Cost-Performance Evaluation of SMP Clusters*

Darshan Thaker, Vipin Chaudhary, Guy Edjlali, and Sumit Roy
Parallel and Distributed Computing Laboratory
Wayne State University
Department of Electrical and Computer Engineering
Detroit, Michigan 48202

Abstract ***Clusters of Personal Computers have been proposed as potential replacements for expensive compute servers. One limitation in the overall performance is the interconnection network. A possible solution is to use multiple processors on each node of the PC cluster. Parallel programs can then use the fast shared memory to exchange data within a node, and access the interconnection network to communicate across multiple nodes. The performance of such a system is likely to be influenced by the runtime support system, whether one uses a standard message passing library or a distributed shared memory system to simulate a large shared memory machine.***

This paper presents a cost-performance evaluation of a cluster of eight dual processor PCs and a ten processor SMP workstation using two benchmark suites. We evaluate the effect of using gigabit ethernet against fast ethernet, single processor PCs versus SMP nodes, and compare the overall performance of the PC cluster to the SMP workstation. It is found that the SMP machine provides the best performance in only 50 % of the programs examined. The PC cluster with a Gigabit network is better in most of the other cases, but the improvement over FastEthernet is marginal.

Keywords: Cluster Computing, Shared Memory Multiprocessors, Performance Analysis, Benchmarks, Message Passing, Distributed Shared Memory

1 Introduction

In recent years it has been found that single processor based computers are not able to solve increasingly large and complex scientific and engineering problems since they are rapidly reaching the limits of possible physical performance. Multiple cooperating processors have instead been used to study *Grand Challenge Problems* like Fuel combustion, Ocean modeling, Image understanding, Rational drug design, etc. From a computer architecture point of view, such machines could use tightly coupled processors that access the same shared memory, known as symmetrical multiprocessors (SMPs), or one could cluster multiple computing nodes with a high performance interconnection network. The latter approach has been extended by combining multiple SMPs to provide a scalable computing environment. A suitable runtime system should allow parallel programs to exploit the fast shared memory when exchanging data within the node, while accessing the slower network only when communication across nodes.

The designer of such SMP clusters has a variety of choices with regards to the node design and the the network technology that is deployed. In this paper we present experimental results with a cluster of eight dual-processor PCs. The machines are connected via a FastEthernet network as well as a Gigabit network. We evaluate the system using programs from two benchmark suites, the NAS Parallel benchmarks which are representative of message passing code, and the SPLASH-2 benchmarks which are written using a shared memory model.

We attempt to show how different SMPs relate to each other with reference to their cost and their performance. Ideally one would like to know the best SMP cluster configuration for a given problem. We have also attempted to answer the

**This work was supported in part by NSF MIP-9309489, NSF EIA-9729828, US Army Contract DAEA-32-93-D-004 and Ford Motor Company Grants 96-136R and 96-628R.*

quintessential question - *For a fixed amount of money, what kind of machines should I buy that satisfactorily meets my performance needs ?* To answer the above question we have conducted some experiments using SMPs. In addition we also try to answer the following questions:

- *What is the cost benefit of using Gigabit over Fast Ethernet ?*
- *How well do the programs scale on a cluster configuration ?*
- *What is the benefit of using SMP nodes within a cluster ?*

The rest of this paper describes the system details, the experimental setup, the results, and our conclusions.

2 System Cost Details

The experiments were carried out on a cluster of eight dual-processor PCs and a ten processor SMP Workstation. The cost analysis reflects prices that were in effect in November 1998.

2.1 Cluster of PCs

Each processor was a 350 MHz Pentium II with 512 Kb of cache, and each node had 256 Mb memory. The machines were connected by a Gigabit Ethernet switch - 3Com SuperStack II Switch 9300 and by a Fast Ethernet switch - NetGear Fast Ethernet Switch Model FS516. Each node of the PC cluster used the DEC 21040 "tulip" based Fast Ethernet cards and the GNIC-II gigabit cards from Packet Engines. The operating system for the PC Cluster was SuSE Linux 6.0, with kernel 2.1.132. The system cost is comprised of the following components:

One Fast Ethernet Switch	$ 800
Base cost per node (Single CPU)	$ 1,500
Dual CPU Motherboard + CPU	$ 400
Gigabit Card + cable	$ 800
One Gigabit Switch	$ 9,000

An eight node, eight CPU cluster costs $ 12,800 and upgrading to dual CPUs raises its price to $ 16,000. The addition of the Gigabit network accounts for another $ 15,400, for a total cost of $ 31,400.

2.2 SMP Workstation

The SMP workstation used for the evaluation was a SUN UltraEnterprise E4000, with ten 250 MHz UltraSparcII processors that have 4 MB external cache each. The system has a unified 256-bit wide Gigaplane system bus with a sustained transfer rate of 2.6 GB/s. The processor/memory boards are integrated with the I/O boards through a high performance Ultra Port Architecture (UPA) crossbar switch. The UPA is fully synchronous internally and runs at the 84 MHz speed of the Gigaplane, thus yielding a sustainable bandwidth of about 1.25 GB/s. The operating system for the SUN machines was Solaris 2.6.

The current UltraEnterprise midrange servers can support upto eight processors in the E3500 configuration [4]. For more processors, one has to use a SUN UltraEnterprise E4500[1] (*this is the SMP that is available from SUN as of the present which the closest to the E4000*). Thus, the system cost is calculated as:

E3500 base cost	$ 22,680
Cost of processors ($>$ 2)	$ 6,300
Upgrade to E4500	$ 39,400

The above costs include the various network cards, memory, and other essentials excluding items like monitors, keyboard etc. The total cost of E3500 with four processors is $ 35,280. The eight processor E3500 costs $ 60480. The ten processor E4500 costs $ 112,480.

3 Experimental Setup

The test programs used in the experimental evaluation consist of programs from the NAS parallel benchmark suite [3] and the SPLASH-2 benchmark suite [2].

[1]We use the price of E3500 for upto eight processors

3.1 NAS Parallel Benchmarks

The NAS Parallel Benchmark (NPB-2.3) consists of a group of programs that are commonly found in Numerical Aerodynamic Simulation codes. We evaluated the performance of five of the NAS benchmarks for two problem sizes, Class W and Class A. The programs that are shown in this paper include:

- BT, this program uses an approach based on block-tridiagonal matrices to solve the Navier-Stokes equation.
- CG, this program uses the conjugate gradient method to solve the Navier-Stokes equation.
- EP, Embarassingly Parallel.
- LU, this program uses a block-lower-triangular block upper-triangular approximate factorization to solve the Navier-Stokes equation.
- SP, this program uses a pentadiagonal matrix-based algorithm for the Navier-Stokes equation.

These are Fortran programs with calls to the Message Passing Interface (MPI) library [5]. The MPI version of these programs were run on the PC cluster and the SUN UltraEnterprise E4000. The MPI implementation we used was the freely available *mpich* from the Argonne National Laboratories. The *ch_p4* device was used for both machines. In case of the dual processor runs, shared memory is used to communicate between processes on the PC. It was found that this method did not work on Solaris, hence only socket level communication was used in this case.

In addition, the SUN Workshop parallelizing Fortran compiler was used to convert the serial version of the NAS Benchmarks for execution on the UltraEnterprise. Note that we did not use the optimizing flags for the compiler.

3.2 SPLASH-2 Benchmarks

The SPLASH-2 benchmark suite [2] has been written for evaluating performance of shared address-space multiprocessors and consists of application kernels as well as full fledged code. The programs used in this evaluation include:

1. LU-c, LU decomposition kernel with contiguous blocks.
2. RAYTRACE, raytracing program.
3. VOLREND, volume rendering program.
4. WATER-sp, simulates system of water molecules with spatial data structure.
5. WATER-n2, simulates system of water molecules with simpler data structure.

The SPLASH-2 suite encapsulates shared memory programming using portable macros. The macros were implemented using POSIX threads on the SUN E4000. The *Strings* [1] Distributed Shared Memory environment was used for executing these programs on the PC cluster. A Distributed Shared Memory is used to give the programmer an illusion that he is using a single large block of shared memory when he is in reality using physically distributed memory over multiple SMPs. *Strings* is implemented as a library that is linked with a shared memory parallel program. A program uses calls to the distributed shared memory allocator to create globally shared memory regions. We have not shown results for program like FFT, RADIX, and OCEAN, since these are known to give very poor performance with Distributed Shared Memory systems [6]. The Sun E4000 outperforms the cluster for these applications.

4 Results

The figures show the results that we obtained from our experiments. Using these we shall try to answer the questions that we had raised earlier. The applications we ran were compute intensive applications that would be run on a medium sized computer system.

4.1 Gigabit vs FastEthernet

We executed the NAS parallel Benchmarks and the Splash-2 suite of benchmarks on the PC cluster using both Fast Ethernet and Gigabit Ethernet. Fig-

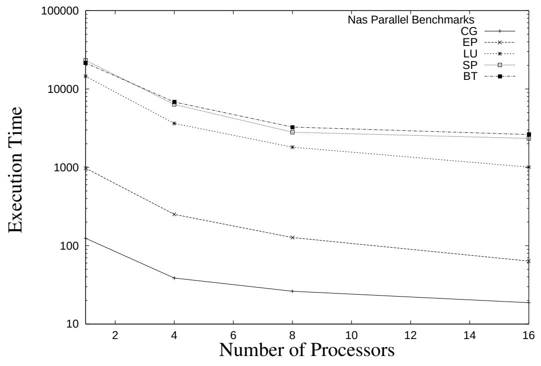

Figure 1: NPB-2.3 with Fast Ethernet

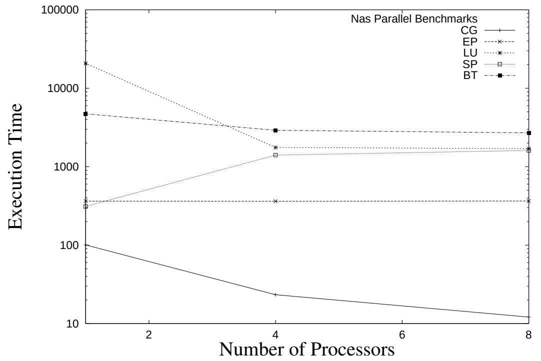

Figure 2: NPB-2.3-serial with parallelizing compiler

ure 1 shows the results of the runs of the NAS parallel benchmarks over Fast Ethernet and Figure 4 shows the results of the runs for the same benchmarks over Gigabit Ethernet. We observed that in most cases the difference in the time taken by the program is negligible when run over the two different interconnects. The reason for this is that the programs we ran did not have the need for the large bandwidth that Gigabit Ethernet provides. The communication needs of these programs were satisfied with the bandwidth offered by Fast Ethernet. The only difference is for SP, which performs much better on Gigabit Ethernet but only when run on 16 processors and this occurs immaterial of the class size of the program. This occurs due to the volume of data that SP tends to send, and the increase in the data transmitted with the increase in the number of processors.

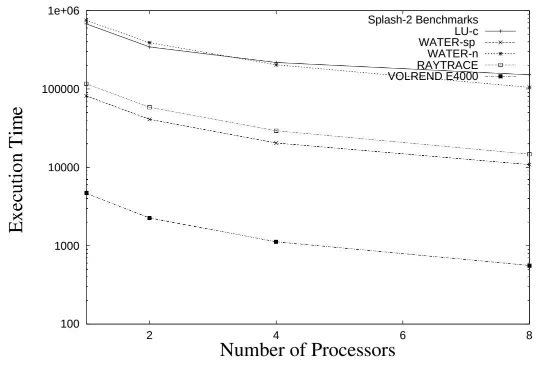

Figure 3: Splash-2 Benchmarks on the SUN E4000

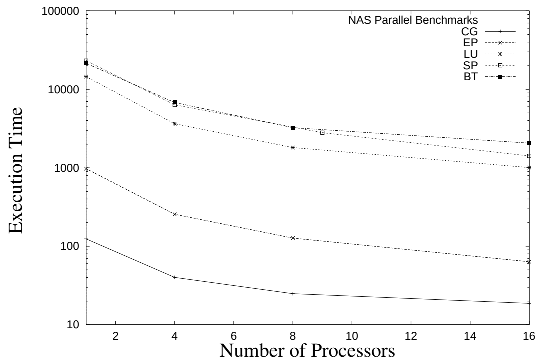

Figure 4: NPB-2.3 with Gigabit Ethernet

Likewise, for the Splash2 benchmarks, it was found that the Gigabit gives very little performance improvement over the Fast Ethernet, as seen in Figure 8 and Figure 6. This leads us to believe that with the nature of applications that we experimented with, the cost of the Gigabit Ethernet is unjustified in terms of its performance.

4.2 Scalability

With regard to scalability we found that the NAS Parallel Benchmarks tend to scale very well up to sixteen processors while the benchmarks of the SPLASH-2 suite undergo a degradation in performance when run over sixteen processors on the PC cluster (it scales well for upto eight processors on the cluster as well as the SMP). From Figure 1 the scalability of the NAS Parallel Benchmarks is apparent. From Figure 6, we can see that the

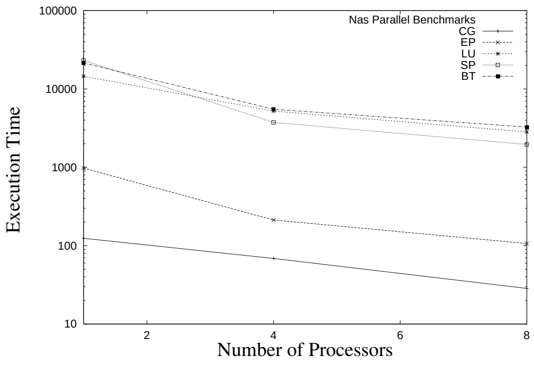

Figure 5: NPB-2.3 on the SUN E4000 with MPI

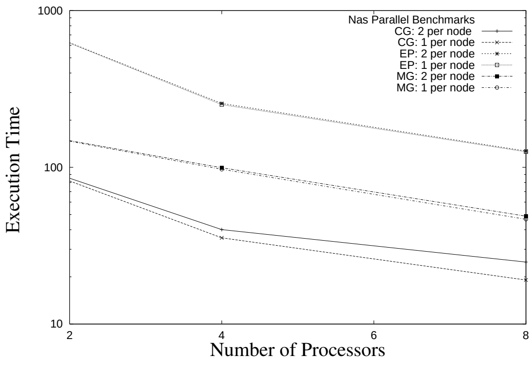

Figure 7: NPB-2.3 Benefit of SMP Nodes

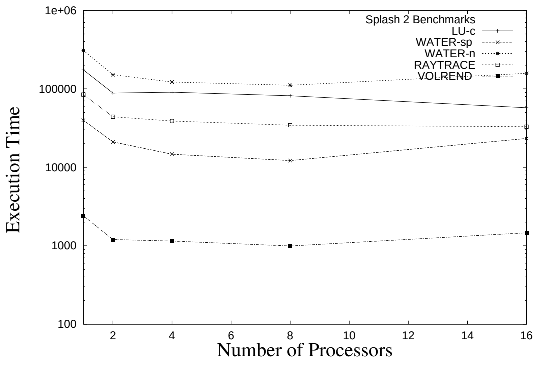

Figure 6: Splash-2 with Gigabit Ethernet

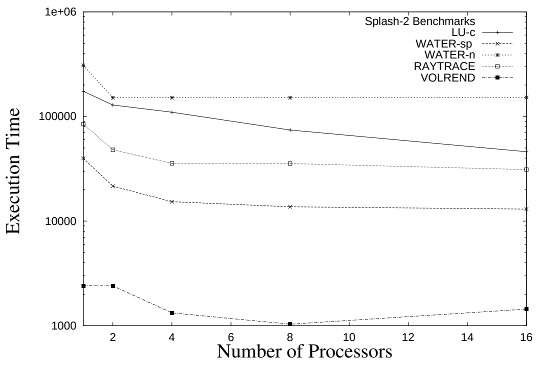

Figure 8: Splash-2 with Fast Ethernet

Splash-2 benchmarks do not scale very well and in some cases tend to give worse performance with the increase in the number of nodes. This is because of the increase in communication with the increase in the number of nodes. These results are also dependent on the inherent nature of the Distributed Shared Memory system. For certain programs from the Splash-2 suite like FFT, Radix, Ocean -c etc, their inherent nature makes them perform badly on a DSM and therefore they give bad performance and scale very badly over the PC cluster. Scalability for these programs could be improved by improvements in the Distributed Shared Memory system.

4.3 Benefit of SMP Nodes

The best performance on the PC cluster for the SPLASH-2 was obtained using four nodes with two threads on each node, as seen in Figure 9. Thus, for performance benefit it would be better to have four nodes with two processors each as compared to eight nodes with one processor each. In terms of cost the four node solution is better because it costs less since one only pays for an extra processor, whereas for eight nodes one has to pay for extra network cards, motherboards and even cables for the Gigabit.

However, from Figure 7, we can see that the NAS Parallel Benchmarks, contrary to our expectations, tend to give better performance when there is only one processor used per node. This we think is because with two processors there is an increase in cache misses, mainly capacity misses, and hence the degradation in performance.

BT	16CPU, PC-Cluster, Gigabit Ethernet	\$31,000
CG	8CPU on E4000, Parallelizing Compiler	\$60,480
EP	16CPU, PC-Cluster, Gigabit Ethernet	\$31,000
LU	16CPU, PC-Cluster, Gigabit Ethernet	\$31,000
SP	4CPU on E4000	\$35,280
MG	16CPU, PC-Cluster, Gigabit Ethernet	\$31,000

Table 1: Best Performance of the NAS Parellel Benchmarks

LU-c	16CPU, PC-Cluster, Fast Ethernet	\$17,000
Water-n	8CPU on E4000	\$60,480
Water-sp	8CPU on E4000	\$60,480
Volrend	8CPU on E4000	\$60,480
Raytrace	8CPU on E4000	\$60,480

Table 2: Best Performance of the SPLASH-2 Benchmarks

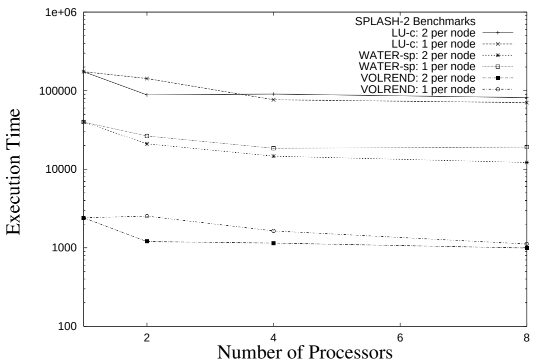

Figure 9: Splash-2 Benefit of SMP Nodes

4.4 Parallelizing Compiler

Figures 2 and 5 show the results of the NAS parallel benchmarks on the SUN E4000 with the parallelizing compiler and the MPI code. We can see that in most cases the parallelizing compiler gives marginally better results compared to the MPI code, while in some cases the MPI code performs better. However, if we take into consideration the fact that the parallelized code is automatically generated and the MPI code is hand parallelized and then from the cost-performance point of view the parallelized code is much better as it comes at a lower cost. Moreover, we have not used any optimizing flag in the parallelizing compiler.

5 Conclusion

The cost of the SUN UltraEnterprise with four processors is almost the same as the whole of the PC cluster with the Gigabit Ethernet as the interconnect. On comparing these for NAS Parallel Benchmarks, we looked at the performance of the complier parallelized code on the SUN with the MPI code on 16 processors of the PC cluster. It was found that the PC cluster performs just as well as the SUN E4000 except for BT and for SP. For the SPLASH-2 benchmarks a similar result was observed, where other than RADIX, FFT and OCEAN-c, the PC cluster (using eight processors) performed as well as the SUN UltraEnterprise E4000.

From the NAS Parallel Benchmarks, the SUN E4000 tends to perform marginally better than the PC cluster for large programs like BT and SP. However the PC cluster outperforms E4000 when it comes to smaller programs. On the PC cluster, Linux is able to exploit the internal shared memory between two processors while for the E4000, the mpich implementation was unable to optimally utilize the shared memory.

However, when we ran the NAS parallel benchmark code, that was parallelized by the SUN native complier, on the SUN E4000, the results show that this parallelized code performs much better than the MPI code for four processors. But the same code lags behind the MPI code when we increase the number of processors to eight.

From the above results it can be seen that the PC cluster performs as well as the SUN UltraEnterprise cluster and the SUN UltraEnterprise E4000 in most of the applications, but in some of the applications the E4000 is far better. The runs from the NAS parallel benchmarks demonstrate that the code parallelized by the native SUN complier is far better to the MPI version of the NAS parallel benchmarks because the parallelized code is able to make use of the shared memory available within the system.

Table 1 and 2 show the configuration which delivers the best performances on the various programs used for the study. In addition, the E4000 SMP performs best on the FFT, RADIX, and OCEAN programs.

Cost-wise it can be seen that the E4000's cost for four processors is same as that of the dual-processor PC cluster (and double that of the PC cluster with eight processors) and while their performance is almost similar for programs run over the DSM, the E4000 (if used with eight processors) can yield greater benefits for programs that can utilize its shared memory capabilities.

References

[1] S. Roy and V. Chaudhary, "*Strings:* A High-Performance Distributed Shared Memory for Symmetrical Multiprocessor Clusters," in *Proceedings of the Seventh IEEE International Symposium on High Performance Distributed Computing*, (Chicago, IL), pp. 90 –97, July 1998.

[2] S. C. Woo, M. Ohara, E. Torri, J. P. Singh, and A. Gupta, "The SPLASH-2 Programs: Characterization and Methodological Considerations," in *Proceedings of the International Symposium on Computer Architecture*, pp. 24–36, June 1995.

[3] D. Bailey, T. Harris, W. Saphir, R. van der Wijngaart, A. Woo, and M. Yarrow, "The NAS Parallel Benchmarks 2.0," Tech. Rep. NAS-95-020, NASA Ames Research Center, http://science.nas.nasa.gov/Software/NPB/, December 1995.

[4] "Sun Enterprise 3500-6500 Server Family: Architecture and Implementation White Paper," http://www.sun.com/servers/white-papers/midrange-arch.ps.

[5] "Message passing interface forum." http://www.mpi-forum.org/.

[6] D. Jiang, H. Shan, and J. P. Singh, "Application Restructuring and Performance Portability on Shared Virtual Memory and Hardware-Coherent Multiprocessors," in *Proceedings of the ACM Symposium on the Principles and Practice of Parallel Programming*, (Las Vegas), pp. 217 – 229, ACM, 1997.