Lattice Gauge Computing Summary

Don Holmgren
Fermilab

Outline

- Communities / Collaborations
- Existing Facilities
- LQCD on Purpose-Built Machines
- LQCD on Clusters

Lattice Gauge Computing Requirements

- Very large floating point requirements:
  - Need 100's to 1000's of Gflop/sec-years for results
  - Primary work is inverting large 4-D sparse matrices
  - Sustained performance is often low (20% of peak)
  - Multiprocessors are essential – clusters or massively parallel
- Inter-processor communications requirements:
  - High bandwidth (100's Mbyte/sec)
  - Low latency - O(μsec)

Communities / Collaborations

- LATFOR (Lattice Forum) (K. Jansen)
  - Forum of German lattice physicists, plus groups in Austria (Graz, Wien) and Switzerland (Bern)
  - Universities: Berlin, Bielefeld, Darmstadt, Leipzig, Műnster, Regensburg, Wuppertal
  - Research Labs: DESY, GSI, NIC
- Efforts:
  - Coordinate physics program
  - Share software
  - Share configurations/propagators (ILDG = International Lattice Data Grid)

Communities / Collaborations

- LATFOR - Requirements
  - Typical application profile:

Lattice Size	32^3 • 64
Memory	100 Gbyte
I/O request	0.1 Gbyte/Gflop
Runtime	15 Teraflops-years

  - Need 12.5 Teraflops sustained

Communities / Collaborations

- SciDAC Lattice Gauge Computing
  - Majority of US community
  - Labs: Brookhaven, Fermilab, Jefferson Lab
  - Universities: Indiana, Utah, UCSB, UCSD, MIT, BU, ASU, Illinois, OSU, Columbia…
  - http://www.lqcd.org/
- Efforts
  - QCDOC (Columbia/BNL)
  - Cluster Prototypes (Fermilab, JLAB)
  - Software (QMP, QLA, QIO, QDP) to allow applications to run on either type of hardware (Chulwoo Jung)
  - O(10 Tflops) in 2004 (QCDOC), 2005/6 (clusters)

Facilities

- SciDAC:
  - Jefferson Lab

    Now: 128 nodes / 128 processors (2.0 GHz Xeon), Myrinet
    Soon: 256 nodes / 256? Processors (Xeon), GigE mesh
  - Fermilab

    Now: 176 nodes / 352 processors (2.0/2.4 GHz Xeon), Myrinet
    Autumn: O(256) processor expansion (Xeon? Itanium2?)
    Myrinet? GigE?

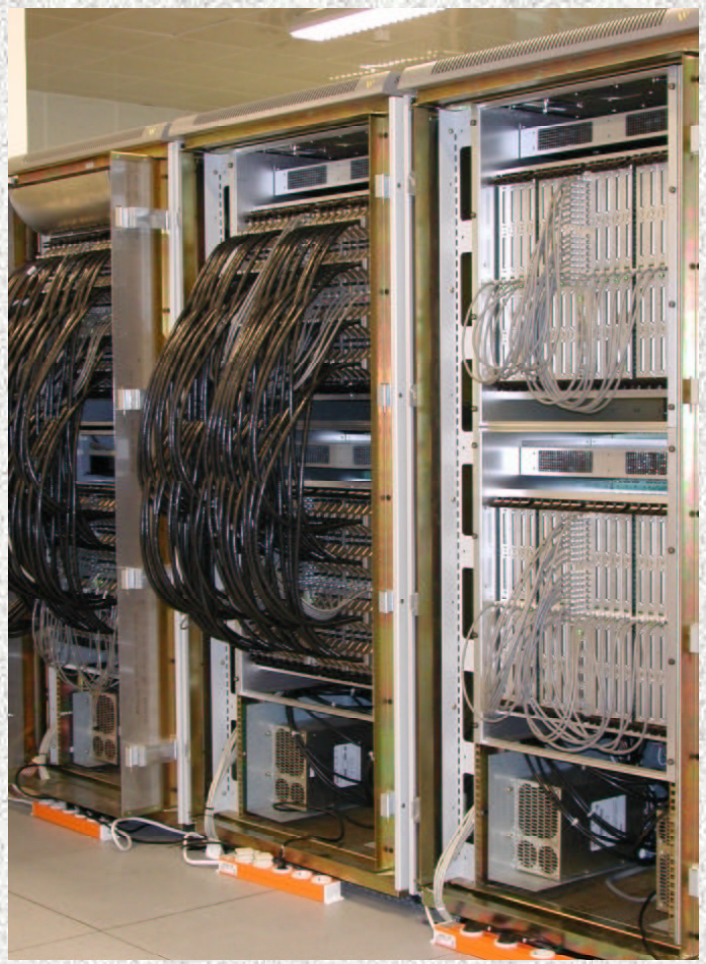
Facilities

- DESY Clusters:
  - DESY Hamburg
    16 Dual 1.7 GHz Xeon
    16 Dual 2.0 GHz Xeon
    Myrinet
  - DESY Zeuthen
    16 Dual 1.7 GHz Xeon
    Myrinet

APEmille installations

- Bielefeld 130 GF (2 crates)
- Zeuthen 520 GF (8 crates)
- Milan 130 GF (2 crates)
- Bari 65 GF (1 crates)
- Trento 65 GF (1 crates)
- Pisa 325 GF (5 crates)
- Rome 1 520 GF (8 crates)
- Rome 2 130 GF (2 crates)
- Orsay 16 GF (1/4 crates)
- Swansea 65 GF (1 crates)

- **Gr. Total ~1966 GF**



LQCD On Purpose-Built Hardware

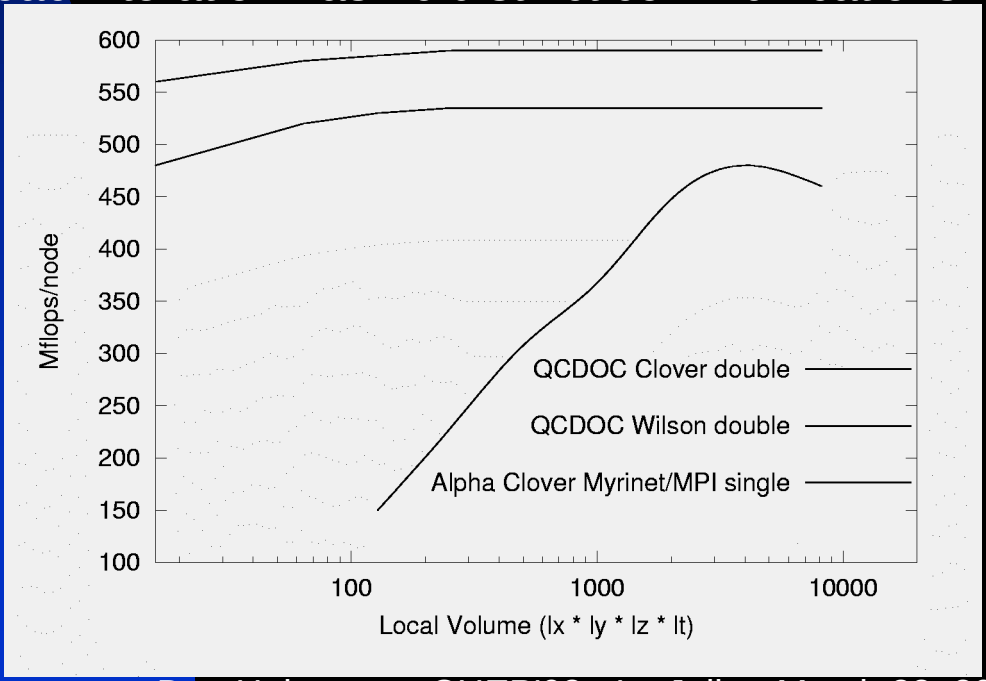
- QCDOC: (T. Wettig)
  - "QCD on a Chip" ASIC (0.18μ):
    - Partner with IBM
    - Based on PPC940 + 64-bit FPU, 1 Gflops peak
    - 4 MB EDRAM + controller for external DDR
    - 6 bidirectional LVDS channels, each 2 X 500 Mbps
    - Ethernet
    - ~ 5 Watts
    - $1/MF assuming 50% of peak
  - Packaging:
    - 2 ASICs/daughterboard, 32 daughterboards per motherboard
    - 8 motherboards per backplane (512 processors)

QCDOC

- Schedule:
  - Funded: 10 Tflops each in late 2003 for Edinburgh, RIKEN-BNL
    5 Tflops for Columbia (SciDAC) [pending]
  - 10 – 20 Tflops for US @ BNL in 2004
  - ASIC design is done, 1st chips in May

QCDOC

- Performance: (P. Boyle)
  - Single node kernels, based on simulations:
    - SU3 x SU3 800 Mflops/node
    - SU3 x 2spinor 780 Mflops/node
  - Multinode:
    - Example – Wilson 2^4, 470 Mflops/node, 22µsec/iteration, each iteration has 16 distinct communications



QCDOC

- Fast global sums via network passthru:

Estimated Scalability

Based on cycle accurate simulation.
Wilson CG performance on a $32^3 \times 64$ Lattice.
Clover and Domain Wall Fermions even more scalable.

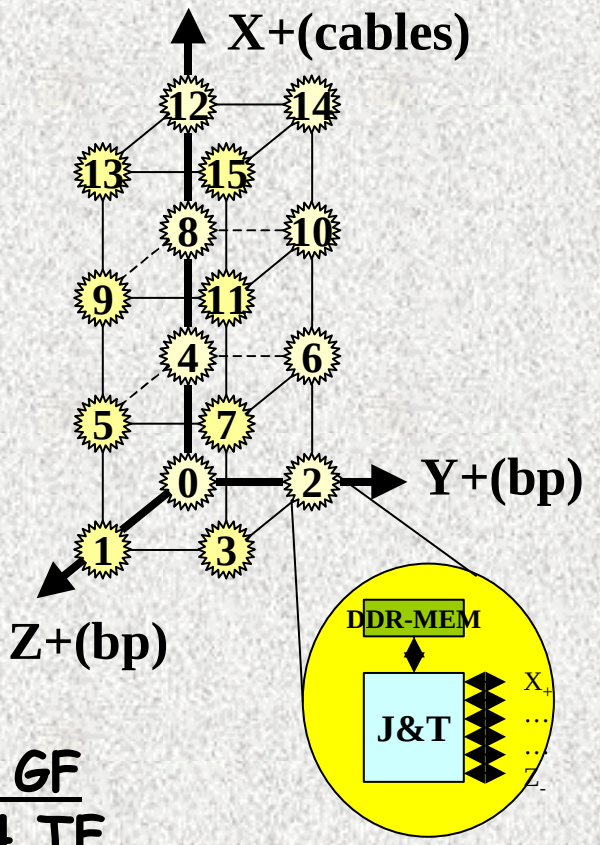
Nodes	$M^\dagger M$	Gsum	Sust. Tflops
4096	2620μs	10 μs	2.15
8192	1310μs	11.5μs	4.2
16384	680 μs	13 μs	8.1
32768	340 μs	15 μs	15.6

APEnext (R. De Pietri)

- In 0.18μ ASIC, native implementation of “fused multiply-add”:
  $a \times b + c$ (complex numbers)
- 256 registers, each holding a complex number (double precision)
- 200 MHz, ~ 5 watts, 1.6 Gflops/chip, $1/Mflop @ 50% of peak
- 6+1 channels of LVDS communications, each 200 Mbyte/sec
  - 3D torus (*x, y, z*)
  - Fast I/O link to UNIX host
- Schedule:
  - 300 to 600 ASIC’s to be delivered in June `03
  - From these, assemble and test 256-processor crate
  - Mass production to start in Sept `03

The apeNEXT architecture (2)

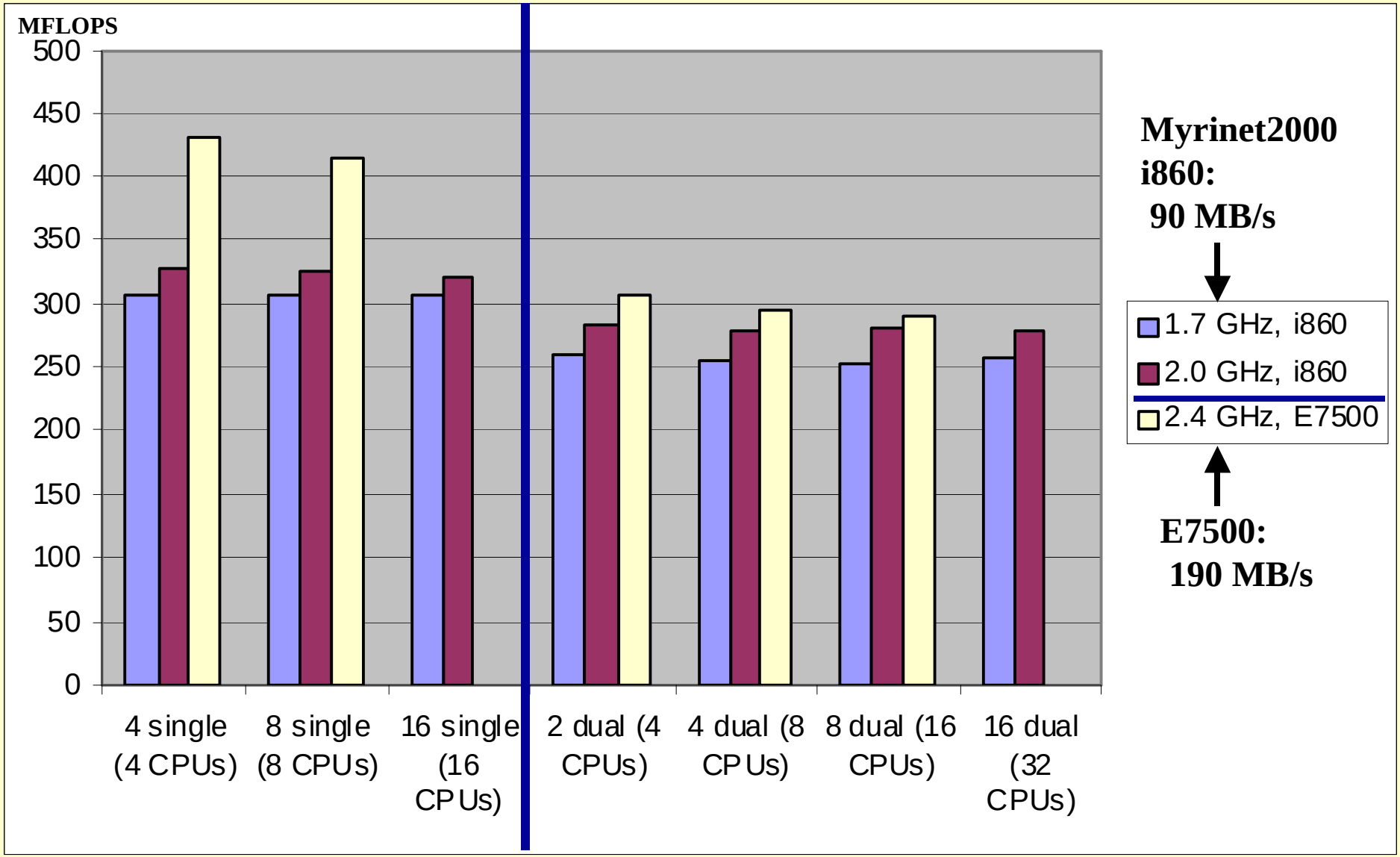
- Two directions (Y,Z) on the backplane
- Direction X through front panel cables
- **System topologies**:
  - Processing Board 4 x 2 x 2 ~ 26 GF
  - subCrate (16 PB) 4 x 8 x 8 ~ 0.4 TF
  - Crate (32 PB) 8 x 8 x 8 ~ 0.8 TF
  - Large systems (8*n) x 8 x 8



LQCD on Clusters

- Performance (DESY: P. Wegner, Fermilab: D. Holmgren)
- Common themes:
  - SSE/SSE2 very important for optimizing performance
  - Large memory bandwidth increase in Pentium 4 critical to LQCD
  - Building clusters with Myrinet, investigating other networks
  - Pay attention to PCI bus performance, varies with chipset

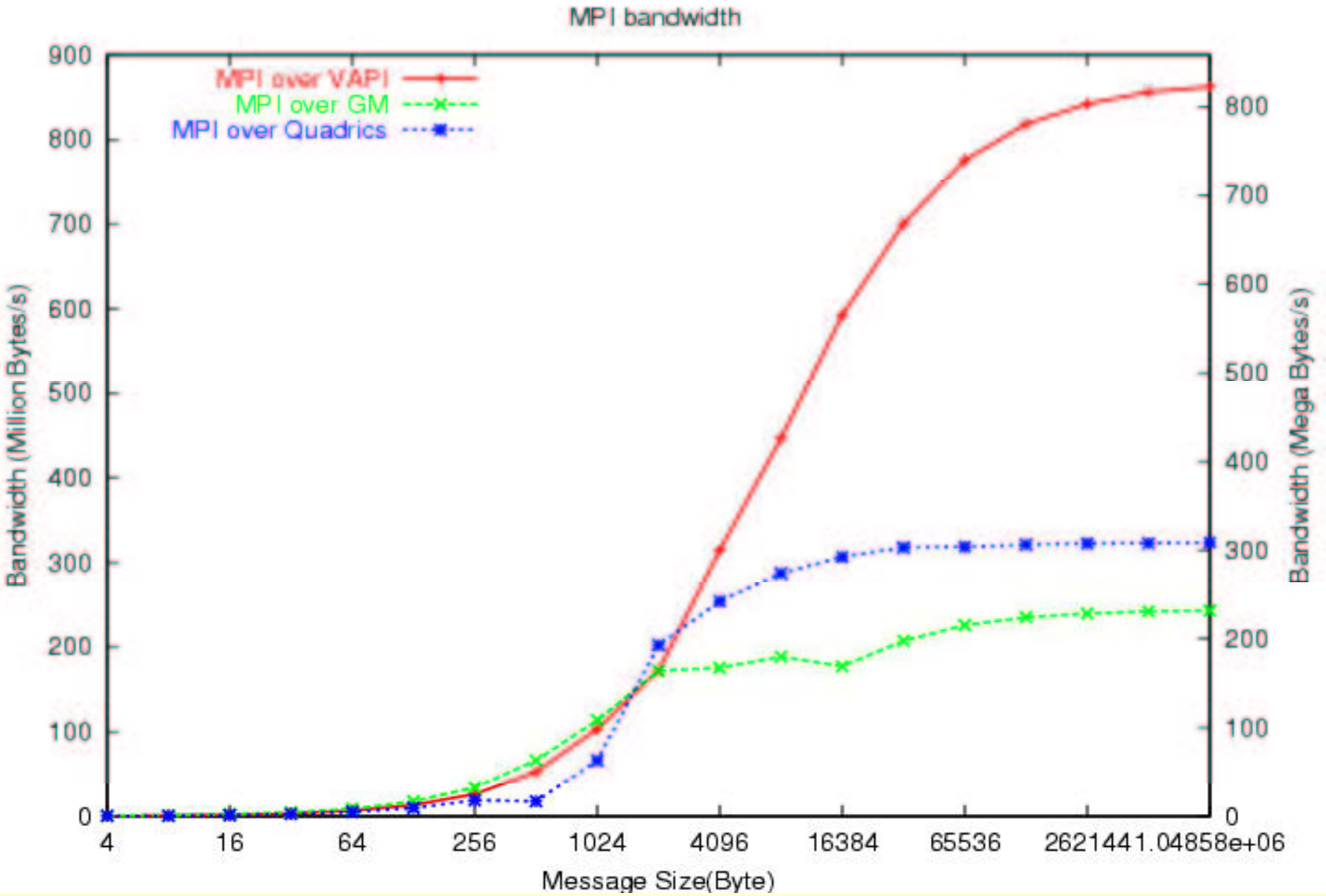
Parallel (1-dim) Dirac Operator Benchmark (SSE), even-odd preconditioned, 2 x 16^3 , XEON CPUs, single CPU performance



Maximal Efficiency of external I/O

	MFLOPs (without communication)	MFLOPS (with communication)	Maximal Bandwidth	Efficiency
Myrinet (i860), SSE	579	307	90 + 90	0.53
Myrinet/GM (E7500), SSE	631	432	190 + 190	0.68
Myrinet/ Parastation (E7500), SSE	675	446	181 + 181	0.66
Myrinet/ Parastation (E7500), non-blocking, non-SSE	406	368	hidden	0.91
Gigabit, Ethernet, non-SSE	390	228	100 + 100	0.58
Infiniband non-SSE	370	297	210 + 210	0.80

Infiniband interconnect



LQCD on Clusters - Management

- See talks by A. Gellrich, A. Singh
- Common themes:
  - Linux, PBS, Myrinet, MPI, private networks
  - Web based monitoring and alarms
  - MRTG history plots (*e.g.* temperatures, fans)
  - Some Myrinet hardware reliability issues