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

Interner Bericht

High Speed Supercomputer Communications in Broadband Networks

Helmut Grund[*], Ferdinand Hommes[*],
Ralph Niederberger, Eva Pless[*]

FZJ-ZAM-IB-9903

April 1999
(letzte Änderung: 20.04.99)

(*) GMD – German National Research Center for Information Technology, Schloss Birlinghoven

Preprint:
To be published in: Proceedings of the TERENA-NORDUnet Networking Conference 1999
7-10 June 1999, Lund, Sweden

High Speed Supercomputer Communications in Broadband Networks

Helmut Grund[2], Ferdinand Hommes[2] , Ralph Niederberger[1], Eva Pless[2]

[1] FZJ - Research Centre Jülich
D - 52425 Jülich
http://www.fz-juelich.de
r.niederberger@fz-juelich.de

[2] GMD – German National Research Center for Information Technology
Schloss Birlinghoven, D-53754 Sankt Augustin
http://www.gmd.de
Eva.Pless@gmd.de , Ferdinand.Hommes@gmd.de , Helmut.Grund@gmd.de

Abstract

High speed communication across wide area networks is currently of great interest in computer communications industry. It is needed wherever a huge amount of data will be created and made available to remote computer systems. Areas of investigation in computer science are the coupling of supercomputers (metacomputing), high-resolution visualisation of data and distributed access to huge amounts of data.

Within the Gigabit Testbed West (GTB-West) a high speed communication link has been installed between the research center Forschungszentrum Jülich GmbH (FZJ) and the National Research Center for Information (GMD) at Sankt Augustin.

Amongst the research fields Metacomputing, Visualization and distributed usage of huge amount of data the installation, management and performance evaluation of the data communication links are main points of interest to the GTBW project.

Possible solutions to connect supercomputers for Metacomputing applications based on different communication technologies like HIPPI, Sonet and ATM are discussed and the pros and cons of the final implementation are presented.

Keywords: Metacomputing, Supercomputer, Broadband Networks, HIPPI, ATM, TCP/IP Tuning

Introduction

In the area of scientific computation many applications have a great demand for extreme computing power, which can be satisfied by parallel computers which are connected over a high speed network. This sort of computation using distributed applications is known as Metacomputing. The connection of supercomputers of different types and vendors adds a further degree of complexity to Metacomputing (parallel computers and vector supercomputers). The availability of broadband communication networks with high speed links provides new solutions to connect supercomputer at remote sites.

Most current supercomputer systems do not support high speed communication interfaces which can be used for remote communication in broadband networks. They use special hardware and software communication infrastructures within their own proprietary world that guarantee the best performance results. When communicating over public wide area networks gateways and specialized switches have to be used which will affect communication performance.

After discussing the different possibilities to connect supercomputers via an ATM network this report will present the final solution which has been used in the GTB West. The solution found offers a great amount of throughput with minimal costs.

The main goal of the Gigabit Testbed West [GTB-West], a project sponsored by the German Research Network Association (DFN-Verein) [GTB-DFN], is to provide a gigabit network infrastructure for scientific computing and to collect experiences operating a high-speed wide-area computer network. Further goals are the implementation of applications with a huge requirement of communication throughput, the coupling of parallel and vector supercomputers (metacomputing), high-resolution visualisation of data and distributed access to huge amounts of data. Another point of interest is to get experience in the coupling of the supercomputers of German supercomputer centers over a high speed WAN for the future. A number of projects like solute transport in ground water, algorithmic analysis of magnetoenzephalography data, complex visualization, multimedia applications, distributed calculations of climate and weather models have been designed and implemented to show the functionality of a metacomputing environment. The Gigabit Testbed West projects started in August 1997 with a 622 Mbps ATM-link between the German National Research Center for Information Technology (GMD) in Sankt Augustin and the Research Center Jülich (FZJ) in Jülich. At the end of July 1998 the link was successfully upgraded to 2.4 Gbps. Just 3 days later a maximum of 2.37 Gbps throughput containing IP user data could be transfered via this new link using a number of SUN workstations with 622 Mbps ATM interfaces.

The connection of the five SGI/CRAY supercomputers at FZJ and the IBM SP2 at GMD is shown in the figure below.

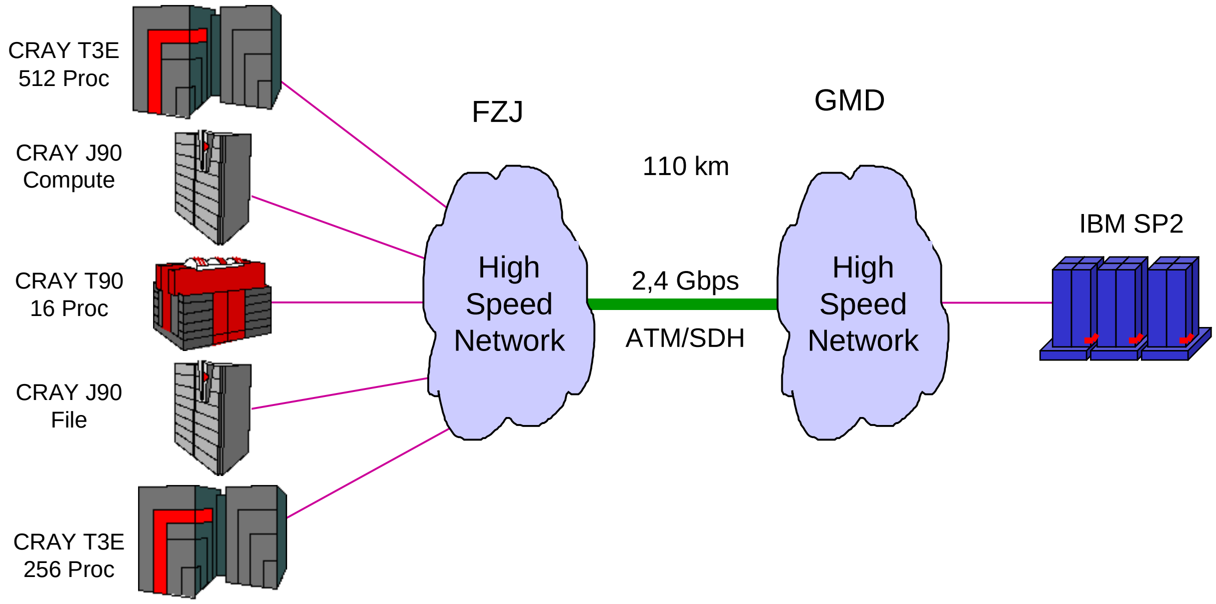

Figure 1 – Logical configuration of the GTB supercomputer networks

CRAY-T3E and IBM-SP2 Communication Techniques

Supercomputers from different vendors normally use their own internal communication hardware and protocols for optimal performance between homogeneous computer systems whereas standard network technologies like Ethernet, HIPPI and ATM and most times the TCP/IP protocol are used to communicate with heterogeneous computer systems.

The following figure shows the current configuration of the IBM SP2 system installed 1995 at the

GMD. The 37 nodes are connected via different networks to each other and to external machines. An IBM High Performance Switch (HP Switch) provides a proprietary high speed local connection between the system nodes. The hardware bandwidth of the switch supports 1200 Mbps. Via this switch up to 380 Mbps can be reached with the native "user space protocol" whereas the TCP/IP protocol delivers only 96 Mbps. To get high throughput to external machines 8 nodes have been equipped with 155 Mbps ATM interfaces. With this type of adapter only up to 100 Mbps throughput can be reached. All nodes of the IBM-SP2 system are equipped with two Ethernet adapters. One is used for internal management operations and the other is used for external connections. A dedicated node is additionally equipped with a HIPPI interface for enabling high speed external connections. Though the native HIPPI protocol allows transfer rates of up to 800 Mbps, it has been examined, that running TCP/IP over HIPPI [RFC-2067] leads to only 370 Mbps throughput because of protocol and software overhead.

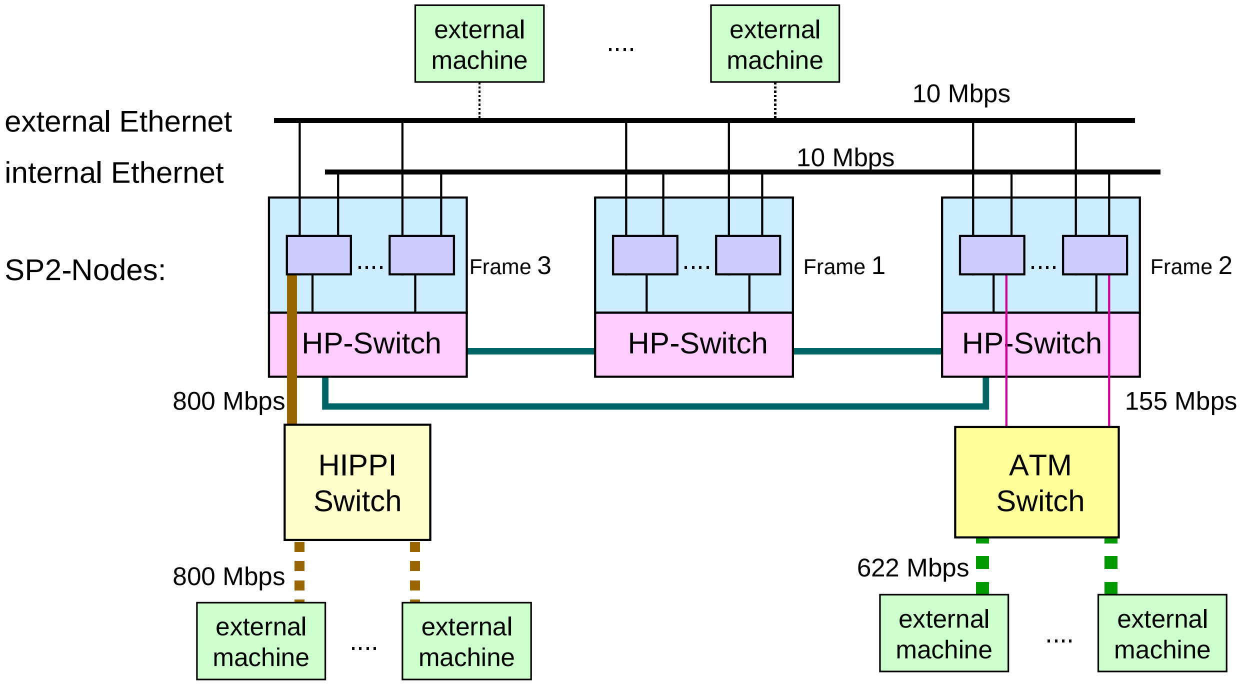

Figure 2 - Network connections of supercomputer SP2 at GMD

The CRAY systems use a different architecture for external communications. This communication architecture is shown in Figure 3. The CRAY nodes (processors) are internally connected via a 3 dimensional torus, which provides a peak link rate of 600 MBps and a data transfer rate of up to 500 MBps. For network I/O operations one of four processors is connected to a GigaRing channel, which supports a *Raw* bandwidth of up to 1.2 GBps throughput (half duplex payload=941 MBps, full duplex payload=2*827 MBps). Network I/O is done using a number of processes running in the T3E system on one or more nodes (Application PE, Communication PE, Packet driver PE; PE = processing engine) depending on system configuration. The optimal internal configuration could not be found until now. External communication can be done using special I/O nodes which are attached to the GigaRing channel. These I/O nodes are realized using Sparc processors with a special operating system providing Ethernet, HIPPI and ATM interfaces.

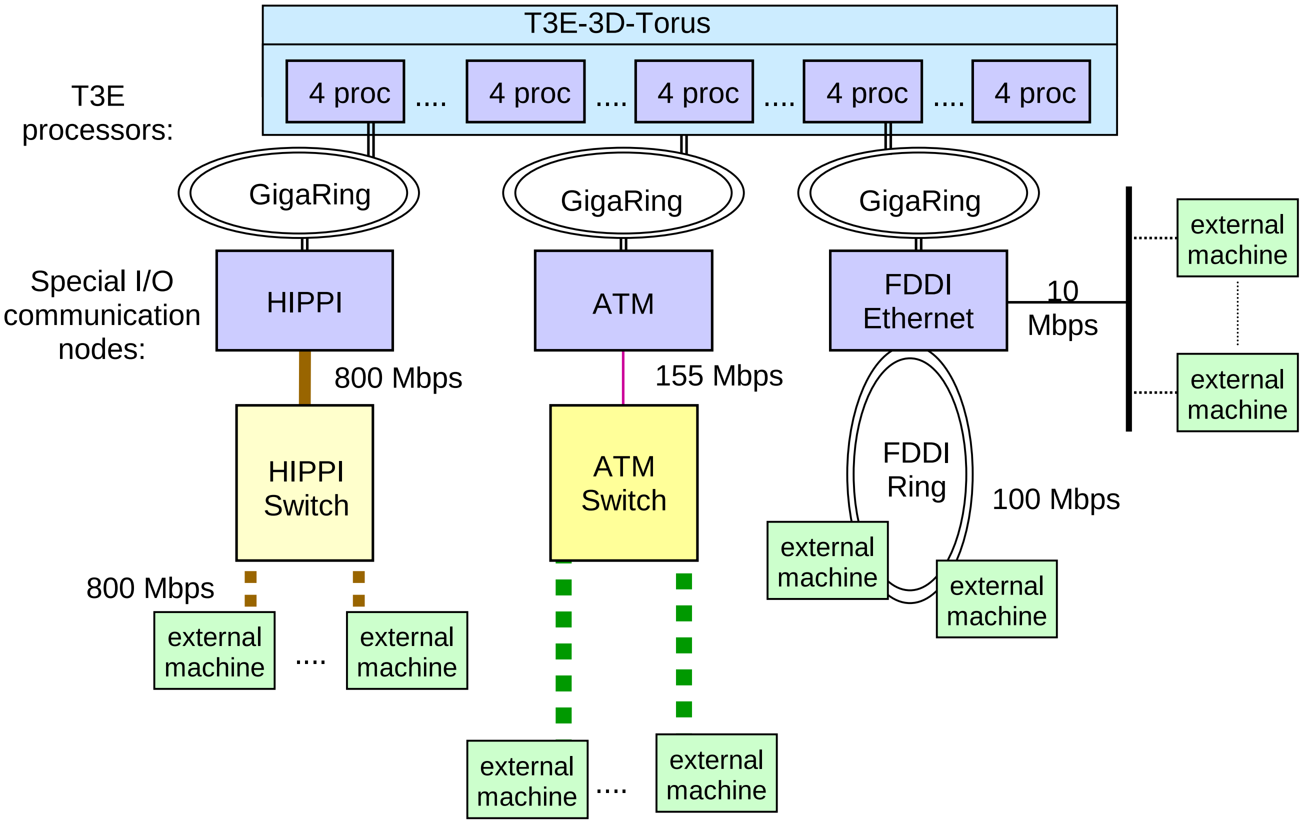

Figure 3 - Network connections of supercomputer CRAY T3E at FZJ

The communication throughput of the supercomputer systems depends, on the selection and implementation of the communication software as mentioned above. The Message Passing Interface [MPI], a de-facto standard of the MPI-Forum [MPI-Forum], is widely used for the communication between parallel processes. This protocol has been also used in the Gigabit Testbed West for the communication between the processors of the same and remote system. Most MPI implementations allow to define point-to-point communication between a fixed number of processes running on different processors of the same system or point-to-point communications between remote processors. Processes communicate by calling library routines that hide the actual physical implementation to the application program. To allow dynamic configuration of communication between remote processors a subset of the MPI-2 library [MPI-2] had to be implemented in the context of the Gigabit Testbed West that delivers additional Metacomputing functionality. It cares for the unique appearance of all nodes at the different supercomputers and therefore communicates internally via the high speed proprietary protocols and externally via TCP/IP. The implemented concept allows the use of one or multiple nodes at each side as communication nodes. This configuration flexibility has been implemented to allow maximum communication throughput between the participating systems.

Possible solution connecting supercomputer systems of different vendors

The main point of interest and basis of any possible solution for the communication between the different supercomputer systems is the underlaying communication infrastructure. The Research Center Jülich (FZJ) and the German National Research Center for Information Technology (GMD) are connected via a 2.4 Gbps ATM connection based on SDH technology. This ATM link is not dedicated for the connection of the supercomputers only but has to be shared with other applications, for example multimedia traffic. Furthermore the main communication protocol used for computer communication is the TCP/IP protocol suite which helps to avoid implementations of proprietary gateways.

Nevertheless there are a lot of interesting possibilities to connect the CRAY systems at FZJ and the IBM-SP2 at GMD.

- Configuring the CRAY systems at FZJ and 1 to 3 nodes of the SP2 with 622 Mbps ATM interfaces would permit a theoretical throughput of 3x622 Mbps (Figure 4). Since CRAY and IBM do not plan to support 622 Mbps ATM interfaces for the systems installed at FZJ and GMD only 155 Mbps ATM interfaces can be used to connect the systems to the GTB-West. Because of promises of the both manufactures to support these interfaces in future first tests with 155 Mbps interfaces have been done. This solution has the disadvantage that the maximum bandwidth available between the supercomputer is 155 Mbps. Furthermore the adapter software of the CRAY systems supports PVC connections only, which leads to additional management overhead in contrast to SVCs (Establishing connections to n systems requires definition of 2*n PVC's on the n+1 systems). Another disadvantage is the maximum MTU size of 9180 byte supported by Classical IP (CIP) [RFC-1577]. Because of architectural restrictions of the interrupt rate at the CRAY T3E systems and the restricted MTU size the communication throughput is limited to 115 Mbps on these adapters.

  To expand the communication throughput at the SP2 system IBM has developed an IBM-Switch 04S and IBM-Switch 16S based on the ASCEND GRF 400/1600 Gigarouter with a 622 Mbps ATM interface. A special board and driver software for this switch allows direct connection to the SP2 High Performance switch. The high expenses required for this hardware solution which are in contradiction to the expected low communication throughput did not allow to use this solution.

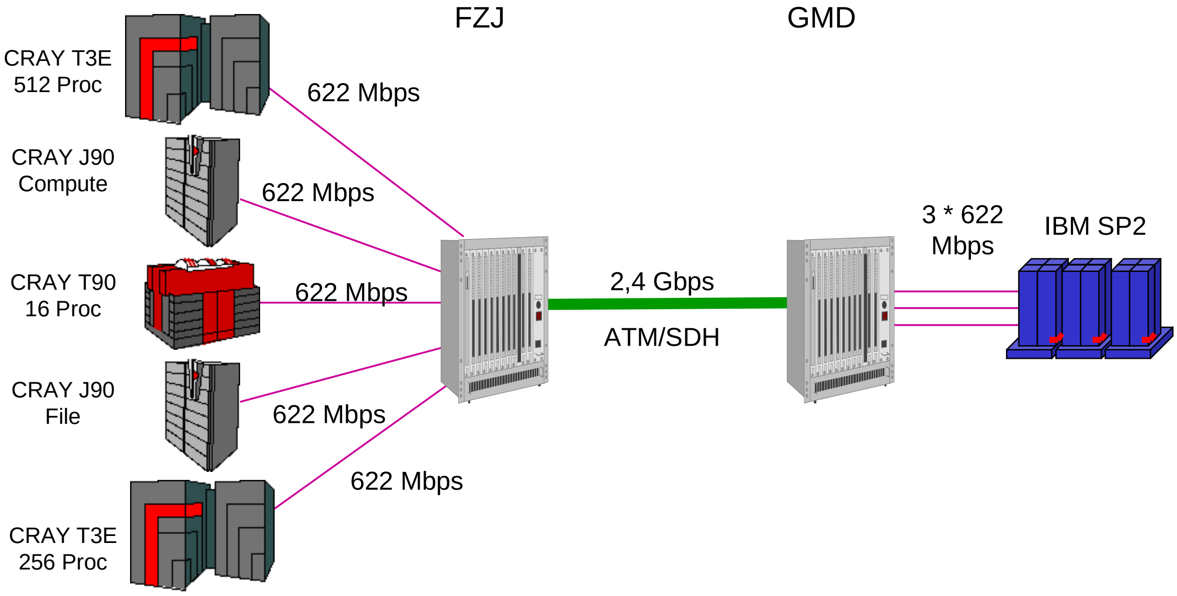

Figure 4 - Hypothetical direct connection via ATM interfaces

- Another solution could be the installation of HIPPI/SONET Extenders from Essential at both locations. Since a dedicated channel has to be reserved for the exclusive use of the HIPPI-SONET connection (155 Mbps up to 932 Mbps) only a reduced bandwidth would be available for the other applications on the Gigabit Testbed. This would require to configure the remaining channels as separate ATM links with additional interfaces in the ATM switch. These channels would have to be collected and distributed by the SDH equipment.

- The usage of the installed HIPPI interfaces seems nevertheless to be a promising solution. There are some other possibilities to use these interfaces:

  – Tunneling of HIPPI over ATM requires the installation of a specialized gateway. A workstation could be used as gateway. This solution needs the development of some software components for packing and unpacking HIPPI blocks into AAL packets.

- Another possibility is the usage of a router solution (CRAY - HIPPI - GRF400 - ATM - GRF400 - HIPPI - IBM/SP2). The only router currently supporting both HIPPI and ATM 622 Mbps interfaces is the ASCEND GRF 400 respectively GRF 1600. This router can be used in two configurations. First configuration is using normal routing functionality. Another possibility is to use the routers tunneling capability which supports HIPPI via ATM tunneling. This would be the same functionality as described above with specialized workstations. At the moment the ASCEND GRF 400 is the only gigabit router which supports this functionality. Other well-known router manufacturers are not interested, because of the reduced market for routers of this very special category. Furthermore this solution is a very expensive one and also demands the use of Classical IP with the implementation specific required MTU size of 9180 bytes.

- The use of UNIX multi-processor computers with PCI-bus acting as routers at both locations is another interesting solution. A number of tests with UNIX multi-processor machines with PCI-bus (Enterprise 5000 / 8 processors, SUN Ultra 60 / 2 processors) showed that the computers were able to route between two 622 Mbps ATM interfaces at full line speed. This leads to the final solution shown in Figure 5 which also shows the nominal bandwidth capacity of the intermediate communication links. The workstations have been equipped with a serial HIPPI interface from Essential and a 622 Mbps ATM interface from FORE and have been configured as IP routers between HIPPI and ATM (H/A-router). This solution is a very simple and cost-effective one, since only a restricted routing functionality is asked. The installation of expensive full-functional routers would have been superfluous.

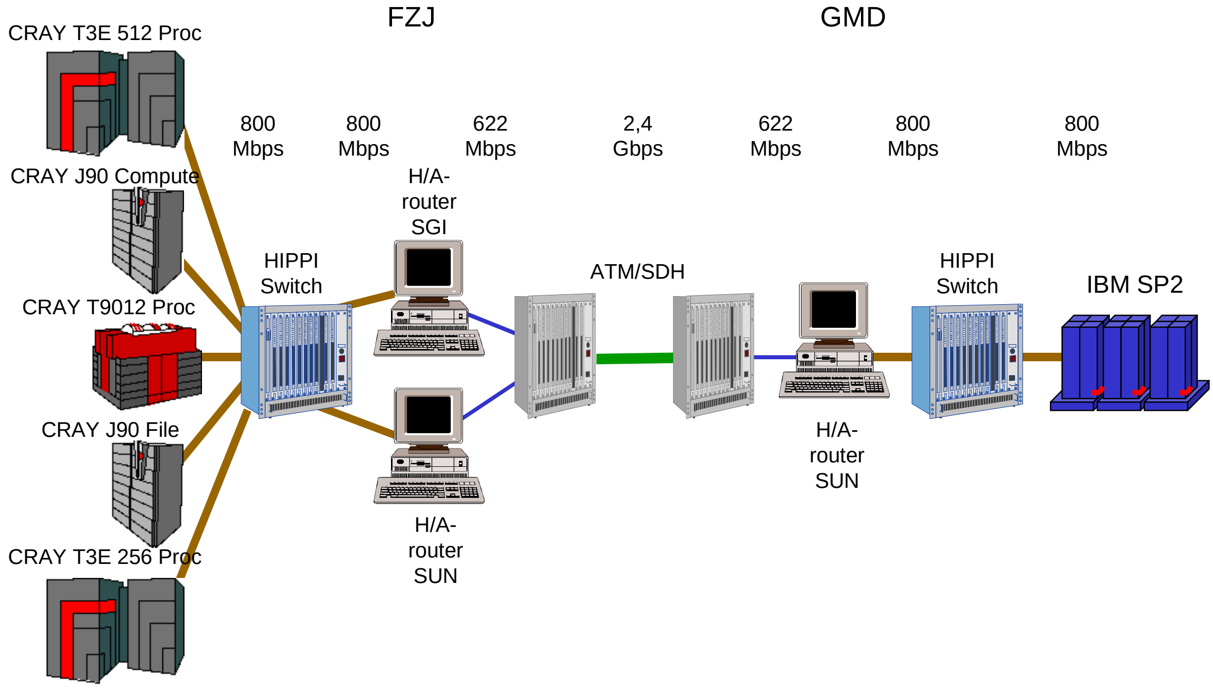

Figure 5 - The current connection of the supercomputers (with nominal bandwidth of the links)

TCP/IP Performance tuning for optimal throughput

Research in gigabit data transfer rates has a long history [Partridge93]. As mentioned above very high speed data transfer rates between supercomputers can only be reached if tuning of the communication parameters has been done [RFC-1323] [HighPerf]. Special tests at the CRAY systems have been made to find out values for the maximum interrupt rates of this systems. The tests have shown, that the number of I/O operations per second a CRAY T3E system can handle is extremely limited. To get high throughput this interrupts have to be minimized, which only can be done if the TCP/IP packets can be made as large as possible. The MTU size for HIPPI networks on CRAY systems is configured to 65280 bytes as default. ATM between the HIPPI-to-ATM routers can be used in two modes. ATM LANE uses the standard Ethernet MTU size of 1500 bytes, whereas the recommended default MTU size for Classical IP is 9180 bytes. Therefore to maximize paket length and minimize network I/O interrupts Classical IP should be preferred.

The SUN and SGI workstations on the FZJ side and the SUN Enterprise on GMD side shown in figure 5 above have been used for the routing between HIPPI and ATM. This implies that the supercomputers are not connected to the same IP network, therefore Path MTU Discovery [RFC-1191] should be supported. Otherwise a default of 512 Bytes for external networks would be used for the whole connection. The CRAY I/O interface node and the H/A-router support this feature. For the SP2 system the tcp_mssdflt value had to be changed from 512 Bytes to 64 KBytes.

Fortunately all 622 Mbps adapters (FORE HE-622) that have been used support configurable MTU sizes up to 64 KBytes. Since no broadcasts are required within the GTB-West and only point-to-point communications will be used the CIP protocol has been choosen. Therefore the maximum MTU size of 64 Kbytes could be configured for the ATM link between the two HIPPI-to-ATM gateways at GMD and FZJ. This implies a 64 KByte MTU size along the whole data path which leads to maximum paket sizes in both communication directions.

Other tuning activities which have to be done are to use the TCP-Winshift option and to increase the *send* and *receive socket buffer sizes* including the *maximum socket buffer space* [RFC-1323]. All involved systems allow socket buffer sizes and window sizes of more than 64 Kbytes which are needed because of the measured round trip delays. The delays are composed of delays within the switches, the start and end systems, the H/A routers and the delays because of the speed of light. The measured total roundtrip delay from SP2 at GMD to CRAY systems at FZJ is about 7 ms, which consists of approximately 1.4 ms for the 110 km 2.4 Gbps ATM connection including the ATM switches on both sides. The local delays are mainly determined by the delay caused by the HIPPI/ATM routing in the H/A-routers, which is about 1ms at each side and by the HIPPI connections, which are 1 ms at the SP2 side and 3 ms at the CRAY side. Using 622 Mbps ATM interfaces the product from bandwidth and delay concerning a delay of 7 ms amounts to 4354 Kbits or 544 KBytes. This is the minimum value which should be used for the send and receive socket buffer sizes. The default maximum value on all machines was set to 1 MByte, thus supporting optimal performance for delays up to 13 ms.

Besides this the receive buffer pool (Type D) of the SP2 HIPPI adapter had to be increased from 3x64 KBytes to 20x64 KBytes to support window sizes of 1 MByte. Unfortunately there were also some software issues with HIPPI adapters of the SP2. These problems result in adapter hang-ups when using small packet sizes. IBM has been informed about these problems.

Throughput measurements have been done with the program Netperf Version 2.1pl3 from HP [Netperf]. Using the TCP/IP tuning described above a total throughput of 370 Mbps was achieved for TCP/IP data transfer between the supercomputers. Figure 6 shows the maximum bandwidth available on the different parts of the link. The throughput between the H/A-routers is close to the theoretical maximum for Classical IP over ATM (538 Mbps). At the moment the HIPPI nodes on the supercomputers are the bottleneck; they allow only up to 370 Mbps on the IBM-SP2 side and 430 Mbps on the CRAY T3E side. Further tests will be made with the CRAY T90 system in FZJ, a vector supercomputer system with 10 cpu's. The T90 system can handle about 5 times higher I/O interrupt rates as seen in the CRAY T3E system. It will be interesting to see if the maximum throughput capacity of the gateway systems, SUN and SGI, can be reached. Special tests with SUN and SGI systems involved showed, that a throughput of 480 Mbps can be reached (SUN-Enterprise5000 - ATM622 - SGI-O200 - HIPPI - SUN-Ultra60).

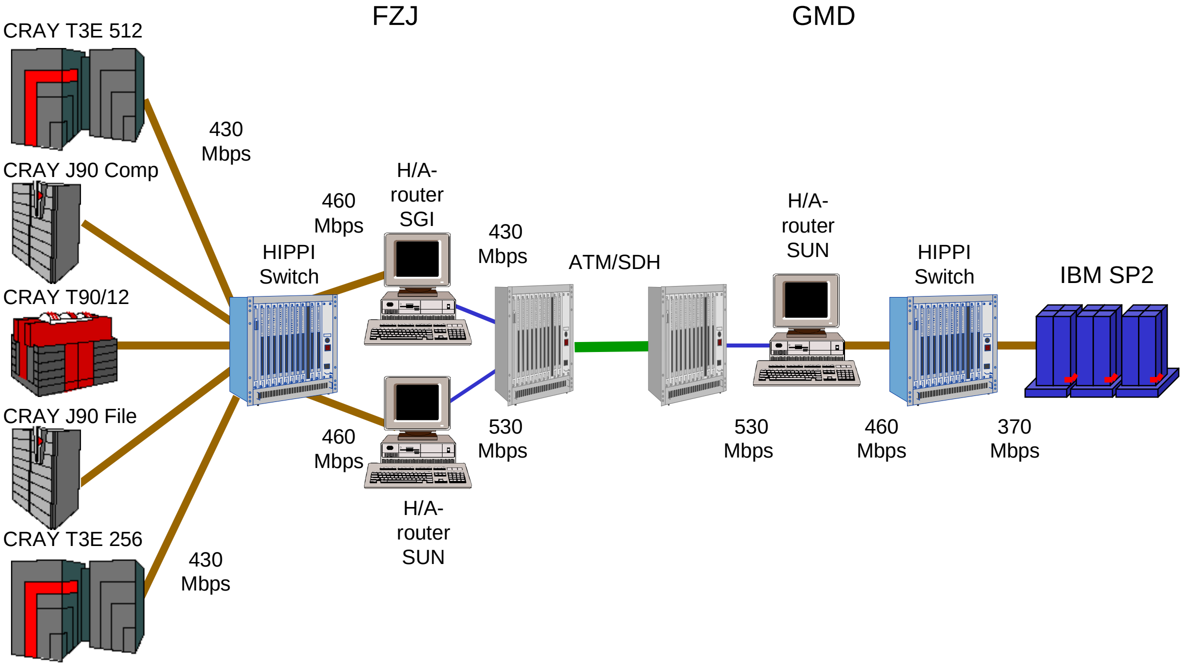

Figure 6 - Actual bandwidths for TCP/IP

Conclusions

The goal of the GTB-West project has been achieved to show that Metacomputing, the coupling of parallel and vector supercomputers, via an ATM gigabit network can be done today. The HIPPI/ATM routers (workstations) could be configured with very cost-effective standard components. To get maximum bandwidth extensive TCP/IP tuning was necessary. This bandwidth seems to be currently an upper limit for the Metacomputing applications in the Gigabit Testbed West. How much of this bandwidth can really be used today depends on the MPI 2 implementation, which is currently under development, and on the applications themselves.

For further enhancement of the maximum bandwidth several aspects can be investigated. One alternative would be to add additional HIPPI nodes with corresponding H/A-routers to the SP2 and to provide further separate H/A-routers for each CRAY. Figure 7 shows an example solution. This would yield a higher bandwidth maybe over 1 Gbps. Currently no tests have been made if other bottlenecks would arise with such configurations. The MPI library which is currently developed will support this multiple-to-multiple-point solution. The second alternative would be to use native HIPPI between the supercomputers. This requires that applications have to be written for the workstation routers, which encapsulate HIPPI-packets into ATM-packets. In order to transmit the 800 Mbps traffic over ATM 622 Mbps two adapters for each H/A-router are required. Using a HIPPI over ATM tunneling mechanism would require special arrangements within the developed software because of HIPPI protocol specifics on long distance communication lines.

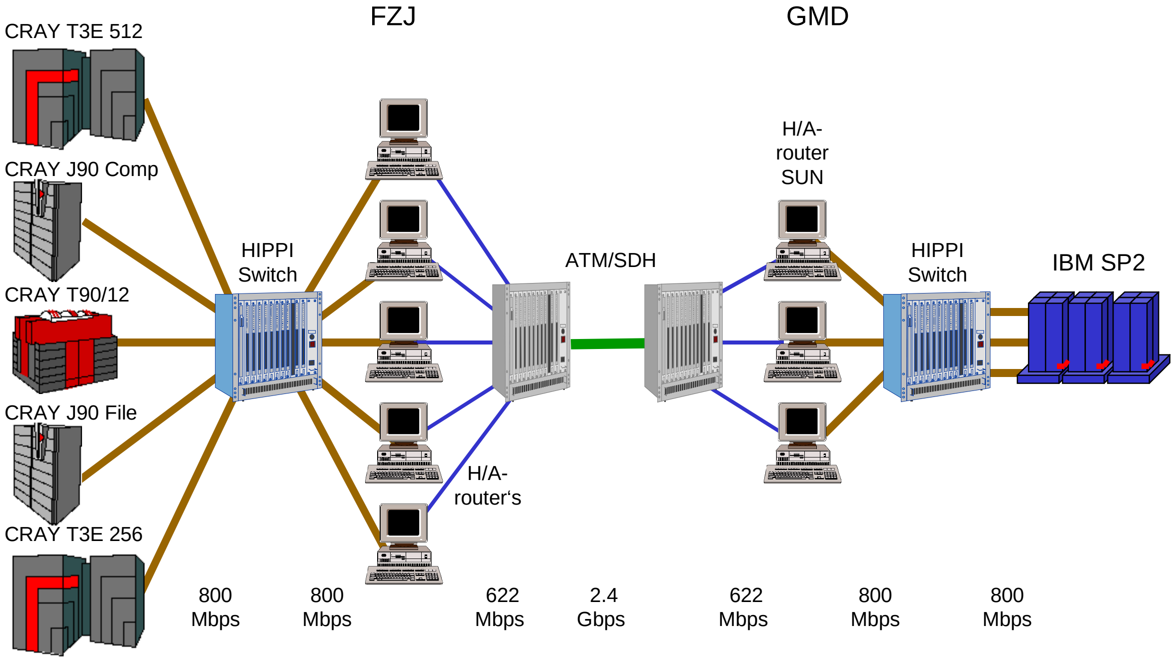

Figure 7 - Increase of bandwidth

The solution implemented was necessary because the supercomputers are currently only optimized for local performance. Proprietary communication solutions (e.g. GigaRing for CRAY, High Performance Switch for SP2) exist setting up local networks. Very high speed communication protocols as HIPPI 6400 with 6.4 Gbps [HIPPI6400-PH] and ATM OC-192 with 10 Gbps are just standardized or in development but currently not available. What is really needed is a common standardized communication protocol which runs over WANs and is supported by all manufacturers. This could be for example a 2.4 Gbps ATM solution.

Acknowledgements

This project has been supported by the BMBF and the German Research Network Association (DFN). The work described in this paper has been performed within the Gigabit Testbed West project in close teamwork with colleagues from the GMD institutes SCAI (A. Klier, R. Völpel) and IMK (U. Eisenblätter, M. Merzbach, P. Wunderling, L. Zier) and the FZJ institute ZAM (T. Eickermann, R. Sczimarowski).

References

[GTB-West] Gigabit Testbed West: http://www.fz-juelich.de/gigabit

[GTB-DFN] Gigabit Projects of DFN: http://www.dfn.de/projekte/gigabit/home.html

[Comer91] D. E.Comer, Internetworking with TCP/IP, Volume I, Principles, Protocols and Architecture, ISBN 0-13-468505-9, Prentice Hall, 1991

[RFC-2067] J.Renwick, IP over HIPPI, http://www.HIPPI.org/rfc2067.pdf , Jan. 1997

[MPI-Forum] MPI Forum: http://www.mpi-forum.org

[MPI] A Message-Passing Interface, Version 1.1

[MPI-2] Extension to the Message-Passing Interface

[RFC-1577] Classical IP and ARP over ATM, ftp://nis.nfs.net/documents/rfc/rfc1577.txt

[Partridge94] G.Partridge, Gigabit networking, ISBN 0-201-56333-9, Addison Wes., 1994

[RFC-1323] TCP Extensions for High Performance, ftp://nis.nfs.net/documents/rfc/rfc1323.txt

[HighPerf] Enabling High Performance Data Transfers on Hosts, http://www.psc.edu/networking/perf_tune.html

[RFC-1191] Path MTU Discovery, ftp://nis.nfs.net/documents/rfc/rfc1191.txt

[Netperf] Benchmark Program Netperf: http://www.netperf.org

[HIPPI6400-PH] HIPPI - 6400 Mbit/s Physical Layer (HIPPI-6400-PH Rev 2.4) , April 1996, http://www.HIPPI.org/c6400PH.html