Bull
HPC BAS3

Installation and Configuration Guide

86 A2 31EG Rev 06

Bull
HPC BAS3

Installation and Configuration Guide

Subject: HPC process for installation and configuration

Special Instructions: Refer to SRB

Software Supported: HPC BAS3 V2

Software/Hardware required: Refer to SRB

Date: November 2005

Bull S.A.
CEDOC
Atelier de reprographie
357, Avenue Patton BP 20845
49008 ANGERS Cedex 01
FRANCE

86 A2 31EG Rev. 06

Your suggestions and criticisms concerning the form, contents and presentation of this manual are invited.
A form is provided at the end of this manual for this purpose.

86 A2 31EG Rev. 06

Preface

Scope and Objectives

The purpose of this guide is to explain how to install the Bull HPC (High Performance Computing) software (Linux® operating system, HPC middleware and distribution) corresponding to "Bull Advanced Server 3V1" for 64 bits NovaScale platforms.

Intended Readers

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

Prerequisites

Structure

This guide is organized as follows:

- Chapter 1 is an overview of the complete procedure.
- Chapter 2 supplies the basic notions of HPC in a LINUX environment with any software. It also gives general information about the hardware configuration of a Bull HPC.
- Chapter 3 supplies the basic notions of storage management for HPC.
- Chapter 4 describes how to install Linux base software.
- Chapter 5 describes how to install commercial tools and other applications.
- Chapter 6 describes how to install Bull Extension Pack for HPC.
- Chapter 7 describes how to install Quadrics Interconnect.
- Chapter 8 describes the process of building reference images, and deploying them on cluster nodes.
- Chapter 9 describes the tasks after preceding installations, in order to make the cluster operational (i.e. specializing the different nodes).
- Appendix A enables you to diagnose some installation problems.
- Appendix B is a comparison between QWERTY and national Keyboard.
- Appendix C gives some recommendation for optimizing the choice of PCI slots for high bandwidth PCI adapters.
- Appendix D is a description of bundles installed during the phase of Linux installation.

- Appendix E describes how to install Digiboard PortServer TS16 for Linux.
- Appendix F is a list of acronyms.

Bibliography

- NovaScale 40xx and NovaScale 5xxx/6xxx documentation.
- Storage Subsystems documentation delivered with the systems.
- Bull HPC BAS3 Administrator's Guide (86 A2 31EM).
- Bull HPC BAS3 User's Guide (86 A2 30EM).
- A Software Release Bulletin (SRB) provides release-specific information and installation instructions (86 A2 31EJ).

Syntax Notation

Text and messages displayed by the system to illustrate explanations are in "Courier New" font. Example:

```
BIOS Intel
```

Text for replies or values to be entered by the user are in "Courier New bold". Example:

COM1

The carriage return or enter key, used to validate an entry, is symbolized by <Enter> where in the context this helps the reader to understand. Example:

COM1<Enter>

Commands to be entered by the user are framed. Example:

```
mount /mnt/cdrom
```

Table of Contents

4. Linux Installation

5. Tools and Applications Installation

6. HPC and Cluster Management CDs Installation

7. Quadrics Interconnect Installation

8. Building Reference Image and Deployment

9. Making the Cluster Operational

A. Error Messages

B. QWERTY Keyboard Comparison

C. Recommendation for PCI Slots Selection

D. Description of Bundles Loaded by Install Process

E. Installation of Digiboard PortServer TS4 and TS16 for Linux

1. Installation Overview

This chapter explains the main steps for installing operating system and environment on a Bull HPC.

1.1 Procedure Overview

For a new Bull HPC platform, the Operating System is preloaded, and you should not have to install it. This document can be used to re-install the system if needed or to update the OS version from a previous one. In such a case, it is the customer's responsibility to save data and software environment before to use this procedure, which installs the OS from scratch, erasing all disk contents. Migration option is not available.

The chart hereafter describes the main steps in the installation of an HPC platform.

Steps order is indicated by the numbers in management and compute nodes colums. It is important to follow this order as some rpms need prerequisites.

For each step the chart indicates in which chapter of this Installation Guide you will find the detail of corresponding procedure.

Installation process *First install management node then compute and I/O nodes*	Management Node	Compute or I/O Node	Remarks	Pointer in installation guide
Cluster hardware Configuration	**1**	**9**	Including Ethernet switch, Asynchronous network and PortServer, Quadrics interconnect, storage resources	**Chapter 2**
Storage hardware Configuration	**2**	**10**		**Chapter 3**
Install Linux Operating System	**3**	**11**		**Chapter 4**
Install Commercial tools and applications (Intel compiler, debug, trace, …)	**4**	**12**	Possible after copying Quadrics rpms	**Chapter 5**
Install Quadrics Interconnect	**5**	**13**	Possible after Deployment (Quadrics not use for deployment)	**Chapter 7**
Install HPC Cd (Bull delivery) and Cluster Management Cd (Bull delivery)	**6**	**14**		**Chapter 6**
Nodes Configuration after Linux Installation	**7**	**15**	Send Information on reference node	**Chapter 4 § 4.5**
Configuration after Bull HPC and Cluster Management CD installation	**8**	**16**		**Chapter 6 § 6.4, 6.5, 6.6, 6.7, 6.8, 6.9, 6.10, 6.11.**
9. Build reference image		**17**	Result of commands run on management node	**Chapter 8**
10. Reboot, Deployment on all other nodes and make cluster operational		**18**	Result of commands run on management node	**Chapter 8 § 8.3**
11. Make the cluster operational	**19**	**20**	Specialize nodes: Compute, management, I/O	**Chapter 9**
12. Reboot and Test conformity	**21**	**22**		

2. Cluster Configuration

2.1 Introduction

A cluster is an aggregation of identical or very similar individual computer systems (each system in the cluster is a "node").

The cluster systems are tightly-coupled using dedicated network connections such as high-performance, low-latency interconnections.

All systems in a cluster share common resources, such as storage, over dedicated cluster file systems.

Cluster systems are generally on a private network so that each system can trust the other systems in the cluster, and avoids the need to manage each of them individually, or to start jobs manually on each node in the cluster.

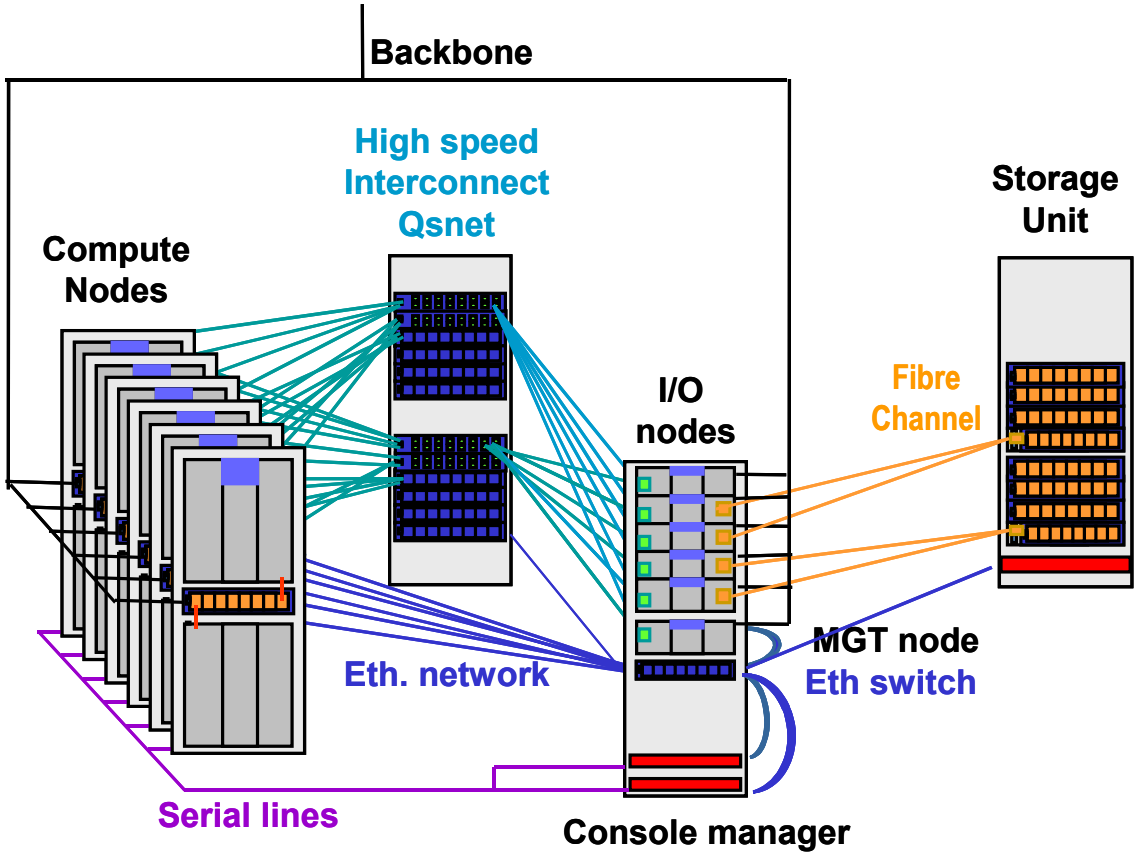
The Message Passing Interface (MPI) is used in order to allow programs to be run across all nodes.

2.2 Hardware Configuration

A typical cluster infrastructure is composed of::

- **Compute nodes** for intensive calculation.
- **Input/Output nodes** to store data in storage units.
- **Management node** to administrate, manage and exploit the cluster.
- **High speed interconnect** switch and boards to transfer data between compute nodes and I/O nodes.
- **Administration Network** including Ethernet and serial networks which are used for cluster management and maintenance.
- **Backbone** is the link between HPC and external word.

This kind of infrastructure is shown in the figure below.



2.3 General NovaScale Node Description

2.3.1 Example of Novascale Configuration

2.3.2 System Console PAP

The Novascale is 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 (Platform Management Board) via the LAN. The PMB runs VxWorks and MAESTRO agents. You are strongly encouraged to read the *"NovaScale administration guide"*.

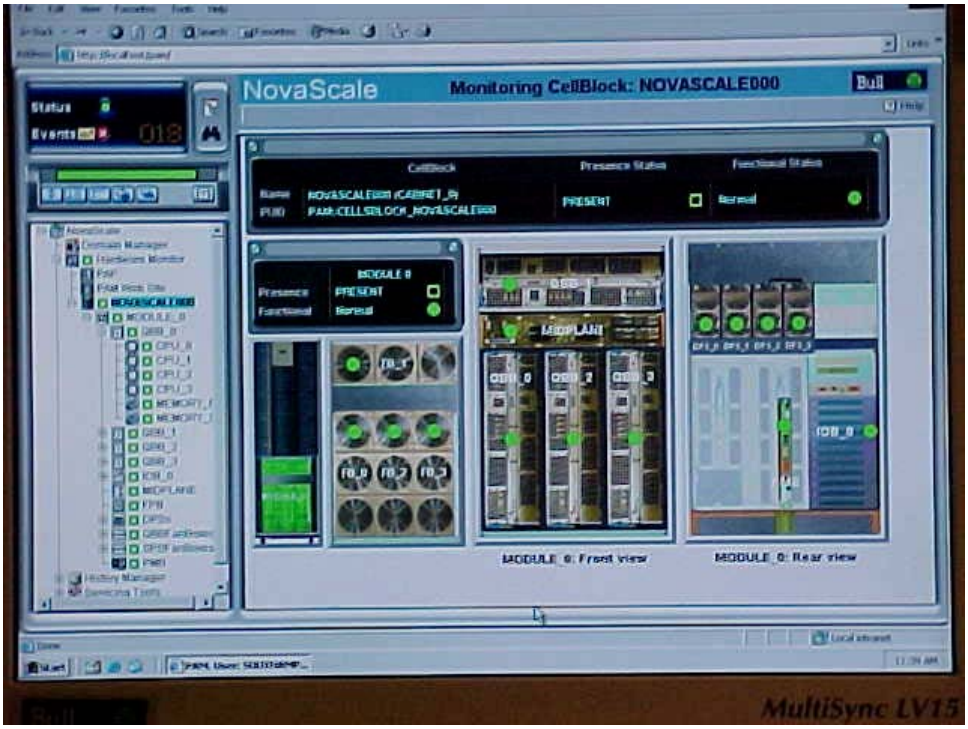
You will use the secured PAM web-based interface to operate, monitor, and configure the Novascale. Once the Windows server 2003 software is booted, you will be prompted to supply a user name and a password to open a Windows 2003 session. The system MUST be started for the first time with the following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

User Name = Administrator
Password = administrator

From the Windows-server 2003 desktop, double-click the Microsoft Internet Explorer icon to launch the Web-based administration tools. These tools will allow you to:

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

The PAM user interface is divided into three main areas within the browser window: A status panel, a PAM tree and a control panel, allowing users to check the system status at a glance.



2.3.3 Keyboard Video Mouse (KVM)

The KVM (Keyboard Video Mouse) switch allows to control the hosts from the front console. You can add other hosts to the KVM using shipped cables.

From the console, you can switch to another system linked to the KVM using a « Ctrl+Ctrl » sequence key. . (Ctrl then Ctrl in a short time)

The system console may also host storage management tools for Bull FDA storage system if no other system is available for this.

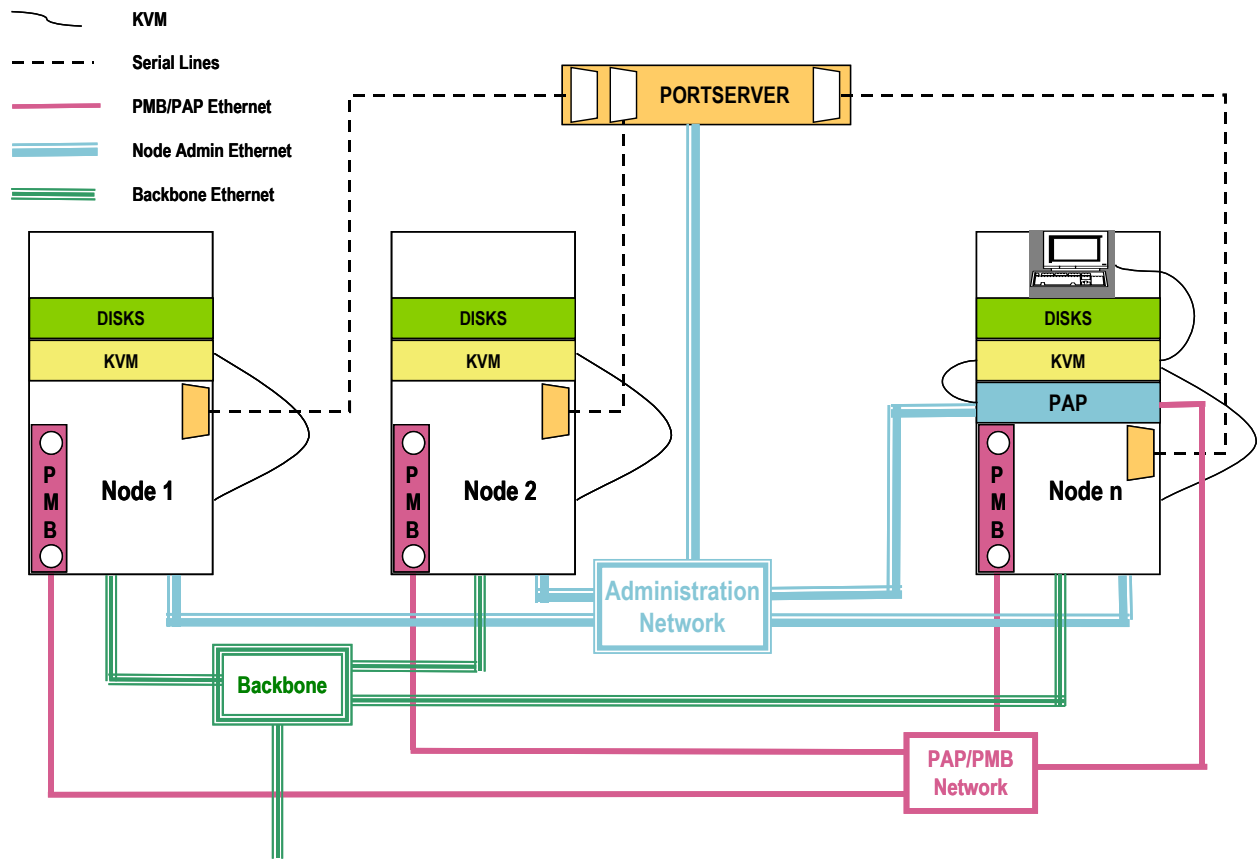
2.4 Administration Network and Backbone

The systems administration needs different kind of network support (ethernet or serial) according to:

- the status of the system (down, running, frozen)
- the nature of the request (hardware, software), we want to send.

The **Administration network** includes **two** separate **Ethernet networks** (First one for general software management of the cluster, the other for it's hardware management) and **one Serial network**. These networks concentrate on the management node or on the PAM all information to control and manage the cluster.

The figure below describes a general scheme of this administration network and backbone



2.4.1 Serial Network

It is impossible in a high density rack of computer to have a graphical console with keyboard and mouse for each node. We use serial ports concentrator on an ethernet line.

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

- It offers consoles management support to all equipment (nodes, disk units, switches, …) from the management node, and is used when Linux is no longer running.
- It provides facilities to get dumps on ethernet administration network when the system is frozen. It's also used to access firmware or to debug the system.
- All the equipment of the cluster supporting a terminal server are connected to the serial line network. Each serial line (asynchronous RS232c) is connected from COM1 to the serial multiport concentrator like Portsrever from Digiboard.
- The PortServer itself is connected to the administration network via Ethernet.

Installation and configuration of this PortServer is fully described in Appendix E.

2.4.2 PAP/PMB Network

PAP/PMB network interfaces all PMB and PAP.

The Platform Administration Processor (PAP) is used to manage (power on or power off a node, …) and get hardware information (state of the processor, temperature, …) of all the hardware components of the cluster. using a separate ethernet network. One PAP can control from 1 to 16 nodes.

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

2.4.3 Actual Administration network

Actual Administration network. This ethernet network allows to manage operating systems, middleware and applications from management node.

This network joins all the eth0 native ports of each node through a 100 Mb/s network .It is also connected to the PAP.

This network has no links to the other networks and it includes 100 Mb/s ethernet switch(es)

2.4.4 Backbone

The **Backbone** is the link of the cluster with the external world

- This network links all the eth1 ports of each node and external networks through a 100 Mb/s network including ethernet switches.

2.4.5 Ethernet Network and Switches

Depending on the model, the switches can be managed:

- directly by Ethernet
- or through a serial line allowing to configure network management over Ethernet. For this you have to plug:
  - a serial line on the switch
  - and all the Ethernet cable defined to be plugged to this switch.

Useful parameters can be set up at this step like multicast management, ARP management and Fast Spanning Tree. Check manufacturer documentation to define which options you have to set on your device.

2.4.6 Main Console and Hardware Management Commands

Hardware management

Hardware management administration and maintenance tools give you immediate insight into system status and configuration. You will use them to operate, monitor, and configure your server. You can use:

- PAM commands available on the PAP platform. For details about PAM commands see: *"Bull - NovaScale 5xx0 & 6xx0 - User's Guide"* (réf 86 A1 94EM).

- NS-commands installed on the management node. These commands invoke the PAM using administration network. They are described in the *Bull HPC BAS3 Administrator's Guide* (86 A2 31EM).

Console management

Conman is a console management program designed to support a large number of console devices and simultaneous users. It currently supports local serial devices and remote terminal servers (via telnet protocol)

Conman and advantages of conman on a simple telnet connection are described in the *Bull HPC BAS3Administrator's Guide* (86 A2 31EM)

Sometimes, accessing the console is the only way to diagnose and correct software failures like kernel debugging: Kdb is active before ethernet driver loading, thus, usable only with an asynchronous line.

Console allows:

- Accessing to firmware shell (BIOS/EFI) in order to get and modify NvRAM information, choose boot parameters: kernel, disk on which node has to boot, boot on a CDROM to make an OS installation
- Boot monitoring.
- Boot interventions like interactive file system check (fsck) at boot.
- Telnet sessions

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

where <serial IP address> is the IP address dedicated to the node on the portserver

2.5 High Speed Interconnect

We describe in this paragraph, the network including Quadrics Interconnect and Elan4, providing data transfer between the nodes of the cluster.

High speed interconnect is using QsNet^{II} technology from Quadrics for interconnection. Effective bandwith of 900 MB/s (of user space to user space), latency lower than 5 µs and possibility to interconnect up to 4096 nodes are the main characteristics.

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

For Elan4 (QM-500) cards:

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

For QsNet^{II} switch (QS5A):

- Connect Ethernet cable to the management board.
- As described before, all nodes with an Elan4 card may have been plugged on 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 plugged directly on the management board. (Also can be set remotely from the console concentrator program.)

2.6 Typical Types of Nodes

2.6.1 Management Node

Management node concentrates on one node all control and management functions.

The Management node may be a NS4040 (or at least a NS4020) and can 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 purpose you will also need a set of screen and keyboard/mouse. Another ethernet connection must be setup to connect to the PortServer for asynchronous connection to nodes terminals. An Elan4 (QM-500) card (or more) may be also plugged, if EIP (Encapsulated IP) or I/O will be used on management node.

The management node stores lots of reference data, and operational data (e.g. for RMS). It is recommended to store these 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 to select the best PCI slots for optimum performances.

2.6.2 Compute Nodes

The **Compute nodes** take benefit of the Itanium2 cpu power, capable of floating point calculation. The Itanium2 has 3 levels of cache and EPIC architecture to deliver very impressive linpack results.

Quadrics Elan4 card(s), serial line to the PortServer and ethernet link to management LAN must be present on the node.

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

If present, the storage system(s) must be configured prior to the configuration of the file system used on the nodes. Refer to Appendix C to select the best PCI slots for optimum performances.

2.6.3 I/O Nodes and Storage Units

Input/Output nodes are dedicated to store and retrieve data.

The Input/Output node is similar to a compute node but, due to its functions this node includes large amounts of storage, , connected through fiber channels links. The storage systems are DDN S2A systems.

The storage system(s) must be configured prior to the configuration of the global file system used on the I/O nodes.

Read also Appendix C to select the best PCI slots for the SCSI and Fiber Channel HBA in order to obtain optimum performances.

3. Storage Systems

3.1 Introduction

Storage Systems can be used for:

- Storing on every node where it- is necessary, (compute, I/O or management nodes) operating system, middleware, applications running on the nodes as well as parameters and administration data.
- Storing and retrieving large amount of data, input or results of the user's applications.

The HPC cluster is delivered with various kinds of storage resources. The **Storage system** can be:

- internal SCSI disks for nodes,
- SCSI JBODs. They are very similar to internal SCSI disk, except that they are packaged in dedicated enclosures with their own power supply: example Bull Storeway Cost Effective JBOD SJ-0812 storage system,
- External RAID storage systems, with SCSI connections to the nodes: example Bull Storeway Cost Effective RAID 8 slots SR-0812,
- External RAID storage systems with Fiber Channel connections to the nodes: example Bull Storeway FDA or Data Direct Networks (DDN) S2A.

This chapter concerns installation and configuration of storage systems in the following environment:

- Hardware: NovaScale 4xxx, NovaScale 5xxx/6xxx,
- Linux distribution: Bull Advanced Server version 3 (BAS3),
- Fibre Channel connections: Emulex LP 9802,
- SCSI connections (LSI 22320-R).

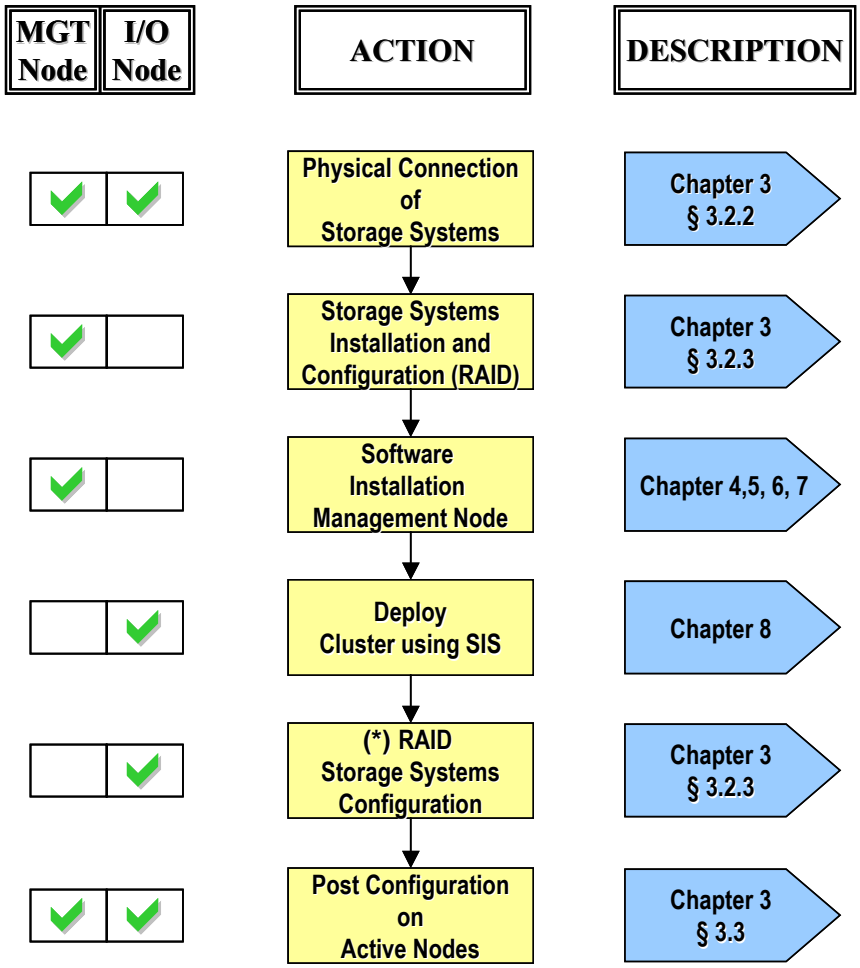
3.2 Connecting and Installing Storage Systems

The technology used for storage system connection, and the configuration utilities depend on the selected storage system. Please refer to the documentation provided with your storage system for more information about "how to connect and configure".

The storage system(s) must be connected and configured prior to the configuration of the global file system used on the I/O nodes.

3.2.1 Global Process for Connecting and Installing Storage Systems

The global process for connecting and installing storage systems is described below:



(*) Generally RAID Storage Systems used for applications data as well as global file systems are configured after deployment.

3.2.2 Connection of the Storage Systems

In order to connect storage sytem to the HPC nodes:

- Please refer to the documentation provided with your storage system.
- Read also Appendix C to select the best PCI slots for the SCSI and Fiber Channel HBA in order to obtain optimum performances.

The procedure is usually the following :

- Connect the hosts ports of the storage systems to the required nodes (management node, IO nodes, ...).
- Connect the serial ports of the storage systems to the PortServer (if this option has been selected). Else, temporarily connect a terminal (or terminal emulator) to the serial port.
- Connect the ethernet interfaces to ethernet administration network of the cluster (if any), and configure the network parameters for the Ethernet ports (usually through the serial port). Update the DNS or the /etc/hosts file of the management station.

3.2.3 Installation and Configuration of the Storage Systems

Software Packages for Storage Systems

Some RAID storage system need software packages to configure their various features. Internal disks and JBODs (such as the Bull Storeway Cost Effective SJ-0812) do not require specific configuration.

The next table summarises the software requirement for various supported storage systems.

Bull Storeway Cost Effective SJ family External SCSI Cabinet (JBOD)	No software installation requirement..
Bull Storeway Cost Effective SR family External SCSI Cabinet with RAID	No software installation requirement. Use the serial port connection to access to a menu driver management tool.
Bull Storeway FDA family External Fiber Cabinet with RAID	A software packages is delivered with the product. This package must be installed on a windows system, such as the PAP of a NovaScale server. Check the documentation supplied with the storage system for guidance. The windows system must have a network access to the ethernet port of the Storeway FDA system(s).
Data Direct Network S2A appliances	The S2A appliances can be configured through the serial port or ethernet port using a command line interface. An optional software package provides a graphical user interface. This GUI should be install on a windows station, such as the PAP for NovaScale servers.

Planning Storage Resource Configuration

For all the storage resources, a planning phase is required to define intended usage: Which node will use which resource, which application? What are the requirements for space, data protection and performance?

The external RAID storage systems include a level of virtualization. The physical HDDs (Hard Disk Drives) are not directly visible from the attached nodes. It is necessary to use dedicated management tools (described in the next paragraph) to create the LUN that will be presented to the node.

The next step is to use Linux storage management tools such as disk partitioning or LVM2 in order to configure the internal SCSI disk, external SCSI JBODs and LUNs in external RAID storage system.

Guidelines for Storage Configuration.

The management node may host very different types of data, and thus the configuration of the storage system may be optimised for each specific usage.

The IO nodes support a global file system. The best results are obtained with an homogeneous configuration of all the LUNs within storage systems: same RAID type, same LUN size, resource balancing between RAID controllers and host ports, ...

RAID Storage Systems Configuration

The configuration tools are specific for each model of storage system. Please refer to the documentation delivered with the storage system for detailed installation and configuration procedures.

The next paragraphs contains guidelines to configure the various storage systems supported with Bull BAS.

Please refer to the documentation of your storage system to understand configuration options.

The process is usually the following:

- Group disk drives in RAID group
- Select hot spare disks
- Define LUNs within RAID groups
- Bind RAID groups or Luns to RAID controlers or host ports
- Configure acces rights for attached servers.

Bull Storeway Cost Effective SR-0812

The Storeway Cost Effective RAID 8 slots SR-0812 storage system must be configured using a menu driven tool accessible through the serial port.

The following steps must be performed:

- Creation of arrays. This includes the selection of groups of HDDs and the definition of the associated RAID protection.
- Definition of partitions within arrays. The partitions will be discovered by the servers as SCSI disks (**sd** devices).

This storage system can be used as a boot system. In that case, two HDDs are reserved for the operating system. They are configured in RAID-1. For such a usage, refer to the *Hardware Installation Guide* of the server.

The next table presents some recommended array configurations for the SR-0812 storage system.

Bull SR-0812	**1 disk drawer** • 1 RAID 5 (6D+1P) + 1 Spare disk = 8 HDDs **or** • 2 RAID 5 (2D + 1P) + 1 Spare Disk = 7 HDDs **2 disk drawers** • 2 RAID 5 (6D+1P) + 1-2 Spare disk = 15-16 HDDs **or** • 3 RAID 5 (3D + 1P) + 1-2 Spare Disk = 13-14 HDDs

Bull Storeway FDA series

The Storeway FDA storage systems can be configured using a graphical tool available on a Microsoft Windows station. Usually, the iStorage Manager is installed on one of the PAP of the cluster.

The main configuration steps are the following:

- Creation of RANKs. This includes the selection of groups of HDDs and the definition of the associated RAID protection.
- Selection of spare HDDs.
- Configuration of LD (Logical Disks), which is the resource that will be discovered by the attached nodes.
- Configuration of access control (if the license is available). Port zoning is recommended when there is a point to point connection of nodes to the singlet's host ports.
- Formatting LUNs.

The next table presents some recommended array configurations for the Storeway FDA family of storage systems.

Storeway FDA 1300/2300	2 RAID 5 (6D+1P) + 1 Spare disk per disk drawer = 15 HDDs 2 or 3 RAID 5 (4D+1P) per disk drawer = 10 or 15 HDDs
Storeway FDA 1400/2400/2800	3 RAID 5 (4D+1P)

DataDirect Networks S2A 8500.

The S2A can be configured using an interactive command line interface available on the singlets. The administrator should connect to the singlet using telnet (if the network has been setup during the installation phase).

The RAID configuration is fixed to groups of 8 data HDDs and 1 parity HDD (8+1). Spare disks are also fixed.

The main configuration steps are the following:

- Definition of LUNs using "LUN ADD" command.
- Definition of zoning to enable nodes to access to the LUNs using "ZONING" command. Port zoning is recommended when there is a point to point connection of nodes to the singlet's host ports.
- Configuration of cache using "CACHE" command. Write back offers the highest performances, but data may be lost in case of singlet failure. It is also recommended to deactivate read prefetch, and to choose 1 MB for the cache segment size.
- Formatting LUNs.

3.3 Post Configuration on Active Nodes

The next operations must be performed on the nodes attached to the storage systems when they are up and running, with a Bull BAS installed.

It is recommended to reboot the nodes attached to the storage systems to discover the newly configured disks.

1. Check that all the HBA are up and running:

```
cd /proc/scsi
ls
```

There should be one directory per adapter family:

- `aicxxxx` for Adaptec SCSI HBA,
- `mptscsi` for LSI SCSI HBA,
- `lpfc` for Emulex HBA

```
ls /proc/scsi/[aic*|mpt*|lpfc*]/*
```

There should be one file per HBA (or native SCSI bus) on the system. File names are numbers.

2. Check that all the disks are discovered:

```
cat /proc/scsi/scsi
```

All the disks should be listed

3. Check that you have a read access to each disk:

```
dd if=/dev/sd<x> of=/dev/null count=100
```

4. Linux Installation

Once the hardware is configured the next step is to install the Linux OS to setup the management node. The same operation must be duplicated for each type of NovaScale server in order to create a reference image for this type of node in the cluster.

This chapter explains how to install the Linux operating system on a Bull HPC node.

This consists of installing the BAS distribution including the Kernel.

4.1 Before Starting Installation

Before starting installation, please, remember:

- The installation procedure is autonomous. Only a few manual operations are required; no tools are needed.
- Check that you have all the required elements described in the SRB (Software Release Bulletin).
- The whole operation takes about 20 minutes.
- The Operating System must be installed on SCSI disk or SCSI storage unit. For Novascale server hosting system disks or SCSI Raid storage units, please refer to *"Novascale xxxx Installation Guide"*.
- If data on a disk have to be saved, the install choice should be "manually partitioning", because other installation will erase all the information on sda and/or sdb, sdc.
- Manual changes to the manufacturing default BIOS settings are possible during installation. The console may also be redirected when installing machines in a cluster.
- Default for page size in kernel is 64 K.
- If it is not possible to boot on a node, think to disconnect the portserver if this portserver has never been configured.
- In some cases you have to verify or configure EFI Menu. The way to configure EFI menu is full y described in
  Bull - NovaScale 5xx5 & 6xx5 - User's Guide - réf 86 A1 41EM
  Bull - NovaScale 5xx0 & 6xx0 - User's Guide - réf 86 A1 94EM

4.2 Procedure Overview

Following is the list of tasks to perform for the installation of Bull Advanced Server 3 (BAS3).

You have to set up (install and configure) the management node first then do the same for compute and I/O reference nodes. From the management node get reference images and deploy them on corresponding nodes.

Phase 1: Boot the system:

- Boot from the OS CD-ROM.
- Setup information for an automatic installation procedure (or manually).
- Perform the installation and reboot the system.

See details in *4.3 Phase 1: Boot the System*.

Phase 2: Configure software environment:

- Select peripheral options and language.
- Configure network and utilities.

See details in *4.4 Phase 2: Configure or Reconfigure (if needed) Software Environment*.

Phase 3: Nodes configuration after Linux Installation:

- Manual operations of Configuration necessary, in order to make nodes available before other operations.

See details in *4.5 Phase 3: Nodes Configuration after Linux Installation*.

4.3 Phase 1: Boot the System

4.3.1 Initializing the Loading

- Power up the machine (you will hear the fans).
- Switch on the monitor (if you have not already done).
- Insert the *"Bull Advanced Server 3"* CD-ROM *(CD1/6)* into the CD-ROM 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 CD-ROM is not inserted during this phase, put it into the drive and under EFI, execute the command

```
map -r
```

- After the various startup phases (Bios, SCSI detection, etc), the screen disappears and the EFI banner is displayed. The output looks like the following:

```
EFI version 1.10 [14.59] Build flags: EFI64 Running on Intel(R)
Itanium Processor EFI_DEBUG
```

- From this point, if the CD has been inserted, 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 press **Enter**

- Choose the installation mode you want to run:

```
1.  BAS3V2 Compute (manually partitioning)
    You will be prompted for the partitioning configuration, bundles
    for Compute node

2.  BAS3V2 Compute 1 disk (all on sda)
    Automatic partitioning on 1 disk, bundles for Compute node

    Sda disk will be erased, the partitioning is :

      /boot/efi     512 Mb

      swap          2 Gb

      /             All Free space

3.  BAS3V2 Management (manually partitioning)
    You will be prompted for the partitioning configuration, bundles
    for Management node
```

4. BAS3V2 Management 1 disk (all on sda)
   Automatic partitioning on 1 disk, bundles for Management node

   Sda disk will be erased, the partitioning is :

   /boot/efi 512 Mb

   swap 2 Gb

   / All Free space

5. BAS3V2 Standalone (manually partitioning)
   You will be prompted for the partitioning configuration, bundles for Standalone node

6. BAS3V2 Custom
   All steps of the installation will be manually set

7. BAS3V2 Rescue
   Allow to boot in rescue mode

8. BACK TO EFI MENU
   Go back to the EFI menu

- One or more supplementary CD can be asked, the CD drive opens and a window is displayed on the console with the message:
  "Please Insert disc n to continue":
  Insert CD *n* and click **OK** (or RC) to continue installation

- After about 20 minutes (total time) the system congratulates you:
  "installation is complete"
  and asks if you want the machine to restart the system.

- Don't forget to remove last CD.

- If you choose to reboot, the system is loaded from the new disk installation.

Refer to *Troubleshooting Errors* section in case of problems.

4.3.2 Automatic Installation

This paragraph refers to choices 2, 4 of installation mode.

- In these choices, all data on the 'sda' disk will be deleted during installation.

4.3.3 Manually Partitioning

This paragraph refers to choices 1, 3, 5, 6 of installation mode.

Only the disk partitioning configuration is requested. The configuration of Linux Partitions are defined by the choice of physical disk, size and type of logical

partitions with menu.

What follows describes the visible part of the automatic installation procedure:

- The Choose a "***Disk Partitioning Setup***" appears.
  Select Automatically or Manual partitioning
  Remove or **Keep all partitions**
  New / Edit / Delete for required partitions

  In case of "***Automatically selection***", you have to modify proposed partitions before confirmation.

- ***Installing Package*** appears.
  After complete installation the machine reboots in order to load the system from the new disk installation.

Here is a partitioning example for a management node (with 3 disks) :

/boot/efi	512 Mb	=> sda
/	15 Gb	=> sda
/var	All Free space	=> sda
/tmp	10 Gb	=> sdb
/home	All Free space	=> sdb
swap	All Free space	=> sdc

Here is a partitioning example for a compute node (with 3 disks) :

/boot/efi	512 Mb	=> sda
/	10 Gb	=> sda
/var	All Free space	=> sda
/tmp	All Free space	=> sdb
swap	All Free space	=> sdc

4.3.4 Custom Installation

This paragraph refers to choice 6 of installation mode.

What follows describes the visible part of the automatic installation procedure.

- The Choose a "***CDROM Found***" appears.
  To begin testing the CD media before installation press OK.
  Choose **Skip** to skip the media test.

  OK is the default option.

  Click (in graphic mode)
  or Press **Enter** and **Next** for continue.

- The Choose a ***Language*** screen appears.
  English is the default language. You can choose another language.

  Press **Enter** and **Next** for continue

- The Choose a ***Keyboard Type*** screen appears.
  us is the default keyboard type. You can choose another type.

  Press **Enter** and **Next** for continue

- The Choose a ***Mouse Configuration*** screen appears.
  Wheel Mouse (PS/2) is the default type.

  Wheel Mouse (USB) is the type hardware present.

  Press **Enter** and **Next** for continue

- The Choose a "***Disk Partitioning Setup***" appears.
  Select Automatically or Manual partitioning
  - Remove or **Keep all partitions**
  - Use: New / Edit / Delete for required partitions

  Next for continue

- The Choose a "***Network Configuration***" appears.
  Select Firewall (default) or No Firewall: No Firewall is current type.
  - IP, Hostname, Router, Name Server are defined here.

  Next for continue

- The Choose an "***Additional Language***" appears.
  Only English is selected, add other language if necessary.

  Next for continue

- The Choose a "***Time Zone Selection***" appears.
  Select Local Time Zone required.

  Next for continue

- The Choose a "***Set root passwd***" appears.
  define root password.

  Next for continue

- The Choose a "***Package Group Selection***" appears.
  select required Packages.

  Next for continue

- ***"Installing Package appears",***
  then system is performing post install configuration, and reboots.

- The Choose a ***"Graphical Interface (X) Configuration"*** appears.
  select the "ATI Mach 64" (selected by default).

  Next for continue

- The Choose a ***"Monitor configuration"*** appears.
  Select your monitor type or select the "Generic CRT/LCD Display", monitor 1024*768 if you don't know it.

  Next for continue

- The Choose a ***"Customize graphical tool"*** appears.
  Choose graphical or text mode.

  Next for continue

After complete installation the machine reboots in order to load the system from the new disk installation.

4.3.5 Warning Messages in install.log

The **/root/install.log** file may contain some warning messges:

```
Installing kernel-2.6.7-B64k.2.1.ia64.
grubby fatal error: unable to find a suitable template
```

or:

```
Installing gstreamer-0.6.0-5.i386.
error: %post(gstreamer-0.6.0-5) scriptlet failed, exit status 127
```

Do not pay attention to these messages, which are of no consequence.

4.3.6 Installed Bundles

The list of **installed bundles** is related in **README-fr** or **README-en** file at the root of installation CD N°1.

A description of each bundle is done in appendix D.

4.4 Phase 2: Configure or Reconfigure (if needed) Software Environment

4.4.1 If you are in text mode

Keyboard: You can change your keyboard configuration with the command:

```
loadkeys <lang>
```

where lang is described in `/lib/kbd/keymaps/i386/` file.

For example to have a French keyboard configuration, type `loadkeys fr`, for qwerty configuration type `loadkeys us` or `loadkeys uk`.

Network: To modify your network configuration, you have to inform the files `ifcfg-eth[1-99]` located in `/etc/sysconfig/network-scripts`.

You can see an example of configuration further in chapter *"4.5.Phase 3: Nodes Configuration after Linux Installation"*.

Language: To modify the system language, you have to modify the parameter **LANG** in file `/etc/sysconfig/i18n` or in your `$HOME/.i18n` if it exists. Values are like **'en_US.UTF-8'**, 'fr_FR.UTF-8' …

You can find a list of supported language by typing:

```
echo $SUPPORTED.
```

On a automatic installation, supported languages are French and English (default).

4.4.2 If you are in graphical mode

Keyboard: You can change your keyboard configuration with the command:

```
redhat-config-keyboard
```

or by clicking Main menu (green tree)->System Settings->Keyboard.

Mouse: You can change your mouse configuration with the command:

```
redhat-config-mouse
```

or by clicking `Main menu (green tree)->System Settings->Mouse`.

Display: You can change your display configuration with the command:

```
redhat-config-xfree86
```

or by clicking `Main menu (green tree)->System Settings-> Display`.

Language: You can change your language configuration with the command:

```
redhat-config-language
```

or by clicking `Main menu (green tree)->System Settings-> Language`.

Network: You can change your network configuration with the command:

```
redhat-config-network
```

or by clicking `Main menu (green tree)->System Settings-> Network`.

When the installation is completed, in a graphical mode, after the 1st reboot you will be prompted to configure some parameters manually, as the date or adding users …

4.5 Phase 3: Nodes Configuration after Linux Installation

After operating system installation you have to configure different parameters on the nodes in order to be operational. The nodes can be pre-configured only by a manual way and require a system administrator with basic knowledge in Linux and TCP/IP networks.

Notes:

1. After the installation you have at your disposal two user logins:
   user: root ; *password*: root
   user: linux ; *password*: linux
2. At this point make sure for systems connected to a RAID storage that it has been configured.

4.5.1 Management Node

This section includes:

- Network definition
- Users, passwords and groups definition
- ssh keys generation
- ssh user creation.

Note: dhcp configuration has to be done just before image deployment (see *Prepare the Image Server*, in the *Building Reference Image and Deployment* chapter in this guide).

1. Set the hostname in file: /etc/sysconfig/network
   ```
   HOSTNAME=ns0
   ```

2. Assign an IP address to the first network interface (ethernet - eth0) by setting it in /etc/sysconfig/network-scripts/ifcfg-eth0 as in example:

   ```
   DEVICE=eth0
   IPADDR=172.16.12.1
   BOOTPROTO=static
   BROADCAST=172.16.12.255
   NETMASK=255.255.255.0
   NETWORK=172.16.12.0
   ONBOOT=yes
   TYPE=Ethernet
   ```

3. The next interface (eth1) may be used for connection to an external network and could be configured as following for a network IP 10.x.x.x, for that you have to change the following scripts /etc/sysconfig/network-scripts/ifcfg-eth1 by:

   ```
   DEVICE=eth1
   IPADDR=10.10.10.1
   BOOTPROTO=static
   BROADCAST=10.10.10.255
   NETMASK=255.255.255.0
   NETWORK=10.10.10.0
   ONBOOT=yes
   TYPE=Ethernet
   ```

   where 10.10.10.1 is an IP address that provides visibility outside of the cluster internal network. If you have an external gateway you may add it in this configuration file using the GATEWAY keyword.

4. You can also configure the DNS for getting access to external domains by adding entries in /etc/resolv.conf by example:

   ```
   nameserver 10.10.10.53
   domainname mydomain.com
   ```

   where:

   10.10.10.53 is the IP of a name server on the local area network.
   mydomain.com is the name of the local domain.

5. Configure manually the nodes that will be in the cluster or at least the nodes that will serve as reference for the deployment (one compute node and one I/O node minimum). For that you have to edit /etc/hosts and add all required nodes:

```
127.0.0.1 localhost.localdomain localhost
172.16.12.1   ns0   rmshost
172.16.12.2   ns1
172.16.12.3   ns2
...
```

   Note: rmshost alias MUST be appended to the line ending with the management node name (ns0).

6. Create a directory for users homes, here named "/home_nfs":

```
mkdir -p /home_nfs
```

7. Share new home directory through NFS by adding corresponding lines in /etc/exports (you can limit the modification to the nodes that will serve as reference for the deployment but you can also add all nodes which are planned to be deployed):

```
/home_nfs      ns1(rw,no_root_squash,sync,no_subtree_check)
 ...           nsX(rw,no_root_squash,sync,no_subtree_check)
```

8. Create a directory that should contain shared programs or data over the cluster, defined name "/opt/envhpc":

```
mkdir -p /opt/envhpc
```

9. Export previous path to the entire cluster in /etc/exports (same note as before, you can add all nodes or only the nodes that will serve as reference for the deployment):

```
/opt/envhpc    ns1(rw,no_root_squash,sync,no_subtree_check)
 ...           nsX(rw,no_root_squash,sync,no_subtree_check)
```

10. If Platform LSF is present share binaries files through NFS in /etc/exports:

```
/usr/share/lsf   ns1(rw,no_root_squash,sync,no_subtree_check)
 ...             nsX(rw,no_root_squash,sync,no_subtree_check)
```

11. Export these NFS shares to the cluster hosts:

```
exportfs -rv
```

12. Restart the network:

```
service network restart
```

13. If you use LSF software add the lsf user (name lsfadmin):

```
groupadd -g 1502 lsf
useradd -u 1502 -g lsf -c "LSF Admin" -s /bin/sh -d /home_nfs/lsf lsfadmin
```

14. Generate SSH keys for root account:

```
/usr/bin/ssh-keygen -b 1024 -t rsa -N "" -f /root/.ssh/id_rsa
/usr/bin/ssh-keygen -b 1024 -t dsa -N "" -f /root/.ssh/id_dsa
```

SSH keys are generated in the rsa or/and dsa mode in order to be compatible with the distant protocol.

15. Install the CDs **in the following order**:

1) Intel compilers
2) Quadrics software
3) Bull extension for HPC
4) Bull extension for cluster management
5) Lustre software.

4.5.2 Compute or I/O Nodes

After operating system installation on management node, you have to build reference images for some nodes (generally one per type of node in the cluster). These images will be used in the deployment when they are achieved.

Important: After having installed Linux on these reference systems don't forget to install all environmental software including Intel compilers, Quadrics software, Bull extension forHPC, Bull extension for cluster management and Lustre software.

Then you have to configure some specifics parameters on these nodes, necessary step in the workflow to make them ready.

1. **The installation of the operating system has configured the first network interface (eth0) in DHCP mode, so you need a DHCP server on the management node configured with the right IP and MAC in its configuration file. At this time when the node reboot after OS installation you get the right IP parameters.**

   Complete /etc/sysconfig/network-scripts/ifcfg-eth0 as in example:

```
DEVICE=eth0
BOOTPROTO=DHCP
ONBOOT=yes
TYPE=Ethernet
```

Configuration of dhcp service is fully described in section Chapter 8, section *Prepare the image server*.

2. Generate SSH keys for root account (Private and public key)

```
/usr/bin/ssh-keygen -b 1024 -t rsa -N "" -f /root/.ssh/id_rsa
/usr/bin/ssh-keygen -b 1024 -t dsa -N "" -f /root/.ssh/id_dsa
```

SSH keys are generated in the rsa or/and dsa mode in order to be compatible with the distant protocol.

3. Copy SSH public key from the management node on this node:

```
/usr/bin/scp root@ns0:/root/.ssh/id_rsa.pub/root/.ssh/authorized_keys2
Password: *******
...
/usr/bin/ssh-keyscan -t rsa -p 22 ns0 2>&1 | tail -n 1 >>
/root/.ssh/known_hosts
```

Note: you may run the same command (ssh-keyscan) on the management node using node name instead of 'ns0' to add this node in the known hosts list of SSH, eg.:

```
ns0 # /usr/bin/ssh-keyscan -t rsa -p 22 nsX 2>&1 | tail -n 1 >>
/root/.ssh/known_hosts
```

4. Copy some essentials files from the management node to this compute or I/O node:

```
/usr/bin/scp root@ns0:/etc/hosts /etc/hosts
/usr/bin/scp root@ns0:/etc/passwd /etc/passwd
/usr/bin/scp root@ns0:/etc/shadow /etc/shadow
/usr/bin/scp root@ns0:/etc/group /etc/group
/usr/bin/scp root@ns0:/etc/gshadow /etc/gshadow
/usr/bin/scp root@ns0:/etc/shells /etc/shells
/usr/bin/scp root@ns0:/etc/securetty /etc/securetty
```

5. Edit /etc/resolv.conf and check that it is empty.

6. Verify the default run level is 3 in /etc/inittab as follows:
   `id:3:initdefault`

7. Create a directory for users homes, here named "/home_nfs":

```
mkdir -p /home_nfs
```

8. Add the corresponding entry in /etc/fstab to create a share for user accounts:
   `ns0:/home_nfs /home_nfs nfs`
   `rsize=8192,wsize=8192,intr,tcp 0 0`

9. Create a directory that should contain shared programs or data over the cluster, defined name `"/opt/envhpc":`

```
mkdir -p /opt/envhpc
```

10. Add the corresponding entry in `/etc/fstab` to create a share for data:

```
ns0:/opt/envhpc /opt/envhpc nfs
rsize=8193,wsize=8192,intr,tcp 0 0
```

11. If Platform LSF is present create the default directory:

```
mkdir -p /usr/share/lsf
```

12. If Platform LSF is present share binaries files through NFS in /etc/fstab:

```
ns0:/usr/share/lsf /usr/share/lsf nfs
rsize=8193,wsize=8192,intr,tcp 0 0
```

13. Mount newly created NFS shares:

```
mount -a
```

14. Remove unused services in xinetd by editing the following files and setting the "disable" parameter to "yes":

```
/etc/xinetd.d/telnet
/etc/xinetd.d/rsh
/etc/xinetd.d/rlogin
/etc/xinetd.d/wu-ftpd
```

and then:

```
/sbin/service xinetd restart
```

15. Reboot the node before continuing installation and after reboot check the configuration (network, NFS mount points...):

```
shutdown -r now
```

5. Tools and Applications Installation

This chapter is a guide to install tools or commercial software from **CDs** or **supplier sites**.

5.1 Intel Compilers

5.1.1 Fortran Compiler

Installation

An installation notice is supplied with the Intel compiler delivered by Bull.

Nevertheless you have to change in the notice, everywhere it appears, the version number , with the real version number of the product.

Moerover, for HPC installation the Fortran compiler installation path recommended is: /opt/intel/compilo_<fc_rel_nb>/l_fc_pc_<fc_pk_version_nb>

Where fc_rel_nb is the release number of the Fortran compiler and fc pk_version_nb is the version of the delivered package.

For example, the path for a 8.1.19 release is: /opt/intel/compilo_8.1/l_fc_c_8.1.019.

5.1.2 C/C++ Compiler

Installation

An installation notice is supplied with the Intel compiler delivered by Bull.

Nevertheless you have to change in the notice, everywhere it appears, the version number, with the real version number of the product.

Moerover, for HPC installation, the C/C++ compiler installation path recommended is: /opt/intel/compilo_<cc_rel_nb>/l_fc_pc_<cc_pk_version_nb>

where cc_rel_nb is the release number of the C/C++ compiler and cc_pk_version_nb is the version of the delivered package.

For example for a 8.1.22 release, the path is: /opt/intel/compilo_8.1/l_cc_bc_8.1.022.

5.1.3 Intel Debugger

The Intel debugger is delivered as a part of Fortran or C package.

Installation

An installation notice is supplied with Intel compilers delivered by Bull.

Nevertheless you have to change in the notice, everywhere it appears, the version number , with the real version number of the product.

For HPC installation, the recommended IDB installation path is: /opt/intel/ intel_idb_<idb_rel_nb>

where intel_idb_nb is the release number of IDB.

5.2 MKL Intel Math Kernel Library

See paragraph *"Intel Math Kernel Library"* in Bull HPC Bas3 User's Guide.

5.3 Performance Analysis and Profiling Tools

5.3.1 Intel Trace Tool

Intel Trace Tool is supplied directly by Intel to the customer. It uses the `flexlm` license scheme.

The recommended path for installation is **/opt/vampir**. Install it as follows:

```
mkdir /opt/vampir
cd /opt/vampir
tar xvfz /ITC-IA64-LIN-MPICH-PRODUCT<rel number 1>tar.gz
tar xvfz /ITA-IA64-LIN-AS21-PRODUCT<rel number 2>tar.gz
```

<rel number 1> and <rel number 2> represent the release number of the product.

– Run the installation command:

```
./install
```

– 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**"

– Define your environment file **vampir-vars.sh**:

```
export PAL_ROOT=/opt/vampir
export PAL_LICENSEFILE=$PAL_ROOT/etc/license.dat
export VT_ROOT=$PAL_ROOT
export PATH=$PATH:$VT_ROOT/bin
MANPATH=$MANPATH:$VT_ROOT/man
```

For more details about the installation procedure you can read the "trace collector and trace analyser user's guide" on the internet site:
http://www.intel.com/software/products/cluster

5.4 Debuggers

5.4.1 TOTALVIEW™

The **totalview** packages are delivered by etnus. They are also available at 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
```

We recommend to install totalview in the **/opt/totalview<rel.nb>** directory**.**

You have received information for license.
In the directory **/opt/totalview<rel.nb>/toolworks/flexlm<relf.nb>,** create the **license.src** file including this information and run the script:

```
. /opt/ /totalview<rel.nb>/toolworks/flexlm<relf.nb>/bin/Configure_License
```

This script creates several files and particularly the **license.dat** file that 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 documentation as **install_guide.pdf**.

<rel.nb> and <relf.nb> represent respectively the release number of totalview .and flexlm.

6. HPC and Cluster Management CDs Installation

6.1 Overview

The purpose of this chapter is to describe installation of the HPC CD(s) and Cluster Management CD(s).

6.2 Installing the Management Node

- If you use an usb cdrom reader, default for first cdrom detected on usb is /dev/scd0. Then you have to run:

```
mount /dev/scd0 /mnt/cdrom
```

- If you use an IDE cdrom reader,
  - Identify available cdrom(s):

```
ll /dev/cdrom*
```

```
lrwxrwxrwx 1 root root 8 May 19 09:18 /dev/cdrom -> /dev/hda
```

  hda is corresponding with IDE cdrom: link between /dev/cdrom and /dev/sda is done at install time if cdrom has been detected.

- Then you have to run:

```
mount /mnt/cdrom
```

- Install HPC V5.3 CD:
  Mount the cdrom and execute:

```
cd /mnt/cdrom
./install.sh
```

- Install Cluster management V1 CD:
  Mount the cdrom and execute:

```
cd /mnt/cdrom
./install.sh
```

- After the console message `"ns commands adds a path into /root/.bashrc, please source it !"`
  run:

```
source /root/.bashrc
```

6.3 Installing a Compute Node

- If you use an usb cdrom reader, default for first cdrom detected on usb is /dev/scd0. Then you have to run:

```
mount /dev/scd0 /mnt/cdrom
```

- If you use an IDE cdrom reader,
  - Identify available cdrom(s):

```
ll /dev/cdrom*
```

```
lrwxrwxrwx 1 root root  8 May 19 09:18 /dev/cdrom -> /dev/hda
```

  hda is corresponding with IDE cdrom: link between /dev/crom and /dev/sda is done at install time if cdrom has been detected.

- Then you have to run:

```
mount /mnt/cdrom
```

- Install HPC V5.3 CD:
  Mount the cdrom and execute:

```
cd /mnt/cdrom
./install.sh
```

- Install Cluster management V1 CD:
  Mount the cdrom and execute:

```
cd /mnt/cdrom
./install.sh
```

6.4 Configuring Ganglia

WARNING This task must be realized before deployment.

- Edit **/etc/gmetad.conf** on the management node:

```
# 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 by:
****

```
# 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 (see note below)

#
#--------------------------------------------------------------------------
# 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
```

Note: `clustername` is the prefix common to all nodes names.

- Edit **/etc/gmond.conf** on the management and reference nodes.

```
>>          name "my_cluster"
```

****
to be replaced by:
****

```
>>          name "clustername"
```

Note: `clustername` being the prefix common to all nodes names.

- Restart
  - gmond service on all the nodes
  - gmetad on management node.

6.5 Configuring Syslog-ng

WARNING This task must be realized before deployment and only if you want to use syslog-ng instead of standard syslog tool.

6.5.1 On management node

- Copy /etc/syslog-ng/syslog-ng.admin on /etc/syslog-ng/syslog-ng.conf
- Modify /etc/syslog-ng/syslog-ng.conf file to add IP address (ethenet eth0 in administration network) on which server will listen

```
# Here you HAVE TO SUBSTITUTE ip("0.0.0.0") with the GOOD Inet
# Address (use ifconfig eth0)
source s_tcp
  { tcp(ip("0.0.0.0") port(5000) keep-alive(yes)); };
```

```
**********************************
Modify for instance in:
**********************************
```

```
# Here you HAVE TO SUBSTITUTE ip("0.0.0.0") with the GOOD Inet Address
# (use ifconfig eth0)
source s_tcp
  { tcp(ip("16.0.0.1") port(5000) keep-alive(yes)); };
```

- Same for:

```
#---------------------------------------------------------------------
#  To catch and send node  I/O status to nagios
#---------------------------------------------------------------------
# Here you HAVE TO SUBSTITUTE ip("127.0.0.1") with the GOOD Inet Address
(use ifconfig eth0)
> source stor_udp                         { udp(ip("127.0.0.1") port(584))
;};
**********************************
Modify for instance in:
**********************************
#---------------------------------------------------------------------
#  To catch and send node  I/O status to nagios
#---------------------------------------------------------------------
# Here you HAVE TO SUBSTITUTE ip("127.0.0.1") with the GOOD Inet Address
(use ifconfig eth0)
> source stor_udp                         { udp(ip("16.0.0.1") port(584))
;};
```

6.5.2 On Compute Node (Golden/reference)

- Modify `/etc/syslog-ng/syslog-ng.conf` file to add server IP address on which log files are centralized.

```
   #--------------------------------------------------------------------
   #  Forward to a loghost server
   #--------------------------------------------------------------------
>                 #destination loghost       { tcp("10.0.0.1" port(514)); };

**********************************
 Modify for instance in:
**********************************
   #--------------------------------------------------------------------
   #  Forward to a loghost server
   #--------------------------------------------------------------------
>                 destination loghost       { tcp("16.0.0.1" port(514)); };
```

6.5.3 On all the I/O Nodes (Compute and/or Management)

- Modify `/etc/syslog-ng/syslog-ng.conf file` to add server IP address on which log files are centralized.

```
#--------------------------------------------------------------------
#  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
(use ifconfig eth0)
>               destination     iologhost       {       udp("127.0.0.1"
port(584)); };

**********************************
 Modify for instance in:*

**********************************
#--------------------------------------------------------------------

#  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
(use ifconfig eth0)
>               destination     iologhost       {       udp("16.0.0.1"
port(584)); };
```

- After having modified thes configuration files, restart syslog-ng service

```
> service syslog-ng restart
```

6.6 Configuring Intel Compilers

Process for compiler installation is defined by Intel. Nevertheless, you will find in Chapter 5 indications concerning this installation.

6.7 Listing the Installed Bundles

The list of **installed bundles** is related in file README-fr or README-en at the root of installation CD #1.
Description of these bundles is given in APPENDIX D.

6.8 Application Post-configuration

This section describes complementary steps to be done after previous phases.

6.8.1 Ganglia

Restart the **gmond** et **gmetad** services.

6.8.2 Syslog-ng

Compute Node (Reference Node) and Management Node

Verify that **syslog-ng** service is running.

6.8.3 Configuring Lustre File System

1. Filling /etc/lustre/storage.conf of management node

This file stores information about the storage devices available on the cluster, describing which are OSTs and which are MDTs. It must reside on the management node.

This file is composed of lines with the following syntax:

```
<ost|mdt>: name=<name> node_name=<node> dev=<device>
[ jdev=<journal device> ]
```

Comments are lines beginning with sharp.
See storage.conf(5) for a more complete explanation.

ost/mdt This device is chosen to be an OST/MDT.

name The name you want to give to the OST or MDT. For example, /dev/sdd on node ns13 can be called ns13_sdd.

node_name The hostname of the node where is the device.

dev The device path (/dev/sdd for example)

jdev The name of the device where the ext3 journal will be stored, if you want it to be outside the main device. This parameter is optional.

This file is filled with information got from the **/proc/partitions** of the I/O nodes. For example, on a cluster where ns13 is an I/O node:

```
>ssh ns13 -l root "cat /proc/partitions"
```

It gives you the following output:

```
major minor  #blocks  name
   8     0   71687372 sda
   8     1     524288 sda1
   8     2   69115050 sda2
   8     3    2048000 sda3
   8    16   71687372 sdb
   8    32   17430528 sdc
   8    48   75497472 sdd
   8    64   17430528 sde
   8    80   75497472 sdf
   8    96   17430528 sdg
   8   112   75497472 sdh
```

We know that sda and sdb are system disks of ns13 so they must NOT be used as Lustre storage devices. Device sdd to sdh are available devices. We will use 17430528 kB disks as journal devices and 75497472 kB disks as main devices. This choice gives the following lines in **/etc/lustre/storage.conf** of the management node:

```
mdt: name=ns13_sdd node_name=ns13 dev=/dev/sdd jdev=/dev/sdc
ost: name=ns13_sdf node_name=ns13 dev=/dev/sdf jdev=/dev/sde
ost: name=ns13_sdh node_name=ns13 dev=/dev/sdh jdev=/dev/sdg
```

The choice of which devices will be mdt or ost is left to the administrator.

This work has to be done for each I/O nodes and new lines append in **/etc/lustre/storage.conf** of management node.

2. /etc/lustre/lustre.cfg settings

a. Edit /etc/lustre/lustre.cfg of the management node.
b. Set LUSTRE_MODE to XML.
c. Set CLUSTERDB to no.
d. Save and quit the editor.

See lustre.cfg(5) for a more complete explanation.

3. File system configuration creation

Run the following command:

```
lustre_config create -s /etc/lustre/models/fs1.lmf
```

This will generate a Lustre configuration file **/etc/lustre/conf/fs1.xml**, which uses all the available OSTs and the first available MDT.

See lustre_config(8) for a more complete explanation about Lustre models files.

For checking you can run:

```
lustre_util info -f fs1
```

This command will print information about fs1 filesystem. It allows you to check that MDT and OSTs are actually those you want to use.

4. Installing the filesystem

Run the following command:

```
lustre_util install -f /etc/lustre/conf/fs1.xml -V
```

This operation is quite long since it formats the underlying filesystem (about 15 mn for a 1TB filesystem). Do not use -V if you want a less verbose output.

For checking you can run:

```
lustre_util status -f fs1
```

Last output lines of previous command should be:

```
Filesystem fs1 is formated and offline
No nodes mount filesystem fs1
```

5. Enabling the filesystem

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 -V if you want a less verbose output.

For checking you can run:

```
lustre_util status -f fs1
```

Last output lines of previous command should be:

```
Filesystem fs1 is formated and online
No nodes mount filesystem fs1
```

6. Mounting the filesystem on clients

Run the following command:

```
lustre_util mount -f fs1 -n <list_of_client_nodes_using_pdsh_syntax>
```

For example, if your client nodes are ns2,ns3,ns4,ns7, you can run:

```
lustre_util mount -f fs1 -n ns[2-4,7]
```

For checking you can run:

```
lustre_util status -f fs1
```

Last output lines of previous command should be:

```
Filesystem fs1 is formated and online
Getting mount information from ns[2-4],ns7
fs1 correctly mounted on ns[2-4],ns7

######################################################################
## AT THIS TIME THE FILESYSTEM IS AVAILABLE AND CAN BE USED TO STORE DATA.##
## THE REST OF THIS DOCUMENT EXPLAINS HOW TO REMOVE THE FILESYSTEM ##
######################################################################
```

7. Unmounting the file system on client

Run the following command:

```
lustre_util umount -f fs1 -n '*'
```

For checking you can run:

```
lustre_util status -f fs1
```

Last output lines of previous command should be:

```
Filesystem fs1 is formated and online
```

```
Getting mount information from ns[2-4],ns7
fs1 correctly unmounted on ns[2-4],ns7
```

8. Stopping the file system

Run the following command:

```
lustre_util stop -f fs1 -V
```

For checking you can run:

```
lustre_util status -f fs1
```

Last output lines of previous command should be:

```
Filesystem fs1 is formated and offline
Getting mount information from ns[2-4],ns7
fs1 correctly unmounted on ns[2-4],ns7
```

9. Removing file system

Run the following command:

```
lustre_util remove -f fs1
```

For checking you can run:

```
lustre_util status -f fs1
```

Last output lines of previous command should be:

```
Filesystem fs1 is not installed and offline
No nodes mount filesystem fs1
```

See lustre_util(8) for a more complete explanation.

6.8.4 Configuring NTP

The Network Time Protocol (NTP) is used to synchronize the time of a computer client to another server or reference time source. This section does not cover time setting to 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 here the reference time source.

6.8.4.1 On the Management Node

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

1. The first line should be under comment:

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

2. The second line should have the following syntax assuming that 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

      ex:
restrict 10.0.0.0 mask 255.255.0.0 nomodify notrap

Leave the line: restrict 127.0.0.1
```

3. Put the following lines in comment:

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

4. Leave the other command lines and parameters 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
```

5. Restart **ntpd** service:

```
service ntpd restart
```

6. Start **ntptrace** assuming 10.0.0.1 being the management node IP address

```
ntptrace 10.0.0.1
   valid0: stratum 11, offset 0.000000, synch distance 0.012515
```

6.8.4.2 On the Reference Node

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

7. Put the following line in comment:

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

and put instead:

```
# 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
```

8. Put in comment the default lines as follows:

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

9. Add management server as reference:

```
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
```

10. Leave the following lines:

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

11. Put under comment:

```
#keys /etc/ntp/keys
```

12. Add the following lines at the end of the file:

```
tinker panic 0
tinker stepout 0
```

6.8.4.3 Restart NTP

13. Restart NTP on each node (management node and reference node):

```
/etc/init.d/ntpd restart
   Shutting down ntpd:                                    [  OK  ]
   Starting ntpd:                                         [  OK  ]
```

14. On the management node, start **ntptrace** and check if management node responds:

```
ntptrace 10.0.0.1
```

```
valid0: stratum 11, offset 0.000000, synch distance 0.012695
.........
```

15. From the management node, check if clocks are identical:

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

```
valid0: Tue Aug 30 16:03:12 CEST 2005
valid1: Tue Aug 30 16:03:12 CEST 2005
```

7. Quadrics Interconnect Installation

7.1 Setting-up Quadrics Interconnect

7.1.1 Assumptions

Hardware:	NovaScale
Linux distribution:	Bull Advanced Server (BAS3)
Quadrics interconnect:	QsNetII (Elan4)
Quadrics Web site:	www.quadrics.com
Quadrics support:	Robin Crook <robin@quadrics.com>

Assumptions made for the installation:

Cluster name:	ns (where ns is the base name).
Nodes to configure:	management node (ns0), reference node (nsX) which could be a compute node or an I/O node.
IP addresses range:	172.16.12.<n+1>, with n ranging from 0 to number of nodes.

7.1.2 Hardware Configuring

The following procedure explains how to add a new Quadrics interconnect equipment in your cluster configuration.

Note: please take all necessary precautions for this hardware installation step. For furthers details, refer to Quadrics documentation (QM500 Installation Manual, QM-S64 or QM-S8 Installation Manual, and QsNetII 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 each node within the cluster and unplug the power leads.
3. Plug one Quadrics card (QM-500) per node into the appropriate PCI slot (see details below and in hardware description).
4. Connect each card to the switch using the provided cables, and label them accordingly.

   For example:

   ns0 on port 0
   ns1 on port 1

   ns<n> on port <n>

After operating system deployment you have to follow these steps to check if the hardware configuration of QsNetII network is operational on each nodes:

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 be become green on both sides.
6. For each node, check that the card is detected on the PCI bus by issuing the lspci command.

The output should be similar to the following:

```
11:01.0 Network controller: Quadrics Ltd QsNetII Elan4 Network
Adapter (rev 01).
```

For switch management purpose you have to connect on each QM-503 board (two in case of QM-S64) the administration network to get information on the switch status:

1. plug a keyboard and a screen on the module to get the login prompt
2. enter default login/password to gain access to the QNX embedded system:
   login: quadrics
   Password: system

For security reason change this default setting as soon as the cluster is in an exploitation state as soon as users can log on.

3. Configure switch network settings:

   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

and follow the instructions in Quadrics QsNetII Installation and Diagnostics Manual (p.104) to configure the network interface.

4. Plug in the ethernet cable to the QM-503 module and then to one administration network switch (ethernet).
5. Check the switch configuration using ping from management node and if it is in working state, remove the keyboard and screen from the QM-503.

Notes:

a. 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 adpater, etc) is present on the same PCI bus (shared controller), a penalty performance will be incurred. It is therefore advised to install each Quadrics adapter on its own PCI bus for performance reasons.

b. The QM-500 card remaps 256MB of PCI memory, and in some configurations, this may prevent the system from booting. In this case, the card should be relocated in another slot, with a bus frequency greater or equal to 133MHz. If this problem occurs, please check the BIOS revision of your NovaScale and ask Bull support if a new revision is available and check in BIOS if "PCI->PCI Gap above 4BG" option is enabled.

c. Quadrics switches doe not support DHCP (dynamic IPs) for now. So you have to set up the IP statically or to use bootp command.

7.2 Installing Quadrics Software Packages

Note: the Quadrics packages will be first installed using Quadrics CDROM and then deployed into the reference image as described in chapter 2. Licenses management and verifications are to be done afterwards.

7.2.1 Install on the Management Node

1. Insert Quadrics CD-ROM and mount it in /mnt/cdrom

```
mount /mnt/cdrom
```

2. Run the installation script to install Quadrics RPMs on reference node:

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

Answer 555 to the questions regarding group and user Ids.

7.2.2 Install on the Reference Node

1. Insert Quadrics CD-ROM and mount it in /mnt/cdrom

```
mount /mnt/cdrom
```

2. Run the installation script to install Quadrics RPMs on reference node:

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

Answer 555 to the questions regarding group and user Ids.
Enter the root@rmshost password, which is requested several times.

3. Reboot the management node before continuing:

```
shutdown -r now
```

7.2.3 Network Installation

You can use this method if you cluster is already operational and that you want to add Quadrics interconnect software. For this you will use a NFS partition previously defined on every node.

The install procedure first copy data in this NFS partition and then install software on the entire cluster.

1. Insert Quadrics CD-ROM and mount it in /mnt/cdrom

```
mount /mnt/cdrom
```

2. Run the installation script to install Quadrics RPMs on all required nodes:

```
cd /mnt/cdrom
./install_quadrics.pl -nodes ns[X-Y,Z]
```

3. Reboot the reference node before continuing:

```
shutdown -r now
```

7.2.4 Licenses Management

1. To obtain the license from Quadrics (if not already configured by Bull) you have to get your FLEXlm host ID doing:

```
/usr/lib/rms/flexlm/bin/lmhostid
```

You will obtain an answer as below

```
lmhostid - Copyright (c) 1989-2003 by Macrovision
Corporation. Allrights reserved.
The FLEXlm host ID of this machine is "0007e993fc4c"
```

and send this output to Quadrics (support@quadrics.com).

2. When you receive the license from Quadrics

Copy "rms.lic" in /usr/lib/rms/flexlm/

```
cp rms.lic /usr/lib/rms/flexlm/
```

Then you can choose running rms with a local license or use a global license manager.

3. For using local license, restart qslmgrd:

```
service qslmgrd restart
```

OR:

4. For running with a global license manager:

```
service rmslmgrd restart
```

7.2.5 Verifying each Installed Node

1. Check 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 status of RMS service:

```
# service rms status
rms module:  loaded
running: rmsmhd rmsd pmanager tlogmgr eventmgr mmanager  swmgr
stopped:  swmserver                                        [  OK  ]
```

3. Check status of mSQL database only on management node:

```
service msqld status
msql3d  (pid 10843) is running...
```

Notes:

If you install Quadrics nodes using deployment software you will have to manually create each node entry in RMS database. For this you need to run the following command after each node deployment on the management node with the corresponding hostname:

```
rcontrol create node nsX
```

8. Building Reference Image and Deployment

The purpose of this chapter is to describe how to prepare:

- an image for each node on the image server
- the list of the nodes on which this image will be deployed.

This primary mechanism can be used for subsidiary purposes such as saving different releases of images for different types of nodes.

Note:

Generally each type of node requires a specific image.

8.1 Prerequisite

At this point of installation we assume that we have completed, according to the process described in preceding chapters, the installation of:

- management node used as the image server,
- one compute and one I/O node minimum These nodes will be cloned by the operations of deployment on the other nodes of the HPC. This cloning uses an image (reference image) of these nodes that will be stocked on the image server,
- quadrics interconnect.

Before deployment it is evidently important to verify on the nodes which will be cloned that all installed services are running correctly (see chapter 9).

8.2 Main Steps for Deployment

The steps for the deployment are:

1. Prepare the Image Server.
2. Prepare Compute or I/O nodes to be cloned and create an image.
3. Get the reference images from the nodes to be cloned on the Image Server using **tksis**.
4. Add clients: this operation consists in building the relations between nodes and images stored on the image server.
5. Deployment: In order to install the nodes, boot the nodes of the cluster through the network.

8.3 System Installer Suite

System Installer Suite (SIS) makes it easy to deploy software distribution, content or data distribution changes, operating system update and software update, through your network of Linux machines.

System Installer Suite is used to ensure safe production deployments. By saving your current production image before updating your new production image you have a highly reliable contingency mechanism. If the new production environment is found to be flawed, simply roll-back to the last production image.

When the valid image is saved, it can be propagated on all cluster nodes through the network.

At this point we assume that System Installer Suite installation is effective for the server and for the client:

8.3.1 Prepare the Image Server

The server configuration file (**/etc/dhcpd.conf**) is pre-initialized with parameters that must be changed to take into account the new network. Edit the **/etc/dhcpd.conf** file and replace values `172.16.110.`*`XX`* by the new network values.

By default a machine range is set from `172.16.110.80` to `172.16.110.90` for client boot request. This value can be increased as needed (for example `172.16.110.1` to `172.16.110.100`).

```
/etc/dhcpd.conf
#
```

```
# "SystemImager"
#
#  Copyright (C) 1999-2001 Brian Elliott Finley
#                          <brian.finley@baldguysoftware.com>
#  Copyright (C) 2002 Bald Guy Software
#                     <brian.finley@baldguysoftware.com>
#
#  This file was created with "mkdhcpserver", which is part of
SystemImager.
#  See http://systemimager.org/ for more information.
#
# This is an ISC DHCP v3 configuration file.

# general options
#ddns-update-style none;
ddns-update-style ad-hoc;

# The "host-checking" option is only used by Curtis Zinzilieta's DHCP
patch.
# The patch adds functionality to the server for removing specified host
# addresses from ranges of available ip addresses at startup.  This
addresses
# the issue where a fixed-address reserved for a specific machine gets
# assigned to a different machine.
#
# If you are running a version of dhcpd with this patch, then you will
want
# the "host-checking" option below to be uncommented.  If you are not
running
# the patched dhcpd, you should leave it commented out.  You can find
the patch
# and the matched DHCP v3.0 source code here:
http://systemimager.org/download/
#host-checking true;


# make network booting the SystemImager autoinstall client possible
allow booting;
allow bootp;

# set lease time to 3 days
default-lease-time 259200;
max-lease-time 259200;

. . . . . .
filename "/elilo.efi";
subnet 172.16.110.0 netmask 255.255.255.0 {
  range 172.16.110.80 172.16.110.90;
  option domain-name "frec.bull.fr";
  option routers 172.16.110.250;
 }
```

Once the **dhcpd.conf** file is modified, dhcpd service must be re-started:

```
# service dhcpd restart
```

To verify that dhcpd daemon is started enter:

```
# service dhcpd status
```

The output should be similar to the following one:

```
dhcpd (pid 1499) is running...
```

8.3.2 Prepare Reference Images

Just remind that the reference images are created on designated nodes (minimum one compute and one I/O node) and are destined to be cloned on the similar nodes of the HPC.

During this operation you must be logged on the node of which you are preparing the image.

Then in order to prepare these nodes:

6. Log on the node
7. According to the **dhcpd.conf** file on the image server, complete the /etc/hosts file to add a range of machines. A base host name has to be chosen (for example `merced`) with a domain name (for example `frec.bull.fr`). The result will be a list of hosts that have to be added to the **/etc/hosts** file:

```
# vi /etc/hosts
```

```
172.16.110.80    merced80.frec.bull.fr  merced80
172.16.110.81    merced81.frec.bull.fr  merced81
172.16.110.82    merced82.frec.bull.fr  merced82
172.16.110.83    merced83.frec.bull.fr  merced83
172.16.110.84    merced84.frec.bull.fr  merced84
172.16.110.85    merced85.frec.bull.fr  merced85
172.16.110.86    merced86.frec.bull.fr  merced86
172.16.110.87    merced87.frec.bull.fr  merced87
172.16.110.88    merced88.frec.bull.fr  merced88
172.16.110.89    merced89.frec.bull.fr  merced89
172.16.110.90    merced90.frec.bull.fr  merced90
```

8. Export the LANG environment:

```
# export LANG=C
```

9. Complete the client preparation using the **/usr/local/sbin/prepareclient** script. You can restart the node preparation using this script. Run the command:

```
# /usr/sbin/prepareclient -yes
```

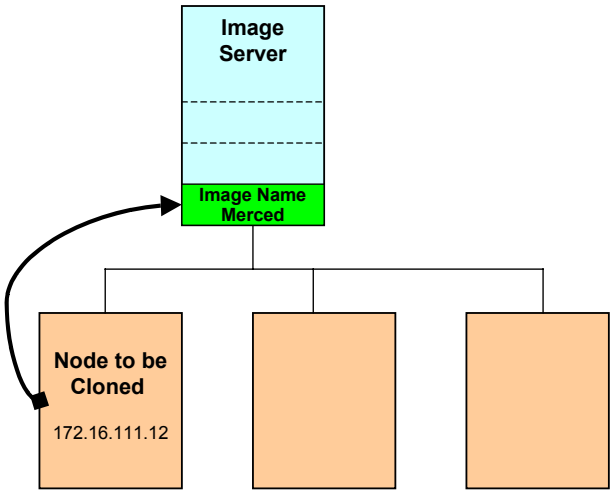
The result of this operation is the creation of this node's image on the node itself in order to be retrieved.

8.3.3 Get Image on the Image Server

This step consists in the creation of a copy of the node image (reference image to be cloned at deployment time) on the management node (Image server).

This operation is done while you are logged on the image server (management node).

In the following example the IP address for the node image to be cloned is `172.16.110.12` and the image name created on the Image server is `MERCED`.



The creation of an image on the image server is done using tksis.

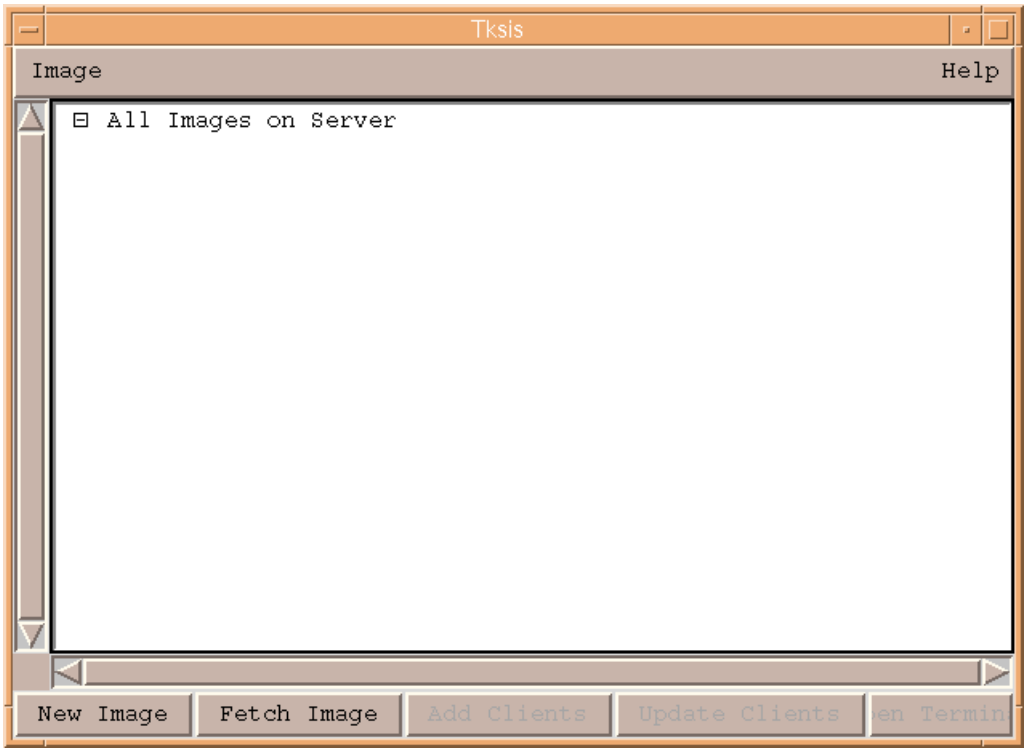
```
#export LANG=C
```

tksis interface consists in a user interface for interactive build image program.

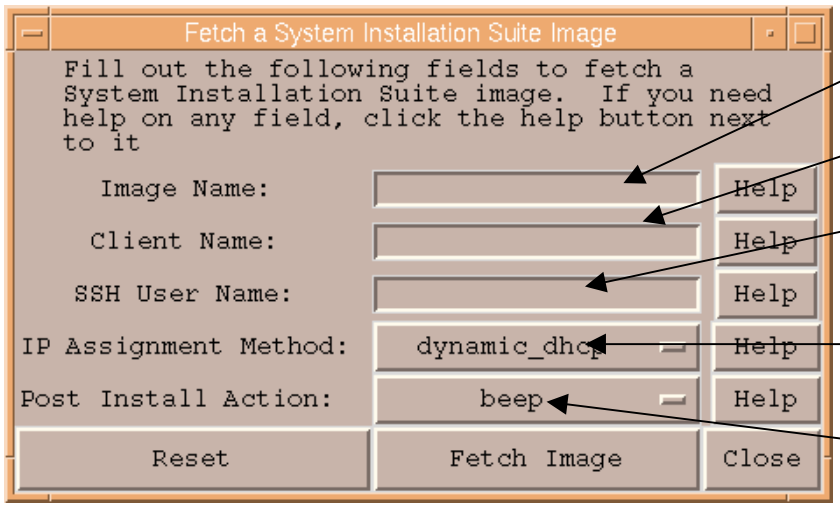
Start tksis:

```
# /usr/bin/tksis
```

The following windows appears:



The first step creates an image from your designated Node. To do this, click the **Fetch Image** button. The following window displays:



For example MERCED

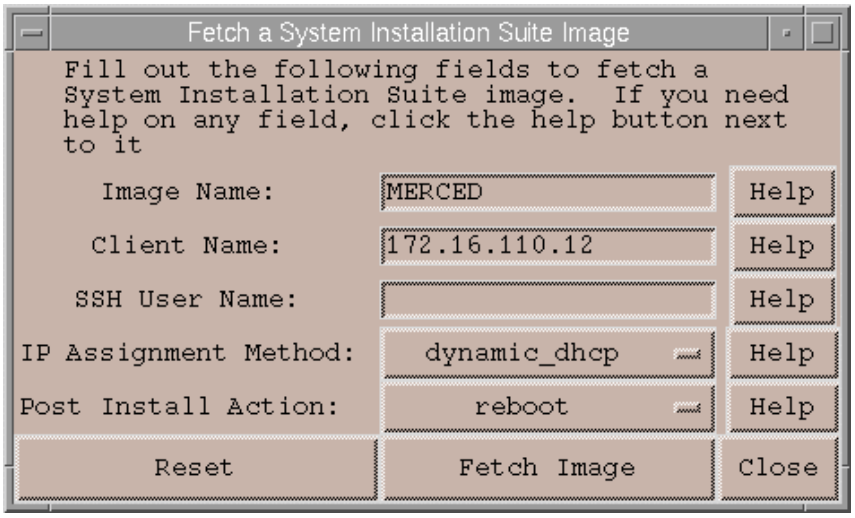
IP address of the node with the reference image

If needed

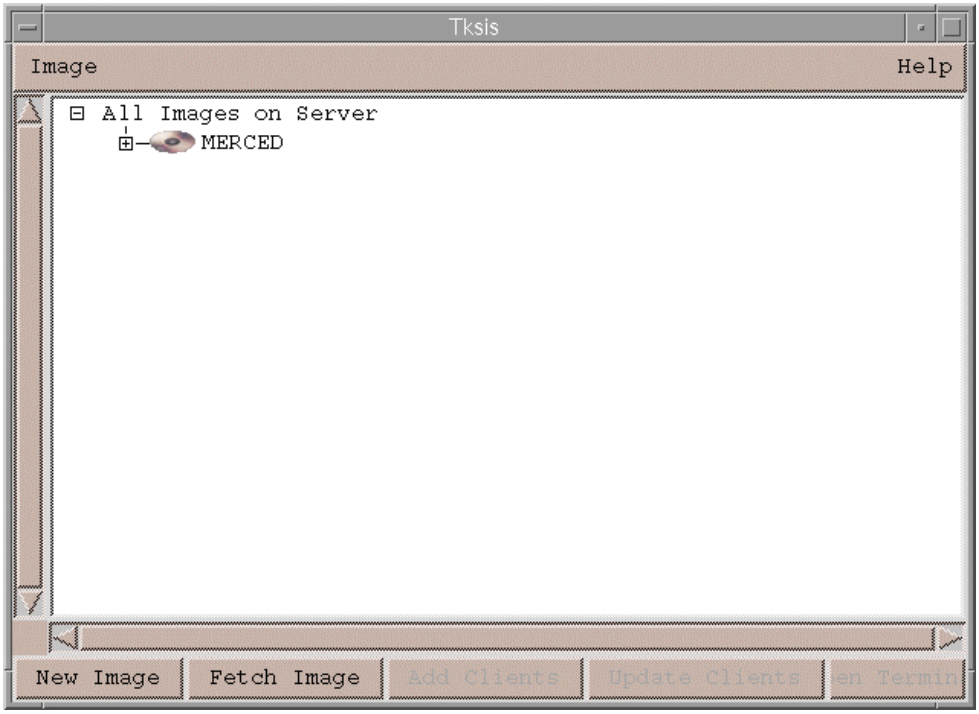
This option is correct

Choose reboot option

Once filled, the window looks as follows:



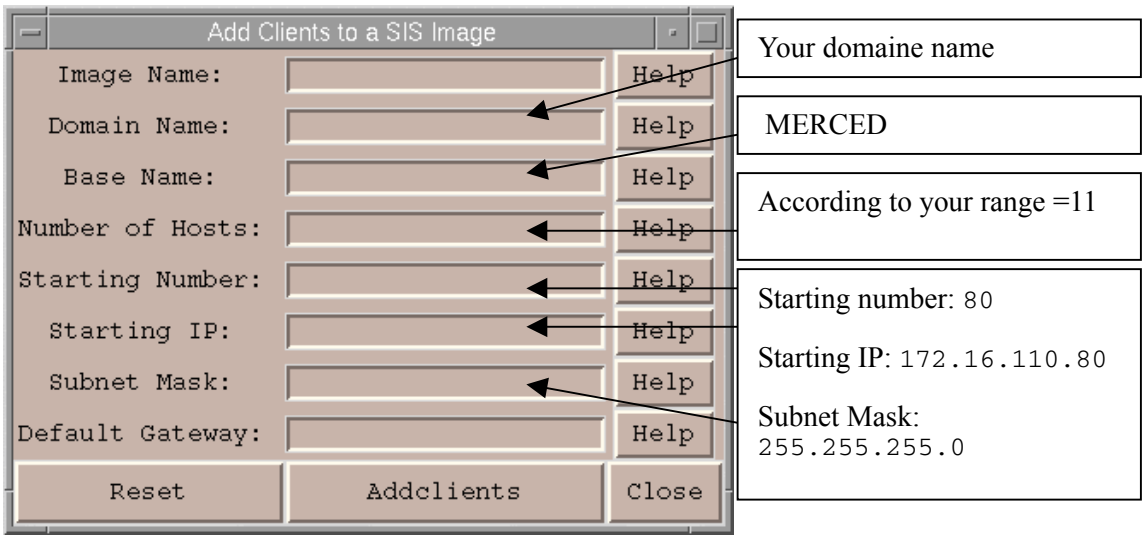
When the image creation is completed tksis displays the created `MERCED` image on your Image server as below.



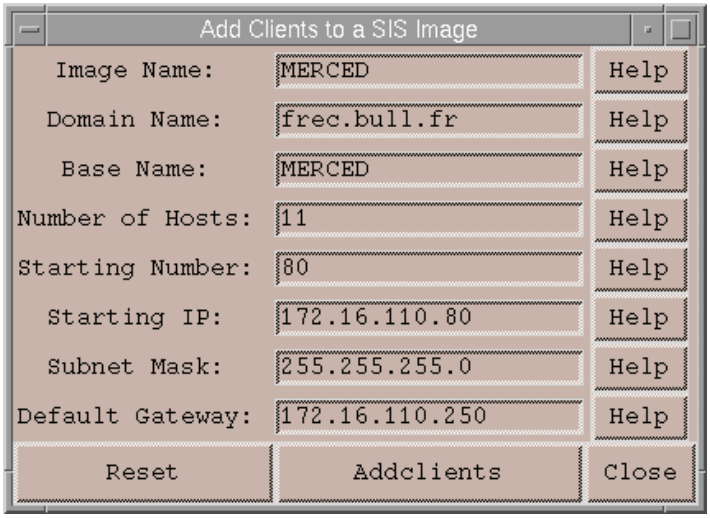
8.3.4 Add Clients

This step consists in linking nodes with the image they have to run with.

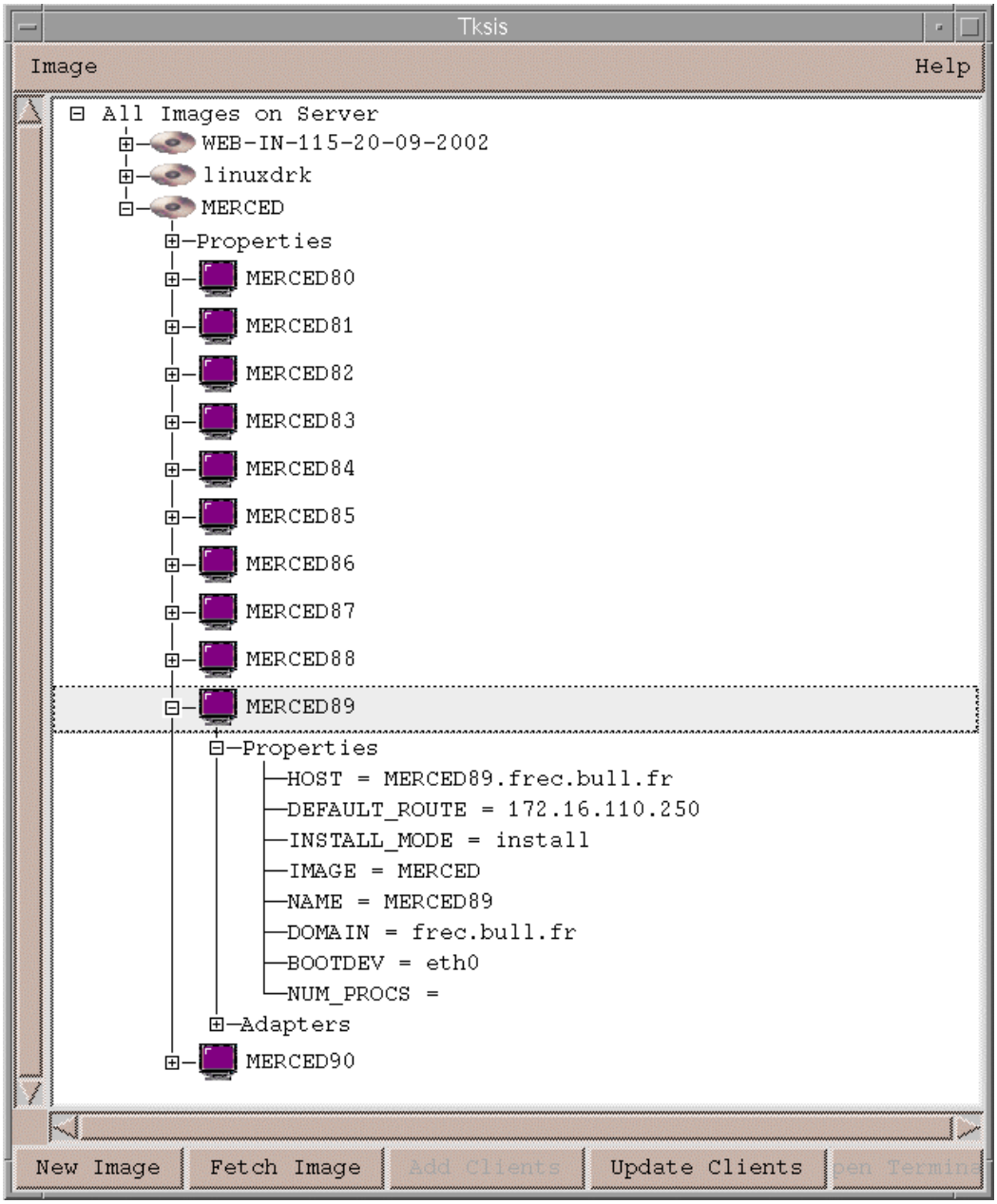
Once `MERCED` image is created you have to link this image to the machine range as indicated in the dhcpd.conf file (*range 172.16.110.80 172.16.110.90;* Number of Hosts=11).



Once filled, the window looks as follows:



Now clients `MERCED80` to `MERCED90` are associated to `MERCED` image.



8.3.5 Deployment

You are now ready to boot and install a new client node across the network.

Every time you boot a node through the network, it loads and installs the image associated with this node on the image server.

Plug the client to the network and boot the system from the network using EFI.

If your node is yet installed with an old version, during the boot phase when boot option is displayed, type the option “boot from the network”.

The client makes a DHCP request and your DHCP Image server responds with the usual information. The installation of the machine starts. At the end, the client machine reboots, network is yet configured. The bootloader installed is EFI.

If the EFI boot fails during DHCP request, just verify that **dhcpd** is running on the image server (service dhcpd status).

If a Time Out TFTP request occurs during network boot you have to restart tftpd on the image server using the command:

```
Service restart xinetd
```

8.3.6 Using SIS After Deployment

Once the nodes have been booted, you can specialize them by adding file system or complementary software in order to obtain an I/O node for example.

When the node has been specialized, to save the modifications two steps are necessary.

1. create an image of this node on the image server as described in § 8.3.3.
2. create a new association between this image and this node (see § 8.3.7).

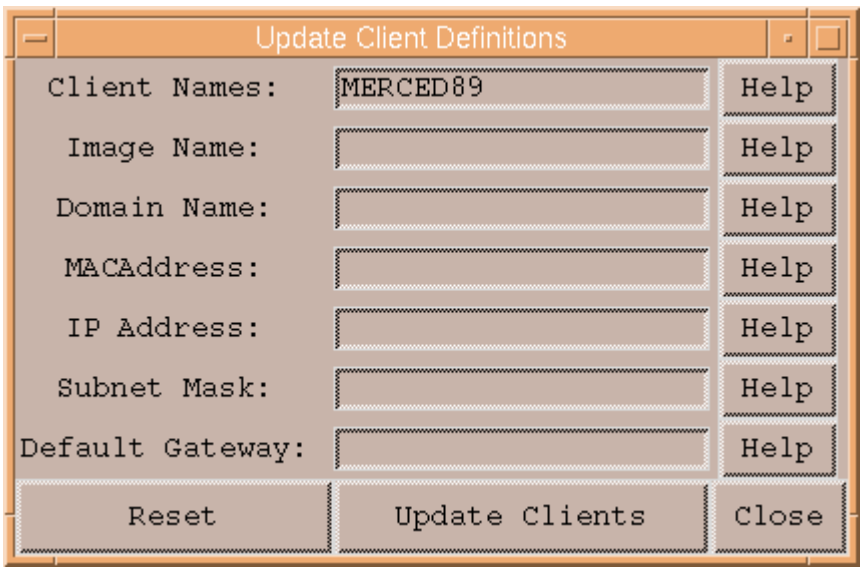
8.3.7 Update client

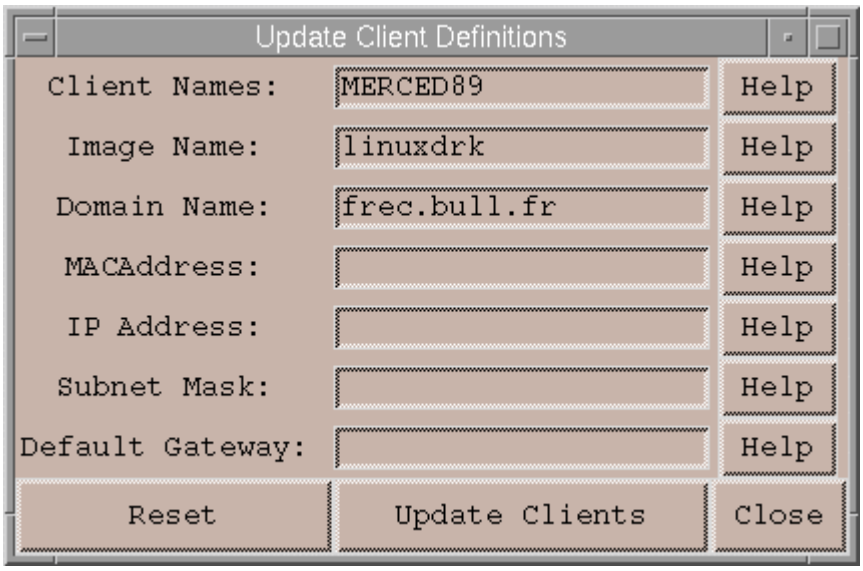
This step consists of modifying the association between image and client.

If more than one image is available on the image server as in the previous windows like WEB-IN-115-20-09-2002 or linuxdrk, you can associate one or more clients to another image. In this example we want to associate `MERCED89` client with `Linuxdrk` image.

In this case, if `MERCED89` reboots using the network, the machine will be automatically installed with this new image.

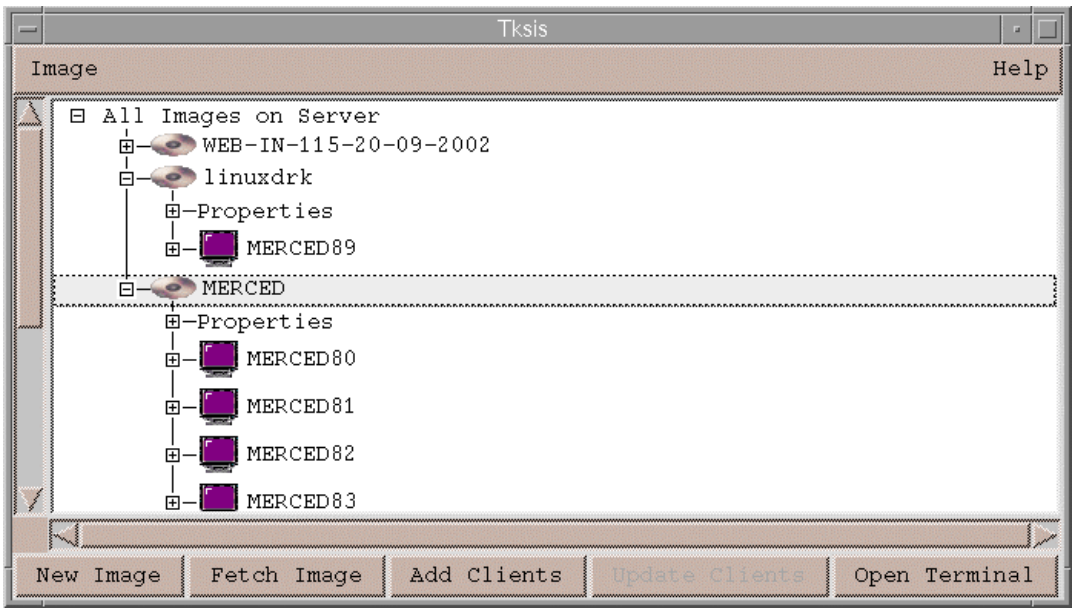
Select your client (MERCED89) and click **Update Clients** button:





The following figure shows that MERCED89 is now associated with the linuxdrk image instead of MERCED.

In this case, if Merced89 reboots using network reboot (PXE), the machine will be automatically installed with this new image.



9. Making the Cluster Operational

This chapter lists the complementary tasks in order to have the cluster operational.

9.1 Compute or I/O Nodes

General verifications of reference image have to be done before deployment.

After deployment it's now time for specializing desired nodes (for instance in I/O nodes).

You have to create an image for each type of node, then get this image on the image server, in order to associate this image with all concerned nodes. The images will be operational on each concerned node after next network reboot.

In I/O nodes it's also necessary to modify Lustre configuration files, mount Lustre file system and activate.

9.2 Management Node

The objective is to check that the cluster is operational. Checking. tasks include:

- complementary necessary services are activated (Torque, Lustre, Conman, Nagios)
- nfs is exported on all the nodes using exportfs command.

To perform a global verification we recommend executing a shell that:

1. Compiles a scientific application using mpi
2. Runs the application on all the nodes.

If the test is successful it means that compilers, Quadrics rms and mpi libraries are running correctly.

After activation of Lustre you can also verify what is mounted on each node.

Bull can send you some script examples that make these verifications on request.

9.3 Backing up the System: Mkcdrec

mkCDrec (make CD-ROM recovery) is an Open Source tool designed to make a bootable system image (including Linux system save), then to recover after a disaster happened, such as a disk crash or system intrusion.

Refer to *HPC BAS3 Administrator's Guide* (86 A2 31EM) for configuring and using.tasks.

9.4 Checking the Nodes: Nodechecking

nodechecking is a tool used for verifying cluster nodes. It is installed on every node on which you want execute the verification.

nodechecking runs tests that require an implementation of MPI, the Intel C compiler, the Intel Fortran compiler and the Intel MKL library. These packages must be installed, so that all tests can run.

Refer to *HPC BAS3 Administrator's Guide* (86 A2 31EM) for configuring and using.tasks.

A. Error Messages

The installation procedures described in the present guide run without error. This appendix describes some errors that you may encounter if a procedure is not correctly applied.

A.1 The Installation Procedure does not start up automatically

During machine startup, after some initial phases (Bios, SCSI detection, etc) the screen doesn't display any more information, then it displays the EFI banner, enabling the user to run the appropriate **.nsh** startup file from the *Bull Kernel Extension for Linux*® CD-ROM. If this banner does not appear, you must exit using the **Ctrl + Alt + Del** key sequence to restart the machine. Then run the standard installation procedure.

If the procedure still doesn't start, launch it manually from the EFI shell (indicated by the **Shell >** prompt).

Note: EFI recognizes only the QWERTY keyboard setup. Be careful when you enter letters and numbers. See keyboard comparison in Appendix B.

1. Insert the Bull Kernel Extension for Linux® CD-ROM.
2. Type the command: **map -r -b** and **Enter**
3. Select the CD-ROM drive from the displayed list (locate the string `CDROM`, with a type `FF`).
4. Take note of the name of the file system of the CD-ROM drive (in the form `fsN:` where `N` is an integer).
5. Type the command **fsN:** (replacing **N** by the correct value: `0`, `1` or other) and **Enter**.
6. Type **dir** and **Enter** to view the Bull Kernel Extension for Linux® CD-ROM files (**elilo.conf**, **elilo.efi**, **initrd-BAS3V20**, **vmlinuz-BAS3V20,** etc).
7. Type the command **elilo** then press **Enter** and choose your install.

The automatic installation procedure of Linux® Bull Advanced Server 3 starts.

If this procedure still does not start, contact your Support representative.

A.2 Message "Error in locating EFI System Partition Protocol"

This message is displayed rapidly in the EFI phase and generally only the first time. Ignore it.

A.3 The Machine freezes during Installation

Several symptoms may appear:

A.3.1 The screen freezes

In the case of a cluster configuration, only the management node connects 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 the management node. However, if you do want to perform your installation without redirecting the report, you must consider this node as the management 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.

A.3.2 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 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 Shell EFI, the procedure will fail to start up.
- Follow the standard installation procedure.

A.3.3 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 the NovaScale 4040 server, remove the disk and install it back.
- On the NovaScale 5080/5160 server, perform a verification using the disk manager.
- Return to the installation procedure.

A.3.4 Message "cu: /dev/ttyD000: Line in use"

Possible causes of the problem:

- A "cu" process is using the line:
  - Run the command: **`/bin/ps -efa |grep cu`** to check if it is the cause of the problem.
  - 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 command "digiConf"

```
rpm -ivv drv-epca-1.50-1.b.1.Bull
/usr/sbin/digiConf
```

  - Answer the series of questions according to your(s) installed Digiboard communication controller(s)
- 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.4 Localization – Messages in English

If the installation is not in English, some messages or menu labels are not translated or partially translated into the local language.

With cu line some French characters are badly interpreted.

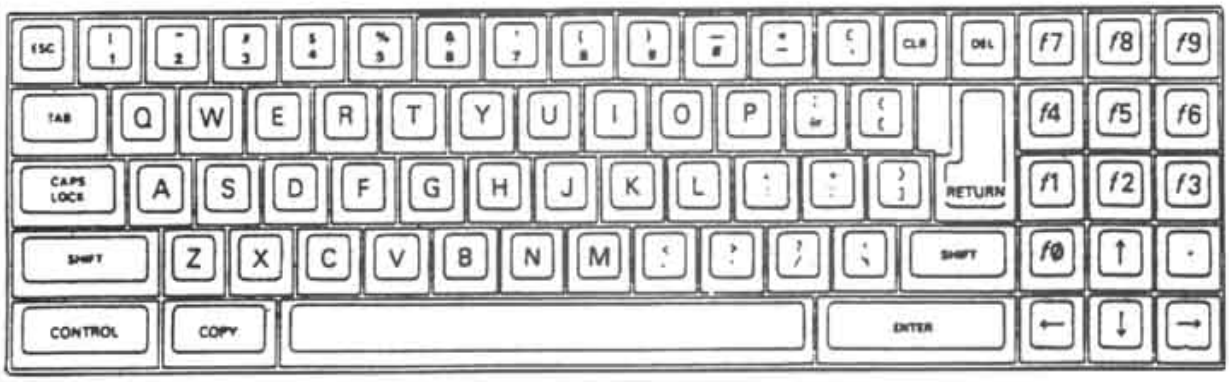
A.5 Power out during installation

If the machine is stopped intentionally or not (power failure) during any phase of the installation process, simply switch on the machine and restart the entire installation process.

B. QWERTY Keyboard Comparison

EFI recognizes only the QWERTY keyboards. This appendix enables you to enter EFI key sequences with an other keyboard by comparison between the key boards.

In addition you can find layout for most of countries on the internet site http://www.translation.net/keyboard.html



C. Recommendation for PCI Slots Selection

This appendix provides detailed information to optimise the choice of PCI slots for high bandwidth PCI adapters. The configuration rules proposed ensure the best performance levels, without IO conflicts, for most type of applications.

C.1 How to optimize IO Performance

The IO performance of a system may be limited by the software, but also by the hardware. The IO architecture of servers usually lead to concentrate data flows from PCI slots to a limited number of internal components, leading to bandwidth bottlenecks.

Thus, it's mandatory to carefully check the installation of PCI adapters within PCI slots to limit as much as possible this kind of limitation. A good practice is to avoid to connect bandwidth hungry adapters to the same PCI bus.

Defining a good adapter installation requires to:

- Know adapter characteristics, maximum theoretical performance and expected performance in the operational context
- Know the detail of the IO architecture of the server.

The next paragraphs covers these aspects and give recommendation to install adapters within servers. The process to follow is quite easy:

- Build a list of adapters to install, sorted from the highest bandwidth requirement to the lowest
- Place these adapters in the server with respect to the priority list defied hereafter for each model of server.

C.2 Building the List of Adapters

The first step is to collect the list of all the adapters which must installed on the system.

Then, if the IO flow for the server is known (bandwidth expectation from the Quadrics interconnect, bandwidth to the disks, ...), it's possible to estimate to bandwidth per adapter and then sort the adapters according to the peculiarity of each operational environment.

Else, when there is no information about real / expected IO flows, the adapter can be sorted according to their theoretical limits.

The following tables provides some numerical values for adapters support on BAS 3:

Adapter	bandwidth/s
Quadrics Elan 4	900 MB/s
SCSI U320 dual channel	640 MB/s (1)
Fibre channel dual ports	400 MB/s (1) (2)
Gigabit ethernet dual port	350 MB/s (1) (2)
SCSI U320 single channel	320 MB/s
Fibre channel single ports	200 MB/s (2)
Gigabit ethernet single port	125 MB/s (2)
Ethernet 100 Mbps	12,5 MB/s

(1) if both channels are used. Else, the adapter must be categorised as a single channel / port adapter

(2) full duplex capability is not take into account. Else double the suggested value

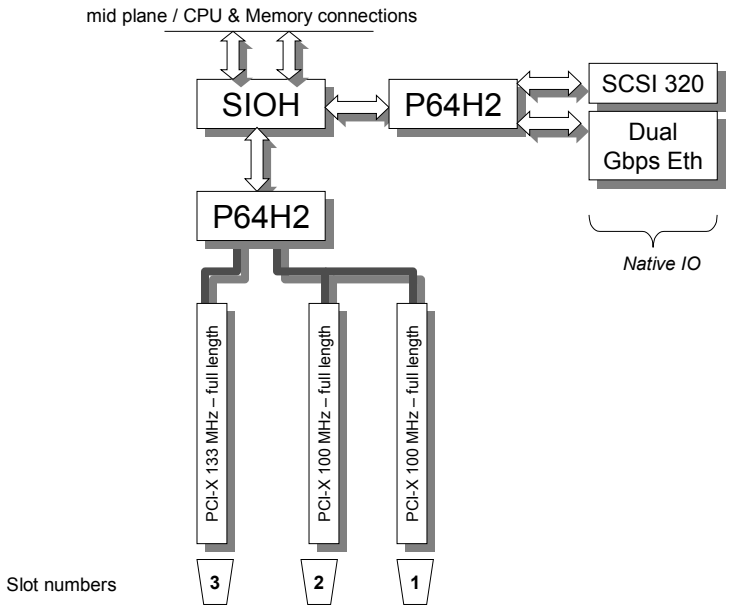
It may be possible to minor these values using the characteristics of the equipment attached to the adapter. For example, and U230 SCSI HBA connected to an U160 SCSI disk subsystem will not deliver more than 160 MB/s.

C.3 Recommendation for NovaScale Servers

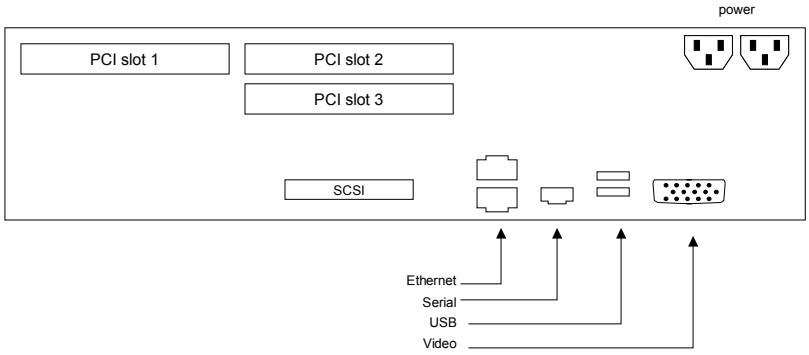
The following paragraphs describe the architecture of the IO subsystem of each family of NovaScale servers. It recommends an ordering to allocate PCI slots to adapters.

C.3.1 NovaScale 4020

The next diagrams explain the entire IO subsystem for this range of server.



NovaScale 4020 – IO subsystem



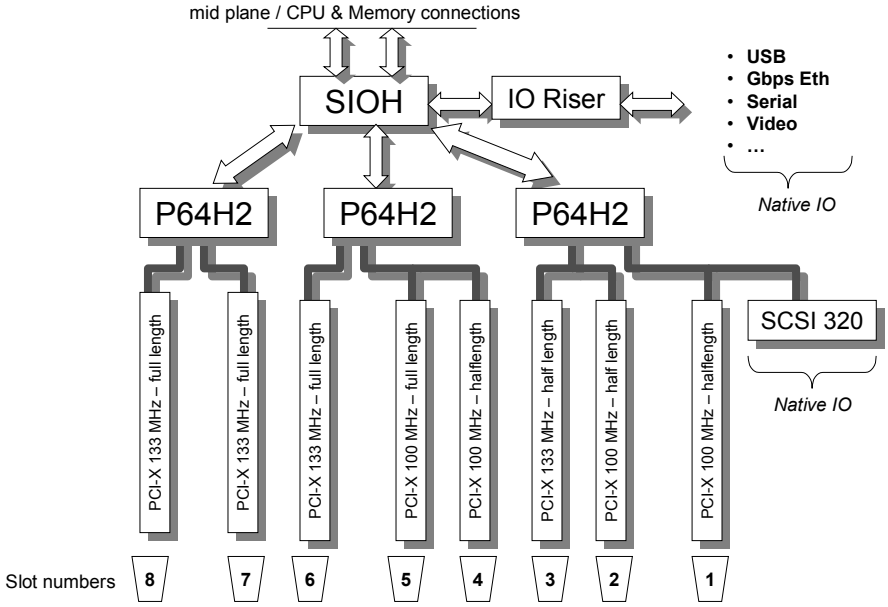
NovaScale 4020 – Pci slot identification

The following table provides the priority list to populate the PCI slots:

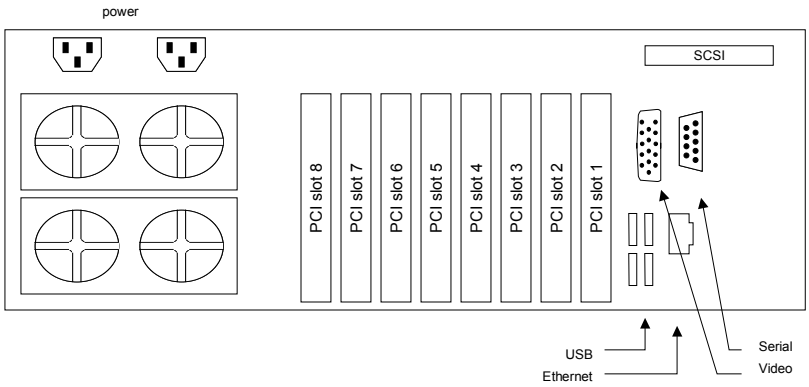
Priority	Slot number
1	3
2	2
3	1

C.3.2 NovaScale 4040

The next diagrams explain the entire IO subsystem for this range of server.



NovaScale 4040 – IO subsystem



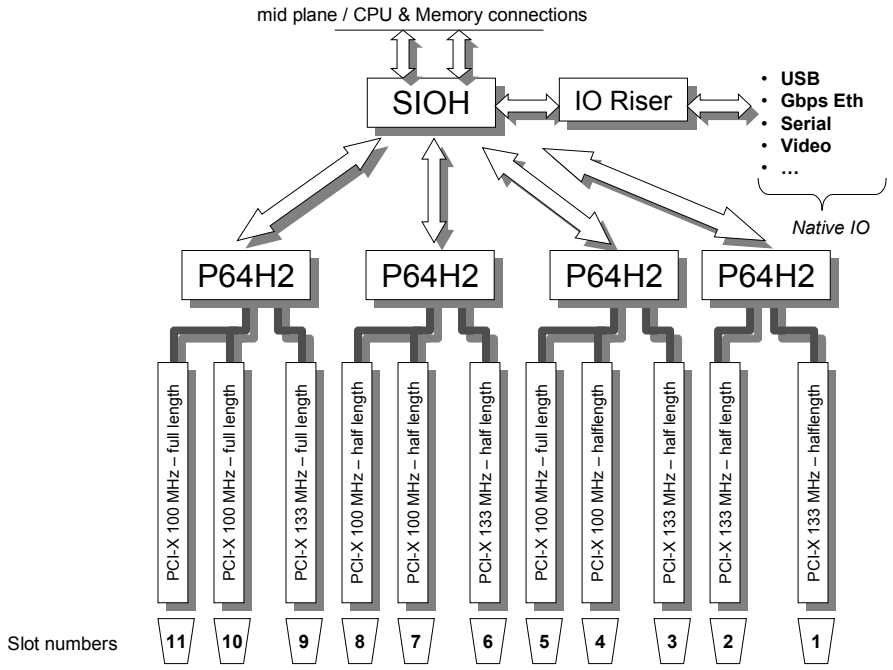
NovaScale 4040 – Pci slot identification

The following table provides the priority list to populate the PCI slots:

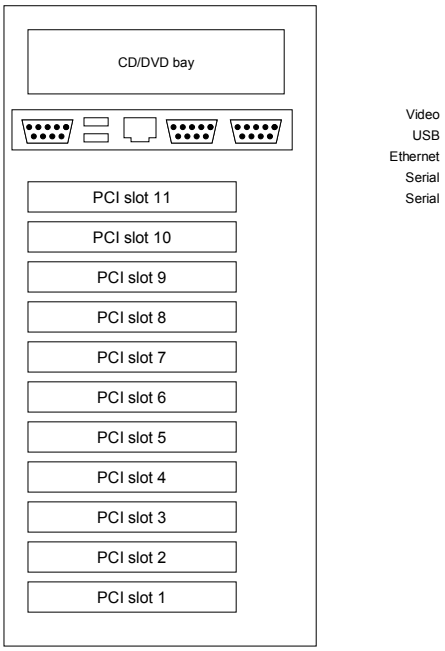
Priority	Slot number
1	8
2	6
3	7
4	5
5	3
6	2
7	4
8	1

C.3.3 NovaScale 5xx0/6xx0

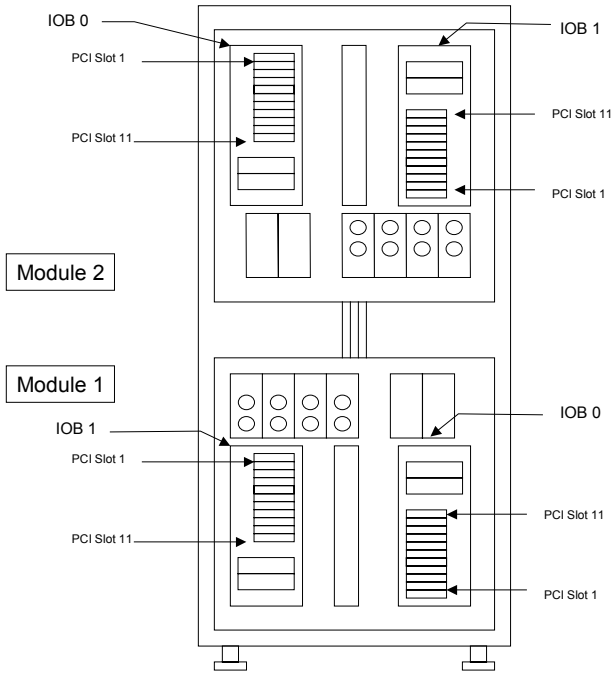
The next diagrams explain the IOB IO subsystem for this range of server. The number of IOB modules per server varies from 1 to 4 depending on the options selected.



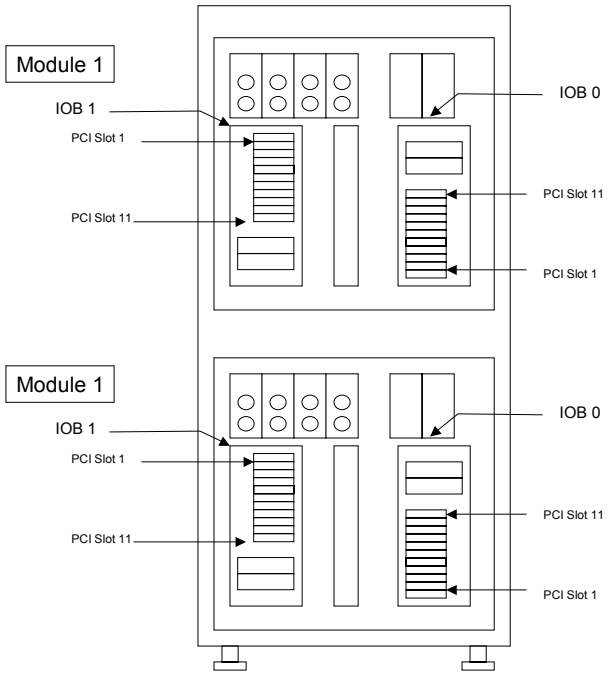
NovaScale 5xx0/6xx0 – IO subsystems per IOB



NovaScale 5xx0/6xx0 – Pci slot identification per IOB



NovaScale 5xx0/6xx0 – IOB identification for dual module / 32w capable



NovaScale 5xx0/6xx0 – IOB identification for 2 single modules in a rack

The next table provides the priority list to populate the PCI slot within an IOB. It's recommended to populate a priority level on 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, 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 large amount of data from the Quadrics interconnect and pushing them to the disk through fibre channel connections. The number of PCI adapters may be different, but the in and out flows and almost identical. The additional rule to apply is to balanced the HBAs with unidirectionalaqaq data flows accross different IOBs.

It the server does not have too many high performance adapters, it's possible to reserve the slot 1 of the module 1 / IOB 0 to the SCSI adapter used for the boot disks. It's a common practice for this family of servers, regardless of the operating system used (Bull BAS, Microsoft Windows, RedHat AS, SuSE SLES, ...).

The following table provides the priority list to populate the PCI slots:

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

D. Description of Bundles Loaded by Install Process

Bundle Name	Description
core	Smallest possible installation
base	
hpc-tools-compute	Install these tools to enable compute node hpc profile.
hpc-tools-mngt	Install these tools to enable management node hpc profile.
printing	Install these tools to enable the system to print or act as a print server.
base-x	Install this group of packages to use the base graphical (X) user interface.
dialup	
gnome-desktop	GNOME is a powerful, graphical user interface which includes a panel, desktop, system icons, and a graphical file manager.
kde-desktop	KDE is a powerful, graphical user interface which includes a panel, desktop, system icons, and a graphical file manager.
graphical-internet	This group includes graphical email, Web, and chat clients.
text-internet	This group includes text-based email, Web, and chat clients. These applications do not require the X Window System.

Bundle Name	Description
sound-and-video	From CD recording to playing audio CDs and multimedia files, this package group allows you to work with sound and video on the system.
graphics	This group includes packages to help you manipulate and scan images.
office	The applications include office suites, PDF viewers, and more.
mail-server	These packages allow you to configure an IMAP or Postfix mail server.
network-server	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.
smb-server	This package group allows you to share files between Linux and MS Windows(tm) systems.
server-cfg	This group contains all of Red Hat's custom server configuration tools.
ftp-server	These tools allow you to run an FTP server on the system.
sql-server	This package group includes packages useful for use with Postgresql.
mysql	This package group contains packages useful for use with MySQL.
web-server	These tools allow you to run a Web server on the system.
dns-server	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.

Bundle Name	Description
engineering-and-scientific	This group includes packages for performing mathematical and scientific computations and plotting, as well as unit conversion.
editors	Sometimes called text editors, these are programs that allow you to create and edit files. These include Emacs and Vi.
emacs	The GNU Emacs text editor.
xemacs	The XEmacs text editor.
ruby	Basic support for the Ruby programming language.
system-tools	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.
admin-tools	This group is a collection of graphical administration tools for the system, such as for managing user accounts and configuring system hardware.
games	Various ways to relax and spend your free time.
ISO8859-2-support	
ISO8859-9-support	
ISO8859-14-support	
ISO8859-15-support	
cyrillic-support	
syriac-support	
afrikaans-support	
british-support	
canadian-support	
catalan-support	
brazilian-support	
czech-support	

Bundle Name	Description
danish-support	
dutch-support	
estonian-support	
finnish-support	
german-support	
greek-support	
hebrew-support	
hungarian-support	
spanish-support	
french-support	
icelandic-support	
italian-support	
korean-support	
norwegian-support	
polish-support	
portuguese-support	
romanian-support	
russian-support	
serbian-support	
slovak-support	
slovenian-support	
swedish-support	
turkish-support	
ukrainian-support	
chinese-support	
japanese-support	
development-tools	These tools include core development tools such as automake, gcc, perl, python, and debuggers.
development-libs	The packages in this group are core libraries needed to develop

Bundle Name	Description
	applications.
kernel-development	Install these packages to recompile the kernel.
legacy-software-development	These packages provide compatibility support for previous releases of Red Hat Enterprise Linux.
compat-arch-support	Multilib support packages
compat-arch-development	Support for developing packages for the non-primary architecture
legacy-software-support	
unsupported-devel-libs	
x-software-development	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.
workstation-common	
server	
gnome	
kde	
miscallvars	

E. Installation of Digiboard PortServer TS4 and TS16 for Linux

The objective of this appendix is to describe how to Configure PS and serial lines in order to access the Linux console and EFI of Fame by COM2 EFI. In this appendix PortServer TS 16 is currently named PS.

E.1 Reminder: Configuration of Linux console and kdb debugger on client Fame

E.1.1 Boot Option in elilo.conf

The boot option in elilo.conf is:

- if no KDB

```
append="console=tty0 console=ttyS1,115200"
```

- if KDB is used

```
append="console=tty0 console=ttyS1,115200 kdb=on"
```

Output will be available on all peripheral mentioned in the "console" option. Last one will be /dev/console.

Validate one getty in /etc/inittab, by adding the line:

```
S1:2345:respawn:/sbin/agetty 115200 ttyS1
```

E.1.2 Access with login root

Update file containing peripheral on serial lines authorized to connect with login root.

Add to /etc/securetty the line:

```
ttyS1
```

E.2 PortServer

E.2.1 Network Configuration

Update /etc/hosts with IP addresses and names for PS and all the serial lines. For one full PS, according to the model there are 5 or 17 IP addresses one for the PS and 4 or 16 for the lines.

For instance:

<PS_address_IP>< PS_name >
< port1_address_IP > < port1_name>
< port2_address_IP > < port2_name>
...

E.2.2 Change to command line mode in order to configure the serial ports

Note: In command line mode, useful keys are following:

Backwards	Control b
Forwards	Control f
Delete character at the left of the cursor	Backspace or Control h
Delete character on the cursor	Delete
Scroll back command	Control p
Scroll forwards next command	Control n
Execute	Enter

On a new PS three ways for change to command-line mode, The first one (1) using an ascii terminal and the two following methods describing how to connect to the PS via Telnet using its IP address automatically attributed For methods 2 and 3 you must be connected with login = root / passwd = dbps

1. Using Ascii Terminal

Use a terminal connected at one serial port with default setting:
emulation. VT 100
speed 9600 bps
character 8-bit

1 stop bit
without parity

2. Using Telnet and MAC address

Connect as root (password dbps).

```
#> set config ip=< IP_adress>
#> set config myname=<PS_name>
```

The MAC address is written on a sticker under the PS.

From the server on which IP address of PS is registered run ARP-Ping:

```
arp -s <IP_address> <MAC_address>
ping <IP_address>|<PS_name>
```

Syntax for MAC address is ii:jj and not ii-jj-..

If time out on ping, try again.

Example:

```
$ arp -s 172.16.110.200 00:40:9D:23:C1:F2
$ ping 172.16.110.200
PING portservm (172.16.110.200) from 172.16.110.112 : 56(84) bytes of data.
64 bytes from portservm (172.16.110.200): icmp_seq=14 ttl=64 time=13.471
msec
64 bytes from portservm (172.16.110.200): icmp_seq=15 ttl=64 time=4.194 msec
64 bytes from portservm (172.16.110.200): icmp_seq=16 ttl=64 time=2.031 msec
```

3. Using Telnet and DHCP

Connect as root (password dbps).

Create an permanent address entry in /etc/dhcpd.conf for the PS:

```
host <PS_name> {
fixed-address <IP_address>;
hardware ethernet <MAC_address>;
}
```

Example:

```
host portservm {
fixed-address 172.16.110.200;
hardware ethernet 00:40:9D:23:C1:F2;
}
```

Switch on PS

"Ping" PS until it answers.

E.2.3 Configure serial lines of PortServer

From a telnet session (command line mode).

Basis command for configuration is "set".
First argument is a level 2 command which combines options in a logical way.
When there is no argument the set command runs in display mode.

If necessary complete the PS general configuration like Network Subnet Mask (netmask) or gateway and display:

```
#> set config submask=<netmask>
#> set config gateway=<gateway>
#> set config
```

Adjust Ethernet communication parameters according to the other part of the network:

```
#> set ethernet duplex=auto speed=auto
```

possible: duplex=half|full|auto , speed=10|100|auto (10/100 Mbps)
default: duplex=half , speed=auto

Configure the lines

In the following commands <n> represents the number of serial ports. According with the model n=4 for TS4 and 16 for TS16.

A multi-port option is available when the value of a parameter is the same for several ports:

```
option range=i-j,k,l-m
```

Note: Normally port number is specified by "range=i". There are some exceptions for instance for "set altip" where portnumber is specified "group=i".

- Define device type connected to port(s)
  - Port 1:

```
#> set port dev=prn range=1
```

  - All Ports:

```
#> set port dev=prn range=1-<n>
```

- Define line(s) speed.
  - Port 1:

```
#> set line baud=115200 range=1
```

  - All Ports:

```
#> set line baud=115200 range=1-<n>
```

- Define flow control for the line(s).
  - Port 1:

```
#> set flow ixoff=off ixon=off rts=on cts=off forcedcd=on range=1
```

  - All Ports:

```
#> set flow ixoff=off ixon=off rts=on cts=off forcedcd=on range=1-<n>
```

- Assign IP address to the ports.
  - Port 1:

```
#> set altip ip=<port1_IP_address> group=1
```

  - Port 2:

```
#> set altip ip=<port2_IP_address> group=2
etc ...
```

- Give a host name to every port.
  - Port 1:

```
#> set host ip=<port1_IP_address> name=<port1_name>
```

  - Port 2:

```
#> set host ip=<port2_IP_address> name=<port2_name>
```

etc ...

E.2.4 Other useful commands

- Back-up PS configuration on a server using tftp.
  On the server, create empty file on /tftpboot with write permission:

```
$ cd /tftpboot
$ > <config_PS>
$ chmod 666 <config_PS>
```

From a telnet session (command line mode):

```
#> cpconf tohost <server_IP_address> <config_PS>
```

Note: Configuration file <config_PS> is in text format . You can modify and correct with a text editor (VI, …).

- It's possible to restore the configuration

```
#> cpconf fromhost <server_IP_address> <config_PS>
```

- Reset PS

```
#> boot action=reset
```

- Restore default parameters

```
#> boot action=factory
```

- View configurations

```
#> show <option>
```

  – show without option gives a list of all options

- View firmware release

```
#> show version
```

- Statistics

```
#> info <option>
```

  – info ? for all the options available

- View statistics for a serial port

```
#> info serial:<i>
```

- View statistics for the network

```
#> info ip
#> info ethernet
#> info network
```

For more details about commands, options of command see document Digiboard *"Command Reference DIGI TS Family" 92000304_L.*

E.2.5 Documentation

In complement with Bull user's guide, see
http://www.kde.org/documentation/userguide/kdebase-applications.html

F. Installation of Digiboard AccelePort C/X and Xr 920 Adapters

F.1 Package installation

The rpm file is drv-epca-1.50-1.b.1.Bull.ia64.rpm

1. Insert the CD-ROM,
2. Run the following commands:

```
mkdir -p /mnt/cdrom
mount /mnt/cdrom
```

then :

```
rpm -ivh /mnt/cdrom/???/drv-epca-1.50-1.b.1.Bull.ia64.rpm
```

When Digiboard AccelePort cards are present on the system, `lspci` command should display something such:

```
lspci | grep -i digi
05:01.0 Communication controller: Digi International AccelePort Xr
(rev 01)
0c:01.0 Communication controller: Digi International AccelePort
C/X (rev 01)
```

F.2 Example of epca driver configuration for a 8 ports "Digi International AccelePort Xr"

Run the following command:

```
/usr/sbin/digiConf
```

You should get the following menu:

```
    How many boards would you like to install? (1-12) 1
    Great! we'll install 1 board/s for you.

    Board #1. What type of board is this? ('L' for list) (1-14) L
           1: Acceleport Xe ISA
           2: Acceleport Xr ISA
           3: Acceleport Xem ISA
           4: Acceleport Xi ISA
           5: Acceleport C/X ISA
           6: Acceleport Xem PCI
           7: Acceleport Xr PCI
           8: Acceleport C/X PCI
           9: Acceleport Xr(PLX) PCI
           10: Acceleport Xr-422
           11: Acceleport 2r-920 PCI
           12: Acceleport 4r-920 PCI
           13: Acceleport 8r-920 PCI     ←
           14: Acceleport EPC/X PCI

    Board #1. What type of board is this? ('L' for list) (1-14) 13
    Great! You've selected to install a Acceleport 8r-920 PCI
board!
    Memory addresses will be read from the PCI card itself.
    This digiBoard has 8 ports
    Do you want to set Altpin on this board?  ('y' or 'n') n
    Great! we have ALL the information we need, to get this
digiBoard running!!
```

F.3 Example of epca driver configuration for a 128 ports "Digi International AccelePort C/X"

On the card there are two connectors corresponding to the term "line" in the configuration dialogue. Let us assume that "line 1" is the lower connector.

Example with 1 RAN (cable on the lower connector)

Run the following command:

```
/usr/sbin/digiConf
```

You should get the following menu:

```
      How many boards would you like to install? (1-12)1
      Great! we'll install 1 board/s for you.

      Board #1. What type of board is this? ('L' for list) (1-14)L
           1: Acceleport Xe ISA
           2: Acceleport Xr ISA
           3: Acceleport Xem ISA
           4: Acceleport Xi ISA
           5: Acceleport C/X ISA
           6: Acceleport Xem PCI
           7: Acceleport Xr PCI
           8: Acceleport C/X PCI              ←
           9: Acceleport Xr(PLX) PCI
           10: Acceleport Xr-422
           11: Acceleport 2r-920 PCI
           12: Acceleport 4r-920 PCI
           13: Acceleport 8r-920 PCI
           14: Acceleport EPC/X PCI

      Board #1. What type of board is this? ('L' for list) (1-14)8
      Great! You've selected to install a Acceleport C/X PCI
board!

      Memory addresses will be read from the PCI card itself.

      How many ports does this digiBoard have?  Possible values:
           1: 8
           2: 16    ← One Ran of 16 ports
           3: 24
           4: 32
           5: 40
           6: 48
           7: 56
           8: 64
           9: 72
```

```
            10: 80
            11: 88
            12: 96
            13: 104
            14: 112
            15: 120
            16: 128
      Board #1. How many ports?  (1-16)2
      Do you want to set Altpin on this board?  ('y' or 'n')n
      Great! we have ALL the information we need, to get this
digiBoard running!!
      cxconf version Version 1.0.6
      Installing support for 1 C/X card
      How many C/CON's are connected to card 1,  line 1? 1

      What type of wiring scheme are you going to use for card 1,
line 1?

      A) 8 Wire Direct    ←
      B) 4 Wire Direct
      C) RS422 Sync
      D) RS232 Sync

      > A
       Enter the communication mode to use on line 1
       (Type 'L' for a list) [14] : L

      Mode    Bit Rate        Clocking Mode
      0       115K    8-wire internal clock
      3       2400    8-wire internal clock
      4       4800    8-wire internal clock
      5       9600    8-wire internal clock
      6       19.2K   8-wire internal clock
      7       38.4K   8-wire internal clock
      8       57.6K   8-wire internal clock
      9       76.8K   8-wire internal clock
      10      115K    8-wire internal clock
      11      230K    8-wire internal clock
      12      460K    8-wire internal clock
      13      920K    8-wire internal clock
      14      1.2M    8-wire internal clock (*** most common for
C/X ***)      ←
      70      1.843M  8-wire internal clock ( EPC/X Only! )
      71      2.458M  8-wire internal clock ( EPC/X Only! )
      72      3.686M  8-wire internal clock ( EPC/X Only! )
      73      7.373M  8-wire internal clock ( EPC/X Only! )
      74      10M     8-wire internal clock ( EPC/X Only! )

          Enter the communication mode to use on line 1
          (Type 'L' for a list) [14] : 14
          How many ports does this C/CON support? (conc #1)
```

```
    (NOTE: The maximum ports here, is 16) [16] : 16
How many C/CON's are connected to card 1,  line 2? 0
```

F.4 Loading the epca driver

Run the following command:

```
/etc/epca start
```

The following messages should be displayed when a 8 ports has been configured:

```
Initializing card #0(Acceleport 8r-920 PCI):
   Downloading xrbios.bin to 0000000000001000 on Acceleport 8r-920
PCI
   Downloading xrfep.bin to 0000000000001000 on Acceleport 8r-920
PCI
Number of ports:
...Card #0(Acceleport 8r-920 PCI): card reports 8 ports found
Creating device nodes:
...Creating device nodes for card #0
 ..From ttyD000 to ttyD007
   digiDload complete.
```

The following messages should be displayed when a 128 ports has been configured:

```
digiDload version 1.3.26
Initializing card #0(Acceleport C/X PCI):
Downloading cxpbios.bin to 0000000000001000 on Acceleport C/X PCI
Downloading cxpfep.bin to 0000000000001000 on Acceleport C/X PCI
Downloading concentrator image:
Downloading /usr/lib/dg/epca/firmware/cxcon.bin to EPCA Memory
space
Number of ports:
.....Card #0(Acceleport C/X PCI): card reports 16 ports found
Creating device nodes:
.....Creating device nodes for card #0
.....From ttyD000 to ttyD015
    digiDload complete.
```

G. Acronyms

API	Application Programmer Interface
BAS	Bull Advanced Server
BIOS	Basic Input Output System
BMC	Baseboard Management Controller
B-SPS	Bull Scalable Port Switch
CMOS	Complementary Metal Oxyde Semiconductor
EFI	Extensible Firmware Interface (Intel)
EIP	Encapsulated IP
EMP	Emergency Management Port
EPIC	Explicit Parallel Instruction set Computing
EULA	End User License Agreement (Microsoft)
FRU	Field Replaceable Unit
FSS	Fame Scalability Switch
GCC	GNU C Compiler
GNU	GNU's Not Unix
GPL	General Public License
GUI	Graphical User Interface
GUID	Globally Unique Identifier
HDD	Hard Disk Drive
HPC	High Performance Computing
HSC	Hot Swap Controller
IDE	Integrated Device Electronics
IPMI	Intelligent Platform Management Interface
KSIS	Utility for Image Building and Deployment

KVM	Keyboard Video Mouse (allows to connect the keyboard, video and mouse either to the PAP, either to the node)
LUN	Logical Unit Number
MPI	Message Passing Interface
NFS	Network File System
NPTL	Native POSIX Thread Library
NTFS	New Technology File System (Microsoft)
NUMA	Non Uniform Memory Access
NVRAM	Non Volatile Random Access Memory
OEM	Original Equipment Manufacturer
OPK	OEM Preinstall Kit (Microsoft)
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
QBB	Quad Brick Block
ROM	Read Only Memory
SDR	Sensor Data Record
SEL	System Event Log
SCSI	Small Computer System Interface
SM	System Management
SMP	Symmetric Multi Processing
SSH	Secure Shell
VGA	Video Graphic Adapter

Index

P

R

S

T

V

X

Vos remarques sur ce document / Technical publications remarks form

Titre / Title :	**Bull HPC BAS3 Installation and Configuration Guide**

N° Référence / Reference No. :	86 A2 31EG Rev06

Date / Dated :	November 2005

ERREURS DETECTEES / ERRORS IN PUBLICATION

AMELIORATIONS SUGGEREES / SUGGESTIONS FOR IMPROVEMENT TO PUBLICATION

Vos remarques et suggestions seront attentivement examinées. Si vous désirez une réponse écrite, veuillez indiquer ci-après votre adresse postale complète.
Your comments will be promptly investigated by qualified personnel and action will be taken as required. If you require a written reply, furnish your complete mailing address below.

NOM / NAME : ..... DATE : .....

SOCIETE / COMPANY : .....

ADRESSE / ADDRESS : .....

.....

Remettez cet imprimé à un responsable Bull S.A. ou envoyez-le directement à :
Please give this technical publications remarks form to your Bull S.A. representative or mail to:

Bull S.A.
CEDOC
Atelier de reprographie
357, Avenue Patton BP 20845
49008 ANGERS Cedex 01
FRANCE

BULL CEDOC
357 AVENUE PATTON
BP 20845
49008 ANGERS CEDEX 01
FRANCE

ORDER REFERENCE
86 A2 31EG Rev06