

bullx cluster suite

Administrator's Guide

Extreme Computing



REFERENCE
86 A2 20FA 03

Extreme Computing

bullx cluster suite

Administrator's Guide

Software

April 2010

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

REFERENCE
86 A2 20FA 03

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 2010

Printed in France

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.

Table of Contents

| | | |
|-------------------|--|------------|
| Chapter 1. | Cluster Management Functions and Corresponding Products | 1-1 |
| Chapter 2. | Initial Configuration Tasks | 2-1 |
| 2.1 | Configuring Services..... | 2-1 |
| 2.2 | Modifying Passwords and Creating Users | 2-1 |
| 2.3 | Configuring Security | 2-2 |
| 2.3.1 | Setting up SSH | 2-2 |
| 2.4 | Running Parallel Commands with pdsh | 2-3 |
| 2.4.1 | Using pdsh..... | 2-3 |
| 2.4.2 | Using pdcp | 2-5 |
| 2.4.3 | Using dshbak | 2-6 |
| 2.5 | Day to Day Maintenance Operations | 2-7 |
| Chapter 3. | Cluster Database Management | 3-1 |
| 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..... | 3-6 |
| 3.3.4 | dbmNode | 3-7 |
| 3.3.5 | dbmHwManager | 3-10 |
| 3.3.6 | dbmGroup | 3-11 |
| 3.3.7 | dbmEthernet | 3-13 |
| 3.3.8 | dbmIconnect..... | 3-14 |
| 3.3.9 | dbmTalim..... | 3-16 |
| 3.3.10 | dbmSerial | 3-17 |
| 3.3.11 | dbmFiberChannel | 3-18 |
| 3.3.12 | dbmServices..... | 3-19 |
| 3.3.13 | dbmDiskArray | 3-20 |
| 3.4 | Managing the ClusterDB | 3-22 |
| 3.4.1 | Saving and Restoring the Database..... | 3-22 |
| 3.4.2 | Starting and Stopping PostgreSQL | 3-23 |

| | | |
|-------------------|---|-------------|
| 3.4.3 | Viewing the PostgreSQL Alert Log | 3-24 |
| 3.5 | ClusterDB Modeling | 3-25 |
| 3.5.1 | Physical View of the Cluster Networks | 3-25 |
| 3.5.2 | Physical View of the Storage | 3-32 |
| 3.5.3 | Machine View | 3-39 |
| 3.5.4 | HWMANAGER View | 3-44 |
| 3.5.5 | Complementary Tables | 3-46 |
| 3.5.6 | Nagios View | 3-48 |
| 3.5.7 | Lustre View | 3-49 |
| Chapter 4. | Software Deployment (KSIS) | 4-1 |
| 4.1 | Overview | 4-1 |
| 4.2 | Configuring and Verifying a Reference Node | 4-2 |
| 4.3 | Main Steps for Deployment | 4-3 |
| 4.4 | Checking Deployed Images | 4-4 |
| 4.4.1 | Checking Principles | 4-4 |
| 4.4.2 | Ksis Tests and Test Groups | 4-5 |
| 4.4.3 | Modifying the Checks Database | 4-6 |
| 4.4.4 | Examining the Check Results | 4-6 |
| 4.5 | Ksis Commands | 4-7 |
| 4.5.1 | Syntax | 4-7 |
| 4.5.2 | Advanced ksis create options | 4-8 |
| 4.5.3 | Creating the Image of the Reference Node | 4-8 |
| 4.5.4 | Deleting an Image or a Patch | 4-8 |
| 4.5.5 | Deploying an Image or a Patch | 4-9 |
| 4.5.6 | Removing a Patch | 4-9 |
| 4.5.7 | Getting Information about an Image or a Node | 4-9 |
| 4.5.8 | Listing Images on the Image Server | 4-9 |
| 4.5.9 | Listing Images by Nodes | 4-10 |
| 4.6 | Building a Patch | 4-11 |
| 4.7 | Checking Images | 4-12 |
| 4.8 | Importing and Exporting an Image | 4-12 |
| 4.9 | Rebuilding ClusterDB Data before Deploying an Image | 4-13 |
| Chapter 5. | Kerberos - Network Authentication Protocol | 5-1 |
| 5.1 | Environment | 5-1 |

| | | |
|-------------------|--|-------------|
| 5.1.1 | Kerberos Infrastructure | 5-1 |
| 5.1.2 | Authentication of the SSHv2 Connections..... | 5-1 |
| 5.2 | KERBEROS Infrastructure Configuration..... | 5-2 |
| 5.2.1 | secu0 Server including KDC Server and Administration Server | 5-2 |
| 5.2.2 | Configuration Files | 5-2 |
| 5.2.3 | Creating the Kerberos Database | 5-3 |
| 5.2.4 | Creating the Kerberos Administrator..... | 5-3 |
| 5.2.5 | Starting the KDC Server | 5-4 |
| 5.2.6 | Adding Access Control List (ACL) Rights for the Kerberos Administrator Created..... | 5-4 |
| 5.2.7 | Starting the Administration Daemon | 5-4 |
| 5.2.8 | Creating Principals Associated with Users | 5-4 |
| 5.2.9 | Creating Principals Associated with Remote Kerberized Services | 5-5 |
| 5.3 | Configuring the secu1 Machine that hosts the Host Principal remote service | 5-6 |
| 5.3.1 | Generating the key associated with the Host Principal remote service | 5-6 |
| 5.4 | Kerberos Authentication and SSH | 5-7 |
| 5.4.1 | Configuring the SSH Server on the secu1 machine..... | 5-7 |
| 5.4.2 | SSH Client | 5-8 |
| 5.5 | Troubleshooting Errors..... | 5-8 |
| 5.6 | Generating Associated Keys for Nodes of a Cluster..... | 5-9 |
| 5.7 | Modifying the Lifespan and Renewal Period for TGT Tickets | 5-10 |
| 5.8 | Including Addresses with Tickets | 5-10 |
| Chapter 6. | Storage Device Management..... | 6-1 |
| 6.1 | Overview of Storage Device Management for Bull Extreme Computing clusters | 6-1 |
| 6.2 | Monitoring Node I/O Status | 6-4 |
| 6.2.1 | Managing I/O Reference Counters | 6-4 |
| 6.2.2 | I/O Counters Definitions | 6-5 |
| 6.2.3 | Managing I/O Resources | 6-6 |
| 6.3 | Monitoring Storage Devices | 6-8 |
| 6.3.1 | Bull System Manager - HPC Edition: Host and Service Monitoring for Storage Devices... .. | 6-8 |
| 6.3.2 | Bull System Manager - HPC Edition: Storage & I/O Information | 6-12 |
| 6.3.3 | Querying the Cluster Management Data Base | 6-17 |
| 6.4 | Monitoring Brocade Switch Status..... | 6-18 |
| 6.5 | Managing Storage Devices with Bull CLI | 6-21 |
| 6.5.1 | Bull FDA Storage Systems | 6-21 |

| | | |
|-------------------|---|-------------|
| 6.5.2 | DataDirect Networks Systems - DDN Commands | 6-22 |
| 6.5.3 | Bull Optima1250 Storage Systems | 6-23 |
| 6.5.4 | EMC/Clariion (DGC) Storage Systems | 6-24 |
| 6.6 | Using Management Tools | 6-24 |
| 6.7 | Configuring Storage Devices | 6-25 |
| 6.7.1 | Planning Tasks | 6-25 |
| 6.7.2 | Deployment Service for Storage Systems | 6-26 |
| 6.7.3 | Understanding the Configuration Deployment Service..... | 6-26 |
| 6.8 | User Rights and Security Levels for the Storage Commands | 6-29 |
| 6.8.1 | Management Node | 6-29 |
| 6.8.2 | Other Node Types | 6-30 |
| 6.8.3 | Configuration Files | 6-30 |
| Chapter 7. | Monitoring with Bull System Manager - HPC Edition | 7-1 |
| 7.1 | Launching Bull System Manager - HPC Edition..... | 7-2 |
| 7.2 | Access Rights | 7-3 |
| 7.2.1 | Administrator Access Rights | 7-3 |
| 7.2.2 | Standard User Access Rights | 7-3 |
| 7.2.3 | Adding Users and Changing Passwords | 7-3 |
| 7.3 | Hosts, Services and Contacts for Nagios | 7-4 |
| 7.4 | Using Bull System Manager - HPC Edition | 7-5 |
| 7.4.1 | Bull System Manager - HPC Edition – View Levels | 7-5 |
| 7.5 | Map Button | 7-6 |
| 7.5.1 | All Status Map View | 7-6 |
| 7.5.2 | Rack View | 7-7 |
| 7.5.3 | bullx blade map view | 7-7 |
| 7.5.4 | Host Services detailed View | 7-8 |
| 7.5.5 | Control view | 7-9 |
| 7.5.6 | Ping Map View | 7-9 |
| 7.6 | Status Button | 7-10 |
| 7.7 | Log Window | 7-11 |
| 7.8 | Alerts Button..... | 7-12 |
| 7.8.1 | Active Checks | 7-13 |
| 7.8.2 | Passive Checks | 7-14 |
| 7.8.3 | Alert Definition..... | 7-14 |
| 7.8.4 | Notifications | 7-14 |

| | | |
|-------------|--|-------------|
| 7.8.5 | Acknowledgments | 7-15 |
| 7.8.6 | Running a Script | 7-15 |
| 7.8.7 | Generating SNMP Alerts | 7-16 |
| 7.8.8 | Resetting an Alert Back to OK..... | 7-16 |
| 7.8.9 | nsmhpc.conf Configuration file | 7-16 |
| 7.8.10 | Comments..... | 7-16 |
| 7.9 | Storage Overview | 7-17 |
| 7.10 | Shell..... | 7-18 |
| 7.11 | Monitoring the Performance - Ganglia Statistics | 7-18 |
| 7.12 | Group Performance View | 7-18 |
| 7.13 | Global Performance View..... | 7-19 |
| 7.13.1 | Modifying the Performance Graph Views..... | 7-20 |
| 7.13.2 | Refresh Period for the Performance View Web Pages | 7-21 |
| 7.14 | Configuring and Modifying Nagios Services | 7-21 |
| 7.14.1 | Configuring Using the Database | 7-21 |
| 7.14.2 | Modifying Nagios Services | 7-21 |
| 7.14.3 | Changing the Verification Frequency..... | 7-22 |
| 7.14.4 | Nagios Services Service | 7-22 |
| 7.14.5 | Nagios Information | 7-22 |
| 7.15 | General Nagios Services..... | 7-23 |
| 7.15.1 | Ethernet Interfaces..... | 7-23 |
| 7.15.2 | Resource Manager Status..... | 7-23 |
| 7.15.3 | Hardware Status | 7-23 |
| 7.15.4 | Alert Log | 7-23 |
| 7.15.5 | I/O Status..... | 7-23 |
| 7.15.6 | Postbootchecker..... | 7-23 |
| 7.16 | Management Node Nagios Services..... | 7-24 |
| 7.16.1 | MiniSQL Daemon | 7-24 |
| 7.16.2 | Resource Manager Daemon | 7-24 |
| 7.16.3 | ClusterDB..... | 7-24 |
| 7.16.4 | Cron Daemon..... | 7-24 |
| 7.16.5 | Compute Power Available..... | 7-24 |
| 7.16.6 | Global File System bandwidth available | 7-24 |
| 7.16.7 | Storage Arrays available | 7-25 |
| 7.16.8 | Global File System Usage | 7-25 |
| 7.16.9 | I/O pairs Migration Alert..... | 7-25 |
| 7.16.10 | Backbone Ports Available | 7-25 |
| 7.16.11 | HA System Status..... | 7-25 |

| | | |
|-------------------|---|-------------|
| 7.16.12 | Kerberos KDC Daemon..... | 7-25 |
| 7.16.13 | Kerberos Admin Daemon..... | 7-25 |
| 7.16.14 | LDAP Daemon (Lustre clusters only)..... | 7-25 |
| 7.16.15 | Lustre file system access..... | 7-26 |
| 7.16.16 | NFS file system access..... | 7-26 |
| 7.16.17 | InfiniBand Links available..... | 7-26 |
| 7.16.18 | CMC Health..... | 7-26 |
| 7.17 | Ethernet Switch Services..... | 7-27 |
| 7.17.1 | Ethernet Interfaces..... | 7-27 |
| 7.17.2 | Fans..... | 7-27 |
| 7.17.3 | Ports..... | 7-27 |
| 7.17.4 | Power supply..... | 7-27 |
| 7.17.5 | Temperature..... | 7-27 |
| 7.18 | Cool Cabinet Door Services..... | 7-28 |
| 7.18.1 | Cool Cabinet Door Functional Status..... | 7-28 |
| 7.18.2 | Cool Cabinet Door Power Consumption..... | 7-28 |
| 7.18.3 | Cool Cabinet Door Delta Pressure..... | 7-28 |
| 7.18.4 | Cool Cabinet Door Ethernet Interfaces..... | 7-29 |
| 7.18.5 | Cool Cabinet Door Power Supply Status..... | 7-29 |
| 7.18.6 | Cool Cabinet Door Temperature Average..... | 7-29 |
| 7.18.7 | Cool Cabinet Door Valve Aperture..... | 7-29 |
| Chapter 8. | Managing PDUs..... | 8-1 |
| 8.1 | Configuring PDUs..... | 8-1 |
| 8.1.1 | Register the PDUs in the Cluster Database..... | 8-1 |
| 8.1.2 | Accessing the PDUs..... | 8-1 |
| 8.1.3 | Customizing the Configuration of the PDUs..... | 8-2 |
| 8.2 | Monitoring PDUs..... | 8-2 |
| Chapter 9. | Cluster Power Management Tool..... | 9-1 |
| 9.1 | Configuring Power Management for High Availability Clusters..... | 9-1 |
| 9.2 | Configuring the Monitoring of Power Consumption..... | 9-2 |
| 9.2.1 | Configuration..... | 9-2 |
| 9.2.2 | Viewing Power Consumption Data..... | 9-3 |
| 9.3 | Configuring Power Management..... | 9-3 |
| 9.4 | Activate - Deactivate Power Management Tool..... | 9-3 |

| | | |
|------------------------------------|---|-------------|
| Chapter 10. | CPU Frequency and Voltage Scaling..... | 10-1 |
| 10.1 | Performance Requirements for Jobs | 10-1 |
| 10.2 | BIOS Power Management Settings | 10-2 |
| 10.3 | Managing Power for the Compute Nodes | 10-2 |
| 10.3.1 | cpuspeed service | 10-2 |
| 10.3.2 | CPUFreq Sub-System Tools | 10-3 |
| 10.3.3 | Changing CPUFreq Settings | 10-4 |
| 10.4 | More Information | 10-4 |
| Glossary and Acronyms | | G-1 |

List of Figures

| | | |
|--------------|--|------|
| Figure 1-1. | bullx cluster suite Management Functions..... | 1-1 |
| Figure 3-1. | bullx cluster suite ClusterDB architecture | 3-1 |
| Figure 3-2. | Cluster Network – diagram 1 | 3-25 |
| Figure 3-3. | Cluster Network – diagram 2 | 3-26 |
| Figure 3-4. | Storage physical view | 3-32 |
| Figure 3-5. | Cluster Database – Machine view 1 | 3-39 |
| Figure 3-6. | Cluster Database – Machine view 2 | 3-40 |
| Figure 3-7. | HWManager view..... | 3-44 |
| Figure 3-8. | Cluster Database – Complementary tables..... | 3-46 |
| Figure 3-9. | Nagios View..... | 3-48 |
| Figure 3-10. | Cluster Database – Lustre view | 3-49 |
| Figure 4-1. | Main steps for deployment..... | 4-3 |
| Figure 6-1. | I/O Status – initial screen | 6-4 |
| Figure 6-2. | Bull System Manager HPC Edition - I/O Status Details | 6-5 |
| Figure 6-3. | Bull System Manager –HPC Edition – I/O Resources of a node..... | 6-7 |
| Figure 6-4. | Detailed service status for a storage host..... | 6-8 |
| Figure 6-5. | Bull System Manager opening console window with the Storage overview icon circled | 6-13 |
| Figure 6-6. | Storage overview | 6-14 |
| Figure 6-7. | Inventory view of faulty storage systems and components | 6-15 |
| Figure 6-8. | Storage detailed view | 6-16 |
| Figure 6-9. | Nodes I/O Overview..... | 6-17 |
| Figure 6-10. | Detailed Service status of a brocade switch | 6-20 |
| Figure 7-1. | Bull System Manager - HPC Edition opening view | 7-5 |
| Figure 7-2. | Map button all status opening view | 7-6 |
| Figure 7-3. | Rack view with the Problems window at the bottom | 7-7 |
| Figure 7-4. | bullx blade map view..... | 7-8 |
| Figure 7-5. | Host Service details | 7-8 |
| Figure 7-6. | Monitoring Control Window | 7-9 |
| Figure 7-7. | Status Overview screen | 7-10 |
| Figure 7-8. | Monitoring - Log Window..... | 7-11 |
| Figure 7-9. | Monitoring Service Status window for a host with the Log Alerts link highlighted..... | 7-11 |

| | | |
|--------------|--|------|
| Figure 7-10. | Alert Window showing the different alert states | 7-12 |
| Figure 7-11. | Monitoring Control Window used to set Active Checks for a Service | 7-13 |
| Figure 7-12. | Hostgroups Reporting Notifications Window showing the Notification Levels | 7-15 |
| Figure 7-13. | Storage overview window | 7-17 |
| Figure 7-14. | Group Performance view | 7-18 |
| Figure 7-15. | Global overview for a host (top screen) | 7-19 |
| Figure 7-16. | Detailed monitoring view for a host (bottom half of screen displayed in Figure 7-15) | 7-20 |
| Figure 7-17. | Ethernet Switch services | 7-27 |
| Figure 7-18. | Cool Cabinet Door Services | 7-28 |

List of Tables

| | | |
|-------------|---|------|
| Table 2-1. | Maintenance Tools | 2-7 |
| Table 3-1. | CLUSTER table | 3-27 |
| Table 3-2. | IP_NW table | 3-27 |
| Table 3-3. | ETH_SWITCH table | 3-28 |
| Table 3-4. | IC_NW table | 3-28 |
| Table 3-5. | IC_SWITCH table | 3-29 |
| Table 3-6. | SERIAL_NW table | 3-29 |
| Table 3-7. | PORTSERVER table | 3-29 |
| Table 3-8. | ETH_VLAN table | 3-30 |
| Table 3-9. | FC_NW table | 3-30 |
| Table 3-10. | CLUSTER_IPV table | 3-30 |
| Table 3-11. | FC_SWITCH table | 3-31 |
| Table 3-12. | TALIM table | 3-31 |
| Table 3-13. | Storage – disk_array table | 3-34 |
| Table 3-14. | Storage – da_enclosure table | 3-34 |
| Table 3-15. | Storage – da_disk_slot table | 3-34 |
| Table 3-16. | Storage – da_controller table | 3-35 |
| Table 3-17. | Storage – da_fc_port table | 3-35 |
| Table 3-18. | Storage – da_serial_port table | 3-35 |
| Table 3-19. | Storage – da_ethernet_port table | 3-36 |
| Table 3-20. | Storage – da_power_supply table | 3-36 |
| Table 3-21. | Storage – da_fan table | 3-36 |
| Table 3-22. | Storage – da_power_fan table | 3-37 |
| Table 3-23. | Storage – da_temperature_sensor table | 3-37 |
| Table 3-24. | da_io_path table | 3-37 |
| Table 3-25. | Storage – da_iocell_component table | 3-37 |
| Table 3-26. | Storage – da_cfg_model table | 3-38 |
| Table 3-27. | Storage – da_power_port table | 3-38 |
| Table 3-28. | Machine view – NODE table | 3-41 |
| Table 3-29. | Machine view – NODE_IMAGE table | 3-42 |
| Table 3-30. | Machine view – NODE_PROFILE table | 3-42 |
| Table 3-31. | Machine view – IC_BOARD table | 3-43 |
| Table 3-32. | Machine view – IPOIB table | 3-43 |
| Table 3-33. | Machine view – SDPOIB table | 3-43 |
| Table 3-34. | Machine view – FC_BOARD table | 3-44 |
| Table 3-35. | HWMANAGER Table | 3-45 |
| Table 3-36. | ADMIN table | 3-46 |
| Table 3-37. | RACK table | 3-46 |

| | | |
|-------------|---|------|
| Table 3-38. | RACK_PORT table | 3-47 |
| Table 3-39. | CONFIG_CANDIDATE table | 3-47 |
| Table 3-40. | CONFIG_STATUS table | 3-47 |
| Table 3-41. | GROUP_NODE table | 3-48 |
| Table 3-42. | SERVICES table | 3-48 |
| Table 3-43. | AVAILABILITY table | 3-49 |
| Table 3-44. | Lustre_FS table | 3-50 |
| Table 3-45. | Lustre OST table | 3-51 |
| Table 3-46. | Lustre_MDT table | 3-51 |
| Table 3-47. | Lustre_IO_node table | 3-51 |
| Table 3-48. | Lustre_mount table | 3-52 |
| Table 4-1. | Standard checks delivered with Ksis | 4-5 |

Preface

Scope and Objectives

The **BAS5 for Xeon** software suite has been renamed as **bullx cluster suite (bullx CS)**. Existing **BAS5 for Xeon** distributions can be upgraded to **bullx cluster suite XR 5v3.1U2**. **bullx cluster suite** is used for the management of all the nodes of a Bull Extreme Computing cluster.

The purpose of this guide is to explain how to configure and manage **Bull extreme computing** clusters, using the administration tools recommended by Bull.

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

Intended Readers

This guide is for administrators of bullx cluster suite system.

Prerequisites

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

Bibliography

Refer to the manuals included on the documentation CD delivered with your system OR download the latest manuals for your **bullx cluster suite** release, and for your cluster hardware, from: <http://support.bull.com/>

The *bullx cluster suite Documentation* CD-ROM (86 A2 12FB) includes the following manuals:

- *bullx cluster suite Installation and Configuration Guide* (86 A2 19FA)
- *bullx cluster suite Administrator's Guide* (86 A2 20FA)
- *bullx cluster suite Application Developer's Guide* (86 A2 22FA)
- *bullx cluster suite Maintenance Guide* (86 A2 24FA)
- *bullx cluster suite High Availability Guide* (86 A2 25FA)
- *InfiniBand Guide* (86 A2 42FD)
- *LDAP Authentication Guide* (86 A2 41FD)
- *SLURM Guide* (86 A2 45FD)
- *Lustre Guide* (86 A2 46FD)

The following document is delivered separately:

- *The Software Release Bulletin (SRB)* (86 A2 80EJ)



Important

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

For **Bull System Manager**, refer to the *Bull System Manager* documentation suite.

For clusters that use the **PBS Professional** Batch Manager, the following manuals are available on the *PBS Professional CD-ROM*:

- *Bull PBS Professional Guide* (86 A2 16FE)
- *PBS Professional Administrator's Guide*
- *PBS Professional User's Guide* (on the *PBS Professional CD-ROM*)

For clusters that use **LSF**, the following manuals are available on the *LSF CD-ROM*:

- *Bull LSF Installation and Configuration Guide* (86 A2 39FB)
- *Installing Platform LSF on UNIX and Linux*

For clusters which include the **Bull Cool Cabinet**:

- *Site Preparation Guide* (86 A1 40FA)
- *R@ck'nRoll & R@ck-to-Build Installation and Service Guide* (86 A1 17FA)
- *Cool Cabinet Installation Guide* (86 A1 20EV)
- *Cool Cabinet Console User's Guide* (86 A1 41FA)
- *Cool Cabinet Service Guide* (86 A7 42FA)

Highlighting

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

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

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

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

- Values to be entered in by the user are in 'Courier New', for example:
COM1
- Commands, files, directories and other items whose names are predefined by the system are in '**Bold**', as shown below:
The **/etc/sysconfig/dump** file.
- The use of *Italics* identifies publications, chapters, sections, figures, and tables that are referenced.
- < > identifies parameters to be supplied by the user, for example:
<node_name>



WARNING

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

Chapter 1. Cluster Management Functions and Corresponding Products

bullx cluster suite is a software distribution that includes a suite of tools and programs to run and monitor **Bull Extreme Computing** clusters.

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

The administration tools are mainly Open Source products that are configurable and adaptable to the requirements of the cluster, and which can be deactivated on demand.

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

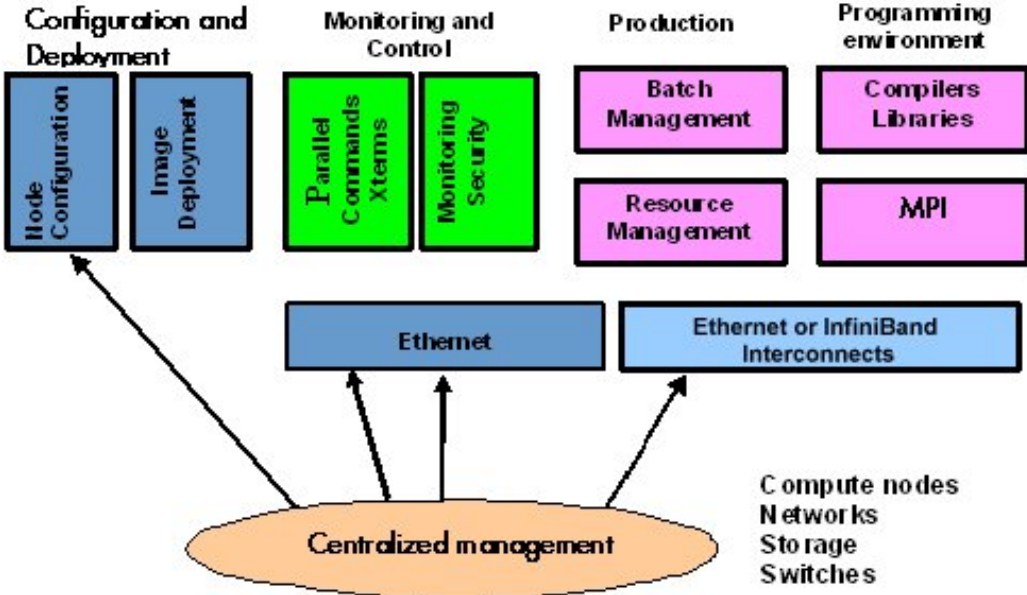


Figure 1-1. bullx cluster suite Management Functions

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 – *Initial Configuration Tasks for Bull HPC Clusters* 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.
- **KSIS** which is used to produce and deploy node images.
See Chapter 4 – *Software Deployment (KSIS)* for more information.
- **Kerberos** (optional) – A Security Suite used to validate users, services and machines for a whole network.
See Chapter 5 – *Kerberos - Network Authentication Protocol* for more information.

Storage and File system Management

- Various tools are available for managing and monitoring the different storage devices which exist for Bull Extreme Computing Clusters.
See Chapter 6 – *Storage Device Management* for more information.
- Parallel file systems ensure high I/O rates for clusters where data is accessed by a large number of processors at once.
See the *Lustre Guide* for more information.

Resource and Batch Management

- **SLURM** (Simple Linux Utilities Resource Manager) an open-source scalable resource manager.
See the *SLURM Guide* for more information.
- **PBS Professional** (optional) is a batch manager that is used to queue, schedule and monitor jobs.
See the *PBS Professional Guide* for more information.



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

Monitoring

- **Bull System Manager - HPC Edition** monitors the cluster and activity and is included in the delivery for all Bull Extreme Computing Clusters.
See Chapter 7 – *Monitoring with Bull System Manager – HPC Edition* for more information.
- **HPC Toolkit** provides a set of profiling tools that help you to improve the performance of the system.
See the *Application Developer's Guide* for more information.

Chapter 2. Initial Configuration Tasks

Most configuration tasks are carried out at the time of installation. This chapter describes how the Administrator carries out some additional configuration tasks. It also covers the security policy for Extreme Computing systems.

The following topics are described:

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

See The *Installation and Configuration Guide*, which describes the initial installation and configuration steps for Bull Extreme Computing clusters, for more information.

2.1 Configuring Services

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

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

- To display the Help details 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**, 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 Configuring Security

This section provides the administrator with basic rules concerning cluster security. Different security policies can be set up according to the cluster's configuration.

The Management Node is the most sensitive part of the cluster from a security point of view. This node submits jobs in batch mode and is the 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.

Compute Nodes 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 individual user should not have access to other partitions.

2.3.1 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** manually for each node responding yes at the prompt to add the node 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.

2.4 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.

2.4.1 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 I/O 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
-----
```

2.4.2 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 copies `/etc/hosts` from the node where `pdcp` executes to all the nodes specified using the `-w` option by copying across the same path with the command.

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

2.4.3 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 output:

- 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, and 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
```


2.5 Day to Day Maintenance Operations

A set of maintenance tools is provided with a Bull Extreme Computing 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 Extreme Computing clusters which may include hundred of nodes.

| Function | Tool | Purpose |
|------------------|-----------------------------------|---|
| Administration | ConMan ipmitool | Managing Consoles through Serial Connection |
| | nsclusterstop / nsclusterstart | Stopping/Starting the cluster |
| | nsctrl | Managing hardware (power on, power off, reset, status, ping checking temperature, changing bios, etc) |
| | Remote Hardware Management CLI | |
| | nsfirm | Obtaining the BMC or BIOS version, upgrading firmware, etc. |
| | syslog-ng | System log Management |
| | lptools (lputils, lpflash) | Upgrading Emulex HBA Firmware (Host Bus Adapter) |
| Backup / Restore | Bull System Backup Restore (BSBR) | Backing-up and restoring data |
| Monitoring | ibstatus, ibstat | Monitoring InfiniBand networks |
| | IBS tool | Providing information about and configuring InfiniBand switches |
| | lsiocfg | Getting information about storage devices |
| | pingcheck | Checking device power state |
| Debugging | ibtracert | Identifying InfiniBand network problem |
| | crash/proc/kdump | Runtime debugging and dump tool |
| | hpcsnap | Collecting cluster information |
| Testing | postbootchecker | Making verifications on nodes as they start |

Table 2-1. Maintenance Tools

See

- The *Maintenance Guide* for more information.
- The *Application Developer's Guide* for details on Bull HPC Toolkit, a set of cluster profiling tools.

Chapter 3. Cluster Database Management

This chapter describes the architecture of the Cluster Database, and the commands and tools which enable the administrator to display, and to change the 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 bullx cluster suite delivery includes the data that is required for the cluster management tools (**BSM – HPC Edition, KSIS, pdsh, syslog-ng, ConMan** etc.). Compared with sequential configuration files, the advantages of using a database are flexibility, and data availability for all the tools, ensuring better integration without the duplication of common data. Cluster database management uses the highly-scalable, **SQL** compliant, Open Source object-relational **PostgreSQL**. The figure below shows the architecture for the **ClusterDB** and its relationship to the cluster management tools.

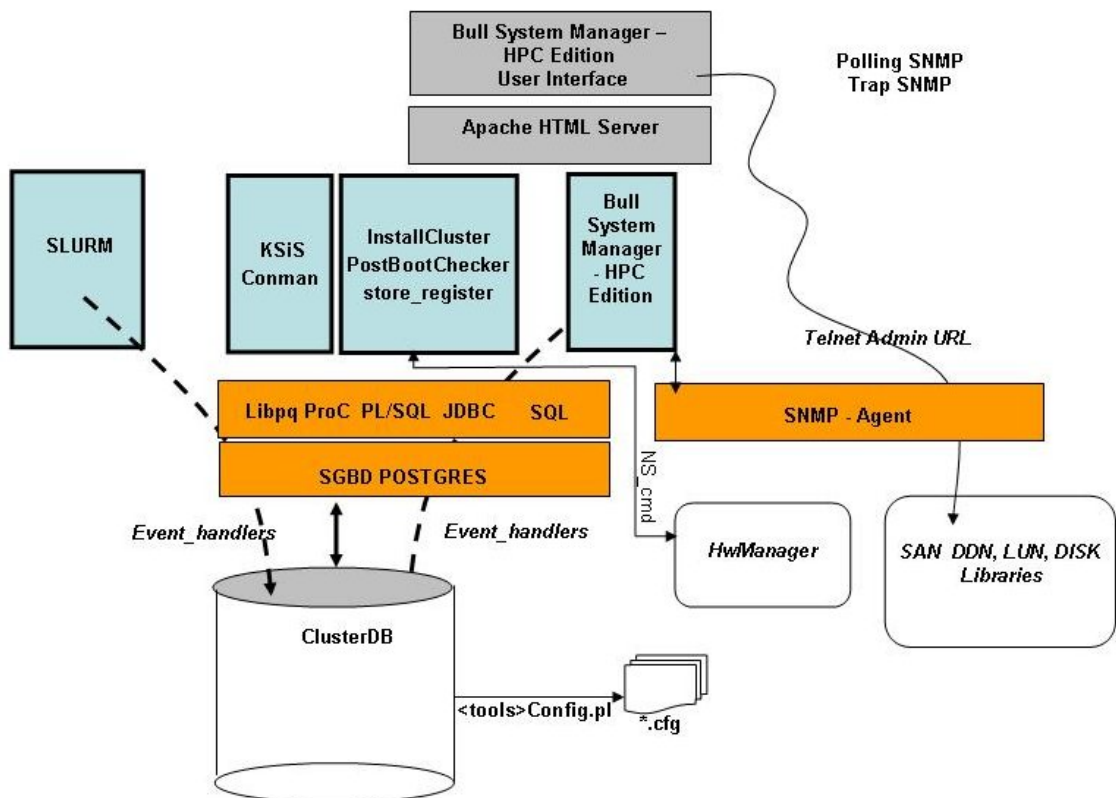


Figure 3-1. bullx cluster suite ClusterDB architecture

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 uses **PostgreSQL** tools (see 3.4 *Managing the ClusterDB*).

3.3 Using Commands

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

| | |
|--|--|
| <u>changeOwnerProperties</u> | Changes the confidentiality parameters |
| <u>dbmConfig</u> | Controls the consistency of the ClusterDB with the system. All database updates are marked to be a "candidate" for synchronization. |
| <u>dbmCluster</u> | Operates on the whole cluster to get information, to check IP addresses and to check rack configuration. |
| <u>dbmNode</u> | Displays information, or change attributes at the node level. |
| <u>dbmHwManager</u> | Displays information, or change attributes at the Hwmanager level. |
| <u>dbmGroup</u> | Manages the groups of nodes. |
| <u>dbmEthernet</u> | Displays information, or change attributes for the Ethernet switches. |
| <u>dbmIconnect</u> | Displays information, or change attributes for the interconnect switches. |
| <u>dbmTalism</u> | Displays information, or change attributes for the remotely controlled power supply devices. |
| <u>dbmSerial</u> | Displays information, or change attributes for the portservers. |
| <u>dbmFiberChannel</u> | Displays information about the Fiber Switches or changes the values of some attributes for a Fiber Switch or a subset of Fiber Switches. |
| <u>dbmServices</u> | Displays information about the Services or changes the values of some attributes for a Service. |
| <u>dbmDiskArray</u> | 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 will have been modified by the client as a result of:

- Running **ECT** (Embedded Configuration Tool) for Interconnect switches
- Running **bmcConfig** for BMC cards
- Running **swtConfig** for Ethernet switches
- The network configuration of the nodes 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 HA 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 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>]
                        [--bkoffset <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 basename . |
| --basename | Specifies the basename of the node. (The node name is constituted of basename + netid). It is also the virtual node name. |
| --adprivacy | 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 adprivacy option can specify two bytes in the form A.V . |
| --icprivacy | 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 icprivacy option can specify three bytes in the form A.V.U . |

- bkprivacy** 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 **bkprivacy** option can specify three bytes in the form **A.V.U**.
- bkgw** Specifies the backbone gateway
- bkdom** Specifies the backbone domain
- bkoffset** Specifies the backbone translation offset. It permits to translate the D.E.U.H backbone ip to D.E.(U + bkoffset).H

Example

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

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

Process

1. Retrieve the current privacy bytes by running.

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

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

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

5. Restart **dbmConfig**.
6. Reconfigure Ethernet switches by running.

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

7. Reconfigure the IP addresses of the BMC cards.

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

8. Manually configure on the FDA (if present).
9. Run **ddn_init** on each DDN and reboot (if DDN storage is used).
10. HA 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**, **snmptt**, **group**, **BSM**).

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

- show** Displays the synchronization state of all the services or of a list of specified services.
- configure** 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 `/var/log/synchro.log` file. It is necessary to check these messages.
Note: The command reports an OK status to indicate that it has completed. This does not mean that no configuration error occurred.
- help** Displays the syntax of the **dbmConfig** command.

Options

- dbname** Specifies the name of the database to which the command applies. Default value: clusterdb.
Note: This option must only be used by qualified people for debugging purposes.
- force** Reconfigures the service and restarts it.
- id** Reloads the configuration of the portserver identified by id. This option applies only to the portserver service (`--service=portserver` option).
- impact** Displays the configuration files and associated services impacted by the next **dbmConfig configure** command.
- nodeps** Forces the reconfiguration, despite the inter service dependencies.
- restart** Restarts the service instead of reloading it.
- service** Specifies the service from the following: **syshosts**, **sysdhcpd**, **conman**, **portserver**, **pdsh**, **nagios**, **snmptt**, **group**, **BSM**. For more information see Updated Configuration Files below.
- timeout** 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 |
|-------------------|---|
| syshosts | Updates the /etc/hosts file with the data available in the administration base |
| sysdhcpd | Updates the /etc/dhcpd.conf file with the data available in the administration base. |
| conman | Updates the /etc/conman.conf file with the data available in the administration base. |
| portserver | Updates the portserver configuration file (/tftpboot/ps16*ConfigTS16 or /tftpboot/ps14*ConfigTS4), reloads the file on the appropriate portserver and reboots it. |
| pdsh | Updates the /etc/genders file with the data available in the administration base. |
| nagios | Updates several configuration files (/etc/nagios/*.cfg) with the data available in the administration base. |
| snmpft | Updates the /etc/snmp/storage_hosts file with the data available in the administration base. |
| group | Creates the predefined groups in the database. (No configuration file is updated.) |
| BSM | 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
```


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

Options

| | |
|-------------------|--|
| --dbname | Specifies the name of the database to which the command applies. Default: clusterdb . Note: this option must be used only by qualified people for debugging purposes. |
| --ipaddr | Checks that the IP addresses are distinct within the cluster. |
| --rack | 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. |
| --unitCell | Checks that the object Unit and Cell number are the same as the Ethernet switch connected to. |
| --profile | Used to set one key/value pair to be changed in table cluster. key must be in [actif_ha, actif_crm, actif_vlan, resource_manager, batch_manager, security_parallel_fs] |

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

| | |
|---------------------|---|
| dbmNode show | [--sysimage [--install_status={installed not_installed in_installed}]] |
| dbmNode show | [--name <node name> --hwmanager --cpu --iproute --serialroute] |
| dbmNode show | [--unit <unit_num> --hwmanager --cpu] [--dbname <database name>] |
| dbmNode set | --name=<node name> --status={managed not_managed} --admin_macaddr <macaddr> --backbone_macaddr <macaddr> |
| dbmNode set | --unit <unit num> --status={managed not_managed} |
| dbmNode set | --nodelist=<node list> --status={managed not_managed} |
| dbmNode set | (--name=<node name> --unit <unit num>) --cpu <total cpu chipset> |

```
dbmNode set ( --name=<node name> | --unit <unit num> ) --hyperthreading={yes | no}
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 BMC / CMC 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 eth1 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----- | BAS5-16K | up |
| node[3] | -C----- | BAS5-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----- | BAS5-16K | up |
| node[3] | -C----- | BAS5-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 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. |
| --all | Deletes all nodes. |
| --addnodelist | Adds a node list in an existing group. |
| --delnodelist | Deletes a node list in an existing group. |
| --dbname | Specifies the name of the database on which the command is applied. Default: clusterdb. Note: this option must only be used 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 compute 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 META_INF [ 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, status information and hwmanager (if any) 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. |
| --name | Specifies the talim name to which the action applies. |
| --iproute | Displays the Ethernet path (the localization and status of all Ethernet switches) between the talim and the admin node |
| --macaddr | Displays the macaddr or changes the macaddr of the Talim. |
| --status | 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. |
| --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 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: This 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/clusmngt/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/clusmngt/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 bullx cluster suite systems. Certain tables will only be exploited by the functionality of BAS4, for bullx cluster suite these tables will be empty.

3.5.1 Physical View of the Cluster Networks

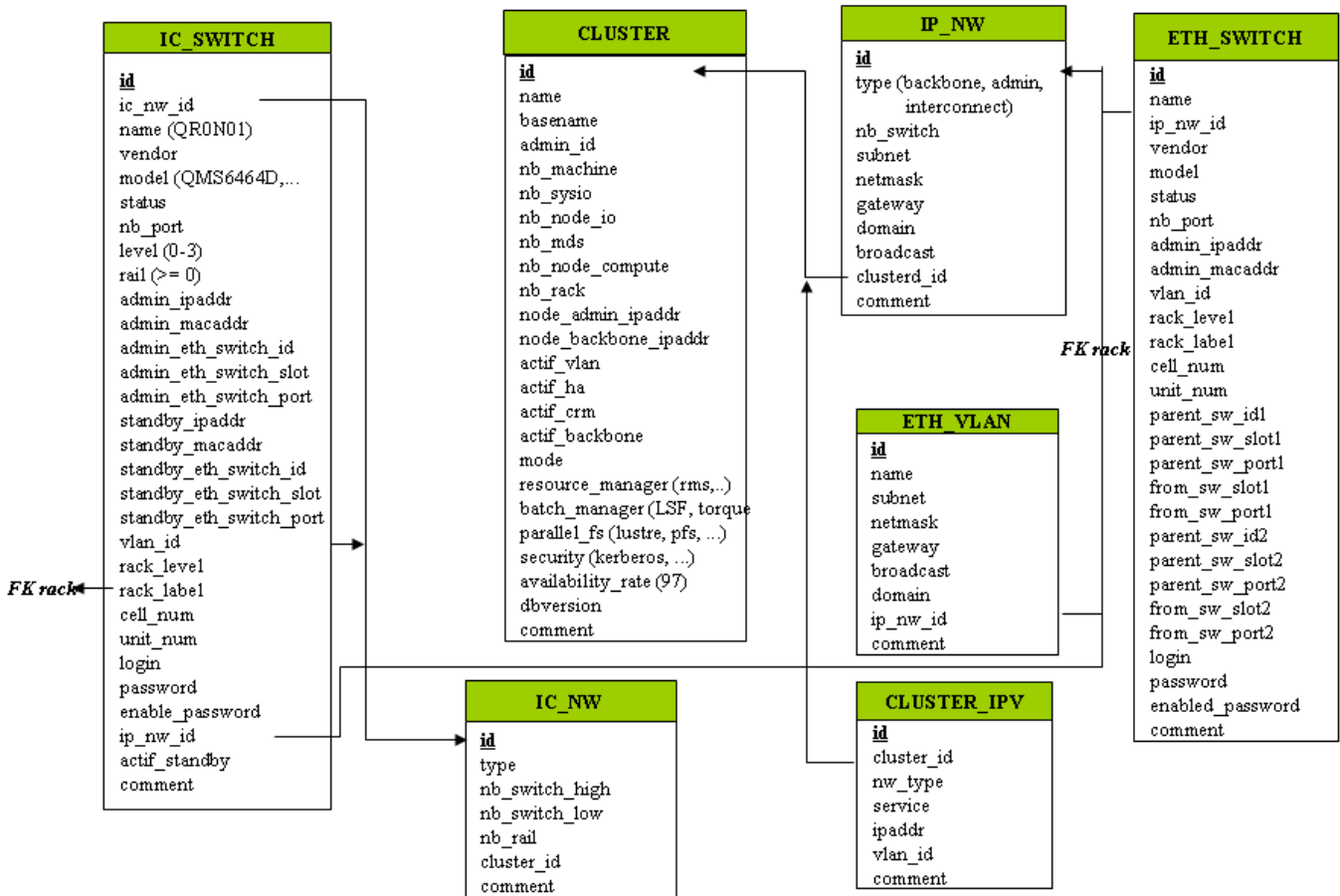


Figure 3-2. Cluster Network – diagram 1

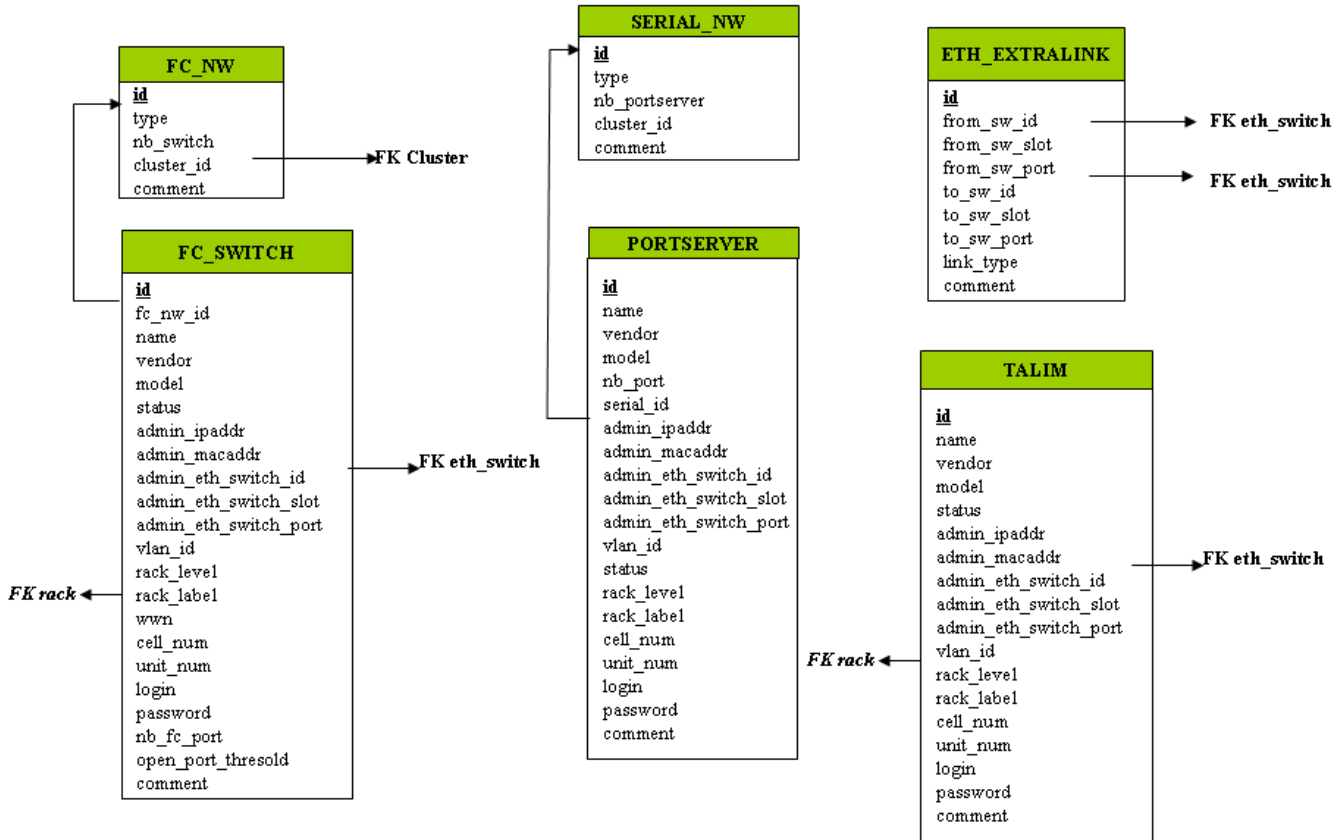


Figure 3-3. Cluster Network – diagram 2

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 |
| 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 |

| Column name | Description | Example | Fill in method |
|-------------------|--|---------|----------------|
| parallel_fs | Lustre | lustre | prelaod |
| security | Kerberos | NULL | preload |
| availability_rate | Availability rate | NULL | Not used |
| dbversion | Development model version for the database | 20.5.0 | DV at 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& loadClusterdb |
| netmask | Sub-network mask | 255.255.0.0 | preload& loadClusterdb |
| 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 | | swtAdmin |
| 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 st parent switch | 0 | preload |
| parent_sw_port1 | Connection port for the 1st switch | 1 | preload |
| from_sw_slot1 | Starting slot number of the 1 st switch | 0 | preload |
| from_sw_port1 | Starting port number of the 1 st switch | 1 | preload |
| parent_sw_id2 | Ethernet switch 2 nd parent | | preload |
| from_sw_slot2 | Starting slot number of the 2 nd switch | | preload |

| Column name | Description | Example | Fill in method |
|------------------|---|---------|----------------|
| parent_sw_port2 | Starting port number for the 2 nd switch | 2 | preload |
| from_sw_slot2 | Starting slot number of the 2 nd switch | | preload |
| from_sw_port2 | Starting port number of the 2 nd switch | | preload |
| login | Administration login | | cmdExpl |
| password | Administration password | | cmdExpl |
| enabled_password | Manufacturer's enabled password | | ECT |
| comment | Free field | | |

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, GbEther | 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 |
| 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 rail | 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 | CO-A16 | preload |
| cell_num | Number of the cell | 1 | preload |

| Column name | Description | Example | Fill in method |
|-----------------|---------------------------------------|---------|----------------|
| 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 =false |
| 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 |
| type | PAP, node, storage, mixed networks | node | preload |
| nb_portserver | Number of PortServer | 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 bullx cluster suite 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 | 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 | 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 CLUSTER_IPV Table

| Column name | Description | Example | Fill in method |
|-------------|-----------------------------------|------------|--------------------|
| id | PK | 1 | preload - sequence |
| cluster_id | FK on the CLUSTER | | preload |
| nw_type | Network type | nfs | preload |
| service | nfs | nfs | preload |
| ipaddr | Virtual IP address of the service | 10.11.0.99 | preload |
| vlan_id | FK on ETH_VLAN | | preload |
| comment | Free field | | |

Table 3-10. CLUSTER_IPV table

3.5.1.11 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 |

| Column name | Description | Example | Fill in method |
|-----------------------|---|------------------------------|------------------|
| vendor | Name of the vendor | BROCADE | preload |
| model | SW model | Silkworm 200 ^E | 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 | NULL | |
| 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 |
| password | Administration Password | | prelaod |
| nb_fc_port | Number of fibre channel ports | | preload |
| open_port_thresold | | | preload |
| comment | Free field | | |

Table 3-11. FC_SWITCH table

3.5.1.12 TALIM Table

| Column name | Description | Example | Fill-in method |
|-----------------------|--------------------------------------|---------|------------------|
| id | PK | | preload-Sequence |
| name | Name of the power switch | | preload |
| vendor | Vendor name | | 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 | NULL | |
| 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 | | prelaod |
| comment | Free field | | |

Table 3-12. TALIM table

3.5.1.13 ETH_EXTRALINK Table

This table is not active in this version.

3.5.2 Physical View of the Storage

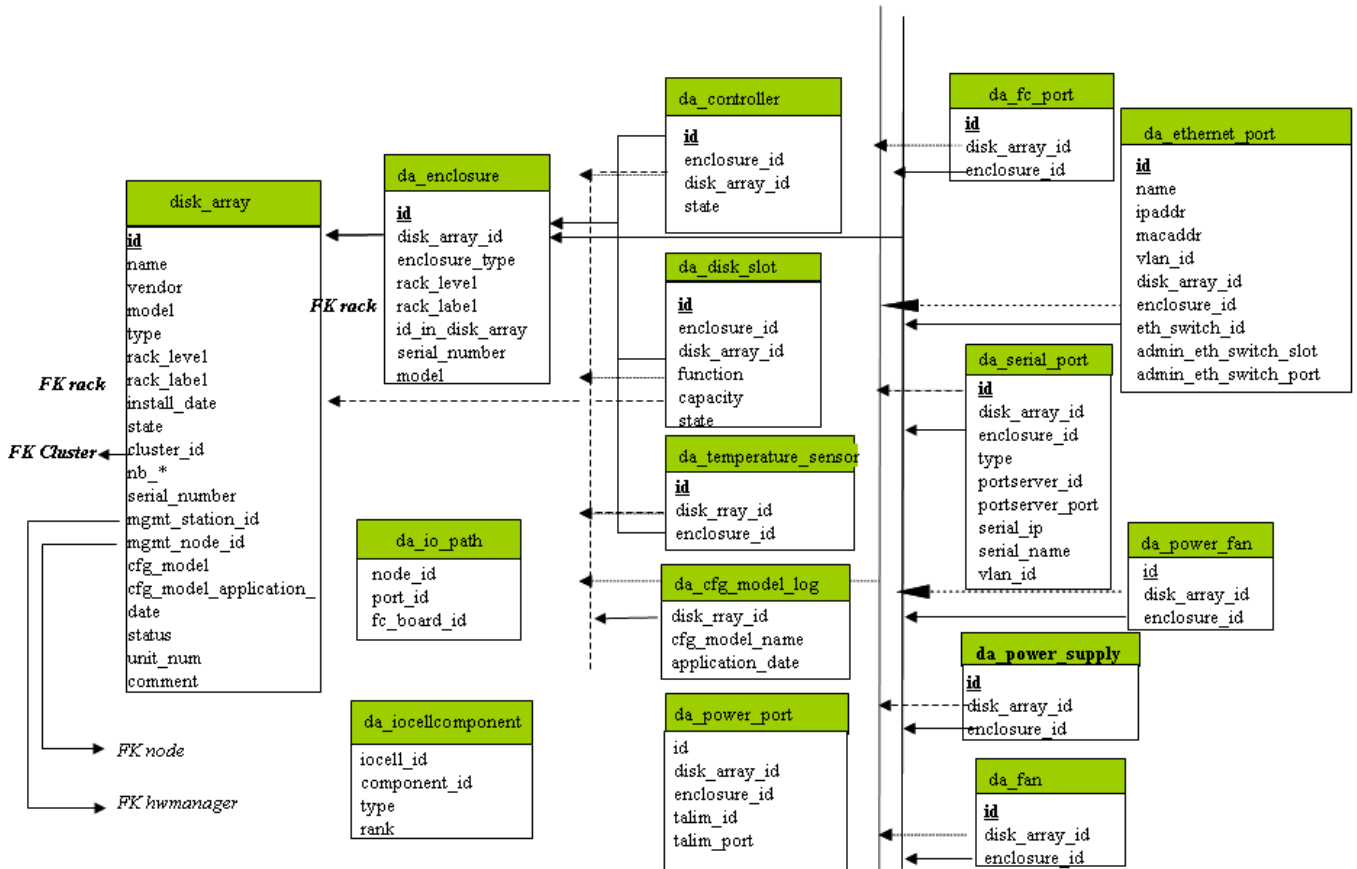


Figure 3-4. Storage physical view

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 - BSM |
| cluster_id | Id of the cluster parent | preload |
| nb_enclosure | Number of disk enclosure | Dynamic (DV=0) - BSM |
| nb_controller | Number of controller | Dynamic (DV=0) - BSM |
| nb_controller_ok | Number of controller in OK state | Dynamic (DV=0) - BSM |
| nb_controller_faulty | Number of controller in FAULTY state | Dynamic (DV=0) - BSM |

| Field name | Field information | Fill in method |
|--------------------------------|--|-----------------------|
| nb_fc_port | Number of FC ports | Dynamic (DV=0) - BSM |
| nb_fc_port_connected | Number of FC ports in CONNECTED state | Dynamic (DV=0) - BSM |
| nb_fc_port_not_connected | Number of FC ports in NOT_CONNECTED state | Dynamic (DV=0) - BSM |
| nb_fc_port_disconnected | Number of FC ports in DISCONNECTED state | Dynamic (DV=0) - BSM |
| nb_fc_port_faulty | Number of FC ports in FAULTY state | Dynamic (DV=0) - BSM |
| nb_serial_port | Number of serial ports | Dynamic (DV=0) - BSM |
| nb_serial_port_connected | Number of serial ports in CONNECTED state | Dynamic (DV=0) - BSM |
| nb_serial_port_not_connected | Number of serial ports in NOT_CONNECTED state | Dynamic (DV=0) - BSM |
| nb_serial_port_disconnected | Number of serial ports in DISCONNECTED state | Dynamic (DV=0) - BSM |
| nb_serial_port_faulty | Number of serial ports in FAULTY state | Dynamic (DV=0) - BSM |
| nb_eth_port | Number of Ethernet ports | Dynamic (DV=0) - BSM |
| nb_ethernet_port_connected | Number of Ethernet ports in CONNECTED state | Dynamic (DV=0) - BSM |
| nb_ethernet_port_not_connected | Number of Ethernet ports in NOT_CONNECTED state | Dynamic (DV=0) - BSM |
| nb_ethernet_port_disconnected | Number of Ethernet ports in DISCONNECTED state | Dynamic (DV=0) - BSM |
| nb_ethernet_port_faulty | Number of Ethernet ports in FAULTY state | Dynamic (DV=0) - BSM |
| nb_disk | Number of disks | Dynamic (DV=0) - BSM |
| nb_disk_slot_ok | Number of disks in OK state | Dynamic (DV=0) - BSM |
| nb_disk_slot_faulty | Number of disks in FAULTY state | Dynamic (DV=0) - BSM |
| nb_disk_slot_empty | Number of disks in EMPTY state | Dynamic (DV=0) - BSM |
| nb_disk_slot_used_spare | Number of disks slots in USED_SPARE state | Dynamic (DV=0) - BSM |
| nb_disk_slot_missing | Number of disks in MISSING state | Dynamic (DV=0) - BSM |
| nb_power_supply | Number of power supplies | Dynamic (DV=0) - BSM |
| nb_power_supply_ok | Number of power supplies in OK state | Dynamic (DV=0) - BSM |
| nb_power_supply_faulty | Number of power supplies in FAULTY state | Dynamic (DV=0) - BSM |
| nb_nb_fan | Number of fans | Dynamic (DV=0) - BSM |
| nb_fan_ok | Number of fans in OK state | Dynamic (DV=0) - BSM |
| nb_fan_faulty | Number of fans in FAULTY state | Dynamic (DV=0) - BSM |
| nb_nb_power_fan | Number of power_fan | Dynamic (DV=0) - BSM |
| nb_power_fan_ok | Number of power_fan in OK state | Dynamic (DV=0) - BSM |
| nb_power_fan_faulty | Number of power_fan in FAULTY state | Dynamic (DV=0) - BSM |
| nb_nb_temperature_sensor | Number of temperature sensors | Dynamic (DV=0) - BSM |
| nb_temperature_sensor_ok | Number of temperature sensors in OK state | Dynamic (DV=0) - BSM |
| nb_temperature_sensor_warning | Number of temperature sensors in WARNING state | Dynamic (DV=0) - BSM |
| nb_temperature_sensor_faulty | Number of temperature sensors in FAULTY state | Dynamic (DV=0) - BSM |
| nb_lun | Number of lun | Dynamic (DV=0) - BSM |
| nb_spare | Number of spare disk | Dynamic (DV=0) - BSM |
| 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 |

| Field name | Field information | Fill in method |
|----------------------------|------------------------------------|---------------------------|
| 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 – BSM (DV="up") |
| unit_num | Unit Number | preload |
| comment | Free field | |

Table 3-13. 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 3-14. 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 |
| 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 – BSM |
| disk_enclosure_id | Disk number in the enclosure | Automatic - storeregister |
| vendor_location | Location of the component expressed in the vendor terms. | Automatic - storeregister |

Table 3-15. 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 – BSM |
| object_num | Controller identifier in the enclosure | Automatic – storeregister |
| vendor_location | Location of the component expressed in the vendor terms. | Automatic – storeregister |

Table 3-16. 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 – BSM |
| object_num | fc_port identifier in the enclosure | preload |
| vendor_location | Location of the component expressed in the vendor terms. | Automatic – storeregister |

Table 3-17. 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 – BSM |
| object_num | Serial port identifier in the enclosure | Preload |
| vendor_location | Location of the component expressed in the vendor terms. | Automatic – storeregister |

Table 3-18. 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 – BSM |
| 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 3-19. 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 – BSM |
| object_num | Power supply identifier in the enclosure | Automatic – storeregister |
| vendor_location | Location of the component expressed in the vendor terms. | Automatic – storeregister |

Table 3-20. 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 – BSM |
| object_num | Fan identifier in the enclosure | Automatic – storegister |
| vendor_location | Location of the component expressed in the vendor terms. | Automatic – storegister |

Table 3-21. 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 – BSM |
| object_num | Power_fan identifier in the enclosure | Automatic- storeregister |
| vendor_location | Location of the component expressed in the vendor terms. | Automatic- storeregister |

Table 3-22. 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 – BSM |
| object_num | Temperature sensor identifier in the enclosure | Automatic – storeregister |
| vendor_location | Location of the component expressed in the vendor terms. | Automatic – storeregister |

Table 3-23. 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 3-24. 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 3-25. 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 3-26. 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 3-27. Storage – da_power_port table

3.5.3 Machine View

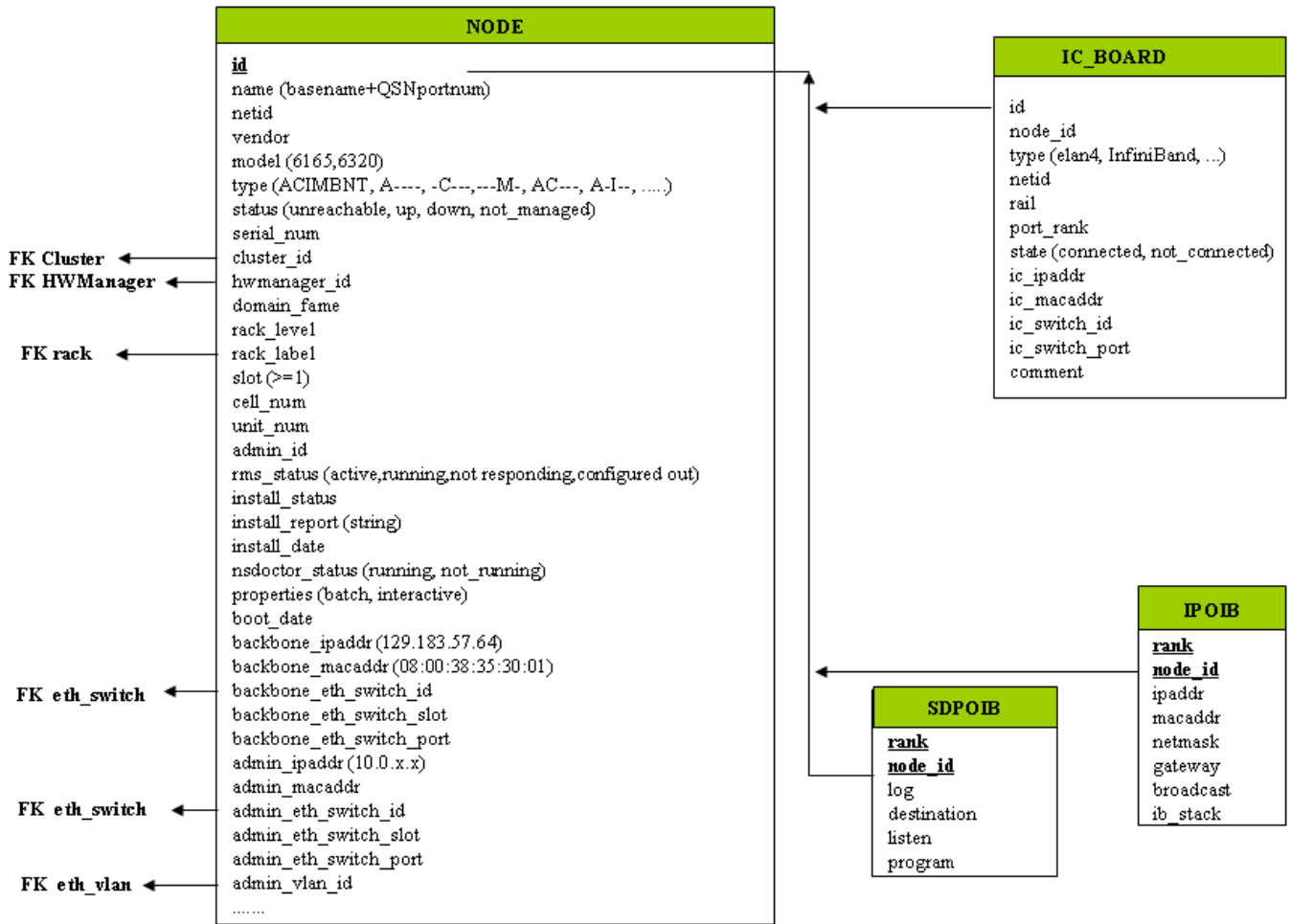


Figure 3-5. Cluster Database – Machine view 1

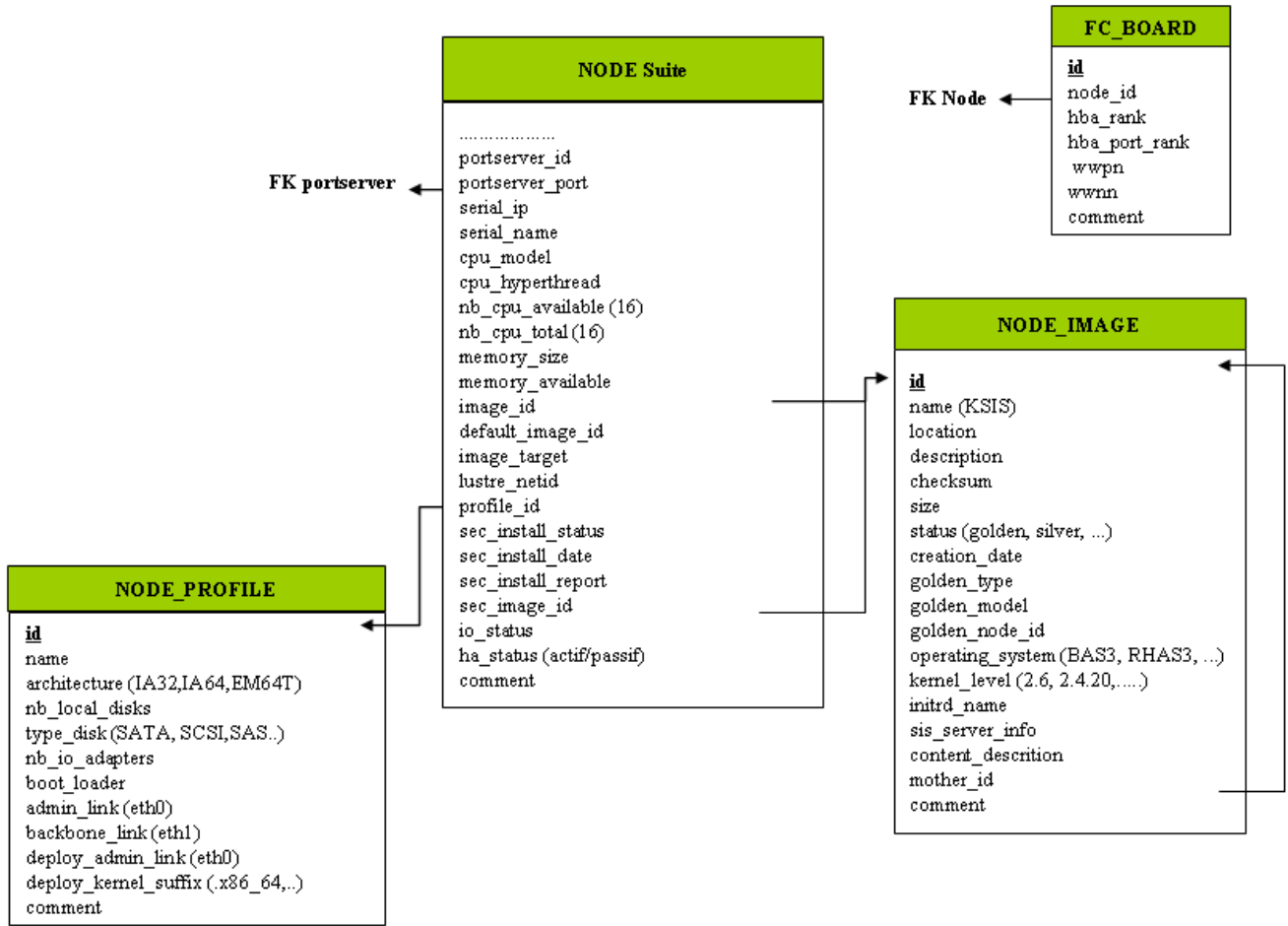


Figure 3-6. Cluster Database – Machine view 2

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 | | deprecated |
| cluster_id | FK on the CLUSTER | | preload |
| hwmanager_id | FK on the HWMANAGER | | preload |
| domain_fame | Machine name PAP side | | deprecated |
| 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 address | NULL | |
| 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 | | | nodeRecord or equipmentRecord |
| 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_hyperthread | 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 |
| post_config_status | PostConfig Status | | KSiS |
| comment | Free field | NULL | |

Table 3-28. 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 | BAS5V2 | 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 3-29. 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 | SERV-A | 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 3-30. 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 |
| rail | Number of rail | 0 | preload |
| port_rank | Port number on the card | 1 | preload |
| state | Status of the port (connected, not_connected) | connected | 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 3-31. 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 3-32. 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 | Destination in sdplib.conf | | updateSDPoIB |
| listen | Listen in sdplib.conf | | updateSDPoIB |
| program | Program in sdplib.conf | | updateSDPoIB |

Table 3-33. 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 3-34. Machine view – FC_BOARD table

3.5.4 HWMANAGER View

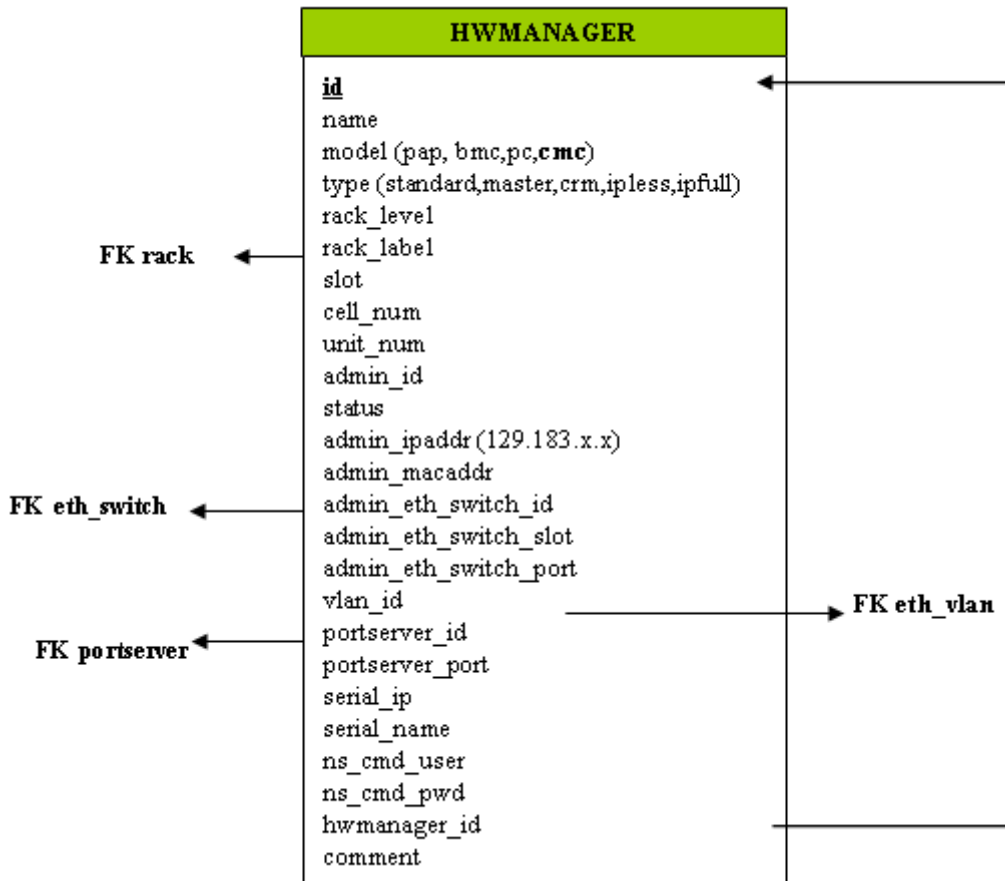


Figure 3-7. HWManager view

3.5.4.1 HWMANAGER Table

| Column name | Description | Example | Fill in method |
|-----------------------|---------------------------------------|---------------|--------------------|
| id | Primary key | | preload - Sequence |
| name | HWMANAGER IP name | papu1c2 | preload |
| model | PAP or BMC or PC or CMC | 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 | | updateMacAddr |
| admin_eth_switch_id | ETH_SWITCH id | | preload |
| admin_eth_switch_slot | Arrival slot number on ETH SW | | preload |
| 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 | HWMANAGER serial name | papu1c2s | preload |
| ns_cmd_user | User NC Commande | nsc | preload |
| ns_cmd_pwd | password | \$nsc | preload |
| hwmanager_id | FK on HWMANAGER | | preload |
| comment | Free field | | |

Table 3-35. HWMANAGER Table

3.5.5 Complementary Tables

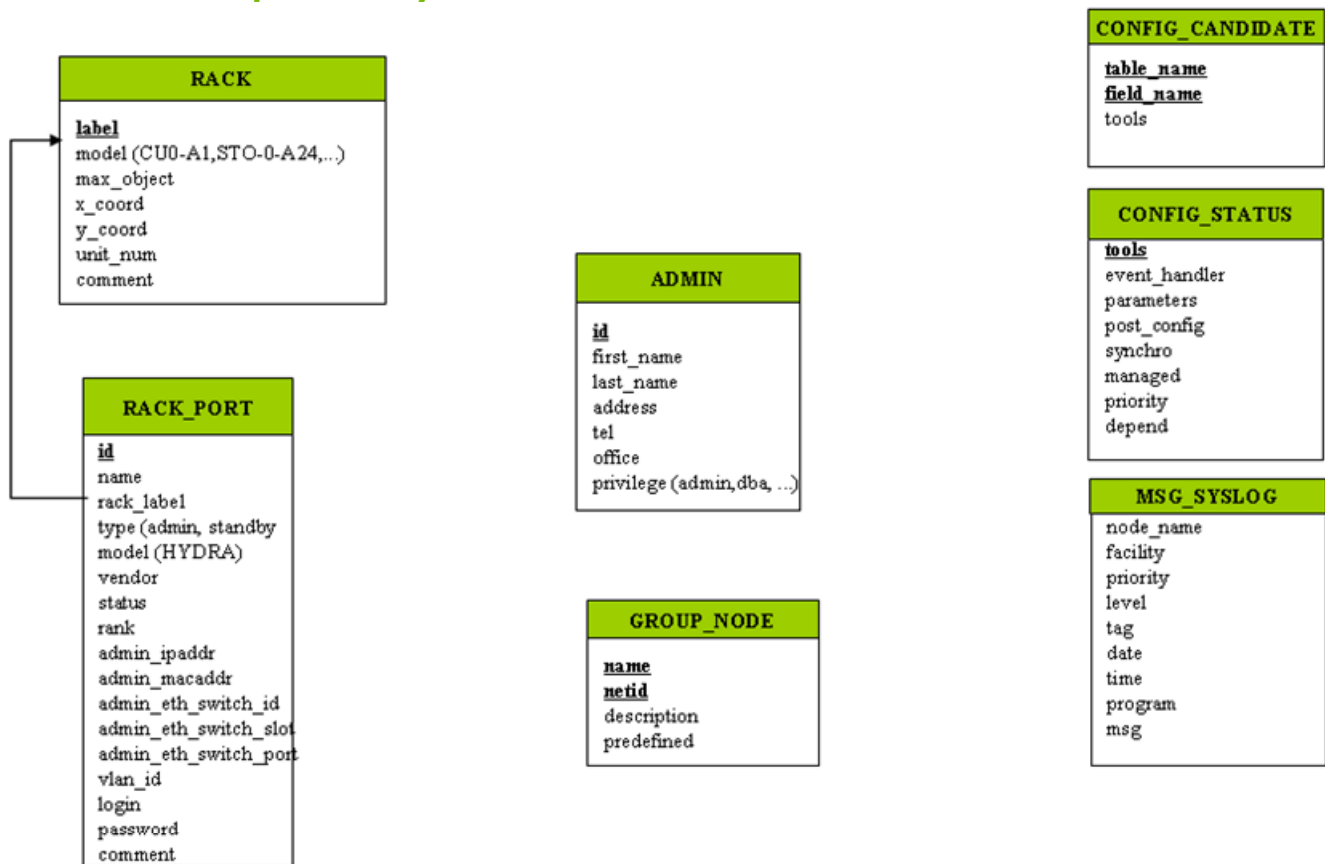


Figure 3-8. Cluster Database – Complementary tables

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, etc. | | admin |

Table 3-36. 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 3-37. RACK table

3.5.5.3 RACK_PORT Table

| Column name | Description | Example | Fill in method |
|-----------------------|-------------------------------|----------------|----------------------|
| id | PK | | Sequence |
| label | FK on RACK | RACK1 | preload |
| type | Port yype | Admin | preload |
| model | Card model | HYDRA | preload |
| vendor | Vendor name | Bull | preload |
| status | Nagios host status | Up | DV = up - Nagios |
| rank | Port instance | 1 | preload |
| admin_ipaddr | Admin IP address | 172.16.118.190 | preload |
| admin_macaddr | Port Mac address | | equipmentRecord |
| admin_eth_switch_id | FK on ETH_SWITCH | | preload |
| admin_eth_switch_slot | Arrival slot number on ETH_SW | | preload |
| admin_eth_switch_port | Connexion port on ETH SW | 3 | preload |
| vlan_id | FK on ETH_VLAN | | preload |
| login | Administration login | | DV = super - preload |
| password | Administration password | | DV = pass - preload |
| comment | Free field | NULL | |

Table 3-38. RACK_PORT table

3.5.5.4 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 3-39. CONFIG_CANDIDATE table

3.5.5.5 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 3-40. CONFIG_STATUS table

3.5.5.6 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 3-41

GROUP_NODE table

3.5.5.7 MSG_SYSLOG Table

This table is not active in this version.

3.5.6 Nagios View

| SERVICES | AVAILABILITY |
|---|--|
| id name (temperature, rms, ..) object (node, eth_switch,) active (true/false) comment | id name (cpu, temp, var,..) enable (true/false) perf_indicator max_threshold min_threshold group_name rule comment |

Figure 3-9. Nagios View

3.5.6.1 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 3-42. SERVICES table

3.5.6.2 AVAILABILITY Table

| Column name | Description | Example | Fill in method |
|----------------|-------------------------|---------|----------------|
| id | Service id | | |
| name | CPU, temp, var | cpu | |
| enable | To check (true / false) | true | |
| perf_indicator | Performance indicator | true | |

| | | | |
|---------------|-------------------|--|--|
| max_threshold | Maximum threshold | | |
| min_threshold | Minimum threshold | | |
| group_name | Application group | | |
| rule | Criterion rule | | |
| comment | comment | | |

Table 3-43. AVAILABILITY table

3.5.7 Lustre View

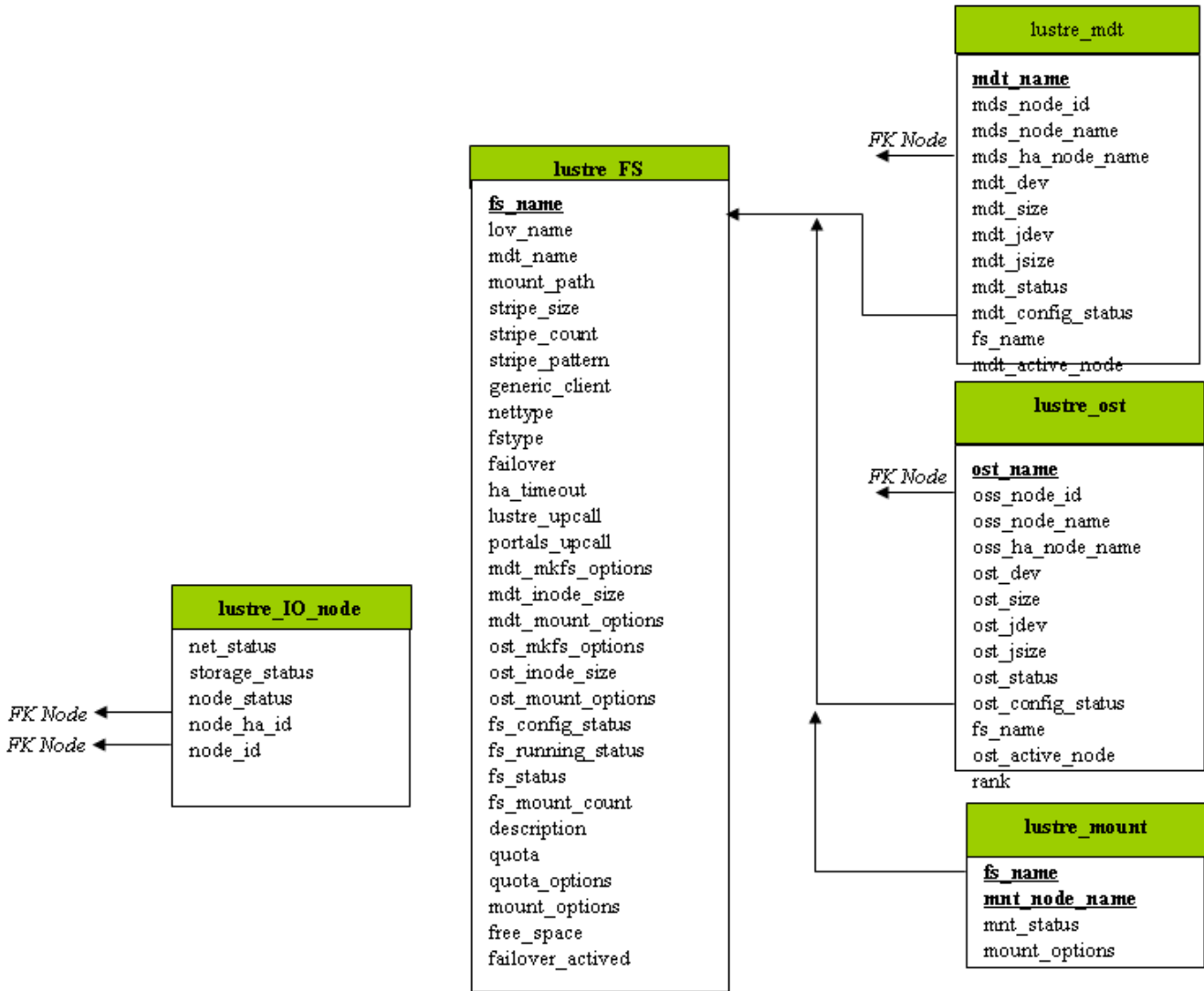


Figure 3-10. Cluster Database – Lustre view

3.5.7.1 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 |
|-------------|------------------------|-------------------|----------------|
| fs_name | File system name | lustre_basic | lustre_config |
| mount_path | File system mount path | /mnt/lustre_basic | lustre_config |
| lov_name | LOV identification | lov_lustre_basic | lustre_config |

| | | | |
|--------------------------------------|---|------------------------|-------------------------|
| mdt_name | MDT reference | mdt_ns44_1 | lustre_config |
| stripe_size | Stripe size | 4MB | lustre_config |
| stripe_count | Number of stripe per file | 0 (all included OSTs) | lustre_config |
| 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 (Deprecated - do not use) | 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 |
| free_space | Size of the file system in GB | 773 | lustre_util |
| failover_activated | For future use | yes | lustre_util |

Table 3-44. Lustre_FS table

3.5.7.2 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 |
| ost_active_node | In case of HA management, current node name support | ns32 | lustre_migrate |

| Column name | Description | Example | Fill in method |
|-------------------|----------------------------------|---------------|-------------------------|
| 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 3-45. Lustre OST table

3.5.7.3 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 3-46. Lustre_MDT table

3.5.7.4 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 3-47. Lustre_IO_node table

3.5.7.5 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 3-48. Lustre_mount table

Chapter 4. Software Deployment (KSIS)

This chapter describes how to use KSIS to deploy, manage, modify and check software images. The following topics are described:

- 4.1 *Overview*
- 4.2 *Configuring and Verifying a Reference Node*
- 4.3 *Main Steps for Deployment*
- 4.4 *Checking Deployed Images*
- 4.5 *Ksis Commands*
- 4.6 *Building a Patch*
- 4.7 *Checking Images*
- 4.8 *Importing and Exporting an Image*
- 4.9 *Rebuilding ClusterDB Data before Deploying an Image*

4.1 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 Extreme Computing systems.

KSIS makes it easy, for a network of Linux machines, to propagate software distributions, content or data distribution changes, operating system and software updates.

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.

The deployment is done using the administration network.

Note The terms **reference node** and **golden node** are interchangeable. The same applies to the terms **reference image** and **golden image**.

4.2 Configuring and Verifying a Reference Node

A reference node is a node which has had all the software installed on to it, and whose image is taken and then stored on the image server. The reference image will be deployed onto the other nodes of the cluster.

Installation and Configuration

Reference nodes have the **bullx cluster suite** software installed on to them in the same way as ordinary COMPUTE/COMPUTEX or I/O nodes. A **KSIS client** is then installed on to these nodes from the bullx cluster suite media. The operating system and applications must be installed and configured to make the node operational.

4.3 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]
```

Note See *Deploying an Image or a Patch*, on page 4-9 for more details about the deployment process.

The following figure shows the creation and deployment of an image.

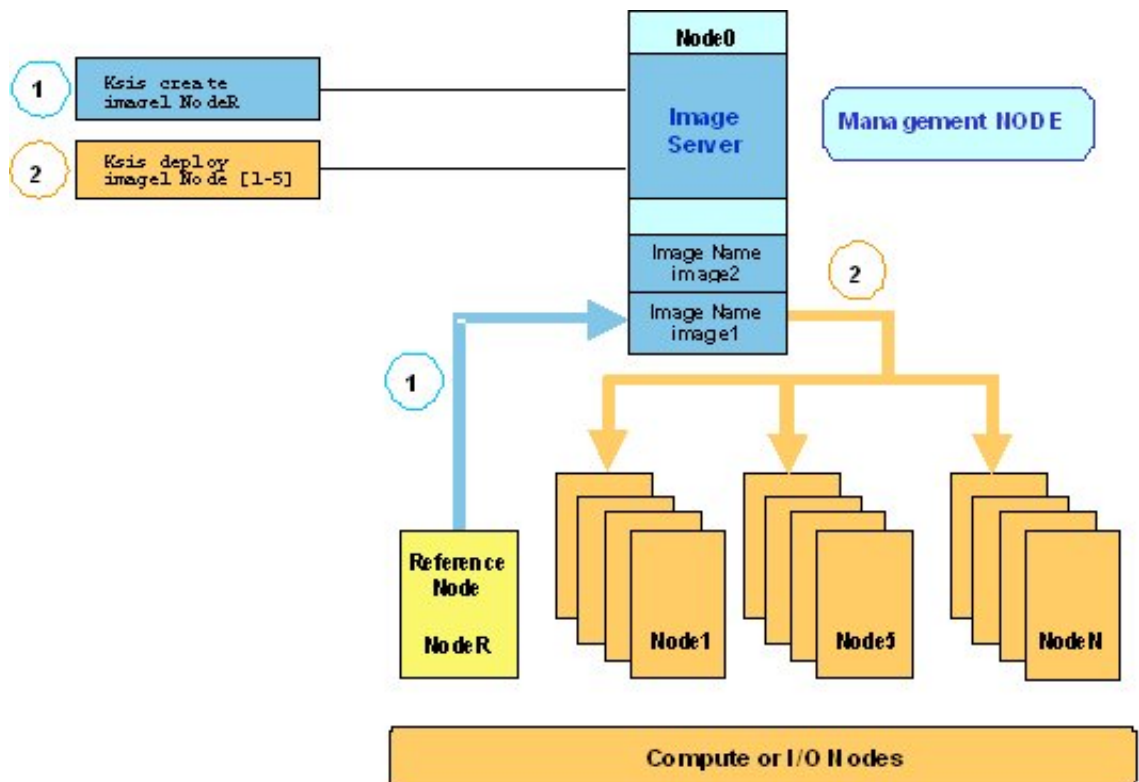


Figure 4-1. Main steps for deployment

4.4 Checking Deployed Images

The **ksis check** command is used to compare the files on a node(s) with the reference image deployed. This is done by listing the discrepancies for tests performed on the node(s), when compared with the results of the same tests on the reference image.

The general form for the **check** command is:

```
ksis check nodeRange [-t testGroup] [-d]
```

Options

- [-t testGroup]** Specify the **Test Group** for the checks. 3 Test Groups exist: **basic**, **basic+** and **sharp**.
- [-d]** View differences between the node(s) and the reference image, if they can be displayed on a few lines.

Examples

```
ksis check nc[2-45]
```

```
ksis check nc[2-45] -t basic -d
```

Note Nodes inside a node range are not always created from the same image.

4.4.1 Checking Principles

The descriptions of the image tests are stored in a database (`/etc/systemimager/ksis_check_Repository/` on the Management Node). When creating an image or a patch, the Administrator specifies the **Test Group** for an image or patch. **KSIS** then executes the commands associated with each check in the **Test Group**, 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 checks defined for the **Test Group** on each node and generates the results. If there is a discrepancy between the results and the *reference* results, the check is set to KO, otherwise it is set to OK. The image server centralizes the node check results.

4.4.2 Ksis Tests and Test Groups

| Test Name | Test Group | OK | KO |
|---------------------|----------------|--|--|
| CheckRpmList | Basic & Basic+ | 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. |
| CheckRpmFiles | Sharp | 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 seems to have been updated regarding contents and/or access rights. |
| CheckFastRpmFiles | Basic & Basic+ | None of the files delivered using RPM seems 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+ | None of the files of the deployed image seems 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 | None of the files of the deployed image seems 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 & Basic+ | 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 & Basic+ | 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 4-1. Standard checks delivered with Ksis

Each test belongs to one, or more, of the 3 **Test Groups** possible; **Basic**, **Basic+** and **Sharp**. Each **Test Group** includes a combination of the tests listed in Table 4-1, for example **Basic** level tests include the following tests: **CheckRpmList**, **CheckFastRpmFiles**, **CheckMandatoryFiles**, and **CheckUsedKernel**. **Sharp** includes all the **Basic**, and **Basic+** checks, in addition to its own checks.

If the `-t` option is not specified all the checks are executed.

4.4.3 Modifying the Checks Database

It is possible to modify the checks in the database, in order to adapt them to the way you use the image.

- To create a new **Test Name**, create a new directory (`/etc/systemimager/ksis_check_Repository/<testName>.vid`) which includes at least the following:
 - **command** file, which contains the command(s) to be run for the test.
 - **Test group**, which defines the Test groups for the test.

The new **Test Name** will be included in the checks database, and will be part of the checks performed on subsequent images.

4.4.4 Examining the Check Results

The result of the checks carried out includes a comparison between the results of the command(s) executed on the reference image and on the node(s) specified. This comparison shows the evolution of the node(s) against the reference image, and can be used to determine the necessity of deploying the reference image again.

If the discrepancies between a node and the reference image are not significant, it may still be useful to analyze their development using the **checkdiff** command.

ksis checkdiff command

The **ksis checkdiff** command displays the discrepancies between the reference image and the results for a given test on a node. The general form for the **checkdiff** command is:

ksis checkdiff testName nodeRange

Example

```
ksis checkdiff CheckSRDir node2
```

4.5 Ksis Commands

4.5.1 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>
```

4.5.2 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.

When the **-d** option is used, if declared in the **/etc/fstab** file, the disk(s) specified (e.g. **/dev/sdj**) will be remounted on all the newly 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.

4.5.3 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 <imagename> <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 4.4 *Checking Deployed Images* for other options).

4.5.4 Deleting an Image or a Patch

This command deletes the defined image or patch from the image server (Management Node).

```
ksis delete <imageNameOrPatchName>
```

4.5.5 Deploying an Image or a Patch

This command deploys an image or a patch on the nodes specified:

```
ksis deploy <imageNameOrPatchName> <nodeRangeOrGroupName> [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

-
- See**
- *Installation and Configuration Guide* for details on the deployment procedure, including post deployment operations.
 - *Maintenance Guide* for more details on the **Ksis** log files.
-

4.5.6 Removing a Patch

This action concerns only the images with the 'patch' status. It removes the last patch deployed from the nodes.

```
ksis undeploy <patchName> <nodeRangeOrGroupName> [options]
```

4.5.7 Getting Information about an Image or a Node

This command displays information for the specified image or node.

```
ksis show <imageNameOrNodeName>
```

4.5.8 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>]
```

- | | |
|-----------------------|--|
| golden | reference image (from a reference node - also called golden node). |
| patch | patch (result of a store command). |
| patched golden | modified reference image (result of a detach command). |
| working patch | 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_u1u2 | golden | 2005-01-14 15:41:25 |
| Compute_hpceth_u1u2.s1.0 | patch | 2005-01-20 13:49:27 |
| Compute_hpceth_u1u2.s1.1 | working patch | 2005-01-22 14:41:03 |

4.5.9 Listing Images by Nodes

This command lists the current images available and their status on the nodes.

```
ksis nodelist [<options>]
```

Example:

```
ksis nodelist
```

```
nc1  unreach  -
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
```

4.6 Building a Patch

ksis buildpatch is used to create a patch from the differences between two images. This can then be used to transform the software structure and content of the first node which has had the first image deployed on it so that it matches a node which has had the second image deployed on it.

Note TextNote **ksis buildpatch** can only be used 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>**.



important **ksis buildpatch** and the use of patches should only be applied for limited image changes. For fundamental image changes the best method remains the creation and the deployment of a new image.

4.7 Checking Images



Important The **KSIS tests** provided must be considered as templates. The tests used by the System Administrator must be adapted so that they are coherent with the cluster node architecture, otherwise they will not work correctly.

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>
```

4.8 Importing and Exporting an Image

KSIS provides a function to export an image to another **KSIS** installation (on another Management 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> [<options>]
```

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**.

The **import** command imports an image previously exported from another cluster.

```
ksis import <imageName> [<options>]
```

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.

4.9 Rebuilding ClusterDB Data before Deploying an Image

There are two cases where it may be necessary to update the reference information before deploying an image:

- Some values have changed in the ClusterDB
- An image has been imported so that its ClusterDB information must be updated.

To do so, use the **builddatanode** command, which updates the images with the latest values in the ClusterDB:

```
ksis builddatanode
```

Nodes context will be updated to take in account new data from DB

Continue (yes/no)

Answer **yes** to the question.

Chapter 5. Kerberos - Network Authentication Protocol

Kerberos is an optional security suite product that can be used to authenticate users, services and machines for a whole network. Kerberos is included within the Linux delivery.

The purpose of this chapter is to describe how to implement Kerberos on a Bull Extreme Computing cluster.

5.1 Environment

5.1.1 Kerberos Infrastructure

There are 3 types of machine within the **Kerberos** infrastructure:

- The **Kerberos** server that includes the Key Distribution Centre (**KDC**) server and administration server, housed on a server called **secu0**. By default, this will be part of the Management Node.
- A set of application servers (SSH, NFS, etc.) which are protected by Kerberos; these servers are named **secui**. The Kerberos configuration file for these servers is shared with the Kerberos server.
- The **Kerberos** client machines. These are not used until Kerberos authenticates the users' rights to access the applications on **secui**.

5.1.2 Authentication of the SSHv2 Connections

The remote **SSH** service (OpenSSH) will be activated on **secu1** with Kerberos support. A remote connection to **secu0** will then be made using Kerberos tickets instead of internal authentication mechanisms.

5.2 KERBEROS Infrastructure Configuration

5.2.1 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 `-without-krb4` option to prevent compatibility with Kerberos 4.

5.2.2 Configuration Files

`/etc/krb5.conf`

This file contains the details of the **KDC** addresses and the administration server, and will be copied on to all the servers containing kerberized applications, as well as on to all the client machines.

```
...
[libdefaults]
  default_realm = DOMAIN.COM
  default_tgs_etypes = des3-hmac-sha1 des-cbc-crc des-cbc-md5
  default_tkt_etypes = des3-hmac-sha1 des-cbc-crc des-cbc-md5
  permitted_etypes = des3-hmac-sha1 des-cbc-crc des-cbc-md5
  forwardable = true
...

[realms]
  DOMAIN.COM = { kdc=secu0:88
                  admin_server = secu0:749
                  default.domain = domain.com
                }

[domain.realm]
  .domain.com = DOMAIN.COM
  domain.com = DOMAIN.COM
  localhost = DOMAIN.COM
...

[login]
  krb4_convert = false
  krb4_get_tickets = false
...
```

`/var/kerberos/krb5kdc/kdc.conf`

This file, containing among other things the information necessary to produce the tickets, is specific to the Kerberos server.

```
...
[realms]
DOMAIN.COM={
  preauth=yes
  admin_keytab = FILE:/etc/krb5.keytab
  max_life = 2d 0h 0m 0s
  max_renewable_life = 10d 0h 0m 0s
  ...
}
```

5.2.3 Creating the Kerberos Database

Use the following command to initialize the Kerberos database.

```
/usr/kerberos/sbin/kdb5_util create -s
enter KDC database master key : XXXX
```

5.2.4 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 initially exist, it is possible to connect to the KDC server the first time as root using the **kadmin.local** command 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 krb5adm/admin
Enter password for principal "krb5adm/admin@DOMAIN.COM": YYYY
```

Now it should be possible to authenticate oneself as **krb5adm** from any Kerberos client machine. The **Unix** system account **krb5adm** must have been created, as shown above, in order to connect to the administrator server and to manage Kerberos, assuming the admin daemon has been launched. See below for more details.



- For security reasons remote administration using **kadmin** is deactivated. To enable it add the **kadmin/admin** and **kadmin/changepw** special entries in the keytabs. However, this setting is not recommended for a cluster environment.
 - The Kerberos administrators which have been created – **krb5adm** in the example above – must belong to the root group in order to have access to, and to be able to modify, Kerberos files.
-

5.2.5 Starting the KDC Server

Use the following command to start the **KDC** server:

```
/sbin/service krb5kdc start
```

Verify the local connection to **Kerberos** on the KDC server using the **krb5adm** administrator access rights:

```
/usr/kerberos/bin/kinit krb5adm/admin
```

```
kinit(V5) : Cannot resolve network address for KDC in requested realm  
while getting initial credentials
```

The problem in the above message is that **krb5adm**'s credentials cannot be confirmed and will only be resolved when **secu0** is replaced by its IP address in the **krb5.conf** file.

```
/usr/kerberos/bin/kinit krb5adm/admin  
  
Password for krb5adm@DOMAIN.COM: YYYY
```

If there is no error message then everything is OK and the **krb5adm** administrator will obtain a Ticket-Granting Ticket (TGT).

5.2.6 Adding Access Control List (ACL) Rights for the Kerberos Administrator Created

In the **/var/kerberos/krb5kdc/kadm5.acl** file, add the line:

```
krb5adm/admin @DOMAIN.COM *
```

5.2.7 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 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 has to be done from the Kerberos server using administrator access rights for **krb5adm/admin**.

5.2.8 Creating Principals Associated with Users

The Kerberos Administrator will create the principals associated with users on the **KDC** server. These users must have associated UNIX accounts on the client machines.

The Kerberos Administrator can create the principals locally on the **KDC** (using the command **kadmin.local**) without needing to authenticate himself. 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
```

The secret key shared between the **KDC** and the client machine for a user principal is derived from the user's password.

The process has to be repeated for all other users.

5.2.9 Creating Principals Associated with Remote Kerberized Services

The principals associated with services have to be created. The **Linux** distribution includes some services that have already been kerberized. The principal associated with **FTP**, **TELNET**, and **RSH** services, included as part of the default installation using the **krb5-workstation** package, is called **host principal**.

The **host principal** name is derived from the name of the machine, and this is used for **Kerberos** Authentication of the basic kerberized services - **RLOGIN**, **TELNET**, etc. residing on the host.

Creation of the host principal for the **secu1** server

Connect to **Kerberos secu0** server and then create the host principal with the **kadmin** command.

```
kadmin.local
addprinc -randkey host/secu1.domain.com
```



Important The hostname has to be the same as in its first appearance in the line associated with the machine in the **/etc/hosts** file.

5.3 Configuring the secu1 Machine that hosts the Host Principal remote service

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 part of the system.

5.3.1 Generating the key associated with the Host Principal remote service

This secret key is shared between the KDC **secu0** server and the server housing the **secu1** remote service. This is essential in order that **secu1** can decipher the **Kerberos** tickets which are transmitted to it. The key can be created on any one of these 2 servers but must then be copied from one to the other.

 **Important** The default file for the keys is as follows:

```
/var/kerberos/krb5kdc/kadm5.keytab
```


Therefore, the file for the keys used by the command `kadmin` is defined in the `realms` section in the `kdc.conf` file:

```
/etc/krb5.keytab
```

Connect as the Kerberos administrator (`krb5adm`) to **secu0**:

```
kadmin
ktadd -k /path/to/krb5.keytab.sec1 host/secu1.domain.com
```

Then recopy the `/path/to/krb5.keytab.sec1` key to **secu1** in the `/etc/krb5.keytab` file.

 **Important** It is recommended to have a `keytab` file for each service, and to store only the keys associated with the remote services that each server hosts, and not the keys for services that are hosted by other servers. However, the KDC server must have its own specific `keytab` file for all the remote service keys.

5.4 Kerberos Authentication and SSH

The SSH remote service is installed on **secu1** with a SSH client connection from **secu0**.

Before using any Kerberized client, such as SSH, you have to request the TGT ticket. In the following example, this request is done for the user connected as *Durand* on **secu0**:

```
kinit
PW : xxxx (password user durand)
klist
....
```

5.4.1 Configuring the SSH Server on the secu1 machine

A typical `sshd_config` configuration file will contain the following:

```
Port 22
Protocol 2
ListenAddress xxx.xxx.xxx.xxx

RSAAuthentication no
PubkeyAuthentication no

RhostsRSAAuthentication no
HostbasedAuthentication no

PasswordAuthentication no
PermitEmptyPasswords no

# 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

UsePAM yes

Subsystem sftp /usr/local/libexec/sftp-server
```

Pre-requisites for the configuration of SSH server

- The `/etc/hosts` file of the remote machine that SSH is connecting to has to have its hostname in the form:
`x.x.x.x secu1.domain.com secu1`
- The hostname of the remote machine may be of the form:
`secu1.domain.com` or `secu1.`
- The principal service associated with this machine has to be the same as its Fully Qualified Domain Name **FQDN**:
`secu1.domain.com.`

5.4.2 SSH Client

On the **secu0** machine, or other machines, a typical **ssh_config** file will appear as follows:

```
RhostsRSAAuthentication no
RSAAuthentication no
PasswordAuthentication no
HostbasedAuthentication no
Port 22
Protocol 2

GSSAPIAuthentication yes

# For tickets forwarding:
GSSAPIDelegateCredentials yes
```

Note TGT ticket forwarding by **SSH** is activated by the **GSSAPIDelegateCredentials yes** parameter in the **SSH** client file.

5.5 Troubleshooting Errors

```
Error : " Permission denied (gssapi-with-mic,keyboard-interactive) "
```

There are various possible causes for this error. Check the following:

1. The target machine has its **full name** in its **/etc/hosts** file as shown below:

```
@IP secu1.domain.com secu1
```

2. If several names are associated with the same IP address, the name used for the connection has to be at the top of **/etc/hosts** file, as shown below:

```
@IP parallel.domain.com parallel
@IP secu1.domain.com secu1
```

3. Check that the **/etc/krb5.conf** file on the **KDC** server and on the **SSH** servers\clients is identical.
4. Check that the keys in the **/etc/krb5.keytab** file are identical on the **KDC** server and on the **SSH** server.
5. Verify that the user has a valid TGT ticket.

5.6 Generating Associated Keys for Nodes of a Cluster

The Perl program, below, generates the **Kerberos** key (keytab) for each node on the **Kerberos** server (hosted on the Management Node), and then transfers the key to the node using Secure Copy (**SCP**), which ensures confidentiality and authentication using a private key/public key.

The pre-requisite here is that the private key / private key infrastructure is in place between the Management Node and each Compute Node.

```
#!/usr/bin/perl -w

print "Lower limit of cluster nodes: ";
$inf = <STDIN>;
chomp ($inf);

print "Upper limit of cluster nodes: ";
$sup = <STDIN>;
chomp ($sup);

# Define constants
#
my $serv = "secu";
my $domain = "domain.com";
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/private keys.
for ($i=$inf; $i <=$sup; $i++) {
    $serv="$serv0$i";
    print("Generate keytab for host : $serv\n");
    system ("rm -f /tmp/$serv$keytab");
    system ("kadmin.local -q 'ktadd -k /tmp/$serv$keytab
    host/$serv.$domain'");
    system ("scp -rp /tmp/$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");
```

5.7 Modifying the Lifespan and Renewal Period for TGT Tickets

The default duration for a Ticket-Granting Ticket (**TGT**) ticket is 10 hours, and this can be renewed while it is still active. In other words its duration must be greater than 0 to be renewed.

The ticket duration and renewal period 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 using this command will be valid for 2 days and it may be renewed at any time during these 2 days to obtain a new ticket which is also valid for 2 days up until the 5 day limit is reached.

The values specified by the user have to be inside the maximum values defined by the Kerberos configuration. To modify the values in the Kerberos `/var/kerberos/krb5kdc/kdc.conf` configuration file do the following:

In the `[realms]` block, add:

```
max_life = 2d
max_renewable_life = 10d
```

Then relaunch the `krb5kdc` and `kadmin` daemons.

5.8 Including Addresses with Tickets

By default tickets do not include addresses.

Use the command below so that the tickets generated include the addresses of the local machine.

```
add noaddresses=no in the paragraph [libdefaults] for the file
/etc/krb5.conf
```

Chapter 6. Storage Device Management

Bull cluster management tools provide services to manage storage systems and a large amount of storage resources. This chapter explains how to setup the management environment, and how to use storage management services.

The following topics are described:

- 6.1 *Overview of Storage Device Management for Bull Extreme Computing clusters*
- 6.2 *Monitoring Node I/O Status*
- 6.3 *Monitoring Storage Devices*
- 6.4 *Monitoring Brocade Switch Status*
- 6.5 *Managing Storage Devices with Bull CLI*
- 6.6 *Using Management Tools*
- 6.7 *Configuring Storage Devices*
- 6.8 *User Rights and Security Levels for the Storage Commands*

6.1 Overview of Storage Device Management for Bull Extreme Computing clusters

Bull Extreme Computing clusters can 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 Extreme Computing 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, and includes:

- Built-in LAN management ports for the storage devices that are connected to the cluster management network.
- Built-in serial ports for the storage devices that are connected to the cluster management network, using terminal servers.

- Management stations or proxy servers (for example Windows stations) hosting device management tools that 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. The storage device status can be monitored using **Bull System Manager – HPC Edition** and by querying the cluster DB with the **storstat** command. These services enable the browsing via a global view covering all the storage devices, and 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 storage system family (for example **nec_admin**, etc.). They enable configuration and status information to be read, 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 included with the storage system. The most useful information and operations are available via these commands. Nevertheless, they do not include 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 who is trained to manage storage devices can also access the management tools for each storage device. The serial ports can be used with **conman** (or telnet). The Ethernet ports can be connected to via 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 configurations. 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 following sections explain how to setup and use this environment.

6.2 Monitoring Node I/O Status

Each node is monitored and any 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 Node using dedicated **syslog** messages.

The I/O status of each node can be verified by displaying the **I/O status** service of the node via **Bull System Manager – HPC Edition**.

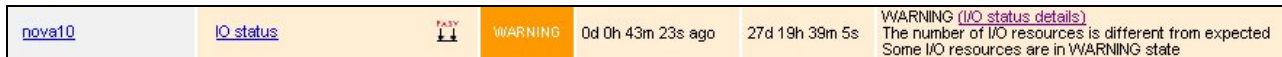


Figure 6-1. I/O Status – initial screen

The semantic of the service is as follows:

| | |
|-----------------|---|
| OK | No problem detected |
| WARNING | An I/O component in WARNING 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. |
| CRITICAL | 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 syslog A fatal HBA error has been reported by a device driver in syslog A link down transition has been notified by a device driver A LUN resource cannot be acceded by a multipath pseudo-device. A device referenced by the persistent binding mechanism (alias) is missing. |
| UNKNOWN | Cannot access the status |
| PENDING | Not yet initialized |

6.2.1 Managing I/O Reference Counters

The I/O status transmitted by each node to the Management Node is a synthesis of the status detected for each I/O resource attached to the node and of the comparison between the I/O counters and their reference values.

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 device classes (named 'I/O counters').

Two reference counters (**nb_io_adapters** and **nb_local_disks**) are stored on the Management Node in cluster DB in the **node_profile** table. 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 for each node by the **I/O status details** link of the **I/O status** service on the node via **Bull System Manager – HPC Edition**.

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/dm*) for LUNs or pseudo-devices | nb_aliases_ok | 8 / 8 |

(Counter Value = Current / Expected)

Figure 6-2. Bull System Manager HPC Edition - I/O Status Details

The **iorefmgmt** command is used to manage I/O device monitoring reference counters.

To obtain 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 using the command:

```
iorefmgmt -c adjust
```

The **nb_io_adapters** and **nb_local_disks** counters cannot be adjusted on a node.

You can manage these counters in the cluster DB node profile table on the Management Node 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 Node, using **ssh** or **pdsh**.

6.2.2 I/O Counters Definitions

nb_io_adapters 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 Number of I/O adapters expected to be configured (driver loaded).

nb_local_disks 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 | Number of physical disks expected to be healthy. |
| nb_io_ports | Expected number of Fibre Channel ports. |
| nb_io_ports_connected | Number of Fibre Channel ports expected to be connected. |
| nb_fixed_luns | 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 | Number of LUNs which are not reconfigurable that are expected to be accessible. |
| nb_reconf_luns | Expected number of reconfigurable LUNs. |
| nb_reconf_luns_ok | 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 | Expected number of paths involved in multi-path to reach LUNs which are reconfigurable. |
| nb_iopaths_ok | Number of paths involved in multipath expected to be alive. |
| nb_aliases | 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 | Number of aliases on Fibre Channel block devices expected to be correctly mapped. |
| nb_pseudos | Expected number of multipath pseudo-devices on a node. |
| nb_pseudos_ok | Number of multipath pseudo-devices expected to be usable. |

6.2.3 Managing I/O Resources

The I/O resources identified for each node are monitored and their status stored on the node. The I/O resources may be displayed in **Bull System Manager – HPC Edition** by looking at the **I/O status** service associated with a node; this is done by clicking on the **I/O status details** link.

| I/O Resources of node : nova5 | | | |
|---|---|--|---|
| Adapter 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 |
| Adapter 24:01 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 6-3. Bull System Manager –HPC Edition – I/O Resources of a node

The list of I/O resources, with their associated status, for each node can also be consulted by using the following command:

```
lsiodev -l
```

On the Management Mode, the equivalent information can be obtained remotely from the nodes by using the following command:

```
iorefmgmt -r <node> -L
```

The current status for each I/O resource is updated automatically by the I/O monitoring mechanism. However, it may be necessary to update the status of a resource manually, for example, to distinguish transient from permanent I/O errors. The status of an I/O resource can be changed remotely from the Management Node by using the following command:

```
iorefmgmt -r <node> -U -t <resource-type> -n <resource-name> -s <new-status> -m "<associated-comment>"
```

Using this command will result in the global I/O status of the node being recomputed and its I/O status service in Bull System Manager – HPC Edition being updated.

6.3 Monitoring Storage Devices

This section explains how the Administrator can monitor and obtain 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 (Bull System Manager – 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 Database information using the **storcheck** command:

```
storcheck -c <cluster_name>
```

This command will check all the storage systems for a cluster. It is possible to reduce the scope to a single storage system:

```
storcheck -c <cluster_name> -n <disk_array_name>
```

6.3.1 Bull System Manager - HPC Edition: Host and Service Monitoring for Storage Devices

Storage device monitoring is integrated in the global monitoring for a cluster. Each storage system is identified by a host and its associated service, regardless of the number of controllers and Ethernet ports.

Bull System Manager - HPC Edition continuously updates the host status and service status values, without any intervention from the Administrator. All **Bull System Manager - HPC Edition** features and services apply to storage devices. Nevertheless, the Administrator using **Bull System Manager - HPC Edition** must be aware of the specific points that are explained next.

| Host ↑ | Service ↑ | Status ↑ | Last Check ↑ | Duration ↑ | Attempt ↑ | Status Information |
|--------|---------------|----------|---------------------|----------------|-----------|--|
| dn1 | 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 6-4. 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 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 6.3.2 *Bull System Manager - HPC Edition: Storage & I/O Information*).

6.3.1.1 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. |

6.3.1.2 Service Semantics

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).

| | |
|----------------|---|
| OK | <p>No problem</p> <p>Criteria:</p> <ul style="list-style-type: none"> • No disk errors • All referenced disks are present |
| WARNING | <p>Maintenance operation must be scheduled</p> <p>Criteria:</p> <ul style="list-style-type: none"> • Some disk failures, and / or removed referenced disks • Does not meet the criteria for critical status. |

| | |
|-----------------|---|
| CRITICAL | Degraded service. Maintenance operation mandatory Criteria: <ul style="list-style-type: none"> The number of faulty / missing disks is higher than the number of spare disks. |
| UNKNOWN | Cannot access the status |
| PENDING | Not yet initialized |

Note The cluster database has been initialized with a detailed status including all disk slots, populated and empty. The Administrator, who decides to permanently remove some HDDs, must manually update the database reference configuration (using the **storstat -u** command). Otherwise, these empty slots 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.

| | |
|-----------------|--|
| OK | No problem Criteria: <ul style="list-style-type: none"> All power supplies and fans are OK All reference power supplies and fans are present |
| WARNING | Maintenance operation must be scheduled Criteria: <ul style="list-style-type: none"> Some power supplies and/or fans are in the warning or critical state Does not meet the criteria for critical status. |
| CRITICAL | Degraded service. Maintenance operation mandatory Criteria: <ul style="list-style-type: none"> 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). |
| UNKNOWN | Cannot access the status |
| PENDING | Not yet initialized |

Temperature Service

Describes the global status for temperature sensors.

| | |
|----------------|--|
| OK | No problem Criteria: <ul style="list-style-type: none"> All temperature sensors are OK |
| WARNING | Maintenance operation must be scheduled Criteria: <ul style="list-style-type: none"> Some temperature sensors are not in the OK state Critical criteria not met |

| | |
|-----------------|--|
| CRITICAL | Degraded service. Maintenance operation mandatory Criteria: <ul style="list-style-type: none"> Some temperature sensors are in the critical state. |
| UNKNOWN | Cannot access the status |
| PENDING | Not yet initialized |

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.

| | |
|-----------------|--|
| OK | No problem Criteria: <ul style="list-style-type: none"> All controllers are OK |
| WARNING | Maintenance operation must be scheduled Criteria: <ul style="list-style-type: none"> Some controllers have a warning state and none are faulty (or missing). |
| CRITICAL | Degraded service. Maintenance operation mandatory Criteria: <ul style="list-style-type: none"> One controller or more is faulty (or missing). |
| UNKNOWN | Cannot access the status |
| PENDING | Not yet initialized |

Fibre Channel Port Service

This service shows the host connectivity status:

| | |
|-----------------|---|
| OK | No problem Criteria: <ul style="list-style-type: none"> All FC ports are OK. |
| WARNING | Maintenance operation must be scheduled Criteria: <ul style="list-style-type: none"> Not in critical status Some ports have a warning status |
| CRITICAL | Degraded service. Maintenance operation mandatory Criteria: <ul style="list-style-type: none"> One or more ports are in a critical status. |
| UNKNOWN | Cannot access the status |
| PENDING | Not yet initialized |

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 together all the problems for 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.

| | |
|----------|---|
| OK | No problem Criteria: <ul style="list-style-type: none">• Disk array semantic. |
| WARNING | Maintenance operation must be scheduled Criteria: <ul style="list-style-type: none">• Some of the other services are warning (but none critical).• The storage system has detected a warning which is not reported by one of the other services. |
| CRITICAL | Degraded service. Maintenance operation mandatory Criteria: <ul style="list-style-type: none">• One of the other services is critical.• The storage system has detected a critical error which is not reported by one of the other services. |
| UNKNOWN | Cannot access the status |
| PENDING | No yet initialized |

6.3.2 Bull System Manager - HPC Edition: Storage & I/O Information

Bull System Manager – 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 **Bull System Manager – HPC Edition** console – see Figure 6-5. A pop-up window appears containing a summary view of the storage systems and hardware component status – see Figure 6-6.

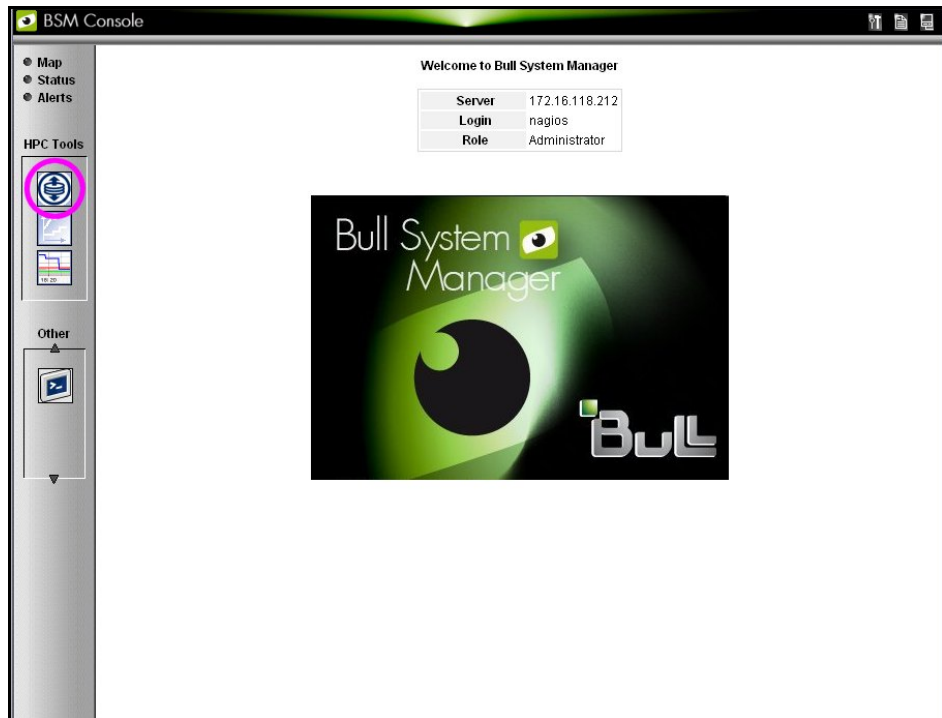


Figure 6-5. Bull System Manager opening console window with the Storage overview icon circled

6.3.2.1 Storage Views

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 **Bull System Manager** 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 provides 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:

| | |
|-----------|--|
| OK | No problem |
| ATTENTION | Maintenance operation must be scheduled, but the system still delivers the expected service. |
| FAILED | Degraded service. Maintenance operation mandatory. |

6.3.2.2

Storage Overview

This view offers a synthesis of the Storage devices monitored in the cluster.



Figure 6-6. 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 of storage systems or hardware components in the selected state – see Figure 6-7.

6.3.2.3

Inventory View of Storage Systems and Components requiring attention

This view - Figure 6-7 - displays the list of faulty components that should either be 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 for the component, within the device and within the cluster, its rack level and label.

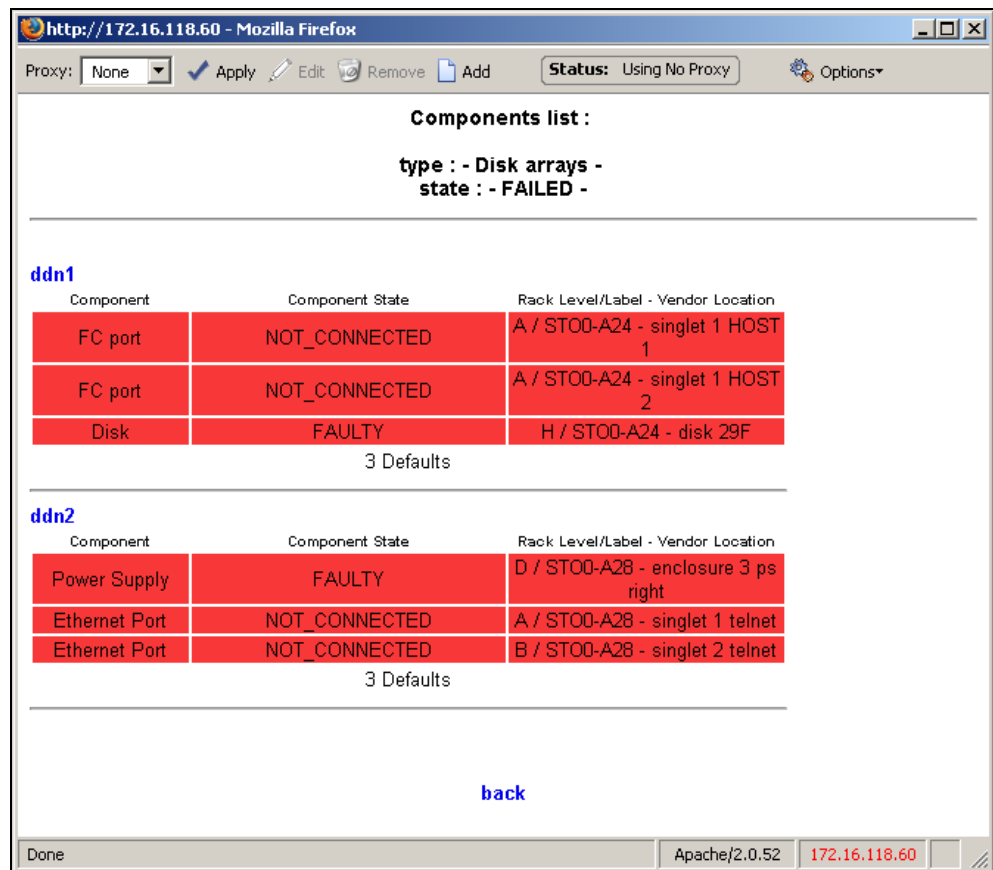


Figure 6-7. Inventory view of faulty storage systems and components

Note TextNote 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.

6.3.2.4 Detailed View of a Storage System

The Storage detailed view - Figure 6-8 - can be displayed by selecting a storage system in the Storage Summary Overview (see Figure 6-6).

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).

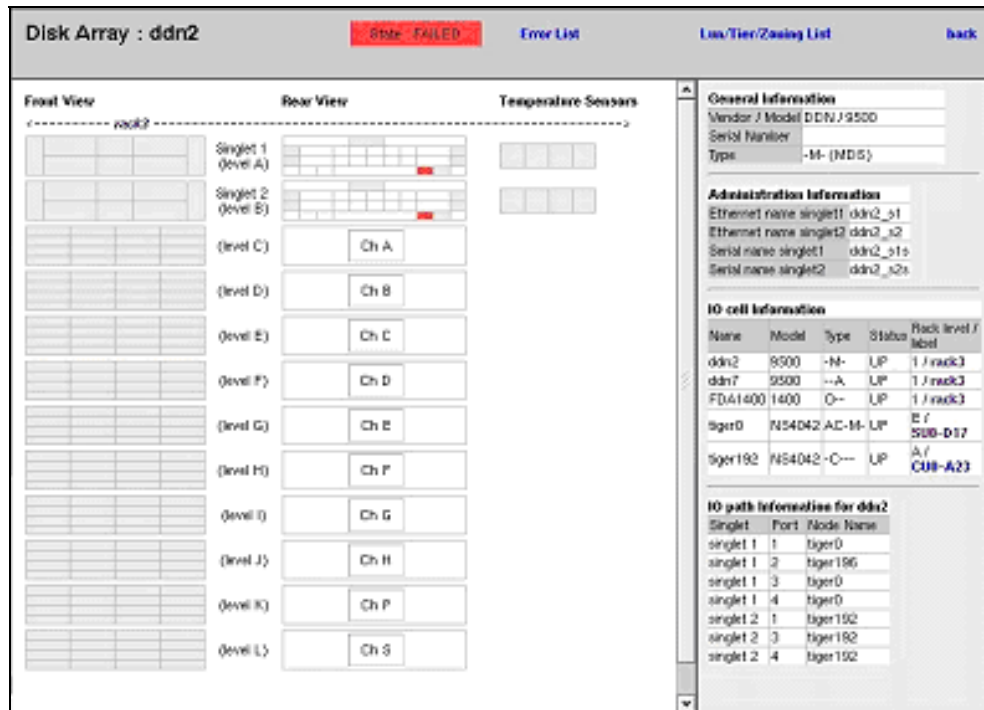


Figure 6-8. Storage detailed view

In the Storage Detailed view the item's description is shown by the use of mouse Tool tips.

6.3.2.5 Nodes I/O Overview

This view – Figure 6-9 – provides a synthesis of the I/O information for the cluster nodes. It shows I/O status statistics and allows the list of nodes to be filtered for 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 its associated I/O counters, to be displayed.

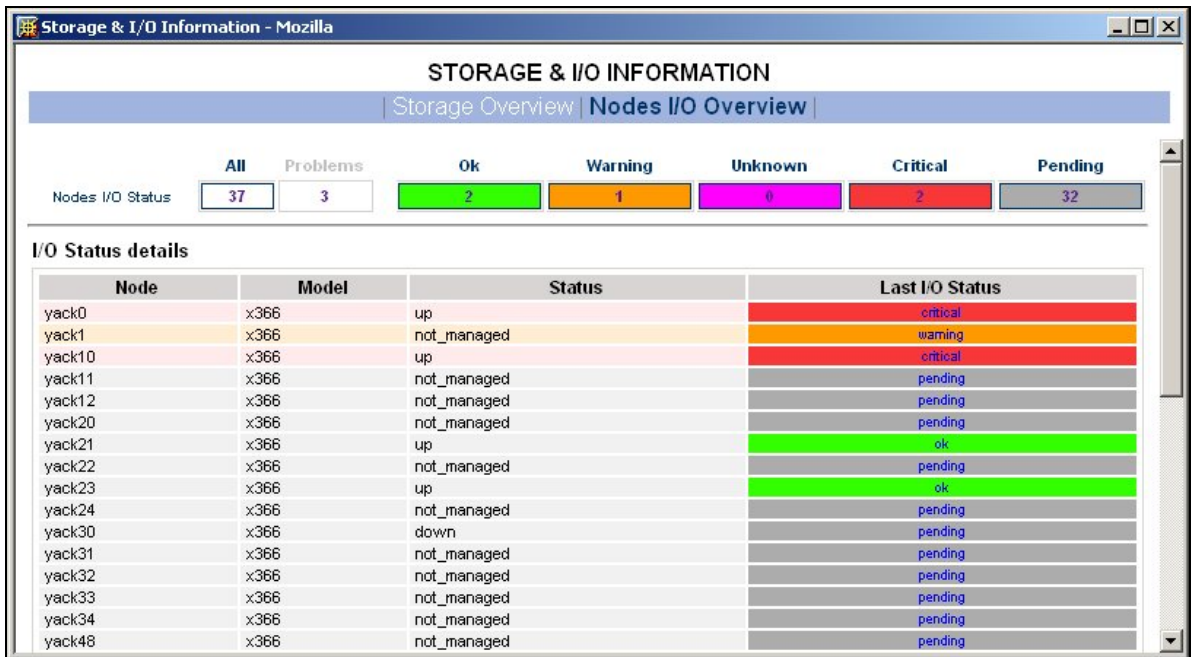


Figure 6-9. Nodes I/O Overview

6.3.3 Querying the Cluster Management Data Base

The **storstat** command obtains status information from the **ClusterDB** and formats the results for storage administrators.

See the help page for this command for more information:

```
storstat -h
```

The following paragraphs describe the most useful options.

6.3.3.1 Checking Storage System Status

Use the command below to display all the registered storage systems with their status and location in the cluster. 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
```

6.3.3.2 Checking Status of Hardware Elements

To display a list of faulty components for all the 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 for all the storage system components:

```
storstat -d -n <disk_array_name>
```

Note If the `-n` flag is omitted the list will include 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>
```

6.4 Monitoring Brocade Switch Status

Each Brocade Fibre Channel switch is monitored by **Bull System Manager - HPC Edition**.

The same check period as for Ethernet switches will be used (10 minutes, this is configurable). No specific configuration is required for the FC switches in order to be able to use the **telnet** interface.

Several generic services are defined for Brocade switches. They reflect the global status of a class of components for the selected switch. A mapping between the SNMP MIB (Management Information Base) values available and returned from the switch, and the **Bull System Manager - HPC Edition** status give the following set of states for each managed services:

Ethernet interface Service

| | |
|----------|--|
| OK | No problem Criteria: <ul style="list-style-type: none">• The Fping of the Ethernet interface is OK |
| CRITICAL | Criteria: <ul style="list-style-type: none">• The Fping of the Ethernet interface is KO |

FC port

| | |
|----------|---|
| OK | No problem Criteria: <ul style="list-style-type: none">All FC ports are OK. |
| WARNING | Maintenance operation must be scheduled Criteria: <ul style="list-style-type: none">Not in critical statusSome ports have a warning statusNumber of operating port higher than expected in the DB (fc_switch.oper_port_threshold) |
| CRITICAL | Degraded service. Maintenance operation mandatory Criteria: <ul style="list-style-type: none">One or more ports are in a critical status.Number of operating ports lower than expected (fc_switch.oper_port_threshold) |
| UNKNOWN | Cannot access the status |
| PENDING | Not yet initialized |

Fans

| | |
|----------|--|
| OK | No problem Criteria: <ul style="list-style-type: none">All fans are present and OK |
| WARNING | Maintenance operation must be scheduled Criteria: <ul style="list-style-type: none">Some fans are in the warning stateDoes not meet the criteria for critical status. |
| CRITICAL | Degraded service. Maintenance operation mandatory Criteria: <ul style="list-style-type: none">At least one of the fan is in a critical state |
| UNKNOWN | Cannot access the status |
| PENDING | Not yet initialized |

Power Supply

| | |
|----------|--|
| OK | No problem Criteria: <ul style="list-style-type: none">All power supplies are present and okNo Power Supply is detected on the switch. |
| WARNING | Maintenance operation must be scheduled Criteria: <ul style="list-style-type: none">Some power supplies are in the warning stateDoes not meet the criteria for critical status. |
| CRITICAL | Degraded service. Maintenance operation mandatory Criteria: <ul style="list-style-type: none">At least one of the power supplies is in a critical state |

| | |
|---------|--------------------------|
| UNKNOWN | Cannot access the status |
| PENDING | Not yet initialized |

Temperature Sensor

| | |
|----------|--|
| OK | <p>No problem</p> <p>Criteria:</p> <ul style="list-style-type: none"> All Temperature sensor are present and OK |
| WARNING | <p>Maintenance operation must be scheduled</p> <p>Criteria:</p> <ul style="list-style-type: none"> Some Temperature sensor are in the warning state Does not meet the criteria for critical status. |
| CRITICAL | <p>Degraded service. Maintenance operation mandatory</p> <p>Criteria:</p> <ul style="list-style-type: none"> At least one of the Temperature Sensor is in a critical state |
| UNKNOWN | Cannot access the status |
| PENDING | Not yet initialized |

Global Status

| | |
|----------|---|
| OK | <p>No problem</p> <p>Criteria:</p> <ul style="list-style-type: none"> Global Brocade switch status is OK. |
| WARNING | <p>Maintenance operation must be scheduled</p> <p>Criteria:</p> <ul style="list-style-type: none"> Some of the other services are warning (but none critical). Switch name (switchX) different as expected (fcswX) |
| CRITICAL | <p>Degraded service. Maintenance operation mandatory</p> <p>Criteria:</p> <ul style="list-style-type: none"> One of the other services is critical. The storage system has detected a critical error which has not been reported by one of the other services. |
| UNKNOWN | Cannot access the status |
| PENDING | No yet initialized |

The different services managed by Bull System Manager - HPC Edition for the Brocade switch are shown below:

| Host | Service | Status | Last Check | Duration | Attempt | Status Information |
|---------|---------------------|----------|---------------------|---------------|---------|--|
| lpwu0c0 | Ethernet interfaces | OK | 28-02-2006 11:01:15 | 0d 3h 4m 35s | 1/1 | down : [] - up : [10.0.0.80] |
| | 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 6-10. Detailed Service status of a brocade switch

6.5 Managing Storage Devices with Bull CLI

This section describes the commands available for each device family.

These commands offer the most useful subset of management features, implemented for 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.

6.5.1 Bull FDA Storage Systems

The Administrator must be familiar with the **FDA** terminology and management tasks.

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.

6.5.2 DataDirect Networks Systems - DDN Commands

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.

6.5.2.1 **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**

6.5.2.2 **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.

6.5.2.3 **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**.

6.5.2.4 **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.

6.5.2.5 **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**.

6.5.2.6 **ddn_check_format**

This command allows you to check the formatting status for a list of DDN systems.

6.5.2.7 **ddn_firmup**

This command automatically upgrades the firmware of the singlets of a DDN system. The Management Node can be used as TFTP server.

6.5.3 **Bull Optima1250 Storage Systems**

The administrator must be familiar with the **OPTIMA1250** Storage System terminology and management tasks.

Note The High Availability solution does not apply for nodes which are connected to **Optima1250** Storage Bays.

See The **StoreWay** OPTIMA1250 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 OPTIMA1250.

The following services are provided by the command:

- **getstatus**
- **list**
- **checkformat**
- **luninfo**
- **zoninfo**
- **poolbind**
- **ldbind**

- sparebind
- setldmap
- setldwn
- poolunbind
- ldunbind
- spareunbind
- unconfig

The OPTIMA1250 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

# OPTIMA1250 Storeway Master Administrator login
xyradmin = admin

# OPTIMA1250 Storeway Master Administrator password
xyrpasswd = password
```

For more information, read the man page or check the command's help.

6.5.4 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

6.6 Using Management Tools

Please refer to the storage system documentation 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.).

6.7 Configuring Storage Devices

6.7.1 Planning Tasks

Storage system configuration requires careful planning. At least two steps are required.

STEP 1 – DEFINE THE DEVICE CONFIGURATION

The storage administrator must define the storage configuration that 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.

Note With some versions of Fibre Channel adapter node drivers, the correct detection of the LUNs for a storage device port is dependent on the accessibility of a LUN numbered LUN 0. It is recommended the Access Control groups for a storage device are configured so that the list of LUNs declared in each group always include an external LUN that is numbered LUN 0.

- 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.

6.7.2 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.

6.7.3 Understanding the Configuration Deployment Service

The configuration deployment service relies on modeling the storage system configuration. The model defines all the configuration parameters (see 6.7.1 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.

6.7.3.1

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.

6.7.3.2 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.

6.7.3.3 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 so the results are reformatted.

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.

6.8 User Rights and Security Levels for the Storage Commands

6.8.1 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 **Bull System Manager - HPC Edition** 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**

Notes • **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
```

- The **PATH** of the **dba** 'username' must be defined 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
```

6.8.2 Other Node Types

All the available storage commands can only be launched as the root user, without exception.

6.8.3 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. These files are named *.conf, for example **storframework.conf**.

Chapter 7. Monitoring with Bull System Manager - HPC Edition

Bull System Manager - HPC Edition provides the monitoring functions for Bull Extreme Computing systems. It uses **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 for the whole cluster. The status of a large number of elements can be monitored.

This chapter covers the following topics:

- *7.1 Launching Bull System Manager - HPC Edition*
- *7.2 Access Rights*
- *7.3 Hosts, Services and Contacts for Nagios*
- *7.4 Using Bull System Manager - HPC Edition*
- *7.5 Map Button*
- *7.6 Status Button*
- *7.8 Alerts Button*
- *7.9 Storage Overview*
- *7.10 Shell*
- *7.11 Monitoring the Performance - Ganglia Statistics*
- *7.12 Group Performance View*
- *7.13 Global Performance View*
- *7.14 Configuring and Modifying Nagios Services*
- *7.15 General Nagios Services*
- *7.16 Management Node Nagios Services*
- *7.17 Ethernet Switch Services*
- *7.18 Cool Cabinet Door Services*

7.1 Launching Bull System Manager - HPC Edition

Note The cluster database (**ClusterDB**) must be running before monitoring is started. See the chapter on *Cluster Data Base Management*.

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 **FireFox** and enter the following URL:

<http://<ManagementNode>/BSM/>

Note **FireFox** is the mandatory navigator for **Bull System Manager – HPC Edition**

7.2 Access Rights

7.2.1 Administrator Access Rights

By default, the Administrator uses the following login and password:

login: **nagios**
password: **nagios**

Once the graphical interface for monitoring has opened, see *Figure 7-1*, the Administrator is able to enter host and service commands, whereas an ordinary user will only be able to consult the interface.

7.2.2 Standard User Access Rights

By default, an ordinary user uses the following login and password:

login: **guest**
password: **guest**

7.2.3 Adding Users and Changing Passwords

The **htpasswd** command is used to create new user names and passwords.

Create additional users for the graphical interface as follows:

1. Enter the following command:

```
htpasswd /opt/BSMServer-Base/core/etc/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 define the user profile in the **/opt/BSMServer-Base/core/share/console/NSMasterConfigInfo.inc** file (either as an Administrator or as an Operator).

Change the password for an existing user as follows:

1. Enter the following command:

```
htpasswd /opt/BSMServer-Base/core/etc/htpasswd.users <login>
```

2. Enter and confirm the new password when prompted.

Note Some of these steps have to be done as the **root** user.

See The **Bull System Manager** documentation for more information on adding users and on account management.

7.3 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.

The **host group** definition is used to group one or more hosts together for display purposes in the graphical interface.

The **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 users logged-in, free disk space, etc.).

Note **Bull System Manager – HPC Edition** will display the services specific to each host when the host is selected within the **Bull System Manager – HPC Edition** interface.

The **contact** definition is used to identify someone who should be contacted in the event of a problem on your network.

The **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 allows greater flexibility in determining who gets notified for particular events.

For more information on the definitions, and the arguments and directives which may be used for the definitions see:

http://nagios.sourceforge.net/docs/3_0/

Alternatively, select the **Documentation** link from the **Bull System Manager** opening screen or select the **Documentation** button in the title bar.

7.4 Using Bull System Manager - HPC Edition

The graphical interface of Bull System Manager - HPC Edition is shown inside a Web browser.

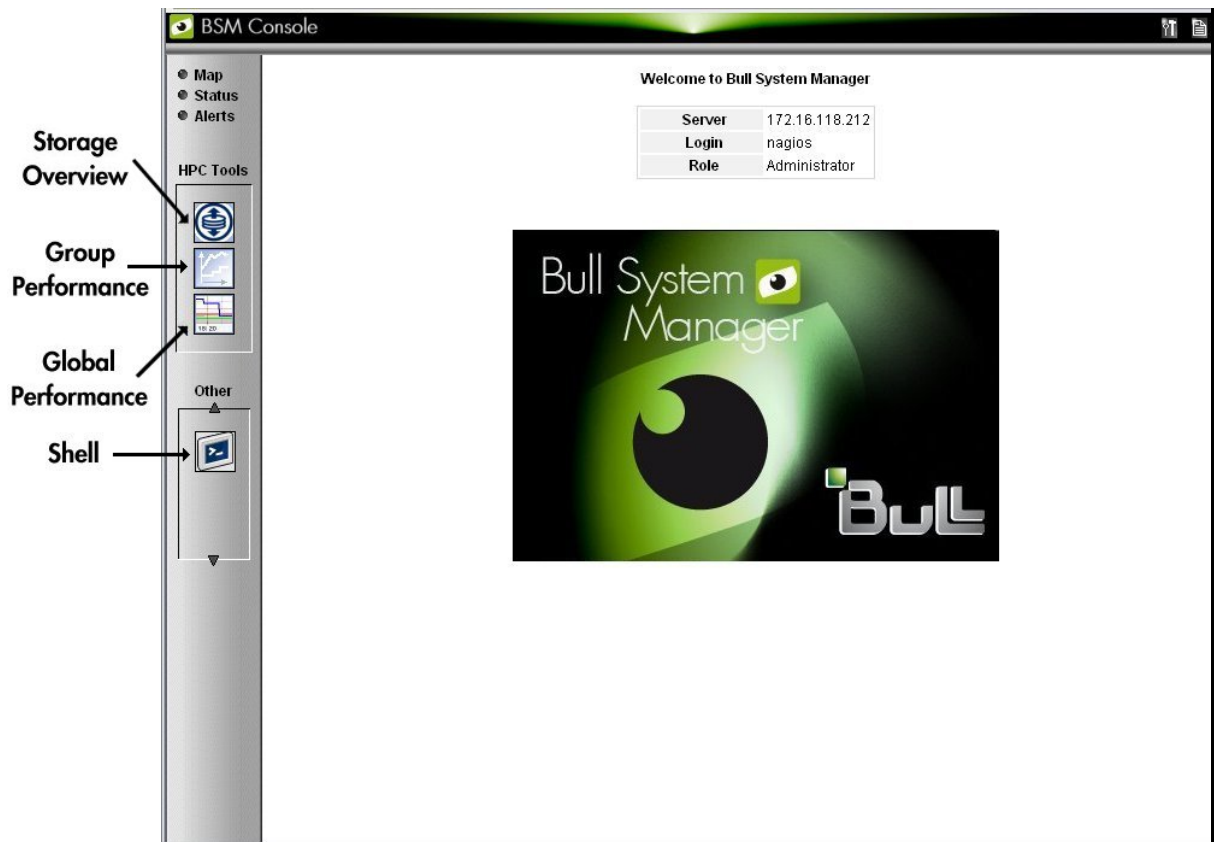


Figure 7-1. Bull System Manager - HPC Edition opening view

7.4.1 Bull System Manager - HPC Edition – View Levels

Initially, the console will open and the administrator can choose to view different types of monitoring information, with a range of granularity levels, by clicking on the icons in the left hand vertical tool bar, and then clicking on the links in the different windows displayed. The information displayed is contextual depending on the host or service selected. Using the links it is possible to descend to a deeper level, to see more detailed information for a particular host, host group, or service. For example, the **Cabinet Rack** map view in *Figure 7-2* leads to the **Rack View** in *Figure 7-3*, which in turns leads to the more detailed **Services** view in *Figure 7-5*, for the host selected in the **Rack View**.

7.5 Map Button

The **Map** button is displayed at the top right hand side of the opening. When this is selected the drop down menu provides two view options, **all status** or **ping**, inside the main window.

7.5.1 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 frame color for each cabinet is determined by the component within it with the highest alarm status, for example if an **Ethernet interface** is in the **critical** status than the status for the whole rack will be **critical**.

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 the **hosts** and **services** and the **Availability Indicators** view window will appear at the top right hand side of the screen – see *Figure 7-2*.



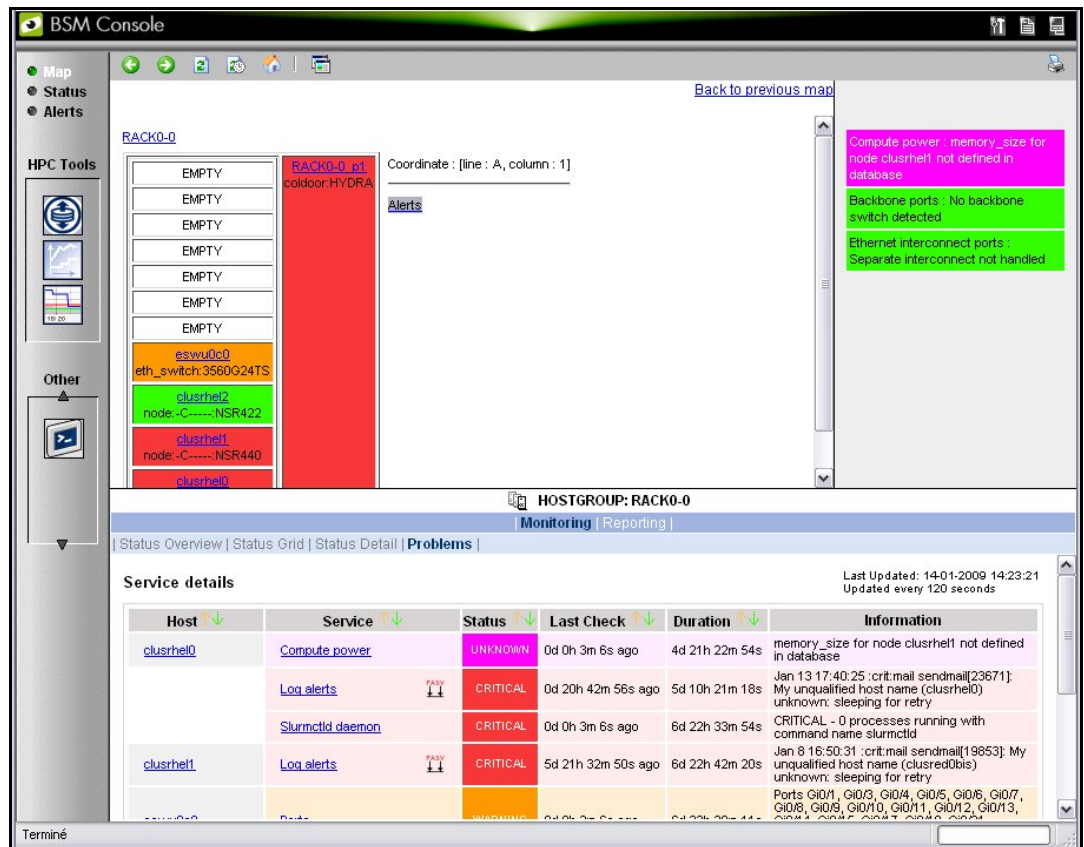
Figure 7-2. **Map** button **all status** opening view

When the cursor passes over a rack, information about it (label, type, and the elements contained in the rack) is displayed. When the user clicks on a cabinet, a detailed view of the cabinet is displayed – see **Rack view** in *Figure 7-3*. This displays additional information, including its physical position and the services which are in a non-OK state.

7.5.2 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 for the rack is displayed at the bottom of the view – see *Figure 7-3*.

Clicking on a component displays a detailed view for it.



The screenshot shows the BSM Console interface. On the left, there are navigation options: Map, Status, Alerts, HPC Tools, and Other. The main area displays a rack layout for RACK0-0. The rack is a vertical column of slots. Most are labeled 'EMPTY'. One slot is highlighted in red and labeled 'RACK0-0_p1 coldoor:HYDRA'. Below it, there are several nodes: 'eswu0c0 eth_switch:3580G24TS', 'clusrhel2 node-C-----NSR422', 'clusrhel1 node-C-----NSR440', and 'clusrhel0'. To the right of the rack, there is a coordinate field 'Coordinate: [line : A, column : 1]' and an 'Alerts' link. Further right, there are three colored boxes with text: a pink box 'Compute power: memory_size for node clusrhel1 not defined in database', a green box 'Backbone ports: No backbone switch detected', and another green box 'Ethernet interconnect ports: Separate interconnect not handled'. At the bottom, there is a 'HOSTGROUP: RACK0-0' section with tabs for 'Monitoring' and 'Reporting'. Below that, there is a 'Service details' window with a table of services for hosts 'clusrhel0' and 'clusrhel1'. The table has columns for Host, Service, Status, Last Check, Duration, and Information.

| Host | Service | Status | Last Check | Duration | Information |
|-----------|------------------|----------|--------------------|----------------|---|
| clusrhel0 | Compute power | UNKNOWN | 0d 0h 3m 6s ago | 4d 21h 22m 54s | memory_size for node clusrhel1 not defined in database |
| | Log alerts | CRITICAL | 0d 20h 42m 56s ago | 5d 10h 21m 18s | Jan 13 17:40:25: crit.mail sendmail[23671]: My unqualified host name (clusrhel0) unknown: sleeping for retry |
| | Slurmctld daemon | CRITICAL | 0d 0h 3m 6s ago | 6d 22h 33m 54s | CRITICAL - 0 processes running with command name slurmctld |
| clusrhel1 | Log alerts | CRITICAL | 5d 21h 32m 50s ago | 6d 22h 42m 20s | Jan 8 16:50:31: crit.mail sendmail[19853]: My unqualified host name (clusred0bis) unknown: sleeping for retry |
| | Ports | UNKNOWN | 0d 0h 3m 6s ago | 6d 22h 33m 54s | Ports Gi0/1, Gi0/3, Gi0/4, Gi0/5, Gi0/6, Gi0/7, Gi0/8, Gi0/9, Gi0/10, Gi0/11, Gi0/12, Gi0/13, Gi0/14, Gi0/15, Gi0/16, Gi0/17, Gi0/18, Gi0/19, Gi0/20, Gi0/21, Gi0/22, Gi0/23, Gi0/24, Gi0/25, Gi0/26, Gi0/27, Gi0/28, Gi0/29, Gi0/30, Gi0/31, Gi0/32, Gi0/33, Gi0/34, Gi0/35, Gi0/36, Gi0/37, Gi0/38, Gi0/39, Gi0/40, Gi0/41, Gi0/42, Gi0/43, Gi0/44, Gi0/45, Gi0/46, Gi0/47, Gi0/48, Gi0/49, Gi0/50, Gi0/51, Gi0/52, Gi0/53, Gi0/54, Gi0/55, Gi0/56, Gi0/57, Gi0/58, Gi0/59, Gi0/60, Gi0/61, Gi0/62, Gi0/63, Gi0/64, Gi0/65, Gi0/66, Gi0/67, Gi0/68, Gi0/69, Gi0/70, Gi0/71, Gi0/72, Gi0/73, Gi0/74, Gi0/75, Gi0/76, Gi0/77, Gi0/78, Gi0/79, Gi0/80, Gi0/81, Gi0/82, Gi0/83, Gi0/84, Gi0/85, Gi0/86, Gi0/87, Gi0/88, Gi0/89, Gi0/90, Gi0/91, Gi0/92, Gi0/93, Gi0/94, Gi0/95, Gi0/96, Gi0/97, Gi0/98, Gi0/99, Gi0/100 |

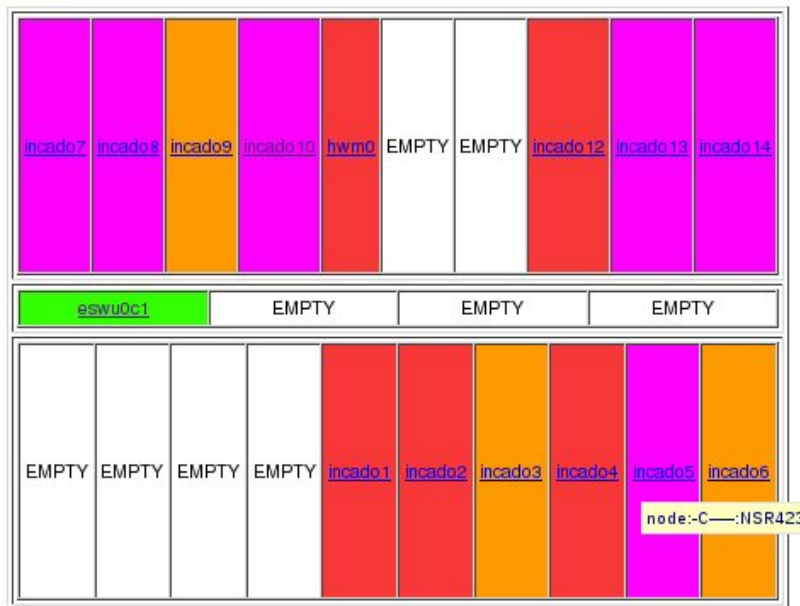
Figure 7-3. Rack view with the Problems window at the bottom

More detailed information regarding the hardware components and services associated with a host appear, when the host in the rack view is clicked. This leads to another pop up window which includes further information for the host and its services – see *Figure 7-5*.

7.5.3 bullx blade map view

For nodes which include **bullx** blades the **Rack View** when selected will open the **bullx blade map view** - see *Figure 7-4*. This displays the individual blades in the **bullx blade chassis**. The Hardware Manager (CMC), shown as **hwn0** below, is displayed, and the **Ethernet/InfiniBand** switches are shown between the two rows of blades. Clicking on an individual blade will open the **Host Services View** - *Figure 7-5*, the same as for any other node.

INCA1



Coordinate : [line : none, column : none]

[Alerts](#)

Figure 7-4. bullx blade map view

7.5.4 Host Services detailed View

Clicking the **Status** or a **Service** links in this window displays more specific information for the component or service.

Monitoring
Reporting

Alert History
Notifications
Availability
Status Trends

| | | | | | | | |
|-------------------------|------------|-----------------|-----------|----------------|----------------|-----------------|----------------|
| | All | Problems | Ok | Warning | Unknown | Critical | Pending |
| Selected Host Services: | 8 | 4 | 3 | 1 | 0 | 3 | 1 |

Click on status links to display the selected services

Service details

Last Updated: 19-02-2008 15:58:28
Updated every 120 seconds

| Service | Status | Last Check | Duration | Information |
|-------------------------------------|----------|---------------------|-----------------|---|
| Ethernet interfaces | OK | 0d 4h 32m 30s ago | 3d 22h 32m 30s | down : [] - up : [13.2.0.5 192.20.0.5] |
| Hardware status | WARNING | 0d 4h 32m 30s ago | 4d 3h 32m 30s | for domain OXAN-S11-00025 functional status is WARNING (domain state is RUNNING) according to papu0c1 PAM manager . |
| IO status | OK | 0d 0h 58m 18s ago | 7d 23h 43m 19s | OK (IO status details) All I/O resources are OK |
| Log alerts | CRITICAL | 18d 0h 10m 33s ago | 26d 16h 55m 40s | Feb 1 15:47:53 : emerg:kern kernel: Kernel panic - not syncing: device_interrupt: interrupt cookie 200000000167 not found |
| NSDoctor | PENDING | 48d 3h 54m 50s+ ago | 48d 3h 54m 50s+ | Service is not scheduled to be checked... |
| Postbootchecker | CRITICAL | 4d 5h 11m 53s ago | 48d 2h 13m 28s | Log file : /var/log/postbootchecker/nova4.log |
| RM status | CRITICAL | 4d 20h 4m 25s ago | 4d 20h 4m 44s | configured out (automatically configured out) |
| Temperature | OK | 0d 4h 32m 12s ago | 4d 3h 32m 24s | All QBBs OK |

8 Matching Service Entries Displayed (filter: Service Status **PENDING OK WARNING UNKNOWN CRITICAL**)

Figure 7-5. Host Service details

By clicking on the links in the windows even more detailed information is provided for the services.

7.5.5 Control view

The **Control** button in the middle of screen provides details for the Management Node and the commands which apply to it - see *Figure 10-5*.

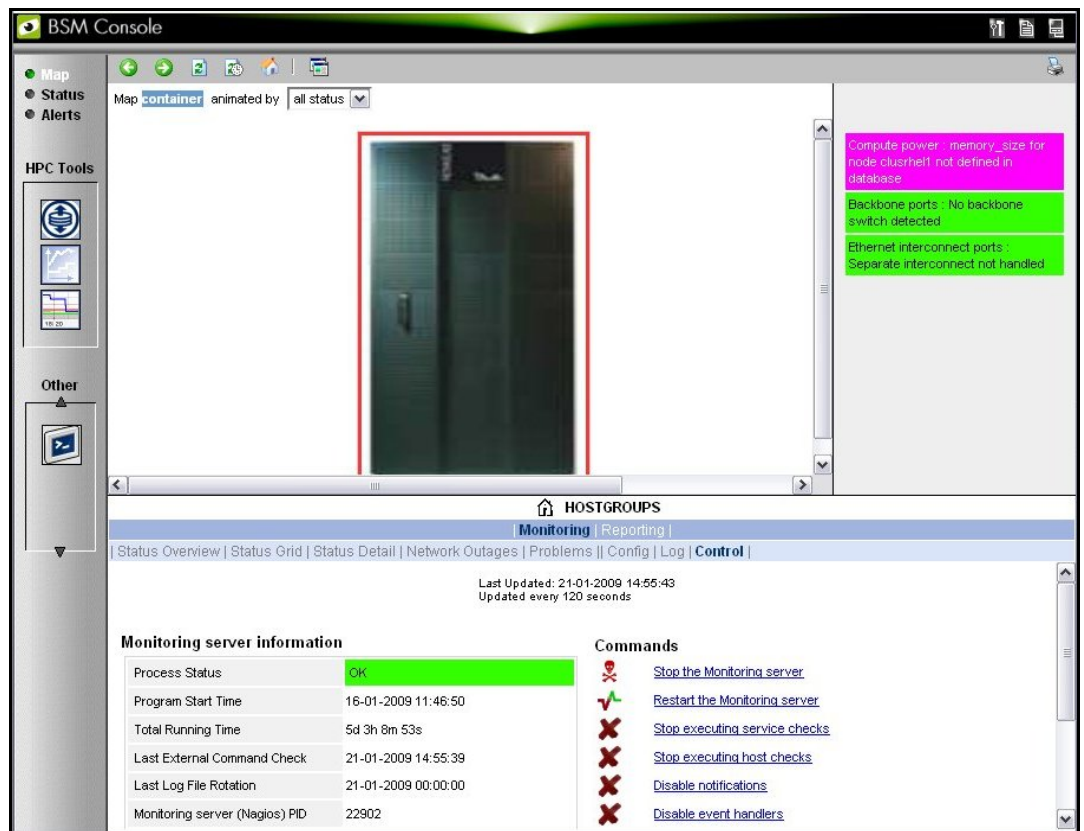


Figure 7-6. Monitoring Control Window

7.5.6 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.

7.6 Status Button

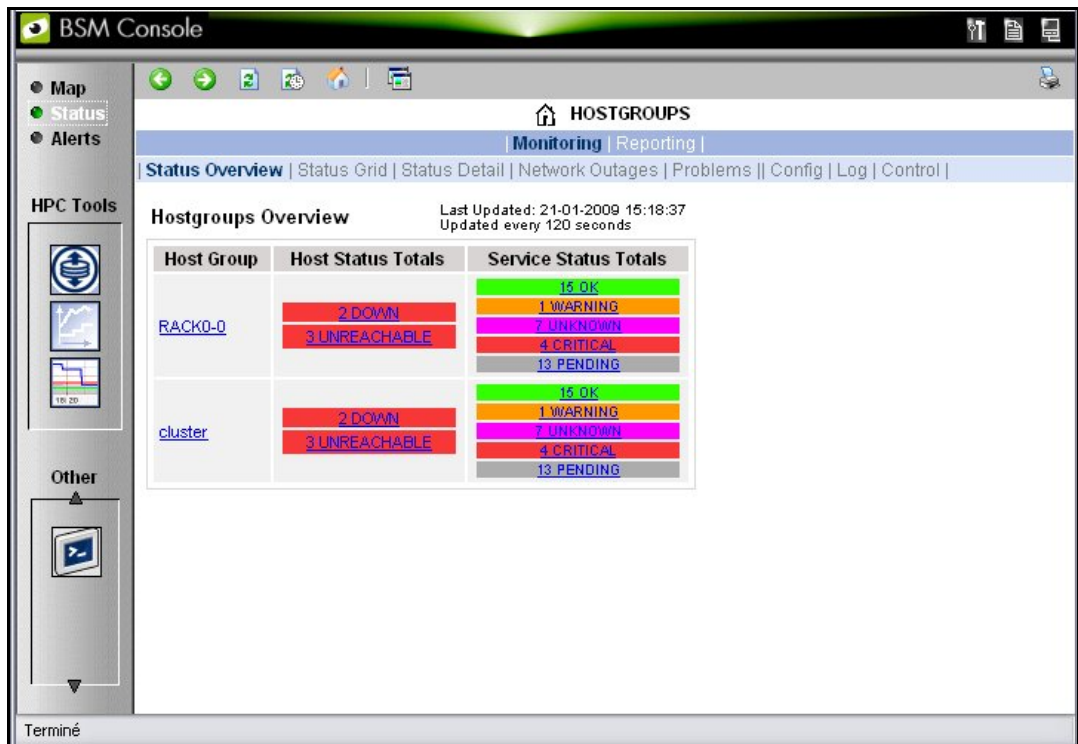


Figure 7-7. Status Overview screen

When the **Status** button is clicked, a screen appears which lists all the hosts, and the status of the services running on them, as shown in Figure 7-7. 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.

7.7 Log Window

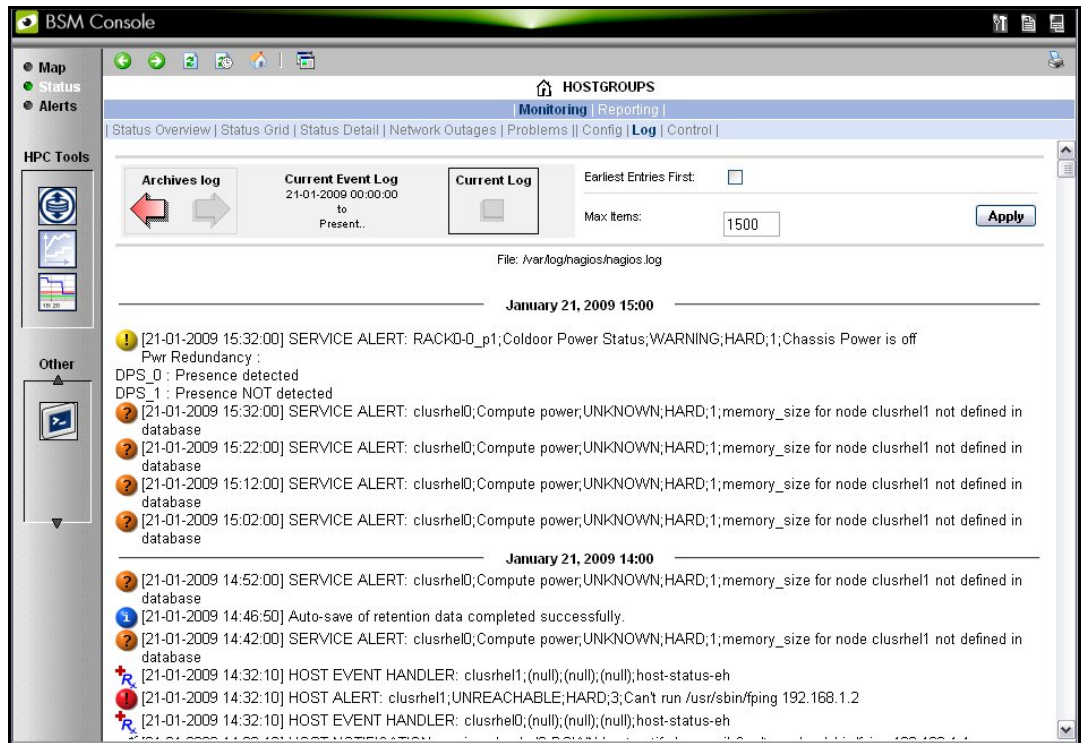


Figure 7-8. Monitoring - Log Window

The **Log Window** which is useful for tracing problems appears when the **Monitoring - Log** button is clicked. This displays a screen similar to that in *Figure 7-8*. 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.

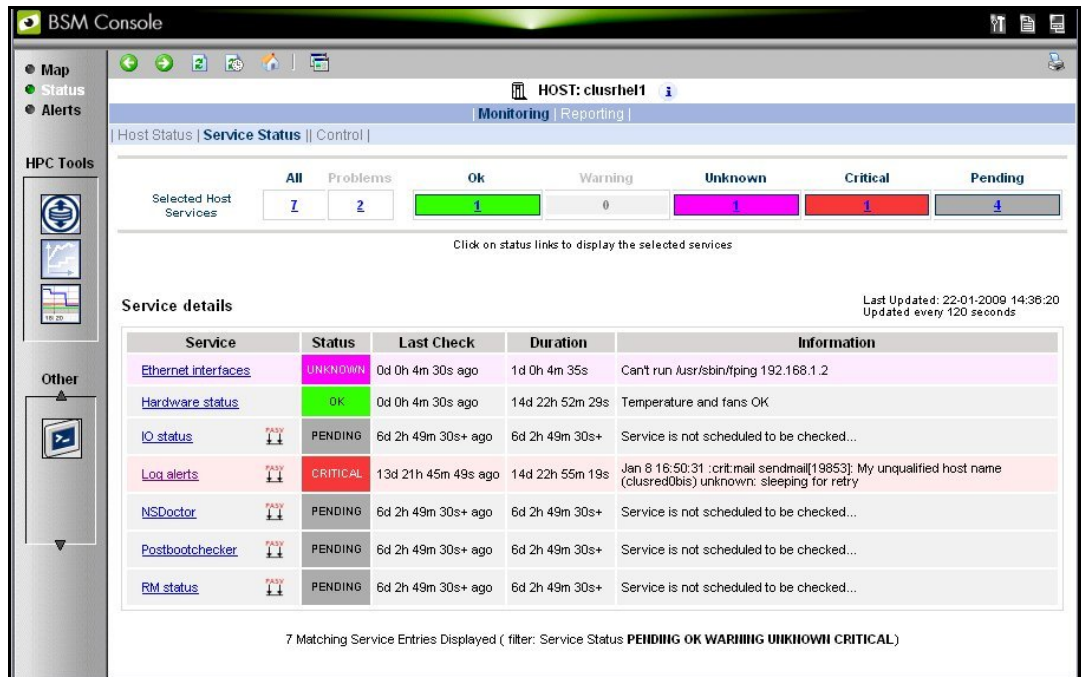


Figure 7-9. Monitoring Service Status window for a host with the Log Alerts link highlighted.

7.8 Alerts Button

The screenshot shows the BSM Console Alert Viewer interface. The window title is "BSM Console" and the main heading is "ALERTS". Below the heading are tabs for "Monitoring" and "Reporting". The "Alert Viewer" section includes filters for hostgroups and hosts, alert type (set to "Hosts and Services"), alert level (set to "All"), report period (set to "Last 24 Hours"), and a "Max Items" field set to 15. There are "Apply" and "Reset" buttons. The "Matching Alerts" section shows a table of alerts with the following data:

| Time | Host | Service | State | Count | Information |
|---------------------|------------|-----------------------------|-------------|-------|---|
| 22-01-2009 09:02:00 | clusrhel0 | Compute power | UNKNOWN | 113 | memory_size for node clusrhel1 not defined in database |
| 22-01-2009 04:03:20 | clusrhel0 | Log alerts | CRITICAL | 719 | Jan 22 04:03:16 :crit.mail sendmail[8554]: My unqualified host name (clusrhel0) unknown: sleeping for retry |
| 21-01-2009 16:37:00 | clusrhel0 | Hardware status | OK | 1 | Temperature and fans OK |
| 21-01-2009 16:32:30 | clusrhel0 | Hardware status | UNKNOWN | 1 | Timeout |
| 21-01-2009 15:32:00 | RACK0-0_p1 | Coldoor Power Status | WARNING | 1 | Chassis Power is off Pwr Redundancy : DPS_0 : Presence detected DPS_1 : Presence NOT detected |
| 21-01-2009 14:32:10 | clusrhel1 | N/A | UNREACHABLE | 2 | Can't run Ausr/sbin/fping 192.168.1.2 |
| 21-01-2009 14:32:10 | clusrhel0 | N/A | DOWN | 3 | Can't run Ausr/sbin/fping 192.168.1.1 |
| 21-01-2009 14:32:10 | clusrhel2 | N/A | UNREACHABLE | 2 | Can't run Ausr/sbin/fping 192.168.1.3 |
| 21-01-2009 14:32:10 | RACK0-0_p1 | N/A | UNREACHABLE | 2 | Can't run Ausr/sbin/fping 192.168.1.251 |
| 21-01-2009 14:32:10 | eswu0c0 | N/A | DOWN | 4 | Can't run Ausr/sbin/fping 192.168.1.200 |
| 21-01-2009 14:31:50 | eswu0c0 | Ethernet interfaces | UNKNOWN | 1 | Can't run Ausr/sbin/fping 192.168.1.200 |
| 21-01-2009 14:31:50 | clusrhel1 | Ethernet interfaces | UNKNOWN | 1 | Can't run Ausr/sbin/fping 192.168.1.2 |
| 21-01-2009 14:31:50 | RACK0-0_p1 | Coldoor Ethernet interfaces | UNKNOWN | 1 | Can't run Ausr/sbin/fping 192.168.1.251 |
| 21-01-2009 14:31:50 | clusrhel0 | Ethernet interfaces | UNKNOWN | 1 | Can't run Ausr/sbin/fping 192.168.1.1 |
| 21-01-2009 14:31:50 | clusrhel2 | Ethernet interfaces | UNKNOWN | 1 | Can't run Ausr/sbin/fping 192.168.1.3 |

(Total alerts : 856, displayed lines : 15, displayed alerts : 853)

Figure 7-10. Alert Window showing the different alert states

The **Bull System Manager Alert Viewer** application displays monitoring alerts (also called events) for a set of **hostgroups**, **hosts** and **services**.

Alerts 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.

Alerts are visible following the selection of the **Alerts** Button, followed by the **Reporting** button, and then by the **Alert Viewer** – see *Figure 7-10*.

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 archived.

Bull System Manager - HPC Edition Alert Viewer utility scans the current monitoring log and archives according to **Report Period** filter settings.

Alerts Level

The following **Alert Level** filters are available:

- **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 or 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.

7.8.1 Active Checks

Active monitoring consists in running a plug-in at regular intervals for a service, this carries out checks and sends the results back to **Nagios**. **Active checks** are set by selecting the **Service** in the **Alert Viewer** window and using the Service Command listed, shown below, to either enable or disable the **Active Check** type.

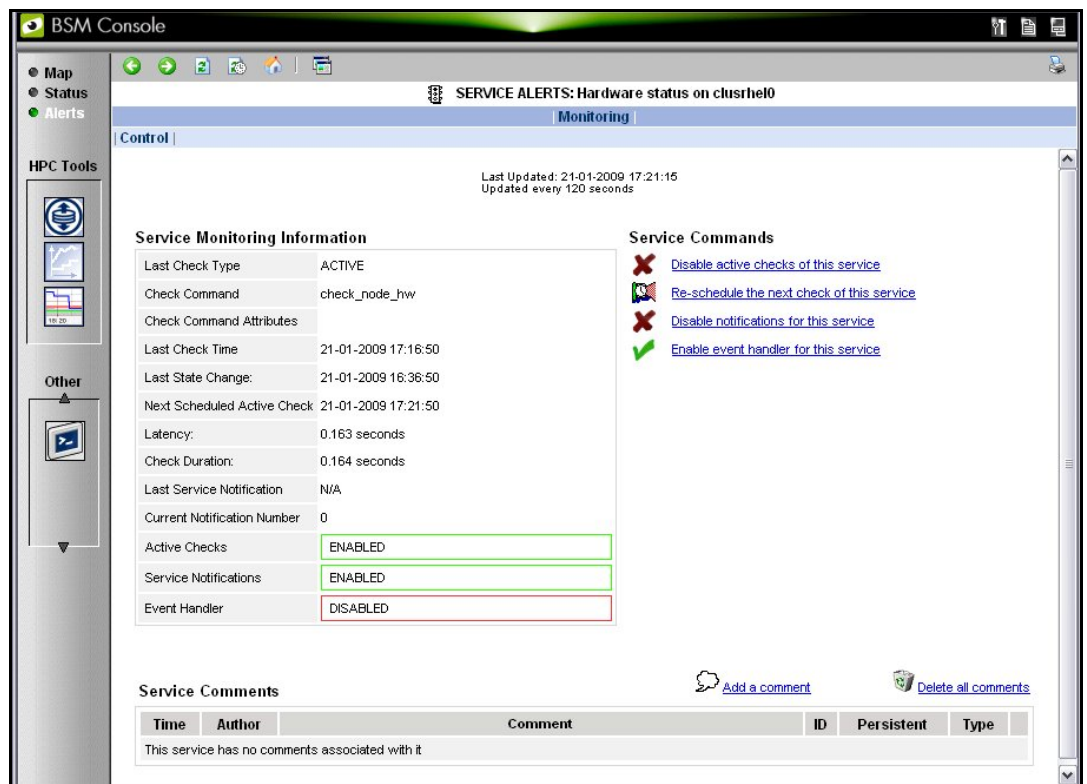


Figure 7-11. **Monitoring Control** Window used to set **Active Checks** for a Service

The **Nagios plug-in** returns a code corresponding to the **Alert** alarm state. The state is then displayed in a colour coded format, in the **Alert Viewer** window - see *Figure 7-10* - as follows:

- 0 for **OK/UP** (Green background)
- 1 for **WARNING** (Orange background)
- 2 for **CRITICAL/DOWN/UNREACHABLE** (Red background)
- 3 for **UNKNOWN** (Violet background)

The plug-in also displays an explanatory text for the alarm level in the adjacent **Information** column.

7.8.2 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/or **Service** concerned, as well as the return code and the explanatory text.

Passive checks appear with a GREY background in the list of alerts.

7.8.3 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 NS423.

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**, **eth_switch**, **ic_switch**).

7.8.4 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 via a User Script. Set the **<notification_interval>** value to 0 to prevent notifications from being sent out more than once for any given problem or change.

The **Monitoring Control** window - see *Figure 7-11* 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 indicated below.



Figure 7-12. Hostgroups Reporting Notifications Window showing the Notification Levels

7.8.5 Acknowledgments

As the **Administrator**, you may choose whether or not alerts are acknowledged, and decide whether they should be displayed or not.

7.8.6 Running a Script

Bull System Manager - HPC Edition can be configured to run a script when a state changes or an alert occurs. User scripts which define events or physical changes to trigger **Nagios** alerts may also be used. More information on scripts or third party plug-ins is available in the documentation from <http://www.nagios.org/docs/>

Below is an example of script.

```
#!/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;
```

In order that e-mail alerts are sent whenever there is a problem, a SMTP server, for example **PostFix** or **Sendmail**, has to be running on the Management node. By default, the e-mail alerts are sent to nagios@localhost on the Management Node.

Normally, by default, only the cluster administrators will receive the alerts for each change for all the Hosts and Services. To send e-mails alerts to other addresses, create the new contacts, and add them to the contact groups. The files to modify are `/etc/nagios/contacts.cfg` and `/etc/nagios/contactgroups.cfg`.

7.8.7 Generating SNMP Alerts

When **Bull System Manager - 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
```

7.8.8 Resetting an Alert Back to OK

To reset an alert back to zero click the Service or the Host concerned, then on the menu **Submit passive check result for this service**. Set the **Check Result** to OK, if this is not already the case, fill in the **Check Output** field with a short explanation, and then click the **Commit** button. The return to the OK state will be visible once Nagios has run the appropriate command.

7.8.9 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 define specific parameter values. A message will inform the administrator if a value is missing.

7.8.10 Comments

Users of a particular host or service can post comments from the **Monitoring Control** window - see *Figure 7-11*

7.9 Storage Overview

Select the **Storage Overview** button in the vertical toolbar on the left hand side to display information similar to that shown below.



Figure 7-13. Storage overview window

More detailed information is provided by clicking on the ATTENTION and FAILED sections of the component summary status bars.

See *Chapter 6 – Storage Device Management* for information on **Bull System Manager - HPC Edition** and storage views.

7.10 Shell

The **Shell** button can be used to open a command shell on the Management Node.

7.11 Monitoring the Performance - Ganglia Statistics

Bull System Manager - HPC Edition provides the means to visualize the performance for the cluster by clicking the icons in the vertical left hand tool bar – see Figure 7-1. 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 define the observation period and display the measurement details for a particular node using the Ganglia interface.

7.12 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 to a particular node.



Figure 7-14. Group Performance view

7.13 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 graph shows changes for the performance metrics over a user defined period of time.

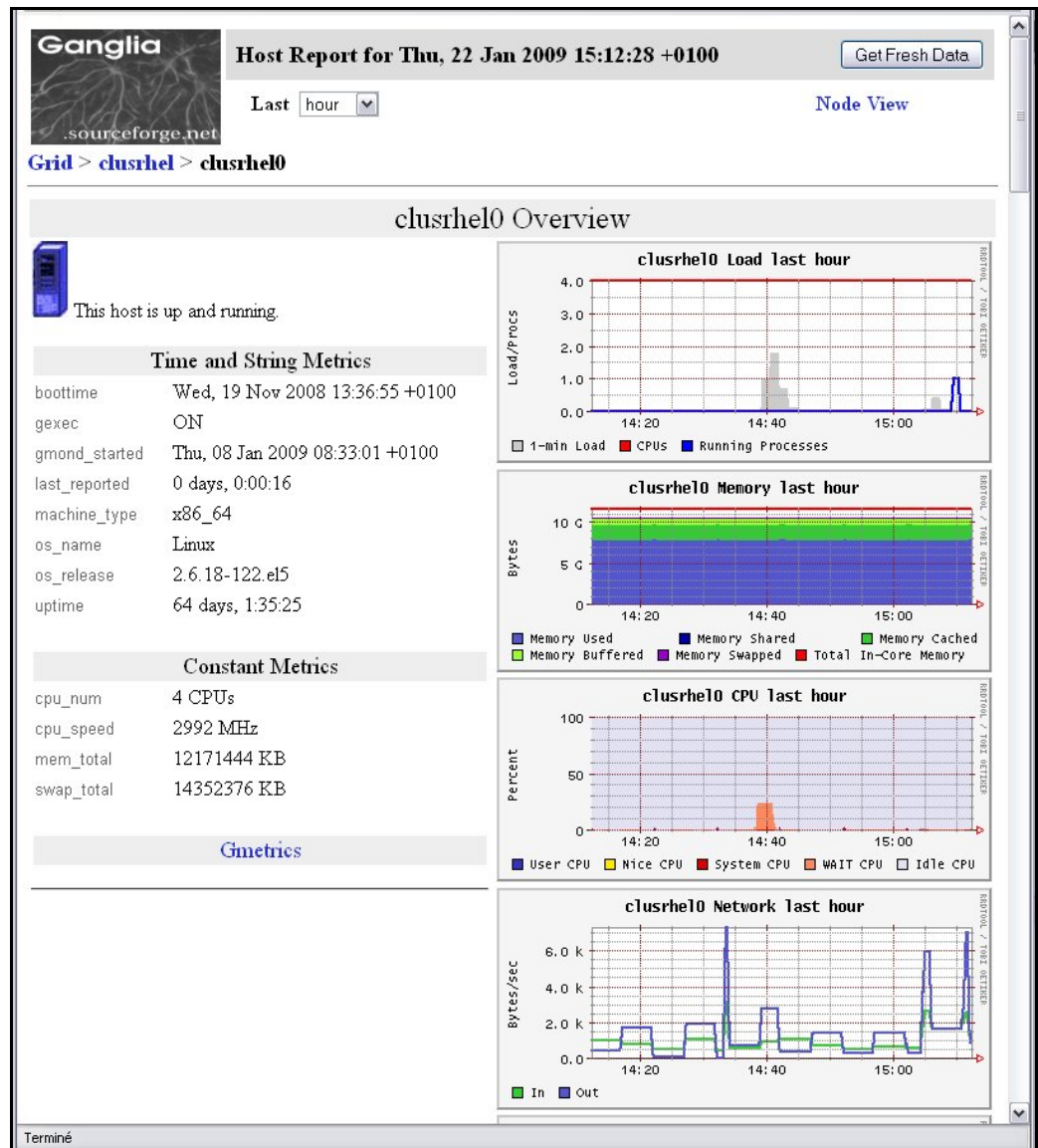


Figure 7-15. Global overview for a host (top screen)

More detailed views are shown by scrolling the window down – see Figure 7-16.



Figure 7-16. Detailed monitoring view for a host (bottom half of screen displayed in Figure 7-15)

7.13.1 Modifying the Performance Graph 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 color code.

<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)
```


7.13.2 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`.

7.14 Configuring and Modifying Nagios Services

7.14.1 Configuring Using the Database

The command 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.

7.14.2 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 require the **Nagios** configuration file to be regenerated using the command:

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

Note 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 the services have not been taken into account then enter the following commands:

```
/usr/lib/clustmngt/clusterdb/bin/nagiosConfig.pl -init  
dbmConfig configure --service nagios
```

Refer to http://nagios.sourceforge.net/docs/3_0/checkscheduling.html for more information on configuring the services.

7.14.2.1 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.

7.14.3 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.

7.14.4 Nagios Services Service

The **Nagios services** service monitors the daemons required for its own usage. If one of them is not up and running, this service will display the CRITICAL state and indicates which daemons are unavailable. The administrator must define a parameter stored in the `/etc/nsmhpc/nsmhpc.conf` file:

`nagios.services`, which defines the daemons which are monitored by the plugin (the default value is `syslog-ng snmpd snmptrapd`).

7.14.5 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/3_0/

In addition look at the **Bull System Manager - 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*.

7.15 General Nagios Services

Nagios includes a wide range of plug-ins, each of which provides a specific monitoring service that is displayed inside the graphical interface. In addition Bull has developed additional monitoring plug-ins which are included within **Bull System Manager – 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 applies to all forms of material/devices.

7.15.1 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**.

7.15.2 Resource Manager Status

The service reports the state of the node as seen by the Resource Manager (for example **SLURM**) which is in place. The service will be updated every time the state of the node changes.

7.15.3 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.

7.15.4 Alert Log

The **Log alerts** passive service displays the last alarm raised by system log for the machine – see *Section 7.7*. 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.

7.15.5 I/O Status

The I/O status reports the global status of HBA, disks and LUNs on a cluster node. Refer to Chapter 6, section *Monitoring Node I/O Status* for more information.

7.15.6 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.

7.16 Management Node Nagios Services

These services are available on the Management Node only.

7.16.1 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.

7.16.2 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.

7.16.3 ClusterDB

This active service uses the `check_clusterdb.pl` plug-in to check that connection to the Cluster 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.

7.16.4 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.

7.16.5 Compute Power Available

A Bull plug-in checks the compute 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 7-2.

This plug-in is specific to the **COMP** group of nodes created by the use of the `dbmConfig` command and which consists of all the Compute Nodes in the Cluster database. Note that Login nodes are considered as Compute Nodes in the **Clusterdb**, and if the Login nodes have not been defined in a Compute partition then the **COMP** group of nodes should be deleted by using the `dbmGroup modify` command – see section 3.3.6 in this guide for more information.

7.16.6 Global File System bandwidth available

A Bull plug-in checks the bandwidth for the Global File System, 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 7-2.

7.16.7 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 7-2.

7.16.8 Global File System Usage

A Bull plug-in checks Global File System 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 7-2.

7.16.9 I/O pairs Migration Alert

A Bull plug-in checks the I/O pairs status, 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 7-2.

7.16.10 Backbone Ports Available

This service calculates the percentage of ports which are usable for 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 7-2.

7.16.11 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**.

7.16.12 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 stops.

7.16.13 Kerberos Admin Daemon

This active service uses the plug-in **check_proc** to check if the **kadmind** 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 stops.

7.16.14 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.

7.16.15 Lustre file system 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 file systems 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).

7.16.16 NFS file system 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.

7.16.17 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 The *InfiniBand Guide* for more information regarding the **IBS** tool.

7.16.18 CMC Health

This active service uses the `check_cmc.pl` plug-in to check if the temperatures and the fans are running correctly for **bullx** blades. It remains at the **OK** alert level while the data is in the right range, but switches to **CRITICAL** if the data is out of the right range. The **WARNING** alert level is displayed if no data is available.

7.17 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 switches are first configured. The services for the switch are displayed when it is selected in either the cluster **HOSTGROUP** or **HOST** window, followed by the selection of **Service Status** window, as shown below.

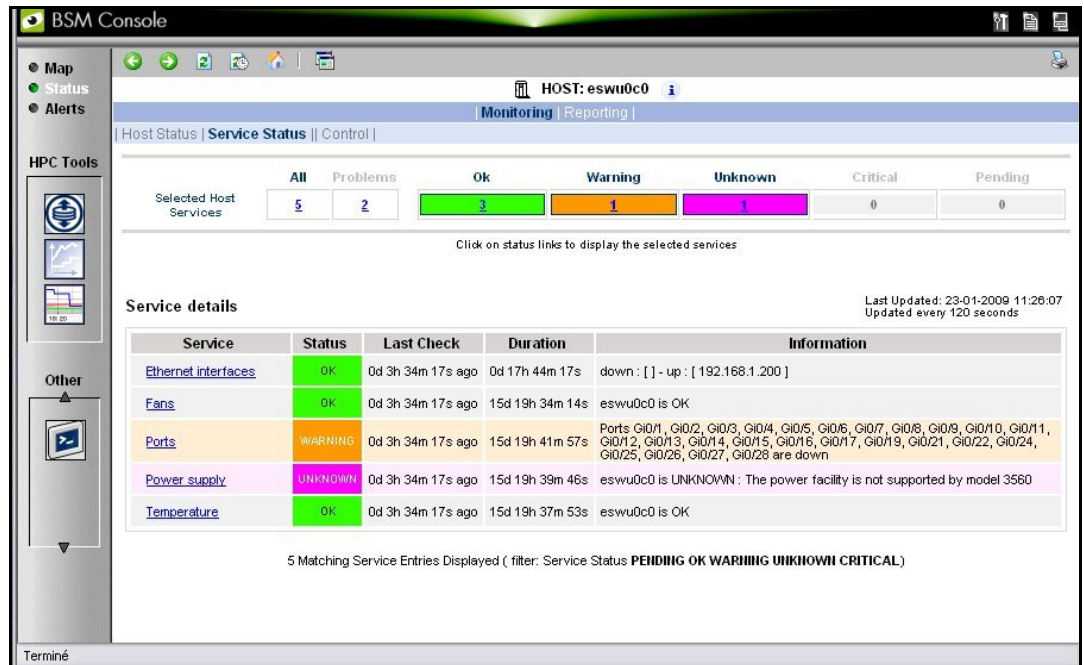


Figure 7-17. Ethernet Switch services

7.17.1 Ethernet Interfaces

The **Ethernet interfaces** service checks that the Ethernet switch is responding by using a ping to its IP address.

7.17.2 Fans

The **Fans** service monitors the fans for the Ethernet switches using the `check_esw_fans.pl` plug-in.

7.17.3 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.

7.17.4 Power supply

The **Power supply** service checks the power supply is functioning properly by using the `check_esw_power.pl` plug-in.

7.17.5 Temperature

The **Temperature** service monitors the temperatures of the Ethernet switches by using the `check_esw_temperature.pl` plug-in.

7.18 Cool Cabinet Door Services

Bull has developed a set of **Nagios** services to monitor the Cool Cabinet Door used to regulate the temperature for racks of servers. These are as follows:

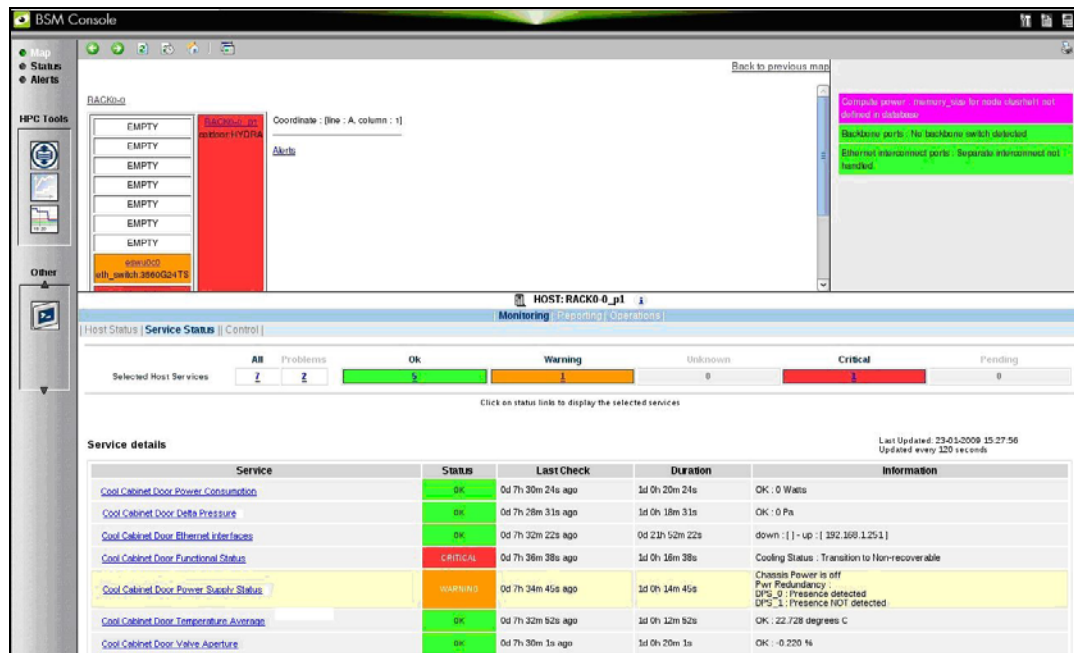


Figure 7-18. Cool Cabinet Door Services

7.18.1 Cool Cabinet Door Functional Status

The **Cool Cabinet Door Functional Status** service monitors the overall status of the cooled water door, with regard to temperature and water pressure through a **BSM CLI (IPMI)**. The status for the service will be displayed as **CRITICAL**, if the value returned is **Transition to Critical** or **Transition to Non-recoverable**. The status for the service will be displayed as **WARNING** when the value returned is **Transition to Non-Critical**.

7.18.2 Cool Cabinet Door Power Consumption

The **Cool Cabinet Door Power Consumption** service monitors the power being used by the Cool Cabinet door through a **BSM CLI (IPMI)**. The status for the service will be displayed as **CRITICAL**, if the value returned is **ABOVE** the **Upper-Critical** or **Upper Non-Recoverable** values, or if the value is **BELOW** the **Lower-Critical** or **Lower Non-Recoverable** values. A **WARNING** status is displayed, if the value is **ABOVE** the **Upper Non-Critical** value or is **BELOW** the **Lower Non-Critical** value.

7.18.3 Cool Cabinet Door Delta Pressure

The **Cool Cabinet Door Delta Pressure** service monitors the difference in water pressure between the water circulating inside the door and the water from the external water source through a **BSM CLI (IPMI)**. The status for the service will be displayed as **CRITICAL**, if the value is **ABOVE** the **Upper-Critical** or **Upper Non-Recoverable** values, or if the value is **BELOW** the **Lower-Critical** or **Lower Non-Recoverable** values. It displays the **WARNING**

state if the value is **BELOW** the **Lower Non-Critical** value or **ABOVE** the **Upper Non-Critical** value.

7.18.4 Cool Cabinet Door Ethernet Interfaces

The **Cool Cabinet Door Ethernet interfaces** service checks that the Cool Cabinet Door is responding by using a ping to its IP address.

7.18.5 Cool Cabinet Door Power Supply Status

The **Cool Cabinet Door Power Supply Status** service monitors the status of the power supply units, including the back-up units for the Cool Cabinet Door.

7.18.6 Cool Cabinet Door Temperature Average

The **Cool Cabinet Door Temperature Average** service monitors the average temperature of the water circulating within the Cool Cabinet Door through a **BSM CLI** (IPMI). The status for the service will be displayed as **CRITICAL**, if the value is **ABOVE** the **Upper-Critical** or **Upper Non-Recoverable** values, or if the value is **BELOW** the **Lower-Critical** or **Lower Non-Recoverable** values. It displays the **WARNING** state if the value is **BELOW** the **Lower Non-Critical** value or **ABOVE** the **Upper Non-Critical** value.

7.18.7 Cool Cabinet Door Valve Aperture

The **Cool Cabinet Door Valve Aperture** service monitors the degree (expressed as a percentage) for the valve opening for the water inlet through a **BSM CLI** (IPMI). The status for the service will be displayed as **CRITICAL**, if the value is **ABOVE** the **Upper-Critical** or **Upper Non-Recoverable** values, or if the value is **BELOW** the **Lower-Critical** or **Lower Non-Recoverable** values. It displays the **WARNING** state if the value is **BELOW** the **Lower Non-Critical** value or **ABOVE** the **Upper Non-Critical** value.

See The *Bull Cool Cabinet* documentation, listed in the *Bibliography* in the *Preface*, for more information regarding the Cool Cabinet Door.

Chapter 8. Managing PDUs

This chapter applies to clusters that include **APC Switched Rack PDU** equipment. It describes how to configure and monitor this equipment.

See the *APC Switched Rack PDU User's Guide* available from <http://www.apc.com> for more information.

8.1 Configuring PDUs

After installation of the **bullx cluster suite XR 5v3.1U2** on the Management Node, the following steps should be followed to configure the administration of the PDUs for the system.

8.1.1 Register the PDUs in the Cluster Database

Use the **equipmentRecord** command to register the PDUs in the Cluster Database.

PreRequisite

All the fields in the **talim** table, in the **ClusterDB**, for all PDUs should be completed except for the **admin_macaddress** field.

equipmentRecord command

Run this command as below:

```
# equipmentRecord pdu
```

Example Output

```
-----  
PDU talimu0c0-1 mac address 00:c0:b7:dX:XX:XX found  
PDU talimu0c0-0 mac address 00:c0:b7:dX:XX:XX found  
  
Wed Nov 18 16:12:15 2009 NOTICE: Begin synchro for sysdhcpd  
Shutting down dhcpd: [ OK ]  
Starting dhcpd: [ OK ]  
Wed Nov 18 16:12:15 2009 NOTICE: End synchro for sysdhcpd  
-----
```

In the example above the PDU MAC addresses are detected, and the **ClusterDB** database updated accordingly. Once this is done, the **DHCP** configuration and **/etc/hosts** files are synchronized with the Cluster Database, so that the PDU entries are taken into account.

Note The registration update may take a few minutes.

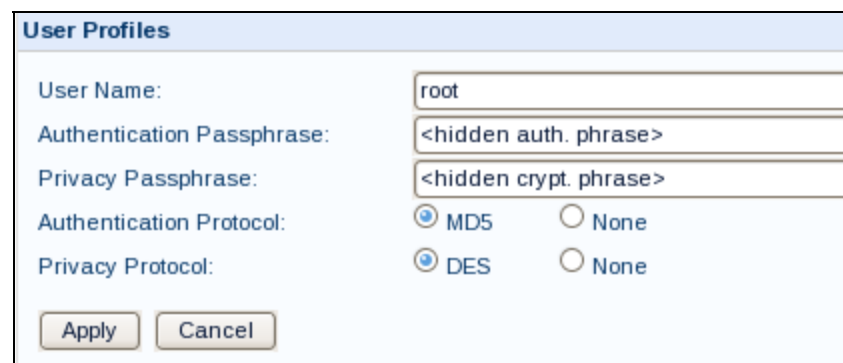
8.1.2 Accessing the PDUs

Check that the PDUs are able to answer **ping** requests and can be reached by their **http** interface, as well as by **telnet** session.

8.1.3 Customizing the Configuration of the PDUs

Only a few customizations are explained in this section, refer to the *APC Switched Rack PDU User's Guide* available from <http://www.apc.com> for more information.

- **Superuser name and password.**
These are defined in the Cluster Database and initially set to **apc/apc** (factory settings).
- **SNMP initial settings.**
When first installed the PDUs enable **SNMPv1** only. The groups defined are **public** (read) and **private** (write).
The initial settings are changed using the http interface for each PDU. This may also be done by connecting to the PDU via **FTP**, and then modifying and reloading its **config.ini** file. See the *APC User's Guide* for more details.
- **Setting SNMPv3 users profiles.**
SNMPv3 User Profiles are created using the interface below. The **Authentication Passphrase** and the **Privacy Passphrase** should include between 15 and 32 different ASCII characters. The System Administrator decides whether or not the **Authentication** and **Privacy Protocols** are set.



- **Setting Access Control.**
This is used to **enable/disable - read/write** access for SNMPv1 users.

 **Important** Do not disable the public read access type, as this is used for monitoring.

SNMPv3 users should have their control access enabled, when they are created. The creation of additional IP addresses for the Management Node is optional.

8.2 Monitoring PDUs

The PDUs are monitored by **Nagios** in **Bull System Manager**. Data for the power consumption is collected and displayed by Nagios, as below:

- **NAGIOS_OK_STATUS:** Measured power consumption is below **warning** overload threshold.
- **NAGIOS_WARNING_STATUS:** Measured power is over **warning** overload and below **critical** overload thresholds.
- **NAGIOS_CRITICAL_STATUS:** Measured power is over **critical** overload threshold or no power is detected.

Chapter 9. Cluster Power Management Tool

The Power Management tool allows the Administrator to:

- Calculate the global power consumption for the cluster or rack.
- To trigger actions when power consumption\temperature thresholds are reached.

The following features are provided:

Power Consumption

- Data for the power consumption of the blades and the PDU(s) is collected by Nagios and then stored in a Round Robin Database (RRD).
- The power consumption for the cluster is calculated using the power consumption data stored in the RRD.
- The power consumption for each rack is calculated using the power consumption data stored in the RRD.
- When the power consumption of the cluster reaches a set threshold, a mail is sent to the System Administrator.

Temperature

- The temperature details for the servers are collected by Nagios and then stored in a RRD.
- When a critical temperature threshold is reached on a Service Node, a mail is sent to the System Administrator.
- When a critical temperature threshold is reached on a blade, a mail is sent to the System Administrator and the blade is stopped.

The data stored in the Round Robin Databases can be visualized graphically.

9.1 Configuring Power Management for High Availability Clusters

Note This section refers to the *bullx cluster suite High Availability Guide*.



Important The changes to the HA configuration files have to be made before configuring High Availability, as described in the *High Availability Guide*.

1. Modify the **HA** configuration files on both Management Nodes as shown below.
 - a. Add the `/etc/init.d/powerManager` line to the following files:
 - `/etc/clustmngt/ha/start`
 - `/etc/clustmngt/ha/stop`
 - `/etc/clustmngt/ha/status`
 - b. Add the `powerManager` line to the `/etc/clustmngt/ha/noboot` file
 - c. Add the following lines to the `/etc/clustmngt/ha/synchro` file:

```
/etc/sec/powerManager/power_master.sec.conf
/etc/sec/powerManager/power_config
/etc/sec/powerManager/power_rack_config.template
```

2. Before carrying out the operations describe in Section 3.2.3 in the *High Availability Guide* run the following command:

```
mkfs.ext3 -L HA_MGMT:pnp4nagios /dev/sdxx
```

3. Add the following line to the `/etc/storageadmin/ha/hafsmgmt.conf` file:

```
LABEL=HA_MGMT: pnp4nagios /var/lib/pnp4nagios ext3
```

9.2 Configuring the Monitoring of Power Consumption

Following the installation of the **bullx cluster suite XR 5v3.1U2** packages, **nagios** has to be started again, using the command below:

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

9.2.1 Configuration

The **pnp4nagios** add-on gathers performance data from the **nagios** services and enters it into round robin databases (**rrd** files). The **rrd** files are stored in the `/var/lib/pnp4nagios/` directory.

Change the **process_perf_data** parameter in the `/etc/nagios/BSM/services-tpl.cfg` file, if there is a need to record, or to stop, the storage of power performance data in **rrd** files for specific components.

Example

```
# Node Power for INCA blades
define service {
name check-blade-power-tpl
use active-check-service
normal_check_interval 300
service_description Power Consumption
notification_period 24x7
check_command check_node_power
process_perf_data 1
register 0
}
```

If the value of **process_perf_data** parameter is set to **1**, **pnp4nagios** stores the data in a **rrd** file, and vice versa if the parameter is set to **0**.

Run the command below, following the modification of the `/etc/nagios/BSM/services-tpl.cfg` file.

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

9.2.2 Viewing Power Consumption Data

The **pnp4nagios** package includes a web GUI that can be used to display the **rrd** file data in a graphical format. The **rrd** graphs are accessed using the menu on the left of **Bull System Manager**.

9.3 Configuring Power Management

Three files are used to configure power management for the cluster:

- **/etc/sec/powerManager/power_config**
This file is used to define the various constants (such as thresholds) used for the consolidation of the power consumption figures for the whole cluster.
- **/etc/sec/powerManager/power_rack_config.template**
This file is used to define the various constants (such as thresholds) used for the consolidation of the power consumption figures for a rack.
- **/etc/sec/powerManager/power_master.sec.conf**
This file is used to define the following parameters for power management.
 - a. **Mail recipient**
Used to define the mail recipients to be automatically e-mailed by the power manager when there is an alarm.
 - b. **\$INCA_POWER_CONSUMPTION_CONSTANT**
This parameter represents the number of watts that should be added to the result of the consolidation figures for the **bullx** blades, in order to take the power consumption of the switches and of the **Chassis Management Controller (CMC)** into account. This value is configurable.

Note Some constants defined in the **power_config** and **power_rack_config.template** files are overridden by the values at the beginning of the **power_master.sec.conf** file (e.g. **node_names** and **rpn_suffix**). This allows the Power Manager to fix values for these constants at runtime, according to the cluster's hardware configuration (number of racks, number of **cmc/rack**).

9.4 Activate - Deactivate Power Management Tool

The Power Management tool must be activated, if your cluster is NOT Highly Available.

To activate the Power Management tool during the boot, run the following command:

```
chkconfig powerManagement on
```

To deactivate the Power Management tool during the boot, run the following command:

```
chkconfig powerManagement off
```

Chapter 10. CPU Frequency and Voltage Scaling

bullx and **NovaScale R4xxE1**, **R4xxE2** platforms support the **Enhanced Intel® Speedstep Technology (EIST)** that includes support for the **ACPI (Advanced Configuration and Power Interface) Performance States (P-states)**.

P-states are used to manage the performance/power consumption for the platform processors. The **P0** state corresponds to the base frequency for the processors (maximum performance). **P_i (i>0)** states correspond to a lower frequency/voltage usage resulting in reduced performance and power consumption.

Xeon 55xx series processors (**QPI** technology) introduce **Intel® Turbo Boost Technology**, which provides a new **P0** state called **Turbo mode** that increases the frequency (up to 3 x 133 MHz) of the processor dynamically beyond the base frequency (over-clocking), when possible according to the global load of the socket and the maximum temperature possible.

This new **P0 Turbo mode** state will be set as the base frequency + 1 MHz. When Turbo mode is activated, the **P1** state corresponds to the static base frequency of the processor.

10.1 Performance Requirements for Jobs

The **Bull Extreme Computing** environment has the following performance requirements:

- The provision of maximum performance for the computation phases.
- Generally, parallel jobs run the same load on all processors of the Compute Nodes, but are limited by the performance level of slowest processors in the cluster. As **Turbo mode** does not guarantee that all the processors increase their frequency uniformly, large parallel jobs may consume more power, without any improvement in performance.
- Performance measurements must be reproducible, that is the same for two runs of the same job. This is not possible when using **Turbo mode**.
- When the **Linux** scheduler enters an idle state, the processors enter **ACPI Power States (C-states)**, which provide a significant power saving.

For the above reasons, **bullx Extreme Computing** clusters are configured by default so that all the processors run at their fixed base frequency, to ensure optimal power management.

This default behaviour can be changed if the applications can handle the power consumption reduction provided by **P-states**, or the additional performance possible with **Turbo mode** for **Xeon 55xx** processors. Scripts may be used to ensure maximum performance (fixed base frequency or **Turbo mode**) when a job is started, and then minimum power consumption when the last job terminates (provided by the **ondemand** or **powersave** governors).

10.2 BIOS Power Management Settings

Intel EIST support and Turbo mode support have to be configured in the BIOS setup for NovaScale R4xxE2 and bullx platforms, using the settings below:

- Advanced-Advanced Processor Options-EIST <enable>
- Advanced-Advanced Processor Options-Turbo Mode <enable>

10.3 Managing Power for the Compute Nodes

The Power Management for the machines supported by bullx cluster suite XR 5v3.1 is organized as below:

1. The **cpufreq** service that manages the **CPUFreq** sub-system in the **Linux** kernel.
2. **CPUFreq** sub-system Tools
 - i. The **cpufreq-info** and **cpufreq-set** tools that manage the **CPUFreq** sub-sub-system directly.
 - ii. The **pwrgov** script to set/display the **governor** and **userspace** governor frequency for groups of processors for a machine.
3. The kernel **CPUFreq** driver interface that is accessed using the **/sys** pseudo file system.
4. The **Advanced Configuration and Power Interface (ACPI)** is used at a low level by the kernel to access the **P-states** provided by the platform.

10.3.1 cpuspeed service

By default, the bullx cluster suite XR 5v3.1 distribution does not start the **cpuspeed** service at boot time. This means that the processors will run at the designated base frequency.

Run the command below to start the **cpuspeed** service at boot time:

```
chkconfig cpuspeed on
```

See The **cpuspeed** man page for more information.

cpuspeed parameters

The following **cpuspeed** parameters can be set in the **cpuspeed** configuration file in the **/etc/sysconfig/cpuspeed** directory.

GOVERNOR=ondemand Allows the system to change the processor's frequencies according to the **CPU** load and to the default settings that are in place.



Important By default, the **ondemand** governor is used to manage the **CPUFreq** sub-system.

GOVERNOR=performance Forces the CPUs to use the maximum frequency that is available (**Turbo mode** if enabled).

Note Stopping the **cpuspeed** service will have the same effect.

GOVERNOR=powersave Forces the CPUs to use the minimum frequency that is available.

GOVERNOR=userspace Starts the **cpuspeed** daemon on each CPU. The daemon will manage the frequency on the CPUs according to the parameters in the **cpuspeed** configuration file.

Note The **cpuspeed** service will need to be restarted if the **GOVERNOR** parameter is changed.

10.3.2 CPUFreq Sub-System Tools

The **cpufreq-utils** RPM provides the **cpufreq-info** and **cpufreq-set** tools to display and set the **cpufreq** parameters, and the **pwrgov** script to set/display the governor and frequency for the **userspace** governor for all the processors of a machine.

See The **cpufreq-info** and **cpufreq-set** man pages for more information about these commands.

pwrgov script usage

pwrgov [<governor> [<frequency>]]

<governor> = powersave | ondemand | userspace | performance

if <governor> = userspace, <frequency> = frequency in Khz

Command Example

```
# pwrgov userspace 1995000
```

Output Example

```
-----
available governors:
powersave userspace ondemand performance
available frequencies:
2528000 2527000 2394000 2261000 2128000 1995000 1862000 1729000
1596000
current governors:
userspace 1995000
userspace 1995000
userspace 1995000
userspace 1995000
current frequencies:
1995000
1995000
1995000
1995000
-----
```

10.3.3 Changing CPUFreq Settings

The CPUFreq settings can be changed using the interface which is accessed via `/sys/devices/system/cpu/cpu<i>/cpufreq/` where `<i>` is the number of the processor, e.g. `/sys/devices/system/cpu/cpu0/cpufreq/` for the first CPU.



Important See the kernel sources/Documentation/cpu-freq/user-guide.txt for details of the CPUFreq parameters that are available for your system, and how they may be changed.

10.4 More Information

-
- See**
- The *Intel® 64 and IA-32 Architectures Software Developer's Manual Volume 3A: \$14 Thermal and Power management* available from <http://www.intel.com/products/processor/manuals/index.htm>
 - The *Advanced Configuration & Power Interface(ACPI) Specification* available from <http://developer.intel.com/technology/iapc/acpi/>
 - <http://www.kernel.org/pub/linux/utils/kernel/cpufreq/cpufreq.html>
-

Glossary and Acronyms

A

ABI

Application Binary Interface

ACL

Access Control List

ACPI

Advanced Configuration and Power Interface

ACT

Administration Configuration Tool

ANL

Argonne National Laboratory (MPICH2)

API

Application Programmer Interface

ARP

Address Resolution Protocol

ASIC

Application Specific Integrated Circuit

B

BAS

Bull Advanced Server

BIOS

Basic Input Output System

Blade

Thin server that is inserted in a blade chassis

BLACS

Basic Linear Algebra Communication Subprograms

BLAS

Basic Linear Algebra Subprograms

BMC

Baseboard Management Controller

BSBR

Bull System Backup Restore

BSM

Bull System Manager

C

CGI

Common Gateway Interface

CLI

Command Line Interface

ClusterDB

Cluster Database

CLM

Cluster Management

CMC

Chassis Management Controller

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.

CUBLAS

CUDA™ BLAS

CUDA™

Compute Unified Device Architecture

CUFFT

CUDA™ Fast Fourier Transform

CVS

Concurrent Versions System

Cygwin

A Linux-like environment for Windows. Bull cluster management tools use Cygwin to provide SSH support on a Windows system, enabling command mode access.

D**DDN**

Data Direct Networks

DDR

Double Data Rate

DHCP

Dynamic Host Configuration Protocol

DLID

Destination Local Identifier

DNS

Domain Name Server:

A server that retains the addresses and routing information for TCP/IP LAN users.

DSO

Dynamic Shared Object

E**EBP**

End Bad Packet Delimiter

ECT

Embedded Configuration Tool

EIP

Encapsulated IP

EIST

Enhanced Intel® Speedstep Technology

EPM

Errors per Million

EULA

End User License Agreement (Microsoft)

F**FDA**

Fibre Disk Array

FFT

Fast Fourier Transform

FFTW

Fastest Fourier Transform in the West

FRU

Field Replaceable Unit

FTP

File Transfer Protocol

G**Ganglia**

A distributed monitoring tool used to view information associated with a node, such as CPU load, memory consumption, and network load.

GCC

GNU C Compiler

GDB

Gnu Debugger

GFS

Global File System

GMP

GNU Multiprecision Library

GID

Group ID

GNU

GNU's Not Unix

GPL

General Public License

GPT

GUID Partition Table

Gratuitous ARP

A gratuitous ARP request is an Address Resolution Protocol request packet where the source and destination IP are both set to the IP of the machine issuing the packet and the destination MAC is the broadcast address `xx:xx:xx:xx:xx:xx`.

Ordinarily, no reply packet will occur. Gratuitous ARP reply is a reply to which no request has been made.

GSL

GNU Scientific Library

GT/s

Giga transfers per second

GUI

Graphical User Interface

GUID

Globally Unique Identifier

H**HBA**

Host Bus Adapter

HCA

Host Channel Adapter

HDD

Hard Disk Drive

HoQ

Head of Queue

HPC

High Performance Computing

Hyper-Threading

A technology that enables multi-threaded software applications to process threads in parallel, within each processor, resulting in increased utilization of processor resources.

IB

InfiniBand

IBTA

InfiniBand Trade Association

ICC

Intel C Compiler

IDE

Integrated Device Electronics

IFORT

Intel® Fortran Compiler

IMB

Intel MPI Benchmarks

INCA

Integrated Cluster Architecture:
Bull Blade platform

IOC

Input/Output Board Compact with 6 PCI Slots

IPMI

Intelligent Platform Management Interface

IPO

Interprocedural Optimization

IPoIB

Internet Protocol over InfiniBand

IPR

IP Router

iSM

Storage Manager (FDA storage systems)

ISV

Independent Software Vendor

K**KDC**

Key Distribution Centre

KSIS

Utility for Image Building and Deployment

KVM

Keyboard Video Mouse (allows the keyboard, video monitor and mouse to be connected to the node)

L**LAN**

Local Area Network

LAPACK

Linear Algebra PACKage

LDAP

Lightweight Directory Access Protocol

LDIF

LDAP Data Interchange Format:

A plain text data interchange format to represent LDAP directory contents and update requests. LDIF conveys directory content as a set of records, one record for each object (or entry). It represents update requests, such as Add, Modify, Delete, and Rename, as a set of records, one record for each update request.

LKCD

Linux Kernel Crash Dump:

A tool used to capture and analyze crash dumps.

LOV

Logical Object Volume

LSF

Load Sharing Facility

LUN

Logical Unit Number

LVM

Logical Volume Manager

LVS

Linux Virtual Server

M**MAC**

Media Access Control (a unique identifier address attached to most forms of networking equipment).

MAD

Management Datagram

Managed Switch

A switch with no management interface and/or configuration options.

MDS

MetaData Server

MDT

MetaData Target

MFT

Mellanox Firmware Tools

MIB

Management Information Base

MKL

Maths Kernel Library

MPD

MPI Process Daemons

MPFR

C library for multiple-precision, floating-point computations

MPI

Message Passing Interface

MTBF

Mean Time Between Failures

MTU

Maximum Transmission Unit

N**Nagios**

A tool used to monitor the services and resources of Bull HPC clusters.

NETCDF

Network Common Data Form

NFS

Network File System

NIC

Network Interface Card

NIS

Network Information Service

NS

NovaScale

NTP

Network Time Protocol

NUMA

Non Uniform Memory Access

NVRAM

Non Volatile Random Access Memory

O**OFA**

Open Fabrics Alliance

OFED

Open Fabrics Enterprise Distribution

OPMA

Open Platform Management Architecture

OpenSM

Open Subnet Manager

OpenIB

Open InfiniBand

OpenSSH

Open Source implementation of the SSH protocol

OSC

Object Storage Client

OSS

Object Storage Server

OST

Object Storage Target

P**PAM**

Platform Administration and Maintenance Software

PAPI

Performance Application Programming Interface

PBLAS

Parallel Basic Linear Algebra Subprograms

PBS

Portable Batch System

PCI

Peripheral Component Interconnect (Intel)

PDSH

Parallel Distributed Shell

PDU

Power Distribution Unit

PETSc

Portable, Extensible Toolkit for Scientific Computation

PGAPACK

Parallel Genetic Algorithm Package

PM

Performance Manager

Platform Management

PMI

Process Management Interface

PMU

Performance Monitoring Unit

pNETCDF

Parallel NetCDF (Network Common Data Form)

PVFS

Parallel Virtual File System

Q**QDR**

Quad Data Rate

QoS

Quality of Service:

A set of rules which guarantee a defined level of quality in terms of transmission rates, error rates, and other characteristics for a network.

R**RAID**

Redundant Array of Independent Disks

RDMA

Remote Direct Memory Access

ROM

Read Only Memory

RPC

Remote Procedure Call

RPM

RPM Package Manager

RSA

Rivest, Shamir and Adleman, the developers of the RSA public key cryptosystem

S**SA**

Subnet Agent

SAFTE

SCSI Accessible Fault Tolerant Enclosures

SAN

Storage Area Network

SCALAPACK

SCALable Linear Algebra PACKage

SCSI

Small Computer System Interface

SCIPOPT

Portable implementation of CRAY SCILIB

SDP

Socket Direct Protocol

SDPOIB

Sockets Direct Protocol over Infiniband

SDR

Sensor Data Record

Single Data Rate

SFP

Small Form-factor Pluggable transceiver - extractable optical or electrical transmitter/receiver module.

SEL

System Event Log

SIOH

Server Input/Output Hub

SIS

System Installation Suite

SL

Service Level

SL2VL

Service Level to Virtual Lane

SLURM

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

SM

Subnet Manager

SMP

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

SNMP

Simple Network Management Protocol

SOL

Serial Over LAN

SPOF

Single Point of Failure

SSH

Secure Shell

Syslog-ng

System Log New Generation

T

TCL

Tool Command Language

TCP

Transmission Control Protocol

TFTP

Trivial File Transfer Protocol

TGT

Ticket-Granting Ticket

U

UDP

User Datagram Protocol

UID

User ID

ULP

Upper Layer Protocol

USB

Universal Serial Bus

UTC

Coordinated Universal Time

V

VCRC

Variant Cyclic Redundancy Check

VDM

Voltaire Device Manager

VFM

Voltaire Fabric Manager

VGA

Video Graphic Adapter

VL

Virtual Lane

VLAN

Virtual Local Area Network

VNC

Virtual Network Computing:
Used to enable access to Windows systems and Windows applications from the Bull NovaScale cluster management system.

W

WWPN

World-Wide Port Name

X

XFS

eXtended File System

XHPC

Xeon High Performance Computing

XIB

Xeon InfiniBand

XRC

Extended Reliable Connection:
Included in Mellanox ConnectX HCAs for memory
scalability

Index

/

- /etc/init.d/powerManager, 9-1
- /etc/krb5.conf, 5-2
- /etc/nagios/BSM/services-tpl.cfg, 9-2
- /etc/nagios/contactgroups.cfg, 7-16
- /etc/nagios/contacts.cfg, 7-16
- /etc/nagios/snmptargets.cfg, 7-16
- /etc/nsmhpc/nsmhpc.conf, 7-16
- /etc/storageadmin/storframework.conf, 6-30
- /var/kerberos/krb5kdc/kadm5.acl, 5-4
- /var/kerberos/krb5kdc/kdc.conf, 5-3
- /var/log/postgres/pgsql, 3-24
- /var/log/synchro.log file, 3-5

A

- ACPI Performance States, 10-1
- administrator
 - postgres (ClusterDB), 3-2
 - root, 2-1
- APC Switched Rack PDU, 8-1
- authorized_keys2 file, 2-2

B

- Backbone ports available alert, 7-25
- Bull System Manager - HPC Edition, 7-1
 - Acknowledgements, 7-15
 - Active checks, 7-13
 - Alert definition, 7-14
 - Alert levels, 7-13
 - Alert types, 7-12
 - All status map view, 7-6
 - bullx blade map view, 7-7
 - Changing passwords, 7-3
 - Comments, 7-16
 - Ganglia, 7-18
 - Global Performance view, 7-19
 - Group Performance view, 7-18
 - Management node Nagios Services Map button, 7-6

- Monitoring performance, 7-18
- Nagios Alert log, 7-23
- Nagios Ethernet interfaces, 7-23
- Nagios IO Status, 7-23
- Nagios logs, 7-11
- Nagios plug-ins, 7-23
- Nagios postbootchecker, 7-23
- Nagios Services, 7-21
- Passive checks, 7-14
- Ping Map view, 7-9
- Rack view, 7-7
- Shell button, 7-18
- SNMP Alerts, 7-16
- Status Button, 7-11
- Storage overview, 7-17
- User password, 7-3

C

- chkconfig command, 2-1

ClusterDB

- administrator (postgres), 3-2
- ChangeOwnerProperties, 3-2
- cluster features, 3-6
- Commands, 3-2
- dbmCluster command, 3-6
- dbmConfig, 3-5
- dbmDiskArray command, 3-20
- dbmEthernet command, 3-13
- dbmFiberChannel command, 3-18
- dbmHwManager command, 3-10
- dbmlconnect command, 3-14
- dbmNode command, 3-7
- dbmSerial command, 3-17
- dbmServices command, 3-19
- Description, 3-1
- managing groups, 3-11
- monitoring, 7-24
- PostgreSQL tools, 3-22
- requisite, 7-2
- save and restore, 3-22
- template files, 3-6

ClusterDB tables

- ADMIN table, 3-46
- AVAILABILITY table, 3-49
- CLUSTER table, 3-27
- CLUSTER_IPV table, 3-30
- CONFIG_CANDIDATE table, 3-47

CONFIG_STATUS table, 3-47
 da_cfg_model table, 3-38
 da_controller table, 3-35
 da_enclosure table, 3-34
 da_ethernet_port table, 3-36
 da_fan table, 3-36
 da_fc_port table, 3-35
 da_io_path table, 3-37
 da_iocell_component table, 3-37
 da_power_fan table, 3-37
 da_power_port table, 3-38
 da_power_supply table, 3-36
 da_serial_port table, 3-35
 da_temperature_sensor table, 3-37
 disk_array table, 3-34
 disk_slot table, 3-34
 ETH_EXTRALINK table, 3-31
 ETH_SWITCH table, 3-28
 ETH_VLAN table, 3-30
 FC_NW table, 3-30
 FC_SWITCH table, 3-31
 GROUP_NODE table, 3-48
 HWMANAGER table, 3-45
 IC_BOARD table, 3-43
 IC_NW table, 3-28
 IC_SWITCH table, 3-29
 IP_NW table, 3-27
 IPOIB table, 3-43
 Lustre_fs table, 3-50
 Lustre_IO_node table, 3-51
 Lustre_MDT table, 3-51
 Lustre_mount table, 3-52
 Lustre_OST table, 3-51
 MSG_SYSLOG table, 3-48
 Node table, 3-41
 Node_image table, 3-42
 Node_profile table, 3-42
 PORTSERVER table, 3-29
 RACK table, 3-46
 RACK_PORT table, 3-47
 SDPOIB table, 3-43, 3-44
 SERIAL_NW table, 3-29
 SERVICES table, 3-48
 TALIM table, 3-31

Commands
 ChangeOwnerProperties, 3-2
 chkconfig, 2-1
 dbmCluster, 3-6
 dbmConfig, 3-5
 dbmDiskArray, 3-20
 dbmEthernet, 3-13
 dbmFiberChannel, 3-18
 dbmGroup, 3-11
 dbmHwManager, 3-10
 dbmlconnect, 3-14
 dbmNode, 3-7
 dbmSerial, 3-17
 dbmServices, 3-19
 dbmTalim, 3-16
 ddn_set_up_date_time, 6-23
 ddn_admin, 6-22
 ddn_check, 6-23
 ddn_conchk, 6-23
 ddn_firmup, 6-23
 ddn_init, 6-22
 ddn_stat, 6-22
 dshbak, 2-3
 iorefmgmt, 6-5
 kadmin, 5-3
 lsiodev, 6-4
 nec_admin, 6-21
 passwd, 2-1
 pdcp, 2-3
 pdsh, 2-3
 stormodelctl, 6-28
 storstat, 6-2, 6-17
 useradd, 2-1

connectivity status, 6-11
 contact groups
 adding, 7-16
 contacts
 adding, 7-16
 controller status, 6-11

D

dbmCluster command, 3-6
 dbmConfig command, 3-5
 dbmDiskArray command, 3-20
 dbmEthernet command, 3-13
 dbmFiberChannel command, 3-18
 dbmGroup command, 3-11
 dbmHwManager command, 3-10
 dbmlconnect command, 3-14
 dbmNode command, 3-7

- dbmSerial command, 3-17
- dbmServices command, 3-19
- dbmTalim command, 3-16
- DDN commands, 6-22
- ddn_set_up_date_time command, 6-23
- ddn_admin command, 6-22
- ddn_check command, 6-23
- ddn_conchk command, 6-23
- ddn_firmup command, 6-23
- ddn_init command, 6-22
- ddn_stat command, 6-22
- deploying software See Ksis
- distributed shell, 2-3
- distribution
 - changing, 4-1
 - updating, 4-1
- distribution software, 4-1
- dropdb command, 3-23
- dshbak command, 2-3

F

- fan status, 6-10
- files
 - /etc/init.d/powerManager, 9-1
 - /etc/nagios/BSM/services-tpl.cfg, 9-2
 - /etc/nagios/contactgroups.cfg, 7-16
 - /etc/nagios/contacts.cfg, 7-16
 - /etc/nagios/snmptargets.cfg, 7-16
 - /etc/nsmhpc/nsmhpc.conf, 7-16
 - /var/log/synchro.log, 3-5
 - authorized_keys2, 2-2
 - genders, 2-4
 - id_dsa.pub, 2-2
 - kadm5.acl, 5-4
 - res_rpm_qsnetmpi, 2-6
 - storframework.conf, 6-30
 - template.model, 6-27

G

- Ganglia
 - data categories, 7-19

- Ganglia
 - Bull System Manager - HPC Edition, 7-1
- genders file, 2-4
- groups of nodes, 3-11

H

- HDD status, 6-9
- HPC Toolkit, 1-2

I

- id_dsa.pub file, 2-2
- image
 - list, 3-7
- InfiniBand links available, 7-26
- iorefmgmt command, 6-5

K

- Kerberos, 1-2, 2-3, 5-1
 - Access Control List, 5-4
 - Admin Daemon, 5-4
 - configuration files, 5-2
 - database, 5-3
 - Host principal, 5-5
 - kadmin command, 5-3
 - KDC, 5-1
 - package, 5-2
 - SSH, 5-7
 - TGT ticket, 5-7

- Kerberos admin daemon, 7-25

- Kerberos KDC daemon, 7-25

- Ksis

- builddatanode command, 4-13
- check command, 4-4, 4-12
- checkdiff command, 4-6, 4-12
- checks database, 4-6
- client node, 4-3
- command file, 4-6
- command options, 4-7
- create commands, 4-8
- delete command, 4-8
- deploy command, 4-9
- help command, 4-7
- image server, 4-1, 4-3
- import command, 4-12
- Ksis server, 4-1

- list command, 4-9
- nodelist command, 4-10
- nodeRange, 4-7
- overview, 1-2, 4-1
- reference node, 4-3
- reference/golden image, 4-1, 4-2
- undeploy command, 4-9

L

- LDAP daemon, 7-25
- linux user, 2-1
- lsiodev command, 6-4
- Lustre filesystems access, 7-26

M

- maintenance tools, 2-7
- MetaData Server migration alert, 7-25
- MiniSQL daemon, 7-24
- model
 - file, 6-27
 - storage system configuration, 6-26
- monitoring the cluster, 7-1

N

- Nagios
 - Contact groups, 7-4
 - Hosts, 7-8
 - Services, 7-4, 7-8
- Nagios
 - Bull System Manager - HPC Edition, 7-1
- Nagios, 8-2
- Nagios Management node plug-ins
 - ClusterDB, 7-24
 - Cron Daemon, 7-24
 - MiniSQL Daemon, 7-24
- Nagios plug-ins
 - Backbone ports available, 7-25
 - Ethernet Switch services, 7-27, 7-29
 - HA system status, 7-25
 - InfiniBand links available, 7-26
 - Kerberos admin daemon, 7-25
 - Kerberos KDC daemon, 7-25
 - LDAP daemon, 7-25
 - Lustre filesystems access, 7-26

- Metadata servers, 7-25
- NFS filesystems access, 7-26

- nec_admin command, 6-21
- nec_admin.conf file, 6-21
- NFS filesystems access, 7-26
- node list, 3-7

O

- oid2name command, 3-23
- OpenSSH, 2-2

P

- parallel commands, 2-3
- passwd command, 2-1
- password
 - user, 2-1
- PBS Professional, 1-2
- pdcp command, 2-3
- pdsh, 1-2
- pdsh command, 2-3
- PDU, 8-1
- pg_dump command, 3-22
- pg_restore command, 3-22
- phpPgAdmin, 3-22
- pnp4nagios, 9-2
- postbootchecker, 7-23
- postgres user, 3-2
- PostgreSQL, 3-22
- Power Consumption, 9-1
- Power Management, 9-1
- power supply status, 6-10
- predefined groups, 3-11
- psql command, 3-22

R

- res_rpm_qsnetmpi file, 2-6
- resource management, 1-2

root user, 2-1

rsh, 2-3

S

security

 Kerberos, 5-1

 policies, 2-2

service

 list, 2-1

 star), 2-1

shell

 distributed, 2-3

 kerberos, 2-3

 pdsh, 2-3

 rsh, 2-3

 ssh, 2-3

SLURM, 1-2

SNMP trap

 response to alert, 7-16

software distribution, 4-1

software update, 4-1

ssh, 2-3

 setting up, 2-2

storage device

 configuration deployment, 6-3

 configuration files, 6-30

 configuration planning, 6-25

 management services, 6-2

 managing, 6-1

 monitoring, using Nagios, 6-8

stormodelctl command, 6-28

storstat command, 6-2, 6-17

system image, 3-7

system status, 6-12

T

Temperature, 9-1

temperature status, 6-10

template.model file, 6-27

U

user

 create, 2-1

 password, 2-1

useradd command, 2-1

V

view

 inventory of storage systems and components,
 6-14

 storage, 6-12

 storage tactical overview, 6-12

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

REFERENCE
86 A2 20FA 03