LQCD Computing at BNL

Robert Mawhinney
Columbia University

BGQ Computers at BNL

USQCD half-rack   2 racks of DD1   1 rack of DD2
(512 nodes)   RBRC   BNL

USQCD use of BNL DD2 BGQ

USQCD has 10% of the available time on the BNL DD2 BGQ (pre-production)

This time is included in the allocations by the SPC

During this allocation year, PI Chris Kelly has run some of his SPC allocated time on 512 nodes of DD2, to use the USQCD 10%.

DD2 rack is running very well. Used extensively by BNL internal users.

USQCD 512 Node BGQ at BNL

USQCD 512 Node BGQ at BNL

Purchased with $1.32 M from USQCD with FY13 Equipment Funds

Delivered in March, 2013, first users (Chulwoo) on Monday, April 15, 2013

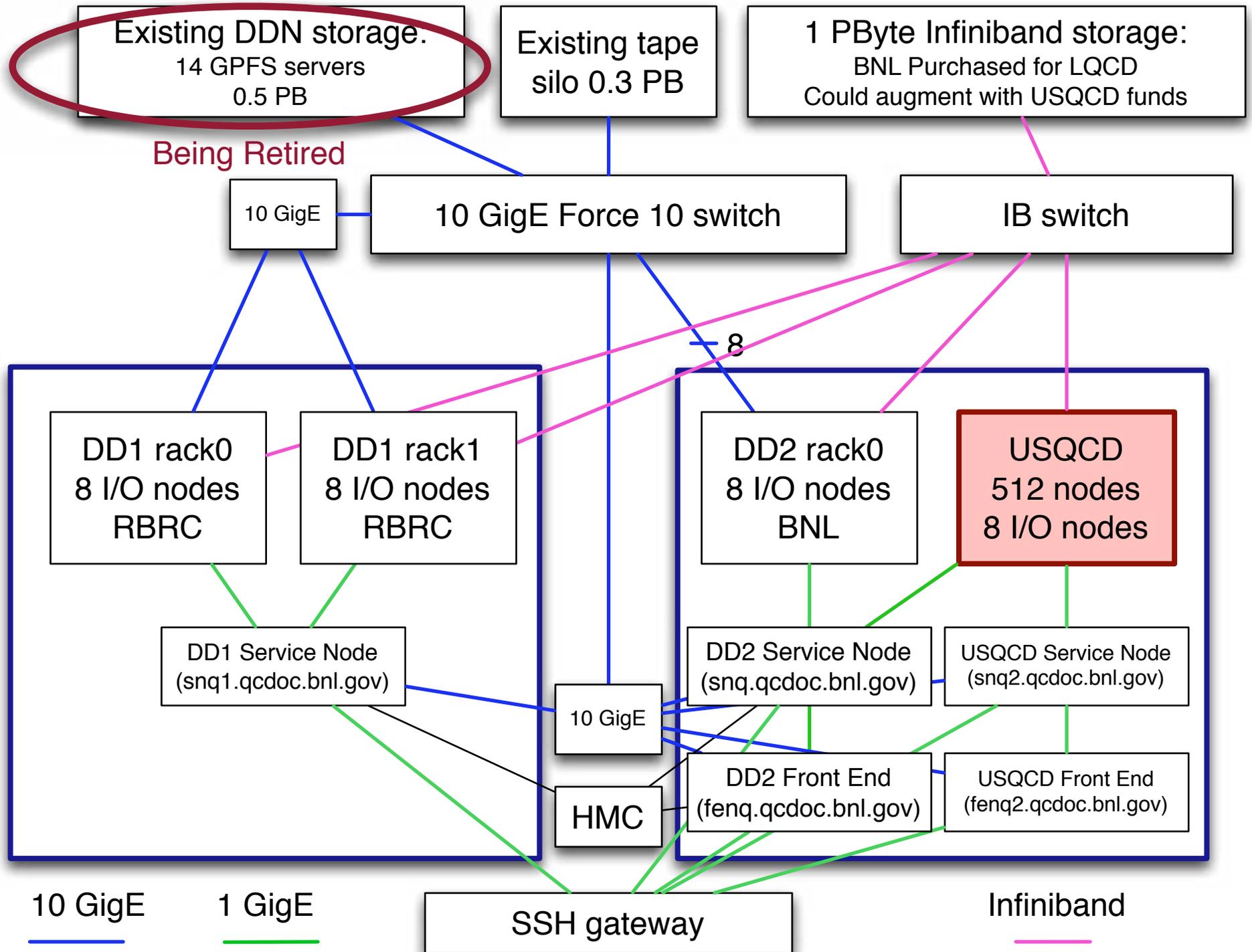USQCD SPC allocated time for 3 projects in 2013-2014

P.I.	Allocated	Used	% Used
Kelly	44.60	48.55	109%
Mackenzie	18.65	22.48	108%
Sugar	7.55	5.71	

Sugar ran early in the allocation year, and once it was clear that extra time was available, it was not convenient to restart those runs. Extra time given to Mackenzie.

USQCD SPC allocated time for 3 projects in 2013-2014. Usage as of May 1, 2015.

P.I.	Allocated	Used	% Used	Max Usage	Max % Usage
Kelly	42.03	47.65	114%	47.65	114%
Kuti	15.42	6.80	44%	17.01	110%
Mackenzie	13.35	12.95	97%	14.72	110%

A maximum of 11.99 M BGQ core hours are available by June 30, 2015

Existing DDN storage:
14 GPFS servers
0.5 PB

Being Retired

Existing tape silo 0.3 PB

1 PByte Infiniband storage:
BNL Purchased for LQCD
Could augment with USQCD funds

10 GigE

10 GigE Force 10 switch

IB switch

8

DD1 rack0
8 I/O nodes
RBRC

DD1 rack1
8 I/O nodes
RBRC

DD2 rack0
8 I/O nodes
BNL

USQCD
512 nodes
8 I/O nodes

DD1 Service Node
(snq1.qcdoc.bnl.gov)

10 GigE

DD2 Service Node
(snq.qcdoc.bnl.gov)

USQCD Service Node
(snq2.qcdoc.bnl.gov)

HMC

DD2 Front End
(fenq.qcdoc.bnl.gov)

USQCD Front End
(fenq2.qcdoc.bnl.gov)

10 GigE   1 GigE

SSH gateway

Infiniband

USQCD BGQ Utilization at BNL 2013-2014

2013-2014 allocation month	Utilization	Comments
July	48%	Faulty compute node, IBM slow to diagnose. No hardware problems from March-June.
August	79%	2 day chilled water outage
September	90%	
October	91%	
November	83%	3 days lost to hardware failure
December	95%	
January	91%	Loadleveler hang
February	99%	
March	95.8%	Legacy file system failure caused brief outage.
April	91.6%	Brief outage to clean filter. I/O drawer software error.
May	98.4%	
June	84.8%	5% of time was lost due to legacy file system problem.

Utilization reported here is the fraction of the time jobs were running divided by the maximum hours available in the month, with no derating

Almost all usage has been a single user running on 512 nodes full time.

USQCD BGQ Utilization at BNL 2014-2015

2014-2015 allocation month	Utilization	Comments
July	90.8%	2% of downtime due to thunderstorms at BNL
August	87.7%	Most of downtime came when single user had to fix a code problem.
September	83.7%	10% of downtime from clogged filter and slow restart of user jobs
October	94.0%	
November	98.1%	
December	92.9%	
January	99.98%	
February	99.8%	
March	80.8%	Scheduled software upgrade, followed by a hardware failure requiring new parts.
April	82.7%	
May		
June		

Utilization reported here is the fraction of the time jobs were running divided by the maximum hours available in the month, with no derating

Conclusions and Outlook

USQCD half-rack is supported by a total of 0.5 FTE at BNL. Cost effectiveness of computing increased by low personnel costs.

USQCD pays IBM for a service contract.

Currently, have not found a way to acquire inexpensive parts to fill up the rest of the BGQ half-rack.

Interest in proposing a USQCD funded Intel Knight's landing based machine for BNL in 2017 fiscal year.

*   Can a tuned communication network balance the KNL local floating point, to produce a more balanced QCD machine for large node-count jobs?

Summary

First year of USQCD BGQ running on track to deliver allocated computing time

Limited number of users - important that they be ready to run to keep machine full.

Cost neutral options for near term doubling of compute power

BNL has retired NY Blue, an IBM BG/L system.

* Lab is engaged in seeking a replacement system - likely a phi cluster

* Possibility for USQCD to augment such a system - more phi boards or next generation accelerators.