**Deploying HPC Clusters using Platform Open Cluster Stack – Dell Edition**

As high-performance computing (HPC) environments continue to grow in size and complexity, cluster deployment and management have become increasingly challenging. Platform™ Open Cluster Stack – Dell Edition helps address these challenges with an open source, modular, hybrid architecture designed to easily integrate third-party software and help simplify deployment and management for cluster administrators.

By Sreeram Vedantham
Arun Rajan
Shivaraj Nidoni
Li Ou, Ph.D.
Gabor Samu
Mark Black

High-performance computing (HPC) environments often comprise a huge data center housing thousands of compute nodes designed to incorporate powerful multi-core processors. To help them deploy and manage these complex environments, administrators often use multiple open source tools for different aspects of administration. Integrating these tools seamlessly, however, can be time-consuming and difficult.

Platform Open Cluster Stack (OCS) – Dell Edition is designed to simplify this process. By providing a unified framework that supports a variety of tools and the seamless integration of third-party products—including both open source software and Dell-specific tools such as the Dell OpenManage™ suite—this comprehensive software stack can help administrators efficiently deploy and manage Linux® OS-based HPC clusters even in highly complex environments.

**Introducing the Kusu-Based Platform OCS – Dell Edition Stack**

Platform OCS – Dell Edition is an open source, modular, hybrid software stack that integrates other open source software as well as commercial software to form a unified cluster environment. This seamless integration of different components helps support the consistent deployment and management of scale-out Linux clusters. Integrated components include the Platform Lava job scheduler, the Platform Load Sharing Facility (LSF®) application (optional), Mellanox OpenFabrics Enterprise Distribution (OFED) (for Dell™ PowerEdge™ servers using Mellanox InfiniBand cards), QLogic OFED (for Dell PowerEdge servers using QLogic InfiniBand cards), Platform OFED, Message Passing Interface (MPI) (MVAPICH and Open MPI), Intel® runtime libraries, and so on. The latest version of the Platform OCS – Dell Edition software stack is based on Red Hat® Enterprise Linux 5 Update 1 and Platform OCS 5.1.

**Architecture**

Platform OCS – Dell Edition is designed as a comprehensive HPC software stack (see Figure 1). Kusu, the middle layer of the stack, comprises the tools that build the cluster and acts as a cluster toolkit foundation for the stack. The bottom layer, which Kusu makes use of, contains different protocols for a variety of functionalities. The upper layer includes the cluster middleware technologies such as job schedulers, MPI applications, external cluster health-monitoring tools, and end-user applications. These layers are closely tied to and integrated within the Kusu framework.
The Kusu framework is built on several basic elements:

- **Installer node**: HPC cluster setup primarily requires a master node known as an installer node. Installer nodes maintain the kusudb MySQL database for all functionalities.

- **Node group**: The node group is the template that defines the various properties of the compute nodes so that one set of compute nodes can be distinguished from the others based on the node grouping. Kusu provides different node groups such as installer, compute-rhel-5-x86_64, compute-imaged-rhel-5-x86_64, compute-diskless-rhel-5-x86_64, and unmanaged node groups. From these node groups, administrators can make copies and create customized node groups—for example, one group for InfiniBand and another for Gigabit Ethernet. This approach enables administrators to distinguish sets of nodes to take specific actions on, including selecting components to install; setting network configurations; changing the disk partitioning, OS, and kernel parameters; or adding custom scripts.

- **Kits**: To help ease the installation of different components, the Kusu framework packages different software components into kits. A kit may contain Red Hat Package Manager (RPM™) packages, scripts, applications, or services that must be installed on compute nodes, and includes meta-RPM components with configuration scripts and dependency information. Administrators can install kits from an ISO image or a mount point. Figure 2 shows an example of a kit framework.

Two advantages of using this mechanism are that the MySQL database stores information related to the different kits, and that it eliminates the need to configure applications and services manually. Installing the kit components automatically generates and updates the configurations the service or application needs, which can help save time.

A variety of kits are available by default with Platform OCS - Dell Edition; administrators can add these kits during installation of the installer node or after installation. Available kits include the following:
- **Base Kit**: Contains the basic components needed for the Kusu framework, as well as the tools for managing the kit and node groups and installing the nodes.
- **OS Kit**: Contains the Red Hat Enterprise Linux 5 Update 1 ISO image on DVD.
- **OFED Kit**: Includes the OFED stack and drivers for InfiniBand configuration, as well as the opensmbd subnet manager.
- **HPC Kit**: Contains math libraries such as Automatically Tuned Linear Algebra Software (ATLAS), Basic Linear Algebra Communication Subprograms (BLACS), and Scalable Linear Algebra Package (ScalAPACK), as well as MPI distributions such as MVAPICH 1 and 2, Open MPI, and MPICH 1 and 2 and precompiled benchmarks such as High-Performance Computing Kit (HPL) and iOzone.
- **Platform Lava Kit**: Provides scheduling for workload management on the compute nodes (based on an open source version of Platform LSF).

The Base Kit and OS Kit are mandatory in Platform OCS – Dell Edition clusters. For more information on management and monitoring kits, see the “Managing and monitoring clusters” section in this article.

During installation, Platform OCS – Dell Edition copies the OS media into a kit it uses to make a repository—a container that can store individual software components, RPM packages, custom scripts, and even a complete OS. To help resolve conflicts between kits and help provide failsafe upgrades, administrators can create custom repositories to distinguish one repository from another. The default repository in Kusu is called “Repo for rhel-5-x86_64”; the default repository path is /depot/repos/1000. Administrators can take snapshots of a repository without affecting the original repository.

During installation of the installer node, associations are created between the kits and node groups—meaning that many applications (provided by kits) require no manual configuration. If administrators create custom repositories, kits, or node groups, they should do so sequentially to help ensure that the components are updated correctly in the MySQL database. Figure 3 shows the correct sequence:

1. Add a new kit A using the kitops --add A command.
2. Specify a repository for the kit components using the repoman command. Documentation and kit component configuration are stored on the installer node.
3. Refresh the repository after adding the new kit.
4. If the kit components are not automatically assigned to node groups defined in the kit, use the ngedit tool to assign the components to node groups.
5. Nodes in the node group are then updated or reprovisioned with the new kit. Adding a new kit typically does not require reprovisioning.

This framework architecture enables administrators to build heterogeneous software bundles for various node groups, and to take advantage of its flexibility in defining the behavior and characteristics of individual components in the cluster.

**Installation and deployment**

Installing cluster nodes using the Kusu framework is a network-based process; only the primary installer node requires administrators to select from various options and start the installation. Administrators can install the other compute nodes using a Preboot Execution Environment (PXE) mechanism, without needing to perform manual intervention.

Platform OCS – Dell Edition uses the Red Hat Anaconda installer for the compute nodes. A kickstart file tells Anaconda the desired RPM packages and the root password needed on compute nodes. During installation, the installer node shows different options to define the partitioning schema, which is completed outside Anaconda. For compute nodes, the ngedit node group editor enables administrators to modify the partitions needed for compute nodes.

The automatic installation mechanism in Platform OCS – Dell Edition enables administrators to easily deploy the components, node groups, and other behavior predefined in the installer node to the requisite compute nodes. By integrating the ability to define functionality and the installation mechanism within one framework, Platform OCS – Dell Edition helps significantly simplify cluster deployment and management.
MANAGING AND MONITORING CLUSTERS

Managing a cluster effectively can be challenging—requirements vary depending on the complexity of cluster setup and operations performed throughout the cluster life cycle. The Dell OpenManage suite offers several options ranging from command-line interfaces to intuitive graphical user interfaces for both in-band and out-of-band systems management. In addition, the Kusu-based Platform OCS – Dell Edition stack makes a variety of utilities available to support complementary cluster software deployment, resource management, and monitoring of cluster-level performance, cluster job state, and resource utilization.¹

Cluster management tools

The Platform OCS – Dell Edition Base Kit contains a number of useful command-line tools and utilities to help simplify cluster management, including the following:

- **addhost**: A tool that allows administrators to add and remove nodes from a specified node group
- **boothost**: A tool to create the PXE configuration files used when starting up a node
- **cfm**: A configuration file manager that synchronizes node group packages and files
- **netedit**: A tool that defines available network interfaces in a cluster, which are then associated with node groups
- **ngedit**: The node group editor
- **nghosts**: A tool that moves hosts between node groups
- **repoman**: A tool that manages the repositories associated with different node groups in a cluster
- **repopatch**: A tool that updates repositories using the yum (Yellowdog Updater, Modified) tool and/or the Red Hat Network (RHN)
- **kitops**: A tool to manipulate kits in a cluster, including adding, removing, and listing kits
- **buildkit**: A tool that allows administrators and developers to build new kits by creating a template workspace for the kit
- **genconfig**: The database-reporting and file-generating tool
- **pdsh**: A high-performance, parallel remote-shell utility used to run commands across a cluster²

Platform OCS – Dell Edition is also bundled with Platform Lava, an open source job scheduler based on Platform LSF that provides a distributed batch system for submitting jobs and managing workloads. It helps manage the day-to-day workload of a whole cluster, providing simplified job execution, management, and accounting. This package is part of the optional Platform Lava Kit and is made up of two components: component-lava-master (which installs Platform Lava on the Platform OCS – Dell Edition primary installer node and sets up the Platform Lava master node with a default configuration) and component-lava-compute (which installs Platform Lava on Platform OCS – Dell Edition cluster compute nodes). Administrators can use Platform Lava to submit jobs within a cluster using the `bsub` command, view job status using the `bjobs` command, and view job output using the `bpeek` command.

Cluster monitoring tools

The Kusu-based Platform OCS – Dell Edition stack bundles several monitoring applications, including open source utilities:

- **Cacti**: A comprehensive network graphing solution that uses the data storage and graphing capabilities of RRDTool, Cacti provides a fast poller, advanced graphing templates, multiple data acquisition methods, and various user management features. Cacti can display almost any kind of data graphically and is part of the optional Cacti Kit, which includes two main components: component-cacti (installed on a node that becomes the Cacti master monitoring node that runs the Cacti Web service and allows administrators to view collected Cacti data) and component-cacti-monitored-node (which instructs Cacti to include the package-based compute node group in the collected data, and to monitor only specified node groups). The Cacti monitoring node performs all Cacti configurations.
- **Nagios**: A network management server that runs on the primary installer node and monitors cluster hosts, services, and networks, Nagios provides a Web interface to display the collected node information and to generate alerts if problems occur within configured thresholds. Nagios monitors the various services (such as disk usage, users, and ping) on both the primary installation and compute nodes in a Platform OCS – Dell Edition cluster. This package is part of the optional Nagios Kit, which includes two main components: component-nagios-installer (which monitors hosts and services and reports exceeded thresholds and problems) and component-nagios-compute (which installs a monitoring agent on compute nodes that collects information and sends it back to the network management service). Nagios provides a Web-based console accessible from compute nodes through component-nagios-installer.
- **Ganglia**: A scalable statistics collector that monitors node availability and displays system load, network usage, and other resource information for HPC clusters, Ganglia displays detailed information about Platform OCS – Dell Edition clusters and their day-to-day functions.


² For a complete list of commands and detailed information on usage and functionality, see the media kit available on the master node and the online documentation available at www.platform.com.
including uptime, available memory, number of processors, and processor usage. Administrators can drill down within the data to view information about each node in a cluster. This package is part of the optional Ganglia Kit, which includes two main components: component-ganglia-server (which acts as the server to which collected cluster information is returned) and component-ganglia-agent (which installs a small daemon that runs on compute nodes and returns collected data to the server). All Ganglia configurations are carried out on the server (by default, the primary installer node) through the Ganglia console.

- **Ntop:** A tool that monitors network bandwidth and analyzes traffic, ntop allows administrators to examine the network patterns of a Platform OCS – Dell Edition cluster and positions the network interfaces into a passive listening mode so administrators can watch traffic to and from each interface. The tool plots the data into a database, then displays this information in the Web console. This package is part of the optional Ntop Kit, which has only one component—component-ntop. Many ntop configurations, such as changing the default listening interface, adding users, restricting user access to pages, and stopping ntop, are made on the ntop Web console.

**USING THE DELL KIT AND DELL OPENMANAGE TOOLS**

The Platform OCS – Dell Edition media includes the Dell Kit (called dell-vendor), which provides drivers and custom scripts for Dell PowerEdge servers. This kit automates the installation and deployment of Dell drivers and utilities on installer nodes and compute nodes. It also contains scripts and RPM packages that administrators can use to install Dell drivers and Dynamic Kernel Module Support (DKMS) packages, deploy Dell OpenManage software, and configure the baseboard management controllers (BMCs) and BIOSs of cluster nodes to support out-of-band systems management and console redirection.3

**Understanding kit components**

The Dell Kit includes four major components: component-Dell-compute, component-Dell-installer, component-Dell-pxeboot, and component-Dell-bmcsetup. Component-Dell-compute and component-Dell-installer are used for installing and configuring Dell drivers and DKMS packages for Dell PowerEdge servers. These components include e1000, igb, bnx2, tg3, mptlinux, and megaraid_sas. During installation, the components check the PowerEdge server model and install appropriate drivers and DKMS packages that are supported and required for that model. Component-Dell-installer supports an automatic installation process for installer nodes. Component-Dell-compute is used to deploy Dell drivers on compute nodes.

Platform OCS – Dell Edition requires that compute nodes have PXE ahead of the hard drive in the BIOS boot sequence, which can be a problem for servers that have PXE after the hard drive in this sequence. To help avoid this problem, component-Dell-pxeboot contains scripts to configure the PXE boot order (for the next boot only) of the compute node using ipmitool. To reinstall compute nodes without manual intervention (pressing F12 for PXE), administrators must pre-enable these components.

Component-Dell-bmcsetup installs Dell OpenManage and configures the BMCs and BIOSs of the compute nodes during deployment. For systems with limited Dell OpenManage support, such as PowerEdge SC servers, the ipmitool utility can only configure the BMCs. For systems that include comprehensive Dell OpenManage support and are supported as part of the Dell HPC program, the script allows configuration of BMCs and BIOSs using the Dell OpenManage omconfig utility. After the system has been configured, it can employ features such as remote power cycling of servers and console redirection. The script must be enabled to execute on the compute nodes during deployment.

**Enabling kit components**

Administrators can add the Dell Kit by using the Kit window in the Kusu front-end installation wizard or by using the kitops command. After kit installation, the component-Dell-compute and component-Dell-installer components are automatically enabled and associated to the default node group (installer-rhel-5-x86_64) and default compute node group (installer-rhel-5-x86_64), respectively. If administrators deploy compute nodes directly without any further configuration, the Dell drivers and DKMS packages included in those components are automatically installed on the servers.

By default, component-Dell-pxeboot is disabled after front-end installation; administrators must perform additional steps to enable the component using the ngedit node group editor for a selected node group. After running ngedit and choosing a node group, administrators would then navigate to the Components window and select “dell-vendor,” then select “component-Dell-bmcsetup,” and then update the nodes in the group. Component-Dell-pxeboot is then activated on the compute nodes without reinstallation.

Administrators must enable component-Dell-bmcsetup manually, following the same steps as for component-Dell-pxeboot; component-Dell-bmcsetup contains the scripts administrators need to automate the installation and the configuration, but not the Dell OpenManage software itself, which allows Platform OCS – Dell Edition to be independent of Dell OpenManage version changes.
Component-Dell-bmcsetup provides a tool that allows administrators to customize the behavior of BMCs and BIOSs before deployment, including defining the BMC IP addresses, enabling and disabling console redirection, and so on. If administrators customize the BMCs and BIOSs, they must then reinstall the existing compute nodes.

Integrating the predefined vendor- and cluster-specific parameters and bundles with the compute node installation process can be challenging. Administrators can use the components of the Dell Kit to help them meet their systems management requirements on Dell clusters.

**MIGRATING FROM PLATFORM OCS 4 TO PLATFORM OCS – DELL EDITION**


Administrators can perform this migration by backing up and then restoring important information in three categories: system information, custom RPM packages, and XML configurations for compute nodes. They can use migration tools to help them back up and restore information during this process. First, they must copy custom RPM packages, scripts, and XML files manually from the Platform OCS 4.x front-end node to the new installer nodes. The optional packages and custom script windows of the ngedit tool in Platform OCS – Dell Edition allow administrators to specify custom RPM packages and scripts deployed on compute nodes. Administrators must also back up and restore all other important data on the existing front-end node.

After all information is ready on the new Platform OCS – Dell Edition installer node, administrators can start the compute node deployment process. To start the deployment, they can simply reboot (to PXE boot) all Platform OCS 4.x compute nodes.

**SIMPLIFYING HPC CLUSTER DEPLOYMENT AND MANAGEMENT**

The Kusu-based Platform OCS – Dell Edition software stack provides a comprehensive set of cluster features and tools that help simplify the deployment and maintenance of HPC clusters. Built on a hierarchical open source architecture, Platform OCS – Dell Edition provides a unified framework that supports the seamless integration of third-party products, including both open source utilities and proprietary software, as well as pre-bundled tools to help administrators monitor, manage, and administer clusters. By taking advantage of tools such as the Dell Kit and Dell OpenManage software, administrators can help simplify the deployment and management of Dell PowerEdge server-based HPC clusters.

Sreeram Vedantham is a development engineer adviser in the Dell High-Performance Computing Group. His current interests include high-speed interconnects, HPC cluster stacks, and schedulers.

Arun Rajan is a lead systems engineer and consultant in the Scalable Systems Group at Dell. His current interests include HPC cluster management, computing packages, and product development. He has a B.E. in Electronics and Communications Engineering from the National Institute of Technology, Tiruchirappalli, and an M.S. in Computer and Information Science from the Ohio State University.

Shivaraj Nidoni is a senior engineering analyst in the Dell High-Performance Computing Group. His current interests include high-end interconnects, parallel file systems, and HPC clustering packages. Shivaraj has an M.Tech. in Computer Network Engineering from Visvesvaraya Technological University.

Li Ou, Ph.D., is a systems engineer in the Scalable Systems Group at Dell. His current interests include parallel rendering, visualization, parallel file systems, high-performance storage systems, and HPC clustering. He has a B.S. in Electrical Engineering and an M.S. in Computer Science, both from the University of Electronics Science and Technology of China, and a Ph.D. in Computer Engineering from Tennessee Technological University.

Gabor Samu is a product manager at Platform Computing. His current interests include cluster management and multi-core programming. He has a B.S. in Computer Science from the University of Toronto.

Mark Black is development manager at Platform Computing. His current interests include cluster management, cluster provisioning, and virtualization. He has a B.E. from Ryerson University and an M.E.S. in Signal Processing from the University of Western Ontario.