Design Considerations

Don Holmgren
Lattice QCD Computing Project Review
Cambridge, MA
May 24-25, 2005

Road Map for My Talks

- Design Considerations
  - Price/performance: clusters vs BlueGene/L
  - Definitions of terms
  - Low level processor and I/O requirements
  - Procurement strategies
  - Performance expectations
- FY06 Procurement
  - FY06 cluster details – cost and schedule
- SciDAC Prototypes
  - JLab and Fermilab LQCD cluster experiences

Hardware Choices

- In each year of this project, we will construct or procure the most cost effective hardware
- In FY 2006:
  - Commodity clusters
  - Intel Pentium/Xeon or AMD Opteron
  - Infiniband

Hardware Choices

- Beyond FY 2006:
  - Choose between commodity clusters and:
    - An updated BlueGene/L
    - Other emerging supercomputers (for example, Raytheon Toro)
    - QCDOC++ (perhaps in FY 2009)
  - The most appropriate choice may be a mixture of these options

Clusters vs BlueGene/L

- BlueGene/L (source: BNL estimate from IBM)
  - Single rack pricing (1024 dual core cpu's):
    - $2M (includes $223K for an expensive 1.5 Tbyte IBM SAN)
    - $135K annual maintenance
  - 1 Tflop sustained performance on Wilson inverter (Lattice'04) using 1024 cpu's
  - Approximately $2/MFlop on Wilson Action
  - Rental costs:
    - $3.50/cpu-hr for small runs
    - $0.75/cpu-hr for large runs
    - 1024 dual-core cpu/rack
    - ~ $6M/rack/year @ $0.75/CPU-hr

Clusters vs. BlueGene/L

- Clusters (source: FNAL FY2005 procurement)
  - FY2005 FNAL Infiniband cluster:
    - ~ $2000/node total cost
    - ~ 1400 Mflop/s-node (14^4 asqtad local volume)
  - Approximately $1.4/MFlop
    - Note: asqtad has lower performance than Wilson, so Wilson would be lower than $1.4/MFlop
- Clusters have better price/performance then BlueGene/L in FY 2005
  - Any further performance gain by clusters in FY 2006 will further widen the gap

Definitions

- "TFlop/s" - average of domain wall fermion (DWF) and asqtad performance.
  - Ratio of DWF:asqtad is nominally 1.2:1, but this varies by machine (as high as 1.4:1)
  - "Top500" TFlop/s are considerably higher

	Top500 Tflop/s	LQCD Tflop/s
BlueGene/L	4.4	1.0
FNAL FY05	1.24	0.36

- "TFlop/s-yr" - available time-integrated performance during an 8000-hour year
  - Remaining 800 hours are assumed to be consumed by engineering time and other downtime

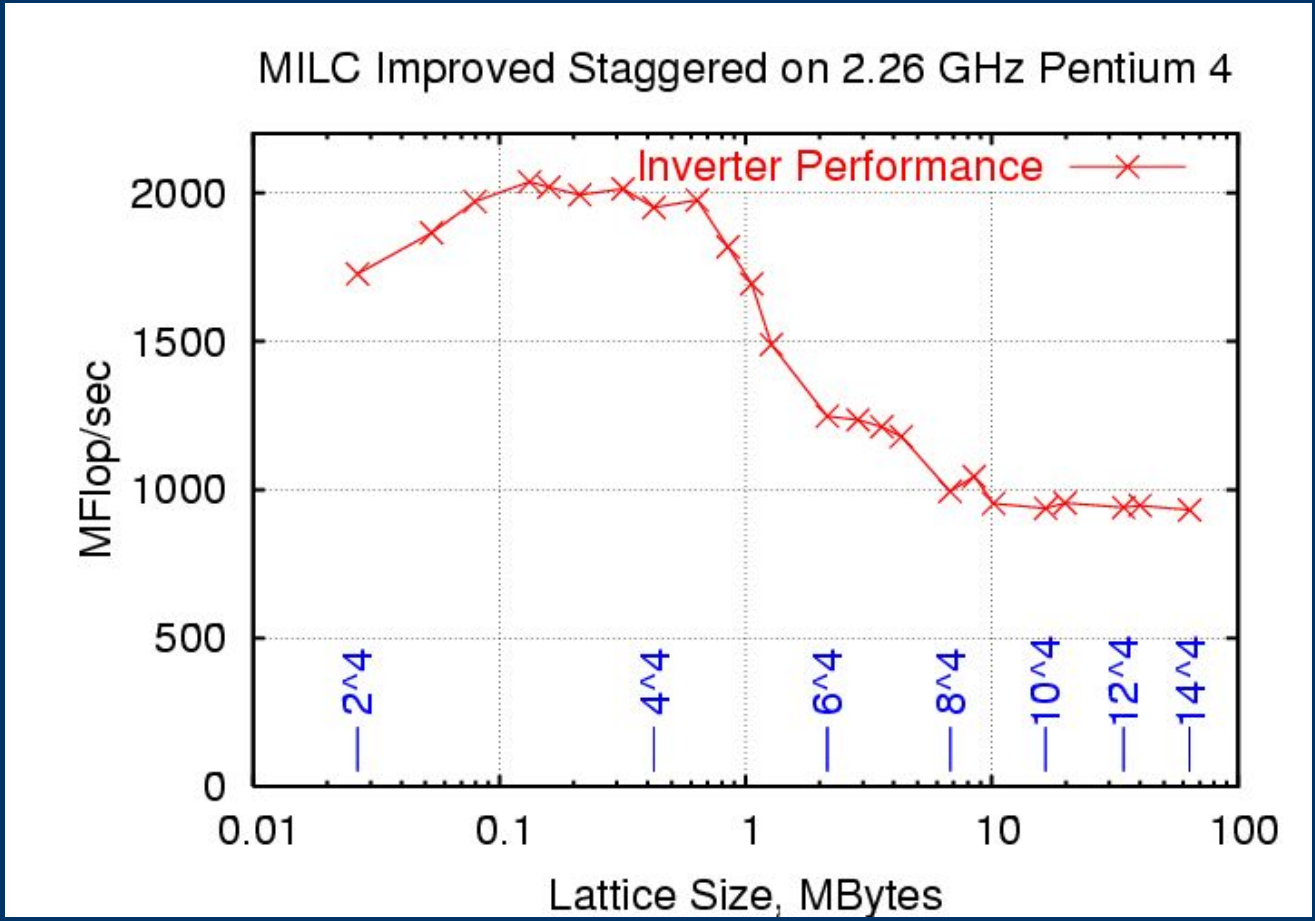
Aspects of Performance

- Lattice QCD codes require:
  - excellent single and double precision floating point performance
  - high memory bandwidth
  - low latency, high bandwidth communications

Balanced Designs
Dirac Operator

- Dirac operator (*Dslash*) – improved staggered action (“asqtad”)
  - 8 sets of pairs of SU(3) matrix-vector multiplies
  - Overlapped with communication of neighbor hypersurfaces
  - Accumulation of resulting vectors
- Dslash throughput depends upon performance of:
  - Floating point unit
  - Memory bus
  - I/O bus
  - Network fabric
- Any of these may be the bottleneck
  - bottleneck varies with local lattice size (surface:volume ratio)
  - We prefer floating point performance to be the bottleneck
    - Unfortunately, memory bandwidth is the main culprit
    - Balanced designs require a careful choice of components

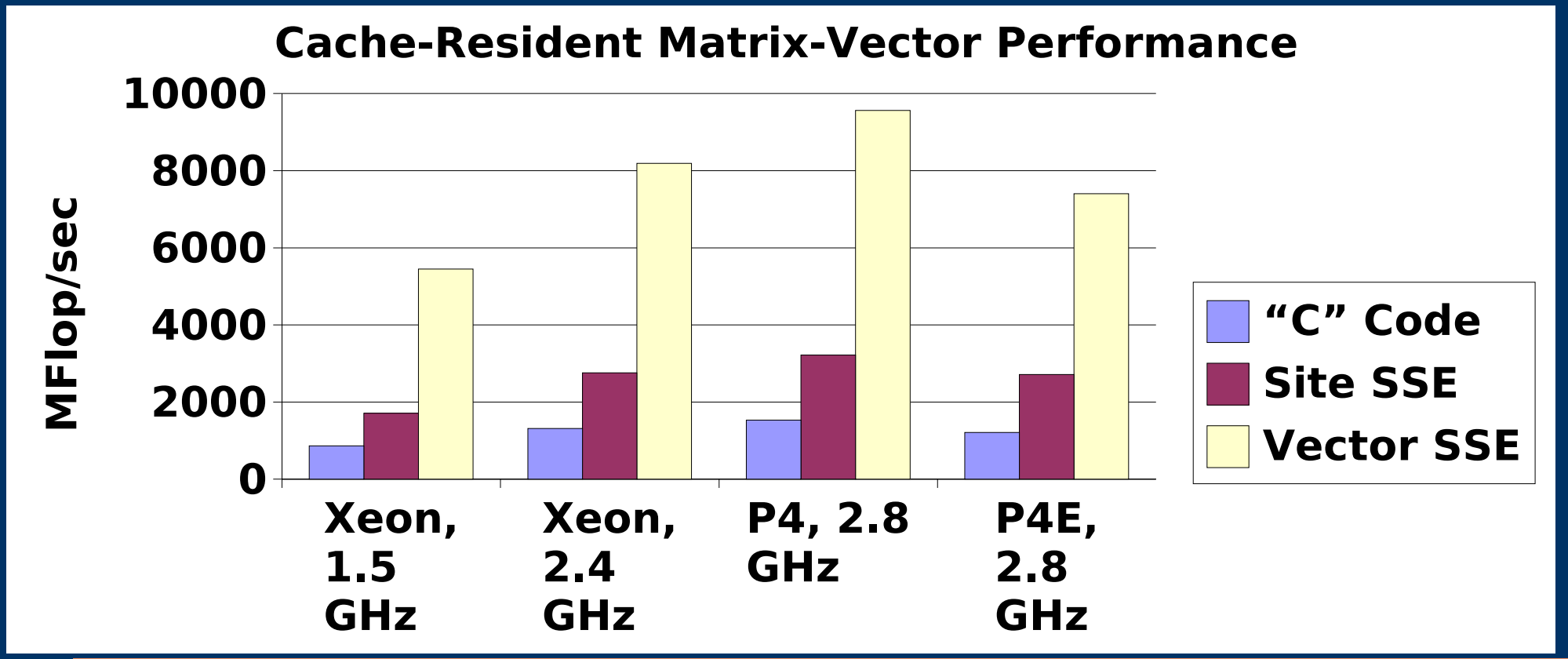
Generic Single Node Performance



- MILC is a standard MPI-based lattice QCD code
- Graph shows performance of a key routine: conjugate gradient Dirac operator inverter
- Cache size = 512 KB
- Floating point capabilities of the CPU limits in-cache performance
- Memory bus limits performance out-of-cache

Floating Point Performance (In cache)

- Most flops are SU(3) matrix times vector (complex)
  - SSE/SSE2/SSE3 can give a significant boost
  - Performance out of cache is dominated by memory bandwidth

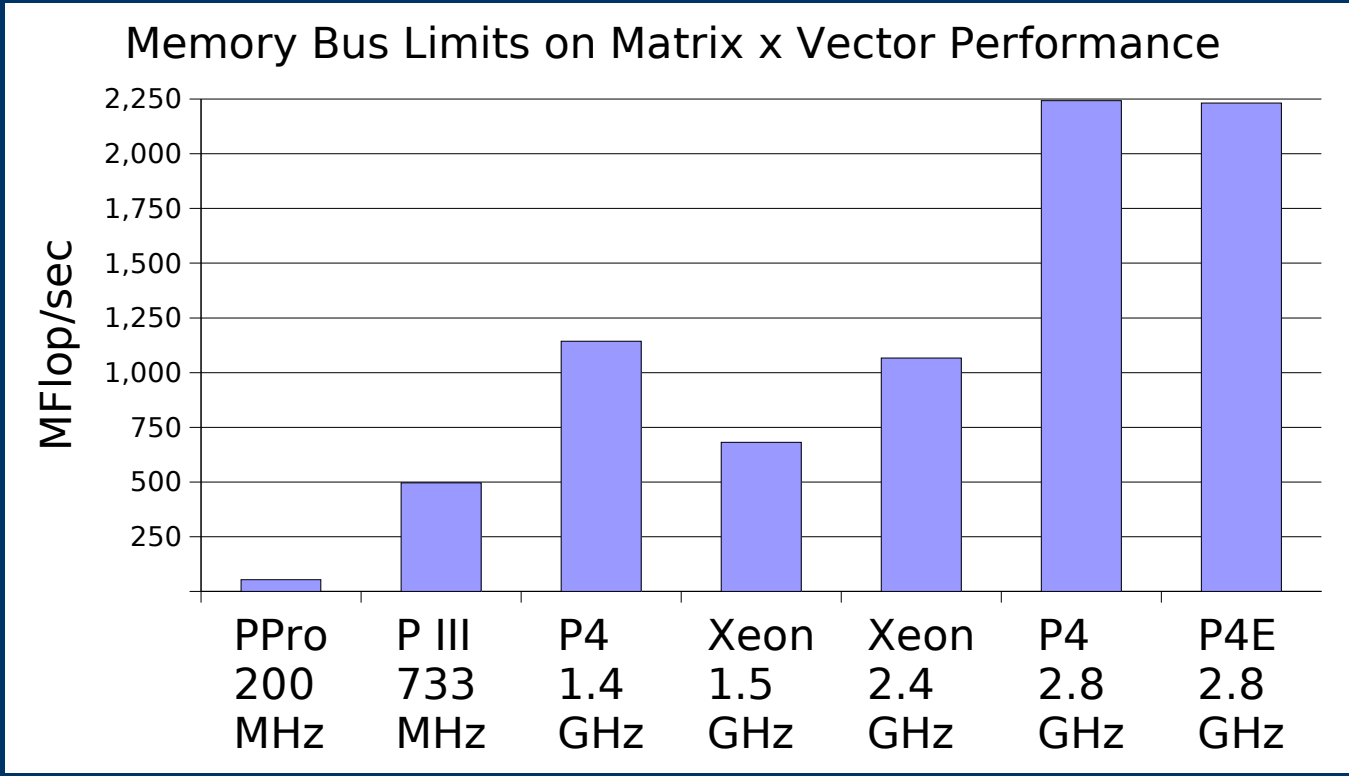


Memory Bandwidth Performance
Limits on Matrix-Vector Algebra

- From memory bandwidth benchmarks, we can estimate sustained matrix-vector performance in main memory
- We use:
  - 66 Flops per matrix-vector multiply
  - 96 input bytes
  - 24 output bytes
  - MFlop/sec = 66 / (96/read-rate + 24/write-rate)
    - read-rate and write-rate in MBytes/sec
- Memory bandwidth severely constrains performance for lattices larger than cache

Processor	FSB	Copy, MB/sec	SSE Read MB/sec	SSE Write MB/sec	M-V MFlop/sec
PPro 200 MHz	66 MHz	98	-	-	54
P III 733 MHz	133 MHz	405	880	1005	496
P4 1.4 GHz	400 MHz	1240	2070	2120	1,144
Xeon 2.4 GHz	400 MHz	1190	2260	1240	1,067
P4 2.8 GHz	800 MHz	2405	4100	3990	2,243
P4E 2.8 GHz	800 MHz	2500	4565	2810	2,232

Memory Bandwidth Performance
Limits on Matrix-Vector Algebra

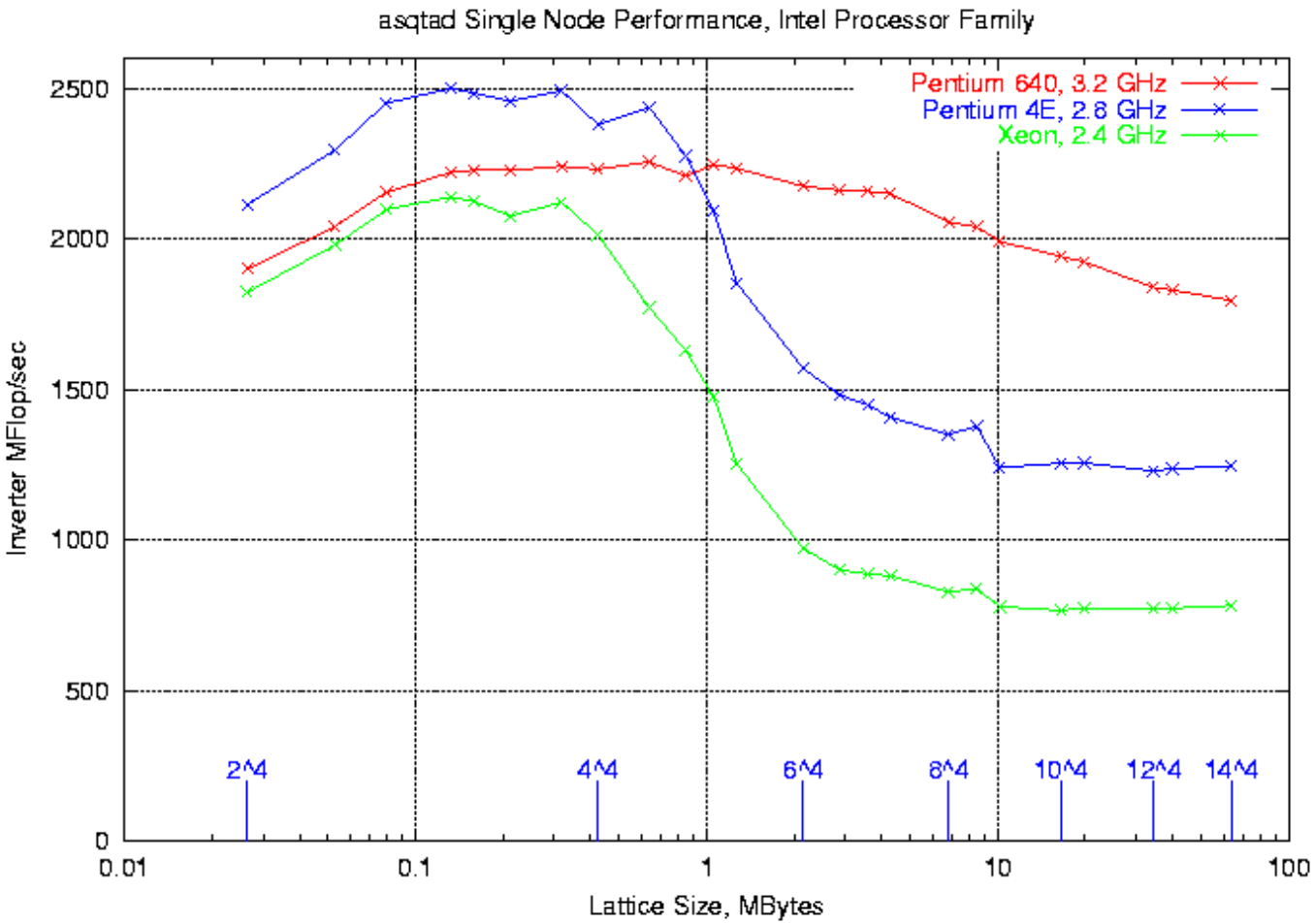


Memory Performance

- Memory bandwidth limits – depends on:
  - Width of data bus (64 or 128 bits)
  - (Effective) clock speed of memory bus (FSB)

- FSB history:
  - pre-1997: Pentium/Pentium Pro, EDO, 66 MHz, 528 MB/sec
  - 1998: Pentium II, SDRAM, 100 MHz, 800 MB/sec
  - 1999: Pentium III, SDRAM, 133 MHz, 1064 MB/sec
  - 2000: Pentium 4, RDRAM, 400 MHz, 3200 MB/sec
  - 2003: Pentium 4, DDR400, 800 MHz, 6400 MB/sec
  - 2004: Pentium 4, DDR533, 1066 MHz, 8530 MB/sec
  - Doubling time for peak bandwidth: 1.87 years
  - Doubling time for achieved bandwidth: 1.71 years
    - 1.49 years if SSE included (tracks Moore's Law)

Performance vs Architecture



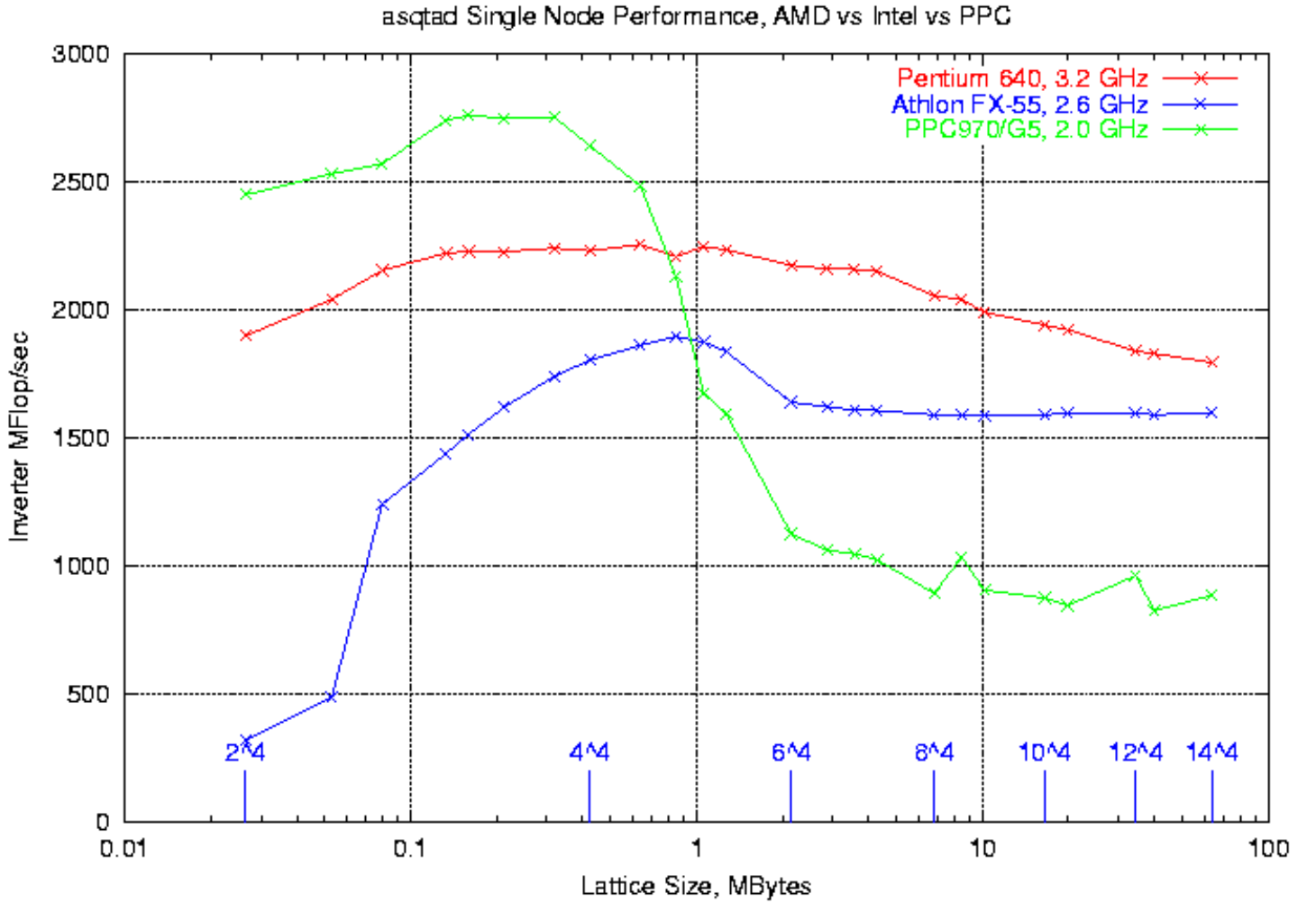
Memory buses:

- Xeon: 400 MHz
- P4E: 800 MHz
- P640: 800 MHz

P4E vs Xeon shows effects of faster FSB

P640 vs P4E shows effects of change in CPU architecture (larger L2 cache)

Performance vs Architecture



Comparison of current CPUs:

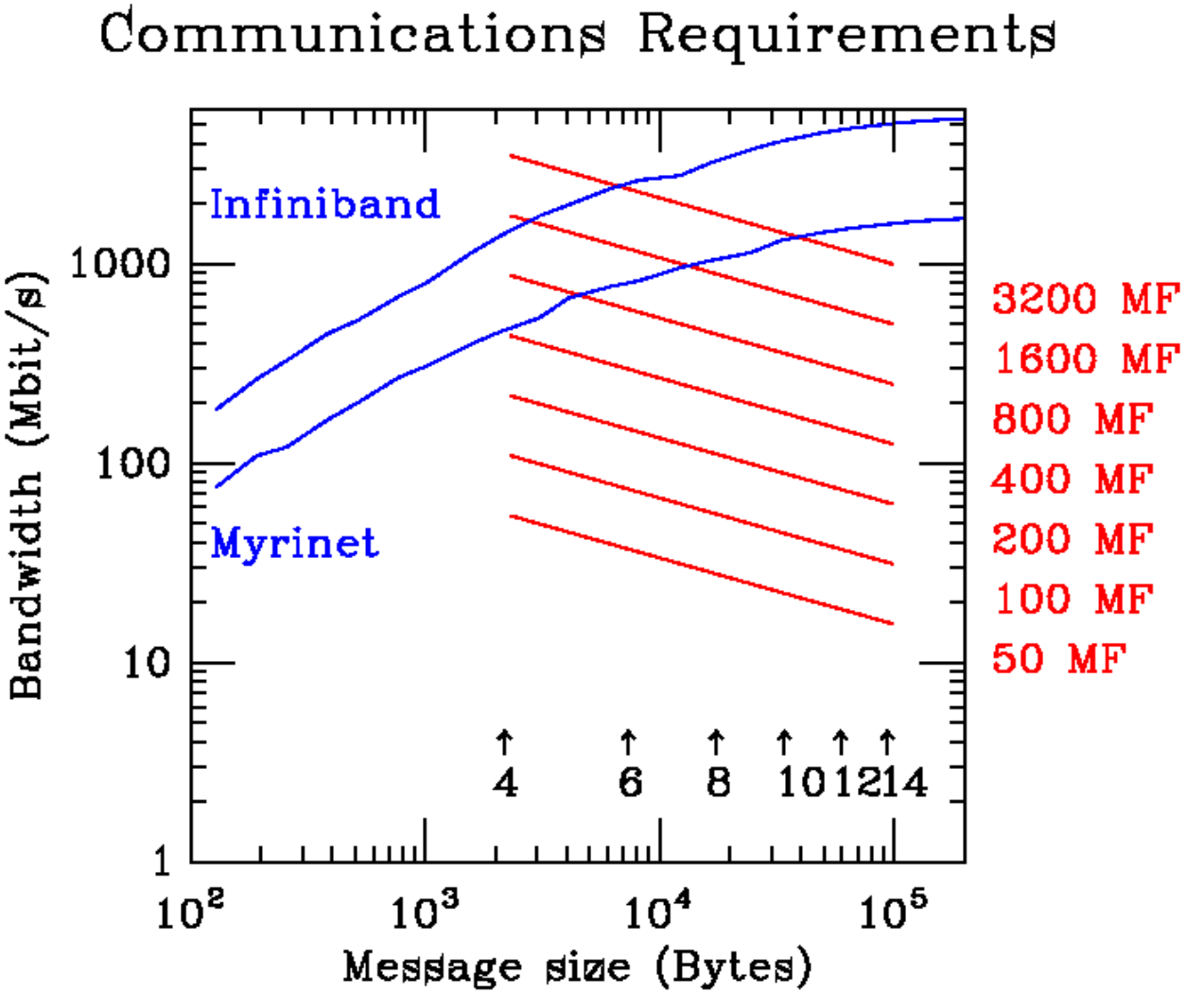
- Pentium 6xx
- AMD FX-55 (actually an Opteron)
- IBM PPC970

Pentium 6xx is most cost effective for LQCD

Communications

- On a cluster, we spread the lattice across many computing nodes
- Low latency and high bandwidths are required to interchange surface data
- Cluster performance depends on:
  - I/O bus (PCI and PCI Express)
  - Network fabric (Myrinet, switched gigE, gigE mesh, Quadrics, SCI, Infiniband)
  - Observed performance:
    - Myrinet 2000 (several years old) on PCI-X (E7500 chipset)
      Bidirectional Bandwidth: 300 MB/sec Latency: 11 usec
    - Infiniband on PCI-X (E7500 chipset)
      Bidirectional Bandwidth: 620 MB/sec Latency: 7.6 usec
    - Infiniband on PCI-E (925X chipset)
      Bidirectional Bandwidth: 1120 MB/sec Latency: 4.3 usec

Network Requirements

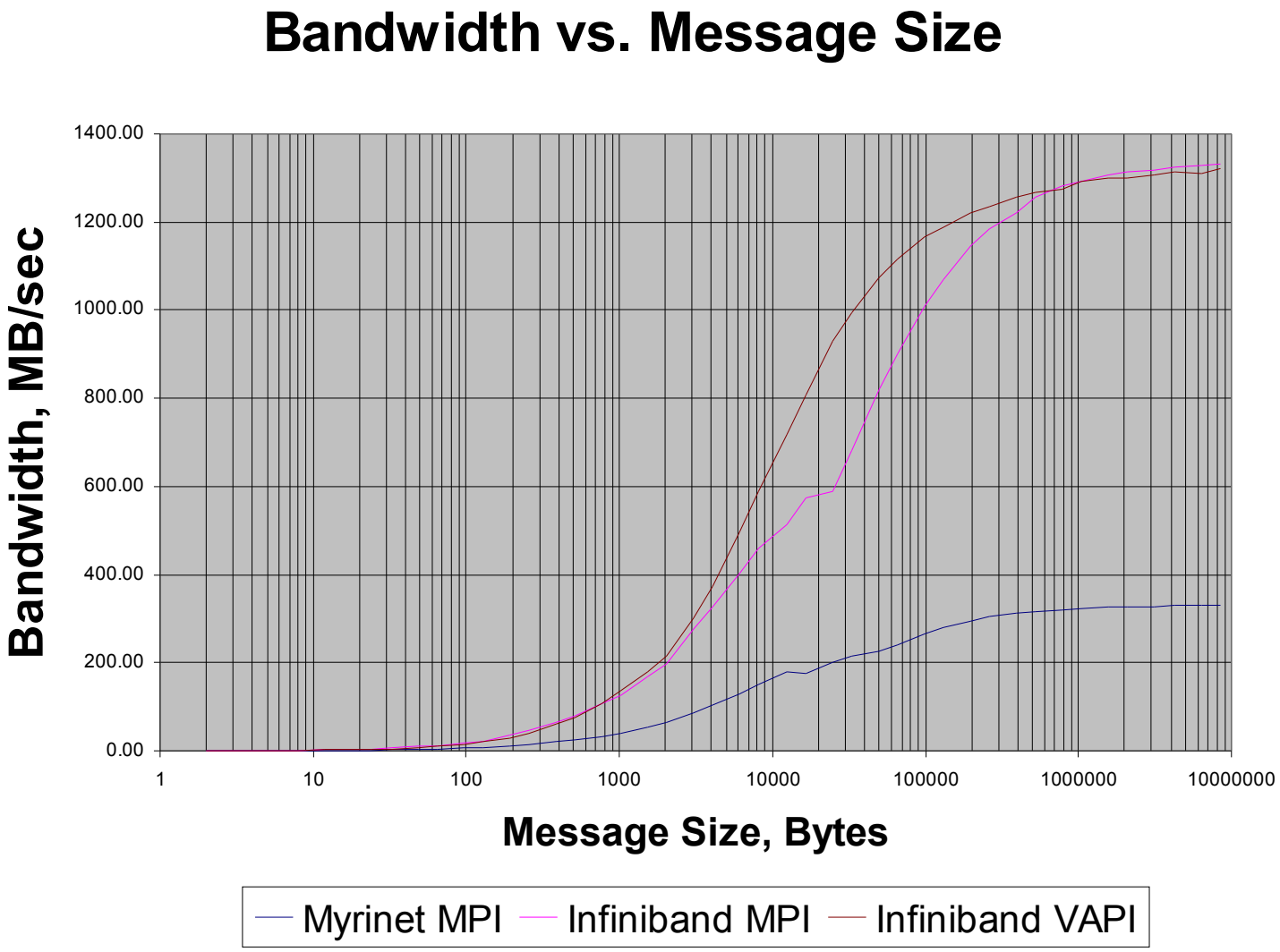


Red lines: required network bandwidth as a function of Dirac operator performance and local lattice size (L^4)

Blue curves: measured Myrinet (LANai-9) and Infiniband (4X PCI-E) unidirectional communications performance

These network curves give very optimistic upper bounds on performance

Measured Network Performance



- Graph shows bidirectional bandwidth
- Myrinet data from FNAL Dual Xeon Myrinet cluster
- Infiniband data from FNAL FY05 cluster
- Using VAPI instead of MPI should give significant boost to performance (SciDAC QMP)

Procurement Strategy

- Choose best overall price/performance
  - Intel ia32 currently better than AMD, G5
  - Maximize deliverable memory bandwidth
    - Sacrifice lower system count (singles, not duals)
  - Exploit architectural features
    - SIMD (SSE/SSE2/SSE3, Altivec, etc.)
  - Insist on some management features
    - IPMI
    - Server-class motherboards

Procurement Strategy

- Networks are as much as half the cost
  - GigE meshes dropped fraction to 25% at the cost of less operational flexibility
  - Network performance increases are slower than CPU, memory bandwidth increases
  - Over design if possible
    - More bandwidth than needed
  - Reuse if feasible
    - Network may last through CPU refresh (3 years)

Procurement Strategy

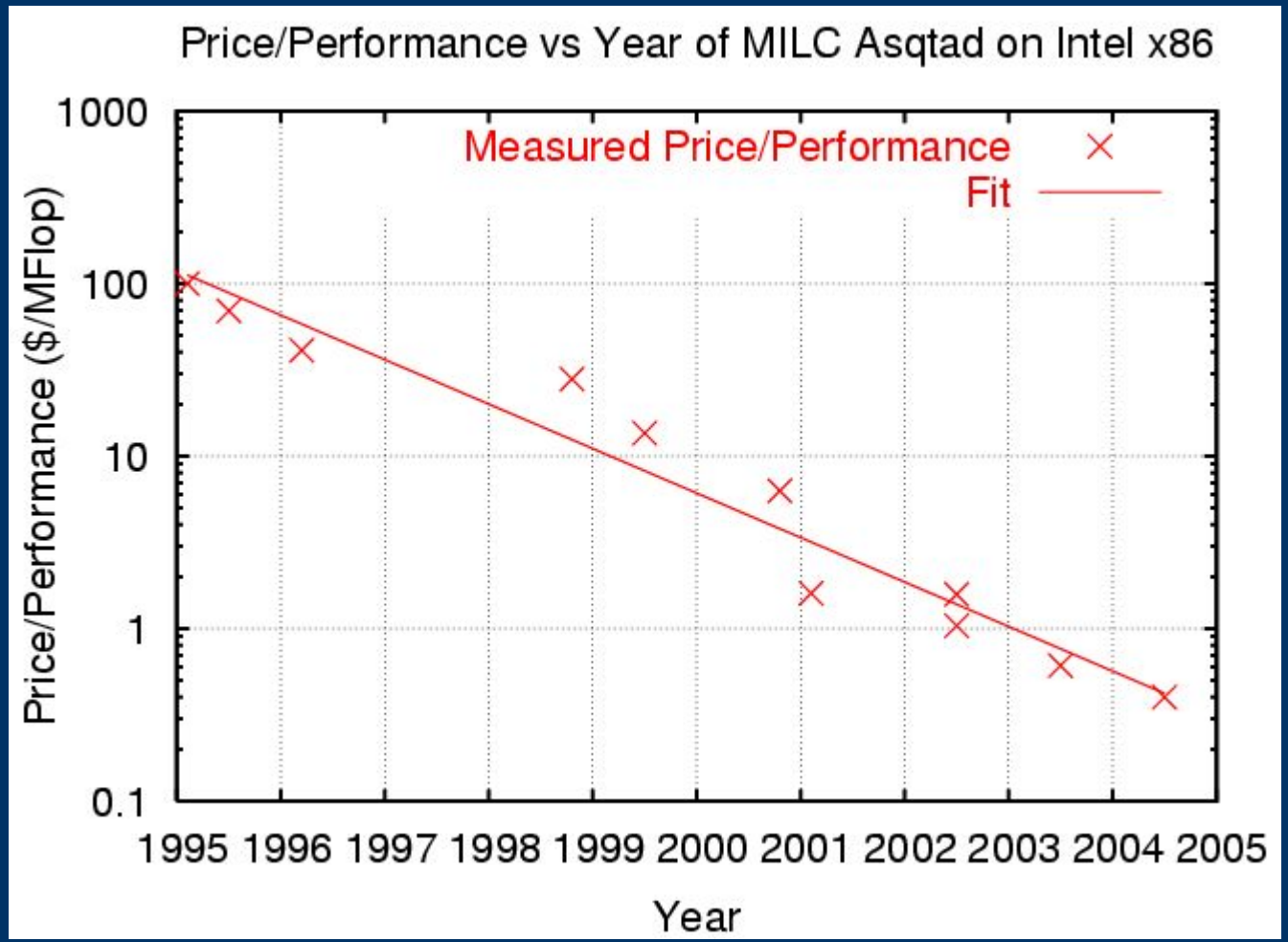
- Prototype!
  - Buy possible components (motherboards, processors, cases) and assemble in-house to understand issues
  - Track major changes – chipsets, architectures

Procurement Strategy

- Procure networks and systems separately
  - White box vendors tend not to have much experience with high performance networks
  - Network vendors (Myricom, the Infiniband vendors) likewise work with only a few OEMs and cluster vendors, but are happy to sell just the network components
  - Buy computers last (take advantage of technology improvements, price reductions)

Expectations

Performance Trends – Single Node



MILC Asqtad

Processors used:

- Pentium Pro, 66 MHz FSB
- Pentium II, 100 MHz FSB
- Pentium III, 100/133 FSB
- P4, 400/533/800 FSB
- Xeon, 400 MHz FSB
- P4E, 800 MHz FSB

Performance range:

- 48 to 1600 MFlop/sec
- measured at 12^4

Halving times:

- Performance: 1.88 years
- Price/Perf.: 1.19 years !!
- We use 1.5 years for planning

Performance Trends - Clusters

Clusters based on:

- Pentium II, 100 MHz FSB
- Pentium III, 100 MHz FSB
- Xeon, 400 MHz FSB
- P4E (estimate), 800 FSB

Performance range:

- 50 to 1200 MFlop/sec/node
- measured at 14^4 local lattice per node

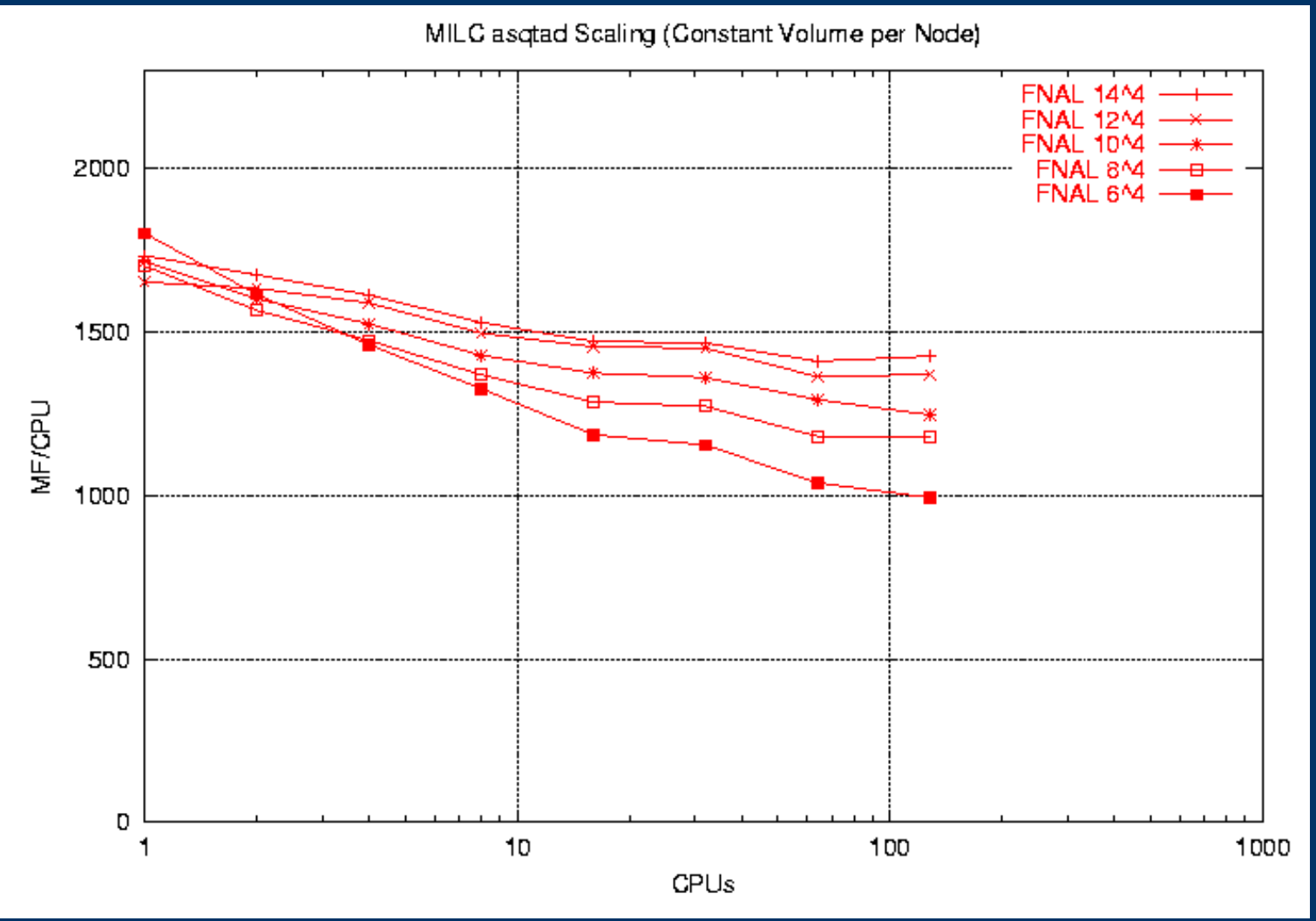
Halving Times:

- Performance: 1.22 years
- Price/Perf: 1.25 years
- We use 1.5 years for planning

Expectations

- FY06 cluster assumptions:
  - Single Pentium 4, or dual Opteron
  - PCI-E
  - Early (JLAB): 800 or 1066 MHz memory bus
  - Late (FNAL): 1066 or 1333 MHz memory bus
  - Infiniband
  - Extrapolate from FY05 performance

Expectations



FNAL FY 2005 Cluster:

- 3.2 GHz Pentium 640
- 800 MHz FSB
- Infiniband (2:1)
- PCI-E

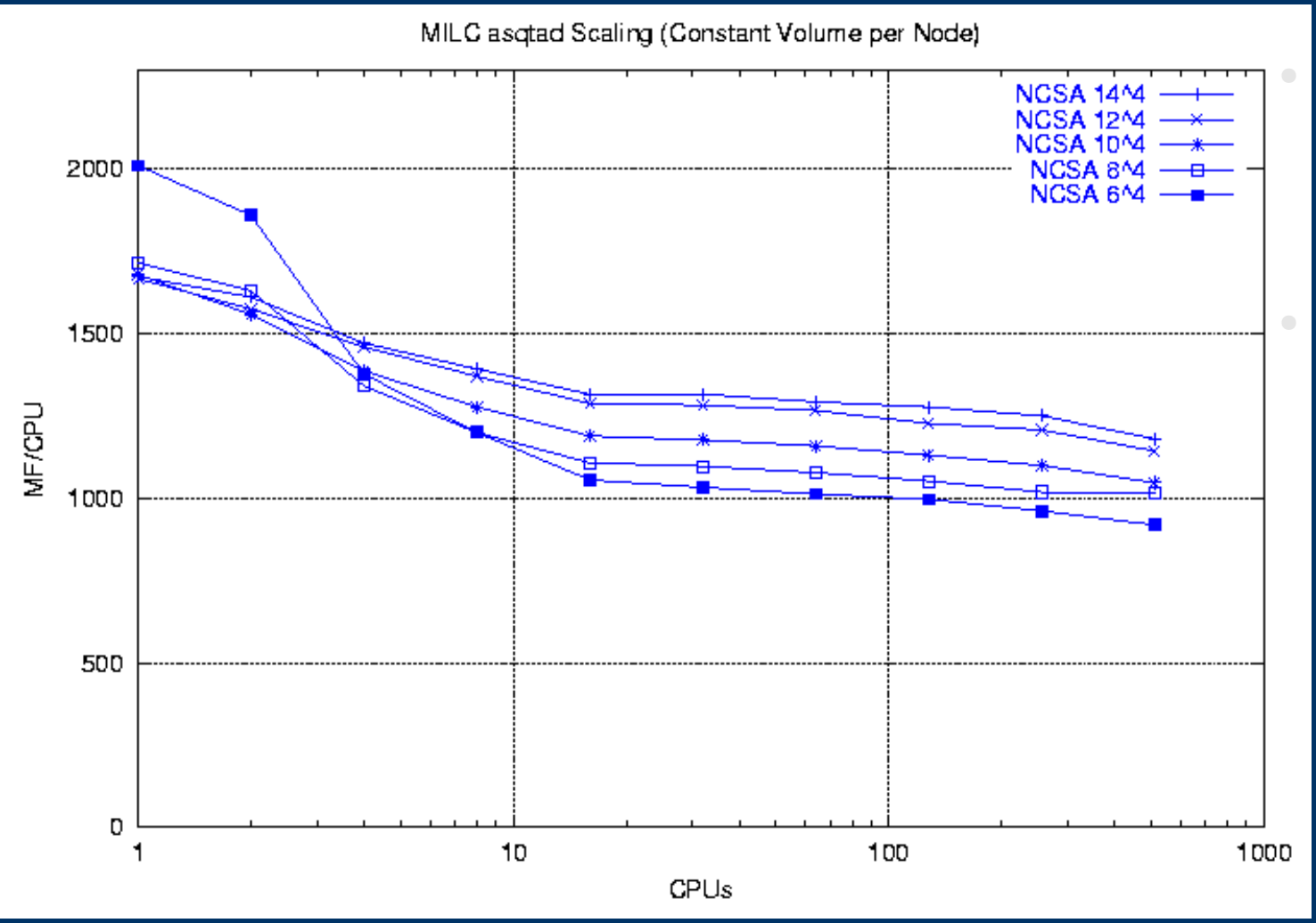
SciDAC MILC code

Cluster still being commissioned

- 256 nodes to be expanded to 512 by October

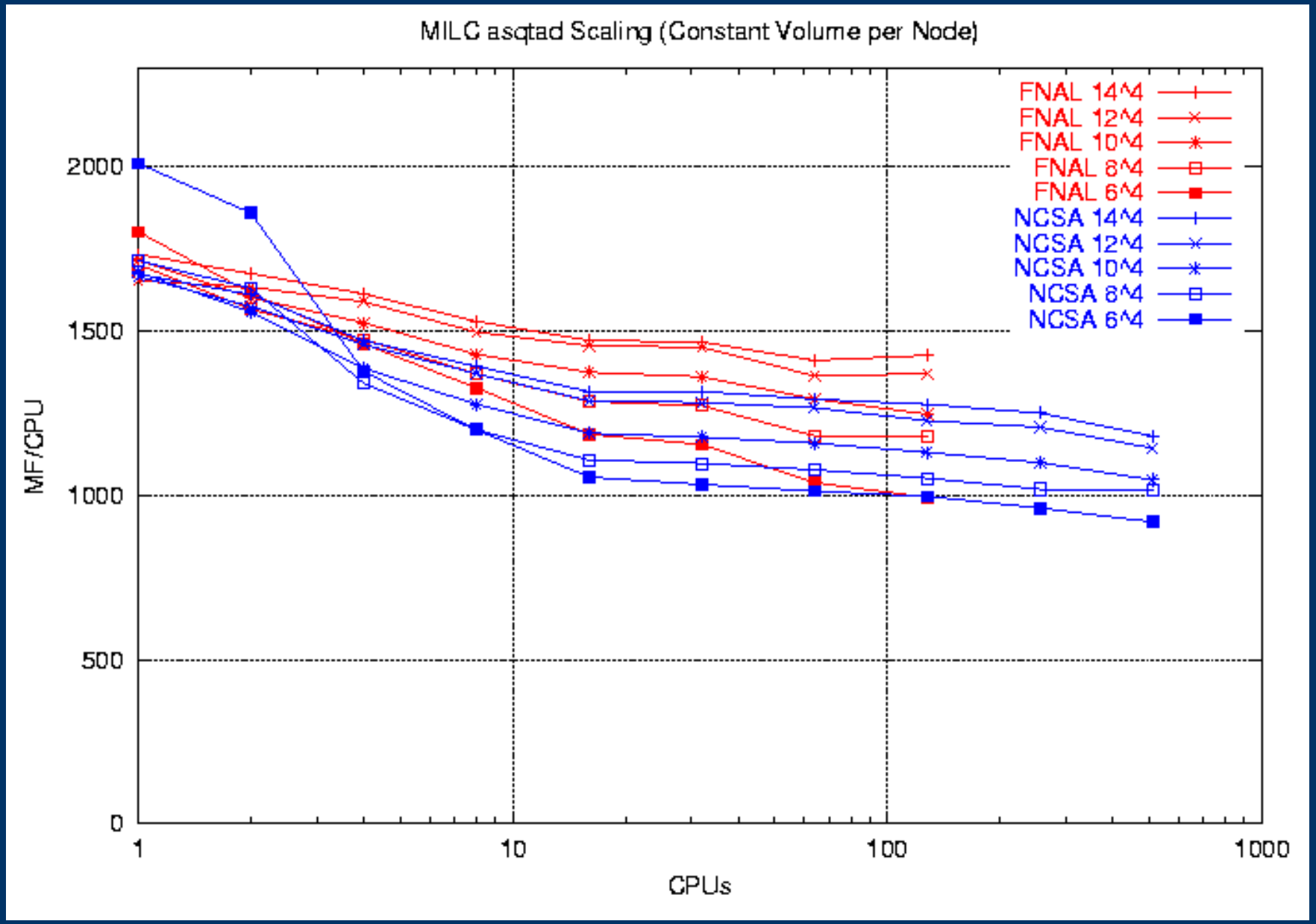
Scaling to O(1000) nodes???

Expectations



- NCSA “T2” Cluster:
  - 3.6 GHz Xeon
  - Infiniband (3:1)
  - PCI-X
- Non-SciDAC version of MILC code

Expectations



Expectations

- Late FY06 (FNAL), based on FY05
  - 1066 memory bus would give 33% boost to single node performance
    - AMD will use DDR2-667 by end of Q2
    - Intel already sells (expensive) 1066 FSB chips
  - SciDAC code improvements for x86_64
  - Modify SciDAC QMP for Infiniband
  - 1700-1900 MFlops per processor
  - $700 (network) + $1100 (total system)
  - Approximately $1/MFlop for asqtad

Predictions

- Large clusters will be appropriate for gauge configuration generation (1 Tflop/s sustained) as well as for analysis computing
- Assuming 1.5 GFlop/node sustained performance, performance of MILC fine and superfine configuration generation:

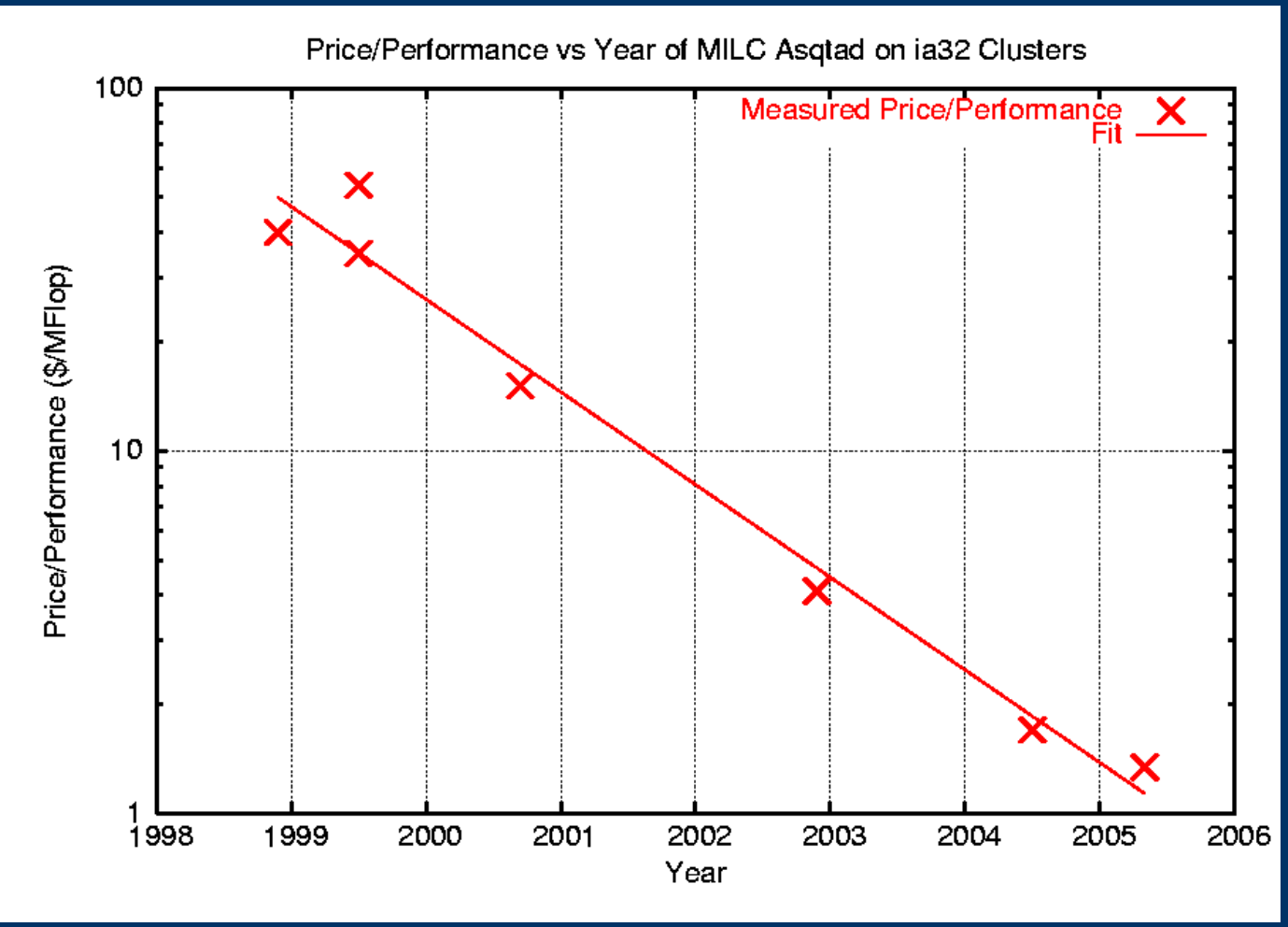
Lattice Size	Sublattice	Node Count	TFlop/sec
40^3 x 96	10^3 x 12	512	0.77
	10^3 x 8	768	1.15
	8^3 x 8	1500	2.25
56^3 x 96	14^3 x 12	512	0.77
	8^3 x 12	2744	4.12
60^3 x 138	12^3 x 23	750	1.13
	10^3 x 23	1296	1.94

Conclusion

- Clusters give the best price/performance in FY 2006
  - We've generated our performance targets for FY 2006 – FY 2009 in the project plan based on clusters
  - We can switch in any year to any better choice, or mixture of choices

Extra Slides

Performance Trends - Clusters



Updated graph

- Includes FY04 (P4E/Myrinet) and FY05 (Pentium 640 and Infiniband) clusters

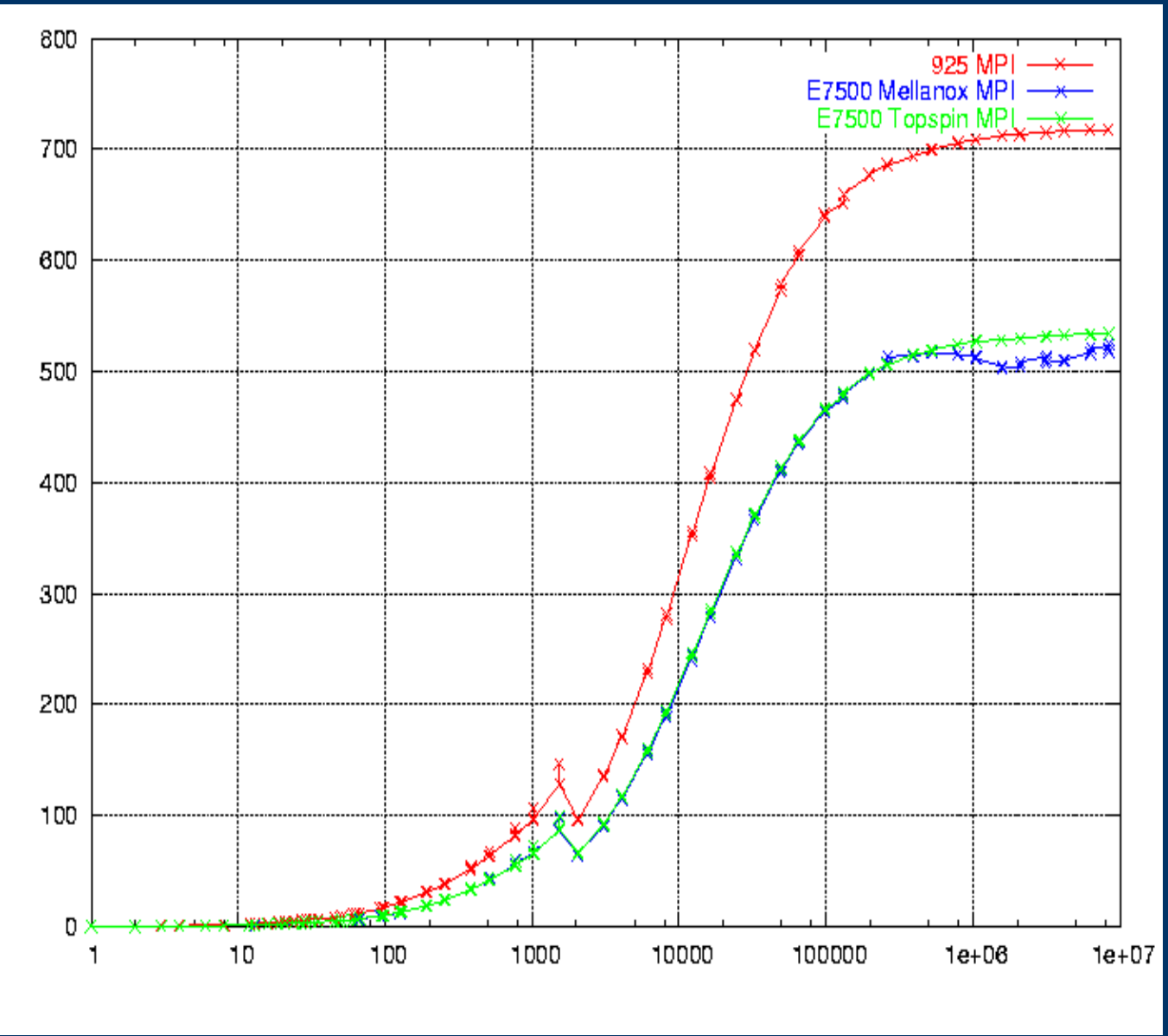
Halving Time:

- Price/Perf: 1.18 years

Beyond FY06

- For cluster design, will need to understand:
  - Fully buffered DIMM technology
  - DDR and QDR Infiniband
  - Dual and multi-core CPUs
  - Other networks

Infiniband on PCI-X and PCI-E



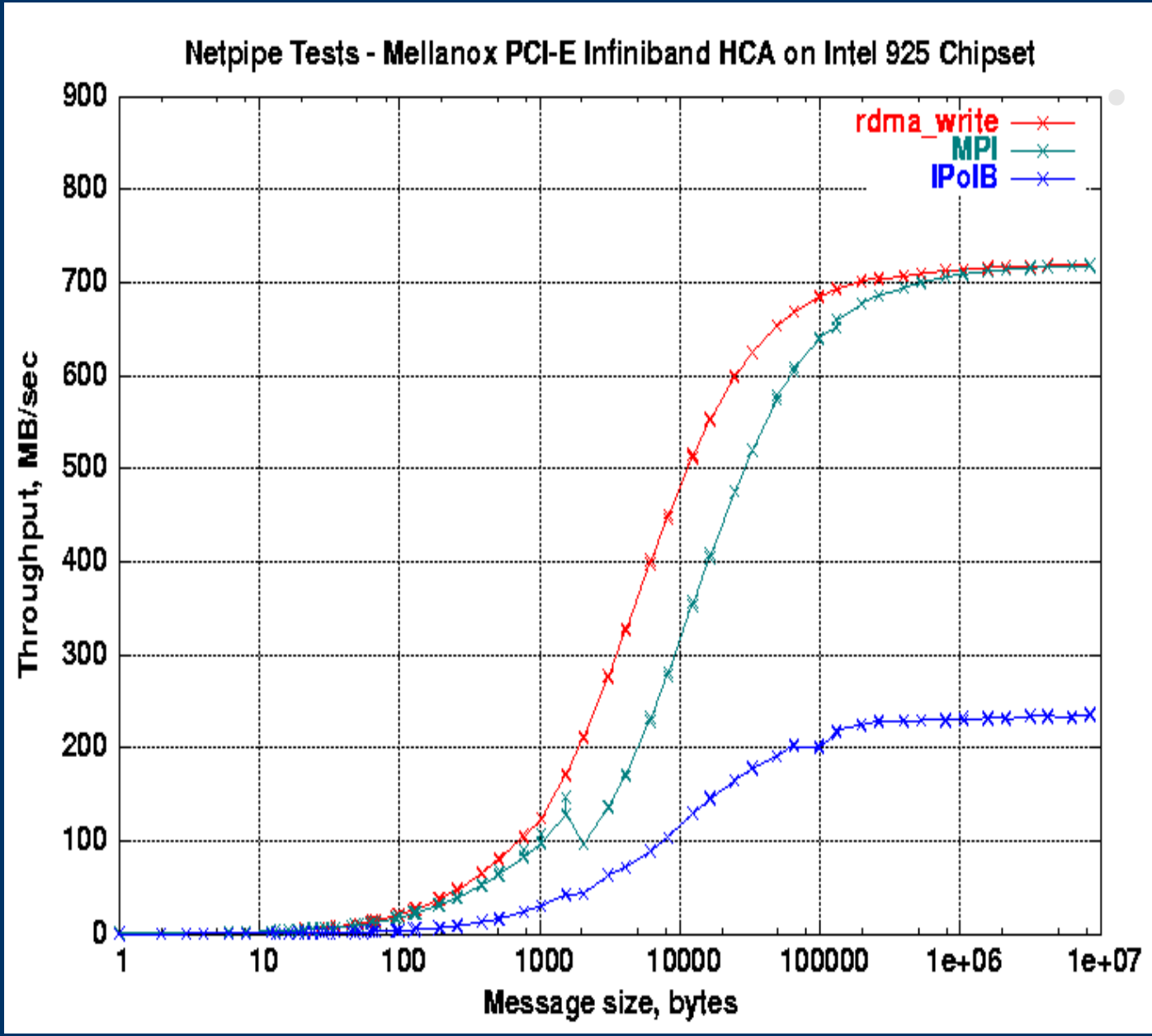
Unidirectional bandwidth (MB/sec) vs message size (bytes) measured with MPI version of Netpipe

- PCI-X (E7500 chipset)
- PCI-E (925X chipset)

PCI-E advantages:

- Bandwidth
- Simultaneous bidirectional transfers
- Lower latency
- Promise of lower cost

Infiniband Protocols



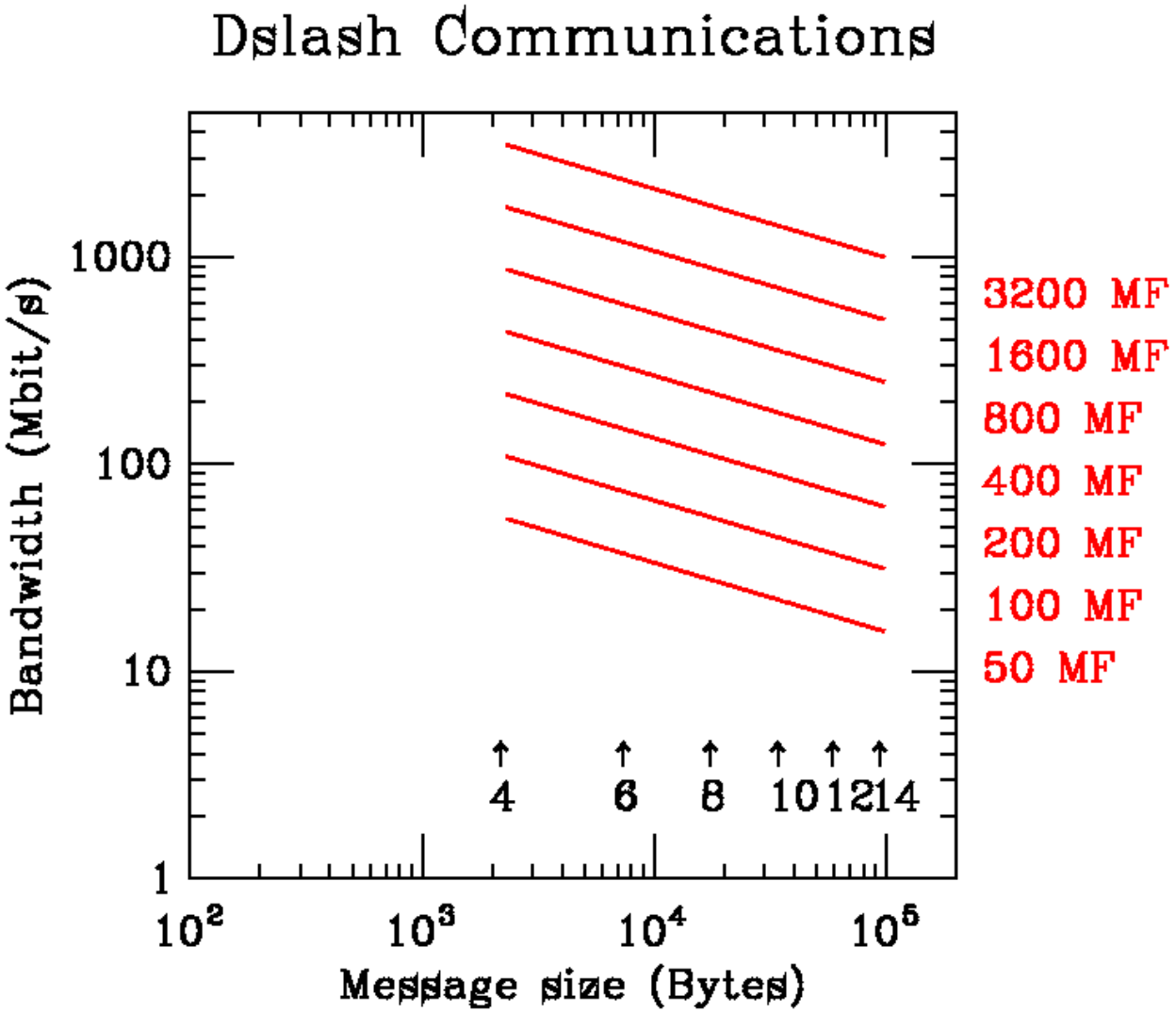
Netpipe results, PCI-E HCA's using these protocols:

- "rdma_write" = low level (VAPI)
- "MPI" = OSU MPI over VAPI
- "IPoIB" = TCP/IP over Infiniband

Recent Processor Observations

- Using MILC “Improved Staggered” code, we found:
  - 90nm Intel chips (Pentium 4E, Pentium 640), relative to older Intel ia32:
    - In-cache floating point performance decrease
    - Improved main memory performance (L2=2MB on '640)
    - Prefetching is very effective
  - dual Opterons scale at nearly 100%, unlike Xeons
    - must use NUMA kernels + *libnuma*
    - single P4E systems are still more cost effective
  - PPC970/G5 have superb double precision floating point performance
    - but – memory bandwidth suffers because of split data bus. 32 bits read only, 32 bits write only – numeric codes read more than they write

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:

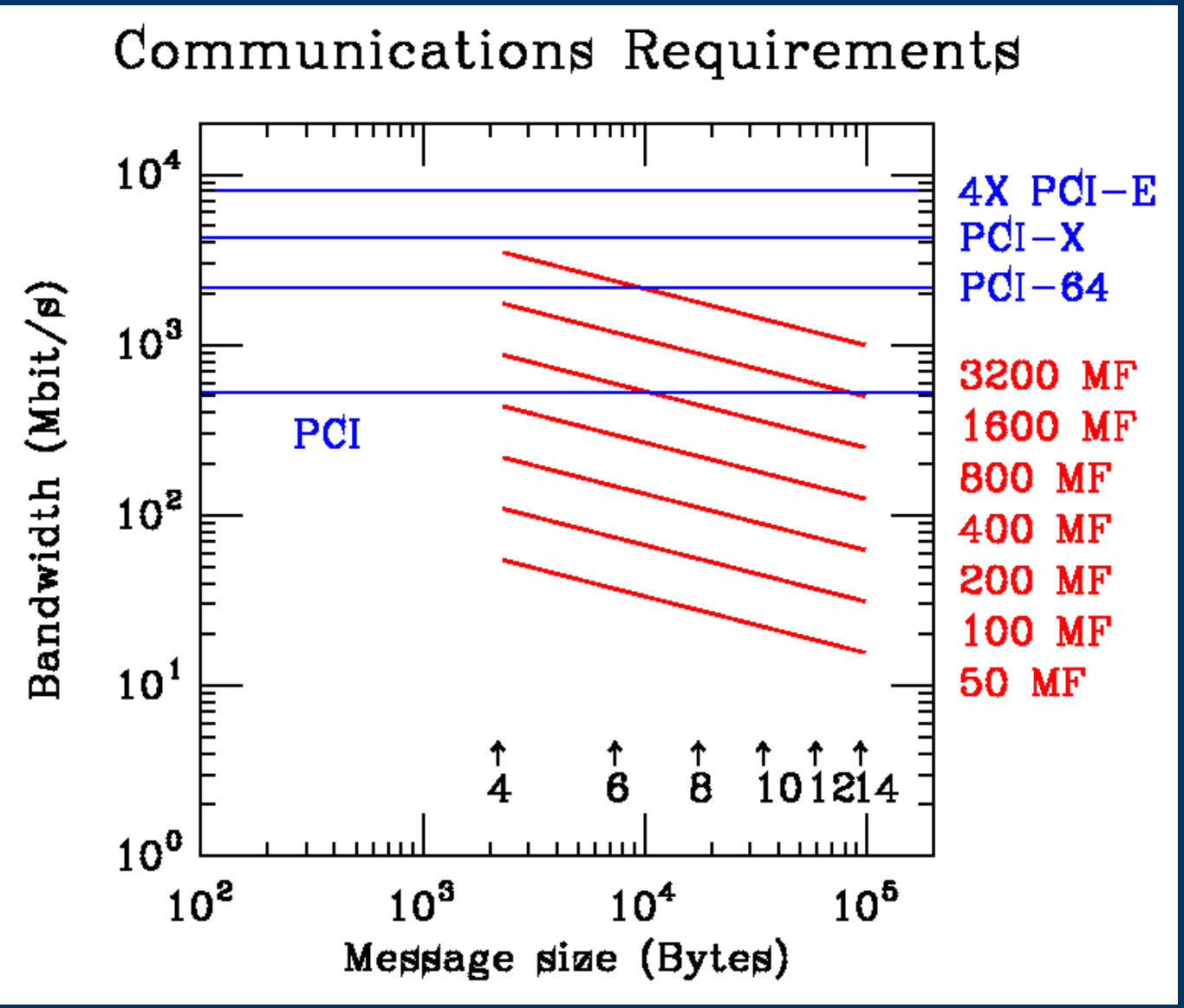
- L implies message size to communicate a hyperplane
- Sustained MFlop/sec together with message size implies achieved communications bandwidth

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

Balanced Design Requirements - I/O Bus Performance

- Connection to network fabric is via the “I/O bus”
- Commodity computer I/O generations:
  - 1994: PCI, 32 bits, 33 MHz, 132 MB/sec burst rate
  - ~1997: PCI, 64 bits, 33/66 MHz, 264/528 MB/sec burst rate
  - 1999: PCI-X, Up to 64 bits, 133 MHz, 1064 MB/sec burst rate
  - 2004: PCI-Express 4X = 4 x 2.0 Gb/sec = 1000 MB/sec
    16X = 16 x 2.0 Gb/sec = 4000 MB/sec
- *N.B.*
  - PCI, PCI-X are *buses* and so unidirectional
  - PCI-E uses *point-to-point pairs* and is bidirectional
    - So, 4X allows 2000 MB/sec bidirectional traffic
- PCI chipset implementations further limit performance
  - See: http://www.conservativecomputer.com/myrinet/perf.html

I/O Bus Performance



Blue lines show peak rate by bus type, assuming balanced bidirectional traffic:

- PCI: 132 MB/sec
- PCI-64: 528 MB/sec
- PCI-X: 1064 MB/sec
- 4X PCI-E: 2000 MB/sec

Achieved rates will be no more than perhaps 75% of these burst rates

PCI-E provides headroom for many years