Technical Design and Proposed Scope for FY09

Don Holmgren

Fermilab

djholm@fnal.gov

LQCD Project Progress Review

Fermi National Accelerator Laboratory

June 4-5, 2009

Introduction

SC LQCD is 4 months from ending (Sep 30)

- All major purchases for FY09 are complete
- In past reviews this talk would have focused on technical design issues and the details of the plans for the upcoming machine purchases
- This year's talk will instead focus on the FY08/FY09 procurement just completed
- Metrics from operations at BNL, TJNAF, and FNAL were given in Bill Boroski's previous talk

Outline

- Hardware requirements imposed by LQCD codes
- SC LQCD project clusters
- Recap from last year's review
- The FY08/FY09 (J/Psi) cluster procurement
- Remaining FY09 technical issues

Hardware Requirements

- Characteristics of production LQCD codes:
  - Computations are dominated by SU(3) algebra (small complex matrices and vectors)
  - High ratio of bytes read/written to FLOPs
    - Single precision complex matrix (3x3) – vector (3x1): 96 bytes read, 24 bytes written, 66 FLOPs → 1.8:1
  - Caches are generally too small to support significant reuse
  - Inter-node communications for message passing require roughly 1 Gbit/sec of bandwidth for each GFLOP/sec of node capability
    - Also, low latency is required for efficient global reductions, and for good strong scaling

Hardware Requirements

- Either memory bandwidth, floating point performance, or network performance (bandwidth at message sizes used) will be the limit on performance on a given parallel machine

- On current single commodity nodes memory bandwidth is the constraint

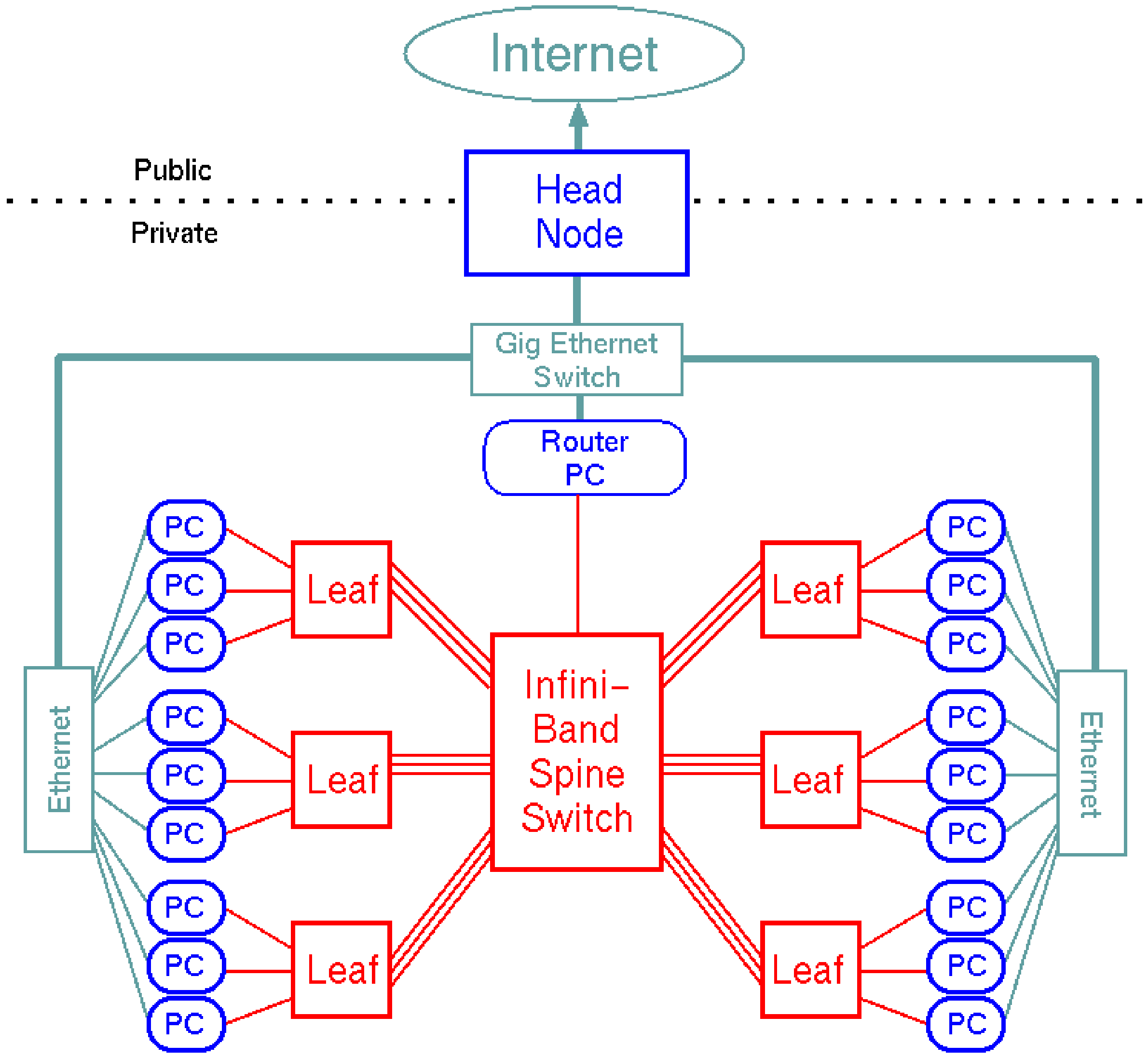
- On current parallel computer clusters, the constraint is either memory bandwidth or network performance, depending upon how many nodes are used on a given job

  – Network performance limits scaling:
  Surface area to volume ratio increases as more nodes are used, causing relatively more communications and smaller messages

Hardware Requirements

- We design and buy clusters with the best LQCD price/performance
- This means:
  - Machines with the best per core memory bandwidth
  - Machines with modest memory size (1 GB/core)
  - High performance interconnects
    (Infiniband now, Myrinet, Gigabit Ethernet meshes previously)

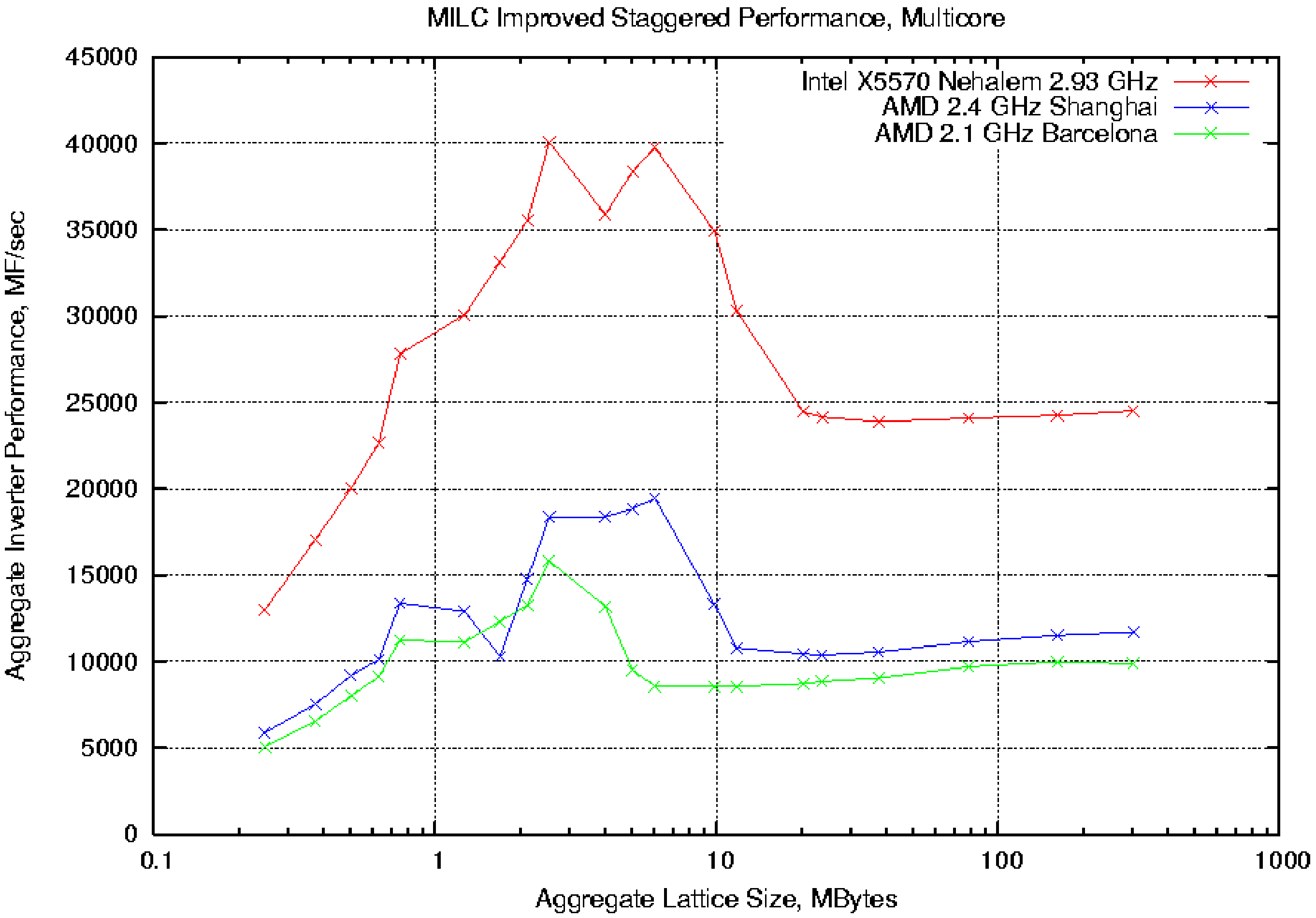
Typical LQCD Cluster Layout



LQCD Project Clusters

Name (FNAL/TJNAF)	6N	Kaon	7N	J/Psi
Speed Processor (Socket Count) (Cores/CPU)	3.0 GHz Pentium D (256) (2)	2.0 GHz Opteron (1200) (2)	1.9 GHz Opteron (792) (4)	2.1 GHz Opteron (1712+16) (4)
Memory Speed	DDR-400	DDR-533	DDR2-667	DDR2-800
Single or Dual Socket	Single	Dual	Dual	Dual
Interconnect Fabric	Infiniband (SDR)	Infiniband (DDR)	Infiniband (DDR)	Infiniband (DDR)
Performance	0.6 Tflop/s asqtad:DWF	2.6 Tflop/s asqtad:DWF	3.0 Tflop/s asqtad:DWF	8.4 Tflop/s asqtad:DWF
Date in Production	3/2006	10/2006	6/2007 (Upgrade 11/2007)	1/2009 (FY08) 4/2009 (FY09)

Performance: Single Node, Using All Available Cores on LQCD Code



Plots show the aggregate performance of eight MPI processes, one on each core

Intel Quad Core Dual Socket (“Nehalem”)

AMD Quad Core Dual Socket (“Shanghai”)

AMD Quad Core Dual Socket (“Barcelona”)

Recap from Last Year's Review

- Combining the FY08 and FY09 acquisitions would yield greater physics production over the first three years of system operation
  - Relative to FY08, the FY09 hardware budget is small ($798K vs. $1,630K)
  - A larger, single, homogenous system has advantages over two systems based on potentially different hardware
  - Procurement requires manpower, both on-project and in-kind
  - Procuring FY09 hardware earlier allows more integrated physics production, even though the capacity might be smaller
  - Goal for FY08-FY09 combined capacity: 6.2 TFlop/s

Recap from Last Year's Review

- Proposed strategy *(FY08 revisions highlighted):*
  - Use a single RFP to solicit bids for FY08 spending
  - Include an option in the purchase contract to buy additional identical hardware in FY09
  - Exercise FY09 option as soon as funds are available *if this maximizes physics production*
  - *If alternate hardware in FY09 (Intel Nehalem) would result in higher physics production, consider a separate new procurement*
    - *The FY09 hardware would be added to the FY08 Infiniband fabric (that is, still a single cluster but heterogeneous)*
    - *Users would submit jobs to different queues corresponding to the FY08 and FY09 hardware sets*

Recap from Last Year's Review

- From the review report:
  - Finding/Comment:
    - *"The strategy of evaluating and possibly acquiring Nehalem hardware in FY 2008-2009 is sound. However, the uncertainty and risk in meeting project milestones are considerable.*
  - Recommendation:
    - *"The schedule contingency and risk associated with the uncertainty in the availability of the Nehalem technology should be clarified."*

The J/Psi Procurement

- At the time of the RFP, we knew:
  - Intel Nehalem would not be available until at least early 2009 (by October we knew the delay was until April 2009)
  - Therefore in FY08 we would buy an AMD Opteron cluster with a goal of 4.2 TF capacity
  - We would need to decide whether to buy Nehalems in FY09, or expand with additional Opterons, with a goal of adding 2.0 TF capacity
  - We might also have to decide how to spend any FY08 funds if 4.2 TF capacity didn’t exhaust the budget: buy more Opterons, wait for Nehalems in FY09, or wait and buy more Opterons in FY09
- Important factors:
  - Nehalem would need a separate RFP (more manpower, delays)
  - A new RFP would result in G&A that we would otherwise avoid

The J/Psi Procurement

2008

- Mar 14 – RFI released to vendors
- Apr 15 – RFI responses received from vendors
- May 13,14 – DOE FY08 Progress Review
- July 11 – RFP released to vendors
- Aug 11 – RFP bids received from vendors
- Aug 21 – RFP award recommendation to purchasing department
- Aug 26 – Purchase order to vendor (commit FY08 funds)
- Nov 5 – Delivery of FY08 equipment complete
- Nov 26 – Friendly user period begins
- Jan 5, 2009 – Release to production FY08 portion (5.75 TFlops)

The J/Psi Procurement

2009

- Late Aug 2008 – “Nehalem” cluster benchmarking
- Oct 15 2008 – Decision to exercise J/Psi expansion option
- Feb 26 – Exercise purchase order option (commit FY09 funds)
- Apr 9 – Delivery of FY09 equipment complete
- Apr 15 – Release to production FY09 portion (2.65 TFlops)

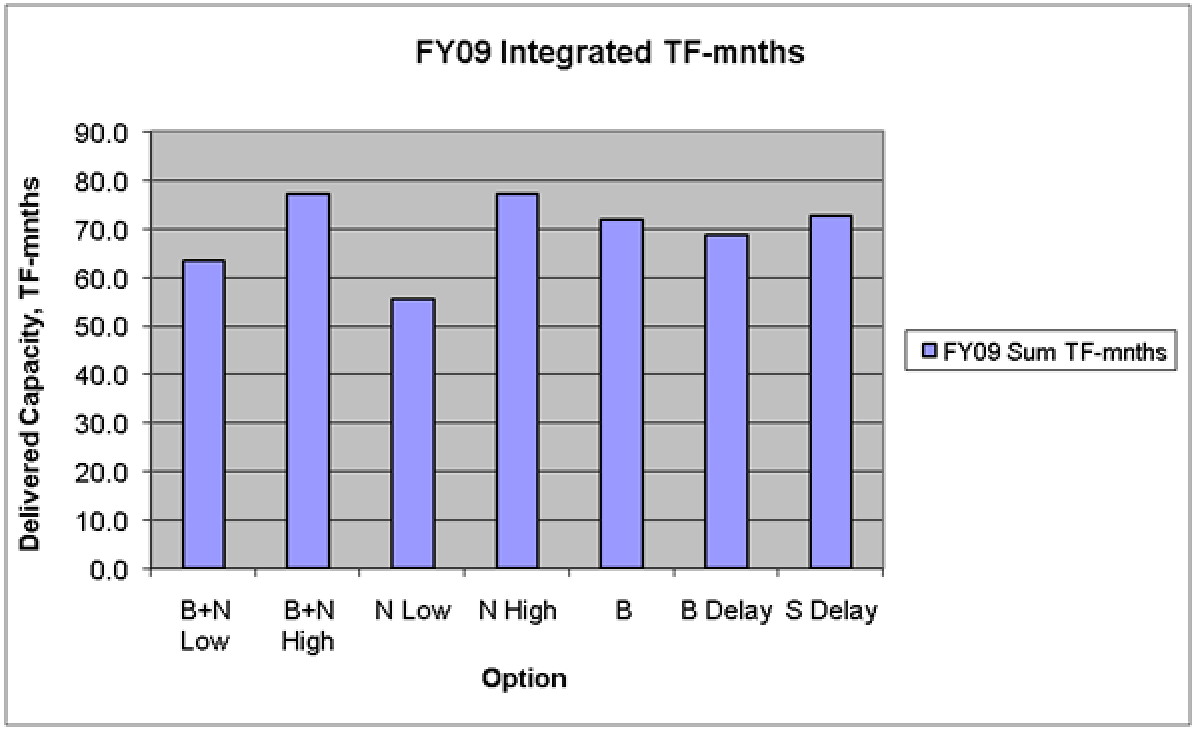
J/Psi Procurement – Details

- RFP
  - Vendors were asked to provide a design and pricing for 4.2 TF, as well as pricing for an option for additional racks, valid through March 2009 (option could be exercised multiple times)
  - Clusters were to be x86_64 hardware using Infiniband (SiCortex was previously excluded because of risk factors)
  - Best value process: Proposals were scored on:
    - Price/performance (60% weighting)
    - Power for 4.2 TF (10%)
    - Footprint for 4.2 TF (10%)
    - Infiniband reuse and reliability (10%)
    - Prior FNAL and LQCD experience with vendor (10%)
    - Preferred components and miscellaneous (5%)

J/Psi Procurement – Details

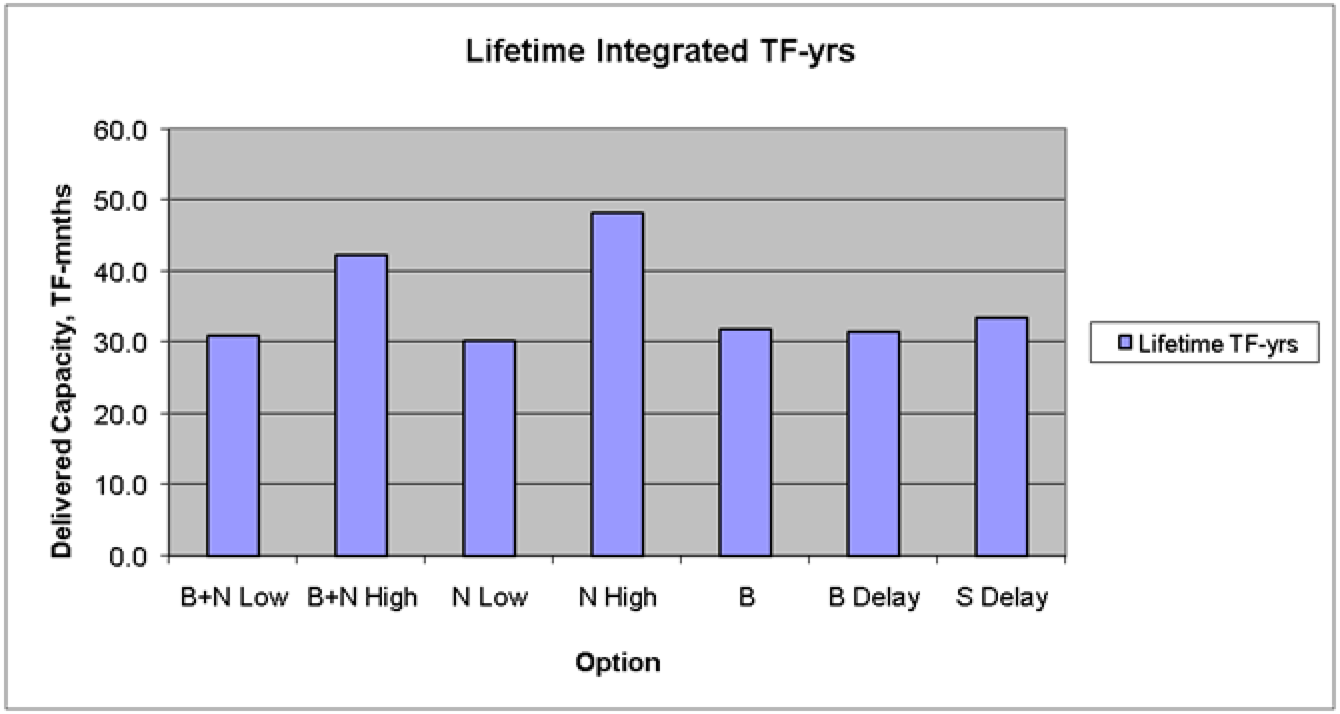
- Vendor information:
  - Solicitation sent to 16 vendors
  - Received 9 proposals from 6 vendors (4 white box, 2 OEM)
  - All proposals used AMD Opteron chips, 7 “Barcelona” and 2 “Shanghai” (with later delivery)
- Winning proposal
  - FY08 funds (Corrected for G&A of $80K): $1.452M
  - Proposal: $0.989M for 4.2 TF, plus $89.6K per option rack (up to 5.9 TF additional, roughly 400 GF per rack)
  - $0.455M left over after the 4.2 TF purchase → how to spend?
    - We examined Nehalem, Shanghai, Barcelona scenarios
    - Conclusion: purchase additional Barcelona racks

J/Psi Procurement – FY08 Scenarios



- "B" – Buy additional Barcelona with FY08 surplus immediately, and buy Barcelona in FY09
- "N Low/High" – based on plausible low/high Flop/$ for Nehalem, assumed production start in June 2009. The FY08 surplus plus all of FY09 to be spent on Nehalem
- "B+N": Spend FY08 surplus immediately on additional Barcelona, spend FY09 funds on Nehalem
- "B/S Delay": delay spending the FY08 surplus, but buy Barcelona or Shanghai with surplus + FY09
- Effects of added G&A costs included for all Nehalem options
- To optimize FY09 TF-months, the best choice was to spend FY08 surplus immediately on additional Barcelona.

J/Psi Procurement – FY08 Scenarios



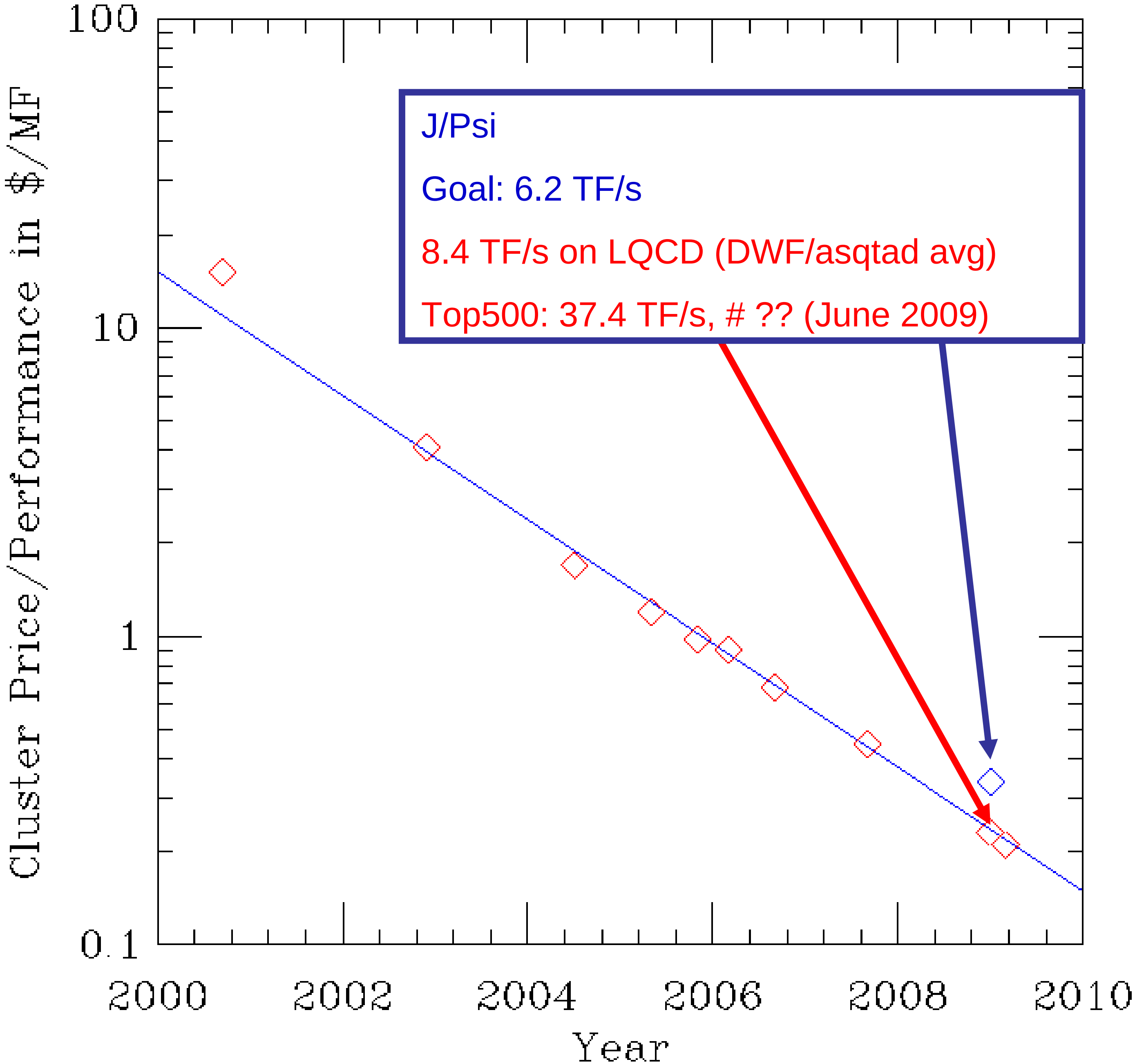
- Over the 3.5 year lifetime of the cluster, the choice was less clear
- Nehalem and Shanghai options both had considerable schedule risk, Nehalem had considerable cost uncertainty and considerable technical risk, and Shanghai had moderate technical risk. Any delayed spending of FY08 surplus had schedule risk because of the anticipated continuing resolution.
- We chose to spend the FY08 surplus on additional Barcelona (1.55 TF) plus storage plus a small holdover to FY09

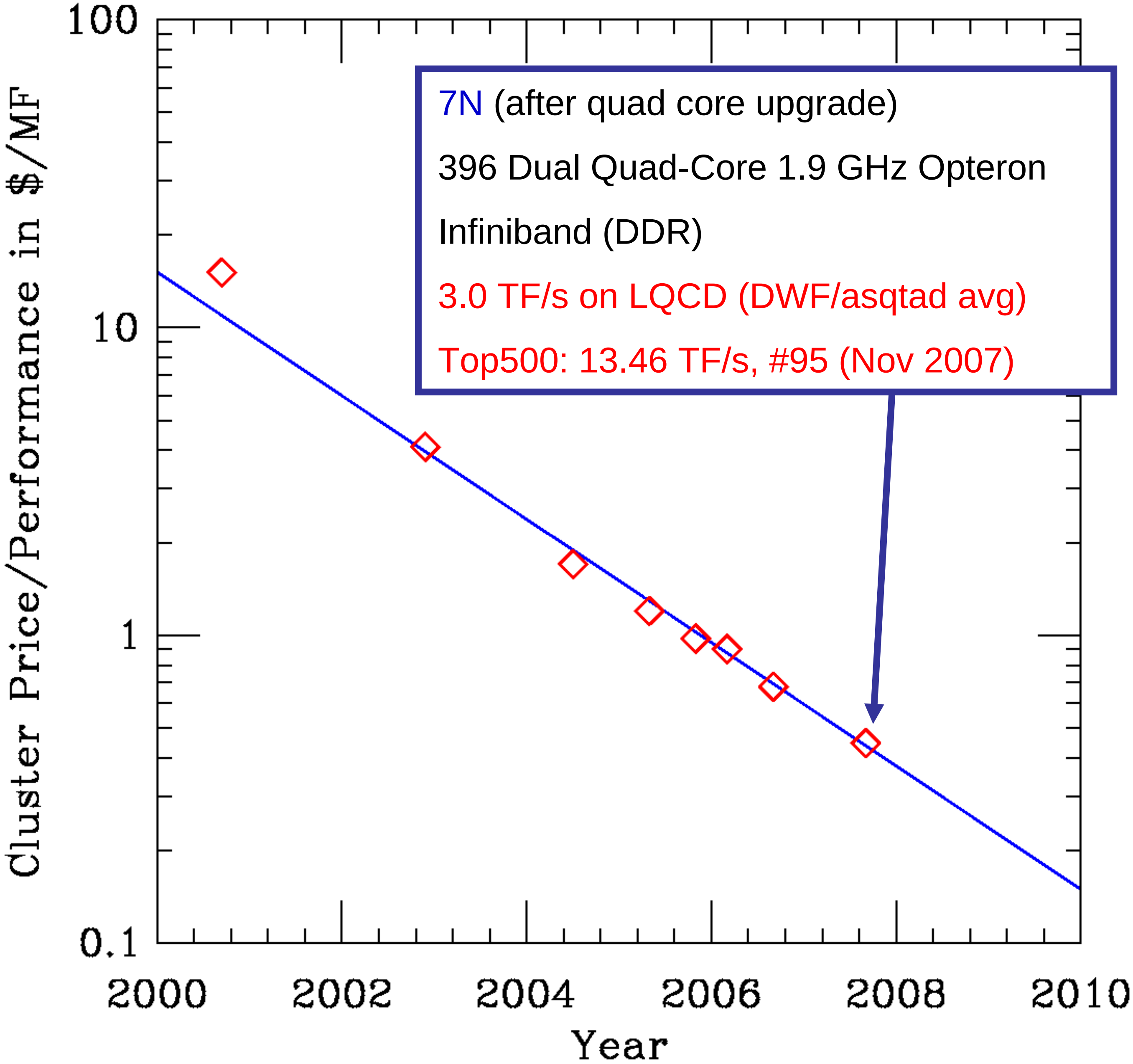
J/Psi Procurement – FY09 Purchase

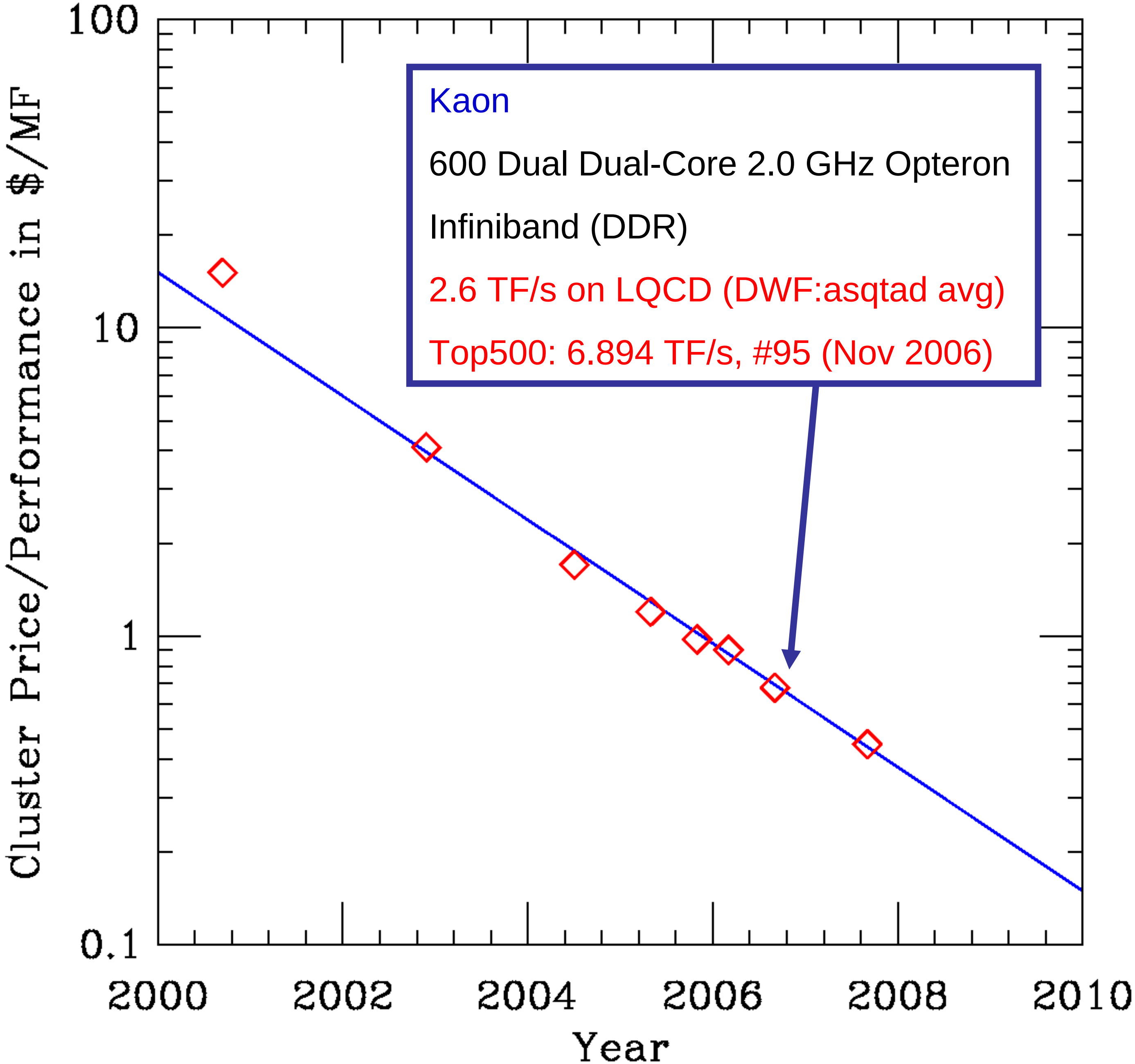
- By mid-October 2008, we had additional information:
  - We had benchmarked two actual Infiniband-based Nehalem clusters, and so had updated performance and scaling figures
  - We also learned that dual-socket Nehalem release would be delayed until at least the end of March
  - Our Nehalem scenarios had assumed production by beginning of June → very difficult to achieve this with the new Nehalem release date
  - Benchmarking resulted in no quantitative changes to our earlier analysis
  - Risks identified earlier remained, with increased Nehalem schedule risk
- Purchase decision:
  - Use FY08 holdover + FY09 funds to buy as many additional racks of Barcelona nodes as would fit on our Infiniband network
  - Remaining FY09 funds to be spent on small GPU cluster, and on additional storage

J/Psi Projected and Achieved Performance

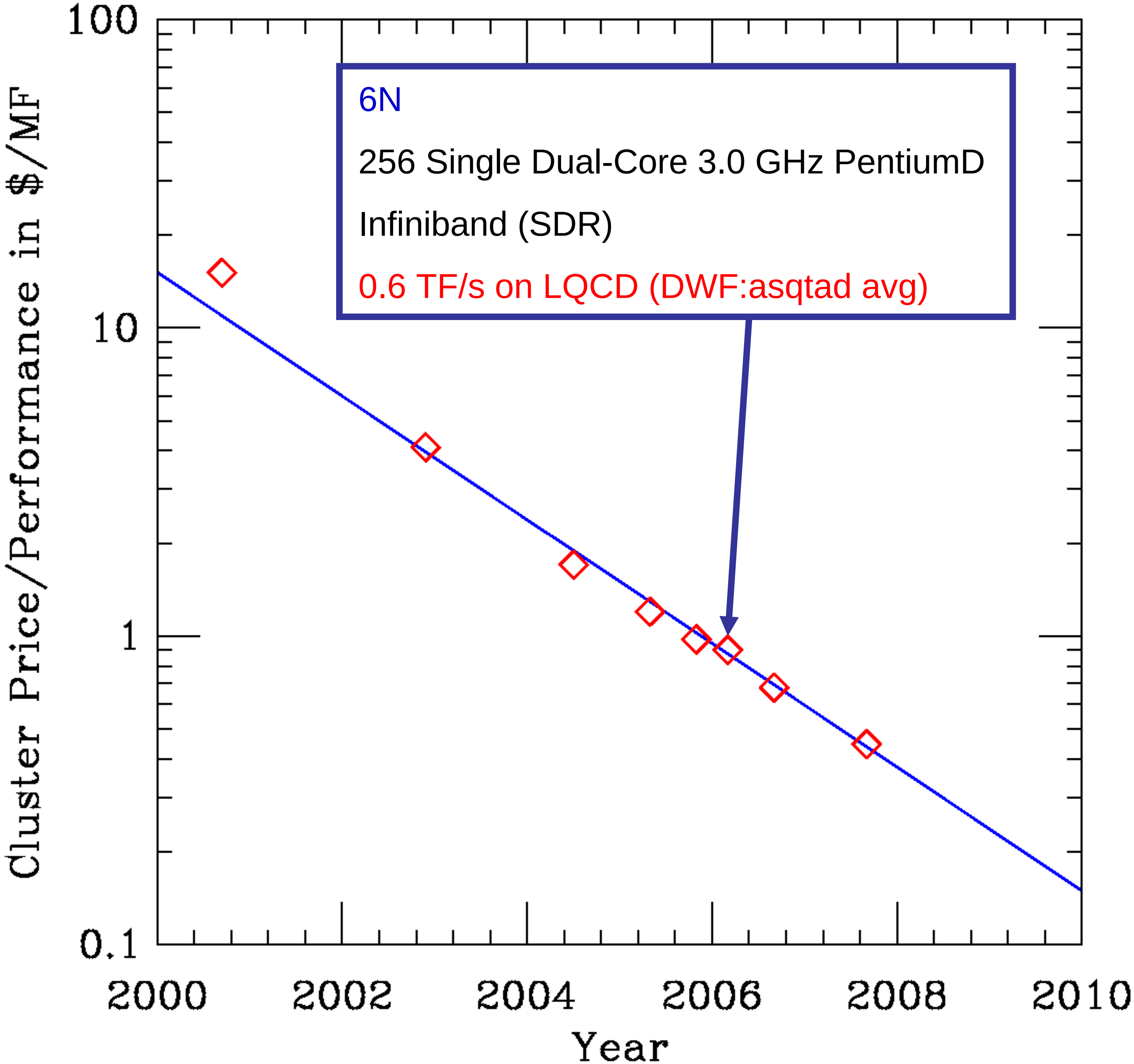
- MILC asqtad single node:
  - 2.0 - 2.3 GHz: 9020 - 9450 MFlops/node
  - 2.1 GHz: 9700 MFlops/node
- DWF single node:
  - 2.0 - 2.3 GHz: 10910 - 11820 MFlops/node
  - 2.1 GHz: 12000 MFlops/node
- Asqtad:DWF single node average
  - 2.0 - 2.3 GHz: 9965 – 10635 MFlops/node
  - 2.1 GHz: 10850 MFlops/node
- Scaling: Factor of 0.92 (8 cores → 64 cores) (0.90)
  - 2.0 - 2.3 GHz cluster: 9170 - 9785 MFlops/node
  - 2.1 GHz cluster: 9812 MFlops/node
- Vendor's unusual choice of faster memory (DDR2-800) and aggressive pricing resulting in exceeding our goal (8.40 TF total vs 6.2 TF total)


100
10
1
0.1
Cluster Price/Performance in $/MF
2000
2002
2004
2006
2008
2010
Year
J/Psi
Goal: 6.2 TF/s
8.4 TF/s on LQCD (DWF/asqtad avg)
Top500: 37.4 TF/s, # ?? (June 2009)


7N (after quad core upgrade)
396 Dual Quad-Core 1.9 GHz Opteron
Infiniband (DDR)
3.0 TF/s on LQCD (DWF/asqtad avg)
Top500: 13.46 TF/s, #95 (Nov 2007)
Cluster Price/Performance in $/MF
100
10
1
0.1
2000
2002
2004
2006
2008
2010
Year



Kaon
600 Dual Dual-Core 2.0 GHz Opteron
Infiniband (DDR)
2.6 TF/s on LQCD (DWF:asqtad avg)
Top500: 6.894 TF/s, #95 (Nov 2006)
Cluster Price/Performance in $/MF
100
10
1
0.1
2000
2002
2004
2006
2008
2010
Year


6N
256 Single Dual-Core 3.0 GHz PentiumD
Infiniband (SDR)
0.6 TF/s on LQCD (DWF:asqtad avg)
Cluster Price/Performance in $/MF
100
10
1
0.1
2000
2002
2004
2006
2008
2010
Year

J/Psi GPU Addition

- Recent results by USQCD physicists at Boston U., as well as earlier work in Europe, duplicated at UCSD, indicated very good cost effectiveness for certain LQCD codes on Nvidia GPUs
- With some of the surplus from FY09, we purchased four Nvidia S1070 quad-GPU systems and attached them to eight J/Psi nodes
  - The current codes use single GPU’s
  - Multi-GPU parallel code can use the J/Psi Infiniband interconnects
- This GPU resource is available to any USQCD scientist
  - Boston U. anistropic clover propagator generation code is ready for production (they anticipate running starting within the next several weeks)
  - Performance is > 100 Gflops per GPU (vs ~ 8 Gflops per J/Psi node)
  - They also have Wilson code available

FY09 Remaining Technical Issues

- Parallel file systems
  - To date FNAL has used NFS and dCache, and TJNAF has used NFS
  - FNAL LQCD team has leveraged off-project Lustre evaluation efforts and has deployed 72 Tbytes of Lustre for LQCD
  - Stability and usability will be carefully monitored as more pilot production is done using Lustre
- GPU cluster
  - Physics production will start in mid-June
  - Looking forward to future acquisitions, we need to understand reliability and usability of this pilot facility

Summary

- We combined FY08 and FY09 purchases into a single procurement
  - Such combined purchases save manpower, and (at FNAL) G&A
  - Very successful
  - We will very likely do more of this in the future
- The resulting cluster, J/Psi, came into production in early January (FY08 portion: 5.75 TF) and mid-April (FY09 portion: 2.65 TF)
  - 8.4 TF total exceeds OMB-300 technical goal of 6.2 TF
  - The winning vendor was creative (used faster memory than spec) and pricing was aggressive – the “best value” process worked very well
  - USQCD cluster price/performance continues to follow the exponential curve that we’ve tracked since 2000

Backup Slides

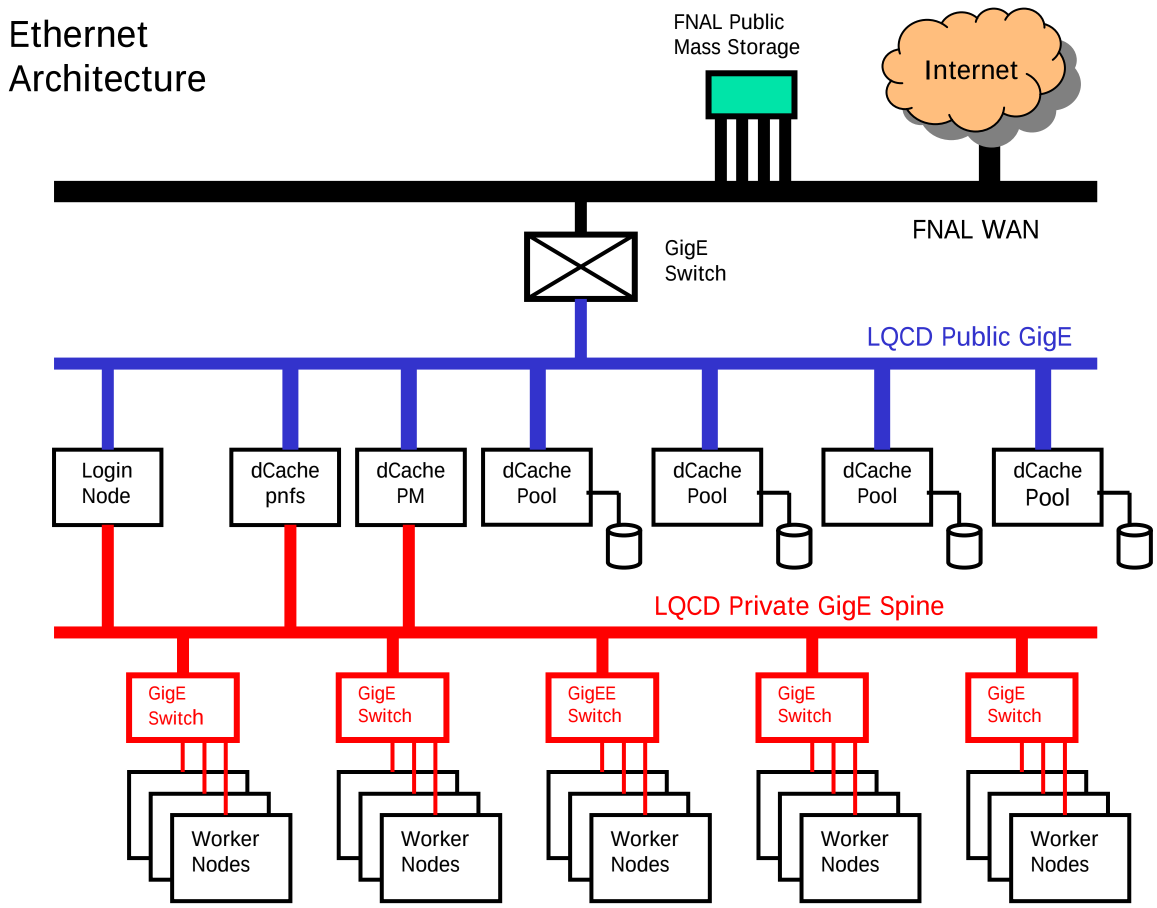
J/Psi Cluster Design

- Followed closely the two prior LQCD clusters (FNAL Kaon, TJNAF 7n):
  - AMD Opteron quad-core
  - Leaf and spine Infiniband DDR fabric, with 3:1 oversubscription
  - Independent service (NFS, job control) and hardware management (IPMI) private Ethernet fabrics
  - Head node for logins, with identical second node for NFS (home areas), Torque, and Maui; both nodes connected to public and private Ethernet networks

J/Psi Cluster Design

- Storage
  - Access via Gigabit Ethernet to existing facility RAID disks (20 TBytes)
  - Access via Infiniband, or Infiniband bridged via 10-Gigabit Ethernet, to existing LQCD facility dCache parallel storage (72 TBytes) and Luster parallel storage (72 TBytes)
  - Access via Gigabit Ethernet to Fermilab mass storage systems: tape robot, and dCache disk front end (currently 315 TBytes on tape)

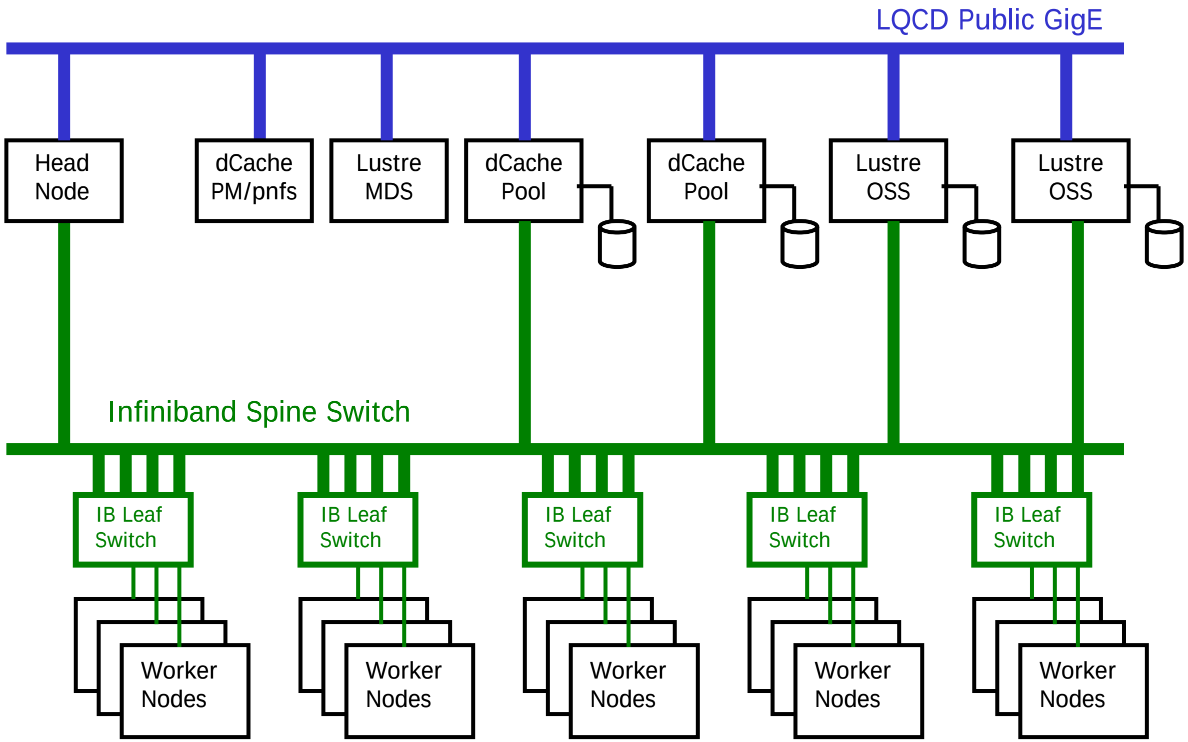
Cluster Layout – Ethernet and Mass Storage



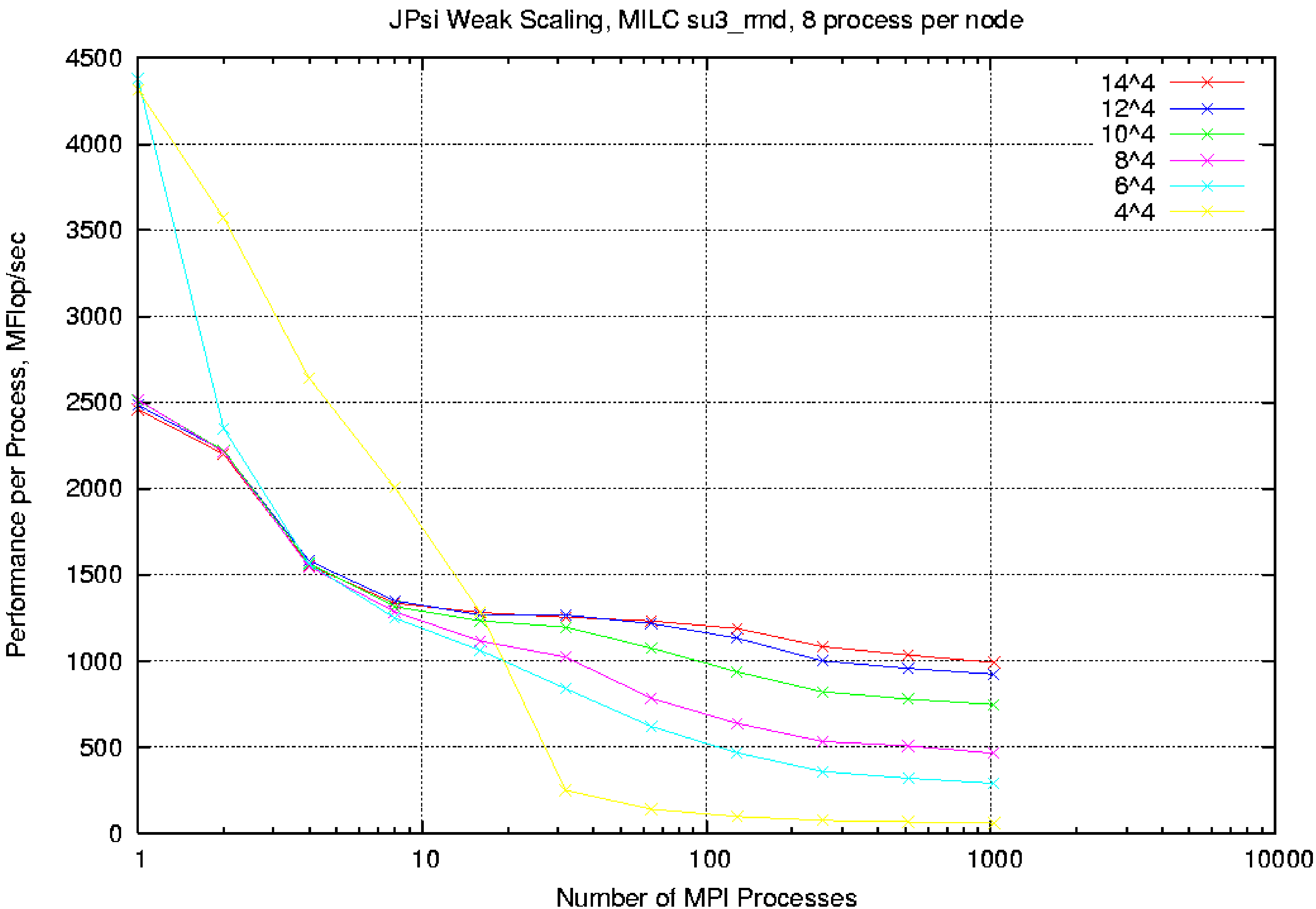
Cluster Layout – Infiniband and Mass Storage

Infiniband

Architecture



Communications



Fermilab J/Psi

- Total cost: $1.89M
  - Includes nodes, Infiniband, ethernet, racks, all incidental cabling
  - 856 nodes, 8.4 TF/s sustained on LQCD code
  - 37.4 TF/s Top500, 57.5 TF/s peak
    - → $32.9K/peak TF
    - → $50.5K/Top500 TF
    - → $225K/LQCD Sustained TF

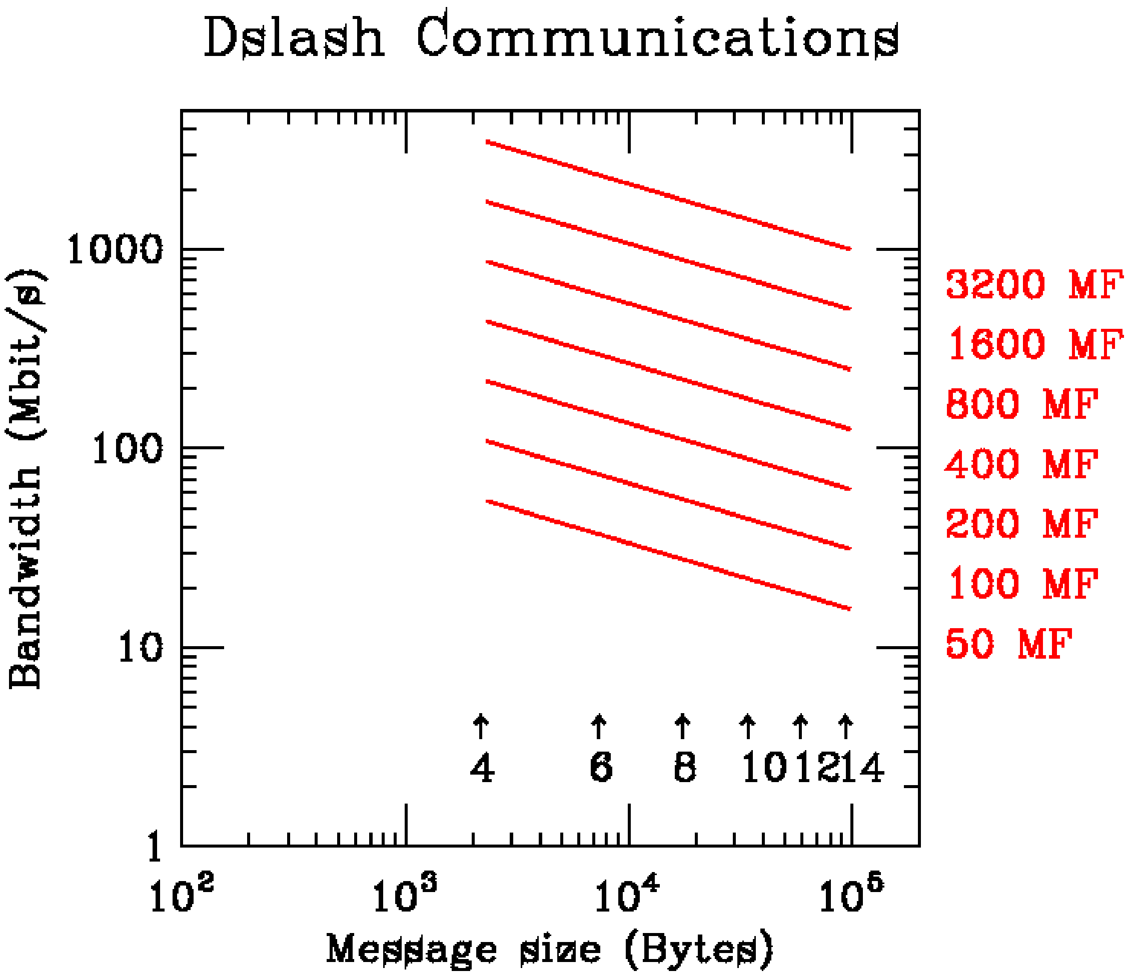
FNAL Kaon

- Total cost: $1.572M
  - Includes nodes, Infiniband, ethernet, racks, all incidental cabling
  - 600 nodes, 2.6 TF/s sustained on LQCD code
  - 6.894 TF/s Top500, 9.6 TF/s peak
    - → $164K/peak TF
    - → $228K/Top500 TF
    - → $605K/LQCD Sustained TF

TJNAF 7n

- Total cost: $1.33M
- Includes nodes, Infiniband, ethernet, racks, all incidental cabling
  - 396 nodes, 2.98 TF/s sustained on LQCD code
  - 13.46 TF/s Top500, 24.1 TF/s peak
    - → $55K/peak TF
    - → $99K/Top500 TF
    - → $446K/LQCD Sustained TF

Balanced Design Requirements
Communications for Dslash



Modified for improved staggered from Steve Gottlieb's staggered model:
physics.indiana.edu/~sg/pcnets/

Assume:

- L^4 lattice
- communications in 4 directions

Then:

- L implies message size to communicate a hyperplane
- Sustained MFlop/sec together with message size implies achieved communications bandwidth

Required network bandwidth increases as L decreases, and as sustained MFlop/sec increases

SDR vs. DDR Infiniband