

# BAS4 for Xeon

## Administrator's Guide





# HPC

# BAS4 for Xeon

## Administrator's Guide

### Software

December 2007

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

REFERENCE  
86 A2 83ET 02

The following copyright notice protects this book under Copyright laws which prohibit such actions as, but not limited to, copying, distributing, modifying, and making derivative works.

Copyright © Bull SAS 2007

Printed in France

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

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

### **Trademarks and Acknowledgements**

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

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

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

*The information in this document is subject to change without notice. Bull will not be liable for errors contained herein, or for incidental or consequential damages in connection with the use of this material.*

---

# Preface

## Scope and Objectives

The purpose of this guide is to explain how to configure and manage Bull High Performance Computing (HPC) **BAS4 for Xeon** clusters, using the administration tools recommended by Bull.

It is not in the scope of this guide to describe the Linux administration functions in depth. For this information, please refer to the standard Linux distribution documentation.

## Intended Readers

This guide is for **BAS4 for Xeon** cluster system administrators.

## Prerequisites

The installation of all hardware and software HPC components must have been completed.

## Structure

This guide is organized as follows:

- |            |                                                                                                                                                                                                             |
|------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Chapter 1. | Explains some <i>General Concepts</i> for Bull's Linux HPC systems.                                                                                                                                         |
| Chapter 2. | <i>HPC Configuration</i><br>Describes some basic configuration tasks including password management and security settings. It also describes how to run parallel commands and some kernel tuning parameters. |
| Chapter 3. | <i>Cluster Database Management</i><br>Describes the commands and the tools which enable the administrator to display and to change the Cluster Database.                                                    |
| Chapter 4. | <i>Parallel File Systems</i><br>Explains how these file systems operate on a Bull HPC system. It describes in detail how to install, configure and manage the <b>Lustre</b> file system.                    |
| Chapter 5. | <i>Software Deployment</i><br>Describes how to use <b>KSIS</b> to deploy, manage, modify and check software images.                                                                                         |
| Chapter 6. | <i>Resource Management</i><br>Explains how the <b>SLURM</b> Resource Manager works to help ensure the optimal management of the resources.                                                                  |
| Chapter 7. | <i>Batch Management with PBS Professional</i><br>Describes post installation checks and some useful commands for the PBS Professional Batch Manager.                                                        |

Chapter 8.	<i>Monitoring</i> Describes the <b>NovaScale Master - HPC Edition</b> monitoring tool for Bull HPC systems.
Chapter 9.	<i>Storage Devices Management</i> Explains how to setup the management environment for storage devices, and how to use storage management services.
Chapter 10.	<i>Kerberos – Network Authentication Protocol</i> Describes how to set up and use Kerberos.
Chapter 11.	<i>Profiling Programs – HPC Toolkit</i> Describes how to use HPC Toolkit. This provides a set of profiling tools that help you to improve the performance of the system.
Chapter 12.	<i>I/O Node and Lustre File System High Availability</i> Explains how to implement High Availability for I/O Nodes which use the <b>Lustre</b> file system.
Appendix A.	Lists the BIOS parameter settings for <b>NovaScale R421</b> and <b>R422</b> Compute Nodes
Glossary	Some of the acronyms and glossary terms for <b>BAS4 for Xeon</b> are detailed in the Glossary

## Bibliography

- Bull *HPC BAS4 for Xeon Installation and Configuration Guide* (86 A2 82ET).
- Bull *HPC BAS4 for Xeon User's Guide* (86 A2 91 ET).
- Bull *HPC BAS4 for Xeon Maintenance Guide* (86 A2 92 ET).
- Bull *NovaScale Master Remote Hardware Management CLI Reference Manual* (86 A2 88EM).
- The *Software Release Bulletin* (SRB) (86 A2 54 EJ) includes additional installation instructions and specific information for each software release.
- Bull *Voltaire Switches Documentation CD* (86 A2 79ET)
- EMC *Navisphere® Command Line Interface (CLI)* (300-003-628)
- StoreWay *Optima 1200 Quick Start Guide* (86 A1 34ET)
- StoreWay *Optima 1200 Installation and User Guide* (86 A1 35ET)
- StoreWay *Master User Guide* (86 A1 38ET)

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

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

## Highlighting

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

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

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



### Warning:

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





---

# Table of Contents

<b>Chapter 1.</b>	<b>General Concepts for Bull BAS4 for Xeon Clusters</b> .....	<b>1-1</b>
1.1	Introduction .....	1-1
1.2	Hardware Configuration .....	1-1
1.3	Typical Types of Nodes .....	1-1
1.4	Service node(s) .....	1-2
1.4.1	Management Node.....	1-3
1.4.2	Login Nodes.....	1-3
1.4.3	I/O Nodes .....	1-3
1.5	Compute Nodes.....	1-4
1.6	Networks .....	1-5
1.6.1	Administration Network .....	1-5
1.6.2	Backbone.....	1-5
1.6.3	Ethernet Network and Switch Management.....	1-6
1.7	Main Console and Hardware Management .....	1-6
1.7.1	System Console .....	1-6
1.7.2	Hardware Management .....	1-6
1.8	High Speed Interconnection.....	1-7
1.8.1	InfiniBand Networks with Voltaire Switching Devices .....	1-7
1.8.2	Ethernet Gigabit Networks .....	1-7
1.9	Program Execution Environment .....	1-8
1.9.1	Resource Management .....	1-8
1.9.2	Parallel processing and MPI libraries.....	1-8
1.9.3	Batch schedulers .....	1-9
1.10	Storage.....	1-9
1.11	BAS4 for Xeon Management Functions and Corresponding Products.....	1-10
<b>Chapter 2.</b>	<b>HPC Configuration</b> .....	<b>2-1</b>
2.1	Configuring Services .....	2-1
2.2	Modifying Passwords and Creating Users .....	2-2
2.3	Managing Partitions .....	2-3
2.4	Creating Swap Partitions.....	2-4
2.5	Configuring Security.....	2-5
2.5.1	Setting up SSH .....	2-5
2.6	Running Parallel Commands with pdsh .....	2-6
2.6.1	Using pdsh.....	2-6
2.6.2	Using pdcp .....	2-9
2.6.3	Using dshbak .....	2-9
2.7	Day to Day Maintenance Operations .....	2-11

<b>Chapter 3.</b>	<b>Cluster Database Management .....</b>	<b>3-1</b>
3.1	Architecture of ClusterDB .....	3-1
3.2	ClusterDB Administrator .....	3-2
3.3	Using Commands .....	3-2
3.3.1	ChangeOwnerProperties .....	3-2
3.3.2	dbmConfig .....	3-5
3.3.3	dbmCluster .....	3-7
3.3.4	dbmNode .....	3-8
3.3.5	dbmHwManager .....	3-11
3.3.6	dbmGroup .....	3-12
3.3.7	dbmEthernet .....	3-14
3.3.8	dbmlconnect .....	3-16
3.3.9	dbmTalim .....	3-17
3.3.10	dbmSerial .....	3-18
3.3.11	dbmFiberChannel .....	3-20
3.3.12	dbmServices .....	3-21
3.3.13	dbmDiskArray .....	3-22
3.4	Managing the ClusterDB .....	3-24
3.4.1	Saving and Restoring the Database .....	3-24
3.4.2	Starting and Stopping PostgreSQL .....	3-26
3.4.3	Viewing the PostgreSQL Alert Log .....	3-26
3.5	ClusterDB Modeling .....	3-27
3.5.1	Physical View of the Cluster Networks .....	3-27
3.5.2	Physical View of the Storage .....	3-35
3.5.3	Machine View .....	3-42
3.5.4	HWMANAGER View .....	3-48
3.5.5	Complementary Tables .....	3-50
3.5.6	NsDoctor View .....	3-52
3.5.7	Nagios View .....	3-53
3.5.8	Lustre View .....	3-55
<b>Chapter 4.</b>	<b>Parallel File Systems .....</b>	<b>4-1</b>
4.1	Parallel File Systems Overview .....	4-1
4.2	Lustre Overview .....	4-2
4.3	Lustre Administrator's Role .....	4-3
4.4	Planning a Lustre System .....	4-4
4.4.1	Data Pipelines .....	4-4
4.4.2	OSS / OST Distribution .....	4-4
4.4.3	MDS / MDT Distribution .....	4-4
4.4.4	File Striping .....	4-5
4.4.5	Lustre File System Limitations .....	4-5
4.5	Lustre System Management .....	4-6
4.5.1	The Lustre Database .....	4-6
4.5.2	/etc/lustre/storage.conf for Lustre Tools without ClusterDB .....	4-8
4.5.3	Lustre Networks .....	4-13
4.5.4	Lustre Management Configuration File: /etc/lustre/lustre.cfg .....	4-13

4.5.5	Lustre Services Definition.....	4-15
4.5.6	Creating Lustre File Systems.....	4-17
4.6	Installing and Managing Lustre File Systems .....	4-21
4.6.1	Installing Lustre File Systems using lustre_util.....	4-21
4.6.2	Removing Lustre File Systems using lustre_util.....	4-21
4.6.3	lustre_util Actions and Options .....	4-21
4.6.4	lustre_util Configuration File /etc/lustre/lustre_util.conf .....	4-32
4.6.5	Lustre Tuning File /etc/lustre/tuning.conf.....	4-34
4.6.6	Lustre Filesystem Reconfiguration.....	4-35
4.6.7	Using Quotas with Lustre File Systems.....	4-36
4.7	Monitoring Lustre System.....	4-39
4.7.1	Lustre System Health Supervision.....	4-39
4.7.2	Lustre Filesystem Indicator .....	4-41
4.7.3	Lustre System Performance Supervision .....	4-43
<b>Chapter 5.</b>	<b>Software Deployment (KSIS).....</b>	<b>5-1</b>
5.1	Overview .....	5-1
5.2	Configuring and Verifying a Reference Node.....	5-2
5.3	Main Steps for Deployment .....	5-2
5.4	Modifying Images and Managing their Release.....	5-3
5.4.1	Methods .....	5-3
5.4.2	Naming Images or Patches with the Workon Mechanism .....	5-5
5.4.3	Image Types.....	5-5
5.5	Checking Deployed Images.....	5-6
5.5.1	Checking Principles.....	5-6
5.5.2	Check Groups .....	5-6
5.5.3	Modifying the Checks Database .....	5-7
5.5.4	Examining the Results .....	5-8
5.5.5	Looking at the Discrepancies .....	5-8
5.6	Importing and Exporting an Image .....	5-9
5.7	Ksis Commands .....	5-10
5.7.1	Syntax .....	5-10
5.7.2	Advanced ksis create options .....	5-11
5.7.3	Creating the Image of the Reference Node .....	5-11
5.7.4	Deleting an Image or a Patch .....	5-12
5.7.5	Deploying an Image or a Patch .....	5-12
5.7.6	Removing a Patch .....	5-13
5.7.7	Getting Information about an Image or a Node.....	5-13
5.7.8	Listing Images on the Image Server .....	5-13
5.7.9	Listing Images by Nodes.....	5-13
5.8	Modifying an Image.....	5-15
5.8.1	Creating a Working Patch Image .....	5-15
5.8.2	Creating a Patch Image.....	5-15
5.8.3	Creating a Patched Golden Image.....	5-16
5.8.4	Building a Patch.....	5-16
5.9	Checking Images .....	5-17

5.10	Importing and Exporting Images.....	5-17
5.11	Rebuilding ClusterDB Data before Deploying an Image.....	5-17
<b>Chapter 6.</b>	<b>Resource Management .....</b>	<b>6-1</b>
6.1	Resource Management with SLURM.....	6-2
6.1.1	SLURM Key Functions .....	6-2
6.1.2	SLURM Components .....	6-3
6.1.3	SLURM Daemons.....	6-3
6.1.4	Scheduler Types.....	6-5
6.2	SLURM Configuration.....	6-7
6.2.1	Configuration Parameters.....	6-8
6.2.2	slurm.conf Example Files .....	6-21
6.2.3	SCONTROL – Managing the SLURM Configuration.....	6-23
6.2.4	Pam_Slurm Module Configuration .....	6-30
6.3	Administrating Cluster Activity with SLURM.....	6-32
6.3.1	Starting the Daemons.....	6-32
6.3.2	SLURMCTLD (Controller Daemon) .....	6-33
6.3.3	SLURMD (Compute Node Daemon) .....	6-34
6.3.4	Scheduler Support.....	6-35
6.3.5	Node Selection .....	6-36
6.3.6	Logging .....	6-36
6.3.7	Corefile Format .....	6-36
6.3.8	Security.....	6-36
6.3.9	SLURM Cluster Administration Examples .....	6-36
<b>Chapter 7.</b>	<b>Batch Management with PBS Professional.....</b>	<b>7-1</b>
7.1	Pre-requisites.....	7-1
7.2	Post Installation checks .....	7-2
7.2.1	Checking the status of the PBS daemons .....	7-2
7.2.2	Adding a Node to the Initial Cluster Configuration.....	7-2
7.3	Useful Commands for PBS Professional.....	7-2
7.4	Essential configuration settings for XBAS4 for Xeon clusters .....	7-3
7.4.1	MPIBull2 and PBS Pro for all clusters (InfiniBand and Ethernet).....	7-3
7.4.2	MPIBull2 and InfiniBand.....	7-4
<b>Chapter 8.</b>	<b>Monitoring with NovaScale Master - HPC Edition .....</b>	<b>8-1</b>
8.1	Launching NovaScale Master - HPC Edition.....	8-2
8.2	Access Rights .....	8-3
8.2.1	Administrator Access Rights.....	8-3
8.2.2	Standard User Access Rights .....	8-3
8.2.3	Adding Users and Changing Passwords .....	8-3
8.3	Hosts, Services and Contacts for Nagios.....	8-4
8.4	Using NovaScale Master - HPC Edition .....	8-5
8.4.1	NovaScale Master - HPC Edition – View Levels.....	8-5
8.5	Map Button.....	8-6

8.5.1	All Status Map View.....	8-6
8.5.2	Rack View.....	8-7
8.5.3	Host Services detailed View .....	8-7
8.5.4	Ping Map View.....	8-8
8.6	Status Button .....	8-9
8.7	Alerts Button .....	8-10
8.7.1	Active Checks.....	8-11
8.7.2	Passive Checks .....	8-11
8.7.3	Notifications.....	8-12
8.7.4	Acknowledgments.....	8-12
8.7.5	Comments.....	8-13
8.7.6	Logs .....	8-14
8.7.7	Alert Definition .....	8-14
8.7.8	Running a Script .....	8-15
8.7.9	Generating SNMP Alerts .....	8-16
8.7.10	Resetting an Alert Back to OK.....	8-16
8.7.11	nsmhpc.conf Configuration file .....	8-16
8.8	Storage Overview .....	8-17
8.9	Shell.....	8-18
8.10	Monitoring the Performance - Ganglia Statistics .....	8-18
8.11	Group Performance View .....	8-18
8.12	Global Performance View .....	8-19
8.12.1	Modifying the Performance Graphics Views.....	8-20
8.12.2	Refresh Period for the Performance View Web Pages .....	8-21
8.13	Configuring and Modifying Nagios Services.....	8-22
8.13.1	Configuring Using the Database.....	8-22
8.13.2	Modifying Nagios Services .....	8-22
8.13.3	Changing the Verification Frequency.....	8-23
8.14	General Nagios Services .....	8-24
8.14.1	Ethernet Interfaces.....	8-24
8.14.2	Resource Manager Status.....	8-24
8.14.3	Hardware Status .....	8-24
8.14.4	Alert Log .....	8-24
8.14.5	I/O Status.....	8-24
8.14.6	Postbootchecker.....	8-24
8.15	Management Node Nagios Services.....	8-25
8.15.1	MiniSQL Daemon .....	8-25
8.15.2	Resource Manager Daemon .....	8-25
8.15.3	ClusterDB.....	8-25
8.15.4	Cron Daemon.....	8-25
8.15.5	Compute Power Available.....	8-25
8.15.6	Global Filesystems bandwidth available .....	8-25
8.15.7	Storage Arrays available .....	8-25
8.15.8	Global Filesystem Usage.....	8-26
8.15.9	I/O pairs Migration Alert.....	8-26
8.15.10	Backbone Ports Available .....	8-26

8.15.11	HA System Status .....	8-26
8.15.12	Kerberos KDC Daemon .....	8-26
8.15.13	Kerberos Admin Daemon .....	8-26
8.15.14	LDAP Daemon (Lustre clusters only) .....	8-26
8.15.15	Lustre filesystems access .....	8-27
8.15.16	NFS filesystems access .....	8-27
8.15.17	InfiniBand Links available .....	8-27
8.16	Ethernet Switch Services .....	8-28
8.16.1	Ethernet Interface .....	8-28
8.16.2	Power supply .....	8-28
8.16.3	Temperature .....	8-28
8.16.4	Fans .....	8-29
8.16.5	Ports .....	8-29
8.17	More Nagios Information .....	8-29
<b>Chapter 9.</b>	<b>Storage Device Management.....</b>	<b>9-1</b>
9.1	Overview of Storage Device Management for Bull HPC Clusters .....	9-2
9.2	Monitoring Node I/O Status.....	9-4
9.2.1	I/O Counters Definitions .....	9-6
9.3	Monitoring Storage Devices.....	9-7
9.3.1	NovaScale Master - HPC Edition: Host and Service Monitoring for Storage Devices .....	9-7
9.3.2	NovaScale Master - HPC Edition: Storage & I/O Information .....	9-12
9.3.3	Querying the Cluster Management Data Base .....	9-17
9.4	Monitoring Brocade Switch Status .....	9-19
9.5	Managing Storage Devices with Bull CLI .....	9-22
9.5.1	Bull FDA Storage Systems.....	9-22
9.5.2	DataDirect Networks Systems - DDN Commands .....	9-23
9.5.3	Bull Optima 1200 Storage Systems .....	9-25
9.5.4	EMC/Clariion (DGC) Storage Systems .....	9-26
9.6	Using Management Tools .....	9-26
9.7	Configuring Storage Devices .....	9-27
9.7.1	Planning Tasks .....	9-27
9.7.2	Deployment Service for Storage Systems .....	9-28
9.7.3	Understanding the Configuration Deployment Service.....	9-28
9.8	User Rights and Security Levels for the Storage Commands.....	9-32
9.8.1	Management Node .....	9-32
9.8.2	Other Node Types .....	9-33
9.8.3	Configuration Files .....	9-33
<b>Chapter 10.</b>	<b>Kerberos - Network Authentication Protocol .....</b>	<b>10-1</b>
10.1	Environment.....	10-1
10.1.1	Kerberos Infrastructure .....	10-1
10.1.2	Validating the Installation .....	10-1
10.1.3	Authentication of the SSH V2 Connections .....	10-1
10.2	KERBEROS Infrastructure Configuration .....	10-2

10.2.1	secu0 Server including KDC Server and Administration Server .....	10-2
10.2.2	Configuration Files .....	10-2
10.2.3	Creating the Kerberos Database .....	10-3
10.2.4	Creating the Kerberos Administrator.....	10-3
10.2.5	Starting the KDC Server.....	10-3
10.2.6	Adding Access Control List (ACL) Rights for the Kerberos Administrator Created.....	10-4
10.2.7	Starting the Administration Daemon .....	10-4
10.2.8	Creating Principals Associated with Users.....	10-4
10.2.9	Creating Principals Associated with Remote Kerberized Services .....	10-5
10.3	Configuring the secu1 Machine Hosting the Remote Service 'host principal'.....	10-6
10.3.1	Generating the Key Associated with the Remote Service 'host principal' .....	10-6
10.4	Validating Kerberos Authentication for the Telnet Service .....	10-7
10.5	Kerberos Authentication and SSH.....	10-8
10.5.1	Configuring the Server SSH on the Machine secu1 .....	10-8
10.5.2	SSH Client .....	10-10
10.6	Troubleshooting Errors .....	10-12
10.7	Generating Associated Keys for Nodes of a Cluster .....	10-13
10.8	Modifying the Lifespan and Renewal Period for TGT Tickets .....	10-14
10.9	Including Addresses with Tickets .....	10-14
<b>Chapter 11.</b>	<b>Profiling Programs - HPC Toolkit.....</b>	<b>11-1</b>
11.1.1	HPC Toolkit Tools.....	11-1
11.1.2	Display Counters.....	11-1
11.1.3	Using HPC Toolkit.....	11-2
11.1.4	More Information .....	11-7
<b>Chapter 12.</b>	<b>I/O Node and Lustre File System High Availability .....</b>	<b>12-1</b>
12.1	Introduction to Lustre File System .....	12-1
12.2	Lustre Failover Mechanism.....	12-2
12.3	Hardware Architecture.....	12-4
12.4	High Availability Policy.....	12-6
12.5	High Availability Management .....	12-7
12.6	Error Detection and Prevention Mechanisms .....	12-9
12.7	Analysis of Failure Modes .....	12-10
12.7.1	I/O Node and Metadata Failures .....	12-10
12.7.2	Storage Failures.....	12-10
12.7.3	Ethernet Network Failures .....	12-11
12.8	Using Cluster Suite .....	12-12
12.8.1	Distributing the cluster.conf file on the I/O Node.....	12-12
12.8.2	Starting / Stopping Cluster Suite's Daemons .....	12-13
12.8.3	Checking the Cluster Suite Status .....	12-13
12.9	Managing Lustre High Availability .....	12-14
12.9.1	ClusterDB Information .....	12-14

12.9.2	LDAP Directory – the lustre_ldap Utility.....	12-14
12.9.3	Failover Tools Configuration – the /etc/lustre/lustre.cfg File.....	12-16
12.9.4	Managing Lustre Failover Services on I/O and Metadata Nodes – the lustre_migrate Tool .....	12-16
12.9.5	Configuring File Systems for Failover .....	12-17
12.10	Lustre High Availability Operations.....	12-18
12.10.1	Service Migration triggered by Cluster Suite.....	12-18
12.10.2	Service Migration triggered by Administrator .....	12-19
12.11	Monitoring Lustre High Availability.....	12-20
12.11.1	Command Line Monitoring .....	12-20
12.11.2	Graphic Monitoring.....	12-21
12.11.3	Traces and Debug.....	12-22

<b>Appendix A.</b>	<b>BIOS Parameter Settings to use for NovaScale R421 and R422 Compute Nodes .....</b>	<b>A-1</b>
A.1	NovaScale R421 BIOS Settings .....	A-1
A.1.1	Example BIOS Parameter Settings for NovaScale R421 .....	A-2
A.2	NovaScale R422 BIOS Settings .....	A-6
A.2.1	Example BIOS Parameter Settings for NovaScale R422 .....	A-6
<b>Glossary and Acronyms .....</b>	<b>G-1</b>	
<b>Index.....</b>	<b>I-1</b>	



---

## List of Figures

Figure 1-1.	NovaScale R440 machine.....	1-2
Figure 1-2.	NovaScale R460 machine.....	1-2
Figure 1-3.	NovaScale R421 machine.....	1-4
Figure 1-4.	NovaScale R422 machine.....	1-5
Figure 3-1.	BAS4 for Xeon ClusterDB architecture .....	3-1
Figure 3-2.	Cluster Network – diagram 1.....	3-27
Figure 3-3.	Cluster Network – diagram 2.....	3-28
Figure 3-4.	Storage physical view .....	3-35
Figure 3-5.	Cluster Database – Machine view 1 .....	3-42
Figure 3-6.	Cluster Database – Machine view 2 .....	3-43
Figure 3-7.	HWManager view.....	3-48
Figure 3-8.	Cluster Database – Complementary tables.....	3-50
Figure 3-9.	Cluster Database – NsDoctor view .....	3-52
Figure 3-10.	ClusterDB –Nagios View .....	3-53
Figure 3-11.	Cluster Database – Lustre view .....	3-55
Figure 4-1.	NovaScale Master Map view.....	4-39
Figure 4-2.	NovaScale Nagios file system indicator .....	4-41
Figure 4-3.	Lustre Management Node web interface .....	4-42
Figure 4-4.	Detailed view of Lustre file systems.....	4-43
Figure 4-5.	Group performance global view pop up window .....	4-44
Figure 4-6.	Dispatched performance view pop up window .....	4-44
Figure 4-7.	Global performance view pop up window .....	4-45
Figure 5-1.	Main steps for deployment.....	5-3
Figure 5-2.	Image modification (workon, store, deploy, detach).....	5-4
Figure 5-3.	Names of derived images or patches.....	5-5
Figure 5-1.	SLURM Simplified Architecture .....	6-3
Figure 5-2.	SLURM Architecture - Subsystems.....	6-4
Figure 8-1.	NovaScale Master - HPC Edition opening view .....	8-5
Figure 8-2.	Map button all status opening view .....	8-6
Figure 8-3.	Rack view with the problems window at the bottom .....	8-7
Figure 8-4.	Host Service details .....	8-8
Figure 8-5.	Status overview screen .....	8-9
Figure 8-6.	Alert Window showing the different alert states.....	8-11
Figure 8-7.	Hostgroups Reporting Notifications Window showing the Notification Levels.....	8-12
Figure 8-8.	Status Monitoring Control window showing the links to add and delete comments .....	8-13
Figure 8-9.	Monitoring Service Status window for a host. ....	8-14
Figure 8-10.	Storage overview window .....	8-17
Figure 8-11.	Group Performance view.....	8-18
Figure 8-12.	Global overview for a host (top screen).....	8-19
Figure 8-13.	Detailed monitoring view for a host (bottom half of screen displayed in Figure 8-12) .....	8-20
Figure 8-14.	Ethernet Switch services.....	8-28
Figure 8-1.	I/O Status Details NovaScale Master HPC Edition example screens .....	9-5
Figure 8-2.	Detailed service status for a storage host .....	9-8
Figure 8-3.	Storage overview .....	9-14
Figure 8-4.	Inventory view of faulty storage systems and components .....	9-15
Figure 8-5.	Storage detailed view .....	9-16
Figure 8-6.	Nodes I/O Overview.....	9-17

Figure 8-7.	Detailed Service status of a brocade switch .....	9-21
Figure 11-1.	View of the counter values, using <b>hpcviewer</b> .....	11-7
Figure 12-1.	Lustre interactions.....	12-1
Figure 12-2.	OST takeover and client recovery .....	12-2
Figure 12-3.	MDT takeover and client recovery.....	12-2
Figure 12-4.	I/O Cell diagram .....	12-4
Figure 12-5.	High Availability/Cluster Suite on NovaScale R440 and R460 IO/MDS nodes .....	12-5
Figure 12-6.	MDT/OST Dispatching on two nodes.....	12-5
Figure 12-7	Lustre High-Availability Management architecture .....	12-7
Figure 12-8	Service migration triggered by Cluster Suite.....	12-18
Figure 12-9	Service migration triggered by the Administrator .....	12-19
Figure 12-10	NovaScale Master Map all status screen .....	12-21
Figure 12-11	Lustre filesystem status indicator in the Host service status window .....	12-22
Figure A-1.	Example BIOS parameter setting screen for NovaScale R421 .....	A-1
Figure A-2.	Example BIOS parameter setting screen for NovaScale R422 .....	A-6

---

## List of Tables

Table 1-1.	BAS4 for Xeon Cluster Types.....	1-2
Table 2-1.	Maintenance Tools .....	2-11
Table 3-1.	Cluster Table.....	3-29
Table 3-2.	IP_NW table.....	3-29
Table 3-3.	ETH_SWITCH Table.....	3-30
Table 3-4.	IC_NW Table .....	3-30
Table 3-5.	IC_Switch Table .....	3-31
Table 3-6.	Serial_NW Table .....	3-31
Table 3-7.	PORTSERVER Table.....	3-32
Table 3-8.	ETH_VLAN table.....	3-32
Table 3-9.	FC_NW table .....	3-33
Table 3-10.	FC_SWITCH table .....	3-34
Table 3-11.	TALIM table .....	3-34
Table 3-12.	Storage – disk_array table.....	3-37
Table 3-13.	Storage – da_enclosure table.....	3-37
Table 3-14.	Storage – da_disk_slot table .....	3-38
Table 3-15.	Storage – da_controller table .....	3-38
Table 3-16.	Storage – da_fc_port.table .....	3-38
Table 3-17.	Storage – da_serial_port table .....	3-39
Table 3-18.	Storage – da_ethernet_port Table .....	3-39
Table 3-19.	Storage – da_power_supply table .....	3-40
Table 3-20.	Storage – da_fan table.....	3-40
Table 3-21.	Storage – da_power_fan table .....	3-40
Table 3-22.	Storage – da_temperature_sensor table.....	3-41
Table 3-23.	da_io_path table .....	3-41
Table 3-24.	Storage – da_iocell_component table .....	3-41
Table 3-25.	Storage – da_cfg_model table .....	3-41
Table 3-26.	Storage – da_power_port table .....	3-42
Table 3-27.	Machine view – node table .....	3-45
Table 3-28.	Machine view – Node_image table .....	3-45
Table 3-29.	Machine view – Node_Profile table .....	3-46
Table 3-30.	Machine view – IC_BOARD table .....	3-46
Table 3-31.	Machine view – IPOIB Table .....	3-47
Table 3-32.	Machine view – SDPOIB table .....	3-47
Table 3-33.	Machine view – FC_BOARD table .....	3-47
Table 3-34.	HWManager Table .....	3-49
Table 3-35.	Cluster Database – Admin table .....	3-50
Table 3-36.	Cluster Database – Rack table.....	3-51
Table 3-37.	Cluster Database – Config Candidate table.....	3-51
Table 3-38.	Cluster database – Config_Status table .....	3-51
Table 3-39.	Cluster Database Group_Node table .....	3-51
Table 3-40.	Cluster Database NsDoctor – Test_Groups table.....	3-52
Table 3-41.	Cluster Database NsDoctor – Tests table .....	3-52
Table 3-42.	Cluster Database NsDoctor – Test_Dependencies table .....	3-53
Table 3-43.	Cluster Database NsDoctor – Test Results table .....	3-53
Table 3-44.	Nagios Services Table .....	3-53
Table 3-45.	Nagios Availability Table.....	3-54

Table 3-46.	Cluster Database – Lustre View – Lustre_fs table .....	3-56
Table 3-47.	Cluster Database – Lustre view – Lustre OST table.....	3-57
Table 3-48.	Cluster Database – Lustre View – Lustre_MDT Table .....	3-57
Table 3-49.	Cluster Database – Lustre View – Lustre_IO_node table .....	3-58
Table 3-50.	Cluster Database – Lustre view – Lustre_mount table.....	3-58
Table 4-1.	Inode Stripe Data.....	4-4
Table 5-1.	Standard checks delivered with Ksis.....	5-7
Table 5-1.	Role Descriptions for SLURMCTLD Software Subsystems.....	6-4
Table 5-2.	SLURMD Subsystems and Key Tasks.....	6-5
Table 5-3.	SLURM Scheduler Types .....	6-6

---

# Chapter 1. General Concepts for Bull BAS4 for Xeon Clusters

## 1.1 Introduction

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

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

## 1.2 Hardware Configuration

**Bull BAS4 for Xeon** High Performance Computing systems feature different **NovaScale Xeon** machines for the nodes.

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



### Note:

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

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

## 1.3 Typical Types of Nodes

The **BAS4 for Xeon HPC** system supports various types of nodes, dedicated to specific activities. Depending on the cluster type, a single Service Node will include the Management Node functions with those of the Login and I/O nodes OR there will be 2 Service Nodes, one will be the dedicated Management Node and the other will be used for I/O and Login functions OR there will be 3 Service Nodes, one will be the dedicated Management Node, one will be used for I/O, and one will be used for Login functions.

Cluster Type	Function Node	Management	Input/Output	Login	Compute
1 Service Node with X Compute Nodes	Management Node	✓	✓		N/A
	Compute Node (s)	N/A	N/A		✓
2 Service Nodes with X Compute Nodes	Management Node	✓	N/A		N/A
	I/O + Login Node	N/A	✓	N/A	
	Compute Node (s)	N/A	N/A		✓
3 Service Nodes with X Compute Nodes	Management Node	✓	N/A		N/A
	I/O Node	N/A	✓	N/A	N/A
	Login Node	N/A	N/A	✓	N/A
	Compute Node (s)	N/A	N/A		✓

Table 1-1. BAS4 for Xeon Cluster Types

## 1.4 Service node(s)

**Bull NovaScale R440 and R460** 2 socket Xeon machines are used for the Service Nodes for **Bull BAS4 for XEON** Clusters.



Figure 1-1. NovaScale R440 machine

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



Figure 1-2. NovaScale R460 machine

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

Service Nodes include the following:

- **Management node(s)** to administrate and run the cluster machines.
- **Login node(s)** to provide access to the cluster and a specific software development environment.
- **I/O (Input/Output) node(s)** to transfer data to and from storage units.

As shown in Table 1-1 these node functions may be combined in 1 or 2 Service Nodes according to the architecture of the cluster.

Service Nodes use a distribution based on **Red Hat Enterprise Linux 4 Advanced Server**.

### 1.4.1 Management Node

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

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

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

### 1.4.2 Login Nodes

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

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

### 1.4.3 I/O Nodes

I/O Nodes provide a shared storage area to be used by the Compute Node when carrying out computations. Either the **NFS** or **Lustre** parallel file systems may be used to carry out the Input Output operations for BAS4 for Xeon clusters.



**Note:**

**Lustre** must use dedicated service nodes for the I/O functions. **NFS** can be used on both dedicated I/O service nodes and on combined Login/IO service nodes.



**Important**

**Lustre** is only possible for Clusters which install the Add-on functionality provided by the Bull **HPCK** software included on the **XHPC** DVD-ROM – see Chapter 2 for installation details.

## 1.5 Compute Nodes

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

- A Minimal Compute Node which includes minimal functionality and is quicker and easier to deploy.
- An Extended Compute Node, which includes additional libraries. Contact your Bull representative for more information on these nodes.

**Bull NovaScale R421** and **R422** machines are used for the Compute Nodes. **Bull NovaScale R422** machine includes 2 nodes.



Figure 1-3. NovaScale R421 machine



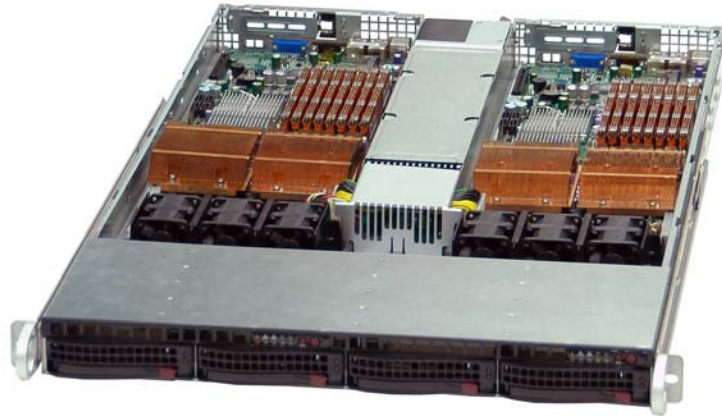


Figure 1-4. NovaScale R422 machine

The **Compute Nodes** are optimized to execute parallel code exclusively.

Interconnect Adapters (**InfiniBand** or Gigabit **Ethernet**) must be installed on these nodes.

Compute Nodes use a distribution based on **Red Hat Enterprise Linux 4 Work Station**.

## 1.6 Networks

The cluster may contain different networks, dedicated to particular functions, including:

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

### 1.6.1 Administration Network

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



**Note:**

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

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

### 1.6.2 Backbone

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

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

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

### 1.6.3 Ethernet Network and Switch Management

The switches can be directly managed by an **Ethernet** network.

Some useful parameters managed by the switch(es) are:

- multicast management
- **ARP** (Address Resolution Protocol) management
- Fast Spanning Tree protocol.

Check the manufacturer's documentation to see which options have to be set for the device.

## 1.7 Main Console and Hardware Management

### 1.7.1 System Console

The system console uses a Keyboard Video Mouse (KVM) switching device to control and administer the different machines within the cluster.

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

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

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

### 1.7.2 Hardware Management

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

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

The consoles, accessed using **ConMan**, provide:

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

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

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

**Bull NovaScale Master HPC Edition** provides all the monitoring functions for **BAS4 for Xeon** clusters using **Nagios**, an open source application for monitoring the status of all the cluster's components and will trigger alerts in the event of any problems. NovaScale Master uses **Ganglia**, a second open source tool, to collect and display graphically performance statistics for each cluster node.

## 1.8 High Speed Interconnection

### 1.8.1 InfiniBand Networks with Voltaire Switching Devices

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

- **400 Ex-D** Double Data Rate (**DDR**) Host Channel Adapters which can provide a bandwidth up to 20 Gbs per second, host device PCI-Express.
- **ISR 9024** switch with 24 DDR ports
- Clusters with up to 288 ports will use **Voltaire® ISR 9096** or **9288** or **2012 Grid Directors** to scale up machines which include **400 Ex-D HCAs** and **ISR 9024** switches.
- Clusters of more than 288 ports will be scaled up using a hierarchical switch structure based on the switches described above.

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

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

### 1.8.2 Ethernet Gigabit Networks

Ethernet Gigabit networks the interconnection generally uses **CISCO** switches as follows:

- The Host Channel Adapter will use one of the two native parts for each node.
- Clusters with less than 288 ports will use **CISCO catalyst 3560** (24 Ethernet + 4 SFP ports, 48 Ethernet +4 SFP ports) switches.

- Clusters with more than 288 ports will use a hierarchical switch structure based on the node switches described above, and with the addition of **Cisco Catalyst 650x** top switches (x= 3,6,9,13) which provide up to 528 ports.



For more information see *Appendix A* in the **BAS4 for Xeon Installation and Configuration Guide**.

## 1.9 Program Execution Environment

### 1.9.1 Resource Management

Both **Gigabit Ethernet** and **InfiniBand BAS4 for Xeon** clusters use the **SLURM** (Simple Linux Utility for Resource Management) open-source, highly scalable cluster management and job scheduling system. **SLURM** allocates compute resources, in terms of processing power and Computer Nodes to jobs for specified periods of time. If required the resources may be allocated exclusively with priorities set for jobs. **SLURM** is also used to launch and monitor jobs on sets of allocated nodes, and will also resolve any resource conflicts between pending jobs. Finally, **SLURM** helps to exploit the parallel processing capability of a cluster.



See the Bull HPC **BAS4 for Xeon Administrator's Guide** and *User's Guide* for more information on **SLURM**

### 1.9.2 Parallel processing and MPI libraries

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

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

Bull provides different **MPI** libraries for use in the HPC environment.

- **MPIBull2**, Bull's second generation **MPI** library, is included in the **Bull BAS4 for Xeon** delivery. This library enables dynamic communication with different device libraries, including **InfiniBand (IB)** interconnects, socket Ethernet/**IB**/**EIB** devices or single machine devices.
- Third party **MPI** libraries are also available. **MPICH\_Ethernet** is provided to allow applications to run using Ethernet interconnects. **BAS4 for Xeon** also uses **Voltaire MPI**, which in turn uses **MVAPICH**, an open-source **MPI** software library designed for **InfiniBand** clusters, which helps to ensure high performance and scalability for **MPI** applications.



See the Bull **BAS4 for Xeon User's Guide** for more information on Parallel Libraries

## 1.9.3 Batch schedulers

Different possibilities exist for handling batch jobs for **BAS4 for Xeon** clusters **PBS-Professional**, a sophisticated, scalable, robust Batch Manager from **Altair Engineering** is supported as a standard. **PBS Pro** can also be integrated with the **MPI** libraries.



See the Bull **BAS4 for Xeon** *User's Guide* for more information on Batch schedulers, the **PBS-Professional** *Administrator's Guide* and *User's Guide* available on the **PBS-Pro CD-ROM** delivered for the clusters which use **PBS-Pro**, and the **PBS-Pro** web site <http://www.pbsgridworks.com>.



### Important

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

## 1.10 Storage

Different storage systems are supported with **BAS4 for Xeon**. These include the following:

### Storeway 1500 and 2500 FDA Storage systems

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

Brocade Fibre Channel switches are supported to connect FDA storage units and help to ensure storage monitoring within **NovaScale Master HPC Edition**

### Storeway Optima 1200 Storage systems

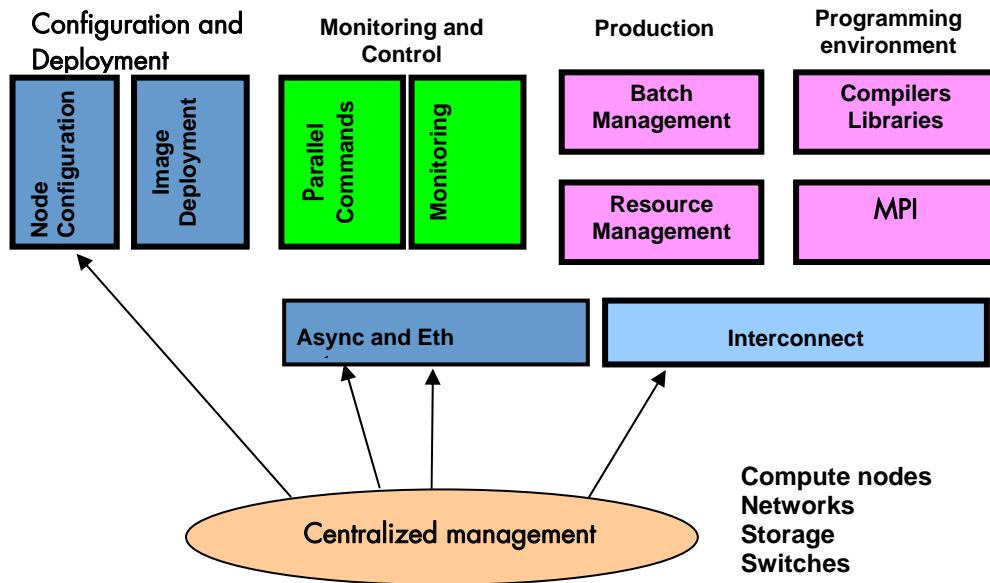
Developed on Fibre Channel standards for server connections and Serial Attached SCSI (SAS) standards for disk connections, the system can support high-performance disks and high-capacity SAS and SATA disks in the same subsystem. 2x 4Gb/s FC host ports per controller with a 3 Gb/s SAS channel with the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

### EMC/Clariion (DGC) CX3-40f storage system

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

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

## 1.11 BAS4 for Xeon Management Functions and Corresponding Products



**Bull** provides a software environment, **BAS4 for Xeon**, which helps to exploit the cluster and makes it extremely efficient. For example:

- The Message Passing Interface (**MPI**) allows the programs to run across all nodes.
- A resource manager controls access to resources distributed through the cluster. **SLURM** – an open-source resource manager is included with the **Bull BAS4 for Xeon** distribution.

The Bull cluster administration scheme is centralized on one node, the Management Node, and the administration products run on this node. All nodes are controlled and monitored from this central point of management, with the objective of ensuring that CPU activity and network traffic on the Compute and I/O nodes run as efficiently as is possible.

The management tools are mainly Open Source products. These products are configurable and adaptable to management needs, and can be deactivated on demand if necessary.

These products have been developed and adapted to Bull platforms and their environments. All management functions are available through a browser interface or through a remote command mode. Users can access the management functions according to their profile.

The management functions are performed by different products which are briefly presented below.

## Configuration and Software Management

- **pdsh** is used to run parallel commands.  
See Chapter 2 –*HPC Configuration* for more information.
- **KSIS** and the Ethernet network enable the deployment of images.  
See Chapter 5–*Software Deployment (KSIS)* for more information.
- The Cluster DataBase - **dbmConfig**, **dbmCluster**, **dbmNode** and other commands are available to manage the Cluster Database.  
See Chapter 3 –*Cluster DataBase Management* for more information.

## Resource Management

- **SLURM** (Simple Linux Utilities Resource Manager) an open-source scalable resource manager.  
See Chapter 6 – *Resource Management* for more information.

## Monitoring

- **NovaScale Master - HPC Edition** monitors the cluster and activity and is included in the delivery for all Bull HPC Clusters.  
See Chapter 8 – *Monitoring with NovaScale Master – HPC Edition* for more information.
- **HPC Toolkit** provides a set of profiling tools that help you to improve the performance of the system.  
See Chapter 11 – *Profiling Programs – HPC Toolkit* for more information

## Maintenance Tools

- **nsctrl** carries out various hardware and firmware tasks from the Management Node.
- **syslog-ng** manages the System Logs.
- **mkCDrec** performs system backups and restores. This function is available from the Service Node.





---

## Chapter 2. HPC Configuration

Most configuration tasks are carried out at the time of installation. This chapter indicates how the Administrator can perform some additional configuration tasks. It also deals with the security policy for HPC systems.

The following topics are described:

- 2.1 *Configuring Services*
- 2.2 *Modifying Passwords and Creating Users*
- 2.3 *Managing Partitions*
- 2.4 *Creating Swap Partitions*
- 2.5 *Configuring Security*
- 2.6 *Running Parallel Commands with pdsh*
- 2.7 *Day to Day Maintenance Operations*

For more information, refer to the *Bull HPC Installation and Configuration Guide*, which describes the different steps for installing and configuring Bull HPC systems.

### 2.1 Configuring Services

- To run a particular functionality when Linux starts enter the command:

```
/sbin/chkconfig --level 235 name_of_service on
```

- To display Help information enter the command:

```
/sbin/chkconfig --help
```

- To display the list of services available, enter the command:

```
/sbin/chkconfig --list
```



**Note:**

Some utilities, such as **sendmail** and **NFS**, are not enabled by default. The administrator is responsible for their configuration.

## 2.2 Modifying Passwords and Creating Users

Two users are created when Linux is installed:

**root**            administrator (password root)

**linux**           ordinary user (password linux)

These passwords must be changed as soon as possible:

- To change the passwords use one of the following commands
  - **passwd user\_name** command for root users
  - **passwd** command for ordinary users.
- To create new users enter the **/usr/sbin/useradd** command

```
useradd -g "group" -d "home login
```

## 2.3 Managing Partitions

This section explains how to add, delete or modify partitions.

Use the Linux `/sbin/parted` command to edit the GPT (GUID Partition Table) format of the disk. By default, the `parted` command loads the first disk `/dev/sda`.

To specify another disk (for example `/dev/sdb`), enter:

```
/sbin/parted /dev/sdb
```

- Run the `print` command to view the partitions table.
- Run the `help` command to view the commands.
- Run the `mkpartfs` command to create one or more partitions. For example:

```
mkpartfs primary ext2 6241.171 7184.955
```



### Note:

The `ext3` fs-type is not implemented in this version of `parted`, but the fs-type can be modified using the Linux `mkfs` command as described below.

- Run the `resize` command to modify the size of a partition.
- Delete a partition using the `rm <minor number>` command corresponding to the partition to be deleted.
- Run `q` to validate the changes.

Use the Linux `/sbin/mkfs` command to modify the file system type.

- For `ext3` file system run: `/sbin/mkfs -j <device>`. For example:

```
mkfs -j /dev/sdb8
```

- For other types, run: `/sbin/mkfs -t <fs-type> <device>`. For example:

```
mkfs -t ext2 /dev/sdc8
```

- Next, the mount points have to be defined in `/etc/fstab` file and these partitions mounted using the `/bin/mount -a` command so that the partitions will be mounted when the system is restarted.

## 2.4 Creating Swap Partitions

The `/sbin/mkswap` command lets you create swap partitions.

- Use the `/sbin/parted` command to edit the GPT format of the disk.
- Use the `mkpartfs` command to create one or more additional swap partitions. For example:

```
mkpartfs primary linux-swap 6241.171 7184.955
```

- Run the `mkswap` command.
- Run the `/sbin/swapon -a` command to take this swap into account.

For example, assuming your swap is on `/dev/sdc1`, run the following to recreate it with a larger size:

```
#!/sbin/swapoff -a
#!/sbin/parted -s /dev/sdc rm 1
#!/sbin/parted -s -- /dev/sdc mkpartfs primary linux-swap 0 <size of your disk in
    Mb, given by 'parted /dev/sdc p' 70000 for a 74 Gb disk for example>
#!/sbin/mkswap -p 65536 -f -v1 -L SWAP-sdc1 /dev/sdc1
#!/sbin/swapon -a
```

## 2.5 Configuring Security

This section provides the administrator with basic rules concerning cluster security. According to the cluster configuration you can set up different security policies.

The Management Node is the most sensitive element from a security point of view. This node will submit jobs in batch mode and it is a central point for management. This is the reason why security has to be enforced regarding access to this node. Very few people should be able to access this node and this access should be made using **OpenSSH** to eliminate eavesdropping, connection hijacking, and other network-level attacks effectively.

Compute node and I/O nodes should not have interactive logins. This means that no user except root should have access to these nodes. Management tools like Nagios will have access to both node types, while a batch manager like **PBS-Pro** will have access to compute nodes only.

If CPU and memory resources are shared among users, each user should not then have access to other partitions.

### 2.5.1 Setting up SSH

Carry out the following steps to set up **SSH** for an admin user:

1. Create a public key:

```
ssh-keygen -t dsa -N ''
```

This creates an **ssh** protocol 2 DSA certificate without passphrase in `~/.ssh/id_dsa.pub`.

2. Append this key to the list of authorized keys in `~/.ssh/authorized_keys2`.
3. Run **ssh** once by hand for each node responding yes at the prompt to add it to the list of known hosts:

```
atlas0: ssh atlas1 hostname
The authenticity of host 'atlas1 (192.168.84.2)' can't be
established.
RSA key fingerprint is
9c:d8:62:b9:14:0a:a0:18:ca:20:f6:0c:f6:10:68:2c.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'atlas1,192.168.84.2' (RSA) to the
list of known hosts.
```



#### NOTE:

For the root user there is an authorized keys file for each node as `~root/.ssh/authorized_keys2` is local. The new key must be appended to each of these files.

Please refer to the chapter in this manual on **Kerberos** for more information on **SSH** and the use of keys.

## 2.6 Running Parallel Commands with pdsh

A distributed shell is a tool that allows the same command to be launched on several nodes. Such a function is essential in a cluster environment so that instructions can be carried out on several nodes instead of running the command manually on each node in turn. Different tools can be used to enable this possibility.

**pdsh** is a utility that runs commands in parallel on all the nodes or on a group of nodes of the cluster. It is a very flexible tool especially for large cluster usage.

**pdsh** is a multi-threaded client for remote shell commands. It can use different remote shell services, such as **rsh**, **ssh** and **kerberos**.

Three utilities are included in **pdsh**:

- **pdsh** is used to run commands in parallel.
- **pdcp** is used to copy files on a group of nodes in parallel.
- **dshbak** is used to format, sort and display the results of a command launched with **pdsh**.

The **pdsh** utility relies on the security and authentication mechanisms provided by **ssh** and / or **Kerberos** V4 layers on which it is configured. See the chapter in this manual on Kerberos.

### 2.6.1 Using pdsh

#### Syntax:

The following commands are the ones which are used most often:

```
pdsh -R <rcmd_module> -w <node_list> -l user -Options Command
```

```
pdsh -R <rcmd_module> -a -x <node_list> -Options Command
```

```
pdsh -R <rcmd_module> -g <group_attributes> -Options Command
```

The most important options are described below. For a complete description of the options, refer to the **pdsh** man page.

#### Standard Target Node List Options:

**-w <node\_list>** Targets the specified list of nodes. Do not use the **-w** option with any other node selection option (**-a**, **-g**). The node list can be a comma-separated list (node 1, node2, etc.); no space is allowed in the list. If you specify only the **'** character, the target hosts will be read from stdin, one per line. The node list can also be an expression such as `host[1-5,7]`. For more information about node list expressions, see the **HOSTLIST EXPRESSIONS** in the **pdsh** man page.

**-x <node\_list>** Excludes the specified nodes. The **-x** option can be used with other target node list options (**-a**, **-g**, **-A**). The node list can be a comma-separated list (node1, node2, etc.); no space is allowed in the list. The node list can also be an expression such as host[1-5,7]. For more information about the node list expressions, see the HOSTLIST EXPRESSIONS in the pdsh man page.

### Standard pdsh Options:

- S** Displays the largest value returned by the remote commands.
- h** Displays commands usage and the list of the available rcmd modules and then quits.
- q** Lists the option values and the target node list and exits without action.
- b** Disables the Ctrl-C status feature so that a single Ctrl-C kills parallel jobs (Batch Mode).
- l <user>** This option is used to run remote commands as another user, subject to authorization.
- t <cnx\_timeout>** Sets the connection timeout (in seconds). Default is 10 seconds.
- u <exec\_time>** Sets a limit on the amount of time (in seconds) a remote command is allowed to execute. Default is no limit.
- f <remote\_cds\_num>**  
Sets the maximum number of simultaneous remote commands. Default is 32.
- R <rcmd\_module>**  
Sets the rcmd module to use. The list of the available rcmd modules can be displayed using the **-h**, **-V**, or **-L** options. The default module is listed with **-h** or **-V** options.  
Note: Instead of using this option, you can set the PDSH\_RCMD\_TYPE environment variable.
- L** Lists information about all loaded **pdsh** modules and then quits.
- d** Includes more complete thread status when SIGINT is received, and displays connection and command time statistics on stderr when done.
- V** Displays pdsh version information, along with the list of currently loaded pdsh modules.

### Group Attributes Options:

The following options use the cluster's group attributes as defined in the **/etc/genders** file.

- A** Targets all nodes defined in the **/etc/genders** file.

- a Targets all nodes in the /etc/genders file except those with the pdsh\_all\_skip group attribute.



**Note:**

The `pdsh -a` command is equivalent to the `pdsh -A -X pdsh_all_skip` command. For example, you can set the `pdsh_all_skip` group attribute to the Service Nodes to exclude these specific nodes from cluster.

- g <gp\_attr1[,gp\_attr2,...]> Targets the nodes that have any of the specified group attributes. This option cannot be used with the -a and -w options.
- X <gp\_attr1[,gp\_attr2...]> Excludes the nodes that have any of the specified group attributes. This option may be combined with any other node selection options (-w, -g, -a, -A).

**Examples:**

- To execute the `pwd` command on all the nodes of the cluster using the `ssh` protocol, enter:

```
pdsh -R ssh -a pwd
```

- To list the system name of all nodes using `ssh` protocol, enter:

```
pdsh -R ssh -A uname -a
```

- To define `ssh` as default protocol, enter:

```
export PDSH_RCMD_TYPE=ssh;
```

- To display the date on all nodes, enter:

```
pdsh -A date
ns1: Mon Dec 13 13:44:48 CET 2004
ns0: Mon Dec 13 13:44:47 CET 2004
ns2: Mon Dec 13 13:44:47 CET 2004
ns3: Mon Dec 13 13:44:46 CET 2004
```

- To display the date on all nodes except on node `ns0`, enter:

```
pdsh -A -x ns0 date
ns1: Mon Dec 13 13:44:48 CET 2004
ns2: Mon Dec 13 13:44:47 CET 2004
ns3: Mon Dec 13 13:44:46 CET 2004
```

- To display the date of the IO group nodes and to merge the output of the nodes whose result is identical, enter:



```

pdsh -g IO -x ns0 date | dshbak -c
-----
ns[2-3]
-----
  Mon Dec 13 14:10:41 CET 2004
-----
ns[1]
-----
  Mon Dec 13 14:10:42 CET 2004

```

## 2.6.2 Using pdcp

**pdcp** is a variant of the **rcp** command. Its syntax is not in the form `remote_user@node:path`. All source files are on the local node. The options which enable the nodes to be reached to be defined are similar to those of **pdsh**.

### Syntax:

**pdcp** **-Options** ... **<source [src2...]>** **<destination>**

### Examples:

```

pdcp -R ssh -w ns[1-5] /etc/hosts /etc/hosts
pdcp -R ssh -g Analyse /tmp/foo

```

In the first example one carries out a copy of `/etc/hosts` from the node where one **pdcp** executes to all the nodes specified using the `-w` option by copying across the same path for the command.

For a complete description of the options please refer to the **pdcp** man page.

## 2.6.3 Using dshbak

One of the problems linked to the execution of commands in parallel on a big cluster, is the exploitation of the results, especially if the command generates a long output. The results of a command executed with **pdsh** are displayed asynchronously and each line is stamped with the node name, as in the following example:

```

pdsh -w ns[0-2] pwd
      ns0 : /root
      ns2 : /root
      ns1 : /root

```

The **dshbak** utility formats the results of a **pdsh** command into a more user friendly form. Note that the results must be directed into a buffer file before being processed by **dshbak**.

### Syntax:

**dshbak** **[-c]** **<buffer\_file>**

**dshbak** can be used to create the following formatting:

- The node name, which was displayed on each line, is removed and replaced by a header containing this name.
- The generated list is sorted according to the node name if this name is suffixed by a number (ns0, ns1, ns2... ns500).
- If the `-c` option is present; **dshbak** will displays the identical results for several nodes once only. In this instance the header contains the node list.

### Examples:

In the following example, the result of the `pdsh` command is not formatted:

```
pdsh -R ssh w ns[0-2] rpm -qa | grep qsnetmpipwd
ns1 : qsnetmpi-1.24-31
      ns2 : qsnetmpi-1.24-31
      ns0 : qsnetmpi-1.24-31
```

In the following example, the `pdsh` output is re-directed to `res_rpm_qsnetmpi` file, then the **dshbak** command formats and displays the results:

```
pdsh -R ssh w ns[0-2] rpm -qa | grep qsnetmpipwd > /var /res_pdsh/res_rpm_qsnetmpi
dshbak -c res_rpm_qsnetmpi
-----
ns[0-2]
-----
qsnetmpi-1.24-31
```

## 2.7 Day to Day Maintenance Operations

A set of maintenance tools is provided with a Bull HPC cluster. These tools are mainly Open Source software applications that have been optimized, in terms of CPU consumption and data exchange overhead, to increase their effectiveness on Bull HPC clusters which may include hundred of nodes.

Function	Tool	Purpose
Administration	ConMan ipmitool	Console Management
	nsclusterstop / nsclusterstart	Stopping/Starting the cluster
	nsctrl	Managing hardware (power on, power off, reset, status, ping)
	syslog-ng	System log Management
	lptools (lputils, lpflash)	Emulex HBA (Host Bus Adapter) Management
Backup / Restore	mkCDrec	Backing-up and restoring data
Monitoring	ibstatus, ibstat	Monitoring <b>InfiniBand</b> networks
	lsiocfg	Getting information about storage devices
	pingcheck	Checking devices power status
Debugging	ibdoctor/ibtracert	<b>InfiniBand</b> network problem diagnosis/ Tracing communication paths in <b>InfiniBand</b> networks
	crash / proc	Runtime debugging
	netdump / diskdump	Dump facilities

Table 2-1. Maintenance Tools



- See the Bull BAS4 for Xeon *Maintenance Guide* for more information
- See Chapter 11 in this manual for details on HPC Toolkit, a set of profiling tools for clusters which include Add on functionality from the Bull HPCK software. See the **BAS4 for Xeon** *Installation and Configuration Guide* for more information.



## Chapter 3. Cluster Database Management

This chapter describes the architecture of the Cluster Database, the commands and the tools which enable the administrator to display and to change this Cluster Database.

The following topics are described:

- 3.1 Architecture of ClusterDB
- 3.2 ClusterDB Administrator
- 3.3 Using Commands
- 3.4 Managing the ClusterDB
- 3.5 ClusterDB Modeling

### 3.1 Architecture of ClusterDB

The Cluster database (**ClusterDB**) of the Bull HPC delivery contains the data that is required for the cluster management tools (**NS Master – HPC Edition, KSiS, pdsh, syslog-ng, ConMan, NsDoctor, etc.**). Compared with sequential configuration files, the advantages of using a database are flexibility and the distribution of the data to all the tools which ensure a better integration whilst at the same time not duplicating common data. Cluster database management uses the highly-scalable, SQL compliant, Open Source object-relational **PostgreSQL**.

The following figure shows an architecture of **ClusterDB** and its relationship to the cluster management tools.

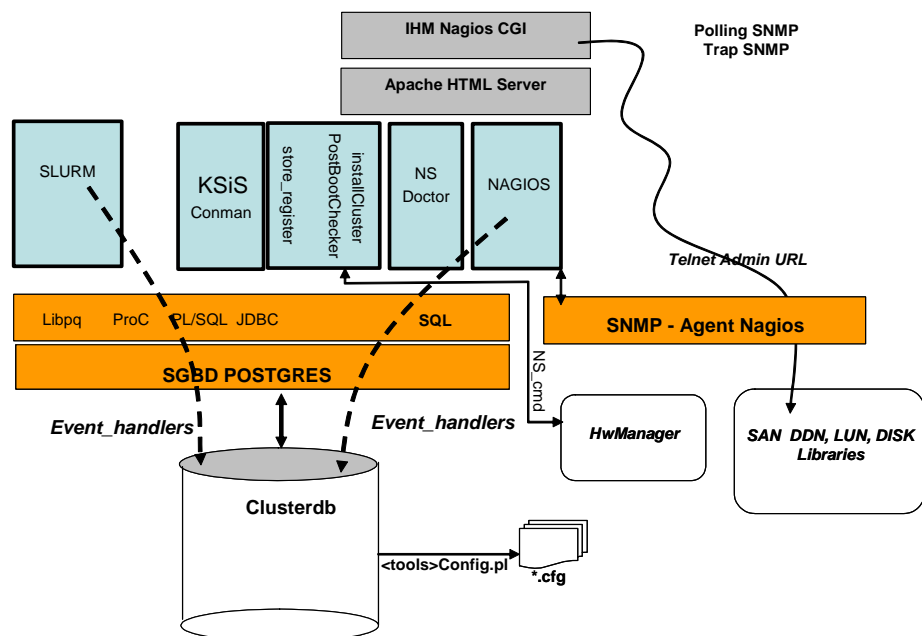


Figure 3-1. BAS4 for Xeon ClusterDB architecture

## 3.2 ClusterDB Administrator

The **ClusterDB** is installed on the Management Node. All operations on the **ClusterDB** must be performed from the Management Node.

The Database administrator is the **postgres** Linux user. This administrator is allowed to display and modify the **ClusterDB**, using the specific commands described in the next section. To manage the database (start, stop, save and restore), the administrator can use **PostgreSQL** tools (see 3.4 *Managing the ClusterDB*).

## 3.3 Using Commands

The administrators can consult or change the **ClusterDB** using the following commands:

- **changeOwnerProperties** changes the confidentiality parameters
- **dbmConfig** controls the consistency of the **ClusterDB** with the system. All database updates are marked to be a "candidate" for synchronization.
- **dbmCluster** operates on the whole cluster to get information, to check IP addresses and to check rack configuration.
- **dbmNode** displays information, or change attributes at the node level.
- **dbmHwManager** displays information, or change attributes at the **Hwmanager** level.
- **dbmGroup** manages the groups of nodes.
- **dbmEthernet** displays information, or change attributes for the Ethernet switches.
- **dbmIconnect** displays information, or change attributes for the interconnect switches.
- **dbmTalim** displays information, or change attributes for the remotely controlled power supply devices.
- **dbmSerial** displays information, or change attributes for the Portservers.
- **dbmFiberChannel** displays information about the Fiber Switches or changes the values of some attributes for a Fiber Switch or a subset of Fiber Switches.
- **dbmServices** displays information about the Services or changes the values of some attributes for a Service.
- **dbmDiskArray** displays information (for example **iproute**, **status**) and manages the disk array (**status**).

### 3.3.1 ChangeOwnerProperties

The cluster is handed over to the client with a name, a basename and IP address defined by Bull.

The IP address syntax used to identify equipment is of the form **A. V. U. H**.

**V** (the second byte) could be used for VLAN identification, **U** for Unit (Storage, Compute or Service) and **H** for Host (Host but also switch, disk subsystem or portserver).

The client may then want to change some of the attributes in keeping with their own security criteria.

These changes will in turn impact the **ClusterDB** Database, the network configuration for all the hosts, the configuration of storage bays and also the Lustre configuration (if installed).

Sometimes, the parameters which have been modified by the client may involve:

- Running **ECT** (Embedded Configuration Tool) for Interconnect switches
- Running **bmcConfig** for BMC cards
- Running **swtConfig** for Ethernet switches
- The network configuration of the nodes that will be done by **KSIS** at the time of the redeployment.
- Reconfiguring the DDN and FDA (Fibre Disk Array) subsystems to update them with the admin IP address and the gateway address.
- Manual operation of the **FDA**
- Running the **ddn\_init** command on each **DDN** and for the reboot.
- Restarting the configuration of the Cluster Suite on I/O nodes, so that each node is aware of its peer node, using the correct names and IP addresses.
- The Lustre system is impacted if the node **basenames** are changed resulting in the obliteration of the file system followed by the creation of a new file system with the new data.
- If there is a change in the node **basenames** and of the admin IP address, the KSIS images are deleted from the database.

Consequently, when using this command, it is necessary to follow the process described below in order to reinitialize the system.

### Syntax:

(This command is installed under `/usr/lib/clustmngt/clusterdb/install`)

```
changeOwnerProperties [--name <clustname>] [--basename <basename>]
                        [--adprivacy <bytes>]
                        [--icprivacy <interconnect privacy bytes (ic over ip)>]
                        [--bkprivacy <bytes>]
                        [--bkgw <ip gateway>] [--bkdom <backbone domain>]
                        [--bkof fset <backbone Unit offset>]
                        [--dbname <database name>] [--verbose]
```

### Options:

- dbname** Specifies the name of the database to which the command applies. Default value: **clusterdb**.  
**Note:** This option must be used only by qualified people for debugging purposes.
- name** Specifies the name of the cluster. By default it is the basemane.

<b>--basename</b>	Specifies the basename of the node. (The node name is constituted of basename + netid). It is also the virtual node name.
<b>--adprivacy</b>	Privacy bytes. According to the admin netmask, one, two or three bytes can be changed. For example, if the admin netmask is 255.255.0.0, then <b>adprivacy</b> option can specify two bytes in the form <b>A.V</b> .
<b>--icprivacy</b>	Privacy bytes. According to the interconnect netmask, one, two or three bytes can be changed. For example, if the interconnect netmask is 255.255.255.0, then <b>icprivacy</b> option can specify three bytes in the form <b>A.V.U</b> .
<b>--bkprivacy</b>	Privacy bytes. According to the backbone netmask, one, two or three bytes can be changed. For example, if the backbone netmask is 255.255.255.0, then <b>bkprivacy</b> option can specify three bytes in the form <b>A.V.U</b> .
<b>--bkgw</b>	Specifies the backbone gateway
<b>--bkdom</b>	Specifies the backbone domain
<b>--bkoffset</b>	Specifies the backbone translation offset. It permits to translate the D.E.U.H backbone ip to D.E.(U + bkoffset).H

#### Examples:

- To change the basename and the byte A of the admin IP address enter:

```
changeOwnerProperties --basename node --adprivacy 15
```

#### Process:

- Retrieve the current privacy bytes by running.

```
dbmEthernet show --nw admin
```

- Change parameters using the command **changeOwnerProperties**. If you changed network parameters then you have to reconfigure the IP addresses of all equipment as follows.
- Reconfigure admin interface of management node (**eth0** and **eth0:0** interfaces).
- Update the **dhcpcd** configuration and restart the service by running.

```
dbmConfig configure --service sysdhcpcd
```

- Restart **dbmConfig**.
- Reconfigure Ethernet switches by running.

```
swtConfig change_owner_properties --oldadprivacy <bytes>
```

- Reconfigure the IP addresses of the BMC cards.

```
/usr/lib/clustmngt/BMC/bmcConfig --oldadprivacy <bytes>
```



8. Manually configure on the FDA (if present).
9. Run **ddn\_init** on each DDN and reboot (if DDN storage is used).
10. Cluster Suite: run **storedepha** (if HA).
11. Syslog: The DDN logs are archived with the base name on the IP address, rename and the log files updated (if DDN is present)
12. For a Lustre configuration if the basename is changed:
  - a. Run **lustre\_util stop**
  - b. Run **lustre\_util remove**
  - c. Truncate the LUSTRE\_OST, LUSTRE\_MDT tables and use **storemodelctl generateost** and **storemodelctl generatemdt** to repopulate the tables with the new information.
  - d. Validate the recreated OSTs / MDTs: **lustre\_investigate check**
  - e. Verify the Lustre models and regenerate the configuration file: **lustre\_config**
  - f. Install new file systems: **lustre\_util install**

### 3.3.2 dbmConfig

The **dbmConfig** command is used to maintain the consistency between the data in the **ClusterDB** and the different services and system files. The **dbmConfig** command shows the synchronization state or synchronizes different cluster services (**syshosts**, **sysdhcpd**, **conman**, **portserver**, **pdsh**, **nagios**, **snmpptt**, **group**, **nsm**).

#### Syntax:

**dbmConfig show**            [--service <name>] [--dbname <database name>] [--impact]

**dbmConfig configure**    [-service <name> [-id <id> -timeout <timeout>] -restart -force  
-nodeps -impact] [-dbname <database name>]

**dbmConfig help**

#### Actions:

- |                  |                                                                                                                                                                                                                                                                                                                                                                                                     |
|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>show</b>      | Displays the synchronization state of all the services or of a list of specified services.                                                                                                                                                                                                                                                                                                          |
| <b>configure</b> | Runs the synchronization between the ClusterDB and all the services or a list of specified services. The configuration errors, if any, are listed on the console and in the <b>/var/log/synchro.log</b> file. It is necessary to check these messages.<br><b>Note:</b> The command reports an OK status to indicate that it has completed. This does not mean that no configuration error occurred. |
| <b>help</b>      | Displays the syntax of the <b>dbmConfig</b> command.                                                                                                                                                                                                                                                                                                                                                |

### Options:

<b>--dbname</b>	Specifies the name of the database to which the command applies. Default value: clusterdb. <b>Note:</b> This option must be used only by qualified people for debugging purposes.
<b>--force</b>	Reconfigures the service and restarts it.
<b>--id</b>	Reloads the configuration of the portserver identified by id. This option applies only to the portserver service (--service=portserver option).
<b>--impact</b>	Displays the configuration files and associated services impacted by the next dbmConfig configure command.
<b>--nodeps</b>	Forces the reconfiguration, despite the inter service dependencies.
<b>--restart</b>	Restarts the service instead of reloading it.
<b>--service</b>	Specifies the service from the following: syshosts, sysdhcpd, conman, portserver, pdsh, nagios, snmptt, group, nsm. For more information see Updated Configuration Files below.
<b>--timeout</b>	Specifies the timeout (in seconds) for restarting the portserver. This option applies only to the portserver service (--service=portserver option). Default value: 240.

### Updated Configuration Files:

According to the specified service, the **dbmConfig configure --service** command updates a configuration file, as described below:

Service	Action
<b>syshosts</b>	Updates the <b>/etc/hosts</b> file with the data available in the administration base
<b>sysdhcpd</b>	Updates the <b>/etc/dhcpd.conf</b> file with the data available in the administration base.
<b>conman</b>	Updates the <b>/etc/conman.conf</b> file with the data available in the administration base.
<b>portserver</b>	Updates the portserver configuration file ( <b>/tftpboot/ps16*ConfigTS16</b> or <b>/tftpboot/ps14*ConfigTS4</b> ), reloads the file on the appropriate portserver and reboots it.
<b>pdsh</b>	Updates the <b>/etc/genders</b> file with the data available in the administration base.
<b>nagios</b>	Updates several configuration files ( <b>/etc/nagios/*.cfg</b> ) with the data available in the administration base.

<b>snmptt</b>	Updates the <code>/etc/snmp/storage_hosts</code> file with the data available in the administration base.
<b>group</b>	Creates the predefined groups in the database. (No configuration file is updated.)
<b>nsm</b>	Updates the authentication file for the HW managers with the data available in the administration base.

If the administrator needs to modify these configuration files, for example, to add a machine that does not belong to the cluster, or to modify parameters, it is mandatory to use the template files created for this usage and to run the **dbmConfig** command again.

The templates files are identified by the **tpl** suffix. For example `/etc/hosts-tpl`, `/etc/dhcpd-tpl.conf`, `/etc/conman-tpl.conf`.

#### Examples:

- To configure the ConMan files, enter:

```
dbmConfig configure --service conman
```

- To list the synchronization state for Nagios, enter:

```
dbmConfig show --service nagios
```

### 3.3.3 dbmCluster

The **dbmCluster** command displays information about the whole cluster, or checks integrity and consistency of some elements of the ClusterDB.

#### Syntax:

```
dbmCluster show [- - dbname <database name>]
```

```
dbmCluster check ((- -ipaddr | - -rack) [- -verbose] ) | - -unitCell [- -dbname <database name>]
```

```
dbmCluster set - -profile <key1>=<value1> ... - -profile <keyN>=<valueN>
[- -dbname <database name>]
```

```
dbmCluster --h | --help
```

key must be in [actif\_ha actif\_crm actif\_vlan resource\_manager batch\_manager security parallel\_fs]

#### Actions:

**show** Displays the features of the cluster in terms of number of nodes and number of disks subsystems, as defined at the time of installation or update of the ClusterDB.

<b>check</b>	Checks integrity and consistency of some data of the ClusterDB: single IP addresses ( <b>--ipaddr</b> option) or consistency of rack equipments ( <b>--rack</b> option) or consistency of Unit Cell equipment ( <b>--unitCell</b> option).
<b>set</b>	Changes the value of some profile fields in the cluster table.
<b>help</b>	Displays the syntax of the <b>dbmCluster</b> command.

#### Options:

<b>--dbname</b>	Specifies the name of the database to which the command applies. Default: <b>clusterdb</b> . <b>Note:</b> this option must be used only by qualified people for debugging purposes.
<b>--ipaddr</b>	Checks that the IP addresses are distinct within the cluster.
<b>--rack</b>	Checks that the amount of equipment set for a rack in the database is not greater than the maximum. Also checks that there are not two sets of equipment on the same shelf.
<b>--unitCell</b>	Checks that the object Unit and Cell number are the same as the Ethernet switch connected to.
<b>--profile</b>	Used to set one key/value pair to be changed in table cluster.

#### Examples:

- To check that each IP address is distinct, enter:

```
dbmCluster check --ipaddr
```

### 3.3.4 dbmNode

The **dbmNode** command displays information about the nodes (type, status, installed image etc.) or changes the values of some attributes for a node or a set of nodes (unit).

#### Syntax:

<b>dbmNode show</b>	<b>[--sysimage [--install_status={installed   not_installed   in_installed}]]</b>
<b>dbmNode show</b>	<b>[--name &lt;node name&gt; --hwmanager   --cpu   --iproute   --serialroute]</b>
<b>dbmNode show</b>	<b>[--unit &lt;unit_num&gt; --hwmanager   --cpu] [--dbname &lt;database name&gt;]</b>
<b>dbmNode set</b>	<b>--name=&lt;node name&gt; --status={managed   not_managed}</b> <b>  --admin_macaddr &lt;macaddr&gt;   --backbone_macaddr &lt;macaddr&gt;</b>
<b>dbmNode set</b>	<b>--unit &lt;unit num&gt; --status={managed   not_managed}</b>
<b>dbmNode set</b>	<b>--nodelist=&lt;node list&gt; --status={managed   not_managed}</b>
<b>dbmNode set</b>	<b>( --name=&lt;node name&gt;   --unit &lt;unit num&gt; ) --cpu &lt;total cpu chipset&gt;</b>
<b>dbmNode set</b>	<b>( --name=&lt;node name&gt;   --unit &lt;unit num&gt; ) --hyperthreading={yes   no}</b>

**dbmNode set** ( --name=<node name> | --unit <unit num> ) --cpu <total cpu chipset>  
--hyperthreading={yes | no} [--dbname <database name>]

**dbmNode -h | --help**

### Actions:

**show** Displays type and status information for all the nodes or a set of nodes (--name option or --unit option). You can display the system images of nodes (using the --sysimage and --installed\_status options), and the CPU or PAP features (using the --cpu and --hwmanager options).

The **Type** parameter specifies the node functions in the form ACIMBNT.

A means ADMIN

C means COMPUTE

I means I/O

M means META

B means INFINIBAND

N means NFS

T means TAPE

For example, the type for a compute node is displayed as "-C-----".

**set** Changes the value of some features for the specified node (--name option) or for all the nodes of the specified unit (--unit option) or for a set of nodes (--nodelist option).

### Options:

**--help** Displays summary of options.

**--admin\_macaddr** Specifies the MAC address of the eth0 interface connected to the administration network.

**--backbone\_macaddr** Specifies the MAC address of the th1 interface connected to the backbone network.

**--cpu** Displays the CPU feature (model and number), or changes the number of CPUs.

**--install\_status** Displays only the nodes that have the specified install status (installed, in\_installed, not\_installed).

**--name** Specifies the node name to which the action applies.

**--iproute** Displays the ethernet path (the localization and status of all Ethernet switches) between the node and the admin node

**--serialroute** Displays the serial path over portserver (the localization and status of all portservers) between the node and the admin node

**--hwmanager** Displays the name of the hwmanager that drives the node.

- status** Changes the status (managed / not\_managed). The "not\_managed" status means that the node has not to be managed by the administration tools.
- sysimage** Displays the nodes and the status of their system image.
- unit** Specifies the unit to which the action applies.
- hyperthreading** Changes the hyperthreading mode.
- dbname** Specifies the name of the database on which the command is applied.  
Default: clusterdb.  
**Note:** this option must be used only by qualified people for debugging purposes.

**Examples:**

- To set the status of the node16 node to "up", enter:

```
dbmNode set --name node16 --status managed
```

- To change the MAC address of the node60 node, enter:

```
dbmNode set --name node60 --admin_macaddr 00:91:E9:15:4D
```

- Below are various examples using the **dbmNode show** command:

```
dbmNode show
```

Nodes names	Type	Status
node[0]	AC-M---	up
node[1-5,9-10]	-C-----	up
node[8]	-C--B--	not_managed
node[6,11]	-CI----	down
node[7]	-CI----	up
node[12-13]	--I-B--	down

```
dbmNode show --sysimage
```

Nodes names	Type	Sys Image	Status
node[4]	-C-----	BAS4-16K	up
node[3]	-C-----	BAS4-FAME	up
node[2,9]	-C-----	ONEDISK	up
node[8]	-C--B--	ONEDISK	up
node[1,5,10]	-C-----	NULL	up
node[6,11]	-CI----	NULL	down
node[7]	-CI----	NULL	up
node[12-13]	--I-B--	NULL	down

```
dbmNode show --sysimage --install_status installed
```

Nodes names	Type	Sys Image	Status
node[4]	-C-----	BAS4-16K	up
node[3]	-C-----	BAS4-FAME	up
node[2,9]	-C-----	ONEDISK	up
node[8]	-C--B--	ONEDISK	up

```
dbmNode show --name ns0 --cpu
```

Name	Cpu model	Cpu total	Cpu available	Hyper threading
ns0	UNDEF	8	0	0

### 3.3.5 dbmHwManager

The **dbmHwManager** command displays information or change status at the level of the HW Manager.

#### Syntax:

```
dbmHwManager show  [--name <hwmanager name> --node | --status | --iproute]
dbmHwManager show  [--unit <unit num> --status] [--dbname <database name>]
dbmHwManager set    --name <hwmanager name> --status ={managed | not_managed} | -
password
dbmHwManager set    --unit <unit_num> --status ={managed | not_managed}
                    [--dbname <database name>]
dbmHwManager        -h | --help
```

#### Actions:

**show** Displays model, type and status information for all the hwmanagers or a subset of hwmanager (--unit option).

**set** Changes the value of some features for the specified hwmanager (--name option) or for all the hwmanagers of the specified unit (--unit option).

#### Options:

**--help** Displays summary of options.

**--name** Specifies the hwmanager name to which the action applies.

**--iproute** Displays the Ethernet path (the localization and status of all Ethernet switches) between the hwmanager and the admin node

**--node** Displays the name of the nodes managed by the hwmanager.

**--status** Changes the status (managed/not\_managed). The "not\_managed" status means that the hwmanager has not to be managed by the administration tools.

**--password** Change the password for a given hwmanager.

**--unit** Specifies the unit to which the action applies.

**--dbname** Specifies the name of the database on which the command is applied. Default: clusterdb.  
**Note:** This option must be used only by qualified people for debugging purposes.

### Examples:

- To change the status of the PAP named `pap1` to "UP", enter:

```
dbmHwManager set --name pap1 --status managed
```

## 3.3.6 dbmGroup

The `dbmGroup` command lets the administrator of the ClusterDB show or modify (add, delete, create) the organization of the groups of nodes.



### Note:

The groups are using commands like `pdsh`, `KSIS`, to perform actions on a set of nodes.

### Syntax:

`dbmGroup`

`dbmGroup show` [--dbname <database name>]

`dbmGroup add` --name <group name> --nodelist <node list> [--comment <description>]  
[--dbname <database name>]

`dbmGroup del` --name <group name> | --all [--dbname <database name>]

`dbmGroup modify` --name <group name> (--addnodelist <node list> | --delnodelist <node list>)  
[--dbname <database name>]

`dbmGroup create` [--dbname <database name>]

`dbmGroup` -h | --help

### Actions:

**show** Displays the group of nodes.

**add** Adds a group to the existing ones.

**del** Deletes one group or all groups.

**modify** Adds or deletes a list of node in an existing group.

**create** Recreates the predefined groups (criterion groups), in the case they have been deleted.

### Options:

**--help** Displays summary of options.

**--name** Specifies the group name.

**--nodelist** List of the netid for the nodes of the group, in the form [x,y-w].

**--comment** Description of the group.



<code>--all</code>	Deletes all nodes.
<code>--addnodelist</code>	Adds a node list in an existing group.
<code>--delnodelist</code>	Deletes a node list in an existing group.
<code>--dbname</code>	Specifies the name of the database on which the command is applied. Default: clusterdb. <b>Note:</b> this option must be used only by qualified people for debugging purposes.

### Predefined Groups:

Once the cluster is configured, some predefined groups are automatically created, depending on the node types defined in the ClusterDB.

The `dbmGroup show` command displays the groups and a short explanation for each one.



#### Note:

A group can be mono-type, or multi-type for the nodes which combine several functions. Seven mono-type groups can be defined: ADMIN, COMPUTE (or COMP), IO, META, IBA, NFS, TAPE. See below examples of mono-type and multi\_type groups.

### Example of Predefined Groups:

In the following example four sorts of groups are defined:

- One Group of all the nodes except the nodes whose type is ADMIN. This group is named ALL.
- The group nodes per type. For instance:

ADMIN	Group of all the nodes whose type is ADMIN (mono-type).
ADMINCOMPMETA	Group of all the nodes whose type is ADMIN, compute or IO (multi-type).
COMPIBA	Group of all the nodes whose type is compute and Infiniband (multi-type).
COMPIO	Group of all the nodes whose type is compute or IO (multi-type).
COMPUTE	Group of all the nodes whose type is compute (mono-type).
IO	Group of all the nodes whose type is IO (mono-type).
IOIBA	Group of all the nodes whose type is IO and Infiniband (multi-type).
META	Group of all the nodes whose type is METADATA (mono-type).

- The groups of COMPUTE nodes for each memory size. For instance:

COMP48GB	Group of all the nodes whose type is <b>compute</b> and with 48GBs of memory (mono-type).
----------	-------------------------------------------------------------------------------------------

COMP128GB Group of all the nodes whose type is **compute** and with 128GB of memory (mono-type).

- The groups of nodes for each memory size. For instance:

NODES16GB Group of all the nodes with 16GBs of memory.

NODES48GB Group of all the nodes with 48GBs of memory.

NODES64GB Group of all the nodes with 64GBs of memory.

NODES128GB Group of all the nodes with 128GBs of memory.

### Examples:

- To display all the groups defined in the ClusterDB, enter:

```
dbmGroup show
```

Group Name	Description	Nodes Name
ADMIN	Nodes by type:ADMIN	node0
ALL	All nodes except node admin	node[4-5,8-10]
COMP	Nodes by type:COMP	node[4,8]
COMP128GB	COMPUTE node with 128GB	node8
COMP48GB	COMPUTE node with 48GB	node4
IO	Nodes by type:IO	node10
META	Nodes by type:META	node[5,9]
NODES128GB	Nodes by memory size:128GB	node8
NODES48GB	Nodes by memory size:48GB	node[4,10]
NODES64GB	Nodes by memory size:64GB	node[0,5,9]

- To add a new group, named GRAPH, which includes the nodes 1 and 4, 5, 6 (netid) into the database, enter:

```
dbmGroup add --name GRAPH --nodelist [1,4-6] --comment 'Graphic Group'
```

- To delete the GRAPH group from the database, enter:

```
dbmGroup del --name GRAPH
```

- To re-create the predefined groups if they have been deleted, enter:

```
dbmGroup create
```

```
=>
Create ALL [ OK ]
Create NODES4GB [ OK ]
Create NODES16GB [ OK ]
Create ADMIN [ OK ]
Create INFNFS [ OK ]
Create INF_TAPE [ OK ]
Create IOINF [ OK ]
Create METAINF [ OK ]
```

## 3.3.7 dbmEthernet

The **dbmEthernet** command displays or change attributes for the Ethernet switches.

### Syntax:

```
dbmEthernet show [--nw ={admin | backbone} ]
dbmEthernet show [--name <switch name> [--status | --macaddr | --iproute | --linkhost]]
dbmEthernet show [--unit <unit num> [--status]] [--dbname <database name>]
dbmEthernet set --name <switch name> [--status ={managed | not_managed}
| --macaddr <macaddr> | ([-password] [-enabled_password]) ]
dbmEthernet set --unit <unit_num> --status ={managed | not_managed}
[--dbname <database name>]
dbmEthernet -h | --help
```

### Actions:

**show** Displays name, network, ip address, Mac address and status information for all the switches or a subset of switches (--unit option).

**set** Changes the value of some features for a specified switch (--name option) or for all the switches of the specified unit (--unit option).

### Options:

**--help** Displays summary of options.

**--name** Specifies the switch name to which the action applies.

**--nw** Displays information about the given network type.

**--iproute** Displays the ethernet path (the localization and status of all ethernet switches) between the switch and the admin node.

**--macaddr** Changes the Macaddr of the Ethernet Switch.

**--status** Changes the status (managed / not\_managed). The "not\_managed" status means that the switch has not to be managed by the administration tools.

**--password** Change the password for a given switch.

**--enabled\_password** Change the enable password for a given switch.

**--unit** Specifies the unit to which the action applies.

**--dbname** Specifies the name of the database on which the command is applied. Default: clusterdb.  
**Note:** This option must be used only by qualified people for debugging purposes.

### Examples:

- To display the features of the administration network, enter:

```
dbmEthernet show --nw admin
```

- To change the mac address of the Ethernet switch named `eswu1c2` to the value `00:91:E9:15:4D`, enter:

```
dbmEthernet set --name eswu1c2 --admin_macaddr 00:91:E9:15:4D
```

### 3.3.8 dbmlconnect

The `dbmlconnect` command displays or change attributes for the interconnect switches.

#### Syntax:

```
dbmlconnect show [--nw={QsNet | InfiniBand | GbEthernet}]
```

```
dbmlconnect show [--name <switch name> [--status | --iproute] | --linkhost]
```

```
dbmlconnect show [--unit <unit num> [--status]] [--dbname <database name>]
```

```
dbmlconnect set --name <switch name> (--status={managed | not_managed} | ([-password] [-enabled_password]))
```

```
dbmlconnect set --unit <unit_num> --status={managed | not_managed} [--dbname <database name>]
```

```
dbmlconnect -h | --help
```

#### Actions:

**show** Displays name, network, admin and standby ip addresses, and status information for all the switches or a subset of switches (--unit option).

**set** Changes the value of some features for a specified switch (--name option) or for all the switches of the specified unit (--unit option).

#### Options:

**--help** Displays summary of options.

**--name** Specifies the switch name to which the action applies.

**--nw** Displays information about the given network type.

**--iproute** Displays the ethernet path (the localization and status of all ethernet switches) between the InterConnect switch and the admin node.

**--linkhost** Displays hosts plugged on a given interconnect switch.

**--status** Changes the status (managed / not\_managed). The "not\_managed" status means that the switch has not to be managed by the administration tools.

**--password** Change the password for a given switch.

**--enabled\_password** Change the enable password for a given switch.

- unit** Specifies the unit to which the action applies.
- dbname** Specifies the name of the database on which the command is applied. Default: clusterdb.  
**Note:** This option must be used only by qualified people for debugging purposes.

#### Examples:

- To display the features of the QsNet interconnect, enter:

```
dbmIconnect show --nw QsNet
```

- To change the status of the interconnect switch named QR0N01 to the value not\_managed, enter:

```
dbmIconnect set --name QR0N01 --status not_managed
```

### 3.3.9 dbmTalim

The **dbmTalim** command displays or change attributes for remotely controlled power supply devices.



#### Note:

**Talim** refers to remotely controlled power supply devices which are used to start and stop equipment.

#### Syntax:

```
dbmTalim show [--name <talim name> [--status | --macaddr | --iproute]]
```

```
dbmTalim show [--unit <unit num> [--status]] [--dbname <database name>]
```

```
dbmTalim set --name <talim name> --status ={managed | not_managed}  
| --macaddr <macaddr>
```

```
dbmTalim set --unit <unit_num> --status ={managed | not_managed}  
[--dbname <database name>]
```

```
dbmTalim -h | --help
```

#### Actions:

**show** Displays name, network, ip address, Mac address and status information for all the talim or a subset of talim (--unit option).

**set** Changes the value of some features for a specified talim (--name option) or for all the talim of the specified unit (--unit option).

#### Options:

**--help** Displays summary of options.

<b>--name</b>	Specifies the talim name to which the action applies.
<b>--iproute</b>	Displays the ethernet path (the localization and status of all ethernet switches) between the talim and the admin node
<b>--macaddr</b>	Displays the Macaddr or changes the Macaddr of the Talim.
<b>--status</b>	Displays the status or changes the status (managed / not_managed). The "not_managed" status means that the talim has not to be managed by the administration tools.
<b>--unit</b>	Specifies the unit to which the action applies.
<b>--dbname</b>	Specifies the name of the database on which the command is applied. Default: clusterdb. <b>Note:</b> This option must be used only by qualified people for debugging purposes.

#### Examples:

- To display the features of the talim named talim2, enter:

```
dbmTalim show --name talim2
```

- To change the mac address of the talim named talim2 to the value 00:91:E9:15:4D, enter:

```
dbmTalim set --name talim2 --macaddr 00:91:E9:15:4D
```

### 3.3.10 dbmSerial



#### Note:

The **dbmSerial** depends on the cluster's configuration and only applies to clusters which include a portserver.

The **dbmSerial** command displays or change attributes for the Portservers.

#### Syntax:

```
dbmSerial show [--nw ={node | pap | storage | mixed}]
```

```
dbmSerial show [--name <portserver name> [--status | --macaddr | --iproute | --linkhost]]
```

```
dbmSerial show [--unit <unit num> [--status]] [--dbname <database name>]
```

```
dbmSerial set --name <portserver name> --status ={managed | not_managed}
| --macaddr <macaddr> | -password
```

```
dbmSerial set --unit <unit_num> --status ={managed | not_managed} [--dbname <database
name>]
```

```
dbmSerial -h | --help
```

### Actions:

- show** Displays name, network, ip address, Mac address and status information for all the Portserver or a subset of portserver (--unit option).
- set** Changes the value of some features for a specified Portserver (--name option) or for all the Portserver of the specified unit (--unit option).

### Options:

- help** Displays summary of options.
- nw** Displays information about the given network type.
- name** Specifies the Portserver name to which the action applies.
- iproute** Displays the Ethernet path (the localization and status of all ethernet switches) between the Portserver and the admin node.
- status** Displays the status or changes the status (managed / not\_managed). The "not\_managed" status means that the Portserver has not to be managed by the administration tools.
- macaddr** Display/changes the Mac address of the Portserver.
- linkhost** Displays hosts plugged on a given portserver.
- password** Change the password for a given switch.
- unit** Specifies the unit to which the action applies.
- dbname** Specifies the name of the database on which the command is applied. Default: clusterdb.  
**Note:** This option must be used only by qualified people for debugging purposes.

### Examples:

- To display the features of all portservers, enter:

```
dbmSerial show
```

- To display the list of the hosts plugged on the portserver named ps16u1c0, enter:

```
dbmSerial show --name ps16u1c0 --linkhost
```

- To change the status of the portserver named ps16u1C0 , enter:

```
dbmSerial set --name ps16u1C0 --status managed
```

- To change the status of all portservers affiliated with unit 0, enter:

```
dbmSerial set --unit 0 --status not_managed
```

### 3.3.11 dbmFiberChannel

Displays the Database information about the Fiber Switches or changes the values of some attributes for a Fiber Switch or a subset of Fiber.

#### Syntax:

```
dbmFiberChannel show [--nw]
dbmFiberChannel show [--name <switch name> [--status | --iproute]]
dbmFiberChannel show [--unit <unit num> [--status]] [--dbname <database name>]
dbmFiberChannel set --name <switch name> --status ={managed | not_managed}
dbmFiberChannel set --unit <unit_num> --status ={managed | not_managed}
                    [--dbname <database name>]
dbmFiberChannel -h | --help
```

#### Actions:

**show** Displays name, network, admin ip address, and status information for all the switches or a subset of switches (--unit option).

**set** Changes the value of some features for a specified switch (--name option) or for all the switches of the specified unit (--unit option).

#### Options:

**--help** Displays summary of options.

**--name** Specifies the switch name to which the action applies.

**--nw** Displays information about all network type.

**--iproute** Displays the ethernet path (the localization and status of all ethernet switches) between the Fiber switch and the admin node.

**--status** Changes the status (managed / not\_managed). The "not\_managed" status means that the switch has not to be managed by the administration tools.

**--unit** Specifies the unit to which the action applies.

**--dbname** Specifies the name of the database on which the command is applied. Default: clusterdb.  
**Note:** This option must be used only by qualified people for debugging purposes.



### Examples:

- To change the FC switch named `fcswu0c1` to up, enter:

```
dbmFiberChannel set --name fcswu0c1 --status managed
```

- To show the hierarchy iproute of the FC switch through Ethernet switches, enter:

```
dbmFiberChannel show --name fcswu0c1 --iproute
```

- To show information about FC switch, enter:

```
dbmFiberChannel show
```

## 3.3.12 dbmServices

Displays the Database information about the Services or changes the values of some attributes for a Service.

### Syntax:

```
dbmServices show --objectlist
```

```
dbmServices show --object <object name> [--name <service name>]  
                [--dbname <database name>]
```

```
dbmServices set  --object <object name> --name <service name> (--enable | --disable)  
                [--dbname <database name>]
```

```
dbmServices     -h | --help
```

### Actions:

**show**            Displays the list of all the objects contained in Services table (**--objectlist** option).  
  
Or displays name, object type and if service is enabled or disabled (**--object --name** options).

**set**             Changes the value of the **actif** field (enable or disable) for a specified service (**--object --name** options).

### Options:

**--help**           Displays summary of options.

**--objectlist**     Displays the list of all the objects contained in Services table.

**--object**         Specifies the object type of service to which the action applies.

**--name**            Specifies the service name to which the action applies.

- enable** Specifies that the service must be activated.
- disable** Specifies that the service must be de-activated.
- dbname** Specifies the name of the database on which the command is applied. Default: clusterdb.  
**Note:** This option must be used only by qualified people for debugging purposes.

**Examples:**

- To print details on the service named "Ethernet interfaces" on object node, enter:

```
dbmServices show --object node --name "Ethernet interfaces"
```

- To change the service named "Ethernet interfaces" on object node to up, enter:

```
dbmServices set --object node --name "Ethernet interfaces" --enable
```

### 3.3.13 dbmDiskArray

**dbmDiskArray** displays information (for example **iproute**, **status**) and manages the disk array (status)

**Syntax:**

**dbmDiskArray show** [-name <diskarray name> -iproute | -serialroute]  
[-dbname <database name>]

**dbmDiskArray set** -name < diskarray name> -status={managed | not\_managed}  
[-dbname <database name>]

**dbmDiskArray** -h | -help

**Actions:**

- show** Displays the type and status information for all the disk arrays or for a specified one (-name option).
- set** Changes the value of some of the features for a specified disk array (-name option).

**Options:**

- help** Displays a summary of options.
- name** Specifies the disk array name to which the action applies.
- iproute** Displays the Ethernet path (including the location and status of all Ethernet switches) between the disk array and the Management Node.

**--serialroute** Displays the serial path which includes a portserver (the location and status of all portservers) between the disk array and the Management Node.



**Note:**

The **--serialroute** option depends on the cluster's configuration and only applies to clusters which include a portserver.

**--status** Changes the status (**managed/ not\_managed**). The **not\_managed** status means that the disk array will not be managed by the administration tools.

**--dbname** Specifies the name of the database to which the command is applied. Default = clusterdb.



**Note:**

This option must be used only by qualified people for debugging purposes.

**Examples:**

- To print details of the disk array named **da0** using Ethernet switches, enter:

```
dbmDiskArray show --name da0 -iproute
```

- To change the status of the disk array named **da0** to up, enter:

```
dbmDiskArray set --name da0 -status managed
```

## 3.4 Managing the ClusterDB

The administrator of the **ClusterDB** must guarantee and maintain the consistency of the data. To view and administrate the database, the ClusterDB administrator can use the following PostgreSQL tools:

- The **PostgreSQL commands**.  
The **psql** command enables the PostgreSQL editor to run. You can run it as follows:

```
psql -U clusterdb clusterdb
```

- The **phpPgAdmin Web interface**.  
You can start it with an URL similar to the following one (admin0 is the name of the Management Node):

```
http://admin0/phpPgAdmin/
```



### Important:

These tools, which let the administrator update the **ClusterDB**, must be used carefully since incorrect usage could break the consistency of the **ClusterDB**.

For more information, refer to the **PostgreSQL** documentation delivered with the product.

### 3.4.1 Saving and Restoring the Database

The database administrator is responsible for saving and restoring the ClusterDB.

The administrator will use the **pg\_dump** and **pg\_restore** PostgreSQL commands to save and restore the database.

#### 3.4.1.1 Saving the Database (pg\_dump)

The **pg\_dump** command has a lot of options. To display all the options, enter:

```
pg_dump --help
```



### Note:

The **pg\_dump** command can run while the system is running.

#### Saving the Metadata and the Data:

It is recommended that the following command is used:

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

### Saving the Data only:

It is recommended that the following command is used:

```
pg_dump -Fc -a -f /var/lib/pgsql/backups/clusterdbdata.dmp clusterdb
```

### Saving Data each Day

When the **clusterdb** rpm is installed, a **cron** is initialized to save the ClusterDB daily, at midnight. The data is saved in the **clusterdball[0-6].dmp** and **clusterdata[0-6].dmp** (0-6 is the number of the day) in the **/var/lib/pgsql/backups** directory. This **cron** runs the **make\_backup.sh** script, located in the directory **/usr/lib/clustmngt/clusterdb/install/**.

## 3.4.1.2 Restoring the Database (pg\_restore)

The **pg\_restore** command has a lot of options. To display all the options, enter:

```
pg_restore --help
```

### Restoring the whole ClusterDB:

Requirement: ClusterDB does not exist anymore.

To list the existing databases, use the **oid2name** command:

```
oid2name
```

If you need to remove an inconsistent **ClusterDB**, enter:

```
dropdb clusterdb
```

When you are sure that the **ClusterDB** does not exist anymore, enter the following command to restore the whole database:

```
pg_restore -Fc --disable-triggers -C -d template1  
/var/lib/pgsql/backups/clusterdball.dmp
```

### Restoring the ClusterDB Data:

Requirement: ClusterDB must exist and be empty.

To create an empty ClusterDB, run these commands:

```
/usr/lib/clustmngt/clusterdb/install/create_clusterdb.sh -nouser  
psql -U clusterdb clusterdb  
clusterdb=> truncate config_candidate;  
clusterdb=> truncate config_status;  
clusterdb=> \q
```

To restore the data, enter:

```
pg_restore -Fc --disable-triggers -d clusterdb /var/lib/pgsql/backups/clusterdbdata.dmp
```

## 3.4.2 Starting and Stopping PostgreSQL

Starting and stopping **postgreSQL** is performed using the **service** Linux command. **postgreSQL** is configured to be launched at levels 3, 4 and 5 for each reboot.



### Note:

Both **root** user and **postgres** user can start and stop PostgreSQL. However it is recommended to use always the **postgres** login.

To start **postgreSQL**, run the following script:

```
/sbin/service postgresql start
```

To stop **postgreSQL**, run the following script:

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

## 3.4.3 Viewing the PostgreSQL Alert Log

The **postgreSQL** log file is **/var/log/postgres/pgsql**. This is read to view any errors, which may exist.



### Note:

This file can increase in size very quickly. It is up to the database administrator to rotate this file when **postgreSQL** is stopped.

## 3.5 ClusterDB Modeling



### Important:

The ClusterDB diagrams and tables which follow are common to both **BAS4** and **BAS4 for Xeon** systems. Certain tables will only be exploited by the functionality of **BAS4**, for **BAS4 for Xeon** these tables will be empty.

### 3.5.1 Physical View of the Cluster Networks

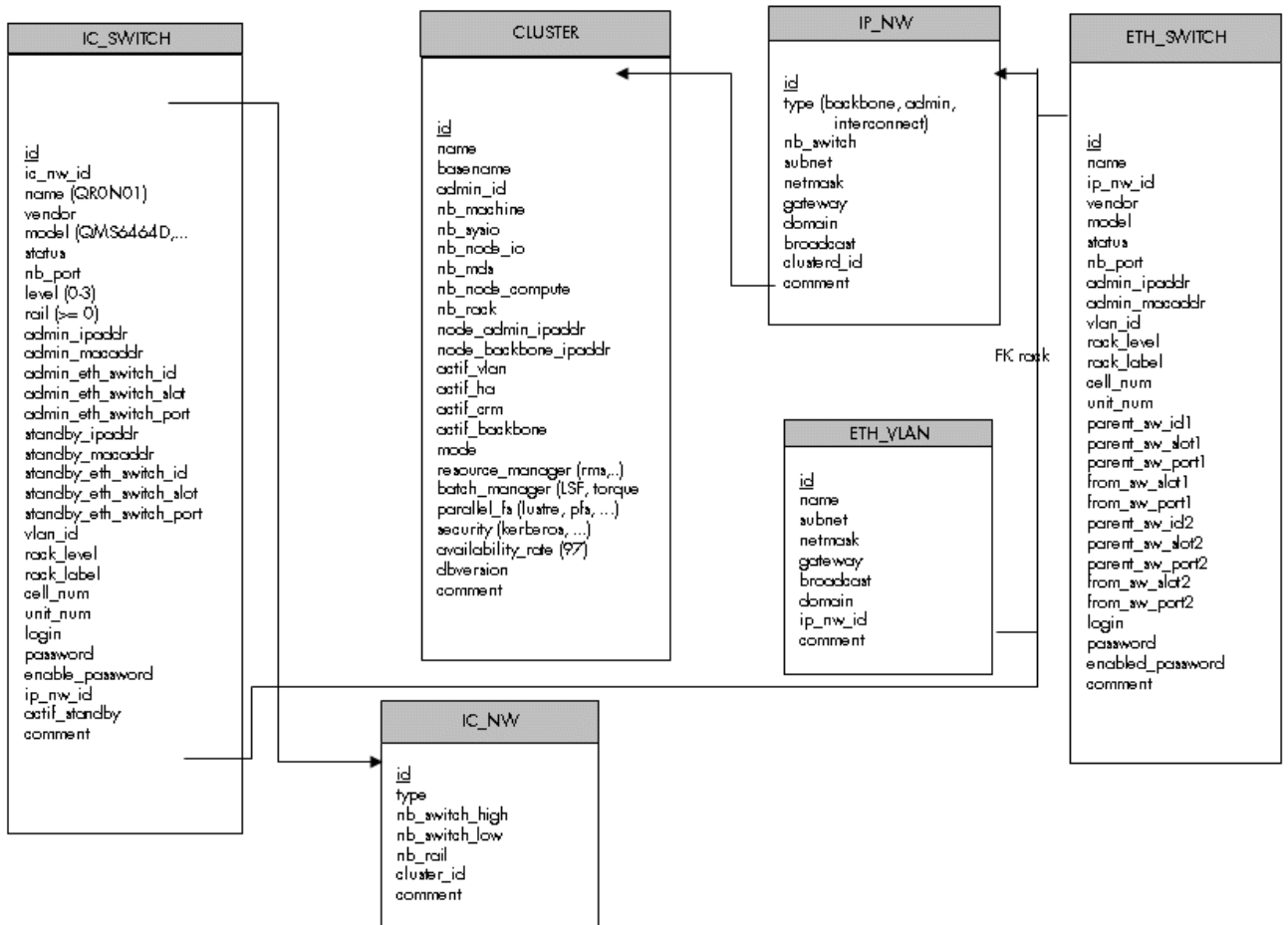


Figure 3-2. Cluster Network – diagram 1

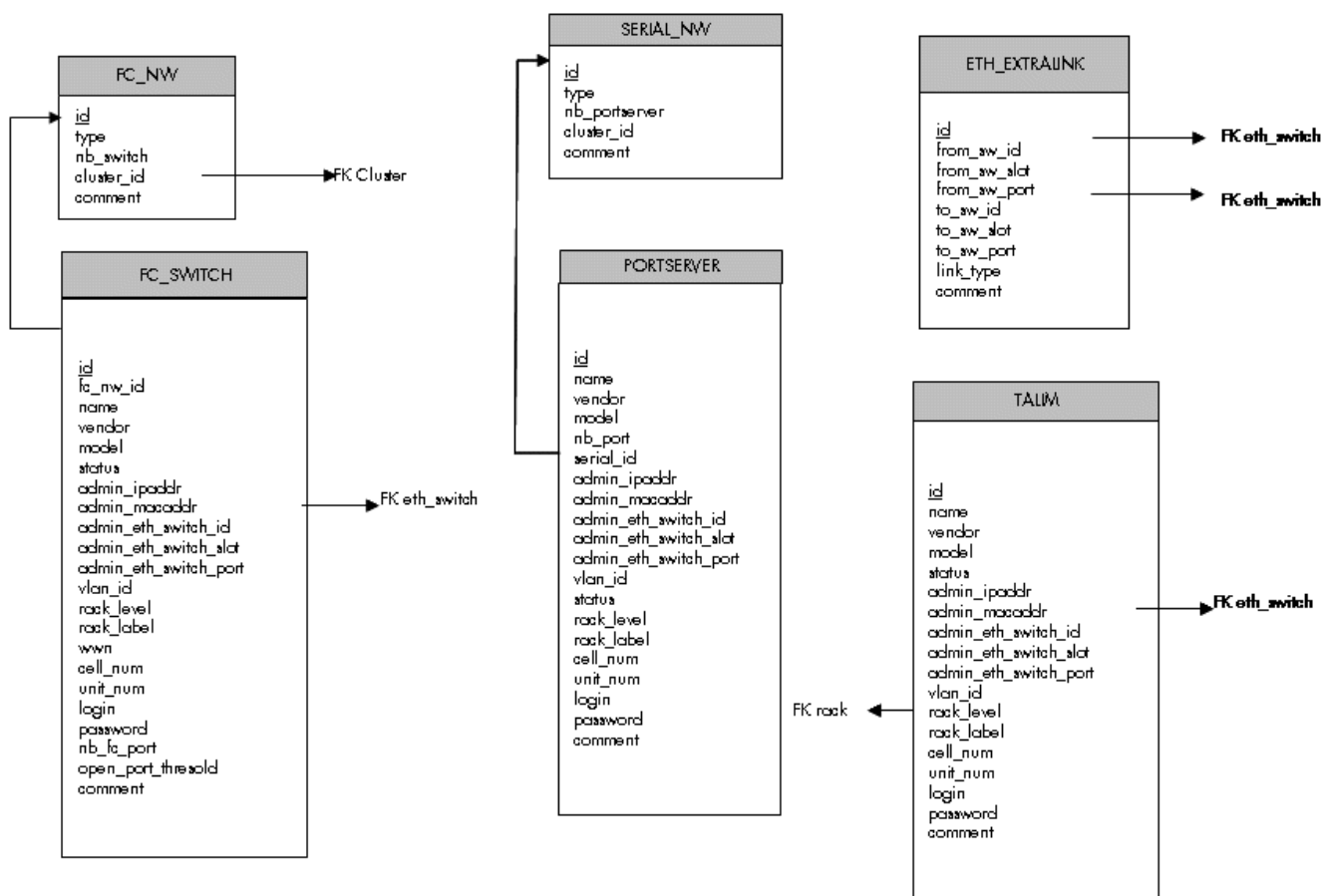


Figure 3-3. Cluster Network – diagram 2

### 3.5.1.1 CLUSTER Table

Column name	Description	Example	Fill in method
id	PK	540	preload - sequence
name	Name of the cluster	molecular	Preload & loadClusterdb
basename	Node basename	node	Preload & loadClusterdb
admin_id	FK table User		preload
nb_machine	Number of Nodes	601	preload – reconfigClusterdb
nb_sysio	Number of disk sub systems	56	preload – reconfigClusterdb
nb_node_io	Number of IO Nodes	54	preload – reconfigClusterdb
nb_mds	Number of MDS	2	preload – reconfigClusterdb
nb_node_compute	Number of Compute Nodes	544	preload – reconfigClusterdb
nb_rack	Number of rack	270	preload – reconfigClusterdb
node_admin_ipaddr	Virtual IP address of the Management node for the backbone network	10.1.0.65	preload
node_backbone_ipaddr	Virtual IP address of the Management node		preload
actif_vlan	Boolean on the VLAN configuration	true	preload
actif_ha	Boolean High Availability	true	Cluster Suite



Column name	Description	Example	Fill in method
actif_crm	CRM Boolean surveillance	true	preload
actif_backbone	Boolean, Use of a backbone	true	DV=true
mode	Mode 100%, 92% or 8%	100	preload – reconfigClusterdb
resource_manager	RMS or SLURM	rms	preload
batch_manager	LSF or TORQUE	torque	preload
parallel_fs	Lustre	lustre	prelaod
security	Kerberos	NULL	preload
availability_rate	Availability rate	20.3.3	preload
dbversion	Development model version for the database	16.2	Creation
comment	Free field		NULL

Table 3-1. Cluster Table

### 3.5.1.2 IP\_NW Table

Column name	Description	Example	Fill in method
id	PK	4	preload – Sequence
type	backbone, admin	backbone	preload
nb_switch	Number of switches	10	preload
subnet	Sub-network	10.0.0.0	preload
netmask	Sub-network mask	255.255.0.0	preload
gateway	IP address of the gateway	10.0.255.254	preload
domain	Name of the domain	frec.bull.fr	preload
broadcast	IP address of the broadcast	NULL	NULL
cluster_id	FK on the CLUSTER		preload
comment	Free field		NULL

Table 3-2. IP\_NW table

### 3.5.1.3 ETH\_SWITCH Table

Column name	Description	Example	Fill in method
id	PK		preload-Sequence
name	Name of the switch		preload
ip_nw_id	FK on IP_NW		preload
vendor	Vendor	CISCO	preload
model	Modele of the SW	CISCO6509	preload
status	Nagios host_status	up	DV = up - Nagios
nb_port	Total number of port		preload
admin_ipaddr	Admin IP address of the Ethernet switch		preload
admin_macaddr	Mac Address of the Switch		ACT

Column name	Description	Example	Fill in method
vlan_id	FK on ETH_VLAN		preload
rack_level	Superposition level in the rack		preload
rack_label	Name of the rack		preload
cell_num	Name of the cell		preload
unit_num	Number of the Unit		preload
parent_sw_id1	Ethernet switch 1st parent		preload
parent_sw_slot1	Arrival slot number of the 1 <sup>st</sup> parent switch	0	preload
parent_sw_port1	Connection port for the 1st switch	1	preload
from_sw_slot1	Starting slot number of the 1 <sup>st</sup> switch	0	preload
from_sw_port1	Starting port number of the 1 <sup>st</sup> switch	1	preload
parent_sw_id2	Ethernet switch 2 <sup>nd</sup> parent		preload
from_sw_slot2	Starting slot number of the 2 <sup>nd</sup> switch		preload
parent_sw_port2	Starting port number for the 2 <sup>nd</sup> switch	2	preload
from_sw_slot2	Starting slot number of the 2 <sup>nd</sup> switch		preload
from_sw_port2	Starting port number of the 2 <sup>nd</sup> switch		preload
login	Administration login		cmdExpl
password	Administration password		cmdExpl
enabled_password	Cisco enabled password		ECT

Table 3-3. ETH\_SWITCH Table

### 3.5.1.4 IC\_NW Table

Column name	Description	Example	Fill in method
Id	PK		preload - Sequence
type	QSNNet, Infiniband	QSNNet	preload
nb_switch_high	Number of high switches	12	preload - reconfigClusterdb
nb_switch_low	Number of low switches	33	preload - reconfigClusterdb
nb_rail	Number of rails	3	preload
cluster_id	FK on the CLUSTER		preload
comment	Free field		

Table 3-4. IC\_NW Table

### 3.5.1.5 IC\_SWITCH Table

Column name	Description	Example	Fill in method
Id	PK		preload - Sequence
ic_nw_id	FK on the IC_NW		preload
name	Name of the Switch Interconnect	QRON01	preload
vendor	Name of the Vendor	QUADRICS	preload
model	Model of the Switch	QMS6464D	preload

Column name	Description	Example	Fill in method
status	Nagios host_status	up	DV = up – Nagios
nb_port	Port number	64	preload
level	Level of the switch	1 – 2	preload
rail	Number of the rails	2	preload
admin_ipaddr	Administration IP address		preload
admin_macaddr	Mac Address of the switch	unused	NULL
admin_eth_switch_id	FK on ETH_SWITCH		preload
admin_eth_switch_slot	Arrival slot number on ETH_SW		preload
admin_eth_switch_port	Connection port on the ETH_SW	5	preload
standby_ipaddr	IP address of the standby controller		preload
standby_macaddr	Mac Address of the controller	unused	NULL
standby_eth_switch_id	FK on the ETH_SWITCH		preload
stanby_eth_switch_slot	Arrival slot number on ETH_SW		preload
standby_eth_switch_port	Connection port on the ETH_SW	6	preload
vlan_id	FK on the ETH_SWITCH		preload
rack_level	Level of superposition in the rack	G	preload
rack_label	Name of the rack	C0-A16	preload
cell_num	Number of the cell	1	preload
unit_num	Number of the Unit	0	preload
Login	Administration login	unused	preload or DV
password	Administration Password	unused	preload or DV
enable_password	Password enable		preload or DV
ip_nw_id	Foreign key on the IP_NW		preload
actif_standby	Configuration of a standby IP address		DV
comment	Free field		

Table 3-5. IC\_Switch Table

### 3.5.1.6 SERIAL\_NW Table

Column name	Description	Example	Fill in method
Id	PK	1	preload – sequence
rype	PAP Network, Node, Storage, Mixed	node	preload
nb_portserver	Number of PS	39	preload
cluster_id	FK on the CLUSTER		preload
comment	Free field		

Table 3-6. Serial\_NW Table

### 3.5.1.7 PORTSERVER Table



**Note:**

This table will not be filled for **BAS4 for Xeon** systems

Column name	Description	Example	Fill in method
id	Primary key		preload - sequence
name	Name of the portserver	ps16u1c0	preload
vendor	Name of vendor	DIGI	preload
model	Model of the PS	TS16	preload
nb_port	Total number of TTY/PS ports	16	preload
serial_id	FK on SERIAL_NW		preload
admin_ipaddr	Administration IP address		preload
admin_macaddr	Mac address of the PS		ACT
admin_eth_switch_id	FK on ETH_SWITCH		preload
admin_eth_switch_slot	Arrival slot number on ETH_SW		preload
admin_eth_switch_port	Connection port on the ETH_SW	10	preload
vlan_id	FK on the ETH_VLAN	40	preload
status	Nagios host_status	down	DV = up – Nagios
rack_level	Height of U in the rack		preload
rack_label	Name of the rack		preload
cell_num	Number of the cell		preload
unit_num	Number of the Unit		preload
login	Administration login		preload
password	Administration password		preload
comment	Free field		

Table 3-7. PORTSERVER Table

### 3.5.1.8 ETH\_VLAN Table

Column name	Description	Example	Fill in method
id	PK	1	preload - sequence
name	Name of the VLAN	pad	preload
subnet	Sub-network IP address	10.4.0.0	preload
netmask	Netmask of the sub-network	255.255.0.0	preload
gateway	IP address of the gateway	10.4.255.254	preload
broadcast	IP address of the broadcast	10.4.255.255	preload
domain	Name of the domain	unused	preload – NULL
ip_nw_id	FW on the IP_NW		preload
comment	Free field		

Table 3-8. ETH\_VLAN table

### 3.5.1.9 FC\_NW Table



**Note:**

This table only applies to systems which include a Storage Area Network (SAN).

Column name	Description	Example	Fill in method
id	PK	1	preload - sequence
type	Role of the network	SAN-META	preload
nb_switch	Number of switches	39	preload
cluster_id	FK on the CLUSTER		preload
comment	Free field		

Table 3-9. FC\_NW table

### 3.5.1.10 FC\_SWITCH Table



**Note:**

This table only applies to systems which include a Storage Area Network (SAN).

Column name	Description	Example	Fill in method
id	PK		preload-Sequence
name	Name of the switch		preload
fc_nw_id	FK on the FC_NW		preload
vendor	Name of the vendor	BROCADE	preload
model	SW model	Silkworm 200 <sup>E</sup>	preload
status	Nagios host_status	up	DV = up - Nagios
admin_ipaddr	IP admin address on the fibre switch channel		preload
admin_macaddr	Mac Address of the Switch		ACT
admin_eth_switch_id	FK on the ETH_SWITCH		preload
admin_eth_switch_slot	Arrival slot number on ETH SW		preload
admin_eth_switch_port	Connection on the ETH SW	3	preload
vlan_id	FK on the ETH_VLAN		preload
rack_level	Superposition level in the rack		preload
rack_label	Name of the rack		preload
cell_num	Number of the cell		preload
unit_num	Number of the unit		preload
login	Administration login		preload

Column name	Description	Example	Fill in method
password	Administration Password		preload
nb_fc_port	Number of fibre channel ports		preload
open_port_threshold			preload
comment	Free field		

Table 3-10. FC\_SWITCH table

### 3.5.1.11 TALIM Table

Column name	Description	Example	Fill-in method
id	PK		preload-Sequence
name	Name of the power switch		preload
vendor	Vendor name	Blackbox	preload
model	Model of the power switch		preload
status	Nagios host_status	up	DV = up - Nagios
admin_ipaddr	Admin IP address of the power switch		preload
admin_macaddr	Mac Address of the power switch		ACT
admin_eth_switch_id	FK on the ETH_SWITCH		preload
admin_eth_switch_slot	Arrival slot number on ETH SW		preload
admin_eth_switch_port	Connection port on the ETH SW	3	preload
vlan_id	FK on the ETH_VLAN		preload
rack_level	Superposition level in the rack		preload
rack_label	Name of the rack		preload
cell_num	Cell number		preload
unit_num	Unit number		preload
login	Administration login		preload
password	Administration password		preload
comment	Free field		

Table 3-11. TALIM table

### 3.5.1.12 ETH\_EXTRALINK Table

This table is not active in this version.

### 3.5.2 Physical View of the Storage

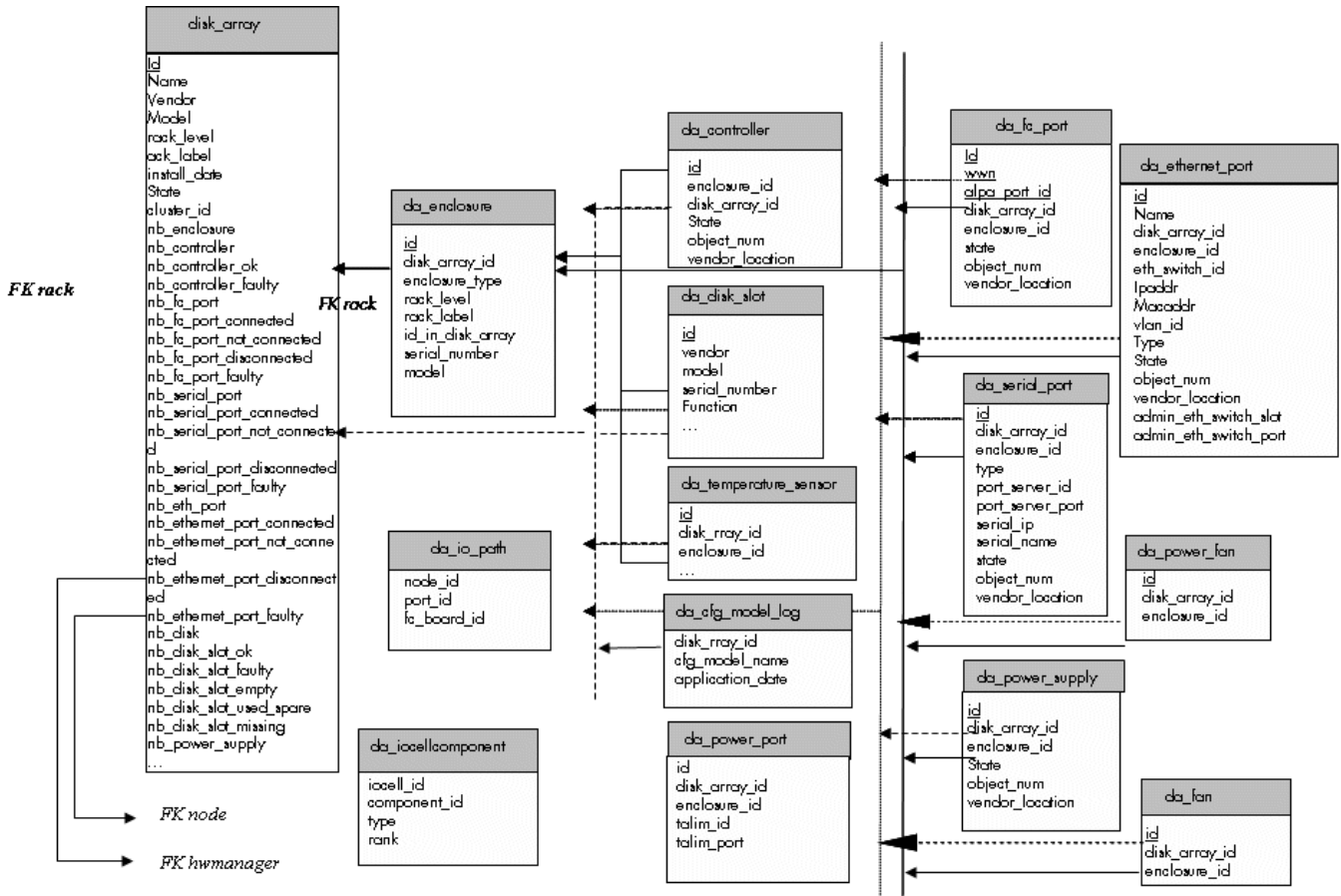


Figure 3-4. Storage physical view

#### 3.5.2.1 disk\_array Table

Field name	Field information	Fill in method
id	Unique identifier for the array in the database	preload - sequence
name	Name of the array (used for host in nagios)	preload
vendor	Vendor name: DDN, NEC, etc.	preload
model	Model name : S2A8500, FDA2300 ...	preload
rack_level	Location in the rack	preload
rack_label	Label of the rack containing the disk array controller	preload
install_date	Date of bay installation	preload – current time
state	UNKNOWN, OK, WARNING, FAULTY, OFF_LINE, OUT_OF_CLUSTER	Preload: OUT_OF_CLUSTER Dynamic - NSM
cluster_id	Id of the cluster parent	preload
nb_enclosure	Number of disk enclosure	Dynamic (DV=0) - NSM

Field name	Field information	Fill in method
nb_controller	Number of controller	Dynamic (DV=0) - NSM
nb_controller_ok	Number of controller in OK state	Dynamic (DV=0) - NSM
nb_controller_faulty	Number of controller in FAULTY state	Dynamic (DV=0) - NSM
nb_fc_port	Number of FC ports	Dynamic (DV=0) - NSM
nb_fc_port_connected	Number of FC ports in CONNECTED state	Dynamic (DV=0) - NSM
nb_fc_port_not_connected	Number of FC ports in NOT_CONNECTED state	Dynamic (DV=0) - NSM
nb_fc_port_disconnected	Number of FC ports in DISCONNECTED state	Dynamic (DV=0) - NSM
nb_fc_port_faulty	Number of FC ports in FAULTY state	Dynamic (DV=0) - NSM
nb_serial_port	Number of serial ports	Dynamic (DV=0) - NSM
nb_serial_port_connected	Number of serial ports in CONNECTED state	Dynamic (DV=0) - NSM
nb_serial_port_not_connected	Number of serial ports in NOT_CONNECTED state	Dynamic (DV=0) - NSM
nb_serial_port_disconnected	Number of serial ports in DISCONNECTED state	Dynamic (DV=0) - NSM
nb_serial_port_faulty	Number of serial ports in FAULTY state	Dynamic (DV=0) - NSM
nb_eth_port	Number of Ethernet ports	Dynamic (DV=0) - NSM
nb_ethernet_port_connected	Number of ethernet ports in CONNECTED state	Dynamic (DV=0) - NSM
nb_ethernet_port_not_connected	Number of ethernet ports in NOT_CONNECTED state	Dynamic (DV=0) - NSM
nb_ethernet_port_disconnected	Number of ethernet ports in DISCONNECTED state	Dynamic (DV=0) - NSM
nb_ethernet_port_faulty	Number of ethernet ports in FAULTY state	Dynamic (DV=0) - NSM
nb_disk	Number of disks	Dynamic (DV=0) - NSM
nb_disk_slot_ok	Number of disks in OK state	Dynamic (DV=0) - NSM
nb_disk_slot_faulty	Number of disks in FAULTY state	Dynamic (DV=0) - NSM
nb_disk_slot_empty	Number of disks in EMPTY state	Dynamic (DV=0) - NSM
nb_disk_slot_used_spare	Number of disks slots in USED_SPARE state	Dynamic (DV=0) - NSM
nb_disk_slot_missing	Number of disks in MISSING state	Dynamic (DV=0) - NSM
nb_power_supply	Number of power supplies	Dynamic (DV=0) - NSM
nb_power_supply_ok	Number of power supplies in OK state	Dynamic (DV=0) - NSM
nb_power_supply_faulty	Number of power supplies in FAULTY state	Dynamic (DV=0) - NSM
nb_nb_fan	Number of fans	Dynamic (DV=0) - NSM
nb_fan_ok	Number of fans in OK state	Dynamic (DV=0) - NSM
nb_fan_faulty	Number of fans in FAULTY state	Dynamic (DV=0) - NSM
nb_nb_power_fan	Number of power_fan	Dynamic (DV=0) - NSM
nb_power_fan_ok	Number of power_fan in OK state	Dynamic (DV=0) - NSM
nb_power_fan_faulty	Number of power_fan in FAULTY state	Dynamic (DV=0) - NSM
nb_nb_temperature_sensor	Number of temperature sensors	Dynamic (DV=0) - NSM



Field name	Field information	Fill in method
nb_temperature_sensor_ok	Number of temperature sensors in OK state	Dynamic (DV=0) - NSM
nb_temperature_sensor_warning	Number of temperature sensors in WARNING state	Dynamic (DV=0) - NSM
nb_temperature_sensor_faulty	Number of temperature sensors in FAULTY state	Dynamic (DV=0) - NSM
nb_lun	Number of lun	Dynamic (DV=0) - NSM
nb_spare	Number of spare disk	Dynamic (DV=0) - NSM
serial_number	Serial number of the array	Dynamic - storegister
type	Type of the array: OSS, MDS, ADMIN. Coded like UNIX rights (OMA, or – instead of the letter when the role does not apply)	preload
cfg_model	Name of the last applied model	Automatic - storemodelctl
cfg_model_application_date	Date of the last model application	Automatic - storemodelctl
mgmt_station_id	FK on HWMANAGER	preload
mgmt_node_id	FK on NODE	preload
status	Nagios status	Dynamic – NSM (DV="up")
unit_num	Unit Number	preload
comment	Free field	

Table 3-12. Storage – disk\_array table

### 3.5.2.2 da\_enclosure Table

Field name	Field information	Fill in method
id	Unique identifier for the disk enclosure in the database	preload –sequence
disk_array_id	Id of the parent array for this enclosure	preload
enclosure_type	Type of the disk enclosure	preload
rack_level	Level in the rack	preload
rack_label	Label of the rack containing the enclosure	preload
id_in_disk_array	Id of the enclosure in the array	preload
serial_number	Serial number of the enclosure	automatic – storeregister
model	Model of the disk enclosure	preload

Table 3-13. Storage – da\_enclosure table

### 3.5.2.3 da\_disk\_slot Table

Field name	Field information	Fill in method
id	Unique identifier for the disk_slot in the database	Automatic - sequence
vendor	Vendor name of disk	Automatic - storeregister
model	Model of disk	Automatic - storeregister

Field name	Field information	Fill in method
serial_number	Serial number of disk	Automatic – storeregister
function	Function of disk: EMPTY, DATA, SPARE (DATA_VAULT, DATA_FLARE, SPARE_VAULT, SPARE_FLARE, etc.)	Automatic – storeregister
capacity	Disk capacity in MBs	Automatic – storeregister
enclosure_id	Id of the parent enclosure	Automatic - storeregister
disk_array_id	Id of the parent array for this disk_slot	Automatic - storeregister
state	State of the disk slot : EMPTY, OK, WARNING, FAULTY, MISSING, USED_SPARE	Dynamic – NSM
disk_enclosure_id	Disk number in the enclosure	Automatic - storeregister
vendor_location	Location of the component expressed in the vendor terms.	Automatic - storeregister

Table 3-14. Storage – da\_disk\_slot table

### 3.5.2.4 da\_controller Table

Field name	Field information	Fill in method
id	Unique identifier for the controller in the database	Automatic – sequence
disk_array_id	Id of the parent array for this controller	Automatic – storeregister
enclosure_id	Id of the parent enclosure	Automatic - storeregister
State	State of the controller : OK , FAULTY, WARNING, OFF_LINE	Automatic – NSM
object_num	Controller identifier in the enclosure	Automatic – storeregister
vendor_location	Location of the component expressed in the vendor terms.	Automatic – storeregister

Table 3-15. Storage – da\_controller table

### 3.5.2.5 da\_fc\_port Table

Field name	Field information	Fill in method
id	Unique identifier for the fc_port in the database	preload – sequence
wwn	World Wide Name of the host port.	Automatic – storeregister
alpa_port_id	Loop address of the port	Automatic – storeregister
disk_array_id	Id of the parent array for this fc_port	preload
enclosure_id	Id of the parent enclosure	preload
State	State of the host port : CONNECTED, NOT_CONNECTED, DISCONNECTED, FAULTY	Dynamic – NSM
object_num	fc_port identifier in the enclosure	preload
vendor_location	Location of the component expressed in the vendor terms.	Automatic – storeregister

Table 3-16. Storage – da\_fc\_port.table

### 3.5.2.6 da\_serial\_port Table

Field name	Field information	Fill in method
id	Unique identifier for the serial port in the database	preload – sequence
disk_array_id	Id of the parent array for this serial port	preload
enclosure_id	Id of the parent enclosure	preload
type	type of serial port	preload
port_server_id	Port_server linked to this serial connection	preload
port_server_port	Index of the port used on the portserver (start at 0)	preload
serial_ip	IP address used to access to this serial port	preload
serial_name	Name of the console for conman	preload
state	State of the serial port : CONNECTED, NOT CONNECTED, DISCONNECTED, FAULTY	Dynamic – NSM
object_num	Serial port identifier in the enclosure	Preload
vendor_location	Location of the component expressed in the vendor terms.	Automatic – storeregister

Table 3-17. Storage – da\_serial\_port table

### 3.5.2.7 da\_ethernet\_port Table

Field name	Field information	Fill in method
id	Unique identifier for the Ethernet port in the database	preload - sequence
name	Name attached to this IP address	preload
disk_array_id	Id of the parent array for this Ethernet port	preload
enclosure_id	Id of the parent enclosure for this Ethernet port	preload
eth_switch_id	Id of the parent Ethernet_switch or parent pap_node	preload
ipaddr	IP address of the Ethernet port	preload
macaddr	MAC address of the Ethernet port	Automatic – storeregister
vlan_id	Id of the VLAN containing this Ethernet port	preload
type	Type of the Ethernet port : PUBLIC, ADMIN	preload
state	State of the Ethernet port : CONNECTED, NOT CONNECTED, DISCONNECTED, FAULTY	Dynamic – NSM
object_num	Ethernet port identifier in the enclosure	preload – storeregister
vendor_location	Location of the component expressed in the vendor terms.	Automatic – storeregister
admin_eth_switch_slot	Arrival slot number on ETH SW	preload
admin_eth_switch_port	Connection port on the ETH SW	preload

Table 3-18. Storage – da\_ethernet\_port Table

### 3.5.2.8 da\_power\_supply Table

Field name	Field information	Fill in method
id	Unique identifier for the power supply in the database	Automatic – sequence
disk_array_id	Id of the parent array for this power supply	Automatic – storeregister
enclosure_id	Id of the parent enclosure for this power supply	Automatic – storeregister
state	State of the power supply : OK, FAULTY,MISSING, [WARNING]	Dynamic – NSM
object_num	Power supply identifier in the enclosure	Automatic – storeregister
vendor_location	Location of the component expressed in the vendor terms.	Automatic – storeregister

Table 3-19. Storage – da\_power\_supply table

### 3.5.2.9 da\_fan Table

Field name	Field information	Fill in method
Id	Unique identifier for the fan in the database	Automatic – sequence
disk_array_id	Id of the parent array for this fan	Automatic – storeregister
enclosure_id	Id of the parent controller for this power supply	Automatic – storeregister
state	State of the power supply: OK, FAULTY, MISSING, [WARNING]	Dynamic – NSM
object_num	Fan identifier in the enclosure	Automatic – storegister
vendor_location	Location of the component expressed in the vendor terms.	Automatic – storegister

Table 3-20. Storage – da\_fan table

### 3.5.2.10 da\_power\_fan Table

Field name	Field information	Fill in method
Id	Unique identifier for the power_fan in the database	Automatic - - sequence
disk_array_id	Id of the parent array for this power_fan	Automatic- storeregister
enclosure_id	Id of the parent enclosure for this power_fan	Automatic- storeregister
State	State of the power_fan: OK, FAULTY, MISSING, [WARNING]	dynamic – NSM
object_num	Power_fan identifier in the enclosure	Automatic- storeregister
vendor_location	Location of the component expressed in the vendor terms.	Automatic- storeregister

Table 3-21. Storage – da\_power\_fan table

### 3.5.2.11 da\_temperature\_sensor Table

Field name	Field information	Fill in method
id	Unique identifier for the temperature sensor in the database	Automatic – sequence
disk_array_id	Id of the parent array for this power supply (if controller_id and enclosure_id are NULL)	Automatic – storeregister
enclosure_id	Id of the parent enclosure for this power supply ( if controller_id and array_id are NULL)	Automatic – storeregister
sensor_name	Name of the temperature sensor	Automatic – storeregister
state	State of the temperature sensor : OK, WARNING, FAULTY	Dynamic – NSM
object_num	Temperature sensor identifier in the enclosure	Automatic – storeregister
vendor_location	Location of the component expressed in the vendor terms.	Automatic – storeregister

Table 3-22. Storage – da\_temperature\_sensor table

### 3.5.2.12 da\_io\_path Table

Field name	Field information	Fill in method
node_id	Id of the node which access to this FC port	preload
port_id	Id of da_fc_port used by the node	preload
fc_board_id	Id of the HBA board	preload

Table 3-23. da\_io\_path table

### 3.5.2.13 da\_iocell\_component Table

Field name	Field information	Fill in method
iocell_id	Id of the IO cell	Preload - sequence
component_id	Id of a node or of a disk array	Preload
Type	Type of the component ("disk_array" or "node")	Preload
Rank	Rank of the node in the IO cell, or rank of the disk array in the IO cell. Start at 0.	preload

Table 3-24. Storage – da\_iocell\_component table

### 3.5.2.14 da\_cfg\_model Table

Field name	Field information	Fill in method
disk_array_id	Id of a disk array	Dynamic - storemodelctl
cfg_model_name	Model of a model which has been applied to the disk array	Dynamic - storemodelctl
application date	Date where the model has been applied	Dynamic - storemodelctl

Table 3-25. Storage – da\_cfg\_model table

### 3.5.2.15 da\_power\_port Table

Field name	Field information	Fill in method
Id	Unique identifier for the power_port in the database	Preload sequence
disk_array_id	FK to disk array	preload
enclosure_id	FK to enclosure id	preload
talim_id	FK to T_ALIM	preload
talim_port	Plug to be powered on/off onT_ALIM	preload

Table 3-26. Storage – da\_power\_port table

### 3.5.3 Machine View

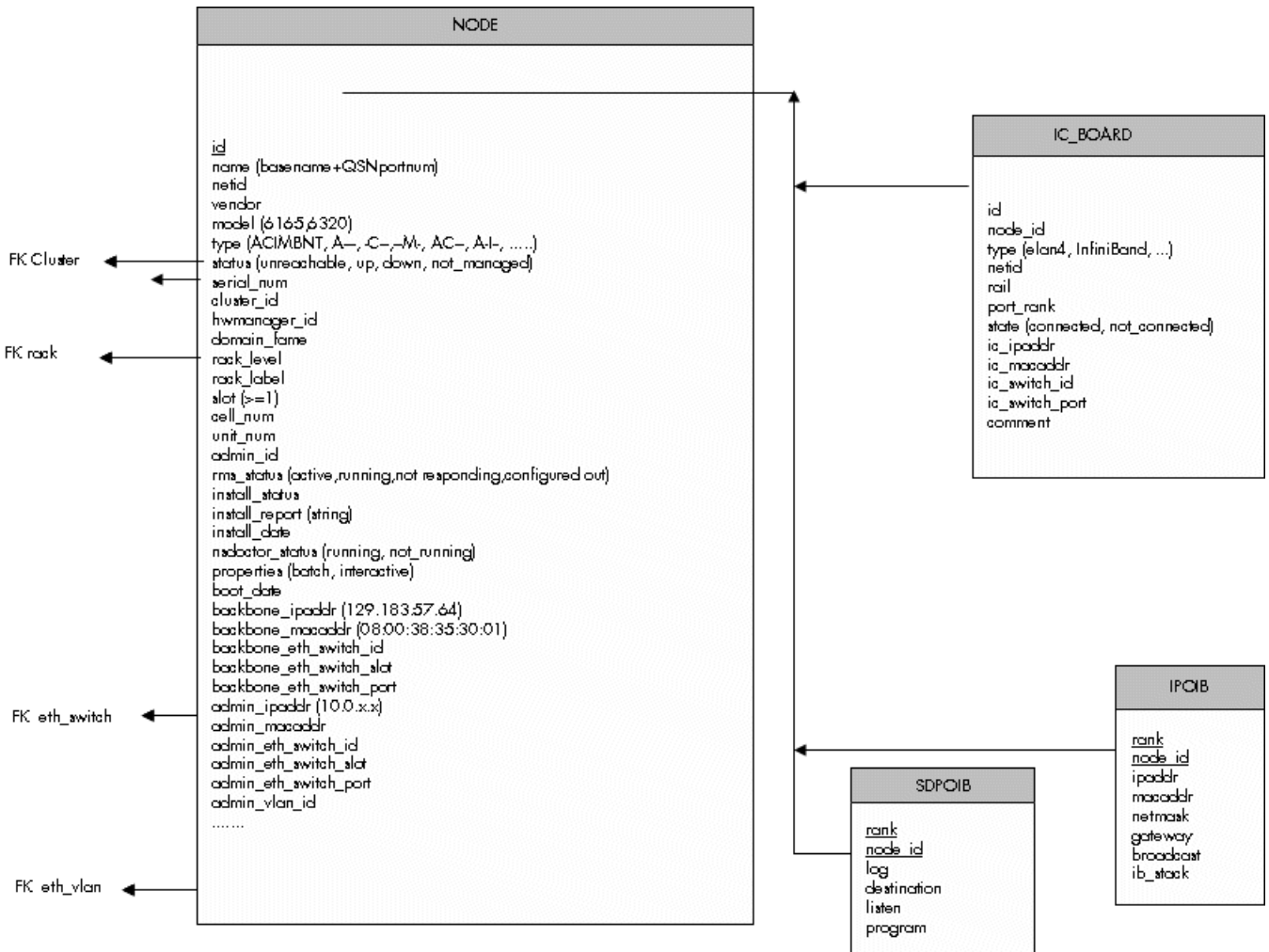


Figure 3-5. Cluster Database – Machine view 1

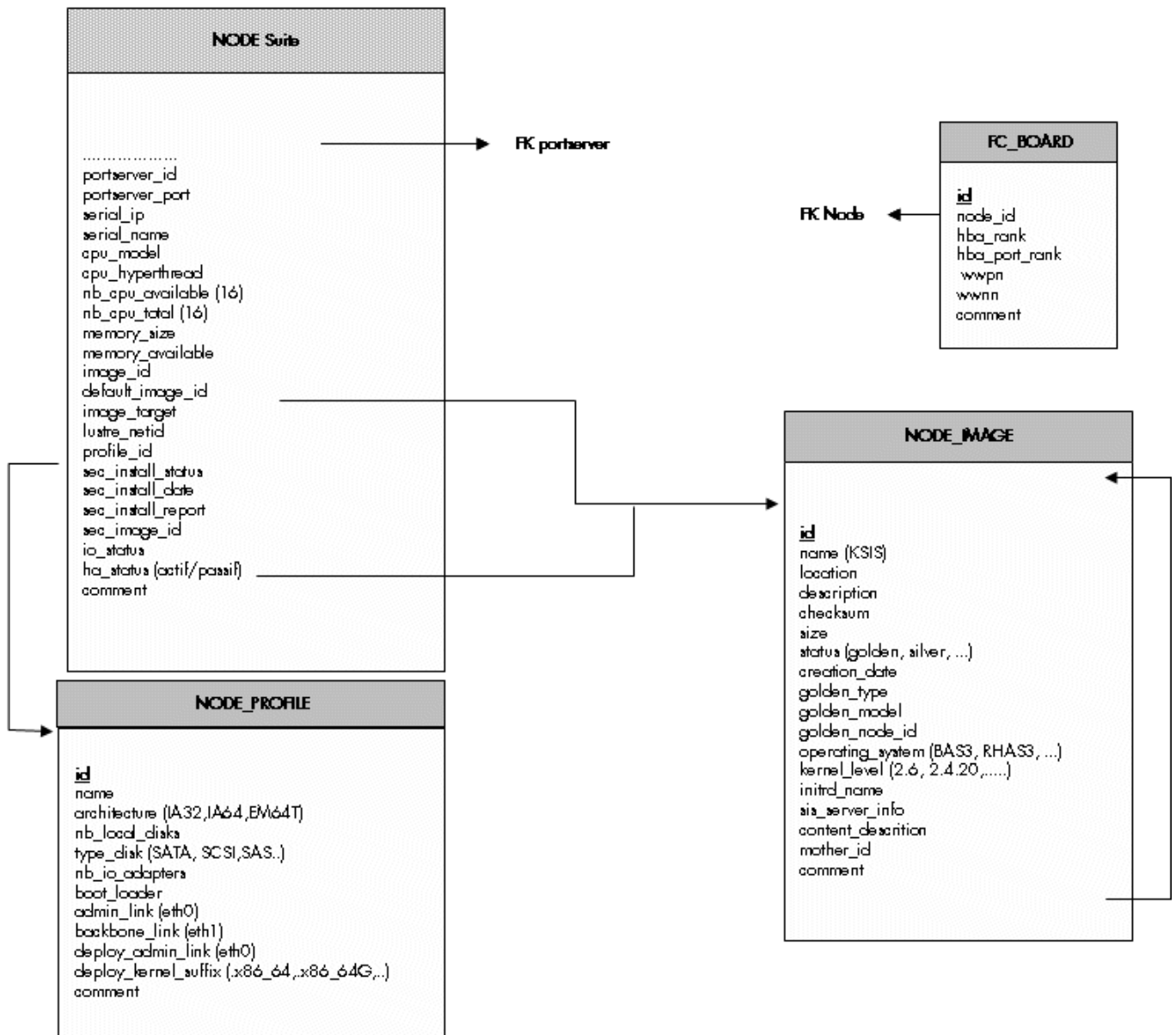


Figure 3-6. Cluster Database – Machine view 2

### 3.5.3.1 NODE Table

Column name	Description	Example	Fill-in method
id	primary key		preload– sequence
name	Node name	ns15	preload
netid	Node identifier number	1	preload
vendor	Name of vendor	Bull	preload
model	Node model	NS6165	preload
type	ACIMBNT node type, A-----, -C-----, - -I----, --M---	A-IM---	preload
status	Nagios host_status up, down, unreachable	down	DV = up – Nagios

serial_num	Serial number		ACT
cluster_id	FK on the CLUSTER		preload
hwmanager_id	FK on the HWMANAGER		preload
domain_fame	Machine name PAP side		ACT
rack_level	Height in the rack	A	preload
rack_label	Rack name	CU0-A5	preload
slot	Number of slot for the node [1-14]	1	preload
cell_num	Cell number	1	preload
unit_num	Unit ID	3	preload
admin_id	FK towards ADMIN		admin
rms_status	RMS status	configure out	event handler
install_status	KsiS Status	not_installed	KsiS
install_report	message	Host not installed	KsiS
install_date	System installation date	13/12/04 10 :30 :10	KsiS
NsDoctor_status	running or not-running	not-running	NsDoctor + DV
properties	Torque properties	Batch	Torque + DV
boot_date	Date of the last boot		PostBootChecker
backbone_ipaddr	Backbone IP Address	129.183.57.64	Preload
backbone_eth_switch_id	FK on the ETH_SWITCH		Preload
backbone_eth_switch_slot	Arrival slot number on ETH SW		Preload
backbone_macaddr	mac adresse	08:00:38:35:30:01	ACT
backbone_eth_switch_port	Connection port for BK_ETH_SW	2	Preload
admin_ipaddr	Admin IP address	10.1.0.1	Preload
admin_eth_switch_id	FK on the ETH_SWITCH	1	Preload
admin_eth_switch_slot	Arrival slot number on ETH SW		Preload
admin_eth_switch_port	Connection port for AD_ETH_SW	5	Preload
admin_vlan_id	FK for ETH_VLAN		Preload
admin_macaddr			ACT
portserver_id	FK on the PORTSERVER		Preload
portserver_port	Port number for the PS		Preload
serial_ip	Serial line access IP address	129.183.75.10	Preload
serial_name	Name of the serial number	ns15s	Preload
cpu_model	CPU model	Montecito	Preload
cpu_hypervisor	Boolean	True	PostBootChecker
nb_cpu_available	Number of CPUs available	15	PostBootChecker
nb_cpu_total	Number of CPUs	16	Preload
memory_size	Memory size	64	Preload
memory_available	Size of memory available	64	PostBootChecker
image_id	FK on the NODE_IMAGE		KsiS
default_image_id	FK on the default image		KsiS



image_target	For future use	NULL	NULL
lustre_netid	For future use	NULL	NULL
profile_id	FK on the NODE_PROFILE		Preload
sec_install_status	Secondary image KSiS status		KSiS
sec_install_date	Secondary Image installation date		KSiS
sec_install_report	Secondary Image message		KSiS
sec_image_id	FK of the NODE_IMAGE		KSiS
io_status	I/O status of the node		storage
ha_status (active/passive)	HA status of the node		Cluster Suite
comment	Free field	NULL	

Table 3-27. Machine view – node table

### 3.5.3.2 NODE\_IMAGE Table

Column name	Description	Example	Fill-in method
id	PK		Sequence
name	Name of the image	try	KSiS
location	localisation	/path/name	KSiS
description	description		KSiS
checksum	checksum	12352425	KSiS
size	Image size		KSiS
status	image status	= golden, silver	KSiS
creation_date	date	=JJ/DD/YY HH :MI :SS	Trigger
golden_type	IO, HPC, MDS, ADMIN		KSiS
golden_model	6165,6320		KSiS
golden_node_id	id of node serving as the golden node		KSiS
operating_system	Distribution type	BAS4V2	KSiS
kernel_level	Kernel level	6.2	KSiS
initrd_name	Initrd name		KSiS
sis_server_info	name/version		KSiS
content_description	description of the image content		KSiS
mother_id	Link to original image		KSiS
comment	Free field		

Table 3-28. Machine view – Node\_image table

### 3.5.3.3 NODE\_PROFILE Table

Column name	Description	Example	Fill in method
id	Primary Key	1	preload sequence

name	Name used to recognise the profile	MGMT	Preload
architecture	Type of architecture IA64, EM64T,etc.	IA64	preload
nb_local_disks	Number of internal disks	3	preload
type_disk	Type of disks (SATA, SCSI,SAS, etc)	SATA	preload
nb_io_adapters	Number of I/O cards	2	preload
boot_loader	elilo, grub	grub	KSIS
admin_link	admin interface (eth0)	eth0	DV
backbone_link	Interface backbone (eth1)	eth1	DV
deploy_admin_link	Deployment interface	eth0	DV
deploy_kernel_suffix	Kernel suffix (.x86_64, .x86_64G, etc.)	NULL	DV
comment	Free field		

Table 3-29. Machine view – Node\_Profile table

### 3.5.3.4 IC\_BOARD Table

This table describes Interconnect parameters (Quadrics, Infiniband or GBEthernet).

Column name	Description	Example	Fill in method
Id	Primary Key	1	preload sequence
node_id	FK on NODE	1	preload
type	type of card	elan4, infiniband	preload
netid	Node identifier number	3	preload
port_rank	Port number on the card	1	preload
ic_ipaddr	IP address of the IC Board	10.0.10.3	preload
ic_macaddr	Mac address	unused	
rail	Number of the rail	2	preload
ic_switch_id	FK on IC_SWITCH		preload
ic_switch_port	Number of the IC_SWITCH port	64	preload
comment	Free field		

Table 3-30. Machine view – IC\_BOARD table

### 3.5.3.5 IPOIB Table

This table describes Infiniband parameters for storage access.

Column name	Description	Example	Fill in method
rank	PK, Rank of the Infiniband adapter	0	updateIPOIB
node_id	PK, reference NODE	10	updateIPOIB

ipaddr	ip address on Infiniband	172.193.1.1	updateIPOIB
macaddr	Mac address		updateIPOIB
gateway	ip address of the gateway		updateIPOIB
broadcast	ip address of the broadcast		updateIPOIB
ib_stack	type of stack IP, SDP, BOTH	SDP	updateIPOIB

Table 3-31. Machine view – IPOIB Table

### 3.5.3.6 SDPOIB Table

Column name	Description	Example	Fill in method
Rank	PK, Rank of the Infiniband adapter	0	updateSDPoIB
node_id	PK, reference NODE	10	updateSDPoIB
Log	Log in sdplib.conf		updateSDPoIB
Destination	Desination in sdplib.conf		updateSDPoIB
Listen	Listen in sdplib.conf		updateSDPoIB
Program	Program in sdplib.conf		updateSDPoIB

Table 3-32. Machine view – SDPOIB table

### 3.5.3.7 FC\_BOARD table



**Note:**

This table only applies to systems which include a Storage Area Network (SAN).

Column name	Description	Example	Fill in method
Id	Primary key		storage
node_id	FK on the node	1	storage
hba_rank	Rank of the adapter		storage
hba_port_rank	Rank of the port		storage
wwpn	World Wide Port Name		storage
wwnn	World Wide Node Name		storage
comment	Free field		

Table 3-33. Machine view – FC\_BOARD table

### 3.5.4 HWMANAGER View

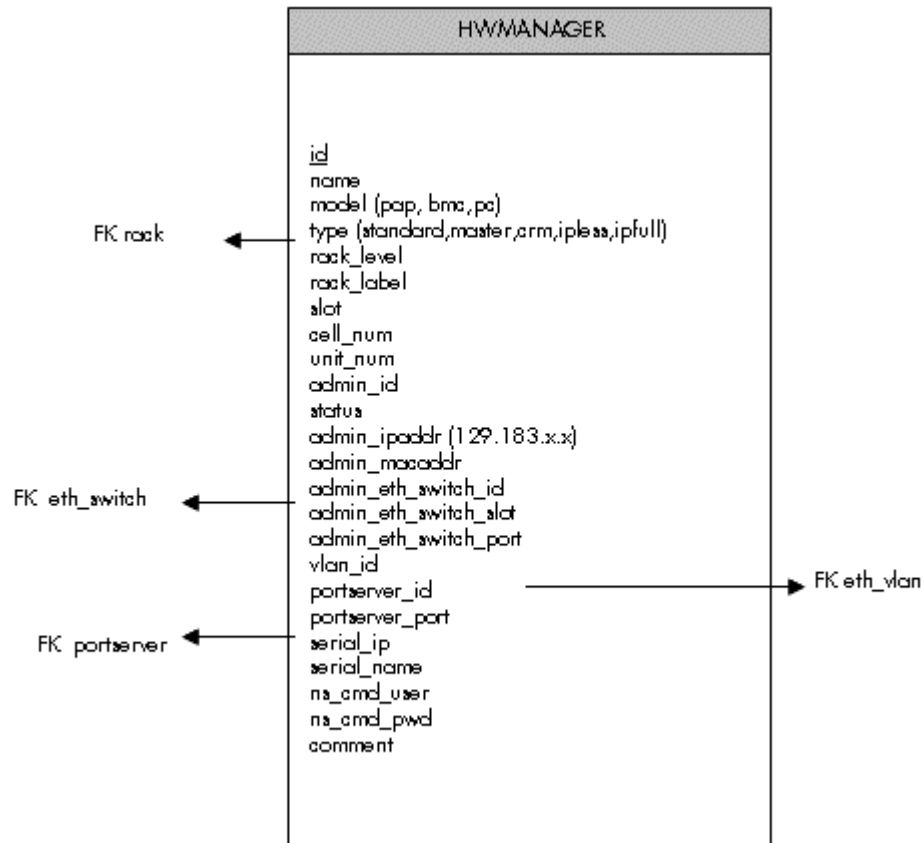


Figure 3-7. HWManager view

#### 3.5.4.1 HWMANAGER Table

Column name	Description	Example	Fill in method
Id	Primary key		preload - Sequence
name	HWMANAGER IP name	pap1c2	preload
model	PAP or BAC	pap	preload
type	standard, master, crm, ipless, ipfull	standard	preload
rack_level	Height in the rack	E	preload
rack_label	Name of the rack	ISO0-H45	preload
cell_num	Number of the cell	3	preload
unit_num	Number of the unit	1	preload
admin_id	ADMIN id		admin
status	Nagios status	unreachable	DV=up – Nagios
admin_ipaddr	Admin IP address		preload
admin_macaddr	Mac address		ACT
admin_eth_switch_id	ETH_SWITCH id		preload
admin_eth_switch_slot	Arrival slot number on ETH SW		preload

Column name	Description	Example	Fill in method
admin_eth_switch_port	ETH_SWITCH connection port	2	preload
vlan_id	ETH_VLAN id		preload
portserver_id	PORTSERVER id		preload
portserver_port	Portserver port number		preload
serial_ip	Serial line access IP address	129.183.75.10	preload
serial_name	Serial line name	ns15s	preload
ns_cmd_user	User NC Commande	nsc	preload
ns_cmd_pwd	password	\$nsc	preload
comment	Free field		

Table 3-34. HWManager Table

## 3.5.5 Complementary Tables

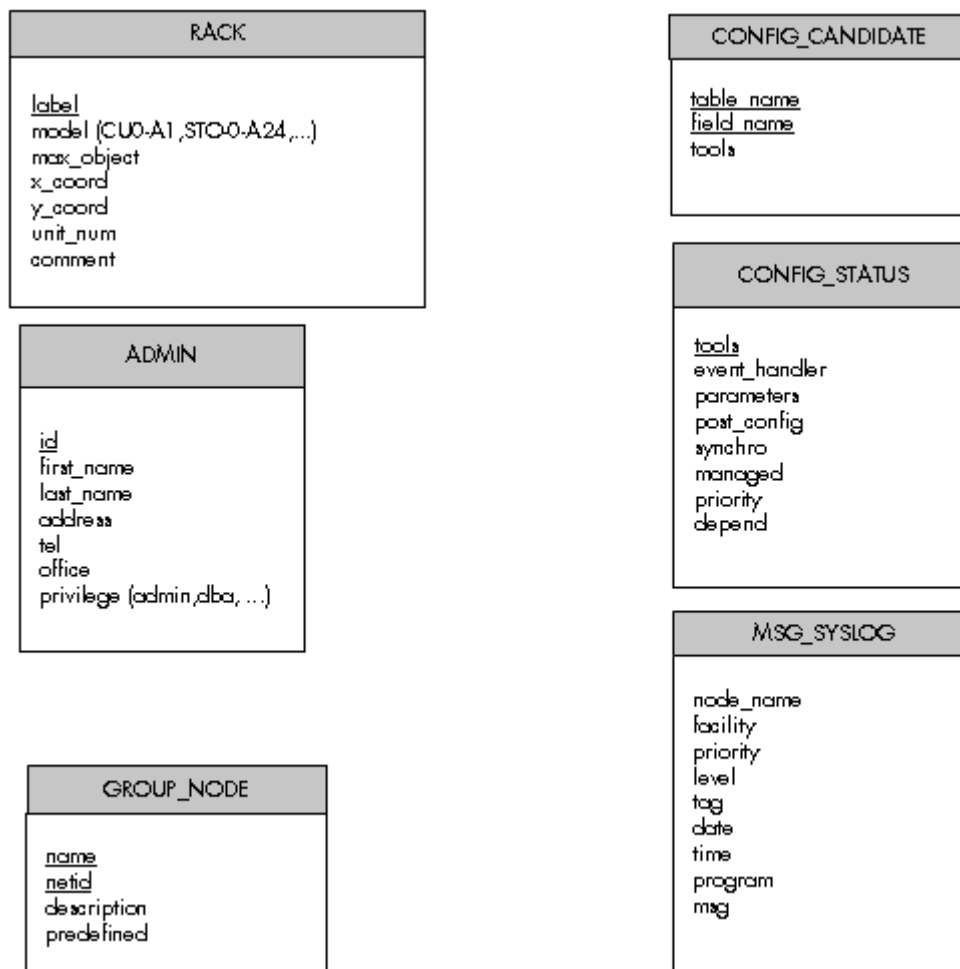


Figure 3-8. Cluster Database – Complementary tables

### 3.5.5.1 ADMIN Table

Column name	Description	Example	Fill in method
Id	PK		Sequence
first_name	First name	Stephane	admin
last_name	surname	Dupont	admin
address	address	...	admin
tel	Phone number		admin
office	office		admin
privilege	admin, dba, ....		admin

Table 3-35. Cluster Database – Admin table

### 3.5.5.2 RACK Table

Column name	Description	Example	Fill in method
Label	PK	RACK1	preload
Model	Type of rack	ARM3	preload
max_object	Maximum number of objects in the rack	3	preload
x_coord	Abscissa in the rows of racks		preload
y_coord	Ordinate in the length of racks		preload
unit_num	Number of theUnit	5	unused
comment	Free field		

Table 3-36. Cluster Database – Rack table

### 3.5.5.3 CONFIG\_CANDIDATE Table

Column name	Description	Example	Fill in method
table_name	PK	node	creation
filed_name	PK	admin_ipaddr	creation
tools	list of the candidates tools	nagios, conman	creation

Table 3-37. Cluster Database – Config Candidate table

### 3.5.5.4 CONFIG\_STATUS Table

Column name	Description	Example	Fill in method
tools	PK	nagios	creation
event_handler	generator of conf file	initNagiosCfg	creation
parameters	parameters of the event handlers	1,5,10	trigger
post_config	service to restart	nagios	creation
synchro	boolean, to be synchronized	True	trigger - dbmConfig
managed	Deactivation of the tool	True	creation
priority	Synchronisation order	1	creation
depend	List of the inter-dependency of the tool	group	creation

Table 3-38. Cluster database – Config\_Status table

### 3.5.5.5 GROUP\_NODE Table

Column name	Description	Example	Fill in method
name	PK	graphique	dbmGroup
netid	PK	10-20,25,30	dbmGroup
description	Comment about the group		dbmGroup
predefined	Predefined group	True	dbmGroup

Table 3-39. Cluster Database Group\_Node table

### 3.5.5.6 MSG\_SYSLOG Table

This table is not active in this version.

## 3.5.6 NsDoctor View

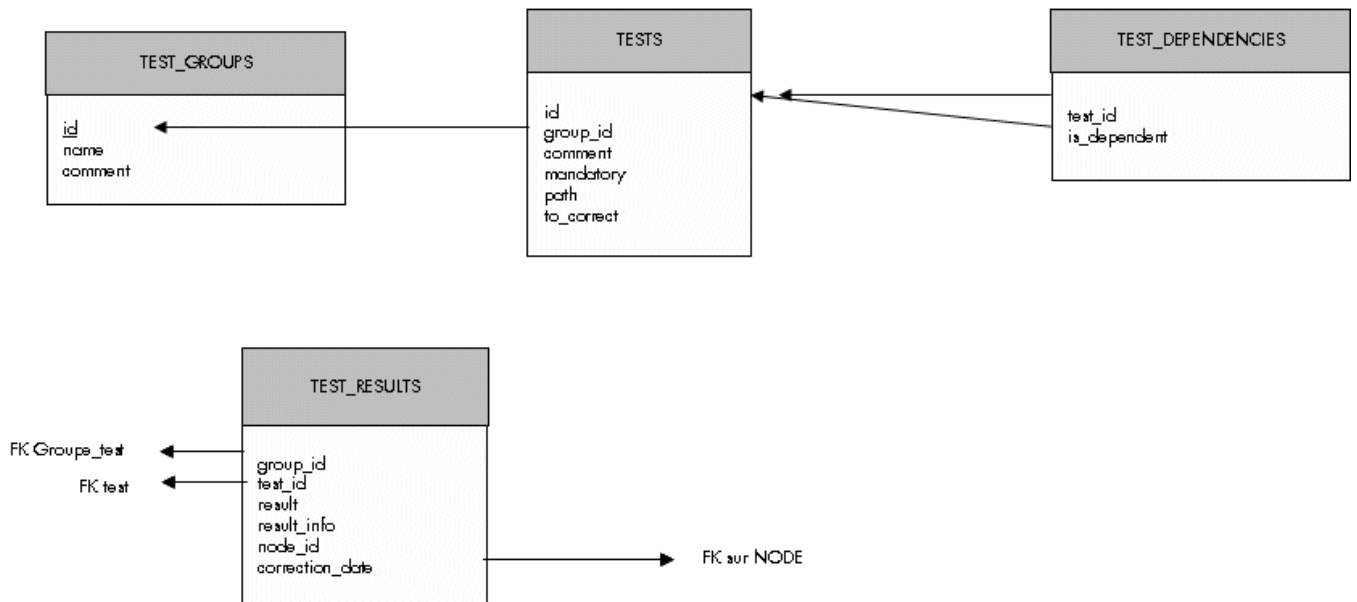


Figure 3-9. Cluster Database – NsDoctor view

### 3.5.6.1 TEST\_GROUPS Table

Column name	Description	Example	Fill in method
id	Primary key		NsDoctor - Sequence
name	Name of the group of tests		preload
comment	Comment about the group		preload

Table 3-40. Cluster Database NsDoctor – Test\_Groups table

### 3.5.6.2 TESTS Table

Column name	Description	Example	Fill in method
Id	Identifier of the test		NsDoctor -sequence
group_id	FK on GROUP_TEST_NSDOCTOR		preload
comment	Definition of the test		preload
mandatory	The test must be run if the group is selected		preload
path	Path Unix		preload
to_correct	Y/N/I		preload

Table 3-41. Cluster Database NsDoctor – Tests table



### 3.5.6.3 TEST\_DEPENDENCIES Table

Column name	Description	Example	Fill in method
test_id	Identifier of the test		preload
is_dependent	Identifier of the test		preload

Table 3-42. Cluster Database NsDoctor – Test\_Dependencies table

### 3.5.6.4 TEST\_RESULTS Table

Column name	Description	Example	Fill in method
group_id	Identifier of the group		NsDoctor
test_id	Identifier of the test		NsDoctor
result	Result of the test (True ou False)		NsDoctor
node_id	FK on NODE		NsDoctor
correction_date	Date of the correction		NsDoctor

Table 3-43. Cluster Database NsDoctor – Test Results table

## 3.5.7 Nagios View

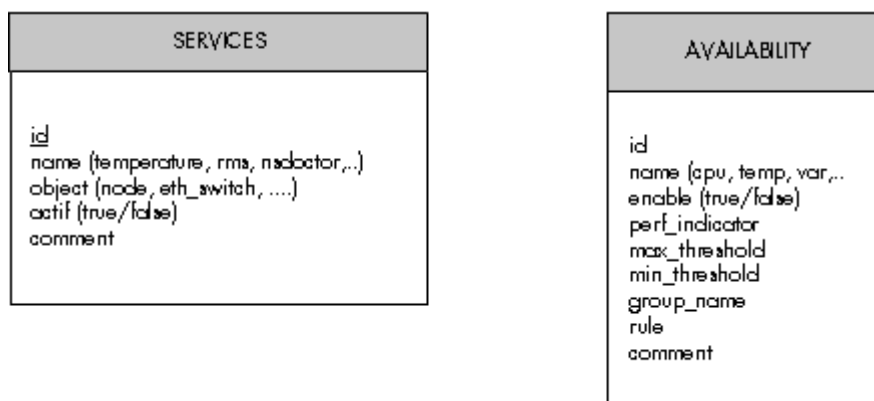


Figure 3-10. ClusterDB –Nagios View

### 3.5.7.1 Nagios SERVICES Table

Column name	Description	Example	Fill in method
id	Service id		dbmConfig
name	Service name	temperature	dbmConfig
object	Node , Eth_switch, portserver, etc.	node	dbmConfig
actif	Status of the service	true	Config & dbmServices
comment	comment	Temperature of the node	dbmConfig

Table 3-44. Nagios Services Table

### 3.5.7.2 Nagios AVAILABILITY Table

Column name	Description	Example	Fill in method
id	Service id		
name	CPU, temp, var	cpu	
enable	To check	true	
perf_indicator	Performance indicator	true	
max_thresold	Maximum threshold		
min_thresold	Minimum threshold		
group_name	Application group		
rule	Criterion rule		
comment	comment		

Table 3-45. Nagios Availability Table

### 3.5.8 Lustre View



**Note:**

These tables will not be filled for **BAS4 for Xeon** systems.

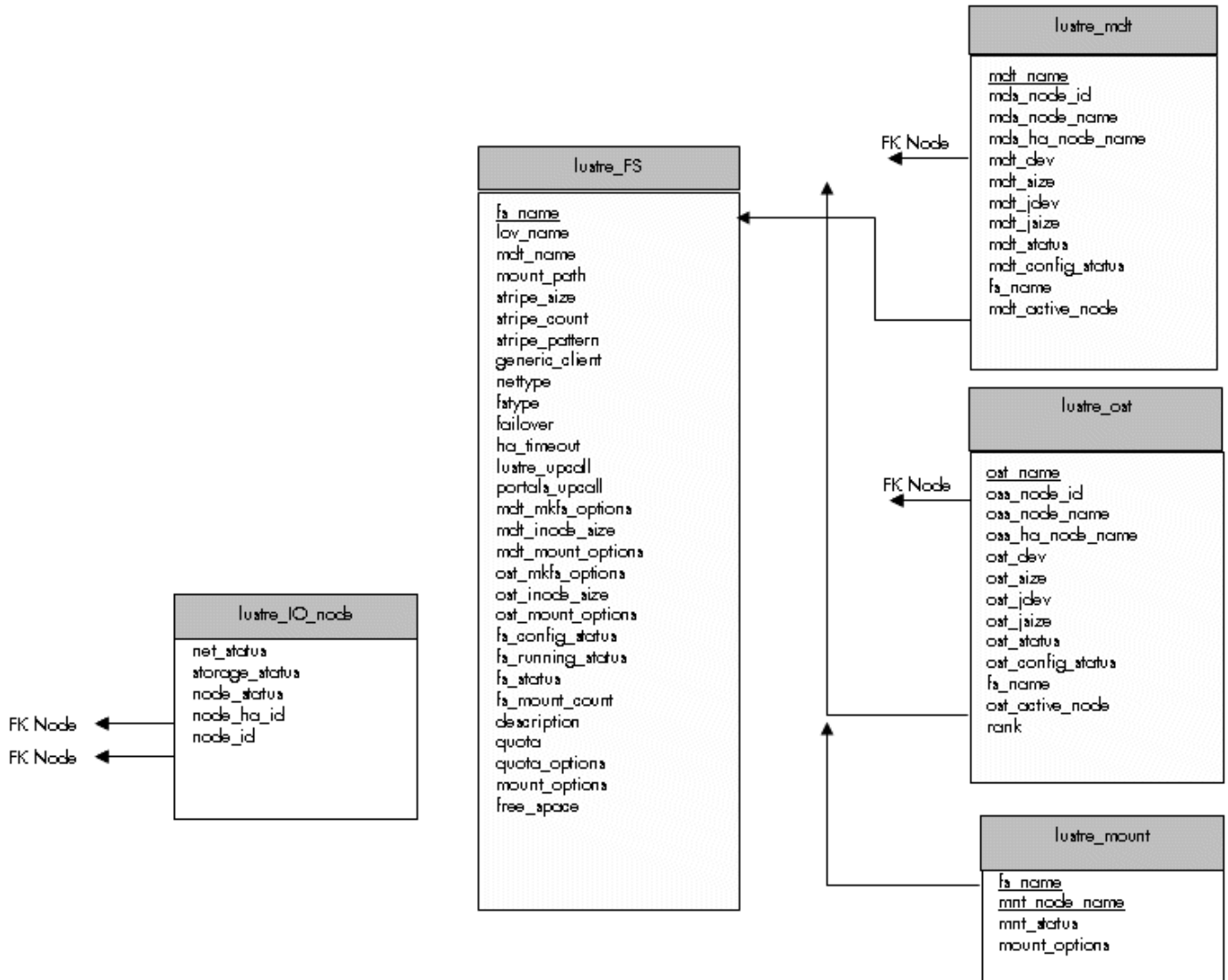


Figure 3-11. Cluster Database – Lustre view

#### 3.5.8.1 lustre\_fs Table

Each entry of the table describes a Lustre file system currently installed on the cluster.

Column name	Description	Example	Fill in method
<code>fs_name</code>	File system name	<code>lustre_basic</code>	<code>lustre_config</code>
<code>mount_path</code>	File system mount path	<code>/mnt/lustre_basic</code>	<code>lustre_config</code>
<code>lov_name</code>	LOV identification	<code>lov_lustre_basic</code>	<code>lustre_config</code>
<code>mdt_name</code>	MDT reference	<code>mdt_ns44_1</code>	<code>lustre_config</code>
<code>stripe_size</code>	Stripe size	4MB	<code>lustre_config</code>
<code>stripe_count</code>	Number of stripe per file	0 (all included OSTs)	<code>lustre_config</code>

stripe_pattern	Striping mode	0 (RAID0)	lustre_config
generic_client	Generic client profile	« client »	lustre_config
Nettype	Network type	elan	lustre_config
Fstype	Back-end file system type	ldiskfs	lustre_config
Failover	High-Availability indicator	« YES »	lustre_config
ha_timeout	High-Availability timeout for compute nodes	30	lustre_config
lustre_upcall	Lustre Exception processing script	/usr/bin/lustre_upcall	lustre_config
Portals_upcall	Portals layer exception processing script	/usr/bin/lustre_upcall	lustre_config
mdt_mkfs_options	MDT formatting options	mkfs command semantic	lustre_config
mdt_inode_size	Inode size for MDT back-end file system	1024	lustre_config
mdt_mount_options	MDT mount options	Mount command semantic	lustre_config
ost_mkfs_options	OSTs common formatting options	mkfs command semantic	lustre_config
ost_inode_size	Inode size for OSTs back-end file systems	1024	lustre_config
ost_mount_options	OSTs mount options	Mount command semantic	lustre_config
fs_config_status	File system configuration status		lustre_config
fs_running_status	File system current running status		Lustre monitoring tools
fs_status	File system status		Lustre monitoring tools
fs_mount_count	File system mount counter	54	lustre_util
description	File system characteristics decription		lustre_config
Quota	User quotas management indicator	"YES"	lustre_config
quota_options	Quotas management tuning options		lustre_config
mount_options	Default mount options for the file system		lustre_config

Table 3-46. Cluster Database – Lustre View – Lustre\_fs table

### 3.5.8.2 lustre\_ost Table

Each entry of the table describes an OST available on the cluster.

Column name	Description	Example	Fill in method
ost_name	OST logical name	OST_ns32_1	lustre_investigate
oss_node_id	OSS ident in the node table	5	lustre_investigate
oss_node_name	Supporting OSS node name	ns32	lustre_investigate
oss_ha_node_name	Secondary OSS node name	ns33	lustre_investigate

Column name	Description	Example	Fill in method
ost_active_node	In case of HA management, current node name support	ns32	lustre_migrate
ost_dev	OST back-end device name	/dev/ldn.45.1	lustre_investigate
ost_size	OST back-end device size	140000000000	lustre_investigate
ost_jdev	External journal device name	/dev/ldn.45.2	lustre_investigate
ost_jsize	External journal device size	100000	lustre_investigate
ost_config_status	OST service configuration status		lustre_config
ost_status	OST service running status		Lustre management tools
fs_name	Proprietary file system name	lustre_basic	lustre_config

Table 3-47. Cluster Database – Lustre view – Lustre OST table

### 3.5.8.3 lustre\_mdt Table

Each entry of the table describes an MDT available on the cluster.

Column name	Description	Example	Fill in method
mdt_name	MDT logical name	MDT_ns32_1	lustre_investigate
mds_node_id	MDS ident in the node table	5	lustre_investigate
mds_node_name	Supporting MDS node name	ns32	lustre_investigate
mds_ha_node_name	Secondary MDS node name	ns33	lustre_investigate
mdt_active_node	In case of HA management, current node name support	ns32	lustre_migrate
mdt_dev	MDT back-end device name	/dev/ldn.45.1	lustre_investigate
mdt_size	MDT back-end device size	140000000000	lustre_investigate
mdt_jdev	External journal device name	/dev/ldn.45.2	lustre_investigate
mdt_jsize	External journal device size	100000	lustre_investigate
mdt_config_status	MDT service configuration status		lustre_config
mdt_status	MDT service running status		Lustre management tools
fs_name	Proprietary file system name		lustre_config

Table 3-48. Cluster Database – Lustre View – Lustre\_MDT Table

### 3.5.8.4 lustre\_io\_node Table

Each cluster node of I/O (I) or metadata (M) type has an entry in this table.

Column name	Description	Example	Fill in method
node_id	Ident of the node in the node table	ns32	preload
node_ha_id	Ident of the HA paired node in the node table	ns33	preload
net_status	Node network status	% available (0 – 33 – 66 – 100)	Lustre monitoring tools

storage_status	Node storage status	% available (0 – 12 – 25 - ... - 100)	Lustre monitoring tools
node_Status	Node lustre status		Failover tools

Table 3-49. Cluster Database – Lustre View – Lustre\_IO\_node table

### 3.5.8.5 lustre\_mount Table

Each entry of this table refers to a couple compute node / mounted Lustre file system.

Column name	Description	Example	Fill in method
mnt_node_name	Compute node name	ns87	lustre_util
nnt_status	Mount point status		lustre_util
fs_name	File system name		lustre_util
mount_options	Lustre file system current mount options for the compute node		lustre_util

Table 3-50. Cluster Database – Lustre view – Lustre\_mount table

---

## Chapter 4. Parallel File Systems

This chapter explains how these file systems operate on a Bull HPC system. It describes in detail how to install, configure and manage the Lustre file system.

The following topics are described:

- 4.1 *Parallel File Systems Overview*
- 4.2 *Lustre Overview*
- 4.3 *Lustre Administrator's Role*
- 4.4 *Planning a Lustre System*
- 4.5 *Lustre System Management*
- 4.6 *Installing and Managing Lustre File Systems*
- 4.7 *Monitoring Lustre System*

### 4.1 Parallel File Systems Overview

Parallel file systems are specifically designed to provide very high I/O rates when accessed by many processors at once.

A parallel file system provides network access to a "virtual" file system distributed across different disks on multiple independent servers or I/O nodes. Real files are split into several chunks of data or stripes, each stripe being written onto a different component in a cyclical distribution manner (striping).

For a parallel file system based on a client/server model, the servers are responsible for file system functionality and the clients provide access to the file system through a "mount" procedure. This mechanism provides a consistent namespace across the cluster and is accessible via the standard Linux I/O API.

I/O operations occur in parallel across multiple nodes in the cluster simultaneously. As all files are spread across multiple nodes (and even I/O buses and disks), I/O bottlenecks are reduced and the overall I/O performance is increased.

## 4.2 Lustre Overview

**Lustre** - a parallel file system - manages the data shared by several nodes, which is dispatched in a coherent way (cyclical distribution) on several disk systems. Lustre works in client / server mode. The server part supplies the functions of the file system, while the client part enables access to the file system through a mounting configuration.

Lustre relies on a set of Data and Meta Data servers which manage the following information related to the files:

- File attributes (name, access rights, hierarchy, etc.).
- File geometry, which means how a file is distributed across different servers.

When a node of the cluster needs access to the global file system, it will mount it locally via the client part. All the nodes can have access to the global file system.

### MDS (MetaData Server)

MDS provides access to services called MDTs (MetaData Target).

A MDT provides a global NameSpace for a Lustre file system: it unifies the directory trees available from multiple file servers to provide a single global directory tree that can be mounted by the Lustre file system clients.

A MDT manages a backend ext3-like file system which contains all the metadata but none of the actual file data for an entire Lustre file system.

### OSS (Object Storage Server)

OSS provides access to services called OST (Object Storage Targets).

An OST contains part of the file data (striping) for a given Lustre file system and very little metadata.

Each OST has its own block device and backend file system where it stores stripes of files in local ext3-like files.

One MDT and several OSTs make up a single Lustre file system and can be accessed through a Logical Object Volume (LOV). This set is managed as a group and can be compared to a NFS export or a LVM logical volume.

The LOV service is replicated on all the client nodes mounting the Lustre file system and distributes the I/O locking load among OSTs.

### Lustre Client

A Lustre client results from the combination of an Object Storage Client (OSC) accessing the LOV.

A client node mounts the Lustre file system over the network and accesses the files with POSIX semantics.

Each client communicates directly with MDS and OSS.



## 4.3 Lustre Administrator's Role

Once the hardware has been setup and the software has been deployed, cluster administrators must perform the following tasks:

Determine how the hardware infrastructure will be used (number of file systems, size, storage resources used, allocation of I/O nodes, accessibility of the various file systems by the Lustre clients, etc.).

If necessary, modify the configuration of the storage devices and the configuration of the Quadrics interconnects (network zoning, etc).

Configure the Lustre service and activate the configured file systems.

During the file system lifetime, administrators may have to perform operations such as stop, start, or repair. They may decide to update a configuration or to change the one loaded. They also need to monitor the file system to check the current performance in case of degradation of service.

## 4.4 Planning a Lustre System

### 4.4.1 Data Pipelines

There are many data pipelines within the Lustre architecture, but there are two in particular which have a very direct performance impact: the network pipe between clients and OSSs, and the disk pipe between the OSS software and its backend storage. Balancing these two pipes maximizes performances.

### 4.4.2 OSS / OST Distribution

The number of clients has no real impact on the number of OSSs to be deployed. To determine the number of OSSs and how to distribute OSTs, two things have to be considered:

- The maximum bandwidth required gives the number of OSSs.
- The total storage capacity needed gives the number of OSTs.

To increase efficiency, it is preferable to distribute OSTs evenly on OSSs and to have fewer larger OSTs in order to use space more efficiently.

When calculating the size of the OSS server nodes, it is recommended that the CPUs are divided into thirds: one third for the storage backend, one third for the network stack and one third for Lustre.

### 4.4.3 MDS / MDT Distribution

The Lustre file system stores the file striping information in extended attributes (**EAs**) on the MDT. If the file system has large-inode support enabled (> 128bytes), then EA information will be stored inline (fast EAs) in the extra space available.

The table below shows how much stripe data can be stored inline for various inode sizes:

Inode Size	#of stripes stored inline
128	0(all EA is stored externally)
256	3
512	13
1024	35
2048	77
4096	163

Table 4-1. Inode Stripe Data

It is recommended that MDT file systems be formatted with the inode large enough to hold the default number of stripes per file to improve performance and storage efficiency. One needs to keep enough free space in the MDS file system for directories and external blocks. This represents ~512 Bytes per inode.

## 4.4.4 File Striping

Lustre stripes the file data across the OSTs in a round robin fashion.

It is recommended to stripe over as few objects as is possible to limit network overhead and to reduce the risk of data loss, in case of OSS failure.

The stripe size must be a multiple of the page size. The smallest recommended stripe size is 512 KB because Lustre tries to batch I/O into 512 KB blocks on the network.

## 4.4.5 Lustre File System Limitations

On the device it manages, an OST reserves up to 400MB for an internal journal and 5% for the root user. This reduces the storage capacity available for the user's data. Like an OST, on the device it manages a MDT reserve.

The encapsulated modified ext3 file system used by MDTs and OSTs relies on the standard ext3 file system provided by the Linux system and optimizes performance and block allocation. It has the same limits in terms of maximum file size and maximum file system size.

## 4.5 Lustre System Management

Bull Lustre management tools provide services to manage large parallel file systems during their whole life cycle. Using these tools the cluster administrator will be able to:

- Configure and install **Lustre** file systems using the Lustre OST/MDT services provided by the storage management model deployment (refer to the Storage Devices Management chapter).
- Perform management operations such as start/stop, mount/umount file systems.

The administrator can monitor and get information about the Lustre system via a graphical user interface for performance and health monitoring.

Status targets of management tools for Lustre file systems and components current activity.

### 4.5.1 The Lustre Database

The Lustre management tools rely on the cluster database (ClusterDB) to store and get information about:

- I/O and Metadata nodes (lustre\_io\_node table),
- Lustre OST/MDT services (lustre\_ost and lustre\_mdt tables),
- File systems currently installed on the cluster (lustre\_fs and lustre\_mount tables).

Some of these tables information is loaded during the cluster deployment phase: those related to the I/O and Metadata nodes implementation and to the OST/MDT services repartition. The rest is maintained by the Lustre management tools.

Specific commands allow the administrator to edit and adjust information when necessary, for example in the case of node replacement due to hardware.



#### **Note:**

Updating the information stored in the Lustre database has direct consequences on the Lustre system behaviour. This must be done only by skilled administrators.

#### 4.5.1.1 lustre\_tables\_dba

##### SYNOPSIS

```
lustre_ost_dba ACTION [options]
lustre_mdt_dba ACTION [options]
lustre_fs_dba ACTION [options]
lustre_io_node_dba ACTION [options]
```

## DESCRIPTION

The `lustre_tables_dba` set of commands allows the administrator to display, parse and update the information of the Lustre tables in the ClusterDB.

<code>lustre_ost_dba</code>	Acts on the <code>lustre_ost</code> table, which describes the OST services
<code>lustre_mdt_dba</code>	Acts on the <code>lustre_mdt</code> table, which describes the MDT services
<code>lustre_fs_dba</code>	Acts on the <code>lustre_fs</code> table, which describes the Lustre file systems currently available on the cluster
<code>lustre_io_node_dba</code>	Acts on the <code>lustre_io_node</code> table, which gives the current status of the cluster I/O and metadata nodes.

These utilities are useful for checking the correctness of **ClusterDB** contents according to the last configuration updates. They allow the further adjustment of values in the instance of mistakes or failures of the Lustre management utilities thus avoiding a full repeat of the operation. They can also be used to force the Lustre services behaviour for some precise and controlled cases.

As these act on the global cluster configuration information, they must be used very carefully. The changes they allow may introduce fatal inconsistencies in the Lustre ClusterDB information.

## ACTIONS

<code>add</code>	Adds an object description in the Lustre table.
<code>update</code>	Updates configuration items of an object description in the Lustre table.
<code>del</code>	Removes an object description from the Lustre table.
<code>attach</code>	Creates a link between Lustre tables objects (i.e. attaches an OST to a file system).
<code>detach</code>	Removes a link between Lustre tables objects (i.e. frees an OST).
<code>list</code>	Displays object information according to the selected criteria provided by the options.
<code>set</code>	Sets one or more status items of an object description.
<code>-h(elp)</code>	Displays this help and exits.
<code>-v(ersion)</code>	Displays the utility version and exits.

## OPTIONS

The options list available for the actions depends on the kind of object they act on and on the action itself. Please, refer to the help of each command for option details.

## 4.5.2 /etc/lustre/storage.conf for Lustre Tools without ClusterDB

The `/etc/lustre/storage.conf` file stores information about the storage devices available on the cluster when ClusterDB is **NOT** present and it records which ones are OSTs and which ones are MDTs. It must be located on the management node. This file is composed of lines with the following syntax:

```
<ost|mdt>: name=<> node_name=<> dev=<> [ ha_node_name=<> ] [ size=<kB> ] [
jdev=<> [ jsize=<kB> ] ]
```

<b>ost/mdt</b>	This device is designated to be either an OST or a MDT.
<b>name</b>	The name given to the OST or MDT.
<b>node_name</b>	The hostname of the node containing the device.
<b>dev</b>	The device path (for example <code>/dev/sdd</code> ).
<b>ha_node_name</b>	The hostname of the failover node. This has to be consistent with the content of <code>/var/lustre/status/lustre_io_nodes</code> .
<b>size</b>	Size of the device in kB.
<b>jdev</b>	The name of the device where the <b>ext3</b> journal will be stored, if this is to be outside the main device. This parameter is optional. Loop devices cannot be used for this purpose.
<b>jsize</b>	The size of the journal device in kB.

Comments are lines beginning with # (sharp).

### 4.5.2.1 Filling /etc/lustre/storage.conf

This file is updated with the information obtained from the `/proc/partitions` or `/sys/block/` of the I/O nodes. For example, on a cluster where **ns13** is an I/O node:

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

```
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
```

**sda** and **sdb** are system disks of **ns13** so they must NOT be used as Lustre storage devices. Devices **sdd** to **sdh** are the devices which are available. 17430528 kB disks will be used as journal devices and 75497472 kB disks as the main devices.

This choice results in the following lines being included in `/etc/lustre/storage.conf` file for the management node:

```
mdt: name=ns13_sdd node_name=ns13 dev=/dev/sdd size=75497472
jdev=/dev/sdc jsize=17430528
ost: name=ns13_sdf node_name=ns13 dev=/dev/sdf size=75497472
jdev=/dev/sde jsize=17430528
ost: name=ns13_sdh node_name=ns13 dev=/dev/sdh size=75497472
jdev=/dev/sdg jsize=17430528
```

The decision as to which devices will be used as **MDTs** and which will be **OSTs** will be left to the administrator. This procedure has to be done for each I/O node and new lines appended to the `/etc/lustre/storage.conf` file of the management node. Bull provides a wizard to help the creation of the `storage.conf` file, this is `/usr/lib/lustre/lustre_storage_config.sh`.

#### 4.5.2.2 Storage Inspection Wizard: `/usr/lib/lustre/lustre_storage_config.sh`

`/usr/lib/lustre/lustre_storage_config.sh` is a script that helps the administrator to complete the `storage.conf` file.

##### SYNOPSIS

```
lustre_storage_config.sh <node> <regexp_on_devices>
<upper_size_limit_of_journal_in_kb>
```

<b>node</b>	This details the node to be inspected
<b>regexp_on_devices</b>	A regular expression on the device name as given in the <code>/dev/</code> directory of the I/O node. Do not forget this is a regular expression, not a globbing expression.
<b>upper_size_limit_of_journal_in_kb</b>	The difference between data devices and journal devices is made according to their size in kB. If a device has a size greater than this parameter, it is assumed to be a data device. If a device has a size smaller than this parameter, it is assumed to be journal device.

The output produced by the `lustre_storage_config.sh` script is a template of lines to be used by `storage.conf`. These lines may require minor modifications. `lustre_storage_config.sh` looks for `/var/lustre/status/lustre_io_nodes` that can be used to fill in the `ha_node` field, therefore `lustre_io_nodes` should have been filled in before running the `lustre_storage_config.sh` script.

For example, for a cluster with two High Availability I/O nodes: **ns6** and **ns7**. The content for the `lustre_io_nodes` is as follows:

```
NODE_NAME=ns6
NODE_HA_NAME=ns7
LUSTRE_STATUS=OK
```

```
NODE_NAME=ns7
NODE_HA_NAME=ns6
LUSTRE_STATUS=OK
```

1. Link to Lustre devices is run using **stordiskname** and the output is as follows:

```
[root@ns6 ~]# ll /dev/ldn.nec.*
lrwxrwxrwx 1 root root 8 May 19 13:23 /dev/ldn.nec.0 -> /dev/sdd
lrwxrwxrwx 1 root root 8 May 19 13:23 /dev/ldn.nec.1 -> /dev/sde
lrwxrwxrwx 1 root root 8 May 19 13:23 /dev/ldn.nec.2 -> /dev/sdn
lrwxrwxrwx 1 root root 8 May 19 13:23 /dev/ldn.nec.3 -> /dev/sdo
```

```
[root@ns7 ~]# ll /dev/ldn.nec.*
lrwxrwxrwx 1 root root 8 May 19 13:23 /dev/ldn.nec.0 -> /dev/sdd
lrwxrwxrwx 1 root root 8 May 19 13:23 /dev/ldn.nec.1 -> /dev/sde
lrwxrwxrwx 1 root root 8 May 19 13:23 /dev/ldn.nec.2 -> /dev/sdn
lrwxrwxrwx 1 root root 8 May 19 13:23 /dev/ldn.nec.3 -> /dev/sdo
```

The same devices can be seen on both **ns6** and **ns7**.

2. All devices that start with **ldn** are to be used however it is not clear for the moment which are data devices and which are journal devices. From the management node the **lustre\_storage\_config.sh** script is run.

```
[root@ns2 /]# /usr/lib/lustre/lustre_storage_config.sh ns6 'ldn.*' 0
#call: ns6 ldn.* 0
```

The resulting output is as follows:

```
ost: name=ost_ns6.nec.0 node_name=ns6 ha_node_name=ns7
dev=/dev/ldn.nec.0 size=262144
ost: name=ost_ns6.nec.1 node_name=ns6 ha_node_name=ns7
dev=/dev/ldn.nec.1 size=262144
ost: name=ost_ns6.nec.2 node_name=ns6 ha_node_name=ns7
dev=/dev/ldn.nec.2 size=46137344
ost: name=ost_ns6.nec.3 node_name=ns6 ha_node_name=ns7
dev=/dev/ldn.nec.3 size=46137344
```

From the output above it can be seen that there are two sizes for the devices, the data devices (46137344 kB) and the journal devices (262144 kB).

3. The size of the journal device has been identified as 262144 kB, this means that the following command can be run:

```
[root@ns2
```

The output is as follows:

```
/]# /usr/lib/lustre/lustre_storage_config.sh ns6 'ldn.*'
262144
#call: ns6 ldn.* 262144
ost: name=ost_ns6.nec.2 node_name=ns6 ha_node_name=ns7
dev=/dev/ldn.nec.2 size=46137344 jdev=/dev/ldn.nec.0 jsize=262144
ost: name=ost_ns6.nec.3 node_name=ns6 ha_node_name=ns7
dev=/dev/ldn.nec.3 size=46137344 jdev=/dev/ldn.nec.1 jsize=262144
```



4. The output is saved in the **storage.conf** file using the following command:

```
[root@ns2 /]# /usr/lib/lustre/lustre_storage_config.sh ns6 'ldn.*' 262144
>>/etc/lustre/storage.conf
```

5. The same operation now has to be run on **ns7**, as below.

```
[root@ns2 /]# /usr/lib/lustre/lustre_storage_config.sh ns7 'ldn.*' 262144
#call: ns6 ldn.* 262144
```

The output is as follows:

```
ost: name=ost_ns7.nec.2 node_name=ns7 ha_node_name=ns6
dev=/dev/ldn.nec.2 size=46137344 jdev=/dev/ldn.nec.0 jsize=262144
ost: name=ost_ns7.nec.3 node_name=ns7 ha_node_name=ns6
dev=/dev/ldn.nec.3 size=46137344 jdev=/dev/ldn.nec.1 jsize=262144
```

6. The output above is now saved in the **storage.conf** file using the following command:

```
[root@ns2 /]# /usr/lib/lustre/lustre_storage_config.sh ns7 'ldn.*' 262144
>>/etc/lustre/storage.conf
```

At this point, the same devices will be stored twice in the **storage.conf** file as shown in the output below.

```
ost: name=ost_ns6.nec.2 node_name=ns6 ha_node_name=ns7
dev=/dev/ldn.nec.2 size=46137344 jdev=/dev/ldn.nec.0 jsize=262144
ost: name=ost_ns6.nec.3 node_name=ns6 ha_node_name=ns7
dev=/dev/ldn.nec.3 size=46137344 jdev=/dev/ldn.nec.1 jsize=262144
ost: name=ost_ns7.nec.2 node_name=ns7 ha_node_name=ns6
dev=/dev/ldn.nec.2 size=46137344 jdev=/dev/ldn.nec.0 jsize=262144
ost: name=ost_ns7.nec.3 node_name=ns7 ha_node_name=ns6
dev=/dev/ldn.nec.3 size=46137344 jdev=/dev/ldn.nec.1 jsize=262144
```

7. A decision has to be made at this point as to which devices will have **ns6** as the master node, and which devices will have **ns7** as the master node. An example is shown below:

```
ost: name=ost_ns6.nec.2 node_name=ns6 ha_node_name=ns7
dev=/dev/ldn.nec.2 size=46137344 jdev=/dev/ldn.nec.0 jsize=262144
ost: name=ost_ns7.nec.3 node_name=ns7 ha_node_name=ns6
dev=/dev/ldn.nec.3 size=46137344 jdev=/dev/ldn.nec.1 jsize=262144
```

8. The first device should be designated as an **mdt**. This is done by replacing **ost** by **mdt**, as shown below:

```
mdt: name=mdt_ns6.nec.2 node_name=ns6 ha_node_name=ns7
dev=/dev/ldn.nec.2 size=46137344 jdev=/dev/ldn.nec.0 jsize=262144
ost: name=ost_ns7.nec.3 node_name=ns7 ha_node_name=ns6
dev=/dev/ldn.nec.3 size=46137344 jdev=/dev/ldn.nec.1 jsize=262144
```

9. **storage.conf** is now ready. If you have more than one pair of High Availability nodes then the same operation will have to be repeated for each pair of nodes.

10. The consistency of the `/etc/lustre/storage.conf` files can be checked using the command below optimal:

```
lustre_util check_storage.
```



**Important:**

`lustre_storage_config.sh` associates the data devices and journal devices in alphabetical order. On some devices, for example DDN, this association is not necessarily optimal and special tuning may be required to improve the performance.



**Note:**

If it is planned to upgrade the cluster from one which does not have a database installed to one which includes a database then `lustre_util check_storage` should not report any errors.

### 4.5.2.3 Loading storage.conf into the Cluster Database using load\_storage.sh

`load_storage.sh` is a script that is used to load `storage.conf` information into the `lustre_ost` and `lustre_mdt` tables of the cluster database. This may be useful:

- If a cluster database is added to your system.
- If there is a database, but no management tools are provided for the storage devices, for example for **NEC** devices.

#### SYNOPSIS

```
/usr/lib/lustre/load_storage.sh < update|crush > <storage.conf file>
```

<b>update</b>	Adds the new devices but does not erase the existing <code>lustre_ost</code> and <code>lustre_mdt</code> tables.
<b>crush</b>	Remove the <code>lustre_ost</code> and <code>lustre_mdt</code> tables, and then add new devices.
<b>storage.conf file</b>	The path to your <code>storage.conf</code> file (this usually <code>/etc/lustre/storage.conf</code> )

### 4.5.2.4 Practical Recommendation

If you use a High Availability MDS as the management node it will be possible to move the `/etc/lustre/storage.conf` file to `/var/lustre/status/`, using the command below and to then make a symbolic link to this file on the 2 MDS nodes:

```
ln -s /var/lustre/status/storage.conf /etc/lustre/storage.conf
```

The same thing can be done for the `/etc/lustre/models` directory. In this way, information does need to be replicated and is available on the node where `/var/lustre/status` is mounted.

### 4.5.3 Lustre Networks

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

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

### 4.5.4 Lustre Management Configuration File: `/etc/lustre/lustre.cfg`

Lustre management tools use this file to get configuration information. This file must reside on all OSS and MDS nodes. Refer to `lustre_util` man page to know how to distribute this file easily.

#### File Syntax:

VARIABLE=VALUE

Lines beginning with `#` are comments.

#### `/etc/lustre/lustre.cfg` contents:

##### **LUSTRE\_MODE=XML**

**XML:** Information about filesystems is given to CFS Lustre tools using the XML format. These files are stored in the directory defined by `LUSTRE_CONFIG_DIR` on OSS, MDS and Management Node.

Default value is XML. This value is mandatory for failover configuration. **HTTP mode is no longer supported.**

##### **CLUSTERDB=yes**

When this variable is set to yes, storage, file systems and mount information is retrieved and stored from the clusterDB tables (`lustre_ost`, `lustre_mdt`, `lustre_mount` and `lustre_fs`).

##### **LUSTRE\_CONFIG\_DIR=/etc/lustre/conf/**

This variable contains the path of the directory where the XML/XMF files are created on the Management Node and where they have to be found on the OSSs and MDSs. The `lustre_util` command uses this path to store and read XML/XMF when required. This directory can be shared using NFS. If `LUSTRE_MODE` is set to XML, `lustre_util` creates this directory on all OSS and MDS nodes in order to copy the XML file associated with filesystems during the install process, as required by CFS Lustre tools (`lconf`).

Default value is `/etc/lustre/conf/`.

##### **LUSTRE\_NET=tcp or elan or o2ib**

This variable specifies the kind of network used by Lustre on the whole cluster.

Default value is `tcp`. It is only used by the `lustre_check` monitoring tool.

#### **LUSTRE\_ADMIN=hostname**

This variable contains the hostname of the I/O server used as central point of management for Lustre in case of cluster not monitored by a management station (`CLUSTERDB="no"`). The primary MDS node is to be chosen for that purpose.

**No default value is defined.**

#### **LUSTRE\_ADMIN2=hostname**

LUSTRE\_ADMIN2 is used only if the HA feature is enabled on the I/O nodes. It provides the hostname of the backup MDS used as alternative Lustre management point.

**No default value is defined.**

#### **LUSTRE\_LDAP\_URL=ldap://hostname/**

This variable contains the address of the ldap server used to store HA information. For example if the ldap server is on a node called `ns2`, then

LUSTRE\_LDAP\_URL=<ldap://ns2/>.

**No default value is defined.**

#### **LUSTRE\_LDAP\_URL2=ldap://hostname/**

LUSTRE\_LDAP\_URL2 is used only when there is no management station supporting the full HA feature. In this case, it provides the LDAP URL of an alternative management station.

**No default value is defined.**

#### **LUSTRE\_DEBUG=yes or no**

If this variable is set to "yes", Lustre management daemons are allowed to log trace information:

- in `/var/log/lustre` directory for failover
- in `/tmp/log/lustre` directory for database daemons

**Default value is no.**

#### **I/O scheduler for block devices**

**LUSTRE\_OST\_DEV\_IOSCHED** = `noop` or `anticipatory` or `deadline` or `cfq`  
(I/O scheduler for OST devices)

**LUSTRE\_OST\_JNR\_IOSCHED** = `noop` or `anticipatory` or `deadline` or `cfq`  
(I/O scheduler for OST ext3 journal devices)

**LUSTRE\_MDT\_DEV\_IOSCHED** = `noop` or `anticipatory` or `deadline` or `cfq`  
(I/O scheduler for MDT devices)

**LUSTRE\_MDT\_JNR\_IOSCHED** = `noop` or `anticipatory` or `deadline` or `cfq`  
(I/O scheduler for MDT ext3 journal devices)

These variables define the I/O scheduler for block devices. For details about I/O schedulers refer to the `/Documentation/block` directory of kernel sources.

Default and recommended values are:

- **deadline** for `LUSTRE_MDT_DEV_IOSCHED`,
- **noop** for `LUSTRE_OST_DEV_IOSCHED`, `LUSTRE_OST_JNR_IOSCHED` and `LUSTRE_MDT_JNR_IOSCHED`.

If OSTs/MDTs are disc partitions (not the whole device) the choice of the scheduler is left to the Administrator.

#### **LUSTRE\_SNMP=yes or no**

If this variable is set to yes, the **snmpd** server will be enabled on the I/O nodes when **lustre\_util set\_cfg** is called (`chkconfig --level 345 snmpd on && service snmpd restart`). This allows the OSS and MDS to send snmp traps to the Management Node when errors occur. These traps force the nagios lustre service to run in order to check the health of the filesystems.

**Default value is no.**

#### **DISABLE\_LUSTRE\_FS\_NAGIOS=yes or no**

Setting this to yes will disable the call of `lustre_fs_nagios` every 15 mn on management node.

**Default value is no**

#### **LUSTRE\_TUNING\_FILE=/etc/lustre/tuning.conf**

This is the path to the tuning file, the default value is `/etc/lustre/tuning.conf`.

## 4.5.5 Lustre Services Definition

The Lustre services MDT(s) and OST(s) rely on the devices created by the storage units configuration deployment. For this reason their distribution schema is tightly dependant of the storage configuration planning and vice versa.

A common model and deployment process is used for both storage units and Lustre services. The model describes the relationship between storage units, nodes, devices and Lustre services.

Refer to the “Storage Administration” chapter for more information.

Each Lustre service defined on the cluster I/O nodes is described by an entry in the **ClusterDB**. During the first cluster deployment phase, the model file is parsed for the storage elements which are created on the nodes and the information related to the Lustre services is stored in the Lustre tables of the **ClusterDB**, **lustre\_mdt** for MDT services and **lustre\_ost** for OST services.

This is theoretical information, which needs to be checked against the node reality using the **lustre\_investigate check** utility. Inconsistencies may come from a model file error or elements configuration failure on nodes.

This check operation must be done after every cluster configuration or reconfiguration operation or every time the Lustre services configuration is modified.

### 4.5.5.1 `lustre_investigate`

#### SYNOPSIS

```
lustre_investigate check [-C <io_cell_list> |-n <nodes_list> |-f <file_system_name>]
```

```
lustre_investigate display [-C <io_cell_list> |-n <nodes_list> |-f <file_system_name>]
```

#### DESCRIPTION

`lustre_investigate` can be used only if the cluster configuration information is managed using the cluster management database, ClusterDB.

It allows the administrator to check the consistency between the information concerning the Lustre services and the real storage configuration available on I/O nodes.

Each Lustre service defined on the cluster I/O nodes is described by an entry in the ClusterDB. This entry provides information about the back-end device used by the service and the primary and the secondary node the service should run on.

Due to failures or cluster reconfiguration operations, this information may become obsolete. An availability status is maintained, which indicates if it is still correct or needs to be updated. This status is updated by running `lustre_investigate`.

`lustre_investigate` must be used from the management station. It issues check commands to each node of the selected range. The returned information is then evaluated against the one stored in the ClusterDB. It relies on the `pdsh` parallel shell to dispatch remote commands.

#### ACTIONS:

**check**                      Parses the Lustre services entries in the ClusterDB according to the select criteria and checks their information consistency.

**display**                    Displays the ClusterDB information about the Lustre services corresponding to the select criteria.

#### OPTIONS:

**-h(elp)**                    Displays this help and exits.

**-v(ersion)**                Displays the current utility version and exits.

**-C**                          Range of I/O cells (format [x,m-n]).

**-n**                          Range of nodes (format <prefix>[x,m-n]).

**-f**                          <file\_system\_name> is the name of the file system to work on.

If neither `-C`, `-n` nor `-f` are provided, all Lustre services declared in the cluster database management are processed.

## 4.5.6 Creating Lustre File Systems

### 4.5.6.1 Prerequisites

- `/etc/lustre/lustre.cfg` is assumed to be updated correctly as described in the section - Lustre Management Configuration File: `/etc/lustre/lustre.cfg`.
- If you are using a cluster database (`CLUSTERDB=yes`) `lustre_ost` and `lustre_mdt` tables are assumed to be updated correctly (use `lustre_investigate check` to verify).
- If you are not using a cluster database (`CLUSTERDB=no`), `storage.conf` must be correctly filled.
- Lustre tools use `ssh` to execute remote commands, so users must be allowed to log into nodes without being prompted for a password. This can be done by appending the right keys in `/root/.ssh/authorized_keys2`.

### 4.5.6.2 Lustre Model File (.lmf)

A Lustre model file describes one or several Lustre filesystems that can be used at the same time. This means they do not share OSTs or MDT. Such files are stored in the `/etc/lustre/models` directory.

#### File Syntax:

keyword: <value>

Lines beginning with `#` are comments.

#### Possible Keywords:

<code>stripe_size</code>	Specify the stripe size in bytes. Default is 1048576 (1M).
<code>stripe_count</code>	Specify the number of OSTs each file should be striped onto. The default value is 2.
<code>stripe_pattern</code>	Only pattern 0 (RAID 0) is supported currently.
<code>nettype</code>	Possible values are <code>tcp</code> or <code>elan</code> or <code>o2ib</code> . The default is <code>elan</code> .
<code>generic_client</code>	The name of the directory from which the filesystem is mounted on the mds. If the network is <code>elan</code> , the default is <code>clientelan</code> . If network is <code>tcp</code> , default is <code>clienttcp</code> .
<code>fstype</code>	File system type. Possible values are <code>ldiskfs</code> or <code>ext3</code> . Default (and recommended) is <code>ldiskfs</code> .
<code>failover</code>	Enable failover support on OST. Possible values are <code>yes</code> or <code>no</code> . The default is <code>no</code> .

<b>ha_timeout</b>	Timeout in seconds before going into recovery. Default is 30s.
<b>lustre_upcall</b>	Location of the Lustre <b>upcall</b> script used by the client for recovery. No default script is defined.
<b>portals_upcall</b>	Location of the Portals <b>upcall</b> script used by the client for recovery. No default script is defined.
<b>mdt_mkfs_options</b>	Optional argument to <b>mkfs</b> for MDT. By default, no option is specified.
<b>mdt_inode_size</b>	Specify new <b>inode</b> size for underlying MDT ext3 file system. The default is self-evaluated.
<b>mdt_mount_options</b>	Optional argument to mount fs locally on the MDS. By default, no option is specified.
<b>ost_mkfs_options</b>	Optional argument to <b>mkfs</b> for OST. By default, no option is specified.
<b>ost_inode_size</b>	Specify new inode size for underlying OST ext3 file system. The default is self-evaluated.
<b>ost_mount_options</b>	Optional argument to mount fs locally on OSS. By default, no option is specified.
<b>cluster_id</b>	Specify the cluster ID (one filesystem uses a single cluster id)
<b>mount_options</b>	Defines the default options to mount the filesystem on clients. Options are separated with ", ". Available options are: ro, rw, user_xattr, nouser_xattr, acl, noacl. Default is no option and the filesystem will be mounted rw. For example, <code>mount_options: ro</code> means that, by default, this filesystem is mounted in read-only mode.
<b>quota</b>	Enables quota support. Possible values are yes or no. The default is no.
<b>quota_options</b>	If quota is set to yes, it describes options for quota support. The default is: <code>quotaon=ug,iunit=5000,bunit=100,itune=50,btune=50</code> . <b>Do not use other settings for the moment.</b>
<b>description</b>	A ONE LINE free description of your filesystem (up to 512 chars). The default is empty string.

If previous keywords are used in the header of the file, before any filesystem definition (this means before any use of the **fs\_name** keyword), they set the new default values which can be locally overloaded for a filesystem.

**fs\_name** This keyword is the starting point of a filesystem definition. It is the name of the filesystem (name of the xml file or the entry in the ldap database). `fsname` must be defined for each filesystem.



<b>mount_path</b>	The mount-point to use to mount Lustre filesystem. Same mount-point must not be used for another filesystem defined in the same model file. Default is /mnt/lustre_<fs_name>.
<b>ost</b>	<p>[ name=&lt;RegExp&gt; ] [ node_name=&lt;RegExp&gt; ]  [ dev=&lt;RegExp&gt; ] [ size=&lt;RegExp&gt; ] [ jdev=&lt;RegExp&gt; ]  [ jsize=&lt;RegExp&gt; ] [ cfg_status=available formatted ]</p> <p>Specify OSTs to use with this filesystem, using regular expressions matching their name, node_name, device, size, journal device, journal size or status. At least one field must be specified. If several fields are specified, only OSTs matching all fields of the lines will be chosen. You can use as many OST lines as you need.  <b>At least one OST line must be defined for each filesystem.</b></p>
<b>mdt</b>	<p>[ name=&lt;RegExp&gt; ] [ node_name=&lt;RegExp&gt; ]  [ dev=&lt;RegExp&gt; ] [ size=&lt;RegExp&gt; ] [ jdev=&lt;RegExp&gt; ]  [ jsize=&lt;RegExp&gt; ] [ cfg_status=available formatted ]</p> <p>Specify MDT of this filesystem. It is the same syntax as for the OSTs. If several MDTs match, then the first will be used.</p>



**Note:**

One and only one MDT line must be defined for each filesystem.

### 4.5.6.3 Extended Model Files (.xmf)

The purpose of the extended model files is to maintain a strict description of a filesystem. It is planned to replace xml files with this format. They have exactly the same syntax as previous model files, except that the OSTs/MDTs are strictly described and each OST/MDT line **MUST** point to one and only one OST/MDT of the **lustre\_ost** and **lustre\_mdt** tables. They can be used in place of lmf files. They are automatically generated in **LUSTRE\_CONFIG\_DIR** when you use **lustre\_util install**, **update** or **rescue** commands.

### 4.5.6.4 Lustre Model Sample File

There follows a model file which describes two filesystems fs1 and fs2, on a cluster with nodes called ns<XX>. Information about OSTs and MDTs can be found using **lustre\_ost\_dba** list and **lustre\_mdt\_dba** list if a cluster database is present, or in **/etc/lustre/storage.conf** if no cluster database is present.

```
#####
# Firstly, the new default values for the 2
# filesystems are defined

# To prevent failover
failover: no

# Set block-size to 4096 for mdt
mdt_mkfs_options: -b 4096
```

```

# Set block-size to 4096 for osts
ost_mkfs_options: -b 4096

# Network is elan
nettype: elan

# New mount options
ost_mount_options: extents,malloc

#####
# First filesystem : fs1

# Filesystem name is fs1
fs_name: fs1

# mount-point of this filesystem will be /mnt/lustre1
# instead of the default /mnt/lustre_fs1
mount_path: /mnt/lustre1

# To specify osts hosted by nodes with names ending by odd numbers, with device
names ending from 2 to 4
ost: node_name=ns.*[1,3,5,7,9] dev=.*[2-4]

# To specify the ost named ost_ns10.ddn1.6
ost: name=ost_ns10.ddn1.6

# The mdt will be the first hosted by ns12 with a name ending with a 3
mdt: node_name=ns12 name=.*3

#####
# Second filesystem : fs2

# Filesystem name is fs2
fs_name: fs2

# mount-point of this filesystem will be /mnt/lustre2
# instead of the default /mnt/lustre_fs2
mount_path: /mnt/lustre2

# To specify osts hosted by nodes with name ending with even numbers, with device
names ending with 1,2,3 and 5
ost: node_name=ns.*[2,4,6,8,0] dev=.*[1-3,5]

# To specify the mdt named mdt_ns13.ddn12.31
mdt: name=mdt_ns13.ddn12.31

# To specify the generic_client to be fs2_client instead of
# clientelan
generic_client: fs2_client

```

## 4.6 Installing and Managing Lustre File Systems

**lustre\_util** is the tool used to install, enable, disable, mount and unmount, one or more Lustre file systems from an administration node.

### 4.6.1 Installing Lustre File Systems using **lustre\_util**

To install lustre file systems, the following tasks must be performed:

- 1 Use **lustre\_util install** command to install the file system.
- 2 Use **lustre\_util start** command to enable the file system.
- 3 Use **lustre\_util mount** command to mount file systems on client nodes.

### 4.6.2 Removing Lustre File Systems using **lustre\_util**

To uninstall lustre filesystems, the following tasks must be performed:

- 1 Use **lustre\_util umount** command to unmount file systems on client nodes.
- 2 Use **lustre\_util stop** command to disable the file systems.
- 3 Use **lustre\_util remove** command to remove the file system.

### 4.6.3 **lustre\_util** Actions and Options

#### SYNOPSIS

```
lustre_util set_cfg [-n <l/O nodes list > | -p <l/O nodes rms partition > ]
```

```
lustre_util install -f < lmf or xmf path > [ --kfeof ] [ --lconf <option>]
```

```
lustre_util update -f < lmf or xmf path > [ --kfeof ] [ --lconf <option>]
```

```
lustre_util fsck -f < fs_name | all >
```

```
lustre_util chk_dev -f < lmf,xmf,xml files or fs_name | all >
```

```
lustre_util rescue -f < fs_name | all >
```

```
lustre_util start -f < fs_name | all > [ --lconf <option>]
```

```
lustre_util tune_servers -f < fs_name | all >
```

```
lustre_util mount -f < fs_name | all > -n <nodes|recover|all> | -p <rms_partition>  
--mount <[+]opt1,opt2,...>
```

```
lustre_util umount -f < fs_name | all > -n <nodes|all> | -p <rms_partition>
```

```

lustre_util status [ -f < fs_name | all > ] [ -n <nodes|all> | -p <rms_partition> ]
lustre_util fs_status [ -f < fs_name | all > ]
lustre_util mnt_status [ -f < fs_name | all > ] [ -n <nodes|all> | -p <rms_partition> ]
lustre_util stop -f < fs_name | all > [ --lconf <option>]
lustre_util remove -f < fs_name | all >
lustre_util info -f < lmf,xmf,xml files or fs_name | all >
lustre_util short_info -f < lmf,xmf,xml files or fs_name | all >
lustre_util lfsck -f < fs_name | all > -n <node> -d <shared_directory>
lustre_util build_mdt_db -f < fs_name | all > -n <node> -d <directory>
lustre_util build_ost_db -f < fs_name | all > -n <node> -d <directory>
lustre_util distribute_coherency -f < fs_name | all > -n <node> -d <directory>
lustre_util set_ost_rank -f < xml file >
lustre_util check_storage
lustre_util show_tuning
lustre_util show_cfg
lustre_util show_conf
lustre_util list

```

## ACTIONS

<b>set_cfg</b>	Copies <code>/etc/lustre/lustre.cfg</code> to OSS and MDS nodes.
<b>install</b>	Checks if filesystems can be installed, and then install them.
<b>update</b>	Updates settings of an installed filesystem that do not require reformatting of <b>previously</b> formatted OSTs/MDT (New OSTs, different network type, etc).
<b>fsck</b>	Runs <code>e2fsck</code> on the OST/MDT. The filesystem must be offline.
<b>chk_dev</b>	Check the devices and their links on I/O nodes.
<b>rescue</b>	Makes a filesystem usable again by formatting OSTs that are assumed to be NOT correctly formatted.

<b>tune_servers</b>	Set the I/O schedulers of filesystems devices regarding <b>lustre.cfg</b> content. Apply the server related tunings of <b>tuning.conf</b> on OSS/MDS.
<b>start</b>	Makes installed filesystems available for mounting.
<b>mount</b>	Mounts filesystems on specified nodes.
<b>umount</b>	Unmounts filesystems on specified nodes.
<b>fs_status</b>	Updates and prints OSS and MDS status information.
<b>mnt_status</b>	Updates and prints information regarding client nodes.
<b>status</b>	Does <b>fs_status</b> AND <b>mnt_status</b> .
<b>stop</b>	Disables filesystems.
<b>remove</b>	Removes filesystems.
<b>info</b>	Prints information (mdt,ost,path,etc.) about filesystems.
<b>short_info</b>	Prints information (mdt,ost,path,etc.) about filesystems, but sort OSTs by nodes.
<b>lfscck</b>	Builds mdt,osts lfscck database and distributes coherency checking of a Lustre filesystem.
<b>build_mdt_db</b>	Build mdt lfscck database (first step of lfscck).
<b>build_ost_db</b>	Build osts lfscck database (second step of lfscck).
<b>distribute_coherency</b>	Distributes coherency checking of a Lustre filesystem (third step of lfscck).
<b>set_ost_rank</b>	Set the rank of the OSTs in the cluster database regarding the rank they have in the XML.
<b>check_storage</b>	Check the consistency of the storage.conf or tables lustre_ost/lustre_mdt.
<b>show_tuning</b>	Display tuning parameter (from tuning.conf).
<b>show_cfg</b>	Display lustre.cfg variable.
<b>show_conf</b>	Display the lustre_util parameter.
<b>list</b>	Prints name of installed filesystems or those filesystems which failed to be installed.

## OPTIONS

<b>-f &lt; filesystem path &gt;</b>	Filesystems to work with, this can be: <ul style="list-style-type: none"><li>- For an install and update: Filesystem path <b>MUST</b> lead to an lmf file (lustre model file) or an xmf file (extended model file).</li><li>- For other operations that require the -f option, the filesystem path <b>MUST</b> be only the name of an installed (or attempted to be installed) filesystem.</li><li>- "all" stands for all installed filesystems.</li></ul>
<b>-n &lt; nodes &gt;</b>	<b>-n &lt;nodes_list&gt;</b> <p>Applies the command to the nodes list using pdsh syntax: name[x-y,z],...,namek.</p> <b>-n all.</b> <p>For mount, stands for "all clients which have mounted this fs as least one time". For umount, stands for "all clients which currently mount this fs".</p> <b>-n recover .</b> <p>For mount, stands for "all clients which are assumed to mount this fs". The main purpose of recover is to mount Lustre clients after a cluster emergency stop (main failure). Clients will be mounted with the same options as their previous mount.</p>
<b>--mount &lt;[+]opt1,opt2,...&gt;</b>	This allows mount options to be specified. For example, if +bar is specified for a filesystem which has foo as a default mount option, mount will be run on the client with -o foo,bar options. If only bar is specified (without +), mount will be run with -o bar options.
<b>--lconf &lt;options&gt;</b>	This provides additional options to lconf for install, update, rescue, start and stop to be used.
<b>-p &lt;rms_partition&gt;</b>	Applies the command to the configured nodes of this running rms partition.
<b>-F</b>	Forces commands execution even though this may be dangerous (no user acknowledgement is asked for).
<b>-t &lt;time_in_second&gt;</b>	Sets the limit on the amount of time a command is allowed to execute. 0 means no timeout.
<b>-u &lt;user&gt;</b>	User name to use to log onto the nodes instead of root.
<b>--fanout</b>	Number of ssh connections allowed to run at the same time, default is 128.

- kfeof**                      Stands for "keep formatting even on failure". Without this option, `lustre_util` returns as soon as the first failure is detected while formatting. With this option, `lustre_util` returns only when all the formatting attempts return for all devices. This can be useful when formatting a large pool of devices. This way you can check the health of all the devices in one shot, and you do not have to reformat devices that succeed in being formatted in a previous pass (using `lustre_util update`).
- V**                              Verbose output.
- v**                              Print version and exits.
- h**                              Print help and exits.

### **set\_cfg: Distributing /etc/lustre/lustre.cfg**

This file must be copied on every OSS and MDS nodes. You can do it using the `set_cfg` command:

```
lustre_util set_cfg [ -n <I/O nodes list > | -p <I/O nodes rms partition> ]
```

If no node parameter is specified, this command copies `/etc/lustre/lustre.cfg` of the Management Node on the nodes that host OST and/or MDT. If nodes are specified, `lustre.cfg` will be only copied on those nodes. If `LUSTRE_SNMP` is set to "yes", and if the variable `disable_chkconfig_for_ldap = no`, snmp server will be enabled on (selected) I/O nodes. If `LUSTRE_LDAP_URL` is set to a server address, this server will be enabled.

### **info: Printing Information about Filesystem**

```
lustre_util info -f < lmf, xmf, xml files or fs_name | all >
```

This command will print information about the file system descriptor you specify. If you specify only a file system name, this fs must be installed and information will be retrieved from the cluster database.

### **short\_info: Printing Information about a Filesystem**

```
lustre_util short_info -f < lmf, xmf, xml files or fs_name | all >
```

Same purpose as the previous command but displays OST sorted by nodes.

### **install: Installing a lustre Filesystem**

```
lustre_util install -f <lmf or xmf path> -V [ --kfeof ] [ --lconf <options>]
```

This command formats the storage devices and performs operations required to install the filesystem such as loading filesystems information into **ldap** database and/or cluster database. If **-F** is used, no user acknowledge is required. If **-F** is not specified, user must enter "yes" to go on if a filesystem with the same name is already installed. xml files required by CFS Lustre tools (lconf) are automatically generated and copied where it is required. An xmf file is also automatically generated in LUSTRE\_CONFIG\_DIR for each filesystem.



**Note:**

This operation is quite long, **-V** (be verbose) option is recommended.

### start: Enabling a Lustre Filesystem

```
lustre_util start -f fs_name -V [ --lconf <option>]
```

This command enables a filesystem and makes it available for mounting (online). Use of **-V** option (be verbose) is recommended.

### mount: Mounting Lustre Filesystem

```
lustre_util mount -f fs_name -n <nodes|all|recover> | -p <rms_partition>  
[ --mount < [+]options> ]
```

This command will mount the filesystem on specified nodes using the mount-path defined in the model file. If this mount-path does not exist, it is automatically created. It is an error if this path is already used to mount another filesystem. If **--mount** is not specified, fs will be mounted with options defined in model file by `mount_options`. If you use **--mount** with a parameter which starts with **+**, fs will be mounted with default options AND with those you give to **--mount**. If the parameter does not start with **+**, fs will be mounted with only those you give to **--mount**.

### umount: Unmounting Lustre Filesystem

```
lustre_util umount -f fs_name -n <nodes|all> | -p <rms_partition>
```

This command unmounts the filesystem on specified nodes. You can use the **-n all** option if you want to unmount the filesystem everywhere it is mounted. If `umount` fails because some processes have their working directories in the mount-path, use `umount` again with **-F** option, in order to kill such processes before the `umount` operation.

### stop: Disabling a Lustre Filesystem

```
lustre_util stop -f fs_name [ --lconf <option>]
```

This command disables a filesystem. It will not be available for mounting any more (offline).



## set\_iosched: Set the I/O Schedulers of Filesystem Devices

```
lustre_util set_iosched -f < fs_name | all >
```

The main purpose of **set\_iosched** is to be used as call-back when migration occurs and to set the I/O schedulers on the nodes where lustre services are restarted. You do not have to use it directly as **lustre\_util start** sets the I/O schedulers automatically.

## remove: Removing a Lustre Filesystem

```
lustre_util remove -f fs_name
```

This command totally removes the file system. All data will be lost. If **-F** is used, the action is done directly without any need of a user acknowledgement. If **-F** is not used, the user is prompted and must answer explicitly "yes".

## fs\_status: Updating Filesystem Status and Printing Filesystem Information regarding OSS and MDS

```
lustre_util fs_status [ -f fs_name ]
```

This command updates the status of OSTs, MDTs, and filesystems. If no filesystem parameters are provided, all installed filesystems are checked. The output appears as follows:

### FILESYSTEM STATUS

filesystem	Config status	Running status	Number of	clts migration
tv2ost8	installed	offline	0	0 OSTs migrated
tv2fs1	installed	online	3	0 OSTs migrated
tv2fs2	installed	online	4	0 OSTs migrated

The **config status** can take one of the following values:

<b>not installed</b>	<b>fs</b> is not installed (it should never be visible).
<b>loaded but not installed</b>	<b>fs</b> information is in a database but <b>lustre_util install</b> failed.
<b>Formatting</b>	<b>lustre_util install</b> is running.
<b>checking</b>	<b>lustre_util fsck</b> is running.
<b>installed:</b>	<b>fs</b> is correctly installed.
<b>not usable:</b>	<b>fs</b> is not correctly installed, because some devices failed to be formatted or <b>fsck</b> failed to repair some devices.

The **Running status** can take one of the following values:

<b>offline</b>	<b>fs</b> is correctly stopped.
<b>Starting</b>	<b>lustre_util start</b> is running.

<b>online</b>	<b>fs</b> is started and OSTs/MDT are healthy.
<b>not correctly started</b>	<b>fs</b> failed to start, some OSTs or MDT may be offline or unhealthy.
<b>CRITICAL</b>	<b>fs</b> started, but for unknown reasons, some OSTs or MDT may be offline or unhealthy.
<b>WARNING</b>	<b>fs</b> is started, but OSS or MDS may not be reachable and their states cannot be checked (elan can work).
<b>stopping</b>	<b>lustre_util stop</b> is running.
<b>not correctly stopped:</b>	<b>fs</b> failed to stop, some OSTs or MDT are still online or are in an unhealthy state.

### **mnt\_status: Updating Clients Status and printing Filesystem Information regarding Clients**

```
lustre_util mnt_status [ -f fs_name ] [-n <nodes|all> | -p <rms_partition> ]
```

This command checks if the filesystem is correctly mounted or unmounted on specified nodes. If no node is specified, **mnt\_status** gives the status of all client nodes that work with this filesystem. If no filesystem parameter is provided, all installed file systems are checked. The output looks similar to the following:

CLIENT STATUS

filesystem	Correctly mounted	should be mounted but are not	Correctly unmounted
tv2ost8	None		tv5
tv2fs1	tv[0-2]		
tv2fs2	tv[0-2,5]	tv[3-4]	

### **status: Updating Status of Servers and Clients, printing Filesystem Information regarding Servers and Clients**

```
lustre_util status [ -f fs_name ] [ -n <nodes|all> | -p <rms_partition> ]
```

This command performs a **fs\_status** AND a **mnt\_status** operation.

### **fsck: running e2fsck on OSTs and MDT**

```
lustre_util fsck -f fs_name [ --lconf <option>]
```

This command runs e2fsck on OSTs and MDT. It reports if some devices have been repaired, if some nodes need to be rebooted, and also if some devices have unrecoverable errors. This command should be applied to offline file systems.

### **chk\_dev: Check Devices and their links on I/O Nodes**

```
lustre_util chk_dev -f < lmf,xfm,xml files or fs_name | all >
```

This command checks devices information on filesystems I/O nodes:

- If the device exists.
- If the device is managed by **stormap**, it checks if device is up or down.
- If size in MBs is the expected size.

### lfck: Builds mdt,osts lfck Database and distributes Coherency Checking of a Lustre Filesystem

```
lustre_util lfck -f < fs_name | all > -n <node> -d <shared_directory>
```

**<node>** is a client which can mount the filesystem, but the **fs MUST NOT** be mounted when you start to use **lfck**.

**<shared\_directory>** is a shared directory where the **lfck** database files will be placed. The I/O nodes and the client node must have read/write access to this directory using the same path.



#### Note:

The database **lfck** files can be large, depending on the number of files in the filesystem (10GB or more for millions of files), so ensure there is enough space in the shared directory before using **lfck**.

**lfck** is to be used **ONLY** when unrecoverable errors have been found on OST devices or when OSTs have been reformatted. It attempts to correct problems such as:

- Inode exists but has missing objects = dangling inode. This normally happens if there was a problem with an OST.
- Inode is missing but OST has unreferenced objects = orphan object. This normally happens if there was a problem with the MDS
- Multiple inodes reference the same objects. This can happen if there was corruption on the MDS, or if the MDS storage is cached and loses some but not all of its writes.

After using **lustre\_util lfck**, you should check **lost+found** in the mountpoint of client.

Using **lfck** is the same as using **build\_mdt\_db**, followed by **build\_ost\_db**, and then **distribute\_coherency**.

### build\_mdt\_db, build\_ost\_db, distribute\_coherency : step by step lfck

```
lustre_util build_mdt_db -f < fs_name | all > -n <node> -d <directory>
lustre_util build_ost_db -f < fs_name | all > -n <node> -d <directory>
lustre_util distribute_coherency -f < fs_name | all > -n <node> -d <directory>
```

These options are to be used:

To restart an **lfck** operation which has failed, avoiding the need to restart the process from the beginning. **Lustre\_util** will provide information regarding which options should be used and when.

If the directory is not a shared directory and there is a need to copy database files, `lustre_util` will provide information regarding which files should be copied and where.

These operations should be done in the following order: `build_mdt_db`, then `build_ost_db`, and then `distribute_coherency`.

### update: Update Filesystems already Installed

```
lustre_util update -f fs_name -V
```

This command allows you to update an **ALREADY INSTALLED** and offline filesystem with new settings (that do not require a reformatting of the **ALREADY FORMATED** devices):

- `stripe_count`
- `nettype`
- `generic_client`
- `failover`
- `mdt_mount_options`
- `ost_mount_options`
- `cluster_id`
- `mount_path`
- `quota`
- `quota_options`
- `description`
- `mount_options`
- `ost` (new OST can be added, previous OSTs must also be included and do not forget that their `cfg_status` should be currently "formatted". OSTs that currently have their `cfg_status` set to "format\_failed" may be removed).

Update is done by updating the model file or the corresponding extended model file with the new settings. The following settings **MUST** be the same:

- `mdt`(mdt line of model file must lead to the same mdt, do not forget that the `cfg_status` of the mdt should be currently "formatted" )
- `ost` that were previously part of the filesystem and that currently have their `cfg_status` set to "formatted".
- `mdt_mkfs_options`
- `mdt_inode_size`
- `ost_mkfs_options`
- `ost_inode_size`
- `fs_name`



#### Important:

An update operation should only be done on a file system which has been stopped correctly.

If High Availability is in use and if the OSTs are distributed on 2 OSSs that are mutually the failover node of each other then all OSTs must be on their primary location otherwise the update will take a long time.

Once the model file is updated, run:

```
lustre_util update -f <path to modified lmf/xmf>.
```

New OSTs will be formatted, new automatically generated xmf/xml will be copied to the right place, and mdt will be updated (write\_conf). Only OSTs that have their `cfg_status` set to "format\_failed" before the update may be removed.



#### Important:

Removing correctly formatted OSTs of a filesystem can cause data loss, Lustre\_util will not allow this to be done.

Update can also be used after the installation of a new release of Lustre, if the underlying way of storing information on the **MDT** has changed.

#### rescue: Try to make the Installed Filesystem Work again

```
lustre_util rescue -f fs_name -V [ --lconf <option>]
```

This command can be used on installed file systems which have stopped:

- If the update failed (may be because new OSTs cannot be formatted)
- If fsck detects devices with unrecoverable errors
- If you lose the XML files
- Or for other reasons.

This command checks which OSTs have been successfully formatted and formats those that are assumed to be not correctly formatted. A new XML file is generated and copied in the right places and the MDT is updated (write\_conf). Theoretically, the file system should be usable again, **but data may be lost**.

#### set\_ost\_rank : set OST rank in Cluster Database regarding the XML file

```
lustre_util set_ost_rank -f < xml file >
```

This option is only used if there has been an upgrade from a release of cluster database that does not have rank field in the `lustre_ost` table to a release that includes this field. The new field is updated with OST information found in the XML file.

#### check\_storage : Checking Consistency of storage.conf or lustre\_ost/lustre\_mdt Tables

```
lustre_util check_storage
```

The main purpose of this option is to check if **storage.conf** has been correctly completed by the administrator. It should not be necessary to use this if a cluster database is used, however, this option can be available if required.

#### show\_tuning: Display the Tuning Parameters

```
lustre_util show_tuning
```

Display the tuning parameters according to the content of **/etc/lustre/tuning.conf**.

#### show\_cfg: Display lustre.cfg Variable

```
lustre_util show_cfg
```

Display lustre.cfg variable.

#### show\_conf: Display lustre\_util Configuration

```
lustre_util show_conf
```

Display lustre\_util configuration, according to the content of **/etc/lustre/lustre\_util.conf**.

#### list: Gives the List of Installed Filesystems

```
lustre_util list
```

This command prints the name of the filesystems which are installed, even if their installation is not yet complete.



#### Note:

An example of the complete process to create and install a Lustre file system is described in Bull HPC BAS4 for Xeon *Installation and Configuration Guide*.

## 4.6.4 lustre\_util Configuration File /etc/lustre/lustre\_util.conf

This file contains some additional settings for **lustre\_util**. The following values are set by default:

```
ssh_connect_timeout=20
```

This is the timeout in seconds given to the **connect\_timeout** parameter of SSH.

```
install_timeout=0
```

This is the timeout in seconds for install, update and rescue operations and can be overwritten by the **-t** option.

```
start_timeout=0
```

Timeout in **s** for the start operation and can be overwritten by the **-t** option.

```
mount_timeout=60
```

Timeout in **s** for the mount operation and can be overwritten by the **-t** option.

```
umount_timeout=60
```

Timeout in **s** for the umount operation and can be overwritten by **-t** option.

```
stop_timeout=0
```

Timeout in **s** for the stop operation and can be overwritten by the **-t** option.

```
status_timeout=30
```

Timeout in **s** for **status**, **fs\_status**, **mnt\_status** operation and can be overwritten by the **-t** option.

```
set_ioscheds_timeout=60
```

Timeout in **s** for setting I/O schedulers on I/O nodes (in start and tune\_servers operation), can be overloaded by **-t** option.

```
set_tuning_timeout=60
```

Timeout in **s** for applying tuning parameters on I/O nodes (in start,tune\_servers and mount operation), can be overloaded by **-t** option.

```
disable_nagios=no [yes]
```

**yes** will disable the update of the nagios pipe by **lustre\_util**.

```
disable_chkconfig_for_ldap=yes [no]
```

**yes** will disable the **chkconfig** of **ldap** service in the **set\_cfg** operation, **no** will allow this operation. It should be set to **yes** if administration node is an HA node.

```
use_stormap_for_chk_dev=yes [no]
```

If **yes**, **lustre\_util** will check health of devices using **stormap -l**. It should only be set to **no** if **stormap** is not installed on I/O nodes. It is not a problem if devices you are using are not managed by **stormap**.

```
allow_loop_devices=no [yes]
```

Unless you explicitly want to use loop device, this should be set to **no**. This way, it prevents **lconf** to create huge loop devices in **/dev/** directory when some LUNS disappear.

```
check_only_mounted_nodes_on_mnt_status=no [yes]
```

If set to **yes**, only nodes that are assumed to mount a filesystem will be checked on **status** and **mnt\_status** operation.

```
default_fanout=128
```

Number of **ssh** connexions allowed to run at the same time. Can be overloaded using **-fanout** option.

## 4.6.5 Lustre Tuning File `/etc/lustre/tuning.conf`

This file contains tuning parameters. The syntax is the following:

**"<string>" <file> <target> [<delay>] [<filesystems>]**

<b>"&lt;string&gt;"</b>	String to write in file, it can contain spaces, MUST be between double-quotes
<b>&lt;file&gt;</b>	Full path to the file where string will be written. Globbing is allowed. 2 macros can be used: <b>\${mdt}</b> stands for the name of the <b>mdt</b> of the filesystem. <b>\${ost}</b> stands for the name of ALL the <b>osts</b> (one line will be generated for each ost).
<b>&lt;target&gt;</b>	A string composed of the <b>OSS</b> , <b>MDS</b> , or <b>CLT</b> , separated by semicolons. <b>OSS</b> , <b>MDS</b> and <b>CLT</b> can be followed by a nodes list ( <b>pdsh syntax</b> ) using colon.
<b>&lt;delay&gt;</b>	A time in ms that we have to wait before continuing setting tuning parameters on a node. This is an optional argument, and the default is 0 ms.
<b>&lt;filesystem&gt;</b>	A list of filesystem separated with semicolons. This is an optional argument, and the default is to allow this tuning for every filesystems.

For OSS and MDS, tuning parameters are set when a filesystem is started. For Clients, tuning parameters are set when the filesystem is mounted, for example:

- **"1" /proc/sys/lnet/panic\_on\_lbug OSS;MDS;CLT**  
This line will enable panic on **lbug** on **ALL** types of node for all filesystems by running **echo "1" >/proc/sys/lnet/panic\_on\_lbug** on all nodes.
- **"0" /proc/sys/lnet/panic\_on\_lbug OSS:ns[5-6];MDS:ns3 fs1;fs2**  
This line will disable panic on **lbug**:
  - on ns5 and ns6, if they are used as an OSS of **fs1** and/or **fs2**,
  - on ns3, if it is used as MDS of **fs1** and/or **fs2**.

String, file and target can be aliased using the following syntax:

**alias <name>=<content>**

**alias** can be declared anywhere in the file, but it also acts on the **WHOLE** file, not only on the lines that follow the declaration.

When you use **alias** on a string, the alias must also be in double quotes.



### Example:

A `tuning.conf` example file is shown below:

```
#### ALIAS DECLARATION #####

alias health_check=/proc/fs/lustre/health_check
alias panic_on_lbug=/proc/sys/lnet/panic_on_lbug
alias ping_osc=/proc/fs/lustre/osc/*${ost}*/ping
alias debug=/proc/sys/lnet/debug

#### TUNING PARAMETER #####

"1"                ping_osc                CLT
"0"                panic_on_lbug            CLT
"0"                panic_on_lbug            OSS;MDS
"524288"           debug                    OSS;MDS;CLT
```

## 4.6.6 Lustre Filesystem Reconfiguration

This procedure allows you to change the distribution of the Lustre services which are defined on the I/O nodes, without having to re-deploy (which involves configuring the DDN storage systems and High Availability). The filesystems involved in the new distribution are stopped; the others continue to be operational.

The following example describes how to stop the `fs1` and `fs2` filesystems.

1. If needed save the data of the `fs1` and `fs2` filesystems.
2. Unmount the `fs1` and `fs2` filesystems:

```
lustre_util umount -f fs1 -n all [-F]
lustre_util umount -f fs2 -n all [-F]
```

3. Stop the `fs1` and `fs2` filesystems:

```
lustre_util stop -f fs1 [-F]
lustre_util stop -f fs2 [-F]
```

4. Remove the `fs1` and `fs2` filesystems:

```
lustre_util remove -f fs1
lustre_util remove -f fs2
```

5. Make the required modifications in the models associated with the filesystems. In our example `fs1` and `fs2` are grouped together in only one `fs3` filesystem.
6. Configure the new `fs3` filesystem (this operation erases the `fs1` and `fs2` filesystems data).

```
lustre_util install -f /etc/lustre/model/fs3.lmf
```

7. Start the new fs3 filesystem:

```
lustre_util start -f fs3
```

8. Mount the new fs3 filesystem:

```
lustre_util mount -f fs3 -p p2
```

9. If needed, restore the saved data.

## 4.6.7 Using Quotas with Lustre File Systems

### 4.6.7.1 Quota Settings in Model Files

Quotas are enabled by setting "quota" to "yes" in lmf file:

```
quota: yes
```

The default quota options are as follows:

```
quota_options: quotaon=ug,iunit=5000,bunit=100,itune=50,btune=50
```

<b>quotaon=&lt;u g ug&gt;</b>	Enable quota for user group user and group.
<b>iunit=&lt;number of inodes&gt;</b>	<b>iunit</b> is the granularity of inodes quotas. Inodes are acquired and released by a slice of iunit. iunit is a int type (>0), the default value in Lustre is 5000 inodes.
<b>bunit=&lt;size in MB&gt;</b>	<b>bunit</b> is the granularity of block quotas. Blocks are acquired and released by a slice of bunit MBs on each OSTs. bunit is expressed in MBs (>0), the default value in Lustre is 100 MBs.
<b>itune=&lt;percentage&gt;</b>	<b>itune</b> sets the threshold to release and acquire iunit inodes. For example, if a user/group owns $n*iunit+m$ inodes, iunit inodes will be acquired for this user as soon as $m$ goes above $itune*iunit/100$ . If a user/group owns $n*iunit-m$ inodes, iunit inodes will be released for this user/group as soon as $m$ goes above $itune*iunit/100$ . itune is a int type ( $100 > itune > 0$ ), the default value in Lustre is 50.
<b>btune=&lt;percentage&gt;</b>	<b>btune</b> sets the threshold to release and acquire bunit block MBs for each OST. For instance, if a user/group owns $n*bunit+m$ MB on one OST, bunit MBs will be acquired on this OST for this user/group as soon as $m$ goes above $btune*bunit/100$ . If a user/group owns $n*bunit-m$ MBs on one OST, bunit MBs will be released on this OST for this user/group as soon as $m$ goes above $btune*bunit/100$ MB. btune is a int type ( $100 > btune > 0$ ), the default value in Lustre is 50.

### 4.6.7.2 Starting Quota: lfs Quotacheck

Once the filesystem is installed, started and mounted, run the following command on a client:

```
lfs quotacheck -<quotaon parameter> <mount_point>
```

This means that if `quota_options` are as follows:

```
quotaon=ug,iunit=5000,bunit=100,itune=50,btune=50 and mountpoint  
is /mnt/lustre,
```

it will be necessary to run:

```
lfs quotacheck -ug /mnt/lustre
```

The time taken by `quotacheck` depends on the size of the biggest device used by the filesystem as OST or MDT. On average, it takes 160s for a 1TB OST/MDT check.

### 4.6.7.3 Setting the Limits: lfs Setquota

`lfs setquota` sets limits on blocks and files.

```
lfs setquota [-u|-g] <name> <block-softlimit> <block-hardlimit> <inode-softlimit>  
<inode-hardlimit> <mount_point>
```

`block-softlimit` and `block-hardlimit` are expressed in kB.

`inode-softlimit` and `inode-hardlimit` are expressed in number of inodes.

Limits on blocks/inodes MUST be greater than `bunit/iunit`. This means, for example, `bunit=100MB`, `block-softlimit` and `block-hardlimit` must be greater than 102400kB. If you have `iunit=5000`, `inode-softlimit` and `inode-hardlimit` must be greater than 5000.

Limits on blocks must be greater than the number of OST \* `bunit`. This means, for example, if there are 9 OSTs and `bunit=100 MBs`, `block-softlimit` and `block-hardlimit` must be greater than  $9 * 100 * 1024 = 921600$  kB.

For example:

```
lfs setquota -u bob 900000 1000000 5000 10000 /mnt/lustre
```

will set a `block-softlimit` to 900MB, `block-hardlimit` to 1GB, `inode-softlimit` to 5000, `inode-hardlimit` to 10000 for user `testfs`, for a lustre filesystem mounted on `/mnt/lustre`.

```
lfs setquota -g dba 900000 1000000 5000 10000 /mnt/lustre
```

The command above will implement the same settings for all users of group `dba`.

#### Restrictions

- At present, soft limits are not supported in Lustre. So set `block-softlimit` and `inode-softlimit` to 0.

- It is strongly recommended to run **setquota** on a Lustre file system which is not busy. Otherwise an incorrect **block-hardlimit** value may be set.

#### 4.6.7.4 Updating/Rescuing a Filesystem with Quota enabled

If a filesystem is rescued, quota will have to be enabled again using the command below.

```
lfs quotacheck -<quotaon parameter> <mount_point>
```

If a filesystem is updated and new OSTs are not added the following command will have to be run again:

```
lfs quotacheck -<quotaon parameter> <mount_point>
```

If a filesystem is updated and **new OSTs are added** then the fs will have to be updated, started and mounted and then run the following command:

```
lfs quotacheck -<quotaon parameter> <mount_point>
```

For **\*ALL\*** groups and users, all the limits may be set to 0 with the following command:

```
lfs setquota -u <user> 0 0 0 0 <mount_point>
lfs setquota -g <group> 0 0 0 0 <mount_point>
```

For **\*ALL\*** groups and users, the limits may be set to their former values with the following command.

```
lfs setquota [-u|-g] <name> <block-softlimit> <block-hardlimit> <inode-softlimit> <inode-hardlimit> <mount_point>
```

## 4.7 Monitoring Lustre System

Status information about the Lustre file system and I/O nodes is kept up to date in the ClusterDB by the Lustre management tools.

Using this information and that collected by performance daemons, the NS-Master HPC Edition supervision tool offers items specific to the Lustre system allowing the health and performance to be monitored from the management station – see the chapter on monitoring for more details.

### 4.7.1 Lustre System Health Supervision

#### 4.7.1.1 The all status Map view

This includes global status indicators which provide the administrator with information about the global I/O system availability.

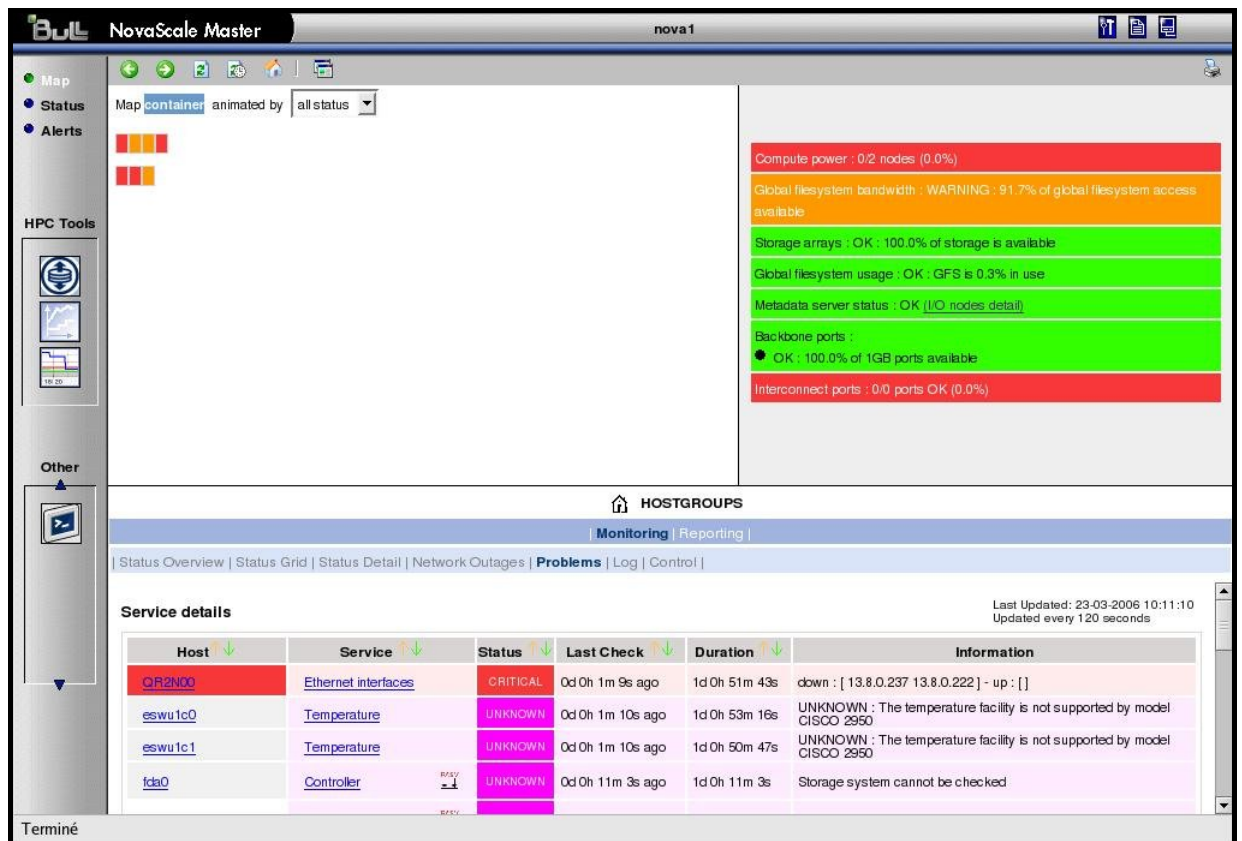


Figure 4-1. NovaScale Master Map view

System Availability Indicators are located at the right top of the topological view to give a status to the administrator at-a-glance. These include:

### Available Global File System Bandwidth as a Percentage

This is indicated as a percentage of I/O nodes available. An I/O node is fully available if it has its three Quadrics rails and its height fibre links up and if its Lustre status is OK. If not, a degradation factor is applied as follows:

- **cancel** the node if Lustre is not OK
- **apply** a 30% factor of degradation per quadrics rail missing
- **apply** a 12% factor of degradation per fibre link missing

### Available Storage Arrays as a Percentage

The ratio of running storage appliances (DDNs) against the total number is indicated.

### Global File System Usage

This gives the current usage rate of the Lustre system for all the Lustre file systems together.

### MDS Migration Alert

If High-Availability is configured, this alerts the administrator to a MDS failover migration. The Lustre system then no longer has the High-Availability status.

## 4.7.1.2 Filesystems Health Monitoring

This is done by the script `/usr/bin/lustre_fs_nagios`. It checks the state of each OSTs/MDTs, and sets the status of the filesystems into the ClusterDB according to whether they are online or offline. This script is called every 15 min on the Management Node using `/etc/cron.d/lustre_fs_nagios.cron`, which is automatically installed and enabled by `lustre_utils` RPM.

`lustre_fs_nagios` should not be used online by the administrator; however, it can be used to force a refresh of `nagios` lustre filesystem status entry.

## 4.7.1.3 The `lustre_check` Tool

The `lustre_check` tool keeps the I/O node availability information up to date in the ClusterDB. It runs on the management station, scheduled by a `cron` every 15 min.

When called, it checks the I/O nodes to collect the network and storage information. This information is stored for each node in the `lustre_io_node` table of the database where it is regularly scanned by the supervision tools.

The `lustre_check` tool is not likely to be used on line by the administrator; however, it can be used to force a refresh of the ClusterDB information and to get an instant status displayed node by node.

## 4.7.2 Lustre Filesystem Indicator

Within NovaScale Master the Nagios service plug-ins include a plug to monitor the health for the Lustre file system.

Service Name	Status	Last Check Time	Next Check Time	Message
Ethernet interfaces	OK	0d 0h 5m 47s ago	1d 0h 59m 57s	down : [] - up : [ 192.20.0.1 13.1.0.1 ]
Global filesystem bandwidth	WARNING	0d 0h 5m 47s ago	0d 0h 5m 47s	WARNING : 91.7% of global filesystem access available
Global filesystem usage	OK	0d 0h 5m 47s ago	1d 0h 55m 12s	OK : GFS is 0.3% in use
HA system status	WARNING	0d 0h 5m 47s ago	1d 0h 52m 49s	clustat could not connect to HA service, possible HA failure
Hardware status	UNKNOWN	0d 0h 0m 47s ago	1d 0h 59m 51s	domain unset for host nova0
IC switch manager	OK	0d 0h 5m 47s ago	1d 0h 57m 28s	OK - 1 processes running with command name swmgr
IO status	OK	1d 1h 3m 24s ago	1d 1h 3m 24s	IOSTAT: IO status details All IO devices are OK
Interconnect ports	CRITICAL	0d 0h 5m 47s ago	1d 0h 55m 5s	IO ports OK (0.0%)
Kerberos KDC daemon	CRITICAL	0d 0h 5m 47s ago	1d 0h 52m 43s	CRITICAL - 0 processes running with command name krb5kdc
Kerberos admin daemon	CRITICAL	0d 0h 5m 47s ago	1d 0h 59m 45s	CRITICAL - 0 processes running with command name kadmind
LDAP daemon	OK	0d 0h 5m 46s ago	0d 19h 7m 22s	LDAP ok - 0 seconds response time
Log alerts	PENDING	1d 1h 3m 28s+ ago	1d 1h 3m 28s+	Service is not scheduled to be checked
Lustre filesystems status	OK	0d 0h 0m 45s ago	0d 17h 27m 36s	(Details) OK - fs1 is installed and online
Metadata server status	OK	0d 0h 5m 47s ago	0d 0h 5m 47s	OK (IO nodes detail)
MiniSQL daemon	OK	0d 0h 5m 47s ago	1d 0h 52m 36s	OK - 1 processes running with command name msq13d
NSDoctor	PENDING	1d 1h 3m 28s+ ago	1d 1h 3m 28s+	Service is not scheduled to be checked...
Postboot checker	PENDING	1d 1h 3m 28s+ ago	1d 1h 3m 28s+	Service is not scheduled to be checked...
RMS daemon	OK	0d 0h 5m 47s ago	1d 0h 59m 39s	OK - 3 processes running with command name rmsd

Figure 4-2. NovaScale Nagios file system indicator

The Lustre file system indicator relates to the Lustre file systems health as a whole. Clicking on the info link will display a detailed status for each file system running.

### Lustre Management Node Web Interface

With a web browser, you can easily check the Lustre filesystem status using the following URL: <http://<management node>/lustre>

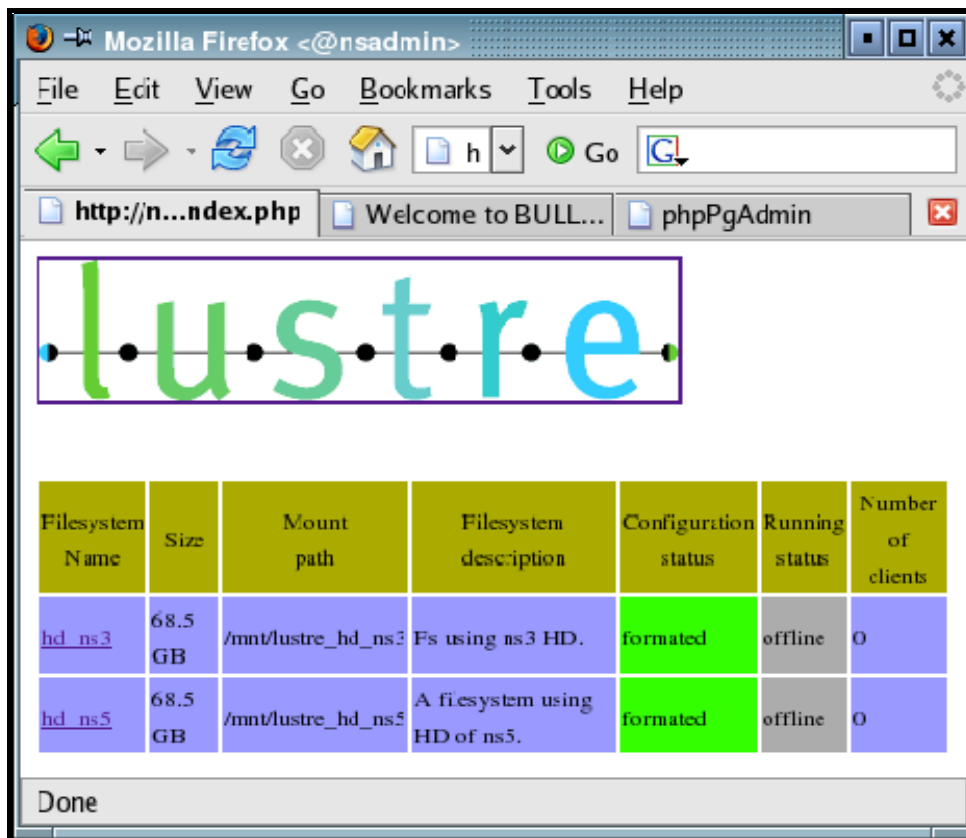


Figure 4-3. Lustre Management Node web interface

By clicking on the filesystem name, you can get details about the filesystem, using an interface that allows you to sort OSTs by name, active node, primary node, secondary node, device size, journal device, Config status, status or migration status.



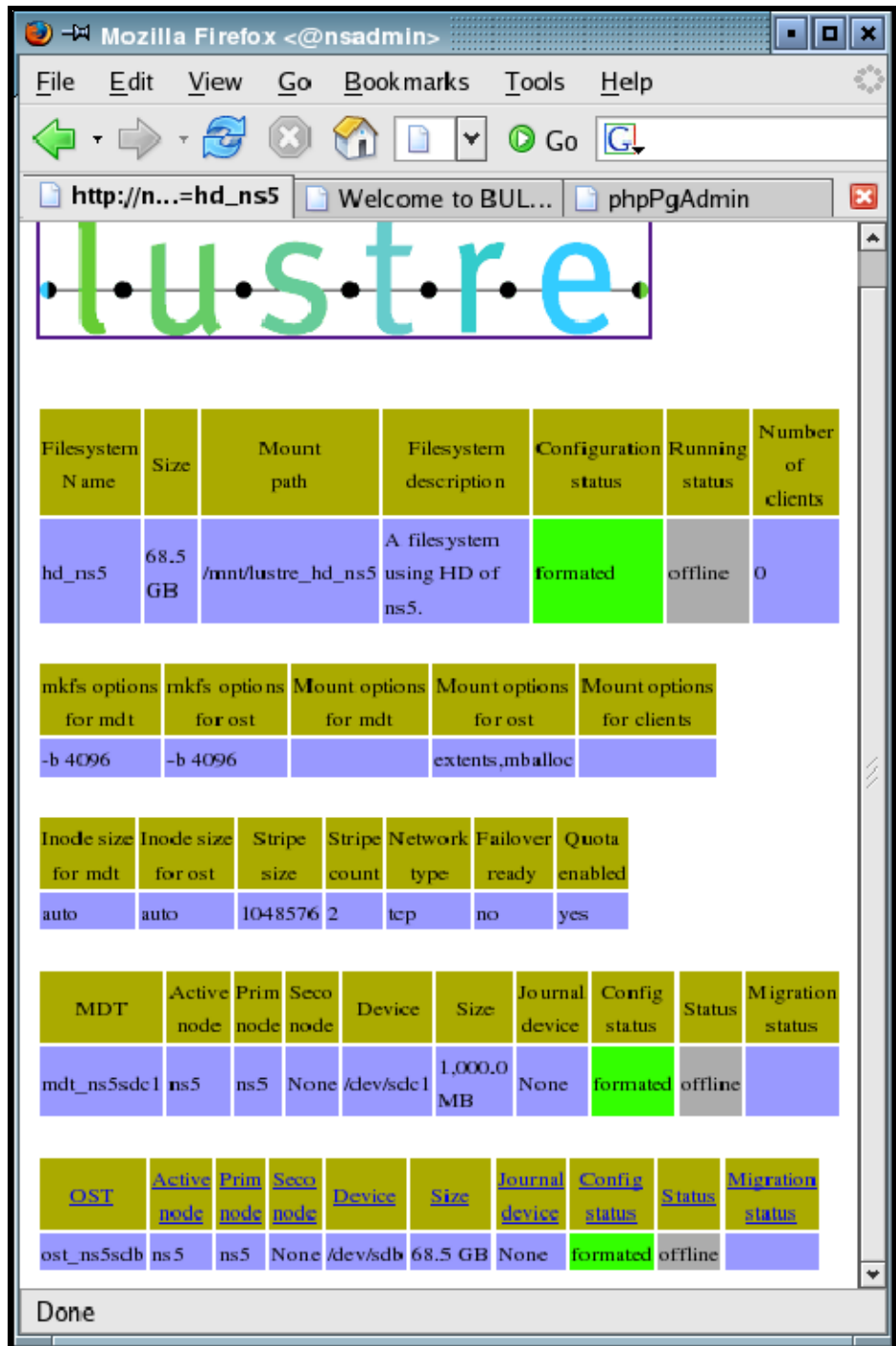


Figure 4-4. Detailed view of Lustre file systems

## 4.7.3 Lustre System Performance Supervision

### 4.7.3.1 Group Performance Views

By clicking on the Group performance button in the NovaScale Master console the administrator is provided with an at-a-glance view of the transfer rates of the Lustre system for the file systems all together. The information period can be specified.

Clicking on the compiled view will display a dispatched view giving the performance rates node by node for the same time period.

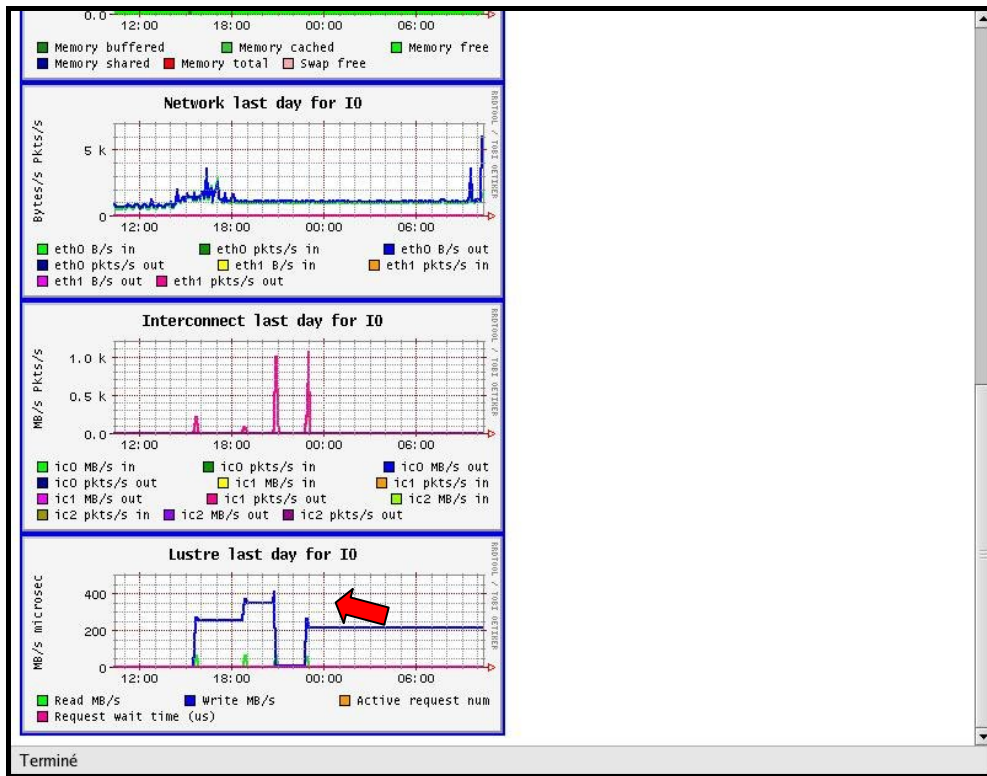


Figure 4-5. Group performance global view pop up window



Figure 4-6. Dispatched performance view pop up window

### 4.7.3.2 Node Performance Views

Views related to Lustre system local transfer and filling rates are available for each I/O node from the Global Performance view in the Nova Scale Master Console.

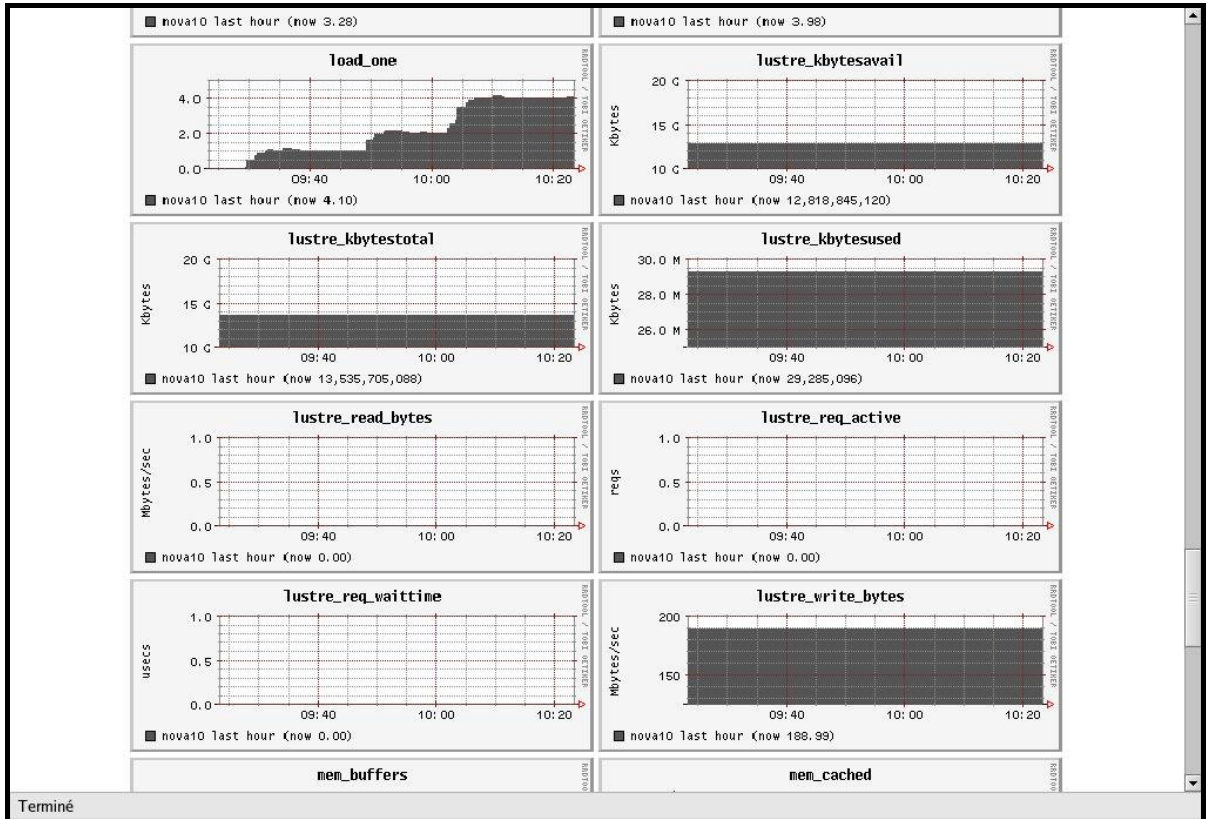


Figure 4-7. Global performance view pop up window



---

## Chapter 5. Software Deployment (KSIS)

This chapter describes how to use KSIS to deploy, manage, modify and check software images. The following topics are described:

- 5.1 Overview
- 5.2 Configuring and Verifying a Reference Node
- 5.3 Main Steps for Deployment
- 5.4 Modifying Images and Managing their Release
- 5.5 Checking Deployed Images
- 5.6 Importing and Exporting an Image
- 5.7 Ksis Commands
- 5.8 Modifying an Image
- 5.9 Checking Images
- 5.10 Importing and Exporting Images
- 5.11 Rebuilding ClusterDB Data before Deploying an Image

### 5.1 Overview

A deployment tool is a piece of software used to install a distribution and packages on several machines at once. For large clusters, such a tool is essential, since it avoids doing the same installation a large number of times. **KSIS** is the deployment tool used on Bull HPC systems.

KSIS makes it easy to propagate software distributions, content or data distribution changes, operating system and software updates, for a network of Linux machines. KSIS is used to ensure safe production deployments. By saving the current production image before updating it with the new production image, a highly reliable contingency mechanism is provided. If the new production environment is found to be flawed, simply roll-back to the last production image.

This chapter describes how to:

- Create an image for each type of node and save it on the image server. These images are called reference/golden images. The image server is on the Management Node and is operated by the KSIS server software.
- Deploy the node images.
- Manage the evolution of the images (**workon** images and patches).
- Check discrepancies between an image on a node and its reference on the image server.



**Note:**

The terms **reference node** and **golden node** are interchangeable. The same applies to the terms **reference image** and **golden image**.

The deployment is done using the administration network.

## 5.2 Configuring and Verifying a Reference Node

A reference node is a node which has had all the software installed on to it and from which the image is taken and stored on the image server. The reference image will be deployed onto the other nodes of the HPC system.

### Installation and Configuration

Reference nodes have the **BAS4 for Xeon** software installed on to them in the same way as ordinary compute or I/O nodes. A **KSIS client** is then installed onto these nodes using the **XHPC** CDROM. The operating system and applications must be installed and configured to make the node operational.

## 5.3 Main Steps for Deployment

Once the image server, reference nodes and client nodes are ready, the steps for the deployment are:

1. Create the image of the reference node to be saved on the Image Server:

```
ksis create <imageName> <ReferenceNodeName>
```

This command requests that a check level is chosen. Choose "basic".

2. Deploy the image:

```
ksis deploy <imageName> node[1-5] -P
```



### Note:

The **-P** option specifies that the deployment will be automatically followed by a post-configuration of the nodes. See *Deploying an Image or a Patch*, on page 5-12 for details.

The following figure shows the creation and deployment of an image.

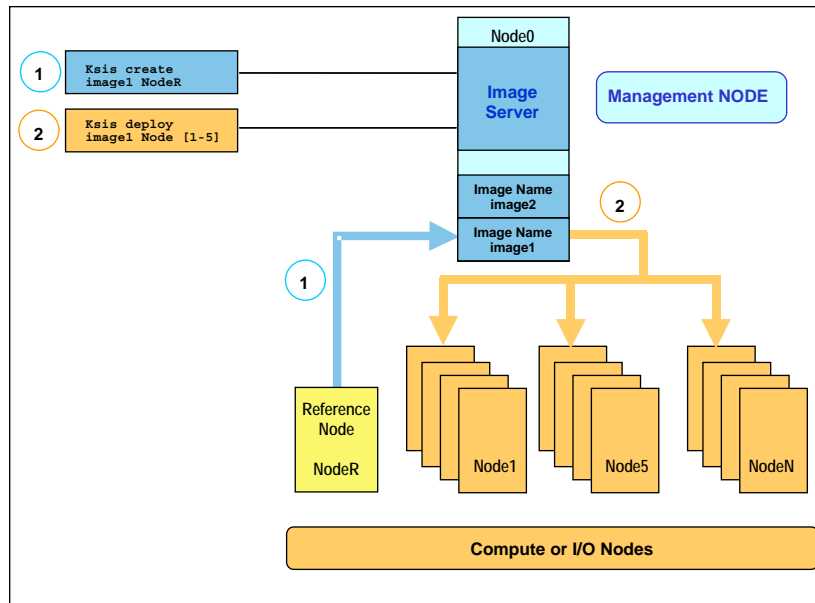


Figure 5-1. Main steps for deployment

## 5.4 Modifying Images and Managing their Release

### 5.4.1 Methods

There are two different methods to modify an image.

1. The image on the Reference Node can be modified and deployed as described in the previous section.



**Note:**

This method is safer and should be used when the modifications are complex or when there are a great number of invoked files.

2. Alternatively use the **workon** mechanism, which consists in directly altering an image on the image server; for instance modifying a configuration file. The **ksis workon** command means that it is possible to "log" on the image and to modify it. This command opens a working environment where the image can be modified using the shell provided.



**Note:**

Environmental modifications are limited to file modifications which are seen when working on a filesystem and not on the nodes which are in use.

When the modifications are finished create a patch containing the modifications using the **ksis store** command.

Example:

```
ksis workon image1
ksis store image1.s1.0
```

It is possible to deploy and apply this patch on the specified nodes without having to deploy the whole image. The **ksis deploy** command deploys the patch.

Example:

```
ksis deploy image1.s1.0 nc[2-45]
```

The **ksis undeploy** command is used to remove the last patch from the nodes specified. This only works if the node has not been completely altered by the patch. For example, **ssh** must be available on the node.

Example:

```
ksis undeploy image1.s1.0 nc[2-6]
```

It is possible to create a new complete image of the node with the applied patch using the **ksis detach** command.

The **ksis list** command creates a list of images and patches available on the image server with their status.

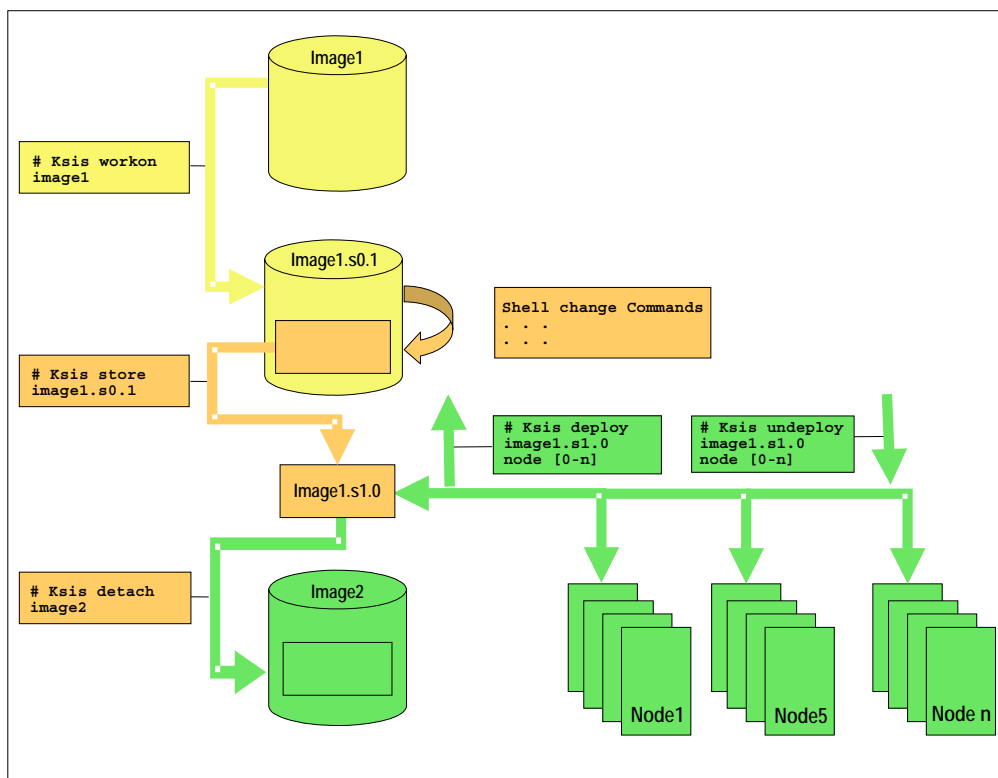


Figure 5-2. Image modification (workon, store, detach)



## 5.4.2 Naming Images or Patches with the Workon Mechanism

The image or patch derivation process follows these rules:

- Only one derivation is possible except at level 1.
- The user can define the name of the *golden* and *patched golden* images only at the time that they are created.
- The name of the patch and working patch image is the name of the mother image suffixed by *.sX.Y* where *X* refers to the patch branch number *X*, and *Y* refers to the patch number *Y* on this branch.

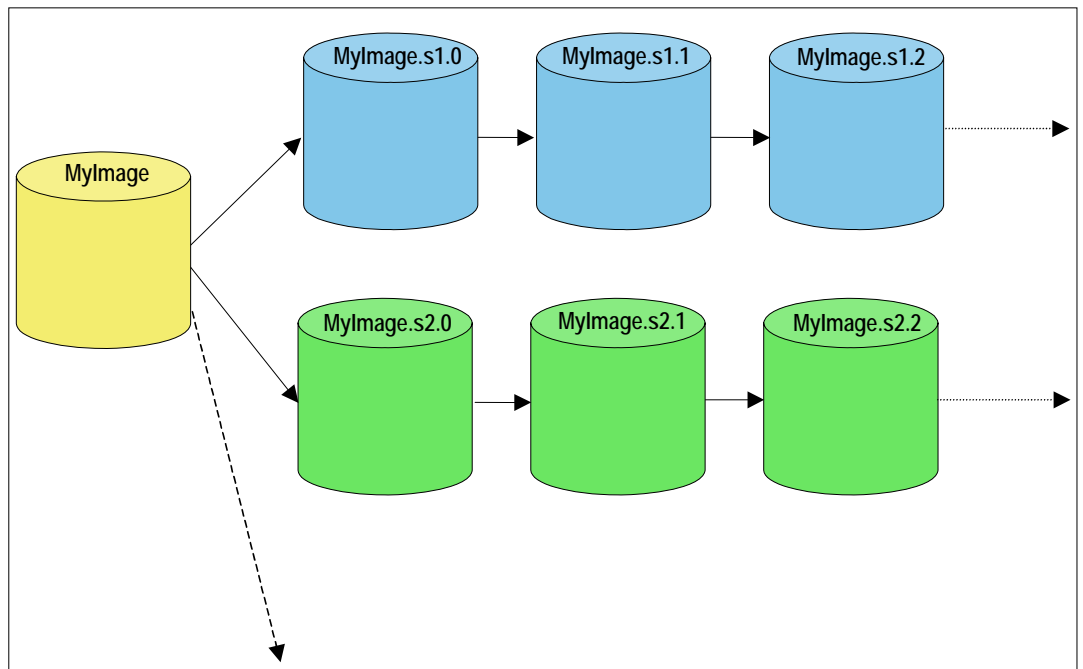


Figure 5-3. Names of derived images or patches

## 5.4.3 Image Types

There are four types of images (result of `ksis list` command):

<b>Golden</b>	Image obtained from a reference node.
<b>Patched golden</b>	Modified golden image (resulting from the <code>detach</code> command).
<b>Working patch</b>	Patch in progress on this image. Cannot be deployed. Waiting for <code>store</code> command.
<b>Patch</b>	Patch (result of a <code>store</code> command).

## 5.5 Checking Deployed Images

The **ksis check** command is used to compare the image deployed on a set of nodes with the corresponding reference image.

This is done by listing the discrepancies which exist for a series of tests performed on the nodes and the results of the same tests on the reference image.

Example:

```
ksis check nc[2-45]
```



### Note:

Nodes inside a node range are not always created from the same image.

### 5.5.1 Checking Principles

The descriptions of the image checks are stored in a database. When creating an image or a patch, the administrator specifies the required check level for this image or patch. Then **KSIS** copies the corresponding checking database to the image and executes the command associated with each check and stores the results as a *reference*. This *reference* is then included in the image.

Each time the **ksis check** command is used, KSIS executes the defined checks on each node and generates the results. If there is a discrepancy between the result and the *reference*, the check is set to KO, otherwise it is set to OK. The image server centralizes the results. In this way, the load for control is spread over the nodes. It is also easy to modify and to add new checks.

### 5.5.2 Check Groups

According to the chosen level, checks for a given image or patch are extracted from the checks database (`/etc/systemimager/ksis_check_Repository/` on the Management Node) and are executed when the image is created. A check level is a particular check group.

Each check belongs to one or more groups defined in the **group** file inside the check directory. If the `-t` option is not specified all the checks are executed.

The checks belonging to the **skip** group are not run.

Name	Level	Group	OK	KO
CheckRpmList	Basic	rpm, fastrpm	List of RPMs installed on the node is the same on the reference image.	List of RPMs installed on the node is not the same on the reference image.

Name	Level	Group	OK	KO
CheckRpmFiles	Sharp	rpm	None of the files delivered using the RPM seems to have been updated regarding contents and/or access rights.	One or more of the files delivered using RPM seem to have been updated regarding contents and/or access rights.
CheckFastRpmFiles	Basic	rpm, fastrpm	None of the files delivered using RPM seem to have been updated regarding length, date, and/or access rights.	One or more of the files delivered using RPM seem to have been updated regarding length, date, and/or access rights.
CheckSRtdir	Basic	lsall	None of the files of the deployed image seem to have been updated regarding length, date, and/or access rights.	One or more of the files of the deployed image seem to have been updated regarding length, date, and/or access rights.
CheckMd5sumDir	Sharp	md5all	None of the files of the deployed image seem to have been updated regarding content (md5 on the content).	One or more of the files of the deployed image seem to have been updated regarding content (md5 on the content).
CheckMandatoryFiles	Basic	ksis	Ksis binaries are present on the node and have the same length as those on the Management Node.	Ksis binaries are not present on the node or have not the same length as those on the Management Node.
CheckUsedKernel	Basic	kernel	Kernel used by the node is the same as the one used on reference/golden node when the image has been created.	Kernel used by the node does not look the same as the one used on reference/golden node when the image has been created.

Table 5-1. Standard checks delivered with Ksis

### 5.5.3 Modifying the Checks Database

It is possible to modify the database checks to adapt them to the way you use the image.

- To change check groups, edit the **group** file.
- To create a new check, add a new directory (`/etc/systemimager/ksis_check_Repository/<testName>.vid`) which includes at least the following:
  - **command** file, which contains the command to be run,
  - **group** file, which defines the group to which the command belongs.

This check will be included in the checks database and will be part of the checks performed on subsequent images.

## 5.5.4 Examining the Results

The checks result is a comparison between a command executed on the reference image and the same command executed on the nodes concerned. This comparison shows the evolution of the node against the reference and means that it is possible to determine if it is necessary to deploy the node again.

## 5.5.5 Looking at the Discrepancies

If the discrepancies between a node and the reference image are not significant, it may still be useful to analyze their development. There are several ways to do this.

- The **ksis checkdiff** command displays the discrepancies between the reference image and the results for a given check.

Example:

```
ksis checkdiff CheckSRTEdir node2
```

- You can also examine the results for the node:
  - `/etc/systemimager/ksis_check_Repository/` for an image,
  - `/usr/ksisClient/PATCH_<patchName>/ksis_check_Repository/` for a patch (name: `<patchName>`).

## 5.6 Importing and Exporting an Image

KSIS provides a function to export an image to another KSIS installation (on another administration node) or to import an image from another KSIS installation.

The **ksis export** command allows you to export a Reference image (not a Patch image). The image will be available as a tar file in the Ksis images directory:

`/var/lib/systemimager/images/<imageName>.tar`

```
ksis export <imageName> [<option>]
```



### Note:

The export operation does not automatically destroy the exported image.

The KSIS import command allows you to import a Reference image from a tar file in the KSIS images directory: `/var/lib/systemimager/images/<imageName>.tar`.

Once the import operation is completed, the image is available and may be listed by using the **ksis list** command.

The import / export feature can be used to archive images that are no longer used on nodes, but that the administrator wants to keep.

## 5.7 Ksis Commands

### 5.7.1 Syntax

```
ksis <action> <parameters> [<options>]
```

#### Options:

- S Step by step
- v Verbose
- g Debug mode
- G Detailed debug mode

#### Format for `nodeRange` or `groupName` parameter:

The nodes, to which the Ksis command applies, are specified either as a range of nodes (**nodeRange**) or as a group name (**groupName**).

- Several formats are possible for the **nodeRange** parameter, as shown in the following examples:
  - `<nodeRange> = host[1]`
  - `<nodeRange> = host[1,2,3,9]`
  - `<nodeRange> = host[1-3]`
  - `<nodeRange> = host[1-3,9]`
- The **groupName** is the name of a group of nodes defined in the ClusterDB. See the *Cluster Database Management* chapter for more information about these groups.

#### Getting Help:

For a complete description of the KSIS commands, enter:

```
ksis help
```

Or:

```
ksis help <action>
```

## 5.7.2 Advanced ksis create options

### -d

The **-d** option is used to define the individual disks of a node, which are to be included in the image.

```
Ksis create <myImage> <myReferenceNode> -d <myDisks>
```

The disks to be included appear after the **-d** option in a comma-separated list, as shown in the example below.

The node disks not listed will not be included in the image.

### Example

```
ksis create MyImage MyGolden -d /dev/sda,/dev/sdb
```

In the command above only disks **sda** and **sdb** will be included in the image.

### -dx

The **-dx** option is used in the similar fashion to the **-d** option. The only difference is that this option is exclusive. In other words, unlike the **-d** option, all the references to the mounted disks which are not included in the image will be deleted and the **/etc/fstab** file which lists the mounts points will be updated.

### When to use the -d and -dx options

The **-dx** option is used, for example, if for some reason it is decided that a particular disk bay (e.g. **/dev/sdj**) connected to the reference node, should not be included in an image when it is deployed.

If the **-d** option is used after deployment then the system will try to remount the **/dev/sdj** disk bay on all the deployed nodes. By using the **-dx** option with the **ksis create** command all references to the **/dev/sdj** bay are deleted, and it will not be remounted after deployment.

## 5.7.3 Creating the Image of the Reference Node

To create an image of the reference node use the **ksis create** command. This operation is done while you are logged onto the image server (Management Node).

```
ksis create <imagenam> <reference_node_name> [options]
```

This command creates a copy of the image of the reference node on the image server (Management Node). The resulting status for this image is "golden".

When using this command the check level associated with this image is requested. Choose **basic** for a standard level (see 5.5 *Checking Deployed Images* for other options).

## 5.7.4 Deleting an Image or a Patch

This command deletes the defined image or patch from the image server (Management Node).

```
ksis delete <imageNameOrPatchName>
```

## 5.7.5 Deploying an Image or a Patch

This command consists in the deployment of an image or a patch on the specified nodes.

```
ksis deploy <imageNameOrPatchName> <nodeRangeOrGroupName> [-P] [options]
```

When you deploy an image the command performs these steps on the nodes concerned:

- Checks the state of the node.
- Reboots the node in network mode.
- Loads the image from the image server using special algorithms to parallelize the loading and to minimize the loading time.
- Checks log files.
- Boots the node with the image loaded.
- If the **-P** option is specified, performs the post configuration of the nodes.



### Note:

See chapter 3 in the Bull *HPC BAS4 for Xeon Maintenance Guide* for more information on the **Ksis** log files.

### Using the **-P** option

The use of the **-P** option is necessary to automatically configure the nodes after deployment of an image. This post configuration is based on a “configuration set”, which defines the services to be configured.

The present delivery provides only one configuration set, named **PostConfig**. It configures **Ganglia**, **Syslog-ng**, **NTP**, **SNMP** and **Pdsh** on the deployed nodes. It also configures the IP over Infiniband interfaces according to the information in the Cluster database.



### Note:

In the future it will be possible to define different configuration sets according to the Cluster needs.



## 5.7.6 Removing a Patch

This action concerns only the images with the "patch" status. It consists in removing the last deployed patch from the nodes.

```
ksis undeploy <patchName> <nodeRangeOrGroupName> [options]
```

## 5.7.7 Getting Information about an Image or a Node

This command displays information for the specified image or node.

```
ksis show <imageNameOrNodeName>
```

## 5.7.8 Listing Images on the Image Server

This command gives the list and status of the images available on the image server. Their status is one of the following:

```
ksis list [<options>]
```

<b>golden</b>	reference image (from a reference node also called golden node).
<b>patch</b>	patch (result of a store command).
<b>patched golden</b>	modified reference image (result of a detach command).
<b>working patch</b>	modification in progress; cannot be deployed, waiting for store command.

Example:

```
ksis list
```

Image Name	Status	Creation Date
BAS3-v13ulu2	golden	2005-01-14 14:33:02
Compute_hpceth_ulu2	golden	2005-01-14 15:41:25
Compute_hpceth_ulu2.s1.0	patch	2005-01-20 13:49:27
Compute_hpceth_ulu2.s1.1	working patch	2005-01-22 14:41:03

## 5.7.9 Listing Images by Nodes

This command lists the current images available and their status on the nodes.

```
ksis nodelist [<options>]
```

Example:

```
ksis nodelist
```

```
nc1  unreachable -  
nc2  up Compute_hpceth_u1u2      2005-01-20 11:28:30  
nc3  up Compute_hpceth_u1u2      2005-01-20 11:29:33  
nc4  up Compute_hpceth_u1u2.s1.0 2005-01-21 12:03:01  
nc5  down Compute_hpceth_u1u2.s1.0 2005-01-21 12:10:43
```

## 5.8 Modifying an Image

There are two means of creating a patch to be used to modify an image:

1. Using the **ksis workon** and **ksis store** commands to modify the image outside of the nodes on which it is deployed. These are reserved for minor modifications such as file changes.
2. **ksis buildpatch**, which is used for more complex image modifications.



### Important:

**ksis buildpatch** and the used of patches should only be used for limited image changes. For fundamental image changes the best method remains the creation and the deployment of a new image.

### 5.8.1 Creating a Working Patch Image

```
ksis workon <ImageNameOrPatchName> [<options>]
```

This **workon** command allows an image to be modified without creating a new one.

The command duplicates the image and creates a workon environment (shell) from which all the modifications required can be performed.

The status for this new image is **working patch**.

### 5.8.2 Creating a Patch Image

```
ksis store <patchname> [<options>]
```

The **store** command is the mandatory step after running the **workon** command.

The command creates a new image containing the differences from the mother image. Using the **-R** option means that a reboot is necessary after installing the patch.

The status for this new image is **patch**.

When using this command the check level associated with this image is requested. Choose 'basic' for a standard level (see 5.5 *Checking Deployed Images* for other options).

## 5.8.3 Creating a Patched Golden Image

```
ksis detach <imagename> [<options>]
```

The **detach** command allows you to modify an image without creating a new one. The command creates a new image resulting in the application of the patch on the source image. The status for this new image is **patched golden**.

## 5.8.4 Building a Patch

**ksis buildpatch** is used to create a patch from the differences between two images that can then be used to transform the software structure and content of a first node which has the first image deployed on it so that it matches a node which has the second image deployed on it.



### Note:

**ksis buildpatch** can only be for two images which are derived from each other and not for images which are unrelated.

The command below would create a patch from the differences between the **<imageName1>** image and the **<imageName2>** image.

```
ksis buildpatch <imageName1> <imageName2>
```

### Using **ksis buildpatch**

1. Make any changes required to the deployed version of the **<imageName1>** image. This is done by logging on to a node **n** which has **<imageName1>** on it and changing whatever needs to be changed. If necessary reboot on the node and check that everything is working OK.
2. Create an image of the node which has the **<imageName1>** image on it.

```
ksis create <imageName1> n
```

3. Create a patch of the differences between the **<imageName1>** and **<imageName2>** images. The patch will be automatically name e.g. **imageName1.s1.0** for the first patch generated for **<imageName1>** image.

```
ksis buildpatch <imageName1> <imageName2>
```

4. Deploy this patch on to the nodes which have **<imageName1>** on them.

```
ksis deploy <patch_name> <nodelist>
```

5. These nodes will now have a software content and structure which matches **<imageName2>**.

## 5.9 Checking Images

The **check** command checks the image deployed on a node set.

```
ksis check <nodeRangeOrGroupName>
```

The **checkdiff** command displays the discrepancies between a reference node and the results for a given check on a given node.

```
ksis checkdiff <testName> <node>
```

## 5.10 Importing and Exporting Images

The **export** command exports an image from one cluster to another cluster.

```
ksis export <imageName>
```

The **import** command imports an image previously exported from another cluster.

```
ksis import <imageName>
```

## 5.11 Rebuilding ClusterDB Data before Deploying an Image

There are two cases where it is necessary to update the reference information before deploying an image:

- some values have changed in the ClusterDB
- or an image has been imported, and its ClusterDB information must be updated.

To do so, use the **bulddatanode** command, which updates the images with the latest values of the ClusterDB:

```
ksis bulddatanode
```

```
Nodes context will be updated to take in account new data from DB  
Continue (yes/no)
```

Answer « yes » to the question.



---

## Chapter 6. Resource Management

Merely grouping together several machines on a network is not enough to constitute a real cluster. Resource Management software is required to optimize the throughput within the cluster according to specific scheduling policies.

A **resource manager** is used to allocate resources, to find out the status of resources, and to collect task execution information. From this information the scheduling policy can be applied. Bull HPC platforms use **SLURM** an open-source, scalable resource manager.

This chapter describes the following topics:

- 6.1 *Resource Management with SLURM*
- 6.2 *SLURM Configuration*
- 6.3 *Administering Cluster Activity with SLURM*

## 6.1 Resource Management with SLURM

### 6.1.1 SLURM Key Functions

As a cluster resource manager, SLURM has three key functions. Firstly, it allocates exclusive and/or non-exclusive access to resources (compute nodes) to users for some duration of time so they can perform work. Secondly, it provides a framework for starting, executing, and monitoring work (normally a parallel job) on the set of allocated nodes. Finally, it arbitrates conflicting requests for resources by managing a queue of pending work.

Users interact with SLURM through four command line utilities:

- **SRUN** for submitting a job for execution and optionally controlling it interactively,
- **SCANCEL** for terminating a pending or running job,
- **SQUEUE** for monitoring job queues, and
- **SINFO** for monitoring partition and overall system state.

System administrators perform privileged operations through an additional command line utility, **SCONTROL**.

The central controller daemon, **SLURMCTLD**, maintains the global state and directs operations. Compute nodes simply run a **SLURMD** daemon (similar to a remote shell daemon) to export control to SLURM.

**SLURM** supports resource management across a single cluster.

**SLURM** is not a sophisticated batch system. In fact, it was expressly designed to provide high-performance parallel job management while leaving scheduling decisions to an external entity. Its default scheduler implements **First-In First-Out (FIFO)**. A scheduler entity can establish a job's initial priority through a plug-in.

An external scheduler may also submit, signal, and terminate jobs as well as reorder the queue of pending jobs via the API.



## 6.1.2 SLURM Components

SLURM consists of two types of daemons and five command-line user utilities. The relationships between these components are illustrated in the following diagram:

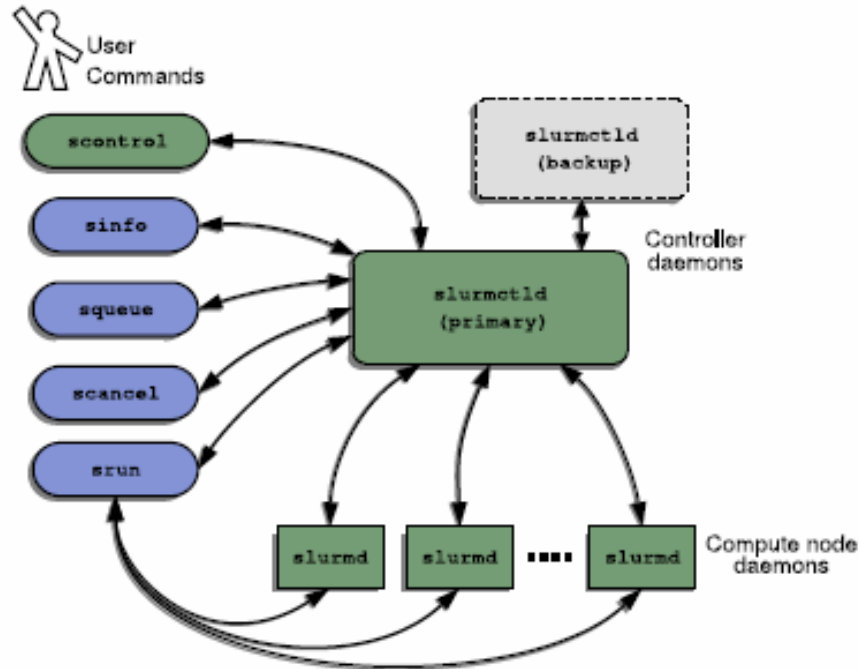


Figure 6-1. SLURM Simplified Architecture

## 6.1.3 SLURM Daemons

### 6.1.3.1 SLURMCTLD

The central control daemon for **SLURM** is called **SLURMCTLD**. **SLURMCTLD** is *multi*-threaded; thus, some threads can handle problems without delaying services to normal jobs that are also running and need attention. **SLURMCTLD** runs on a single management node (with a fail-over spare copy elsewhere for safety), reads the **SLURM** configuration file, and maintains state information on:

- Nodes (the basic compute resource)
- Partitions (sets of nodes)
- Jobs (or resource allocations to run jobs for a time period)
- Job steps (parallel tasks within a job).

The **SLURMCTLD** daemon in turn consists of three software subsystems, each with a specific role:

Software Subsystem	Role Description
<b>Node Manager</b>	Monitors the state and configuration of each node in the cluster. It receives state-change messages from each compute node's SLURMD daemon asynchronously, and it also actively polls these daemons periodically for status reports.
<b>Partition Manager</b>	Groups nodes into disjoint sets (partitions) and assigns job limits and access controls to each partition. The partition manager also allocates nodes to jobs (at the request of the Job Manager) based on job and partition properties. SCONTROL is the (privileged) user utility that can alter partition properties.
<b>Job Manager</b>	Accepts job requests (from SRUN or a metabatch system), places them in a priority-ordered queue, and reviews that queue periodically or when any state change might allow a new job to start. Resources are allocated to qualifying jobs and that information transfers to (SLURMD on) the relevant nodes so the job can execute. When all nodes assigned to a job report that their work is done, the Job Manager revises its records and reviews the pending-job queue again.

Table 6-1. Role Descriptions for SLURMCTLD Software Subsystems

The following figure illustrates these roles of the SLURM Software Subsystems.

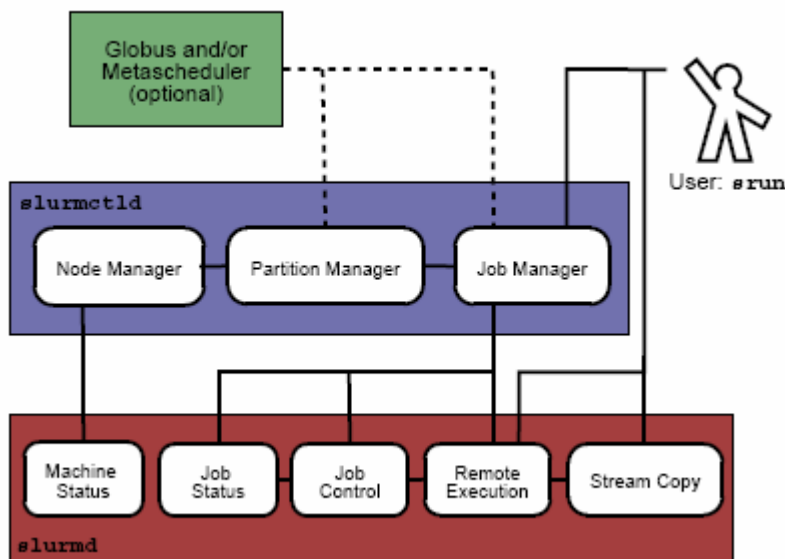


Figure 6-2. SLURM Architecture - Subsystems

### 6.1.3.2 SLURMD

The **SLURMD** daemon runs on all the compute nodes of each cluster that **SLURM** manages and performs the lowest level work of resource management. Like **SLURMCTLD** (previous subsection), **SLURMD** is multi-threaded for efficiency; but, unlike **SLURMCTLD**, it runs with root privileges (so it can initiate jobs on behalf of other users).

SLURMD carries out five key tasks and has five corresponding subsystems. These subsystems are described in the following table.

SLURMD Subsystem	Description of Key Tasks
Machine Status	Responds to <b>SLURMCTLD</b> requests for machine state information and sends asynchronous reports of state changes to help with queue control.
Job Status	Responds to <b>SLURMCTLD</b> requests for job state information and sends asynchronous reports of state changes to help with queue control.
Remote Execution	Starts, monitors, and cleans up after a set of processes (usually shared by a parallel job), as decided by <b>SLURMCTLD</b> (or by direct user intervention). This can often involve many changes to process-limit, environment-variable, working-directory, and user-id.
Stream Copy Service	Handles all <b>STDERR</b> , <b>STDIN</b> , and <b>STDOUT</b> for remote tasks. This may involve redirection, and it always involves locally buffering job output to avoid blocking local tasks.
Job Control	Propagates signals and job-termination requests to any SLURM-managed processes (often interacting with the Remote Execution subsystem).

Table 6-2. SLURMD Subsystems and Key Tasks

## 6.1.4 Scheduler Types

The system administrator for each machine can configure **SLURM** to invoke one of several alternative local job schedulers. To determine which scheduler SLURM is currently invoking on any machine, execute the following command:

```
scontrol show config |grep SchedulerType
```

where the returned string will have one of the values described in the following table.

Returned String Value	Description
<b>builtin</b>	A first-in-first-out scheduler. SLURM executes jobs strictly in the order in which they were submitted (for each resource partition), unless those jobs have different priorities. Even if resources become available to start a specific job, SLURM will wait until there is no previously-submitted job pending (which sometimes confuses impatient job submitters). This is the default.

<b>backfill</b>	<p>Modifies strict FIFO scheduling to take advantage of resource islands that may appear as earlier jobs complete. SLURM will start later-submitted jobs out of order if resources become available, <i>and</i> if doing so does not delay the expected execution time of any earlier-submitted job. To increase the job's chances of benefiting from such backfill scheduling:</p> <p>(1) specify reasonable time limits (the default is the same time limit for all jobs in the partition, which may be too large), and</p> <p>(2) avoid requiring or excluding specific nodes by name.</p>
<b>wiki</b>	<p>Uses the Maui Scheduler, with a sophisticated set of internal scheduling algorithms. This choice can be configured in several ways to optimize job throughput. Details are posted on a support web site at the following URL:</p> <p style="text-align: center;"><a href="http://supercluster.org/maui">http://supercluster.org/maui</a></p>

Table 6-3. SLURM Scheduler Types

## 6.2 SLURM Configuration

The SLURM configuration file, **slurm.conf**, is an ASCII file that describes the following:

- General SLURM configuration information
- The nodes to be managed
- Information about how those nodes are grouped into partitions
- Various scheduling parameters associated with those partitions.

The SLURM configuration file includes a wide variety of parameters. This configuration file must be available on each node of the cluster. A full description of the parameters is included in the **slurm.conf** man page. The **slurm.conf** file should define at least the configuration parameters defined in the samples provided and likely additional ones. Any text following a **#** is considered a comment. The keywords in the file are not case sensitive, although the argument typically is (e.g., "**SlurmUser=slurm**" might be specified as "**slurmuser=slurm**"). Port numbers to be used for communications are specified as well as various timer values.

A description of the nodes and their grouping into partitions is required. A simple node range expression may optionally be used to specify ranges of nodes to avoid building a configuration file with large numbers of entries. The node range expression can contain one pair of square brackets with a sequence of comma separated numbers and/or ranges of numbers separated by a **-** (e.g. "linux[0-64,128]", or "lx[15,18,32-33]").

Node names can have up to three name specifications: **NodeName** is the name used by all **SLURM** tools when referring to the node, **NodeAddr** is the name or IP address SLURM uses to communicate with the node, and **NodeHostname** is the name returned by the command `/bin/hostname -s`. Only **NodeName** is required (the others default to the same name), although supporting all three parameters provides complete control over naming and addressing the nodes. See the **slurm.conf** man page for details on all configuration parameters.

Nodes can be in more than one partition and each partition can have different constraints (permitted users, time limits, job size limits, etc.). Each partition can thus be considered a separate queue. Partition and node specifications use node range expressions to identify nodes in a concise fashion. The Example #2 configuration file (see Section 6.2.2 **slurm.conf** Example Files) defines a 1154-node cluster for **SLURM**, but it might be used for a much larger cluster by just changing a few node range expressions. Specify the minimum processor count (**Procs**), real memory space (**RealMemory**, megabytes), and temporary disk space (**TmpDisk**, megabytes) that a node should have to be considered as available for use. Any node lacking these minimum configuration values will be considered **DOWN** and not scheduled. An annotated sample configuration file for SLURM is provided with this distribution as `/etc/slurm/slurm.conf.example`. Edit this configuration file to suit the needs of the user site and cluster, and then copy it to `/etc/slurm/slurm.conf`.

## 6.2.1 Configuration Parameters

Three types of SLURM configuration parameters are described in this section.

- General configuration parameters
- Node configuration parameters
- Partition configuration parameters

### 6.2.1.1 General Configuration Parameters

This section describes the overall configuration parameters available for SLURM.

#### AuthType

Define the authentication method for communications between SLURM components. Acceptable values at present include **auth/none**, **auth/authd** and **auth/munge**. The default value is **auth/none**, which means the UID included in communication messages is not verified. This may be fine for testing purposes, but do not use **auth/none** if any security is needed.

- **auth/authd** indicates that Brett Chun's **authd** is to be used (see <http://www.theether.org/authd/> for more information).
- **auth/munge** indicates that Chris Dunlap's **munge** is to be used (this is the best supported authentication mechanism for **SLURM**. The **munge** application will need to be installed in order to use this functionality ( - see <http://www.llnl.gov/linux/munge/> for more information).

All SLURM daemons and commands must have been terminated prior to changing the value of **AuthType** and later restarted (SLURM jobs can be preserved).

#### BackupAddr

Name to use when referring to the **BackupController** for establishing a communications path. This name will be used as an argument to the **gethostbyname()** function for identification.

For example, `e1x0000` might be used to designate the Ethernet address for node `1x0000`. By default the **BackupAddr** will be identical in value to **BackupController**.

#### BackupController

The name of the machine where SLURM control functions are to be executed in the event that **ControlMachine** fails. This node may also be used as a compute server if so desired. It will come into service as a controller only upon the failure of **ControlMachine** and will revert to a "*standby*" mode when the **ControlMachine** becomes available once again. This should be a node name without the full domain name (e.g. `1x0002`). While not essential, it is recommended that a backup controller be specified.

#### CacheGroups

If set to 1, the **SLURMD** daemon will cache **/etc/groups** entries. This can improve performance for highly parallel jobs if NIS servers are used and unable to respond very quickly. The default value is 0 to disable caching group data.

### CheckpointType

Define the system-initiated checkpoint method to be used for user jobs. The **SLURMCTLD** daemon must be restarted for a change in **CheckpointType** to take effect. Acceptable values at present include "checkpoint/none" and "checkpoint/aix" (only on AIX systems). The default value is "checkpoint/none".

### ControlAddr

Name to use when referring to the **ControlMachine** for establishing a communications path. This name will be used as an argument to the **gethostbyname()** function for identification. For example, "e1x0000" might be used to designate the Ethernet address for node "1x0000". By default the **ControlAddr** will be identical in value to **ControlMachine**.

### ControlMachine

The name of the machine where SLURM control functions are executed. This should be a node name without the full domain name (e.g. "1x0001"). This value must be specified.

### Epilog

Fully-qualified pathname of a script to execute as user root on every node when a user's job completes (e.g. **/usr/local/slurm/epilog**). This may be used to purge files, disable user login, etc. By default there is no epilog.

### FastSchedule

If set to 1 (the default), then the configuration of each node will be considered to be that which is specified in the configuration file. If set to 0, then scheduling decisions will be based upon the actual configuration of each individual node. If the number of node configuration entries in the configuration file is significantly lower than the number of nodes, setting **FastSchedule** to 1 will permit much faster scheduling decisions to be made. (The scheduler can check only the values in a few configuration records instead of possibly thousands of node records. If a job cannot be initiated immediately, the scheduler may execute these tests repeatedly.) Note that on systems with hyper-threading, the processor count reported by the node will be twice the actual processor count. Review these values for scheduling purposes.

### FirstJobId

The job id to be used for the first job submitted to **SLURM** without a specific requested value. Job id values generated will be incremented by 1 for each subsequent job. This may be used to provide a meta-scheduler with a job id space, which is disjoint from the interactive jobs. The default value is 1.

### HeartbeatInterval

Obsolete parameter. Interval of heartbeat for **SLURMD** daemon is half of **SlurmdTimeout**. Interval of heartbeat for **SLURMCTLD** daemon is half of **SlurmctldTimeout**.

### **InactiveLimit**

The interval, in seconds, a job or job step is permitted to be inactive before it is terminated. A job or job step is considered inactive if the associated **SRUN** command is not responding to **SLURM** daemons. This could be due to the termination of the **SRUN** command or the program being in a stopped state. A batch job is considered inactive if it has no active job steps (e.g. periods of pre- and post-processing). This limit permits defunct jobs to be purged in a timely fashion without waiting for their time limit to be reached. This value should reflect the possibility that the **SRUN** command may be stopped by a debugger or considerable time could be required for batch job pre- and post-processing. This limit is ignored for jobs running in partitions with the **RootOnly** flag set (the scheduler running as root will be responsible for the job). The default value is unlimited (zero). May not exceed 65533.

### **JobAcctType**

Define the job accounting mechanism type. Acceptable values at present include **jobacct/aix** (for AIX operating system), **jobacct/linux** (for Linux operating system) and **jobacct/none** (no accounting data collected). The default value is **jobacct/none**. In order to use the **SACCT** tool, **jobacct/aix** or **jobacct/linux** must be configured.

### **JobAcctLogFile**

Define the location where job accounting logs are to be written. For **jobacct/none** this parameter is ignored. For **jobacct/linux** this is the fully-qualified file name for the data file.

### **JobAcctFrequency**

Define the polling frequencies to pass to the job accounting plug-in. For **jobacct/none** this parameter is ignored. For **jobacct/linux** the parameter is a number of seconds between polls.

### **JobCompLoc**

The interpretation of this value depends upon the logging mechanism specified by the **JobCompType** parameter.

### **JobCompType**

Define the job completion logging mechanism type. Acceptable values at present include **jobcomp/none**, **jobcomp/filetxt**, and **jobcomp/script**. The default value is **jobcomp/none**, which means that upon job completion the record of the job is purged from the system. The value **jobcomp/filetxt** indicates that a record of the job should be written to a text file specified by the **JobCompLoc** parameter. The value **jobcomp/script** indicates that a script specified by the **JobCompLoc** parameter is to be executed with environment variables indicating the job information.

### **JobCredentialPrivateKey**

Fully qualified pathname of a file containing a private key used for authentication by **SLURM** daemons.

### **JobCredentialPublicCertificate**

Fully qualified pathname of a file containing a public key used for authentication by **SLURM** daemons.



### KillTree

This option is mapped to `ProctrackType=proctrack/linuxproc`. It will be removed from future releases.

### KillWait

The interval, in seconds, given to a job's processes between the `SIGTERM` and `SIGKILL` signals upon reaching their time limit. If the job fails to terminate gracefully in the interval specified, it will be forcibly terminated. The default value is 30 seconds. May not exceed 65533.

### MaxJobCount

The maximum number of jobs `SLURM` can have in its active database at one time. Set the values of `MaxJobCount` and `MinJobAge` to insure the `SLURMCTLD` daemon does not exhaust its memory or other resources. Once this limit is reached, requests to submit additional jobs will fail. The default value is 2000 jobs. This value may not be reset via `scontrol reconfig`. It only takes effect upon restart of the `SLURMCTLD` daemon. May not exceed 65533.

### MinJobAge

The minimum age of a completed job before its record is purged from `SLURM`'s active database. Set the values of `MaxJobCount` and `MinJobAge` to insure the `SLURMCTLD` daemon does not exhaust its memory or other resources. The default value is 300 seconds. A value of zero prevents any job record purging. May not exceed 65533.

### MpiDefault

Identifies the default type of `MPI` to be used. `SRUN` may override this configuration parameter in any case. Currently supported versions include: `lam` (which supports `LAM MPI` and `Open MPI`, but specify `none` instead and let `LAM MPI` and `Open MPI` select the plug-in using an option of the `SRUN` command), `mpich-gm`, `mvapich`, and `none` (default, which works for many other versions of `MPI`).

### PluginDir

Identifies the places in which to look for `SLURM` plug-ins. This is a colon-separated list of directories, like the `PATH` environment variable. The default value is `/usr/local/lib/slurm`.

### PlugStackConfig

Location of the configuration file for `SLURM` stackable plug-ins that use the Stackable Plug-in Architecture for Node job (K)control (`SPANK`). This provides support for a highly configurable set of plug-ins to be called before and/or after execution of each task spawned as part of a user's job step. Default location is `plugstack.conf` in the same directory as the system `slurm.conf`. For more information on `SPANK` plug-ins, see the `spank(8)` manual.

### ProctrackType

Identifies the plug-in to be used for process tracking. The **SLURMD** daemon uses this mechanism to identify all processes that are children of processes it spawns for a user job. Acceptable values at present include **proctrack/aix** (which uses an AIX kernel extension and is the default for AIX systems), **proctrack/linuxproc** (which uses Linux process tree), **proctrack/rms** (which uses Quadrics kernel patch and is the default if **SwitchType=switch/elan**) and **proctrack/pgid** (which is the default for all other systems). The **SLURMD** daemon must be restarted for a change in **ProctrackType** to take effect.



#### Note:

**proctrack/linuxproc** is not compatible with **switch/elan**.

### Prolog

Fully-qualified pathname of a script to execute as user root on every node when a user's job begins execution (e.g. **/usr/local/slurm/prolog**). This may be used to purge files, enable user login, etc. By default there is no prolog.

### PropagatePrioProcess

Setting **PropagatePrioProcess** to "1", will cause a users job to run with the same priority (aka nice value) as the users process which launched the job on the submit node. If set to "0", or left unset, the users job will inherit the scheduling priority from the **SLURM** daemon.

### PropagateResourceLimits

A list of comma-separated resource limit names. The **SLURMD** daemon uses these names to obtain the associated (soft) limit values from the users process environment on the submit node. These limits are then propagated and applied to the jobs that will run on the compute nodes. This parameter can be useful when system limits vary among nodes. Any resource limits that do not appear in the list are not propagated. However, the user can override this by specifying which resource limits to propagate with the **SRUN** commands **"--propagate"** option. If neither of the propagate resource limit's parameters are specified, then the default action is to propagate all limits. Only one of the parameters, either **PropagateResourceLimits** or **PropagateResourceLimitsExcept**, may be specified.

### PropagateResourceLimitsExcept

A list of comma-separated resource limit names. By default, all resource limits will be propagated, (as described by the **PropagateResourceLimits** parameter), except for the limits appearing in this list. The user can override this by specifying which resource limits to propagate with the **SRUN** commands **"--propagate"** option.

### ReturnToService

If set to 1, then a **DOWN** node will become available for use upon registration. The default value is 0, which means that a node will remain in the **DOWN** state until the system administrator explicitly changes its state (even if the **SLURMD** daemon registers and resumes communications).

### SchedulerAuth

An authentication token, which must be used in a scheduler communication protocol. The interpretation of this value depends upon the value of **SchedulerType**. In the Wiki scheduler plug-in, this value must correspond to the checksum seed with which Maui was compiled.

### SchedulerPort

The port number on which **SLURMCTLD** should listen for connection requests. This value is only used by the Maui Scheduler (see **SchedulerType**).

### SchedulerRootFilter

Identifies whether or not **RootOnly** partitions should be filtered from any external scheduling activities. If set to 0, then **RootOnly** partitions are treated like any other partition. If set to 1, then **RootOnly** partitions are exempt from any external scheduling activities. The default value is 1. Currently only used by the built-in backfill scheduling module **sched/backfill** (see **SchedulerType**).

### SchedulerType

Identifies the type of scheduler to be used. Acceptable values include **sched/builtin** for the built-in FIFO scheduler, **sched/backfill** for a backfill scheduling module to augment the default FIFO scheduling, **sched/hold** to hold all newly arriving jobs if a file **/etc/slurm.hold** exists otherwise use the built-in FIFO scheduler, and **sched/wiki** for the Wiki interface to the Maui Scheduler. The default value is **sched/builtin**. Backfill scheduling will initiate lower-priority jobs if doing so does not delay the expected initiation time of any higher priority job. Note that this backfill scheduler implementation is relatively simple. It does not support partitions configured to share resources (run multiple jobs on the same nodes) or support jobs requesting specific nodes. When initially setting the value to **sched/wiki**, any pending jobs must have their priority set to zero (held). When changing the value from **sched/wiki**, all pending jobs should have their priority change from zero to some large number. The **SCONTROL** command can be used to change job priorities. The **SLURMCTLD** daemon must be restarted for a change in scheduler type to become effective.

### SelectType

Identifies the type of resource selection algorithm to be used. Acceptable values include **select/linear** for allocation of entire nodes assuming a one-dimensional array of nodes in which sequentially ordered nodes are preferable, **select/cons\_res** for allocation of individual processors within the available nodes. The default value is **select/linear**.

### SlurmUser

The name of the user under which the **SLURMCTLD** daemon executes. For security purposes, a user other than "**root**" is recommended. The default value is "**root**".

### SlurmctldDebug

The level of detail to provide **SLURMCTLD** daemon's logs. Values from 0 to 7 are legal, with "0" being "**quiet**" operation, and "7" being extremely verbose. The default value is 3.

### SlurmctldLogFile

Fully-qualified pathname of a file into which the **SLURMCTLD** daemon's logs are written. The default value is none (performs logging via **syslog**).

### SlurmctldPidFile

Fully-qualified pathname of a file into which the **SLURMCTLD** daemon may write its process id. This may be used for automated signal processing. The default value is `"/var/run/slurmctld.pid"`.

### SlurmctldPort

The port number that the **SLURM** controller, **SLURMCTLD**, listens to for work. The default value is **SLURMCTLD\_PORT** as established at system build time.



#### Note:

Either **SLURMCTLD** and **SLURMD** daemons must not execute on the same nodes or the values of **SlurmctldPort** and **SlurmdPort** must be different.

### SlurmctldTimeout

The interval, in seconds, that the backup controller waits for the primary controller to respond before assuming control. The default value is 120 seconds. May not exceed 65533.

### SlurmdDebug

The level of detail to provide **SLURMD** daemon's logs. Values from 0 to 7 are legal, with "0" being "quiet" operation, and "7" being extremely verbose. The default value is 3.

### SlurmdLogFile

Fully-qualified pathname of a file into which the **SLURMD** daemon's logs are written. The default value is none (performs logging via **syslog**). Any "%h" within the name is replaced with the hostname on which the **SLURMD** is running.

### SlurmdPidFile

Fully-qualified pathname of a file into which the **SLURMD** daemon may write its process id. This may be used for automated signal processing. The default value is `"/var/run/slurmd.pid"`.

### SlurmdPort

The port number that the **SLURM** compute node daemon, **SLURMD**, listens to for work. The default value is **SLURMD\_PORT** as established at system build time.

### SlurmdSpoolDir

Fully-qualified pathname of a directory into which the **SLURMD** daemon's state information and batch job script information is written. This must be a common pathname for all nodes, but should represent a directory that is local to each node (reference a local file system). The default value is `"/var/spool/slurmd"`.

**Note:**

This directory is also used to store SLURMD's shared memory lockfile, and should not be changed unless the system is being cleanly restarted. If the location of **SlurmdSpoolDir** is changed and **SLURMD** is restarted, the new daemon will attach to a different shared memory region and lose track of any running jobs.

**SlurmdTimeout**

The interval, in seconds, that the SLURM controller waits for **SLURMD** to respond before configuring that node's state to DOWN. The default value is 300 seconds. A value of zero indicates the node will not be tested by **SLURMCTLD** to confirm the state of **SLURMD**, the node will not be automatically set to a DOWN state indicating a non-responsive **SLURMD**, and some other tool will take responsibility for monitoring the state of each compute node and its **SLURMD** daemon. The value may not exceed 65533.

**StateSaveLocation**

Fully-qualified pathname of a directory into which the SLURM controller, **SLURMCTLD**, saves its state (e.g. **/usr/local/slurm/checkpoint**). **SLURM** state will be saved here to recover from system failures. **SlurmUser** must be able to create files in this directory. If a **BackupController** is configured, this location should be readable and writable by both systems. The default value is **/tmp**. If any SLURM daemons terminate abnormally, their core files will also be written into this directory.

**SrunEpilog**

Fully-qualified pathname of an executable to be run by **SRUN** following the completion of a job step. The command line arguments for the executable will be the command and arguments of the job step. This configuration parameter may be overridden by **srun's --epilog** parameter.

**SrunProlog**

Fully-qualified pathname of an executable to be run by **SRUN** prior to the launch of a job step. The command line arguments for the executable will be the command and arguments of the job step. This configuration parameter may be overridden by the **SRUN --prolog** parameter.

**SwitchType**

Identifies the type of switch or interconnect used for application communications. Acceptable values include **switch/none** for switches not requiring special processing for job launch or termination (**Myrinet**, **Ethernet**, and **InfiniBand**), **switch/elan** for Quadrics Elan 3 or Elan 4 interconnects. The default value is **switch/none**. All SLURM daemons, commands and running jobs must be restarted for a change in **SwitchType** to take effect. If running jobs exist at the time **SLURMCTLD** is restarted with a new value of **SwitchType**, records of all jobs in any state may be lost.

**TaskEpilog**

Fully qualified pathname of a program to be executed as the SLURM job's owner after termination of each task. See **TaskPlugin** for execution order details.

## TaskPlugin

Identifies the type of task launch plug-in, typically used to provide resource management within a node (e.g. pinning tasks to specific processors). Acceptable values include **task/none** for systems requiring no special handling or **tasks/affinity** to enable the **--cpu\_bind** and/or **--mem\_bind** affinity **SRUN** options. The default value is **task/none**. The order of task prolog/epilog execution is as follows:

1. **pre\_launch()**: function in **TaskPlugin**
2. **TaskProlog**: system-wide per task program defined in **slurm.conf**
3. user prolog: job step specific task program defined using the **SRUN --task-prolog** option or **SLURM\_TASK\_PROLOG** environment variable
4. Execute the job step's task user epilog: job step specific task program defined using the **SRUN --task-epilog** option or **SLURM\_TASK\_EPILOG** environment variable
5. **TaskEpilog**: system-wide per task program defined in **slurm.conf**
6. **post\_term()**: function in **TaskPlugin**

## TaskProlog

Fully-qualified pathname of a program to be executed as the **SLURM** job's owner prior to the initiation of each task. Aside from the normal environment variables, this has **SLURM\_TASK\_PID** available to identify the process ID of the task being started. Standard output from this program of the form **export NAME=value** will be used to set environment variables for the task being spawned.



See: **TaskPlugin** for execution order details.

## TmpFS

Fully-qualified pathname of the file system available to user jobs for temporary storage. This parameter is used in establishing a node's **TmpDisk** space. The default value is **/tmp**.

## TreeWidth

**SLURMD** daemons use a virtual tree network for communications. **TreeWidth** specifies the width of the tree (i.e. the fanout). The default value is 50.

## UseCPUSETS

When set to 1, this determines whether **CPUSETS** (Multiprocessor partitioning for Linux) will be used.

**CPUSETS** are lightweight objects in the Linux kernel that enable users to partition their multiprocessor servers by creating execution areas. The ultimate objective is to contain processes to a certain number of well-identified processors.

**CPUSETS**:

- Allow the creation of sets of CPUs on the system, and bind applications to them.
- Provide a way to create sets of CPUs inside a specific CPU set: hence a system administrator can partition a system among users, and users can then further partition their partition among their applications.
- Memory used by a **CPUSET** may be restricted to some of the nodes of a **NUMA** system.



#### Note:

This parameter is only effective if the `TaskPlugin` parameter (see above) is set to **task/affinity**. The use of this parameter also requires that the `libcpuset.so` library be installed on the compute nodes.

#### Example

```
UseCPUSETS=1 or UseCPUSETS=Yes # default is not to use CPUSETS
```

#### UsePAM

If set to 1, **PAM (Pluggable Authentication Modules for Linux)** will be enabled. **PAM** is used to establish the upper bounds for resource limits. With **PAM** support enabled, local system administrators can dynamically configure system resource limits.

Changing the upper bound of a resource limit will not alter the limits of running jobs, only jobs started after a change has been made will pick up the new limits. The default value is 0 (not to enable **PAM** support).

Remember that **PAM** also needs to be configured to support **SLURM** as a service. For sites using **PAM**'s directory based configuration option, a configuration file named **SLURM** should be created. The module-type, control-flags, and module-path names that should be included in the file are: **auth required pam\_localuser.so**, **auth required pam\_shells.so**, **account required pam\_unix.so**, **account required pam\_access.so** and **session required pam\_unix.so**. For sites configuring **PAM** with a general configuration file, the appropriate lines (see above), where **SLURM** is the service-name, should be added.

#### WaitTime

Specifies how many seconds the **SRUN** command should by default wait after the first task terminates before terminating all remaining tasks. The `--wait` option on the **SRUN** command line overrides this value. If set to 0, this feature is disabled. May not exceed 65533.

### 6.2.1.2 Node Configuration Parameters

The configuration of nodes (or machines) that will be managed by **SLURM** is also specified in `/etc/slurm/slurm.conf`. Only the **nodeName** must be supplied in the configuration file. All other node configuration information is optional. It is advisable to establish baseline node configurations, especially if the cluster is heterogeneous. Nodes that register to the system with less than the configured resources (e.g. too little memory) will be placed in the "DOWN" state to avoid having any jobs scheduled on them. Establishing baseline configurations will also speed **SLURM**'s scheduling process by permitting it to compare job requirements against these (relatively few) configuration parameters and possibly avoid having to check job requirements against every individual node's configuration.

The resources checked at node registration time are: **Procs**, **RealMemory** and **TmpDisk**. While baseline values for each of these can be established in the configuration file, the actual values upon node registration are recorded and these actual values may be used for scheduling purposes (depending upon the value of **FastSchedule** in the configuration file).

Default values can be specified with a record in which **nodeName** is DEFAULT. The default entry values will apply only to lines that follow it in the configuration file; the default values can be reset multiple times in the configuration file with multiple entries where "**nodeName=DEFAULT**". The "**nodeName=**" specification must be placed on every line describing the configuration of nodes. In fact, it is possible and desirable to define the configurations of all nodes in only a few lines. This convention permits significant optimization in the scheduling of larger clusters. In order to support the concept of jobs requiring consecutive nodes on the some architecture, node specifications should be placed in this file in consecutive order. No single node name may be listed more than once in the configuration file.

Use "**DownNodes=**" to record the state of nodes which are temporarily in a DOWN or DRAINED state without altering permanent configuration information. A job step's tasks are allocated to nodes in the order in which the nodes appear in the configuration file. There is presently no capability within SLURM to arbitrarily order a job step's tasks.

A simple node range expression may optionally be used to specify ranges of nodes to avoid building a configuration file with a large numbers of entries. The node range expression can contain one pair of square brackets with a sequence of comma-separated numbers and/or ranges of numbers separated by a "-" (e.g. "**linux[0-64,128]**", or "**lx[15,18,32-33]**"). Presently the numeric range must be the last characters in the node name (e.g. "**unit[0-31]rack1**" is invalid).

The node configuration specifies the following information:

#### **nodeName**

The Name that SLURM uses to refer to a node. Typically this would be the string that **/bin/hostname -s** returns; however, it may be an arbitrary string if **NodeHostname** is specified. If the **nodeName** is "DEFAULT", the values specified with that record will apply to subsequent node specifications unless explicitly set to other values in that node record or replaced with a different set of default values. For architectures in which the node order is significant, nodes will be considered consecutive in the order defined. For example, if the configuration for **nodeName=charlie** immediately follows the configuration for **nodeName=baker** they will be considered adjacent in the computer.

#### **NodeHostname**

The string that **/bin/hostname -s** returns. A node range expression can be used to specify a set of nodes. If an expression is used, the number of nodes identified by **NodeHostname** on a line in the configuration file must be identical to the number of nodes identified by **nodeName**. By default, the **NodeHostname** will be identical in value to **nodeName**.

#### **NodeAddr**

Name by which a node should be referred when establishing a communications path. This name will be used as an argument to the **gethostbyname()** function for identification. If a node range expression is used to designate multiple nodes, they must exactly match the entries in the **nodeName** (e.g. "**nodeName=lx[0-7] NodeAddr=elx[0-7]**"). **NodeAddr** may also contain IP addresses. By default, the **NodeAddr** will be identical in value to **nodeName**.



### DownNodes

Any node name, or list of node names, from the **NodeName=** specifications. The **DownNodes=** configuration permits marking certain nodes as being in a **DOWN** or **DRAINED** state without altering the permanent configuration information listed under a **NodeName=** specification.

### Feature

A comma-delimited list of arbitrary strings indicative of some characteristic associated with the node. There is no value associated with a feature at this time, a node either has a feature or it does not. If desired, a feature may contain a numeric component indicating, for example, processor speed. By default a node has no features.

### Procs

Number of processors on the node (e.g. "2"). The default value is 1.

### RealMemory

Size of real memory on the node in MegaBytes (e.g. "2048"). The default value is 1.

### Reason

Identifies the reason for a node being in state **DOWN** or **DRAINED** or **DRAINING**. Use quotes to enclose a reason having more than one word.

### State

State of the node with respect to the initiation of user jobs. Acceptable values are **BUSY**, **DOWN**, **DRAINED**, **DRAINING**, **IDLE**, and **UNKNOWN**.

- **BUSY** indicates the node has been allocated work and should not be used in the configuration file.
- **DOWN** indicates the node failed and is unavailable to be allocated work.
- **DRAINED** indicates the node was configured unavailable to be allocated work and is presently not performing any work.
- **DRAINING** indicates the node is unavailable to be allocated new work, but is completing the processing of a job.
- **IDLE** indicates the node available to be allocated work, but has none at present
- **UNKNOWN** indicates the node's state is undefined, but will be established when the SLURMD daemon on that node registers.

The default value is **UNKNOWN**.

### TmpDisk

Total size of temporary disk storage in **TmpFS** in MegaBytes (e.g. "16384"). **TmpFS** (for "Temporary File System") identifies the location that jobs should use for temporary storage. Note that this does not indicate the amount of free space available to the user on the node, only the total file system size. The system administration should ensure that this file system is purged as needed, thus allowing user jobs to have access to most of this space. The **Prolog** and/or **Epilog** programs (specified in the configuration file) might be used to ensure that the file system is kept clean. The default value is 1.

## Weight

The priority of the node for scheduling purposes. All things being equal, jobs will be allocated the nodes with the lowest weight that satisfies their requirements. For example, a heterogeneous collection of nodes might be placed into a single partition for greater system utilization, responsiveness and capability. It would be preferable to allocate smaller memory nodes rather than larger memory nodes if either will satisfy a job's requirements. The units of weight are arbitrary, but larger weights should be assigned to nodes with more processors, memory, disk space, higher processor speed, etc. Weight is an integer value with a default value of 1.

### 6.2.1.3 Partition Configuration Parameters

The partition configuration permits different job limits or access controls to be established for various groups (or partitions) of nodes. Nodes may be in more than one partition, making partitions serve as general-purpose queues. For example, the same set of nodes may be put into two different partitions, each with different constraints (time limit, job sizes, groups allowed to use the partition, etc.). Jobs are allocated resources within a single partition.

The partition configuration file contains the following information:

#### AllowGroups

Comma-separated list of group IDs that may execute jobs in the partition. If at least one group associated with the user attempting to execute the job is in AllowGroups, he will be permitted to use this partition. Jobs executed as user root can use any partition without regard to the value of AllowGroups. If user root attempts to execute a job as another user (e.g. using SRUN'S `--uid` option), this other user must be in one of the groups identified by **AllowGroups** for the job to successfully execute. The default value is "ALL".

#### Default

If this keyword is set, jobs submitted without a partition specification will utilize this partition. Possible values are "YES" and "NO". The default value is "NO".

#### Hidden

Specifies if the partition and its jobs are to be hidden by default. Hidden partitions will by default not be reported by the SLURM APIs or commands. Possible values are "YES" and "NO". The default value is "NO".

#### RootOnly

Specifies if only user ID zero (i.e. user root) may allocate resources in this partition. User root may allocate resources for any other user, but the request must be initiated by user root. This option can be useful for a partition which is to be managed by some external entity (e.g. a higher-level job manager) and prevents users from directly using those resources. Possible values are "YES" and "NO". The default value is "NO".

#### MaxNodes

Maximum count of nodes that may be allocated to any single job. The default value is "UNLIMITED", which is represented internally as -1. This limit does not apply to jobs executed by **SlurmUser** or user root.

### MaxTime

Maximum wall-time limit for any job in minutes. The default value is "UNLIMITED", which is represented internally as -1. This limit does not apply to jobs executed by **SlurmUser** or user root.

### MinNodes

Minimum count of nodes that may be allocated to any single job. The default value is 1. This limit does not apply to jobs executed by **SlurmUser** or user root.

### Nodes

Comma-separated list of nodes that are associated with this partition. Node names may be specified using the node range expression syntax described above. A blank list of nodes (i.e. "Nodes=") can be used if one wants a partition to exist, but have no resources (possibly on a temporary basis).

### PartitionName

Name by which the partition may be referenced (e.g. "Interactive"). Users can specify this name when submitting jobs.

### Shared

Ability of the partition to execute more than one job at a time on each node. Shared nodes will offer unpredictable performance for application programs, but can provide higher system utilization and responsiveness than otherwise possible. Possible values are **FORCE**, **YES**, and **NO**.

- **FORCE** makes all nodes in the partition available for sharing without user means of disabling it.
- **YES** makes nodes in the partition available for sharing if and only if the individual jobs permit sharing (see the SRUN "**--shared**" option).
- **NO** makes nodes unavailable for sharing under all circumstances. The default value is **NO**.

### State

State of partition or availability for use. Possible values are **UP** or **DOWN**. The default value is **UP**.

## 6.2.2 slurm.conf Example Files

This section provides two examples of **slurm.conf** files.

### Example #1

```
ControlMachine=linux0
ControlAddr=linux0
SlurmctldLogFile=/var/log/slurm/slurmctld.log
SlurmdLogFile=/var/log/slurm/slurmd.log.%h
StateSaveLocation=/var/log/slurm/log_slurmctld
SlurmdSpoolDir=/var/log/slurm/log_slurmd/
SlurmUser=slurm
SlurmctldDebug=3      # default is 3
SlurmdDebug=3        # default is 3
SlurmctldPort=6817
```

```

SlurmdPort=6818
AuthType=auth/none
SelectType=select/linear
SchedulerType=sched/builtin # default is sched/builtin
#JobCompType=jobcomp/filetxt # default is jobcomp/none
#JobCompLoc=/var/log/slurm/slurm.job.log
SwitchType=switch/none
ProctrackType=proctrack/pgid
#JobAcctType=jobacct/linux # default is jobacct/none
#JobAcctLogFile=/var/log/slurm/slurm_acct.log

FastSchedule=1 # default is `1'
FirstJobid=1000 # default is `1'
ReturnToService=0 # default is `0'
MpiDefault=none # default is "none"

JobCredentialPrivateKey=/etc/slurm/slurm.key
JobCredentialPublicCertificate=/etc/slurm/slurm.cert

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

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

```

## Example #2

```

#
# Sample slurm.conf for mcr.llnl.gov
#
ControlMachine=mcri ControlAddr=emcri
BackupMachine=mcrj BackupAddr=emcrj
#
AuthType=auth/munge
Epilog=/usr/local/slurm/etc/epilog
FastSchedule=1
JobCompLoc=/var/tmp/jette/slurm.job.log
JobCompType=jobcomp/filetxt
JobCredentialPrivateKey=/usr/local/etc/slurm.key
JobCredentialPublicCertificate=/usr/local/etc/slurm.cert
PluginDir=/usr/local/slurm/lib/slurm
Prolog=/usr/local/slurm/etc/prolog
SchedulerType=sched/backfill
SelectType=select/linear
SlurmUser=slurm
SlurmctldPort=7002
SlurmctldTimeout=300
SlurmdPort=7003
SlurmdSpoolDir=/var/tmp/slurmd.spool
SlurmdTimeout=300
StateSaveLocation=/tmp/slurm.state
SwitchType=switch/elan
TreeWidth=50
#

```

```

# Node Configurations
#
NodeName=DEFAULT Procs=2 RealMemory=2000 TmpDisk=64000
State=UNKNOWN
NodeName=mcr[0-1151] NodeAddr=emcr[0-1151]
#
# Partition Configurations
#
PartitionName=DEFAULT State=UP
PartitionName=pdebug Nodes=mcr[0-191] MaxTime=30 MaxNodes=32
Default=YES
PartitionName=pbatch Nodes=mcr[192-1151]

```

## 6.2.3 SCONTROL – Managing the SLURM Configuration

**SCONTROL** manages available nodes (for example, by "draining" jobs from a node or partition to prepare it for servicing). It is also used to manage the **SLURM** configuration and the properties assigned to nodes, node partitions and other SLURM-controlled system features.



### Note:

Most **SCONTROL** options and commands can only be used by System Administrators. Some **SCONTROL** commands *report* useful configuration information or manage job *checkpoints*, and any user can benefit from invoking them appropriately.

### NAME

SCONTROL - Used to view and modify SLURM configuration and state.

### SYNOPSIS

```
SCONTROL [OPTIONS...] [COMMAND...]
```

### DESCRIPTION

SCONTROL is used to view or modify the SLURM configuration including: job, job step, node, partition, and overall system configuration. Most of the commands can only be executed by user root. If an attempt to view or modify configuration information is made by an unauthorized user, an error message will be printed and the requested action will not occur. If no command is entered on the execute line, SCONTROL will operate in an interactive mode and prompt for input. It will continue prompting for input and executing commands until explicitly terminated. If a command is entered on the execute line, SCONTROL will execute that command and terminate. All commands and options are case-insensitive, although node names and partition names are case-sensitive (node names "LX" and "lx" are distinct). Commands can be abbreviated to the extent that the specification is unique.

### 6.2.3.1

## OPTIONS

### **-a, --all**

When the show command is used, then it displays all partitions, their jobs and jobs steps. This causes information to be displayed about partitions that are configured as hidden and partitions that are unavailable to user's group.

### **-h, --help**

Print a help message describing the usage of SCONTROL.

### **--hide**

Do not display information about hidden partitions, their jobs and job steps. By default, neither partitions that are configured as hidden nor those partitions unavailable to a user's group will be displayed (i.e. this is the default behavior).

### **-o, --oneline**

Print information one line per record.

### **-q, --quiet**

Print no warning or informational messages, only fatal error messages.

### **-v, --verbose**

Print detailed event logging. This includes time-stamps on data structures, record counts, etc.

### **-V, --version**

Print version information and exit.

### 6.2.3.2

## Commands

### **all**

Show all partitions, their jobs and jobs steps. This causes information to be displayed about partitions that are configured as hidden and partitions that are unavailable to a user's group.

### **abort**

Instruct the SLURM controller to terminate immediately and generate a core file.

### **checkpoint CKPT\_OP ID**

Perform a checkpoint activity on the job step(s) with the specified identification. CKPT\_OP may take one of the following values: "disable" (disable future checkpoints), "enable" (enable future checkpoints), "able" (test if presently not disabled, report start time if checkpoint in progress), "create" (create a checkpoint and continue the job step), "vacate" (create a checkpoint and terminate the job step), "error" (report the result for the last checkpoint request, error code and message), or "restart" (restart execution of the previously checkpointed job steps). ID can be used to identify a specific job (e.g. "<job\_id>", which applies to all of its existing steps) or a specific job step (e.g. "<job\_id>.<step\_id>").

**completing**

Display all jobs in a COMPLETING state along with associated nodes in either a COMPLETING or DOWN state.

**delete SPECIFICATION**

Delete the entry with the specified SPECIFICATION. The only supported SPECIFICATION presently is of the form PartitionName=<name>.

**exit**

Terminate the execution of SCONTROL.

**help**

Display a description of SCONTROL options and commands.

**hide**

Do not display partition, job, or job-step information for partitions that are configured as hidden or partitions that are unavailable to the user's group. This is the default behavior.

**oneliner**

Print information one line per record.

**pidinfo PROC\_ID**

Print the SLURM job id and scheduled termination time corresponding to the supplied process id, PROC\_ID, on the current node. This will only work for processes that SLURM spawns and their descendants.

**ping**

Ping the primary and secondary SLURMCTLD daemon and report if they are responding.

**quiet**

Print no warning or informational messages, only fatal error messages.

**quit**

Terminate the execution of SCONTROL.

**reconfigure**

Instruct all SLURM daemons to re-read the configuration file. This command does not restart the daemons. This mechanism would be used to modify configuration parameters (**Epilog**, **Prolog**, **SlurmctldLogFile**, **SlurmdLogFile**, etc.), register the physical addition or removal of nodes from the cluster or recognize the change of a node's configuration, such as the addition of memory or processors. The SLURM controller (**SLURMCTLD**) forwards the request to all other daemons (**SLURMD** daemon on each compute node). Running jobs continue execution. Most configuration parameters can be changed by just running this command, however, SLURM daemons should be shutdown and restarted if any of the following parameters are to be changed: **AuthType**, **BackupAddr**, **BackupController**, **ControlAddr**, **ControlMach**, **PluginDir**, **StateSaveLocation**, **SlurmctldPort** or **SlurmdPort**.

**resume job\_id**

Resume a previously suspended job.

**show ENTITY ID**

Display the state of the specified entity with the specified identification. ENTITY may be config, daemons, job, node, partition or step. ID can be used to identify a specific element of the identified entity: the configuration parameter name, job ID, node name, partition name, or job step ID for entities config, job, node, partition, and step respectively. Multiple node names may be specified using simple node range expressions (e.g. "lx[10-20]"). All other ID values must identify a single element. The job step ID is of the form "job\_id.step\_id", (e.g. "1234.1"). By default, all elements of the entity type specified are printed.

**shutdown**

Instruct all SLURM daemons to save current state and terminate. The SLURM controller (SLURMCTLD) forwards the request all other daemons (SLURMD daemon on each compute node).

**suspend job\_id**

Suspend a running job. Use the resume command to resume its execution. User processes must stop on receipt of SIGSTOP signal and resume upon receipt of SIGCONT for this operation to be effective. Not all architectures and configurations support job suspension.

**update SPECIFICATION**

Update job, node or partition configuration per the supplied specification. SPECIFICATION is in the same format as the SLURM configuration file and the output of the show command described above. It may be desirable to execute the show command (described above) on the specific entity that is to be updated, use cut-and-paste tools to enter updated configuration values to the update. Note that while most configuration values can be changed using this command, not all can be changed using this mechanism. In particular, the hardware configuration of a node or the physical addition or removal of nodes from the cluster may only be accomplished through editing the SLURM configuration file and executing the reconfigure command (described above).

**verbose**

Print detailed event logging. This includes time-stamps on data structures, record counts, etc.

**version**

Display the version number of SCONTROL being executed.

**!!**

Repeat the last command executed.



### 6.2.3.3

## Specifications for the Update Command - Jobs

### **Account=<account>**

Account name to be changed for this job's resource use. Value may be cleared with blank data value, "Account=".

### **Contiguous=<yes | no>**

Set the job's requirement for contiguous (consecutive) nodes to be allocated. Possible values are "YES" and "NO".

### **Dependency=<job\_id>**

Defer job's initiation until specified job\_id completes. Cancel dependency with job\_id value of "0", "Dependency=0".

### **Features=<features>**

Set specified features to the job's required features on nodes. Multiple values may be comma separated if all features are required (AND operation) or separated by "|" if any of the specified features are required (OR operation). Value may be cleared with blank data value, "Features=".

### **JobId=<id>**

Identify the job to be updated. This specification is required.

### **MinMemory=<megabytes>**

Set the job's minimum real memory required per node to the specified value.

### **MinProcs=<count>**

Set the job's minimum number of processors per node to the specified value.

### **MinTmpDisk=<megabytes>**

Set the job's minimum temporary disk space required per node to the specified value.

### **Name=<name>**

Set the job's name to the specified value.

### **Partition=<name>**

Set the job's partition to the specified value.

### **Priority=<number>**

Set the job's priority to the specified value. Note that a job priority of zero prevents the job from ever being scheduled. By setting a job's priority to zero, it is held. Set the priority to a non-zero value to permit it to run.

### **Nice[=delta]**

Adjust job's priority to the specified value. Default value is 100.

### **ReqNodeList=<nodes>**

Set the job's list of required nodes. Multiple node names may be specified using simple node range expressions (e.g. "lx[10-20]"). Value may be cleared with blank data value, "ReqNodeList=".

**ReqNodes=<count>**

Set the job's count of required nodes to the specified value.

**ReqProcs=<count>**

Set the job's count of required processors to the specified value.

**Shared=<yes | no>**

Set the job's ability to share nodes with other jobs. Possible values are "YES" and "NO".

**StartTime=<time\_spec>**

Set the job's earliest initiation time. It accepts times of the form HH:MM:SS to run a job at a specific time of day (seconds are optional). (If that time is already past, the next day is assumed.) You may also specify midnight, noon, or teatime (4pm) and you can have a time-of-day suffixed with AM or PM for running in the morning or the evening. It is also possible to specify the day on which the job will be run, by giving a date in the form MMDDYY or MM/DD/YY or MM.DD.YY. It is also possible to give times, such as now + count time-units, where the time-units can be minutes, hours, days, or weeks and SLURM can be told to run the job today with the keyword today and to run the job tomorrow with the keyword tomorrow.

**Notes for date/time specifications:**

- Although the 'seconds' field of the HH:MM:SS time specification is allowed by the code, the poll time of the SLURM scheduler is not precise enough to guarantee dispatch of the job on the exact second. The job will be eligible to start on the next poll following the specified time. The exact poll interval depends on the SLURM scheduler (e.g. 60 seconds with the default **sched/builtin**).
- If no time (HH:MM:SS) is specified, the default is (00:00:00).
- If a date is specified without a year (e.g. MM/DD) then the current year is assumed, unless the combination of MM/DD and HH:MM:SS has already passed for that year, in which case the next year is used.

**TimeLimit=<minutes>**

Set the job's time limit to the specified value.

**Connection=<type>**

Reset the node connection type.

**Geometry=<geo>**

Reset the required job geometry.

**Rotate=<yes | no>**

Permit the job's geometry to be rotated. Possible values are "YES" and "NO".

### 6.2.3.4 Specifications for the Update Command - Nodes

**NodeName=<name>**

Identify the node(s) to be updated. Multiple node names may be specified using simple node range expressions (e.g. "lx[10-20]"). This specification is required.

**Reason=<reason>**

Identify the reason why the node is in a "DOWN" or "DRAINED" or "DRAINING" state. Use quotes to enclose a reason having more than one word.

**State=<state>**

Identify the state to be assigned to the node. Possible values are "NoResp", "DRAIN", "RESUME", "DOWN", "IDLE", "ALLOC", and "ALLOCATED". "RESUME" is not an actual node state, but it will return a DRAINED, DRAINING, or DOWN node to service, either IDLE or ALLOCATED state as appropriate. The "NoResp" state will only set the "NoResp" flag for a node without changing its underlying state.

### 6.2.3.5 Specifications for Update and Delete Commands - Partitions

**AllowGroups=<name>**

Identify the user groups that may use this partition. Multiple groups may be specified in a comma-separated list. To permit all groups to use the partition specify "AllowGroups=ALL".

**Default=<yes | no>**

Specify if this partition is to be used by jobs that do not explicitly identify a partition to use. Possible values are "YES" and "NO".

**Hidden=<yes | no>**

Specify if the partition and its jobs should be hidden from view. Hidden partitions will by default not be reported by SLURM APIs or commands. Possible values are "YES" and "NO".

**Nodes=<name>**

Identify the node(s) to be associated with this partition. Multiple node names may be specified using simple node range expressions (e.g. "lx[10-20]"). Note that jobs may only be associated with one partition at any time. Specify a blank data value to remove all nodes from a partition: "Nodes=".

**PartitionName=<name>**

Identify the partition to be updated. This specification is required.

**RootOnly=<yes | no>**

Specify if only allocation requests initiated by user root will be satisfied. This can be used to restrict control of the partition to some meta-scheduler. Possible values are "YES" and "NO".

**Shared=<yes | no | force>**

Specify if nodes in this partition can be shared by multiple jobs. Possible values are "YES", "NO", and "FORCE".

**State=<up | down>**

Specify if jobs can be allocated nodes in this partition. Possible values are "UP" and "DOWN". If a partition allocated nodes to running jobs, those jobs will continue execution even after the partition's state is set to "DOWN". The jobs must be explicitly canceled to force their termination.

#### **MaxNodes=<count>**

Set the maximum number of nodes that will be allocated to any single job in the partition. Specify a number or "INFINITE".

#### **MinNodes=<count>**

Set the minimum number of nodes that will be allocated to any single job in the partition.

### 6.2.3.6 ENVIRONMENT VARIABLES

Some SCONTROL options may be set via environment variables. These environment variables, along with their corresponding options, are listed below. (Note: Command-line options will always override these settings.)

SLURM_CONF	The location of the SLURM configuration file.
SCONTROL_ALL	-a, --all

### 6.2.3.7 SCONTROL EXAMPLE

```
# scontrol
scontrol: show part class
PartitionName=class TotalNodes=10 TotalCPUs=20 RootOnly=NO
  Default=NO Shared=NO State=UP MaxTime=30 Hidden=NO
  MinNodes=1 MaxNodes=2 AllowGroups=students
  Nodes=lx[0031-0040] NodeIndices=31,40,-1
scontrol: update PartitionName=class MaxTime=99 MaxNodes=4
scontrol: show job 65539
JobId=65539 UserId=1500 JobState=PENDING TimeLimit=100
  Priority=100 Partition=batch Name=job01 NodeList=(null)
  StartTime=0 EndTime=0 Shared=0 ReqProcs=1000
  ReqNodes=400 Contiguous=1 MinProcs=4 MinMemory=1024
  MinTmpDisk=2034ReqNodeList=lx[3000-3003]
  Features=(null) JobScript=/bin/hostname
scontrol: update JobId=65539 TimeLimit=200 Priority=500
scontrol: quit
```

## 6.2.4 Pam\_Slurm Module Configuration

This section describes how to use the **pam\_slurm** module. This module restricts access to Compute Nodes in a cluster where Simple Linux Utility for Resource Management (SLURM) is in use. Access is granted to root, any user with a SLURM-launched job currently running on the node, or any user who has allocated resources on the node according to the SLURM database.

Use of this module is recommended on any Compute Node where it is desirable to limit access to just those users who are currently scheduled to run jobs.

For `/etc/pam.d/` style configurations where modules reside in `/lib/security/`, add the following line to the PAM configuration file for the appropriate service(s) (eg, `/etc/pam.d/system-auth`):

```
account    required    /lib/security/pam_slurm.so
```

If it is necessary to always allow access for an administrative group (e.g., `wheel`), stack the `pam_access` module ahead of `pam_slurm`:

```
account    sufficient    /lib/security/pam_access.so
account    required    /lib/security/pam_slurm.so
```

Then edit the `pam_access` configuration file (`/etc/security/access.conf`):

```
+ :wheel:ALL
- :ALL:ALL
```

When access is denied because the user does not have an active job running on the node, an error message is returned to the application:

```
Access denied: user foo (uid=1313) has no active jobs.
```

This message can be suppressed by specifying the `no_warn` argument in the PAM configuration file.

## 6.3 Administrating Cluster Activity with SLURM

SLURM consists of two types of daemons.

- **SLURMCTLD** is sometimes called the "controller" daemon. It orchestrates **SLURM** activities, including queuing of job, monitoring node states, and allocating resources (nodes) to jobs. There is an optional backup controller that automatically assumes control in the event that the primary controller fails. The primary controller resumes control when it is restored to service. The controller saves its state to disk whenever there is a change. This state can be recovered by the controller at startup time. State changes are saved so that jobs and other states can be preserved when the controller moves (to or from a backup controller) or is restarted.

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

- The **SLURMD** daemon executes on all Compute nodes. It resembles a remote shell daemon which exports control to **SLURM**. Because **SLURMD** initiates and manages user jobs, it must execute as the user **root**.

### 6.3.1 Starting the Daemons

The **SLURM** daemons are initiated at node startup time, provided by the `/etc/init.d/slurm` script. If needed, the `/etc/init.d/slurm` script can be used to check the status of the daemon, start, startclean or stop the daemon on the node.

Once a valid configuration has been set up and installed, the **SLURM** controller, **SLURMCTLD**, should be started on the primary and backup control machines, and the **SLURM** compute node daemon, **SLURMD**, should be started on each compute server. The **SLURMD** daemons need to run as root for production use, but may be run as a user for testing purposes (obviously no jobs should be running as any other user in the configuration). The **SLURM** controller, **SLURMCTLD**, must be run as the configured **SlurmUser** (see the configuration file).

For testing purposes it may be prudent to start by just running **SLURMCTLD** and **SLURMD** on one node. By default, they execute in the background. Use the `-D` option for each daemon to execute them in the foreground and logging will be done to the terminal. The `-v` option will log events in more detail with more `v`'s increasing the level of detail (e.g. `-vvvvvv`). One window can be used to execute `slurmctld -D -vvvvv`, whilst `slurmd -D -vvvv` is executed in a second window. Errors such as "Connection refused" or "Node X not responding" may be seen when one daemon is operative and the other is being started. However, the daemons can be started in any order and proper communications will be established once both daemons complete initialization. A third window can be used to execute commands such as, `srun -N1 /bin/hostname`, to confirm functionality.

Another important option for the daemons is `-c` to clear the previous state information. Without the `-c` option, the daemons will restore any previously saved state information: node state, job state, etc. With the `-c` option all previously running jobs will be purged and the node state will be restored to the values specified in the configuration file. This means that a node configured down manually using the **SCONTROL** command will be returned to service unless also noted as being down in the configuration file. In practice, **SLURM** restarts with preservation consistently.

The `/etc/init.d/slurm` script can be used to start, startclean or stop the daemons for the node on which it is being executed.

## 6.3.2 SLURMCTLD (Controller Daemon)

### NAME

**SLURMCTLD** - The central management daemon of SLURM.

### SYNOPSIS

`slurmctld [OPTIONS...]`

### DESCRIPTION

SLURMCTLD is the central management daemon of SLURM. It monitors all other SLURM daemons and resources, accepts work (jobs), and allocates resources to those jobs. Given the critical functionality of SLURMCTLD, there may be a backup server to assume these functions in the event that the primary server fails.

### OPTIONS

`-c`

Clear all previous SLURMCTLD states from its last checkpoint. If not specified, previously running jobs will be preserved along with the state of **DOWN**, **DRAINED** and **DRAINING** nodes and the associated reason field for those nodes.

`-D`

Debug mode. Execute SLURMCTLD in the foreground with logging to stdout.

`-f <file>`

Read configuration from the specified file. See NOTE under ENVIRONMENT VARIABLES below.

`-h`

Help; print a brief summary of command options.

`-L <file>`

Write log messages to the specified file.

- v  
Verbose operation. Using more than one v (e.g., -vv, -vvv, -vvvv, etc.) increases verbosity.
- V  
Print version information and exit.

## ENVIRONMENT VARIABLES

The following environment variables can be used to override settings compiled into **SLURMCTLD**.

### SLURM\_CONF

The location of the SLURM configuration file. This is overridden by explicitly naming a configuration file in the command line.



#### Note:

It may be useful to experiment with different **SLURMCTLD**-specific configuration parameters using a distinct configuration file (e.g. timeouts). However, this special configuration file will not be used by the **SLURMD** daemon or the **SLURM** programs, unless each of them is specifically told to use it. To modify communication ports, the location of the temporary file system, or other parameters used by other **SLURM** components, change the common configuration file, **slurm.conf**.

## 6.3.3 SLURMD (Compute Node Daemon)

### NAME

**SLURMD** - The compute node daemon for SLURM.

### SYNOPSIS

**slurmd** [OPTIONS...]

### DESCRIPTION

**SLURMD** is the compute node daemon of SLURM. It monitors all tasks running on the compute node, accepts work (tasks), launches tasks, and kills running tasks upon request.

### OPTIONS

- c  
Clear system locks as needed. This may be required if **SLURMD** terminated abnormally.
- D  
Run **SLURMD** in the foreground. Error and debug messages will be copied to **stderr**.



- M**  
Lock **SLURMD** pages into system memory using **mlockall** to disable paging of the **SLURMD** process. This may help in cases where nodes are marked **DOWN** during periods of heavy swap activity. If the **mlockall** system call is not available, an error will be printed to the log and **SLURMD** will continue as normal.
- h**  
Help; print a brief summary of command options.
- f <file>**  
Read configuration from the specified file. See NOTES below.
- l <file>**  
Write log messages to the specified file.
- v**  
Verbose operation. Using more than one v (e.g., -vv, -vvv, -vvvv, etc.) increases verbosity.
- V**  
Print version information and exit.

## ENVIRONMENT VARIABLES

The following environment variables can be used to override settings compiled into **SLURMD**.

### **SLURM\_CONF**

The location of the **SLURM** configuration file. This is overridden by explicitly naming a configuration file on the command line.



#### **Note:**

It may be useful to experiment with different **SLURMD**-specific configuration parameters using a distinct configuration file (e.g. timeouts). However, this special configuration file will not be used by the **SLURMD** daemon or the **SLURM** programs, unless each of them is specifically told to use it. To modify communication ports, the location of the temporary file system, or other parameters used by other **SLURM** components, change the common configuration file, **slurm.conf**.

## 6.3.4 Scheduler Support

The scheduler used by **SLURM** is controlled by the **SchedType** configuration parameter. This is meant to control the relative importance of pending jobs. **SLURM**'s default scheduler is **FIFO** (First-In First-Out). A backfill scheduler plug-in is also available. Backfill scheduling will initiate a lower-priority job if doing so does not delay the expected initiation time of higher priority jobs; essentially using smaller jobs to fill holes in the resource allocation plan. **SLURM** also supports a plug-in for use of the **Maui** Scheduler, which offers sophisticated scheduling algorithms. Motivated users can even develop their own scheduler plug-in if so desired.

## 6.3.5 Node Selection

The node selection mechanism used by SLURM is controlled by the **SelectType** configuration parameter. If you want to execute multiple jobs per node, but apportion the processors, memory and other resources, the **cons\_res** (consumable resources) plug-in is recommended. If you tend to dedicate entire nodes to jobs, the **linear** plug-in is recommended.

## 6.3.6 Logging

SLURM uses the **syslog** function to record events. It uses a range of importance levels for these messages. Be certain that your system's **syslog** functionality is operational.

## 6.3.7 Corefile Format

SLURM is designed to support generating a variety of core file formats for application codes that fail (see the **--core** option of the **srun** command).

## 6.3.8 Security

Unique job credential keys for each site should be created using the **openssl** program **openssl must be used (not ssh-genkey) to construct these keys**. An example of how to do this is shown below.

Specify file names that match the values of **JobCredentialPrivateKey** and **JobCredentialPublicCertificate** in the configuration file. The **JobCredentialPrivateKey** file must be readable only by **SlurmUser**. The **JobCredentialPublicCertificate** file must be readable by all users. Both files must be available on all nodes in the cluster. These keys are used by **slurmctl** to construct a job credential, which is sent to **srun** and then forwarded to **slurmd** to initiate job steps.

```
> openssl genrsa -out /path/to/private/key 1024
> openssl rsa -in /path/to/private/key -pubout -out /path/to/public/key
```

## 6.3.9 SLURM Cluster Administration Examples

**SCONTROL** may be used to print all system information and modify most of it.

Only a few examples are shown below. Please see the **SCONTROL** man page for full details. The commands and options are all case insensitive.

- Print detailed state of all jobs in the system.

```
adev0: scontrol
scontrol: show job
JobId=475 UserId=bob(6885) Name=sleep JobState=COMPLETED
  Priority=4294901286 Partition=batch BatchFlag=0
  AllocNode:Sid=adevi:21432 TimeLimit=UNLIMITED
  StartTime=03/19-12:53:41 EndTime=03/19-12:53:59
  NodeList=adev8 NodeListIndecies=-1
```

```
ReqProcs=0 MinNodes=0 Shared=0 Contiguous=0
MinProcs=0 MinMemory=0 Features=(null) MinTmpDisk=0
ReqNodeList=(null) ReqNodeListIndecies=-1
```

```
JobId=476 UserId=bob(6885) Name=sleep JobState=RUNNING
Priority=4294901285 Partition=batch BatchFlag=0
AllocNode:Sid=adevi:21432 TimeLimit=UNLIMITED
StartTime=03/19-12:54:01 EndTime=NONE
NodeList=adev8 NodeListIndecies=8,8,-1
ReqProcs=0 MinNodes=0 Shared=0 Contiguous=0
MinProcs=0 MinMemory=0 Features=(null) MinTmpDisk=0
ReqNodeList=(null) ReqNodeListIndecies=-1
```

- Print the detailed state of job 477 and change its priority to zero. A priority of zero prevents a job from being initiated (it is held in "pending" state).

```
adev0: scontrol
scontrol: show job 477
JobId=477 UserId=bob(6885) Name=sleep JobState=PENDING
Priority=4294901286 Partition=batch BatchFlag=0
more data removed....
scontrol: update JobId=477 Priority=0
```

- Print the state of node adev13 and drain it. To drain a node, specify a new state of **DRAIN**, **DRAINED**, or **DRAINING**. SLURM will automatically set it to the appropriate value of either **DRAINING** or **DRAINED** depending on whether the node is allocated or not. Return it to service later.

```
adev0: scontrol
scontrol: show node adev13
NodeName=adev13 State=ALLOCATED CPUs=2 RealMemory=3448
TmpDisk=32000
Weight=16 Partition=debug Features=(null)
scontrol: update NodeName=adev13 State=DRAIN
scontrol: show node adev13
NodeName=adev13 State=DRAINING CPUs=2 RealMemory=3448
TmpDisk=32000
Weight=16 Partition=debug Features=(null)
scontrol: quit
Later
adev0: scontrol
scontrol: show node adev13
NodeName=adev13 State=DRAINED CPUs=2 RealMemory=3448
TmpDisk=32000
Weight=16 Partition=debug Features=(null)
scontrol: update NodeName=adev13 State=IDLE
```

- Reconfigure all SLURM daemons on all nodes. This should be done after changing the SLURM configuration file.

```
adev0: scontrol reconfig
```

- Print the current SLURM configuration. This also reports if the primary and secondary controllers (slurmctld daemons) are responding. To just see the state of the controllers, use the command ping.

```
adev0: scontrol show config
Configuration data as of 03/19-13:04:12
AuthType      = auth/munge
BackupAddr    = eadevj
```

```

BackupController = adevj
ControlAddr      = eadevi
ControlMachine  = adevi
Epilog          = (null)
FastSchedule    = 1
FirstJobId      = 1
InactiveLimit   = 0
JobCompLoc      = /var/tmp/jette/slurm.job.log
JobCompType     = jobcomp/filetxt
JobCredPrivateKey = /etc/slurm/slurm.key
JobCredPublicKey = /etc/slurm/slurm.cert
KillWait        = 30
MaxJobCnt       = 2000
MinJobAge       = 300
PluginDir       = /usr/lib/slurm
Prolog          = (null)
ReturnToService = 1
SchedulerAuth   = (null)
SchedulerPort   = 65534
SchedulerType   = sched/backfill
SlurmUser       = slurm(97)
SlurmctldDebug  = 4
SlurmctldLogFile = /tmp/slurmctld.log
SlurmctldPidFile = /tmp/slurmctld.pid
SlurmctldPort   = 7002
SlurmctldTimeout = 300
SlurmdDebug     = 65534
SlurmdLogFile  = /tmp/slurmd.log
SlurmdPidFile  = /tmp/slurmd.pid
SlurmdPort     = 7003
SlurmdSpoolDir = /tmp/slurmd
SlurmdTimeout  = 300
TreeWidth      = 50
JobAcctLogFile = /tmp/jobacct.log
JobAcctFrequency = 5
JobAcctType    = jobacct/linux
SLURM_CONFIG_FILE = /etc/slurm/slurm.conf
StateSaveLocation = /usr/local/tmp/slurm/adev
SwitchType     = switch/elan
TmpFS          = /tmp
WaitTime       = 0

```

Slurmctld(primary/backup) at adevi/adevj are UP/UP

- Shutdown all SLURM daemons on all nodes.

```
adev0: scontrol shutdown
```

---

## Chapter 7. Batch Management with PBS Professional

PBS Professional is the professional version of the Portable Batch System (PBS), a flexible resource and workload management system, originally developed to manage aerospace computing resources at NASA.

PBS is a distributed workload management system which has three primary roles:

### Queuing

The collecting together of jobs or tasks to be run on a computer. Users submit tasks or jobs to the resource management system which places them in a queue until the system is ready to run them.

### Scheduling

The process of selecting which jobs to run, where and when, according to predetermined policies. Sites balance competing needs and goals on the system(s) in order to maximize the efficient use of resources (both computer time and people time).

### Monitoring

The act of tracking and reserving system resources and enforcing usage policy. This covers both user-level and system-level monitoring, as well as monitoring the jobs that are being run. Tools are provided to help the human monitoring of the PBS system as well.



See the PBS Professional 9.0 *Administrator's Guide* (on the **PBS Pro CD-ROM** delivered for clusters which use **PBS PRO**) for more detailed information on using PBS PRO, including descriptions of some of the different configurations possible, with examples, plus descriptions of the PBS PRO Administrator commands.

This chapter describes some specific details which apply to **BAS4 for Xeon** clusters.

## 7.1 Pre-requisites



### Important

**SLURM** should not run on the same clusters as **PBS Professional**. If necessary deactivate **SLURM** by running the command `chkconfig -- level 345 slurm off` on the Management Node and on all the Compute Nodes.

1. The root user, administrator, should have direct access to all the Compute Nodes from the Management Node, and vice versa, without having to use a password. `ssh` is used to protect this access, see *section 2.5.1.* in this manual for more information.

## 7.2 Post Installation checks

The `/etc/pbs.conf` file will have been created automatically during the installation of PBS PRO. This will contain the `PBS_EXEC` path (`/usr/pbs` by default) and the `PBS_HOME` directory (`/var/spool/PBS` by default).



See the *Configuring Administration Software* step in Chapter 2 of the **BAS4 for Xeon Installation and Configuration Guide** for more information about the installation and configuration of PBS Pro.

### 7.2.1 Checking the status of the PBS daemons

Run the following command on the Management and Compute Node to check the status of the PBS daemons

```
/etc/init.d/pbs status
```

On the Management Node output similar to that below should appear:

```
pbs_server is pid xxxx
pbs_sched is pid xxxx
```

On the Computes Nodes output similar to that below should appear:

```
pbs_mom is pid xxxx
```

### 7.2.2 Adding a Node to the Initial Cluster Configuration

Use the `qmgr` option, as below, to add a Compute Node to the list of Compute Nodes for a cluster:

```
/usr/pbs/bin/qmgr -c "create node <node_name>"
```

Use the follow command to verify that the node has been created and added to the Compute Node list:

```
/usr/pbs/bin/pbsnodes -a
```

## 7.3 Useful Commands for PBS Professional

The following commands, which are in the `/usr/pbs/bin` directory, may be used to test that **PBS Professional** is up and running correctly:

#### **pbsnodes -a**

Used to display the status of the nodes in cluster

#### **qsub**

Used to submit a job

**qdel**

Used to delete a job

**qstat**

Used to display the job, queue and server status

**Tracejob**

Used to extract job info from the log files



See the **PBS Professional** *Administrator's Guide* and *User's Guide* included on the PBS Pro CD ROM for more detailed information on these and on other commands.

## 7.4 Essential configuration settings for XBAS4 for Xeon clusters

This section describes some essential configuration settings which are required to ensure that **PBS PRO** runs smoothly on **XBAS4 for Xeon** clusters.

### 7.4.1 MPIBull2 and PBS Pro for all clusters (InfiniBand and Ethernet)

To use **MPIBull2** with **PBS Pro** run the following commands on both the Management Node and on all the Compute Nodes:

```
cd /usr/pbs/bin
```

```
pbsrun_wrap /opt/mpi/mpibull2-<version>/bin/mpirun pbsrun.mpich2
```

This will give output similar to that below:

```
pbsrun_wrap: EXECUTED: "mv /opt/mpi/mpibull2-<version>/bin/mpirun
/opt/mpi/mpibull2-1.2.1-4.t/bin/mpirun.actual"
```

```
pbsrun_wrap: EXECUTED: "cp /usr/pbs/bin/pbsrun
/usr/pbs/bin/pbsrun.mpich2"
```

```
pbsrun_wrap: EXECUTED: "chmod 755 /usr/pbs/bin/pbsrun.mpich2"
```

```
pbsrun_wrap: EXECUTED: "ln -s /usr/pbs/bin/pbsrun.mpich2
/opt/mpi/mpibull2-<version>/bin/mpirun"
```

```
pbsrun_wrap: EXECUTED: "ln -s /opt/mpi/mpibull2-
<version>/bin/mpirun.actual /usr/pbs/lib/MPI/pbsrun.mpich2.link"
```

```
pbsrun_wrap: EXECUTED: "chmod
644/usr/pbs/lib/MPI/pbsrun.mpich2.init"
```

## 7.4.2 MPIBull2 and InfiniBand

To ensure that MPIBULL2 runs correctly on **InfiniBand** clusters, using the **ibmr\_gen2** device, the PBS launching script will need to be modified on all the Compute Nodes. This is done as follows:

1. Stop PBS on all the Compute Nodes:

```
pdsh -w <node_list> /etc/init.d/pbs stop
```

2. Edit the PBS launching script on each Compute Node:

```
vi /etc/init.d/pbs
```

3. Edit the **start\_pbs()** function on this file by adding the following line:

```
ulimit -l unlimited
```

4. Restart PBS on all the Compute Nodes

```
pdsh -w <node_list> /etc/init.d/pbs start
```



See the **PBS Professional Administrator's Guide** on the **PBS PRO CD-ROM** for more details about the configuration settings



---

## Chapter 8. Monitoring with NovaScale Master - HPC Edition

**NovaScale Master - HPC Edition** provides the monitoring functions for Bull HPC systems. It relies on Nagios and Ganglia open source software. Nagios is used to monitor the operating status for the different components of the cluster. Ganglia collects performance statistics for each cluster node and displays them graphically on a cluster scale. The status of a large number of elements can be displayed.

This chapter covers the following topics:

- 8.1 *Launching NovaScale Master - HPC Edition*
- 8.2 *Access Rights*
- 8.3 *Hosts, Services and Contacts for Nagios*
- 8.4 *Using NovaScale Master - HPC Edition*
- 8.5 *Map Button*
- 8.6 *Status Button*
- 8.7 *Alerts Button*
- 8.8 *Storage Overview*
- 8.9 *Shell*
- 8.10 *Monitoring the Performance - Ganglia Statistics*
- 8.11 *Group Performance View*
- 8.12 *Global Performance View*
- 8.13 *Configuring and Modifying Nagios Services*
- 8.14 *General Nagios Services*
- 8.15 *Management Node Nagios Services*
- 8.16 *Ethernet Switch Services*
- 8.17 *More Nagios Information*

## 8.1 Launching NovaScale Master - HPC Edition



### Note:

The cluster database (**ClusterDB**) must be running before starting monitoring. See the *Cluster Data Base Management* Chapter.

1. If necessary restart the **gmond** and **gmetad** services:

```
service gmond restart
service gmetad restart
```

2. Start the monitoring service:

```
service nagios start
```

3. Start Mozilla and go to the following URL:

<http://<ManagementNode>/NSMaster/>



### Note:

Mozilla is the mandatory navigator for **NovaScale Master – HPC Edition**

## 8.2 Access Rights

### 8.2.1 Administrator Access Rights

By default the Administrator will use the following login and password:

login: **nagios**  
password: **nagios**

The graphical interface for monitoring opens, see Figure 8-1.

The Administrator will be able to enter host and service commands via the interface, whereas an ordinary user will only be able to consult the interface.

### 8.2.2 Standard User Access Rights

By default the ordinary user will use the following login and password:

login: **guest**  
password: **guest**

### 8.2.3 Adding Users and Changing Passwords

The `htpasswd` command is used to create new user names and passwords.

To create additional users for the graphical interface, do as follows:

1. Enter the following command:

```
htpasswd /etc/nagios/htpasswd.users <login>
```

This command will prompt you for a password for each new user, and will then ask you to confirm the password.

2. You must also:
  - a. Define the user in the `/opt/NSMaster/core/share/etc/rbm/missions.xml` file (either as an Administrator or as an ordinary user).
  - b. Add the password to the `/opt/NSMaster/core/etc/htpasswd.users` file.

To change the password for an existing user, do as follows:

1. Enter the following command:

```
htpasswd /etc/nagios/htpasswd.users <login>
```

Enter and confirm the new password when prompted.

2. Change the password listed for the user in the `/opt/NSMaster/core/etc/htpasswd.users` file.



**Note:** Some of these steps have to be done as the root user.



See the NS Master documentation for more information on adding users and on account management.

## 8.3 Hosts, Services and Contacts for Nagios

Nagios defines two entities: **hosts** and **services**.

A **host** is any physical server, workstation, device etc. that resides on a network.

A **host group** definition is used to group one or more hosts together for display purposes in the CGIs.

A **service** definition is used to identify a "service" that runs on a host. The term "service" is used very loosely. It can mean an actual service that runs on the host (POP, SMTP, HTTP, etc.) or some other type of metric associated with the host (response to a ping, number of logged in users, free disk space, etc.).



**Note:**

NovaScale Master – HPC Edition will display the services specific to each host when the host is selected within the different parts of the NovaScale Master – HPC Edition interface.

A **contact** definition is used to identify someone who should be contacted in the event of a problem on your network.

A **contact group** definition is used to group one or more contacts together for the purpose of sending out alert/recovery notifications. When a host or service has a problem or recovers, Nagios will find the appropriate contact groups to send notifications to, and notify all contacts in these contact groups. This may sound complex, but for most people it doesn't have to be. It does, however, allow for flexibility in determining who gets notified for particular events.

For more information on these definitions and the arguments and the directives which may be used to format the definitions see:

[http://nagios.sourceforge.net/docs/2\\_0/](http://nagios.sourceforge.net/docs/2_0/)

Alternatively, select the **Documentation** link on the Nova Scale Master opening screen or by selecting the Documentation button in the title bar.

## 8.4 Using NovaScale Master - HPC Edition

The graphical interface of NovaScale Master - HPC Edition is shown inside a Web browser.

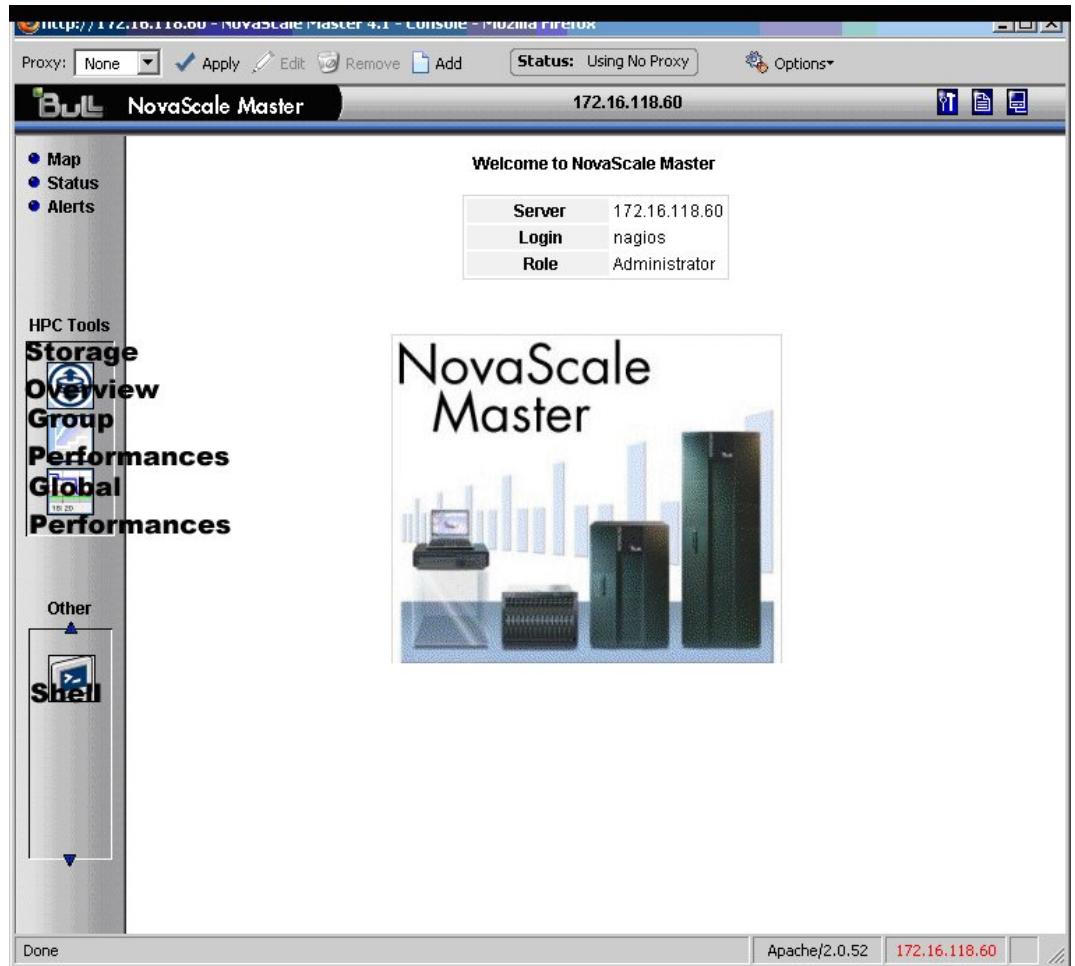


Figure 8-1. NovaScale Master - HPC Edition opening view

### 8.4.1 NovaScale Master - HPC Edition – View Levels

Initially the console will open and the administrator can then choose to view different types of monitoring information with a range of granularity levels either by clicking on the icons in the left hand vertical tool bar and then by clicking on the links in the different windows displayed. Using the links it is possible to descend to a deeper level for more detailed information for a particular host or service. For example, the Cabinet Rack map view in Figure 8-2 leads to the Rack View in Figure 8-3, which in turns leads to a more detailed Services view for the host selected in Figure 8-4.

## 8.5 Map Button

The Map Button is displayed at the top right hand side of the opening, when it is selected the drop down menu provides two options inside the main window. The map container can either be animated by **all status** or by **ping** views.

### 8.5.1 All Status Map View

The **all status** map view presents a chart of the cluster representing the various server rack cabinets in the room.

The color of each cabinet is determined by the component within it which is in the worst state.

By default, in addition to the view of the rack cabinets in the room, the Monitoring Problems window will appear at the bottom of the screen with a status for all hosts and services and the Availability Indicators view window will appear on the top right hand side – see Figure 8-2.

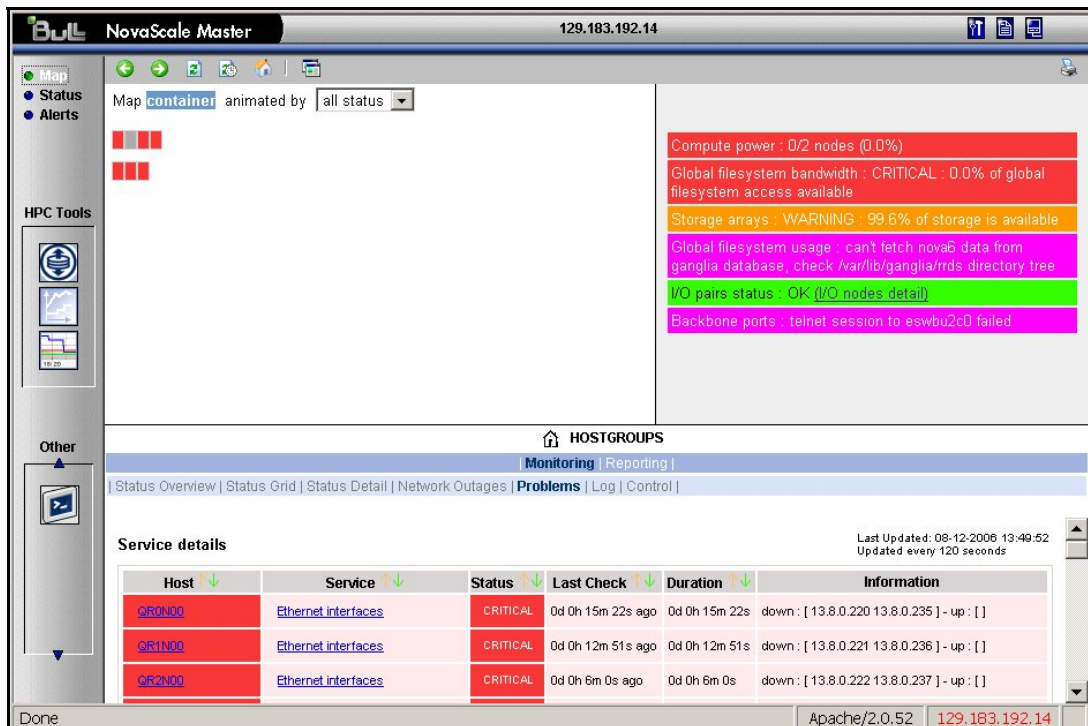


Figure 8-2. Map button all status opening view

When the cursor passes over a rack, information (its label, its type and the elements contained in the rack) about the rack is displayed. When the user clicks on a cabinet, a detailed view of the cabinet is displayed – see Rack view in Figure 8-3. This displays additional information, including its physical position and the services which are in a non-OK state.

## 8.5.2 Rack View

The Rack view details the contents of the rack: the nodes, their position inside the rack, their state, with links to its Alert history, etc. The list of the problems encountered is displayed at the bottom of the view – see Figure 8-3.

Clicking on a component displays a detailed view for it.

The screenshot shows the NovaScale Master console interface. The main content area displays the 'RACK-A1' view, which includes a rack diagram with nodes 'nova5' and 'nova6'. The 'nova5' node is highlighted in red, indicating a problem. The 'nova6' node is also highlighted in red. The rack diagram shows the position of the nodes: 'nova6' is at [line : A, column : 1] and 'nova5' is at [line : M, column : 1]. The 'Alerts' link is visible below the rack diagram.

On the right side of the rack view, there is a list of system metrics and alerts:

- Compute power : 0/2 nodes (0.0%)
- Global filesystem bandwidth : CRITICAL : 0.0% of global filesystem access available
- Storage arrays : WARNING : 99.6% of storage is available
- Global filesystem usage : can't fetch nova5 data from ganglia database, check /var/lib/ganglia/rds directory tree
- I/O pairs status : OK (I/O nodes detail)
- Backbone ports : telnet session to eswbu2c0 failed

Below the rack view, the 'HOSTGROUP: RACK-A1' section is visible, with tabs for 'Monitoring' and 'Reporting'. The 'Problems' tab is selected, showing a table of service details.

Host	Service	Status	Last Check	Duration	Information
nova5	Ethernet interfaces	CRITICAL	0d 0h 9m 59s ago	0d 0h 9m 59s	down : [ 13.2.0.6 192.20.0.6 ] - up : [ ]
	Hardware status	UNKNOWN	0d 0h 2m 12s ago	0d 0h 38m 4s	Timeout while polling PAP papu0c1
	Temperature	UNKNOWN	0d 0h 2m 12s ago	0d 0h 45m 56s	Timeout while running AusrNSMasterHWbin/rmsinfo.sh
nova6	Ethernet interfaces	CRITICAL	0d 0h 9m 59s ago	0d 0h 9m 59s	down : [ 13.2.0.7 192.20.0.7 ] - up : [ ]

Figure 8-3. Rack view with the problems window at the bottom

More detailed information regarding the hardware components and services associated with a host appear, when the host in the rack view is clicked on, in the top right pane of the Rack view. This leads to another pop up window which includes further information for the host and its services – see Figure 8-4.

## 8.5.3 Host Services detailed View

Clicking the **Host Status** or a **Service Status** link in this window displays more specific information for the component or service.

The control button in the middle of screen will provide the details for the Service monitoring information and the Service commands for the hardware component.

HOST: nova6 Monitoring | Reporting |

| Host Status | **Service Status** | Control |

Last Updated: 08-12-2006 11:22:39  
Updated every 90 seconds

Service	Status	Last Check	Duration	Information
<a href="#">Ethernet interfaces</a>	CRITICAL	0d 0h 7m 3s ago	0d 0h 17m 3s	down : [ 13.2.0.7 192.20.0.7 ] - up : [ ]
<a href="#">Hardware status</a>	UNKNOWN	0d 0h 1m 44s ago	0d 0h 45m 9s	Timeout while polling PAP papu0c1
<a href="#">IO status</a>	PENDING	0d 0h 57m 8s+ ago	0d 0h 57m 8s+	Service is not scheduled to be checked...
<a href="#">Log alerts</a>	PENDING	0d 0h 57m 8s+ ago	0d 0h 57m 8s+	Service is not scheduled to be checked...
<a href="#">NSDoctor</a>	PENDING	0d 0h 57m 8s+ ago	0d 0h 57m 8s+	Service is not scheduled to be checked...
<a href="#">Postbootchecker</a>	PENDING	0d 0h 57m 8s+ ago	0d 0h 57m 8s+	Service is not scheduled to be checked...
<a href="#">RMS status</a>	PENDING	0d 0h 57m 8s+ ago	0d 0h 57m 8s+	Service is not scheduled to be checked...
<a href="#">Temperature</a>	UNKNOWN	0d 0h 1m 44s ago	0d 0h 45m 8s	Timeout while running /usr/NSMasterHW/bin/nsminfo.sh

8 Matching Service Entries Displayed ( filter: Service Status **PENDING OK WARNING UNKNOWN CRITICAL**)

Figure 8-4. Host Service details

By clicking on the links in the windows even more detailed information is provided for the services.

## 8.5.4 Ping Map View

The **ping** map view is similar to the **all status** map view except that it only shows the state of the pings sent to the different components in the cabinets. The state of the services associated with the nodes is not taken into account.

By default the Monitoring Problems window will appear at the bottom of the screen.



## 8.6 Status Button

By clicking on the Status button a screen appears which lists all the hosts and the services running on each one of them as shown in Figure 8-5. More detailed information may be seen for each Host Group by selecting either the individual Host Group or by selecting the links in the Host Status Totals or Service Status Totals columns.

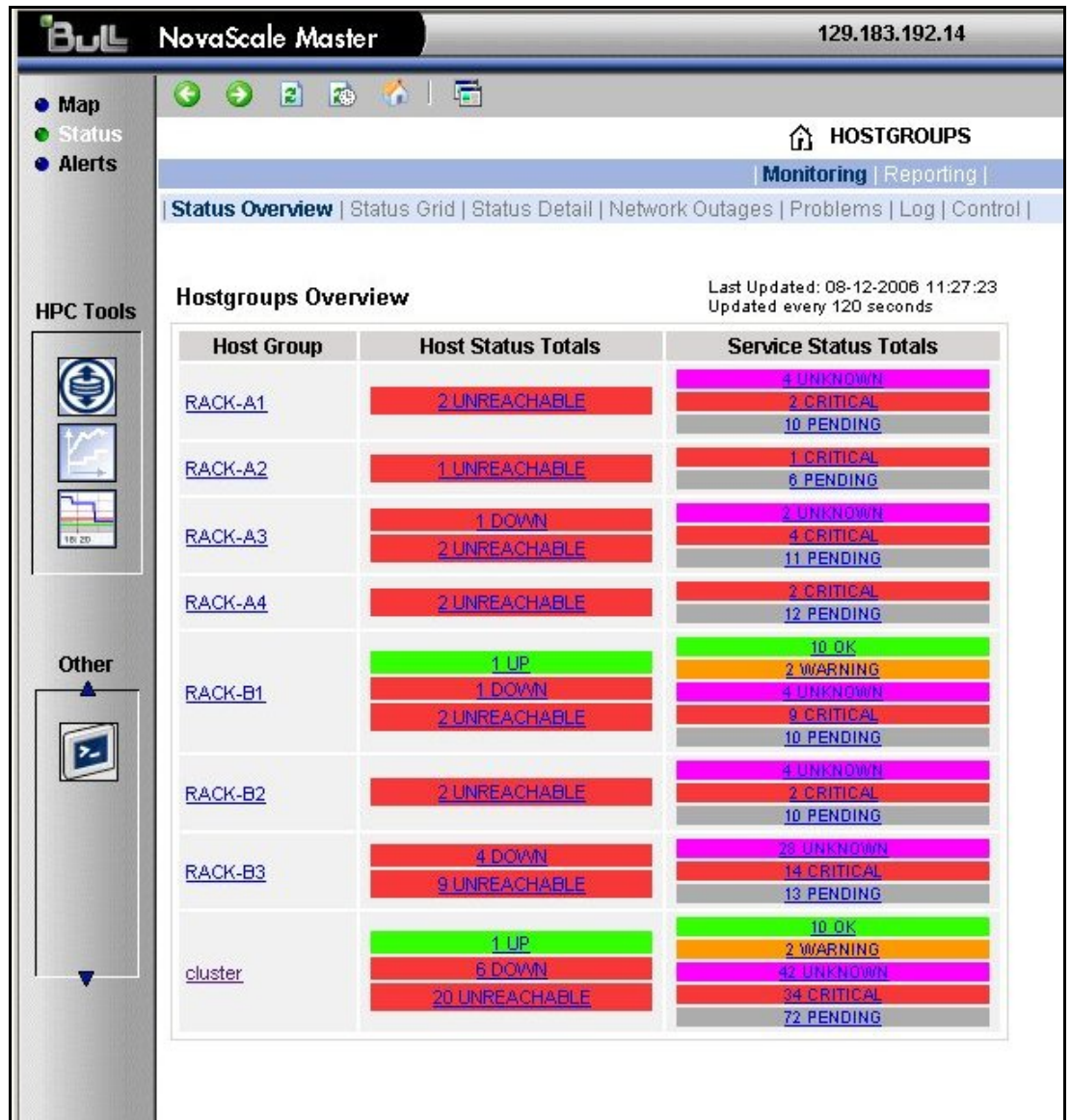


Figure 8-5. Status overview screen

## 8.7 Alerts Button

The **Nova Scale Master Alert Viewer** application displays monitoring alerts (also called events) concerning a set of **hostgroups**, **hosts** and **services**. The application provides filter functions in order to display alerts on all monitored resources or on only a subset of these resources.

Alerts are visible following the selection of the Alert Button followed by the Reporting link tab, and then by the Alert Viewer – see Figure 8-6.

Whenever a service or host status change takes place, the monitoring server generates an alert, even when status passes from **CRITICAL** to **RECOVERY** and then to **OK**. Alerts are stored in the current monitoring log and are then archived.

The NovaScale Master - HPC Edition Alert Viewer application scans the current monitoring log and archives according to filter **report period** settings.

### Alert Types

The alerts can be filtered according to the following alert types:

- Hosts and Services
- Hosts
- Services.



**Note:** By default, Hosts and Services is selected.

### Alert Level

Alerts can be filtered according to the following alert levels:

- **All** – Displays all alerts.
- **Major and Minor problems** - Displays host alerts with DOWN or UNREACHABLE status levels or displays service alerts with WARNING, UNKNOWN or CRITICAL status levels.
- **Major problems** -Displays host alerts with DOWN or UNREACHABLE status levels displays service alerts with UNKNOWN or CRITICAL status levels.
- **Current problems** -Display alerts with a current non-OK status level. When this alert level is selected, the Time Period is automatically set to 'This Year' and cannot be modified.



**Note:** By default, All is selected.

### Report Period

This setting can be changed using the drop down menu provided.

## 8.7.1 Active Checks

Active monitoring consists in running at regular intervals a plug-in program which will carry out checks and send the results back to Nagios. The plug-in will send various codes which correspond to the Alert alarm level.

These are 0 for OK/UP (Green background), 1 for WARNING (Orange background), 2 for CRITICAL/DOWN (Red background), 3 for UNKNOWN (Violet background). The plug-in will also display an explanatory text for the alarm level.

The screenshot shows the 'ALERTS' window in the NovaScale Master console. It includes a filter section for 'Alert Viewer' and a table of 'Matching Alerts'. The table lists various alerts with their respective states and counts.

Time	Host	Service	State	Count	Information
07-03-2006 13:10:43	tiger0	Storage arrays	CRITICAL	55	CRITICAL : No storage available
07-03-2006 13:10:23	tiger0	Global filesystem usage	UNKNOWN	54	can't fetch tiger6 data from ganglia database, check /var/lib/ganglia/frds directory tree
07-03-2006 13:10:03	tiger0	Global filesystem bandwidth	UNKNOWN	54	node_ha_id not defined in table lustre_io_node
07-03-2006 08:00:07	tiger12	Ethernet interfaces	CRITICAL	1	down : [ 172.16.1.13 10.2.1.13 ] - up : [ ]
07-03-2006 08:00:07	tiger12	N/A	DOWN	3	down : [ 172.16.1.13 10.2.1.13 ] - up : [ ]
06-03-2006 16:40:14	tiger12	Ethernet interfaces	OK	1	down : [ 172.16.1.13 ] - up : [ 10.2.1.13 ]
06-03-2006 16:40:14	tiger12	N/A	UP	1	down : [ 172.16.1.13 ] - up : [ 10.2.1.13 ]
06-03-2006 14:01:04	ddn1	System status	CRITICAL	1	at least one of the services is CRITICAL on this disk_array
06-03-2006 14:01:04	ddn1	Temperature	OK	1	All 58 temperature sensors are ok
06-03-2006 14:01:04	ddn1	Power fan	OK	1	All 76 power_supply(jes), power_fan(s) or fans are ok
06-03-2006 14:01:04	ddn1	FC port	CRITICAL	1	2 FC port(s) is/are faulty
06-03-2006 14:01:03	ddn1	Disk	WARNING	1	1 disk_slot(s) is/are missing or faulty / 31 spare disk left

Figure 8-6. Alert Window showing the different alert states

## 8.7.2 Passive Checks

With this form of monitoring a separate third-party program or plug-in will keep Nagios informed via its external command file (`/var/spool/nagios/nagios.cmd`). It submits the result in the form of a character string which includes a timestamp, the name of the host and service concerned as well as the return code and the explanatory text.

Passive checks appear with a grey background in the list of alerts.

### 8.7.3 Notifications

Notifications are sent out if a change or a problem occurs. The Notification may be one of 3 types- e-mail, SNMP Trap or using a user Script. Host and service notifications occur in the following instances:

#### When a hard state change occurs

When a host or service remains in a hard non-OK state and the time specified by the **<notification\_interval>** option in the host or service definition has passed since the last notification was sent out (for that specified host or service). In order to prevent recurring notifications, set the **<notification\_interval>** value to 0 - this stops notifications from being sent out more than once for any given problem.

The Monitoring Control window see Figure 8-8 provides the facility to enable or disable notifications.

The Notification level is set in the Maps → Hostgroups → Reporting → Notifications window. The different notification levels are as indicated below.

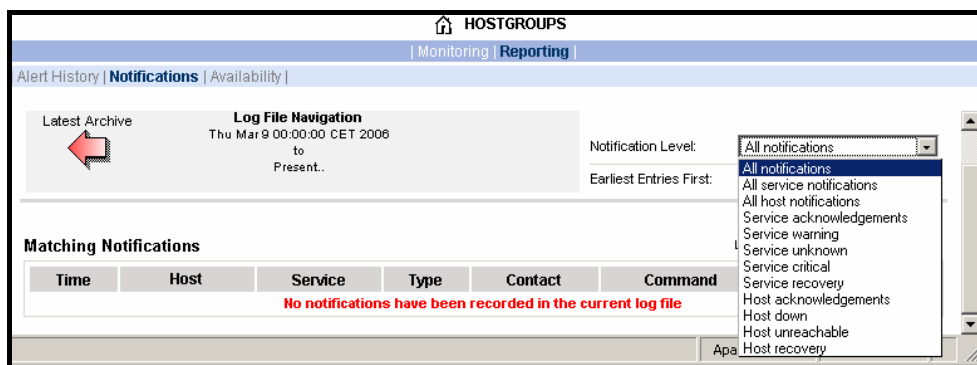


Figure 8-7. Hostgroups Reporting Notifications Window showing the Notification Levels

### 8.7.4 Acknowledgments

As the **Administrator**, you may acknowledge alerts and decide whether they should be displayed or not.

## 8.7.5 Comments

Users of a particular host can post comments using the Monitoring Control window as shown below.

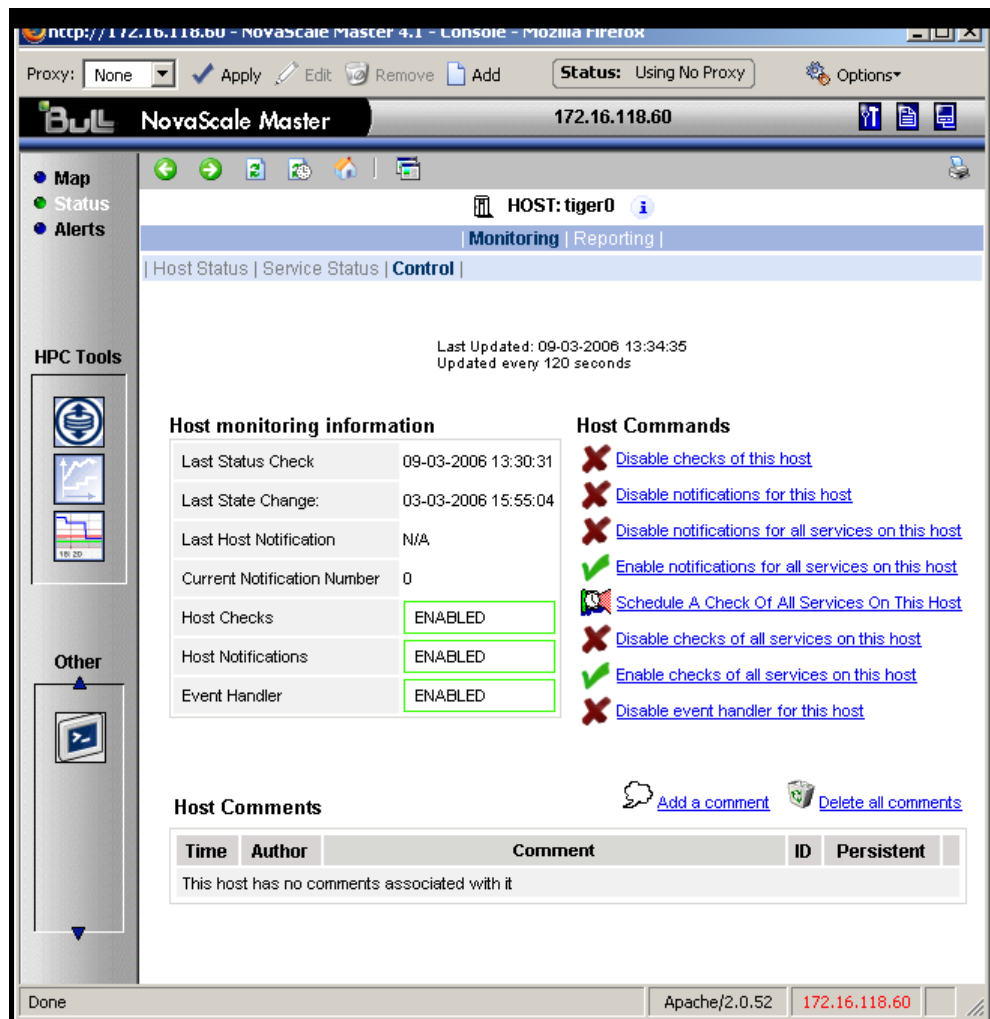


Figure 8-8. Status Monitoring Control window showing the links to add and delete comments

## 8.7.6 Logs

The current Nagios log file is `/var/log/nagios/nagios.log`. The log archives for the preceding weeks is saved `/var/log/nagios/archives`. The Service Log Alert window may be displayed by selecting it in the Service Status window as shown below.

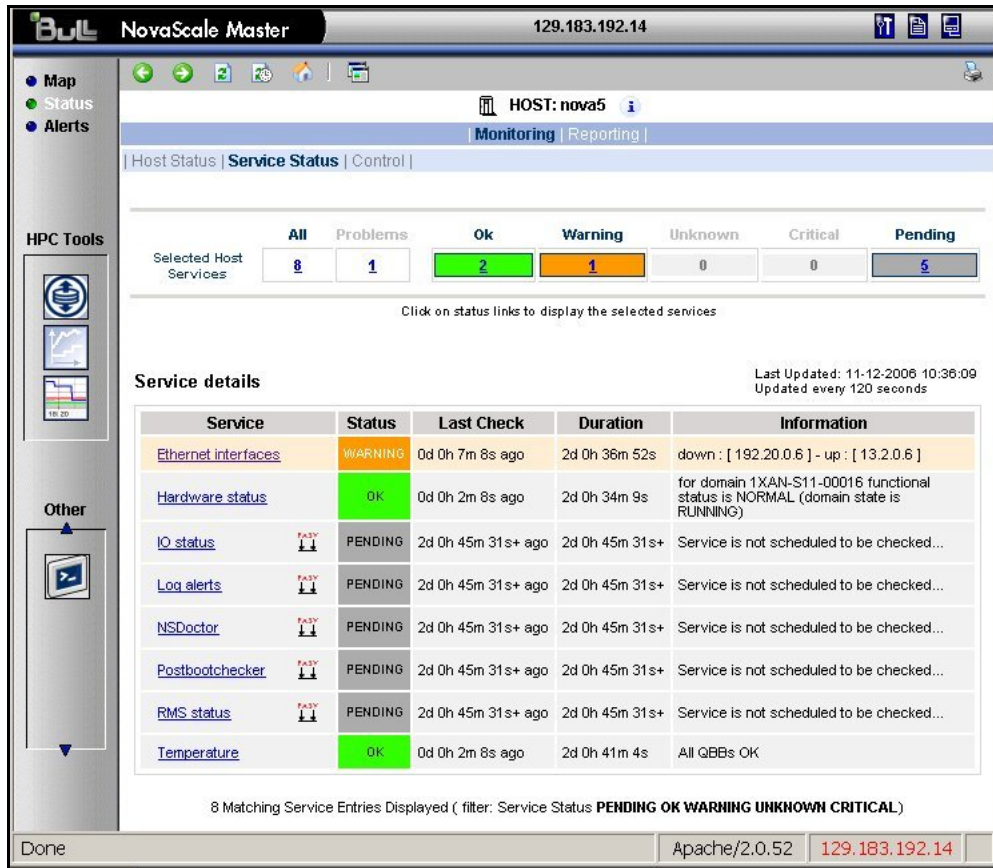


Figure 8-9. Monitoring Service Status window for a host.

More details for the Log alerts are available by selecting the Log alerts link in the middle of the screen.

## 8.7.7 Alert Definition

The different parameters which may be used for an alert are as follows:

**\$HOSTNAME\$**: The name of the host from which the alert is returned.

**\$HOSTALIAS\$**: The content of the comma separated field '!'

For a node this is: `node:<type>:<model>`  
 with `<type>` = for example A-, -C-, AC-M-  
 with `<model>` = for example NS6160.

For a PAP this is: `pap:<type>`  
 with `<type>` = master, standard.

For an Ethernet switch: **eth\_switch:<model>**  
with **<model>** = for example. CISCO 3750G24TS.

For an interconnect switch : **ic\_switch:<model>**  
with **<model>** = for example the type of material (**node, pap, eth\_switch, ic\_switch**).

## 8.7.8 Running a Script

NovaScale Master - HPC Edition can be configured to run a script when a state changes or an alert occurs.

Below is an example of script, which is run when **RMS** sends a **configure-out** event on a node.

```
#!/usr/bin/perl -w

# Arguments : $SERVICESTATE$ $STATETYPE$ $HOSTNAME$ $HOSTSTATE$ $OUTPUT$

$service_state = shift;
$state_type = shift;
$host_name = shift;
$host_state = shift;
$output = join(" ", @ARGV);

# Sanity checks
if ($state_type !~ "HARD") { exit 0; }
if ($service_state !~ "WARNING" && $service_state !~ "CRITICAL") {
    exit 0;
}

# Launch NSDoctor if needed
if ($host_state =~ "UP" &&
    $output =~ /automatically configured out|no response/) {
    system("/usr/sbin/nsdoctor.pl $host_name");
}

exit 0;
```

User scripts which define events or physical changes to trigger Nagios alerts may also be used.

More information on scripts or third party plugins is available in the documentation from <http://www.nagios.org/docs/>

In order that e-mail alerts are sent whenever there is a problem, a SMTP server such as sendmail or postfix has to be in running on the Management node.

By default, the e-mail alerts are sent to [nagios@localhost](mailto:nagios@localhost) on the Management Node.

Normally, by default, only the cluster administrators will receive the alerts for each change for all hosts and services. To send the alert e-mails to other addresses, it is necessary to create new contacts and to add them to the contact groups. The files to modify are `/etc/nagios/contacts.cfg` and `/etc/nagios/contactgroups.cfg`.

## 8.7.9 Generating SNMP Alerts

When **NovaScale Master - HPC Edition** receives an alert (service in a WARNING or CRITICAL state, host in DOWN or UNREACHABLE state), the event handler associated with the service or host sends an SNMP trap, using the `snmptrap` command.

The Management Information Base (MIB) is available in the file `/usr/share/snmp/mibs/NSMASTERTRAPMIB.txt`. This describes the different types of traps and the information that they contain.

In order that an SNMP trap is sent the following actions should be performed:

1. Add the IP address of the host(s) that will receive the traps in the `/etc/nagios/snmptargets.cfg` file (one address per line).
2. Add the contact that will receive the traps to a contact group. To do this, edit the `/etc/nagios/contactgroups.cfg` file and change the line:

```
members nagios
in:
members nagios,snmpt1
```

3. Restart nagios:

```
service nagios reload
```

## 8.7.10 Resetting an Alert Back to OK

To reset an alert back to zero click on the service or the host concerned, then on the menu **Submit passive check result for this service**. Set the **Check Result** to OK, if it is not already the case, fill in the field **Check Output** with a short explanation then click on the button **Commit**. The return to the OK state will be visible after the next command reading by the Nagios demon.

## 8.7.11 nsmhpc.conf Configuration file

The `/etc/nsmhpc/nsmhpc.conf` file contains several configuration parameters. Most of them have default values, but for some services the administrator may have to specify parameter values. A message will inform the administrator if a value is missing.



## 8.8 Storage Overview

By selecting the Storage overview button in the vertical toolbar on the left hand side a window containing the information similar to that shown below is displayed.

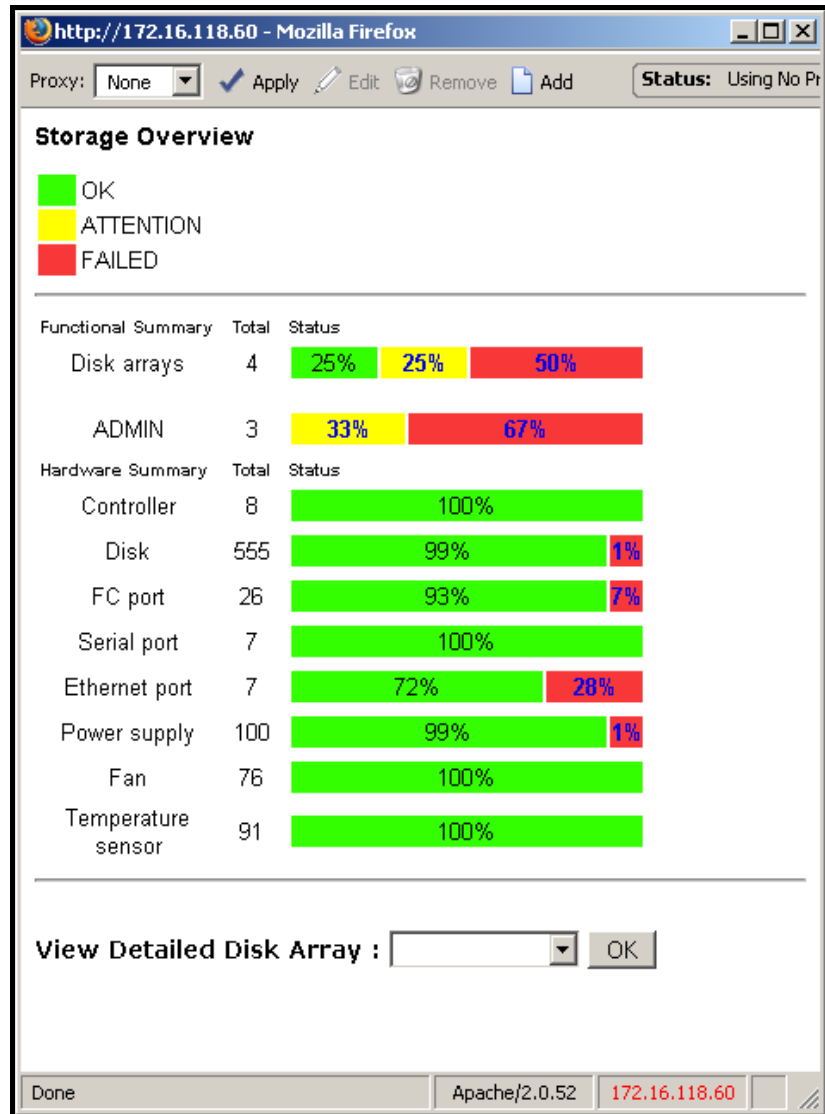


Figure 8-10. Storage overview window

More detailed information is provided by clicking on the ATTENTION and FAILED sections of the component summary status bars.

See *Chapter 9 – Storage Device Management* for information on **NovaScale Master - HPC Edition** and storage views.

## 8.9 Shell

The Shell button can be used to open a shell and enter commands on the administration node.

## 8.10 Monitoring the Performance - Ganglia Statistics

NovaScale Master - HPC Edition provides the means to visualize the performance for the cluster by selecting the icons in the vertical left hand tool bar – see Figure 8-1. This can be done either for a Global Performance View, which displays data either for a complete cluster or on a node by node basis, or in a Group Performance View. These views enable the statistical examination of a predefined group of nodes in the database.

The parameters which enable the calculation of the performance of the cluster are collected on all the nodes by Ganglia and are displayed graphically. One can also choose the observation period and display the measurement details for a particular node using the Ganglia interface.

## 8.11 Group Performance View

This view displays the group performance for 6 different metric types for the complete cluster as shown below. Using this view it is possible to see view the nodes in groups and then to zoom on a particular node.



Figure 8-11. Group Performance view

## 8.12 Global Performance View

The global performance view gives access to the native interface for Ganglia and provides an overall view of the cluster. It is also possible to view the performance data for individual nodes.

Five categories of data collected. These are:

- Load for CPUS and running processes
- Memory details
- Processor activity
- Network traffic in both bytes and packets
- Storage.

Each diagram shows changes for the performance metrics over a user defined period of time.



Figure 8-12. Global overview for a host (top screen)

More detailed views are shown by scrolling the window down – see Figure 8-13.



Figure 8-13. Detailed monitoring view for a host (bottom half of screen displayed in Figure 8-12)

## 8.12.1 Modifying the Performance Graphics Views

The format of the graphs displayed in the performance views can be modified by editing the file `/usr/share/nagios/conf.inc`. The section which follows the line `Metrics` enumeration defines the different graphs; each graph is created by a call to the producer of the `Graph` class. To create a new graph, it is necessary to add the line:

```
$myGraph = new Graph("<graphname>")
```

`<graphname>` is the name given to graph.

To specify a metric to the graph, the following command must be edited as many times as there are metrics to be added or changed:

```
$myGraph->addMetric(new Metric("<metricname>", "<legende>", "<fonction>",
"<couleur>", "<trait>"))
```

`<metricname>` the name given by Ganglia for the metric.

`<legende>` text displayed on the graph to describe the metric.

`<fonction>` aggregating function used to calculate the metric value for a group of nodes, currently the functions `sum` and `avg` are supported.

`<couleur>` HTML code color.

**<trait>** style for feature displayed (**LINE1**, **LINE2**, **AREA**, **STACK**), See the man page for **rrdgraph** for more details.

Use the command below to add the graph to those which are displayed:

```
graphs:$graphSet->addGraph($myGraph)
```

## 8.12.2 Refresh Period for the Performance View Web Pages

By default the refresh period is 90 seconds. This can be modified by changing the value for the parameter **refresh\_rate** in the file **/etc/nagios/cgi.cfg**.

## 8.13 Configuring and Modifying Nagios Services

### 8.13.1 Configuring Using the Database

The command to be used to regenerate the Nagios services database configuration files is:

```
/usr/sbin/dbmConfig configure --service Nagios --restart
```

This command will also restart Nagios after the files have been regenerated.

Use the following command to test the configuration:

```
service nagios configtest
```



#### Important

The services are activated dynamically according to the Cluster type and the functionalities which are detected. For example, the services activated for **Quadrics** clusters will be different from those which are activated for **InfiniBand** clusters

### 8.13.2 Modifying Nagios Services

The list and configuration of Nagios services is generated from the database and from the file `/etc/nagios/services-tpl.cfg`. This file is a template used to generate the complete files.

All template modifications necessitate the Nagios configuration file to be regenerated using the following command:

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



#### Note:

To check that all services have been taken into account, you can use the **dbmServices** command (this command is described in the *Cluster Database Management* chapter in the present guide). If it is not the case, enter the following commands:

```
/usr/lib/clustmgt/clusterdb/bin/nagiosConfig.pl --init  
dbmConfig configure --service nagios
```

Refer to [http://nagios.sourceforge.net/docs/2\\_0/checkscheduling.html](http://nagios.sourceforge.net/docs/2_0/checkscheduling.html) for more information on configuring the services.

### 8.13.2.1 Clients without Customer Relationship Management software

If a CRM product is not installed then the Nagios configuration files will have to be changed to prevent the system from being overloaded with error messages. This is done as follows:

1. Edit the `/etc/nagios/contactgroups` file and change the line which reads `members nagios,crmwarn,crmcrit` so that it reads `members nagios`
2. In the `/etc/nagios/nagios.cfg` file change the status of the line `process_performance_data=1` so that it is commented.

### 8.13.3 Changing the Verification Frequency

Usually the application will require that the frequencies of the Nagios service checks are changed. By default the checks are carried out once every ten minutes, except on certain services. To change this frequency, the `normal_check_interval` parameter has to be added to the body of the definition of the service and then modified accordingly.

## 8.14 General Nagios Services

Nagios includes a wide range of plug-ins, each of which provides a specific monitoring service which is displayed inside the graphical interface. In addition Bull has developed additional monitoring plug-ins which are included within NovaScale Master – HPC edition. The plug-ins and corresponding monitoring services are listed below. The services listed in this section apply to all node types. The Ethernet Interfaces service also applies to all forms of material/devices.

### 8.14.1 Ethernet Interfaces

The Ethernet interfaces service indicates the state of the Ethernet interfaces for a node. The plug-in associated with this service is **check\_fping** which runs the **fping** command for all the Ethernet interfaces of the node. If all the interfaces respond to the ping, the service posts OK. If **N** indicates the total number of Ethernet interfaces and at least **1** or at most **N-1** interfaces do not answer, then the service will display **WARNING**.

### 8.14.2 Resource Manager Status

The service reports the state of the node as seen by the Resource Manager (**SLURM**) which is in place. The service will be updated every time the state of the node changes.

### 8.14.3 Hardware Status

The material status (temperature and fan status) of each node is posted to the passive Hardware status service, resulting from information from the **check\_node\_hw.pl** plug-in which interfaces with the **BMC** associated with the node.

### 8.14.4 Alert Log

The passive service **Log alerts** displays the last alarm raised by system log for the machine – see Figure 8-9. A mapping is made between the **syslog** severity levels and the Nagios alarm levels: **OK** gathers info, debug and notice alarms; **WARNING** gathers warn and err alarms; **CRITICAL** gathers **emerg**, **crit**, **alert**, **panic** alerts.

### 8.14.5 I/O Status

The I/O status reports the global status of HBA, disks and LUNs on a cluster node. Refer to Chapter 9, section *Monitoring Node I/O Status* for more information.

### 8.14.6 Postbootchecker

The **postbootchecker** tool carries out various analyses after a node is rebooted. It communicates the results of its analyses to the corresponding passive service.



## 8.15 Management Node Nagios Services

These services are available on the management node only.

### 8.15.1 MiniSQL Daemon

This active service uses the `check_proc` plug-in to verify that the `msql3d` process is functioning correctly. It remains at the **OK** alert level whilst the daemon is running but switches to **CRITICAL** if the daemon is stopped.

### 8.15.2 Resource Manager Daemon

This active service uses the `check_proc` plug-in to verify that the **RMSD** process (**Quadrics** clusters), or the **SLURMCLTD** (**InfiniBand** clusters) process, is functioning correctly. It remains at the **OK** alert level whilst the daemon is running but switches to **CRITICAL** if the daemon is stopped.

### 8.15.3 ClusterDB

This active service uses the plug-in `check_clusterdb.pl` to check that connection to the database is being made correctly. It remains at the **OK** alert level whilst the connection is possible but switches to **CRITICAL** if the connection becomes impossible.

### 8.15.4 Cron Daemon

This active service uses the `check_proc` plug-in to verify that the `cron` daemon is running on the system. It remains at the **OK** alert level whilst the daemon is running but switches to **CRITICAL** if the daemon is stopped.

### 8.15.5 Compute Power Available

A Bull plug-in checks the computer power available and the alert level associated with it and then displays the results in the Availability Indicators view pane on the top right hand side of the opening window for the Map button as shown in Figure 8-2.

### 8.15.6 Global Filesystems bandwidth available

A Bull plug-in checks the bandwidth for the global filesystem and the alert level associated with it and then displays the results in the Availability Indicators view pane on the top right hand side of the opening window for the Map button as shown in Figure 8-2.

### 8.15.7 Storage Arrays available

A Bull plug-in checks how much space is available for the storage arrays and the alert level associated with it and then displays the results in the Availability Indicators view pane on the top right hand side of the opening window for the Map button as shown in Figure 8-2.

## 8.15.8 Global Filesystem Usage

A Bull plug-in checks the global filesystem usage and the alert level associated with it and then displays the results in the Availability Indicators view pane on the top right hand side of the opening window for the Map button as shown in Figure 8-2.

## 8.15.9 I/O pairs Migration Alert

A Bull plug-in checks the I/O pairs status and the alert level associated with them and then displays the results in the Availability Indicators view pane on the top right hand side of the opening window for the Map button as shown in Figure 8-2.

## 8.15.10 Backbone Ports Available

This service calculates the percentage of ports which are usable on the backbone switches. All the ports which are not usable have to be in the state *administratively down*.

The results are displayed in the Availability Indicators view pane on the top right hand side of the opening window for the Map button as shown in Figure 8-2.

## 8.15.11 HA System Status

This service is based on the output of the **clustat** command. It displays the state of the management nodes which are running with High Availability. As soon as one or more management nodes rocks to the 'offline' state the service displays a list of all the nodes in the 'offline' state and returns an alert level of **CRITICAL**. If all the management nodes are 'online' then the service returns **OK**.

## 8.15.12 Kerberos KDC Daemon

This active service uses the plug-in **check\_proc** to check if the daemon **krb5kdc** is running on the system. It remains at the **OK** alert level whilst the daemon is running but switches to **CRITICAL** if the daemon is stopped.

## 8.15.13 Kerberos Admin Daemon

This active service uses the plug-in **check\_proc** to check if the daemon **kadmin** is running on the system. It remains at the **OK** alert level whilst the daemon is running but switches to **CRITICAL** if the daemon is stopped.

## 8.15.14 LDAP Daemon (Lustre clusters only)

This active service checks if the **check\_ldap** plug-in which the Lightweight Directory Access Protocol (**LDAP**) uses with **Lustre** is working correctly. This plug-in makes a connection to **LDAP** using **fs=lustre** as root for the naming hierarchy.

### 8.15.15 Lustre filesystems access

This is a passive service which is run every 10 minutes by a cron. The cron connects to a client node taken from a specified group at random, for example a Compute Node, and attempts to create and write (stripe) a file on all the **Lustre** file system directories that are listed in the Cluster DB and that are mounted on the node. The file is deleted at the end of the test. If the operation is successful an **OK** code is sent to Nagios with the message '*All Lustre filesystems writable*'. If not, a **CRITICAL** code is returned with the message '*Lustre problem detected*'.

The service uses the **lustreAccess.group** parameter defined in the **/etc/nsmhpc/nsmhpc.conf** file to specify the group containing the nodes that can be used for the test (default: COMP).

### 8.15.16 NFS filesystems access

This is a passive service which is run every 10 minutes by a cron. The cron connects to a client node taken from a specified group at random, for example a Compute Node, and looks for all the NFS filesystems mounted on this node. Then it tries to create and write a file in a specified sub-directory, on all NFS filesystems. The file is deleted at the end of the test. If the operation is successful an **OK** code is sent to Nagios. If not, a **CRITICAL** code is returned with detailed information.

The service uses three parameters, defined in the **/etc/nsmhpc/nsmhpc.conf** file:

- **nfsAccess.group**, which specifies the group containing the nodes that can be used for the test (default: COMP).
- **nfsAccess.directory**, which specifies an existing sub-directory in the filesystem where the test file will be created.
- **nfsAccess.user**, which specifies a user authorized to write in the sub-directory defined in the **nfsAccess.directory** parameter.

### 8.15.17 InfiniBand Links available

This service calculates the percentage of links that are usable for the **InfiniBand** switches.

The results are displayed in the Availability indicators view pane on the top right hand side of the opening window for the Map button as shown in Figure 8-2.

The administrator must specify two parameters in the **/etc/nsmhpc/nsmhpc.conf** file:

- **indicator.ib.numUpLinks**, which specifies the number of installed up links (top switches <-> bottom switches)
- **indicator.ib.numDownLinks**, which specifies the number of installed down links (bottom-switches <-> nodes)

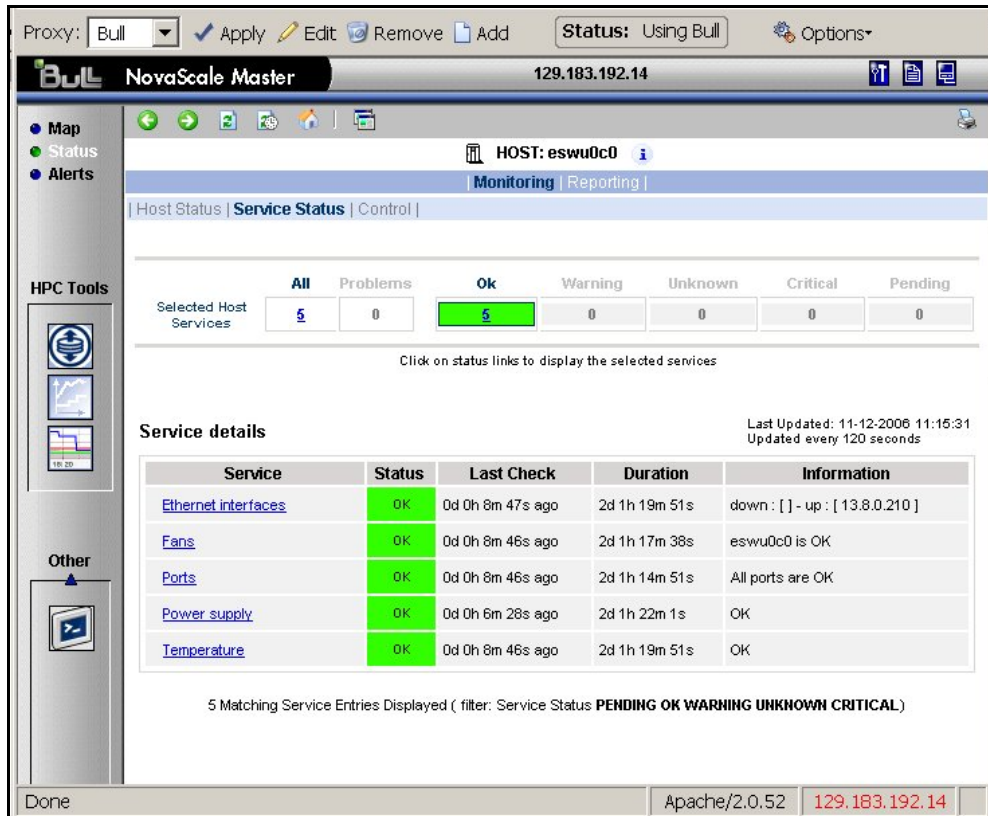
According to these values and the values returned by the IBS tool, the service will be able to define the availability of the **InfiniBand** interconnects.



See Chapter 2 in the BAS4 for Xeon *Maintenance Guide* for more information on the IBS tool.

## 8.16 Ethernet Switch Services

The Ethernet switches which are not used should be set to *disabled* so that Ethernet switch monitoring works correctly. This is usually done when the switch is first configured. The services for each Ethernet switch are displayed when the switch is selected in either the cluster host group or the host window.



The screenshot shows the Bull NovaScale Master interface for host `eswu0c0`. The interface includes a navigation sidebar with 'Map', 'Status', and 'Alerts' options. The main content area displays 'Service Status' for the host, with a summary table showing 5 services in 'Ok' status and 0 in other categories. Below this is a 'Service details' table listing services like Ethernet interfaces, Fans, Ports, Power supply, and Temperature, all with 'OK' status.

Service	Status	Last Check	Duration	Information
<a href="#">Ethernet interfaces</a>	OK	0d 0h 8m 47s ago	2d 1h 19m 51s	down : [ ] - up : [ 13.8.0.210 ]
<a href="#">Fans</a>	OK	0d 0h 8m 46s ago	2d 1h 17m 38s	eswu0c0 is OK
<a href="#">Ports</a>	OK	0d 0h 8m 46s ago	2d 1h 14m 51s	All ports are OK
<a href="#">Power supply</a>	OK	0d 0h 6m 28s ago	2d 1h 22m 1s	OK
<a href="#">Temperature</a>	OK	0d 0h 8m 46s ago	2d 1h 19m 51s	OK

Figure 8-14. Ethernet Switch services

### 8.16.1 Ethernet Interface

The **Ethernet interfaces** service checks that the Ethernet switch is responding by using a ping to its IP address.

### 8.16.2 Power supply

The **Power supply** service checks the power supply is functioning properly by using the `check_esw_power.pl` plug-in.

### 8.16.3 Temperature

The **Temperature** service monitors the temperatures of the Ethernet switches by using the `check_esw_temperature.pl` plug-in.

## 8.16.4 Fans

The **Fans** service monitors the fans for the Ethernet switches using the `check_esw_fans.pl` plug-in.

## 8.16.5 Ports

The **Ports** service monitors the ports for the switches. If one or more ports are detected as being in a *notconnect* state, this service will display the WARNING state and indicate which ports are unavailable.

## 8.17 More Nagios Information

See the Nagios documentation for more information, in particular regarding the configuration. Look at the following web site for more information  
[http://nagios.sourceforge.net/docs/2\\_0/](http://nagios.sourceforge.net/docs/2_0/)

In addition look at the **NovaScale Master - HPC Edition** documentation suite, this includes an *Installation Guide*, a *User's Guide*, an *Administrator's Guide* and a *Remote Hardware Management CLI Reference Manual*.



---

## Chapter 9. Storage Device Management

Bull cluster management tools provide services to manage a large amount of storage systems and storage resources. This chapter explains how to setup the management environment, and how to use storage management services.

The following topics are described:

- *9.1 Overview of Storage Device Management for Bull HPC Clusters*
- *9.2 Monitoring Node I/O Status*
- *9.3 Monitoring Storage Devices*
- *9.4 Monitoring Brocade Switch Status*
- *9.5 Managing Storage Devices with Bull CLI*
- *9.6 Using Management Tools*
- *9.7 Configuring Storage Devices*
- *9.8 User Rights and Security Levels for the Storage Commands*

## 9.1 Overview of Storage Device Management for Bull HPC Clusters

Bull HPC clusters may contain various kinds of storage devices. Thus, storage device management may quickly become a complex task, due to the variety and the number of management interfaces.

Using Bull storage management services the cluster administrator will be able to:

- Monitor the status of storage devices
- Monitor storage within cluster nodes
- Get information about faulty components
- Get synthetic reports for the storage resources
- Automate the deployment of storage device configurations
- Ensure consistency between storage systems and I/O nodes
- Configure individual storage devices using a command line interface from the cluster management station
- Obtain access to the management tools for each storage device, regardless of its user interface.

Bull HPC clusters are deployed with both a specific hardware infrastructure and with software packages to simplify and unify these management tasks.

The hardware infrastructure enables the management of all the storage devices from the cluster Management Nodes:

- Built-in LAN management ports of the storage devices are connected to the cluster management network.
- Built-in serial ports of the storage devices are connected to the cluster management network, using terminal servers.
- Management stations or proxy servers (for example Windows stations) hosting device management tools are connected to the cluster management network, or are reachable from the Management Nodes.

The software packages installed on the cluster Management Node and on other cluster nodes provide various device management services:

- **Device monitoring.** A device inventory is performed and detailed descriptions of all the storage devices are stored in the cluster data base. The storage devices are monitored by the cluster Management Node, using standardized protocols such as **SNMP**, **syslog**, or proprietary interfaces. The Management Node waits for event notification from the devices. To prevent silent failures, forced updates are scheduled by the Management Node. All the events are automatically analyzed and the cluster DB is updated to reflect status changes. Storage device status can be monitored using **NovaScaleMaster – HPC Edition** and by querying the cluster DB with the **storstat** command. These services enable the browsing of a global view covering all the storage devices to a more detailed view focusing on a single storage device.



- **Advanced device management.** Administrators trained to manage the storage devices and familiar with the terminology and operations applicable to each kind of storage device can use the command line interfaces available on the cluster Management Node. These commands are specific to a family of storage system (for example **nec\_admin**, etc.). They enable the reading of configuration and status information and also configuration tasks to be performed. The syntax and output are as close as is possible to the information provided by the device management tools provided with the storage system. The most useful information and operations are available through these commands. Nevertheless, they do not offer all the management services for each device. Their advantage is that they provide a command line interface on the cluster Management Node. They can also be used to build custom tasks, by parsing command outputs or creating batches of commands.



**Warning:**

Changing the configuration of a storage device may affect all the cluster nodes using this device.

- **Access to management tools.** The storage administrator trained to manage storage devices can also access to the management tool for each storage device. The serial ports can be used with `conman` (or `telnet`). The Ethernet ports can be connected with `telnet` or a web browser. Management software on proxy UNIX servers can be used with `ssh` (command mode) or `X11` (graphical applications). Similarly, an `ssh` service and a VNC server are provided for Windows, in order to enable access to the management software on proxy Windows servers, either in command mode or in graphical mode.
- **Storage device configuration deployment.** For small clusters, the administrator can use either the device specific commands installed on the cluster Management Node, or the tools for each storage device. For medium to large clusters, there are often lots of storage systems with the same hardware and logical configuration. For these kinds of complex environments, configuration deployment services are provided.

These services are only available in command mode.



**Warning:**

System Administrators must be trained to manage the storage devices, and be familiar with the terminology and operations applicable to each kind of storage device. They must be aware of the impact of updating a storage device configuration

The next sections explain how to setup and use this environment.

## 9.2 Monitoring Node I/O Status

Each node is monitored and many I/O errors reported in **syslog** are tracked. A global I/O status is computed locally on each node and is reported to the management station using dedicated **syslog** messages.

The current I/O status of each node can be verified by displaying the "I/O status" service of the node using NovaScale Master for HPC.

The semantic of the service is as follows:

<b>OK</b>	No problem detected
<b>WARNING</b>	An I/O component in <b>WARNING</b> state is in an unstable state but the resource is still available. It may also indicate that the current number of I/O components is higher than its expected reference number.
<b>CRITICAL</b>	Degraded service. Maintenance operation mandatory Criteria: Hereafter is a list of possible critical errors:  A fatal disk error has been reported by the linux I/O stack in <b>syslog</b> A fatal <b>HBA</b> error has been reported by a device driver in <b>syslog</b> A link down transition has been notified by a device driver A <b>LUN</b> resource cannot be acceded by a multipath pseudo-device. A device referenced by the persistent binding mechanism (alias) is missing.
<b>UNKNOWN</b>	Can't access the status
<b>PENDING</b>	Not yet initialized

The I/O status transmitted by each node to the management station results of the synthesis of multiple controls. This detailed information is available on each node using the **lsiodev** command:

```
lsiodev -l
```

The I/O status monitoring service builds a reference during its initial startup, usually at the first boot of the node.

The reference contains the expected number of various classes of devices (named "I/O counters").

Two reference counters (**nb\_io\_adapters** and **nb\_local\_disks**) are stored on the management station in cluster DB in the table `node_profile`. The other reference counters are stored on the local node.

At boot time the **nb\_io\_adapters** and **nb\_local\_disks** counters are automatically adjusted from the cluster DB node I/O profile.

You can view details of I/O status reference counter values of each node by the link **I/O status details** of the **I/O status** service on the node using NovaScale Master for HPC.

**I/O Status Details of node : nova5**

- The number of I/O resources is different from expected
- === Global I/O Status is WARNING ===

**I/O Counters of node : nova5**

Status	Counter	Value	Definition	OK State Counter	Value
WARNING	nb_io_adapters	2 / 5	I/O adapters and internal chips	nb_io_adapters_configured	2 / 5
WARNING	nb_local_disks	3 / 4	Physical disks	nb_local_disks_ok	3 / 4
OK	nb_io_ports	1 / 1	I/O ports	nb_io_ports_connected	1 / 1
OK	nb_fixed_luns	3 / 3	Fixed LUNs (/dev/sd*) directly mapped to local disks	nb_fixed_luns_ok	3 / 3
WARNING	nb_reconf_luns	10 / 8	Reconfigurable LUNs (/dev/sd*) from external storage or RAID adapter	nb_reconf_luns_ok	10 / 8
OK	nb_pseudos	0 / 0	Multipath pseudo-devices (/dev/dm-*, /dev/emcpower*)	nb_pseudos_ok	0 / 0
OK	nb_iopaths	0 / 0	Multipath I/O paths (under pseudo-devices)	nb_iopaths_ok	0 / 0
OK	nb_aliases	8 / 8	Device aliases (/dev/ldn*) for LUNs or pseudo-devices	nb_aliases_ok	8 / 8

(Counter Value = Current / Expected)

**I/O Resources of node : nova5**

Adapter	Host	Resource	Physical Disk
03:01 LSI LSI53C1030 Driver: mptspi CONFIGURED	host0	sdb (0:0:10:0) OK (Fixed LUN)	Physical Disk sdb OK SEAGATE SPI 286102MB
		sdc (0:0:11:0) OK (Fixed LUN)	Physical Disk sdc OK SEAGATE SPI 286102MB
	host1	sda (0:0:9:0) OK (Fixed LUN)	Physical Disk sda OK SEAGATE SPI 286102MB
2d:01 Emulex LP11000 Driver: lpfc CONFIGURED	host2 (Port) WWN: 10:00:00:00:c9:4b:c0:9a CONNECTED	sdd (2:0:0:0) OK DDN 1000MB (Reconfigurable LUN, FC) ← Alias ldn.ddn0.24	
		sde (2:0:0:1) OK DDN 1048576MB (Reconfigurable LUN, FC) ← Alias ldn.ddn0.25	
		sdl (2:0:0:12) OK DDN 49896MB (Reconfigurable LUN, FC)	
		sdm (2:0:0:13) OK DDN 49896MB (Reconfigurable LUN, FC)	
		sdf (2:0:0:2) OK DDN 1000MB (Reconfigurable LUN, FC) ← Alias ldn.ddn0.26	
		sdg (2:0:0:3) OK DDN 1048576MB (Reconfigurable LUN, FC) ← Alias ldn.ddn0.27	
		sdh (2:0:0:4) OK DDN 1000MB (Reconfigurable LUN, FC) ← Alias ldn.ddn0.28	
		sdi (2:0:0:5) OK DDN 1048576MB (Reconfigurable LUN, FC) ← Alias ldn.ddn0.29	
		sdj (2:0:0:6) OK DDN 1000MB (Reconfigurable LUN, FC) ← Alias ldn.ddn0.30	
		sdk (2:0:0:7) OK DDN 1048576MB (Reconfigurable LUN, FC) ← Alias ldn.ddn0.31	
ldn.ddn0.24 Alias OK linked to sdd			
ldn.ddn0.25 Alias OK linked to sde			
ldn.ddn0.26 Alias OK linked to sdf			
ldn.ddn0.27 Alias OK linked to sdg			
ldn.ddn0.28 Alias OK linked to sdh			
ldn.ddn0.29 Alias OK linked to sdi			
ldn.ddn0.30 Alias OK linked to sdj			
ldn.ddn0.31 Alias OK linked to sdk			

Figure 9-1. I/O Status Details NovaScale Master HPC Edition example screens

The **iorefmgmt** command is used to manage I/O device monitoring reference counters.

To get the list of the reference counter enter:

```
iorefmgmt -g
```

Use the help or the man page to obtain a description of the counters used, alternatively see the definitions in the section below.

If the reference is wrong, it can be updated as follows:

```
iorefmgmt -s -n <counter_name> -v <value>
```

You can adjust reference counters to the current discovery value by the command:

```
iorefmgmt -c adjust
```

The counters **nb\_io\_adapters** and **nb\_local\_disks** cannot be adjusted on a node.

You can manage these counters in the cluster DB node profile table on the administration station by using the command:

```
iorefmgmt -c dbset|dbget|dbdel
```

For more information use the **iorefmgmt** man page or help.

All these operations can be done from the management station, using **ssh** or **pdsh**.

## 9.2.1 I/O Counters Definitions

- **nb\_io\_adapters** is the expected number of I/O adapters on the node (a multi-port adapter is counted as 1, an internal I/O chip is also counted as one adapter).
- **nb\_io\_adapters\_configured** is the number of I/O adapters expected to be configured (driver loaded).
- **nb\_local\_disks** is the expected number of physical disks on a node.  
A physical disk may be:
  - an internal disk which is directly attached,
  - a physical disk in a SCSI JBOD,
  - a physical disk behind a RAID controller.
- **nb\_local\_disks\_ok** is the number of physical disks expected to be healthy.
- **nb\_io\_ports** is the expected number of Fibre Channel ports.
- **nb\_io\_ports\_connected** is the number of Fibre Channel ports expected to be connected.
- **nb\_fixed\_luns** is the expected number of LUNs which are not reconfigurable.  
A LUN which is not reconfigurable is directly mapped to a physical disk.
- **nb\_fixed\_luns\_ok** is the number of LUNs which are not reconfigurable that are expected to be accessible.
- **nb\_reconf\_luns** is the expected number of reconfigurable LUNs.
- **nb\_reconf\_luns\_ok** is the number of reconfigurable LUNs expected to be accessible.  
A “reconfigurable LUN” is typically a LUN in an external storage system (usually a RAID system) or a LUN presented by a RAID HBA, on top of RAIDed local disks.
- **nb\_iopaths** is the expected number of paths involved in multi-path to reach LUNs which are reconfigurable.
- **nb\_iopaths\_ok** is the number of paths involved in multipath expected to be alive.
- **nb\_aliases** is the expected number of aliases on Fibre Channel block devices.  
Aliases are used to obtain a persistent device naming scheme, regardless of the order that the FC devices are detected.
- **nb\_aliases\_ok** is the number of aliases on Fibre Channel block devices expected to be correctly mapped.
- **nb\_pseudos** is the expected number of multipath pseudo-devices on a node.
- **nb\_pseudos\_ok** is the number of multipath pseudo-devices expected to be usable.

## 9.3 Monitoring Storage Devices

This section explains how the administrator can monitor and get information about all the managed storage systems of the cluster, using a unified interface. The two following interfaces are available for the administrator:

- Graphical User Interface (NovaScale Master – HPC Edition):
  - Hosts and service monitoring for storage devices.
  - Storage views, providing detailed information regarding the storage systems.
- Command line interface:
  - **storstat** command, to query the ClusterDB for storage information.
  - Archiving of **syslog** messages.



### Note:

The monitoring relies on information stored in the **ClusterDB**. This information is updated periodically and also when failures or repairs are notified by storage devices. The monitoring information is therefore not updated in real-time when silent state changes occur, such as repairs.

The administrators can force a refresh of the Data Base information using the **storcheck** command:

```
storcheck -c <cluster_name>
```

This command will check all the storage systems of the cluster. It is possible to reduce the scope to a single storage system:

```
storcheck -c <cluster_name> -n <disk_array_name>
```

### 9.3.1 NovaScale Master - HPC Edition: Host and Service Monitoring for Storage Devices

Storage device monitoring is integrated in the global monitoring of the cluster. Each storage system is identified by a host and associated service, regardless of the number of controllers and Ethernet ports.

**NovaScale Master - HPC Edition** continuously updates the host status and service status values, without any administrator intervention. All NovaScale Master - HPC Edition features and services apply to storage devices. Nevertheless, the administrator using NovaScale Master - HPC Edition must be aware of some specificities, which are explained after this.

Host	Service	Status	Last Check	Duration	Attempt	Status Information
ddn1	Controller	OK	03-09-2004 09:22:23	1d 23h 21m 40s	1/1	All 2 controllers are ok
	Disk	OK	03-09-2004 09:22:23	1d 18h 29m 27s	1/1	All 74 disk_slots are ok (6 is/are set as empty)
	FC_port	WARNING	03-09-2004 09:22:23	0d 0h 21m 26s	1/1	8 FC ports(s) is/are warning
	Power_fan	CRITICAL	03-09-2004 09:22:23	0d 0h 10m 15s	1/1	4 power_supply(ies), power_fan(s) or fans is/are faulty or missing
	System status	OK	03-09-2004 09:22:23	1d 23h 21m 39s	1/1	Global disk_array status is ok
	Temperature	OK	03-09-2004 09:22:23	1d 16h 34m 55s	1/1	All 8 temperature sensors are ok

Figure 9-2. Detailed service status for a storage host

The host and service monitoring offers uniform monitoring for all the cluster components, with history and statistical capabilities. It provides for each storage system a general view of the major functional domains.

However, this monitoring does not allow the easy identification of the storage devices among other cluster components nor individual faulty hardware components to be identified. These limitations are compensated by the use of Storage Views (see 9.3.2 *NovaScale Master - HPC Edition: Storage & I/O Information*).

### 9.3.1.1 Host Semantic

The host name is a logical name, which uniquely identifies a storage system. But caution, it is not bound to an IP address; it is not possible to ping using this parameter.

The host status indicates whether the storage system is manageable or not:

UP	The storage system responds through the management interfaces
UNREACHABLE	Some network problems prevent the management interface from being reached.
DOWN	The management interfaces of the storage system do not answer to requests. But note that from a storage point of view, the storage system may process I/O requests from attached hosts.

### 9.3.1.2 Service Semantic

Several generic services are defined for storage systems. They reflect the global status of a class of components in the selected storage system:

- Disk
- Power-Fan
- Temperature
- Controller
- FC ports
- System status.

## Disk Service

This service describes the global status for the **HDDs**. It monitors both disk failures and if any disks have been removed (for example for maintenance purpose).

<b>OK</b>	<b>No problem</b> Criteria: No disk errors All referenced disks are present
<b>WARNING</b>	<b>Maintenance operation must be scheduled</b> Criteria: Some disk failures, and / or removed referenced disks Does not meet the criteria for critical status.
<b>CRITICAL</b>	<b>Degraded service. Maintenance operation mandatory</b> Criteria: The number of faulty / missing disks is higher than the number of spare disks.
<b>UNKNOWN</b>	<b>Can't access the status</b>
<b>PENDING</b>	<b>Not yet initialized</b>



### Note:

The cluster database has been initialized with a detailed status including all populated and empty disk slots. The administrator, who decides to permanently remove some HDDs, must manually update the database reference configuration (using the **storstat -u** command). Otherwise, empty slots due to a permanent removal will lead to a permanent **WARNING** status.

## Power-Fan Service

Describes the global status for the power supply and fan modules. These two kinds of hardware parts are grouped and monitored using a single service.

<b>OK</b>	<b>No problem</b> Criteria: <ul style="list-style-type: none"><li>• All power supplies and fans are ok</li><li>• All reference power supplies and fans are present</li></ul>
<b>WARNING</b>	<b>Maintenance operation must be scheduled</b> Criteria: <ul style="list-style-type: none"><li>• Some power supplies and/or fans are in the warning or critical state</li><li>• Does not meet the criteria for critical status.</li></ul>

	<b>Degraded service. Maintenance operation mandatory</b> Criteria:
<b>CRITICAL</b>	<ul style="list-style-type: none"> <li>The percentage of faulty/missing power supplies or fans objects has reached the threshold defined in <code>/etc/storageadmin/storframework.conf</code> (<code>service_power_fan_critical_threshold</code> parameter).</li> </ul>
<b>UNKNOWN</b>	<b>Can't access the status</b>
<b>PENDING</b>	<b>Not yet initialized</b>

### Temperature Service

Describes the global status for temperature sensors.

	<b>No problem</b> Criteria:
<b>OK</b>	<ul style="list-style-type: none"> <li>All temperature sensors are OK</li> </ul>
	<b>Maintenance operation must be scheduled</b> Criteria:
<b>WARNING</b>	<ul style="list-style-type: none"> <li>Some temperature sensors are not in the OK state</li> <li>Critical criteria not met</li> </ul>
	<b>Degraded service. Maintenance operation mandatory</b> Criteria:
<b>CRITICAL</b>	<ul style="list-style-type: none"> <li>Some temperature sensors are in the critical state.</li> </ul>
<b>UNKNOWN</b>	<b>Can't access the status</b>
<b>PENDING</b>	<b>Not yet initialized</b>

### Controller Service

This service shows the controller status. The controller refers to the storage system elements in charge of host connection and I/O processing.

	<b>No problem</b> Criteria:
<b>OK</b>	<ul style="list-style-type: none"> <li>All controllers are OK</li> </ul>
	<b>Maintenance operation must be scheduled</b> Criteria:
<b>WARNING</b>	<ul style="list-style-type: none"> <li>Some controllers have a warning state and none are faulty (or missing).</li> </ul>
	<b>Degraded service. Maintenance operation mandatory</b> Criteria:
<b>CRITICAL</b>	<ul style="list-style-type: none"> <li>One controller or more is faulty (or missing).</li> </ul>
<b>UNKNOWN</b>	<b>Can't access the status</b>
<b>PENDING</b>	<b>Not yet initialized</b>



## Fibre Channel Port Service

This service shows the host connectivity status:

<b>OK</b>	<b>No problem</b> Criteria: <ul style="list-style-type: none"><li>• All FC ports are OK.</li></ul>
<b>WARNING</b>	<b>Maintenance operation must be scheduled</b> Criteria: <ul style="list-style-type: none"><li>• Not in critical status</li><li>• Some ports have a warning status</li></ul>
<b>CRITICAL</b>	<b>Degraded service. Maintenance operation mandatory</b> Criteria: <ul style="list-style-type: none"><li>• One or more ports are in a critical status.</li></ul>
<b>UNKNOWN</b>	<b>Can't access the status</b>
<b>PENDING</b>	<b>Not yet initialized</b>



### Note:

If the FC link is connected to a switch, and the link is broken 'after' the switch and not between the controller and the switch, the failure is not detected by the disk array and therefore will not be displayed by the FC port service.

## System Status Service

This service is a collector and gathers all problems of the storage system. If one of the services described above is warning or critical, the system status service will be critical. This service also reflects the other problems which may arise but are not classified in one of the previously defined services. For example, all the other services may be OK, while the system status is warning or critical.

<b>OK</b>	<b>No problem</b> Criteria: <ul style="list-style-type: none"><li>• Disk array semantic.</li></ul>
<b>WARNING</b>	<b>Maintenance operation must be scheduled</b> Criteria: <ul style="list-style-type: none"><li>• Some of the other services are warning (but none critical).</li><li>• The storage system has detected a warning which is not reported by one of the other services.</li></ul>

	<b>Degraded service. Maintenance operation mandatory</b>
<b>CRITICAL</b>	Criteria: <ul style="list-style-type: none"> <li>• One of the other services is critical.</li> <li>• The storage system has detected a critical error which is not reported by one of the other services.</li> </ul>
<b>UNKNOWN</b>	<b>Can't access the status</b>
<b>PENDING</b>	<b>No yet initialized</b>

### 9.3.2 NovaScale Master - HPC Edition: Storage & I/O Information

NovaScale Master – HPC Edition contains specific views, which focus on the monitoring of storage devices and I/O systems for the nodes connected to these devices. It enables administrators to pinpoint faulty hardware components, and provides detailed reporting and configuration information for storage systems.

The Storage and I/O information view is selected by clicking on the **Storage overview** icon on left hand side of the **NovaScale Master – HPC Edition** console – see Figure 9-3. A pop-up window appears containing a view pre-selected on the summary view of the storage systems and hardware component status – see Figure 9-4.

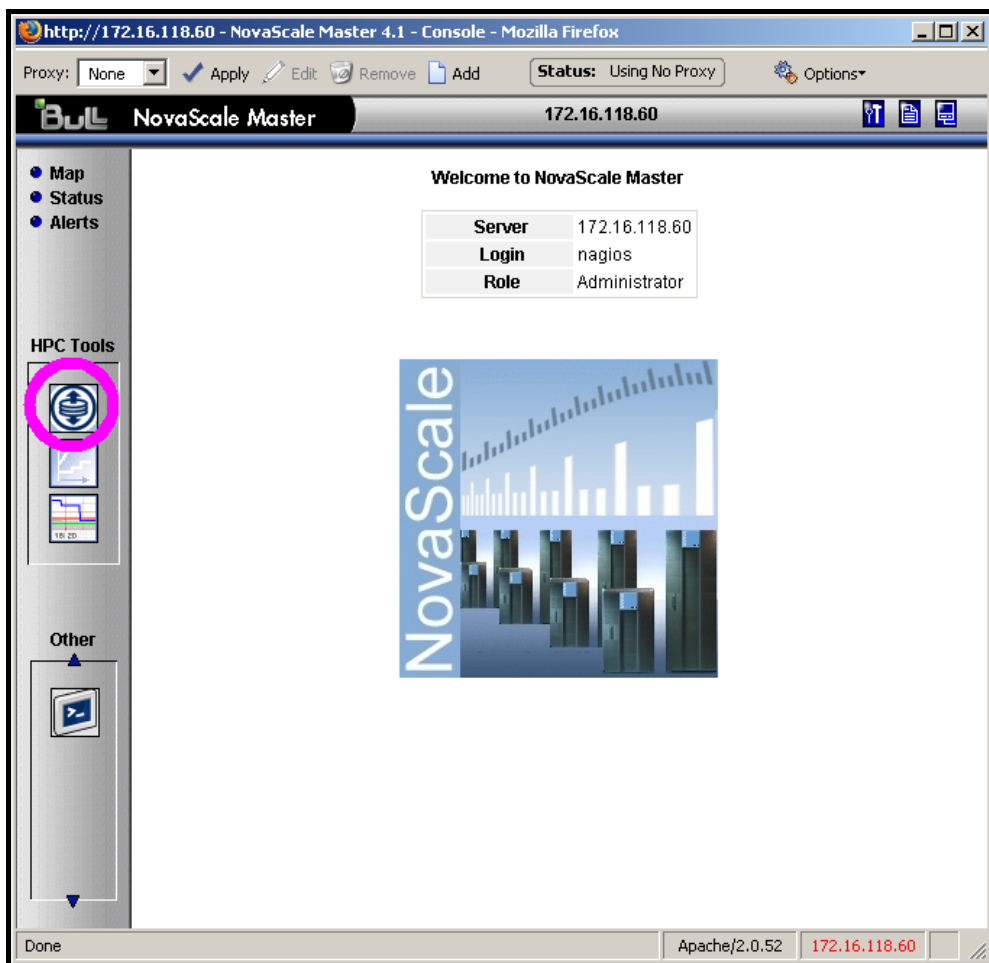


Figure 9-3. Nova Scale Master opening console window with the Storage overview icon circled

### 9.3.2.1

## Storage Views

The storage views provide information about:

- **Disk arrays.** Their status refers to the last known operational status of the storage system under review. It is similar to the 'system status' service in NovaScale Master host and service views. For example a storage system that does not answer to management requests is considered as faulty.
- **Individual hardware components** (Disk, FC port, RAID controller, and so on). There is no equivalent in the host and service monitoring services that provide a single service for all the disks of a storage system.



#### Note:

The disk array status is a superset of the individual hardware components status. It is usually managed by the disk array and is not limited to the hardware components managed by storage views. Therefore the disk array status may be more severe than the worst status of the individual hardware components.

The status used in the storage views are the following ones:

<b>OK</b>	No problem
<b>ATTENTION</b>	Maintenance operation must be scheduled, but the system still delivers the expected service.
<b>FAILED</b>	Degraded service. Maintenance operation mandatory.

### 9.3.2.2 Storage Overview

This view offers a synthesis of the Storage devices monitored in the cluster.

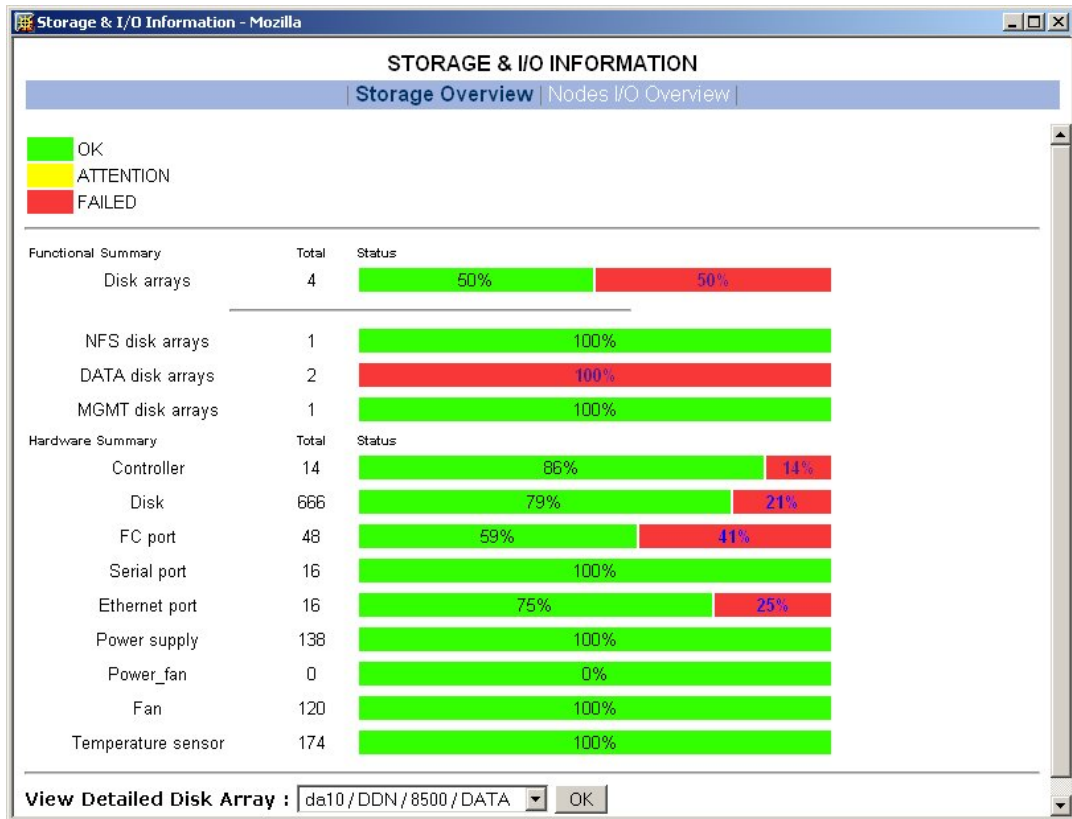


Figure 9-3. Storage overview

#### Functional Summary

This diagram refers to storage systems. It sorts the storage systems according to their operational status and to their respective roles.

#### Hardware Summary

This diagram provides statistics on low level hardware components such as HDDs, Fibre Channel ports, RAID controllers, power supplies, FANs, etc. The diagram is displayed by family of components, sorted by state.

The administrator clicks the ATTENTION and FAILED percentages links in the Storage overview pop-up window to get an inventory list of storage systems or hardware components in the selected state – see Figure 9-4.

### 9.3.2.3 Inventory View of Storage Systems and Components requiring attention

This view – Figure 9-4 -displays the list of faulty components that should be either examined or replaced. The components are grouped by storage system. For each component, the view displays:

- The description of the component

- Its status
- Location information of the component within the device and within the cluster, its rack level and label.

**Components list :**

**type : - Disk arrays -**  
**state : - FAILED -**

---

**ddn1**

Component	Component State	Rack Level/Label - Vendor Location
FC port	NOT_CONNECTED	A / ST00-A24 - singlet 1 HOST 1
FC port	NOT_CONNECTED	A / ST00-A24 - singlet 1 HOST 2
Disk	FAULTY	H / ST00-A24 - disk 29F

3 Defaults

---

**ddn2**

Component	Component State	Rack Level/Label - Vendor Location
Power Supply	FAULTY	D / ST00-A28 - enclosure 3 ps right
Ethernet Port	NOT_CONNECTED	A / ST00-A28 - singlet 1 telnet
Ethernet Port	NOT_CONNECTED	B / ST00-A28 - singlet 2 telnet

3 Defaults

[back](#)

Figure 9-4. Inventory view of faulty storage systems and components



**Note:**

The hardware components whose status is OK are not listed

This view is useful for planning maintenance operations for the components that are to be examined or replaced.

### 9.3.2.4 Detailed View of a Storage System

The Storage detailed view - Figure 9-5- can be displayed by selecting a storage system in the Storage Summary Overview (see Figure 9-3).

This view provides detailed information for the selected storage system:

- Technical information (disk array status, firmware version, addressing information for management purposes, etc.).
- Front and rear diagram view, where the status of all the hardware components is represented by a color code.

- I/O cell and I/O path information:
  - An I/O cell is a set of nodes and storage systems functionally tied together.
  - An I/O path is a logical path between a node and the host port of a storage system. When a point-to-point connection is used, the I/O path is physically represented by a cable. In SAN environment, the I/O path represents both the I/O initiator (the node) and I/O target (the host port of the storage system).
- “Error List” hyperlink (list of faulty components).
- “Lun / Tier / Zoning List” hyperlink (information about the logical configuration of the storage system).

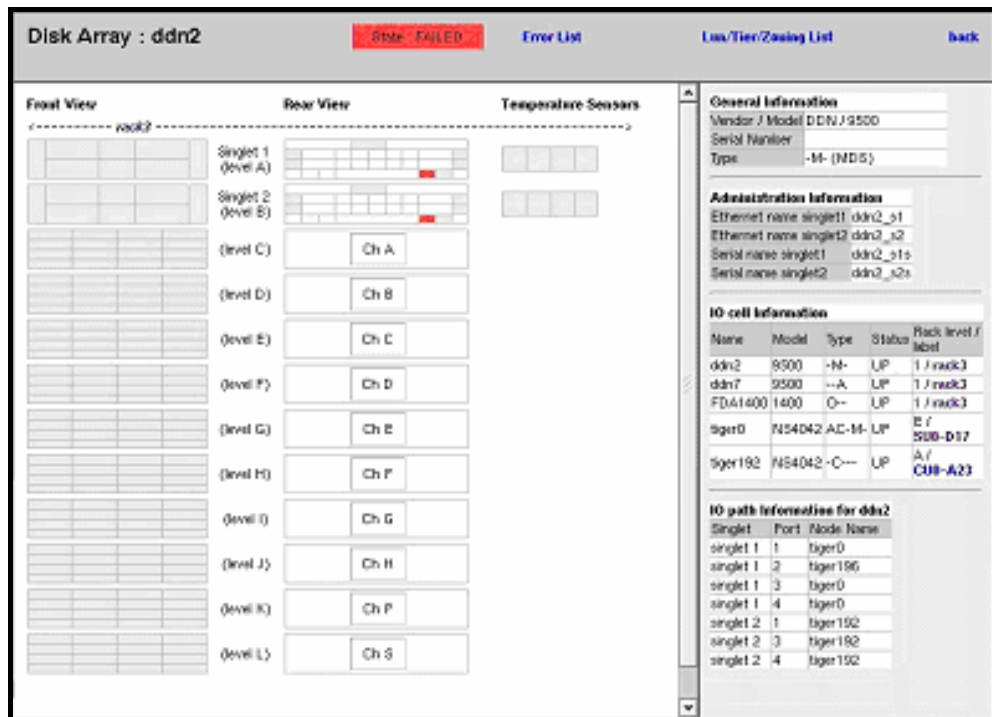


Figure 9-5. Storage detailed view

In the Storage Detailed view the item’s description is shown through the use of mouse Tool tips.

### 9.3.2.5 Nodes I/O Overview

This view – Figure 9-6 – offers a synthesis of the I/O information about the nodes of the cluster.

It shows I/O status statistics and allows the list of nodes to be filtered on a selected I/O status value.

Clicking on the I/O status value of a node allows detailed information about the I/O resources of the node and the associated I/O counters to be displayed.

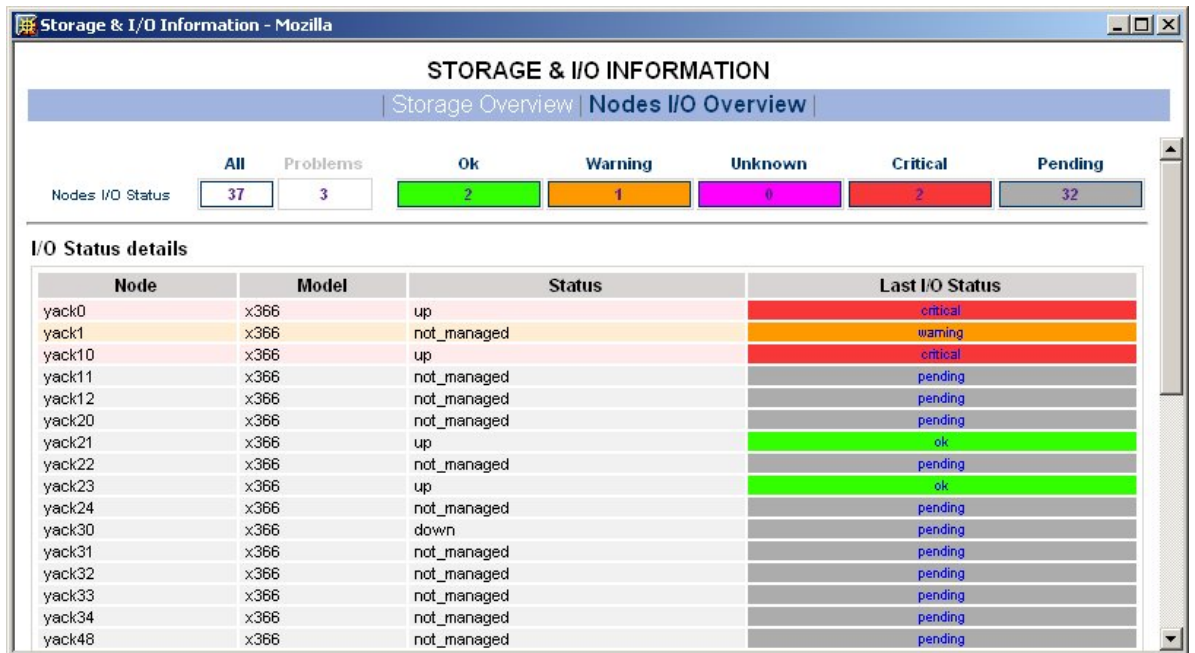


Figure 9-6. Nodes I/O Overview

### 9.3.3 Querying the Cluster Management Data Base

The **storstat** command obtains status information from the **ClusterDB** and formats the results for storage administrators.

Please refer to the help page for this command for more information:

```
storstat -h
```

The following paragraphs describe the most useful options.

#### 9.3.3.1 Checking Storage System Status

To display all the registered storage systems with their status and location in the cluster use the command below. The location is based on rack label and position in the rack:

```
storstat -a
```

To display a list of faulty storage systems:

```
storstat -a -f
```

To check the status of a storage system using the name identifying the storage system:

```
storstat -a -n <disk_array_name> -H
```

### 9.3.3.2

## Checking Status of Hardware Elements

To display a list of faulty components for all registered storage systems:

```
storstat -d -f -H
```

For each element, the following information is displayed:

- Disk array name
- Enclosure of the disk array housing the component
- Type of the component
- Status of the component
- Location of the component within the enclosure or disk array. This location uses vendor specific terminology
- Location of the enclosure (or disk array) in the cluster.

The `-n <disk_array_name>` flag can be used to restrict the list to a single storage system.

To display a list of all the components for a storage system:

```
storstat -d -n <disk_array_name>
```



#### Note:

If the `-n` flag is omitted, the list is extended to all the registered storage systems.

To check the number of available or faulty elements in the cluster (or in a selected storage system):

```
storstat -c
```

or

```
storstat -c -n <disk_array_name>
```



## 9.4 Monitoring Brocade Switch Status

Each Brocade Fibre Channel switch is monitored by **NovaScale Master - HPC Edition**.

The same check period as for Ethernet switches will be used (10 minutes, possibly configurable). No specific configuration is required on the FC switches in order to be able to use the telnet interface.

Several generic services are defined for brocade switch. They reflect the global status of a class of components of the selected switch.

A mapping between SNMP MIB (Management Information Base) values available and returned from the switch and NS MASTER HPC status give the following set of states for each managed services:

### Ethernet interface Service

---

<b>OK</b>	<b>No problem</b> Criteria: <ul style="list-style-type: none"><li>The Fping of the Ethernet interface is OK</li></ul>
<b>CRITICAL</b>	Criteria: <ul style="list-style-type: none"><li>The Fping of the Ethernet interface is KO</li></ul>

---

### FC port

---

<b>OK</b>	<b>No problem</b> Criteria: <ul style="list-style-type: none"><li>All FC ports are OK.</li></ul>
<b>WARNING</b>	<b>Maintenance operation must be scheduled</b> Criteria: <ul style="list-style-type: none"><li>Not in critical status</li><li>Some ports have a warning status</li><li>Number of operating port higher than expected in the DB (fc_switch.oper_port_threshold)</li></ul>
<b>CRITICAL</b>	<b>Degraded service. Maintenance operation mandatory</b> Criteria: <ul style="list-style-type: none"><li>One or more ports are in a critical status.</li><li>Number of operating ports lower than expected (fc_switch.oper_port_threshold)</li></ul>
<b>UNKNOWN</b>	<b>Can't access the status</b>
<b>PENDING</b>	<b>Not yet initialized</b>

---

## Fans

---

OK	<b>No problem</b> Criteria: <ul style="list-style-type: none"><li>• All fans are present and OK</li></ul>
WARNING	<b>Maintenance operation must be scheduled</b> Criteria: <ul style="list-style-type: none"><li>• Some fans are in the warning state</li><li>• Does not meet the criteria for critical status.</li></ul>
CRITICAL	<b>Degraded service. Maintenance operation mandatory</b> Criteria: <ul style="list-style-type: none"><li>• At least one of the fan is in a critical state</li></ul>
UNKNOWN	<b>Can't access the status</b>
PENDING	<b>Not yet initialized</b>

---

## Power Supply

---

OK	<b>No problem</b> Criteria: <ul style="list-style-type: none"><li>• All power supplies are present and ok</li><li>• No Power Supply is detected on the switch.</li></ul>
WARNING	<b>Maintenance operation must be scheduled</b> Criteria: <ul style="list-style-type: none"><li>• Some power supplies are in the warning state</li><li>• Does not meet the criteria for critical status.</li></ul>
CRITICAL	<b>Degraded service. Maintenance operation mandatory</b> Criteria: <ul style="list-style-type: none"><li>• At least one of the power supplies is in a critical state</li></ul>
UNKNOWN	<b>Can't access the status</b>
PENDING	<b>Not yet initialized</b>

---

## Temperature Sensor

---

OK	<b>No problem</b> Criteria: <ul style="list-style-type: none"><li>• All Temperature sensor are present and OK</li></ul>
WARNING	<b>Maintenance operation must be scheduled</b> Criteria: <ul style="list-style-type: none"><li>• Some Temperature sensor are in the warning state</li><li>• Does not meet the criteria for critical status.</li></ul>
CRITICAL	<b>Degraded service. Maintenance operation mandatory</b> Criteria: <ul style="list-style-type: none"><li>• At least one of the Temperature Sensor is in a critical state</li></ul>

---

UNKNOWN	Can't access the status
PENDING	Not yet initialized

### Global Status

OK	<p><b>No problem</b></p> <p>Criteria:</p> <ul style="list-style-type: none"> <li>Global brocade switch status is ok.</li> </ul>
WARNING	<p><b>Maintenance operation must be scheduled</b></p> <p>Criteria:</p> <ul style="list-style-type: none"> <li>Some of the other services are warning (but none critical).</li> <li>Switch name (switchX) different as expected (fcswwX)</li> </ul>
CRITICAL	<p><b>Degraded service. Maintenance operation mandatory</b></p> <p>Criteria:</p> <ul style="list-style-type: none"> <li>One of the other services is critical.</li> <li>The storage system has detected a critical error which is not reported by one of the other services.</li> </ul>
UNKNOWN	Can't access the status
PENDING	No yet initialized

The different services managed by NovaScale Master HPC for the brocade switch are shown below:

Host	Service	Status	Last Check	Duration	Attempt	Status Information
fcsww0c0	Ethernet interfaces	OK	28-02-2006 11:01:15	0d 3h 4m 35s	1/1	down : [ ] - up : [ 10.0.0.90 ]
	FC ports	CRITICAL	28-02-2006 11:01:15	0d 3h 4m 35s	1/1	8 FC ports - OK [ 0 1 3 4 5 6 7 ] - WARNING [ 2 ], Number of operating ports (2) lower than expected (4)
	Fans	OK	28-02-2006 11:01:15	0d 3h 4m 35s	1/1	All 3 Fans are OK
	Power supply	OK	28-02-2006 10:56:50	0d 18h 5m 23s	1/1	All 1 Power Supplies are OK
	Status	CRITICAL	28-02-2006 11:01:15	0d 0h 13m 20s	1/1	Global switch status is CRITICAL
	Temperature	OK	28-02-2006 11:01:15	0d 3h 4m 35s	1/1	All 4 Temperature Sensors are OK

Figure 9-7. Detailed Service status of a brocade switch

## 9.5 Managing Storage Devices with Bull CLI

This section introduces the commands provided for each family of devices.

These commands offer the most useful subset of management features, implemented in each storage system.

For Storage systems not listed in the next paragraph the administration will be done via the tools delivered with the Storage System.

### 9.5.1 Bull FDA Storage Systems

The administrator must be familiar with the **FDA** terminology and management tasks.



**Note:**

See the Bull FDA documentation for the **StoreWay FDA** model for more information on the options, parameters and possible values.

The **nec\_admin** command usually requires at least two input parameters:

- The IP address (or host name) of the Windows system which hosts the FDA Storage Manager for the target FDA system.
- The name of the target FDA system.

The following services are provided by the command:

- **rankbind**
- **ldbind**
- **addldset**
- **addldsetld**
- **sparebind**
- **sparerebuild**
- **dellldset**
- **ldunbind**
- **rankunbind**
- **spareunbind**
- **unconfig**
- **getstatus**
- **direct**

All the FDA arrays are supposed to be manageable using a single login / password. The **nec\_admin** command enforces the parameters defined in the **/etc/storageadmin/nec\_admin.conf** file as follows:

```
# NEC CLI Command path

# On Linux iSMpath="/opt/iSMSMC/bin/iSMcmd"
# On Windows iSMpath="/cygdrive/c/Program
\Files/FDA/iSMSM_CMD/bin/iSMcmd"
iSMpath = /opt/iSMSMC/bin/iSMcmd
#iSMpath="/cygdrive/c/Program\ Files/FDA/iSMSM_CMD/bin/iSMcmd"
```

```
# NEC iStorage Manager host Administrator
hostadm = administrator
# NEC iStorage Manager administrator login
necadmin = admin
# NEC iStorage Manager administrator password
necpasswd = adminpassword
```

For more information, read the man page or check the command's help.

## 9.5.2 DataDirect Networks Systems - DDN Commands



### Note:

This section only applies to clusters which include DDN Storage and have had the Bull **BAS4 for Xeon Add-on CD** software installed and configured on the Management Node.

The administrator must be familiar with the DDN S2A terminology and management tasks. If necessary the administrator must refer to the documentation provided with S2A storage systems in order to understand the options, parameters and possible values.

The DDN specific commands usually require at least one input parameter:

- The IP address (or host name) of the target singlet for the command.

### 9.5.2.1 `ddn_admin`

This command allows you to get information from a singlet, or to configure the singlet. The following services are provided by the `ddn_admin` command:

- `deletelun`
- `formatlun`
- `getinfo`
- `getfmtstatus`
- `getstatus`
- `setlun`
- `setzoning`
- `shutdown`
- `showall`
- `setcache`

### 9.5.2.2 `ddn_stat`

This command is used to collect statistical information.

The following services are provided by the `ddn_stat` command:

- `getbasic`
- `getlength`
- `repeatIO`
- `repeatMB`

For more information, read the man page or check the command's help.

### 9.5.2.3 `ddn_init`

This command is used for the initial setup of a singlet or a couplet. It must be used very carefully as it usually restarts the singlet(s).

The command uses the information preloaded in the ClusterDB. Some parameters may be overwritten using the command line.

`ddn_init` connects to each singlet through the serial port, using `conman`. Thus, it may be necessary to provide the name of the `conman` console.

A login / password is required to modify the singlet configuration. `ddn_init` attempts to connect with factory defaults login / password, using a command line supplied login / password, and with the login / password defined in `/etc/storageadmin/ddn_admin.conf`. The `ddn_admin` command then enforces the login / password defined in `ddn_admin.conf`.

### 9.5.2.4 `ddn_conchk`

This command checks the connections to a DDN system, and compares them with the connections predefined in the **ClusterDB**.

`Conman`, the serial network and the LAN must be ready for use in order to check the Serial/Ethernet consistency.

Attached nodes must be up, running, and reachable from the management station to check the fibre channel consistency.

### 9.5.2.5 `ddn_set_up_date_time`

This command is used to update the date and time of DDN subsystems with the UTC date and time of the management station. The administrator can specify a list of DDN systems to be synchronized.

A recommended practice, which is the installation default, is to periodically synchronize all DDN systems using a daily `cron`.

### 9.5.2.6 `ddn_check_format`

This command allows you to check the formatting status for a list of DDN systems.

### 9.5.2.7 `ddn_firmup`

This command automatically upgrades the firmware of the singlets of a DDN system. The Management Node can be used as TFTP server.

### 9.5.3 Bull Optima 1200 Storage Systems

The administrator must be familiar with the OPTIMA 1200 Storage System terminology and management task.



See the BULL OPTIMA 1200 Storage System documentation for more information on the options, parameters and possible values.

The **xyr\_admin** command usually requires at least one input parameter:

- The IP address of the controller of the target OPTIMA 1200.

The following services are provided by the command:

- getstatus
- list
- checkformat
- luninfo
- zoninfo
- poolbind
- ldbind
- sparebind
- setldmap
- setldwwn
- poolunbind
- ldunbind
- spareunbind
- unconfig

The OPTIMA 1200 are managed using a single login/password. The **xyr\_admin** command uses the parameters that are defined in the `/etc/storageadmin/xyr_admin.conf` file as follows:

```
# XYRATEX host Administrator (where the CLI is installed)
xyr_cli_ip = 127.0.0.1
xyr_cli_user = root

# OPTIMA 1200 Storeway Master Administrator login
xyradmin = admin

# OPTIMA 1200 Storeway Master Administrator password
xyrpasswd = password
```

For more information, read the man page or check the command's help.

## 9.5.4 EMC/Clariion (DGC) Storage Systems

The administrator must be familiar with EMC/Clariion terminology and management tasks. See the **Navisphere®** CLI documentation for more information on options, parameters and possible values.

The **dgc\_admin** command is used to get information or configure an EMC/Clariion disk array.

The storage system to be managed is recognized using one of the identifiers below:

- The IP address (or IP name) of one of the Service Processors
- The name of the storage system

The following services are provided by the **dgc\_admin** command:

- **unconfig all** - to delete the current configuration
- **unconfig zoning** - to delete the LUN access control configuration only
- **checkformat** - to check if a formatting operation is in progress
- **direct <Navisphere CLI command>** - pass-through mode for the original **Navisphere®** CLI commands

## 9.6 Using Management Tools

Please refer to the documentation of the storage systems to understand which management tools are available. Then determine how they can be accessed from Bull cluster Management Node using Linux utilities (**conman**, **telnet**, **web browser**, **X11**, **rdesktop client**, **ssh client**, etc.).



## 9.7 Configuring Storage Devices

### 9.7.1 Planning Tasks

Storage system configuration requires several planning operations. At least two steps are required.

#### STEP 1 – DEFINE THE DEVICE CONFIGURATION

The storage administrator must define the storage configuration which is required for the cluster. It is especially important for RAID storage systems, which enable the creation of logical disks (LUNs) with full flexibility in terms of number and size.

Usually, the storage configuration is a compromise of several parameters:

- The available storage resources and configuration options for the storage systems.
- The performance requirements (which may drive the choice of RAID types, LUN numbers, LUN size, striping parameters, memory cache tuning, etc.).
- The file systems and applications requirements. It is thus necessary to identify which cluster nodes will use the storage resources, the applications and/or services running on these nodes, and the system requirements for each one.

At the end of this planning phase, the administrator must be able to define for each storage system:

- The grouping of hardware disks (HDD) and the **RAID** modes to use.
- The **LUNs** to be created on each RAID volume, with their size and, if necessary, tuning parameters.
- The **LUN** access control rules. This means how the storage system should be configured to ensure that a LUN can be accessed only by the cluster node which is supposed to use this storage resource. Depending on the way the nodes are connected to a storage system, two methods of LUN access control can be used:
  1. **Port-mode LUN** access control: describes the visibility of the LUNs on each port of the storage system
  2. **WWN-mode LUN** access control: describes the visibility of the LUNs according to the initiator's worldwide name (WWN of the host fibre channel adapter). This method requires the collection of WWN information on nodes before applying the configuration on the storage systems.
- Miscellaneous tuning parameters.

#### STEP 2 – DEPLOY THE STORAGE CONFIGURATION

Changing the configuration of a storage system may not be a transparent operation for the cluster nodes using storage resources which have been configured previously.

Thus the storage administrator is advised to respect the following process when deploying a storage configuration:

- Stop all the applications accessing data on the selected storage systems.

- Unmount the file systems accessing data on the selected storage systems and, if possible, shutdown the nodes.
- Modify the storage system configuration.
- Restart the attached nodes, or force them to re-discover the modified storage resources.
- Update the node's configuration.
- Mount file systems, restart applications.

## 9.7.2 Deployment Service for Storage Systems



### Note:

This service is currently supported for FDA storage systems.

Medium and large clusters are usually built with multiple storage systems with the same hardware configuration. The purpose of the deployment service is to simplify the configuration tasks by:

- Automatically deploying the same logical configuration on multiple storage systems.
- Forcing I/O nodes to discover the storage resources and to setup a deterministic disk naming to simplify resource discovery on I/O nodes. This mechanism also ensures a persistent device naming.

This deployment service is well suited for storage systems and nodes dedicated to a single function, such as the I/O system of the cluster. It is hazardous to use it on storage systems or nodes which have multiple functions, such as nodes which are simultaneously Management Nodes and I/O nodes. Read the explanation and warnings of the next paragraphs carefully, to determine if this powerful and automated process is suitable for your cluster.

## 9.7.3 Understanding the Configuration Deployment Service

The configuration deployment service relies on modeling the storage system configuration. The model defines all the configuration parameters (see 9.7.1 Planning Tasks, Step 1). The model contains the list of the target storage systems to be configured.

The recommended process to modify the storage configuration in a large cluster, using the storage configuration deployment service, follows.



### Warning:

The administrators must follow the 3 step process described in the following paragraphs. Otherwise, there is a high risk of inconsistency between storage systems and nodes, leading to a non operational file system

## STEP 1 – DEFINE THE STORAGE CONFIGURATION

The administrator must either create a model to specify the storage configuration to deploy, or use an existing model.

The administrators can define multiple models. They are responsible for managing versions and for remembering the purpose of each model.

## STEP 2 – DISABLE THE GLOBAL FILE SYSTEM

If necessary, backup all the data that must be preserved.

Release the storage resources used on the I/O nodes. Typically, unmount and stop the global file system.

## STEP 3 – CONFIGURE THE STORAGE SYSTEMS USED BY THE GLOBAL FILE SYSTEM

The model contains all the directives to configure the storage systems. When multiple storage systems must be configured with the same configuration, the configuration operations are performed in parallel.



### Warning:

The application of a model on a storage system is destructive. The configuration deployment is not an incremental process that modifies only the differences between the current configuration and the model. The first step erases the existing configuration, and then the new configuration is built using a known reference. All data will be lost.

The application of the model stops when all the commands have been acknowledged by the target storage systems. A synthetic report is provided to quickly identify which storage devices have been successfully configured and which ones have failed.

Usually, the configuration does not complete, and tasks such as disk formatting continue to run. Another command is used to check that these tasks complete.

### 9.7.3.1

## STEP 1 - Preparing and Managing Configuration Models

The configuration model is a text file. It uses an XML syntax style. To obtain details about the syntax, the administrator can refer to the **template.model** file, delivered with the rpm in `/usr/share/doc/storageadmin-framework-<version>`.

Another way to obtain a model template is to use the following command:

```
stormodelctl -c showtemplate
```

This template describes one LUN model for each supported storage system vendor (some parameters are vendor-specific).

A model is identified by its file name. The **.model** suffix is mandatory and a recommended practice is to store all the models in the same directory. The ClusterDB contains a history of the models applied to the storage systems. Thus the administrators should not change the contents of a model without changing its name.

A global model is made up of a list of LUN models.

A LUN model is a description of the configuration to be applied to a storage system; it includes:

- A description of LUNs using an associated label.
- LUN Access control rules describing the LUNs visibility for host ports.
- Storage system tuning parameters.
- A list of the storage systems to configure using the LUN model.

### 9.7.3.2 STEP 2 – Disabling the Global File System

Before changing the configuration of storage systems, it is mandatory to stop I/O activity, stop the global file system and unmount the local file systems on the nodes attached to the storage systems.

### 9.7.3.3 STEP 3 - Applying a Model to Storage Systems



**Note:**

It is possible to skip the storage system configuration phase and to use only the I/O Node configuration phases. In this case the administrator must manually configure the storage system, in accordance with the configuration defined in the model. This way of operating is also useful when the administrator does not want to erase the existing configuration (for example to safeguard existing data), or for the storage systems that do not support the automatic configuration.

The application of a configuration model to storage systems is performed in two phases:

1. The configuration of storage resources and tuning of parameters
2. The application of LUN access control directives

If the LUN access control method used is the **WWN**-mode (use of <NodePort> directives in the model file, see the model template for detailed description), it is necessary to update the cluster database with information about the Fibre Channel adapters of the cluster nodes before applying the configuration model. This is done using the following command:

```
ioregister -a
```

If the LUN access control method used is the **Port**-mode (use of <StoragePort> directives only in the model file), there is no need to use this command.

A model contains a list of storage systems to configure. The **stormodelctl** command checks the state of the storage systems in the **ClusterDB** before attempting to configure them.

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



**Warning:**

This operation destroys all the data stored in the storage systems listed in the model.



**Important:**

It may be necessary to wait several minutes for the completion of the command. No message will be displayed for the intermediate steps.

The administrator can exclude storage systems from the list (**-e** flag), or add storage systems (**-i** flag).

The **stormodelctl** command returns a configuration message in the following format:

```
<disk array name> : <message>
```

The output may be processed by the **dshbak** command to reformat the results.

The administrator must check the output of the command. If errors are reported for some disk arrays, detailed analysis is required to diagnose and resolve the problem. The **stormodelctl** command can then be used to apply selectively the model on the disk arrays that have not been configured, using the **-i** flag.

The **-v** flag provides a verbose output, giving better control of the operations performed on the storage system.

The command only transmits the configuration information to the target storage systems. LUN formatting is a background task. To control the formatting process, use the **checkformat** sub-command:

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



**Important:**

Wait for the command to complete before running the next step.

Please refer to the help of the **stormodelctl** command for additional options.

## 9.8 User Rights and Security Levels for the Storage Commands

### 9.8.1 Management Node

#### Situation 1: superuser (= root) user

All the storage commands are available but it is not recommended to launch any of them as the root user for obvious security reasons.

#### Situation 2: non root user

**Nagios user:** The storage views, like all the NSMASTER- HPC web pages, are only accessible for the Nagios user who is automatically created during the installation/configuration of the cluster – see Chapter 3 *Cluster Database Management* for more details.

Any specific security rules/access rights will have been applied to the storage commands. Therefore, the non root users, for example, admin, must be part of the **dba** group, and the Nagios supplementary group, in order to be able to launch storage commands.

For example:

```
useradd -g dba -G nagios <username>
```

Some of these **dba** restricted access commands must be used with the **sudo** command in order to have root privileges. The reason why this privilege is restricted is that these commands may access other nodes, in addition to the MANAGEMENT node, by using **ssh**, to get or set information.

The following commands must be launched with **sudo**:

- **iorefmgmt**
- **ioregister**
- **lsiodev**
- **lsiocfg**
- **stordepha**
- **storioha**
- **stordepmap**
- **stormap**
- **stormodelctl**



**Note:**

**sudo** is a standard linux command. It allows a permitted user/group to execute a command as the superuser or as another user, as specified in the **/etc/sudoers** file which is managed by the superuser only. This file contains a list of groups/commands which have these root privileges. Refer to the **sudo** man pages for more information. To use a command with **sudo**, the command has to be prefixed by the word 'sudo' as in the following example:

```
<prompt>: sudo /usr/sbin/iorefmgmt
```



**Note:**

The PATH of the **dba** "username" must be completed in order to access these root commands without the absolute PATH in the sudo command :

```
export PATH=$PATH:/usr/sbin in the $HOME/.bashrc of login "username"
```

– The sudo command is:

```
<prompt>: sudo iorefmgmt
```

## 9.8.2 Other Node Types

All the available storage commands can only be launched as the root user, without exception.

## 9.8.3 Configuration Files

The configuration files, which an administrative user of the **dba** group can modify manually, are located in the **/etc/storageadmin/** directory of the management node.

The files are:

<b>nec_admin.conf</b>	Specific to a NEC configuration
<b>storframework.conf</b>	General configuration file for storage management





---

# Chapter 10. Kerberos - Network Authentication Protocol

**Kerberos** is an optional security suite product which can be used to validate the identity of users, services and machines for a whole network. Kerberos is included within the **Red Hat Enterprise Linux 4** delivery.

The purpose of this chapter is to describe how to implement Kerberos on a HPC cluster.

## 10.1 Environment

### 10.1.1 Kerberos Infrastructure

There exist 3 types of machine within the Kerberos infrastructure:

- The Kerberos server with the Key Distribution Centre (**KDC**) keys server and administration server housed on a server called **secu0** (by default, for a HPC cluster, this will be part of the Management Node).
- The servers containing one or more applications which are protected by Kerberos; these servers are named **secui**. A Kerberos configuration file is shared with the Kerberos server.
- The Kerberos client machines. These are not used until Kerberos authenticates the user's rights to access the applications on **secui** and to **secu0**.

### 10.1.2 Validating the Installation

The settings made by Kerberos will be validated by the telnet access authentication. The server **secu1** hosts the remote telnet service. The telnet connection to **secu1** will be made by **secu0**.

### 10.1.3 Authentication of the SSH V2 Connections

The remote service SSH will be installed on **secu1** with a connection to a SSH client from **secu0**.

## 10.2 KERBEROS Infrastructure Configuration

### 10.2.1 secu0 Server including KDC Server and Administration Server

Verify the installation of the latest version of the Kerberos RPM on **secu0**.



#### Important:

For security reasons, the Kerberos package is compiled with the option **-without-krb4** to prohibit compatibility with Kerberos 4.

### 10.2.2 Configuration Files

#### [/etc/krb5.conf](#)

This file containing the details of the KDC addresses and the administration server will be recopied on all the servers containing kerberized applications as well as on all the client machines.

```
...
[libdefaults]
default_realm = DOMAIN.COM
...

[realms]
DOMAIN.COM = { kdc=secu0:88
                admin_server = secu0:749
                default.domain=domain.com
                }

[domain.realm] .domain.com=DOMAIN.COM
                domain.com=DOMAIN.COM
...
```

#### [/var/kerberos/krb5kdc/kdc.conf](#)

This file, containing amongst other things the information necessary to produce the tickets, is specific to the Kerberos server.

```
...
[realms]
DOMAIN.COM={
preauth=yes
max_life= 2d
max_renewable_life= 10d
...
}
```

## 10.2.3 Creating the Kerberos Database

Use the following command the Kerberos database.

```
/usr/kerberos/sbin/kdb5_util create -s
enter KDC database master key : XXXX
```

## 10.2.4 Creating the Kerberos Administrator

The KDC server may be administered from any network machine using the command **kadmin** as long as the user's identity is authenticated.

As the Kerberos administrator node does not exist initially it is possible to connect the first time using root with the command **kadmin.local** on the KDC server. It is not possible to authenticate oneself with this command as one is logged onto the KDC server.

```
/usr/kerberos/sbin/kadmin.local
kadmin.local : addprinc user/admin
enter PW : YYYY
```

Now one will be able to authenticate oneself as «user» from any Kerberos client machine (the account Unix «user» must have been created as above) in order to connect to the administrator server, and to manage Kerberos assuming the admin demon has been launched. See below for more details.



### Important:

The Kerberos administrators which have been created – «user» in the example above - must belong to the root group in order to be able to reach and modify Kerberos files.

## 10.2.5 Starting the KDC Server

Use the following command to start the KDC server:

```
/sbin/service krb5kdc start
```

Verifying the local connection to Kerberos on the KDC server using «user»'s administrator access rights:

```
/usr/kerberos/bin/kinit user/admin
```

```
kinit(V5) : Cannot resolve network address for KDC in requested
realm while getting initial credentials
```

The problem in the above message is one of confirming «user»'s credentials and should be resolved by replacing **secu0** by its IP address in the **krb5.conf** file.

```
/usr/kerberos/bin/kinit user/admin  
  
PW : YYYY
```

If there is no error message then everything is OK and the administrator «user» will obtain a Ticket-Granting Ticket (TGT).

## 10.2.6 Adding Access Control List (ACL) Rights for the Kerberos Administrator Created

In the file `/var/kerberos/krb5kdc/kadm5.acl`, add the line by:

```
user/admin @DOMAIN.COM *
```

## 10.2.7 Starting the Administration Daemon

Use the following command to start the administration daemon.

```
/sbin/service kadmin start
```

It should now be possible to connect to the system and to administer the KDC server with a view to specifying the principals. A principal is an entity in the Kerberos realm – every user, instance and service in the Kerberos realm has a designated principal. The creation of principals can be done from any network machine using – in the example above - the administrator's access rights for user/admin.

## 10.2.8 Creating Principals Associated with Users

The Kerberos Administrator will create on the KDC server the principals associated with users. These users must have associated UNIX accounts existing on the client machines.

The Kerberos Administrator can create the principals either:

- Locally on the KDC (using the command `kadmin.local`) without needing to authenticate himself.

or

- From another network machine (for example a client machine) using the command `kadmin` as long as he has authenticated himself as a principal with administration rights and the administration demon has been launched. For example, for user Durand:

```
kadmin.local  
PW : YYYY  
  
kadmin : addprinc durand  
PW : ZZZZ (add the user password on the client machines)  
Principal " durand@DOMAIN.COM " created
```

For a principal user the secret key shared between the KDC and the client machine is derived from the user's password.

The process has to be repeated for all other users.

## 10.2.9 Creating Principals Associated with Remote Kerberized Services

The principals associated with services have to be created. However, MIT provides some basic services which have already been kerberized in their Kerberos distribution. For the **ftp**, **telnet**, and **rsh** services (included as part of the default installation using the package `krb5-workstation`), the principal associated with them generic and is called **host principal**.

This **host principal**, whose name is derived from the name of the machine, can be used for Kerberos Authentication of the basic kerberized services - `rlogin`, `telnet`, etc. residing on this host.

Creation of the host principal for the server **secu1**:

Connect to **secu0** or to **secu1** with the **kadmin** command.

```
kadmin.local
addprinc -randkey host/secu1.domain.com
```



### Important:

The hostname has to correspond with its first occurrence in the line associated with the machine in the file `/etc/hosts`.

## 10.3 Configuring the secu1 Machine Hosting the Remote Service 'host principal'

Verify the installation of the latest version of the Kerberos RPMs on **secu1**.

Copy the configuration file `/etc/krb5.conf` from **secu0** to **secu1**, and to any other machines which may be on the system.

### 10.3.1 Generating the Key Associated with the Remote Service 'host principal'

This secret key is shared between the KDC server **secu0** and the server housing the remote service **secu1**. This is necessary so that **secu1** can decipher the Kerberos tickets which are transmitted to it. The key can be created on any of these 2 servers but must then be copied from one to the other.



#### Important:

The file for the keys defined in the configuration file `kdc.conf` is as follows:

```
/var/kerberos/krb5kdc/kadm5.keytab
```

The file for the keys used by the command `kadmin` is as follows:

```
/etc/krb5.keytab
```

Therefore the name of the file has to be modified in the `kdc.conf` file or a link has to be made between the files as follows:

```
ln -s /var/kerberos/krb5kdc/kadm5.keytab /etc/krb5.keytab
```

Connect as a Kerberos administrator («user») to **secu0**:

```
kadmin  
ktadd host/secu1.domain.com
```

Then recopy the key to **secu1**.



#### Important:

When working it is recommended to have a keytab file for each service and to store on each server only the keys associated with the remote services hosted on the server and not the keys whose services are hosted by other servers. However, the KDC server must of course have its own specific keytab file for all the keys of the remote services

## 10.4 Validating Kerberos Authentication for the Telnet Service

Command to create the TGT ticket for a user connected as Durand on **secu0**:

```
kinit
PW : xxxx (password user durand)
klist
....
```

### Launching a telnet server

**telnetd** has to be launched on **secu1**.

For the command:

```
/etc/xinetd.d/krb5-telnet
```

Add the arguments **server\_args = -a user** to force Kerberos Authentication when the client connects using telnet. Set **disable = no** to launch the **krb5-telnet** demon when next starting **xinetd**.

It is also possible to set **disable = yes** in the file **/etc/xinetd.d/krb5-telnet** to prevent telnet from starting.

Restart **xinetd** using the command:

```
/etc/rc.d/init.d/xinetd restart
```

Using the command **chkconfig --list** verify that the daemon **krb5-telnet** is being used and not the usual demon **krb5**.

To test the telnet connection from the client machine client **secu0** to the server **secu1** hosting the telnet service for the user Durand on **secu0**, use the following command:

```
telnet -a secu1
```

The connection has to be confirmed – without a password being provided – by the following message:

```
Kerberos accepts you as " durand@DOMAIN.COM "
```

If a password is used then a **Kerberos** Authentication is not made and the password will be available across the network.

## 10.5 Kerberos Authentication and SSH

The remote service SSH is installed on **secu1** with a SSH client connection from **secu0**.

### 10.5.1 Configuring the Server SSH on the Machine secu1

A typical **sshd\_config** configuration file will contain the following:

```
#      $OpenBSD: sshd_config,v 1.69 2004/05/23 23:59:53 dtucker Exp $

# This is the sshd server system-wide configuration file.  See
# sshd_config(5) for more information.

# This sshd was compiled with PATH=/usr/bin:/bin:/usr/sbin:/sbin:/usr/local/bin

# The strategy used for options in the default sshd_config shipped with
# OpenSSH is to specify options with their default value where
# possible, but leave them commented.  Uncommented options change a
# default value.

Port 22
Protocol 2
ListenAddress xxx.xxx.xxx.xxx
#ListenAddress ::

# HostKey for protocol version 1
#HostKey /usr/local/etc/ssh_host_key
# HostKeys for protocol version 2
#HostKey /usr/local/etc/ssh_host_rsa_key
#HostKey /usr/local/etc/ssh_host_dsa_key

# Lifetime and size of ephemeral version 1 server key
#KeyRegenerationInterval 1h
#ServerKeyBits 768

# Logging
#obsoletes QuietMode and FascistLogging
#SyslogFacility AUTH
#LogLevel VERBOSE

# Authentication:

#LoginGraceTime 2m
#PermitRootLogin yes
#StrictModes yes
#MaxAuthTries 6

RSAAuthentication no
PubkeyAuthentication no
```



```

#AuthorizedKeysFile .ssh/authorized_keys

# For this to work you will also need host keys in /usr/local/etc/ssh_known_hosts
RhostsRSAAuthentication no
# similar for protocol version 2
HostbasedAuthentication no
# Change to yes if you don't trust ~/.ssh/known_hosts for
# RhostsRSAAuthentication and HostbasedAuthentication
#IgnoreUserKnownHosts no
# Don't read the user's ~/.rhosts and ~/.shosts files
#IgnoreRhosts yes

# To disable tunneled clear text passwords, change to no here!
PasswordAuthentication no
PermitEmptyPasswords no

# Change to no to disable s/key passwords
#ChallengeResponseAuthentication yes

# Kerberos options
KerberosAuthentication yes
# If the Kerberos authentication is denied, an Authentication password is not
# provided for the user :
KerberosOrLocalPasswd no
KerberosTicketCleanup yes

# GSSAPI options
GSSAPIAuthentication yes
GSSAPICleanupCredentials yes

# Set this to 'yes' to enable PAM authentication, account processing,
# and session processing. If this is enabled, PAM authentication will
# be allowed through the ChallengeResponseAuthentication mechanism.
# Depending on your PAM configuration, this may bypass the setting of
# PasswordAuthentication, PermitEmptyPasswords, and
# "PermitRootLogin without-password". If you just want the PAM account and
# session checks to run without PAM authentication, then enable this but set
# ChallengeResponseAuthentication=no
UsePAM yes

#AllowTcpForwarding yes
#GatewayPorts no
#X11Forwarding no
#X11DisplayOffset 10
#X11UseLocalhost yes
#PrintMotd yes
#PrintLastLog yes
#TCPKeepAlive yes
#UseLogin no

```

```

#UsePrivilegeSeparation yes
#PermitUserEnvironment no
#Compression yes
#ClientAliveInterval 0
#ClientAliveCountMax 3
#UseDNS yes
#PidFile /var/run/sshd.pid
#MaxStartups 10

# no default banner path
#Banner /some/path

# override default of no subsystems
Subsystem sftp /usr/local/libexec/sftp-server

```



### Important:

The following pre-requisites apply:

- The file `/etc/hosts` of the remote machine with which the ssh connection is being made has to have its hostname in the form:  
x.x.x.x secul.domain.com secul
- The hostname of the remote machine may be of the form:  
secul.domain.com or secul.
- The principal service associated with this machine has to be the same as its Fully Qualified Domain Name **FQDN**:  
secul.domain.com.

## 10.5.2 SSH Client

On the machine **secu0** or another machine a typical `ssh_config` file appears as follows:

```

#           $OpenBSD: ssh_config,v 1.19 2003/08/13 08:46:31 markus Exp $

# This is the ssh client system-wide configuration file.  See
# ssh_config(5) for more information.  This file provides defaults for
# users, and the values can be changed in per-user configuration files
# or on the command line.

# Configuration data is parsed as follows:
# 1. command line options
# 2. user-specific file
# 3. system-wide file
# Any configuration value is only changed the first time it is set.
# Thus, host-specific definitions should be at the beginning of the
# configuration file, and defaults at the end.

```

```
# Site-wide defaults for various options

# Host *
# ForwardAgent no
# ForwardX11 no
RhostsRSAAuthentication no
RSAAuthentication no
PasswordAuthentication no
HostbasedAuthentication no
# BatchMode no
# CheckHostIP yes
# AddressFamily any
# ConnectTimeout 0
# StrictHostKeyChecking ask
# IdentityFile ~/.ssh/identity
# IdentityFile ~/.ssh/id_rsa
# IdentityFile ~/.ssh/id_dsa
# Port 22
# Protocol 2,1
# Cipher 3des
# Ciphers aes128-cbc,3des-cbc,blowfish-cbc,cast128-cbc,arcfour,aes192-
cbc,aes256-cbc
# EscapeChar ~
Port 22
Protocol 2
GSSAPIAuthentication yes

# Pour le forwarding des tickets :
GSSAPIDelegateCredentials yes
```



**Note:**

The forwarding of TGT tickets by **ssh** is activated by the parameter **GSSAPIDelegateCredentials yes** for the ssh client file.

## 10.6 Troubleshooting Errors

```
Error : " Permission denied (gssapi-with-mic,keyboard-interactive) "
```

There are several possible causes for this error:

- The target machine must have its **full name** in its **/etc/hosts** file as shown below:

```
@IP      secul.domain.com  secul
```

If several names are associated with the same IP address the name with which one is connecting has to be at the top of **/etc/hosts** as shown below:

```
@IP parallel.domain.com  parallel
@IP secul.domain.com    secul
```

- Verify that the file **/etc/krb5.conf** is identical on the KDC server and the SSH servers and clients.
- Verify that the keys in the file **/etc/krb5.keytab** are identical on the KDC server and the SSH server.
- Verify that the user has a valid TGT ticket.

## 10.7 Generating Associated Keys for Nodes of a Cluster

The Perl program, below, generates on the Kerberos server, hosted on the management node, the Kerberos key (keytab) associated with each node and then transfers it to the node using Secure Copy (SCP) security (confidentiality and Authentication ensured by private key / public key).

The pre-requisite here is the preliminary installation of a private key / private key infrastructure between the management node and each compute node thus ensuring the secure transfer of the Kerberos keys.

The file `/etc/ssh/ssh_config_public` is the configuration file for the SSH client which uses Authentication by the public key/private key protocol.

```
#!/usr/bin/perl -w

print "Lower limit of cluster nodes: ";
$inf = <STDIN>;
chomp ($inf);
print "Upper limit of cluster nodes: ";
$sup = <STDIN>;
chomp ($sup);

my $serv="secu";
my $serv0="secu";
my $keytab="_keytab";
my $krb5_keytab=":/etc/krb5.keytab";

# Key creation for each node of the cluster
# Each key is generated on the management node and is stored in a temporary
# file (and also in the KDC base) ; this file will then be recopied on the
# associated node;
# The remote recopy by SCP will be secured by public key/ private key.

for ($i=$inf; $i <=$sup; $i++) {
    $serv="$serv0$i";
    print("Création keytab pour $serv\n");
    system ("rm -f /tmp/$serv$keytab");
    system ("kadmin.local -q 'ktadd -k /tmp/$serv$keytab
        host/$serv.domain.com'");
    system ("scp -rp /tmp/$serv$keytab $serv$krb5_keytab");
    system ("rm -f /tmp/$serv$keytab");
}

print("\n----> The new keys for the nodes secu$inf to secu$sup have been
    generated \n\n");
```

## 10.8 Modifying the Lifespan and Renewal Period for TGT Tickets

The default the duration for a Ticket-Granting Ticket (**TGT**) ticket is 10 hours and it cannot be renewed while it is still active. In other words its duration must be at 0 before it is renewed.

These two time periods can be modified by a user.

For example the command below is used to change the duration of a ticket to 2 days and its renewal period to 5 days.

```
kinit -l 2d -r 5d
```

The ticket obtained with this command will be valid for 2 days and may be renewed at any time during these 2 days in order to obtain a new ticket which is also valid for 2 days up until the 5 day limit is reached.

However, the values specified by the user have to be inside the maximum values defined by the Kerberos configuration.

To modify these values in the Kerberos configuration file `/var/kerberos/krb5kdc/kdc.conf` do as follows:

In the block `[realms]`, add:

```
max_life = 2d
max_renewable_life = 10d
```

Then relaunch the `krb5kdc` and `kadmin` daemons.

- To authorize the creative entity principal for `krbtgt` tickets to deliver tickets with the values above, use the command below:

```
kadmin
modprinc -maxlife "2 days" krbtgt/DOMAIN.COM
modprinc -maxrenewlife "10 days" krbtgt/DOMAIN.COM
```

- To authorize the values for the user concerned use the command below:

```
kadmin
modprinc -maxlife "2 days" durand
modprinc -maxrenewlife "10 days" durand
```

## 10.9 Including Addresses with Tickets

By default tickets do not include addresses.

Use the following command in order that tickets generated include the addresses of the local machine.

```
add noaddresses=no in the paragraph [libdefaults] for the file
/etc/krb5.conf
```

---

## Chapter 11. Profiling Programs - HPC Toolkit

**HPC Toolkit** provides a set of profiling tools that help you to improve the performance of the system. These tools perform profiling operations on the executables and display information in a user-friendly way.

The main advantage of HPC Toolkit over other profiling tools is that you do not need to include profiling options and to re-compile the executable.



### Note:

In this section, the term “executable” refers to a Linux program file, in ELF (Executable and Linking Format) format.

### Prerequisites:

- The executable must contain debugging information (if not, there will be no correspondence between counters and code at source line level)
- The executable must be dynamically linked because HPC Toolkit overloads the default initialization functions to call PAPI.
- The executable must not use ANSI libstdc++. (Using HPC Toolkit with this type of executable produces a SIGSEGV).

### 11.1.1 HPC Toolkit Tools

HPC Toolkit includes the following tools:

**hpcrun** profiles the execution of an executable, by statistically measuring the hardware counters.

**hpcprof** interprets the profile files produced by **hpcrun**, and associates them with the source code.

**bloop** analyzes the executables to determine their structure. Its goal is to search for execution loops and to identify the corresponding source code.

**hpcview** generates high level metrics from raw profiling data and correlates it with logical source code abstractions. By default, it generates an Experiment database (ExperimentXML format) for use with **hpcviewer**.

**hpcviewer** is a graphical user interface used to view the information gathered by **hpcview** easily, particularly the links between the source code and the performance.

### 11.1.2 Display Counters

The **hpcrun** tool uses the hardware counters as parameters. To know which counters are available for your configuration, use the **papi\_avail** command:

```
papi_avail
```

Available events and hardware information.

```
-----  
Vendor string and code   : GenuineIntel (1)  
Model string and code   : 32 (1)  
CPU Revision : 0.000000  
CPU Megahertz: 1600.000122  
CPU's in this Node : 6  
Nodes in this System: 1  
Total CPU's : 6  
Number Hardware Counters : 12  
Max Multiplex Counters : 32  
-----
```

The information displayed below corresponds to fields in the `PAPI_event_info_t` structure.

Name	Code	Avail	Deriv	Description (Note)
PAPI_TOT_CYC	0x8000003b	Yes	No	Total cycles
PAPI_L1_DCM0	x80000000	Yes	No	Level1 data cache misses
PAPI_L1_ICM0	x80000001	Yes	No	Level 1 instruction cache misses
PAPI_L2_DCM0	x80000002	Yes	Yes	Level 2 data cache misses
...				
PAPI_FSQ_INS	0x80000064	No	No	Floating point square root instructions
PAPI_FNV_INS	0x80000065	No	No	Floating point inverse instructions
PAPI_FP_OPS	0x80000066	Yes	No	Floating point operations

The following counters are particularly interesting: `PAPI_TOT_CYC` (number of CPU cycles) and `PAPI_FP_OPS` (number of floating point operations).

To display more details use the `papi_avail -d` command.

## 11.1.3 Using HPC Toolkit



### Important

All the above tools should have been launched before the different results are gathered and analyzed.

### 11.1.3.1 Analyzing the executable code (bloop)

#### Syntax

```
bloop executable_name > executable_name.psxml
```

#### Output

`bloop` generates an XML file, whose type is PGM (program), which specifies the structure of the program and can be used by `hpcview` or `hpcquick`.



### 11.1.3.2 Retrieving the execution information (hpcrun)

#### Syntax

```
hpcrun -e event1[:period1] -e event2[:period2] . . . executable_name
```

**-e eventx** Specify counter names.

**periodx** Retrieve a counter during a specific time period.



#### Note:

Some counters are not compatible. To resolve this problem, specify a period of time for each counter to run using the **:period** parameter. When this option is specified **hpcrun** retrieves each counter in sequence, thus avoiding conflicts.

#### Output

**hpcrun** generates a file named `executable_name.eventx-etc.hostname.pid.tid`. This file contains all the counter values in the form of a histogram.

#### Examples

```
hpcrun -e IA64_INST_RETIRED -e L3_MISSES -e PAPI_TOT_CYC /bin/ls
```

To retrieve the `IA64_INST_RETIRED` counter for 3000 events, enter:

```
hpcrun -e IA64_INST_RETIRED:3000 -e L3_MISSES -e PAPI_TOT_CYC /bin/ls
```

### 11.1.3.3 Analyzing the execution (hpcprof)

#### Syntax

```
hpcprof [options] executable_name executable_name.event1-etc.hostname.pid.tid
```

`executable_name.event1-etc.hostname.pid.tid` is the name of the file generated by the **hpcrun** command.

#### Options

**-e, --everything** Show all information

**-f, --files** Show all files

**-r, --funcs** Show all functions

**-l, --lines** Show all lines

**-p, --profile** Specify that the output is issued in the form of a “profile” file, which can be used by **hpcview**. By default the output is in ASCII.

### Example 1

```
hpcprof executable executable.B.L3_MISSES-etc.hostname.pid.tid -f
```

```
L3_MISSES:32767 - L3 Misses (70 samples)
PAPI_TOT_CYC:32767 - Total cycles (126483 samples)

File Summary:
64.3% 44.7%
<</root/NPB3.2.1/NPB3.2-SER/bin/bt.B>>/root/NPB3.2.1/NPB3.2-SER/BT/exact_rhs.f
0.0% 40.3%
<</root/NPB3.2.1/NPB3.2-SER/bin/bt.B>>/root/NPB3.2.1/NPB3.2-
SER/BT/exact_solution.f
35.7% 12.5%
<</root/NPB3.2.1/NPB3.2-SER/bin/bt.B>>/root/NPB3.2.1/NPB3.2-SER/BT/initialize.f
0.0% 2.2%
<</lib/libgcc_s-3.4.6-20060404.so.1>>.././gcc/config/ia64/liblfuncs.asm
0.0% 0.2% <</root/NPB3.2.1/NPB3.2-SER/bin/bt.B>><unknown>
```

In this example the utilization time is specified for each file because the **-f** option was used. Use the other options to see alternative information.

### Example 2

To obtain the results in the form of a "profile" file, use the **-p** option, as follows:

```
hpcprof executable executable.B.L3_MISSES-etc.hostname.pid.tid -e
-p > profile
```

## 11.1.3.4

### Creating a HPCVIEW configuration file (hpcquick)

The HPCVIEW configuration file is required by the **hpcview** tool. It describes which data should be examined, the type of performance data to be calculated from these results, and how they should be displayed. An HPCVIEW configuration file can be created easily using the **hpcquick** tool.

#### Syntax

```
hpcquick [options] [ -I dir1 dir2 ... dirN ] [ -S struct1 struct2 ...
structJ ] [ -G group1 group2 ... groupK ] -P prof1 prof2 ... profM
```

- I dir1 dir2 ...** Specify the directories pertaining to the executable source being analyzed.
- S struct1 struct2 ....** Specify the PGM output files generated by **bloop** (.psxml).
- P prof1 prof2 ...** Specify the profile files containing the counter values analyzed by **hpcprof**.
- G group1 group2 ...** Specify the group files and/or any or all of the shared libraries that can be used by the program.
- n** This option generates the configuration file, which is used by **hpcview**.

## Example

```
hpcquick -I ../BT/ -S resbloop -P bt.B.L3_MISSES-etc.dedea.21173.0x52b5 -n
```

```
Canonicalizing performance data...
hpcprof -p bt bt.B.L3_MISSES-etc.dedea.21173.0x52b5 >
bt.B.L3_MISSES-etc.dedea.21173.0x52b5.hpcquick.pxml
* Collecting metrics from performance data...
* Generating hpcview configuration file...
  Adding source path: '../BT/'
  Adding structure file: 'resbloop'
  Adding metric 'L3_MISSES' from
  'bt.B.L3_MISSES-etc.dedea.21173.0x52b5.hpcquick.pxml'
  Adding metric 'PAPI_TOT_CYC' from
  'bt.B.L3_MISSES-etc.dedea.21173.0x52b5.hpcquick.pxml'
* Sending command to create browsable database to log...
hpcview -o experiment-db-21368 hpcquick.xml
* Created files: 'hpcquick.xml', 'hpcquick.log'
```



### Note:

The HPCVIEW configuration file is editable, in terms of the data you want to analyze and the results you want to display.

## 11.1.3.5 Collecting the results (hpcview)

**hpcview** generates high level metrics from the raw profiling data and correlates it with logical source code abstractions. By default, it generates an Experiment database file (Experiment XML format) that is used with **hpcviewer**.

### Syntax

```
hpcview [options] <Config_File> [<Profile_File1 . . .>]
```

**Profile\_File1 . . .** Name of profile files generated by **hpcprof**.

**-x File\_Name** Name of the XML file (PGM type) generated by **hpcview** from the executable's analysis.

**Config\_File** Name of the HPCVIEW configuration file describing which data should be examined, the type of performance data to be calculated from these results and how it should be displayed.

### General options

**-v, --verbose [<n>]** Verbose mode; generate progress messages to **stderr** (standard error output) at verbosity level <n>. Default: 1.  
Use n=2 to debug path replacement if metric and program structure are not properly matched.

**-V, --version** Print version information.

**-h, --help** Print the help file for the command.

## Output options

**-o <db-path>, --db <db-path>, --output <db-path>**

Specify Experiment database name <db-path>.  
Default: /experiment-db.

**--src [yes | no], --source [yes | no]**

Specify whether or not to copy the source code files into the Experiment database. By default, **hpcprof** copies source files with performance metrics that can be reached by PATH/REPLACE statements, resulting in a self-contained dataset that does not rely on an external source code repository. Note that if copying is prevented, the database is no longer self-contained.

## Output formats

Select different output formats and optionally specify the output filename **<fname>** (located within the Experiment database). The output is sparse in the sense that it ignores program areas without profiling information. (Set **<fname>** to **'!'** to write to **stdout**)

**-x [<fname>], --experiment [<fname>]**

Experiment XML format. Default: **experiment.xml**.  
Note: to disable, set <fname> to 'no'.

**--csv [<fname>]** Comma-separated-value format. Default: **experiment.csv**. (Flat scope tree; Loop level.) (Useful for downstream external tools.)

**--tsv [<fname>]** Tab-separated-value format. Default: **experiment.tsv**. (Flat scope tree; line level.) (Useful for downstream external tools.)

## Example

```
hpcview hpcquick.xml bt.B.L3_MISSES-etc.dedea.21173.0x52b5.hpcquick.pxml
```

### 11.1.3.6 Viewing the results (hpcviewer)

The **hpcviewer** tool displays the counters values for each line of code (See following figure). **hpcviewer** uses the XML file generated by **hpcquick** or **hpcview**.

The following figure shows an example.

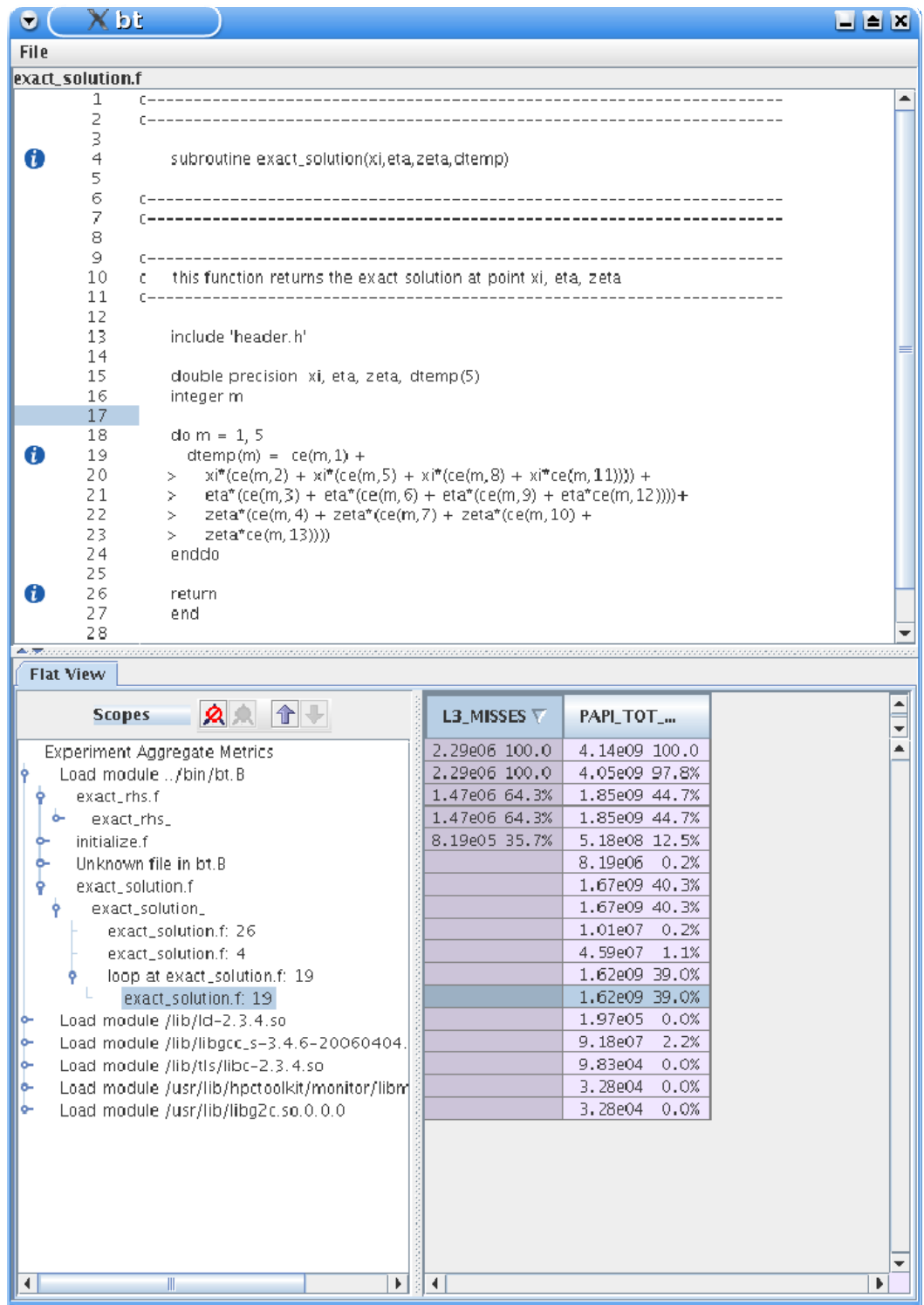


Figure 11-1. View of the counter values, using hpcviewer

## 11.1.4 More Information

For more information about HPC Toolkit go to:

<http://hipersoft.cs.rice.edu/HPC Toolkit/documentation.html>



## Chapter 12. I/O Node and Lustre File System High Availability

This chapter explains how to implement High Availability for I/O Nodes and **Lustre** file system and only applies to clusters which have installed the **Lustre** file system from the HPCK CDROM.

### 12.1 Introduction to Lustre File System

**Lustre** uses object based disks for storage. Metadata servers are used for storing file system metadata. This design provides a substantially more efficient division of labor between computing and storage resources. Replicated, failover metadata Servers (**MDSs**) maintain a transactional record of high-level files and file system changes. Distributed Object Storage Targets (**OSTs**) are responsible for actual file system I/O operations and for interfacing with storage devices. This division of labor, and of responsibility, leads to a truly scalable file system and more reliable recoverability from failures by providing a combination of the advantages of journaling and distributed file systems. **Lustre** supports strong file and metadata locking semantics to maintain total coherency of the file systems even when there is concurrent access. File locking is distributed across the storage targets (**OSTs**) that constitute the file system, with each **OST** handling locks for the objects that it stores.

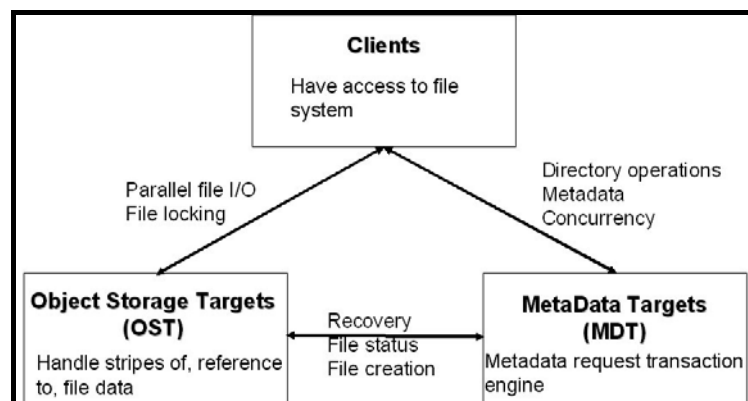


Figure 12-1. Lustre interactions



For more information, see:

Lustre: A Scalable, High-Performance File System Cluster File Systems, Inc.  
<http://www.lustre.org/docs/whitepaper.pdf>

## 12.2 Lustre Failover Mechanism

Lustre supports the notion of failover pairs. Two nodes which are connected to shared storage can function as a failover pair, in which one node is the active provider of the service (**OST** or **MDT**), and the second node is the passive secondary server.

The Lustre services are declared on both nodes with the same name. The **MDT** is configured with a list of servers (**OSSs**) for clients to pass through in order to connect to the **OSTs**. The Lustre servers must have distinct network addresses.

The failover mechanism of the Lustre system is based on the capacity to enable client reconnection when the **OSTs** and **MDTs** are moved to other nodes.

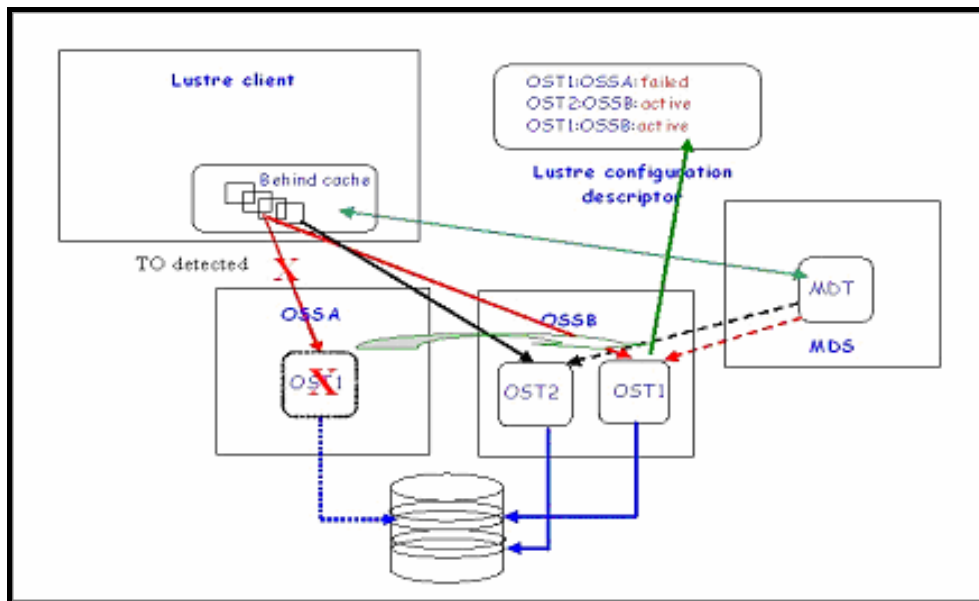


Figure 12-2. OST takeover and client recovery

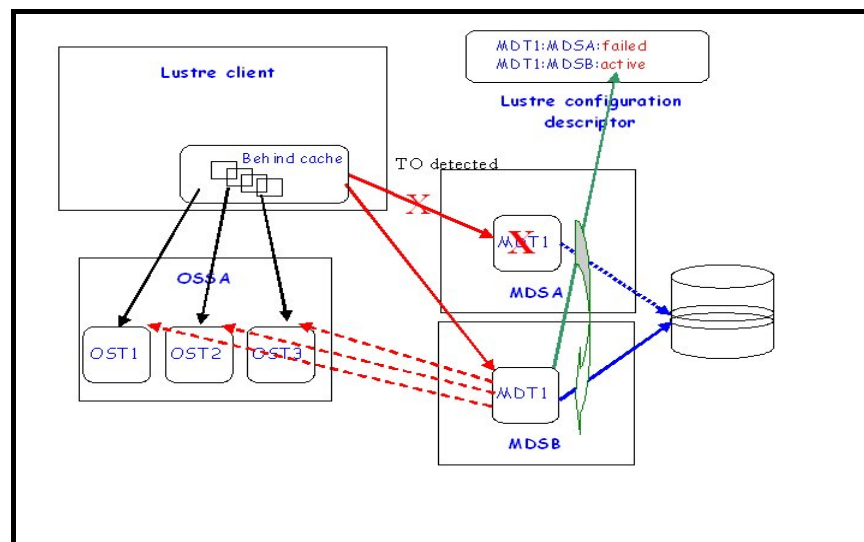


Figure 12-3. MDT takeover and client recovery



Lustre targets work in a synchronous mode: a client request is executed on the storage device and acknowledgement provided for the client only after it has been acknowledged by the device. In the case of a failure, the storage devices will ensure that committed data is preserved. Uncommitted data, meaning data not acknowledged, is kept by the clients in their "reserve cache" and can be resent when the system recovers.

When a request to a target (MDT or OST) is not acknowledged in time, the Lustre client suspects a failure and starts the recovery procedure as follows:

- Checks routing information in the Lustre configuration descriptor
- Reconnects to the migrated target
- Lost transactions are replayed
- Locks are re-acquired
- Partially completed multi-node operations are restarted.

On the server side, the target failover mechanism is similar to the High Availability service migration: the service is "stopped" on one node, and then restarted on the rescue node which has access to the same storage devices.

Transactions requested on metadata local to the targets (**ext3** log files) and those which are global for the Lustre system (MDT) are logged in log files. These files are replayed when there is a service migration.

Lustre does not prevent simultaneous access. This means that an external mechanism must ensure that the shared storage will only be accessed by one node at a time. This can be done by powering off the failed node.

## 12.3 Hardware Architecture

Bull High-Availability management of the Lustre system relies on a specific hardware architecture.

The I/O nodes operating in HA mode are grouped in I/O cells which brings together two I/O nodes that access one or more disk arrays. The I/O cell contains either **OSTs** or **MDTs**, but both are exclusive.

Usually, nodes are directly connected to the storage systems ports without intermediate switches or HUBs. This “point to point link” avoids having additional active components which may become other SPOFs.

The LUNs within the storage systems are accessible for both nodes of the I/O cell, enabling OSSs and MDSs to retrieve their data when they are moved to the other node. But each LUN must be used by only one node to avoid data corruption.

An I/O fencing mechanism is implemented so that the faulty node can not access the LUNs again after the OSSs or MDSs are restarted on the peer node of the I/O cell. In any case the failing node is powered off.

The underlying mechanism which ensures OSS and MDS migration and I/O fencing is provided by **Cluster Suite**. The failover process relies on basic entities known as failover services. When a node fails, **Cluster Suite** determines how each service should be relocated.

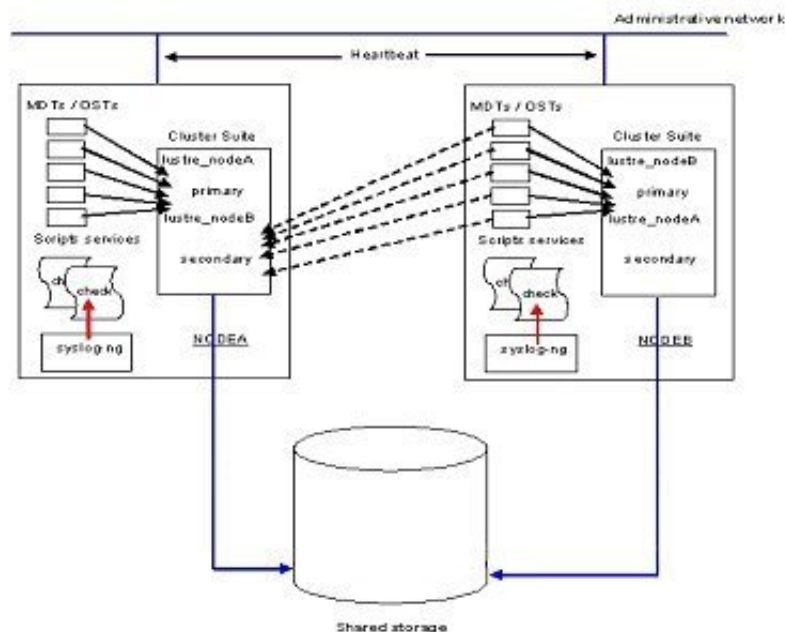


Figure 12-4. I/O Cell diagram

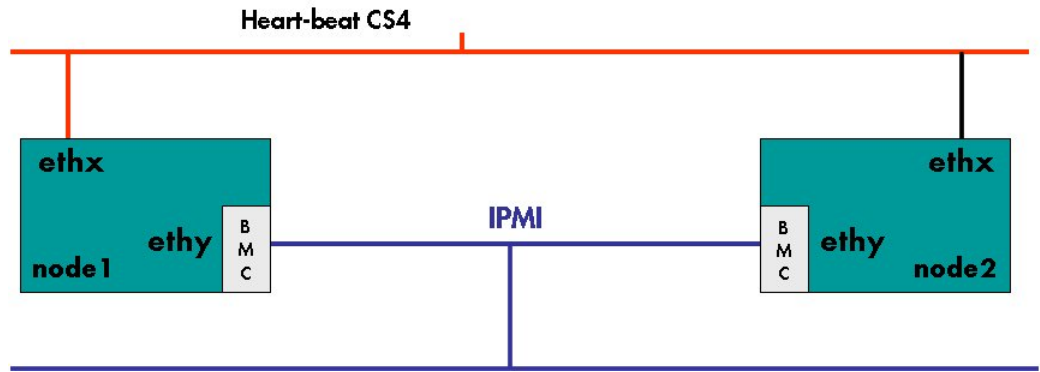


Figure 12-5. High Availability/Cluster Suite on NovaScale R440 and R460 IO/MDS nodes

In the case of multi-types I/O nodes (nodes which serve as both OSSs and MDSs), the Lustre file systems must be configured so that the same node does not support both MDT and OSTs services for the same file system. If not, a failure of this type of node constitutes a double failure (MDS + OSS) for the Lustre file system and its recovery is not guaranteed. The following figure illustrates how you can position OSTs and MDTs for two file systems **FS1** and **FS2**.

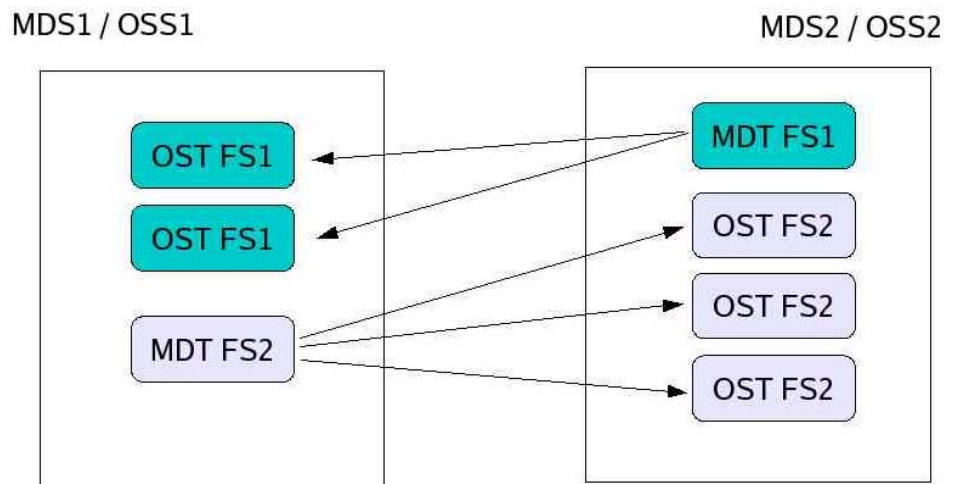


Figure 12-6. MDT/OST Dispatching on two nodes

## 12.4 High Availability Policy

In a Cluster Manager environment the customer can simply spread the application services across the clustered servers in any way that seems appropriate. All nodes will be actively running a share of the total load.

This is called **Active/Active clustering**, as all servers are active. If one server shuts down, the other servers will pick up the running load of its services.

The High Availability mode to be applied in the Bull HPC context is **mutual takeover** for each node of a pair of nodes in the same I/O cell. In standard state, each I/O node supports its own Lustre services (MDTs or OSTs), whereas in a failure state, one node manages its own services plus all the services from the failing node.

Firstly, all Lustre services from the failing node will be migrated to the second node, even if the failure concerns only one Lustre service. It means that for each I/O node, one failover service is defined which includes all the currently installed Lustre services.

Cluster Suite requires a service script for each failover service that will be managed. The script must be able to stop, start and report status for the service.

On each unitary High Availability cluster based on the I/O cells, two failover services are defined: one for the Lustre services of each node.

In the I/O Cell above we have:

**lustre\_nodeA** with primary node NODEA and secondary node NODEB

**lustre\_nodeB** with primary node NODEB and secondary node NODEA

A different failover script is associated with each of the two services provided they are not composed of the same Lustre components (MDTs / OSTs).

At any moment, the Lustre failover service on an I/O node is composed of all the Lustre services (MDTs / OSTs) associated with the **active** file systems. Its composition is subject to change according to the Lustre file systems activation.

On an I/O node, the Lustre services (MDTs / OSTs) are not started for the boot but only by Lustre administrative tools by means of file system start. This to ensure consistency of the Lustre file system services start on all the nodes it relies on.

For a reboot of a failed MDS or I/O node, the Lustre services are not automatically relocated. This is may be done only by the administrator using a Lustre management tool. This mechanism is chosen to avoid inopportune Lustre services migration when there is a partial repair of the primary node.



### Important:

A simultaneous migration of the management station and of the metadata server is considered as a double failure and is not supported.

## 12.5 High Availability Management

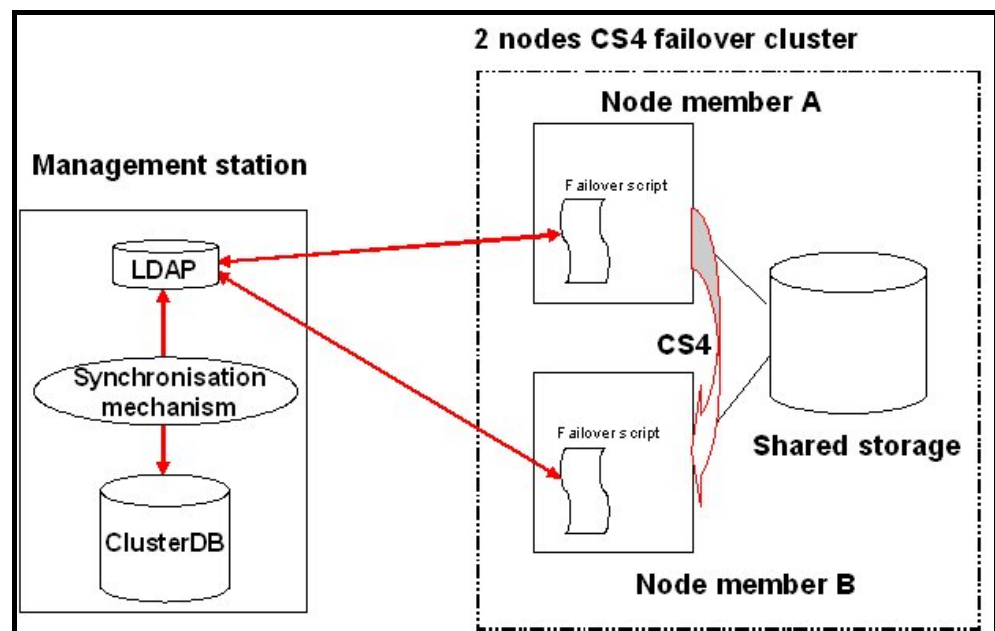


Figure 12-7 Lustr High-Availability Management architecture

When targets (MDT/OST) failover is configured, not only the information about paired targets (a cell) is stored in the Lustr configuration information of the ClusterDB, but also which instance of that target is the currently active one. It is the High Availability application (Cluster Suite script) that has to dynamically update the active instance information of the Lustr tables into the ClusterDB.

### LDAP Directory

The LDAP directory is the pivot of the Lustr High Availability architecture. It has been proved to be the more flexible and more efficient way to share Lustr configuration information on a large cluster. It is the repository of the actual file systems distribution on the cluster:

- which file systems are active,
- Lustr services (OST/MDT) migrations.

The information status it contains is updated by the Lustr failover script every time a High Availability event occurs.

This information status has to be synchronized with the clusterDB one in order to ensure file systems status and Lustr services migrations monitoring by the Lustr administration tools.

**Using such a mechanism allows Lustr file systems relying on some migrated nodes, to be stopped and restarted respecting the actual Lustr services node distribution.**

### Clients /Servers Reconnections

An 'epoch' number on every storage controller, an incarnation number on the metadata server/cluster, and a generation number associated with connections between clients and other systems form the infrastructure for Lustr recovery, enabling clients and servers to detect restarts and select appropriate and equivalent servers.

## Failover Scripts

The sole difference between both failover scripts of a HA unitary cluster is the set of Lustre services it includes. The specificity of this set is to evolve according to the Lustre file systems activation / de-activation.

A generic failover script `/usr/sbin/failover/lustre_failover` is implemented on the I/O nodes. The set of Lustre services to manage is dynamically determined by requesting the LDAP directory for active file systems having a group of services on this node.

The group to check for is always the same for a node but different for each one. To deal with that, the failover script associated with a failover service is a symbolic link to this generic script which name includes the primary node name:

```
/usr/sbin/lustre_failover_<primary_node_name>
```

The group of Lustre services to check for is determined by parsing the script calling name. The symbolic link is configured once at cluster deployment time.

A **naming convention** is established which is to be taken in account by Lustre file systems configuration and the Cluster Suite configuration:

```
failover service name = lustre_<primary_node_name>
```

## 12.6 Error Detection and Prevention Mechanisms

### Lustre Single Point Of Failure (SPOF) tracking

Tracked Lustre services SPOF include:

1. Service crash (no longer running on the node).
2. Lustre services in an unavailable state (hanging, starting, etc.).
3. Repetitive abnormal comportment (systematic client eviction, etc.).
4. I/O errors on the back-end device.

The Lustre services failures (points 1 to 3) detection relies on the intrinsic Lustre health monitoring system. This internal failures management mechanism maintains diagnose items in the local `/proc/sys/lustre` and `/proc/fs/lustre` directories of each I/O or metadata node.

A regular check of Lustre services sanity is scheduled by the Cluster Suite through the status target of the failover script. This check will process the diagnose items maintained by Lustre.

I/O errors on the back-end device (point 4) detection relies on the storage management monitoring daemon **storfilter**. When it detects a problem, this daemon warns the Lustre failover script using a dedicated target.

When one these SPOF is detected, a Lustre services migration is triggered followed by a node power off.

## 12.7 Analysis of Failure Modes

### 12.7.1 I/O Node and Metadata Failures

#### I/O Node Panic/Hang

When a node hangs or encounters a panic, it does not send its heartbeat messages in the authorized period. This silence is detected by the peer HA node which fences the silent node and takes over the cluster services when the fencing is completed.

#### I/O Node Power down

The node does not send its heartbeat messages in the authorized period. This silence is detected by the peer HA node which fences the silent node and takeover the cluster services when the fencing is completed.

#### Lustre Software Failure

Two scenarios can take place leading to the same action:

- The health monitoring mechanism of the Lustre system detects the failure and update the health information in the `/proc` directory. The next time the status target of the Lustre failover script is activated by the Cluster Suite, it will detect the problem and power off the failing node.
- The status target of the Lustre failover script detects that some Lustre services are missing or not in the correct state. It first tries to restart them. In case of failure it will power off the failing node.

Both scenarios trigger the **I/O node power down** failure treatment.

### 12.7.2 Storage Failures

#### Fibre Channel Adapter Errors

Fibre channel adapters are used to access to external storage systems, shared by the two nodes of the HA I/O cell. They store the OSS and MDS data.

Fibre channel adapter errors are ignored. Typically, link events may be transient on fibre channel links, and should not lead to a node failover. If the adapter error is real, it will lead to a linux disk error, which is monitored.

#### Disk Subsystem Failure

When a disk subsystem, despite its internal redundancy, encounters errors which prevent the processing of node's I/O requests, the node will detect a disk error.



Disk errors are monitored by the I/O status service; they are notified to the Management Node, and to Lustre management locally on the node. Lustre will execute the necessary actions, and then stops the node. The node being silent, the Cluster Suite will take over the service on the peer HA node.



**Note:**

Lustre verifies that the device generating the I/O error is being used by Lustre. If it is not then no corrective action will be taken.

### SCSI Adapter Error or Hang

SCSI adapters are usually used to store systems data, binaries, swap, and temporary files. A SCSI adapter failure leads to a kernel hang or panic, or to a lustre service failure. These two types of failure have already been described.

## 12.7.3 Ethernet Network Failures

A failure of the heartbeat network will stop heartbeat exchanges leading each node to initiate to service take over. A fencing race starts between both nodes.



**Note:**

Only the heartbeat network is monitored by Cluster Suite. A failure on another Ethernet network than the one used for heartbeat will not lead to service takeover; however the failure will be displayed on the management node via NovaScale Master - HPC Edition.

### Ethernet Network Access Failure (NIC or link Failure)

The management network is also used to send fence requests to the appropriate PAP, this is the only way with the FAME architecture to stop a node remotely.

The node which is unable to use the management network cannot fence its peer node.

Thus, the peer node wins the fencing race, and takes over the cluster services.

There is no risk of split brain (i.e. both nodes of the I/O cell running simultaneously the same Lustre service).

### Management Network Failure

If the management network is unavailable for both nodes of the HA I/O cell, none will be able to fence its peer node. The Cluster Suite does not initiate any failover action.

If one the node of the HA pair is able to fence the peer node, it wins the fencing race and takes over the services.

## 12.8 Using Cluster Suite

Large clusters may contain multiple I/O cells, and within each I/O cell, the Cluster Suite must be configured. This process is fully automated by Bull cluster management tools. All the necessary information are extracted from the cluster DB, and read to use Cluster Suite configuration files are pushed to each node which must be controlled by the Cluster Suite.



### Important:

Cluster Suite commands not described in the present paragraph must not be used, as they may lead to fatal inconsistency for the Lustre file system. The GUI must not be used as well. All the Cluster Suite setup is predefined to enable the failover process expected by the Lustre file system, and prevent any risk of split brain (i.e. both nodes of the I/O cell running simultaneously the same Lustre service). Administrator must not attempt to modify the Cluster Suite's configuration within I/O cells.

The management tasks for Cluster Suite are:

- Distributing the configuration file
- Starting the Cluster Suite.

By default, there is not automatic start at boot time, and it is not recommended to enable this.

### 12.8.1 Distributing the cluster.conf file on the I/O Node

The `/etc/cluster/cluster.conf` is generated using node HA pair defined in the cluster DB. The following options are selected, and must not be changed:

- Name of the services to be managed.
- Manual start of services (to avoid split brain if a node can not join its peer node).
- List of nodes.
- For NovaScale R440 and R460:
  - Heartbeat through the Management Network

The Cluster Suite configuration files are automatically generated and deployed on each node by the `stordepha` command.

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

The `-a` flag means all nodes. It is possible to exclude some nodes (`-e` flag) or to specify a list of target nodes (`-i` flag, exclusive with `-a`). See the man page of the command for more information.



### Note:

The `cluster.conf` file is not preserved when a node is reinstalled by KSIS. It must also not be integrated in a node image. After each node's deployment, the `stordepha` command must be used to restore the node's configuration.

## 12.8.2 Starting / Stopping Cluster Suite's Daemons

The Cluster Suite's daemons can be configured from the Management Node on all or a subset of the HA I/O nodes, or locally on each node. In both case, all the required daemons are started and stopped consistently, in the right order.

Starting the Cluster Suite starts the Cluster Suite daemons. But the Cluster Suite services do not start, because the automatic start is disabled.

Stopping the Cluster Suite on a node causes its services to fail on the peer node.

- From the management station:

```
stordepha -c start|stop -a
```

The `-a` flag means all nodes. It is possible to exclude some nodes (`-e` flag) or to specify a list of target nodes (`-i` flag, exclusive with `-a`). See the man page of the command for more information.

- Locally on a node:

```
storioha -c start|stop
```

## 12.8.3 Checking the Cluster Suite Status

The Cluster Suite's status can be verified from the Management Node on all or a subset of the HA I/O nodes, or locally on each node.

- From the management station:

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

The `-a` flag means all nodes. It is possible to exclude some nodes (`-e` flag) or to specify a list of target nodes (`-i` flag, exclusive with `-a`). See the man page of the command for more information.

- Locally on a node:

```
storioha -c status
```

Alternatively, it is also possible to use the Cluster Suite's `clustat` command:

```
clustat
```

or:

```
clustat -i <refresh period>
```

## 12.9 Managing Lustre High Availability

Lustre High-Availability management is included in the Lustre management framework under the form of add-ons and specific tools. It is operated from the management station.

The management tasks for Lustre failover are:

- Setting up the hardware and software configurations information.
- Enabling file systems for failover support.
- Dealing with the nodes migrations and Lustre services take over.

### 12.9.1 ClusterDB Information

Two kinds of information are included into the Lustre tables of the ClusterDB to allow Lustre services failover management and monitoring:

- Static information linked with the cabling schema of the paired node.
- Dynamic information about the actual nodes migrations and Lustre services distribution.

**lustre\_io\_node** table for each node of M(etadata) and I(/O) type gives

- Its paired node identity, pre-loaded at cluster install
- Its current migration status, maintained up to date by the failover management tools.

**lustre\_ost** and **lustre\_mdt** tables for each Lustre service give

- Its primary and secondary nodes loaded by the storage/Lustre deployment process
- Its currently supporting node (active) dynamically set by the failover management tools.

This information can be accessed and if necessary very carefully updated using the standard **lustre\_tables\_dba** tools.



**Note:**

The **mds\_ha\_node** and **oss\_ha\_node** are initialized by the contents of the **lustre\_io\_node** tables.

### 12.9.2 LDAP Directory – the **lustre\_ldap** Utility

The Lustre LDAP directory contains the description of each Lustre file system currently installed on the cluster I/O nodes with its current services distribution on the cluster I/O nodes. It is located on the management station.

When a file system is started, it is noted as active in the LDAP directory enabling the management of the takeover of its services by the Lustre failover scripts of the I/O nodes. This is done automatically by the **lustre\_util** utility.

The Lustre failover scripts check it each time a High-Availability event occurs on the nodes to get the Lustre services list they are supposed to act on. They update it with the result of the migration operations. A synchronization mechanism ensures the transfer of this information to the **ClusterDB**.

Configure and start the Lustre LDAP directory.

1. Create the `/var/lib/ldap/lustre` directory:

```
mkdir -p /var/lib/ldap/lustre
chown ldap.ldap /var/lib/ldap/lustre
```

2. Enable and start the LDAP service:

```
chkconfig --level 345 ldap
service ldap start
```

3. In the `/etc/sudoers` configuration file, verify that the `ldap` user has access to the `lustre_tables_dba` commands:

```
Cmd_Alias LUSTRE_DB=/usr/sbin/lustre_ost_dba *, /usr/sbin/lustre_mdt_dba *,
/usr/sbin/lustre_io_node_dba *
ldap ALL = NOPASSWD: LUSTRE_DB
```

If not, use the **visudo** tool to update the `/etc/sudoers` file.

The management of the LDAP directory is performed using the **lustre\_ldap** utility:

- Callback for **lustre\_util** and the LDAP server for ClusterDB synchronization
- Online interface for the administrator to display the LDAP directory contents.

```
lustre_ldap show [-f <file_system_name>]
```

Display the current status of the file system as seen by the High-Availability system:

- **active** for a started file system,
- **unactive** for a stopped file system.

Without any parameters, the command will show the status of all the file systems loaded in the LDAP directory.

```
lustre_ldap list [-f <file_system_name>]
```

List the LDAP descriptor of the file system in LDIF format. Without any parameters, the command will list all the file systems names loaded in the LDAP directory. Regarding the LDIF format of the display, it is mainly useful for maintenance process.



**Important:**

**lustre\_ldap** can also be used to update the LDAP directory contents for punctual corrections. This has to be done very carefully provided the failover information consistency could be broken.

### 12.9.3 Failover Tools Configuration – the /etc/lustre/lustre.cfg File

Edit the configuration file of the management tools `/etc/lustre/lustre.cfg`, to include the modifications below.

To use failover file systems, set `LUSTRE_LDAP_URL` according to the name of the Management Node (`ldap://<mgmt node>/`).

To enable the failover tools trace feature, set `LUSTRE_DEBUG` to “yes”.

Verify that `LUSTRE_NET` is set according to the cluster type. By default this will be set to `tcp` and it may be necessary to change it to `elan` or `O2ib`.

On each I/O node, the Lustre failover scripts will log events in the `/var/log/lustre` directory. On the management station, the `lustre_ldap` daemon will log events in the `/tmp/log/lustre` directory.

### 12.9.4 Managing Lustre Failover Services on I/O and Metadata Nodes – the `lustre_migrate` Tool

Lustre failover services are used by the Cluster Suite to control the Lustre OST/MDT services migration.



**Warning:**

**The failover services have to be started before the Lustre file systems are started. They can be stopped only when all Lustre file systems are stopped**

The `lustre_migrate` command allows the failover Lustre services on the cluster to be managed.

Without any parameters, the command acts on all the I/O and metadata nodes.

#### Failover Services `start/stop/status`

```
lustre_migrate hastat [-n <node_list>]
```

Display the status of the Lustre failover services on the I/O and metadata nodes of the list. Without any parameters, it acts on all the I/O and metadata nodes.

```
lustre_migrate hastart [-n <node_list>]
```

Start the Lustre failover services on the I/O and metadata nodes of the list. Without any parameters, it acts on all the I/O and metadata nodes.

```
lustre_migrate hastop [-n <node_list>]
```

Stop the Lustre failover services on the I/O and metadata nodes of the list. Without any parameters, it acts on all the I/O and metadata nodes.

## Failover Services Migration Control

For maintenance purposes, it may be useful to migrate Lustre services of a node to its HA paired, so that the node can be stopped without disturbing the Lustre system.

```
lustre_migrate export -n <node_name>
```

Stop the `lustre_<node_name>` failover service for which the node `<node_name>` is primary, and restart it on its secondary node. The secondary node information is taken from the ClusterDB. If the `lustre_<node_name>` failover service was already running on its secondary node, the command has no effect.

To relocate a failover service on its primary node once it is repaired:

```
lustre_migrate relocate -n <node_name>
```

Stop the `lustre_<node_name>` failover service for which the node `<node_name>` is primary, on its secondary node and restart it on its primary node. The secondary node information is taken from the ClusterDB. If the `lustre_<node_name>` failover service was already running on its primary node, the command has no effect.

## 12.9.5 Configuring File Systems for Failover

Configuring file systems for failover means configuring two paired OSS/MDS as possible support for each OST/MDT, one being the primary node, the other being the secondary node.

The failover feature is declared in the `/etc/lustre/models/<file_system_name>.lmf` model file. In the file system model update the following parameters:

- **failover=yes** enables failover configuration generation,
- **timeout=<xx>** sets the recovery time-out value. This time-out is used by Lustre to manage its recovery process. Recommended value = 60.

A file system is then described as usual, composed of one MDT and several OSTs taken from the ClusterDB.

The secondary node declared in the ClusterDB will be taken in account for a second OST/MDT declaration.

The file system is then managed using the `lustre_util` utility in a standard way. The failover specificity (LDAP directory interaction, status display, alternative mount , etc.) being automatically supported by the Lustre management tools.

## 12.10 Lustre High Availability Operations

### 12.10.1 Service Migration triggered by Cluster Suite

This process is conducted on node failure detection by the High-Availability system.

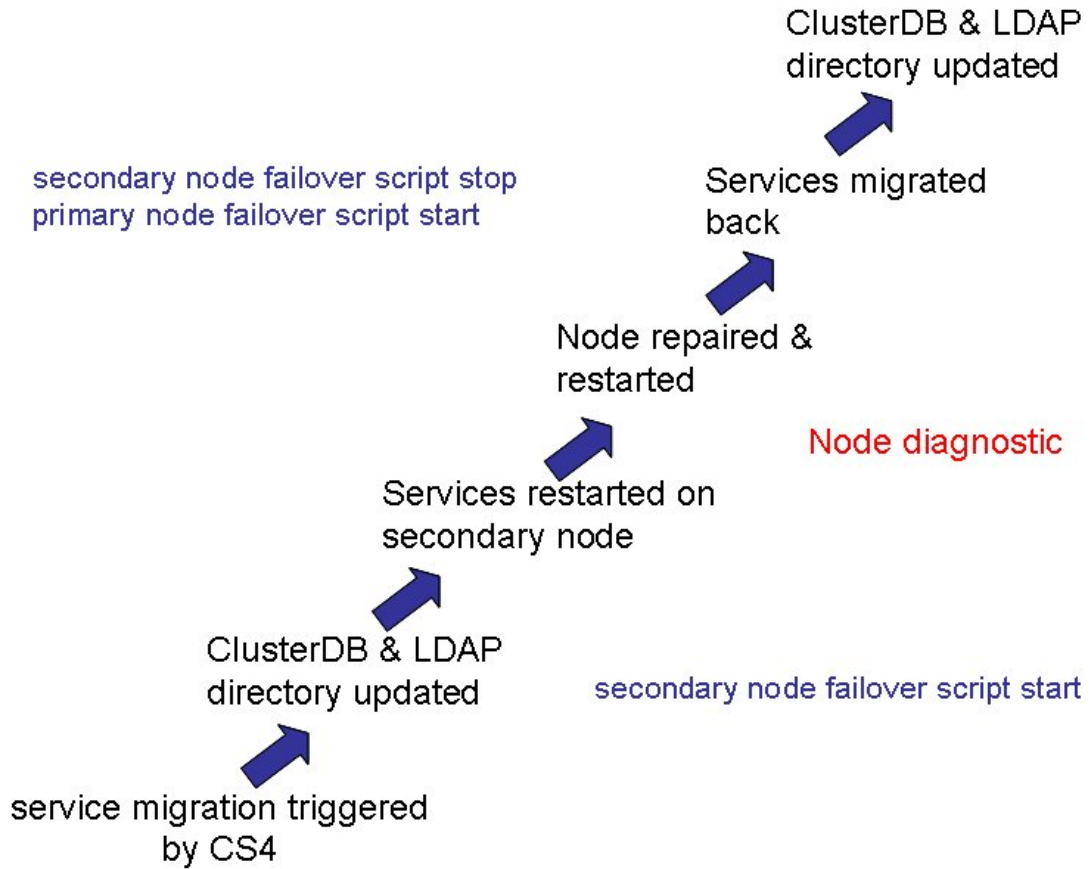


Figure 12-8 Service migration triggered by Cluster Suite



## 12.10.2 Service Migration triggered by Administrator

This process is conducted when the administrator needs to insulate a node without stopping the Lustre system.

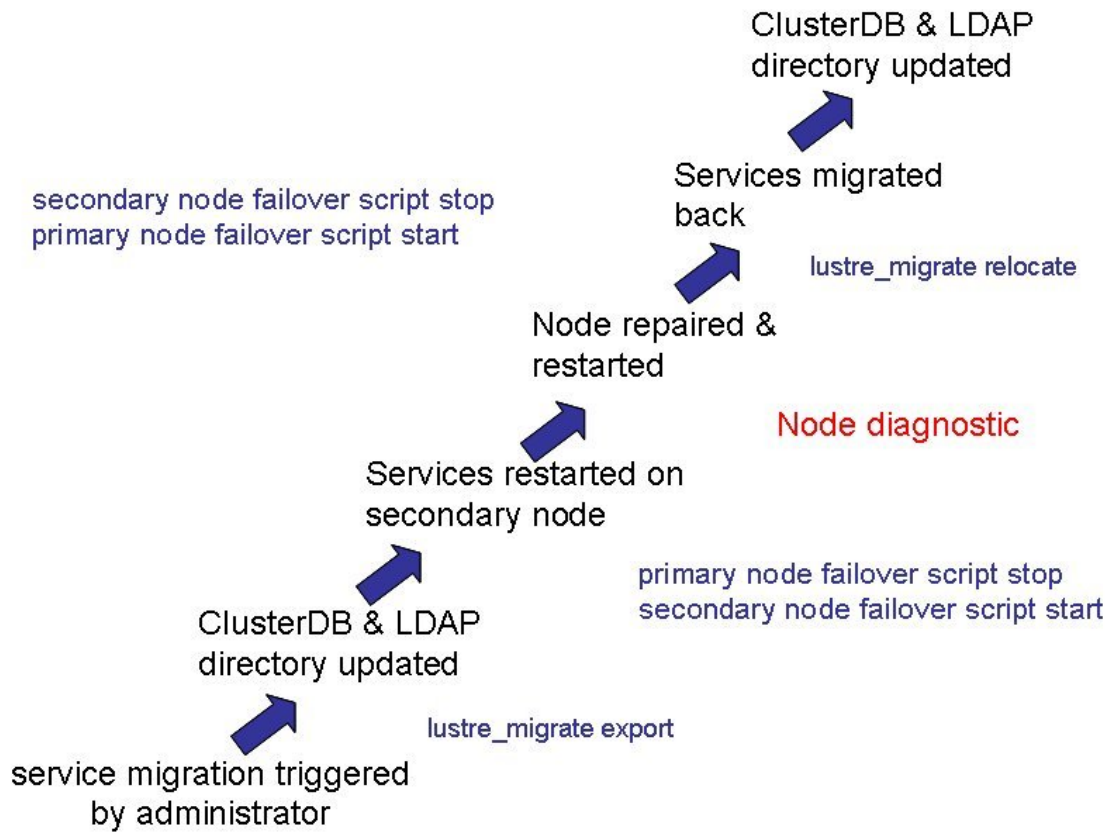


Figure 12-9 Service migration triggered by the Administrator

## 12.11 Monitoring Lustre High Availability

Two approaches are available from the monitoring tools: nodes migrations and the resulting OST/MDT file systems actual distribution.

On line commands allow the administrator to get an instant status of the Lustre High-Availability system.

If the cluster has a management node, important global health indicators are available via **NovaScale Master - HPC Edition** main view. They constitute a warning system for the administrator.

A trace system can be activated for debug and problem resolution purpose.

### 12.11.1 Command Line Monitoring

The following command displays the current failover paired nodes status under the form of an array with one line for each pair of nodes, as follows:

```
lustre_migrate nodestat
```

node name	node status	node HA name	node HA status
ns6	OK	ns7	MIGRATED

For each node, the status is that of the Lustre failover service it is primary for:

- KO** the Lustre failover service is UP
- WARNING** the Lustre failover service is UP some Lustre services are missing. A node migration may be in progress
- MIGRATED** the Lustre failover service has successfully migrated to the paired node and is now running on it
- CRITICAL** the Lustre failover service is no longer operating. The node migration has failed

The following command displays the current Lustre failover services distribution and status as seen by the Cluster Suite.

```
lustre_migrate hastat
```

For each Lustre file system installed on the cluster, the following command displays the detailed distribution of the MDTs and OSTs.

```
lustre_util info -f <File system name>
```

## 12.11.2 Graphic Monitoring

The Graphic Monitoring feature is available only if the cluster has a management node.

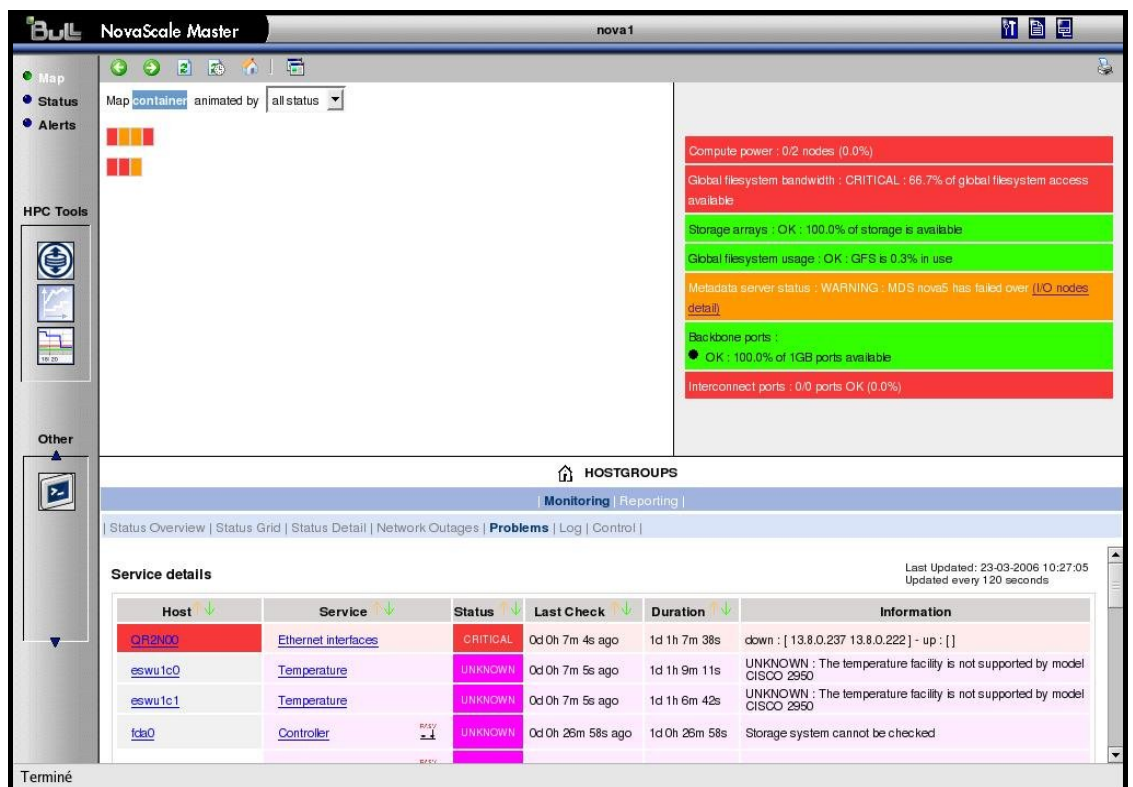


Figure 12-10 NovaScale Master Map all status screen

The I/O pairs status alert indicates if a migration of the metadata server has occurred. In this case, the Lustre system is no longer Highly -Available and an intervention is highly urgent.

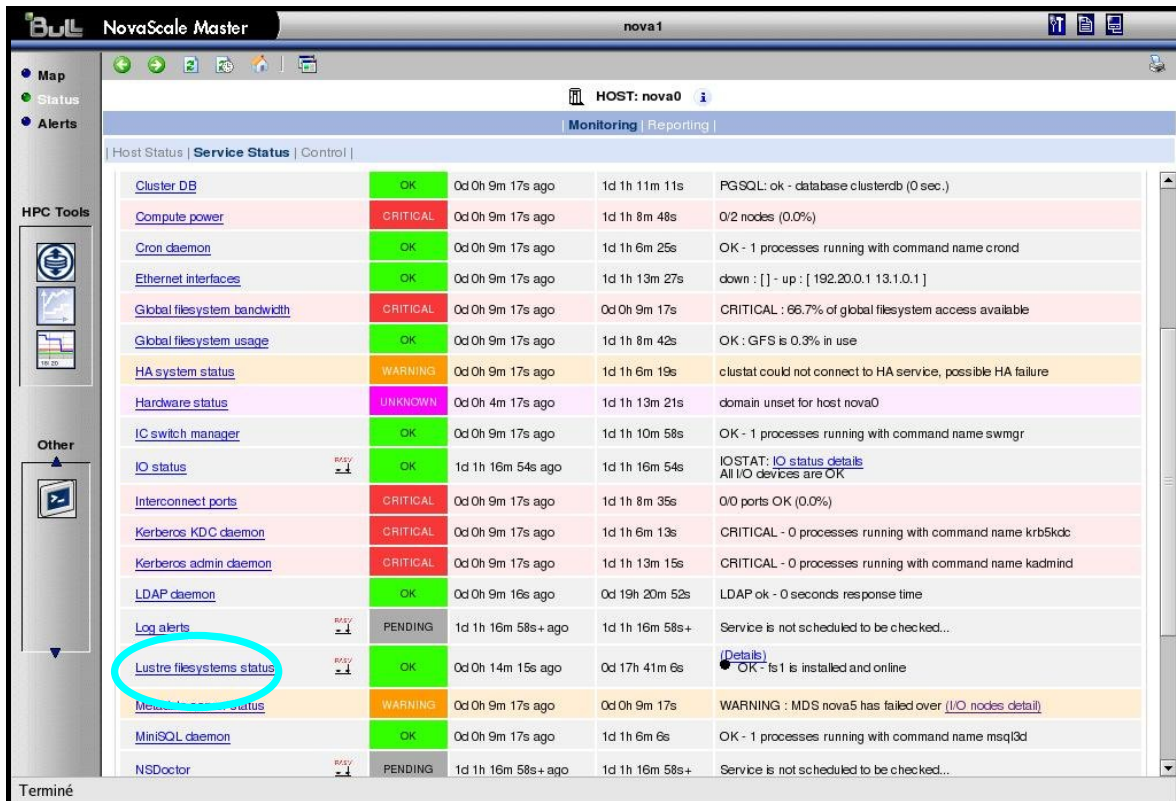


Figure 12-11 Lustre filesystem status indicator in the Host service status window

The Lustre file system indicator warns about failures. Clicking on the info link will display MDTs/OSTs detailed status.

## 12.11.3 Traces and Debug

### Failover Tools Traces

These are enabled by setting the LUSTRE\_DEBUG parameter of the `/etc/lustre/lustre.cfg` file to yes.

On the management station, a daily log file, for example `//tmp/log/lustre/LDAP-<dd mm>.log`, is recorded under the `/tmp/log/lustre` directory by the `lustre_ldap` daemon. It gives information about migration events transmitted to the LDAP directory.

On the I/O and metadata nodes, a daily log file is recorded under the `/var/log/lustre` directory by the `lustre_failover` scripts. It gives information about failover events and their management.

### System Log Files

On each I/O and metadata node, the Cluster Suite and the failover scripts log events in the `/var/log/messages` and the `/var/log/syslog` files. These files are centralized on the management station by the `syslog-ng` system.

# Appendix A. BIOS Parameter Settings to use for NovaScale R421 and R422 Compute Nodes

The BIOS parameter settings for the NovaScale R421 and R422 Compute Nodes will normally be configured in the factory before the machines are delivered. However, if the cluster set up is changed, the following settings can be used to reset the machines back to their original state.



**Note:**

Some of these settings, for example for the storage, will vary according to the cluster and will differ from the settings shown in the tables and screen grabs.

## A.1 NovaScale R421 BIOS Settings

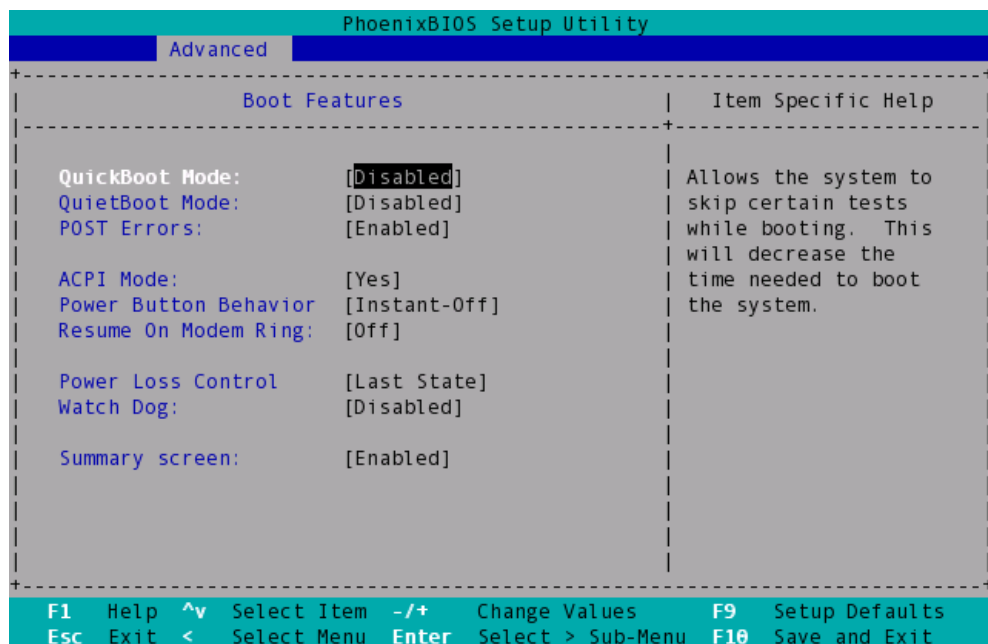


Figure A-1. Example BIOS parameter setting screen for NovaScale R421

## A.1.1 Example BIOS Parameter Settings for NovaScale R421



**Note:**

The settings shown are for a cluster which uses SATA devices. The parameters may vary according to cluster.

BIOS setup section		parameter	value
Main		System Time	<Current local time>
		System Date	<Current date>
		Serial ATA	Enabled
		Native Mode Operation	Serial ATA
		SATA Controller Mode Option	Compatible
Advanced	Boot Features	QuickBoot Mode	Disabled
		QuietBoot Mode	Disabled
		POST Errors	Enabled
		ACPI Mode	Yes
		Power Button Behaviour	Instant-Off
		Resume On Modem Ring	Off
		Power Loss Control	Last State
		Watch Dog	Disabled
		Summary screen	Enabled
	Memory Cache	Cache System BIOS area	Write Protect
		Cache Video BIOS area	Write Protect
		Cache Base 0-512k	Write Back
		Cache Base 512k-640k	Write Back
		Cache Extended Memory Area	Write Back
		Discrete MTRR Allocation	Disabled
	PCI Configuration	Onboard G-LAN1 OPROM Configure	Enabled
		Onboard G-LAN2 OPROM Configure	Disabled
		Default Primary Video Adapter	Onboard
		Emulated IRQ Solution	Disabled
		PCI-e I/O Performance	Payload 256B
		PCI Parity Error Forwarding	Disabled
		ROM Scan Ordering	Onboard First
		PCI Fast Delayed Transaction	Disabled
		Reset Configuration Data	No
	Frequency for PCIX#1-#2/MASS	Auto	

BIOS setup section		parameter		value
		SLOT1 PCI-X 100MHz	Option ROM Scan	<i>Enabled</i>
			Enable Master	<i>Enabled</i>
			Latency Timer	<i>Default</i>
		SLOT2 PCI-X 100MHz ZCR	Option ROM Scan	<i>Enabled</i>
			Enable Master	<i>Enabled</i>
			Latency Timer	<i>Default</i>
		SLOT2 PCI-X 100MHz ZCR	Option ROM Scan	<i>Enabled</i>
			Enable Master	<i>Enabled</i>
			Latency Timer	<i>Default</i>
		SLOT3 PCI-Exp x8	Option ROM Scan	<i>Enabled</i>
			Enable Master	<i>Enabled</i>
			Latency Timer	<i>Default</i>
		SLOT4 PCI-Exp x8	Option ROM Scan	<i>Enabled</i>
			Enable Master	<i>Enabled</i>
			Latency Timer	<i>Default</i>
	SLOT5 PCI-Exp x8	Option ROM Scan	<i>Enabled</i>	
		Enable Master	<i>Enabled</i>	
		Latency Timer	<i>Default</i>	
	Advanced Chipset Control	Large Disk Access Mode		<i>DOS</i>
		SERR signal condition		<i>Single bit</i>
		4GB PCI Hole Granularity		<i>256 MB</i>
Memory Branch Mode		<i>Interleave</i>		
Branch 0 Rank Interleave		<i>« 4:1 »</i>		
Branch 0 Rank Sparing		<i>Disabled</i>		
Branch 1 Rank Interleave		<i>« 4:1 »</i>		

BIOS setup section		parameter	value
		Branch 1 Rank Sparing	<i>Disabled</i>
		Enhanced x8 Detection	<i>Enabled</i>
		High Bandwidth FSB	<i>Enabled</i>
		High Temp DRAM OP	<i>Disabled</i>
		AMB Thermal Sensor	<i>Disabled</i>
		Thermal Throttle	<i>Disabled</i>
		Global Activation Throttle	<i>Disabled</i>
		Crystal Beach Feature	<i>Enabled</i>
		Route Port 80h cycles to	<i>LPC</i>
		Clock Spectrum Feature	<i>Disabled</i>
		High Precision Event Timer	<i>No</i>
		USB Function	<i>Enabled</i>
		Legacy USB Support:	<i>Enabled</i>
		Advanced Processor Options	Frequency Ratio
	Core Multi-Processing		<i>Enabled</i>
	Machine Checking		<i>Enabled</i>
	Thermal Management 2		<i>Enabled</i>
	C1 Enhanced Mode		<i>Disabled</i>
	Execute Disable Bit		<i>Enabled</i>
	Adjacent Cache Line Prefetch		<i>Enabled</i>
	Hardware Prefetcher		<i>Enabled</i>
	Direct Cache Access		<i>Disabled</i>
	Intel(R) Virtualization Technology		<i>Disabled</i>
	Intel EIST support	<i>Disabled]</i>	
	I/O Device Configuration	KBC Clock Input	<i>12MHz</i>
		Serial port A	<i>Enabled]</i>
		Base I/O address (Serial port A)	<i>3F8</i>
		Interrupt (Serial port A)	<i>IRQ 4</i>
		Serial port B	<i>Enabled</i>
		Mode	<i>Normal</i>
		Base I/O address (Serial port B)	<i>2F8</i>
		Interrupt (Serial port B)	<i>IRQ 3</i>
		Floppy disk controller	<i>Enabled</i>
	Base I/O address	<i>Primary</i>	
	DMI Event Logging	Event Logging	<i>Enabled</i>
		ECC Event Logging	<i>Enabled</i>
	Console Redirection	Com Port Address	<i>On-board COM B</i>
		Baud Rate	<i>115.2K</i>
		Console Type	<i>VT100+</i>



BIOS setup section		parameter	value
		Flow Control	None
		Console connection	Direct
		Continue C.R. after POST	Off
	Hardware Monitor	CPU Temperature Threshold	75oC
		Fan Speed Control Modes	1)Disable(Full spe
	IPMI	System Event Logging	Enabled
		Clear System Event Log	Disabled
		SYS Firmware Progress	Disabled
		BIOS POST Errors	Enabled
		BIOS POST Watchdog	Disabled
		OS boot Watchdog	Disabled
		Timer for loading OS (min)	10
		Time out action	No Action
Security		Supervisor Password Is	Clear
		User Password Is	Clear
		Password on boot	Disabled
Boot		1	USB FDC
		2	USB CDROM
		3	USB KEY
		4	PCI BEV: IBA GE Slot 0400 v1236
		5	IDE 4: WDC WD1600YS- 01SHB1-(S2)
		6	
		7	
		8	

## A.2 NovaScale R422 BIOS Settings

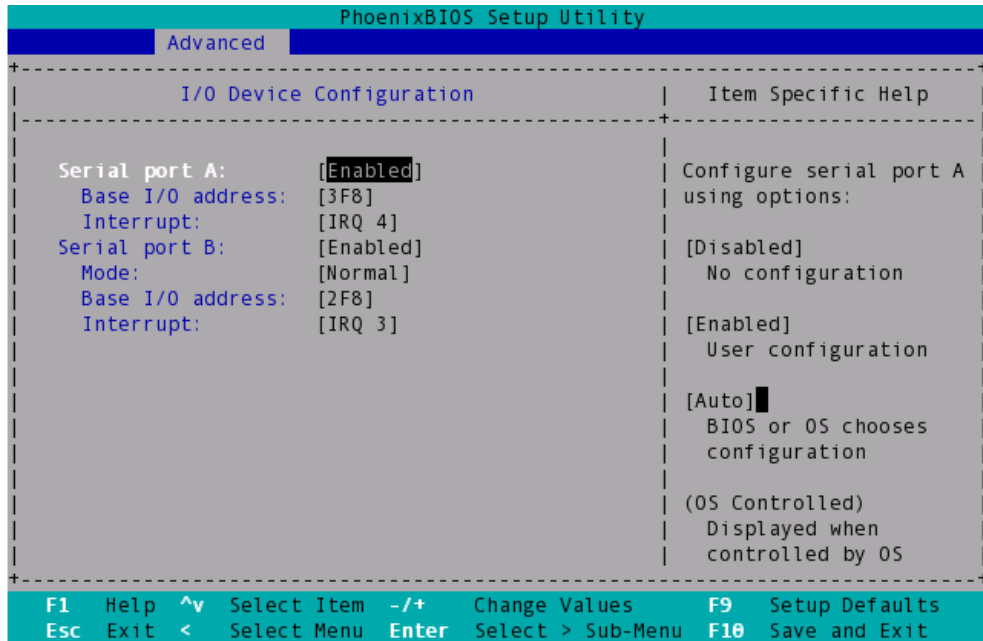


Figure A-2. Example BIOS parameter setting screen for NovaScale R422

### A.2.1 Example BIOS Parameter Settings for NovaScale R422



**Note:**

The settings shown are for a cluster which uses SATA devices. The parameters may vary according to cluster.

BIOS setup section		parameter	value
Main		System Time	<Current local time>
		System Date	<Current date>
		Serial ATA	Enabled
		Native Mode Operation	Serial ATA
		SATA Controller Mode Option	Compatible
Advanced	Boot Features	QuickBoot Mode	Disabled
		QuietBoot Mode	Disabled
		POST Errors	Enabled
		ACPI Mode	Yes
		Power Button Behaviour	Instant-Off
		Resume On Modem Ring	Off
		Power Loss Control	Last State
	Watch Dog	Disabled	

BIOS setup section	parameter	value	
	Summary screen	Enabled	
Memory Cache	Cache System BIOS area	Write Protect	
	Cache Video BIOS area	Write Protect	
	Cache Base 0-512k	Write Back	
	Cache Base 512k-640k	Write Back	
	Cache Extended Memory Area	Write Back	
	Discrete MTRR Allocation	Disabled	
PCI Configuration	Onboard G-LAN1 OPROM Configure	Enabled	
	Onboard G-LAN2 OPROM Configure	Disabled	
	Default Primary Video Adapter	Onboard	
	Emulated IRQ Solution	Disabled	
	PCI-e I/O Performance	Payload 256B	
	PCI Parity Error Forwarding	Disabled	
	ROM Scan Ordering	Onboard First	
	Reset Configuration Data	No	
	SLOT1 PCI-Exp x8	Option ROM Scan	Enabled
		Enable Master	Enabled
Latency Timer		Default	
Large Disk Access Mode	DOS		
Advanced Chipset Control	SERR signal condition	Single bit	
	4GB PCI Hole Granularity	256 MB	
	Memory Branch Mode	Interleave	
	Branch 0 Rank Interleave	« 4:1 »	
	Branch 0 Rank Sparing	Disabled	
	Branch 1 Rank Interleave	« 4:1 »	
	Branch 1 Rank Sparing	Disabled	
	Enhanced x8 Detection	Enabled	
	High Bandwidth FSB	Enabled	
	High Temp DRAM OP	Disabled	
	AMB Thermal Sensor	Disabled	
	Thermal Throttle	Disabled	
	Global Activation Throttle	Disabled	
	Crystal Beach Feature	Enabled	
	Route Port 80h cycles to	LPC	
	Clock Spectrum Feature	Disabled	
	High Precision Event Timer	No	
	USB Function	Enabled	
Legacy USB Support:	Enabled		
Advanced Processor	Frequency Ratio	Default]	

BIOS setup section	parameter	value	
Options	Core Multi-Processing	Enabled	
	Machine Checking	Enabled	
	Thermal Management 2	Enabled	
	C1 Enhanced Mode	Disabled	
	Execute Disable Bit	Enabled	
	Adjacent Cache Line Prefetch	Enabled	
	Hardware Prefetcher	Enabled	
	Direct Cache Access	Disabled	
	Intel(R) Virtualization Technology	Disabled	
	Intel EIST support	Disabled	
	I/O Device Configuration	Serial port A	Enabled
		Base I/O address (Serial port A)	3F8
		Interrupt (Serial port A)	IRQ 4
		Serial port B	Enabled
		Mode	Normal
		Base I/O address (Serial port B)	2F8
		Interrupt (Serial port B)	IRQ 3
	DMI Event Logging	Event Logging	Enabled
		ECC Event Logging	Enabled
	Console Redirection	Com Port Address	On-board COM B
		Baud Rate	115.2K
		Console Type	VT100+
		Flow Control	None
		Console connection	Direct
		Continue C.R. after POST	Off
	Hardware Monitor	CPU Temperature Threshold	75oC
		Fan Speed Control Modes	2)3-pin(Server)
	IPMI	System Event Logging	Enabled
		Clear System Event Log	Disabled
		SYS Firmware Progress	Disabled
		BIOS POST Errors	Enabled
		BIOS POST Watchdog	Disabled
		OS boot Watchdog	Disabled
Timer for loading OS (min)		10	
Time out action		No Action	
Security	Supervisor Password Is	Clear	
	User Password Is	Clear	
	Password on boot	Disabled	
Boot	1	USB FDC	

BIOS setup section	parameter	value
	2	USB CDROM
	3	USB KEY
	4	PCI BEV: IBA GE Slot 0400 v1236
	5	IDE 4: WDC WD1600YS- 01SHB1-(S2)
	6	
	7	
	8	



---

# Glossary and Acronyms

---

## A

### ACL

Access Control List.

---

## B

### Bisectional Bandwidth

The bandwidth flowing through a fabric while half the nodes send and receive a full duplex stream of data to the other half of the nodes.

---

## C

### CGI

Common Gateway Interface.

### ConMan

A management tool, based on telnet, enabling access to all the consoles of the cluster.

### Cron

A UNIX command for scheduling jobs to be executed sometime in the future. A cron is normally used to schedule a job that is executed periodically - for example, to send out a notice every morning. It is also a daemon process, meaning that it runs continuously, waiting for specific events to occur.

### Cygwin

A Linux-like environment for Windows. The Bull cluster management tools use Cygwin to provide ssh support on a Windows system, enabling access in command mode from the Cluster management system.

---

## D

### DNS

Domain Name Server. A server that retains the addresses and routing information for TCP/IP LAN users.

---

## G

### Ganglia

A distributed monitoring tool used to view information associated with a node, such as CPU load, memory consumption, network load.

### GID

Group ID.

### GPT

GUID Partition Table.

---

## H

### HBA

Host Bus Adapter.

### Hyper-Threading

Hyper-Threading technology is an innovative design from Intel that enables multi-threaded software applications to process threads in parallel within each processor resulting in increased utilization of processor execution resources. To make it short, it is to place two logical processors into a single CPU die.

### HPC

High Performance Computing.

---

## K

### KSIS

Utility for image building and development.

---

## L

### LDAP

Lightweight Directory Access Protocol.

---

## LKCD

Linux Kernel Crash Dump. A tool capturing and analyzing crash dumps.

## LOV

Logical Object Volume.

## LVM

Logical Volume Manager.

---

## M

### MIB

Management Information Base.

### MDS

MetaData Server.

### MDT

MetaData Target.

### MkCDrec

Make CD-ROM Recovery. A tool making bootable system images.

### MPI

Message Passing interface.

### MTBF

Mean Time Between Failures.

---

## N

### Nagios

A powerful monitoring tool, used to monitor the services and resources of Bull HPC clusters.

### NFS

Network File System.

### NIC

Network Interface Card.

### NTP

Network Time Protocol.

---

## O

### OpenSSH

Open Source implementation of the SSH protocol.

### OSC

Object Storage Client.

### OSS

Object Storage Server.

### OST

Object Storage Targets.

---

## P

### PAM

Platform Administration & Maintenance.

### PDSH

A parallel distributed shell.

---

## R

### RMS

Resource Management System. Manages the cluster resources.

---

## S

### SAN

Storage Area Network.

### SIS

System Installation Suite.

### SLURM

Simple Linux Utility for Resource Management.

### SSH

Secure Shell. A protocol for creating a secure connection between two systems.



## **Syslog-ng**

Syslog New Generation, a powerful system log manager.

---

## **T**

### **TGT**

Ticket-Granting Ticket.

---

## **U**

### **UID**

User ID

---

## **V**

### **VNC**

Virtual Network Computing. It is used to enable access to Windows systems and Windows applications from the Bull NovaScale cluster management system.

---

## **W**

### **WWPN**

World Wide Port Name- a unique identifier in a Fibre Channel SAN.

---

## **X**

### **XFS**

eXtended File System.

### **XHPC**

Xeon High Performance Computing

### **XIB**

Xeon InfiniBand



---

# Index

## /

- /etc/krb5.conf, 10-2
- /etc/lustre/storage.conf file, 4-8
- /etc/nagios/contactgroups.cfg, 8-16
- /etc/nagios/contacts.cfg, 8-16
- /etc/nagios/snmptargets.cfg, 8-16
- /etc/nsmhpc/nsmhpc.conf, 8-16
- /etc/storageadmin/nec\_admin.conf, 9-33
- /etc/storageadmin/storframework.conf, 9-33
- /var/kerberos/krb5kdc/kadm5.acl, 10-4
- /var/kerberos/krb5kdc/kdc.conf, 10-2
- /var/log/postgres/pgsql, 3-26
- /var/log/synchro.log file, 3-5

## A

- administrator
  - postgres (ClusterDB), 3-2
  - root, 2-2
- authorized\_keys2 file, 2-5

## B

- backbone network, 1-5
- Backbone ports available alert, 8-26
- batch management, 1-11
- Batch Management, 7-1
- bloop tool, 11-2

## C

- chkconfig command, 2-1
- cluster
  - definition, 1-1
- Cluster Suite, 12-12
- ClusterDB
  - administrator (postgres), 3-2
  - ChangeOwnerProperties, 3-2
  - cluster features, 3-7

- Commands, 3-2
  - dbmCluster command, 3-7
  - dbmConfig, 3-5
  - dbmDiskArray command, 3-22
  - dbmEthernet command, 3-14
  - dbmFiberChannel command, 3-20
  - dbmGroup command, 3-12
  - dbmHwManager command, 3-11
  - dbmlconnect command, 3-16
  - dbmNode command, 3-8
  - dbmSerial command, 3-18
  - dbmServices command, 3-21
  - dbmTalim command, 3-17
- Description, 3-1
- managing groups, 3-12
- monitoring, 8-25
- PostgreSQL tools, 3-24
- requisite, 8-2
- save and restore, 3-24
- template files, 3-7

## ClusterDB tables

- Admin table, 3-50
- AVAILABILITY table, 3-54
- CLUSTER table, 3-29
- Config\_Candidate table, 3-51
- Config\_Status table, 3-51
- da\_cfg\_model table, 3-41
- da\_controller table, 3-38
- da\_enclosure table, 3-37
- da\_ethernet\_port table, 3-39
- da\_fan table, 3-40
- da\_fc\_port table, 3-38
- da\_io\_path table, 3-41
- da\_iocell\_component table, 3-41
- da\_power\_fan table, 3-40
- da\_power\_port table, 3-42
- da\_power\_supply table, 3-40
- da\_serial\_port table, 3-39
- da\_temperature\_sensor table, 3-41
- disk\_array table, 3-37
- disk\_slot table, 3-38
- ETH\_EXTRALINK table, 3-34
- ETH\_SWITCH table, 3-30
- ETH\_VLAN table, 3-32
- FC\_NW table, 3-33
- FC\_SWITCH table, 3-34
- Group\_Node table, 3-51
- HwManager table, 3-49
- IC\_BOARD table, 3-46

- IC\_NW table, 3-30
- IC\_SWITCH table, 3-31
- IP\_NW table, 3-29
- IPOIB table, 3-47
- Lustre\_fs table, 3-56
- Lustre\_IO\_node table, 3-58
- Lustre\_MDT table, 3-57
- Lustre\_mount table, 3-58
- Lustre\_OST table, 3-57
- MSG\_SYSLOG table, 3-52
- Node table, 3-45
- Node\_image table, 3-45
- Node\_profile table, 3-46
- PORTSERVER table, 3-32
- Rack table, 3-51
- SDPOIB table, 3-47
- SERIAL\_NW table, 3-31
- SERVICES table, 3-53
- TALIM table, 3-34
- Test\_Dependencies table, 3-53
- Test\_Groups table, 3-52
- Test\_Results table, 3-53
- Tests table, 3-52

#### Commands

- ChangeOwnerProperties, 3-2
- chkconfig, 2-1
- dbmCluster, 3-7
- dbmConfig, 3-5
- dbmDiskArray, 3-22
- dbmEthernet, 3-14
- dbmFiberChannel, 3-20
- dbmGroup, 3-12
- dbmHwManager, 3-11
- dbmIconnect, 3-16
- dbmNode, 3-8
- dbmSerial, 3-18
- dbmServices, 3-21
- dbmTalim, 3-17
- ddn\_set\_up\_date\_time, 9-24
- ddn\_admin, 9-23
- ddn\_check, 9-24
- ddn\_conchk, 9-24
- ddn\_firmup, 9-25
- ddn\_init, 9-24
- ddn\_stat, 9-23
- dshbak, 2-6
- iorefmgmt, 9-5
- kadmin, 10-3
- lfs quotacheck, 4-37
- lfs setquota, 4-37

- lsiodev, 9-4
- lustre\_investigate, 4-16
- lustre\_tables\_dba, 4-6
- lustre\_util, 4-21
- mkfs, 2-3
- mkpartfs, 2-3, 2-4
- mkswap, 2-4
- mount, 2-3
- nec\_admin, 9-22
- parted, 2-3, 2-4
- passwd, 2-2
- pbsnodes, 7-2
- pdcp, 2-6
- pdsh, 2-6
- qdel, 7-3
- qstat, 7-3
- qsub, 7-2
- resize, 2-3
- rm, 2-3
- stormodelctl, 9-31
- storstat, 9-2, 9-17
- swapon, 2-4
- Tracejob, 7-3
- useradd, 2-2

- Conman, 1-6

- connectivity status, 9-11

- contact groups
  - adding, 8-16

- contacts
  - adding, 8-16

- controller status, 9-10

- counters
  - display, 11-1
  - papi\_avail -d command, 11-2
  - PAPI\_FP\_OPS, 11-2
  - PAPI\_TOT\_CYC, 11-2

## D

- dbmCluster command, 3-7

- dbmConfig command, 3-5

- dbmDiskArray command, 3-22

- dbmEthernet command, 3-14

- dbmFiberChannel command, 3-20

- dbmGroup command, 3-12

- dbmHwManager command, 3-11

- dbmIconnect command, 3-16
- dbmNode command, 3-8
- dbmSerial command, 3-18
- dbmServices command, 3-21
- dbmTalim command, 3-17
- DDN commands, 9-23
- ddn\_set\_up\_date\_time command, 9-24
- ddn\_admin command, 9-23
- ddn\_check command, 9-24
- ddn\_conchk command, 9-24
- ddn\_firmup command, 9-25
- ddn\_init command, 9-24
- ddn\_stat command, 9-23
- deploying software See Ksis
- distributed shell, 2-6
- distribution
  - changing, 5-1
  - updating, 5-1
- distribution software, 5-1
- dropdb command, 3-25
- dshbak command, 2-6

## E

- Ethernet network, 1-6

## F

- fan status, 9-9
- file system
  - parallel, 4-1
  - striping, 4-1
- files
  - /etc/lustre/storage.conf, 4-8
  - /etc/nagios/contactgroups.cfg, 8-16
  - /etc/nagios/contacts.cfg, 8-16
  - /etc/nagios/snmptargets.cfg, 8-16
  - /etc/nsmhpc/nsmhpc.conf, 8-16
  - /var/log/synchro.log, 3-5
  - authorized\_keys2, 2-5
  - fstab, 2-3
  - genders, 2-7

- id\_dsa.pub, 2-5
- kadm5.acl, 10-4
- lustre.cfg, 4-13
- lustre\_util.conf, 4-32
- nec\_admin.conf, 9-33
- res\_rpm\_qsnetmpi, 2-10
- storframework.conf, 9-33
- template.model, 9-29
- tuning.conf, 4-34

- fsck, 1-7

- fstab file, 2-3

## G

- Ganglia

- data categories, 8-19

- Ganglia

- NovaScale Master HPC Edition, 8-1

- genders file, 2-7

- GPT format (disk), 2-3, 2-4

- groups of nodes, 3-13

## H

- HDD status, 9-9

- High Availability

- Failure mode analysis, 12-10
- I/O nodes hardware architecture, 12-4
- LDAP directory, 12-7
- Lustre Cfg file, 12-16
- Lustre debug, 12-22
- Lustre failover, 12-17
- Lustre File Systems, 12-14
- Lustre File Systems, 12-1
- Lustre LDAP directory, 12-14
- Lustre management, 12-18
- Lustre SPOF, 12-9
- Monitoring Lustre, 12-20

- hpcprof tool, 11-3

- hpcquick tool, 11-4

- hpcrun tool, 11-3

- HPCToolkit, 11-1

- hpcview tool, 11-5

- hpcviewer tool, 11-6

## I

id\_dsa.pub file, 2-5

image  
list, 3-8

InfiniBand links available, 8-27

Infiniband Networks, 1-7

iorefmgmt command, 9-5

## J

JobCredentialPrivateKey, 6-36

JobCredentialPublicCertificate, 6-36

## K

Kerberos, 2-6, 10-1

Access Control List, 10-4

Admin Daemon, 10-4

configuration files, 10-2

database, 10-3

Host principal, 10-5

kadmin command, 10-3

KDC, 10-1

package, 10-2

SSH, 10-8

TGT ticket, 10-7

Kerberos admin daemon, 8-26

Kerberos KDC daemon, 8-26

Ksis

builddatanode command, 5-17

buildpatch command, 5-16

check command, 5-6, 5-17

check group, 5-6

checkdiff command, 5-8, 5-17

checks database, 5-7

client node, 5-2

command file, 5-7

command options, 5-10

create commands, 5-11

delete command, 5-12

deploy command, 5-4, 5-12

detach command, 5-4, 5-16

export command, 5-17

groupfile, 5-6

help command, 5-10

image server, 5-1, 5-2

import command, 5-17

Ksis server, 5-1

list command, 5-4, 5-13

nodelist command, 5-13

nodeRange, 5-10

overview, 1-11, 5-1

patch, 5-3

patch image, 5-15

patched golden image, 5-16

reference node, 5-2

reference/golden image, 5-1, 5-2

store command, 5-4, 5-15

undeploy command, 5-4, 5-13

working patch image, 5-15

workon command, 5-3, 5-15

workon mechanism, 5-3

## L

LDAP daemon, 8-26

linux user, 2-2

LOV (Logical Object Volume), 4-2

lsiodev command, 9-4

Lustre, 4-2

administrator tasks, 4-3

Creating File systems, 4-17

database, 4-6

Extended model file, 4-19

Installing Lustre file systems, 4-21

lfs quotacheck, 4-37

load\_storage.sh, 4-12

lustre.cfg file, 4-13

lustre\_check tool, 4-40

lustre\_investigate command, 4-16

lustre\_storage\_config.sh, 4-9

lustre\_util, 4-21

lustre\_util.conf file, 4-32

Management Node interface, 4-42

model file, 4-17

Monitoring, 4-39

Nagios filesystem indicator, 4-41

networks, 4-13

NovaScale Group Performance view, 4-43

NovaScale Master monitoring, 4-39

NovaScale Node Performance view, 4-45

planning, 4-4

Quota settings, 4-36

Rescuing a file system, 4-38

Services, 4-15

- Setting limits, 4-37
  - striping, 4-5
  - system limitations, 4-5
  - tuning.conf file, 4-34
- Lustre filesystems access, 8-27
- NovaScale Master, 4-39
- lustre.cfg file, 4-13

## M

- maintenance tools, 2-11
- Maui Scheduler, 6-35
- MDS (MetaData Server), 4-2
- MDT (MetaData Target), 4-2
- Message Passing Interface See MPI
- MetaData Server migration alert, 8-26
- MiniSQL daemon, 8-25
- mkfs command, 2-3
- mkpartfs command, 2-3, 2-4
- mkswap command, 2-4
- model
  - file, 9-29
  - storage system configuration, 9-28
- monitoring the cluster, 8-1
- mount command, 2-3
- MPI, 1-10
- MPI libraries
  - MPIBull2, 1-8
  - MPICH\_Ethernet, 1-8

## N

- Nagios
  - Contact groups, 8-4
  - Hosts, 8-7
  - Services, 8-4, 8-7
- Nagios
  - NovaScale Master HPC Edition, 8-1
- Nagios Management node plug-ins
  - ClusterDB, 8-25
  - Cron Daemon, 8-25
  - MiniSQL Daemon, 8-25

- Nagios plug-ins
  - Backbone ports available, 8-26
  - Ethernet Switch services, 8-28
  - HA system status, 8-26
  - InfiniBand links available, 8-27
  - Kerberos admin daemon, 8-26
  - Kerberos KDC daemon, 8-26
  - LDAP daemon, 8-26
  - Lustre filesystems access, 8-27
  - Metadata servers, 8-26
  - NFS filesystems access, 8-27
- NameSpace, 4-2
- nec\_admin command, 9-22
- nec\_admin.conf file, 9-22
- network
  - administration network, 1-5
  - backbone, 1-5
  - Ethernet network, 1-6
  - switches, 1-6
- NFS filesystems access, 8-27
- node
  - compute node, 1-5
  - login node, 1-3
  - Management Node, 1-3
- node list, 3-8
- NovaScale Master HPC Edition
  - Acknowledgements, 8-12
  - Active checks, 8-11
  - Alert definition, 8-14
  - Alert levels, 8-10
  - Alert types, 8-10
  - Alerts button, 8-10
  - All status map view, 8-6
  - Changing passwords, 8-3
  - Comments, 8-13
  - Ganglia, 8-18
  - Global Performance view, 8-19
  - Group Performance view, 8-18
  - Management node Nagios Services
  - Map button, 8-6
  - Monitoring performance, 8-18
  - Nagios Alert log, 8-24
  - Nagios Ethernet interfaces, 8-24
  - Nagios IO Status, 8-24
  - Nagios logs, 8-14
  - Nagios plug-ins, 8-24
  - Nagios postbootchecker, 8-24

- Nagios Services, 8-22
- Notifications, 8-12
- Passive checks, 8-11
- Ping Map view, 8-8
- Rack view, 8-7
- Scripts, 8-15
- Shell button, 8-18
- SNMP Alerts, 8-16
- Status Button, 8-9
- Storage overview, 8-17
- User password, 8-3

NovaScale Master HPC Edition, 1-11, 8-1

NovaScale R421

- BIOS settings, A-1

NovaScale R422

- BIOS settings, A-6

nsctrl, 1-11

## O

oid2name command, 3-25

OpenSSH, 2-5

openssl, 6-36

OSC (Object Storage Client), 4-2

OSS (Object Storage Server), 4-2

OST (Object Storage Target), 4-2

## P

parallel commands, 2-6

parted command, 2-3, 2-4

partition

- add, delete, modify, 2-3
- swap, 2-4

passwd command, 2-2

password

- user, 2-2

PBS Professional Batch Manager, 7-1

- Commands, 7-2
- Daemons, 7-2
- Ethernet, 7-3
- InfiniBand, 7-3, 7-4
- MPIBull2, 7-3

pdcp command, 2-6

pdsh, 1-11

pdsh command, 2-6

pg\_dump command, 3-24

pg\_restore command, 3-24

phpPgAdmin, 3-24

pipeline (data), 4-4

postbootchecker, 8-24

post-configuration (Ksis), 5-2

postgres user, 3-2

PostgreSQL, 3-24

power supply status, 9-9

predefined groups, 3-13

profiling tools, 11-1

psql command, 3-24

## R

res\_rpm\_qsnetmpi file, 2-10

resize command, 2-3

resource management, 1-11, 6-1

rm command, 2-3

root user, 2-2

rsh, 2-6

## S

security

- Kerberos, 10-1
- policies, 2-5

service

- list, 2-1
- star), 2-1

shell

- distributed, 2-6
- kerberos, 2-6
- pdsh, 2-6
- rsh, 2-6
- ssh, 2-6

SLURM, 6-1

- Configuration Parameters
  - AuthType, 6-8
  - BackupAddr, 6-8



- BackupController, 6-8
- CacheGroups, 6-8
- CheckpointType, 6-9
- ControlAddr, 6-9
- ControlMachine, 6-9
- Epilog, 6-9
- FastSchedule, 6-9
- FirstJobId, 6-9
- HeartbeatInterval, 6-9
- InactiveLimit, 6-10
- JobAcctFrequency, 6-10
- JobAcctLogFile, 6-10
- JobAcctType, 6-10
- JobCompLoc, 6-10
- JobCompType, 6-10
- JobCredentialPrivateKey, 6-10
- JobCredentialPublicCertificate, 6-10
- KillTree, 6-11
- KillWait, 6-11
- MaxJobCount, 6-11
- MinJobAge, 6-11
- MpiDefault, 6-11
- PluginDir, 6-11
- PlugStackConfig, 6-11
- ProctrackType, 6-12
- Prolog, 6-12
- PropagatePrioProcess, 6-12
- PropagateResourceLimits, 6-12
- PropagateResourceLimitsExcept, 6-12
- ReturnToService, 6-12
- SchedulerAuth, 6-13
- SchedulerPort, 6-13
- SchedulerRootFilter, 6-13
- SchedulerType, 6-13
- SelectType, 6-13
- SlurmctlDebug, 6-13
- SlurmctlLogFile, 6-14
- SlurmctlPidFile, 6-14
- SlurmctlPort, 6-14
- SlurmctlTimeout, 6-14
- SlurmdDebug, 6-14
- SlurmdPort, 6-14
- SlurmdSpoolDir, 6-14
- SlurmdTimeout, 6-15
- SlurmLogFile, 6-14
- SlurmPidFile, 6-14
- SlurmUser, 6-13
- SrunEpilog, 6-15
- SrunProlog, 6-15
- StateSaveLocation, 6-15
- SwitchType, 6-15
- TaskEpilog, 6-15
- TaskPlugin, 6-16
- TaskProlog, 6-16
- TmpFS, 6-16
- TreeWidth, 6-16
- UseCPUSETS, 6-16
- UsePAM, 6-17
- WaitTime, 6-17
- Draining a node, 6-37
- Functions, 6-2
- Node Configuration Parameters, 6-17
  - DownNodes, 6-19
  - Feature, 6-19
  - NodeAddr, 6-18
  - NodeHostname, 6-18
  - NodeName, 6-18
  - Procs, 6-19
  - RealMemory, 6-19
  - Reason, 6-19
  - State, 6-19
  - TmpDisk, 6-19
  - Weight, 6-20
- NodeAddr, 6-7
- NodeHostname, 6-7
- NodeName, 6-7
- Partition Configuration Parameters, 6-20
  - AllowGroups, 6-20
  - Default, 6-20
  - Hidden, 6-20
  - MaxNodes, 6-20
  - MaxTime, 6-21
  - MinNodes, 6-21
  - Nodes, 6-21
  - PartitionName, 6-21
  - RootOnly, 6-20
  - Shared, 6-21
  - State, 6-21
- SCANCEL, 6-2
- SchedType configuration parameter, 6-35
- Scheduler Support, 6-35
- SCONTROL, 6-2, 6-23
- Scontrol examples, 6-36
- SelectType configuration parameter, 6-36
- SINFO, 6-2
- slurm.conf, 6-7
- slurm.conf example files, 6-21
- SLURMCTLD Controller daemon, 6-32, 6-33
- SLURMCTLD daemon, 6-2, 6-3
- SLURMD, 6-2, 6-4
- SLURMD Compute node daemon, 6-32, 6-34
- SQUEUE, 6-2

- SRUN, 6-2
- SLURM and openssl, 6-36
- SLURM and Security, 6-36
- SLURM and syslogr, 6-36
- SNMP trap
  - response to alert, 8-16
- software distribution, 5-1
- software update, 5-1
- ssh, 2-6
  - setting up, 2-5
- storage device
  - configuration deployment, 9-3
  - configuration files, 9-33
  - configuration planning, 9-27
  - management services, 9-2
  - managing, 9-1
  - monitoring, using Nagios, 9-7
- stormodelctl command, 9-31
- storstat command, 9-2, 9-17
- swap partition, 2-4
- swapon command, 2-4
- syslog-ng, 1-11

- system image, 3-8
- system logs See syslog-ng
- system status, 9-11

## T

- temperature status, 9-10
- template.model file, 9-29
- TORQUE, 1-11

## U

- user
  - create, 2-2
  - password, 2-2
- useradd command, 2-2

## V

- view
  - inventory of storage systems and components, 9-14
  - storage, 9-12
  - storage tactical overview, 9-12
- Voltaire Switching Devices, 1-7

## Technical publication remarks form

<b>Title:</b>	BAS4 for Xeon Administrator's Guide
---------------	-------------------------------------

<b>Reference:</b>	86 A2 83ET 02
-------------------	---------------

<b>Date:</b>	December 2007
--------------	---------------

### ERRORS IN PUBLICATION

------------------------------------------

### SUGGESTIONS FOR IMPROVEMENT TO PUBLICATION

------------------------------------------

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

NAME: \_\_\_\_\_ DATE: \_\_\_\_\_

COMPANY: \_\_\_\_\_

ADDRESS: \_\_\_\_\_

---

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

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





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

REFERENCE  
86 A2 83ET 02