PC Clusters for Lattice QCD

Don Holmgren
djholm@fnal.gov
Fermilab

Lattice'04
June 23, 2004

PC Clusters for Lattice QCD

- The charge from the committee:
  - Last year's talk reviewed clusters for QCD deployed around the world, so this year please focus on new issues:
    - How has price/performance improved?
    - How is price/performance expected to improve in the future?
    - What are the challenges to building clusters of many thousands of nodes?

Outline

- Brief update on new QCD clusters
- Following the charge, examine:
  - Performance Trends
  - Explaining the Trends - Requirements for Balanced Designs
  - Costs
  - Limits to Cluster Sizes
  - Predictions

New Clusters since Lattice'03

- At Lattice'03, Thomas Lippert gave a very good and thorough review of current deployments:
  http://www.rccp.tsukuba.ac.jp/lat03/Ana/Ple-Dat/transparency/lippert/lippert.html
- Major new cluster deployments:
  - University Budapest, Hungary
    - 320 nodes, Pentium 4 processors
    - 2-dimensional gigabit Ethernet mesh
  - Wuppertal, Germany
    - 512 nodes, 1.5 GHz dual Athlon Opteron processors
    - 2 GB memory per node
    - 2-dimensional gigabit Ethernet mesh
    - Additional hierarchal switched gigabit Ethernet network
    - Parastation software – simultaneous use of mesh and switched networks
    - Approximately $2K/node

New Clusters since Lattice'03 – continued

- Jefferson Lab, Virginia, USA
  - 256 nodes, 2.66 GHz Xeon processors, E7501 chipset
  - single processors used in dual motherboards
  - 256 MB memory per node
  - 3-dimensional gigabit Ethernet mesh
  - additional switched gigabit Ethernet control network
  - approximately $1950/node including mesh
- Fermilab, Illinois, USA
  - 128 nodes, single 2.8 GHz Pentium 4E processor
  - 1 GB memory per node
  - reusing Myrinet LANai-9 fabric purchased in 2000
  - $900/node without Myrinet
  - Myrinet cost today would be $850/node

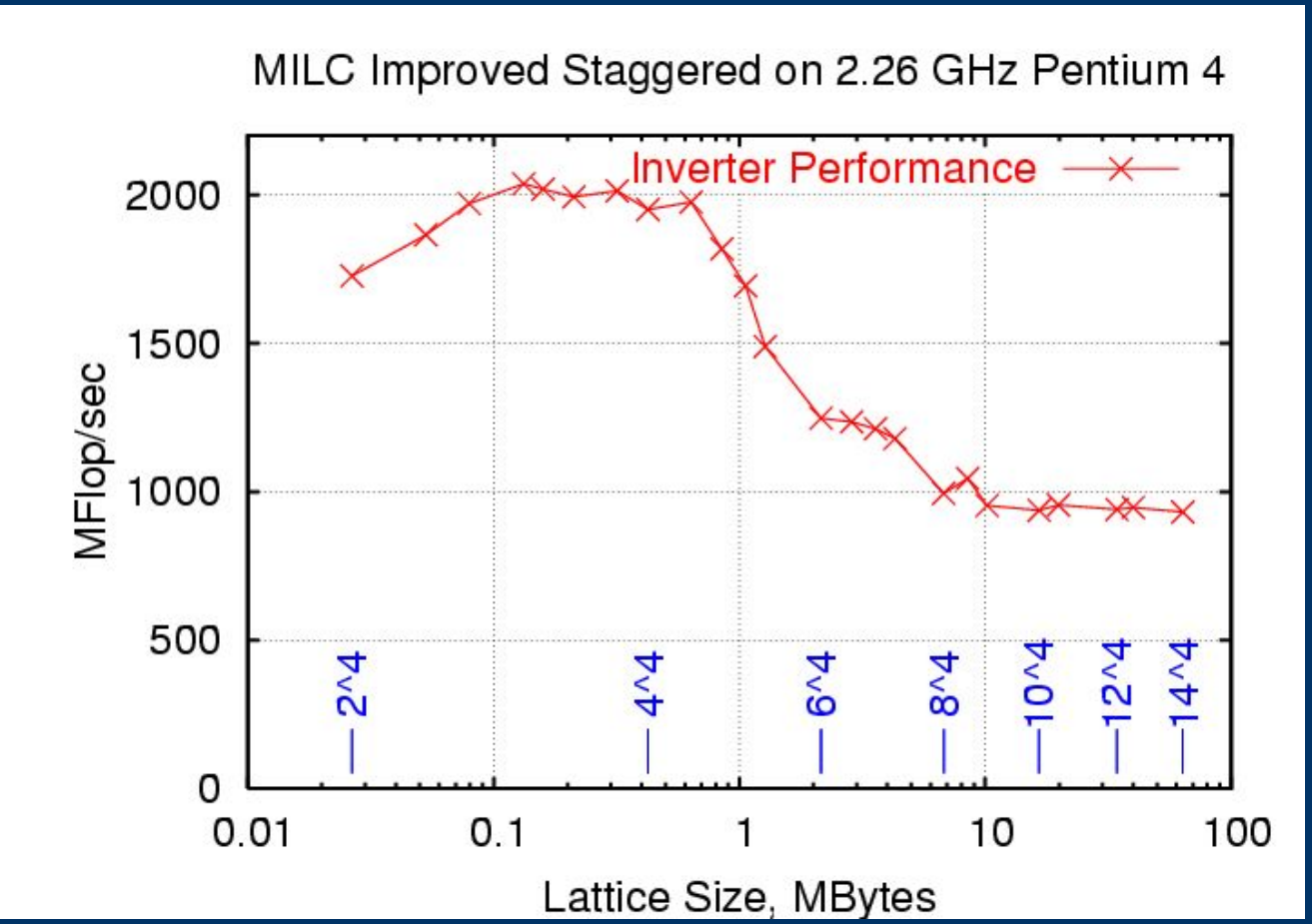
Boundary Conditions

- Simplifications to make the talk fit the time:
  - I will only discuss in detail Intel processors
    - AMD will be mentioned
    - I'm happy to discuss other processors at the break
    - For other processor and network results, see http://lqcd.fnal.gov/benchmarks/
  - Performance results will be from MILC “asqtad” codes
    - Single precision only - see Carleton Detar's talk: http://thy.phy.bnl.gov/www/scidac/presentations/detar.pdf
  - The trends discussed are not dependent upon the specific choices of hardware or action

Some Definitions

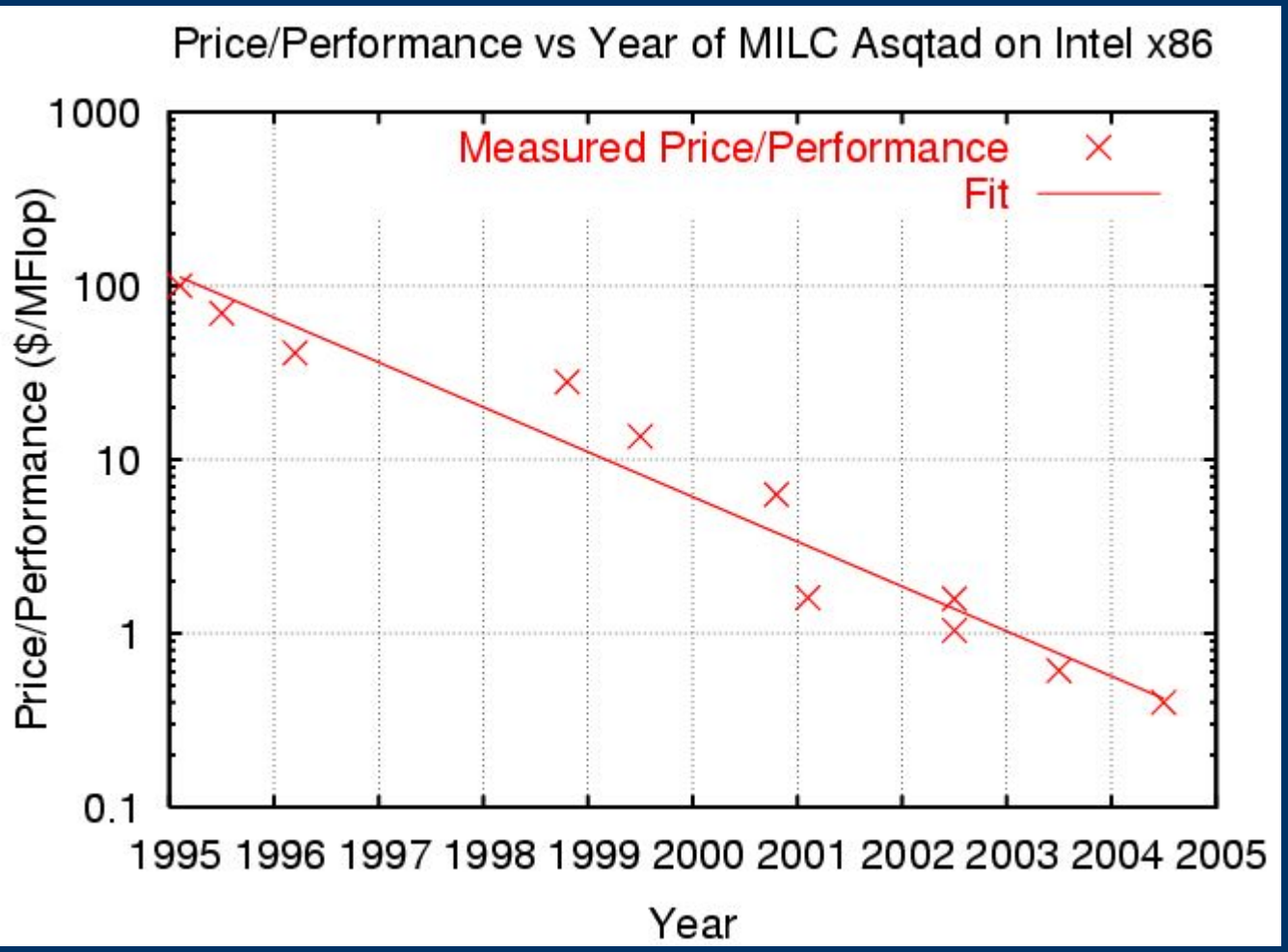
- Common jargon used in PC discussions:
  - “FSB” = front side bus
    = effective clock speed of the memory bus
  - “P4” = Pentium 4, always uniprocessor
  - “P4E” = Pentium 4E, or “Prescott”, always uniprocessor
  - “Xeon” = SMP-capable P4
  - “SSE” = Intel's SIMD instruction set (also SSE2 & SSE3)

Generic Single Node Performance



- Cache size = 512 KB
- Floating point capabilities of the CPU limits in-cache performance
- Memory bus limits performance out-of-cache

Performance Trends – Single Node



MILC Improved Staggered Code ("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

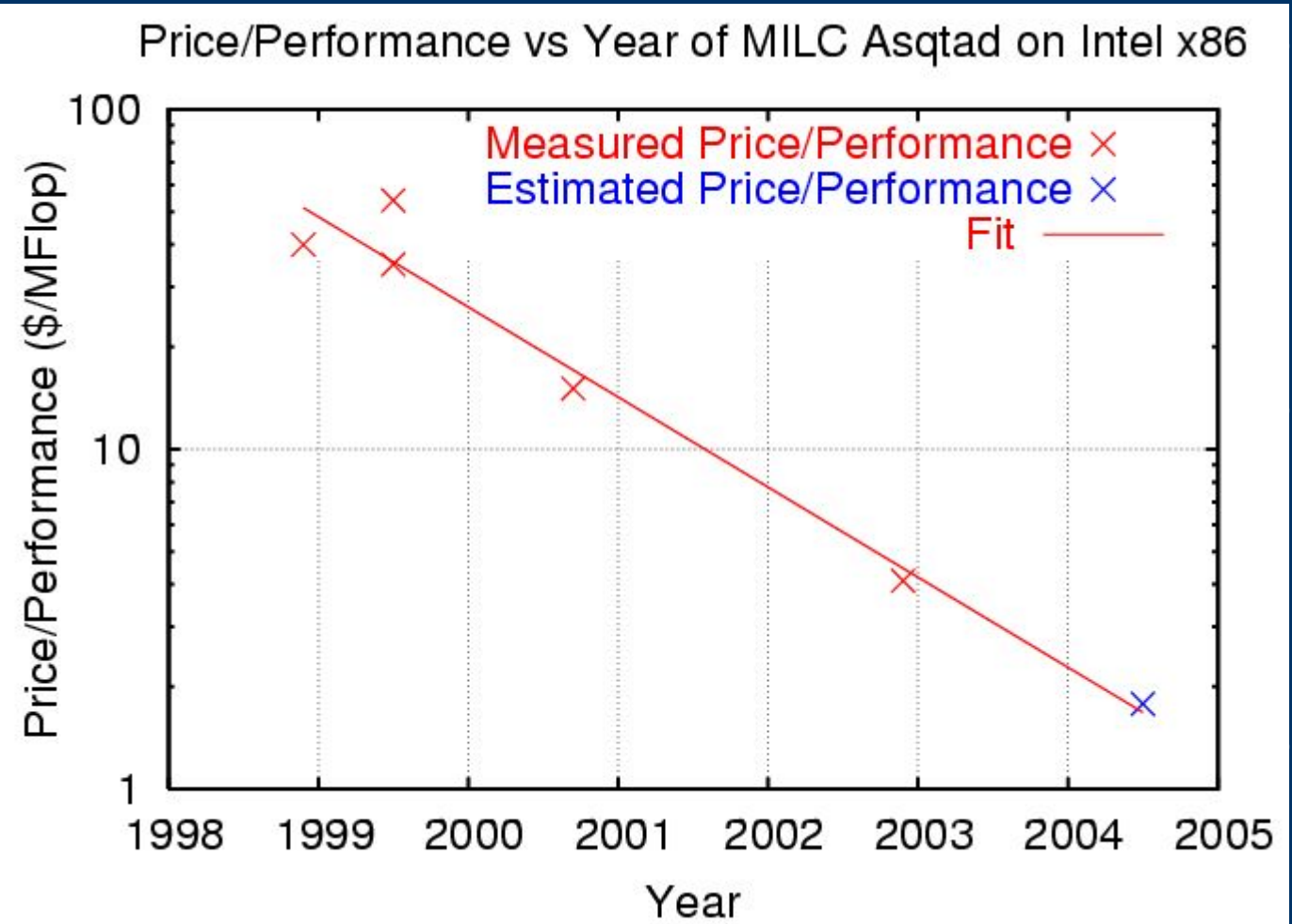
Performance range:

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

Doubling times:

- Performance: 1.88 years
- Price/Perf.: 1.19 years !!

Performance Trends - Clusters



Clusters based on:

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

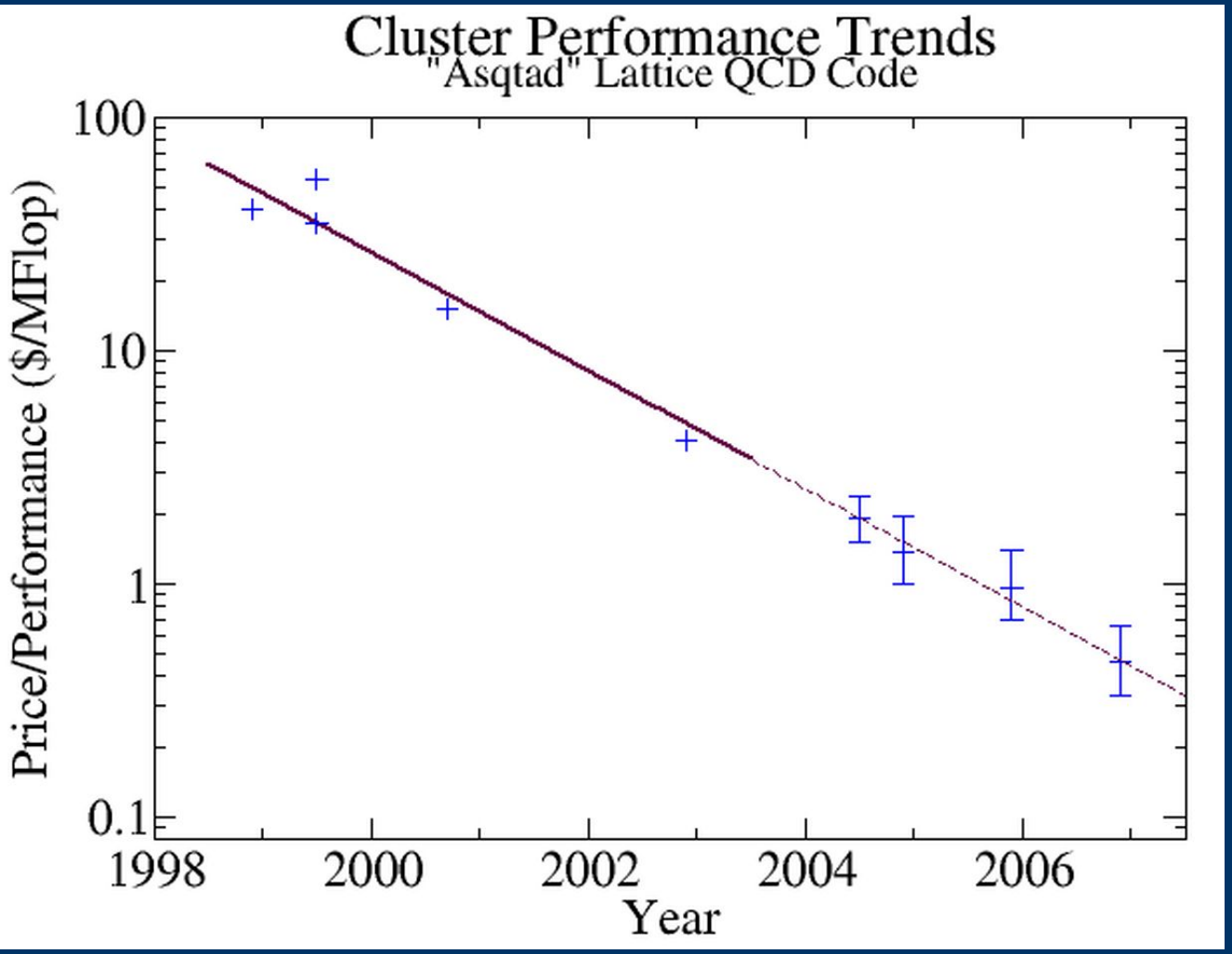
Performance range:

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

Doubling Times:

- Performance: 1.22 years
- Price/Perf: 1.25 years

Predictions



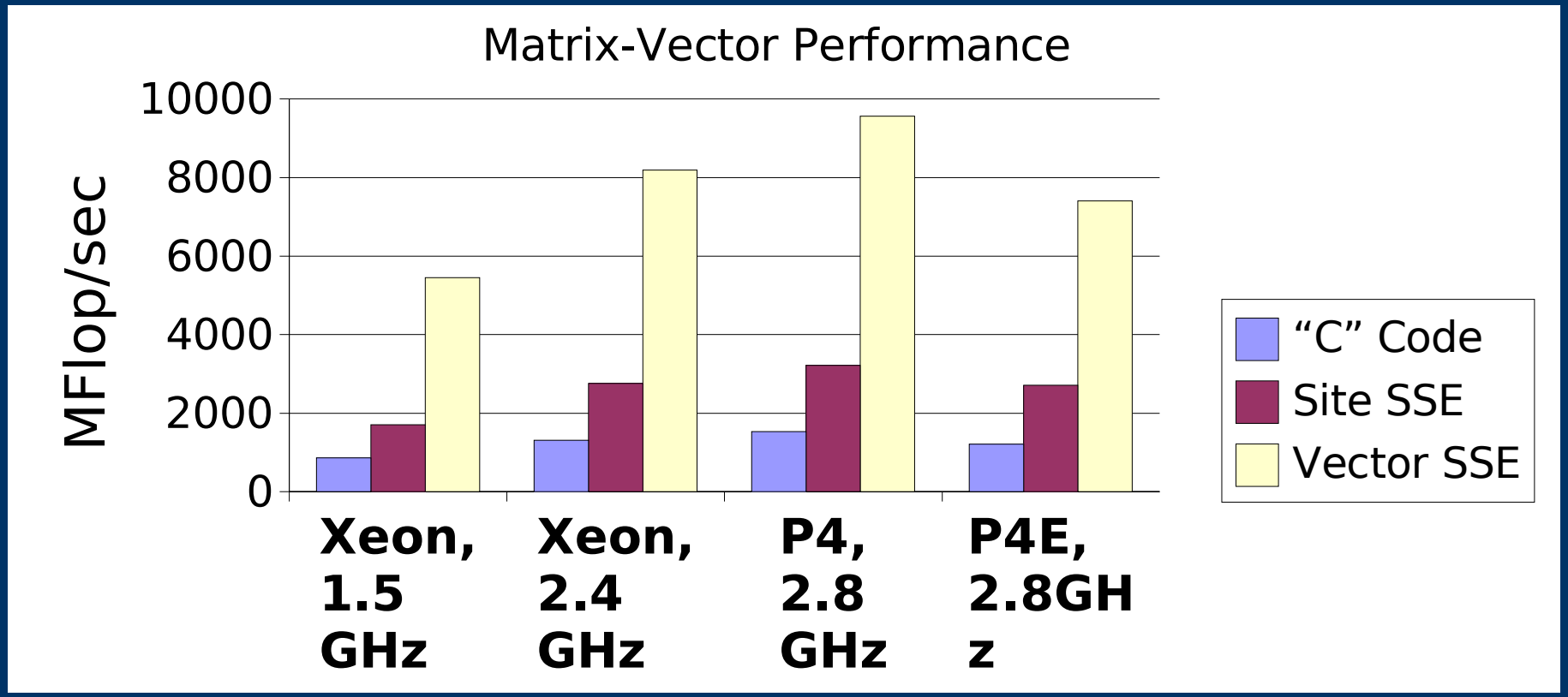
- The four extrapolated points are conservatively based upon vendor roadmaps, and upon historical trends
- The rest of the talk explains these trends and predictions

Balanced Design Requirements
Dirac Operator

- Dirac operator (*Dslash*) – improved staggered action ("asqtad")
  - 8 sets of 2 matrix-vector multiplies in negative direction
  - 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
  - The bottleneck varies with local lattice size, algorithm
  - We prefer floating point performance to be the bottleneck
    - Unfortunately, memory bandwidth is the main culprit
    - Balanced designs require a careful choice of components

Balanced Design Requirements
Floating Point Performance

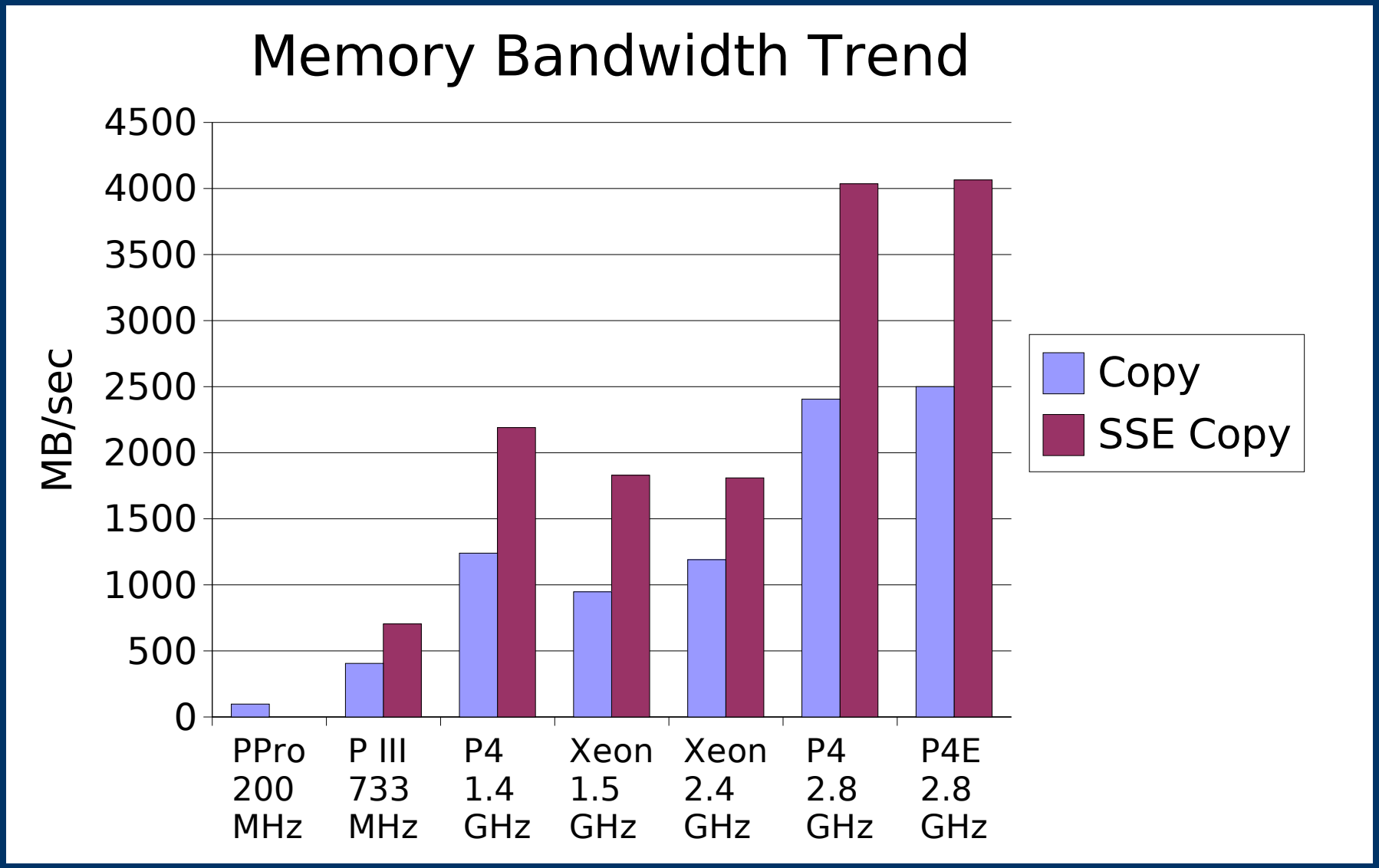
- Most flops are SU3 matrix times vector
  - SSE/SSE2/SSE3 can give a significant boost
    - Site-wise (M. Lüscher)
    - Fully vectorized (A. Pochinsky)



Balanced Design Requirements - Memory Performance

- Memory bandwidth limits – depends on:
  - Width of data bus
  - (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

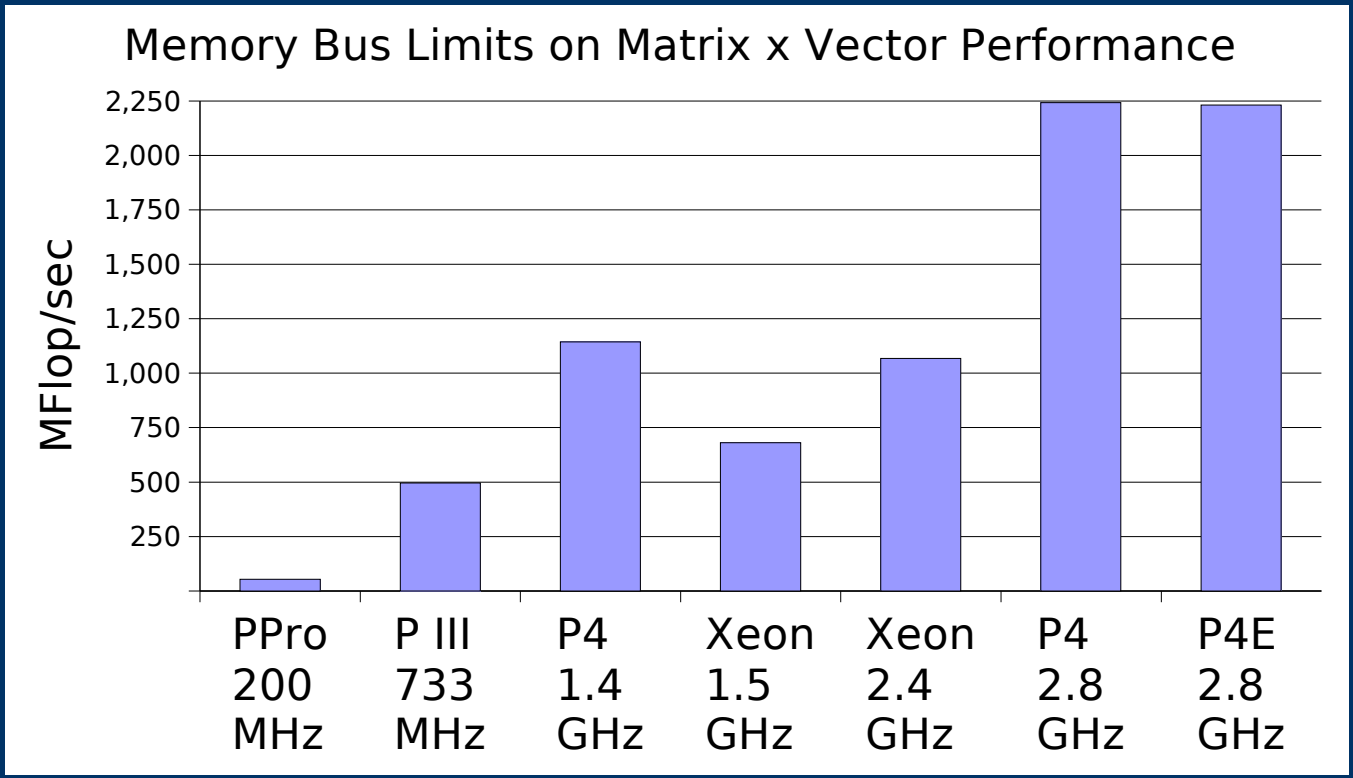

Memory Bandwidth Trend
MB/sec
4500
4000
3500
3000
2500
2000
1500
1000
500
0
PPro 200 MHz
P III 733 MHz
P4 1.4 GHz
Xeon 1.5 GHz
Xeon 2.4 GHz
P4 2.8 GHz
P4E 2.8 GHz
Copy
SSE Copy

Memory Bandwidth Performance
Limits on Matrix-Vector Algebra

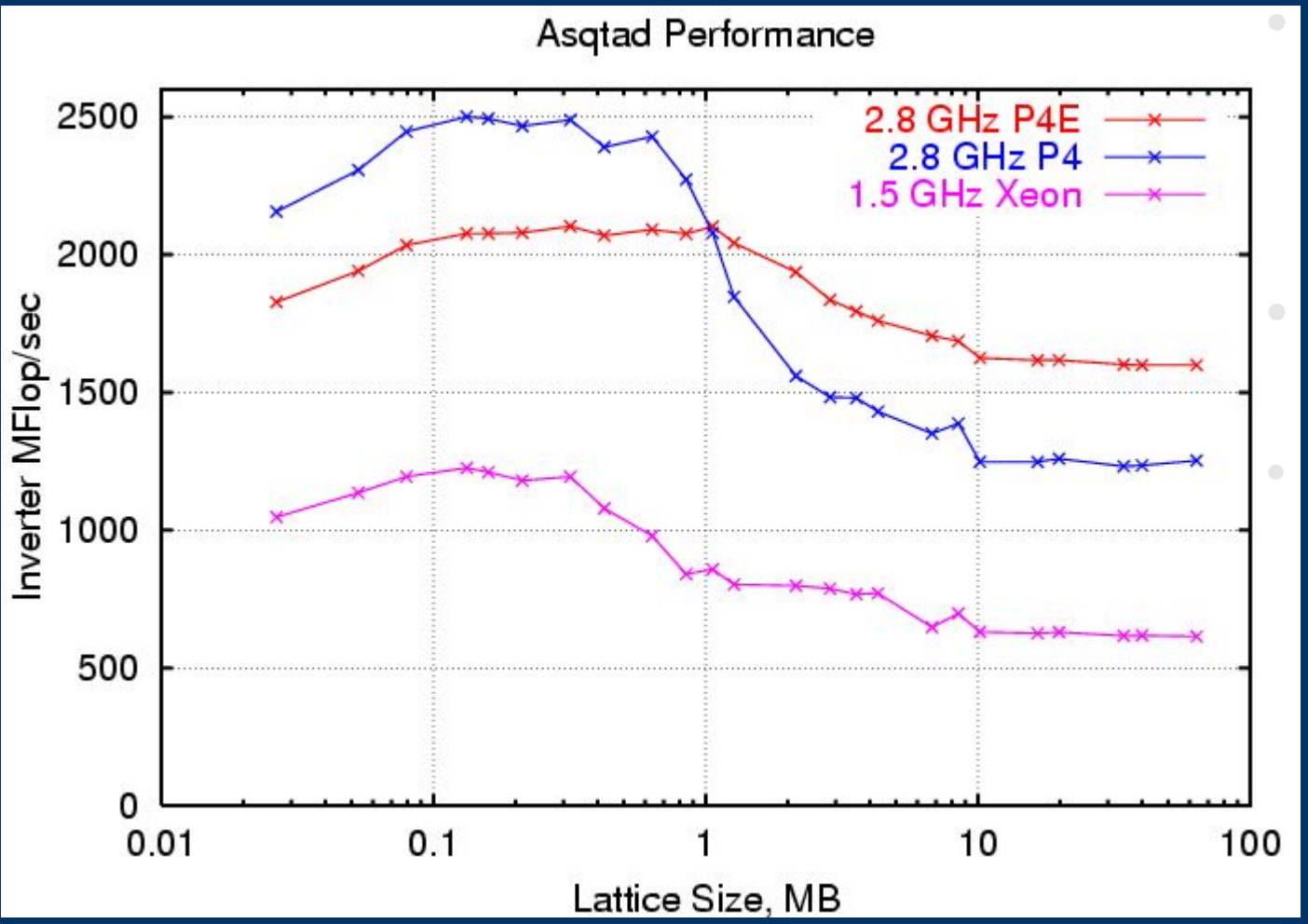
- From memory bandwidth benchmarks, we can estimate sustained matrix-vector performance in main memory
- We use:
  - 66 Flops per matrix-vector multiply
  - 96 input bytes
  - 24 output bytes
  - MFlop/sec = 66 / (96/read-rate + 24/write-rate)
    - read-rate and write-rate in MBytes/sec
- Memory bandwidth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

Processor	FSB	Copy	SSE Read	SSE Write	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

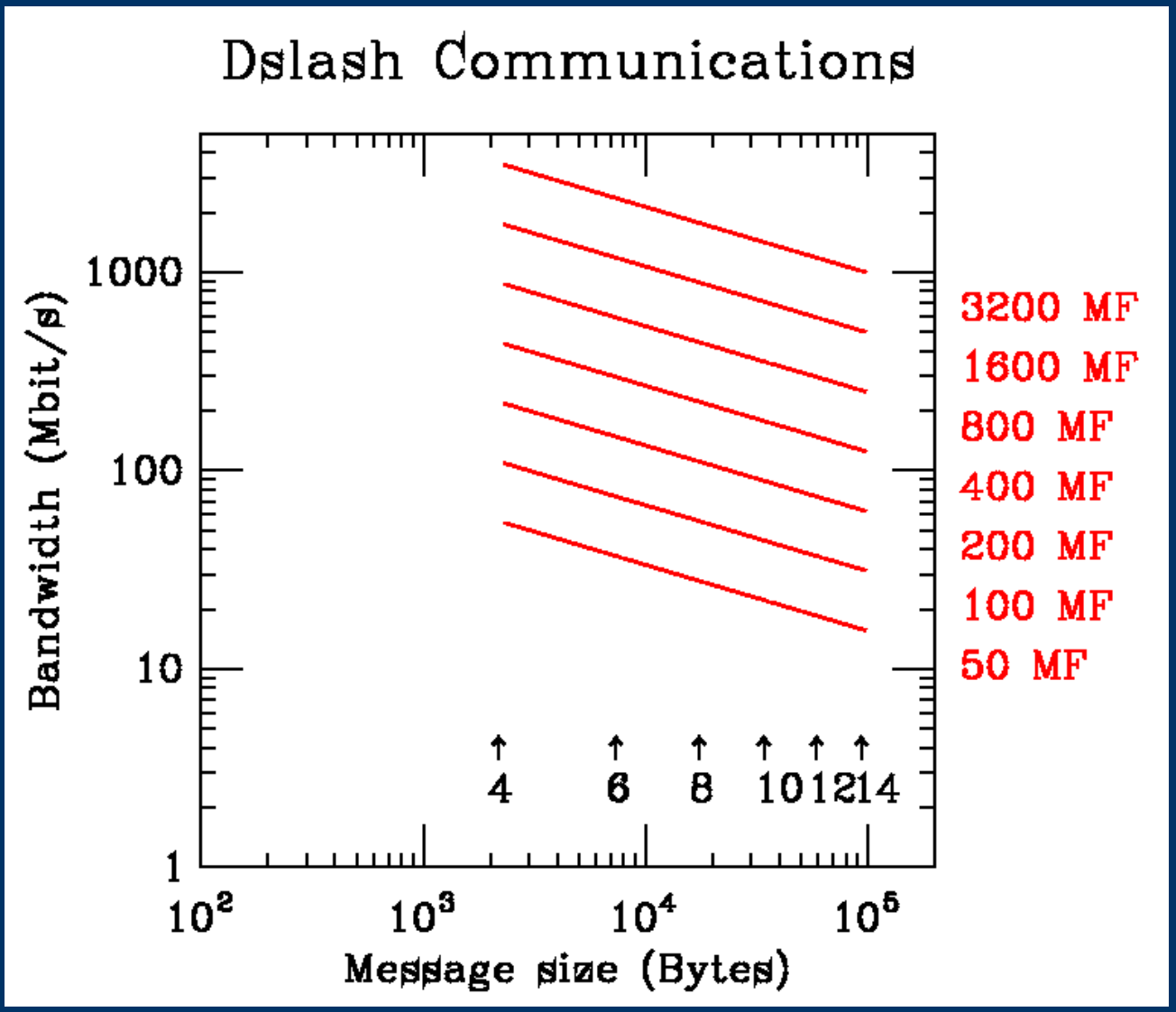


Performance vs Architecture



- Memory buses:
  - Xeon: 400 MHz
  - P4: 800 MHz
  - P4E: 800 MHz
- P4 vs Xeon shows effects of faster FSB
- P4 vs P4E shows effects of change in CPU architecture
  - P4E has better heuristics for hardware memory prefetch, but longer instruction latencies

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

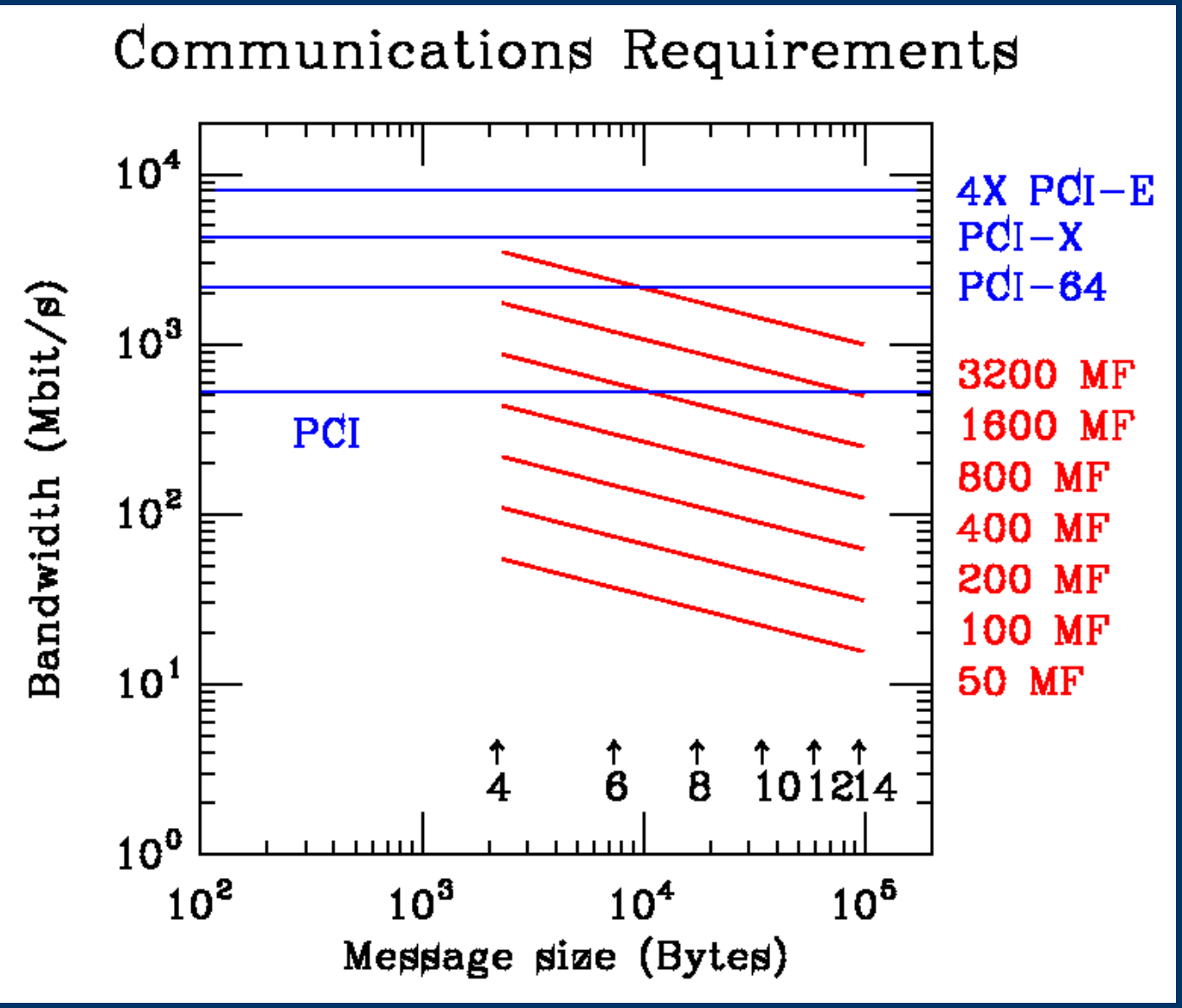
Then:
- L implies message size to communicate a hyperplane
- Achieved MFlop/sec with message size implies achieved communications bandwidth

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 peak burst rates

PCI-E provides headroom for many years

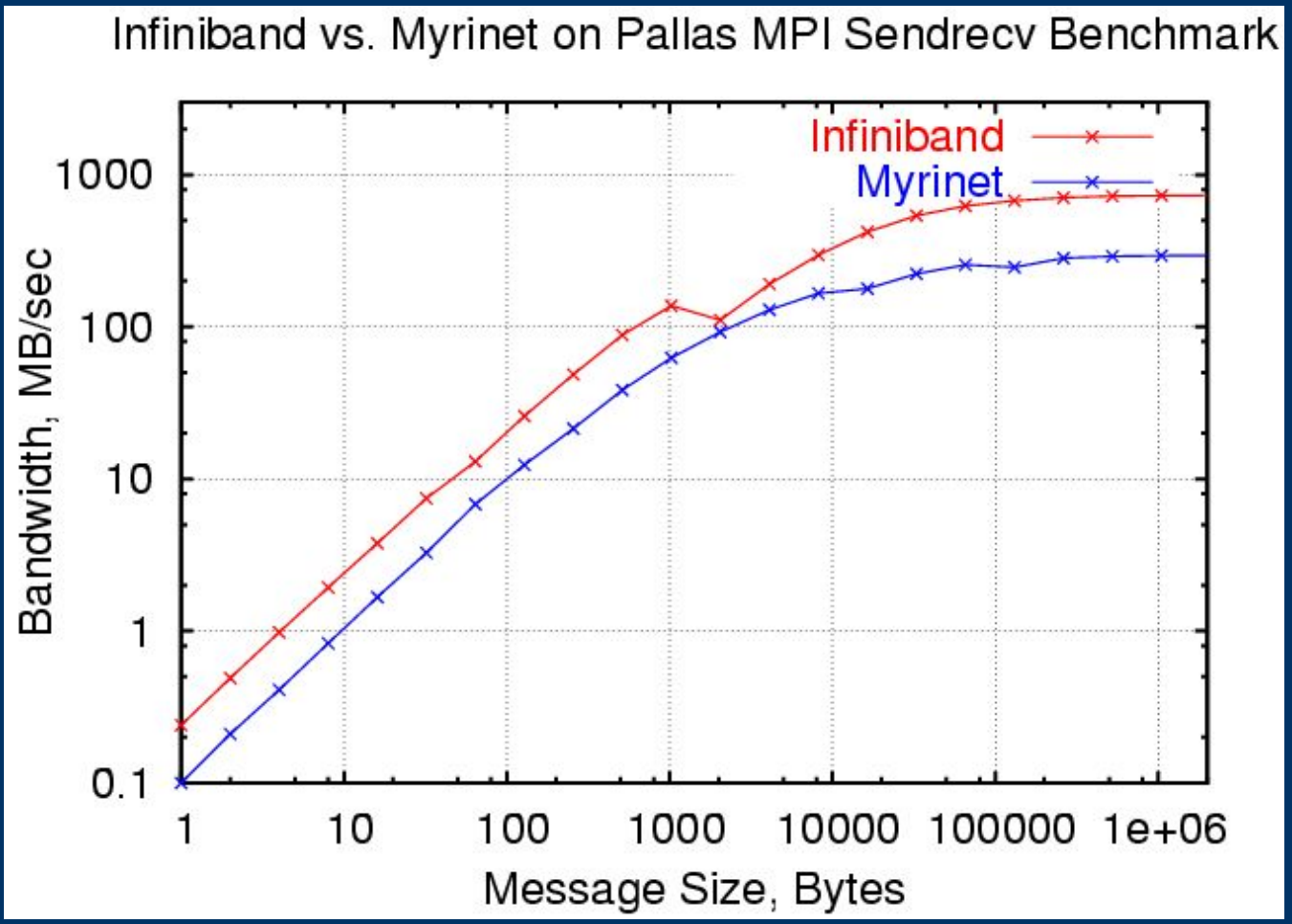
Balanced Design Requirements
Network Performance

- Network fabric choices:
  - Ethernet (switched or mesh fabric)
    - GigE now (125 MB/sec bidirectional)
    - 10 GigE - emerging but expensive (1250 MB/sec/dir)
    - TCP/IP bypass such as VIA needed for low latency and to lower processor overhead
    - Switches add latency, large switches are costly
    - Meshes have good latency (< 20µsec), bandwidth
  - Myrinet (switched fabric)
    - 2.0 Gb/sec physical layer = 250 MB/sec/direction
    - Channel bond for higher rates
    - MPI latencies as low as 6.3µsec now, 3.5µsec soon
  - Quadrics (switched fabric)
    - ELAN4: 1.8 µsec latency, 1000 MB/sec/direction

Balanced Design Requirements - Network Performance

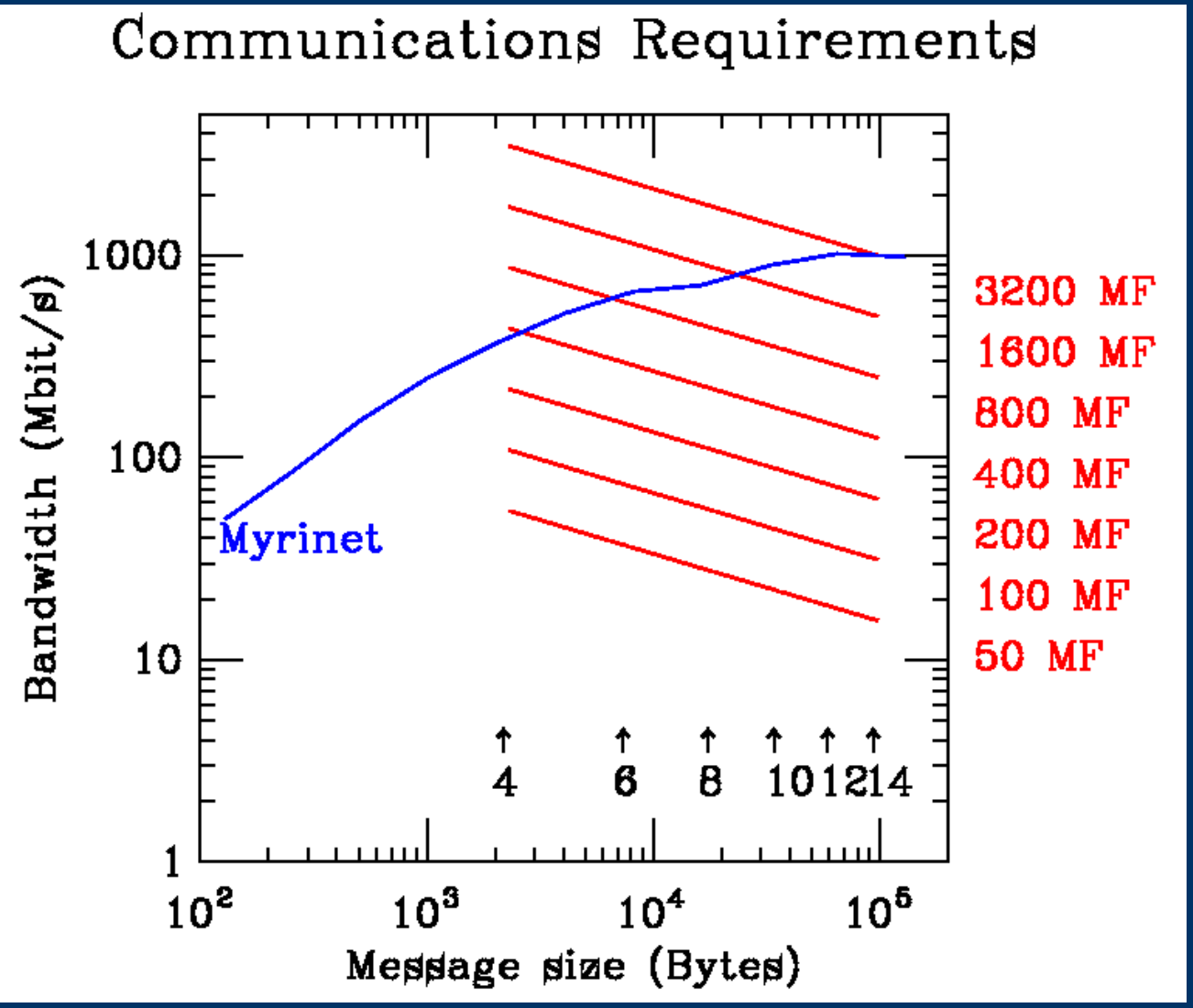
- (Slowly) emerging fabric: Infiniband (switched fabric)
  - 4X = 8.0 Gb/sec = 1000 MB/sec/direction
  - 12X = 3000 MB/sec/direction
  - 4X cards (“HCA” = host channel adapter), most with two 4X ports
  - 12X available now to interconnect switches
  - MPI latencies now about 6 µsec (PCI-X)
    - 4 µsec expected for PCI-E
  - Multiple applications, unlike Quadrics, Myrinet:
    - HPC fabrics
    - Storage (fiber channel, iSCSI)
    - Bridging (gigE)
    - Vendors believe data mining will be biggest market
  - Has driven HPC network fabric costs down
  - HCAs may be integrated on motherboards soon, further dropping cost

Network Performance
Bandwidth



- Typical performance curves:
  - Myrinet LANai-9 on E7500 PCI-X
  - Infiniband 4X on E7501 PCI-X
- Infiniband feature at 2048 byte message size is MPI “eager-rendezvous” threshold
- Performance drops with message size

Balanced Design Requirements
Dslash and the Network



Blue curve: measured Myrinet (LANai-9) performance on Fermilab dual Xeon cluster

This gives a very optimistic upper bound on performance – actual performance will be affected by:

- actual message sizes are smaller than modeled
- competition for memory bus
- competition for I/O bus
- processor overheads for performing the communication

Curvature of network performance curve limits the practical cluster size

Costs
Node Costs

- Single CPU systems
  - Cheapest type of system – sold in huge volumes as desktops and home machines
  - By far the best price/performance for single node calculations
  - Fastest memory bus of all Intel x86 choices
    - 800 MHz FSB since 2003
    - 1066 MHz FSB in 3rd quarter 2004
  - Prior to 2004, often a poor choice for clusters because of low performance (32 bit, 33 MHz) I/O bus
  - Current price (May 2004 Fermilab purchase): $900
    - 2.8 GHz P4E processor
    - 1 GB DDR3200 memory
    - PCI-X (less than 64-bit, 66 MHz performance)
    - “2U” case

Costs
Node Costs

- SMP (dual CPUs)
  - Less than 2X the cost of a uniprocessor node
  - Lower cost/processor than uniprocessor nodes – sold in volume as low- and mid-range servers
  - Excellent I/O bus implementations
    - but, always measure before buying!
  - Slower memory bus than Intel x86 uniprocessor nodes
    - 533 MHz FSB since 2003
    - 800 MHz FSB this month (June 2004)
    - FSB speed is limited because CPUs share the bus
      - AMD Opteron fixes this problem – 1 bus/processor
    - Now: ~ $1600
      - 2.66 GHz Xeon processors
      - 256 MB DDR2100 memory
      - PCI-X
      - “1U” system case

Costs
Network Costs

- Ethernet
  - Network interfaces are free (integrated on motherboard)
  - For meshes, 2-way cards are about $150 each
  - Large switches are expensive
- Myrinet
  - 256-port fabric: switch, network interface cards, cables
  - List: $950/node Street: $850/node
  - Note that fast processors may require bonded ports
    - boosts price/node by ~ $700
- Infiniband
  - Building blocks are 24, 72, 144, and 288 port switches
  - No real market yet, so prices may fluctuate
  - Expect $900 - $1000 per node now
  - Excess bandwidth

Limits to Cluster Size

- Network limits
  - Large node counts require cascaded switches, driving up costs
  - Strong scaling is limited by latencies (small message bandwidth, global sums)
  - These are hard limits – no solution except to wait for better hardware

- Reliability
  - MTBF (mean time between failures) is O(100K) hours
    - For 1000 nodes, O(1) failure per 100 hours
    - Switched networks are failure tolerant, meshes are failure intolerant
  - Soft limit – use job length restrictions to protect results

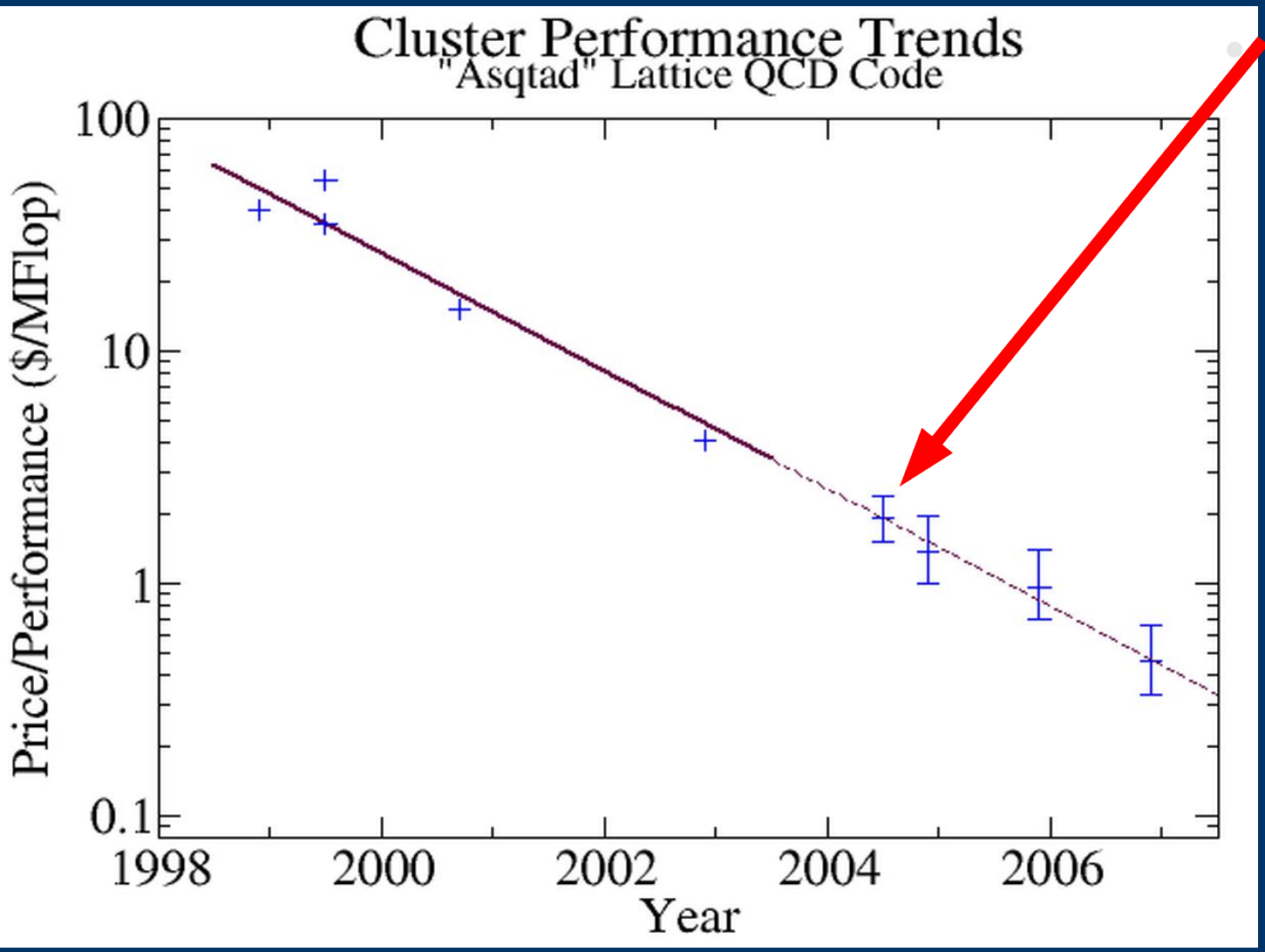
Limits to Cluster Size

- Operating system
  - Mutually asynchronous periodic service interruptions
  - On very large clusters, this will put a lower bound on CG iteration time and hence on performance
  - Soft limit - can be solved with effort
    - This problem is well understood in real time applications, such as triggers
- Power consumption
  - Typical x86 machine consumption is 180 Watts
  - Assuming matching cooling requirement, 1000 nodes require 360 KWatt
  - At $0.045/KWatt-HR, this is ~ $140,000/year (5 to 10% of cluster cost)

Predictions

- Extrapolating from trends, I make some educated guesses about what we will buy in the next few years
- About the predictions:
  - I am only assuming benefits from faster or cheaper hardware
  - I am not assuming benefits from software improvements
    - software used here was MILC “C” code, with site SSE matrix-vector routines (following Lüscher)
    - SciDAC “level-2” and “level-3” routines could give increases of 10-30%
  - I assume hardware improvements slip a year from current vendor roadmaps
    - I don't show these error bars on the time-axis

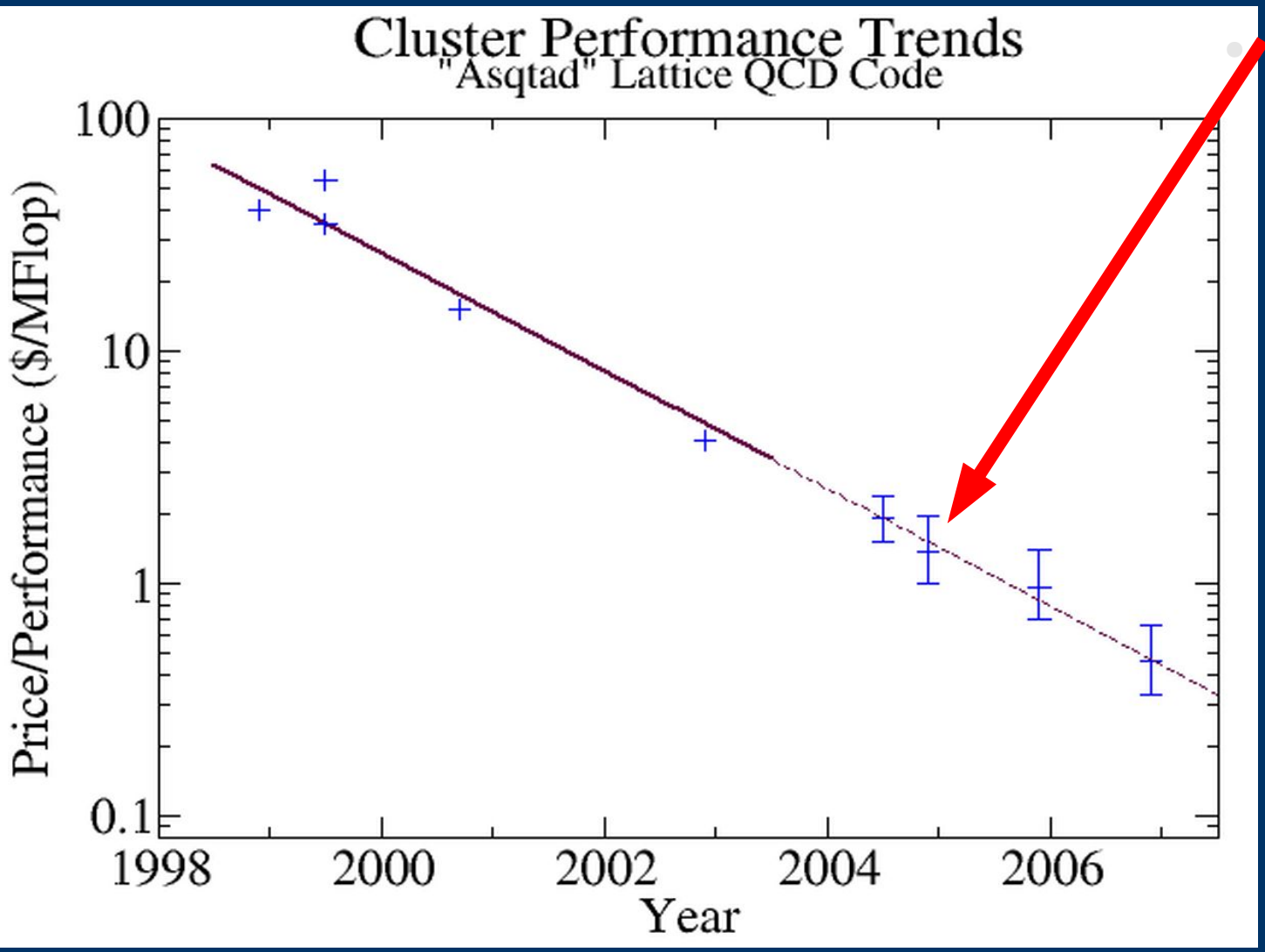
Predictions



Current (June 2004) Fermi purchase:

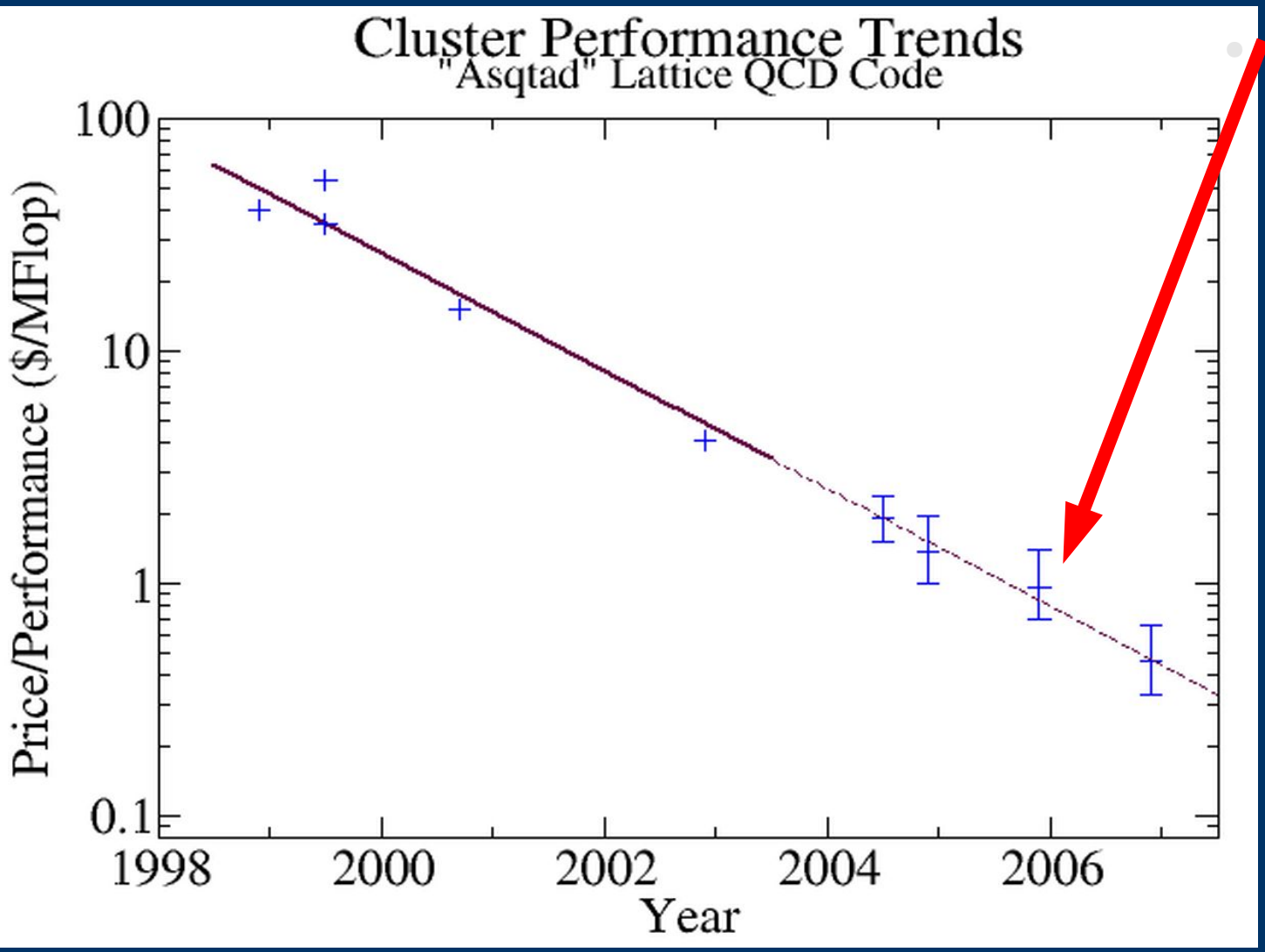
- 2.8 GHz P4E
- PCI-X
- 800 MHz FSB
- Myrinet (reusing existing fabric)
- $900/node
- 1.2 GFlop/node, based on 1.65 GF single node performance (preliminary measurement: 1.10 GFlop/node)

Predictions



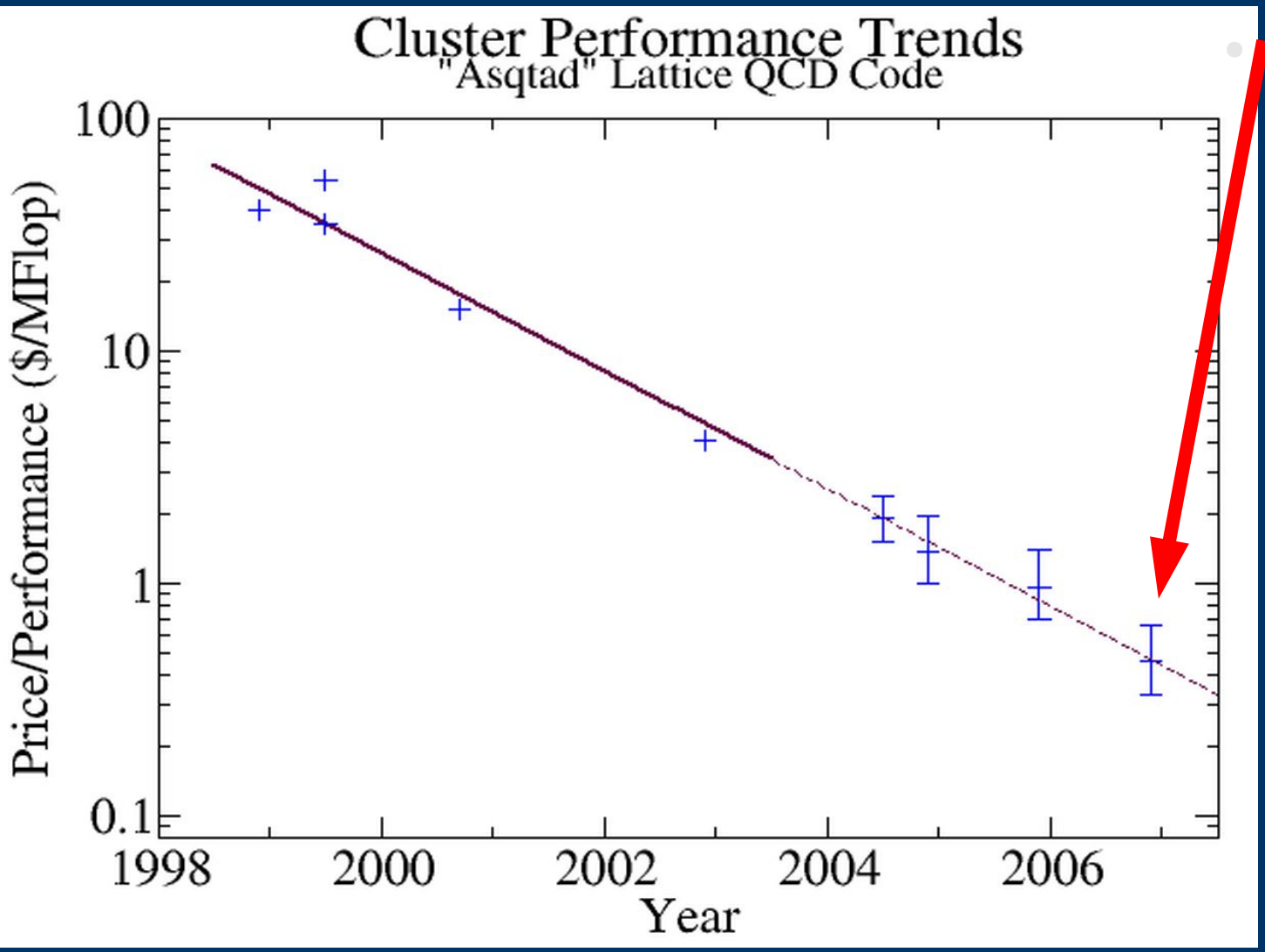
- Late 2004:
  - 3.4 GHz P4E
  - 800 MHz FSB
  - PCI-Express
  - Infiniband
  - $900 + $1000 (system + network per node)
  - 1.4 GFlop/node, based on faster CPU and better network

Predictions



- Late 2005:
  - 4.0 GHz P4E
  - 1066 MHz FSB
  - PCI-Express
  - Infiniband
  - $900 + $900 (system + network per node)
  - 1.9 GFlop/node, based on faster CPU and higher memory bandwidth

Predictions



- Late 2006:
  - 5.0 GHz P4 (or dual core equivalent)
  - >> 1066 MHz FSB ("fully buffered DIMM technology")
  - PCI-Express
  - Infiniband
  - $900 + $500 (system + network per node)
  - 3.0 GFlop/node, based on faster CPU, higher memory bandwidth, cheaper network

Summary

- Since 1999, cluster price/performance has steadily dropped with a halving time of about 1.25 years
- With careful design choices, we can achieve balanced designs:
  - faster CPUs have fortunately also been matched to faster memory buses
  - the transition from PCI-X to PCI-Express should provide I/O headroom for many years
  - multiple, competing network fabric choices are available, with performance increases (for now) pacing processor improvements

Backup Slides

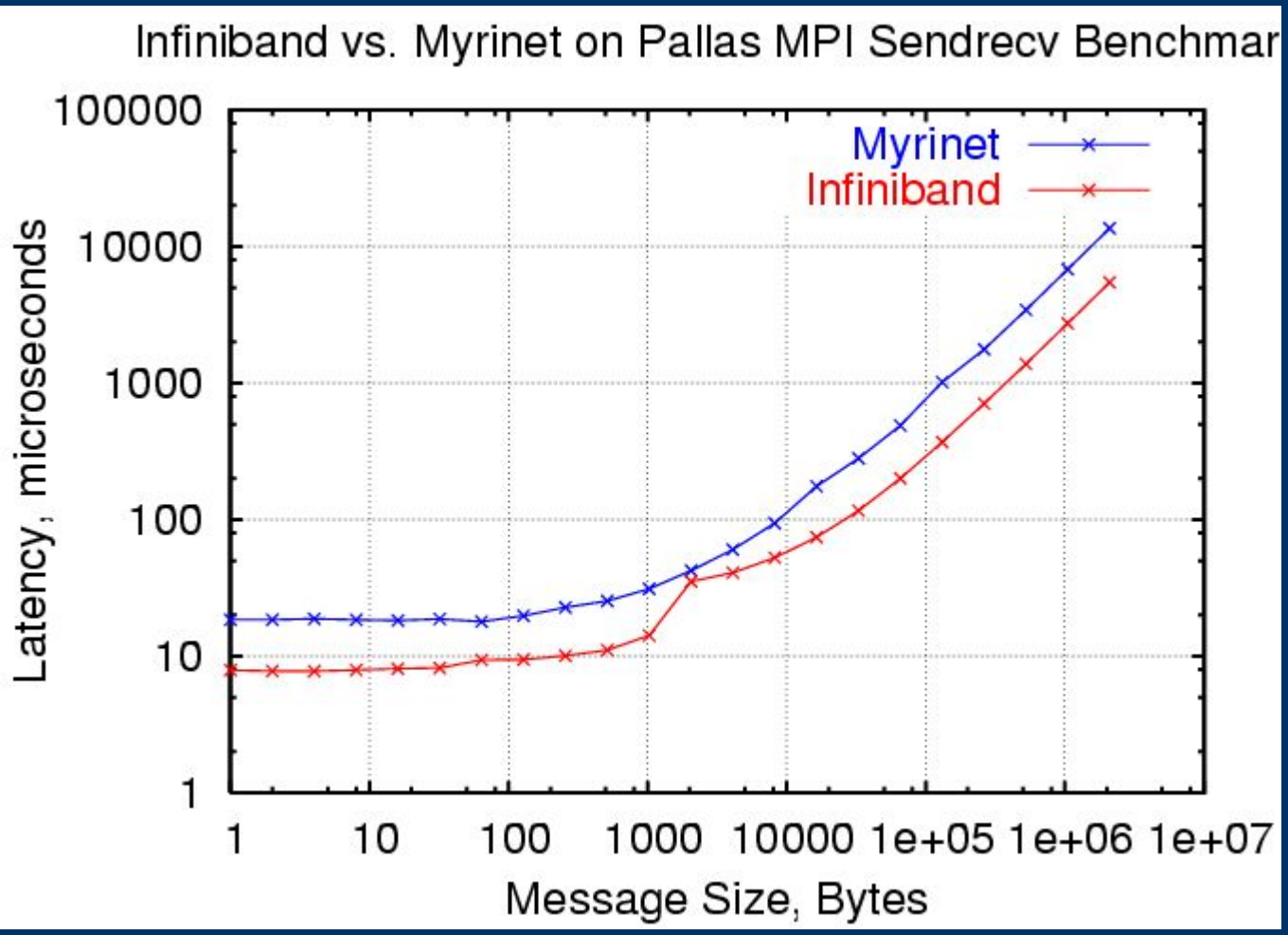
Hardware Roadmap

- Processors
  - Xeon: improve to 800 Mhz FSB now (July)
  - P4E: to 1066 Mhz FSB (September), and to 3.8/4.0 Ghz
  - Intel, IBM have hit the clock speed wall
    - Leakage currents are dominating power consumption
    - Switching to dual core processors
  - Memory bus improvements
    - AMD: integrated memory controllers + hypertransport
    - Intel: “fully buffered DIMMs”

Hardware Roadmap

- Networks
  - Infiniband: dual 4X HCA's now (2 x 1000 MB/sec/dir)
    - 12X HCA's when needed (3000 MB/sec/dir)
    - PCI-E will have to keep pace (8X now, 16X soon)
    - Switches just transitioned from 8-way to 24-way xbar
  - Myricom: faster physical layer eventually
  - Ethernet: 10 gigE emerging

Network Performance
Latency



- Typical performance curves for same networks
- Latency is affected by:
  - network type
  - I/O bus
  - software

Predictions

- 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

Node Costs

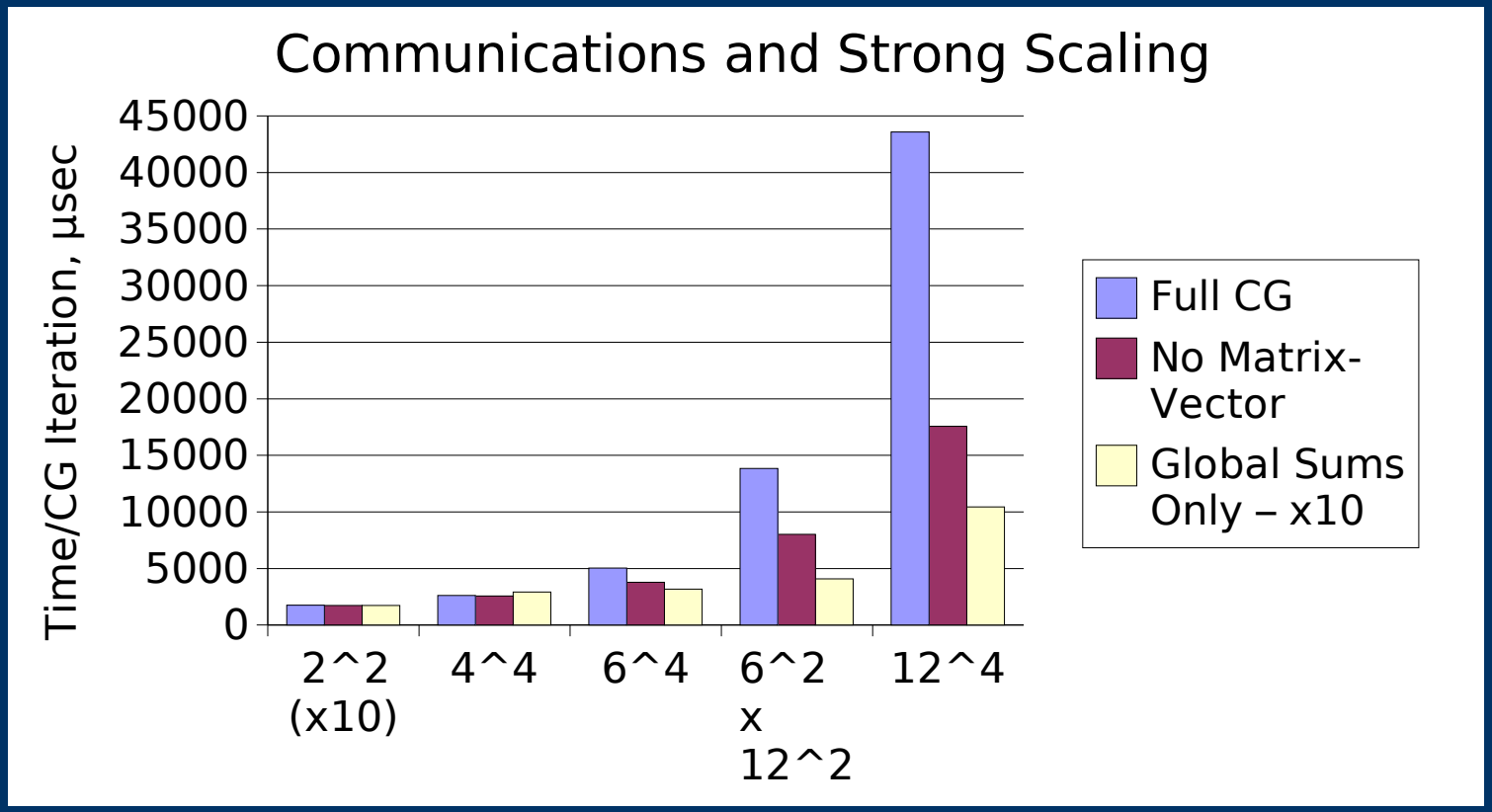
- MP (quad and 8-way)
  - Premium cost – high-end, high-availability servers sold in low volume
  - Usually excellent I/O bus implementation
  - Poorer memory bus than SMP (Intel x86)
    - AMD OpteronMP fixes this
  - Now: $4000+

Balanced Design Requirements
CG Inverson of Dslash

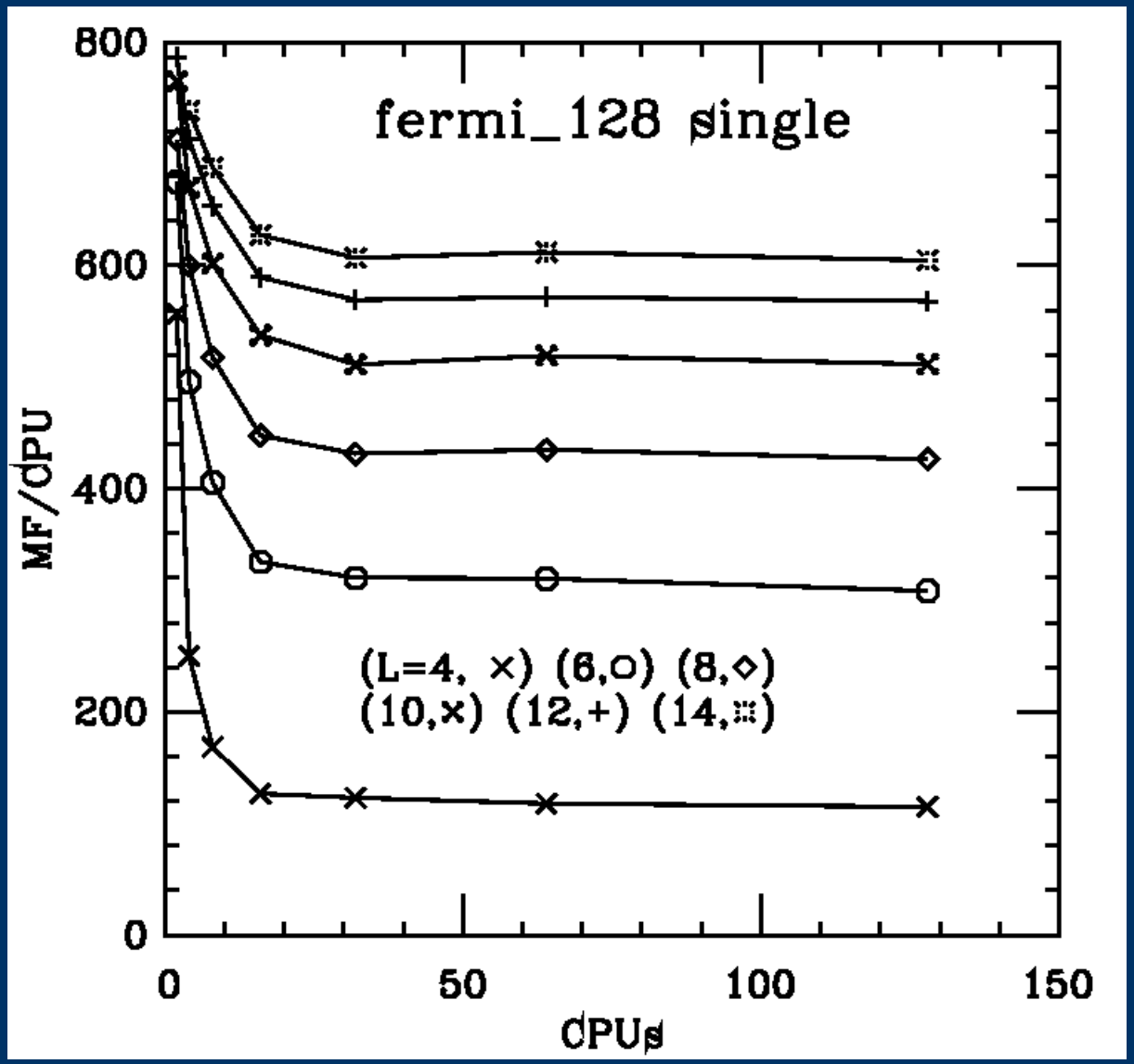
- Conjugate gradient algorithm:
  - Evaluate *Dslash* for even, odd sites
  - Refine estimates using inner products accumulated from all nodes via global sums
- Global sums expose network latency
  - Execution time goes as *log N*, where *N* = node count
  - Fermilab Myrinet cluster:
    - 64-node global sum of a double takes 155 µsec
  - Limits strong scaling (relative time to solve a problem of constant size as node count increases):
    - The global sum communications cannot overlap with computations
    - As node count increases, the time to perform the global sums will approach and pass the time to do the computations

Balance Design Requirements - CG Inversion of Dslash

- Communications set the strong scaling limit
- Example:
  - Asqtad on 64-node, 2.4 GHz Xeon cluster with Myrinet
  - Modified MILC code times the full CG, CG without matrix-vector operations, CG with only global sums
  - For small local lattices, performance is bound by communications

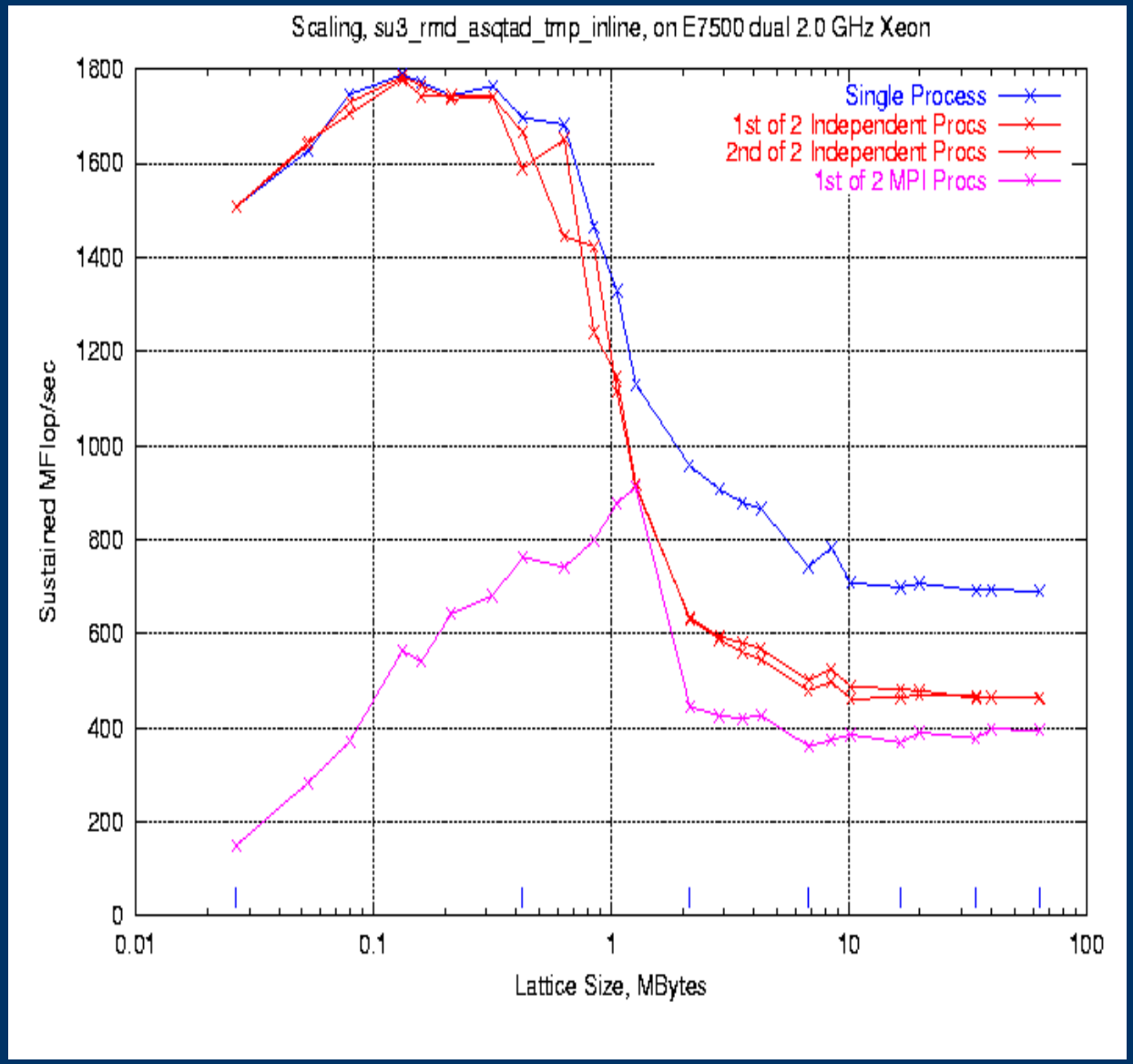


Weak Scaling Behavior



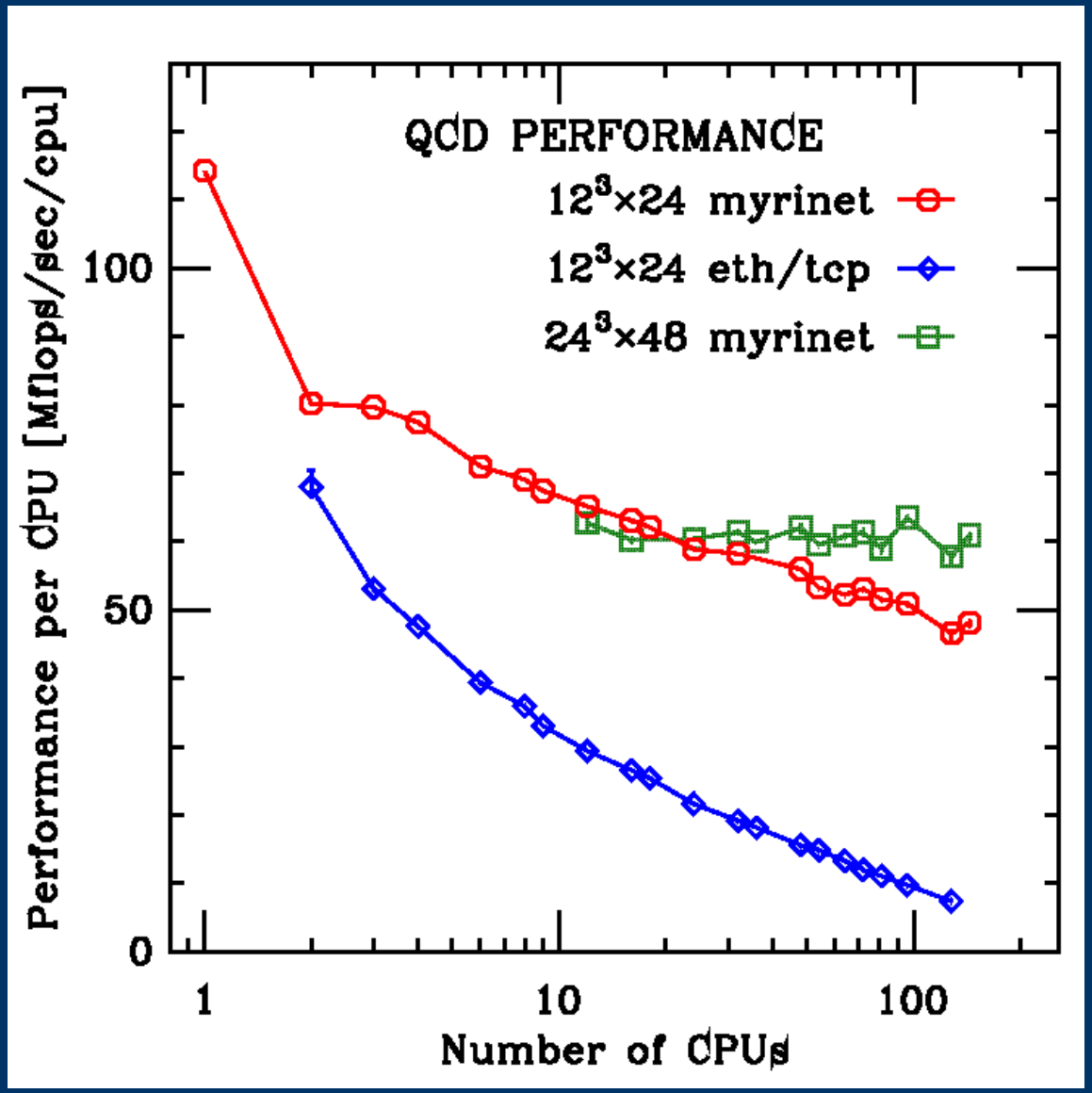
- “Weak Scaling” - relative performance as node count is increased, where local lattice volume on each node is kept constant

SMP Scaling - Xeon



- Shared memory bus limits aggregate performance of dual Xeon processors

Strong Scaling Behavior



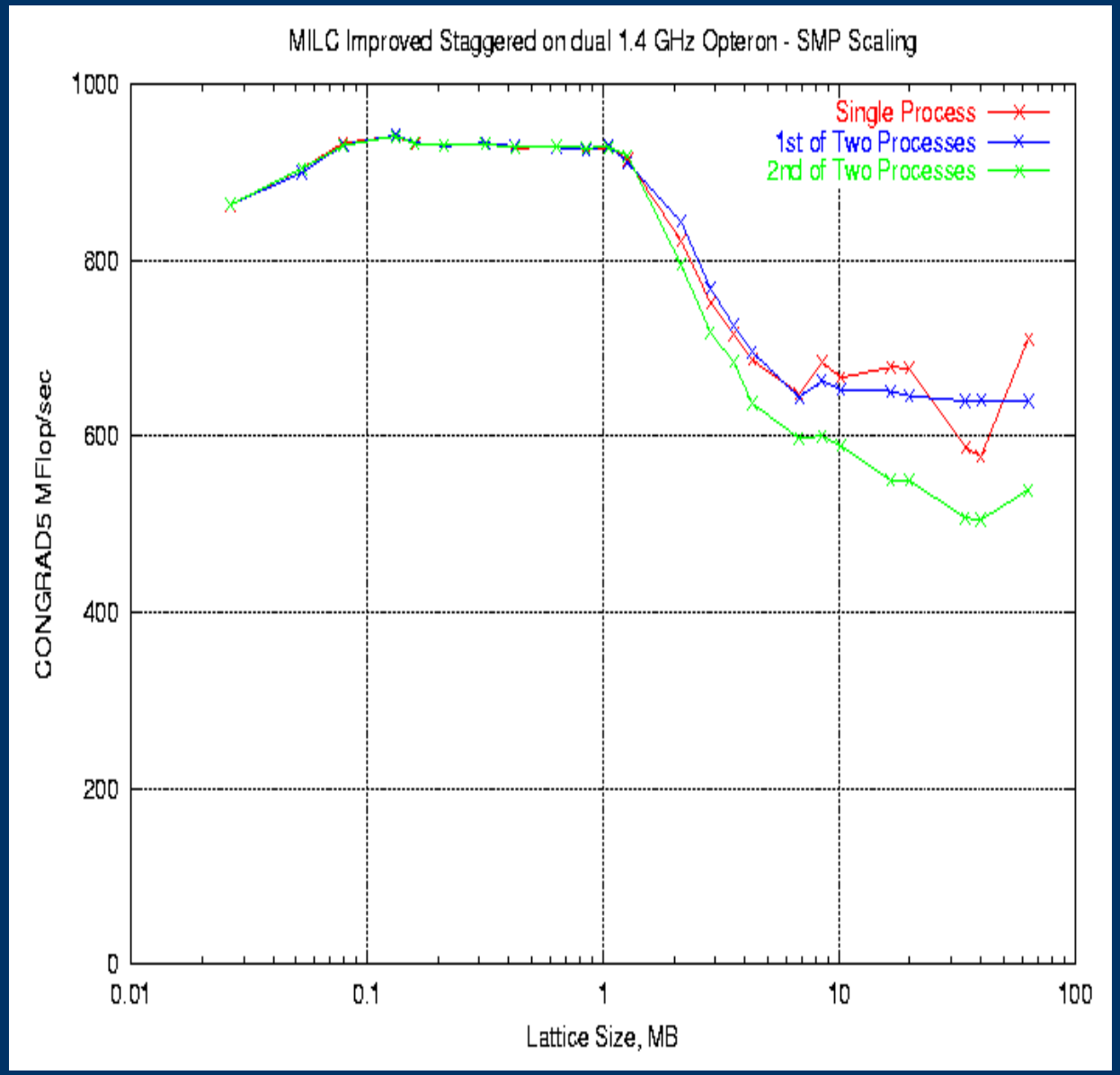
- "Strong Scaling" - relative performance as node count is increased and the lattice size is kept constant
  - smaller sublattices (local lattice on each node) as count increases

Price/Performance

Clusters included:

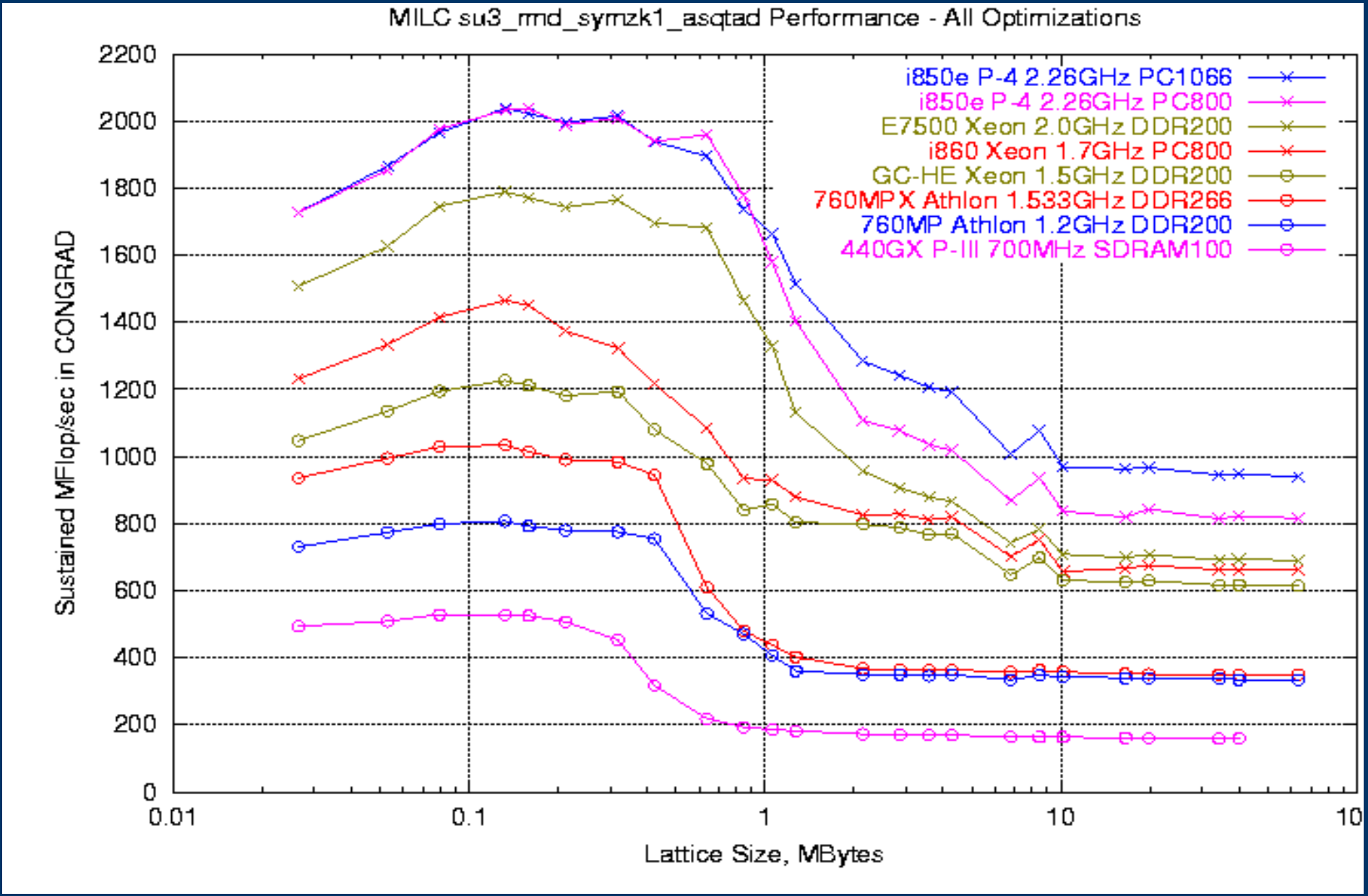
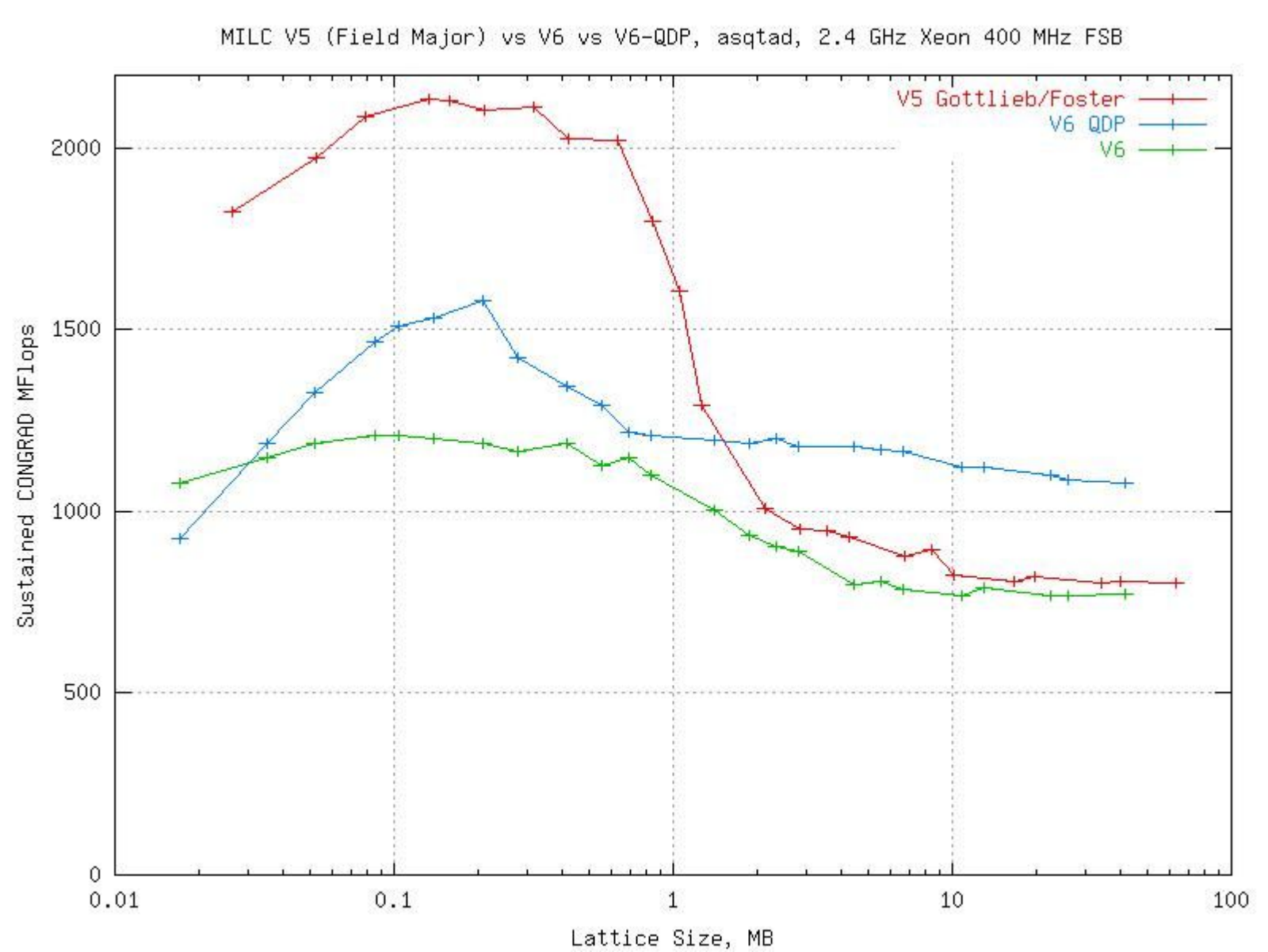
- Pentium II fast ethernet and Myrinet (~ 1999)
- Pentium III Myrinet (2000)
- 2.4 GHz Xeon, Myrinet (2003)
- 2.8 GHz P4E, Myrinet

SMP Scaling - Opteron



- Opterons have embedded memory controllers, resulting in scalable SMP systems

Performance vs Clock Speed



Performance vs Optimization