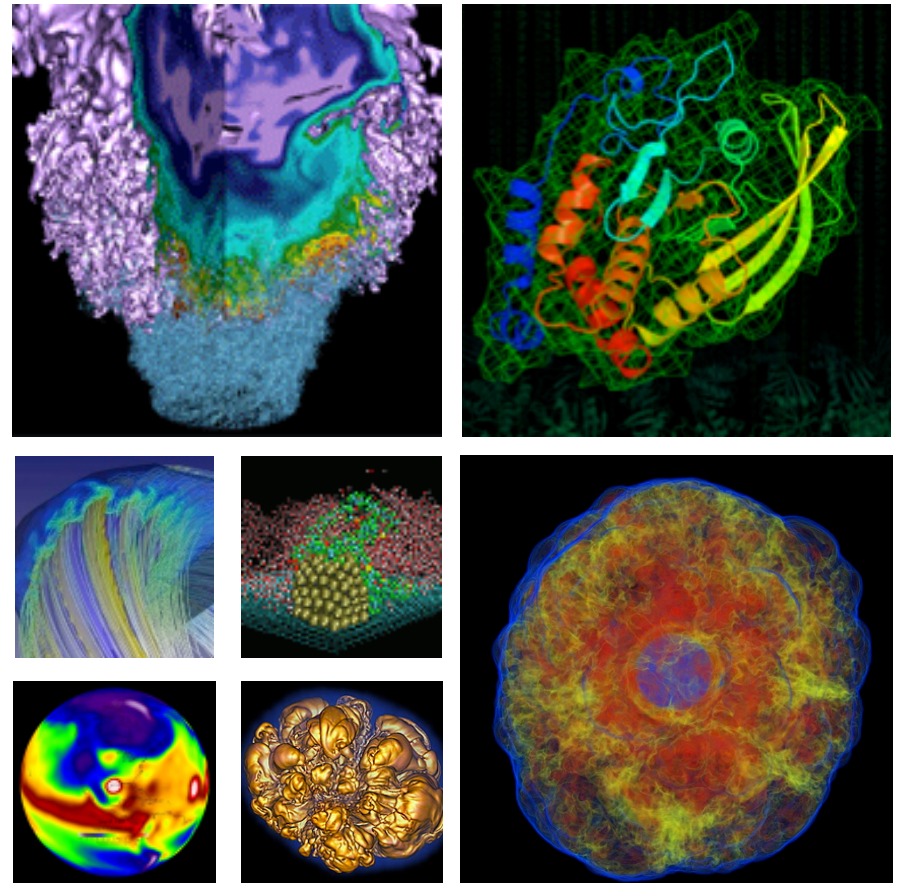


NUG Webinar January 2014



Richard Gerber
NERSC Senior Science Advisor
User Services Group Lead

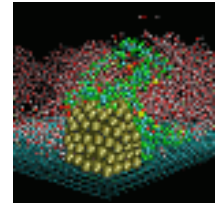
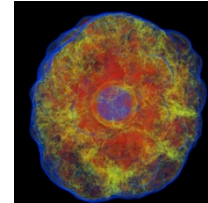
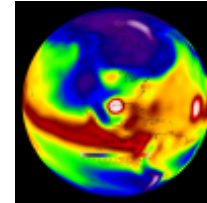
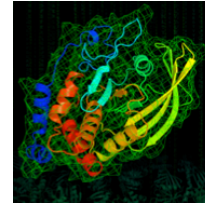
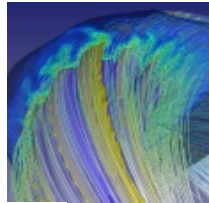
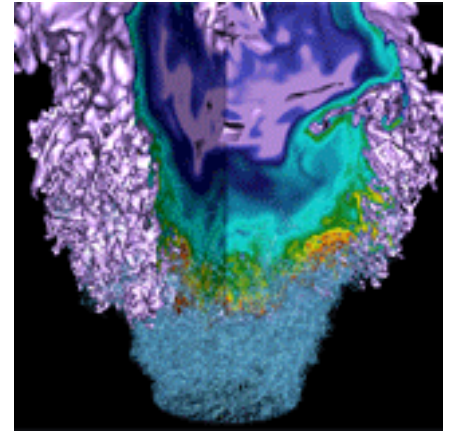
January 9, 2014

Agenda



- **Edison Configuration for AY 2014**
- **Allocation Year Rollover Issues**
- **Project Directories per Project**
- **NUG 2014 User Group Meeting**
- **Automatic I/O Profiling for Your Code**
- **Give / Take Utility for Sharing Files**
- **NUG Queue Committee Update**
- **User Survey Needs Your Input**
- **Open Discussion**

Edison Update



Jeff Broughton
NERSC Deputy for Operations
Systems Department Head

Edison at a Glance



- First Cray XC30
- Intel Ivy Bridge 12-core, 2.4GHz processors
- Aries interconnect with Dragonfly topology for great scalability
- Performs 2-4 x Hopper per node on real applications
- 12 x 512GB login nodes to support visualization and analytics
- 3 Lustre scratch file systems configured as 1:1:2 for capacity and bandwidth
- Access to NERSC's GPFS global file system via DVS
- Easy adoption for users because of similarity to Hopper
- Ambient cooled for extreme energy efficiency

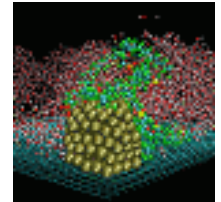
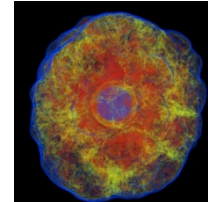
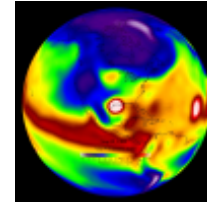
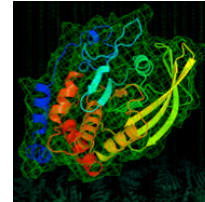
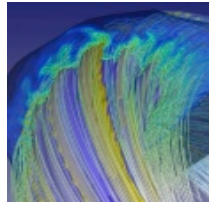
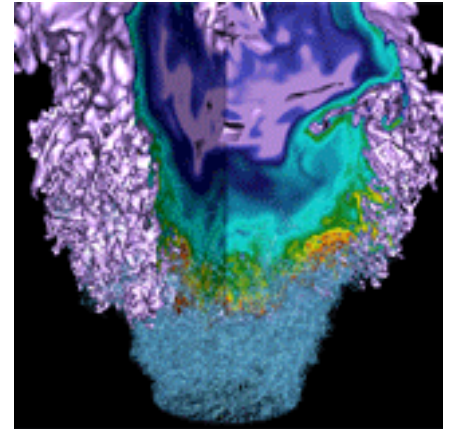
- **Memory configuration was changed in December as a workaround for a problem under investigation**
 - Memory BW was reduced from 102GB/s to 89GB/s per node
 - Performance impact on applications range from ~4% for most apps up to ~10% for memory intensive apps
 - An additional 384 nodes will be installed in February to compensate for performance loss
 - We hope to raise memory speed after a fix is found
- **I/O upgrades**
 - /scratch1 and /scratch2 have been upgraded to add 15GB/s and 500TB each
- **Edison accepted on December 27, 2013**
- **Edison will enter production on January 14, 2014 at the start of the NERSC ERCAP Allocation Year**
 - Frequency of maintenance outages will be substantially reduced
 - Charge factor will be 2 (relative to Hopper)

Vital Statistics



	Hopper	Edison (current)	Edison (Feb)
Cabinets	68	28	30
Compute Nodes	6,384	5,200	5,576
CPU Cores (<i>Total / Per-node</i>)	152,408 / 24	124,800 / 24	133,824 / 24
CPU Frequency (GHz)	2.1	2.4	2.4
Peak Flops (PF)	1.29	2.4	2.6
Memory (TB) (<i>Total / Per-node</i>)	217 / 32	333 / 64	357 / 64
Memory (Stream) BW* (TB/s)	331	462.8	496.3
Memory BW/node* (GB/s)	52	89	89
File system(s)	2 PB @ 70 GB/s	7.4 PB @ 180 GB/s	7.4 PB @ 180 GB/s
Peak Bisection BW (TB/s)	5.1	11.0	12.6
Power (MW Linpack)	2.9	1.9	2.0

Allocation Year Rollover



David Turner
HPC Consultant
NERSC User Services Group

Allocation Year Rollover 2013/2014



- **AY13 ends at 23:59:59 on Monday, January 13, 2014**
- **AY14 runs from Tuesday, January 14, 2014 through Monday, January 12, 2015**
- **No planned system outages associated with rollover**
 - Batch jobs will run uninterrupted
 - Time accrued before midnight will be charged to AY13 repos
 - Time accrued after midnight will be charged to AY14 repos
 - Running jobs associated with non-continuing repos will have the AY14 component of their usage charged to a NERSC overhead repo
 - Queued jobs associated with non-continuing repos will fail upon startup in AY14.

- **AY14 allocation award letters were emailed to PIs on December 13**
 - All awards are available on website
<http://www.nersc.gov/users/accounts/awarded-projects/2014-allocation-awards/>
- **New allocations (CPU time and mass storage units) *replace* those from AY13**
 - Unused allocations do *not* carry over into the new allocation year

AY14 Edison Changes



- **Edison becomes production resource**
 - Discontinue “fairshare” batch job scheduling
 - Implement age-based scheduling
 - Similar to approach used on Hopper and Carver
 - Begin repo charging
 - Same charging algorithm as Hopper and Carver
 - Edison charge factor is 2.0, versus 1.0 for Hopper and 1.5 for Carver
 - Premium and Low queues will be enabled
 - Maximum node count for reg_small, ccm_queue, low, and killable rises from 512 to 682
 - Minimum node count for reg_med rises to 683
 - Any job requesting 683 or more nodes receives a 25% charging discount
 - Unless run in premium
 - Also receives “one-day” scheduling priority boost

AY14 Hopper Changes



- **Eliminate charging discounts for large jobs**
 - Greater than 682 nodes
 - “One-day” scheduling priority boost remains in effect
 - Helps with scheduling large jobs
- **Premium queue temporarily disabled**
 - 08:00, Wednesday, January 8 through 08:00, Tuesday, January 14
 - No premium jobs will start
 - Jobs submitted to premium will be rejected
 - To prevent repos with large balances from dominating

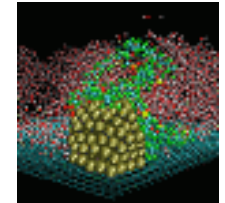
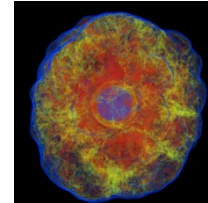
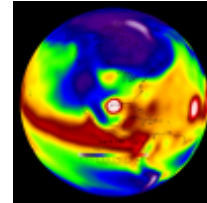
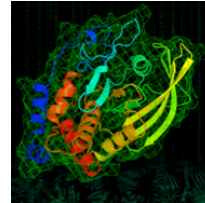
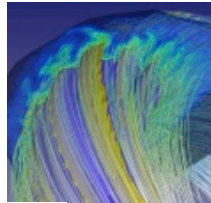
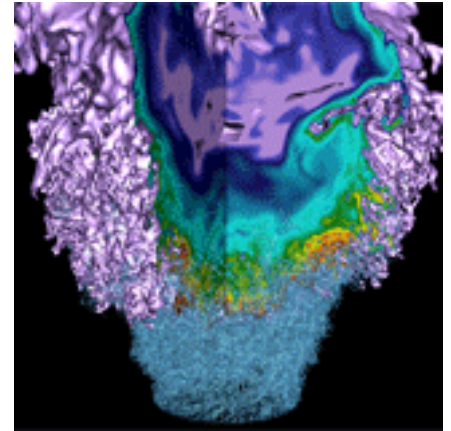
Additional Information



- **Web refreshing**
- **New repository membership**
- **Account creation/validation moratorium**
- **Discontinued users**
- **Deadline for AY13 startup requests**

<http://www.nersc.gov/users/announcements/allocation-year-rollover>

Project Directories for All



David Turner
HPC Consultant
NERSC User Services Group

Default Project Directories



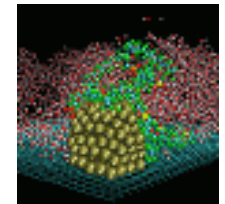
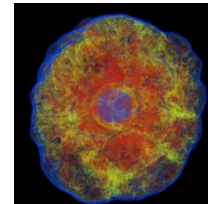
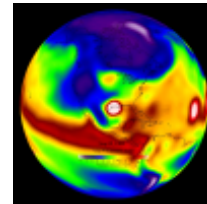
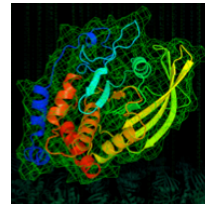
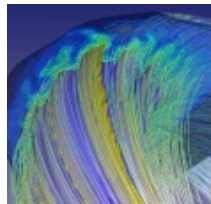
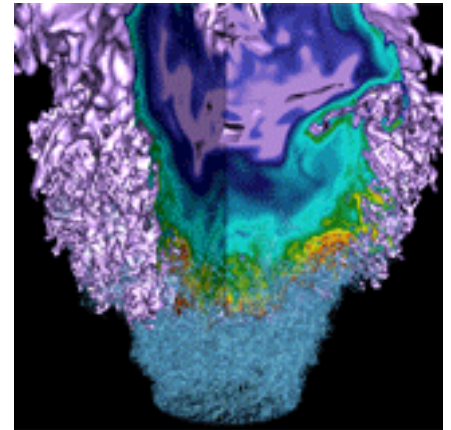
- **All MPP repos will automatically receive a project directory**
 - `/project/projectdirs/repo_name`
 - Repo members have access via repo Unix group
 - *Can* add non-repo members to repo Unix group for wider sharing
 - But a “designer” project directory might be more appropriate
 - Default quota of 1TB and 1M inodes
 - Can be increased by request
 - Usage and quotas available via NIM
 - And command-line (`prjquota`)

“Designer” Project Directories



- **Based on previous approach to project directories**
- **Requested by PI or PI Proxy**
 - Up to 80-character name
 - First 8 characters, when converted to lower case, must be unique
 - Unix group limitation imposed by NIM
- **PI/Proxy must manage group membership**
 - Simple NIM operation to add/remove users to/from group
- **Default quota of 1TB and 1M inodes**
 - Can be increased by request
 - Pre-AY14 project directories will retain current quotas
 - 4TB and 4M inodes

Give and Take



Jack Deslippe
HPC Consultant
NERSC User Services Group

What is Give and Take?

When you want to quickly share a file or directory with another user without messing with directory permissions, creating a joint project directory etc...

```
give -u <username> <file>  
take -u <username> <file>
```

Files given are temporarily stored in spool area and **count against your global scratch quota.**

How to Use Give

Give Usage:

```
$ give
Usage: give -u <username> <files>
-u is a required option
```

```
$ give -u jdeslip BGW-1.0.6.tar.gz
Giving BGW-1.0.6.tar.gz
    9166601 100% 106.23MB/s    0:00:00 (xfer#1, to-check=0/1)
Gave BGW-1.0.6.tar.gz
```

```
$ give -u jdeslip test_dir
Creating tar archive from directory test_dir
test_dir/
test_dir/dgemm.f90
test_dir/fftttest3.f90
test_dir/fftttest_omp.f90
Gave test_dir.tar.
```

How to Use Take

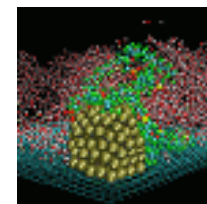
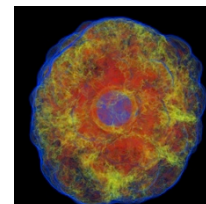
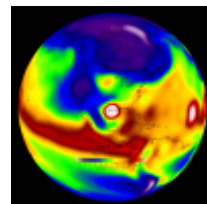
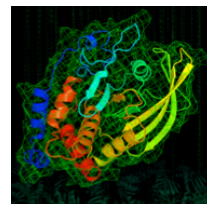
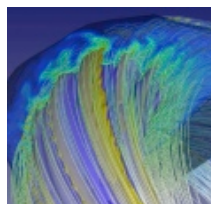
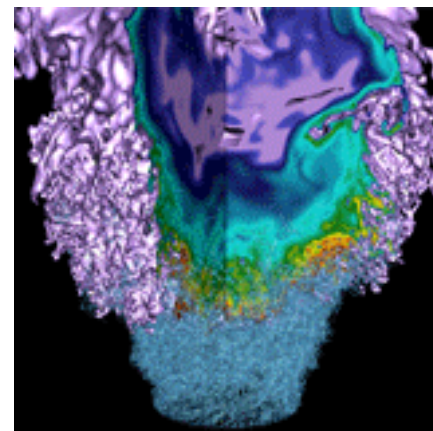
Take Usage:

```
$ take
Usage:
See available files from user: take -u <user>
Take one file from user:  take -u <user> -d <destination> <files>
Take all files from user: take -u <user> -d <destination> -a -d is
optional
```

```
$ take -u jdeslip
Files available from jdeslip:
8.8M2014-01-08 08:27 BGW-1.0.6.tar.gz
256K2014-01-08 08:32 test_dir.tar
```

```
$ take -u jdeslip BGW-1.0.6.tar.gz
Taking BGW-1.0.6.tar.gz
    9166601 100%  946.90kB/s    0:00:09 (xfer#1, to-check=0/1)
Ta Took BGW-1.0.6.tar.gz
```

Automatic I/O Profiling Your Code Using Darshan

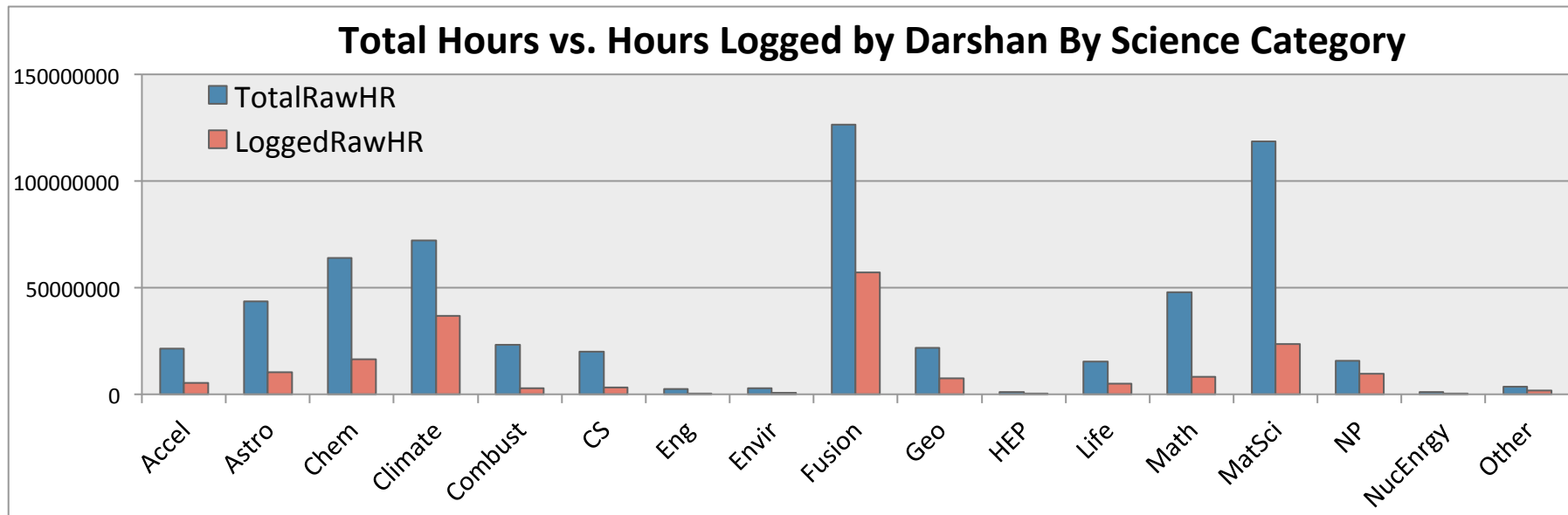


Yushu Yao
HPC Consultant
NERSC User Services Group

Darshan monitors user I/O statistics on Hopper since Nov. 2012



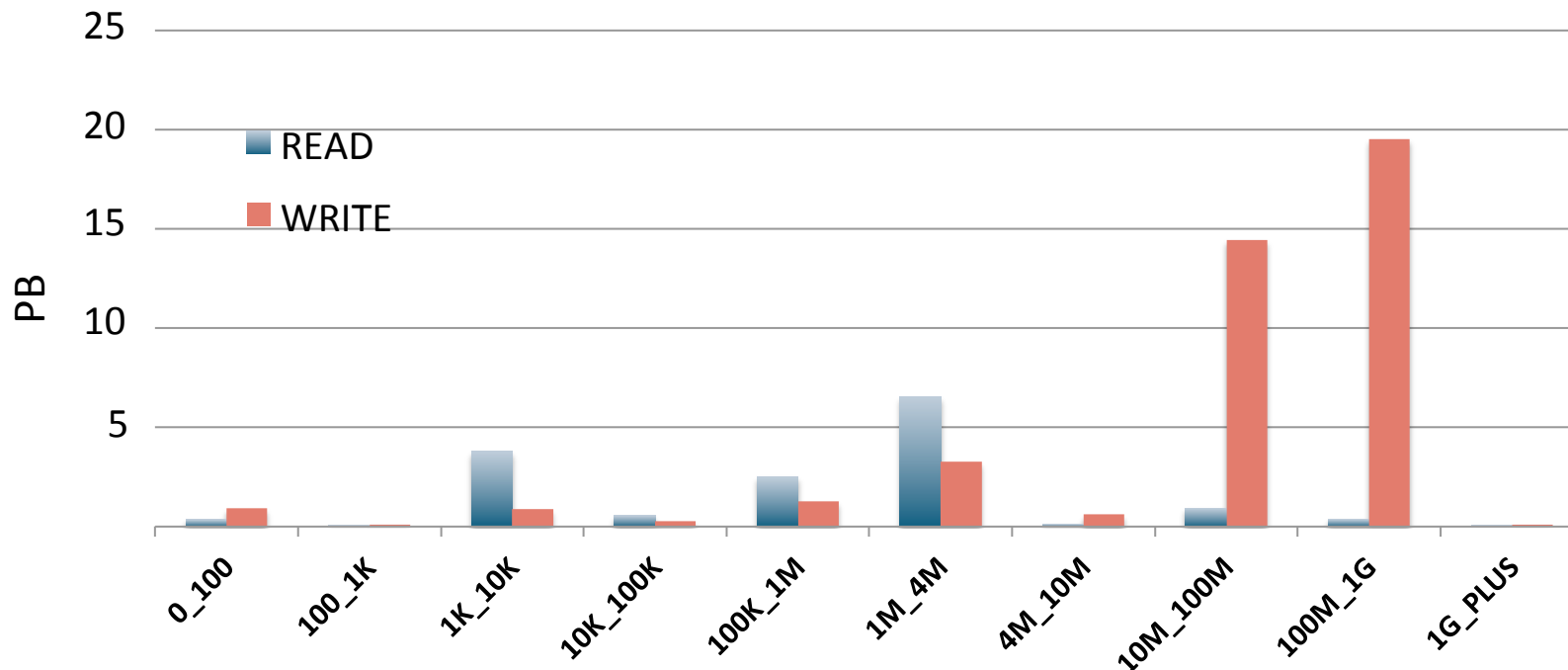
- Loaded by default into users environment, users must relink applications against Darshan library
- Captures I/O traffic of applications using MPI
- Approximately 40% of raw hours on Hopper are monitored by Darshan
- Captures I/O rates, transaction sizes and I/O library usage



Hopper I/O workload by transaction size

Distribution of I/O transactions Weighted by Size

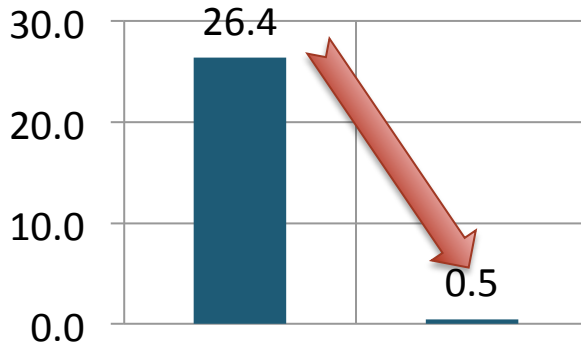
(Data taken from Mar 1 – Oct 31, 2013, approx 40% of Hopper raw hours captured)



- NERSC jobs in aggregate write 100MB per raw hour and read 70MB per raw hour
- NERSC applications write larger block sizes than they read

User spending too much time in I/O due to using the wrong file system

Walltime Minutes Spent in IO • Behavior:



- A 35K core job uses MPI-IO to write a small restart.dat file in Global Home
- Only a handful of seek/write per core, writing a very small amount (~100MB) of data

• Problem

- Spent 30% time writing this file
- Global Home will not perform at scale
- Using the wrong File System really hurts

• Suggestion

- Use Lustre (SCRATCH) instead of HOME

• Outcome

- Saved 30% (18K) MPP Hours Per Job

Saved 30% (18K) MPP Hours PER JOB

Darshan I/O Profiling Information is now available online

Locate the Completed Jobs Page



<https://www.nersc.gov/users/job-logs-and-analytics/completed-jobs/>



Enable Darshan IO Option

NERSC Completed Batch Jobs

https://www.nersc.gov/users/job-logs-and-analytics/completed-jobs/

COMPLETED BATCH JOBS

Select a time period

Show jobs that completed after Jan 1 2014 @ 0 : 0 :00 Pacific Time

and completed on or before Jan 8 2014 @ 11 : 29 :59 Pacific Time

Limit the number of jobs to display

Max. Number of results to show ☐ 50 ☐ 100 ☒ 500 ☐ 1,000 ☐ 5,000 ☐ 10,000

Refine search

Host Hopper User Repo Not Exec Queue All

Charge Class All JobID Node Name

Select columns to display

☒ Hostname	☒ Job ID	☒ Job Name	☒ User	☐ Exec Queue	☐ Charge Class	☐ Repo
☐ Status	☒ Nodes	☐ Cores	☐ Submitted	☐ Start	☒ Complete	
☒ Wallclock	☒ Raw Machine Hours*	☐ Raw MPP Hours**	☐ Wall Request	☐ Wait Time		
☐ Repeat column headings	☐ IPM jobs only					

* - Raw Machine Hours = Wallclock * nodes * cores/node
** - Raw MPP Hours = Raw Machine Hours * Machine Charge Factor (Hopper=1, Edison=2, Carver=1.5)
Neither of these numbers include charges associated with the queue in which the job ran (e.g. premium, reg_big, etc.)

Update Display Update with CSV Export Reset Form Load Default Display

See Your IO Information

NERSC Completed Batch Jo x

← → ↺ ⌂ <https://www.nersc.gov/users/job-logs-and-analytics/completed-jobs/ProcessRestfulQueryForm/62>

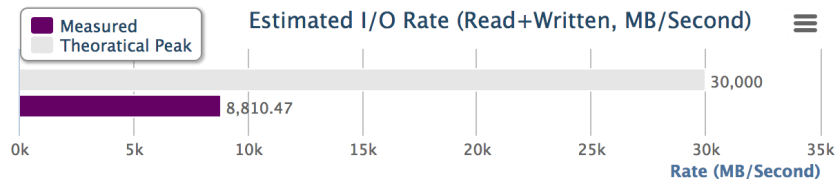
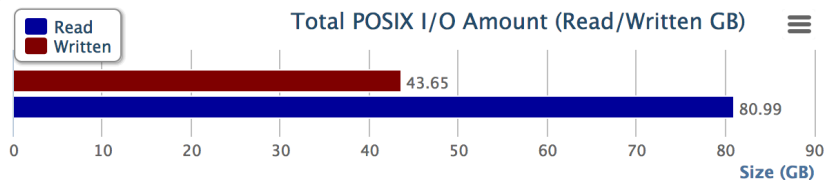
Displaying 500 jobs that completed between Jan-01-14 00:00 and Jan-08-14 11:29

Show 50 entries Search:

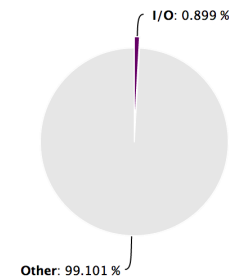
#	Darshan I/O Score Click For Details	Host	JobID	Job Name	User	Nds	Complete	Wall hrs	Raw Mach Hrs*
496		hopper	6963862	LZn0_009a	jbreslau	1	01/08/14 08:23	0.014	0.35
97		hopper	6964318	gest1260nscf	vormar	5	01/08/14 10:56	0.124	14.83
441		hopper	6951224	ca_orig	xhqu	100	01/08/14 08:45	22.356	53,654.67
398		hopper	6963873	lyf	yefeil	20	01/08/14 08:59	0.453	217.33
490		hopper	6963806	lyf	yefeil	20	01/08/14 08:27	0.452	216.93
115		hopper	6964264	bout_hoppe... [Full Name]	jingfei	43	01/08/14 10:49	0.123	126.71
0	No Log Found	hopper	6964529	GemiDbgPgi00	hargrove	2	01/08/14 11:26	0.028	1.32
1	No Log Found	hopper	6964526	timeprop	shish	1	01/08/14 11:25	0.017	0.41
2	No Log Found	hopper	6865647	Nb-Y.nph79	brightzh	8	01/08/14 11:25	8.226	1,579.31
3	No Log Found	hopper	6964517	run_pw	sahar	36	01/08/14 11:24	0.049	42.48
6	No Log Found	hopper	6964251	G4BLjob	aoliu	50	01/08/14 11:23	0.803	963.00
8	No Log Found	hopper	6963476	100B	peelaers	13	01/08/14 11:23	0.318	99.15

Drill Down to Job Details

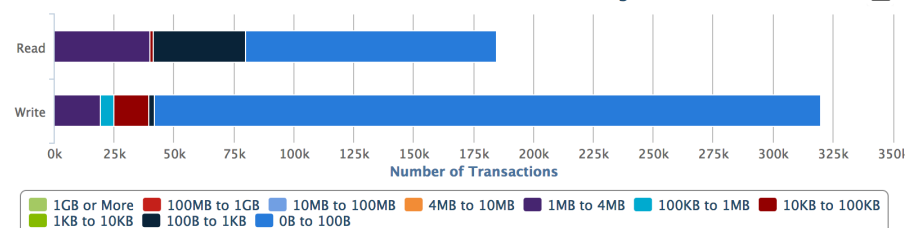
Host: hopper
JobID: 6963806
Command Line: pw.x
Number of PEs: 480
Start Time: 1/8/2014 8:00:46 AM
End Time: 1/8/2014 8:27:38 AM
Walltime: 1612 Seconds
Estimated I/O Rate: 8810.468533
Estimated Percentage Time Spent in I/O: 0.899 Percent



Percentage time spent on I/O



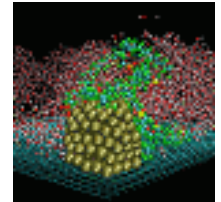
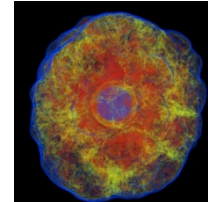
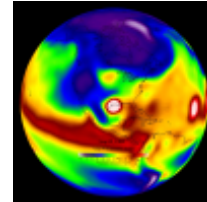
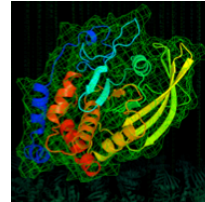
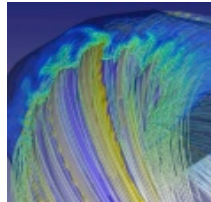
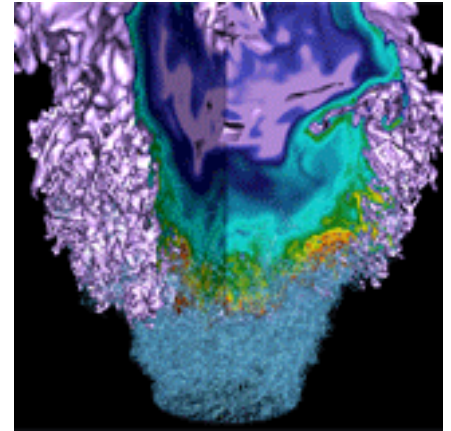
Number of Transaction in Each Size Range



Summary

- Darshan is useful for aggregate I/O behavior for Hopper
- Darshan is useful to help individual users to improve their I/O
- Now job analysis are available on the web
- We will continue to use Darshan on Hopper, and will deploy Darshan on Edison

Miscellaneous



NUG Queue Committee



- Being lead by Anubhav Jain from Berkeley Lab
- Contact ajain@lbl.gov if you want to participate
- Teleconference planned for mid January
- Topics
 - thruput queue - what's the policy and system for high-throughput jobs at NERSC?
 - debug queue - do current limitations in walltime/ncores cause any problems in trying to debug?
 - special queue and 2X charging - is this the best system for priority jobs?

NERSC User Survey for 2013



- **Survey is ongoing**
 - 301 respondents
 - Representing 19% of all usage
- **2012 Response Rate**
 - 552 respondents **We want to meet this target!**
 - Representing 61% of all usage
- **The survey is extremely important to NERSC, DOE and you, please give us your feedback at**
<http://www.nersc.gov/news-publications/publications-reports/user-surveys/2013/>
- **Large users' responses urgently needed!**

Other Issues or Questions?





Thank you.