# Report on the Clusters at Fermilab

Don Holmgren
USQCD All-Hands Meeting
JLab
May 6-7, 2011

# Outline

- Current Hardware

- Storage

- FY11 Deployments (clusters, GPUs, storage)

- Statistics

- Policies

- User Survey

# Hardware – Current Clusters

| Name | CPU | Nodes | Cores | Network | DWF | Asqtad | Online |
|------|-----|-------|-------|---------|-----|--------|--------|
| Kaon | Dual 2.0 GHz Opteron 240 (Dual Core) | 600 | 2400 | Infiniband Double Data Rate | 4696 MFlops per Node | 3832 MFlops per Node | Oct 2006 2.56 TFlops |
| J/$\psi$ | Dual 2.1 GHz Opteron 2352 (Quad Core) | 856 | 6848 | Infiniband Double Data Rate | 10061 MFlops per Node | 9563 MFlops per Node | Jan 2009 / Apr 2009 8.40 TFlops |
| Ds (2010) | Quad 2.0 GHz Opteron 6128 (8 Core) | 245 | 7840 | Infiniband Quad Data Rate | 51.2 GFlops per Node | 50.5 GFlops per Node | Dec 2010 11 TFlops |
| Ds (2011) | Quad 2.0 GHz Opteron 6128 (8 Core) | 176 | 5632 | Infiniband Quad Data Rate | 51.2 GFlops per Node | 50.5 GFlops per Node | 50% June, 50% Sept |
| GPU (2011) | NVIDIA C2050 | 2/host | 128 GPUs | Infiniband Quad Data Rate | | | Oct 2011 |

# Storage

- Disk storage options:

  – 387 TB  Lustre filesystem at /lqcdproj

  – 3.2 TB  total "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

# Storage – Planned Changes

1. Move /project to the Lustre filesystem –  June or July 2011

   - Currently /project is on 5-year old disk array

   - New home will be under /lqcdproj, but via symbolic links this change will be transparent to your scripts and programs

   - We need to know from all users when the move can occur

   - /project will continue to be backed up nightly, but as it inevitably grows we will not be able to provide restores from as far back as we can now (1 year)

2. Rearrange and enforce group quotas on /lqcdproj – July 2011

   - We must rearrange directory layouts to allow us to fairly charge projects for usage, and to control usage via group quotas

   - We will do this during Lattice'11 quiet time, and we will also upgrade the JPsi cluster to be binary compatible with Ds

3. Deploy additional Lustre storage (+ 200 TB → 587 TB total) – added gradually during the next allocation year

# Storage – Date Integrity

- Some friendly reminders:

  – Data integrity is your responsibility

  – With the exception of home areas and /project, backups are not performed

  – Make copies on different storage hardware of any of your data that are critical

  – Data can be copied to tape using dccp commands.  Please contact us for details.  We can also show you how to make multiple copies that are guaranteed to be on different tapes.  We have never lost LQCD data on Fermilab tape (750 TiB and growing).

  – At 110 disk pools, the odds of a partial failure will eventually catch up with us – please don't be the unlucky project that loses data when we lose a pool.

# Or 12 pools…

# Storage - Utilization
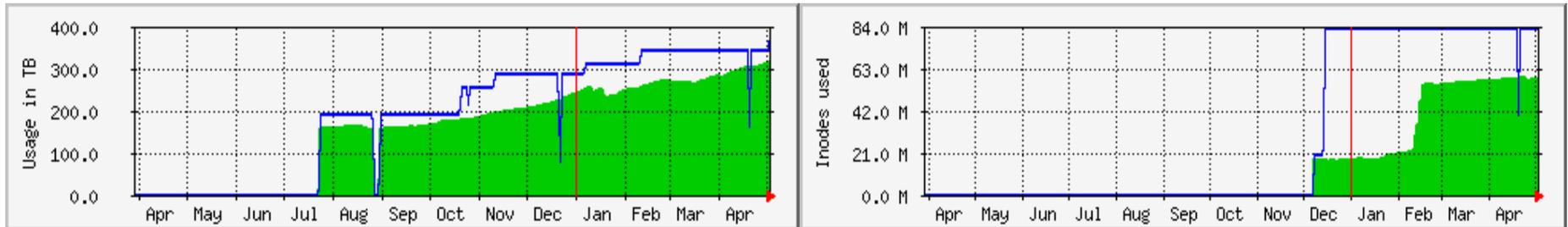


- Utilization of /lqcdproj will always increase to fill all space.  This is a good thing (disk is expensive – we don't mind you using it).

- But:

  - Lustre misbehaves when the pools get to 95% fill.  Please be responsive to our requests to clear space.  If users prefer, we can set up a scratch partition similar to JLab in which older files are automatically deleted to clear space.

  - Last week we reached a 95% fill state.  One user detected a file that was truncated when copied from one part of /lqcdproj to another part.  If you notice any problems, please let us know (lqcd-admin@fnal.gov)

  - For our planning purposes, it is critical that in your proposals that storage requests are reasonably (factor of 2) accurate.  We have instances of both large overruns (20 TiB when zero was requested) and under-utilization.  We can adjust budgets annually, but we need reliable data.
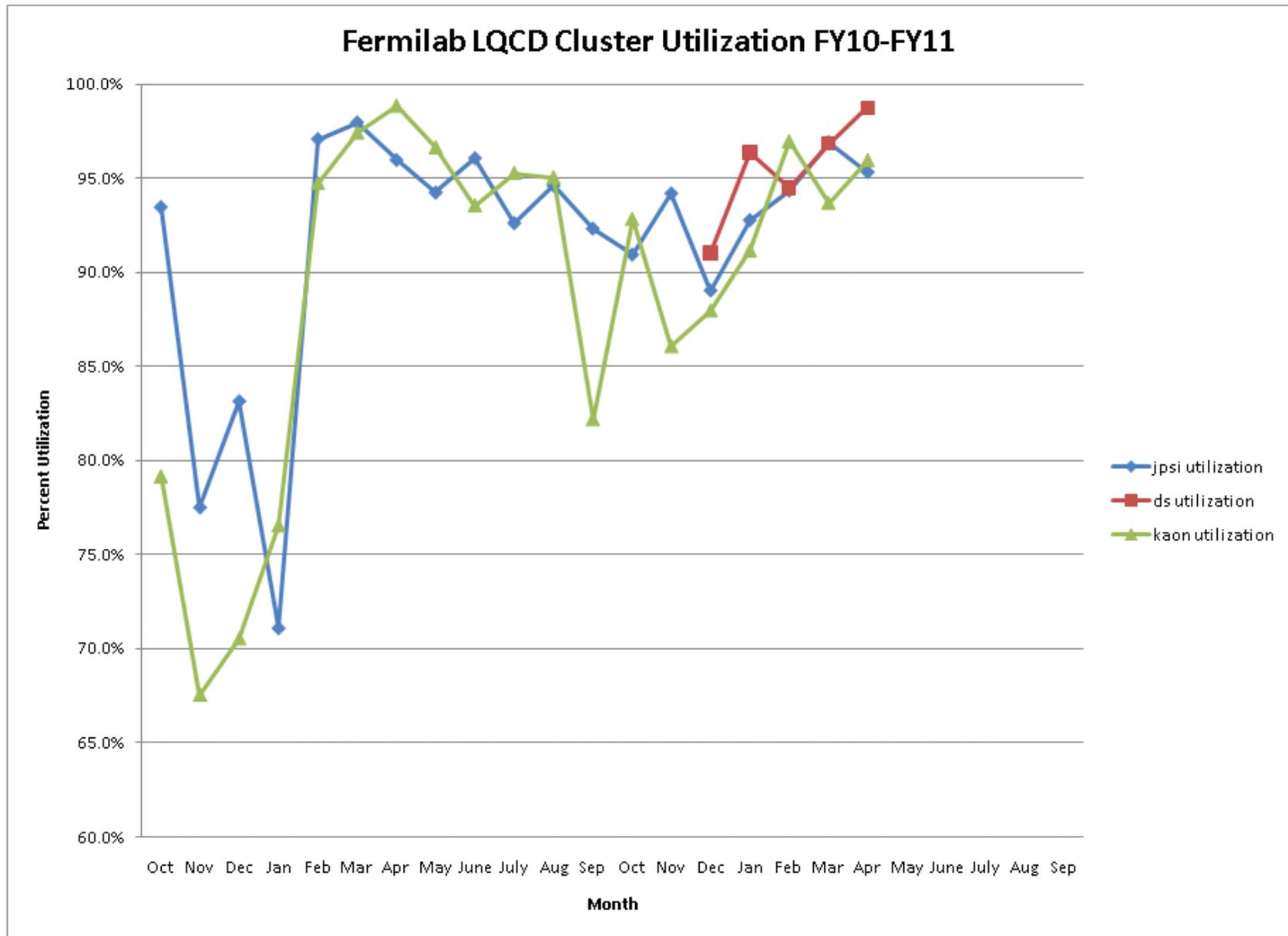
# FY11 Deployments

- There have been a total of 8 continuing budget resolutions

  – Fermilab spending was throttled because of these CR's

  – We planned to order 176 additional Ds nodes in January, but were only able to order 88 nodes in March (arriving now)

  – As soon as Fermilab receives final FY11 budget guidance, we will order the other 88 nodes

  – The CR's have also prevented us from buying the planned GPU cluster – we will do so once budget is available.
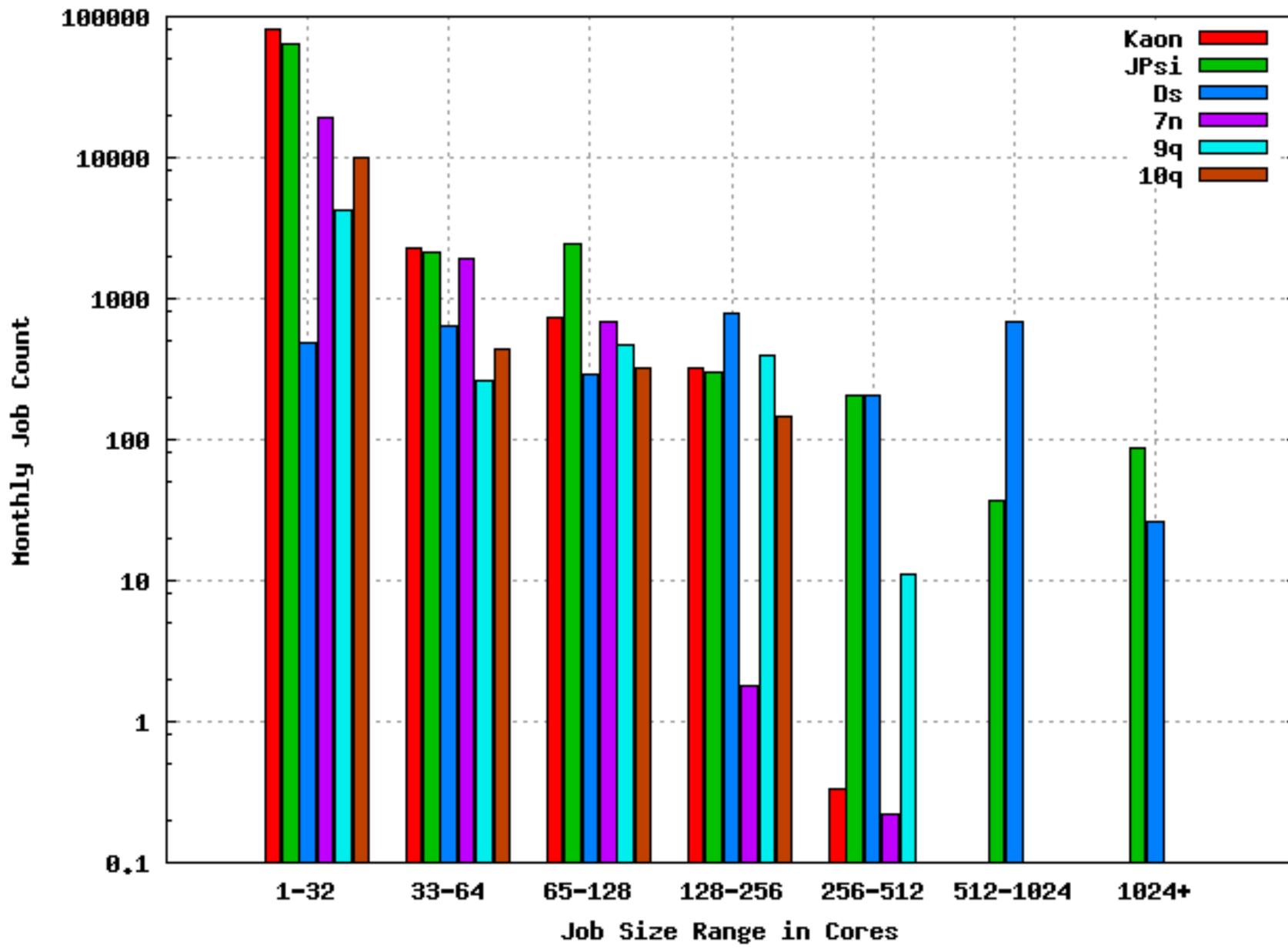
# GPU Cluster Plans

- Preliminary design:

  - 128 Tesla C2050 GPUs, two per host machine

  - Hosts will be dual socket, 8 cores/host, 24 GiB or 48 GiB host memory

  - QDR Infiniband

  - This design will allow running jobs of significant size (64 to 128 GPUs) with sufficient inter-node bandwidth to give reasonable strong scaling with cutting along more than just the time axis

  - GPUs with ECC will allow safe non-inverter calculations

- Possible variations (we need your advice):

  - 3 or 4 GPUs per host

  - Larger host memory and/or 4 socket hosts (32 to 48 cores)

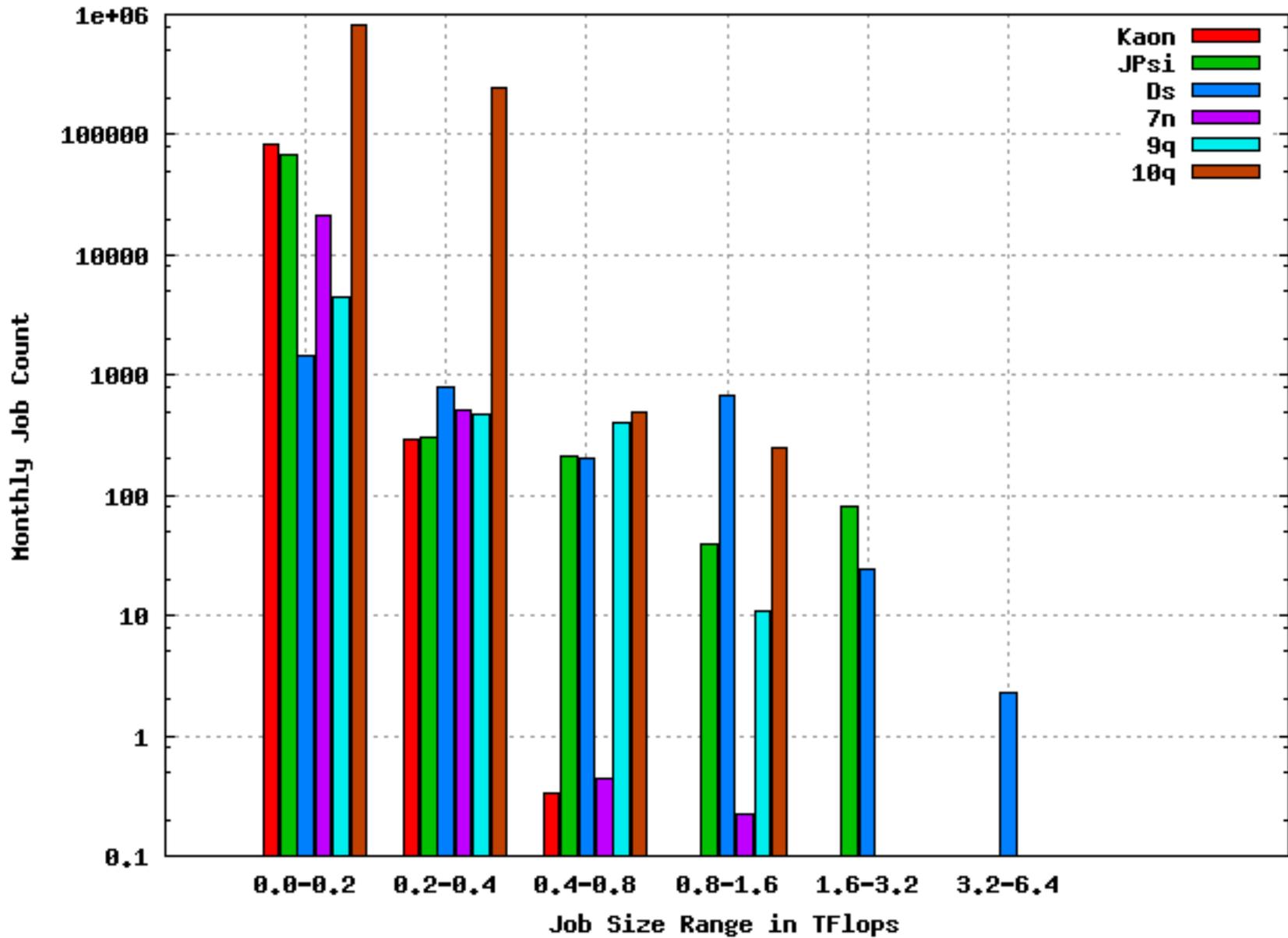  - 6 GiB GPU memory (C2070) instead of 3 GiB (C2050)

# Statistics

- Since April 1, 2010, including Kaon, JPsi, Ds (Dec→now)
  - 1,836,894 jobs
  - 10.2M node-hours → 94.8M JPsi-core-hours
  - We have not charged for Kaon since Oct 1 (6.6M JPsi-core-hours)

- Unique USQCD users submitting jobs:
  - FY10: 56
  - FY11 to date: 51

- Lustre filesystem (/lqcdproj)
  - 387 TiB capacity, 318 TiB used, 110 disk pools
  - 59.2M files
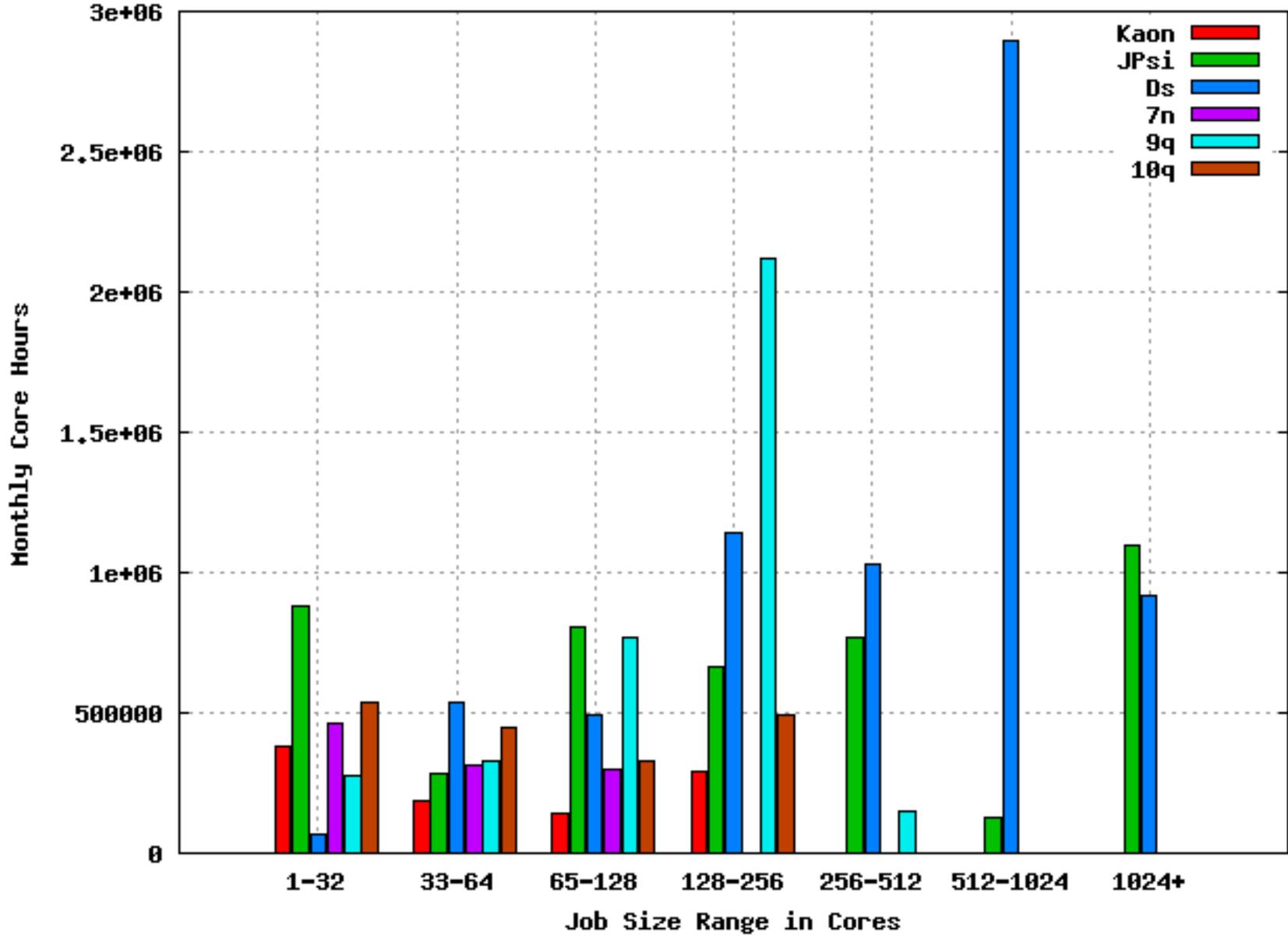  - File sizes: 210 GiB maximum, 5.64 MiB average

Fermilab LQCD Cluster Utilization FY10-FY11

USQCD 2011 AHM    Fermilab Report

USQCD 2011 AHM    Fermilab Report

USQCD 2011 Job Statistics, July 2010 - March 2011

USQCD 2011 Job Statistics, July 2010 - March 2011
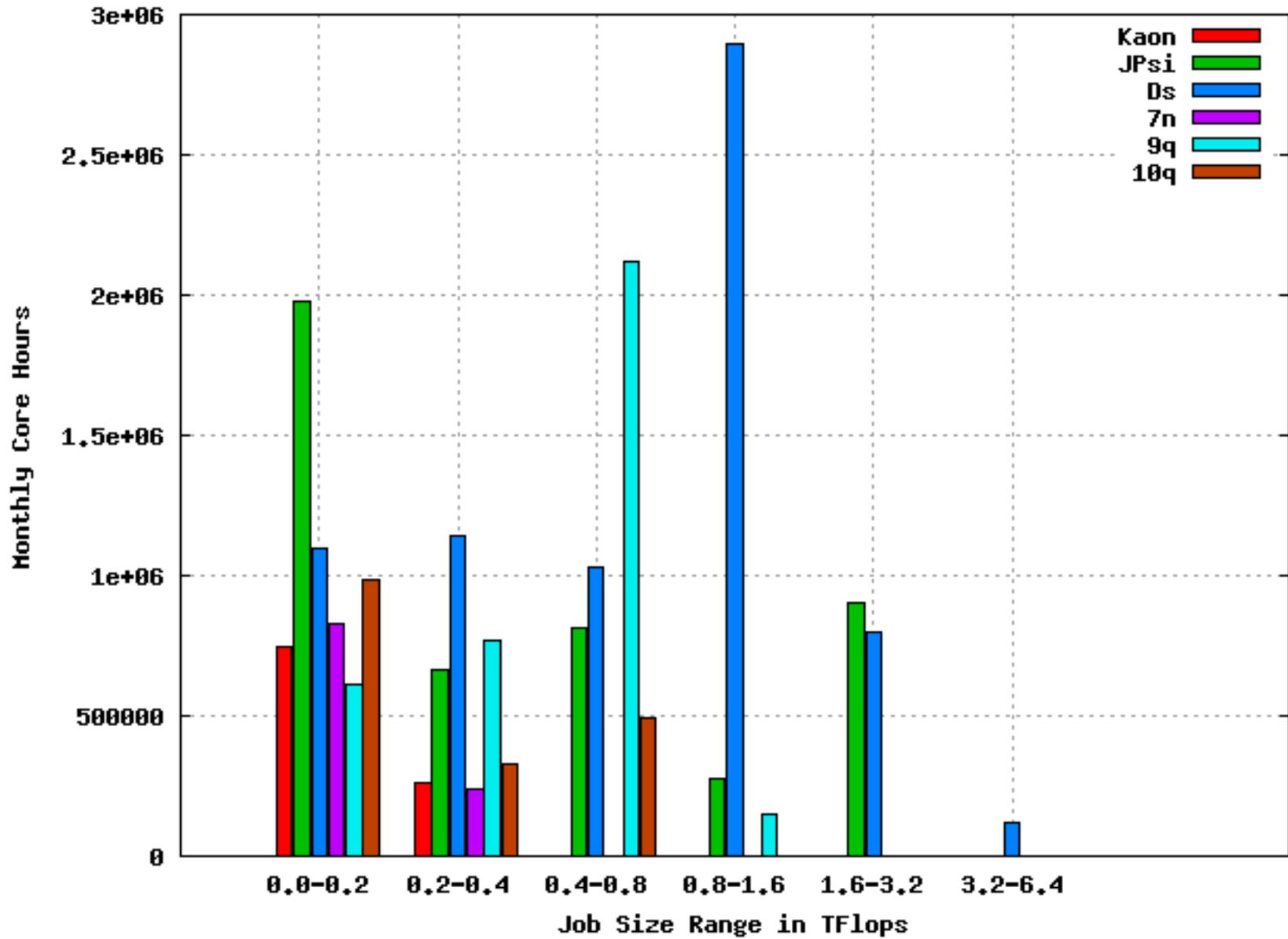
# Progress Against Allocations

- Total Fermilab allocation: 103.3M JPsi core-hrs

- Delivered to date: 80.4M  (77.6%, at 83.8% of the year)
  - Includes disk and tape utilization (2.14M)
  - Does not include 12.6M delivered without charge on Kaon
  - Does not include 4.5M delivered in November on Ds (friendly user period)

- Anticipated delivery through June 30:
  - 22.0M on JPsi and Ds
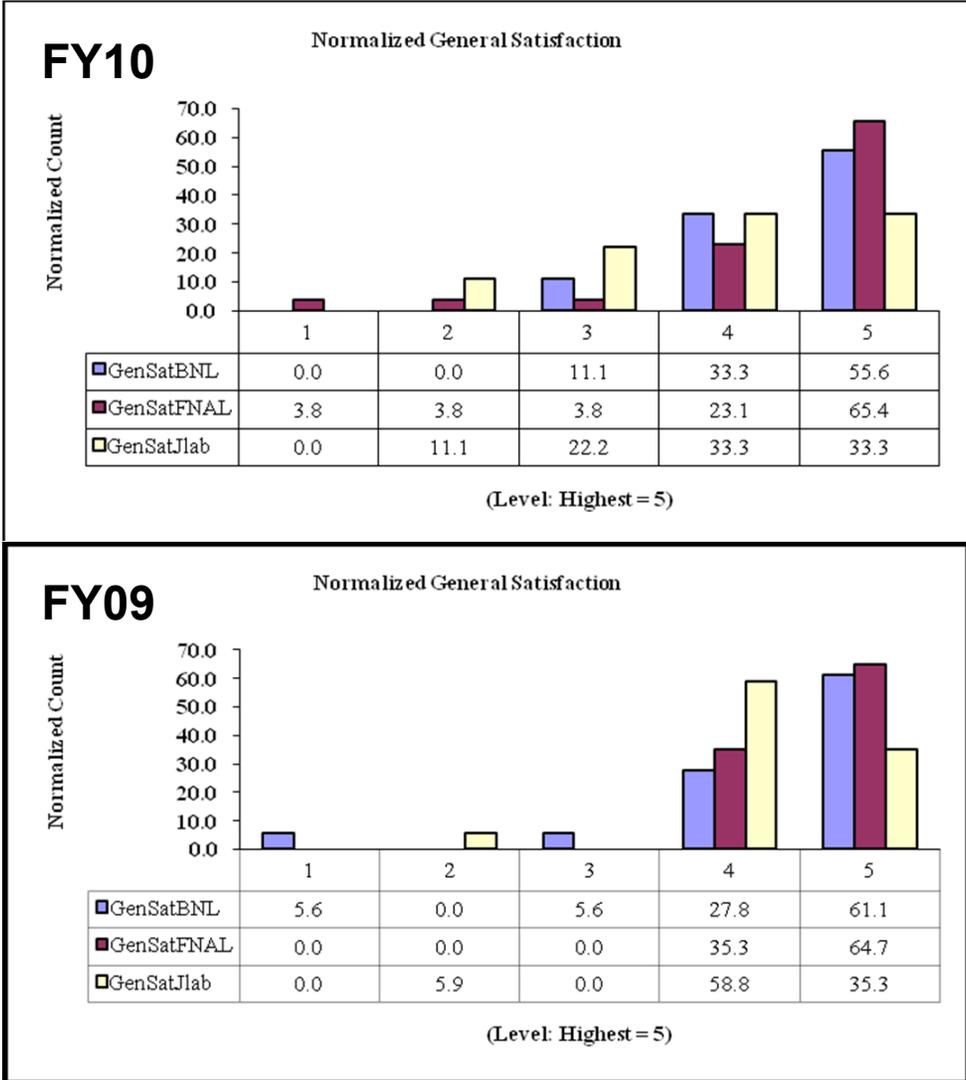  - 4.1M on the new Ds nodes (88 nodes starting June 1)

# Policies

- ## Directory permissions

  - By default not group writeable, but visible to group members *and* to non-group members.  We can help you restrict access to group members, and/or to add write access.

  - This applies to home areas, Lustre, and tape storage

- ## Access to batch queue information

  - We allow all users to see all queued jobs

  - We could restrict the view to only your jobs, but this would affect all users

- ## Web information

  - Not restricted

  - We could restrict with user authentication

# User Support

Fermilab points of contact:

- Best choice: lqcd-admin@fnal.gov

- Don Holmgren, djholm@fnal.gov

- Amitoj Singh, amitoj@fnal.gov

- Kurt Ruthmansdorfer, kurt@fnal.gov

- Nirmal Seenu, nirmal@fnal.gov

- Jim Simone, simone@fnal.gov

- Ken Schumacher, kschu@fnal.gov

- Rick van Conant, vanconant@fnal.gov

- Bob Forster, forster@fnal.gov

- Paul Mackenzie, mackenzie@fnal.gov

# User Survey



**FY10**

Normalized General Satisfaction

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| GenSatBNL | 0.0 | 0.0 | 11.1 | 33.3 | 55.6 |
| GenSatFNAL | 3.8 | 3.8 | 3.8 | 23.1 | 65.4 |
| GenSatJlab | 0.0 | 11.1 | 22.2 | 33.3 | 33.3 |

(Level: Highest = 5)

**FY09**

Normalized General Satisfaction

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| GenSatBNL | 5.6 | 0.0 | 5.6 | 27.8 | 61.1 |
| GenSatFNAL | 0.0 | 0.0 | 0.0 | 35.3 | 64.7 |
| GenSatJlab | 0.0 | 5.9 | 0.0 | 58.8 | 35.3 |

(Level: Highest = 5)

# Questions?