CHOOSING THE RIGHT HARDWARE FOR THE BEOWULF CLUSTERS CONSIDERING PRICE/PERFORMANCE RATIO

Muhammet ÜNAL
e-mail: muhunal@gazi.edu.tr

Selma YÜNCÜ
e-mail: syuncu@gazi.edu.tr

Gazi University, Faculty of Engineering and Architecture, Department of Electrical & Electronics Engineering
06570, Maltepe, Ankara, Turkey

Key words: Computer Hardware Architecture, Parallel Computers, Distributed Computing, Networks of Workstations (NOW), Beowulf, Cluster, Benchmark

ABSTRACT

In this study, reasons of choosing Beowulf Clusters from different types of supercomputers are given. Then Beowulf systems are briefly summarized, and some design information about the hardware system of the High Performance Computing (HPC) Laboratory that will be build up in Gazi University Electrical and Electronics Engineering department is given.

I. INTRODUCTION

The world of modern computing potentially offers many helpful methods and tools to scientists and engineers, but the fast pace of change in computer hardware, software, and algorithms often makes practical use of the newest computing technology difficult. Just two generations ago based on Moore's law [1], a plethora of vector supercomputers, non-scalable multiprocessors, and MPP clusters built from proprietary nodes and networks formed the market. But none of them can survive up to date. Within the past three years there has been a rapid increase in the deployment and application of computer cluster to expand the range of available system. Capabilities beyond those of conventional desktop and server platforms: Clustering is a powerful concept and technique for deriving extended capabilities from existing classes of components in nature. A parallel system must have several properties: one of these properties is scalability which makes system cost effective and easyly programmable.

IMPORTANCE OF SCALABLE COMPUTER ARCHITECTURES

A common feature of modern computer is parallelism. Even within a single processor, parallelism has been exploited in many ways. More recently, the concept of scalability has been reemphasized, which includes parallelism.

Hwang and Xu [2] made a definition of stability as:
A computer system, including all its hardware and software resources, is called scalable if it can scale up to accommodate ever-increasing performance and functionality demand and/or scale down to reduce cost.

Scalability requires:

- Functionality and Performance: The scaled-up system should provide more functionality or better performance.
- Scaling in Cost: The cost paid for scaling up must be reasonable.
- Compatibility: The same components, including hardware, system software, and application software should still be usable with little change.

II. SYSTEMS PERFORMANCE ATTRIBUTES

Basic performance attributes are summarized in table 1. Machine size n is the number of processors in a parallel computer. Workload W is the number of computation operations in a program P_{peak} is the peak speed of a processor.

Terminology	*Notation*	Unit
Machine Size	n	Dimensionless
Clock Rate	F	Mhz.
Workload	W	Mflop
Sequential Execution Time	T_1	S
Parallel Execution Time	T_n	S
Speed	$P_n=W/T_n$	Mflop/s
Speedup	$S_n=T_1/T_n$	Dimensionless
Efficiency	$E_n=S_n/n$	Dimensionless
Utilization	$U_n=P_n/(nP_{peak})$	Dimensionless
Startup time	t_0	µs
Asymptotic bandwidth	r_∞	MB/s

Table 1. Performance Attributes of Parallel System

III. PHYSICAL MACHINE MODEL

Large-scale computer systems are generally classified into six practical machine models:

The single-instruction multiple-data (SIMD) machines, the parallel vector processors (PVP), the symmetric multiprocessors (SMP), the massively parallel processors (MPP), the cluster of workstations (COW), and the distributed shared memory (DSM) multiprocessors. SIMD computers are used mostly for special-purpose applications. The remaining models are all MIMD machines as summarized in Table 2

Attributes	PRAM	PVP/SMP	DSM	MPP/COW
Homogeneity	MIMD	MIMD	MIMD	MIMD
Synchrony	Instruction-level synchronous	Asynchronous or loosely synchronous	Asynchronous or loosely synchronous	Asynchronous or loosely synchronous
Interaction Mechanism	Shared variable	Shared variable	Shared variable	Message Passing
Address Space	Single	Single	Single	Multiple
Access Cost	UMA	UMA	NUMA	NORMA
Memory Model	EREW, CREW or CRCW	Sequental Consistency	Weak ordering is widely used	N/A
Example Machines	Theoretical model	IBM R50, Cray T-90	Stanford DASH, SGI Origin 2000	Cray T3E, Berkeley NOW, Beowulf

Table 2. Semantic Attributes of Parallel Machine Models.

The boundary between MPP's and COW's is becoming fuzzy these days. Clusters have many cost / performance advantages over MPP's. So clustering is becoming a trend in developing scalable parallel computers as what we are also going to do so.

IV. BASIC CONCEPTS OF CLUSTERING

A cluster is a collection of complete computers (nodes), which are physically interconnected by a high-performance network or a local-area network (LAN). Typically each computer node is an SMP server, or a workstation, or a personal computer. More importantly, all cluster nodes must be able to work together collectively as a single integrated, computing resource; in addition to filling the conventional role of using each node by interactive users individually. A conceptual architecture of a cluster is shown in Fig 1.

Five important architectural concepts are merged into a cluster, as an interconnected set of whole computers (nodes) that work collectively as a single system to provide uninterrupted (availability) and efficient (performance) services.

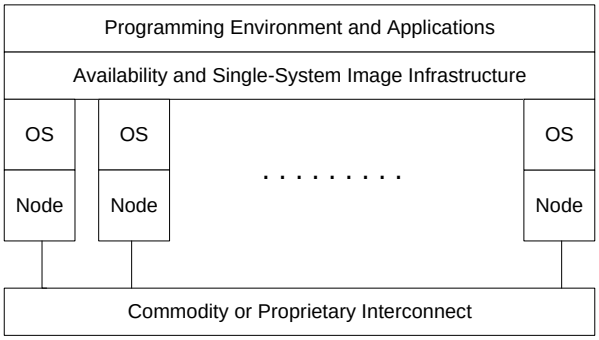

Figure 1. Typical architecture of a cluster of multiple computers.

V. BEOWULF CLUSTERS

Famed was this Beowulf: far flew the boast of him, son of Scyld, in the Scandian lands. So becomes it a youth to quit him well with his father's friends, by fee and gift, that to aid him, aged, in after days, come warriors willing, should war draw nigh, liegemen loyal: by lauded deeds shall an earl have honor in every clan. Beowulf is the earliest surviving epic poem written in English. It is a story about a hero of great strength and courage who defeated a monster called Grendel. [9]

As the performance of commodity computer and network hardware increase, and their prices decrease, it becomes more and more practical to build parallel computational systems from off-the-shelf components, rather than buying CPU time on very expensive Supercomputers. In fact, the price per performance ratio of a Beowulf type machine is between three to ten times better than that for traditional supercomputers. Beowulf architecture scales well, it is easy to construct and you only pay for the hardware as most of the software is open-source and free. So software engineering and programming is easier than ever.[3,6-8]

VI. HPC LABORATORY

In Gazi University Electrical & Electronics Engineering Department, a high performance computing (HPC) laboratory is planned to be founded. It will be used for research purposes in several research aspects. The Project is funded by Gazi University Scientific Research Projects Department.

VII. ENABLING TECHNOLOGIES

Computing with a Beowulf cluster engages four distinct but interrelated areas of consideration.

1) Hardware system structure
2) Resource administration and management environment
3) Distributed programming libraries and tools
4) Parallel algorithms

Hardware system structure encompasses all aspects of the hardware node components and their capabilities, the

dedicated network controllers and switches and interconnection topology that determines the system's global organization. The resource management environment is the battery of system software and tools that govern all phases of system operation from installation, configuration, and initialization, through administration and task management, to system status monitoring, fault diagnosis, and maintenance. The distributed programming libraries and tools determine the paradigm by which the end user coordinates the distributed computing resources to execute simultaneously and cooperatively the many concurrent logical components constituting the parallel application program. Finally, the domain of parallel algorithms provides the models and approaches for organizing a user's application to exploit the intrinsic parallelism of the problem while operating within the practical constrains of effective performance. Hardware system structure will be covered briefly below.

VIII. HARDWARE SYSTEM STRUCTURE

The most visible and discussed aspect of cluster computing systems are their physical components and organization. The two principal subsystems of a Beowulf cluster are its constituent compute nodes and its interconnection network that integrates the nodes into a single system. Design schematic of our hardware system is given in Figure 2.

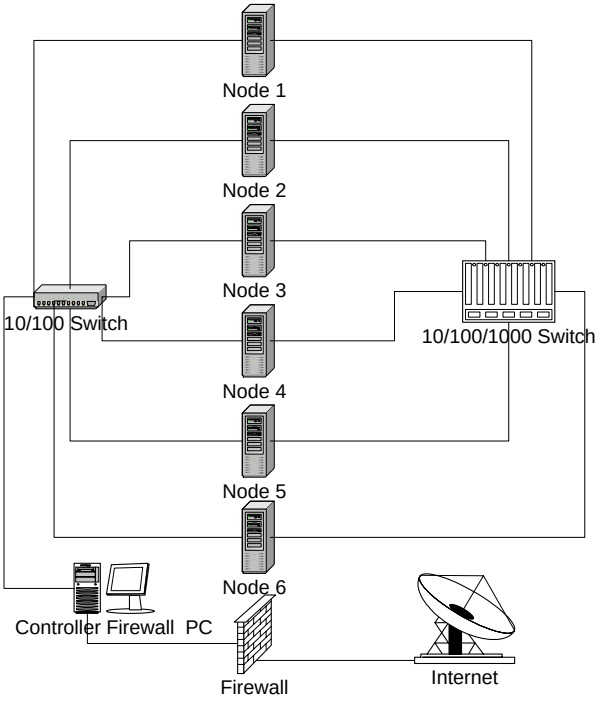

Figure 2. Design schematic of hardware system of Beowulf Cluster of Gazi University Electrical and Electronics Engineering Department HPC Laboratory.

BEOWULF COMPUTE NODES

The compute or processing nodes incorporate all hardware devices and mechanisms responsible for program execution, including performing the basic operations, holding the working data, providing persistent storage, and enabling external communications of intermediate results and use command interface. Five key components make up the compute node of a Beowulf cluster: the microprocessor, main memory, the motherboard, secondary storage, and packaging.

- The Microprocessor provides the computing power of the node with its peak performance measured in Mips(millions of instructions per second) and Mflops(millions of floating-point operations per second). While choosing the microprocessor the Mflops is important for the scientists because they use it for complex arithmetic operations. In our project, Athlon XP with Barton core(512Kb L2 cache) will be used as the microprocessor of our Beowulf system, because it has 3 floating-point pipeline besides Intel P4 has only 2 floating-point units(FPU) [10,11]. Although Itaniums have 4 FPU and also they have 64 bits architecture, but their clock frequency is slow and unfortunately there is no reseller in Turkey yet. In real, Athlon microprocessors' working clock rate is *f* 2133 Mhz [10] while they are named as 2800+. The peak performance per processor will be 2133x3=6400 Mflops, while Intel's P4 microprocessors' peak performance will be 3066x2=6132 Mflops.[11] And the cost of the microprocessors doubles the cost of Athlons [12,14].
- Main Memory stores the working data set and programs used by the microprocessors during job execution. We will going to use at least 1GB Double Data Rate (DDR) 266(133x2)Mhz SDRAM, in each node to provide enough space for the microprocessors to work comfortably.
- The Motherboard is the medium of integration that combines all the components of a node into a single operational system. Far more than just a large printed circuit board, the motherboard incorporates a sophisticated chipset almost as complicated as the microprocessor itself. The chipset manages all the interfaces between components and controls the bus protocols. Dual CPU system is going to be used for performance and cost. Unfortunately there is only one chipset, manufactured by AMD for dual Athlon system and that chipset (AMD-760 MPX) [13] will be used for this purpose. Also there are a lot of motherboard manufacturers, so the selection is difficult. One of the important issues is bios update support. After making some internet search bios update support is only provided by MSI [15] and Tyan [16] motherboards, which support AMD Athlon XP 2800+ CPUs. But Tyan works only with registered DDR SDRAMs and that makes system costlier. MSI K7D-L Master was chosen as motherboard. One of the reasons to buy this board is it will give us chance to change the clock ratio and CPU multiplier if

the CPU's are unlocked [17]. It will be very important in our research because we will going to test the system to find out the overheads of the system in different clock-rate speeds.

- Secondary storage provides high capacity persistent storage. While memory loses all its contents when the system is powered off, Secondary stage fully retains its data in the power down state. EIDE hard disks are going to be used for this purpose. In design of the system, each node is going to have a raid card to supply enough bandwidth as SCSI and this will be a cheaper solution [5,14]. Each node will have dual 8 MB cached 7200 RPM 80x2=160 MB storage [18]. This is very important when working with large data sets. As an example a 3D matrix with 100000 elements in each dimension will require a huge storage space.
- Packaging for PC's originally was in the form of tower cases. These cases must be cooled well for proper operation of the system. Another important issue is about the electricity system in Turkey. Line interactive uninterrupted power supply (UPS) is going to be used for each node to provide protection and fail free system.

INTERCONNECTION NETWORK

Without the ability of moderate cost short hand network technology, Beowulf cluster computing would never have happened. Ethernet was developed as a local area network for interconnecting distributed single user and community computing resources with shared peripherals and file servers. A network is a combination of physical transport and control mechanisms associated with a layered hierarchy of massage encapsulation

The network is characterized primarily in terms of its bandwidth and its latency. Bandwidth is the rate at which the massage bits are transferred usually cited in terms of peak throughput as bits per second. [19]

Latency is the length of time required to send the message. Both bandwidth and latency is important [4]. The HPC system is planned to have two individual networks. One of them will be Fast Ethernet(100 Mbit) used for controlling, managing and other purposes, while the other Gigabit Ethernet(1000 Mbit) SAN(System Area Network) will be dedicated to running parallel programs which share data by MPI (Message Passing Interface) or PVM (Parallel Virtual Machines) or any other method. Also this way will give opportunity to try different protocols types in the networks.
The system will be connected to internet via a firewall on the fast Ethernet network. So the users from all over the world will connect to our supercomputer using SSH telnet connection and line up to run their programs easily with this project.

IX. CONCLUSION

After the hardware system is settled up, benchmark programs bill be run to find behaviors and problems on the system. By changing clock rates, network throughputs and latencies, we will test the behaviors of Beowulf clusters experimentally. A speed-up of 10 to 12 is expected from the fastest computer sold nowadays. This system will be pioneer in our university. After we will have finished this project successfully we wish to start a new project to build up a supercomputer that -we wish- will be in the top 500 supercomputer list of the world.

REFERENCES

1. G. Bell, "Observation on Supercomputing Alternatives: Did the MPP Bandwagon Lead to a Cul-de-Sac?", Communications of the ACM 39, no.3 (March 1996) 11-15, 1995.
2. K. Hwang, Z. Xu, "Scalable Parallel Computing Technology, Architecture, Programming", Mc Graw-Hill, 1998.
3. Ridge, D., Becker, D., Merkey, P., Sterling T., "Beowulf: Harnessing the Power of Parallelism in a Pile-of-PCs" Proceedings, IEEE Aerospace, 1997
4. Reschke, C., Sterling, T., Ridge., Savarese, D., Becker, D., Merkey, P., "A Design study of alternative Network Topologies for the Beowulf Parallel Workstation", Proceedings, High Performance and Distributed Computing, 1996.
5. Sterling, T., Becker, D., Savarese, D., Berry, M., Res, C., "Achieving a Balanced Low-Cost Architecture for Mass storage Management through Multiple Fast ethernet Channels on the Beowulf Parallel Workstation", Proceedings, International Parallel Processing Symposium, 1996.
6. Becker, D., Sterling, T., Savarese, D., Dorband, J., Ranawak, U.A., Packer, C.V., "BEOWULF: A Parallel Workstation for Scientific Computation", Proceedings, International Conference on Parallel Processing, 1995.
7. http://www.beowulf.org
8. http://www.netlib.org/
9. http://legends.dm.net/beowulf/index.html
10. Advanced Micro Devices, Inc., "AMD Athlon MP Processor Model 10 Data Sheet for Multiprocessor Platforms", 2003.
11. Intel Corporation,Intel® Xeon™ Processor with 533 MHz Front Side Bus at 2 GHz to 3.06 GHz", 2003.
12. http://www.mavibilgisayar.com
13. Advanced Micro Devices, Inc., "AMD-760™ MPX Chipset Overview", December 2001.
14. http://www.tomshardware.com
15. http://www.msi.com.tw
16. http://www.tyan.com
17. http://fab51.fc2web.com/pc/mother/index.html
18. http://www.wdc.com
19. R. Çölkesen, B. Örencik, "Bilgisayar Haberleşmesi ve Ağ Teknolojiler", Papatya Yayıncılık, Haziran 1999.