Building a Large Scalable Internet Superserver for Academic Services with Linux Cluster Technology

Putchong Uthayopas, Surasak Sanguanpong,
Yuen Poovarawan
Department of Computer Engineering,
Faculty of Engineering, Kasetsart University,
Bangkok, Thailand 10900.
E-Mail: pu,nguan,yuen@ku.ac.th
Phone (662) 9428555 EXT:1416 Fax (662) 5796245

Abstract

With the speed and bandwidth offered by the next generation Internet technology, there is a need for large and scalable Internet server that can provides an adequate computing power and storage for the new generation Internet applications. This requires a huge investment in a very large and expensive commercial server system. Recently, the emergence of Linux PC clustering or so-called Beowulf Cluster Computing technology has provided a very low cost, scalable, and high-performance alternative to using these commercial superserver system in many class of applications. This paper presents our experiences learned from the construction of a 72 nodes Beowulf Class Cluster called PIRUN (Pile of Inexpensive and Redundant Universal Nodes) Cluster at Kasetsart University, Bangkok, Thailand. The purpose of this project is to explore the use Linux PC clustering technology to build a central Internet superserver and supercomputing class system for academics support at the university level. In this paper, the planning of the system is presented along with the discussion about issues in both hardware and software. This system when fully exploited will cost the same as current same main server but offer about 30-40 times more computing power and storage. Moreover, PIRUN System will becomes the largest supercomputing system in Thailand and deliver about 3 times higher performance than the previous largest supercomputing system in Thailand. We strongly believe that this type of system and technology is very suitable for developing countries in Asia Pacific Region.

1 Introduction and Motivation

The increasing bandwidth and speed of the next generation high-speed Internet provides many new opportunities for Internet based services and applications. To provides such services, one need to have a very powerful server system with huge amount of storage and computing power. This system must also be very scalable to keep up with the rapidly increasing demand. However, such a powerful system is still very costly. Hence,

this is potentially becomes a major limiting factor to the widespread implementation of new Internet services in developing countries in Asia Pacific Region.

The newly emerging technology of Linux PC clustering, which is the building of large, scalable, superserver or supercomputing class platform from PC and Linux operating system seems to be a viable alternative for this purpose. This technology has been employ by many large organization such as NASA, National Laboratories in US, and Universities around the world[1, 2]. There is a project called Beowulf Project at NASA[3, 4] that is an important contributor and major driving force for this technology.

Realizing the significant impact of this technology, Kasetsart University decided to build a large 72 nodes PC cluster system called PIRUN (Pile of Inexpensive and Redundant Universal Nodes) Beowulf Cluster as a central computing facility and large scale test based for clustering technology. This project is a collaboration between KU Research and Development Institute, Faculty of Engineering, and Computing Service Center. The design and construction has been under the responsibility of the staffs from Computer and Network System Research Laboratory (SRU), Department of Computer Engineering, Faculty of Engineering. The goals of this project are:

- Building a system that serve as a centralize scalable Internet superserver for more than 15,000-20,000 registered Internet users including both students and staffs of Kasetsart University(KU).
- Providing a world class supercomputing facility for researchers in Kasetsart University in the area such as Computational Chemistry, Computational Fluid Dynamics, Bioinformatics, and Computer science researches.
- Building a large Intel based PC Cluster to be used as a test-bed for cluster computing technology.

The organization of this paper is as follows. Section 1 discusses about the hardware selection. Next section, Section 2, presents the ideas about software selection followed by Section 3 that gives the detail on system installation. Section 4 discusses the planned application of the system. Finally, Section 5 presents the conclusion.

2 Hardware Selection

The process of constructing PIRUN Beowulf Cluster started by the selection of hardware component. The criteria that we taken into consideration are:

- All hardware must be pure commodity part to minimize the cost.
- However, extra cost can be spent on component that we consider crucial to the reliable operation of the system.
- These hardware must deliver high enough performance for large scale scientific computing needs.

To make the decision easier, we decided to organize the nodes in our system as follows:

1. CSN (Computer Service Node) : These nodes are the system that users log-on to do their work or submit their computation tasks to be execute by the system.

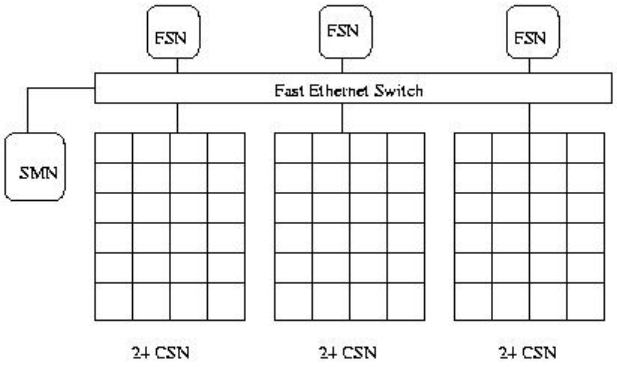

Figure 1: PIRUN Cluster Organization

2. FSN (File Server Node) : FSN is a set of computer with large and fast storage. All users data will be distributed among FSN nodes. Only root can login to these nodes to config them.

3. SMN (System Management Node) : This class of node is a console to the system and also perform some functions such and system performance collection, centralize user account management.

All CSN are diskless nodes. This allows us to easily manage the kernel image , module, CSN root file system centrally. Moreover, this is also cut down the price of the system dramatically. Since the target application is compute intensive application with minimal I/O. This approach does not degrade the system performance in anyway. We decided to base PIRUN implementation on Intel processor due to the low cost and high software compatibility. We decide to have a mother board with wake-on LAN and hardware monitoring feature. This allows us, with appropriate software, to remotely turn on and off unused nodes. Also, hardware monitoring support helps detect many problems such as malfunction CPU fan, insufficient air flow before they can cause the damage to the system. For FSN, we decided to install RAID system and hot swap power supply. Since there are only 3 FSN, this helps increase the system performance with slightly additional system cost.

For the interconnection, the major requirement is that all nodes must be on the same physical network to minimize the latency and maximize the communication bandwidth. This is achieved using a set of stackable fast Ethernet switches with Gigabit connection that linked the switch fabric together. For PIRUN, the switch used is 3COM SuperstackII which support up to four 24 ports switches. This provides us 96 ports which is enough for current configuration. For the future system expansion, gigabit Ethernet can be used as a backbone that link more switches together. Also, with large enough budget, MYRINET seems to be a much faster and more scalable choice. The list of components used are:

1. CSN: Pentium III 500 MHz, 128 Mbytes RAM per nodes, ASUStech P2B series mother board that support wake on LAN and hardware monitoring chip. There are 72 CSN in PIRUN system.

2. FSN : DUAL Pentium Xeon 500 MHz Server with hardware RAID. Currently, each FSN has 6 of 9 Gbytes Ultra SCSI disk which total to 54 Gbytes per node. There are 3 FSN in the system.

3. SMN : Pentium III 500 MHz , the same as CSN but with some local hard disk to store software and configuration data. Only 1 SMN is used.

4. KVM (Keyboard/Video/Mouse) switch with daisy chained capability are used. Total of 10 KVM has been chained to centralize the console access to SMN.

5. Four 24 port stackable Fast Ethernet switches are used.

6. A large set of of CAT5 Ethernet cable.

3 Software Selection and System Planning

For software installation there are plenty of good opensource software available. Part of the software that we used are:

- Linux is used as an operating system. Currently, Linux seems to be a clear choice in this area due to its strong presence and technical strength. We decide to use Red Hat 6.1 distribution for PIRUN due to its fine collection of tools.

- Parallel programming support is also a crucial element for the system. We install PVM4.3 and MPICH1.1.2 for this purpose. Also, we consider using Portland Group C,C++, and High Performance compiler if possible. PetSc and Scalabpack math library is also installed to help ease the application development on PIRUN.

- Managing such a large scale system is not a trivial task. So, for the system management tool, we install our tools called SCMS (Smile Cluster Management System) which we have developed for several years now. SCMS consists of an alarm system for system malfunction, real-time monitoring, hardware component browser and many necessary software tools for system management[5]. We have developed many tools for setting up and managing large ensemble of machines, these software are available on Internet at (http://smile.cpe.ku.ac.th) Furthermore, to control the usage of the hardware and the sharing of resources, we plan to use PBS (Portable Batch System) on our system.

- Finally, Visualization is now an important part of scientific computing. VIS5D and IBM Data Explorer is installed to be used for scientific visualization in our system

For PIRUN, we wants to create a single system image view so that user can use this system as a single large system. Usually, NIS is a tools that is widely used for this purpose. However, we found that as the system become larger NIS will slow down many activities substantially due to its centralize nature. So, the prefer method for us is replication of the major files. This has been done by our tool that we developed.

Figure 2: PIRUN Beowulf Cluster

4 System Installation

The installation steps are:

1. Install the hardware, cabling, switch , rack mount in-place. This takes us about 2 days to finish.

2. Install Linux System on the Server and prepare the file system for diskless nodes. We develop a special software tool for this process. This tool will be available as an opensource next year.

3. Prepare the boot floppy disk for each nodes and boot up all nodes one by one, test whether it works with the system or not.

4. Install selected software on server nodes. As most of the software has been in RPM, an easy to install format, the installation of software happened quickly at this point. Next step is to install test applications, we use PVMPOVRAY, popular parallel rendering software to see whether the nodes works together as a large group.

5 Application of PIRUN Beowulf Cluster System

There are many planed applications for this systems. At the first stage, the system will be split into two parts: an Internet server part and supercomputing part.

For Internet server part, every normal university's users will have an account and use this system to store their data, read email, store web page.

In order to use the supercomputing part, user needs a special password to gain an access to PBS batch system.

There will be 16-24 nodes with local hard disk. These nodes will be used to run parallel web software and used as a scalable high-load web server for the university tasks. We have a development project to build a parallel web crawler that collect,

process and index various information. The power of this system will be used to drive that application. The data collected can be searched using our search engine and parallel search engine under development by Applied Network Research Group, Computer and Network System Research Laboratory, Department of Computer Engineering, Kasetsart University. More detail about this effort are as listed:

1. Parallel Search Engines(Nontri Search): NontriSearch is a robot-based search engine, originally designed to serve as Thai enabled campus network search engine inside Kasetsart University. Recently, NontriSearch capabilities are extended to serve as Internet search engines as well. Due to the limitation of bandwidth, NontriSearch collects only th domain for now. We have already found 1231 unique domains that have th domain with 79101 URLs. NontriSearch next generation is enhanced with parallel architecture based on PIRUN Beowulf cluster. Parallel Web Crawler, Indexer and queries processing has also been developed. The crawlers are running totally independence and have separated collection of web pages. We are currently testing and measuring the system performance.

2. Parallel Web Server: Popular web site like CNN, Microsoft, etc may received millions of http requests per day. These web servers need a sophisticated way to handle the traffic at the higher rate than a single server can offer. Web server clustering liked Beowulf can add scalability and high availability to the services. Simple round robin DNS may be easily to implement without any modification but it does not provide availability when one of the machines is down. We investigate multi-ways to balance the load and provide automatic fail-over of the system.

Besides using PIRUN system as an Internet Superserver, this system will be employed as a supercomputing facility for research in area such as computational chemistry, computational fluid dynamics, parallel software tools and environment. By having this system, researchers at Kasetsart University will have a very powerful tools to help them tackle difficult research problem.

6 Conclusion

From this project, we have learned many valuable experiences. A careful planning proof to be very valuable. For this kind of system, all system components can be put in place within a short period of time (two days). For the software installation, the the system can be quickly installed due to our expertise in cluster software tool. With out this kinds of automated script, the installation will be very tedious and long. Currently, there are no such tool available openly. Hence, this is one thing that need to be addressed. Therefore, we plan to release our software tool set for benefit the community soon.

Finally, we feel that this technology will create a significant impact in the future because it allows various organizations to have a large, scalable powerful Internet superserver at the affordable prize. In our case, we are able to cut the cost for having large server dramatically. The system also very scalable which make a future expansion of the system capacity very easy to accomplish. But a certain level of expertise is needed to build and operate such system.

References

[1] R. Reisen, R. Brightwell, L. A. Fisk, T. Hudson, and J. Otto, "Cplant*," in *Extreme Linux Workshop, USENIX Annual Technical Conference*, (Monterey, California), June 8-11 1999.

[2] D. M. Halstead, B. Bode D. Turner, and V. Lewis, "Giga-plant scalale cluster," in *Extreme Linux Workshop, USENIX Annual Technical Conference*, (Monterey, California), June 8-11 1999.

[3] T. Sterling, D. J. Becker, D. Savarese, J. E. Dorband, U. A. Ranawake, and C. E. Packer, "Beowulf: A Parallel Workstation for Scientific Computation ," in *Proceedings of ICPP 95*, 1995.

[4] D. Ridge, T. Sterling, D. J. Becker, and P. Merkey, "Beowulf: A Parallel Workstation for Scientific Computation ," in *Proceedings of IEEE Aerospace 1997*, 1997.

[5] T. Angsakul and P. Uthayopas, "Basic support systems for distributed application on smile beowulf cluster," in *Proceedings of ANSCSE 3*, (Chulalongkorn University), National Electronics and Computer Technology Center, March 1999.