Cluster Distributions Review

Emir Imamagić Damir Danijel Žagar

Department of Computer Systems,
University Computing Centre, Croatia
{eimamagi, dzagar}@srce.hr

Abstract

Computer cluster is a set of individual computers, connected with high-speed networks, functioning and acting as a single computer. Cluster distribution is a software suite used to install a computer cluster. Specially, group of applications needed to provide cluster functionality is called cluster middleware.

In this paper, we will describe set of components used as building blocks for cluster middleware that are supposed to be contained in any cluster distribution. Furthermore, we will discuss few Linux-based cluster distributions. Based on our experiences and evaluation results, we will recommend some distributions.

1. Introduction

Computer cluster is a set of network-interconnected individual computers that function as single computer. Cluster installation and configuration may be a very difficult task. It begins with hardware resources deployment and configuration. Following that, Operating System (OS) is installed on the main (server) computer and compute nodes. Installation of the cluster middleware tools is the last step in the cluster deployment process.

Cluster middleware enables job execution on different computers within the cluster. Usually it includes job management system (JMS), parallel execution libraries, runtime environment and tools for node management and monitoring.

Cluster maintenance may impose even bigger problem than the initial installation. Once everything is configured, administrator's duty is to regularly update software packages and install various security patches. Updating cluster middleware usually demands performing synchronized and simultaneous updates on many different systems and large number of computers.

Phrase "cluster distribution" comes from the analogy with OS distribution. Idea of cluster distribution is to provide integrated set of software components needed for cluster installation, configuration and control. Usually it consists of cluster middleware tools, typical and most common cluster applications and other software components. Some cluster distributions integrate OS itself, but most of them allow user to choose their own.

For cluster administrators, cluster distribution provides and simplifies cluster setup, management and maintenance tasks.

2. Cluster Middleware

Cluster middleware is a set of applications needed for a single computer to transparently cooperate with other

computers thus creating an impression of the single computer.

Cluster middleware (shown in Figure 1) consist of:

- job management system
- cluster monitoring system
- parallel execution libraries
- cluster management tools
- global process space

Job Management System (JMS) is cluster component responsible for user's job control, their dispatching and scheduling. In an ideal case, user accesses and communicates with the cluster only through JMS. JMS provides functionality for user to describe their jobs, monitor jobs in progress, control jobs and fetch results. Administrator uses JMS to implement various usage policies, such as fair sharing, limiting resource consumption, making advance reservation, etc. Furthermore, JMS is used by cluster owner to monitor cluster usage and gather usage statistics information which may be later used to charge users for consumed time or analyze cluster utilization.

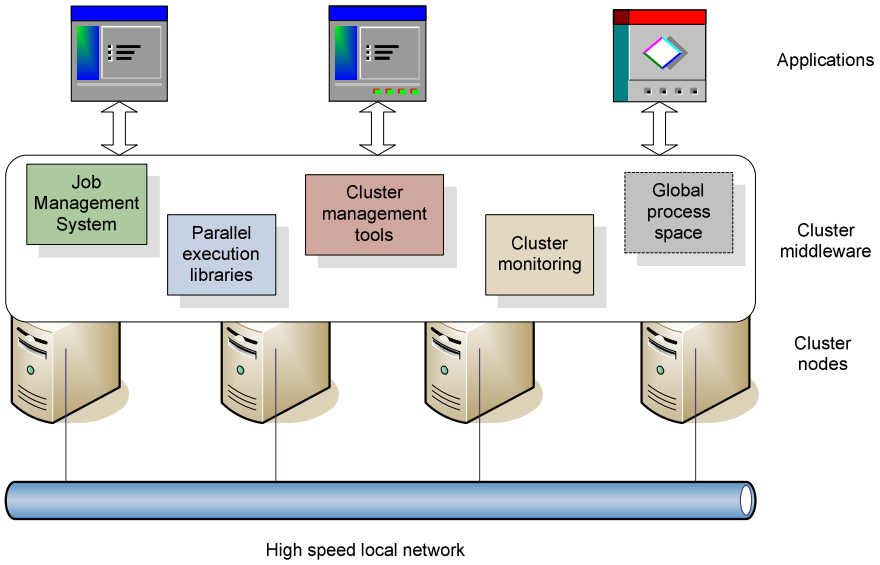

Figure 1 Cluster Middleware

Cluster monitoring system collects various information concerning cluster nodes - system load, amount of free memory, amount of free disk, active processes, etc. Collected information is used by cluster administrator to identify potential problems within the cluster or to record cluster utilization. Ordinary users may use these data to analyze and debug execution of their jobs. Furthermore, information about nodes is used by JMS in the process of job scheduling.

Parallel execution libraries provide application developers methods for achieving application parallelization and synchronization among different tasks (processes) executing on the same or different computer. Most common parallel libraries are MPI and PVM. Various parallel programming languages can also be used for parallel application development.

Cluster management system enables control over cluster nodes. It is supposed to

be used by the administrator for recovering from error conditions or for additional system configuration.

One of the most important tools is automatic installer – system used for installing software stack on remote computers.

Global process space functionality provides control of all processes running on the cluster from any node within the cluster. Having it is not necessary for a cluster to function properly, thus many clusters do not provide it, but its advantage is to allow higher level of process control over the entire cluster.

Instead of integrating different software components on top of the existing OS, cluster middleware can also be implemented as integral part of the operating system (one example is MOSIX cluster). In such case middleware subsystems are not distinguished and cluster functionality is achieved through set of different kernel modules. In this paper, we are limited to cluster distributions that provide cluster middleware in form of software components built on top of the operating system.

3. Cluster Distributions Overview

We have evaluated following Linux-based cluster distributions: Rocks, OSCAR, OpenSCE, Scyld Beowulf, Clustermatic, Warewulf, xCAT and SCore. Many other cluster distributions are available but not widespread as those previously mentioned. Numerous enterprise solutions exist as well. Comments here stated are based on available documentation and our own experiences.

3.1. NPACI Rocks

One of nowdays most popular cluster distributions is NPACI Rocks [2]. Rocks cluster distribution is based on Red Hat Linux distribution. It uses Red Hat kickstart mechanism to deploy software packages to compute nodes. Software packages are arranged in rolls, every roll containing set of specific packages. Rocks enables usage of several Job Management Systems – SGE, OpenPBS and Condor. Rocks is updated regularly and roll-based system provides scalable installation.

Rocks major advantage is cluster middleware software scalability and full automation of the cluster installation process, being almost the the same as the RedHat installation on the single computer.

Integrating and depending with the specific operating system is also Rocks main disadvantage as update of the cluster distribution on the front-end node will probably result in complete (from-the-scratch) cluster installation. Update option is available but it does not differ in many details from the complete reinstall.

3.2. OSCAR

Open Source Cluster Application Resources (OSCAR) [4] consists of typical cluster middleware software. In contrast to Rocks, OSCAR is not bound to any specific OS. User can choose to install Red Hat or Mandrake Linux on the main (server) node and OSCAR will provide set of cluster middleware RPMs and automatic installation tool – System Installation Suite. Cluster middleware package contains OpenPBS JMS, LAM/MPI, MPICH and PVM parallel execution libraries and Clumon and Ganglia node monitoring system.

3.3. OpenSCE

Compared with other cluster distributions, OpenSCE [3] has slightly different approach. While other cluster distributions are trying to integrate various

existing solutions, OpenSCE provides its own set of tools: parallel execution library, job management system, monitoring system and automatic installation tool. OpenSCE provides special cluster middleware tool - KSIX. KSIX creates global process space for processes on all cluster nodes.

Disadvantage of OpenSCE is that its set of new cluster tools lacks evaluation and tests that other tools have already passed.

3.4. Scyld Beowulf

Scyld Beowulf [7] is commercial cluster solution. Scyld uses BProc system to create global process space. In contrast to Rocks and OSCAR, automatic installation tool installs very limited (lightweight) set of software and OS packages on the compute nodes. Scyld comes with standard set of parallel execution libraries. Advantage of Scyld is global process space, which assures that there will not be any uncontrolled "runaway" processes on compute nodes.

Scyld Beowulf does not integrate any job management system, but it is possible to use some of the existing job management systems (PBSPro).

3.5. Clustermatic

Similar to Scyld, Clustermatic [1] uses BProc to achieve global process space. It provides tool for automatic compute nodes installation, which installs very limited set of packages on compute nodes. Clustermatic comes with ZPL programming language and Supermon monitoring system. Disadvantage of Clustermatic is lack of advanced job management system.

3.6. Warewulf

Warewulf Cluster Project [9] is one of the youngest cluster distribution projects. Main goal of Warewulf cluster distribution is to achieve functionality somewhere between Scyld Beowulf and Rocks or OSCAR. Warewulf uses minimal OS stack on compute nodes (similar to Scyld Beowulf) in the same time enabling usage of advance cluster middleware systems (like Rocks or OSCAR). Warewulf uses SGE system for job management, Ganglia for node monitoring and LAM MPI and PVM parallel execution libraries.

Its major advantage is small OS stack on compute nodes and capability for multiple networks usage - nodes haveing multiple network interfaces can use specific interface for node management, shared file system and parallel communication. Being ‘new on the cluster scene’ it is still not mature enough and the user community is relatively small.

3.7. xCAT

xCAT (Extreme Linux Cluster Administration Toolkit) [8] is IBM's cluster distributions for IBM servers. xCAT, like Rocks, it uses RedHat as base OS and RedHat kickstart mechanism for nodes installation. Job management is performed through OpenPBS and Maui. For monitoring xCAT uses Ganglia monitoring system and integrates MPICH, LAM MPI and PVM parallel execution libraries.

As xCAT distribution is tightly integrated with IBM Management Processor Network system for node management and monitoring xCAT is considered to be hardly portable. Beside that, we consider that xCAT's functionalities are inferior to NPACI Rocks. Major disadvantage is that process of cluster installation is poorly automated -

user has to configure everything by editing configuration files.

3.8. SCore

One of the oldest cluster distributions - SCore [6] is made by Real World Computing Partnership (RCWP) group. At the year 2002, RCWP group ended their work and SCore was taken over by PC Cluster Consortium group. SCore, like OpenSCE, has many of its own components. The core of the system is SCore-D. SCore-D, like BProc and KSIX, is layer above node OS's that achieves global process space. Second most important component of SCore cluster distributions is PMv2 library for optimization of network communication. SCore also has PVM and MPICH implementation based on PMv2 library. For job management and monitoring SCore uses external systems: OpenPBS and Ganglia.

Major weakness of SCore distribution is lack of future development milestones and directions. We expect that SCore will eventually become a commercial product.

4. Conclusion

Biggest advantage of the cluster distribution is that they integrate all necessary cluster software components and their regular updates and patches. Having broad user base and based on their feedback, comments and wishes, cluster distributions are constantly being developed and evolving.

Most mature open source cluster distributions currently available are NPACI Rocks and OSCAR. Mainly because of the global process space functionality - Scyld Beowulf is one of the most interesting commercial solutions. In addition, we consider Warewulf as the most promising cluster distribution on the horizon.

5. References

[1] Clustermatic, http://www.clustermatic.org

[2] NPACI Rocks, http://www.rocksclusters.org

[3] OpenSCE, http://www.opensce.org

[4] OSCAR, http://oscar.openclustergroup.org

[5] P. M. Papadopoulos, M. J. Katz, G. Bruno: **"NPACI Rocks: Tools and Techniques for Easily Deploying Manageable Linux Clusters"**

[6] SCore, http://www.pccluster.org

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

[8] xCAT, http://xcat.org/

[9] The Warewulf Cluster Project, http://warewulf-cluster.org