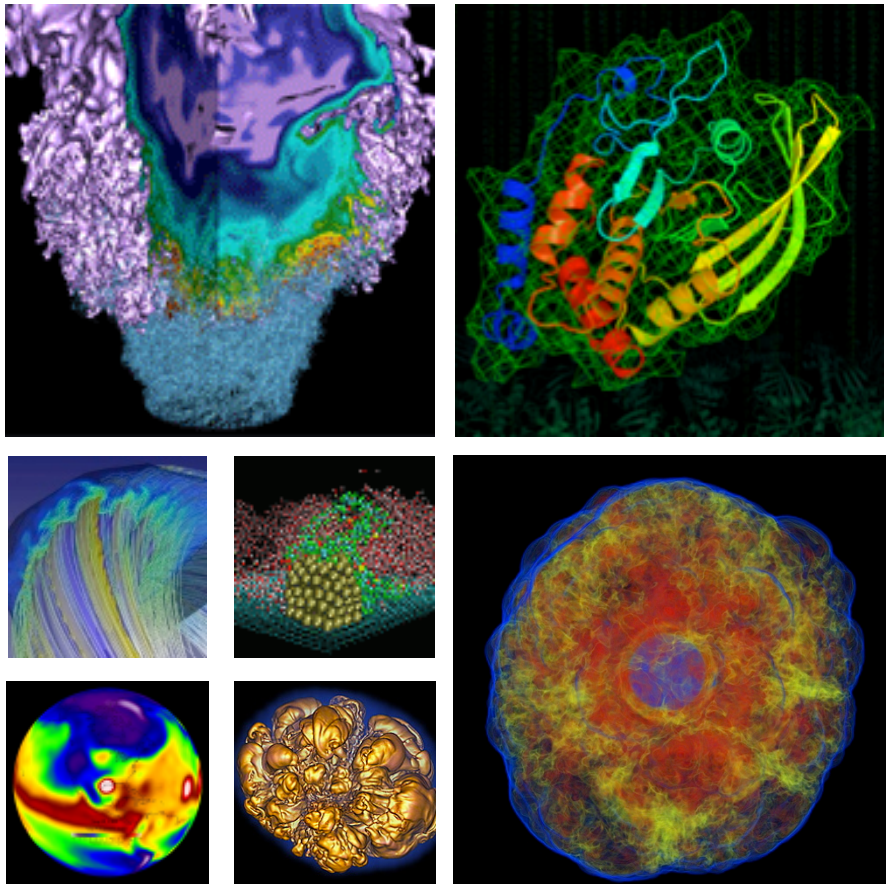
NERSC Users Group Webinar - May 2015



Speaker Name(s)

Title

May 14, 2015

Agenda

- NERSC Move Timeline Update
- Carver Retirement Reminder
- Hopper and Edison Status Updates
- HPSS Archival Storage System Enhancements
- Plans for Global Scratch
- NESAP Update
- New Batch Scheduler at NERSC: SLURM
- NIM (accounting interface) Enhancements
- Annual User Group Meeting: Your Opinion Wanted
- What is "Shifter"? (Think user-defined images; webinar Friday.)
- NERSC is Hiring!

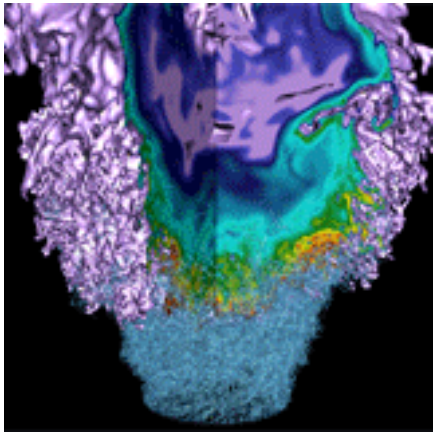
NERSC Move to CRT

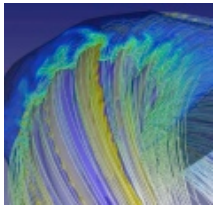
- **NERSC will be moving systems and staff from downtown Oakland (OSF) to Berkley Lab main campus this year (CRT building)**
- **Timeline Highlights**
  - Cori Phase 1 (Cray XC40/Haswell) user availability October 2015
  - Carver retires 9/30/2015
  - Global scratch retires 9/30/2015 (+14 days to retrieve files)
  - Hopper retirement after Cori Phase 1 stable and available to all users; expected December 2015
  - Edison moves from OSF to CRT beginning sometime in Nov.-Dec., unavailable for ~6 weeks

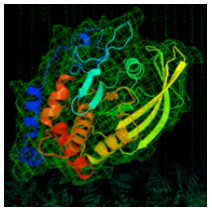
Carver Retirement Reminder

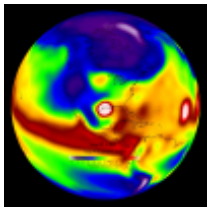
- **Carver will be shut down on Sep. 30, 2015**
- **Running jobs will be terminated beginning at noon on Sep. 30**
- **Software stack frozen on July 1**
- **14 days to retrieve files on scratch**
  - $GSCRATCH on Edison
- **Please move your work to Edison**
- **Contact NERSC consultants if you need help or advice**
  - consult@nersc.gov
  - https://help.nersc.gov
  - https://my.nersc.gov
  - https://www.nersc.gov/users/computational-systems/carver/retirement-plans/

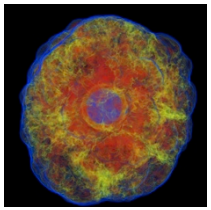
Edison Updates

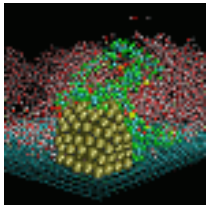












Zhengji Zhao
NERSC User Services

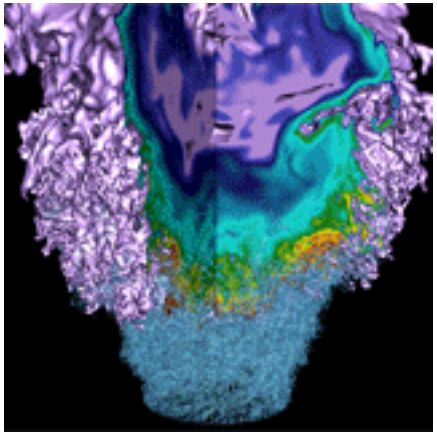
Edison Updates

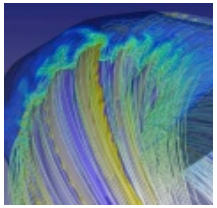
- **Queue change to decrease wait for debug, medium and large jobs (end of March, 2015):**
  - Debug, reg_xbig, reg_big, and reg_med queues priority increase
  - reg_small jobs can move to Hopper, which has shorter wait times
- **Edison compute nodes can now access remote networks (RSIP configuration change, end of March 2015)**
  - Provides remote (outside of NERSC) database access from compute nodes.
  - Web portals, ipython notebook server, etc., can run on Edison compute nodes (via remote secure tunneling).
  - Working to increase the number of ports per RSIP node
- **TMPDIR=/tmp on Edison login nodes (May 8, 2015)**
  - Compilations should be faster than before

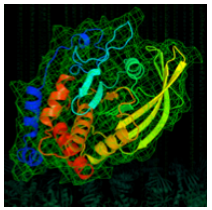
Edison Updates - continued

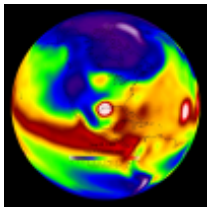
- **Increasing Minimal Bias (IMB) setting for Aries is in place now to address the performance variation on Edison (experimental, May 8, 2015)**
  - env MPICH_GNI_ROUTING_MODE=ADAPTIVE_1. Slight non-minimal bias at injection point, but increasing bias toward the minimal path as the packet traverses the network. It does help with intermediate group interference.
- **VTUNE is now available on Edison (Feb, 2015)**
  - VTUNE provides a rich set of performance insight into hotspots, threading, locks & waits, bandwidth and more. Use powerful analysis to sort, filter and visualize results on the timeline and on your source. VTUNE is a preferred performance analysis tool for on-node code optimization for Cori and Edison.
  - To use, http://www.nersc.gov/users/software/debugging-and-profiling/vtune/

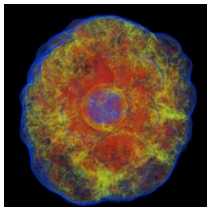
Hopper Updates

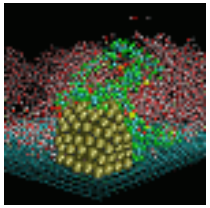












Helen He
NERSC User Services

Recent Changes on Hopper

- **Scratch file systems updated to Lustre 2.4.1**
  - Required so we could keep OS up to date
  - Feb 3-4: /scratch2
  - Feb 18: /scratch
- **Mar 11: OS upgraded to CLE52UP02**
  - New default Cray programming environment software
  - NERSC Software and user applications rebuilt
- **May 12: Batch queue changes for reg_long queue**
  - Max global run limit increased from 50 to 100
  - Max user run limit increased from 16 to 32

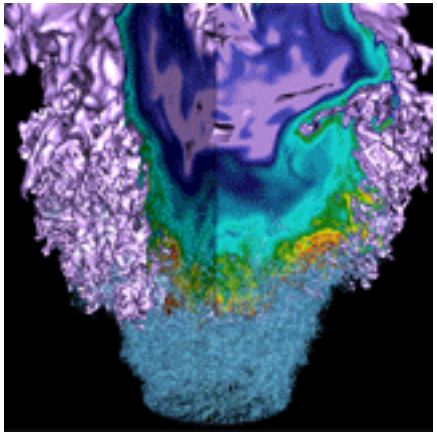
Current Issues on Hopper

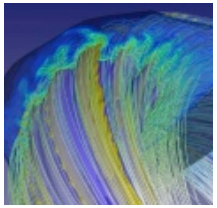
- **A handful of users reported jobs got run time error of compute nodes OOM (out-of-memory) after the OS upgrade.**
- **Narrowed down to a bug in the Lustre client triggered by heavy IO.**
- **Available workarounds are:**
  - Use a non-Lustre file system, such as /project or /global/scratch2
  - Use fewer cores per node
  - Use larger memory nodes
  - Use Edison

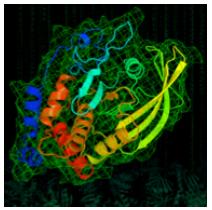
Future Plans for Hopper

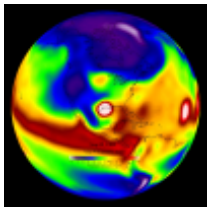
- **Keep system stable to maximize your productivity**
  - No more major system upgrades
  - Plan to change RSIP configuration to allow outgoing network connection on the compute nodes as recently done on Edison
- **Hopper retirement expected December 2015**
  - Exact date TBA
  - After Cori Phase 1 is stable and available to all users

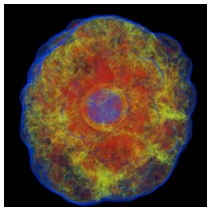
HPSS Enhancements and Global Scratch Plans

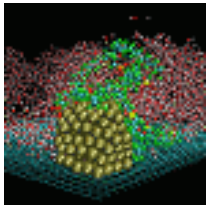












Lisa Gerhardt, Data & Analytics Services

Archive Disk Cache Increased

- **In looking at how our archive (HPSS) is being used, we observe:**
  - High read-rate of files, about 40% of files read occur within 30 days of being archived
  - Growing at a total of about 1.5 PB per month
- **Previously, the disk cache was optimized exclusively for writes**
  - 5 peak days of data ingest
  - Total aggregate bandwidth was 12GB/sec
- **With the increase, we have sized the disk cache for both reads and writes**
  - Retains data for about 30 days
  - Total aggregate bandwidth is 40GB/sec

/global/scratch Retirement and Status

- **/global/scratch will retire with Carver system on 9/30/2015**
  - Will remain read-only on other systems through 10/14/2015
  - Archive needed files to HPSS or move to Edison or Hopper scratch
- **Purge policy reduced from 12 to 8 weeks (5/11/2015)**
  - This will reduce usage and allow some hardware to aid in the relocation of the /project file system to CRT
  - Remainder of /global/scratch capacity will be added to project file system after relocation to CRT
- **NERSC plans to provide a global scratch file system when Cori and Edison are co-located in CRT**

The compute and storage systems 2015

Hopper: 1.3PF, 212 TB RAM

Cray XE6, 150K Cores

Edison: 2.5PF, 357 TB RAM

Cray XC30, 130K Cores

Sponsored Compute Systems
Carver, PDSF, JGI, KBASE, HEP
8 x FDR IB

Vis & Analytics, Data Transfer Nodes, Adv. Arch., Science Gateways

2.2 PB Local Scratch 70 GB/s

16 x QDR IB

6.4 PB Local Scratch 140 GB/s

16 x FDR IB

42 GB/s

50 GB/s

5 GB/s

40 GB/s

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

/global/scratch — ***3 PB***
3 x SFA12KE

/project — **5 PB**
DDN9900 & NexSAN

/home — ***250 TB***
NetApp E5460

HPSS — ***70 PB stored, 240 PB capacity, 40 years of community data***

2 x 10 Gb

1 x 100 Gb

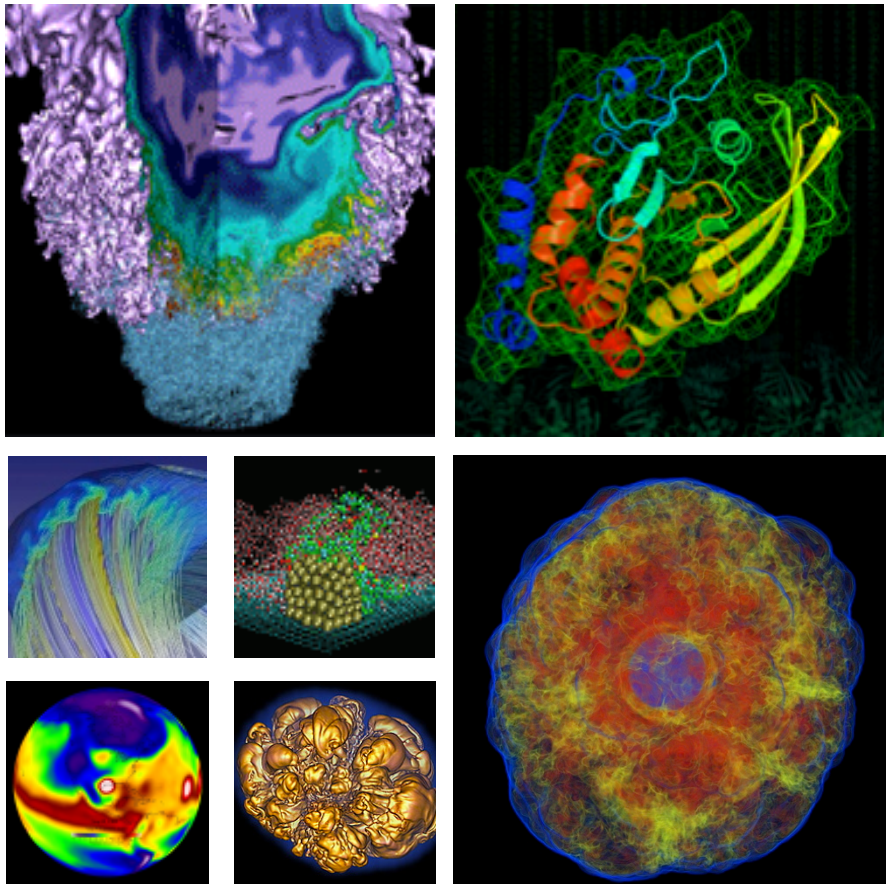
Science Data Network

ESnet Energy Sciences Network

NESAP Update

- NERSC Exascale Science Application Program
- NERSC continues to actively engage with code teams and vendors (Intel, Cray) to prepare codes for Cori
- Trainings, teleconferences, visits, dungeon sessions
- http://www.nersc.gov/users/computational-systems/cori/nesap/
- Three postdocs have been selected (of ultimately 8)
  - Optimization of the BoxLib Adaptive Mesh Refinement Framework for Scientific Application Codes, PI: Ann Almgren (Lawrence Berkeley National Laboratory)
  - High-Resolution CFD and Transport in Complex Geometries Using Chombo-Crunch, David Trebotich (Lawrence Berkeley National Laboratory)
  - Materials Science using Quantum Espresso, Paul Kent (Oak Ridge National Laboratory)
  - http://cs.lbl.gov/careers/careers-and-fellowships/

SLURM @NERSC



Computational Systems Group

NUG Meeting
May 14, 2015

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

SLURM Deployment at NERSC

- **The Cori Phase 1 system will be using SLURM as the Workload Manager (WLM)**
- **SLURM will provide both Resource Manager (RM) functionality and Scheduler functionality**
- **WLMs have traditionally interfaced with ALPS on the Crays**
  - ALPS is a lower-level application placement scheduler
- **SLURM can be run in "native" mode – i.e. without the use of ALPS**
  - More details later

SLURM at NERSC

- **SLURM provides all of the same functionality as Torque/Moab (a few differences).**
- **SLURM is fully open source**
- **In the Cori time frame, SLURM will be able to better support our mixed Data-HPC needs.**
- **SLURM is extensible (plugin architecture).**
- **SLURM provides a PBS translator**
  - Allows scripts written for Torque to be submitted to SLURM

SLURM vs Torque/Moab

Torque/Moab (#PBS)

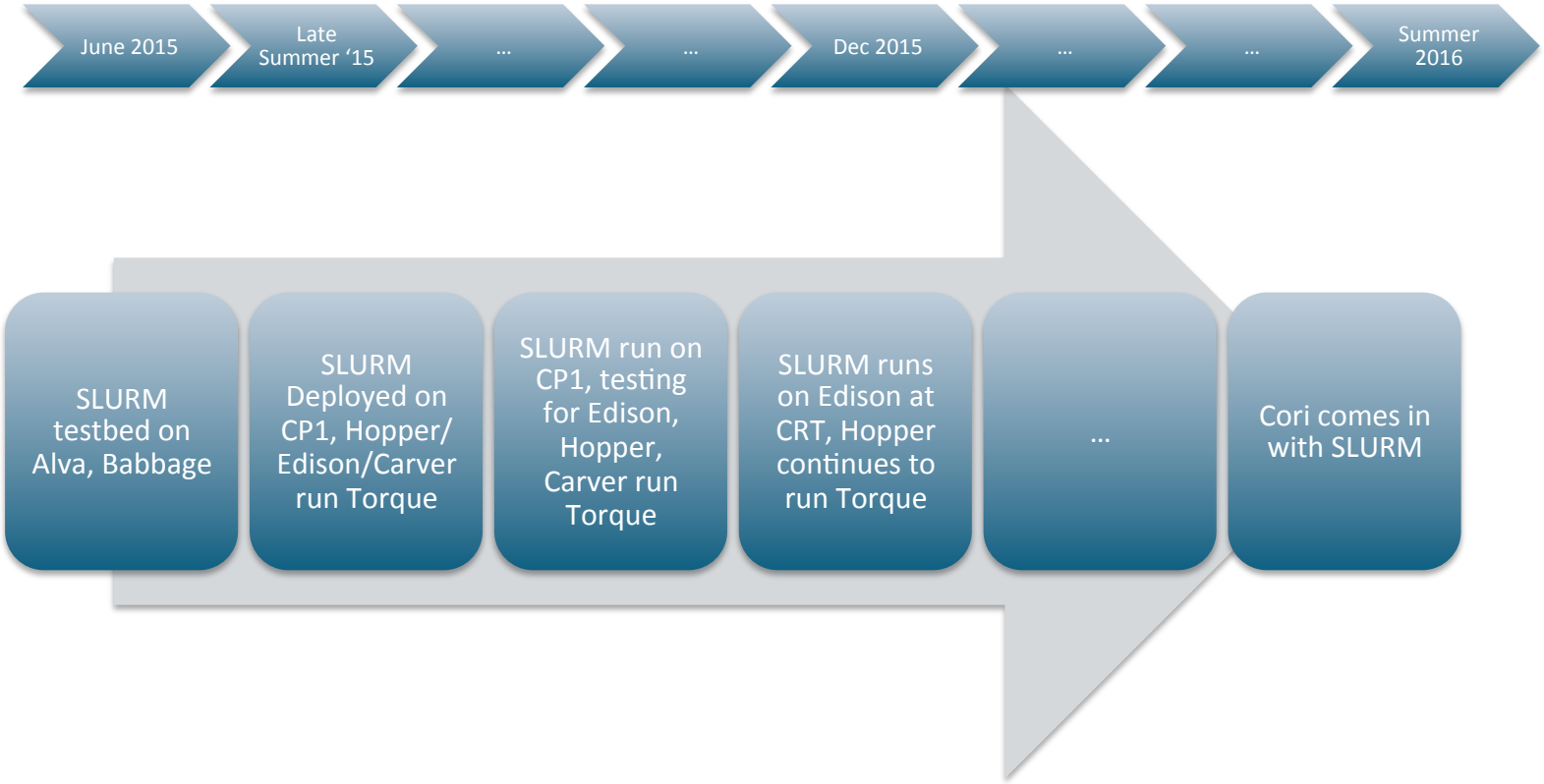
- qsub/qdel/qstat
- qstat –a
- -l nodes/mppwidth
- -l walltime
- -t [array]
- -Wdepend=

SLURM (#SBATCH)

- sbatch/scancel/squeue
- sinfo
- -N (nodes) / -n (PEs)
- -t (min)
- --array=[array]
- --depend=

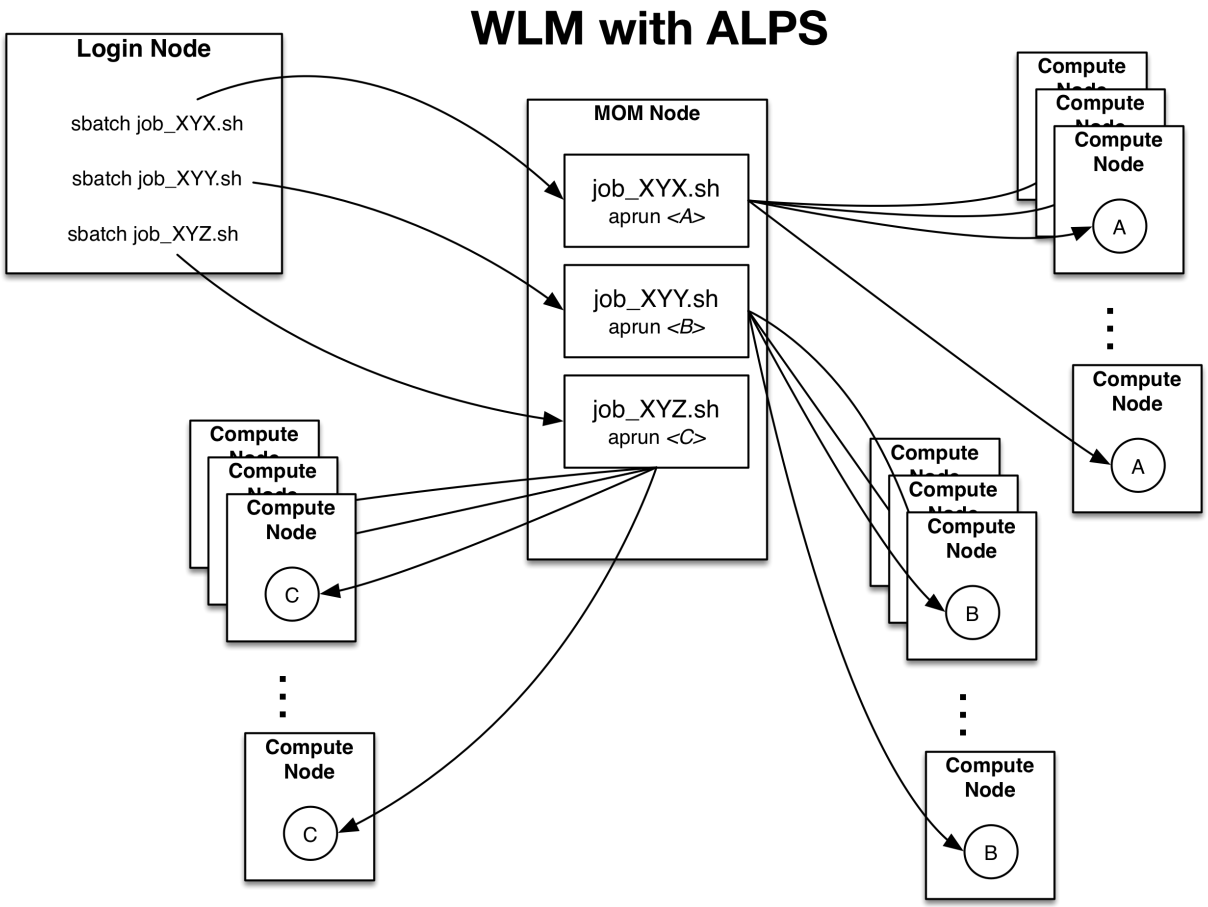
Not exhaustive, See: http://slurm.schedmd.com/rosetta.pdf for details

High Level Plan – 2015-2016



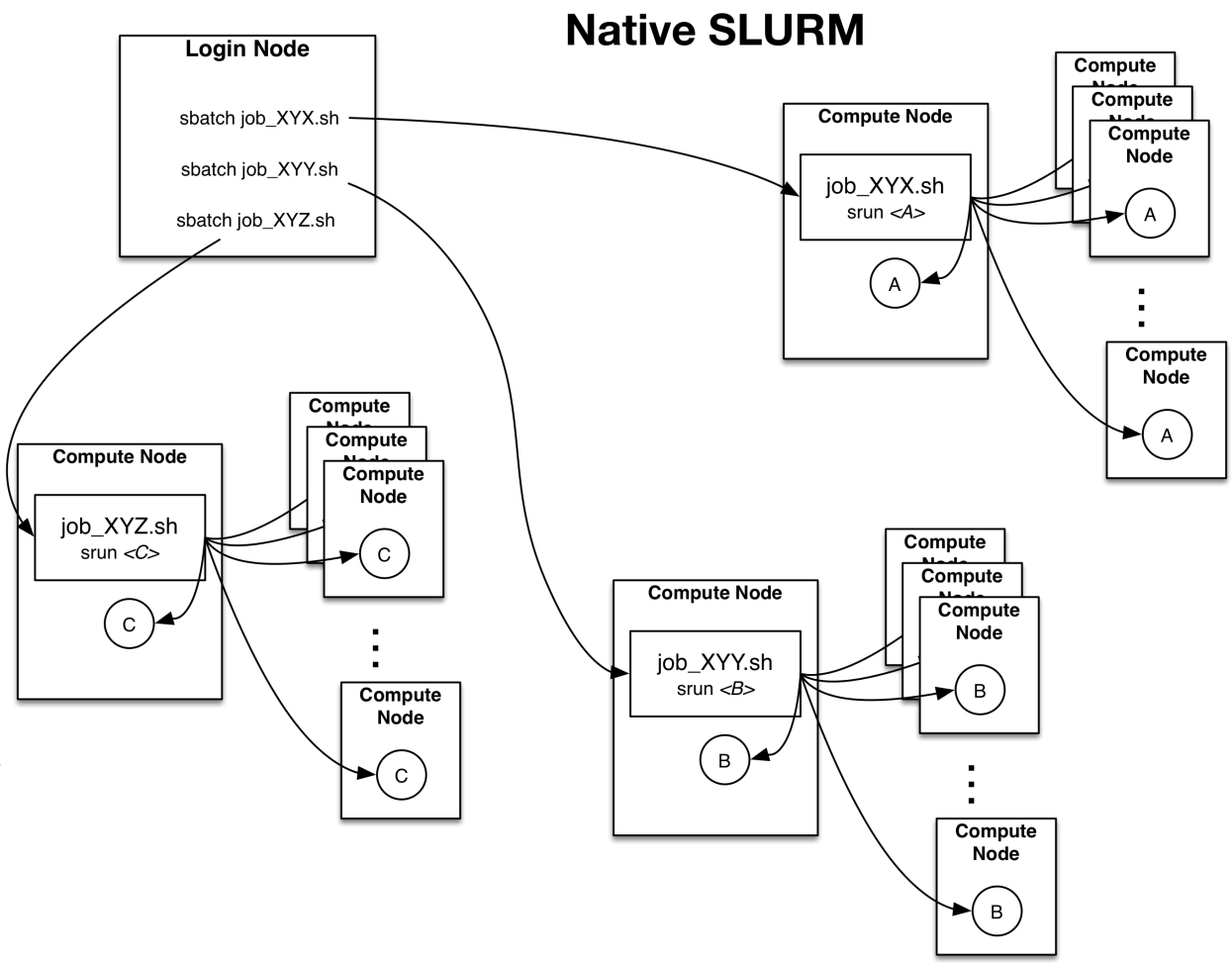
WLM/ALPS Interactions on the Cray

- **Batch Script runs on a shared service node**
  - Level of indirection to access compute nodes
  - "CCM" required to run applications unable to utilize hierarchical compute model
  - Resources directly managed by ALPS, indirect communication with WLM
- **ALPS manages compute node resources, application placement, Aries High-speed Network (HSN) access**



Native SLURM on the Cray

- **Batch Script runs directly on a compute node**
  - Simpler access to compute resources (especially for x86_64 and KNL environments)
  - No bottleneck in shared "MOM" node
  - Implicit "CCM" functionality for many applications; NERSC adding ssh-based access for others
  - Resources directly managed by SLURM on each compute node (memory, processes)
- **Interaction with Native Cray networking libraries for full access to Aries HSN**
- **Reduced complexity by *not* interacting with separate resource manager (ALPS)**



Testing Plans

- **Continual use of Alva, Babbage**
  - test development
  - test new SLURM releases
- **Full Scale tests of SLURM on Hopper/Edison**
  - production queue structure
  - slurm Unit Test suite developed at NERSC
    - reservations * serial jobs * preemption * job dependencies * job arrays * routing queues * MPI * PBS emulation
  - simulated workload reflecting the spectrum of job sizes on Edison and create a backlog of several thousand jobs
- **Allows us to tune scheduler settings to maximize efficiency and fairness**

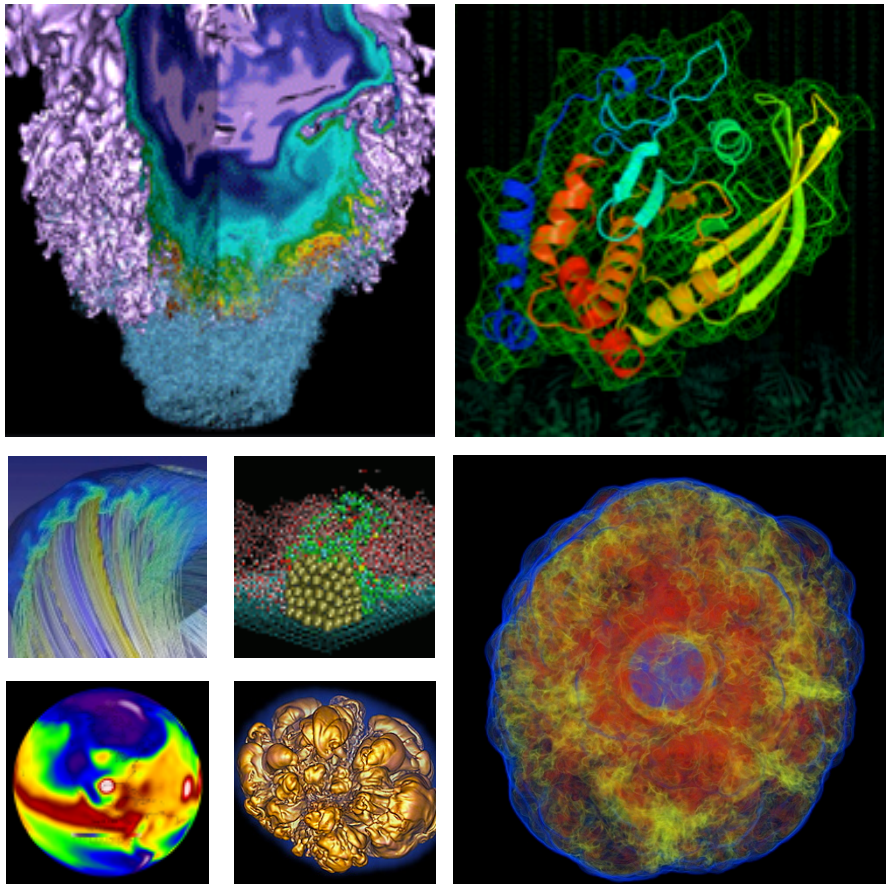
User Resources for SLURM

- **User Services Group will have**
  - Documentation
  - SLURM usage examples
  - Sample Job Script Files
- **NERSC Documentation**
  - https://www.nersc.gov/users/computational-systems/edison/alva-test-and-development-machine-for-edison/ (Restricted to authenticated users)
  - Using SLURM on Babbage and Cori Phase 1 pages are forthcoming
- **References**
  - http://slurm.schedmd.com/documentation.html
  - http://slurm.schedmd.com/pdfs/summary.pdf

What do We Need From You?

- **Check to see how you use the WLM**
  - Simple submission/query
  - Complicated workflows?
  - Workflow tools?
    - Fireworks
    - Qdo
    - Others?
- **Simple use cases will translate to SLURM easily**
- **If you have complicated workflows, please contact us (consult@nersc.gov) to test out SLURM**

Allocating Fixed Hours to Users Instead of Repo Percentage in NIM



Clayton Bagwell

NERSC Account & Allocations Support
NIM Development Team

Oakland Scientific Facility
May 14, 2015

U.S. DEPARTMENT OF ENERGY | Office of Science

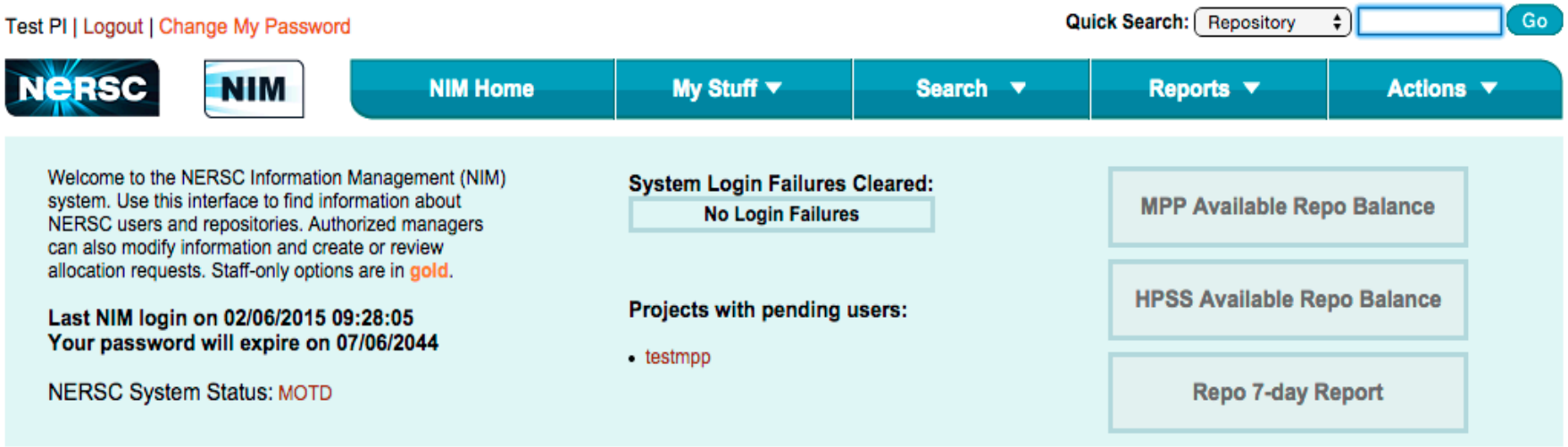
BERKELEY LAB

Background

- **The original model for allocating time to users within a repository has been to allow them access to a percentage of the repo's total allocation.**
  - This meant that the number of hours a user could access would fluctuate with the changes to a repo's allocation
- **Many of our PIs have requested the ability to allocate a fixed number of hours to their users.**
- **On April 15th, the NIM Development Team released modifications to NIM to accommodate Fixed Hour Allocations.**

User Requests an Account

- New user submits an account request
- PIs and Proxies get email notification and reminder upon logging into NIM



Pending Users tab info

- **Clicking on the reponame takes you to the Pending Users tab for the project**

Project Information | User Roles & Contact Info | User Status by Repo | MPP Usage & Quotas | HPSS Usage & Quotas | Transfer History | Pending Users | Project Access

Project	Project Repos	Project Unix Groups	PI	Last Active
testmpp#15	testmpp testpdsf	c_claveyrv c_pishouldownthis c_testtpidir c_troutcrk includef pishould richards_dir testint testint2 testmpp testpdsf testtpid	Test Pi	2015
Test project/repositories				

Pending Account Requests

Name	Uname	Organization Label	Email	Workphone	Remarks	Submit Date	Reponame	Resource Type	% Allowed	Hours Allowed		
User, Test Fixed Hour	tfhu	NERSC	nobody@nersc.gov	510-486-8612	Testing the Fixed Hour Allocation process.	13-MAY-2015	testmpp	REPO	100		Update Allocation	Approve Deny
							testmpp	STR	100		Update Allocation	

2 records found

Updating Allocation before Approval

- **You can either adjust the % Allowed or enter a value for Hours Allowed**

Return to List

Full Name	User, Test Fixed Hour
UNIX User Name	tfhu
Project Name	testmpp
Repository Name	testmpp
Resource Type	REPO
Total Hours Allocated	1,000
Total Usage Charged	0
Remaining Allocation	1,000
% Allowed	100
Hours Allowed	500

Submit

update successful.
Return To Form
Return to List

Approve Account Request

- **Clicking Return to List takes you back to the Pending Account Requests where you can now approve the account**

Project Information | User Roles & Contact Info | User Status by Repo | MPP Usage & Quotas | HPSS Usage & Quotas | Transfer History | Pending Users | Project Access

Project	Project Repos	Project Unix Groups	PI	Last Active
testmpp#15	testmpp testpdsf	c_claveyrv c_pishouldownthis c_testtpidir c_troutcrk includef pishould richards_dir testint testint2 testmpp testpdsf testtpid	Test Pi	2015
Test project/repositories				

Pending Account Requests

Name	Uname	Organization Label	Email	Workphone	Remarks	Submit Date	Reponame	Resource Type	% Allowed	Hours Allowed		
User, Test Fixed Hour	tfhu	NERSC	nobody@nersc.gov	510-486-8612	Testing the Fixed Hour Allocation process.	13-MAY-2015	testmpp	REPO		500	Update Allocation	Approve Deny
							testmpp	STR	100		Update Allocation	

2 records found

Manually Adding a User to Your Repo

- **When you use Add/Revive User to add a user to your repo, you can allocate either by % Allowed or Hours Allowed**

Add a new NERSC User

For NERSC Principal Investigators, Account Managers, and NERSC staff.

Please fill out a separate request for each new NERSC user that you want added to your repositories (repos), and then click the "Submit" button at the bottom of the form.

Your request will be reviewed by NERSC Account Support. After it has been received and processed, the user will need to submit a Computer Use Policy form and then the Ac
an email regarding their account password information.

User First Name:
Middle Initial:
User Last Name:
NERSC Username: *(If user does not have a NERSC username, enter a preferred username.)*
Citizenship: Not Specified
Email Address:
Telephone:
Organization: UNKNOWN: Not Yet Set
Mail Stop (optional):

Repository Information

Choose the platforms on which you would like an account created, select the repository name
and the percentage or amount of the total allocation you would like the user to be able to use (where applicable).

Add ?	Host	Repository name	% Allowed	Hours allowed
	MPP: carver,edison,hopper,matcomp	testmpp		
	HPSS	testmpp		
	pdsf	testpdsf		

Allocating to Existing Users

- **You can set or adjust Fixed Hours for existing users through the MPP Usage & Quotas tab**

Project Information | User Roles & Contact Info | User Status by Repo | MPP Usage & Quotas | HPSS Usage & Quotas | Transfer History | Pending Users | Project Access

Project	Project Repos	Project Unix Groups	PI	Last Active
testmpp#15	testmpp testpdsf	c_claveyrv c_pishouldownthis c_testtpidir c_troutcrk includef pishould richards_dir testint testint2 testmpp testpdsf testtpid	Test Pi	2015
Test project/repositories				

Format: Read-only <--> Edit user allocations

NOTE: all hours displayed below are user hours, not repo hours.

testmpp MPP Users, AY 2015 <--> Show users for prior AY

Login	Name	User Hrs Used	User Charged	Avg CF	% Used	% Allowed	Hours Allowed	User Balance	Repo User Status	Base Repo?	Dflt Now?
bagwell	Bagwell, Clayton	0	0		0		100	100	Active	N	N
tbutler	Butler, Tina	0	0		0	10		100	Active	N	N
shreyas	Cholia, Shreyas	0	0		0	10		100	Active	N	N
tinad	Declerck, Tina	0	0		0	10		100	Active	N	N
wayneh	Hurlbert, Wayne	0	0		0	10		100	Active	N	N
clant	Lant, Craig	0	0		0	10		100	Active	N	N
toffmgr	OFFMGR, Test	0	0		0	0		0	Restricted - negative	Y	Y
rkowen	Owen, R.K.	0	0		0	10		100	Active	N	N
tpi	PI, Test	0	0		0	50		500	Active	Y	Y
tpiproxy	PIPROXY, Test	0	0		0		1,000	1,000	Active	Y	Y
sakrejda	Sakrejda, Iwona	0	0		0	10		100	Active	N	N
dskinner	Skinner, David	0	0		0	10		100	Active	N	N
tuser	USER, Test	0	0		0	100		1,000	Active	Y	Y
projuser	User, Project	0	0		0	10		100	Active	Y	Y
whitney	Whitney, Cary	0	0		0	10		100	Active	N	N
Total:		0	0				1,100				

15 records found

Adjusting User Allocations

- **Click on the Edit user allocations link, enter new % Allowed or Hours Allocated values, click on Save All Rows**

Format: Read-only <--> **Edit user allocations**

NOTE: all hours displayed below are user hours, not repo hours.

testmpp MPP Users, AY 2015 <--> Show users for prior AY

Login	Name	User Hrs Used	User Charged	Avg CF	% Used	% Allowed	Hours Allowed	User Balance	Repo User Status	Base Repo?	Dflt Now?
bagwell	Bagwell, Clayton	0	0		0		100	100	Active	N	N
tbutler	Butler, Tina	0	0		0	10	25	100	Active	N	N
shreyas	Cholia, Shreyas	0	0		0	10		100	Active	N	N
tinad	Declerck, Tina	0	0		0	10	25	100	Active	N	N
wayneh	Hurlbert, Wayne	0	0		0	10	25	100	Active	N	N
clant	Lant, Craig	0	0		0	10	25	100	Active	N	N
toffmgr	OFFMGR, Test	0	0		0	0		0	Restricted - negative	Y	Y
rkowen	Owen, R.K.	0	0		0	10		100	Active	N	N
tpi	PI, Test	0	0		0	50		500	Active	Y	Y
tpiproxy	PIPROXY, Test	0	0		0		1,000	1,000	Active	Y	Y
sakrejda	Sakrejda, Iwona	0	0		0	10		100	Active	N	N
dskinner	Skinner, David	0	0		0	10		100	Active	N	N
tuser	USER, Test	0	0		0	100		1,000	Active	Y	Y
projuser	User, Project	0	0		0	10		100	Active	Y	Y
whitney	Whitney, Cary	0	0		0	10		100	Active	N	N

15 records found

Save All Rows

Instructions in the NIM User's Guide for PIs

- **You can find detailed instructions in the NIM Guide for PIs and Project Managers**

- **http://www.nersc.gov/users/accounts/nim/nim-guide-for-pis/**

Annual Users Group Meeting

- **Floating a trial balloon:**
- **What do you think about an annual user group meeting for the three ASCR facilites (NERSC, ALCF, OLCF)**
  - Hosts move on a rotating basis
  - Would you be more/less likely to attend in person?
  - Advantages?
  - Disadvantages?
  - Let us know what you think.

What is Shifter?

User Defined Images/Containers in HPC

- **Data Intensive computing often require complex software stacks**
- **Efficiently supporting these in HPC environments offers many challenges**
- **shifter – Prototype containers in HPC**
  - NERSC R&D effort, in collaboration with Cray, to support User-defined, user-provided Application images
  - "Docker-like" functionality on the Cray
  - Efficient job-start & Native application performance

Want to Know More?

Please come see our upcoming talk:

Contain This, Unleashing Docker for HPC

Douglas Jacobsen and Shane Canon

Friday, May 15, 2015, 12:00PM – 1:00PM
NERSC, Oakland Scientific Facility, Room 238

Remote Access Info:

https://nersc-training.webex.com

Password: shifter

NERSC is Hiring!

- http://cs.lbl.gov/careers/careers-and-fellowships/
- Application readiness/HPC consultants
- High Energy Physics/Nuclear Physics consultant
- NESAP Postdocs
- Network Engineer
- Data Analytics
- Security Analyst
- Computer Systems Engineer

- NERSC Users Make Great NERSC Staff!

National Energy Research Scientific Computing Center

Section Title