Fermilab Facilities report

Amitoj Singh

USQCD All-Hands Collaboration Meeting, BNL

29-30 April 2016

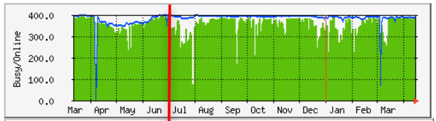
Hardware – Current Clusters

Name	CPU	Nodes	Cores GPUs	Network	Equivalent Jpsi core or Fermi gpu-hrs	Online
Ds	Quad 2.0 GHz Opteron 6128 (8-core)	421	13472	Infiniband QDR	**1.33** Jpsi	Dec 2010 Aug 2011
Dsg	*Dual NVIDIA M2050 GPUs+Intel 2.53 GHz E5630 (4-core)*	*76*	*608 Cores* *152 GPUs*	*Infiniband QDR*	***1.1*** *Fermi*	*Mar 2012*
Bc	Quad 2.8 GHz Opteron 6320 (8-core)	224	7168	Infiniband QDR	**1.48** Jpsi	July 2013
Pi0	Dual 2.6 GHz Xeon E2650v2 (8-core)	314	5024	Infiniband QDR	**3.14** Jpsi	Oct 2014 Apr 2015
Pi0g	*Dual NVIDIA K40 GPUs+Intel 2.6 GHz E2650v2 (8-core)*	*32*	*512 Cores* *128 GPUs*	*Infiniband QDR*	***2.6*** *Fermi*	*Oct 2014*

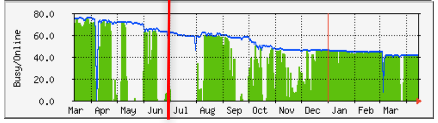
Progress Against Allocations

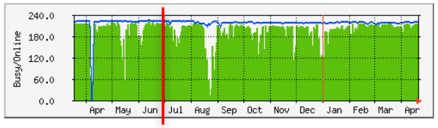
- Total FNAL allocation: 329 M Jpsi core-hrs, 3522 GPU-KHrs, 19M Jpsi core-hrs for storage
- Delivered to date: 263 M (89%), 2359 GPU-KHrs (78%), at 82% of the year
  - Does not include disk and tape utilization (roughly 19M + 2.4M)
  - Class A (25 total): 5 finished, 7 at or above pace
  - Class B (6 total): 2 finished, 2 at or above pace
  - Class C: 5 for conventional, 3 for GPUs
  - Opportunistic: 1 for conventional, 4 for GPUs

Ds

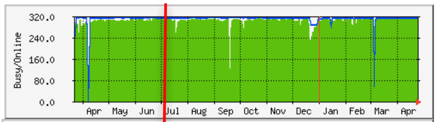


Dsg

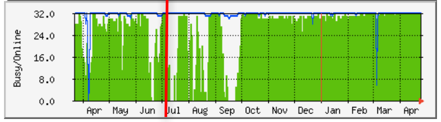




Bc



Pi0



Pi0g

Storage

- Global disk storage:
  - 1.1 PB Lustre file-system at /lqcdproj.
  - 7.25 TB "project" space at /project (backed up nightly)
  - 6 GB per user at /home on each cluster (backed up nightly)
- Robotic tape storage is available via *dccp* commands against the dCache filesystem at /pnfs/lqcd.
  - Some users will benefit from direct access to tape by using *encp* commands on lqcdsrm.fnal.gov
  - Please email us if writing TB-sized files. With 8.5TB tapes, we may want to guide how these are written to avoid wasted space.
- Worker nodes have local storage at /scratch.
- Globus Online endpoint:
  - lqcd#fnal - for transfers in or out of our Lustre file system.

Storage – Data integrity

- Some friendly reminders:
  - **Data integrity is your responsibility.**
  - With the exception of home area and /project, backups are not performed.
  - Make copies on different storage hardware of any of your critical data.
  - Data can be copied to tape using *dccp* or *encp* commands. Please contact us for details. We have never lost LQCD data on Fermilab tape.
  - At 111 disk pools and growing on Lustre, the odds of a partial failure will eventually catch up with us.

Lustre File-System

- Lustre Statistics:
  - Capacity: 1.1 PB available, 923 TB used (82% used)
  - Files: 76 million (108M last year)
  - File sizes: largest file is 4.91 TB, average size is 14 MB

- Please email us if writing TB-sized files. For Lustre there will be tremendous benefit in striping such files across several OSTs both for performance and for balancing space used per storage target.

- NOTE: No budget till FY18 to grow disk storage capacity. Please remove or copy old data to tape.

Lustre File-System

- Lustre Migration
  - **Thank you** for your patience as we migrated ~600TB worth of data from old (1.8.9) to new (2.5.3) Lustre.
  - ZFS under Lustre 2.5.3 allows compression of data and we were able to recover some storage space in a few cases.
  - ZFS under Lustre 2.5.3 provides adequate redundancy which allows bit rot detection and corrective action.
  - Last step in the migration process: We still have to migrate 190TB worth of actively accessed data. A few projects responsible for this data will need to pause in order for us to complete this migration.

Upcoming upgrades and major changes

- Ds and Dsg clusters:
  - For the 2016-17 program year, the Ds and Dsg clusters will be available to you as an unallocated resource.
  - As of now there are 389 Ds and 40 Dsg worker nodes in good to fair condition.

- Bc cluster relocation:
  - At the beginning of the 2016-17 program year we plan to replace Ds worker nodes with Bc cluster worker nodes.
  - Combined cluster will consist of: 200Ds + 40Dsg + 224Bc
  - All clusters will be binary compatible and use Mellanox based Infiniband for high speed interconnect.

Upcoming upgrades and major changes

- OS upgrade:
  - Once Lustre data migration is complete, we will no longer need older Lustre clients (1.8.9) which prohibited us from upgrading the Infiniband fabric software.
  - Upgrading the IB software stack will enable all available Mellanox/NVIDIA GPUDirect optimizations on Pi0g.
  - The above upgrade will require rebuilding of MPI libraries and user binaries will have to be rebuilt.
- Fermilab Service Desk:
  - As part of the Fermilab ISO20K certification for IT services, in early June we will be required to use the Service NOW ticketing facility to track user incidents and requests. You will continue to email us at lqcd-admin@fnal.gov.

User Support

Fermilab points of contact:

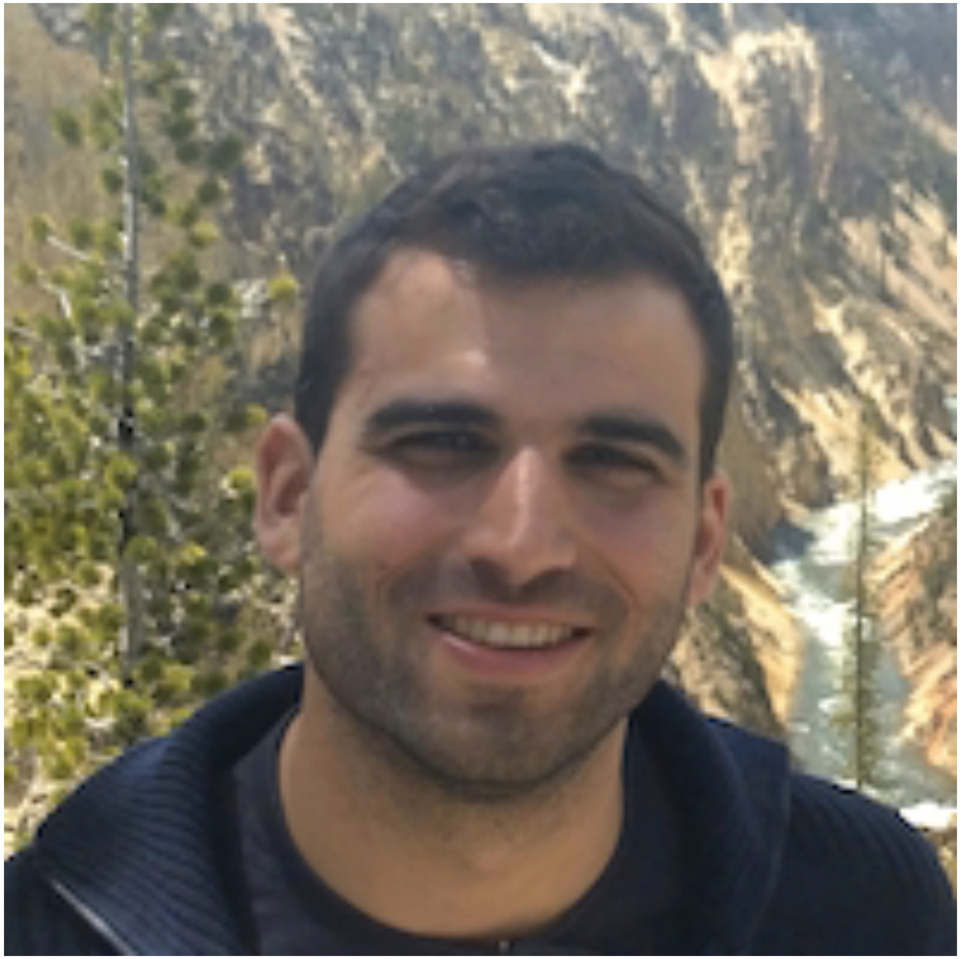
Please avoid sending support related emails directly to the POCs.

- Jim Simone, simone@fnal.gov
- Amitoj Singh, amitoj@fnal.gov
- Gerard Bernabeu, gerard1@fnal.gov
- Alexei Strelchenko, astrel@fnal.gov (GPUs)
- Alex Kulyavtsev, aik@fnal.gov (Mass Storage and Lustre)
- Ken Schumacher, kschu@fnal.gov
- Rick Van Conant, vanconant@fnal.gov
- Paul Mackenzie, mackenzie@fnal.gov

Please use lqcd-admin@fnal.gov **for incidents or requests.**

Don is retiring

Welcome Gerard Bernabeu Altayo



Questions?