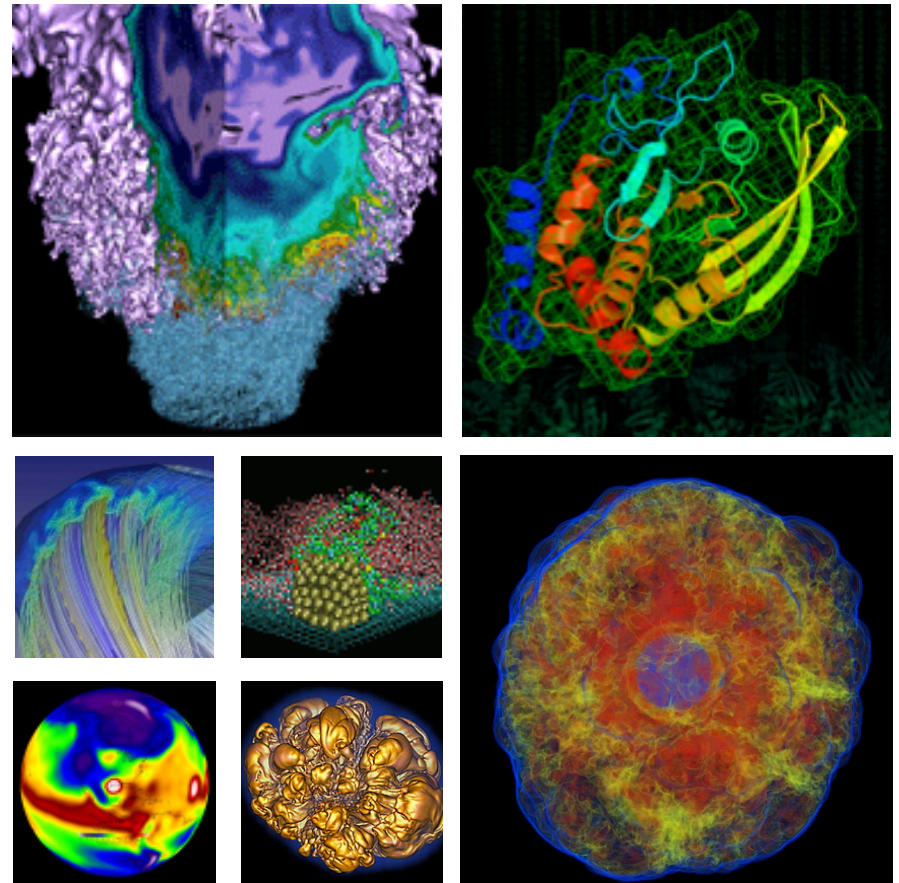


NERSC Operational Assessment Review Highlights



Katie Antypas
NERSC Deputy for Data Science

March 24, 2016

NERSC Operational Review Details



- **Every year NERSC prepares a report with operational highlights for DOE**
- **Every 3rd to 4th year, we are reviewed onsite**
- **2016 NERSC Operational Review February 16-17**
- **5 Reviewers**
 - Brett Bode – Chair, NCAR
 - Buddy Bland – OLCF
 - Susan Coghlan – ALCF
 - Steve Hammond – NREL
 - Bailent Joo - JLab

Charge Questions



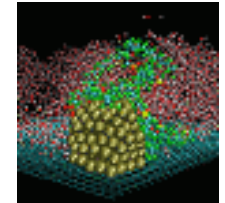
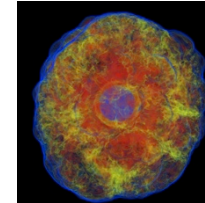
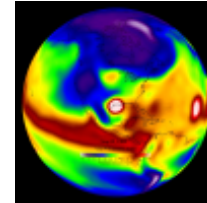
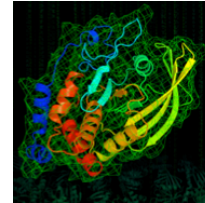
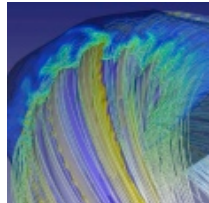
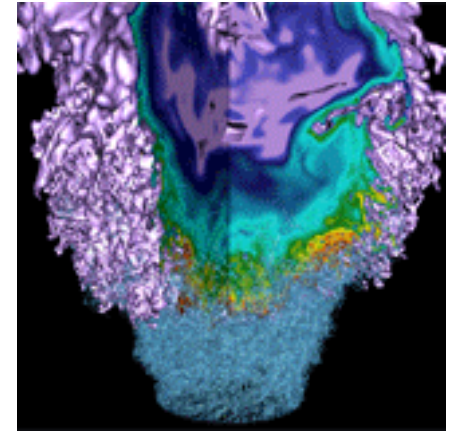
- **Are the processes for supporting the users/customers, resolving problems, and outreach effective?**
 - [Katie's User Results Presentation, Sudip's overview](#)
- **Is the science output commensurate with NERSC's mission to 'accelerate the pace of scientific discovery through high performance computing and data analysis'?**
 - [Richard's Science and Strategic Results presentation](#)
- **Is NERSC optimizing the use of its resources consistent with its mission? Was the FY2015 NERSC operations budget reasonable? Is the Spend Plan for FY 2016 reasonable?**
 - [Jackie's systems results presentation](#)
 - [Richard will address allocations and usage in Strategic results](#)
 - [Jeff will give tour of CRT Wang Hall](#)
 - [Katie will cover budget at end of the day](#)

Charge Questions



- **What innovations have been implemented that have improved NERSC operations?**
 - Innovations presentation (Shane, Glenn, Liz)
- **Is NERSC effectively managing risk and ensuring safety?**
 - Katie's presentation at end of the day
- **Are the performance metrics used for the review year and proposed for future years sufficient and reasonable for assessing NERSC Operational performance.**
 - Katie's presentation at the end of the day

Are the processes for supporting the users/customers, resolving problems, and outreach effective?



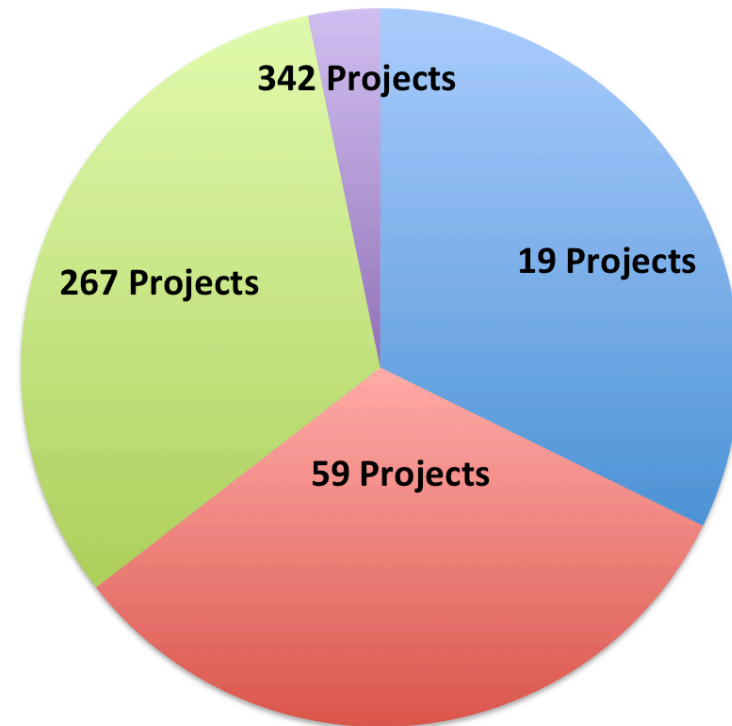
A Deeper Dive into our Users



- **Nearly 6000 users**

Category of Users	Count
Total NERSC Users (NERSC Program, JGI, PDSF)	6,568
Total JGI users	296
JGI users who are also NERSC Program users	184
Total PDSF users	731
PDSF users who are also NERSC Program users	273
Storage only users	671
NERSC Program Users	5,998
NERSC Users who submitted jobs	3,467

Breakdown of 3.17 Billion Hours by Project



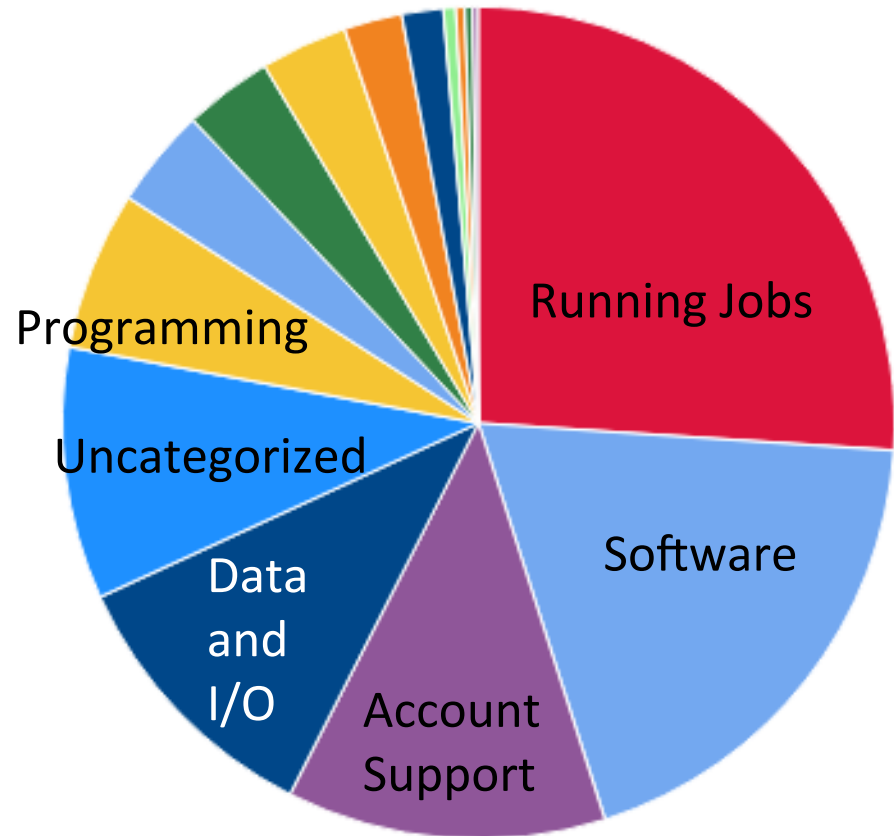
- In recent years we've significantly increased the number of user facing staff
- As recently as 2013 we had ~8 consultants and 2 account support staff handling majority of engagements with users
- Deeper engagements are necessary to prepare for advanced architectures (Cori) and new data intensive workflows
- We now have 4 groups interacting with users
 - User Engagement Group (7 staff)
 - Application Performance Group (5 staff)
 - Data and Analytics Services Group (9 staff)
 - Data Science Engagement Group (3 staff, new group)
- A cross group team provides consulting services to NERSC's ~6000 users

Tickets by Category



- 8,422 tickets were submitted by NERSC program users in 2015
- Compared to 2014 consultants are fielding more questions about hardware, performance, profiling, and data and I/O
- **Our service level agreement**
 - Respond to all tickets within 4 business hours
 - Assure 80% of tickets are resolved or have a communicated path to resolution within 3 business days

2015 Tickets by Category



- **Goal: Prepare DOE Office of Science user community for Cori manycore architecture**
- **Partner closely with ~20 application teams and apply lessons learned to broad SC user community**
- **NESAP activities include:**



Strong support from vendors

Developer Workshops for 3rd-Party SW

Early engagement with code teams

Leverage existing community efforts

Postdoc Program

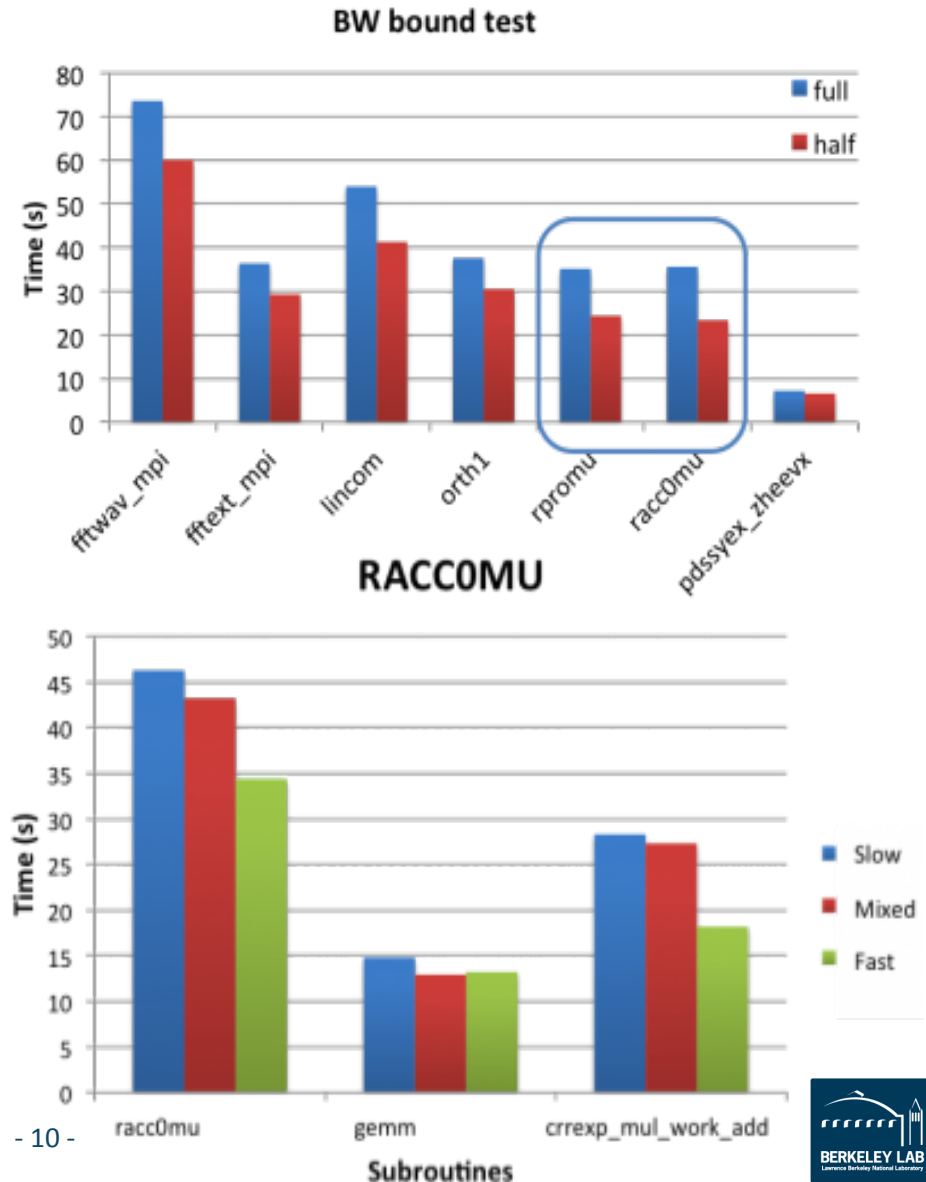
NERSC training and online modules

Early access to KNL technology

VASP – Studying the effects of high bandwidth memory on performance



- Widely used materials science code, consumes most time at NERSC
- Dungeon session examined benefits from high bandwidth memory
- Significant speed ups when only key arrays are placed in fast memory – promising outcome



Work by Martijn Marsman, Zhengji Zhao

Intel Xeon Phi User Group (IXPUG)



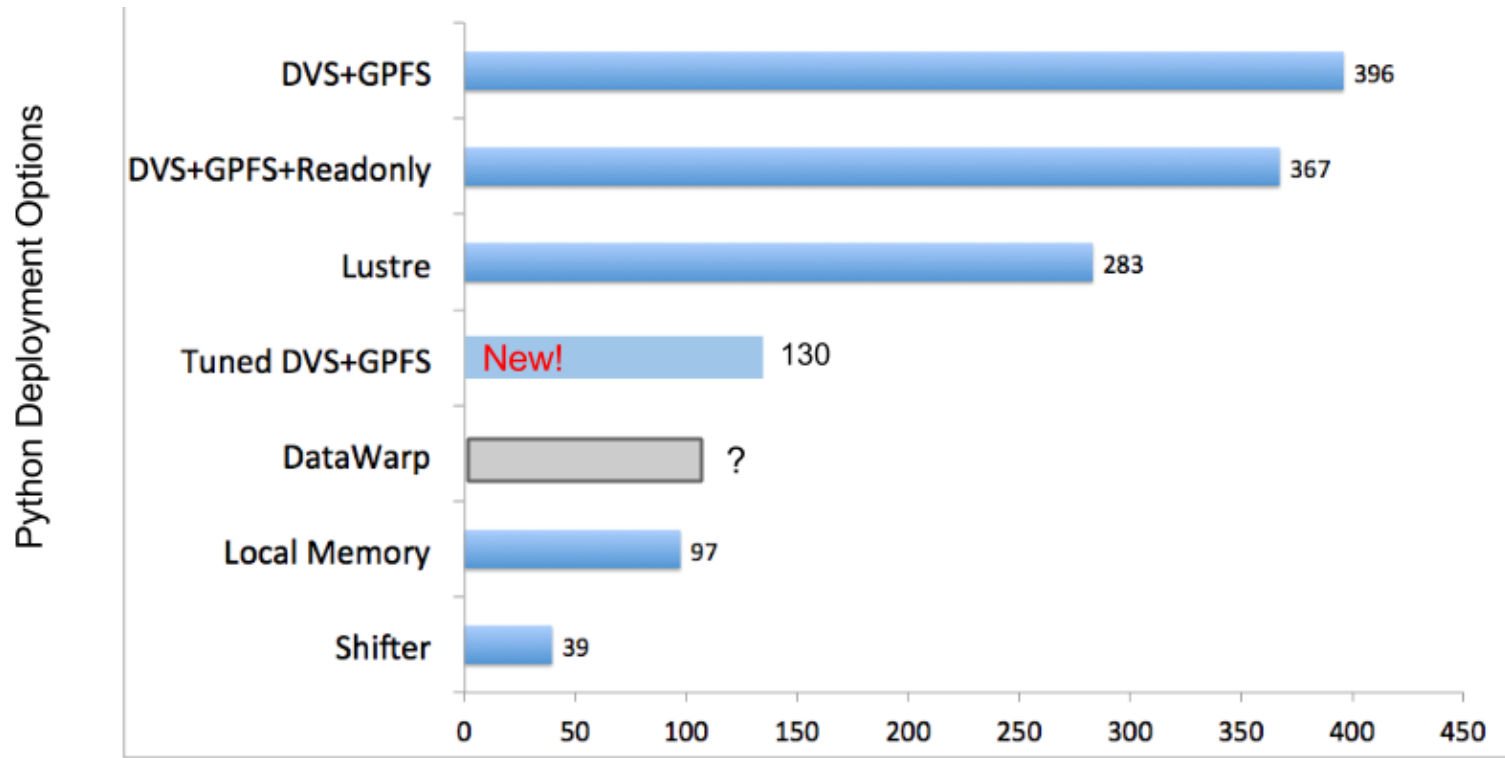
- **NERSC hosted IXPUG 2015 at the CRT facility**
- **Over 100 attendees**
- **Week long community event with training sessions, hackathons and technical briefings and community meetings**
- **DFT for Exascale community workshop on last day**



Improving Python Performance



- Python is a critical tool for many data intensive applications
- We have tried various ways to improve python performance on Cray systems



Time (seconds) to complete Python Performance (Test on 300 nodes, 24 cores per node)

Transition to SLURM to better support data intensive science



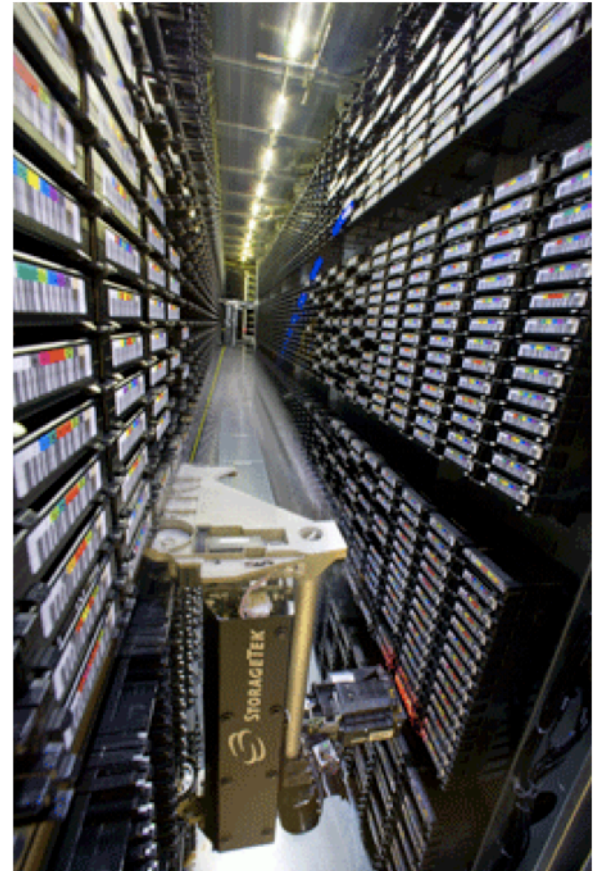
- **NERSC made the switch from the Torque/Moab scheduler to SLURM on both Edison and Cori**
- **Open source, NERSC can contribute to development**
- **Enables a number of features for data intensive science**
 - Real time queues
 - High throughput queues



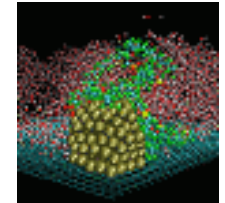
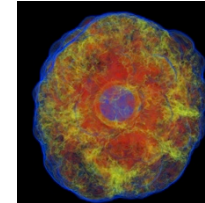
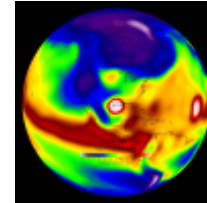
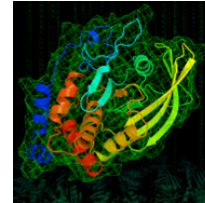
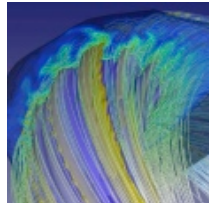
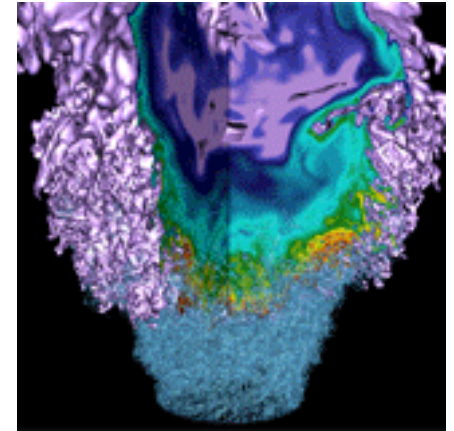
Increased Archive Disk Cache improves HPSS performance for users



- Disk cache increased by 10x from 200TB to over 2PB
- Before the increase, files stayed on disk cache for 2.5 days and now stay on 24.5 days (10x improvement)
- Impact for users is enormous, latency to tape is 90 seconds while disk cache is < 1 sec
- Of the files read, 75% are read within 30 days of writing – disk cache close to optimal capacity



Is the science output commensurate with
NERSC's mission to 'accelerate the pace of
scientific discovery through high
performance computing and data analysis'?



Focus on Science



- NERSC supports the broad mission needs of the six DOE Office of Science program offices
- 6,000 users and 750 projects
- MPP (supercomputing) and data-only users
- NERSC science engagement team provides outreach and POC

2,078 refereed publications in 2015




Table format?

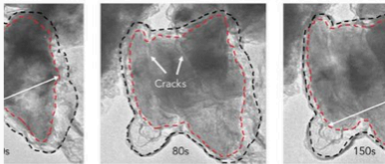
NERSC Science in the News




Social Media

 **Richard Gerber** @ragerber 2d
 #NERSC Coupling 2 'tabletop' laser-plasma accelerators: A step toward ultrapowe...
bit.ly/20vcEoR via @BerkeleyLab @EurekAlertAAAS
 Open

 **NERSC** @NERSC 2d
 RT @gizmag: Lithium-ion battery boost could come from "caging" silicon in graphene -
gizm.ag/1Vye4M6
pic.twitter.com/dn5i0kdKhr

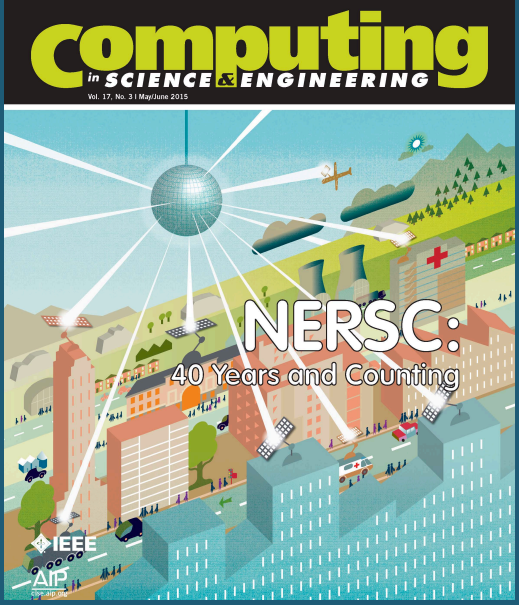


Open

 **Glenn K. Lockwood** @glenn... 2d
 NERSC hosting Advanced OpenMP workshop on Feb 4, led by members of OpenMP Lang Committee. Webcast avail, reg: nersc.gov/users/training...

NERSC's Impact on Advances of Global Gyrokinetic PIC Codes for Fusion Energy Research, Ethier, S. ; Choon-Seock Chang ; Seung-Hoe Ku ; Wei-li Lee ; Weixing Wang ; Zhihong Lin ; Tang, W.

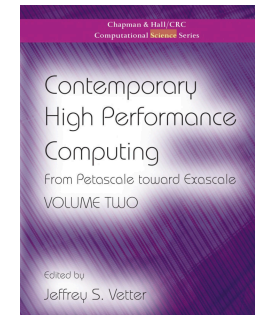
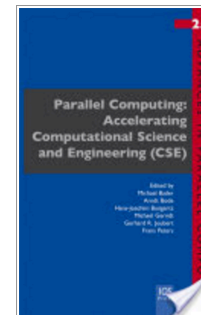
Big Bang, Big Data, Big Iron: Fifteen Years of Cosmic Microwave Background Data Analysis at NERSC, Borrill, J. ; Keskitalo, R. ; Kisner, T.



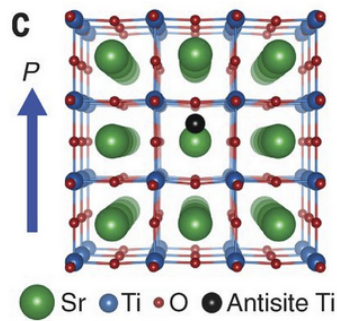
NERSC Annual Reports



NERSC staff author book chapters on science accomplishments

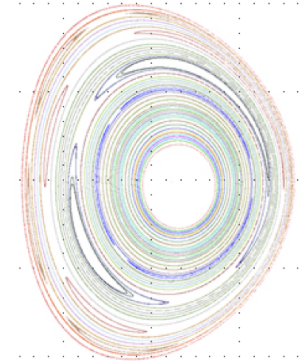


NERSC Sends Quarterly Highlights to DOE



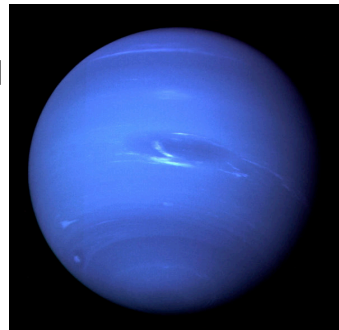
Materials Science
Theoretical calculations help provide evidence of room-temperature ferroelectricity in nanometer-thick films (Xifan Wu, Temple Univ., *Science*)

Fusion Energy
3D simulations run at NERSC help gain new insights into fusion plasma behavior that will improve the ability to stabilize a tokamak reactor (S. Jardin, Princeton Plasma Physics Lab, *Phys. Rev. Lett.*)



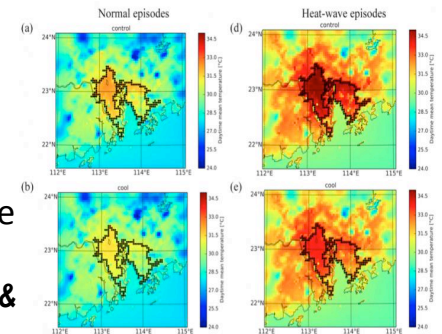
Chemistry

Simulations run at NERSC lead to the prediction of a new phase of superionic ice, a special form of ice that could exist on Uranus and Neptune (Roberto Car, Princeton U., *Nature Comm.*)



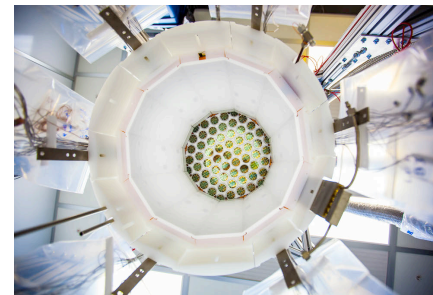
Energy

Computer models run at NERSC determine that, during a heat wave, white roofs can help mitigate the urban heat island effect (Dev Millstein, LBNL, *Env. Sci. & Tech.*)



High Energy & Nuclear Physics

Simulations run at NERSC are helping the Large Underground Xenon (LUX) dark matter experiment better focus their search for dark matter particles (R. Jacobsen, LBNL, *Phy. Rev. Lett.*)



Industrial Partnerships

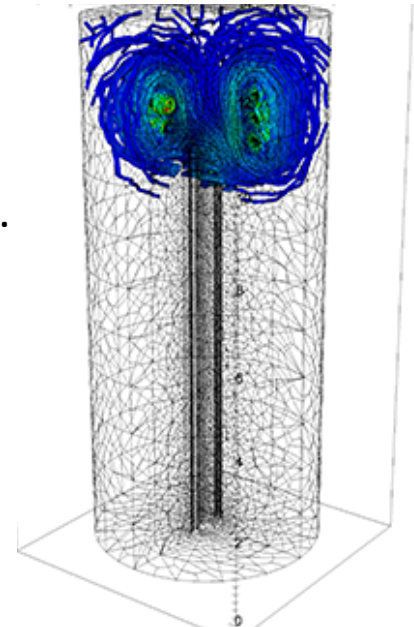


- **Separate SBIR allocation pool now in place**
- **130 industry users**
- **50 companies**
- **Added 9 new projects in energy-specific fields in 2015**

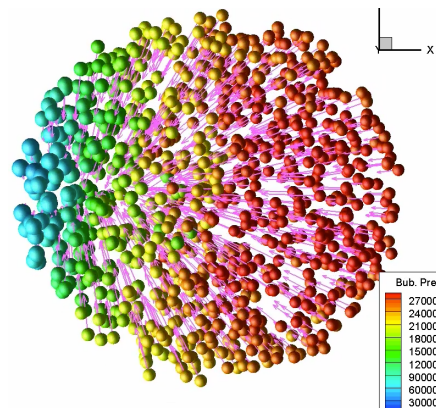
Vertum Partners



Tech-X, Inc.



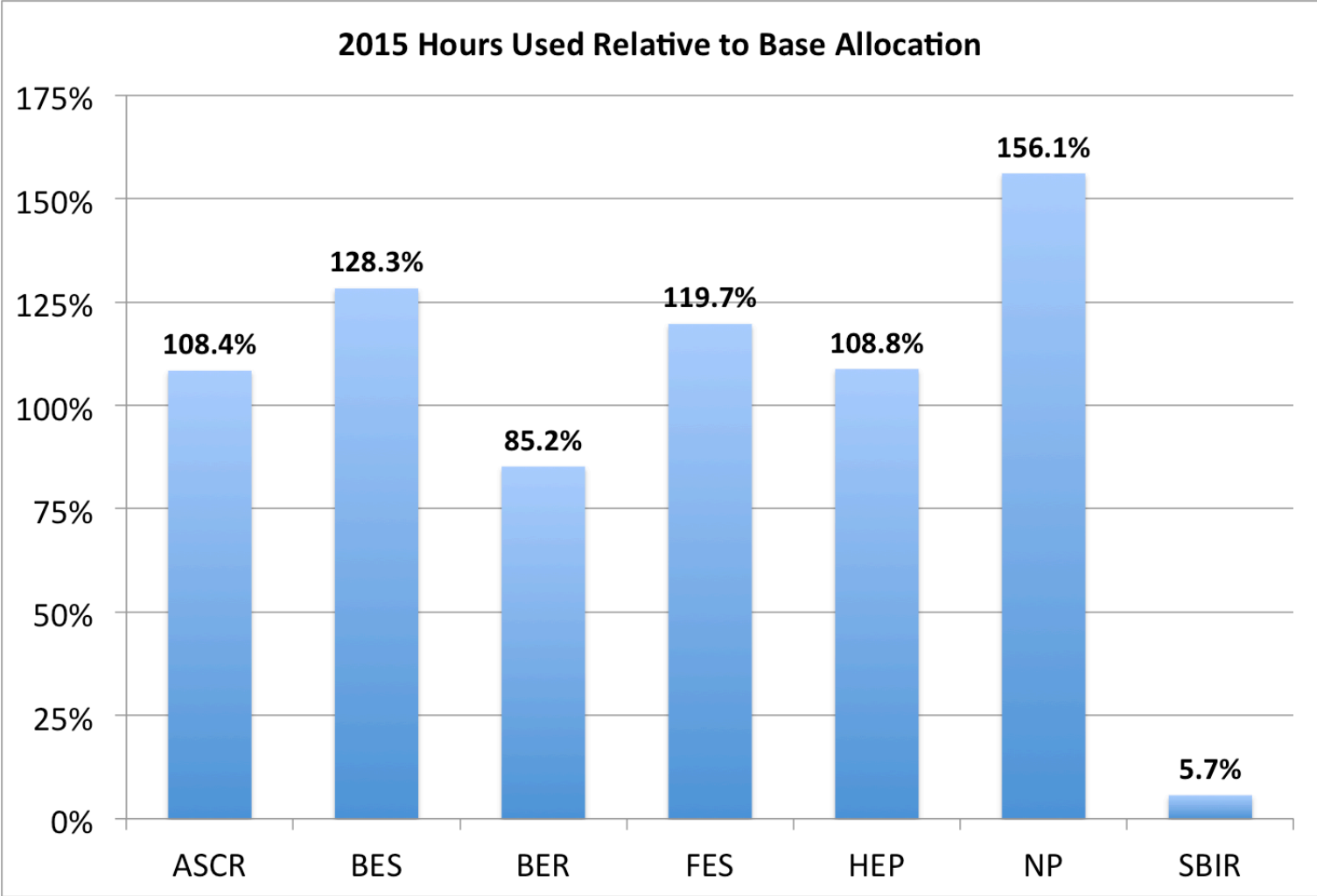
DynaFlow, Inc.



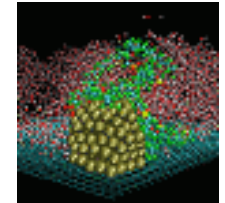
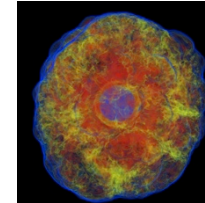
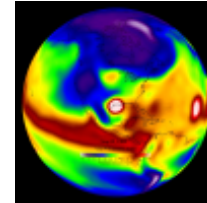
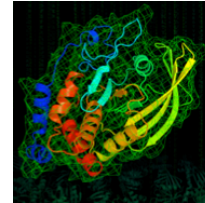
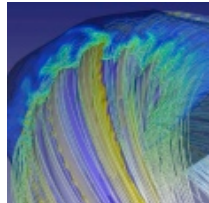
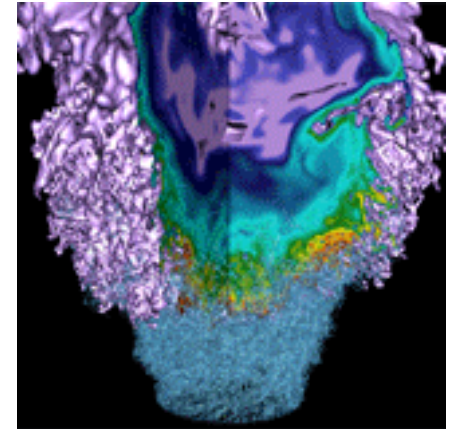
Hours Used Relative to DOE Base Allocation



NERSC
Delivered
3.17B hours
to DOE
users in
2017



Is NERSC optimizing the use of its resources consistent with its mission?



Impact of move on users



- **Network had only 50 minutes of downtime**
 - Upgrade of backplane of core routers to support 400Gpbs link between Wang Hall and Oakland
- **Zero downtime on the GPFS file systems**
 - Only modest performance degradation for a few days while data was replicated over the 400Gbps link
- **Edison was down for 5 weeks**
 - Up to 6 weeks were exempted from the availability calculation per agreement with program manager
- **There was always at least one system in production either in Oakland or Wang Hall during the move**

Key events affecting system availability



- **System changes**
 - Cori Phase I installed (8/26) ; opened to users (11/11)
 - Edison moved from OSF to CRT (11/30 to 1/4)
 - Hopper decommissioned (12/15)
- **Significant unscheduled outages**
 - Two power outages due to PG&E failure in Oakland (8/12 and 11/14) – 34 hours total downtime

Mean Time to Interrupt and Failure



Metric Description	Edison	Hopper	GPFS	HPSS
Mean Time to Interrupt				
2014 MTTI Actual	11:18:54	17:01:17	51:21:39	20:02:59
2015 MTTI Actual	12:10:35	13:4:52	60:14:40	19:46:00
Mean Time to Failure				
2014 MTTF Actual	20:14:58	25:18:01	91:02:54	182:10:03
2015 MTTF Actual	19:10:7	16:10:35	72:19:17	60:12:44

[Numbers are reported in days:hours:minutes]

- **Hopper MTTI and MTTF dropped mostly due to lustre file system issues**
- **The 2 power outages in 2015 did affect the MTTF on most of our systems**

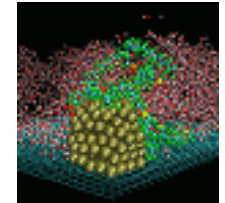
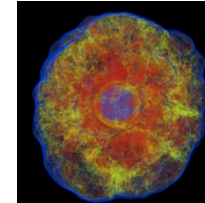
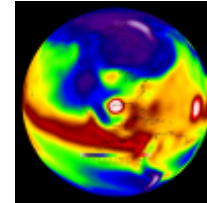
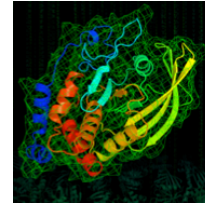
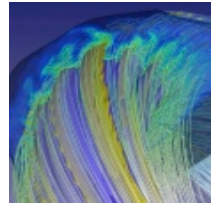
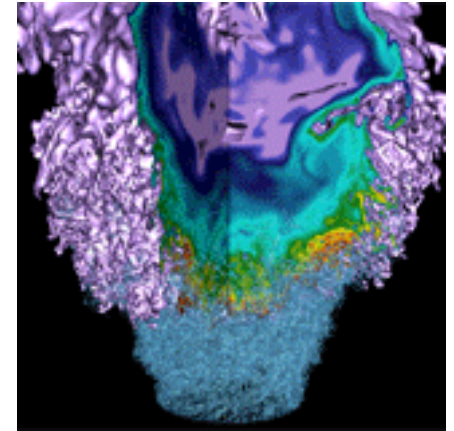
Resource Utilization Metrics



Metric Description	2014	2015
Edison Utilization	91.4%	92.2%
Hopper Utilization	90.1%	89.6%
GPFS (Global file system) Usage	4.5PB	4.9PB
HPSS Usage	50.1PB	70.7PB
Total MPP Hours	3.31B	3.17B

- **Hopper users were migrated to Edison in preparation of the move**
- **The centerwide storage systems are important to the workflows at the facility – HPSS increased in usage 20PB in 2015**

What innovations have been implemented that have improved NERSC operations?

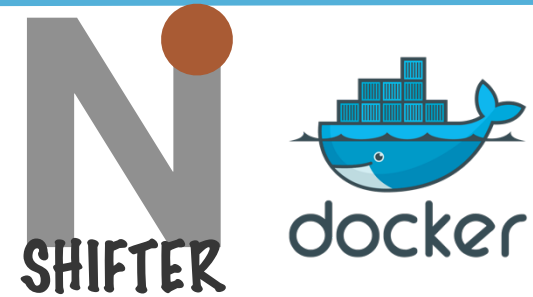


Shifter: Containers for HPC



Challenge and Opportunity

- Data Intensive computing often require large, complex software stacks that are difficult to support in HPC.
- Docker is rapidly becoming a new way to package and run applications.



Innovation

- Shifter is a NERSC R&D effort, in collaboration with Cray, to support User-created Application images.
- Shifter provides “Docker-like” functionality for HPC

Impact and Early Successes

- Shifter has already enabled multiple projects to quickly make use of NERSC (e.g. LCLS, LHC)
- Shifter can improve job-startup times and application performance (e.g. Python)
- Shifter will be supported by Cray and is already being evaluated by other HPC centers



Supporting Real-Time Computing on HPC

NERSC

Challenge and Opportunity

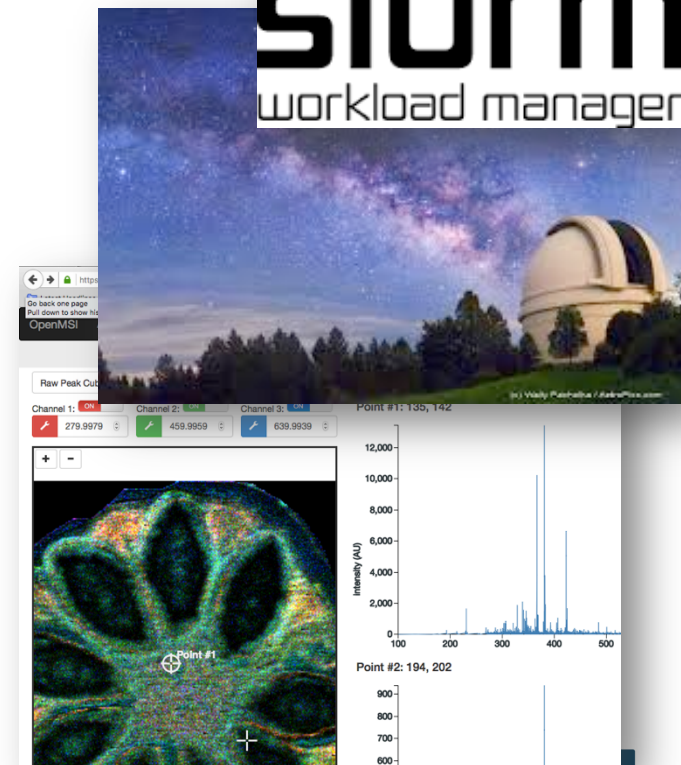
- Many users require 'real-time' access to the system.
- With the adoption of SLURM, NERSC has the capability to offer 'immediate' or 'real-time' access on Cori Phase 1.
- NERSC Developed a process to select and judge projects.
- 15 of 19 submission were approved for access.

Innovation

- Starting with modest real-time resources, 32 nodes
- Jobs that exceed the reserved node allocation are scheduled on a high-priority queue

Impact and Early Successes

- The Palomar Transient Factory uses the RTQ to schedule real-time analysis and classification of telescope images. 95% of new transients are processed in less than 6 minutes.
- OpenMSI and Metabolite Atlas are web-based portals that use real-time queues to analyze raw experimental data to identify new molecules



Data Management: DataMap



Challenge

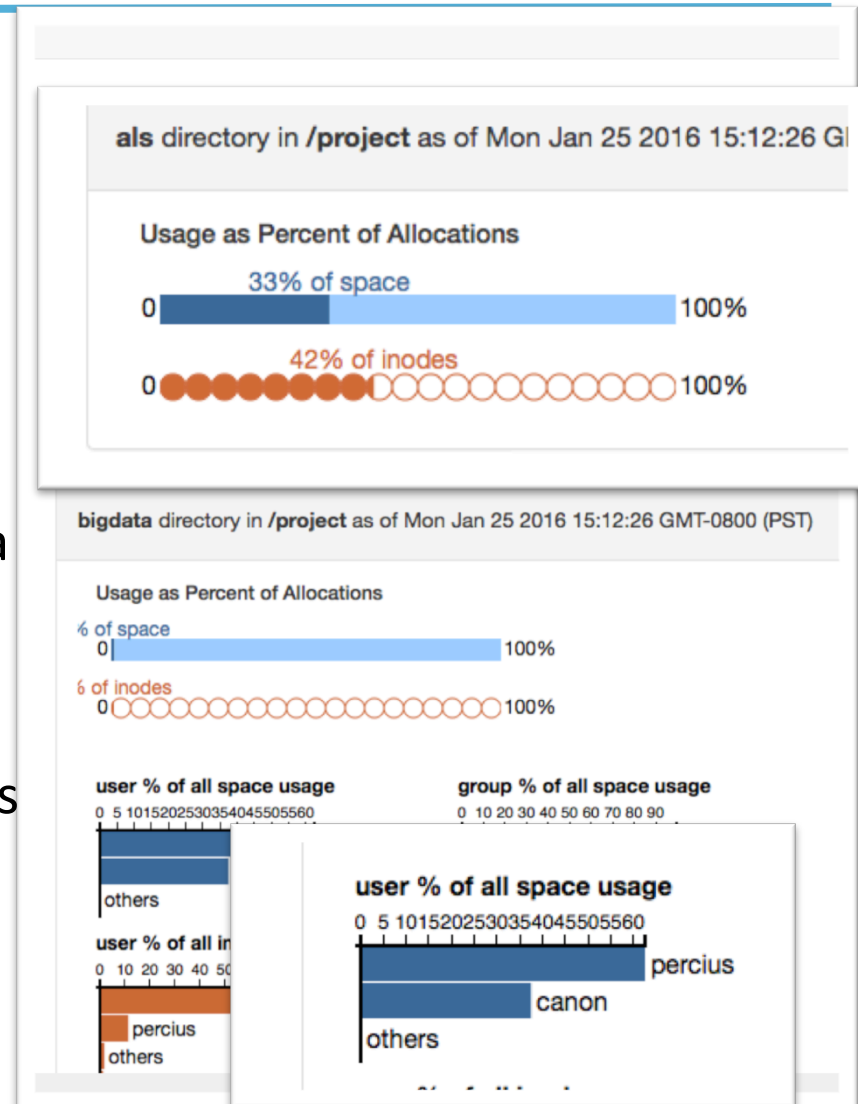
- Users increasingly struggle to organize and manage data.

Innovation

- Deploy or develop tools to help users manage, find, and easily archive data
- First phase: developed Data Map, a web-based tool to visualize data usage

Impact

- Avoids the need for individual users to “crawl” the file system to build indexes.
- Enables PIs and user to quickly see who is consuming space or files



Evaluating and Deploying 400Gb



Challenge and Opportunity

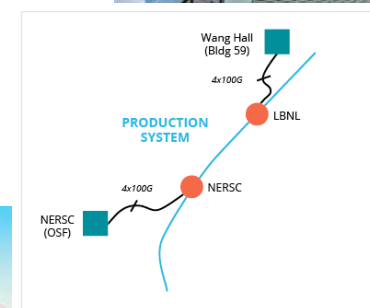
- Move Petabytes of data from the Oakland Scientific Facility to Wang Hall
- Minimize Impact to Users
- Get it done fast!

Implementation

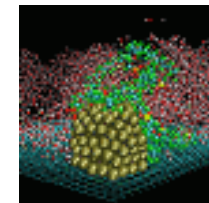
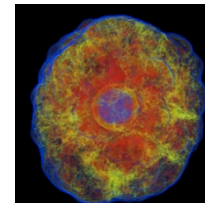
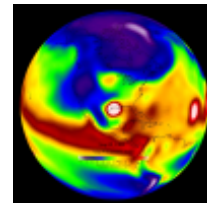
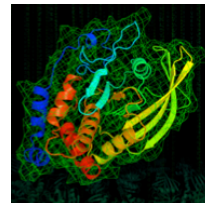
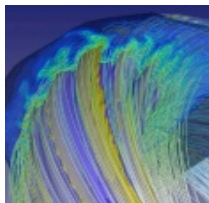
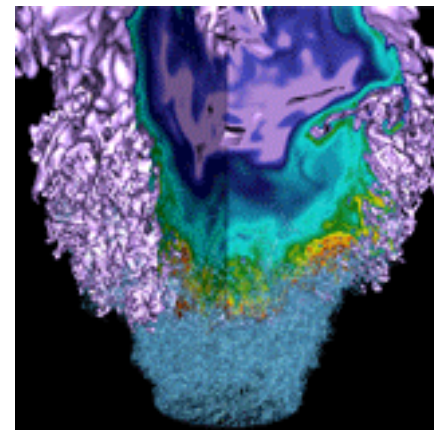
- NERSC and ESnet deployed one of the first 400Gb networks to be put into production by a research and education network
- NERSC architected and deployed 14 advanced Ethernet-to-Infiniband Routers at CRT to route between the WAN link and the internal storage network

Impact and Early Successes

- 400Gb link was used to live-migrate data from OSF to CRT and achieved sustained transfer speeds of 170 Gbps, roughly 1 PB per day (disk-to-disk)
- No Downtime!



Outbrief Report



Overall Assessment

Overall NERSC is an exceedingly well-managed and professionally operated Center.

- Staff are enthusiastic and clearly focused on supporting science.
- Multiple disruptive changes were well-managed minimizing their impact on users including:
 - The transition from OSF to CRT was carefully planned and well executed.
 - A substantial turnover in staff
 - The deployment of Cori phase one and retirement of Hopper.
- As a result user satisfaction with NERSC remained very high.
- Noteworthy innovations:
 - Innovations such as Shifter, python workflow support, SSI and IOR benchmarks and operations automation
 - The switch to Slurm and changes to allocation management including real-time queues and a scavenger queue.
- NERSC has a robust user support effort that is strengthened by the new organization:
 - The new emphasis on data-driven science.
 - Broad user engagement
 - The NESAP program is particularly important in preparing for Cori-2
- NERSC has received a substantial budget increase over the last several years. This has enabled the move from the Oakland facility to the new Wang Hall facility on the campus, and will increase the size of the NERSC-8 system from what was previously possible.

Keep up the good work!

Highlights of some of the OAR review committee's comments

- NERSC is to be congratulated on the manner in which the move to CRT was planned and executed. In particular, for the filesystem relocation only a 50 minute network downtime was incurred, and only a minor performance degradation could be seen while the data was mirrored/migrated.
- We commend NERSC for their excellent and wide ranging user communication and interaction which prepared the users for the CRT move well in advance (weekly mailings, long campaign of information to prepare users for pending changes).
- The panel believes that the Dungeon Session “Application form” is a great way to prepare new application teams for a deep dive engagement with the vendors.
- The “Edison queue wait time” survey result was the only one below the target 5.25 score, however even this area has improved since last year. NERSC continues to monitor and improve this situation through appropriate queue incentivization.
- It was clear from the presentations and the OA report that Scientific Discovery is the driving mission for NERSC.

Highlights of some of the OAR review committee's comments

- The NESAP post-doc program not only assists application readiness efforts, and serves as a pool of talent for NERSC staff, it is having an immediate positive effect on Science output.
- The level of engagement with others in many of the innovative efforts is impressive and significantly helps drive adoption and long-term community support.
- The benchmark development effort has good involvement within APEX, but can be expanded with broader participation from other sites. NERSC is well-positioned to lead a workshop on benchmark development with an emphasis on data and workflow benchmarks.
- NERSC should collect detailed usage statistics on Cori Phase 2 usage and offer incentives to get users from Xeon to KNL.
- NERSC should collect detailed data on their burst buffer resource usage.
- NERSC may want to work with other DOE labs on workflow benchmarks.