

Large scale commodity clusters for lattice QCD*

A. Pochinsky^a, W. Akers^b, R. Brower^c, J. Chen^b, P. Dreher^a, R. Edwards^b, S. Gottlieb^{d,e},
D. Holmgren^e, P. Mackenzie^e, J. Negele^a, D. Richards^b, J. Simone^e, W. Watson^b

^aCenter for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

^bThomas Jefferson National Accelerator Facility, Newport News, VA 23606, USA

^cPhysics Department, Boston University, Boston, MA 02215, USA

^dDepartment of Physics, Indiana University, Bloomington, IN 47405, USA

^eTheory Group, Fermi National Accelerator Center, Batavia, IL 60510, USA

We describe the construction of large scale clusters for lattice QCD computing being developed under the umbrella of the U.S. DoE SciDAC initiative. We discuss the study of floating point and network performance that drove the design of the cluster, and present our plans for future multi-Terascale facilities.

1. INTRODUCTION

Over the last several years, the intense competition for high performance desktop computers has lead to enormous gains in processor performance. At the same time, the market for specialized interconnects to support parallel computing has made it possible to assemble cost effective clusters from commodity components with the performance required for QCD calculations.

Advantages of the cluster approach are manifold:

- Leverage computer industry's continued development of commodity processors, system boards and computer network engineering.
- Exploit commodity software and benefit from the Free Software Movement (Linux, GNU compilers and development tools, MPI, ...).

*This work is supported in part by the U.S. Department of Energy Cooperative Agreement DE-FC02-94ER40818, Contract DE-AC05-84ER40150 under which the Southeastern Universities Research Association (SURA) operates the Thomas Jefferson National Accelerator Facility (TJNAF), Grant Proposal DE-FG02-91ER40676, Grant DOE FG02-91ER 40661, and Contract DE-AC02-76CH03000 under which Universities Research Association, Inc. operates the Fermi National Accelerator Laboratory.

- Easily adopt improvements in silicon technology and processor architecture.
- Run legacy codes effortlessly.
- Support a wide range of computational models.
- Attain price/performance on lattice codes down to \$1/MF on single node Pentium 4's.
- Clone, upgrade continuously and cheaply.
- Maintain a steady state: continual upgrades, several thousand nodes, replace oldest third of system each year.

2. EXISTING CLUSTERS

Both Cornell/FNAL/MILC and the JLab/MIT collaborations deployed pilot clusters in recent years. The Cornell/FNAL/MILC QCD80 cluster is as follows:

- Dual Pentium III nodes with 700MHz processors and 100MHz SDRAM
- 80 node cluster
- Myrinet 2000 switch (128 ports): up to 64 simultaneous links with a bandwidth of 90 MB/sec each and latency under 10 μ s.

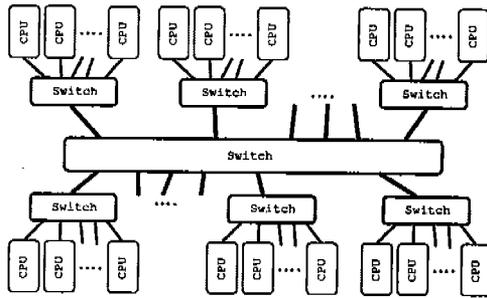


Figure 1. Switch based cluster

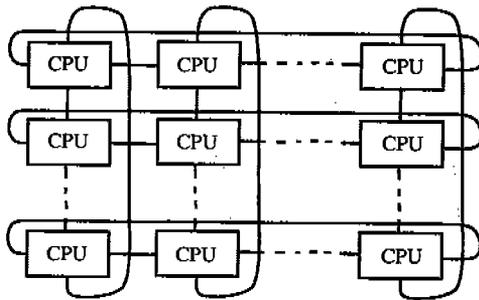


Figure 2. Point-to-point grid cluster

JLab in collaboration with MIT prototyped DEC Alpha based clusters. Three machines were built and used extensively for lattice calculations:

- 12 node system at MIT: 4-way alpha SMP nodes, 64GFlops peak. System in production use since summer 1999,
- 12 node system at JLab: 2-way alpha SMP nodes, 21GFlops peak. System in production use since 2000,
- 16 node cluster of single processor alphas at JLab, 21 GFlops peak. System in production use since 1999.

In all prototype clusters Myrinet is used as the cluster network.

3. SciDAC CLUSTER FACILITIES

Both FNAL and JLab are extending their clusters significantly.

3.1. FNAL

- Initial procurement—July 2002
 - 48 dual 2.0 GHz Xeon nodes
 - 1GB memory per node, interleaved 200MHz DDRAM
 - E7500 chipset
 - Myrinet LANai 9 NICs (133MHz)
 - 2-D GigE mesh by late summer 2002.
- Second procurement—Fall 2002
 - 128 dual 2.4GHz Xeon nodes
 - 1GB memory each, interleaved 233MHz DDRAM
 - Myrinet LANai 9 NICs (133MHz)
- FY03: further procurement, considering 256 dual Xeons with Myrinet as one possibility.

3.2. JLab

- Initial procurement — installed August 2002
 - 128 single processor 2.0 GHz Xeon nodes
 - 512MB memory each
 - Myrinet
- Late FY02 procurement — based on GigE studies will procure 192 node GigE cluster ($4 \times 6 \times 8$ 3-D torus)
- FY03 procurement — considering 384 node GigE cluster ($6 \times 8 \times 8$ 3-D torus).

3.3. System Administration

To make the most efficient uses of the clusters and make them into user facilities for the lattice community, both laboratories provide extensive operational support:

- 1 system administrator
- 1 technician
- 2 system programmers

	FY02	FY04	FY05	FY06
CPUs/Node	2	2	4	4
GFlops/Node	2.0	3.2	8.0	10.0
Nodes		192	384	512
Link Bandwidth (MB/s)		300+300	2(400+400)	2(500+500)
Link Latency μ sec	10	6	5	4
Performance (TFlops)		0.6	2.5	4.5
Hardware Cost (\$M)		0.7	2.5	3.2
\$/MFlops	2.0	1.2	0.9	0.7

Table 1
Cluster Parameters.

4. COMMUNICATION NETWORK

Until now, Myrinet network has been used for intercluster communication, because it presented the most cost effective solution for typical LQCD communication problems. Logically Myrinet is a fat tree with the following characteristics:

- Bandwidth 110MB/sec, going up to 240MB/sec on PCI64B card,
- Latency 9μ sec, going down to 7μ sec on PCI64B card.

Since fixed size switches are used to construct the tree, the latency grows logarithmically with the cluster size.

Fast development of Gigabit Ethernet and its adoption as a commodity network provides an alternative cluster topology [1]. Pentium 4 motherboards are available in a configuration with up to six PCI slots, which allows one to construct two and three dimensional tori, as shown on Fig. 2. The present generation of GigE NIC's reduces the cost per node by roughly 30% and shows promising performance [2]:

- Bandwidth upto 110+110MB/sec
- Latency 2.4μ sec, possibly going down to below 1μ sec.
- Upto 6 links per node for aggregate bandwidth of 660+660MB/sec.

5. PERFORMANCE

Compared to previous generations of IA32, Pentium III features the Streaming SIMD Extension (SSE) which has been further extended to double precision (SSE2) on Pentium 4. Once the utility of SSE for LQCD had been pointed out [3], it became the *de facto* standard for high performance codes. For a problem residing in L2 cache, the Dirac operator runs close to 2GFlops on 1.7 GHz Pentium 4. There are indications that sustained performance near 5GFlops on 1.8 GHz Pentium 4 is possible, see [2] for details.

6. FUTURE PLANS

Table 1 gives the most significant cluster parameters, and the expected size and cost of the hardware at each of the cluster sites. Actual procurement may differ, using a build-to-cost strategy and market optimizations. The price/performance ratio in \$/MFlops and the overall performance assumes a full-cluster job with Wilson fermions. Performance estimates are based on Moore's law.

REFERENCES

1. Z. Fodor, S.D. Katz, G. Papp, *A scalable PC-based parallel computer for lattice QCD*, in these proceedings. See also hep-lat/0209115.
2. //www.mit.edu/~avp/lqcd/GigE/report.pdf
3. M. Lüscher, private communication.

