

PRELIMINARY COMPREHENSIVE CLUSTER REPORT

The Khwarzimidic Science Society 5 Feb 2000

Submitted by:

Jawad Mahmood

Lead Developer Khwarzimidic Cluster Team

supercompute@khwarziic.org

Bilal Muddassir

Resource Lead Khwarzimidic Cluster Team

noorh@brain.net.pk

Submitted to:

Dr. Saadat Anwar Siddiqi Project Supervisor and President

Khwarzimidic Science Society

saadat@khwarzimidic.org

Feb 5 2000

<http://khwarzimidic.org/beowulf/report01.pdf>

Contents

1. Introduction
2. Progress Made
3. Problems Faced
4. The Requirements
5. Future Directions
6. Conclusions

1. Introduction.

This preliminary report comprehensively summarises the first phase of the Beowulf Cluster Project initiated by the Khwarzimidic Science Society. It methodically mentions the problems faced and the progress made as well as recommendations and requirements of the future. In short, it can be epitomised that the Cluster has been successfully conceived by our Cluster Developer (Mr. Jawad Mahmood), myself and the rest of our Team. The infrastructure has been laid down and future progress is imperative.

2. Progress Made.

2.1 After an initial donation from extra sources, to the Khwarzimidic Science Society and partial funding from the Khwarzimidic Science Society, 3 Celeron 400 processors, 3 mainboards with 32 MB RAMs, Network interface cards and 1 FDD have been purchased.

2.2. The necessary workstation and network hardware has been deployed to give final shape to a two-node

Cluster as of now. The 100baseT LAN is operational through a direct link between the two BayLan 100BaseT Ethernet NICs.

2.3. The RedHat 5.1 OS is up and running without any immediate problem on the master node. The slave node is up and running after successful boot off the floppy disk. The slave node is successfully mounting the disk filesystem off the master's hard disk via NFS.

2.4. The Cluster is running PVM. We have demonstrated the run-time of a couple of pre-compiled PVM-enabled programs for the purpose of comparison. We could see that the single-CPU run-time took fairly large amount of time to complete against the Cluster run-time. With an addition of only 1 node, we observed a 180 % increase in operation time for a very simple code.

2.5. We have developed a very sound theoretical understanding of many hardware and software processes for our Cluster, including diskless workstation bootup khwarzimidic.org/beowulf/diskless.pdf, channel bonding, PVM and MPI libraries that can be used for benchmarking and complex codes that can be run and tested on the Cluster as well as code design to enhance Cluster gain.

2.6. Moreover, our international correspondence has revealed a number of interesting and worthwhile ideas. We have found many useful precompiled codes in the fields of molecular modeling, climatology modeling, image processing etc., but most of them can only be optimised for Clusters larger than the one we have established now. We are also looking into the possibilities of database management with our Cluster in administrative and government offices for the social sector.

2.7. We are also working on a research paper titled: "Beowulf Clusters for a Socially Developing Environment" that is intended to be presented at the International Distributed Computing Conference arranged by the Tokyo Institute of Technology at Islamabad. We have also received announcements for other national and international conferences and are looking into the possibilities for presenting our recent findings in the pertinent fields.

3. *Problems Faced.*

Upto now, the main problem encountered was in setting up the diskless system. There was only one FDD available and there were a lot of non-preconceived problems in the scripts and methods discussed in the diskless *how-to* documents that were available with us. So we had to devise our own method.

4. *The Requirements.*

4.1. We fervently suggest that the Cluster needs a 100baseT Ethernet hub. We cannot hook up all the four nodes through a cable because the NICs do not have cable interfaces. Only RJ-45 interfaces are provided that can only be linked through either a hub or a switch. Hence only two nodes are operational at the moment. Before going on with further code-evaluation or benchmarking, we must complete the whole Cluster.

4.2. A 100baseT Ethernet Switch is certainly preferable. Only a switch would allow simultaneous talks between pairs of hosts on the Cluster. Otherwise with a hub, which merely emulates a shared network cable, only one pair of nodes can talk at one time. Therefore we again recommend the *immediate* purchase of a 100baseT Ethernet Switch.

4.3. At the time of up-scaling the Cluster, we must replace the NICs we are currently using with better models. People all around the world prefer high-end NICs for large clusters. High-end favourably translates to the ability to run in full-duplex mode. Full-duplex mode means a NIC can send as well receive Ethernet frames to and from the physical layer at the same time. This increases the communication speed and hence the overall processing power of the cluster. Currently only two companies are popular among the Beowulfers. DEC for their Tulip family and Intel for their IntelXPress family. So we recommend the future purchase of full duplex NICs.

4.4. We must also think about using bootable NICs. We must discard the floppy-boot method of booting the slave nodes. People have written scripts and binaries that can help us in making bootable NICs. They are available as different packages. One such package is called Etherboot.

4.5. For future upscaling of the Cluster to achieve supercomputing speeds we must seek satisfactory funding and grants. We are already looking into the trajectory that we must adopt to attain speeds in the order of 1 billion floating point operations per second.

5. *Future Directions.*

5.1. Our first priority is to make the 4 node - Cluster ready and complete hardware-wise with an *Ethernet Switch and bootable full-duplex NICs* as mentioned in 4.2., 4.3. and 4.4.

5.2. After the Cluster is all-node ready, we must benchmark it using some standard tool. We have been receiving praises on the Beowulf mailing list about a tool called Linpack which is available to us. We can start using it as soon as the Cluster is ready. Moreover we shall use the following libraries for code-testing and benchmarking: LIPS, MPICH, Linda/BaLinda, Aurora, Arunja and SpeedesComm.

5.3. Indigenous software development for specific scientific calculations of *Pakistan's* science community is a matter of thinking over. We must engage programmers of parallel code for this purpose. For this purpose we must contact the software houses and the institutes teaching theoretical computer science for volunteer programmers. We must engage people who can write customized code for the Cluster. Only this way the Cluster would better serve its purpose. Many undergraduate and graduate level projects in code development can be designed out of this once the Cluster is ready.

5.4. We suggest the use of Java for a parallel computing environment as ours. At the present, Java, due to many of its inherent properties, such as inability to allow direct memory access, and so forth is unable to throttle a Cluster environment. But many people around the world have taken up projects to investigate this issue and they are trying to make Java compatible with parallel-computing schemes.

5.5. We are already considering the possibility of providing authorised access to our Cluster through an HTTP interface. Java could also be

useful in this regard or we can have some other hybrid interface for the Cluster. The Cluster cannot reach everybody everytime. We believe that it should be placed on a nation-wide network and people should be able to submit their parallel jobs via the above-mentioned interface.

6. *Conclusions.*

In the end, the Khwarzimidic Beowulf Cluster Team envisages that the Cluster must have its seeds in the country's soil and seek to fulfill its social needs. Towards this end, we should start asking various scientific organizations in our country if they have problems that they think need large computational power and give them the opportunity to talk to us. Moreover, in the long run, we can reap many social and administrative benefits out of the Cluster in terms of management of huge databases and statistical analysis thereof.