HPC

BAS4 V5.1

Fix11 to Fix12 Upgrade Procedure

REFERENCE
86 A2 43FD 00

HPC

BAS4 V5.1

Fix11 to Fix12 Upgrade Procedure

Software

November 2009

BULL CEDOC
357 AVENUE PATTON
B.P.20845
49008 ANGERS CEDEX 01
FRANCE

REFERENCE
86 A2 43FD 00

Printed in France

Trademarks and Acknowledgements

Preface

Scope and Objectives

BAS4 V5.1 Fix12 includes new versions and fixes for some of the **BAS4 V5.1 Fix 11** software components.

See The *BAS4 V5.1 FIX12 New Features Guide,* the *Lustre Guide and* the *InfiniBand Guide* for more information about the new software components for this release.

Intended Readers

This manual is aimed at Systems Administrators of **BAS4 V5.1 Fix11** clusters who wish to upgrade to **BAS4V5.1 Fix 12**.

Prerequisites

Refer to the *BAS4 V5.1 Fix12 Software Release Bulletin* (SRB) (Ref. 86 A2 52EJ 12) for details of any restrictions which apply to your release.

The Software Release Bulletin contains the latest information regarding BAS4V5.1 Fix12. This should be read first. Contact Bull Technical Support for more information.

Bibliography

- BAS4 V5.1 FIX12 – SRB (Software Release Bulletin) - 86 A2 52EJ 12
- The *BAS4 V5.1 Documentation* CD-ROM (86 A2 97ER 09), delivered with FIX11, includes the following manuals:
  - *BAS4 Installation and Configuration Guide* (86 A2 28ER 11)
  - *BAS4 Administrator's Guide* (86 A2 30ER 12)
  - *BAS4 User's Guide* (86 A2 29ER 09)
  - *BAS4 Application Tuning Guide* (86 A2 19ER 06)
  - *BAS4 Maintenance Guide* (86 A2 46ER 06)
- The *BAS4 V5.1 FIX12 Documentation* CD-ROM (86 A2 47FD 00) includes the following manuals:
  - *BAS4 V5.1 FIX11 to FIX12 Upgrade Procedure* - 86 A2 43FD 00
  - *BAS4 V5.1 FIX12 New Features Guide* - 86 A2 44FD 00
  - *Lustre Guide* - 86 A2 46FD 00
  - *InfiniBand Guide* - 86 A2 42FD 01

Note	The Bull Support Web site may be consulted for product information, documentation, downloads, updates and service offers: http://support.bull.com

Table of Contents

List of Figures

Introduction

The upgrade to **BAS4 V5.1 Fix12** involves an upgrade of the components listed in the *BAS4 V5.1 Fix 12 New Features Guide*, in the *Lustre Guide* and in the *InfiniBand Guide*. Follow the process described in this chapter carefully.

Pre-Requisite

The upgrade to **BAS4 V5.1 Fix12** is only supported for clusters which have **BAS4V5.1 Fix 11** already installed, and with no addition of any new hardware for the upgrade.

See The *Installation and Configuration Guide* (Reference 86 A2 28ER 11) for details on how to install and configure **BAS4 V5.1 Fix 11.**

As a precaution, in case there is a problem when upgrading to BAS4V5.1 Fix 12 it is recommended to back up:

- **The Management Node system**
- **A copy of the Reference Node image for each type of node.**

WARNING

All activity, including jobs scheduled by the Resource Manager/Batch Manager, on the cluster must be stopped, softly and cleanly, before starting the upgrade process.

Upgrade Process Overview

The process to upgrade **BAS4 V5.1 Fix11** clusters to **BAS4 V5.1 Fix12** consists of the following STEPs:

STEP 1	Pre-Installation Operations 1) Save the **SSH** Keys for the Nodes and for the root user. 2) Backup the Cluster Database data files. 3) Specific procedures for **Lustre** and **NFS** clusters.
STEP 2	Install and Configure BAS4 V5.1 Fix 12 on the Management Node(s) 1) Installation of **BAS4 V5.1 Fix 12 DVD.** 2) Configuration of **MPI** on the Management Node.
STEP 3	Install and Configure BAS4 V5.1 Fix 12 on Compute, Login and I/O Reference Nodes 1) Installation and configuration of **BAS4 V5.1 Fix 12** on the Reference Nodes. 2) Reboot of the Reference Node.
STEP 4	Deploy the BAS4V5.1 Fix12 Reference Nodes 1) Installation and configuration of the image server - as necessary. 2) Creation of the reference image of a **Compute** (the same image will be used for the I/O nodes) and **Login** Node. 3) Deployment of the reference images.
STEP 5	Post Installation Configuration Operations 1) Configuration of **InfiniBand** or **Quadrics** interfaces. 2) Optional - Post configuration of **Lustre**. 3) Compilers on the Login Node.
STEP 6	Final Cluster Checks Final checks for the **pdsh**, **NTP**, **syslog-ng**, **Nagios**, **nsctrl** and **conman** services.

STEP 1. Pre-installation Operations

1.1 BAS4 V5.1 Fix 11 files to be saved

Important

As a precaution, in case of any problems with the upgrade to BAS4V5.1 Fix12, the following files should be saved and backed-up.

1.1.1 Save the SSH Keys for the Nodes and for the root User

To avoid RSA identification changes, the **SSH** keys must be kept.

- To keep the node SSH keys, save the **/etc/ssh** directory for each node type (Management Node, Compute Node, Login Node, etc.), assuming that the SSH keys are identical for all nodes of the same type.
- To keep the root user SSH keys, save the **/root/.ssh** directory on the Management Node, assuming that its content is identical on all nodes.

1.1.2 Save the Cluster Database

1. Login as the root user on the Management Node.
2. Enter:

```
su - postgres
```

3. Enter the following commands:

```
cd /var/lib/pgsql/backups
pg_dump -Fc -C -f/var/lib/pgsql/backups/<name_of_clusterdball.sav> clusterdb
pg_dump -Fc -a -f/var/lib/pgsql/backups/<name_of_clusterdbdata.sav> clusterdb
```

For example, **<name_of_clusterdbdata.sav>** might be **clusterdbdata-2006-1105.sav**.

4. Copy the two **.sav** files onto a non-formattable media outside of the cluster.

1.1.3 Save the Service Configuration files

/etc/inittab file

If the **/etc/inittab** file has been customised, then save and back-up the file for each node type.

Nagios

Save the following **BAS4 V5.1 Fix11 Nagios** configuration files:

/etc/nagios/services-tpl.cfg
/etc/nagios/hpccommands.cfg
/etc/nagios/hosts-tpl.cfg

/etc/nagios/hostgroups-tpl.cfg

Other Files

As a precaution it is recommended to save and back-up the configuration files for the following services: **pdsh**, **ganglia**, **syslog-ng**, **ntp**, **snmp**, **postfix**, **SLURM** and **InfiniBand.**

1.2 Stop the Jobs

Check that the Batch Manager/Resource Manager has stopped all jobs that were running, or are scheduled to run on the cluster.

1.3 Services to be stopped

1.3.1 Optional - Lustre Clusters

Important

This section applies to clusters with the Lustre file system installed, and that include data that has to be kept. All data stored in the Lustre file systems should be backed up before Lustre is migrated.

For Lustre 1.6.7.1 and above, the following upgrades are supported:

- Latest **Lustre 1.4.x** version to latest **Lustre 1.6.x** version.
- One minor version to the next (for example, 1.6.2 > 1.6.3 and 1.4.10 > 1.4.11).

Actions to be performed before updating

WARNING

The *Lustre 1.6 Operations Manual,* available from http://www.lustre.org states that a 'rolling upgrade' is possible, meaning that the file system is not taken out of commission for the migration. However, Bull only supports a Lustre migration which has been carried out on a system which has been completely stopped. This ensures that the migration will be risk free and is simpler to carry out

1. Stop Lustre activity.

   Ensure **Lustre** is stopped correctly for all **Lustre** file systems:

```
lustre_util umount -f <fsname> -n <client nodes list | all>
lustre_util stop -f <fsname>
```

Actions specific to Lustre High Availability Clusters

Stop Cluster Suite

a. If necessary, relocate the **Lustre** services on their Primary Node by using the commands below:

```
lustre_migrate hastat -n <io_node_list>
lustre_migrate relocate -n <node>
```

b. Stop the **Lustre** services:

```
lustre_migrate hastop -n <io_node_list>
```

c. Stop **Cluster Suite**:

```
stordepha -c stop -i <all | node list>
```

Backup your LDAP Directory

```
ldapsearch -LLL -x -D cn=Manager,fs=lustre -w secret -H ldap:/// -b
fs=lustre>/somewhere/on/the/management/node/lustre_ldap_backup.ldif
```

Be careful to put the LDIF file somewhere that will be backed up.

Save the ldap backend files

```
service ldap stop
```

The back-end files are in the **/var/lib/ldap/lustre** folder.

Important

The ldap backend files will not be altered by the RPM upgrade. The files are saved as a precaution:

```
cp -r/var/lib/ldap/lustre /var/lib/ldap/lustre.bkp
```

2. Make a backup copy of the **/etc/lustre** directory before continuing.

```
cp -r /etc/lustre /somewhere/on/the/management/node/lustre.bkp
```

WARNING

The directory where these backup files are copied to must not be lost when the Management Node is reinstalled.

3. Save the storage configuration files
   On all I/O nodes: save the **/etc/storageadmin/disknaming.conf** file. This file is not modified by the upgrade RPMs, but if it is lost you will have to manually upgrade the **OST** and **MDT** mapping (Using the **lustre_ost_dba update** and **lustre_mdt_dba update** commands, or by updating the **/etc/lustre/storage.conf** file) to maintain coherency with the mapping provided by the **stormap -l** command.

1.3.2 Optional - NFS Clusters

Stop **NFS** activity before upgrading to **BAS4 V5.1 Fix12.**

```
service nfs stop
```

STEP 2. Install and Configure BAS4 V5.1 Fix12 on the Management Node

Note The installation procedure is interactive. Pay close attention to the messages which appear.

2.1 Install the BAS4 V5.1 Fix 12 DVD-ROM on the Management Node

To identify the media mount points, look at **/etc/fstab** file:

- USB CD-ROMs look like `/dev/scd.../media/...`
- IDE CD-ROMs look like `/dev/hd..../media/...`

Assuming that **/media/cdrom** is the mountpoint for the media.

Mount the **BAS4 V5.1 Fix 12 DVD-ROM**, and then run:

```
cd /media/cdrom
./install.sh
```

Note Once the installation has finished check the root log files to see if there have been any problems. These log files provide details of the status of the machine before and after the installation.

2.1.1 syslog-ng

The **/etc/syslog-ng/syslog-ng.conf** file will need reconfiguring. Add the IP address (Ethernet **eth0** for the administration network) which the server will use for tracking.

a. Search for all the lines which contain the `SUBSTITUTE` string; for example:

```
# Here you HAVE TO SUBSTITUTE ip("127.0.0.1") with the GOOD Inet
Address "X.1.0.65" (use eth0:1 given by "ip addr show")
```

b. Make the changes as explained in the messages (3 substitutions).

2.1.2 MPIBull2

Any **MPIBull2** environment shell scripts previously installed in the **/etc/profile.d/** directories will need to be upgraded with the new versions. These scripts are upgraded on the reference nodes as follows:

```
cp /opt/mpi/mpibull2-<version>/share/mpibull2.*sh /etc/profile.d
```

2.1.3 Restore the Nagios files

Restore the following **BAS4 V5.1 Fix11** files:

/etc/nagios/services-tpl.cfg
/etc/nagios/hpccommands.cfg
/etc/nagios/hosts-tpl.cfg
/etc/nagios/hostgroups-tpl.cfg

Note These files will overwrite the newly installed versions.

2.1.4 lkcd

Run the command below for **lkcd**:

```
chkconfig --add dumputils
```

2.1.5 Optional - Restore the /etc/inittab file

If this file has been saved then restore it on the Management Node.

2.1.6 Restart IPMI

Run the commands below to restart the **IPMI** service.

```
chkconfig --level 2345 ipmi on
service ipmi start
```

2.1.7 Restart Nagios

Run the command below to restart **Nagios**.

```
dbmConfig configure --service nagios
```

2.1.8 Restart all the Services

Run the command below to start all the services.

```
dbmConfigure configure --restart --force
```

2.2 Check the Management Node Services

Run the command below:

```
dbmConfig show
```

If all the config files for the services are **OK**, then reboot the Management Node. If a service is shown as **KO**, check the **config** file for the service and copy over the details from the **BAS4 V5.1 Fix11** config file saved previously (see *Section 1.1.3)* into the file.

Run the **dbmConfig show** command again and if all the services are OK, reboot the Management Node.

See The *BAS4 Installation and Configuration Guide* (Ref: 86 A2 28ER 11), *Chapter 2, STEP 5* for more information on the Management Node services.

STEP 3. Install and Configure BAS4 V5.1 Fix 12 on the Reference Nodes

BAS4 V5.1 Fix 12 is installed on top of the existing **BAS4 V5.1 Fix 11** Reference Nodes, and then a new image, which includes **Fix 12**, is deployed using **KSIS**.

- An existing **BAS4 V5.1 Fix11** Compute Node must be identified as the Compute Reference Node. The **BAS4 V5.1 Fix 12** software is installed and configured on this node, and then an image of this node is deployed onto the Compute Nodes of the cluster.

The same Compute Node image is used for the I/O Nodes. For Lustre clusters the upgrade_lustre_layout.sh script (See *Section 5.7*) configures the upgraded I/O Nodes following the deployment.

- An existing **BAS4 V5.1 Fix11** Login Node must be identified as the Login Reference Node. The **BAS4 V5.1 Fix 12** software is installed and configured on this node, and then an image of this node is deployed onto the other Login Nodes of the cluster.

3.1 Creation of the Compute Reference Node

3.1.1 Install BAS4 V5.1 Fix12 on the Compute Reference Node

To identify the media mount points for the Compute Reference Node, look at the **/etc/fstab** file:

- USB CD-ROMs look like `/dev/scd.../media/...`
- IDE CD-ROMs look like `/dev/hd..../media/...`

Assuming that **/media/cdrom** is the mount point for the media, mount the **BAS4 V5.1 Fix 12 DVD-ROM** and then run:

```
cd /media/cdrom
./install.sh
```

Once the installation has finished check the root log files to see if there have been any problems. These log files provide details of the status of the machine before and after the installation.

3.1.2 syslog-ng

The **/etc/syslog-ng/syslog-ng.conf** file will need reconfiguring to add the IP address of the server on which log files are centralized.

Search for all the lines with contain the SUBSTITUTE string; for example:

```
#-------------------------------------------------------------------
#  Forward to a loghost server
#-------------------------------------------------------------------
# Here you HAVE TO SUBSTITUTE ip("127.0.0.1") with the GOOD Inet Address
"X.1.0.65" (use eth0:1 given by "ip addr show" on the admin station)
destination loghost       { tcp("127.0.0.1" port(5000)); };

. . .

#2) Sending node  I/O status to the admin station
#-------------------------------------------------------------------
#  To send  I/O node status coming from the logger command to the admin
station
#-------------------------------------------------------------------
# Here you HAVE TO SUBSTITUTE ip("127.0.0.1") with the GOOD Inet Address
"X.1.0.65" (use eth0:1 given by "ip addr show" on the admin station)
destination     iologhost     {     udp("127.0.0.1" port(584)); };
```

Make the changes as explained in the messages (2 substitutions).

3.1.3 Intel Version 11.1 Compilers and Runtime Libraries

If the **Intel** version 11.1 compilers are to be installed on the cluster, then run the command below to remove the version 9 Intel runtime libraries installed by **BAS4 V5.1 Fix11**.

```
rpm -e intelruntime-cc_fc-b.91044_91039.Bull
```

Note The **Intel** version 11.1 runtime libraries are included in the **Intel Compiler Suite** version 11.1 and are installed on the nodes that do not include the compilers.

3.1.4 lkcd

Run the command below for **lkcd**:

```
chkconfig -- add dumputils
```

3.1.5 Optional - Restore the /etc/inittab file

If this file has been saved then restore it on the Compute Reference Node.

3.2 Creation of the Login Reference Node

3.2.1 Install BAS4 V5.1 Fix12 on the Login Reference Node

Assuming that **/media/cdrom** is the Login Reference Node media mount point, mount the **BAS4 V5.1 Fix 12 DVD-ROM** and then run:

```
cd /media/cdrom
./install.sh
```

Once the installation has finished check the root log files to see if there have been any problems. These log files provide details of the status of the machine before and after the installation.

3.2.2 syslog-ng

See Configure the **syslog-ng.conf** on the Login Reference Node as described in Section 3.1.2.

3.2.3 Configure the MPI User environment

Note If any other node, e.g. Management Node, is used for the Login Node functions, then the **MPI** environment should be configured on this node.

MPIBull2 comes with different communication drivers and with different process manager communication protocols.

When using the **InfiniBand** OFED/SLURM pairing, the System Administrator has to verify that:

- Users are able to find the **OFED** libraries required
- User jobs can be linked with the **SLURM PMI** library and then launched using the **SLURM** process manager.

The **MPIBull2** RPMs include 2 automatic setup files **/opt/mpi/mpibull2-1.3.9-10.s/share/mpibull2.*sh**, which are used to define default settings for the cluster.

User access to MPIBull2

The administrator has a choice of 3 different way of making **MPIBull2** available to all users:

1. Copying the **mpibull2.*** environment initialization shell scripts from **/opt/mpi/mpiBull2-<version>/share** to the **/etc/profile.d/** directory on the Login Node, according to the environment required. For example:

   For **MPI**:

```
cp /opt/mpi/mpibull2-1.3.9-10.s/share/mpibull2.*sh /etc/profile.d
```

For **Intel** C:

```
cp /opt/intel/cce/<compiler_version>/bin/iccvars.*sh /etc/profile.d
```

For **Intel** Fortran:

```
cp /opt/intel/fce/<compiler_version>/bin/ifortvars.*sh /etc/profile.d
```

2. Use the command below to enable the module with the profile files:

```
test -e /opt/mpi/modulefiles/mpibull2/1.3.9-10.s && echo "export
MODULEPATH=\$MODULEPATH:/opt/mpi/modulefiles/mpibull2/" >>
/etc/profile
```

Then the end user can load their environment by running the command below:

```
module load your_mpi_version
```

3. Asking users to customize their environment by sourcing the `/opt/mpi/mpiBull2_your_version/share/setenv_mpiBull2.*` files.

   Depending on the setup solution chosen, the Administrator must define two things: a default communication driver for their cluster and the default libraries to be linked to, according to the software architecture.

   In all the files mentioned above, the following must be specified:

   a. A **MPIBull2_COMM_DRIVER**, this can be done by using the mpiBull2-devices -d= command to set the default driver. For **InfiniBand** systems, the name of the driver is **ibmr_gen2**.

   b. **MPIBull2_PRELIBS** variable must be exported to the environment containing the reference to the **SLURM** PMI library.

Some examples are provided in the files.

For a cluster using the **OpenIB** InfiniBand communication protocol, the following line must be included in the mpiBull* file:

```
mpibull2-devices -d=ibmr_gen2
```

For a cluster using SLURM, set the following line, and if necessary add the path to the PMI library:

```
export MPIBULL2_PRELIBS="-lpmi
```

When using the **MPI InfiniBand** communication driver, memory locking must be enabled. There will be a warning during the InfiniBand RPM installation if the settings are not correct. The **/etc/security/limits.conf** file must specify both **soft memlock** and **hard memlock** settings, according to the memory capacity of the hardware. These should be set around 4GBs or unlimited.

Note It is mandatory to restart the **sshd** daemons after changing these limits.

3.2.4 Intel Version 11.1 Compilers and Runtime Libraries

If the **Intel** version 11.1 compilers are to be installed on the cluster, then run the command below to remove the version 9 Intel runtime libraries installed by **BAS4 V5.1 Fix11**.

```
rpm -e intelruntime-cc_fc-b.91044_91039.Bull
```

Note The **Intel** version 11.1 runtime libraries are included in the **Intel Compiler Suite** version 11.1 and are installed on the nodes that do not include the compilers.

3.2.5 lkcd

Run the command below for **lkcd**:

```
chkconfig -- add dumputils
```

3.2.6 Optional - Restore the /etc/inittab file

If this file has been saved then restore it on the Login Reference Node.

3.3 Pre Deployment Operations

3.3.1 Optional - NIS Clusters Only

The **NISDOMAIN** definition line has to be added manually to the **/etc/sysconfig/network** file on the Reference Nodes before deployment, as follows:

```
NISDOMAIN=<DOMAIN>
```

3.4 Reference Node Reboot

Important

When the **BAS4 V5.1 Fix 12** distribution is installed on a Reference Node, check and modify the kernel version listed in the **/tftpboot/WWXXYYZZ.conf** file for the Reference Node on the Management Node **BEFORE** the Reference Node is rebooted.

Note **WWXXYYZZ** is the **IP address** of the Management Node in hexadecimal format. For example, if Reference Node **node5** has **10.0.0.6** as its **eth0 IP** address, the name of its **conf** file on the Management Node will be **/tftpboot/0A000006.conf**.

1. Search for the kernel version references in the Reference Node **/tftpboot/WWXXYYZZ.conf** file. These will be listed as below:

```
image=scsi0:/efi/redhat/vmlinuz-<kernel_version>

initrd=scsi0:/efi/redhat/initrd-<kernel_version>.img
```

2. Change the kernel version references to the **BAS4 V5.1 Fix12** kernel version (**2.6.18-B64k.1.26**).

3. Reboot the Reference Node on the network using the **tftpd** protocol.

STEP 4. Deployment of the BAS4 V5.1 Fix12 Reference Nodes

4.1 Deployment Pre-Requisites

The following pre-requisites should be in place before the **BAS4 V5.1 Fix 12** images are created and deployed by **Ksis**:

- **Ksis Image Server** has been installed on the Management Node.
- The **systemimager** service must be running. If not, run the command:

```
service systemimager start
```

- The node boot entry must be configured in the **EFI** Menu. Each node must be configured to boot from the network via the **eth0** interface. If necessary edit the **EFI** menu by deleting all the existing lines and creating one entry for the network boot via **eth0**.

Note Do not change the **EFI** boot configuration for the Reference Node because the image should NOT be deployed to this node before the deployment of the image to the other nodes has been successfully completed.

- The Cluster Database is accessible. This can be checked by running the command:

```
ksis list
```

  The result must be "*no data found*" or an image list with no error messages.

- All the nodes that will receive a particular image, for example the Compute Node image, are hardware equivalent, that is use the same **NovaScale** platform, disks and network interfaces.
- All system files are on local disks and not on the disk subsystem.
- All the nodes for the deployment are powered on. This can be checked by running the **nsctrl** command, for example:

```
nsctrl status node[1-100]
```

  Any nodes that are shown as **inactive** will need to be powered on.

- All the nodes for the deployment must be **up**. This can be checked using the command below from the Management Node:

```
ksis nodelist
```

  – If the status for any of the nodes is different from **up**, then restart **Nagios** by running the following command from the root prompt on the Management Node:

```
service nagios restart
```

- For clusters that use the **RMS** resource manager:
  - Check that the value reported in the **RMS** column is **OUT**, which means that the node is not currently used for computing purposes.
  - Check also that the **status** of each node is **not UNREACH**, which means that there is a problem with network access for the node.

Note **Ksis** cannot be used for a system already in operation with **RMS** running. In this situation the node has to be released from the Management Node.

4.1.1 Interconnect Interface Files

Before the deployment of the **BAS4 V5.1 Fix12** node images, the existing versions of the interconnect interface files in the **/etc/sysconfig/network-scripts/ifcfg-ib0** directory on all nodes must be deleted; otherwise the **config_ipoib** script used to configure the interconnect interfaces will not work, when the newly deployed nodes are rebooted.

The **config_ipoib** script configures the interconnect interfaces and is launched every time a node is booted. **Compute** and **Login** nodes reboot automatically, following a deployment of a new reference image, and new interconnect interfaces will be created on each node.

See The Section 2.7 in the *Installation and Configuration Guide* (Reference 86 A2 28ER 11) and the *InfiniBand Guide* for more information regarding **InfiniBand** interconnect interfaces.

4.2 Create the Images

Create an image of each **BAS4 v5.1 Fix 12** Reference Node.

```
ksis create <image_name> <reference_ node_name>
```

Example

```
ksis create image1 ns1
```

The **ksis create** command will also ask for a check level. Select the **basic** level. If no level is selected, the **basic** level will be selected automatically by default, after the timeout.

4.3 Deploy the Images on the Cluster

Deploy the Reference Node images by running the command:

```
ksis deploy <image_name> node[n-m]
```

If, for example, 3 Compute Nodes are listed as ns[2-4], then enter the following command for the deployment:

```
ksis deploy image1 ns[2-4]
```

Notes

- The reference nodes may be kept as reference nodes and not included in the deployment. Alternatively, the image may be deployed on to them, so that they are included in the cluster. It is recommended that this second option is chosen.
- For the initial configuration verify in the **EFI** menu that there is only one entry which boots from the **eth0** network.

STEP 5. Post Installation Operations

5.1 Restore the I/O Node aliases

The symbolic links (aliases) must be recreated on each I/O node using the information contained within the **disknaming.conf** file. To do this, run the **stormap** command, as below.

If the node is NOT in a High-Availability pair

```
ssh root@<node_name> “stormap –c”
```

If the node is in a High-Availability pair

```
ssh root@<node1_name> “stormap –c”
ssh root@<node2_name> “stormap –c”
```

5.2 Check the Interconnect Interfaces

Check the settings of interconnect interfaces are OK by running the command:

```
pdsh –w node[n,m] cat /etc/sysconfig/network-scripts/ifcfg-ib0
```

Alternatively, to see the interface parameter settings separately for a group of nodes, use the commands below:

```
pdsh –w node[n,m] cat /etc/sysconfig/network-scripts/ifcfg-ib0 |grep IPADDR
```

```
pdsh –w node[n,m] cat /etc/sysconfig/network-scripts/ifcfg-ib0 |grep NETMASK
```

```
pdsh –w node[n,m] cat /etc/sysconfig/network-scripts/ifcfg-ib0 |grep BROADCAST
```

```
pdsh –w node[n,m] cat /etc/sysconfig/network-scripts/ifcfg-ib0 |grep NETWORK
```

```
pdsh –w node[n,m] cat /etc/sysconfig/network-scripts/ifcfg-ib0 |grep ONBOOT
```

Reconfigure those settings, where the values returned by these commands do not match what is required for the cluster.

5.3 Restart syslog-ng

Restart the **syslog-ng** service on all nodes:

```
service syslog-ng restart
```

5.4 Install the Intel Compilers and Tools on the Login Nodes

Follow the installation routine below to install the new **Intel**© **C++** and the **Fortran** compilers, together with the **Intel**© **Math Kernel Library** and the **Intel**© **Debugger**. These tools are installed on the node which contains the Login functionality (this may be a dedicated node or one which is combined with the Management functionalities).

Note Compilers and tools must be installed on each Login Node separately.

1. Install the **Intel** Compilers (**Fortran**, **C/C++**) on the Login Node.
2. Install the **Intel MKL** on the Login Node.
3. Install the **Intel Debugger (IDB)** on the Login Node.

See The **INSTALL.txt** file provided by **Intel** for more details regarding the installation of the **Compilers**, **MKL** and **IDB**.

4. Export the **/opt/intel** directory via **NFS** and mount it on the **Compute Nodes.**

See The *BAS4 V5.1 Fix 12 New Features Guide* for more information on the Intel tools.

5.5 Optional - Lustre Clusters

1. After the Lustre version **1.6.7** packages have been installed save **/etc/lustre/lustre.cfg** as **/etc/lustre/lustre.cfg.rpmnew** and save **/etc/lustre/models/fs1.lmf.ia64** as **/etc/lustre/models/fs1.lmf.rpmnew**.
2. Copy the contents of the backed up **lustre.bkp** directory into **/etc/lustre/** on the Management Node:

```
cp -r /somewhere/on/the/management/node/lustre.bkp/* /etc/lustre/
```

3. Check the storage configuration.

   a. If necessary, restore the **/etc/storageadmin/disknaming.conf** files on the I/O nodes.

Note The **RPM** upgrade to **BAS4 v5.1 Fix12** does not modify the **disknaming.conf** files, therefore in most situations this operation will not be necessary.

   b. If there is a problem and it is not possible to restore the previous version of the **disknaming.conf** file, then regenerate the **disknaming.conf** file on each I/O node.

      i. If a model file is used, run the command, below, on the Management Node.

```
stordepmap -m </etc/storageadmin/models/model file> -p -c
```

         Quit this step only when the **stormap -l** command indicates all **I/O** node devices are **UP**.

      ii. If a model file is NOT used, run the commands below:

```
stordiskname -u -r <IO nodes>
stormap -c
```

4. Restore and update the Lustre configuration files.

   a. **lustre.cfg** file:

      After the upgrade to **BAS4 v5.1 Fix12**, edit the newly installed **/etc/lustre/lustre.cfg.rpmnew** file, and add any modifications that have been previously made to the old **/etc/lustre/lustre.cfg** file. Use the **diff** command to compare the existing **lustre.cfg** file and the new **lustre.cfg.rpmnew** file.

      After backporting the changes into the **lustre.cfg.rpmnew** file, rename it as the **/etc/lustre/lustre.cfg** file.

      There are 3 new groups:

      LUSTRE_MGS (3 items) - All clusters
      DUMP (1 item) - HA clusters only
      LUSTRE_DB_DAEMON (2 items) - HA clusters only

      Read the documentation included in the **lustre.cfg** file and fill the item details as necessary.

   b. File system model files:

      After the upgrade to **BAS4 v5.1 Fix12**, edit the newly installed **fs1.lmf** file, and add any modifications that have been previously made to the old **/etc/lustre/models/fs1.lmf.rpmsave** file. Use the **diff** command to compare the existing **fs1.lmf.rpmsave** file and the new **fs1.lmf** file.

Important

When the newly installed **fs1.lmf** file and your current **lmf** files are compared, one of the differences will be the inclusion of the **-m 0** option in the **mdt_mkfs_options** and **ost_mkfs_options** fields in the new file. The **lustre_util update** command does not work if these fields are modified in any way in your existing file. Please do not add this option to your current **lmf** file.

Actions specific to Lustre High Availability Clusters

Start the Lustre daemons and test SSH connectivity

a. Test **SSH** connectivity by running the command below:

```
pdsh -w <IO node list> "ssh <management node> echo 'OK'" | dshbak -c
```

If there is a problem with **SSH** reconfigure it so that it works.

b. Actions specific to the **LDAP** service.

i. Launch the **LDAP** service

ii. Restore the **LDAP** backend files, as and when needed, by running the command:

```
service ldap start
```

iii. Verify the **LDAP** content by running the command:

```
lustre_ldap show
```

This command will show details of the **Lustre** High Availability file systems that are installed.

Note If an error occurs and no data is displayed, the **LDAP** directory has to be populated with the content that has been backed-up previously, as described in the pre-install section, using the commands below.

```
lustre_ldap init
lustre_ldap override -l
/somewhere/on/the/management/node/lustre_ldap_backup.ldif
lustre_ldap dump -l /tmp/temporary.ldif
```

This last command is used to regenerate the **LDAP** internal indexes.

iv. Add the **LDAP** service to the **chkconfig** file:

```
chkconfig --add ldap
```

c. Launch **lustredbd**

```
service lustredbd.sh start
```

If the Management Node is NOT Highly Available, add the **lustredbd** to the **chkconfig** file:

```
chkconfig --add lustredbd.sh
```

Setup and start Cluster Suite

d. The **Cluster Suite** configuration files must be regenerated. Run the command below to do this:

```
stordepha -c configure -i <all | IO node list> -o lustre [-H]
```

It is recommended that the **Heuristic** functionality of **stordepha** command (option **-H**) is used for High Availability node pairs.

e. Start the **Cluster Suite** daemons:

```
stordepha -c start -i <all | IO node list>
```

5. Check the new **lustre.cfg** file on the Management Node contains the new **MGS** related directives, i.e.

LUSTRE_MGS_HOST
LUSTRE_MGS_NET
LUSTRE_MGS_ABSOLUTE_LOOPBACK_FILENAME

If your cluster is based on a Quadrics interconnect, it is not possible to use 'elan' as the LUSTRE_MGS_NET parameter. This parameter must be set to 'tcp'. This parameter only impacts the network used between the MGS and the other Lustre nodes for administrative communication, it does not define the network used for Lustre data transfers.

6. Run the command below on the Management Node to distribute the **lustre.cfg** file on all servers.

```
lustre_util set_cfg
```

Actions specific to Lustre High Availability Clusters

Enable Lustre High Availability. For all Lustre file systems run the command:

```
lustre_ldap active -f <fsname>
```

After running these commands, it is strongly recommended to wait for 3 minutes. This corresponds to the default duration for the Lustre HA timeout feature, and will ensure that the commands are taken into account correctly.

7. From the Management Node run the **clean_extents_on_dirs.sh** script on all Lustre file systems to remove version 1.4 extents (these are not supported for version 1.6.x).

```
clean_extents_on_dirs.sh <fsname>
```

8. Launch the **upgrade_lustre_layout.sh** script from the Management Node to setup the new **MGS** and upgraded **Lustre** layout.

```
upgrade_lustre_layout.sh
```

   This script will automatically find all the **Lustre** file systems on your cluster, so it only needs to be launched once.

9. Update the **Lustre** file system descriptions. For each Lustre file system, run the command:

```
lustre_util update -f </path/to/lmf_file>
```

10. Restart the Lustre file systems.

```
lustre_util start –f all
```

11. Stop the Lustre file systems

```
lustre_util stop –f all
```

12. Launch the **set_ostrank_from_label.py** script from the Management Node to fix the **OST** labels.

```
set_ostrank_from_label.py
```

   This script runs for all existing file systems, so it only needs to be launched once.

13. Update the **Lustre** file system descriptions. For each Lustre file system, run the command:

```
   lustre_util update -f </path/to/lmf_file>
```

14. Re-initialize the **MGS** server:

```
service mgs stop
service mgs reinstall
service mgs start
```

Actions specific to Lustre High Availability Clusters

Start the Lustre High Availability services:

```
lustre_migrate hastart -n <all | IO node list>
```

15. Restart the Lustre file systems.

```
lustre_util start –f all
lustre_util mount –f all –n all
```

5.5.1 Known issues for the Lustre Upgrade

Lustre Performance Loss

Problem description: If the Lustre **stripe_size** parameter was set to a value lower than **1MB** with **4KB** pages, performance loss may result after updating Lustre to the new version. This is due to the fact that for the previous **Lustre** version, the **stripe_size** parameter was automatically (and silently) adjusted regarding the page size: 1MB minimum on 4KB page size kernels.

Solution: The recommended solution is to comment the **stripe_size** line in the **Lustre** model file corresponding to your filesystem, and run the command **lustre_util update -f <path to .lmf file>**.

5.6 Optional - Start NFS

Start **NFS** by running the command below:

```
service nfs start
```

5.7 Bull Scientific Studio

The **Bull Scientific Studio** RPMs are installed automatically on all the Reference Nodes.

See The *BAS4 V5.1 Fix12 New Features Guide* and *System Release Bulletin* for more information on the libraries included in **Scientific Studio**.

STEP 6. Final Cluster Checks

6.1 Test pdsh

pdsh is a utility that runs commands in parallel on all the nodes or on a group of nodes for a cluster. This is tested as follows:

All nodes

1. Run a command similar to that below from the Management Node as root:

```
pdsh -w ns[8-10] hostname
```

This will give output similar to that in the example below:

```
ns10: ns10
ns9: ns9
ns8: ns8
```

Groups of nodes

2. Run the **dbmGroup** command

```
dbmGroup show
```

This will give output similar to that in the example below:

```
Group Name  Description                    Nodes Name
-------------------------------------------------------
ADMIN       Nodes by type:ADMIN            ns[0,12]
ALL         All nodes except node admin    ns[1-10]
Burning     Burning group                  ns5
COMP        Nodes by type:COMP             ns[1-4,7-8]
COMP128GB   COMPUTE node with 128GB        ns8
COMP48GB    COMPUTE node with 48GB         ns4
Deploy      Deploy group                   ns3
HwRepair    HwRepair group                 ns8
IO          Nodes by type:IO               ns[6,10]
META        Nodes by type:META             ns[5,9]
MYFAME      ensemble des fame du cluster   ns[0,4-6,8-10]
NODES128GB  Nodes by memory size:128GB     ns8
NODES16GB   Nodes by memory size:16GB      ns[1-3,7]
NODES48GB   Nodes by memory size:48GB      ns[4,6,10]
NODES64GB   Nodes by memory size:64GB      ns[0,5,9,12]
QxTest      QxTest group                   ns[0,6]
TEST        TEST group                     ns[5,9]
UnitTest    UnitTest group                 ns[1,9]
```

3. Run a test command for a group of nodes, as shown below:

```
pdsh -g IO date | dshbak  -c
```

4. If **pdsh** is functioning correctly this will give output similar to that in the example below:

```
ns[6,10]
Thu Aug 7 15:35:27 CEST 2008
```

See *Section 2.5.2* in the *BAS4 Installation and Configuration Guide* (Ref. 86 A2 28ER 11) if there are any problems.

6.2 Check NTP

1. Run the following command on a **COMPUTE**, **LOGIN** and **I/O** node:

```
ntpq -p
```

   Check that the output returns the name of the NTP server, and that values are set for the **delay** and **offset** parameters.

2. On the Management Node, start **ntptrace** and check if the Management Node responds:

```
ntptrace 172.17.0.99
```

```
ns0: stratum 11, offset 0.000000, synch distance 0.012695
```

3. From the Management Node, check that the node clocks are identical:

```
pdsh -w ns[0-1] date
```

```
ns0: Tue Aug 30 16:03:12 CEST 2005
ns1: Tue Aug 30 16:03:12 CEST 2005
```

See *Section 2.5.5* in the *BAS4 Installation and Configuration Guide* (Reference 86 A2 28ER 11) if there are any problems.

6.3 Check syslog-ng

1. Check on the Management Node and node host that the **syslog-ng** service has started on both hosts:

```
service syslog-ng status
```

   The output should be:

```
syslog-ng (pid 3451) is running...
```

2. On the node host, run the command below to test the configuration:

```
logger "Test syslog-ng"
```

3. On the node host, verify that the *'Test syslog-ng'* message is included in the **/var/log/messages** file.

4. On the Management Node, verify that the *'Test syslog-ng'* message is included **/var/log/HOSTS/<node_hostname>/messages** file.

See *Section 2.5.4* in the *BAS4 Installation and Configuration Guide* (Reference 86 A2 28ER 11) if there are any problems.

6.4 Check Nagios

Both **nagios** and **httpd** services have to be running on the Management Node, check these as follows:

```
service nagios status
```

Example output:

```
bsm_nagios (pid 31356 31183 19413) is running...
```

```
service httpd status
```

Example output:

```
> httpd (pid 18258 18257 18256 18255 18254 18253 18252 18251 5785) is running
```

1. Start a web browser (Firefox, Mozilla, etc.) and enter the following URL:
   `http://<Management_Node_name>/BSM`

2. Then, left click the **Start Console** button.

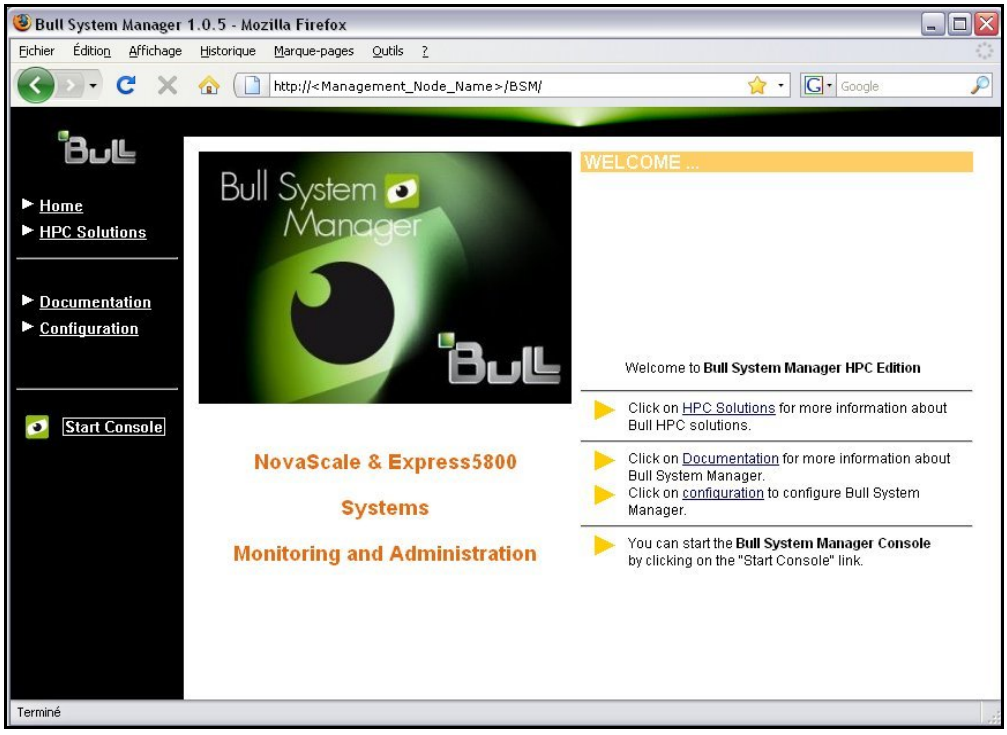

Figure 6-1. Bull System Manager Welcome screen

An authentication window appears asking for a user name and a password.

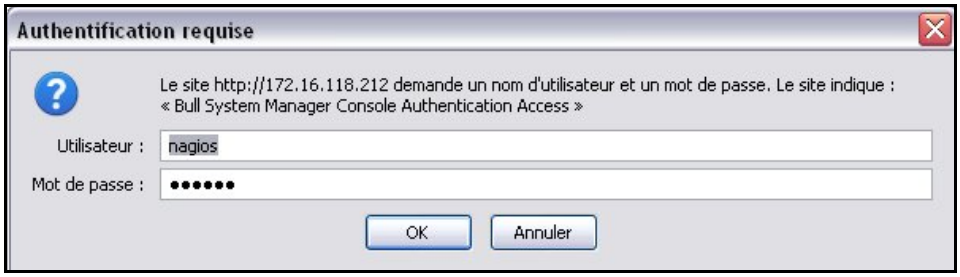

Figure 6-2. **Bull System Manager** Authentication Window

3. Once authenticated, the **Bull System Manager** console appears.

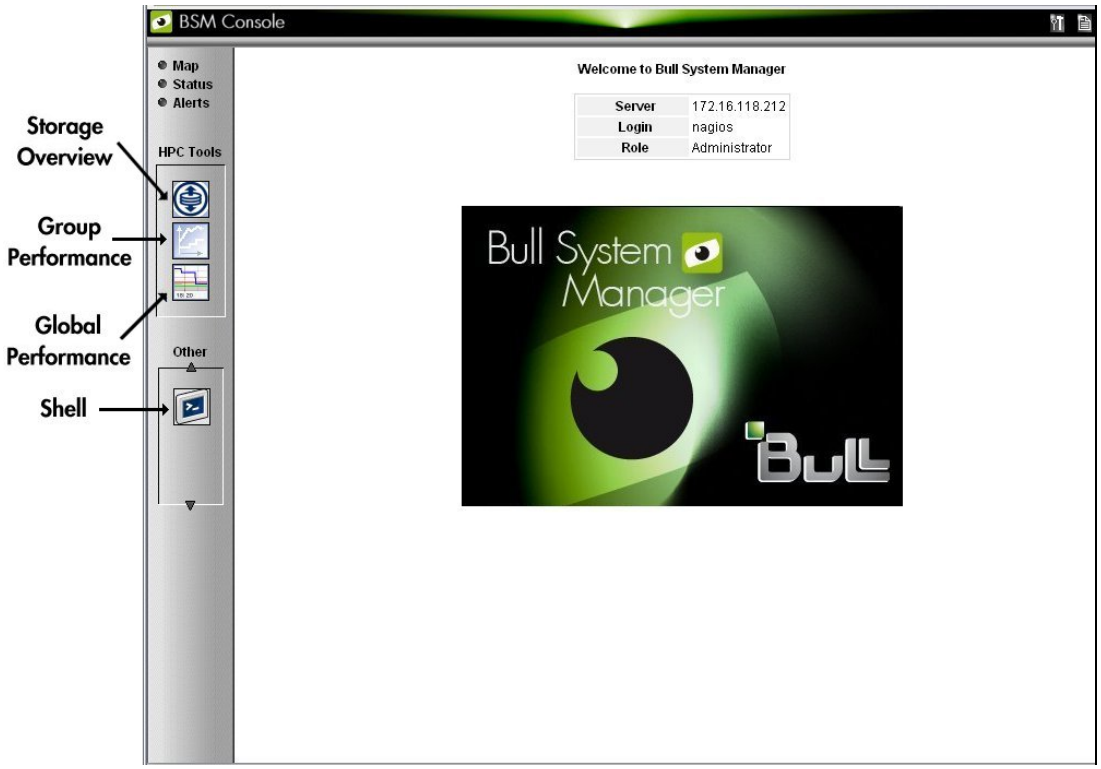

Figure 6-3. The **Bull System Manager** console

Click the **Map** link (top left) to display all the elements that are being monitored.

Figure 6-4. **Bull System Manager Monitoring** Window

See *Chapter 3* in the *BAS4 V5.1 Fix12 New Features Guide* (Reference 86 A2 44FD 00) if there are any problems.

6.5 Check nsctrl

nsctrl is a command that allows administrators to issue commands to the node **PAM** or **PAP**.

usage:
/usr/sbin/nsctrl [options] action nodes

The available actions are: **reset, poweron, poweroff, poweroff_force, status, ping, dump, temperature, infoNS56**

To test **nsctrl**, run a command similar to that below:

```
[root@ns0 ~]# nsctrl status ns[1-5]
```

This will give output similar to that below:

```
ns2 : Chassis Power is on
ns1 : Chassis Power is on
ns3 : Chassis Power is on
ns5 : Chassis Power is on
ns4 : Chassis Power is on
[root@ns0 ~]#
```

See *Chapter 2* in the *BAS4 V5.1 Fix12 New Features Guide* (Reference 86 A2 44FD 00) if there are any problems.

6.6 Check conman

conman is a command that allows administrators to connect to the node consoles.

Usage: conman [OPTIONS] [CONSOLES]

It runs via the **conmand** daemon, and the **dbmConfig** command is used to configure it.

1. Run the command below to check the **conmand** daemon:

```
[root@ns0 ~]# service conman status
```

```
conmand (pid 5943) is running...
[root@ns0 ~]#
```

2. Run a command similar to the one below to check **conman.**

```
[root@ns0 ~]# conman ns2
```

```
 <ConMan> Connection to console [ns2] opened.
 Red Hat Enterprise Linux Server release 5.3 (Tikanga)
Kernel 2.6.18-53.1.21.el5.Bull.1 on an x86_64
 ns2 login:
```

See Section 3.2.2 *in* the *BAS4 Maintenance Guide* (Reference 86 A2 46 ER 06) if there are any problems.

Glossary

A

ACT
Administration Configuration Tool

API
Application Programmer Interface

ARP
Address Resolution Protocol

B

BAS
Bull Advanced Server

BIOS
Basic Input Output System

BLAS
Bull Linux Advanced Server

C

CMOS
Complementary Metal Oxide Semi Conductor

D

DDN
Data Direct Networks

DHCP
Dynamic Host Configuration Protocol

DIB
Device Interface Board

DDR
Double Data Rate

E

ECT
Embedded Configuration Tool

EFI
Extensible Firmware Interface

EIP
Encapsulated IP

EMP
Emergency Management Port

EPIC
Explicitly Parallel Instruction set Computing

F

FCR
Fibre Channel Router

FDA
Fibre Disk Array

FSS
Fame Scalability Switch

FTP
File Transfer Protocol

G

GCC
GNU C Compiler

GNU
GNU's Not Unix

GPL
General Public License

Gratuitous ARP

A gratuitous ARP request is an Address Resolution Protocol request packet where the source and destination IP are both set to the IP of the machine issuing the packet and the destination MAC is the broadcast address xx:xx:xx:xx:xx:xx. Ordinarily, no reply packet will occur. Gratuitous ARP reply is a reply to which no request has been made.

GUID

Globally Unique Identifier

H

HA

High Availability

HPC

High Performance Computing

HSC

Hot Swap Controller

I

IB

Infiniband

IDE

Integrated Device Electronics

IOB

Input/Output Board with 11 PCI Slots

IOC

Input/Output Board Compact with 6 PCI Slots

IPD

Internal Peripheral Drawer

IPMI

Intelligent Platform Management Interface

IPR

IP Router

iSM

Storage Manager (FDA storage systems)

K

KSIS

Utility for Image Building and Deployment

KVM

Keyboard Video Mouse (allows the keyboard, video monitor and mouse to be connected to the PAP or to the node)

L

LAN

Local Area Network

LDAP

Lightweight Directory Access Protocol

LUN

Logical Unit Number

M

MAC

Media Access Control (a unique identifier address attached to most forms of networking equipment)

MDS

MetaData Server

MDT

MetaData Target

MKL

Maths Kernel Library

MPI

Message Passing Interface

N

NFS
Network File System

NPTL
Native POSIX Thread Library

NS
NovaScale

NTFS
New Technology File System (Microsoft)

NTP
Network Time Protocol

NUMA
Non Uniform Memory Access

NVRAM
Non Volatile Random Access Memory

O

OPK
OEM Preinstall Kit (Microsoft)

OST
Object Storage Target

P

PAM
Platform Administration and Maintenance Software

PAP
Platform Administration Processor

PAPI
Performance Application Programming Interface

PCI
Peripheral Component Interconnect (Intel)

PDU
Power Distribution Unit

PMB
Platform Management Board

PMU
Performance Monitoring Unit

PVFS
Parallel Virtual File System

PVM
Parallel Virtual Machine

Q

QBB
Quad Brick Board – The QBB is the heart of the **NovaScale 5xxx/6xxx Series** platforms, housing 4 Itanium™ 2 processors.

R

RAID
Redundant Array of Independent Disks

RMS
Resource Management System (Quadrics)

ROM
Read Only Memory

RSA
Rivest, Shamir and Adleman, the developers of the RSA public key cryptosystem

S

SAFTE
SCSI Accessible Fault Tolerant Enclosures

SDP
Socket Direct Protocol

SDPOIB
Sockets Direct Protocol over Infiniband

SDR
Sensor Data Record

SEL
System Event Log

SCSI
Small Computer System Interface

SIOH
Server Input/Output Hub

SLURM
Simple Linux Utility for Resource Management – an open source, highly scalable cluster management and job scheduling system.

SM
System Management

SMP
Symetric Multi Processing. The processing of programs by multiple processors that share a common operating system and memory.

SMT
Symetric Multi Threading

SNMP
Simple Network Management Protocol

SOL
Serial Over LAN

SSH
Secure Shell

T

TFTP
Trivial File Transfer Protocol

U

USB
Universal Serial Bus

UTC
Coordinated Universal Time

V

VDM
Voltaire Device Manager

VFM
Voltaire Fabric Manager

VGA
Video Graphic Adapter

VLAN
Virtual Local Area Network

VNC
Virtual Network Computing

W

WWPN
World – Wide Port Name

Index

BULL CEDOC
357 AVENUE PATTON
B.P.20845
49008 ANGERS CEDEX 01
FRANCE

REFERENCE
86 A2 43FD 00