# Clusters at Fermilab

Paul Mackenzie
Fermilab Theoretical Physics
mackenzie@fnal.gov

SciDAC Lattice QCD All-Hands Meeting
Brookhaven
March 26-27, 2004

Additional material at:
http://lqcd.fnal.gov.

Fermilab  lattice cluster effort is led by Don Holmgren.

The clusters are currently housed in the New Muon Lab.

The 176 node Pentium 4 system. ~100 GFlops.

# Current clusters

2000:

80 node Pentium III, 700 MHz.
100 MHz memory bus.
Myrinet.

2002:

48 node Pentium 4, 2.0 GHz.
400 MHz memory bus.
Myrinet.

128 node Pentium 4, 2.4 GHz.
Myrinet.

# http://lqcd.fnal.gov

Fermilab Lattice Gauge Theory Computational Facility

http://lqcd.fnal.gov/

Apple .Mac Amazon eBay Yahoo! News ▾ lqcd

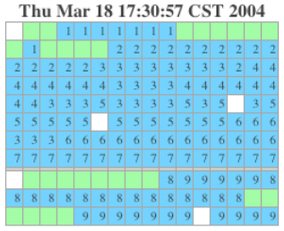
## Fermilab Lattice Gauge Theory Computational Facility

Fermilab Fermilab at Work Theoretical Physics Dept. Distributed Systesms Projects Group Integrated Systems Development Dept. Fermilab Computing Division

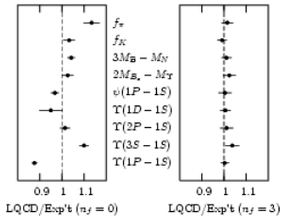
Fermilab operates large clusters of computers for lattice quantum chromodynamics, as part of the national computational infrastructure for lattice QCD established by the Department of Energy. Their goal is the understanding of the strong dynamics of quarks and gluons, which is beyond the reach of the traditional perturbative methods of quantum field theory. A central goal of the groups using the computers is the accomplishment of the calculations required to extract from experiment the fundamental parameters of the Standard Model of particle physics.

LQCD System Status

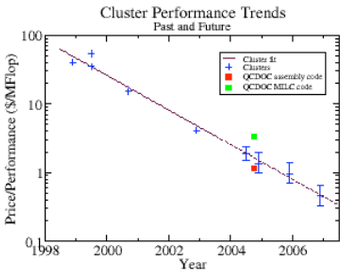
**Thu Mar 18 17:30:57 CST 2004**



The QCD clusters




The QCD cluster future.




Fermilab Lattice QCD

Lattice 2004

Commodity Hardware for Lattice QCD

DOE, SciDAC Support for Lattice QCD Computing

LQCD Information:

- New accounts
- Notes for Users
- Kerberos Notes for Users
- Additional Kerberos Information
- LQCD-Users Mail Archive
- Draft Run Time Environment Specification

System status

- Cluster Status
- Node Status
- Help: lqcd-admin@fnal.gov
- LQCD-Admin Mail Archive

Useful code

Cluster work at Fermilab

Lattice links

Original qcdhome.fnal.gov pages (non-SciDAC Lattice Cluster)

lqcd.fnal.gov PBS Cluster Status

http://lqcd.fnal.gov/cgi-bin/stat

Apple .Mac Amazon eBay Yahoo! News ▾ lqcd

## lqcd.fnal.gov Status

Thu Mar 18 17:34:21 CST 2004

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w101 | | w102 | | w103 | | w104 | 1 | w105 | 1 | w106 | 1 | w107 | 1 | w108 | 1 | w109 | 1 | w110 | 1 | w111 | | | | | | | | | | | | | | | | |
| w112 | | w113 | | w114 | | w115 | | w116 | | w117 | | w118 | 1 | w119 | | w120 | | w121 | | w122 | | | | | | | | | | | | | | | | |
| w123 | 2 | w124 | 2 | w201 | 2 | w202 | 2 | w203 | 2 | w204 | 2 | w205 | 2 | w206 | 2 | w207 | 2 | w208 | 2 | w209 | 2 | | | | | | | | | | | | | | | |
| w210 | 2 | w211 | 2 | w212 | 2 | w213 | 2 | w214 | 3 | w215 | 3 | w216 | 3 | w217 | 3 | w218 | 3 | w219 | 3 | w220 | 3 | | | | | | | | | | | | | | | |
| w221 | 3 | w222 | 2 | w223 | 4 | w224 | 4 | w301 | 4 | w302 | 4 | w303 | 4 | w304 | 4 | w305 | 4 | w306 | 4 | w307 | 3 | | | | | | | | | | | | | | | |
| w308 | 3 | w309 | 3 | w310 | 3 | w311 | 4 | w312 | 4 | w313 | 4 | w314 | 4 | w315 | 4 | w316 | 4 | w317 | 4 | w318 | 4 | | | | | | | | | | | | | | | |
| w319 | 3 | w320 | 3 | w321 | 3 | w322 | 5 | w323 | 3 | w324 | 3 | w401 | 3 | w402 | 3 | w403 | 5 | w404 | 3 | w405 | 5 | | | | | | | | | | | | | | | |
| w406 | | w407 | 3 | w408 | 5 | w409 | 5 | w410 | 5 | w411 | 5 | w412 | 5 | w413 | 5 | w414 | | w415 | 5 | w416 | 5 | | | | | | | | | | | | | | | |
| w417 | 5 | w418 | 5 | w419 | 5 | w420 | 5 | w421 | 5 | w422 | 6 | w423 | 6 | w424 | 6 | w501 | 3 | w502 | 3 | w503 | 3 | | | | | | | | | | | | | | | |
| w504 | 6 | w505 | 6 | w506 | 6 | w507 | 6 | w508 | 6 | w509 | 6 | w510 | 6 | w511 | 6 | w512 | 6 | w513 | 6 | w514 | 6 | | | | | | | | | | | | | | | |
| w515 | 6 | w516 | 6 | w517 | 7 | w518 | 7 | w519 | 7 | w520 | 7 | w521 | 7 | w522 | 7 | w523 | 7 | w524 | 7 | w601 | 7 | | | | | | | | | | | | | | | |
| w602 | 7 | w603 | 7 | w604 | 7 | w605 | 7 | w606 | 7 | w607 | 7 | w608 | 7 | nqcd0101 | | nqcd0102 | | nqcd0103 | | nqcd0104 | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| nqcd0105 | | nqcd0106 | | nqcd0201 | | nqcd0202 | | nqcd0203 | | nqcd0204 | 8 | nqcd0205 | 9 | nqcd0206 | 9 | nqcd0301 | 9 | nqcd0302 | 9 | nqcd0303 | 9 | | | | | | | | | | | | | | | |
| nqcd0304 | 8 | nqcd0305 | 8 | nqcd0306 | 8 | nqcd0401 | 8 | nqcd0402 | 8 | nqcd0403 | 8 | nqcd0404 | 8 | nqcd0405 | 8 | nqcd0406 | 8 | nqcd0501 | 8 | nqcd0502 | 8 | | | | | | | | | | | | | | | |
| nqcd0503 | 8 | nqcd0504 | 8 | nqcd0505 | 8 | nqcd0506 | 8 | nqcd0601 | | nqcd0602 | | nqcd0603 | | nqcd0604 | | nqcd0605 | | nqcd0606 | | nqcd0701 | 9 | | | | | | | | | | | | | | | |
| nqcd0702 | 9 | nqcd0703 | 9 | nqcd0704 | 9 | nqcd0705 | 9 | nqcd0706 | 9 | nqcd0801 | 9 | nqcd0802 | | nqcd0803 | 9 | nqcd0804 | 9 | nqcd0805 | 9 | nqcd0806 | 9 | | | | | | | | | | | | | | | |

free : 27 down : 0 offline : 5 reserve : 0 job-exclusive : 144 job-sharing : 0 Usage 84%

Temperature COOL

### Job List

| Ref Id | Job Id | Job Name | User | Time Used | Time Req | Status | Queue | Nodes |
|---|---|---|---|---|---|---|---|---|
| | 1168 | pQuen | Jim Simone | | 09:50 | Queued | workq | 16 |
| | 1261 | pQuen | Jim Simone | | 09:50 | Queued | workq | 16 |
| 8 | 1270 | pQuen | Jim Simone | 08:16 | 09:50 | Running | workq | 16 |
| 9 | 1274 | pQuen | Jim Simone | 07:22 | 09:45 | Running | workq | 16 |
| 3 | 1301 | gluon_stuf | Matthew Nobes | 16:49 | 24:00 | Running | workq | 24 |
| 7 | 1326 | pQuen | Jim Simone | 04:35 | 09:45 | Running | workq | 16 |
| 5 | 1327 | pQuen | Jim Simone | 04:34 | 09:45 | Running | workq | 16 |
| | 1329 | pQuen | Jim Simone | | 09:45 | Queued | workq | 16 |
| | 1340 | pQuen | Jim Simone | | 09:45 | Queued | workq | 16 |
| | 1341 | pQuen | Jim Simone | | 09:45 | Queued | workq | 16 |
| | 1342 | pQuen | Jim Simone | | 09:45 | Queued | workq | 16 |
| 4 | 1344 | pQuen | Jim Simone | 04:34 | 09:45 | Running | workq | 16 |
| 6 | 1345 | par_4sc.sh | Alan Gray | 04:34 | 24:00 | Running | workq | 16 |

Job list a mixture of perturbation theory and valence propagator calculation and analysis;
1 node, 4 node, 16 node jobs, etc.

# Current program.

Projects approved for running in 2003 on current lqcd.fnal.gov (~100 GF):

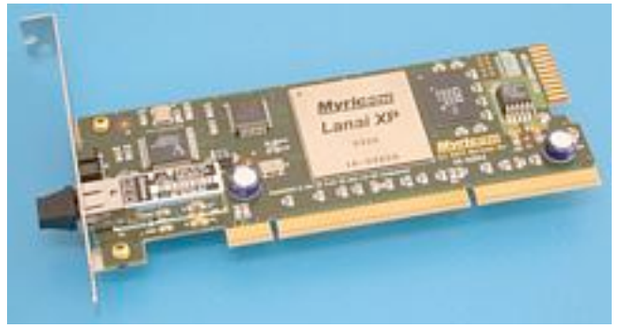
2003-1 C. Bernard et al., B and D Meson Decays with Unquenched Improved Staggered Fermions. 270 x 128 p4 days.
2003-2 C. De Tar et al., Topological Susceptibility and Hybrid Exotic Quarkonium in Full QCD. 23 x 128 p4 days.
2003-3 M. Di Pierro et al., Quarkonium Physics with Unquenched Improved Staggered Fermions. 90 x 128 p4 days.
2003-4 G. P. Lepage, HPQCD, High-Precision Heavy-Quark Physics . 60 x 128 p4 days.

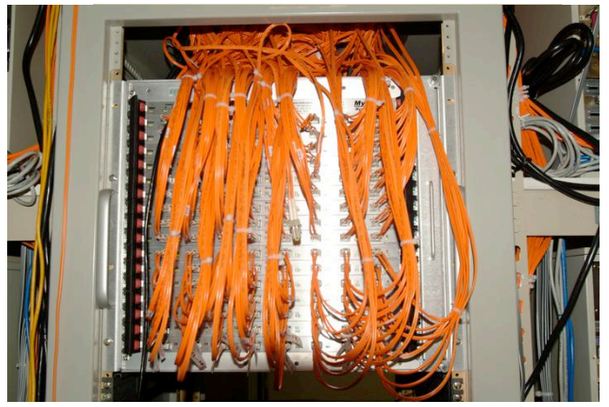
| Configurations | Conf. generation time | Valence and analysis time |
|---|---|---|
| Milc "coarse" ($20^3$ x 64) | 100 GF years | 30 GF years |
| Milc "fine" ($28^3$ x 96) | 640 GF years | 200 GF years? |

# Communications.

## Myrinet:

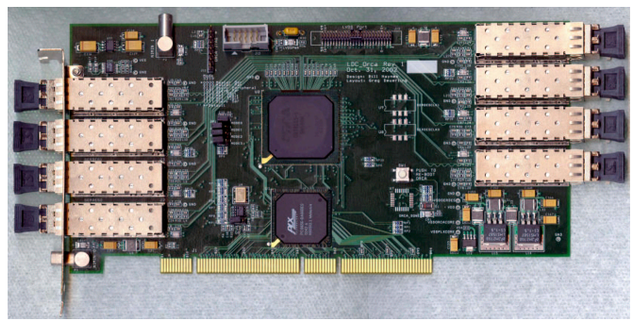
4 sec latency, 450 MB/sec bandwidth.



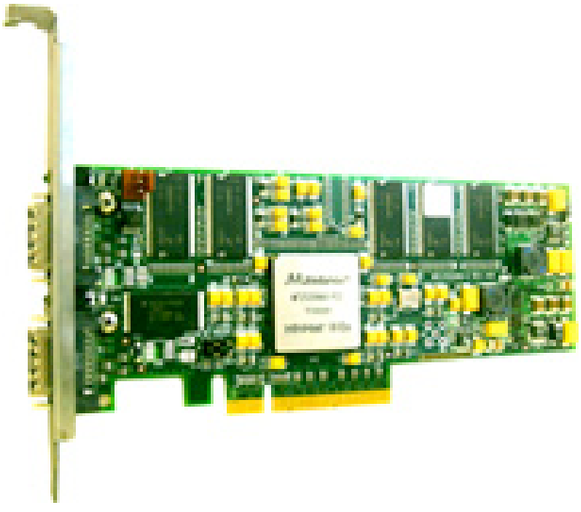
## FPGA:

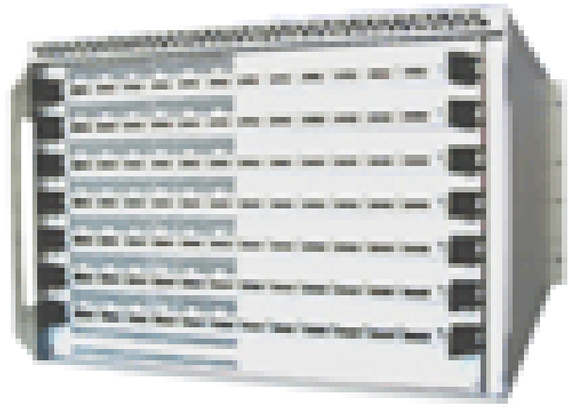
Investigated at Fermilab.
Routable toroidal mesh,
potential bandwith 4 GB/sec,
limited by I/O bus, good latency.



## Infiniband 4X:

6 sec latency, 2 GB/sec.





## GigE mesh: Under investigation at JLab.

Toroidal mesh, 1 GB/sec. Good latency.

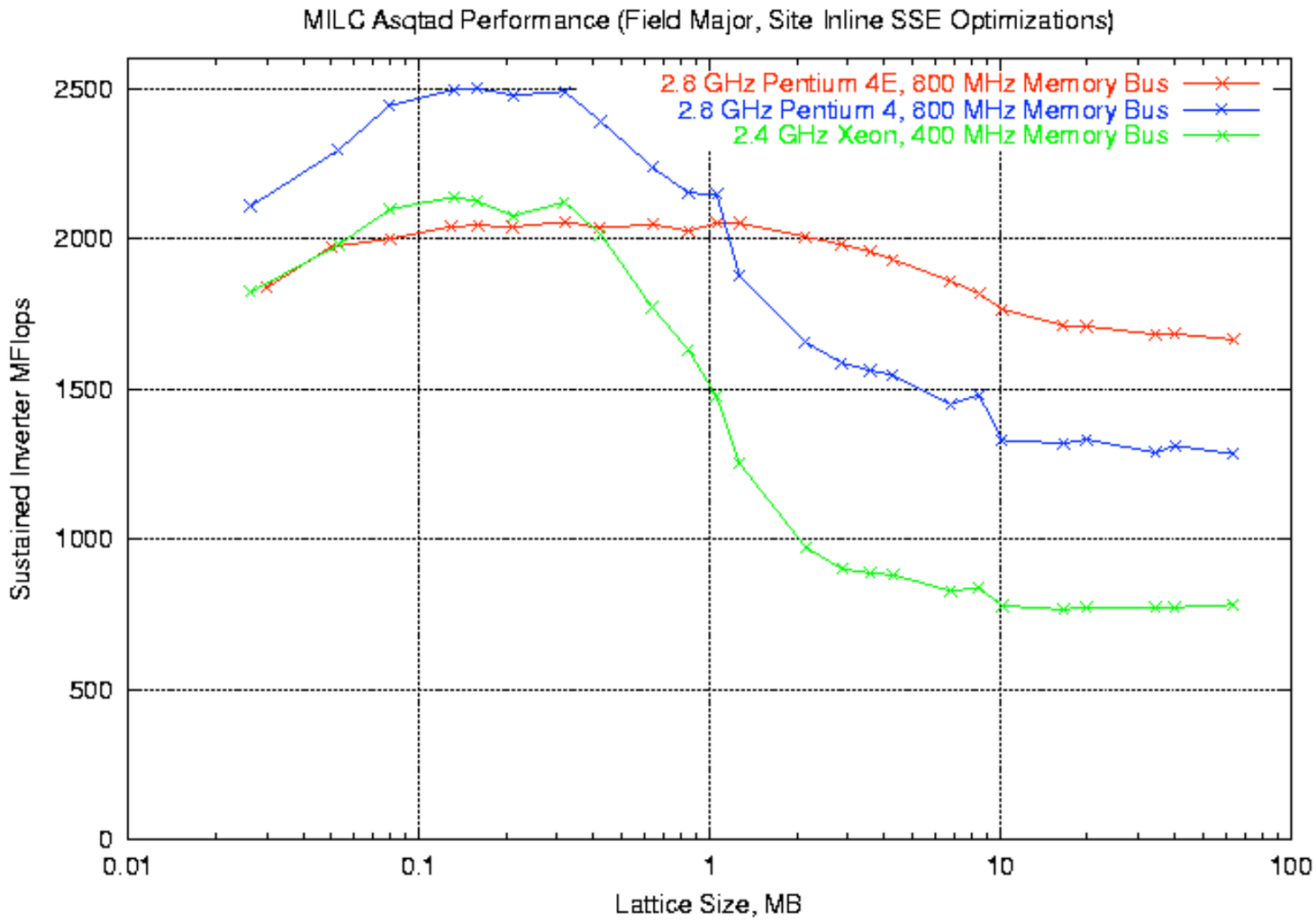
## I/O buses: PCI 32/33 132 MB/sec.; PCI-X 64/133 1 GB/sec.; PCI Express 4X 2 GB/sec.

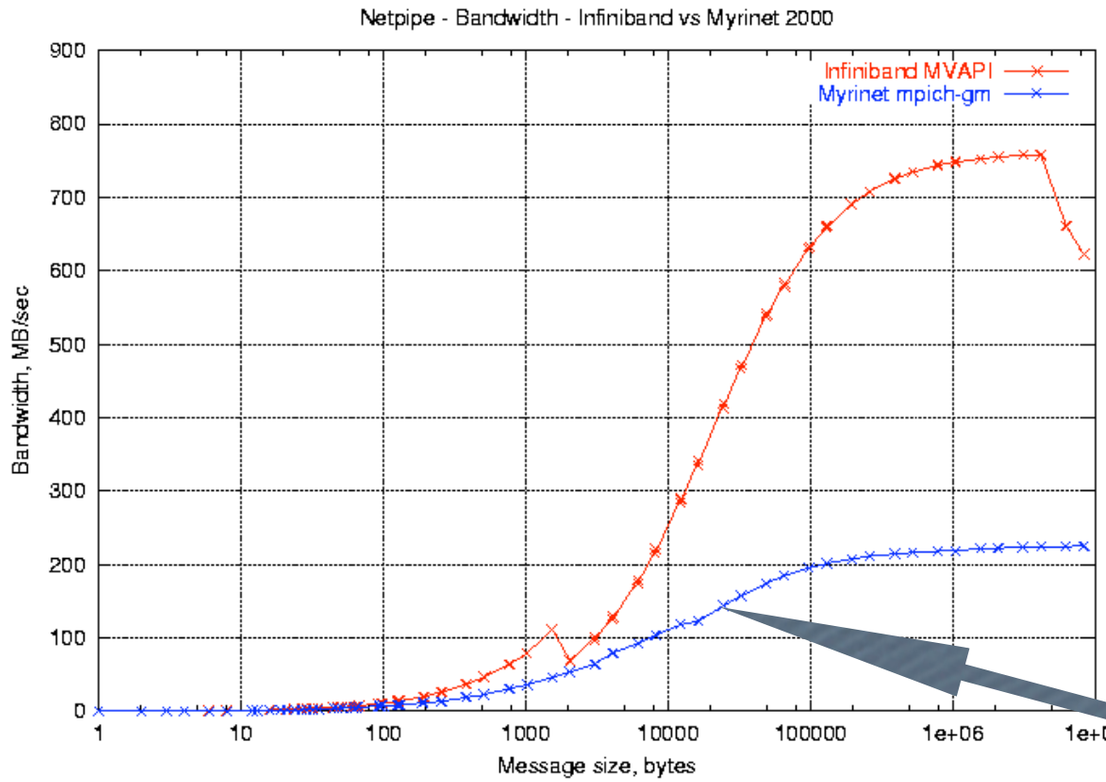
# Benchmarks

2.8 GHz P4E

2.8 GHz P4

2.4 GHz Xeon




## Infiniband and Myrinet with PCI-X.

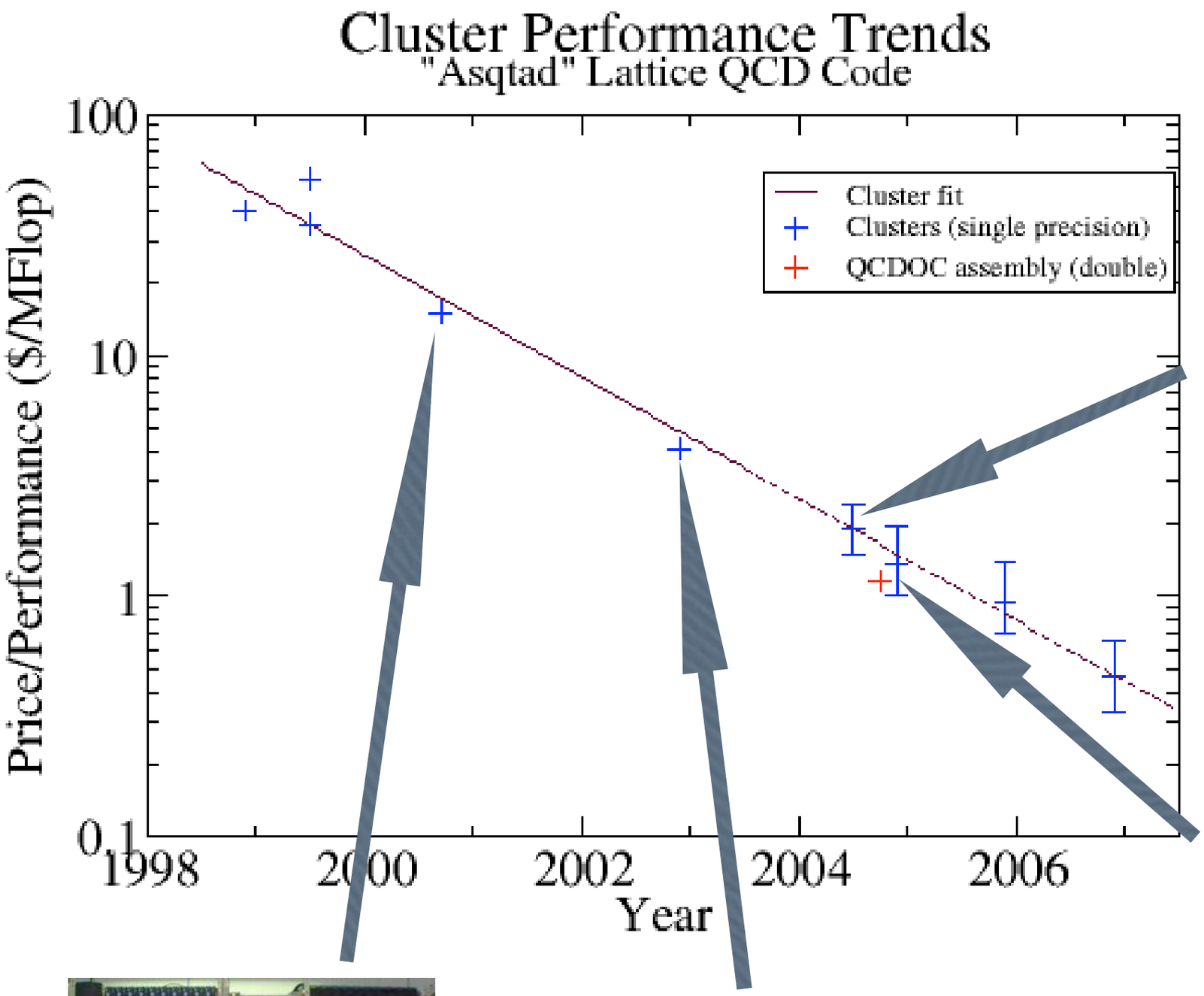
Our typical message size: few $10^4$.

Vendor Roadmaps:

2004: Q1 P4E,
Q3 1 →2 GB/sec, PCI Express.

2005: H1 800→1066 MHz memory bus.

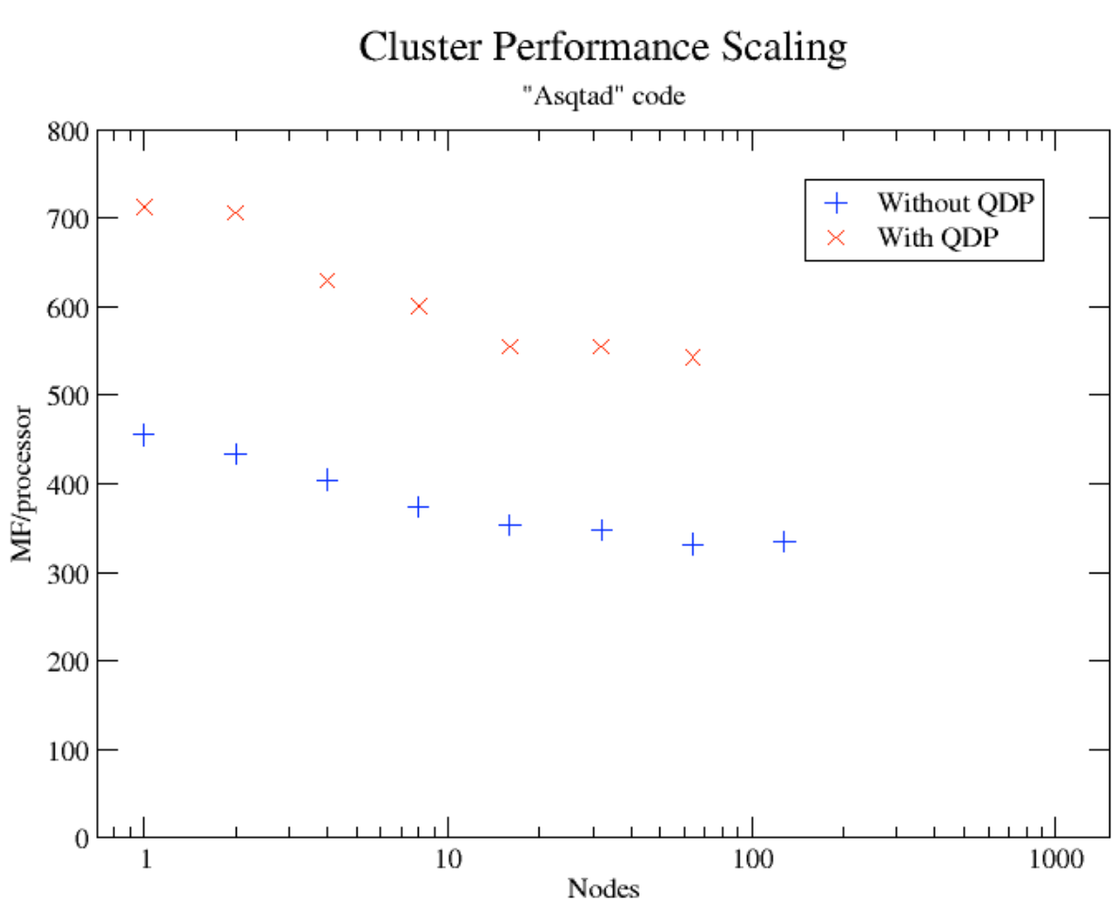
2006: H1 Fully buffered dimm memory bus, dual core processors. Infiniband integrated in motherboard?

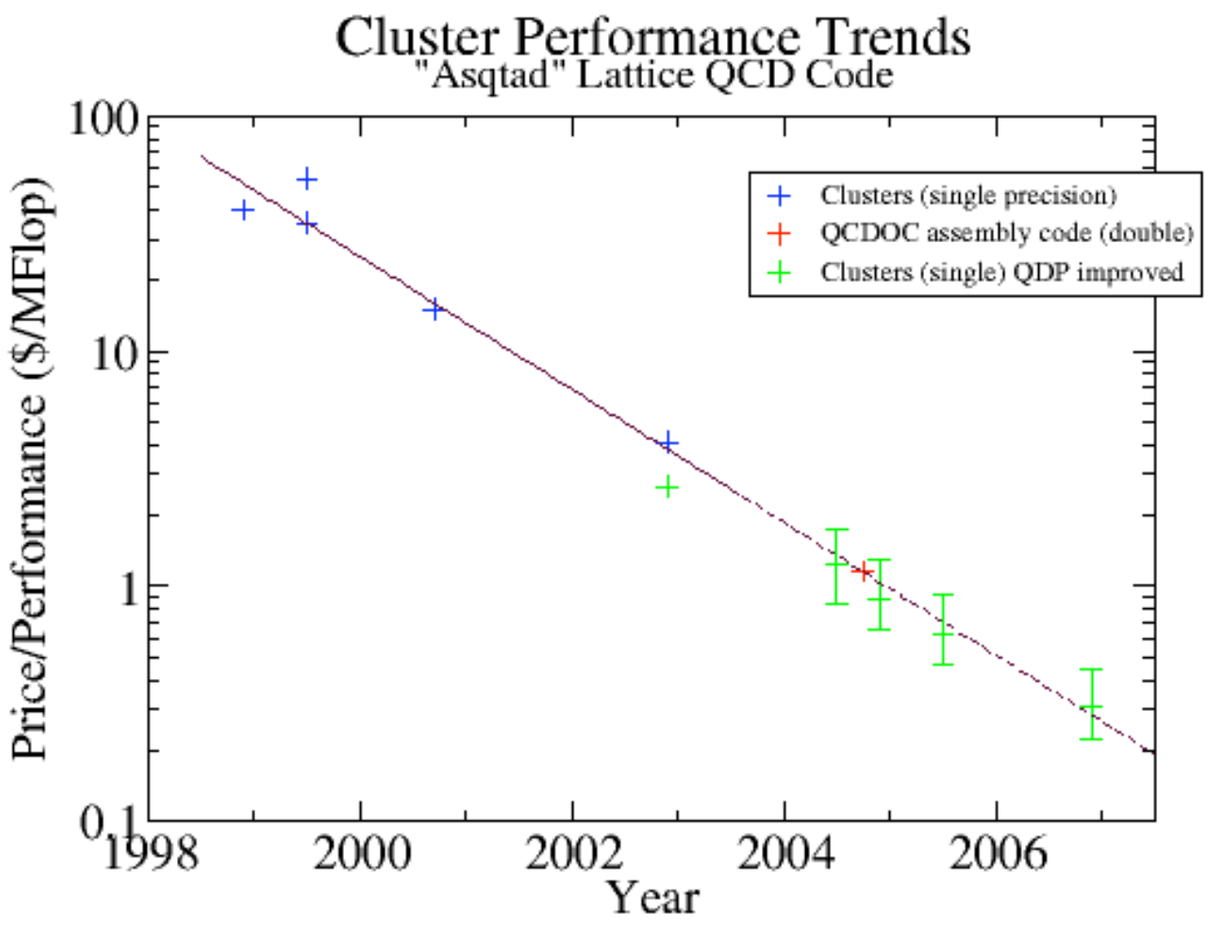



2004:
128 P4E singles,
reuse Myrinet switch.
Incremental cost:
$1/MF.

256 node, P4E singles?
Infiniband switch.

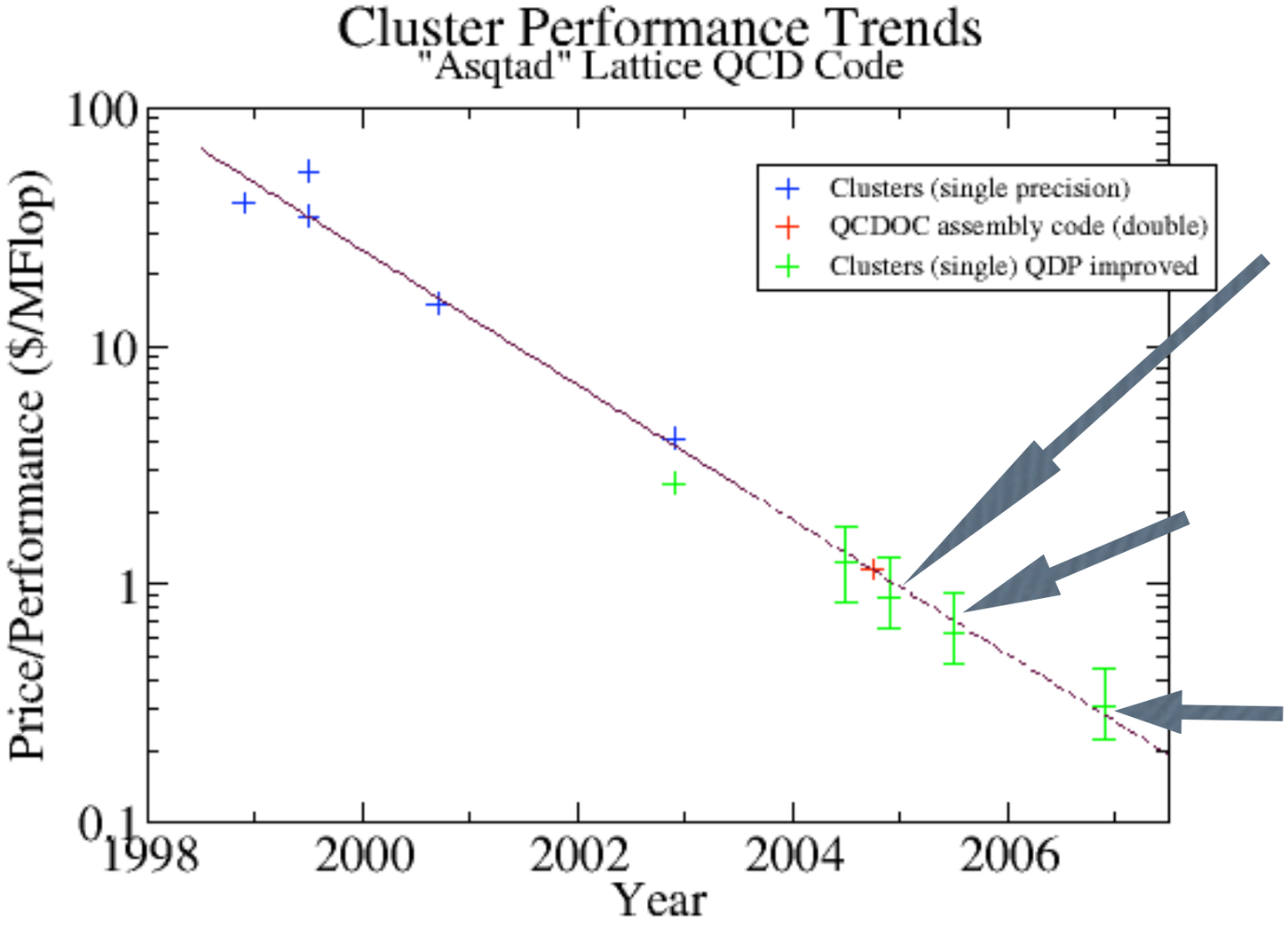
# QDP optimizations (Osborn): reorder loops, eliminate redundant communications, reorder communications...




Uncertainties in extrapolations. Vendor roadmaps, Infiniband acceptance, pricing...

# Cluster Roadmap




2004: 256 node,
$0.5M, 0.35 TF. (SciDAC.)

2005: 512 node,
$1M, 1 TF.

2006: 1024 node,
$1.5 M, 3 TF (or double).

In February, Bob Sugar told HEPAP that the minimum support required for the health of the field was $3M/year. In 04/05, this would be spent on:

- a $1M, 1 TF cluster (less damages the cluster program),
- a $5M, 5 TF QCDOC (ready to go for big machines right now).

In 06 and 07, clusters look to be ready to provide larger systems even more cost-effectively.

We at Fermilab strongly support this program, and hope that we can all help our friends in DoE-HEP to convince DoE Nuclear and Computing that they should support it, too.