

LQCD-ext Technical Performance of FY14 Cluster Deployment

Amitoj Singh

Fermilab

amitoj@fnal.gov

SC LQCD-ext II Annual Progress Review
Brookhaven National Laboratory

May 21-22, 2015

Talk Outline

- Overview of SC LQCD-ext FY14 and FY15 acquisitions
- FY14 and FY15 cluster deployment and performance
- Questions

Overview of SC LQCD-ext Acquisitions

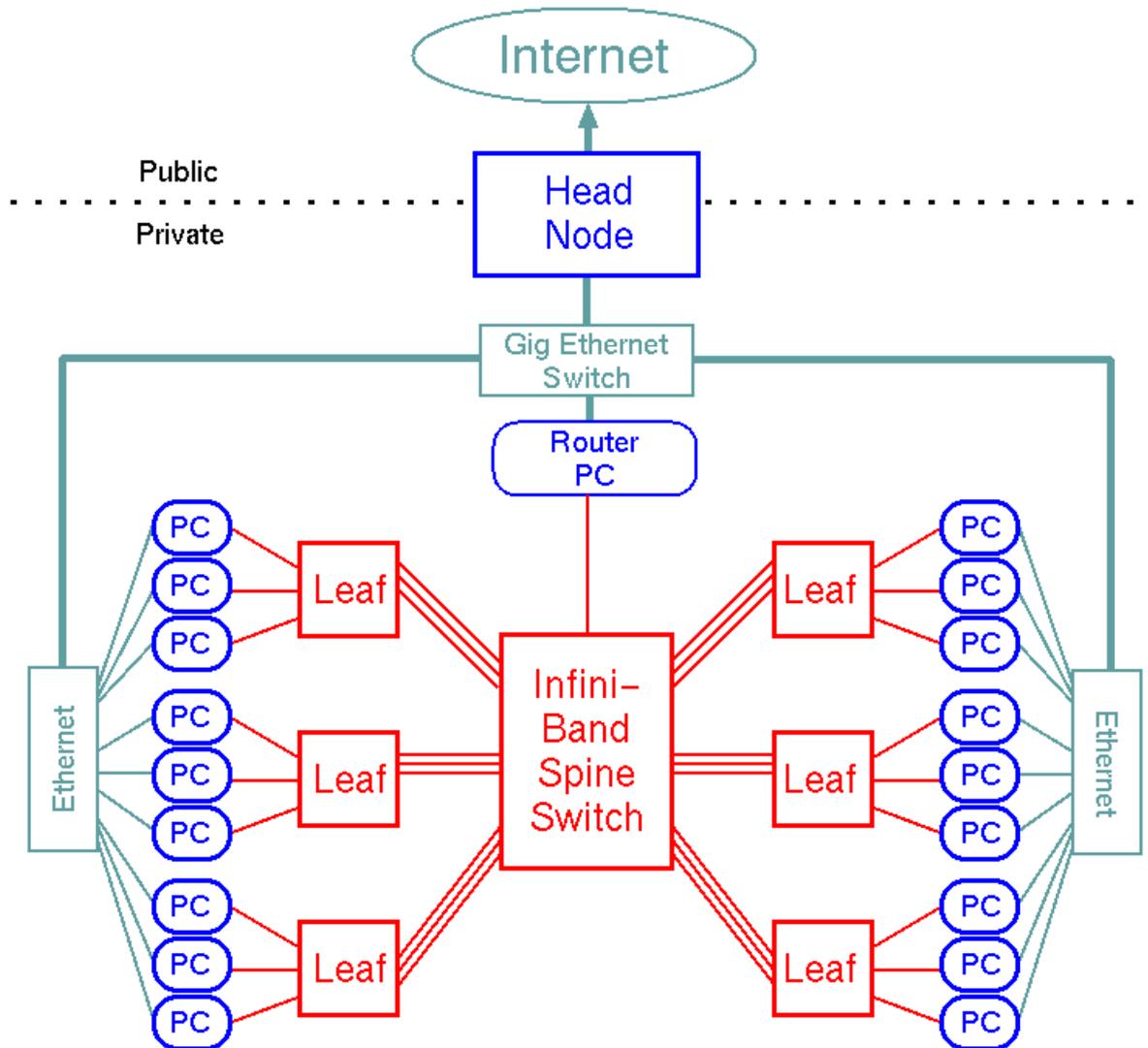
Computational capacity goals for SC LQCD-ext FY14 and FY15 acquisitions:

	FY2014	FY2015	TOTAL
Baseline computing hardware budget (not including storage) Planned / <i>Revised</i>	\$2.26M / <i>\$1.8M</i>	0 / <i>\$0.45M</i>	\$2.26M / <i>\$2.25M</i>
Capacity of new cluster deployments (Tflop/s) Planned / <i>Revised</i> / <i>Achieved</i> (conventional)	57 / <i>22-33</i> / <i>13</i>	0 / <i>0</i> / <i>6</i>	57 / <i>22-33</i> / <i>19</i>
Million “Fermi” GPU-hrs/yr Planned / <i>Revised</i> / <i>Achieved</i> (accelerated)	0 / <i>7.5-11.2</i> / <i>2.9</i>	0 / <i>0</i> / <i>0</i>	0 / <i>7.5-11.2</i> / <i>2.9</i>

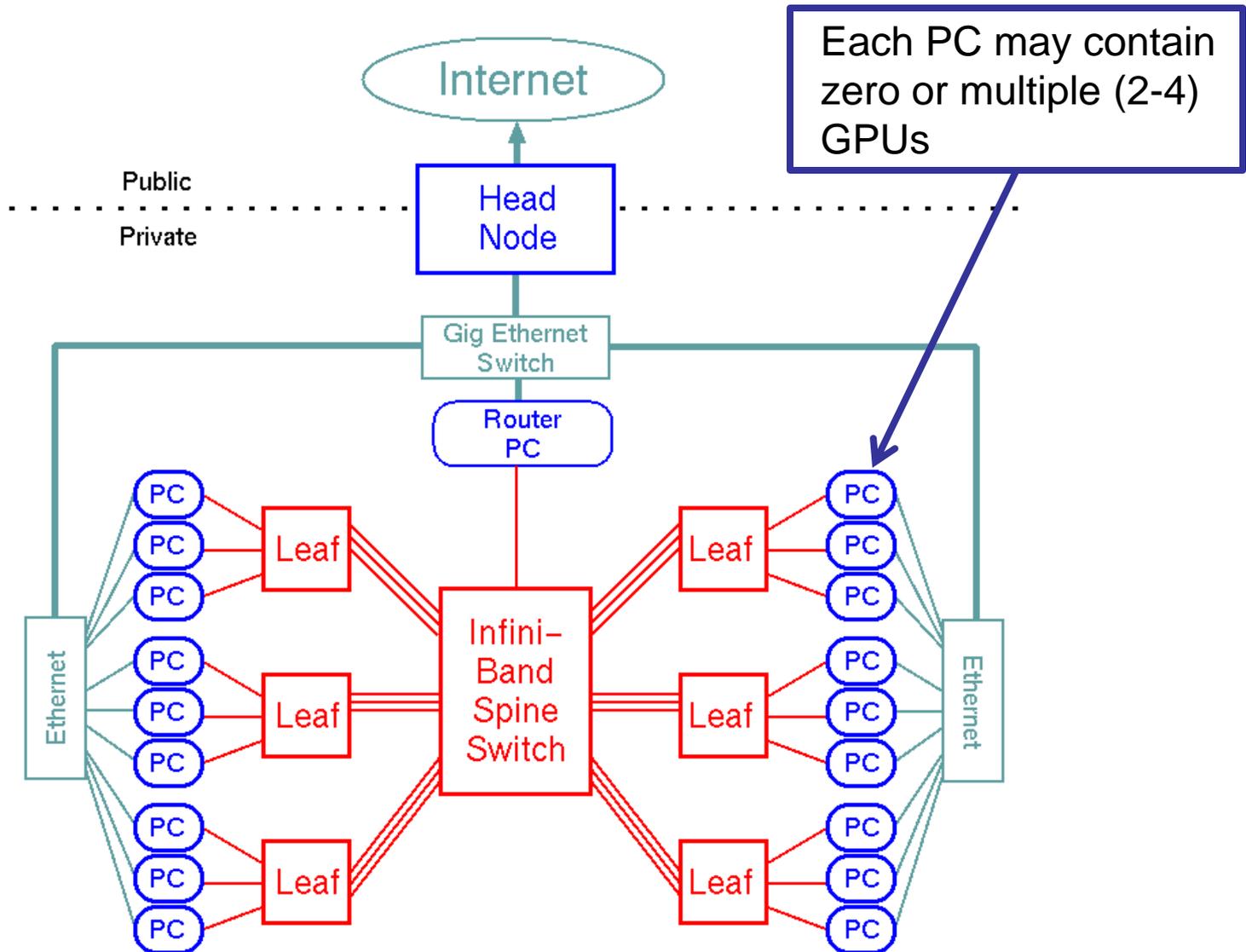
- FY2012-FY2014: revised goals to reflect 40%-60% ranges in budget allocated to conventional and accelerated clusters.
- FY2014: Ranges in **green** based on baseline budget of \$2.26M which was reduced to \$1.8M to defer funds due to uncertainty in LQCD-ext II operations budget. Project deployed pi0 “*conventional*” (**13 TF**) and pi0g “*accelerated*” (**2.9M “Fermi” GPU-hrs/yr**) clusters. GPU price/performance extrapolations from the FY2011 purchase using the observed Moore’s Law halving time for conventional hardware were too optimistic.
- FY2015: \$451K was used to expand pi0 “*conventional*” cluster with additional **6 TF**.

FY14 and FY15 Cluster Deployment and Performance

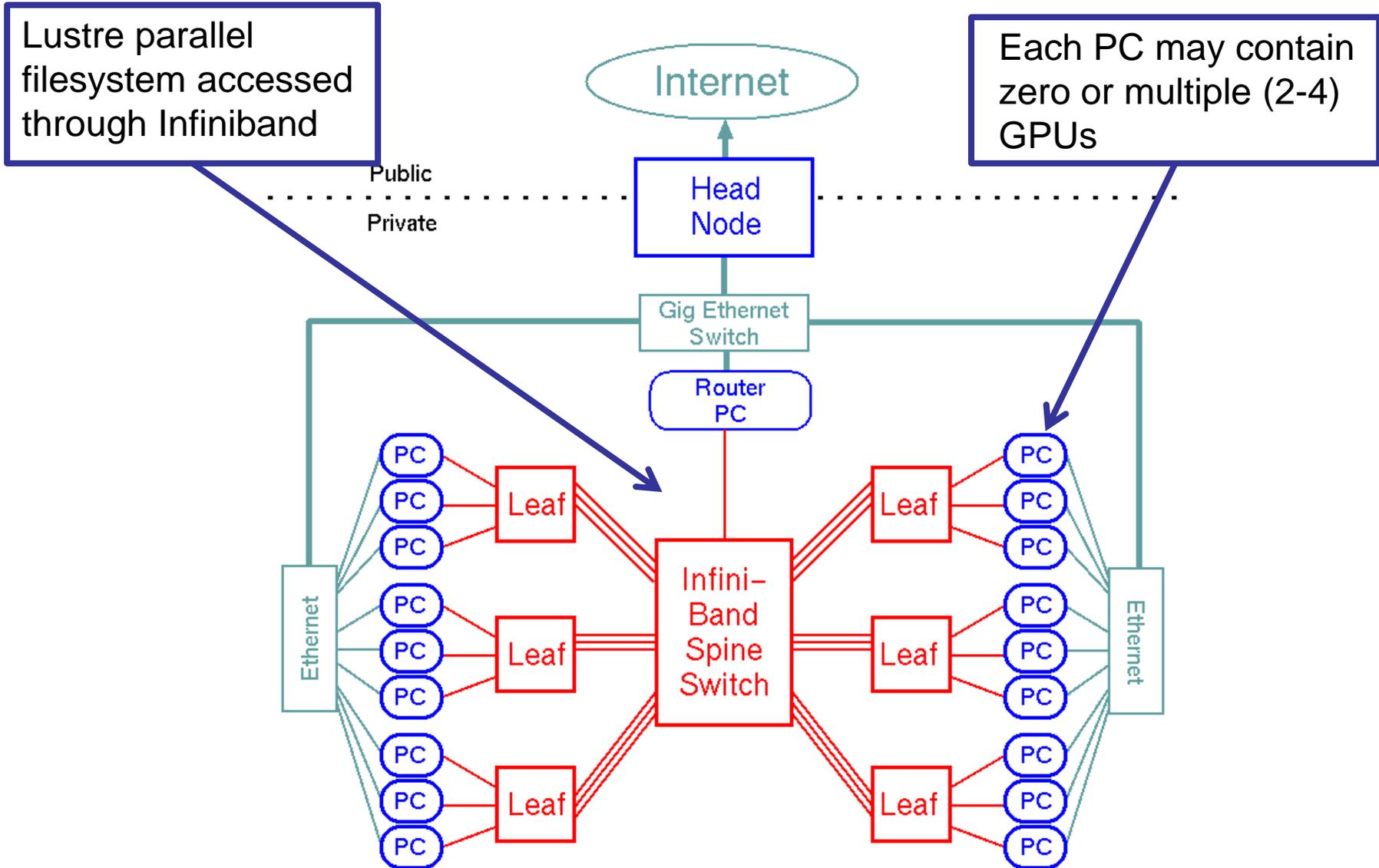
Typical LQCD Cluster Layout



Typical LQCD Cluster Layout



Typical LQCD Cluster Layout



FY14 pi0 “conventional” Details

- Award was to best value bid, based on price, LQCD application performance, power efficiency, space efficiency, vendor qualifications and past performance. Purchase contract included options to purchase additional hardware till Dec 31, 2014.
- Hardware details (FY2014):
 - Dual-socket eight-core Intel 2.6 GHz “Ivy Bridge” processors,
 - Upgraded to 128 GB memory per node to support *deflation* algorithm for the g-2 experiment and others,
 - QDR Infiniband with 2:1 oversubscription,
 - Upgraded to include 4th- and 5th- year warranties,
 - 216 worker nodes, 2 worker nodes were re-purposed as head nodes,
 - \$1.17M including G&A
- Performance
 - 214 worker nodes (3,424 cores)
 - Asqtad:DWF 61 Gflop/node (128-process MPI run)
 - 13 Tflop/s machine

FY15 pi0 “conventional” Details

- Award was to best value bid, based on price, LQCD application performance, power efficiency, space efficiency, vendor qualifications and past performance. Purchase contract included options to purchase additional hardware till June 30, 2015.
- Hardware details (FY2015):
 - Dual-socket eight-core Intel 2.6 GHz “Ivy Bridge” processors,
 - Upgraded to 128 GB memory per node to support *deflation* algorithm for the g-2 experiment and others,
 - QDR Infiniband with 2:1 oversubscription,
 - Upgraded to include 4th- and 5th- year warranties,
 - 100 worker nodes,
 - \$0.45M not subject to G&A
- Performance
 - 100 worker nodes (1,600 cores)
 - Asqtad:DWF 61 Gflop/node (128-process MPI run)
 - 6 Tflop/s machine

FY14 pi0g “accelerated” Details

- Award was to best value bid, based on price, LQCD application performance, power efficiency, space efficiency, vendor qualifications and past performance. Purchase contract included options to purchase additional hardware till Dec 31, 2014.
- Hardware details (FY2014): :
 - Dual-socket eight-core Intel 2.6 GHz “Ivy Bridge” processors,
 - 128 Gigabytes memory per node,
 - 4 NVIDIA K40m (Kepler Tesla) GPUs per node,
 - QDR Infiniband with no oversubscription,
 - 32 worker nodes and 128 GPUs,
 - Upgraded to include 4th- and 5th- year warranties,
 - \$0.78 M including G&A
- Performance
 - 128 GPUs (368,640 CUDA cores)
 - DWF:HISQ:Clover 10.4 GPU-hr/node-hr
 - 2.9M GPU-hrs/yr (26 effective TF)

pi0 & pi0g Details

- Vendors had the option to re-use existing racks and contract with on-site DELL Managed Services for the installation and support.
- Using space vacated by de-commissioning the FY2009 “*Jpsi*” cluster and re-used “*Jpsi*” racks.
- Added QDR Infiniband adapters to the existing Lustre servers to connect to the pi0 & pi0g Infiniband spine network fabric.



pi0 & pi0g FY14 Procurement Timeline

- Apr 11 RFP released to vendors
 - May 9 Proposal due date
 - May 28 Purchase Order to vendor
 - June 27 Purchase Order issued
 - Aug 6 Delivery of 98 pi0 and 32 pi0g nodes to Fermilab
 - Aug 7 Completion of integration. Acceptance test begins
 - Sep 15 Release to production
 - Sep 15 Delivery of remaining 118 pi0 nodes to Fermilab
 - Sep 17 Completion of integration. Acceptance test begins
 - Oct 8 Acceptance test completes for all equipment
 - Oct 10 Release to production
-
- The diagram uses colored arrows to show dependencies between milestones. Black arrows indicate planned dependencies: from 'Purchase Order to vendor' to 'Purchase Order issued', from 'Release to production' (Sep 15) to 'Purchase Order issued', and from 'Release to production' (Oct 10) to 'Purchase Order issued'. Blue arrows indicate achieved dependencies: from 'Delivery of 98 pi0 and 32 pi0g nodes to Fermilab' to 'Release to production' (Sep 15), from 'Completion of integration. Acceptance test begins' (Aug 7) to 'Release to production' (Sep 15), from 'Delivery of remaining 118 pi0 nodes to Fermilab' to 'Release to production' (Oct 10), and from 'Completion of integration. Acceptance test begins' (Sep 17) to 'Release to production' (Oct 10).

■ Achieved ■ Planned

pi0 & pi0g FY15 Procurement Timeline

- Feb 9 Purchase Order to vendor
- Mar 19 Delivery of 100 pi0 nodes to Fermilab
- Mar 20 Completion of integration. Acceptance test begins
- Apr 10 Release to production

Questions?