HPC

HPC BAS4

Installation and Configuration Guide

REFERENCE
86 A2 28ER 09

HPC

HPC BAS4

Installation and Configuration Guide

Software

July 2007

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

REFERENCE
86 A2 28ER 09

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

Preface

Scope and Objectives

This guide describes how to install the **Bull** HPC software distribution **BAS4 (Bull Advanced Server)** and all other associated software on Bull High Performance Computing clusters. It also describes all the configuration tasks necessary to make the cluster operational.

Intended Readers

This guide is for Administrators of Bull HPC systems that need either to re-install their systems, to update software with a newer version, or to install a new application.

Prerequisites

Refer to the *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 HPC system.

Chapter 2. *Installing BAS4 Software on HPC Nodes*
Details the software installation process, for all types of cluster nodes. It covers the case of a migration from BAS4V4.3 to BAS4V5.1.

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 the Lustre File System*
Describes how to configure the storage devices that the Lustre File System relies on, how to configure the Lustre file systems and, if necessary, configure the High Availability mechanism.

Chapter 5. *Installation of a Standalone Configuration*
Describes the installation of standalone configuration which consists of a single node.

Chapter 6. *Installing Tools and Applications*
Describes how to install commercial tools (Intel Compilers and MKL, TotalView debugger, Torque) and other applications (Modules).

Chapter 7. *Installing and Configuring Quadrics Interconnects*
Describes the tasks for the installation of Quadrics Interconnects.

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

Chapter 9. *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. *Installation Errors*
Helps to diagnose some installation problems.

Appendix B. *Configuring Switches*
Describes how to configure CISCO, Brocade, Quadrics and Voltaire switches.

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

Appendix D. *BAS4 Bundles*
Describes the bundles installed during the Linux installation phase.

Appendix E. *Installing TS4 and TS16 Digiboard PortServers for Linux*
Describes how to install Digiboard PortServer TS4 and TS16 for Linux.

Appendix F. *QWERTY/AZERTY Keyboard Comparison*
Provides a figure of a QWERTY keyboard, enabling you to enter EFI key sequences.

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

Bibliography

- Bull HPC BAS4 *Administrator's Guide* (86 A2 30ER)
- Bull HPC BAS4 *User's Guide* (86 A2 29ER)
- Bull HPC BAS4 *Application Tuning Guide* (86 A2 19ER)
- Bull HPC BAS4 *Maintenance Guide* (86 A2 46ER)
- A *Software Release Bulletin* (SRB) provides release-specific information and installation instructions.
- NovaScale *3xx5 Series Installation and User's Guide* (86 A1 02ET)
- NovaScale *4040 User's Guide* (86 A1 26EG)
- NovaScale *4020 User's Guide* (86 A1 72EG)
- NovaScale *5xx5 & 6xx5 User's Guide* (86 A1 41EM)
- NovaScale *5000 and 6000 User's Guide* (86 A1 94EM)
- 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)

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 HPC system.

The following topics are described:

- 1.1 *Introduction*
- 1.2 *Hardware Configuration*
- 1.3 *Administration Network and Backbone*
- 1.4 *NovaScale Administration Networks*
- 1.5 *Main Console and Hardware Management*
- 1.6 *High Speed Interconnection*
- 1.7 *Typical Types of Nodes*
- 1.8 *Installing Software and Configuring Nodes*

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 cluster file systems.

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

1.2 Hardware Configuration

Bull High Performance Computing systems feature different NovaScale Series machines for the nodes.

The 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 and application development I/O activity.

Note:

The System Administrators must be fully aware of 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 typical cluster infrastructure consists of **compute nodes** for intensive calculation and **service nodes** for management, storage and software development services.

- **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.
- **Service node(s)** cover the following functionalities:
  - **Management node(s)** administrate and run the cluster machines.
  - **Login node(s)** provide access to the cluster and a specific software development environment.
  - **I/O (Input/Output) node(s)** transfer data to and from storage units.
  - **Other node(s)** may support services or act as servers for either parallel or distributed file-systems, for example MetaData Nodes.

See 1.7 *Typical Types of Nodes* for more details.

Different networks, dedicated to particular functions, may be used, including:

- **High speed interconnects,** consisting of switches and cable/boards to transfer data between Compute Nodes and I/O Nodes.
- **Administration Networks** that include Ethernet and serial networks, and which are used for cluster management and maintenance.
- A **Backbone** to link the HPC system and the external world.

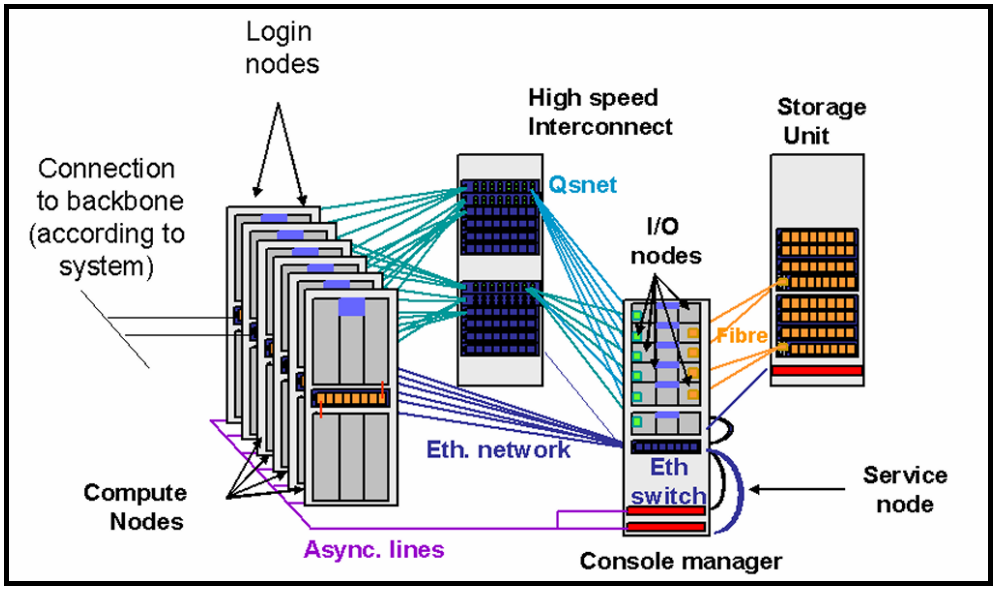

Figure 1-1. A typical HPC architecture

1.3 Administration Network and Backbone

The systems administration uses different networks (Ethernet or serial) according to:

- The status of the system (down, running or frozen).
- The nature of the request (hardware or software) that the system wants to send.

The **Administration network** includes **two** separate **Ethernet networks** (one for general software management of the cluster, the other for hardware management) and **one Serial network**. These networks concentrate on the Management node or by using the Platform Administration and Maintenance (**PAM**) software all the information necessary to control and manage the cluster.

Note:

An optional Ethernet link is necessary to connect cluster's login node(s) to a LAN backbone that is external to the cluster.

1.3.1 Serial Network

It is impossible in a high density rack of computers to have a graphical console with a keyboard and mouse for each node. Instead a serial ports concentrator for the NovaScale 5xx5 Series platform, or a Serial over LAN (SOL) connection using a maintenance coprocessor for the NovaScale 3xx5 series platform, has been used via an Ethernet line.

The **Serial network** is used for hardware and software services:

- It offers console management support for all equipment (nodes, disk units, switches, etc) from the Management node, and is used when Linux is no longer running.
- It provides the facility to obtain dumps from the Ethernet administration network when the system is frozen. It is also used to access firmware or to debug the system.

For NovaScale 40x0/5xxx/6xx0 series all the cluster equipment supporting a terminal server is connected to the serial line network. Each serial line (asynchronous RS232c) is connected from COM1 to the serial multiport concentrator (**PortServer** from **Digiboard**). The PortServer itself is connected to the administration network via Ethernet. The installation and configuration of this PortServer is fully described in Appendix E.

NovaScale 3xx5 Series platform use Serial Over LAN (**SOL**) functionalities and the **IPMI2** protocol on maintenance coprocessors (**BMC**, Baseboard Management Controller). These configurations do not need serial networks and **nport** servers but use the Ethernet link. Cluster Management is done using Server Manager and **IPMI_tools** via the BMC.

1.3.2 Administration Network

The **Administration network** is an Ethernet network which allows the management of operating systems, middleware and applications from the Management node.

This network joins all the eth0 native ports of each node through a 100/1000 Mb/s network. It is also connected to the **PAP** (if there is any). This network has no links to the other networks and it includes 100/1000 Mb/s Ethernet switch(es)

1.3.3 Backbone

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

This network links all the eth1 ports of the login nodes and external networks through a LAN network which includes Ethernet switches.

For performance and cluster security reasons it is advised to connect backbone only to login nodes.

1.3.4 Ethernet Network and Switch Management

Depending on the model, the switches can be managed:

- Directly by Ethernet
- Or through a serial line which allows the network management to be configured over Ethernet. For this:
  - A serial line has to be connected to each switch
  - The Ethernet cables need to be connected to the switches.

Some useful parameters which may be set up at this step include multicast management, **ARP** (Address Resolution Protocol) management and the Fast Spanning Tree protocol. Check the manufacturer's documentation to see which options have to be set for the device.

1.4 NovaScale Administration Networks

Bull High Performance Computing clusters feature different NovaScale Series machines and have different Administration Networks as indicated in this section.

1.4.1 PAP/PMB Network (NovaScale 5xxx/6xxx Series)

The Platform Administration Processor (**PAP**) is used to manage the node (power on or power off for example) and to get hardware information (state of the processor, temperature, etc.) for all the hardware components of the cluster. It uses a separate Ethernet network. One PAP can control 1 to 16 nodes. In addition a master PAP connected to PAP(s) managing nodes is necessary. This network has no links with the other networks. It includes:

- **PAP** (Platform Administration Processor).
- The **PMB** (Platform Management Board) of each node.
- 100 Mb/s Ethernet switch.

The **PAP/PMB network** interfaces all **PMB**s and the **PAP**.

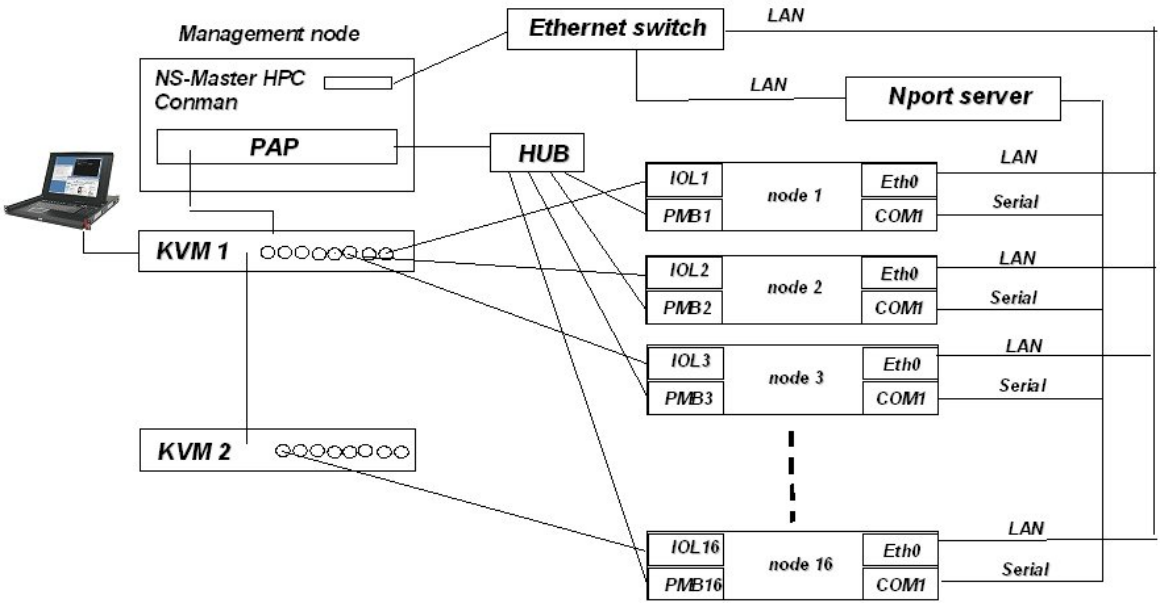

Figure 1-2. Administration network for a NovaScale 5000/6000 cluster (example)

1.4.2 NS Commands/ISM/BMC Network (NovaScale 40x0 Series)

The **NS Commands** and Service Manager software interface nodes through **BMC** maintenance coprocessors (**IPMI1.5**).

Console concentration is with **conman** through Eth switch and the **Nport** server.

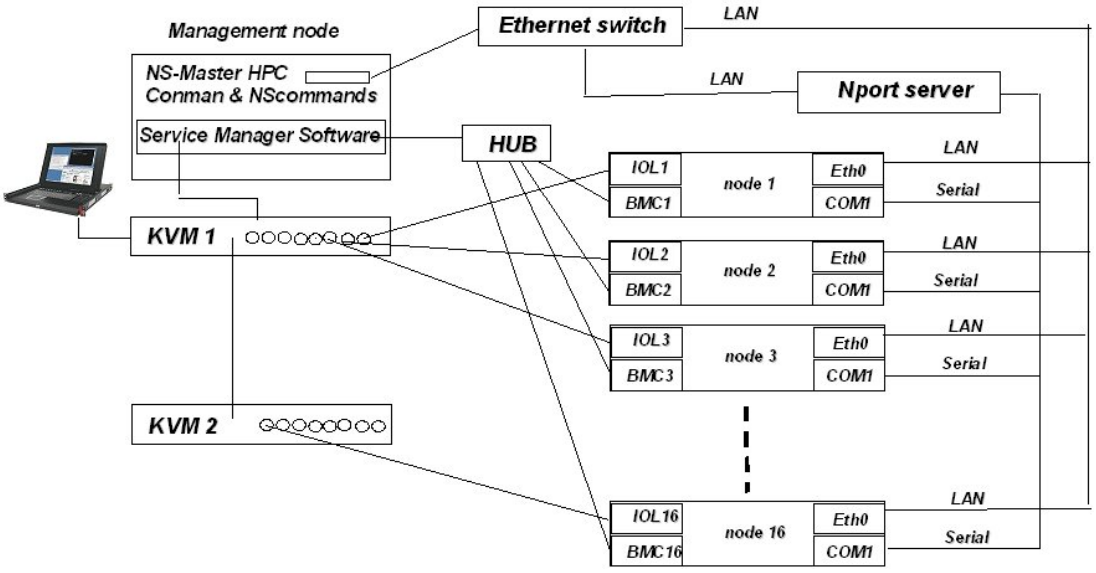

Figure 1-3. Administration network for a 40x0 cluster (example)

1.4.3 NS Commands/IPMI tools/BMC Network (NovaScale 3005 Series)

The NovaScale 3005 Series BMCs support **IPMI2**, this implementation supports common interfaces that allow IT managers to receive status alerts, send instructions to servers and run diagnostics over a LAN network. IPMI 2.0 is designed to extend customers' IT capabilities and further improve remote management by introducing enhanced security, remote access and configuration capabilities, while maintaining compatibility with previous IPMI versions.

New features include:

- **Serial Over LAN** - supports remote interaction with serial-based applications, BIOS and operating system
- **SMBus System Interface** - provides low-pin count connection for low-cost management controllers
- **New User Login and configuration options** - enable user access-rights and security configuration capabilities to be tailored to the needs of the user's facility.

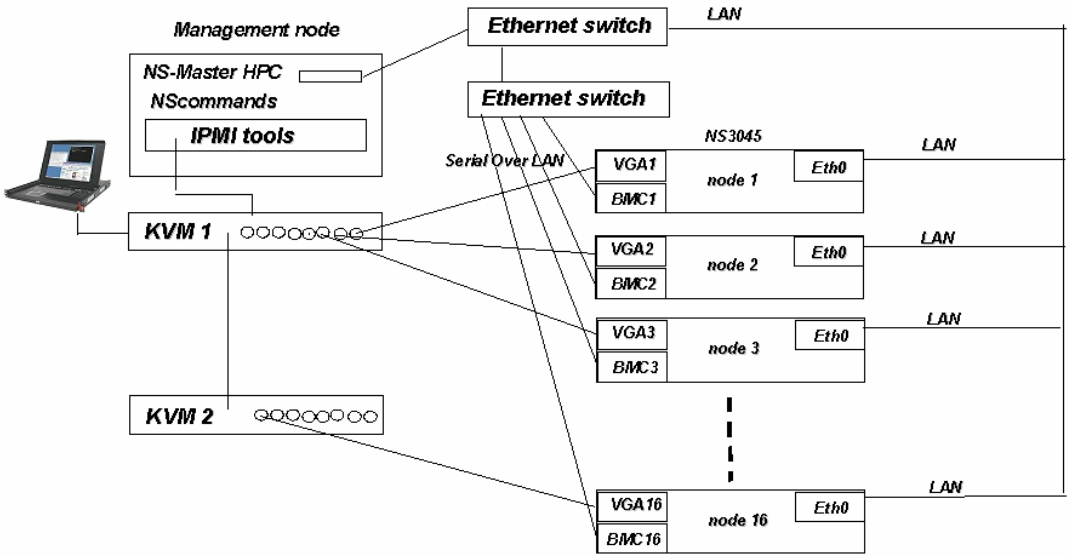

Figure 1-4. Administration network for a NovaScale 3045 (Compute Nodes) cluster

1.5 Main Console and Hardware Management

1.5.1 System Console

PAP (NovaScale 5xxx / 6xxx Series)

NovaScale 5xxx/6xxx nodes are delivered with an integrated administration tool named **PAM** (Platform Administration and Maintenance). The PAM software runs on the **PAP** (Platform Administration Processor) unit under Windows. From the PAP you access all the NovaScale System **PMB**s (Platform Management Board) via the LAN. The PMB runs **VxWorks** and **MAESTRO** agents. It is strongly advised that the NovaScale administration documentation is read.

The secured PAM web-based interface is used to operate, monitor, and configure the NovaScale nodes. Once the Windows Server 2003 software has booted, the user will be prompted to supply a user name and a password to open a Windows Server 2003 session. The system MUST be started for the first time with the following factory defaults for the user name and password:

User Name = Administrator
Password = administrator.

From Windows Server 2003 desktop, double-click the Microsoft Internet Explorer icon to launch the web-based administration tools. These tools are used to:

- Power ON/ Power OFF (Force Power Off)
- Check the hardware configuration
- Check the BIOS /Firmware environment.

The **PAM** user interface provides a status panel, a PAM tree and a control panel, which allow users to check the system status at a glance.

Server Manager or IPMI_tools and IPMI 1.5 & 2 (NovaScale 40x0 and NovaScale 3045 Series)

The management node runs the management software servicing the IPMI protocol to access the **BMC**s of the managed nodes. These tools are used to:

- Power ON/ Power OFF (Force Power Off)
- Check the hardware configuration.

1.5.2 Keyboard Video Mouse (KVM)

The **KVM** (Keyboard Video Mouse) is an optional equipment switch allows hosts to be controlled from the front console (better use the mandatory serial network and associated tools instead). Other hosts can be added to the KVM using the cables provided.
From the console, it is possible to switch to another system linked to the KVM using a « Ctrl+Ctrl » key sequence (Ctrl quickly followed by Ctrl).
The system console may also host storage management tools for Bull **FDA** storage systems if no other tools are available.

1.5.3 Hardware Management

Hardware management administration and maintenance tools give immediate insight into the status and configuration of the system. These can be used to operate, monitor, and configure the server. The following commands may be used:

- **PAM** commands available on the **PAP** platform. For details about **PAM** commands see the *NovaScale 5xx0 & 6xx0 - User's Guide*.
- **NS-commands** installed on the Management node. These commands invoke the PAM or the **Server Management** Software or the **IPMI_tools** according to the platform type you have and interact with nodes to manage equipment through the LAN network. These are described in the *NovaScale Master Remote HW Management CLI Reference Manual*.

1.5.4 Console Management

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 telnet protocol). **ConMan**, and its advantages over a simple telnet connection, are described in *HPC BAS4 Maintenance Guide*.

Note: ConMan is not applicable for NovaScale 3xx5 series platforms.

Sometimes, accessing the console is the only way to diagnose and correct software failures, for example when kernel debugging: **Kdb** is active before the Ethernet driver is loaded, thus, usable only with an asynchronous line.

The console allows:

- Access to the firmware shell (**BIOS/EFI**) to get and modify **NvRAM** information, choice of boot parameters e.g. for the kernel, the disk on which the node boots, a CDROM used to make an OS installation.
- Boot monitoring.
- Boot interventions like interactive file system check (**fsck**) at boot.
- Telnet sessions, using the command:

```
telnet <serial IP address>
```

Note:
It is not possible to use the telnet command if conman is active. Use instead this command:

```
conman <serial IP address>
```

NovaScale 3xx5 Series clusters which use ipmitools instead of ConMan

Example:

```
ipmitool -I lanplus -C 0 -U <BMC_user> -P <BMC_pwd> -H <BMC_Ipaddr> sol activate
```

1.6 High Speed Interconnection

1.6.1 Quadrics Networks

These networks may include **Gigabit Ethernet** switches or **Quadrics QsNet^{II}** Interconnects and **Elan4** cards to provide data transfer between the nodes of the cluster.

Using **QsNet^{II}** technology from **Quadrics** for interconnection provides an effective bandwidth of 900 MB/s (from a user space to user space), latency lower than 5 µs and the possibility to interconnect up to 4096 nodes.

- The **QsNet^{II}** network is composed of a **QsNet^{II}** switches and **Elan4 (QM-500)** boards.
- The **QsNet^{II}** switch offers more features including packet error correction and load balancing through dynamic routing. Transfer latency is around 21 nsec.
- Elan4 boards are able to support 64 bit virtual addressing and 900 MB/s in both directions at the same time.

To install Elan4 (QM-500) cards

- Plug the card into a short **PCI-X** 133 Mhz slot (64 bits). Refer to **Appendix C** for more information on how to select the best PCI slots for optimum performance.
- Connect one end of the link cable to the QM-500 card.
- Connect the other end to the corresponding QS5A port number.

To install QsNet^{II} (QS5A) switches

- Connect an Ethernet cable to the management board.
- As previously described, all the nodes with an Elan4 card will have been connected to the defined port (corresponding to the node name).
- Power on the switch.
- Configure the network interface of the switch using a keyboard and a screen directly connected to the management board. This can also be remotely set from the console concentrator program.

Quadrics networks will use the **RMS** Resource Manager – see the HPC *BAS4 Administrator's Guide* for more details.

1.6.2 InfiniBand Networks with Voltaire Switching Devices and the SLURM Resource Manager

For **InfiniBand** Networks the interconnection generally uses **Voltaire®** devices including:

- **400 Ex-D** Double Date Rate (**DDR**) Host Channel Adapters which can provide up to 20 Gbs per second bandwidth.
- Smaller clusters will use the **Voltaire® ISR 9024** Grid Switches to route up to 24 **InfiniBand** DDR ports per switch.
- The **Voltaire® ISR 9096** and **9288** Grid Directors are used to scale up clusters which include **400 Ex-D** HCAs and **ISR 9024** switches. The **ISR 9096** and **ISR 9288 Grid Directors** house 96 and 288 **InfiniBand** DDR ports respectively and can be used to build extremely large clusters featuring thousands of nodes.

The InfiniBand\Voltaire solution uses a FAT Tree (Clos) topology and provides full bisectional bandwidth for each port.

For more information on installing and configuring Voltaire devices refer to *Chapter 6* in this manual and to the documentation provided by Voltaire.

These clusters will use the **SLURM** Resource Manager - see the HPC *BAS4 Administrator's Guide* for more details.

1.6.3 Ethernet Gigabit Networks

It is possible to use the Ethernet Administration Network for the interconnection if the administration traffic, following deployment, is low enough (somewhere between 5 – 10% of Gigabit bandwidth). If the administration traffic uses more than 10% of the available bandwidth then, as a rule, there will not be enough bandwidth for interconnect traffic, and performance will be impacted. In this situation a dedicated Ethernet network can be configured.

See *Appendix C* for details of the Ethernet connectors for each hardware platform.

1.7 Typical Types of Nodes

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

1.7.1 Management Node

The **Management Node** is dedicated to providing services and to running cluster management software. This node concentrates on one node all control and management functions. The services, for example, may be those from the **NIS, NTP, Lustre LDAP** and **Lustre Services, Cluster DataBase, Kerberos, KDC, HA system status, MiniSQL, RMS daemon, IC switch manager(Quadrics), snmtrapd, ganglia, dhcpd, httpd, conman** etc.

The Management Node can also be configured as a gateway for the cluster. You need to connect it to the external LAN and also to the management LAN using two different Ethernet cards. For management purposes you will need a screen and a keyboard/mouse. Another Ethernet connection must be setup to connect to the **PortServer** for asynchronous connection to node terminals. An **Elan4** (QM-500) card (or higher) may also be plugged in if **EIP** (Encapsulated IP) or I/O operations is going to be done from the Management Node.

The Management Node stores a lot of reference data, and operational data, which may be used by the Resource Manager and other administration tools. It is recommended to store data on a **RAID** storage system. Configure this storage system before creating the file system for the management data on the Management node. Refer to *Appendix C* in this manual for information on how to select the best PCI slots for optimum performance.

Note:
In addition the Management Node can also be configured to provide the login node functions.

High availability feature:

The management node can be protected against failures by coupling it with a secondary passive management node. In case of failure of the active management node, all the operations are taken over by the second management node. For more information about this feature, please refer to the *HPC BAS4 Administrator's Guide* and Chapter 2 in this manual.

1.7.2 Compute Node

The **Compute nodes** are exclusively optimized to execute parallel code. These nodes benefit from the Itanium2® CPU power, with its high floating point calculation performance. The Itanium2 has 3 levels of cache and uses the **EPIC** (Explicitly Parallel Instruction set Computing) architecture to deliver very impressive **linpack** results.

Network Adapters (Quadrics QM500 or Voltaire 410 Ex cards), a serial line to the PortServer and an Ethernet link to the management LAN must also be present on the node.

In the case of compute nodes without **PAP** it is necessary to configure the **PMB** ID to setup the LAN connection to the main **PAP** used to manage multiple NovaScale systems. For this you have to select a unique ID for each compute node on the same **PAP** and then plug the Ethernet into a dedicated hub (or switch).

If present, the storage system(s) must be configured prior to the configuration of the file system used by the nodes.

1.7.3 Login Node

Login nodes are dedicated to providing cluster access and a software development environment for the cluster's users. These are used to:

- Login
- Develop, edit and compile programs
- Run programs in a reduced environment which may consist of just one machine
- Debug parallel code programs.

1.7.4 I/O Node

The I/O nodes are optimized to run the **Lustre** file system. They are connected to high speed cluster interconnects and to high capacity external **RAID** storage arrays. They provide the **OSS** and **MDS** services for Lustre file systems. For large clusters, these nodes usually operate in pairs in order to ensure High Availability using mutual take over capabilities.

Specific configuration is necessary for the Lustre file system, as detailed in the *Installation Process* chapter of this manual. See also the *HPC BAS4 Administrator's Guide* for more information about file system configuration.

1.8 Installing Software and Configuring Nodes

Before installing software on the nodes, you must know the node distribution that has been planned for your HPC system (Management Nodes, Compute Nodes, Login Nodes, I/O Nodes).

Chapter 2 explains how to install software either on a Management node, on a Compute Node or on a Login Node.

The software installed on a **Compute Node** may be used by **Ksis** - a utility for image building and deployment - for the creation of a reference image to be deployed throughout the cluster to create other Compute Nodes. The **Reference Node** designates the node from which the reference image is taken.

The software installed on a **Login Node** is the same as for a Compute Node with the addition of the development environment.

To create an **I/O node** after deployment of a reference image, you will have to configure file systems (for example Lustre).

See: Appendix D *BAS4 Bundles* describes the bundles installed on each type of node.

For a **Standalone configuration**, where only one node is configured, see:
Chapter 3 *Installation of a Standalone Configuration.*

Chapter 2. Installing BAS4 Software on the HPC Nodes

This chapter describes the complete process to install the software environment on all the nodes of a Bull HPC configuration. It covers different installation situations: **first installation**, **migration** and **re-installation** of BAS4 software with the preservation of the cluster database.

Installation Process Overview

The process to install the software on the cluster's nodes is divided into several steps:

STEP 1	Saving the database and the configuration files. **Skip this step if you are installing for the first time.** Use this step only in the case of migration or re-installation, when the cluster is already configured (or partially configured) and there is the desire to save this configuration.	Page 2-6
STEP 2	Installing the software on the Management node(s): 1) Installation of the Bull Linux Advanced Server software. 2) Configuration of the network. 3) Installation of External Storage System 4) Installation of additional software 5) Configuration of the database.	Page 2-9
STEP 3	Configuring equipment and initializing tools: 1) Configure equipment manually in some specific cases 2) Configure the management tools 3) Configure BMCs on NovaScale 3005 Series.	Page 2-31
STEP 4	Installing the software on a Compute Node or a Login Node: 1) Installation of the software: either from a CD or DVD (provided the node has a drive), or from a server, which may be the Management Node, via NFS (Network File System). 2) Reboot of the system node and post-boot installation of some software.	Page 2-35
STEP 5	Configuring Administration Software ssh, PDSH, ganglia, syslog-ng, NTP, SNMP Server, postfix, SLURM, InfiniBand are configured on the Management Node(s) and on the Compute or Login Node, as necessary.	Page 2-44
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.	Page 2-62
STEP 7	Post Installation Configurations for InfiniBand Clusters Configuration of InfiniBand interfaces	Page 2-65

Special Installation Features

Some features will have an impact on the installation process:

VLAN (Virtual Local Area Network) Configuration

The configuration of the Network interfaces differs for clusters configured **with VLAN** from those **without VLAN**.
(Clusters configured without VLAN are generally smaller clusters, consisting of less than 20 nodes. Clusters configured with VLAN are generally larger clusters - more than 20 nodes).

The installation and configuration specificities for these features are clearly indicated throughout this chapter.

Management Node High Availability

This feature requires two Management Nodes, known as the **Primary Node** and the **Secondary Node**. The installation and configuration for these two nodes is not the same. This is described all along this chapter.

See the *HPC BAS4 Administrator's Guide* for details about the concepts, installation, configuration and operation of the High Availability feature.

Standalone Configuration

For a **Standalone configuration**, where only one node is configured, see Chapter 5, *Installation of a Standalone Configuration* in the present guide.

Note:

If a **PortServer** is not configured it may prevent a node from booting. In this case, you must disconnect the PortServer.

RAID Configuration for BAS Installation Devices

BAS can be deployed by taking advantage of some of RAID's features, according to the SCSI **HBA** used and the number of SCSI disks delivered with the NovaScale system. RAID can be used for:

- **LSI 1068** (onboard or **LSI SAS 3442X HBA**)
- **LSI MegaRAID 320-2x HBAs**

It is necessary to configure the RAID volume before installing BAS. The procedure is described below for each HBA:

LSI 1068 (onboard or LSI SAS 3442X HBA)

If the system has to be installed on SAS disks grouped in RAID volumes managed by the **LSI 1068** chip or by an additional LSI SAS **3442X** board the RAID volume has to be configured before installing the system using the EFI menu as described below:

Using the EFI menu, list the EFI drivers to obtain the **LSI MPT SAS** driver ID by using the command below:

```
Shell> drivers
```

The drivers command result appears as follows:

```
Shell> drivers
                T   D
D               Y C I
R               P F A
V  VERSION      E G G #D #C DRIVER NAME                          IMAGE NAME
== ========     = = = == == ================================== ==========
0F 00000010     B - -  4 190 PCI Bus Driver                      PciBus
10 00000010     D - -  1  - PC-AT ISA Device Enumeration Driver  PcatIsaAcpi
11 00000010     B - -  1  3 ISA Bus Driver                       IsaBus
12 00000010     B - -  1  1 ISA Serial Driver                    IsaSerial
13 00000000     D - -  1  - BIOS[INT10] VGA Mini Port Driver     BiosVgaMiniPort
14 00000010     D - -  1  - VGA Class Driver                     VgaClass
15 00000010     ? - -  -  - UGA Console Driver                   GraphicsConsole
16 00000010     B - -  1  1 Serial Terminal Driver               Terminal
17 00000010     D - -  4  - Usb Uhci Driver                      UsbUhci
18 00000010     D - -  4  - USB Bus Driver                       UsbBus
19 00000010     ? - -  -  - Usb Cbi0 Mass Storage Driver         UsbCbi0
1A 00000010     ? - -  -  - <UNKNOWN>                            UsbCbi1
1B 00000010     ? - -  -  - Usb Keyboard Driver                  UsbKeyboard
1C 00000010     ? - -  -  - Usb Bot Mass Storage Driver          UsbBot
1D 00000010     ? - -  -  - Generic USB Mass Storage Driver      UsbMassStorage
1E 00000010     ? - -  -  - Usb Mouse Driver                     UsbMouse
1F 00000010     D - -  2  - Platform Console Management Driver   ConPlatform
20 00000010     D - -  1  - Platform Console Management Driver   ConPlatform
21 00000010     B - -  1  1 Console Splitter Driver              ConSplitter
22 00000010     ? - -  -  - Console Splitter Driver              ConSplitter
23 00000010     B - -  2  2 Console Splitter Driver              ConSplitter
24 00000010     B - -  2  2 Console Splitter Driver              ConSplitter
EC 00000010     B X X  1  1 PCI IDE/ATAPI Bus Driver             Ide
ED 00000010     ? - -  -  - ATAPI SCSI Pass Thru Driver          AtapiScsiPassThru
EE 00000010     D - -  7  - Generic Disk I/O Driver              DiskIo
EF 00000010     B - -  2  4 Partition Driver(MBR/GPT/El Torito)  Partition
F0 00000010     D - -  1  - FAT File System Driver               Fat
F1 00000010     B - -  1  1 Intel(R) PRO 100 UNDI Driver         Undi
F2 00000000     ? - -  -  - BIOS[UNDI] Simple Network Protocol   BiosSnp16
F3 00000010     D - -  3  - Simple Network Protocol Driver       Snp3264
F4 00000010     D - -  3  - PXE Base Code Driver                 PxeBc
F5 00000010     D - -  3  - PXE DHCPv4 Driver                    PxeDhcp4
F6 FFFFFFF0     ? - -  -  - Serial Mouse Driver                  SerialMouse
F7 00031115     D X X  3  - Emulex SCSI Pass Thru Driver        Elxcli311a5
F8 02000900     B X X  2  2 LSI Logic Fusion MPT SAS Driver      EFIDriver
F9 03001900     B X X  2  2 Intel(R) PRO/1000 3.0.19 EFI-64      EFIDriver
```

Figure 2-1. The drivers command screen

Invoke the **LSI EFI** driver using the **drvcfg -s** command with the SAS driver ID:

```
Shell> drvcfg -s F8
```

The **LSI MPT** Setup utility screen is displayed:

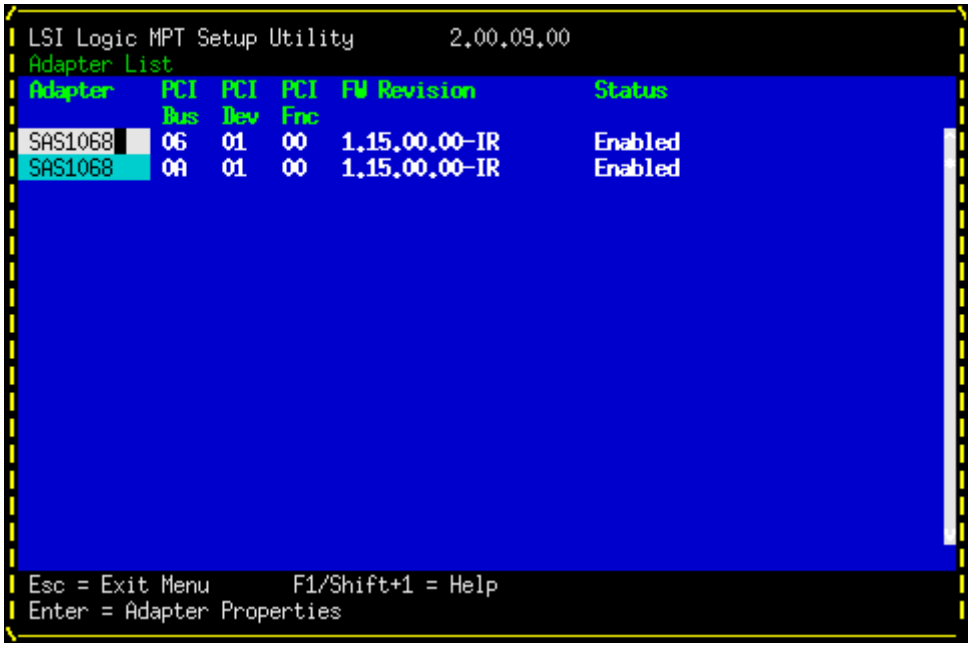

Figure 2-2. The LSI MPT Utility Setup – screen 1

The RAID volumes may now be configured by entering the RAID properties menu of one SAS 1068 controller (see LSI documentation for further details).

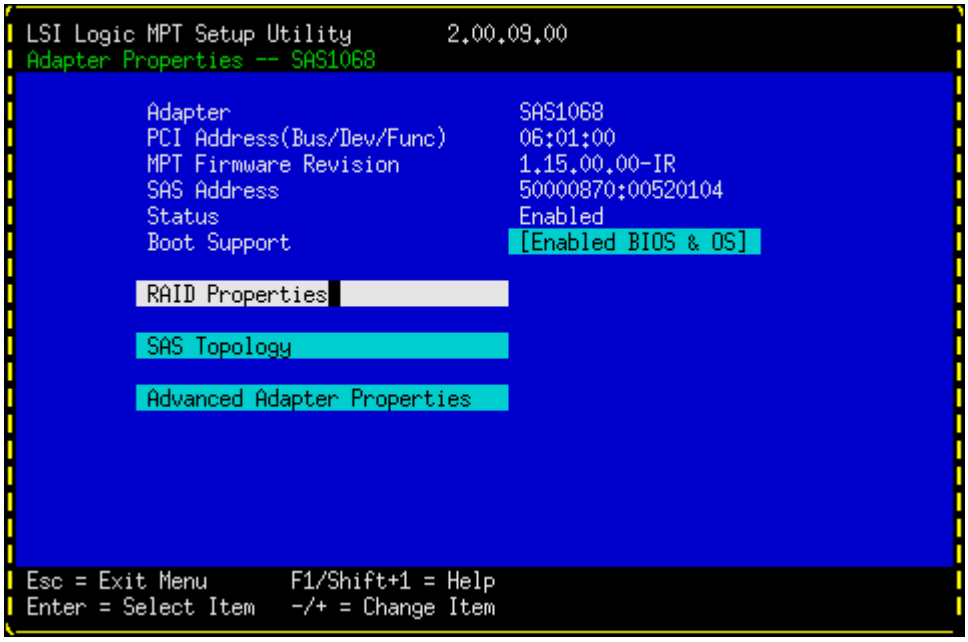

Figure 2-3. The LSI MPT Utility Setup – screen 2

LSI MegaRAID 320-2x HBAs

Ultra 320 disks are used on the 2 SCSI buses by this HBA model (up to 30 disks). RAID0, RAID1, RAID5, RAID10 and RAID50 volumes can also be provided. RAID volumes have to be created before installing the **BAS** system.

When starting the NovaScale server, wait for the **EFI** boot menu and select EFI shell. Then, use the MegaRAID **EFI** configuration tool (provided on CD-Rom **76679543-004**) to create the RAID volumes. When RAID volumes have been created, BAS can then be installed (RAID Volumes are shown with Vendor='**MegaRAID**' and Model='**MegaRAID LD X RAIDY ZG**', where **X=logical volume** ID, **Y=raid level** and and **Z=volume size in Gbyte**s).

Note:
For more detailed information see the *MegaRAID® Configuration Software User's Guide* available from http://www.lsi.com

About EFI (Extensible Firmware Interface)

The EFI (Extensible Firmware Interface) menus are fully described in the server documentation, for example:
NovaScale 5xx5 & 6xx5 User's Guide and *NovaScale 5xx0 & 6xx0 User's Guide*.

Managing the EFI Boot entries using efibootmgr

The **efibootmgr** command allows the EFI boot entries to be displayed and their order changed.

- To display the EFI boot entries, enter:

```
efibootmgr
```

- To change the boot entries order, use the **–o** option. For example:

```
efibootmgr -o 0001,0003,0000,0002
```

Refer to *HPC BAS4 Maintenance Guide* for more details about **efibootmgr**.

2.1 STEP 1: Saving the Database and the Configuration Files

This step describes how to save the ClusterDB database and important configuration files.

Use this step only in the case of migration or re-installation, when the cluster is already configured (or partially configured) and there is the desire to save this configuration.

Skip this step when installing for the first time.

Warning:

The Operating System is installed from scratch, erasing all disk contents in the process.

It is the customer's responsibility to save data and the 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 should be saved.

Important:

All the data must be saved onto a **non-formattable** media.
It is recommended to use the **tar** or **cp -a** command, which keeps the files permissions.

2.1.1 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`.

Note

For the case of a migration, a script is used which requires that the ClusterDB save file is called **clusterdbdata.sav** (see 2.2.6.2 *BAS4 Migration with ClusterDB Preservation*).

4. Copy the two **.sav** files onto a non-formattable media.
5. Take note of 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 and the *Management Node High Availability* feature is **NOT** implemented.

2.1.2 Saving SSH Keys of the Nodes and of root User

To avoid RSA identification change, you can keep the SSH keys.

- To keep the **Node SSH keys**, save the **/etc/ssh** directory on each type of node (Management Node, Compute Node, Login Node, …), 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.

You will restore these directories after installation (see 2.5.1 *Configuring SSH*).

2.1.3 Saving the Storage Configuration Information

The following configuration files, under the **/etc/storageadmin** directory of the Management Node, are used by the storage management tools. It is recommended that these files are saved onto a non-formattable media, as there is no automatic saving of these files in the case of migration or re-installation.

- **storframework.conf** configured for traces, etc
- **ddn_admin.conf** configured for any **DDN** disk array administration access
- **nec_admin.conf** configured for any **FDA** disk array administration access
- **dgc_admin.conf** configured for any **DGC** disk array administration access

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

If the **I/O** & **MDS** node symbolic links for the devices have been created from the Management Node using the **stordiskname** command with the **–r** option (remote), and not by using the **stordepmap** command, the **/etc/storageadmin/disknaming.conf** files must also be saved.

The **disknaming.files** directory will include the **disknaming.conf.<node_name>** files for all the **I/O** & **MDS** nodes

See Chapter 3 for more information.

If the **stordiskname** command has not been used with the **–r** option (remote), the Administrator will have managed the backup of the specific file generated by the **stordiskname** command himself. All the backup files must have been saved separately, wherever they may be.

2.1.4 Saving the Lustre File Systems

The following files are used by the **Lustre** system administration framework. It is 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 High-Availability capability is enabled: **/var/lib/ldap/lustre** directory.

2.1.5 Saving the SLURM Configuration

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

2.2 STEP 2: Installing Software on the Management Node(s)

This step describes how to install the software on the Management Node(s). It includes the following sub-tasks:

1. Installation of the Bull Linux Advanced Server software.
2. Configuration of the network.
3. Installation of External Storage System.
4. Installation of additional software.
5. Configuration of the database.

If the *Management Node High Availability* feature is implemented:
in this case the cluster has two Management Nodes (Primary and Secondary Management Nodes); install the software and configure these Nodes as described in this step.

2.2.1 Bull Linux AS4 (BLAS) Installation

2.2.1.1 System Boot

Bull Linux AS and BAS
Bull Linux AS (*Bull Linux Advanced Server*) consists of Kernel software from kernel.org and Open Source Linux distribution only, whereas **BAS** (*Bull Advanced Server*) designates the whole Bull HPC software offer, including Bull Linux AS and Bull Cluster Management software. The versions of Bull Linux AS and BAS may be different.

Important:
Before starting the installation it is recommended to read all the procedure details carefully.

Start with the following operations:

1. Power up the machine.
2. Switch on the monitor, if necessary.
3. Insert the DVD of **Bull Linux Advanced Server 4** into the 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. If this happens, run `map -r` under EFI (this command enables a system to search for devices).

After the initialization phases (BIOS, SCSI detection, etc), the screen disappears and the **EFI** banner is displayed, similar to that below:

4. At `"EFI Boot Manager menu"` and `"Please select a boot option"` use the key pad to go to:

   ```
   CD/DVD ROM/Pci(1F|1)/Ata(Primary,Master)
   ```

and then select Enter.

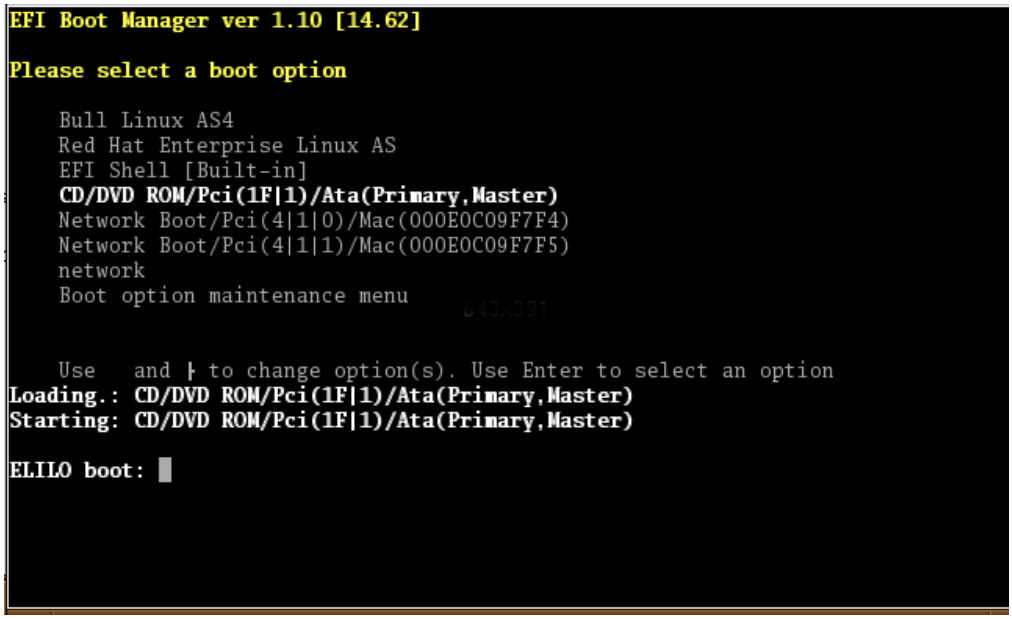

Figure 2-4. EFI Boot Manager – Choice of Boot Option

5. At the ELILO boot prompt press the Enter key to start the boot. Alternatively, the boot will start automatically after a timeout of 50 seconds.

6. Select the keyboard model for the system:

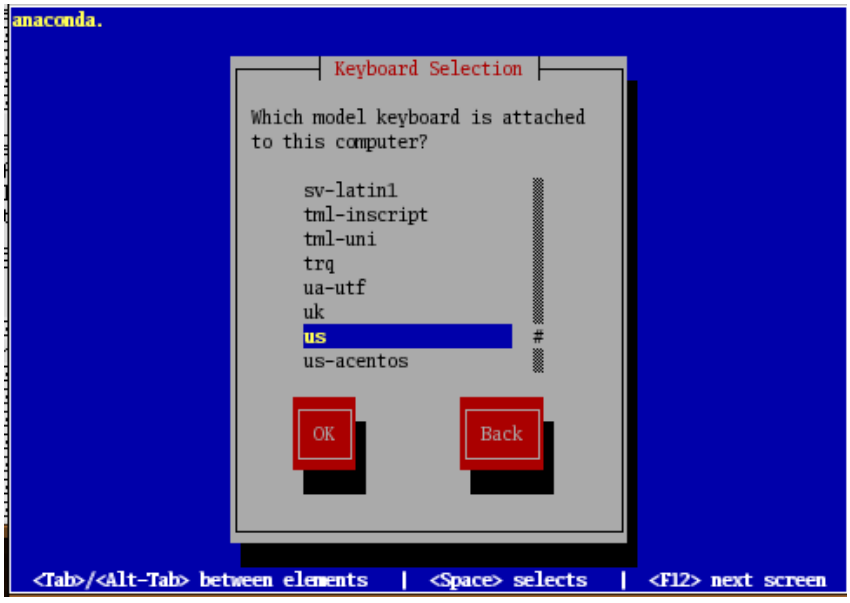

Figure 2-5. Keyboard selection screen

2.2.1.2 BLAS General Installation Procedure

A suite of screens helps you to install BLAS software on a Management Node, Compute Node, Login Node or Standalone Node.

Important:
The installation screens which follow apply to a Compute Node. Different settings, as indicated, have to be used for the Management Node and other types of nodes.

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

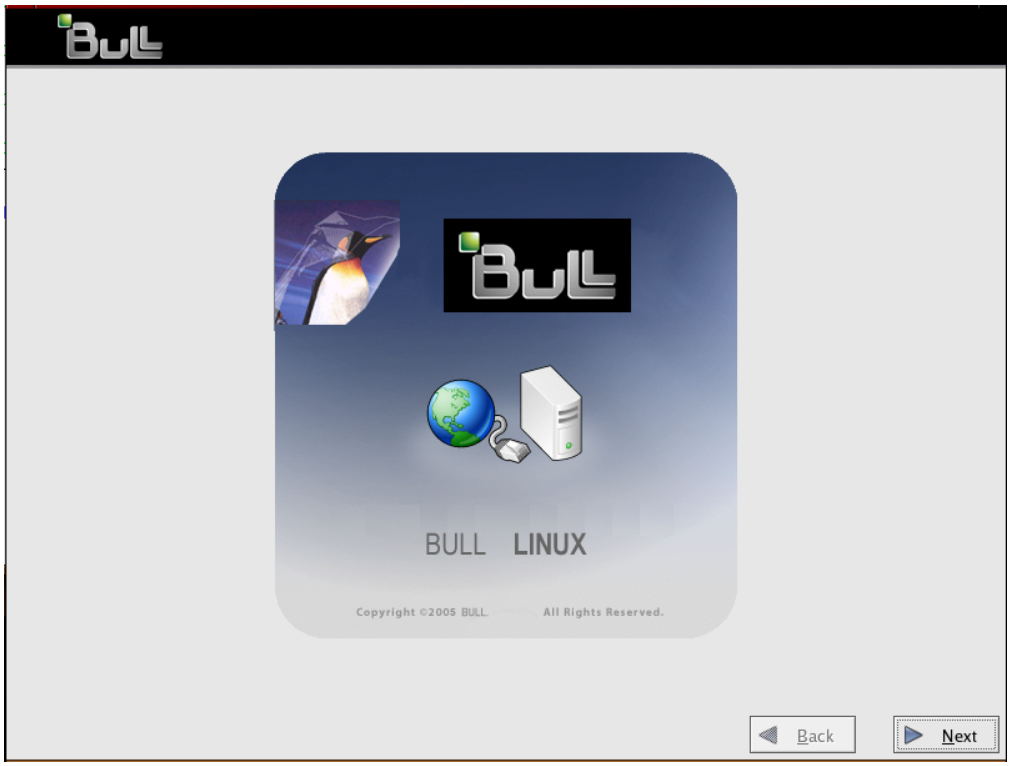

Figure 2-6. The Welcome Screen

Disk partitioning

2. Click on the **Next** button. The disk partitioning screen will appear as shown below. By default this shows the number of disks detected. Any partitioning modifications which are necessary can be made using the following screens.

Note:

For HPC, the automatic partitioning and **Remove all partitions** options on the system have been pre-defined and you will be advised, that ALL DATA will be removed. However, the options provided allow you to have complete control concerning which data (if any) is removed from your system.

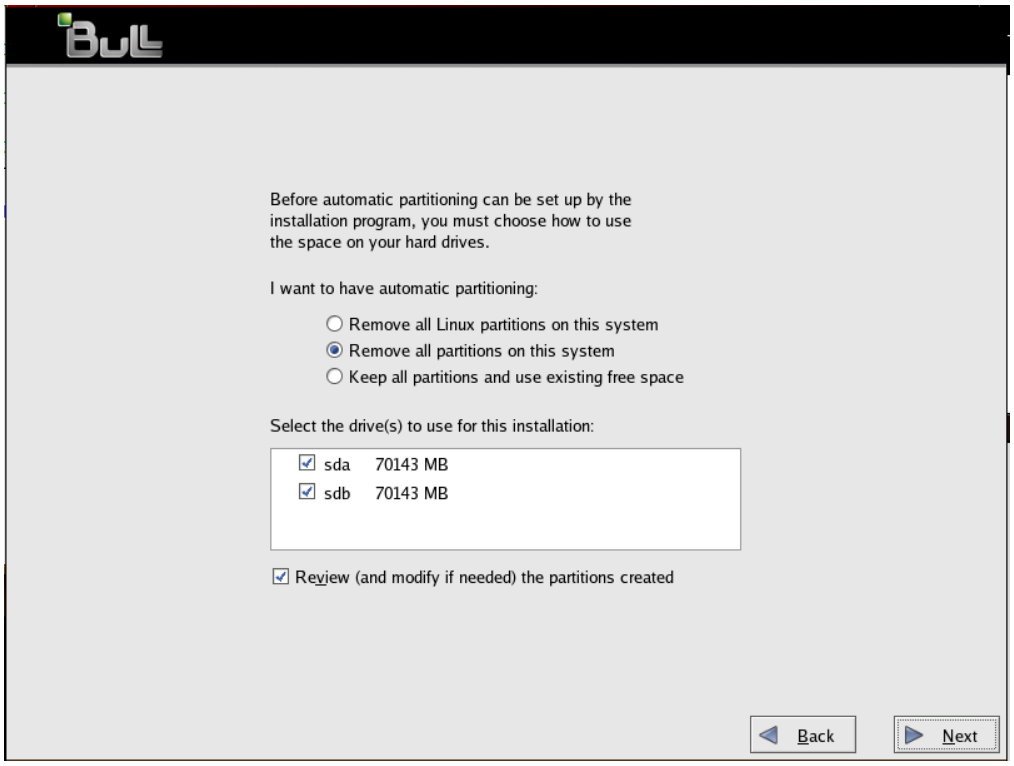

Figure 2-7. Drive and Partition screen

Important

Do NOT change the default settings above. It will be possible to select only one of the disks in step 4, below, should you want to.

3. Click on the **Next** button.

Note

During the installation process, the system detects the storage devices connected via the fibre channel and not named **sda**, **sdb**, **sdc**, and requests that they are formatted. Choose "**ignore**" until **install on SDA** option appears. Accept this and continue.

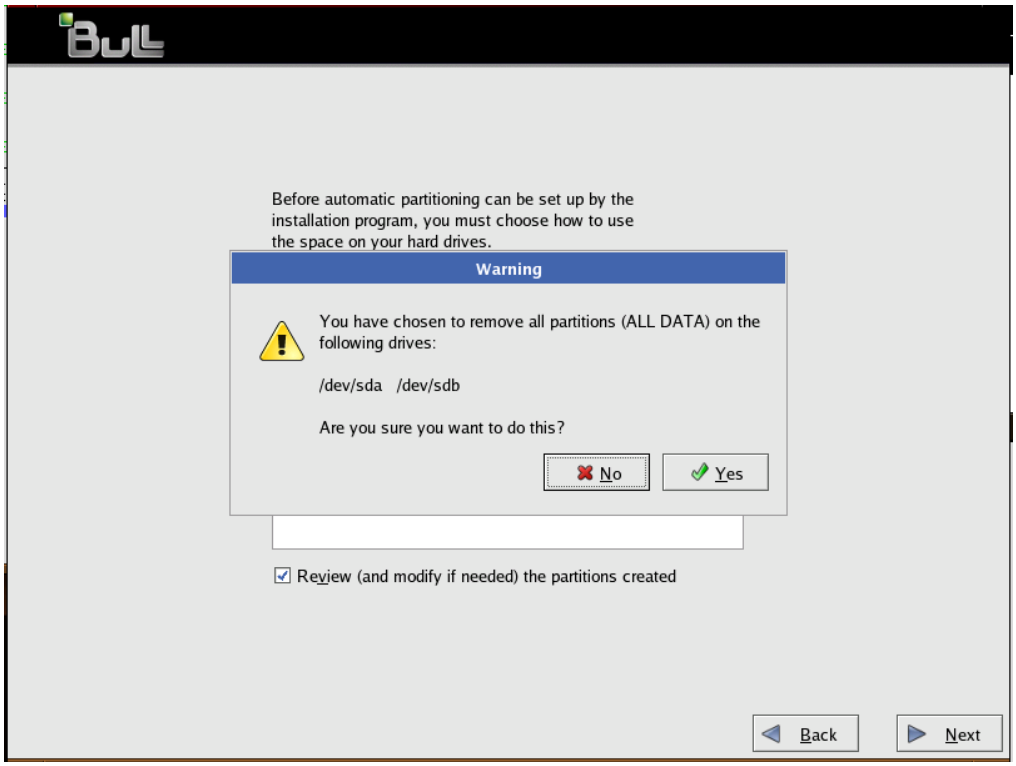

Figure 2-8. Data Removal Confirmation Screen

4. Select **Yes**, as at this stage there is no risk that data will be deleted. Click on the **Next** button. A screen similar to that below appears.

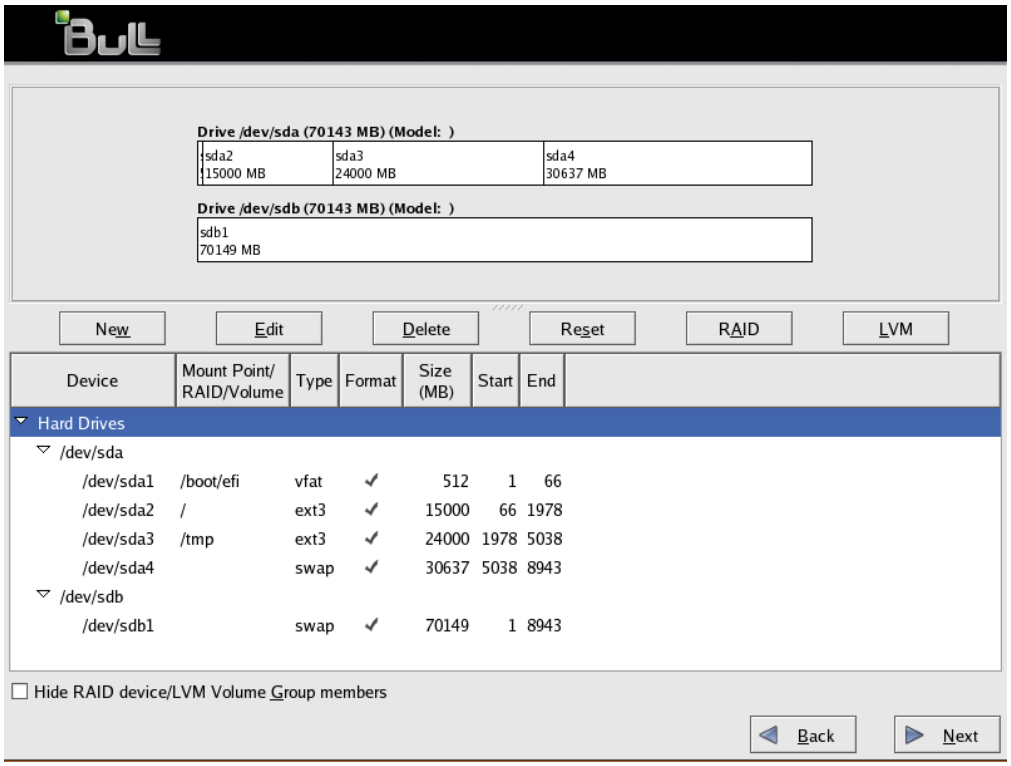

Figure 2-9. Default partitioning for a Compute Node

Note:
If you want to modify the original disk partitioning (not recommended for HPC systems), then click on the **Reset** button and create the new partitioning.

5. Click on the **Next** button.

Important
When the **Next** button is selected (Figure 2-9) any data on the partitions selected will be removed.

Note:
See the Bull *HPC BAS4 V5.1 System Release Bulletin* for partitioning examples for the different types of nodes.

Network access Configuration

6. The next step is to configure the access to the network using the screen below. The network configuration screen which appears is for a Compute Node.

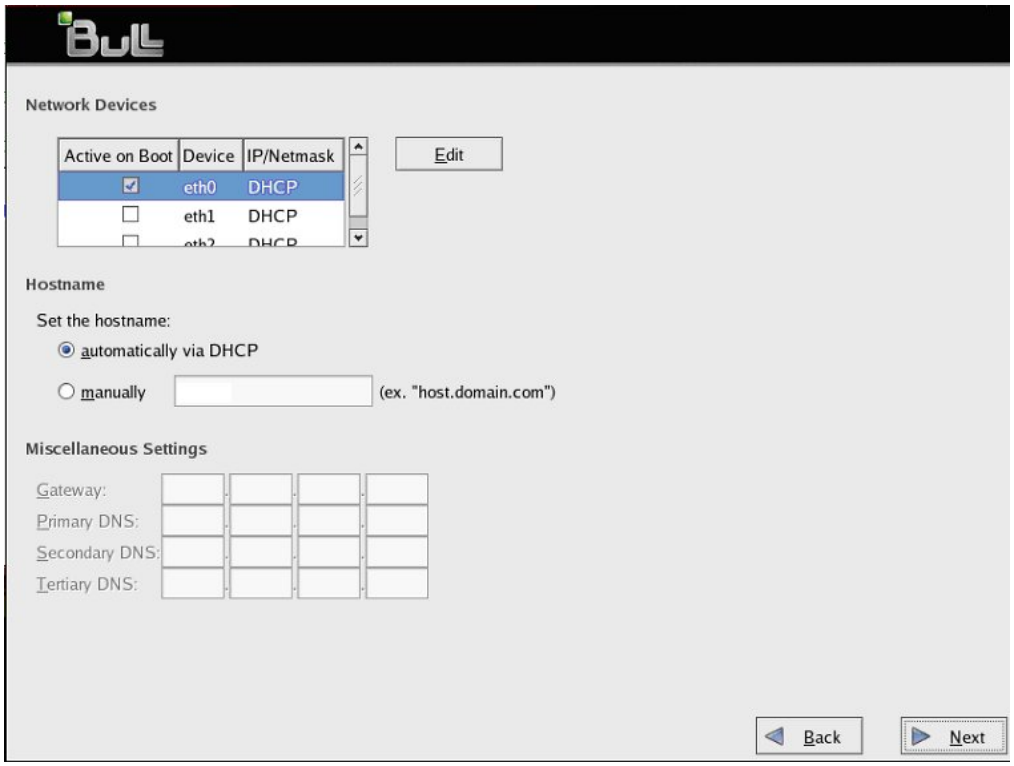

Figure 2-10. Network Configuration Screen for a Compute Node

Note:
Other nodes are configured manually by clicking on the **Edit** button, shown in Figure 2-10, and changing the settings accordingly.

7. If DHCP is not used then other Device Interfaces, Hostnames and Gateway configuration settings can be set manually – see Figure 2-11.

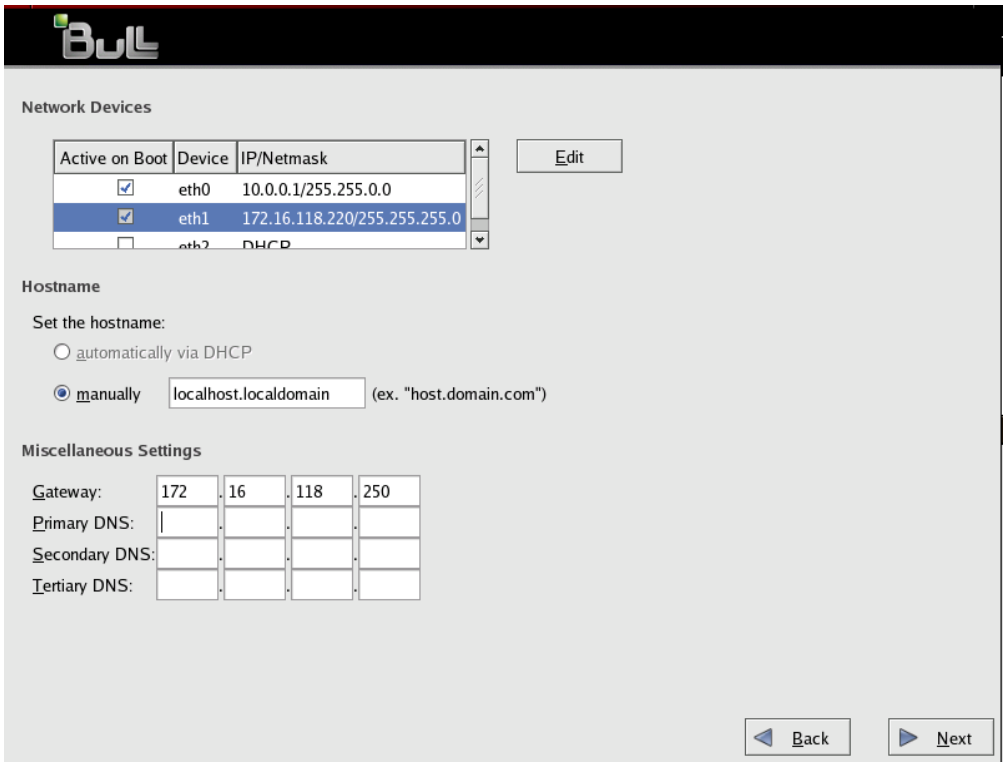

Figure 2-11. Network Configuration Settings

Language and Timezone Settings

8. Choose the Language for the system.

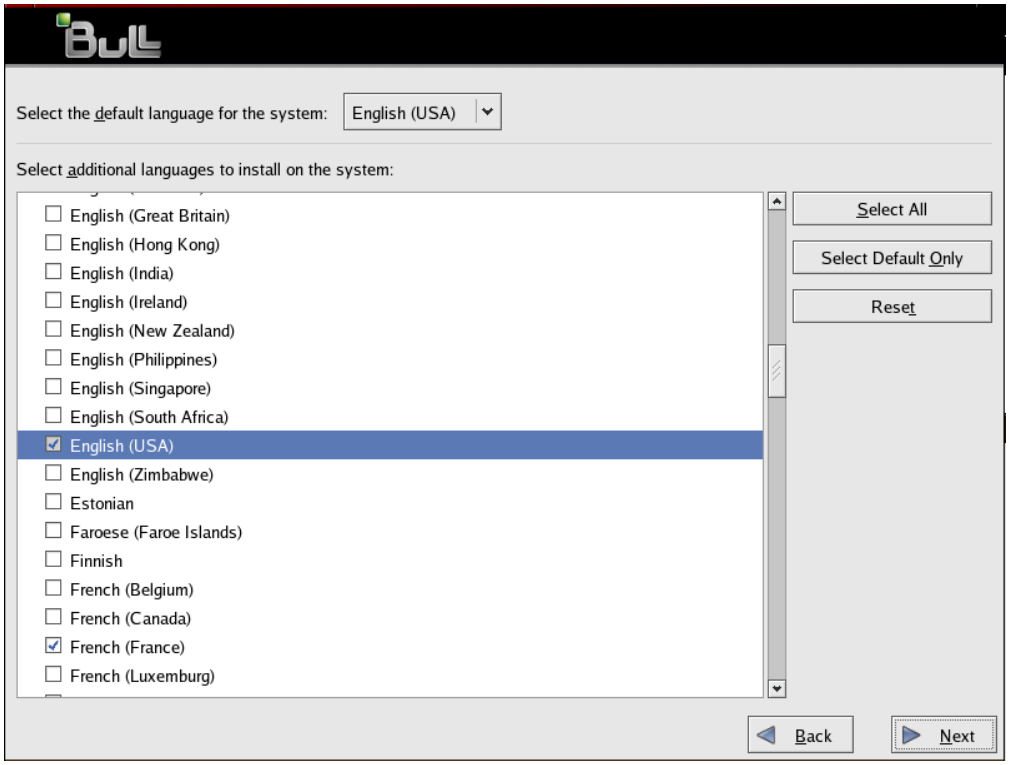

Figure 2-12. Language Selection screen

9. Choose the timezone.

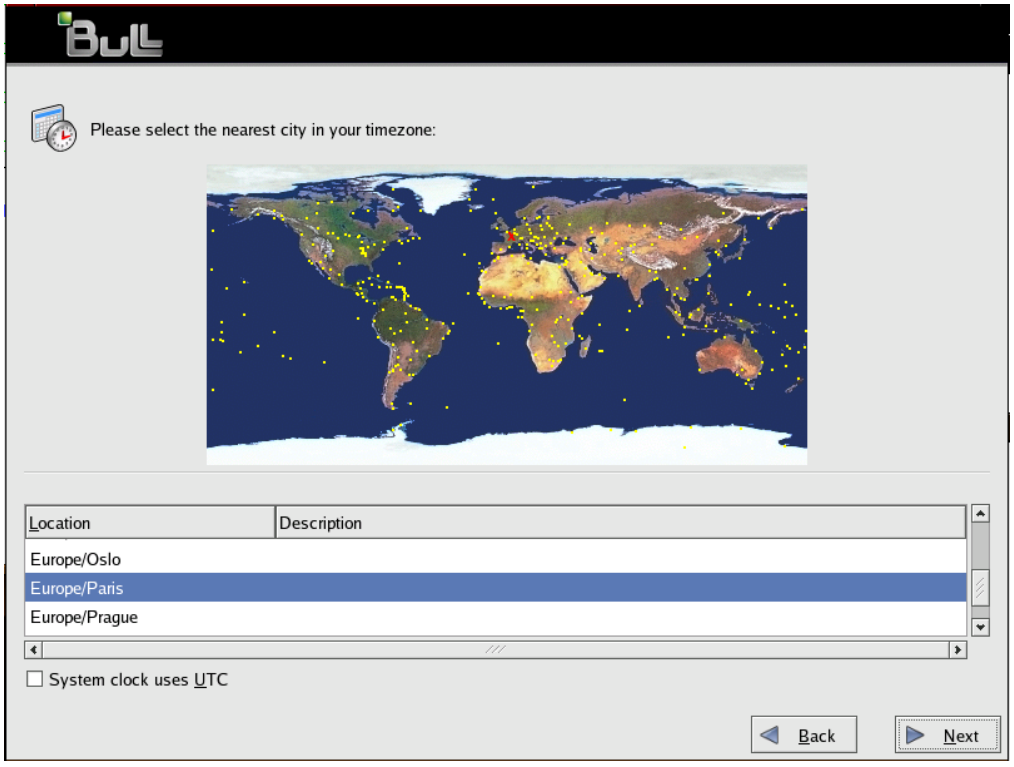

Figure 2-13. Timezone Selection screen

Note:
To use UTC (Coordinated Universal Time), thus helping to avoid problems of seasonal time changes, etc., check the **System clock uses UTC** box in the screen above.

Node Profile Installation

10. Select the Node profile installation required – see Figure 2-14. By default the **High Performance Computing Group for COMPUTE Node profile** is selected. This will install all the system packages required for this type of node. There is no need to select any additional tools or packages from the options listed below the list of HPC Group Node profiles.

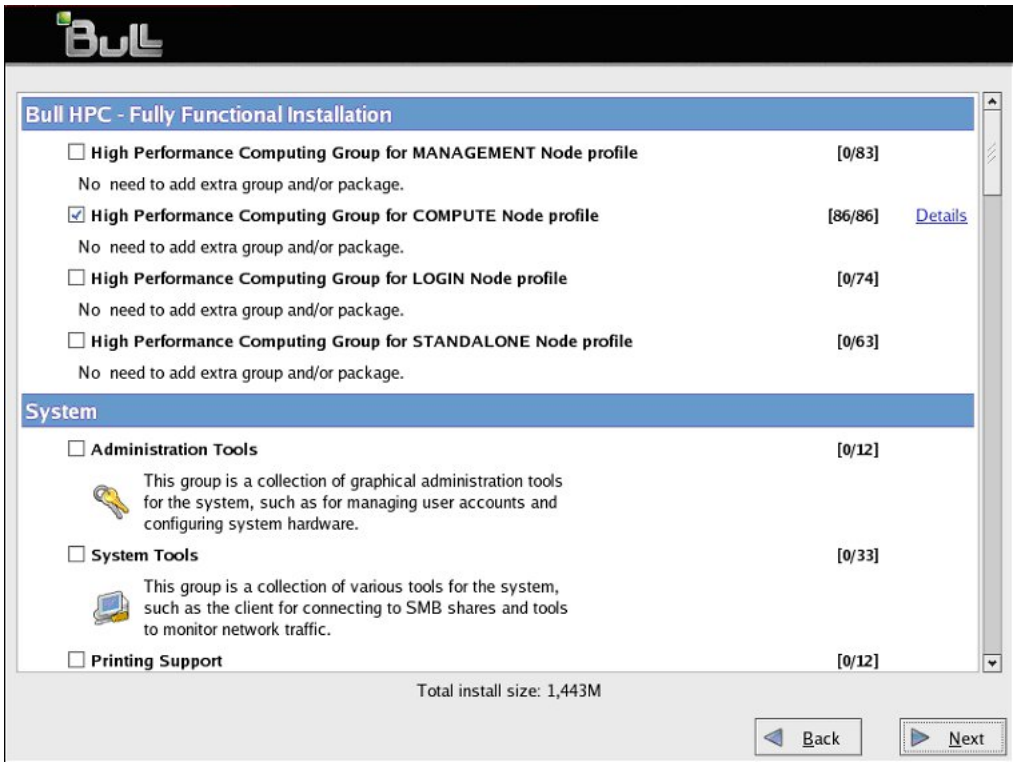

Figure 2-14. Node profile selection screen – by default a Compute Node is selected.

Important

The same process as described in points 1 to 9 is followed for the installation of the other Node profiles. At this point the **High Performance Computing Group for COMPUTE Node profile** check box **MUST** be deselected and the relevant profile selected.

Note

To install a **Login Node** on the **Management Node** select **BOTH** the Management Node and Login Node profile check boxes. Do not forget to deselect the Compute Node profile.

11. Finally, the confirmation screen for the node profile appears, as shown below. Click on Next to start the installation.

Figure 2-15. Installation Confirmation Screen

During the installation, the default password for the root user is set to "**root**". This password may be changed after the next reboot.

Once the Node profile installation has finished the node will reboot automatically.

NovaScale 5xxx Series – Workaround for CD/DVD drives

On **NovaScale 5xxx Series**, after the **BLAS** installation has finished, the CD and DVD drivers will not be active. The solution for this is as follows:

1. Edit the **/etc/udev/rules.d/50-udev.rules** file.

2. Uncomment line number 343 so that:
   ```
   #       RUN+="/sbin/modprobe sr_mod"
   ```
   is replaced with:
   ```
          RUN+="/sbin/modprobe sr_mod"
   ```

3. Reboot the node.

2.2.2 Disk Health Monitoring Configuration

By default, the disk health monitoring feature (**smartctl**) is configured for **sda**, **sdb** and **sdc** on the Management Node. But if you have other local disks you must manually configure **smartctl** for these disks.

Note: Do not configure **smartctl** on RAID HBA devices or external RAID storage systems.

Use the following procedure:

1. Open the **/etc/smartd.conf** configuration file;
   a. For each disk, add the line:

```
/dev/<your-disk-ex:sda> -H -I 194 -I 231 -l selftest -f -s S/../../.06
```

   b. For **NovaScale 3045 COMPACT** platforms with SATA disks, the following line should be added.

```
/dev/<your-disk-ex:sda> -d sat -H -I 194 -I 231 -l selftest -f -s
S/../../.06
```

2. Run the **smartctl** command;
   a. For each disk

```
smartctl -s on /dev/<your-disk-ex:sda>
```

   b. For **NovaScale 3045 COMPACT** platforms with SATA disks

```
smartctl -d sat -s on /dev/<your-disk-ex:sda>
```

3. Start the daemon by running:

```
service smartd start
```

Note:
When using **LSI MegaRAID 320-2x HBA**s, use the tools provided on the CD supplied (76679543) for monitoring and managing the SCSI disks and RAID volumes.

2.2.3 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 in this section.

Run the command:

```
system-config-network
```

Note:
If the network has already been configured for the Management Node during the Installation Procedure, go to sections 2.2.3.2 and 2.2.3.3.

2.2.3.1 Administration Network Configuration

Configure the **administration network** (device **eth0**) as follows:

1. In the **Devices** panel select device **eth0**.
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:
   - For a cluster configured with VLAN:

     ```
     IP ADDRESS        XXX.YYY.0.1
                       XXX.YYY.0.13 for a Secondary Management node
     NETMASK           255.255.0.0
     DEFAULT GATEWAY   none
     ```

   - For a cluster configured without VLAN:

     ```
     IP ADDRESS        XXX.1.0.1
                       XXX.1.0.13 for a Secondary Management node
     NETMASK           255.0.0.0
     DEFAULT GATEWAY   none
     ```

Important
The address settings used for the IP addresses should 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.

2.2.3.2 Route Configuration (only for Cluster configured with VLAN)

1. In the **Devices** panel select device **eth0**.
2. Click **Edit**.
3. Select **panel route** -> function **add and set**:

   ```
   IP ADDRESS    XXX.0.0.0
   NETMASK       255.0.0.0
   GATEWAY       XXX.1.255.254
   ```

4. If necessary, after the system installation:
   a. Delete the route using the command below:

```
route delete 169.254.0.0
```

   b. Disable the **zeroconf** route for the system boot by editing the **/etc/sysconfig/network** file, and by adding the following NOZEROCONF value at the end of the file, as shown:

```
NETWORKING=YES
HOSTNAME=localhost.localdomain
NOZEROCONF=yes
```

2.2.3.3 Alias Creation on eth0

1. In the **Devices** panel select device **eth0**.
2. Click **New**.
3. For device type select **Ethernet connection**.
4. Click **Forward**.
5. Select **eth0**.
6. Click **Forward**.
7. Select **statically set IP addresses** and set the following values, according to your cluster type:
   – For a cluster configured with VLAN:

```
IP ADDRESS        XXX.YYY.0.65
NETMASK           255.255.0.0
DEFAULT GATEWAY   none
```

   – For a cluster configured without VLAN:

```
IP ADDRESS        XXX.1.0.65
NETMASK           255.0.0.0
DEFAULT GATEWAY   none
```

The alias is named: eth0:1.

2.2.3.4 Backbone Network Configuration

1. In the Devices panel select device **eth1**.
2. Click **Edit**.
3. Select **activate device** when computer start.

4. Select **statically inet.ip** addresses and set:

```
IP ADDRESS    YYY.ZZZ.0.1
              YYY.ZZZ.0.13 for a Secondary Management node
NETMASK       255.255.0.0
GATEWAY       "customer backbone related"
Host Name Settings
```

5. In the DNS panel set the hostname to: `<clustername>0`

6. Select **Activate**.

Important:
For the new hostname to be recognized, restart X server as follows:

```
logout
login (graphical)
```

2.2.3.5 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.xxx.xxx.xxx/24 #only for internal use
  bind            = xxx.xxx.1.1  #local IP address
}
```

Note:

The configurations above can be adapted and used by other services.

2.2.3.6 Configuration Check

1. Check the **/etc/sysconfig/network-scripts/ifcfg-eth0** file. You should read the following values, according to your cluster type:
   - For a cluster configured with VLAN:

     ```
     DEVICE=eth0
     ONBOOT=yes
     TYPE=Eth
     NETMASK=255.255.0.0
     IPADDR=XXX.1.0.1
     NOZEROCONF=no
     ```

   - For a cluster configured without VLAN:

     ```
     DEVICE=eth0
     ONBOOT=yes
     TYPE=Eth
     NETMASK=255.0.0.0
     IPADDR=XXX.1.0.1
     NOZEROCONF=no
     ```

2. For a large cluster, check also the **/etc/sysconfig/network-scripts/route-eth0** file. It should have the following format:
   `XXX.0.0.0/8 via XXX.1.255.254`

2.2.3.7 Miscellaneous & Verification

1. Edit the **/etc/hosts** file:
2. Change the first line as follows:
   `<localhost IP address> localhost.localdomain localhost`
3. Add the line:
   `<IP address> <clustername>0`
   where the IP address is: `XXX.1.0.1`

2.2.4 External Storage System Installation

The Management Node may be connected to an external storage system, and if so, it should be configured as described in this section.

Note: External RAID storage system is mandatory for the *Management Node High Availability* feature.

2.2.4.1 Installing for the First time - Storage System Installation and Configuration

Important:

Go to the next paragraph if carrying out a cluster migration with the preservation of data.

If an external storage system such as a **FDA** is planned for the storage of application data then this device must be configured first with any appropriate file system configuration changes made for the Management Node before installing additional software.

Please refer to the documentation provided with the storage system for details on how to configure this device. The process is summarized below using **FDA** storage systems as example. Other models of storage systems can be used.

If the *Management Node High Availability* feature is implemented, refer to *HPC BAS4 Administrator's Guide* for guidelines on how to create LUNs (size and number) on the RAID storage system.

	Generic process	FDA storage system
1	Initial storage system setup.	Connect to the management port using Internet Explorer to setup the IP address of the management port.
2	Install the management software.	Get the Storage Manager CD-ROM and the electronic licenses. The Storage Manager server and the **CLI** software may be installed either on Windows (usually a **PAP**) or on the Linux system planned for **FDA** management. The Storage Manager client (**GUI**) may only be installed on Windows.
3	Configure the storage system.	Create **RAID** groups and **LUN** for the file systems required by the applications. Configure **LUN** access from host ports or attached nodes.
4	Restart the Management Node to detect additional LUNs.	
5	Check that the new LUNs have been detected: **cat /proc/scsi/scsi**	
6	Configure the file systems on the Management Node.	

2.2.4.2 Disk Configuration for ClusterDB

If the **ClusterDB** is to be installed on an external storage system, and the *Management Node High Availability* feature is not implemented, the disks on which the ClusterDB data will be installed have to be configured.

If the *Management Node High Availability* feature is implemented, skip this section and go to: 2.2.5 *Other Software Installation*.

Note: Follow the procedure which corresponds to the situation either *First Installation* or *Migration*.

First Installation

1. Stop the **postgresql** service if it has already been started (after **postgres** rpm installation the service is not started):

```
/sbin/service postgresql stop
```

2. Save **/var/lib/pgsql** (after postgres installation, this directory should exist and be the home directory for the postgres user).

```
mv /var/lib/pgsql /tmp
```

3. Create a partition and a file system with a minimum size of 10 GB (for example on **/dev/sdv**):

```
parted
select /dev/sdv1
mkpartfs (partition type) primary
(File system type) ext2
(start)
(end)
```

Note:
The best practice is to create a LUN of the appropriate size in the storage system. If this is done, then there is no need to partition the Linux disk with the parted command.

4. Set FS journalization on the partition created above:

```
mkfs -j /dev/sdv1
```

5. Create FS label (to avoid problems if the name of the disk device changes):

```
e2label /dev/sdv1 /clusterdb_data
```

6. Create the mount directory:

```
mkdir /var/lib/pgsql
```

7. Edit **/etc/fstab** and add the following line:

```
LABEL=/clusterdb_data   /var/lib/pgsql ext3  defaults 0 0
```

8. Mount FS:

```
mount -a
```

9. Change access rights:

```
chmod 700  /var/lib/pgsql
chown postgres.dba /var/lib/pgsql
```

10. Copy **/tmp/pgsql** to **/var/lib/pgsql**:

```
cp -a /tmp/pgsql/* /var/lib/pgsql/
cp /tmp/pgsql/.bash_profile /var/lib/pgsql/
```

Migration

1. Stop **postgresql** service if it has already started (after **postgres** rpm installation the service is not started):

```
/sbin/service postgresql status (stop)
```

2. Save **/var/lib/pgsql** (after postgres installation, this directory should exist and be the home directory for the postgres user).

```
mv /var/lib/pgsql /tmp
```

3. Create the mount directory:

```
mkdir /var/lib/pgsql
```

4. Edit **/etc/fstab** and adding the following line:

```
LABEL=/clusterdb_data   /var/lib/pgsql ext3  defaults 0 0
```

5. Mount the partition that has been created during the first installation of the cluster (see 2.1.1 *Saving the ClusterDB*):

```
mount /dev/sdv
```

6. Remove all files, except **/var/lib/pgsql/backups**:

```
rm -rf /var/lib/pgsql/<file_name>
```

Then un-mount the partition:

```
umount /dev/sdv
```

7. Mount FS:

```
mount -a
```

8. Move data:

```
chmod 700  /var/lib/pgsql
chown postgres.dba /var/lib/pgsql
```

9. Copy **/tmp/pgsql** to **/var/lib/pgsql**:

```
cp -a /tmp/pgsql/* /var/lib/pgsql/
cp /tmp/pgsql/.bash_profile /var/lib/pgsql/
```

2.2.5 Other Software Installation

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/...`

Assuming that **/media/cdrom** is the mountpoint for the CD-ROM.

1. Install the **Quadrics CD** (if the configuration includes Quadrics).
   Mount the CD-ROM, and then run:

```
cd /media/cdrom
./install_quadrics.pl -mgmt
```

   Answer 555 to the questions regarding group and user IDs.
   See Chapter 5, *Quadrics Interconnect Installation*, for more details.

2. Install the **HPC CD**.
   Mount the CD-ROM, and then run:

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

Note:
The HPC CD installation script checks to see if the **Quadrics** RPMs have been installed. If these have not been installed, then the **InfiniBand** RPMs are automatically installed.

3. After the following console message:

```
intel runtime adds a path for LD_LIBRARY_PATH into /etc/profile,
please source it !
```

   run:

```
source /etc/profile
```

4. Install the **Cluster Management CD**.
   Mount the CD-ROM, and then run:

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

   The installer will ask you if you want to install **Torque**.

5. After the following console message:

```
ns commands adds a path into /etc/profile, please source it !
```

run:

```
source /etc/profile
```

6. Install the **Lustre CD** (if Lustre is included in your delivery).
   Mount the CD-ROM, and then run:

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

7. Install the **Bull Cluster Management Data CD** specific to your configuration and containing the **cluster.data** files. This CD-ROM contains the database preload for your configuration which is required when installing for the first time, and also the tools needed to upgrade the **ClusterDB** Database. Contact your Bull representative for more information.
   Mount the CD-ROM and install the RPMs by running the command:

```
rpm -ivh <rpm-file name>
```

8. If your delivery includes an **update CD** of fixes, install it as follows:

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

9. Reboot the system.

2.2.6 Database Configuration

Important:
If the *Management Node High Availability* feature is implemented configure the Database **on the Primary Management Node only.**

Please go to the section corresponding to your case and follow the instructions carefully:

- 2.2.6.1 First Installation - Initialize Cluster Management Database.
- 2.2.6.2 BAS4 Migration with ClusterDB Preservation
- 2.2.6.3 Re-installation of BAS4 with ClusterDB Preservation

2.2.6.1 First Installation - Initialize Cluster Management Database

Note: This paragraph applies only when performing the **first installation** of BAS4.

1. Run the following commands:

```
su - postgres
cd /usr/lib/clustmngt/clusterdb/install
loadClusterdb --basename <clustername> --adnw xxx.0.0.0/255.0.0.0
--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 ipoverinterconnectnetwork

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
```

2.2.6.2 BAS4 Migration with ClusterDB Preservation

Note: This paragraph applies only in the case of a migration to the next version of BAS4. Typical example of migration: from BAS4V4.3 to BAS4V5.1.

1. Retrieve the Cluster DB file previously saved (**clusterdbdata.sav**) – see section 2.1.1 – and copy it to **/var/lib/pgsql/backups**.

Important:
The Cluster database save file must be named **clusterdbdata.sav**. If necessary rename the file so that it is called **clusterdbdata.sav**.

2. Run the commands:

```
su - postgres
cd /usr/lib/clustmngt/clusterdb/install
./upgradeClusterdb
pg_dump -Fc -C -f /var/lib/pgsql/backups/clusterdball.dmp clusterdb
pg_dump -Fc -a -f /var/lib/pgsql/backups/clusterdbdata.dmp clusterdb
```

2.2.6.3 Re-installation of BAS4 with ClusterDB Preservation

Note:

This paragraph applies only to the case of a re-installation of the **existing version** of BAS4.

1. Run the commands:

```
su - postgres
psql -U clusterdb clusterdb
clusterdb=> truncate config_candidate;
clusterdb=> truncate config_status;
clusterdb=> \q
```

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 BAS4 Administrator's Guide*.

2.3 STEP 3: Configuring Equipment and Initializing Tools

Use this step in order to:
- Configure equipment manually in some specific cases
- Configure some management tools
- Configure BMCs on NovaScale 3005 Series.

2.3.1 Configuring Equipment Manually (Small Clusters only)

Except for small configurations, the cluster will have been delivered with a database which is already initialized. All equipment (**Switches**, **Portserver**, **PAP**) will have been recognized and tested from the Management Node using the ClusterDB. Tests will have been done by Manufacturing and the database saved at delivery time. Therefore this database initialisation task should not be required.

So only perform this task on a small cluster, and either during the **first installation** or if **new equipment** has been installed.

Note:

If the *Management Node High Availability* feature is implemented, perform this task on the Primary Management Node only.

1. Collect the mac address of each node, portserver, switch Ethernet Quadrics of the cluster. To collect the information you must:
   - 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.

2. 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**

```
portserver psu1c1 00:40:9d:25:9c:b4
eth_switch eswu1c1 00:11:20:d9:ea:01
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
```

3. Run the command:

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

4. Collect the domain name of each node of the cluster. This information is included in the PAM interface.

5. 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**.

6. For NovaScale 5xxx Series servers only, load the PAP domain name:

```
updateFameDomain <file>
```

<file> is the name of a file that must have been created previously in a directory where it can be easily retrievable. It must be backed up. Its format is defined as follows :
`<node name> <PAP domain identity>`

An example is available: **/usr/lib/clustmngt/clusterdb/install/domfame.exp**

```
valid0 VALID0
valid1 VALID1
valid2 VALID2
valid3 VALID3
valid4 VALID4
```

Note: Respect lower-/upper-case letter: they must be the same than in the PAM.

7. Configure the switches (*see Appendix B*).

2.3.2 Configuring Management Tools Using Database Information (for all Clusters)

Note: This task is mandatory for **all clusters**, and in **all installations** cases (first installation, migration, re-installation).

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

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

2. Configure the tools with the following commands:

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

3. Save the ClusterDB again:

```
su - postgres
pg_dump -Fp -C -f /var/lib/pgsql/clusterdball-xx.dmp clusterdb
```

2.3.3 Configuring BMCs from the Management Node (NovaScale 3005 Series Only)

If the Management Node uses NovaScale 3005 Series servers with **Ethernet Cisco** switches, then the **Baseboard Management Controllers (BMC)** have to be configured at this stage of the installation procedure. The **BMC** manages the interface between system management software and the platform hardware and is used for the monitoring of system parameters such as temperature, cooling fan speeds and the power mode. **BMC**s are integral to the **IPMI** (Intelligent Platform Management Interface) architecture.

The **BMC**s for all the cluster nodes will be configured from the Management Node using the **configure-bmc.pl** script.

2.3.3.1 Pre-requisites

- Each **BMC** has to be connected to the **Ethernet** network and be reachable from the Management Node. The Management Node has to be connected via an Ethernet network that provides a telnet connection for each switch.
- All the switches have to be interconnected and have individual IP addresses.
- The **MAC** address of each **BMC** has to correlate with the corresponding MAC address for the Ethernet interface for the switch. If necessary the *mac-address tables* for the Cisco switches will be updated by Frames issued by the BMCs.
- Gratuitous ARP has to be activated. This will usually be done by manufacturing.
- In order that the BMC configuration script functions correctly, the cluster database has to be properly configured on the Management Node. The ports have to be identified by their number in the **ClusterDB** and which corresponds to the cabling, and not by their name. The format for the number will be dependent on the corresponding Cisco switches.
- The IPMI modules have to be loaded on the nodes.

2.3.3.2 configure-bmc.pl Command and Output

The **BMC**s will be configured by running the script below from the root account:

```
configure-bmc.pl -n <nodelist>
```

This will give output similar to that below. The values for the various settings will be added as the command executes and are taken those that are already defined in the Cluster Database. The command has to complete without any errors.

```
Learning eswu12c0
Can't reach eswu12c0
        Ignoring this switch and connected BMC at ./configure-bmc.pl line 77.
Learning eswu1c0
Calling ipmitool -I lan -P "" -H 10.2.2.74 raw 0x2e 0x23

Calling ipmitool -I lan -P "" -H 10.2.2.74 bmc reset cold
Sent cold reset command to MC
Calling ipmitool -I lan -P "" -H 10.2.2.74 lan set 1 ipaddr 10.2.2.74
Setting LAN IP Address to 10.2.2.74
Calling ipmitool -I lan -P "" -H 10.2.2.74 lan set 1 netmask 255.255.0.0
```

```
Setting LAN Subnet Mask to 255.255.0.0
Calling ipmitool -I lan -P "" -H 10.2.2.74 lan set 1 macaddr 00:00:87:e2:db:38
Setting LAN MAC Address to 00:00:87:e2:db:38
Calling ipmitool -I lan -P "" -H 10.2.2.74 lan set 1 arp respond on
Enabling BMC-generated ARP responses
Calling ipmitool -I lan -P "" -H 10.2.2.74 lan set 1 arp generate on
Enabling BMC-generated Gratuitous ARPs
Calling ipmitool -I lan -P "" -H 10.2.2.74 user set name 2 administrator
Calling ipmitool -I lan -P "" -H 10.2.2.74 user set password 2 administrator
Calling ipmitool -I lan -P "" -H 10.2.2.74 user enable 2
Calling ipmitool -I lan -P "" -H 10.2.2.74 raw 0x06 0x43 0xf1 0x02 0x04 0x00

Calling ipmitool -I lan -P "" -H 10.2.2.74 lan set 1 defgw ipaddr 10.0.255.254
Setting LAN Default Gateway IP to 10.0.255.254
```

In order to check that everything has executed correctly run the following command from the Management Node:

```
nsctrl status <node>

ipmitool -I lanplus –C 0 -U administrator -P administrator -H <@_ip BMC> sol
activate
```

2.3.3.3 configure-bmc.pl Command Options

The following options are available:

-h, --help Displays on-line help

-o <file name>, --out <file name> Once the file has finished running the status for each BMC is saved in a XML format.

-n <node list>, --nodes <node list> List the nodes whose BMCs should be configured. For example: ns10, ns[1-50], ns[80-90, 100]

-s <switch list>, --switches <switch list> Only configure those BMCs which are connected to the switches listed. For example: eswu1c0, eswu1c8

-i <file name>, **--in <file name>** Indicate the name of the XML files generated by a previous attempt at configuring the BMCs. A configuration attempt will only be done for those BMCs whose status read *error*.

2.4 STEP 4: Installing Software on a Compute Node or a Login Node

This step describes how to install the software required to create a Compute Node or a Login Node.

Notes:

- The software installed on a Compute Node will be used by **Ksis** to create a reference image that can be deployed on the other nodes of the cluster (see *STEP 6: Creating and Deploying an Image Using Ksis)*.
- The software installed on a **Login Node** is the same as a Compute Node with the addition of the development environment. **Ksis** can also create and deploy an image of a Login Node.
- For other node types (I/O for example) additional software has to be installed by the administrator.

Choose the installation media:

You can use either the **CD / DVD** media or **NFS** (Network File System) to install the Login Node or a Reference Node. (For CD / DVD installation the node must have a DVD drive and a system console and keyboard). Depending on the choice made, refer to:

- 2.4.1 *Installing a Compute Node or Login Node Using CD/DVD*

or:

- 2.4.2 *Installing a Compute or Login Node Using* NFS.

2.4.1 Installing a Compute Node or Login Node Using CD/DVD

1. Install the **BLAS software**.
   To do this, follow all the steps of the *BLAS General Installation Procedure* in section 2.2.1.2 of this chapter.
   Install either the **High Performance Computing Group for Compute Node profile** or **High Performance Computing Group for Login Node profile** from the Node profile selection screen (see Figure 2-14).

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

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

3. Install the **Intel Compilers** (if required). If your cluster includes a Login Node, the compilers should be installed on this node only.
   Follow the instructions written in the Bull notice supplied with the compiler.

4. Install the **Intel MKL** and **MKLCLUSTER** (if required) libraries, on the Compute Node and Login Node.
   Follow the instructions written in the Bull installation notice supplied with the library.

5. Install the **Quadrics CD (if the configuration includes Quadrics)**.
   Mount the CDROM, and then run:

```
cd /media/cdrom
./install_quadrics.pl -node
```

   See Chapter 7, *Quadrics Interconnect Installation*, for more details.

6. Install the **HPC CD**.
   Mount the CDROM, and then run:

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

Note:
The HPC CD installation script checks to see if the **Quadrics** RPMs have been installed. If these have not been installed, then the **InfiniBand** RPMs are automatically installed.

7. After the following console message:
   `intel runtime adds a path for LD_LIBRARY_PATH into /etc/profile, please source it !`
   run:

```
source /etc/profile
```

8. Install the **Cluster Management CD**.
   Mount the CDROM, and then run:

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

   The installer will ask you if you want to install **Torque**.

9. After the following console message (only on Compute Node):
   `ns commands adds a path into /etc/profile, please source it !`
   run:

```
source /etc/profile
```

10. Install the **Lustre CD (If Lustre is part of your delivery)**.
    Mount the CDROM, and then run:

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

11. If your delivery includes an **update CD** of fixes, install it as follows:

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

12. Reboot the system.

13. Go to 2.4.3 *Disk Health Monitoring Configuration on all Nodes*.

2.4.2 Installing a Compute or Login Node Using NFS

2.4.2.1 Prepare the Installation Server

You need a system with DHCP, TFTP and NFS running. The remote node's mac address should be known and present in the **/etc/dhcpd.conf** file.

1. Install the **preparenfs** rpm.
   From a server, which can be the Management Node, install the **preparenfs** rpm (this will have already been installed on the Management Node by the Cluster Management CD). If necessary, you can find it on the *Cluster Management* CD in the **TOOLS/rpm** directory.
   The **preparenfs** rpm installs the following files:
   - /usr/bin/prepare_nfs_install_bas.sh
   - /usr/bin/change_eth_default.sh
   - /usr/share/doc/preparenfs-X/README.

2. Run **prepare_nfs_install_bas.sh**.
   The **prepare_nfs_install_bas.sh** command will execute a pre-configuration script on your Management Node so that it can be used as an installation server. This script will request the following details. These should be to hand before running it.
   - The IP address of the **NFS** install server.
   - The IP address of the remote node you want to install.
   - The path where the media (ISOs images, CDROM or DVDROM) will be mounted when the Bull Linux Distribution is copied to the hard disk.
   - The media which contains the Linux Distribution.
   - The serial line profile you will use (NovaScale Series 3xx5, 40x0, 5xxx or 6xxx or None).

   If the administrator would like to install additional media (for example HPC or CLM) then the details for this media should also be to hand.
   Stop the **prepare_nfs_install_bas.sh** script with **Ctrl+C** if there is a need to exit. It may be restarted later skipping the steps already completed. If the script is run and the Linux distribution is not installed then the label of the distribution will be requested so that its location can be identified.

Note:

The Management Node automatically mounts the CD or DVD when it is inserted.

The **preparenfs** script automatically unmounts the CDs, DVDs and ISOs.

3. Run the **preparenfs POST** script after the remote node has been installed:
   Once the remote node has been partially installed and booted, run the script **golden_integrate.sh** (this should be from the same server as used for the installation of the **preparenfs rpm** in point 1.) to declare this node as the remote node from which the reference image will be taken.
   This script will request the following details. These should be to hand before running it.
   - The IP address of the **remote node** that has just been installed.

   This script makes the remote node boot on its local disk via the network (in a similar fashion to **KSIS**) and informs the database which node is the remote node.

4. If necessary, run the **change_eth_default.sh** command:
   The reference node will be installed via one of its Ethernet interfaces (eth0 by default). To install it via another interface (eth1 or eth2) the **change_eth_default.sh** command has to be used. This command will ask for:
   - The directory where the Linux distribution has been copied to in **/release**; check that this directory exists in **/tftpboot**.
   - The Ethernet card to be used in the kickstart (for example eth1, eth2 etc).

elilo File Modification

If a node has already been deployed, there may already exist an **elilo** file named **<ip_node_address_in_hexa>.conf** in the **/tftpboot** directory of the install server.
(For example, if the IP address of the node to install is `10.0.0.10`, the **elilo** file name is `0A00000A.conf`.)

In this case, the **preparenfs** script will save this file under **<ip_node_address_in_hexa>.conf.preparenfs.sav** before creating the new one.

The saved file will have to be restored after the installation of the reference node.

About the prepare_nfs_install_bas.sh script

The **prepare_nfs_install_bas.sh** script carries out the following tasks:

- It creates a directory in **/release** and **/tftpboot**. The name of this directory is found in the **.discinfo** file of the MEDIA of the distribution.
- It copies files across the distribution MEDIA (according to the number of media you have specified) as follows
  - For all media it copies all rpm to **/release/<dist-name>**.

  - For the 1st media, it copies the entire directory to **/release/<dist-name>**, it copies the **textmenu** and **elilo.efi** files to **/tftpboot**, the **elilo.conf** file to **/tftpboot/elilo.nfs**, the **initrd.img** and **vmlinuz** files to **/tftpboot/<dist-name>**.

- It asks if other media (for example **HPC, CLM**) should be copied across from a directory in **/release**. The name of this directory is given in the .**discinfo** file of the MEDIA.
- It updates the **/etc/exports** file to create **/release** exportable, and it runs the **exportfs** command.
- It updates the **/etc/xinetd.d/tftp** file to enable **tftp**, and restarts the **xinetd** service.
- It modifies all the NFS kickstart of the **initrd.img** adding the IP address given as the IP address of the install server, and it adds the directory where all the files can be found (**/release/<dist-name>).**
- It copies the **/tftpboot/elilo.nfs** file to **/tftpboot/<ip-reference-in-hexa>.conf** with the IP address of the machine you want to install to.
  - It updates this file by adding the path of **initrd.img** and **vmlinuz** (**/tftpboot/<dist-name>**).
  - It updates this file by adding vnc and serial redirection options.
- It checks if **nfs** and **dhcpd** services are running and it creates a summary at the end of the installation.

2.4.2.2 Install the Compute or Login Node

Reboot the node which will have the software installed onto it using, for example. the **nsctrl** command:

```
nsctrl reset <node_name>
```

Choose “network boot” under EFI.

Follow the Installation procedure described in section 2.2.1.2.
When the Node profile selection screen appears select **High Performance Computing Group for Login Node profile**.

Note:

Conman can be used to start the installation and for this the kickstart has been modified to use serial lines for the installation.

In addition the kickstart has been modified to use **vnc** in graphical mode. It is advisable to be connected via `"vncviewer <node_address>:1"` once the graphical part of the installation begins (a Linux prompt will appear requesting this).
If manual partitioning or custom installation is not used, then the installation will be automatic and no information needs to be entered. At the end the system will reboot.

2.4.2.3 Install the Remaining CD-ROMs

Warning:

If the Reference Node is identified within the database as an installable node, it will be erased by KSIS at the next EFI network boot.

If the CD-ROMs have been copied onto an install server for installation on the Compute or Login Node, then mount the directory as follows:

```
ssh <compute_or_login_node>
mkdir /release
mount -t nfs <install_server_ip>:/release /release
```

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

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

2. Install the **Intel Compilers** (if required). If your cluster includes a Login Node, the compilers should be installed on this node only.
   Follow the instructions written in the Bull notice supplied with the compiler.
   See Chapter 6, *Installing Tools and Applications* for more details.

3. Install the **Intel MKL and MKLCLUSTER** (if required), on the Compute Node and Login Node.
   Follow the instructions written in the Bull installation notice supplied with the library.
   See Chapter 6, *Installing Tools and Applications*, for more details.

4. Install the **Quadrics CD** (if the configuration includes Quadrics):

```
cd /release/<product_directory_name>
./install_quadrics.pl -node
```

Note: Quadrics may ask you to start again the installation. This is normal. Follow the displayed indications.

See Chapter 7, *Quadrics Interconnect Installation*, for more details.

5. Install the **HPC CD**.

```
cd /release/<product_directory_name>
./install.sh
```

Note:
The HPC CD installation script checks to see if the **Quadrics** RPMs have been installed. If these have not been installed, then the **InfiniBand** RPMs are automatically installed.

6. After the following console message (only on a Compute Node) :

   ```
   intel runtime adds a path for LD_LIBRARY_PATH into
   /etc/profile, please source it !
   ```

   run:

```
source /etc/profile
```

7. Install the **Cluster Management CD**

```
cd /release/<product_directory_name>
./install.sh
```

   The installer will ask you if you want to install **Torque**.

8. After the following console message (only on a Compute Node):

   ```
   ns commands adds a path into /etc/profile, please source it !
   ```

   run:

```
source /etc/profile
```

9. Install **Lustre CD** (if Lustre is included in your delivery):

```
cd /release/<product_directory_name>
./install.sh
```

Note:

Once a Compute Node has been installed, a new entry will appear in the **EFI** menu (**Bull Linux AS4**). This is now the default entry. When there is a need to deploy an image onto this node then this entry will have to be deleted, leaving the Network entry as the default.

2.4.2.4 Installing Other Types of Nodes

To install other types of nodes (for example to install a Login node after a Compute Node), you do not need to perform again the preparation of the installation server. You only have to:

1. Copy the **<ipnode-address-in-hexa>.conf** file corresponding to the installed node to **<ip-new-node-address-in-hexa>.conf** file.

2. Check that this node is configured to “boot on network” in EFI menu.

3. Continue installation from 2.4.2.2 *Install the Compute or Login Node*.

2.4.3 Disk Health Monitoring Configuration on all Nodes

By default, the disk health monitoring feature (**smartctl**) is configured for **sda.** But if there are other local disks, then **smartcl** must be manually configured for them.

Complete the following operations:

1. Open the **/etc/smartd.conf** configuration file;
   a. Check that the following line is present for each disk:

```
/dev/<your-disk-ex:sda> -H -I 194 -I 231 -l selftest -f -s S/../../.06
```

      Add the line if not present.

   b. For **NovaScale 3045 COMPACT** platforms with SATA disks, the following line should be present.

```
/dev/<your-disk-ex:sda> -d sat -H -I 194 -I 231 -l selftest -f -s
S/../../.06
```

      Add the line if not present.

2. Run the **smartctl** command;
   a. For each disk

```
smartctl -s on /dev/<your-disk-ex:sda>
```

   b. For **NovaScale 3045 COMPACT** platforms with SATA disks

```
smartctl -d sat -s on /dev/<your-disk-ex:sda>
```

3. Start the daemon by running:

```
service smartd start
```

2.4.4 SJ0812 Monitoring Configuration

Note:

The **Bull SJ0812 JBOD** storage system is only found on older hardware configurations. If SJ0812 is not used on your system, skip this paragraph.

If some nodes are connected to a SJ0812 storage system, the **safte-monitor** service must be enabled. See Chapter 3, Section *Enabling Bull SJ0812 Management* for more information.

2.4.5 MPIO Configuration for I/O Multi-pathing

If multi-pathing is expected to be used after the deployment of the reference image, then the following operations have to be done on the Compute node:

The following lines should be added to the **/etc/modprobe.conf** file:

```
options lpfc lpfc_nodev_tmo=10
options scsi_transport_fc dev_loss_tmo=2
```

The **initrd** file must be rebuilt:

```
mkinitrd -v -f /boot/efi/efi/redhat/initrd-VERSION-KERNEL.img VERSION-KERNEL
```

The multi-path service will have to be launched for each boot sequence and so the following command must be executed:

```
chkconfig --level 235 multipathd on
```

Important:
If this setting is retained when **MPIO** is not in use, then **SCSI** devices will be set to an unavailable state sooner (10 seconds instead of the default 30 seconds), and system will be less fault tolerant to Fibre Channel failures.

Note:
For a FC multipath configuration, the LUN access control on the storage sub-system must be defined accordingly.

2.5 STEP 5: Configuring Administration Software

This step describes how to configure **SSH**, **PDSH**, **ganglia**, **syslog-ng**, **NTP**, **SNMP Server**, **postfix**, **SLURM** and **InfiniBand** on the Management Node(s) and on the Compute or Login Node if necessary.

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 1: Saving the Database and the Configuration Files*. 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 other 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**).

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

Important:

If the *Management Node High Availability* feature is implemented, run the following commands **on the Primary Management node only.**

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 this configuration:

```
ssh localhost uname
```

```
The authenticity of host 'localhost (10.0.0.5)' 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,10.0.0.5' (RSA) to the list of known hosts.
   Linux
```

Then enter:

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

```
   Linux
```

On the Compute, Login Node and Secondary Management Node

1. From the Management Node copy the **/root/.ssh** directory on to the Compute and Login Node.

```
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 (10.0.0.5)' 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 'ns1,10.0.0.5' (RSA) to the list of known hosts.
   Linux
```

Then enter:

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

```
   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 PDSH

Copy the **/etc/genders** file from the Management Node to the Compute and Login Nodes, so that **pdsh** can be used with the node group option on these nodes.

2.5.3 Configuring Ganglia

Important
This task must be performed before deployment.

2.5.3.1 On the Management Node

1. From the Management Node, copy the file:
   /usr/share/doc/ganglia-gmond-3.0.1/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 name of the cluster.
     `name = "xxxxx"  /* replace with your cluster name */`
   - In line 24 replace `x.x.x.x` with the alias IP address of the Management Node.
     `host = x.x.x.x  /* replace with your administration node ip address */`

3. Restart **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 "<clustername>" 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. Restart **gmetad**:

```
service gmetad start
```

2.5.3.2 On the Compute, Login Node and Secondary Management Node

1. Copy the file
   /usr/share/doc/ganglia-gmond-3.0.1/templates/gmond.conf
   into **/etc** for all other nodes apart from the Management Node.

2. Edit the **/etc/gmond.conf** file:
   - In line 18, replace `xxxxx` with the name of the cluster.
     ```
     name = "xxxxx"  /* replace with your cluster name */
     ```
   - In line 24 replace `x.x.x.x` with the alias IP address of the Management Node.
     ```
     host = x.x.x.x  /* replace with your administration node ip
     address */
     ```

3. Restart **gmond** service:

```
service gmond start
```

2.5.4 Configuring Syslog-ng

Warning:

This task must be performed before deployment.

2.5.4.1 Syslog Ports Usage

514 / udp This port is used to log DDN storage system syslog messages. It is intentionally chosen as a non standard port. This value must be consistent with the value defined in **/etc/storageadmin/ddn_admin.conf** and with the current DDN storage system configuration.

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.

2.5.4.2 On the Management Node(s)

1. 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.
   a. Search for all the lines which contain the `SUBSTITUTE` string; for example:

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

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

2. If the cluster includes **DDN storage system**, run the command:

```
cp -p /etc/syslog-ng/logrotate/logrotate_ddn.conf /etc/logrotate.d/ddn
```

3. If the *Management Node High Availability* feature is implemented, run the command:

```
cp -p /etc/syslog-ng/logrotate/logrotate_syslog-ng_admin_ha.conf
/etc/logrotate.d/syslog-ng-ha
```

2.5.4.3 On the Compute and Login Node

Modify the **/etc/syslog-ng/syslog-ng.conf** file, as follows, to add the IP address of the server on which log files are centralized.

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

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

. . .

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

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

2.5.4.4 Restart syslog-ng

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

```
service syslog-ng restart
```

2.5.5 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 synchronize the Management Node with an external time source.

2.5.5.1 On the Management Node(s)

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

The first line should be marked as a comment:

```
#restrict default nomodify notrap noquery
```

The second line should have the following syntax assuming that the IP address is the management network with associated netmask:

```
# Permit all access over management network
restrict <mgt_network_IP_address> mask <mgt_network_mask
nomodify notrap>
```

For example, if the IP address of the Management Node is `10.1.0.1`:

```
restrict 10.1.0.0 mask 255.255.0.0 nomodify notrap
```

Leave the line:

```
restrict 127.0.0.1
```

Put the following lines in as comments:

```
# --- OUR TIMESERVERS -----
#server 0.pool.ntp.org
#server 1.pool.ntp.org
#server 2.pool.ntp.org
```

Leave the other command lines and parameters unmodified, as follows:

```
server  127.127.1.0     # local clock
fudge   127.127.1.0 stratum 10
driftfile /var/lib/ntp/drift
broadcastdelay  0.008
keys            /etc/ntp/keys
```

Restart **ntpd** service:

```
service ntpd restart
```

Start **ntptrace** with 10.0.0.1 as the Management Node IP address:

```
ntptrace 10.0.0.1
```

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

Test installation: run the following command on each node:

```
ntpq -p
```

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

2.5.5.2 On the Compute and Login Nodes

Configure the **/etc/ntp.conf** file on the node as follows.

1. Change the following line to a comment:

   ```
   #restrict default nomodify notrap noquery
   ```
   and add:
   ```
   # Authorize all access over management network
   restrict default ignore
   restrict <mgt_network_IP_address> mask <mgt_network_mask>
   ```

   Example:
   ```
   restrict 10.0.0.0 mask 255.255.0.0
   ```
   should match network addresses defined in Management Node ntp.conf file.

   Leave the line:

   ```
   restrict 127.0.0.1
   ```

2. Change the following lines to comments as shown below:

   ```
   # --- OUR TIMESERVERS -----
   #server 0.pool.ntp.org
   #server 1.pool.ntp.org
   #server 2.pool.ntp.org
   ```

3. Add the management server as the reference (if the *Management Node High Availability* feature is implemented, use the IP alias address):

   ```
   server <mgt_node_IP_address>
   ```

   Example:
   ```
   server 10.0.0.1
   ```

   The "local" configuration should become comment lines, since local backup is not required:
   ```
   #server 127.127.1.0     # local clock
   #fudge  127.127.1.0 stratum 10
   ```

4. Leave the following lines:

   ```
   driftfile /var/lib/ntp/drift
   broadcastdelay  0.008
   ```

5. Put as a comment:
   ```
   #keys /etc/ntp/keys
   ```

6. Add the following lines at the end of the file:
   ```
   tinker panic 0
   tinker stepout 0
   ```

2.5.5.3 Restart NTP

1. Restart NTP on each node (Management Node, Reference Node, Login Node):

```
/etc/init.d/ntpd restart
```

```
Shutting down ntpd:                                        [  OK  ]
Starting ntpd:                                             [  OK  ]
```

2. On the Management Node, start **ntptrace** and check if the Management Node responds (if the *Management Node High Availability* feature is implemented, use the IP alias address):

```
ntptrace 10.0.0.1
```

```
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
```

2.5.6 Configuring the SNMP Server

To know where the SNMP traps must be sent, add these lines to **/etc/snmp/snmpd.conf** on the Compute and Login Nodes.

```
trap2sink <mgmt hostname> public
rocommunity     public
```

Where `mgmt hostname` is the host name of the Management Node, for HA configurations use the virtual address of the Management Node. **snmpd** will later be enabled by Lustre administration tools on the I/O nodes only.

2.5.7 Configuring Postfix

Note: If the *Management node High Availability* feature is implemented, this operation has to be done on both Management Nodes.

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. Restart the postfix service:

```
# service postfix restart
```

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

```
# chkconfig postfix on
```

2.5.8 Configuring SLURM

The SLURM resource manager is mandatory for **InfiniBand** configurations. The **SLURM** files on the **Bull BAS4 Cluster Management CD** are installed under the **/usr** and **/etc** directories.

Note:
Steps 2.5.8.1 and 0 should be carried out from the Management Node. The configuration file is then copied to the other nodes and the files on these nodes are checked (step 2.5.8.3)

2.5.8.1 Configure the SLURM job credential keys

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.

From the Management Node, when you are within the directory where the keys will reside, run the commands below:

```
openssl genrsa -out private.key 1024
openssl rsa -in private.key -pubout -out public.key
```

The path of these keys must be provided as values for the **JobCredentialPrivateKey** and **JobCredentialPublicCertificate** parameters in the **slurm.conf** configuration file. Usually, these keys are located in **/etc/slurm/** directory, as shown below:

Example lines from Slurm.conf

```
SlurmUser=slurm
...
JobCredentialPrivateKey=/etc/slurm/private.key
```

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

The **JobCredentialPrivateKey** 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 **JobCredentialPublicCertificate** 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
```

2.5.8.2 Create and Modify the SLURM configuration file

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

From the Management Node modify the parameters to suit the needs of the cluster.

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

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

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. The path directories must be created on all of the nodes, if they do not already exist. (step 2.5.8.3)

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). (step 2.5.8.3)

```
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. (step 2.5.8.3)

```
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 the 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
```

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

Copy the following files from the Management Node to the Compute and Login Nodes.

- **/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)

Check that the directory used by the SLURM daemon (typically **/var/log/slurm**) exists on the Compute and Login 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 Login Nodes, and the Compute 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).

2.5.8.4 Checking and Starting the SLURM Daemons on Cluster Nodes

Check to see if the **Slurmctld** daemon has started on the Management Node and the **Slurmd** daemon has started on the Compute and Login Nodes 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.

2.5.8.5 Starting the SLURM Daemons on a Single Node

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

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

or

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

Note:

The **startclean** argument will start the daemon on that node without preserving saved state information (all previously running jobs will be purged and node state will be restored to the values specified in the configuration file).

2.5.8.6 More Information

See the BULL HPC BAS4 *Administrator's Guide* for more information on SLURM (security, the creation of job credential keys, the **slurm.conf** file, the **slurm.sh** script and stopping and starting daemons).

2.5.9 Using MPIBull2 with SLURM

Note:
MPI_Bull1 is not supported on **InfiniBand/SLURM** software stacks.

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 may be linked with the **SLURM PMI** library and then launched with the **SLURM** process manager.

All this should be easy for the user.

The **MPIBULL2** RPMs include the **/etc/profile.d/mpibull2*** file which is used to define the default settings for the cluster.

The System Administrator should check the configuration settings on the Login Node, and decomment or add lines to the **/etc/profile.d/mpibull2*** file. Setting examples are provided in the **/etc/profile.d/mpibull2*** file.

For a cluster using **OpenIB InfiniBand** the following line must be found in the **mpibull*** file:

```
mpibull2-devices -d=ibmr_gen2 &> /dev/null
```

and the path must be added for the communication libraries

```
MPIBULL2_PRELIBS="-L/opt/ofed-1.1/lib/ $MPIBULL2_PRELIBS"
```

For a cluster which uses **SLURM** set the following line, and add the path to the **PMI** library, as necessary.

```
MPIBULL2_PRELIBS="-lpmi $MPIBULL2_PRELIBS"
```

When using the **MPI InfiniBand** communication driver, memory locking should 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, and should be set between 1GB and 4GBs.

Notes:

- These initializations, which are necessary to run **MPIBULL2**, should be declared in a file in the **/etc/profile.d** directory. It is essential that this is done in order to ensure that the environment is always available.
- Ensure that the compiler paths are defined. To do this, create a file, for example **compilo.sh**, in the **/etc/profile.d** directory which includes the paths.

  See *Chapter 6* in this manual for more information on the compiler.

compilo.sh example

This defines the environment and path for the compilers. The script has to be installed on the Login nodes in the **/etc/profile.d** directory. For example:

```
. /opt/intel/compilo_9/l_cc_c_9.1.044/bin/iccvars.sh
```

```
. /opt/intel/compilo_9/l_fc_c_9.1.039/bin/ifortvars.sh
```

The path for the **MKL** librairies has to be defined, if they are required.

```
export LAdir=/opt/intel/mkl/8.1/lib/64
```

2.5.10 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 should 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 should 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).

2.5.10.1 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 should be given 0400 permissions and owned by the user that the munged daemon will run as. Securely propagate this file (e.g. via ssh) to all other hosts within the same security realm.

2.5.10.2 Starting the Daemon

On each host within the security realm, invoke the daemon directly (`/usr/sbin/munged`) or 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.

2.5.10.3 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 locally decoded:

```
$ 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`).

2.5.11 Installing InfiniBand Interconnect Software on the NovaScale 3005 Series Platform

Installing InfiniBand rpms on each node

The **InfiniBand** software has to be installed on each node. Run the command below to check to see if the software has been installed:

```
rpm -qa lgrep kernel_ib
```

If the software is present then go to the next step.

If the software is not present run the commands below:

1. Go to the OFED directory on each node using the command below:

```
cd /media/cdrecorder/HPCV7.1/OFED/rpm
```

2. The **InfiniBand** rpms on the BAS4 HPC CD have to be installed in the **OFED** directory on each node by using the command below:

```
rpm -ivh <rpm_file_name>
```

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 a Compute or Login Node installed previously
3. Deployment of this image on cluster nodes.

These operations have to be performed **from the Management Node**. (From the **Primary Management Node** if the *Management Node High Availability* feature is implemented).

Note:

To create and deploy a node image using Ksis, all system files should be on local disks and not on the disk subsystem. To create an I/O node image for example, all disk subsystems should 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 3045)
- Disks (same number, controller, size)
- Network interface.

See:
Refer to the *HPC BAS4 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 **Cluster Management CD**. It uses NovaScale commands and the cluster management database, which are installed from the BAS and cluster management software.

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 **Ksis** is run.

2.6.1.3 Verifying the Ksis Server

In order to deploy an image using **Ksis**, various conditions should 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. The **systemimager** service must be running. If not, run the command:

```
service systemimager start
```

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

Note:

Do not change the EFI boot configuration for the Reference or Login Node because the image should NOT be deployed to these nodes before the deployment of the image to the other nodes has been successfully completed.

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
```

For cluster that use RMS resource manager:

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

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

2.6.2 Creating an Image

Create a reference image of a Compute or Login Node previously installed.

```
ksis create <image_name> <reference_or_login_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
```

This will give an output similar to that shown below:

```
Node             Status  RMS     Image                    Boot Mode
=================================================================
ns0              up      in      manual installation
ns1              up      in      -
ns2              up      in      -
```

2. If the status for any of the nodes is different from **up** then restart the system by running the following command from the root prompt:

```
service nagios restart
```

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

4. Start the deployment by running the command:

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

5. If, for example, 3 compute nodes are listed as `ns[2-4]`, then enter the following command for the deployment:

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

Note:

For the initial installation verify in the EFI menu that there is only one entry which boots from the eth0 network.

2.7 STEP 7: Post Installation Configurations for InfiniBand Clusters

2.7.1 Configuring InfiniBand Interfaces

An **InfiniBand** interface (IP-over-IB) has to be configured for each node. This is named **ifcfg-ib0** when the **InfiniBand** cable is connected to **port0** of the node adapter, and **ifcfg-1b1** when the **InfiniBand** cable is connected to **port1** of the node adapter.

An example is shown below for a node (ns3):

```
# cat /etc/sysconfig/network-scripts/ifcfg-ib0
DEVICE=ib0
ONBOOT=yes
BOOTPROTO=static
NETMASK=255.255.0.0
BROADCAST=17.0.0.255
NETWORK=17.0.0.0
IPADDR=17.0.0.4
```

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.

The config_ipoib script

The script **config_ipoib** is provided to help configure the **InfiniBand** interfaces. The different setting values in the script may have to be updated, depending on the configuration of the cluster.

If this tool is not used, then the **InfiniBand** interface will have to be created on each node manually.

The **config_ipoib** script is launched every time a node is booted. During the boot this script will check to see if the **InfiniBand** interface exists. If the interface is not found then it will be created.

The Compute and Login nodes automatically reboot, following a deployment of a reference image onto them by **Ksis**. This means if a reference image has been deployed onto a node, then the **InfiniBand** interface will have been created on the node by the **config_ipoib** script.

InfiniBand interfaces for the Management Node, Standalone Node or Reference Nodes

For a Management Node, a Standalone Node, or for the Compute/Login Nodes used as Reference Nodes for the creation of the reference image, the System Administrator has the choice of either rebooting the node, or running the **config-ipoib** script manually, to create the **InfiniBand** interfaces on them.

Checking the InfiniBand interfaces

It is recommended that the configuration of the **InfiniBand** interfaces is verified to ensure that all the settings are OK. This is done by running the command:

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

Alternatively, to see the interface settings separately in groups for a set of nodes, use the commands below:

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

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

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

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

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

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

Important

If at some point a **Ksis** reference image is recreated and redeployed onto the nodes, the existing versions of the **/etc/sysconfig/network-scripts/ifcfg-ib0** files must be deleted; otherwise the **config_ipoib** script will not work.

2.7.2 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.

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 saved before being overwritten by the new files. They are renamed **.conf.old**. Thus, the administrators can manually check the differences, and list the customized values when they have changed the default values delivered with the rpms.

For more information about setting up the storage management services, refer to *HPC BAS4 Administrator's Guide*, chapter *Storage Devices Management*.

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

Perform 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 between regular checks of storage device status. 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 the regular time updates. Check 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 then the line above must be commented.

3.2 Enabling Bull SJ0812 Management

The SJ0812 is a SCSI JBOD storage system connected to some cluster nodes and hosting the node's local disks.

Note:

The **Bull SJ0812 JBOD** storage system is only found on older hardware configurations. If SJ0812 is not used on your system, skip this paragraph.

The management is limited to temperature monitoring by the **safte-monitor** daemon.

1. On Management and Compute Nodes, check that the **safte-monitor** service is enabled:

```
chkconfig --list safte-monitor
```

2. If the service has not been activated during the installation, activate it now:

```
chkconfig -level 2345 safte-monitor on
```

Note:

The service is enabled on all nodes, and the daemon will be started at boot time, even if the node does not contain a SJ0812. But if no SJ0812 is discovered, the daemon automatically stops, preventing unnecessary memory and CPU consumption.

The administrators can change the default value for the device polling period (**Period** parameter, defined in seconds, in the **/etc/safte-monitor.conf** file).

The alarm temperature threshold can be customized as follows:

- Edit **/etc/sysconfig/safte-monitor** and change the value for the temperature (`-t <value in Celsius degrees>`).

3.3 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* for the **StoreWay FDA** model, which is being installed and configured.

The management of the 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 information. 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. When the cluster contains **NovaScale 5xx5 Series** models the FDA software must be installed on a **PAP**. Otherwise, the software must be installed on the Management Node. For systems with more than 32 **FDA** disk arrays, the FDA software must be installed on multiple systems, either PAPs or cluster service nodes.

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. For clusters including **NovaScale 5xx5 Series** models, it is recommended to install the GUI on a PAP, with the FDA manager server and CLI components. For clusters without Windows systems, the GUI can be installed on an external Windows management station.

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.
- A 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.

3.3.1 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 **Linux RPMs**.
   Linux RPMs are IA-32 compiled, so on IA-64 systems it may initially be necessary to create a symbolic link from **/emul/ia32-linux/bin/sh** to **/bin/sh** by entering the command:

```
ln -s /bin/sh /emul/ia32-linux/bin/sh
```

   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:

```
# 3fda1300
# An IP address is defined
diskarray1 =(
```

```
ip =(172.116.213.8)
)
# 4fda2500
# Two IP address are defined
diskarray2 =(
ip =(172.116.213.80, 172.116.213.90)
)
```

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 another 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 FDA storage systems. Ssh must be enabled so that FDA management tools operating 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

```
sshkeygen  -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.

5. Connect to the Linux system FDA manager.

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

6. Do not destroy the **~/.ssh/authorized_keys** file. Run:

```
mkdir -p .ssh
cat id_rsa.pub >> .ssh/authorized_keys
rm id_rsa.pub
```

Note:
If necessary, repeat this operation for other pairs of Linux and FDA manager Linux 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.

3.3.2 Installing and Configuring FDA software on a Windows system

For Windows, the **disk_array** table contains the **mgmt_station_id** field which is the foreign key for the **hwmanager** table. This table contains information about the FDA storage manager Windows station.
Install the **ISM** server and the **ISM** Integration Base (CLI) software, which is delivered on the CDs provided with the storage arrays.

Enabling the CLI on the FDA Windows management station

1. Change the name of the file

   C:\Program_Files\FDA\iSMSM_CMD\conf\iSMSM.sample
   to
   C:\Program Files\FDA\iSMSM_CMD\conf\iSMSM.conf.

2. Start the 'Integration Base Service'.

3. Run the CLI to check that it is correctly installed:

```
iSMcmd -host <ip_of_local_windows_station>
```

Note:

This command should be installed under
/cygdrive/c/Program\ Files/FDA/iSMSM_CMD/bin/

Enabling ssh Access from the Management Node on a Windows System

Note:

This part of the process is required for the **FDA** storage system management from the cluster Management Node only. **ssh** is used by the management application to monitor **FDA** storage systems. **ssh** must be enabled in order that the FDA management tools on the cluster Management Node function correctly.

4. Check that **copssh** is installed on the Windows system. If it is not installed launch the **copssh_x.x.x_Installer** software located in the **\ADMIN_FDA** directory of the **ClusterMngt CDROM.** Once the installation has finished check that the **cygwin sshd** service has started.

5. Configure **copssh** to authorize access from the Management Node.

The following configuration phases enable login and remote command execution, without the need of typing in a password. This configuration applies to all Linux users who send **ssh/scp** commands to the Windows system and should be root users.

6. Configure the Windows server side:
   On the Windows system side, log in to the administrator account and enable the **local <administrator> user** for **copssh** using the menu:
   start > all programs > copssh > add a user

7. Distribute the **RSA** keys to enable password-less connections from the cluster Management Node:

   a. Log in as root on the cluster Management Node and generate asymmetric RSA keys.

b. Move to the directory where the RSA keys are stored. Usually, it is '**~/.ssh**'. You should find the **id_rsa** and **id_rsa.pub** files there. The **.pub** file must be appended to the **authorized_keys** file on the Windows system. The **authorized_keys** file defined in the **/etc/sshd_config** file, (by default: **~/.ssh/authorized_keys**) must be used.

c. If no key has been generated, generate a key with the **ssh-keygen** command:

```
sshkeygen -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.

d. The public key to the Windows system should be copied using sshd:

```
scp id_rsa.pub <administrator>@<windowshost>:~
```

<windowhost> can be a host name or an IP address. Substitute **<administrator>** with the current administrator login.

e. Then connect to the Windows system:

```
ssh <administrator>@<windowshost>
```

Do not destroy the file **~/.ssh/authorized_keys**. 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 users and Windows users.

It should now be possible to connect to the system through ssh without entering a password. Try from the root accounts on the Linux system:

```
ssh <administrator>@<windowshost> ls -al
```

If prompted for a password, then the environment is not properly setup. Try again.

Enabling password-less ssh execution for Apache Servers 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.

Enabling GUI Access from the Management Node

GUI access is recommended for configuration and diagnostic purposes. GUI access is not used by monitoring tools, and thus the configuration of graphical access is optional. It only runs on Windows, either on a **PAP** or an external Windows system which must have network access to the **FDA** manager.

1. Install the **FDA iSM** client software that is on the NEC CDROM onto a Windows system.
2. Configure the **iSM** client with the FDA disk arrays you want to manage.
3. Find a user account on the Windows station hosting the FDA client software.
4. Connect to this Windows system and check that the Windows Terminal services are active.
5. Check the connection for a private session:

```
rdesktop <windows_system>
```

6. Start an **iSM** (Storage Manager) client to verify the FDA management capability.

3.3.3 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
```

3.3.4 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.

   If the **iSM** GUI is not available, it is possible to connect to the server via the browser using one of the FDA Ethernet IP addresses, for example, http://<ip_address>. Enter 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
```

3.3.5 Updating the ClusterDB with FDA Storage System information

1. For each FDA system, enter:

```
storregister -u -n <fda_name>
```

   The ClusterDB should have been populated with details of disks, disk serial numbers, WWPN for host ports, and so on.

2. Check that the operation was successful:

```
storstat -d -n <fda_name> -H
```

   If the registration has been successful, the command should display all the information for the disks - manufacturer, model, serial number, and so on.
   If the FDA does not appear in NovaScale Master GUI, run the following command:

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

3.4 Enabling DataDirect Networks (DDN) S2A Storage Systems Management

Note: All these operations are done on the cluster Management Node.

If a DDN storage device is connected the following tasks have to be performed.

3.4.1 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.

3.4.2 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 **/etc/logrotate.d/ddn**. See the **logrotate** man page for more details.

3.4.3 Enabling Management Access for Each DDN

1. List the storage system 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:

```
fda0 |     NEC |       1400 |    UNKNOWN |     |     RACK-B1 |  G
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 /etc/hosts. The conman consoles are automatically generated in **/etc/conman.conf**. Otherwise, refer to the **dbmConfig** command.

3.4.4 Initializing the DDN Storage System

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

3.4.4.1 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>
```

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 **ConMan** to access serial ports and configure IP network information for those architectures which include a PortServer. **ConMan** is not available for **NovaScale 3005 Series** models.
- 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 and 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.

3.4.4.2 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, carefully check the full configuration.

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
```

3.4.5 Updating the ClusterDB

1. Store in the database the exact information concerning each **DDN**:

```
storregister -u -n <ddn_name>
```

This operation has to be run for each **DDN**.
In case of error, run again the command.

2. Check for each **DDN** that the ClusterDB has been populated:

```
storstat -d -n <ddn_name> -H
```

If the registration has been successful, the command should display all the disks, with manufacturer, model, serial number, and so on.
If the DDN does not appear in NovaScale Master GUI, run the following command:

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

3.5 Enabling Brocade Fibre Channel Switches Management

3.5.1 Enabling Access from Management Node

The ClusterDB is preloaded with configuration information for the Brocade switches. Refer to the **fc_switch** table. If it is not the case, information must be entered by the administrator.

Each Brocade switch must be configured with the correct IP address/netmask/gateway, and switch name, login and password, in order to match the information of the ClusterDB.

Please refer to Appendix B for more information about the switch configuration. You can also refer to Brocade's documentation.

3.5.2 Updating the ClusterDB

When the Brocade switches have been initialized, they must be registered in the ClusterDB running the following command from the Management Node (for each switch):

```
fcswregister -n <fibrechannel switch name>
```

3.6 Storage Management Services

The purpose of this phase is to build and distribute on the cluster nodes attached to fibre channel storage systems a data file which contains a human readable description for each **WWPN**. This file is very similar to **/etc/hosts**. It is used by the **lsiocfg** command to display a textual description of each fibre channel port instead of a 16 digit **WWPN**.

1. Build a list of **WWPN**s on the management station:

```
lsiocfg -W > /etc/wwn
```

Note:

This file must be rebuilt if a singlet is changed, or if FC cables are switched, or if new LUNs are created.

2. Distribute the file on all the nodes connected to fibre channel systems (for example all the I/O nodes).

The file can be included in a KSIS patch of the Compute nodes. The drawback is that there are changes to the **WWPN** then a new patch will have to be distributed on all the cluster nodes.

Another option is to copy the **/etc/wwn** file on the target nodes using the **pdcp** command:

```
pdcp -w <target_nodes> /etc/wwn /etc
```

Chapter 4. Configuring the Lustre File System

Note: This chapter only applies if **Lustre** is included in your delivery.

This chapter 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 BAS4 Administrator's Guide*. This document also contains additional information about High Availability for I/O nodes and the ClusterDB.

4.1 Enabling Lustre Management Services on the Management Node

1. In the case of software migration only, restore the Lustre system configuration information:
   - **/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 High-Availability feature is available, the following command will display the HA paired nodes configuration:

```
lustre_migrate nodestat
```

In case of 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 following ones:

```
# lustre_check is called every 15 mn
*/15 * * * * root /usr/sbin/lustre_check >> /var/log/lustre_check.log 2>&1
```

4.2 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 in the case of software migration, if the Lustre configuration and data must be preserved.

4.2.1 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 with the model file to deploy the configuration specified in the model file on a range of storage systems.

The steps required to logically configure the storage systems are fully described in the *HPC BAS4 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 adapter. 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 which 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 <ddn_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.

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.

4.2.2 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 BAS4 *Administrator's Guide* for more information.

4.3 Making the Storage Systems Operational for Lustre

Depending on your configuration go to:

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

or:

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

4.3.1 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. This means this command does not apply to NovaScale 3005 models.

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. Then, the node forces a verification of the storage resources and checks them against the expected LUNs defined in the model file.

Important:

Do not run **stordepmap** if Lustre is running. Read carefully the *Bull HPC Administrator's Guide* before running this command, to fully understand the risks and prerequisites.

```
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:

For the case of 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.

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

```
lustre_investigate check
```

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

When the storage systems attached to a node are configured, the node can be rebooted to discover the new storage resource. That's the easiest way to perform LUN discovery.

Lustre expects persistent naming of storage devices and 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 carefully the *Bull HPC Administrator's Guide* 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:

- Either from a Management Node (or "centralized node"):

```
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:

- Either from a Management Node (or "centralized node"):

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

and after completion:

```
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 completion:

```
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 **other than 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 described in the Administration Guide of the storage subsystem.

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.

This 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 not used with the **–r** option, the administrator has to manage the backup of the **/etc/storageadmin/disknaming.conf** file himself.

When used remotely (**–r**), immediately following 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 as previously used before 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:
Include the **–m** mode option, if this was specified when the **stordiskname** was previously executed. This applies to both High Availability and non High Availability nodes.

Once the **disknaming.conf** file has been copied over, the symbolic links must be created again using the information contained within it. 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.

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

This phase should be done in the following cases:

- 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 which are being used. This file allows you to populate the **lustre_ost** and **lustre_mdt** tables using the script **/usr/lib/lustre/load_storage.sh**.

Important:

Skip this phase in the case of software migration, when the **/etc/lustre** directory has been saved and restored.

Please refer to the *HPC BAS4 Administrator's Guide* for more details about the **storage.conf** file.

4.5 Configuring and Starting Cluster Suite on I/O Nodes

This phase only applies if High-Availability is to be implemented on the cluster. Depending on your configuration, go to:

- 4.5.1 *Cluster with a Management Node*

or:

- 4.5.2 *Cluster without Management Node*

4.5.1 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 cluster I/O nodes.

The **stordepha** allows the distribution of these configuration files:

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

Note:

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

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

The following steps are specific to NovaScale 40xx nodes:

The NovaScale 40xx nodes require a specific configuration process. In order to ensure correct failover, the two nodes of an HA pair are connected by a direct Ethernet cross cable, used for the heartbeat traffic. This dedicated Ethernet interface is not managed in the ClusterDB, and thus must be configured by the administrator after each KSIS installation.

The eth0 interface is used for the IPMI protocol (for example to power off the node), and is connected to the management network of the cluster.

1. Connect to each node of the HA pairs. Look at the **/etc/cluster/cluster.conf** file and search for the first line which contains:

   ```
   <clusternode name="<host_name>_hb" votes="1">
   ```

2. Add a line in **/etc/hosts** to define IP address 10.0.0.1 for that <host_name>_hb.
   If the host_name is the local node, configure the Ethernet interface used for the heartbeat with this IP address.

3. Search for the second line which contains:

   ```
   <clusternode name="<host_name>_hb" votes="1">
   ```

4. Add a line in **/etc/hosts** to define IP address 10.0.0.21 for that <host_name>_hb
   If the host_name is the local node, configure the Ethernet interface used for the heartbeat with this IP address.

5. To avoid problems using two Ethernet interfaces with Lustre services, Lustre has to be locked on the eth0 (interconnect network access) by way of the Lustre lnet layer configuration.
   On each I/O node, update the **/etc/lustre_modprobe.conf** configuration file as follows

   ```
   options lnet networks=elan0,tcp0(eth0)
   ```

The following steps apply to all kinds of I/O nodes, including NovaScale 40xx nodes configured as above:

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 repeat this sequence of actions.

6. Check that the services of the Cluster Suite are started (**ccsd**, **fenced**, **rgmanager**, **cman**), using the following command:

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

   The output should be similar to the following one:

```
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 is not started on a node, the output is similar to the following one:

```
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
```

4.5.2 Cluster without Management Node

For a cluster without Management Node, the cluster and LDAP databases and their associated tools are not available. Instead, there is a zone on the shared storage of the MDS nodes to manage Lustre configuration and status information. This information is recorded in regular ext3 files.

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

Important:

In the case of software migration, if the Lustre configuration and data must be preserved:

- Backup and restore the **/etc/lustre** directory
- Do not erase the ext3 filesystem installed on the shared storage.

I/O nodes pair configuration

The automatic mode (**stordepha** command) is not available for I/O nodes pairs configuration, so each pair of I/O nodes has to be manually configured for HA.

The Cluster Suite configuration file (**/etc/cluster/cluster.conf**) must be customized and installed for each node. Both I/O nodes of the same pair use the same **cluster.conf** file.

For each pair of nodes do as follows:

1. Select the appropriate **cluster.conf** template in the **/etc/storageadmin** directory.
   - For NovaScale 5xxx/6xxx: **cluster.conf.tpl**
   - For NovaScale 40xx: **cluster.conf.tpl2**
2. Copy the template in **/tmp/cluster.conf**.
3. Uncomment the **/var/lustre/status** file system declaration:
   - Lines between `<!-- <resources> -->` and `<!--</resources> -->` must appear as follows:

     ```
     <resources>
          <fs device="DEVICE_SYMB_LINK_NAME" force_unmount="1"
     fstype="ext3" mountpoint="/var/lustre/status"
     name="lustre_status" options=""/>
     </resources>
     ```

   - Line `<!-- <fs ref="lustre_status"/> -->` must appear as follows:

     ```
     <fs ref="lustre_status"/>
     ```

4. Choose a name for each pair of I/O nodes and replace the **HA_CLUSTER_NAME** keyword with this name, which must be different for each pair.

5. Replace each other keyword with the appropriate value (keywords are self-explanatory).

6. Create the **/etc/cluster** directory for both nodes of the HA pair using the following command:

```
mkdir -p /etc/cluster
```

7. Copy the new **cluster.conf** file on both nodes of the HA pair in **/etc/cluster/cluster.conf**.

The following steps are specific to NovaScale 40xx nodes:

If the I/O nodes pair are **NovaScale 40xx** nodes, a specific configuration is necessary for the Cluster Suite heart-beat additional Ethernet interface. This interface cannot be eth0, which is dedicated to **ipmitools**. It must be a point-to-point interface.

1. On both paired nodes, configure and start the ethx interface (ethx can be eth1 or another interface **but not** eth0) with IP addresses, according to the values set in **/etc/cluster/cluster.conf** file for the NODE_NAME_hb and NODE_HA_NAME_hb keywords:

   ```
   <primary node>  IP@10.0.0.1
   <backup node>  IP@10.0.0.2
   ```

2. Declare the additional interfaces into the **/etc/hosts** table so that Cluster Suite can find them. For that, add the lines:

   ```
   # For Cluster Suite Heart_Beat
    10.0.0.1  <primary_node_hostname>_hb
    10.0.0.2  <backup_node_hostname>_hb
   ```

   The "_hb" suffix is used to easily differentiate each Ethernet interface on nodes.

3. To avoid problems using two Ethernet interfaces with Lustre services, Lustre has to be locked on the eth0 (interconnect network access) by the way of the Lustre **lnet** layer configuration.

   On each I/O node, update the **/etc/lustre_modprobe.conf** configuration file as follows:

   ```
   options lnet networks=elan0,tcp0(eth0)
   ```

The following steps apply to all kinds of I/O nodes, including NovaScale 40xx nodes configured as above:

Lustre failover scripts are self-customized from the generic **/usr/sbin/lustre_failover** script by the way of symbolic links. On both nodes of the I/O pair create the links:

```
ln -s /usr/sbin/lustre_failover /usr/sbin/lustre_failover_<primary_node_hostname>
ln -s /usr/sbin/lustre_failover /usr/sbin/lustre_failover_<backup_node_hostname>
```

Configuration of a Central point on the MDS nodes pair

Important:

When undertaking a software migration which involves the preservation of the Lustre configuration and data, recreate the mount point in the **/etc/fstab** file. Then go to *High Availability operations* section below.

Prepare an **ext3** file system on a partition of the shared storage:

1. Create the **/var/lustre/status** directory and add the following line to the **/etc/fstab** file on both MDS nodes:
   ```
   <symbolic_link_name> /var/lustre/status ext3 sync,noauto,data=journal
   0 0
   ```

2. Mount the **/var/lustre/status** device on the primary MDS node.

The I/O nodes connection configuration and migration status is described by the **lustre_io_nodes** template provided in the **/etc/lustre** directory. This file has to be manually initialized by the administrator with the information related to the paired connections. It is updated by the Lustre failover scripts when node migrations occur. It must contain one descriptor by node. Both paired nodes have their own descriptor:

```
NODE_NAME=<node_hostname>
NODE_HA_NAME=<paired_node_hostname>
LUSTRE_STATUS=OK or MIGRATED (must be initialized to OK)
```

To initialize the **lustre_io_nodes** template, do as follows :

3. Copy the template from **/etc/lustre** directory to the **/var/lustre/status** directory mounted on the MDS primary node.

4. Initialize the node descriptors using your favourite editor.

The Lustre management configuration file (**/etc/lustre/lustre.cfg**) provides the environment variables for nodes to connect to the *central point* of Lustre management. This file has to be dispatched over all the I/O nodes.

5. Update the **/etc/lustre/lustre.cfg** file contents so that you can run Lustre without a management station:

   CLUSTERDB=no (sets the mode "Dbless" for the Lustre management tools).

   `LUSTRE_ADMIN_NODE=<primary_MDS_hostname>` (gives the central management point address).

   `LUSTRE_ADMIN_NODE2=<backup_MDS_hostname>` (gives the backup management address).

   `LUSTRE_DEBUG=yes` (turns on the login mode for Lustre failover scripts).

6. Dispatch the file on all the I/O nodes using the **pdcp** utility.

High Availability operations

These operations are conducted from the *central point* of management, identified by the **/var/lustre/status** mountpoint.

Cluster Suite activation is performed on each node using the **storioha –c start** command. It can be parallelized from the central point of management using the **pdsh** utility.

The Lustre failover services are started using the **lustre_migrate hastart** command.

Lustre file systems are now managed in a standard way using only the **lustre_util** and **lustre_migrate** utilities.

4.6 Configuring Lustre File System (with a Management Node)

1. Change the Lustre user password.

   The **lustre_mgmt rpm** creates the user « lustre » on the Management node with « lustre » as the password. It is strongly advised to change this password, running the following from the root command line on both Primary and Secondary Management node for High Availability systems.

```
passwd lustre
```

   The **lustre** user is allowed to perform the most common operations on Lustre filesystems 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 **/etc/lustre/lustre.cfg**.

   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>/).

   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** starts run the command below:

```
service ldap start
```

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

3. Check database consistency.

```
lustre_investigate check
```

   This command checks which storage devices of **lustre_ost** and **lustre_mdt** tables can be used. A clean output means the command succeeds.

   Refer to the **lustre_investigate** man page or the *HPC BAS4 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**'.

4. Set **rank** field of the file systems installed previously.

Note:
This only applies when migrating from **BAS4V4.2** to **BAS4V4.3**

If filesystems have been already installed and if the release of the cluster database was prior to 20.1, run the following command for each filesystem:

```
lustre_util set_ost_rank -f /etc/lustre/conf/<fs_name>.xml
```

This will set the **rank** field in **lustre_ost** table regarding contents of the XML file.

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 BAS4 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 Lustre OST/MDT services migration.

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 be stopped only when all Lustre file systems are 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 File system configuration.

   The **/etc/lustre/models/fs1.lmf** file is a default model file which comes with Lustre RPMs. It describes 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 Administrator Guide or to the **lustre_util** man page for more details about Lustre model files.

   Run the following command:

```
lustre_util info -f /etc/lustre/models/fs1.lmf
```

   This command prints information about **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 or if there are any problems.

9. Install the file system.

Important:
Do not perform this step 9 in the case of software migration, if 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 since it formats the underlying file system (about 15 mn 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  : formating
   Status      : offline
   Mounted     : 0 times
```

Wait until the following appears:

```
Filesystem fs1:
   Cfg status  : installed
   Status      : offline
   Mounted     : 0 times
```

The last printed line of the execution terminal must be:

```
Filesystem fs1 SUCCESSFULLY installed
```

10. Enable the file system.

Run the following command:

```
lustre_util start -f fs1 -V
```

This operation is quite long (about 10 mn for a 1TB filesystem). Do not use the **-V** option if you want a less verbose output.

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 printed lines of 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 printed line of the execution terminal must be:

```
Mounting filesystem fs1 succeeds on ns[0,2]
```

The file system is now available. As administrator it will be possible to create user's 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. Installation of a Standalone Configuration

This chapter describes the installation of standalone configuration which consists of a single node.

5.1 Bull Linux AS4 Installation

Note:

It is recommended to read carefully all the procedure details before starting the installation.

Start with the following operations:

- Power up the machine.
- Switch on the monitor.
- Insert the DVD of *Bull Linux Advanced Server 4* into the drive.

Note:

This operation must be done during the initial phases of the internal tests (while the screen is displaying either the logo or the diagnostic messages). If the DVD is not inserted during these phases, then insert it and, under EFI, execute the following command:
map -r

Then it is possible to login as root/root or linux/linux user and configure the mouse with the **system-config-mouse** tool and the network with the **system-config-network** tool.

During installation, the default password for the root user is set to "**root**".

5.1.1 Installation Procedure

Install the **BLAS software**:

To do this, follow all the steps of the *2.2.1.2* (BLAS General Installation Procedure) in Chapter 2.

Follow the procedure until step 10 "Node Profile Installation". At this point de-select the **High Performance Computing Group for COMPUTE Node profile**, and select the **High Performance Computing Group for STANDALONE Node profile**. The package installation will begin, and the system will reboot automatically once the installation has completed.

5.2 Disk Health Monitoring Configuration

By default, the disk health monitoring feature (**smartctl**) is configured for **sda**, **sdb** and **sdc** on the Management Node. But if you have other local disks you must manually configure **smartcl** for these disks.

Note: Do not configure **smartcl** on RAID HBA devices or external RAID storage systems.

Use the following procedure:

1. Open the **/etc/smartd.conf** configuration file;
   a. For each disk, add the line:

```
/dev/<your-disk-ex:sda> -H -I 194 -I 231 -l selftest -f -s S/../../.06
```

   b. For **NovaScale 3045 COMPACT** platforms with SATA disks, the following line should be present.

```
/dev/<your-disk-ex:sda> -d sat -H -I 194 -I 231 -l selftest -f -s
S/../../.06
```

2. Run the **smartctl** command;
   a. For each disk

```
smartctl -s on /dev/<your-disk-ex:sda>
```

   b. For **NovaScale 3045 COMPACT** platforms with SATA disks

```
smartctl -d sat -s on /dev/<your-disk-ex:sda>
```

3. Start the daemon by running:

```
service smartd start
```

5.3 Other Software Installation

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

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

Assuming that **/media/cdrom** is the mount point for the CD-ROM.

1. Install the **Intel Compilers** (if required). If your cluster includes a Login Node, the compilers should be installed on this node only.
   Follow the instructions written in the Bull notice supplied with the compiler.
   See Chapter 6, *Installing Tools and Applications* for more details.

2. Install the **HPC** CD.
   Mount the CD-ROM, and then run:

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

3. After the following console message :

```
intel runtime adds a path for LD_LIBRARY_PATH into /etc/profile, please
source it !
```

   run:

```
source /etc/profile
```

4. Install the **Cluster Management CD**.
   Mount the CD-ROM, and then run:

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

   The installer will ask you if you want to install **Torque**.

Note:
See the *Installing Tools and Applications* chapter in this manual for information on installing compilers.

5. Reboot the system.

Chapter 6. Installing Tools and Applications

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

The following topics are described:

- 6.1 *Intel Products*
- 6.2 *TOTALVIEWTM Debugger*
- 6.3 *TORQUE*
- 6.4 *Configuring Modules on the Login Node*

6.1 Intel Products

6.1.1 Intel Libraries Delivered

Some applications delivered with the Bull HPC CD-ROM have been compiled with Intel compilers. The Bull HPC CD-ROM installs the **intelruntime-cc_fc** rpm, which contains various free distribution Intel libraries that are needed for these applications to work, on all node types (Management, Compute, Login and Standalone). These libraries are installed in the **/opt/intelruntime/** folder.

If there is a desire to install a different version of an **Intel** compiler from that which is provided, then this has to be copied on to the other nodes, in order to ensure coherency.

Note:

The **Intel** compilers should only be installed if you wish to compile.

6.1.2 Fortran Compiler

Installation

Follow the instructions of the Bull notice, which is supplied with the **Intel** compiler provided by Bull.

6.1.3 C/C++ Compiler

Installation

Follow the instructions of the Bull notice, which is supplied with the **Intel** compiler provided by Bull.

6.1.4 Intel Debugger

The package used to install the Intel debugger is located in either the Fortran or C tar archive.

Installation

Follow the instructions of the Bull notice, which is supplied with the **Intel** debugger provided by Bull.

6.1.5 Intel Math Kernel Library (MKL)

The MKL and MKLCLUSTER libraries must be installed on Compute and Login Nodes.

Installation

Follow the instructions of the Bull notice, which is supplied with the **Intel MKL / MKLCLUSTER** library provided by Bull.

6.1.6 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

6.2 TOTALVIEW™ Debugger

The **Totalview** proprietary packages are delivered by **Etnus**. They are also available from the Etnus web site: http://www.etnus.com

Unpack the packages and install them:

```
mkdir /tmp/totalview
cd /tmp/totalview
tar xvf  totalview<rel.nb>linux-ia64.tar
tar xvf totalview<rel.nb>doc.tar
cd totalview.<rel.nb>
./Install
rm -rf /tmp/totalview
```

It is recommended that the Totalview is installed in the **/opt/totalview<rel.nb>** directory.

License details will be provided.
In the directory **/opt/totalview<rel.nb>/toolworks/flexlm<relf.nb>**, create the **license.src** file which includes this information and run the script:

```
. /opt/ /totalview<rel.nb>/toolworks/flexlm<relf.nb>/bin/Configure_License
```

This script creates several files, including the **license.dat** file which is the file needed by **Flexlm** to manage the license.

To set the environment, create the **/opt/totalview<rel.nb>/totalview-vars.sh** file including:

```
PATH=/opt/totalview<rel.nb>/totalview/bin/:$PATH
LM.LICENSE_FILE=/opt/totalview<rel.nb>/toolworks/flexlm<relf.nb>/license.dat
```

The complete installation procedure is available in the *Totalview Installation Guide*.

<rel.nb> and **<relf.nb>** represent the release numbers of **Totalview** and **Flexlm** respectively.

For more information see Chapter *7 Debugging Tools* in the HPC BAS4 *User's Guide*.

6.3 TORQUE

Torque is installed from the Cluster Management CD, as described in the *Installation Process* chapter in this manual.

Optional RPMs

Not all the Torque **RPM**s are installed (for example the documentation) and the customer can decide whether or not to install them. These RPMs are available in the **TORQUE/rpm** directory of the Cluster Management CD.

The following optional Torque RPMs can be installed on a Management or Login node:

torque-devel-<release_number>.rpm
torque-docs-<release_number>.rpm
torque-gui-<release_number>.rpm

6.4 Configuring Modules on the Login Node

The **Modules** package allows the environment of the user via module files to be modified dynamically. Once **modules-oscar** is installed on the login node (available on Linux Distribution CDs), you can then run the **/opt/modules/modules-oscar-bull-post-conf.sh** script, which automatically creates modules files for Intel **cc**, **fc** and **mkl**.

Important:
Please verify in the files created by this script that the **cc**, **fc** and **mkl** versions are as expected.

Chapter 7. Installing and Configuring Quadrics Interconnects

This chapter describes how to configure a **Quadrics** interconnect, install the Licenses for it, and verify the nodes and network.

The following topics are described:

7.1 Switch Naming Convention

All switches are connected to a management network by an individual switch controller card. The switch controller has a name that identifies its position in the network. The naming scheme for the switch modules is as follows:

`name = QR<railNumber>[N|T]<switchNumber>`

where:

railNumber	Indicates the rail number: 0 for single rail systems; 0 or 1 for dual rail systems; and so on
N|T	Indicates the type of switch: N for node switch; T for top switch
switchNumber	Is the switch number: the numbering of both node and top switches starts at 0

Examples:

```
QR0N00
QR1N00
```

7.2 Setting-up a Quadrics Interconnect

In the installation process described in this section, the following is assumed:

Cluster name	**ns** (where ns is the base name).
Nodes to configure	Management Node (**ns0**), reference node (**nsX**), which can be a compute node or an I/O node.
IP addresses range	**A.B.C. <n+1>**, with **n** ranging from **0** to number of nodes.

The Quadrics interconnect relies on **QsNet^{II}** (Elan4).

7.2.1 Configuring Hardware

The following procedure explains how to add new **Quadrics** interconnect equipment to the cluster configuration.

Note:
Take all the necessary precautions for this hardware installation step. For furthers details, refer to the **Quadrics** documentation (*QM50x Installation Manual*, *QM-S64* or *QM-S8 Installation Manual*, and *QsNet^{II} Installation and Diagnostics Manual*).

Procedure

1. Install and turn on the **Quadrics** switch (**QM-S64** - 64 ports max. or **QM-S8** - 8 ports).

2. Stop all the nodes within the cluster and unplug the power leads.

3. Plug in one **QM-500 Quadrics** card for each node into the appropriate PCI slot (see details in *Appendix C*).

4. Connect each card to the switch using the cables provided, and label them accordingly. For example:
   ns0 on port 0
   ns1 on port 1
   …
   ns<n> on port **<n>**

After the operating system has been deployed the hardware configuration of **QsNet^{II}** network has to be checked to see if it is operational on each node. Do as follows:

5. Turn on all nodes in the same order as their names (ns1, then ns2, and so on).
   For each connected node, a red and a green led should be lit on the switch side, and a red led should be lit on the card side. After OS boot and Quadrics modules load, the led should become green on both sides.

6. For each node, check that the card is detected on the PCI bus using the **lspci** command. The output should be similar to the following:

   ```
   11:01.0 Network controller: Quadrics Ltd QsNetII Elan4 Network
   Adapter (rev 01).
   ```

The switch uses a **QM-503** Quadrics card (two in the case of redundant **QM-S64**). For switch management purposes you have to connect the administration network to each Quadrics card to get information about the switch status:

7. Plug a keyboard and a screen on the module to get the login prompt.

8. Enter default login/password to gain access to the QNX embedded system:

   login: `quadrics`
   password: `system`

   For security reasons change this default setting as soon as the cluster is in an exploitable state (as soon as users can log on).

9. Configure switch network settings as follows:

   Quadrics Switch Control – (QR0N00)

   1. Show network settings
   2. Change network settings
   3. Run **jtest**
   4. Set module mode
   5. Firmware upgrade
   6. Quit
   7. Reboot
   8. Access Settings
   9. Self Test

   Enter 1,2,3,4,5,6,7,8 and press return: 2

   Follow the instructions in *Quadrics QsNet*^{II} *Installation and Diagnostics Manual* to configure the network interface.

10. Plug in the Ethernet cable to the **QM-503** module and then to an administration network switch (Ethernet).

11. Check the switch configuration using a ping from Management Node and if it is in working state, remove the keyboard and screen from the **QM-503**.

Notes:

- In order to maximize throughput and performance, each Quadrics card should be inserted into a high speed PCI-X slot, with a bus frequency greater or equal to 133MHz. If another card (network, SCSI adapter, etc.) is present for the same PCI bus (shared controller), a penalty performance will be incurred. It is therefore advisable to install each Quadrics adapter with its own PCI controller.
- The **QM-500 card** remaps 256 MBs of PCI memory, and in some configurations, this may prevent the system from booting. In this case, the card should be relocated into another slot, with a bus frequency greater or equal to 133MHz. If this problem occurs, please check the BIOS version of your NovaScale system and ask Bull support if a new version is available. Also check in the BIOS to see if the '**PCI->PCI Gap above 4GB**' option is enabled.
- Quadrics switches do not support **DHCP** (Dynamic Host Configuration Protocol) - **dynamic Ips -** for now. So the IP address has to be set up statically or the **bootp** command used.

7.3 Installing Quadrics Software Packages

The **Quadrics** packages are first installed using the **Quadrics** CD-ROM and then deployed by means of a reference image as described in Chapter 2.

License management and verifications should be done later. This is described in sections *7.3.1 License Management*, *7.3.2 Verifying each Node Installed* and *7.3.3 Verifying the Quadrics Network Status*.

In some cases, it might be useful to use a different version of the **qsnet2libs** package (cluster bringup library). See *7.3.4 Using alternative qsnet2 libraries*.

7.3.1 License Management

1. To obtain the licenses from **Quadrics,** the host ID must be sent to Quadrics. Enter the following command to obtain the host ID:

```
/usr/lib/rms/flexlm/bin/lmhostid
```

   The output is similar to that below:

```
lmhostid - Copyright (c) 1989-2003 by Macrovision Corporation.
Allrights reserved.
The FLEXlm host ID of this machine is "0007e993fc4c"
```

   Send this host ID to Quadrics.

2. When you have received the license from Quadrics, copy **rms.lic** into the **/usr/lib/rms/flexlm/** folder:

```
cp rms.lic /usr/lib/rms/flexlm/
```

3. You can choose to run **RMS** via a global license manager. In this case, the following command must be run:

```
service qslmgrd restart
```

4. Restart the **RMS** service:

```
service rms restart
```

5. To display information about the license, enter:

```
# rinfo -LM
```

```
RMS License Management Information

Feature         Tokens   Expires
RMS/partition   demo     18-apr-2007

FLEXlm license manager is not running
```

7.3.2 Verifying each Node Installed

1. Check the status of interconnect modules:

```
service qsnet status
```

```
modules loaded    : qsnet elan elan3 elan4 rms ep
modules not loaded: eip

elan4 device 0: NodeId=0 Rev=<unknown> Build=B1 Serial=K8A2CB1BFRH916
elan4 device 1: NodeId=<unknown> Rev=<unknown> Build=B1
Serial=K8A2CB1BFRH922
elan4 device 2: NodeId=<unknown> Rev=<unknown> Build=B1
Serial=K8A2CB1BFRH881

ep       : MachineId=0x2400
ep rail 0: Device=elan4 NodeId=0 NumNodes=1024 NodeSet=[0-24,26-31]
ep rail 1: Device=elan4 NodeSet=<not running>
ep rail 2: Device=elan4 NodeSet=<not running>
eip interface 0: down
default library: elan4
```

2. Check the status of RMS service:

```
service rms status
```

```
rms module:  loaded
running: rmsmhd rmsd pmanager tlogmgr eventmgr mmanager  swmgr
stopped:  swmserver                                          [  OK  ]
```

3. Check the status of mSQL database (only on the Management Node):

```
service msqld status
```

```
msql3d  (pid 10843) is running...
```

Note:

If you install **Quadrics** nodes using deployment software you have to manually create each node entry in the **RMS** database. For this, run the following command on the Management Node with the corresponding hostname after each node deployment:

```
# rcontrol create node nsX
```

7.3.3 Verifying the Quadrics Network Status

The **qsctrl** command checks the **Qsnet** network and issues a report according to the database that was initialized previously by means of the **qsdiagadm** command. To make sure that the switches are correctly configured it is recommended to run the **qsctrl –f** command, which enables the system to fill out the **IP** and **MAC** address entries in the database. Below is the series of commands to run:

```
# qsdiagadm -c <type> -r <rails_nb> . . .
# qsctrl -f
# qsctrl
```

Note:

See *7.4 qsdiagadm Command* for details about the flags that must be specified for **qsdiagadm**.

Example:

In the following example, the **rmsquery** is run followed by **qsctrl –f.** Initially the output shows that the switches have not been identified:

```
# rmsquery
```

```
sql> verbose
sql> select name,macaddr from switch_modules
name    macaddr
--------------------------
QR0N00
QR1N00
sql> quit
```

Now run:

```
# qsctrl -f
```

```
qsctrl: switch network manager must be halted, continue [yes/no]: y
qsctrl: addresses updated with 0 errors
qsctrl: switch network manager should now be restarted
```

```
# rcontrol restart server swmgr
```

```
swmgr: ns0
```

```
# rmsquery
```

```
sql> select name,macaddr from switch_modules
QR0N00  00:40:53:0a:25:45
QR1N00  00:40:53:0a:25:4e
sql> quit
```

The output shows that the switches are now identified.

7.3.4 Using alternative qsnet2 libraries

The **qsnet2libs** package provides the user level libraries for using the QsNet^{II} hardware. It includes high level interfaces such as **TPORT**s and **SHMEM** as well as lower level interfaces for direct access to the QsNet^{II} hardware. This package is only required on machines with QsNet^{II} (Elan4) hardware. It is automatically installed on these machines.

In some cases, it might be useful to use a version of the **qsnet2libs** package different from the standard version. This version, named **cluster bringup qsnet2libs** library, is available in the **BONUS** directory of the **QUADRICS** CD.

This section describes how to install and use this library.

Notes:

- The following example shows how to install the alternative library in a file system named `/fs1`. This file system must be accessible on all machines; it may be, for example, an NFS mount point.
- The version numbers of the RPM used in this example may be different for your system.

1. Copy the libraries to /fs1:

```
cp /media/cdrom/BONUS/QSNET2LIBS.2.2.11-3.BULL.tar.gz /fs1
```

2. Enter the /fs1 directory:

```
cd /fs1
```

3. Unpack the archive:

```
tar -xzf QSNET2LIBS.2.2.11-3.BULL.tar.gz
```

4. Enter the rpm directory:

```
cd QSNET2LIBS.2.2.11-3.BULL/RPMS/ia64/
```

5. Convert the RPM package into a CPIO archive:

```
rpm2cpio < qsnet2libs-2.2.11-3.BULL.ia64.rpm > qsnet2libs-2.2.11-3.BULL.ia64.cpio
```

6. Unpack the CPIO archive

```
cpio -idmvu < qsnet2libs-2.2.11-3.BULL.ia64.cpio
```

```
./usr/bin/elan4perf
./usr/bin/elan4probe
[ ... ]
./usr/share/man/man1/padb.1
56024 blocks
```

At this stage, the content of the RPM package is unpacked into the local usr directory. These libraries can be used as follows:

```
export LD_LIBRARY_PATH=/fs1/QSNET2LIBS.2.2.11-
3.BULL/RPMS/ia64/usr/lib/qsnet/elan4/lib:$LD_LIBRARY_PATH
prun <args> job ...
```

The **$PATH** variable may also be set to point to the alternative **qsnet2libs** binaries as follows:

```
export PATH=/fs1/QSNET2LIBS.2.2.11-3.BULL/RPMS/ia64/usr/bin:$PATH
```

Use **nm** to make sure that the new libraries are used by the system:

```
ldd `which tping`
```

```
linux-gate.so.1 =>  (0xa000000000000000)
libelan.so.1 => /fs1/QSNET2LIBS.2.2.11-
3.BULL/RPMS/ia64/usr/lib/qsnet/elan4/lib/libelan.so.1
(0x2000000000050000)
libc.so.6.1 => /lib/tls/libc.so.6.1 (0x2000000000150000)
libelanctrl.so.2 => /usr/lib/libelanctrl.so.2
(0x20000000003c0000)
librmscall.so.1 => /usr/lib/librmscall.so.1
(0x20000000003e0000)
libelan4.so.1 => /fs1/QSNET2LIBS.2.2.11-
3.BULL/RPMS/ia64/usr/lib/qsnet/elan4/lib/libelan4.so.1
(0x2000000000400000)
/lib/ld-linux-ia64.so.2 (0x2000000000000000)
```

7.4 qsdiagadm Command

The **qsdiagadm** command is used to create the diagnostics database. It creates tables in the RMS database which describe the status of the network.

Usage:

qsdiagadm [-fhv] [-c type] [-d database] [-m name] [-n ns] [-r rails] [-t ts]

Flags:

-b opt	Network bandwidth option (default: `full`)
-c type	Create a diags database for a given network type
-d database	Specify an alternate diags database
-f	Force creation of a new network description
-m name	Specify the machine name (cluster name)
-n ns	Number of node switches per rail (defaults to max)
-r rails	Number of rails
-t ts	Number of top switches per rail (defaults to max)
-v	Verbose

The default diags database is **qsdiagshost:qsnet**. This refers to a database called **qsnet** on the **qsdiagshost** server.

Standalone network types are: **QMS8**, **QMS16**, **QMS32**, **QMS64** and **QMS128**.
Federated network types are: **QMF256** and **QMF1024**.
Federated network bandwidth options are: **full (default)**, **half** and **reduced**.

Half and reduced bandwidth networks have half the maximum number of top switches. Reduced bandwidth networks also have half the maximum number of **QM502** cards in their node switches.

Machine names are assumed to be a short character string. Node names are constructed from the machine name and the **ELAN id** of the node. Sets of nodes are specified using a range string. For example: **m**, **m0** and **m[0-1023]**.

Example:

This example shows the flags to set according to the following configuration. The **–m** flag is used to specify the name of the cluster.

- A half bandwidth network => `-b half`
- A federated network 1024 node => `-c QMF1024`
- 11 node switches per rail => `-n 11`
- 4 top switches per rail => `-t 4`
- 3 rails => `-r 3`

```
qsdiagadm -b half -c QMF1024 -n 11 -t 4 -r 3 -m <cluster name>
```

7.5 qsctrl Command

qsctrl is used to check, show and modify the **Qsnet** network.

Usage:

qsctrl [-cfhlsv] [-b route] [-d db] [-D name] [-i link|hosts] [-o link|hosts] [-r rail] [-p on|off]

qsctrl [-S param value] [-G param]

Flags:

-a	List barcodes for switch cards
-b route	Set the broadcast tree on the selected rail. It is highly recommended to not use this flag, which should be reserved for Bull support usage.
-c	Clear link errors on the embedded controllers
-D name	Describe a named component
-d db	Select a database
-e	Show environmental data from all switch modules
-f	Fill out IP and MAC address entries in database
-G param	Get firmware parameters from the embedded controllers
-h	Print help
-i	Configure in a link or a list of hosts. This flag should be used carefully and only by skilled administrators.
-l	Long format descriptions
-n	Just testing, don't make changes
-o	Configure out a link or a list of hosts. This flag should be used carefully and only by skilled administrators.
-p on|off	Power ON/OFF switch cards (default: ON)
-r rail	Select rail (default 0)
-s	Show database state, configured out links and broadcast tree
-S param value	Set firmware parameters on the embedded controllers
-u	Upgrade the firmware on the switch modules
-v	Verbose operation

With no option, **qsctrl** checks the network against the definition in the database and displays a report.

Important:

It is highly recommended **NOT** to use the **-b** flag, which should be reserved for Bull support usage.
The **-i** and **–o** flags should be used carefully and only by skilled administrators.

Available parameters

The following parameters can be specified with the **–G** and **–S** flags:

swmserver.auto-reset	swmserver.disable-vc1	network.ip
swmserver.resetlink-interval	swmserver.ignore-unconnected	network.name
swmserver.linkport-update	swmserver.power-control-state	network.netmask
swmserver.check-gctrl	swc_failover_enabled	network.protocol
swmserver.top-switch	swmserver.update-elan-errors	network.tftpserver_ip
swmserver.poll-interval	swmserver.update-stats	network.upgrade_filename
swmserver.resend-interval	swmserver.linkstate-eeprom	network.v_mac
swmserver.error-report-threshold	swmserver.auto-cauterize	watchdog.last_error
swmserver.error-free-threshold	swmserver.module-type-override	watchdog.processlist
swmserver.ref_clk_freq	network.gateway	watchdog.reboot_count
swmserver.phase-delay		watchdog.swmserver
		upgrade.version

Examples:

- The **-a** flag reports Quadrics identifications for RMA using the syntax:

  `switch: module serial_number type revision`

```
# qsctrl -a
```

```
QR0N00:00  K67B6B2BFQ3211  QM501_CA B2
QR0N00:01  K67B6B2BFQ3213  QM501_CA B2
QR0N00:02  K67B6B2BFQ3204  QM501_CA B2
QR0N00:03  K67B6B2BFQ3210  QM501_CA B2
QR0N00:04  K9825A1BFRV515  QM502_CA A1
QR0N00:05  K9825A1BFRV543  QM502_CA A1
...
```

- The command without arguments reports the Quadrics network status:

```
# qsctrl
```

```
qsctrl: passed power control check (on)
qsctrl: passed population check (ok)
qsctrl: passed bus control check (ok)
qsctrl: passed gateway check (10.1.0.65)
qsctrl: passed module heartbeat check
qsctrl: passed firmware version check (44-5061099)
qsctrl: passed tftp server check (10.1.0.65)
qsctrl: passed upgrade file check (503-upgrade.tar)
qsctrl: passed broadcast top check
qsctrl: passed route up check (4,4)
qsctrl: passed link state check
```

```
qsctrl: passed temperature check (32-35)
qsctrl: passed fan status check
qsctrl: passed PSU status check
```

- The **-e** flag reports environmental data, including the temperature:

```
# qsctrl -e
```

```
Name    IPAddr     Type       Build      Clock      PSU   FanSp   Temp

QR0N00  10.8.0.220 QS2_64U64D 44-5061099 local/656  O/O   OOOOOO  41'C
QR1N00  10.8.0.221 QS2_64U64D 44-5061099 local/656  O/O   OOOOOO  37'C
QR2N00  10.8.0.222 QS2_64U64D 44-5061099 local/656  O/O   OOOOOO  38'C
```

- To display the **network.ip** parameter on the switch modules, enter:

```
# qsctrl -G network.ip
```

```
Module            Primary        Secondary
QR0N00         13.8.0.220       13.8.0.235
QR1N00         13.8.0.221       13.8.0.236
```

- The following example describes how to disconnect/reconnect the ports of the **ns8** node.

Important:
This operation must be done very carefully and only by skilled administrators.

a. Before disconnection, the **nodeset** command shows that **ns8 (node ID = 8)** was present:

```
cat /proc/qsnet/ep/rail*/nodeset
```

```
[0,4-6,8,10]
[0,4-6,8,10]
```

b. Disconnect **ns8**:

```
# qsctrl -o ns8
```

```
qsctrl: QR0N00:00:2:0 configured out with 0 errors
qsctrl: QR1N00:00:2:0 configured out with 0 errors
qsctrl: Node(s) ns8 configured out
```

c. The **nodeset** command shows that **ns8** is not present any more:

```
cat /proc/qsnet/ep/rail*/nodeset
```

```
[0,4-6,10]
[0,4-6,10]
```

d. Reconnect **ns8**:

```
qsctrl -i ns8
```

```
qsctrl: QR0N00:00:2:0 configured in with 0 errors
qsctrl: QR1N00:00:2:0 configured in with 0 errors
qsctrl: Node(s) ns8 configured in
```

e. The **nodeset** command shows that **ns8** is back:

```
cat /proc/qsnet/ep/rail*/nodeset
```

```
[0,4-6,8,10]
[0,4-6,8,10]
```

7.6 More Information

For more information, please refer to Quadrics Web site: www.quadrics.com

Chapter 8. Installing and Configuring InfiniBand Interconnects

Bull is known for its high-performance computing systems based on open-source components, and contributes actively to the Linux software community.

Bull HPC Clusters may include **InfiniBand** interconnect networks with **Voltaire**® adapters and switches to provide fully integrated hardware and software stacks which include the latest Double Data Rate (**DDR**) technology.

This chapter describes how to install and configure **Voltaire** devices (those used may vary according to the size and type of cluster) and how to configure IP-over-IB.

The following topics are described:

- 8.1 *Installing HCA-400 Ex-D Interface Cards*
- 8.2 *Configuring the Voltaire ISR 9024 Grid Switch*
- 8.3 *Configuring the Voltaire ISR 9096/9288 Grid Director*
- 8.4 *Configuring Passwords*
- 8.5 *Verifying the Voltaire Configuration*
- 8.6 *More Information on Voltaire Devices*

8.1 Installing HCA-400 Ex-D 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.

Note:
Installation should be performed by an authorized user who logs in to the system as the root user.

8.2 Configuring the Voltaire ISR 9024 Grid Switch

8.2.1 Connecting to a Console

Connect the Management node, with a terminal emulation program, to the RS-232 console interface according to the instructions in the *Hardware Installation Guide*. Make sure that the terminal emulation program is configured as follows:

Setting	Value
Terminal Mode	VT-100
Baud	38400
Parity	No Parity
Stop Bits	1 Stop Bit
Flow Control	None

Table 8-1. Voltaire ISR 9024 Switch Terminal Emulation Configuration

8.2.2 Starting a CLI Management Session

To start a Command Line Interface management session for the switch via a HyperTerminal connection, do as follows:

1. Connect the switch via its serial port, using the cable supplied by Voltaire.
2. Start HyperTerminal client.
3. Configure the terminal emulation parameters as described in section 8.2.1 *Connecting to a Console*.
4. Type in the appropriate password at the logon prompt. Admin default password is: `123456`.

To change to Privileged mode:

1. Once in admin mode, enter: enable.
2. Enter the following password at the prompt: voltaire

8.2.3 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, 2006.

```
clock set 062108222006
```

8.2.4 Entering in the IP address and Default Gateway for the Management Interface

The switch requires an IP address for the management interface before it can be configured. At the command prompt, enter in the management interface IP address. Use the following commands:

1. Enter Privileged mode (from Exec mode).

```
enable <password>
```

2. Enter Config mode from the Privileged mode. Type the password when prompted.

```
config
```

3. Enter the fast interface configuration.

```
interface fast
```

4. Set the IP address of the current interface mode.

```
ip-address set <192.0.0.1> <255.255.255.0>
```

5. Exit configuration mode.

```
exit
```

6. Enter the route configuration.

```
route
```

7. Set the default gateway IP address for the fast Ethernet interface.

```
default-gw fast set
```

8. Exit configuration mode

```
exit.
```

8.2.5 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:

1. Once in admin mode, enter: **enable**.
2. Enter the following password at the prompt: **voltaire**
3. Enter the appropriate CLI commands to complete the required actions.

8.3 Configuring the Voltaire ISR 9096/9288 Grid Director

The **Voltaire**® **Grid Director**™ **ISR 9096/9288 InfiniBand** multiprotocol switching solution provides unprecedented levels of performance and scalability for large **InfiniBand** clusters and grids, enabling high performance applications to run using distributed server, storage and network resources.

Up to 96 or 288 **InfiniBand** 4X ports may be connected to provide a potential bisectional bandwidth of 10 Gbps. 4/12 slots are provided and may house different types of Line Boards or Router Blade Drawers (**sRBD**). Up to 3 **InfiniBand** Form Factor Router Modules may be installed in each Router Blade Drawer. The configuration of IP Routers (**IPR**) and Fibre Channel Routers (**FCR**) is described below.

8.3.1 Configuring the InfiniBand Address

IPR and **FCR** devices require an **InfiniBand** address so that they can be addressed by other devices in the **InfiniBand** network. This is configured as follows:

1. Enter Privileged mode (from Exec mode)

```
enable <password>
```

2. Enter Config mode from Privileged mode.

```
config
```

3. Enter the InfiniBand interface configuration.

```
interface IB
```

4. Set the IB IP address for the current interface mode.

```
ip-address-IB set 192.0.0.1 255.255.255.0
```

5. Exit the Privileged mode.

```
exit
```

8.3.2 Configuring the GbE Address

The IPR requires an IP address for the Gigabit Ethernet interfaces in order that the router may be addressed by other IP devices on the network.

1. Enter Privileged mode (from Exec mode). The default password is **voltaire**.

```
enable <password>
```

2. Enter Config mode from Privileged mode.

```
config
```

3. Enter the GBE IP interface configuration.

```
interface GBE
```

4. Set the Gigabit Ethernet IP address for the current interface.

```
ip-address-IB set 192.0.0.1 255.255.255.0
```

5. Exit the Privileged mode.

```
exit
```

8.4 Configuring Passwords

Use the following procedure for configuring passwords for Exec and Privileged mode access to the RS-232 console interface and to the Ethernet management interface (used for establishing a CLI session via Telnet; see section 8.2.5 *Starting a CLI Management Session via Telnet*).

Note:

The default password for Privileged mode is `123456` and for Exec mode is `voltaire`.

1. Enter Privileged mode (from Exec mode).

```
enable <password>
```

2. Set the Privileged and Exec mode passwords

```
password update [admin | enable]
```

3. Exit Privileged mode.

```
exit
```

8.5 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 **InfiniBand** interface IP address and configuration.

```
Ib-interface show
```

5. To display the GbE interface IP address and configuration.

```
gbe-interface show
```

6. To display the system clock.

```
clock show
```

8.6 More Information on Voltaire Devices

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

HCA 400 User Manual
Voltaire Switch User Manual ISR 9024, ISR 9096, and ISR 9288 Switches
ISR 9024 Installation Manual
IPR User Manual
FCR User Manual

Note:
For more information on the **SLURM** Resource Manager used in conjunction with InfiniBand stacks and Voltaire switches see Chapter 6 in the HPC BAS4 *Administrator's Guide* and Chapter 6 in the HPC BAS4 *User's Guide*.

Chapter 9. Checking and Backing-up Cluster Nodes

This chapter describes the following topics:

- 9.1 *Checking the Management Node*
- 9.2 *Checking Other Nodes*
- 9.3 *The List of the Installed Bundles*
- 9.4 *Checking the Release*
- 9.5 *Backing up the System*

9.1 Checking the Management Node

Check the following:

- The required services (**Torque**, **Lustre**, **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 resource manager **Quadrics RMS** or **SLURM**, and the **MPI** libraries are all running correctly. Once **Lustre** has been activated it will then be possible to verify what is mounted on each node.

9.2 Checking Other Nodes

9.2.1 Nodechecking

nodechecking is a tool used for verifying cluster nodes. It should be installed on all the nodes to be verified. An implementation of **MPI**, the Intel C compiler, the Intel Fortran compiler and the Intel MKL library have to be installed on the nodes so that the all the tests can be executed.

Refer to the HPC BAS4 *Maintenance Guide* for details of the tests and the command options used by **nodechecking**.

9.2.2 I/O status

I/O status is a **Nagios** service which runs continuously for an I/O node. The results are reported to **NovaScale Master – HPC Edition**, through 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 – see the HPC BAS4 *Maintenance Guide* for more details.

Refer to the chapter on Storage Device Management in the HPC BAS4 *Administrator's Guide* for more information about the I/O status service and **NovaScale Master – HPC Edition**.

9.3 The List of the Installed Bundles

The list of installed bundles is included in the file **README-fr** or **README-en** in the root directory of installation CD1 or of the DVD. These bundles are described in Appendix D.

9.4 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)
```

9.5 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...) This ensures that the **ClusterDB** is up to date on the system save disk.

To save and restore the system use the Cloning Method. This method of saving and restoring the Management Node is based on system disk cloning using the **dd** command, once the RMS (if used) and ClusterDB databases on the system disk have been saved.

See:
HPC BAS4 Maintenance Guide for details on how to save and restore the system.

Appendix A. Installation Errors

The installation procedures described in this guide should run without any errors. This appendix describes some errors that you may encounter if a procedure is not correctly applied.

A.1 Message 'Error in Locating EFI System Partition Protocol'

This message is displayed rapidly during the EFI phase and usually only for the first time. Ignore it.

A.2 The Machine Freezes during Installation

Several symptoms may appear:

The screen freezes

Within a cluster configuration, only a Service Node connects to a monitor or console. The installation process of a client node is reported via this monitor and line "cu".

Even if a monitor or console is connected directly to a client node, nothing will be displayed on the screen because the BIOS has been modified to redirect the console to the **tty** of a Service Node. However, if you want to perform your installation without redirecting the report, you must consider this node as the Service Node and type **admin** instead of **node**.

If, despite the checks described above, the procedure still appears to be blocked for some time, switch off and then switch on again.

Message "Error opening: kickstart file"

The installation freezes on the **Kickstart Error** screen with the message:

```
Error opening: kickstart file
/tmp/<kickstart file>: no such file or directory
```

Cause of the problem: error on the Bull CD-ROM.

- Click OK: the machine will reboot. This will cause the CD-ROM to be ejected. **Do not push it back in**.
- Remove the CD-ROM to check it, clean it and insert it again.
- Under the **EFI** shell, the procedure will fail to start up.
- Follow the standard installation procedure.

Message "Can't determine device capacity"

Cause of the problem: the disk is badly inserted or moving the machine has resulted in a connection issue.

- Switch off the machine (recommended).
- On **NovaScale 40xx** servers, remove the disk and install it back.
- On **NovaScale 5xxx/6xxx servers**, perform a verification using the disk manager.
- Return to the installation procedure.

Message "cu: /dev/ttyD000: Line in use"

Possible causes of the problem:

- A "**cu**" process is using the line:
  - Run the following command to check if it is the cause of the problem:

```
/bin/ps -efa |grep cu
```

  - If it is the case, ask the user to quit the **cu** session or kill the processes with **kill –9 pid_no** for cu processes (input and output).
- The **drv-epca-1.50-1.b.1.Bull** package for the communication controllers "Digi International" AccelePort Xr" (8 ports) or the "AccelePort C/X"(128 ports) has not been installed or the driver has not been configured.
  - Check if the driver exists, using the following commands:

```
ls /lib/modules/`uname -r`/epca.ko
rpm -q drv-epca
```

  - If the driver does not exist, install the **drv-epca-1.50-1.b.1.Bull package** and configure the driver by running the **digiConf** command.

```
rpm -ivv drv-epca-1.50-1.b.1.Bull
/usr/sbin/digiConf
```

  - Answer the series of questions according the Digiboard communication controller(s) which is installed.
- The driver has not been loaded:
  - Check if the driver is loaded, using the following command:

```
/sbin/lsmod |grep epca
```

    The response should be similar to:

    ```
    epca       102608 . . .
    ```

  - If it is not the case, run the following command:

```
/sbin/service epca.rc start
```

A.3 Localization: Messages in English

If the installation is not in English, some messages or menu items may not be translated or are only partially translated into the local language.

With **cu** line some French characters are badly interpreted.

A.4 Power out During Installation

If the machine stops, either intentionally or not, as a result of a power failure for example, during any phase of the installation process, then simply switch on the machine and restart the entire installation process.

A.5 Kernel Warning Messages

Some warning messages, similar to the following, may be displayed at the console or in the **/var/log/messages** file.

```
kernel: <application-name> (<pid-number>): floating-point assist
fault at ip ...
```

For example:

```
kernel: ntpd(9638): floating-point assist fault at ip 2000000800047962,
isr 0000020000007001
```

This message is due to Network Time Protocol (NTP) bad portability code for the IA64 architecture. This application comes directly from Linux community and the code is not modified by Bull.

This explanation is valid for all warning messages of the same type.

A.6 The Installation of the Quadrics CD Fails

To install the Quadrics CD on the management node correctly, the administration network has to be assigned to **eth0**. So, if problems are encountered during Quadrics CD installation, check to see if this condition has been met.

Appendix B. Configuring Switches

This appendix describes how to configure different switches, CISCO, Brocade, Quadrics and Voltaire.

The following topics are described:

- B.1 *Configuring a CISCO Switch*
- B.2 *Configuring a Brocade Switch*
- B.3 *Configuring a Quadrics Switch*
- B.4 *Configuring Voltaire Devices*

B.1 Configuring a CISCO Switch

Pre-Requisites

Before a CISCO 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 to the Linux machine:
   This must be done using a Linux machine and a serial cable.
   Using the serial cable, connect an available serial port to the CONSOLE port at the rear of the Cisco switch (model 2950 or 3750). Note the number of the serial port as this will be needed later.

2. Establish the connection with the switch, from the Linux machine:
   - 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, run:

```
cu -s 9600 -l /dev/ttyS0
```

If the serial cable connects using port 1, run:

```
cu -s 9600 -l /dev/ttyS1
```

Enter "no" to any questions which may appear in until the following message is displayed.

```
Connected.
Switch>
```

3. Configure the switch:

– Set the enable mode:

```
Switch>enable
```

– Enter the configuration mode:

```
Switch#configure terminal
Enter configuration commands, one per line.  End with CNTL/Z.
Switch(config)#
```

– Set the name of the switch in the form: *hostname < switch_name>*. For example:

```
Switch(config)#hostname myswitch
myswitch(config)#
```

– Enter the **SVI Vlan 1** interface configuration mode:

```
myswitch(config)#interface vlan 1
myswitch(config-if)#
```

– 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
```

– Exit the interface configuration:

```
myswitch(config-if)#exit
myswitch(config)#
```

– 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.
```

– Set a password for the enable mode. For example:

```
myswitch(config)#enable password myswitch
```

– 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
```

– 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
```

– Exit the configuration :

```
myswitch(config)#exit
```

– Save the configuration in RAM:

```
myswitch#copy running-config startup-config
```

4. Update the switch boot file on the Management Node

– Run the following commands from the Management Node console.

Note:
The switch configure file name must include the switch name followed by '**-confg**', for example, **eswu1c0-confg.** Using this convention each time a switch configure file is saved avoids the risk of having lots of different versions saved with different names.

```
  touch /tftpboot/<switch_configure_file>
chmod ugo+w /tftpboot/< switch_configure_file>
```

5. 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.

6. 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.

7. 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:

```
#sh conf
Using 1053 out of 32768 bytes
!
version 12.1
no service pad
service timestamps debug uptime
service timestamps log uptime
no service password-encryption
!
hostname myswitch
!
enable password myswitch
!
ip subnet-zero
!
!
spanning-tree mode pvst
spanning-tree portfast default
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
!
!
interface FastEthernet0/1
 no ip address
!
interface FastEthernet0/2
 no ip address
!
interface FastEthernet0/3
 no ip address
!
interface FastEthernet0/4
 no ip address
!
interface FastEthernet0/5
 no ip address
!
interface FastEthernet0/6
 no ip address
!
interface FastEthernet0/7
 no ip address
!
interface FastEthernet0/8
 no ip address
!
interface FastEthernet0/9
 no ip address
!
interface FastEthernet0/10
 no ip address
!
interface FastEthernet0/11
 no ip address
!
interface FastEthernet0/12
 no ip address
!
interface Vlan1
```

```
 ip address 10.0.0.254 255.0.0.0
 no ip route-cache
 shutdown
!
ip http server
!
!
line con 0
 password admin
 login
line vty 0 4
 password admin
 login
line vty 5 15
 password admin
 login
!
end
```

B.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
```

B.3 Configuring a Quadrics Switch

A direct connection to the switch controller is used during configuration or when remote connections are not available. To make a direct connection, connect a serial cable to the console port or connect a standard PS2 keyboard and VGA monitor to their respective ports.

The switch control menu is then displayed as follows:

```
Quadrics Switch Control -- (QxxNxx Slot X)
        1. Show network settings
        2. Change network settings
        3. Run jtest
        4. Set module mode
        5. Firmware upgrade
        6. Quit
        7. Reboot
        8. Access Settings
        9. Self Test
```

Each switch must be assigned an IP address. The configuration files can be edited by hand, as described in the **Quadrics** documentation. This menu option allows you to assign IP (static or BOOTP), the type of switch (N for node, and T for top (federated)), the location in the Quadrics network (position of the switch in case of multi-switches configuration), the IP netmask, gateway, TFTP server (for firmware update), the firmware name (keep default value).

Set module mode

This option allows the administrator of the switch to set the module mode, the switch can be toggled between being a standalone switch, or a switch which is part of a federated network composed of more than one switch unit. The switch can also be toggled to enable or disable the redundant mode. Some options are restricted to **QS64** switches and not available on low cost switches (QS8).

Access settings

This option allows the administrator of the switch to change the connection protocol to **ssh** and set telnet or **ssh** password.

See:

- Chapter 5, *Quadrics Interconnect Installation*, in the present guide
- *QsNet Installation and Diagnostics Manual* (Quadrics documentation) for details.

B.4 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 and may be used to monitor **Voltaire® Grid Director™ ISR 9096/9288** 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** 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 Switches* manual provided on the *Voltaire Switches Documentation CD*.

Appendix C. 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, particularly with regard to the Ethernet connections which can be used for both the Administration network and interconnection.

The following topics are described:

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

C.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.

C.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 BAS4. These are sorted by throughput, giving the HBA slotting rank.

Adapter	Bandwidth
SCSI SAS 3442X	1200 MB/s
Quadrics Elan 4	900 MB/s
Fibre channel dual ports	800 MB/s (1) (2)
SCSI MegaRAID 320 X2 dual channel	640 MB/s (1)
Fibre channel single ports	400 MB/s (2)
SCSI U320 single channel	320 MB/s
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 C-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 C-2. PCI-Express Table

C.3 Recommendations for NovaScale Servers

The following paragraphs describe the architecture of the I/O subsystem for each family of NovaScale servers. Recommendations for the adapter/PCI slot allocation order are also provided.

C.3.1 The NovaScale 3045 series platform

The diagrams below illustrate the entire I/O subsystem for this range of server.

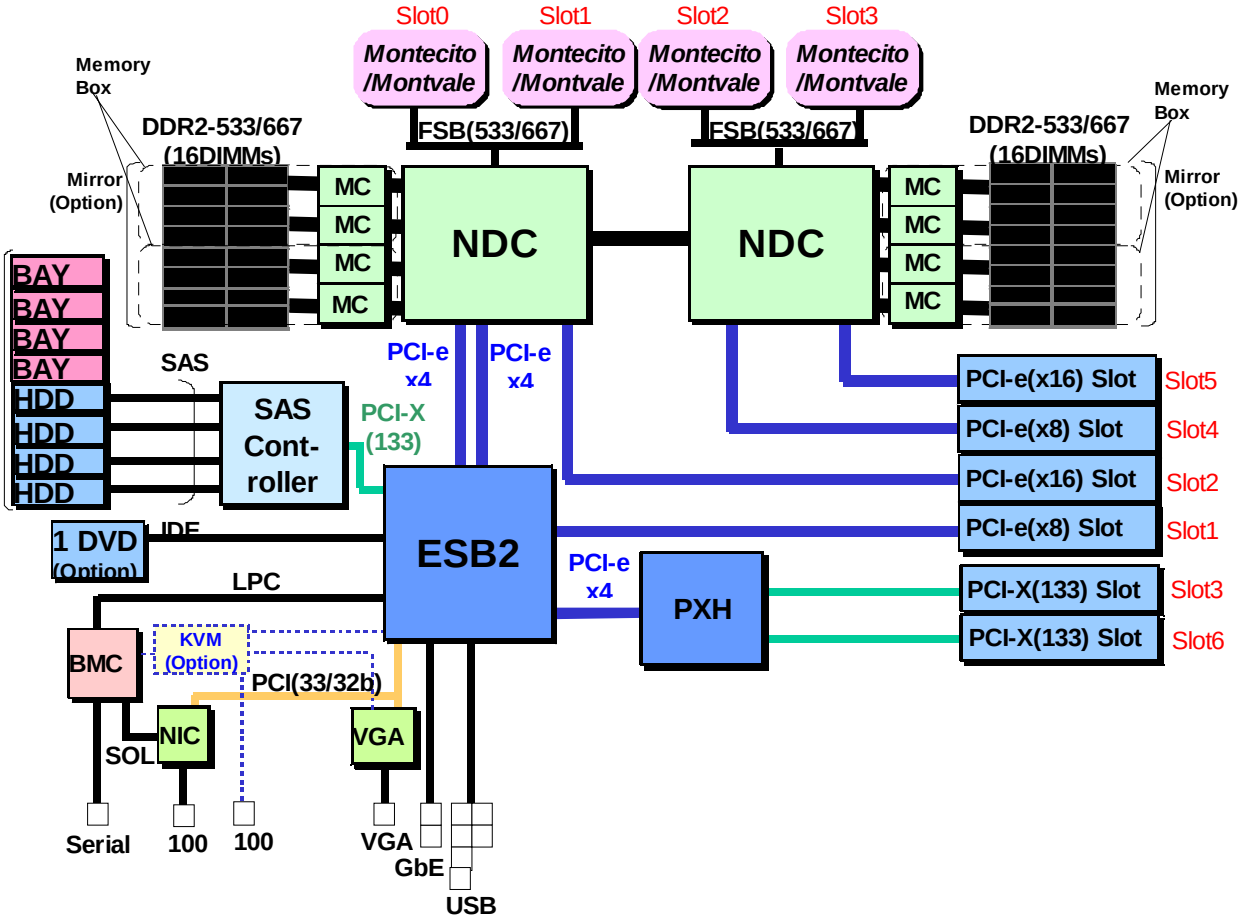

Figure C-1. NovaScale 3045 series complete I/O subsystem

C.3.1.1 NovaScale 3045 – I/O subsystem and slotting

The next figure shows the rear view of the NovaScale 3045 drawer, including the PCI card slots.

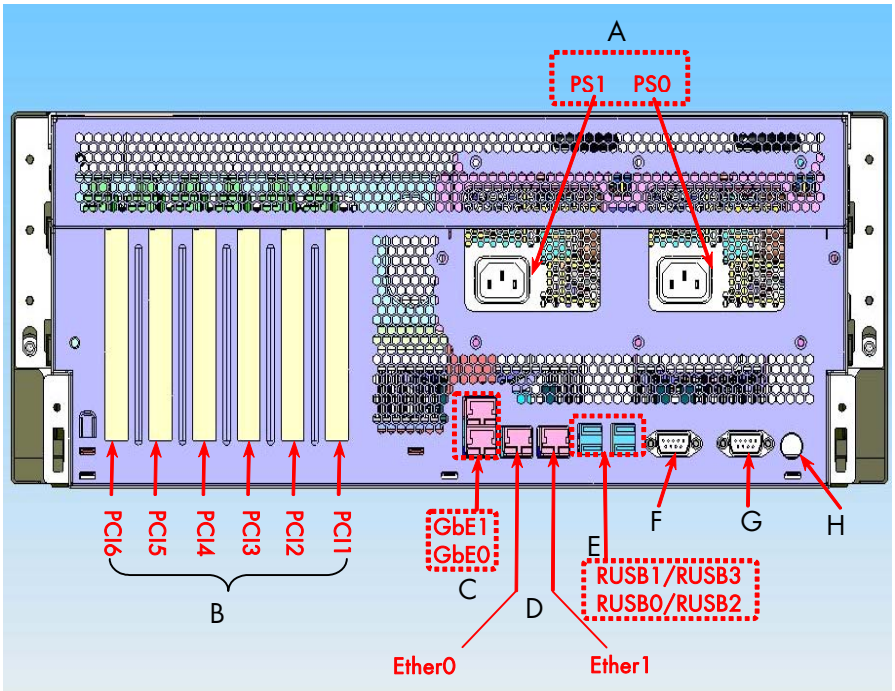

Figure C-2. NovaScale 3045 series – PCI slot identification

#	Item	Description
A	AC input power connectors	
B	PCI Slots	All slots support hot-plug PCI add-in cards: - PCI 1: PCIe x 8 slot, short - PCI 2: PCIe x 16 slot, short - PCI 3: 133 MHz, 64-bit PCI-X slot, short - PCI 4: PCIe x 8 slot, short - PCI 5: PCIe x 16 slot, medium - PCI 6: 133 MHz, 64-bit PCI-X slot, long
C	GbE ports	GbE0 – Admin Network, GbE1
D	100 MbE ports	Ether0 – BMC Management, Ether1
E	Four USB ports	4-pin connectors
F	Video port	Standard VGA compatible – KVM LAN
G	Serial port	9-pin RS-232 connector
H	Identification Switch	Toggles Server drawer ID LED On/Off.

Table C-3. NovaScale 3045 Series rear connections

See the Bull NovaScale 3005 Series documentation, including the *NovaScale 3005 Series Installation and User's Guide,* for more hardware information regarding NovaScale 3045 Series platforms.

C.3.1.2 NovaScale 3045 Series PCI Slot Priority lists

The table below shows the priority list for **PCI Express** slot population.

Adapter Ranking	PCI-e Slot
1	5
2	2
3	4
4	1

Table C-4. NovaScale 3045 PCI Express slot priorities

The table below shows the priority list for **PCI-X** slot population. The throughputs possible for slots 3 and 6 are the same but slot #3 is a short slot and slot #6 is a long slot.

Adapter Ranking	PCI-X Slot
1	6
2	3

Table C-5. NovaScale 3045 PCI-X slot priorities

C.3.2 The NS 3045 Compact Series Platform

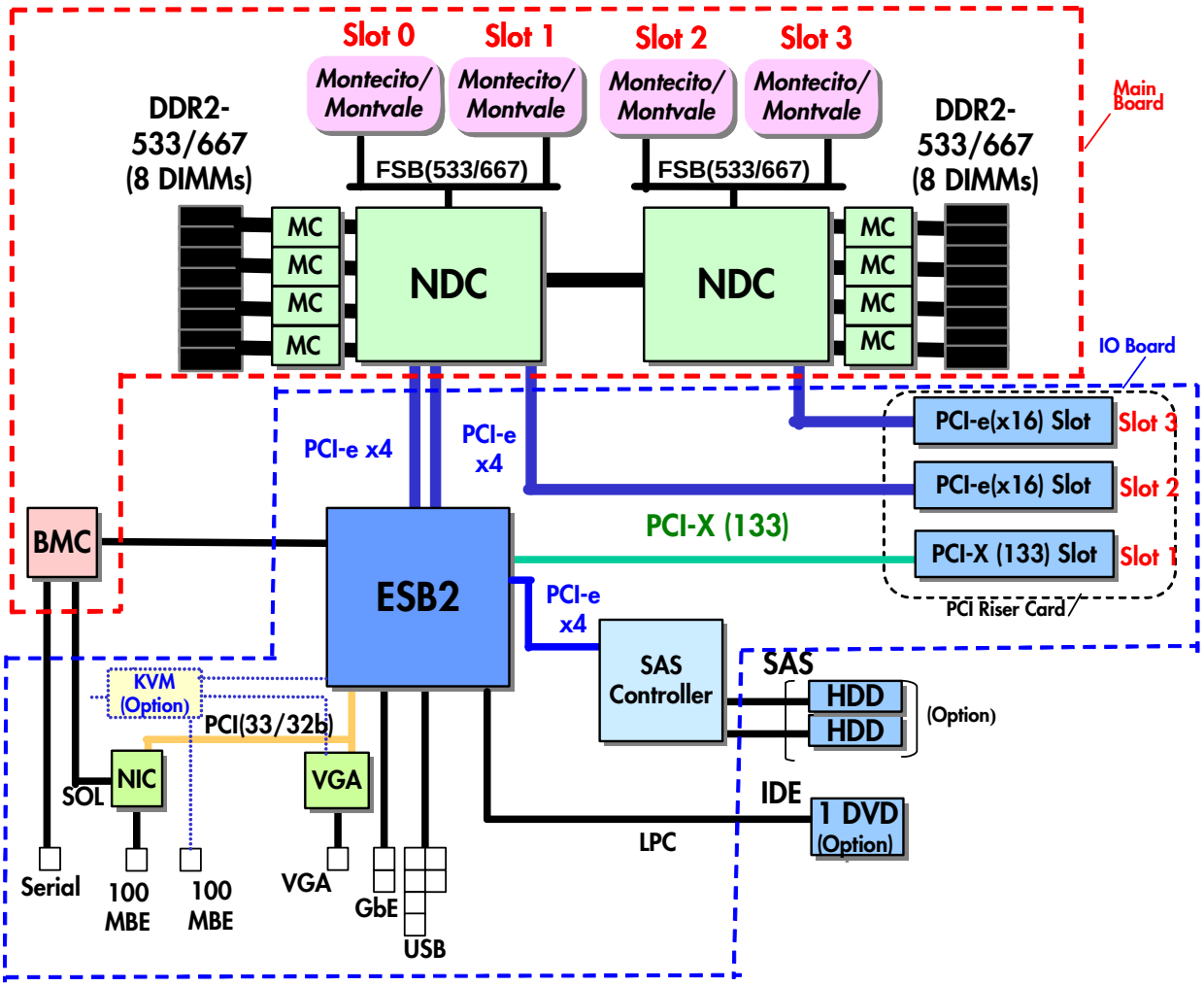

Figure C-3. NovaScale 3045 Compact – I\O architecture

The next figure shows the features found on the server drawer, including the slots for PCI boards, for the NovaScale 3045 compact series platforms.

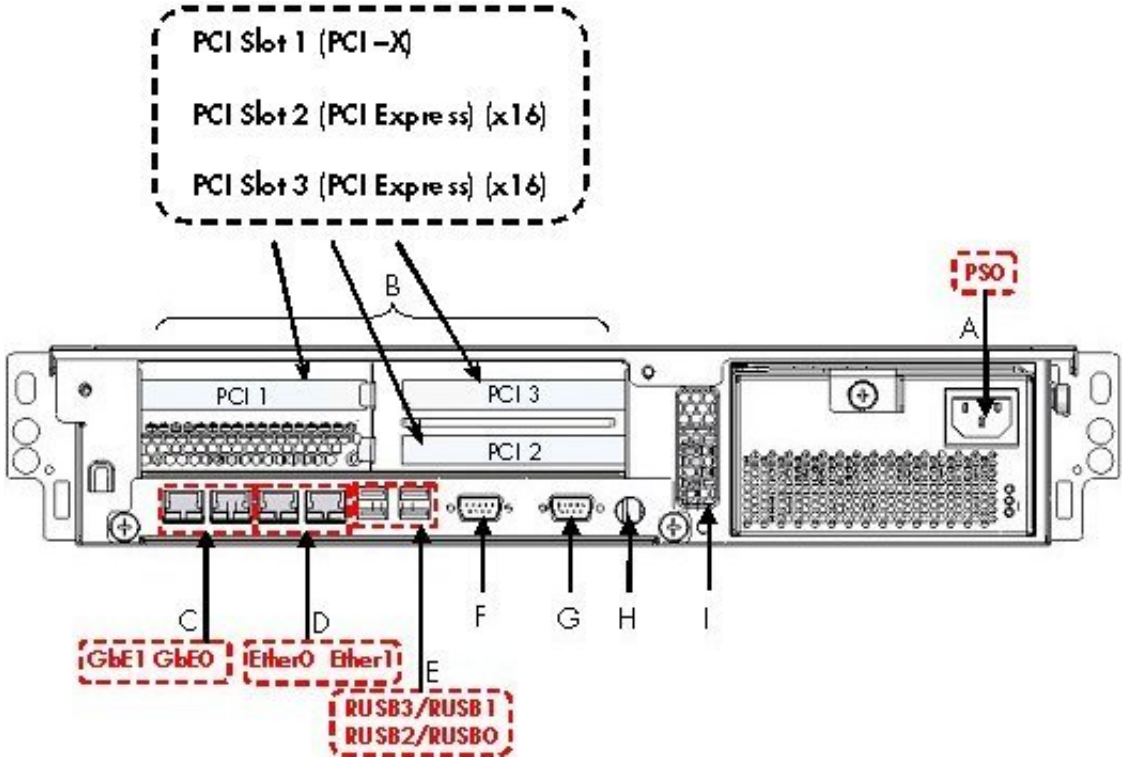

Figure C-4. NovaScale 3045 Compact Series PCI Slots

#	Components	Description
A	AC input power connector	
B	PCI Slots	All slots do not support hot-plug. - Slot 1: 133 MHz, 64-bit PCI-X slot, long length - Slot 2: PCIe(x16) slot, long length - Slot 3: PCIe(x16) slot, short length
C	GbE ports	GbE0 – Admin Network, GbE1
D	100 MbE ports	Ether0 – BMC Management, Ether1
E	Four USB ports	4-pin connectors
F	Video port	Standard VGA compatible – KVM LAN
G	Serial port	9-pin RS-232 connector
H	Identification Switch	Toggles Chassis ID LED On/Off.
I	Air Chamber	Holes for air evacuation

Table C-6. NovaScale 3045 COMPACT Server Components – Rear view

C.3.2.1 NovaScale 3045 Compact Series – PCI Express Slot Priority

The following table provides the priority list for the population of the PCI Express slots for NovaScale 3045 compact models.

Adapter Ranking	PCI-e Slot
1	2
2	3

Table C-7. NovaScale 3045 compact PCI Express slot priorities

C.3.3 The NovaScale 4020 series platform

The following diagram shows the entire I/O subsystem for NovaScale 4020 servers.

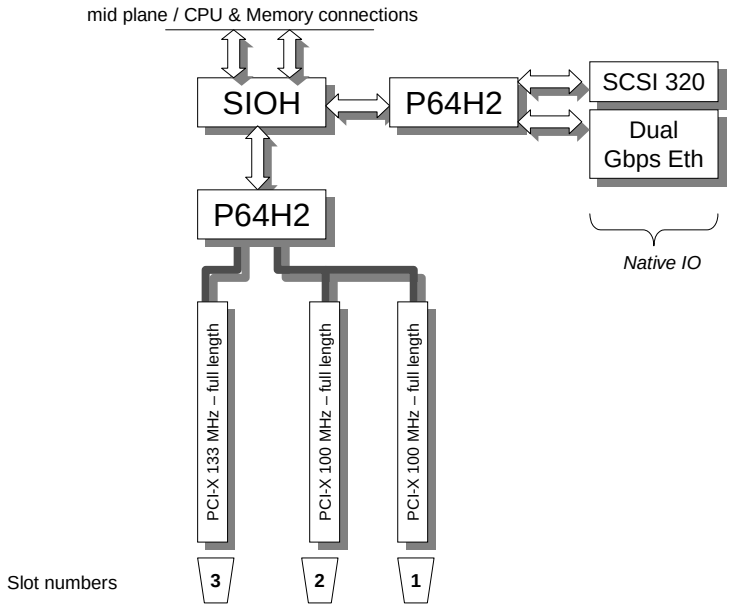

Figure C-5. NovaScale 4020 – I/O subsystem

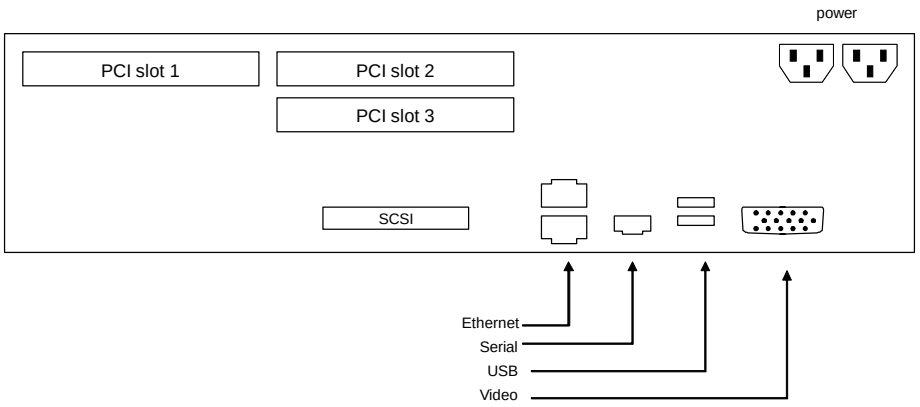

Figure C-6. NovaScale 4020 Model – PCI slot identification

The priority for the population of the PCI slots is shown in the table below.

Adapter Ranking	Slot number
1	3
2	2
3	1

Table C-8. NovaScale 4020 Model.

C.3.4 NovaScale 4040 Series Platform

The following diagram shows the entire I/O subsystem for NovaScale 4040 servers.

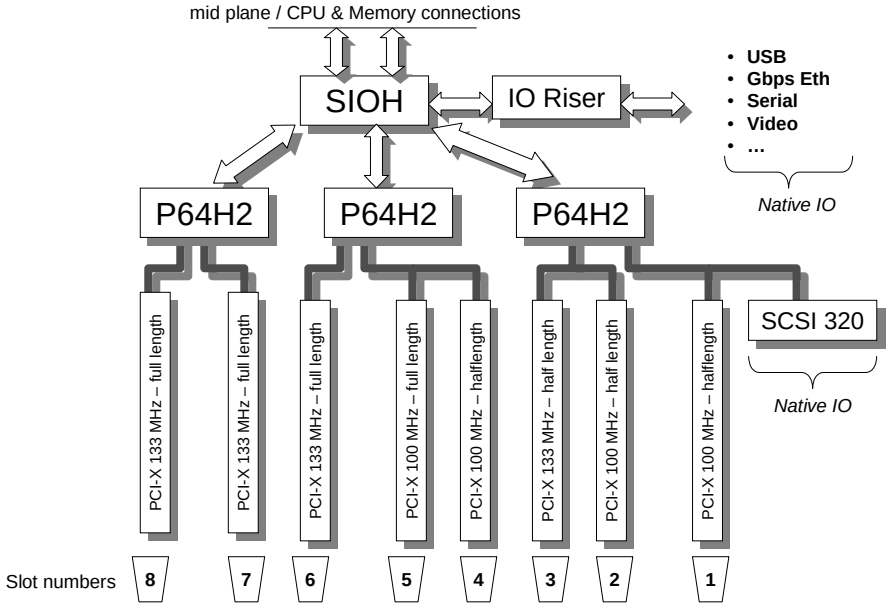

Figure C-7. NovaScale 4040 Series – I/O subsystem

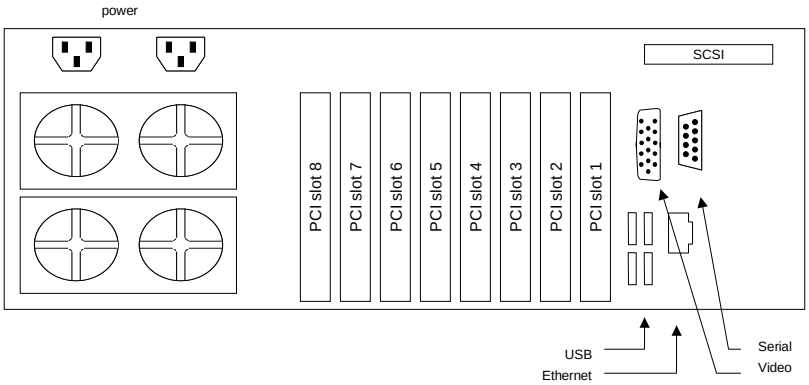

Figure C-8. NovaScale 4040 Series – PCI slot identification

The following table provides the priority list for the population of the PCI slots for NovaScale 4040 Series platforms.

Adapter Ranking	Slot number
1	8
2	6
3	7
4	5
5	3
6	2
7	4
8	1

Table C-9. NovaScale 4040 Series Platform PCI Slot priorities

C.3.5 NovaScale 5xx0/6xx0 Series models

The following diagram shows the IOB I/O subsystem for NovaScale 5xx0 /6xx0 servers. The number of IOB modules per server ranges from 1 to 4 depending on the selected options.

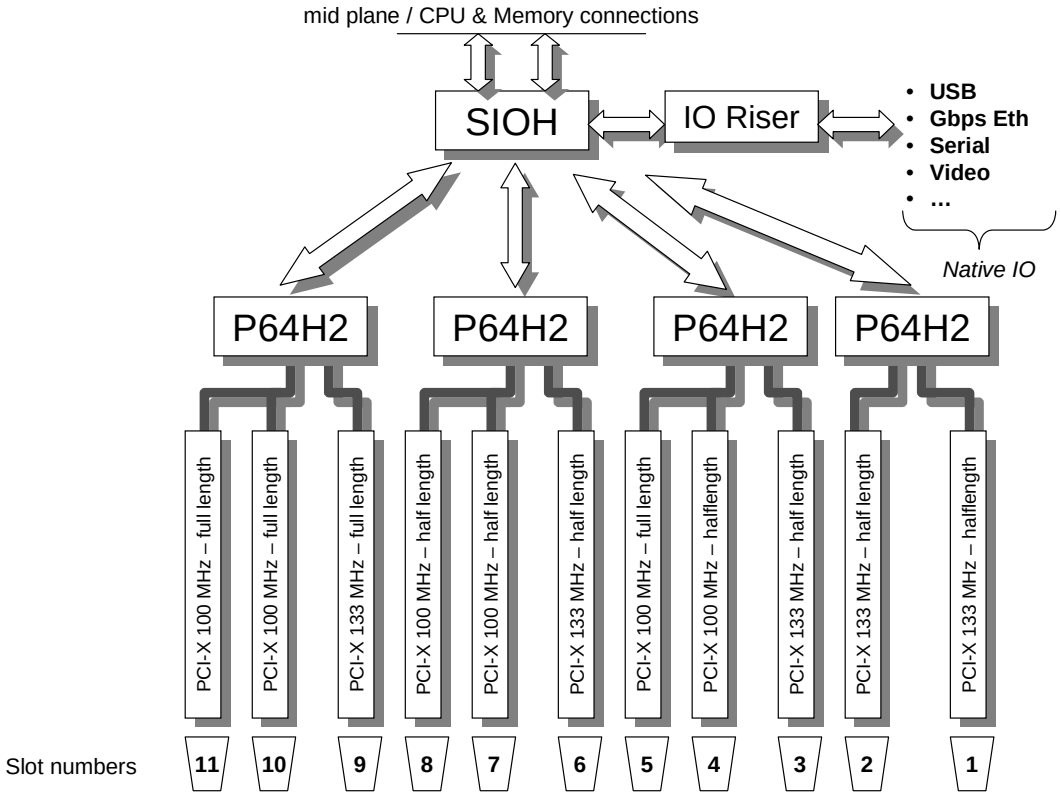

Figure C-9. NovaScale 5xx0/6xx0 – I/O subsystems per IOB

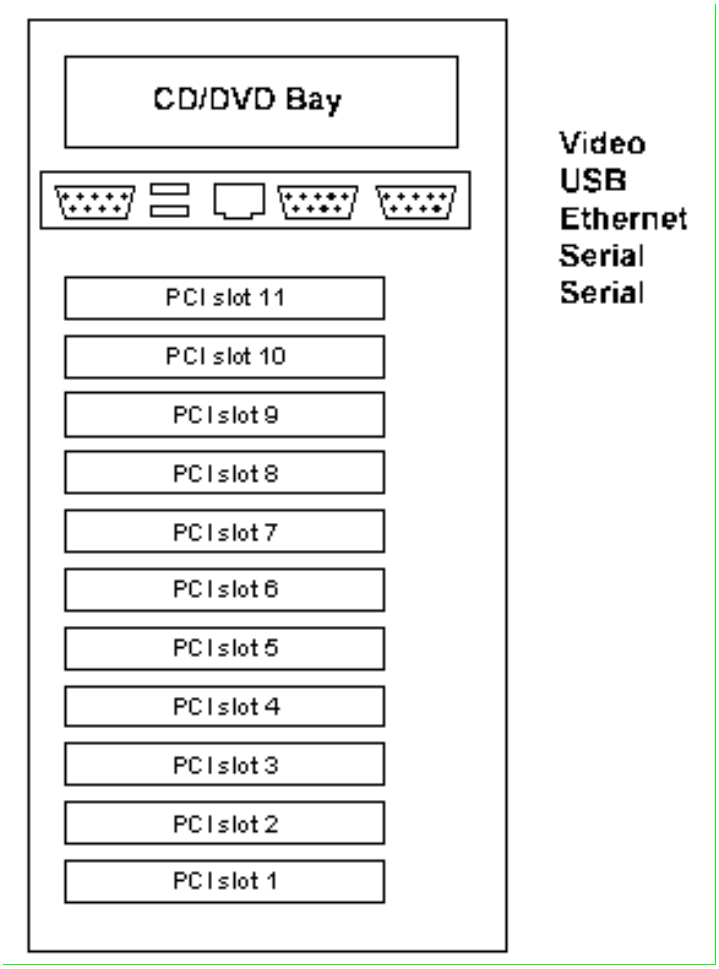

Figure C-10. NovaScale 5xx0/6xx0 – PCI slot identification per IOB

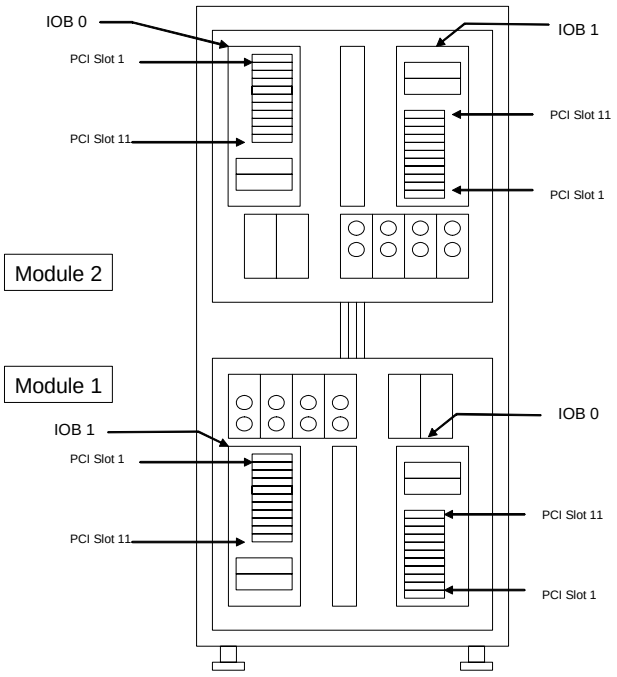

Figure C-11. NovaScale 5xx0/6xx0 – IOB identification for dual module / 32w capable

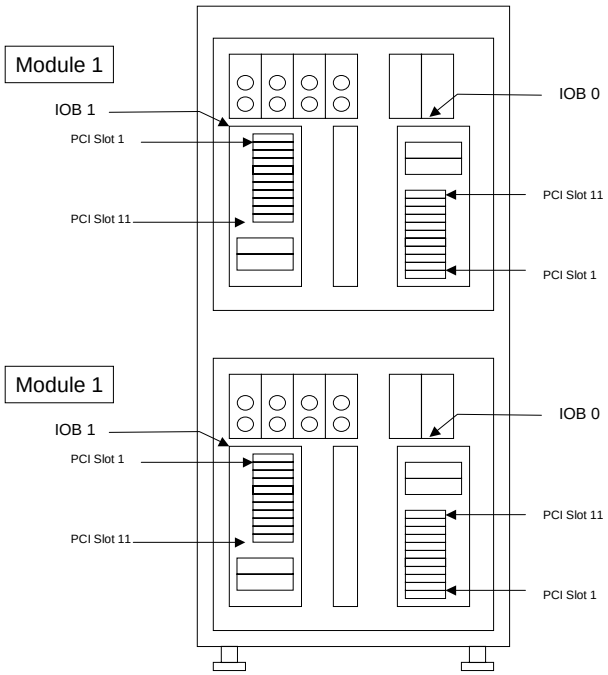

Figure C-12. NovaScale 5xx0/6xx0 – IOB identification for 2 single modules in a rack

The following table provides the priority list for the population of PCI slots within an IOB. It is recommended to populate the PCI Slots of the same priority level for each IOB of the server before populating the next priority level. Within a given PCI slot priority level, the IOBs must be populated in the following order: IOB 0/Module 1, IOB 0/Module 2, IOB 1/Module 1, and IOB 1/Module 2.

Furthermore, another rule should be applied when the environment has to sustain significant unidirectional data flows. A typical example is a server reading a large amount of data from the Quadrics interconnects, and pushing it to the disks through fibre channel connections. The number of PCI adapters may be different, but the in and out flows are almost identical. The additional rule to apply is to balance the HBAs with unidirectional data flows across different IOBs.

It the server does not have many high performance adapters; it is possible to reserve slot 1 of the module 1/IOB 0 for the SCSI adapter used for the boot disks. This is a common practice for this family of servers, regardless of the operating system used (Bull BAS, Microsoft Windows, RedHat AS, SuSE SLES, etc.).

Priority	Slot number
1	3
2	6
3	9
4	1
5	2
6	5
7	8
8	11
9	4
10	7
11	10

Table C-10. NovaScale 5xx0/6xx0 Series PCI Slot priorities

C.3.6 NovaScale 5xx5 Series Models

The following diagrams illustrate the **IOB** I/O subsystems for this range of servers. The number of IOB subsystems per server varies from 1 to 8 depending on the options selected. A single server may be composed of 1, 2, 3 or 4 modules. The first module includes 1 **QBB**, 1 **IOB**, 1**DIB** and optionally a second IOB and another DIB. The other modules may be delivered with or without IOBs and DIBs. The largest configurations contain 2 IOB and 2 DIBs chained per module.

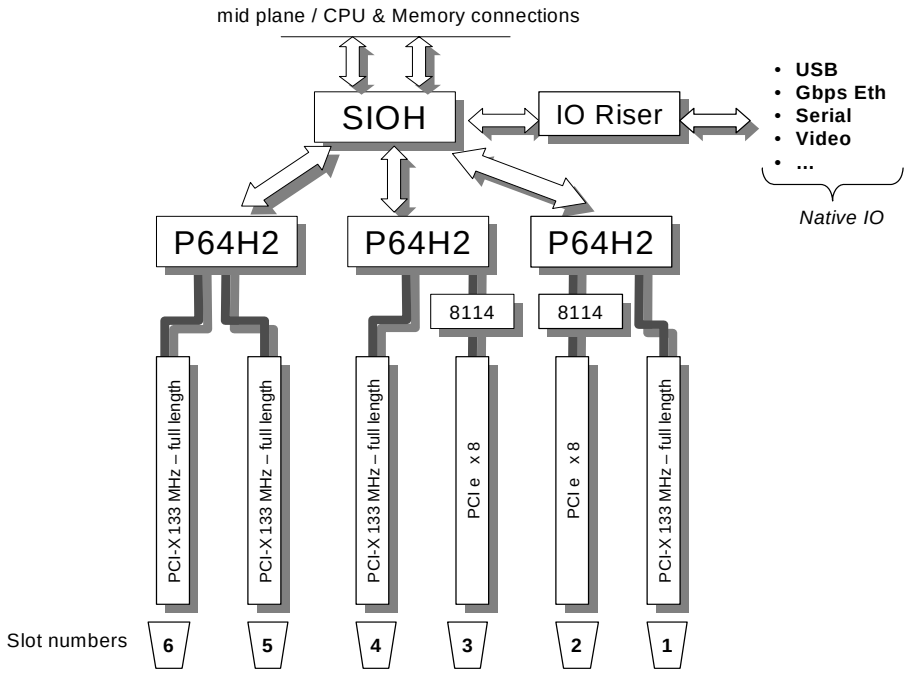

Figure C-13. NovaScale 5xx5 – I /O subsystem slotting

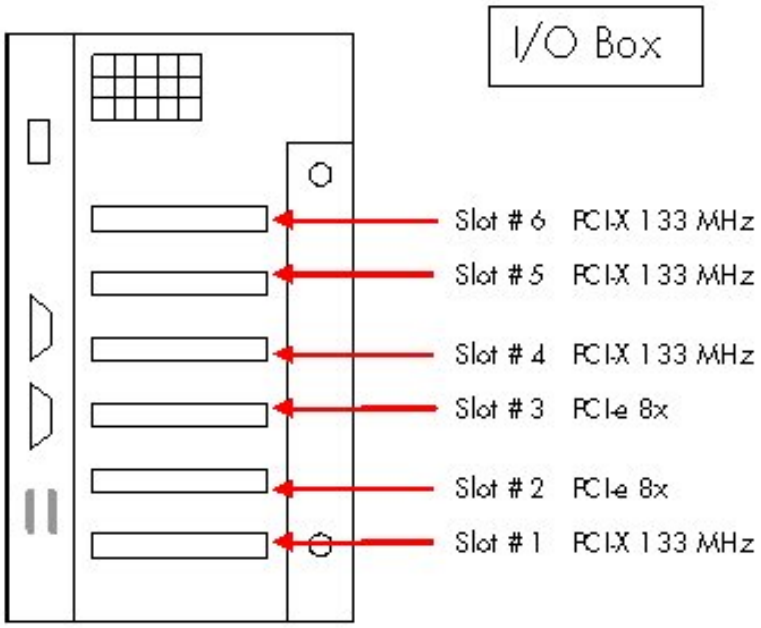

Figure C-14. NovaScale 5xx5 Platform PCI Slot identification

The following table provides the priority list for the population the PCI-X slots. It is preferable to avoid putting a PCI-Express HBA next to a PCI-X HBA, which is connected to the same P64H2, as the total throughput may then exceed the P64H2 capacity (1 GB/s).

Adapter Ranking	PCI-X Slot
1	6
2	4 (skip if HBA is in PCI –e #3 slot)
3	1 (skip if HBA is in PC1-e #2 slot)
4	5

Table C-11. NovaScale 5xx5 platform PCI-X slot priorities

The following table provides the priority list for the population of the PCI-Express slots for NovaScale 5xx5 platforms.

Adapter Ranking	PCI-Express Slot
1	2 (skip if HBA is in slot PCI-X #1)
2	3 (skip if HBA is in slot PCI-x # 4)

Table C-12. NovaScale 5xx5 platform PCI-Express priorities

As stated above, it is preferable to avoid putting a PCI-Express HBA next to a PCI-X HBA, which is connected to the same P64H2, as the total throughput may then exceed the P64H2 capacity (1 GB/s). However, if a PCI-X HBA and a PCI-Express HBA have a total input which is inferior to 1 GB/s (for example, a PCI-e FC single port adapter and a Gibabit PCI-X adapter), then they can be slotted into slots connected to the same P64H2 (1& 2, 3 & 4).

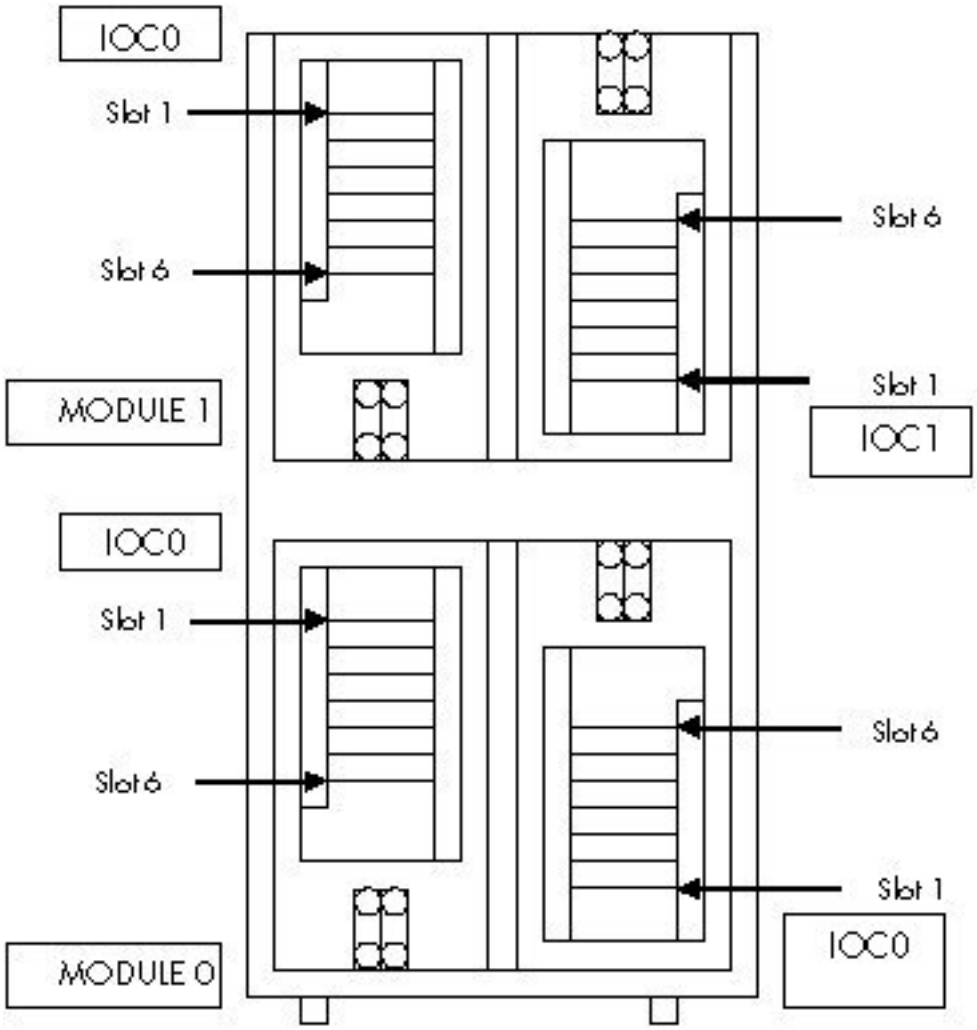

Figure C-15. NovaScale 5xx5 – IOC identification for dual module/32w capable

It is recommended to populate the IOC PCI Slots with the same priority level before populating the next priority level. Within a given PCI slot priority level, the IOCs must be populated in the following order: IOC 0/Module 1, IOC 0/Module 2, IOC 1/Module 1, IOC 1/Module 2. For example, if there are 2 PCI-e HBAs to be plugged into a NovaScale server with 2 IO Boxes, then the two slots which are ranked first on the two IO Boxes should be used, as opposed to using the first and second slot on the same IO Box.

An alternative, for servers which are partitioned, is to use each partition as a single machine and populate the IO boxes of the partition according to the adapter rankings in the tables above.

The server boots on the disks included in a drawer connected via a Device Interface Board (DIB), or on FC disks, or on SAS disks. In any case a slot has to be reserved for the HBA which provides access to the disks (SCSI U 320, FC Dual or single port, MegaRAID SCSI 320-2X, SAS 3442X). This is a common practice for this family of servers, regardless of the operating system used (Bull BAS, Microsoft Windows, RedHat AS, SuSE SLES, etc.).

Appendix D. BAS4 Bundles

The following table lists the bundles available on BAS4 CDs. It also indicates which bundles are installed by the automatic installation on the **Management node**, **Compute node**, **Login node** or **Standalone node**.

D.1 Bull Linux AS4 Media Bundles

Bull Linux AS (*Bull Linux Advanced Server* or *BLAS*) consists of Kernel software from kernel.org and Open Source Linux distribution only, whereas **BAS** (*Bull Advanced Server*) designates the whole Bull HPC software offer, including Bull Linux AS, Bull HPC and Bull Cluster Management software. The versions of Bull Linux AS and BAS may be different.

Bundle Name	Management Node	Compute Node	Login Node	Standalone Node	Description
High Performance Computing Tools For Compute Node		X			Install these tools to enable compute node HPC profile.
High Performance Computing Tools For Management Node	X				Install these tools to enable Management Node HPC profile.
High Performance Computing Tools For Login Node			X		Install these tools to enable Login Node HPC profile.
High Performance Computing Tools For Standalone Node				X	Install these tools to enable Standalone Node HPC profile.
Printing Support	X			X	Install these tools to enable the system to print or act as a print server.
X Window System	X		X	X	Install this group of packages to use the base graphical (X) user interface.
GNOME Desktop Environment	X		X	X	GNOME is a powerful, graphical user interface which includes a panel, desktop, system icons, and a graphical file manager.
KDE (K Desktop Environment)					KDE is a powerful, graphical user interface which includes a panel, desktop, system icons, and a graphical file manager.
Graphical Internet	X		X	X	This group includes graphical email, Web, and chat clients.
Text-based Internet					This group includes text-based email, Web, and chat clients. These applications do not require the X Window System.
Sound and Video	X			X	From CD recording to playing audio CDs and multimedia files, this package group allows you to work with sound and video on the system.

Bundle Name	Management Node	Compute Node	Login Node	Standalone Node	Description
Graphics	X			X	This group includes packages to help you manipulate and scan images.
Office/Productivity	X		X	X	The applications include office suites, PDF viewers, and more.
Mail Server					These packages allow you to configure an IMAP or Postfix mail server.
Network Servers	X				These packages include network-based servers such as DHCP, Kerberos and NIS.
Legacy Network Server					These packages include servers for old network protocols such as rsh and telnet.
News Server					This group allows you to configure the system as a news server.
Windows File Server	X			X	This package group allows you to share files between Linux and MS Windows systems.
Server Configuration Tools	X		X	X	This group contains all of custom server configuration tools.
FTP Server	X		X	X	These tools allow you to run an FTP server on the system.
PostgreSQL Database	X				This package group includes packages useful for use with Postgresql.
MySQL Database					This package group contains packages useful for use with MySQL.
Web Server	X			X	These tools allow you to run a Web server on the system.
DNS Name Server	X			X	This package group allows you to run a DNS name server (BIND) on the system.
Authoring and Publishing					These tools allow you to create documentation in the DocBook format and convert them to HTML, PDF, Postscript, and text.
Engineering and Scientific	X		X	X	This group includes packages for performing mathematical and scientific computations and plotting, as well as unit conversion.
Editors	X		X	X	Sometimes called text editors, these are programs that allow you to create and edit files. These include Emacs and Vi.
Emacs			X		Emacs text editor
System Tools	X			X	This group is a collection of various tools for the system, such as the client for connecting to SMB shares and tools to monitor network traffic.
Administration Tools	X		X	X	This group is a collection of graphical administration tools for the system, such as for managing user accounts and configuring system hardware.
Games and Entertainment					Various ways to relax and spend your free time.

Bundle Name	Management Node	Compute Node	Login Node	Standalone Node	Description
Development Tools	X		X	X	These tools include core development tools such as automake, gcc, perl, python, and debuggers.
Development Libraries			X		
Legacy Software Development		X	X	X	These packages provide compatibility support for previous releases of Linux.
Compatibility Arch Support			X		Multilib support packages
Compatibility Arch Development Support			X		Multilib development support packages
X Software Development			X	X	These packages allow you to develop applications for the X Window System.
GNOME Software Development					Install these packages in order to develop GTK+ and GNOME graphical applications.
KDE Software Development					Install these packages to develop QT and KDE graphical applications.

D.2 Bull HPC Media Bundles

The Bull HPC media contains the following features.

Bundle Name	Management Node	Compute Node	Login Node	Standalone Node	Description
FEATURES					
CPUSET	X	X	X	X	Development tool
PTOOLS	X	X	X	X	Development tool
PERFTOOLS					
PAPI	X	X	X	X	Performance tool
BULLMATH	X	X	X	X	Performance tool
CRT	X	X	X	X	Performance tool
MPI					
MDM		X	X	X	Data Mover Module
MPIBULL	X	X	X	X	ONLY for QUADRICS configuration.
MPIBULL2	X	X	X	X	
MPICH_ETHERNET		X	X		ONLY for ETHERNET configuration.
bull_mono_libs				X	
SCIENTIFIC LIBRARIES					
BLACS		X	X	X	
BLOCKSOLVE95		X	X	X	
FFTW		X	X	X	
netCDF		X	X	X	
PETSc		X	X	X	
SCALAPACK		X	X	X	
SuperLU		X	X	X	
INFINIBAND					
INFINIBAND	X	X		X	ONLY for INFINIBAND configuration.

D.3 CLM (Cluster Management) Media Bundles

The Bull CLM media contains the following features.

Bundle Name	Management Node	Compute Node	Login Node	Standalone Node	Description
lptools	X	X	X		
HIGH AVAILABILITY	X				High Availability Feature
ConMan	X		X	X	Console Management
casimir	X				Log Files Analyzing
NsDoctor	X				Node Analyzing
nodechecking	X	X			Node Testing
clusterdb	X				Database
phpPgAdmin	X				Database Management
ganglia	X	X	X		Monitoring tool
TORQUE	X	X	X	X	Batch Management
genders	X	X	X		Cluster Configuration Management
KSIS	X	X	X		Image Building and Deployment
nsctrl	X				Hardware Management
ACT	X				Administration Configuration Tool
ECT					Embedded Configuration Tool
NS Master - HPC Edition	X				HPC Monitoring
pdsh	X	X			Parallel Distributed SHell
STORAGEADMIN	X	X	X		Storage Management
syslog-ng	X	X	X		Log Management
postbootchecker	X	X	X		Nodes verifications
blbs	X				Bull Load Balancing System
SLURM	X	X	X	X	Resource Management

Appendix E. Installing TS4 and TS16 Digiboard PortServers for Linux

This appendix describes how to configure a PortServer and serial lines to access the Linux console and EFI using COM2 EFI.

The following topics are described:

- E.1 *Configuring a Linux Console and a kdb Debugger on NovaScale Clients*
- E.2 *Connecting the PortServer with a Serial Line*

Note:

In this Appendix TS4 or 16 PortServers are sometimes referred to as “PS”.

E.1 Configuring a Linux Console and a kdb Debugger on NovaScale Clients

E.1.1 Boot Option in the elilo.conf file

The boot option in the **elilo.conf** file is:

- if KDB is not used

```
append="console=tty0 console=ttyS1,115200"
```

- if KDB is used

```
append="console=tty0 console=uart,io,0x2f8,115200n8 kdb=on"
```

The output will be available on all peripherals specified in the "console" option. The last one will be **/dev/console**.

Validate a **getty** in **/etc/inittab**, by adding the line:

```
S1:2345:respawn:/sbin/agetty 115200 ttyS1
```

E.1.2 Access with root Login

Update the file containing peripheral on serial lines authorized to connect with root login.

Add the following line in the **/etc/securetty** file:

```
ttyS1
```

E.2 Connecting the PortServer with a Serial Line

Connect the PortServer using one of the commands below:

If the serial cable connects using port 0, run:

```
cu -s 9600 -l /dev/ttyS0
```

If the serial cable connects using port 1, run:

```
cu -s 9600 -l /dev/ttyS1
```

In response, the following message is displayed:

```
Connected.
Switch>
```

To check the configuration, run the command:

```
show config
```

Appendix F. QWERTY/AZERTY Keyboard Comparison

EFI recognizes only the QWERTY keyboards. The following figure shows a QWERTY keyboard. It enables you to enter EFI key sequences using another keyboard, by comparing your keyboard with the QWERTY keyboard.

Moreover you can find layout for most countries on the internet site http://www.translation.net/keyboard.html .

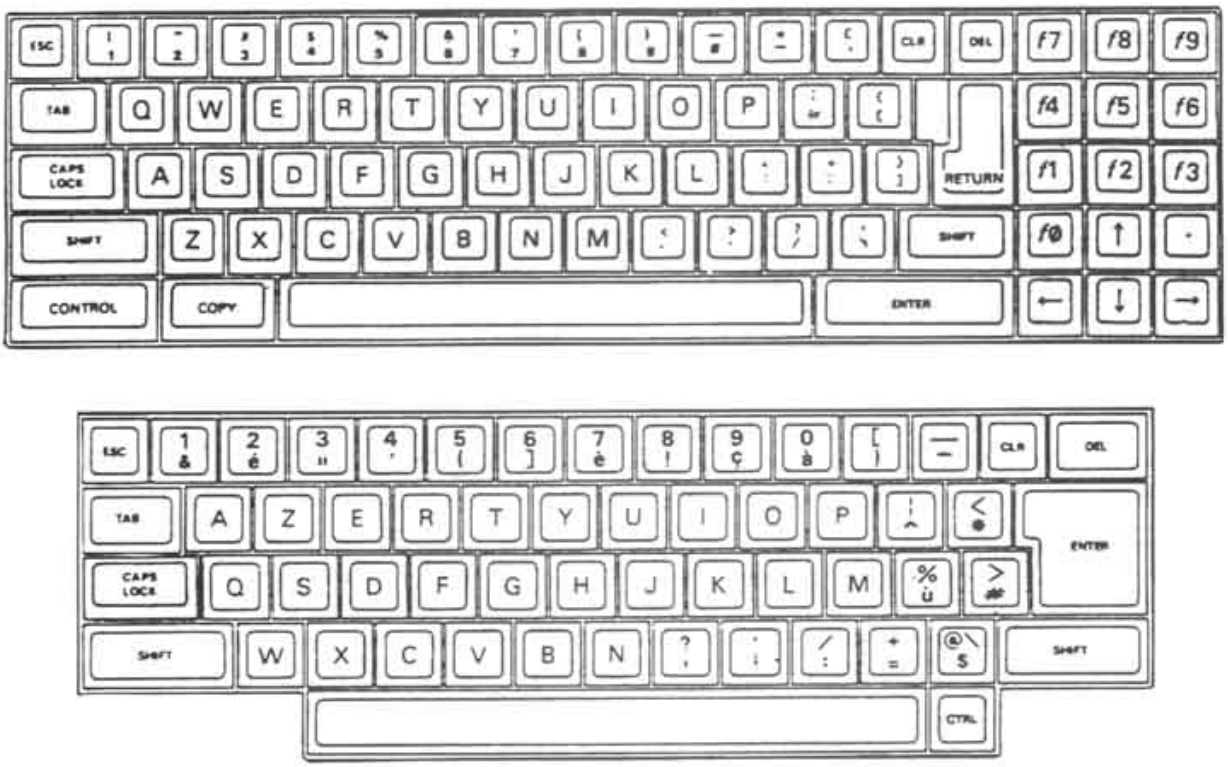

Figure F-1. QWERTY Keyboard Comparison

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

BLAS
Bull Linux Advanced Server

C

CMOS
Complementary Metal Oxide Semi Conductor

D

DDN
Data Direct Networks

DHCP
Dynamic Host Configuration Protocol

DIB
Device Interface Board

DDR
Double Data Rate

E

ECT
Embedded Configuration Tool

EFI
Extensible Firmware Interface

EIP
Encapsulated IP

EMP
Emergency Management Port

EPIC
Explicitly Parallel Instruction set Computing

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

HA
High Availability

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

KDE
K Desktop Environment

KSIS
Utility for Image Building and Deployment

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

L

LAN
Local Area Network

LDAP
Lightweight Directory Access Protocol

LUN
Logical Unit Number

M

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

MDS
MetaData Server

MDT
MetaData Target

MKL
Maths Kernel Library

MPI
Message Passing Interface

N

NFS
Network File System

NPTL
Native POSIX Thread Library

NS
NovaScale

NTFS
New Technology File System (Microsoft)

NTP
Network Time Protocol

NUMA
Non Uniform Memory Access

NVRAM
Non Volatile Random Access Memory

O

OEM
Original Equipment Manufacturer

OPK
OEM Preinstall Kit (Microsoft)

OST
Object Storage Target

P

PAM
Platform Administration and Maintenance Software

PAP
Platform Administration Processor

PAPI
Performance Application Programming Interface

PCI
Peripheral Component Interconnect (Intel)

PDU
Power Distribution Unit

PMB
Platform Management Board

PMU
Performance Monitoring Unit

PVFS
Parallel Virtual File System

PVM
Parallel Virtual Machine

Q

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

R

RAID
Redundant Array of Independent Disks

RMS
Resource Management System (Quadrics)

ROM
Read Only Memory

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

S

SAFTE
SCSI Accessible Fault Tolerant Enclosures

SDP
Socket Direct Protocol

SDPOIB
Sockets Direct Protocol over Infiniband

SDR
Sensor Data Record

SEL
System Event Log

SCSI
Small Computer System Interface

SIOH
Server Input/Output Hub

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

SM
System Management

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

SMT
Symetric Multi Threading

SNMP
Simple Network Management Protocol

SOL
Serial Over LAN

SSH
Secure Shell

T

TFTP
Trivial File Transfer Protocol

U

USB
Universal Serial Bus

UTC
Coordinated Universal Time

V

VDM
Voltaire Device Manager

VFM
Voltaire Fabric Manager

VGA
Video Graphic Adapter

VLAN
Virtual Local Area Network

VNC
Virtual Network Computing

W

WWPN
World – Wide Port Name

Index

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Technical publication remarks form

Title:	HPC BAS4 Installation and Configuration Guide

Reference:	86 A2 28ER 09

Date:	July 2007

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 28ER 09