HPC

BAS5 for Xeon

Installation and Configuration Guide

REFERENCE
86 A2 87EW 00

HPC

BAS5 for Xeon

Installation and Configuration Guide

Hardware and Software

April 2008

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

REFERENCE
86 A2 87EW 00

Printed in France

Suggestions and criticisms concerning the form, content, and presentation of this book are invited. A form is provided at the end of this book for this purpose.

To order additional copies of this book or other Bull Technical Publications, you are invited to use the Ordering Form also provided at the end of this book.

Trademarks and Acknowledgements

We acknowledge the rights of the proprietors of the trademarks mentioned in this manual.

All brand names and software and hardware product names are subject to trademark and/or patent protection.

Quoting of brand and product names is for information purposes only and does not represent trademark misuse.

Preface

Scope and Objectives

This guide describes how to install the Bull HPC **BAS5 for Xeon v1.1 (**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 HPC systems that either need to re-install the BAS5 for Xeon software, or to update existing software with the newer version, or to install a new application.

Prerequisites

Refer to the **BAS5 for Xeon v1.1** *Software Release Bulletin* (SRB).

Structure

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

Chapter 3. *Configuring Storage Management Services*
Describes how to configure the storage management software to manage the storage systems of the cluster.

Chapter 4. *Configuring and Sharing 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 5. *Installing Tools and Applications*
Describes how to install commercial tools (Intel Compilers and MKL) and other applications (Modules).

Chapter 6. *Installing and Configuring InfiniBand Interconnects*
Describes the tasks for the installation and configuration of different Voltaire Devices.

Chapter 7. *Configuring Switches and Card*
Describes how to configure CISCO and Foundry Ethernet switches, Voltaire InfiniBand, and Brocade switches.

Chapter 8. *Checking and Backing-up Cluster Nodes*
Describes how to check the nodes, the installed software, the release and how to make a backup.

Appendix A. *Interconnect Interfaces*
Describes the Interface description file.

Appendix B. *Recommendation for PCI Slot Selection*
Gives some recommendations to optimize the choice of PCI Slots for high bandwidth PCI adapters.

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

Appendix D. *Activating your Red Hat Account*

Glossary and Acronyms
Lists the Acronyms used in the manual.

Bibliography

- Bull *BAS5 for Xeon Administrator's Guide* (86 A2 88EW)
- Bull *BAS5 for Xeon User's Guide* (86 A2 89EW)
- Bull *BAS5 for Xeon Maintenance Guide* (86 A2 90EW)
- A *Software Release Bulletin* (SRB) (86 A2 64EJ) includes additional installation instructions and specific information for each software release.
- NovaScale Master *Remote HW Management CLI Reference Manual* (86 A2 88EM)
- NovaScale Master *Installation Guide* (86 A2 48EG)
- NovaScale Master *Administrator's Guide* (86 A2 50EG)
- FDA Storage Manager - *Configuration Setting Tool Users Manual* (GUI) (86 A2 88EG)
- Bull *Voltaire Switches Documentation CD* (86 A2 79ET 00)
- Bull *NovaScale R421 Installation and User's Guide* (86 A1 94ET)
- Bull *NovaScale R422 Installation and User's Guide* (86 A1 95ET)
- Bull *NovaScale R422-E1/R422-INF-E1 Installation and User's Guide* (86 A1 93EW)
- Bull *NovaScale R421-E1 Installation and User's Guide* (86 A1 94EW)
- Bull *NovaScale R423 Installation and User's Guide* (86 A1 95EW)
- CLARiiON *CX3-40f Setup Guide* (300-004-208)
- EMC *Navisphere® Command Line Interface (CLI)* (300-003-628)
- StoreWay *Optima 1250 Quick Start Guide* (86 A1 34ET)
- StoreWay *Optima 1250 Installation and User Guide* (86 A1 35ET)
- StoreWay *Master User Guide* (86 A1 38ET)

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

- PBS Professional 9.0 *Administrator's Guide* (on PBS Pro CD-ROM)
- PBS Professional 9.0 *User's Guide* (on PBS Pro CD-ROM)

Highlighting

- Commands entered by the user are in a frame in "Courier" font. Example:

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

- Commands, files, directories and other items whose names are predefined by the system are in "Bold". Example:
  The **/etc/sysconfig/dump** file.
- Text and messages displayed by the system to illustrate explanations are in "Courier New" font. Example:
  `BIOS Intel`
- Text for values to be entered in by the user is in "`Courier New`". Example:
  `COM1`
- *Italics* Identifies referenced publications, chapters, sections, figures, and tables.
- < > identifies parameters to be supplied by the user. Example:
  `<node_name>`

Warning:

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

Table of Contents

List of Figures

List of Tables

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 systems feature different **NovaScale Xeon** machines for the nodes.

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 HPC** system supports various types of nodes, dedicated to specific activities.

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 includes minimal functionality, are quicker and easier to deploy, and requires 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 includes additional libraries and require more disk space for their installation. These are used for applications that require a graphical environment (XWindows), and also for most ISV applications. They are also installed if there is a need for **Intel Cluster Ready** compliance.

In addition to the Compute Nodes, a certain number of **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 different functions as shown in the diagrams below.

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.

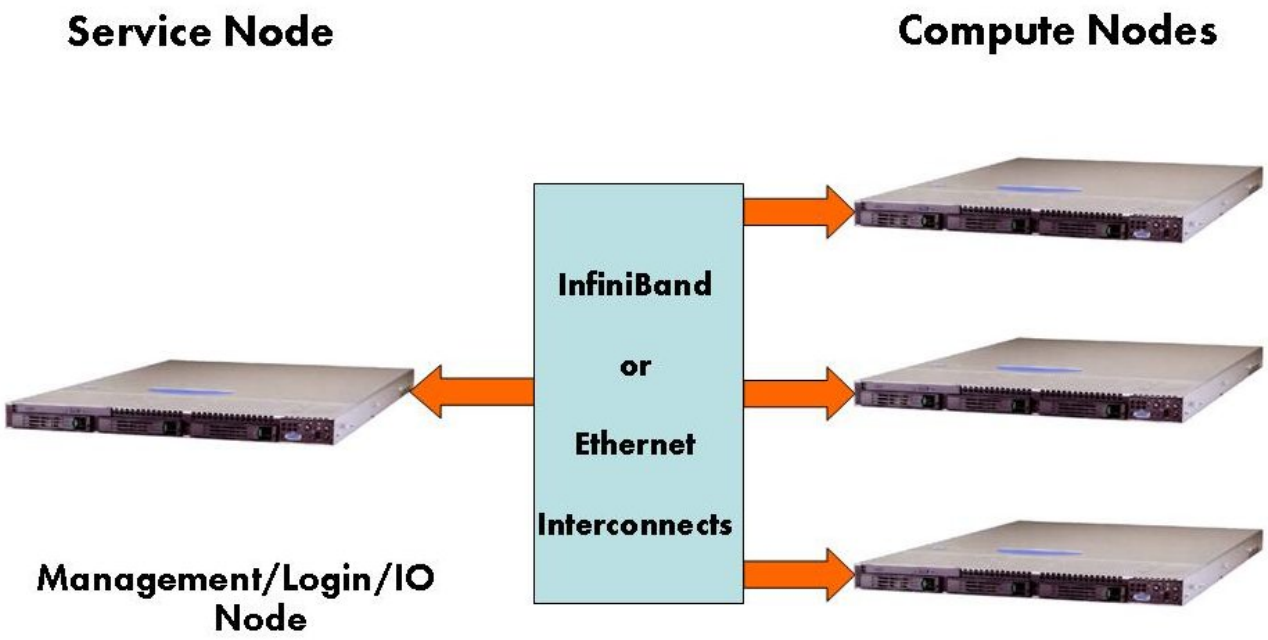

Figure 1-1. Small Cluster Architecture

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.

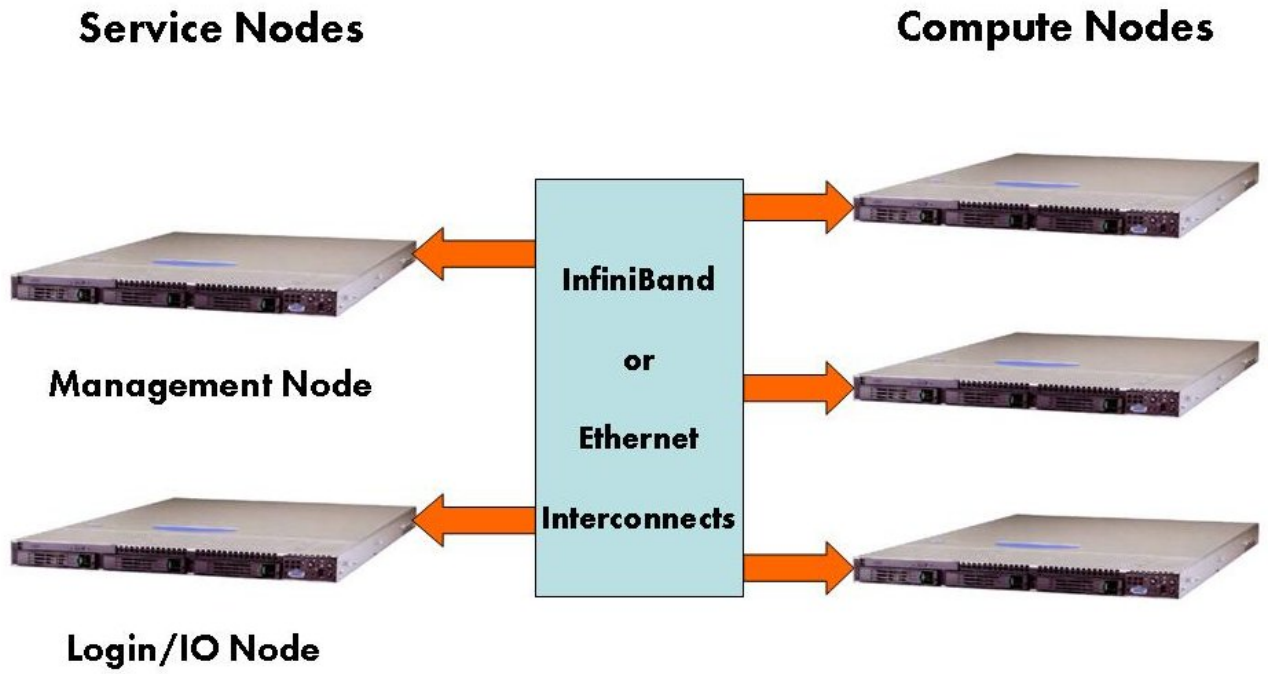

Figure 1-2. Medium-sized Cluster Architecture

1.2.2.3 Large clusters

On large clusters the cluster management services run on dedicated nodes. The Login and I/O services will 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.

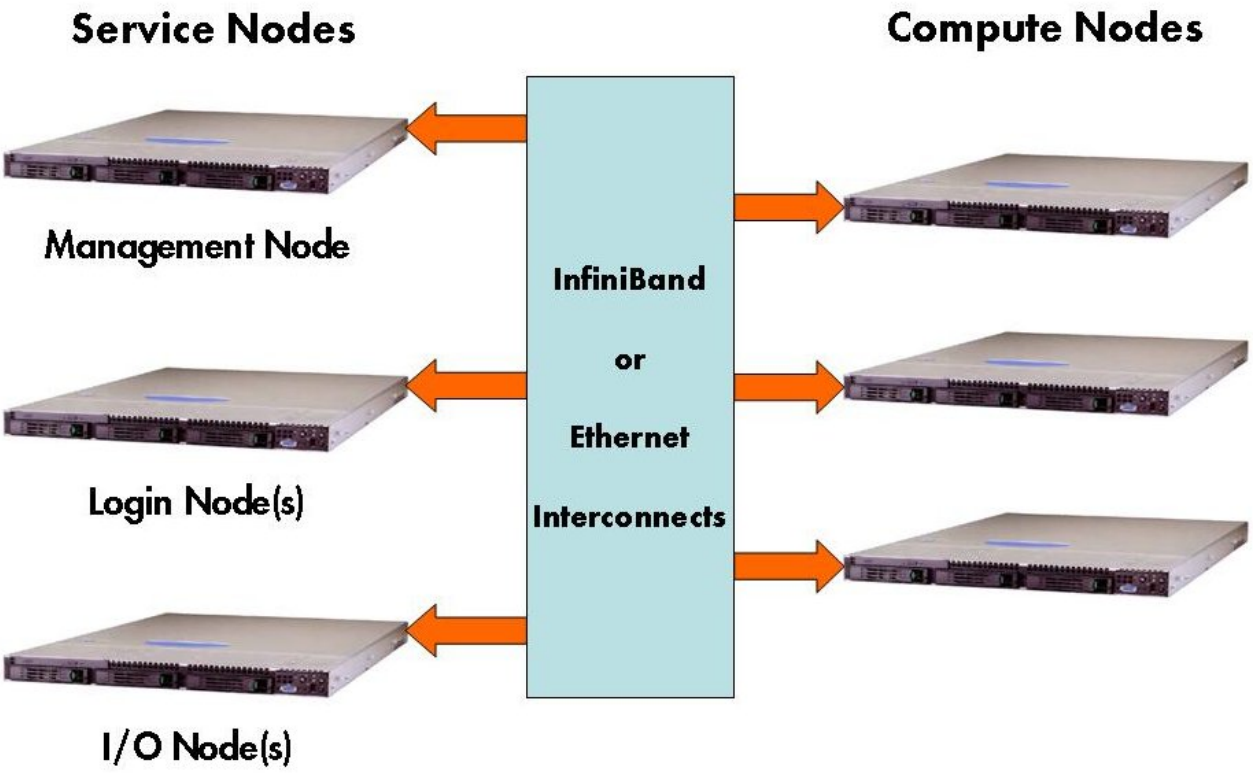

Figure 1-3. Large Cluster Architecture

1.2.3 Service node(s)

Bull **NovaScale R440, R460** and **R423** 2 socket Xeon machines are used for the Service Nodes for Bull **BAS5 for Xeon** Clusters.

Figure 1-4. NovaScale R440 machine

NovaScale R440 machines support **SATA2, SAS, SAS 2.5** storage systems.

Figure 1-5. NovaScale R460 machine

NovaScale R460 machines support **SAS** and **SATA2** storage systems.

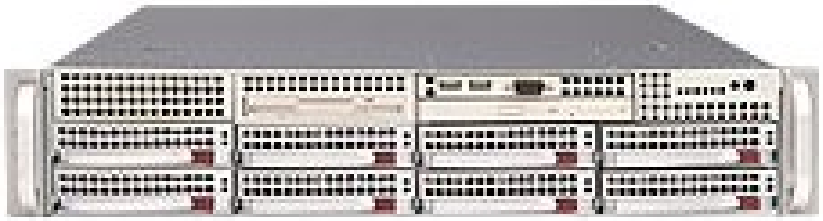

Figure 1-6. NovaScale R423 machine

NovaScale R423 machines support **SAS,** and **SATA2** 3.5 inch storage systems.

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, snmtrapd, 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 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.

Refer to *Appendix B* in this manual for information on how to select the best PCI slots for optimum performance.

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

Bull **NovaScale R421, R421 E1**, **R422,** and **R422 E1** machines may all be used for the Compute Nodes for **BAS5 for Xeon v1.1**.

Bull **NovaScale R422** and **R422 E1** machine includes 2 nodes.

Figure 1-7. NovaScale R421 machine

Figure 1-8. NovaScale R421 E1 machine

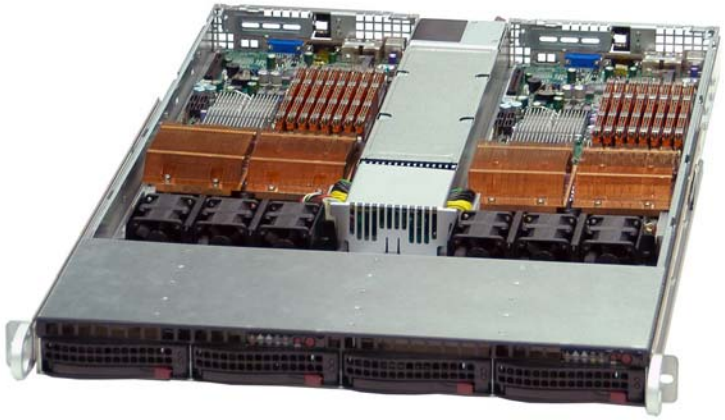

Figure 1-9. NovaScale R422, R422 E1 machine

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

1.2.5 Networks

The cluster may contain 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 which allows the management of operating systems, 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 9024** 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.

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 Connect™ InfiniBand cards support Dual 4X ports providing a bandwith of 10 or 20 Gb/s per port. They support PCI Express 2.0 but are compatible with PCI-Express 1.1 and fit x8 or x16 slots.

Important:
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 combination RJ45-SFP ports in which the port may be used as either a 10/100/1000 Mbps copper Ethernet or as a fiber 100/1000 Mbps when using an SFP transceiver in the corresponding SFP port. The **FastIron LS 624** includes three 10-Gigabit Ethernet slots that are configurable with single-port 10-Gigabit Ethernet pluggable modules.
- The **FastIron LS 648** supports forty-eight 10/100/1000 Mbps RJ-45 Ethernet ports. Four of these ports are implemented as combination RJ45-SFP ports in which the port may be used as either a 10/100/1000 Mbps copper Ethernet or as a fiber 100/1000 Mbps 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 a back –up power operation 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.
- For more information on configuring Ethernet switches see *Appendix A* in the **BAS5 for Xeon** *Installation and Configuration Guide*.

1.2.7 Storage

Different storage systems are supported by **BAS5 for Xeon**. These include the following:

Storeway 1500 and 2500 FDA Storage systems

Based on the 4Gb/s FDA (Fibre Disk Array) technology, the networked 1500 and 2500 FDA Storage systems support transactional data access, associated with fibre and SATA disk media hierarchy. RAID6 double-parity technology enables continued operation even in the case of two disk drive failures, thus providing 100 times better data protection than RAID5.

Brocade Fibre Channel switches are supported 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. 2x 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) CX3-40f

The **CX3-40f** model benefits from the high performance, cost-effective and compact UltraScale architecture. It supports Fibre Channel connectivity, with 8 GB cache memory, and fits perfectly within SAN infrastructures; it offers a complete suite of advanced storage software, in particular **Navisphere Manager,** to simplify and automate the management of the storage infrastructure. 8 x 4 Gb/s FC front-end and back-end ports are included.

Note:

The **EMC Clariion CX300** storage system 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 960TB 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 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 which enable the operation and monitoring of the cluster, including:

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

The consoles, accessed using **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.

NSCommands 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 and will trigger alerts in the event of any problems. NovaScale Master uses **Ganglia**, a second open source tool, to collect and graphically display performance statistics for each cluster node.

1.3.2 Program Execution Environment

1.3.2.1 Resource Management

Both **Gigabit Ethernet** and **InfiniBand BAS5 for Xeon** clusters use the **SLURM** (Simple Linux Utility for Resource Management) open-source, highly scalable cluster management and job scheduling system. **SLURM** allocates compute resources, in terms of processing power and Computer 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. Finally, SLURM helps to exploit the parallel processing capability of a cluster.

See the Bull HPC 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 with each other through calls to subroutines of the MPI library.

Bull provides **MPIBull2,** Bull's second generation MPI library in the Bull **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 Bull **BAS5 for Xeon** *User's Guide* for more information on Parallel Libraries

1.3.2.3 Batch schedulers

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

See the Bull **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.

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

1.4 Bull BAS5 for Xeon software distribution

1.4.1 Installing Software and Configuring Nodes

Before installing the **BAS5 for XEON** software on the nodes, the node distribution architecture planned for your **HPC** system (Management Nodes, Compute Nodes, Login Nodes, I/O Nodes) must be known.

Chapter 2 explains how to install **BAS5 for Xeon** distribution on a Management Node and how to use the **Prepare NFS** script to install the **RPM**s for the specific services required for each type of node.

The software installed on a **Compute**, **Login** or **I/O Node** 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 **Reference Node** designates the node from which the reference image is taken.

Chapter 2. Installing BAS5 for Xeon V1.1 Software on the HPC Nodes

This chapter describes the complete installation process for the FIRST installation of the **BAS5 for Xeon** v1.1 software environment on all nodes of a Bull **HPC** cluster. The same process can also be used for a reinstallation of **BAS5 for Xeon** v1.1 using existing configuration files – see section *2.0*

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.

Important:
Read this chapter carefully and install the **BAS5 for Xeon v1.1** software that applies to your cluster.

Installation Process Overview

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

Pre-installation Operations when Re-installing BAS5 for Xeon **Skip this step if you are installing for the first time.** 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**.		
STEP 1	Installing the RHEL5.1 software on the Management node 1) Installation of the **Red Hat Enterprise Linux 5** Management Node System software 2) Configuring Disk Health Monitoring 3) Configuring the Network 4) Installing an external Storage System	Page 2-5
STEP 2	Installing Bull BAS5 for Xeon software on the Management Node 1) Installing Bull **XHPC**, **XIB** and **XLustre** software 2) Database Configuration	Page 2-21
STEP 3	Configuring equipment and Installing other software on the Management Node 1) Configuring Equipment Manually (small clusters only) 2) Configuring Ethernet switches 3) Configuring some of the management tools 4) Installing and configuring **Ganglia**, **Syslog-ng**, **NTP**, **Postfix**, **Kdump**, **SLURM** and **PBS Pro** on the Management Node 5) Installing compilers (only on Management Nodes which include Login functionality) 6) Configuring the User environment, and in particular **MPI**	Page 2-28
STEP 4	Installing RHEL5.1, BAS5v1.1 for Xeon Software, and optional HPC software products 1) Specifying the software and the nodes to be installed 2) Running the **preparenfs** script	Page 2-44
STEP 5	Configuring Administration Software on Login, I/O, Compute and Computex Nodes 1) Installing and configuring **ssh**, **Kdump, SLURM** and **PBS Pro** as necessary 2) Installing compilers on Login Nodes 3) Configuring the User environment, and in particular **MPI**	Page 2-51
STEP 6	Creating an image and deploying it on the cluster nodes using Ksis 1) Installation and configuration of the image server 2) Creation of the image of a Compute or Login Node previously installed 3) Deployment of this image on cluster nodes 4) Post Deployment tests	Page 2-63

2.0 Pre-installation Operations when Re-installing BAS5 for Xeon

This step describes how to save the **ClusterDB** database and other important configuration files. Use this step only in the case of a **re-installation**, when the cluster has already been configured (or partially configured), and there is the need to save its configuration.

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.

Important:

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.

2.0.1 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.
5. Note the name of the file system where the ClusterDB mount point is (for example `/dev/sdv`). This only applies when there is an external storage system.

2.0.2 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 2.5.1 *Configuring SSH*).

2.0.3 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
- **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.

2.0.4 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.

2.0.5 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.

2.1 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. Configuration of Disk Health Monitoring.
2. Configuration of the network.
3. Installation of External Storage System.

2.1.1 Red Hat Enterprise Linux 5 System Installation

2.1.1.1 Initial Steps

Important:
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.

2.1.2 Red Hat Enterprise 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 2-1. The Welcome Screen

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

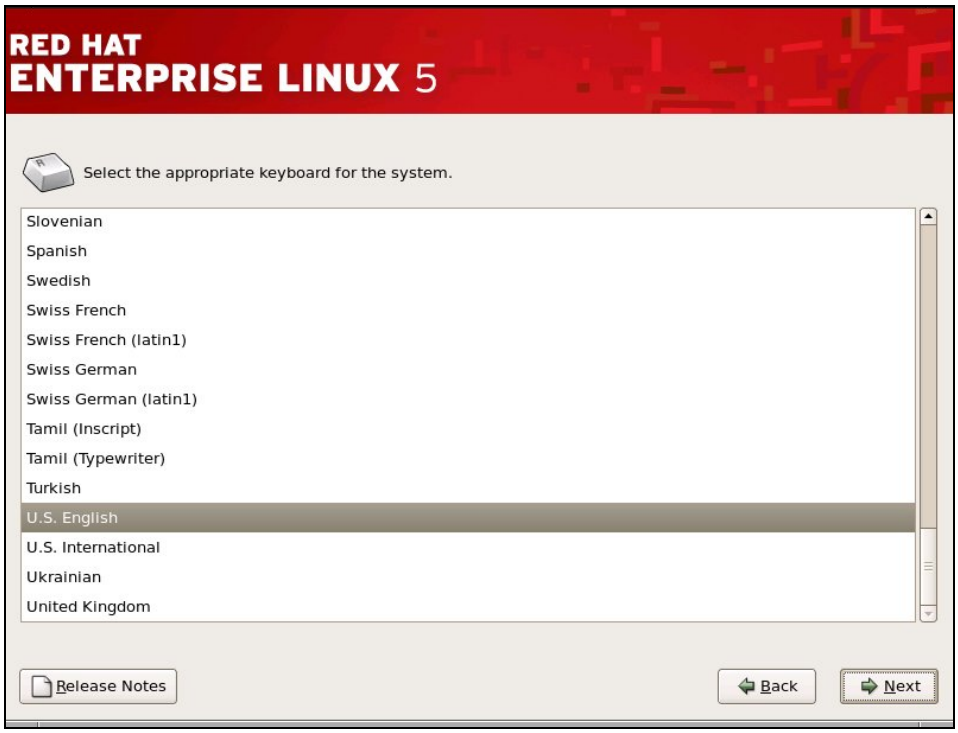

Figure 2-2. Keyboard installation screen

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.

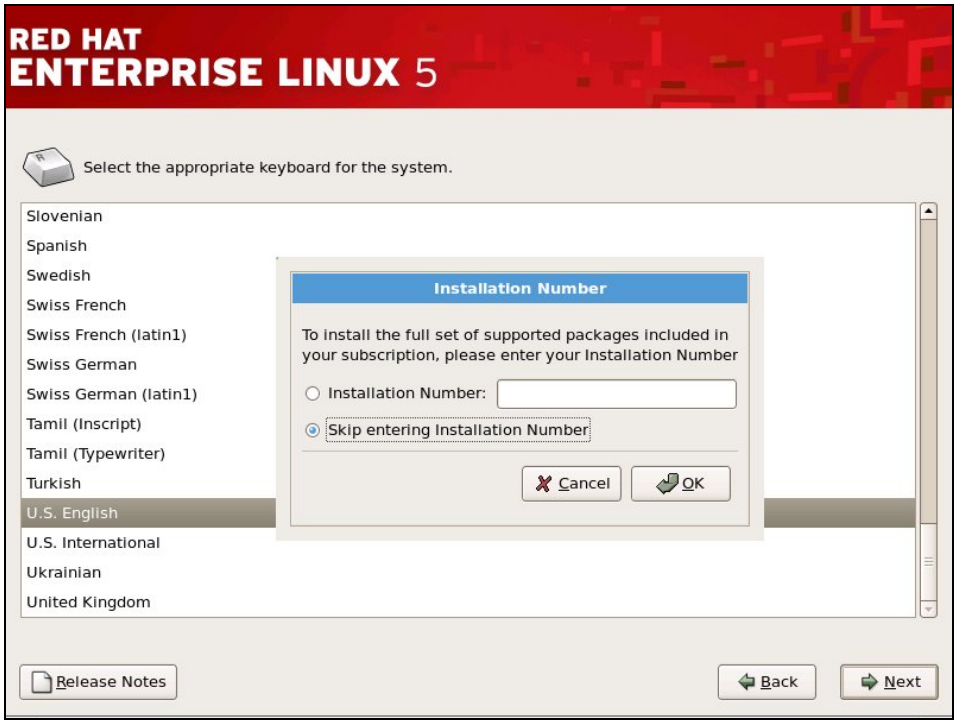

Figure 2-3. RHEL5 installation number dialog box

Figure 2-4. 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 2.4. Click on **Next**.

Important:
See Appendix D - for important information regarding the use of installation numbers.

Figure 2-5. First RHEL5 installation screen

4. Select the option **Install Red Hat Enterprise Linux Server** as shown in Figure 2-5.

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.

2.1.3 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.

2.1.3.1 Default partitioning

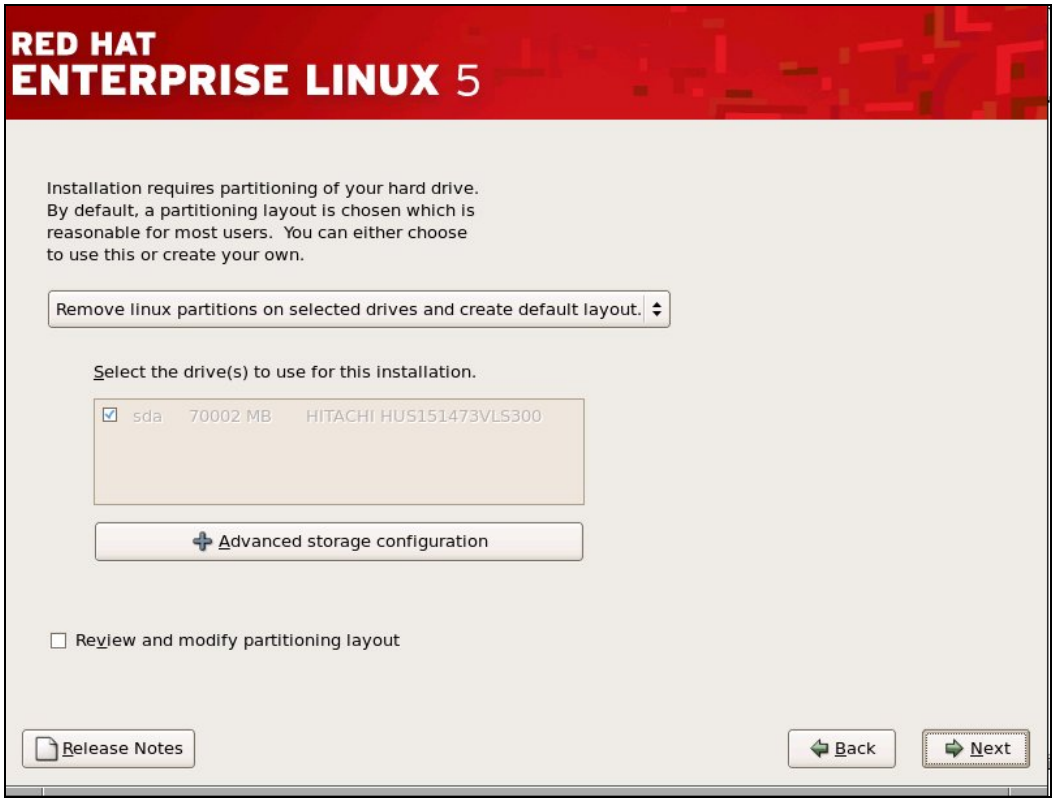

Figure 2-6. Partitioning screen

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 2-6 and will be reformatted and the Red Hat software installed on them. **Deselect those disks where your wish to preserve the existing data.**

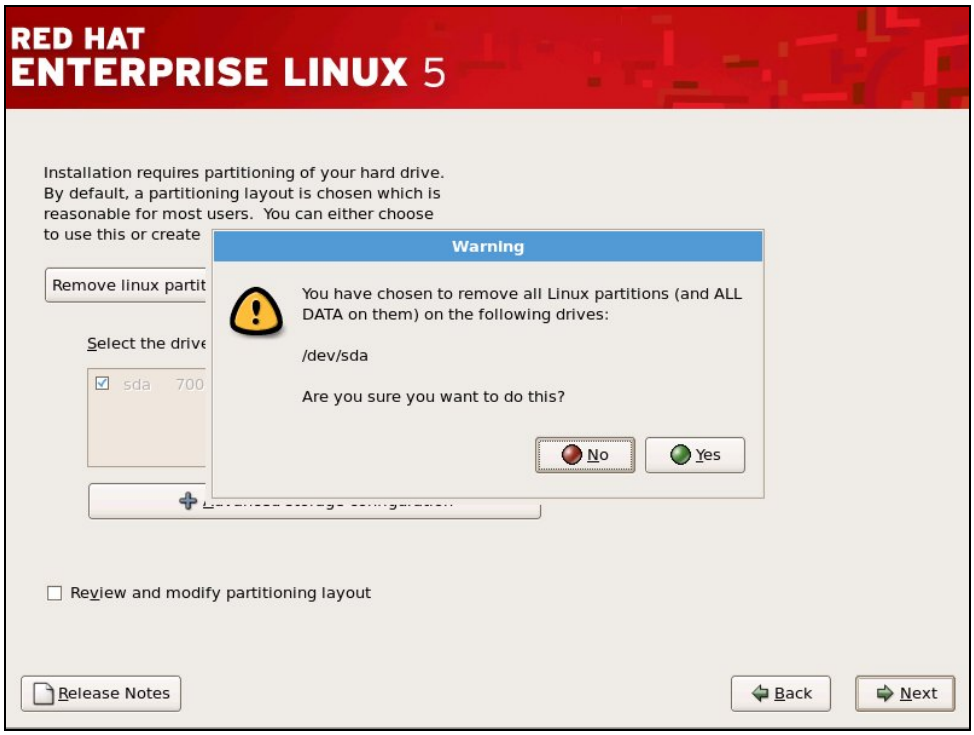

Figure 2-7. Confirmation of the removal of any existing partitions

Select **Yes** to confirm the removal of any existing partitions as shown in Figure 2.7, if this screen appears.

If the default partitioning is to be left in place go to section 2.1.3.3 *Network access Configuration.*

2.1.3.2 Reinstallation using the existing partition layout

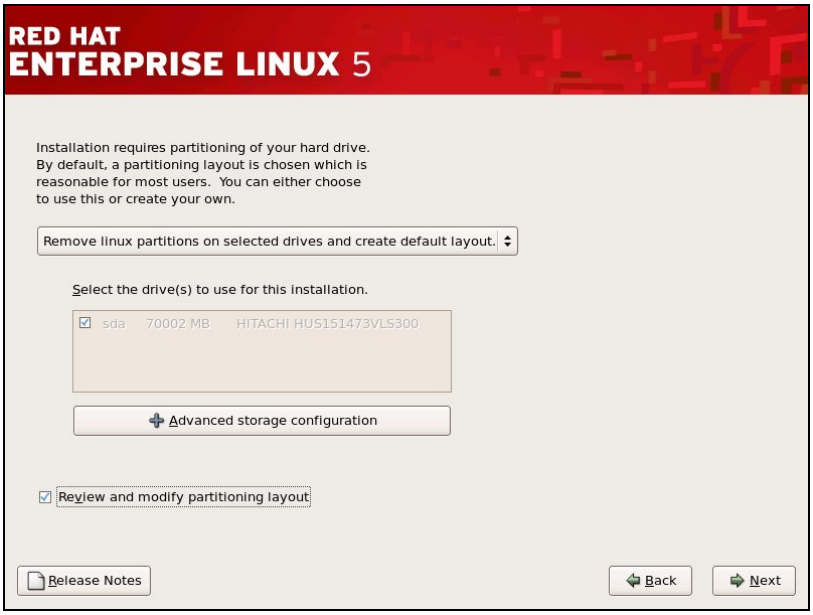

Figure 2-8. Modifying the partitioning layout – 1st screen

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

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

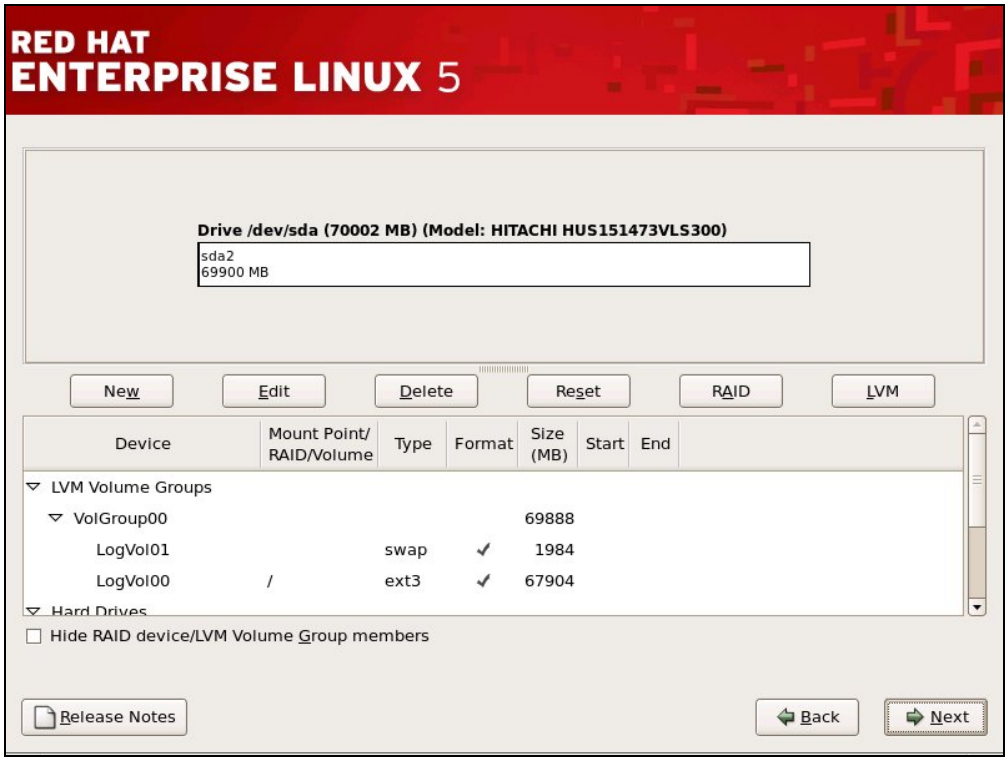

Figure 2-9. RHEL5 Partitioning options screen

c. If you wish to keep the partitioning options as they were previously, click on **Reset** in the screen above, as shown in Figure 2-9, and confirm the settings, including the mount point, that appear.

Figure 2-10. Confirmation of previous partitioning settings

2.1.3.3 Network access Configuration

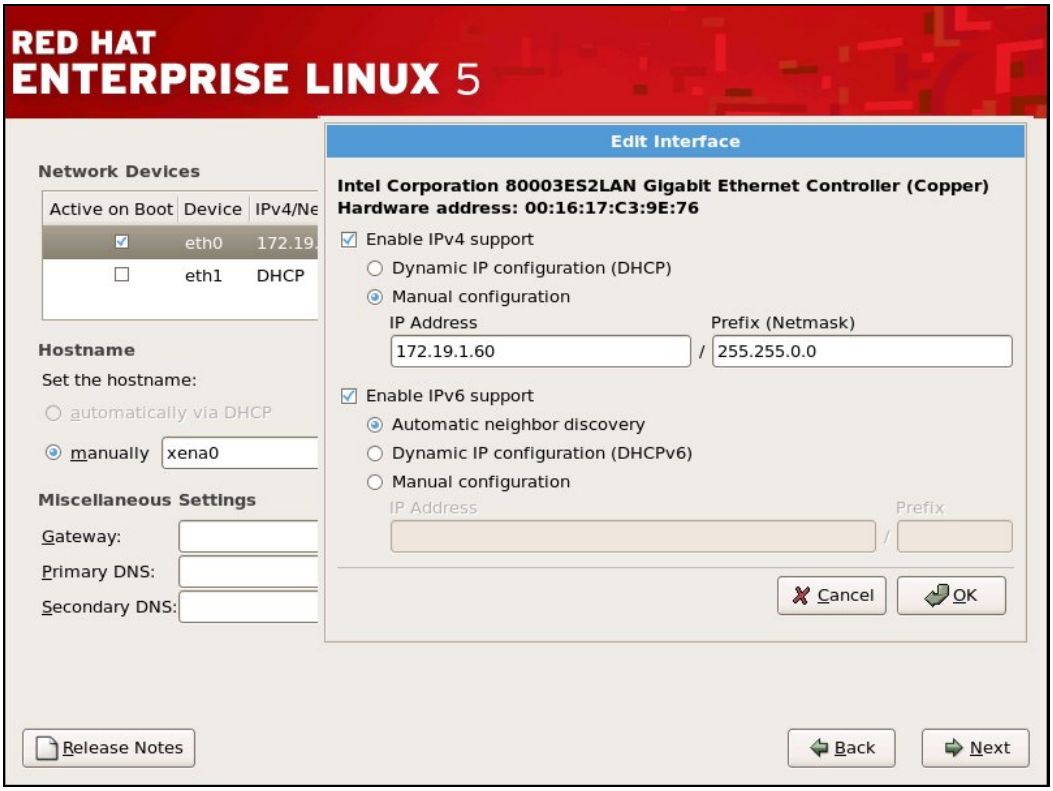

Figure 2-11. Network Configuration Screen

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 2-11.

   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 2-11 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.

2.1.3.4 Time Zone Selection and Root Password

Figure 2-12. Time Zone selection screen.

7. Select the Time Zone settings required, as shown in Figure 2-12, and click on **Next.**

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

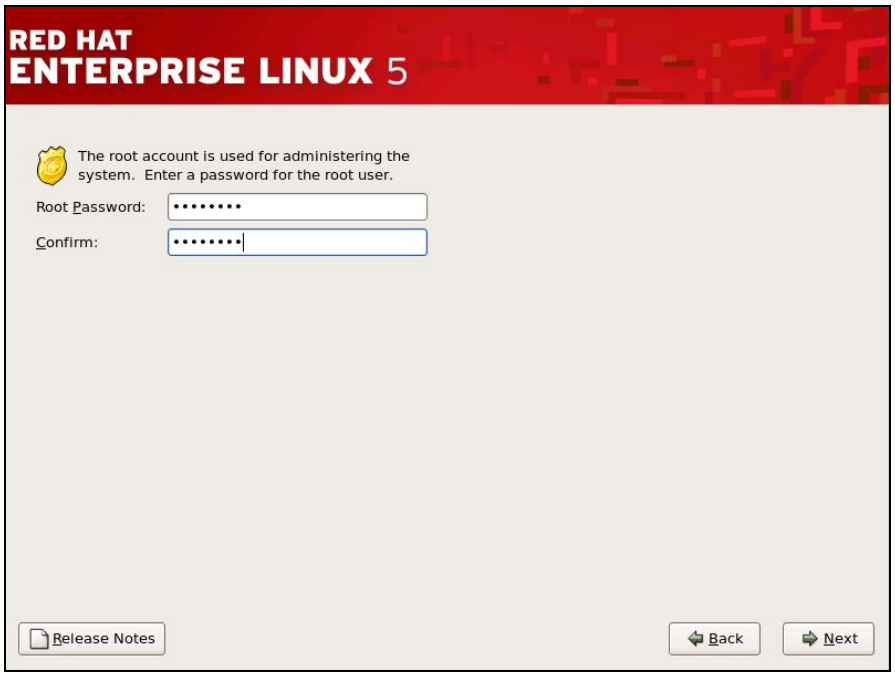

Figure 2-13. Root Password Screen

8. Set the Root password as shown in Figure 2-13. This must use a minimum of 6 characters.

2.1.3.5 Red Hat Enterprise Linux 5 Package Installation

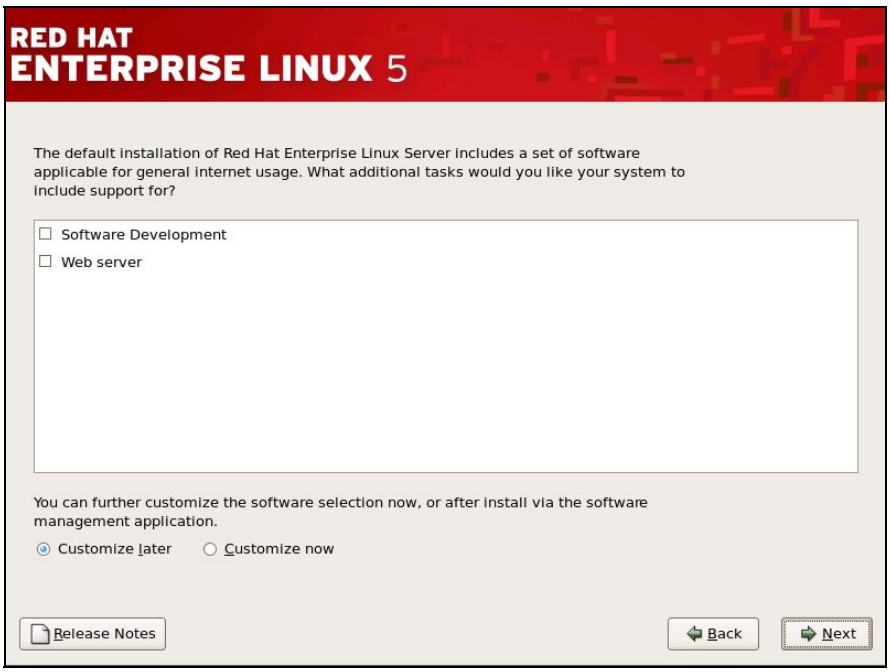

Figure 2-14. Software selection screen

9. Leave the screen with the additional tasks **deselected**, as shown in Figure 2-14. Click on **Next.**

Figure 2-15. Installation screen

10. Click on **Next** in Figure 2-15 to begin the installation of **Red Hat Enterprise Linux Server.**

11. Following the reboot after the **Congratulations the installation is complete** screen there may be problems with the screen appearance with the bottom of the screen removed. Carry out the procedure below to avoid these problems **before rebooting**.

    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.

2.1.3.6 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. Select **No** for in reply to the question **Are there are any additional CDs to install?**.
18. Click on **Finish**
19. Click on **Reboot**

2.1.4 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. This standard command opens the graphical tool used for this section.

Run the command:

```
system-config-network
```

2.1.4.1 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
```

Important:
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 **HPC BAS5 for Xeon** clusters do not support VLAN.

2.1.4.2 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 2.5.

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
```

2.1.4.3 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.

2.1.4.4 Updating the /etc/hosts file

1. Edit the **/etc/hosts** file:
2. Remove the server name before `localhost` (xena0 in the example below):

   ```
   127.0.0.1 xena0 localhost.localdomain localhost
   ```

3. Add the **<IP address> <basename>0** line which applies to the server. This has been set previously – see section *2.1.3.3*:

2.1.4.5 Restarting the network service

Run the command:

```
service network restart
```

2.1.5 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. Refer to the documentation provided with the storage system for details on how to install the storage system.

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

This step describes how to install the Bull **BAS5 for Xeon** 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 XHPC** software
3. Preparation for the Installation of the **BAS5 for Xeon** optional software
4. Installation of Bull **BAS5 for Xeon** 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 will have been loaded that are specifically required for Bull **BAS5 for Xeon** HPC Clusters. The following section describes how these packages are installed along with the Bull **XHPC** software, and optional **InfiniBand, XLustre** and **XToolkit** software.

2.2.1 Preparing the Installation of the Red Hat software on other cluster nodes

1. Create the directory for the software:

```
mkdir –p /release/RHEL5.1
```

2. Create a mount point for the **BAS5 for Xeon XHPC** 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.

Important

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. Run the command below:

```
mkdir –p /media/cdrecorder
```

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

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

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

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

11. Eject the DVD:

```
umount /dev/cdrom
```

or use the **eject** command:

```
eject
```

2.2.2 Preparing the Installation of the BAS5 for Xeon XHPC software

1. Create the directory for the **BAS5v1.1 XHPC** software:

```
mkdir –p /release/XBAS5V1.1
```

2. Insert the **BAS5v1.1 XHPC** DVD-ROM into the DVD reader and mount it:

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

3. Copy the **BAS5v1.1 XHPC** DVD-ROM contents into the **/release** directory:

```
cp -a /media/cdrecorder/* /release/XBAS5V1.1/
```

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

2.2.3 Preparing the Installation of the BAS5 for Xeon XIB 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/XBAS5V1.1** 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 **BAS5v1.1 XIB** DVD-ROM into the DVD reader and mount it:

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

2. Copy the **BAS5v1.1 XIB** DVD-ROM contents into the **/release** directory, as shown below:

```
unalias cp
cp -a /media/cdrecorder/* /release/XBAS5V1.1
```

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 **BAS5v1.1 XLUSTRE** DVD-ROM into the DVD reader and mount it:

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

2. Copy the **BAS5v1.1 XLUSTRE** DVD-ROM contents into the **/release** directory:

```
unalias cp
cp -a /media/cdrecorder/* /release/XBAS5V1.1
```

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

2.2.4 Installing the Bull BAS5 for Xeon software

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

```
cd /release/XBAS5V1.1
```

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 the description of the different architecture and functions possible.

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

```
./install -func MNGT [IO] [LOGIN] -prod RHEL XHPC [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:

- **RHEL** to install the mandatory **RHEL** RPMs
- **XHPC** to install the Bull **BAS5 for Xeon** software
- **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 **Red Hat** software and the Bull **BAS5v1.1 for Xeon XHPC** software when there is a dedicated Management Node, with additional Service Nodes for the I/O and Login functions:

```
./install -func MNGT -prod RHEL XHPC
```

By default all available **BAS5 for Xeon** products and mandatory **RHEL** packages will be installed if the **–prod** option is not specified.

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>.rpm
```

   For example:

```
yum install java-1.5.0-bea-1.5.0.08-1jpp.5.el5.x86_64.rpm
```

See Chapter 11 in the Bull **BAS5 for Xeon** *Administrator's Guide* for details on configuring and using **HPC Toolkit.**

2.2.5 Database Configuration

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

- *First Installation - Initialize Cluster Management Database*
- *Re-installation of BAS5 for Xeon with ClusterDB Preservation*

2.2.5.1 First Installation - Initialize Cluster Management Database

Note:
This paragraph applies only when performing the **first installation** of **BAS5 for Xeon**.

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
```

2.2.5.2 Re-installation of BAS5 for Xeon with ClusterDB Preservation

Note:

This paragraph applies when re-installing an **existing version** of **BAS5 for Xeon** withthe restoration of its existing ClusterDB.

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 ClusterDB 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`.

For more details about restoring data, refer to the HPC BAS5 for Xeon *Administrator's Guide*.

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

2.3 STEP 3: Configuring Equipment and Installing other software on the Management Node

This step describes how to:

- Configure equipment
- Configure Ethernet switches
- Configure some of the management tools
- 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**

Important:

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 section 2.3.4 *Configuring Management Tools Using Database Information*.

2.3.1 Configuring Equipment

Important:

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

Collect the MAC address for each node in the cluster and configure the hardware manager for these nodes.

Look for 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
- Update the database with the node and hardware manager MAC addresses for the rack by running the command:

```
/usr/lib/clustmngt/clusterdb/install/updateMacAdmin <file name>
--rack <rack label>
```

- Configure the IP addresses for the **BMC** of the rack by running the command:

```
/usr/lib/clustmngt/BMC/bmcConfig --input <file name>
```

2.3.2 Configuring Equipment Manually

If equipment has been installed and the MAC address files have not been found you have to collect the MAC address 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 is available: **/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. Collect the domain name of each node of the cluster. 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.

2.3.3 Configuring Ethernet Switches

Important:
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 in their initial set (factory settings).

Install Ethernet switches by running the command:

```
swtAdmin auto
```

For more details see Chapter *7 Configuring Switches*.

2.3.4 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:

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

3. Save the ClusterDB:

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

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

5. Reboot the Management Node

```
exit
reboot
```

2.3.5 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 (x.x.0.99) of the Management Node.
     `host = x.x.x.x  /* replace with your administration node ip address */`

3. Start **gmond** service:

```
service gmond start
```

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

```
# data_source "my cluster" 10 localhost  my.machine.edu:8649
# 1.2.3.5:8655
# data_source "my grid" 50 1.3.4.7:8655 grid.org:8651
# grid-backup.org:8651
# data_source "another source" 1.3.4.7:8655  1.3.4.8
```

```
>>    data_source "mycluster" localhost

#
#------------------------------------------------------------------------
# Scalability mode. If on, we summarize over downstream grids, and
# respect
# authority tags. If off, we take on 2.5.0-era behavior: we do not
# wrap our output
```

to be replaced with:

```
# data_source "my cluster" 10 localhost  my.machine.edu:8649

# 1.2.3.5:8655
# data_source "my grid" 50 1.3.4.7:8655 grid.org:8651
# grid-backup.org:8651
# data_source "another source" 1.3.4.7:8655  1.3.4.8

>>    data_source "<basename>" localhost

#
#------------------------------------------------------------------------
# Scalability mode. If on, we summarize over downstream grids, and
# respect
# authority tags. If off, we take on 2.5.0-era behavior: we do not
# wrap our output
```

5. Start **gmetad**:

```
service gmetad start
```

2.3.6 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 – the x.x.0.99 alias IP address).

Restart syslog-ng

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

```
service syslog-ng restart
```

2.3.7 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 a 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
```

2.3.8 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
```

5. To activate the mail relaying at each reboot, run:

```
# chkconfig postfix on
```

2.3.9 Configuring the kdump kernel dumping tool

Kdump will have been enabled during the Red Hat installation on the Management Node – see section *2.1.3.6*

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
   ```

   a. 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 available from

http://people.redhat.com/duffy/debuginfo/index-js.html in the **kernel-debuginfo-<kernel_release>.rpm**

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
```

2.3.9.1 Testing kdump

In order to test that **kdump** is working correctly a dump may 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
```

2.3.9.2 Configuring systems to take dumps from the Management Network

In addition to forcing a dump for a kernel crash, it is possible to force a dump using the **ipmitool** command from the Management Node. This is done as follows:

Add `nmi_watchdog=0` to the kernel boot options in the **/boot/grub/menu.lst** file in order to deactivate the **NMI** watchdog used by **RHEL**, so that the other NMIs can be put into effect. An example of the menu.1st file is shown below:

```
kernel /vmlinuz-2.6.18-53.d5.ELsmp ro root=LABEL=/ nmi_watchdog=0
console=tty0 console=ttyS1,115200n8 console=ttyS0,1152,00n8 rhgb
quiet
```

Once the system has been restarted the kernel has to be reconfigured so that a panic is launched when an unknown NMI is received. This can be set to happen automatically by configuring the **kernel.unknown_nmi_panic = 1** option in the **/etc/sysct1.conf** file

Alternatively, this can be done manually by using the command.

```
echo 1 > /proc/sys/kernel/unknown_nmi_panic
```

An **NMI** dump may be launched using **IPMI** via the command:

```
ipmitool -H <bmc_address> -U <user_name> -P <pwd> chassis power diag
```

or by using the **nsctrl** command.

For more information see http://kbase.redhat.com/faq/FAQ_105_9036.shtm

Notes:

- If watchdog is still active after the **kernel.unknown_nmi_panic = 1** option is set the machine will no longer boot.
- For this release of **BAS5 for Xeon** the **IPMI power diag** command will launch a dump for **NovaScale R423**, **NovaScale R440** and **NovaScale R460** series machines.
- There is also a dump button on the back of the **NovaScale R460** series machines that will launch an **NMI** dump for these machines.

Further information can be found in the **kdump** man pages.

2.3.10 Configuring SLURM - optional

Important

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.

2.3.10.1 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, 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** number does not have to match the one indicated above, but it has to be the same on all the nodes that are included in the SLURM cluster.

The **uid** number does not have to match the one indicated above, but it has to be the same on all the nodes that are included in the SLURM cluster.

The user name in the example above is '**slurm**', another name can be used, but it has to be the same on all the nodes that are included in the SLURM cluster.

2.3.10.2 Installing SLURM RPMS

Run the command below to install the SLURM rpms:

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

2.3.10.3 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 **srun** and then forwarded to **slurmd** to initiate job steps.

Important:
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 only by **SlurmUser**. Use the commands below to change the setting, if this is not the case.

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

The **Public.Key** file must be readable by all users. Use the commands below to change the setting, if this is not the case.

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

2.3.10.4 Create and Modify the SLURM configuration file

A **SLURM** configuration file must be created from 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.

1. Provide the name of the machine where the **SLURM** control functions will execute. This will be the Management Node.

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

2. Provide the SlurmUser and the authentication method for communications:

   ```
   SlurmUser=slurm
   ```
   AuthType=auth/munge (as shown in the example file)
   or
   ```
   AuthType=auth/none
   ```

3. Provide the type of switch or interconnect used for application communications.

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

4. Provide 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, (see step 2.5.3.3).

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) (see step 2.5.3.3).

```
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.

   ```
   #JobCompType=jobcomp/filetxt # default is jobcomp/none
   #JobCompLoc=/var/log/slurm/slurm.job.log
   #JobAcctType=jobacct/linux   # default is jobacct/none
   #JobAcctLogFile=/var/log/slurm/slurm_acct.log
   ```

   Uncomment these lines if job accounting is to be undertaken.

7. Provide the paths to the job credential keys. The keys must be copied to all of the nodes (see step 2.5.3.3).

   ```
   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 manually change the state of the node 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.

slurm.conf file example

```
ControlMachine=bali0
ControlAddr=bali0
SlurmUser=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
```

```
#JobAcctType=jobacct/linux   # default is jobacct/none
#JobAcctLogFile=/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
```

2.3.10.5 More Information

See the Bull HPC BAS5 for Xeon *Administrator's Guide* for more information on SLURM (Security, the creation of job credential keys, the **slurm.conf** file, and stopping and starting daemons).

2.3.10.6 Installing and Configuring Munge for SLURM Authentication

This software is required if the authentication method for the communication between SLURM components is munge (where AuthType=auth/munge). On most platforms, the munged daemon does not require root privileges. If possible, the daemon must be run as a non-privileged user. This can be controlled by the init script as detailed in the *Starting the Daemon* section below.

See:
For additional information about munge software, refer to http://home.gna.org/munge/.

By default, the munged daemon uses the following system directories:

- `/etc/munge/`
  This directory contains the daemon's secret key. The recommended permissions for it are 0700.
- `/var/lib/munge/`
  This directory contains the daemon's PRNG seed file. It is also where the daemon creates pipes for authenticating clients via file-descriptor-passing. If the file-descriptor-passing authentication method is being used, this directory must allow execute permissions for all; however, it must not expose read permissions. The recommended permissions for it are 0711.
- `/var/log/munge/`

This directory contains the daemon's log file. The recommended permissions for it are 0700.

- `/var/run/munge/`
  This directory contains the Unix domain socket for clients to communicate with the daemon. It also contains the daemon's pid file. This directory must allow execute permissions for all. The recommended permissions for it are 0755.

These directories must be owned by the user that the munged daemon will run as. They cannot allow write permissions for group or other (unless the sticky-bit is set). In addition, all of their parent directories in the path up to the root directory must be owned by either root or the user that the munged daemon will run as. None of them can allow write permissions for group or other (unless the sticky-bit is set).

2.3.10.7 Creating a Secret Key

A security realm encompasses a group of hosts having common users and groups. It is defined by a shared cryptographic key. Credentials are valid only within a security realm. All munged daemons within a security realm must possess the same secret key.

By default, the secret key resides in **/etc/munge/munge.key**. This location can be overridden using the munged command-line, or via the init script as detailed in the *Starting the Daemon* section below.

A secret key can be created using a variety of methods:

- Use random data from **/dev/random** or **/dev/urandom**:

```
$ dd if=/dev/random bs=1 count=1024 >/etc/munge/munge.key
```

or

```
$ dd if=/dev/urandom bs=1 count=1024 >/etc/munge/munge.key
```

- Enter the hash of a password:

```
$ echo -n "foo" | sha1sum | cut -d' ' -f1 >/etc/munge/munge.key
```

- Enter a password directly (not recommended):

```
$ echo "foo" >/etc/munge/munge.key
```

This file must be given 0400 permissions and owned by the user that the munged daemon will run as.

2.3.10.8 Starting the Daemon

Start the daemon by using the init script (`/etc/init.d/munge start`). The init script sources `/etc/sysconfig/munge`, if present, to set the variables recognized by the script.

The OPTIONS variable passes additional command-line options to the daemon; for example, this can be used to override the location of the secret key (`--key-file`) or set the number of worker threads (`--num-threads`). If the init script is invoked by root, the USER variable causes the daemon to execute under the specified username; the 'daemon' user is used by default.

2.3.10.9 Testing the Installation

Perform the following steps to verify that the software has been properly installed and configured:

1. Generate a credential on stdout:

```
$ munge -n
```

2. Check if a credential can be decoded locally:

```
$ munge -n | unmunge
```

3. Run a quick benchmark:

```
$ remunge
```

If problems are encountered, verify that the munged daemon is running (`/etc/init.d/munge status`). Also, check the log file (`/var/log/munge/munged.log`) or try running the daemon in the foreground (`/usr/sbin/munged --foreground`).

Some error conditions can be overridden by forcing the daemon (`/usr/sbin/munged --force`).

2.3.11 Installing and Configuring the PBS Professional Batch Manager - optional

Important:
PBS Professional does not work with **SLURM**. **SLURM** must not have been installed.

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

Important:
The **FlexIm** License Server has to be installed before **PBS Professional** is installed.

2.3.11.1 Downloading, Installing and Starting the FLEXlm License Server

1. Create a directory where the license server will reside, normally this is on the Management Node. This location is known as **<install loc>**. The command to create this is similar to that below:

```
mkdir /root/PBS
```

2. Copy all tarballs and the documentation into the **<install loc>** directory.

3. Uncompress and extract the files, using the commands below:

```
cd /root/PBS
tar -xvzf altair_flexlm.amd64_s8.tar.gz
```

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

5. Install the **install_altairlm.sh** startup script:

```
cd /root/PBS/altair/scripts
./install_altairlm.sh
```

6. Modify the startup script by editing the **/etc/init.d/altairlmgnd** file. Comment out the line which starts **su $ DAEMON**.

7. Start the license server using the command below:

```
/etc/init.d/altairlmgrd start
```

2.3.11.2 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 both the Management Node and on the Compute Reference Node, using a command similar to that below.

```
cd /root/PBS
tar -xvzf PBSPro_9.0.0linux24_x86_64_AS3.tar.gz
```

2. Go to the installation directory on each node and run:

```
cd PBSPro_9.0.0
```

3. Start the installation process:

```
./INSTALL
```

2.3.11.3 PBS Professional Installation Routine

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

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 **/root/PBS/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**'.

2.3.11.4 Initial configuration on the Management Node

See Chapter 4 in the PBS Professional *Administrator's Guide* for more information on configuring and starting **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=cluster0
PBS_SCP=/usr/bin/scp
```

2. Run the **vi /etc/init.d/pbs** command and add the following line in the **start_pbs()** function

```
ulimit -l unlimited
```

2.3.11.5 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
```

2.3.12 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.

2.3.13 Configuring the 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.

Three techniques can be used in order to make **MPIBull2** available to all users. The administrator should choose the one they prefer:

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:

4. 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**.

5. **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.

2.4 STEP 4: Installing RHEL5.1, BAS5v1.1 for Xeon Software, and optional HPC software products

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.

Important:
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 correctly followed then these prerequisites will already be in place.

2.4.1 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.

2.4.2 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
```

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
         - 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. All the **Linux** installation steps will be pre-filled if you chose to use **interactive** mode, and will have to be confirmed, or otherwise, for each step

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

5. The path that includes the **BAS5v1.1 for Xeon** software installer. This will be something like **/release/XBAS5V1.1**. 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/XBAS5V1.1

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

6. 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:
-----------------------------------------------------------------
```

7. 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/XBAS5V1.1 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 :
-----------------------------------------------------------------
```

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

9. 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 Bull **BAS5 for Xeon** optional HPC products can be installed later manually (see Appendix C)

10. 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/XBAS5V1.1
        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.

11. 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 :
-------------------------------------------------------------------
```

2.4.3 Launching the NFS Installation of the BAS5v1.1 for Xeon software

1. The Bull **BAS5v1.1 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 C - *Manually Installing Bull BAS5v1.1 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.

2.5 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**, **InfiniBand** and **PBS Pro** as necessary for the Reference Nodes to be deployed. It also describes the installation of compilers on the Login Nodes and how to configure the User environment.

2.5.1 Configuring SSH

Important:
These tasks must be performed before deployment.

2.5.1.1 When re-installing

In the case of a re-installation, you can retrieve the SSH keys of the nodes and of the root user, which have been saved during STEP 0. 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\Computex and Login I/O Nodes.

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

- Restart the SSH service on each type of node:

```
service sshd restart
```

Note:
The SSH keys of the users can be restored from the files saved by the administrator (for example **/<username>/.ssh**).

Note:
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.

2.5.1.2 When installing for the first time

In the case of a first installation, you must create the SSH keys for the **root** user, first on the Management Node, then on the other nodes, as described below:

On the Management Node

1. Change to the root directory:

```
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
--------------------------------------------------------------------------------
```

On the Compute\Computex and\or combined Login I/O or dedicated Login Reference Nodes

1. Copy the **/root/.ssh** directory from the Management Node on to the Compute and combined Login I/O or dedicated Login Nodes.

```
scp -r .ssh <reference_or_login_or_secondary-management_node>:.
```

2. Test this configuration:

```
> ssh <reference_or_login_or_secondary-management_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.

2.5.2 Configuring the kdump kernel dumping 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:

   b. 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
```

   a. 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 available from

http://people.redhat.com/duffy/debuginfo/index-js.html in the **kernel-debuginfo-<kernel_release>.rpm**

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
```

2.5.2.1 Testing kdump

In order to test that **kdump** is working correctly a dump may 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
```

2.5.2.2 Configuring systems to take dumps from the Management Network

In addition to forcing a dump for a kernel crash, it is possible to force a dump using the **ipmitool** command from the Management Node. This is done as follows:

Add `nmi_watchdog=0` to the kernel boot options in the **/boot/grub/menu.lst** file in order to deactivate the **NMI** watchdog used by **RHEL,** so that the other NMIs can be put into effect. An example of the menu 1st file is shown below:

```
kernel /vmlinuz-2.6.18-53.d5.ELsmp ro root=LABEL=/ nmi_watchdog=0
console=tty0 console=ttyS1,115200n8 console=ttyS0,1152,00n8 rhgb
quiet
```

Once the system has been restarted the kernel has to be reconfigured so that a panic is launched when an unknown NMI is received. This can be set to happen automatically by configuring the **kernel.unknown_nmi_panic = 1** option in the **/etc/sysct1.conf** file

Alternatively, this can be done manually by using the command.

```
echo 1 > /proc/sys/kernel/unknown_nmi_panic
```

An **NMI** dump may be launched by using **IPMI** via the command:

```
ipmitool -H <bmc_address> -U <user_name> -P <pwd> chassis power diag
```

or by using the **nsctrl** command.

For more information see http://kbase.redhat.com/faq/FAQ_105_9036.shtm

Notes:

- If watchdog is still active after the **kernel.unknown_nmi_panic = 1** option is set the machine will no longer boot.
- For this release of **BAS5 for Xeon** the **IPMI power diag** command will launch a dump for **NovaScale R423**, **NovaScale R440** and **NovaScale R460** series machines.
- There is also a dump button on the back of the **NovaScale R460** series machines that will launch an **NMI** dump for these machines.

Further information can be found in the **kdump** man pages.

2.5.3 Configuring SLURM - optional

Important

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 Reference Node for the following node types - COMPUTE, COMPUTEX and Login.

2.5.3.1 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, 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** number does not have to match the one indicated above, but it has to be the same on all the nodes that are included in the SLURM cluster.

The **uid** number does not have to match the one indicated above, but it has to be the same on all the nodes that are included in the SLURM cluster.

The user name in the example above is "slurm", another name can be used, but it has to be the same on all the nodes that are included in the SLURM cluster.

2.5.3.2 Installing SLURM

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-auth-munge slurm-auth-none slurm-devel
```

2.5.3.3 Copying the SLURM configuration file and checking files on the other nodes

Copy the following files from the Management Node to the Reference Nodes – COMPUTE, COMPUTEX, combined Login/IO or dedicated Login.

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

Check that the directory used by the SLURM daemon (typically **/var/log/slurm**) exists on the Compute\Computex, combined Login/IO or dedicated Login Reference Nodes.

Setting appropriate 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 Reference Nodes.

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

2.5.3.4 More Information

See the Bull HPC BAS5 for Xeon *Administrator's Guide* for more information on SLURM (security, the creation of job credential keys, the **slurm.conf** file, and stopping and starting daemons).

2.5.4 Installing and Configuring Munge for SLURM Authentication

This software is required if the authentication method for the communication between SLURM components is munge (where AuthType=auth/munge). On most platforms, the munged daemon does not require root privileges. If possible, the daemon must be run as a non-privileged user. This can be controlled by the init script as detailed in the *Starting the Daemon* section below.

See:
For additional information about munge software, refer to http://home.gna.org/munge/.

By default, the munged daemon uses the following system directories:

- `/etc/munge/`
  This directory contains the daemon's secret key. The recommended permissions for it are 0700.
- `/var/lib/munge/`
  This directory contains the daemon's PRNG seed file. It is also where the daemon creates pipes for authenticating clients via file-descriptor-passing. If the file-descriptor-passing authentication method is being used, this directory must allow execute permissions for all; however, it must not expose read permissions. The recommended permissions for it are 0711.
- `/var/log/munge/`
  This directory contains the daemon's log file. The recommended permissions for it are 0700.
- `/var/run/munge/`
  This directory contains the Unix domain socket for clients to communicate with the daemon. It also contains the daemon's pid file. This directory must allow execute permissions for all. The recommended permissions for it are 0755.

These directories must be owned by the user that the munged daemon will run as. They cannot allow write permissions for group or other (unless the sticky-bit is set). In addition, all of their parent directories in the path up to the root directory must be owned by either root or the user that the munged daemon will run as. None of them can allow write permissions for group or other (unless the sticky-bit is set).

2.5.4.1 Deploying the Secret Key on other Nodes

```
$ echo "foo" >/etc/munge/munge.key
```

Securely propagate this file (e.g. via ssh) to all other hosts within the same security realm.

2.5.4.2 Starting the Daemon

On each host within the security realm use the init script (`/etc/init.d/munge start`). The init script sources `/etc/sysconfig/munge`, if present, to set the variables recognized by the script.

The `OPTIONS` variable passes additional command-line options to the daemon; for example, this can be used to override the location of the secret key (`--key-file`) or set the number of worker threads (`--num-threads`). If the init script is invoked by root, the USER variable causes the daemon to execute under the specified username; the "daemon" user is used by default.

2.5.4.3 Testing the Installation

Perform the following steps from the Management Node and from each Reference Node to verify that the software has been properly installed and configured:

1. Generate a credential on stdout:

```
$ munge -n
```

2. Check if a credential can be decoded locally:

```
$ munge -n | unmunge
```

3. Check if a credential can be remotely decoded:

```
$ munge -n | ssh somehost unmunge
```

4. Run a quick benchmark:

```
$ remunge
```

If problems are encountered, verify that the munged daemon is running (`/etc/init.d/munge status`). Also, check the logfile (`/var/log/munge/munged.log`) or try running the daemon in the foreground (`/usr/sbin/munged --foreground`).

Some error conditions can be overridden by forcing the daemon (`/usr/sbin/munged --force`).

2.5.5 Installing and Configuring the PBS Professional Batch Manager - optional

Important
PBS Professional does not work with **SLURM**.

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

Important
The **FlexIm** License Server has to be installed on the Management Node before **PBS Professional** is installed – see section *2.3.11.1*

2.5.5.1 Starting the installation of PBS Professional

The commands for the installation have to be carried 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 both the Management Node and on the Compute\Computex Reference Node, using a command similar to that below.

```
cd /root/PBS
tar -xvzf PBSPro_9.0.0linux24_x86_64_AS3.tar.gz
```

2. Go to the installation directory on each node and run:

```
cd PBSPro_9.0.0
```

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:

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**

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

2.5.5.2 Initial configuration on a Compute, Computex or Login Reference Node

See Chapter 4 in the PBS Professional *Administrator's Guide* for more information on configuring and starting **PBS Professional.**

2.5.5.3 Initial configuration on the Compute and Computex Reference Node

1. Modify the **/etc/pbs.conf** file 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. Run the **vi /etc/init.d/pbs** command and add the following line in the **start_pbs()** function

```
ulimit -l unlimited
```

2.5.5.4 Initial configuration on the Login Reference Node

Modify the **/etc/pbs.conf** file 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=<basename>0
PBS_SCP=/usr/bin/scp
```

2.5.6 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.

2.5.7 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.

2.5.8 Configuring the 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/mpi**Bull**2/mpi**Bull**2.*sh**, which are used to define default settings for the cluster.

Three techniques can be used in order to make **MPIBull2** available to all users. The administrator should choose the one they prefer:

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_mpi**Bull**2.*** 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:

1. 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**.

2. **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.

2.6 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\Computex Node and Login or I/O or Login/IO Reference Node previously installed
3. Deployment of these images on cluster nodes.
4. Post Deployment Tests

These operations have to be performed **from the Management Node**.

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.

Important:
It is only possible to deploy an image to nodes that are equivalent and have the same hardware architecture:

- Platform, (for example NovaScale R421)
- Disks (same number, controller, size)
- Network interface.

See:
Refer to the *HPC BAS5 for Xeon Administrator's Guide* for more information about **Ksis**.

2.6.1 Installing, Configuring and Verifying the Image Server

2.6.1.1 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.

2.6.1.2 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 finished before running **Ksis**.

2.6.1.3 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.

Note:

Do not change the BIOS boot configuration for the Reference Nodes, because the image must NOT be deployed to these nodes before the deployment of the image to the other nodes has completed successfully.

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
```

2.6.2 Creating an Image

Create a reference image of the Compute\Computex Node and the Login or I/O or Login/IO Reference Node (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.

2.6.3 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] -P
```

   The use of the **–P** option is mandatory as it enables **Ganglia**, **Syslog-ng**, **NTP**, **SNMP** and **Pdsh** to be configured automatically on the machines. This option also allows the **IP over InfiniBand** interfaces to be configured according to the information in the Cluster database.

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] –P
```

2.6.4 Post Deployment Tests

2.6.4.1 Testing NTP

1. Test installation: run the following command on a Compute or a Computex node and on a combined I/O Login or dedicated Login nodes:

```
ntpq -p
```

   Check that the output returns the name of the NTP server, and that values are set for **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 if 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
```

2.6.4.2 Checking and Starting the SLURM Daemons on Compute and Login/IO Reference 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 or Computex 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.

2.6.4.3 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).

Chapter 3. 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:

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.

For more information about setting up the storage management services, refer to the *Storage Devices Management* chapter in the Bull *HPC 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.

3.1 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
```

3. 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.

3.2 Enabling FDA Storage System Management

Important:
This section only applies when installing for the first time.

Note:

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.

3.2.1 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
```

Important

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.

4. Connect to the Linux system FDA manager.

```
ssh <administrator>@< LinuxFDAhost >
```

5. 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.

3.2.2 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
```

3.2.3 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 on Windows. 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 Disk Array Unit User's Guide* 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
```

3.3 Enabling DataDirect Networks (DDN) S2A Storage Systems Management

3.3.1 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.

3.3.2 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, edit the file found in the **/etc/logrotate.d/ddn** directory. See the **logrotate** man page for more details.

3.3.3 Enabling Management Access for Each DDN

1. List the storage systems as defined in the cluster management database:

```
storstat -a
```

   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.

3.3.4 Initializing the DDN Storage System

Initialize each DDN storage system either from the cluster Management Node or from a laptop, as described below.

3.3.4.1 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.

Note:
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).

Note:
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.

3.3.4.2 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
```

3.4 Enabling the Administration of an Optima 1250 Storage System

Important

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.

3.4.1 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.

3.4.2 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
```

3.5 Enabling the Administration of an EMC/Clariion (DGC) CX3-40f storage system

Note:

The information in this section is also valid for the **EMC/Clariion (DGC) CX300** storage system.

3.5.1 Initial Configuration

See the *CLARiiON CX3-40f Setup Guide* for more details on configuring the CX3-40f storage system. 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 **Clariion CX3-40f** 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.

3.5.2 Complementary Configuration Tasks

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.

3.5.3 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 CX3-40f** 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
```

3. In order to use the http interface for the **EMC Navisphere Management Suite** the following 2 RPMs found in the **BONUS** directory on the Bull **XHPC DVD** must be installed:
   `XHPC/BONUS/jre-<version>-linux-i586.rpm`
   `XHPC/BONUS/firefox-<version>-Bull.0.i386.rpm`

These are installed by running the commands below:

```
cd /release/XBAS5V1.1/XHPC/BONUS
rpm -i jre-<version>-linux-i586.rpm firefox-<version>-Bull.0.i386.rpm
```

3.6 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.

3.7 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 WWPNs 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
```

3.8 Enabling Brocade Fibre Channel Switches

3.8.1 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 7* for more information about the switch configuration. You can also refer to Brocade's documentation.

3.8.2 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 4. Configuring the Lustre File System

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.

4.1 Setting up NIS to share user accounts

Important

For those clusters which include dedicated I/O + LOGIN nodes there is no need to use **NIS** on the Management Node.

4.1.1 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
```

4.1.2 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’
```

4.2 Configuring NFS v3 to share the /home_nfs and /release directories

4.2.1 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.

Note:

The following instructions only apply if dedicated disks or storage arrays are being used for the **NFS** file system.

Note:

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

```
mount LABEL=<label for home_nfs dedicated block device> /home_nfs
mount LABEL=<label for release dedicated block device> /release
```

if labels have been applied to the file systems.

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>
```

4.2.2 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
```

4.3 Configuring the Lustre File System

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 the **Lustre** file systems
- Register detailed information about each Lustre File System component in the ClusterDB
- If necessary, configure the High Availability mechanism.

Important
These tasks must be performed after 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.

If there are problems setting up the Lustre File System, and for more information about Lustre commands, refer to the HPC BAS5 for Xeon *Administrator's Guide*. This document also contains additional information about High Availability for I/O nodes and the ClusterDB.

4.3.1 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 High-Availability capacity is enabled.

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 filesystems are started for the first time. If the High Availability feature is available, the following command will display the configuration of the HA paired nodes:

```
lustre_migrate nodestat
```

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
```

4.3.2 Configuration of Storage Systems in the Cluster

This phase configures the disk arrays connected to nodes other than the Management Node.

Important:

Skip this phase if a software migration is being undertaken where the Lustre configuration and data must be preserved.

4.3.3 Configure the Storage Systems Using the Storage Configuration Deployment Service

The Storage Configuration Deployment service consists of:

- Using a model file which includes all the information needed to configure a storage system.
- Applying the **stormodelctl** command to deploy the configuration details specified in the model file on a range of storage systems.

The steps required to logically configure the storage systems are described fully in the **HPC BAS5 for Xeon** *Administrator's Guide*. A summary of the configuration process is provided below.

Note:

This phase requires the definition of the logical configuration of the storage systems used by the **Lustre** file system. Since it requires some thought, it may be postponed until there is a need to configure Lustre file system.

Initial conditions

If **WWN-mode LUN** access control is used in the model file, the ClusterDB will need to be updated with the HBA WWN information. This is done by using the command below:

```
ioregister –a
```

The **ioregister –a** command scans each node, to produce a list of adapters. This information is then stored in the Cluster Database.

Note:

The collection of I/O information may fail for some nodes which are not yet operational in the cluster. Check that it succeeded, at least for the nodes referenced by the Mapping directives of the model file (i.e. the nodes in the I/O cell of the storage system that are linked to it with an I/O path).

Configuration process

1. Copy the config model for the storage system into **/etc/storageadmin** on the Management Node. Run the command:

```
cd /etc/storageadmin
```

2. Apply a model to the storage systems (formatting the disks):

```
stormodelctl -c applymodel -m <model_name>
```

model_name is the name of the file containing the storage configuration rules.

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 -c applymodel -m <model_name> -i <disk_array_name> -v
```

3. Check the status of format operations on the storage systems. When the **applymodel** command is completed, the disk array proceeds to format operations using the model that has been applied. This operation can take a long time. The progress of the format should be checked periodically with the following command:

```
stormodelctl -c checkformat -m <model_name>
```

Warning:

Ensure that all formatting operations are completed on all storage systems before doing anything else on these systems.

4. For DDN storage systems, the following command displays the LUN formatting status:

```
ddn_admin -i < singlet IP-name or IP-address> -c getinfo -o logical
```

Wait for the 'ready' status for each LUN.

5. Once the storage systems are fully configured, reboot the all nodes that are connected to them in so that the new storage systems can be detected.

Warning:

It is essential that ALL the nodes connected to the storage system are rebooted at this point.

Notes:

- The message 'no formatting operation', which may appear following the command above, indicates that the formatting has finished and is OK.
- The **stormodelactl** and **ddn_admin** formatting status checking commands listed above may be run in parallel.

4.3.4 Configuring Storage Systems without Using the Storage Configuration Deployment Service

Please refer to the documentation provided with the storage system to understand how to use the management tools: all the RAID LUNs must be created and formatted, and the operational parameters tuned, using the native tool.

Most of the configuration operations can be performed from the Management Node, using the **CLI**. Please refer to the HPC BAS5 for Xeon *Administrator's Guide* for more information.

4.3.5 Making the Storage Systems Operational for Lustre

Depending on your configuration go to:

- 4.3.5.1 *Making the Storage Systems Operational for Lustre Using the Storage Configuration Deployment Service*

or:

- 4.3.5.2 *Making a Storage System Operational for Lustre without Using the Storage Configuration Deployment Service*

4.3.5.1 Making the Storage Systems Operational for Lustre Using the Storage Configuration Deployment Service

Note:

This phase requires that all the storage systems are configured and their LUNs formatted. It may be postponed until there is a need to configure Lustre file system.

1. Check that each I/O node is connected to the correct storage system. For DDN storage systems check the connection of each one with the following command (use **storstat –a** to get the list of DDN names):

```
ddn_conchk -I <ddn_name> -f
```

Note:

This command can only be used if **ConMan** is available for the **DDN** storage systems.

2. I/O nodes post configuration.

   Prerequisite: **ssh** must have been configured "password-less".

   This operation transmits configuration information to each node attached to a storage system defined in the specified model. The node then forces a verification of the storage resources and checks them against the LUNs defined in the model file.

Important:

Do not run **stordepmap** if Lustre is running. Read the Bull HPC BAS5 for Xeon *Administrator's Guide* carefully to fully understand the risks and prerequisites, before running this command. In particular, see section 4.6 – Installing and Managing Lustre File Systems - in the *Administrator's Guide* for details of the **lustre_util** tool which is used to stop **Lustre**.

```
stordepmap -m <model_name>
```

model_name is the name of the file containing the storage configuration rules.

Warning:
This command is silent and long. Be sure to wait until the end.

Note:
stordepmap should not display any errors.

Important:
When performing a software migration, do not run the two following **stormodelctl** commands, if the Lustre configuration and data have to be preserved.

3. OST Lustre configuration:

```
stormodelctl -c generateost -m <model_name>
```

model_name is the name of the file containing the storage configuration rules.

4. MDT Lustre Configuration:

```
stormodelctl -c generatemdt -m <model_name>
```

model_name is the name of the file containing the storage configuration rules.

6. Run the **lustre_investigate check** command to make the OST and MDT available for the filesystem:

```
lustre_investigate check
```

4.3.5.2 Making a Storage System Operational for Lustre without Using the Storage Configuration Deployment Service

When storage systems attached to a node are configured, the node can be rebooted to discover the new storage resource. This is the easiest way to ensure LUN discovery.

Lustre functions on the basis that persistent naming of storage devices exists with the same device name for a LUN on each node of an HA pair. Thus, the following operations are mandatory for Lustre.

To ensure persistency of Storage Systems devices naming on the node system, the following commands must be performed, depending on the configuration.

Important:

1. Do not run **stordiskname** or **stormap** if Lustre is running. Read the Bull HPC BAS5 for Xeon *Administrator's Guide* carefully before running these commands, to fully understand the risks and prerequisites.

2. If I/O multipathing has been configured, ensure that all paths to all devices are in an 'alive' state (using the **lsiocfg –x** command), if not then **stordiksname** will exit in error.

Note:

It is highly recommended to use the **stordiskname** command with the **–r** option (remote) from the Management Node, in order to take advantage of its automatic backup/restore functionalities.

If the node is NOT in a High Availability pair:

- From the Management Node run:

```
stordiskname –c –r <node_name>
```

and after this finishes:

```
ssh root@<node_name> "stormap –c"
```

where <node_name> is the IP name of the target node.

- Or locally on the I/O node:

```
stordiskname –c
```

and after this finishes:

```
stormap –c
```

If the node is in a High Availability pair:

- From the Management Node run:

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

and when this has finished, run:

```
ssh root@<node1_name> "stormap –c"
ssh root@<node2_name> "stormap –c"
```

where **<node1_name>** is the name of one node in the HA pair, and **<node2_name>** is the name of the other node in the HA pair.

OR:

```
pdsh -w <node1_name>,<node2_name> "stormap -c" | dshbak -c
```

- Or locally on any node of the HA pair

Prerequisite: **ssh** must have been configured 'password-less'. This means that the RSA keys must be installed on all the nodes.

```
stordiskname –c –n <peer_node_name>
```

and after this has finished, run:

```
stormap –c
ssh root@<peer_node_name> "stormap –c"
```

where <peer_node_name> is the name of the adjacent node in the HA pair.

Note:

For some storage subsystems apart from FDA and DDN, the **stordiskname** command might return such an error:

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

In this case, refer to the procedures for the storage subsystem described in the Bull BAS5 for Xeon *Administration Guide*

Important:

The **stordiskname** command builds a **/etc/storageadmin/disknaming.conf** file which contains information, including symbolic link names, the LUN UIDs and the LUN's WWPN access. Only the **stordiskname** command can create or modify this file to include information specific to each node.

Note that if one or more LUNs on a storage system have been configured as quorum disks for Cluster Suite that they will also have been linked and therefore it is most important NOT to use these LUNs for other purposes, apart from that of being quorum disks. Use **mkqdisk –L** and **stormap –l** to check that this is so.

This **disknaming.conf** file will be erased when redeploying the **ksis** reference image, or when the system is restored for a node.

Therefore, **stordiskname**, if used with the **–r** option (remote) from the Management Node, will enable backups and restorations of the **/etc/storageadmin/disknaming.conf** file to be managed automatically. It is highly recommended that this is done. If this option is not used, the administrator has to manage the backup of the **/etc/storageadmin/disknaming.conf** file himself.

When used remotely (**–r**) immediately after a **ksis** image re-deployment, or a node system restoration, the following commands must be used in order that the **LUN**s are addressed by the same symbolic link names used previously to avoid the need to reconfigure **Lustre**.

The **stordiskname** command should be executed from the Management Node using the **–u** (update) option 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:

If the **–m** mode option was specified when the **stordiskname** was previously executed, then this should be included as well. This applies to both High Availability and non High Availability nodes.

The symbolic links must be recreated using the information contained within the **disknaming.conf** file, once it has been copied over. Therefore, run the **stormap** command as described previously.

Note

If a node has been rebooted after the **disknaming.conf** file was copied over, the symbolic links will have been created automatically at boot time, therefore there is no need to run **stormap** again.

4.3.6 Adding Information into the /etc/lustre/storage.conf File

This phase should be done in the following situations:

- If there is a need to use **Lustre** filesystems and no cluster database is available.
- 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.

Important:

Skip this phase if a software migration has been carried out, as the **/etc/lustre** directory will have been saved and restored.

Please refer to the HPC BAS5 for Xeon *Administrator's Guide* for more details about the **storage.conf** file.

4.3.7 Configuring and Starting Cluster Suite on I/O Nodes

Important

This section only applies to clusters where the High Availability feature is to be implemented.

4.3.7.1 Cluster with a Management Node

The configuration files for the Cluster Suite (**/etc/cluster/cluster.conf**) are automatically generated using information preloaded in the ClusterDB. They are also distributed on all the I/O nodes for the cluster.

Note:

The quorum disk must have been defined in the model file in order that a LUN is dedicated to it. If a model file is not used and **stordiskname** has been used as an alternative, then a LUN must be configured as a quorum disk (see **mkqdisk** help for more information)

If quorum disk is not included in the Cluster Suite configuration then **stordehpa** can be used as follows:

```
stordepha -a -c configure
```

If quorum disk is included in the Cluster Suite configuration, **stordepha** must be used with the **–q** option, as below:

```
stordepha -a -c configure -q
```

Note:

The **–a** option indicates that the command applies to all I/O nodes set as HA pairs in the ClusterDB. You can use other options (**-e, -i**) to specify some I/O nodes. See the help command for details.

These configuration files do not depend on the I/O configuration nor on the Lustre file system configuration.

The following steps apply to all kinds of I/O nodes,

Once the I/O nodes have been configured, Cluster Suite can be started on the nodes as follows:

```
stordepha -a -c start
```

If a node is re-installed by KSIS, it is mandatory to carry out the actions below.

Check the Cluster Suite services (**ccsd**, **fenced**, **rgmanager**, **cman**) have started by using the following command:

```
stordepha -a -c status
```

The output should be similar to that below:

```
Status for ccsd
================================
ccsd (pid 7004) is running...
```

```
Status for fenced
==================================
fenced (pid 10209) is running...

Status for rgmanager
==================================
clurgmgrd (pid 12776) is running...

Status for cman
==================================
cman is running...
```

If the Cluster Suite has not started on a node, the output will be similar to that below:

```
Status for ccsd
==================================
ccsd is stopped

Status for fenced
==================================
fenced dead but pid file exists

Status for rgmanager
==================================
clurgmgrd is stopped

Status for cman
==================================
cman is stopped
```

4.3.8 Configuring the Lustre File System

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.
   By default **Lustre** runs on all network layers that may be active in the kernel, for example **InfiniBand** or **Ethernet**. If you do not want **Lustre** to run on certain network layers, these network layers must be deactivated for the nodes in question.

If **Ethernet** is used as the **Lustre** network layer, it is possible to select the link on which Lustre will run. This is done by editing the **/etc/lustre_modprobe.conf** file. For details see the *Lustre Operations Manual* from CFS (Section *Multihomed Servers*, sub-section *modprobe.conf*) at http://manual.lustre.org/

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).
   d. If you want to use failover filesystems, set **LUSTRE_LDAP_URL** according to the name of the Management Node (ldap://<mgmt node>/).

      Set the **LUSTRE_DB_DAEMON_HOST** and **LUSTRE_DB_DAEMON_PORT** lines, as follows, so that **LDAP** is also replicated in the **clusterDB** :

      LUSTRE_DB_DAEMON_HOST=hostname
      LUSTRE_DB_DAEMON_HOST2=hostname
      LUSTRE_DB_PORT=tcp port

      LUSTRE_DB_DAEMON_HOST2 is to be used when the Management Node does not support the High Availability feature. An alternative **LUSTRE_DB_DAEMON** hostname will be provided as a backup.

      No default values are defined for the hostna**mes. The default value for the tcp port is 56283, e.g. 0xDBDB so this appears as** LUSTRE_DB_DAEMON_PORT=56283
   e. Save and quit the editor.
   f. Once the **lustre.cfg** file has been edited copy it to the Secondary Management node for High Availability systems.
   g. To be certain that the **LDAP** and the ClusterDB start, run the command below:

```
service ldap start
```

   h. **LDAP** has to be initialized using the commands below when installing a new cluster:

```
ldapadd -W -D cn=Manager,fs=lustre -w secret -x -H ldap://hostname/ -f
/usr/share/lustre/top.ldif
service lustredbd.sh start
```

```
--------------------------------------------------------------------
starting lustredbd:
 .....
lustredbd started [OK]

--------------------------------------------------------------------
```

   i. Check that the service has launched correctly, using the command below:

```
service lustredbd.sh status
```

```
------------------------------------------------------------------
lustredbd is running
```

j. Refer to the **lustre.cfg** man page for more details.

4. 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.

Refer to the **lustre_investigate** man page or the HPC BAS5 for Xeon *Administrator's Guide* for more details.

Checking:

```
lustre_ost_dba list
```

This command displays the list of OSTs. You must have at least one OST with **cfg_stat** set to "**available**".

```
lustre_mdt_dba list
```

This command displays the list of MDTs. You must have at least one MDT with **cfg_stat** set to '**available**'.

5. Set the Lustre configuration on 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 >
------------------------------------------------------------------
```

Note:

SNMP and **LDAP** will only be enabled if the required parameters have been set. See the HPC BAS5 for Xeon *Administrator's Guide* for more details.

6. Start Lustre failover services on I/O and metadata nodes.

This step can be skipped if the High-Availability feature is not needed. Failover Lustre services are used by the **Cluster Suite** to control the migration of Lustre OST/MDT services.

Important:

The Lustre failover services have to be started before the Lustre file systems are started. Otherwise there is a risk that the services will not be managed by Lustre failover.

The Lustre failover services can only be stopped when all the Lustre file systems have been stopped. The **lustre_migrate** command allows you to manage these services on the cluster.

– To display the status of the Lustre failover services on all I/O and Metadata Nodes:

```
lustre_migrate hastat
```

– To start the Lustre failover services on all I/O and Metadata Nodes:

```
lustre_migrate hastart
```

Note:
Refer to the **lustre_migrate** man page for more information or if there are any problems with the Lustre failover services

7. 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.

8. 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.

Note:
Refer to the **lustre_util** man page for more information.

9. Install the file system.

Important:
Do not perform this step when performing a software migration and the **Lustre** configuration and data must be preserved.

   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
```

10. 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 |
+-----------+----------+--------+--------+----------------+
```

11. 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 atthe 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 set access rights accordingly.

It is possible to check the health of the filesystem, 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 5. Installing Intel Tools and Applications

This chapter describes how to install tools or commercial software from CDs or supplier sites.

5.1 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.

5.2 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 2. 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

5.2.1 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.

5.2.2 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.

5.3 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.

5.4 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.

5.5 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

5.6 Updating Intel Compilers and BAS5 for Xeon v1.1

BAS5 for Xeon V1.1 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 6. Installing and Configuring InfiniBand Interconnects

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:

6.1 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.

6.2 Configuring the Voltaire ISR 9024 Grid Switch

6.2.1 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 6-1. Voltaire ISR 9024 Switch Terminal Emulation Configuration

6.2.2 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**

6.2.3 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.

6.2.4 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
```

6.2.5 Hostname setup

6.2.5.1 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)#
```

6.2.5.2 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
```

6.2.6 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).

6.2.6.1 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)#
```

6.2.6.2 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
----------------------------------------------------------------------
```

6.2.7 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>
```

6.2.8 Route setup

6.2.8.1 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)#
```

6.2.8.2 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.

6.2.9 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.

6.2.10 Subnet manager (SM) setup

6.2.10.1 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)#
```

6.2.11 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 6.2.3 *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
```

6.3 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)#
```

6.3.1 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.

6.3.2 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).

6.3.2.1 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.

6.3.2.2 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
-------------------------------------------------------------------------
```

6.3.3 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.

Important:
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 type of switches have 4 x 3 = 12 spines.

This will provide output similar to that below for a cluster with 12 spines:

6.3.3.1 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
```

6.3.3.2 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
```

6.4 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.

6.4.1 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)#
```

6.4.2 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

-----------------------------------------------------------------
```

6.5 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.

6.5.1 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)#
```

6.5.2 Setting up FTP

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
```

Note:

The FTP server must have been configured on the Management Node

6.6 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
```

6.6.1 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)#
```

6.6.2 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
-----------------------------------------------------------------
```

6.6.3 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
```

6.6.4 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
```

6.7 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
```

6.8 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.

6.9 More Information on Voltaire Devices

For specific instructions, refer to the manuals available on the Bull *Voltaire Switches Documentation CD* or from www.voltaire.com :

Note:
For more information on the **SLURM** Resource Manager used in conjunction with **InfiniBand** stacks and Voltaire switches see the HPC BAS5 for Xeon *Administrator's Guide* and the HPC BAS5 for Xeon *User's Guide*.

Chapter 7. Configuring Switches and Cards

This chapter describes how to configure **BAS5 for Xeon** switches and cards.

The following topics are described:

- *7.1 Configuring Ethernet Switches*
- *7.2 Configuring a Brocade Switch*
- *7.3 Configuring Voltaire Devices*
- *7.4 Installing Additional Ethernet Boards*

7.1 Configuring Ethernet Switches

The Ethernet switches are configured automatically using the ClusterDB database information and the configuration file. See section *7.1.5 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 7.1.6 Ethernet Switches Initial Configuration)

7.1.1 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.

7.1.2 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 /tfpboot 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

7.1.3 Automatic Installation of Ethernet Switches

A fully automatic installation of the Ethernet switches is carried out by running the command:

```
swtAdmin auto
```

All the steps below in the Ethernet Switch Configuration Procedure (1–6) are executed in order, with no user interaction. If the automatic installation fails at any stage, you will need to execute only the steps which remain (including the one that failed).

Another option is to perform an interactive step-by-step installation using the command below:

```
swtAdmin step-by-step --switch_number <number_of_new_switches>
```

Again, all the installation steps (1-6) are executed in order, but the user is asked to continue after each one.

7.1.4 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 is run in order to discover automatically the network topology.

This action 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 that 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

At this stage, the topology discovered 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 action 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 again and will implement their final configuration 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 at previous step.

This action 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. 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 transfer to the Management Node **/tftpboot** directory.

This action 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.
-----------------------------------------------------------------
```

7. Checking the configuration of a switch

The configuration of a switch is displayed by using the command:

```
swtConfig status --name <name_of_switch>
```

7.1.5 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 **/usr/lib/clustmngt/ethswitch-tools/bin/config** script.

7.1.6 Ethernet Switches Initial Configuration

7.1.6.1 CISCO Switches

CISCO switches must be reset to the same settings that were in place when they left the factory. This is done manually.

1. **Hardware reinitialization**

Hold down the mode button located on the left side of the front panel, while 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.

7.1.6.2 Foundry Network Switches

Foundry Network switches must be configured with the IP address: 192.168.1.200/24. In order to set this IP address for each switch, follow the procedure described below in section 7.1.7.

7.1.7 Basic Manual Configuration

Please use this method when initially configuring **Foundry Network** switches 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 is displayed.

```
Connected.
Switch>
```

7.1.7.1 Configuring a CISCO Switch

1. Set the enable mode:

```
Switch>enable
```

2. Enter the 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 by 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-confg.**

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
```

7.1.7.2 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#
```

7.2 Configuring a Brocade Switch

1. Set the Ethernet IP address for the brocade switch.

Note:

The Real Value (IP address, name of the switch) to be used may be found in the cluster database (FC_SWITCH table).

Use a portable PC to connect the serial port of the switch.

Note:

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
```

7.3 Configuring Voltaire Devices

The **Voltaire® Command Line Interface (CLI)** is used for all the commands necessary to perform all management functions including software upgrades and maintenance.

The **Voltaire Fabric Manager (VFM)** provides **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 a **Performance Manager (PM)** which may be 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 the **Voltaire® Grid Switch™ ISR 9024** devices. It can also be used to monitor and configure device parameters.

For more detailed information on configuring the devices, a description of all 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*.

7.4 Installing Additional Ethernet Boards

When installing an additional Ethernet card, the IP addresses of the Ethernet interfaces may end up 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, you should:

1. Before installing a new Ethernet board, obtain the MAC addresses of the on-board Ethernet interfaces 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.

Chapter 8. Checking and Backing-up Cluster Nodes

This chapter describes the following topics:

- 8.1 *Checking the Management Node*
- 8.2 *Checking Other Nodes*
- 8.3 *Checking the Release*
- 8.4 *Backing up the System*

8.1 Checking the Management Node

Check the following:

- The required services (**Conman**, **Nagios**, **gmond**, **gmetad**, **syslog-ng**) are activated.
- **nfs** is exported to all the nodes using the **exportfs** command.

To perform a global verification it is recommended that a shell is executed which:

1. Compiles scientific applications using **MPI**
2. Runs the application on all the nodes.

If these checks are OK it means that the compilers, the **SLURM** resource manager and the **MPI** libraries are all running correctly.

8.2 Checking Other Nodes

8.2.1 I/O status

I/O status is a **Nagios** service which monitors the I/O nodes. The results are reported to **NovaScale Master – HPC Edition**, via the **I/O status** service.

If a node returns an I/O status which is not 'OK', the administrator should connect to the node and run diagnostic tests. The problem may be a hardware issue, an incorrect configuration for the devices or for the monitoring service

Refer to the chapter on *Storage Device Management* in the HPC BAS5 for Xeon *Administrator's Guide* for more information about the I/O status service and **NovaScale Master – HPC Edition**.

8.3 Checking the Release

The **Bull-infos** and **Bull-release** commands provide information about the current release.

```
# cat /etc/bull-infos
# Don't modify this file.
# Release Created on 13 Dec 2005
Bull Linux Advanced Server release 4AS (Bull V4.0)
kernel-2.6.12-B64k.2.9
installation type : REFERENCE NODE
```

```
# cat /etc/bull-release
Bull Linux Advanced Server release 4AS (Bull V4.0)
```

8.4 Backing up the System

The Management node system should be saved after installation, once the Management Node is fully operational.

Note:
It is recommended that the system is saved whenever the cluster is modified either for software updates, or for hardware modifications (for example, distribution upgrade, new or removed users, new nodes or equipment installation, etc.) This ensures that the **ClusterDB** is up to date on the system save disk.

Two methods are available to save and restore the system:

- Cloning Method.
  This method of saving and restoring the Management Node is based on system disk cloning using the **dd** command, once the ClusterDB databases on the system disk have been saved.
- Using the **mkcdrec** tool.
  mkCDrec (make CD-ROM recovery) is an Open Source tool used to make a bootable system image (including Linux system save) which can then be used for system recovery after a problem occurs, such as a disk crash or a system intrusion.

Appendix A. Interconnect Interfaces

In normal circumstances the configuration of the **Ethernet** or **InfiniBand** Interconnect interfaces is carried out automatically by **Ksis** when the images of the Compute and Login\IO node are deployed. Therefore there will not be a nee to configure these except when there is a problem and a debugging operation is being carried out, or a manual installation is being carried out for additional nodes, or there is manual reinstallation of existing nodes. In these situations an IP (Internet Protocol) has to be configured for each interface for each node as follows.

A.1 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.

A.1.1 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.

A.1.2 Starting the InfiniBand interfaces

The following commands may be 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 as they are loaded automatically following a reboot.

Appendix B. 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:

- B.1 *How to Optimize I/O Performance*
- B.2 *Creating the list of Adapters*
- B.3 *Connections for NovaScale R4xx* Servers

B.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.

B.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 B-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 B-2. PCI-Express Table

B.3 Connections for NovaScale R4xx Servers

The following paragraphs illustrate the I/O subsystem architecture for each family of **NovaScale** Rxx servers.

B.3.1 NovaScale R421 Series – Compute Node

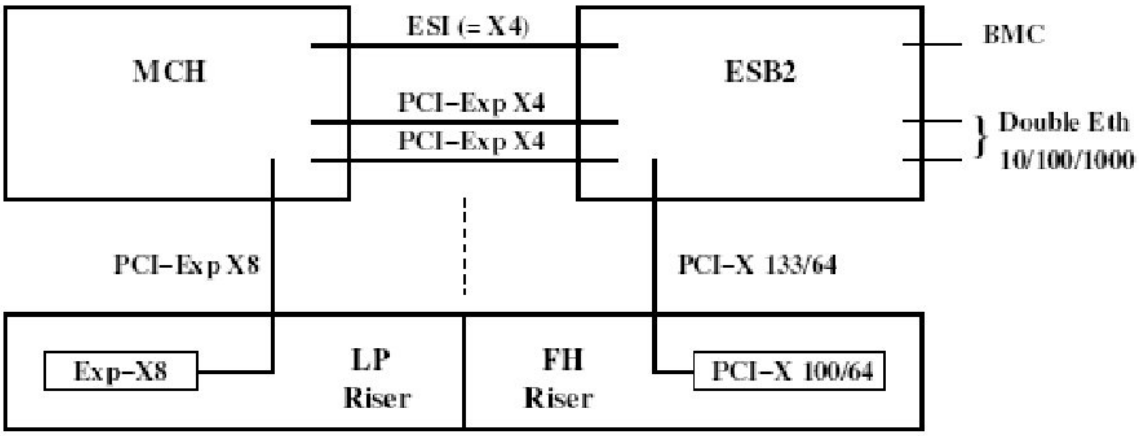

Figure B-1. NovaScale R421 rear view of Riser architecture

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**).

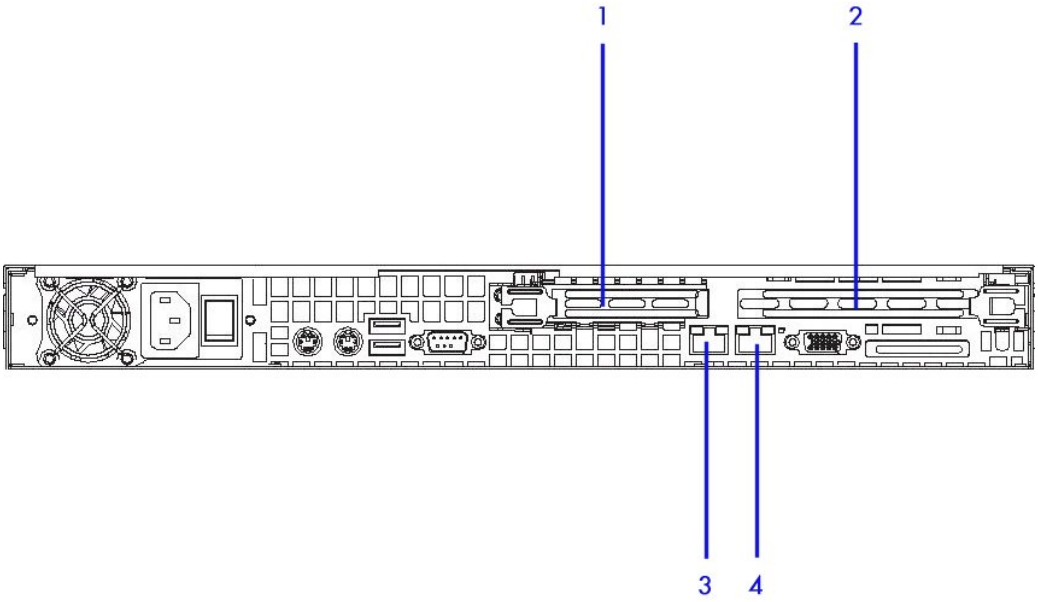

Figure B-2. NovaScale R421 rear view connectors

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 B-3. NovaScale R421 Slots and Connectors

B.3.2 NovaScale R422 Series – Compute Node

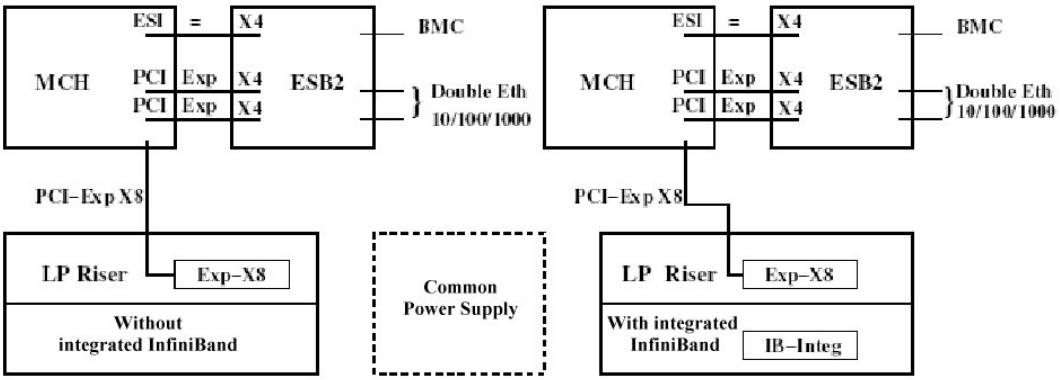

Figure B-3. NovaScale R422 rear view of Riser architecture

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.

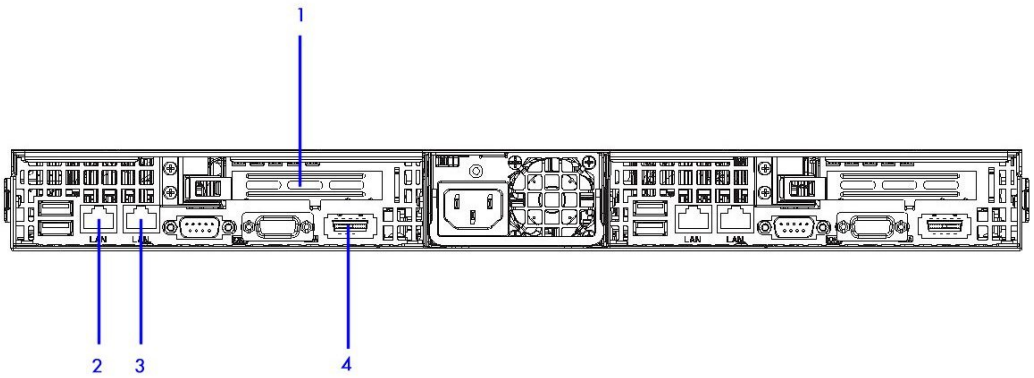

Figure B-4. NovaScale R422 Rear view connectors

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 B-4. NovaScale R422 Slots and Connectors

B.3.3 NovaScale R460 Series – Service Node

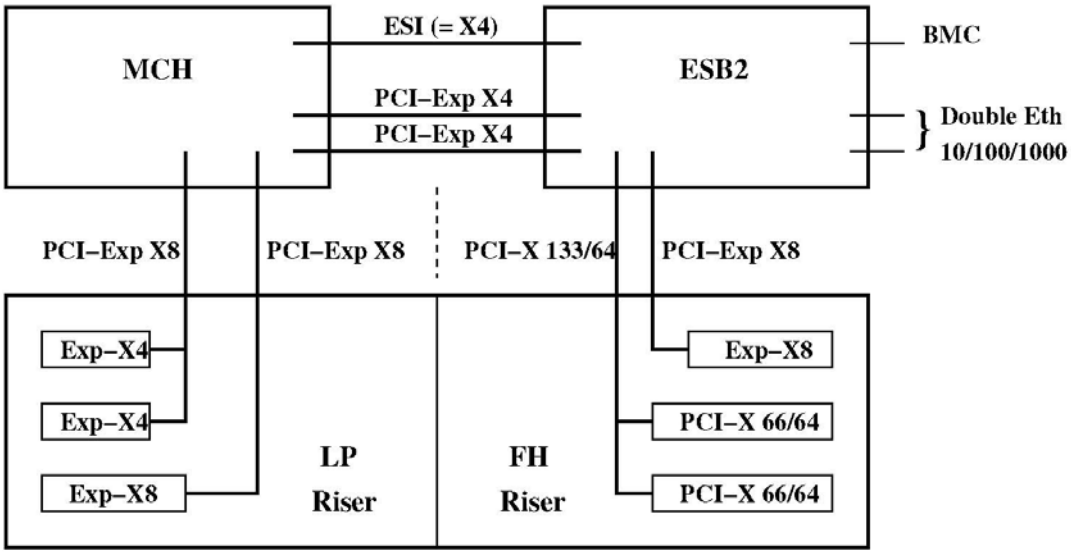

Figure B-5. NovaScale R460 risers and I/O subsystem slotting

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).

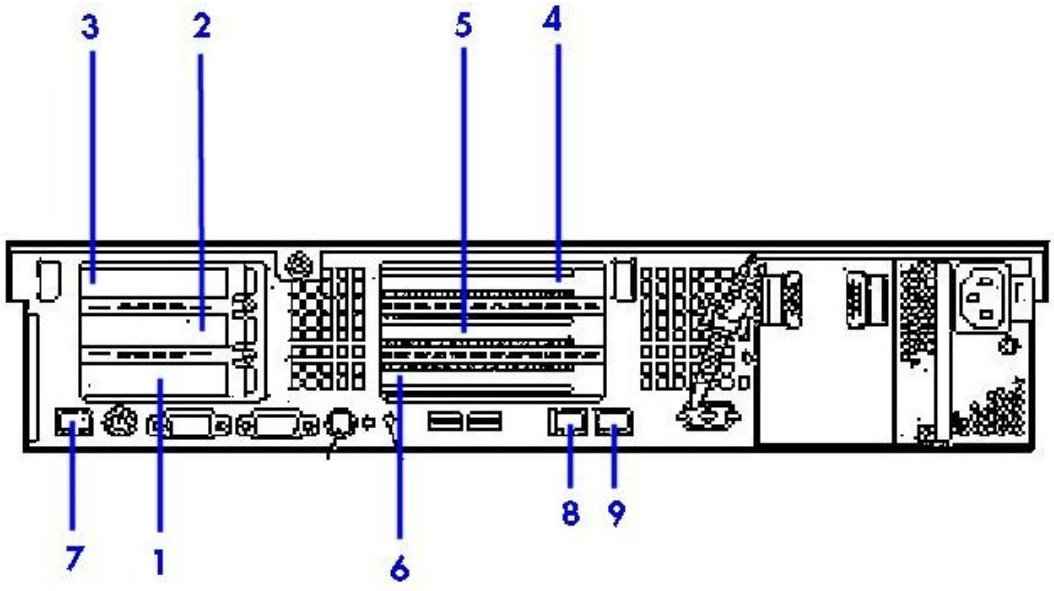

Figure B-6. Rear view of NovaScale R460 Series

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 B-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 C. 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 Chapter 11 in the *Bull BAS5 for Xeon Administrator's Guide* for details on configuring and using **HPC Toolkit**

Appendix D. 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

SDP
Socket Direct Protocol

SDPOIB
Sockets Direct Protocol over Infiniband

SDR
Sensor Data Record

SEL
System Event Log

SCSI
Small Computer System Interface

SIOH
Server Input/Output Hub

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

SM
System Management

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

SMT
Symetric Multi Threading

SNMP
Simple Network Management Protocol

SOL
Serial Over LAN

SSH
Secure Shell

T

TFTP
Trivial File Transfer Protocol

U

USB
Universal Serial Bus

UTC
Coordinated Universal Time

V

VDM
Voltaire Device Manager

VFM
Voltaire Fabric Manager

VGA
Video Graphic Adapter

VLAN
Virtual Local Area Network

VNC
Virtual Network Computing

W

WWPN
World – Wide Port Name

X

XHPC
Xeon High Performance Computing

XIB
Xeon InfiniBand

Index

G

H

I

K

L

M

N

Technical publication remarks form

Title:	BAS5 for Xeon Installation and Configuration Guide

Reference:	86 A2 87EW 00

Date:	April 2008

ERRORS IN PUBLICATION

SUGGESTIONS FOR IMPROVEMENT TO PUBLICATION

Your comments will be promptly investigated by qualified technical personnel and action will be taken as required.
If you require a written reply, please include your complete mailing address below.

NAME: _____ DATE: _____

COMPANY: _____

ADDRESS: _____

_____

Please give this technical publication remarks form to your BULL representative or mail to:

Bull - Documentation Dept.
1 Rue de Provence
BP 208
38432 ECHIROLLES CEDEX
FRANCE
info@frec.bull.fr

Technical publications ordering form

To order additional publications, please fill in a copy of this form and send it via mail to:

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

Phone: +33 (0) 2 41 73 72 66
FAX: +33 (0) 2 41 73 70 66
E-Mail: srv.Duplicopy@bull.net

Reference	Designation	Qty
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
[_ _] : The latest revision will be provided if no revision number is given.		

NAME: ______________________________ DATE: ____________

COMPANY: ______________________________

ADDRESS: ______________________________

____

PHONE: ____________________ FAX: ____________________

E-MAIL: ______________________________

For Bull Subsidiaries:
Identification: ______________________________

For Bull Affiliated Customers:
Customer Code: ______________________________

For Bull Internal Customers:
Budgetary Section: ______________________________

For Others: Please ask your Bull representative.

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

REFERENCE
86 A2 87EW 00