Performance of an Infiniband cluster running MPI applications

José Miguel-Alonso[1], Txema Mercero[2], Edu Ogando[2]

[1]Department of Computer Architecture and Technology, UPV/EHU
[2]Scientific Computing Service, UPV/EHU

Technical Report EHU-KAT-IK-03-07

Introduction

In this paper we show the results of running a collection of well-known benchmarks on the Arina/Maiz cluster managed by the Scientific Computing Service of the University of the Basque Country. For the purpose of putting the numbers in perspective, we compare the results with those obtained from the MareNostrum supercomputer managed by the Barcelona Supercomputing Center.

The benchmarks of choice are

- Intel MPI Benchmarks [1], a collection of small programs that test the performance of the different MPI primitives, including point-to-point as well as collective operations. We focus on the performance of uni-directional transferring operations.
- NAS Parallel benchmarks [2], a collection of applications representative of the class of codes executed in supercomputing centers.

Experiments in Arina/Maiz have been run in an empty system. However, in the MareNostrum they have been executed when the machine was in production. This difference means that results are somewhat biased against the MareNostrum. Note, though, that our intention was to provide some figures useful for comparison, not to perform an exhaustive evaluation of both systems.

Results show that the Arina/Maiz cluster, although limited in size, is a powerful platform to run parallel applications.

Details of the characteristics of the Arina/Maiz cluster are given in Appendix A. Details of the MareNostrum are provided in Appendix B.

IMB Micro-benchmarks

The Intel MPI Bechmarks, version 2.3, are designed to measure the performance (in terms of latency and/or throughput) of the most representative MPI primitives running on a given platform. They are not really useful to measure actual applications, but give interesting indicators about the performance of the interconnection mechanism.

In this test we will only use the benchmark that measures uni-directional data transfers, involving just two MPI processes. As we always allocate one process to one CPU, we can use both terms without distinction. We have run the tests on different networks. In order to put numbers in context, these are the main characteristics of the interconnection networks:

- **VAPI**. Infiniband network [3], at 10 Gb/s. VAPI is the name of the libraries and API used to access this network. The MPI implementation is configured to use VAPI.
- **TCP-GE**. Gigabit Ethernet network (1 Gb/s). The MPI implementation is configured to use TCP connections on top of this network.
- **GM**. Myrinet-2000 network [4], at 2 Gb/s. GM is the name of the libraries and API used to access this network. The MPI implementation (MPICH-GM) is configured to use GM.
- **SHMEM**. Internal communication between processors in the same node. MPI messages are interchanged using shared memory structures.

We have measured throughput and delay for these combinations of node / network of the Arina/Maiz cluster:

- VAPI-I. Itanium processors in different nodes, connected via the Infiniband network.
- SHMEM-I. Itanium processors in the same node, connected via shared memory.
- VAPI-O. Opteron processors in different nodes, connected via the Infiniband network.
- SHMEM-O. Opteron processors in the same node, connected via shared memory.
- TCP-GE-O. Opteron processors in different nodes, connected via Gigabit Ethernet.

Compilers and MPI implementation are provided by HP.

For comparison purposes, we also include these configurations that do not correspond to the Arina/Maiz cluster, but to the MareNostrum:

- GM-MN. Two IBM PowerPC processors in different nodes, connected via the Myrinet-2000 network.
- SHMEM-MN. Two IBM PowerPC processors in the same node, connected via shared memory.

The MPI implementation is MPICH-GM, provided by Myricom. Compilers are of the GCC family.

Results of experiments are summarized in Fig. 1 (delay) and in Fig. 2 (throughput). We can see that:

- The performance of the MPI implementations over shared memory is very good in all cases. Delays are very small, and throughput is very high—especially for small and medium messages. So it is a good idea to run small-scale MPI applications in a single node.
- The VAPI interconnection provides very good levels of performance. The implementation for the Opteron provides less delay than that for the Itanium, but the peak throughput is slightly lower.
- Although Gigabit Ethernet can be used for MPI applications, the performance is quite low, compared to Infiniband.

- The Myrinet-2000 network used in the MareNostrum provides much better results than Gigabit Ethernet; however, it does not reach the good results of Infiniband.

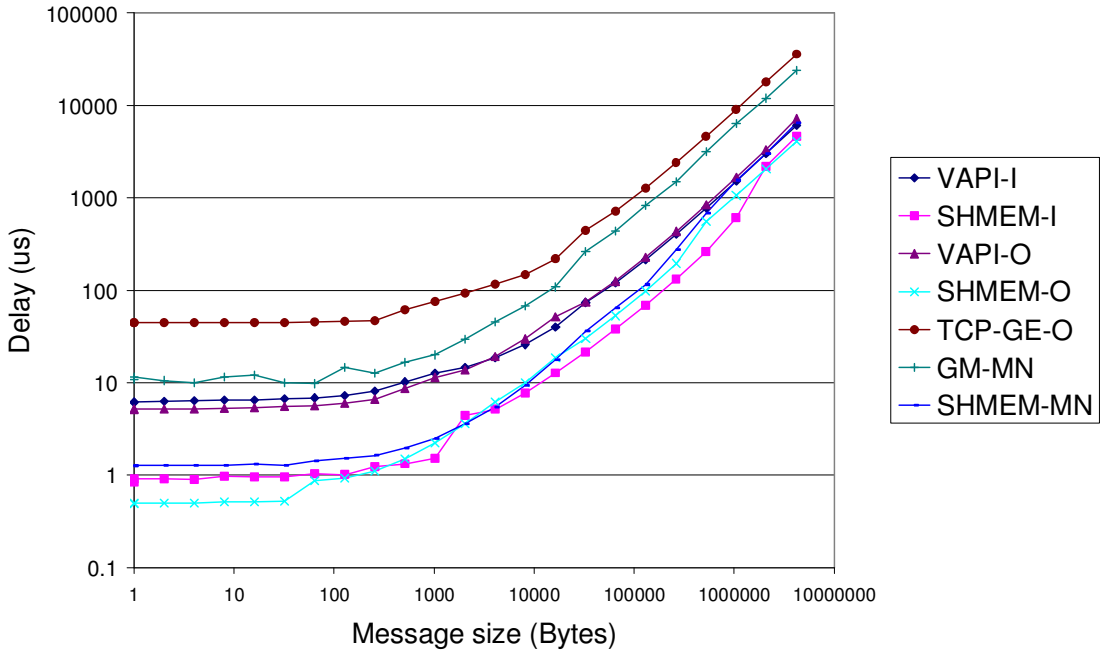

Fig. 1 Summary of experiments with the IMB. Delays for different message sizes. Logarithmic scales in both axes. The lower, the better.

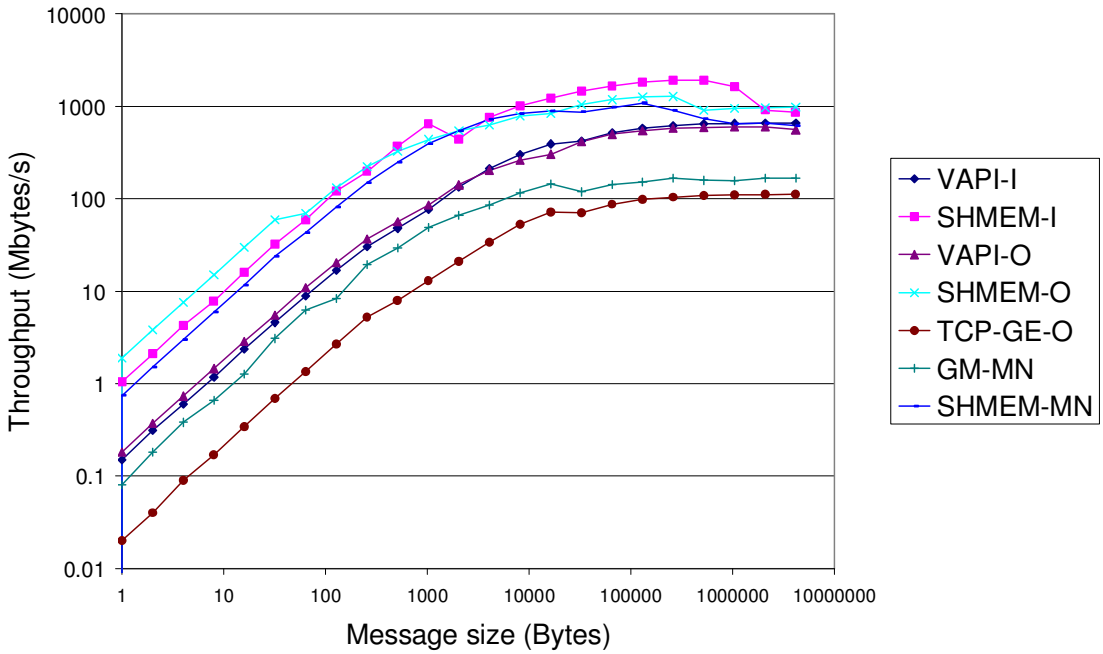

Fig. 2 Summary of experiments with the IMB. Throughput for different message sizes. Logarithmic scales in both axes. The higher, the better.

NAS Parallel Benchmarks

The NAS Parallel Benchmarks (NPB) are a small set of programs designed to help evaluate the performance of parallel supercomputers. The benchmarks, which are derived from computational fluid dynamics (CFD) applications, consist of five kernels and three pseudo-applications. The NPB come in several "flavors." In particular, the NPB-2 are MPI-based source-code implementations written and distributed by NAS. They are intended to be run with little or no tuning, and approximate the performance a typical user can expect to obtain for a portable parallel program.

NPB applications are compiled for a given number of processes, and a given "problem size". We have chosen 64 processes (and always allocate one CPU per MPI process), and class A – quite small, but more adequate to test the impact of the network on application speed.

In Arina, applications have been compiled using versions 7 and 9 of Intel compilers. We have run the experiments on 64 (16x4) Itanium2 processors, using the MPI over VAPI (Infiniband) network. Results are given as generated by the benchmarks: in Mop/s. They are summarized in Fig. 3. The same experiments have been run in 16 nodes (64 PowerPC processors) of the Mare Nostrum.

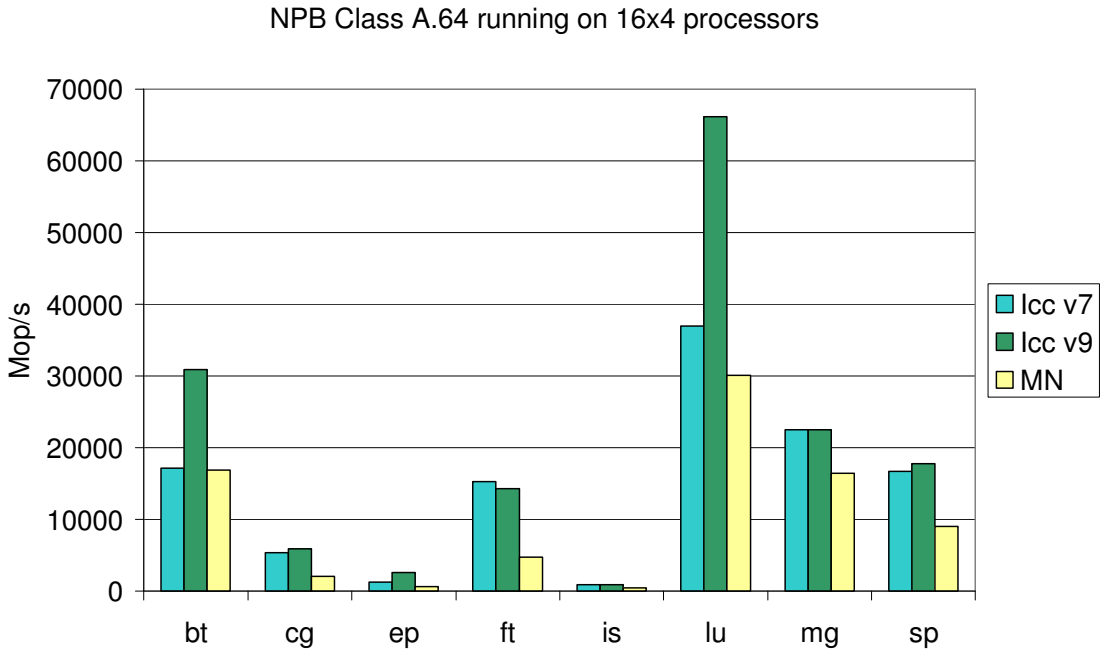

Fig. 3 Summary of experiments with the NPB

It is interesting to see the importance of the compiler in the achieved performance. With the only exception of FT, programs compiled with version 9 are faster (or much faster) than those compiled with version 7.

Conclusions

The MareNostrum and Arina/Maiz represent different classes of machines: current Arina has 88 powerful Itanium2 processors, while the MareNostrum incorporates 10240 PowerPC processors. Still, most of the technology used in both clusters is similar, and we can perform some comparison tests.

In terms of raw communication performance, the Infiniband network used in the Arina/Maiz clusters offers very good levels of performance, which can be efficiently used by MPI-based parallel applications. The MPI implementation works very well using shared memory; small applications (with up to 4 processes) work the best when allocated to the same node.

Gigabit Ethernet is an option when building cheap clusters, but its performance is not as good as that of a specific-purpose network, such as Myrinet or Infiniband. It would be of interest a comparison of the new Myrinet-10G with Infiniband, but the only implementation of Myrinet available runs at 2 Gb/s.

Regarding the experiments with the NPB, the comparison of Arina vs. MareNostrum shows how the higher performance of Arina's CPU and interconnection network results in much better results for Arina. Obviously, the potential of the MareNostrum can only be exploited by large applications.

Acknowledgements

We gratefully acknowledge the support of these institutions, that provided the machines evaluated in this work:

- Barcelona Supercomputing Center / Centro Nacional de Supercomputación (Spain)
- Scientific Computing Service, The University of the Basque Country

Work done with the support of the Spanish Ministerio de Educación y Ciencia, grant TIN2004-07440-C02-02.

References

[1] Intel® MPI Benchmarks. Available (February 16, 2007) at http://www.intel.com/cd/software/products/asmo-na/eng/219848.htm
[2] NAS Parallel Benchmarks. Available (February 16, 2007) at http://www.nas.nasa.gov/Resources/Software/npb.html
[3] Infiniband Trade Association. http://www.infinibandta.org/home
[4] Myricom, Inc. http://www.myri.com/

Appendix A: characteristics of Arina and Maiz

Information taken from http://www.ehu.es/SGI/, February 16, 2007.

Arina is a Cluster HP Integrity Server with 88 CPUs Itanium2 and Red Hat Linux AS 4 (update 4) operating system. These are some detailed characteristics of Arina:

- 1 service node for login and compilation Rx2600 Chipset HP.
- 18 compute nodes
  - 10 compute nodes, Rx4640 Chipset HP, each with 4 Itanium 2 processors running at 1.3 Ghz. 4 GB of RAM per node.
  - 4 compute nodes, Rx4640 Chipset HP, each with 4 Itanium 2 processors running at 1.3 Ghz. 16 GB of RAM per node.
  - 4 compute nodes, Rx7640 Chipset, each with 4 Itanium 2 processors running at 1.6 Ghz. 16 GB of RAM per node.
- Infiniband interconnection network (Voltaire).
- Gigabit Ethernet interconnection network.
- Peak performance (Rpeak): 496 Gflops.

Maiz is a Cluster DL585 Server with 40 CPUs Opteron and Red Hat Linux AS 4 (update 4) operating system. These are some detailed characteristics of Maiz:

- 1 service node for login and compilation, DL385.
- 4 compute nodes, DL585, each with 4 dual core Opterons running at 2.4 Ghz. 16 GB of RAM per node.
- Infiniband interconnection network (Voltaire)
- Gigabit Ethernet interconnection network
- Peak performance (Rpeak): 192 Gflops.

Appendix B: characteristics of the Mare Nostrum

Information taken from http://www.bsc.es, February 16, 2007.

In March 2004 the Spanish government and IBM signed an agreement to build one of the fastest computers in Europe. In July 2006 its capacity has been increased due to be the large demand of scientific projects.

MareNostrum is a supercomputer based on processors PowerPC, the architecture BladeCenter, a Linux system and a Myrinet interconnection.

See below a summary of the system:

- Peak Performance of 94,21 Teraflops
- 10240 IBM Power PC 970MP processors at 2.3 GHz (2560 JS21 blades)
- 20 TB of main memory
- 280 + 90 TB of disk storage
- Interconnection networks:
  - Myrinet-2000 and Gigabit Ethernet
- Linux: SuSe Distribution

MareNostrum has 44 racks and takes up a space of 120m^2.