AN OVERVIEW OF THE AGILA BEOWULF CLUSTER

WILLIAM EMMANUEL S. YU

Abstract. This paper aims to present a rapid outline of the design, installation, construction and configuration of the Ateneo High Performance Computing Group’s AGILA Cluster. A brief summary of its current features and configuration will be presented. Some suggestions for improvement will also be proposed. (Sept. 21, 2000)

1. Hardware Configuration

The AGILA cluster was constructed with commodity-off-the-shelf hardware and free and open source software. All the hardware presented below can be found with relative ease in any sufficiently stocked local computer vendor. Here is the current hardware specification of the AGILA cluster:

Master Node Configuration(1):

- AMD Athlon K7 600Mhz
- Freetech K7M 200Mhz Motherboard
- 256 MB SDRAM
- 6.4 GB IDE Hard Drive
- (2) Intel Ethernet Express Pro 100+ Network Interface Card
- 21” Monitor
- CDROM Drive

Processing Node Configuration(7):

- AMD Athlon K7 600Mhz
- Freetech K7M 200Mhz Motherboard
- 128 MB SDRAM
- 6.4 GB IDE Hard Drive
- Intel Ethernet Express Pro 100+ Network Interface Card

Networking Hardware:

- (1) 24 port 100mbps Intel Express Pro 410T Switch
- Category 5 UTP cable and RJ-45 connectors

Others:

- APC SmartUPS Pro 3000
- 8-way Monitor-Keyboard-Mouse Sharing Switch

The nodes are connected together in a star topology to the Intel 410T switch. A switch was used since the prices of these Fast Ethernet Network Switches have already been greatly reduced due to the boom in the local internet industry that has caused an increase in the demand for these devices thus driving down their prices. The AGILA team is also working on an alternative network topology since there
are some inherent limitations with the star topology. The most obvious of which is the limit in the port density of the network switches in the market. The master node will have two network interface cards. One card connects the master node to the internal network and the other to the Internet. This is done to isolate the compute nodes from the Internet. This enables us to save on external IP address of the university and at the same time isolates the cluster from external network traffic reducing the probability of getting network collisions. The other network interface allows the master to enable connections from the external network that allows users to remotely connect to the cluster from anywhere in the world. The Intel 410T is Intel’s baseline switch that is reputed to provide sufficient throughput and is able to operate at full duplex mode together with the Intel Ethernet Express Pro 100+ Network Interface Cards. The Intel Ethernet Express Pro 100+ Network Interface Cards were selected because of its reputation for being the most efficient 100mbps NIC in the market. Aside from this, the Intel Ethernet Express Pro 100+ Network Interface Card provides wire speed performance and has PXE extension that eliminate the need for bootroms.

The 8-way Monitor-Keyboard-Mouse sharing switch was installed in order to monitor the compute nodes during the initial cluster configuration. The presence of this device enables the administrator to debug configuration and installation problems with relative ease. In the future, this device will not be necessary. The APC SmartUPS Pro 3000 can handle a load of up to 3000VA. Aside from providing backup power in case of power failure, the UPS is used to ensure that clean power is fed to the processing units. As of the moment, the eight node cluster is only taking up one-third of the UPS power capacity. There have been reports from other beowulf builders that ”dirty” and ”noisy” power causes some compute nodes to fail during some operations. This will be unacceptable especially when some operations need to run for continuous amounts of time. The selection of the Athlon as the processor of choice was easy because of studies and feedback from other beowulf builders that report of the positive performance results they get from the Athlons. The Athlon’s 3DNOW extensions can also be used to enhance cluster performance. There have also been reports that the Athlon has one of the best CPU benchmarks results in the industry today even beating its rival Pentium III class processor in nearly every category based on the SPECCPU benchmarks.

2. Software Configuration

The selection of software for the AGILA was relatively simple task. The software currently selected for the cluster is either free or open source software. These are the software packages currently installed in the system:
<table>
<thead>
<tr>
<th>Software Used</th>
<th>Linux Distribution</th>
<th>Redhat Linux 6.2 (Zoot)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linux Kernel</td>
<td>Linux Kernel v. 2.2.16 w/ NFS patches</td>
<td></td>
</tr>
<tr>
<td>Message Passing Libraries</td>
<td>Parallel Virtual Machine(PVM 3.4.3), Message Passing Interface(MPI-MPICH 1.2.1), Local Area Multi-computer MPI(MPI-LAM 6.3.2)</td>
<td></td>
</tr>
<tr>
<td>Scientific Libraries</td>
<td>Linear Algebra Package(LAPACK), (Scalable Linear Algebra Package(SCALAPACK), Portable Extensible Toolkit for Scientific Computation (PETSc)</td>
<td></td>
</tr>
<tr>
<td>Benchmark Suites</td>
<td>High Performance Linpack(HPL), NAS Parallel Benchmarks(NPB)</td>
<td></td>
</tr>
<tr>
<td>Support Software</td>
<td>NFS, BOOTP, RSH, NTPD</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>MPI-POVRAY Raytracer</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Software Currently Installed in the AGILA Cluster


3. Installation Procedure

3.1. Master Node. Linux was installed to the Master node off a CDROM drive. Great care was taken to partition the hard drive in such a way that there would be a separate partitions for the /home directory and the /tftpboot directory. This is done since the processing nodes will be accessing their data via Networked File System(NFS) to the master node. The /home directory will contain the users directory and the clustering binaries and libraries. In the future, there are even plans to make the slave nodes diskless and use the Network Interface Card to do a BOOTP and root over NFS on startup. Installation was straight forwarding since nearly everything is autodetected by Linux. During package selection screen, the unnecessary packages where removed from the installation. Unnecessary daemons where also removed like httpd, ftpd, squid and others.

After Linux is installed, there was a problem with Intel Ethernet Express Pro 100+ Network Interface Card. The NIC was detected automatically during boot-up. However, the network device did not seem to work. The problem was solved by statically compiling into a new Linux kernel the driver for the Intel Ethernet Express Pro 100+ NIC. It seems that this solved the problem. It might have been that case the Intel Ethernet Express Pro 100+ Kernel Modules cannot properly configure the network device. The newly compiled kernel was stripped of the items that were not necessary for the cluster as of the moment such as uninstalled hardware drivers, SCSI support and others.

The next phase was to setup NFS. The /home directory was shared across the entire cluster via NFS. This is done by modifying the /etc/exports file. The next step was to ensure that the entries in the /etc/hosts contain the names of the
nodes in the cluster and their proper mappings. In the /etc/hosts.equiv, the entries for the cluster nodes where placed in. This is done in order to ensure the remote shell(RSH) would work. RSH is commonly used to start the cluster jobs and daemons across the cluster. This is also a very convenient tool for configuring the other nodes in the cluster. MPICH 1.2.1 was compiled into the cluster and the binaries and the configuration files are kept at the /home/apps folder which is shared across the cluster. The same was done for LAM 6.3.2. PVM 3.4.3 was also installed via RPM. LAPACK(Linear Algebra Package) and ATLAS(Automatically Tuned Linux Algebra Software) where installed, configured and compiled. These libraries provide the cluster with Linux Athlon optimized LAPACK(Linear Algebra Package) and BLAS(Basic Linear Algebra Subroutines) routines. The BLACS communications libraries where then installed. SCALAPACK(Scalable Linear Algebra Package) which is a parallel version of the LAPACK libraries was also installed. PETSc( Portable, Extensible Toolkit for Scientific Computation) was also installed since it provides tools for computing linear algebra, differential equations problems as well as tools for visualizing and profiling them.

3.2. Processing Nodes. The installation of the processing nodes where trivial. The Linux CD was installed into the processing node an a Redhat Linux KickStart Script was used to automate the installation. All seven nodes where installed with relative ease. Their IP addresses and other network parameters where configured by hand but BOOTP support is being tested at this moment for automatic network configuration and bootup.

3.3. System Management, Maintenance and Use. In order to ease the configuration and upgrade of the cluster, the /home directory is shared via NFS to the other slaves nodes. A script called /bin/slave-rsh was created to control the entire cluster with one command from the master node. For example, if I had the add a user to all the nodes instead of doing a adduser newuser in every node, I would simply do a /bin/slave-rsh adduser newuser on the master node. The code for this script file and other script files will soon be made available in our website.

This script provides a means to do the following:
- Single point for adding and removing users from the cluster
- Shutdown and Reboot of the entire cluster
- Execute commands on all the nodes
- Ease of system maintenance

The sharing of the /home directory via NFS provides:
- Enables users to simple programming, compile and execute on a single node without worrying about the other nodes.
- Binaries are automatically propagated since the entire /home directory is share via NFS.
- Newer versions of the Message Passing and Scientific Libraries are automatically propagated across the cluster since the /home/apps directory contains these libraries which are shared.
- Setting of filesystem permission are made easier since they are only in one location
There are other benefits of this approach to cluster management. Since, message passing libraries and scientific computing libraries are provided in the master node. Applications can be compiled by the user in a single machine. The user simply logs on, writes code, and compiles it. The user need not migrate the binaries to the other nodes. After compiling, the user would simply execute the compiled application using mpirun. In order to ease compiling of code and ensuring that they will compile and link with the proper libraries, sample Makefiles were generated to be distributed to the cluster users.

4. Benchmarking

In order to check the performance of the cluster and detect possible bottlenecks, a set of benchmarks were run on the cluster. There are the High Performance Linpack(HPL) and the NAS Parallel Benchmarks(NPB). High Performance Linpack (HPL) is a freely available implementation of the High Performance Computing Linpack Benchmark for distributed-memory computers. It is a software package that solves a (random) dense linear system in double precision (64 bits) arithmetic on distributed-memory computers. The HPL package provides a testing and timing program to quantify the accuracy of the obtained solution as well as the time it took to compute it. This serves as a replacement for the old LINPACK benchmarks used in determining the Top 500 supercomputer in the world. The NPB are a set of 8 programs designed to help evaluate the performance of parallel supercomputers. The benchmarks, which are derived from computational fluid dynamics (CFD) applications, consist of five kernels and three pseudo-applications.

These are the initial HPL results taken on Sept 19-21, 2000:

<table>
<thead>
<tr>
<th>N</th>
<th>NB</th>
<th>P</th>
<th>Q</th>
<th>Gflops</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>64</td>
<td>2</td>
<td>4</td>
<td>1.02</td>
</tr>
<tr>
<td>5000</td>
<td>64</td>
<td>2</td>
<td>4</td>
<td>2.22</td>
</tr>
<tr>
<td>8000</td>
<td>64</td>
<td>2</td>
<td>4</td>
<td>2.88</td>
</tr>
<tr>
<td>10000</td>
<td>64</td>
<td>2</td>
<td>4</td>
<td>3.16</td>
</tr>
</tbody>
</table>

Table 2.A: Results for the AHPC AGILA 8-node Athlon Cluster(128MB) - Fast Ethernet

<table>
<thead>
<tr>
<th>N</th>
<th>NB</th>
<th>P</th>
<th>Q</th>
<th>Gflops</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>64</td>
<td>2</td>
<td>4</td>
<td>1.76</td>
</tr>
<tr>
<td>5000</td>
<td>64</td>
<td>2</td>
<td>4</td>
<td>2.32</td>
</tr>
<tr>
<td>8000</td>
<td>64</td>
<td>2</td>
<td>4</td>
<td>2.51</td>
</tr>
<tr>
<td>10000</td>
<td>64</td>
<td>2</td>
<td>4</td>
<td>2.58</td>
</tr>
<tr>
<td>15000</td>
<td>64</td>
<td>2</td>
<td>4</td>
<td>2.72</td>
</tr>
<tr>
<td>20000</td>
<td>64</td>
<td>2</td>
<td>4</td>
<td>2.73</td>
</tr>
</tbody>
</table>

Table 2.B: Results for the UTK/ICL Torc 8-Dual Intel PIII 550Mhz(512MB) - Myrinet

The performance measure for different systems may vary with different factors. The most important of such factors are the processor, memory and network connection. HPL assumes that the power of every node in the cluster is the least powerful of all the nodes. HPL also contains restrictive assumptions on the interconnection network. The AGILA cluster is only using 100mbps Fast Ethernet as compared to
the Torc Cluster which uses Myrinet(1.2Gbps). Since HPL maximizes this bandwidth, placing AGILA as a disadvantage. However, for an 8-node cluster, 100mbps Fast Ethernet is more than enough. HPL test scalability is limited to the amount of memory that each node has. For example, the AGILA cluster with only 128MB per node is only able to solve matrices of size 10000x10000 as compared to the Torc cluster with 512MB that can solve matrices of size 20000x20000.

However, the AGILA is able to performance better than the Torc Cluster when it comes to computing 8-processor grids. This is expected since the Torc Cluster is using PIII 550Mhz while the AGILA is using AMD K7 ATHLON 600Mhz. The Torc Cluster will definitely perform better than the AGILA in computing 16-processor grids. Based on these preliminary results, it can be concluded that the cluster is performing as it should.

5. Next Steps

There are still a lot of things to be since AGILA itself is less than a month old. However, there has been a wealth of experience gained from this endeavor here are some of the points of improvements and the possible improvements to be made:

- Improve network topology and routing - since current network topology and other available network topologies are unable to scale beyond 64 or 128 nodes with a Fast Ethernet switch. Other network topologies simple have too large latency values or too little bandwidth.
- Easier Cluster Construction, Configuration and Maintenance - under development right now is a method to remove all configurations and software from the processing nodes. All the nodes will be diskless, diskette drive less and bootrom less if possible. The node will simply connect via NIC bootrom or for Intel Card the PXE to a BOOTP and NFS server to automate the network configuration and startup of the processing nodes.
- Other Clustering Technologies - Try other clustering emerging cluster concepts and idea such as MOSIX, PVFS and others.
- Develop Courseware and Modules - this is in line with the team’s goal to promote the use of parallel computing in the university in particular and in the country as a whole.

**Ateneo High Performance Computing Group, Ateneo de Manila University, Loyola Heights, Quezon City, 1108 Philippines**

*E-mail address: william.s.yu@ieee.org*