Bluemin: A Suite for Management of PC Clusters

Hai Jin, Hao Zhang, Qincheng Zhang, Baoli Chen, Weizhong Qiang
School of Computer Science and Engineering
Huazhong University of Science and Technology
Wuhan, 430074
CHINA

Abstract: - In this paper, a suite for management of PC clusters is proposed. PC cluster are becoming more and more popular in high performance computing because of its high performance-costs ratio. But it is a hard work for the administrator to manager so many computers and let them work as whole. To solve the problem, a software suite called Bluemin is developed. Bluemin provides a way to manage the pc clusters - the installation of the cluster, the user management, the system configuration, the system monitor, the system maintenance are all include in the suite. This paper introduces the basic architecture of Bluemin and describes some components of the suite.

Key-Words: - Cluster computing, Linux, Management systems, Cluster configuration

1 Introduction

Nowadays PC clusters are getting more and more popular. Many research institutes use clusters for HPC servers and networks servers. PC clusters [1] are PCs interconnected by network to work as one computational resource. But for the system administrator, the work of management and configuration of the clusters are really complex for that the clusters is not stable due to inconsistent configuration and sometimes we do not know the system status because it contains so many machines. The system administrators have to master too much system knowledge to help them be competent for the job. So, some methods are raised and some management packages are released to help the administrator to do the work in these years.

There are three main problems involved in the managements of PC clusters.

- **Installation of the cluster** - there are one head node which provides a single user interface and lots of slave nodes in a cluster. The head node servers the uses and dispatches the tasks to different slave nodes, also the management package is installed on the head node. Commonly, the administrator has to install the head node individually because it's quite different from the slave nodes, but a way must be provided to install all of the slave nodes once, which can help the administrator do the work as whole and rapidly.
- **System management** - there are many things involved in the system managements of a PC cluster: the single point of system management; the user managements; the health of the cluster and the monitor; the software package managements of the cluster; the reconfiguration of the cluster; the uniform configuration and etc. A good PC management suite should solve the problems to help the administrator managing the cluster just like managing a single server. Not like the single UNIX server, the configuration files of a cluster are always distributed on the different nodes, this make things become much complex and then it is a hard and meticulous work to develop a good cluster management package.
- **Managements of distribute resources** - the cluster is used to compute or to provide services. Then, the distribute resources of the bound PCs must work together properly to make the cluster server the

users efficiently. So the management software package must provide a way to achieve cluster resource scheduling, job submission, the monitor and control of the whole system.

The developments of Bluemin do much work on the problems to achieve it as a complete cluster management suite. The paper is organized as following: in section 2, we provide an overview of some cluster management tool kit. In section 3, we describe Bluemin architecture and the details of the different components. In section 4 we summarize the characteristic of Bluemin and draw a conclusion.

2 Related Works

In this section, we introduce some PC clusters management packages, and also we point the excellence and the pitfalls of them.

- **OSCAR -** OSCAR is release by the Open Cluster Group [2]. OSCAR provides a graphical wizard for the installation of clusters. The whole process is based on the kickstart [3] of Redhat Linux and OSCAR logs the configuration in Oscar Data Repository. The main insufficiency of OSCAR is that OSCAR does not provide a single system managements interface when the cluster is running, then the administrator have to configure the cluster by configuring the nodes one by one and the administrator have to re-install the whole cluster when errors occur or a new software package would be installed.
- **Rocks -** Rocks is a clustering management tool developed by the NPACI [4], which includes many work of the clustering management and maintenance. Rocks designed to produce customized distributions that define the complete set of software for a particular node. REXEC [5] from UC Berkeley is used to provide remote process control when MPI programs run in Rocks. Rocks provide a MYSQL database instance for some records of the status of cluster management but Rocks does not provide the auto-maintenance of some tables and so increase the burden of the system administrators.
- **Scyld Beowulf -** Scyld Beowulf is a product of Scyld Computing Corporation [6], which provides a single system image by modify the kernel of the linux and glib. Oscar, Rocks and Bluemin are all not SSI system. The installation and maintenance of Scyld clusters are very easy, but because Scyld makes so many changes on linux kernel, the suite has some limitation when users want to do some updates.
- **ClusterWorX -** ClusterWorX is release by the LinuxNetworX Computing Corporation [7]. ClusterWorX is an administration tool for monitor and management of Linux-based cluster systems. ClusterWorX pays more attention on the cluster monitor, and provides a very nice GUI to the administrator to manage the cluster. The installation of the cluster nodes is based on the disk cloning. Bluemin does not use the disk cloning, because such a technology does not suitable for the heterogeneous hardware and software cluster environment.
- **Scalable Cluster Environment -** The OpenSCE.org deploy their SCE project [8]. SCE is a collection of the cluster toolset. The packages for system installation, system monitor, job schedule are all included in the SCE. The pitfall of SCE is that the components are not combined to an integrated system. The administrator has to configure the different component carefully to make them work together. This decrease efficiency of all the packages.

3 Bluemin Architecture

3.1 Overview of Bluemin

Bluemin is a system to install and manage the PC clusters. The system administrator could control the cluster easily and effectively with Bluemin. In Figure 1, we define Bluemin a four level architecture.

In level one, we provide a web based user interface to integrate all management GUI. The administrator can do most of their work in a uniform user interface.

In level two, a database constructed with PostgreSQL is used to record most of the information of the cluster. Many fields of the tables are filled

automatically by Bluemin, so the tables can be kept correct and uniform. The administrator can get more information from the database and also logs can be print out to provide more information to help the administrator diagnose the errors of the cluster.

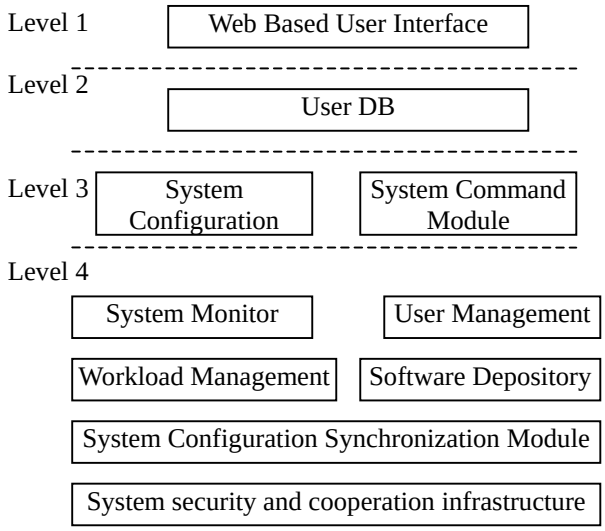

Fig.1 The Architecture of Bluemin

Level three contain two components: one is the system Configuration module, which defines the configuration interface to the low lever system configuration. The other is the system command module providing a cluster command interface to the system administrator.

Level four includes six modules: system monitor module is used to provide system information and status to the user; user management module provides functions for the user profile maintenance, user adding and deleting. Some packages are integrated into Bluemin to provide the workload balance. The other important part of this level is the software depository, which store the copies of the software that should be installed on the nodes. System configuration synchronization module examines and compares all the configuration changes of the cluster and keeps the system being consistent and convergent. System security and corporation model is a low level infrastructure defining the basic security and corporation protocol, and thus the cluster can work together smoothly.

3.2 The Cluster Database

The most successful design in Bluemin is the cluster database. Most of the changes and operates on the clusters are logged into the database. The database keeps the information of the cluster being uniform and consistent. The managements based on the database will be convergent because there just one copy of data is believable. All changes are based on the information in the database. On the other hand, if one or more slave nodes crash, the administrator could recover the data and configuration based on the information provided by the database.

Bluemin configuration database uses eleven tables to keep the basic information of a cluster, which cover the physical information of node, the user profile, the records of the software packages, the status logs of all the nodes and etc. As there are no a common format configure files in UNIX system, in Bluemin, many tables of the database refer to a certain configuration file of the cluster system, but the system guarantee the consistent of changes. For an example, Bluemin use DHCP to define the IP address of the slave node, so, when the user change the IP address of a slave nodes in the web base GUI, the system will change the DHCP configuration file automatically.

3.3 The System Installation

Bluemin cluster is based on Redhat Linux. The goal of the installation module is to make the installation and reinstallation easier, all the slave nodes should be installed once and automatically. [9] lists some protocols we should use in the installation process, and a simple implement based on Debian Linux.

The installation of Bluemin is based on Redhat linux. At first we have to install the head node independently, and then copy all the rpm packages to the head node. Bluemin provide a configuration file to let the administrator specify the packages that will be installed on the nodes. The configuration file is a standard Redhat kickstart file, which is a text-based description of all the packages and some basic configure file of a system. Then a binary file for

nodes' installation will be produced based on the configuration file.

The whole installation process is described in Figure 2. Here, The PXE client is a program in the ROM of network interface card. The program will launch when the system boot from the network. The DHCP service offer the client node an IP address and specify the IP of he TFTP [10] server, the location and the name of the PXE boot file, and also the linux boot kernel file to the PXE client. Then the node get the linux boot kernel and release it to the memory, the kernel run and request the kickstart file through the NFS [11] service and get the RPM packages through the FTP service on the head node.

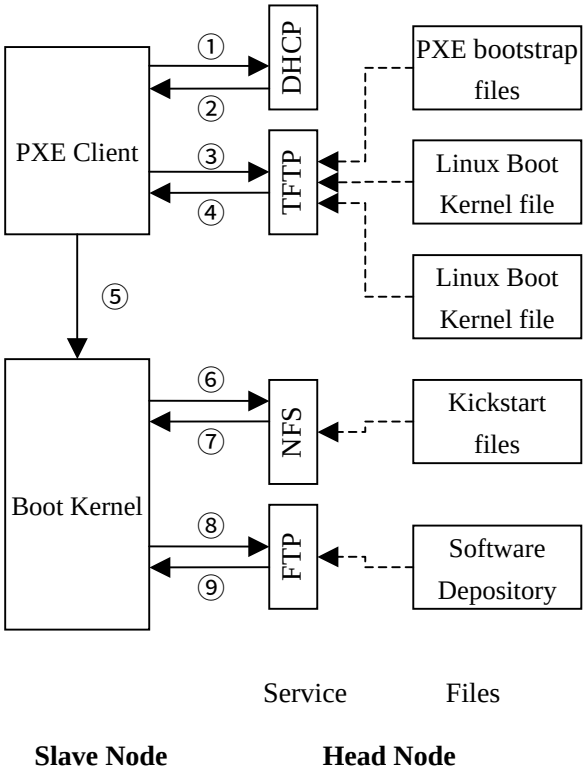

Fig.2 Process of System Installation

All the packages installed on the nodes are logged in Bluemin database for the further maintenance. Bluemin also provide an interface for the installation of new packages, administrator can decide on which nodes the packages will be installed and also all the results will be filled in the table of the cluster database. By this way, the system achieves the software heterogeneous.

Because most of the system information, such as the software package list, the user profile, are all logged into the system database, so in most of circumstance, Bluemin can help the system administrator recover one or more nodes when they crash and need not to reinstall the whole clusters.

3.4 The System Security and Cooperation

The clusters are independent PCs before we define some methods to support internal cooperation between nodes, namely, a node should be trusted to run programs on the other nodes if it is be authorized.

There are two ways to achieve such infrastructure, one is the rsh/rlogin that is known that the way has some security vulnerability, and the other way is SSH, which define a security shell protocol. In Bluemin, OpenSSH is used. OpenSSH encrypts all traffic to effectively eliminate eavesdropping, connection hijacking, and other network-level attacks. Additionally, OpenSSH provides a myriad of secure tunneling capabilities, as well as a variety of authentication methods.

In Bluemin, OpenSSH is configured to support the cooperation between nodes. The nodes trusted each other could communication and operation without user's interactive input. An OpenSSH configuration interface is also provided to help the system to update and dispatch the key pairs.

Bluemin provides a web based GUI for the system administrator. The system must guarantee the security when the administrator works with the user interface. Bluemin implement the interaction between the web browser and the web server with the SSL (*Secure Sockets Layer*) protocol, which encrypts the traffic to achieve the transfer security.

3.5 Configuration of the Cluster

There are many problem involved in the system configurations. In the single workstation UNIX environment, the Cfengine [12] is widely used, which keep the consistency of system by examination the whole OS. Bluemin partly adopts

the idea of the Cfengine, and on the other hand, Bluemin provides a single management point, the changes of the system are all make updates of the certain configuration files.

The configuration of the cluster focus on the IP, hostname, MAC address, Net mask and also the configuration files of some services such as Apache, FTP, SSH, DHCP. Sometimes the change of one item will make the whole cluster work incorrectly. Bluemin provides the database tables to keep the uniform data information and as the same time every change is spread to all the relevant configuration files. The administrator can do all the management work on the web based GUI.

The IP addresses of the slaver nodes are specified by the head node's DHCP service. The nodes will get an IP address when they boot. When the system administrator changes the IP addresses or host name of a node, it means that the DHCP configuration file and the static route file will all be update as the same time. To avoid the mismatch of the undesirable changes of the configuration files, Bluemin scan all of the related files and collect all the difference between the configuration files and the cluster database periodically, the result will be sent to the administrator. The administrator would decide which changes should be noticed and be carried out and which changes should be recovered.

3.6 User Management

For HPC cluster, the user managements are very important, because that many different users will login in the cluster to run their computing jobs.

Bluemin installs MPICH for all users to compile and run MPI programs. The execution of MPI program need the user on head node can run programs on slaver nodes, Bluemin achieve this character by using SSH but not rsh. Bluemin set the SSH to omit the interactive between trusted nodes and users. When a user is added, the SSH configuration files will be produced and dispatched automatically, which defined the id of the trusted nodes and the keys pairs of the trusted user.

The other important thing is the synchronization of the user profile between all the nodes. Bluemin uses a program running periodically on the head node to push the user profile related configuration files to all the nodes.

3.7 The System Monitor

Bluemin has the responsibility to tell the system administrator the status of the whole cluster and also to warn the administrator when some errors happen, such as the disk is full on some nodes, a certain node crashed.

Bluemin achieves a system monitor with the help of ganglia from UC Berkeley [13]. But some changes have been made to let ganglia become a real part of Bluemin. Ganglia is a lightweight distributed, multicast-based monitor system. A daemon called gmond runs on all the nodes, which monitor the status of the local node. A daemon called gmeta runs on the head node. This daemon collects all the information from each node and produces a XML file, and then pushes the system information into the RRD database.

Bluemin provides a web based front-end to show the status of the cluster. All the information is presented in graphical mode. So the system administrator can see the system status clearly.

The RRD database is not Bluemin cluster database. Bluemin enhance the functions of ganglia to helps Bluemin fill some fields of the cluster database automatically. We set the monitor collect the status of all nodes to make sure that they are all alive every ten minutes, and the result will be send to the database. All the nodes unavailable will be logged into the database and when the administrator work on the web based GUI would see that the unavailable nodes are all marked as a red color.

3.8 Other Tools

Some tools for high performance computing and some basic cluster operation are also included in Bluemin packages.

MPICH-1.2.5 and PVM-3.4 are two widely used massage-passing programming interfaces in parallel development environments. SSH is used to provide the remote process control support of the cluster, so the MPICH is configured to use the SSH protocol.

To support the job management and schedule in cluster environment, OpenPBS (*Portable Batch System*) is integrated in the whole package.

C3 [14] (*Cluster Command and Control tool*) is a toolset used to control the cluster with command line. C3 provide about ten commands such as cls (list the user's files on all the nodes), cps (list the process running on all the nodes). C3 also uses the SSH as its base infrastructure. In Bluemin, we bind the most of the C3 command to the web based GUI to help the administrator to do some works.

4 Conclusions

Bluemin achieves a complete suite of a cluster management tools. The packages cover the cluster installation, the system configuration, the user management, the system monitor and the integrated softwares. Bluemin also integrates some protocols to provide a standard interface. The system provides a database to log most of the information of the cluster and also provides a uniform web based user interface to integrate all managements GUI. Bluemin collects many excellences of the existing methods and software packages. All these characteristics make Bluemin work effectively.

In the development Bluemin, we look forward to integrate more software packages and more management functions into Bluemin, and on the other hand, we want to do more work on the suite to let it be suit for the development of new hardware and software.

References:

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

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

[3] Redhat Linux 9.0: The official Redhat linux customization guide, *http://www.redhat.com/docs/manuals*, 2003.

[4] P. M. Papadopoulos, M. J. Katz, and G. Bruno, NPACI Rocks: Tools and Techniques for Easily Deploying Manageable Linux Clusters. *Proceedings of IEEE Cluster 2001*, pp.258-267.

[5] B. N. Chun and D. E. Culler, REXEC: A Decentralized, Secure Remote Execution Environment for Clusters, *Proceedings of 4th Workshop on Communication, Architecture, and Applications for Network-based Parallel Computing*, 2000.

[6] Scyld Beowulf, *http://www.scyld.com/.*

[7] ClusterworX, *http://www.linuxnetworx.com/.*

[8] P. Uthayopas, T. Angsakul, and J. Maneesilp, System management framework and tools for beowulf cluster, *Proceedings of HPC Asia2000*, 2000.

[9] T. Hiroyasu, M. Miki, K. Kodama, J. Uekawa, and J. Dongarra, A Simple Installation and Administration Tool for the Large-scaled PC Cluster System, *Proceedings of ClusterWorld Conference and Expo*, 2003.

[10] TFTP protocol (revision 2), *http://www.ietf.org/rfc/rfc1534.html.*

[11] B. Callaghan, B. Pawlowski, and P. Staubach, RFC 1813: NFS version 3 protocol specification, 1995.

[12] M. Burgess, Cfengine: a site configuration engine, *USENIX Computing Systems*, Vol.8, 1995.

[13] Ganglia cluster monitoring toolkit, *http://www.millennium.berkeley.edu/ganglia/*

[14] C3, *http://www.csm.ornl.gov/torc/C3/.*

Same with pc, pc version made to use all yr pcs resources and game run smothly with better picture and better fps. And on other hand u can use emulation for running game but mobile games optimized for mobiles iron its totally different platform and when u use emulator yr pc uses most of resources to make emulation working. Its not about game its all about good emulator and emulation where u waste yr pcs resourses. The MegaRAID Management Suite Data Sheet is an overview of the next-generation MegaRAID Management tool suite, including a detailed description of MegaRAID Storage Manager and overviews of MegaCLI and WebBIOS utilities. Version: 1.0.Â This document explains how to set up high-availability direct-attached storage (HA-DAS) clustering on a Syncro CS 9361-8i and Syncro CS 9380-8e configuration after you configure the hardware and install the operating system. The Syncro CS solution provides fault tolerance capabilities as a key part of a high-availability data storage system. The S Version Microsoft Download Manager. Manage all your internet downloads with this easy-to-use manager.Â Windows Server 2008, Windows Server 2012 R2, Windows Server 2003, Windows Server 2012. The Windows Server Cluster Management Pack for Operations Manager is designed for the following versions of System Center Operations Manager: â€¢ System Center Operations Manager 2012 â€¢ System Center Operations Manager 2012 SP1 â€¢ System Center Operations Manager 2012 R2. Install Instructions. See MP Operations Guide. Using the --cluster-type=none option allows users to skip all cluster-related checks or modifications to the sosreport command that gets run on the nodes, and simply collect from a static list of nodes passed through the --nodes parameter. Red Hat Satellite is now a supported cluster type to allow collecting sosreports from the Satellite and any Capsules.