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

Interner Bericht

Super Computer Communications

Ralph Niederberger

FZJ-ZAM-IB-9905

April 1999
(letzte Änderung: 21.04.99)

Preprint:
To be published in: Proceedings of the Cray User Group Conference 1999
24-28 May 1999, Minneapolis, USA

Super Computer Communications

Ralph Niederberger
R.Niederberger@FZ-Juelich.de

Abstract

In the last decade the communication throughput of computer networks has increased in more and more shorter timeperiods. Furthermore new networking protocols with essential higher bandwidths get available. Concerning this evolution the distribution of parallel applications which are connected via a high speed computer network becomes increasingly interesting.

This paper will examine what currently available supercomputer systems can contribute to such metacomputing environments.

Introduction and Motivation

Since many years computer simulation has become attractive as another scientific method beneath experiment and theory for the solution of scientific technical problems. This method known as *Computational Science and Engineering* can best be used to solve mathematical complex and non linear systems [li01].

Concerning this new challenge in 1995 the Research Center Jülich has decided to install a new supercomputer complex one of the most powerful systems world wide at this time. This supercomputer complex provides new opportunities to the scientific staff of the Research Center Jülich, to partners in industry and economy as well as to the projects of the *John von Neumann Institut für Computing* (NIC) (formerly HLRZ) distributed across Germany to solve today's g*rand challenges* using computer simulation.

Until mid of the 90^{th} the development of vector supercomputers has been main point of interest. Currently massively parallel systems have become focus of scientific investigation [li03].

Beneath the system architecture, node performance and network technology extended analysis of parallel algorithms, programming methods and programming models will become highly recommendable. Today large application packages get more complex and will become increasingly heterogeneous with respect to algoritms. This implies to solve this problems with heterogeneous computer system complexes where the contributing systems are connected via a high speed network.

Introducing Metacomputing

Many problems can be divided into different parts which can be best solved on the one side by massively parallel systems and on the other side by vector systems. It is also best practice to divide similar parts onto different systems of the same architecture to get a speed up because of more computing power (more capacity).

Connecting many of these computer systems classes of applications can be managed which would be not solvable with currently available standalone supercomputer systems. This kind of problem solving is known as metacomputing.

Metacomputing in general means distributing an application onto two or more computer systems with similar or heterogeneous architecture which are dynamically connected by an external network.

An interesting question is which prerequisites are needed to get a significant speedup in computation time or to enlarge the problem classes which can be solved.

There are essentially two different problem classes.

- Solving a problem with 10 processors which can be parallalized by 10 % only will lead to a speedup of only 9 %. This is because 90 % of the application have to be solved in a scalar fashion. Using 100 processor elements instead of 10 processors will give another 0.9 % speedup compared to the single processor solution.
- Scaling the parameter space of a problem by a factor of *n* will most times increase the computing time to the order of $O(n^2)$ or $O(n^3)$. Using many massively parallel computer systems seems to be the best solution to get these problems solved. Here of course the metacomputing concept is extremely limited by the number of systems available.

Both problem classes are highly dependent on the required communication between the processors. High communication throughput is necessary. This throughput is dependent on the communication medium, communication protocol, length of communication link, number of participating network components and the power and capacity of the end systems.

Unfortunately there is a great difference between internal communication throughput in massively parallel systems and communication throughput to external systems. New SGI/CRAY systems support up to 2048 processor elements with a bidirectional internal communication bandwidth of up to 500 MB/s going into 3 dimensions (3D torus). Traditional communication media currently do not support such high communication bandwidths. Here we see HiPPI 800 Mb/s - 1600 Mb/s, ATM 622 Mb/s – 2.5 Gb/s and in the near future Gigabit-Ethernet with 1 Gb/s. Most times this communication media are not supported by the supercomputer systems or only announced for future releases.

Although supercomputer systems have been designed mainly to compute and solve large problems and not to communicate with other systems there are some classes of applications which may have a benefit using a set of parallel systems.

These applications:

- have high compute power requirements,
- are highly parallel , i.e. consist of many independent processes,
- have low need of communication requirement between this processes
- have a need of symmetrical external communication between subprocesses and
- must use large blocks of data traffic to minimize communication overhead and interrupt rate.

CRAY Systems at Research Center Jülich

Three parallel vector processor systems (PVP's), one CRAY T90 and two CRAY J90, two massively parallel systems, CRAY T3E-512 and CRAY T3E-256 are currently installed at the Research Center Jülich and operated by the Central Institute for Applied Mathematics (ZAM). The CRAY T90 with IEEE floating point arithmetic has 10 cpu's installed with an overall peak performance of 18 GFLOPS and 8 Gbyte (1024 mega words) main storage.

A CRAY J90 system is used as a compute server for interactive and batch processing. This system is equipped with 16 cpu's and 8 Gbyte of main memory. The second system is used as file server for all CRAY systems within the research center.

The massively parallel CRAY T3E systems are equipped with DEC Alpha processors. The CRAY-256 installed with 256 Compute nodes (450 MHz) and 256 Compute nodes (600 MHz) leads to an overall peak performance of 530 Gigaflops. Distributed memory of 32 GBytes (128 MB per node) plus 128 GBytes (512 MB per node) is available. The CRAY T3E-512 delivers up to 300 Gigaflops and has 64 Gbytes (128 MB per node) of distributed memory.

The processing elements (PE's) of the T3E system are connected by an internal network designed as a 3 dimensional torus. Application PE's (APP-PE's) can be used for parallel programs. Command PE's (CMD PE's) are used for UNIX functionality, compilation and program development whereas OS PE's are reserved for server processes of the distributed operating system. The CRAY complex can be used by scientists of the Research Center Jülich, by users of the NIC projects and by cooperation partners from industry.

The actual configuration of the CRAY systems at Research Center Jülich (number of processors, main storage, disc capacity, etc.) can be found at http://www.fz-juelich.de/zam/CompServ/services/sco.html.

CRAY networking environment

At time of installation it has been considered to configure the five CRAY systems into one single GigaRing as soon as this will be supported by CRAY software. Until now this solution has not been implemented because of technical considerations.

Currently an interim solution has been implemented. All CRAY systems are connected to an Essential HiPPI switch. External partners can be reached via 100 Mb/s FDDI. Some 155 Mb/s ATM interface cards have been installed. These interface cards have been thought of as beta testing external high speed communication of the CRAY systems. In 1996 SGI/CRAY had planned to support 622 Mb/s ATM and in future also 2.5 Gb/s ATM interface cards.

Until now no real solutions for native ATM 622 Mb/s interfaces have been offered. It appears that no high speed ATM interfaces will be available in future. Therefore in mid of 1997 ZAM started a project to find alternatives for external high speed communication possibilities for CRAY GigaRing systems. This solutions are also of high interest for the future connection of the supercomputer centers in Germany [li05].

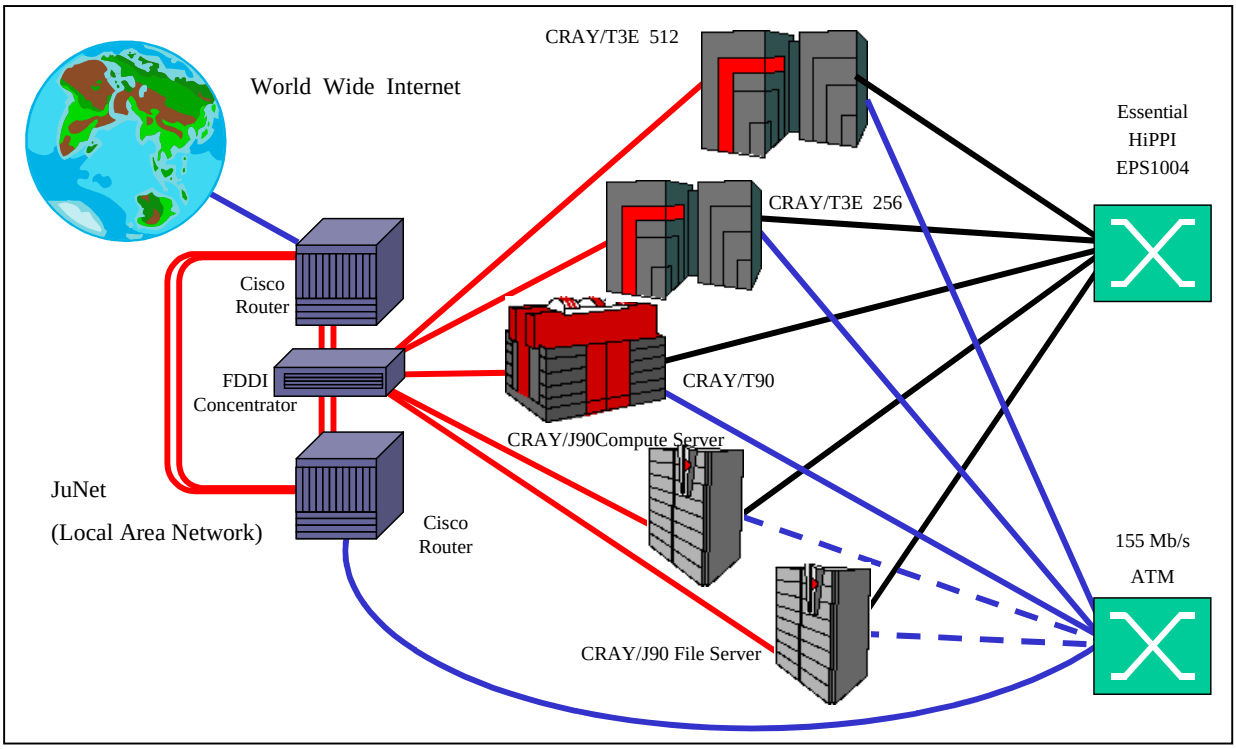

Fig. 1: CRAY supercomputer complex at Research Centwer Jülich

Further projects have been started.

The goal of the project UNICORE [li06] granted by BMBF is to deliver software that allows users to submit jobs to remote high performance computing resources without having to learn details of the target operating system, data storage conventions and techniques, or

administrative policies and procedures at the target site. Existing Web-based technologies will be exploited wherever possible. The user interface will be based on Java and modern browser technology to allow access to UNICORE resources from anywhere in the Internet for properly authorized users and eliminate software distribution.

The project HPCM (High Performance Computer Management) [li07] develops software for remote access to supercomputer resources using Java and WWW technologies.

The Gigabit Testbed West

High speed communication in a WAN environment and scientific applications requiring such communication lines are main point of interest within the project *"Gigabit-Testbed West"* [li08] sponsored by the German Bundesministerium für Bildung und Forschung (BMBF) and the Deutsches Forschungsnetz - German Research Network (DFN). Partners of the project are the Research Center Jülich, GMD - National Research Center for Information Technology, Deutsches Klimarechenzentrum, Alfred Wegner Institute for Polar and Marine Research, Pallas GmbH und o.tel.o communications GmbH. Goal of investigation is the coupling of architectural different computer systems leading to a new kind of metacomputer. The project has started in August 1997 and will end in January 2000.

The following sub-projects have been defined :

- GIGAnet - Configuration, Management and Performance Analysis of the Gigabit Testbed
- Methods and Tools, Software Support
- Solute Transport in Ground Water
- Algorithmic Analysis of Magnetoenzephalography Data
- Complex Visualization over a Gigabit-WAN
- Multimedia applications in a Gigabit-WAN
- Distributed computation of climate- and weather models
- Porting Parallel and Distributed Applications from CEC CISPAR Project

A 622 Mb/s ATM communication line (SDH, STM4 - Synchronous Transfer Mode) between the Reseach Center Jülich and the GMD installed by o.tel.o was operational just a few days after the installation. Two Fore ASX1000 ATM switches have been used as communication devices to connect the two installations over the 70 miles link.

In July 1998 the link has been upgraded to 2.5 Gb/s. Two new Fore ASX4000 switches have been installed which support this high speed communication bandwidth.

Beneath the installation, test, management and performance analysis of the communication lines there is one task of the sub-project *Giganet* to realize the connection of the supercomputer systems.

The CRAY/T3E systems at the Research Center Jülich GmbH have been connected to the Gigabit-Testbed-West network via FORE 155 Mb/s ATM interface cards. PVC connections [li09] *"everyone to everyone"* had to be configured because SVC connections for CRAY systems are currently not implemented. This introduced a considerable configuration overhead because n connections lead to n×(n-1) configuration entries. Because of this overhead only some test entries have been configured.

An IBM/SP2 with 9 ATM 155 Mb/s interfaces cards has been connected to the Gigabit-Testbed-West at GMD Sankt Augustin.

All other systems, mainly SUN and DEC systems, have been installed and configured with SVC connections.

A privat class C network 192.168.110 has been used to connect the systems to the IP network.

Main point of interest to network administration are data streams and communication profiles expected to appear by the Gigabit-Testbed-West applications.

Most of the commonly used network based applications do not need high speed data rates. Exceptions are mainly archiving, backup, NFS, FTP, Video Codec, Digital HDTV, MPEG, framebuffer and metacomputing with e.g. memory to memory transfers [li10] which normally will only locally be used.

This limitation to local communication is not the fact at the Gigabit-Testbed-West. The intent was to investigate if these applications can be used on high speed WAN networks. This implies also high I/O performance at the supercomputer systems.

Super Computer Communication with CRAY systems

One important goal of the sub-project GIGAnet was to establish a high speed connection between the supercomputer systems at Research Center Jülich and at GMD Sankt Augustin.

A set of tests has been done within the Gigabit-Testbed-West to find out about weaknesses, bottlenecks and optimizations concerning supercomputer communications. Tests have been made in a dedicated environment and in production environment. The test programs used have been:

Program	**Client / Server**	**Source**	**Communication protocol**
tcpspray	tcpspray / Echo-Daemon bzw. Discard-Daemon	Greg Christy gmc@quotron.com	TCP oder UDP
sockbench	sockclient / sockserver	Research Center Jülich	TCP
hippibench	hippiclient / hippiserver	Research Center Jülich	RawHiPPI
netperf	netperf / netserver	HP	TCP oder UDP

Fig. 2: Test programs used

First tests using the public-domain program *tcpspray* led to communication throughput for CRAY/T3E systems via HiPPI of 50-60 Mb/s at a maximum in production environment.

With the test program *sockbench* developed at Research Center Jülich throughput values of up to 61.5 Mb/s via ATM 155 Mb/s could be reached. A startup time of approximately 2 msec has been measured.

```
$ sockclient 134.94.2.29
argc: 2, *argv: 134.94.2.29
Connecting to server; 134.94.2.29
Client socket sucsessfully opened.
snd/rcv:      2 bytes   ttime:   1.920  msec ->      8.138 Kb/sec   err:0
snd/rcv:      4 bytes   ttime:   1.895  msec ->     16.491 Kb/sec   err:0
snd/rcv:      8 bytes   ttime:   1.902  msec ->     32.864 Kb/sec   err:0
snd/rcv:     16 bytes   ttime:   1.902  msec ->     65.714 Kb/sec   err:0
snd/rcv:     32 bytes   ttime:   1.897  msec ->    131.756 Kb/sec   err:0
snd/rcv:     64 bytes   ttime:   1.903  msec ->    262.788 Kb/sec   err:0
snd/rcv:    128 bytes   ttime:   1.970  msec ->    507.503 Kb/sec   err:0
snd/rcv:    256 bytes   ttime:   2.000  msec ->    999.760 Kb/sec   err:0
snd/rcv:    512 bytes   ttime:   1.947  msec ->   2054.870 Kb/sec   err:0
snd/rcv:   1024 bytes   ttime:   2.039  msec ->   3923.925 Kb/sec   err:0
snd/rcv:   2048 bytes   ttime:   2.150  msec ->   7443.349 Kb/sec   err:0
snd/rcv:   4096 bytes   ttime:   2.273  msec ->  14076.546 Kb/sec   err:0
snd/rcv:   8192 bytes   ttime:   2.725  msec ->  23485.075 Kb/sec   err:0
snd/rcv:  16384 bytes   ttime:   4.338  msec ->  29510.018 Kb/sec   err:0
snd/rcv:  32768 bytes   ttime:   5.570  msec ->  45959.100 Kb/sec   err:0
snd/rcv:  65536 bytes   ttime:   8.932  msec ->  57323.914 Kb/sec   err:0
snd/rcv: 131072 bytes   ttime:  17.798  msec ->  57533.358 Kb/sec   err:0
snd/rcv: 262144 bytes   ttime:  33.328  msec ->  61450.533 Kb/sec   err:0
snd/rcv: 524288 bytes   ttime:  67.129  msec ->  61017.025 Kb/sec   err:0
snd/rcv: 1048576 bytes  ttime: 150.482  msec ->  54438.303 Kb/sec   err:0
```

Fig. 3: First Tests: CRAY/T3E 155 Mb/s ATM communication with program sockbench

Special tests with the program *sockbench* between the two CRAY/T3E systems have shown that throughput rates of up to 430 Mb/s could be measured using TCP/IP via HiPPI technology (nominal 800 Mb/s) [li11]. Throughput values within production environment vary extremely depending on system usage. The program *sockbench* has already been optimized by enlargement of SocketBufferSize, use of TCP-WinShift option and use of TCP-Nodelay.

Sending data with a special program *hippibench (hippiclient und hippiserver)* adapted from *sockbench* and using only Raw-HiPPI protocol [li12, li13, li14, li15] (no TCP/IP) 530 Mb/s can be reached.

```
Testprogramme: hippibench: hippiclient und hippiserver
Transfers von zam003 (t3e256) nach zam006 (t3e512)

----------------------------Starting loop-------------

snd/rcv:      8 bytes   ttime:   3.493  msec ->     17.891 Kb/sec   err:0
snd/rcv:     16 bytes   ttime:   3.048  msec ->     41.005 Kb/sec   err:0
snd/rcv:     32 bytes   ttime:   3.039  msec ->     82.253 Kb/sec   err:0
snd/rcv:     64 bytes   ttime:   3.117  msec ->    160.419 Kb/sec   err:0
snd/rcv:    128 bytes   ttime:   3.062  msec ->    326.540 Kb/sec   err:0
snd/rcv:    256 bytes   ttime:   3.049  msec ->    656.010 Kb/sec   err:0
snd/rcv:    512 bytes   ttime:   3.237  msec ->   1235.733 Kb/sec   err:0
snd/rcv:   1024 bytes   ttime:   3.301  msec ->   2423.534 Kb/sec   err:0
snd/rcv:   2048 bytes   ttime:   3.078  msec ->   5198.535 Kb/sec   err:0
snd/rcv:   4096 bytes   ttime:   3.074  msec ->  10409.466 Kb/sec   err:0
snd/rcv:   8192 bytes   ttime:   3.113  msec ->  20555.909 Kb/sec   err:0
snd/rcv:  16384 bytes   ttime:   3.354  msec ->  38161.510 Kb/sec   err:0
snd/rcv:  32768 bytes   ttime:   3.841  msec ->  66651.479 Kb/sec   err:0
snd/rcv:  65536 bytes   ttime:   4.850  msec -> 105577.786 Kb/sec   err:0
snd/rcv: 131072 bytes   ttime:   6.083  msec -> 168325.954 Kb/sec   err:0
snd/rcv: 262144 bytes   ttime:   7.965  msec -> 257111.041 Kb/sec   err:0
snd/rcv: 524288 bytes   ttime:  11.767  msec -> 348103.364 Kb/sec   err:0
snd/rcv: 1048576 bytes  ttime:  19.045  msec -> 430131.691 Kb/sec   err:0
snd/rcv: 2097152 bytes  ttime:  33.752  msec -> 485417.980 Kb/sec   err:0
snd/rcv: 4194304 bytes  ttime:  64.641  msec -> 506919.400 Kb/sec   err:0
snd/rcv: 8388608 bytes  ttime: 122.915  msec -> 533181.987 Kb/sec   err:0
```

Fig. 4: CRAY/T3E Raw-HiPPI communication with program hippibench

SGI corporation has stated that the low transmission speed is related to a low interrupt rate within the T3E systems. Communicating via ATM (MTU 9180 Byte, (Maximum Transmission Unit)) 15000 packets per second have to be processed using a bidirectional 622 Mb/s ATM interface. The following table shows the interrupt rates measured by SGI with simple test programs for GigaRing systems CRAY/T90, CRAY/J90 and CRAY/T3E related to this packet size.

System	Packets/s = Interrupts/s	Percent	Throughput effective	Throughput nominal
CRAY/T90	12988	85 %	529 Mb/s	622 Mb/s
CRAY/J90	4216	27 %	172 Mb/s	622 Mb/s
CRAY/T3E	2786	18 %	113 Mb/s	622 Mb/s

Fig. 5: Throughput of CRAY GigaRing systems (9180 Byte packets)

This interrupt rate implies that higher data throughput can only be reached if data transmission is done using large blocks. This can only be realized using large MTU values as this is possible e.g. with the HiPPI technology (65280 Byte). Furthermore a number of operating system parameters have to be optimized.

When connecting CRAY systems to the Gigabit-Testbed West different configurations are possible.

As a first approach the HiPPI data stream can be tunneled by an Ascend GRF Router [li16]. This type of router supports HiPPI as well as 622 Mb/s ATM interface cards. Here both end systems do not see the ATM link on the communication path.

A major disadvantage is that the tunnel protocol requires to configure a MTU of 9180 Byte. This implies that the large MTU size of the HiPPI protocol can not be utilized. Furthermore experiences at other sides have shown that an optimal support of this product by the manufacturer could not be guaranteed until now.

A second alternative is to extend the HiPPI connection which is limited to 25 m by standard specification via a HiPPI-Sonet extender available e.g. from Essential Communications Corp. However this solution would require an explicitely Sonet connection between the two installations. This would require to install an additional link or to divide the 2.5 Gb/s connection into 4 * 622 Mb/s. This configuration has not been possible in the scope of the project.

Another alternative is to install an IP router or a gateway system using HiPPI technology. This solution has been analyzed in detail in the Gigabit-Testbed-West.

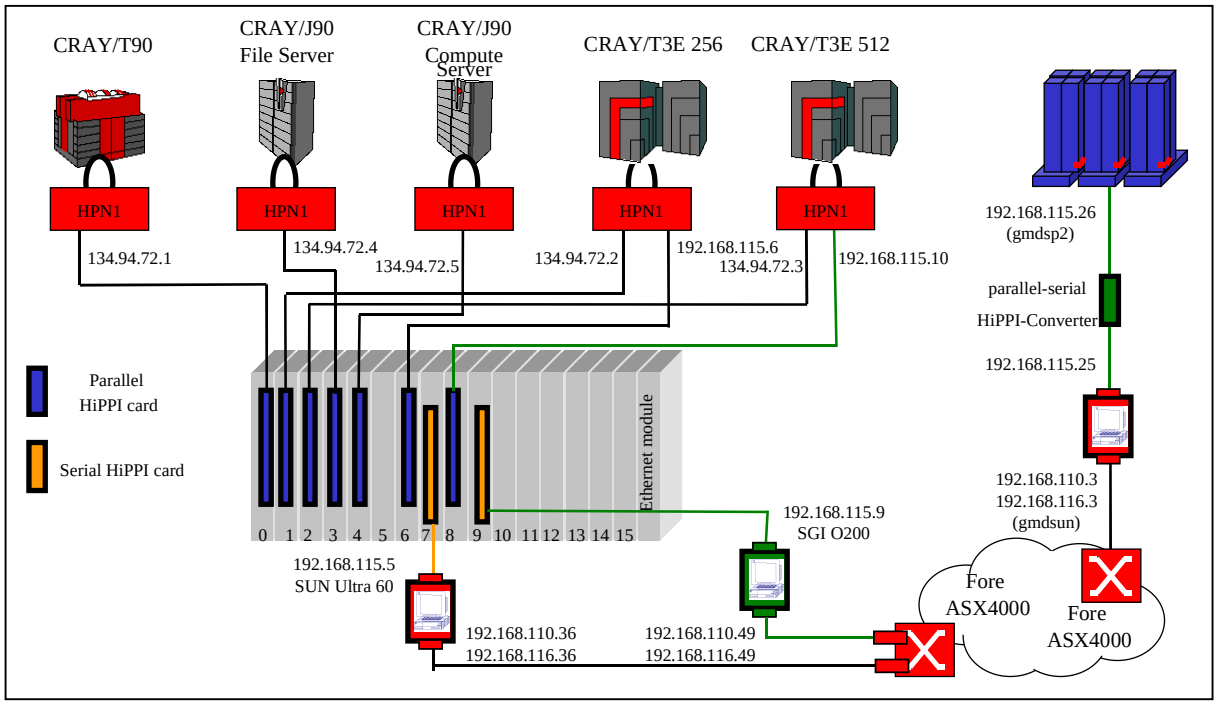

Fig. 6: CRAY HiPPI Testbed Konfiguration

As mentioned above in 1996 GigaRing communication between different CRAY systems has not been supported. Therefore the five CRAY systems at the Research Center Jülich have been coupled via an Essential HiPPI switch. The CRAY internal communication node has been connected via a GigaRing to an HPN1 node. This sort of node supports two independent

HiPPI interfaces each with 800 Mb/s throughput capacity. The second HiPPI link has been used to connect the two T3E systems to the Gigabit-Testbed-West.

In the beginning of 1998 a SUN Ultra 60 workstation has been installed in the Research Center Jülich as a HiPPI-to-ATM gateway using this second HiPPI link. At the GMD a SUN Enterprise 5000 system has been installed as counterpart. Since mid of 1998 a SGI O200 system has been configured in a beta test environment to the second T3E system to compare the SGI and SUN high speed gateway solutions.

Further tests have to be done to see if dedicated communication nodes and dedicated GigaRing connections will lead to improved throughput capacity.

A similar installation with HiPPI-to-ATM gateway at the GMD prevented from hardware and software configuration overhead at ATM and IP level. Otherwise 9 ATM interfaces with 155 Mb/s would have been to be configured. Routing decisions would have been to be made at IP or MPI application level. Because of the similar configuration on both sides of the communication path a main advantage is the possibility to use a MTU size of 65280 byte on all communication links instead of the standard size of 9180 byte for ATM connections. Using the Path MTU Discovery mechanism at the end systems large packet lengths can be used which is of great benefit at the T3E systems. Herewith higher data transfer rates can be reached.

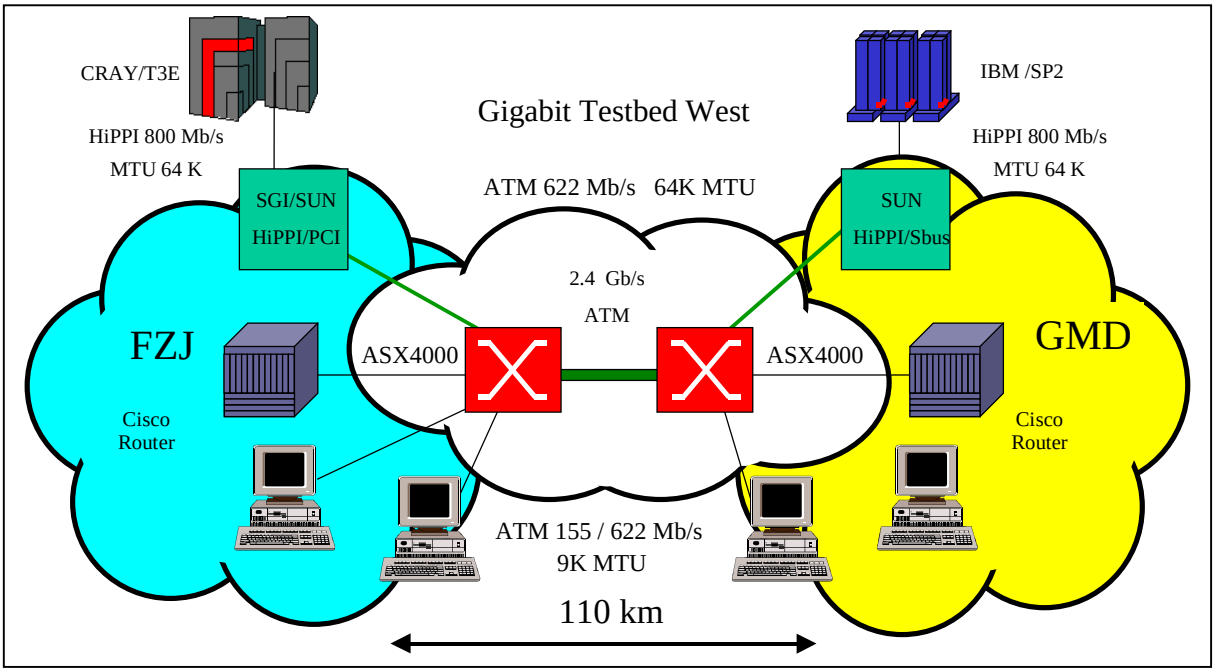

Fig. 7: Metacomputer configuration at Gigabit-Testbed-West

The figure above shows the configuration established at the Gigabit-Testbed-West between the CRAY/T3E at the Research Center Jülich and the IBM/SP2 at GMD. As a local testbed the connection between the two T3E systems via the Sun Ultra 60 and SGI O200 gateway

systems could be used (see Fig. 6). Within this test environment the communication throughput will not be effected by the minimal link length (lower than 25 m) and minimal delay because of speed of light (100 km conforms to 0.5 ms delay).

Throughput tests have shown that the performance supported by the implemented HiPPI-to-ATM gateways will be sufficient to connect the CRAY/T3E systems to a high speed data network. Using the HiPPI-to ATM gateways up to 400 Mb/s throughput can be reached. (622 Mb/s ATM supports up to approximately 540 Mb/s user data). To get this throughput the user applications have to transfer data with large blocksizes. (Using MTU 65280 byte, *TCP Window Shift Option* and increased *Socket Buffer Size*) [li17]. Throughput values will decrease drastically if one of this requirements will not be used. As mentioned above the interrupt rate of the T3E systems is the limiting factor. Using smaller data packets per interrupt leads to decreased throughput.

	155 Mb/s	**622 Mb/s**	**2,4 Gb/s**	**%**
Capacity	155,520	622,080	2.488,320	100
SDH Overhead	5,760	23,040	92,160	3,7
ATM Overhead	14,128	56,513	226,053	9,1
ATM User Data	135,632	542,527	2.170,107	87,2
ATM cells	353.208	1.412.830	5.651.321	
CIP TCP 9140 Overhead	1,118	4,474	17,896	0,7
CIP TCP 9140 User Data	134,513	538,053	2.152,211	86,5
CIP TCP 9140 Packets/s	1840	7.358	29.434	
LANE TCP 1460 Overhead	6,711	26,844	107,375	4,3
LANE TCP 1460 User Data	128,921	515,683	2.062,732	82,9
LANE TCP 1460 Packets/s	11.038	44.151	176.604	

Values in Mbit/s

Fig. 8: Protocol overhead in ATM and TCP/IP networks

Figure 9 below shows the throughput values measured using the program *netperf*. The SUN and SGI systems have been used as gateways systems as well as end systems in the ATM network. The throughput measured connecting the T3E systems directly via the HiPPI switch is signed with ***direct.*** The throughput values signed with ***gate*** have been measured using the path CRAY ⇒ HiPPI ⇒ SUN ⇒ ATM ⇒ SGI ⇒ HiPPI ⇒ CRAY (and vice versa).

If the SUN and SGI systems are used as end systems within the ATM network with a MTU size of 9180 then this MTU size will be used for the whole transaction because of the Path MTU Discovery mechanism [li18] resulting into a 6 times interrupt rate. Therefore a decreased performance is observed.

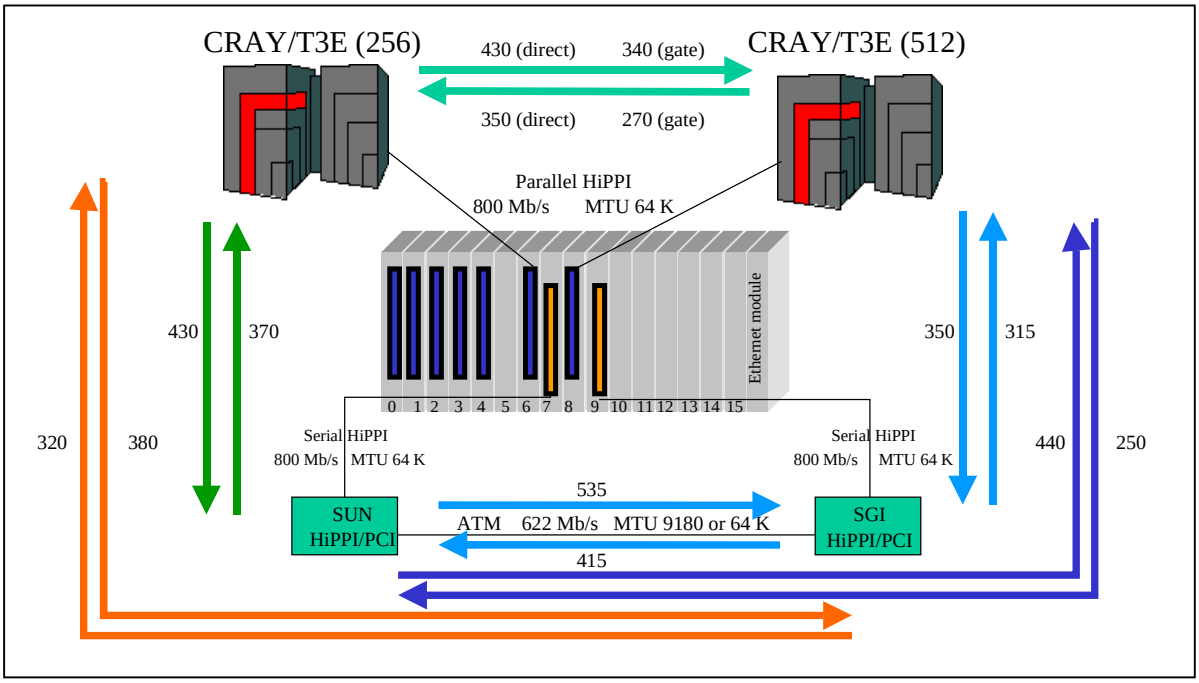

Fig. 9: Throughput rates using HiPPI-to-ATM gateways

Further Investigations

The investigations done until now have shown that high speed supercomputer communications are not only dependent on the interfaces used but also on networking and system configuration issues. This design considerations at the end systems as well as the gateway systems include as mentioned above: choise of network media and configuration, MTU, Path MTU Discovery, SocketBufferSize, WindowScalingOption.

Furthermore hardware configuration details of the involved systems are relevant. Using a HiPPI-to-ATM gateway with two internal PCI busses leads to enhanced throughput if the ATM and HiPPI interface cards have their own PCI bus. Furthermore interface specific options are of great importance (64 bit, 66 MHz or 32 bit, 33 MHz). The FORE HE622 ATM interface card (622 Mb/s) supports both slots and therefore can use the faster one whereas the Essential HiPPI interface card only can be put into the slower slot.

Further investigations concerning the supercomputer systems are necessary. The HiPPI-to-ATM gateway connection can be configured to use a dedicated or a shared HPN. Using a dedicated HPN a dedicated or shared OS/PE (Operating System/ Processor Element) within the T3E can be used. Moreover the shared usage of a OS/PE with other interfaces like ATM, FDDI and Ethernet can be configured. Tests using separate OS/PE's lead to better communication throughput. Furthermore the position of the Communications PE's within the T3E torus may effect the communication throughput because external and internal

communication flows may hinder one another (communication , I/O, ...) on some internal 3D torus specific links.
Last but not least tests have to be set up using a single HiPPI-to-ATM gateway system as shared gateway for multiple CRAY systems. If all systems want to communicate at the same time the ATM 622 Mb/s will be a bottleneck. If there are more bottenecks e.g. queue length, mbufs, paging problems within the gateway systems will be of great interest. Using this sort of configuration should lead to no problems if the high speed communication link is used only alternately and temporarily. This would imply a cost effective solution.

Currently communication between both Gigabit Testbed West locations can be done with a maximum of 2.5 Gb/s communication throughput. Future considerations have to be done how this bandwidth can be utilized by supercomputer systems. Until now only communication bandwidths of 800 Mb/s (HiPPI at CRAY and IBM/SP2) have been considered.

State of the art routers do not support 1.6 Gb/s HiPPI. Super-HiPPI [li19] with 6,4 Gb/s has just been developed and standardized. 2,4 Gb/s interface cards for computers and supercomputers are not available today. To get higher throughput values multiple parallel pathes have to be used. Metacomputing applications which need this high bandwidth have to address this interfaces explicitely or have to use a software interface which has been incorporated into the operating system or an intermediate software level (PACX, MPI, ...).

There are many configurations possible for a final scenario. Multiple HiPPI and/or ATM interfaces with or without gateway systems are possible. Major problems arise finding the right communication node within the local system to communicate with nodes on remote systems and which remote communication node to contact to reach a special remote node. Until now no mechanisms are available to solve this routing decisions in an optimal way.

An optimal configuration for all metacomputing applications can not be determined. This is dependent on the kind of application and the communication profile (length and number of messages).

Summary

The investigations done at the Research Center Jülich within the Gigabit-Testbed-West describe a snapshot how supercomputer systems currently can be used within a metacomputing environment. Today many special configurations have to be considered to get reasonable throughput values.

The Giganet sub-project of the Gigabit-Testbed-West has shown that gigabit communication can become reality today. Workstation systems can communicate with testprograms at gigabit throughput. Real applications needing this throughput are currently developed.

Metacomputing becomes reality in LAN as well as WAN environments. So supercomputer system developers have to realize that:

"The net is the computer and the computer is the net".

Today it must be summarized:

(SuperComputer)Communications != Super(ComputerCommunications)

Literatur

[li01] F.Hoßfeld, B.Mertens - Grand Challenges - Höchstleistungen mit Supercomputern, HGF Jahresheft 1997, 1996, S. 11-12

[li02] B.Mertens, Das neue Supercomputersystem der KFA: Chancen und Herausforderungen, Internal report of Forschungszentrums Jülich, KFA-ZAM-IB-9621, Sep. 1996

[li03] F.Hoßfeld, Der "Digitale Windkanal": Herausforderung der Computersimulation komplexer Systeme, Proceedings 14.ITG/GI-Fachtagung "Architektur von Rechensystemen", Workshop 4: Optimierung in Physik und Informatik - Konzepte und Realisierungen, pp. 99-110, Sep. 1997 Rostock

[li05] *Hoßfeld, Friedel; Nagel, Wolfgang E. (1997)* - Verbund der Supercomputer-Zentren in Deutschland - eine Machbarkeitsanalyse, Bericht der Arbeitsgruppe zur Erstellung der VESUZ-Machbarkeitsanalyse, BMBF-Förderkennzeichen O1 IR 602/9, Oktober 1997

[li06] V.Sander, High Performance Computer Management, Workshop Hypercomputing, 14. ITG/GI-Fachtagung von Rechensystemen, ARCS '97, Rostock, Sept.. 1997

[li07] J.Almond, M.Romberg, The UNICORE Project: Uniform Access Supercomoputing over the Web, Proceedings of Cray User Group Meeting, Stuttgart, June 1998

[li08] Th.Eickermann, P.Wunderling, R.Völpel, R.Niederberger, Aufbruch ins Jahr 2000 - Start des Gigabit Testbed West, DFN-Mitteilungen, Heft 49, pp. 13-15, Nov. 1997

[li09] J.Halpern, M.Laubach, Classical IP and ARP over ATM, RFC 2225, April 1998

[li10] P.W.Haas, High - Performance Communications in Large MTU Networks

[li11] J.Renwick, IP over HiPPI, RFC 1374, Jan. 1997

[li12] HiPPI - Mechanical, Electrical and Signalling Protocol Specification (HiPPI-PH), ANSI X3T9.3/92- Rev 8.2

[li13] HiPPI - Physical Switch Control (HiPPI-SC), ANSI X3.222-1993

[li14] HiPPI - Framing Protocol (HiPPI-FP), ANSI X3.210-1992

[li15] HiPPI - Link Encapsulation (HiPPI-LE), ANSI X3.218

[li16] HiPPI - Mapping to ATM with Annex A HiPPI to ATM IP Router, ANSI X3T11/Project 1026/Rev 1.6

[li17] D.Borman, R.Braden, V.Jacobson, TCP Extensions for High Performance, RFC1323, 1992

[li18] J.Mogul, S,Deerung, Path MTU Discovery, RFC 1191, Nov. 1990

[li19] D.Tolmie - Project 1249-D - HiPPI-6400 Mbit/s Optical Specification, http://www.t11.org, Los Alamos National Lab, September 1998