HPC

# BAS5 for Xeon

## Installation and Configuration Guide

REFERENCE
86 A2 87EW 01

# HPC

# BAS5 for Xeon

## Installation and Configuration Guide

### Hardware and Software

October 2008

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

REFERENCE
86 A2 87EW 01

Printed in France

## Trademarks and Acknowledgements

# Preface

## Scope and Objectives

This guide describes how to install, or re-install, the Bull HPC **BAS5 for Xeon v1.2** (Bull Advanced Server) software distribution, and all other associated software, on Bull High Performance Computing clusters. It also describes the configuration tasks necessary to make the cluster operational.

## Intended Readers

This guide is for Administrators of Bull **BAS5 for Xeo**n systems.

## Prerequisites

Refer to the **BAS5 for Xeon v1.2** *Software Release Bulletin* (SRB) for details of any restrictions which apply to your release. Use this manual in conjunction with the **BAS5 for Xeon** *High Availability Guide* if your cluster includes any form of High Availability.

## Structure

This manual is organised as follows:

Chapter 1. *Cluster Configuration*
Explains the basics of High Performance Computing in a LINUX environment. It also provides general information about the hardware and software configuration of a Bull **BAS5 for Xeon** HPC system.

Chapter 2. *Updating BAS5 for Xeon v1.1 clusters to BAS5 for Xeon v1.2*
Describes how to update existing BAS5 for Xeon v1.1 clusters to BAS5 for Xeon v1.2.

Chapter 3. *Installing BAS5 for XEON Software on HPC Nodes*
Details the software installation processes possible for the different types of cluster nodes.

Chapter 4. *Configuring Storage Management Services*
Describes how to configure the storage management software to manage the storage systems of the cluster.

Chapter 5. *Configuring I/O Resources for the Cluster*
Describes the use of storage model configuration files.

Chapter 6. *Configuring File Systems*
Describes how to configure NIS on the Login and Compute Nodes, setting NFSv3 file systems and configuring the Lustre Parallel File System.

Chapter 7. *Installing Tools and Applications*
Describes how to install commercial tools (Intel Compilers and MKL) and other applications (Modules).

Chapter 8. *Installing and Configuring InfiniBand Interconnects*
Describes the tasks for the installation and configuration of different Voltaire Devices.

Chapter 9. *Configuring Switches and Card*
Describes how to configure CISCO and Foundry Ethernet switches, Voltaire InfiniBand, and Brocade switches.

Appendix A. *Default Logins for different cluster elements*
Details the default logins for different cluster elements.

Appendix B. *Migrating and Reinstalling the Cluster Database*
Describes how to migrate the Cluster Database

Appendix C. *Migrating Lustre*
Describes how to migrate to Lustre v1.6.x

Appendix D. *Manually Installing BAS5 for Xeon Additional Software*

Appendix E. *Configuring Interconnect Interfaces*
Describes the **config_ipoib** command for Ethernet interconnects Interface description file.

Appendix F. *Binding Services to a Single Network*
Describes the use of the bind attribute in the **/etc/xinetd.conf** file to restrict

Appendix G. *Configuring AOC-USASLP-S8iR RAID Adapters for NovaScale R423 and R425 machines*

Appendix H. *PCI Slot Selection and Server Connectors*

Appendix I. *Activating your Red Hat Account*

*Glossary and Acronyms*
Lists the Acronyms used in the manual.

## Bibliography

Refer to the manuals included on the documentation CD delivered with you system OR download the latest manuals for your Bull Advanced Server **(BAS)** release, and for your cluster hardware, from: http://support.bull.com/

The Bull *BAS5 for Xeon Documentation* CD-ROM (86 A2 91EW) includes the following manuals:

- Bull *HPC BAS5 for Xeon Installation and Configuration Guide* (86 A2 87EW).
- Bull *HPC BAS5 for Xeon Administrator's Guide* (86 A2 88EW).
- Bull *HPC BAS5 for Xeon User's Guide* (86 A2 89EW).
- Bull *HPC BAS5 for Xeon Maintenance Guide* (86 A2 90EW).
- Bull *HPC BAS5 for Xeon Application Tuning Guide* (86 A2 16FA).

- Bull *HPC BAS5 for Xeon High Availability Guide* (86 A2 21FA).

The following document is delivered separately:

- The *Software Release Bulletin* (SRB) (86 A2 71 EJ)

---

**Important**

**The Software Release Bulletin contains the latest information for your BAS delivery. This should be read first. Contact your support representative for more information.**

---

In addition, refer to the following:

- Bull *Voltaire Switches Documentation CD* (86 A2 79ET)
- *NovaScale Master* documentation

For clusters which use the **PBS Professional** Batch Manager:

- PBS Professional 9.2 *Administrator's Guide (*on the *PBS Professional CD-ROM)*
- PBS Professional 9.2 *User's Guide* (on the *PBS Professional CD-ROM*)

## Highlighting

- Commands entered by the user are in a frame in '`Courier`' font, as shown below:

```
mkdir /var/lib/newdir
```

- System messages displayed on the screen are in '`Courier New`' font between 2 dotted lines, as shown below.

```
Enter the number for the path :
```

- Values to be entered in by the user are in '`Courier New`', for example:

  `COM1`

- Commands, files, directories and other items whose names are predefined by the system are in '**Bold**', as shown below:

  The **/etc/sysconfig/**dump file.

- The use of *Italics* identifies publications, chapters, sections, figures, and tables that are referenced.

- < > identifies parameters to be supplied by the user, for example:

  `<node_name>`

**WARNING**

**A Warning notice indicates an action that could cause damage to a program, device, system, or data.**

# Table of Contents


**Chapter 1. Cluster Configuration ..... 1-1**

1.1 Introduction ..... 1-1

1.2 Hardware Configuration ..... 1-1
- 1.2.1 BAS5 for Xeon Cluster architecture ..... 1-1
- 1.2.2 Different architectures possible for BAS5 for Xeon ..... 1-2
- 1.2.3 Service node(s) ..... 1-4
- 1.2.4 Compute Nodes ..... 1-7
- 1.2.5 Networks ..... 1-10
- 1.2.6 High Speed Interconnection ..... 1-10
- 1.2.7 Storage ..... 1-12

1.3 Software Environment ..... 1-14
- 1.3.1 Main Console and Hardware Management ..... 1-14
- 1.3.2 Program Execution Environment ..... 1-15

1.4 Bull BAS5 for Xeon software distribution ..... 1-16
- 1.4.1 Installing Software and Configuring Nodes ..... 1-16

**Chapter 2. Updating BAS5 for Xeon v1.1 clusters to BAS5 for Xeon v1.2 ..... 2-1**

2.1 BAS5 for Xeon v1.1 Files ..... 2-1

2.2 High Availability ..... 2-1
- 2.2.1 Optional - for SLURM clusters only ..... 2-2
- 2.2.2 BAS5 for Xeon v1.1 Configuration files ..... 2-3

2.3 Pre-installation Operations for BAS5 for Xeon v1.2 XHPC Software ..... 2-3

2.4 Pre-installation Operations for BAS5 for Xeon v1.2 Optional Software ..... 2-4

2.5 Install BAS5 for Xeon v1.2 on the Management Node ..... 2-5
- 2.5.1 Configure BAS5 for Xeon v1.2 Management Node ..... 2-5

2.6 Install BAS5 for Xeon v1.2 on the Reference Nodes ..... 2-6

2.7 Deploy the BAS5 for Xeon v1.2 Reference Node Images ..... 2-7
- 2.7.1 Deployment Pre-Requisites ..... 2-7
- 2.7.2 Create an Image ..... 2-8
- 2.7.3 Deploy the Image on the Cluster ..... 2-8
- 2.7.4 Post Deployment Configuration ..... 2-8
- 2.7.5 Configuring Interconnect Interfaces ..... 2-9
- 2.7.6 Post Deployment Operations ..... 2-9
- 2.7.7 Restoring I/O Node aliases ..... 2-9
- 2.7.8 Reconfiguring Cluster Suite on High Availability I/O Nodes ..... 2-11

2.8 Post Deployment Checks ..... 2-11
- 2.8.1 Optional - for SLURM clusters ..... 2-11

**Chapter 3. Installing BAS5 for Xeon v1.2 Software on the HPC Nodes ..... 3-1**

Installation Process Overview ..... 3-2

3.0 Pre-installation Backup Operations when Re-installing BAS5 for Xeon v1.2 ..... 3-3

3.0.1 Saving the ClusterDB ..... 3-3
3.0.2 Saving SSH Keys of the Nodes and of root User ..... 3-4
3.0.3 Saving the Storage Configuration Information ..... 3-4
3.0.4 Saving the Lustre File Systems ..... 3-4
3.0.5 Saving the SLURM Configuration ..... 3-4

3.1 STEP 1: Installing Red Hat Enterprise Linux Software on the Management Node ..... 3-5
3.1.1 Configure Internal RAID discs for BAS5 for Xeon clusters - optional ..... 3-5
3.1.2 Red Hat Enterprise Linux 5 Installation ..... 3-5
3.1.3 Red Hat Linux Management Node Installation Procedure ..... 3-6
3.1.4 Disk partitioning ..... 3-9
3.1.5 Network access Configuration ..... 3-13
3.1.6 Time Zone Selection and Root Password ..... 3-14
3.1.7 Red Hat Enterprise Linux 5 Package Installation ..... 3-15
3.1.8 First boot settings ..... 3-17
3.1.9 Network Configurations ..... 3-17
3.1.10 External Storage System Installation ..... 3-19

3.2 STEP 2: Installing BAS5 for Xeon software on the Management Node ..... 3-20
3.2.1 Preparing the Installation of the Red Hat software ..... 3-20
3.2.2 Preparing the Installation of the BAS5 for Xeon XHPC software ..... 3-21
3.2.3 Preparing the Installation of the BAS5 for Xeon optional software ..... 3-22
3.2.4 Installing the Bull BAS5 for Xeon software ..... 3-22
3.2.5 Database Configuration ..... 3-24

3.3 STEP 3: Configuring Equipment and Installing Utilities on the Management Node ..... 3-26
3.3.1 Generate the SSH keys ..... 3-26
3.3.2 Configuring Equipment ..... 3-27
3.3.3 Configuring Equipment Manually ..... 3-28
3.3.4 Configuring Ethernet Switches ..... 3-29
3.3.5 Configuring Postfix ..... 3-29
3.3.6 Configuring Management Tools Using Database Information ..... 3-29
3.3.7 Configuring Ganglia ..... 3-31
3.3.8 Configuring Syslog-ng ..... 3-31
3.3.9 Configuring NTP ..... 3-32
3.3.10 Configuring the kdump kernel dump tool ..... 3-33
3.3.11 Installing and Configuring SLURM - optional ..... 3-34
3.3.12 3-39
3.3.13 Installing and Configuring PBS Professional Batch Manager – optional ..... 3-39
3.3.14 Installing Intel Compilers and Math Kernel Library ..... 3-42
3.3.15 Configuring the MPI User environment ..... 3-42

3.4 STEP 4: Installing RHEL5.1, BAS5v1.2 for Xeon Software, and optional HPC software products on other nodes ..... 3-45
3.4.1 Preparenfs script prerequisites ..... 3-45
3.4.2 Preparing the NFS node software installation ..... 3-45
3.4.3 Launching the NFS Installation of the BAS5v1.2 for Xeon software ..... 3-48

3.5 STEP 5: Configuring Administration Software on LOGIN, I/O, COMPUTE and COMPUTEX Reference Nodes ..... 3-49
3.5.1 Configuring SSH and /etc/hosts ..... 3-49
3.5.2 Configuring Ganglia ..... 3-50
3.5.3 Configuring the kdump kernel dump tool ..... 3-51
3.5.4 Installing and Configuring SLURM - optional ..... 3-52

3.5.5 Installing and Configuring the PBS Professional Batch Manager – optional ..... 3-56
3.5.6 Installing Compilers ..... 3-58
3.5.7 Intel Math Kernel Library (MKL) ..... 3-58
3.5.8 Configuring the MPI User Environment ..... 3-58
3.5.9 Bull Scientific Studio ..... 3-59
3.5.10 NVIDIA Tesla Graphic Card accelerators – optional ..... 3-59
3.5.11 NVIDIA CUDA Toolkit – optional ..... 3-59
3.5.12 Installing RAID Monitoring Software - optional ..... 3-60

3.6 STEP 6: Creating and Deploying an Image Using Ksis ..... 3-61
3.6.1 Installing, Configuring and Verifying the Image Server ..... 3-61
3.6.2 Creating an Image ..... 3-62
3.6.3 Deploying the Image on the Cluster ..... 3-63
3.6.4 Post Deployment Node configuration ..... 3-63

3.7 STEP 7 : Final Cluster Checks ..... 3-65
3.7.1 Testing pdsh ..... 3-65
3.7.2 Checking NTP ..... 3-66
3.7.3 Checking Syslog-ng ..... 3-66
3.7.4 Checking Nagios ..... 3-66
3.7.5 Checking nsctrl ..... 3-69
3.7.6 Checking Conman ..... 3-69
3.7.7 Testing PBS Professional – Basic setup ..... 3-70
3.7.8 Checking and Starting the SLURM Daemons on COMPUTE(X) and Login/IO Nodes ..... 3-72
3.7.9 Testing kdump ..... 3-72

**Chapter 4. Configuring Storage Management Services ..... 4-1**

4.1 Enabling Storage Management Services ..... 4-2

4.2 Enabling FDA Storage System Management ..... 4-3
4.2.1 Installing and Configuring FDA software on a Linux system ..... 4-4
4.2.2 Configuring FDA Access Information from the Management Node ..... 4-6
4.2.3 Initializing the FDA Storage System ..... 4-6

4.3 Enabling DataDirect Networks (DDN) S2A Storage Systems Management ..... 4-8
4.3.1 Enabling Access from Management Node ..... 4-8
4.3.2 Enabling Date and Time Control ..... 4-8
4.3.3 Enabling Event Log Archiving ..... 4-8
4.3.4 Enabling Management Access for Each DDN ..... 4-8
4.3.5 Initializing the DDN Storage System ..... 4-9

4.4 Enabling the Administration of an Optima 1250 Storage System ..... 4-12
4.4.1 Optima 1250 Storage System Management Prerequisites ..... 4-12
4.4.2 Initializing the Optima 1250 Storage System ..... 4-12

4.5 Enabling the Administration of EMC/Clariion (DGC) storage systems ..... 4-14
4.5.1 Initial Configuration ..... 4-14
4.5.2 Complementary Configuration Tasks for EMC/Clariion CX series storage devices ..... 4-15
4.5.3 Complementary Configuration Tasks for EMC/Clariion AX4-5 storage devices ..... 4-16
4.5.4 Configuring the EMC/Clariion (DGC) Access Information from the Management Node 4-16

4.6 Updating the ClusterDB with Storage Systems Information ..... 4-17

4.7 Storage Management Services ..... 4-17

4.8 Enabling Brocade Fibre Channel Switches ..... 4-18
4.8.1 Enabling Access from Management Node ..... 4-18
4.8.2 Updating the ClusterDB ..... 4-18

Chapter 5. Configuring I/O Resources for the Cluster ..... 5-1

5.1 Automatic Deployment of the I/O Configuration ..... 5-1
5.1.1 Storage Model Files ..... 5-1
5.1.2 Automatic Configuration of a Storage System ..... 5-2
5.1.3 Automatic Deployment of the configuration of I/O resources for the nodes ..... 5-4

5.2 Manual Configuration of I/O Resources ..... 5-5
5.2.1 Manual Configuration of Storage Systems ..... 5-5
5.2.2 Manual Configuration of I/O resources for Nodes ..... 5-5

Chapter 6. Configuring File Systems ..... 6-1

6.1 Setting up NIS to share user accounts ..... 6-1
6.1.1 Configure NIS on the Login Node (NIS server) ..... 6-1
6.1.2 Configure NIS on the Compute or/and the I/O Nodes (NIS client) ..... 6-2

6.2 Configuring NFS v3 to share the /home_nfs and /release directories ..... 6-3
6.2.1 Preparing the LOGIN node (NFS server) for the NFSv3 file system ..... 6-3
6.2.2 Setup for NFS v3 file systems ..... 6-4

6.3 Configuring the Lustre file system ..... 6-5
6.3.1 Enabling Lustre Management Services on the Management Node ..... 6-5
6.3.2 Configuring I/O Resources for Lustre ..... 6-6
6.3.3 Adding Information to the /etc/lustre/storage.conf file ..... 6-7
6.3.4 Configuring the High Availability services (Lustre High Availability clusters only) ..... 6-8
6.3.5 Lustre Pre Configuration Operations ..... 6-8
6.3.6 Configuring the Lustre MGS service ..... 6-9
6.3.7 Lustre Pre-Configuration Checks ..... 6-11
6.3.8 Configuring Lustre ..... 6-12

Chapter 7. Installing Intel Tools and Applications ..... 7-1

7.1 Intel Libraries Delivered ..... 7-1

7.2 Intel Compilers ..... 7-1
7.2.1 Fortran Compiler for Intel® 64 architecture (formerly Intel® EM64T) ..... 7-1
7.2.2 C/C++ Compiler for Intel® 64 architecture (formerly Intel® EM64T) ..... 7-1

7.3 Intel Debugger ..... 7-2

7.4 Intel Math Kernel Library (MKL) ..... 7-2

7.5 Intel Trace Tool ..... 7-2

7.6 Updating Intel Compilers and BAS5 for Xeon v1.2 ..... 7-3

Chapter 8. Installing and Configuring InfiniBand Interconnects ..... 8-1

8.1 Installing HCA-400 Ex-D and Mellanox ConnectX™ Interface Cards ..... 8-1

8.2 Configuring the Voltaire ISR 9024 Grid Switch ..... 8-2
8.2.1 Connecting to a Console ..... 8-2
8.2.2 Starting a CLI Management Session using a serial line ..... 8-2

8.2.3 Starting a CLI Management Session via Telnet ... 8-2
8.2.4 Configuring the Time and Date ... 8-3
8.2.5 Hostname setup ... 8-3
8.2.6 Networking setup ... 8-3
8.2.7 Setting up the switch IP address ... 8-4
8.2.8 Route setup ... 8-4
8.2.9 Routing Algorithms ... 8-5
8.2.10 Subnet manager (SM) setup ... 8-5
8.2.11 Configuring Passwords ... 8-6
8.3 Configuring Voltaire switches according to the Topology ... 8-7
8.3.1 Setting the Topology CLOS stage ... 8-7
8.3.2 Determining the node GUIDs ... 8-8
8.3.3 Adding new Spines ... 8-9
8.4 Performance manager (PM) setup ... 8-11
8.4.1 Performance manager menu ... 8-12
8.4.2 Activating the performance manager ... 8-12
8.5 FTP setup ... 8-13
8.5.1 FTP configuration menu ... 8-13
8.5.2 Setting up FTP ... 8-13
8.6 The Group menu ... 8-13
8.6.1 Group Configuration menu ... 8-14
8.6.2 Generating a group.csv file ... 8-14
8.6.3 Importing a new group.csv file on a switch running Voltaire 3.X firmware ... 8-14
8.6.4 Importing a new group.csv file on a switch running Voltaire 4.X firmware ... 8-15
8.7 Verifying the Voltaire Configuration ... 8-16
8.8 Voltaire GridVision Fabric Manager ... 8-16
8.9 More Information on Voltaire Devices ... 8-16
Chapter 9. Configuring Switches and Cards ... 9-1
9.1 Configuring Ethernet Switches ... 9-1
9.1.1 Ethernet Installation scripts ... 9-1
9.1.2 swtAdmin Command Option Details ... 9-2
9.1.3 Automatic Installation andConfiguration of the Ethernet Switches ... 9-2
9.1.4 Ethernet Switch Configuration Procedure ... 9-3
9.1.5 Ethernet Switches Configuration File ... 9-6
9.1.6 Ethernet Switches Initial Configuration ... 9-6
9.1.7 Basic Manual Configuration ... 9-8
9.2 Configuring a Brocade Switch ... 9-15
9.3 Configuring Voltaire Devices ... 9-16
9.4 Installing Additional Ethernet Boards ... 9-17
Appendix A. Default Logins for different cluster elements ... A-1
Appendix B. Cluster Database Operations ... B-1
B.1 Migrating to BAS5 for Xeon v1.2 ... B-1

B.1.1 Migrating Cluster DB Data from BAS5 for Xeon v1.1 ..... B-1
B.1.2 Migrating Cluster DB Data from BAS4 for Xeon v1.2 ..... B-1
B.1.3 Migrating Cluster DB Data from BAS4 for Xeon v1.1 ..... B-2
B.2 Saving and Reinstalling the Cluster DB data ..... B-2
B.2.1 Saving the Data files ..... B-2
B.2.2 Reinstalling the Data files ..... B-3
B.3 Initializing the Cluster Database using the preload file ..... B-3
Appendix C. Migrating Lustre ..... C-1
C.1 Migrating Lustre from version 1.4 to version 1.6 ..... C-1
C.1.1 Pre-Configuration for Migration ..... C-1
C.1.2 Installation and Configuration of Lustre version 1.6.x RPMS ..... C-2
C.1.3 Post-Configuration operations ..... C-2
Appendix D. Manually Installing BAS5 for Xeon Additional Software ..... D-1
Appendix E. Configuring Interconnect Interfaces ..... E-1
E.1 The config_ipoib command ..... E-1
E.2 Interface Description file ..... E-2
E.2.1 Checking the interfaces ..... E-2
E.2.2 Starting the InfiniBand interfaces ..... E-3
Appendix F. Binding Services to a Single Network ..... F-1
Appendix G. Configuring AOC-USASLP-S8iR RAID Adapters for NovaScale R423 and R425 machines G-1
Appendix H. PCI Slot Selection and Server Connectors ..... H-1
H.1 How to Optimize I/O Performance ..... H-1
H.2 Creating the list of Adapters ..... H-2
H.3 Connections for NovaScale R4xx Servers ..... H-3
H.3.1 NovaScale R421 Series – Compute Node ..... H-3
H.3.2 NovaScale R422 Series – Compute Node ..... H-5
H.3.3 NovaScale R460 Series – Service Node ..... H-7
Appendix I. Activating your Red Hat account ..... I-1
Glossary and Acronyms ..... G-1
Index ..... I-1

# List of Figures


Figure 1-1. Small Cluster Architecture ..... 1-3
Figure 1-2. Medium-sized Cluster Architecture ..... 1-3
Figure 1-3. Large Cluster Architecture ..... 1-4
Figure 1-4. NovaScale R423 server ..... 1-5
Figure 1-5. NovaScale R440 server ..... 1-5
Figure 1-6. NovaScale R460 server ..... 1-5
Figure 1-7. NovaScale R421 server ..... 1-7
Figure 1-8. NovaScale R421 E1 server ..... 1-7
Figure 1-9. NovaScale R422, R422 E1 machine ..... 1-8
Figure 1-10. NVIDIA Tesla S1070 accelerator ..... 1-8
Figure 1-11. NovaScale R425 machine ..... 1-9
Figure 1-12. NVIDIA Tesla C1060 internal graphic card ..... 1-9
Figure 1-13. NovaScale R480 E1 machine ..... 1-9
Figure 3-1. The Welcome Screen ..... 3-6
Figure 3-2. Keyboard installation screen ..... 3-7
Figure 3-3. RHEL5 installation number dialog box ..... 3-7
Figure 3-4. Skip screen for the installation number ..... 3-8
Figure 3-5. First RHEL5 installation screen ..... 3-9
Figure 3-6. Partitioning screen ..... 3-10
Figure 3-7. Confirmation of the removal of any existing partitions ..... 3-11
Figure 3-8. Modifying the partitioning layout – 1st screen ..... 3-11
Figure 3-9. Confirmation to remove existing partitions ..... 3-12
Figure 3-10. RHEL5 Partitioning options screen ..... 3-12
Figure 3-11. Confirmation of previous partitioning settings ..... 3-13
Figure 3-12. Network Configuration Screen ..... 3-13
Figure 3-13. Time Zone selection screen. ..... 3-14
Figure 3-14. Root Password Screen ..... 3-15
Figure 3-15. Software selection screen ..... 3-15
Figure 3-16. Installation screen ..... 3-16
Figure 3-17. Launching NovaScale Master ..... 3-67
Figure 3-18. NovaScale Master Welcome screen ..... 3-67
Figure 3-19. NovaScale Master Authentication Window ..... 3-68
Figure 3-20. The NovaScale Master console ..... 3-68
Figure 3-21. NovaScale Master Monitoring Window ..... 3-69
Figure G-1. Boot screen with Adaptec RAID BIOS ..... G-1
Figure G-2. RAID Configuration Utility Options menu -> Array Configuration Utility ..... G-1
Figure G-3. Array Configuration Utility Main Menu ..... G-2
Figure G-4. Example of Array Properties for a RAID 5 Array ..... G-2
Figure G-5. Example of Array Properties for a RAID 1 array ..... G-3
Figure G-6. Example of drive list for a server ..... G-3
Figure G-7. Selection of drives of the same size for new RAID array ..... G-4
Figure G-8. Array Properties - Array Type ..... G-4
Figure G-9. Array Properties - Write caching ..... G-5
Figure G-10. Array Properties - Confirmation screen ..... G-5
Figure G-11. RAID Configuration Utility - Options Menu ..... G-6
Figure G-12. RAID Configuration Utility - Options Menu -> Controller Configuration ..... G-6

Figure G-13. SMC AOC-USAS-S8iR Controller settings....G-7
Figure G-14. SAS PHY Settings....G-7
Figure G-15. RAID Configuration Utility - Options Menu -> Disk Utilities....G-8
Figure G-16. An example of a drive list for an Adaptec controller....G-8
Figure G-17. RAID Configuration Utility - Exit Utility menu....G-9
Figure H-1. NovaScale R421 rear view of Riser architecture....H-3
Figure H-2. NovaScale R421 rear view connectors....H-4
Figure H-3. NovaScale R422 rear view of Riser architecture....H-5
Figure H-4. NovaScale R422 Rear view connectors....H-5
Figure H-5. NovaScale R460 risers and I/O subsystem slotting....H-7
Figure H-6. Rear view of NovaScale R460 Series....H-7

# List of Tables


Table 8-1. Voltaire ISR 9024 Switch Terminal Emulation Configuration ..... 8-2
Table H-1. PCI-X Adapter Table..... H-2
Table H-2. PCI-Express Table ..... H-2
Table H-3. NovaScale R421 Slots and Connectors..... H-4
Table H-4. NovaScale R422 Slots and Connectors..... H-6
Table H-5. NovaScale R460 Slots and Connectors..... H-8

# Chapter 1. Cluster Configuration

This chapter explains the basics of High Performance Computing in a LINUX environment. It also provides general information about the hardware and software configuration of a Bull **BAS5 for Xeon** HPC system.

The following topics are described:

- 1.1 *Introduction*
- 1.2 *Hardware Configuration*
- 1.3 *Software Environment*
- 1.4 *Bull BAS5 for Xeon software distribution*

## 1.1 Introduction

A cluster is an aggregation of identical or very similar individual computer systems. Each system in the cluster is a 'node'. Cluster systems are tightly-coupled using dedicated network connections, such as high-performance, low-latency interconnects, and sharing common resources, such as storage via dedicated file systems.

Cluster systems generally constitute a private network; this means that each node is linked to the other nodes in the cluster. This structure allows nodes to be managed collectively and jobs to be launched on several nodes of the cluster at the same time.

## 1.2 Hardware Configuration

Bull **BAS5 for Xeon** High Performance Computing cluster nodes use different **NovaScale Xeon** servers.

Cluster architecture and node distribution differ from one configuration to another. Each customer must define the node distribution that best fits his needs, in terms of computing power, application development and I/O activity.

**Note** The System Administrators must have fully investigated and confirmed the planned node distribution, in terms of Management Nodes, Compute Nodes, Login Nodes, I/O Nodes, etc. before beginning any software installation and configuration operations.

A **BAS5 for Xeon** cluster infrastructure consists of **Service Nodes** for management, storage and software development services and **Compute Nodes** for intensive calculation operations.

### 1.2.1 BAS5 for Xeon Cluster architecture

The **BAS5 for Xeon** clusters feature nodes that are dedicated to specific activities.

**Service Nodes** are configured to run the **cluster services**. The cluster services supported by **BAS5 for Xeon** are:

- **Cluster Management**, including installation, configuration changes, general administration and monitoring of all the hardware in the cluster.
- **Login**, to provide access to the cluster and a specific software development environment.
- **I/O**, to transfer data to and from storage units using a powerful shared file system service, either NFS or Lustre (ordered as an option)

Depending on the size and the type of cluster, a single Service Node will cover all the Management, Login and I/O Node functions OR there will be several Service Nodes providing the different functions as shown in the diagrams that follow.

**Compute Nodes** are optimized for code execution; limited daemons run on them. These nodes are not used for saving data but instead transfer data to Service Nodes. There are two types of Compute Nodes possible for Bull **BAS5 for Xeon**.

- Minimal Compute or **COMPUTE** Nodes which include minimal functionality, are quicker and easier to deploy, and require less disk space for their installation. These are ideal for clusters which work on data files (non graphical environment).
- Extended Compute or **COMPUTEX** Nodes which include additional libraries and require more disk space for their installation. These are used for applications that require a graphical environment (X Windows), and also for most ISV applications. They are also installed if there is a need for **Intel Cluster Ready** compliance.

## 1.2.2 Different architectures possible for BAS5 for Xeon

### 1.2.2.1 Small Clusters

On small clusters all the cluster services – Management, Login, and I/O – run on a single Service Node as shown in Figure 1.1.

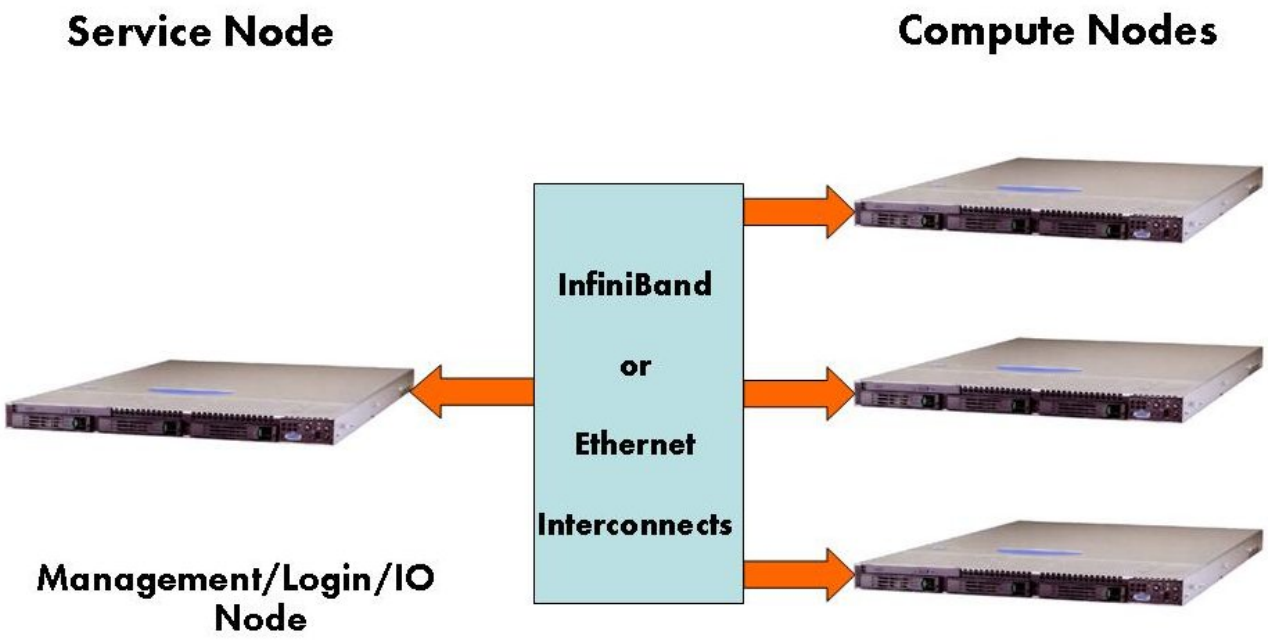


Figure 1-1. Small Cluster Architecture

### 1.2.2.2 Medium-sized Clusters

On medium-sized clusters, one Service Node will run the cluster management services and a separate Service Node will be used to run the Login and I/O services.

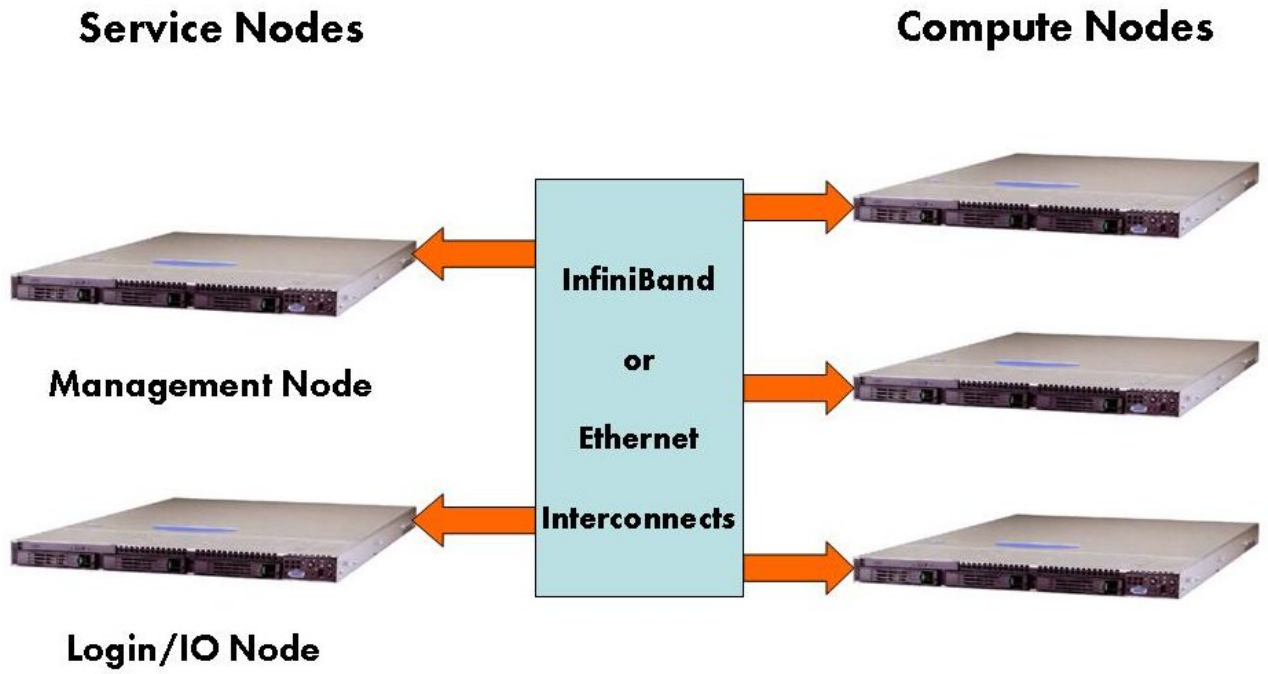


Figure 1-2. Medium-sized Cluster Architecture

### 1.2.2.3 Large clusters

On large clusters the cluster management services run on dedicated nodes. The Login and I/O services also run on separate dedicated nodes. Clusters which use the **Lustre** parallel file system will need at least two separate Service Nodes dedicated to it.

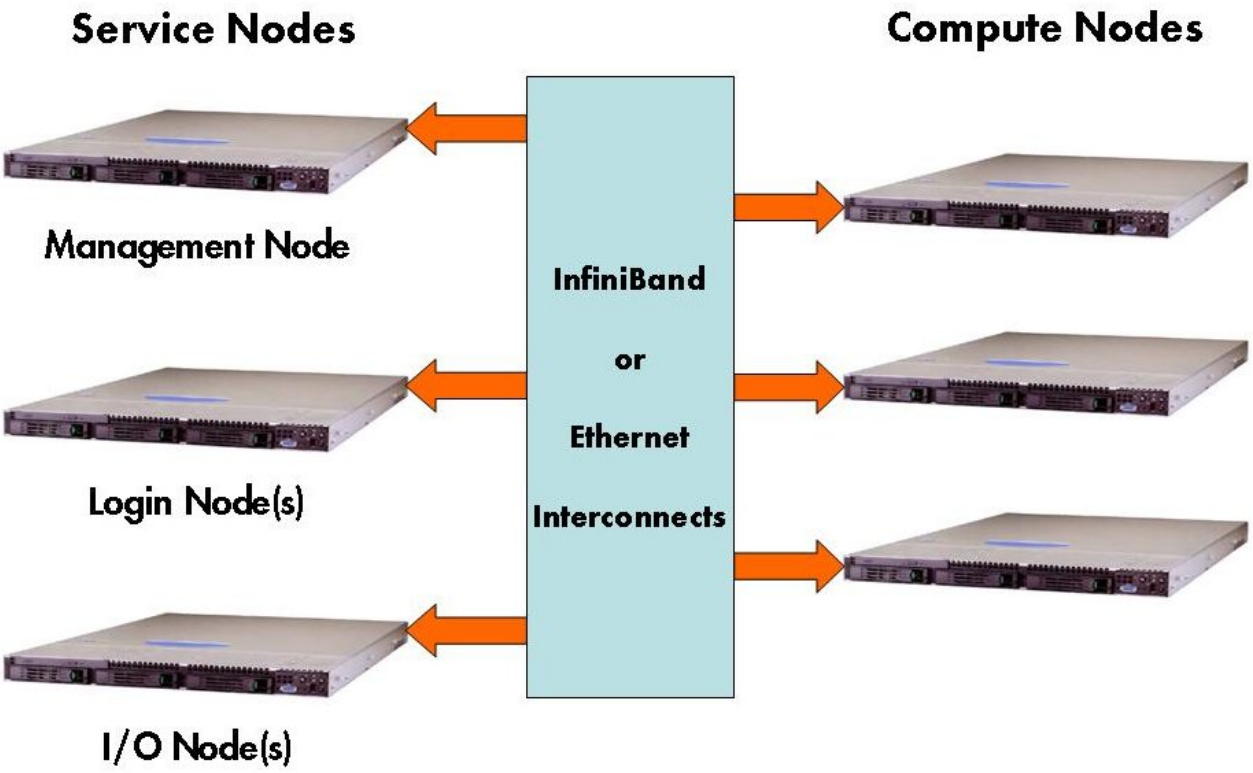


Figure 1-3. Large Cluster Architecture

## 1.2.3 Service node(s)

Bull **NovaScale R423** 2 socket Xeon servers, **R440,** and **R460** and can all be used for the Service Nodes for Bull **BAS5 for Xeon v1.2** Clusters.

### NovaScale R423 servers

Figure 1-4. NovaScale R423 server

**NovaScale R423** servers are double socket, dual or quad core machines that support **SAS,** and **SATA2** 3.5 inch storage disks.

### NovaScale R440 servers

Figure 1-5. NovaScale R440 server

**NovaScale R440** servers are double socket, dual core machines that support **SATA 3.5, SAS 2.5** and **SAS 3.5** storage disks.

### NovaScale R460 servers

Figure 1-6. NovaScale R460 server

**NovaScale R460** servers are double socket, dual core machines that support **SAS** and **SATA2** storage disks.

**Note** From this point onwards the Service Node running the management services will be known as the Management Node. For small clusters, as explained, this node may also include Login and I/O services

### 1.2.3.1 Management Node Services

The **Management Node** is dedicated to providing services and to running the cluster management software. All management and monitoring functions are concentrated on this one node. For example, the following services may be included: **NTP, Cluster DataBase, Kerberos, snmptrapd, ganglia, dhcpd, httpd, conman** etc.

The Management Node can also be configured as a gateway for the cluster. You will need to connect it to the external LAN and also to the management LAN using two different Ethernet cards. A monitor, keyboard and mouse will also need to be connected to the Management Node.

The Management Node houses a lot of reference and operational data, which can then be used by the Resource Manager and other administration tools. It is recommended to store data on an external **RAID** storage system. The storage system should be configured **BEFORE** the creation of the file system for the management data stored on the Management node.

### 1.2.3.2 Login Node Services

**Login Node(s)** are used by cluster users to access the software development and run-time environment. Specifically, they are used to:

- Login
- Develop, edit and compile programs
- Debug parallel code programs.

### 1.2.3.3 I/O Node Services

I/O Nodes provide a shared storage area to be used by the Compute Node when carrying out computations. Either the **NFS** or the **Lustre** parallel file systems may be used to carry out the Input Output operations for **BAS5 for Xeon** clusters.

**Important**

**Lustre must use dedicated service nodes for the I/O functions and NOT combined Login/IO service nodes. NFS can be used on both dedicated I/O service nodes and on combined Login/IO service nodes.**

## 1.2.4 Compute Nodes

The **Compute Nodes** are optimized to execute parallel code. Interconnect Adapters (**InfiniBand** or Gigabit **Ethernet)** must be installed on these nodes.

Bull **NovaScale R421, R421 E1**, **R422, R422 E1, R425** and **R480 E1** servers may all be used as Compute Nodes for **BAS5 for Xeon v1.2**.

### NovaScale R421 and R421 E1 servers

Bull **NovaScale R421** and **R421 E1** servers are double socket, dual or quad core machines.

Figure 1-7. NovaScale R421 server

Figure 1-8. NovaScale R421 E1 server

### NovaScale R422 and R422 E1 servers

Bull **NovaScale R422** and **R422 E1** servers are double socket, dual or quad core machines.

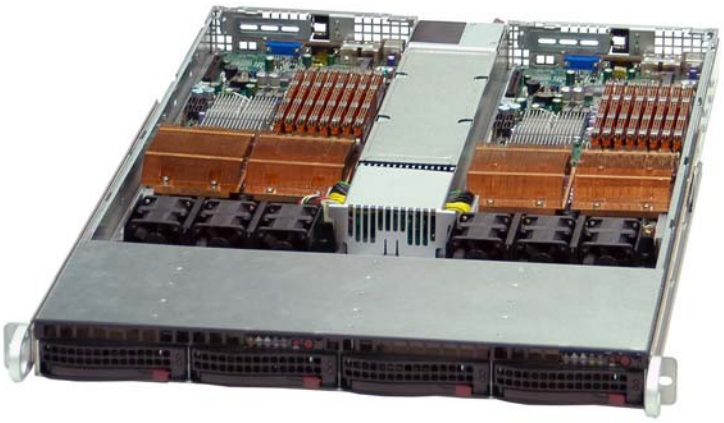

Figure 1-9. NovaScale R422, R422 E1 machine

### NVIDIA Tesla S1070 accelerators

**NovaScale R422 E1** and **R425** servers can be connected with external **NVIDIA Tesla S1070** accelerators resulting in vast improvements in calculation times. Each accelerator card is connected via an external port and 2 PCI cards to the server.

Figure 1-10. NVIDIA Tesla S1070 accelerator

### NovaScale R425 servers

Bull **NovaScale R425** servers are double socket, dual or quad core machines and include a powerful PSU to support internal **NVIDIA Tesla C1060** accelerator cards.

Figure 1-11. NovaScale R425 machine


Figure 1-12. NVIDIA Tesla C1060 internal graphic card

### NovaScale R480 E1 servers

Bull **NovaScale R480 E1** servers are quad socket, dual or quad core machines.

Figure 1-13. NovaScale R480 E1 machine

## 1.2.5 Networks

The cluster contains different networks, dedicated to particular functions, including:

- An **Administration Network**.
- **High speed interconnects,** consisting of switches and cable/boards to transfer data between Compute Nodes and I/O Nodes.

### 1.2.5.1 Administration Network

The **Administration network** uses an **Ethernet** network for the management of the operating system, middleware, hardware (switches, fibre channel cabinets, etc.) and applications from the Management Node.

---

**Note** An optional Ethernet link is necessary to connect the cluster's Login Node(s) to a LAN backbone that is external to the cluster.

---

This network connects all the **LAN1** native ports and the **BMC**, for the nodes using a 10/100/1000 Mb/s network. This network has no links to other networks and includes 10/100/1000 Mb/s Ethernet switch(es).

### 1.2.5.2 Backbone

The **Backbone** is the link between the cluster and the external world.

This network links the Login Node to the external network through a LAN network via Ethernet switches.

For performance and cluster security reasons it is advised to connect the backbone to the Login and Management Nodes only.

## 1.2.6 High Speed Interconnection

### 1.2.6.1 InfiniBand Networks

The following devices may be used for **InfiniBand** clusters.

#### Voltaire Switching Devices

For **InfiniBand** Networks the following **Voltaire**® devices may be used:

- **400 Ex-D** Double Date Rate (**DDR**) Host Channel Adapters which can provide a bandwidth up to 20 Gbs per second, host device PCI-Express.
- **ISR 9024D** switch with 24 DDR ports
- Clusters with up to 288 ports will use **Voltaire® ISR 9096** or **9288** or **2012 Grid Directors** to scale up machines which include **400 Ex-D HCA**s and **ISR 9024** switches.

- Clusters of more than 288 ports will be scaled up using a hierarchical switch structure based on the switches described above.

**See** For more information on installing and configuring Voltaire devices refer to the Chapter on *Installing and Configuring InfiniBand Interconnects* in this manual, and to the Bull *Voltaire Switches Documentation CD*.

#### Mellanox ConnectX™ Dual-Port Cards

**Mellanox ConnectX™ InfiniBand** cards support Dual 4x ports providing a bandwidth of 10 or 20 or 40 Gb/s per port. They support PCI Express 2.0 but are compatible with **PCI-Express 1.1** and fit x8 or x16 slots.

**Card part number DCCH406-DPOO should be used with NovaScale R421, R422, R421 E1 and R422 E1 Compute Nodes.**

### 1.2.6.2 Ethernet Gigabit Networks

**BAS5 for Xeon** Ethernet Gigabit networks can use either **CISCO** or **FOUNDRY** switches as follows:

#### Cisco Switches

- The Host Channel Adapter will use one of the two native ports for each node.
- Clusters with less than 288 ports will use **Cisco Catalyst 3560** (24 Ethernet + 4 SFP ports, 48 Ethernet + 4 SFP ports) switches.
- Clusters with more than 288 ports will use a hierarchical switch structure based on the node switches described above, and with the addition of **Cisco Catalyst 650x** top switches (x= 3,6,9,13) which provide up to 528 ports.

#### Foundry Switches

**BAS5 for Xeon** supports two **FastIron LS** base model switches, **LS 624** and **LS 648,** and the **BIGIRON RX-4, RX-8** and **RX-16** layer 2/3 Ethernet switch rack.

- The **FastIron LS 624** supports twenty-four 10/100/1000 Mbps RJ-45 Ethernet ports. Four ports are implemented as RJ45-SFP combination ports in which the port may be used as either a 10/100/1000 Mbps copper Ethernet port or as a 100/1000 Mbps fiber port when using an SFP transceiver in the corresponding SFP port. The **FastIron LS 624** includes three 10-Gigabit Ethernet slots that are configurable with 10-Gigabit Ethernet single-port pluggable modules.

- The **FastIron LS 648** supports forty-eight 10/100/1000 Mbps RJ-45 Ethernet ports. Four of these ports are implemented as RJ45-SFP combination ports in which the port may be used as either a 10/100/1000 Mbps copper Ethernet or as a fiber 100/1000 Mbps port when using an SFP transceiver in the corresponding SFP port. The **FastIron LS 648** includes two 10-Gigabit Ethernet slots that are configurable with single-port 10-Gigabit Ethernet pluggable modules.
- The **FastIron LS** switches include an integral, non-removable AC power supply. An optional one rack unit high AC power supply unit can be used to provide back–up power for up to four FastIron LS switches.
- The **BIGIRON RX-4, RX-8** and **RX-16** racks include 4, 8 or 16 I/O modules that in turn can accommodate either 1-Gigabit Ethernet or 10-Gigabit Ethernet ports.

---

**See** The www.cisco.com and www.foundry.com for more details regarding these switches.

*Chapter 8* in the **BAS5 for Xeon** *Installation and Configuration Guide* for more information on configuring Ethernet switches.

---

## 1.2.7 Storage

The storage systems supported by **BAS5 for Xeon** include the following:

### Storeway 1500 and 2500 FDA Storage systems

Based on the 4Gb/s FDA (Fibre Disk Array) technology, the Storeway 1500 and 2500 networked FDA Storage systems support transactional data access, associated with fibre and SATA disk media hierarchies. RAID6 double-parity technology enables continued operation even in the case of double disk drive failures, thus providing 100 times better data protection than for RAID5.
Brocade Fibre Channel switches are used to connect FDA storage units and help to ensure storage monitoring within **NovaScale Master HPC Edition**

### Storeway Optima 1250 Storage systems

Developed on Fibre Channel standards for server connections and Serial Attached SCSI (SAS) standards for disk connections, the system can support high-performance disks and high-capacity SAS and SATA disks in the same subsystem. 2 x 4Gb/s FC host ports per controller with a 3 Gb/s SAS channel via SAS and SATA protocol interfaces to the disks.

### EMC/CLARiiON (DGC) Storage systems

The **CX3 Series** models benefit from the high performance, cost-effective and compact **UltraScale** architecture. They support Fibre Channel connectivity, and fit perfectly within **SAN** infrastructures; they offer a complete suite of advanced storage software, in particular **Navisphere Manager**, to simplify and automate the management of the storage infrastructure.

They offer RAID protection levels 0, 1, 1/0, 3, 5 and 6, all of which can co-exist in the same array to match the different protection requirements of data.

They also include a write mirrored cache, a battery backup for controllers and cache vault disks to ensure data protection in the event of a power failure.

The **CX3-40f** model has 8 GB cache memory, 8 x 4 Gb/s FC front-end ports and 8 x 4 Gb/s FC back-end disk ports. It supports up to 240 drives (FC or SATA).

The **CX3-20f** model has 4 GB cache memory, 12 x 4 Gb/s FC front-end ports and 2 x 4Gb/s FC back-end disk ports. It supports up to 120 drives (FC or SATA).

The **CX3-10c** model has 2 GB cache memory, 4 x 4 Gb/s FC front-end ports and 2 x 4 Gb/s FC back-end disk ports and supports full SATA high-capacity disk drive configuration (60 drives).

The **AX4-5** model is a cost-effective solution delivering performance, scalability and advanced data management features. It comes with **Navisphere Express** which simplifies installation, configuration and operation. It offers RAID protection levels 1/0, 3 and 5.

It has 2 GB cache memory, 4 x 4 Gb/s FC front-end ports and 2 x 3 Gb/s SAS back-end expansion ports.

It supports up to 60 SATA or SAS mixed drives.

**Note** The **EMC/CLARiiON CX300** model is supported on older systems.

### DDN S2A 9550 Storage systems

The S2A9550 Storage Appliance is specifically designed for high-performance, high-capacity network storage applications. Delivering up to 3 GB/s large file performance from a single appliance and scaling to 960 TBs in a single storage system.

# 1.3 Software Environment

## 1.3.1 Main Console and Hardware Management

### 1.3.1.1 System Console

The Management Node uses management software tools to control and run the cluster. These tools are used for:

- Power ON/ Power OFF (Force Power Off)
- Checking and monitoring the hardware configuration.
- Serial over LAN

The **IPMI** protocol is used to access the Baseboard Management Controllers (**BMC**) which monitor the hardware sensors for temperature, cooling fan speeds, power mode, etc.

### 1.3.1.2 Hardware Management

Bull **Advanced Server for Xeon** software suite includes different hardware management and maintenance tools that enable the operation and the monitoring of the cluster, including:

**ConMan**: a console management program designed to support a large number of console devices and users connected at the same time. It supports local serial devices and remote terminal servers (via the telnet protocol) and can also use Serial over LAN (via the **IPMI** protocol).

Consoles when managed by **ConMan** provide:

- Access to the firmware shell (**BIOS**) to obtain and modify **NvRAM** information, to choose the boot parameters for the kernel, for example, the disk on which the node boots.
- Visualization of the BIOS operations for a console, including boot monitoring.
- Boot interventions including interactive file system check (**fsck**) at boot.

**NS Commands** : these may be used to configure starting and stopping operations for cluster components. These commands interact with the nodes using the **LAN** administration network to invoke **IPMI_tools** and are described in the *NovaScale Master Remote HW Management CLI Reference Manual*.

**Ksis**: is used to create and deploy software images.

Bull **NovaScale Master HPC Edition**: provides all the monitoring functions for **BAS5 for Xeon** clusters using **Nagios**, an open source application for monitoring the status of all the cluster's components that will trigger an alert if there is a problem. **NovaScale Master** uses **Ganglia**, a second open source tool, to collect and display performance statistics for each cluster node graphically.

## 1.3.2 Program Execution Environment

### 1.3.2.1 Resource Management

Both **Gigabit Ethernet** and **InfiniBand BAS5 for Xeon** clusters can use the **SLURM** (Simple Linux Utility for Resource Management) open-source, highly scalable cluster management and job scheduling program. **SLURM** allocates compute resources, in terms of processing power and Compute Nodes to jobs for specified periods of time. If required the resources may be allocated exclusively with priorities set for jobs. **SLURM** is also used to launch and monitor jobs on sets of allocated nodes, and will also resolve any resource conflicts between pending jobs. **SLURM** helps to exploit the parallel processing capability of a cluster.

**See** The **BAS5 for Xeon** *Administrator's Guide* and *User's Guide* for more information on **SLURM**.

### 1.3.2.2 Parallel processing and MPI libraries

A common approach to parallel programming is to use a message passing library, where a process uses library calls to exchange messages (information) with another process. This message passing allows processes running on multiple processors to cooperate.

Simply stated, a **MPI** (Message Passing Interface) provides a standard for writing message-passing programs. A **MPI** application is a set of autonomous processes, each one running its own code, and communicating other processes through calls to subroutines of the MPI library.

Bull provides **MPIBull2,** Bull's second generation MPI library in the **BAS5 for Xeon** delivery. This library enables dynamic communication with different device libraries, including InfiniBand **(IB)** interconnects, socket Ethernet/IB/EIB devices or single machine devices.

**See** The **BAS5 for Xeon** *User's Guide* for more information on Parallel Libraries.

### 1.3.2.3 Batch schedulers

Different possibilities are supported for handling batch jobs for **BAS5 for Xeon** clusters including **PBS-Professional**, a sophisticated, scalable, robust Batch Manager from **Altair Engineering**. **PBS Pro** works in conjunction with the **MPI** libraries.

**Important**

**PBS Pro does not work with SLURM and should only be installed on clusters which do not use SLURM.**

**See** The **BAS5 for Xeon** *User's Guide* for more information on Batch schedulers, the PBS-Professional *Administrator's Guide* and *User's Guide* available on the **PBS-Pro CD-ROM** delivered for the clusters which use PBS-Pro, and the PBS-Pro web site http://www.pbsgridworks.com

## 1.4 Bull BAS5 for Xeon software distribution

### 1.4.1 Installing Software and Configuring Nodes

The Node distribution architecture planned for your **HPC** system (Management Nodes, Compute Nodes, Login Nodes, I/O Nodes) must be known before installing the **BAS5 for XEON** software.

Chapter 3 explains how to install **BAS5 for Xeon** distribution on a Management Node and how to use the **Prepare NFS** script to install the node function and product **RPM**s required for each type of node.

The software installed on the nominated **Compute**, **Login** or **I/O Nodes** is then used by **Ksis** - a utility for image building and deployment – to create a reference image that is deployed throughout the cluster to create other **Compute, Login** or **I/O** Nodes. The term **Reference Node** designates the node from which the reference image is taken.

# Chapter 2. Updating BAS5 for Xeon v1.1 clusters to BAS5 for Xeon v1.2

**BAS5 for Xeon v1.1** clusters can easily be updated to **BAS5 for Xeon v1.2** using the configuration files that are already in place. Follow the procedure described in this chapter to carry out this update.

**See** The **BAS5 for Xeon v1.2** *Software Release Bulletin* for details of any restrictions which apply either to the updating procedure or for High Availability.

WARNING
**All activity on the cluster must be stopped before carrying out the updating procedure - see the BAS5 for Xeon v1.2 Software Release Bulletin for more information.**

**It is the customer's responsibility to back-up data and their software environment, including configuration files, before using the procedure described in this chapter. For example the /etc/passwd, /etc/shadow files, /root/.ssh directory and the home directory of the users must be saved.**

**All the data must be saved onto a non-formattable media outside of the cluster.**
**It is recommended to use the tar or cp -a command, which maintains file permissions.**

## 2.1 BAS5 for Xeon v1.1 Files

## 2.2 High Availability

For clusters which include some form of High Availability this chapter must be used in conjunction with the **BAS5 for Xeon** *High Availability Guide*. For example, if your cluster includes High Availability for the **Lustre** file system refer to the chapter in the High Availability Guide which refers to the configuration of High Availability for Lustre as well as this chapter.

**The BAS5 for Xeon v1.1 haionfs.conf file for NFS clusters is overwritten when upgrading to BAS5 for Xeon v1.2. The High Availability packages for NFS will need to be reinstalled and the haionfs.conf file edited, as described in Chapter 9 in the High Availability Guide.**

## 2.2.1 Optional - for SLURM clusters only

### 2.2.1.1 SLURM state files

WARNING
**All jobs that are running should be saved and backed up before they are cancelled.**

SLURM state files for version 1.3.2 are different from those for version 1.1.19. This means that it will not be possible to reuse previously saved job and node state information from version 1.1.19. Therefore all version 1.1.19 jobs must be cancelled cleanly before upgrading to version 1.3.2.

This is done by using the commands, below:

```
scancel --state=pending
```

and,

```
scancel --state=running
```

### 2.2.1.2 Uninstall existing version of SLURM

For clusters which include versions of SLURM earlier than 1.3.2, all files, including the Pam_Slurm module, RPMs, config files, must be completely uninstalled before starting the updating operation.

**Note** Save the **slurm.conf** file as the information that it contains can be re-used when regenerating the new **slurm.conf** file.

The command below can be used to check the version of the SLURM files that are installed:

```
rpm -qa \*slurm\*
```

The existing SLURM files can be deleted using the command below:

```
rpm -qa \*slurm\* | xargs rpm -e
```

### 2.2.1.3 Uninstall Munge - optional

If the MUNGE authentication type is used then the existing versions of the MUNGE files will have to be uninstalled.

The command below can be used to check the version of the MUNGE files that are installed:

```
rpm -qa \*munge\*
```

The existing **MUNGE** files can be deleted using the command below:

```
rpm -qa \*munge\* | xargs rpm -e
```

#### 2.2.1.4 SLURM Configuration file

It is recommended that the **slurm.conf** file is rebuilt using the **configurator.html** tool that comes with **SLURM** version **1.3.2**. The cluster information included in the existing **slurm.conf** file can be reused, however new parameters and extra options have been added, for example for the partition parameters.

**See** **STEP 3** in the **Chapter 3** in this manual regarding the use of the **SLURM configurator.html** file to generate the new **slurm.conf** file.

#### 2.2.1.5 SLURM User Scripts

User scripts that previously invoked **srun --allocate**, **--attach** and **-batch** mode options in **SLURM** version **1.1.19** will have to be modified, as these options have been removed and now exist separately as the **salloc**, **sattach**, and **sbatch** commands in **SLURM** version **1.3.2**

**See** The *What's New* chapter in the *Software Release Bulletin* for **BAS5 for Xeon V1.2** for details of the version 1.3.2 changes for **SLURM**.

### 2.2.2 BAS5 for Xeon v1.1 Configuration files

#### Syslog-ng.conf

The **BAS5 for Xeon v1.1 syslog-ng.conf** file must be saved on an external back up device, as this will be used later, before **BAS5 for Xeon v1.2** is installed.

**The BAS5 for Xeon v1.1 syslog-ng.conf file will be overwritten when BAS5 for Xeon v1.2 is installed.**

## 2.3 Pre-installation Operations for BAS5 for Xeon v1.2 XHPC Software

1. Create the directory for the **BAS5 for Xeon v1.2 XHPC** software on the Management Node:

```
mkdir -p /release/XBAS5V1.2
```

2. Insert the **BAS5 for Xeon v1.2 XHPC** DVD-ROM into the DVD reader and mount it:

```
mount /dev/cdrom /media/cdrecorder/
```

3. Copy the **BAS5 for Xeon v1.2 XHPC** DVD-ROM contents into the **/release** directory:

```
cp -a /media/cdrecorder/* /release/XBAS5V1.2/
```

4. Eject the **XHPC** DVD-ROM:

## 2.4 Pre-installation Operations for BAS5 for Xeon v1.2 Optional Software

According to the cluster type and the software options purchased, the preparation for the installation of the Bull **XIB** software and/or the **XLustre** software must now be done.

The **/release/XBAS5V1.2** directory already created for the **XHPC** software on the Management Node will be used, so the only thing to do is copy the **XIB** and **XLustre** software across, as follows:

### XIB software installation

1. Insert the **BAS5 for Xeon v1.2 XIB** DVD-ROM into the DVD reader and mount it:

```
mount /dev/cdrom /media/cdrecorder/
```

2. Copy the **BAS5 for Xeon v1.2 XIB** DVD-ROM contents into the **/release** directory, as shown below:

```
unalias cp
cp -a /media/cdrecorder/* /release/XBAS5V1.2
```

**Note** If the **unalias cp** command has already been executed, the message that appears below can be ignored:
`-bash: unalias: cp: not found`

3. Eject the **XIB** DVD-ROM.

### XLustre software installation

1. Insert the **BAS5 for Xeon v1.2 XLUSTRE** DVD-ROM into the DVD reader and mount it:

```
mount /dev/cdrom /media/cdrecorder/
```

2. Copy the **BAS5 for Xeon v1.2 XLUSTRE** DVD-ROM contents into the **/release** directory:

```
unalias cp
cp -a /media/cdrecorder/* /release/XBAS5V1.2
```

**Note** If the **unalias cp** command has already been executed, the message that appears below can be ignored:
`-bash: unalias: cp: not found`

3. Eject the **XLustre** DVD-ROM

## 2.5 Install BAS5 for Xeon v1.2 on the Management Node

1. Go to the **/release/XBAS5V1.2** directory.

```
cd /release/XBAS5V1.2
```

2. Execute the install command.

```
./install
```

3. Confirm all the installation options that appear.

4. **Optional** - for clusters which use **SLURM**.

**Note** Check that the operations described in *Section 2.2.1* have been carried out before starting the installation and configuration of **SLURM**.

Install and configure **SLURM** on the Management Node as described in STEP 3 in Chapter 3 - *Installing BAS5 for Xeon v1.2 Software on the HPC Nodes*.

**Note** **Munge** will be included in the **SLURM** installation in STEP 3, above, for clusters which use this authentication type for the **SLURM** components.

### 2.5.1 Configure BAS5 for Xeon v1.2 Management Node

The **BAS5 for Xeon v1.2** Management Node will be configured automatically except for the files listed below, where a manual intervention is required.

#### Syslog-ng

The **BAS5 for Xeon v1.2 syslog-ng.conf** file must be manually updated with the cluster details contained in **BAS5 for Xeon v1.1 syslog-ng.conf** file, saved previously. The **BAS5 for Xeon v1.2 syslog-ng.conf** file contains bug fixes. This file should be used and NOT the **BAS5 for Xeon v1.1** file.

#### Nagios.cfg

When **BAS5 for Xeon v1.1** is updated to **BAS5 for Xeon v1.2, Nagios** will not start. Use the old version of the **nagios.cfg** file, which has been renamed as **nagios.cfg.rpmsave**, and relaunch **dmbConfig**.

Important

**This procedure is not fully supported for this release. Contact Bull Technical Support for more information**

### Conman

Run the command below from the Management Node to force **conman** to start for the newly installed **BAS5 for Xeon v1.2** cluster.

```
dbmConfig configure --restart --force
```

## 2.6 Install BAS5 for Xeon v1.2 on the Reference Nodes

Install the **BAS5 for Xeon v1.2** software on the reference nodes for the cluster.

**Note** According to the cluster architecture reference nodes will exist for the following types of nodes:

- COMPUTE or COMPUTEX
- LOGIN/IO or LOGIN and I/O

1. Mount **NFS** from the **/release** directory on the Management Node to the **/release** directory on the Reference Node:

```
ssh <Reference_Node>
mount -t nfs <Management_Node_IP>:/release /release
```

2. Go to the **/release/XBAS5V1.2** directory.

```
cd /release/XBAS5V1.2
```

3. Execute the install command.

```
./install
```

4. Confirm all the installation options that appear.

5. **Optional** - for clusters which use **SLURM**.

**Note** Check that the operations described in *Section 2.2.1* have been carried out before starting the installation and configuration of **SLURM**.

6. Install and configure **SLURM** on the Reference Nodes as described in STEP 5 in Chapter 3 - *Installing BAS5 for Xeon v1.2 Software on the HPC Nodes.*

**See** STEP 5, above, for details on installing **Munge** for clusters which use this authentication type for the **SLURM** components.

## 2.7 Deploy the BAS5 for Xeon v1.2 Reference Node Images

### 2.7.1 Deployment Pre-Requisites

The following pre-requisites should be in place before the new **BAS5 for Xeon v1.2** images are created and deployed by **Ksis**:

- **Ksis Image Server** has been installed on the Management Node.
- The node descriptions and administration network details in the cluster database are up to date and correct.
- 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 COMPUTEX 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. Before creating an I/O node image, for example, all disk subsystems must be unmounted and disconnected.
- Each node is configured to boot from the network via the **eth0** interface. If necessary edit the BIOS menu and set the Ethernet interface as the primary boot device.
- All the nodes for the deployment are powered on. This can be checked by running the **nsctrl** command, for example:

```
nsctrl status xena[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 BAS5 for Xeon v1.1 clusters with High Availability for NFS

1. In the **/etc/modprobe.conf** file, add the **options lpfc lpfc_nodev_tmo=5** line before the lines below :

```
install lpfc modprobe -i lpfc; logger -p local7.info -t "IOCMDSTAT" "LOAD lpfc";
remove lpfc logger -p local7.info -t "IOCMDSTAT" "UNLOAD lpfc"; modprobe -ir lpfc;
```

2. Identify the kernel version installed on the node by running the command:

```
uname -r
```

3. Save the old **initrd** image using the kernel version, identified above:

```
mv /boot/initrd-<kernel_version>.img /boot/initrd-<kernel_version>.img-orig
```

4. Generate a new **initrd** image:

```
mkinitrd -v /boot/initrd-<kernel_version>.img <kernel_version>
```

## 2.7.2 Create an Image

Create an image of each **BAS5 for Xeon v1.2** reference node.

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

### Example

```
ksis create image1 ns1
```

This command will 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.

## 2.7.3 Deploy the Image on the Cluster

Start the deployment 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]
```

**Note** 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.

## 2.7.4 Post Deployment Configuration

### 2.7.4.1 Edit the postconfig script

Before running the **postconfig** command in the section below, the postconfig script will need editing as follows:

1. Run the command below to disable the configuration of the interconnect interface:

```
ksis postconfig disable CONF_60_IPOIB
```

2. Recompile the **postconfig** sript by running the command below:

```
ksis postconfig buildconf
```

#### 2.7.4.2 postconfig command

Once the image deployment has finished, the cluster nodes will need to be configured according to their type - Compute, I/O, etc. Post deployment configuration is mandatory as it configures **Ganglia, Syslog-ng, NTP, SNMP** and **Pdsh** on the nodes.

The **Ksis postconfig** command configures each node that the image has been deployed to in the same way, ensuring that all the cluster nodes of a particular type are homogenous.

**Ksis** post-configuration is carried out by running the command:

```
ksis postconfig run PostConfig <cluster_name>[nodelist]
```

**For example**

```
ksis postconfig run PostConfig xena[1-100]
```

### 2.7.5 Configuring Interconnect Interfaces

Use the **config_ipoib** command to configure the interconnect interfaces for both **InfiniBand** and **Ethernet** networks.

---

**See** *Appendix E - Configuring Interconnect Interfaces* for details on using the **config_ipoib** command.

---

### 2.7.6 Post Deployment Operations

### 2.7.7 Restoring I/O Node aliases

Once the **BAS5 for Xeon v1.2** I/O Reference Nodes have been deployed, the aliases have to be restored, on each I/O Node. According to whether or not a storage model exists for the cluster, either **a.** or **b.**, below, is used to restore the aliases.

a. Where a storage model exists, then use the deployment command from the Management Node, as shown below:

```
stordepmap –m <model_name> -i <nodelist>
```

b. If no storage model exists, use the **stordiskname** command to create a new **disknaming.conf** file, as shown below.

**Important**

The existing **disknaming.conf** file will be erased when the new I/O nodes are deployed. The **stordiskname** command should be used with the **–r** option (remote) from the Management Node enabling backups and restorations of the **/etc/storageadmin/disknaming.conf** file to be managed automatically. If the **-r** option is not used, the Administrator will have to manage the backup of the **/etc/storageadmin/disknaming.conf** file manually.

When used remotely (**–r** option) - immediately after the I/O node deployment - the **stordiskname** command must be used in **update** mode (**-u** option). This ensures that the LUNs are addressed by the same symbolic link names, as used previously, and avoids having to configure the file system again.

i. The **stordiskname** command should be executed from the Management Node as shown below.

If the node is NOT in a High-Availability pair

```
stordiskname –u –r <node_name>
```

If the node is in a High-Availability pair

```
stordiskname –u –r <node1_name>,<node2_name>
```

**Note** For some storage systems, not including **FDA** and **DDN**, the **stordiskname** command may return an error similar to the one below:

```
Error : -= This tool does not manage configuration where a given UID
appears more than once on the node = -
```

If this happens try running it with the **–m SCSI_ID** option.

ii. The symbolic links (aliases) must be recreated on each node using the information contained within the **disknaming.conf** file, newly created by **stordiskname**. 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”
```

### 2.7.8 Reconfiguring Cluster Suite on High Availability I/O Nodes

**See** The **BAS5 for Xeon** *High Availability Guide* for details of how to use the **stordepha** command (Lustre nodes) and the **stordepha.nfs** command (NFS nodes) for clusters which have High Availability in place for the I/O nodes.

## 2.8 Post Deployment Checks

Carry out the post deployment checks that are described in **STEP 7** in Chapter 3 in this manual.

### 2.8.1 Optional - for SLURM clusters

Once **SLURM** version **1.3.2** has been installed following the system update to **BAS5 for Xeon v1.2** then all previously saved state information must be cleared using the **-c** option, for example.:

```
slurmctld -c <job_name>
```

or use the command:

```
/etc/init.d/slurm startclean
```

The node state information for **SLURM** version 1.3.2 will be taken from the new configuration file.

# Chapter 3. Installing BAS5 for Xeon v1.2 Software on the HPC Nodes

**Read this chapter carefully and install the BAS5 for Xeon v1.2 software that applies to your cluster.**

This chapter describes the complete installation process for the **FIRST** installation from scratch of the **BAS5 for Xeon v1.2** software environment on all nodes of a Bull HPC cluster. The same process can also be used for a **reinstallation** of **BAS5 for Xeon v1.2** using the existing configuration files – see section *3.0*.

Different installation options are possible:

- **Red Hat Enterprise Linux Server 5** distribution – all clusters
- Bull **BAS5 for Xeon** distribution – all clusters
- Bull **HPC Toolkit** monitoring tools – all clusters
- Bull **XIB** software – for clusters which use **InfiniBand** interconnects
- Bull **XLustre** software – for clusters which use the Lustre Parallel file system

In addition there are two installation possibilities for the Compute Nodes. These are:

- A Minimal Compute or **COMPUTE** Node, which includes minimal functionality and is quicker and easier to deploy.
- An Extended Compute or **COMPUTEX** Node, which includes additional libraries and will take longer to deploy. These nodes are used for most ISV applications and for applications that require a graphical environment (X Windows). They are also installed if there is a need for Intel Cluster Ready compliance.

**This chapter describes BAS5 for Xeon v1.2 installation process for clusters without any form of High Availability in place. For clusters which include some form of High Availability this manual must be used in conjunction with the BAS5 for Xeon *High Availability Guide*. For example, if your cluster includes High Availability for the Lustre file system, refer to the chapter in the High Availability Guide which refers to the configuration of High Availability for Lustre as well as this chapter.**

**See** The **Software Release Bulletin** delivered with your **BAS5 for Xeon** release for details of any restrictions which may apply.

# Installation Process Overview

The process to install Bull **BAS5 for Xeon v1.2** on the HPC cluster's nodes is divided into different steps, to be carried out in the order shown below:

| | | |
|---|---|---|
| Backups Operations when Re-installing BAS5 for Xeon v1.2<br>**Skip this step if you are installing for the first time.**<br>This step only applies in the case of a re-installation, when the cluster has already been configured (or partially configured) and there is the desire to save and reuse the existing configuration files for the re-installation of **BAS5 for Xeon v1.2**. | | |
| STEP 1 | Installing the RHEL5.1 software on the Management node<br>1) RAID configuration – optional.<br>2) Installation of the **Red Hat Enterprise Linux 5** Server software<br>3) First boot settings<br>4) Configuring the Network<br>5) Installing an external Storage System | Page 3-5 |
| STEP 2 | Installing Bull BAS5 for Xeon software on the Management Node<br>1) Installing Bull **XHPC**, **XIB** and **XLustre** software<br>2) Database Configuration | Page 3-20 |
| STEP 3 | Configuring equipment and installing utilities on the Management Node<br>1) Configuring Equipment Manually (small clusters only)<br>2) Configuring Ethernet switches<br>3) Installing and configuring **Ganglia**, **Syslog-ng**, **NTP**, **Postfix**, **Kdump**, **SLURM** and **PBS Pro** on the Management Node<br>4) Installing compilers (only on Management Nodes which include Login functionality)<br>5) Configuring the **MPI** User environment. | Page 3-26 |
| STEP 4 | Installing RHEL5.1, BAS5v1.2 for Xeon Software, and optional HPC software products on other nodes<br>1) Specifying the software and the nodes to be installed<br>2) Running the **preparenfs** script | Page 3-42 |
| STEP 5 | Configuring Administration Software on Login, I/O, COMPUTE and COMPUTEX Nodes<br>1) Installing and configuring **ssh**, **Kdump**, **SLURM** and **PBS Pro** as necessary<br>2) Installing compilers on Login Nodes<br>3) Configuring the **MPI** User environment<br>4) Installing **RAID** monitoring software - optional | Page 3-49 |
| STEP 6 | Creating an image and deploying it on the cluster nodes using Ksis<br>1) Installation and configuration of the image server<br>2) Creation of the image of a COMPUTE(X) or LOGIN Node previously installed<br>3) Deployment of this image on cluster nodes | Page 3-61 |
| STEP 7 | Final Cluster Checks | Page 3-65 |

# 3.0 Pre-installation Backup Operations when Re-installing BAS5 for Xeon v1.2

This step describes how to save the ClusterDB database and other important configuration files. Use this step only when re-installing **BAS5 for Xeon v1.2** where the cluster has already been configured (or partially configured), and there is the need to save and reuse the existing configuration files.

Skip this step when installing for the first time.

**WARNING**
**The Operating System will be installed from scratch, erasing all disk contents in the process.**

It is the customer's responsibility to save data and their software environment, before using the procedure described in this chapter. For example the **/etc/passwd**, **/etc/shadow files**, **/root/.ssh** directory and the **home** directory of the users must be saved.

**All the data must be saved onto a non-formattable media outside of the cluster.**
**It is recommended to use the tar or cp -a command, which maintains file permissions.**

## 3.0.1 Saving the ClusterDB

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.

### 3.0.2 Saving SSH Keys of the Nodes and of 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.

These directories must be restored once the installation has finished (see 3.5.1 Configuring SSH).

### 3.0.3 Saving the Storage Configuration Information

The following configuration files, in the **/etc/storageadmin** directory of the Management Node, are used by the storage management tools. It is strongly recommended that these files are saved onto a non-formattable media, as they are not saved automatically for a re-installation.

- **storframework.conf** configured for traces, etc
- **stornode.conf** configured for traces, etc.
- **nec_admin.conf** configured for any **FDA** disk array administration access
- **ddn_admin.conf** configured for any **DDN** disk array administration access
- **xyr_admin.conf** configured for any **OPTIMA 1250** disk array administration access
- **dgc_admin.conf** configured for any **EMC/Clariion (DGC)** disk array administration access

Also save the storage configuration models (if any) used to configure the disk arrays. Their location will have been defined by the user.

### 3.0.4 Saving the Lustre File Systems

The following files are used by the Lustre system administration framework. It is strongly recommended that these files are saved onto a non-formattable media (from the Management Node):

- Configuration files: **/etc/lustre** directory
- File system configuration models (user defined location; by default **/etc/lustre/models**)
- **LDAP** directory if the High-Availability capability is enabled: **/var/lib/ldap/lustre** directory.

### 3.0.5 Saving the SLURM Configuration

The**/etc/slurm/slurm.conf** file is used by the **SLURM** resource manager. It is strongly recommended that this file is saved from the Management Node onto a non-formattable media.

# 3.1 STEP 1: Installing Red Hat Enterprise Linux Software on the Management Node

This step describes how to install the Red Hat Enterprise Linux software on the Management Node(s). It includes the following sub-tasks:
1) RAID configuration – optional.
2) Installation of the Red Hat Enterprise Linux 5 Server software
3) First boot settings
4) Configuring the Network
5) Installing an external Storage System (small clusters only)

## 3.1.1 Configure Internal RAID discs for BAS5 for Xeon clusters - optional

### 3.1.1.1 Configure RAID for AOC-USASLP-S8iR Adapters

This kind of adapter is installed on **NovaScale R423** and **NovaScale R425** machines only. Each machine has to be configured individually.

**See** Appendix G - *Configuring AOC-USASLP-S8iR RAID Adapters for NovaScale R423 and R425 machines* in this manual for details on how to configure these adapters.

### 3.1.1.2 Configure RAID for the LSI 1064 chip

**Note** This kind of adapter is only installed on **NovaScale R421-E1** machines only.

## 3.1.2 Red Hat Enterprise Linux 5 Installation

### 3.1.2.1 Initial Steps

**Before starting the installation read all the procedure details carefully.**

Start with the following operations:

1. Power up the machine.
2. Switch on the monitor.

3. Insert the **Red Hat Enterprise Linux Server 5 DVD** into the slot-loading drive.

**Note** The media must be inserted during the initial phases of the internal tests (whilst the screen is displaying either the logo or the diagnostic messages); otherwise the system may not detect the device.

4. Select all the options required for the language, time, date and keyboard system settings.

5. Skip the media test.

### 3.1.3 Red Hat Linux Management Node Installation Procedure

A suite of screens helps you to install RHEL5 software on the Service Node that includes the Management Node Services.


Figure 3-1. The Welcome Screen

1. The Welcome screen will appear at the beginning of the installation process.

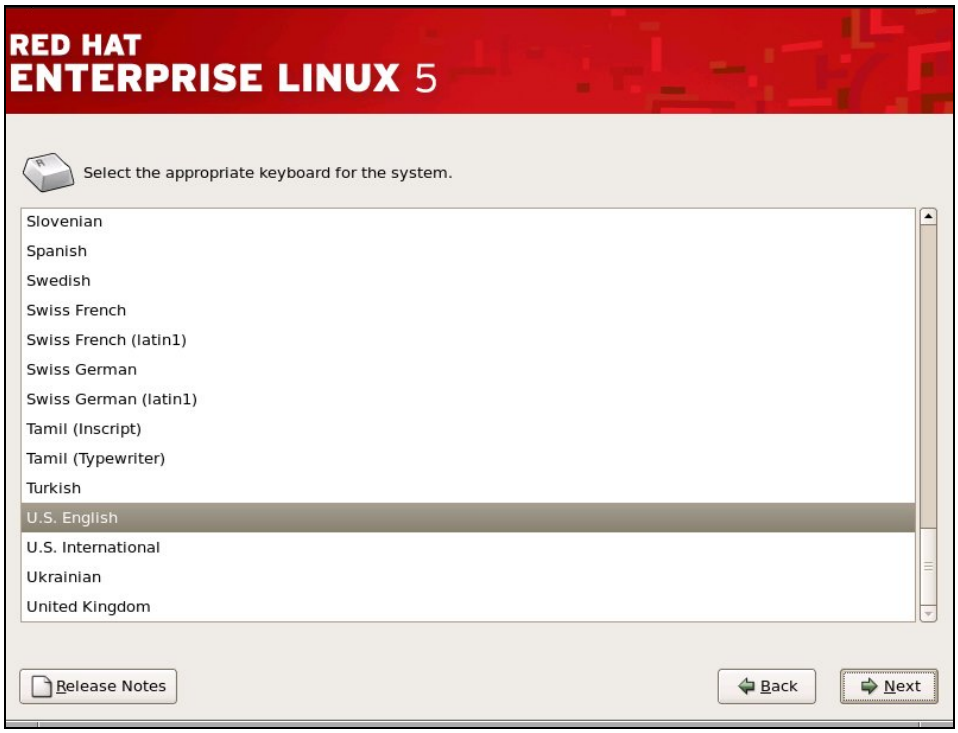


Figure 3-2. Keyboard installation screen

2. Select the language to be used for installation. Click on the **Next** button. Select the keyboard that is used for your system. Click on the **Next** button.

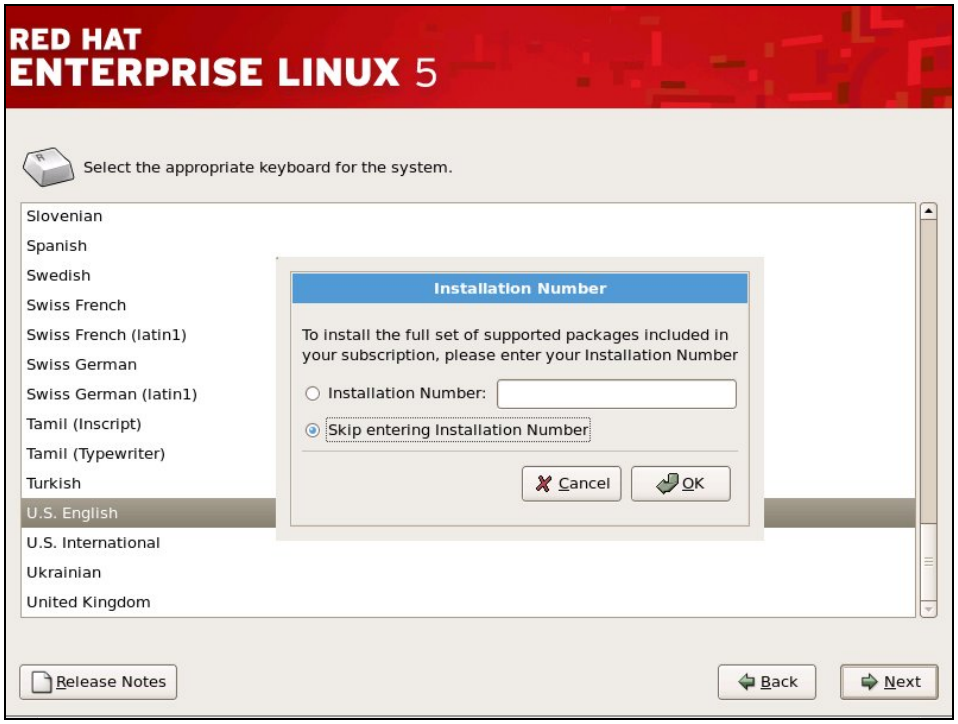


Figure 3-3. RHEL5 installation number dialog box


Figure 3-4. Skip screen for the installation number

3. The **BAS5 for Xeon** installation procedure requires that the **Red Hat** Installation Number is NOT entered now. The Installation Number can be entered later so that you can benefit from the **Red Hat** support network. Select **Skip entering Installation Number**. You will also have to click on **Skip**, as shown in Figure 3.4. Click on **Next**.

---

**Important**

See *Appendix I Activating your Red Hat account* - for important information regarding the use of installation numbers.

---


Figure 3-5. First RHEL5 installation screen

4. Select the option **Install Red Hat Enterprise Linux Server** as shown in Figure 3-5.

**Important**

**The Upgrade an existing installation option is not described in this manual. Contact Bull technical support for more information.**

**Note** For new clusters which are installing **BAS5 for Xeon** for the first time the **Upgrade an existing installation** option will not be in place.

### 3.1.4 Disk partitioning

There are different disk partitioning options available according to whether you are installing for the first time and using the default partitioning provided by LVM OR are carrying out a reinstallation and wish to use the partitioning that already exists.

### 3.1.4.1 Default partitioning

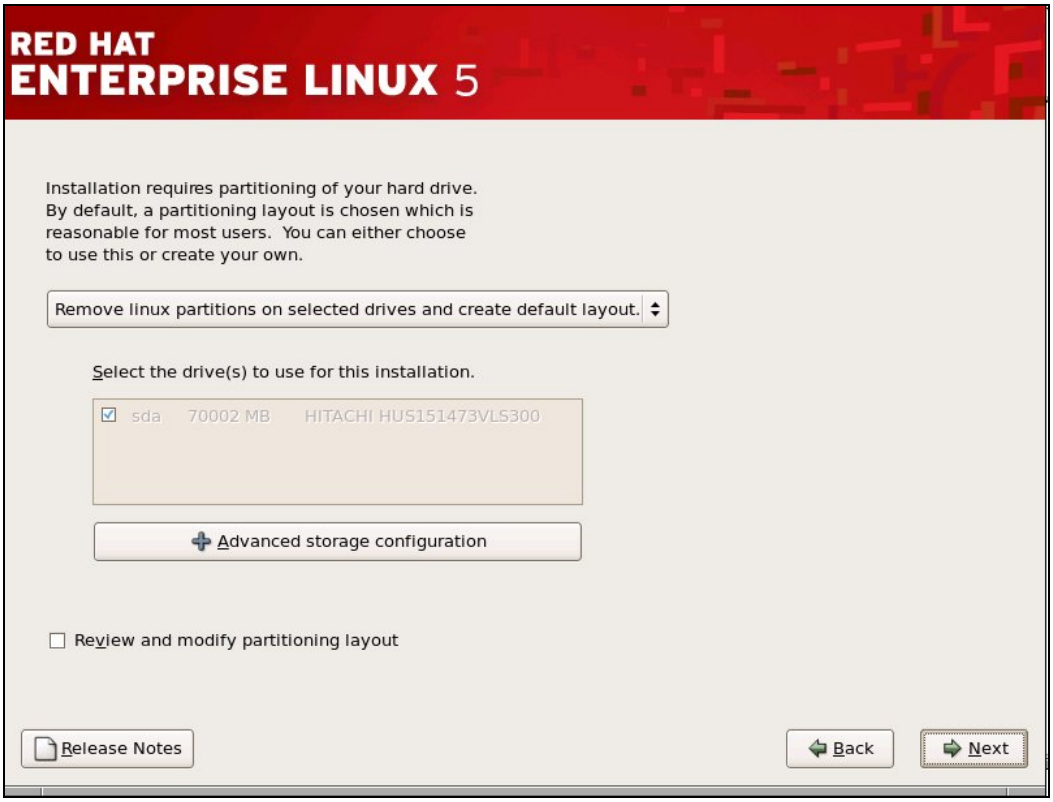


Figure 3-6. Partitioning screen

5. The default disk partitioning screen will appear as shown above. Usually, all the default options can be left as shown above, as the partitioning will be handled automatically by Logical Volume Manager (LVM). Click on **Next**.

**Note** If there is more than one disk for the Management Node, they will all appear checked in the drive list in Figure 3-6 and will be reformatted and the Red Hat software installed on them. Deselect those disks where your wish to preserve the existing data.


Figure 3-7. Confirmation of the removal of any existing partitions

Select **Yes** to confirm the removal of any existing partitions as shown in Figure 3.7, if this screen appears.

If the default partitioning is to be left in place go to section 3.1.5 *Network access Configuration*.

### 3.1.4.2 Reinstallation using the existing partition layout

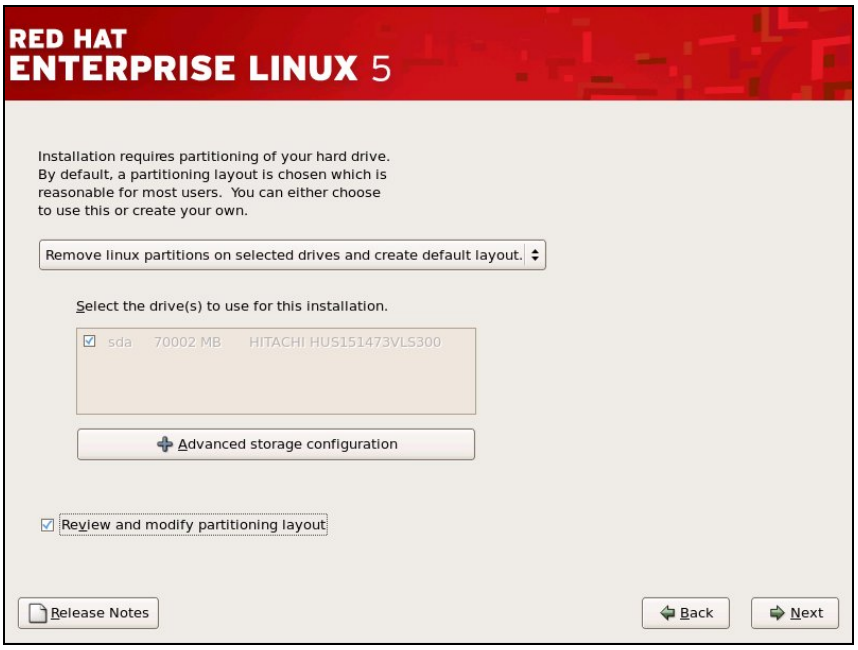


Figure 3-8. Modifying the partitioning layout – 1st screen

a. Tick the **Review and modify partitioning layout** box, as shown above.


Figure 3-9. Confirmation to remove existing partitions

b. Click **Yes**, above, to confirm the removal of all existing Linux partitions.

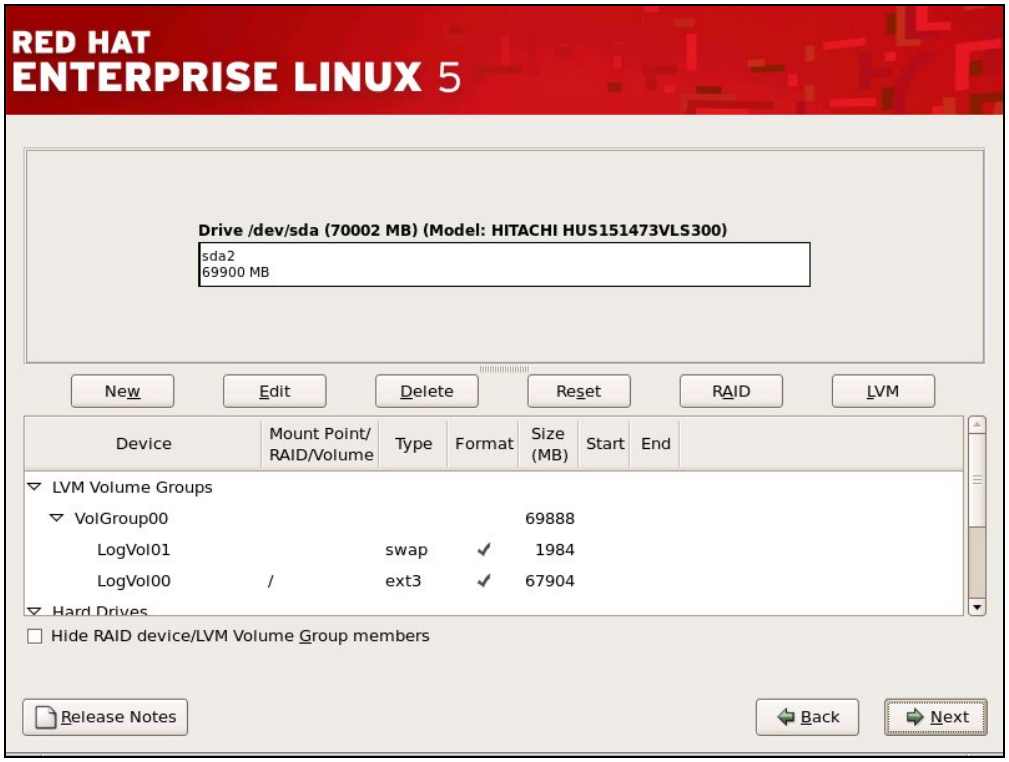


Figure 3-10. RHEL5 Partitioning options screen

c. If you wish to keep the partitioning options as they were previously, click on **Reset** in the screen above, as shown in Figure 3-9, and confirm the settings, including the mount point, that appear.


Figure 3-11. Confirmation of previous partitioning settings

## 3.1.5 Network access Configuration

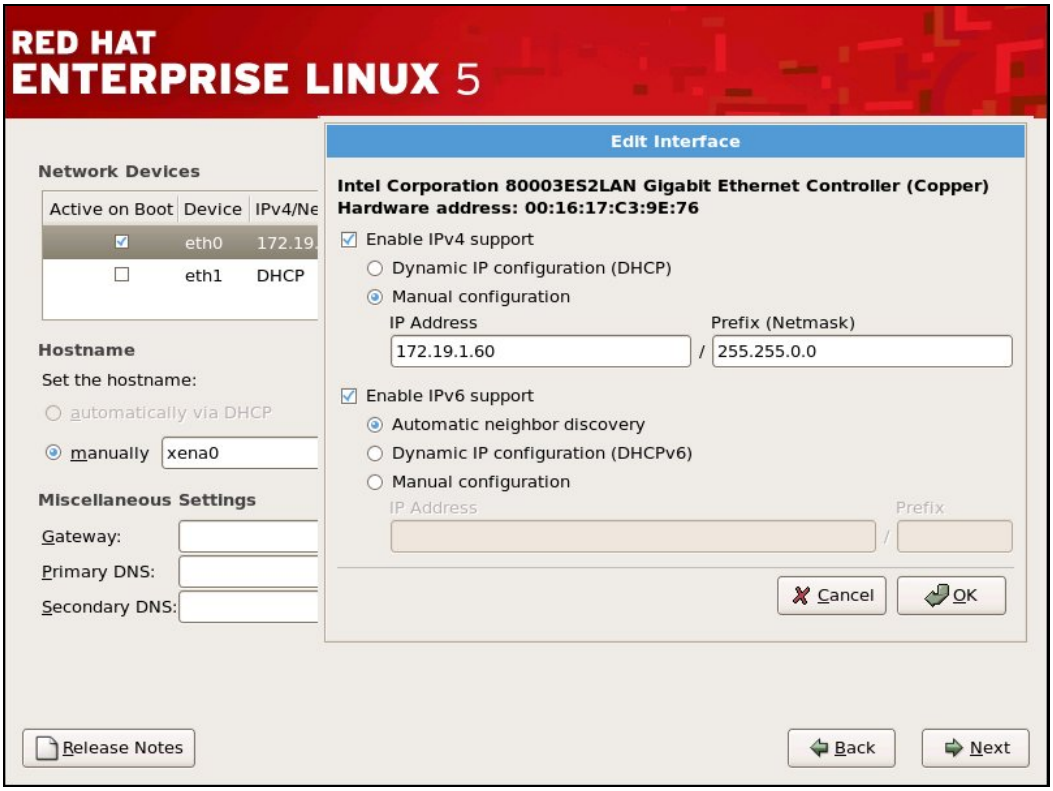


Figure 3-12. Network Configuration Screen

6. The next step is used to configure network access for the Management Node. Click on **manually** and enter the hostname (this is shown as xena0 in the example above). Select the device connected to the cluster management network (normally this is eth0) and click on the **Edit** button. Enter the IP address and NetMask configuration settings– see Figure 3-11.

   The miscellaneous settings for the Gateway, Primary DNS and Secondary DNS can be configured, if necessary. Warning messages may appear if this is not done and can be ignored.

   Click on the **OK** and **Next** buttons in Figure 3-11 when all the network configurations have been set.

**Note** The host name in the screen grab must be replaced by the name of the Management Node. The IP addresses in the screen above are examples and will vary according to the cluster.

## 3.1.6 Time Zone Selection and Root Password


Figure 3-13. Time Zone selection screen.

7. Select the Time Zone settings required, as shown in Figure 3-12, and click on Next.

**Note** Bull recommends using UTC, check the **System clock uses UTC** box to do this.

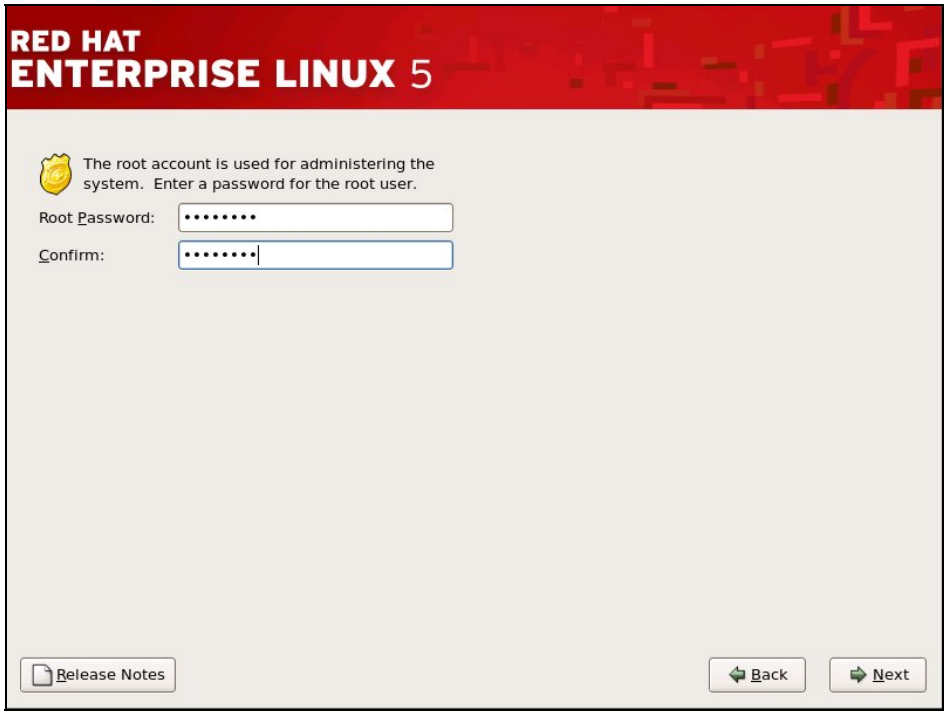


Figure 3-14. Root Password Screen

8. Set the Root password as shown in Figure 3-13. This must use a minimum of 6 characters.

## 3.1.7 Red Hat Enterprise Linux 5 Package Installation


Figure 3-15. Software selection screen

9. Leave the screen with the additional tasks deselected, as shown in Figure 3-14. Click on **Next**.


Figure 3-16. Installation screen

10. Click on **Next** in Figure 3-15 to begin the installation of **Red Hat Enterprise Linux Server**.

11. When the **Congratulations the installation is complete** screen appears carry out the procedure below to avoid problems later (There may be problems with the graphic display: the bottom part of the screen does not appear on some machines).

    a. Hold down the **Ctrl Alt F2** keys to go to the shell prompt for console 2.

    b. Save the **xorg.conf** file by using the commands below:

```
cd /mnt/sysimage/etc/X11
cp -p xorg.conf xorg.conf.orig
```

    c. Edit the **xorg.conf** file by using the command below:

```
vi /mnt/sysimage/etc/X11/xorg.conf
```

    d. Go to the Screen section, subsection Display and after the Depth 24 line add the following line.

```
                Modes    "1024x768" "832x624"
```

    e. Save the file and exit vi.

    f. Confirm that the modifications have been registered by running the command:

```
diff xorg.conf.orig xorg.conf
```

This will give output similar to that below:

```
27a28
>                Modes    "1024x768" "832x624"
```

g. Check the screen appearance is OK by holding down the **Ctrl Alt F6** keys.

h. Click on the **Reboot** button.

**Note** The screen resolution can be changed if there are any display problems by holding down **Ctrl Alt -** or **Ctrl Alt +** on the keyboard.

## 3.1.8 First boot settings

12. After the system has rebooted the Administrator must configure the list of post boot settings which appear. In particular the follow settings MUST be made:
    - Disable the firewall
    - Disable SELinux
    - Enable Kdump and select 128 MBs of memory for the kernel dump
13. The time and date must be set.
14. Select **Register later** for the software update.
15. The option **Create the Linux user** appears and can be set if required.
16. Ignore the No sound card screen which appears.
17. Ignore the Additional CDs screen
18. Click on **Finish.**
19. Click on **Reboot.**

## 3.1.9 Network Configurations

**Note** The IP addresses used will depend on the address plan for the system. Those used in this section are examples.

To configure the network use the **system-config-network** command, as below, this will launch the graphical tool used for the configuration.

```
system-config-network
```

### 3.1.9.1 Administration Network Configuration

**Note** The section only applies for those devices which have not been configured earlier, or if you wish to change an existing address.

Configure other network interfaces, e.g. **eth1**, **eth2** if required.

### Example

1. In the **Devices** panel select device eth1.
2. Click **Edit**.
3. Select **Activate device** when computer starts.
4. Select Statically set IP addresses and set the following values, according to your cluster type:

| | |
|---|---|
| IP ADDRESS | XXX.YYY.0.1 |
| SUBNETMASK | 255.255.0.0 |
| DEFAULT GATEWAY | none |

**The address settings used for the IP addresses must match the addresses declared in the Management Database (ClusterDB). If these are not known please contact Bull technical support. The IP addresses given in this section are examples and are for information only.**

**Note** Bull **BAS5 for Xeon** clusters do not support VLAN.

## 3.1.9.2 Alias Creation on eth0 (Management Node)

Aliases provide hardware independent IP addresses for cluster management purposes. The alias created below is used by the administration software – see section 3.5.

1. Go to the **/etc/sysconfig/network-scripts/** directory.
2. Copy the **ifcfg-eth0** file to the **ifcfg-eth0:0** file.
3. Edit the **ifcfg-eth0:0** file and modify the **DEVICE** setting so that it reads eth0:0 as shown.

```
DEVICE=eth0:0
```

4. Modify **IPADDR** with the alias IP address.

## 3.1.9.3 Restarting the network service

Run the command:

```
service network restart
```

### 3.1.10 External Storage System Installation

The Management Node may be connected to an external storage system, when the I/O and Login functions are included in the same Service Node as the Management functions.

**See** Chapter 4 *Configuring Storage Management Services,* in this manual, for more information regarding the installation, and also refer to the documentation provided with the storage system for details on how to install the storage system.

## 3.2 STEP 2: Installing BAS5 for Xeon software on the Management Node

This step describes how to install the Bull **BAS5 for Xeon v1.2** software on the Management Node(s). It includes the following sub-tasks:
1. Preparation for the Installation of the **Red Hat** software on other cluster nodes
2. Preparation for the Installation of the **BAS5 for Xeon v1.2 XHPC** software
3. Preparation for the Installation of the **BAS5 for Xeon v1.2** optional software
4. Installation of Bull **BAS5 for Xeon v1.2** software
5. Configuration of the Database

To identify the CD-ROM mount points, look at /etc/fstab file:

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

**Note** The examples in this section assume that **/media/cdrecorder** is the mount point for the CD-ROM.

During the installation procedure for **Red Hat Enterprise Linux Server 5** some software packages are loaded that are specifically required for Bull **BAS5 for Xeon** clusters. The following section describes the installation of these packages along with the Bull **XHPC**, and optional **InfiniBand**, **XLustre** and **XToolkit** software.

### 3.2.1 Preparing the Installation of the Red Hat software

1. Create the directory for the software:

```
mkdir –p /release/RHEL5.1
```

2. Create a mount point for the **RHEL5.1 DVD** by running the command below:

```
mkdir –p /media/cdrecorder/
```

3. Insert the RHEL5.1 DVD into the DVD reader and mount it:

```
mount /dev/cdrom /media/cdrecorder/
```

4. Copy the RHEL5.1 files to the /release/RHEL5.1 directory:

```
cp -a /media/cdrecorder/* /media/cdrecorder/.discinfo /release/RHEL5.1
```

**Note** This step will take approximately 7 minutes.

5. Eject the DVD:

```
umount /dev/cdrom
```

or use the eject command:

```
eject
```

6. If the **RHEL5.1-Supplementary-for-EM64T CDROM** is part of your delivery, carry out steps 7 to 11, below.

**The Java virtual machine rpm on the RHEL5.1-Supplementary-for-EM64T CDROM has to be installed later on clusters that use the hpcviewer tool included in HPC Toolkit.**

7. Create the directory:

```
mkdir –p /release/RHEL5.1-Supplementary
```

8. Insert the **RHEL5.1-Supplementary-for-EM64T CDROM** into the CD reader and mount it:

```
mount /dev/cdrom /media/cdrecorder/
```

9. Copy the **RHEL5.1** supplementary files into the **/release/RHEL5.1-Supplementary** directory:

```
cp -a /media/cdrecorder/* /release/RHEL5.1-Supplementary/
```

10. Eject the DVD:

```
umount /dev/cdrom
```

or use the eject command:

```
eject
```

## 3.2.2 Preparing the Installation of the BAS5 for Xeon XHPC software

1. Create the directory for the **BAS5 for Xeon v1.2 XHPC** software:

```
mkdir –p /release/XBAS5V1.2
```

2. Insert the **BAS5 for Xeon v1.2 XHPC** DVD-ROM into the DVD reader and mount it:

```
mount /dev/cdrom /media/cdrecorder/
```

3. Copy the **BAS5 for Xeon v1.2 XHPC** DVD-ROM contents into the /release directory:

```
cp -a /media/cdrecorder/* /release/XBAS5V1.2/
```

4. Eject the **XHPC** DVD-ROM.

### 3.2.3 Preparing the Installation of the BAS5 for Xeon optional software

According to cluster type and the software options purchased, the preparation of the installation of the Bull **XIB** software and/or the **XLustre** software will now need to be done.

The **/release/XBAS5V1.2** directory already created for the **XHPC** software will be used, so the only thing to do is copy the **XIB** and **XLustre** software across as follows:

#### Preparation for XIB software installation

1. Insert the **BAS5 for Xeonv1.2 XIB** DVD-ROM into the DVD reader and mount it:

```
mount /dev/cdrom /media/cdrecorder/
```

2. Copy the **BAS5 for Xeon v1.2 XIB** DVD-ROM contents into the **/release** directory, as shown below:

```
unalias cp
cp -a /media/cdrecorder/* /release/XBAS5V1.2
```

**Note** If the **unalias cp** command has already been executed, the message that appears below can be ignored:
`-bash: unalias: cp: not found`

3. Eject the **XIB** DVD-ROM.

#### Preparation for XLustre software installation

1. Insert the **BAS5 for Xeon v1.2 XLUSTRE** DVD-ROM into the DVD reader and mount it:

```
mount /dev/cdrom /media/cdrecorder/
```

2. Copy the **BAS5 for Xeon v1.2 XLUSTRE** DVD-ROM contents into the **/release** directory:

```
unalias cp
cp -a /media/cdrecorder/* /release/XBAS5V1.2
```

**Note** If the **unalias cp** command has already been executed, the message that appears below can be ignored:
`-bash: unalias: cp: not found`

3. Eject the **XLustre** DVD-ROM

### 3.2.4 Installing the Bull BAS5 for Xeon software

**Important**

**The mandatory RHEL packages and general BAS5 for Xeon products will be installed automatically by default.**

Go to the **/release/XBAS5V1.2** directory:

```
cd /release/XBAS5V1.2
```

The software installation commands for the Management Node correspond to the Function/Product combination applicable to the Service Node which includes the Management Node.

**See** *Chapter 1* for a description of the different architectures and functions possible.

The **BAS5 for Xeon** install command syntax is shown below.

```
./install -func MNGT [IO] [LOGIN] [-prod [XIB] [XLUSTRE] [XTOOLKIT]]
```

The **–func** option is used to specify the node function(s) to be installed and can be a combination of the following:

- **MNGT** for management functions
- **IO** for **IO/NFS** functions
- **LOGIN** for login functions

Different combinations of products can be installed using the –prod flag. The –prod options include the following:

- **XIB** to install the **BAS5 for Xeon InfiniBand** software (This needs to be purchased separately)
- **XLUSTRE** to install the **BAS5 for Xeon Lustre** software (This needs to be purchased separately)
- **XTOOLKIT** to install the **BAS5 for Xeon HPC** Toolkit software

For example, use the command below to install the MNGT, IO, LOGIN functions with the InfiniBand software:

```
./install -func MNGT IO LOGIN -prod XIB
```

The install script installs the software which has been copied previously into the **/release** directory on the NFS server.

### hpcviewer for HPC Toolkit

If **HPC Toolkit** has been installed and you wish to use the **hpcviewer** tool on the Management Node carry out the following procedure:

The Java virtual machine rpm included on the **RHEL5.1-Supplementary-for-EM64T CDROM** must be installed so that the **hpcviewer** tool included in **HPC Toolkit** can function. This is done as follows:

1. Go to the **/release/RHEL5.1-Supplementary** directory:

```
cd /release/RHEL5.1-Supplementary/
```

2. Manually install the public key for the verification of the Java virtual machine RPM by using the command below:

```
rpm --import ./RPM-GPG-KEY-redhat-release
```

3. Install the Java virtual machine by running a command similar to the one below:

```
yum install <java_virtual_machine_version>
```

For example:

```
yum install java-1.5.0-bea-1.5.0.08-1jpp.5.el5
```

**See** The Bull **BAS5 for Xeon** *Application Tuning Guide* for details on configuring and using HPC Toolkit.

## 3.2.5 Database Configuration

Please go to the section, below, that corresponds to your installation and follow the instructions carefully:

- *First Installation - Initialize the Cluster Database*
- *Re-installation of BAS5 for Xeon v1.2 with ClusterDB Preservation*

### 3.2.5.1 First Installation - Initialize the Cluster Database

**Note** This paragraph applies only when performing the first installation of **BAS5 for Xeon v1.2** and the cluster has been delivered with no Cluster DB preloaded by Bull. Contact Bull Technical Support to obtain the Cluster DB preload file.

1. Run the following commands (the IP addresses and netmasks below have to be modified according to your system):

```
su - postgres
cd /usr/lib/clustmngt/clusterdb/install
loadClusterdb --basename <clustername> --adnw xxx.xxx.0.0/255.255.0.0
--bknw xxx.xxx.0.0/255.255.0.0 --bkgw <ip_gateway> --bkdom
<domain_name>
--icnw xxx.xxx.0.0./255.255.0.0
--preload <load_file>
```

Where:

**basename** (mandatory) designates both the node base name, the cluster name and the virtual node name

**adnw** (mandatory) is administrative network

**bknw** (option) is backbone network

**bkgw** (option) is backbone gateway

**bkdom** (option) is backbone domain

**icnw** (option) is ip over interconnect network

**Note** See the **loadClusterdb** man page and the preload file for details of the options which apply to your system.

Preload sample files are available in:
**/usr/lib/clustmngt/clusterdb/install/preload_xxxx.sql**
(xxxx in the path above corresponds to your cluster).

2. Save the database:

```
pg_dump -Fc -C –f /var/lib/pgsql/backups/clusterdb.dmp clusterdb
```

### 3.2.5.2 Re-installation of BAS5 for Xeon v1.2 with ClusterDB Preservation

**Note** This paragraph applies when re-installing an existing version of **BAS5 for Xeon v1.2** with the restoration of the existing Cluster Database.

1. Run the commands:

```
su - postgres
psql -U clusterdb clusterdb

<Enter Password>
clusterdb=> truncate config_candidate;truncate config_status;\q
TRUNCATE TABLE
TRUNCATE TABLE
```

2. Restore the Cluster DB files which have been stored under **/var/lib/pgsql/backups**:

```
pg_restore -Fc --disable-triggers -d clusterdb
   /var/lib/pgsql/backups/<name_of_ClusterDB_saved_file>
```

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

**See** Section *3.0.1 Saving the ClusterDB* for details of the Cluster database files that have been saved. See the **BAS5 for Xeon** *Administrator's Guide* for more details about restoring data.

3. Go back to root by running the **exit** command.

## 3.3 STEP 3: Configuring Equipment and Installing Utilities on the Management Node

This step describes how to:

- Configure equipment
- Configure Ethernet switches
- Install and configure **Ganglia**, **Syslog-ng**, **NTP**, **Postfix**, **Kdump**, **SLURM** and **PBS Pro**
- Install compilers (only on Management Nodes which include Login functionality)
- Configure the User environment and in particular **MPI**

**If your cluster has been delivered with the ClusterDB preload in place or if you have saved your cluster database from a previous installation go to the section *Configuring Management Tools Using Database Information*.**

### 3.3.1 Generate the SSH keys

1. Change to the root directory on the Management Node:

```
cd /root
```

2. Enter the following commands:

```
ssh-keygen -t rsa
```

Accept the default choices and do not enter a pass-phrase.

```
cat .ssh/id_rsa.pub >> .ssh/authorized_keys
```

3. Test the configuration:

```
ssh localhost uname
```

```
The authenticity of host 'localhost (127.0.0.1)' can't be established.
RSA key fingerprint is
91:7e:8b:84:18:9c:93:92:42:32:4a:d2:f9:38:e9:fc.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost,127.0.0.1' (RSA) to the list of
known hosts.
   Linux
```

Then enter:

```
ssh <clustername>0 uname
```

```
Linux
```

## 3.3.2 Configuring Equipment

**Only carry out this task during the very first installation.**

The purpose of this part is to collect the MAC address for each node in the cluster and to configure the hardware manager (often called the **BMC**) for these nodes.

Look for the **MAC** address files in the **/usr/lib/clustmngt/clusterdb/install/** directory. These files will have been provided by manufacturing and are named **Type_Rack+Xan_Rack.final**. The format of a MAC address file is as follows:

```
<rack_level> <level_slot> <mac addr of node> <mac addr of bmc> <ip addr of bmc> <comment>
```

For each **MAC** address file:

- Identify the **rack_label** from **rack** table of the ClusterDB which corresponds to the file. For example <**Type_Rack+Xan.final**> might be **SNRXA2.final**, where '**Type_Rack**' is **SNR**, '**a**' (the **x_coord** of rack) is **A** and '**n**' (the **y_coord** of rack) is **2**. Execute the command below as the **postgres** user in order to retrieve the **rack_label**:

```
$ psql -c "select label from rack where x_coord='A' and y_coord='2'"clusterdb
label
```

```
RACK1
(1 row)
```

- Update the database with the node and hardware manager **MAC** addresses for the rack by running the command below as the **postgres** user:

```
$ /usr/lib/clustmngt/clusterdb/install/updateMacAdmin
<Type_Rack+Xan.final> --rack <rack label>
```

### Example

```
$ /usr/lib/clustmngt/clusterdb/install/updateMacAdmin SNRXA2.final
--rack RACK1
```

- Configure the IP addresses for the **BMC**s of the rack by running the command below as the root user:

```
# /usr/lib/clustmngt/BMC/bmcConfig --input <Type_Rack+Xan.final>
```

### Example

```
# /usr/lib/clustmngt/BMC/bmcConfig --input SNRXA2.final
```

### 3.3.3 Configuring Equipment Manually

If a node has been installed and the MAC address files have not been found you can collect the MAC addresses of the admin Ethernet cards for each node as follows:

– Start the **DHCPD** service by running the command:

```
dbmConfig configure --service sysdhcpd
```

– Configure the nodes so that they boot on the network.

– Reboot the equipment individually and collect their MAC addresses in the **/var/log/**messages file.

Create the file which contains the MAC addresses, IP addresses and cluster elements. Its format is as follows:

```
<type> <name> <mac address>
```

An example, similar to that below, is available from:

**/usr/lib/clustmngt/clusterdb/install/mac_file.exp**

```
node valid0 00:04:23:B1:DF:AA
node valid1 00:04:23:B1:DE:1C
node valid2 00:04:23:B1:E4:54
node valid3 00:04:23:B1:DF:EC
```

1. Run the command:

```
su - postgres
```

2. Run the command:

```
cd /usr/lib/clustmngt/clusterdb/install
```

3. Run the following command to collect the domain name of each node of the cluster and load the MAC addresses for the network cards for the administration network:

```
updateMacAdmin <file>
```

`<file>` is the name of a file that must have been created previously – see point 2. The full path must be included so that it can be easily retrieved, for example **updateMacAdmin /root/cluster-mac-address**.

4. Go back to root by running the exit command.

### 3.3.4 Configuring Ethernet Switches

**Only carry out this task during the first installation or if new Ethernet switches have been added to the cluster. The Ethernet switches should be as initially set (factory settings).**

Install Ethernet switches by running the command, as root:

```
# swtAdmin auto
```

**See** Chapter 9 - *Configuring Switches and Cards* in this manual for more details.

### 3.3.5 Configuring Postfix

1. Edit the **/etc/postfix/main.cf** file.
2. Uncomment or create or update the line that contains myhostname
   `myhostname = <adminnode>.<admindomain>`

   You must specify a domain name.

   **Example**
   `myhostname = node0.cluster`
3. This step ONLY applies to configurations which use **CRM** (Customer Relationship Management); for these configurations the Management Node is used as Mail Server, and this requires that Cyrus is configured.
   Uncomment the line:

   `mailbox_transport = cyrus`
4. Start the postfix service:

```
# service postfix start
```

### 3.3.6 Configuring Management Tools Using Database Information

1. Run the following commands and check to see if any errors are reported. These must be corrected before continuing.

```
dbmCluster check –-ipaddr
dbmCluster check --rack
```

2. Configure the tools with the following command, as root:

```
# dbmConfig configure --restart --force
```

An output example for this command is below:

```
Wed Jul 30 09:09:06 2008 NOTICE: Begin synchro for syshosts
Wed Jul 30 09:09:06 2008 NOTICE: End synchro for syshosts
Wed Jul 30 09:09:06 2008 NOTICE: Begin synchro for sysdhcpd
Shutting down dhcpd: [  OK  ]
Starting dhcpd: [  OK  ]
Wed Jul 30 09:09:07 2008 NOTICE: End synchro for sysdhcpd
Wed Jul 30 09:09:07 2008 NOTICE: Begin synchro for group
INSERT group ALL                [ OK ]  (xena[1-18,30-33,140-141])
INSERT group IO                 [ OK ]  (xena[1-2,11,17-18,140-141])
INSERT group COMP               [ OK ] xena[3-8,14]
INSERT group META               [ OK ]  (xena[10,12])
INSERT group NODES8GB           [ OK ]  (xena[0-18,30-33,140-141])
INSERT group ADMIN              [ OK ]  (xena0)
Wed Jul 30 09:09:08 2008 NOTICE: End synchro for group
Wed Jul 30 09:09:08 2008 NOTICE: Begin synchro for pdsh
Wed Jul 30 09:09:08 2008 NOTICE: End synchro for pdsh
Wed Jul 30 09:09:08 2008 NOTICE: Begin synchro for conman
Stopping ConMan: conmand[  OK  ]
Starting ConMan: conmand[  OK  ]
Wed Jul 30 09:09:08 2008 NOTICE: End synchro for conman
Wed Jul 30 09:09:08 2008 NOTICE: Begin synchro for snmptt
Wed Jul 30 09:09:08 2008 NOTICE: End synchro for snmptt
Wed Jul 30 09:09:08 2008 NOTICE: Begin synchro for nagios
INITIALIZATION of the services
Running configuration check...done
Resetting host status in DB, update by Nagios will take a few minutes
Stopping NovaScale Master nagios ...[  OK  ]
Starting NovaScale Master nagios ...Resetting host status in DB,
update by Nagios will take a few minutes
[  OK  ]syslog-ng (pid 2998) is running...
Reloading syslog-ng: [  OK  ]

syslog-ng (pid 2998) is running...
Reloading syslog-ng: [  OK  ]
Wed Jul 30 09:09:10 2008 NOTICE: End synchro for nagios
Wed Jul 30 09:09:10 2008 NOTICE: Begin synchro for nsm

Wed Jul 30 09:09:10 2008 NOTICE: End synchro for nsm
```

3. Switch to **postgres**:

```
su - postgres
```

4. Save the ClusterDB:

```
pg_dump -Fp -C -f /var/lib/pgsql/backups/clusterdball-
<name_of_clusterdbdata.sav>.dmp clusterdb
```

5. Go back to root by running the **exit** command.

6. Reboot the Management Node:

```
exit
reboot
```

## 3.3.7 Configuring Ganglia

1. Copy the file:
   **/usr/share/doc/ganglia-gmond-3.0.5/templates/gmond.conf**
   into /etc.

2. Edit the **/etc/gmond.conf** file:
   - In line 9, replace "`deaf = yes`" with "`deaf = no`".
   - In line 18, replace xxxxx with the basename of the cluster.
     ```
     name = "xxxxx"  /* replace with your cluster name */
     ```
   - In line 24 replace x.x.x.x with the alias IP address of the Management Node.
     ```
     host = x.x.x.x  /* replace with your administration node ip
     address */
     ```

3. Start the **gmond** service:

```
service gmond start
chkconfig --level 235 gmond on
```

4. Edit the **/etc/gmetad.conf** file:

In Line 39, replace
"`data_source "mycluster" localhost`" with `data_source "basename" localhost`

Example `data_source "nova" localhost`

5. Start **gmetad**:

```
service gmetad start
chkconfig --level 235 gmetad on
```

## 3.3.8 Configuring Syslog-ng

### Syslog Ports Usage

**584 / udp** This port is used by cluster nodes to transmit I/O status information to the Management Node. It is intentionally chosen as a non standard port. This value must be consistent with the value defined in the **syslog-ng.conf** file on cluster nodes and this ensured by Bull tools. There is no need for action here.

### Modify the syslog-ng.conf file

Modify the **/etc/syslog-ng/syslog-ng.conf** file, as follows, adding the IP address (Ethernet **eth0** in the administration network) which the server will use for tracking.

1. 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 "<Management_Node_IP_address>"
```

2. Make the changes as explained in the messages (3 substitutions with the alias IP address).

### Restart syslog-ng

After modifying the configuration files, restart the **syslog-ng** service:

```
service syslog-ng restart
```

## 3.3.9 Configuring NTP

The Network Time Protocol (NTP) is used to synchronize the time of a computer client with another server or reference time source. This section does not cover time setting with an external time source, such as a radio or satellite receiver. It covers only time synchronization between the Management Node and other cluster nodes, the Management Node being the reference time source.

**Note** It is recommended that the System Administrator synchronizes the Management Node with an external time source.

Modify the **/etc/ntp.conf** file on the Management Node as follows.

The first two lines must be marked as comments:

```
#restrict default kod nomodify notrap nopeer noquery
#restrict -6 default kod nomodify notrap nopeer noquery
```

Leave the lines:

```
restrict 127.0.0.1
restrict -6 ::1
```

The next line should have the following syntax assuming that the parameters used are for the management network with an associated netmask:

```
restrict <mgt_network_IP_address> mask <mgt_network_mask nomodify notrap>
```

For example, if the IP address of the Management Node alias is 172.17.0.99:

```
restrict 172.17.0.0 mask 255.255.0.0 nomodify notrap
```

Put the following lines in as comments:

```
#server 0.rhel.pool.ntp.org
#server 1.rhel.pool.ntp.org
#server 2.rhel.pool.ntp.org
```

Leave the other command lines and parameters unmodified.

Restart **ntpd** service:

```
service ntpd restart
```

Start **ntptrace** with the IP address as the Management Node alias (x.x.0.99):

**Example**

```
ntptrace 172.17.0.99
```

```
ns0: stratum 11, offset 0.000000, synch distance 0.012515
```

## 3.3.10 Configuring the kdump kernel dump tool

**Kdump** will have been enabled during the Red Hat installation on the Management Node – see section 3.1.8

1. The following options must be set in the **/etc/kdump.conf** configuration file:
   a. The path and the device partition where the dump will be copied to should be identified by its **LABEL**, **/dev/sdx** or **UUID** label either in the **/home/** or **/** directories.

      **Examples**

      ```
      path /var/crash
      ext3 /dev/volgroup00/logvol00
      ```

   b. The tool to be used to capture the dump must be configured. Uncomment the `core_collector` line and add `-d 1`, as shown below:

      ```
      core_collector makedumpfile -c -d 1
      ```

      **-c** indicates the use of compression and **-d 1** indicates the dump level.

**Important**

It is essential to use non-stripped binary code within the kernel. Non-stripped binary code is included in the debuginfo RPM, **kernel-debuginfo-<kernel_release>.rpm**, available from: http://people.redhat.com/duffy/debuginfo/index-js.html

This package will install the kernel binary in the **/usr/lib/debug/lib/modules/<kernel_version>/** folder

**Note** The size for the dump device must be larger than the memory size if no compression is used.

Use the command below to launch **kdump** automatically when the system restarts:

```
chkconfig kdump on
```

## 3.3.11 Installing and Configuring SLURM - optional

**SLURM does not work with the PBS-Professional Batch manager and must only be installed on clusters which do not use PBS-Professional.**

The SLURM files are installed under the **/usr** and **/etc** directories.

**Note** This step applies to the Management Node only. The same configuration file will be copied later to the other nodes in the cluster – see STEP 5.

### 3.3.11.1 Install the SLURM RPMS

Run the command below to install the SLURM RPMs:

```
yum install slurm pam_slurm slurm-munge slurm-auth-none slurm-devel
```

**Note** **Munge** and **munge-libs** are included within the **slurm-munge** RPM and will not need to be installed separately.

**See** The *Software Release Bulletin* for **BAS5 for Xeon v1.2** for details on how to install **SLURM** version **1.0.15.** This version is required to ensure compatibility with the LSF Batch Manager

### 3.3.11.2 Create and Modify the SLURM configuration file

A SLURM configuration file must be created using the parameters that describe the cluster. The **/etc/slurm/slurm.conf.example** file can be used as a template to create the **/etc/slurm/slurm.conf** file for the cluster.

The **slurm.conf** file can be created manually from the template described above, **OR** the tool found at **/usr/share/doc/slurm-1.3.2/html/configurator.html** can be used to help define the necessary parameters. This tool is an HTML file that, when loaded into a browser (e.g. Firefox), will generate a **slurm.conf** file in text format using the parameter supplied by the user. The generated file can be saved, or cut/pasted into a text editor if the configuration details need to be modified.

Whether generated manually, or by the **configurator.html** tool, the **slurm.conf** file must contain the following information:

1. The name of the machine where the **SLURM** control functions will run. This will be the Management Node, and will be set as shown in the example below.

   ```
   ControlMachine=<basename>0
   ```

```
ControlAddr=<basename>0
```

2. The **SlurmUser** and the authentication method for the communications:

   ```
   SlurmUser=slurm
   ```
   AuthType=auth/munge (as shown in the example file)

   or

   ```
   AuthType=auth/none
   ```

3. The type of switch or interconnect used for application communications.

   ```
   SwitchType=switch/none  # used with Ethernet and InfiniBand
   ```

4. Any port numbers, paths for log information and **SLURM** state information. If they do not already exist, the path directories must be created on all of the nodes.

   ```
   SlurmctldPort=6817
   SlurmdPort=6818
   SlurmctldLogFile=/var/log/slurm/slurmctld.log
   SlurmdLogFile=/var/log/slurm/slurmd.log.%h
   StateSaveLocation=/var/log/slurm/log_slurmctld
   SlurmdSpoolDir=/var/log/slurm/log_slurmd/
   ```

5. Provide scheduling, resource requirements and process tracking details:

   ```
   SelectType=select/linear
   SchedulerType=sched/builtin   # default is sched/builtin
   ProctrackType=proctrack/pgid
   ```

6. Provide accounting requirements. The path directories must be created on all of the nodes, if they do not already exist. For Job completion:

   ```
   #JobCompType=jobcomp/filetxt # default is jobcomp/none
   #JobCompLoc=/var/log/slurm/slurm.job.log
   ```

   For accounting type for **SLURM v1.0.15** use:

   ```
   #JobAcctType=jobacct/linux   # default is jobacct/none
   #JobAcctLogFile=/var/log/slurm/slurm_acct.log
   ```

   For accounting type for **SLURM v1.3.2** use:

   ```
   #JobAcctGatherType=jobacct/linux   # default is jobacct/none
   #AccountingStorageLoc=/var/log/slurm/slurm_acct.log
   ```

   Uncomment the appropriate lines if job accounting is to be included.

7. Provide the paths to the job credential keys. The keys must be copied to all of the nodes.

   ```
   JobCredentialPrivateKey=/etc/slurm/private.key
   JobCredentialPublicCertificate=/etc/slurm/public.key
   ```

8. Provide Compute Node details:

```
NodeName=bali[10-37] Procs=8 State=UNKNOWN
```

9. Provide information about the partitions. **MaxTime** is the maximum wall-time limit for any job in minutes. The state of the partition may be UP or DOWN.

```
PartitionName=global Nodes=bali[10-37] State=UP Default=YES
PartitionName=test Nodes=bali[10-20] State=UP MaxTime=UNLIMITED
PartitionName=debug Nodes=bali[21-30] State=UP
```

10. In order that **Nagios** monitoring is enabled inside **NovaScale Master – HPC Edition**, the **SLURM** Event Handler mechanism has to be active. This means that the following line in the **SLURM.conf** file on the Management Node has to be uncommented, or added if it does not appear there.

```
SlurmEventHandler=/usr/lib/clustmngt/slurm/slurmevent
```

**Note** If the value of the **ReturnToService** parameter in the **slurm.conf** is set to 0, then when a node that is down is re-booted, the administrator will have to change the state of the node manually with a command similar to that below, so that the node appears as idle and available for use:

```
$ scontrol update NodeName=bass State=idle Reason=test
```

To avoid this, set the **ReturnToService** parameter to **1** in the **slurm.conf** file.

**See**

- The **slurm.conf** man page for more information on all the configuration parameters, including the **ReturnToService** parameter, and those referred to above.
- https://computing.llnl.gov/linux/slurm/documentation.html for an example of the **configurator.html** tool for SLURM version 1.3.2.and the parameters that it includes.

### slurm.conf file example

```
ControlMachine=bali0
ControlAddr=bali0
SlurmUser=slurm
SlurmUID=105
SlurmGroup=slurm
SlurmGID=105
SlurmHome=/home/slurm
AuthType=auth/munge
SlurmctldPort=6817
SlurmdPort=6818
SlurmctldLogFile=/var/log/slurm/slurmctld.log
SlurmdLogFile=/var/log/slurm/slurmd.log.%h
StateSaveLocation=/var/log/slurm/log_slurmctld
SlurmdSpoolDir=/var/log/slurm/log_slurmd/
SlurmctldDebug=3    # default is 3
SlurmdDebug=3       # default is 3
SelectType=select/linear
SchedulerType=sched/builtin   # default is sched/builtin
```

```
#JobCompType=jobcomp/filetxt  # default is jobcomp/none
#JobCompLoc=/var/log/slurm/slurm.job.log
SwitchType=switch/none
ProctrackType=proctrack/pgid

#valid below for SLURM v1.0.15
#JobAcctType=jobacct/linux  # default is jobacct/none
#JobAcctLogFile= /var/log/slurm/slurm_acct.log

# Valid below for SLURM v1.3.2
JobAcctGatherType=jobacct/linux # default is jobacct/none
AccountingStorageLoc=/var/log/slurm/slurm_acct.log
FastSchedule=1      # default is `1'
FirstJobid=1000     # default is `1'
ReturnToService=1   # default is `0'
MpiDefault=none     # default is "none"
SlurmEventHandler=/usr/lib/clustmngt/slurm/slurmevent

JobCredentialPrivateKey=/etc/slurm/private.key
JobCredentialPublicCertificate=/etc/slurm/public.key

# NODE CONFIGURATION
NodeName=bali[10-37] Procs=8 State=UNKNOWN

# PARTITION CONFIGURATION
PartitionName=global Nodes=bali[10-37] State=UP Default=YES
PartitionName=test Nodes=bali[10-20] State=UP MaxTime=UNLIMITED
PartitionName=debug Nodes=bali[21-30] State=UP
```

### 3.3.11.3 Final Configuration Steps

After the **SLURM** RPMs have been installed, and all the necessary parameters for the cluster have been defined in the **slurm.conf** file, a few steps still remain before the configuration of **SLURM** is complete on the Management Node. These steps can either be done using the **slurm_setup.sh** script - see section *3.3.11.4* OR manually - see section *3.3.11.5*.

### 3.3.11.4 Using the slurm_setup.sh Script

The **SLURM** setup script is found in **/etc/slurm/slurm_setup.sh** and is used to automate and customize the installation process. The script reads the **slurm.conf** file created previously and does the following:

1. Creates the **SlurmUser**, using the **SlurmUID**, **SlurmGroup**, **SlurmGID**, and **SlurmHome** optional parameter settings in the **slurm.conf** file to customize the user and group. It also propagates the identical Slurm User and Group settings to the reference nodes.

2. Validates the pathnames for log files, accounting files, scripts, and credential files. It then creates the appropriate directories and files, and sets the permissions. For user supplied scripts, it validates the path and warns if the files do not exist. The directories and files are replicated on both the Management Node and reference nodes.

3. Creates the **job credential** validation private and public keys on the Management and reference nodes.

4. If **auth/munge** is selected as the authorization type (AuthType) in the **slurm.conf** file, it validates the functioning of the **munge** daemon and copies the munge key file from the Management to the reference nodes.

5. Copies the **slurm.conf** file from the Management Node to the reference nodes.

**Note** The script is also used to configure the LOGIN and COMPUTE(X) Reference nodes, as described in **STEP 5**.

Skip the next section, which describes how to complete the configuration of SLURM manually, if the **slurm_setup.sh** script has been used.

### 3.3.11.5 Completing the Configuration of SLURM on the Management Node Manually

These manual steps must be carried out before **SLURM** is started on any of the cluster nodes

**Note** The files and directories used by **SLURMCTLD** must be readable or writable by the user **SlurmUser** (the SLURM configuration files must be readable; the log file directory and state save directory must be writable).

#### Create a SlurmUser

The **SlurmUser** must be created before **SLURM** is started. The **SlurmUser** will be referenced by the **slurmctld** daemon. Create a **SlurmUser** on the COMPUTE(X), Login/IO or LOGIN Reference nodes with the same **uid gid** (105 for instance):

```
groupadd -g 105 slurm
useradd -u 105 -g slurm slurm
mkdir -p /var/log/slurm
chmod 755 /var/log/slurm
```

The **gid** and **uid** numbers do not have to match the one indicated above, but they have to be the same on all the nodes in the cluster.

The user name in the example above is **slurm**, another name can be used, however it has to be the same on all the nodes in the cluster.

#### Configure the SLURM job credential keys as root

Unique job credential keys for each job should be created using the **openssl** program. These keys are used by the **slurmctld** daemon to construct a job credential, which is sent to the **srun** command and then forwarded to **slurmd** to initiate job steps.

**openssl must be used (not ssh-genkey) to construct these keys.**

When you are within the directory where the keys will reside, run the commands below:

```
cd /etc/slurm
openssl genrsa -out private.key 1024
openssl rsa -in private.key -pubout -out public.key
```

The **Private.Key** file must be readable by **SlurmUser** only. If this is not the case then use the commands below to change the setting.

```
chown slurm.slurm /etc/slurm/private.key
chmod 600 /etc/slurm/private.key
```

The **Public.Key** file must be readable by all users. If this is not the case then use the commands below to change the setting.

```
chown slurm.slurm /etc/slurm/public.key
chmod 644 /etc/slurm/public.key
```

#### 3.3.11.6 More Information

**See** The Bull **BAS5 for Xeon** *Administrator's Guide* for more information on SLURM (Munge configuration, security, the creation of job credential keys and the **slurm.conf** file). See **man slurm.conf** for more information on the parameters of the **slurm.conf** file, and **man slurm_setup.sh** for information on the **SLURM** setup script.

### 3.3.12

## 3.3.13 Installing and Configuring PBS Professional Batch Manager – optional

- **PBS Professional does not work with SLURM.**
- **The PBS license file (altair_lic.dat) must be available as a prerequisite**
- **The FLEXlm License Server has to be installed before PBS Professional is installed.**

**See** Chapter 4 in the PBS Professional *Administrator's Guide*, available on the **PBS Professional CD-ROM,** for more information on the installation and configuration routines for PBS Professional, described below.

### 3.3.13.1 Downloading, Installing and Starting the FLEXlm License Server

**This step applies to the Management Node (standard installation), or to a node which is dedicated as the Flexlm Server(s).**

**This section only applies to clusters which do NOT feature High Availability for the Management Node NOR redundancy for PBS Pro.**

**See** The **BAS5 for Xeon** *High Availability Guide* and the **PBS Professional** *Administrator's Guide*, available on the **PBS Professional CD-ROM,** if High Availability for the Management Node and High Availability (redundancy) for **PBS Pro** are in place.

1. Copy all tarballs and documentation from the **PBS Professional** CD ROM on to the Management Node.

2. Uncompress and extract the files, using the command below:

```
tar -xvzf altair_flexlm-<version>.<architecture>.tar
```

For example

```
tar -xvzf altair_flexlm-9.0.amd64_s8.tar
```

3. Run the command, below, to start the installation process.

```
./licsetup.sh
```

4. Respond to the questions as they appear identifying the location where the licensing package will be installed (**/opt** is recommended). This location is known as **<install loc>.**

5. Copy the license file, provided by Bull technical support, to the folder **<install loc>/altair/security/altair_lic.dat**.

6. Run the following commands to start the **FLEXlm** license server:

```
cd <install loc>/altair/security/
./altairlm.init.sh start
```

7. To install the license startup script, run the following command:

```
<install loc>/altair/security/install_altairlm.sh
```

### 3.3.13.2 Starting the installation of PBS Professional

The commands for the installation have to be carried out by the cluster Administrator logged on as root.

1. Extract the package from the **PBS Pro CD ROM** to the directory of choice on the Management Node, using a command similar to that below.

```
cd /root/PBS
tar -xvzf PBSPro_9.2.0-RHEL5_x86_64.tar.gz
```

2. Go to the installation directory on the Management Node and run:

```
cd PBSPro_9.2.0
```

3. Start the installation process:

```
./INSTALL
```

### 3.3.13.3 PBS Professional Installation Routine

During the **PBS Professional** installation routine, the Administrator will be asked to identify the following:

#### Execution directory

The directory into which the executable programs and libraries will be installed, for example, **/usr/pbs**

#### Home directory

The directory into which the **PBS Pro** daemon configuration files and log files will be installed, for example, **/var/spool/PBS**.

#### PBS installation type

The installation type depends on the type of node that PBS Professional is being installed on:

- On the Management Node: type 1

#### Do you want to continue?

Answer **Yes**.

#### License file location

In the example above this is **/opt/altair/security/altair_lic.dat**

#### Would you like to start?

When the Installation complete window appears, the installation program offers to start PBS Professional, enter 'n' for 'no'.

### 3.3.13.4 Initial configuration on the Management Node

**See**

- Chapter 4 in the PBS Professional *Installation and Upgrade Guide*, available on the **PBS Professional CD ROM** for more information on the installation for PBS Professional.
- Chapter 2 in the PBS Professional *Administrator's Guide* for more information on configuring PBS Professional.

**See the BAS5 for Xeon *High Availability Guide* for the pbs.conf file configuration details if High Availability is in place for PBS Professional.**

1. Modify the **/etc/pbs.conf** file as follows:

```
PBS_EXEC=/usr/pbs
PBS_HOME=/var/spool/PBS
PBS_START_SERVER=1
PBS_START_MOM=0
PBS_START_SCHED=1
PBS_SERVER=basename0
PBS_SCP=/usr/bin/scp
```

### 3.3.13.5 Starting PBS Professional

Run the PBS start script using a command with the following format, **<path to script>/pbs start**, for example:

```
/etc/init.d/pbs start
```

## 3.3.14 Installing Intel Compilers and Math Kernel Library

Install the **Intel®** Compilers and Math Kernel Library (if required) on a Service Node if it includes BOTH the Management services and the Login services. Intel **MKL** is included with the Professional Editions of Intel version 10 compilers.

Follow the instructions written in the Bull notice supplied with the computer.

## 3.3.15 Configuring the MPI User environment

**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/modulefiles/mpiBull2/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, according to the environment required. For example:

   For **MPI**:
   ```
   cp /opt/mpi/mpibull2-1.2.1-4.t/share/mpibull2.sh /etc/profile.d
   ```

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

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

2. Using the module command with the profile files to load the MPIBull2 module for the end users:

   ```
   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 with, 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 add, if necessary, 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.

## 3.4 STEP 4: Installing RHEL5.1, BAS5v1.2 for Xeon Software, and optional HPC software products on other nodes

The Management Node has to be configured to be the **NFS** server that will install the **Red Hat Linux** distribution and the Bull **BAS5 for Xeon HPC** software on all the other nodes of the cluster. Once the NFS environment has been correctly set, all that is required is that the individual nodes are booted for the Linux distribution to be installed on them.

**Only one node of each type has to be installed as KSIS will be used for the deployment, for example, install a single COMPUTE or COMPUTEX Node and then deploy it, and/or install a single IO/LOGIN Node and then deploy it. See STEP 6.**

Before running the **preparenfs** script, the prerequisites, below, must be satisfied.

**Note** If the steps in the previous section have been followed correctly, these prerequisites will already be in place.

### 3.4.1 Preparenfs script prerequisites

- The node(s) that are to be installed must have been configured in the **dhcpd.conf** file in order that an IP address is obtained on **DHCP** request.
- The option next-server, and the option filename for each host, has to be set correctly.
- The **DHCPD** service must be running, if not the script will try to start it.
- The **XINETD** service must be running and configured to run tftp, if not the preparenfs script will try to configure **tftp** and start the service.
- The BMCs of the nodes must be already configured.

### 3.4.2 Preparing the NFS node software installation

Run the **preparenfs** command:

```
preparenfs
```

Use the **–verbose** option for a more detailed trace of the execution of the preparenfs script to be stored in the **preparenfs** log file:

```
preparenfs --verbose
Use the --interactive option to force the installation to run in
interactive mode. All the Linux installation steps will be pre-filled,
and will have to be confirmed or changed:preparenfs –-interactive
```

The script will ask for the following information:

1. The path containing the operating system you want to use to prepare the PXE boot, for example, **/release/RHEL5.1/**. In the example, below, number 2 would be entered from the options displayed.

```
The following Operating System(s) have been found in the /release
directory:
     0 : Choose Custom PATH
     1 : Red Hat Enterprise Linux Server 5 (/release/TEST2)
     2 : Red Hat Enterprise Linux Server 5 (/release/RHEL5.1)
     3 : Red Hat Enterprise Linux Server 5 (/release/TEST1)

Select the line for the Operating System you want to use for the
installation:
```

2. The partitioning method to be used for the installation.

```
Select the partitioning method you want to use for the installation :
         - manual : user defined partitioning (you will be asked
                    interactively for the partitioning)
         - auto   : kickstart will use a predefined partitioning
```

The **auto** option will only handle the **sda** disk, and will leave other node disks as previously partitioned. Use the **manual** partitioning option if other previously partitioned disks need to be repartitioned.

The auto kickstart options are shown below:

| | / | /usr | /opt | /tmp | /var |
|---|---|---|---|---|---|
| swap | ext3 | ext3 | ext3 | ext3 | ext3 |
| 16 GBs | 10 GBs | 10 GBs | 10 GBs | 10 GBs | The remaining disk space |
| sda | sda | sda | sda | sda | sda |

3. The question *Do you want to enable vnc mode?* will appear. If you answer no, it will be possible to follow the installation via a serial line (conman).

4. The path that includes the **BAS5v1.2 for Xeon** software installer. This will be something like **/release/XBAS5V1.2**. A list of potential paths will be displayed, as shown below.

```
Select the path for the Bull HPC installer:

         0 : Choose Custom PATH
         1 : NONE
         2 : /release/XBAS5V1.2

Enter the number for the path :
```

5. The HPC node functions that you want to install. The possible options are: **IO, LOGIN, COMPUTE, COMPUTEX** – See *Chapter 1* for more details regarding the different **BAS5 for Xeon** architectures. Some of these functions may be installed together, as shown for the group C functions below:

```
Select the node functions to be installed. Node functions from the
same group can be added together, for example IO and LOGIN. Node
functions from different groups are exclusive.
        1 : COMPUTE     (group A)
        2 : IO          (group C)
```

```
        3 : LOGIN       (group C)
        4 : COMPUTEX    (group B)
Enter the node functions required using a comma separated list, when
more than one product is to be installed, for example: 2,3,4:
```

6. The Bull **BAS5 for Xeon** optional HPC product(s) to be installed for the cluster, as shown below. By default, the Bull **XHPC** software is always installed.

```
Select any optional Bull HPC software product(s) to be installed.
N.B. The media corresponding to your choice(s) must have been copied
into the /release/XBAS5V1.2 directory.

        0 : NONE
        1 : XIB
        2 : XLUSTRE
        3 : XTOOLKIT

Enter the product(s) to be installed using a comma separated list when
more than one product is to be installed, for example : 1,2 :
```

7. The IP address of the **NFS** server node. This must be the same node as the one on which the script runs.

8. A list of the different nodes that are included in the Cluster database will be displayed, as shown in the example below. The node name(s) of the node(s) to be installed must then be entered using the following syntax : basename[2-15,18]. The use of square brackets is mandatory.

```
Node names                Type     Status
-------------------------------------------
basename1                  A------  not_managed
basename0                  A------  up
basename[1076-1148]        -C-----  not_managed
basename[26-33,309-1075]   -C-----  up
basename[2-23]             --I----  up

The nodes that are included in the Cluster database are shown above.
Enter the list of nodes to be installed using NFS (syntax example -
basename[2-15,18]) :
```

**Note** The **BAS5 for Xeon** optional HPC products can be installed later manually (see Appendix B)

9. A detailed summary is then displayed listing the options to be used for the installation, as shown in the example below. The Administrator has to confirm that this list is correct or exit the installation.

```
INSTALLATION SUMMARY:

        PXE boot files will be copied from :
/release/RHEL5.1/images/pxeboot
        Path containing Linux Distro : /release/RHEL5.1
        NFS Server IP address is : 10.30.1.99
        Serial Line option is : ttyS1,115200
        Partitioning method is : auto
        The following hexa file(s) will be generated in
/tftpboot/pxelinux.cfg : 0A1F0106
        The path containing Bull HPC installer : /release/XBAS5V1.2
        Installation function(s): IO LOGIN
        Optional HPC product(s) : XIB XLUSTRE
```

```
Please confirm the details above or exit : [confirm] | exit :
```

**Note** Some hexa files will be created in the **/tftpboot/pxelinux.cfg** directory. These files are called hexa files because their name represents an IP address in hexadecimal format, and they are required for the PXE boot process. Each file corresponds to the IP address of a node.
For convenience the **preparenfs** script creates links to these files using the node names.

10. A line appears regarding the use of **nsctrl** commands to reboot the node where the software is going to be installed, as shown below. Before you click yes to confirm this, check that the **BMC** for the node is reachable. If this is not the case, answer no and manually reboot your node later.

```
Do you want prepareNFS to perform a hard reboot, via the
/usr/sbin/nsctrl command, on the node(s) listed for the installation?
[y] | n :
```

## 3.4.3 Launching the NFS Installation of the BAS5v1.2 for Xeon software

1. The Bull **BAS5v1.2 for Xeon** software will be installed immediately after the reboot. The progress of the install can be followed using conman via a serial line, and/or by using **vncviewer** if you have chosen to use **VNC**.

2. Once the **Linux** distribution has been installed, the **kickstart** will then manage the installation of the optional **HPC** product(s) selected for the installation, and the node will then reboot. The node can then be accessed to carry out any post-installation actions that are required using the **ssh** command (the **root** password is set to root by default).

3. The **preparenfs** script will generate a log file: **/root/preparenfs.log** on the Management Node that can be checked in case of any problems.

**See** Appendix D - *Manually Installing Bull BAS5v1.2 for Xeon Additional Software*, in this manual, if there is a need to install any of the additional software options **(XIB, XLUSTRE** and **XTOOLKIT)** later after completing this step.

## 3.5 STEP 5: Configuring Administration Software on LOGIN, I/O, COMPUTE and COMPUTEX Reference Nodes

This step describes how to install and configure **SSH, kdump, SLURM,** and **PBS Pro** as necessary for the Reference Nodes to be deployed. It also describes the installation of compilers on the Login Nodes and the configuration of the MPI User environment.

### 3.5.1 Configuring SSH and /etc/hosts

**These tasks must be performed before deployment.**

#### 3.5.1.1 For a reinstallation of BAS5 for Xeon v1.2

Retrieve the **SSH** keys of the nodes and of the root user, which have been saved previously – see *section 3.0.2*. To do this:

- Restore the **/etc/ssh** directory of each type of node to its initial destination.
- Restore the **/root/.ssh** directory on the Management Node.
- Go to the root directory:

```
cd /root
```

- From the management Node copy the **/root/.ssh** directory on to the COMPUTE(X) and LOGIN and I/O Nodes.

```
scp -r .ssh <node_name>:/root/
```

- Restart the **SSH** service on each type of node:

```
service sshd restart
```

**Notes**

- The **SSH** keys of the users can be restored from the files saved by the administrator (for example **/<username>/.ssh**).
- The **sudo** configuration will have been changed during Bull **XHPC** software installation to enable administrators and users to use the **sudo** command with **ssh**. By default, **sudo** requires a pseudo-tty system call to be created in order to work, and this is set by the **requiretty** option in the **/etc/sudoers** configuration file. In order that the automated commands run over **ssh/sudo**, the installer will have modified the default configuration file by commenting out this option.

#### Copy the /etc/hosts file onto the Reference Node

Copy the **/etc/hosts** file from Management Node using the **scp** command with the IP address of the Management Node as the source parameter.

#### Example

```
scp root@<Management_Node_IP_address>:/etc/hosts /etc/hosts
```

### 3.5.1.2 For a first installation of BAS5 for Xeon v1.2

#### On the COMPUTE /COMPUTEX and combined LOGIN I/O or dedicated LOGIN and I/O Reference Nodes

1. Copy the **/root/.ssh** directory from the Management Node on to the Reference Nodes.

```
scp -r .ssh <reference_node>:.
```

2. Test this configuration:

```
> ssh <reference_node> uname
```

```
The authenticity of host 'ns1 (127.0.0.1)' can't be established.
RSA key fingerprint is
91:7e:8b:84:18:9c:93:92:42:32:4a:d2:f9:38:e9:fc.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ns1,127.0.0.1' (RSA) to the list of known
hosts.
   Linux
```

**Note** With this SSH configuration, no password is required for root login from the Management Node to the other HPC nodes.

#### Copy the /etc/hosts file onto the Reference Node

Copy the **/etc/hosts** file from Management Node using the **scp** command with the IP address of the Management Node as the source parameter.

#### Example

```
scp root@<Management_Node_IP_address>:/etc/hosts /etc/hosts
```

## 3.5.2 Configuring Ganglia

1. Copy the file:
   **/usr/share/doc/ganglia-gmond-3.0.5/templates/gmond.conf**
   into /etc.

2. Edit the **/etc/gmond.conf** file:

– In line 18, replace xxxxx with the basename of the cluster.

```
name = "xxxxx"  /* replace with your cluster name */
```

– In line 24 replace x.x.x.x with the alias IP address of the Management Node.

```
host = x.x.x.x  /* replace with your administration node ip
address */
```

3. Start the **gmond** service:

```
service gmond start
chkconfig --level 235 gmond on
```

## 3.5.3 Configuring the kdump kernel dump tool

1. Reserve memory in the kernel that is running for the second kernel that will make the dump by adding '**crashkernel=128M@16M**' to the grub kernel line, so that 128MBs of memory at 16MBs is reserved in the /boot/grub/grub.conf file, as shown in the example below:

```
kernel /vmlinuz-2.6.18-53.el5 ro root=LABEL=/ nodmraid
console=ttyS1,115200 rhgb quiet crashkernel=128M@16M
```

   It will be necessary to reboot after this modification.

2. The following options must be set in the **/etc/kdump.conf** configuration file:

   a. The path and the device partition where the dump will be copied to should be identified by its **LABEL, /dev/sdx** or **UUID** label either in the **/home/** or **/** directories.

   Examples

```
path /var/crash
ext3 /dev/sdb1
#ext3 LABEL=/boot
#ext3 UUID=03138356-5e61-4ab3-b58e-27507ac41937
```

   b. The tool to be used to capture the dump must be configured. Uncomment the `core_collector line` and add `–d 1`, as shown below:

```
core_collector makedumpfile -c -d 1
```

   **`-c`** indicates the use of compression and **`-d 1`** indicates the dump level:

**Important**

It is essential to use non-stripped binary code within the kernel. Non-stripped binary code is included in the debuginfo RPM, kernel-debuginfo-<kernel_release>.rpm, available from http://people.redhat.com/duffy/debuginfo/index-js.html i

This package will install the kernel binary in the folder **/usr/lib/debug/lib/modules/<kernel_version>/**

**Note** The size for the dump device must be larger than the memory size if no compression is used.

Use the command below to launch **kdump** automatically when the system restarts:

```
chkconfig kdump on
```

## 3.5.4 Installing and Configuring SLURM - optional

**SLURM does not work with the PBS-Professional Batch manager and must only be installed on clusters which do not use PBS-Professional.**

The **SLURM** files are installed under the **/usr** and **/etc** directories.

**Note** These steps must be carried out for each **COMPUTE(X)** and **LOGIN** Reference Node.

### 3.5.4.1 Installing SLURM on the Reference Nodes

1. Mount NFS from the **/release** directory on the Management Node to the **/release** directory on the Node:

```
mount -t nfs <Management_Node_IP>:/release /release
```

2. Run the command below to install the **SLURM** RPMs:

```
yum install slurm pam_slurm slurm-munge slurm-auth-none slurm-devel
```

**See** The *Software Release Bulletin* for **BAS5 for Xeon v1.2** for details on how to install **SLURM** version **1.0.15**. This version is required to ensure compatibility with the LSF Batch Manager

After the **SLURM** RPMs have been installed, some steps remain before the configuration of **SLURM** is complete on the Reference Nodes. These steps can either be done using the **slurm_setup.sh** script - see section *3.5.4.3* OR manually - see section *3.5.4.4*

### 3.5.4.2 Installing Munge on the Reference Nodes - optional

**Munge** is installed as follows on clusters which use this authentication type for the **SLURM** components.

1. Run the command below on the **COMPUTE** and **I/O** reference nodes:

```
yum install munge munge-libs
```

2. Run the command below on the **COMPUTEX** and **LOGIN** reference nodes:

```
yum install munge munge-libs munge-devel
```

**Note** **munge** and **munge-libs** are installed by default as part of the standard **SLURM** installation and are included in the commands above as a check.

### 3.5.4.3 Configuring SLURM on the Reference Nodes using the Setup script tool

The **Slurm** setup script is found in **/etc/slurm/slurm_setup.sh** and is used to automate and customize the installation process.

**See** Section *3.3.11.4* for a description of the basic functions of the **slurm_setup.sh** script.

**Notes**

- The **slurm.conf** file must have been created on the Management Node and all the necessary parameters defined **BEFOR**E the script is used to propagate the information to the Reference Nodes.
- The use of the script requires **root** access, and depends on the use of the **ssh**, **pdcp** and **pdsh** tools.

#### Running the slurm_setup.sh script

As the **root** user on the Management Node, execute the script, supplying the names of the **LOGIN** and **COMPUTE(X)** Reference Nodes to be configured, for example:

```
/etc/slurm/slurm_setup.sh –N login0,compute0,computex0
```

The script will run, and in the process read the **slurm.conf** file, copy it and other required files to the Reference Nodes, create the **SlurmUser**, create the job credential keys, and create the log files as needed.

#### Additional slurm_setup.sh script options

The following additional options are available for greater control of the **slurm_setup.sh** script or for debugging purposes:

**/etc/slurm/slurm_setup.sh **
**-N <reference node list> **
**[-p <slurm user password>] **
**[-b <slurm base pathname>] **
**[-v] **
**[-d] **
**[-F] **
**[-r uid,gid]**

Parameters

| | |
|---|---|
| -N | Comma separated list of reference nodes, not including the node on which the script was invoked. After running the script on the local node, the script and other files will be copied to the Reference Nodes and the SLURM configured there as well. |
| -p <password> | Optional. If there is a need to create a logon for the slurmuser user name, a password can be specified that will be applied for slurmuser on all the nodes of the cluster. |
| -b <base_pathname> | Optional. If SLURM is installed in a directory other than the /usr default, the path to the install directory should be specified here, (e.g. /opt/slurm). This also affects the location of the SLURM configuration file: if -b is not specified, the SLURM configuration file will be accessed using the default /etc/slurm/slurm.conf path. If -b is specified, the configuration file will be accessed at <base_pathname>/etc/slurm.conf. |
| -v | Optional verbose option. If set, additional progress messages are outputted when the script is executed. |
| -d | Optional debug option. If set, parameters and variable names are outputted when the script is executed to help debugging. |
| -F | Optional force option. If slurmuser or slurmgroup already exist on any of the nodes, this option may be used to force the deletion and recreation of the user name and group name. |
| -r | Internal script option. Set for subordinate scripts, in order to inhibit actions that are only performed by the main script: e.g., generation of credential keys. |

### 3.5.4.4 Manually configuring SLURM on the Reference Nodes

If there is a problem with the SLURM setup script, then SLURM can be configured manually on the Reference Nodes. The following steps are necessary to complete the configuration of SLURM:

1. Create a SlurmUser

The SlurmUser must be created before SLURM is started. SlurmUser will be referenced by the slurmctld daemon. Create a SlurmUser on the COMPUTE(X), Login/IO or LOGIN Reference nodes with the same uid gid (105 for instance):

```
groupadd -g 105 slurm
useradd -u 105 -g slurm slurm
mkdir -p /var/log/slurm
chmod 755 /var/log/slurm
```

The gid and uid numbers do not have to match the one indicated above, but they have to be the same on all the nodes in the cluster.

The user name in the example above is slurm, another name can be used, however it has to be the same on all the nodes in the cluster.

2. **Copy the SLURM configuration file on to the reference nodes**

Copy the following files from the Management Node to the **COMPUTE(X)**, and combined **LOGIN/IO** or dedicated **LOGIN** Reference Nodes.

- **/etc/slurm/slurm.conf**
- **public.key** (using the same path as defined in the **slurm.conf** file)
- **private.key** (using the same path as defined in the **slurm.conf** file)

**Note** The public key must be on the **KSIS** image deployed to ALL the **COMPUTE/COMPUTEX** Nodes otherwise **SLURM** will not start.

3. **Check SLURM daemon directory**

Check that the directory used by the SLURM daemon (typically **/var/log/slurm**) exists on the **COMPUTE(X)**, combined **LOGIN/IO** or dedicated **LOGIN** Reference Nodes.

4. **Check access rights**

Check that all the directories listed in the **slurm.conf** file exist and that they have the correct access rights for the **SLURM** user. This check must be done on the Management Node, the combined **LOGIN/IO** or dedicated **LOGIN** and **COMPUTE(X)** Reference Nodes.

The files and directories used by **SLURMCTLD** must have the correct access rights for the **SLURM** user. The **SLURM** configuration files must be readable; the log file directory and state save directory must be writable.

### 3.5.4.5 Starting the SLURM Daemons on a Single Node

If for some reason an individual node needs to be rebooted, one of the commands below may be used.

```
/etc/init.d/slurm start  or  service slurm start
```

or

```
/etc/init.d/slurm startclean  or  service slurm startclean
```

**Note** The **startclean** argument will start the daemon on that node without preserving saved state information (all previously running jobs will be purged and node state will be restored to the values specified in the configuration file).

### 3.5.4.6 More Information

**See** The Bull **BAS5 for Xeon** *Administrator's Guide* for more information on SLURM (Munge configuration, security, the creation of job credential keys and the **slurm.conf** file). See **man slurm.conf** for more information on the parameters of the **slurm.conf** file, and **man slurm_setup.sh** for information on the **SLURM** setup script.

## 3.5.5 Installing and Configuring the PBS Professional Batch Manager – optional

- **PBS Professional does not work with SLURM.**
- **The Flexlm License Server has to be installed before PBS Professional is installed – see *section 3.3.13.1***
- **PBS Professional has to be installed on the Management Node before it is installed on the COMPUTE(X)/LOGIN reference nodes.**

**See**

- Chapter 4 in the PBS Professional *Installation and Upgrade Guide*, available on the **PBS Professional CD ROM** for more information on the installation for PBS Professional, described below.
- Chapter 3 in the PBS Professional *Administrator's Guide*, available on the **PBS Professional CD ROM** for more information on the configuration routine for PBS Professional, described below.

### 3.5.5.1 Starting the installation of PBS Professional

The commands for the installation have to be performed by the cluster Administrator logged on as root.

1. Copy and extract the package from the **PBS Pro CD-ROM** to the directory of choice on the **COMPUTE(X)** Reference Node, using a command similar to that below.

```
cd /root/PBS
tar -xvzf PBSPro_9.2.0-RHEL5_x86_64.tar.gz
```

2. Go to the installation directory on each node and run:

```
cd PBSPro_9.2.0
```

3. Start the installation process:

```
./INSTALL
```

#### Follow the installation program

During the PBS Professional installation routine, the Administrator will be asked to identify the following:

#### Execution directory

The directory into which the executable programs and libraries will be installed, for example, **/usr/pbs**

### Home directory

The directory into which the PBS Pro daemon configuration files and log files will be installed, for example, **/var/spool/PBS**

### PBS installation type

The installation type depends on the type of node that PBS Professional is being installed on and are as follows:

- On the COMPUTE Node : type 2
- On the Login Node : type 3 (This has to be a separate dedicated Login Node)

### Do you want to continue?

Answer **Yes**

### You need to specify a hostname for the Server

Give the hostname of the node where the **PBS** server has been installed, normally this is the Management Node.

### Would you like to start?

When the Installation complete window appears, the installation program offers to start PBS Professional, enter '**n**' for '**no**'.

## 3.5.5.2 Initial configuration on a COMPUTE(X) or LOGIN Reference Node

**See** Chapter 3 in the **PBS Professional** *Administrator's Guide* for more information on configuring and starting PBS Professional.

## 3.5.5.3 Initial configuration on the COMPUTE(X) Reference Node

1. Modify the **/etc/pbs.conf** file for the node as follows:

```
PBS_EXEC=/usr/pbs
PBS_HOME=/var/spool/PBS
PBS_START_SERVER=0
PBS_START_MOM=1
PBS_START_SCHED=0
PBS_SERVER=<server_name>0
PBS_SCP=/usr/bin/scp
```

2. Start PBS on the Compute Node:

```
/etc/init.d/pbs start
```

3. Run the following command on the Management Node to ensure the Compute Node is visible for the PBS server:

```
qmgr –c "create node <compute_node_name>"
```

4. Modify the initial script by removing the **–s –P** options from the **options_to_pbs_attach** line (line 177). This should appear as below:

```
vi /usr/pbs/lib/MPI/pbsrun.mpich2.init.in

                options_to_pbs_attach="-j $PBS_JOBID”
```

5. Add the MPIBull2 wrapper, by using the command below.

```
/usr/pbs/bin/pbsrun_wrap  /opt/mpi/mpibull2-<xxx>/bin/mpirun pbsrun.mpich2
```

#### 3.5.5.4 Initial configuration on the LOGIN Reference Node

Modify the **/etc/pbs.conf** file for the node as follows:

```
PBS_EXEC=/usr/pbs
PBS_HOME=/var/spool/PBS
PBS_START_SERVER=0
PBS_START_MOM=0
PBS_START_SCHED=0
PBS_SERVER=<server_name>0
PBS_SCP=/usr/bin/scp
```

**Important**

**The Login Nodes have to be defined on the Management Node by creating a /etc/hosts.equiv file containing the Login Node names (one per line).**

### 3.5.6 Installing Compilers

Install the Intel Compilers on the LOGIN Reference Nodes (if required).

Follow the instructions written in the Bull notice supplied with the compiler.

### 3.5.7 Intel Math Kernel Library (MKL)

Install the **Intel® MKL** libraries on the Compute, Extended Compute and Login Reference Nodes (if required). Intel MKL is included with the Professional Editions of **Intel** version 10 compilers

Follow the instructions written in the Bull notice supplied with the compiler.

### 3.5.8 Configuring the MPI User Environment

**See** Section *3.3.15* in this chapter for details.

## 3.5.9 Bull Scientific Studio

The Bull Scientific Studio RPMs are installed automatically on the **COMPUTE(X)/LOGIN** reference nodes.

**See** The **BAS5 for Xeon** *User's Guide* and *System Release Bulletin* for more information on the libraries included in Scientific Studio.

## 3.5.10 NVIDIA Tesla Graphic Card accelerators – optional

The drivers for both the **NVIDIA Tesla C1060** card and for the **NVIDIA Tesla S1070** accelerator are installed automatically on the **COMPUTE(X)/LOGIN** reference nodes.

**The NVIDIA Tesla C1060 card is used on NovaScale R425 servers only and the NVIDIA Tesla S1070 accelerator can be used by both NovaScale R422 E1 and R425 servers.**

## 3.5.11 NVIDIA CUDA Toolkit – optional

The **NVIDIA CUDA™ Toolkit** and **Software Development Kit** are installed automatically on the **LOGIN**, **COMPUTE** and **COMPUTEX** reference nodes for clusters which include Tesla graphic accelerators, so that the **NVIDIA** compilers and the **NVIDIA** mathematical and scientific libraries are in place for the application.

### Configuring NVIDIA CUDA Toolkit

The **PATH** and **LD_LIBRARY_PATH** environmental variables should be modified to give access to the directories where the **CUDA** Toolkit has been installed, as shown in the examples below:

### Examples

**PATH**=/usr/kerberos/bin:/opt/intel/fce/10.1.015/bin:/opt/intel/cce/10.1.013/bin:/opt/cuda/bin:/usr/local/bin:/bin:/usr/bin

**LD_LIBRARY_PATH**=/usr/local/cuda/lib:/opt/intel/cce/10.1.013/lib:/opt/intel/mkl/9.0/lib/em64t:/opt/intel/fce/10.1.015/lib:/opt/cuda/lib

**See** The **BAS5 for Xeon** *User's Guide* and *System Release Bulletin* for more information on the NVIDIA compilers and libraries.

The **NVIDIA CUDA** *Compute Unified Device Architecture Programming Guide,* and the other documents in the **/opt/cuda/doc** directory for more information.

## 3.5.12 Installing RAID Monitoring Software - optional

### 3.5.12.1 Monitoring using the LSI MegaRAID 8408E Adapter

**Note** This kind of adapter is only installed on **NovaScale R440** and **NovaScale R460** machines.

Install the **MegaCli-xxxx.i386.rpm** package which is available on the *Bull Extension Pack CD-ROM*, below, delivered with the machines which use these adapters:

*Bull Extension Pack for NovaScale Universal Rack-Optimized & Tower Series with RHEL5.1*

No further configuration is required for the **NovaScale R440** and **R460** machines once the **MegaCli-xxxx.i386.rpm** is installed.

### 3.5.12.2 Monitoring using the AOC-USASLP-S8iR Adapter

**Note** This kind of adapter is installed on **NovaScale R423** and **NovaScale R425** machines only.

1. Install the **StorMan-xxxx.x86_64.rpm** package which is available on the CD-ROM, below delivered with the machines which use these adapters:

   *SUPERMICRO AOC-USAS-SRL*

2. Then run the commands below:

```
service stor_agent stop
chkconfig stor_agent off
```

3. Check that **RAID** has been configured correctly by running the command:

```
lsiocfg -cv |more
```

4. Look for the host which has **aacraid** displayed against it. Verify that the detailed information for the Logical and Physical disks displays correctly, as shown in the example below.

```
host7  aacraid   0        256     -     Optimal             SMC AOC-USAS-S8iR-LP
       DRV= 1.1-5 (2437)
       FW= 5.2-0 (15575)
       Interface=SAS/SATA
       Slot=7
       SN=4FAFF0
       LogicalDisks=2
              Number=1 Name=RD1 Device=sdd Status=Optimal Raid=1 Size=239190 DiskLocations="0,3 0,1"
              Number=2 Name=RD5SAS Device=sde Status=Optimal Raid=5 Size=280188 DiskLocations="0,4 0,5 0,6"
       PhysicalDisks=8
              Device=0 SN=WD-WCAPW5321110 WWN=Unknown State=Ready Location=0,0 Vendor=WDC Size=476940 LogicalDisk= Role=
              Device=1 SN=WD-WCANY3792200 WWN=Unknown State=Online Location=0,1 Vendor=WDC Size=239372 LogicalDisk=RD1 Role=
              Device=2 SN=WD-WCANY3792290 WWN=Unknown State=Ready Location=0,2 Vendor=WDC Size=239372 LogicalDisk= Role=
              Device=3 SN=WD-WCANY3600678 WWN=Unknown State=Online Location=0,3 Vendor=WDC Size=239372 LogicalDisk=RD1 Role=
              Device=4 SN=DQ00P65004LP WWN=500000E01203CF60 State=Online Location=0,4 Vendor=FUJITSU Size=140272 LogicalDisk=RD5SAS Role=
              Device=5 SN=DQ00P65004ND WWN=500000E01203D2A0 State=Online Location=0,5 Vendor=FUJITSU Size=140272 LogicalDisk=RD5SAS Role=
              Device=6 SN=DQ00P65004L9 WWN=500000E01203CE80 State=Online Location=0,6 Vendor=FUJITSU Size=140272 LogicalDisk=RD5SAS Role=
              Device=7 SN=WD-WCANY3792380 WWN=Unknown State=Ready Location=0,7 Vendor=WDC Size=239372 LogicalDisk= Role=
```

## 3.6 STEP 6: Creating and Deploying an Image Using Ksis

This step describes how to perform the following tasks:
1. Installation and configuration of the image server
2. Creation of an image of the COMPUTE(X) Node and Login or I/O or Login/IO Reference Node installed previously.
3. Deployment of these images on cluster nodes.

These operations have to be performed from the Management Node.

**Please refer to *BAS5 for Xeon High Availability Guide* if High Availability is to be included for any part of your cluster to check that all the High Availability configurations necessary are in place on the Reference Node image.**

**Note** To create and deploy a node image using **Ksis**, all system files must be on local disks and not on the disk subsystem. To create an I/O node image, for example, all disk subsystems must be unmounted and disconnected.

**It is only possible to deploy an image to nodes that are equivalent and have the same hardware architecture:**
- **Platform**
- **Disks**
- **Network interface**

**See** The **BAS5 for Xeon** *Administrator's Guide* for more information about **Ksis**.

### 3.6.1 Installing, Configuring and Verifying the Image Server

#### 3.6.1.1 Installing the Ksis Server

The **Ksis** server software is installed on the Management Node from the **XHPC CDROM**. It uses **NovaScale** commands and the cluster management database.

#### 3.6.1.2 Configuring the Ksis Server

**Ksis** only works if the cluster management database is correctly loaded with the data which describes the cluster (in particular with the data which describes the nodes and the administration network).

The preload phase which updates the database must have been completed before using Ksis.

### 3.6.1.3 Verifying the Ksis Server

In order to deploy an image using **Ksis**, various conditions must have been met for the nodes concerned. If the previous installation steps have been completed successfully then these conditions will be in place. These conditions are listed below.

1. Start the **systemimager** service by running the command.

```
service systemimager start
```

2. Each node must be configured to boot from the network via the **eth0** interface. If necessary edit the BIOS menu and set the Ethernet interface as the primary boot device.

3. The access to cluster management database should be checked by running the command:

```
ksis list
```

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

4. Check the state of the nodes by running the **nsctrl** command:

```
nsctrl status ip_node_name
```

   The output must not show nodes in an inactive state meaning that they are not powered on.

5. Check the status of the nodes by running the **ksis nodelist** command:

```
ksis nodelist
```

## 3.6.2 Creating an Image

Create an image of the **COMPUTE(X)** and **LOGIN** and **I/O** or combined **LOGIN/IO** reference nodes (according to cluster type) installed previously.

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

Example:

```
ksis create image1 ns1
```

This command will 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.

### 3.6.3 Deploying the Image on the Cluster

**Note** Before deploying the image it is mandatory that the equipment has been configured – see *STEP 3*.

1. Before deploying check the status of the nodes by running the command `ksis nodelist`:

```
ksis nodelist
```

2. 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
```

3. Each node must be configured to boot from the network via the **eth0** interface. If necessary edit the BIOS menu and set the Ethernet interface as the primary boot device.

4. Start the deployment by running the command:

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

5. 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]
```

**Note** 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.

### 3.6.4 Post Deployment Node configuration

#### 3.6.4.1 Edit the postconfig script - Clusters with Ethernet interconnects only

Before running the **postconfig** command below the script will need editing as follows for **Ethernet** clusters:

1. Run the command below to disable the configuration of the interconnect interfaces:

```
ksis postconfig disable CONF_60_IPOIB
```

2. Recompile the **postconfig** script by running the command below:

```
ksis postconfig buildconf
```

**Important**

**Do not edit the postconfig script when installing on InfiniBand clusters.**

### 3.6.4.2 postconfig command

Once the image deployment has finished, the cluster nodes will need to be configured according to their type, Compute, I/O, etc. Post deployment configuration is mandatory as it configures **Ganglia, Syslog-ng, NTP, SNMP** and **Pdsh** automatically on these machines. It also allows the IP over **InfiniBand** interfaces to be configured according to the information in the Cluster database.

The **Ksis postconfig** command configures each node that the image has been deployed to, ensuring that all the cluster nodes of a particular type are homogenous.

**Ksis** post-configuration is carried out by running the command:

```
ksis postconfig run PostConfig <cluster_name>[nodelist]
```

For example

```
ksis postconfig run PostConfig xena[1-100]
```

### 3.6.4.3 Configure the Interconnect Interfaces - Ethernet clusters only

The interconnect **Ethernet** interface description file is generated from the Management Node for each node by using the **config_ipoib** command.

**See** *Appendix E - Configuring Interconnect Interfaces* for more details regarding the use of the **config_ipoib** command.

# 3.7 STEP 7 : Final Cluster Checks

## 3.7.1 Testing 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 nova[8-10] hostname
```

2. This will give output similar to that below:

```
nova10: nova10
nova9: nova9
nova8: nova8
```

### Groups of nodes

1. Run the **dbmGroup** command

```
dbmGroup show
```

2. This will give output similar to that below:

```
Group Name  Description                    Nodes Name
-------------------------------------------------------
ADMIN       Nodes by type:ADMIN            nova[0,12]
ALL         All nodes except node admin    nova[1-10]
Burning     Burning group                  nova5
COMP        Nodes by type:COMP             nova[1-4,7-8]
COMP128GB   COMPUTE node with 128GB        nova8
COMP48GB    COMPUTE node with 48GB         nova4
Deploy      Deploy group                   nova3
HwRepair    HwRepair group                 nova8
IO          Nodes by type:IO               nova[6,10]
META        Nodes by type:META             nova[5,9]
MYFAME      ensemble des fame du cluster   nova[0,4-6,8-10]
NODES128GB  Nodes by memory size:128GB     nova8
NODES16GB   Nodes by memory size:16GB      nova[1-3,7]
NODES48GB   Nodes by memory size:48GB      nova[4,6,10]
NODES64GB   Nodes by memory size:64GB      nova[0,5,9,12]
QxTest      QxTest group                   nova[0,6]
TEST        TEST group                     nova[5,9]
UnitTest    UnitTest group                 nova[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 below:

```
nova[6,10]
```

```
Thu Aug 7 15:35:27 CEST 2008
```

### 3.7.2 Checking NTP

1. Run the following command on a COMUTE(X) node and on a combined LOGIN/IO Login or dedicated LOGIN nodes:

```
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
```

### 3.7.3 Checking Syslog-ng

1. Check on the admin 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, check in the **/var/log/messages** file if the message is present.

4. On the admin host, check in the **/var/log/HOSTS/<node_hostname>/messages** file if the message is present:

### 3.7.4 Checking Nagios

Both **nagios** and **httpd** services have to be running on the Management Node:

```
service nagios status
nsm_nagios (pid 31356 31183 19413) is running...
service httpd status
> 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://<admin_node_name>/NSMaster`

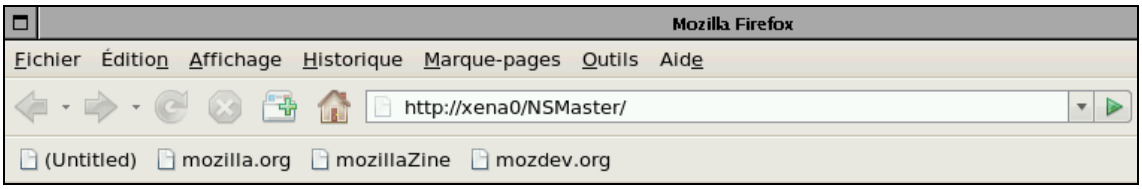


Figure 3-17. Launching NovaScale Master

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

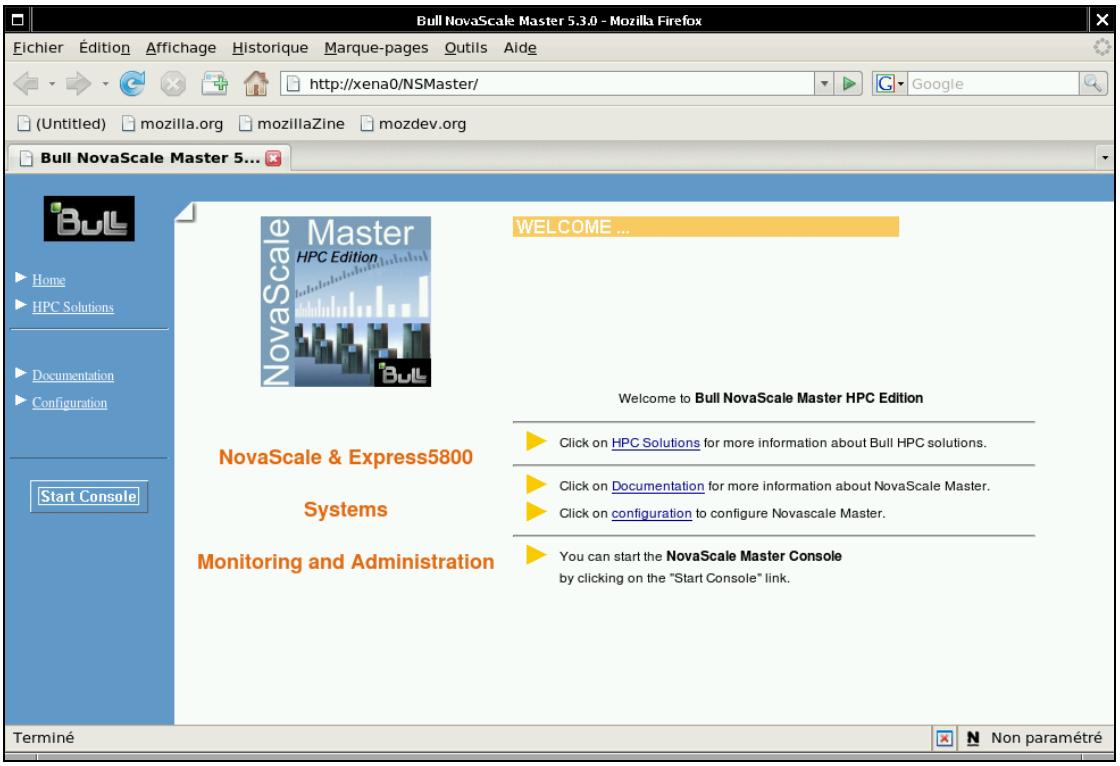


Figure 3-18. NovaScale Master Welcome screen

An authentication window appears asking for user name and password.

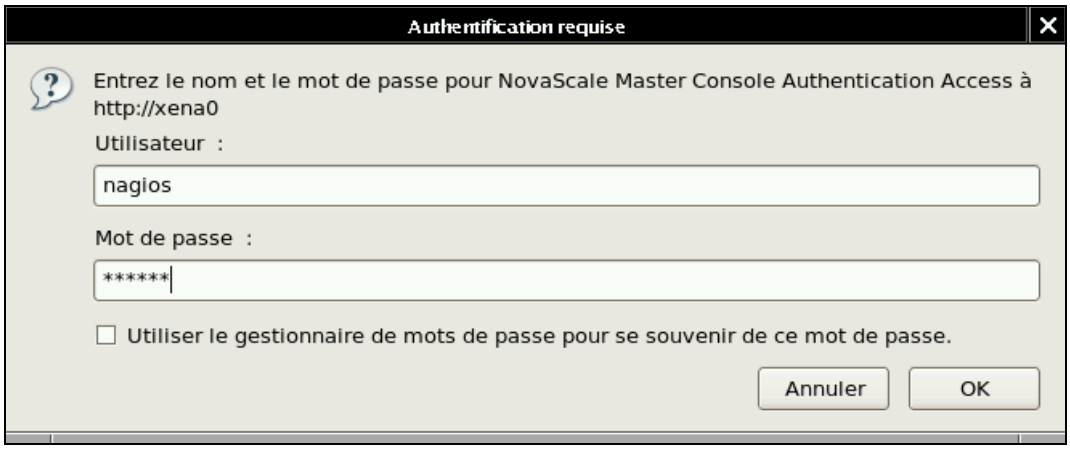


Figure 3-19. NovaScale Master Authentication Window

3. Once authenticated, the **NovaScale Master** console appears.

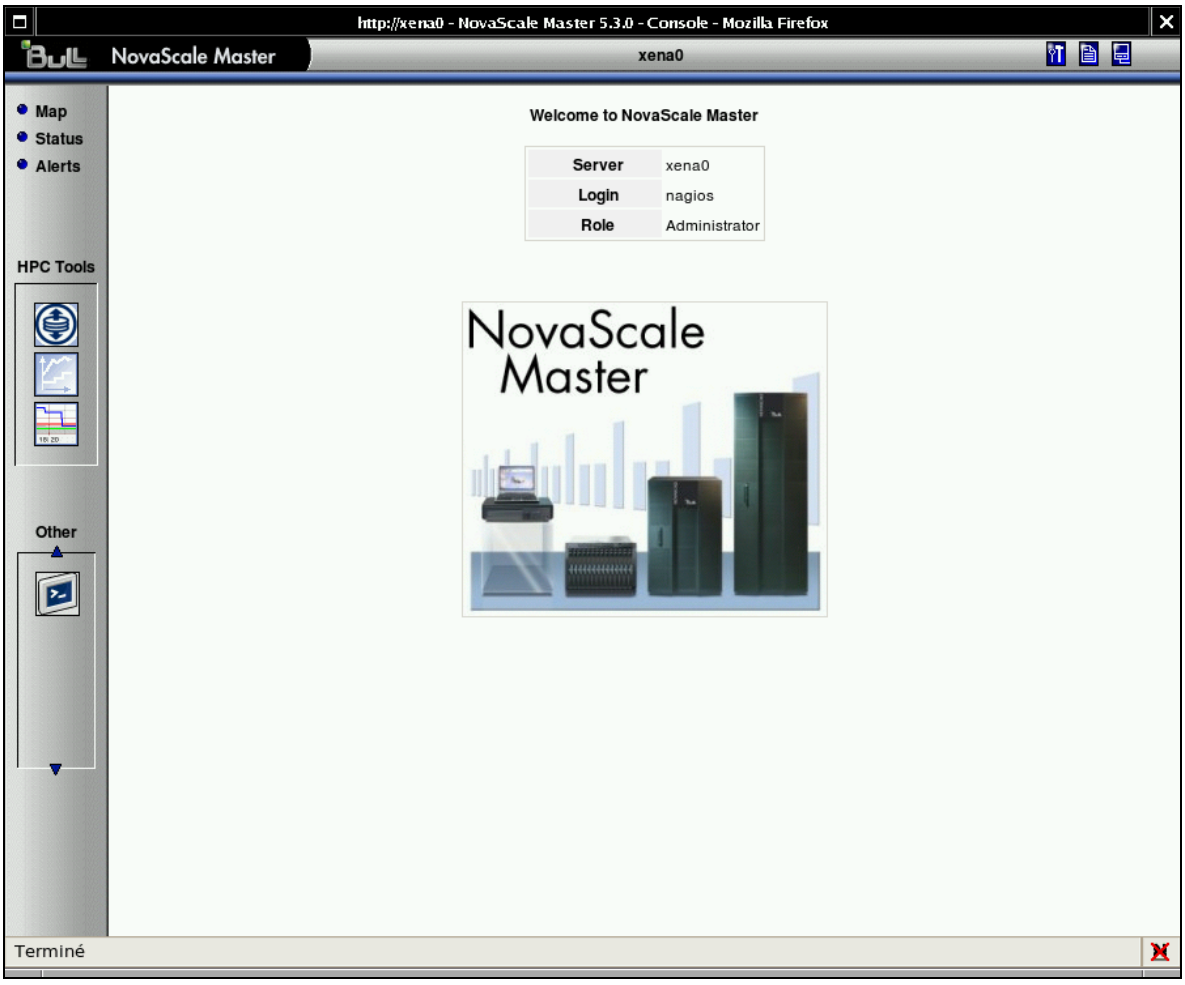


Figure 3-20. The NovaScale Master console

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

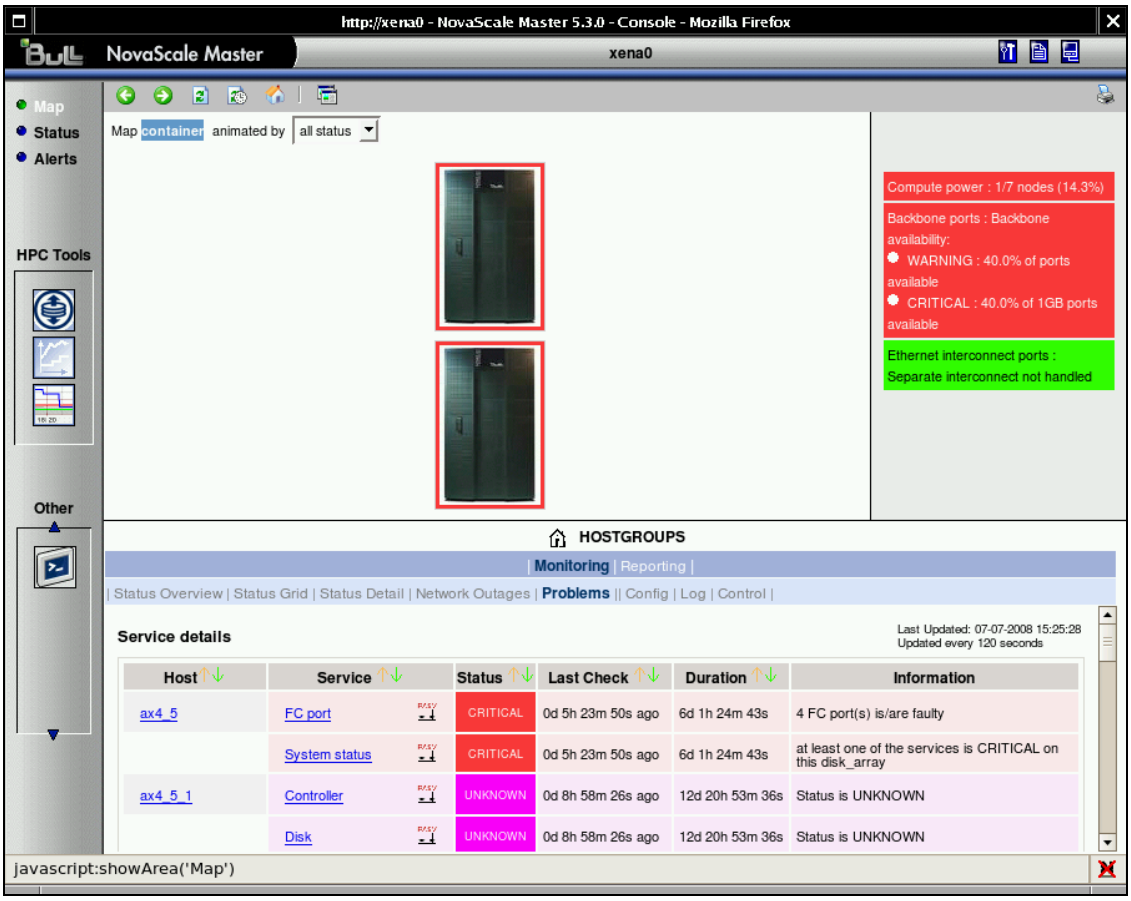


Figure 3-21. NovaScale Master Monitoring Window

## 3.7.5 Checking nsctrl

**nsctrl** is a command that allows administrators to issue commands to the node BMCs

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

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

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

```
[root@xena0 ~]# nsctrl status xena[1-5]
```

This will give output similar to that below:

```
xena2 : Chassis Power is on
xena1 : Chassis Power is on
xena3 : Chassis Power is on
xena5 : Chassis Power is on
xena4 : Chassis Power is on
[root@xena0 ~]#
```

## 3.7.6 Checking Conman

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

**Usage**: conman [OPTIONS] [CONSOLES]

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

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

```
[root@xena0 ~]# service conman status
```

```
conmand (pid 5943) is running...
[root@xena0 ~]#
```

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

```
[root@xena0 ~]# conman xena2
```

```
 <ConMan> Connection to console [xena2] opened.
 Red Hat Enterprise Linux Server release 5.1 (Tikanga)
Kernel 2.6.18-53.1.21.el5.Bull.1 on an x86_64
 xena2 login:
```

## 3.7.7 Testing PBS Professional – Basic setup

1. A user should be created on all the nodes for testing purposes, in the examples below this is referred to as **user-test**. This is done as follows:

```
useradd –g <group> -d <home login>
```

2. The **ssh** keys for the user should have been dispatched to all nodes, normally this will have done at STEP *5* during the installation procedure - see section *3.5.1.2* for more information.

3. Launch a test job from either the Management Node or the Login Node as the test user, using a command similar to that below:

```
echo "sleep 60" | /usr/pbs/bin/qsub
```

4. Check the execution of the job using the **qstat** command, as shown below:

```
qstat -an
```

5. This will give output in format similar to that below:

```
nova0:
                                                            Req'd    Req'd       Elap
Job ID          Username    Queue    Jobname    SessID NDS TSK Memory   Time    S  Time
--------------- ----------- -------- ---------- ------ --- --- ------ ------- - -----
0.nova0         user-test   workq    STDIN      8424   1    1   --       --     R  0:00
nova8/0
```

6. Once the job has finished check that no errors are listed in the output, as in the example below:

```
cat STDIN.e0
```

```
cat STDIN.o0
```

7. If there are problems run the **tracejob** command so that the problem can be identified.

```
tracejob <job_ID>
```

This will give output similar to that below, where no errors are reported:

```
Job: 0.nova0

07/17/2008 16:24:31  L    Considering job to run
07/17/2008 16:24:31  S    enqueuing into workq, state 1 hop 1
07/17/2008 16:24:31  S    Job Queued at request of bench@nova0, owner
= user-test@nova0, job name = STDIN, queue = workq
07/17/2008 16:24:31  S    Job Run at request of Scheduler@nova0 on
hosts (nova8:ncpus=1)
07/17/2008 16:24:31  S    Job Modified at request of Scheduler@nova0
07/17/2008 16:24:31  L    Job run
07/17/2008 16:25:31  S    Obit received momhop:1 serverhop:1 state:4
substate:42
07/17/2008 16:25:31  S    Exit_status=0 resources_used.cpupercent=0
resources_used.cput=00:00:00
resources_used.mem=2796kb resources_used.ncpus=1
resources_used.vmem=167888kb resources_used.walltime=00:01:00
07/17/2008 16:25:31  S    dequeuing from workq, state 5
```

8. If errors are reported then look at the **STDIN.e0** output file for PBS Professional problems, and the **STDIN.o0** output file for other problems. See the **PBS Professional** *Administrator's Guide* for more information regarding PBS Professional problems.

### Testing a job launched in parallel

1. Give the test job a name, in the example that follows this is 'HelloWorld'.

2. Execute the **cat run.pbs** command:

```
cat run.pbs
```

3. This will give output similar to that below:

```
#! /bin/bash
#PBS -l select=2:ncpus=4:mpiprocs=4
#PBS -l place=scatter
#PBS -N HelloWorld

source /opt/intel/fce/<version>/bin/ifortvars.sh
source /opt/intel/cce/<version>/bin/iccvars.sh
source /opt/mpi/mpibull2-<version>/share/setenv_mpibull2.sh
mpibull2-devices -d=ibmr_gen2

mpirun –n 8 ./helloWorld
```

4. Check that the test job was launched successfully across all the CPUs requested, as in the example above.

5. If errors are reported then look at the **run.e<job_ID>** output file for PBS Professional problems and the **run.o<job_ID>** output file for other problems. See the **PBS Professional** *Administrator's Guide* for more information regarding PBS Professional problems.

### 3.7.8 Checking and Starting the SLURM Daemons on COMPUTE(X) and Login/IO Nodes

Check to see if the **Slurmctld** daemon has started on the Management Node and the **Slurmd** daemon has started on the combined LOGIN/IO or dedicated LOGIN and on a COMPUTE(X) Node by using the command:

```
scontrol show node --all
```

If NOT then start the daemons using the commands below:

- For the Management Node:

```
service slurm start
```

- For the Compute Nodes:

```
service slurm start
```

Verify that the daemons have started by running the **scontrol show node –all** command again.

### 3.7.9 Testing kdump

**Important**

It is essential to use non-stripped binary code within the kernel. Non-stripped binary code is included in the debuginfo RPM, kernel-debuginfo-<kernel_release>.rpm, available from http://people.redhat.com/duffy/debuginfo/index-js.html i
This package will install the kernel binary in the folder
**/usr/lib/debug/lib/modules/<kernel_version>/**

In order to test that **kdump** is working correctly a dump can be forced using the commands below.

```
echo 1 > /proc/sys/kernel/sysrq
echo c > /proc/sysrq-trigger
```

The end result can then be analysed using the crash utility. An example command is shown below. The **vmcore** dump file may also be found in the **/var/crash** folder.

```
crash /usr/lib/debug/lib/modules/<kernel_version>/vmlinux vmcore
```

# Chapter 4. Configuring Storage Management Services

This chapter describes how to:

- Configure the storage management software installed on the Management Node
- Initialize the management path to manage the storage systems of the cluster
- Register detailed information about each storage system in the ClusterDB.

The following topics are described:

4.1 *Enabling Storage Management Services*

4.2 *Enabling FDA Storage System Management*

4.3 *Enabling DataDirect Networks (DDN) S2A Storage Systems Management*

4.4 *Enabling the Administration of an Optima 1250 Storage System*

4.5 *Enabling the Administration of EMC/Clariion (DGC) storage system*

4.6 *Updating the ClusterDB with Storage Systems Information*

4.7 *Storage Management Services*

4.8 *Enabling Brocade Fibre Channel Switches*

**Note** When installing the **storageadmin-xxx** rpms in update mode (**rpm –U**), all the configuration files described in this section and located in **/etc/storageadmin** are not replaced by the new files. Instead the new files are installed and suffixed by **.rpmnew**. Thus, the administrators can manually check the differences, and update the files if necessary.

**See** For more information about setting up the storage management services, refer to the *Storage Devices Management* chapter in the Bull **BAS5 for Xeon** *Administrator's Guide.*

Unless specified, all the operations described in this section must be performed on the cluster management station, using the root account.

## 4.1 Enabling Storage Management Services

Carry out these steps on the Management Node.

1. Configure ClusterDB access information:
   The ClusterDB access information is retrieved from the **/etc/clustmngt/clusterdb/clusterdb.cfg** file.

2. Edit the **/etc/cron.d/storcheck.cron** file to modify the period for regular checks of the status for storage devices. This will allow a periodic refresh of status info by pooling storage arrays. Four (4) hours is a recommended value for clusters with tens of storage systems. For smaller clusters, it is possible to reduce the refresh periodicity to one (1) hour.

```
0 */2 * * * root /usr/bin/storcheck > /var/log/storcheck.log 2>&1
```

## 4.2 Enabling FDA Storage System Management

**This section only applies when installing for the first time.**

**See** The *Bull FDA User's Guide* and *Maintenance Guide* specific to the **StoreWay FDA** model that is being installed and configured.

The management of **FDA** storage arrays requires an interaction with the FDA software, (delivered on the CDs provided with the storage arrays). The Cluster management software installed on the cluster Management Node, checks the FDA management software status. Several options are available regarding the installation of this FDA software.

### The FDA manager server and CLI

These two components are mandatory for the integration of FDA monitoring in the cluster management framework. A **FDA** manager server is able to manage up to 32 storage arrays. The server and **CLI** components must be installed on the same system, for as long as the cluster contains less than 32 FDA systems.

### The FDA Manager GUI client

The GUI client provides an easy to use graphical interface, which may be used to configure, and diagnose any problems, for FDA systems. This component is not mandatory for the integration of the FDA in a cluster management framework.

**Note** The external Windows station must have access to the FDA manager server.

The Linux **rdesktop** command can be used to provide access to the GUI from the cluster Management Node.

### FDA Storage System Management prerequisites

- A laptop is available and is connected to the maintenance port (MNT) using an Ethernet cross cable. Alternatively, a maintenance port of the FDA is connected to a Windows station.
- The electronic license details are available. These have to be entered during the initialisation process.
- Knowledge of installing and configuring FDA storage systems.
- The User manuals for this storage system should be available.
- The **FDA** name must be the same as in the disk array table for the **ClusterDB** and for the **iSM** server.

- The FDA Manager user name and password have to have been transferred to the respective **necadmin** and **necpasswd** fields in the **/etc/storageadmin/nec_admin.conf** file.
- The addresses predefined in the **ClusterDB** for the management ports. These may be retrieved using the **storstat** command.

### 4.2.1 Installing and Configuring FDA software on a Linux system

On Linux, the **disk_array** table in the **ClusterDB** contains the **mgmt_node_id** field which is the foreign key for the node table. This table contains information, for example the IP address for the FDA storage manager.

The Storage Manager server and the CLI software may be installed on a Linux system planned for FDA management.

**Note** The Storage Manager GUI client can only be installed on Windows

1. Install the RPMs.

```
rpm –iv ISMSMC.RPM ISMSVR.RPM
```

   – The **ISMSMC.RPM** is located on the *FDA series – StoreWay Manager Integration Base CDROM*.

   – The **ISMSVR.RPM** is located on the *FDA series – StoreWay ISM Storage Manager CDROM*.

2. **FDA Manager** Configuration.

   a. Copy the **/etc/iSMsvr/iSMsvr.sample** file into the **/etc/iSMsvr/iSMsvr.conf** file. Add the lines that define the disk arrays to be managed, using the syntax shown in the example below:

```
# 3fda1500
# Two IP addresses are defined
diskarray1 =(
ip =(172.17.0.200, 172.17.0.201)
)
# 4fda2500
# Two IP addresses are defined
diskarray2 =(
ip =(172.17.0.210, 172.17.0.211)
)
```

   b. Add the following line in the client section after the default line for login1 in the **iSMsvr.conf** file. Note that the **<admin user>** and the **<admin password>** details must be consistent with the corresponding fields in the **/etc/storageadmin/nec_admin.conf** file.

```
login2 = (<admin>, <password>, L3)
```

   c. Then restart the **iSM** manager service:

```
/etc/init.d/iSMsvr restart
```

3. **FDA CLI** Configuration.

   a. Copy the **/etc/iSMSMC/iSMSM.sample** file into the **/etc/iSMSM/iSMSM.conf** file.

   b. Restart the CLI manager service:

```
/etc/init.d/iSMSMC restart
```

### Enabling ssh access from the Management Node on a Linux System

**Note** This part of the process is only required when the **FDA** software is installed on a system other than the Management Node. There is no need to enable **ssh** access if the NEC software is located locally on the Management Node. If this is the case, skip this paragraph.

**ssh** is used by the management application to monitor the FDA storage systems. **ssh** must be enabled so that FDA management tools operate correctly on the cluster Management Node.

Distribute **RSA** keys to enable password-less connections from the cluster Management Node:

1. Log on as root on the cluster Management Node and generate asymmetric **RSA** keys.

2. Go to the directory where the RSA keys are stored. Usually, it is **"~/.ssh"**. You should find **id_rsa** and **id_rsa.pub** files. The **.pub** file must be appended to the **authorized_keys** file on the Linux FDA manager system. The **authorized_keys** file defined in the **/etc/sshd_config** file, (by default: **~/.ssh/authorized_keys**) must be used.

3. If no key has been generated, generate a key with the **ssh-keygen** command

```
ssh-keygen  -b 1024 -t rsa
```

**The default directory should be accepted. This command will request a passphrase to retrieve the password. Do not use this function; press the return key twice to ignore the request.**

4. The public key for the FDA manager Linux system should be copied with **sshd**:

```
scp id_rsa.pub <administrator>@<LinuxFDAhost>:~
```

   **< LinuxFDAhost >** can be a host name or an IP address. Replace **<administrator>** with the existing administrator login details.

5. Connect to the Linux system FDA manager.

```
ssh <administrator>@< LinuxFDAhost >
```

6. Do not destroy the **~/.ssh/authorized_keys** file. Run:

```
mkdir –p .ssh
cat id_rsa.pub >> .ssh/authorized_keys
rm id_rsa.pub
```

**Note** If necessary, repeat this operation for other pairs of Linux and FDA manager users.

### Enabling password-less ssh execution for the Apache server for the Management Node

**ssh** may also be activated from the Linux Apache account. For this specific user, **sudo** must be configured.

Check that the appropriate rights have been set for the **nec_admin** command:

```
grep nec_admin /etc/sudoers
```

This command should return the following line:

```
%apache ALL=(root)NOPASSWD:/usr/sbin/nec_admin
```

If this does not happen, run **visudo** to modify the sudoers file and add the line above.

## 4.2.2 Configuring FDA Access Information from the Management Node

1. Obtain the Linux or Windows host user account, and the **iSM** client user and password which have been defined. All the FDA arrays should be manageable using a single login/password.

2. Edit the **/etc/storageadmin/nec_admin.conf** file, and set the correct values for the parameters:

```
# On Linux iSMpath="/opt/iSMSMC/bin/iSMcmd"
# On Windows iSMpath="/cygdrive/c/Program\
Files/FDA/iSMSM_CMD/bin/iSMcmd"
iSMpath = /opt/iSMSMC/bin/iSMcmd
# iSMpath="/cygdrive/c/Program\ Files/FDA/iSMSM_CMD/bin/iSMcmd"
# NEC iStorage Manager host Administrator
hostadm = administrator
# NEC iStorage Manager administrator login
necadmin = admin
# NEC iStorage Manager administrator password
necpasswd = password
```

## 4.2.3 Initializing the FDA Storage System

1. Initialise the storage system using the maintenance port (MNT). The initial setting must be done through the Ethernet maintenance port (MNT), using the Internet Explorer browser. Refer to the documentation provided with the **FDA** storage system to perform the initial configuration.

Important

**The IP addresses of the Ethernet management (LAN) ports must be set according to the values predefined in the ClusterDB.**

```
storstat –d –n <fda_name> -i -H
```

2. Carry out the following post configuration operations using the **iSM** GUI.
   Start the **iSM** GUI and verify that the FDA has been discovered. Make the following settings:
   - Set a FDA name which is the same as the name already defined in the ClusterDB **disk_array** table.
   - Enable the **SNMP** traps, and send the traps to the cluster Management Node.

   It is possible to connect to the server via the browser using one of the **FDA** Ethernet IP addresses if the **iSM** GUI is not available. Use the password '**C**' to access the configuration menu.

**See** The *User's Guide* for the **FDA** storage system for more information.

3. Check that end-to-end access is correctly setup for the cluster Management Node:

```
nec_admin -n <fda_name> -i <ip-address-of-the-Windows-FDA-management-
station> -c getstatus -all
```

## 4.3 Enabling DataDirect Networks (DDN) S2A Storage Systems Management

### 4.3.1 Enabling Access from Management Node

Edit the **/etc/storageadmin/ddn_admin.conf** file to configure the singlet connection parameters.

```
# Port number used to connect to RCM API server of ddn
port = 8008

# login used to connect to ddn
login = admin

# Password used to connect to ddn
password = password
```

The configuration file uses the factory defaults connection parameters for the S2A singlets. The **login** and **password** values may be changed.

### 4.3.2 Enabling Date and Time Control

If the HPC cluster includes DDN storage systems check, and if necessary update, the **/etc/cron.d/ddn_set_up_date_time.cron** file to modify regular time checks. Ensure that the default period (11 pm) is acceptable for your environment:

```
0 23 * * * root /usr/sbin/ddn_set_up_date_time -s all -f –l
```

This cron synchronizes times for **DDN** singlets daily.

**Note** If the configuration does not include **DDN** storage systems then the line above must be commented.

### 4.3.3 Enabling Event Log Archiving

The **syslog** messages generated by each DDN singlet are stored in the **/var/log/DDN** directory or in the **/varha/log/DDN** directory if the Management Node is configured for High Availability.

**Note** The log settings, for example, size of logs are configured by default. Should there be a need to change these sizes, edit the **/etc/logrotate.d/syslog-ng** file. See the **logrotate** man page for more details.

### 4.3.4 Enabling Management Access for Each DDN

1. List the storage systems as defined in the cluster management database:

```
storstat –a |grep DDN
```

This command returns the name of the DDNs recorded in the cluster management database. For example:

```
ddn0 |     DDN |       9500 |    WARNING |    |     RACK-A2 |  K
No faulty subsystem registered !
```

The next operation must be done once for each DDN system.

2. Retrieve the addressing information:

```
storstat -d -n <ddn_name> -i -H
```

**Tip**: To simplify administrative tasks, Bull preloads the **ClusterDB** with the following conventions:

| DDN system name | IP name for singlet 1 | IP name for singlet 2 | Console name for singlet 1 | Console name for singlet 2 |
|---|---|---|---|---|
| <ddn_name> | <ddn_name>_s1 | <ddn_name>_s2 | <ddn_name>_s1s | <ddn_name>_s2s |

IP names and associated IP address are automatically generated in the **/etc/hosts** directory. The conman consoles are automatically generated in the **/etc/conman.conf** file. Otherwise, refer to the **dbmConfig** command.

## 4.3.5 Initializing the DDN Storage System

Initialize each DDN storage system either from the cluster Management Node or from a laptop, as described below.

### 4.3.5.1 Initialization from a Cluster Management Node with an existing Serial Interface between the Management Node and the DDNs

Check that **ConMan** is properly configured to access the serial ports of each singlet:

```
conman <console name for the singlet>
```

When you hit return, a prompt should appear.

#### ddn_init command

The **ddn_init** command has to be run for each DDN. The target DDN system must be up and running, with 2 singlets operational. The serial network and the Ethernet network must be properly cabled and configured, with **ConMan** running correctly, to enable access to both serial and Ethernet ports, on each singlet.

**Notes**

- The **ddn_init** command is not mandatory to configure DDN storage units. The same configuration can be achieved via other means such as the use of DDN CLI (**ddn_admin**) or DDN telnet facilities (to configure other items).
- The **ddn_init** command can only be run at the time of the first installation or if there is a demand to change the IP address for some reason

```
ddn_init -I <ddn_name>
```

This command performs the following operations:

- Set the IP address on the management ports
- Enable telnet and API services
- Set prompt
- Enable syslog service, messages directed to the Management Node, using a specific UDP port (544)
- Enable SNMP service, traps directed to the Management Node
- Set date and time
- Set common user and password on all singlets
- Activate SES on singlet 1
- Restart singlet
- Set self heal
- Set network gateway.

#### ddn_init command tips

- The **ddn_init** command should not be run on the DDN used by the cluster nodes, as this command restarts the DDN.
- Both singlets must be powered on, the serial access configured (conman and portserver) and the LAN must be connected and operational before using the **ddn_init** command.
- Randomly, the DDN may have an abnormally long response time, leading to time-outs for the **ddn_init** command. Thus, in case of error, try to execute the command again.
- The **ddn_init** command is silent and takes time. Be sure to wait until it has completed.

WARNING

**The ddn_init command does not change the default tier mapping. It does not execute the save command when the configuration is completed.**

### 4.3.5.2 Initialization from a Laptop without an existing Serial Interface between the Management Node and the DDNs

Connect to the laptop to each serial port and carry out the following operations:

- Set the IP address on the management ports according to the values of the ClusterDB.
- Enable telnet and API services.
- Set prompt.
- Configure and enable the syslog service and transmit the messages to the Cluster Management Node, using a specific UDP port (544).
- Configure and enable SNMP service, traps directed to the Cluster Management Node.
- Set date and time.
- Set admin user and password and all singlets, according to the values defined in **/etc/storageadmin/ddn_admin.conf** file.
- Activate SES on singlet 1.
- Set the tier mapping mode.
- Enable the couplet mode.

- Activate cache coherency.
- Disable cache write back mode.
- Set self heal.
- Set network gateway.

**Notes**

- The laptop has to be connected to each one of the 2 **DDN** serial ports in turn. This operation then has to be repeated for each DDN storage unit
- The administrator must explicitly turn on the 8 and 2 mode on DDN systems where dual parity is required. This operation is not performed by the **ddn_init** command.

**Important**

**SATA systems may require specific settings for disks. Consult technical support or refer to the DDN *User's Guide* for more information.**
**When the default command has been performed on the system, it is recommended to restart the complete initialisation procedure.**
**After a power down or a reboot, check the full configuration carefully.**

Check that initialization is correct, that the network access is setup, and that there is no problem on the DDN systems:

```
ddn_admin -i <ip-name singlet 1> -c getinfo -o HW
ddn_admin -i <ip-name singlet 2> -c getinfo -o HW
```

## 4.4 Enabling the Administration of an Optima 1250 Storage System

**This section only applies when installing for the first time.**

**Note** The High Availability solution does not apply for nodes which are connected to Optima 1250 Storage Bays

**See** The **Storeway** *Optima 1250 Quick Start Guide* for more details on the installation and configuration

**Storeway Master** is a web interface module embedded into the Optima 1250 controllers.

It allows an Optima 1250 storage system to be managed and monitored from a host running **StoreWay Master** locally using a web browser across the internet or an intranet.

There is no particular software which needs to be installed to manage an Optima 1250 storage system.

### 4.4.1 Optima 1250 Storage System Management Prerequisites

- If the initial setup was not done by manufacturing, a laptop should be available and connected to the Ethernet Port of the **Optima**1250 storage system via an Ethernet cross cable.
- The **SNMP** and **syslogd** electronic licenses sent by e-mail should be available. The Global Licence is included in the standard product.
- The **Storeway** *Optima 1250 Quick Start Guide* specific to the storage system should be available.
- The addresses predefined in the **ClusterDB** must be the same as those set in **Storeway Master** for the Optima 1250. These may be retrieved using the **storstat -di** command.

### 4.4.2 Initializing the Optima 1250 Storage System

1. The network settings of the Optima 1250 storage system will need to be configured for the first start up of the **StoreWay Master** module, if this has not already been done by manufacturing.
   - Configure you LAPTOP with the local address 10.1.1.10
   - Connect it to the Ethernet Port of the Optima 1250 storage system using an Ethernet cross cable

- Insert the Software and manual disk, delivered with the Optima 1250 storage system, into you CD drive. The autorun program will automatically start the navigation menu.
- Select **Embedded Storeway Master set up**
- Review the information on the screen and click the next button. The program searches the embedded master module using the addresses 10.1.1.5 and 10.1.1.6
- Use the embedded module MAC address for each controller whose network settings are being configured. The IP addresses of the Ethernet management (LAN) ports must be set according to the values predefined in the ClusterDB.
- Enter and confirm the new password and then click the configure button.

---

**See** The *Storeway Optima 1250 Quick Start Guide* for more information.

---

2. Once the network settings are configured, you can start **StoreWay Master** using a web browser by entering the explicit IP address assigned to the embedded StoreWay Master server followed by the port number (9292), for example **http://<IP_address>:9292**

3. If the default settings are changed (user name =admin, password = password), then the user name and password settings in the **xyradmin** and **xyrpasswd** fields of the **/etc/storageadmin/xyr_admin.conf** file will have to be updated.

4. Configure **SNMP** using the **StoreWay Master** GUI, firstly select the **Settings** button and then the **SNMP** button. If this is the first time that SNMP has been set you will be asked for the paper licence details that are included with the Optima 1250 storage system. Using the **SNMP** menu enter the IP address of the management station and deselect the information level box for this trap entry (leave the warning and error levels checked).

5. Check that end-to-end access has been correctly set up for the cluster Management Node using the command below:

```
xyr_admin -i <optima_1250_IP_address> -c getstatus -all
```

# 4.5 Enabling the Administration of EMC/Clariion (DGC) storage systems

## 4.5.1 Initial Configuration

**See** The appropriate *EMC CLARiiON CX3-Series Model XX (40 or 20 or 10c) Setup Guide* delivered with the storage system for more details on the initial configuration. A Windows laptop and a RS232 cable will be required.

The initialization parameters are saved in the cluster database (**da_ethernet_port** table) and can be retrieved as follows:

1. Run the command below to see the **EMC/Clariion** storage system information defined in the cluster management database.

```
storstat -a | grep DGC
```

   This command will list the **DGC** disk arrays to be configured on the cluster.

2. For each DGC storage system retrieve the IP addressing information by using the command below.

```
storstat -d -n <dgc_name> -i -H
```

3. For each Service Processor (SPA and SPB) of each CX3-40f set the IP configuration parameters for the:
   - IP address
   - Hostname (for SPA : <dgc_name>_0, for SPB : <dgc_name>_1)
   - Subnet Mask
   - Gateway
   - Peer IP address (IP address of the other SP of the same DGC disk array)

   Once these settings have been made, the Service Processor will reboot and its IP interface will be available.

4. The Java and Firefox plugins are installed and linked by default, so that the http interface for the **EMC Navisphere Management Suite** can be used for the complementary configuration tasks.

   Start the **Firefox** browser by running the command:

```
/usr/bin/firefox-<Bull version>
```

   However, if there is a problem follow the procedure defined below to install these plugins:

   a. Install the following 2 RPMs from the **BONUS** directory on the Bull **XHPC DVD.**

   **XHPC/BONUS/jre-<version>-linux-i586.rpm**

XHPC/BONUS/firefox-<Bull version>-<release>.i386.rpm

b. These are installed by running the commands below:

```
cd /release/XBAS5V1.2/XHPC/BONUS
rpm -i jre-<version>-linux-i586.rpm firefox-<Bull version>-<release>.i386.rpm
```

c. Declare the java plugin for the newly installed **Firefox** version as follows:

```
cd /usr/lib/firefox-<Bull version>/plugins
ln –s /usr/java/latest/plugin/i386/ns7/libjavaplugin_oji.so .
```

d. Restart the **/usr/bin/firefox-<Bull version>** browser to take the java plugin into account.

## 4.5.2 Complementary Configuration Tasks for EMC/Clariion CX series storage devices

The disk array is configured via the **Navisphere Manager** interface in a web browser using the following URLs:

**http://<SPA-ip-address>** or **http://<SPB-ip-address>**

1. Set the disk array name by selecting the disk array and opening the properties tab.
2. Set the security parameters by selecting the disk array and then selecting the following option in the menu bar:

   **Tools -> Security -> User Management**

   Add a username and a role for the administrator.
3. Set the monitoring parameters as follows
   a. Using the **Monitors** tab, create a Monitoring template with the following parameters:

      **General** tab:
      - **Events** = General
      - **Event Severity** = Warning + Error + Critical
      - **Event Category =** Basic Array Feature Events

      **SNMP** Tab:
      - **SNMP Management Host** = <IP address of the HPC Storage Management station>
      - **Community** = public

   b. Using the **Monitors** tab, associate the new template to each Service Processor by selecting the **Monitor Using Template** option.

### 4.5.3 Complementary Configuration Tasks for EMC/Clariion AX4-5 storage devices

The disk array is configured via the **Navisphere Express** interface in a web browser using the following URLs:

http://<SPA-ip-address> or http://<SPB-ip-address>

1. Set the disk array name in the *Manage / Storage System* page
2. Set the security parameters in the *System / Settings / User Management* page:
   Add a username and a password for the administrator.
3. Set the monitoring parameters in the *System / Settings / Event Notification* page:
   Set *SNMP Trap Destination* = <IP address of the Management node>

### 4.5.4 Configuring the EMC/Clariion (DGC) Access Information from the Management Node

1. Install the **Navisphere CLI rpm** on the Administration Node.

---

**Note** This package is named **navicli.noarch.rpm** and is available on the *EMC CLARiiON Core Server Support* CD-ROM, which is delivered with an **EMC/Clariion** storage system.

---

2. Edit the **/etc/storageadmin/dgc_admin.conf** file, and set the correct values for the security parameters, including:
   - Navisphere CLI security options (for naviseccli only)
   - The same user and password must be declared on each disk array by using the command below.

```
dgc_cli_security = -User <user> -Password <password> -Scope 0
```

## 4.6 Updating the ClusterDB with Storage Systems Information

1. For each storage system, run the command below.

```
storregister -u -n <disk_array_name>
```

   As a result the **ClusterDB** should now be populated with details of disks, disk serial numbers, **WWPN** for host ports, and so on.

2. Check that the operation was successful by running the command below.

```
storstat -d -n <disk_array_name> -H
```

   If the registration has been successful, all the information for the disks, manufacturer, model, serial number, and so on should be displayed.

## 4.7 Storage Management Services

The purpose of this phase is to build, and distribute on the cluster nodes attached to fibre channel storage systems, a data file which contains a human readable description for each **WWPN**. This file is very similar to **/etc/hosts**. It is used by the **lsiocfg** command to display a textual description of each fibre channel port instead of a 16 digit **WWPN**.

1. Build a list of **WWPN**s on the management station:

```
lsiocfg -W > /etc/wwn
```

**Note** This file must be rebuilt if a singlet is changed, or if FC cables are switched, or if new LUNs are created.

2. Distribute the file on all the nodes connected to fibre channel systems (for example all the I/O nodes).

The file can be included in a **KSIS** patch of the Compute Nodes. The drawback is that there are changes to the **WWPN** then a new patch will have to be distributed on all the cluster nodes.

Another option is to copy the **/etc/wwn** file on the target nodes using the **pdcp** command:

```
pdcp -w <target_nodes> /etc/wwn /etc
```

## 4.8 Enabling Brocade Fibre Channel Switches

### 4.8.1 Enabling Access from Management Node

The ClusterDB is preloaded with configuration information for **Brocade** switches. Refer to the **fc_switch** table. If this is not the case, then the information must be entered by the administrator.

Each Brocade switch must be configured with the correct IP/netmask/gateway address, switch name, login and password, in order to match the information in the ClusterDB.

Please refer to *Chapter 9* for more information about the switch configuration. You can also refer to Brocade's documentation.

### 4.8.2 Updating the ClusterDB

When the Brocade switches have been initialized, they must be registered in the ClusterDB by running the following command from the Management Node for each switch:

```
fcswregister -n <fibrechannel switch name>
```

# Chapter 5. Configuring I/O Resources for the Cluster

The configuration of I/O resources for the cluster consists of two phases:

**Phase 1: The configuration of the storage systems**

- Definition of the data volumes (LUNs) with an acceptable fault tolerance level (RAID)
- Configuration of the data access control rules for the I/O nodes
- Configuration of specific parameters (cache size, cache policy, watermarks, etc.)

**Phase 2: The configuration of coherent naming for I/O node resources**

- Definition of logical names (aliases) for LUNs that maintain device names following reboots.
- Configuration of Quorum disks (optional) for High-Availability.

The I/O configuration can either be automatically deployed (with some exceptions) or configured manually.

## 5.1 Automatic Deployment of the I/O Configuration

**Automatic deployment of the I/O configuration is not possible for Optima 1250 and EMC/CLARiiON AX4-5 storage systems. These systems must be configured manually.**

The automatic deployment of the storage configuration uses a *model* file that describes the data volumes that have to be created, and how the nodes can access them.

**See** The **BAS5 for Xeon** *Administrator's Guide* for more detailed information about configuration models and the deployment process.

### 5.1.1 Storage Model Files

A template for the creation of a storage configuration model can be obtained with the following command:

```
stormodelctl –c showtemplate
```

This template contains declaration examples for storage systems supported from the different storage vendors. A model file is specific to storage systems of the same type from a specific vendor.

The model file contains the following information:

- The storage vendor name
- The list of storage system names to which the model is applicable
- Vendor-specific information (cache configuration, watermarks, etc.)
- Declaration of RAID groups (grouping disks in pools)
- Declaration of spare disks
- Declaration of LUNs
- Declaration of LUN access control groups and mappings of internal/external LUN numbers
- **LUSTRE** specific declarations for storage systems which use the **LUSTRE** global file system deployment.

**Note** With some versions of Fibre Channel adapter node drivers, the correct detection of the LUNs for a storage device port is dependent on the accessibility of a LUN numbered 0. It is recommended the Access Control groups for a storage device are configured so that the list of LUNs declared for each group always includes an external LUN that is numbered 0.

A model file is created by manual by editing the file, and its syntax is checked when the model is deployed to the storage systems.

Although there is no constraint about the location of storage model files, a good practice is to store them in the */etc/storageadmin* directory of the Management Node.

**The Administrator should backup storage model files as model files may be reused later to reinstall a particular configuration.**

## 5.1.2 Automatic Configuration of a Storage System

The automatic configuration of storage system using a model file requires that the storage devices declared in the model are initialized correctly and are accessible via their management interface.

**When a storage model is deployed any existing configuration details that are in place are overwritten. All previous data will be lost.**

### Initial conditions

For some storage systems (**EMC/CLARiiON** ), the LUNs can only be accessed using authorized Fibre Channel adapters (HBAs) for the hosts connected to the storage system. This access control is based on the Worldwide Names (**WWN**) of the FC adapters. So these WWN details must be collected and stored in the Cluster Database using the following command:

```
ioregister –a
```

The collection of I/O information may fail for those nodes which are not yet operational in the cluster. Check that it succeeded for the nodes referenced by the Mapping directives in the model file (i.e. for the nodes that are supposed to be connected to the storage system).

### Configuration process

1. Create or reuse a storage configuration model and copy it into the **/etc/storageadmin** directory on the Management node:

```
cd /etc/storageadmin
```

2. Apply the model to the storage systems:

```
stormodelctl -m <model_name> -c applymodel
```

**WARNING**
**This command is silent and long. Be certain to wait until the end.**

To have better control when applying the model on a single system it is possible to use the verbose option, as below:

```
stormodelctl -m <model_name> -c applymodel -i <disk_array_name> -v
```

3. Check the status of formatting operations on the storage systems.

   When the **applymodel** command has finished, the disk array proceeds to LUN formatting operations. Depending on the type of storage system, this operation can take a long time (several hours). The progress of the formatting phase can be checked periodically using the following command:

```
stormodelctl -m <model_name> -c checkformat
```

   The message '*no formatting operation*' indicates that the formatting phase has finished and is OK.

**WARNING**
**Ensure that all formatting operations are completed on all storage systems before using these systems for other operations.**

4. Once the storage systems have been fully configured, reboot all the nodes that are connected to them so that the storage systems and their resources can be detected.

.

**Note** The LUN Access control information (zoning) can be reconfigured, using the **stormodelctl –c applyzoning** option, once the configuration model has been deployed. The LUN configuration and all other parameters are preserved.

### 5.1.3 Automatic Deployment of the configuration of I/O resources for the nodes

**Note** All the storage systems connected to the nodes must have been configured, their LUNs formatted, and the nodes rebooted before this phase is carried out.

1. Check that each node is connected to the correct storage system.

   Check the connection of each DDN storage system using the following command.

```
ddn_conchk -I <ddn_name> -f
```

**Note** This command can only be used if **ConMan** is in place for the DDN storage systems.

Check that the LUNs are accessible for the storage systems connected to each node by using the command below:

```
lsiocfg –dv
```

2. Deploy the aliases for the I/O resources from the Management Node.

   As a prerequisite **ssh** must have been configured "password-less" to allow the Management Node to run remote operations on the nodes connected to storage systems. Run the command below, using the model file created previously when the storage system was automatically configured:

```
stordepmap -m <model_name>
```

**WARNING**
**This command is silent and long. Be sure to wait until the end.**

This operation transmits configuration information to each node attached to the storage system defined in the specified model file. A check is made to ascertain which storage resources are accessible from each node compared with the LUNs defined in the model file for it. A symbolic link (alias) is then created for each disk resource that corresponds to a storage system LUN declared in the model file for the node.

3. Check aliases created for I/O resources.

   Use the following command on each node to check that the aliases have been created correctly:

```
stormap –L
```

All device aliases listed must return an '*up*' status.

### Quorum disks

If one or more LUNs for a storage system have been declared as quorum disks for **Cluster Suite**, the configuration/formatting of these devices as quorum disks is done automatically.

Use the command below on each node that is included in a High Availability pair to check this.

```
mkqdisk –L
```

### Restoring a node

After restoring the system on a node, the aliases also have to be restored using the deployment command, below, from the Management Node:

```
stordepmap -m <model_name> -i <node_name>
```

## 5.2 Manual Configuration of I/O Resources

**Important**

**It is not recommended to configure the I/O resources manually except for those storage systems where automatic configuration is not supported i.e. Optima 1250 and EMC/CLARiiON AX4-5.**

### 5.2.1 Manual Configuration of Storage Systems

Please refer to the documentation provided with the storage system to understand how to use the storage vendor's management tools. Most of the configuration operations can also be performed from the Management Node using the CLI management commands (**ddn_admin, nec_admin, dgc_admin, xyr_admin commands**) provided by the storage administration packages .

**See** The **BAS5 for Xeon** *Administrator's Guide* for more information.

### 5.2.2 Manual Configuration of I/O resources for Nodes

**Note** All the storage systems connected to the nodes must have been configured, their LUNs formatted, and the nodes rebooted before this phase is carried out.

1. Check that each node is connected to the correct storage system.

   Check the connection of each DDN storage system using the following command.

```
ddn_conchk -I <ddn_name> -f
```

**Note** This command can only be used if **ConMan** is in place for the DDN storage systems.

Check that the LUNs are accessible for the storage systems connected to each node by using the command below:

```
lsiocfg –dv
```

2. Create aliases from the Management Node without using a model file.

   An alias must be created for each LUN of a storage system connected to a node. If I/O multipathing has been configured, ensure that all paths to all devices are in the *alive* state by using the **lsiocfg –x** command.

   **If the node is NOT in a High-Availability pair:**

   From the Management Node, run the command:

```
stordiskname –c –r <node_name>
```

   Then run the command:

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

   **If the node is in a High-Availability pair (node1,node2):**

   From the Management Node run the command:

```
stordiskname –c –r <node1_name>,<node2_name>
```

   Then run the command:

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

3. Check the aliases created for the I/O resources.

   Use the following command on each node to check that the aliases have been created correctly:

```
stormap –L
```

   All device aliases listed must return an ‘*up*’ status.

**Note** For some storage systems, not including **FDA** and **DDN**, the **stordiskname** command may return an error similar to the one below:

```
Error : -= This tool does not manage configuration where a given
UID appears more than once on the node = -
```

If this happens try running it with the **–m SCSI_ID** option.

The **stordiskname** command builds a **/etc/storageadmin/disknaming.conf** file which contains, among other things, details of symbolic link names, the LUN UIDs and the WWPN access for the LUN's. Only the **stordiskname** command can create or modify the node specific information in this file.

### Quorum disks

If one or more LUNs on a storage system have been configured as quorum disks for **Cluster Suite**, aliases will also be created for these LUNs but it is important NOT to use these LUNs for other purposes apart from quorum disks.

On each node that is included in a High Availability pair use the commands, below, to check this.

```
mkqdisk –L
stormap –L
```

### Restoring a node

The **disknaming.conf** file will be erased when redeploying the **ksis** reference image, or when the system is restored for a node. Therefore, the **stordiskname** command should be used with the **–r** option (remote) from the Management Node enabling backups and restorations of the **/etc/storageadmin/disknaming.conf** file to be managed automatically. This is highly recommended.

If the **-r** option is not used, the Administrator will have to manage the backup of the **/etc/storageadmin/disknaming.conf** file himself.

When used remotely (**–r** option) - immediately after a **ksis** image re-deployment, or a node system restoration - the **stordiskname** command must be used in **update** mode (**-u** option). This ensures that the LUNs are addressed by the same symbolic link names as used previously, and avoids having to configure the file system again.

The **stordiskname** command should be executed from the Management Node as shown below (possibly with the **–m SCSI_ID** option, see *Note* above)

### If the node is NOT in a High-Availability pair

```
stordiskname –u –r <node_name>
```

### If the node is in a High-Availability pair

```
stordiskname –u –r <node1_name>,<node2_name>
```

The symbolic links (aliases) must be recreated on each node using the information contained within the **disknaming.conf** file newly created by **stordiskname**. To do this, run the **stormap** command as described previously:

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

# Chapter 6. Configuring File Systems

Three types of file structure are possible for sharing data and user accounts for **BAS5 for Xeon** clusters:

- **NIS** (Network Information Service) can be used so that user accounts on Login Nodes are available on the Compute Nodes.
- **NFS** (Network File System) can be used to share file systems in the home directory across all the nodes of the cluster.
- **Lustre** Parallel File System

This chapter describes how to configure these three file structures.

## 6.1 Setting up NIS to share user accounts

**For those clusters which include dedicated I/O + LOGIN nodes there is no need to use NIS on the Management Node.**

### 6.1.1 Configure NIS on the Login Node (NIS server)

1. Edit the **/etc/sysconfig/**network file and add a line for the **NISDOMAIN** definition.

```
NISDOMAIN=<DOMAIN>
```

   Any domain name may be used for **<DOMAIN>**, however, this name should be the same on the Login node, which is acting as the NIS server, and on all the Compute Nodes (NIS clients).

2. Start the **ypserv** service

```
service ypserv start
```

3. Configure **ypserv** so that it starts automatically whenever the server is started.

```
chkconfig ypserv on
```

4. Initialize the **NIS** database.

```
/usr/lib64/yp/ypinit –m
```

**Note** When a new user account is created the YP database should be updated by using the command:

```
cd /var/yp
make
```

## 6.1.2 Configure NIS on the Compute or/and the I/O Nodes (NIS client)

1. Edit the **/etc/sysconfig/**network file and add a line for the NISDOMAIN definition.

```
NISDOMAIN=<DOMAIN>
```

   Any domain name may be used for **<DOMAIN>**, however, this name should be the same on the Login node, which is acting as the NIS server, and on all the Compute or I/O Nodes (NIS clients).

2. Edit **/etc/yp.conf** and add a line to set the Login Node as the NIS domain server

```
domain <DOMAIN> server <login_node>
```

3. Modify the **/etc/nsswitch.conf** file so that **passwd**, **shadow** and **group** settings are used by **NIS**.

```
passwd: files nisplus nis
shadow: files nisplus nis
group: files nisplus nis
```

4. Connect to the **NIS** YP server.

```
service ypbind start
```

5. Configure the **ypbind** service so that it starts automatically whenever the server is restarted.

```
chkconfig ypbind on
```

**Note** The **NIS** status for the Compute or I/O Node can be verified by using the **ypcat hosts** command. This will return the list of hosts from the **/etc/hosts** file on the **NIS** server.

### Nodes which use an image deployed by Ksis

The **/etc/sysconfig/network** file is not included in an image that is deployed from the reference node to the other Compute or I/O Nodes. This means that the **NISDOMAIN** definition has to be added manually to the files that already exist on the Compute or I/O Nodes by using the command below.

```
pdsh –w cluster[x-y] ‘echo NISDOMAIN=<DOMAIN> >>
/etc/sysconfig/network’
```

The **restart ypbind** service then has to be restarted so that the NIS domain is taken into account.

```
pdsh –w cluster[x-y] ‘service ypbind restart’
```

## 6.2 Configuring NFS v3 to share the /home_nfs and /release directories

### 6.2.1 Preparing the LOGIN node (NFS server) for the NFSv3 file system

Firstly, create a dedicated directory (mount point) for the **NFS** file system which is dedicated to *'home'* usage. As the **/home** directory is reserved for local accounts, it is recommended that **/home_nfs** is used as the dedicated *'home'* directory for the NFS file system.

#### Recommendations

- Use dedicated devices for **NFS** file systems (one device for each file system that is exported).
- The **lsiocfg -d** command will provide information about the devices which are available.
- Use the **LABEL** identifier for the devices.
- Use disks that are partitioned.

**Important**

If a file system is created on a disk which is not partitioned, then mount cannot be used with the **LABEL** identifier. The disk device name (e.g. **/dev/sdX**) will have to be specified in the **/etc/fstab** file.

**Notes**

- The following instructions only apply if dedicated disks or storage arrays are being used for the **NFS** file system.
- The following examples refer to configurations that include both **home_nfs** and **release** directories. If the '**release'** NFS file system has already been exported from the Management Node, ignore the operations which relate to the **release** directory in the list of operations below.

1. Create the directories that will be used to mount the physical devices:

```
mkdir /home_nfs
mkdir /release
```

2. Mount the physical devices:

```
mount <home_nfs dedicated block device> /home_nfs
mount <release dedicated block device> /release
```

or, if labels have been applied to the file systems:

```
mount LABEL=<label for home_nfs dedicated block device> /home_nfs
mount LABEL=<label for release dedicated block device> /release
```

3. Edit the **/etc/fstab** file and add the following lines for the settings which are permanent:

```
# these are physical devices (disks) dedicated to NFS usage
LABEL=release /release auto defaults 0 0
LABEL=home_nfs /home_nfs auto defaults 0 0
```

4. Use the **adduser** command with the **-d** flag to set the **/home_nfs** directory as the home directory for new user accounts.

```
adduser -d /home_nfs/<NFS user login> <NFS user_login>
```

## 6.2.2 Setup for NFS v3 file systems

### Configuring the NFSv3 Server

1. Edit the **/etc/exports** file and add the directories that are to be exported.

```
/release *(ro,sync)
/home_nfs *(rw,sync)
```

2. Restart the **NFS** service

```
service nfs restart
```

3. Configure the **NFS** service so that it is automatically started whenever the server is restarted.

```
chkconfig nfs on
```

**Note** Whenever the **NFS** file systems configuration is changed (**/etc/exports** modified), then the **exportfs** command is used to configure the **NFS** services with the new configuration.

```
exportfs -r
exportfs -f
```

### Configuring the NFSv3 Client

1. Create the directories that will be used to mount the **NFS** file systems.

```
mkdir /release
mkdir /home_nfs
```

2. Edit the **/etc/fstab** file and add the **NFSv3** file system.

```
<nfs server>:/release /release nfs defaults 0 0
<nfs server>:/home_nfs /home_nfs nfs defaults 0 0
```

3. Mount the **NFS** file systems.

```
mount /release
mount /home_nfs
```

## 6.3 Configuring the Lustre file system

**Important**

**For clusters which include High Availability for Lustre this section should be read alongside the *Configuring High Availability for Lustre* chapter in the BAS5 for Xeon *High Availability Guide*. Lustre HA pointers are included throughout this section. These indicate when and where the additional configurations required for Lustre High Availability should be carried out.**

This section describes how to:

- Initialize the information to manage the **Lustre** File System
- Configure the storage devices that the **Lustre** File System relies on
- Configure **Lustre** file systems
- Register detailed information about each Lustre File System component in the Cluster DB

**Important**

**These tasks must be performed after the deployment of the I/O Nodes.**

Unless specified, all the operations described in this section must be performed on the cluster Management Node, from the root account.

**See** If there are problems setting up the **Lustre** File System, and for more information about Lustre commands, refer to the **BAS5 for Xeon** *Administrator's Guide*. This document also contains additional information about High Availability for I/O nodes and the Cluster DB.

### 6.3.1 Enabling Lustre Management Services on the Management Node

1. Restore the Lustre system configuration information if performing a software migration:
   - **/etc/lustre** directory,
   - **/var/lib/ldap/lustre** directory if Lustre High Availability is included
2. Verify that the I/O and metadata nodes information is correctly initialized in the ClusterDB by running the command below:

```
lustre_io_node_dba list
```

This will give output similar to that below, displaying the information specific to the I/O and metadata nodes. There must be one line per I/O or metadata node connected to the cluster.

```
IO nodes caracteristics
id name type netid clus_id HA_node net_stat stor_stat lustre_stat
4 ns6 --I-- 6 -1 ns7 100.0 100 OK
5 ns7 --IM- 7 -1 ns6 100.0 100 OK
```

The most important things to check are that:

- ALL the I/O nodes are listed with the right type: I for OSS and/or M for MDS.
- The High Availability node is the right one.

It is not a problem if **net_stat, stor_stat, lustre_stat** are not set. However, these should be set when the file systems are started for the first time.

In there are errors, the ClusterDB information can be updated using the command:

```
lustre_io_node_dba set
```

**Note** Enter `lustre_io_node_dba --help` for more information about the different parameters available for **lustre_io_node_dba**

3. Check that the file **/etc/cron.d/lustre_check.cron** exists on the Management Node and that it contains lines similar to the ones below:

```
# lustre_check is called every 15 mn
*/15 * * * * root /usr/sbin/lustre_check >> /var/log/lustre_check.log 2>&1
```

## 6.3.2 Configuring I/O Resources for Lustre

**Skip this phase when carrying out an update to BAS5 for Xeon v1.2, or if BAS5 for Xeon v1.2 is being reinstalled, as the Lustre configuration and data files will have been saved.**

At this point of the installation, the storage resources should have already been configured, either automatically or manually, and in accordance with the type of storage system.

**See** Chapter 5 - *Configuring I/O Resources for the cluster* - in this manual for configuration details (manual and automatic) for each type of storage system.

#### 6.3.2.1 Configuring I/O Resources for Lustre after Automatic Deployment of I/O Configuration

This phase must take place after executing the procedures described in the *Automatic Configuration of a Storage System* and *Automatic Deployment of the configuration of I/O resources for the nodes* sections in Chapter 5.

**Important**

When carrying out an update to **BAS5 for Xeon v1.2**, or if **BAS5 for Xeon v1.2** is being reinstalled, do not run the following two **stormodelctl** commands, as the **Lustre** configuration and data files will have been saved.

The automated configuration of **Lustre** I/O resources uses the storage model file described in *Automated Deployment of the I/O Configuration* section in chapter 5. This model file details how Lustre uses the configured **LUN**s (description of **OST** and **MDT** data and journal LUNs).

The **Lustre** tables in the Cluster database should be populated with the information found in the model file, as described in this section.

1. Declare the Lustre OST configuration:

```
stormodelctl -m <model_name> -c generateost
```

2. Declare the Lustre MDT configuration:

```
stormodelctl -m <model_name> -c generatemdt
```

3. Make the OSTs and MDTs available for the Lustre filesystem:

```
lustre_investigate check
```

#### 6.3.2.2 Configuring I/O Resources for Lustre after Manual I/O Configurations

This phase must take place after executing the procedures described in the *Manual Configuration of Storage Systems* and *Manual Configuration of I/O resources for Nodes* in Chapter 5.

The **Lustre** tables in the Cluster database must be populated using the **/etc/lustre/storage.conf** file.

### 6.3.3 Adding Information to the /etc/lustre/storage.conf file

**See** The **BAS5 for Xeon** *Administrator's Guide* for more details about the **storage.conf** file

This phase should be done in the following situations:

- If there is a need to use the **Lustre** file system and no cluster database is available, as may be the case for clusters which do not include **Lustre** High Availability.
- If there is a cluster database but no management tools are provided for the storage devices being used. This file allows you to populate the **lustre_ost** and **lustre_mdt** tables using the **/usr/lib/lustre/load_storage.sh** script.

Skip this phase for a migration to **BAS5 for Xeon v1.2**, or if **BAS5 for Xeon v1.2** is being reinstalled, as the **/etc/lustre** directory will have been saved.

### 6.3.4 Configuring the High Availability services (Lustre High Availability clusters only)

**Lustre HA** Carry out the actions indicated in the *Checking the Cluster Environment* and the *Using Cluster Suite* sections, in the *Configuring High Availability for Lustre* chapter, in the **BAS5 for Xeon** *High Availability Guide*.

### 6.3.5 Lustre Pre Configuration Operations

1. Change the Lustre user password.

   The **lustre_mgmt rpm** creates the « lustre » user on the Management node with « lustre » as the password. It is strongly advised to change this password by running the following from the root command line on both Primary and Secondary Management nodes for High Availability systems.

   ```
   passwd lustre
   ```

   The « lustre » user is allowed to carry out most common operations on Lustre filesystems by using **sudo**. In the next part of this document, the commands can also be run as **lustre** user using the **sudo <command>**. For example:

   ```
   sudo lustre_util status
   ```

2. Set Lustre Network layers.

   **Lustre** runs on all network layers that can be activated in the kernel, for example **InfiniBand** or **Ethernet**.

By default the **Lustre** model file delivered is set to the **elan** nettype. The nettype parameter in the **/etc/lustre/models/fs1.lmf** file must be changed to **o2ib** for **InfiniBand** networks, and **tcp** for **Ethernet** networks.

If **Ethernet** is used as the **Lustre** network layer, and there are several physical links, you must select the links to be used by **Lustre**. This is done by editing the **/etc/modprobe.d/lustre** file.

**See** The *Lustre Operations Manual* from CFS (Section *Multihomed Servers*, sub-section *modprobe.conf*) available from http://manual.lustre.org/, for more details.

3. Set the **/etc/lustre/lustre.cfg** file.
   a. Edit the **/etc/lustre/lustre.cfg** file of the Management Node.
   b. Set **LUSTRE_MODE** to **XML**. (This should already have been done).
   c. Set **CLUSTERDB** to **yes** (if not already done).

**Lustre HA** Carry out the actions indicated in the *Installing the Lustre LDAP Directory* and the *Cluster DB Synchronisation using lustredbd* sections, in the *Configuring High Availability for Lustre* chapter, in the **BAS5 for Xeon** *High Availability Guide*.

### 6.3.6 Configuring the Lustre MGS service

**Important**

The **Lustre MGS** service must be installed and configured on the Management Node before Lustre is installed.

The Lustre **MGS** service is not managed by the **lustre_util** tool. It is an independent service which has to be run separately. You can only have one MGS running per node.

1. When you are configuring your **/etc/lustre/lustre.cfg** file, there are some fields that have to be filled in to link the MGS with the Lustre core.

**Important**

Before the **LUSTRE_MGS_HOST** and **LUSTRE_MGS_NET** fields are filled, check that the host node is valid by running the command **gethostip -dn <host_name>.** This will list the host name and its IP address. This is particularly recommended when there are multiple interfaces for a node.

- **LUSTRE_MGS_HOST**=name of the Management Node where the MGS service is installed. This value is used by the **lustre_util** tool to link the **MGS** with other Lustre entities, for example, **MDS, OSS**.
- **LUSTRE_MGS_NET**= the name of the network used to read the **MGS**, for example, TCP or o2ib. When the **o2ib** net type is used the **LUSTRE_MGS_HOST** name value has to be suffixed with'-ic0' which is hostname suffix for IB networks.

For example, if you need to use an **InfiniBand** network to reach the MGS entity that runs on the node **zeus6** you have to:
  - set LUSTRE_MGS_NET to o2ib
  - set LUSTRE_MGS_HOST to zeus6-ic0

– **LUSTRE_MGS_ABSOLUTE_LOOPBACK_FILENAME** = file for mgs loop device. The default is **/home/lustre/run/mgs_loop**.
When High Availability exists for the Management Node, select a directory which is shared for the Management Node pairs. This value is used by the **MGS** service when **lustre_util** is not used.

2. **Verify your network.**
**Lustre** requires IP interfaces to be in place. On your **Lustre MGS** node, make sure that **IPOIB** is configured if the **InfiniBand** modules are available.

3. **Introduction to the MGS service**.
MGS is delivered as a service matching the cluster suite layout. The service is located in /etc/init.d/mgs

```
service mgs help
```

**Usage: mgs {start|stop|restart|status|install|erase|reinstall|clear}**

Start the **MGS** service on this node

**start** Start the MGS service using the **mount.lustre** command
The mount point is: **/mnt/srv_lustre/MGS**
This returns 0 if successful or if the MGS service is already running

**stop** Stop the mgs service using the **umount.lustre** command. This returns 0 if successful or if the **MGS** service has already stopped.

**status** Status of the **MGS** service resulting from the **mount -t lustre** command. This returns 0 if successful.

**restart** Restart the **MGS** service using the stop and start target. This returns 0 if successful.

**install** Installs the **MGS** service if the service is not already installed or running. Creates a folder and file for the loopback device.
- Format using **mkfs.lustre**
- Size for loopback file is **512 MBs**
- Loopback file name is given by **/etc/lustre/lustre.cfg** file :
- target : **LUSTRE_MGS_ABSOLUTE_LOOPBACK_FILENAME**
- default value is : **/home/lustre/run/mgs_loop**

Returns 0 if successful.

**erase** Erase/Remove the MGS backend using the **rm** remove command on the loopback file. Check if service is stopped before. This returns 0 if successful.

**reinstall** Reinstall the MGS service using the erase and install target. Free the loopback reservation using the **losetup –d** command. This returns 0 if successful.

**clear** Clean the loopback map using **losetup -a** and **losetup –d** commands. Returns 0 if successful.

4. **Installation of the MGS service on the Management station** :
   You must apply this section before running **lustre** and before running the **lustre_util install** command. Ensure that the **lustre.cfg** file is completed correctly and dispatched. Use the **lustre_util set_cfg** tool.Run the command below to install the MGS service

```
service mgs install
```

```
--------------------------------------------------------------------
.mgs installed [OK]
```

5. **Start the MGS service on the Management Node.**

```
service mgs start
```

```
--------------------------------------------------------------------
Starting mgs: on xena0
mgs xena0 is not running
mgs started [OK]
--------------------------------------------------------------------
```

When there is no High Availability on the Management Node, the service must be started at boot time.

Run the command below in order to ensure that the **MGS** service restarts:

```
chkconfig --add mgs
```

## 6.3.7 Lustre Pre-Configuration Checks

Save the **lustre.cfg** file and quit the editor.

1. Once the **lustre.cfg** file has been edited copy it to the Secondary Management node for clusters which feature High Availability for the Management Node.

2. Use the **service mgs status** command to check that **mgs service** is running on the Management Node,

```
service mgs status
```

```
--------------------------------------------------------------------
/dev/loop0 on /mnt/srv_lustre/MGS type lustre (rw)

mgs xena0 is running
--------------------------------------------------------------------
```

3. Check the consistency of the database.

```
lustre_investigate check
```

This command checks which storage devices in the **lustre_ost** and **lustre_mdt** tables can be used. A clean output means that the command has been successful.

**See** The **lustre_investigate** man page or the **BAS5 for Xeon** *Administrator's Guide* for more details.

Run the command below to list the OSTs. There must have be at least one OST with **cfg_stat** set to "**available**":

```
lustre_ost_dba list
```

Run the command below to list the MDTs. There must have be at least one MDT with **cfg_stat** set to "**available**":

```
lustre_mdt_dba list
```

**Lustre HA** Carry out the actions indicated in the *Managing Lustre Failover Services on the I/O and Metadata Nodes - the lustre_migrate Tool* section, in the *Configuring High Availability for Lustre* chapter, in the **BAS5 for Xeon** *High Availability Guide*.

## 6.3.8 Configuring Lustre

1. **Configure Lustre on the I/O nodes.**

   Run the following command, and answer 'yes':

```
lustre_util set_cfg
```

An output similar to the following is displayed:

```
lustre.cfg copied on < I/O nodes >
snmpd enabled on < I/O nodes >
ldap database enabled on < mgmt node >
```

2. **Create the file system configuration.**

   The **/etc/lustre/models/fs1.lmf** file is a default model file which comes with the Lustre RPMs. It implements a file system which uses all the available OSTs and the first available MDT, with no failover. If you want to create more than one file system and/or with failover capability, refer to Bull **BAS5 for Xeon** *Administrator Guide* or to the **lustre_util** man page for more details about the Lustre model files.

   Run the following command:

```
lustre_util info -f /etc/lustre/models/fs1.lmf
```

This command prints information about the **fs1** file system. It allows you to check that the MDT and OSTs are actually those you want to use. Ensure that no warning occurs.

**Lustre HA** Carry out the actions indicated in the *Configuring File Systems for Failover* section, in the *Configuring High Availability for Lustre* chapter, in the **BAS5 for Xeon** *High Availability Guide*.

3. **Check what happened.**

   At this point it is possible to run the following command on a second terminal (checking terminal) to see what happened during the installation process.

```
watch lustre_util info -f all
```

   The following message should be displayed:

```
No filesystem installed
```

   It is also possible to look at http://<mngt_node>/lustre from a Web browser.

**See** The **lustre_util** man page for more information.

4. **Install the file system.**

**Important**

Do not perform this step when performing a software migration as the **Lustre** configuration details and data will have been preserved.

5. **Run the following command:**

```
lustre_util install -f /etc/lustre/models/fs1.lmf -V
```

   This operation is quite long as it formats the underlying file system (about 15 minutes for a 1 TB file system). Do not use the **-V** option if a less verbose output is required.

   At the top of the checking terminal, the following should appear:

```
Filesystem fs1:
   Cfg status  : formating
   Status      : offline
   Mounted     : 0 times
```

   Wait until the following appears:

```
Filesystem fs1:
      Cfg status  : installed
      Status      : offline
   Mounted     : 0 times
```

   The last line printed at the execution terminal must be:

```
Filesystem fs1 SUCCESSFULLY installed
```

6. **Enable the file system by running the following command:**

```
lustre_util start -f fs1 -V
```

This operation is quite long (about 10 minutes for a 1TB file system). Do not use the **-V** option if a less verbose output is required.

At the top of the checking terminal, the following should appear:

```
Filesystem fs1:
      Cfg status  : installed
      Status      : starting
      Mounted     : 0 times
```

Wait until the following appears:

```
   Filesystem fs1:
      Cfg status  : installed
      Status      : online
      Mounted     : 0 times
```

The "running status" of the OSTs/MDT must also be 'online'.

The last lines printed at the execution terminal must be:

```
   FILESYSTEMS STATUS
   +-----------+----------+--------+--------+----------------+
   |filesystem |  config  |running | number |   migration    |
   |           |  status  | status |of clts |                |
   +-----------+----------+--------+--------+----------------+
   |fs1        |installed |online  |0       |0 OSTs migrated |
   +-----------+----------+--------+--------+----------------+
```

7. **Mount the file system on clients.**

   Run the following command:

```
lustre_util mount -f fs1 -n <list_of_client_nodes_using_pdsh_syntax>
```

For example, if the client nodes are `ns0` and `ns2`, then run:

```
lustre_util mount -f fs1 -n ns[0,2]
```

At the top of the checking terminal, the following should appear:

```
   Filesystem fs1:
      Cfg status  : installed
      Status      : online
      Mounted     : 2 times
```

The last line printed at the execution terminal must be:

```
Mounting filesystem fs1 succeeds on ns[0,2]
```

**The file system is now available. As administrator it will be possible to create user directories and to set access rights.**

It is possible to check the health of the file system, at any time, by running:

```
lustre_util status
```

This will display a status, as below:

```
FILESYSTEMS STATUS
+-----------+----------+--------+--------+----------------+
|filesystem |  config  |running | number |   migration    |
|           |  status  | status |of clts |                |
+-----------+----------+--------+--------+----------------+
|fs1        |installed |online  |2       |0 OSTs migrated |
+-----------+----------+--------+--------+----------------+
  CLIENTS STATUS
+-----------+----------+
|filesystem |correctly |
|           | mounted  |
+-----------+----------+
|fs1        |ns[0,2]   |
+-----------+----------+
```

If more details are required, then run:

```
lustre_util all_info -f all
```

The file system health can also be checked in the **Nagios** view of the Management Node.

# Chapter 7. Installing Intel Tools and Applications

This chapter describes how to install tools or commercial software from CDs or supplier sites.

## 7.1 Intel Libraries Delivered

Some applications delivered with the Bull **XHPC** CD-ROM have been compiled with Intel compilers. The Bull **XHPC** CD-ROM installs the **intelruntime–<version>-Bull.X.x86_64.rpm**, which contains various free distribution Intel libraries that are needed for these applications to work on all node types (Management, I/O, Login, COMPUTEX and COMPUTE). These libraries are installed in the **/opt/intelruntime/<version>** folder, where version equals the compiler version number for these libraries. For example, for applications which have been compiled with version 10.1.011 compilers the folder is named 10.1.011.

The **/opt/intelruntime/<version>** path should be added to the **LD_LIBRARY_PATH** environment variable in the shell configuration file so that the applications delivered on the Bull **XHPC** CDROM can run.

If there is a desire to install a different version of an **Intel** compiler, then this has to be copied on to the other nodes, in order to ensure coherency. At the same time the path in the **LD_LIBRARY_PATH** variable has to be modified to include the new version reference.

## 7.2 Intel Compilers

Install the **Intel Compilers** as and when required. This is not necessary if they have been systematically deployed previously – see STEP 5 in Chapter 3. The compilers must be installed on the node which contains the Login functionality (this may be a dedicated node or one which is combined with the I/O and/or Management functionalities).

Follow the instructions written in the Bull notice supplied with the compiler

### 7.2.1 Fortran Compiler for Intel® 64 architecture (formerly Intel® EM64T)

#### Installation

Follow the instructions contained in the Bull notice, which is supplied with the **Intel** compiler provided by Bull and use the default path proposed by the installation routine.

### 7.2.2 C/C++ Compiler for Intel® 64 architecture (formerly Intel® EM64T)

#### Installation

Follow the instructions contained in the Bull notice, which is supplied with the **Intel** compiler provided by Bull and use the default path proposed during the installation routine.

## 7.3 Intel Debugger

The package used to install the Intel debugger is located in either the **Fortran** or C tar archive.

### Installation

Follow the instructions contained in the Bull notice, which is supplied with the **Intel** compiler provided by Bull.

## 7.4 Intel Math Kernel Library (MKL)

The **Intel MKL** libraries must be installed on Compute, Extended Compute and Login Nodes.

### Installation

An installation notice is supplied with the **Intel MKL** provided by Bull.

## 7.5 Intel Trace Tool

Intel Trace Tool is supplied directly by **Intel** to the customer. Intel Trace Tool uses the FlexLM license scheme. The recommended path for installation is **/opt/intel/itac/<rel number1>**.

Install it as follows:

```
cd /tmp
tar -zxvf /l_itac_<rel number 2>tar.gz
```

`<rel number 1>` and `<rel number 2>` represent the release numbers of the product.

- Run the installation command:

```
./install.sh
```

  Answer the questions with "**y**".

- Save the license in the **etc** subdirectory:

```
cp /license.dat ./etc/
```

- Run the command:

```
./install.sh
```

  Answer the questions with "**y**"

- Run the command

```
opt/intel/itac/rel_number_1/etc/itacvars.sh
```

For more details about the installation procedure you can read the *Intel® Trace Collector User's Guide* on the internet site:
http://www.intel.com/software/products/cluster

## 7.6 Updating Intel Compilers and BAS5 for Xeon v1.2

**BAS5 for Xeon V1.2** has been validated with **Intel C/C++** and **Fortran** version 10.1.011 compilers for **Linux.** It will work with later **10.x** compiler and **MKL** releases provided that the Bull **intelruntime-10.1.011 RPM** is NOT installed, and the **Intel** runtime for the compilers and MKL libraries, is made available for all the Compute or Extended Compute Nodes.

If the **intelruntime-10.1.011 RPM** has been installed it can be uninstalled using the following command:

```
rpm –u intelruntime-10.1.011 RPM
```

Two possible methods exist for updating compiler and MKL versions:

- Install the **Intel** compilers and **MKL** libraries on the reference COMPUTE or COMPUTEX Node and redeploy the reference node image using the **KSIS** tool.
- Install the Intel compilers on the Login Nodes. Then export the **/opt/intel** directory via **NFS** and mount it on the COMPUTE or COMPUTEX Nodes.

If an **Intel** license is not available for the node, the compiler will not work BUT the runtime libraries can be used by applications previously compiled with the compiler.

# Chapter 8. Installing and Configuring InfiniBand Interconnects

**Important**

**The information in this chapter only applies to switches with 3.x version firmware. Refer to the Voltaire documentation, available on the Bull *Voltaire Switches Documentation CD*, or from www.voltaire.com, if the firmware version is later, and/or the switch models are different.**

This chapter describes how to install and configure **InfiniBand** interconnects including **Voltaire®** devices (these vary according to the size and type of cluster) and **Mellanox ConnectX™** Interface Cards.

The following topics are described:

- 8.1 *Installing HCA-400 Ex-D and Mellanox ConnectXTM Interface Cards*
- 8.2 *Configuring the Voltaire ISR 9024 Grid Switch*
- 8.3 *Configuring Voltaire switches according to the Topology*
- 8.4 *Performance manager (PM) setup*
- 8.5 *FTP setup*
- 8.6 *The Group menu*
- 8.7 *Verifying the Voltaire Configuration*
- 8.8 *Voltaire GridVision Fabric Manager*
- 8.9 *More Information on Voltaire Devices*

## 8.1 Installing HCA-400 Ex-D and Mellanox ConnectX™ Interface Cards

**Note** Refer to the safety information prior to performing the installation.

1. Ensure that the host is powered down and disconnect the host from its power source.
2. Locate the **PCI-Express** slot and plug the Host Channel Adapter into the slot, handling the **HCA** carefully by the bracket.
3. Press the **HCA** firmly into the **PCI - Express** slot by applying pressure on the top edge of the bracket.
4. Re-install any fasteners required to hold the HCA in place.
5. Connect the **InfiniBand** cable to either of the HCA ports and to the switch.
6. Reconnect the host to its power source and power up the system.

## 8.2 Configuring the Voltaire ISR 9024 Grid Switch

### 8.2.1 Connecting to a Console

Connect the Management Node, with a terminal emulation program, to the RS-232 console interface according to the instructions in the *Hardware Installation Guide*. Make sure that the terminal emulation program is configured as follows:

| Setting | Value |
|---|---|
| Terminal Mode | VT-100 |
| Baud | 38400 |
| Parity | No Parity |
| Stop Bits | 1 Stop Bit |
| Flow Control | None |

Table 8-1. Voltaire ISR 9024 Switch Terminal Emulation Configuration

### 8.2.2 Starting a CLI Management Session using a serial line

To start a Command Line Interface management session for the switch via a HyperTerminal connection, do the following:

1. Connect the switch via its serial port, using the cable supplied by **Voltaire**.
2. Start the HyperTerminal client.
3. Configure the terminal emulation parameters as described in the section above.
4. Type in the appropriate password at the logon prompt. The Admin default password is: `123456`.

To change to Privileged mode:

1. Once in admin mode, enter: **enable.**
2. Enter the following password at the prompt: **voltaire**

### 8.2.3 Starting a CLI Management Session via Telnet

1. Establish a Telnet session with the Voltaire device.
2. At the Login prompt, type the user name: **admin**.
3. At the Password prompt, type the default password: **123456**.

To change to Privileged mode:

4. Once in admin mode, enter: **enable**.
5. Enter the following password at the prompt: **voltaire**
6. Enter the appropriate CLI commands to complete the required actions.

### 8.2.4 Configuring the Time and Date

Use the command sequence below to configure the time and date parameters for the switch. The time and date will appear on event reports that are time stamped.

1. Enter Privileged mode (from Exec mode).

```
enable <password>
```

2. Set the time and date. For example, time.8:22 AM; date, June 21, 2008.

```
clock set 062108222008
```

### 8.2.5 Hostname setup

#### 8.2.5.1 Names configuration menu

Enter the switch name configuration menu as follows:

```
ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
Connecting
```

```
switchname # config
switchname (config)# names
switchname (config-names)#
```

#### 8.2.5.2 Setting up the system name

The switch name can be set as follows:

```
switchname (config-names)# system-name set <switch hostname>
```

It can be checked as follows:

```
switchname (config-names)# system-name show
```

### 8.2.6 Networking setup

The following section describes how to set up the switch IP address for the **Ethernet** interface. The configuration of the IP address over the **Infiniband** network is not described.

**Note** The default IP address for a **Voltaire** switch is 192.168.1.2. If the switch cannot be reached using this address, then use a serial line (speed: 38600, no parity, 1 stop bit, no flow control).

#### 8.2.6.1 Networking configuration menu

Enter the networking configuration menu as follows:

```
ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
Connecting
```

```
switchname # config
switchname (config)# interface fast
switchname (config-if-fast)#
```

#### 8.2.6.2 Determining the current IP setup

The IP address that is currently configured can be seen as follows:

```
switchname (config-if-fast)# ip-address-fast show
```

```
fast interface ip is 172.20.2.20
ip mask is 255.255.0.0
broadcast ip is 172.20.255.255
management interface is eth1
link speed is auto-negotiation
The DHCP client is disabled
```

### 8.2.7 Setting up the switch IP address

The switch IP address is set as follows:

```
switchname (config-if-fast)# ip-address-fast set <ip address> <network mask>
```

Also make sure that the broadcast address is configured properly:

```
switchname (config-if-fast)# broadcast-fast set <broadcast IP address>
```

### 8.2.8 Route setup

#### 8.2.8.1 Route configuration menu

The route can be set from the following menu:

```
ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
Connecting
```

```
switchname # config
switchname (config)# route
switchname (config-route)#
```

#### 8.2.8.2 Setting up the route

Set the route as follows:

```
switchname (config-route)# default-gw fast set <gateway ip address>
```

Check that the route is fine:

```
switchname (config-route)# default-gw show
```

**Important**

**It is strongly advised to reboot the switch after modifying the route parameter.**

### 8.2.9 Routing Algorithms

The following routing algorithms are possible: Balanced-routing, Rearrangable, or Up-down.

```
switchname (config-sm)# sm-info algorithm set <algorithm>
```

- Balanced-routing is good for CLOS topologies when using a pruned network.
- Up-down is the best routing algorithm on fully non blocking networks.
- Rearrangeable routing may impact performance.

### 8.2.10 Subnet manager (SM) setup

#### 8.2.10.1 Subnet Manager Configuration menu

Enter the subnet manager configuration menu as follows:

```
ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
Connecting
```

```
switchname # config
switchname (config)# sm
```

```
switchname (config-sm)#
```

## 8.2.11 Configuring Passwords

Use the following procedure for configuring passwords for Exec and Privileged mode access to the RS-232 console interface and to the Ethernet management interface (used for establishing a CLI session via Telnet; see section 8.2.3 *Starting a CLI Management Session via Telnet*).

**Note** The default password for Privileged mode is `123456` and for Exec mode is `voltaire`.

1. Enter Privileged mode (from Exec mode).

```
enable <password>
```

2. Set the Privileged and Exec mode passwords

```
password update [admin | enable]
```

3. Exit Privileged mode.

```
exit
```

## 8.3 Configuring Voltaire switches according to the Topology

It is essential that the topology settings for the **Voltaire** switches are correct, otherwise the performance of the cluster will suffer. **InfiniBand** networks support **3-stage-CLOS** and **5-stage-CLOS** topologies.

- If the network consists of a single **ISR9024 [DM] Voltaire** switch, then it is not a CLOS network. The topology parameter is not taken into account in this case.
- If the network only uses **ISR9024 [DM] Voltaire** switches, the topology is most likely CLOS 3. While it is technically feasible to build a CLOS 5 network using these switches, it does not make much sense economically.
- If the network only uses **ISR9096 [DM], ISR9288 [DM], ISR2012 [DM]** chassis switches, the topology is most likely CLOS 3.
- If the network uses both kinds of switches, then the topology is certainly CLOS 5.

**Important**

**The System Administrator should know which topology applies to his cluster. If not contact Bull for more information.**

### Pre-requisite

All the following switch configuration commands take place inside the **config-sm** menu. To enter this menu, proceed as follows:

```
ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
connecting
```

```
switchname # config
switchname (config)# sm
switchname (config-sm)#
```

## 8.3.1 Setting the Topology CLOS stage

1. Use the **sm-info show** command and look at the **topology** and **active topology** fields to check which topology setting is in place for the cluster. This should match the setting required for the cluster.

```
<switchname>(config-sm)# sm-info show
```

```
subnet manager info is:
smName=
port guid= 0008f1040041254a
topology= 5-stage-CLOS
active topology= 5-stage-CLOS              <==========
algorithm= up-down
active algorithm= up-down
```

```
sm KEY = 0000000000000000
sm priority = 3
sm sweep interval (seconds)= 15
sm verbosity mode = error
sm topology verbosity  = none
sm mads-pipeline = 16
sm polling-retries = 12
sm activity = 98663
sm state = master
sm mode = enable
sm LMC = 0
sm hoq = 16
sm slv = 16
sm mopvl = vl0-14
subnet-prefix = 0xfe80000000000000
port-state-change-trap = enable
bad ports mode = disable
pm mode = enable
grouping mode = enable
```

2. To change the topology setting to 3 stage CLOS run the command below;

```
<switchname>(config-sm)# sm-info topology set 3
```

or to change the topology setting to 5 stage CLOS.

```
<switchname>(config-sm)# sm-info topology set 5
```

The changes will take effect after the next Fabric Reconfiguration.

3. For both CLOS 3 and CLOS 5 topologies, some of the switches or switch ASICs will need to be declared as spines, as shown in the sections which follow.

## 8.3.2 Determining the node GUIDs

**Important**

**Before starting the Administrator should know which Voltaire switches are the top switches. Contact Bull if this information is not available.**

All the top switches must be defined as spines. Each top switch is identified using its node GUID. There are 2 possible cases:

- The top switch is an **ISR9024 [DM]**
- The top switch is not an **ISR9024 [DM]**, i.e. the switch is a chassis switch (**ISR9096 [DM], ISR9288 [DM], ISR2012**, etc).

### 8.3.2.1 Determining the node GUIDs for a Voltaire ISR9024 [DM] switch

Look for the NODEGUID fields of all top switches.

1. For **Voltaire ISR 9024** switches make a note of the NODEGUID identifier which is shown when the **ibs topo action** command is run, as shown in the example below. See Chapter 2 in the **BAS5 for Xeon** *Maintenance Guide* for more information on the **IBS** tool.

```
ibs -a topo -s <subnet manager IP address or hostname>
```

```
-------------------------------------------------------------------------------------------
DESCRIPTION          | HOSTNAME          | NODEGUID           | NODELID | LOCATION       |
-------------------------------------------------------------------------------------------
ISR9024D Voltaire    | iswu0c0-2         | 0x0008f10400411946 | 0x0017  | [A,2] RACK1/B  |
-------------------------------------------------------------------------------------------
```

In this case, the node GUID is 0x0008f10400411946.

### 8.3.2.2 Determining the node GUIDs for a chassis switch

Find the 'Spine' lines and look out for the NODEGUID field.

Use the IBS tool, as below, to identify the node GUID:

```
ibs -a topo -s <subnet manager IP address or hostname>
```

```
-------------------------------------------------------------------------------------------
 PART         | ASIC     | NODESYSTEMGUID      | NODEGUID           | NODELID | CHASSIS     |
-------------------------------------------------------------------------------------------
 Spine 4      | 3        | 0x0008f10400401e60 | 0x0008f10400401e1b | 0x0001  | iswu0c0     |
-------------------------------------------------------------------------------------------
```

In this case, the node guid is 0x0008f10400401e1b. Repeat for all spines on all switches.

Alternatively, the IBS tool (version > 0.2.8) can be used to produce the same information as follows:

```
[user@host ~]# ibs -a showspines -s <subnet manager IP address or hostname>
```

```
Available spines:
0x0008f10400401e1b
0x0008f10400401e1c
```

## 8.3.3 Adding new Spines

Each spine is specified using an (index, nodeguid) tuple as follows.

Note that the index can be any positive integer and its value does not impact performance:

```
switchname (config-sm)# spines add 1 0x0008f10400411946
```

The change will take effect after the next fabric reconfiguration.

**Note** If the switch firmware is **Voltaire** version 3.X , remove the '0x' part of the node GUID, as shown below. For interconnects which use **Voltaire** 4.x firmware you should always prepend 0x to the NodeGUID.

```
switchname (config-sm)# spines add 1 0008f10400411946
```

The change will take effect after the next reconfiguration of the fabric. Repeat this procedure for all spines.

**The NodeGUID has to be declared for each spine included in the Switch topology by running the add option separately for each spine.**

**Note** An **ISR 9288/2012** switch has 4 fabric boards, each of them using 3 ASICs, so these types of switches have 4 x 3 = 12 spines.

This will provide output similar to that below for a cluster with 12 spines:

### 8.3.3.1 Listing configured spines

Once the NodeGUIDS have been declared, check that the GUID details have been updated by running the command below.

```
switchname (config-sm)# spines show
```

#### Sample output for 1 spine

```
entry  GUID
|------|----------------
1      0008f10400411946
```

#### Sample output for 12 spines

```
entry   GUID
|------|----------------
1      0x0008f10400401e61
2      0x0008f10400401e62
3      0x0008f10400401e63
4      0x0008f104004018d5
5      0x0008f104004018d6
6      0x0008f104004018d7
7      0x0008f10400401e4d
8      0x0008f10400401e4e
9      0x0008f10400401e4f
10     0x0008f10400401e19
11     0x0008f10400401e1a
12     0x0008f10400401e1b
```

Alternatively, the IBS tool (version > 0.2.8) can be used to produce the same information as follows:

```
ibs -a showspines -s <subnet manager IP address or hostname>
```

```
Spine nodeguids currently configured in the subnet manager:
0x0008f10400411946
```

### 8.3.3.2 Activating changes

Now that the topology and the spines have been defined, activate the changes as follows:

```
switchname(config-sm)# sm-info sm-initiate-fabric-configuration set
switchname(config-sm)# sm-info sm-initiate-fabric-reconfiguration set
switchname(config-sm)# sm-info sm-initiate-routing-reconfiguration set
```

**Note** These commands will interrupt all **InfiniBand** traffic, so be sure to stop all the jobs that are running before using them.

Confirm that the new settings have been implemented by running the **sm-info show** command:

```
switchname(config-sm)# sm-info show
```

#### Example output

```
subnet manager info is:
smName= zeus
port guid= 0008f1040041254a
topology= 3-stage-CLOS
active topology= 3-stage-CLOS
algorithm= up-down
active algorithm= up-down
sm KEY = 0000000000000000
sm priority = 3
sm sweep interval (seconds)= 15
sm verbosity mode = error
sm topology verbosity  = none
sm mads-pipeline = 16
sm polling-retries = 12
sm activity = 66049
sm state = master
sm mode = enable
sm LMC = 3
sm hoq = 16
sm slv = 16
sm mopvl = vl0-14
subnet-prefix = 0xfe80000000000000
port-state-change-trap = enable
bad ports mode = disable
pm mode = enable
grouping mode = enable
```

## 8.4 Performance manager (PM) setup

The performance manager is a daemon running on a managed switch that collects error and bandwidth statistics. It is essential to ensure that it is running with the correct setup.

### 8.4.1 Performance manager menu

Enter the FTP configuration menu as follows:

```
ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
Connecting
```

```
switchname # config
switchname (config)# pm
switchname (config-pm)#
```

### 8.4.2 Activating the performance manager

Performance Manager is activated as follows:

```
switchname (config-pm)# pm mode set enable
```

Once activated configure the performance manager to enable reporting:

```
switchname (config-pm)# pm report-enable set enable
```

Check that everything is OK by using the **pm show** command:

```
switchname (config-pm)# pm show
```

```
pm mode                                          enable
Trap mask                                        [29294560]
Polling interval                                 180
Scope                                            all
Reset-scope                                      all
Counter operation                                delta
Symbol error counter threshold                   200
Link error recovery counter threshold            1
Link downed counter threshold                    1
Port rcv errors threshold                        5
Port rcv remote physical errors threshold        5
Port rcv switch relay errors threshold           0
Port xmit discards threshold                     5
Port rcv constraint errors threshold             5
Port xmit constraint errors threshold            5
Local link integrity errors threshold            5
Excessive buffer overrun errors threshold        5
Vl15 dropped threshold                           5
Port xmit data threshold                         0
Port rcv data threshold                          0
Port xmit pkts threshold                         0
Port rcv pkts threshold                          0
Report mode                                      enable
alert join                                       enable
alert ATS                                        enable
```

## 8.5 FTP setup

The switch management software allows the administrator to upload or download files to or from the switch. For this to happen it is vital to have a working FTP setup. The FTP server is installed automatically on the Management Node but is not active. Run the command below to start the service.

```
service vsftpd start
```

### 8.5.1 FTP configuration menu

Enter the FTP configuration menu as follows:

```
ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
Connecting
```

```
switchname # config
switchname (config)# ftp
switchname (config-ftp)#
```

### 8.5.2 Setting up FTP

**Notes**

- The FTP server must have been configured on the Management Node.
- The username and password shown below are examples only. Use the username and password which applies to your cluster.

The following settings define the node 172.20.0.102 as the FTP server. The switch logs onto this server using Joe's account with the specified password (yummy).

```
switchname (config-ftp)# server 172.20.0.102
switchname (config-ftp)# username joe
switchname (config-ftp)# password yummy
```

## 8.6 The Group menu

The group menu is used to import host details from a **group.csv** file. The **group.csv** file is used to supply data to the switch subnet manager. This data is used to create the mapping GUID. Therefore recognisable hostnames should be used to make switch identification easier. In addition, it also contains geographical information that may be useful when using **Voltaire Fabric Manager**.

#### Sample from an existing group.csv:

```
Type,Id/guid,name,Don't show in group,Rack Id,Location in rack,U
HCA,2c9020024b8f4,zeus14,0,2,0,U
```

### 8.6.1 Group Configuration menu

Enter the group configuration menu as follows:

```
ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
Connecting
```

```
switchname # config
switchname (config)# group
switchname (config-group)#
```

### 8.6.2 Generating a group.csv file

The **group.csv** file can be generated automatically by using the **IBS** command as follows:

```
[user@host /tmp ] ibs -a group -s switchname -NE
```

```
Successfully generated configuration file group.csv
To update a managed switch with a firmware version 4.X, proceed as
follows:
 - Log onto the switch
 - Enter the 'enable' mode
 - Enter the 'config' menu
 - Enter the 'group' menu
 - Type the following command: group import /tmp
To update a managed switch with a firmware version 3.X, proceed as
follows:
 - Log onto the switch
 - Enter the 'enable' mode
 - Enter the 'config' menu
 - Enter the 'ftp' menu
 - Type the following command: importFile group /tmp
 - Leave the 'ftp' menu by typing 'exit'
 - Enter the 'group' menu
 - Type the following command: group import
```

### 8.6.3 Importing a new group.csv file on a switch running Voltaire 3.X firmware

Assuming the FTP server is set up properly, import the **group.csv** file located in **/tmp**:

```
switchname (config-ftp)# importFile group /tmp
```

**Note** This action takes place using the **config-ftp** menu.

Once this is done, enter the group menu and import the file as follows:

```
switchname (config-group)# group import
```

```
Summary report:
```

```
     Racks         :3
     Elements      :20
     Normal events :3
     Warning events:18
     Error events  :0
```

### 8.6.4 Importing a new group.csv file on a switch running Voltaire 4.X firmware

Assuming the FTP server is set up properly, import the **group.csv** file located in **/tmp**:

```
switchname (config-group)# group import /tmp
```

```
Summary report:
     Racks         :3
     Elements      :20
     Normal events :3
     Warning events:18
     Error events  :0
```

## 8.7 Verifying the Voltaire Configuration

The following Command Line Interface commands can be used to verify basic system parameters.

1. To display the version of the current software.

```
version show
```

2. To display the **ftp** server configuration.

```
ftp show (Optional)
```

3. To display the management interface IP address and configuration.

```
fast-interface show
```

4. To display the system clock.

```
clock show
```

5. To check the hardware including serial numbers, etc.

```
vital product data
vpd show
```

## 8.8 Voltaire GridVision Fabric Manager

For details of configuring routing using the **GridVision Fabric Manager** GUI see section 12.6 in the **Voltaire® GridVision™ Integrated User Manual for Grid Directors ISR 9096 and ISR 9288 and the Grid Switch ISR 9024.** This is included on the **Voltaire** documentation CD provided.

## 8.9 More Information on Voltaire Devices

**See** The manuals available on the Bull *Voltaire Switches Documentation CD,* or from www.voltaire.com, for more information regarding other switch models and for switches which use a firmware later than version 3.x.

**Note** For more information on the **SLURM** Resource Manager used in conjunction with **InfiniBand** stacks and Voltaire switches, see the **BAS5 for Xeon** *Administrator's Guide* and the **BAS5 for Xeon** *User's Guide*.

# Chapter 9. Configuring Switches and Cards

This chapter describes how to configure **BAS5 for Xeon** switches and cards.

The following topics are described:

- 9.1 *Configuring Ethernet Switches*
- 9.2 *Configuring a Brocade Switch*
- 9.3 *Configuring Voltaire Devices*
- 9.4 *Installing Additional Ethernet Boards*

## 9.1 Configuring Ethernet Switches

The Ethernet switches are configured automatically using the ClusterDB database information and the configuration file- see section 9.1.5 *Ethernet Switches Configuration File*.

### Prerequisites

- The Management Node must be installed. In particular, the Ethernet interface of the Administration Network and its alias must be configured and the **netdisco** package installed.
- The **ClusterDB** database must be preloaded and reachable.
- **CISCO** switches must remain as configured initially (factory settings). **Foundry Network** switches must have the default IP address preinstalled (see section 9.1.6 Ethernet Switches Initial Configuration)

### 9.1.1 Ethernet Installation scripts

The tool is supplied in the form of a RPM package (**ethswitch-tools1.0-0.Bull.noarch.rpm**) on the Cluster Management CD. It should be installed on the Management Node.

This package includes the following scripts:

**/usr/sbin/swtAdmin**: The main script used to install switches

**/usr/sbin/swtConfig**: A script that enables configuration commands to be run on the switches.

Also, the package includes the **/usr/lib/clustmngt/ethswitch-tools** directory which contains the following directories:

**bin** Perl scripts, called by the **swtAdmin** main script.

**lib** The libraries required to execute the scripts.

**data** The configuration file and DTD files.

## 9.1.2 swtAdmin Command Option Details

```
/usr/sbin/swtAdmin auto|step-by-step|generate|preinstall|
                   netdisco|mac-update|install|save|clear
               [--switch_number <number of new switches> ]
               [--netaddress <network ip for temporary config.> ]
               [--netmask <netmask for temporary configuration> ]
              [--network <admin|backbone> ]
               [--first <device name to start netdisco> ]
               [--dbname <database name> ]
               [--logfile <logfile name> ]
               [--verbose ]  [--help ]
```

### Example

```
/usr/sbin/swtAdmin auto --switch_number 4 --network backbone
```

### Actions

| | |
|---|---|
| **generate** | Generate configuration files |
| **preinstall** | Copy configuration files in the **/tftpboot** and restart **DHCPD** for the pre-installation of the switches |
| **netdisco** | Run **netdisco** in order to discover new switches |
| **mac-update** | Update database with the MAC address of the new switches |
| **install** | Install new switches |
| **save** | Save the configuration of the new switches |
| **auto** | Full configure and installation of switches |
| **step-by-step** | Interactive configuration and installation of switches |
| **clear** | Delete temporary configuration files |

### Options

| | |
|---|---|
| **help** | Display this message |
| **dbname** | Specifies the name of the database (default value: ClusterDB ) |
| **verbose** | Debug mode |
| **logfile** | Specifies the **logfile** name (default **/var/log/switchcfg.log**) |
| **switch_number** | Number of switches to install (default 1) |
| **first** | Specifies the IP address or name of device to start **netdisco** |
| **netaddress** | Specifies the network IP to use for the pre-install configuration |
| **netmask** | Specifies the netmask to use for the pre-install configuration |
| **network** | Specifies the type of network to be installed, admin or backbone |

## 9.1.3 Automatic Installation andConfiguration of the Ethernet Switches

The Ethernet switches can be configured automatically by running the command:

```
swtAdmin auto
```

All the steps (1–6), below, in the *Ethernet Switch Configuration Procedure* are executed in order, with no user interaction. If the automatic installation fails at any stage, you will only need to execute the steps which remain (including the one that failed).

Alternatively, the switches can be installed and configured interactively by using the command below:

```
swtAdmin step-by-step --switch_number <number_of_new_switches>
```

All the installation and configuration steps (1-6) are executed in order, but the user is asked to continue after each one.

## 9.1.4 Ethernet Switch Configuration Procedure

1. **Generating Configuration Files**

There are two kinds of configuration files: (1) files for the temporary configuration of the network and **DHCPD** services on the Service Node and (2) configuration files for the switches.

The switch configuration files are generated by running the command:

```
swtAdmin generate [--dbname <database name> ]
          [--netaddress <network ip for temporary config.> ]
          [--netmask <netmask for temporary configuration> ]
          [--network <admin|backbone> ]
          [--logfile <logfile name> ]
          [--verbose ]  [--help ]
```

While this command is being carried out the following message will appear.

```
Generate configuration files
/tmp/CfgSwitches/eswu0c1-confg
/tmp/CfgSwitches/eswu1c0-confg
/tmp/CfgSwitches/eswu1c1-confg
Temporary configuration files will start
 with 192.168.101.1 ip address (255.255.255.0 netmask)
```

2. **Pre-installation of switches**

At this stage, the following actions are carried out:

- Temporary configuration of the **eth0** network interface aliases and reconfiguration of the DHCPD service on the Service Node
- The configuration files are copied to the **/tftpboot/** directory
- The **DHCP** service is reconfigured and restarted

These actions are carried out by running the command:

```
swtAdmin preinstall [--dbname <database name> ]
          [--network <admin|backbone> ]
          [--logfile <logfile name> ]
          [--verbose ]  [--help ]
```

While this command is being carried out the following message will appear.

```
Pre-installation of switches
copy configuration files in /tftpboot/ directory
```

```
WARNING: we are looking for uninstalled switches. Please wait ...
Pre-installed X new switches.
```

**Note** After this step has finished, the switches will use the temporary configuration.

3. **Discovering new switches on the network**

If the cluster includes more than one switch, the **netdisco** application runs automatically in order to discover the network topology.

This is carried out by running the command:

```
swtAdmin netdisco [--first <device name to start netdisco> ]
                  [--network <admin|backbone> ]
                  [--dbname <database name> ]
                  [--logfile <logfile name> ]
                  [--verbose ]  [--help ]
```

While this command is being carried out a message similar to the one below will appear.

```
Discover new switches on the network
clear netdisco database
network discovering by netdisco application starting from
192.168.101.5 ip
WARNING: not all new switches has been discoved, retry ...
netdisco discovered X new devices.
```

4. **Updating MAC address in the eth_switch table**

When the topology has been discovered it is compared with the database topology. If there are no conflicts, the corresponding MAC addresses of switches are updated in the **eth_switch** table of the database.This is done by running the command:

```
    swtAdmin mac-update [--dbname <database name> ]
                [--logfile <logfile name> ]
                [--verbose ]  [--help ]
```

The following message will appear:

```
Update MAC address in the eth_switch table
Updating mac address values in clusterdb database ...
```

5. **Restarting Switches and final Installation Configuration**

At this step, all the switches are restarted and their final configuration is implemented by **TFTP** according to the parameters in the **DHCP** configuration file.

The **DHCP** configuration file is regenerated and will now include the MAC addresses of the switches, obtained during the previous step.

This is carried out by running the command:

```
    swtAdmin install [--dbname <database name> ]
                [--network <admin|backbone> ]
```

```
              [--logfile <logfile name> ]
              [--verbose ]  [--help ]
```

This will display a message similar to that below:

```
Final install and restart dhcp service
stop the dhcpd service
Shutting down dhcpd:                                        [  OK  ]
Installing switches ...
installing eswu1c0 switch (192.168.101.5 fake ip)
installing eswu0c0 switch (192.168.101.4 fake ip)
installing eswu1c1 switch (192.168.101.3 fake ip)
installing eswu0c1 switch (192.168.101.2 fake ip)
installed eswu1c0 switch
installed eswu0c0 switch
installed eswu1c1 switch
installed eswu0c1 switch
switches installed.
dbmConfig configure --service sysdhcpd --force --nodeps --dbname
clusterdb
Tue Oct 16 12:48:33 2007 NOTICE: Begin synchro for sysdhcpd
Shutting down dhcpd: [FAILED]
Starting dhcpd: [  OK  ]
Tue Oct 16 12:48:34 2007 NOTICE: End synchro for sysdhcpd
```

6. **Delete the temporary configuration files**

```
swtAdmin clear
```

7. **Saving the switches configuration**

Finally, when the switches have been installed, the configuration parameters will be stored locally in their memory and also sent by TFTP to the Management Node **/tftpboot** directory.

This is carried out by running the command:

```
   swtAdmin save [--dbname <database name> ]
              [--logfile <logfile name> ]
              [--verbose ]  [--help ]
```

This will display a message similar to that below:

```
Save configuration of switches
Saving switches configuration ...
saving configuration of eswu0c0 switch
saving configuration of eswu0c1 switch
saving configuration of eswu1c1 switch
saving configuration of eswu1c0 switch
saved configuration of eswu0c0 switch
saved configuration of eswu0c1 switch
saved configuration of eswu1c1 switch
saved configuration of eswu1c0 switch
save done.
```

8. **Checking the configuration of a switch**

The configuration of a switch is displayed by running the command:

```
swtConfig status --name <name_of_switch>
```

## 9.1.5 Ethernet Switches Configuration File

This file describes the parameters used to generate the switches configuration file.

A configuration file is supplied with the package as **/usr/lib/clustmngt/ethswitch-tools/data/cluster-network.xml**. The file structure is defined by **/usr/lib/clustmngt/ethswitch-tools/data/cluster-network.dtd** file.

The file contains the following parameters:

```
<!DOCTYPE cluster-network SYSTEM "cluster-network.dtd">
<cluster-network>
    <mode type="any">
        <login acl="yes" />
        <netadmin name="admin" />
        <vlan id="1" type="admin" dhcp="yes" svi="yes" />
        <mac-address logger="yes" />
        <logging start="yes" level="warnings" facility="local0" />
        <ntp start="yes" />
     </mode>
</cluster-network>
```

It specifies that:

- Only the workstations of the administration network are allowed to connect to the switches
- **DHCP** requests are forwarded
- The Management IP address is configured
- Log warnings are sent to the node service **syslog** server
- The switches system clock is synchronized with the **NTP** server for the node

For clusters configured with **VLAN** (Virtual Local Area Network,) or with the virtual router configuration, additional parameters must be defined using the **/usr/lib/clustmngt/ethswitch-tools/bin/config** script.

## 9.1.6 Ethernet Switches Initial Configuration

### 9.1.6.1 CISCO Switches

**CISCO** switches must be reset to the factory settings. This is done manually.

1. **Hardware reinitialization**

Hold down the mode button located on the left side of the front panel, as you reconnect the power cable to the switch.

For **Catalyst 2940, 2950** Series switches, release the Mode button after approximately 5 seconds when the Status **(STAT) LED** goes out. When you release the Mode button, the **SYST LED** blinks amber.

For **Catalyst 2960, 2970** Series switches, release the Mode button when the **SYST LED** blinks amber and then turns solid green. When you release the Mode button, the **SYST LED** blinks green.

For **Catalyst 3560, 3750** Series switches, release the Mode button after approximately 15 seconds when the **SYST LED** turns solid green. When you release the Mode button, the **SYST LED** blinks green.

2. **From a serial or Ethernet connection**

   Enter the following commands:

```
switch>enable
```

   Enter the password [admin] when requested

```
switch#delete flash:/config.text
```

   Answer the default questions (ENTER)

```
switch#reload
```

   Confirm without saving (ENTER).

   Ignore the question *"Would you like to enter the initial configuration dialog? [yes/no]"* and disconnect.

### 9.1.6.2 Foundry Network Switches

**Foundry Network** switches must be configured with the IP address: 192.168.1.200/24.

1. **Erase the configuration**

   From a serial or Ethernet connection enter the following commands:

```
switch>enable
```

   Enter the password [admin] when requested

```
switch#erase startup-config
```

   Answer the default questions (ENTER)

```
switch#reload
```

   Confirm without saving (ENTER).

2. **Configure the 192.168.1.200/24 IP address**

```
FLS648 Switch>enable
No password has been assigned yet...
```

```
FLS648 Switch#configure terminal
FLS648 Switch(config)#
```

a. on FastIron **FLS624** or **FLS648** models:

```
FLS648 Switch(config)#ip address 192.168.1.200 255.255.255.0
FLS648 Switch(config)#end
FLS648 Switch#write mem
```

b. on BigIron **RX4**, **RX8** and **RX16** models:

```
RX Switch(config)#vlan 1
RX Switch(config-vlan-1)# router-interface ve 1
RX Switch(config-vlan-1)#interface ve 1
RX Switch(config-vif-1)#ip address 192.168.1.200 255.255.255.0
RX Switch(config-vif-1)# end
RX Switch# write mem
```

## 9.1.7 Basic Manual Configuration

Please use this method when configuring the **Foundry Network** switches initially with the IP address 192.168.1.200/24 or for a temporary configuration of an Ethernet switch (**Cisco** or **Foundry**).

### Pre-Requisites

Before an Ethernet switch can be configured ensure that the following information is available:

- The name of the switch
- The IP address of the switch
- The IP address of the Netmask
- Passwords for the console port and the enable mode. These must be consistent with the passwords stored in the **ClusterDB** database.

1. **Connect the Console port of the switch to the Linux machine**

   Using a serial cable, connect a free serial port on a Linux machine to the CONSOLE port of the switch. Make a note of the serial port number, as this will be needed later.

2. **From the Linux machine establish a connection with the switch:**
   - Connect as **root**.
   - Open a terminal.
   - In the **/etc/inittab** file, comment the **tty** lines that enable a connection via the serial port(s) ; these lines contain `ttyS0` and `ttyS1`:

     ```
     # S0:2345:respawn:/sbin/agetty 115200 ttyS0
     # S1:2345:respawn:/sbin/agetty 115200 ttyS1
     ```

   Run the command:

```
kill -1 1
```

Connect using one of the commands below:

– If the serial cable connects using port 0, then run:

```
cu -s 9600 -l /dev/ttyS0
```

– If the serial cable connects using port 1, then run:

```
cu -s 9600 -l /dev/ttyS1
```

Enter 'no' to any questions which may appear until the following message, below, is displayed.

```
Connected.
Switch>
```

### 9.1.7.1 Configuring a CISCO Switch

1. Set the enable mode:

```
Switch>enable
```

2. Enter configuration mode:

```
Switch#configure terminal
Enter configuration commands, one per line.  End with CNTL/Z.
Switch(config)#
```

3. Set the name of the switch in the form: *hostname <switch_name>*. For example:

```
Switch(config)#hostname myswitch
myswitch(config)#
```

4. Enter the **SVI Vlan 1** interface configuration mode:

```
myswitch(config)#interface vlan 1
myswitch(config-if)#
```

5. Assign an IP address to the **SVI** of Vlan 1, in the form:
*ip address <ip : a.b.c.d> <netmask : a.b.c.d>*

```
myswitch(config-if)#ip address 10.0.0.254 255.0.0.0
myswitch(config-if)#no shutdown
```

6. Exit the interface configuration:

```
myswitch(config-if)#exit
myswitch(config)#
```

7. Set the *portfast* mode as the default for the spanning tree:

```
myswitch(config)#spanning-tree portfast default
```

```
%Warning: this command enables portfast by default on all interfaces.
You should now disable portfast explicitly on switched ports leading
to hubs, switches and bridges as they may create temporary bridging
loops.
```

8. Set a password for the enable mode. For example:

```
myswitch(config)#enable password myswitch
```

9. Set a password for the console port:

```
myswitch(config)#line console 0
myswitch(config-line)#password admin
myswitch(config-line)#login
myswitch(config-line)#exit
```

10. Enable the telnet connections and set a password:

```
myswitch(config)#line vty 0 15
myswitch(config-line)#password admin
myswitch(config-line)#login
myswitch(config-line)#exit
```

11. Exit the configuration:

```
myswitch(config)#exit
```

12. Save the configuration in RAM:

```
myswitch#copy running-config startup-config
```

13. Update the switch boot file on the Management Node

    Run the following commands from the Management Node console.

```
touch /tftpboot/<switch_configure_file>
chmod ugo+w /tftpboot/< switch_configure_file>
```

**Note** The switch configure file name must include the switch name followed by '**-config**', for example, **myswitch-config**.

14. Save and exit the switch configuration from the switch prompt.

```
myswitch#copy running tftp
myswitch#exit
```

    Enter the information requested for the switch. For the **tftp** server, indicate the IP address of the Service Node, which is generally the **tftp** server.

15. Disconnect the **CISCO** Switch

Once the switch configuration has been saved and the Administrator has exited from the interface it will then be possible to disconnect the serial line which connects the switch to the Linux Management Node.

16. You can check the configuration as follows:

    From the Management Node run the following command:

```
telnet 10.0.0.254
```

    Enter the password when requested.

    Set the enable mode

```
enable
```

    Enter the password when requested.

    Display the configuration with the show configuration command. An example is shown below:

```
#show configuration
```

```
Using 2407 out of 65536 bytes
!
version 12.2
no service pad
service timestamps debug uptime
service timestamps log uptime
no service password-encryption
!
hostname eswu0c1
!
enable secret 5 $1$ljvR$vnD1S/KOUD4tNmIm.zLTl/
!
no aaa new-model
ip subnet-zero
!
no file verify auto
spanning-tree mode pvst
spanning-tree portfast default
spanning-tree extend system-id
!
vlan internal allocation policy ascending
!
 interface GigabitEthernet0/1
!
interface GigabitEthernet0/2
!
interface GigabitEthernet0/3
!
interface GigabitEthernet0/4
!
interface GigabitEthernet0/5
!
interface GigabitEthernet0/6
!
interface GigabitEthernet0/7
!
interface GigabitEthernet0/8
!
interface GigabitEthernet0/9
!
interface GigabitEthernet0/10
!
interface GigabitEthernet0/11
!
interface GigabitEthernet0/12
!
interface GigabitEthernet0/13
```

```
!
interface GigabitEthernet0/14
!
interface GigabitEthernet0/15
!
interface GigabitEthernet0/16
!
interface GigabitEthernet0/17
!
interface GigabitEthernet0/18
!
interface GigabitEthernet0/19
!
interface GigabitEthernet0/20
!
interface GigabitEthernet0/21
!
interface GigabitEthernet0/22
!
interface GigabitEthernet0/23
!
interface GigabitEthernet0/24
!
interface Vlan1
 ip address 10.0.0.254 255.0.0.0
 no ip route-cache
!
ip http server
logging history warnings
logging trap warnings
logging facility local0
snmp-server community public RO
!
control-plane
!
line con 0
 password admin
 login
line vty 0 4
 password admin
 login
line vty 5 15
 password admin
 login
!
end
```

### 9.1.7.2 Configure a Foundry Networks Switch

The following procedure works for the **FastIron** and **BigIron** models

1. Set the enable mode:

```
FLS648 Switch>enable
No password has been assigned yet...
FLS648 Switch#
```

2. Enter the configuration mode:

```
FLS648 Switch#configure terminal
FLS648 Switch(config)#
```

3. Set the name of the switch in the form: *hostname <switch_name>*. For example:

```
FLS648 Switch(config)#hostname myswitch
myswitch(config)#
```

4. Assign a management IP address, in the form:

   a. on **FastIron FLS624** or **FLS648** models

      - Assign IP address to the switch:
        *ip address <ip : a.b.c.d> <netmask : a.b.c.d>*

```
myswitch(config)#ip address 10.0.0.254 255.0.0.0
myswitch(config)#
```

   b. on **BigIron RX4**, **RX8** and **RX16** models

      - Enter the **Vlan 1** interface configuration mode:

```
myswitch(config)#vlan 1
myswitch(config-vlan-1)#
```

      - Set the corresponding virtual interface (this allows the management IP address to be configured)

```
myswitch(config-vlan-1)#router-interface ve 1
```

      - Enter the virtual interface **ve 1** interface configuration mode:

```
myswitch(config-vlan-1)#interface ve 1
myswitch(config-vif-1)#
```

      - Assign an IP address to the virtual interface ve 1:
        *ip address <ip : a.b.c.d> <netmask : a.b.c.d>*

```
myswitch(config-vif-1)#ip address 10.0.0.254 255.0.0.0
```

      - Exit the interface configuration:

```
myswitch(config-vif-1)#exit
myswitch(config)#
```

5. The *portfast* mode for the spanning tree is the default mode:

```
myswitch(config)# fast port-span
```

6. Set a password for the enable mode. For example:

```
myswitch(config)#enable password myswitch
```

7. Enable the telnet connections and set a password:

```
myswitch(config)# enable telnet password admin
```

8. Exit the configuration:

```
myswitch(config)#exit
```

9. Save the configuration in RAM:

```
myswitch#write memory
```

10. Update the switch boot file on the Management Node

11. Run the following commands from the Management Node console.

```
 touch /tftpboot/<switch_configure_file>
chmod ugo+w /tftpboot/< switch_configure_file>
```

**Note** The switch configure file name must include the switch name followed by '**-confg**', for example, **myswitch-confg**.

12. Save and exit the switch configuration from the switch prompt.

```
myswitch#copy running tftp <tftp server> <switch_configure_file>
myswitch#exit
```

Indicate the IP address of the Service Node for the **tftp** server, this is generally the same as the tftp server.

13. Disconnect the Foundry Networks Switch.

Once the switch configuration has been saved and the Administrator has exited from the interface it will then be possible to disconnect the serial line which connects the switch to the Linux Management Node.

14. The configuration can be checked as follows:

From the Management Node run the following command:

```
telnet 10.0.0.254
```

Enter the password when requested.

Set the enable mode:

```
enable
```

Enter the password when requested.

Display the configuration with the **show configuration** command. Two examples are shown below:

```
Model FLS648:
telnet@myswitch#show configuration
```

```
!
Startup-config data location is flash memory
!
Startup configuration:
!
ver 04.0.00T7e1
fan-threshold mp speed-3 50 90
!
module 1 fls-48-port-copper-base-module
!
hostname myswitch
ip address 10.0.0.254 255.0.0.0
!
end
```

```
Model RX4 :
telnet@myswitch#show configuration
```

```
!
Startup-config data location is flash memory
!
Startup configuration:
!
ver V2.3.0dT143
module 1 rx-bi-10g-4-port
module 2 rx-bi-10g-4-port
module 3 rx-bi-1g-24-port-copper
!

vlan 1 name DEFAULT-VLAN
 router-interface ve 1
!
enable telnet password .....
enable super-user-password .....
logging facility local0
hostname myswitch
!
interface management 1
 ip address 209.157.22.254/24
!
interface ve 1
 ip address 172.17.18.210/16
!
end

telnet@myswitch#
```

## 9.2 Configuring a Brocade Switch

1. Set the Ethernet IP address for the brocade switch. Use a portable PC to connect the serial port of the switch.

**Notes**

- The Real Value (IP address, name of the switch) to be used may be found in the cluster database (**FC_SWITCH** table).
- It is mandatory to use the serial cable provided by Brocade for this step.

The initial configuration of the **Brocade** Fibre Channel Switch is made using a serial line (see *Silkworm 200E Hardware Reference Manual*).

2. Open a serial session:

```
cu –s 9600 –l /dev/ttyS0
```

```
login : admin
Password: password
switch:admin>
```

3. Initialize the IP configuration parameters (according to the addressing plan).
   - Check the current IP configuration:

```
switch:admin> ipAddrShow
```

```
Ethernet IP Address: aaa.bbb.ccc.ddd
Ethernet Subnetmask: xxx.yyy.zzz.ttt
Fibre Channel IP Address: none
Fibre Channel Subnetmask: none
Gateway Address: xxx.0.1.1
```

   - Set the new IP configuration.

```
s3800:admin> ipAddrSet
```

```
Ethernet IP Address [aaa.bbb.ccc.ddd]: <new-ip-address>
Ethernet Subnetmask [xxx.yyy.zzz.ttt]: <new-subnet-mask>
Fibre Channel IP Address [none]:
Fibre Channel Subnetmask [none]:
Gateway Address [none]: <new-gateway-address>
```

4. Initialize the switch name, using the name defined in the ClusterDB.

```
switch:admin> switchName "<new_switch_name>"
```

Then:

```
exit
```

## 9.3 Configuring Voltaire Devices

The **Voltaire® Command Line Interface (CLI)** is used for all the commands required, including those for software upgrades and maintenance.

The **Voltaire Fabric Manager (VFM)** provides the **InfiniBand** fabric management functionality including a colour-coded topology map of the fabric indicating the status of the ports and nodes included in the fabric. **VFM** may be used to monitor **Voltaire® Grid Director™ ISR 9096/9288/2012** and **Voltaire® Grid Switch™ ISR 9024** devices. **VFM** includes **Performance Manager (PM), a tool** which is used to debug fabric connectivity by using the built-in procedures and diagnostic tools.

The **Voltaire Device Manager (VDM)** provides a graphical representation of the modules, their LEDS and ports for **Voltaire® Grid Director™ ISR 9096/9288/2012** and **Voltaire® Grid Switch™ ISR 9024** devices. It can also be used to monitor and configure device parameters.

**See** For more detailed information on configuring the devices, updating the firmware, the **Voltaire CLI** commands and management utilities refer to the *Voltaire Switch User Manual ISR 9024, ISR 9096, and ISR 9288/2012 Switches* manual provided on the *Voltaire Switches Documentation CD*.

## 9.4 Installing Additional Ethernet Boards

When installing additional Ethernet cards, the IP addresses of the Ethernet interfaces may end up by being misconfigured:

The Ethernet interfaces are named (eth0, eth1, eth2, etc.) according to the PCI bus order. So when a new Ethernet board is added, the Ethernet interface names may be changed if the PCI bus detects the new board before the existing on-board Ethernet interfaces (PCI bus detection is related to the position of the PCI slots).

To avoid misconfiguration problems of this type, before installing a new Ethernet board, you should:

1. Obtain the MAC addresses of the on-board Ethernet interfaces by using the **ifconfig eth0** and **ifconfig eth1** commands.

2. After the new Ethernet board has been installed, obtain the MAC addresses of the new Ethernet interfaces (obtain all the MAC addresses using the **ifconfig** command)

3. Edit each **/etc/sysconfig/network-scripts/ifcfg-ethX** file (ethX = eth0, eth1, etc.) and add an **HWADDR=<MAC_ADDRESS>** attribute for each interface in each file, according to the Ethernet interface name and the MAC address obtained in Step 2, above.

# Appendix A. Default Logins for different cluster elements

| Element | Login | Password | Comments |
|---|---|---|---|
| **Baseboard Management Controller** | administrator | administrator | |
| **InfiniBand switches** | enable | voltaire | Equivalent to root → used for configuration switch |
| | admin | 123456 | Read only |
| **Ethernet switches** | admin | admin | |
| | admin | admin | Same login and password for root |
| **DDN Storage subsystems** | admin | password | The same logins are defined in the **/etc/storageadmin/ddn_admin.conf** file. |
| **NEC Storage subsystems** | iSM | iSM | Change to **admin** and **password** to match logins defined in the **/etc/storageadmin/nec_admin.conf** file. |
| **Xyratex Optima 1200 Storage subsystems** | admin | password | The same logins are defined in the **/etc/storageadmin/xyr_admin.conf** file. |
| **EMC/DGC (CX3-40F) Storage systems** | User defined at the first connection | User defined at the first connection | It is recommended to use **admin** and **password** in the same way as for other systems. |

# Appendix B. Cluster Database Operations

## B.1 Migrating to BAS5 for Xeon v1.2

### B.1.1 Migrating Cluster DB Data from BAS5 for Xeon v1.1

The Cluster Database data will be migrated automatically when an existing **BAS5 for Xeon v1.1** cluster is upgraded to **BAS5 for Xeon v1.2**.

---

**See** *Chapter 2 – Updating BAS5 for Xeon v1.1 clusters to BAS5 for Xeon v1.2* in this manual for a description of the upgrade procedure for clusters without any form of High Availability.

---

### B.1.2 Migrating Cluster DB Data from BAS4 for Xeon v1.2

WARNING

**All working activity for the cluster must have been stopped before migrating the cluster database on the Management Node.**

1. Log on as root on the **BAS4 for Xeon v1.2** Management Node and install the **clusterdb-data-ANY412-20.4.1-b.x.Bull rpm,** available either from the **BAS5 for Xeon v1.2 XHPC DVD ROM** or from Bull Technical Support, by running the command below:

```
rpm -ivh clusterdb-data-ANY412-20.4.1-b.x.Bull --nodeps
```

2. Change to **postgres** :

```
su - postgres
```

3. Go to the install directory on the Management Node:

```
cd /usr/lib/clustmngt/clusterdb/install
```

4. Run the command **preUpgradeClusterdb**. This command modifies and creates dumps of the Cluster DB data and schema.

```
./preUpgradeClusterdb.
```

5. If the **preUpgradeClusterdb** command completes without any errors, copy the **preclusterdball2041.dmp** and **preclusterdbdata2041.dmp** files onto an external storage device. Contact Bull Technical Support if there are any errors.

6. Stop the **postgresql** service to prevent the cluster database from being modified.

```
service postgresql stop
```

7. Install **BAS5 for Xeon v1.2** on the cluster by following the installation procedure described in *Chapter 3* in the **BAS5 for Xeon** *Installation and Configuration Guide.*

WARNING

**Read Chapter 3 carefully before installing BAS5 for Xeon. Be sure that all data has been backed up onto non-formattable media outside of the cluster. Check that all the necessary configuration files necessary have been saved and backed up as described in the Pre-installation Operations when Re-installing BAS 5 v1.2 section.**

8. Copy across the dump files saved in point 5 from the external storage device to the **~/backups** directory on the Management Node.

9. Change to **postgres** and restore the cluster database data dump files by running the commands below:

```
su - postgres
cd ~/backups
pg_restore -Fc --disable-triggers -d clusterdb preclusterdbdata2041.dmp
exit
```

### B.1.3 Migrating Cluster DB Data from BAS4 for Xeon v1.1

The procedure is exactly the same as described in section B.1.2 for **BAS4 for Xeon v1.2** clusters **EXCEPT** the **BAS4 for Xeon v1.1** specific Cluster DB data RPM, **clusterdb-data-ANY411-20.4.1-b.x.Bull**, must be installed on the **BAS4 for Xeon v1.1** Management Node. This RPM is available on the **BAS5 for Xeon v1.2 XHPC DVD ROM** or from Bull Technical Support.

## B.2 Saving and Reinstalling the Cluster DB data

Follow the procedure, described below, to save and to restore cluster database data for a **BAS5 for Xeon v1.2** clusters.

### B.2.1 Saving the Data files

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-2008-1105.sav.

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

### B.2.2 Reinstalling the Data files

1. Switch to **postgres**:

```
su - postgres
```

2. Go to the install directory:

```
cd /usr/lib/clustmngt/clusterdb/install
```

3. Remove the existing cluster DB:

```
dropdb clusterdb
```

4. Create a new cluster DB schema:

```
create_clusterdb.sh --nouser
```

5. Truncate the default values:

```
psql -U clusterdb  -c "truncate config_status; truncate
config_candidate" clusterdb
```

6. Run the command:

```
psql -U clusterdb  -c " alter table ic_switch alter column
admin_ipaddr drop not null" clusterdb
```

7. Restore the **.sav** files saved previously

```
pg_restore -Fc --disable-triggers -d clusterdb
/var/lib/pgsql/backups/<name_of_clusterdb_saved__file>
```

8. Go back to root by entering the exit command:

```
exit
```

## B.3 Initializing the Cluster Database using the preload file

Contact Bull Technical Support to obtain the Cluster DB preload file for **BAS5 for Xeon v1.2**, and then follow the procedure described in *section 3.2.5.1* in the *BAS5 for Xeon Installation and Configuration Guide* for the initialization of the Cluster Database.

# Appendix C. Migrating Lustre

For Lustre 1.6.3 and above, the following upgrades are supported:

- 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).

The complete migration procedure is described in the *Upgrading Lustre* chapter in the *Lustre 1.6 Operations Manual* available from http://www.lustre.org

**Important**

- **This chapter assumes an existing BAS4 for Xeon cluster has migrated to BAS5 for Xeon without the XLustre 1.6.x RPMS being installed when the system was migrated.**
- **Lustre has to be migrated from version 1.4.x to version 1.6.x on all clusters which install BAS5 for Xeon.**
- **All data stored in the Lustre file systems should be backed up before Lustre is migrated.**

WARNING

**The *Lustre 1.6 Operations Manual* 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.**

## C.1 Migrating Lustre from version 1.4 to version 1.6

### C.1.1 Pre-Configuration for Migration

1. Disable High Availability for **Lustre,** if it is in place. For all Lustre file systems, run the command:

```
lustre_ldap unactive -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.

2. Stop all the file systems from the Management Node:

```
lustre_util umount –f all –n all
lustre_util stop –f all
```

3. 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.

## C.1.2 Installation and Configuration of Lustre version 1.6.x RPMS

1. Mount **NFS** from the **/release** directory on the Management Node to the **/release** directory on the Lustre Service Node :

```
ssh <Service_Node>
mount -t nfs <Management_Node_IP>:/release /release
```

2. Install the **XLustre** software as shown below

```
cd /release/XBAS5V1.2
./install -prod XLUSTRE
```

3. Configure the new version of Lustre, as detailed in the *Configuring Lustre* section in Chapter 6 in the **BAS5 for Xeon** *Installation and Configuration Guide*

**Important**

**Stop at the *Install the file system* step as the Lustre configuration details and data will have been saved previously.**

## C.1.3 Post-Configuration operations

1. After installing the Lustre version 1.6 packages, copy the contents of the backed up **lustre.bkp** directory into **/etc/lustre/**:

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

2. Check the new **lustre.cfg** file contains the new MGS related directives, i.e.

LUSTRE_MGS_HOST
LUSTRE_MGS_NET
LUSTRE_MGS_ABSOLUTE_LOOPBACK_FILENAME

3. 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).

```
clean_extents_on_dirs.sh
```

4. Set up the new **MGS** entity on the Management Node and upgrade the Lustre layout by running the **upgrade_lustre_layout.sh** script from the Management Node.

```
upgrade_lustre_layout.sh
```

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

```
lustre_util update -f <fsname>
```

6. Restart the Lustre file systems.

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

7. Enable Lustre High Availability, if it is in place. 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.

# Appendix D. Manually Installing BAS5 for Xeon Additional Software

If the **preparenfs** command was NOT used to install the additional software options (**XIB** and/or **XLUSTRE** and/or **XTOOLKIT**), the process to install them manually is described below.

1. Mount **NFS** from the **/release** directory on the Management Node to the **/release** directory on the Service Node :

```
ssh <Service_Node>
mount -t nfs <Management_Node_IP>:/release /release
```

2. Install the optional **BAS5 for Xeon** software products required. The products to be installed for the cluster must be listed after the **–prod** option, as shown in the example below. In this example all the software products will be installed:

```
cd /release/XBAS5V1.1
./install -prod XIB XLUSTRE XTOOLKIT
```

**Important**

**Lustre must use dedicated service nodes for the I/O functions and NOT combined Login/IO service nodes. NFS can be used on both dedicated I/O service nodes and on combined Login/IO service nodes.**

**See** The Bull BAS5 for Xeon *Application Tunings Guid*e for details on configuring and using **HPC Toolkit**.

# Appendix E. Configuring Interconnect Interfaces

### First installation or reinstallation of BAS5 for Xeon v1.2

The configuration of the **InfiniBand** Interconnect interfaces is carried out automatically by **Ksis** when the images of the Compute and Login/IO nodes are deployed.

### Update from BAS5 for Xeon v1.1 to BAS5 for Xeon v1.2

The **config_ipoib** command is used to configure both **InfiniBand** and **Ethernet** interfaces.

## E.1 The config_ipoib command

The interconnect interface description file is generated from the Management Node for each node by using the **config_ipoib** command.

The interfaces parameters are obtained from the **/etc/hosts** file on the Management Node.

Different options have to be set for the **config_ipoib** command according to the configuration of the cluster. The command options are shown below:

### Usage

```
config_ipoib -n node[a-b,x] [-d device] [-m netmask] [-s suffixe]
```

### Command options

**-h –help** - print this message

**-n <node>** - node to update, pdsh form node[a-b,x] or ssh form root@node

**-d <device>** - ip device (default ib0)

**-m <masque>** - ip net mask (default 255.255.0.0)

**-s <suffixe>** - name suffix in /etc/hosts (default -ic0)

In the example below, the command will create the configuration file **ifcfg-eth1** on the nodes **zeus8** to **zeus16**, to configure the **eth1** interface for these nodes, using the IP addresses listed in the **/etc/hosts** file for the **zeus8-ic1** to **zeus16-ic1** interfaces.

```
config_ipoib -n zeus[8-16] -d eth1 -m 255.255.0.0 -s –ic1
```

## E.2 Interface Description file

### Ethernet Adapters

The **Ethernet** interconnect adapter will be identified by a logical number by using the format **eth[1/2/...]**, for example **eth1** and **eth2**. The IP properties (address, netmask, etc.) for the Ethernet adapter are configured using a description file named: **/etc/sysconfig/network-script/ifcfg-eth[1/2/...]**

### InfiniBand Adapters

The **InfiniBand** interconnect adapter will be identified by a logical number by using the format **ib[0/1/2/...]**, for example **ib0** and **ib1**. The IP properties (address, netmask, etc.) for the InfiniBand adapter are configured using a description file named **/etc/sysconfig/network-script/ifcfg-ib[0/1/2/...]**

### Example

An example of a description file is shown below for a node with an **InfiniBand** interface:

```
# cat /etc/sysconfig/network-scripts/ifcfg-ib0
```

```
DEVICE=ib0
ONBOOT=yes
BOOTPROTO=static
NETWORK=172.18.0.0
IPADDR=172.18.0.4
```

**Note** The value of last byte (octet) of the IPADDR address is always 1 more than the value for the machine number. For example, in the interface above the machine number is 3 (ns3) and so the last byte in the IPADDR setting is 4.

## E.2.1 Checking the interfaces

It is recommended that the configuration of the **Ethernet** and **InfiniBand** interfaces are verified to ensure that all the settings are OK. This is done by running the command below for **InfiniBand** interfaces:

```
pdsh –w node[n,m] cat /etc/sysconfig/network-scripts/ifcfg-ib[0/1/2…]
```

or the command below for **Ethernet** interfaces:

```
pdsh –w node[n,m] cat /etc/sysconfig/network-scripts/ifcfg-eth[1/2/3…]
```

Alternatively, to see the interface settings separately in groups for a set of nodes, use the commands below:

**Note** The examples below show the commands to be used for **InfiniBand** interfaces. For **Ethernet** interfaces replace the adapter interface identifier accordingly, for example replace **ifcfg-ib0** with **ifcfg-eth1**.

```
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.

### E.2.2 Starting the InfiniBand interfaces

The following commands are used to load all the modules, and to start all the **InfiniBand** interfaces, on each node:

```
/etc/init.d/openibd start
```

or

```
service openibd start
```

These commands have to be executed for each node individually.

**Note** A node reboot may be used to load the **InfiniBand** modules automatically.

# Appendix F. Binding Services to a Single Network

The bind attribute in the **/etc/xinetd.conf** file is used to bind a service to a specific IP address. This may be useful when a machine has two or more network interfaces; for example, a backbone computer which is part of a cluster administration network and is at the same time connected to the customer LAN through a separate interface. In this situation there may be backbone security concerns coupled with a desire to limit the service to the LAN.

For example, to bind the ftp service to the LAN, the **/etc/xinetd.conf** file has to be configured as follows:

### LAN network configuration

```
{
  id            = ftp-local
  wait          = no
  user          = root
  server        = /usr/sbin/in.ftpd
  server_args   = -l
  instances     = 4
  nice          = 10
  only_from     = 0.0.0.0/0 #allows access to all clients
  bind          = xxx.xxx.xxx.xxx #local IP address
}
```

### Administration network configuration

```
{
  id            = ftp-admin
  socket_type   = stream
  wait          = no
  user          = root
  server        = /usr/sbin/in.ftpd
  server_args   = -l
  only_from     = xxx.yyy.0.0/16 #only for internal use
  bind          = xxx.yyy.0.99  #local IP address
}
```

**Note** The configurations above can be adapted and used by other services.

# Appendix G. Configuring AOC-USASLP-S8iR RAID Adapters for NovaScale R423 and R425 machines

**Note** The operations described in this chapter have to be carried out individually on each **NovaScale R423** and **R425** machine included in the cluster

1. Reboot the machine via **conman** from the Management Node. Press **Ctrl A** after the **Adaptec RAID BIOS** line appears to enter the **Adaptec RAID Configuration Utility**, as shown below.

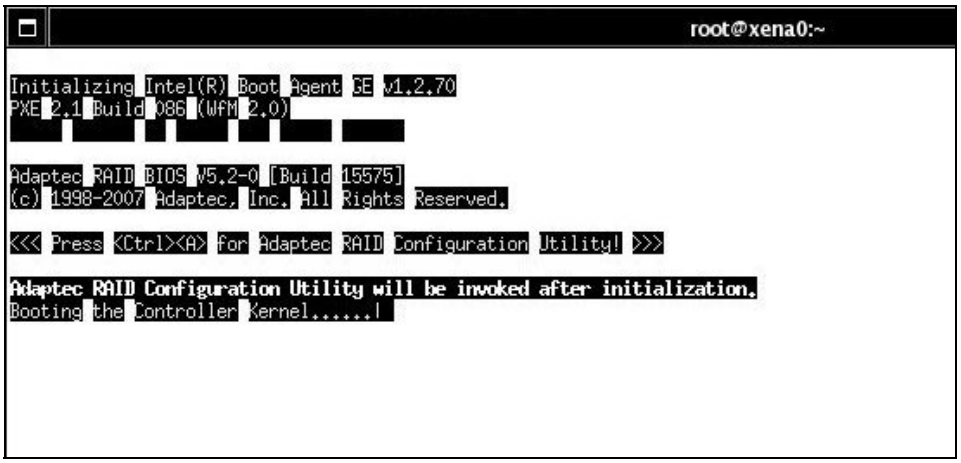


Figure G-1. Boot screen with Adaptec RAID BIOS

2. Select the **Array Configuration Utility** from the Adapter RAID Configuration Utility **Options** menu, as shown below.


Figure G-2. RAID Configuration Utility Options menu -> Array Configuration Utility

3. Select **Manage Arrays** from the Array Configuration Utility **Main Menu**, as shown below.


Figure G-3. Array Configuration Utility **Main Menu**

4. The **List of Arrays** already installed will be displayed. Select an array from the list to see its properties, as shown in the examples below.

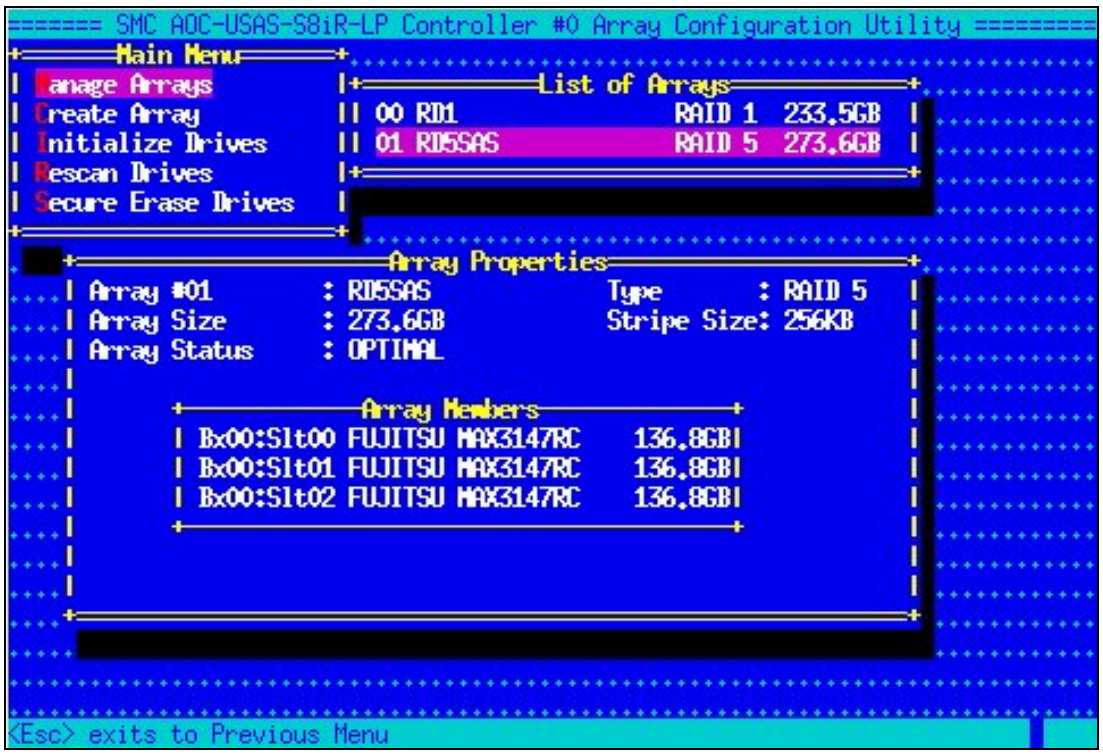


Figure G-4. Example of **Array Properties** for a RAID 5 Array

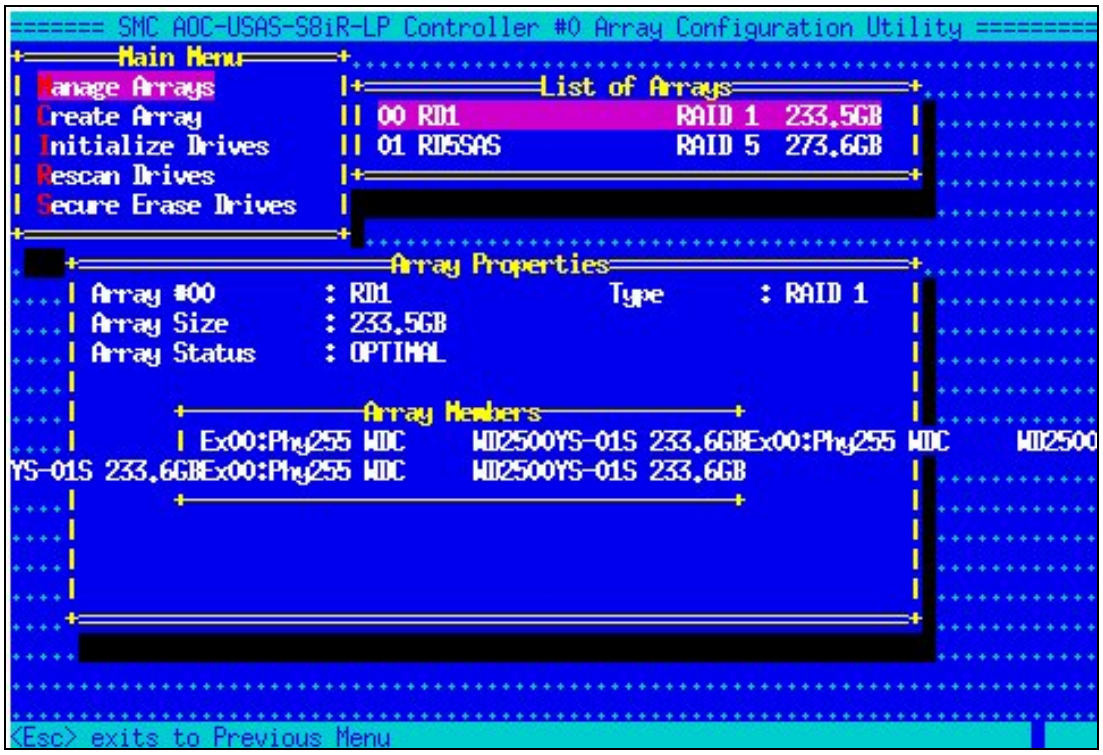


Figure G-5. Example of **Array Properties** for a RAID 1 array

5. Press the Escape key to return to the previous screen and select **Create Array** from the **Main Menu**. All the drives connected to the server will be displayed, those that are shown with 0KB in the final column - see example below - will not be accessible, as they are already included in an array.

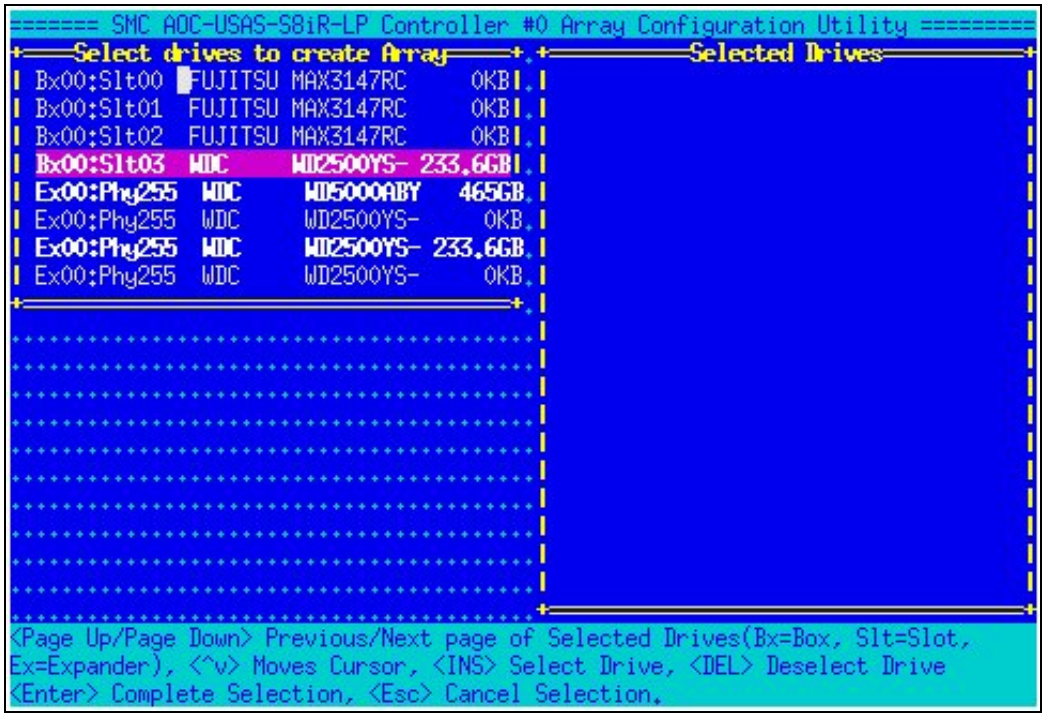


Figure G-6. Example of drive list for a server

6. Press **F7** to select the drives to be included in the new array. **Only drives of the same size can be selected for the new array, see figure below.**

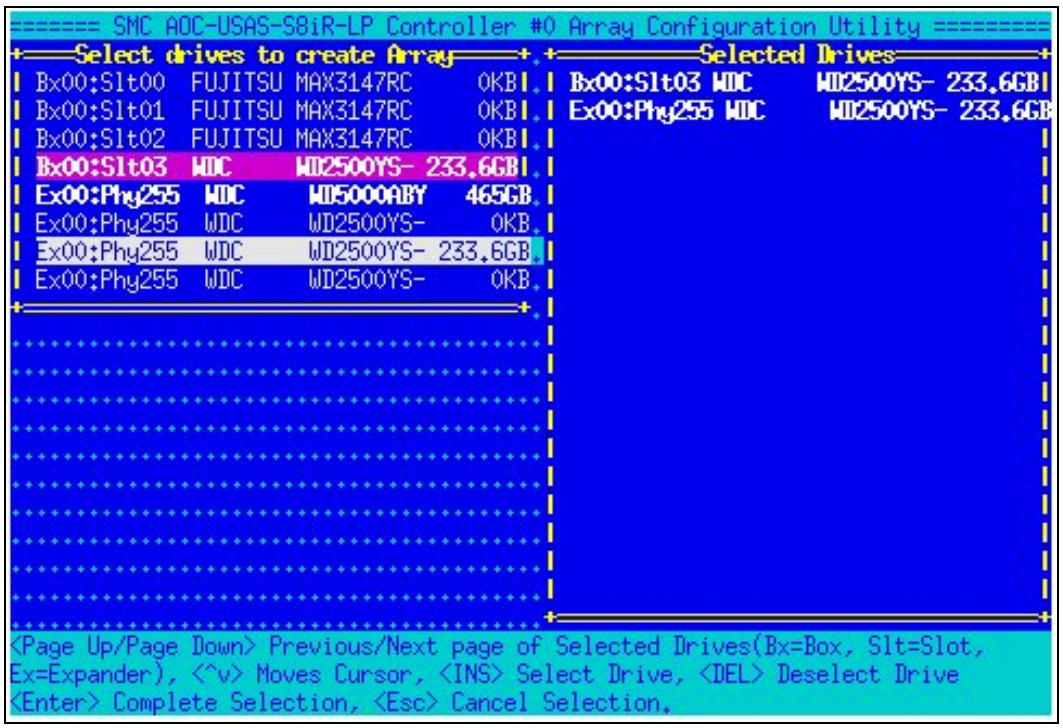


Figure G-7. Selection of drives of the same size for new RAID array

7. Press **Enter** when all the drives have been selected for the new array. The **Array Properties** screen appears- see Figures below. Select the **Array Type** to be configured, followed by the other properties, size, label, etc. for the array.

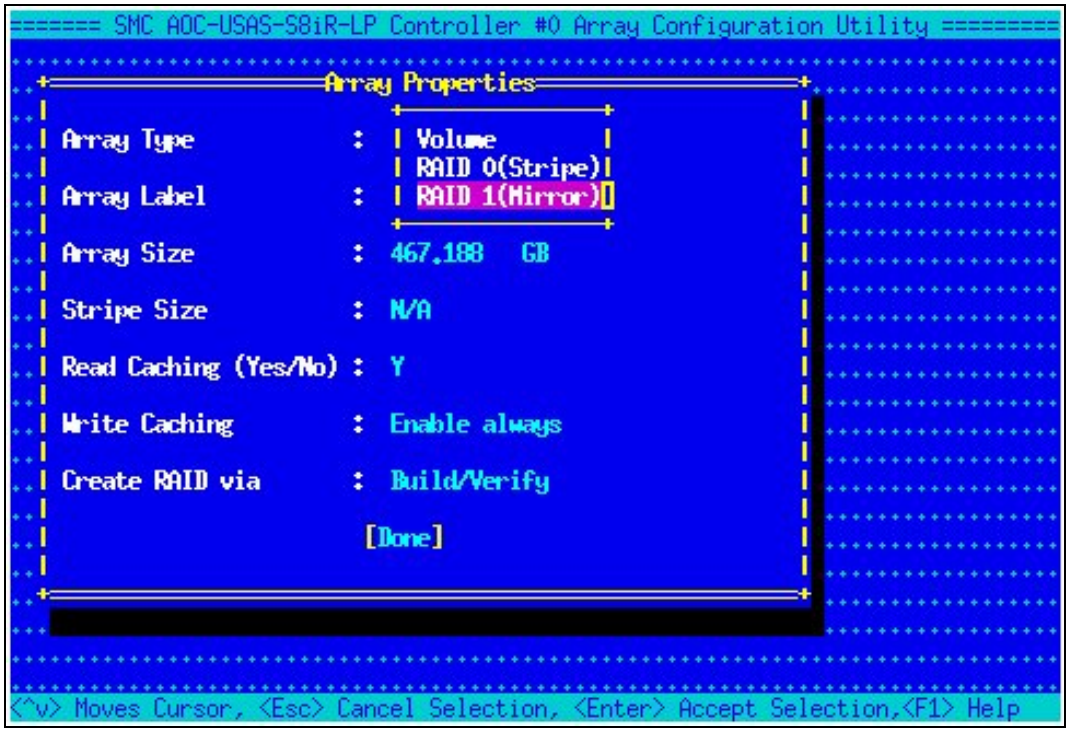


Figure G-8. Array Properties - Array Type

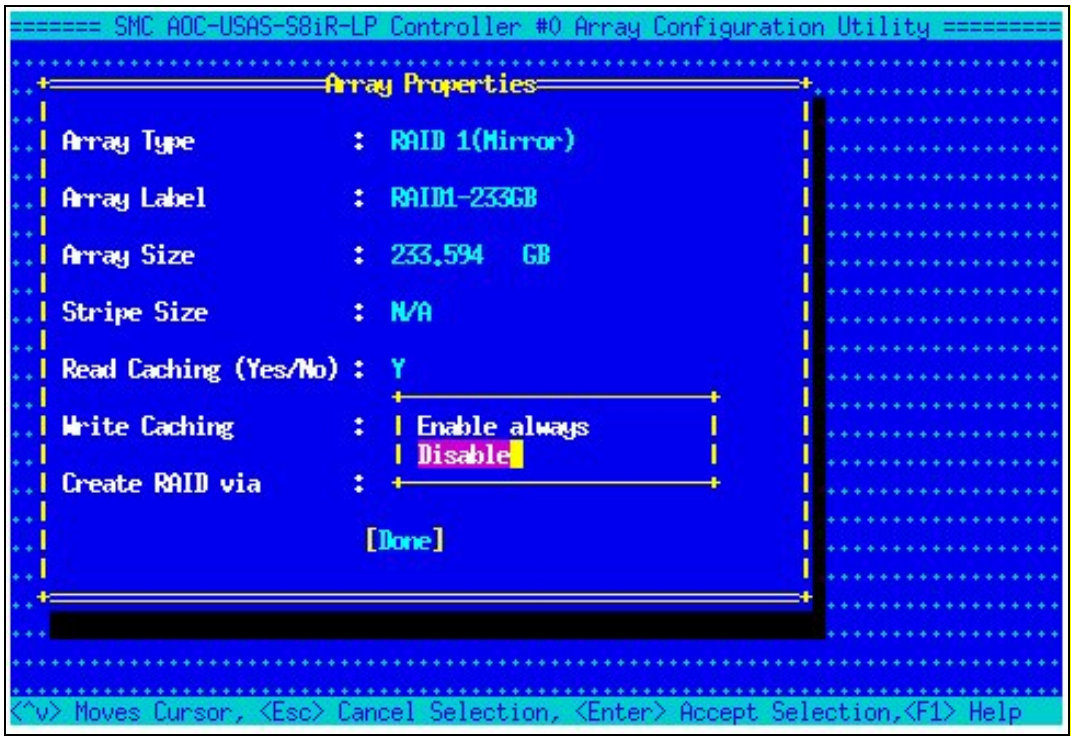


Figure G-9. Array Properties - Write caching

**Note** It is recommended that **Write Caching** is disabled, however this is not obligatory.

8. Confirm all the values for the new RAID array by selecting **Done**, as shown in the Figure below. The settings below are an example only.

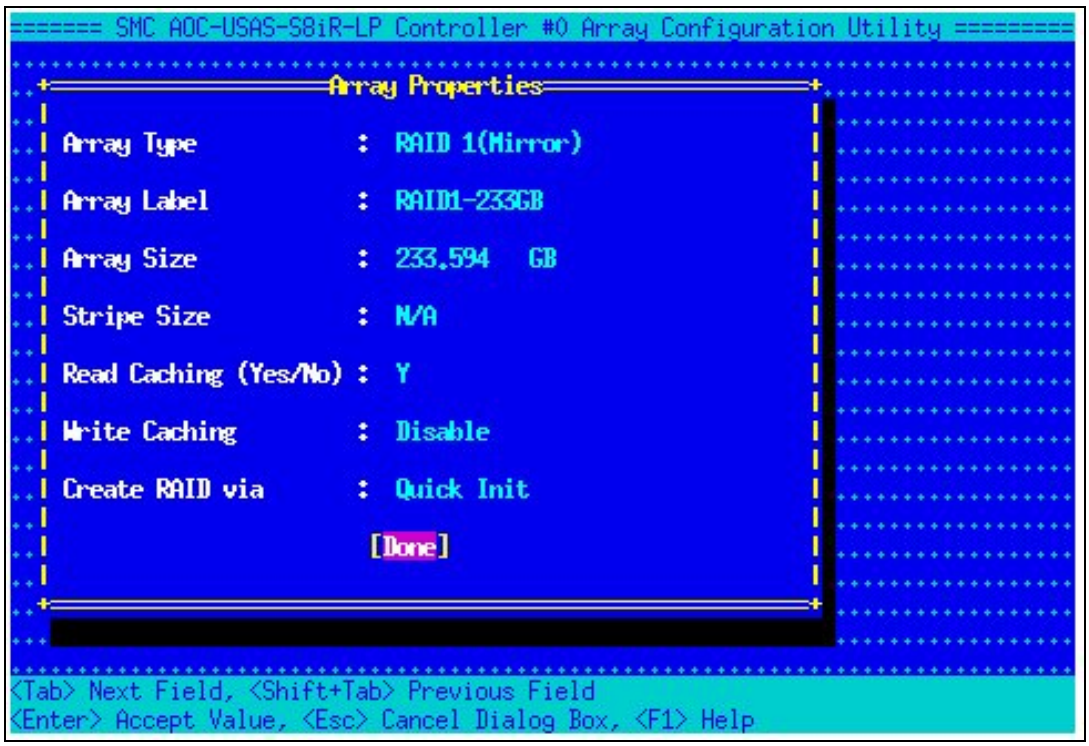


Figure G-10. Array Properties - Confirmation screen

9. Exit the **Array Configuration Utility**. Press **Escape** several times until the **Options** screen appears and select **SerialSelect Utility**, as shown below.

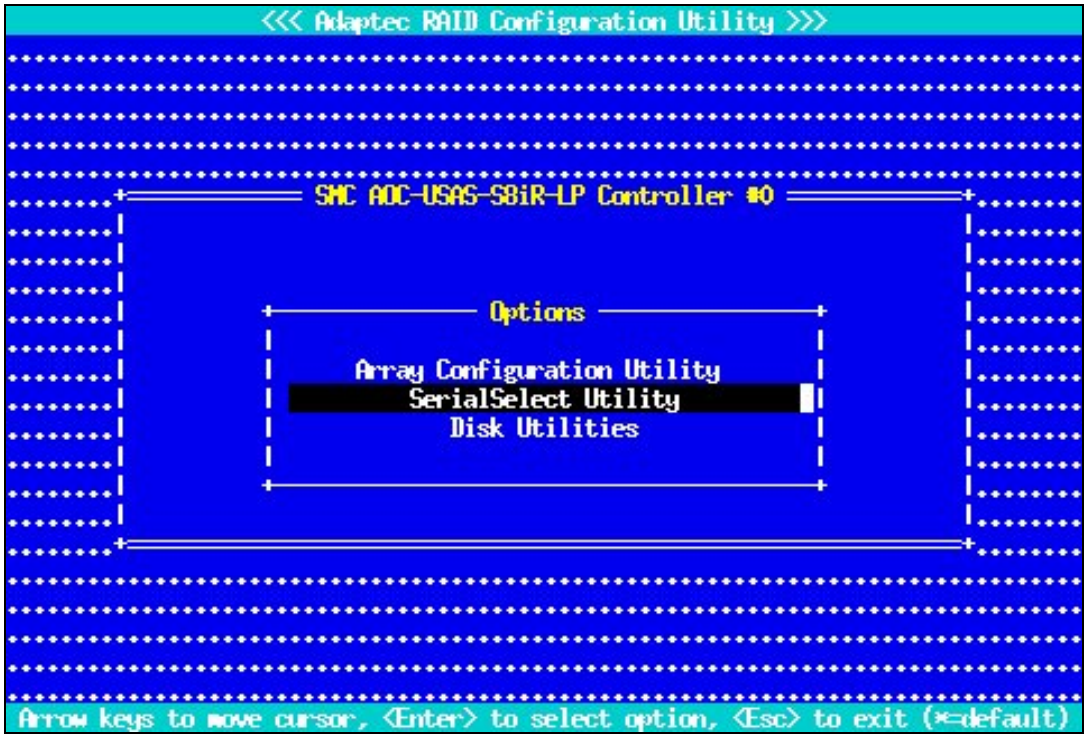


Figure G-11. RAID Configuration Utility - Options Menu

10. Select **Controller Configuration**, as shown below.

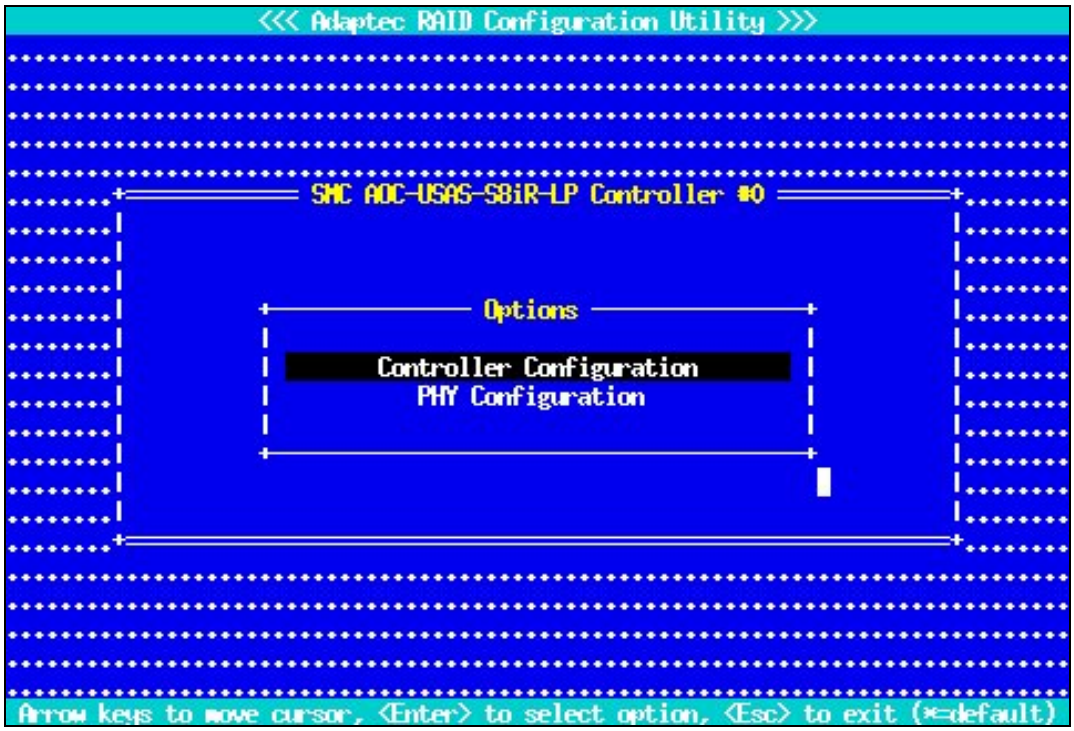


Figure G-12. RAID Configuration Utility - Options Menu -> Controller Configuration

11. Check all the settings for the Controller - see Figure below


Figure G-13. SMC AOC-USAS-S8iR Controller settings

12. Once all the settings are in place press **Escape** to exit and select **PHY Configuration** from the **Options** menu - see Figure G-12.

13. Check the Physical Layer settings that are in place - see Figure below. The settings below are examples only.


Figure G-14. SAS PHY Settings

14. Press **Escape** several times until the **Options** menu appears, select **Disk Utilities** , as shown below.

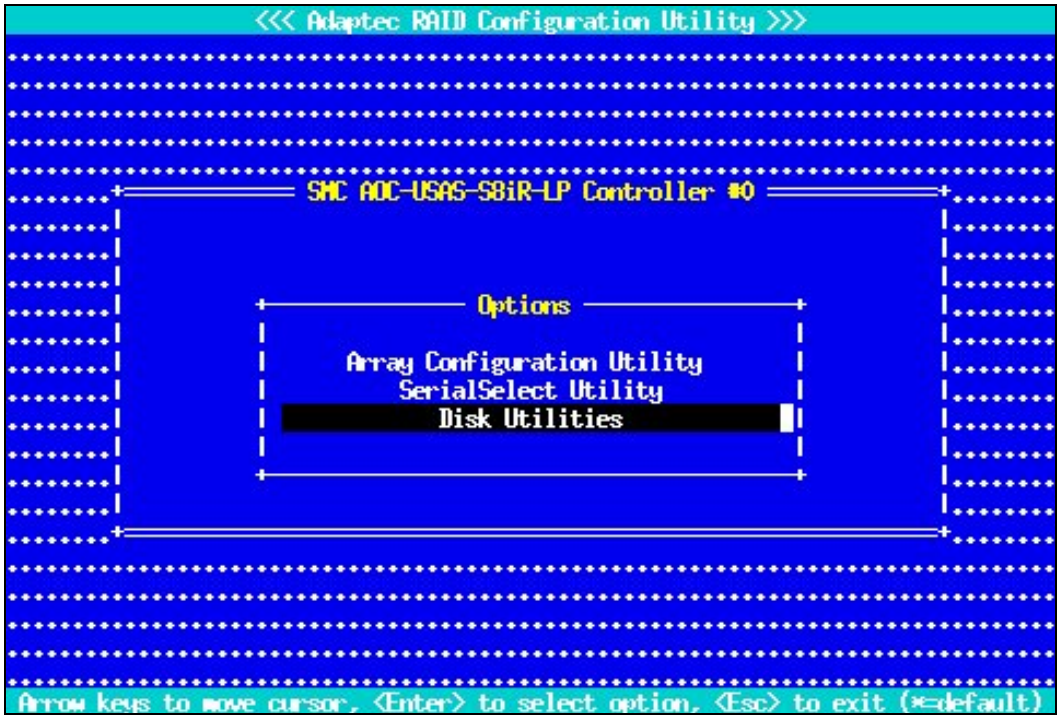


Figure G-15. RAID Configuration Utility - Options Menu -> Disk Utilities

15. Check that all the drives are present - see Figure below.

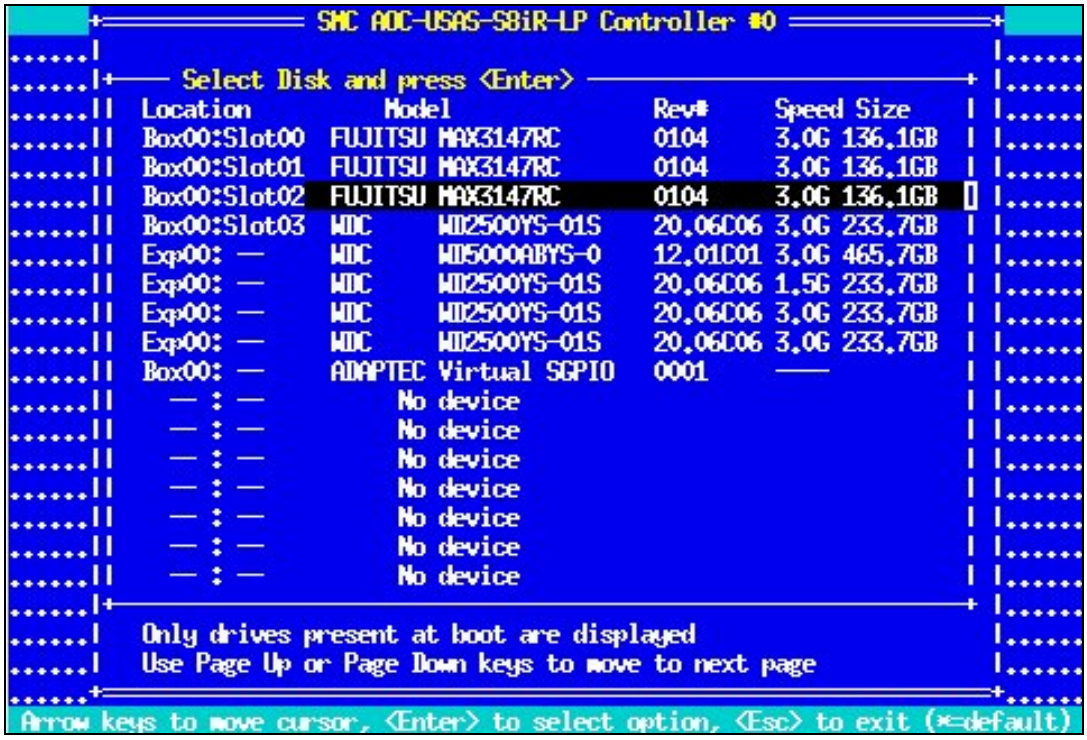


Figure G-16. An example of a drive list for an Adaptec controller

16. If everything is OK press **Escape** several times to go back until the **Exit Utility** menu appears, as shown below.


Figure G-17. RAID Configuration Utility - Exit Utility menu

17. Select **Yes** from the **Exit Utility** menu to confirm settings and press **Enter**. The **Rebooting the system** message will appear. Once the system has rebooted the new RAID will have been configured.

# Appendix H. PCI Slot Selection and Server Connectors

This appendix provides detailed information regarding the choice of PCI slots for high bandwidth PCI adapters. The configuration rules put forward ensure the best performance levels, without I/O conflicts, for most type of applications. System diagrams are included which may be used to configure the hardware connections.

The following topics are described:

- H.1 *How to Optimize I/O Performance*
- H.2 *Creating the list of Adapters*
- H.3 *Connections for NovaScale R4xx* Servers

## H.1 How to Optimize I/O Performance

The I/O performance of a system may be limited by the software, and also by the hardware. The I/O architecture of servers can lead to data flows from PCI slots being concentrated on a limited number of internal components, leading to bandwidth bottlenecks.

Thus, it is essential to look at the installation of PCI adapters, and slot selection, carefully, to reduce any limitations as much as is possible. One good practice is to avoid connecting bandwidth hungry adapters to the same PCI bus.

The following details should be ascertained, in order to ensure the highest possible performance for the adapter installation:

- Adapter characteristics, maximum theoretical performance and expected performance in the operational context.
- The I/O architecture of the server.

The following paragraphs cover these aspects, and provide recommendations for the installation of adapters for different NovaScale servers. The process to follow is quite easy:

1. Create a list of the adapters to be installed, sorted from the highest bandwidth requirement to the lowest.

2. Place these adapters in each server using the priority list specific to the platform, as defined in this Appendix.

## H.2 Creating the list of Adapters

The first step is to make a list of all the adapters that will be installed on the system.

Then, if the I/O flow for the server is known (expected bandwidth from the Interconnect, bandwidth to the disks, etc.), it will be possible to estimate the bandwidth required from each adapter, and then sort the adapters according to the requirements of the operational environment.

If there is no information about real/expected I/O flows, the adapters should be sorted according to their theoretical limits. As both PCI Express adapters and PCI-X adapters may be connected, 2 tables are provided for the adapters supported by BAS5 for Xeon. These are sorted by throughput, giving the HBA slotting rank.

| Adapter | Bandwidth |
|---|---|
| Fibre channel dual ports | 800 MB/s (1) (2) |
| Fibre channel single ports | 400 MB/s (2) |
| Gigabit Ethernet dual port | 250 MB/s (1) (2) |
| Gigabit Ethernet single port | 125 MB/s (2) |
| Ethernet 100 Mbps | 12,5 MB/s |

Table H-1. PCI-X Adapter Table

(1) If both channels are used. Otherwise, the adapter must be categorised as a single channel/port adapter
(2) Full duplex capability is not taken into account. Otherwise, double the value listed.

It may be possible that these values will be reduced, due to the characteristics of the equipment attached to the adapter. For example, a U230 SCSI HBA connected to a U160 SCSI disk subsystem will not be able to provide more than 160 MB/s bandwidth.

| Adapter | Bandwidth |
|---|---|
| Infiniband Voltaire 400 or 410-EX-D | 1500 MB/s |
| Fibre channel dual ports | 800 MB/s |
| Fibre channel single ports | 400 MB/s (2) |
| Gigabit Ethernet dual port | 250 MB/s |
| Gigabit Ethernet single port | 125 MB/s (2) |

Table H-2. PCI-Express Table

## H.3 Connections for NovaScale R4xx Servers

The following paragraphs illustrate the I/O subsystem architecture for each family of **NovaScale** Rxx servers.

### H.3.1 NovaScale R421 Series – Compute Node

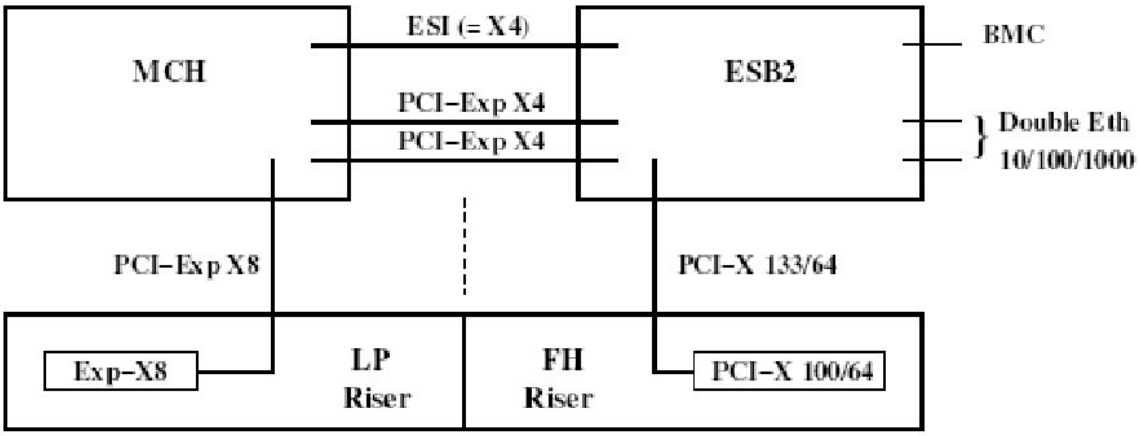


Figure H-1. NovaScale R421 rear view of Riser architecture

The ports attached to the North Bridge or the Memory Controller Hub (**MCH**) offer a higher performance than those attached to the Enterprise South Bridge (**ESB**).

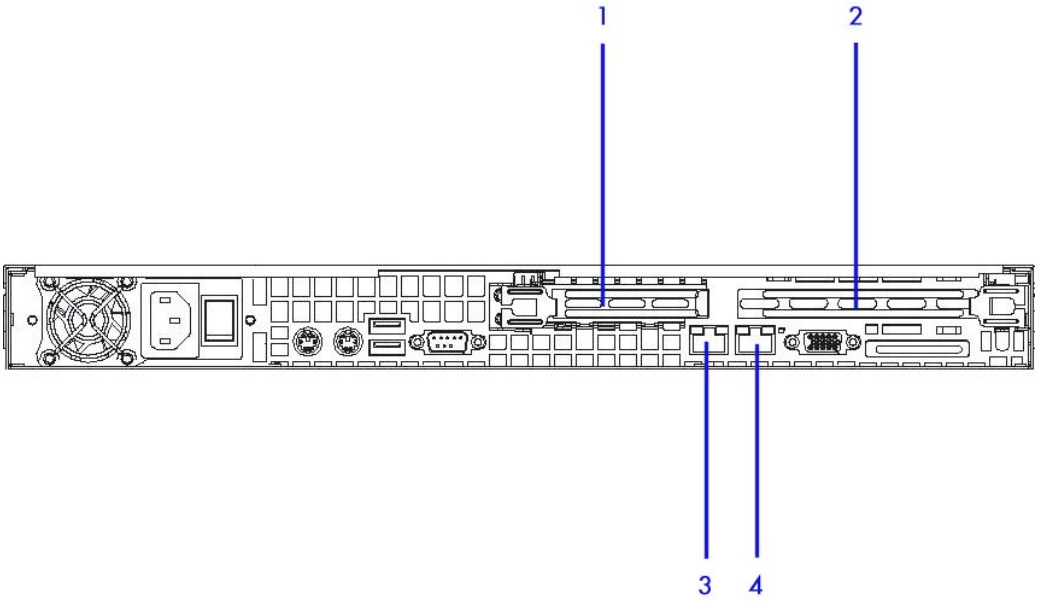


Figure H-2. NovaScale R421 rear view connectors

| Connector number | Port/Slot | Use |
|---|---|---|
| 1 | PCI-Express x8 | InfiniBand interconnect or Ethernet 1000 Backbone (when slot 4 is used for Ethernet 1000 interconnect) |
| 2 | PCI-X 100MHz / 64 bit | |
| 3 | Ethernet | Administration Network or BMC Network |
| 4 | Gbit Ethernet | Ethernet 1000 interconnect or Ethernet Backbone (when slot 1 is used for InfiniBand interconnects) |

Table H-3. NovaScale R421 Slots and Connectors

## H.3.2 NovaScale R422 Series – Compute Node

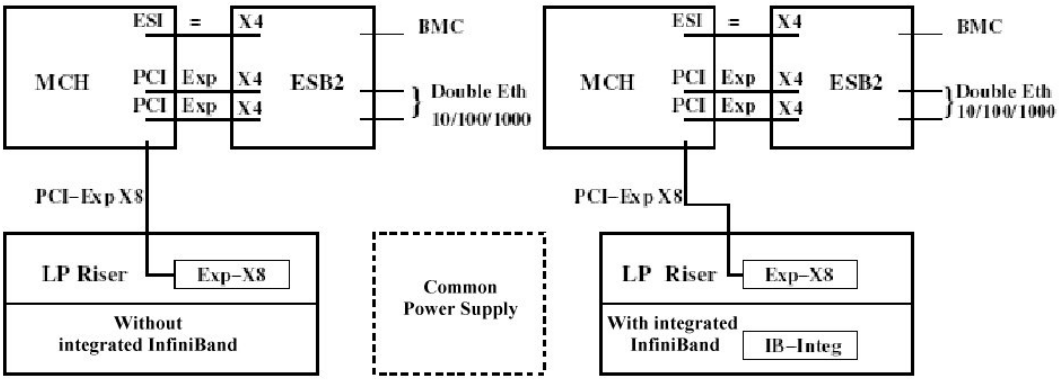


Figure H-3. NovaScale R422 rear view of Riser architecture

The ports attached to the North Bridge or the Memory Controller Hub (**MCH**) offer a higher performance than those attached to the Enterprise South Bridge (**ESB**).

---

**Note** Depending on the model, an on-board **InfiniBand** controller with a dedicated port may be included. The two servers within a **NovaScale R422** machine are identical, they either both include the **InfiniBand** controller or they both do not.

---

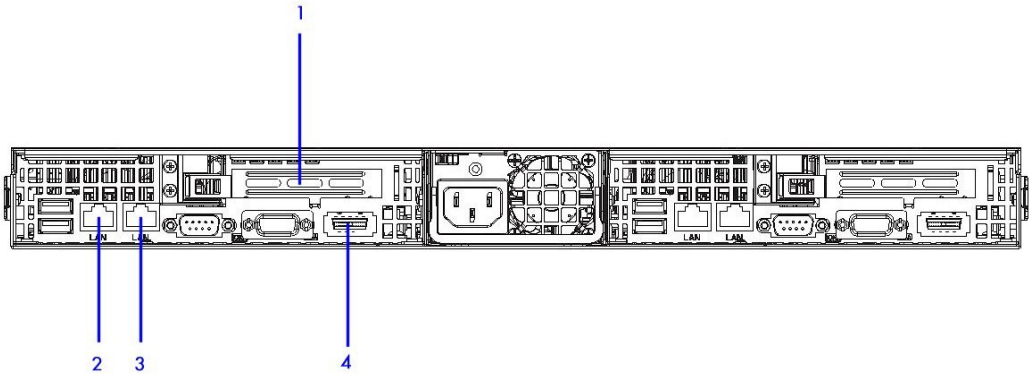


Figure H-4. NovaScale R422 Rear view connectors

| Connector number | Port/Slot | Use |
|---|---|---|
| 1 | **PCI - Express x8** | InfiniBand Interconnect or Ethernet 1000 Backbone |
| 2 | **LAN port** | Management Network or BMC Network |
| 3 | **LAN port** | Gbit Ethernet or Gbit Ethernet Interconnect or Ethernet 1000 backbone |
| 4 | **InfiniBand** port (optional) | InfiniBand Interconnect |

Table H-4. NovaScale R422 Slots and Connectors

## H.3.3 NovaScale R460 Series – Service Node

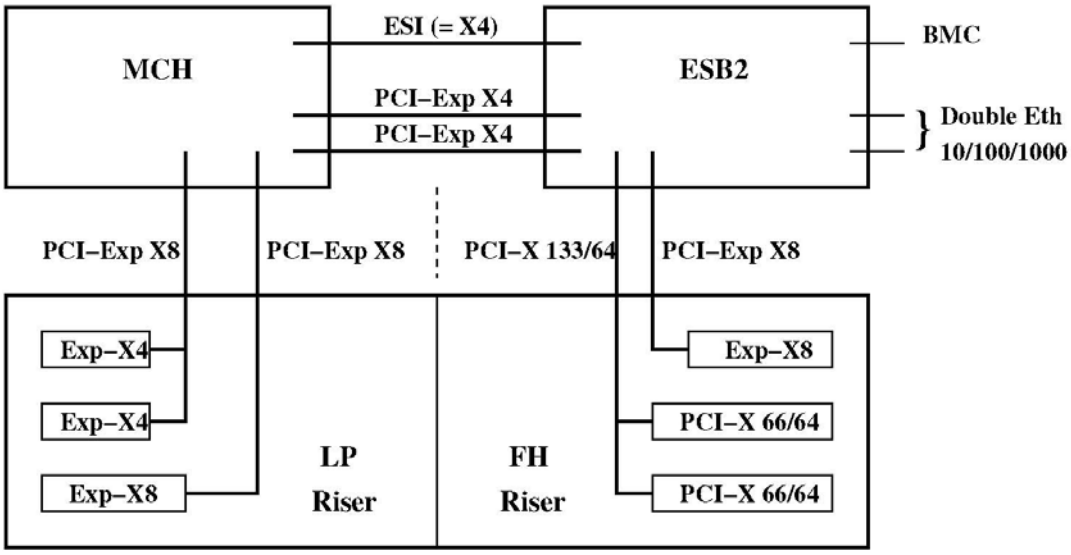


Figure H-5. NovaScale R460 risers and I/O subsystem slotting

The ports attached to the North Bridge or the Memory Controller Hub (MCH) offer a higher performance than those attached to the Enterprise South Bridge (ESB).

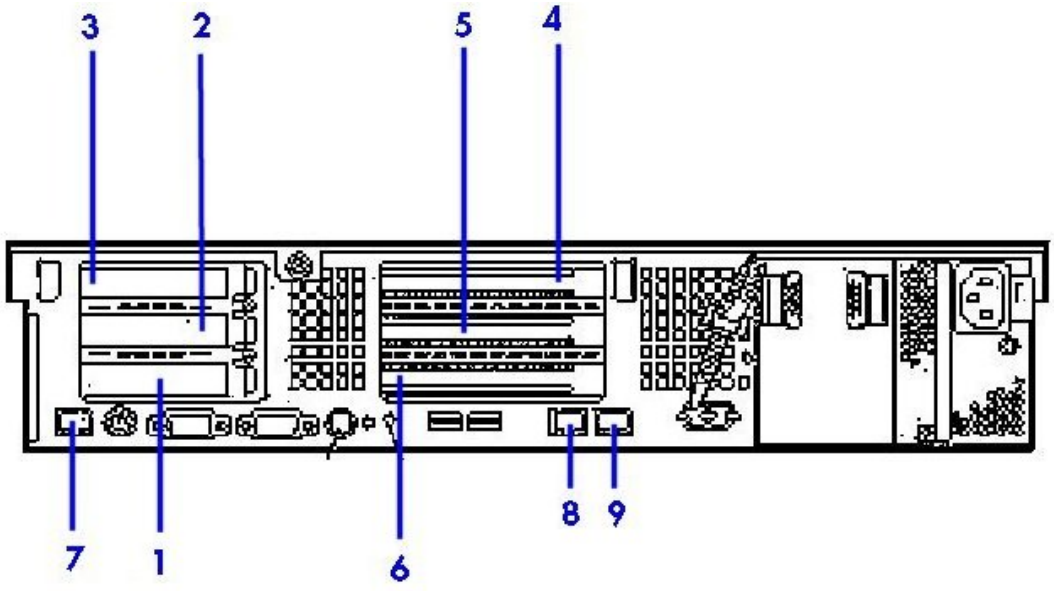


Figure H-6. Rear view of NovaScale R460 Series

| Connector number | Port/Slot | Use |
|---|---|---|
| 1 | PCI-Express x8 | InfiniBand Double Data Rate Adapter |
| 2 | PCI-Express x4 | Fibre Channel Disk Rack |
| 3 | PCI-Express x4 | Fibre Channel Input\Output |
| 4 | PCI-Express x8 | Optional backbone - 10 Gigabit Ethernet **Myricom Myri-10G** (x8) **OR** 1 Gbit Ethernet **Intel 82571** Ethernet Controller (x4) |
| 5 | PCI-X 66 MHz / 64 bit | |
| 6 | PCI-X 66 MHz /64 bit | |
| 7 | Ethernet | Dedicated Board Management Controller (**BMC**) connector for the BMC network. |
| 8 | Ethernet | Administration Ethernet Connector |
| 9 | Ethernet | Gigabit Ethernet Interconnect |

Table H-5. NovaScale R460 Slots and Connectors

**Note** Either slot number 1 is used for **InfiniBand** interconnects OR connector number 9 is used for Gigabit **Ethernet** interconnects. These networks are exclusive.

# Appendix I. Activating your Red Hat account

The command **rhnreg_ks** can be used to activate your Red Hat account. For full details regarding installation numbers and activating your Red Hat account see:

http://www.redhat.com/support/resources/faqs/installation_numbers/index.html#what_is

WARNING

**Do not update the Red Hat RPMs from the Red Hat web site as Bull cannot guarantee the continued functioning of your BAS5 for Xeon cluster. Contact Bull technical support for more information regarding when the Red Hat and Bull RPMs can be updated.**

# Glossary and Acronyms

## A

**ACT**
Administration Configuration Tool

**API**
Application Programmer Interface

**ARP**
Address Resolution Protocol

## B

**BAS**
Bull Advanced Server

**BIOS**
Basic Input Output System

## 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

**EIP**
Encapsulated IP

**EPIC**
Explicitly Parallel Instruction set Computing

**EULA**
End User License Agreement (Microsoft)

## 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.

**GUI**
Graphical User Interface

**GUID**
Globally Unique Identifier

## H

**HDD**
Hard Disk Drive

**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 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

**OEM**
Original Equipment Manufacturer

**OPK**
OEM Preinstall Kit (Microsoft)

**OST**
Object Storage Target

## P

**PAM**
Platform Administration and Maintenance Software

**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

## Q

## R

**RAID**
Redundant Array of Independent Disks

**ROM**
Read Only Memory

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

## S

**SAFTE**
SCSI Accessible Fault Tolerant Enclosures

**SCSI**
Small Computer System Interface

**SDP**
Socket Direct Protocol

**SDPOIB**
Sockets Direct Protocol over Infiniband

**SDR**
Sensor Data Record

**SFP**
Small Form-factor Pluggable transceiver - extractable optical or electrical transmitter/receiver module.

**SEL**
System Event Log

**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

## X

**XHPC**
Xeon High Performance Computing

**XIB**
Xeon InfiniBand

# Index


## A

Adaptec
  RAID Configuration Utility, G-1

adapters placement, H-1

Apache server, 4-6

## B

backbone network, 1-10

BAS4 for Xeon v1,2, B-1

bind attribute, F-1

Brocade switch
  configuration, 9-15
  enabling, 4-18

## C

CISCO Switch
  configuration, 9-8

CLOS, 8-7

cluster
  definition, 1-1

Cluster DB, B-1
  Migration, B-1

ClusterDB
  Reinstalling, B-3
  Saving, B-2

clusterdb.cfg, 4-2

Commands
  config_ipoib, 3-64, E-1
  preparenfs, D-1
  rhnreg_ks, I-1

Compilers
  Fortran, 7-1, 7-2
  installation, 7-1
  Intel, 7-1

config_ipoib command, 3-64, E-1

configuration
  Ganglia, 3-31, 3-50
  Lustre file system, 6-5
  network, 3-17
  NTP, 3-32
  overview, 3-2
  postfix, 3-29
  switches, 9-1

Configuring FTP, 8-13

Conman, 1-14

## D

database
  dump, 3-25
  initialization, 3-24
  register storage information, 4-1

ddn_admin.conf file, 4-8

ddn_init command, 4-10

ddn_set_up_date_time.cron file, 4-8

debuggers (Intel)
  installation, 7-2

disk partitioning, 3-10

## E

Ethernet adapters, E-2

## F

fcswregister command, 4-18

FDA Storage Systems
  Configuring, 4-3
  GUI Client, 4-3
  iSMsvr conf file, 4-4
  Linux
    ssh access, 4-5
  Linux Systems, 4-4
  Storage Manager server, 4-4

File
  Group.csv, 8-14

Fortran
  installation, 7-1, 7-2

fsck, 1-14

fstab file, 3-20

## G


Ganglia
  configuration, 3-31, 3-50

gmetad.conf file, 3-31

gmond.conf file, 3-31, 3-50

golden image
  creating, 3-62


## H


*HCA-400 Ex-D Interface, 8-1*


## I


InfiniBand, 8-1

InfiniBand adapters, E-2

InfiniBand interfaces
  Configuring, E-1

Infiniband Networks, 1-10

installation
  Ksis server, 3-61
  management Node, 3-5
  overview, 3-2

Intel debugger
  installation, 7-2

Intel libraries, 7-1

Intel Trace Tool
  installation, 7-2

intelruntime-cc_fc rpm, 7-1

ISR 9024 Grid Switch, 8-2


## K


Ksis server
  installation, 3-61


## L


Linux
  rdesktop command, 4-3

load_storage.sh, 6-8

lsiocfg command, 4-17

Lustre
  upgrade_lustre_layout.sh script, C-2

Lustre 1.4.11, C-1

Lustre 1.6.3, C-1

Lustre file system
  configuration, 6-5

Lustre MGS entity, C-2

Lustre Migration, C-1

Lustre Post configuration, C-2

lustre.cfg file, 6-9

lustre_investigate command, 6-12


## M


mount points (cdrom), 3-20

MPI libraries
  MPIBull2, 1-15


## N


nec_admin command, 4-6

nec_admin.conf file, 3-4, 4-6

network
  administration network, 1-10
  administration network, 3-18
  backbone, 1-10
  configuration, 3-17

Network Time Protocol (NTP), 3-32

node
  compute node, 1-7
  login node, 1-6
  Management Node, 1-6

NTP
  configuration, 3-32

ntp.conf file, 3-32


## O


oppensl, 3-38


## P


partitioning
  disk, 3-10

PCI slots selection, H-1

postfix configuration, 3-29

postfix/main.cf file, 3-29

-prod option, D-1

## S

saving
  ClusterDB, 3-3
  Lustre file system, 3-4
  ssh keys, 3-4
  storage information, 3-4

SLURM and openssl, 3-38

SLURM and Security, 3-38

ssh
  saving keys, 3-4

ssh-keygen, 4-5

storageadmin directory, 3-4

storcheck.cron file, 4-2

storframework.conf file, 3-4

switch configuration, 9-1

syslog-ng
  port usage, 3-31
  service, 3-32

syslog-ng.conf file, 3-31

syslog-ng/DDN file, 4-8

system-config-network command, 3-17

## T

Trace Tool (Intel)
  installation, 7-2

## V

Voltaire device
  configuration, 9-16

Voltaire Device Manager (VDM), 9-16

Voltaire Fabric Manager (VFM), 9-16

Voltaire GridVision Fabric Manager, 8-16

Voltaire Performance Manager, 8-11

Voltaire switch topology, 8-7

Voltaire Switching Devices, 1-10

## W

wwn file, 4-17

WWPN description, 4-17

## X

xinetd.conf file, F-1

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

REFERENCE
86 A2 87EW 01