LATTICE QCD CLUSTERS AT FERMILAB

Don Holmgren, Paul Mackenzie, Jim Simone, Amitoj Singh,
Fermi National Accelerator Laboratory, Batavia, IL 60510, USA

Abstract

As part of the DOE SciDAC "National Infrastructure for Lattice Gauge Computing" and DOE "Lattice QCD Computing" projects, Fermilab builds and operates production clusters for lattice QCD simulations for the US community. We currently operate two clusters: a 128-node Pentium 4E Myrinet cluster, and a 520-node Pentium 640 Infiniband cluster. We discuss the performance of these systems. We also discuss the 1000-processor Infiniband cluster planned for summer of 2006.

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[2].

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. Details of the SciDAC systems built prior to CHEP06 and discussions of their performance have been described elsewhere[3][4]. This paper discusses the Fermilab production clusters.

PROCESSOR AND NETWORK PERFORMANCE

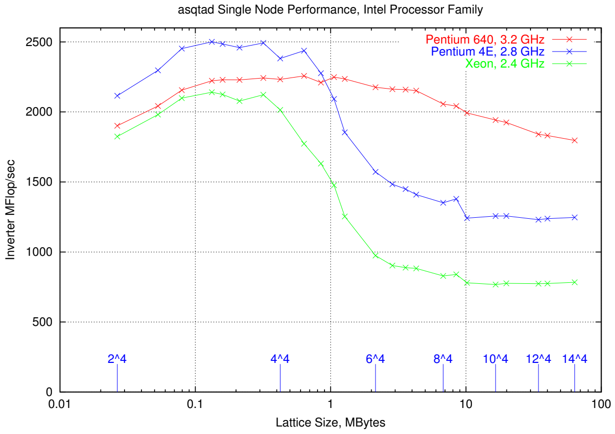

Figure 1: Single node MILC "asqtad" (improved staggered action) inverter performance as a function of lattice size.

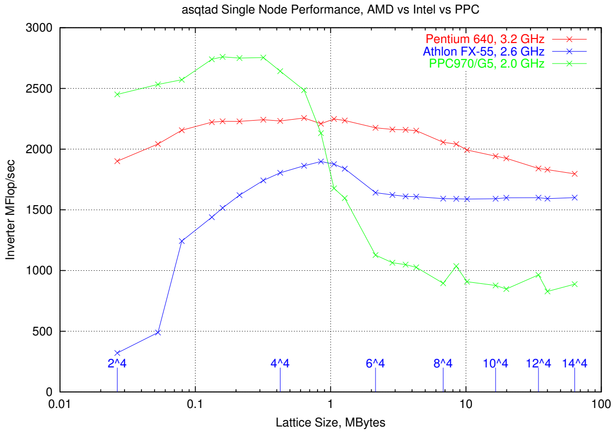

Figure 2: Single node MILC "asqtad" (improved staggered action) inverter performance as a function of lattice size.

In the past 18 months, 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 ("Extreme Edition") 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 a 2.6 GHz AMD Athlon FX-55 processor and a 2.0 GHz G5 processor. Note that 2.5 GHz G5 processors are now available, as well as faster Athlon FX processors.

In the last 10 months, both Intel and AMD have introduced dual-core versions of their processors. Recent benchmarking on these processors is very encouraging, showing strong scaling. Fig. 3 shows the relative performance of systems built with two AMD 280 processors (2.4

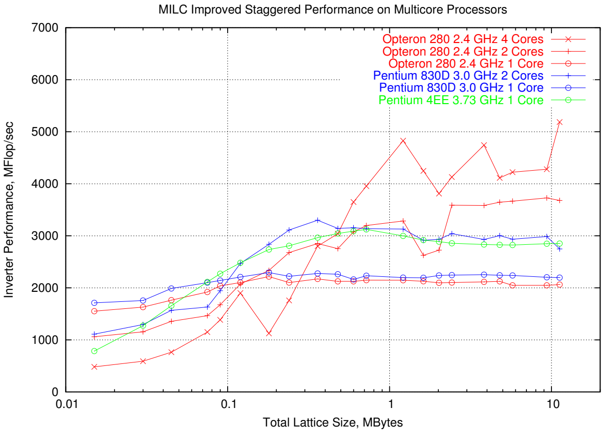

Figure 3: Aggregate MILC "asqtad" (improved staggered action) inverter performance as a function of lattice size and number of cores used during MPI runs. The strong variance in the Opteron runs is due to lack of application control over physical memory allocation on these NUMA systems.

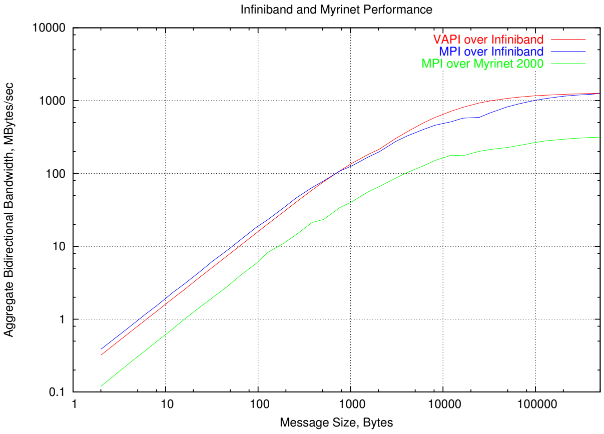

Figure 4: Bidirectional bandwidth performance of Myrinet 2000 and Infiniband networks measured with NetPIPE.

GHz), a single Pentium 830D processor (3.0 GHz), and a Pentium 4 Extreme Edition processor (3.73 GHz). The same binary was used for each test. On the dual-core systems, MPI was used to launch multiple cooperative processes on the indicated number of cores.

Network fabrics, like processors, have continued to increase in performance and decrease in cost. Fig. 4 shows the bidirectional bandwidth as a function of message size on single-data-rate (SDR) 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.

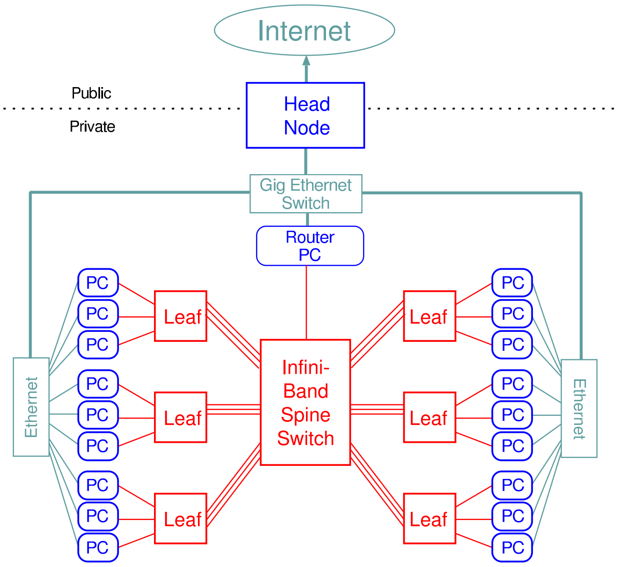

Figure 5: Fermilab Infiniband cluster schematic.

THE FERMILAB CLUSTERS

The newest SciDAC cluster at Fermilab, "Pion", came online in June 2005 with 260 nodes, followed by an expansion to 520 nodes in November 2005. Its specifications are:

- 3.2 GHz Intel Pentium 4 640 processors, 2 MB L2 cache, 800 MHz FSB
- 520 compute nodes, Intel SE7221BK1 motherboard
- 1 GByte memory per node, 40 GB local disk per node
- 4X Infiniband network (SDR) using PCI-Express host channel adapters, 4:1 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 was expanded to 520 nodes, this 2:1 oversubscription was increased to 4:1 with a negligible drop in application performance. Communications between the head node and the worker nodes can either use a fast ethernet network, or IP over Infiniband for higher performance via the "Router PC" shown in the drawing.

The other Fermilab production cluster, "QCD", came online in June 2004. Its specifications are:

- 2.8 GHz Intel Pentium 4E processors, 1 MB L2 cache, 800 MHz FSB
- 128 compute nodes, Intel SE7210TP1 motherboard
- 1 GByte memory per node, 20 GB local disk per node

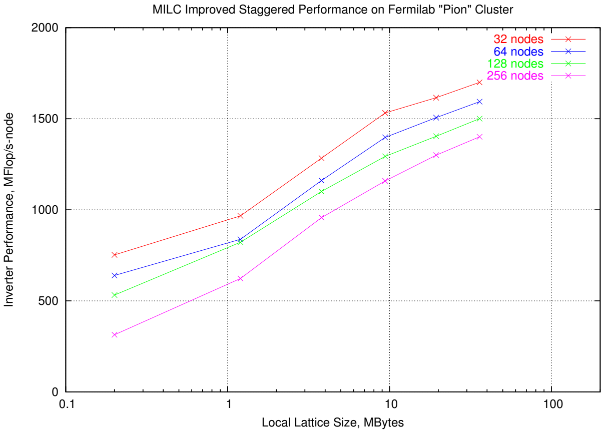

Figure 6: MILC "asqtad" inverter performance on "Pion".

- Myrinet 2000 network

The Fermilab "QCD" and "Pion" clusters share a common head node. For the batch system, we use the TORQUE resource manager combined with the Maui scheduler. dCache is used both as a local parallel file system, and as an interface to the Fermilab Enstore mass storage facility. Myricom GM version 2.1.23 is used on the Myrinet fabric. OpenIB version 1.8.1 stack is used on the Infiniband fabric. Three MPI versions are available: MPICH_GM 1.2.6..14b for "QCD", MVAPICH 0.9.6 for "Pion", and MPICH-VMI 1.2.6 over VMI 2.1 for both clusters.

Figures 6 and 7 show the weak scaling performance of the Fermilab "Pion" and "QCD" clusters on MILC "asqtad" code. In production, the "Pion" cluster sustains 1.34 GFlops per node while generating $40^3 \times 96$ gauge configurations using 256 processors. The total cost per node of the "Pion" cluster including Infinband was $1661 ($1838 per node for the first half, $1484 per node for the second half five months later because of price reductions on the Infiniband equipment and on the Intel processors), or $1.24/MFlop/s on "asqtad" (improved staggered action) simulations.

THE NEXT CLUSTER

Funded by the Department of Energy's Lattice QCD Computing Project, a four year project begun in October 2005, Fermilab will build an 800 processor cluster in summer 2006, with an additional 200 processors funded by the SciDAC project and supplemental grants. This cluster will likely use the forthcoming Intel dual-core, dual-socket systems which support the fully buffered DIMM memory architecture, with an Infiniband network fabric. Alternate designs include dual-core, dual-socket Opteron systems with an InfiniPath network fabric, and dual-core single-socket systems based on the Intel Pentium 9xx processor family. This new cluster will be released to production by the end of September 2006.

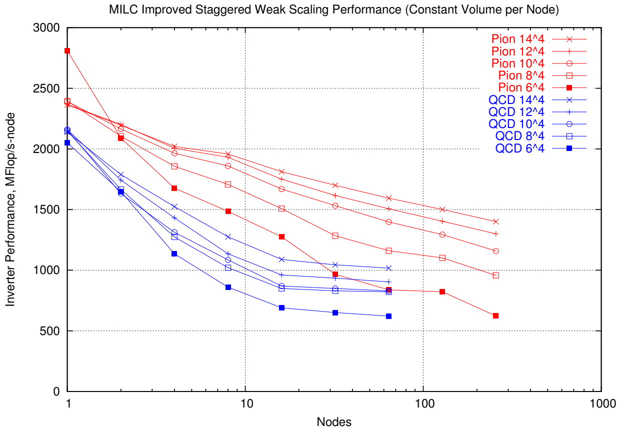

Figure 7: MILC "asqtad" inverter performance on the Fermilab 3.2 GHz Infiniband ("Pion") and 2.8 GHz Myrinet ("QCD") clusters as a function of the number of processors, using constant lattice volume per processor.

REFERENCES

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

[2] http://www.usqcd.org/usqcd-software/

[3] D. Holmgren, *PoS (LAT2005) 105*

[4] D. Holmgren et al., *CHEP03 TUIT004*