A Comparison of Provisioning Systems for Beowulf Clusters

Mario Trangoni, Matías Cabral

{mario.trangoni,matias.a.cabral}@intel.com
Argentina Software Design Center (Intel)

Abstract. Provisioning Systems were developed to reduce the effort required to install the several components included in the hardware and software stack of High Performance Computing (HPC) clusters. These systems are designed to be turnkey solutions, providing predefined configurations and a myriad of extra tools for management and development. However, as the volume cluster ecosystem grows, so does the number of provisioning systems, and the prospective user has to decide which system is the most adequate. This paper reports a comparative analysis of five provisioning systems for HPC clusters. The analysis was realized as part of the Intel® Cluster Ready program, but the core of the comparison between systems is useful for any organization that wants start using HPC clusters.

1 Introduction

A modern High Performance Computing (HPC) cluster consists of a diverse stack of hardware and software. The hardware includes servers acting as nodes, one or more communication networks, and a storage system. The software stack, whose general structure is shown in Figure 1, usually includes: (1) an operating system, in most cases a Linux[*] distribution, (2) a provisioning system middleware that allows to install the software stack and configure the cluster, (3) tools for monitoring and managing the cluster's status, configuration and resources, (4) software development tools and libraries, (5) a parallel file system (optional), and (6) the applications, the programs that the users uses to do their job.

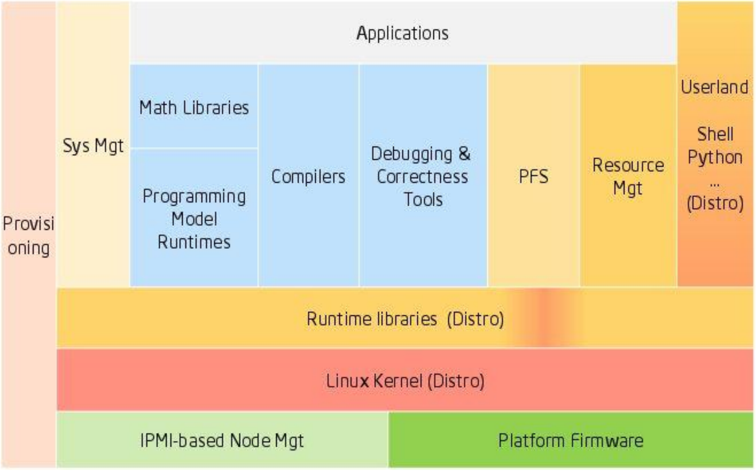

Fig. 1. HPC components Stack.

Installing and managing all these components is a complex task. In the past, system administrators often took on the responsibility for hardware and software provisioning on the cluster. This approach

[*] Other names and brands may be claimed as the property of others.

was time consuming and only useful as a “one time” custom solution. In order to reduce the effort required to install and maintain an HPC cluster and to make easy to replicate the tasks involved, a series of semi-automated tools were developed, called provisioning systems. These systems are designed to be turnkey solutions, providing predefined configurations and tools that allow easy deployment of an HPC cluster.

In this paper we report our analysis and comparison of several provisioning system. As part of our work on the Intel® Cluster Ready program [23], we needed establish the “state of the art” on provisioning systems and determine, under certain specific conditions, what systems are better suited than the others. The basis of our analysis is a comparison with the HPC standard established in the Intel Cluster Ready Specification v.1.2 [19] and the ability of being integrated with our own tools. However, since most of the criteria used for our analysis are based on common needs of all HPC users, we report in this paper the results of comparing five current provisioning systems, without including the few Intel Cluster Ready-specific criteria.

2 Evaluated Provisioning Systems

There are plenty of provisioning systems, but only a handful of them are currently supported by a company or community. We decide to analyze the five provisioning systems that we know that are supported. The final list of systems analyzed is the following.

1. *Bright Cluster Manager* [1], of Bright Computing, is a proprietary product with an intuitive graphical interface.
2. *Platform HPC* [3], of Platform Computing, now a subsidiary of IBM, is one of the pioneers in this type of tools.
3. *Rocks+* [5], of StackIQ, a derivation of Rocks, one of the oldest and most popular open source provisioning systems. Currently, their market strategy seems to be toward cloud computing.
4. *Warewulf* [9], an open source system derived from CAOS/Perceus and developed by the Lawrence Berkeley National Laboratory [14].
5. *xCAT* [11], an open source provisioning system from IBM, oriented to large systems (up to 100,000 nodes in a hierarchical routing infrastructure).

Other provisioning systems have been left out of the comparison, even when they are well known or currently in use in several clusters around the world. For example: OSCAR [15], Rocks [7], Scyld from Penguin Computing [16], and CHAOS [17].

3 Methodology and Analysis

In this section we report the results obtained of analyzing the twenty criteria that we consider of interest to general HPC cluster users. Each criterion is briefly described and the results obtained for each provisioning system is shown on a table.

3.1 Licensing model

A proprietary software license usually involves paying a fee for its use during a limited time. In some cases, per cores fees are really huge. A free/open source licensing model usually allows access to the source code, a must for Academic and R&D institutions. On the other hand, companies prefer a strong support offering.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
Paid, time limited with support.	Paid, time limited wit support	Paid, free until 16 nodes	BSD License [21]	Eclipse Public License [20]

3.2 Supported Linux Distributions

A provisioning system that supports different Linux distributions provides flexibility to cluster vendors to cater more users.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
− Scientific Linux 5/6 − RHEL 5/6 (*) − CentOS 5/6 − SLES 11(*)	− Scientific Linux 5.5 − RHEL 5/6(*) − CentOS 5.6 − SUSE 11(*)	− Oracle Linux 6 − RHEL 5/6(*) − CentOS 5/6	− RHEL 5/6(*) − CentOS 5/6 − Debian − Ubuntu	− SLES 10/11(*) − RHEL 5/6(*) − CentOS 5/6 − Fedora 8/9/12/13/14 − AIX 5/6/7(*)

(*) Extra license, must be paid.

3.3 Scalability

This criterion refers to the number of compute nodes that can be provisioned within a reasonable time, and at the same time still be properly managed by the cluster's managing tool.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
1,023 nodes	5,000 nodes	1,000 nodes	10,000 nodes	100,000 nodes

Note: Numbers obtained from documentation or hardware vendors. Not tested.

3.4 Provisioning Method

It is the method by which all compute nodes are installed. This criterion also influences the system scalability and how versatile against cluster configuration changes a provisioning system can be.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
rsync, load kernel and ramdisk via PXE, create the local file system and then compare with head what to be provisioned	Packages (default) and Images	Packages, Avalanche Ad-Hoc Peer-to-Peer Package Serving Network	Images, VNFS images created by Administrators	Images (default) and Packages

Most of research here is related to scalability and provision's speed. Imaged methods suffer when referring to scalability. Keep in mind that provisioning systems must provide also manageability.

3.5 Job Schedulers automatically configured

When clusters are used by relatively big number of users, it is vital to have a Job Scheduler (JS). A JS automate submission of executions and define priorities and/or queues to control the execution order of unrelated jobs. Because of those reasons, when different users share resources (cluster), a JS simplify administrator work.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
– Oracle Grid Scheduler (v6.2u5p2)(*) – Torque (v2.5.5) – Maui – Moab(*) – Pbs Pro (v11.0.2) (*) – Slurm (v2.2.4)	– Platform LSF-Master v7.0.6 (*)	– SGE(Oracle)(*) – Moab(*) – Univa Grid Engine(*) – PBS Pro(*) – LSF Roll (Platform)(*)	No	– Torque – Moab(*) – PBS Pro(*)

(*) Extra license, must be paid.

3.6 Support for third party software add-on

Support additional software as a plug-in, shortens configuration time and the level of expertise required for cluster administrators is reduced.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
Yes	Yes	Yes	No	Yes

3.7 Add-on for Software Development Tools

Sometimes clusters are required to provide development tools such as compilers and profilers.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
– GNU Compilers – Intel Compilers(*) – PGI High-Performance Compilers(*) – AMD Open64 Compiler Suite	– GNU Compilers	– GNU Compilers – CUDA roll – Absoft Roll* (Absoft Compilers) (*) – Intel Developer Roll(*) – PGI Roll(*) – Totalview Roll(*)	– GNU Compilers	– Intel Compilers(*) – PGI Compilers(*) – GNU Compilers

(*) Extra license, must be paid.

3.8 Add-on for MPI libraries

This criterion refers to the libraries required by the Message Passing Interface standard. Provisioning systems could offer open source libraries (example: MPICH, OpenMPI) and proprietary libraries (example: Intel MPI, Platform MPI).

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
– OpenMPI – MPICH – MPICH2 – MVAPICH – MVAPICH2 – MPICH-MX	– Platform MPI(*) – MPICH1 – MPICH2 – MVAPICH1	– HPC Roll (open libraries) – Extra with Intel Roll(*) – Extra with PGI Roll(*)	No	– MPICH – MPICH-GM

(*) Extra license, must be paid.

3.9 Add-on for Mathematical libraries

In the same way as above, this criterion refers to mathematical libraries of third parties that offer hardware-optimized versions (example: Intel MKL) or open source libraries (example: BLAS, LAPACK).

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
– Intel MKL(*) – ScaLAPACK – GotoBLAS – ATLAS – GMP – FFTW – AMD Core Math Libraries – Intel IBB(*) – Intel IPP(*) – GlobalArrays – HDF5 – NetCDF – PETSc	– ScaLAPACK – ATLAS – FFTW – NetCDF – HDF5	– Intel MKL(*) – Open Source math libraries	No	– Intel MKL(*) – AMD Core Math Libraries – Goto Libraries – ATLAS libraries

(*) Extra license must be paid.

3.10 Disk-less Provisioning support

Many times, depending of necessities of the administrator (time to deploy, simpler nodes images, size of memory), disk-less installation is required or desired. Disk-less installation is more versatile, staying only in compute nodes RAM.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
Not validated (1)	Not functional (2)	Not functional (2)	Yes	Yes

(1) Vendor says that it is supported, but it was not tested.
(2) Vendor says that it is supported, but we could not make it work.

3.11 Command Line Interface Cluster Management Tool

This criterion analyzes different implementations of command line interface. Provisioning systems usually have specific commands that are different from usual Linux sentences. They are focused in automation possibilities.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
Cmsh	kusu-ngedit	ROCKS commands	wwsh	XCAT commands

3.12 Graphical User Interface Cluster Management Tool

To make tools more user friendly, some provisioning systems offer graphical interfaces, adapted to the characteristics of their products.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
cmgui	Platform Web Portal	No	No	No

3.13 Built-in Monitoring Tools

Monitoring tools offer the possibility of observing cluster operation, showing the status of jobs that are currently running, and resources in use and/or available.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
– cmsh – cmgui	Platform HPC Web Portal	Ganglia Monitor	No	– Ganglia – IBM's RSCT(*) – xcatmon

(*) Extra license must be paid.

3.14 Built-in Parallel Shells

Parallel shells allow running standard Linux commands on all compute nodes at the same time. They are developed aiming at a better scalability in the execution compared to using standard SSH.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
pexec	pdsh	Tentakel	No	psh, xdsh

3.15 User management and administration Tool

This management tool is one of the most important components of provisioning systems, because adding users to a cluster impacts across its entire infrastructure and is very useful to have an automated method of administration.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
cmsh, cmgui	kusu-cfmsync	ROCKS commands	No	Yes

3.16 Cluster security features

This criterion refers to the software dedicated to protect the access to the cluster. Its importance lies in the protection of the users’ information and ensuring the proper execution of the tasks by not allowing unruly interruption of the jobs running in the cluster.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
OS tools configured	Platform ISF(*)	rocks sync host sec_attr	No	IBM Tivoli Security Tools(*)

(*) Extra license, must be paid.

3.17 Database used

On a provisioning system, the database stores all the configuration management data of the cluster. The data are store at the head node and a goo database simplifies the re-provisioning of individual compute nodes.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
MySQL	PostgreSQL	MySQL	MySQL	– SQLite (Default) – MySQL – PostgreSQL

3.18 Commercial support

This criterion takes in account the availability of paid support, offered by product development companies.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
Yes	Yes	Yes	No	Yes, from IBM

3.19 Configured InfiniBand support

InfiniBand (IB) interconnections offer normally more performance than an Ethernet. Most of the Top 500 [22] systems are connected with IB (462 were listed in June 2012). One important aspect is that they must be correctly configured and this depends on the configuration and libraries installed.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
Yes	Yes	OFED Roll (Mellanox binaries)	No	Yes

3.20 Cloud Support

This criterion analyzes the options offered to burst into a public cloud, such as Amazon EC2. This characteristic allows for extra flexibility, compared to that user's hardware can provide.

BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
Amazon EC2	Amazon EC2	Amazon EC2	No	Moab Adaptive Computing Suite

To analyze these criteria, we carried out default installation of the systems under the most uniform conditions possible, taking as OS for the installation RedHat / CentOS version 6.2 Linux distributions. Some of the information published here was obtained from the documentation provided by each provisioning system, while certain points required exploring the system and its components available. Provisioning systems owners were contacted to discuss support level.

4 Results of the Evaluation

When analyzing the different components of an HPC cluster, including the provisioning systems as proposed in this work, there is no absolute and unquestioned best option, but a better option according to the needs of the user. Those needs should be formally established through a Decision Analysis and Resolution (DAR) process. This process requires defining the evaluation criteria according to the specific needs of the cluster user and giving to each of the criterion a “weight” according to the impact it has. Table 1 shows the weight we assigned to each criterion, from 1 (least relevant) to 10 (most relevant), based on the needs of the Intel Cluster Ready program.

Licensing Model	3	Command Line Interface Cluster Management Tool	9
Supported Linux Distributions	5	Graphical User Interface Cluster Management Tool	2
Scalability	6	Built-In Monitoring Tools	4
Provisioning Method	1	Built-In Parallel Shells	9
Job Schedulers Automatically Configured	3	User Management and Administration Tool	9
Support for Third Party Software Add-On	7	Cluster Security Features	4
Add-On for Software Development Tools	2	Database Used	2
Add-On for MPI Libraries	7	Commercial Support	7
Add-On for Mathematical Libraries	7	Configured InfiniBand Support	4
Disk-Less Provisioning Support	6	Cloud Support	2

Table 2. Weights for DAR.

Then, each criterion must be evaluated assigning it a value from 1 to 10 based on how good each provisioning is with respect to that criterion. At the end of the process there is a table with pondered results that allows a rational comparison of the provisioning systems considered. Table 2 shows the results of our analysis, establishing the Bright Cluster Manager version 5.2 as the most convenient solution for the needs of our project. However, we are not saying that this provisioning system is the best. We are including this analysis step only to show an example of how the prospective user of an HPC cluster should analyze the results of our comparison in order to obtain the best option for her needs.

Provisioning System	BCM 5.2	Platform HPC 3.0.1	ROCKS+ 6	Warewulf 3.1	xCAT 2.6.9
Score	1006	852	816	378	745

Table 2. Results of the DAR.

5 Conclusions

During the execution of the Intel Cluster Ready program, we were faced with the selection of a modern provisioning system that met certain criteria. In this work we report the analysis done over five of the currently most used provisioning systems, focused on 20 criteria that we consider relevant for most of the users of HPC clusters.

Our contribution is a complete analysis of the most used provisioning systems for HPC clusters that can be used for prospective users to feed their DAR process, where they should feed the process with the particular weights assigned to each criterion.

6 References

[1] Bright Computing,Inc. R , Official website, http://www.brightcomputing.com/, March 2012.
[2] Bright Cluster Manager version 5.2, Administrator Manual, Revision 2029, December 2011.
[3] Platform Computing an IBM Company, Official website, http://www.platform.com/, March 2012.
[4] Administering Platform HPC, Platform HPC Version 3.0.1, September 2011.
[5] StackIQ Inc., Official website, http://www.stackiq.com/ , March 2012.
[6] Stack IQ Base Users Guide, StackIQ ROCKS+ 6.0 Edition, StackIQ Inc. & University of California, November 2011.
[7] ROCKS, Official website, http://www.rocksclusters.org/, March 2012.
[8] ROCKS Discussion List, Discussion of Rocks Clusters, https://lists.sdsc.edu/mailman/listinfo/npaci-rocks-discussion/, March 2012.
[9] Warewulf, Official website, http://warewulf.lbl.gov/, March 2012.
[10] General Warewulf Discussions, “”Users Discussion List”, https://groups.google.com/a/lbl.gov/group/warewulf/, March 2012.
[11] xCAT Extreme Cloud Administration Toolkit, “Official website”, http://xcat.sourceforge.net/, March 2012.
[12] xCAT Users Mailing list, xCAT-user, https://lists.sourceforge.net/lists/listinfo/xcat-user, March 2012.
[13] xCAT Developers Community, xCAT 2 Cookbook for Linux, http://docs.huihoo.com/xcat/xCAT2.pdf, July 2008.
[14] Lawrence Berkeley National Laboratory, “Official website”, http://www.lbl.gov/, 2012.
[15] OSCAR, “Open Source Cluster Application Resources, “Official website”, http://svn.oscar.openclustergroup.org/trac/oscar, 2012.
[16] Penguin Computing, “Official website”, http://www.penguincomputing.com/Products/ClusterManagement, 2012.
[17] Chaos-release, “Official website”, http://code.google.com/p/chaos-release/wiki/CHAOS_Description , 2012.
[18] Decision Analysis and Resolution (DAR), “Software-Quality-Assurance.org Website”, http://www.software-quality-assurance.org/cmmi-decision-analysis-and-resolution.html, 2012.
[19] Intel® Cluster Ready - Document Library, “Intel® Cluster Ready Specification 1.2”,http://software.intel.com/file/36320, 2011.
[20] Eclipse Foundation, “Eclipse Public License v10”, http://www.eclipse.org/legal/epl-v10.html , 2012.
[21] Linux Information Project, “BSD License”, http://www.linfo.org/bsdlicense.html , 2012.
[22] TOP500.Org, “Top 500 List June 2012”, http://www.top500.org/list/2012/06/100, 2012.
[23] Intel® Cluster Ready Program – http://www.intel.com/go/cluster, 2012.