Lattice QCD Clusters at Fermilab

Don Holmgren, Paul Mackenzie, Amitoj Singh, Jim Simone

Fermilab

CHEP'04
September 27, 2004

http://lqcd.fnal.gov/chep04.pdf

Outline

- Current Production Clusters
- Aspects of Performance
  - floating point
  - memory bandwidth
  - communications – networks and buses
    - Infiniband details
  - processors
- Performance Trends
  - single node price/performance
  - cluster price/performance
  - extrapolations and plans

Current Production Clusters

- Cluster architecture:
  - head node:
    - on public internet, also on private ethernet and high performance fabrics (Myrinet)
    - serves /home, /usr/local via NFS
  - worker nodes:
    - on private networks, only accessible via head node
    - data files are staged by user job scripts to local disk via TCP/IP over high performance network
  - PBS batch system with Maui scheduler
  - large data areas to be served to workers by dcache

Current Production Clusters

- December 2004, 2.4 GHz SciDAC Dual Xeon cluster:
  - Sustains ~ 815 MFlop/node, or $3.9/MFlop
  - 128 nodes based on SuperMicro P4DPE motherboards
  - E7500 chipset, 400 MHz front side bus
  - Myrinet 2000 fabric
    - M3F-PCI64B interfaces (LANai 9 chips)
  - 1 GB memory
  - 20 GB local disk
  - Price:
    Computers: $1750/node
    Myrinet: $1400/node

Current Production Clusters

- July 2004, 2.8 GHz SciDAC Pentium 4E cluster:
  - Sustains ~ 1 GFlop/node
    - $0.90/MFlop incremental cost
  - 128 nodes based on Intel SE7210TP1-E motherboard
  - E7210 chipset: 800 MHz FSB, 64/66 PCI-X
  - Reused fabric from old dual Pentium III cluster (Myrinet 2000, LANai 9)
  - 1 GB memory
  - 40 GB local disk
  - Price:
    - Computers: $900/node
    - Network (estimated): $900/node

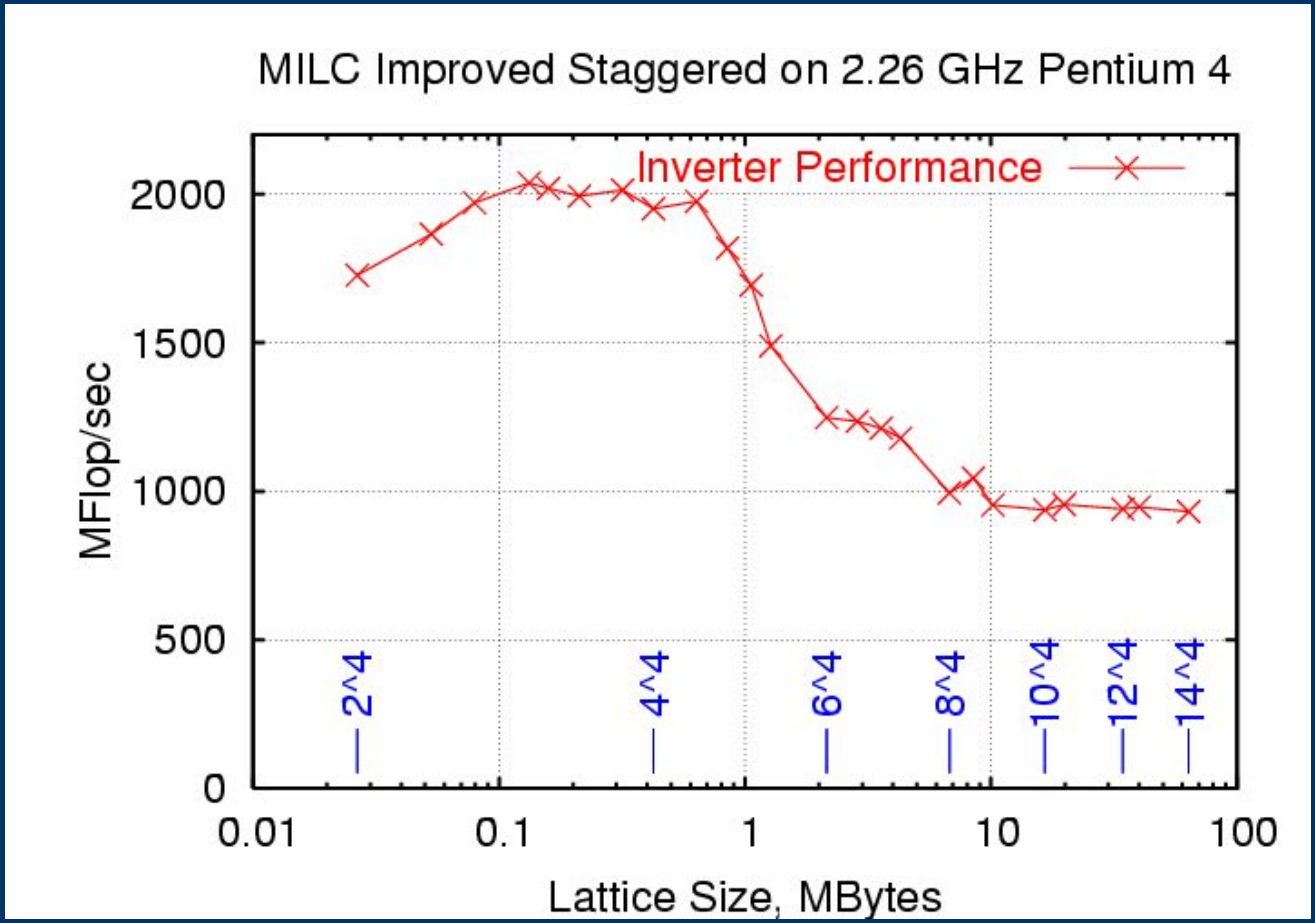
Protoype Infiniband Cluster

- Two TopSpin 120 24-port (X4 ports) Infiniband switches
- 32 nodes, 2.0 GHz dual Xeon, SuperMicro P4DPE motherboard, TopSpin PCI-X Infiniband host channel adapters each with two X4 Infiniband ports
- 2 nodes, 3.2 GHz single Pentium 4E, Abit AA8 motherboard, Mellanox PCI-E Infiniband HCA
- Costs:
  - TopSpin 120 switches: $4000 ($167/port)
  - PCI-X HCA's: $490
  - PCI-E HCA's: $735
  - Dual Xeon systems (July 2002): $1700
  - P4E systems (July 2004): $960

Aspects of Performance

- Lattice QCD codes require:
  - excellent single and double precision floating point performance
    - majority of Flops are consumed by small complex matrix-vector multiplies (SU3 algebra)
  - high memory bandwidth (principal bottleneck)
  - low latency, high bandwidth communications
    - typically implemented with MPI or similar message passing APIs

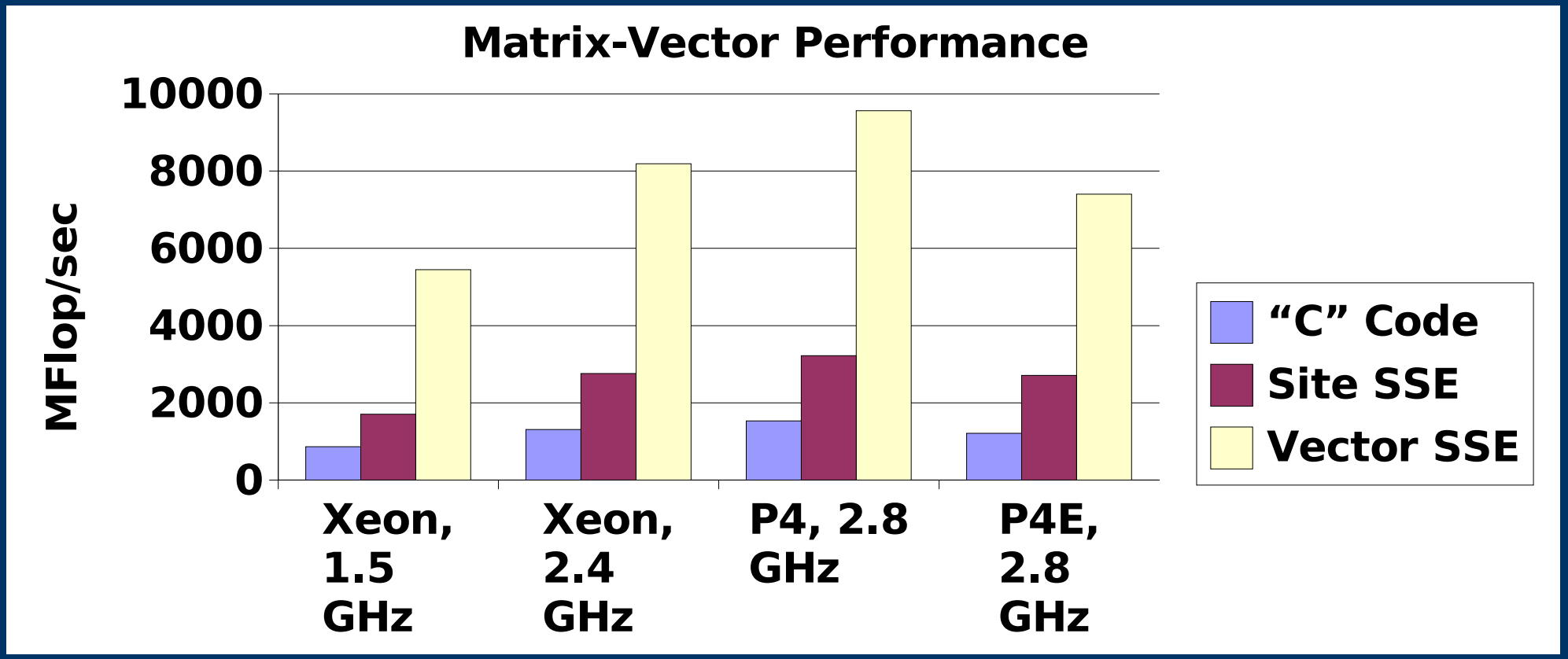
Generic Single Node Performance



- MILC is a standard MPI-based lattice QCD code
- "Improved Staggered" is a popular "action" (discretization of the Dirac operator)
- 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)

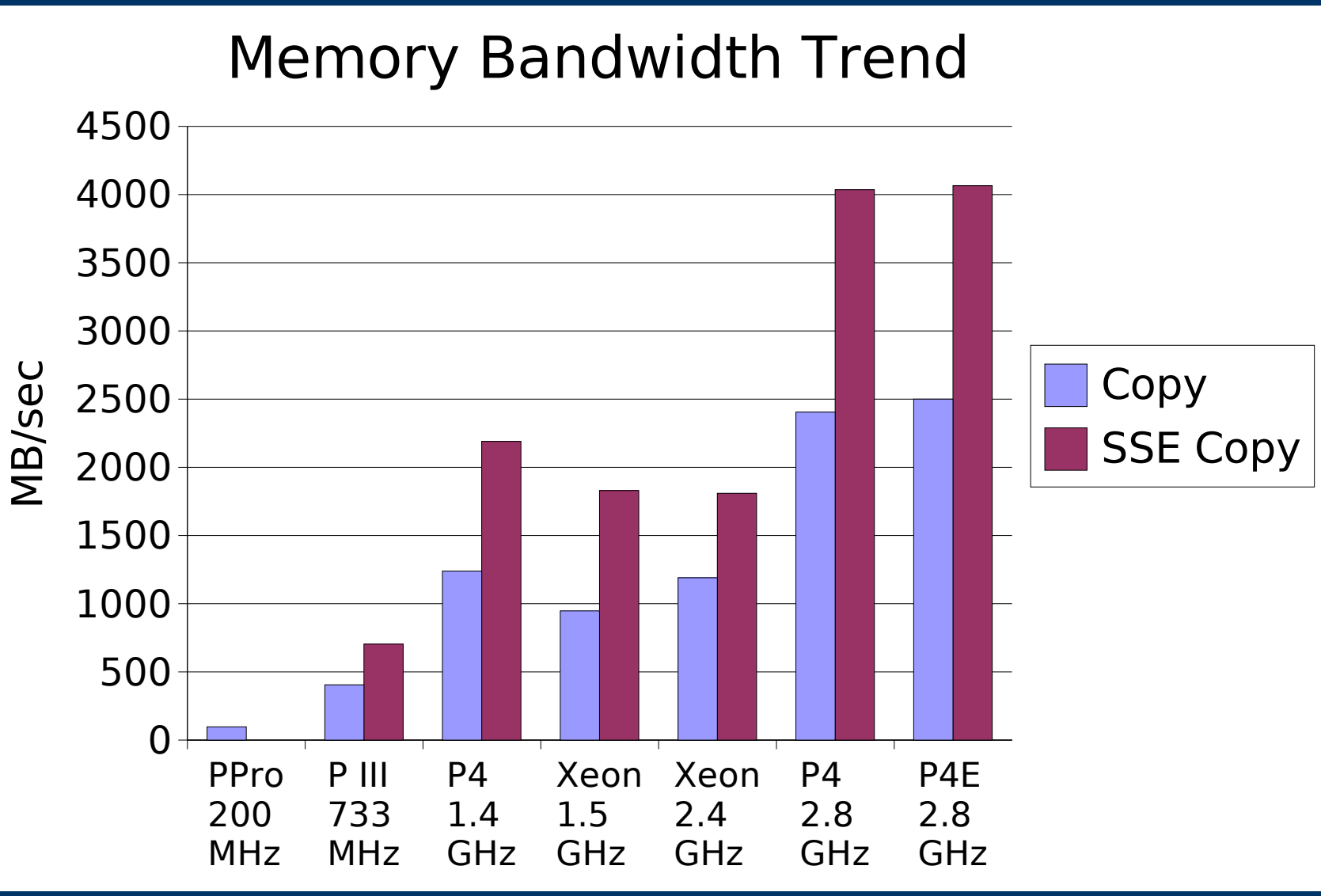
- Most flops are SU3 matrix times vector (complex)
  - SSE/SSE2/SSE3 can give a significant boost
    - Site-wise (M. Lüscher)
    - Fully vectorized (A. Pochinsky)
      - requires a memory layout with 4 consecutive reals, then 4 consecutive imaginaries



Memory Performance

- Memory bandwidth limits – depends on:
  - Width of data bus – always 64 bits for x86 processors
  - (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

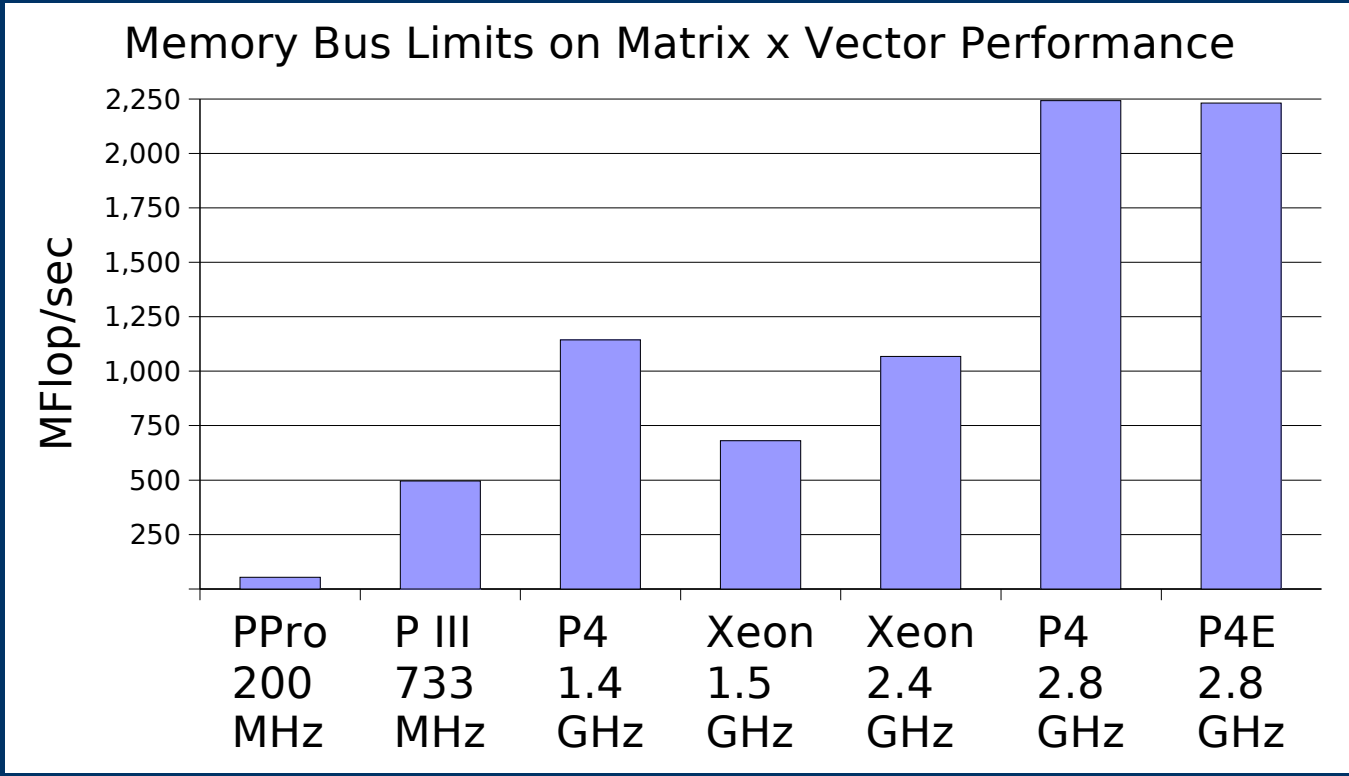

Memory Bandwidth Trend
MB/sec
4500
4000
3500
3000
2500
2000
1500
1000
500
0
PPro 200 MHz
P III 733 MHz
P4 1.4 GHz
Xeon 1.5 GHz
Xeon 2.4 GHz
P4 2.8 GHz
P4E 2.8 GHz
Copy
SSE Copy

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



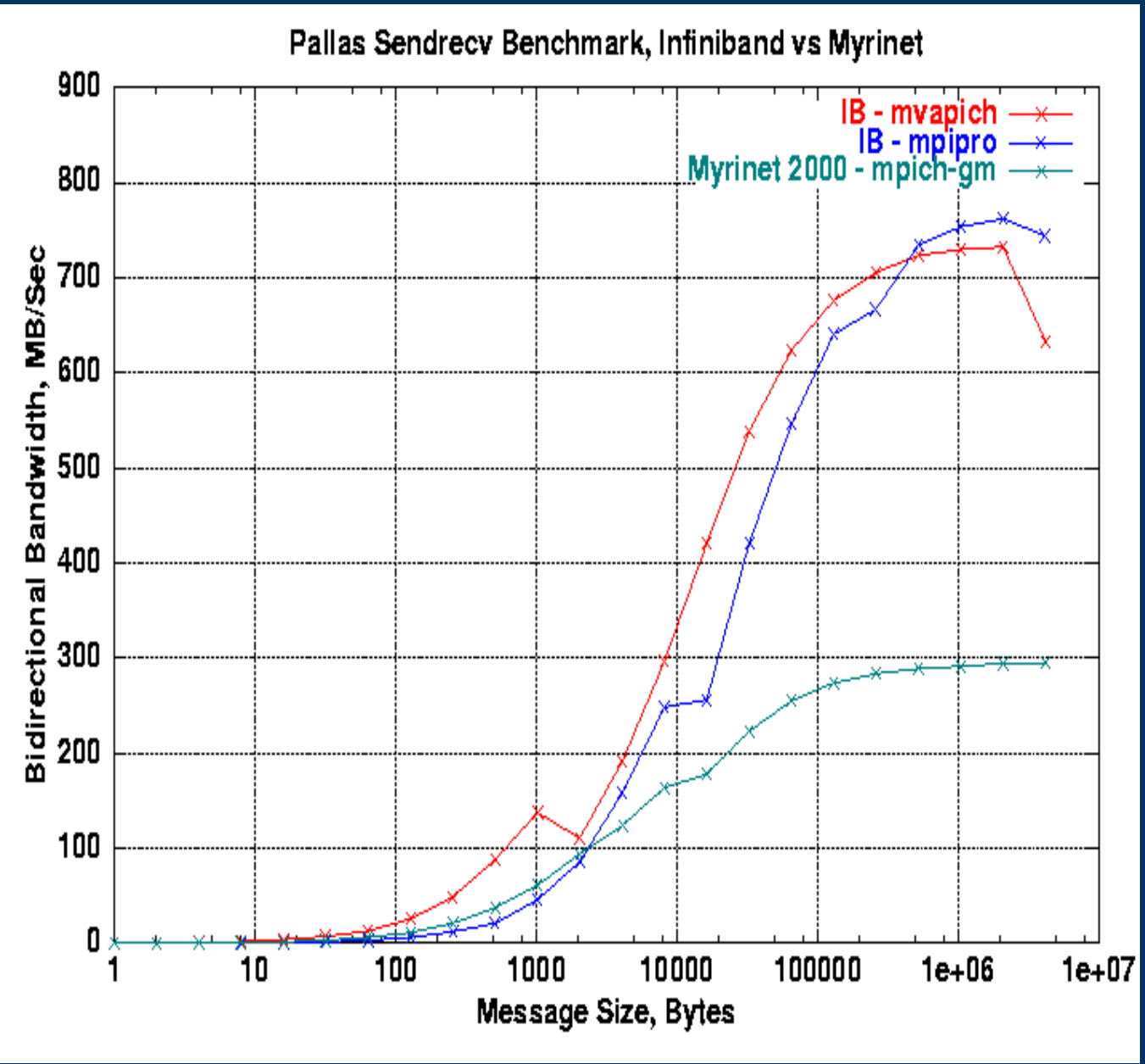
Communications – I/O Buses

- Low latency and high bandwidths are required
- Performance depends on I/O bus:
  - at least 64-bit, 66 MHz PCI-X is required for LQCD
  - PCI Express (PCI-E) is now available
    - not a bus, rather, one or more bidirectional 2 Gbit/sec/direction (data rate) serial pairs
    - for driver writers, PCI-E looks like PCI
    - server boards now offer X8 (16 Gbps/direction) slots
    - we've used desktop boards with X16 slots intended for graphics – but, Infiniband HCAs work fine in these slots
    - latency is also better than PCI-X
    - strong industry push this year, particularly for graphics (thanks, DOOM 3!!)
    - should be cheaper, easier to manufacture than PCI-X

Communications - Fabrics

- Existing Lattice QCD clusters use either:
  - Myrinet
  - Gigabit ethernet (switched or multi-D toroidal mesh)
  - Quadrics also a possibility, but historically more expensive
  - SCI works as well, but has not been adopted
- Emerging (finally) is Infiniband
  - like PCI-E, multiple bidirectional serial pairs
  - all host channel adapters offer two independent X4 ports
  - rich protocol stacks, now available in open source
  - target HCA price of $100 in 2005, less on motherboard
- Performance (measured at Fermilab with Pallas MPI suite):
  - 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

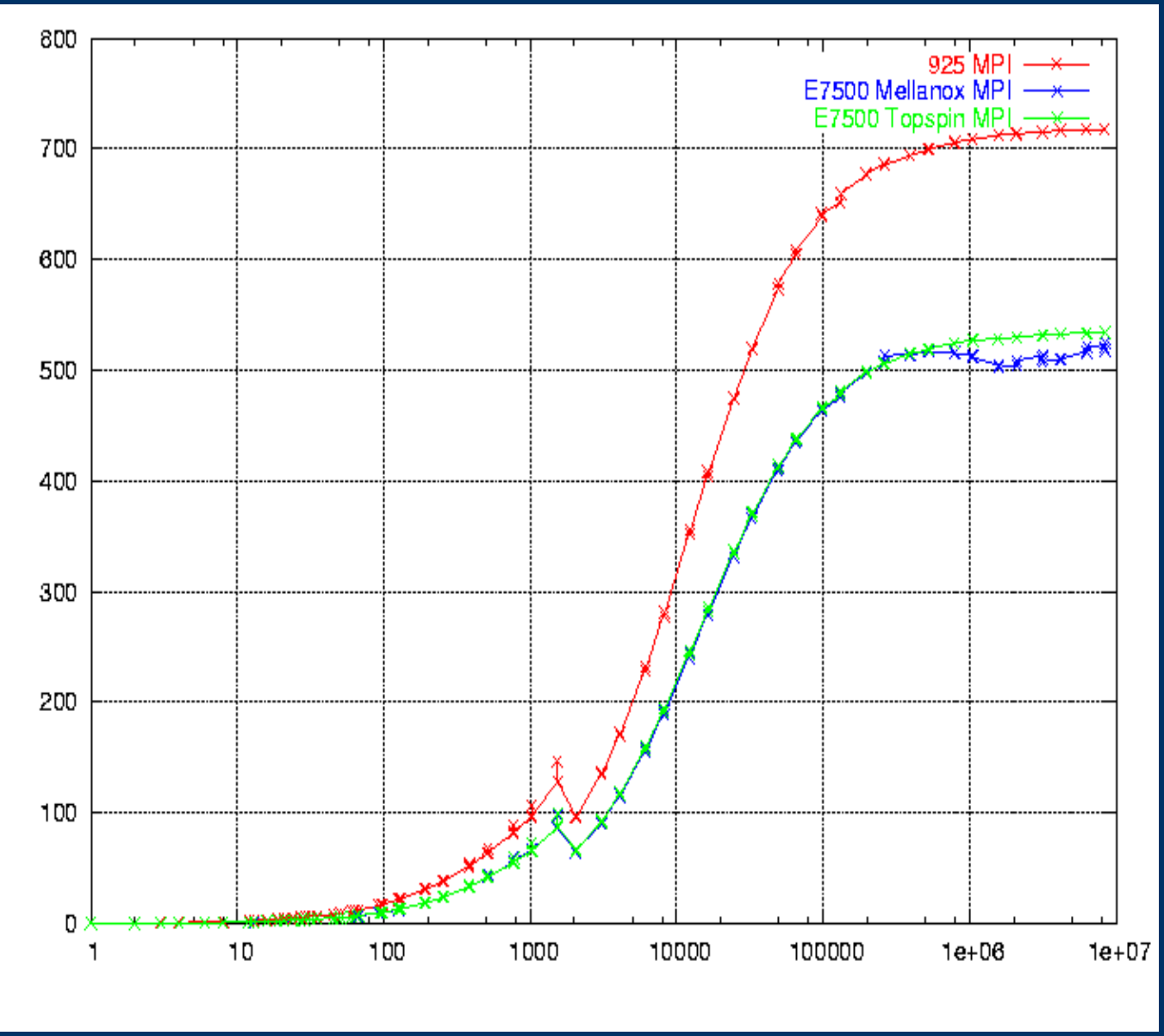
Networks – Myrinet vs Infiniband



Network performance:

- Myrinet 2000 on E7500 motherboards
  - Note: much improved bandwidth, latency on latest Myricom hardware
- Infiniband PCI-X on E7501 motherboards
- Important message size region for lattice QCD is O(1K) to O(10K)

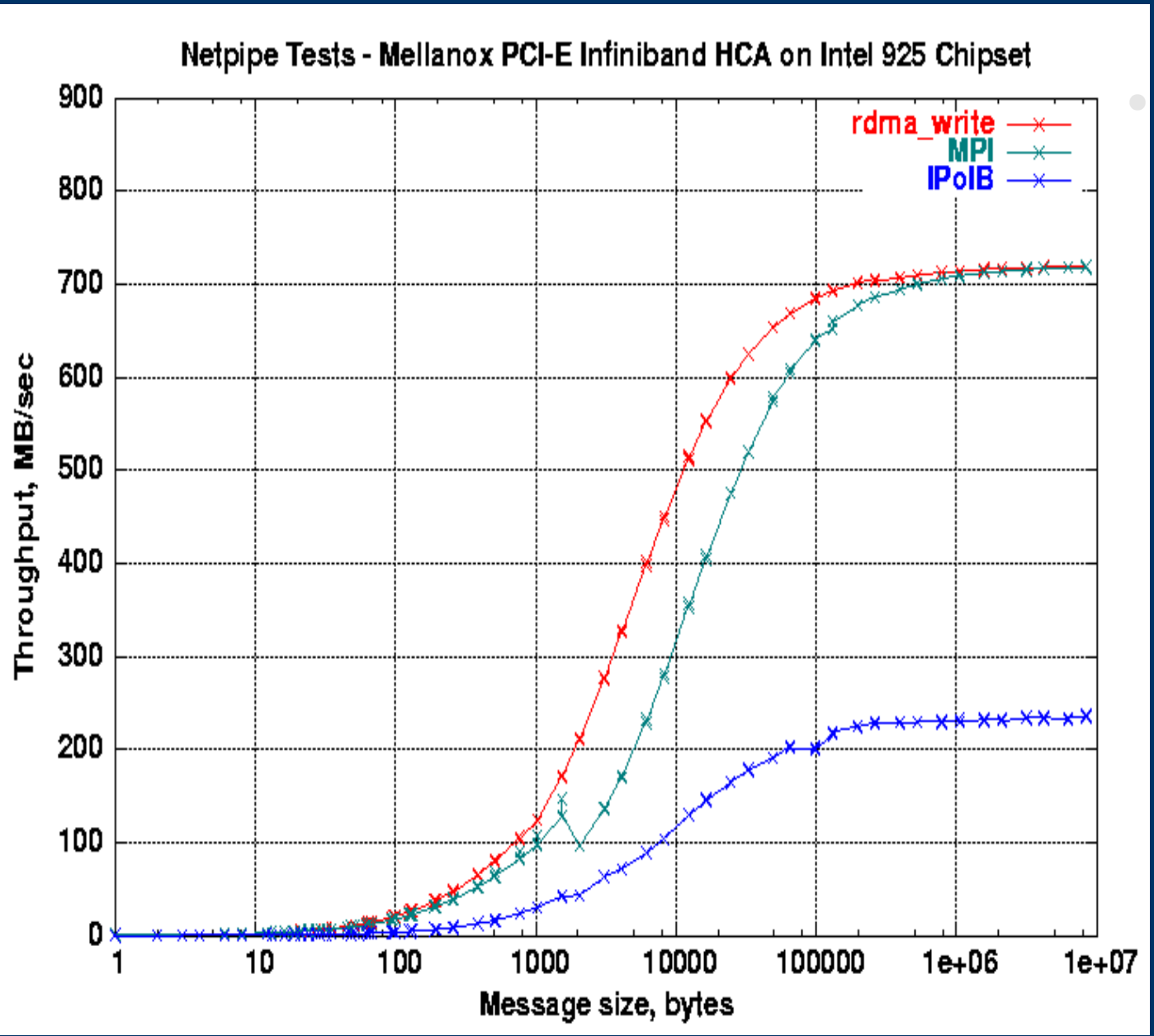
Infiniband on PCI-X and PCI-E



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

- PCI-X on E7500
  - "TopSpin MPI" from OSU
  - "Mellanox MPI" from NCSA
- PCI-E on 925X
  - NCSA MPI
  - 8X HCA used in 16X "graphics" PCI-E slot

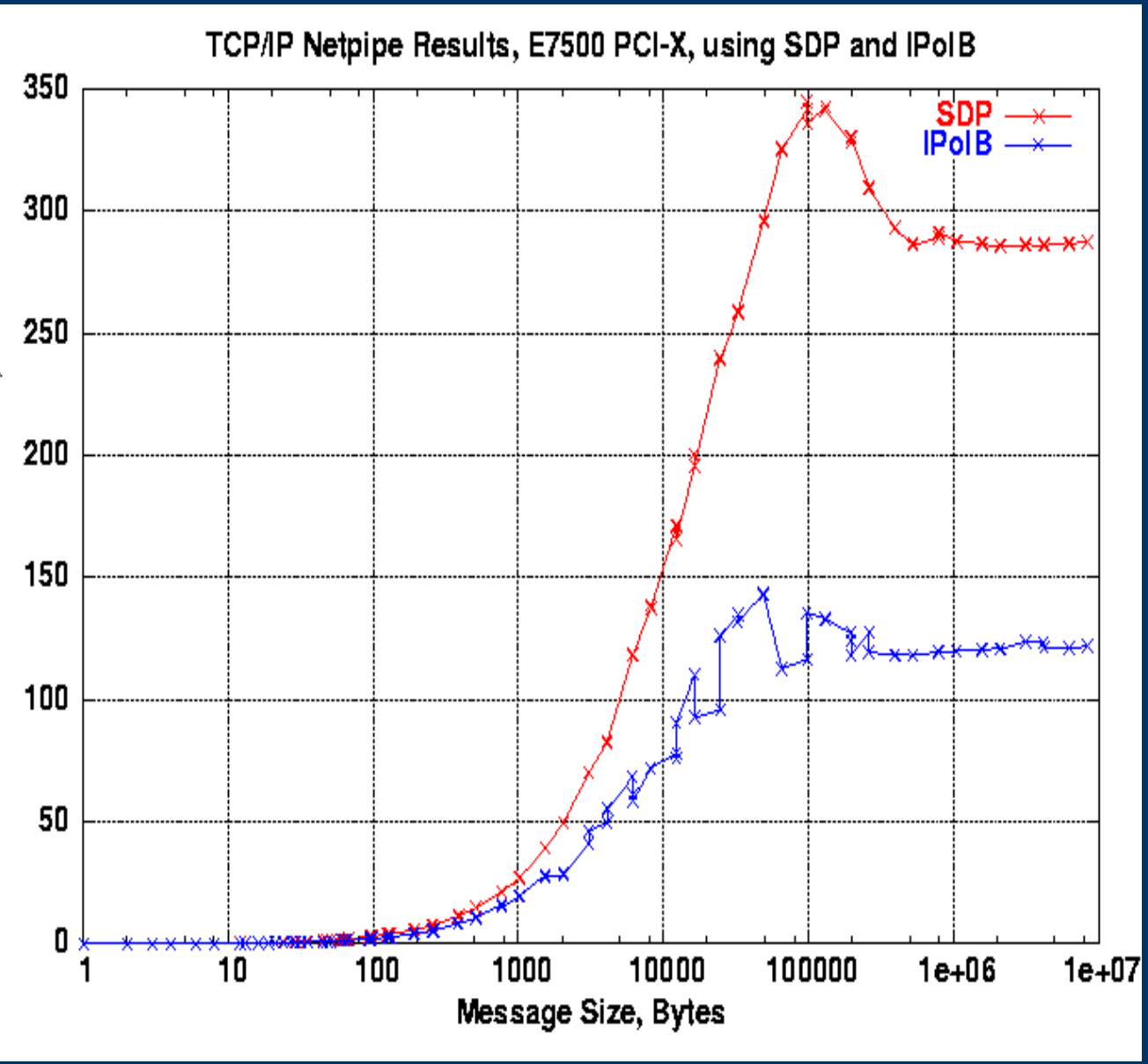
Infiniband Protocols



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

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

TCP/IP over Infiniband



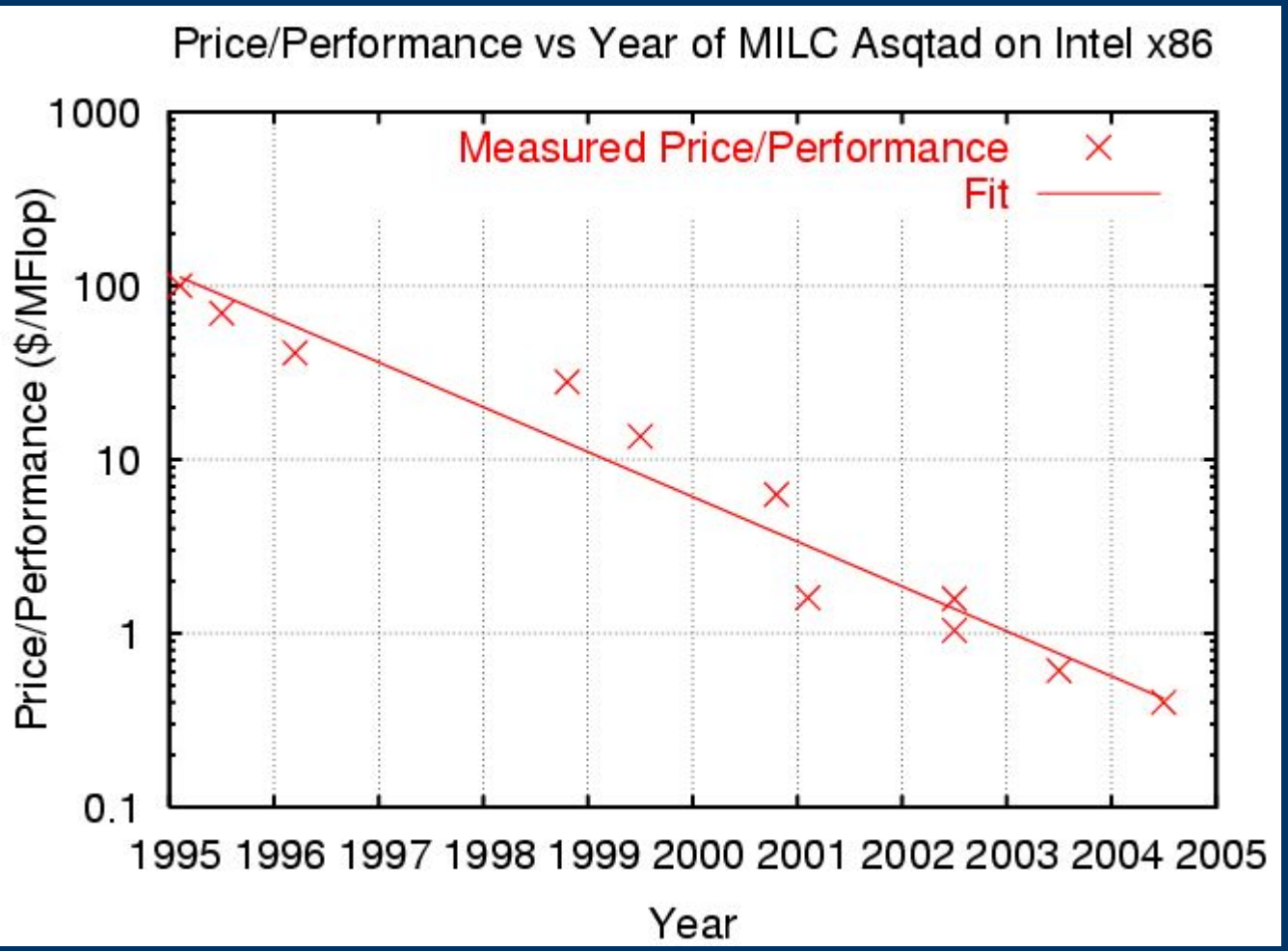
Options for codes which stream data over sockets:

- "IPoIB" - full TCP/IP stack in Linux kernel
- "SDP" - new protocol, AF_SDP, instead of AF_INET
  - Bypasses kernel TCP/IP stack
  - Socket-based code has no other changes
  - Data here were taken with the same binary, using LD_PRELOAD for libsdp

Processor Observations

- Using MILC “Improved Staggered” code, we found:
  - the new 90nm Intel chips (Pentium 4E, Xeon “Nacona”) have lower floating point performance at the same clock speeds because of longer instruction latencies
    - but – better performance in main memory (better hardware prefetching?)
  - dual Opterons scale at nearly 100%, unlike Xeons
    - but – must use NUMA kernels + *libnuma*, and alter code to lock processes to processors and allocate only local memory
    - 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
    - power consumption very high for the 2003 CPUs (dual G5 system drew 270 Watts, vs 190 Watts for dual Xeon
    - we hear that power consumption is better on 90nm chips

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

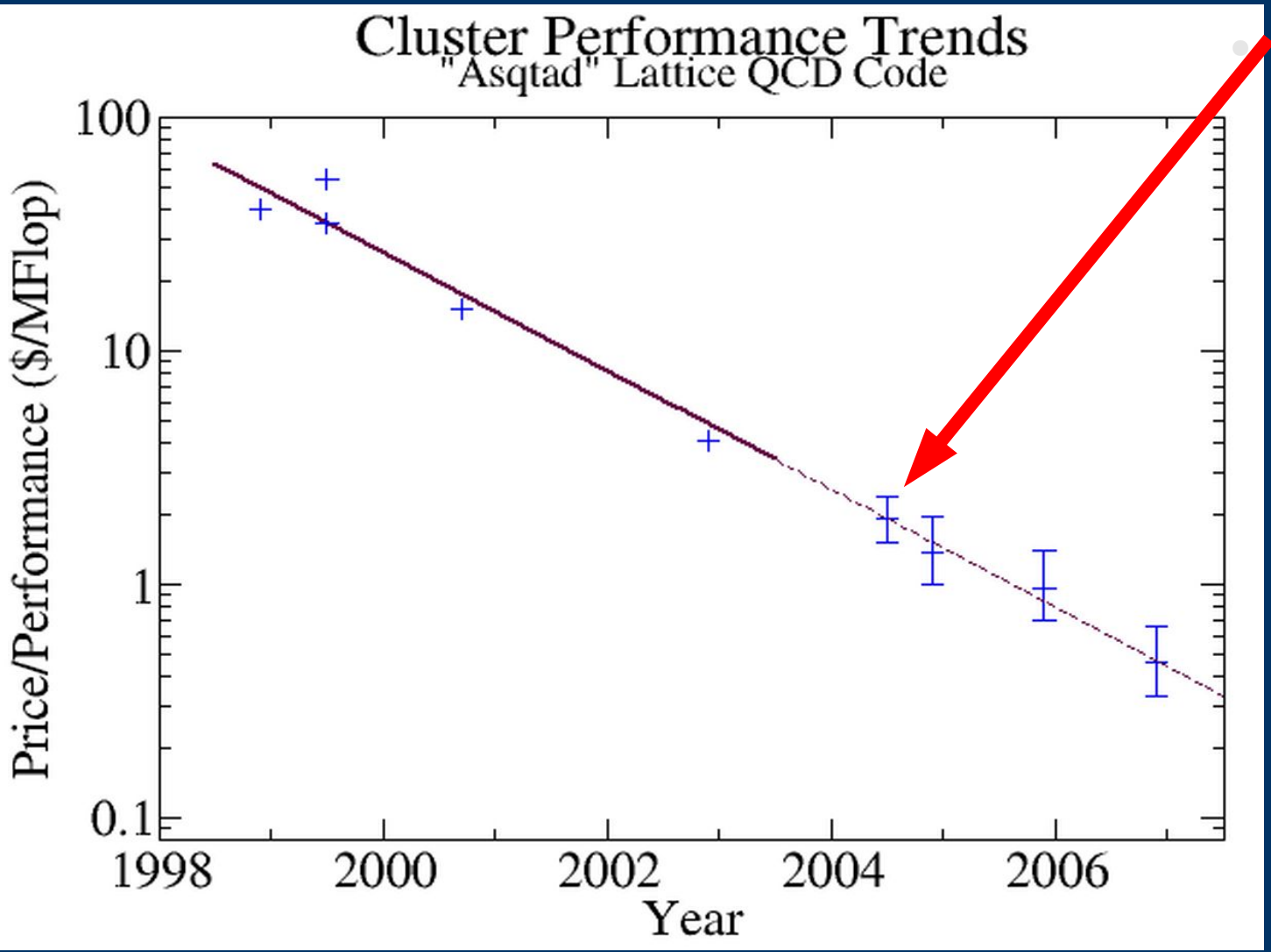
Performance range:

- 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

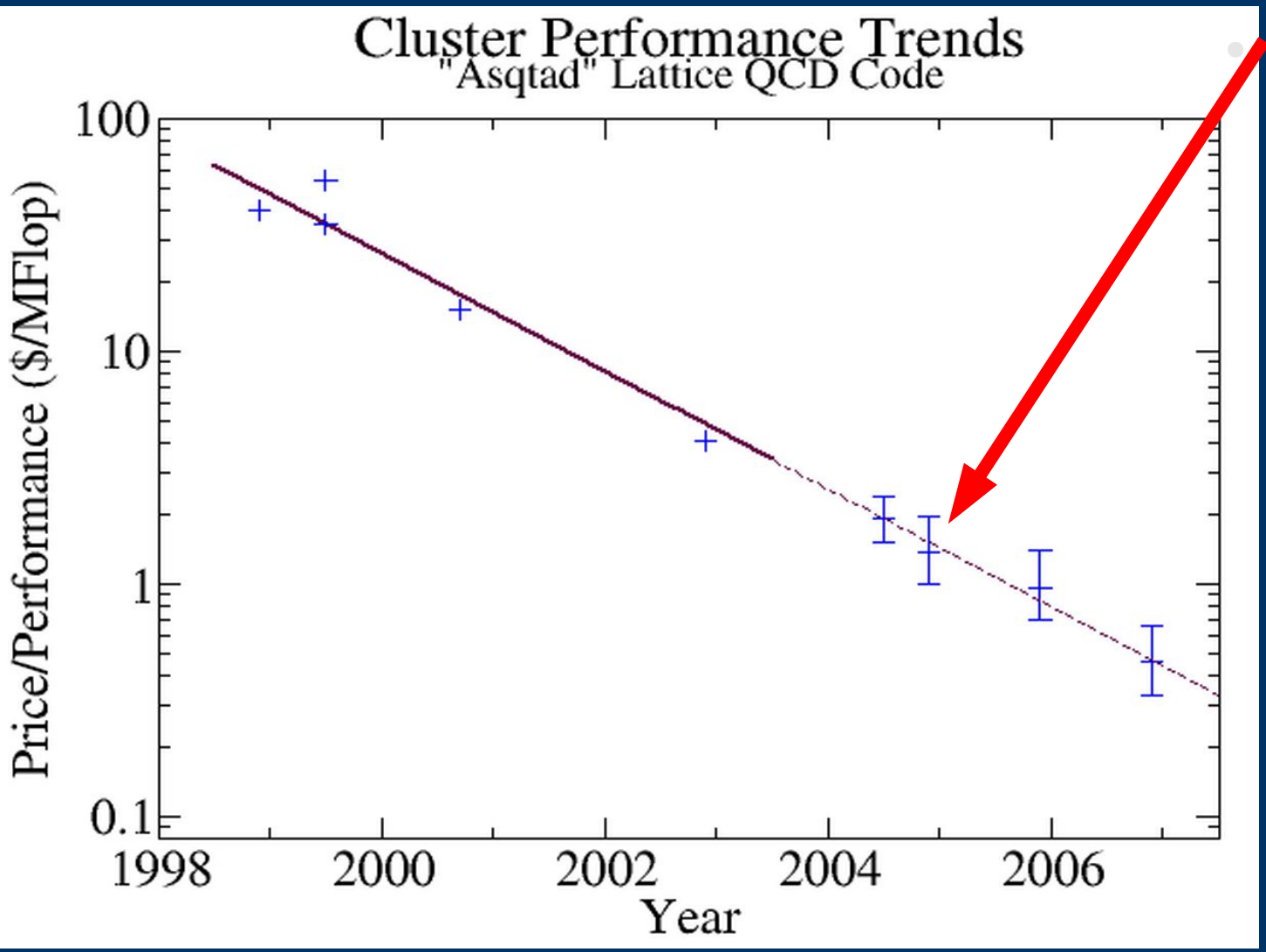
Predictions



Latest (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 (measured: 1.0 – 1.1 GFlop/node, depending on 2-D or 3-D communications)

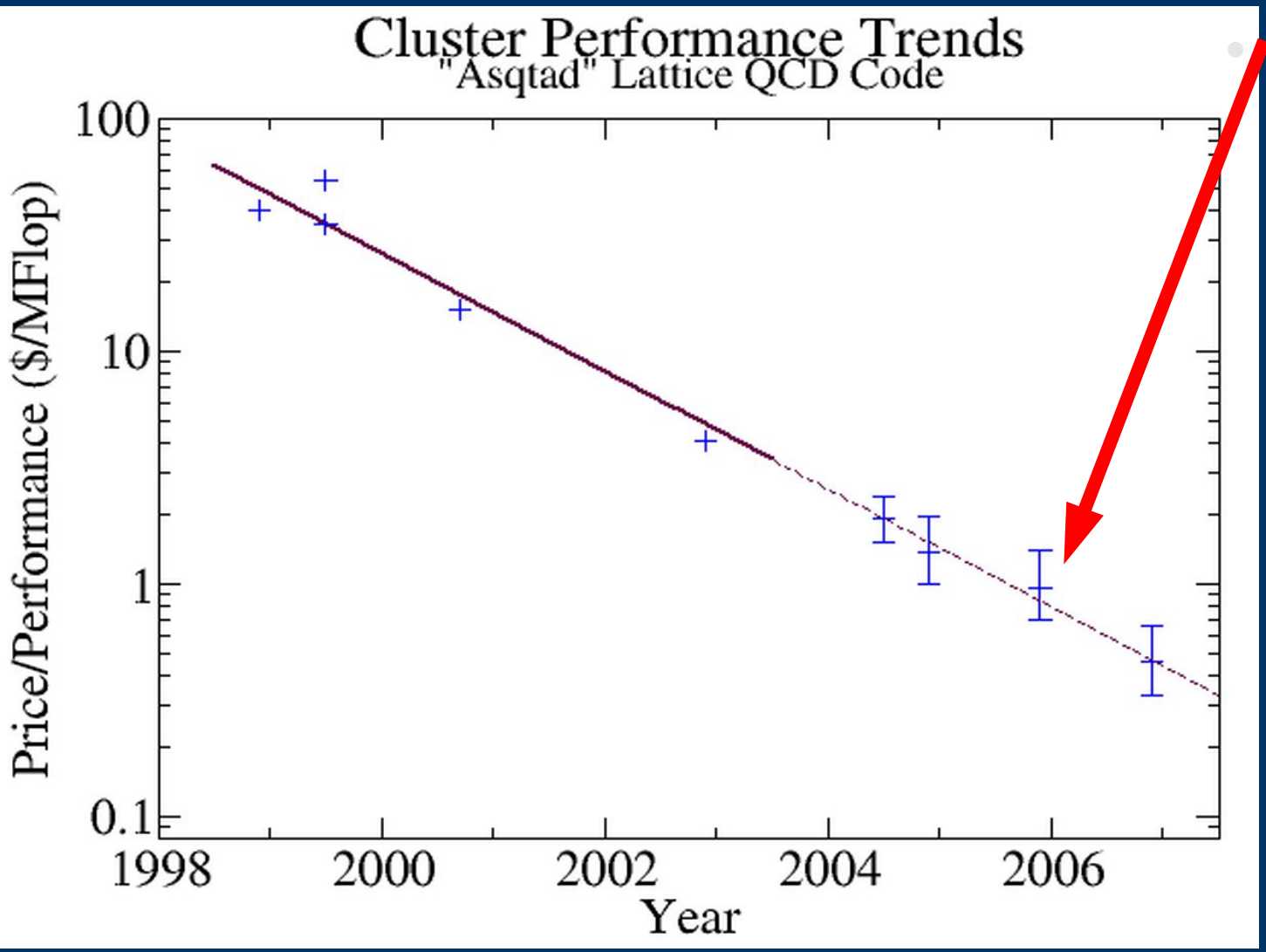
Predictions



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

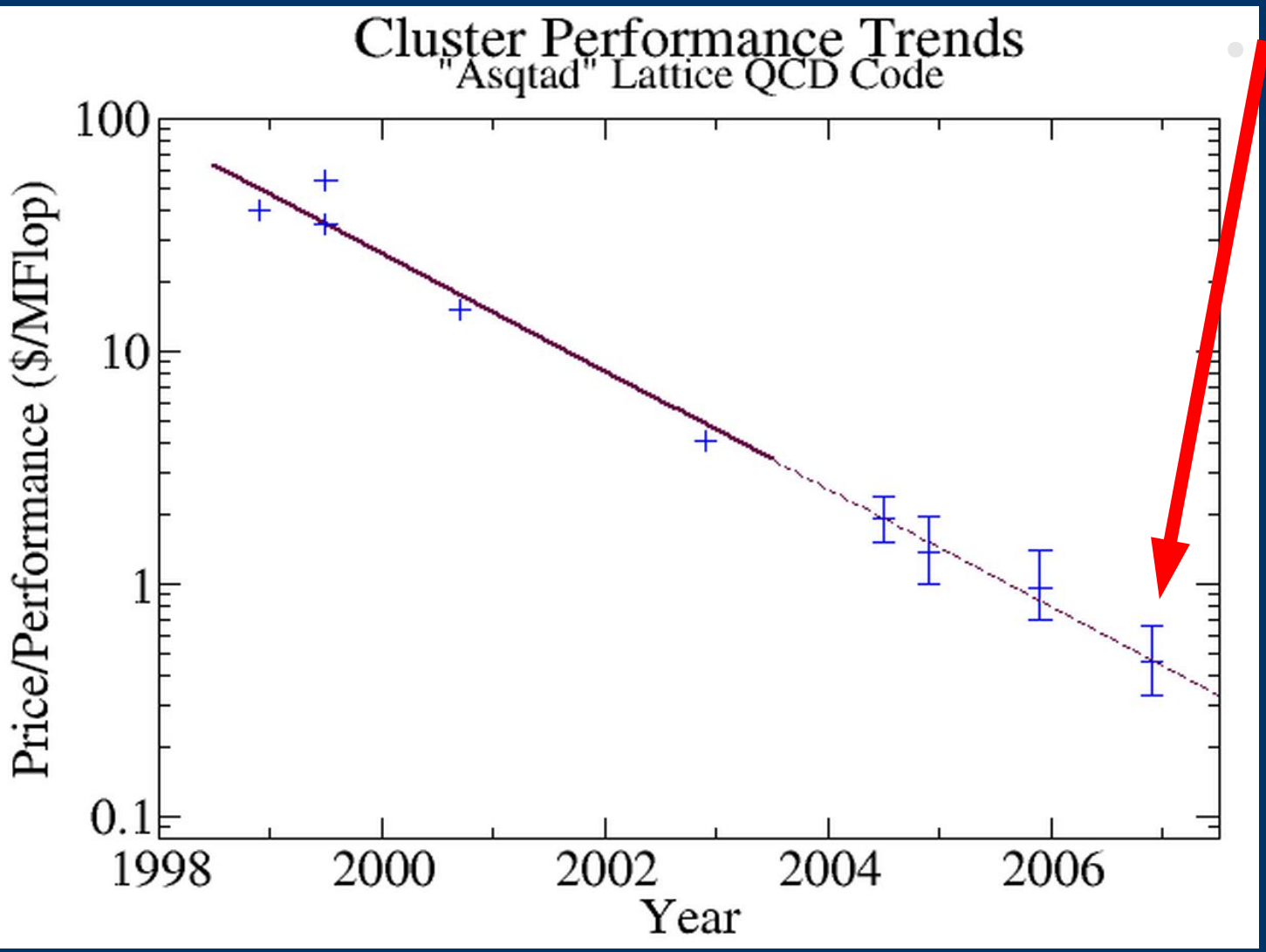
Predictions



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

Future Plans

- Through SciDAC, DOE is funding lattice QCD centers at Fermilab, Jefferson Lab, Brookhaven
- Plans below are for Fermilab systems
- Late 2004:
  - 256 Pentium 4E nodes, PCI-E, Infiniband
  - at least 800 MHz FSB, 1066 MHz if available
- mid-2005:
  - 256 additional P4E nodes (1066 MHz FSB)
  - will expand Infiniband fabric so that jobs can use as many as 512 processors
- 2006:
  - 1024 nodes on Infiniband
    - Dual core Pentium 4?
- 2007:
  - 1024 nodes
  - start yearly refresh, replacing 1/3rd of existing systems
  - note that network fabrics only need upgrading every two refreshes

For More Information

- Fermilab lattice QCD portal:
  http://lqcd.fnal.gov/
- Fermilab benchmarks:
  http://lqcd.fnal.gov/benchmarks/
- US lattice QCD portal:
  http://www.lqcd.org/