Readme for Platform Open Cluster Stack (OCS) Rolls

Version 4.1.1-2.0

October 25 2006

Platform Computing

Contents

[Top]

Platform Roll

The Platform roll includes some useful tools that help make cluster management easier once you have set up your initial cluster.

- Add/remove roll tool
- Patch management tool
- Extend-compute CLI tools

Add-hosts tool

The add-hosts tool lets you quickly and easily add nodes to your cluster without using the `Insert-Ethers` command. By pre-populating the Platform OCS database from an XML configuration file, you can pre-load all the host information and simply startup your machines to start the compute node installation process.

The add-hosts tool is bundled with the Platform OCS software.

How it works

The add-hosts tool allows you to define an XML configuration file that describes the information for each host you want to add to your cluster. For example, when you list your hosts in the configuration file, your host attributes (like IP address, MAC address, and host name) are inserted directly into the Platform OCS database. By populating the database using the add-hosts tool, you no longer need to start your machines in a pre-determined order.

You need to do two things before you use the add-hosts tool:

1. Enter the MAC addresses of your nodes into a text file
2. Enter your host information into the XML configuration file

The XML configuration file allows you to define information on a host or a subnet (group of hosts) level.

The host section of the configuration file lets you list your hosts and host attributes individually.

The subnet section lets you collect your nodes into a virtual grouping with the same appliance type so that you do not have to list individual hosts. When you specify a beginning IP address and the number of nodes in the subnet section, the add-hosts tool can extrapolate a group of nodes, name them, and deduce their IP addresses. If your IP addresses are not incremental or if you have gaps in your IP addresses, you can also specify any IP addresses to be excluded.

The add-hosts tool processes the <host> and <subnet> sections in the order that they occur in the XML configuration file. The MAC addresses in the MAC address file must be listed in the same order that the <host> and <subnet> sections occur. Also, these MAC addresses must correspond to the physical order of your hosts. The XML configuration file is processed in tandem with the MAC address file. Each host is paired up with the next occurring MAC address in the MAC address file.

Note: The add-hosts tool only loads host information into the database. You may need to perform additional configuration to set up a network (for example, you may need to set up switches to route from one subnet to another).

- Test mode

Run this tool in test mode to verify the your results before actually populating the database and changing any configuration files. To run this tool in test mode, follow the steps below but use the `--testmode` option of the `add-hosts` command:

```
# add-hosts --testmode
```

Installing a new Platform OCS cluster

- Assumptions
  - You have acquired MAC addresses for all your hosts from your hardware vendor and listed them in a text file.
  - There are no empty slots or gaps between hosts in a rack.
  - You run this tool as root.
- Using the add-hosts tool
  1. List each MAC address to a separate line in `/opt/rocks/etc/mac.txt`.

     The order in this text file must be the same as the physical order of your nodes. Comments (added to the beginning or end of a line, or on their own lines, marked by a "#") can be useful in tracking groups of MAC addresses.

  2. Fill in values in the XML configuration file `/opt/rocks/etc/add-hostsrc`.

     The order of hosts in this file must match the order of MAC addresses in the `mac.txt` file.

     The skeleton XML configuration file:

```
<?xml version="1.0" standalone="yes"?>

<add-hosts>

    <mac_addr_file value = file path />
    <num_hosts_per_rack value = number />
    <order_by_rack value = "yes" |"no" />
    <netmask value = address />
    <host_num_length value = number />
    <host_start_num value = "0" | "1" />

    <host>
        <name value = string />
        <ip value = string />
        <appliance value = string />
    </host>

    <subnet>
        <host_prefix value = string />
```

```
        <host_suffix value = string />
        <baseip value = string />
        <num_hosts_in_subnet value = number />
        <appliance value = string />
        <ip_exclude_list value = list of IP addresses />
        <restart_numbering value = "yes" | "no" />
        <gen_descending_ip value = "yes" | "no" />
    </subnet>

</add-hosts>
```

You can have zero or more <host> sections and zero or more <subnet> sections. Note that <host> sections must be defined before <subnet> sections.

Parameter	Description
Global Variables	
mac_addr_file**:**	The absolute path name to the location of the MAC address text file. By default, located at /opt/rocks/etc/mac.txt.
num_hosts_per_rack:	The number of hosts per rack.
order_by_rack**:**	"Yes" (default) specifies that you want to order by rack and rank (node-0-0, node-0-1...node-2-4, node-2-5) and "No" indicates that you want to order by rank only (node-1, node-2, node-3).
netmask:	The netmask for this network.
host_num_length (optional):	Specifies the number of digits in your host names. Applies only if order_by_rack value="no". This option is useful for restricting your host names to a fixed length. For example, if you have 20 hosts with each host using the same prefix: ■ If you do not set host_num_length (or set host_num_length=0) and you have 20 hosts in your cluster with each host using the same prefix, your host names will have different lengths, for example, host-0, host-1, ..., host-10, host-11, ... ■ If you set host_num_length=2, you can restrict the numeric portion of your host names to a fixed length of 2, for example, host-00, host-01, ..., host-10, host-11, ...
host_start_num (optional):	Specifies whether nodes are numbered beginning with 0 (default) or 1.
Host Section Parameters: All host sections must precede all subnet sections.	
host:	Section used to define individual hosts.
name:	The name of the host.

ip:	The IP address of the host.
appliance:	Valid appliance type names with a default Platform OCS installation are: ▪ Compute ▪ Ethernet Switches ▪ Power Units ▪ Remote Management ▪ NAS Appliance This list can change depending on the rolls you install with Platform OCS. ***Please see PVFS2 meta server if you have a PVFS2 meta server appliance type.*** To see your valid appliance types: 1. Run # ***insert-ethers*** A list of appliances is displayed. 2. Select any appliance type 3. Press **F9** to exit
Subnet section parameters	
subnet:	Section used to define a group of hosts in one <subnet>.
host_prefix:	Host name not including the number that will be assigned. For example, if host_prefix value ="alpha", then the first host is named alpha0-0. Note that a dash is automatically added between the prefix and the assigned number.
host_suffix (optional):	Host name suffix. For example, if host_suffix value ="A", hosts are named charlie0-0A.
base_ip:	The first IP address in this subnet.
num_hosts_in_subnet:	Number of hosts in the subnet.
appliance:	Valid names are the same as for the Host Section Parameter. Each subnet section can only contain one appliance type. Set up different subnets for different appliances types.
ip_exclude_list (optional):	IP addresses to be excluded from the subnet. List IP addresses separated by a space. Use this option if you have gaps in your IP addresses.
restart_numbering:	"Yes" restarts host numbering for the current subnet. "No" (default) continues numbering from the last subnet.
gen_descending_ip:	Specifying "yes" will generate the IP numbers in descending order, and specifying "no" will generate the IP numbers in ascending order. The default value is "no"

1. Save your configuration file `/opt/rocks/etc/add-hostsrc`.
4. From the command line, run **`add-hosts`** or **`add-hosts --testmode`** to test your outcome before making any configuration file changes.
5. Check `./add-hosts.log` to see if you were successful. Each host added successfully has its own line and indicates "SUCCESS" at the end. If you failed to add a host, the line indicates "FAILED".
6. Perform a PXE startup on your compute nodes. For each host that was added to the Platform OCS database by the

add-hosts tool, the compute node installation begins automatically.

- PVFS2 meta server

If you have a PVFS2 meta server appliance type, you must name your hosts as follows: **`pvfs2-meta-server-0-0`**. The PVFS2 roll only recognizes PVFS2 meta server hosts with this host name.

Adding a new host to your existing cluster

- Assumptions
  - The host is added to the end of a rack.
  - You have run the add-hosts tool at least once before.
- Using the add-hosts tool
  1. Add the new MAC address to the list in `/opt/rocks/etc/mac.txt`. Make sure the order in this text file is the same as the physical order of your nodes.
  2. In `/opt/rocks/etc/add-hostsrc`, add a host section or edit a subnet section to include the new host. For example, when adding a host to a subnet, increase the value of `num_hosts_in_subnet` by 1.
  3. Save the configuration file.
  4. From the command line, run **`add-hosts`** or **`add-hosts --testmode`** to test your outcome before making any configuration file changes.
  5. An error will occur indicating trouble adding the first host. This error is expected since you have already added this host to your database.
  6. When prompted, select **`all`** to skip all subsequent errors.
  7. Check `./add-hosts.log` to see if you were successful. Each host added successfully has its own line and indicates "SUCCESS" at the end. If you failed to add a host, the line indicates "FAILED".

Replacing a host

Once you have run the add-hosts tool, you can easily replace one host with another in the same physical location with the same IP address:

1. Replace the old MAC address with the new one in `/opt/rocks/etc/mac.txt`.
2. Remove the old MAC address in the Platform OCS database. For example:

```
# dbreport ethers | grep 00:c0:9f:45:02:16
00:c0:9f:45:02:16 compute-0-3.local
# insert-ethers --remove "compute-0-3"
```

3. From the command line, run **`add-hosts`** or **`add-hosts --testmode`** to test your outcome before making any configuration file changes.

   An error will occur indicating trouble adding the first host. This error is expected since you have already added this host to your

database.

4. When prompted, select **`all`** to skip all subsequent errors.
5. Check `./add-hosts.log` to see if you were successful. Each host added successfully has its own line and indicates "SUCCESS" at the end. If you failed to add a host, the line indicates "FAILED".

Upgrading your Platform OCS installation

If you are upgrading your Platform OCS installation, you can transfer your existing host configuration to the new installation.

1. Before upgrading the front end, copy `/opt/rocks/etc/add-hostsrc` and `/opt/rocks/etc/mac.txt` onto a disk or shared directory on another file server.
2. After you have upgraded your front end, copy `/opt/rocks/etc/add-hostsrc` and `/opt/rocks/etc/mac.txt` to `/opt/rocks/etc`.
3. Run the add-hosts tool.
4. Check `./add-hosts.log` to see if you were successful. Each host added successfully has its own line and indicates "SUCCESS" at the end. If you failed to add a host, the line indicates "FAILED".

Add/remove roll tool

This command-line tool allows you to dynamically add or remove a roll on the front end. You can also use this tool to specify if you want a roll to install only on the front end and not on all compute nodes. Once you have added or removed a roll on the front end, you must reinstall the compute nodes.

Requirements

- You must run this tool as `root`.

Limitations

- This tool only adds or removes rolls supported by Platform Computing.
- This tool does not prevent you from removing rolls that should not be removed. Check the list below to see which rolls can be safely removed.
- The tool now supports upgrading certain supported rolls.
- Installing 32-bit rolls on a 64-bit front end is not supported at this time.

Platform Lava and Platform LSF HPC cannot run on the same node. The Platform LSF HPC roll will disable Platform Lava

---

Different versions of the Intel compilers roll cannot be used together. They will overwrite each other.

Rolls you can add and remove

The following is a list of rolls that you can add or remove after you have installed on the front end.

WARNING: If you try to add or remove a roll that is not supported, you could seriously damage or impair your system. Check the list of rolls that can be safely added and removed before using this tool.

Rolls that can be added and removed:

- Intel® Software Tools
- Cisco® Topspin®
- CluMon
- Extra Tools
- intel_mpirt
- Lava
- LSF HPC
- mattool
- ntop
- PVFS2

Rolls that can be added but **not** removed:

Do not attempt to remove the following rolls. They will not uninstall correctly.

- Ganglia
- Intel®
- Myrinet®
- SGE6

To see a list of installed rolls

In the command line, run the following command:

```
# rollops -l
```

To add a new roll

You can add a new roll by either using the DVD or having a copy of the ISO file for the roll you want.

1. Make sure you have not already installed the roll. To see a list of installed rolls, run the following command:

   ```
   # rollops -l
   ```

2. Add the roll.
   - If you have the DVD with the roll you want to add, insert the DVD and run the following command:

     ```
     # rollops -a
     ```

   - If you have a meta roll, type **`rollops -a`** `roll_name`. For example:

     ```
     # rollops -a lava
     ```

   - If you have an ISO file, type **`rollops -a -i`** `iso_file_name`. For example:

     ```
     # rollops -a -i lava-4.0.0-i386-disk1.iso
     ```

   If you do not specify a roll name and the DVD or ISO contains a meta roll, you will be prompted to select from a list of available rolls.

   If you are adding the CluMon roll, you are prompted for your `root` password.

3. The tool gives you a message indicating your success or failure to add the new roll. For example:

   ```
   The lava roll was added successfully.
   ```

4. If you have added any of the following rolls, you must restart your front end:
   - CluMon
   - Cisco® Topspin®
   - Ganglia
   - Lava
   - LSF HPC
   - Myrinet®
   - PVFS2
   - SGE6
5. Reinstall your compute nodes.

To remove a roll

To remove a roll from the front end, type **`rollops -e`** <*roll_name*>. For example:

```
# rollops -e lava
```

If you are removing the CluMon roll, you are prompted for your root password.

The tool gives you a message indicating your success or failure to remove the roll. For example:

```
The lava roll has been sucessfully removed.
```

If you have removed this roll on your front end node, you must reinstall your compute nodes for the removal to take effect.

To upgrade a roll

To upgrade a roll in the front end, type **`rollops -u`** <*roll_name*>. For example:

```
# rollops -u dell
```

If you have the media with the roll you want to upgrade, insert the media and run the following command without specifying the roll name:

```
# rollops -u
```

Note: If you have a meta roll, you will be prompted with a list of the rolls found in the meta roll, if available. Specify the name of the roll you want to upgrade, as listed in the meta roll.

If you have an ISO image, type **`rollops -u`** <*optional_roll_name*> **`-i`** <*iso_file_name*>. For example:

```
# rollops -u -i dell-4.1.1-0.x86_64.disk1.iso
```

Note: If you specified a meta roll ISO image and you do not specify a roll name, you will be prompted with a list of the rolls found in the meta roll, if available.

Only the Dell roll supports upgrading at this time.

The tool gives you a message indicating your success or failure to add the new roll. For example:

```
The dell roll was added successfully.
```

If you upgraded the Dell roll, you must restart your front end node.

If you upgraded the Dell roll on your front end node, you must reinstall your compute nodes for them to have the latest version installed.

To prevent a roll from installing on compute nodes

You can choose to install a roll on the front end only and not on the compute nodes.

Type **rollops -r** <*roll_name*> **-p no**. For example:

```
# rollops -r lava -p no
```

When you install on your compute nodes, the roll for which you specified **-p no** is not installed.

Patch management tool

The patch management tool lets you update your cluster with new packages from your operating system's update network. With this tool, you no longer have to reinstall your whole cluster to get the latest patches and enhancements for your OS.

You may use the patch management tool regardless of whether you have a central install server or not.

- With the add-hosts tool

  If you used the add-hosts tool to pre-populate the Platform OCS database with your host information, you can transfer your existing host configuration to the new installation. See Upgrading your Platform OCS installation **before** you use the patch management tool to ensure you save your host configuration.

How it works

The patch management tool checks for updates and enhancements to your operating system available from your operating system's update network, directs the download of the updates, and tracks the version that you are working with. Any appliance is patched as long as the packages on the appliance exist on both the front end and appliance type itself.

Requirements

- For upgrades from Red Hat Network with Platform OCS Enterprise Edition, you must have a subscription to Red Hat Network and registration with `up2date`.

Limitations

- Patching for cross-kickstarted front ends and compute nodes is not supported.

- You can patch any appliance type, but be aware that RPM packages installed on the front end may have dependencies on appliance packages of a specific version. If you break this dependency, your front end may break your appliance.
- If you completely upgrade your front end, any updates made with the Patch Management tool will be overwritten. The version number that tracks your upgrade path will return to the original: 4.1.1-2.0.
- You can not select from the packages to download because some items will have dependencies. Do not deselect any packages when you register with Red Hat.
- Rollbacks are not supported.
- You cannot selectively choose packages to download from a front end installed from a central server.
- The Patch Management tool will not allow you to update kernel-dependent packages.

Downloading an update

Run the following command:

```
# rocks-update -d <packagename>
```

For Platform OCS Enterprise Edition, if you have not registered with Red Hat Network, you will be prompted to register. Enter your Red Hat Network account information and follow the prompts. Do not de-select any of the packages listed for download. When the registration is complete, `rocks-update` will download the update required.

`rocks-update` downloads the update to your front end. If there are new packages available, the patch management tool returns the following message indicating the new repository version and that you can proceed to update your appliances.

```
rocks-update: repository for updates is now version
4.1.1-2.0. You can now install/update your compute node or
front end appliances.
```

The repository version database is also updated. It does not install the packages on the front end.

Downloading or patching updates from the central server front end

Before downloading or patching updates from the central server front end, change the following settings:

1. On the central install server, enable the Apache user access to the `app_globals` table for your new front end using the following commands:

   ```
   # mysql -u root -p cluster
   mysql> GRANT SELECT on cluster.app_globals to
   apache@<new_front_end_external_ip>
   ```

   where <new_front_end_external_ip> is the external IP address of the new front end.

2. On the new front end, edit the `/etc/front end.repo` file with `vi` and replace the `baseurl` line with the following line:

   baseurl=http://*<central_install_server_ip>*/updates/

   where `<central_install_server_ip>` is the IP address of the central server.

Follow these steps to download updates from a central server front end:

1. Run the following command:

   ```
   # rocks-update -g
   ```

   The tool displays the download details and the following message:

   ```
   rocks-update: repository for updates is now version
   4.0.0.1. You can now install/update your compute node or
   front end appliances.
   ```

2. If you want to reinstall your compute nodes, you may do so now they will have the new updated applied to them.

Follow these steps to patch updates from a central server front end:

1. Run the following command:

   ```
   # rocks-update -f
   ```

   The tool displays the installed packages and the following message:

   ```
   rocks-update: 4 Update(s) installed successfully on front
   end!
   ```

2. If you want to reinstall your compute nodes, you may do so now they will have the new updated applied to them.

Patching a compute node

From `root`, run `rocks-update -c`.

The tool checks for new updates and returns a message indicating whether there was a patch available or not. The default is to perform updates on 64 nodes at a time. You can change this default by specifying the number of nodes to update concurrently, up to a maximum of 250 nodes.

For example, `rocks-update -c 128`. 128 compute nodes will be patched concurrently.

Installing or reinstalling a compute node

Run `insert-ethers --replace=<host_designation>`

OR

Use the `add-hosts` tool to add a host. See Adding a new host to your existing cluster and Replacing a host.

Listing versions of installed updates

From your front end, run `rocks-update -g`.

Extend-compute CLI tools

Two new CLI tools have been added to ease the process of customizing your compute nodes. These tools are as follows:

- `rocks-compute`: this tool allows you to customize the RPM packages and lets you add a script to customize the post-installation for compute node installations.
- `custom-partition`: this tool allows you to customize the `root` and `swap` partition sizes on your compute nodes.

Requirements

- You must run this tool as `root`.
- You must run this tool on the front end node.

Using the rocks-compute tool

This tool customizes compute nodes by modifying the `extend-compute.xml` file. This XML file is located in `/export/home/install/site-profiles/4.1.1/nodes`. The tool allows you to add, list, or remove RPM packages and post-install scripts.

The extend-compute tool does not handle package dependencies. If you wish to install a new package to your compute nodes you need to keep in mind of any dependencies the package may need. You will need to add all of them using the extend-compute tool.

You may occasionally need to update `extend-compute.xml` manually to add, remove, or update the contents. User-added changes are preserved as long as they are not affected by operations carried out by the `rocks-compute` tool.

- Adding a custom package

  You can add your own RPM packages to a compute node installation by running the following command:

```
# rocks-compute -a -p <path_to_rpm_package>
```

where <*path_to_rpm_package*> can be an absolute or relative path to the RPM package.

This will add the package to the Platform OCS distribution for your native architecture. If you want to add packages to a distribution for a non-native architecture, run the following command:

```
# rocks-compute -a -p <path_to_rpm_package>
--arch <architecture_type>
```

where <*architecture_type*> can be **i386**, **x86_64**, or **ia64**.

To automatically rebuild your Platform OCS distribution (`rocks-dist` tree) after adding your package, run the commands above with the `-b` option.

- Removing a custom package

  To delete a custom package that you added to your compute nodes, you need to know its ID and the architecture of the Platform OCS distribution from where you want to remove it. Obtain the RPM ID by running the following command:

  ```
  # rocks-compute -l p
  ```

  Remove the package from a native architecture by running the following command:

  ```
  # rocks-compute -d -p <rpm_id>
  ```

  Remove the package from a non-native architecture by running the following command:

  ```
  # rocks-compute -d -p <rpm_id> --arch <architecture_type>
  ```

  To automatically rebuild your Platform OCS distribution (`rocks-dist` tree) after removing your package, run the commands above with the `-b` option.

- Listing custom packages

  To display a list of all the custom packages you added to your compute nodes, run the following command:

  ```
  # rocks-compute -l p
  ```

  This command gives the name and ID of each custom package.

- Adding a custom post-install script

  In a typical Platform OCS compute node installation, a post-install script executes after all RPM packages are installed. A post-install script allows you to perform custom actions on the system, such as service configuration or executing installation scripts. The specified script must be a valid `bash` shell script.

  To add a script to the Platform OCS distribution for your native architecture, run the following command:

  ```
  # rocks-compute -a -s <path_to_script>
  ```

  where <*path_to_script*> can be an absolute or relative path to the post-install script.

  To add a script to a Platform OCS distribution for a non-native architecture, run the following command:

  ```
  # rocks-compute -a -s <path_to_script>
  --arch <architecture>
  ```

  To add a script to a Platform OCS distribution for an unspecified architecture, run the following command:

  ```
  # rocks-compute -a -s <path_to_script> --no-arch-script
  ```

  To automatically rebuild your Platform OCS distribution (`rocks-dist` tree) after adding your script, run the commands above with the `-b` option.

- Removing a custom post-install script

  To delete a script, you need to know its ID and the architecture of the Platform OCS distribution from where you want to remove it. Obtain the script ID by running the following command:

  ```
  # rocks-compute -l s
  ```

  If you are removing the script for your native architecture, run the following command:

  ```
  # rocks-compute -d -s <script_id>
  ```

  If you are removing the script for a non-native architecture, run the following command:

  ```
  # rocks-compute -d -s <script_id>
  --arch <architecture_type>
  ```

  If you are removing the script that was added to a Platform OCS distribution for an unspecified architecture, run the following

command:

```
# rocks-compute -d -s <script_id> --no-arch-script
```

As with the add operation, you can rebuild your Platform OCS distribution (rocks-dist tree) by running the deletion commands above with the "-b" option.

To automatically rebuild your Platform OCS distribution (`rocks-dist` tree) after removing a script, run the commands above with the `-b` option.

- Listing the contents of a custom post-install script

  To list the contents of a custom post-install script that you added, run the following command:

  ```
  # rocks-compute -l s
  ```

  You will be prompted to select one of the scripts. Make a selection by specifying the number next to the post-script ID.

Using the custom-partition tool

This tool customizes the partition sizes for compute nodes by modifying the `extend-auto-partition.xml` file. This file is located in `/export/home/install/site-profiles/4.0.0/nodes`.

You may occasionally need to update `extend-auto-partition.xml` manually to add, remove, or update the contents. User-added changes are preserved as long as they are not affected by operations carried out by the `custom-partition` tool.

- Changing root/swap partition sizes

  The default `root` partition size of a Platform OCS cluster is 10 GB (10 000 MB). You can change this size by running the following command:

  ```
  # custom-partition -r <root_partition_size>
  ```

  where <*root_partition_size*> is specified in MB and must be at least 6 000 MB.

  The default `swap` partition size of a Platform OCS cluster is 1 GB (1 000 MB). You can change this size by running the following command:

  ```
  # custom-partition -s <swap_partition_size>
  ```

  where <*swap_partition_size*> is specified in MB and must be at least 1 000 MB.

Note that you can use the `-r` and `-s` options together in the same command.

- Listing current root/swap partition sizes

  To list the current `root` and `swap` partition sizes, run the following command:

  ```
  # custom-partition -l
  ```

- Restoring default root/swap partition sizes

  To restore the root and swap partition sizes to the Platform OCS default values, run the following command:

  ```
  # custom-partition -d
  ```

[Top]

Platform OCS LSF HPC

Platform OCS LSF HPC® is an optional roll for managing and accelerating High Performance Computing (HPC) mission-critical workload.

With Platform LSF HPC you can intelligently schedule parallel and serial workload providing the capability of solving large, challenging problems while utilizing the available computing resources at maximum capacity.

For more information about Platform LSF HPC, see the Platform Web site: http://www.platform.com/products/HPC.

Product support

For Platform OCS hardware, operating system support, and CD distributions, see *Readme for Platform OCS*.

Platform LSF HPC will disable Platform Lava. Do not install Platform LSF HPC unless you intend to use it.

If you have already installed Platform LSF HPC, but want to use Platform Lava, you must use the `rollops` tool to remove Platform LSF

HPC, then re-enable Platform Lava by running `chkconfig -add lava` on all affected nodes. On the front end node, you must also run `chkconfig --add lavagui`.

Configuring and managing Platform OCS LSF HPC

Once you have installed Platform OCS and the LSF HPC roll, get a Platform LSF HPC license and start license manager daemons. Then set up and start your Platform LSF HPC cluster.

Setting up a Platform LSF HPC license

1. Decide which machine will be a license server machine.

   The following steps use `front end-0` as the license server.

2. Get FLEXlm `hostid`

   Use the `lmhostid` command on the FLEXlm server host to get the

   hardware identifier of your FLEXlm license server host. For example:

   ```
   # lmhostid
      lmhostid - Copyright (c) 1989-2003 by Macrovision Corporation. All rights
   reserved.
      The FLEXlm host ID of this machine is "0006296d1f2c"
   ```

   In this example, send the code "0006296d1f2c" to Platform.

3. Contact license@platform.com to get a permanent Platform LSF HPC license.

   Send the following information to Platform at license@platform.com:

   - Host name of the license server host
   - Host identifier of the license server host (`lmhostid` output)
   - Products required
   - Number of licenses required for your cluster

4. When you receive your license file, save it as `/opt/lsfhpc/conf/license.dat`.

   The following is an example of a permanent license:

   ```
   SERVER front end-0 0006296f1f2c 1700
   DAEMON lsf_ld /opt/lsfhpc/6.1/linux2.4-glibc2.3-x86/etc/lsf_ld
   ```

```
FEATURE lsf_base lsf_ld 6.000 1-sep-0000 10 CCF7C3C92A5471A12345 "Platform"
FEATURE lsf_manager lsf_ld 6.000 1-sep-0000 10 4CF7C37944B023A12345 "Platform"
FEATURE lsf_sched_fairshare lsf_ld 6.000 1-sep-0000 10 8CA763A93AC825C12345
   "Platform"
FEATURE lsf_sched_parallel lsf_ld 6.000 1-sep-0000 10 3C77F30945F7FBC12345
   "Platform"
FEATURE lsf_sched_preemption lsf_ld 6.000 1-sep-0000 10 3C0733892C1683812345
   "Platform"
FEATURE lsf_sched_resource_reservation lsf_ld 6.000 1-sep-0000 10
   ECD7C369072CA3812345 "Platform"
FEATURE platform_hpc lsf_ld 6.000 1-sep-0000 10 CA6CBE08B635EAC765EC "Platform"
```

5. Copy the license.dat file to `/opt/lsfhpc/conf/license.dat`.

   `LSF_LICENSE_FILE` in `lsf.conf` is set automatically during installation to `LSF_LICENSE_FILE=/opt/lsfhpc/conf/license.dat`.

6. Start the license daemons (`lmgrd`):
   a. Log on to the license server host.
   b. Use the `lmgrd` command to start the license server daemon. For example:

      ```
      % lmgrd -c /opt/lsfhpc/conf/license.dat -l /tmp/license.log
      ```

      DO NOT run `lmgrd` as root.

   LSF installation puts the `lmgrd` command in LSF_SERVERDIR. For example: `/opt/lsfhpc/6.1/linux2.4-glibc2.3-x86/etc/lmgrd`.

   You should include LSF_SERVERDIR in your PATH environment variable. You should also include the full `lmgrd` command line in your system startup files on the license server host, so that `lmgrd` starts automatically during system restart.

7. Check the license daemons (`lmstat`):

```
License server status: 1700@front end-0
License file(s) on front end-0: /opt/lsfhpc/conf/license.dat:
front end-0: license server UP (MASTER) v7.0
Vendor daemon status (on front end-0):
lsf_ld: UP v7.0
Feature usage info:
Users of lsf_base:  (Total of 4 licenses available)
Users of lsf_manager:  (Total of 4 licenses available)
```

```
Users of Platform_HPC:  (Total of 4 licenses available)
```

See *Licensing Platform LSF* for detailed information about configuring Platform LSF licenses.

Configure compute nodes in your Platform LSF HPC cluster:

Compute nodes are automatically added to the cluster either when the `lsf` service is restarted or if `insert-ethers` exists. New compute nodes are added to the `lsf.cluster.lsfhpc` file. Computer nodes are assigned to the `lammpi` and `mpichp4` resources by default. You can change the default by changing the value of `DefaultLSFHostResource` in the Platform OCS database, as follows:

```
# mysql -u apache cluster
mysql> insert into app_globals (service,component,value)
values (`Info','DefaultLSFHostResource','<resource_list>') ;
```

where *<resource_list>* is a list of all the Resources for a node. LSF HPC includes support for the following MPI implementations:

- Infiniband: `mvapich`
- Myrinet: `mpich_gm`
- LAM: `lammpi`
- Mpich p4: `mpichp4`
- Scali MPI: `sca_mpimon`
- Mpich MX MPI: `mpich_mx`
- MPICH Shared Memory: `mpichsharemem`

See the `lsf.shared` file for a full list of all supported MPI implementations.

The following is a sample host entry:

```
Begin Host
HOSTNAME   model  type   server r1m  mem  swp  RESOURCES
...
compute-0-1  !      !        1   3.5  ()  ()    (lammpi mpichp4)
...
End Host
```

Setting up and starting your Platform LSF HPC cluster.

1. Log on to the front end node as `root`.
2. Install the Platform LSF HPC license as described above.
3. Restart the LSF HPC services:

```
# service lsf restart
```

The compute nodes start LSF HPC at boot time.

After starting your cluster, run a few basic LSF commands (`lsid`, `lshosts`, `bhosts`). For example:

```
% lsid
   Platform LSF HPC 6.1 for Linux, Sep 1 2005
   Copyright 1992-2005 Platform Computing Corporation
   My cluster name is lsfhpc
   My master name is front end-0.public
```

Configuring your Platform LSF HPC cluster for master failover

The `LSF HPC Master` appliance type is a node type for LSF HPC master candidates. These nodes are used to offload the LSF Master to a less busy node. You should use the `LSF HPC Master` appliance type in larger clusters to install one or more nodes as LSF master candidates.

Configure your cluster for master failover using the following procedure:

1. Run `insert-ethers` and select the `LSF HPC Master` appliance type.
2. Install one or more of the `LSF HPC Master` nodes.
3. Exit `insert-ethers`.

   This is needed to update the `lsf.cluster.lsfhpc` file.

4. Verify that the NFS filesystem can be mounted on the `lsfhpc-0-0` node by running the following commands:
   a. `# ssh lsfhpc-0-0`
   b. `# mount NFS_server_name:export_name /mnt`
   c. `# unmount /mnt`
5. Run the `config-lsf-master` script in `/home/install/upgrades/lsfhpc`
6. Answer the dialog questions when prompted to by the script.

After completing the questions, the LSF HPC cluster should be using `lsfhpc-0-0` node as the primary LSF master, and the front end node should be the last node to which it will fail over.

You should run the `config-lsf-master` script each time you add or remove an `LSF HPC Master` appliance node.

Removing a compute node

To remove a compute node, you should shut down the entire compute node, or at least shut down the Platform LSF HPC daemons, then remove the compute node from LSF and the Platform OCS cluster.

1. `# ssh compute_node_name`

2. # **shutdown**

   OR

   # **/etc/init.d/lsf stop**

3. On the front end node, run remove the related host line from the `lsf.cluster.lsfhpc` file.
4. # **lsadmin reconfig**
5. # **badmin reconfig**

Removing a host from the Platform OCS cluster:

insert-ethers --remove="*<compute_node_name>***"**

Troubleshooting

Solution:

1. **Symptom**: Unable to launch parallel jobs using LAM MPI when the number of processes increases (approximately 32).

   Explanation: lamboot is failing to ssh to the nodes in a timely manner. The issue can be further traced to name resolution in ssh. The cluster may be trying to use a non-existent domain or DNS server.

   To test if the issue exists:

   a. Connect to a compute node and try to ssh to another compute node.

      Take note of the time required.

   b. Log out of the other compute node and try to ssh to the IP address of another compute node.
   c. If the result is dramatically faster, the issue exists.
   d. To verify the problem is with name resolution, edit the `/etc/resolv.conf` file on a compute node. Change the search line to only include the private domain.
   e. ssh to another compute node. It should respond faster.

   Solution: Fix this problem using any one of the following:

   - Update the front end's `/etc/resolv.conf` file to use a real DNS server.
   - Update the database and set the `PublicDNSDomain` in the `app_globals` table to be blank, and then reinstall the compute nodes.
   - As a temporary fix, change the search parameter in the `/etc/resolv.conf` of all the compute nodes.
2. **Symptom**: A suspended lammpi job cannot be terminated by the owner with `bkill`.

Solution: Resume the job to be killed; the job will go to `EXIT`.

3. **Symptom**: If an `MPI` job is launched on a compute node to run across other compute nodes, and other compute nodes are not accessible from the launching node, the job execution goes to failure, but `bjobs` shows the job is `DONE`.

   Solution: Make sure all compute nodes are accessible to each other.

4. **Symptom**: An `mpich_gm` job can be dispatched to a host without Myrinet card installed.

   Solution: Configure a host group `gm_hosts` for those hosts with Myrinet card properly installed and specify the option `-m gm_hosts` for `bsub` when submitting the job.

5. **Symptom**: When a job failed to launch or a job is terminated during execution, the related temp files are left over under `HOME`.

   Solution: Remove the files manually.

6. **Symptom**: **Ctrl-C** cannot terminate a started interactive `lammpi` job.

   Solution: Use `bkill` instead.

7. **Symptom**: `lammpi` application fails if run across the nodes mixed with and without Myrinet card installed.

   Solution: Set the environment variable `LAM_MPI_SSI_rpi=tcp` before submitting the job.

[Top]

Platform Lava

Platform Lava is a distributed batch system for submitting jobs and managing the workload on a Platform OCS™ cluster. Platform Lava is free and is based on Platform LSF.

Platform Lava lets you easily manage the day-to-day worklfront endad of a whole cluster, providing simplified job execution, management, and accounting.

Platform LSF HPC will disable Platform Lava. Do not install Platform LSF HPC unless you intend to use it.

If you have already installed Platform LSF HPC, but want to use Platform Lava, you must use the `rollops` tool to remove Platform LSF HPC, then re-enable Platform Lava by running `chkconfig -add lava` on all affected nodes. On the front end node, you must also run `chkconfig --add lavagui`.

About the Platform Lava GUI

The Platform Lava GUI gives you the ability to monitor and control your Platform Lava jobs, queues, and hosts. This browser-based interface is installed on your front end only; no components are installed on your compute hosts.

The Platform Lava GUI is not intended to be production quality. We hope to get feedback on its usefulness as a tool for future releases.

Supported browsers

- Microsoft® Internet Explorer 5.0 and higher
- Netscape 6.22 and higher
- Mozilla 1.7.3 and higher
    - You may have to resize the window if you use Mozilla. Buttons may not be easily visible if you do not resize the window.

The Platform Lava GUI Web address

Use your browser to navigate to your Platform Lava GUI Web address which is **http://**<*host_name*>**:**<*port_number*>**/Platform**.

The word "Platform" in the URL is case sensitive. Ensure that it appears exactly as it appears here. The default port number is 8080.

Logging in to the Platform Lava GUI

To log in to the Platform Lava GUI:

1. Navigate to **http://**<*host_name*>**:**<*port_number*>**/Platform** with your browser. The default port is 8080.
2. Log in as the Platform Lava Administrator or as a listed user with your OS user name and password.
    a. To login as a Platform Lava Administrator, run `passwd lavaadmin` to set the password. The username is **lavaadmin**, with no password as default. For more information, see *Administering Platform Lava.*

You can not log in as **root**.

Using the MPI job submission scripts

Platform Lava includes two scripts to aid in submitting LAM over Ethernet and MPICH over Ethernet MPI jobs. The scripts are wrappers to `mpirun` which handle the setup of the `machinefile` for MPICH, or the `bhosts` for LAM. The scripts are included with the Lava roll and are located in the `$LSF_BINDIR` directory. The scripts are for use with Lava only, and are named according to the type of MPI:

- `lam-mpirun`: for use with LAM MPI
- `mpich-mpirun`: for use with MPICH MPI.

lam-mpirun

Run the following command to run the `lam-mpirun` script:

```
% bsub -n <num_processors> lam-mpirun -np <num_processors>
<MPI_JOB> <ARGS>
```

Note: A properly configured LAM environment is required before using this wrapper. The `lam-mpirun` will also call `lamboot` and `lamhalt`.

mpich-mpirun

Run the following command to run the `mpich-mpirun` script:

```
% bsub -n <num_processors> mpich-mpirun -np <num_processors>
<MPI_JOB> <ARGS>
```

Known issues

Lava GUI

- Relative priority

  When selecting scheduling options with the submit job screen, **Relative Priority** has the following description: **1 is the highest**. This should read: **1 is the lowest**.

- Host information after restarting Platform Lava

  If you stop Platform Lava and then restart it, the `EXEC_HOST` parameter (the host on which a job ran) for recently run jobs is displayed as `lost_and_found`.

- Host page delay

When monitoring jobs through the **Hosts** page, there is a substantial delay in displaying updated information. Use the **Jobs** page to monitor job status.

- Logging error

  The file `/opt/lava/webgui/log/gabd.log` logs an error every time you log in to the Platform Lava GUI. The error is "Select Interrupted system call." Ignore this error.

- Platform Lava GUI service

  When you check the status of the Platform Lava GUI service in the command line, you get the following descriptive sentence: Script for starting up and shutting down gabd. The word "gabd" should be "Platform Lava GUI."

- Number of disks

  If your host has more than one disk, the Platform Lava GUI displays only one in the host details.

- Mozilla window

  When using Mozilla as your browser, the window does not close properly when you change your working directory by going to the **Jobs** page, choosing **Tools** | **Job Environment**. To fix, close the **Job Environment** window and open it again.

- Email

  When you request an email to be sent to you when a job completes, the actual email sender is LSF and never the Platform Lava GUI. `SMTP_SERVER` and `MAIL_SENDER` in `ga.conf` are ignored.

- Maximum remote pending jobs

  When you look at **Queue** | **Details** for a specific queue, the **Maximum Remote Pending Jobs** has the value **Infinite**. The value should be **5000**.

- Upload/Download

  The Upload/Download feature has been disabled for security reasons.

Troubleshooting

1. **Symptom**: The Lava GUI (`lavagui`) link is broken in the main page after reinstall the Lava Roll using the `rollops` tool.

   Solution: Restart the `lavagui` daemon as follows:

a. `# /etc/init.d/lavagui stop`
b. `# /etc/init.d/lavagui start`

The Lava GUI should function normally after you restart the `lavagui` daemon.

2. **Symptom**: After removing a compute node from your Platform OCS cluster, the `bhosts` or `lsload` command shows that the Platform Lava daemons are still running on the host. This occurs after you run the following command:

   ```
   # insert-ethers --remove host_name
   ```

   Solution: Restart the daemons on the master Platform Lava host by running the following command:

   ```
   # /etc/init.d/lava stop
   # /etc/init.d/lava start
   ```

3. **Symptom**: After physically disconnecting a compute node from your Platform OCS cluster, the `bhosts` or `lsload` command shows that the host is UNKNOWN.

   Solution: Restart the daemons on the master Platform Lava host by running the following commands:

   ```
   # /etc/init.d/lava stop
   # /etc/init.d/lava start
   ```

[Top]

Intel® Software Tools Roll

The Intel® Software Tools roll contains the Intel C++ compiler and the Intel MPI Library 2.0 packages, including library and integrated performance primitives, LAM MPI, and MPICH.

Getting a license for Intel compiler and tools

To get a license (either evaluation or commercial) for the Intel compiler and tools, follow these steps:

1. Go to the following URL:

http://www.intel.com/software/products/distributors/rock_cluster.htm

2. Select the tool for which you want to obtain a license by clicking on the appropriate link under the Intel Software Development Products section.

   You need a separate license for each Intel tool.

3. Get the license:
   - To get a demo/evaluation license, click the **Free Evaluation Software** link in the Evaluate/Purchase table, and follow the instructions.
   - To get a commercial license, click the **Buy/Renew** link in the Evaluate/Purchase table, and follow the instructions.

Troubleshooting

1. **Symptom**: When trying to run or compile applications using the X86, or EM64T Intel MPI package, you get link time and runtime errors.

   Explanation: This occurs because the MPI libraries are not set in the system library path.

   Solution: Run the following commands on a front end and compute nodes.

   - On the front end node, run:
     - For x86 :

       ```
       # echo "/opt/intel_mpi_10/lib" >> /etc/ld.so.conf
       # ldconfig
       ```

     - For Intel EM64T:

       ```
       # echo "/opt/intel_mpi_10/lib64" >> /etc/ld.so.conf
       # ldconfig
       ```

   - For your compute nodes, log into the front end as root and run:
     - For x86:

       ```
       # cluster-fork 'echo "/opt/intel_mpi_10/lib" >> /etc/ld.so.conf; ldconfig'
       ```

     - For Intel EM64T:

       ```
       # cluster-fork 'echo "/opt/intel_mpi_10/lib64" >> /etc/ld.so.conf; ldconfig'
       ```

You will see warnings about shared libraries being too small. They can be safely ignored.

---

2. **Symptom**: Applications compiled with the Intel MKL library cannot run. The following error is encountered when running applications compiled with the Intel MKL:

   ```
   xhpl: error while loading shared libraries: libmkl_lapack64.so: cannot open
   shared object file: No such file or directory
   ```

   Solution: Add the MKL library path to `/etc/ld.so.conf` as follows and run `ldconfig`:

   - For x86, add `/opt/intel/mkl701cluster/lib/32` to `/etc/ld.so.conf`
   - For Intel EM64T, add `/opt/intel/mkl701cluster/lib/em64t` to `/etc/ld.so.conf`

   To avoid the following error:

   ```
   xhpl: relocation error: /usr/lib64/libguide.so: undefined symbol:
   _intel_fast_memset
   ```

   make sure the line `/opt/intel/mkl701cluster/lib/em64t` comes *before* the `/lib64` and `/usr/lib64` lines in `/etc/ld.so.conf` to make it link correctly: For example:

   ```
   /opt/gm/lib
   /usr/X11R6/lib64
   /usr/kerberos/lib
   /usr/X11R6/lib
   /usr/kerberos/lib64
   /usr/lib64/mysql
   /usr/local/topspin/lib64
   /usr/local/topspin/mpi/mpich/lib64
   /opt/gridengine/lib/lx24-amd64
   /opt/intel/mkl701cluster/lib/em64t
   /lib64
   /usr/lib64
   /usr/kerberos/lib64
   /opt/sge/lib/glinux
   /opt/nmi/lib
   /usr/lib64/qt-3.1/lib
   /usr/lib64/mysql
   /usr/X11R6/lib64
   /opt/intel_fce_80/lib
   /opt/intel_cce_80/lib
   ```

CluMon Roll

CluMon (Beta) is an open source cluster monitoring system developed at the National Center for Supercomputing Applications (NCSA) to keep track of its Linux® clusters. CluMon is a tunable system that can be made to work for almost any set of Linux machines.

For more information on CluMon, see http://clumon.ncsa.uiuc.edu.

The optional CluMon roll fully integrates the CluMon monitoring application with Platform Lava or Platform LSF HPC.

The CluMon roll is for both Standard and Enterprise Editions of Platform OCS version 4.1.1-2.0.

CluMon roll architecture

CluMon roll supports x86 Red Hat Enterprise Linux® 4.0 and Intel EM64T Red Hat Enterprise Linux® 4.0.

Known issues

The CluMon roll is beta quality and has the following known issues:

Installation

- CluMon needs Platform Lava or Platform LSF HPC

  CluMon is hard coded to work only with Platform Lava or Platform LSF HPC. In the case that both Platform Lava and Platform LSF HPC are not installed, CluMon does not work with the rest of Platform OCS.

- PCP RPM does not install on Segment Server nodes

  CluMon PCP RPM does not install on Segment Server nodes. You can not use CluMon to monitor these types of nodes.

CluMon Interface

- No scale units on Host List page

  On the **Host List** page of CluMon graphical user interface (GUI), for each host, the related "1 Min Load", "Mem Usage" charts do not have scale units defined. The "Disk Usage" chart is missing all numerical information.

- Incorrect information on **Host List** page

  The **Host List** page does not show all running jobs for each host. In addition, the job's Name, Acct, and Owner information fields are empty.

- Incorrect information on **Queues Information** page

  The queues table on the **Queues Information** page is always empty for all queues. On the **Information** page for each queue, many values are incorrect (for example very large or negative numbers). In addition, in **Jobs Currently Queued for <queue>**, all jobs have the state **Unknown** and all the other columns are empty.

- Closing host

  With a Platform Lava cluster, if you run `badmin hclose` to close a host, the CluMon GUI will not show the data for that host.

- Incorrect information on **Job Information** page

  After a Platform Lava job is submitted, the **Job Information** Page shows that the job is running on 0 nodes. However, a job always runs on at least one node.

- Values on **Job Information** page for Platform Lava jobs

  Platform Lava jobs display unusual values on the Job Information page in the CluMon GUI. For example, the following values are displayed for a Platform Lava job:

  - **submitTime: 1111002560**
  - **startTime: 1111002563**
  - **cpuFactor: 0.000000**
  - **loadSched:-1.997742**
  - **loadStop: -1.997742**
  - **jobPriority: -1**
- Refresh rate

  The CluMon GUI refresh rate is hard-coded to 120 seconds. If you have short-running jobs, you can not easily monitor them using the GUI.

- Error messages on **Job Information** page

  During and after a Platform Lava job runs, the Job Information page displays unclear and incomplete messages indicating that no information for the job is available (in the job information and job variable sections). The queue name is missing.

- Incomplete error message

When the PCP (Performance Co-Pilot) is down at the compute node, the CluMon GUI shows the following incomplete error message: "There were problems encountered while collecting data. The errors were encountered while:".

- No updates after daemons are stopped

  After the CluMon and PCP daemons are stopped, the CluMon GUI can no longer be updated to reflect the fact that CluMon is shut down.

[Top]

PVFS2 Roll

The PVFS2 (Parallel Virtual Filesystem 2) roll is a bundle of all the components you need to run a high-performance distributed file system.

The following groups have collaborated on or supported the development of PVFS:

- Parallel Architecture Research Laboratory at Clemson University
- Mathematics and Computer Science Division at Argonne National Laboratory
- Ohio Supercomputer Center
- NASA Goddard Space Flight Center Code 931
- National Computational Science Alliance (NCSA)

There are three components in PVFS2 1.3.2:

1. Meta server
2. Data server
3. Client

The roll creates a sample distributed file system that should then be fine tuned for your own configuration and hardware. PVFS2 allows the disk space in each node to be accessible to all nodes as a single file system, creating a high-speed file system ideal for datasets and job information. There are some limitations and an administrator should understand them before configuring it. The latest documentation is available from the PVFS2 Web site at: http://www.pvfs.org/pvfs2/documentation.html.

Requirements

One host must be dedicated as the meta server and named **`pvfs2-meta-server-0-0`**. Once this host is installed, all the others will be able to access a sample PVFS2 file system under `/mnt/pvfs2`.

Installation

During installation, an `autofs` configuration file is installed along with the binaries and source code on all clients. On the first startup after installation, the kernel module will be built. The source code is included so that a more optimal kernel module can be built. The default kernel module includes support for Ethernet. You can add InfiniBand and Myrinet® support by rebuilding the kernel module.

If the installation process was interrupted (for example, from a power outage), **`pvfs2-meta-server`** will not reinstall.

Clients

All nodes in the cluster are clients by default.

Data servers and disk space

The data servers provide the disk space that is combined together to form the distributed file system. For example ten data servers with 10 GB of free space could create a distributed file system size of approximately 100 GB. Generally more data servers will provide more space and more speed. Loss of a data server will cause the loss of the portion of the file system it contained.

Compute nodes may be used as data servers with some additional configuration. If a machine is to be a dedicated data server, it should be installed as a PVFS2 meta server appliance. The batch queuing systems SGE, Lava, and LSF HPC will be disabled in this appliance type.

Meta server

The PVFS2 meta server appliance provides both a meta server and a data server. The configuration is intended only as a demonstration.

A real production installation will require the use of one meta server and one or more data servers. Adding additional data servers after the file system is in use is difficult, so they should be allocated during cluster configuration.

The meta server is responsible for maintaining the distributed file system index. This is a critical component of the distributed file system. Currently PVFS2 only allows one meta server per file system. If this host goes down, or if you re-install, all the data is lost.

Naming the meta server hosts

Name all meta server hosts with the following naming convention: **`pvfs2-meta-server-0-0`**.

Do not use the `add-hosts` tool (from the Platform Roll) to add PVFS2 meta or data servers.

MPI, Myrinet, and Infiniband

Support for MPI (Message Passing Interface), Myrinet, and Infiniband is not built in to this roll. For information, please see http://www.pvfs.org/pvfs2/pvfs2-quickstart.html.

If you have Myrinet and Cisco Topspin drivers and want to use them in PVFS2, you need to rebuild the package. Run the configure script and provide one or more of the following options:

`--with-gm=`*<Location of GM installation>*

`--with-ib=`*<Location of Infiniband installation>*

Production cluster configuration

Consult the PVFS2 Web site for more detailed instructions. The following steps outline what is necessary to configure a production cluster.

1. Select one system to be meta server, and install the PVSF2 meta server roll. If possible create a mirrored file system to hold the PVFS2 meta data.
2. Select a number of machines to be data servers.
3. Install using the appropriate appliance:
   - If you do not intend the data servers to be compute nodes, install the PVSF2 meta server appliance on them.
   - If you intend the data servers to be compute nodes, install using the compute appliance. Disable automatic re-installation by running the following commands:
     - # **`service rocks-grub stop`**
     - # **`chkconfig --del rocks-grub`**
4. Run **`pvfs2-genconfig`** to generate the configuration files on the meta server.
5. Copy the configuration files to the data servers.
6. Format the file systems on the data servers by running the following command:

   # **`pvfs2-server pvfs2-fs.config pvfs2-server.conf-$HOSTNAME -f`**

7. Start the meta and data servers: **`/etc/init.d/pvfs2-server start`**.
8. Mount the PVFS2 file system.

[Top]

Modules Roll

The Modules roll contains the Environment Modules package and various add-on tools and module files for LSF HPC environments. The Environment Modules package enables you to dynamically modify your environment via modulefiles.

For more information about the Modules tool, refer to http://modules.sourceforge.net.

module

Provides for the dynamic modification of a user's environment.

Add-ons for modules

Included in this roll are various add-on tools to simply the loading of modules. The tools are as follows:

- An auto-complete capability has been added to the show to allow you to instantly obtain a list of the available modules from the command-line using the **TAB** key.
- The `savemodules` tool allows you to take a snapshot of the modules they have loaded. The user can have all of these modules re-loaded automatically if they log out and log back in. A corresponding `unsavemodules` tool lets a user remove any snapshots taken of the loaded modules.

Module files for HPC environments

Included in this roll are various module files for loading some common HPC environments. These environments include the following:

- MPICH (GNU, Intel)
- MPICH-GM
- LAM (GNU, Intel)
- MVAPICH (Topspin Infiniband)

The following is an example of loading a module for MPICH (GNU):

```
% module load hpc/mpich-ethernet-gnu
% which mpirun
/opt/mpich/gnu/bin/mpirun
```

The following is an example of unloading the module for MPICH (GNU):

```
% module unload hpc/mpich-ethernet-gnu
% which mpirun
mpirun not found
```

[Top]

Extra Tools Roll

The Extra Tools roll is a collection of tools that may be useful for cluster management. It may also contain tools for users.

coNCePTuaL

coNCePTuaL is a network correctness and performance testing language.

A frequently reinvented wheel among network researchers is a suite of programs that test a network's performance. A problem with having umpteen versions of performance tests is that it leads to a variety in the way results are reported. coNCePTuaL is a domain-specific programming language (with associated compiler and other tools) designed to facilitate writing communication benchmarks and reporting the results in as scientific a manner as possible. The language is English-like and easy to read, even by a non-expert in the area of high-performance communication.

coNCePTuaL has been compiled with MPICH support. To use another MPI implementation it will be necessary to recompile the package. Source code is provided in /opt/extras/src

coCNePTuaL is from the Los Alamos National Laboratory.

All available info and man pages are installed with the roll.

whatelse

`whatelse` reports what else is running on a computer.

Sometimes a computer seems slow for no particular reason. whatelse helps determine the source of the problem. The program waits quietly for a length of time then reports what happened on the computer while it was waiting. Among other things, `whatelse` reports process state changes, network activity, memory behavior, and hardware interrupts that occurred.

`whatelse` is from the Los Alamos National Laboratory.

All available info and man pages are installed with the roll.

oddmanout

`oddmanout` runs a program on a group of nodes and reports which nodes gave different output

Homogeneity is a desirable attribute for workstation clusters used for high-performance computing. However, it can be difficult to ensure that all nodes in the cluster are exactly the same. Nodes may have hung processes, filesystems that failed to mount, modules that failed to load, etc. On very large clusters, there may even be nodes with different CPU speeds or amounts of memory. oddmanout helps find nodes that are different from the rest. The idea behind `oddmanout` is to run a command (or set of commands) on every node of a cluster, find the most common output across all nodes, and report those nodes whose output is different from that.

`oddmanout` is from the Los Alamos National Laboratory.

All available info and man pages are installed with the roll. Run **`man oddmanout`** for more information.

[Top]

---

ntop Roll

`ntop` is a network bandwidth monitoring and traffic analysis tool. It allows you to examine the network patterns of your cluster.

`ntop` is bundled with the Platform OCS software.

`ntop` puts the network interfaces into a passive listening mode watching all traffic coming to and from each interface. The tool plots the data into a database and displays this information in the web GUI. There is no manual configuration involved. You can configure `ntop` from the web GUI. For example, you can use the web GUI to change the default listening interface that `ntop` should watch.

Known issue

`ntop` has the following known issue:

- Web page rendering

Some web browsers do not properly render the HTML output of the the `ntop` help and data dump pages.

Using ntop

Connect to the `ntop` web GUI:

In your web browser, go to http://frontend:3000 or https://frontend:3001. The default administrator username and password are both `admin` should you wish to configure the web interface.

You can use the web GUI to perform tasks such as add users, restrict access to pages users can see, and stop `ntop`.

[Top]

Dell Roll

The Dell roll, available in the Enterprise Edition, provides drivers and utilities for Dell™ PowerEdge servers. It provides support for IPMI and Dell OpenManage™ (OM).

The Dell roll also contains a script that can be used to configure the BMC and BIOS of the cluster nodes during deployment. For systems that are not supported by OM, such as the PE SC 1435, only the BMC can be configured using the Linux `ipmitool` utility. For systems that are supported by OM, such as the PE1950, 1955, 2950, and 6950, the script allows configuration of the system BMC and BIOS using the `omconfig` utility provided by OM. Once the system is configured, features such as remote power cycling of servers and console redirection can be used. The script must be enabled to execute on the compute nodes during deployment.

More information on the Dell roll can be found in the Dell Roll readme on the front end through the following link: http://localhost/homepage/list-rolls.cgi.

[Top]

Learn About Platform Products

World Wide Web

The latest information about all supported releases of any Platform product is available on the Platform Web site at http://www.platform.com. Look in the Online Support area for current README files, Release Notes, Upgrade Notices, Frequently Asked Questions (FAQs),

and other helpful information. Also visit http://my.platform.com.

If you have problems accessing the Platform Web site or the Platform FTP site, contact support@platform.com.

Platform training

Platform's Professional Services training courses can help you gain the skills necessary to effectively install, configure and manage your Platform products. Courses are available for both new and experienced users and administrators at our corporate headquarters and Platform locations worldwide. Customized on-site course delivery is also available.

Find out more about Platform Training at http://www.platform.com/Services/Training, or contact trng@platform.com for details.

Technical support

Contact Platform Computing at support@platform.com for technical support. When contacting Platform, please include the full name of your company.

See the Platform Web site at http://www.platform.com/Company/Contact.Us.htm for other contact information.

Platform documentation

Roll readmes are packaged with the rolls, and are accessible from the Roll Call page in your front end node: http://localhost/homepage/list-rolls.cgi.

Platform OCS

The *Readme for Platform OCS* describes the supported architecture and special installation instructions for Platform OCS and is available from our Platform OCS web site:

http://www.platform.com/products/ocs.

Documentation for other Platform products is available in HTML and PDF format on the Platform Web site:

http://www.platform.com/services/support/docs_home.asp.

CentOS

CentOS documentation is available from the following:

http://mirror.centos.org/centos/4.3/os/i386/RELEASE-NOTES-en.html.

Dell OpenManage - OM

Dell OpenManagement documentation is available from the following:

http://support.dell.com/support/edocs/software/smsom/

Platform LSF HPC

Using Platform LSF HPC for Linux® is a comprehensive guide to using LSF HPC.

Running Jobs with Platform Lava

The Platform Lava user's guide contains the concepts and procedures for working with the Platform Lava batch application workload processing software that comes with Platform OCS:

- Scheduling and dispatching jobs
- Workload scheduling on compute nodes
- Managing queues
- The job execution environment

Inside Platform Lava

The Platform Lava administrator's guide describes how to configure and manage a Platform OCS cluster with Platform Lava. This is available from our Platform OCS web site:

http://www.platform.com/products/ocs.

PVFS2

The latest documentation is available from the PVFS2 web site at: http://www.pvfs.org/pvfs2/documentation.html.

ntop

The latest documentation is available from the ntop web site at: http://www.ntop.org/documentation.html.

Modules

The latest documentation is available from the Modules web site at: http://modules.sourceforge.net/.

[Top]

Get Technical Support

Contact Platform

Contact Platform Computing or your Platform OCS vendor for technical support. Use one of the following to contact Platform technical support:

Email

support@platform.com

World Wide Web

http://www.platform.com

Mail

Platform Support
Platform Computing Corporation
3760 14th Avenue
Markham, Ontario
Canada L3R 3T7

When contacting Platform, please include the full name of your company.

See the Platform Web site at http://www.platform.com/Company/Contact.Us.htm for other contact information.

Get patch updates and hotfixes

Obtain the latest patches and hotfixes for Platform OCS from the following page:

http://my.platform.com/products/platform-ocs

To obtain a user name and password, contact Platform Computing technical support at support@platform.com

[Top]

Copyright and Trademarks

Although the information in this document has been reviewed, Platform Computing Corporation ("Platform") does not warrant it to be free of errors or omissions. Platform reserves the right to make corrections, updates, revisions or changes to the information in this document.

UNLESS OTHERWISE EXPRESSLY STATED BY PLATFORM, THE PROGRAM DESCRIBED IN THIS DOCUMENT IS PROVIDED "AS IS" AND WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT WILL PLATFORM COMPUTING BE LIABLE TO ANYONE FOR SPECIAL, COLLATERAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES, INCLUDING WITHOUT LIMITATION ANY LOST PROFITS, DATA, OR SAVINGS, ARISING OUT OF THE USE OF OR INABILITY TO USE THIS PROGRAM.

Trademarks

This product includes software developed by the Rocks Cluster Group at the San Diego Supercomputer Center at the University of California, San Diego and its contributors.

™ Platform Lava is a trademark of Platform Computing Corporation in the United States and in other jurisdictions.

® LSF and LSF HPC are trademarks or registered trademark of Platform Computing Corporation in the United States and in other jurisdictions.

™ ACCELERATING INTELLIGENCE, PLATFORM COMPUTING, and the PLATFORM and Platform OCS logos are trademarks of Platform Computing Corporation in the United States and in other jurisdictions.

PVFS2 is an open source tool developed and maintained by a number of collaborators and supporters including Parallel Architecture Research Laboratory at Clemson University, Mathematics and Computer Science Division at Argonne National Laboratory, Ohio Supercomputer Center, NASA Goddard Space Flight Center Code 931, and National Computational Science Alliance (NCSA). See http://www.pvfs.org for more information.

CluMon is an open source tool developed and maintained by the National Centre for Computing Applications (NCSA).

® Linux is the registered trademark of Linus Torvalds in the U.S. and other countries.

® Red Hat is a registered trademark of Red Hat, Inc.

® Intel and Itanium are registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

® Cisco Systems and Topspin are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the U.S. and in other countries

® Myrinet and Myricom are registered trademarks of Myricom, Inc. in the United States and other countries.

®™ Dell and Dell OpenManage are service marks or registered trademarks of Dell Inc. in the United States and in other countries.

ntop is a GPL open source tool developed and maintained by a number of collaborators and supporters. See http://www.ntop.org for more information.

Other products or services mentioned in this document are identified by the trademarks or service marks of their respective owners.

[Top]

Date Modified: July 06, 2006
Platform Computing: www.platform.com

Platform Support: support@platform.com
Platform Information Development: doc@platform.com