

Platform Rocks: A Cluster Software Package for Dell HPC Platforms

Administrators can employ cluster solution packages such as Platform Rocks to help deploy, maintain, and manage high-performance computing (HPC) clusters. Based on the Linux® OS and NPACI Rocks, Platform Rocks includes drivers and other features that can provide comprehensive cluster management tools for Dell™ HPC platforms.

BY RIZWAN ALI, RINKU GUPTA, GARIMA KOCHHAR, AND BILL BRYCE

Related Categories:

Cluster management

Clustering

Dell PowerEdge servers

High-performance computing (HPC)

NPACI Rocks

Platform Computing

Systems management

Visit www.dell.com/powersolutions for the complete category index.

Twice a year, a group of manufacturers, computational scientists, experts in high-performance computing (HPC), and members of the Internet community compile and disseminate a list of sites that operate the 500 most powerful computer systems in the world. In the last five years, the percentage of clusters in the TOP500 Supercomputer Sites list has grown from 2.2 percent to 60.8 percent.¹ During the same time frame, the typical size of these clusters has grown from a few hundred processors to thousands of processors. Today, the largest Intel® processor-based cluster comprises more than 10,000 processors.²

As clusters expand, cluster deployment, maintenance, monitoring, and management can become complex and time-consuming processes. Before an application can tap the tremendous computing power available to an HPC

cluster, administrators must install and configure the cluster for monitoring and management. Manually deploying even a small HPC cluster is no small task. The job demands that each cluster node and its OS be identically configured and installed, including proper drivers, networks, parallel libraries, and management software.

A cluster solution package is a semiautomated tool that helps administrators accomplish deployment, maintenance, and management tasks on an HPC cluster.³ NPACI Rocks^{4,5} is an open source, Linux-based software stack for building and maintaining HPC clusters. Platform Rocks⁶ is an enterprise-level version of NPACI Rocks—developed by Platform Computing Inc.—that has been validated and verified on Dell PowerEdge™ servers by the Dell HPC Cluster team and Platform Computing.⁷

¹ For more information, see the TOP500 Supercomputer Sites Web site at www.top500.org. Results are available at www.top500.org/lists/2000/06 for the June 2000 list and at www.top500.org/lists/2005/06 for the June 2005 list.

² This cluster is NASA's Columbia Supercomputer; for more information, visit www.top500.org/sublist/System.php?id=7288.

³ For more information, see "Felix, Scali, and Rocks: An Introduction to Cluster Computing Solutions" by Baris Guler, Rinku Gupta; Saeed Iqbal, Ph.D.; and Monica Kashyap in *Dell Power Solutions*, October 2004, www.dell.com/downloads/global/power/ps4q04-20040139-Gupta.pdf.

⁴ For more information, see "Streamlining Beowulf Cluster Deployment with NPACI Rocks" by Rinku Gupta, Yung-Chin Fang, and Munira Hussain in *Dell Power Solutions*, February 2005, www.dell.com/downloads/global/power/ps1q05-20040176-Gupta.pdf.

⁵ For more information about NPACI Rocks or to download Rocks, visit www.rocksclusters.org/Rocks.

⁶ For more information about Platform Rocks or to download Platform Rocks, visit www.platform.com/products/Rocks.

⁷ For more information, visit www.dell.com/hpcc.

Introduction to NPACI Rocks

The National Partnership for Advanced Computational Infrastructure (NPACI) designed the NPACI Rocks toolkit in November 2000 to help simplify the building and management of clusters. Rocks—a simple, self-contained, cluster-aware management system that is scalable and upgradeable—is slowly becoming the de facto cluster package. In October 2005, registered users of NPACI Rocks reported computational power that totals 149 trillion floating-point operations per second (TFLOPS).⁸

NPACI Rocks version 4.0.0 appeared in June 2005. Previous versions of Rocks contained two components—a base CD and Roll CDs. The base CD included minimal components for installation, including a recompiled Red Hat® Enterprise Linux OS.

Rolls are add-on components that provide specific capabilities to a cluster. Administrators can create Rolls to contain domain-specific software and applications; third-party vendors can customize Rolls to answer specific needs and requirements. Rather than using the base CD and assorted Rolls, the latest Rocks release employs a Roll-based framework. With Rocks 4.0.0, even core software packages for building the cluster are furnished in the form of a Roll—for example, Base Roll, Kernel Roll, HPC Roll, OS Roll, and so forth—and are considered essential for an HPC Rocks installation. An OS Roll separate from the core Rocks components allows organizations to create and use their preferred OS, such as Red Hat Enterprise Linux or Scientific Linux, to install their clusters. Apart from the OS Roll, Rocks 4.0.0 also allows administrators to use regular OS CDs to install the cluster.

Platform Rocks functionality

Platform Rocks is designed to improve upon NPACI Rocks by providing additional functionality and enhanced support. The Dell HPC Cluster team and Platform Computing have tested Platform Rocks on Dell hardware,⁹ and both Platform Computing and Dell provide support for enterprise customers.

As with NPACI Rocks, Platform Rocks has “base” components and optional modules packaged in “Rolls.” Similarly, the base components of Platform Rocks *Standard* edition are designed as an open

source software stack to automate and streamline cluster installation management. The *Standard* edition is designed for the Community Enterprise Operating System (CentOS)¹⁰ and has no support option. However, it features the same set of tools, options, and configurations as the *Enterprise* edition.

The Platform Rocks *Enterprise* edition comes packaged with Red Hat Enterprise Linux AS on the front-end node and Red Hat Enterprise Linux WS on the compute nodes. It also provides a support option through Annual Cluster Care, which is a subscription service available from Platform Computing.

The *Enterprise* edition of Platform Rocks features a set of cluster tools, including the following:

- Message Passing Interface (MPI) libraries and drivers
- Cluster management tools and a workload manager
- Interconnect drivers and libraries to support interconnects such as Gigabit Ethernet, Topspin InfiniBand, and Myricom Myrinet
- Drivers for Dell hardware

Rocks management tools include Anaconda and kickstart for hosting images; Ganglia for monitoring; Cluster Top for sensing the precise activity of cluster nodes; and the 411 user management system.

Integrated with Platform Rocks as a Roll is an entry-level workload management tool called Platform Lava. This tool can be a critical component for operational management of the entire cluster, enabling job execution, management, and metrics for demanding enterprise environments in a simple, user-friendly package. Platform Lava is compatible with Platform Load Sharing Facility (LSF) HPC, providing organizations with a clear migration path from cluster workload management to enterprise workload management and enterprise grid deployments.

Platform Rocks 4.0.0 introduces several features that facilitate the administration and maintenance of a Platform Rocks cluster. Some enhancements include:

- Inclusion of Red Hat Enterprise Linux 4
- Ability to add and remove Rocks Rolls
- Integration with Red Hat Network (RHN)
- A database pre-population tool
- National Center for Supercomputing Applications (NCSA) Cluster Monitoring (Clumon) Roll¹¹
- Platform Lava Web-based graphical user interface (GUI)

⁸ For a registry of Rocks clusters, see www.rocksclusters.org/rocks-register.

⁹ For specific Dell hardware, visit www.dell.com/hpcc.

¹⁰ For more information about CentOS, visit www.centos.org.

¹¹ For more information about NCSA and Clumon, visit clumon.ncsa.uiuc.edu.

Integration with Red Hat Enterprise Linux

NPACI Rocks relies on CentOS to maintain compatibility with Red Hat Enterprise Linux. However, many enterprises require full Red Hat licenses and support. Platform Rocks contains Red Hat Enterprise Linux AS edition on the front-end node and Red Hat Enterprise Linux WS on the compute nodes. Adding Red Hat Enterprise Linux to Platform Rocks enables support for a complete stack: The hardware can be supported by Dell, the OS can be supported by Red Hat, and the Rocks framework can be supported by Platform Computing.

Ability to add and remove Rocks Rolls

Earlier versions of NPACI Rocks lacked a way to add or remove a Roll. Installing a new Roll on a cluster required rebuilding the cluster, including the front-end node, from scratch. This was consistent with the philosophy of NPACI Rocks that compute nodes are stateless. Although statelessness helps ensure compute-node consistency, cluster administrators may want to add Rolls to a cluster without having to reinstall the front end. Version 4.0.0 of NPACI Rocks introduced the capability to add Rolls to the cluster, but this version still does not allow Rolls to be removed—which can make ongoing cluster maintenance difficult.

The capability to add and remove Rolls requires an administrative command on the front end. A command called `rollops` in Platform Rocks 4.0.0 builds upon the `kroll` utility provided by NPACI Rocks. This command enables Rocks administrators to control Rolls on the front-end and compute nodes. Using `rollops`, an administrator can:

- Add Rolls to the cluster, without reinstalling the front-end node
- Remove Rolls, provided that the Rolls do not have kernel dependencies
- List Rolls currently installed on the front-end node
- Manage permissions for installations and uninstallations through a configuration file

The `rollops` command leverages the NPACI Rocks `kroll` command to install Rolls. It also provides a file called `/opt/rocks/etc/rollopsrc`, through which administrators can block access to key Rolls to prevent accidental overwriting.

Platform Lava Web-based GUI

Platform Rocks 4.0.0 includes a Web-based GUI for all Platform Lava users. The Web-based GUI is an alternative to the command-line interface and allows simple job submission, control, and monitoring for end users and administrators. This GUI can be used as a portal on the Platform Rocks front-end node and provides access to the cluster from JavaScript-enabled Web browsers.

Annual Cluster Care

The Annual Cluster Care program for Platform Rocks is available to subscribers at a per-node fee. It includes support, maintenance, upgrades, fixes, access to resources, and other services to help organizations manage their Platform Rocks clusters.

Platform Rocks and Red Hat Network

NPACI Rocks regularly updates the Rocks distribution with CentOS; Red Hat Enterprise Linux OS updates are provided by Red Hat through RHN. Platform Rocks 4.0.0 introduces the `rocks-update` command, which updates a Platform Rocks cluster using either the `up2date` (for the Platform Rocks *Enterprise* edition) or `yum` (for the Platform Rocks *Standard* edition) command. The `up2date` command obtains updated Red Hat Package Manager (RPM™) packages from RHN, and the `yum` command obtains updates from a Yum repository such as the Fedora repository.

The `rocks-update` command helps ensure that the Red Hat Enterprise Linux distribution installed by Platform Rocks remains up-to-date and has the latest security patches downloaded. It also maintains a list of RPM modules that should not be automatically updated because doing so could cause problems with the Platform Rocks framework. The `rocks-update` command connects to RHN and downloads the latest packages to `/var/spool/rpms`. However, this command does not automatically update the front-end and compute nodes. To do this, the administrator must run `% rocks-update --patch-compute` and `% rocks-update --patch-frontend`.

Downloading from RHN changes the build number of the installed version. For example, if the current version number is 4.0.0.0, the new version number advances to 4.0.0.1. In any distributed system such as Platform Rocks, the possibility always exists that some compute nodes will not update properly. The `rocks-update` command helps identify these compute nodes by generating a `rocks-update -list` option that displays the compute nodes and the current installed version. Sample output generated by entering `%rocks-update -list` is shown in Figure 1.

```
Current Repository Version: 4.0.0.2
Current Frontend Install Version: 4.0.0.1

Appliance Install Version

compute-0-0 4.0.0.1
compute-0-1 4.0.0.2
compute-0-2 4.0.0.1
compute-0-3 4.0.0.1
compute-0-4 4.0.0.1
compute-0-5 4.0.0.2-incomplete
compute-0-6 4.0.0.1
```

Figure 1. Sample output generated by `rocks-update -list` command

In the output shown in Figure 1, the current repository version represents the Rocks version available for installation on the front-end or compute nodes. The currently installed version on the front-end node is displayed as 4.0.0.1 and the versions for each compute node are listed. Some compute nodes have version 4.0.0.1 and some have 4.0.0.2. This means that not all compute nodes have been patched or reinstalled. One compute node bears an “incomplete” tag, meaning that a reimage of the compute node failed. The recommended recovery from this situation is to issue a `shoot-node` command, which reinstalls the compute node.

Rocks database pre-population tool

NPACI Rocks traditionally relies on Preboot Execution Environment (PXE) booting to add compute nodes to a cluster, but in NPACI Rocks 4.0.0, the `insert-ethers` command has additional options that can be used to update the database directly with a new node’s IP and Media Access Control (MAC) address. Platform Rocks extends this concept with a Platform Rocks tool, `add-hosts`, which compiles an XML configuration file for all nodes in the

cluster and pre-populates the Rocks database with the required information. This makes it easier for administrators to plan a Rocks deployment and determine in advance the IP addresses, subnets, node names, and rack locations for the new compute nodes. For example, if an administrator needs to install 256 nodes in two subnets, the XML configuration file `/opt/rocks/etc/add-hostsrc` shown in Figure 2 must be created.

Once the configuration file has been compiled and the administrator has run the `add-hosts` command, the new compute node values will enter the Platform Rocks database. This process does not install Rocks on the compute nodes; when each node is turned on, however, Rocks is designed to recognize the node by MAC address and install Rocks on the node automatically.

Optional Rolls

Optional Rolls extend the capability of Platform Rocks and provide administrators with the flexibility to choose desired modules to operate their cluster. The Platform Rocks Rolls should integrate with any Rocks implementation that does not modify the original

```
<?xml version="1.0" standalone="yes"?>
<add-hosts>
  <!-- the MAC addresses for hosts are contained in /opt/rocks/etc/mac-addr-list -->
  <mac_addr_file value="/opt/rocks/etc/mac-addr-list" />
  <!-- order_by_rack -->
  <order_by_rack value="no" />
  <!-- netmasks for all compute nodes will be 255.255.255.0 -->
  <netmask value="255.255.255.0" />
  <subnet>
    <!-- in the first subnet start naming the nodes with prefix node-, the base ip address is
    10.1.2.1, there are 128 nodes in this subnet and they are all compute appliances -->
    <host_prefix value="node-"/>
    <baseip value="10.1.2.1" />
    <num_hosts_in_subnet>128</num_hosts_in_subnet>
    <appliance>Compute</appliance>
  </subnet>
  <subnet>
    <!-- in the second subnet name the nodes with prefix 'node-', the base ip is 10.1.3.1, there are
    128 nodes in this subnet and they are all compute appliances.-->
    <host_prefix value="node-"/>
    <baseip value="10.1.3.1" />
    <num_hosts_in_subnet>128</num_hosts_in_subnet>
    <appliance>Compute</appliance>
  </subnet>
</add-hosts >
```

Figure 2. XML configuration file for populating the Platform Rocks database

NPACI source code. Rolls currently available from Platform Computing include:

- Clumon Roll
- Platform Lava Roll, which provides the freeware Platform Lava workload management tool
- Platform LSF HPC Roll, which provides the proprietary Platform LSF workload manager
- Intel Tools Roll, which provides compilers to optimize x86 and Intel Extended Memory 64 Technology (EM64T) environments¹²
- IBRIX Fusion File System Roll
- Topspin InfiniBand Roll

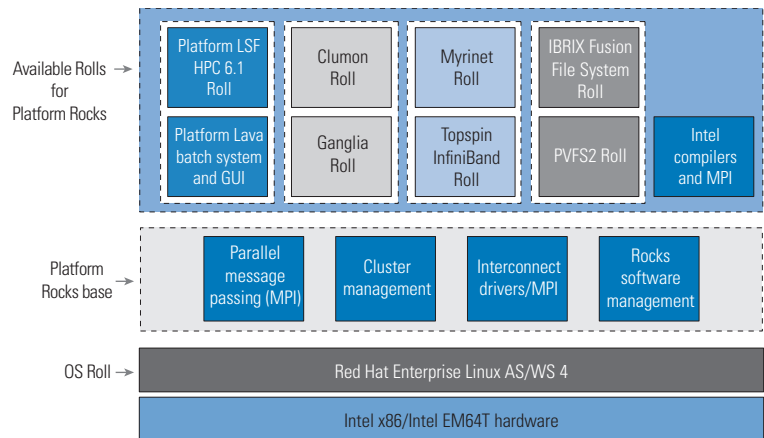



Figure 3 shows the Platform Rocks framework, including the different types of Rolls.

Figure 3. Platform Rocks framework and available Rolls

NCSA Clumon Roll

NPACI Rocks uses Ganglia to provide system information and metrics that help monitor the status and the load of compute nodes in the Rocks cluster. Ganglia permits easy viewing of summarized information for the entire cluster. The NCSA, located at the University of Illinois at Champaign-Urbana, developed the Clumon performance monitoring system for Linux-based clusters. Clumon displays cluster information for each node so concisely that the status and load of each node can be easily displayed on one screen, even for large clusters. This simple method for viewing a cluster provides an instant survey of the status of each node and lets administrators examine the circumstances when the state of a node visibly changes. Platform Computing is working with the developers of Clumon at NCSA to augment the functionality of Clumon and create a Clumon Roll, which is integrated with the Platform LSF and Platform Lava workload managers. Enhancements to Clumon are open source.

Comprehensive, efficient HPC cluster management

Dell and Platform Computing have collaborated to provide a turnkey HPC cluster solution that enables IT organizations to efficiently deploy and administer a cluster. Platform Rocks has been validated and is supported on 8- to 256-node Dell HPC cluster bundles for monolithic, rack-mountable servers and on 10- to 260-node Dell HPC cluster bundles for blade servers.¹³ The comprehensive cluster management capabilities provided by Platform Rocks can help make HPC cluster environments an attractive architecture for the enterprise data center. 

Rizwan Ali is a systems engineer with the Scalable Systems Group at Dell. His current research interests include performance benchmarking, cluster architecture, parallel applications, and high-speed interconnects. Rizwan has a B.S. in Electrical Engineering from the University of Minnesota.

Rinku Gupta is a systems engineer and advisor in the Scalable Systems Group at Dell. Her current research interests are middleware libraries, parallel processing, performance, and interconnect benchmarking. Rinku has a B.E. in Computer Engineering from Mumbai University in India and an M.S. in Computer Information Science from The Ohio State University.

Garima Kochhar is a systems engineer in the Scalable Systems Group at Dell. She has a B.S. in Computer Science and Physics from Birla Institute of Technology and Science (BITS) in Pilani, India, and an M.S. in Computer Science from The Ohio State University, where she worked on job scheduling.

Bill Bryce is the senior product manager for Platform Rocks at Platform Computing, where he has worked for the past 10 years. His interests include distributed computing, parallel programming, communications protocols, and operating systems. Bill has a B.S.C. in Computer Science from the University of Waterloo.

FOR MORE INFORMATION

Dell HPC clusters:

www.dell.com/hpcc

Platform Rocks:

www.platform.com/products/Rocks

¹² For more information about Intel software development products and cluster tools, visit www.intel.com/software/products/index.htm?iid=HPAGE+low_prod_software&.

¹³ For more information about these bundles, visit www1.us.dell.com/content/topics/global.aspx/solutions/en/clustering_hpcc?c=us&cs=555&l=en&s=biz&-tab=4.