HPC

BAS5 for Xeon

Maintenance Guide

REFERENCE
86 A2 24FA 00

HPC

BAS5 for Xeon

Maintenance Guide

Hardware and Software

March 2009

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

REFERENCE
86 A2 24FA 00

Trademarks and Acknowledgements

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

Preface

Intended Readers

This guide is intended for use by qualified personnel, in charge of maintaining and troubleshooting the Bull HPC clusters of NovaScale R4xx nodes, based on Intel® Xeon® processors.

Prerequisites

Readers need a basic understanding of the hardware and software components that make up a Bull HPC cluster, and are advised to read the documentation listed in the Bibliography below.

Bibliography

Refer to the manuals included on the documentation CD delivered with you system OR download the latest manuals for your Bull Advanced Server **(BAS)** release, and for your cluster hardware, from: http://support.bull.com/

The Bull *BAS5 for Xeon Documentation* CD-ROM (86 A2 12FB) includes the following manuals:

- Bull *HPC BAS5 for Xeon Installation and Configuration Guide* (86 A2 19FA).
- Bull *HPC BAS5 for Xeon Administrator's Guide* (86 A2 20FA).
- Bull *HPC BAS5 for Xeon User's Guide* (86 A2 22FA).
- Bull *HPC BAS5 for Xeon Maintenance Guide* (86 A2 24FA).
- Bull *HPC BAS5 for Xeon Application Tuning Guide* (86 A2 23FA).
- Bull *HPC BAS5 for Xeon High Availability Guide* (86 A2 25FA).

The following document is delivered separately:

- The *Software Release Bulletin* (SRB) (86 A2 68EJ)

mportant

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

In addition, refer to the following:

- Bull *Voltaire Switches Documentation CD* (86 A2 79ET)
- *Bull System Manager* documentation

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

- PBS Professional *10.0 Administrator's Guide* (on the *PBS Professional CD-ROM)*
- PBS Professional *10.0 User's Guide* (on the *PBS Professional CD-ROM*)

For clusters which use **LSF**:

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

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 Door Installation Guide (86 A1 20EV)
- Cool Cabinet Door Console User's Guide (86 A1 41FA)
- Cool Cabinet Door 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.

CAUTION
A *Caution* notice indicates the presence of a hazard that has the potential of causing moderate or minor personal injury.

Table of Contents

List of Figures

List of Tables

Chapter 1. Stopping/Restarting Procedures

This chapter describes procedures for stopping and restarting Bull HPC cluster components, which are mainly used for maintenance purposes.

The following procedures are described:

- 1.1 *Stopping/Restarting a Node*
- 1.2 *Stopping/Restarting an Ethernet Switch*
- 1.3 *Stopping/Restarting a Backbone Switch*
- 1.4 *Stopping/Restarting the Bull Cool Cabinet Door*
- 1.5 *Stopping/Restarting the HPC Cluster*
- 1.6 *Checking Nodes after the Boot* Phase

1.1 Stopping/Restarting a Node

1.1.1 Stopping a Node

Follow these steps to stop a node:

1. Stop the customer's environment. Check that the node is not running any applications by using the SINFO command on the management node. All customer applications and connections should be stopped or closed including shells and mount points.

2. Un-mount the file system.

3. Stop the node:
   From the Management Node enter:

```
nsctrl poweroff <node_name>
```

   This command executes an Operating System (OS) command. If the OS is not responding it is possible to use:

```
nsctrl poweroff_force <node_name>
```

   Wait for the command to complete.

4. Check the node status by using:

```
nsctrl status <node_name>
```

The node can now be examined, and any problems which may exist diagnosed and repaired.

1.1.2 Restarting a Node

To restart a node, enter the following command from the Management Node:

```
nsctrl poweron <node_name>
```

Note If during the boot operation the system detects an error (temperature or otherwise), the node will be prevented from rebooting.

Check the node status

Make sure that the node is functioning correctly, especially if you have restarted the node after a crash:

- Check the status of the services that must be started during the boot. (The list of these services is in the /etc/rc.d file).
- Check the status of the processes that must be started by a **cron** command.
- The mail server, **syslog-ng** and **ClusterDB** must be working.
- Check any error messages that the mails and log files may contain.

Restart SLURM and the filesystems

If the previous checks are successful, reconfigure the node for **SLURM** and restart the file systems.

1.2 Stopping/Restarting an Ethernet Switch

- Power-off the Ethernet switch to stop it.
- Power-on the Ethernet switch to start it.
- If an Ethernet switch must be replaced, the MAC address of the new switch must be set in the Cluster Database. This is done as follows:

1. Obtain the MAC address for the switch (generally written on the switch, or found by looking at **DHCP** logs).
2. Use the **phpPgAdmin** Web interface of the DATABASE to update the switch MAC address (http://IPadressofthemanagementnode/phpPgAdmin/ user=`clusterdb` and password=`clusterdb`).
3. In the **eth_switch** table look for the **admin_macaddr** row in the line corresponding to the name of your switch. Edit and update this MAC address. Save your changes.
4. Run a **dbmConfig** command from the management node:

```
dbmConfig configure --service sysdhcpd --force -nodeps
```

5. Power-off the Ethernet switch.
6. Power-on the Ethernet switch.

The switch issues a **DHCP** request and loads its configuration from the management node.

See The **BAS5 for Xeon** *Administrator's Guide* for information about how to change the management of the Cluster Database.

1.3 Stopping/Restarting a Backbone Switch

The backbone switches enable communication between the cluster and the external world. They are not listed in the **ClusterDB**. It is not possible to use **ACT** for their reconfiguration.

1.4 Stopping/Restarting the Bull Cool Cabinet Door

1.4.1 Using the GUI of the Bull Cool Cabinet Door

Use the **GUI** Console of the Bull Cool Cabinet Door to power on/off the Cool Cabinet Door.

Important

Check the power off/on states of hardware equipment included in the rack before stopping/starting the Bull Cool Cabinet Door, in order to avoid overheating issues.

See The *Cool Cabinet Door Console User's Guide* for details about the GUI console.

1.4.2 Using nsclusterstart and nsclusterstop Commands

The Bull Cool Cabinet Doors are stopped / started when the **nsclusterstart** / **nsclusterstop** commands are used to stop /start the HPC Cluster, as it is the case for the cluster nodes.

See Section 1.5 *Stopping/Restarting the HPC Cluster* for more information.

1.5 Stopping/Restarting the HPC Cluster

The **nsclusterstop/nsclusterstart** scripts are used to stop or start the whole HPC cluster. These scripts launch the various steps, in sequence, making it possible to stop/start the cluster in complete safety. For example, the stop process includes the following steps:

1. Checking the various equipment,
2. Stopping the file systems (Lustre for example),
3. Stopping the storage devices,
4. Stopping the nodes, except the Management Node(s),
5. Stopping the Bull Cool Cabinet Doors if any.

Important

In order to ensure that the hardware has time to cool down, delays are included for the stopping sequence for the HPC Cluster. Before stopping/starting the HPC Cluster the power off/on status for the hardware in the cabinet should be checked to ensure there is no risk of overheating.

1.5.1 Stopping the HPC Cluster

To stop the whole cluster in complete safety it is necessary to launch the different steps in sequence. The **nsclusterstop** script includes all the required steps.

1. From the Management Node, run:

```
nsclusterstop
```

2. Stop the Management Node.

1.5.2 Starting the HPC Cluster

To start the whole cluster in complete safety it is necessary to launch different stages in sequence. The **nsclusterstart** script includes all the required stages.

1. Start the Management Node.

2. From the Management Node, run:

```
nsclusterstart
```

1.5.3 Configuring and Using nsclusterstop and nsclusterstart

The **nsclusterstop** and **nsclusterstart** commands use configuration files to define:

- The delay parameters between the different stages required to stop/start the cluster
- The sequence in which the group of nodes should be stopped/started. (You can run **dmbGroup show** to display the groups configured.)

By default the configuration files are respectively **/etc/clustmngt/nsclusterstop.conf** and **/etc/clustmngt/nsclusterstart.conf**. The **--file** option allows you to specify another configuration file.

Usage:

/usr/sbin/nsclusterstop [-h] | [-f, --file <filename>]

/usr/sbin/nsclusterstart [-h] | [-f, --file <filename>]

Options:

--file <filename>, -f Specify a configuration file (default: **/etc/clustmngt/nsclusterstop.conf** and **/etc/clustmngt/nsclusterstart.conf**).

-h Display **nsclusterstart/nsclusterstop** help.

--only_test , -o Display the commands that would be launched according to the specified options. This is a testing mode, no action is performed.

--verbose, -v Verbose mode.

/etc/clustmngt/nsclusterstart.conf Configuration file

```
##############################################################################
#
# First Part is used to control the power on safety delay for the Cool Cabinet
Door
#
##############################################################################

# time to wait for the Cool Cabinet Doors to be started
coldoor_StartDelay = 30

##############################################################################
#
# Second Part is used to control the power supply of DDN and servers
#
##############################################################################

# time to wait for all diskarrays ok, before powering the powerswitches on
disk_arrays_StartDelay = 300

# time to wait for all powerswitches being ON after a poweron

couplets_StartDelay = 60

# time to wait after poweron for all servers being effectively operational
servers_StartDelay = 480
```

```
##############################################################################
#
# Following part is used to control the order to start nodes groups
#
##############################################################################

# GROUP <nb simultaneous poweron> <time to wait> <period to wait> <time to
wait after this GROUP>
IO 5 1 5 5
META 5 1 5 5
COMP 5 1 5 5
```

/etc/clustmngt/nsclusterstop.conf Configuration file

```
##############################################################################
#
# First Part is used to control the power off safety delay for the Cool
Cabinet Doors
#
##############################################################################

# time to wait for before the Cool Cabinet Doors are stopped
coldoor_StopDelay = 120

##############################################################################
#
# Second Part is used to controls the power supply of DDN and servers
#
##############################################################################

# time to wait after poweroff for all servers being effectively down
servers_StopDelay = 180

# time to wait for ddn processing shutdown
ddnShutdown_Time = 180

# time to wait after poweroff for all powerswitches being OFF
couplets_StopDelay = 30

#############################################################################
#
# Following part is used to control the order to stop nodes groups
#
##############################################################################

# GROUP <nb simultaneous poweron> <time to wait> <period to wait> <time to
wait after this GROUP>
COMP 5 1 5 5
META 5 1 5 5
IO 5 1 5 5
```

1.6 Checking Nodes after the Boot Phase

This section describes how to use **postbootchecker** to check nodes after boot. **postbootchecker** detects when a Compute Node is starting and runs check operations on this node after its boot phase. The objective is to verify that **CPU** and memory parameters are coherent with the values stored in the **ClusterDB**, and if necessary to update the **ClusterDB** with the real values.

1.6.1 Prerequisites

- **syslog-ng** must be installed and configured as follows:
  - Management Node: management of the logs coming from the cluster nodes.
  - Compute nodes: detection of the compute nodes as they start.
- The **postbootchecker** service must be installed before the RMS service, to avoid jobs being disturbed.

1.6.2 Checking the Compute Nodes

The **postbootchecker** service (**/etc/init.d/postbootchecker**) detects every time a Compute Node starts. Whilst the node is starting up, **postbootchecker** runs three scripts to retrieve information about processors and memory. These scripts are the following:

Script name	Description
procTest.pl	Retrieves the number of CPUs available for the node.
memTest.pl	Retrieves the size of memory available for the node.
modelTest.pl	Retrieves model information for the CPUs available on the node.

Then **postbootchecker** returns this information to the Management Node using **syslog-ng**.

1.6.3 Checking the Management Node

On the Management Node, the **postbootchecker** server gets information returned from the Compute Nodes and compares it with information stored in the **ClusterDB**:

- The number of CPUs available for a node is compared with the **nb_cpu_total** value in the ClusterDB.
- The size of memory available for a node is compared with the **memory_size** value in the ClusterDB.
- The CPUs model type for a node is compared with the **cpu_model** value in the ClusterDB.

If discrepancies are found, the ClusterDB is updated with the new values. In addition, the Nagios status of the **postbootchecker** service is updated as follows:

- If the discrepancies concern the number of CPUs or the memory size the service is set to **CRITICAL**.
- If the discrepancies concern the model of the CPUs the service is set to **WARNING**.
- If no discrepancies were found, the service is **OK**.

Chapter 2. Administrating the Cluster

This chapter describes the following procedures:

- 2.1 *Managing Consoles through Serial Connections (conman, ipmitool)*
- 2.2 *Managing Hardware (nsctrl, BSM Commands)*
- 2.3 *Collecting Information for Resolving Problems*

2.1 Managing Consoles through Serial Connections (conman, ipmitool)

The serial lines of the servers are the communication channel to the firmware and enable access to the low-level features of the system. This is why they play an important role in the system **init** surveillance, or in taking control if there is a crash or a debugging operation is undertaken.

The serial lines are brought together with Ethernet/Serial port concentrators, so that they are available from the Management Node.

- **ConMan** can be used as a console management tool.
  See 2.1.1 *Using ConMan*.
- **ipmitool** allows you to use a Serial Over Lan (**SOL**) link.
  See 2.1.2 *Using ipmi Tools*.

Note Storage Units may also provide console interfaces through serial ports, allowing configuration and diagnostics operations.

2.1.1 Using ConMan

The **ConMan** command allows the administrator to manage all the consoles, including server consoles and storage subsystem consoles, on all the nodes. It maintains a connection with all the lines that it administers. It provides access to the consoles and uses a logical name. It supports the key sequences that provide access to debuggers or to dump captures (Crash/Dump).

ConMan is installed on the Management Node.

The advantages of ConMan on a simple telnet connection are as follows:

- Symbolic names are mapped per physical serial line.
- There is a log file for each machine.
- It is possible to join a console session or to take it over.
- There are three modes for accessing the console: monitor (read-only), interactive(read-write), broadcast(write only).

Syntax:

conman <OPTIONS> <CONSOLES>

-b	Broadcast to multiple consoles (write-only).
-d HOST	Specify server destination. [127.0.0.1:7890]
-e CHAR	Specify escape character. [&]
-f	Force connection (console-stealing).
-F FILE	Read console names from file.
-h	Display this help file.
-j	Join connection (console-sharing).
-l FILE	Log connection output to file.
-L	Display license information.
-m	Monitor connection (read-only).
-q	Query server about specified console(s).
-Q	Be quiet and suppress informational messages.
-r	Match console names via regex instead of globbing.
-v	Be verbose.
-V	Display version information.

Once a connection is established, enter "&." to close the session, or "&?" to display a list of currently available escape sequences.

See the **conman** man page for more information.

Examples:

- To connect to the serial port of NovaScale `bull47`, run the command:

```
conman bull47
```

Configuration File:

The **/etc/conman.conf** file is the conman configuration file. It lists the consoles managed by conman and configuration parameters.

The **/etc/conman.conf** file is automatically generated from the ClusterDB information. To change some parameters, the administrator should only modify the **/etc/conman-tpl.conf** template file, which is used by the system to generate **/etc/conman.conf**. It is also possible to use the **dbmConfig** command. See the *Cluster Data Base Management* chapter for more details.

See the **conman.conf** man page for more information.

Note The **timestamp** parameter, which specifies the watchdog frequency, is set to 1 minute by default. This value is suitable for debugging and tracking purposes but generates a lot of messages in the **/var/log/conman** file. To disable this function, comment the line `SERVER timestamp=1m` in the **/etc/conman-tpl.cfg** file.

2.1.2 Using ipmi Tools

The **ipmitool** command provides a simple command-line interface to the **BMC** (Baseboard Management Controller).
To use **SOL** (Serial Over Lan) interface, run the following command:

```
ipmitool –I lanplus –C O –U <BMC_user_name> –P <BMC_password>
–H <BMC_IP_Address> sol activate
```

`BMC_user_name`, `BMC_password` and `BMC_IP_Address` are values defined during the configuration of the BMC and are taken from those in the **ClusterDB**. The standard values for user name/password are `administrator/administrator`.

ipmitool Command Useful Options

Note If –H is not specified, the command will address the BMC of the local machine.

- To start a remote SOL session (to access the console):

```
ipmitool -I lanplus –C 0 -H <ip addr> sol activate
```

- To reset the BMC and return to BMC shell prompt:

```
ipmitool -I lanplus –C 0 -H <ip addr> bmc reset cold
```

- To edit the FRU of the machine:

```
ipmitool -H <ip addr> fru print
```

- To edit the network configuration:

```
ipmitool -I lan -H <ip_addr> lan print 1
```

- To trigger a dump (signal INIT):

```
ipmitool -H <ip addr> power diag
```

- To power down the machine:

```
ipmitool -H <ip addr> power off
```

- To perform a hard reset:

```
ipmitool -H <ip addr> power reset
```

- To display the events recorded in the System Event Log (SEL):

```
ipmitool -H <ip addr> sel list
```

- To display the MAC address of the BMC:

```
ipmitool -I lan -H <ip addr> raw 0x06 0x52 0x0f 0xa0 0x06 0x08 0xef
```

- To know more about the **ipmitool** command, enter:

```
ipmitool –h
```

2.2 Managing Hardware (nsctrl, BSM Commands)

2.2.1 Using nsctrl

The **nsctrl** command carries out various tasks related to hardware. This command must be run from the Management Node. The tasks can be performed on any type of node (Compute Node, I/O Node, etc.) except the Management Node.

Usage:

```
/usr/sbin/nsctrl [options] <action> [<nodes>]
```

General Options:

--debug	Debug mode (more than verbose).
--dbname name	Specify database name.
--force, -f	Do not ask for confirmation or state checking.
--group, -g	Specify a group of nodes. You can use the **dbmGroup show** command to display the defined groups.
--help, -h	Display **nsctrl** help.
--interval, -i	Specify the number of bsm calls before waiting the period defined by the **--time** option.
--jobs, -j	Number of simultaneous bsm actions (for example, with -j 5 you can run 5 simultaneous **bsmpower** processes). Default = 30.
--only_test, -o	Display the BS Commands that would be launched according to the specified options and action. This is a testing mode, no action is performed.
--time, -t	Time to wait after the number of bsm calls defined by the **--interval** option.
--verbose, -v	Verbose mode.

Specifying nodes:

The nodes are specified as follows: **basename[i,j-k]** .
If no nodes are explicitly specified, **nsctrl** uses the nodes defined by the **--pap** or **--group** option.

Actions:

poweron
poweroff
poweroff_force
reset
status
ping

Examples:

Note In the following examples the **–o** option (**--only_test**) is used to display which BSM commands would be launched for the specified action.

- To power off node `ns1`, enter:

```
nsctrl -o poweroff_force ns1
```

```
ns1 : /opt/BSMHW/bin/bsmpower.sh -a off_force -m ipmilan -H ns1 -u
user2
```

- To ping node `ns1`, enter:

```
# nsctrl -o ping ns1
```

```
ns1 : ping -c 1  ns1
```

2.2.2 Using Remote Hardware Management CLI (BSM Commands)

The Remote Hardware Management **CLI** (Command Line Interface) is a set of commands that perform hardware tasks on Bull HPC, these are also known as BSM Commands. These commands provide the administrator with an easy way to automate scripts to power on/off and to get hardware information about the nodes.

See Refer to **Bull System Manager** documentation for more information about Remote Hardware Management CLI.

2.3 Collecting Information for Resolving Problems

The **hpcsnap** tool collects HPC cluster information. This information can be sent to Bull Support and used for problem analysis.

To facilitate the analysis, the **hpclog** tool can extract lines corresponding to a given period.

The package is delivered as an alpha version (0.2.5) in the *BAS5 for Xeon v3.1 XHPC* DVD-ROM.

Installation

```
yum localinstall /release/XBAS5V<version_nb>/XHPC/BONUS/hpcsnap-0.2.5-0.noarch.rpm
```

Usage

Using **hpcsnap** with default options:

```
# cd /usr/local/hpcsnap.0.2.5

# ./hpcsnap
```

If you have opened a request to Bull Support System (D1 or PARM ticket) use the ticket number as TAG parameter with the **–t** option, as follows:

```
# ./hpcsnap -t <TAG>
```

To list all available options enter:

```
# ./hpcsnap -h
```

Reporting Problems

If you find problems using **hpcsnap**, please send an email to team-linux@support.frec.bull.fr

Chapter 3. Managing System Logs

This chapter describes syslog-ng and how to configure it.

3.1 Introduction to syslog-ng

For security and tracking purposes, and also to decrease the amount of administration work resulting from the size of the cluster, all the system logs are centralized on the Management Node. There are two ways to send system log information to the Management Node:

- The logs are collected on each node, using standard mechanisms for archival and log file permutation. Various utilities ensure compression, transfer and archival of these log files on the Management Node in asynchronous mode. A centralized operation is performed on the Management Node, in order to extract and search events according to the criterion required for example date, type, gravity, and so on.

  This asynchronous process facilitates curative actions for the incidents that have occurred on the cluster.

- Some events are immediately reported to the Management Node. Filters are used, which specify the type and gravity level of the events that have to be transferred immediately.

  This synchronous process instantaneously gives the administrator a global view of system events.

syslog-ng (Syslog New Generation) is the powerful system log manager used on Bull HPC clusters to manage cluster system logs and includes the following features:

- The ability to filter messages based on content using regular expressions.
- Encoding and authentication of the network traffic.
- Forwarding logs using TCP and UDP protocols.
- Log compression.

3.2 Configuring syslog-ng

syslog-ng is installed on the cluster using the default configuration. The scripts used to transfer log files are also installed. The administrators can modify the default configuration according to their needs.

The **/etc/syslog-ng/syslog-ng.conf** file contains the configuration parameters for syslog-ng. This file is divided into five sections:

options section	General options
source section	Source events
destination section	Log destinations
filter section	Filter definitions
log section	Actions to be performed on messages

3.2.1 options Section

Any general parameters may be configured in the options section. An example is below:

```
# Start of options area
options {
 sync (0);      # Number of events before writing in the logs
 time_reopen (10);   # Wait 10s before reconnecting if the connection
     failed. Used when logs are centralized through network
 #time_reap (number);# Closes a log file that is not accessed after
     "number" seconds
 log_fifo_size (1000); # number of event lines stored, before writing them.
     Enables events to be taken quickly into account
     and to free the process that has generated them.
 long_hostnames (off);  # Usage of long names
 use_dns (no)   # Usage of DNS to find addresses
 use_fqdn (no); # Usage of machine short name
 owner("root"); # logs owner
 group("root"); # logs group
 perm("644");   # logs rights mask
 keep_hostname (yes);#
 create_dir (yes);   # Create directories for log storage
 use_time_recv(no);  # Local time will be used instead of the time written
in the logs
 #gc_idle_threshold(100); # The garbage collector is started after 100
       events if syslog-ng is inactive.
 #gc_busy_threshold(100); # The garbage collector is started after 3000
events if syslog-ng is active.
};
```

3.2.2 source Section

The source section defines the log source from the following: network, local files, peripheral, pipe, stream.

Syntax:

source <identifier>
{source-driver(params); source-driver(params); etc.};

For example, the following lines are suitable for a Linux system. They enable the **/dev/log** stream to be read and also to receive syslog-ng internal messages and to handle kernel starting messages:

```
source src {
unix-stream("/dev/log");
internal();
file("/proc/kmsg");
};
```

Possible sources are as follows:

unix-stream(<filename>)	Stream pipes (used in Linux).
file(<filename>)	File data (Linux kernel messages for example).
pipe(<filename>)	Named pipes (for interfacing with Nagios for example).
tcp(<ip>,<port>) and **udp(<ip>,<port>)**	To listen on an address and a port.
internal()	syslog-ng internal messages.

3.2.3 destination Section

This section defines the destination of the logs.

Syntax:

destination <identifier>
{ destination-driver(params); destination-driver(params); etc.};

The possible destinations are the following ones:

file(<filename>) To send to a file.

tcp(<ip>,<port>) and udp(<ip>,<port>)
To send the logs on the network to another machine.

unix-stream(<filename>) To send to stream pipes (used in Linux).

userttyr(<user>) To send to the <user > consoles, but only if this user is connected. You can use the "*" character to specify that the messages have to be sent to all users.

program(<commandtorun>) To send towards a program.

Examples :

You can specify several destination directives in a destination section, as in the following example:

```
destination debug {file("/var/log/debug.log"); };
destination messages {file("/var/log/messages.log"); };
destination console {usertty("root"); };
destination xconsole {pipe("/dev/xconsole"); };
destination mail2admin {program("/usr/bin/MailToAdmin"); };
destination full{
file("/dev/tty12");
file("/var/log/full.log" log_fifo_size(2000));
};
```

Note You can add specific options such as `log_fifo_size(2000)` as shown in the example above.

In the following example, all the logs will be sent to the Management Node, whose address is `192.168.0.100`:

```
destination central_log {tcp ("192.168.0.100" port(514); }
```

Using Macros:

It may be useful to use macros to set intelligible names for your destination files. Predefined macros exist, such as FACILITY, PRIORITY or LEVEL, DATE, FULLDATE, ISODATE, YEAR, MONTH, DAY, HOUR, MIN, SEC, FULLHOST, HOST. Some examples are below:

```
destination full {
file("/dev/tty12");
file("/var/log/full_$DAY-$MONTH-$YEAR.log"
owner("root")
group("adm")
perm(0640));
};
```

```
destination hosts {
file("/var/log/HOSTS/$HOSTS/$FACILITY/$YEAR/$MONTH/$DAY/$FACILITY$YEAR
$MONTH$DAY"
owner("root")
group("adm")
perm(0600)
dir_perm(0700)
create_dirs(yes));
};
```

Note Do not forget to remove or archive older files regularly.

3.2.4 filter Section

This section describes the filtering mechanism for events.

Syntax:

filter <identifier> {expression; };

The filters are defined by the following keywords:

facility(facility[,facility]) To filter by type.

level(pri[,pri1, .. pri2 [,pri3]]) To filter by priority or level.

program(regexp) To filter by the name of the program that has generated the message.

host(regexp) To filter by the regular expression of the name of the host that has sent the message.

match(regexp) To filter by a regular expression.

filter(filtername) To use another filter.

All keywords may be used several times. The expressions can contain the AND, OR and NOT operators.

Examples:

```
filter f_iptables { match("IN=.*OUT=.*MAC=.*"); };
filter f_snort { match("snort: "); };
filter f_full { not filter(f_snort) AND NOT filter(f_iptables); };
filter f_messages { level(info..warn) AND NOT facility(auth, authpriv,
mail, news); };
```

3.2.5 log Section

In this section you define how the messages will be processed using source, destination and filters commands defined in the previous sections.

Syntax:

log { source(s1); source(s2); ...
filter(f1); filter(f2); ...
destination(d1); destination(d2);
flags(flag1[, flag2...]; };

Examples:

```
log { source(src);
filter(f_news); filter(f_notice);
destination(newsnotice);
};
log { source(src);
destination(full);
};
```

Chapter 4. Saving and Restoring the System

This chapter describes **BSBR** (Bull System Backup Restore) and the following tasks:

- *4.1 Introduction to BSBR*
- *4.2 Installing BSBR*
- *4.3 Configuring BSBR*
- *4.4 Backing up a system*
- *4.5 Restoring a System*
- *4.6 Optional - for Lustre clusters only*

4.1 Introduction to BSBR

To save and restore the Management Node system, use **BSBR** (Bull System Backup Restore). **BSBR** is based on the **mkCDrec** (make recovery CD-ROM) Open Source tool, and is used to create a bootable Linux system image so that the system can be restored following a problem, such as a disk crash or a system intrusion.

BSBR is available on the *Bull Extension Pack* CD delivered with the **Red Hat** media.

See The Bull *System Backup/Restore User's Guide* (86 A2 73EV) available on the *Bull Extension Pack* CD for more information about installing and using the product.

Important

- **BSBR** is designed to back up the operating system in place on a node
  BSBR should NOT be used for data backups. A different method should be used for this.
- The disk used for the restoration must not be smaller than the original disk and must have the same identification (e.g. /dev/sdx)
- This section highlights the points that must be kept in mind when **BSBR** is used in a **HPC** environment.

The system backups are saved on **DVDs** or on **NFS** mounted disks or tape. It is possible to restore the system by either booting on the DVD or by booting on the network. To be able to boot on the network the system must have been saved via **NFS**.

See Section 5.3 in the Bull *System Backup/Restore User's Guide* for more information on booting on the network. In particular refer to section 5.3.1 which indicates the need to use lower case text when entering the MAC address details.

The Management Node system files should be backed up regularly once **Bull Advanced Server** has been installed on the Management Node, and the different node image files have been deployed.

4.2 Installing BSBR

BSBR uses the RPM included on the *Bull Extension Pack* CD. To install **BSBR** insert the *Bull Extension Pack* CD in the drive and type the following commands:

```
cd /<mntdir>/tools/mkcdrec
.install.sh
```

where `<mntdir>` = the mount directory for the DVD/CD (see **/etc/fstab**).

Note Ignore the Warning message related to **Webmin**.

4.3 Configuring BSBR

The **/var/opt/mkcdrec/Config.sh** file contains the configuration parameters for **BSBR**. All parameters have a default value. By default, BSBR is configured to save either on DVD or on the network via **NFS**.

The following values should be checked, either to verify that they fit your needs, or to define values specific to your cluster.

ISOFS_DIR The temporary directory where all the files to be backed up are stored.
The default is **/tmp/backup**
WARNING: The content of the ISOFS_DIR will be deleted when the make clean command is used.

CDREC_ISO_DIR The location where 'CDrec.iso' ISO9660 images will be made.
Default is **/tmp**

KERNEL_APPEND Add the **nmi_watchdog=0** parameter in the variable, as in the following example:
`KERNEL_APPEND = “nmi_watchdog=0”`

EXCLUDE_LIST List of directories which should be excluded for the backup. Obviously, these directories cannot be restored later.

The default includes the **/tmp/***, **/proc/***, **/mnt/*** directories.
IMPORTANT
Assuming that /test is a directory:
- "/test" will exclude all items in /test, also /test will NOT exist upon restore.
- "/test/*" will exclude all items in the directory /test, but /test will be created upon restore.

It is essential to exclude as many directories as possible in order to reduce the number of DVDs used for the backup.
For Bull **HPC** Clusters the following directories can be excluded:

- **/release**
- **/tmp/*** (excluded by default)
- **/mnt/*** (excluded by default)
- **/test** (if it exists)

If the **KSIS** images do not need to be saved, you can also exclude:

- /var/lib/systemimager/images/*
- /var/lib/systemimager/scripts/*
- /var/lib/systemimager/overrides/*

Note The RPMs that are installed on the Management Node are in the **/release** directory. It is not necessary to save this directory because these RPMs can be retrieved from the installation CDs.

4.4 Backing up a system

4.4.1 Un-mount the Mounted Drives

It is recommended to un-mount the mounted drives, assuming the mounted data does not need to be saved.

4.4.2 Stop Services

Important

All activity on the Management Node must be stopped when creating the backup. The ClusterDB must not be used during the backup operation.

The following services should be stopped before running **BSBR**.

- Lustre
- ganglia
- postgresql

Lustre File System

1. Stop Lustre activity

   Ensure **Lustre** is stopped correctly for all **Lustre** file systems:

```
lustre_util umount -f <fsname> -n <client nodes list | all>
lustre_util stop -f <fsname>
```

2. Stop Cluster Suite - Lustre High Availability Clusters only

   a. If necessary, relocate the **Lustre** services onto their Primary Node by using the commands below:

```
lustre_migrate hastat -n <io_node_list>
lustre_migrate relocate -n <node>
```

   b. Stop the **Lustre** services:

```
lustre_migrate hastop -n <io_node_list>
```

c. Stop **Cluster Suite**:

```
stordepha -c stop -i <all | node list>
```

3. Stop the Lustre daemons and save the backend files

a. **MGS** service

```
service mgs stop
```

The **MGS** back-end can be saved at this point. However, this is optional as **MGS** is able to rebuild itself when **Lustre** starts. The back-end file is configured in the **/etc/lustre/lustre.cfg** file using the command below:

```
grep LUSTRE_MGS_ABSOLUTE_LOOPBACK_FILENAME /etc/lustre/lustre.cfg
```

b. **Ldap** - Lustre High Availability Clusters only

```
service ldap stop
```

The back-end files are in the **/var/lib/ldap/lustre** folder.

c. **Lustredbd** - Lustre High Availability Clusters only

```
service lustredbd.sh stop
```

There is no backend file for this daemon.

Ganglia

To stop the **ganglia** service, run the following commands:

```
service gmond stop
service gmetad stop
```

Postgresql

To stop the **postgresql** service, run the following commands:

```
service postgresql stop
```

Important **Do not forget to restart these services once the system has been backed up.**

4.4.3 Creating the Backup

Carry out these operations on the Management Node.

1. Log on as the root user (preferably in single mode).

2. Go to the base directory, by default this is **/var/opt/mkcdrec**:

```
cd /var/opt/mkcdrec
```

3. Check that the system is operational:

```
make test
```

Warning messages are displayed if some elements are missing for the backup. If this happens, make the appropriate corrections and restart **make test** until the test is successful.

Note Ignore the "`/bin/mt not found`" warning message issued by **test2** if there is no tape drive.

4. Launch the backup operation:

```
bsbr
```

A menu is displayed, similar to that below:

```
Enter your selection:
 1) Create rescue CD-ROM only (no backups)
 2) Create ISO backup images in /tmp
   (to burn on CDROM or DVD)
 3) Create backup on disk
   (mounted harf disk, NFS mount point, SMB mount point)
 4) Create backup on tape device /dev/nst0
 5) Quit

Please choose from the above list [1-5]:
```

Select one of the options displayed (1 to 5) and follow the instructions that are displayed on the screen.

5. The ISO images to be burnt to DVD will be created in the directory specified in the configuration file (**CDREC_ISO_DIR** parameter). By default this is **/tmp**.

Note Check the **mkcdrec.log** file in case of problems.

6. It is recommended to burn the ISO image on DVD immediately. Refer to the *Bull System Backup / Restore User's Guide* for details on how this should be done.

4.5 Restoring a System

To restore a system, boot on the first DVD-ROM. An automatic procedure starts, which is sufficient for most cases.

For more control over the restoration procedure, you can stop the automatic restoration by pressing **Enter** when the following message is displayed:

```
Automatic Disaster Recovery (AUTODR) Mode is active !
Press “Enter” key to interrupt AUTODR mode (within 20 seconds)…
```

Then launch the restore manually using the following commands:

```
cd /etc/recovery
./start-restore.sh
```

When the restore is completed, reboot the machine using the **reboot** command.

4.6 Optional - for Lustre clusters only

If you are restoring the Management Node and redeploying the I/O node images, carry out all the actions listed below.

If you are restoring the Management Node and are NOT redeploying the I/O node images, carry out the actions below from point 3.

1. **Regenerate the disknaming.conf file on each I/O node.**
   Run the command, below, on the Management Node.

```
stordepmap -m </etc/storageadmin/models/model file> -p -c
```

   Quit this step only when the **stormap** -l command indicates all **I/O** node devices are **UP**.

2. **Distribute the lustre.cfg file onto the I/O nodes**

   Use the **lustre_util set_cfg** command to do this.

3. **Start the Lustre daemons and test SSH connectivity.**

   a. Test **SSH** connectivity by running the command below:

```
pdsh -w <IO node list> "ssh <management node> echo 'OK'" | dshbak -c
```

   If there is a problem with **SSH** reconfigure it so that it works.

   b. Launch the **MGS** service
   Restore the **MGS** backend, as and when needed, by running the command:

```
service mgs start
```

If the Management Node is NOT Highly Available, add the **MGS** service to the **chkconfig** file:

```
chkconfig --add mgs
```

c. Launch the **LDAP** service - **Lustre High Availability** Clusters only.

Restore the **LDAP** backend, as and when needed, by running the command:

```
service ldap start
```

Verify the **LDAP** content by running the command:

```
lustre_ldap show
```

This command will show details of the High Availability **Lustre** file systems that are installed. If the Management Node is **NOT** Highly Available, add the **LDAP** service to the **chkconfig** file:

```
chkconfig --add ldap
```

d. Launch **lustredbd** - **Lustre High Availability** Clusters only.

```
service lustredbd.sh start
```

If the Management Node is NOT Highly Available, add the **lustredbd** to the **chkconfig** file:

```
chkconfig --add lustredbd.sh
```

4. **Lustre High Availability Clusters only - Setup and start Cluster Suite.**

a. Regenerate the **Cluster Suite** configuration files:

Clusters without Quorum Disk

i. Run the command below to regenerate the **Cluster Suite** configuration files:

```
stordepha -c configure -i <all | IO node list>
```

OR

Clusters with Quorum Disk

ii. If the cluster includes **Quorum disk** partitions run the command below to regenerate the Cluster Suite configuration files:

```
stordepha -q -c configure -i <all | IO node list>
```

b. Start the **Cluster Suite** daemons:

```
stordepha -c start -i <all | IO node list>
```

c. Start the **Lustre** High Availability services:

```
lustre_migrate hastart -n <all | IO node list>
```

5. Start Lustre - All Lustre clusters

   a. Start **Lustre**:

```
lustre_util start -f <fsname> [-V]
```

   b. Mount the **Lustre** clients:

```
lustre_util mount -f <fsname> -n <all | client nodes list> [-V]
```

Chapter 5. Monitoring Devices

This chapter describes the monitoring for the following devices:

- 5.1 *Checking the status of InfiniBand Networks* (ibstatus, ibstat)
- 5.2 *Diagnosing InfiniBand Fabric* Problems (IBS tool)
- 5.3 *Monitoring Voltaire Switches*
- 5.4 *Getting Information about Storage Devices (lsiocfg)*
- 5.5 *Checking Device Power State* (pingcheck)
- 5.6 *Setting Up Outlet Air Temperature*

5.1 Checking the status of InfiniBand Networks (ibstatus, ibstat)

5.1.1 ibstatus Command

ibstatus displays basic information obtained from each **InfiniBand** driver for the local adapter included in an **InfiniBand** network.

Normal output includes LID, Subnet Manager LID, port state (UP or DOWN), port physical state and the link width in terms of transfer rate. **-v** enable verbose mode which includes all **sysfs** supported parameters for the port interface and port.

Syntax:

ibstatus [-h] [devname[:port]]...

Examples:

- To display status of all IB ports, enter:

```
ibstatus
```

- To display status of mthca1 ports, enter:

```
ibstatus mthca1
```

- To show status of specified ports, enter:

```
ibstatus mthca1:1 mthca0:2
```

Output example for a mthca dual port HCA

```
Infiniband device 'mthca0' port 1 status:
        default gid:        fe80:0000:0000:0000:0008:f104:0397:7ca5
        base lid:          0x0
        sm lid:            0x0
        state:             1: DOWN
        phys state:        2: Polling
        rate:              2.5 Gb/sec (1X)
```

```
Infiniband device 'mthca0' port 2 status:
        default gid:      fe80:0000:0000:0000:0008:f104:0397:7ca6
        base lid:         0x2d
        sm lid:           0x3
        state:            4: ACTIVE
        phys state:       5: LinkUp
        rate:             10 Gb/sec (4X)
```

5.1.2 ibstat Command

ibstat works in a similar fashion to the **ibstatus** utility but is implemented as a binaries and not a script, and is more useful than **ibstatus** as more detailed information is provided. It includes options to list Channel Adapters and/or Ports.

Syntax:

ibstat [-d(ebug) -l(ist_of_cas) -p(orts_list) -s(hort)] <ca_name> [portnum]

ibstat command examples:

- To display status of all IB ports, enter:

```
ibstat
```

- To display status of mthca1 ports, enter:

```
ibstat mthca1
```

- To show status of specified ports, enter:

```
ibstat mthca1 2
```

- To list the port guids of mthca0, enter:

```
ibstat -p mthca0
```

- To list all CA names, enter:

```
ibstat -l
```

5.2 Diagnosing InfiniBand Fabric Problems (IBS tool)

This tool is used from the Management Node to diagnose problems for **InfiniBand** fabric using the cluster switch topology information contained in the **NetworkMap.xml** file, and the error checking counters contained in the **PortCounters.csv** file. Alternatively, an IBS database, **IBSDB**, containing all the switch information can be created and then used as the data source to diagnose the problems

5.2.1 IBS Command Syntax

ibs -a <action> [-hvCNE] [-l |-s <switch>] [-f <networkmap-file>] [-c <counters-file>]

-h Help file

-v Verbose mode

-C Disable colored text output

-a Action (one of: **topo**, **bandwidth**, **errors**, **config**, **group**, **dbpopulate**, **availability**, **dbcreate**, **dbdelete**, **dbupdate**, **dbupdatepc**).

OFED related options

When working from the cluster Management Node, and provided this node is fitted with an **InfiniBand** adapter that is connected to an InfiniBand interconnect, it is recommended that the **–N** and **–E** options are used as the OFED software view of the cluster is more reliable than that provided by data taken directly from the switch.

-N Query the IB subnet manager to obtain and update the hostname details.

-E Query the IB subnet manager to obtain and update data using the error and traffic counters.

Data related options

By default IBS analyses the data contained in the IBSDB database unless the **–s** or **–l** flags are used. This default mode is known as 'database mode'.

-s <switch> 'Connected mode'. Connect to the switch specified by its hostname or IP address and then retrieve the **NetworkMap.xml** and **PortCounters.csv** files for this switch.

-l 'Local mode'. Use the **NetworkMap.xml** and **PortCounters.csv** files that are available locally or that are specified by the **-f** and **-c** flags for the analysis. These files can then be analysed separately on a machine which is not part of the cluster. However, as stated above it is better to work within the OFED stack using the **–N** and **–E** options to obtain the latest data.

-f <networkmap-file> Specify the file to be used when loading or saving the network map file, **NetworkMap.xml**. When used in conjunction with the **-s switch** option, the file downloaded from the switch will be saved to file **<filename>**. When used in conjunction with the **-l** flag, the specified file will be used as the input file.

-c <counters-file> Specify the file to be used when loading or saving the port counters file (**PortCounters.csv** file). When used in conjunction with the **-s** switch option, the file downloaded from the switch will be saved to the file **<filename>**. When used in conjunction with the -l flag, the specified file will be used as the input file.

topo action

The **topo** action for the **– a** option provides detailed topology details for the switch.

```
ibs -s <switch_name> -a topo -NE
```

This will give output that includes a description of the switches, the hostnames, the GUID for the Nodes, the LID for the Nodes, the physical location of the switches. The port details, including any errors, are shown in the bottom half of the screen for both local ports and for ports which are connected to remotely – see the screen example on Figure 5-1

Use the command below to obtain the fabric topology using the data stored in the IBS database. The hostnames and traffic counters are updated using the OFED tools:

```
ibs -a topo -NE
```

Use the command below to dump the fabric topology using the local map file test/**NetworkMap.xml** and test/**portcounters.csv**. The data read from these files is updated using the OFED tools:

```
ibs -l -f test/NetworkMap.xml -c test/portcounters.csv -a topo -NE
```

bandwidth action

The syntax for the bandwidth action is shown below. This action is very useful when benchmarking in order to monitor the performance of switch and to identify any bottlenecks.

```
ibs -s <switch_name> -a bandwidth -NE
```

Details of packets sent and received for the switch for both local and remote connections are displayed, as shown in Figure 5-2.

errors action

The errors action can be used to produce a short report containing details of the faulty links for a switch. This is very useful for troubleshooting and will help to pinpoint any problems for the interconnects.

```
ibs -s <switch_name> -a errors -NE
```

This will give output, similar to that shown in Figure 5-3. **EPM** indicates the error rate in the form of Errors per Million packets sent.

See FAQ ID – F10040 *"How to debug and clear InfiniBand fabric errors using FVM PM Counters CSV file?"* available from www.voltaire.com for details of the different Port Counter error messages.

config action

This action manually creates the instruction sequence needed to configure the hostname mapping for a switch.

Note This option only applies to Voltaire switches which use 4.0 or later firmware versions.

```
ibs -s <switch_name> -vNE -a config
```

group action

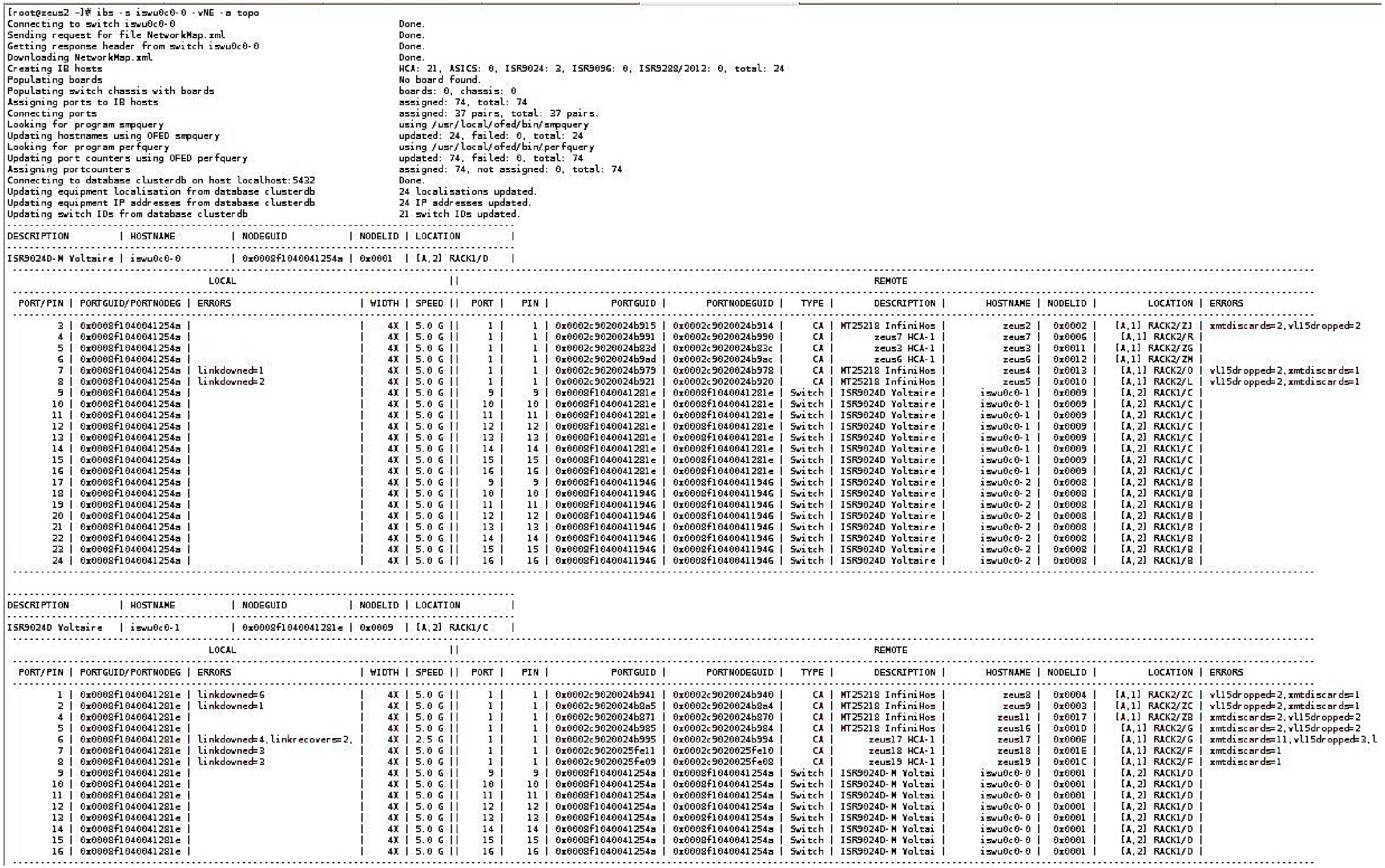
This action generates the **group.csv** file that includes the hostname mapping configuration details for all the switches, this can then be imported into a switch in order to configure it. For large clusters, this is quicker than running the **config** action (as detailed above), to generate and import the cluster switch configuration details into a switch.

Note This option only applies to Voltaire switches which use 4.0 or later firmware versions.

```
ibs -s iswu0c0-0 -a group
```

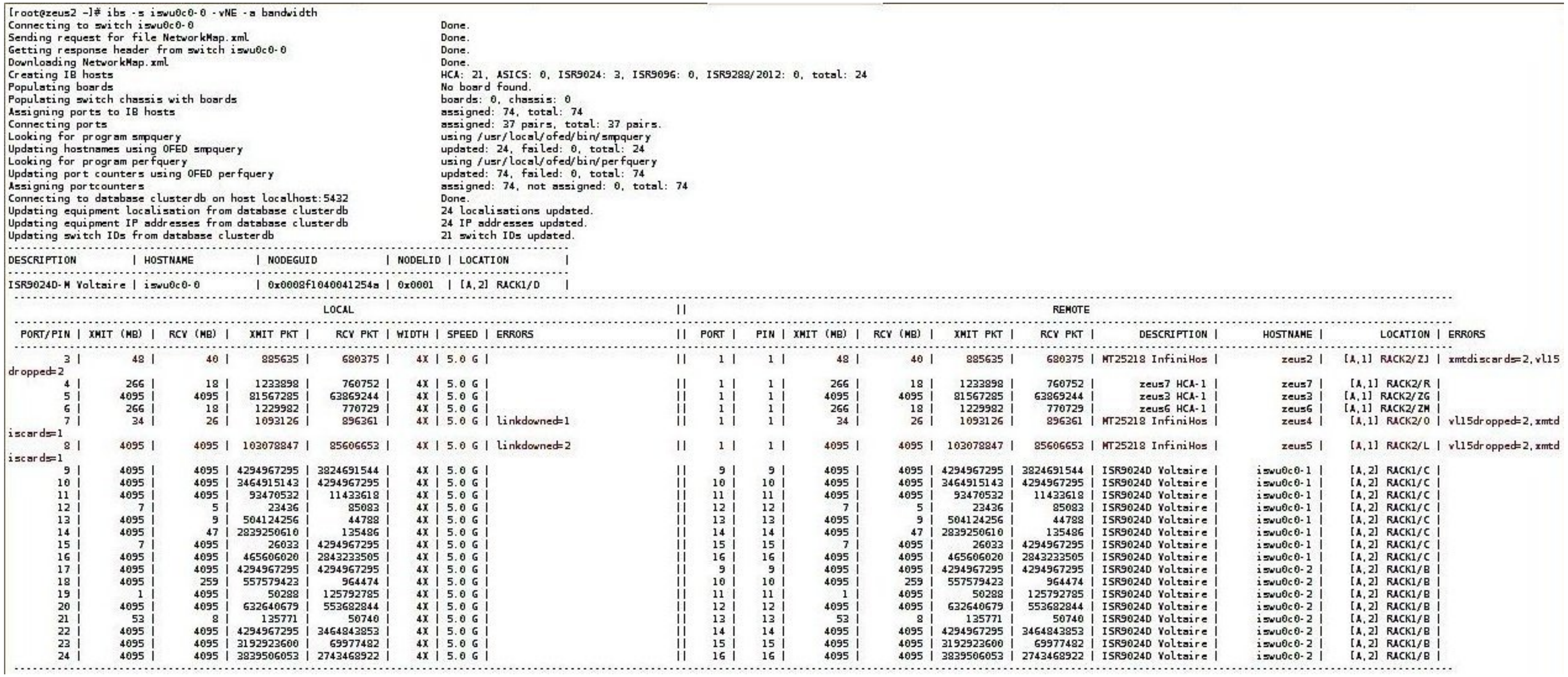
While the command is being carried out a message similar to that below will appear:

```
Successfully generated configuration file group.csv
To update a managed switch, proceed as follows:
 - Log onto the switch
 - Enter the 'enable' mode
 - Enter the 'config' menu
 - Enter the 'group' menu
 - Type the following command: group import /home/user/path
```



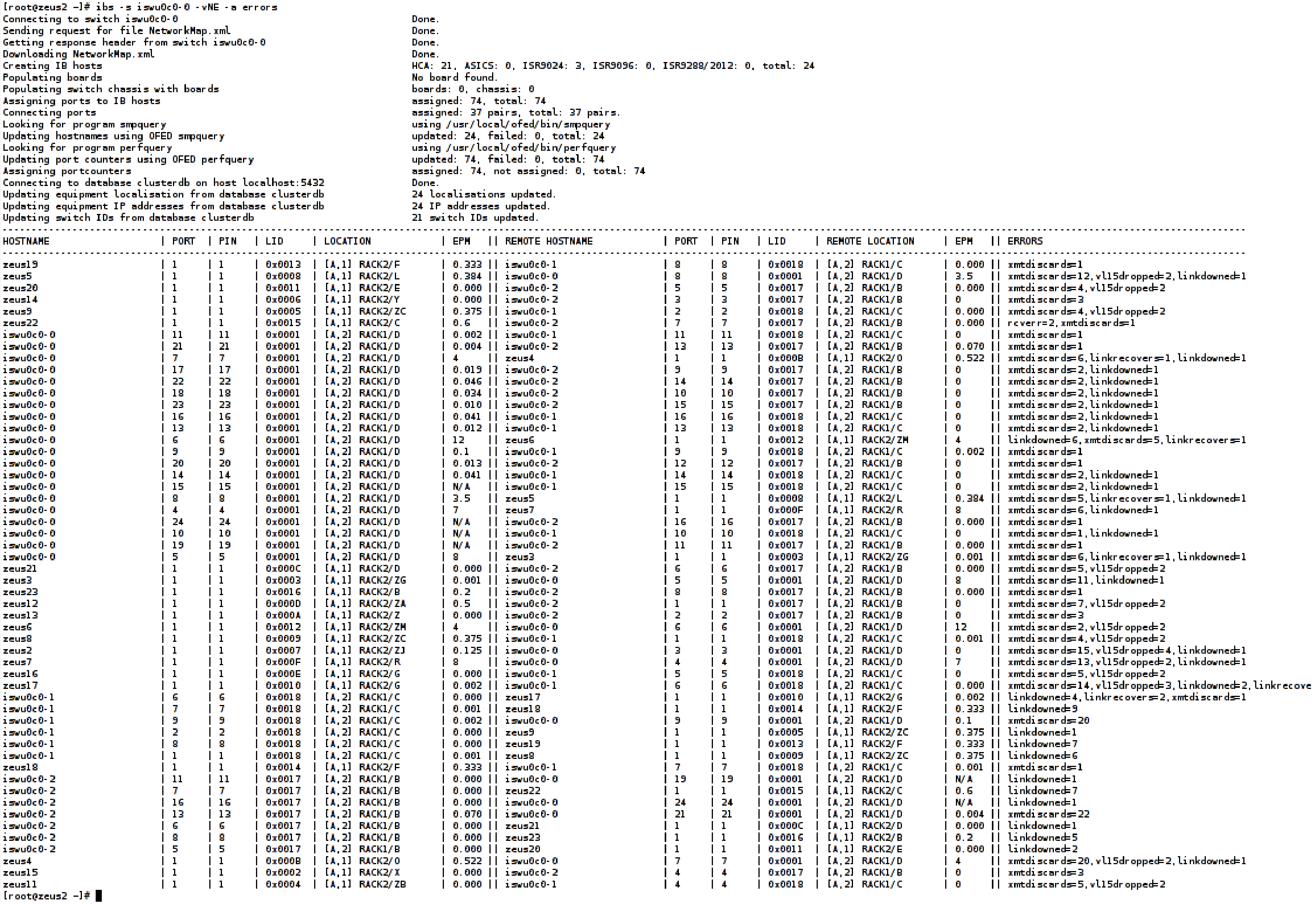
```
[root@zeus2 ~]# ibs -s iswu0c0-0 -vNE -a topo
Connecting to switch iswu0c0-0                          Done.
Sending request for file NetworkMap.xml                 Done.
Getting response header from switch iswu0c0-0           Done.
Downloading NetworkMap.xml                              Done.
Creating IB hosts                                       HCA: 21, ASICS: 0, ISR9024: 3, ISR9096: 0, ISR9288/2012: 0, total: 24
Populating boards                                       No board found.
Populating switch chassis with boards                   boards: 0, chassis: 0
Assigning ports to IB hosts                             assigned: 74, total: 74
Connecting ports                                        assigned: 37 pairs, total: 37 pairs.
Looking for program smpquery                            using /usr/local/ofed/bin/smpquery
Updating hostnames using OFED smpquery                  updated: 24, failed: 0, total: 24
Looking for program perfquery                           using /usr/local/ofed/bin/perfquery
Updating port counters using OFED perfquery             updated: 74, failed: 0, total: 74
Assigning portcounters                                  assigned: 74, not assigned: 0, total: 74
Connecting to database clusterdb on host localhost:5432 Done.
Updating equipment localisation from database clusterdb 24 localisations updated.
Updating equipment IP addresses from database clusterdb 24 IP addresses updated.
Updating switch IDs from database clusterdb             21 switch IDs updated.
-------------------------------------------------------------------------------------
DESCRIPTION         | HOSTNAME          | NODEGUID           | NODELID | LOCATION      |
-------------------------------------------------------------------------------------
ISR9024D-M Voltaire | iswu0c0-0         | 0x0008f1040041254a | 0x0001  | [A,2] RACK1/D |
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
                                LOCAL                                  ||                                                                  REMOTE
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
 PORT/PIN | PORTGUID/PORTNODEG | ERRORS                           | WIDTH | SPEED || PORT |   PIN |           PORTGUID |       PORTNODEGUID |   TYPE |        DESCRIPTION |       HOSTNAME | NODELID |        LOCATION | ERRORS
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
        3 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||    1 |     1 | 0x0002c9020024b915 | 0x0002c9020024b914 |     CA | MT25218 InfiniHos |          zeus2 |  0x0002 |  [A,1] RACK2/ZJ | xmtdiscards=2,vl15dropped=2
        4 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||    1 |     1 | 0x0002c9020024b991 | 0x0002c9020024b990 |     CA |       zeus7 HCA-1 |          zeus7 |  0x0006 |   [A,1] RACK2/R |
        5 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||    1 |     1 | 0x0002c9020024b83d | 0x0002c9020024b83c |     CA |       zeus3 HCA-1 |          zeus3 |  0x0011 |  [A,1] RACK2/ZG |
        6 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||    1 |     1 | 0x0002c9020024b9ad | 0x0002c9020024b9ac |     CA |       zeus6 HCA-1 |          zeus6 |  0x0012 |  [A,1] RACK2/ZM |
        7 | 0x0008f1040041254a | linkdowned=1                     |    4X | 5.0 G ||    1 |     1 | 0x0002c9020024b979 | 0x0002c9020024b978 |     CA | MT25218 InfiniHos |          zeus4 |  0x0013 |   [A,1] RACK2/O | vl15dropped=2,xmtdiscards=1
        8 | 0x0008f1040041254a | linkdowned=2                     |    4X | 5.0 G ||    1 |     1 | 0x0002c9020024b921 | 0x0002c9020024b920 |     CA | MT25218 InfiniHos |          zeus5 |  0x0010 |   [A,1] RACK2/L | vl15dropped=2,xmtdiscards=1
        9 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||    9 |     9 | 0x0008f1040041281e | 0x0008f1040041281e | Switch | ISR9024D Voltaire |      iswu0c0-1 |  0x0009 |   [A,2] RACK1/C |
       10 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   10 |    10 | 0x0008f1040041281e | 0x0008f1040041281e | Switch | ISR9024D Voltaire |      iswu0c0-1 |  0x0009 |   [A,2] RACK1/C |
       11 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   11 |    11 | 0x0008f1040041281e | 0x0008f1040041281e | Switch | ISR9024D Voltaire |      iswu0c0-1 |  0x0009 |   [A,2] RACK1/C |
       12 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   12 |    12 | 0x0008f1040041281e | 0x0008f1040041281e | Switch | ISR9024D Voltaire |      iswu0c0-1 |  0x0009 |   [A,2] RACK1/C |
       13 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   13 |    13 | 0x0008f1040041281e | 0x0008f1040041281e | Switch | ISR9024D Voltaire |      iswu0c0-1 |  0x0009 |   [A,2] RACK1/C |
       14 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   14 |    14 | 0x0008f1040041281e | 0x0008f1040041281e | Switch | ISR9024D Voltaire |      iswu0c0-1 |  0x0009 |   [A,2] RACK1/C |
       15 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   15 |    15 | 0x0008f1040041281e | 0x0008f1040041281e | Switch | ISR9024D Voltaire |      iswu0c0-1 |  0x0009 |   [A,2] RACK1/C |
       16 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   16 |    16 | 0x0008f1040041281e | 0x0008f1040041281e | Switch | ISR9024D Voltaire |      iswu0c0-1 |  0x0009 |   [A,2] RACK1/C |
       17 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||    9 |     9 | 0x0008f10400411946 | 0x0008f10400411946 | Switch | ISR9024D Voltaire |      iswu0c0-2 |  0x0008 |   [A,2] RACK1/B |
       18 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   10 |    10 | 0x0008f10400411946 | 0x0008f10400411946 | Switch | ISR9024D Voltaire |      iswu0c0-2 |  0x0008 |   [A,2] RACK1/B |
       19 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   11 |    11 | 0x0008f10400411946 | 0x0008f10400411946 | Switch | ISR9024D Voltaire |      iswu0c0-2 |  0x0008 |   [A,2] RACK1/B |
       20 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   12 |    12 | 0x0008f10400411946 | 0x0008f10400411946 | Switch | ISR9024D Voltaire |      iswu0c0-2 |  0x0008 |   [A,2] RACK1/B |
       21 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   13 |    13 | 0x0008f10400411946 | 0x0008f10400411946 | Switch | ISR9024D Voltaire |      iswu0c0-2 |  0x0008 |   [A,2] RACK1/B |
       22 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   14 |    14 | 0x0008f10400411946 | 0x0008f10400411946 | Switch | ISR9024D Voltaire |      iswu0c0-2 |  0x0008 |   [A,2] RACK1/B |
       23 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   15 |    15 | 0x0008f10400411946 | 0x0008f10400411946 | Switch | ISR9024D Voltaire |      iswu0c0-2 |  0x0008 |   [A,2] RACK1/B |
       24 | 0x0008f1040041254a |                                  |    4X | 5.0 G ||   16 |    16 | 0x0008f10400411946 | 0x0008f10400411946 | Switch | ISR9024D Voltaire |      iswu0c0-2 |  0x0008 |   [A,2] RACK1/B |
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

-------------------------------------------------------------------------------------
DESCRIPTION         | HOSTNAME          | NODEGUID           | NODELID | LOCATION      |
-------------------------------------------------------------------------------------
ISR9024D Voltaire   | iswu0c0-1         | 0x0008f1040041281e | 0x0009  | [A,2] RACK1/C |
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
                                LOCAL                                  ||                                                                  REMOTE
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
 PORT/PIN | PORTGUID/PORTNODEG | ERRORS                           | WIDTH | SPEED || PORT |   PIN |           PORTGUID |       PORTNODEGUID |   TYPE |        DESCRIPTION |       HOSTNAME | NODELID |        LOCATION | ERRORS
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
        1 | 0x0008f1040041281e | linkdowned=6                     |    4X | 5.0 G ||    1 |     1 | 0x0002c9020024b941 | 0x0002c9020024b940 |     CA | MT25218 InfiniHos |          zeus8 |  0x0004 |  [A,1] RACK2/ZC | vl15dropped=2,xmtdiscards=1
        2 | 0x0008f1040041281e | linkdowned=1                     |    4X | 5.0 G ||    1 |     1 | 0x0002c9020024b8a5 | 0x0002c9020024b8a4 |     CA | MT25218 InfiniHos |          zeus9 |  0x0003 |  [A,1] RACK2/ZC | vl15dropped=2,xmtdiscards=1
        4 | 0x0008f1040041281e |                                  |    4X | 5.0 G ||    1 |     1 | 0x0002c9020024b871 | 0x0002c9020024b870 |     CA | MT25218 InfiniHos |         zeus11 |  0x0017 |  [A,1] RACK2/ZB | xmtdiscards=2,vl15dropped=2
        5 | 0x0008f1040041281e |                                  |    4X | 5.0 G ||    1 |     1 | 0x0002c9020024b985 | 0x0002c9020024b984 |     CA | MT25218 InfiniHos |         zeus16 |  0x001D |   [A,1] RACK2/G | xmtdiscards=2,vl15dropped=2
        6 | 0x0008f1040041281e | linkdowned=4,linkrecovers=2,     |    4X | 2.5 G ||    1 |     1 | 0x0002c9020024b995 | 0x0002c9020024b994 |     CA |      zeus17 HCA-1 |         zeus17 |  0x000E |   [A,1] RACK2/G | xmtdiscards=11,vl15dropped=3,l
        7 | 0x0008f1040041281e | linkdowned=3                     |    4X | 5.0 G ||    1 |     1 | 0x0002c9020025fe11 | 0x0002c9020025fe10 |     CA |      zeus18 HCA-1 |         zeus18 |  0x001E |   [A,1] RACK2/F | xmtdiscards=1
        8 | 0x0008f1040041281e | linkdowned=3                     |    4X | 5.0 G ||    1 |     1 | 0x0002c9020025fe09 | 0x0002c9020025fe08 |     CA |      zeus19 HCA-1 |         zeus19 |  0x001C |   [A,1] RACK2/F | xmtdiscards=1
        9 | 0x0008f1040041281e |                                  |    4X | 5.0 G ||    9 |     9 | 0x0008f1040041254a | 0x0008f1040041254a | Switch | ISR9024D-M Voltai |      iswu0c0-0 |  0x0001 |   [A,2] RACK1/D |
       10 | 0x0008f1040041281e |                                  |    4X | 5.0 G ||   10 |    10 | 0x0008f1040041254a | 0x0008f1040041254a | Switch | ISR9024D-M Voltai |      iswu0c0-0 |  0x0001 |   [A,2] RACK1/D |
       11 | 0x0008f1040041281e |                                  |    4X | 5.0 G ||   11 |    11 | 0x0008f1040041254a | 0x0008f1040041254a | Switch | ISR9024D-M Voltai |      iswu0c0-0 |  0x0001 |   [A,2] RACK1/D |
       12 | 0x0008f1040041281e |                                  |    4X | 5.0 G ||   12 |    12 | 0x0008f1040041254a | 0x0008f1040041254a | Switch | ISR9024D-M Voltai |      iswu0c0-0 |  0x0001 |   [A,2] RACK1/D |
       13 | 0x0008f1040041281e |                                  |    4X | 5.0 G ||   13 |    13 | 0x0008f1040041254a | 0x0008f1040041254a | Switch | ISR9024D-M Voltai |      iswu0c0-0 |  0x0001 |   [A,2] RACK1/D |
       14 | 0x0008f1040041281e |                                  |    4X | 5.0 G ||   14 |    14 | 0x0008f1040041254a | 0x0008f1040041254a | Switch | ISR9024D-M Voltai |      iswu0c0-0 |  0x0001 |   [A,2] RACK1/D |
       15 | 0x0008f1040041281e |                                  |    4X | 5.0 G ||   15 |    15 | 0x0008f1040041254a | 0x0008f1040041254a | Switch | ISR9024D-M Voltai |      iswu0c0-0 |  0x0001 |   [A,2] RACK1/D |
       16 | 0x0008f1040041281e |                                  |    4X | 5.0 G ||   16 |    16 | 0x0008f1040041254a | 0x0008f1040041254a | Switch | ISR9024D-M Voltai |      iswu0c0-0 |  0x0001 |   [A,2] RACK1/D |
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
```

Figure 5-1. Example of IBS command topo action output



```
[root@zeus2 ~]# ibs -s iswu0c0-0 -vNE -a bandwidth
Connecting to switch iswu0c0-0                                  Done.
Sending request for file NetworkMap.xml                         Done.
Getting response header from switch iswu0c0-0                   Done.
Downloading NetworkMap.xml                                      Done.
Creating IB hosts                                               HCA: 21, ASICS: 0, ISR9024: 3, ISR9096: 0, ISR9288/2012: 0, total: 24
Populating boards                                               No board found.
Populating switch chassis with boards                           boards: 0, chassis: 0
Assigning ports to IB hosts                                     assigned: 74, total: 74
Connecting ports                                                assigned: 37 pairs, total: 37 pairs.
Looking for program smpquery                                    using /usr/local/ofed/bin/smpquery
Updating hostnames using OFED smpquery                          updated: 24, failed: 0, total: 24
Looking for program perfquery                                   using /usr/local/ofed/bin/perfquery
Updating port counters using OFED perfquery                     updated: 74, failed: 0, total: 74
Assigning portcounters                                          assigned: 74, not assigned: 0, total: 74
Connecting to database clusterdb on host localhost:5432         Done.
Updating equipment localisation from database clusterdb         24 localisations updated.
Updating equipment IP addresses from database clusterdb         24 IP addresses updated.
Updating switch IDs from database clusterdb                     21 switch IDs updated.
-------------------------------------------------------------------------------------------
DESCRIPTION          | HOSTNAME          | NODEGUID           | NODELID | LOCATION        |
-------------------------------------------------------------------------------------------
ISR9024D-M Voltaire | iswu0c0-0          | 0x0008f1040041254a | 0x0001  | [A,2] RACK1/D   |
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
                                       LOCAL                                               ||                                                     REMOTE
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
 PORT/PIN | XMIT (MB) |  RCV (MB) |  XMIT PKT |    RCV PKT | WIDTH | SPEED | ERRORS                  ||  PORT |   PIN | XMIT (MB) |  RCV (MB) |  XMIT PKT |    RCV PKT |        DESCRIPTION |      HOSTNAME |         LOCATION | ERRORS
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
        3 |        48 |        40 |    885635 |     680375 |    4X | 5.0 G |                         ||     1 |     1 |        48 |        40 |    885635 |     680375 | MT25218 InfiniHos |         zeus2 |   [A,1] RACK2/ZJ | xmtdiscards=2,vl15
dropped=2
        4 |       266 |        18 |   1233898 |     760752 |    4X | 5.0 G |                         ||     1 |     1 |       266 |        18 |   1233898 |     760752 |       zeus7 HCA-1 |         zeus7 |    [A,1] RACK2/R |
        5 |      4095 |      4095 |  81567285 |   63869244 |    4X | 5.0 G |                         ||     1 |     1 |      4095 |      4095 |  81567285 |   63869244 |       zeus3 HCA-1 |         zeus3 |   [A,1] RACK2/ZG |
        6 |       266 |        18 |   1229982 |     770729 |    4X | 5.0 G |                         ||     1 |     1 |       266 |        18 |   1229982 |     770729 |       zeus6 HCA-1 |         zeus6 |   [A,1] RACK2/ZM |
        7 |        34 |        26 |   1093126 |     896361 |    4X | 5.0 G | linkdowned=1            ||     1 |     1 |        34 |        26 |   1093126 |     896361 | MT25218 InfiniHos |         zeus4 |    [A,1] RACK2/O | vl15dropped=2,xmtd
iscards=1
        8 |      4095 |      4095 | 103078847 |   85606653 |    4X | 5.0 G | linkdowned=2            ||     1 |     1 |      4095 |      4095 | 103078847 |   85606653 | MT25218 InfiniHos |         zeus5 |    [A,1] RACK2/L | vl15dropped=2,xmtd
iscards=1
        9 |      4095 |      4095 | 4294967295 | 3824691544 |   4X | 5.0 G |                         ||     9 |     9 |      4095 |      4095 | 4294967295 | 3824691544 | ISR9024D Voltaire |    iswu0c0-1 |    [A,2] RACK1/C |
       10 |      4095 |      4095 | 3464915143 | 4294967295 |   4X | 5.0 G |                         ||    10 |    10 |      4095 |      4095 | 3464915143 | 4294967295 | ISR9024D Voltaire |    iswu0c0-1 |    [A,2] RACK1/C |
       11 |      4095 |      4095 |  93470532 |   11433618 |    4X | 5.0 G |                         ||    11 |    11 |      4095 |      4095 |  93470532 |   11433618 | ISR9024D Voltaire |    iswu0c0-1 |    [A,2] RACK1/C |
       12 |         7 |         5 |     23436 |      85083 |    4X | 5.0 G |                         ||    12 |    12 |         7 |         5 |     23436 |      85083 | ISR9024D Voltaire |    iswu0c0-1 |    [A,2] RACK1/C |
       13 |      4095 |         9 | 504124256 |      44788 |    4X | 5.0 G |                         ||    13 |    13 |      4095 |         9 | 504124256 |      44788 | ISR9024D Voltaire |    iswu0c0-1 |    [A,2] RACK1/C |
       14 |      4095 |        47 | 2839250610 |    135486 |    4X | 5.0 G |                         ||    14 |    14 |      4095 |        47 | 2839250610 |    135486 | ISR9024D Voltaire |    iswu0c0-1 |    [A,2] RACK1/C |
       15 |         7 |      4095 |     26033 | 4294967295 |    4X | 5.0 G |                         ||    15 |    15 |         7 |      4095 |     26033 | 4294967295 | ISR9024D Voltaire |    iswu0c0-1 |    [A,2] RACK1/C |
       16 |      4095 |      4095 | 465606020 | 2843233505 |    4X | 5.0 G |                         ||    16 |    16 |      4095 |      4095 | 465606020 | 2843233505 | ISR9024D Voltaire |    iswu0c0-1 |    [A,2] RACK1/C |
       17 |      4095 |      4095 | 4294967295 | 4294967295 |   4X | 5.0 G |                         ||     9 |     9 |      4095 |      4095 | 4294967295 | 4294967295 | ISR9024D Voltaire |    iswu0c0-2 |    [A,2] RACK1/B |
       18 |      4095 |       259 | 557579423 |     964474 |    4X | 5.0 G |                         ||    10 |    10 |      4095 |       259 | 557579423 |     964474 | ISR9024D Voltaire |    iswu0c0-2 |    [A,2] RACK1/B |
       19 |         1 |      4095 |     50288 |  125792785 |    4X | 5.0 G |                         ||    11 |    11 |         1 |      4095 |     50288 |  125792785 | ISR9024D Voltaire |    iswu0c0-2 |    [A,2] RACK1/B |
       20 |      4095 |      4095 | 632640679 |  553682844 |    4X | 5.0 G |                         ||    12 |    12 |      4095 |      4095 | 632640679 |  553682844 | ISR9024D Voltaire |    iswu0c0-2 |    [A,2] RACK1/B |
       21 |        53 |         8 |    135771 |      50740 |    4X | 5.0 G |                         ||    13 |    13 |        53 |         8 |    135771 |      50740 | ISR9024D Voltaire |    iswu0c0-2 |    [A,2] RACK1/B |
       22 |      4095 |      4095 | 4294967295 | 3464843853 |   4X | 5.0 G |                         ||    14 |    14 |      4095 |      4095 | 4294967295 | 3464843853 | ISR9024D Voltaire |    iswu0c0-2 |    [A,2] RACK1/B |
       23 |      4095 |      4095 | 3192923600 |   69977482 |   4X | 5.0 G |                         ||    15 |    15 |      4095 |      4095 | 3192923600 |   69977482 | ISR9024D Voltaire |    iswu0c0-2 |    [A,2] RACK1/B |
       24 |      4095 |      4095 | 3839506053 | 2743468922 |   4X | 5.0 G |                         ||    16 |    16 |      4095 |      4095 | 3839506053 | 2743468922 | ISR9024D Voltaire |    iswu0c0-2 |    [A,2] RACK1/B |
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
```

Figure 5-2. Example of IBS command bandwidth action output



```
[root@zeus2 ~]# ibs -s iswu0c0-0 -vNE -a errors
Connecting to switch iswu0c0-0                                        Done.
Sending request for file NetworkMap.xml                               Done.
Getting response header from switch iswu0c0-0                         Done.
Downloading NetworkMap.xml                                            Done.
Creating IB hosts                                                     HCA: 21, ASICS: 0, ISR9024: 3, ISR9096: 0, ISR9288/2012: 0, total: 24
Populating boards                                                     No board found.
Populating switch chassis with boards                                 boards: 0, chassis: 0
Assigning ports to IB hosts                                           assigned: 74, total: 74
Connecting ports                                                      assigned: 37 pairs, total: 37 pairs.
Looking for program smpquery                                          using /usr/local/ofed/bin/smpquery
Updating hostnames using OFED smpquery                                updated: 24, failed: 0, total: 24
Looking for program perfquery                                         using /usr/local/ofed/bin/perfquery
Updating port counters using OFED perfquery                           updated: 74, failed: 0, total: 74
Assigning portcounters                                                assigned: 74, not assigned: 0, total: 74
Connecting to database clusterdb on host localhost:5432               Done.
Updating equipment localisation from database clusterdb               24 localisations updated.
Updating equipment IP addresses from database clusterdb               24 IP addresses updated.
Updating switch IDs from database clusterdb                           21 switch IDs updated.
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
HOSTNAME                    | PORT | PIN  | LID    | LOCATION         | EPM   || REMOTE HOSTNAME          | PORT | PIN  | LID    | REMOTE LOCATION  | EPM   || ERRORS
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
zeus19                      | 1    | 1    | 0x0013 | [A,1] RACK2/F    | 0.333 || iswu0c0-1                | 8    | 8    | 0x0018 | [A,2] RACK1/C    | 0.000 || xmtdiscards=1
zeus5                       | 1    | 1    | 0x0008 | [A,1] RACK2/L    | 0.384 || iswu0c0-0                | 8    | 8    | 0x0001 | [A,2] RACK1/D    | 3.5   || xmtdiscards=12,vl15dropped=2,linkdowned=1
zeus20                      | 1    | 1    | 0x0011 | [A,1] RACK2/E    | 0.000 || iswu0c0-2                | 5    | 5    | 0x0017 | [A,2] RACK1/B    | 0.000 || xmtdiscards=4,vl15dropped=2
zeus14                      | 1    | 1    | 0x0006 | [A,1] RACK2/Y    | 0.000 || iswu0c0-2                | 3    | 3    | 0x0017 | [A,2] RACK1/B    | 0     || xmtdiscards=3
zeus9                       | 1    | 1    | 0x0005 | [A,1] RACK2/ZC   | 0.375 || iswu0c0-1                | 2    | 2    | 0x0018 | [A,2] RACK1/C    | 0.000 || xmtdiscards=4,vl15dropped=2
zeus22                      | 1    | 1    | 0x0015 | [A,1] RACK2/C    | 0.6   || iswu0c0-2                | 7    | 7    | 0x0017 | [A,2] RACK1/B    | 0.000 || rcverr=2,xmtdiscards=1
iswu0c0-0                   | 11   | 11   | 0x0001 | [A,2] RACK1/D    | 0.002 || iswu0c0-1                | 11   | 11   | 0x0018 | [A,2] RACK1/C    | 0     || xmtdiscards=1
iswu0c0-0                   | 21   | 21   | 0x0001 | [A,2] RACK1/D    | 0.004 || iswu0c0-2                | 13   | 13   | 0x0017 | [A,2] RACK1/B    | 0.070 || xmtdiscards=1
iswu0c0-0                   | 7    | 7    | 0x0001 | [A,2] RACK1/D    | 4     || zeus4                    | 1    | 1    | 0x000B | [A,1] RACK2/O    | 0.522 || xmtdiscards=6,linkrecovers=1,linkdowned=1
iswu0c0-0                   | 17   | 17   | 0x0001 | [A,2] RACK1/D    | 0.019 || iswu0c0-2                | 9    | 9    | 0x0017 | [A,2] RACK1/B    | 0     || xmtdiscards=2,linkdowned=1
iswu0c0-0                   | 22   | 22   | 0x0001 | [A,2] RACK1/D    | 0.046 || iswu0c0-2                | 14   | 14   | 0x0017 | [A,2] RACK1/B    | 0     || xmtdiscards=2,linkdowned=1
iswu0c0-0                   | 18   | 18   | 0x0001 | [A,2] RACK1/D    | 0.034 || iswu0c0-2                | 10   | 10   | 0x0017 | [A,2] RACK1/B    | 0     || xmtdiscards=2,linkdowned=1
iswu0c0-0                   | 23   | 23   | 0x0001 | [A,2] RACK1/D    | 0.010 || iswu0c0-2                | 15   | 15   | 0x0017 | [A,2] RACK1/B    | 0     || xmtdiscards=2,linkdowned=1
iswu0c0-0                   | 16   | 16   | 0x0001 | [A,2] RACK1/D    | 0.041 || iswu0c0-1                | 16   | 16   | 0x0018 | [A,2] RACK1/C    | 0     || xmtdiscards=2,linkdowned=1
iswu0c0-0                   | 13   | 13   | 0x0001 | [A,2] RACK1/D    | 0.012 || iswu0c0-1                | 13   | 13   | 0x0018 | [A,2] RACK1/C    | 0     || xmtdiscards=2,linkdowned=1
iswu0c0-0                   | 6    | 6    | 0x0001 | [A,2] RACK1/D    | 12    || zeus6                    | 1    | 1    | 0x0012 | [A,1] RACK2/ZM   | 4     || linkdowned=6,xmtdiscards=5,linkrecovers=1
iswu0c0-0                   | 9    | 9    | 0x0001 | [A,2] RACK1/D    | 0.1   || iswu0c0-1                | 9    | 9    | 0x0018 | [A,2] RACK1/C    | 0.002 || xmtdiscards=1
iswu0c0-0                   | 20   | 20   | 0x0001 | [A,2] RACK1/D    | 0.013 || iswu0c0-2                | 12   | 12   | 0x0017 | [A,2] RACK1/B    | 0     || xmtdiscards=1
iswu0c0-0                   | 14   | 14   | 0x0001 | [A,2] RACK1/D    | 0.041 || iswu0c0-1                | 14   | 14   | 0x0018 | [A,2] RACK1/C    | 0     || xmtdiscards=2,linkdowned=1
iswu0c0-0                   | 15   | 15   | 0x0001 | [A,2] RACK1/D    | N/A   || iswu0c0-1                | 15   | 15   | 0x0018 | [A,2] RACK1/C    | 0     || xmtdiscards=2,linkdowned=1
iswu0c0-0                   | 8    | 8    | 0x0001 | [A,2] RACK1/D    | 3.5   || zeus5                    | 1    | 1    | 0x0008 | [A,1] RACK2/L    | 0.384 || xmtdiscards=5,linkrecovers=1,linkdowned=1
iswu0c0-0                   | 4    | 4    | 0x0001 | [A,2] RACK1/D    | 7     || zeus7                    | 1    | 1    | 0x000F | [A,1] RACK2/R    | 8     || xmtdiscards=6,linkdowned=1
iswu0c0-0                   | 24   | 24   | 0x0001 | [A,2] RACK1/D    | N/A   || iswu0c0-2                | 16   | 16   | 0x0017 | [A,2] RACK1/B    | 0.000 || xmtdiscards=1
iswu0c0-0                   | 10   | 10   | 0x0001 | [A,2] RACK1/D    | N/A   || iswu0c0-1                | 10   | 10   | 0x0018 | [A,2] RACK1/C    | 0     || xmtdiscards=1,linkdowned=1
iswu0c0-0                   | 19   | 19   | 0x0001 | [A,2] RACK1/D    | N/A   || iswu0c0-2                | 11   | 11   | 0x0017 | [A,2] RACK1/B    | 0.000 || xmtdiscards=1
iswu0c0-0                   | 5    | 5    | 0x0001 | [A,2] RACK1/D    | 8     || zeus3                    | 1    | 1    | 0x0003 | [A,1] RACK2/ZG   | 0.001 || xmtdiscards=6,linkrecovers=1,linkdowned=1
zeus21                      | 1    | 1    | 0x000C | [A,1] RACK2/D    | 0.000 || iswu0c0-2                | 6    | 6    | 0x0017 | [A,2] RACK1/B    | 0.000 || xmtdiscards=5,vl15dropped=2
zeus3                       | 1    | 1    | 0x0003 | [A,1] RACK2/ZG   | 0.001 || iswu0c0-0                | 5    | 5    | 0x0001 | [A,2] RACK1/D    | 8     || xmtdiscards=11,linkdowned=1
zeus23                      | 1    | 1    | 0x0016 | [A,1] RACK2/B    | 0.2   || iswu0c0-2                | 8    | 8    | 0x0017 | [A,2] RACK1/B    | 0.000 || xmtdiscards=1
zeus12                      | 1    | 1    | 0x000D | [A,1] RACK2/ZA   | 0.5   || iswu0c0-2                | 1    | 1    | 0x0017 | [A,2] RACK1/B    | 0     || xmtdiscards=7,vl15dropped=2
zeus13                      | 1    | 1    | 0x000A | [A,1] RACK2/Z    | 0.000 || iswu0c0-2                | 2    | 2    | 0x0017 | [A,2] RACK1/B    | 0     || xmtdiscards=3
zeus6                       | 1    | 1    | 0x0012 | [A,1] RACK2/ZM   | 4     || iswu0c0-0                | 6    | 6    | 0x0001 | [A,2] RACK1/D    | 12    || xmtdiscards=2,vl15dropped=2
zeus8                       | 1    | 1    | 0x0009 | [A,1] RACK2/ZC   | 0.375 || iswu0c0-1                | 1    | 1    | 0x0018 | [A,2] RACK1/C    | 0.001 || xmtdiscards=4,vl15dropped=2
zeus2                       | 1    | 1    | 0x0007 | [A,1] RACK2/ZJ   | 0.125 || iswu0c0-0                | 3    | 3    | 0x0001 | [A,2] RACK1/D    | 0     || xmtdiscards=15,vl15dropped=4,linkdowned=1
zeus7                       | 1    | 1    | 0x000F | [A,1] RACK2/R    | 8     || iswu0c0-0                | 4    | 4    | 0x0001 | [A,2] RACK1/D    | 7     || xmtdiscards=13,vl15dropped=2,linkdowned=1
zeus16                      | 1    | 1    | 0x000E | [A,1] RACK2/G    | 0.000 || iswu0c0-1                | 5    | 5    | 0x0018 | [A,2] RACK1/C    | 0     || xmtdiscards=5,vl15dropped=2
zeus17                      | 1    | 1    | 0x0010 | [A,1] RACK2/G    | 0.002 || iswu0c0-1                | 6    | 6    | 0x0018 | [A,2] RACK1/C    | 0.000 || xmtdiscards=14,vl15dropped=3,linkdowned=2,linkrecove
iswu0c0-1                   | 6    | 6    | 0x0018 | [A,2] RACK1/C    | 0.000 || zeus17                   | 1    | 1    | 0x0010 | [A,1] RACK2/G    | 0.002 || linkdowned=4,linkrecovers=2,xmtdiscards=1
iswu0c0-1                   | 7    | 7    | 0x0018 | [A,2] RACK1/C    | 0.001 || zeus18                   | 1    | 1    | 0x0014 | [A,1] RACK2/F    | 0.333 || linkdowned=9
iswu0c0-1                   | 9    | 9    | 0x0018 | [A,2] RACK1/C    | 0.002 || iswu0c0-0                | 9    | 9    | 0x0001 | [A,2] RACK1/D    | 0.1   || xmtdiscards=20
iswu0c0-1                   | 2    | 2    | 0x0018 | [A,2] RACK1/C    | 0.000 || zeus9                    | 1    | 1    | 0x0005 | [A,1] RACK2/ZC   | 0.375 || linkdowned=1
iswu0c0-1                   | 8    | 8    | 0x0018 | [A,2] RACK1/C    | 0.000 || zeus19                   | 1    | 1    | 0x0013 | [A,1] RACK2/F    | 0.333 || linkdowned=7
iswu0c0-1                   | 1    | 1    | 0x0018 | [A,2] RACK1/C    | 0.001 || zeus8                    | 1    | 1    | 0x0009 | [A,1] RACK2/ZC   | 0.375 || linkdowned=6
zeus18                      | 1    | 1    | 0x0014 | [A,1] RACK2/F    | 0.333 || iswu0c0-1                | 7    | 7    | 0x0018 | [A,2] RACK1/C    | 0.001 || xmtdiscards=1
iswu0c0-2                   | 11   | 11   | 0x0017 | [A,2] RACK1/B    | 0.000 || iswu0c0-0                | 19   | 19   | 0x0001 | [A,2] RACK1/D    | N/A   || linkdowned=1
iswu0c0-2                   | 7    | 7    | 0x0017 | [A,2] RACK1/B    | 0.000 || zeus22                   | 1    | 1    | 0x0015 | [A,1] RACK2/C    | 0.6   || linkdowned=7
iswu0c0-2                   | 16   | 16   | 0x0017 | [A,2] RACK1/B    | 0.000 || iswu0c0-0                | 24   | 24   | 0x0001 | [A,2] RACK1/D    | N/A   || linkdowned=1
iswu0c0-2                   | 13   | 13   | 0x0017 | [A,2] RACK1/B    | 0.070 || iswu0c0-0                | 21   | 21   | 0x0001 | [A,2] RACK1/D    | 0.004 || xmtdiscards=22
iswu0c0-2                   | 6    | 6    | 0x0017 | [A,2] RACK1/B    | 0.000 || zeus21                   | 1    | 1    | 0x000C | [A,1] RACK2/D    | 0.000 || linkdowned=1
iswu0c0-2                   | 8    | 8    | 0x0017 | [A,2] RACK1/B    | 0.000 || zeus23                   | 1    | 1    | 0x0016 | [A,1] RACK2/B    | 0.2   || linkdowned=5
iswu0c0-2                   | 5    | 5    | 0x0017 | [A,2] RACK1/B    | 0.000 || zeus20                   | 1    | 1    | 0x0011 | [A,1] RACK2/E    | 0.000 || linkdowned=2
zeus4                       | 1    | 1    | 0x000B | [A,1] RACK2/O    | 0.522 || iswu0c0-0                | 7    | 7    | 0x0001 | [A,2] RACK1/D    | 4     || xmtdiscards=20,vl15dropped=2,linkdowned=1
zeus15                      | 1    | 1    | 0x0002 | [A,1] RACK2/X    | 0.000 || iswu0c0-2                | 4    | 4    | 0x0017 | [A,2] RACK1/B    | 0     || xmtdiscards=3
zeus11                      | 1    | 1    | 0x0004 | [A,1] RACK2/ZB   | 0.000 || iswu0c0-1                | 4    | 4    | 0x0018 | [A,2] RACK1/C    | 0     || xmtdiscards=5,vl15dropped=2
[root@zeus2 ~]# █
```

Figure 5-3. Example of IBS command errors action output

5.2.2 IBSDB Database

It is possible to create a database, which includes all the hardware and InfiniBand traffic details for all the switches, with the **IBS** tool. This database is specific to **InfiniBand** hardware.
The following commands apply to the **IBSDB** Database.

dbcreate

To create an empty, new IBS database (ibsdb) use the **dbcreate** command. Only the '**postgres**' user is allowed to create an empty database.

```
postgres@admin$ ibs -a dbcreate
```

While the command is being carried out a message similar to that below will appear:

```
Looking for program createdb                        using /usr/bin/createdb
Looking for program psql                            using /usr/bin/psql
Creating database ibsdb                             Done.
Loading table definitions into database ibsdb       Done.
```

dbdelete

To delete an IBS database (ibsdb) use the **dbdelete** command. Only the '**postgres**' user is allowed to delete an empty database.

```
postgres@admin$ ibs -a dbdelete
```

While the command is being carried out a message similar to that below will appear:

```
Looking for program dropdb                              using /usr/bin/dropdb
Deleting database ibsdb                                Done.
```

dbpopulate

Use the **dbpopulate** action to populate a new database. In the example below data is supplied from the **iswu0c0-0** managed switch from the Management Node, and the hostnames and traffic counters are populated using the OFED tools:

```
ibs -s iswu0c0-0 -a dbpopulate –vNE
```

While the command is being carried out a message similar to that below will appear:

```
-----------------------------------------------------------------------------------------
Connecting to switch iswu0c0-0                          Done.
Sending request for file NetworkMap.xml                 Done.
Getting response header from switch iswu0c0-0           Done.
Downloading NetworkMap.xml
Creating IB hosts   HCA: 21, ASICS: 0, ISR9024: 3, ISR9096: 0, ISR9288/2012: 0, total: 24
Populating boards                                       No board found.
Populating switch chassis with boards                   boards: 0, chassis: 0
Assigning ports to IB hosts                             assigned: 74, total: 74
Connecting ports                                        assigned: 37 pairs, total: 37 pairs.
Looking for program smpquery                            using /usr/local/ofed/bin/smpquery
```

```
Updating hostnames using OFED smpquery              updated: 24, failed: 0, total: 24
Looking for program perfquery                       using /usr/local/ofed/bin/perfquery
Updating port counters using OFED perfquery      updated: 74, failed: 0, total: 74
Assigning portcounters                           assigned: 74, not assigned: 0, total: 74
Connecting to database clusterdb on host localhost:5432          Done.
Updating equipment localisation from database clusterdb         24 localisations updated.
Updating equipment IP addresses from database clusterdb          24 IP addresses updated.
Updating switch IDs from database clusterdb                      21 switch IDs updated.
Connecting to database ibsdb on host localhost:5432              Done.
Populating table 'chassis' in database ibsdb                     0 chassis stored.
Populating tables 'asic' and 'chassis' in database ibsdb         3 ISR9024 switch stored.
Populating table 'board' in database ibsdb                       0 boards stored.
Populating table 'asic' in database ibsdb                        0 ASICs stored.
Populating table 'hca' in database ibsdb                         21 HCAs stored.
Populating tables 'asic_port' and 'hca_port' in database ibsdb   74 ports stored.
Populating tables 'asic_portcounters' and 'hca_portcounters      74 portcounters stored.
-----------------------------------------------------------------------------------------
```

dbupdate

Use the **dbupdate** action to update an existing IBSDB database.
In the example below the topology and traffic counter details for the **iswu0c0-0** managed switch from the Management Node, is updated using the OFED tools:

```
ibs -s iswu0c0-0 -a dbupdate -NE
```

In order to ensure that the data is always up to date, add the following line to the **cron** table (using **crontab -e**).

```
*/10 * * * * PATH=/usr/local/ofed/bin:$PATH /usr/bin/ibs -s iswu0c0-0 -a
dbupdate -vNE >> /var/log/ibs.log 2>&1
```

The traffic and error counters as well as the **InfiniBand** equipment stored in the **IBS** database will be refreshed every 10 minutes using the data supplied by the **iswu0c0-0** switch.

Note The user needs to know which switch is running the subnet manager as master for **InfiniBand** clusters that include multiple managed switches. This switch should always be the one that is specified as the argument of the **-s** flag. Assuming that the data is refreshed by the **cron** daemon, then if another switch becomes the subnet manager master the data details contained in the database would then be incorrect, as it would use data from what is the slave switch as defined in the cron script.
Use the **sminfo** command as follows to know which subnet manager is running as the master. Output in a form similar to that below will be provided:

```
sminfo: sm lid 1 sm guid 0x8f1040041254a, activity count 544113
priority 3 state 3 SMINFO_MASTER
```

The **guid** that is identified can then be used to find the corresponding switch name in the ibsdb '**chassis**' table.

dbupdatepc

Use the **dbupdatepc** action to update the port counters for an existing IBSDB database. Use the command below:

```
ibs -a dbupdatepc -vNE
```

availability

Use the **availability** action to see which ports and links are available for the **InfiniBand** interconnects. This action will not work unless the IBSDB database has been created and populated.

```
ibs -s iswu0c0-0 -a availability
```

This will give results in a similar format to that below.

```
Active ports: 74
Active uplinks: 16
Active downlinks: 21
```

5.2.3 Return Values

IBS returns 0 for success. Any other value indicates a failure.

5.3 Monitoring Voltaire Switches

Different options exist for monitoring and maintaining the performance of **Voltaire** switches.

5.3.1 Launching Utilities

To begin with enter the utilities menu as follows:

```
[user@host ~]# ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
Connecting
```

```
switchname # utilities
switchname (utilities)#
```

5.3.2 Resetting the counters

The counters (volume and errors) can be reset through the **zero-counters** command as follows:

```
switchname (utilities) zero-counters
```

```
Zero All Counters
Zero lid 8 port 255 mask 0xffff
[ ... ]
```

5.3.3 Finding bad ports

The **find_bad_ports** command can be used to detect faulty ports:

```
switchname (utilities) find_bad_ports
```

```
Found bad link/port:
node_guid:......................0008f10400411946
node_desc:......................'ISR9024D Voltaire'
lid:............................152
smlid:..........................8
Port 4
direct path from self switch: 0,1 4
```

5.3.4 Verifying the ports

The whole **Infiniband** fabric can be checked using the **port-verify** command as follows:

```
switchname (utilities) port-verify
```

```
#
# Topology file: generated on Thu Oct  4 20:19:24 2007
#
devid=0x5a31
switchguids=0x8f1040041254a
Switch  24 "S-0008f1040041254a"   # "ISR9024D-M Voltaire" smalid 8
[1] "S-0008f10400411946"[13] width 4X speed 5.0 Gbs
[2] "S-0008f10400411946"[14] width 4X speed 5.0 Gbs
[3] "S-0008f10400411946"[15] width 4X speed 5.0 Gbs
[ ... ]
devid=0x6282
hcaguids=0x2c9020024b940
Hca  2 "H-0002c9020024b940"    # "zeus8 HCA-1"
[1] "S-0008f1040041281e"[1]  # lid 72 lmc 3 width 4X speed 5.0 Gbs
SUMMARY: NO PROBLEMS DETECTED.
```

5.3.5 Checking the port width

To ensure the best performance, check that the ports are running in 4x mode as follows:

```
switchname (utilities) width-check
```

```
Verify / every error found - will be printed
lid 8 guid 0008f1040041254a ports 24
lid 160 guid 0008f1040041281e ports 24
lid 152 guid 0008f10400411946 ports 24
```

5.3.6 Dealing with a faulty port

When a faulty port is diagnosed, it can be disabled or reset using the **port-manage** command, as below:

```
iswu0c0-0(utilities) port-manage
```

Description:

port-manage.sh is used to trigger a physical state change for the port specified. This is useful when the active width/speed of a specific port must be changed without the cable being reconnected.

Syntax:

port-manage.sh [-v] [-f] <-d|-e|-r> <LID> <PORT>

Options:

-v Increase output verbosity level

-f Force disabling or resetting a port even when the port is located on the Access Path (path/way to the specific port)

-d lid port Disable the port

-e lid port Enable the port (set port state machine to polling state)

-r lid port	Reset the port
-S lid port	Reset the port and set Enabled Speed to SDR
-D lid port	Reset the port and set Enabled Speed to SDR/DDR
-h	Show this help

Example:

```
#port-manage.sh -r 17 21 (reset LID=17 PORT=21)
```

5.4 Getting Information about Storage Devices (lsiocfg)

lsiocfg is a tool used for reporting information about storage devices. It is mainly dedicated to external storage systems (DDN and FDA disk arrays) and their dedicated Host Board Adapters (Emulex FC adapters), but it can also be used with internal system storage (system disks) and their Host Board Adapters tools.

Reported information is related to several inventories:

- Host Board Adapters (-c flag)
- Disks (-d flag)
- Disk partitions (-p flag)
- Disk usages.

5.4.1 lsiocfg Command Syntax

According to needed information, **lsiocfg** can be used with options related to each inventory.

- **lsiocfg [-P] [-v] -c [HBAs IDs]**
  Gives information about all SCSI controllers. If HBAs IDs are specified, only applies to this list of HBAs.
- **lsiocfg [-P] [-v] -d [-u] [devices names]**
  Gives information about SCSI devices. [-u] has to be used to display non disk devices. If devices are specified, only applies to this list of devices.
- **lsiocfg -p**
  Displays partitions.
- **lsiocfg [-P] [-v] -a**
  Dsplays all (= -cdp).
- **lsiocfg [-r user] -n remote node [-P] [-v] [-c|-d|-a]**
  Gives information from remote node about controllers/disks.
- **lsiocfg -M [devices names]**
  Gives information about SCSI devices usage.
- **lsiocfg <-l|-L> <wwpn>**
  Reports WWPN owner. The –l flag uses **/etc/wwn** file, and the –L flag uses cluster manager database.
- **lsiocfg <-w|-W>**
  Displays all WWPN owners. The –w flag uses **/etc/wwn** file, and the –W flag uses cluster manager database.

General flags:

-P No headers (before -[a|c|d] commands).

-v Verbose (before -[a|c|d] commands). WWPN verbose information is extracted from **/etc/wwn** file.

-h Help message. Exclusive with other options.

-V Display the version. Exclusive with other options.

Online help and a man page give information about **lsiocfg** usage.

5.4.2 HBA Inventory

Using the **lsiocfg** HBA inventory option, you can get basic information about Host Board Adapters:

- model,
- link up or down.

When getting HBA inventory in verbose mode, more details are available:

- firmware levels,
- serial number,
- WWNN and WWPN (for fibre channel HBAs).

Example:

```
# lsiocfg -cv
```

```
------ HOST/CHANNEL INVENTORY ----------------------------------------
Host   Driver     Unique_id Cmd/Lun HostQ  State               Model
----------------------------------------------------------------------

host0  mptbase    0         7       -                          -
host1  mptbase    1         7       -                          -
host2  lpfc       0         30      -     LINK_UP             LP11000
        DRV=8.0.30_p1
        FW=2.10A7 (B2D2.10A7)
        Bus-Number=26
        SN=VM53824841
        Host-WWNN=20:00:00:00:c9:4b:e7:02
        Host-WWPN=10:00:00:00:c9:4b:e7:02
        FN=20:00:00:00:c9:4b:e7:02
        speed=2 Gbit
host3  usb-storage 0         1       -                          -
```

5.4.3 Disks Inventory

Using the **lsiocfg** Disk inventory option, you can get basic information about the available disks:

- system location
- vendor
- state

- disk size.

When getting the disk inventory in verbose mode, more details are shown:

- model
- serial number
- firmware revision
- WWPN (fiber channel devices).

```
# lsiocfg -dv
```

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

----- DISK INVENTORY ----------------------------------------------------
Dev   Location  Maj:Min  Vendor        state    Size (MB) QueueDepth  Lname
  (location= Host:Channel:Id:LUN)
-------------------------------------------------------------------------

sdb   0:0:10:0  8:16     SEAGATE       running  286102    31
        MODEL=SEAGATE ST3300007LC
        FWREV=0003
        SERIAL=3KR0KTPH00007547TR0P
        TRANSPORT=SPI
sdc   0:0:11:0  8:32     SEAGATE       running  286102    31
        MODEL=SEAGATE ST3300007LC
        FWREV=0003
        SERIAL=3KR0KTHM000075475NWC
        TRANSPORT=SPI
sda   0:0:9:0   8:0      SEAGATE       running  286102    31
        MODEL=SEAGATE ST3300007LC
        FWREV=0003
        SERIAL=3KR0JT0T00007548GUXA
        TRANSPORT=SPI
sdd   2:0:0:0   8:48     DDN           running  10000     30  /dev/ldn.ddn0.13
        MODEL=DDN S2A 8500
        FWREV=5.20
        SERIAL=02A820510D00
        TRANSPORT=FC
        WWPN=24:00:00:01:ff:03:02:a8
        NAME=unknown
sde   2:0:0:1   8:64     DDN           running  125000    30  /dev/ldn.ddn0.14
        MODEL=DDN S2A 8500
        FWREV=5.20
        SERIAL=02A820540E00
        TRANSPORT=FC
        WWPN=24:00:00:01:ff:03:02:a8
        NAME=unknown
sdf   2:0:0:2   8:80     DDN           running  10000    30  /dev/ldn.ddn0.15
        MODEL=DDN S2A 8500
        FWREV=5.20
        SERIAL=03E020570F00
        TRANSPORT=FC
        WWPN=24:00:00:01:ff:03:02:a8
        NAME=unknown
sdg   2:0:0:3   8:96     DDN           running  125000    30  /dev/ldn.ddn0.16
        MODEL=DDN S2A 8500
        FWREV=5.20
        SERIAL=03E0205A1000
        TRANSPORT=FC
        WWPN=24:00:00:01:ff:03:02:a8
        NAME=unknown
---------------------------------------------------------------------------------
```

5.4.4 Disk Usage and Partition Inventories

These inventories give information about system and logical use of the devices. Such information is mostly used for system administration needs.

5.5 Checking Device Power State (pingcheck)

The **pingcheck** command checks the power state (on or off) of the specified devices.

Usage:

pingcheck [options] --Type <device type> command devices

Options:

--dbname name Specify database name.

--debug, -d Debug mode (more than verbose).

--help, -h Display **pingcheck** help.

--interval, -i Specify the number of bsm calls before waiting the period defined by the **--time** option.

--jobs, -j Number of simultaneous bsm actions (for example, with -j 5 you can run 5 simultaneous **bsmpower** processes). Default: 30.

--only_test, -o Display the NS Commands that would be launched according to the specified options and action. This is a testing mode, no action is performed.

--time, -t Time to wait after the number of bsm calls defined by the **--interval** option.

--verbose, -v Verbose mode.

Parameters

--Type <device type> Type of devices to be «pinged »: **disk_array** or **server**.

command **on** or **off**.

devices Specify the name of the devices, using the **basename[i,j-k]** or **lc-like** syntax.

Examples:

- The following command verifies that all the power supplies for disk_array `10` to `15` are in on state and indicates those which are not.

```
pingcheck --Type disk_array on da[10-15]
```

- The following command verifies that servers nova`5` to `7` are in `off` state and indicates those which are not.

```
pingcheck --Type server off nova[5-7]
```

5.6 Setting Up Outlet Air Temperature

Use the GUI Console of the Bull Cool Cabinet Door to modify the default outlet air temperature value of the Computer Centre.

See The *Cool Cabinet Door Console User's Guide* for details.

Chapter 6. Debugging Tools

This chapter describes the following debugging tasks:

- 6.1 *Modifying the Core Dump Size*
- 6.2 *Identifying InfiniBand Network Problems*
- 6.3 *Using dump tools with RHEL5* (crash, proc, kdump)
- 6.4 *Configuring systems to take dumps from* the Management Network
- 6.5 *Identifying problems in the different parts of a kernel*

6.1 Modifying the Core Dump Size

By default the maximum size for core dump files for Bull HPC systems is set to 0 which means that no resources are available and core dumps cannot be done. In order that core dumps can be done the values for the **ulimit** command have to be changed.

For more information refer to the options for the **ulimit** command in the **bash** man page.

6.2 Identifying InfiniBand Network Problems (ibtracert)

ibtracert uses Subnet Manager Protocols (**SMP**) to trace the path from a source GID/LID to a destination GID/LID. Each hop along the path is displayed until the destination is reached or a hop does not respond. By using the **-mg** and/or **-ml** options, multicast path tracing can be performed between the source and destination nodes.

Syntax:

ibtracert [options] <src-addr> <dest-addr>

Flags

-n Simple format; no additional information is displayed.

-m <mlid> Show the multicast trace of the specified mlid.

Examples

- To show trace between lid 2 and 23, enter:

```
ibtracert 2 23
```

- To show multicast trace between lid 3 and 5 for mcast lid 0xc000, enter:

```
ibtracert -m 0xc000 3 5
```

Output:

The output for a command between two points is displayed in both hexadecimal format and in human-readable format – as shown in the example below for the trace between the two lids 0x22 and 0x2c. This is very useful in helping to identify any port/switch problems in the **InfiniBand** Fabric.

```
ibtracert 0x22 0x2c
```

```
>From ca {0008f10403979958} portnum 1 lid 0x22-0x22 "lynx13 HCA-1"
[1] -> switch port {0008f104004118e2}[8] lid 0x4-0x4 "ISR9024D Voltaire"
[13] -> switch port {0008f104004118e8}[16] lid 0x3-0x3 "ISR9024D-M Voltaire"
[21] -> switch port {0008f104004118e4}[13] lid 0x1-0x1 "ISR9024D Voltaire"
[4] -> ca port {0008f10403979985}[1] lid 0x2c-0x2c "lynx19 HCA-1"
To ca {0008f10403979984} portnum 1 lid 0x2c-0x2c "lynx19 HCA-1"


In short:
=> OUT  lynx13 (lid 0x22 / port 1
=> INTO node switch (lid 0x4) / port 8
=> OUT  node switch (lid 0x4) / port 13
=> INTO top switch  (lid 0x3) / port 16
=> OUT  top switch  (lid 0x3) / port 21
=> INTO node switch (lid 0x1) / port 13
=> OUT  node switch (lid 0x1) / port 4
=> INTO lynx 19 (lid 0x2c) / port 1
```

6.3 Using dump tools with RHEL5 (crash, proc, kdump)

Various tools allow problems to be analysed whilst the system is in operation:

- **crash** portrays system data symbolically using the possibilities provided by the **GDB** debugger. The commands which it offers are system oriented, for example, the list of tasks, tracing function calls for a task which is waiting, etc.
  See the **crash** man page for more information.
- The system file **/proc** may be used to view, and if necessary modify, system information. In particular it can be used to examine system information for different tasks, the state of the memory allocation, etc.
  See the **proc** man page for more information.
- In the event of a system crash, memory will be written to the configured disk location using **kdump**. Upon subsequent reboot, the data will be copied from the old memory and formatted into a **vmcore** file and stored in the **/var/crash/** subdirectory. The end result can then be analysed using the **crash** utility. An example command is shown below.

```
crash /usr/lib/debug/lib/modules/<kernel_version>/vmlinux vmcore
```

6.4 Configuring systems to take dumps from the Management Network

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

Add **nmi_watchdog=0** to the kernel boot options in the **/boot/grub/menu.lst** file in order to deactivate the NMI watchdog used by **RHEL**, so that the other NMIs can be put into effect.

An example of the menu.1st file is shown below:

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

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

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

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

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

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

or by using the **nsctrl** command.

See http://kbase.redhat.com/faq/FAQ_105_9036.shtm for more information

Notes

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

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

Important

It is essential to use non-stripped binary code within the kernel. Non-stripped binary code is included in the debuginfo RPM available from:
http://people.redhat.com/duffy/debuginfo/index-js.html
This package installs the kernel binary in the folder
/usr/lib/debug/lib/modules/<kernel_version>/

6.5 Identifying problems in the different parts of a kernel

Various configuration parameters enable traces or additional checks to be used on different kernel operations, for example, locks, memory allocation and so on.

It is usually possible to focus the debug mode on the problematic part of the kernel which has been identified after recompilation. It is also possible to insert code, e.g. **printk**, to help examine the problematic part.

The different compilation tasks for a machine – stopping, starting, resetting, creating a dump, bootstrapping a compiled system and debugging may be carried out from a remote work station, connected to a development machine configured as a DHCP server.

Chapter 7. Troubleshooting

Troubleshooting deals with the unexpected and is an important contribution towards maintaining a cluster in a stable and reliable condition. This chapter is aimed at helping you to develop a general, comprehensive methodology for identifying and solving problems on- and off-site.

The following topics are described:

7.1 Troubleshooting Voltaire Networks

7.1.1 Voltaire's Fabric Manager

Voltaire's Fabric Manager enables **InfiniBand** fabric connectivity debugging using the built-in **Performance Manager** (PM). **PM** has two major capabilities:

Port Counters Monitoring and Report
The **PM** generates a periodic port counters report file (in **CSV** format) that can be loaded to Excel and further analyzed by the user. It also monitors port counters errors and reports every port that passes its error threshold limit (as configured by the user).

Event Logging
This creates an event log file for both **IB** traps and **SubNet** internal events. The user may filter the events using a **GUI** and or a **CLI**. The filtering policy determines whether an event is logged and whether a trap is generated.

It is essential to identify any problem ports and node connectivity problems prior to running application as well as during standard operation.

Note See the *Voltaire Switch User Manual ISR 9024, ISR 9096, and ISR 9288/2012 Switches* for details on how to configure and use Port Counters and the Performance Manager. This manual also includes a description of all the **PortCounter** fields and counter values.

7.1.2 Fabric Diagnostics

Diagnostic is recommended in the following cases:

- During Fabric installation and during startup.
- Before running an application.
- Performance problems (by locating discarded packets and link integrity problems).
- MPI job run problem, to locate malfunctioning nodes and get the overall fabric structure.
- Additional problems related to fabric stability, blocking or other.

7.1.3 Debugging Tools

Tools available to perform diagnostic:

- Use the Topology Map to see current problems.
- The Error Log.
- The Bad Ports Log.
- The Current Alarms Table.
- The Fabric Statistics **portcounters.csv** file.

7.1.4 High-Level Diagnostic Tools

1. Enable the SM Fabric Inspect preferences for debugging Fabric Failure.

2. Use the VFM/VDM **Port Counters Information and Graph** window to check a specific port counter's health.

3. Use the **Event Log** to discover that there is a problem in the fabric. In the VFM, right click and select View Event to get information to help identify where problem is located. Alternatively, you can show the Event Log from the CLI.

4. Use the **Current Alarms** Table to see current problems. In the VFM, right click and select Alarm Data to get information to help identify where the problem is located.

5. Use the **Topology Map** to identify nodes with a current alarm.

6. Proactively look for increasing error counters using the statistics feature and running the Diagnostic scripts using the **CLI.**

Note See the *Voltaire Switch User Manual ISR 9024, ISR 9096, and ISR 9288/2012 Switches* for full details on using these tools.

7.1.5 CLI Diagnostic Tools

7.1.5.1 zero-counters script

To clear out all the errors across the fabric, use the **zero-counters** script to traverse the fabric and clear out all the port counters on both the switches and HCAs. This script is very easy to use and is helpful if you want to start off with a clean baseline of your fabric after many changes have occurred.

```
 ISR9288(utilities) zero-counters
Zero All Counters
lid 1 ports 24
************************
lid 5 ports 24
************************
lid 4 ports 24
************************
lid 3 ports 24
************************
lid 2 ports 24
************************
lid 11 ports 24
************************
....
```

Note See the *Voltaire Switch User Manual ISR 9024, ISR 9096, and ISR 9288/2012 Switches* for full details on the CLI commands.

7.1.5.2 width-check script

Another valuable script is the **width-check** script which allows you to easily check the fabric for 1X connections links. While the fabric will work over a 1X connection, it will however create a bottleneck and hurt performance within the fabric. All links should report no 1X connections when the script is ran. Nothing else will be reported other than the LID and GUID if it's a full 4X link.

```
ISR9288(utilities) width-check

Verify / every error found - will be printed

lid 1 guid 0008f104004004d7 ports 24

lid 5 guid 0008f104003f0723 ports 24
lid 4 guid 0008f104003f0722 ports 24
lid 3 guid 0008f104003f071f ports 24
lid 2 guid 0008f104003f071e ports 24
lid 11 guid 0008f104003f0747 ports 24
lid 10 guid 0008f104003f0746 ports 24
lid 7 guid 0008f104003f073b ports 24
    ...
```

7.1.5.3 error-find script

The easiest way to look for errors on all ports in the fabric is to run the error-find script. It will report any non-zero port counters found throughout the fabric on both switches and HCAs.

```
--------------------------------------------------------------------
    ISR9288(utilities) error-find

Show All Counter Errors / every error found - will be printedlid 1 guid
    0008f104004004d7 ports 24
lid 5 guid 0008f104003f0723 ports 24
port 22 xmitdiscards:...................4
port 10 linkdowned:.....................1
port 13 lid 4 guid 0008f104003f0722 ports 24
port 14 errs.sym:.......................83
--------------------------------------------------------------------
```

7.1.6 Event Notification Mechanism

Fabric related events can be generated by both the **PM** (Performance Monitor) and by the **SM** (Subnet Manager).

The **PM** periodically scans the error counters of all IB elements in the fabric and reports if a counter exceeds its threshold.

The **SM** monitors the fabric, detects configuration changes and dynamically configures the new elements and new routes in the fabric. The **SM** can detect fabric errors/warnings/informative events and report them.

Both, the **PM** and the **SM** generate events and report them to the event notification mechanism. In addition, events may be generated in the fabric and sent to the **SM** by fabric elements. The **SM** reports those events as well.

The event mechanism can do the following actions with each event:

a. Log the event in the event log.
b. Issue a trap to the GUI session.
c. If the event corresponds to an alarm, it is also sent to the current alarm mechanism.

The GUI Color coding is defined according to traps and events severity, as described below.

GUI Color-Coding	Event Severity	Description	Examples
Red	Critical / Major	Critical means that the system or a system component fails to operate.	Invalid link Duplicate or conflicting ports or path
Yellow	Warning / Minor	Warning/minor reflects a problem in the fabric but does not prevent its operation. A warning is asserted when an event is exceeding a predefined threshold.	Broken link Illegal connections between two sLB ports
Green	Normal	Information/Notification provided to the user of normal operating state or a normal system event.	Complete subnet reconfiguration Create/Delete Multicast group Applied routing scheme Port State Change

7.2 Troubleshooting InfiniBand Stacks

A suite of **InfiniBand** diagnostic tools are provided with the Bull Advanced Server.

ibstatus, **ibtracert** and **ibs** (a tool developed by Bull), are described in chapter 5 – *Monitoring Devices*. Some of the more useful troubleshooting tools are described below.

7.2.1 smpquery

Subnet Manager Query (**smpquery**) includes a subset of standard SMP query options which may be used to bring up information – in a human readable format - for different parts of the network including nodes, ports and switches.

Syntax:

```
smpquery [options] <op> <dest_addr> [op_params]
```

nodeinfo example:

An example of use of this command including the Local ID and the port number is below:

```
smpquery nodeinfo 45 1
```

The resulting information output will be similar to that displayed below:

```
--------------------------------------------------------------------------
 BaseVers:........................1
 ClassVers:.......................1
 NodeType:........................Channel Adapter
 NumPorts:........................2
 SystemGuid:......................0x0008f10403977ca7
 Guid:............................0x0008f10403977ca4
 PortGuid:........................0x0008f10403977ca6
 PartCap:.........................64
 DevId:...........................0x5a04
 Revision:........................0x000000a1
 LocalPort:.......................2
 VendorId:........................0x0008f1
--------------------------------------------------------------------------
```

portinfo example:

An example of use of this command including the Local ID and the port number is below:

```
smpquery portinfo 45 1
```

The resulting information output will be similar to that displayed below:

```
--------------------------------------------------------------------------
 Mkey:............................0x0000000000000000
 GidPrefix:.......................0xfe80000000000000
 Lid:.............................0x002d
 SMLid:...........................0x0003
 CapMask:.........................0x500a68
                                  IsTrapSupported
                                  IsAutomaticMigrationSupported
```

```
                                   IsSLMappingSupported
                                   IsLedInfoSupported
                                   IsSystemImageGUIDsupported
                                   IsVendorClassSupported
                                   IsCapabilityMaskNoticeSupported
  DiagCode:........................0x0000
  MkeyLeasePeriod:.................0
  LocalPort:.......................2
  LinkWidthEnabled:................1X or 4X
  LinkWidthSupported:..............1X or 4X
  LinkWidthActive:.................4X
  LinkSpeedSupported:..............2.5 Gbps
  LinkState:.......................Active
  PhysLinkState:...................LinkUp
  LinkDownDefState:................Polling
  ProtectBits:.....................0
  LMC:.............................0
  LinkSpeedActive:.................2.5 Gbps
  LinkSpeedEnabled:................2.5 Gbps
  NeighborMTU:.....................2048
  SMSL:............................0
  VLCap:...........................VL0-7
  InitType:........................0x00
  VLHighLimit:.....................0
  VLArbHighCap:....................8
  VLArbLowCap:.....................8
  InitReply:.......................0x00
  MtuCap:..........................2048
  VLStallCount:....................7
  HoqLife:.........................13
  OperVLs:.........................VL0-7
  PartEnforceInb:..................0
  PartEnforceOutb:.................0
  FilterRawInb:....................0
  FilterRawOutb:...................0
  MkeyViolations:..................0
  PkeyViolations:..................0
  QkeyViolations:..................0
  GuidCap:.........................32
  ClientReregister:................0
  SubnetTimeout:...................18
  RespTimeVal:.....................1
  LocalPhysErr:....................15
  OverrunErr:......................0
  MaxCreditHint:...................0
  RoundTrip:.......................0
--------------------------------------------------------------------------
```

switchinfo example:

An example of use of this command including the Local ID is below:

```
smpquery switchinfo 0x4
```

The resulting information output will be similar to that displayed below:

```
--------------------------------------------------------------------------
 LinearFdbCap:....................49152
 RandomFdbCap:....................0
 McastFdbCap:.....................1024
 LinearFdbTop:....................46
 DefPort:.........................0
 DefMcastPrimPort:................0
 DefMcastNotPrimPort:.............0
 LifeTime:........................15
 StateChange:.....................0
 LidsPerPort:.....................0
 PartEnforceCap:..................32
 InboundPartEnf:..................1
 OutboundPartEnf:.................1
 FilterRawInbound:................1
 FilterRawInbound:................1
 EnhancedPort0:...................0
--------------------------------------------------------------------------
```

7.2.2 perfquery

perfquery uses Performance Management General Services Management Packets (**GMP**) to obtain the PortCounters (basic performance and error counters) from the Performance Management Attributes at the node specified.

Syntax:

```
perfquery [options]  [<lid|guid> [[port] [reset_mask]]]
```

Non standard flags:

-a Show aggregated counters for all port of the destination lid.

-r Reset counters after read.

-R Only reset counters.

Examples

- To read local port's performance counters, enter:

```
perfquery
```

- To read performance counters from lid 32, port 1, enter:

```
perfquery 32 1
```

- To read node aggregated performance counters, enter:

```
perfquery -a 32
```

- To read performance counters and reset, enter:

```
perfquery -r 32 1
```

- To reset performance counters of port 1 only, enter:

```
perfquery -R 32 1
```

- To reset performance counters of all ports, enter:

```
perfquery -R -a 32
```

- To reset only non-error counters of port 2, enter:

```
perfquery -R 32 2 0xf000
```

Example output

The resulting information output will be similar to that displayed below

```
-------------------------------------------------------------------------
 # Port counters: Lid 45 port 2
 PortSelect:......................2
 CounterSelect:...................0x0000
 SymbolErrors:....................0
 LinkRecovers:....................0
 LinkDowned:......................0
 RcvErrors:.......................0
 RcvRemotePhysErrors:.............0
 RcvSwRelayErrors:................0
 XmtDiscards:.....................2
 XmtConstraintErrors:.............0
 RcvConstraintErrors:.............0
 LinkIntegrityErrors:.............0
 ExcBufOverrunErrors:.............0
 VL15Dropped:.....................0
 XmtBytes:........................458424
 RcvBytes:........................1908363
 XmtPkts:.........................6367
 RcvPkts:.........................41748
-------------------------------------------------------------------------
```

7.2.3 ibnetdiscover and ibchecknet

ibnetdiscover is used to scan the topology of the subnet and converts the output into a human readable form. Global IDs, node types, port numbers, port Local IDs and NodeDescriptions are displayed. The full topology is displayed including all nodes and links with the option of highlighting those which are currently connected. The output may be printed to a topology file.

Syntax:

ibnetdiscover [options] [<topology-filename>]

Non standard flags:

-1	List of connected nodes
-H	List of connected HCAs
-S	List of connected switches

ibchecknet uses a topology file which has been created by **ibnetdiscover** to scan the network validating the connectivity and reporting errors detected by the port counters. The command runs as follows.

```
ibchecknet
```

A sample output is displayed below:

```
-------------------------------------------------------------------------
 #warn: counter SymbolErrors = 65535     (threshold 10)
 #warn: counter LinkRecovers = 26        (threshold 10)
 #warn: counter LinkDowned = 16  (threshold 10)
 #warn: counter RcvErrors = 21   (threshold 10)
 #warn: counter RcvSwRelayErrors = 54810         (threshold 100)
 #warn: counter XmtDiscards = 65535      (threshold 100)
 Error check on lid 2 port all:  FAILED
 #warn: counter RcvSwRelayErrors = 3995  (threshold 100)
 Error check on lid 2 port 4:  FAILED
 # Checked Switch: nodeguid 0x0008f104004118d8 with failure

 # Checking Ca: nodeguid 0x0008f10403979970

 # Checking Ca: nodeguid 0x0008f10403979860

 # Checking Ca: nodeguid 0x0008f104039798ec

 # Checking Ca: nodeguid 0x0008f1040397996c

 # Checking Ca: nodeguid 0x0008f104039798e8

 # Checking Ca: nodeguid 0x0008f10403979910

 # Checking Ca: nodeguid 0x0008f104039798e4

 # Checking Ca: nodeguid 0x0008f10403979920

 # Checking Ca: nodeguid 0x0008f10403979948

 # Checking Ca: nodeguid 0x0008f104039798f4

 # Checking Ca: nodeguid 0x0008f104039798d0

 # Checking Ca: nodeguid 0x0008f10403977ca4

 ## Summary: 13 nodes checked, 0 bad nodes found
 ##          24 ports checked, 0 bad ports found
 ##          1 ports have errors beyond threshold
-------------------------------------------------------------------------
```

7.2.4 ibcheckwidth and ibcheckportwidth

ibcheckwidth checks all nodes, using the complete topology file which was created by **ibnetdiscover**, to validate the bandwidth for links which are active and will also identify ports with 1X bandwidth.

```
ibcheckwidth
```

Output Example

```
## Summary: 40 nodes checked, 0 bad nodes found

##         140 ports checked, 0 ports with 1x width in error found
```

ibcheckportwidth checks connectivity and the link width for a given port lid and will indicate the actual bandwidth being used by the port. This should be checked against the maximum which is possible. For example, if the port supports 4 x bandwidth then this should be used. Similarly, if the adapter supports DDR then this should be used.

Syntax:

ibcheckportwidth [-h] [-v] [-G] <lid|guid> <port>

Example:

```
ibcheckportwidth -v 0x2 1
```

Output:

```
Port check lid 0x2 port 1:  OK
```

7.2.5 More Information

Please refer to the man pages for more information on the all tools described in this section and also on the other **OpenIB** tools which are available.

7.3 Troubleshooting Node Deployment

ksis is the deployment tool used to deploy node images on Bull HPC systems. This section describes how deployment problems are logged by **ksis** for different parts of the deployment procedure.

7.3.1 ksis deployment accounting

Following each deployment **ksis** take stock of the nodes, and identifies those that have had the image successfully deployed onto them, and those that have not.

This information is listed in the files below, and remains available until the next image deployment:

- List of nodes successfully deployed to - **/tmp/ksisServer/ksis_nodes_list**
- List of nodes not deployed to - **/tmp/ksisServer/ksis_exclude_nodes_list**

When the image has failed to be deployed to a particular node, **Ksis** adds a line in the **ksis_exclude_nodes_list** file to indicate:

a. The name of the node (between square brackets)

b. The consequences of the problem for the node.
   Three states are possible:
   - **not touched** The node was excluded by the deployment with no impact (for the node).
   - **restored** The configuration of the node was modified, but its initial configuration was able to be restored.
   - **corrupt** The node was corrupted by the operation.

c. The circumstance which led to the deployment problem.

Example:

```
[node2] not touched: node is configured-in
```

Most of the time, the information in the excluded node list allows the source of the problem to be identified, without the need for further analysis.

7.3.2 Possible Deployment Problems

There are 2 areas where deployment problems may occur.

7.3.2.1 Pre-check problems

Before the image is deployed, node states are verified in the **ClusterDB** Database, and through the use of **bsm** commands. If there are any problems, the nodes in question will be excluded for the deployment.
The error will be displayed once the deployment has finished, and will also be logged in the **/tmp/ksisServer/ksis_exclude_nodes_list** file.

7.3.2.2 Image transfer problems

Problems may occur during the phase when the image is being transferred onto the target nodes. These problems are logged and centralised by **Ksis** on the Management Node.

The errors will be displayed once the deployment has finished, and will also be logged in the **/tmp/ksisServer/ksis_exclude_nodes_list** file.

ksis image server logs

ksis server logs are saved on the Management Node in
/var/lib/systemimager/overrides/ka-d-server.log

and
Ksis server traces are saved on the Management Node in
/var/lib/systemimager/overrides/server_log

Note Traces are only possible for the ksis server, and for client nodes, if the **ksis deploy** command is executed using the **–g** option.

ksis image client logs

ksis client logs on the Management Node in
/var/lib/systemimager/overrides/imaging_complete_<nodeIP>
or
/var/lib/systemimager/overrides/patching_complete_<nodeIP>
or
/var/lib/systemimager/overrides/unpatching_complete_<nodeIP>

and ksis client traces on the Management Node in
/var/lib/systemimager/overrides/imaging_complete_error_<nodeIP>

These traces will only be logged if the deployment error occurs on the client side.

Patch deployment client traces on the Management Node in
/var/lib/systemimager/overrides/patching_complete_error_<nodeIP>
or
/var/lib/systemimager/overrides/unpatching_complete_error_<nodeIP>

The client log files will be used during the post-check phase. **Ksis** client and image server errors are compared in order to identify the source of any problems which may occur.

The trace files are kept for support operations.

7.4 Troubleshooting Storage

This section provides some tips to help the administrator troubleshoot a storage configuration.

7.4.1 Verbose Mode (-v Option)

Some of the storage commands have a **–v** (verbose) option, which provides more output information during the processing of the command.

See BAS5 for Xeon *Administrator's Guide* for an inventory of storage commands supporting the **–v** option.

7.4.2 Log/Trace System

Principle

If the verbose mode is not enough, a system of traces can also be configured to obtain more information on some commands. To activate these traces you can set the trace level in the appropriate **/etc/storageadmin/*.conf** file

There are two lines in these files to set the trace. These lines look as follows, where `<command_name>` is the name of the command to debug:

```
#<command_name>_TRACE_STDOUT_LEVEL =
#<command_name>_TRACE_LOG_FILE_LEVEL =
```

The first line is used to activate traces on stdout, the second one is used to generate traces in a **/tmp/storregister.PID.traces** log file. By default the two lines are in comment.

Note It is recommended to use this trace tool only for temporary debugging because there is no automatic cleaning of the **/tmp/<command_name>.PID.traces** log files.

Four levels of traces are available:

- 4 => TRACE_LEVEL_DEBUG
- 3 => TRACE_LEVEL_INFO
- 2 => TRACE_LEVEL_WARNING
- 1 => TRACE_LEVEL_ERROR

Level 4 is the most verbose level, level 1 traces only error messages.

Note It is not possible to add new commands. All the commands accepting this system of traces are listed in the corresponding ***.conf** file.

See BAS5 for Xeon *Administrator's Guide* to identify the right configuration file.

Example

The following example explains how to obtain log file and/or stdout traces on **storregister** command.

1. Find the right **/etc/storageadmin/*.conf** file to modify. In the case of the **storregister** command, it is **storframework.conf** because of the presence of these two lines:
   ```
   # storregister_TRACE_STDOUT_LEVEL =
   # storregister_TRACE_LOG_FILE_LEVEL =
   ```

2. Edit the **storframework.conf** file:
   - Uncomment one of the two previous lines.
   - Choose a level of trace between 1 (lowest) and 4 (highest) level.

   For example, to add traces of debug level (4 = highest level) on stdout only , the **storframework.conf** file must contain the following lines:

   ```
   # STDOUT trace level configuration :
   …
   storregister_TRACE_STDOUT_LEVEL = 4
   …
   # log file trace level configuration :
   # storregister_TRACE_LOG_FILE_LEVEL =
   ```

3. Save the **storframework.conf** file.

4. Relaunch **storregister**. New traces will appear on the stdout.

7.4.3 Available Troubleshooting Options for Storage Commands

The following table sums up the available troubleshooting options for the storage commands.

Command	User Command	-v option	Log/Traces	Name of the corresponding .conf File
fcswregister	Yes			
iorefmgmt	Yes			
ioshowall	Yes			
lsiocfg	Yes	Yes		
lsiodev	Yes			
nec_admin	Yes		Yes	nec_admin.conf
nec_stat	Yes			
stordepha	Yes			
storcheck	Yes		Yes	storframework.conf
stordepmap	Yes	Yes		
stordiskname	Yes			
storiocellctl	Yes		Yes	storframework.conf
storioha	Yes			

Command	User Command	-v option	Log/Traces	Name of the corresponding .conf File
storiopathctl	Yes		Yes	storframework.conf
stormap	Yes	Yes		
stormodelctl	Yes		Yes	storframework.conf
storregister	Yes		Yes	storframework.conf
storstat	Yes		Yes	storframework.conf
stortrapd	No		Yes	storframework.conf
stortraps	No		Yes	storframework.conf

Table 7-1. Troubleshooting options available for storage commands

7.4.4 nec_admin Command for Bull FDA Storage Systems

The **nec_admin** command is used to manage Bull FDA Storage Systems This command interacts with the FDA CLI. A retry mechanism has been implemented to manage the fact that the CLI may reject commands when overloaded. If, despite default setting, the **nec_admin** command occasionally fails, you may change the timeout and retry values defined in the **/etc/storageadmin/nec_admin.conf** file.

```
# Number of retries in case of iSMserver Busy (Not Mandatory)
retry = 3

# If "retry" is set: time in second between two retries (Not
Mandatory)
rtime = 5

# Timeout value : when timeout is reached, the command is considered
as failed
# If number of retries does not exceed the "retry" value, the
# command is launched again, otherwise it is failed.
cmdtimeout = 300
```

See BAS5 for Xeon *Administrator's Guide* for more details about the **nec_admin** command.

7.5 Troubleshooting Lustre

The following section helps you troubleshoot some of the problems affecting your Lustre file system. Because typographic errors in your configuration script or your shell script can cause many kinds of errors, check these files first when something goes wrong.

First be sure your File-system is mounted and you have mandatory user rights.

7.5.1 Hung Nodes

There is no way to clear a hung node except by rebooting. If possible, un-mount the clients, shut down the MDS and OSTs, and shut down the system.

7.5.2 Suspected File System Bug

If you have rebooted the system repeatedly without following complete shutdown procedures, and Lustre appears to be entering recovery mode when you do not expect it, take the following actions to cleanly shut down your system.

1. Stop the login nodes and all other Lustre client nodes. Include the **-F** option with the **lustre_util** command to un-mount the file system.

```
#lustre_util umount -F -f <file_system> -n <node_name>
```

2. Shut down the rest of the system.

3. Run the **e2fsck** command.

7.5.3 Cannot re-install a Lustre File System if the status is CRITICAL

If the status of a file system is CRITICAL (according to the **lustre_util status** command), and if the file system needs to be re-installed (for instance if some nodes of the cluster have been deployed and reconfigured), it is possible that the file system description needs to be removed from the cluster management database, as shown below:

1. Run the following command to install the `fs1` file system:

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

   The command may issue an output similar to:
   `file system already installed, do "remove" first`

2. Run the following command to remove the `fs1` file system:

```
lustre_util remove -f fs1
```

   The command may fail with a message similar to:
   `file system not loaded, try to give the full path`
   If it is not possible to re-install neither remove the file system with force option (-F).

The **lustre_fs_dba** command can then be used to remove the file system information from the cluster management database.
For example, to remove the `fs1` file system description from the cluster management database, enter the following command:

```
lustre_fs_dba del -f fs1
```

After this command the file system can be re-installed using the **lustre_util install** command.

7.5.4 Cannot create file from client

If you get the following error message when you try to create a new file from a Lustre client, it simply means that the user (UID) you use to create the file is not recognized by the Lustre filesystem:

```
touch: cannot touch `/mnt/lustre/myfile': Identifier removed
```

To avoid such problems, all users (UID) that exist on the Lustre client nodes must also exist on the MDS server.

7.5.5 No such device

If the start of the Lustre filesystem fails with the following message, most of the time it is due to the fact that InfiniBand is not properly configured on the Lustre nodes:

```
mount.lustre: mount /dev/ldn.lustrefda2500.4 at
/mnt/srv_lustre/scratch/scratch-OST0003 failed: No such device
Are the lustre modules loaded?
```

This is confirmed by the following lines in the system logs of the machine from which the problem is coming:

```
LustreError: 11602:0:(o2iblnd.c:1569:kiblnd_startup()) Can't query
IPoIB interface ib0: it's down
LustreError: 105-4: Error -100 starting up LNI o2ib
```

Please pay particular attention to the fact that the **IPoIB** interface has to be fully functional in order to start and run Lustre. Despite that fact that Lustre data is not transmitted on the IPoIB interface, IPoIB is used by Lustre to create and manage InfiniBand connections.

7.6 Troubleshooting Lustre File System High Availability

Before using a Lustre file system configured with the High Availability (HA) feature, or in the event of abnormal operation of HA services, it is important to perform a check-up of the Lustre HA file system. This section describes the tools that allow you to make the required checks.

7.6.1 On the Management Node

The following tools must be run from the management node.

lustre_check

This command updates the **lustre_io_nodes** table in the ClusterDB. The **lustre_io_nodes** table provides information about the availability and the state of the I/O nodes and metadata nodes.

lustre_migrate nodestat

This command provides information about the node migrations carried out. It indicates which nodes are supposed to support the OST/MDT services.

In the following example, the MDS are nova5 and nova9, the I/O nodes are nova6 et nova10. nova5 and nova6 have been de-activated, so their services have migrated to their pair-nodes (nova9 and nova10).

```
lustre_migrate nodestat
```

```
HA paired nodes status
---------------------
node name  node status    HA node name  HA node status
    nova5      MIGRATED         nova9              OK
    nova6      MIGRATED         nova10             OK
```

Note This table is updated by the **lustre_check** command.

lustre_migrate hastat [-n <node_name>]

This command indicates how the Lustre failover services are dispatched, after CS4 software has been activated.
Each node has a view on the paired failover services (the failover service dedicated to the node and the failover service dedicated to its pair node). If the pair-node has switched roles, the owner column of the command output will show that this node supports the two lustre_HA services.

In the following example, nova6 and nova10 are paired I/O nodes. The lustre_nova6 service is started on nova10 (owner node). This status is consistent on both nova6 and nova10 nodes.

```
lustre_migrate hastat -n nova[6,10]
```

```
----------------
nova10
----------------
Member Status: Quorate, Group Member
  Member Name                      State      ID
  ------ ----                      -----      --
  nova6                            Online     0x0000000000000001
  nova10                           Online     0x0000000000000002
  Service Name       Owner (Last)               State
  ------- ----       ----- ------               -----
  lustre_nova10      nova10                     started
  lustre_nova6       nova10                     started
----------------
nova6
----------------
Member Status: Quorate, Group Member
  Member Name                      State      ID
  ------ ----                      -----      --
  nova10                           Online     0x0000000000000002
  nova6                            Online     0x0000000000000001
  Service Name       Owner (Last)               State
  ------- ----       ----- ------               -----
  lustre_nova10      nova10                     started
  lustre_nova6       nova10                     started
```

To return to the initial configuration, you should stop `lustre_nova6` which is running on `nova10` and start it on `nova6`, using the `lustre_migrate relocate` command.

lustre_util status

This command displays the current state of the Lustre file systems.

Important

- Sometimes this command can simply indicate that the recovery phase has not finished; in this situation the status will be set to "WARNING" and the remaining time will be displayed.
- When an I/O node have been completely re-installed following a system crash, the Lustre configuration parameters will have been lost for the node. They need to be redeployed from the Management Node by the system administrator. This is done by coping all the configuration files from the Management Node to the I/O node in question by using the scp command as shown below:
  `scp/etc/lustre/conf/<fs_name>.xml<io_node_name>:/etc/lustre/conf/<fs_name>.xml`
  <fs_name> is the name for each file system that was included on the I/O node before the crash.

lustre_util info

This command provides detailed information about the current distribution of the OSTs/MDTs. The services and their status are displayed, along with information about the primary, secondary and active nodes.

/tmp/log/lustre/lustre_HA-*ddmm*.log

This file provides a trace of the commands issued by the nodes to update the LDAP and ClusterDB databases. This information should be compared with the actions performed by CS5.

Note In **lustre_HA-*ddmm*.log**, *dd* specifies the day and *mm* the month of the creation of the file.

/var/log/lustre/HA-DBDaemon=*yy-mm-dd*.log

This file provides a trace of any ClusterDB updates that result from the replication of LDAP. This could be useful if **Lustre** debug is activated at the same time.

7.6.2 On the Nodes of an I/O Pair

The following tools must be run from the I/O nodes.

ioshowall

This command allows the configuration to be checked.
Look at the **/etc/cluster/cluster.conf** file for any problems if the following error is displayed:

```
-- cannot connect to < PAP address> or HWMANAGER
```

Check if the node is an inactive pair-node if the following error appears, otherwise start the node again:

```
-- service lustre_ha inactif
```

clustat

Displays a global status for Cluster Suite 4, from the HA cluster point of view.

Important

If there is a problem, the two pair nodes may not have the same view of the HA cluster state.

storioha -c status

This command checks that all the Cluster Suite 4 processes are running properly ("running state").

Notes
- This command is equivalent to the following one on the Management Node:
  `stordepha -c <status> -i <node>`
- This command is included in the global checking performed by the **ioshowall** command.

stormap -l

This command checks the state of the virtual links.

Note This command is included in the global checking performed by the **ioshowall** command.

lctl dl

This command checks the current status of the OST/MDT services on the node.
For example:

```
1 UP lov fs1_lov-e0000047fcfff680 b02a458d-544e-974f-8c92-23313049885e 4
2 UP osc OSC_nova9_ost_nova6.ddn0.11_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
3 UP osc OSC_nova9_ost_nova10.ddn0.5_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
4 UP osc OSC_nova9_ost_nova6.ddn0.3_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
5 UP osc OSC_nova9_ost_nova10.ddn0.21_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
6 UP osc OSC_nova9_ost_nova6.ddn0.19_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
7 UP osc OSC_nova9_ost_nova10.ddn0.7_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
8 UP osc OSC_nova9_ost_nova6.ddn0.1_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
9 UP osc OSC_nova9_ost_nova10.ddn0.23_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
10 UP osc OSC_nova9_ost_nova6.ddn0.17_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
11 UP osc OSC_nova9_ost_nova10.ddn0.13_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
12 UP osc OSC_nova9_ost_nova6.ddn0.9_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
13 UP osc OSC_nova9_ost_nova10.ddn0.15_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
14 UP mdc MDC_nova9_mdt_nova5.ddn0.25_MNT_clientelan-e0000047fcfff680
b02a458d-544e-974f-8c92-23313049885e 4
```

The last line indicates the state of the MDC, which is the client connecting to the MDT (on the MDS).
The other lines indicate the state of the OSC, which are the clients connecting to each OST (on the `nova6` and `nova10` OSS).

/var/log/lustre/HA_yy-mm-dd.log

This file provides a trace of the calls made by CS5 to the Lustre failover scripts.

Note In the **HA_yy-mm-dd.log** file, yy specifies the year, *mm* the month and *dd* specifies the day of the creation of the file.

/var/log/syslogfile

This file provides a trace of the events and activity of CS5 and Lustre.

Pair Node Consistency

In some very specific cases, it may be necessary to reset the HA system to a state which ensures consistency across the pair-nodes, **without stopping** the Lustre system.

1. Disconnect the `fs1` Lustre File System from the HA system:

```
lustre_ldap unactive -f fs1
```

2. Run **clustat** to view the location of the services:

```
clustat
```

3. Perform one of the following actions:

   – To switch a node from primary state to pair-node state, run:

```
lustre_migrate export -n <node_name>
```

   – Or, to reset the switched node back to its primary state, run:

```
lustre_migrate relocate -n <node_name>
```

4. Re-connect the Lustre File System to the Lustre HA system:

```
lustre_ldap active -f fs1
```

7.7 Troubleshooting SLURM

7.7.1 SLURM does not start

Check that all the RPMs have been installed on the Management Node by running the command below.

```
rpm –qa | grep slurm
```

The following RPMs should be listed:

slurm-x.x.xx-x.Bull

pam_slurm-x.x- x.x.xx-.x.Bull

slurm-munge- x.x.xx-x.Bull

slurm-auth-none- x.x.xx-x.Bull

slurm-devel- x.x.xx-x .Bull

Note The version numbers depend on the release and are indicated by the letter x above.

7.7.2 SLURM is not responding

1. Run the command **scontrol ping** to determine if the primary and backup controllers are responding.

2. If they respond, then there may be a Network or Configuration problem – see section *7.7.5 Networking and Configuration Problems*.

3. If there is no response, log on to the machines to rule out any network problems.

4. Check to see if the **slurmctld** daemon is active by running the following command:

```
ps -ef | grep slurmctld
```

a. If **slurmctld** is not active, restart it as the root user using the following command.

```
service slurm start
```

b. Check the **SlurmctldLogFile** file in the **slurm.conf** file for an indication of why it failed.

c. If **slurmctld** is running but not responding (a very rare situation), then kill and restart it as the root user using the following commands:

```
service slurm stop
service slurm start
```

d. If it hangs again, increase the verbosity of debug messages by increasing **SlurmctldDebug** in the **slurm.conf** file, and restart. Again, check the log file for an indication of why it failed.

5. If SLURM continues to fail without an indication of the failure mode, stop the service, add the controller option "-c" to the **/etc/slurm/slurm.sh** script, as shown below, and restart.

```
service slurm stop
```

```
SLURM_OPTIONS_CONTROLLER="-c"
```

```
service slurm start
```

Note All running jobs and other state information will be lost when using this option.

7.7.3 Jobs are not getting scheduled

1. This is dependent upon the scheduler used by **SLURM**. Run the following command to identify the scheduler.

```
scontrol show config | grep SchedulerType
```

See the Bull HPC *Administrator's Guide* for a description of the different scheduler types.

2. For any scheduler, the priorities of jobs can be checked using the following command:

```
scontrol show job
```

7.7.4 Nodes are getting set to a DOWN state

1. Check to determine why the node is down using the following command:

```
scontrol show node <name>
```

This will show the reason why the node was set as down and the time when this happened. If there is insufficient disk space, memory space, etc. compared to the parameters specified in the **slurm.conf** file, then either fix the node or change **slurm.conf**.

For example, if the temporary disk space specification is `TmpDisk=4096`, but the available temporary disk space falls below 4 GB on the system, **SLURM** marks it as `down`.

2. If the reason is '*Not responding*', then check the communication between the Management Node and the DOWN node by using the following command:

```
ping <address>
```

Check that the <address> specified matches the **NodeAddr** values in the **slurm.conf** file. If ping fails, then fix the network or the address in the **slurm.conf** file.

3. Login to the node that **SLURM** considers to be in a DOWN state and check to see if the **slurmd** daemon is running using the following command:

```
ps -ef | grep slurmd
```

4. If **slurmd** is not running, restart it as the root user using the following command:

```
service slurm start
```

5. Check **SlurmdLogFile** file in the **slurm.conf** file for an indication of why it failed.

   a. If **slurmd** is running but not responding (a very rare situation), then kill and restart it as the root user using the following commands:

```
service slurm stop
service slurm start
```

6. If the node is still not responding, there may be a Network or Configuration problem – see section *7.7.5 Networking and Configuration Problems*.

7. If the node is still not responding, increase the verbosity of debug messages by increasing **SlurmdDebug** in the **slurm.conf** file, and restart. Again, check the log file for an indication of why it failed.

8. If the node is still not responding without an indication as to the failure mode, stop the service, add the daemon option "-c" to the **/etc/slurm/slurm.sh** script, as shown below, and restart.

```
service slurm stop
```

```
SLURM_OPTIONS_DAEMONS=”-c”
```

```
service slurm start
```

Note All running jobs and other state information will be lost when using this option.

7.7.5 Networking and Configuration Problems

1. Use the following command to examine the status of the nodes and partitions:

```
sinfo --all
```

2. Use the following commands to confirm that the control daemons are up and running on all nodes:

```
scontrol ping
scontrol show node
```

3. Check the controller and/or **slurmd** log files (**SlurmctldLog** and **SlurmdLog** in the **slurm.conf** file) for an indication of why a particular node is failing.

4. Check for consistent **slurm.conf** and credential files on the node(s) experiencing problems.

5. If the problem is a user-specific problem, check that the user is configured on the Management Node as well as on the Compute Nodes. The user does not need to be able to login, but his user ID must exist. User authentication must be available on every node. If not, non-root users will be unable to run jobs.

6. Verify that the security mechanism is in place, see chapter 6 in the *Bull HPC BAS5 for Xeon Administrator's Guide* for more information on SLURM and security.

7. Check that a consistent version of SLURM exists on all of the nodes by running one of the following commands:

```
sinfo -V
```

   or

```
rpm -qa | grep slurm
```

   If the first two digits of the version number match, it should work fine. However, version 1.1 commands will not work with version 1.2 daemons or vice-versa.

   Errors can result unless all these conditions are true.

8. Each node must be synchronized to the correct time. Communication errors occur if the node clocks differ.

   Execute the following command to confirm that all nodes display the same time:

```
pdsh -a date
```

   To check a group of nodes use the following command:

```
pdsh w <node list> date
```

   A matter of a few seconds is inconsequential, but SLURM is unable to recognize the credentials of nodes that are more than 5 minutes out of synchronization. See the *Bull HPC BAS5 for Xeon Installation and Configuration Guide* for information on setting node times using the **NTP** protocol.

7.7.6 More Information

For more information on SLURM Troubleshooting see the *Bull HPC BAS5 for Xeon Administrator's Guide*, *Bull HPC BAS5 for Xeon User's Guide* and http://www.llnl.gov/linux/slurm/slurm.html

7.8 Troubleshooting FLEXlm License Manager

7.8.1 Entering License File Data

You can edit the hostname on the server line (first argument), the port address (third argument), the path to the vendor-daemon on the VENDOR line (if present), or any right half of a string (b) of the form a=b where (a) is all lower case. Any other changes will invalidate the license.
Be cautious when transferring data received by Mailers. Many Mailers add characters at the end-of-line that may confuse the reader about the real license data.

7.8.2 Using the lmdiag utility

The **lmdiag** command analyzes a license file with respect to the SERVER, the FEATUREs, license counts and dates. It may help you to understand problems that may occur. **lmdiag** attempts to checkout all FEATUREs and explains failures. You may run extended diagnostics attempting to connect to the license manager on each port on the host.

7.8.3 Using INTEL_LMD_DEBUG Environment Variable

Setting this environment variable will cause the application to produce product diagnostic information at every checkout.

Daemon Startup Problems.

Cannot find license file. Most products have a default location in their directory hierarchy (or use **/opt/intel/licenses/server.lic**). The environment variable INTEL_LICENSE_FILE names this directory. Startup may fail if these variables are set wrong, or the default location for the license is missing.

No such Feature exists

The most common reason for this is that the wrong license file, or an outdated copy of the file, is being used.

Retrying Socket Bind

This means the TCP port number is already in use. Almost always, this means an **lmgrd.intel** is already running, and you have tried to start it twice. Sometimes it means that another program is using this TCP port number. The number is listed on the SERVER line in the license file as the last item. You can change the number and restart **lmgrd.intel**, but only do this if you do not already have an **lmgrd.intel** running for this license file.

INTEL: cannot initialise

```
(INTEL) FLEXlm version 7.2
(lmgrd) Please correct problem and restart daemons
```

You may be starting the **lmgrd.intel** from the wrong directory, or with relative paths. Use the following lines in the start up and add a full root path to 'INTEL' to the end of the VENDOR line in the license file:

```
cd <installation-directory>
`pwd`/lmgrd.intel -c `pwd`/server.lic -l `pwd`/lmgrd.intel.log
```

License manager: cannot initialize: Cannot find license file

You have started **lmgrd.intel** on a non-existent file. The recommended way to specify the file for **lmgrd.intel** to use -c <license>:

```
cd <installation-directory>
`pwd`/lmgrd.intel -c `pwd`/server.lic -l `pwd`/lmgrd.intel.log
```

Invalid license key (inconsistent encryption code for 'FEATURE')

This happens for 3 different reasons:

1. The license file has been typed in incorrectly.

   (Cutting and pasting from email is a safe way to avoid this). Or the data have been altered by the end user. See "Entering License File Data" above.

2. The license is generated incorrectly. Your vendor will have to generate a new license if this is the case.

3. The license vendor has changed encryption seeds (rare).

MULTIPLE vendor-daemon-name servers running

There are 2 **lmgrd** and vendor-daemons running for this license file. Only one process per vendor-daemon/per node is allowed to run. Sometimes this can happen because the **lmgrd** was killed with a -9 signal (which should not be done!). The **lmgrd** was then not able to bring the vendor-daemon process down, so it's still running, although not able to serve licenses.

If **lmgrd** is killed with a -9, the vendor-daemons also then must be killed with a -9 signal. In general, **lmdown** should be used.

Vendor daemon cannot talk to lmgrd

This means a pre-version-3.0 **lmgrd** version is being used with a 3.0+ vendor daemon. Simply use the latest version of **lmgrd** (MUST be a version equal to or greater than the vendor daemon version). This can also happen if TCP networking does not function on the node where you are trying to run **lmgrd** (rare).

No licenses to serve

The license file has only 'uncounted' licenses, and these do not require a server. Uncounted licenses have a '0' or 'uncounted' in the 'number-of-licenses' field on the FEATURE line.

Other Starting **lmgrd.intel** from a remote directory may lead to unknown results. If **lmgrd.intel** is started from a remote directory the license file line:

```
VENDOR INTEL
```

Should be modified to include the root directory where the 'INTEL' vendor daemon resides:

```
VENDOR INTEL <root-directory-path>
```

The **lmgrd.intel** daemon MUST be started with the -c argument:

```
cd <installation-directory>
`pwd`/lmgrd.intel -c `pwd`/server.lic -l `pwd`/lmgrd.intel.log
```

Application Execution Problems

```
Cannot connect to license server
```

Usually this means the server is not running. It can also mean the server is using a different copy of the license file, which has a different port number than the license file you are currently using indicates. You can use the **lmdiag** utility to more fully analyze this error.

License Server does not support this Feature

This means the server is using a different copy of the license file than the application. They should be synchronized. This error will also report "UNSUPPORTED" in the debug log file.

Invalid Host

You may be attempting to run the application on a host not listed in the "HOSTID" field of your license. Use **lmhostid** to find the hostid number for the current host.

```
Cannot find license file.  No such file or directory
Expected license file location: <path>
```

The application was not able to find a license file. It gives you the location(s) where it was looking for a license file.

Check that the named file exists. To use a file at a different location, use the environment variable INTEL_LICENSE_FILE.

No such Feature exists

The license manager cannot find a ‘FEATURE’ line in the license file.

Feature has expired

Your license has expired. The system time may be set incorrectly. Run the 'date' command to make sure the date is not later than the Expiration Date listed in the license file.

```
<FEATURE name>: Invalid (inconsistent) license key
```

The license-key and data for the feature do not match. This usually happens when a license file has been altered. See "Entering License File Data" above.

System Bootup Problems

For reasons unknown some bootup files (/etc/rc, /sbin/rc2.d, etc) refuse to run **lmgrd** with the simple commands indicated above. Here are two workarounds:

1. Use 'nohup su username -c 'umask 022;lmgrd -c ...' (It is not recommended to run **lmgrd** as root; the "su username" is used to run **lmgrd** as a non-privileged user.)

2. Add 'sleep 2' after the **lmgrd** command.

7.9 Troubleshooting the Bull Cool Cabinet Door

7.9.1 /usr/sbin/coldoorRecord CLI not found

Check that the RPM have been installed on the Management Node by running the command below.

```
rpm –qa | grep coldoor-record
```

The following RPM should be listed:

coldoor-record-x.x-Bull.x

Note The version number depends on the release and are indicated by the letter x above.

7.9.2 No Cool Cabinet Door found

1. Check the Cool Cabinet Door is electrically plugged-in.

2. Run the commands:

```
su - postgres
psql -U clusterdb clusterdb
```

```
<Enter Password>

clusterdb=> SELECT e.admin_ipaddr, rp.id, rp.admin_eth_switch_id,
rp.admin_eth_switch_slot, rp.admin_eth_switch_port, rp.admin_ipaddr
FROM rack_port rp, eth_switch e WHERE e.id = rp.id and e.status !=
'not_managed';
```

This should return the Cool Cabinet Doors configured if any, in the following format:

```
admin_ipaddr|id|admin_eth_switch_id|admin_eth_switch_slot
        |admin_eth_switch_port|admin_ipaddr
```

Example:

```
---------------+----+------------+--------------+----+------------
172.17.0.210   |  0 |           0 |              0 | 23 | 172.17.0.103
```

It means that the Cool Cabinet Door whose IP address is `172.17.0.103` is connected to switch `172.17.0.210` on port `23`, slot `0`.

3. Check the wiring configuration:

   The Cool Cabinet Doors must be connected to the appropriate switch, as defined in the Cluster Database, and as returned by the previous **psql** command, above.

Chapter 8. Upgrading Emulex HBA Firmware

This chapter describes the following tasks:

- 8.1 *Upgrading Emulex Firmware on a Node*
- 8.2 *Upgrading Emulex Firmware on Multiple Nodes*

8.1 Upgrading Emulex Firmware on a Node

8.1.1 Using lptools

lptools is a set of two utilities for upgrading Emulex HBA firmware. These two utilities are:

- **lputil**: low level tool used to interact with Emulex HBA
- **lpflash**: high level script used to upgrade firmware of a set of Emulex HBA.

Emulex driver (**lpfc** module) has to be loaded when using **lptools** (check with **lsmod**). Firmware updates are available from Emulex Web site.

On a node, you can get the current FW level from all the Emulex HBA using the **lsiocfg** tool ("getting information about storage devices").

WARNING
Be sure that FC devices are not being used when upgrading the Emulex HBA firmware.

8.1.2 lputil

This low level tool should not be used in standalone mode. Please refer to on-line help when using this tool.

8.1.3 lpflash

lpflash flashes Emulex HBAs with the specified firmware file. **lpflash** may be used to upgrade in one shot all the HBAs on a server.

Syntax:

```
lpflash <-m LP_Model -f path_to_firmware [-v]> | <-h> | <-V>
```

Flags:

-m model	Emulex HBA model to flash (case insensitive)
-f file	firmware file
-v	verbose mode
-h	displays help
-V	displays version

Example:

```
lpflash -m lp11000 -f  /tmp/bd210a7.all
```

This command will upgrade all LP11000 HBA to `2.10A7` firmware.

8.2 Upgrading Emulex Firmware on Multiple Nodes

Running the **pdcp** / **pdsh** commands, Emulex firmware can be upgraded in one shot on a set of nodes:

- use **pdcp** to copy the new firmware file on all the nodes
- use **pdsh** to run **lpflash** on these nodes.

Example:

The following commands copy the Emulex firmware file on to nodes `node1`, `node2` and `node3`, and then upgrade all Emulex LP11000 HBA on these nodes with firmware 2.10A7:

```
pdcp -w “node1,node2,node3” bd210a7.all /tmp/
pdsh -w “node1,node2,node3” lpflash -m lp11000 -f /tmp/bd210a7.all
```

Chapter 9. Updating the InfiniBand Switches Firmware

Voltaire switches should be properly configured to ensure maximum performance. For example, **Voltaire** switch firmware version 00.08.06 ASIC does not utilise Double Data Rate transfer for those links which include **Mellanox** cards and should be upgraded. The **Voltaire** switch firmware upgrade procedure is described below.

9.1 Checking which Firmware Version is running

Go to the **utilities** menu as follows:

```
ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
Connecting
```

```
switchname # utilities
switchname (utilities)#
```

Once in the **utilities** menu, check which firmware version is installed:

```
switchname(utilities)# firmware_verify_anafa_II
```

```
Scan Fabric
Default fw_version is 00.08.06
```

9.2 Configuring FTP for the firmware upgrade

If the switch firmware requires an upgrade, the FTP options for the switch will need to be set. These may already be in place following the initial Installation and Configuration of the cluster. If not, they are put into place as follows:

9.2.1 Installing the FTP Server

To install the FTP server (**vsftpd**), proceed as follows:

```
rpm -ivh /<path_to_vsftpd-<version>-<arch>.rpm>
```

By default, the **vsftpd** daemon will not allow root access to the FTP server. For security reasons, it is advised to create a dedicated user for this purpose. However, if you wish to enable root access to the FTP server, **vsftpd** can be enabled to allow this as follows:

1. Edit **/etc/vsftpd.ftpusers** file and comment out the line that starts by root, as follows:

```
# Users that are not allowed to login via ftp
# root
Bin
```

2. Edit **/etc/vsftpd.ftpuser_list** and comment out the line that starts by root, as follows:

```
/etc/vsftpd.user_list
# vsftpd userlist
# If userlist_deny=NO, only allow users in this file
# If userlist_deny=YES (default), never allow users in this file, and
# do not even prompt for a password.
# Note that the default vsftpd pam config also checks
/etc/vsftpd.ftpusers
# for users that are denied.
# root
bin
```

3. Start the **vsftpd** server as follows:

```
[root@host ~]# service vsftpd start
```

```
Starting vsftpd for vsftpd:          [  OK  ]
```

4. Check that FTP is working correctly:

```
[root@host ~]# ftp host
```

```
Connected to host.
220 (vsFTPd 2.0.1)
530 Please login with USER and PASS.
530 Please login with USER and PASS.
KERBEROS_V4 rejected as an authentication type
Name (host:root): root
331 Please specify the password.
Password:
230 Login successful.
Remote system type is UNIX.
Using binary mode to transfer files.
ftp> quit
221 Goodbye.
```

9.2.2 Configuring the FTP server options for the InfiniBand switch

Enter the FTP configuration menu as follows:

```
ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
connecting
```

```
switchname # config
switchname (config)# ftp
switchname (config-ftp)#
```

The following settings define the node 172.20.0.102 as the FTP server. The switch logs onto this server using Joe's account using the 'yummy' password.

```
switchname (config-ftp)# server 172.20.0.102
switchname (config-ftp)# username joe
switchname (config-ftp)# password yummy
```

Once FTP is set-up on the switch, make sure the FTP server is running on the Management Node:

```
ftp host
```

If ftp fails to connect to the host (as in the example above), it probably means that the FTP server has not been installed on the host.

```
ftp: connect: Connection refused
ftp> quit
```

9.3 Upgrading the firmware

In the following example, it is assumed that the end user stored the firmware in the existing **/path/to/firmware** directory.

1. Extract the firmware archive to the **/path/to/firmware** directory as follows:

```
cd /path/to/firmware
tar -xvf Ver_10.06_fw.1.0.0.tar
```

```
voltaire_fw_images.tar
voltaire_fw_ini.tar
howto_upgrade_voltaire_switch.txt
```

2. Once the firmware has been extracted, log-on to the switch and proceed with the upgrade.

   a. Upgrading the firmware for the whole switch:

```
[user@host ~]# ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
Connecting
```

```
switchname # update firmware chassis /<path_to_firmware>
```

   b. Upgrading the firmware for a specific line-board (line board 4 in the example below):

```
[user@host ~]# ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
connecting
```

```
switchname # update firmware line 4 /<path_to_firmware>
```

   c. Upgrading a fabric board (fabric board number 2 in the example below):

```
[user@host ~]# ssh enable@switchname
```

```
enable@switchname's password: voltaire
Welcome to Voltaire Switch switchname
Connecting
```

```
switchname # update firmware spine 2 /path/to/firmware
```

Note Whenever a line board or a fabric board is replaced, always ensure that it is using the correct firmware.

3. Check that the firmware has upgraded correctly by running the **firmware_verify_anafa_II** command.

```
switchname(utilities)# firmware_verify_anafa_II
```

Chapter 10. Updating the MegaRAID Card Firmware

The **MegaRAID SAS** driver for the **8408E** card is included in the **BAS5 for Xeon** delivery. The **MegaRAID** card will be detected and the driver for it installed automatically during the installation of the **BAS5 for Xeon** software suite.

The **MegaCLI** tool used to update the firmware for the **MegaRAID** card and is available on the **Bull** support CD. The latest firmware file should be downloaded from the **LSI** web site.

Follow the procedure described below to update the firmware:

1. Check the version of the firmware already installed by running the command:

```
/opt/MegaCli -AdpAllInfo -a0
```

This will provide full version and manufacturing date details for the firmware, as shown in the example below:

```
Adapter #0
================================================================
                    Versions
                ================
Product Name    : MegaRAID SAS 8408E
Serial No       : P088043006
FW Package Build: 5.0.1-0053
                    Mfg. Data
                ================
Mfg. Date       : 01/16/07
Rework Date     : 00/00/00
Revision No     : (

                Image Versions In Flash:
                ================
Boot Block Version : R.2.3.2
BIOS Version       : MT25
MPT Version        : MPTFW-01.15.20.00-IT
FW Version         : 1.02.00-0119
WebBIOS Version    : 1.01-24
Ctrl-R Version     : 1.02-007

                Pending Images In Flash
                ================
None
```

Note The following **MegaRAID** card details are also provided when the **AdpAllInfo** command runs: PCI slot info, Hardware Configuration, Settings and Capabilities for the card, Status, Limitations, Devices present, Virtual Drive and Physical Drive Operations supported by the card, Error Counters, and Default Card Settings.

2. Decompress and extract the firmware by running the command below:

```
unzip ~/lsi/5.1.1-0054_SAS_FW_Image_1.03.60-0255.zip
```

```
Archive:  /root/lsi/5.1.1-0054_SAS_FW_Image_1.03.60-0255.zip
  inflating: sasfw.rom
  inflating: 5.1.1-0054_SAS_FW_Image_1.03.60.0255.txt
 extracting: DOS_MegaCLI_1.01.24.zip
```

3. Update the firmware using the MegaCLI tool using the command below:

```
/opt/MegaCli -adpfwflash -f sasfw.rom -a0
```

```
Adapter 0: MegaRAID SAS 8408E
Vendor ID: 0x1000, Device ID: 0x0411

FW version on the controller: 1.02.00-0119
FW version of the image file: 1.03.60-0255
Flashing image to adapter...
Adapter 0: Flash Completed.
```

4. Reboot the server so that the new firmware is activated for the card.

Appendix A. Tips

A.1. Replacing Embedded Management Board (OPMA) in Bull Cool Cabinet Door

Refer to the *R@ck'n Roll & R@ck-to-Build Installation & Service Guide* and the *Cool Cabinet Door Service Guide* for details on replacing the OPMA board.

Important

The ClusterDB should be updated with new Bull Cool Cabinet Door MAC address. Refer to BAS5 for Xeon *Installation and Configuration Guide* for details on the procedure.

Glossary and Acronyms

A

ACT
Administration Configuration Tool

B

BAS
Bull Advanced Server

BIOS
Basic Input Output System

BMC
Baseboard Management Controller

BSBR
Bull System Backup / Restore

BSM
Bull System Manager

C

CLI
Command Line Interface

ClusterDB
Cluster DataBase

D

DDN
Data Direct Networks

DHCP
Dynamic Host Configuration Protocol

E

ECT
Embedded Configuration Tool

F

FDA
Fibre Disk Array

FRU
Field Replaceable Unit

FTP
File Transfer Protocol

G

GCC
GNU C Compiler

GNU
GNU's Not Unix

GPL
General Public License

GUI
Graphical User Interface

GUID
Globally Unique Identifier

H

HBA
Host Bus Adapter

HPC
High Performance Computing

I

IB
InfiniBand

IPMI
Intelligent Platform Management Interface

K

KSIS
Utility for Image Building and Deployment

L

LAN
Local Area Network

LDAP
Lightweight Directory Access Protocol

LUN
Logical Unit Number

M

MAC
Media Access Control (address)

MPI
Message Passing Interface

N

NFS
Network File System

NIS
Network Information Service

NS
NovaScale

NTP
Network Type Protocol

O

OPMA
Embedded Management Board for the Bull Cool Cabinet Door

P

PBS
Portable Batch System

PCI
Peripheral Component Interconnect (Intel)

R

RAID
Redundant Array of Independent Disks

S

SCSI
Small Computer System Interface

SLURM
Simple Linux Utility for Resource Management

SMP
Symmetric Multi Processing

SMT
Symmetric Multi Threading

SNMP
Simple Network Management Protocol

SOL
Serial Over LAN

SSH
Secure Shell

T

TCP

Transmission Control Protocol

TFTP

Trivial File Transfer Protocol

U

UDP

User Datagram Protocol

USB

Universal Serial Bus

W

WWPN

World – Wide Port Name

Index

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

REFERENCE
86 A2 24FA 00