*CLEI Electronic Journal 3 (1), 2000*

# Low cost cluster architectures for parallel and distributed processing

*Giancarlo C. Mai and César A.F. De Rose*

Catholic University of Rio Grande do Sul, Porto Alegre, Brazil

## Abstract

*Cluster based architectures are standing out in the last years as an alternative for the construction of versatile, low cost parallel machines. This versatility permits their use as much as a teaching tool or as a research environment in the field of parallel and distributed processing. This paper describes some of the possibilities found today on the market for the construction of cluster based parallel machines and proposes different configurations based on cost and application areas.*

***Key Words*** — Computer architectures, parallel and distributed processing, cluster based parallel machines.

## 1 Introduction

Parallel processing systems have become more popular in the last few years due to a growing demand for high performance computing. Unfortunately the systems that offer the processing capacity to satisfy this demand are usually either overly expensive, or too difficult to program, or both.

A high sum of money has been invested in the last years in the research of parallel machines based on clusters, because of their advantages over the three other classes of machines, resulting in a lower cost and more flexible machine that could be configured according to the application used.

The main characteristics of parallel machines based on clusters as well as an overview of the main interconnection technologies for this class of available machines in the market will be presented. Based in these technologies, different architectures are proposed for the creation of a high performance computing laboratory, where teaching as well as the research in the area of parallel and distributed processing can be developed. The cost factor and the possible activities in teaching and in the research are outstanding for each presented architecture.

The configurations and conclusions contained in this article are part of a study that has been accomplished at the Institute of Computer Science of the Catholic University of Rio Grande do Sul - PUCRS (Brazil) for the creation of a high performance computing laboratory.

This paper is organized as follows. Following a brief description of the state of the art in the area of parallel systems for high performance computing, section 3 describes the main characteristics of the cluster based parallel machines and presents its main interconnection technologies for this class of machines available today in the market. Section 4 proposes different configurations the creation of cluster machines. Finally, section 5 presents the conclusions of the paper.

## 2 State of the Art

High performance processing is considered a fundamental tool for the areas of science and technology, because several areas of science require high performance (e.g. molecular biology , chemistry, meteorology). The research and development financed by governments' world-wide demonstrate the strategic importance of this area.

High performance processing depends fundamentally on parallel processing techniques, which are capable to provide the necessary performance to high performance applications [1].

Parallel processing systems have become more popular because of the growing need for powerful computational systems. However, in most cases, these systems are either too expensive, or too difficult to program or both.

Parallel systems available today can be divided into three classes [2][3]:

- Symmetrical Multiprocessors with shared memory (e.g. SGI Power Challenge);
- Massively parallel systems with distributed memory based on high speed network (e.g. Intel Paragon, IBM R6000/SP, Cray T3E, Thinking Machines CM-5).
- Network of workstations (NOW) (e.g. SUN workstations connected by Ethernet).

The three system classes have advantages and disadvantages. For example, the Symmetrical multiprocessors (SMPs) programming is very simple, but these systems have limited size because of their low scalability, thus the bus that interconnect these processors would quickly become the critical point of communication with the increase of the number of processors. Massively Parallel Systems (MPPs) have a good scalability, however they are highly expensive (high speed communication network is required) and very hard to program since there is no shared memory and the communication is by message passing. One of the most important advantages about NOW is its low cost, either for hardware or software. Software like PVM (Parallel Virtual Machine [4]) are available and allow to explore distributed computation for parallel applications without additional costs. However, the network of workstations has low performance of communication.

Networks that use message passing have communication mechanisms usually based on heavy protocols for secure data transference (e.g. TCP/IP). This kind of protocol causes a high latency (the necessary time for sending or receiving short messages), of a magnitude order about three times worse than in the multiprocessors. It is interesting to observe that even in high-speed networks (based on ATM) the communication using TCP/IP protocols still has high latency [5].

An approach that has been used to proliferate more the use of the parallel processing is the adoption of networks of PCs to work in parallel, because PCs have excellent cost/performance rates. Besides, operating systems such as Linux and Windows NT provide reliable support for communication on LANs (Local Area Network). Software like PVM are available for Linux and, more recently, for Windows, therefore, networks of PCs are becoming attractive platforms for parallel processing.

## 3 Cluster Machines

In spite of the presented problems with low performance of communication due to the new technologies of high speed local networks, PCs have become more attractive for parallel processing. With this combination called Cluster Computing, the PC Cluster tries to ally the advantages of the three other classes presented, building parallel machines with the following characteristics:

- Behaves as a network of workstations, the nodes of the network are workstations or normal PCs and they can also be used in conventional applications;
- With the use of high speed communication boards the communication system has a performance close to MPPs (throughput in order of hundreds of Mbytes/s and latency of some few (sec);
- PCs and communication boards are produced in large scale that resulting in a total cost smaller than a MPP, either in the purchase or in the maintenance;

- Shared memory and coherent caches are implemented in some communication boards available in the market, which allow the shared memory programming model as in SMP's, resulting in a easier programming.

In the last years a lot of research has been developed in more efficient mechanisms of communication for cluster machines. This research is based on the development of a dedicated interconnection hardware for network interface that removes the operating system and the software processing in general from the critical line of communication (figure 1). State of the art in the architectures based on clusters has been obtaining communication latencies of the order of few microseconds (μs).

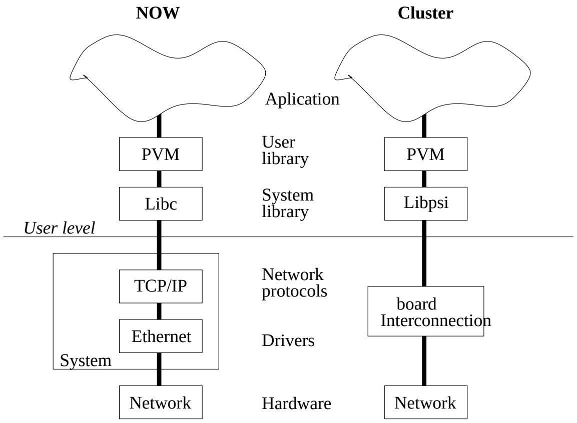

**Figure 3** Communication Layers.

To guarantee a better performance in the cluster machine communication, several high speed interconnection network and protocols are being developed to connect the nodes of these machines, some of them mainly referred in the literature: Myrinet [6], SCI [7][8], and the ParaStation [9].

### 3.1 Myrinet

Myrinet is a new type of high speed local area network based on the technology used for packet communication and switching within MPPs (Massively Parallel Processors). The Myrinet message passing network send and receive data in the form of packets. Any node may send a packet to any other node.

A packet consists of a sequence of bytes starting with a routing header, which is examined by routing circuits that steer the packet through the network. The packet is followed by a payload, which is the data delivered to the destination. However, in contrast with LANs, a Myrinet link is composed of a full-duplex pair of 640 Mb/s channels and is reasonably referred to as a 1.28 Gb/s link in comparison with 100 Mb/s Ethernet (in that Ethernet channels carry packets in only one direction at once). Conventional networks such as Ethernet can be used to built clusters, but do not provide the high performance required for cluster machines. Myrinet can provide high data rate and low latency (about 5 microseconds) communication.

The Myrinet was originally developed to be used in MPP systems. However, due to its characteristics, Myrinet is a cost-effective, high performance, packet communication and switching technology that is widely used to interconnect clusters of workstations or clusters of PCs.

A Myrinet can have an arbitrary topology, but the most common is a two-dimensional mesh. A Myrinet cable can connect two Myrinet host interface boards, or connect a Myrinet host interface board to any port of a switch. A Myrinet cable can also connect two switches.

### 3.2 ParaStation

The programming interface presented by the ParaStation consists of a UNIX socket emulation and widely used parallel programming environment such as PVM. As well as *Myrinet*, *ParaStation* removes the kernel and the common protocols from the communication path. Some initial implementation of ParaStation achieved a communication latency as low as 2 microseconds and a sustained bandwidth of more than 15 Mbyte/s per channel.

The goal of the ParaStation is to support a standardized and efficient programming interface on the top of the network. The ParaStation network is dedicated to parallel applications and is not intended as a replacement for a common LAN, so associated protocols (e.g. TCP/IP) can be eliminated. These properties allow using specialized network features and controlling the network at user level without operating system interaction. The ParaStation protocol software implements multiple logical communication channels on a physical link. In contrast with other high-speed network, as Myrinet for example, in the ParaStation there is not additional cost for components of central switching.

The network topology is based on a two-dimensional mesh. For small systems, a ring topology is sufficient

### 3.3 SCI (Scalable Coherent Interface)

SCI is a recent standard that specifies innovative interconnect hardware and protocols to connect up to 64K nodes (e.g. multiprocessors, PCs, network interfaces) into a high speed network. SCI defines bus-like services. The most notable of these services is a physical 64-bit address space across SCI nodes. Distributed cache coherence protocols for the DSM (Distributed Shared Memory) have been developed, so SCI systems with NUMA (Non Uniform Memory Access) as well as CC-NUMA (Cache Coherence NUMA) [2] characteristics can be built.

SCI avoids the physical limitations of computer busses by employing unidirectional point-to-point links only. Thus, there are no principal impediments to scalability.

The major benefit of SCI network and protocols is that communication can be performed at user level via the hardware DSM, by load and store operations into memory regions mapped from remote memories.. This translates into latencies in the low microseconds range even in a cluster environment.

The basic building blocks of SCI networks are small rings. Large systems are built of rings of rings, interconnected by SCI switches.

## 4 Machine Configurations

The main characteristics of our design are the way it separates safety, functionality, and efficiency concerns among a set of CA actions, which thus can be designed, and v In this section three configurations are proposed for the construction of a cluster machine for a high performance laboratory. This laboratory should be able to support teaching and research activities in the field of parallel and distributed processing.

The main goal is to take advantage of the cluster architecture and use the available resources as processing nodes to build the machine (PC´s or Workstations), investing just in its interconnection. The three interconnection technologies presented in section 3 were used here to interconnect the machine nodes in three different machine configurations. If extra resources are available, they can be invested to increase the power of processing nodes, using last generation workstations with multiple processors. Linux is recommended as the operating system because it is free of charge and its code can be altered if necessary.

The three proposed configurations will be analyzed regarding their programming potential, possible uses as a teaching tool and a research environment and their cost, always indicated in American currency (US $). Considerations about the number of nodes are not made, because the main goal is the academic use o the machines in teaching and research activities and not the performance aspect. The configurations presented will always have 4 nodes and expansion possibilities will be indicated where available.

The workstation used to connect the parallel machine to the external world is called the host and is not considered as being part of the parallel machine in the proposed configurations. This machine is responsible for all the parallel machine’s I/O including loading processes and resource monitoring. These tasks imply a considerable load to this machine but it can be still used as a processing node.

### 4.1 Minimum Configuration

Figure 2 presents what can be called a minimum configuration for a cluster machine. In this case a low latency Fast-Ethernet switch is used for the interconnection of machine nodes.

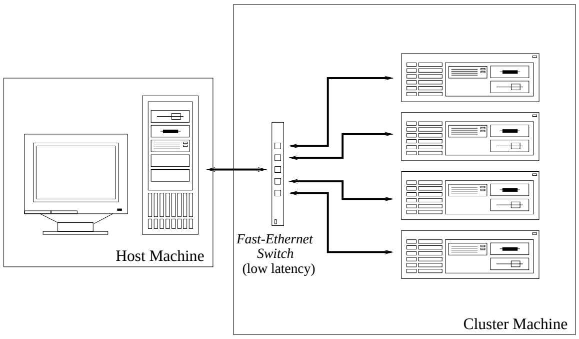

**Figure 3** Minimum configuration

It is important to stand out that in spite of the similarity to a conventional Ethernet local network, the special switch used in this configuration guarantees a reduced communication latency, through the emulation of several point-to-point connections among machine nodes. This is decisive to classify this configuration as a cluster and not as a NOW (Network of Workstations). The machine nodes are connected to the switch with Fast-Ethernet network cards, with a nominal throughput of 100 Mb/s. The use of conventional network cards implies in software implementation of the network layers (Figure 1), what compromises the communication latency. In the other configurations proposed in this work, network layers are implemented in hardware, reducing the latency significantly.

Table 1 presents the cost involved in building a cluster machine with 4 nodes based on the minimum configuration. The switch used has 8 ports allowing a machine expansion up to 8 nodes with the purchase of more network cards. Switches with more ports are also available, and can be used for greater clusters.

| **Description** | **Unit Cost** | **Quantity** | **Cost** |
|---|---|---|---|
| Fast-Ethernet Network Card | 80 $ | 5 | 400 $ |
| Cable (twisted pair, 2 m) | 4 $ | 5 | 20 $ |
| 8 port Fast-Ethernet low latency Switch | 2500 $ | 1 | 2500 $ |
| | | **Total Cost** | **2920 $** |

**Table 1:** Minimum configuration cost with 4 nodes (US$)

These machines can be programmed with parallel programming libraries like MPI and PVM that are available for the Linux operating system free of charge. Both work with message passing that adapts well the distributed memory of this configuration.

This configuration, in spite of its simplicity, already allows the exploration of the parallel and distributed programming paradigms can be explored in teaching activities. With the use of a load monitor in the host, that constantly shows the load of the processing nodes and the message traffic in the machine, topics like application modeling and performance, and load balancing in parallel applications can be better visualized by students.

The implementation of a distributed global memory in software and its implications is a possible research subject in this configuration. Another interesting subject is the implementation of the load monitor previously mentioned and its implications in the operation of the machine, since monitor and applications messages share the same communication channels.

### 4.2 Basic Configuration

Figure 3 presents what can be called a basic configuration for a cluster machine. In this case a low latency network is used for the interconnection of machine nodes.

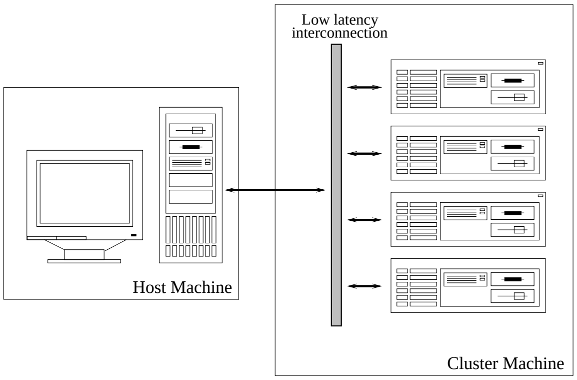

**Figure 3:** Basic configuration

In this configuration the use of a Myrinet or ParaStation interconnection board is suggested. Figure 4 shows two interconnection possibilities depending on the chosen board. The Myrinet board requires a central switch which is connected to all nodes with point-to-point cables (Figure 4a). The ParaStation boards do not require any additional hardware and can be interconnected directly with point-to-point cables. For a small number of nodes (2-10) the recommended interconnection topology is the ring (Figure 4b).

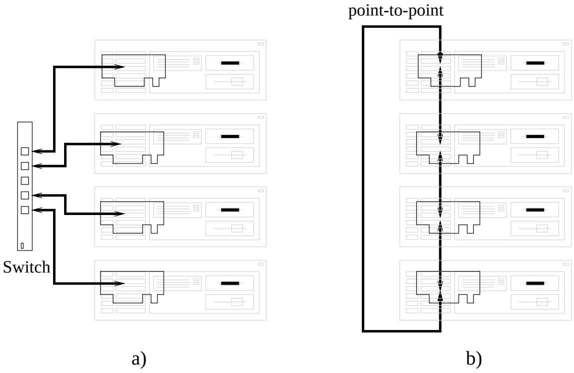

**Figure 4:** Interconnection possibilities

The main difference to the minimum configuration is that the network layers are implemented in hardware (in the board), and not in software as in the previous configuration, improving the communication latency. With this special interconnection boards, latency is reduced to few microseconds, considerably smaller than the previous configuration (depending on the used switch witch, around 20 microseconds).

Tables 2 and 3 presents the cost involved in building a cluster machine with 4 nodes based on the basic configuration depending on the interconnection board. The Myrinet switch in table 2 has 8 ports allowing a machine expansion up to 8 nodes with the purchase of more network cards. Switches with more ports are also available, and can be used for greater clusters.

| **Description** | **Unit Cost** | **Quantity** | **Cost** |
|---|---|---|---|
| *Myrinet* SAN Card with PCI interface (cables included) | 1300 $ | 5 | 6500 $ |
| 8 port *Myrinet* SAN Switch | 6000 $ | 1 | 6000 $ |
| | | **Total Cost** | **12,500 $** |

**Table 2:** Basic configuration cost with 4 Myrinet nodes (US$)

| Description | Unit Cost | Quantity | Cost |
|---|---|---|---|
| *Parastation* Card with PCI interface (cables included) | 1400 $ | 5 | 7000 $ |
| | | **Total Cost** | **7000 $** |

**Table 3:** Basic configuration cost with 4 ParaStation nodes (US$)

As in the previous configuration, these machines can be programmed with parallel programming libraries like MPI and PVM that are available for the Linux operating system free of charge. Both work with message passing that adapts well the distributed memory of this configuration.

Because of the distributed memory, the implementation of a distributed global memory in software and its implications is here also a possible research subject in this configuration. All of the teaching and research activities of the previous configuration can be utilized here as well. Because of the similar latency values, performance comparisons with MPP´s are already possible with these configurations.

### 4.3 Advanced Configuration

Figure 3 presents what can be called a basic configuration for a cluster machine. In this case a low latency network is used for the interconnection of machine nodes.

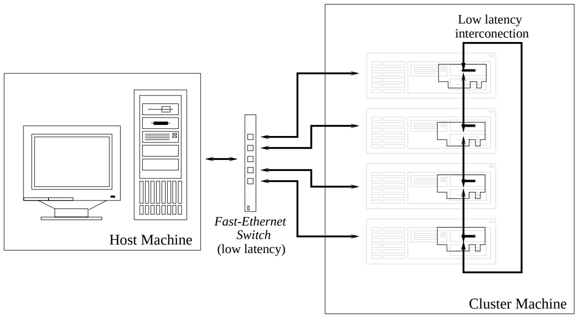
Low latency interconection
Host Machine
Fast-Ethernet Switch (low latency)
Cluster Machine

**Figure 3:** Advanced configuration

In this configuration the use of a Myrinet or ParaStation interconnection board is suggested. Figure 4 shows two interconnection possibilities depending on the chosen board. The Myrinet board requires a central switch which is connected to all nodes with point-to-point cables (Figure 4a). The ParaStation boards do not require any additional hardware and can be interconnected directly with point-to-point cables. For a small number of nodes (2-10) the recommended interconnection topology is the ring (Figure 4b).

The main difference to the minimum configuration is that the network layers are implemented in hardware (in the board), and not in software as in the previous configuration, improving the communication latency. With this special interconnection boards, latency is reduced to few microseconds, considerably smaller than the previous configuration (depending on the used switch witch, around 20 microseconds).

Tables 2 and 3 presents the cost involved in building a cluster machine with 4 nodes based on the basic configuration depending on the interconnection board. The Myrinet switch in table 2 has 8 ports allowing a machine expansion up to 8 nodes with the purchase of more network cards. Switches with more ports are also available, and can be used for greater clusters.

As in the previous configuration, these machines can be programmed with parallel programming libraries like MPI and PVM that are available for the Linux operating system free of charge. Both work with message passing that adapts well the distributed memory of this configuration.

| Description | Unit Cost | Quantity | Cost |
|---|---|---|---|
| *SCI* Card with PCI interface (cables included) | 1600 $ | 4 | 6400 $ |
| Fast-Ethernet Network Card | 80 $ | 5 | 400 $ |
| Cable (twisted pair, 2 m) | 4 $ | 5 | 20 $ |
| 8 port Fast-Ethernet low latency Switch | 2500 $ | 1 | 2500 $ |
| | | **Total Cost** | **9,320 $** |

**Table 3:** Advanced configuration cost with 4 *SCI* nodes and a low-latency *Fast-Ethernet* Switch (US$)

Because of the distributed memory, the implementation of a distributed global memory in software and its implications is here also a possible research subject in this configuration. All of the teaching and research activities of the previous configuration can be utilized here as well. Because of the similar latency values, performance comparisons with MPP´s are already possible with these configurations.

## 5 Conclusions

In this paper cluster architectures are presented as an alternative to the construction of parallel machines. The resulting cluster brings together the advantages of the other three machine classes (MPP, SMP and NOW), providing a versatile low cost parallel machine. Due to the fact that the new local network interconnection technologies are the driving force behind this class of machine, three of these low latency interconnection schemes, Myrinet, ParaStation and SCI, are described and compared. In all three, low latency is achieved by implementing communication software layers in hardware and with fast point-to-point links among machine nodes.

Based on this emerging machine class, three machine configurations are proposed for a parallel and distributed processing’s laboratory. These configurations are analyzed regarding their research and teaching potential, and the parts needed to assemble the system are listed with quantities and costs for a mid range parallel machine.

The proposed machine configurations are ranged from what was called the minimum configuration, which uses a special low latency Fast-Ethernet switch for node interconnection, to an advanced configuration, where two interconnection networks are used. The advanced configuration is able to eliminate system messages interference in the application traffic, since one low latency Fast-Ethernet switch is used for I/O and system services and a ring interconnected SCI boards are dedicated to the communication among the parallel application processes. The SCI board was chosen in the advanced configuration for being the most versatile of the three technologies, once it allows both communication paradigms, shared memory and message passing.

It is important to emphasize that cluster technology is still in an evolution stage. Only few manufacturers are producing what could be called the first generation of low latency interconnection boards, and new standards for these boards are still being developed. In researching and implementing programming environments and operating systems for this architecture, the scientific community will trace the development path of this emerging class of parallel machines.

### Acknowledgements

This work was partially supported by FAPERGS and PUC-RS.

### References

[1] Ted G. Lewis; H. El-Rewini. **Introduction to Parallel Computing**. Prentice-Hall International, Englewood Cliffs, 1992

[2] Kai Hwang. **Advanced computer architecture: parallelism, scalability, programmability**. MacGraw-Hill Series in Computer Science. McGraw-Hill, 1993.

[3] Albert Y. Zomaya, editor. **Parallel and Distributed Computing Handbook**. McGraw-Hill, New York, 1996

[4] Al Geist, et al. **PVM: Parallel Virtual Machine**. Cambridge, MA:MIT Press, 1994.

[5] T.von Eicken, A.Basu, V. Buch. Low-Latency Communication Over ATM Networks Using Active Messages. **IEEE Micro**, Feb.1995, 46-53

[6] C. L. Seitz et. al. Myrinet - A Gigabit-per-Second Local-Area Network. **IEEE-Micro**, vol. 15, n. 1, February 1995, pp. 29-36.

[7] IEEE: IEEE Standart for Scalable Coherent Interface (SCI). **IEEE standart 1596-1992**, New York, 1993

[8] H. Hellwanger, W. Karl, M. Leberecht. **Enabling a PC Cluster for High Performance Computing**. LRR-TUM Muenchen, 1998

[9] T. M. Warschko et. al. The ParaStation Project: Using Workstations as Building Blocks for Parallel Computing. **In proceedings of the International Conference on Parallel and Distributed Processing, Techniques and Aplications (PDPTA "96)**, Aug. 1996, CA, vol 1, pp. 375-386.

[10] C. A. F. De Rose and P. A. O. Navaux. Um modelo distribuído para a gerência de processadores em multicomputadores. **Anais do IX Simpósio Brasileiro de Arquitetura de computadores e processamento paralelo – SBAC/PAD**. Campos do Jordão, SP, 1997.

[11] Cristiana Amza et. al. TreadMarks: Shared Memory Computing on Networks of Workstations. **ACM Computer**, 1995

[12] Arthur Dumas. **Programming WinSock**. Sams Publishing, 1995.