ORGANISATION EUROPEAN POUR LA RECHERCHE NUCLEAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN-IT-2003 - 02
12 May 2003

Macro-Clusters, a Way to Adapt New Technologies in Existing Large Computing Clusters

Arie Van Praag, Mathias Gug, Andras Horvath, Fabian Collin,
CERN, division IT/ADC/LE
12 Mai 2003
http://hsi.web.cern.ch/HSI/HSI/people/Arie/AboutMacroClusters.pdf

Abstract:
A series of new interconnect technologies are emerging on the market, HyperTransport PCI-Express, Infiniband and 10 Gigabit Ethernet. Some are very interesting as cluster interconnects and some other not at all. A description is made of methods to use this cluster oriented interconnects in different ways. In order to adapt easily to existing clusters the term of Macro-Clusters is introduced to define small sub-clusters with a defined function that can be used independent or coupled to large cluster networks.
Some notice is given on how to interconnect hardware and software of the different technologies and some remarks about cost performance relations are included. In the Appendix are mentioned test results for bandwidth and connections to the CERN storage system.

INDEX

Macro-Clusters, a Way to Adapt New Technologies in Existing Large Computing Clusters

Arie Van Praag, Mathias Gug, Andras Horvath, Fabian Collin,
CERN, division IT/ADC/LE
12 Mai 2003

Abstract:
A series of new interconnect technologies are emerging on the market, HyperTransport PCI-Express, Infiniband and 10 Gigabit Ethernet. Some are very interesting as cluster interconnects and some other not at all. A description is made of methods to use this cluster oriented interconnects in different ways. In order to adapt easily to existing clusters the term of Macro-Clusters is introduced to define small sub-clusters with a defined function that can be used independent or coupled to large cluster networks.
Some notice is given on how to interconnect hardware and software of the different technologies and some remarks about cost performance relations are included. In the Appendix are mentioned test results for bandwidth and connections to the CERN storage system.

Some definitions:

The definitions of different computing clusters used in this description are as follows:

- **Micro-Clusters:** This are computing engines with an architecture that can not be influenced after manufacturing, but are defined either by multiple processors on a single motherboard or by multiple processors on a single semiconductor chip. Micro-clusters are mostly found between 2 and 8 nodes on motherboard designs and up to 512 and more in monolithic designs.

- **Macro Clusters:** A small computing cluster with an architecture that contains several individual computing engines that are coupled by an interconnecting medium to form a cluster that can be used either individually or in large computer clusters. Macro clusters can have specialized functions such as storage number crunching or I/O and networking.

- **Meta Clusters** is a term introduced by Utah University and stays for functional groups as *Number Cruncher group, Data Base Group* , *Admin & Control Group* or *others*. They can be part of a single large cluster or consist of a number of physically separated clusters under a single administrator function.

- **Computing Clusters** or simply **Clusters:** The overall architecture of a large cluster.

Introduction:

With a number of new interconnect technologies coming on the market it is of interest to have a look on what kind of modifications to exiting cluster architecture can be made to gain in performance pure, price performance, relation flexibility, scalability and management. By introducing the term of "Macro Clusters" as a definition of small cluster structures that can be used to build large computing architectures. This macro clusters with sizes of 4, 8 or 16 nodes can be pure computing engines or can have specialized functions coupled to theme for Storage or I/O and Networking. Each macro cluster should also be self sufficient for management and reporting to a larger management system, eventually coupled with an auto-reset system. In many cases a self supporting Blade System or rack architecture can already be seen as a kind of Macro-Cluster. The term Macro-Clusters and their closed/open architecture should make it easy to upgrade stepwise to new, and at first view only marginally compatible, technologies.

The Old Interconnects:

Recently a number of old interconnects with extended bandwidth or new interconnects come to the market place with higher bandwidth and facilities to minimize latency. They can be classified as follows:

Ethernet:
Introduced 10 Gigabit Ethernet, with an actual user bandwidth of 10 Gigabit. The interface is optical only and divided in short and long-distance versions, demanding expensive lasers. Discussions about a copper connection are going on without results up to now. The IEEE Ethernet specifications keeps the 1500 byte data block as the only MTU allowed. With as result that processing power and I/O and Memory bandwidth limits the useful bandwidth in standard mode to 150 – 200 MByte/s. Except for local backbone connections to 10 GE specification foresees to couple directly to SONET OC 192 for long distance connections.

Fibre Channel:
As in the past Fibre Channel has shown to be successful in the networked storage field for NAS (Network Attached Storage) and SAN (Storage Attached Networks). High quality disks and RAID arrays are both upgrading to the 2 Gbit/s variant. Fibre Channel announces FC4G, backward compatible and FC 10G not backward compatible for Summer 2004. Attempts have been made in the past to use Fibre Channel as a universal interconnect, but the large number of protocol options and bandwidth variants have rapidly discouraged people to continue this path.
Some of new FC 2G Raid arrays use more economic IDE disks inside. Fibre Channel equipped storage is still expensive and the need for switches in large networked storage arrays increases the cost performance relation even more. The increase in bandwidth does not correspond with an increase in Disk speed but more disks can be put on an arbitrated loop.

Myrinet:
Myrinet has upgraded to 2 Gbit/s and announces 10 Gbit/s bandwidth. As in the past it is a relatively cheap network made for low latency that excels in combination with MPI.

The New Interconnects:

Old Interconnects							
Interconnect	**Name**	**Bandwidth Gbit/s**	**Media**	**Protocols**	**Storage**	**Price**	**Available**
Ethernet	**10 GE**	**10**	**Fibre**	**TCP/IP, UDP**	**iSCSI**	**Very high**	**Yes**
Fibre-Channel	**FC 2G FC 10 G**	**2 10**	**Fibre, Copper**	**FC, TCP/IP, SCSI**	**FC-SCSI**	**High**	**Yes No**
Myrinet	**M 2G M 10G**	**2 10**	**Copper**	**Appropriate, TCP/IP, MPI**	**---**	**Low**	**Yes No**
New Interconnects							
GSN		**10**	**Fibre, Copper**	**ST, SCSI-ST, TCP/IP**	**SCSI-ST**	**High**	**Not anymore**
Infiniband	**IB X1 IB X4 IB X12**	**2.5 10 30**	**Copper, Fibre**	**IB, TCP/IP, MPI, foreseen SCSI**	**Foreseen native**	**Start prices as GE**	**Yes Yes No**
HyperTransport	**---**	**12.8 GByte/s**	**Copper**	**Appropriate**	**??**	**Low**	**3Q 2003**
PCI-Express		**2.5 / channel**	**Copper**	**??**	**??**	**Low**	**3Q 2004**

Table 1: Overview of 10 Gbit/s interconnects

GSN:
The Gigabyte System Network, also known, as ANSI standard HIPP-6400 was the first one of the 10 Gbit/s networks developed together with a low latency protocol ST. If technically a success able to move data at 760 MByte/s with 5% processor use, the system is not a commercial success.

Infiniband:
After a long introductory time Infiniband is arriving on the market now. IB X1 with 2.5 Gbit bandwidth shows to be obsolete already and switches and interfaces start immediately with the 10 Gbit/s IB X4. With a theoretical bandwidth of about 1 GByte/s, bandwidth of up to 750 MByte/s are attained with standard Pentium 4 equipment and latency can be as low as 7 usec. Infiniband can be used with TCP/IP4 and TCP/IP6 but handling those protocols will influence the bandwidth negatively, mainly due to processor limits in handling the stacks. For this reasons Infiniband uses its own protocols, including RDMA and Secure Transfer. ANSI T10 is working on a standard for SCSI over IB that will probably be accepted during 3Q 2003. Infiniband shows up to have lower latency as Myrinet using MPI. The fact that IB X4 has 4 parallel streams of 2.5 Gbit/s each, eases coupling to 10 GE, using the XIAU interface that is part of the 10 GE

specifications. Which makes it possible to have one or more 10 GE channels on the larger Infiniband Switches to serve backbone connections and long distance traffic.

HyperTransport:

HyperTransport is specified by an independent industrial consortium and is used by the MIPS 9000 processor. AMD introduced HyperTransport as internal interconnect for the Opteron processor that replaces the classical chip-set It is able to interconnect multiple processors with memory and I/O channels with a peak bandwidth of up to 19.2 GByte/s. The HyperTransport specification also includes a back-panel connector such that small macro-clusters can easily be built. However using LVDS signals, cable length will be limited. The advantage of such a system is its enormous bandwidth and the large address space for 256 TByte of memory. HyperTransport is made not only as an internal bus, but also with native cluster connections in mind.

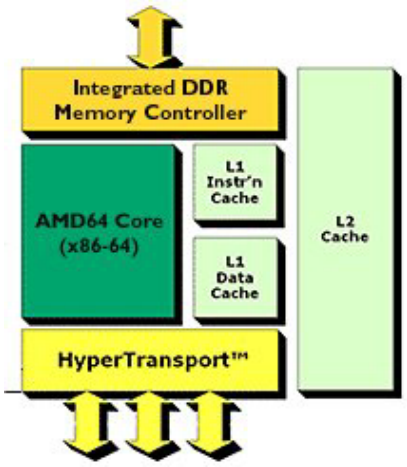


Fig 1 The AMD Opteron-HyperTransport Architecture

PCI-Express:

PCI-Express will be in parallel or replace the existing PCI type interfaces. The basic interface is a 2.5 Gbit/s serial stream. Multiple of those streams can be in parallel to obtain higher bandwidth. Early specifications suggest that PCI-Express will also be available on the back-panel for peer-to-peer connections, however recent publications talk only about an internal I/O bus to replace the parallel PCI-2 and PCI-X. Even if the PCI-Express specification gives room for 256 channels, practical values combined with available memory and I/O bandwidth will limit this number between 4 and 8 channels. Hardware wise PCI-Express is an extension to the PCI-X connector and can as such be available simultaneously.

An overview of different 10 Gbit/s interconnects that will play a roll in future systems is given in Table 1

Architectural Variants:

To get the maximum possibilities out of a macro-cluster the architecture has to be adapted to the technology chosen and the function it should fulfill in the cluster. As such a clear difference can be made between a production or number crunching macro-cluster and small- and large storage macro-clusters. But also specialized clusters for I/O or print servers are thinkable.

Properties of a macro-cluster should be that it can be used independent or in a large cluster. It should have its own management either by one of the machines in the cluster that has a kind of master function or by an independent small machine. This management

function should report to a general cluster manager, that decides about an automatic reset or excluding the node via the macro-clusters management function. As given this structure of macro-clusters to build larger clusters, a more hierarchical architecture will be the result.
This in contradiction to tightly coupled super-computers that use many processors with shared storage and shared memory. Even if the HyperTransport variant comes near to it with its coherent memory space over at least a limited number of double processor nodes.

Using HyperTransport:
HyperTransport is the only interconnect directly made to interconnect processors. For that it is part of the Northbridge where a non-blocking crossbar switch handles the communication facilities. As such it is adapted by two processor manufacturers: by MIPS and Sierra for the 9000 RISC 64 bit processor and by AMD for the 32/64 bit Opteron PC compatible processor. Part of the HyperTransport specification is a back-panel connection with separate input and output connections, such that the direct connection has a serial character (fig 2 and 3).

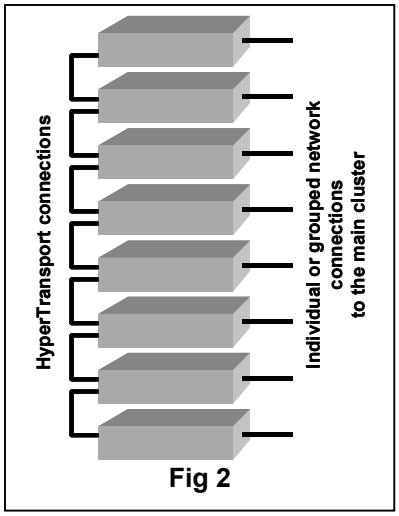


Fig 2

The direct HyperTransport interconnect is for the moment limited to 8 processors (Fig 2). In the future some silicon chip designs for HyperTransport external switches will extend beyond this limit. Connections to a larger cluster can be by one or more high throughput network connections. The bandwidth of HyperTransport is given as 19.8 MByte/s, and the connection is a 64 bit copper cable with limited length. With the separate input and output ports it should in principle be possible to construct a ring architecture. However, no word is told about this in the specifications.

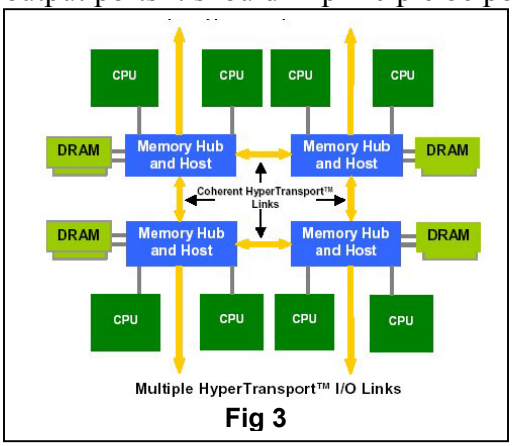


Fig 3

A big advantage of this interconnect is that the all nodes coupled to each other can access each others memory as being a single large memory space. This may not be needed in the actual CERN computer centre for data collection and storage. However it can be a big advantage for new analysis software for LHC data. And HyperTransport interfaces comes for free with the AMD Opteron processors.

If HyperTransport switches (API Networks) become available such 8 node macro-clusters can be coupled together to form larger clusters. It has to be seen if this switches allows extending the memory addressing range over more than one 8 node clusters as the total 64 bit addressing space (256 TByte) would permits (Fig 4). Or that it needs a specific transfer or messaging protocol. Once in production the switches should not be too expensive as simple and well-known level two technologies are sufficient using the

HyperTransport protocol. Using local mode for connections and a tunneling protocol for transfers, the latency should be low.
HyperTransport clusters as described here are only the basic forms. Variants such as storage clusters as described for Infiniband can be built with this kind of clusters to. However it is very questionable if direct connecting RAID boxes will ever see the market. Such for storage connection there is a PCI type interface necessary, SCSI, Gigabit Ethernet, Fibre Channel or Infiniband. With an architecture as in Fig 7 each node can even have 2 Infiniband RAID boxes without limiting the bandwidth.

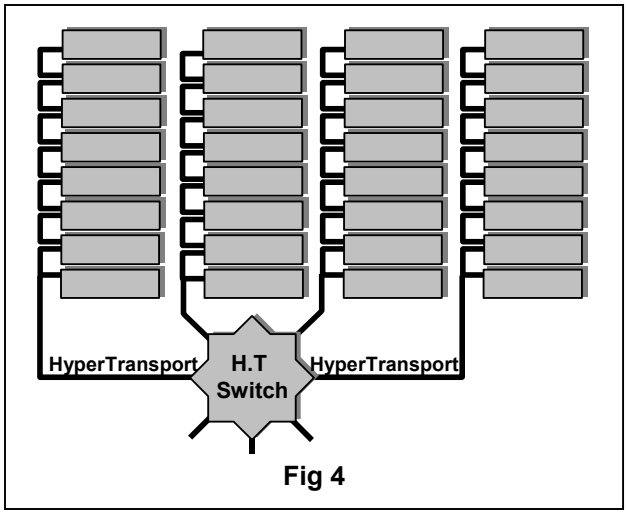


Fig 4

Using Infiniband:
As commercial Infiniband products come to the market actually it is in the IB X4 variant with 10 Gbit/s bandwidth. Interfaces cards use PCI-X and normally have two channels. The next generation of interface cards that is already announced will be capable to use either PCI-X or PCI-Express depending on what is available at the motherboard..
IB X4 switches are available in version running with 8 and 16 channels and are announced with up to 128 channels. Due to the switch silicon used the MTU is for the moment limited to 2048 Bytes. This will be corrected in the next generation of switches that allow the IB native MTU of 4096 Bytes. This does not necessarily results in higher bandwidth, but will make transfers less processor intensive.
Infiniband has foreseen the use of TCP with Ipv4 and with Ipv6, but will be far more efficient with its own native protocols. Latency, as proved with use of MPI can be as low as 7 µsec.

Already RAID arrays are available in versions that use low cost IDE disks and interface with Gigabit Ethernet or Fibre Channel 2G. As interface silicon for IB is relatively cheap compared to Fibre Channel it can be expected that, with growing popularity of IB, budget aware RAID Arrays with Native IB X4 Interfaces will come to the market soon (already at "API Storage", but expensive). Use of TCP/IP over IB brings a bandwidth penalty, but software under development allows socket service that converts IB oriented addressing to IB native protocol.

This makes that a variety of specialized macro-clusters can be build that are not as easy to realize in other technologies.

The IB computing Macro-Cluster:
A simple number crunching or computing macro-cluster is given in Fig 5. With the high bandwidth a simple connection to the switch should be sufficient, leaving the secondary interface channel free as spare for other purposes. An example is machine 1 declared for

background local management and has as such an extra connection to the group switch S2, to report to the general cluster management. As results in case of failure a single machine can be remotely reset or excluded from the cluster. If it is the management machine that fails the hole cluster would be doomed. For that reasons it is advisable to have a fallback algorithm to a second machine. Especially as the second IB connections are available. The size of a computing cluster is given by the size of the switch, but depends also on the organizational flexibility whished.

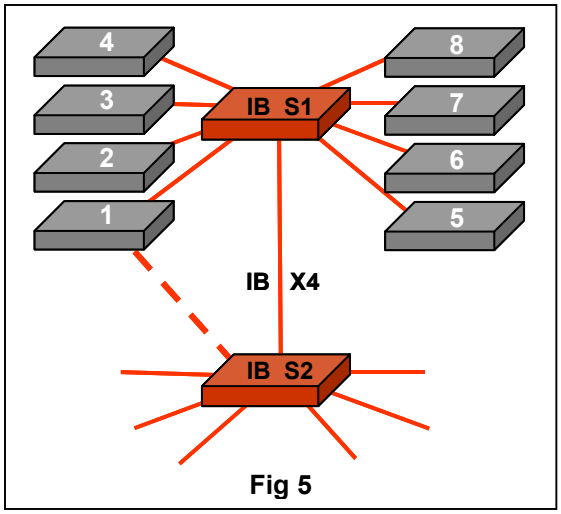


Fig 5

An IB Macro-Cluster for Background Storage:

By equipping a node with two disks, one can be used for OS and local storage, and the second one can be used as part of a storage array, using software RAID (fig 6). The storage array can be accessed by either iSCSI, but the use of TCP/IP will strongly punish the IB bandwidth. Or by SCSI over IB. In that case the disk have to be accessed locally by the SCSI to IDE converter that is part of LINUX. The storage bandwidth is depending on the workload of the hosts. If the maximum values of Serial-ATA class 1 of 150 MByte/s per disk is taken, a Raid system with 4 data disks and 1 parity disk can theoretically give total storage bandwidth as high as 600 MByte/s. No latency is calculated here for software layers for RAID and SCSI to IDE conversion. This is easily in the ranch of a single Infiniband IB X4 connection.

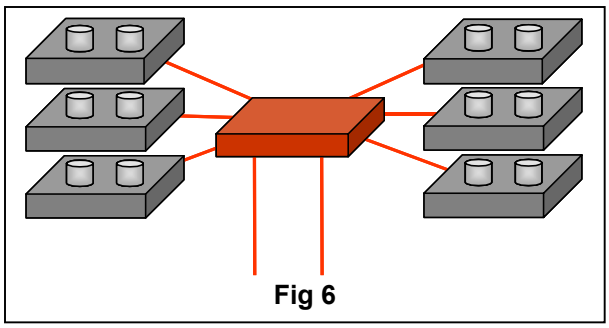


Fig 6

Most recent IDE Disks nowadays have a realistic bandwidth of around 40 MByte/s, which gives a realistic bandwidth of 4 X 40 or 160 MByte/s.

Using for the second (storage) disks 200 GByte models there is already 1 TByte of storage available and one disk as hot spare. Extension to more storage is possible by attaching more nodes. For small capacities a single disk per node with multiple partitions can be used.

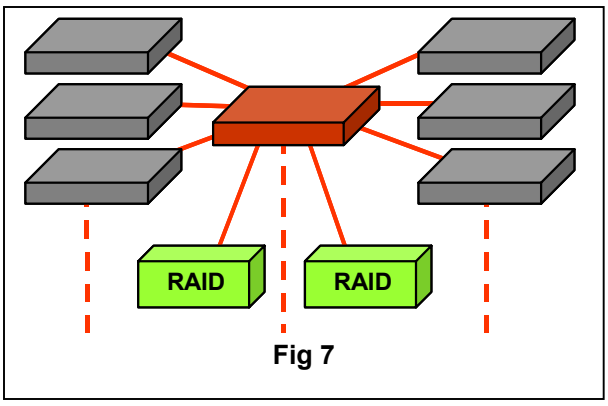


Fig 7

Macro-Clusters with Network Attached Storage:

Building with macro-clusters will be a good opportunity to move from single node disc servers to storage macro-clusters that behave like a disk server but are in reality small-networked storage devices, either SAN or NAS.

Calculating with a RAID box with 12 disks, two groups of 4 Data disks + Parity disk can be formed, leaving two hot spare disks for fall over. Theoretical bandwidth using Serial-ATA class1 disks with up to 150 MByte/s a single box can have two groups with a bandwidth of each 600 MByte/s, or a theoretical total of 2.4 GByte/s in the example of Fig7. With 200 GByte disk such a group has 1 TByte storage capacity. With two RAID boxes, the available storage is 4 TByte with a maximum bandwidth following Serial ATA class 1 near to 3 GByte/s.
If the actual value for IDE disks of 40 MByte/s is accounted there is still a bandwidth of around 640 MByte/s.. A value can be handled by the single Infiniband IB X4 connections to the RAID boxes and to outside connections. However it needs a double connection to the outside as soon as better Serial ATA disks become available.. This can be either from the switch or, if the higher latency is not a problem, from the double ports at one or more of the nodes (dotted connections).

High Capacity Storage Macro-Cluster:
To get higher storage capacity it seems easy to connect more RAID boxes to the previous architecture. However the connectivity runs rapidly out of bandwidth. Two architectures are possible.
The first one as given in fig 8 and is in principle a copy of Fibre Channel SAN structures, and gives the same possibilities. The individual storage connections to the IB switch have far more bandwidth as needed, but the connections to the network side need to be fast, as bandwidth, referred again to Serial-ATA specifications can go as high as 8 X 2 X 600 MByte/s or 9.6 GByte/s. This is to much on the limit for a single IB X4 channel largely in the range for the 30 MByte/s bandwidth of IB X12 connection. Using the practical values of 40 MByte/s the bandwidth drops to 2.5 GByte/s, and as such again in range for an IB X4 connection. The processors shown can be used for Meta-Data or for a file system. The capacity, excluding one disk per group for parity storage only, using 200 GByte disks of 8 X 10 X 200 GByte or almost 16 TByte.

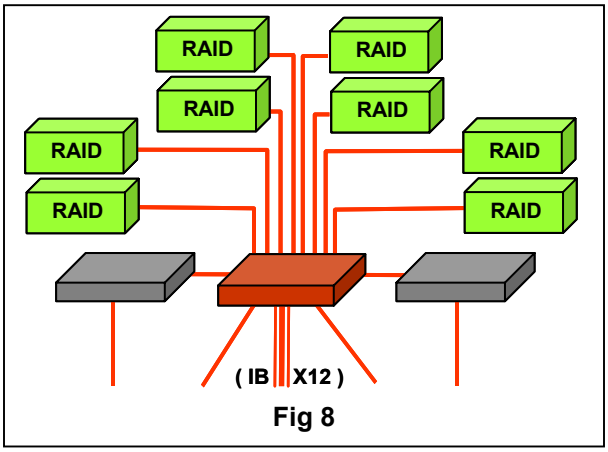


Fig 8

Avariant with different possibilities for large capacity storage is to use the second channel of the Interface to connect RAID boxes. As I/O run through the same channel simultaneously there is a limit in bandwidth depending largely on the efficiency of the PCIbus at the nodes. If the bandwidth of a single RAID box up to

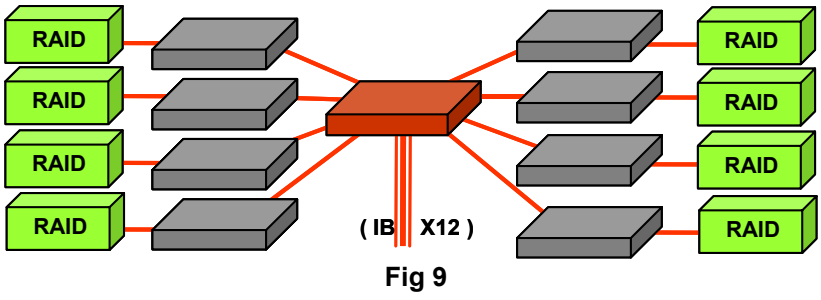


Fig 9

the IB switch is taken as 350 MByte/s there is still a total bandwidth of 2.8 GByte/s. Due to the channel limits the penalty for slow disks is more or less balanced. The capacity, excluding one disk per group for parity storage only, using 200 GByte disks of 8 X 10 X 200 GByte or almost 16 TByte. This kind of architecture can profitably be used with a distributed file system such as flavors of GFS, or if it turns out to be successful with Lustre.

Infiniband Price Comparison:

Price comparisons are in some way difficult due to differences in physical specifications. So is 10 GE extremely expensive because of the price of optical components. Fibre Channel has always been high priced by its complexity and small production numbers.

	Bandwidth	Host card	Host Latency	Switch/port	Switch Latency
10 GE Ethernet	10 Gbit/s	$ 7995	Not given	$ 12 000	35 – 45 µsec.
2 G Fibre Channel	2 Gbit/s	$ 999 - $ 1400	40 µsec.	$ 600 - $ 2200	1 – 2 µsec.
Infiniband IB X4	10 Gbit/s	$ 475 - $ 955	7 µsec. (MPI)	$ 600 - $ 1000	200 nsec
Gigabit Ethernet Prices 1998	1 Gbit/s	$ 900 - $ 1200	60 µsec.	$ 2000 - $ 2800	
Gigabit Ethernet Prices 2003	1 Gbit/s	$ 150 - $ 200$	60 µsec.	$ 400 - $ 650	

Table 2 Price Comparison; Infiniband Prices of 2003 (June 2003) should be compared with the Gigabit Ethernet introduction prices of 1998

The introduction of small the pluggable connectivity modules called GBITS and the miniaturized GBICS keep prices also high for the future.

The best comparison may be to look at the Infiniband X4 prices and compare theme with Gigabit Ethernet's introduction prices as in table 2. Corresponding to the Infiniband Roadmap (fig10) it will be only towards the end of 2004 that popularity and with it sales volume will increment drastically, leading to mass production and with it sharp drop in prices. So it seems only a matter of time that Infiniband pricing drops to the same level as Gigabit Ethernet, or lower. Also the simpler hardware constraints point in this direction. And it should be taken in account that

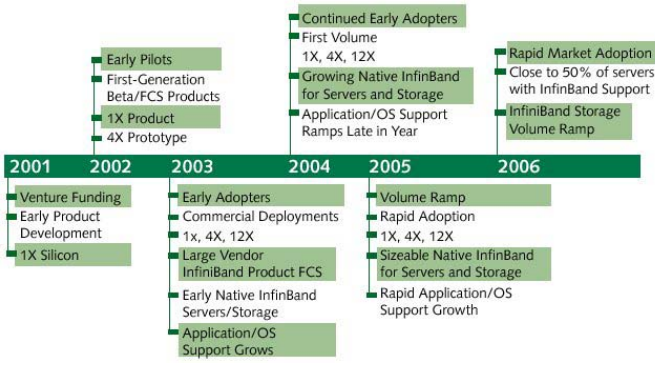


Fig 10 Infiniband Complete Roadmap

Infiniband as a cluster interconnect has, besides the bandwidth a number of advantages such as low latency and processor efficiency over all other interconnects except GSN.

Network Interconnects between Technologies:

HyperTransport connections are not made for networking. If back-panel connections come the cable length will be limited to short distances only. Interconnects between HyperTransport and TCP/IP oriented Ethernet networks is not directly possible. However the HyperTransport internal switch is also used for the connection between memory and the PCI type interface channels. Due to this characteristic the transfer is of a DMA type with a bandwidth limited by either the memory or the PCI bus. Reasonable transfer speeds should be obtainable with 10Gigabit Ethernet.

Infiniband copper connections are given for up to 17 m, with a single switch in between distances of up to 30 m can be covered, which is enough for Macro-Cluster interconnects. For longer distances Infiniband foresees fiber optics connections. For short-range connections 850 nm light sources are used on parallel fiber. IB 4 (IB-4X-SX) uses 2 X 4 fibers in a single MPO connector and IB 12 (IB-12X-SX) with 2 X 12 Fibers uses dual MPO connectors. Long distances are only available in IB 1 and use 1350 nm optics. With Single mode fiber a distance of 10 Km can be covered. However stretching the distance to the maximum will introduce non-negligible distance latency by the logical handshakes. In order to introduce flexible use of switch and interface ports a modular system is proposed in the form of plug-in units called IBPAKs. They should be available by the end of 2003.

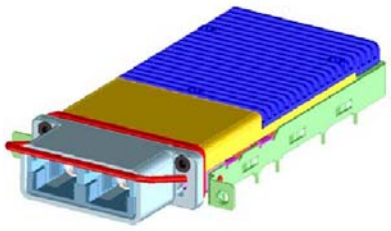


Fig 11 X2-MSA Plug In Module

To connect Infiniband with Ethernet type networks is at a first look more problematic as protocols are not the same. Looking at both standards physical specifications however shows that IB X 4 is composed of 4 streams of 2.5 Gbit/s. The intermediate interface of 10 GE called XAUI also works with 4 streams of 2.5 Gbit/s, and has at the receiving end a deskew mechanism. Both quadruple serial streams are synchronous. Coupling can be at two levels, either at the clocked XGMII level or at the self-clocked bidirectional differential signal level, that are part of each of the systems. Due to this similarity network devices will come that have either IB 4X channels between 10 GE or the opposite. Also 10 GE material starts to use pluggable I/O modules called X2-MSA (fig 11). The size and other specifications are in many respects similar to the IBPAK. So it is very well thinkable that an IB X4

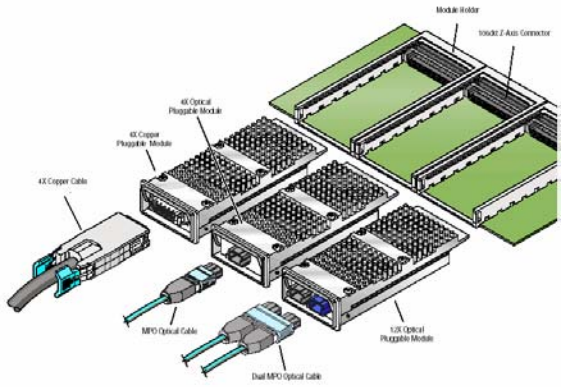


Fig 12 The Proposed IBPAK Modules

version makes switch I/O channel connectivity corresponding to the module used, either 10 GE or IB X4. The results will be that after the first series of Infiniband switches there will come a generation to the market with pluggable port interfaces and the possibility to have one or more ports that can also used for 10 GE. Which means that for coupling to existing Ethernet type TCP/IP networks installations and Infiniband and for longer distance connections the best way may be via 10 Gbit/s Ethernet backbones.

There are differences in the protocols; Ethernet uses IP and Infiniband IQ. They can be chosen to equal, but the way they are handled in the protocols is different. A mechanism is needed that can decide if a transfer uses Ethernet or IB, which needs either look-up tables or some kind of name server. Industrial solutions are already available in the form of routers for IB to Infiniband and FC to Infiniband (Voltaire)

The remaining problems are the difference in frame size and the place and size of the CRC in the protocols. Going from the larger IB frames to Ethernet frames without loosing bandwidth needs to use Ethernet "Virtual Concatenation" protocol, as defined for the Resilient Packet Ring. In the opposite direction Byte stuffing can place multiple Ethernet frames in an IB packet.
It is shown in the past that those problems can be solved with relatively simple hardware without creating extra latency.

Some Software Aspects to Couple Between Technologies:

Infiniband has its own protocol structure, and its own stacks that are certainly oriented towards Ipv4 and Ipv6, but are at the same time sufficiently different to be not compatible. It is certainly possible to use standard TCP/IP over Infiniband but on the price of very high software latency. The IpoIB (Ip over IB) Interface is a level 2 network interface that uses the TCP/IP stack before converting to a IB transfer (Fig 13). Infiniband HCA's are normally delivered with a set of drivers and a transport layer. For LINUX a set of ß level software is available that couples IB to the socket direct level in the form of the SDP (Socket Direct protocol). To keep latency as low as possible SDP allows bypassing the OS resident TCP stack for all stream connections between any endpoints on the Infiniband fabric. The SDP stack itself depends on an IPoIB link driver for local IP assignments and for IP address resolution. Commercial solutions to the problem of TCP/IP over Infiniband are becoming available a is the case with NDIS and Remote NDIS (Microsoft) For network storage an iSCSI like protocol is available in the form of SCSI over RDMA Protocol (SRP) defined by the SCSI standardization body ANSI T-10. This project will define a Linux block storage SCSI driver that implements the SRP. It plugs into the Linux SCSI mid-layer. As SRP uses RDMA it can be expected that software latency is low, combined with low processor use. It depends largely on memory and I/O bandwidth if there will also be an increase in bandwidth. For CERN applications application layers such as SocketDirect must be able to communicate with the existing infrastructure. A way has to be found to use the data structures as used in castor and NSF. Interfacing with a script level using Posix seems a simple way out but will have its penalty in achievable bandwidth.

By the use of macro-clusters management functions get a more hierarchical orientation with a macro manager for local management, who function it is to report to a central management node. This may change the way run status is handled. The way to obtain this status will not be different from known procedures. The same is for reset and exclude procedures; they can be handled centrally or at the macro level.

Granularity will change for distribution of nodes to different users. In a large flat cluster it is possible to distribute nodes to users in units. Using macro clusters this is not impossible but it will be more efficient to distribute in units of a macro cluster. If storage macro clusters are used this distribution can include storage capacity.

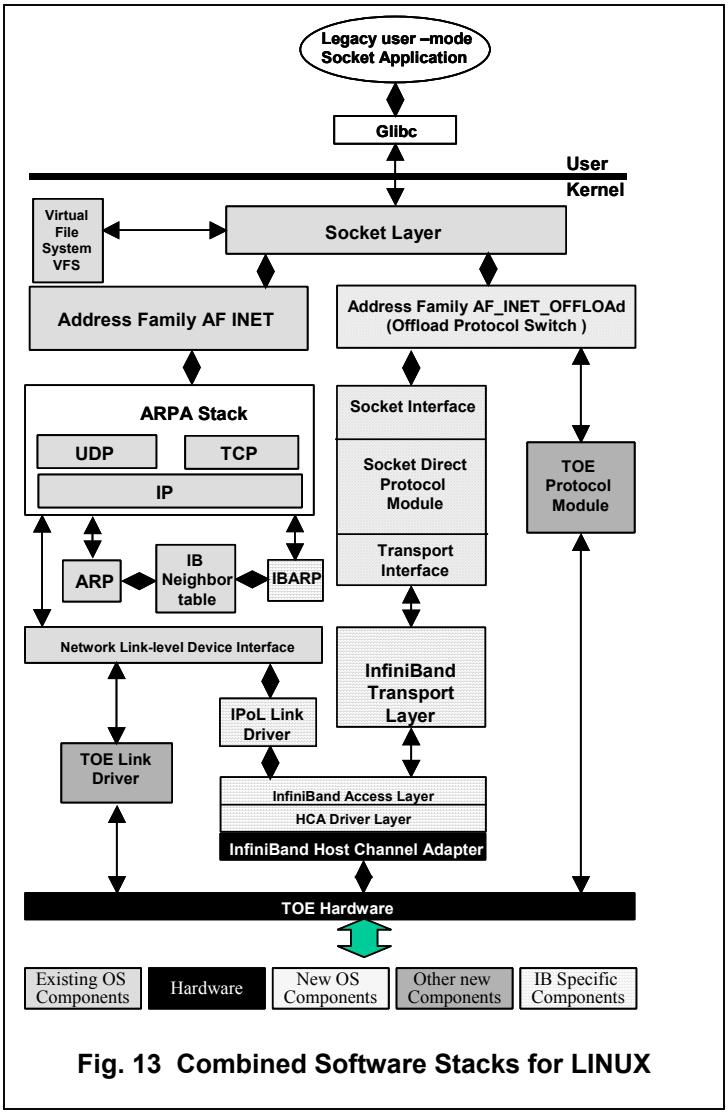


Fig. 13 Combined Software Stacks for LINUX

Conclusion:

The term Macro-Clusters is introduced here to make conversion from a TCP/IP cluster to a partly or completely mixed cluster oriented towards HyperTransport or Infiniband or both. This is a process that can be executed as an evolutionary process. In this way macro clusters can first be applied in places where it really brings advantages because of the increments in bandwidth and flexibility. The CERN computer centre as a GRID tier 1 station has to take in account what happens at other places that mostly use MPI. And it is well possible that LHC analyzing software needs low latency message passing. Both HyperTransport and Infiniband will both excellent solutions

Infiniband is a way that leads to more flexible and faster interconnects with excellent characteristics for MPI. Some problems have to be solved on the level of protocol adaptation and coupling to CERN production software

Also from the point of storage macro servers may it be centralized network storage or distributed storage servers, both technologies can be used. HyperTransport can use IDE-

RAID boxes that couple with Gigabit Ethernet or Fibre-Channel (Infortrend). Infiniband can bring higher bandwidth with simpler an architecture, especially the moment cheap RAID boxes come on the market with native Infiniband connections.

Infiniband just comes to the market but shows already aggressive pricing that is under the values that ware shown with the introduction of Gigabit Ethernet. Compared other 10 Gbit/s networks IB X4 is factor 10 less budget intensive. So it may be that on the long term Infiniband turns out to be less budget intensive as Gigabit Ethernet connections, with 10 X the bandwidth. This brings for certain the best cost performance relation for 10 Gbit/s interconnects.

HyperTransport will still be TCP/IP oriented, but will bring macro clusters with shared memory possibilities and interfaces to Ethernet and Infiniband by means of PCI type plug-ins. HyperTransport has foreseen to integrate in a later phase native Infiniband.

A step further will be if HyperTransport connectivity becomes available, that makes possible that a certain number of nodes couple their memory space.
Seen from the price performance relation, Opteron HyperTransport, 32/64 bit machines are already available at the same price level as 32 bit machines.

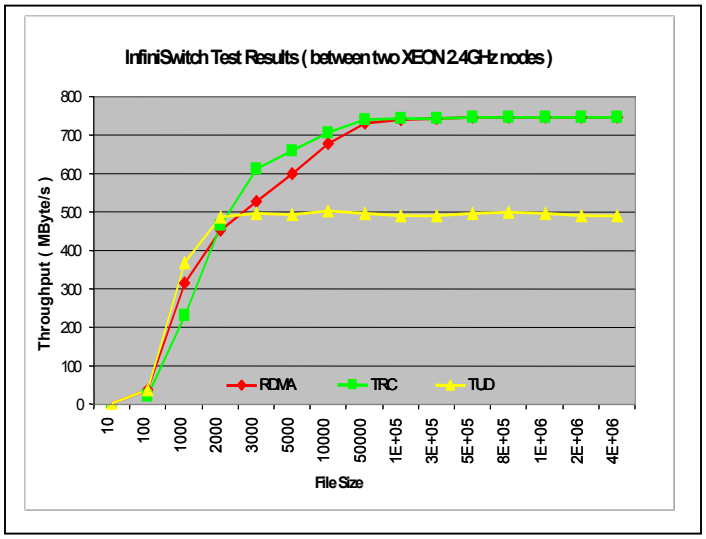
Appendix

Some Test Results

Test results from 6 June 2003 on 32 bit XEON machines;

This bandwidth tests are done by Andras Horvath and Mathias Gug on Dual Processor XEON 2.4 MHz machines running Red-Had Liniux. Test software used is perf_main from Mellanox, Updated HCA firmware 1.18 and drivers, Processor 100%

File Size Bytes	Infiniband Bandwidth MByte/s		
	UD Mode	RC Mode	RDMA Mode
10	3.9	2.3	2.9
100	38.6	23.4	33.7
1000	385.7	232.7	460.6
2000	538.7	464.7	604.6
3000	*n.a.	611.8	649.4
5000	n.a.	658.7	688.6
10 000	n.a.	706.3	724.9
50 000	n.a.	739.8	741.8
100 000	n.a.	744.0	745.1
250 000	n.a.	744. 7	743.3
500 000	n.a.	746.1	745.4
750 000	n.a.	746.2	744.9
1 000 000	n.a.	747.9	747.2
2 000 000	n.a.	747.4	745.4
4 000 000	n.a.	748.5	746.4



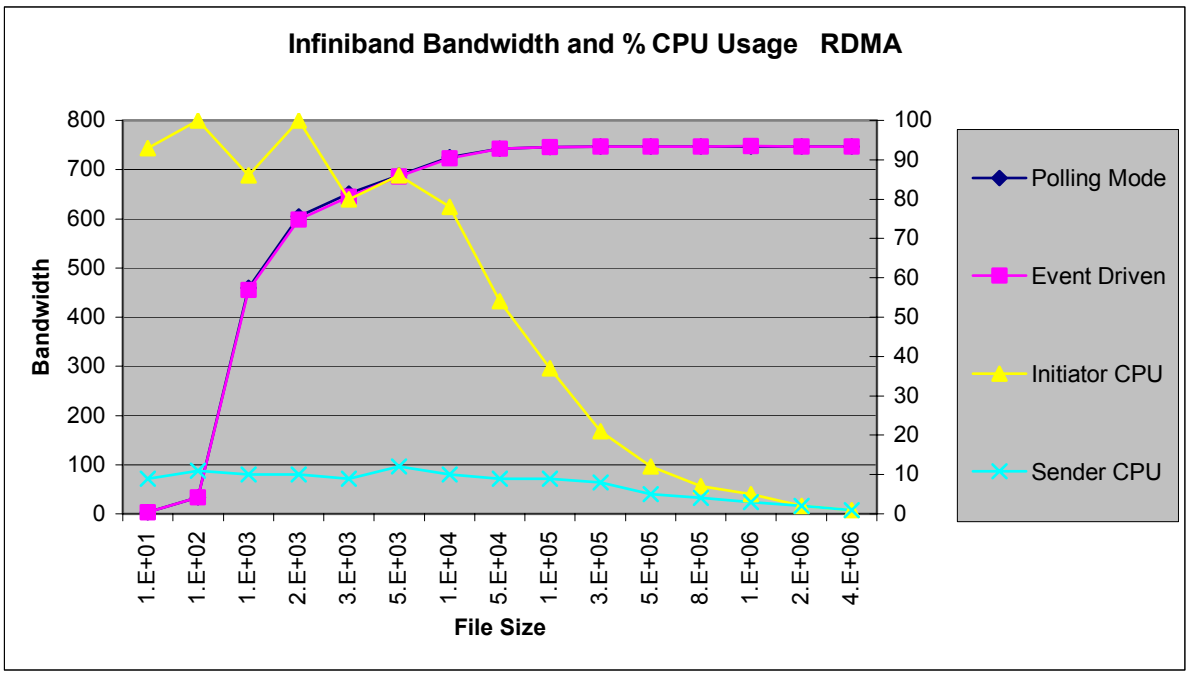
Remarks: * UD mode is limited to small size frames of max, 2048 Bytes.

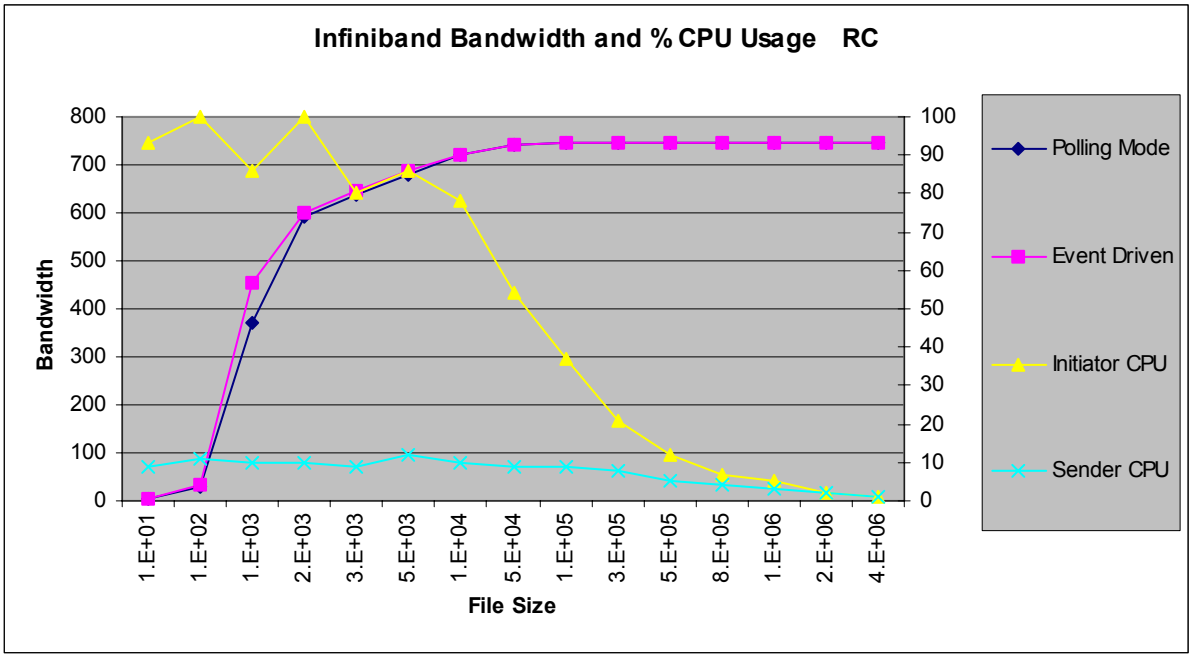
Processor Usage This bandwidth test uses polling to find the end of transfer, and as such will always show 100% processor use.
With the Mellanox test tool in event driven mode and processor usage measured with PO; overall CPU utilization is around 1-2% on the sender and 3-4% on the initiator (with an occasional peak of 10% on the latter for small message sizes)

Remarks: This bandwidth test uses polling to find the end of transfer, and as such will always show 100% processor use.

Infiniband Tests 17-7-2003

File Size	Infiniband Bandwidth MByte/s								
	UD Mode	RC Mode				RDMA Mode			
Bytes		Polling	Event driven	CPU Initiater	CPU Sender	Polling	Event Driven	CPU Initiator	CPU Sender
10	3.9	2.9	2.9	93	9	2.9	2.9	93	9
100	38.6	29.7	33.3	100	11	33.6	33.3	100	11
1000	385.7	370.3	455.0	86	10	459.7	455.0	86	10
2000	538.7	589.9	598.2	100	10	605.4	598.2	100	10
3000	*n.a.	636.3	645	80	9	651.5	645.0	80	9
5000	n.a.	681.2	686.2	86	12	688.3	686.2	86	12
10 000	n.a.	718.9	722.7	78	10	725.4	722.7	78	10
50 000	n.a.	742.1	742.2	54	9	742.6	742.2	54	9
100 000	n.a.	745.4	745.3	37	9	745.6	745.3	37	9
250 000	n.a.	747.1	746.8	21	8	746.9	746.8	21	8
500 000	n.a.	747.6	747.2	12	5	747.3	747.2	12	5
750 000	n.a.	747.8	747.3	7	4	747.4	747.3	7	4
1 000 000	n.a.	747.8	747.5	5	3	747.4	747.5	5	3
2 000 000	n.a.	747.5	747.1	2	2	747.2	747.1	2	2
4 000 000	n.a.	747.7	747.3	1	1	747.3	747.3	1	1

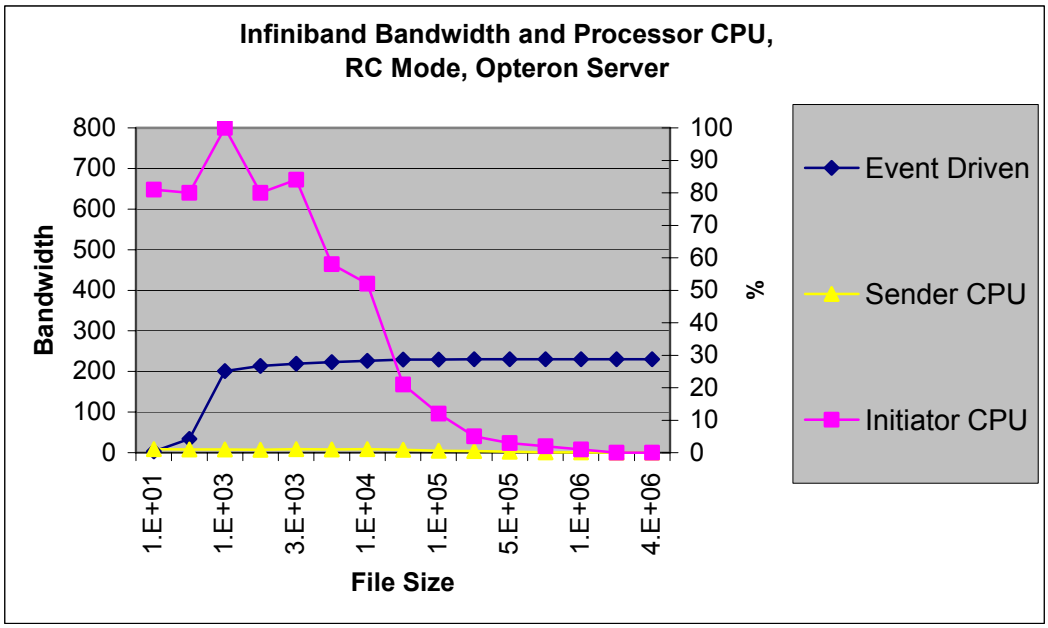


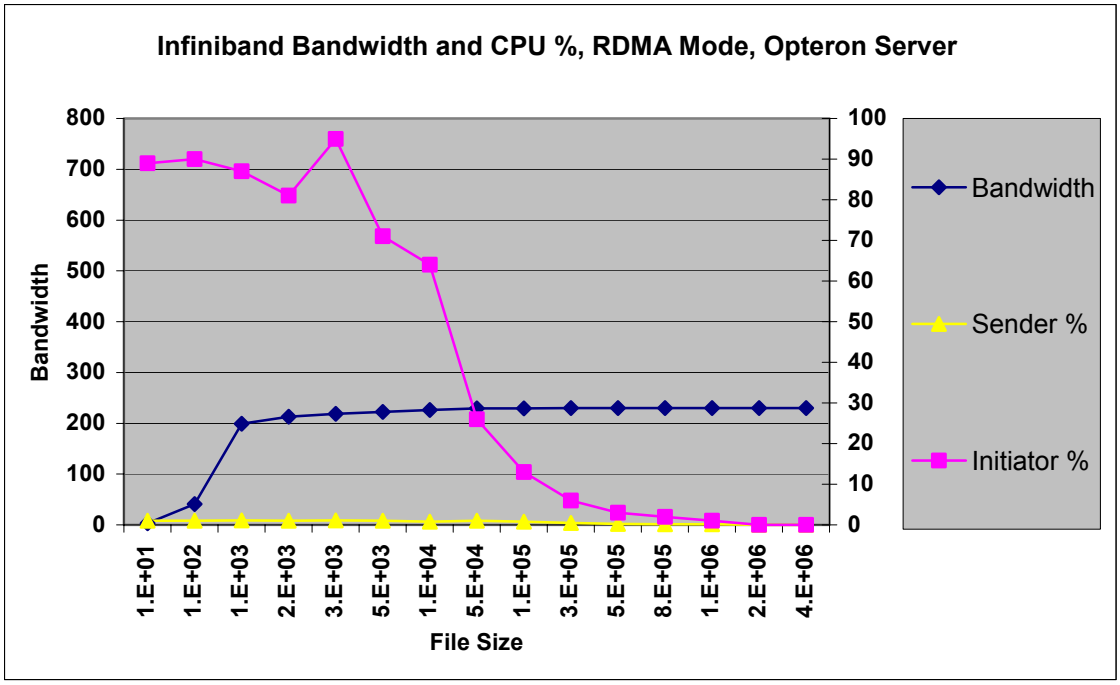

Infiniband Bandwidth and % CPU Usage RC
Bandwidth
800
700
600
500
400
300
200
100
0
100
90
80
70
60
50
40
30
20
10
0
1.E+01
1.E+02
1.E+03
2.E+03
3.E+03
5.E+03
1.E+04
5.E+04
1.E+05
3.E+05
5.E+05
8.E+05
1.E+06
2.E+06
4.E+06
File Size
Polling Mode
Event Driven
Initiator CPU
Sender CPU

Tests on Opteron

Test on Opteron machines are done by Andras Horvath on Dual Opteron 1.4 GHz machines. System OS is Suse Linux Server ver. 8.2
The 1.4 GHz processors used belong to a early series and are known to have a problems concerning I/O bandwidth.
Such that it need to be mentioned that this testresults show functionality but are not typical for the Opteron processor. New tests will be done as soon as we have the latest version Opteron of 2 GHz or better available.

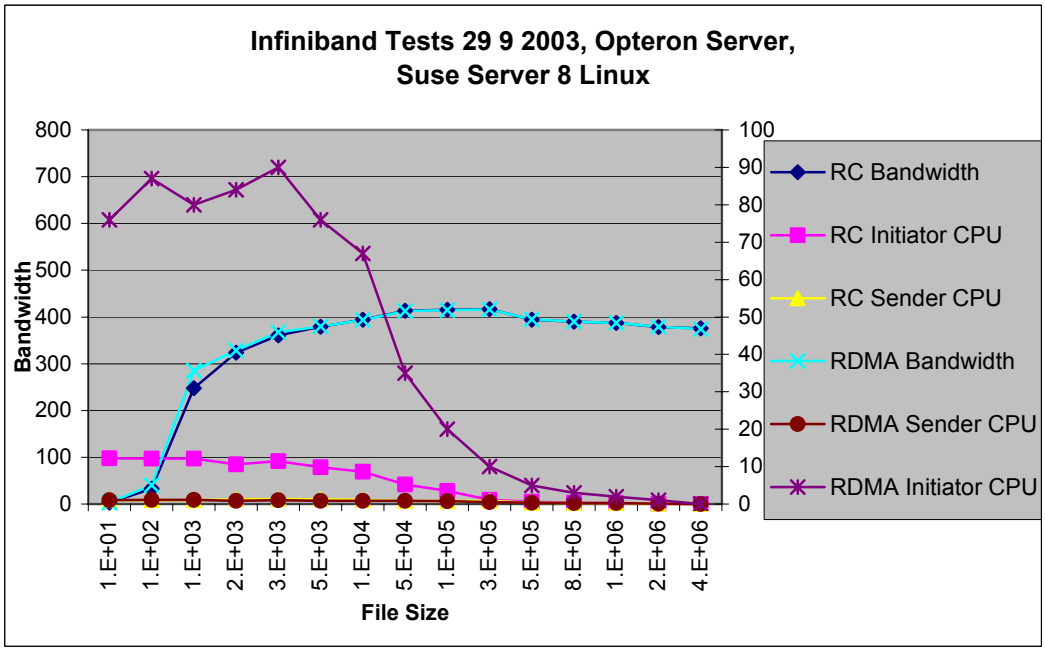
File Size	Infiniband Bandwidth Mbyte/s and CPU Usage Opteron Server 16-9-2003						
	UD Mode	RC Mode			RDMA Mode		
Bytes	Mbyte/s	MByte/s	CPU Initiator	CPU Sender	MByte/s	CPU Initiator	CPU Sender
10	3.4	2.9	81	9	2.9	89	8
100	37.2	33.5	80	8	40.8	90	8
1000	193.4	200.6	100	8	199	87	9
2000	194.7	213.8	80	7	213.1	81	8
3000	180.8	218.8	84	9	218.4	95	9
5000	n.a.	223.1	58	8	222.6	71	8
10000	n.a.	226.5	52	9	226.1	64	6
50000	n.a.	229.4	21	7	229.2	26	8
100000	n.a.	229.7	12	5	229.6	13	6
250000	n.a.	229.9	5	4	229.8	6	4
500000	n.a.	230	3	2	229.9	3	2
750000	n.a.	230.1	2	1	229.9	2	1
1000000	n.a.	230.1	1	1	229.9	1	1
2000000		230.1	0	0	229.9	0	0
4000000		230.1	0	0	230	0	0





Tests 29 September 2003

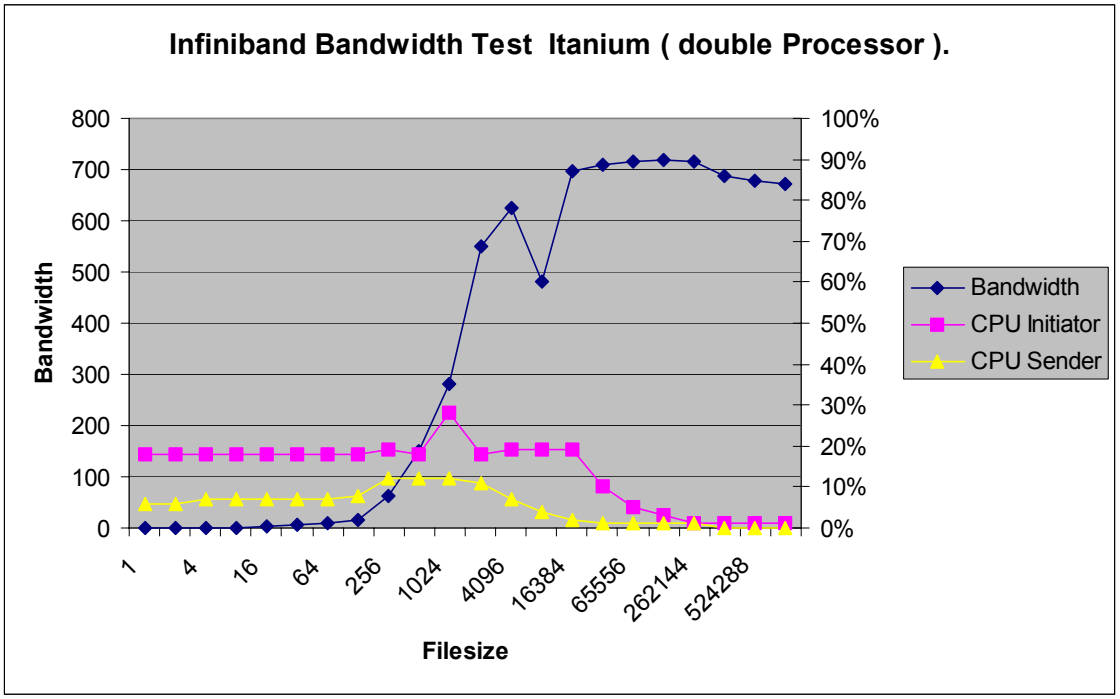
File Size	Infiniband Bandwidth Mbyte/s and CPU Usage Opteron Server 10-9-2003						
	UD Mode	RC Mode			RDMA Mode		
Bytes	Mbyte/s	MByte/s	CPU Initiator	CPU Sender	MByte/s	CPU Initiator	CPU Sender
10	3.3	2.4	98	8	2.9	76	8
100	37.2	32.8	97	8	40.2	87	9
1000	192.6	248.0	97	8	285.3	80	9
2000	281.6	323.8	85	10	328.7	84	7
3000	288.8	360.6	92	10	367.7	90	8
5000	288.9	379.4	79	10	379.7	76	7
10000	n.a.	394.4	69	9	393.5	67	7
50000	n.a.	413.4	42	7	412.8	35	7
100000	n.a.	415.1	28	6	414.7	20	6
250000	n.a.	416.5	9	7	416.2	10	4
500000	n.a.	394.4	5	3	394.6	5	3
750000	n.a.	390.4	3	2	389.9	3	2
1000000	n.a.	387.4	2	2	387.1	2	2
2000000	n.a.	378.5	1	1	379.2	1	1
4000000	n.a.	375.9	0	0	374.1	0	0


Infiniband Tests 29 9 2003, Opteron Server,
Suse Server 8 Linux
Bandwidth
800
700
600
500
400
300
200
100
0
100
90
80
70
60
50
40
30
20
10
0
1.E+01
1.E+02
1.E+03
2.E+03
3.E+03
5.E+03
1.E+04
5.E+04
1.E+05
3.E+05
5.E+05
8.E+05
1.E+06
2.E+06
4.E+06
File Size
RC Bandwidth
RC Initiator CPU
RC Sender CPU
RDMA Bandwidth
RDMA Sender CPU
RDMA Initiator CPU

Tests On Itanium

The tests on Itanium are done by Fabian Collin on Dual Itanium 2 1.5 GHz Machines using Linux Kernel 2.4.20-1

Dual Itanium 2 1.5 Ghz		Kernel 2.4.20-1	vmstat output (CPU % relative to 2 CPUs!!!)				
perf_main benchmark from Mellanox Software kit 2.0					17 -10-2003		
Msg size	Datarate MB/s	CPU (receiver)	int/s	ctxt sw/s	CPU (sender)	int/s	ctxt sw/s
1	0.2	18%	29000	54000	6%	3800	3600
2	0.3	18%	29000	54000	6%	3800	3600
4	0.6	18%	29000	54000	7%	3800	3600
8	1	18%	29000	54000	7%	3800	3600
16	2.4	18%	29000	54000	7%	3800	3600
32	5.2	18%	29000	54000	7%	3800	3600
64	9.8	18%	29000	54000	7%	3800	3600
128	16.6	18%	29000	54000	8%	3800	3600
256	63.1	19%	19000	30000	12%	4700	5300
512	148.7	18%	9800	15500	12%	5200	6200
1024	280.3	28%	4600	5200	12%	5500	6800
2048	550.1	18%	12700	21000	11%	4900	5700
4096	624	19%	29800	55500	7%	3700	3200
8192	480.1	19%	40000	75000	4%	2900	1700
16384	695.8	19%	45800	87500	2%	2500	900
32768	709.2	10%	24800	45500	1%	2200	470
65556	716.2	5%	13500	23000	1%	2200	250
131072	719.8	3%	7800	11500	1%	2100	130
262144	715.4	1%	4900	5700	1%	2100	65
393216	686.8	1%	3900	3600	0%	2100	45
524288	677.8	1%	3400	2700	0%	2000	40
786432	672.4	1%	2900	1800	0%	2000	40



Test for coupling to the CERN storage system.

Test on connections to the CERN storage system CASTOR using RFIO are done by Andras Horvath in collaboration with the informatics group from Karlsruhe Institute for Nuclear Physics.

Results with AI64 double processor Itanium machines running LINUX:

Throughput Memory to Memory	436 MByte/s
Processor Usage	21 %

The following is the result script:

```
ahorvath@oplapro29:/tmp$ /usr/bin/time rfcp oplapro30:/data2/try/512m.file /dev/null
536870912 bytes in 1 seconds through VAPI (in) and local (out) (524288 KB/sec)
0.11user 0.02system 0:01.23elapsed 11%CPU (0avgtext+0avgdata 0maxresident)k
0inputs+0outputs (194major+485minor)pagefaults 0swaps
ahorvath@oplapro29:/tmp$ /usr/bin/time rfcp oplapro30:/data2/try/512m.file /dev/null
536870912 bytes in 2 seconds through VAPI (in) and local (out) (262144 KB/sec)
0.11user 0.02system 0:01.23elapsed 10%CPU (0avgtext+0avgdata 0maxresident)k
0inputs+0outputs (194major+485minor)pagefaults 0swaps
ahorvath@oplapro29:/tmp$ /usr/bin/time rfcp oplapro30:/data2/try/512m.file /dev/null
536870912 bytes in 1 seconds through VAPI (in) and local (out) (524288 KB/sec)
0.11user 0.02system 0:01.24elapsed 11%CPU (0avgtext+0avgdata 0maxresident)k
0inputs+0outputs (194major+485minor)pagefaults 0swaps
ahorvath@oplapro29:/tmp$

raas@pcitadc01:~/tmp$ bc
bc 1.06
Copyright 1991-1994, 1997, 1998, 2000 Free Software Foundation, Inc.
This is free software with ABSOLUTELY NO WARRANTY.
For details type `warranty'.
536870912/1.23
436480416
```

Results with AI32 double processor Itanium machines running LINUX:

Throughput Memory to Memory	300 MByte/s
Processor Usage	30 %

References

1. 10 Gigabit Ethernet - an introduction, 10 Gigabit Ethernet Alliance, September 10, 2000Rev 1.0b
2. AMD HyperTransport-based System Architecture, White Paper, Advanced Micro Devices, May 2002
3. HyperTransport Technology, White Paper, Advanced Micro Devices, July 20 2001
4. Infiniband Architecture Specification 1, Infiniband Trade Association, June 19, 2001
5. XAUI: An Overview, John D'Ambrosio, Shawn Rogers, Jim Quilici, 10 GEA Whitepaper, 1 March 2002.
6. Linux InfiniBand, http://infiniband.sourceforge.net/
7. SCSI RDMA Protocol (SRP), http://infiniband.sourceforge.net/
8. Sockets Direct Protocol (SDP), http://infiniband.sourceforge.net/
9. Remote NDIS and Windows.
   http://www.microsoft.com/whdc/hwdev/tech/network/rmNDIS.mspx#XSLTsection124121120120

Useful Web Addresses

1.	10 Gigabit Ethernet:	http://www.10gea.org/
2.	Myrinet	http://www.myri.com/
	Fibre Channel	http://www.fibrechannel.org/
	HyperTransport	http://www.hypertransport.org/ http://www.amd.com/us-en/Corporate/VirtualPressRoom/0,,51_104_4699_4741%5E4752,00.html
3.	Linux Software for IB	http://infiniband.sourceforge.net/

Infiniband Manufacturers

1.	http://www.fabricnetworks.com/	before Infiniswitch, Switches, HBA's, software
2.	http://www.mellanox.com/	Chipsets, HBA's, Switches, Driver software
3.	http://www.infiniconsys.com/	HBA's, Switches,
4.	http://www.voltaire.com/	HBA's, Modular Switches, IB to GigE Routers, Software
5.	http://www.divergenet.com/	HBA's, Switches, IB Native to Storage routers
6.	http://www.jni.com/	HBA's, HCA's
7.	http://www.voltaire.com/index.html	Routers, Infiniband to IP, Infiniband to FC