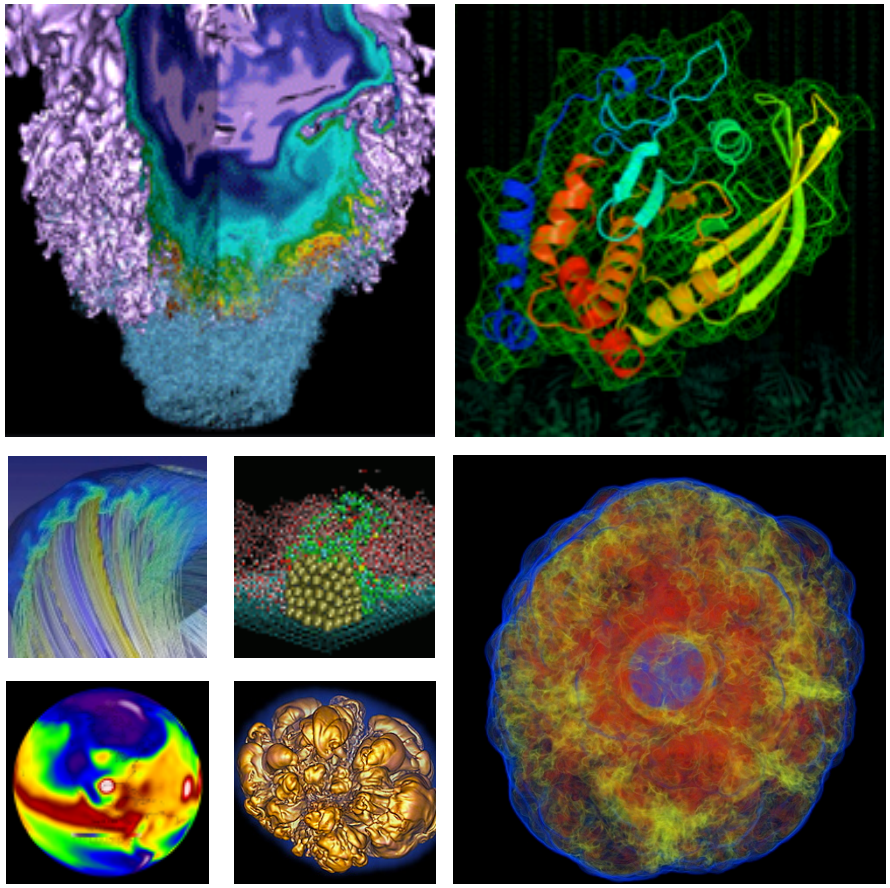
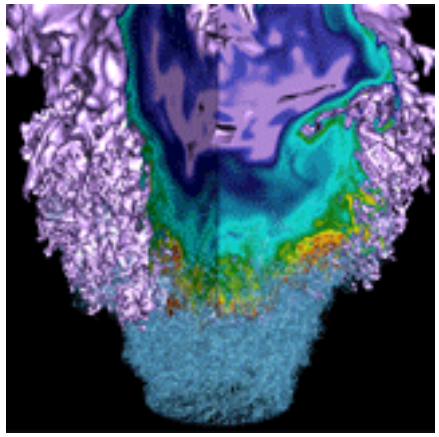
Edison and Cori: User Update

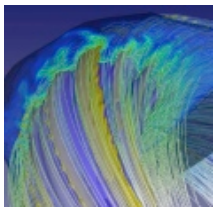


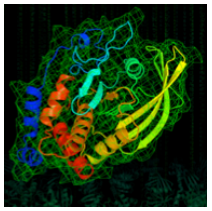
Zhengji Zhao, Helen He, Wahid Bhimji

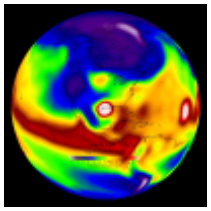
NERSC User Group Meeting
Berkeley, CA, March 24, 2016

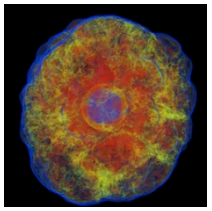
Edison Update

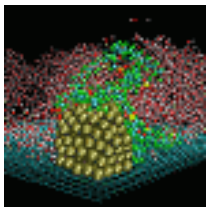












Zhengji Zhao

NERSC 40 YEARS at the FOREFRONT 1974-2014

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Edison upgrades (11/30/2015-3/15)

- **Edison move 11/30-12/23/2015**
  - Edison disassembled, reassembled, integrated, reconfigured and tested at CRT
  - 1/4/2016 users were enabled
  - Free charging period 1/4 – 1/10/2016
- **Switch to Slurm**
  - Slurm configuration has been in continuous improvement and adjustment
  - Users needed a lot of help with running jobs and workflow switch
  - Favor largely to big jobs
  - Major issue is the slow queue turnaround – we are working on it
- **/scratch3 upgrade to Grid Raid**
  - I/O performance issue is still in investigation

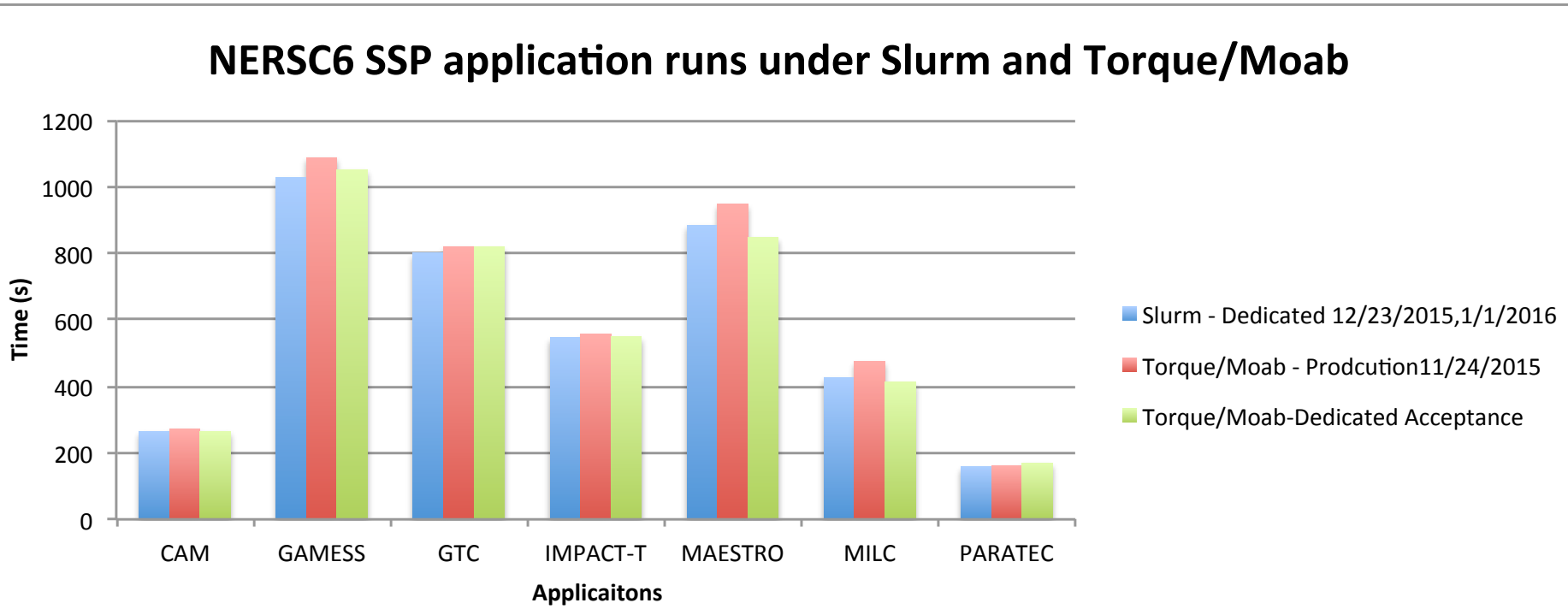
Edison upgrades

- **Host IP change**
  - Users had ssh issues to login
- **NEW SSH authentication mechanism (1/12/2016)**
  - Login issue as well
- **Edison experienced multiple planned and unplanned down times (power outage) during Jan - Mar, 2016.**
  - User jobs affected
- **CDT upgrades on 12/23/2015 (15.12), 2/3/2016(16.01), 3/22/2016 (16.03)**
  - Encountered a few major bugs;Workarounds provided for all bugs, and major bugs were fixed as of 1/15; A remaining bug will be fixed in CDT 16.03. Fixes were in place on 3/22/2016.
  - Extended the CDT testing script to include more tests
  - Default option --craype-buildtools-check

Edison upgrades

- **A shorter purging period will be in place effective 4/1/2016**
  - Purging period will be 8 weeks (from 12 weeks)
  - 84%, 81%full on /scratch1 and 2 file systems currently
- **/scratch3 quota in place as of 3/17/2016**
  - Quota 100TB disk space, 50,000,000 inode
  - Quota check will be in place in the job submission filter, fail the submission if over quota
  - 74% full
- **/scratch1 and /scratch2 will be upgraded to Grid Raid, time TBD**
  - Depending on when the current /scratch3 performance bug is resolved

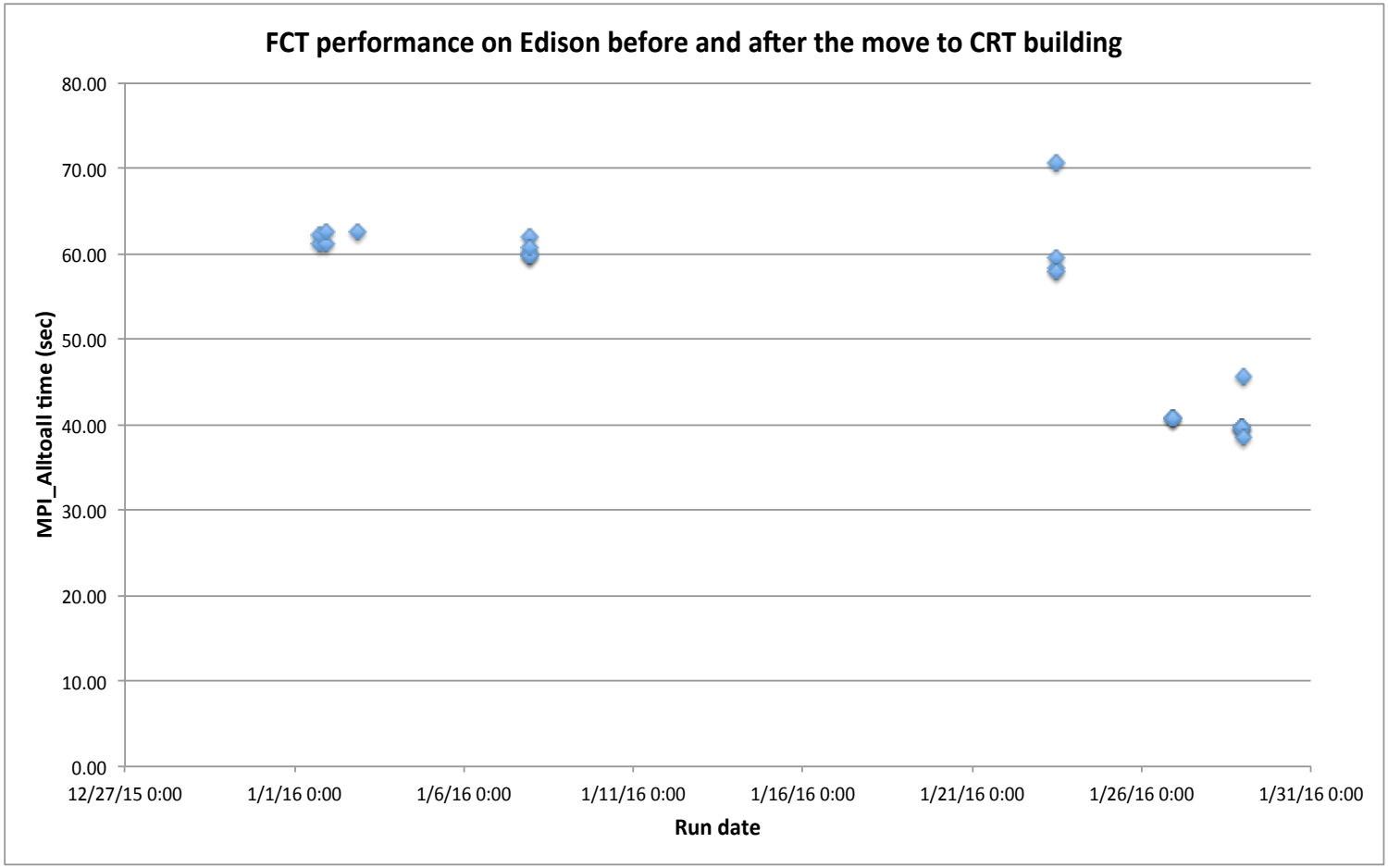
SSP benchmark performance after the move



Edison performance monitoring:

https://my.nersc.gov/benchmarks-cs.php

FCT performance regression is resolved



Before the move

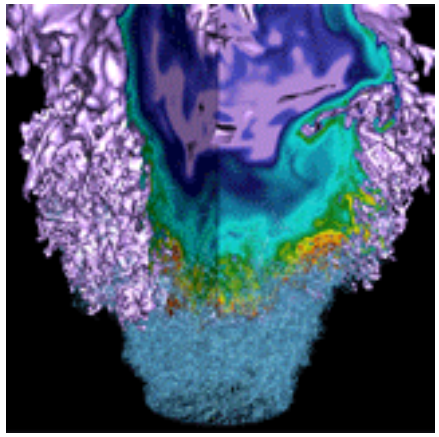
Run date	output/KEY:tag	ntasks	MPI_Alltoall time(sec)
2/19/14	fct99p1.o785758	132367	40.94
1/15/15	fct100p1.o2274657	133296	33.36

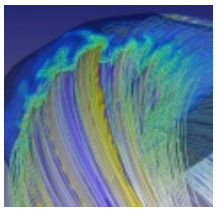
I/O performance degradation after the Grid Raid upgrade is still in investigation

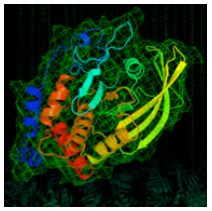
Date	Time	Job description: File per process IOR						Write (MB/s)	Read (MB/s)	Comment
12/24/15	2:22	FS3	1m2	1152ranks	8fpo	24ppn	144osts	21718	23952	GRIDRAID
12/24/15	7:48	FS3	1m2	288ranks	8fpo	24ppn	36osts	17602	22911	GRIDRAID
3/26/15	2:39	FS3	1m2	1152ranks	8fpo	24ppn	144osts	62663	54416	MDRAID

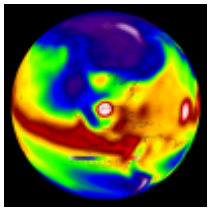
- **This is roughly 3 times performance degradation.**
- **NERSC needs the recommendation from Cray and Seagate about how to run IOR benchmark to compare with the MDGRID performance.**

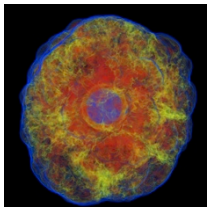
Cori Update

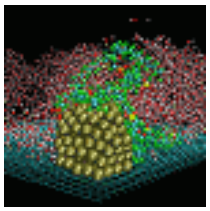






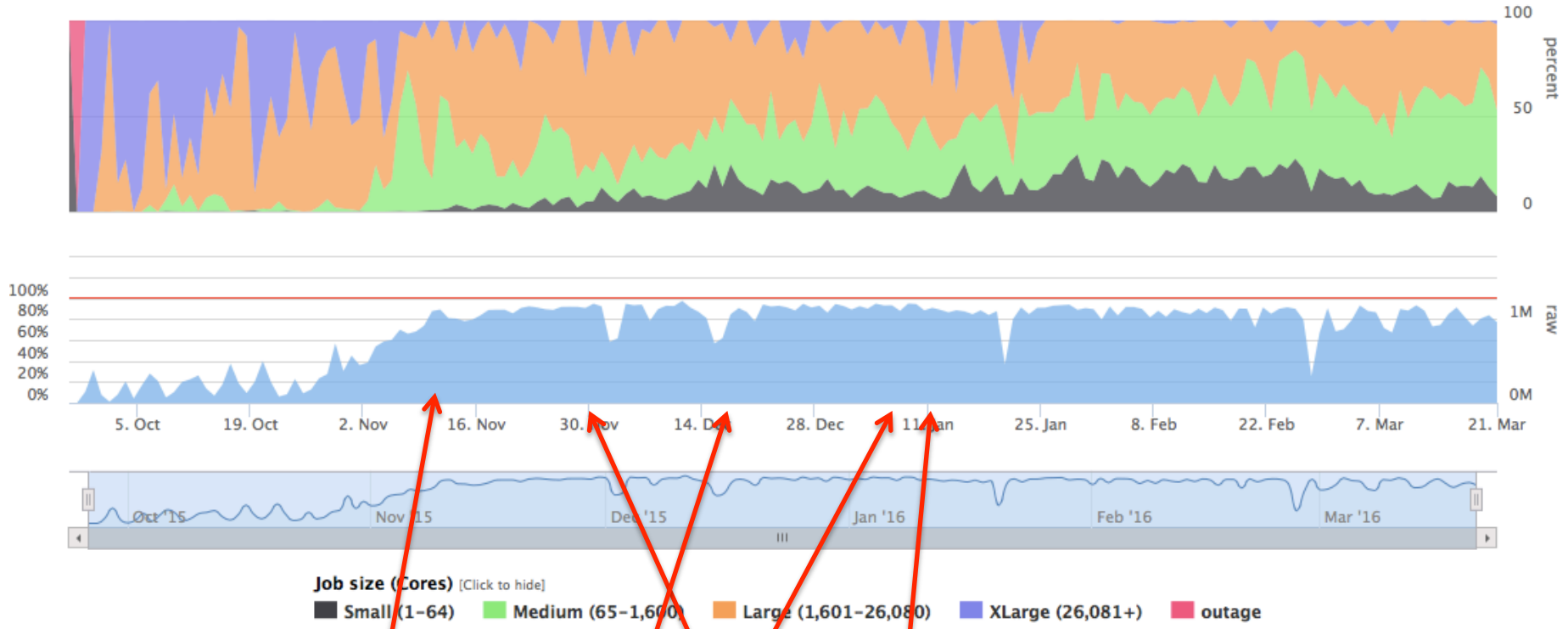






Helen He and Wahid Whimji

Cori Usage Info



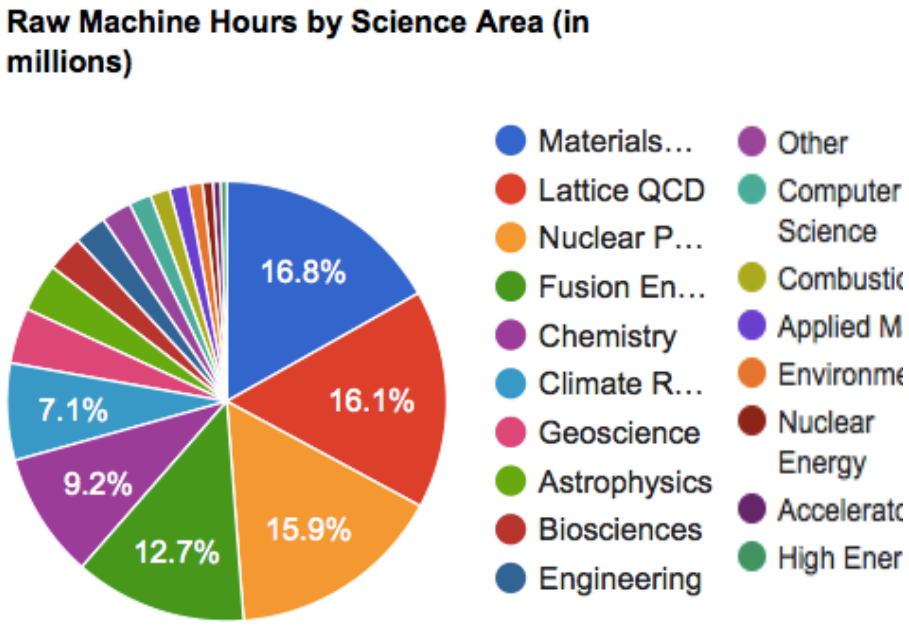
- 11/12/2015: All users enabled
- 11/30/2015 – 1/4/2016: Edison offline
- 12/15/2015: Hopper retired.
- 1/12/16: Cori started charging when AY16 began

More large jobs during free time ☺

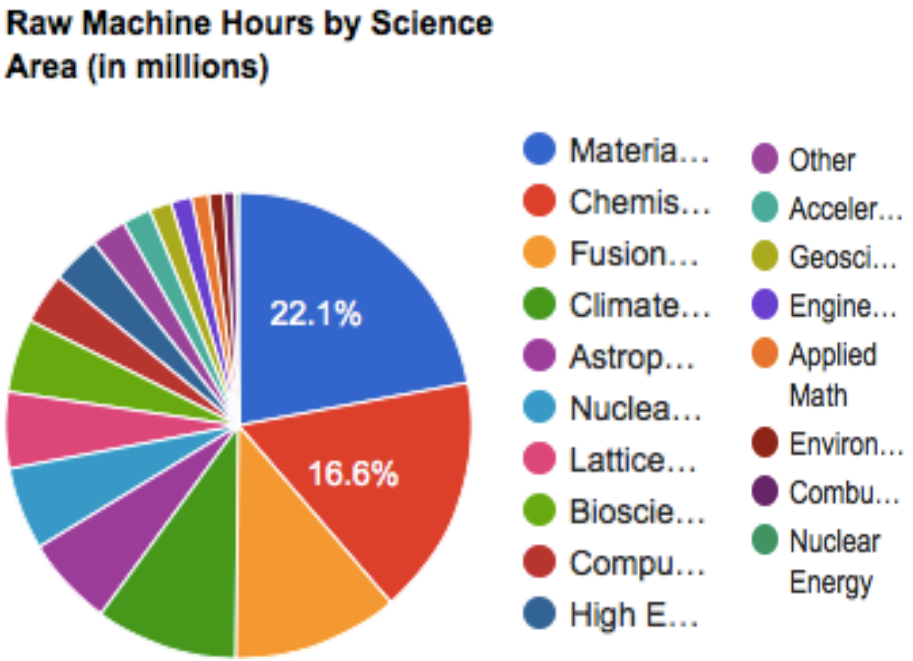
Cori Usage Info: Free Period and AY16

- **Early users were enabled in 7 phases:**
  - **Allow Cori system became ready in various aspects (networking, programming environment, batch system, etc.)**

162M MPP hours used (10/29/15-1/11/16)



75.8M MPP hours used (1/12/16-3/22/16)



Cori Phase 1 Data Features

- **File Systems**
  - Burst Buffer for high bandwidth, low latency I/O
  - High performance Lustre file system: 28 PB of disk, >700 GB I/O bandwidth
  - Cross mounting of file systems (Cori scratch on Edison and DTNs) (TBA)
  - Large amount of memory per compute node (128 GB) as well as some high memory login nodes (775 GB).
- **Networking**
  - Improved outbound Internet connections (eg. to access a database in another center)
  - Software Defined Networking R&D for high bandwidth transfers in and out of the compute node (TBA)
- **On node software**
  - Improved shared library performance
  - User-defined images/Shifter

Cori Phase 1 Data Features (SLURM)

- **Cori Phase 1 also known as the "Cori Data Partition"**
- **Designed to accelerate data-intensive applications, with high throughput and "real time" need.**
  - "shared" partition. Multiple jobs on the same node. Larger submit and run limits. 40 nodes set aside
  - The 1-2 node bin in the "regular" for high throughput jobs. Large submit and run limits.
  - "realtime" partition for jobs requiring real time data analysis. Highest queue priority. Special permission only.
  - Internal sshd (CCM mode) in any queue
  - Large number of login/interactive nodes to support applications with advanced workflows
  - "burst buffer" usage integrated in SLURM, in early user period.
  - Encourage users to run jobs using 683+ nodes on Edison with queue priority boost and 40% charging discount there.

Transition from Hopper/Edison to Cori

- Programming environment is very similar to Hopper/Edison. Porting to Cori is straightforward in regards to software building.
- The aspect that users need to adjust the most is the transition from Torque/Moab to SLURM.
- Provided detailed documentations on SLURM transition guide, example batch scripts, and tutorials.
- Worked with some specific applications and users for the porting. CESM is one such example. It is a new machine port, with bfb required.

SLURM Batch Scheduler Adoption

- **Overall SLURM adoption is smooth.**
- **Easy to use "premium", "ccm", good support and usage for "shared" and "realtime".**
- **A few traps (with user education):**
  - Hyperthreading is on by default
    - SLURM sees 64 CPUs per node
    - Asking nodes with "#SBATCH –n", but without "#SBATCH –N" may get half the node desired
    - Need to set OMP_NUM_THREADS=1 explicitly to run with pure MPI (for hybrid MPI/OpenMP program compiled with openmp enabled)
  - Automatic process and thread affinity is good. Can explore with advanced settings for more complicated binding options.

Batch Job Wait Time

- **Users reported about LONG wait time for jobs**
- **Monitoring and tuning SLURM configuration is an ongoing task**
- **Changes made on Jan 15**
  - Added max number of backfill jobs per partition (on top of max number of backfill jobs per user)
  - Decreased max size of debug from 128 to 112.
  - Communicated with individual users to use the “shared” partition, job arrays, and bundling jobs.
  - Jobs do not plan to run in AY16 were deleted
  - Most debug jobs then started within 30 min instead of hours, many now start in a few min.
  - The regular jobs wait time are significantly smaller too
- **Changes made on Mar 22 for the scheduling algorithm greatly increased system utilization (keep watching ☺)**

NERSC Custom Queue Monitoring Script

- **Original "sqs" provides basic batch job info plus the job ranking based on start time provided by the backfill scheduler.**
- **A new version of "sqs" was deployed on Jan 19 with two columns of ranking values to give users more perspective of their jobs in queue.**
  - Added job priority ranking with absolute priority value (a function of partition, QOS, job wait time, and fair share)

A Few Tips to Get Faster Job Turnaround

- Request shorter wall time, do not use allowed max wall time.
- Use “shared” partition for serial jobs or very small parallel jobs.
- Bundle jobs (multiple “srun”s in one script, sequential or simultaneously)
- Use Job Arrays (better managing jobs, not necessary faster turnaround). Each array task is considered a single job for scheduling.
- Use job dependency feature for managing workflow.

Resolved: Cray HDF5 with Intel16

- **Internal compiler error for Fortran codes when using cray-hdf5, and cray-hdf5-parallel/1.8.14 with intel/16.0.0.109**
- **Two workarounds:**
  - Use NERSC built hdf5/1.8.14 and hdf5-parallel/1.8.14 with Intel/16.0.0.109 compiler
  - Use cray-hdf5/1.8.14, but swap intel compiler version from 16.0.0.109 to 15.0.1.133.
- **cray-hdf5/1.8.16 has been installed and set to default which resolved this issue (Feb 27, 2016)**

Workaround: Node Voltage Fault

- Compute node voltage fault only seen with one specific Quantum Espresso application “pw.x”.
- By default, hyperthreading is used. And the application generates a very close sequence of current spikes that may cause the Voltage Converter to self-protect and shut down.
- Workaround by user education to use 1 thread per MPI task. Also modified the NERSC provided module file to set OMP_NUM_THREADS=1. (Jan 16, 2016)

Resolved: /project IO performance

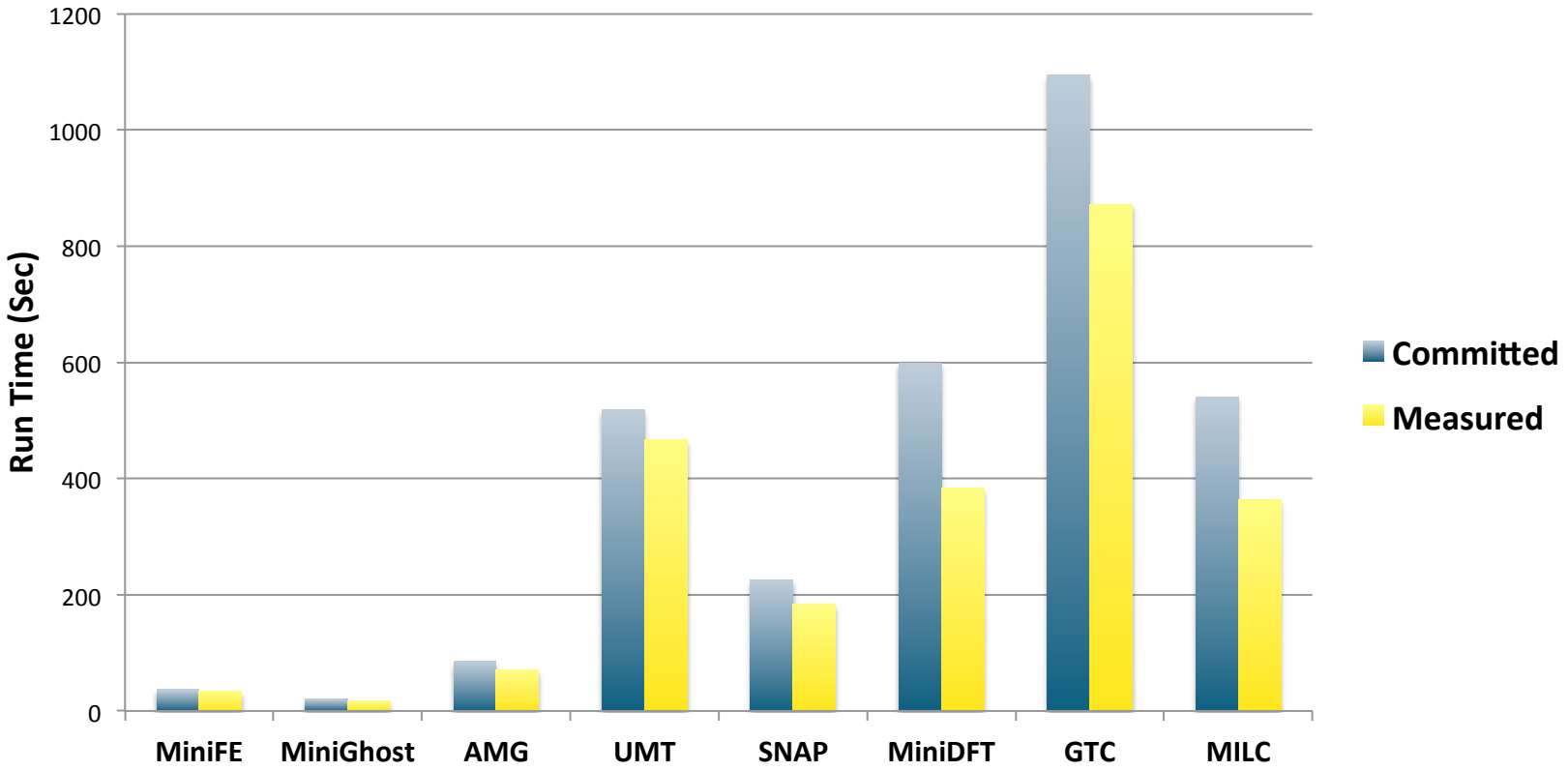
- **Two applications reported 10x parallel IO performance slowdown in /project, seen after Dec 25, 2015.**
- **Fixed during system reboot with scheduled maintenance on Jan 20, 2016.**
- **Exact cause of slowdown unknown**
  - Unlikely due to "Cori DVS nodes GPFS IB cable not used"

Current Issues

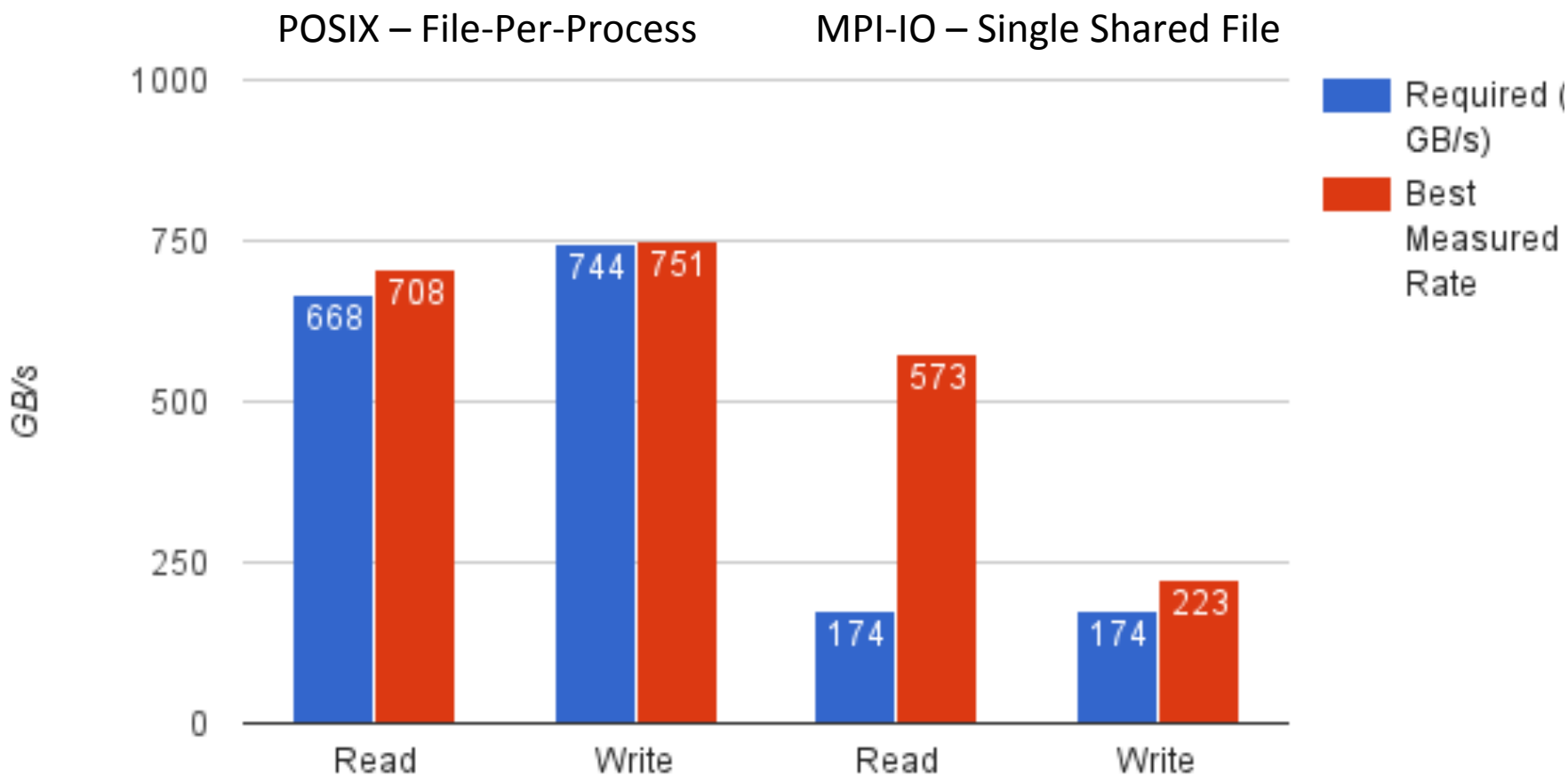
- Login nodes crash when hitting Lustre file system bug
- Compute nodes stuck in completing states from certain Burst Buffer jobs
- Compute nodes went down with out-of-memory error from certain applications
- Burst Buffer still in early user period

Cori Phase 1 SSP Performance

Committed SSP: 68.2
Measured SSP: 83.0



Peak Cori Scratch Lustre I/O Performance



Thank you.