FORSCHUNGSZENTRUM JÜLICH GmbH
Zentralinstitut für Angewandte Mathematik
D-52425 Jülich, Tel. (02461) 61-6402

Interner Bericht

Performance Issues of Distributed MPI Applications in a German Gigabit Testbed

Thomas Eickermann, Helmut Grund, Jörg Henrichs***

FZJ-ZAM-IB-9912

September 1999

(letzte Änderung: 13.09.99)

Preprint: In J. Dongarra, E. Luque, and T. Margalef (eds.),
Recent Advances in Parallel Virtual Machine and Message Passing Interface,
Proc. of the 6th European PVM/MPI Users' Group Meeting,
Barcelona, September 1999

(*) Institut für Algorithmen und Wissenschaftliches Rechnen, GMD
Schloß Birlinghoven, D–53754 Sankt Augustin, Germany

(**) Pallas GmbH, Hermülheimer Str. 10, D–50321 Brühl, Germany

Performance Issues of Distributed MPI Applications in a German Gigabit Testbed

T. Eickermann[1], H. Grund[2], and J. Henrichs[3]

[1] Zentralinstitut für Angewandte Mathematik, Forschungszentrum Jülich, D–52425 Jülich, Germany
[2] Institut für Algorithmen und Wissenschaftliches Rechnen, GMD – Forschungszentrum Informationstechnik, Schloß Birlinghoven, D–53754 Sankt Augustin, Germany
[3] Pallas GmbH, Hermülheimer Str. 10, D–50321 Brühl, Germany

Abstract. The Gigabit Testbed West is a testbed for the planned upgrade of the German Scientific Network B-WiN. It is based on a 2.4 Gigabit/second ATM connection between the Research Centre Jülich and the GMD – National Research Center for Information Technology in Sankt Augustin. This contribution reports on those activities in the testbed that are related to metacomputing. It starts with a discussion of the IP connectivity of the supercomputers in the testbed. The achieved performance is compared with MetaMPI, an MPI library that is tuned for the use in metacomputing environments with high–speed networks. Applications using this library are briefly described.

1 Introduction

Traditionally, massively parallel (MPP) and vector–supercomputers are used as stand–alone machines. The availability of high–speed wide–area networks makes it feasible to distribute applications with high demands of CPU and/or memory over several such machines. This approach is followed by projects all over the world [1–3]. However, to make this 'metacomputing' approach successful, a couple of problems have to be solved. The first of them is to attach the supercomputers to the network with reasonable performance. This is a nontrivial task, since such machines are often not optimized for external communication. Secondly, there is a need for communication libraries like MPI that offer a high level of abstraction to the application programmer. In order to fully utilize especially a high–speed network some attention has to be paid to the overhead that is introduced by such a library. Finally, the applications have to take into account that the performance characteristics of the external network can't compete with those of the MPP-internal communication.

A German project dealing with these issues is the 'Gigabit Testbed West'. It started in August 1997 as a joint project of the Research Centre Jülich (FZJ) and the GMD – National Research Center for Information Technology in Sankt Augustin and is aimed to prepare the upgrade of the German Scientific Network B-WiN to Gigabit per second capacity which is scheduled for spring 2000. It is

funded by the DFN, the institution that operates the B-WiN. In the first year of operation the two research centers — which are approximately 120 km apart — were connected by an OC–12 ATM link (622 Mbit/s) based upon Synchronous Digital Hierarchy (SDH/STM4) technology. In August 1998 this link has been upgraded to OC–48 (2.4 Gbit/s).

Besides several institutes in the research centers in Jülich and Sankt Augustin other institutions participate in the testbed with their applications. These are the Alfred Wegener Institute for Polar and Marine Research (AWI), the German Climate Computing Center (DKRZ), the Universities of Cologne and Bonn, the National German Aerospace Research Center (DLR) in Cologne, the Academy of Media Arts in Cologne, and the industrial partners Pallas GmbH and echtzeit GmbH.

In this contribution, we report on our activities to set up a metacomputing environment with the parallel computers in Jülich and Sankt Augustin. We first describe the actions we took to improve the connectivity of those computers. Then we give some preliminary performance data of an MPI library, which is currently developed by Pallas for the use in the testbed. We end with a list of the applications that make use of this library, briefly discussing their communication needs.

2 Supercomputer Connectivity

Like in the 155 Mbit/s German Scientific Network, the backbone of the Gigabit Testbed West is an ATM network. In both research centers, a FORE Systems ATM switch (ASX–4000) connects the local networks to the OC–48 line. Since each router in a communication path results in an extra delay of up to a millisecond, the best solution with respect to performance would be to have all machines of the metacomputer in a single classical IP ATM network. Unfortunately, ATM connectivity for supercomputers has evolved quite slowly. While 622 Mbit/s interfaces are now available for all common workstation platforms, solutions are still outstanding for the major supercomputers used in the testbed. In Jülich, this is a Cray complex, consisting of two 512–node T3E MPPs and a 10–processor T90 vector–computer. For all of them, only 155 Mbit/s ATM interfaces are available (and will be in the foreseeable future). The same holds for a 35–node IBM SP2 in Sankt Augustin.

Therefore a different solution had to be found. The best–performing network connection of the Cray supercomputers is the 800 Mbit/s 'High Performance Parallel Interface' (HiPPI). About 430 Mbit/s have been achieved in the IP over HiPPI network of the Cray complex in Jülich. This is mainly due to the fact that HiPPI networks allow IP-packets of up to 64 KByte size (MTU size) and that a limiting factor for the performance is the rate at which the Crays can process IP packets.

In order to make use of that advantage for the testbed, it is essential that the router between the HiPPI and ATM network supports large MTUs on both media. In Jülich, dedicated workstations (currently an SGI O200 and a SUN

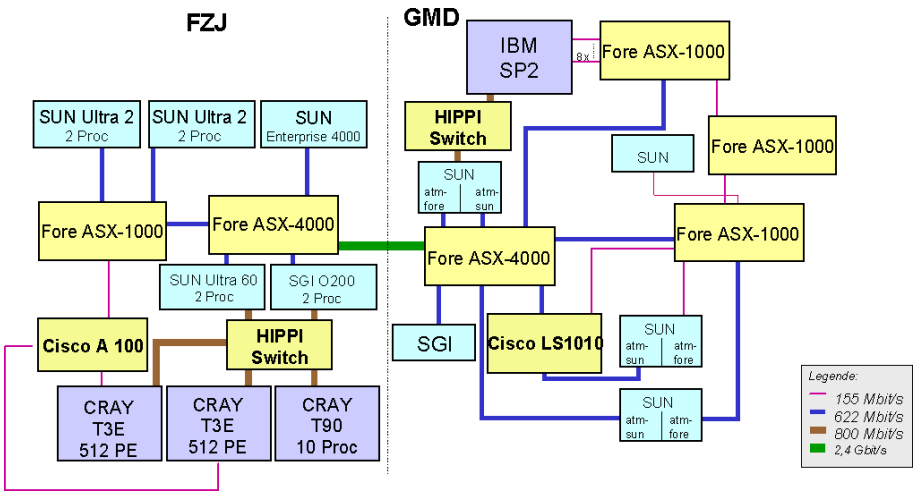

Fig. 1. Configuration of the Gigabit Testbed West in spring 1999. The FZJ in Jülich and the GMD in Sankt Augustin are connected via a 2.4 Gbit/s ATM link. The supercomputers are attached to the testbed via HiPPI–ATM gateways (and additionally via 155 Mbit/s ATM adapters), several workstations via 622 or 155 Mbit/s ATM interfaces.

Ultra 30) are used for that purpose. Both machines are equipped with a FORE Systems 622 Mbit/s ATM adapater supporting an MTU of 64 KByte.

A similar solution was chosen to connect the 35–node IBM SP2 in Sankt Augustin to the testbed. 8 SP2-nodes are equipped with 155 Mbit/s ATM adapters and one with a HiPPI interface. The ATM adapters are connected to the testbed via a FORE ASX 1000. The HiPPI network is routed by an 8–processor SUN E5000 which has also a FORE 622 Mbit/s ATM adapter. A 12–processor SGI Onyx 2 visualization server that is also used by applications in the testbed is equipped with a 622 Mbit/s ATM–interface. Figure 1 shows the current configuration of the testbed as described above.

Depending on the application, latency can be as important as bandwidth. An unavoidable delay is introduced by the travelling time of the signal in the wide–area network. In our case, this delay is about 0.9 msec for the 120 km SDH/ATM line. Each ATM switch adds about 10-20 μsec. As our measurements show, a substantial delay can be introduced by the attached computers. Table 1 lists the results of delay/throughput measurements between various machines in the testbed. More details are given elsewhere [4]. The SUN Ultra 60 – SUN E5000 numbers show that the workstations and servers which are used as HiPPI–ATM gateways can make good use of the available bandwidth without adding a significant delay. The situation is different for the supercomputers. Even in the local HiPPI network in Jülich, a latency of 3 msec is measured on the application level. In order to estimate the effect of the HiPPI–ATM gateways the T3E – T3E measurements were repeated, routing the data the following way: T3E via HiPPI to Ultra 60 via ATM to O200 via HiPPI to the second T3E. This

Table 1. Throughput and delay measurements with a ping–pong program using TCP–stream sockets. The delay is half of the round–trip time of a 1 Byte message. For throughput measurements, messages of up to 10 MByte size were used. The bandwidth available for a TCP connection over the given network is listed as 'theoretical bandwidth'.

	network	theoretical bandwidth [MByte/s]	measured throughput [MByte/s]	untuned throughput [MByte/s]	delay [msec]
T3E — T3E	HiPPI	95.3	44.7	10.2	3.3
	HiPPI/ATM 622	67.3	34.4	6.4	4.0
	Ethernet	1.2	0.97	0.97	4.3
Ultra 60 – E5000	ATM 622	67.3	56.9	33.1	0.91
T3E — SP2	HiPPI/ATM 622	67.3	32.4	1.95	3.1
	ATM 155	16.8	12.5	6.5	2.6

adds only a delay of about 0.7 msec but substantially reduces the throughput. The performance of the T3E – SP2 connection is mainly limited by the I/O capabilities of the MicroChannel based SP2 nodes. For all measurements, large TCP buffers and TCP windows were used. The comparison with untuned sockets (using the operating system defaults) show the importance of socket tuning.

3 MPI communication in the testbed

Many programs written for parallel computers use MPI [5], which offers a portable and efficient API. Therefore, to make metacomputing usable for a broader range of users, the availability of a metacomputing–aware MPI implementation is a must. With metacomputing–aware we mean that the communication both inside and between the machines that form the metacomputer should be efficient. Furthermore, a couple of features that are useful for metacomputing applications are part of the MPI–2 [6] definition. Dynamic process creation and attachment e.g. can be used for realtime–visualization or computational steering; language–interoperability is needed to couple applications that are implemented in different programming languages. When the project started, no MPI implementation with this set of features was available (to our knowledge, this is still true today). Therefore the development of such a library, named 'MetaMPI' was assigned to Pallas GmbH. The implementation is based on MPICH and uses different devices for the communication inside and between the coupled machines. The communication between the machines is handled by so–called router–PEs (processing elements). A message that has to travel from one machine to another is at first handed over to a local router–PE. From there it is transferred to a router–PE on the target machine using a TCP–socket connection. Finally this router–PE sends the message to the target PE. The current status of this development is

Table 2. Throughput and delay measurements with a ping–pong program using MetaMPI. The 'native' values refer to communication inside the machines.

	network	measured throughput [MByte/s]	delay [msec]
T3E	native	333	0.030
SP2	native	42	0.14
T3E — T3E	HiPPI	35.2	4.8
	HiPPI/ATM 622	28.7	5.5
	Ethernet	0.96	5.8
T3E — SP2	HiPPI/ATM 622	16.5	5.1
	ATM 155	7.7	4.3

that a full MPI–1.2 library which supports the Cray T3E, the IBM SP2 and Solaris 2 machines is available. Tuning of the external communication and incorporation of the selected MPI–2 features is under way. A detailed description of the MetaMPI implementation is given in [7]. Here we focus on point–to–point communication performance and compare it with the TCP/IP measurements of the previous section.

A comparison of the delays shown in table 1 and 2 shows that MetaMPI introduces an additional delay of about 1.5 msec for external communication. The bandwidth that is available for external messages is limited by the fact that this message has to take three hops to reach its destination: from the sender to local router–PE, from there to a remote router–PE, and finally to the target. Assuming that the time needed for each hop is determined by the available bandwidth on that path leads to a net bandwidth that is good agreement with our measurements. This means that, whereas careful tuning of the MetaMPI implementation might further reduce the delay, the throughput is near optimal. It should be noted that the measurements used the `MPI_DOUBLE` datatype, which can be exchanged between the T3E and the SP2 without any conversion. Such conversions are necessary e.g. for integer datatypes and significantly reduce the available bandwidth.

4 Distributed MPI Applications

With MetaMPI, the structure of the metacomputer is completely transparent to the application. Nevertheless, it is generally not a good idea to run a tightly coupled parallel application on a metacomputer without taking care of the hierarchy of bandwidths and latencies. Increased communication time often results in reduced efficiency. A class of applications that do not suffer from this problem are so called 'coupled fields' simulations. Here, two or more space- and time–dependent fields interact with each other. When the fields are distributed

over the machines of the metacomputer these components are often only loosely coupled – leading to moderate communication requirements. A second class of applications benefits from being distributed over supercomputers of different architecture, because they contain partial problems which can best be solved on massively parallel or vector–supercomputers. For other applications, real–time requirements are the reason to connect several machines. We now briefly describe the MPI based applications that currently use the Gigabit Testbed West, with a focus on their communication patterns. Details on these applications will be given in seperate publications.

1. **Solute Transport in Ground Water**
   Partners: Institute for Petroleum and Organic Geochemistry, FZJ.
   Synopsis: Two independent programs for ground water flow simulation (TRACE, FORTRAN 90) and transport of particles in a given water flow (PARTRACE, C++) are coupled. The distributed version will allow for larger simulations, since both applications have high CPU and memory requirements.
   Communication: The 3–D water flow field is transferred from the IBM SP2 (TRACE) to the Cray T3E (PARTRACE) at the beginning of every timestep. For typical parameters, 10–100 MByte are transferred in a burst every 2–10 seconds.
2. **MEG Analysis**
   Partners: Institute of Medicine, FZJ.
   Synopsis: A parallel program (pmusic) that estimates the position and strength of current dipoles in a human brain from magnetoenzephalography measurements using the MUSIC algorithm is distributed over a massively parallel and a vector supercomputer. One part of the MUSIC algorithm can use up to 100 CPUs very efficiently, while another needs a single fast CPU. Therefore superlinear speedup is expected when the application is run on a heterogeneous metacomputer.
   Communication: Only few data are transferred, but the algorithm is sensible to latency, because one cycle takes only a few milliseconds.
3. **Distributed Climate and Weather Models**
   Partners: Alfred Wegener Institute for Polar and Marine Research, German Climate Computing Center, and the Institute for Algorithms and Scientific Computing (SCAI), GMD.
   Synopsis: A parallel ocean–ice model (based on MOM–2) running on Cray T3E and a parallel atmospheric model (IFS) running on IBM SP2 are coupled with the CSM flux coupler that is also run on the T3E. Both programs have high CPU requirements.
   Communication: Exchange of 2–D surface data in every timestep (typically 2 seconds), up to 1 MByte in short bursts. Although the average network load is small, high bandwidth is needed since the application blocks until all data are exchanged.

4. **Distributed Fluid–Structure Interaction**
   Partners: Pallas GmbH and SCAI, GMD.
   Synopsis: An open interface (COCOLIB) that allows the coupling of industrial structural mechanics and fluid dynamics codes has been developed in the EC–funded project CISPAR. This is ported to the metacomputing environment.
   Communication: Depends on the coupled applications.
5. **Multiscale Molecular Dynamics**
   Partners: Institute for Applied Mathematics, University of Bonn.
   Synopsis: Two existing parallel programs that simulate molecular dynamics with short–range (MolGrid) and long–range (TreeMol) interactions will be coupled and shall run distributed on the Cray T3E in Jülich and the PARNASS parallel computer in Bonn.
   Communication: Atom positions have to be exchanged in every timestep of the long–range/short–range interaction, resulting in an average load of 40 MByte/s. Latency is hidden by the algorithm.
6. **Coupled Litospheric Processes**
   Partners: Institute for Applied Mathematics and Institute of Geodynamics, University of Bonn.
   Synopsis: Coupled litospheric processes involving fluid dynamics, structural mechanics, heat exchange and chemical processes will be simulated by four parallel programs running on the Cray T3E, the IBM SP2 and PARNASS.
   Communication: About 2 MByte (after an application–specific compression) must be transferred at each timestep. Such a timestep is expected to take 0.1 seconds.

5 Related Activities

The applications in cooperation with the University in Bonn rely on an extension of the testbed to Bonn that will be operable in spring 1999. Further extensions to the University, the National German Aerospace Research Center, and the Academy of Media Arts in Cologne are currently being installed and will be used by new application projects from the areas of multimedia and metacomputing. Another metacomputing application in the testbed deals with realtime analysis and visualization of functional magnetic resonance imaging (fMRI). It does not use MPI for the external communication and is described in more detail elsewhere [9].

6 Conclusions

In this contribution, we presented first results of our efforts to establish efficient communication for MPI based metacomputing applications in the Gigabit Testbed West. The underlying 2.4 Gbit/s SDH and ATM technology for the wide–area backbone is mature, a neccessary condition for the upgrade of the

German Scientific Network that is planned for early 2000. The implementation of the HiPPI–ATM gateways for the Cray T3E and IBM SP2 resulted in a significant enhancement of their networking capabililities. Still there seems to be room for improvements in this area. The MetaMPI library imposes only little overhead to the raw TCP/IP communication (in terms of bandwidth), yet delivering a standard API for message passing programs. Nevertheless, there is a dramatic difference between the performance of the MPP–internal and the external communication. The applications of 'coupled fields' and 'heterogenous metacomputing' type will have to prove that this difference can be compensated for.

7 Acknowledgements

Many of the activities that are reported in this contribution are not the work of the authors but of several persons in the institutions that participate in the Gigabit Testbed West project. The authors wish to thank R. Niederberger, M. Sczimarowsky, from the Research Centre Jülich, U. Eisenblätter, F. Hommes, P. Wunderling, and L. Zier at the GMD, J. Worringen, M. Pöppe, T. Bemmerl RWTH Aachen, to mention but a few. We also wish to thank the BMBF for partially funding the Gigabit Testbed West and the DFN for its support.

References

1. I. Foster and C. Kesselman, The Globus Project: A Status Report. Proc. IPPS/SPDP '98 Heterogeneous Computing Workshop, pp 4-18, 1998.
2. H.E. Bal, A. Plaat, T. Kielmann, J. Maassen, R. van Nieuwpoort, and R. Veldema, Parallel Computing on Wide-Area Clusters: the Albatros Project, Proc. Extreme Linux Workshop, pp. 20-24, Monterey, CA, June 8-10, 1999.
3. E. Gabriel, M. Resch, T. Beisel, R. Keller, Distributed computing in a heterogeneous computing environment, in V. Alexandrov and J. Dongarra (eds.) Recent Advances in PVM and MPI, pp 180-197, Springer 1998.
4. H. Grund, F. Hommes, R. Niederberger, and E. Pless, High Speed Supercomputer Communications in Broadband Networks, TERENA–NORDUnet Networking Conference, Sweden, June 1999.
5. Message Passing Interface Forum, MPI: A Message-Passing Interface Standard, University of Tennessee, http://www.mcs.anl.gov/mpi/index.html, 1995.
6. Message Passing Interface Forum, MPI-2: Extensions to the Message-Passing Interface, University of Tennessee, http://www.mcs.anl.gov/mpi/index.html, 1997.
7. J. Worringen, M. Pöppe, T. Bemmerl, J. Henrichs, T. Eickermann, and H. Grund, MetaMPI – an extension of MPICH for Metacomputing Environments, submitted to ParCo99, Delft, August 1999.
8. G.D. Burns, R.B. Daoud, and J.R. Vaigl, LAM: An Open Cluster Environment for MPI, Supercomputing Symposium '94, Toronto, Canada, June 1994.
9. T. Eickermann, W. Frings, S. Posse, and R. Völpel, Distributed Applications in a German Gigabit WAN, accepted for High Performance Distributed Computing, Los Angeles, August 1999.