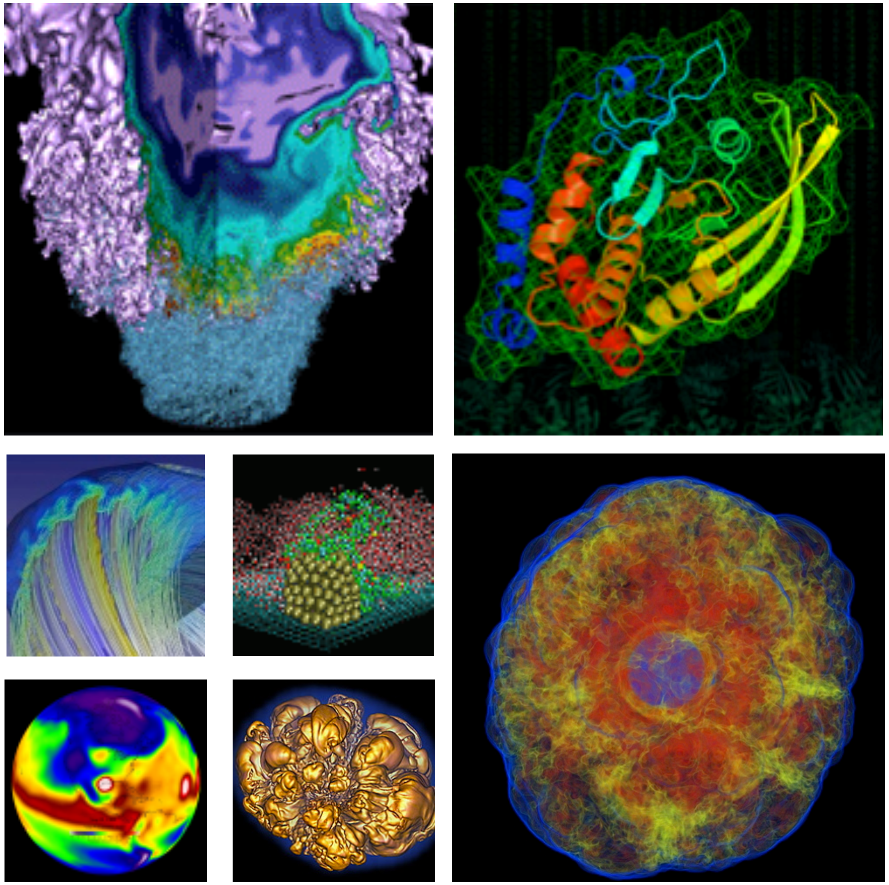
NERSC-8 Project

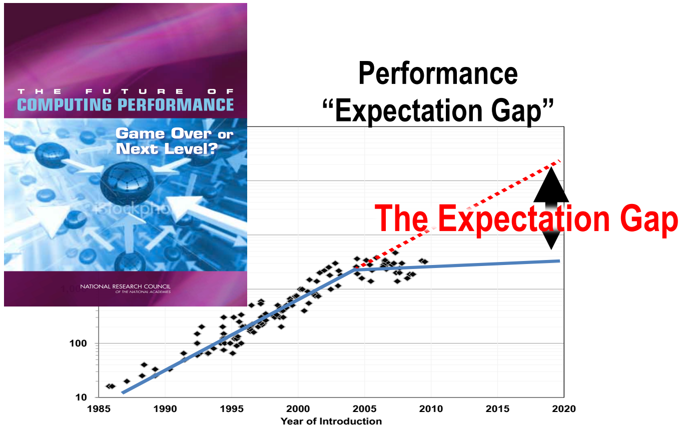


NUG Meeting

Katie Antypas – NERSC-8 Project Lead

Feb. 12, 2013

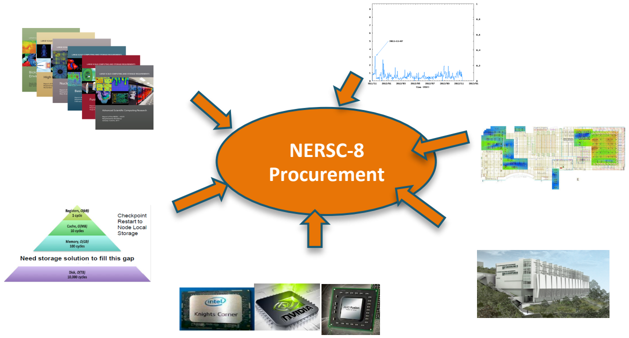
Topics and Discussion for Today



HPC landscape in NERSC-8 time frame: power and programming challenges

Trinity/NERSC-8 Collaboration

NERSC-8 Project Information


NERSC-8 Procurement
Checkpoint Restart to Node Local Storage
Need storage solution to fill this gap
Knights Corner

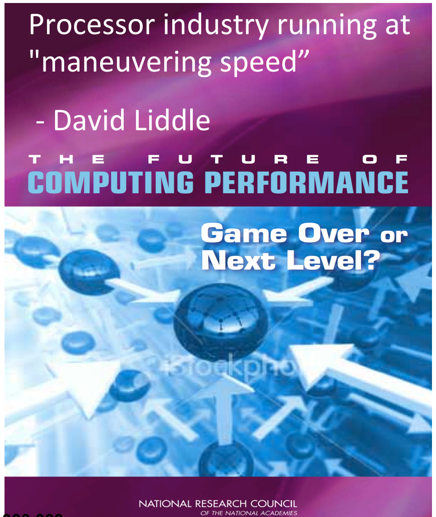
Related projects providing input to NERSC-8 project

NERSC-8 Mission Need

The Department of Energy Office of Science requires an HPC system to support the rapidly increasing computational demands of the entire spectrum of DOE SC computational research.

- Provide a significant increase in computational capabilities, at least 10 times the sustained performance of the Hopper system on a set of representative DOE benchmarks
- Delivery in the 2015/2016 time frame
- Provide high bandwidth access to existing data stored by continuing research projects.
- Platform needs to begin to transition users to more energy-efficient many-core architectures.

To sustain historic performance growth, the DOE community must prepare for new architectures



Processor industry running at "maneuvering speed"

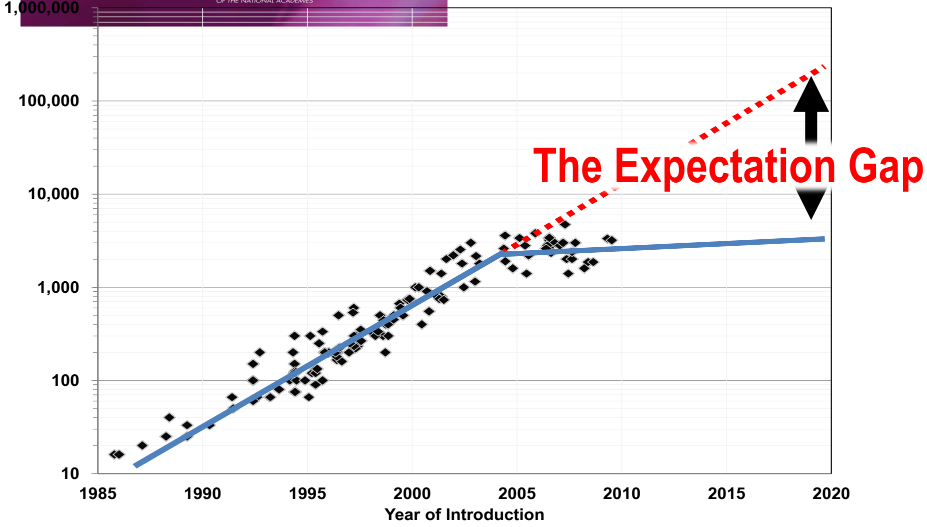
- David Liddle

THE FUTURE OF COMPUTING PERFORMANCE

Game Over or Next Level?

NATIONAL RESEARCH COUNCIL OF THE NATIONAL ACADEMIES

Performance "Expectation Gap"

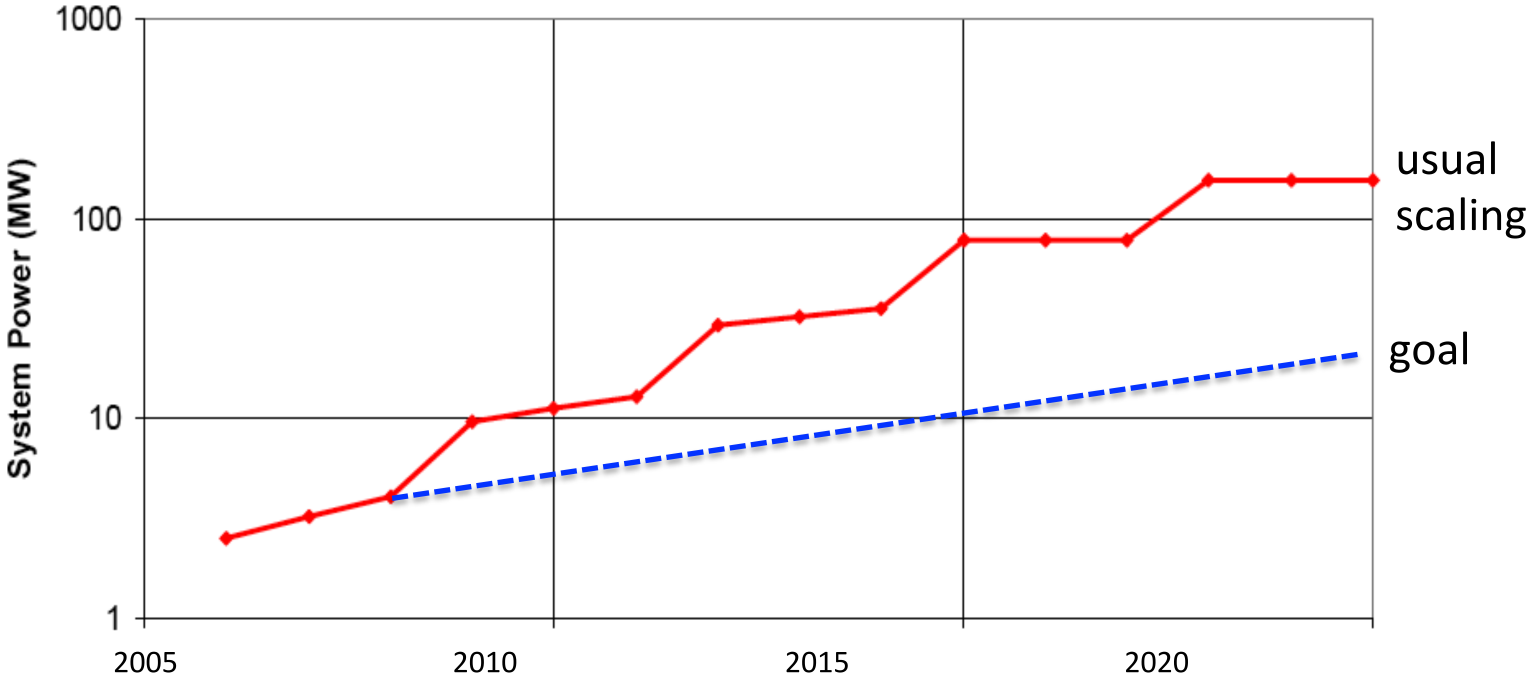


- Processor speeds have stalled though transistor density continues to increase
- Vendors are increasing the number of cores on a chip
- To take advantage of these emerging architectures, applications must
  - Exploit parallelism at deeper levels
  - Manage data placement and movement
  - Accommodate less memory per process space

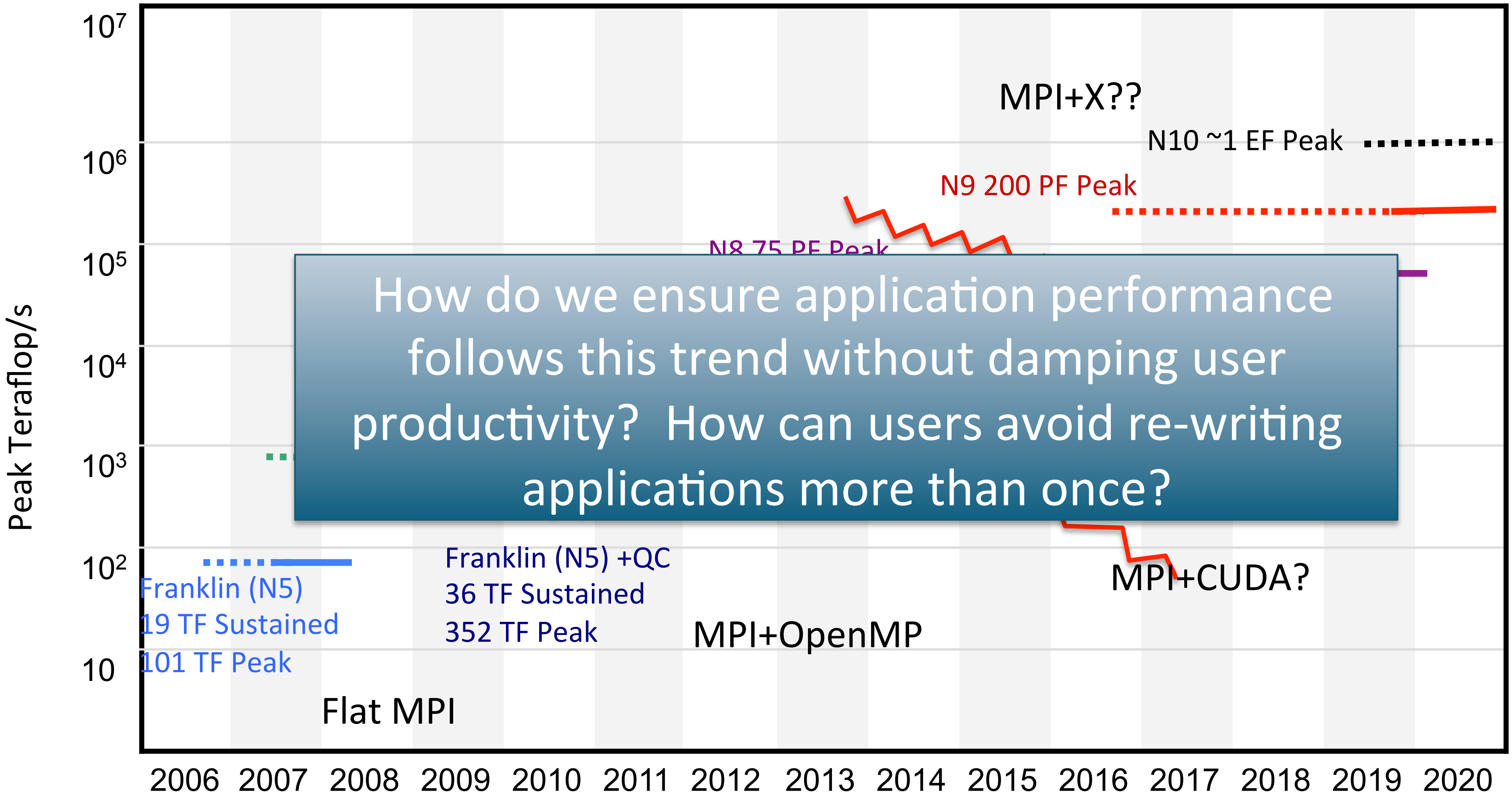
New energy efficient architectures are necessary for achieving system power goals

- 1 petaflop (10^{15} ops) requires ~3MW

  → 3 GW for 1 Exaflop (10^{18} ops/sec)
- DARPA committee suggested 200 MW with “usual” scaling



The path to Exascale: How and When to Move Users?



How do we ensure application performance follows this trend without damping user productivity? How can users avoid re-writing applications more than once?

The ACES Trinity team and the NERSC-8 team are collaborating

NERSC

- Teams worked together on Hopper/Cielo and found interactions useful
- Strengthen alliance between SC/NNSA on road to exascale
- Share technical expertise between Labs

Trinity/NERSC-8 Collaboration

Sandia National Laboratories

Los Alamos NATIONAL LABORATORY EST. 1943

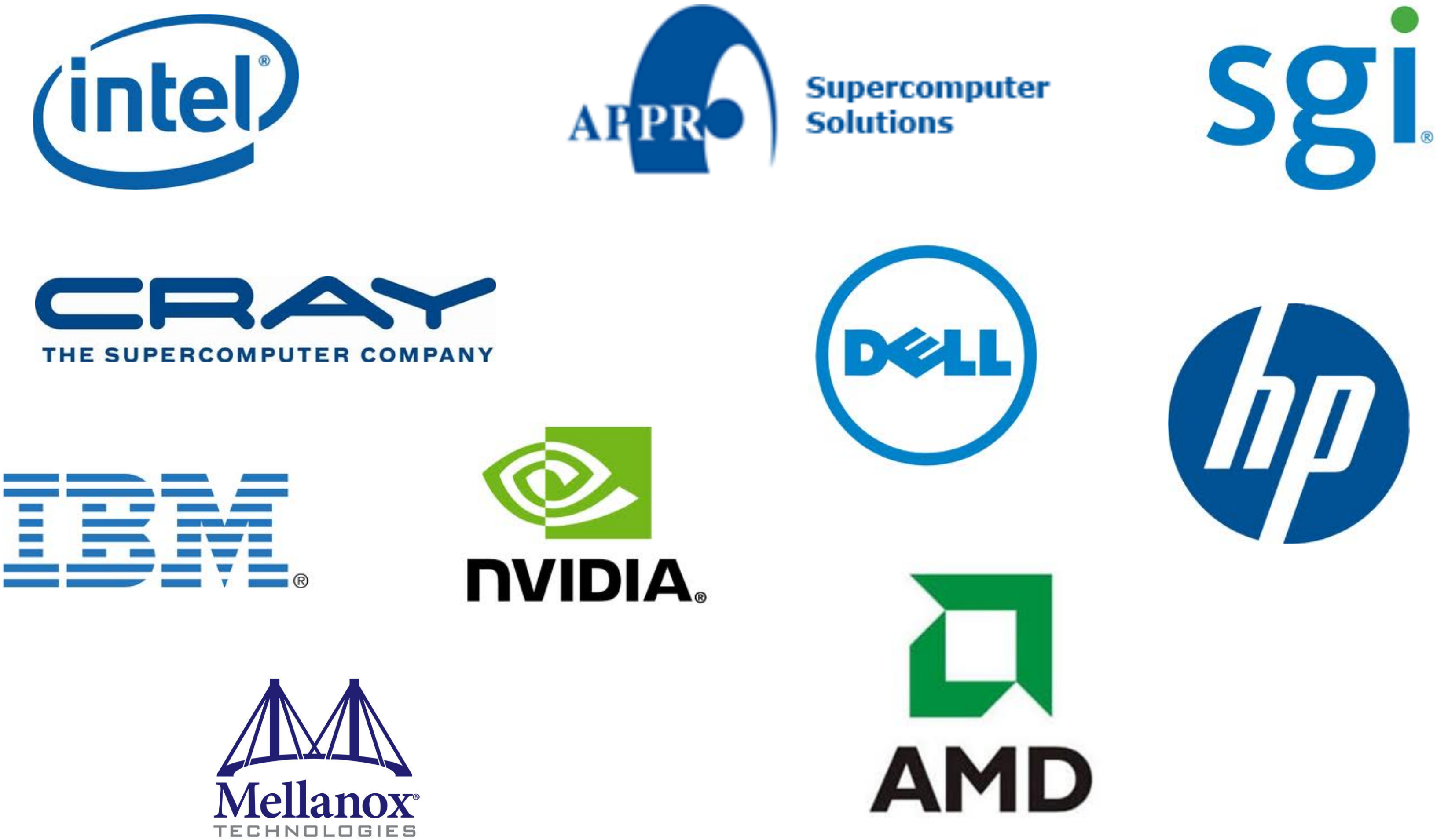
NERSC

Plans are for a joint Trinity/NERSC-8 RFP calling for two distinct systems of similar technology with the intention to award both systems to the same vendor.

Approximate NERSC-8 timeline

Requirements Gathering	Fall 2012	✓ On-going
Vendor Market Surveys	Fall 2012	✓ On-going
Mission Need Approval	Nov 2012	✓
Release Draft Technical Specifications	Dec 2012	✓
Release Draft Benchmarks	Dec 2012-March 2013	
Design and Lehman Review	Spring 2013	
Release final RFP	Summer 2013	
Vendor Selection	Summer 2013	
Vendor Negotiations	Fall 2013	
Mini-Lehman Review	Fall 2013	
System Delivery	Late 2015	
System Acceptance	2016	

We kicked off the procurement with a round of vendor market surveys



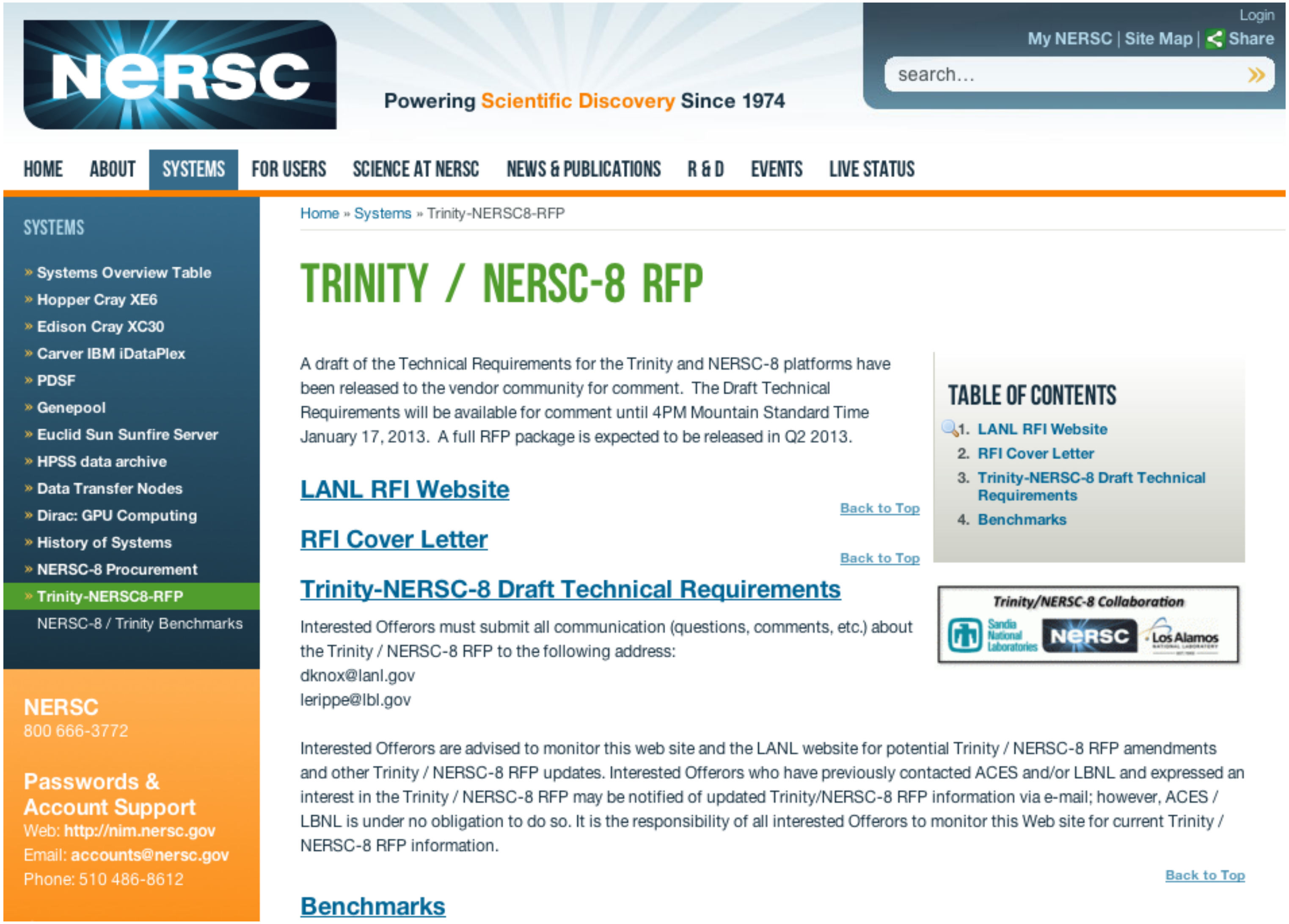
NERSC-8 Draft Design Targets and Limitations

- >10x application performance over Hopper system based on SSP metric
- Aggregate memory: 1-2PB
- Disk capacity: >20x memory
- I/O Bandwidth: dump 80% of memory in 35 minutes
- Maximum power – 6MW
- Delivery in 2015/2016 time frame

Many other target requirements in the following areas

- Performance for real applications
- Application portability
- Ease of programming
- Scalable interconnect
- System resilience and reliability
- System facility integration
- User, system and management software

Draft Technical Requirements released to vendor community in December 2012



NERSC
800 666-3772

Passwords & Account Support
Web: http://nim.nersc.gov
Email: accounts@nersc.gov
Phone: 510 486-8612

TRINITY / NERSC-8 RFP

A draft of the Technical Requirements for the Trinity and NERSC-8 platforms have been released to the vendor community for comment. The Draft Technical Requirements will be available for comment until 4PM Mountain Standard Time January 17, 2013. A full RFP package is expected to be released in Q2 2013.

TABLE OF CONTENTS

LANL RFI Website

RFI Cover Letter

Trinity-NERSC-8 Draft Technical Requirements

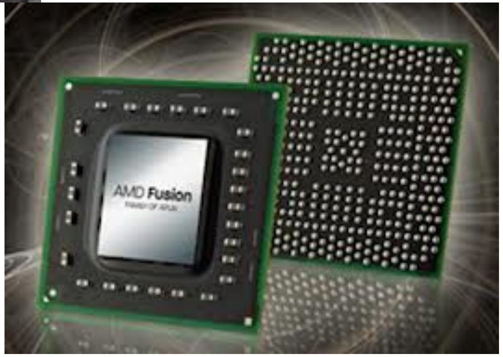
Interested Offerors must submit all communication (questions, comments, etc.) about the Trinity / NERSC-8 RFP to the following address:
dknox@lanl.gov
leripe@lbl.gov

Interested Offerors are advised to monitor this web site and the LANL website for potential Trinity / NERSC-8 RFP amendments and other Trinity / NERSC-8 RFP updates. Interested Offerors who have previously contacted ACES and/or LBNL and expressed an interest in the Trinity / NERSC-8 RFP may be notified of updated Trinity/NERSC-8 RFP information via e-mail; however, ACES / LBNL is under no obligation to do so. It is the responsibility of all interested Offerors to monitor this Web site for current Trinity / NERSC-8 RFP information.

Benchmarks

Although architecture for NERSC-8 is not yet known, trend is toward manycore processors

- **Regardless of chip vendor chosen for NERSC-8, users will need to modify applications to achieve performance**
- **Multiple levels of code modification may be necessary**
  - Expose more on-node parallelism in applications
  - Increase application vectorization capabilities
  - For co-processor architectures, locality directives must be added

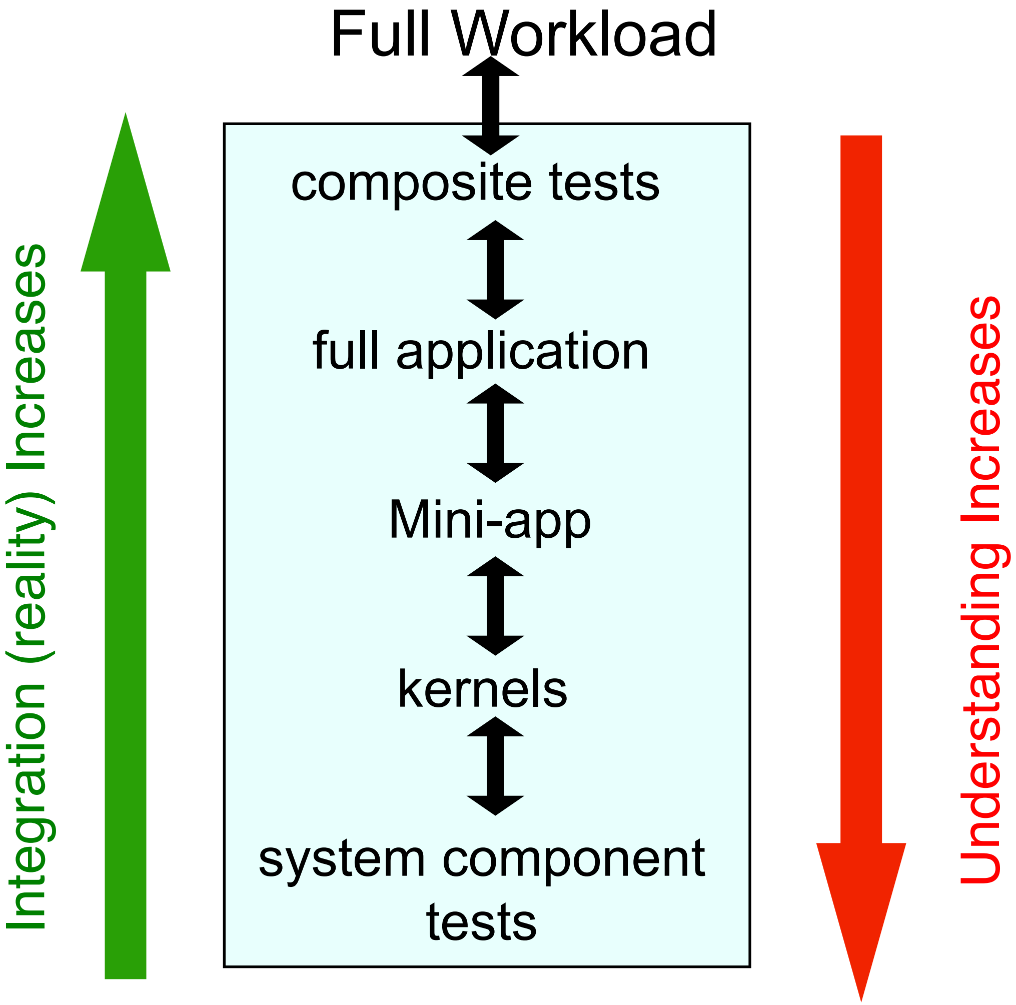


Current technology landscape makes benchmark packaging and selection challenging

- **Variety of chip architectures (CPU, GPU, MIC)**
- **Uncertain programming model (MPI, OpenMP, OpenACC, CUDA)**
- **Limited staff to assemble benchmarks, infeasible to release benchmarks compatible for all programming models**
- **Cognizant of effort required by vendors to run benchmarks**
- **Collaboration with Trinity means we must find overlap in benchmark selection**

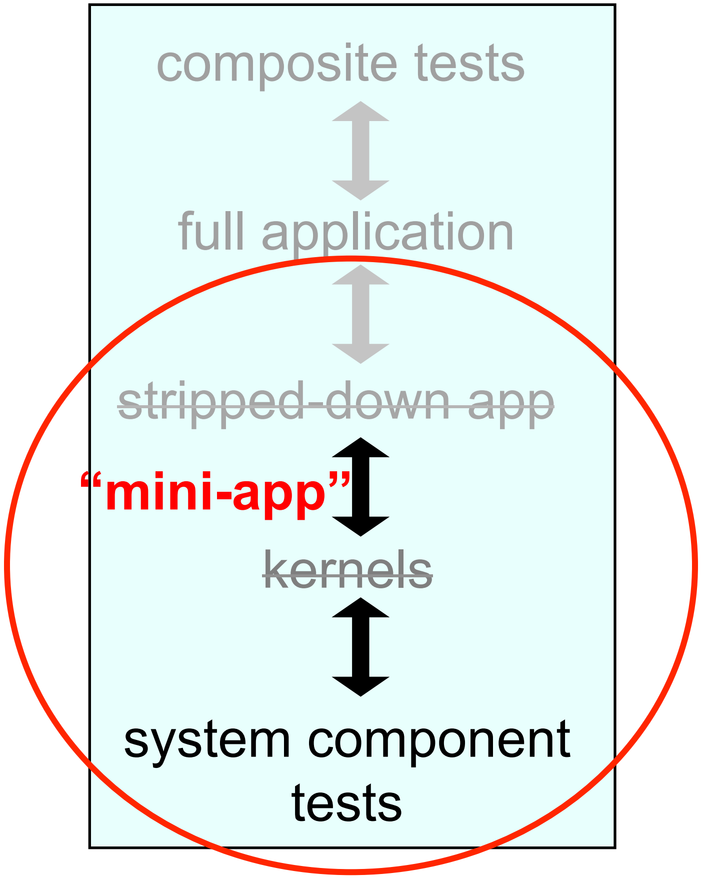
In the past NERSC has released benchmarks of various levels of complexity with the RFP

- Distinguish performance of systems
- Compare price/ performance metrics Flops/$ and Flops/Watt
- Represent scientific workload on system
- Give confidence that chosen system will perform well for NERSC workload
- Used throughout lifetime of the system



NERSC-8/Trinity plan to use "mini-apps", some full apps and micro-benchmarks for system evaluation

- **Releasing full benchmarks for all architectures and programming models infeasible**
- **Plan is to release mini-apps with MPI+OpenMP**



NERSC-8/Trinity plan to use “mini-apps”, some full apps for system evaluation

NERSC

MiniApp	Descirption
miniDFT (Quantum Espresso)	Density Functional Theory (DFT)
MILC	Lattice Quantum Chromodynamics (QCD). Sparse matrix inversion, CG
GTC	Particle-in-cell magnetic fusion
AMG	Algebraic Mult-Grid linear system solver for unstructured mesh physics packages
UMT	Unstructured-Mesh deterministic radiation Transport
miniFE	Unstructured implicit finite element
miniContact	Structural mechanics contact search
miniGhost	Finite difference stencil
SNAP	Neutral particle transport application

Note: this list is not final and could change before final RFP release

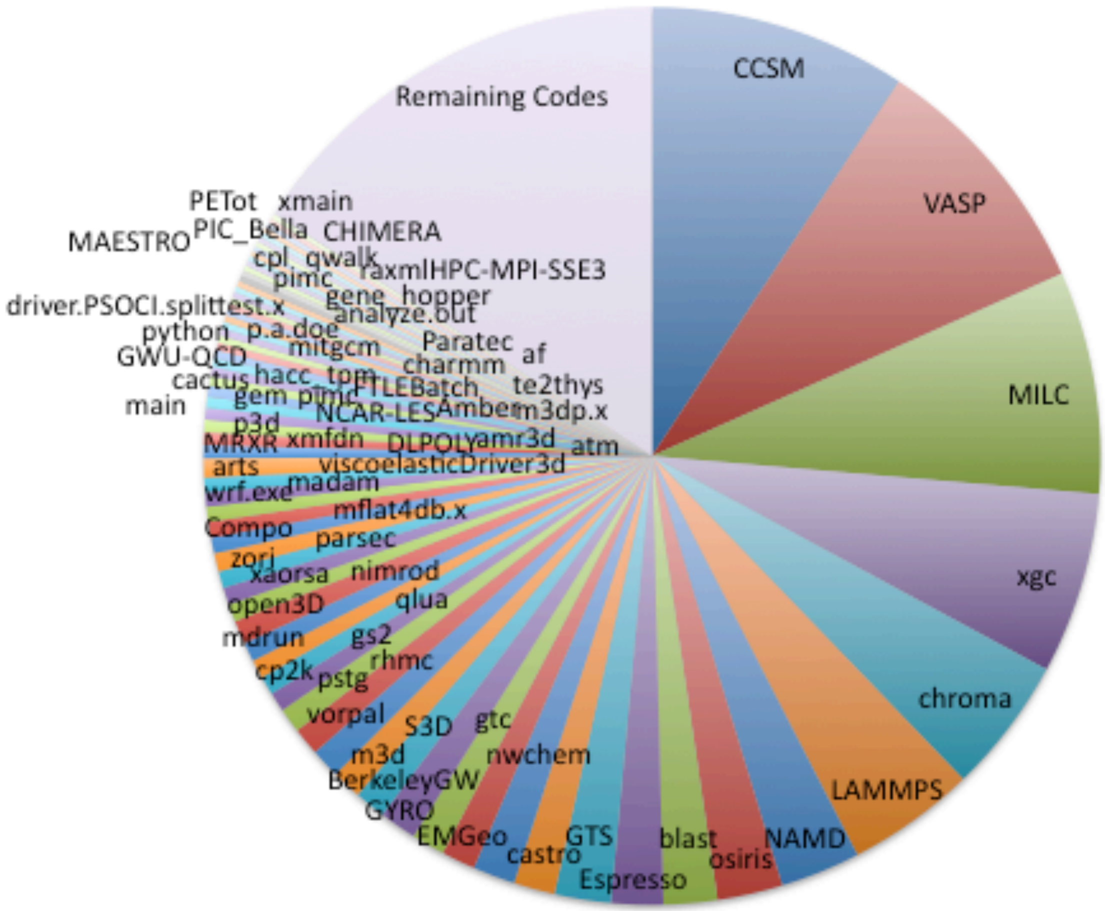
Methods covered by NERSC-8/Trinity Benchmarks

Codes	Dense Linear Algebra	Sparse Linear Algebra	FFTs	Particle Methods	Structured Grids	Unstructured Grids/AMR
miniDFT	X		X		X	
MILC		X		X	X	
GTC				X	X	
UMT / AMG		X				X
miniFE		X				X
miniContact		X				
miniGhost		X			X	
SNAP					X	X

Science Area coverage by NERSC-8/Trinity Benchmarks

Codes	Accel Sci	Astro physics	Chem	Climate	Combustion	Fusion	Lattice Gauge	Material Science
miniDFT			X					X
MILC							X	
GTC						X		
UMT / AMG		X			AMG only	X		
miniFE	X	X		X	X	X		
miniContact								
miniGhost	X	X	X	X	X	X	X	X
SNAP		X						X

While N8 benchmark apps represent workload, there are hundreds of codes that need to migrate to new architectures.



Top Application Codes

Jan – Nov 2012

- **10 codes make up 50% of workload**
- **25 codes make up 66% of workload**
- **75 codes make up 85% of workload**
- **remaining codes make up bottom 15% of workload**

Approximately 80% of the workload needs to transfer to NERSC-8 (20% can remain on Edison)

Edison plays a key role in the NERSC-8 strategy

- Moving entire workload to NERSC-8 platform will be a challenge

- Workloads that have difficulty moving to NERSC-8 can still work productively on Edison while the code is adapted

- In 2016 Edison will likely provide ~15-20% of NERSC's cycles

Architecture Evaluation and Application Readiness team

Nick Wright: Lead
Amber (Molecular Dynamics)
(proxy: NAMD, DLPOLY, LAMMPS, Gromacs)

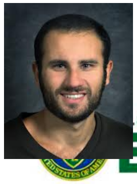
Katie Antypas: FLASH (explicit hydro)
(proxy: Castro, MAESTRO, S3D?, AMR)

Harvey Wasserman: POP
(proxy: CESM)

Jack Deslippe: Quantum Espresso, Berkeley GW (proxy: VASP, PARATEC, Abinit, PETot, Qbox, p2k)

Woo-Sun Yang: CAM (Spectral Element)
(proxy: CESM)

Matt Cordery: MPAS (Scalable Ocean Model)
(proxy: CESM)

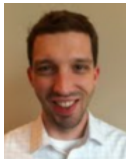


Lenny Oliker: GTC (PIC - Fusion)
(proxy: GTS, XGC, Osiris, g2s)

Brian Austin: Zori (QMC)
(proxy: qwalk)

Kirsten Fagnan: BLAST**,** Allpaths
(Bioinformatics)

Burlen Loring: Vis, Iso-surface



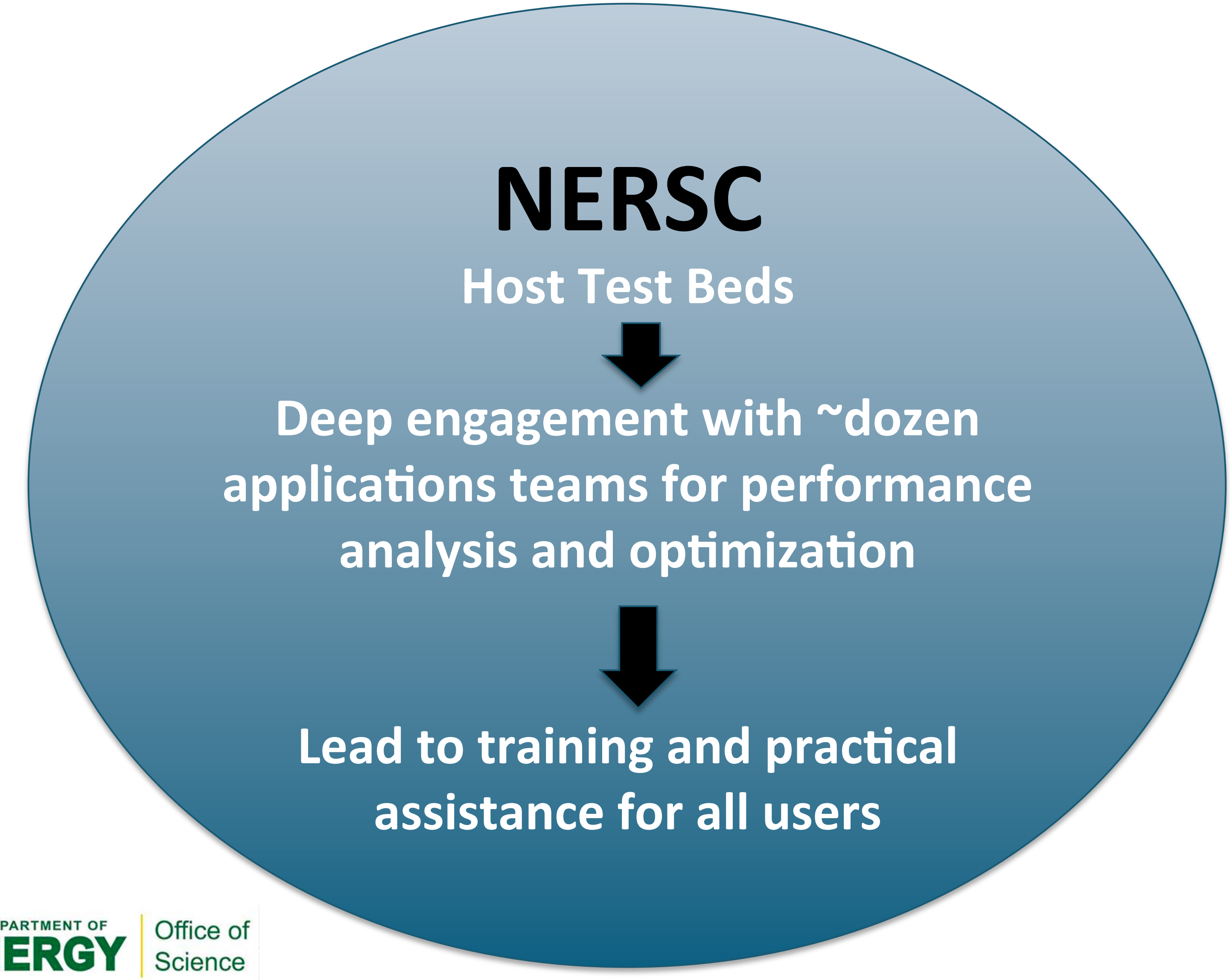
Aaron Collier: MADAM-toast (CMB), Gyro (Fusion)

Hongzhang Shan: NWChem
(proxy: qchem, Gamess)

Helen He: WRF (Regional Climate)

Hari Krishnan: Vis, Iso-surface

NERSC will help users transition to new architectures



NERSC

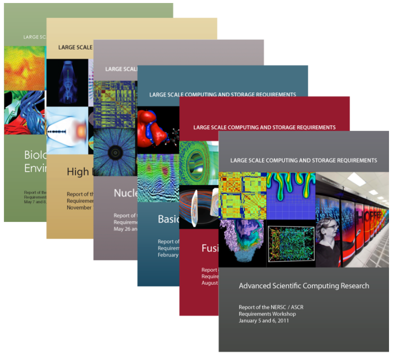
Host Test Beds

Deep engagement with ~dozen applications teams for performance analysis and optimization

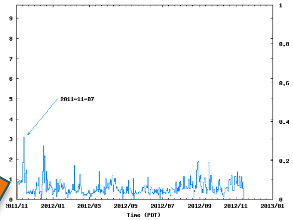
Lead to training and practical assistance for all users

Many people, teams and activities providing input to NERSC-8 procurement team

Requirements Workshops

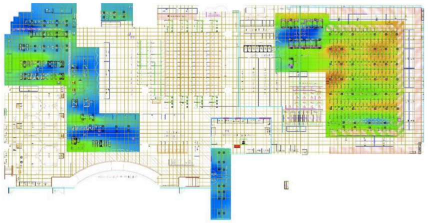

LARGE SCALE COMPUTING AND STORAGE REQUIREMENTS
Biolo
Envir
High
Nucle
Basic
Fusi
Advanced Scientific Computing Research
Report of the NERSC / ASCR
Requirements Workshop
January 5 and 6, 2011

Workload Analysis

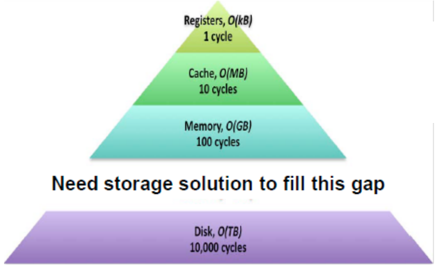


NERSC-8 Procurement

Power Management Team



NVRAM team



Registers, *O(kB)*
1 cycle

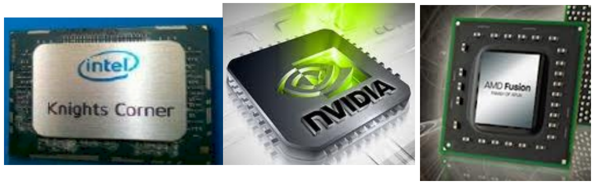
Cache, *O(MB)*
10 cycles

Memory, *O(GB)*
100 cycles

Need storage solution to fill this gap

Disk, *O(TB)*
10,000 cycles

Architecture Evaluation/ Application Readiness Team



CRT Building Project (NERSC-8 Site Prep)

HPC motifs are well represented by App-readiness applications (or proxies)

Science Areas	Dense Linear Algebra	Sparse Linear Algebra	Spectral Methods (FFTs)	Particle Methods	Structured Grids	Unstructured or AMR Grids
Accelerator Science	Vorpal	Vorpal	**IMPACT**, Vorpal	**IMPACT**, Vorpal	**IMPACT**, Vorpal	Vorpal
Astrophysics		MAESTRO, CASTRO, **FLASH**			MAESTRO, CASTRO, **FLASH**	MAESTRO, CASTRO, **FLASH**
Chemistry	GAMESS, **NWChem**, qchem, QWalk, **Zori**	QWalk, **Zori**		Qwalk, **Zori**, NAMD, **AMBER**, Gromacs, LAMMPS		
Climate		lesmpi, global_fcst	**CAM**, MITgcm		**CAM, POP, WRF, MPAS**, MITgcm, lesmpi, global_fcst	
Combustion					MAESTRO, **S3D**	
Fusion	Xaorsa		Xaorsa, NIMROD	**GTC**, XGC, Osiris, gs2, GTS	**GTC**, Xaorsa, **gyro**, NIMROD, XGC, Osiris, gs2	
Lattice Gauge	MFD	MFD, **MILC**, chroma, hmc, qlua	**MILC**, chroma, hmc, qlua	**MILC**, chroma, hmc, qlua	**MILC**, chroma, hmc, qlua	
Material Science	PARATEC, cp2k, **QE**, VASP, **BerkeleyGW**, Abinit, qbox, PETot		PARATEC, cp2k, **QE**, VASP, **BerkeleyGW**, Abinit, qbox, PETot		PARATEC, cp2k, **QE**, VASP, **BerkeleyGW**, Abinit, qbox, PETot	

Work in progress

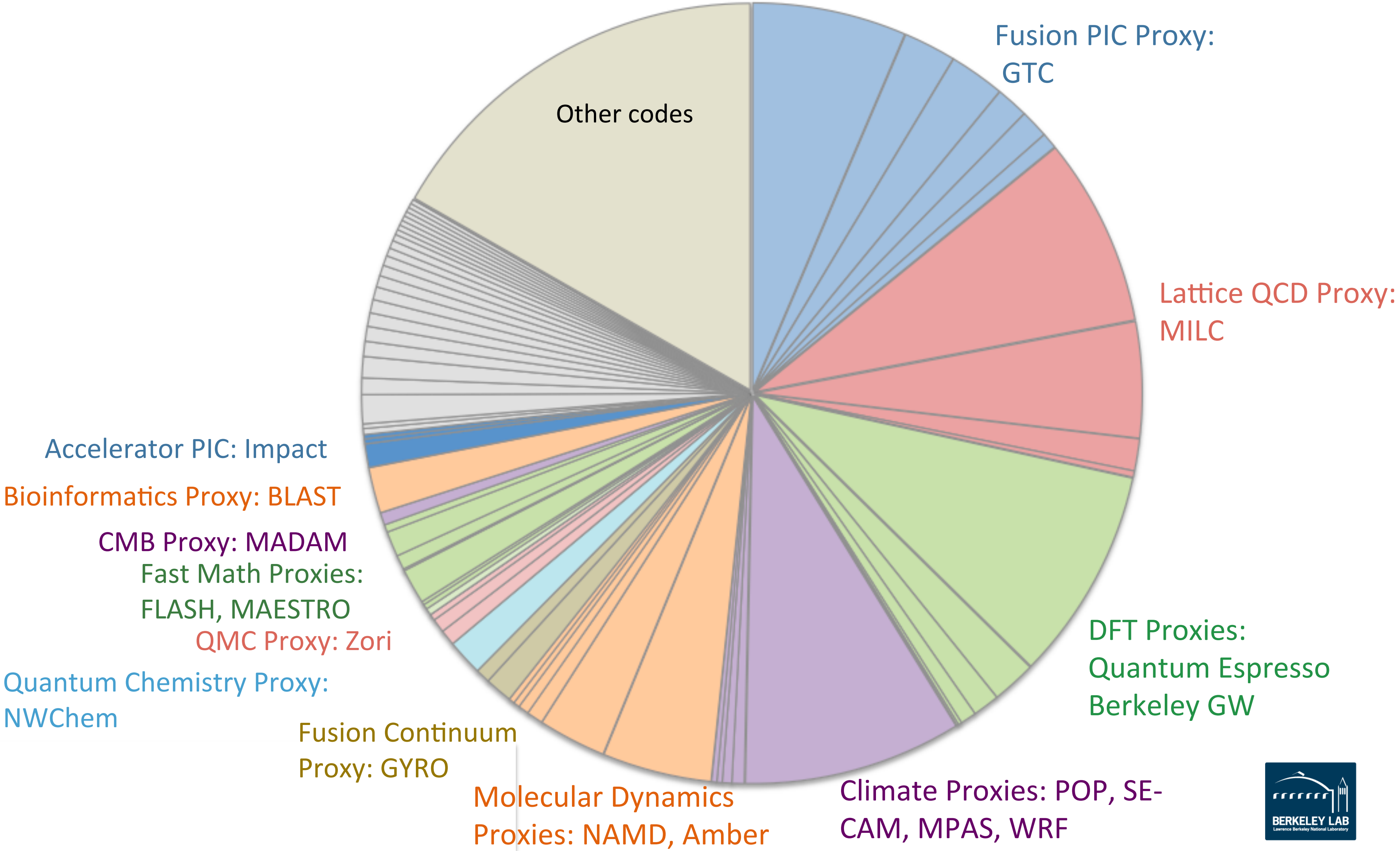
Application Readiness Coverage by Area Expertise

Work in progress

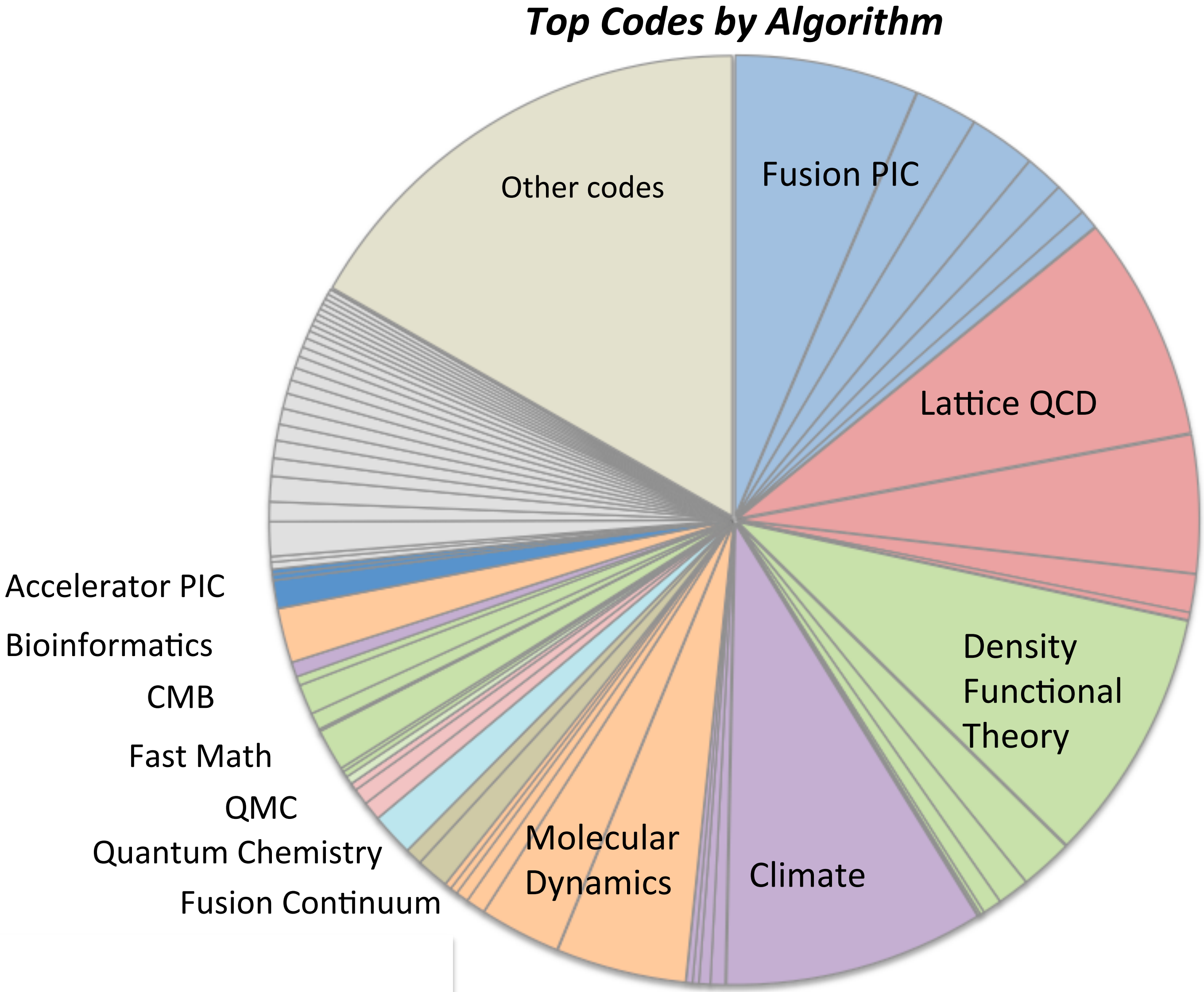
Science Area	Algorithm	Codes	NERSC / LBL Familiarity
Accelerator Physics	PIC	**IMPACT**, Vorpal	Brian, Sherry Li
Astrophysics	Explicit and implicit hydro-dynamics, multi-physics, block structured grids.	MAESTRO, CASTRO, **FLASH,** CHIMERA	Katie, JBB, Almgren, Kirsten
Bioinformatics	Sequence Alignment	BLAST, **AllPaths**	Kirsten, Shane/JGI
Materials Science	Density Functional Theory	CP2K, **QE**, VASP, **BerkeleyGW,** Paratec, Petot, qbox, Abinit	Jack, Zhengji
Chemistry	Electronic Structure (ab initio)	GAMESS, **NWChem**, qchem	Brian, Shan
	Electronic Structure (QMC)	QWalk, **Zori**	Brian
	Molecular Dynamics	NAMD, **AMBER**, Gromacs, LAMMPS	Brian, Nick
Climate	Spectral element (CAM) Explicit Finite difference/Finite Volume	CCSM/**CAM**, global_fcst, lesmpi, mitcgm, **WRF**, **POP**, **MPAS**	Matt, Helen, Woo-Sun, Harvey
Combustion	DNS Navier-Stokes	S3D	JBB, Shalf
Fusion	PIC	XGC, **GTC**, Osiris, gs2, GTS	Lenny
	MHD	NIMROD, M3D	🟨
	Eulerian gyrokinetic	**Gyro**	Aaron Colier
	EM	Xaorsa, Vorpal	🟥
Geoscience	EM / Geo	EM3D_Inv	Commer/Newman
Nuclear Physics	QCD	**MILC**, chroma, hmc, qlua	(Active community)
	Lanczos eigensolver	MFD	Shan? (similar to Ab Initio?)
Data Analysis /Vis	Iso-surface rendering	**MADAM**	Burlen Loring/Hari Krishnan Aaron Collier

The App Readiness team proxies cover almost 75% of the workload

Top Codes by Algorithm and Application Readiness Coverage



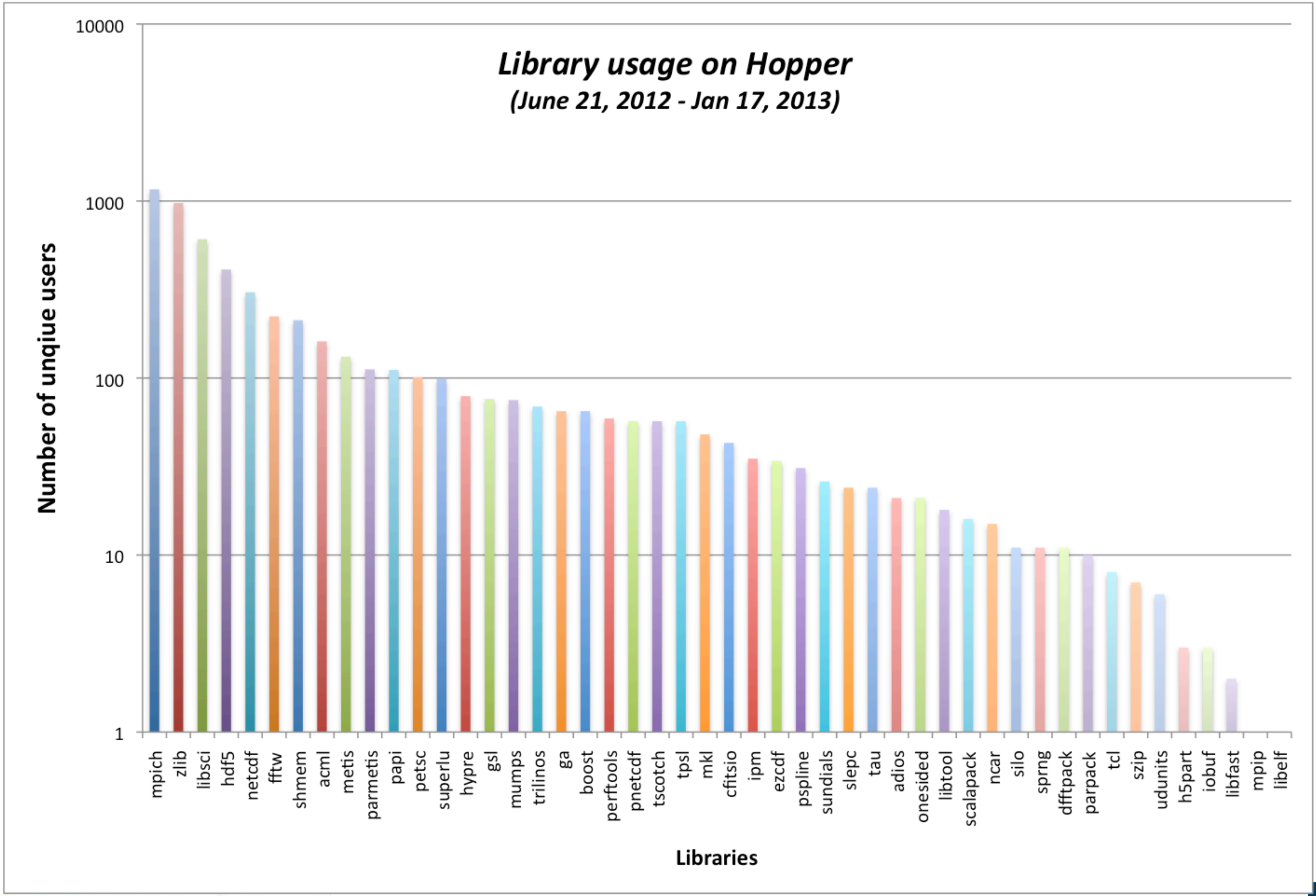
The App Readiness team proxies cover almost 75% of the workload



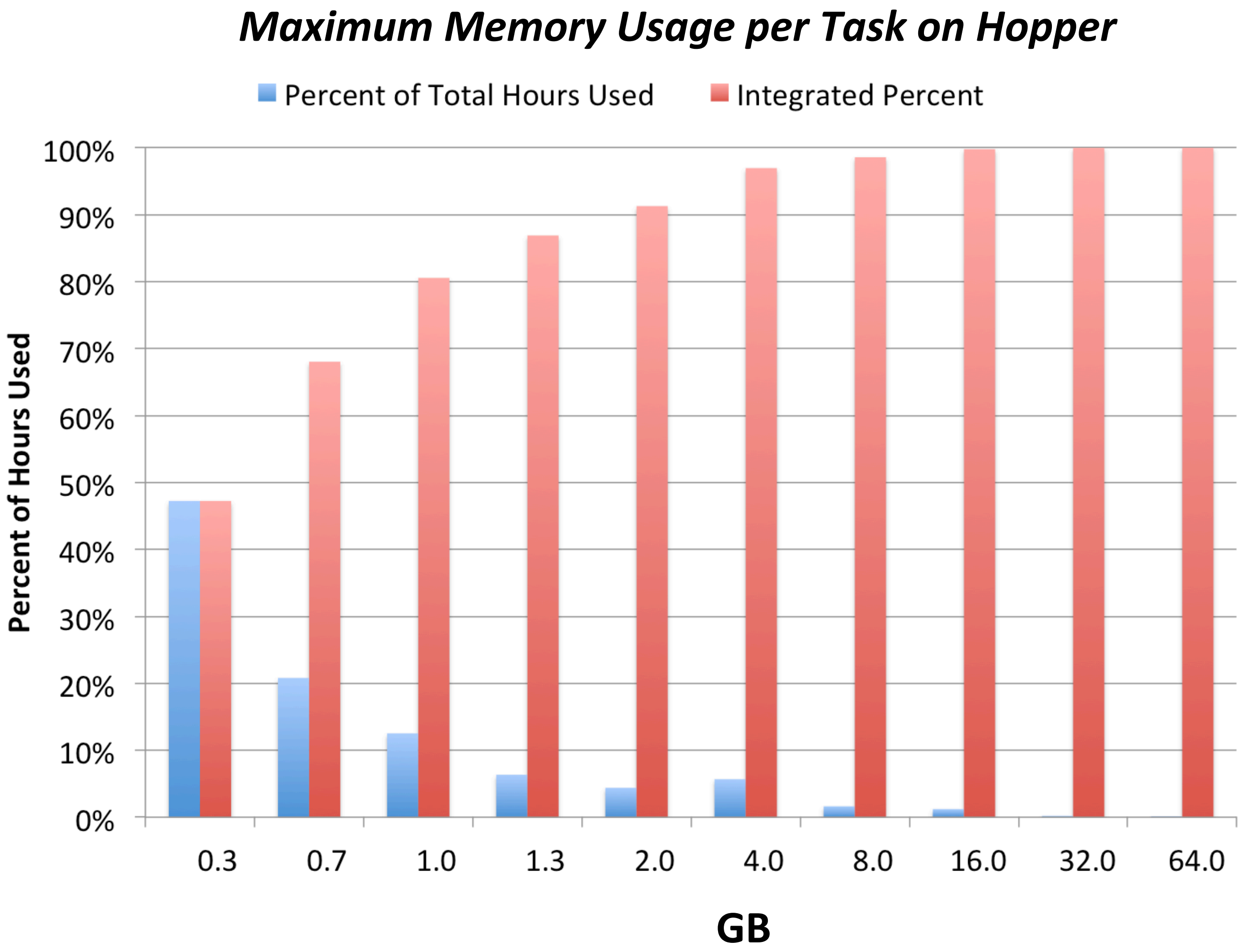
Many activities going on in the center to help better understand the NERSC workload

- I/O Characterization
- Memory usage on Hopper
- Library usage with ALTD

ALTD Data

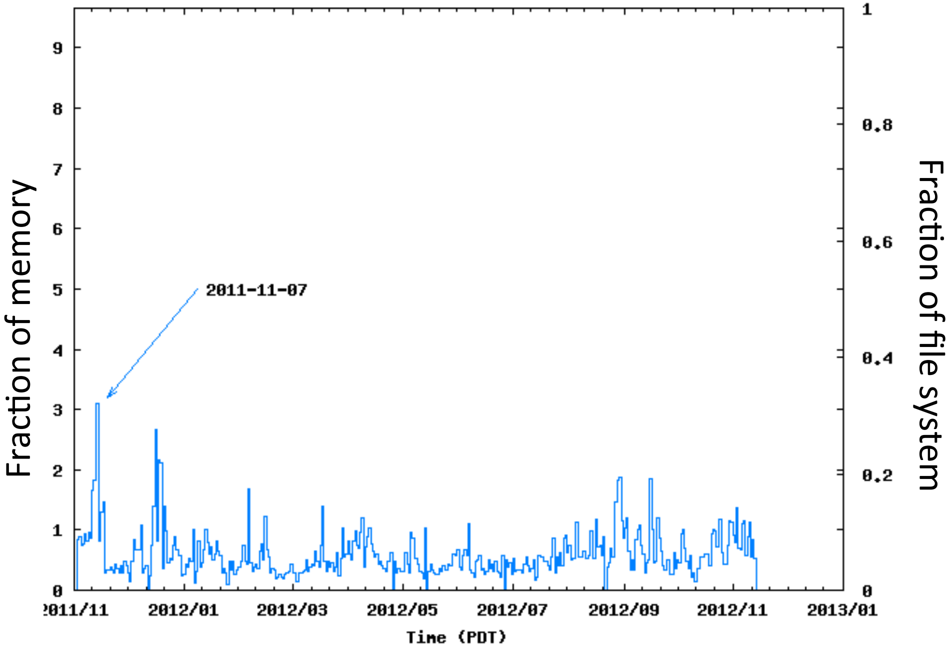

Library usage on Hopper
(June 21, 2012 - Jan 17, 2013)
Number of unqiue users
10000
1000
100
10
1
Libraries
mpich
zlib
libsci
hdf5
netcdf
fftw
shmem
acml
metis
parmetis
papi
petsc
superlu
hypre
gsl
mumps
trilinos
ga
boost
perftools
pnetcdf
tscotch
tpsl
mkl
cfitsio
ipm
ezcdf
pspline
sundials
slepc
tau
adios
onesided
libtool
scalapack
ncar
silo
sprng
dfftpack
parpack
tcl
szip
udunits
h5part
iobuf
libfast
mpip
libelf

Memory usage on Hopper

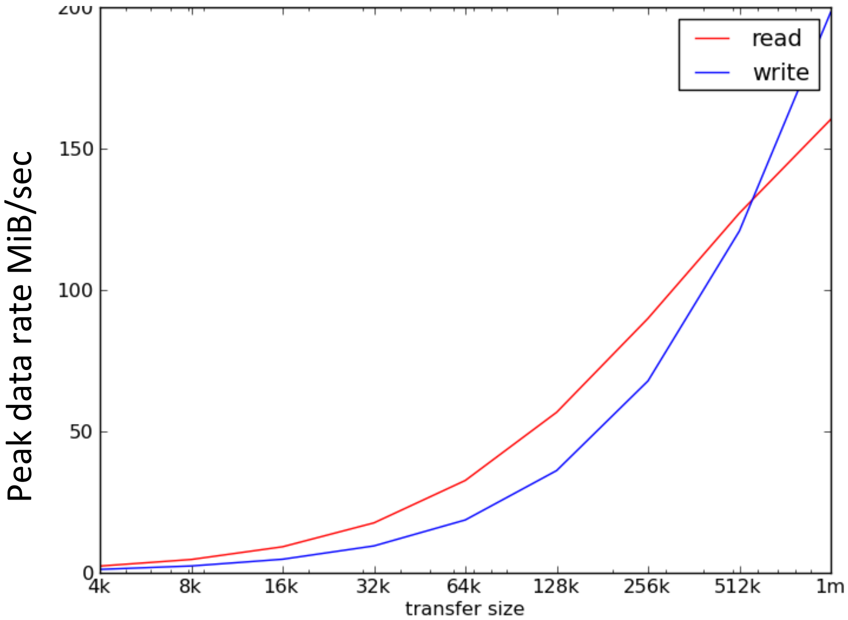


Use LMT to understand file system utilization as a function of I/O transfer size

Fraction of memory written per day



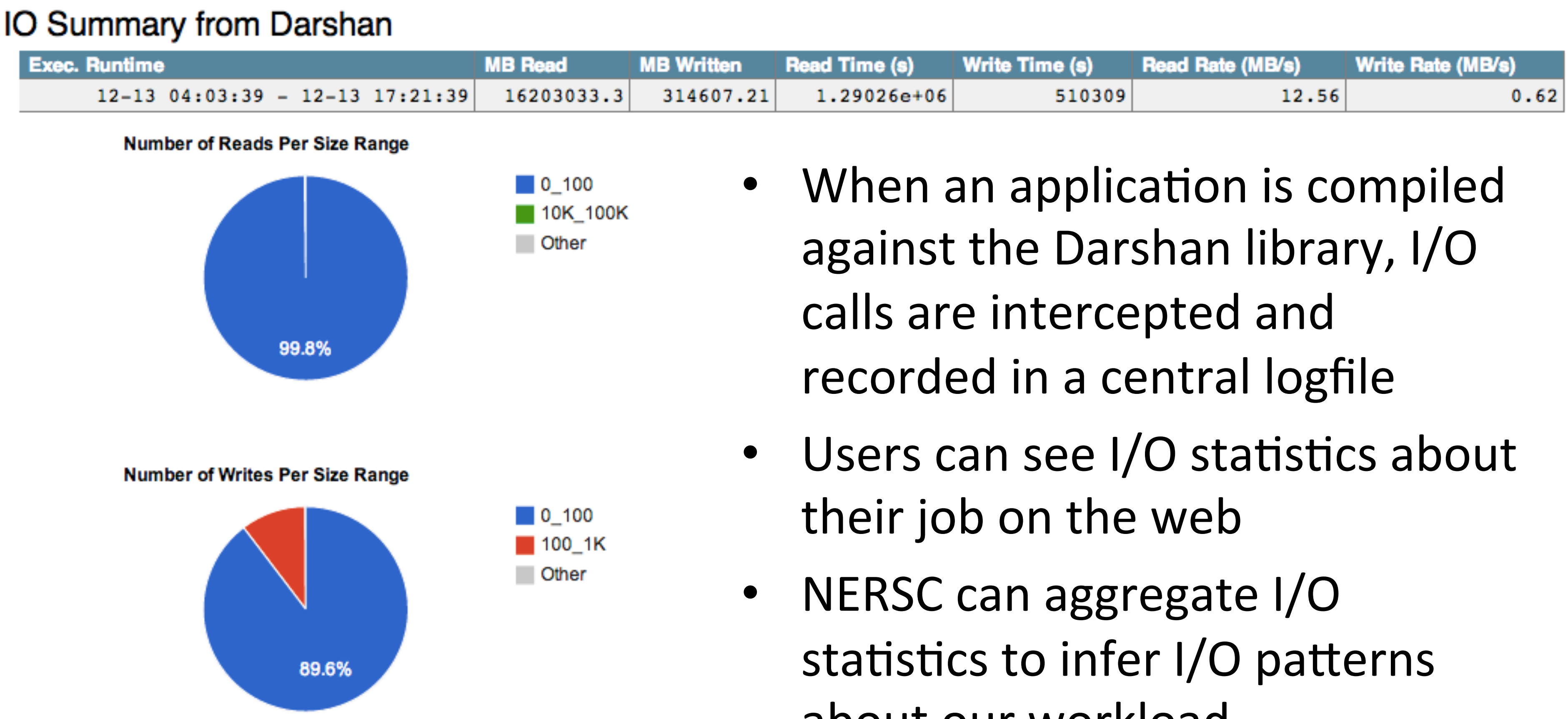
The Effect of Transfer size on Peak Aggregate I/O Rate



- •Use LMT to understand relationship between file system utilization and I/O transfer size
- •Is the Hopper file system more heavily utilized than appears due to un-optimal transfer sizes and distributed I/O servers on Lustre?

Andrew Uselton, Nick Wright

Application I/O statistics are now available with Darshan



IO Summary from Darshan

Exec. Runtime	MB Read	MB Written	Read Time (s)	Write Time (s)	Read Rate (MB/s)	Write Rate (MB/s)
12-13 04:03:39 - 12-13 17:21:39	16203033.3	314607.21	1.29026e+06	510309	12.56	0.62

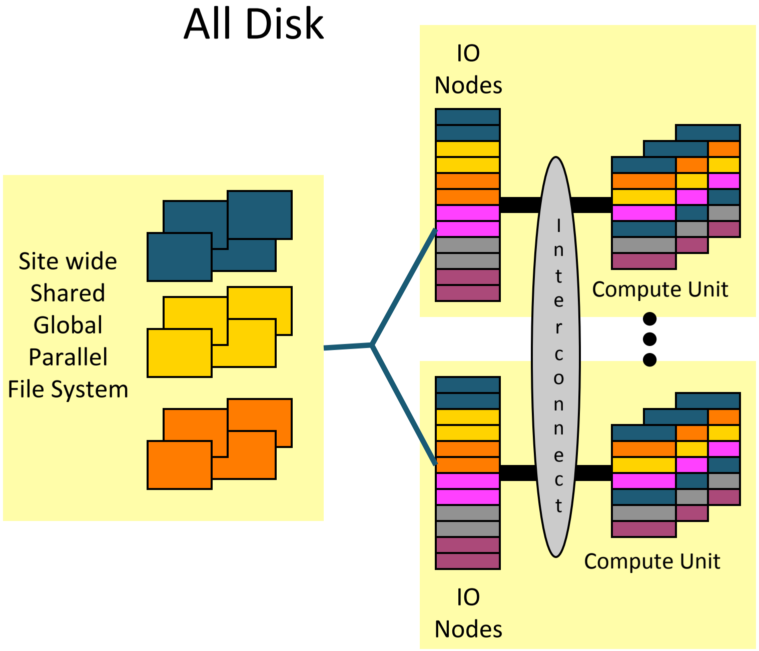
Number of Reads Per Size Range
99.8%
0_100
10K_100K
Other
Number of Writes Per Size Range
89.6%
0_100
100_1K
Other

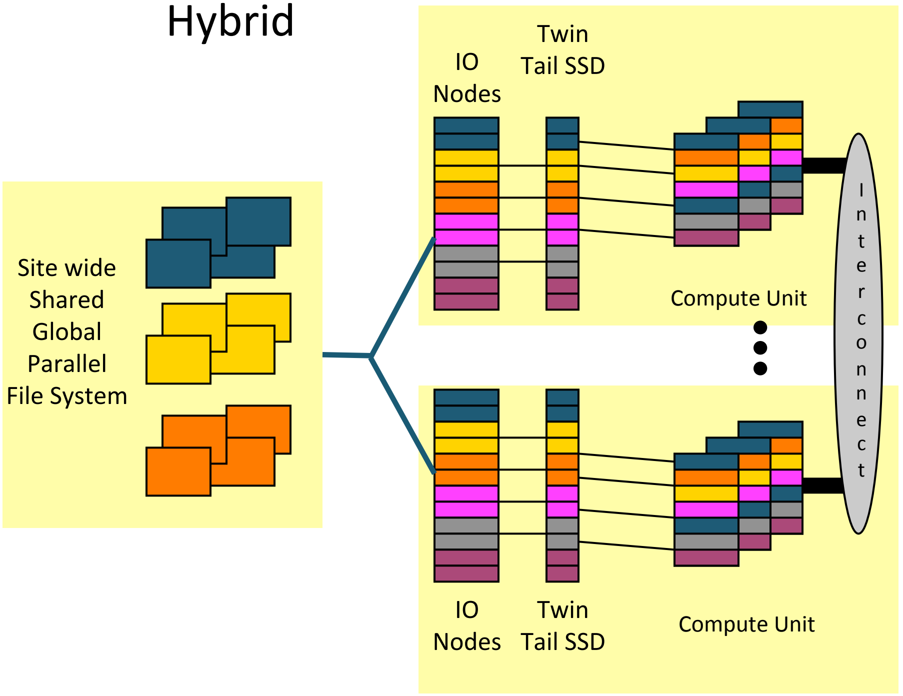
- When an application is compiled against the Darshan library, I/O calls are intercepted and recorded in a central logfile
- Users can see I/O statistics about their job on the web
- NERSC can aggregate I/O statistics to infer I/O patterns about our workload.

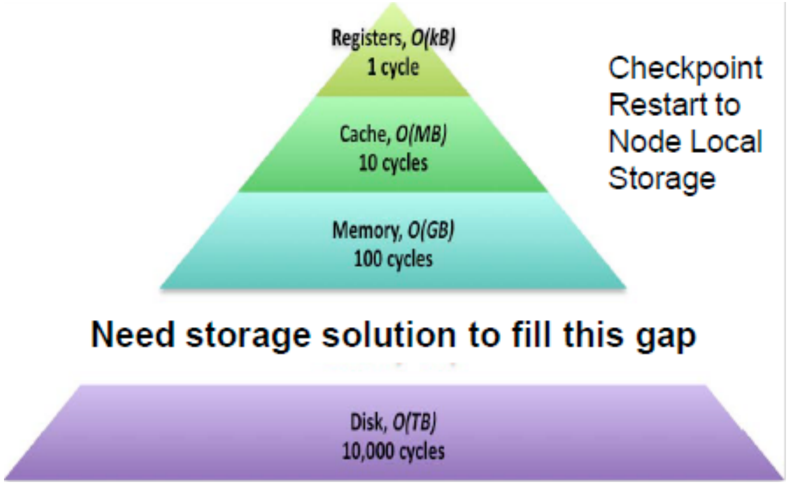
Could NERSC workload benefit from an intermediate I/O layer of NVRAM?

NVRAM Team	*Questions*
• Nick Wright • Matt Andrews • Rei Lee • Shane Canon • Andrew Uselton • Yushu Yao • Katie Antypas • Jeff Broughton • Jason Hick • Brian Austin • David Skinner • Nick Cardo	• What are the use cases for NVRAM in a supercomputer? • How much of the NERSC workload could benefit from NVRAM in an HPC system? • What are the cost/benefit trade-offs in terms of disk capacity, I/O bandwidth and compute? • At what level should NVRAM sit? (compute node? I/O node? Disk?) • What software development would be needed to allow productive use of NVRAM on NERSC-8?

Hypothesis: purchase disk for capacity and NVRAM for bandwidth







- Previously purchased disk for capacity. Now we purchase disk for bandwidth
- Too expensive to purchase NVRAM for capacity
- Hybrid approach would offer best value

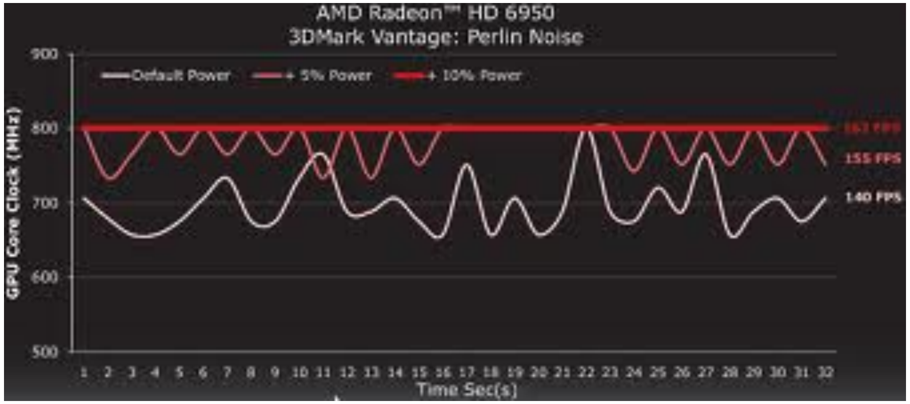
Image Gary Grider, LANL

Possible NVRAM Use Cases

- Checkpoint/restart 'burst buffer' to improve reliability and application time to solution for large-scale applications (Trinity primary use case)
- Speed-up I/O read/write operations (NERSC primary use case?)
- In-situ visualization
- Stage shared libraries
- Database driven workloads
- Relieve stress to disk file system by streaming I/O in more favorable pattern
- Fast I/O for out-of-core applications

Power Team

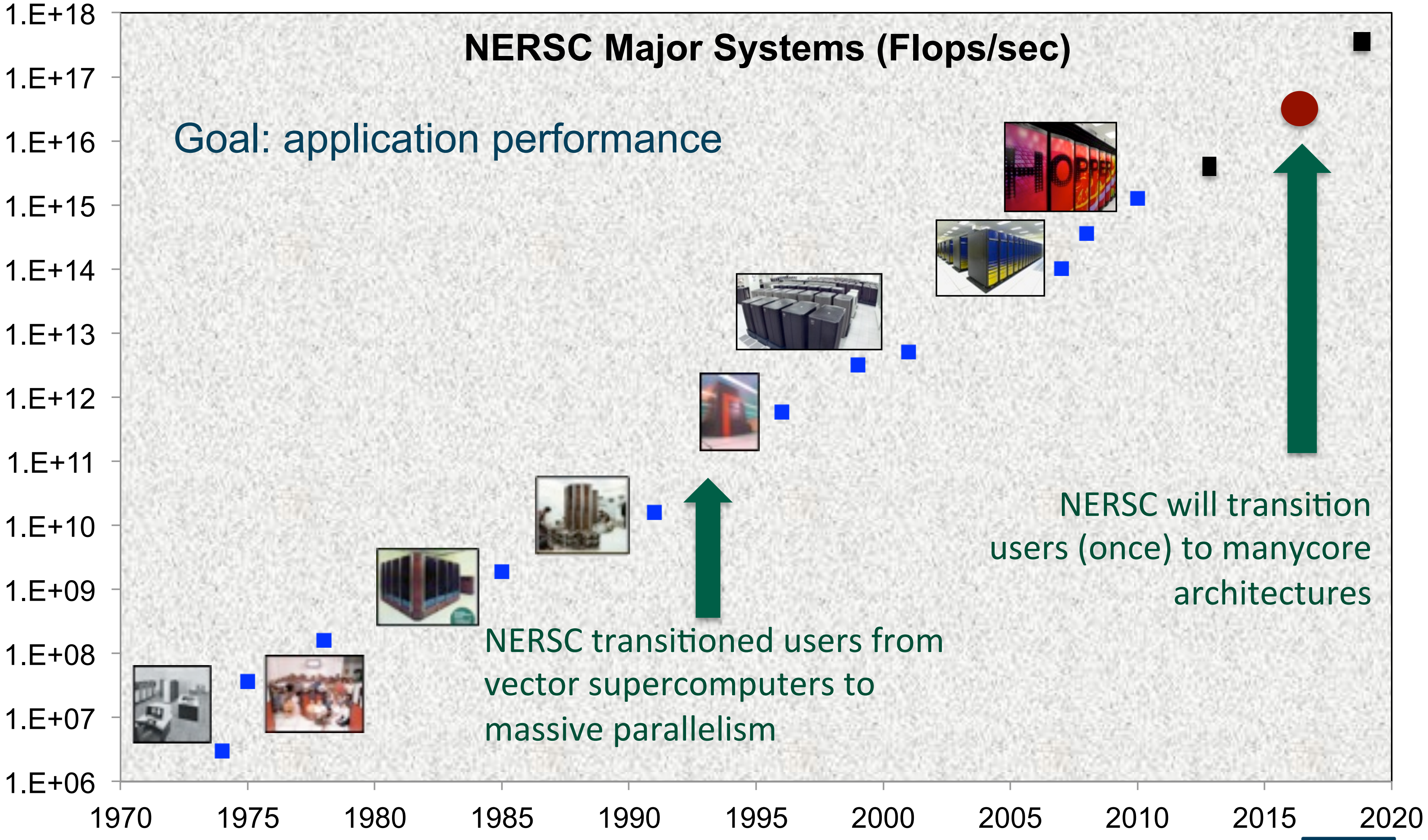
- Team Members
  - Jim Craw
  - Tom Davis
  - Tina Butler
  - Nick Cardo
  - Nick Wright
  - Jeff Broughton



- Questions
  - Understand future vendor capabilities for active power management and power capping
  - Explore power management and monitoring experiments on current systems

Image from AMD

NERSC plan will take scientists through technology transition



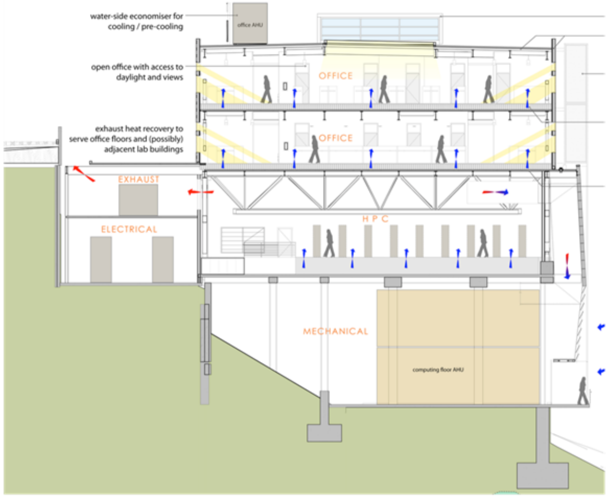
Goal: application performance

NERSC transitioned users from vector supercomputers to massive parallelism

NERSC will transition users (once) to manycore architectures

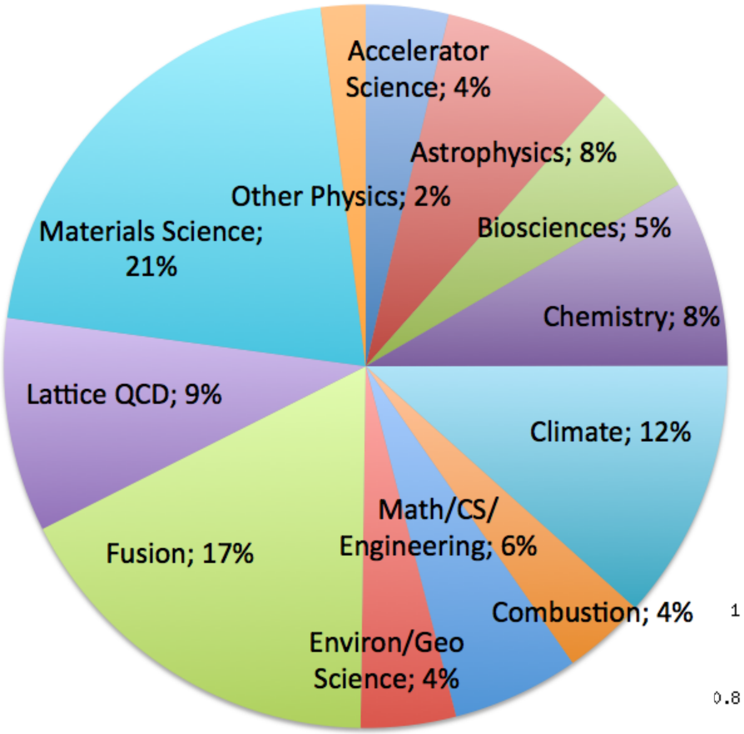
We are deploying the CRT facility to meet the ever growing computing and data needs of our users

- **Four story, 140,000 GSF**
  - Two 20Ksf office floors, 300 offices
  - 20K -> 29Ksf HPC floor
  - Mechanical floor
- **42MW to building**
  - 12.5MW initially provisioned
  - WAPA power: Green hydro
- **Energy efficient**
  - Year-round free air and water cooling
  - PUE < 1.1
  - LEED Gold
- **Occupancy Early 2015**


water-side economiser for cooling / pre-cooling
office AHU
open office with access to daylight and views
OFFICE
exhaust heat recovery to serve office floors and (possibly) adjacent lab buildings
OFFICE
EXHAUST
ELECTRICAL
HPC
MECHANICAL
computing floor AHU

NERSC's challenge is to transition diverse and broad workload to future architectures

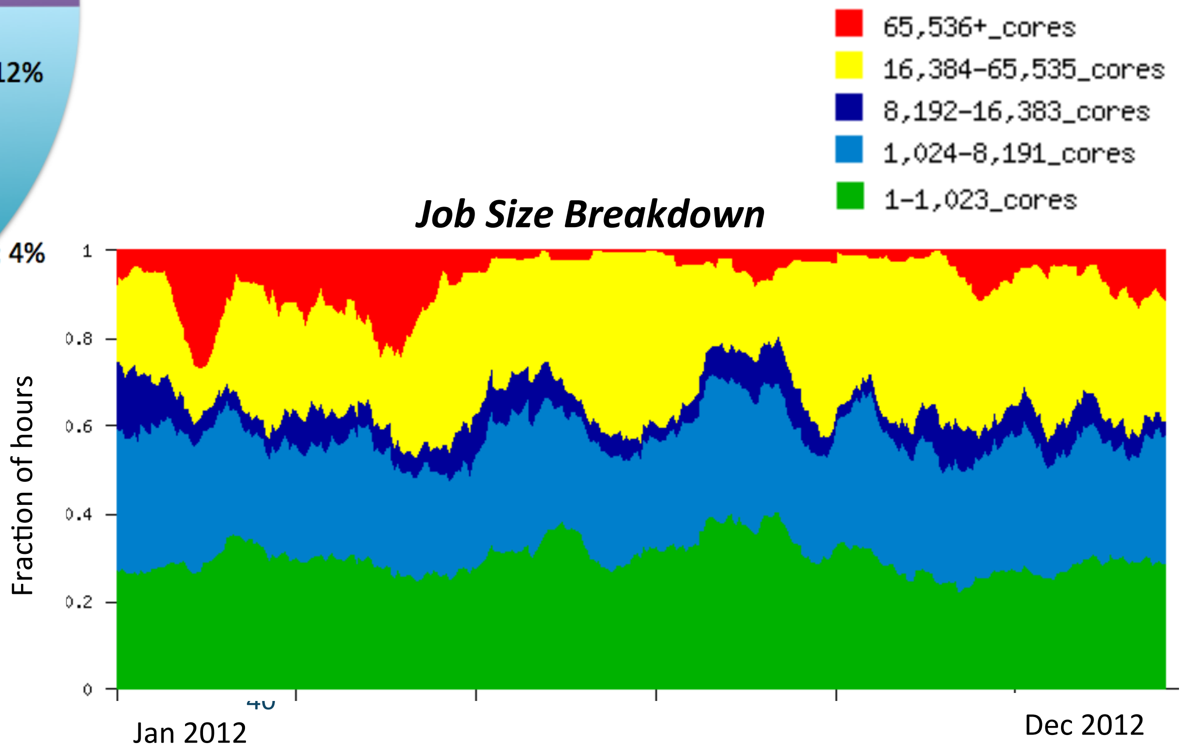
2011 Allocation Breakdown



NERSC serves:

- Over 4500 users
- Over 650 projects

Job Size Breakdown



Options in the final RFP will allow each site to further customize a system

- Visualization partition
- Burst Buffer (more later)
- Advanced power management
- Application transition support
- Early access development systems and testbeds

Conclusions

- NERSC-8 project is moving forward!
- Challenges working with another Lab (communication, distance), but also benefits (expanded technical breadth, outside perspective)
- Many people at NERSC outside the core NERSC-8 procurement team working to make project a success
- We expect a busy (hectic) spring with 2 project reviews before the final RFP release in the summer