# Cluster Interconnection Networks

Andrew Griffiths and Glenn Metherall

School of Computer Science and Software Engineering
Monash University
Clayton Campus, Melbourne, Australia

Email: {andrewg, glennm}@csse.monash.edu.au

**Abstract**

With huge advances in computing power being made all the time the performance capabilities of nodes in clusters is becoming increasing great. With this increase in computing power comes a problem in that the bottlenecks in the system shift from the nodes to the actual networks which connect them.

In order to reduce bottlenecks in the networks between nodes new Interconnection schemes are always being developed and evaluated. In order to chose the best Interconnection scheme (comprising of topology, protocols and link types) it is important to understand the current technologies on offer and the likely ones of the future.

Burns and Wallace's paper evaluates the performance abilities of SCI for creating Interconnection Networks, Akyildiz and Seong paper discusses using ATM over Satellite networks and Almus gives an overview of a Pan-European network of LAN's using ATM.

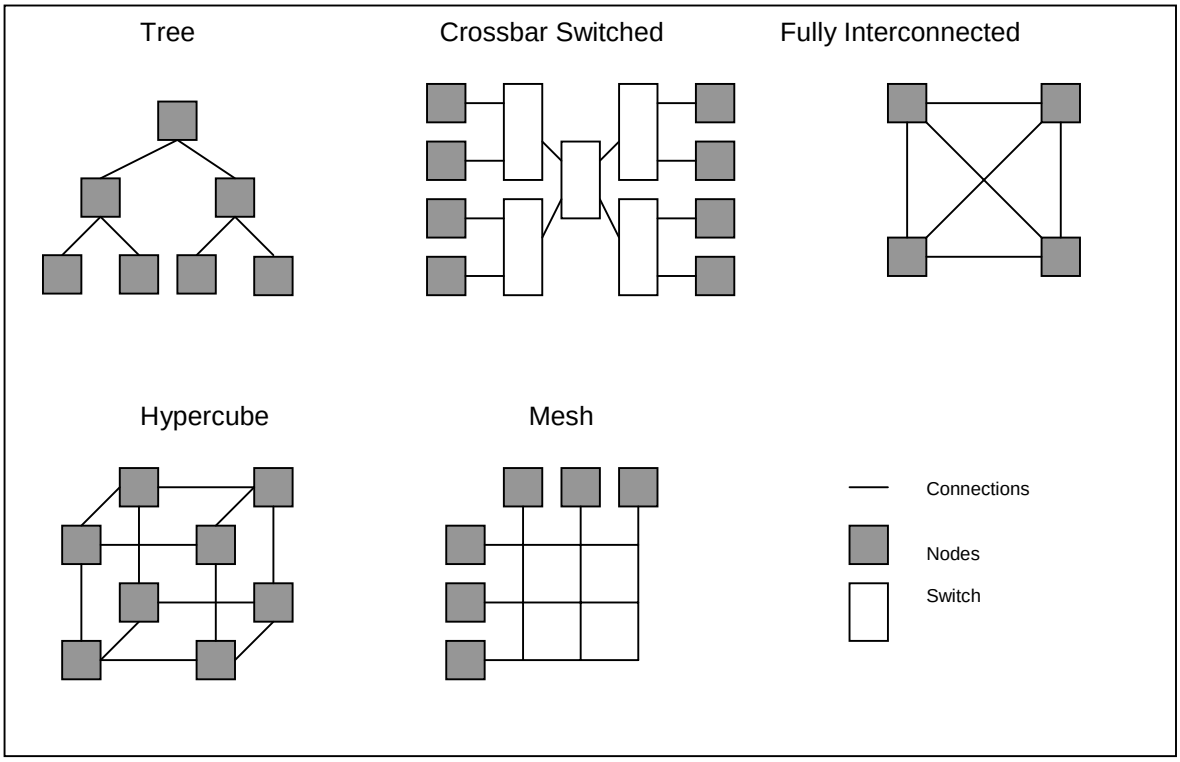
## 1. Introduction

Vital to the cluster systems in operation today (and most likely those in the future) is the existence of a network connecting the nodes and allowing them to communicate information with each other – or a master server controlling the cluster.

Although communication abilities between the nodes (either directly or indirectly) are vital to the operation of a cluster, the exact communication abilities and indeed the way in which these are implemented can change from cluster to cluster or indeed in one cluster over time.

A general principle adhered to is the provision of communication abilities through the use of an interconnection network between the nodes. Just like a standard network these interconnection networks may be set up with many different topologies, some of which are in current use… but many of which are still being developed and researched in order to evaluate the merits of all the different possibilities.

Current Popular interconnection network topologies include Tree, Crossbar, Hypercube and Mesh Topologies which are represented diagrammatically below;




Each of these different topology's provide their own merits and drawbacks in terms of communication abilities. Whilst the Fully interconnected one provides the most abilities in terms of communication, it is also the most expensive to implement in most cases and doesn't scale well in terms of cost.

In addition to the many different topology choices available for the implementation of an interconnected network, there are also a number of different protocols currently available for communication on the network, and many more are in development.

While most initial development of Clusters used IP over Ethernet to implement communication, new protocols and systems such as ATM, SCI and even Myranet are beginning to receive widespread use due to the extra speed and (in some cases) capabilities they provide. Most clusters still continue to use IP as a preferred higher level protocol for interconnection network communication but these new link level protocols are receiving more attention.

While once 10mbps Ethernet was the most popular interconnection protocol, new 100mbps and 1000mbps Ethernet systems are now being used in order to provide extra speed without much extra overhead required to change network protocols.

ATM Also provides a good system with increased speed and lower latency than the 10mbps Ethernet systems – primarily due to the use of smaller, fixed length datagrams instead of Ethernet's large variable sized packets.

SCI systems provide similar advantages as 1000mbps Ethernet systems except that they inherently offer a cache coherency scheme which can be used for the created of shared memory systems across a number of nodes on a network.

Finally Myranet systems are also beginning to be developed based on Myranet communications protocols which were designed with cluster technology in mind and as such provides features useful in the implementation of cluster systems.

In this report a number of Parallel cluster system reports investigating the use of these new technologies will be listed, then several discussed in more detail in order to provide an insight into the research currently taking place.

## 2. Works on this area

| Network | Work | Description | Url |
|---|---|---|---|
| ATM | Berlin ATM LAN Connection | This work, discussed in more detail below, looks at a project in Berlin and throughout Europe which hopes to extensively test ATM. | http://www.prz.tu-berlin.de/docs/html/EANTC/PROJ-ECTS/BALI/index.html |
| | Networks Evaluation and Benchmarking Project | A network being run through Melbourne, Adelaide and Canberra to demonstrate the need for reserved bandwidth and investigate how it can best be managed. | http://www.dhpc.adelaide.edu.au/projects/network/index.html |
| Myrinet | The Berkeley NOW Project. | The Berkeley (NOW) project seeks to harness the power of clustered machines via myrinet switched network. | http://now.cs.berkeley.edu/ |
| SCI | A performance Evaluation of SCI as a Parallel Computing Interconnection Network | Compares performance achieved using different configurations of number of nodes & node type using SCI Interconnected Network | http://www.hcs.ufl.edu/~wallace/EEL5718/project.html - http://www.hcs.ufl.edu/~wallace/EEL5718 |
| Ethernet | The Beowulf Project | Site discusses history and the creation of the beowulf machine and it's uses with Ethernet and Fast Ethernet networks. | http://www.beowulf.org/ |
| Gigabit Ethernet | Clustering-in Search for Scalable Commodity Supercomputing | Various papers discussing different aspects of clustering computers and related issues. | http://www.dgs.monash.edu.au/~rajkumar/papers/informatica.html |
| | CWRU Beowulf Introduction | Talks about evolution of the CWRU Beowulf cluster | http://home.cwru.beowulf/documentation |

| | | | |
|---|---|---|---|
| Miscellaneous | A simulation Research on Multiprocessor Interconnection Networks with Wormhole Routing | Discusses issues arising when selecting appropriate network type for clusters | Abstract available online at |
| | Cluster Information | Describes the clusters implemented by those at Gatech University | http://www.cc.gatech.edu/projects/ihpcl/clusters.html |
| | Designing Scalable Parallel Architectures with Processor Clustering | Discusses development of Interconnection Network Topology designed for Scalability using modern advances | http://www.cc.gatech.edu/projects/ihpcl/clusters.html |
| | Scalable Networked Information Processing Environment (SNIPE) | Describes a metacomputing system providing numerous functionality's which aid in development of fault tolerant, distributed systems | http://www.supercomp.org/sc97/program/TECH/MOORE/INDEX.HTM |

## 3. A performance Evaluation of SCI as a Parallel Computing Interconnection Network

Mark W. Burns, Brad Wallace

Computer Communications

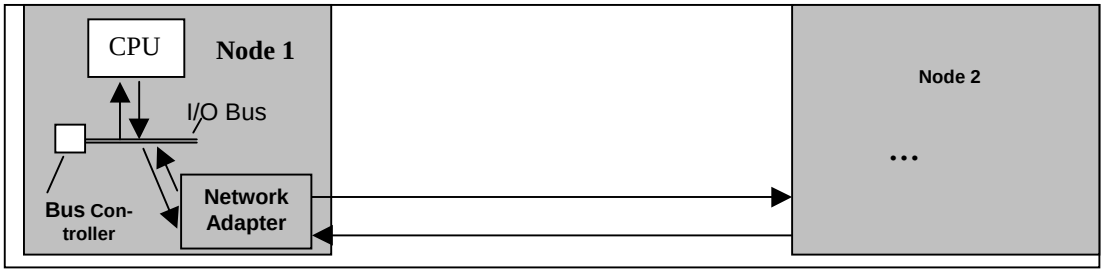
The goal of this project is currently to evaluate the performance achieved with the use of different numbers and different types of nodes when using an SCI Interconnection Network. Throughout the development of this project however the exact specifications have changed due to problems encountered.

Some of these problems Burns and Wallace have encountered in the creation of this SCI Network and the solutions they have found and presented provide valuable information to other wishing to create and use systems of clusters running over SCI or indeed those creating parallel computing systems in general.

Initially Burns and Wallace hoped to create an Interconnected Network cluster using Myinet instead of SCI for communication between the nodes but discovered that although Myrinet was designed with clusters in mind, some features which they felt necessary such as the ability to provide shared memory access weren't available in Myrinet without proprietary extensions.

For this reason SCI was chosen as the communication protocol to be used and was then integrated into the computer systems via a connection to the I/O busses of the nodes.

This was accomplished through the design of a network interface which acted as a secondary bus controller for the Nodes I/O busses. This allows the node to treat the Network adapter as an I/O device instead of using IP software to send packets through a typical network card.




Preliminary performance results achieved using their systems to perform matrix multiplication tasks half way through their study provided very promising results for Large scale Clusters, as well as for Clusters of SMP Machines each containing two processors (although in this case only one IO bus and one network adapter).

In the final report comparisons are made between SMP Clusters, Uniprocessor clusters using SCI implemented through the I/O bus and Clusters which implement a shared memory bus through proprietary architecture.

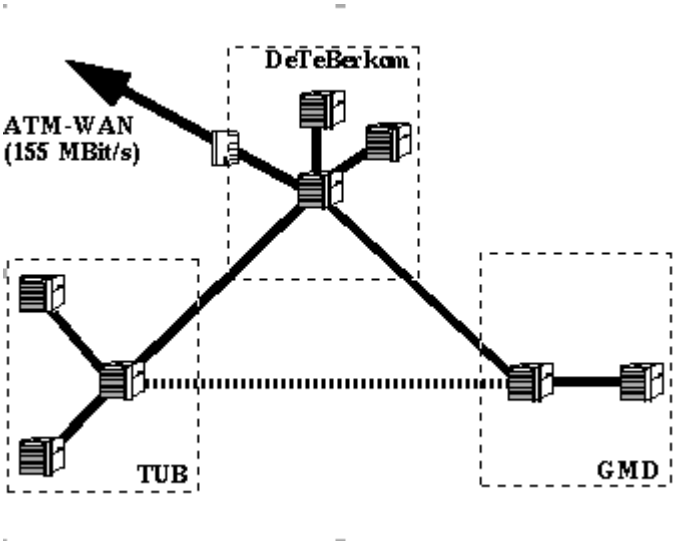
As expected their results show that for clusters with a small number of nodes the use of SCI significantly helped performance when doing parallel computations such as matrix multiplication.

Against what they expected however is the fact that the SMP machines performed worse than their uni-processor counterparts using the same number of CPU's. A suggested explanation is that the choice of a single I/O bus and network adapter configuration to be used with the SMP machines severely degraded performance by creating a bottleneck at the I/O Bus [1].

## 4. The Berlin ATM LAN Interconnection (BALI)

The BALI project was initiated by three German universities as a way of connecting a major area to an ATM network. This network is also connected to the Pan-European ATM Network which extends all the way through Europe, from Madrid to Dublin, Vienna and even Helsinki.

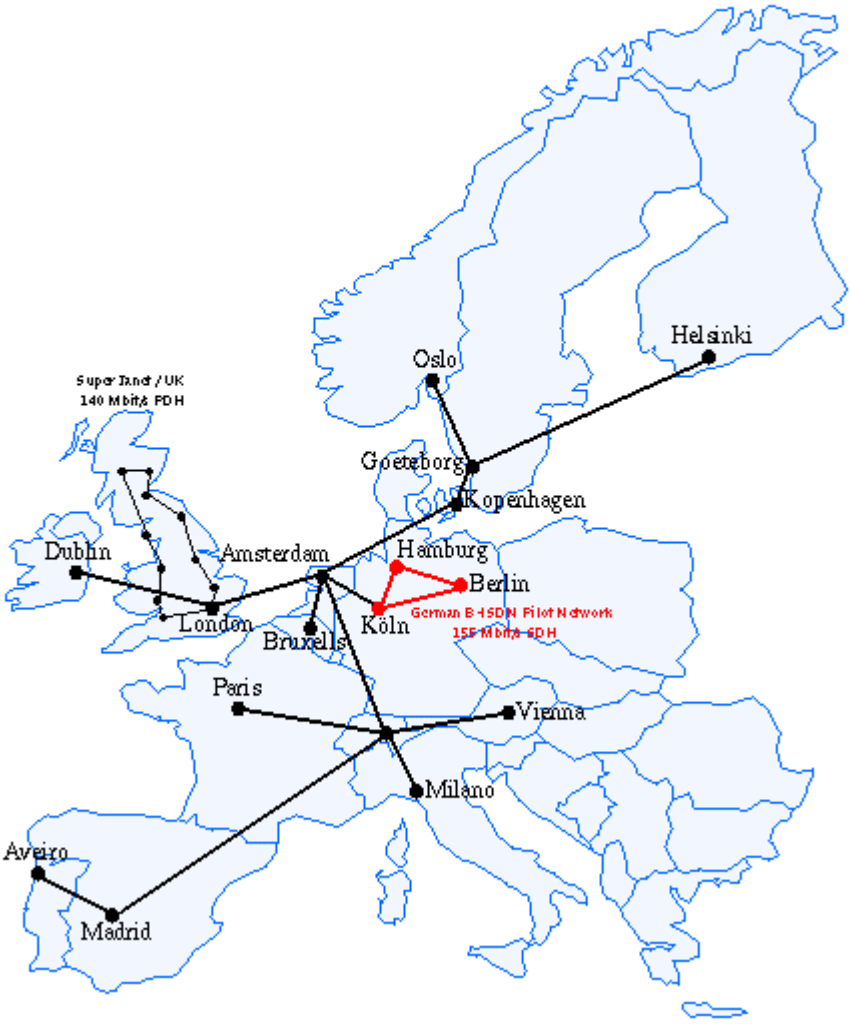
The goal of the network was to exam ATM in a metropolitan area environment. The network consists of three LAN's, one at each university, where there lies a conglomeration of smaller networks which perform a number of tasks depending on the requirements of the site. These LAN's are interconnected via two 155 Mbit/s SONET/SDH connections. One of these is then connected through to Amsterdam where it joins the Pan-European ATM Network. The following diagram shows how the BALI network is linked and how it joins to the rest of Europe. Over the page is a diagram of the Pan-European ATM Network.




At each site the ATM switches are connected via 100 or 140 Mbit/s TAXI interfaces using FDDI's 4B/5B coding schemes. The machines running on these networks are all Hewlett-Packards and workstations from SUN.

The network was tested for use with services such as multimedia mail, multimedia collaboration and joint viewing and teleoperation services. The rates of information was to be tested also by varying the traffic parameters including saturated loads and very bursty traffic.

## The Pan-European ATM Network




There was also a DFN which was implemented over the ATM-trial of the BALI network. This DFN-trial included tests in each of the following areas:

- Interconnection of the local ATM-networks at TU Berlin, DeTeBerkom and GMD.
- Connection to the ATM-trial of the German Telekom.
- Installation of ATM-capable endsystems.
- Provision of routers and gateways between ATM-network and conventional LAN's.
- Connection to available MANs.
- Connection to the ATM-trial of the European PTTs.
- Connection to Telekom's Narrowband-ISDN neetwork.
- Measurement and evaluation of relevant performance parameters.

Overall, with the number of tests that were performed in this Nationwide experiment it could almost be seen as the best possible way of comparing and evaluating ATM technology. Unfortunately no results were given with the document so we cannot analyse the findings. This project

was still worth reporting though as it really demonstrates the scope and potential that ATM technology has [2].

## 5. Satellite ATM Networks

A satellite network has many benefits to offer. It can cover a wide geographical area, it allows flexible network configuration and capacity allocation, has alternative channels for different traffics to maximize resource utilization and has broadcast and multipoint-to-multipoint capabilities and fast network setup. On the surface satellites don't seem to have much to do with clustering technologies, however as they have the potential to offer access to more LANs at a cheaper cost than an ordinary WAN they could possibly give more clustering power.

ASIU (ATM satellite interworking unit) is the key component of the architecture in that it interconnects both ATM and satellite networks and is able to manage and control system resources. It has a disadvantage it has elements which can affect NT performance and so the system should be chosen carefully with consideration to the requirements.

The system has a complex error control mechanism which can handle a large number of differing errors. An interleaving error detection mechanism is used to help stop the problems of burst errors. By using an ATM header error check single bit errors can be detected and are distributed to increase the speed at which they're corrected. An error recovery mechanism is also in place using standard networking techniques such as stop-and-wait ARQ, Go-back-N ARQ, Selective-repeat ARQ. Selective-repeat gives the best performance but is more complex to implement on a cluster system.

Traffic and congestion management is critical throughout the network due to the slow data rates of satellite links. A selective cell implementation iterates through any priority messages and schedules them appropriately. A buffering technique is used to try and store any messages that can't be immediately sent. Prioritization can sometimes be an integral part of a cluster system, depending on the amount of message passing throughout the network.

When considering multimedia a satellite network isn't as efficient due to the low data rates. This can be overcome however by using a dynamic satellite channel and a transponder. Video compression hasn't really been tested thoroughly tested yet to pass judgement but it probably wouldn't run very well unless the connection has been boosted.

Overall, with respect to satellite links as a cluster alternative, it seems to be very good. Cluster projects tend not to deal with much multimedia and so the large transfer rates aren't required. But with computer projects requiring more and more processing power, a satellite ATM network could be a feasible solution to get access to more computing power. The connections are

relatively inexpensive so we may see a growth in the use of satellite ATM networks as a way of clustering a large number of computers and workstations [3].

## 6. Networks Evaluation and Benchmarking Project

The main aim of this project is to give Australian research centers much wider access to distributed computing resources. To do this they will try to demonstrate the utility of dedicated networked access to high-performance resources and quantify these effects in terms of cumulative and peak bandwidth within the latency constraints imposed by inter-city distances in Australia. To cover a broad range of computing requirements a set of long and short term demonstrations are planned.

A set list of objectives are below:

- to investigate the usefulness of dedicated bandwidth for various applications and activities that we believe are becoming crucial to the way scientific research is carried out over wide-area networks.
- to demonstrate this importance to those authorities who have the power to make such dedicated bandwidth available.
- to demonstrate the capabilities of this new technology to other research and development organisations including commercial ones, who may wish to acquire such dedicated bandwidth networks for their own use in the near future.

Demonstrations will involve streaming video and other multimedia devices, parallel computations using PVM and MPI and perhaps some data mining demonstrations using data from the Gravitational Wave Observatory. Overall this project, should it succeed, will provide much better internet access to not only research centers but to large companies and thus could help the general public in the long term. This project is being run by the Department of Computer Science at the University o Adelaide. By visiting the website given in the table above you will also find links to papers evaluating the performance of ATM networks [4].

## 7. Success Stories

The Beowulf project has received major international attention thanks to the ease and cost effectiveness with which techniques they propose allow Interconnection network cluster creation.

The System is based on use of machines running Linux to act as nodes in cluster interconnection networks usually using Ethernet for Communication. The choice of Linux as an operating system provides many advantages thanks to the fact that it’s open source and as such freely available for use and modification.

Further information is available at www.beowulf.org

## 8. Summary and Conclusions

As the use of clusters of computers becomes more prevalent and computing power achievable by clusters becomes greater and greater there becomes a serious need to research the best methods for connecting nodes in the clusters in order to avoid or reduce bottlenecks.

The studies presented give an insight into some of the problems encountered when creating these interconnection networks and ways in which these problems can be tackled and workable solutions developed. Many of the solutions rely on new technologies such as ATM and SCI and hopefully in the future the works in this area will lead to the development of even better tools and methods for creating high speed, low latency interconnection networks.

## References


[1] Burns M, Wallace, B., A performance Evaluation of SCI as a Parallel Computing Interconnection Network, - http://www.hcs.ufl.edu/~wallace/EEL5718/project.html

[2] http://www.prz.tu-berlin.de/docs/html/EANTC/PROJECTS/BALI/index.html

[3] http://giro.ccaba.upc.es/~jmasip/esquemes/ATMsatnt/ATMsatnt.html

[4] http://www.dhpc.adelaide.edu.au/projects/network/aarnet2spec.html

General References used throughout:

[5] Buyya, R., Abramson, D., and Giddy, J., *Nimrod/G: An Architecture for a Resource Management and Scheduling System in a Global Computational Grid*, Proceedings of the 4th International Conference on High Performance Computing in Asia-Pacific Region (HPC Asia'2000), IEEE Computer Society Press, USA.

[6] Berman F. and Wolski R., *The AppLeS Project: A Status Report*, Proceedings of the Eight NEC Research Symposium, Germany, May 1997.

[7] Grid Forum - http://www.gridforum.org/