A Simple Installation and Administration Tool for the Large-scaled PC Cluster System

Tomoyuki HIROYASU[1], Mitsunori MIKI[1], Kenzo KODAMA[2], Junichi UEKAWA[2] & Jack DONGARRA[3]
[1]*Department of Knowledge Engineering and Computer Science, Doshisha University, Japan.*
[2]*Graduate School of Engineering, Doshisha University, Japan.*
[3]*Department of Computer Science, University of Tennessee*

Abstract

In this paper, a new setup/administration tool for PC cluster systems is proposed. Recently, in the high performance computing field, PC cluster systems are becoming popular. PC cluster systems consist of PCs connected via a network and are used for parallel and distributed computing. PC cluster systems achieve a good cost to performance ratio by using commodity hardware to construct the cluster. However, it is very hard to install and configure a PC cluster because many nodes exist and a large amount of knowledge is required for the installation and configuration of the cluster. In this paper, to solve this problem, a simple installation and administration tool for PC cluster called "Doshisha Cluster Auto Setup Tool: DCAST" is developed. DCAST has the following features: DCAST can be used for both diskless and diskfull clusters; DCAST is targeted for Linux; there is no interactive operation during the installation; slave nodes are booted over the network; and the whole system is reinstalled when reconfiguring the system. The targets, philosophy, and operations of DCAST are described.

1 Introduction

Recently, in the engineering and technical computing areas, requirements for large-scale computation is increasing, leading to a high demand for better computational performance in computers. Currently, PC clusters[1, 2] are one of high performance computing equipments and they are attractive due to their relatively low

installation cost. PC clusters are parallel computers constructed by multiple PCs interconnected with special network components. Because they are constructed using parts that are widely available in the consumer market, they have a very good cost to performance ratio. However, although PC clusters are relatively cheap to install, system construction and tuning is not easy. Also, for installing new software to the cluster system, and updating software for security problems, all machines that are part of a cluster needs to be updated at the same time. The task of installing new software is often cause of trouble and mistakes when it is done manually.
In this research, Doshisha Cluster Auto Setup Tool (DCAST) is developed, to allow easy system construction and system updating for PC clusters. DCAST targets at university students without much prior knowledge of PC clusters, and aims to construct large-scale PC cluster systems based on Debian GNU/Linux operating system. Also, when upgrading the whole system, DCAST takes the approach of reinstalling all machines in the cluster system instead of upgrading individual nodes. In this paper, we will discuss problems in PC clusters, and propose DCAST, and explain the problems with users that DCAST targets at, the design policy of DCAST, and the actual operation when constructing a PC cluster.

2 PC clusters

2.1 Overview of PC Clusters

PC Clusters are parallel computers that are constructed by multiple PCs, and interconnected by network to operate as one computational resource as a whole. Individual nodes in a PC Cluster systems are PCs of which the components are mass-produced for the consumer market.
In this paper, we assume a basic construction of a PC cluster to be as shown in Figure 1. As shown in this diagram, in a PC cluster system, only the master node can communicate with outside network, and the slave nodes construct an internal network that is independent from the outside network. It is not possible to access the slave nodes from outside of the PC cluster. This scheme has an advantage that even when number of nodes in the PC cluster increased, one host can be used as a firewall, to help ease the security risk.

2.2 Problems with PC cluster systems

In a PC cluster system, the following problems exist.

- Requires deep specific knowledge for constructing and maintenance:
  Individual nodes that construct a PC cluster system require an operating system of its own. For the operating system, Linux is typically used in the recent years. PC clusters take the form of multiple Linux PCs linked by special network devices. This demands knowledge of the operating system, and of the special networking, and there is a high barrier of knowledge for

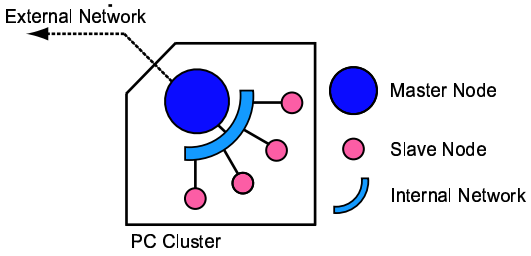

Figure 1: Basic PC cluster structure

novices.

- Installation effort:
  When the PC clusters approach a scale of more than several hundred nodes, it will take a large amount of time and effort just installing the operating system and configuring.

- Maintenance cost:
  Individual nodes that construct a PC cluster is each a computer on its own, and similar amount of maintenance to a standalone work PC is required, such as software installation, hardware checking, system upgrading, and replacement when parts of hardware break. In addition, maintenance as a parallel computer, such as configuring the network, installation of parallel computing library, introduction of networking software and middleware are required.

To solve these problems, we propose `DCAST`.

3 DCAST

3.1 Overview of DCAST

`DCAST` is a system to construct PC clusters with ease and effectiveness. The target users of this system are students who do not have technologies specific to PC clusters. To those students without such technologies, large-scale PC cluster can be constructed using this system.
The target users of this system have the following problems:

- Cannot answer interactive questions at install time.

- Cannot provide information that are specific to characteristics of individual hardware.

- Cannot check system integrity of individual PCs or cluster-wide integrity.

- Does not know what software needs to be updated.
- Does not know how to improve security to an acceptable level.

DCAST was designed to be a helpful tool for users with such problems.

3.2 DCAST design goals

DCAST has the following design goals:

1. Use Debian GNU/Linux as operating system:
   Debian GNU/Linux has an advanced packaging system, and it allows to easily update installed software and install security updates. Also, it automatically detects conflicts and dependencies in packages, and system integrity can be maintained. These packaging systems are specific to Debian GNU/Linux, and compared to other Linux distributions, security updates and upgrading versions of software is easier.

2. Omission of interactive questions at installation time:
   For installation of an operating system, interactive operations are required. For large-scale systems such as PC clusters, interactive operation requires a large amount of effort. Also, a novice administrator cannot reasonably answer questions. DCAST does not require interactive operations, and automates installation, to simplify the installation process, and reduces the workload.

3. Reinstalling of PC cluster for upgrading software versions:
   It is possible to take in latest software technology with PC cluster systems. To keep on using the latest technology, frequent updating and upgrading is required. However, frequent upgrading makes it harder to keep the integrity of inter-node software versions. As a countermeasure of such problems, DCAST will reinstall all the nodes in the PC cluster system for upgrading.

4. Does not assume heterogeneous computing environment:
   DCAST assumes students without specific knowledge as user, and does not assume a complex heterogeneous cluster. DCAST assumes a homogeneous x86 architecture cluster.

5. Supports both diskless and diskfull installations:
   There are cases where having many hard disks is undesirable, due to the fact that it is a moving part and may break more often than static parts. There is a form of PC clusters without having a hard disk on each host, called “diskless clusters”. For PC clusters constructed with PCs with hard disk, they are called “diskfull clusters” in this paper. DCAST will construct both types of clusters.

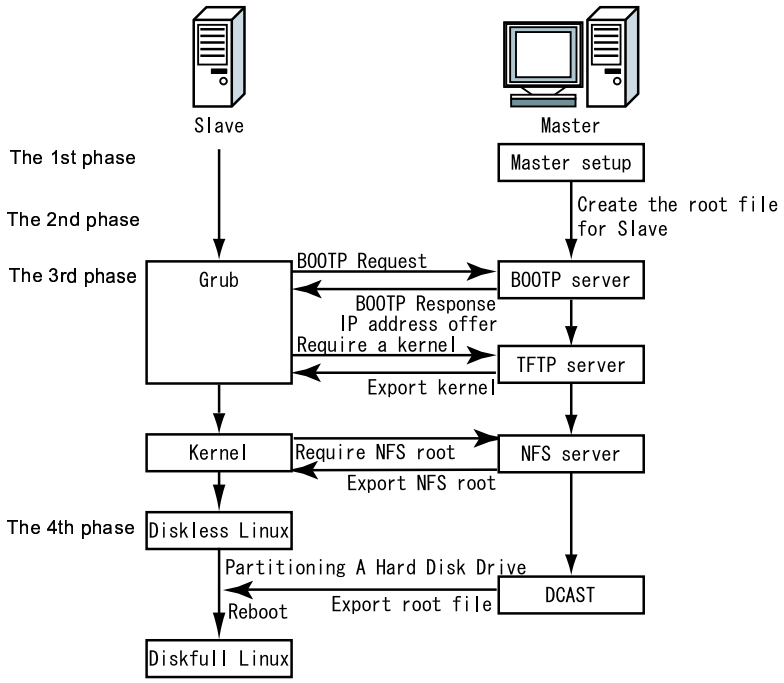

Figure 2: Operation of DCAST

6. Integration with existing software:
   DCAST is constructed with several existing services. By using multiple software, DCAST operation is divided, and it makes a structurer that is easier for bugfixes and upgrading.

By using this design, DCAST provides easy PC cluster construction and maintenance.

3.3 The flow of constructing a PC cluster using DCAST

When constructing a PC cluster using DCAST, a node for running DCAST is prepared, and operating system is installed. Using that node as the master node, all slave nodes are set up. The DCAST process at the master node is constructed from bootp[3] server for providing IP address to slave nodes, tftp[4] server for providing kernel to slave nodes, and NFS[5, 6] server for providing the root filesystem to slave nodes. The configuration for these software are done by DCAST at the beginning.
DCAST operation can be divided in four steps. The operation of DCAST is shown in Figure 2.

```
#Enter PARTITION size.
FPRT    /dev/hda1       128     boot    *
SPRT    /dev/hda2       512     swap
TPRT    /dev/hda3       -       /
4PRT    /dev/hda4       -       etc
HOSTNAME
NISDOMAIN       nis.org
LOCALETHCARD    eth0
#       NETWORK         NETMASK         BROADCAST
NET     192.168.1.1     255.255.255.0   192.168.1.255
#MASTER Master's name   Master's IP
MASTER  host01          192.168.1.11
SMASTER host_master     192.168.1.2
GATEWAY 192.168.1.254
#slave's name   slave's IP      slave's MACaddress
#Autogenerated by update-cluster
host02 192.168.1.12 009027D0A80B
host03 192.168.1.13 004005A06C67
host04 192.168.1.14 004005A886A5
host05 192.168.1.15 004005A06427
host06 192.168.1.16 004005A40DEE
host07 192.168.1.17 004005A40DEF
#End update-cluster
```

Figure 3: slave.lst

In the first step, preparation and configuration required for proper operation of DCAST is done at the master node. The configuration is done once by DCAST, and it is not required to do this part afterwards. The command used in this setup is provided as `dcast-setup`. `dcast-setup` does the setup and configuration according to `slave.lst` (Figure 3). `slave.lst` is the only configuration file for DCAST, and it contains the partition information, network configuration, host names, IP address, and network card MAC addresses.

In the second step, root directory for starting the slave nodes are prepared. The master node prepares root directories for all the slave nodes. Slave nodes will boot using the directories as their root directory. Root directories for slave nodes are rewritten by DCAST for diskless node configuration. The slave node root directory is basically the same as the master node.

In the third step, slave nodes are booted. By booting the slave nodes, DCAST sends bootp request to the master node. Master node provides an IP address to slave node in response to the bootp request. The slave node with the IP address will then send a request for a kernel from the master node. tftp server that is operating on the master node will provide a kernel to the slave node. The slave node will use the kernel to boot up, and will NFS-mount the provided directory on the master node as the root file system and operate as a diskless node. For the operation as a diskless PC cluster, the process completes here.

As a fourth step, diskfull machine is constructed. DCAST provides its own version of init program on the slave node that starts up with the operating system provided on the third step. DCAST init will partition the hard disk of the slave node, and copy the operating system files. Operating system information is basically the same as

the master node, and there is no need to reconfigure the parts which function in the same manner as the master node, except for some specific hardware details. The copying is done via NFS. Finally, the configuration file for setting the filesystem mounting information, `/etc/fstab`, is rewritten, to change the root file system to the hard disk. Also, the configuration for the boot loader grub[7] is rewritten to change network booting to booting from local harddisk.

3.4 Addition of module scripts

For software used in PC clusters, for those which are able to have the exact same configuration over all nodes, DCAST will duplicate all information on all nodes, and individual configuration is not required. However, there are some software where operation differ on master node and slave nodes, and configuration needs to be modified accordingly. To solve this problem, DCAST provides module scripts for setting some software. Also, it is possible for a cluster administrator to provide additional scripts.
DCAST will run all scripts in `/usr/lib/dcast` by default, and by placing scripts in the directory, arbitrary software configuration can be done.

3.5 Comparison with existing software

There are several other cluster construction and maintenance software, and Oscar and NPACI Rocks are popular. However most of these software are specific to Red Hat Linux, and there are little which support Debian GNU/Linux.
Oscar[8] is a cluster construction and administration tool that is being developed by Open Cluster Group. At the Open Cluster Group, PC cluster construction is aimed at being a Computational Grid[9] end-point. Therefore it is only possible to construct a PC cluster that takes in consideration a Computational Grid environment. With DCAST, it is possible to modify the configuration files to be more flexible about the design of PC clusters.
NPACI Rocks[10] aims at people who have enough technology level. NPACI Rocks supports architecture and platform independence. DCAST restricts architecture to x86, and restricts the operating system to Debian GNU/Linux. By paying this price DCAST allows installation and configuration of PC cluster without requiring much knowledge and technology on administrator side.

4 Conclusion

In this paper, a cluster set up tool DCAST is presented, and its actions and characteristics are described. It is a tool to set up PC clusters that is attracting attention in parallel and distributed computing in recent years.
DCAST has the following features.

- Target users are university students who do not have much skills specific to

PC clusters.

- Allows automatic PC cluster construction without interactivity.
- Debian GNU/Linux is the target operating system, and targets the x86 architecture.
- For upgrading the whole system, the whole system is reinstalled.
- Allows construction of both diskless and diskfull clusters.
- By adding module scripts, `DCAST` operation can be customized.

By these characteristics, `DCAST` is a useful software for creating large-scale PC clusters which may be a hybrid of diskfull or diskless.

References

[1] T. Sterling, D. Savarese, D. J. Beeker, J. E. Dorband, U. A. Renawake, and C. V. Packer. BEOWULF: A parallel workstation for scientific computation. *In Proceedings of the 24th International Conference on Parallel Processing*, pp. 11–14, 1995.

[2] SCYLD COMPUTING CORPORATION. *http://www.scyld.com/*.

[3] Interoperation Between DHCP and BOOTP. *http://www.faqs.org/rfcs/rfc1534.html*.

[4] THE TFTP PROTOCOL (REVISION 2). *http://www.ietf.org/rfc/rfc1350.txt*.

[5] NFS: Network File System Protocol Specification. *http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1094.html*.

[6] NFS Version 3 Protocol Specification. *http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1813.html*.

[7] GNU GRUB GNU Project Free Software Foundation (FSF). *http://www.gnu.org/software/grub/*.

[8] Open Cluster Group. OSCAR: A packaged cluster software stack for high performance computing. *http://oscar.sourceforge.net/*.

[9] Ian Foster and Carl Kesselman. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann Pub, 1998.

[10] Philip M. Papadopoulos, Mason J. Katz, and Greg Bruno. NPACI Rocks: Tools and Techniques for Easily Deploying Manageable Linux Clusters. *IEEE Cluster 2001*, pp. 258–267, 2001.