PROCEEDINGS OF SCIENCE

U.S. Lattice Clusters and the USQCD Project

Donald Holmgren*
Fermi National Accelerator Laboratory
Batavia, Illinois, USA
E-mail: djholm@fnal.gov

We describe the lattice QCD clusters constructed at Fermilab and at Jefferson Lab since Lattice 2004. We also describe the plans for the next several years of the USQCD Project for new computer hardware and machine operations.

XXIIIrd International Symposium on Lattice Field Theory
25-30 July 2005
Trinity College, Dublin, Ireland

*Speaker.

PoS(LAT2005)105

1. Introduction

Since 2001, the U.S. Department of Energy has supported the development of prototype commodity clusters for lattice QCD calculations through the SciDAC (Scientific Discovery through Advanced Computing) program[1]. This program has also supported the design and implementation of software API's that allow lattice QCD applications to run without modification on a great variety of hardware platforms, including all parallel computers supporting MPI, clusters based on switched and toroidal mesh networks, and the QCDOC.

The SciDAC prototype machines housed at Fermilab and Jefferson Lab that are currently used for lattice QCD production have all been based on Intel ia32 microprocessors. These machines have used a variety of network architectures, including Myrinet, Infiniband, and toroidal gigabit ethernet meshes.

In 2004 and 2005, via a separate supplemental grant, the Department of Energy funded the construction of the U.S. QCDOC at Brookhaven National Laboratory. The High Energy Physics, Nuclear Physics, and Advanced Scientific Computing Research program offices contributed to the grant. In 2005, this same grant, along with SciDAC and Fermilab contributions, is funding the construction of a large Infiniband cluster at Fermilab.

Beginning in fiscal year 2006, the U.S. Department of Energy will fund the four year Lattice QCD Computing Project (USQCD). This project will build or purchase a series of computing systems for lattice QCD, and it will operate these systems as well as the QCDOC and the SciDAC clusters.

2. Processor and Network Performance

In the past year, Intel introduced a number of new ia32 microprocessors. Two were of particular interest for lattice QCD clusters. The "Nocona" Xeon cpu is the first SMP-capable processor with an 800 MHz memory bus. Its performance and specifications match those of the earlier Pentium 4 "Prescott" (P4E) processor. The "6xx" Pentium 4 series is available with either 800 MHz or 1066 MHz memory buses, though the latter is cost prohibitive for lattice QCD clusters. These processors have 2 MB of L2 cache, compared with the 1 MB on "Prescott" and "Nocona". Fig. 1 shows the relative performance of "6xx" and P4E processors compared with the 400 MHz memory bus Xeon from several years ago. Fig. 2 shows the relative performance of the "6xx" Pentium 4 compared with the fastest available AMD Opteron processor and an older 2.0 GHz G5 processor (2.5 GHz G5 processors are now available).

Network fabrics, like processors, have continued to increase in performance and decrease in cost. Fig. 3 shows the bidirectional bandwidth as a function of message size on Infiniband equipment purchased in 2005, and on Myrinet 2000 equipment purchased in 2002. The two Infiniband curves show the performance difference between a high level protocol, MPI, and a network specific protocol, VAPI. VAPI has a distinct advantage in the message size region of interest to lattice QCD (order 1K to order 100K bytes). The SciDAC project will implement a VAPI native version of their QMP communications API to take advantage of this.

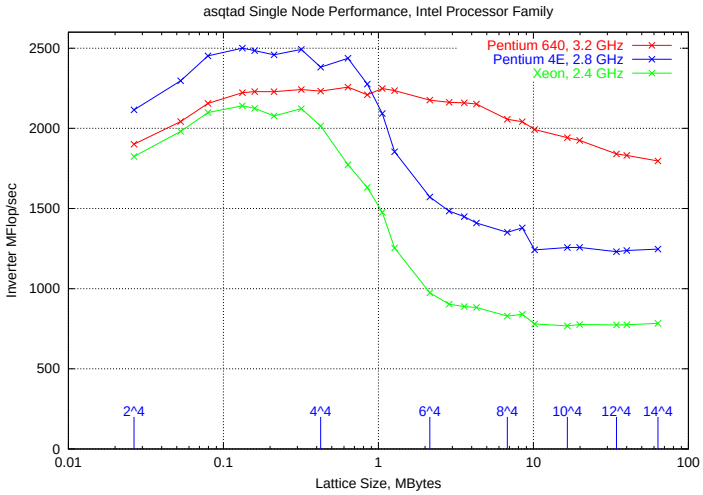

Figure 1: Single node MILC "asqtad" inverter performance as a function of lattice size.

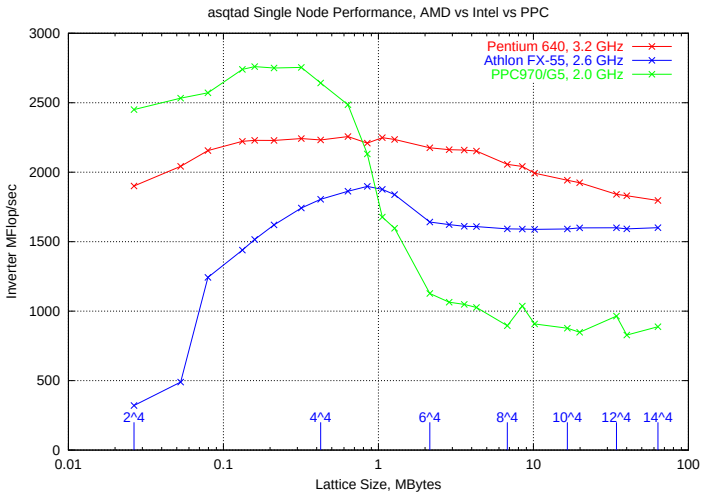

Figure 2: Single node MILC "asqtad" inverter performance as a function of lattice size.

3. Jefferson Lab

The newest SciDAC cluster at Jefferson Lab came online in December 2004. Its specifications are:[2]

- 2.8 GHz Intel Xeon "Nocona" processors, 1 MB L2 cache, 800 MHz FSB
- Based on Dell PowerEdge dual Xeon servers, but populated with only 1 CPU
- 384 compute nodes
- 512 MByte memory per node, 36 GB local disk per node
- 5-dimensional toroidal GigE mesh, $6 \times 8 \times 2^3$

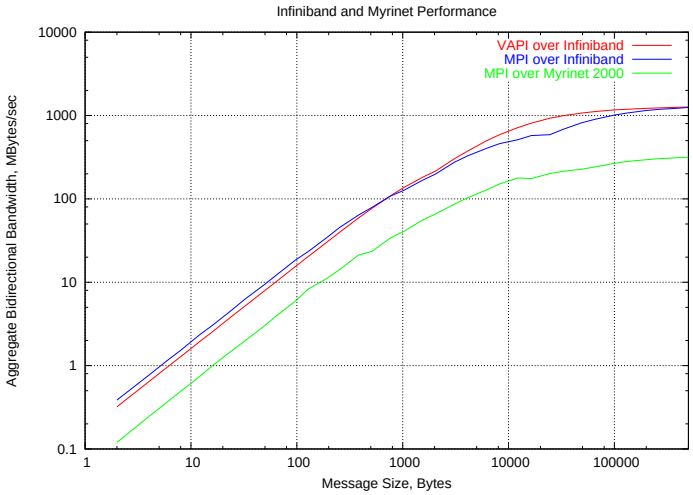

Figure 3: Bidirectional bandwidth performance of Myrinet 2000 and Infiniband networks measured with NetPIPE.

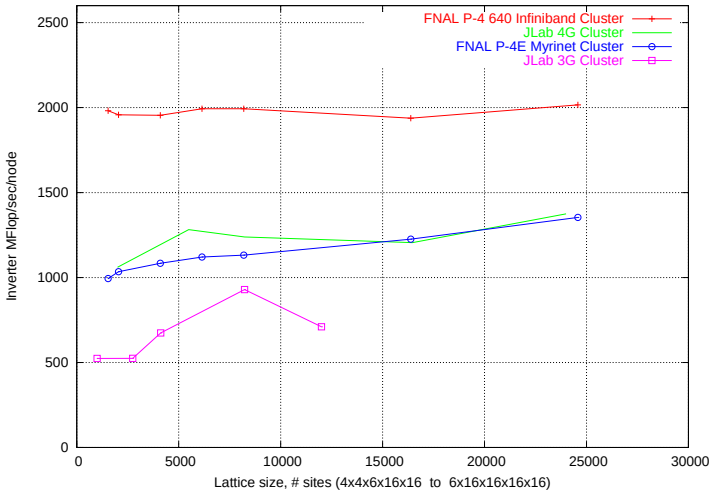

Figure 4: Performance of DWF inverter on the Jefferson Lab 3G (2.67 GHz) and 4G (2.8 GHz) clusters, and on the Fermilab 2.8 GHz Myrinet and 3.2 GHz Infiniband clusters.

This cluster is generally operated as three 128-node partitions. Only one processor is present in each compute node, as the Xeon shared memory bus architecture provides insufficient memory bandwidth for two processors on lattice QCD code. During the evaluation of systems for the purchase of this cluster, Opterons were found to have good SMP scaling, but single processor Xeon systems had lower price/performance ratios. The aggregate performance of this cluster is approximately 0.75 TFlops on DWF code, with a cost of slightly less than $1/MFlops. Fig. 4 shows DWF inverter performance as a function of local lattice size on the two most recent clusters at Jefferson Lab[3] and at Fermilab.

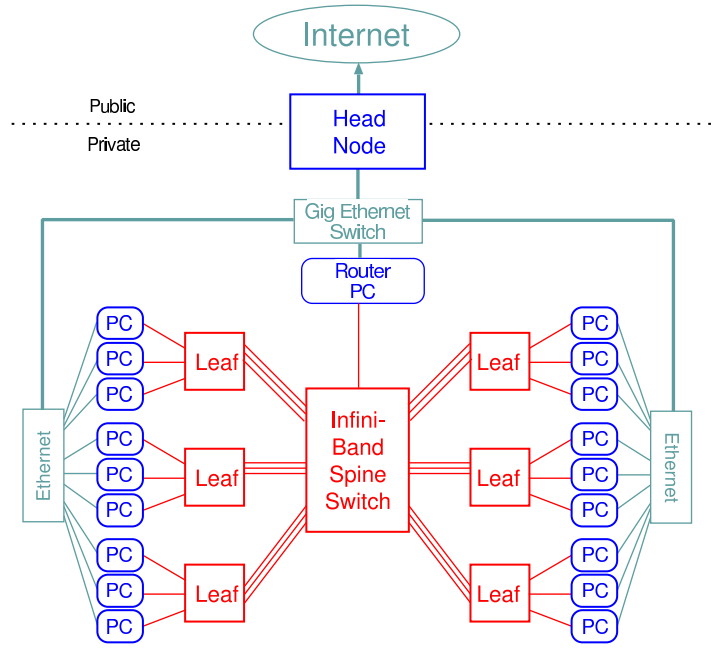

Figure 5: Fermilab Infi niband cluster schematic.

4. Fermilab

The newest SciDAC cluster at Fermilab came online in June 2005. Its specifications are:

- 3.2 GHz Intel Pentium 4 640 processors, 2 MB L2 cache, 800 MHz FSB
- 260 compute nodes, to grow to 520 nodes in September 2005
- 1 GByte memory per node, 40 GB local disk per node
- 4X Infiniband network using PCI-Express host channel adapters, 2:1 over-subscription

Fig. 5 shows the architecture of the cluster. A cascaded Infiniband network is used, with a 144-port "spine" switch connected to 24-port "leaf" switches. The initial wiring configuration used eight uplinks per 24-port switch, with the remaining 16-ports connected to compute nodes. When the cluster is expanded in the coming months, this 2:1 oversubscription will increase to 4:1.

Fig. 6 shows the weak scaling performance of the Fermilab Infiniband cluster on MILC "asqtad" code. In production, this cluster sustains over 1.3 GFlops per node while generating $40^3 \times 96$ gauge configurations using 256 processors. DWF performance (see Fig. 4) is approximately 2.0 GFlops per node. The total cost per node including Infiniband was \$1900, or \$1.45/MFlops for "asqtad" and \$0.95/MFlops for DWF. This cluster will be expanded to 520 total systems during the next three months. The total cost per node for this expansion including Infiniband is \$1550, with the reduction due to price decreases for the processors and Infiniband. This price corresponds to \$1.18/MFlops for "asqtad" and \$0.78/MFlops for DWF.

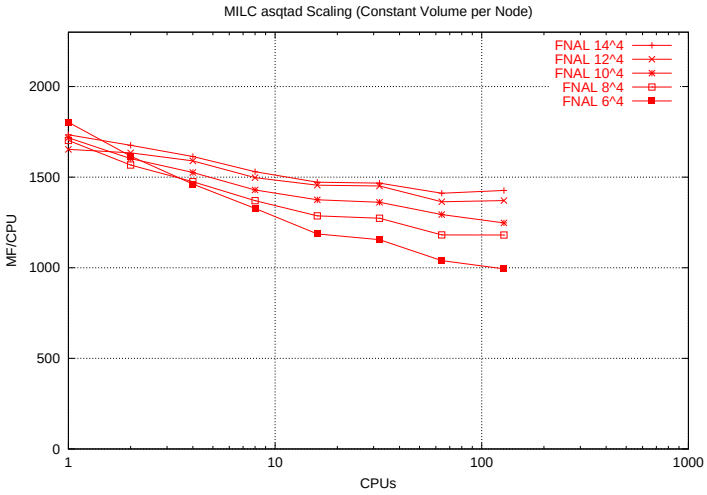

Figure 6: MILC "asqtad" inverter performance on the Fermilab 3.2 GHz Infi niband cluster as a function of the number of processors, using constant lattice volume per processor.

5. USQCD

The Department of Energy's Lattice QCD Computing Project (USQCD) will begin in October 2005. This four year project will have a total funding of approximately $9.2 million, beginning with $2.5 million in fiscal year 2006. USQCD will build or purchase a series of computing systems for lattice QCD, and it will operate these systems as well as the U.S. QCDOC and the SciDAC clusters.

USQCD enters its first project year with a total sustained computing capacity on lattice QCD codes of approximately 6 TFlops. In the first year, Infiniband clusters sustaining an aggregate of approximately 2 TFlops will be constructed at Fermilab and at Jefferson Lab. In each of the subsequent years, the most cost effective computing solution will be purchased. By the end of the project in 2009, an estimated total sustained capability of 17 TFlops will be in operation.

The USQCD facilities will all support the SciDAC programming environment[4]. A uniform user environment will be enforced on all of the systems, so that conforming Makefiles and scripts will run without modification at all sites. As the International Lattice Data Grid (ILDG) project proceeds, the USQCD facilities will support the file formats, metadata specifications, and grid middleware software necessary to be able to easily exchange data with other sites and collaborations worldwide.

References

[1] `http://www.scidac.org/`

[2] W.Watson, private communication

[3] D.G.Richards, private communication

[4] `http://www.usqcd.org/usqcd-software/`