Improved LS-DYNA Performance on Sun Servers

Youn-Seo Roh, Ph.D. And Henry H. Fong
Sun Microsystems, Inc.

Abstract

Current Sun platforms which are very competitive in price/performance include Linux servers using either the Intel Xeon or AMD Opteron processors. Benchmark results using the industry-standard Neon model are presented. Performance and scalability up to 32 CPU's are discussed, as well as a comparison of use of gigabit Ethernet (GBE) interconnect versus Myrinet in a Linux Xeon cluster. Current status of Solaris x86 porting of LS-DYNA is also presented.

Introduction

In line with Sun Microsystems' recent announcements of support of x86-based hardware products, the company recently announced server products based on Intel's Xeon processors. In addition, in a more recent alliance with AMD, servers based on AMD's Opteron CPU's are also being shipped. In this report, the scalability results of LS-DYNA-MPP on the cluster of these servers – currently dual CPU models – are presented. Also the effects of interconnects – Gigabit Ethernet and Myrinet – to the scalability are compared.

Sun Fire V60/65x Servers

The Sun Fire V65x server[1] is a data center-class, x86-based, dual-processor, 2U-rackmount server designed for high availability and expandability. The server is an entry-level rack-optimized (or cluster-oriented) server designed to run standard Linux distributions or Solaris 9 Operating System x86 Platform Edition. In addition to dual Xeon processors and up to 12GB of memory, the V65x server supports up to six Ultra320 SCSI disks and six PCI-X expansion slots.

The target markets of the server are among many:

- High performance technical computing (HPTC) and grid computing.
- Web infrastructure.
- Small workgroup server and clustered database.
- Software development including custom development.

Being used for most computationally intensive applications, ls-dyna is one of the most representative HPTC workload. And a rack-mounted cluster[2] of V65x servers interconnected with high bandwidth, low-latency interconnects such as Myrinet, and cluster management software such as Sun Control Station and Sun Grid Engine, becomes a very appropriate computational platform for LS-DYNA.

Sun Fire V65X Cluster Benchmark

A scalability benchmark of ls-dyna Linux binary running Dodge Neon data from the public domain “topcrunch” benchmark site [3] was performed on a V65x cluster with both

Myrinet interconnects and Gigabit Ethernet(GBE).

Hardware Setup

- 16-node Sun Fire V65x servers
- Dual Intel Xeon 3.06GHz processors
- Myrinet interconnect with PCI-X cards
- Optional interconnect with Gigabit Ethernet

Software Setup

- Myrinet binary: ls-dyna-mpp version 970 release 3858 object files are linked with Myrinet GM library version 2.0.6 that utilizes the latest PCI-X capability. For the linking, Intel compiler version 7.0 was used, the same compiler used for compiling the object files. These ls-dyna object files make use of Intel's SSE2 (Streaming SIMD Extension Version 2) library.
- For MPI and cluster environment, MPICH [4] was selected because Myrinet GM library only supports MPICH during the time of the benchmark test. Later, by the time of writing this paper, LAM-MPI version 7 and later also came to support Myrinet GM.

Executable	mpp970.3858, Linux 2.4 SSE
Operating System	RedHat Linux2.4.18-24.7.xbigmem #1
Clustering	MPICH 1.2.5..10
Other Library	Myrinet GM library v.2.0.6

Scalability Benchmark Results

MPICH command line used to run the scalability benchmark was:
$ mpirun -np <NP> mpp970 i=neon.refined.k p=pfile

where the pfile is the parallel options. Below is the table that compares the scalabilities of Myrinet and GBE clusters. Last column is the ratio of elapsed times between the two clusters.

NCPU	Scalability Myrinet	Scalability GBE	Elapsed Time Ratio GBE / Myrinet
1	1	1	1.03
2	1.97	1.95	1.03
4	3.73	3.62	1.05
8	6.67	6.33	1.08
12	9.07	8.1	**1.15**
16	11.18	9.48	**1.21**
32	15.48	12.28	**1.29**

The table shows the effect of the low latency and high-bandwidth of Myrinet as compared to GBE starts to show up noticeably at processor counts of greater than eight, or node

count of four. At 32 processors or 16 nodes, the difference becomes even up to 30%. This means Myrinet is appropriate for performance-critical jobs utilizing large number of nodes.

On a different perspective, for jobs targeted for cluster throughput running at relatively low processor counts like up to eight, GBE interconnect presents a cost-effective means for clustering the nodes especially when the price/performance is considered for the system.

It should be noted that the scalability comparison is limited for the given dataset of Neon model, which is a frontal crash simulation. The results should depend on the dataset, and other datasets might very well show different scalability behaviors.

Sun Fire V20z Servers

The dual-processor Sun Fire V20z server [5], first in a new line of AMD Opteron-based servers from Sun, delivers high performance, reliability, and scalability in a low-cost, ultra-dense, rack-optimized 1U form factor. With a single architecture, the Sun Fire V20z server supports both 32-bit and 64-bit computing, allowing the users to maintain existing x86 infrastructure while still enabling a smooth migration to next-generation 64-bit operating systems and applications.

The Sun Fire V20z server features an integrated memory controller, up to 16 GB of memory, up to two Ultra320 SCSI hard drives, dual on-board Gigabit Ethernet, and Lights Out Management. This set of features is ideal for a wide range of applications, including high-performance technical computing, Web or application serving, database management, and grid computing.

Sun Fire V20z Cluster Benchmark

A scalability benchmark similar to that on Sun Fire V65x was performed of Neon dataset on a Myrinet cluster of 16 node V20z's.

Hardware setup

- 16-node Sun Fire V20z servers
- Dual AMD Opteron 246 processors (2.0GHz) with 4GB of memory
- Myrinet interconnect with PCI-X cards

Software setup

- Executable: It is 64bit Opteron-optimized binary, compiled with PGI compiler[6], allowing full benefit of the latest version of Myricom's proprietary GM library. It uses LAM-MPI 7.0.3 [7].
- Operating System: 64-bit SuSe Linux SLES 8. The executable at hand was built on SuSe8, and it turned out it didn't run on RedHat 9 which is another O/S available on V20z.
- On the current cluster, Myrinet GM library of 2.0.9 was used.
- MPI/Cluster Environment: LAM-MPI 7.0.3. This version of lam-mpi has the capability to select and tune the communication module between gm and tcp during run-time. By using a command line option of mpirun, users can select the configurations between one that uses Myrinet and one using GBE.

- Compiler Used: At current configuration, the executable requires lam-mpi which in turn requires the compiler that was used to built the executable, which was PGI f77 compiler. Lam-mpi configuration step requires the path of the compiler of the same kind that was used for building executable.

	V20Z Cluster	V65X Cluster
Executable	Mpp970.4842, amd64 lam703	mpp970.3858, Linux 2.4 SSE
Operating System	SuSe Linux 8 64bit 2.4.21-127-smp #1 SMP	RedHat Linux 2.4.18-24.7.xbigmem #1
MPI/Clustering	LAM-MPI 7.0.3	MPICH 1.2.5..10
Other Library	Myrinet GM library v.2.0.9	Myrinet GM library v.2.0.6
Other Dependency	Compiler PGI f77	-

Scalability Benchmark Results

Below is the scalability benchmark result comparing V20z cluster and V65x cluster.

NCPU	Scalability V20z/Myrinet	Scalability V65x/Myrinet	Elapsed Time Ratio V65x/V20z
1	1	1	1.27
2	1.8	1.97	1.17
4	3.53	3.73	1.22
8	7.18	6.67	1.37
16	16.09	11.18	1.38
32	19.29	15.48	1.59

- Elapsed times of V20z cluster with Opteron 246 2.0GHz processors show superiority over those the SFV65x cluster with Xeon 3.06GHz processors over the whole range of CPU counts.
- Scalability of V20z cluster improves significantly over that of V65x cluster with cross-over point of 8-CPU's. At full 32-CPU's, the difference in scalability is 25%, which is a significant one. This trend is thought to be coming from difference in parallel option of mpp-dyna job. At the time of the writing, a further benchmark for V65x cluster for correcting this discrepancy is under preparation.

Solaris x86 Porting of LS-DYNA

Both V65x and V20z servers come out with both standard Linux (RedHat and SuSe Linux) as well as Solaris Operating Environment x86 Platform Edition [8], or simply Solaris x86. While Linux are fully supported on these servers, by having Solaris x86 installed, users will benefit from well-established stability and technical advantages of Solaris. In order to support

LS-DYNA on the Solaris x86 operating environment, a porting project from Linux to Solaris has been initiated.

For this porting project, Sun Studio 9 compiler collection [9] which is currently in early access stage is being used. Initial porting attempt produced successful build of ls-dyna-mpp binary that uses lam-mpi 7.0.3, successfully completing a number of testcases. Full scale quality assurance runs will be in place, but overall impression of the porting was rather straightforward.

Current version of Solaris x86 9 only supports 32-bit environment, but the next update which is expected first half of 2004 will be capable of 64-bit environment. This will allow the generation of executable that is runnable on 64-bit V20z servers.

Conclusions

Sun's new server products based on AMD and Intel's x86 processors provide for adequate platform for muti-processor MPI jobs of ls-dyna. Ls-dyna-mpp executables runs with excellent scalabilities up to 32 processors when clustered together with high-bandwidth, low-latency interconnects. Porting process of ls-dyna to Solaris x86 operating environment is straightforward, and will offer extended user choices for ls-dyna computing platforms together with the stability of Solaris.

Acknowledgments

Authors would like to thank Dr. Jason Wang of LSTC for his support in providing MPP executables and other technical expertise. They also acknowledge Eduardo Pavon, Larry Olson, Qinghuai Gao and Steve Nash of Sun Microsystems for their support on the clusters that were used in these benchmarks.

References

[1] Sun Fire V65x Server Overview web page:
http://www.sun.com/servers/entry/v65x/
[2] Compute Grid Rack System web page:
http://www.sun.com/servers/computegrid/
[3] Top Crunch website:
http://www.topcrunch.org/
[4] MPICH – A Portable Implementation of MPI web page:
http://www-unix.mcs.anl.gov/mpi/mpich/
[5] Sun Fire V20z Server Overview web page:
http://www.sun.com/servers/entry/v20z/
[6] PGI Workstation Compiler web page:
http://www.pgroup.com/products/workindex.htm
[7] LAM/MPI Parallel Computing web page:
http://www.lam-mpi.org/
[8] Solaris Operating System for x86 Platforms web page:
http://wwws.sun.com/software/solaris/x86/
[9] Sun Studio 9 Early Access web page:
http://developers.sun.com/prodtech/cc/ea/ss9/

- AMD and AMD Opteron are trademarks or registered trademarks of Advanced Micro Devices, Inc. in the United States and other countries.
- Myrinet and Myricom is a registered trademark of Myricom, Inc.
- PGI is a registered trademark of STMicroelectronics Group.
- SuSe is a registered trademark of SUSE AG.
- Linux is a registered trademark of Linus Torvalds.