Technical Design and Scope for the FY08-FY09 (“J/Psi”) Procurement

Don Holmgren

Fermilab

djholm@fnal.gov

LQCD Project Progress Review

Brookhaven National Laboratory

May 13-14, 2008

Outline

- Commodity cluster hardware requirements imposed by LQCD codes
- Existing LQCD project clusters
- Performance of currently available hardware
- Recap of decisions from last year's review
- FY08/FY09 cluster design
- Procurement status and schedule
- Projected performance

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 required for efficient global reductions, and for good strong scaling

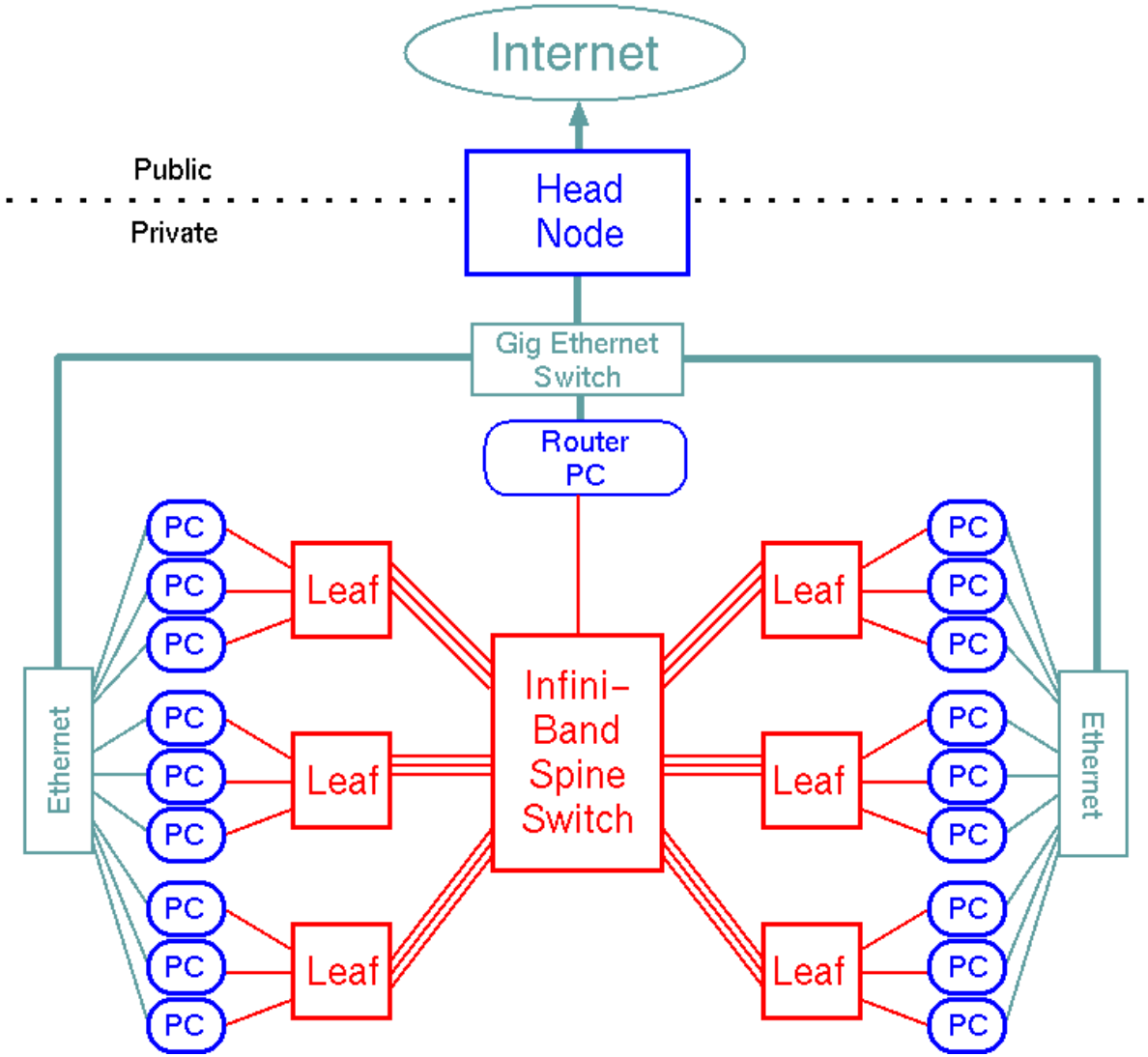
Hardware Requirements

- 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 (0.5-1.0 GB/core)
  - High performance interconnects
    (Infiniband now, Myrinet, Gigabit Ethernet meshes previously)

LQCD Cluster Layout



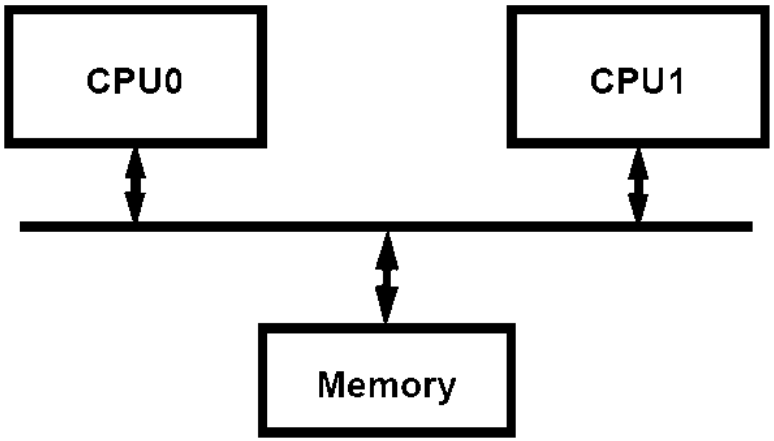
LQCD Project Clusters

Name (FNAL/TJNAF)	6N	Kaon	7N	J/Psi (planned)
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)	2.0 – 2.8 GHz Xeon/Opteron? (1200?) (4)
Memory Bus Speed	800 MHz	1066 MHz	1066 MHz	1333 -1600 MHz
Single or Dual Socket	Single	Dual	Dual	Single or 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	6.2 Tflop/s asqtad:DWF
Date in Production	3/2006	10/2006	6/2007 (Quad Core Upgrade 11/2007)	1/2009 (planned)

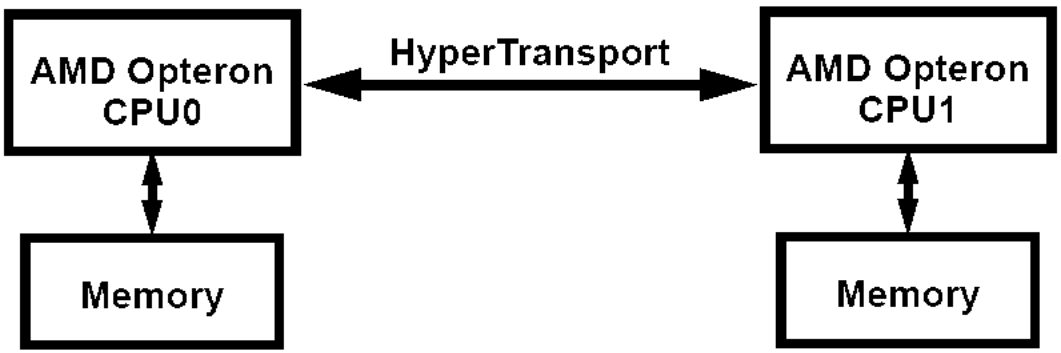
Currently Available Hardware

- Processors suitable for LQCD
  - AMD Opteron
    - Quad Core (“Barcelona”)
    - NUMA design
    - Used on JLab 7n
  - Intel Xeon
    - Quad Core (45nm “Penryn”)
    - Only shared memory bus currently available for multi-socket
      - Insufficient memory bandwidth for LQCD
      - Potentially single socket systems could be competitive with AMD on price/performance

Memory Architectures

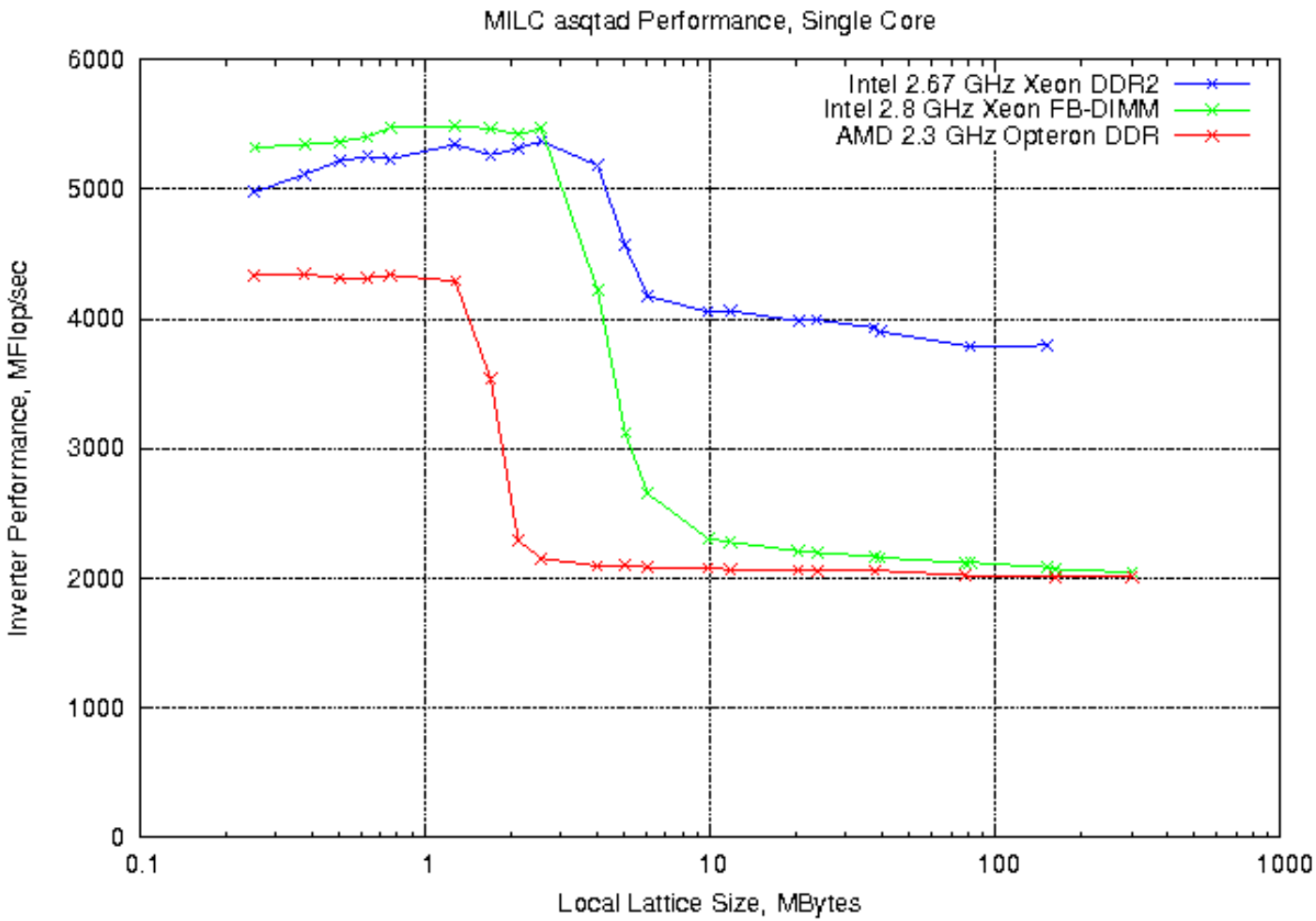


Shared Memory Bus Architecture



Non-Uniform Memory Architecture (NUMA)

Performance: Single Node, Using a Single Core on LQCD Code



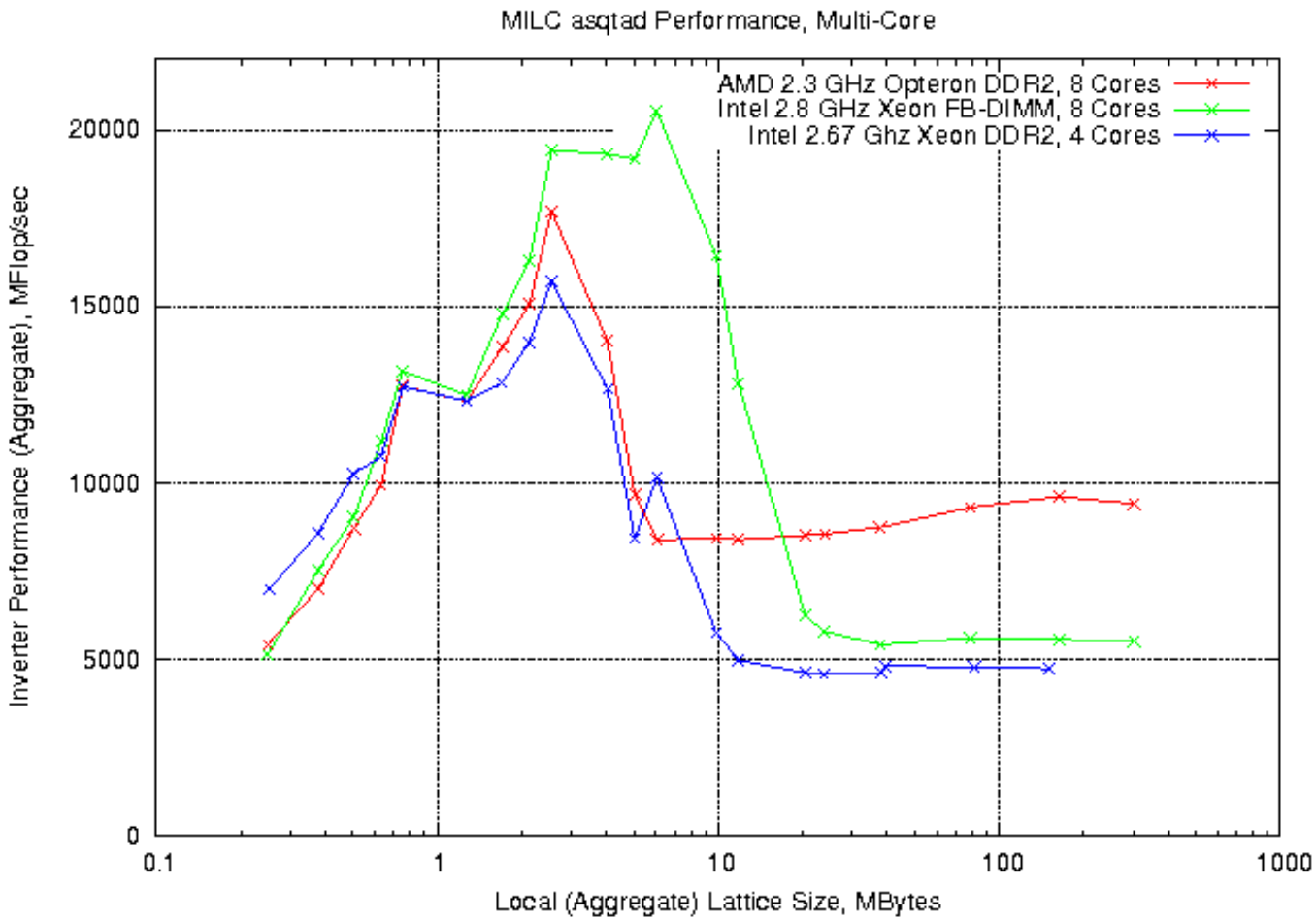
Plots show performance of single code instance

Intel Quad Core Single Socket

Intel Quad Core Dual Socket

AMD Quad Core (Barcelona)

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



Plots show aggregate performance of one MPI process on each core

Intel Quad Core Single Socket

Intel Quad Core Dual Socket

AMD Quad (Barcelona)

Intel “Nehalem”

- 2nd–generation 45nm processor design
  - NUMA architecture
  - Faster memory (1600 MHz, DDR3), more channels than Opteron (3 per socket vs 2 per socket)
  - HyperTransport-like inter-socket link (“QuickPath”)
  - If architecture is successful, should easily be the best performance choice for LQCD clusters (but not necessarily the best price/performance choice)
- Timing
  - No volume shipments expected until Jan 2009
  - Not in time for our FY08 procurement

Communications

- Since 2005, LQCD has used Infiniband
  - Prior clusters used Myrinet, or gigE meshes
  - Infiniband gave us better overall performance plus scaling to large clusters (512+ nodes)
- FY08 choices – use best price/performance
  - Double data rate (DDR) Infinihost III Lx host channel adapters (HCA's): either add-in, or "landed" on motherboard
  - DDR ConnectX HCA's (significantly more expensive)
    - Better latency, better bandwidth at message sizes of interest

Communications

- Switches
  - Current: based on InfiniScale-III silicon
    - 24-port leaf switches
    - 144-port and 288-port spine switches
  - New (likely in time): based on InfiniScale-IV silicon
    - 36-port leaf switches
    - Spines not available in time – would use a stack of 36-port leaf switches
    - Adaptive routing, better latency, simplified designs

Recap of Decisions 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 of Decisions from Last Year's Review

- Proposed strategy *(with possible revisions):*
  - 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*

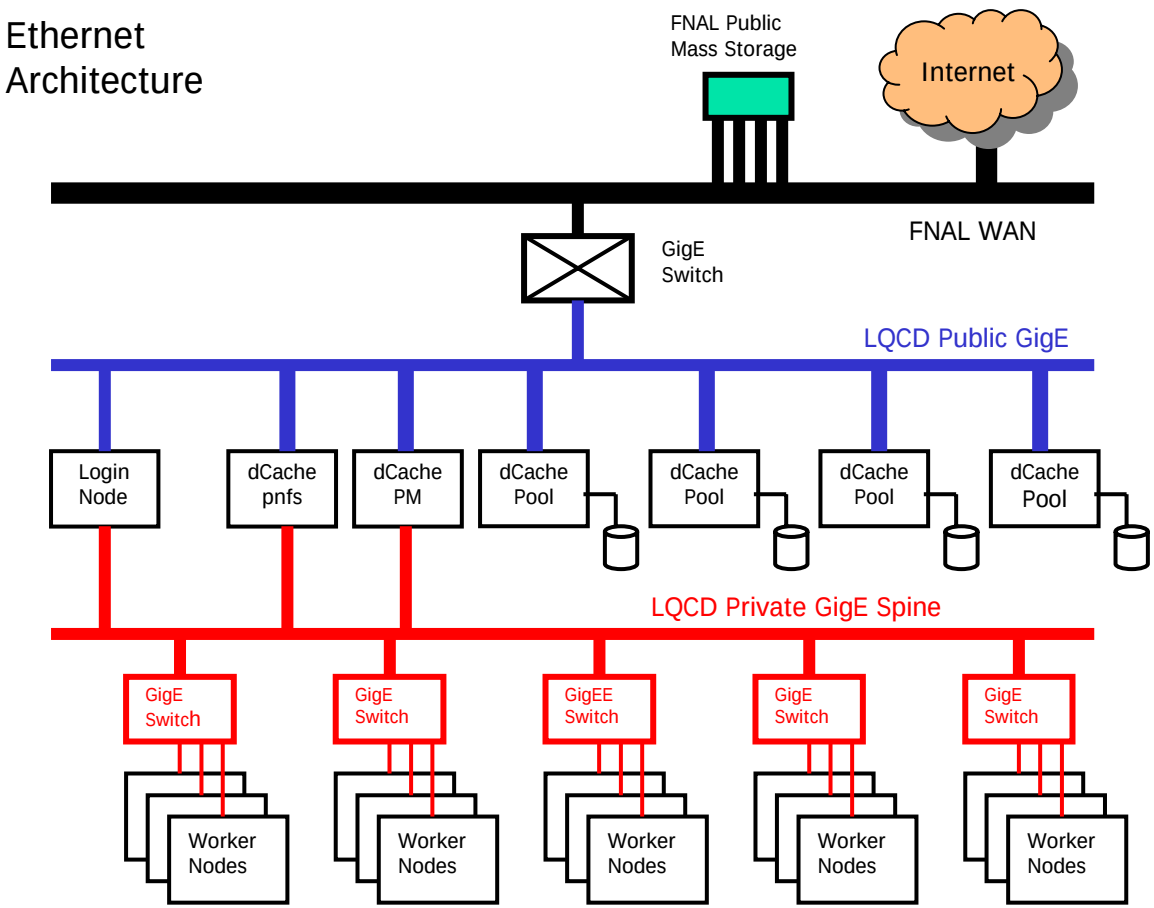
FY08-FY09 Cluster Design

- Following closely the two prior LQCD clusters (FNAL Kaon, JLab 7n):
  - AMD Opteron or Intel Xeon
  - 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

FY08-FY09 Cluster Design

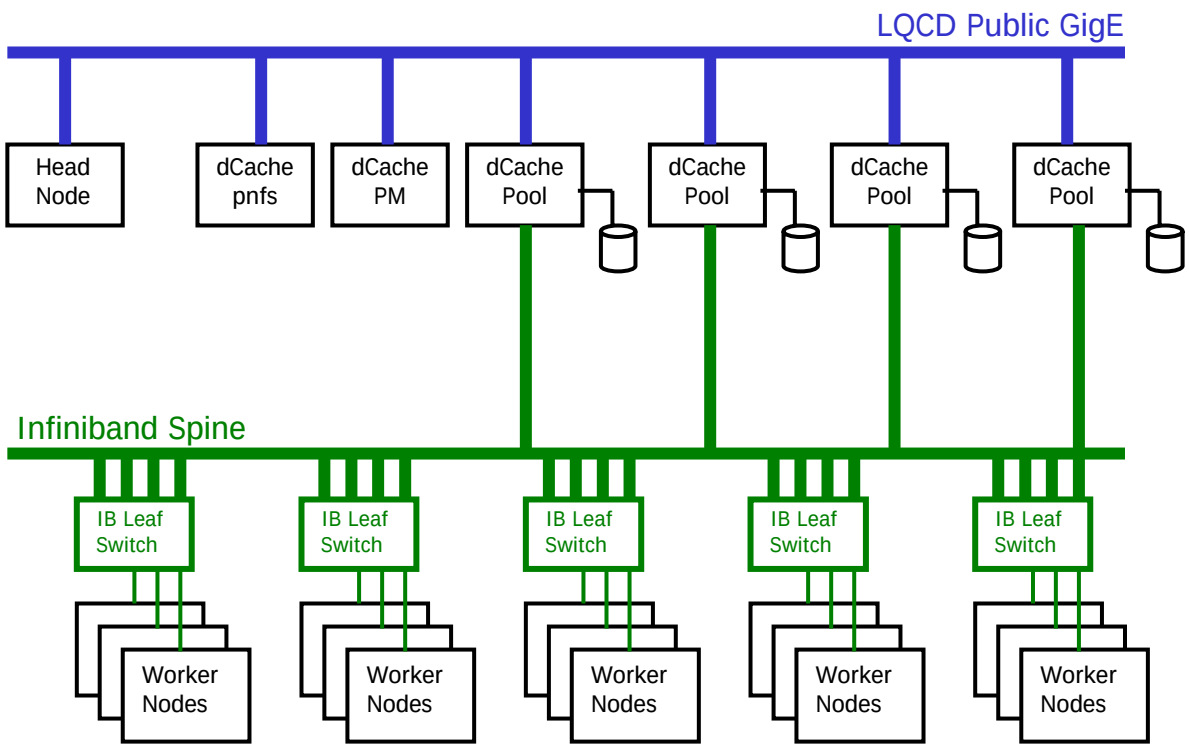
- Storage
  - Access via Gigabit Ethernet to existing facility RAID disks (20 TBytes)
  - Access via Infiniband, or Infiniband bridged via 10-Gigabit Ethernet, to existing LQCD facility dCache parallel storage (60 TBytes)
  - Access via Gigabit Ethernet to Fermilab mass storage systems: tape robot, and dCache disk front end (currently 250 TBytes on tape)

Cluster Layout – Ethernet and Mass Storage



Cluster Layout – Infiniband and Mass Storage

Infiniband
Architecture



Procurement Details

- Request for Information (RFI) was released to 24 vendors and manufacturers on March 14
  - Outline of technical specification was given, suggesting single, dual, or quad-socket AMD or Intel nodes and DDR Infiniband
  - Facility restrictions given (power type and amount, cooling, space)
  - A benchmark package was released (http://lqcd.fnal.gov/fermi_bench_64_mpich2.tar.bz2)
  - Vendors were asked for design recommendations, capabilities, testing capacity, and willingness to accept FY09 purchase option
  - 15 responses: 14 vendors who intend to bid conventional clusters according to our specifications, plus 1 other (SiCortex) who would like to bid a special purpose machine
    - We are awaiting LQCD application performance data from SiCortex to determine if their solution is viable

Procurement Details

- A series of non-disclosure meetings has been held with manufacturers and vendors:
  - Intel
  - AMD
  - Mellanox
  - SuperMicro
  - Tyan
  - SiCortex
  - Penguin
  - HP
  - Sun
- Benchmarking (by FNAL or vendor/manufacturer personnel)
  - Mellanox ConnectX
  - Intel “Penryn” and “Nehalem”
  - AMD “Barcelona”
  - SuperMicro Opteron motherboards with landed Infiniband

Schedule

2008

- Mar 14 – RFI released to vendors
- Apr 15 – RFI responses received from vendors
- May 30 – RFP release to vendors
- June 15 – Beneficial occupancy in new computer room
- June 30 – RFP bids due
- July 9 – RFP award recommendation
- July 15 – Access to new computer room
- July 23 – Contract award (commit FY08 funds)
- Sep 8 – Delivery of FY08 equipment complete
- Sep – Integration and testing begins
- Oct – Friendly user period begins
- Dec 1 – Release to production FY08 portion

Schedule

- July-Sept – Evaluate Nehalem hardware
- Oct 1 – Decision: exercise FY09 option or start new procurement for Nehalem hardware

FY09 Option:
- Oct 1+n – FY09 funds available
- +2 weeks – commit FY09 funds
- +7 weeks – delivery of hardware
- +11 weeks – release to production

Earliest possible release: Dec 17
Probable release: Jan 15

Nehalem Option:
- Oct 30 – release RFP
- Dec 1 – vendor bids due
- Dec 10 – RFP award recommendation
- Dec 31 – contract award
- Feb 15 – delivery of hardware
- Mar 15 – friendly user mode
- Apr 15 – release to production

All dates could slip!

The Nehalem option would be viable if the additional physics production during the latter half of 2009 exceeded the loss of production from Jan 15 – Apr 15.
→ Nehalem performance/price would have to be 1.5 times that of the FY08 hardware or better.
We would also consider total life cycle physics production.

Projected Performance

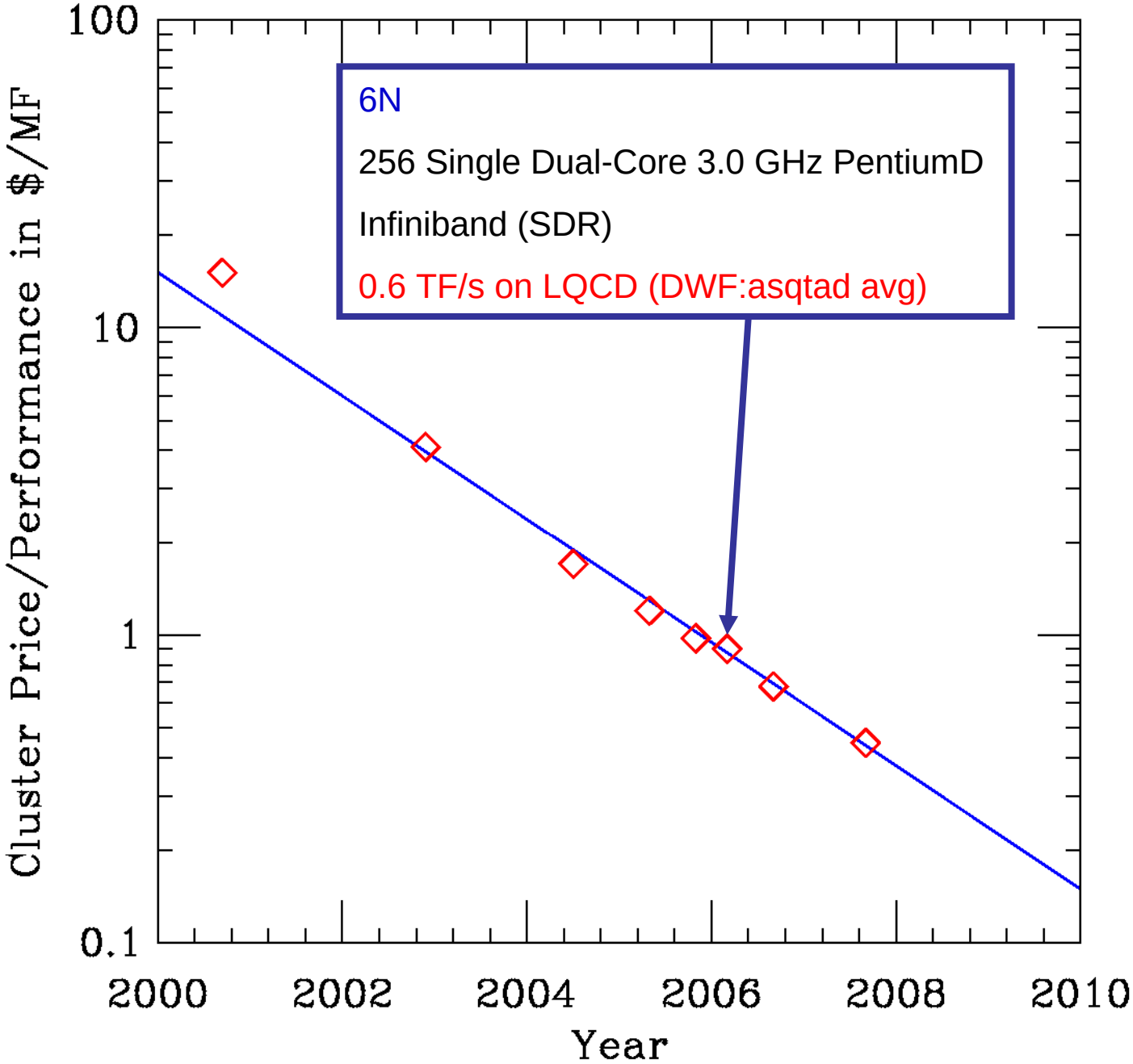
- MILC asqtad single node:
  - 2.0 - 2.3 GHz Barcelona: 9020 - 9450 MFlops/node
- DWF single node:
  - 2.0 - 2.3 GHz Barcelona: 10910 - 11820 MFlops/node
- Asqtad:DWF single node average
  - 2.0 - 2.3 GHz Barcelona: 9965 – 10635 MFlops/node
- Scaling: Factor of 0.92 (8 cores → 64 cores)
  - 2.0 - 2.3 GHz Barcelona cluster: 9170 - 9785 MFlops/node

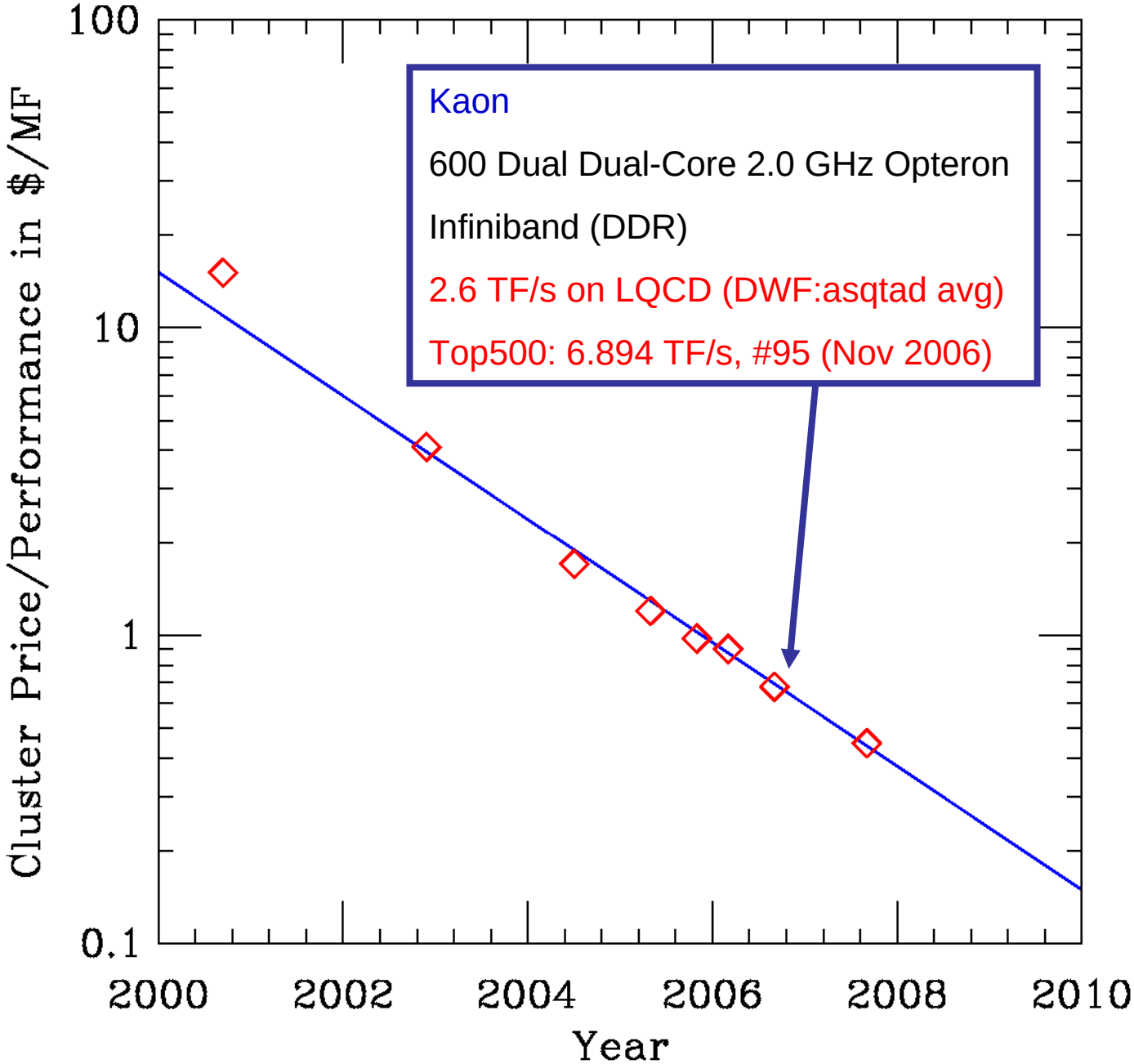
Projected Performance

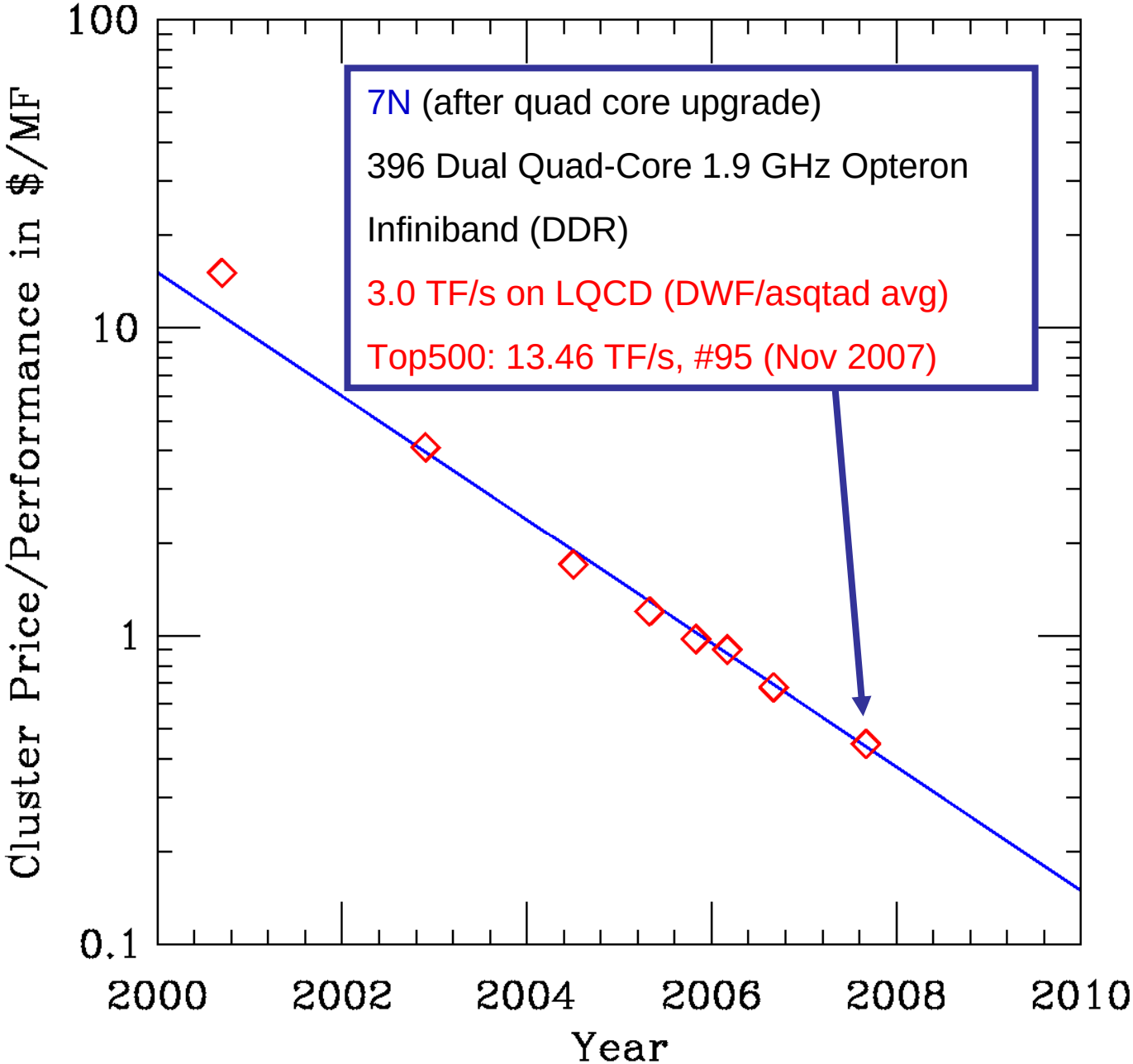
- One RFI response from a mid-price vendor estimated the total cost of a 612-node 2.1 GHz Barcelona cluster with 3:1 DDR oversubscription at $2,140K
- Total FY08+FY09 less G&A is $2,090K
- 600 node performance range (2.0 - 2.3 GHz) based on benchmarking to date is 5.5 – 5.9 TFlops
- Goal is 6.2 TFlops
- Expect at least 10% better pricing than this estimate through the RFP process
  - Cost savings through Infiniband on motherboard ($200 per node)
  - Effect of competition
  - Range is 6.0 – 6.5 TFlops

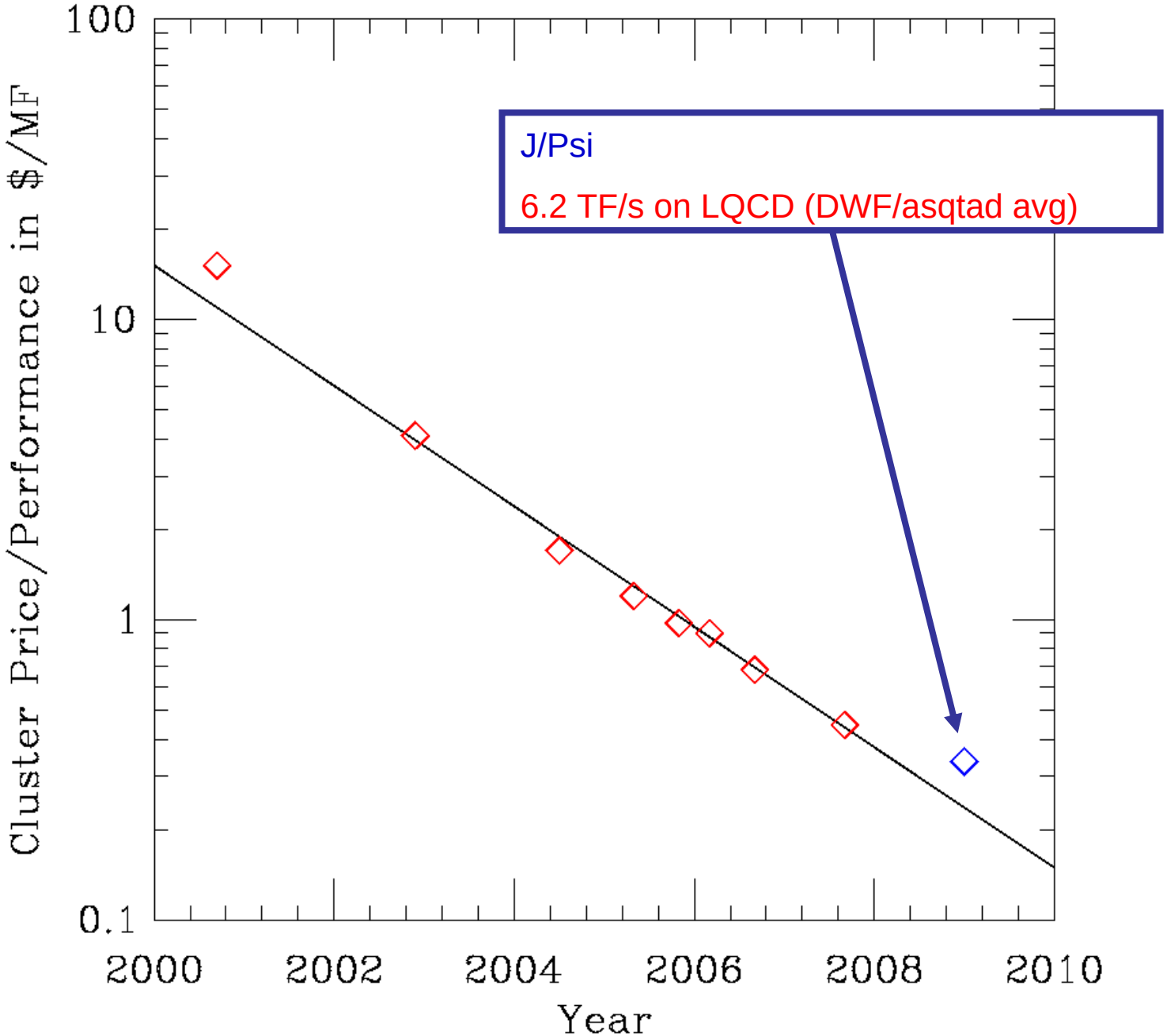
Impact on OMB-300 Delivered Tflops-yrs Milestones

- Currently deployed Tflops (BNL + FNAL + JLab):
  4.20 + 3.58 + 3.97 = 11.75
- Delivered Tflops-yrs milestones (FY06-FY09):
  6.2, 9.0, 12.0, 15.0
- A conservative, achievable J-Psi estimate is 6.2 TFlops
- Achieved (Predicted) Tflops-yrs:
  - FY06: 6.21
  - FY07: 9.68
  - FY08: 12.03(based on performance through April 30, 2008)
  - FY09: 16.34 (assume J-Psi = 6.2 Tflops, JLab decommissioned 3G/4G, J-Psi operates for 9 months, only 90% uptime)


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


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


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

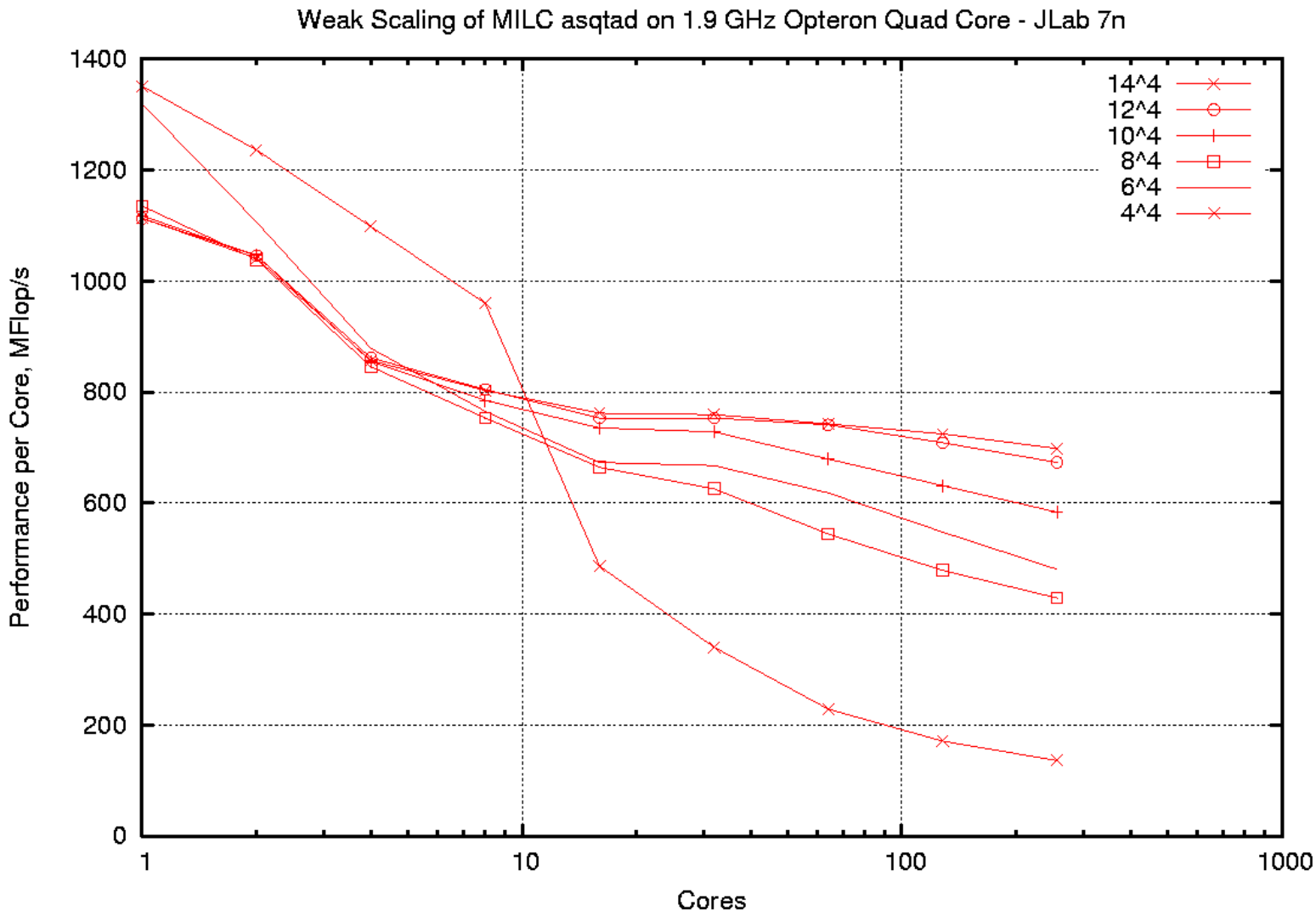

J/Psi
6.2 TF/s on LQCD (DWF/asqtad avg)
Cluster Price/Performance in $/MF
100
10
1
0.1
2000
2002
2004
2006
2008
2010
Year

Summary

- We are proceeding on plans to combine FY08 and FY09 purchases into a single procurement, J/Psi
- Based on benchmarking, an RFI, and price estimates, J/Psi computing capacity will meet or exceed the 6.2 TF goal
- Assuming J/Psi delivers 6.2 TF, 90% uptime, and availability for 9 months, the LQCD project will exceed the 15.0 TF-yrs goal in FY09 (estimate: 16.3 TF-yrs)
- The most likely hardware for J/Psi is AMD quad-core Opteron
- The project will continue to watch Intel Nehalem, potentially switching the FY09 procurement to this processor

Backup Slides

Communications



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

GCC-C Computer Room

Room Characteristics:

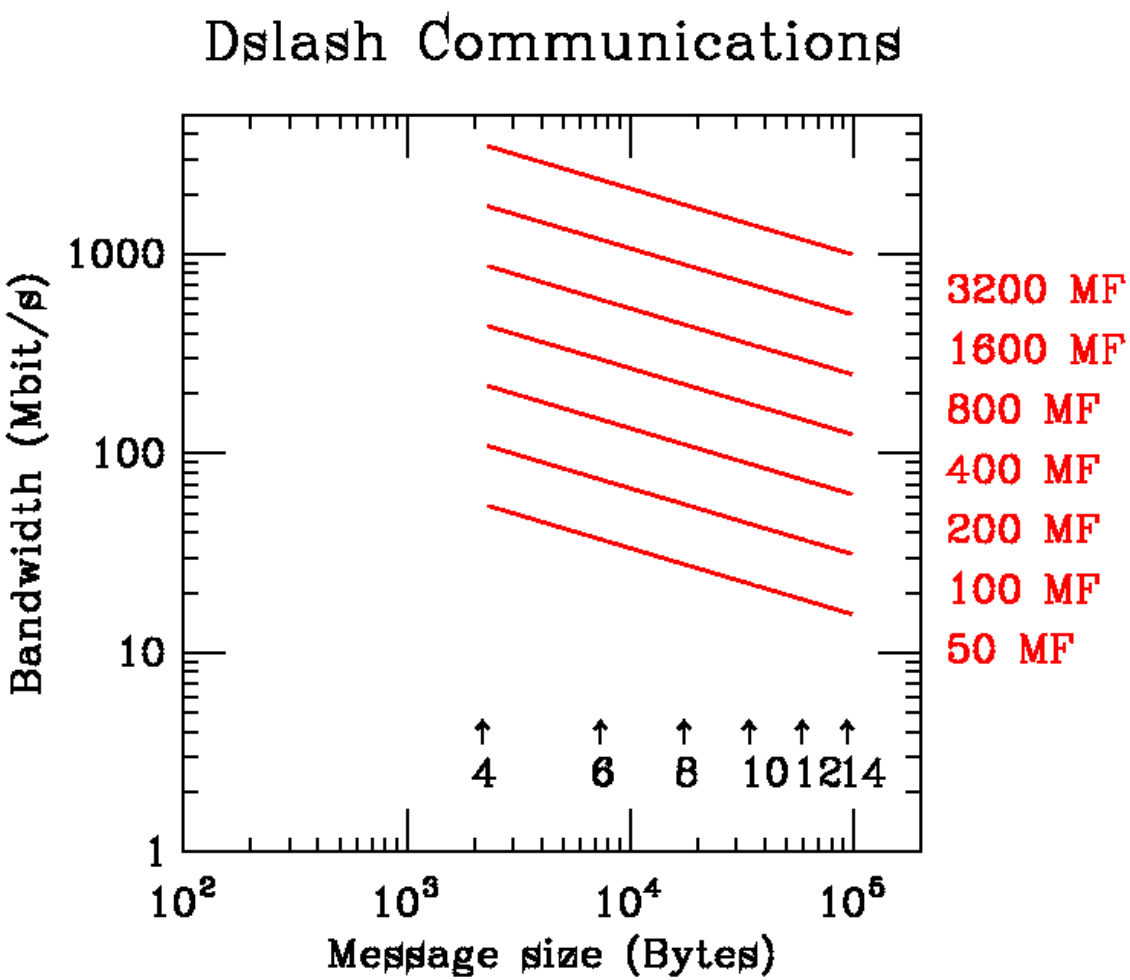
- 3 Foot raised-floor plenum
- 1000 KVA entrance (not incl. cooling)
- 208 VAC overhead service (L6-30R)
- 64 rack positions
- Total sensible CRAC capacity: 880 KW

TJNAF 7N

FNAL Kaon

Balanced Design Requirements

Communications for Dslash



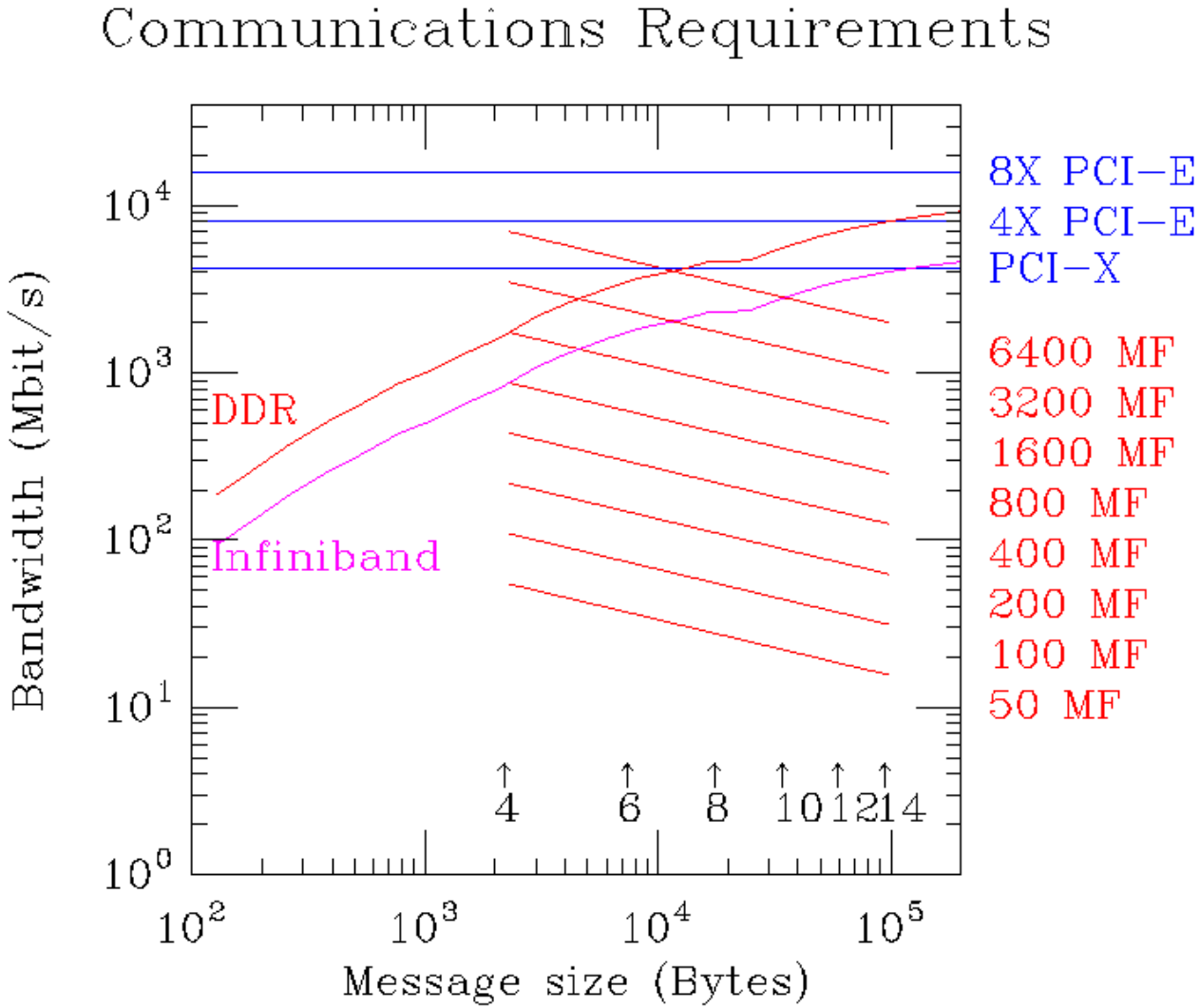
Modified for improved staggered from Steve Gottlieb's staggered model: physics.indiana.edu/~sg/pcnets/

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