SGI® Altix® XE1200 Cluster Quick Reference Guide

007-4933-001

CONTRIBUTORS

Written by Mark Schwenden

Illustrated by Chrystie Danzer

Production by Mark Schwenden

Additional contributions by Matthias Blankenhaus, Dick Brownell, Edward Mascarenhas, Bradley Palmer, James Rada, Keith Schilling and Louise Westoby

Record of Revision

Version	Description
-001	December 2006 First publication

Contents

SGI XE1200 Cluster Quick Reference Information

Overview

Your SGI Altix XE1200 cluster system ships with a variety of hardware and software documents in both hard copy and soft copy formats. Hard copy documents will be in the packing box and soft copy documents are located on your system hard disk. Additional third-party documentation may be shipped on removable media (CD/DVD) included with your shipment.

This document is intended as an overview of some of the common operations that system administrators may have to perform to set-up, boot, re-configure (upgrade) or troubleshoot the SGI Altix XE1200 cluster. Additional helpful documents shipped with your cluster include:

- *Manufacturing Audit Checklist* (P/N 007-4942-00*x*)
- *Manufacturing Configuration Summary* (P/N 007-4943-00*x*)
- *Manufacturing System Diagram* (P/N 007-4944-00*x*)

The SGI Altix XE1200 cluster is a set of XE servers (called nodes), networked together, that can run parallel programs using a message passing tool like the Message Passing Interface (MPI). The XE1200 cluster is a distributed memory system as opposed to a shared memory system like that used in the SGI Altix 450 or Altix 4700 high-performance compute servers. Instead of passing pointers into a shared virtual address space, parallel processes in an application pass messages and each process has its own dedicated processor and address space.

Just like a multi-processor shared memory system, a cluster can be shared among multiple applications. For instance, one application may run on 16 processors in the cluster while another application runs on a different set of 8 processors. Very large clusters may run dozens of separate, independent applications at the same time.

Typically, each process of an MPI job runs exclusively on a processor. Multiple processes can share a single processor, through standard Linux context switching, but this can have a significant effect on application performance. A parallel program can only finish when all of its sub-processes have finished. If one process is delayed because it is sharing a processor and memory with another application, then the entire parallel program will be delayed. This gets slightly more complicated

when systems have multiple processors (and/or multiple cores) that share memory, but the basic rule is that a process will run on a dedicated processor core.

There are three primary hardware component types in the rackmounted cluster:

- The head node(s)
- Compute nodes,
- Network interconnect components (switches, PCI cards and cables)

The head node is connected to the interconnect network and also to the "outside world", typically via the local area network (LAN). The head node is the point of submittal for all MPI application runs in the cluster. An MPI job is started from the head node and the sub-processes are distributed to the cluster compute nodes from the head node. The main process on the head node will wait for the sub-processes to finish. For large clusters or clusters that run many MPI jobs, multiple head nodes may be used to distribute the load.

The compute nodes are identical computing systems that run the primary processes of MPI applications. These compute nodes are connected to each other through the interconnect network.

The network interconnect components are typically Ethernet or InfiniBand, and MPI messages are passed across this network between the processes. This compute node network does not connect directly to the "outside world" because mixing external and internal cluster network traffic could impact application performance.

Site Plan Verification

Ensure that all site requirements are met before you install and boot your system. If you have questions about the site requirements or you would like to order full-size floor templates for your site, contact a site planning representative by e-mail (site@sgi.com).

Unpacking and Installing a Cluster Rack

When your system is housed in a single rack, the cluster components come rackmounted and cabled together and a document describing how to unpack and install the rack should be included with the system. See the *SGI Altix XE System Rack Installation Instructions* (P/N 007-4902-00*x*). Follow the instructions provided in that manual to safely and properly unpack and install your rack system. Ensure all rack power distribution units are properly plugged in and the circuit breakers are switched to **(On)**. All units within the rack should be connected to power before booting.

Multi-rack cluster systems require connection of special interconnect cables between racks. The *Manufacturing System Diagram* document (P/N 007-4944-00*x*) shipped with your cluster system describes the inter-rack cable connections.

If you have arranged for SGI field personnel to install the system rack(s), contact your service representative. After your cluster rack(s) are installed, refer back to this guide to continue working with your SGI cluster system.

Booting the XE1200 Cluster

Power on any mass storage units attached to your cluster, then press the power button on the front of the head node (see callout C in Figure 1-1) and let it fully boot. Repeat the process on all the other nodes (compute nodes) in the cluster.

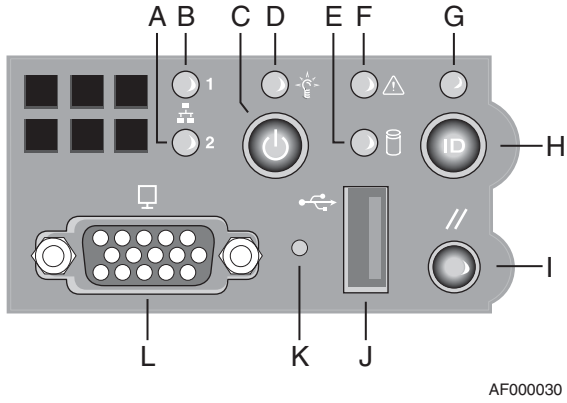

Figure 1-1 Front Controls on the Cluster Nodes

Table 1-1 Cluster node control panel functions

Callout	Feature	Functional description
A	NIC 2 Activity LED	Continuous green light indicates a link between the system and the network interface card to which it is connected.
B	NIC 1 Activity LED	Blinking green light indicates network interface card 1 activity
C	Power/Sleep button	Powers the system On/Off. Puts the system in an ACPI sleep state.

Table 1-1 (continued) Cluster node control panel functions

Callout	Feature	Functional description
D	Power/Sleep LED	Constant green light indicates the system has power applied to it. Blinking green indicates the system is in S1 sleep state. No light indicates the power is off or is in ACPI S4 or S5 state.
E	Hard disk drive activity LED	Blinking green light indicates hard disk activity (SAS or SATA). Unlighted LED indicates no hard disk drive activity.
F	System status LED	Solid green indicates normal operation. Blinking amber indicates degraded performance. Solid amber indicates a critical or non-recoverable condition. No light indicates the system POST is running or the system is off.
G	System Identification LED	Solid blue indicates system identification is active. No light indicates system identification is not active.
H	System Identification Button/LED	Press this button once to activate the System Identification LED. Press the button again to de-activate the System Identification LED. Solid blue indicates system identification is active.No light indicates system identification is not active.
I	Reset Button	Reboots and initializes the system.
J	USB 2.0 port	Allows attachment of a USB component to the front of the node.
K	NMI button	Puts the node in a halt-state for diagnostic purposes.
L	Video Port	Allows attachment of a video monitor to the front of the chassis. Note the front and rear video ports **cannot** be used at the same time.

Cluster Configuration Overview

The following four figures are intended to represent the general types of cluster configurations used with SGI XE1200 systems. Since each cluster shipped from the SGI factory is configured to the customer's specification, these configuration drawings are for informational purposes only and are not meant to represent any specific cluster system.

Figure 1-2 on page 5 diagrams a basic Gigabit Ethernet configuration using a single Ethernet switch for node-to-node communication.

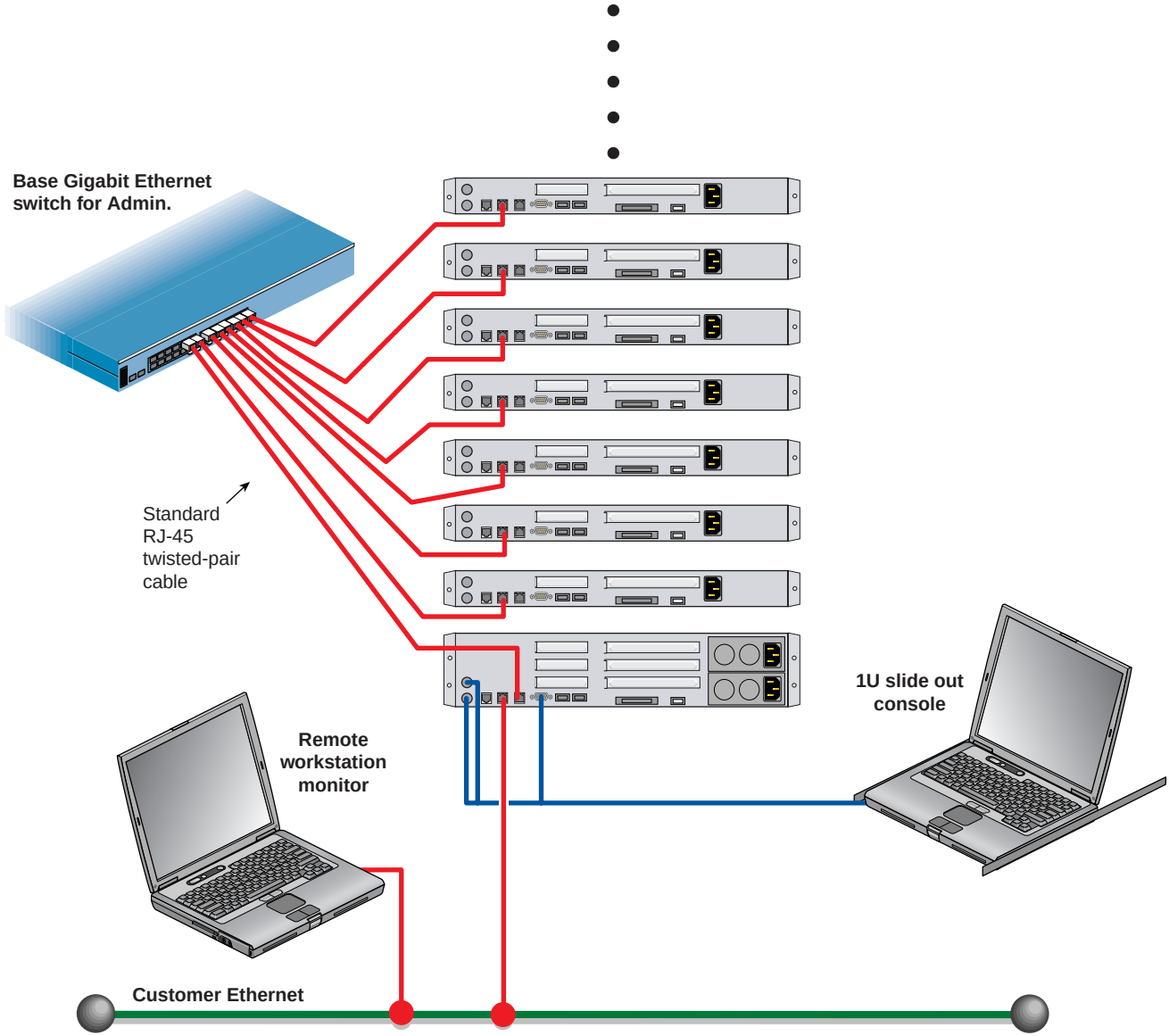

Figure 1-2 Basic Cluster Configuration Example Using a Single Ethernet Switch

Figure 1-3 on page 6 illustrates a dual-switch cluster configuration with one switch handling MPI traffic and the other used for basic cluster administration and communication.

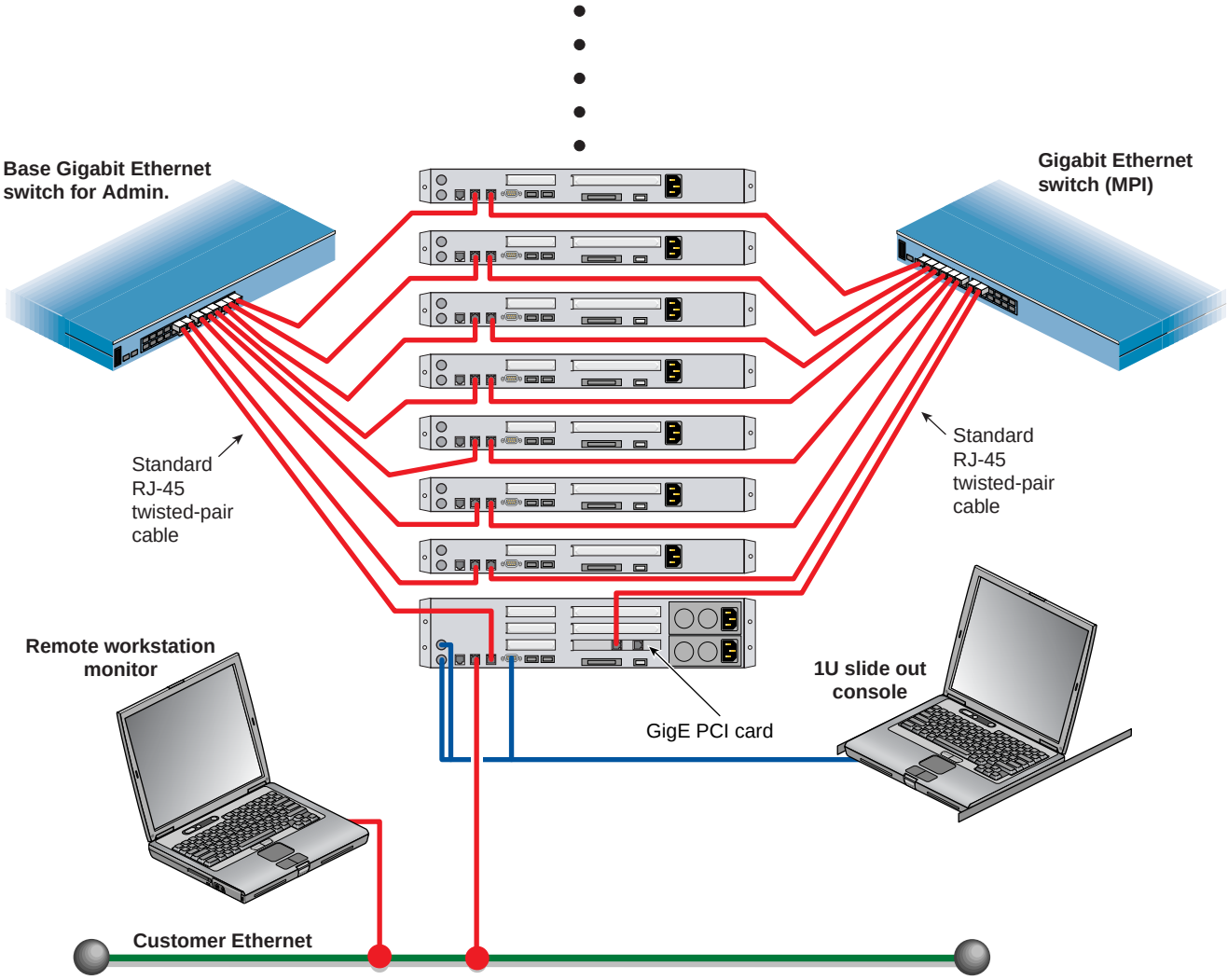

Figure 1-3 Dual-Ethernet Switch Based Cluster Example

Figure 1-4 on page 7 is an example configuration using one Ethernet switch for general administration and one InfiniBand switch for MPI traffic. Figure 1-5 on page 8 shows a configuration with one Ethernet switch for administration, one for NAS and an Infiniband switch for MPI.

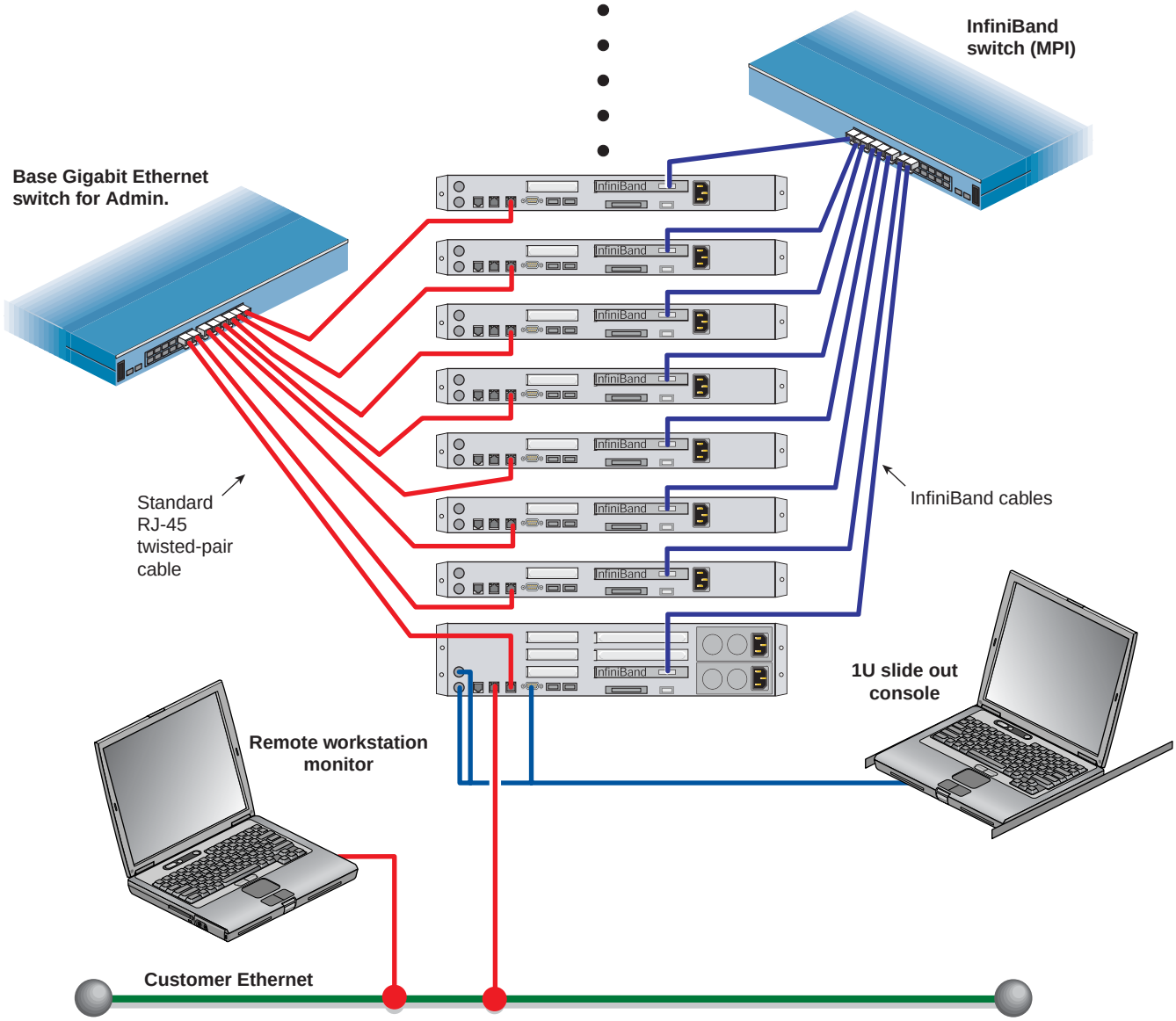

Figure 1-4 Single Ethernet and Single InfiniBand Switch Configuration Example

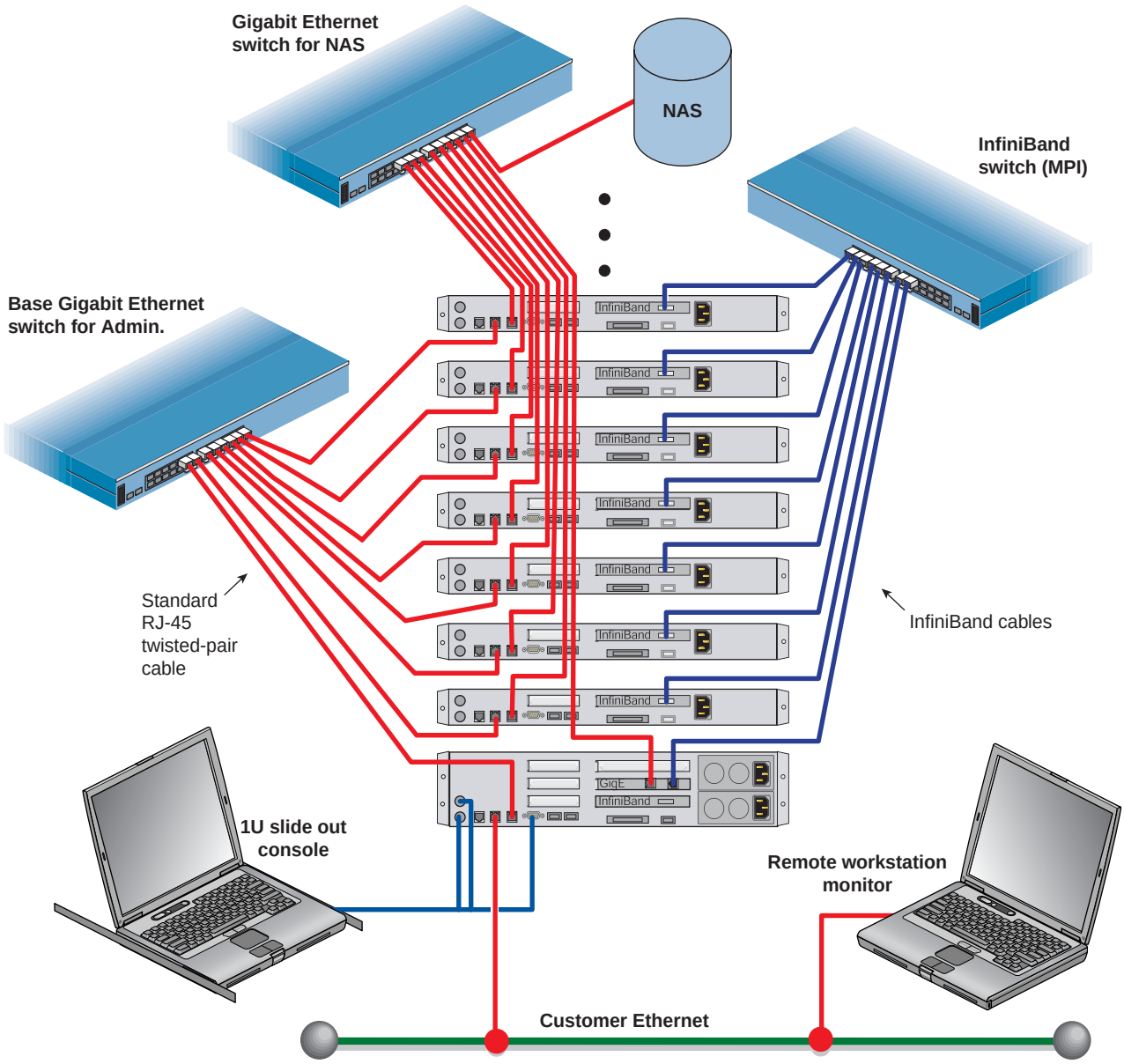

Figure 1-5 Dual Ethernet Plus Infiniband Switch Cluster Configuration Example

Power Down the Cluster

If your cluster uses the Scali Manage administrative software (release 5.3 or later), you can power-off specific nodes or the entire system using the graphical user interface. Select `Management Menu>Power Mgt>Power Off`. The compute nodes can be halted from the Scali GUI by selecting the nodes and choosing "halt system" and "power down" from the System Management menu. A command line interface is also available to power-on/off or check status.

See the *Scali Manage User's Guide* for more information. You must have root privileges to perform these types of tasks.

Powering Off Manually

To power off your cluster system manually, follow these steps:

Caution: If you power off the cluster before you halt the operating system, you can lose data.

1. Shut down the operating system by entering the following command:

   ```
   # init 0
   ```

2. Press the power button on the head node(s) that you want to power off. You may have to hold the button down for up to 5 seconds. You may power off the nodes in any order.
3. To power off the compute nodes, press the power button (for up to 5 seconds) on the front panel of each unit (see Figure 1-1 on page 3).
4. To power off optional storage units in the cluster rack, press the power button(s) on their rear panel to the OFF (**O**) position.

Ethernet Network Interface Card (NIC) Guidelines

While Ethernet ports are potentially variable in a cluster, the following rules generally apply to the cluster head node:

- The server motherboard's nic1 is always the public IP in the head node.
- The server motherboard's nic2 is always the private administrative network connection.
- Nic3 is always a PCI expansion controller port. It is typically used to handle MPI traffic.

Cluster Manager Node (Head Node) IP Addresses

The primary head node of the cluster (head node1) is also known as the cluster management head node. Head node 1 is where the cluster management software is installed and it has the following technical attributes:

- On-board network interface (nic1) IP address is variable (used as public Ethernet access).

  Important: The on-board network interface 1 (nic1) IP address is the factory IP address setting. This setting needs to be changed to reflect the customer domain IP address before connection to the LAN. See the section "Changing the NIC1 (Customer Domain) IP Address" on page 11.

- On-board network interface 2 (nic2) (10.0.10.1) is always used as the management and administration (internal) network port on the primary head node of the cluster.

  Note: In the case of a GigEnet solution, nic3 is used for MPI traffic. In this case nic3 is on an optional I/O module or PCI card.

- The optional Infiniband HCA IP address is 192.168.10.1.
- Board Management Control (BMC) static IP address. The Intelligent Platform Management Interface (IPMI) uses IP address 10.0.30.1 to make controller connections to the BMC.

It is possible to have additional head nodes on the cluster. Table 1-2 lists examples of the headnode port IP address information for more than one head node. Baseboard Management Control routes through nic1 in any additional head nodes added to the cluster. Each fourth octet number in an address iterates by one number as a head node is added.

Table 1-2 Head node Ethernet address listings

Head node number	Internal management IP address nic2	(GigEnet) MPI NAS/SAN option nic3	Infiniband IP address	Baseboard Management Control or IPMI address nic1
1	10.0.10.1	172.16.10.1	192.168.10.1	10.0.30.1
2	10.0.10.2	172.16.10.2	192.168.10.2	10.0.30.2
3	10.0.10.3	172.16.10.3	192.168.10.3	10.0.30.3
4	10.0.10.4	172.16.10.4	192.168.10.4	10.0.30.4

Changing the NIC1 (Customer Domain) IP Address

The “external” IP address assigned to NIC1 must be changed to reflect the new network environment. In addition a set of network parameters specific to your networking environment need to be specified. This can be accomplished by passing a network configuration file to a script. The script is named **config_headnode** and is located at: /usr/local/Factory-Install/Scripts.

The network configuration file contains all relevant parameters used by the nic1 port to communicate with your local area network.

The file is located at: /usr/local/Factory-Install/Scripts/network.cfg

You should edit this file with your favorite Unix or Linux text editor. **Do not** use a Windows based text editor as it may leave carriage return characters at the end of each line causing the script to fail.

The configuration file format is line-oriented, e.g. you cannot break any of the lines into two or more lines. In addition, it adheres to the following formatting rules:

- White spaces are supported.
- Empty lines are supported.
- Comments are line-oriented and begin with the # sign.
- Everything after the hash sign is ignored.
- The variable identifier (all identifiers left to the assignment operators) need to be all uppercase, e.g. NEW_HEAD_NODE_NAME.
- The assignment operator is mandatory.
- Every variable needs to have a value.
- List declarations, (as in DNS_SERVER_LIST in the example configuration file that follows), need to be comma separated.

Only minimal error checking is done to ensure the configuration file has the right format. The user is responsible for the updated local content. You must have root privilege to execute the script.

After you have edited the file with your local addresses and domain information, launch the following script at your head node console to change Scali's network settings:

```
# cd/usr/local/Factory-Install/Scripts/
Then execute the script like this:
# ./config_headnode -f network.cfg
```

After the script has made the necessary changes, it will automatically reconfigure all the compute nodes. Then it will reboot the head node. An example of the configuration file looks like this:

```
# This is a comment.
NEW_HEAD_NODE_NAME=dab
ETH_DEV=eth0
# another comment
NEW_SUBNET_NAME=163.154.16.0/24
NEW_SUBNET_ADDR=163.154.16.0
NEW_SUBNET_MASK=255.255.255.0
NEW_IP_ADDR=163.154.16.197
NEW_GW_ADDR=163.154.16.1
NIS_DOMAIN=engr.sgi.com
NIS_SERVER_LIST=broadcast
DNS_DOMAIN_LIST=engr.sgi.com
DNS_SERVER_LIST=192.26.80.2, 163.154.16.4, 163.154.16.5
NTP_SERVER_LIST=127.127.1.0
```

Cluster Compute Node IP Addresses

The cluster system can have multiple compute nodes that each use up to three IP address points (plus the Infiniband IP address). As with the head nodes, each fourth octet number in an address iterates by one number as a compute node is added to the list. Table 1-3 shows the factory assigned IP address settings for compute nodes one through four.

Table 1-3 Compute node Ethernet address listings

Compute node number	Management IP address nic1	Infiniband IP address	GigEnet solution nic2	Baseboard Management (BMC) or IPMI address nic1
Compute node1	10.0.0.1	192.168.1.1	172.16.1.1	10.0.40.1
Compute node2	10.0.0.2	192.168.1.2	172.16.1.2	10.0.40.2
Compute node3	10.0.0.3	192.168.1.3	172.16.1.3	10.0.40.3
Compute node4	10.0.0.4	192.168.1.4	172.16.1.4	10.0.40.4

Note: The management (internal cluster administration port) IP address and the BMC/IPMI address are shared by the same network interface port (nic1). The circuitry allows the same physical Ethernet port to share two separate IP address references.

SMC Switch Connect and IP Address

Table 1-4 lists the factory IP address for SMC switches that may be used with your cluster.

Web or Telnet Access to the Switch

Your SMC switch(s) setup is configured in the factory before shipment and should be accessible via telnet or a web browser. You can connect to a console directly from the head node through the administration network using telnet.

To access the first SMC switch via telnet:
`telnet 10.0.20.1`

Login as the administrator:
login **`admin`**
passwd: **`admin`**

Web access would be:
http://10.0.20.1

Note: The fourth IP octet grows sequentially for each additional switch. For example, access to SMC switch 2 would be at IP address 10.0.20.2 via telnet or the web.

Table 1-4 SMC Switch IP Addresses

SMC switch number	IP address
SMC switch1 (stacked or single)	10.0.20.1
SMC switch2 (stacked or single)	10.0.20.2
SMC switch3 (stacked or single)	10.0.20.3

SMC Gigabit Ethernet Switch Addressing for NAS/SAN Option

When used with a NAS/SAN option, the SMC switch is configured with the IP addresses shown in Table 1-5. The fourth IP octet grows sequentially for each additional switch used.

Table 1-5 SMC GigEnet NAS/SAN Switch IP Addresses

SMC GigEnet NAS/SAN switch number	IP address
SMC switch1 (stacked or single)	172.16.20.1
SMC switch2 (stacked or single)	172.16.20.2
SMC switch3 (stacked or single)	172.16.20.3

Serial Access to the SMC Switch

Use of a serial interface to the switch should only be needed if the factory assigned IP address for the switch has been somehow deleted, altered or corrupted. Otherwise, use of the web or telnet access procedure is recommended. To use a serial interface with the switch, connect a laptop, or PC to the switch's console port. See Figure 1-6 for the location of the console port.

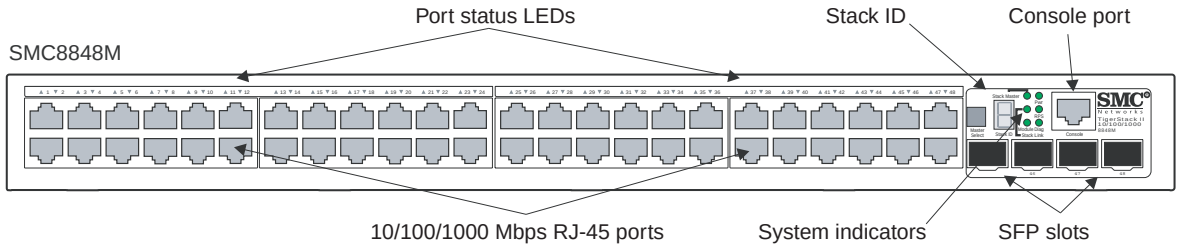

Figure 1-6 SMC Switch Connectors Example

1. Establish a command line interface (CLI) and list the port connection settings:

```
Port Settings
 Bits Per Second=19200
 Data bits=8
 Parity=None
 Stop Bits=1
 Flow Control=none
```

2. In order to verify and save any new settings type the following:

console# **`show running-config`** (make sure your settings are intact)

console# **`copy running-config startup-config`** (it will ask for a file name)

console# **`file name?  startup`**

Note: Any changes made to the switch port settings through the serial interface or Web interface are **not** saved unless the previous steps have been executed.

3. Power cycle the switch by disconnecting and reconnecting its power cable.

InfiniBand Switch Connect and IP Address

Table 1-6 on page 16 lists the factory IP address settings for your InfiniBand switch(s) used with the cluster. For clusters with greater than 288 network ports, consult SGI Professional Services for specific IP address configuration information.

Web or Telnet Access to the Switch

Your InfiniBand switch(s) setup is configured in the factory before shipment and should be accessible via telnet or a web browser.

To access the first InfiniBand switch via telnet:
`telnet 10.0.21.1`

Login as the administrator:
login **`admin`**
passwd: **`123456`**

Web access would be:
http://10.0.21.1

`javaws` (java Webstart) is required for use of the InfiniBand fabric GUI.

SLES 9 service pack 3 location of `javaws` is : /usr/java/j2re1.4.2_12/javaws/javaws

SLES 10 location of `javaws` is : /usr/bin/javaws

Note: The fourth IP octet grows sequentially for each additional switch. For example, access to InfiniBand switch 2 would be at IP address 10.0.21.2 via telnet or the web, see Table 1-6.

Table 1-6 InfiniBand Switch IP Address Listings Example

InfiniBand switch number	IP address
InfiniBand switch1	10.0.21.1
InfiniBand switch2	10.0.21.2
InfiniBand switch3	10.0.21.3

Serial Access to the Switch

You should connect a Voltaire serial cable (either DV-9 to DB-9 or DB-9 to DB-9) that comes with the 24-port switch, from a PC/laptop directly to the switch for serial access. Use of a serial interface to the switch should only be needed if the factory assigned IP address for the switch has been somehow deleted, altered or corrupted. Otherwise, use of the web or telnet access procedure is recommended.

Note: For Voltaire switches 96-ports or larger always use a DB-9 serial cable.

To interface with the switch, use the connected laptop or other PC to:

1. List the port connection settings. Default settings are:

```
Port Settings
Bits Per Second=38400
Data bits=8
Parity=None
Stop Bits=1
Flow Control=xon/xoff
```

Note: For clusters with InfiniBand switches, the fourth Octet IP address will increment for each InfiniBand switch added. See Table 1-6 for an example list.

2. Click "ok" if the settings are acceptable. In the serial interface window on the PC, press enter several times until you see the “ISR-xxxx login:” prompt, then enter the following:
   ISR-xxxx login: **admin**
   ISR-xxxx login: Password: **123456**
   ISR-xxxx> **enable**
   ISR-xxxx> Password: **voltaire**

3. Set up the network for your InfiniBand switch cluster configuration using the following information and Table 1-6 on page 16.

   Enter the following commands to set up the network:
   ISR-xxxx# **config**
   ISR-xxxx(config)# **interface fast**
   ISR-xxxx(config-if-fast)# **ip-address-fast set [10.0.20.x] 255.255.0.0**
   ISR-xxxx(config-if-fast)# **broadcast-fast set 10.0.255.255**
   ISR-xxxx(config-if-fast)# **exit**
   ISR-xxxx(config)# **exit**
   ISR-xxxx# **reset software** (This will reboot the 24-port InfiniBand switch)
   For a 96-port or larger switch:

4. ISR-xxxx# **reload software**
   ISR-xxxx# **fast-interface show** (This will show the IP address)

5. Power cycle the switch by disconnecting its power cable from the plug and then plugging it back in.

Using the 1U Console Option

The SGI optional 1U console is a rackmountable unit that includes a built-in keyboard/touchpad, and uses a 17-inch (43 cm) LCD flat panel display of up to 1024 x 768 pixels. The 1U console attaches to the headnode using PS/2 and HD15M connectors or to a KVM switch (not provided by SGI). The 1U console is basically a “dumb” VGA terminal, it cannot be used as a workstation or loaded with any system administration program.

Note: While the 1U console is normally plugged into the head node on the cluster, it can be connected to any node in the system for terminal access purposes.

The 27-pound (12.27kg) console automatically goes into sleep mode when the monitor cover is closed down.

Installing or Updating Software

Scali Manage offers a mechanism to upload and install software across the cluster. This upload and installation process requires that the software installation be in RPM format. Tarball software distributions can be installed across a cluster. Please see the Scali "scarcp" (cluster remote copy) and the "scash" (cluster remote shell) commands in the *Scali Manage User's Guide*.

Instructions for installing software options or uploading additional software for your cluster using the Scali GUI are covered in Chapter 3 of the *Scali Manage User's Guide*.

Your integrated cluster also comes with a NFS mounted filesystem. The head node exports a /data1 directory. Each compute node mounts this exported filesystem on /cluster. This can be used as a mechanism to install software across the cluster as well.

Customers with support contracts needing BIOS or Firmware updates, should check the SGI Supportfolio Web Page at:

https://support.sgi.com/login

Accessing BIOS Information

BIOS Setup Utility options are used to change server configuration defaults. You can run BIOS Setup with or without an operating system being present.

Important: The BIOS comes preconfigured with the SGI recommended settings. Changes to any of the BIOS settings can impact the performance of your cluster.

You can enter and start the BIOS Setup Utility after you apply power to a head node or compute node (with a console attached) and the Power-On Self Test (POST) completes the memory test. During the POST, you will see this prompt:

```
Press <F2> to enter SETUP
```

When CMOS/NVRAM has been corrupted, you will see other prompts but not the <F2>prompt:

```
Warning: CMOS checksum invalid
Warning: CMOS time and date not set
```

Under these circumstances, you should contact your SGI service representative.

Scali Manage Troubleshooting Tips

This section describes some general guidelines as well as emergency procedures.

Whenever a Scali cluster parameter is changed, it is necessary to apply the configuration. This can be done either through the GUI (Provisioning > Apply All Configuration Changes) or via CLI: scalimanage-cli reconfigure all. Changes can be made in batches and then applied all at once.

There are situations when the GUI does not reflect the cluster configuration properly. Restarting the GUI may solve this problem.

In rare cases the Scali product enters an inconsistent state. In this state it shows abnormal behavior and refuses to take any input. In this case try to reinitialize the head node via /etc/init.d/scance restart.

This must be run on the head node. If this does not change Scali's state, then you should reboot the head node. This should ensure that Scali will be in a consistent state. If you have trouble that is more hardware related, see “Customer Service and Removing Parts” on page 20.

NFS Quick Reference Points

The cluster head node exports an NFS, compute nodes import NFS on the head node. The cluster comes with a preconfigured NFS mount. The headnode exports the /data filesystem.The compute nodes mount head node /data1 on /cluster.

You need to execute the following commands to export a filesystem via NFS from the head node:

```
# scalimanage-cli addnfsexport <head_node> <filesystem>
# /etc/init.d/scance restart
```

To import this filesystem on a particular compute node:

```
# scalimanage-cli addremotefs <compute_node> nfs <head_node:/filesystem> <mount_point>
# scalimanage-cli reconfigure <compute_node>
```

If the compute nodes need to mount filesystems located outside the cluster, then NAT must be enabled on the head node. You need to execute the following commands on the head node:

```
# scalimanage-cli addnatservice <head_node> <ethernet_dev>
# /etc/init.d/scance restart
```

Now you can access nodes outside the cluster from your compute nodes.

To mount a remote filesystem residing outside the cluster on a particular compute node you need to do the following:

```
# scalimanage-cli addremotefs <compute_node> nfs <external_node:/filesystem> <mount_point>
# scalimanage-cli reconfigure <compute_node>
```

Customer Service and Removing Parts

If you are experiencing trouble with the cluster and determine that a replacement part will be needed, please contact your SGI service representative using the information in “Contacting the SGI Customer Service Center”. Return postage information is included with replacement parts.

Removal and replacement of the hardware components that make up the head and compute nodes within the cluster are fully documented in:
SGI Altix XE240 User’s Guide (P/N 007-4873-00*x*)
and
SGI Altix XE210 User’s Guide (P/N 007-4870-00*x*)

These documents can be used to help troubleshoot node-level hardware problems and are included as soft copy (PDF format) on the head node’s system disk at:
/usr/local/Factory-Install/Docs

You can also down-load these documents via internet, from the SGI publications library at:

http://docs.sgi.com

If you need to replace a node within your cluster, go to the SGI Supportfolio web page:

https://support.sgi.com/login

Contacting the SGI Customer Service Center

To contact the SGI Customer Service Center, call 1-800-800-4SGI, or visit:
http://www.sgi.com/support/customerservice.html

From outside the United States contact your local SGI sales office.

Cluster Administration Training from SGI

SGI offers customer training classes covering all current systems, including clusters. If you have a maintenance agreement in place with SGI, contact SGI Customer Education at 1-800-361-2621 for information on the time, location and cost of the applicable training course you are interested in. Or, go to the following URL site for more education information:

http://www.sgi.com/support/custeducation/

Customers with support contracts can also obtain information from:

https://support.sgi.com/login

Administrative Tips and Adding a Node

This chapter provides general administrative information section and information on starting and using the Scali Manage GUI to add a node in a Scali managed cluster. For information on using the Scali Manage command line interface to add a node, see Chapter 8 in the *Scali Manage User's Guide*. Basic information on starting Scali Manage, administrative passwords and factory installed files and scripts are covered in the first section of this chapter, "Administrative Tips" on page 24.

Add a node to the cluster using the following sections and accompanying screen snaps:

Set a node failure "alarm" using the information in:

Administrative Tips

Root password and administrative information includes:

- Root passwd = **sgisgi** (head node and compute nodes)
- Ipmitool user/password info: User = **admin** Password = **admin**

See Table 1-1 on page 3 and Table 1-2 on page 10 for listings of the IPMI IP addresses for nodes.

The ipmitool command syntax for compute nodes (run via the Scali Manage head node):

-ipmitool –I lanplus –o intelplus –H [ip address] -U admin –P admin [command]
Note that to access the BMC (ipmi) directly from the host system, (as opposed to IPMI over LAN), you must load the following kernel modules:

 # modprobe ipmi_msghandler

 # modprobe ipmi_devintf

 # modprobe ipmi_si

Following is the ipmitool command syntax for running directly from the head node (or any node) on itself. (note the absence of the –I –o and –H options): **ipmitool U admin –P admin [command**

If you want these loaded automatically at boot time, For SLES: add them to the `MODULES_LOADED_ON_BOOT` in the /etc/sysconfig/kernel file.

The Scali Manage installer directory**: /usr/local/Scali###**
Is the location of the bits used to install Scali Cluster management Software. Note that most of the software packages used to install your cluster environment (Scali Manage and supporting cluster applications) are stored on the head node system disk.

The Factory install directory is located on the head node server at **/usr/local/Factory-Install.** The /Factory-Install directory contains software files that support the cluster integration and many files and scripts that may be helpful, including:

/Factory-Install/Apps: Scali, ibhost, intel compiler/mpi runtime libraries, ipmitool, etc.

/Factory-Install/ISO: CD iso’s of the base OS; for installing Scali Cluster Manage software

/Factory-Install/Docs: Cluster documentation manuals (Scali, PBS Pro, Voltaire, SMC, SGI)

/Factory-Install/Firmware: Voltaire HCA and Voltaire switch firmware files, etc.

/Factory-Install/CFG: Cluster configuration files

Start the Scali Manage GUI

Login to the Scali Manage interface as root, the factory password is **sgisgi**. Use your system name and log in as root. See Figure 2-1 for an example.

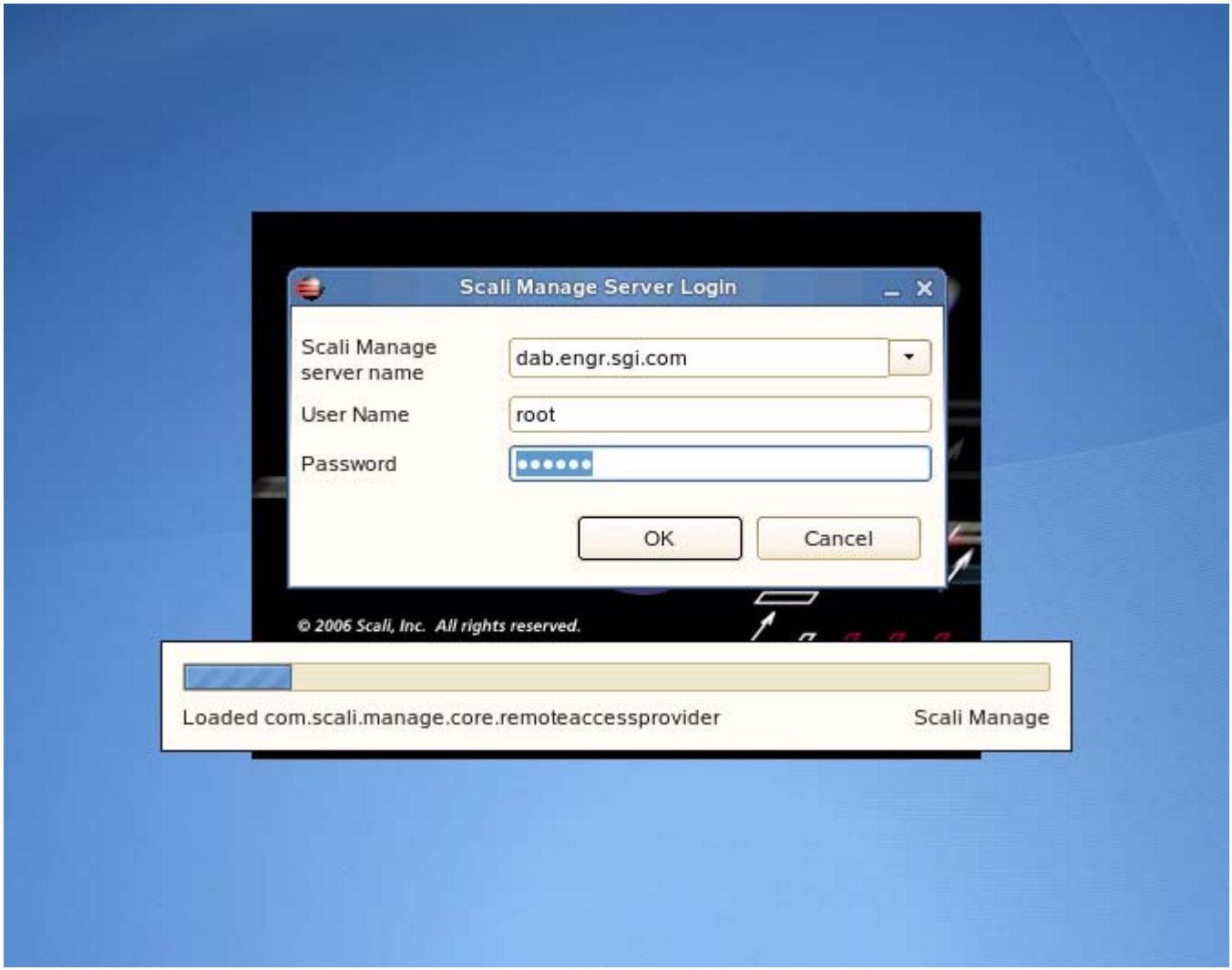

Figure 2-1 Example Starting Screen for the Scali Manage GUI

Head Node Information Screen

You can view and confirm the head node information from the main GUI screen. Click on the node icon (three red stripes) for name and subnet information on your cluster head node.

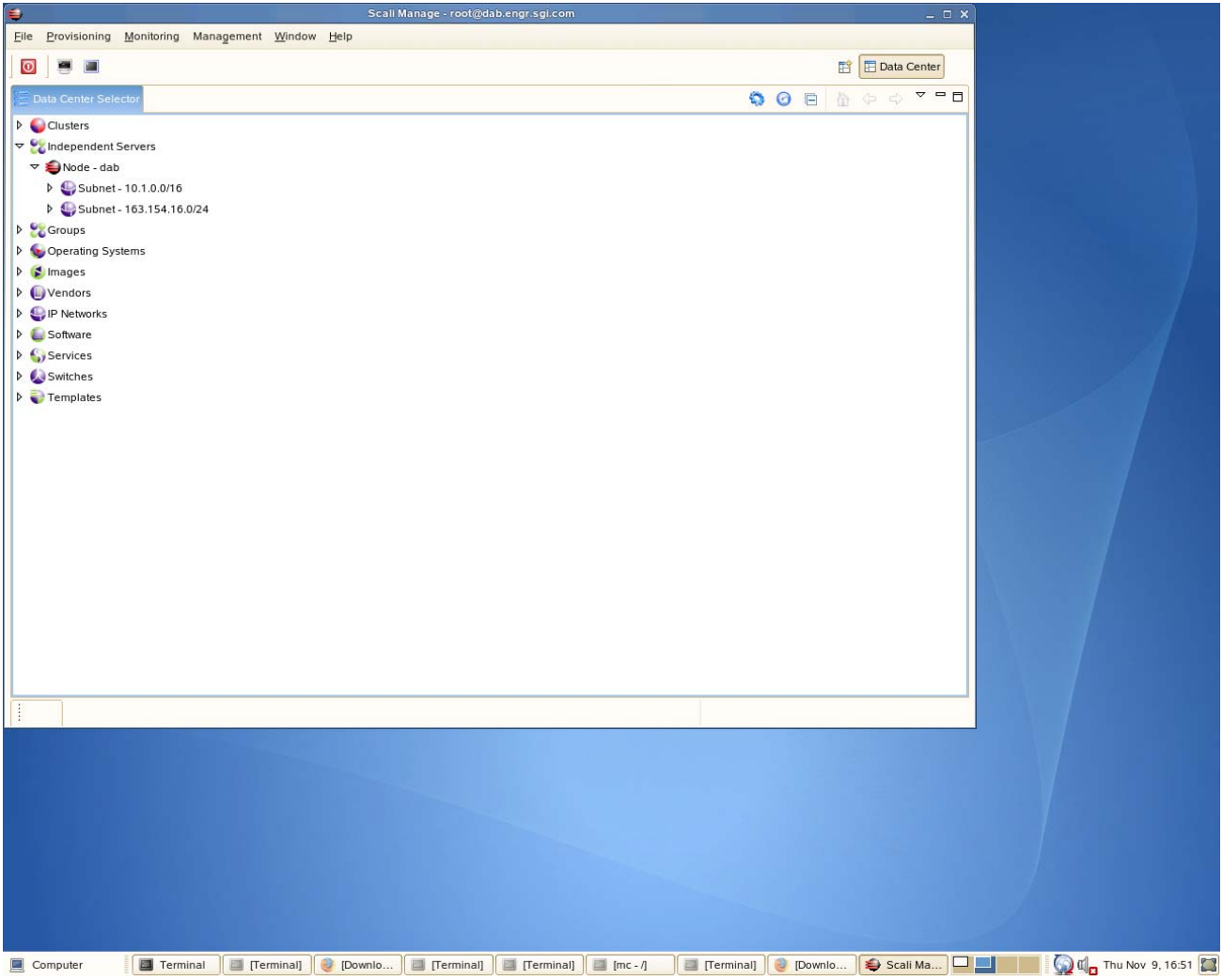

Figure 2-2 Head Node Information Screen Example

Adding a Node Starting from the Main GUI Screen

Add a node when you need to upgrade. To add a cluster node, open the Clusters tree by clicking the right mouse button. Move your cursor over the cluster tree (cluster cl1 in the example screen), and click the right mouse button. Then click the left mouse button on the “New” popup window.

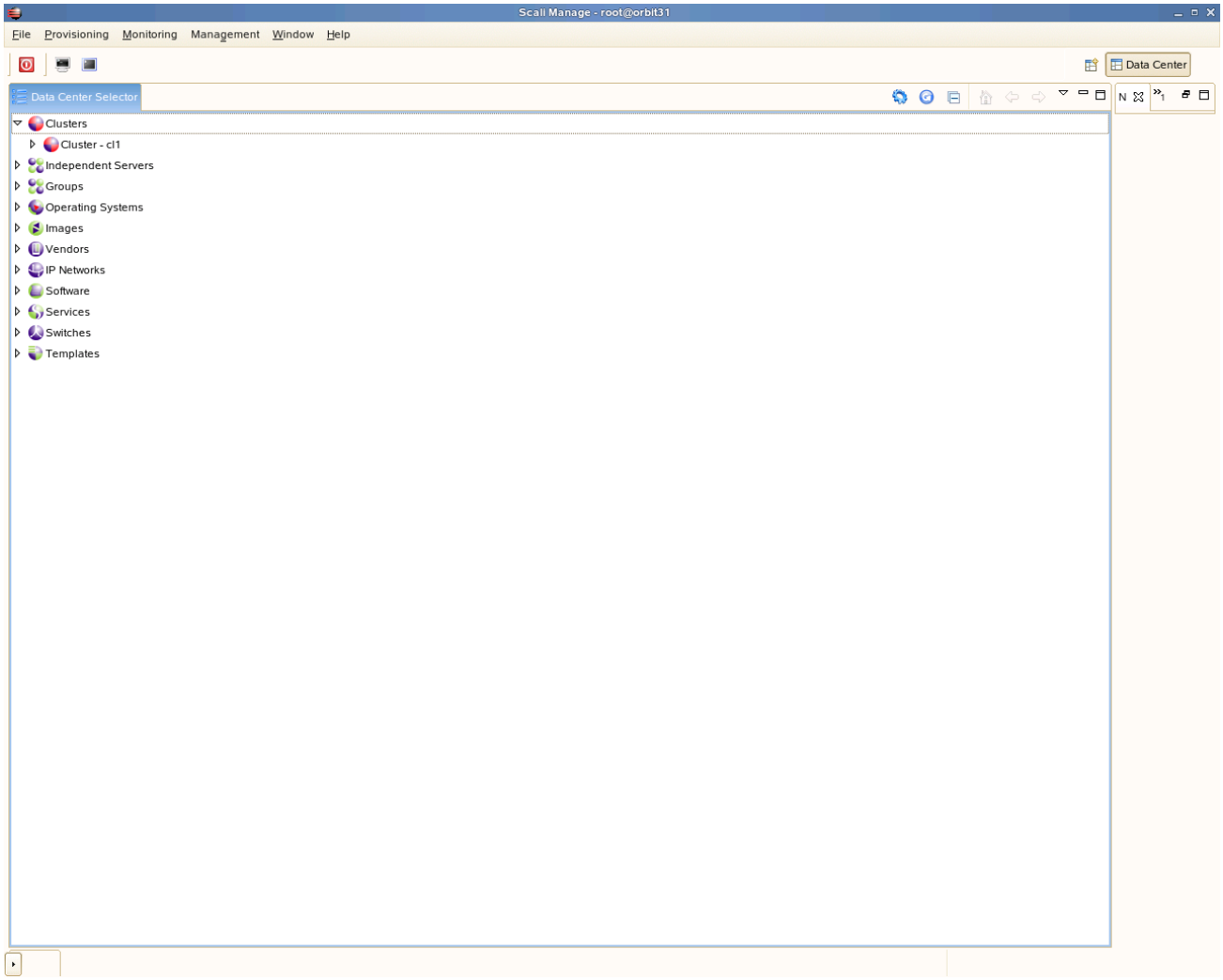

Figure 2-3 Scali Manage Main Screen Selections Example

Adding a Cluster Compute Node

These steps should only be taken if the cluster needs to be upgraded or re-created. Select the option “Extend existing cluster” and provide the number of servers (1 in the example). Then select the “Cluster Name” (cl1 in the example). Click “Next” to move to the following screen.

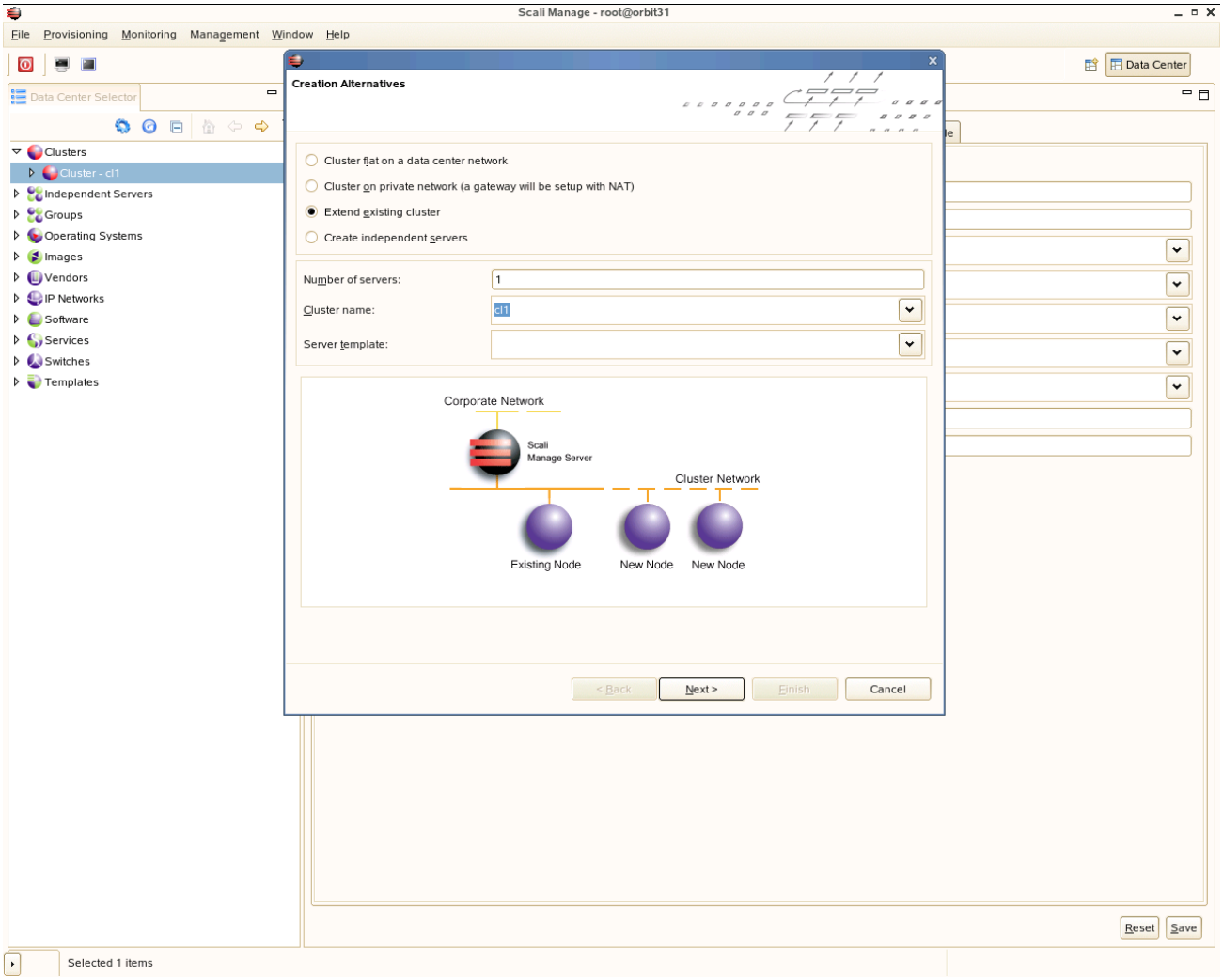

Figure 2-4 New Cluster Node Selection Example

Selecting the Server Type

Scroll down the menu and select the server type you are adding.

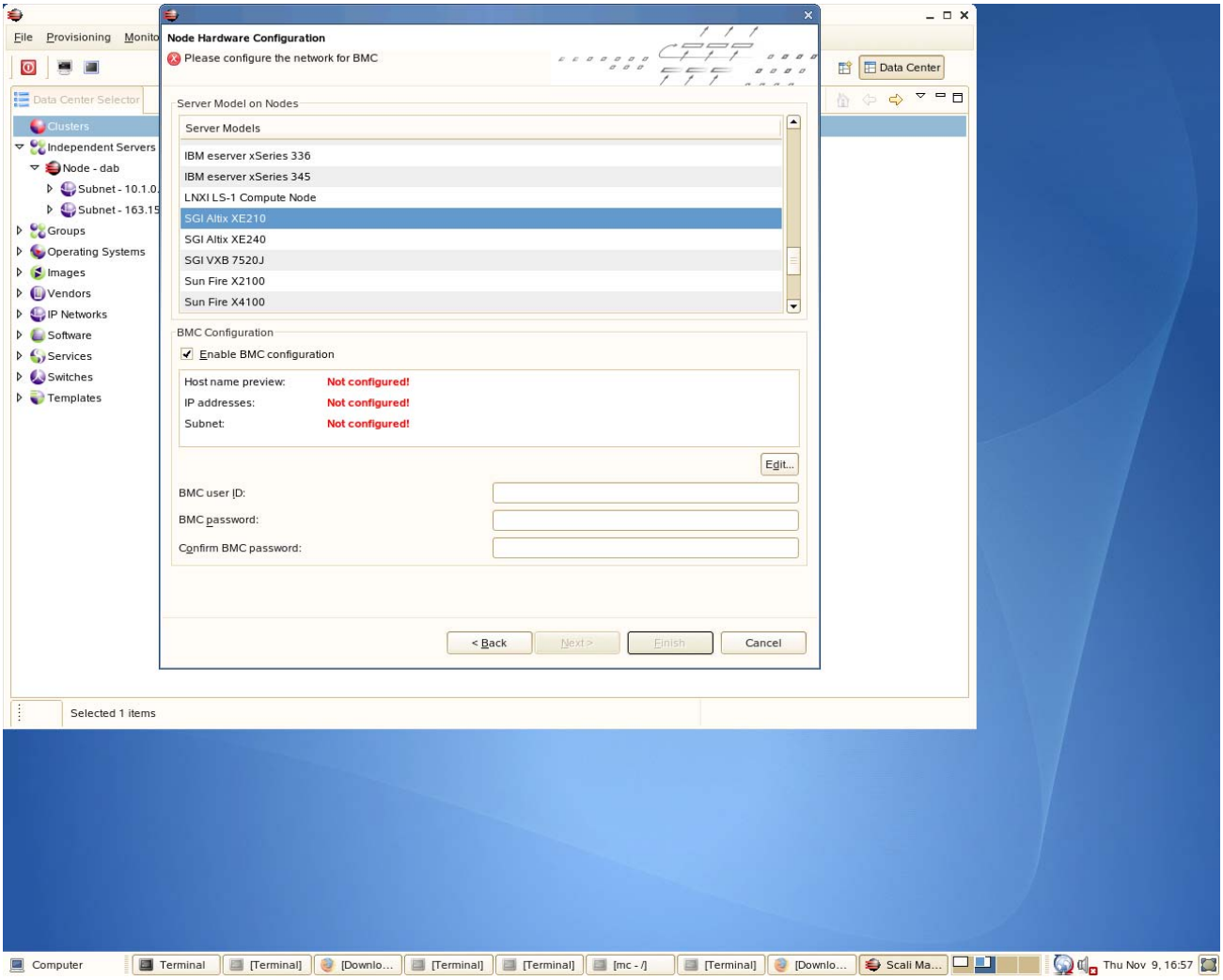

Figure 2-5 Node Server Type Selection Screen Example

Network BMC Configuration

Assign the new BMC IP address, stepping and BMC host name. Click OK when the appropriate information is entered. Then enter the BMC user id (**admin**) and the password (**admin**). Click "Next" to move to the following screen.

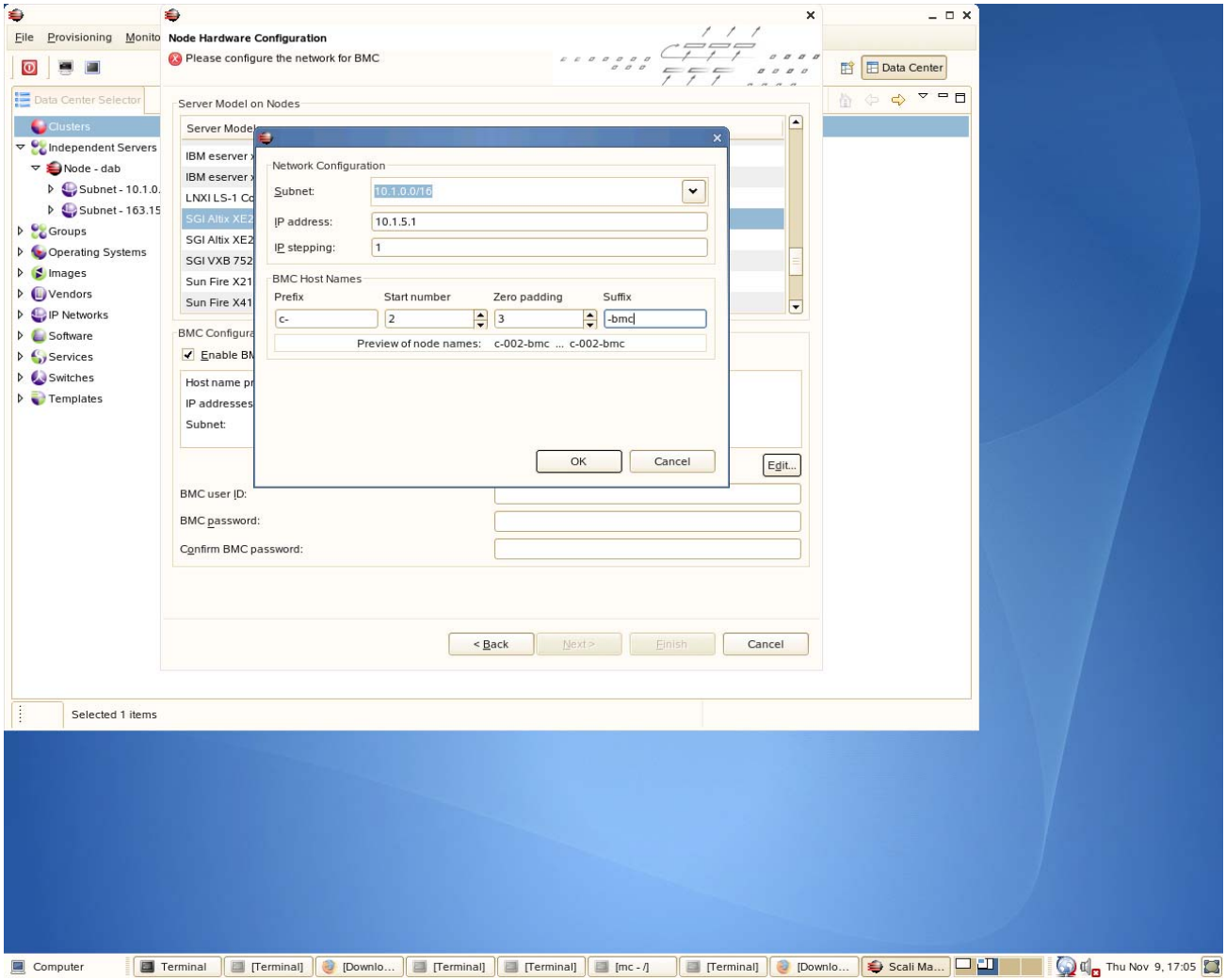

Figure 2-6 BMC Network Configuration Screen Example

Select Preferred Operating System

Click on the option to select the new node's operating system. Enter the **sgisgi** factory password or whatever new password may have been assigned. Click "Next" to move to the following screen.

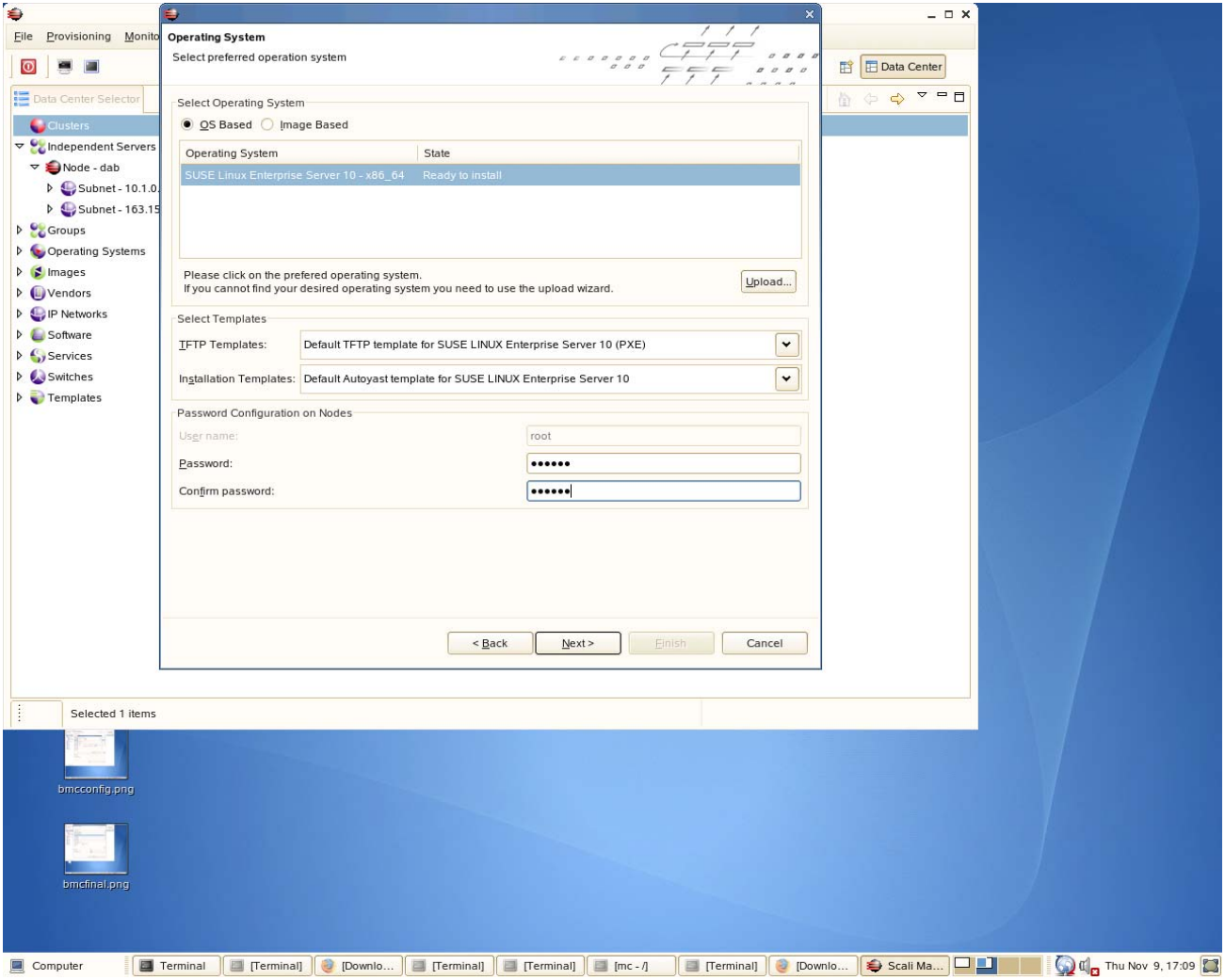

Figure 2-7 Preferred Operating System Screen Selection Example

Node Network Configuration Screen

Use this screen to assign Ethernet 0 (eth0) as your network interface port. Fill in the additional information as it applies to your local network. Click “OK” to continue.

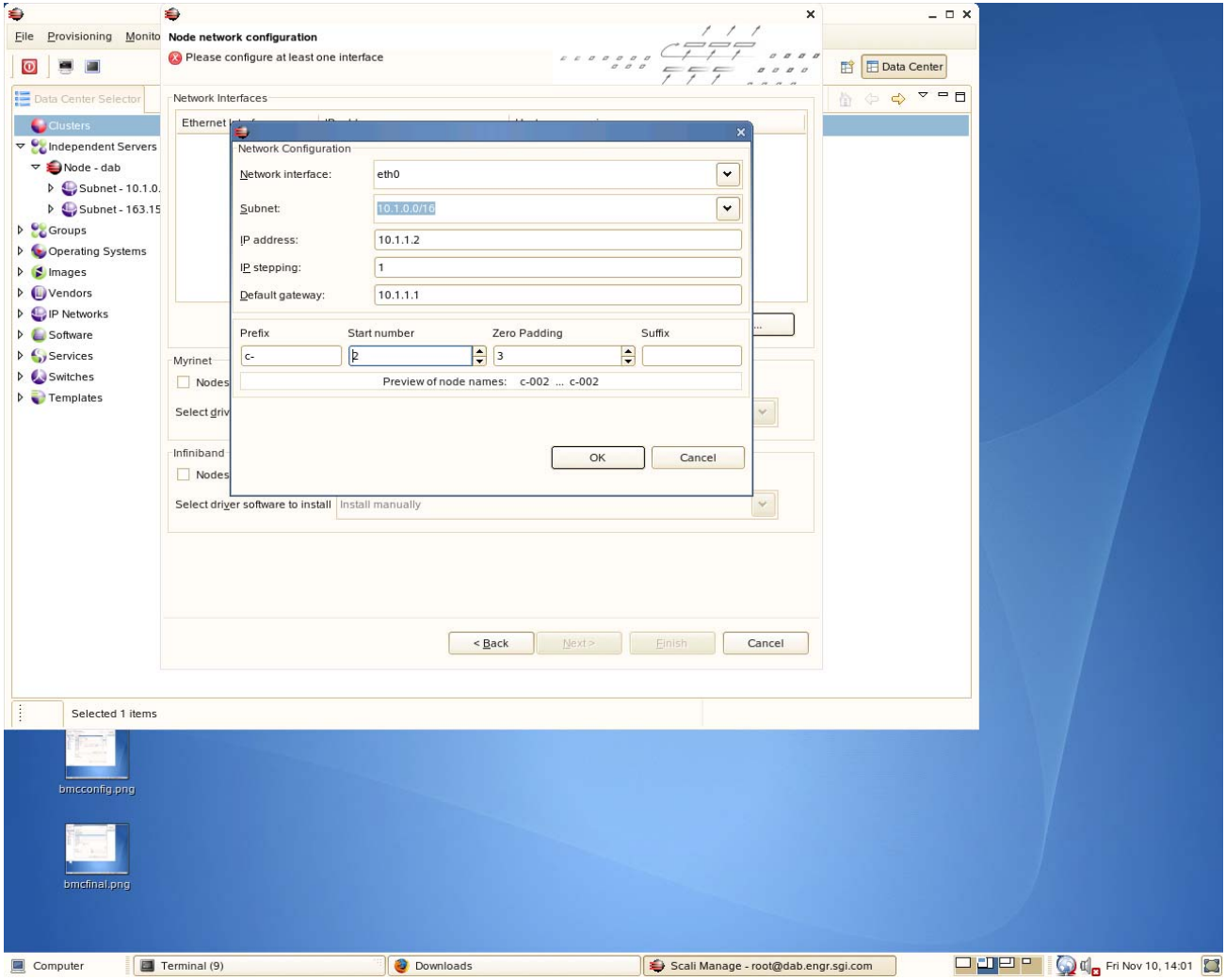

Figure 2-8 Node Network (Ethernet 0) Screen Example

DNS and NTP Configuration Screen

This screen extracts the name server numbers for use with the system configuration files. In this example, the domain name is engr.sgi.com with NTP enabled. Click “Next” when complete.

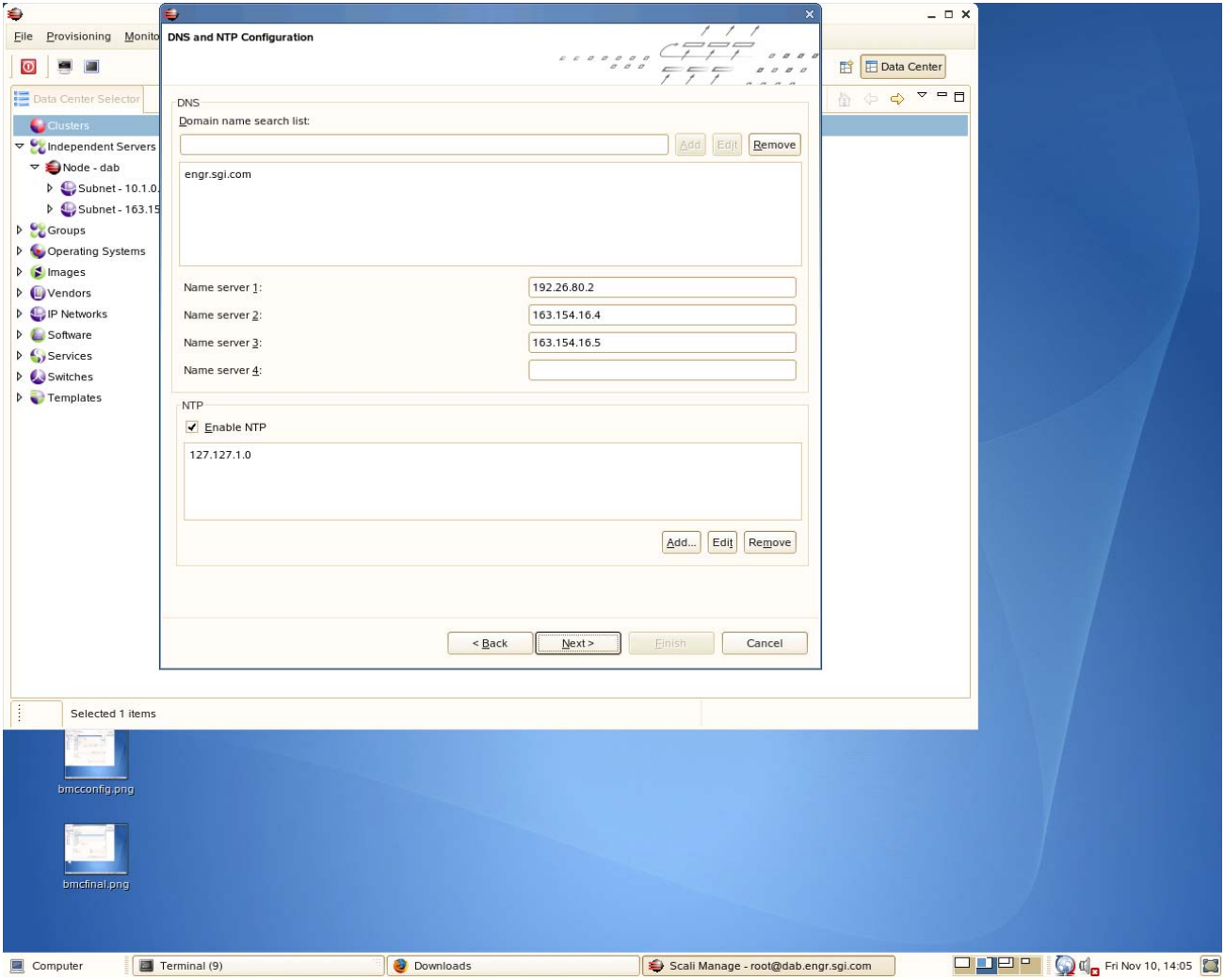

Figure 2-9 DNS and NTP Configuration Screen Example

NIS Configuration Screen

This screen allows you to specify, enable or disable a Network Information Service (NIS) for the new node. Assign your domain name and click “Next” to go to the following screen.

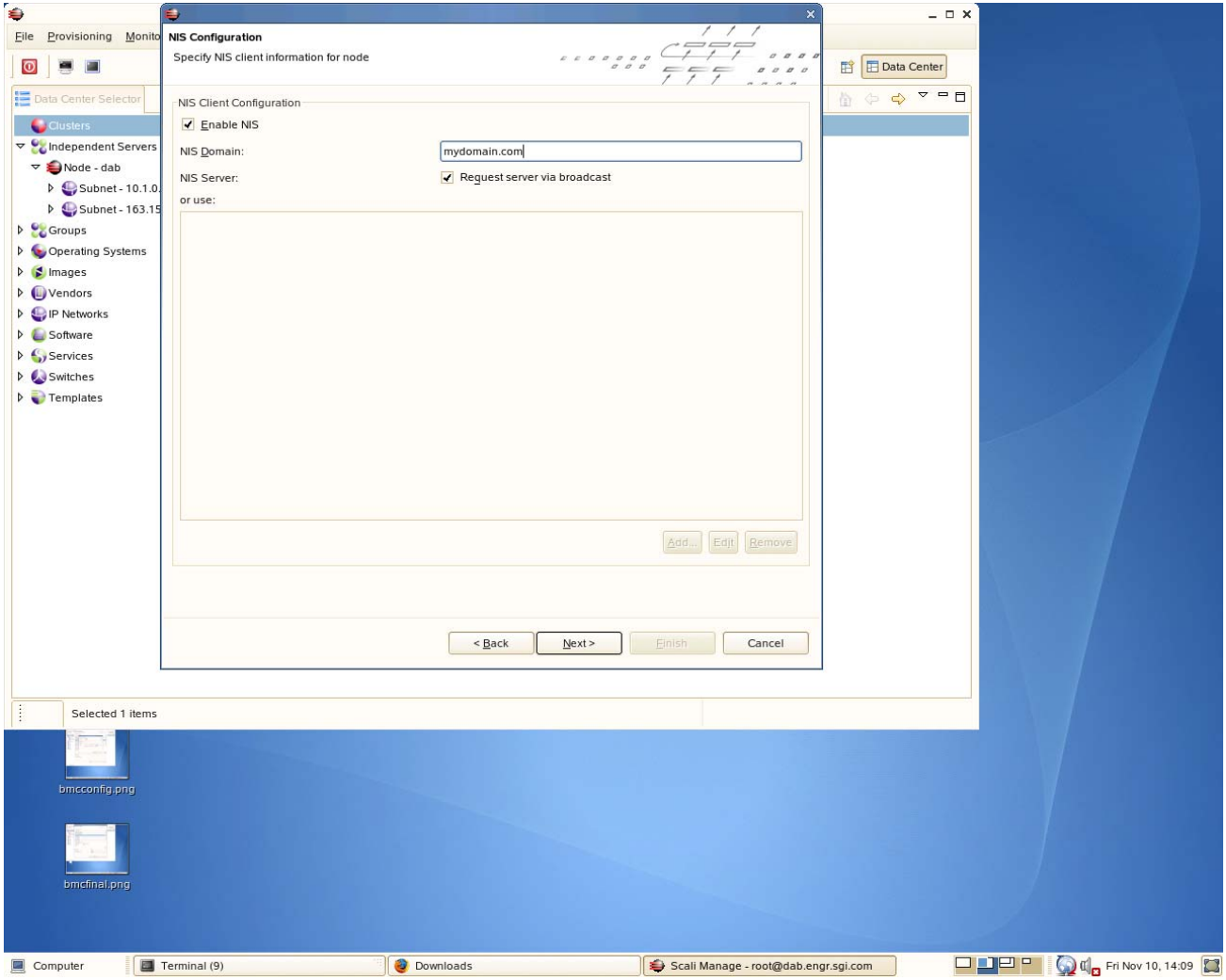

Figure 2-10 NIS Configuration Screen Example

Scali Manage Options Screen

This screen provides the options shown, including installation of MPI, your software version, monitor options and more.Click “Next” to move to the following screen.

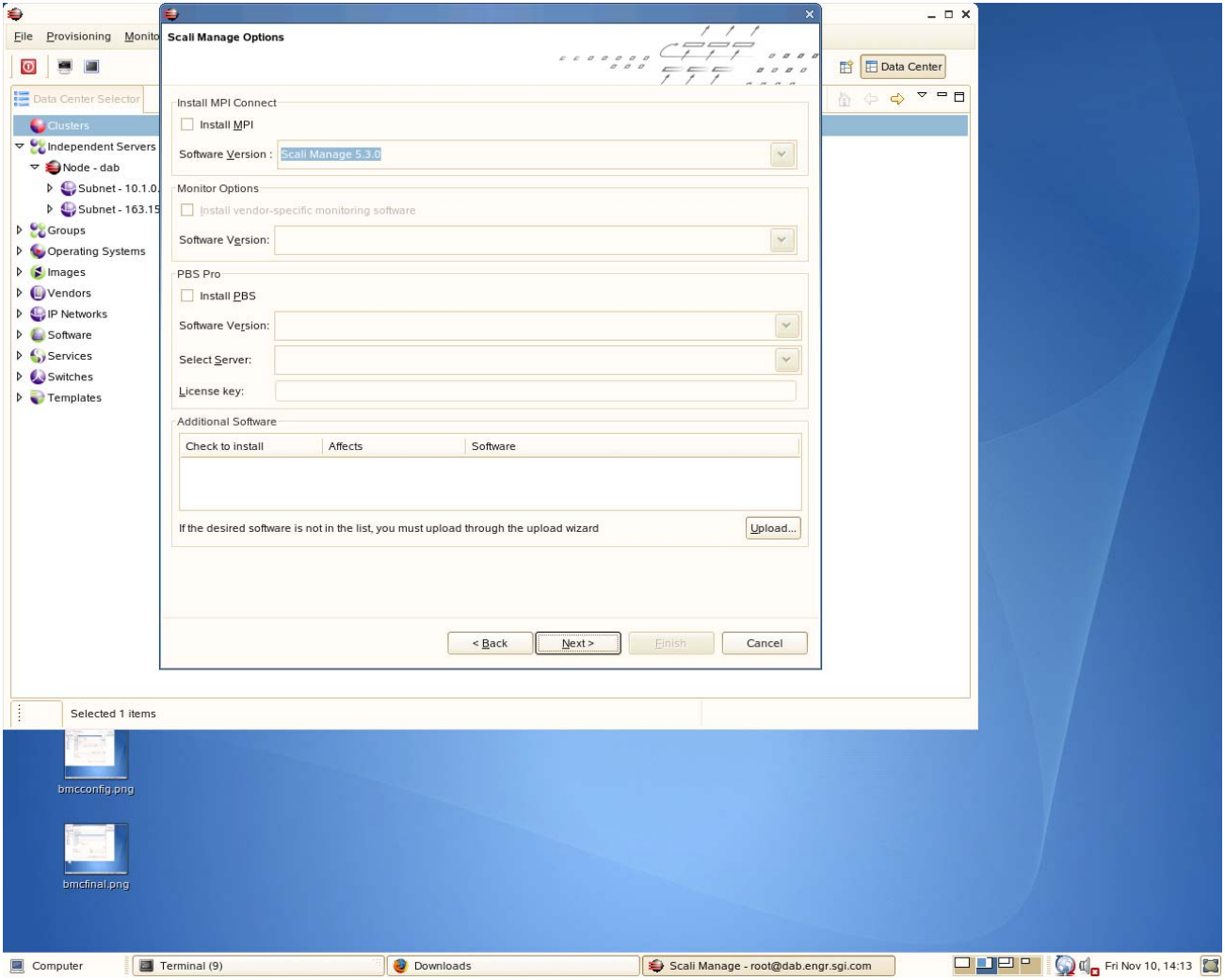

Figure 2-11 Scali Manage Options Screen Example

Configuration Setup Complete Screen

This screen allows you to install the operating system and Scali Manage immediately, or store the configuration for later use. Click “Finish” after you make your selection.

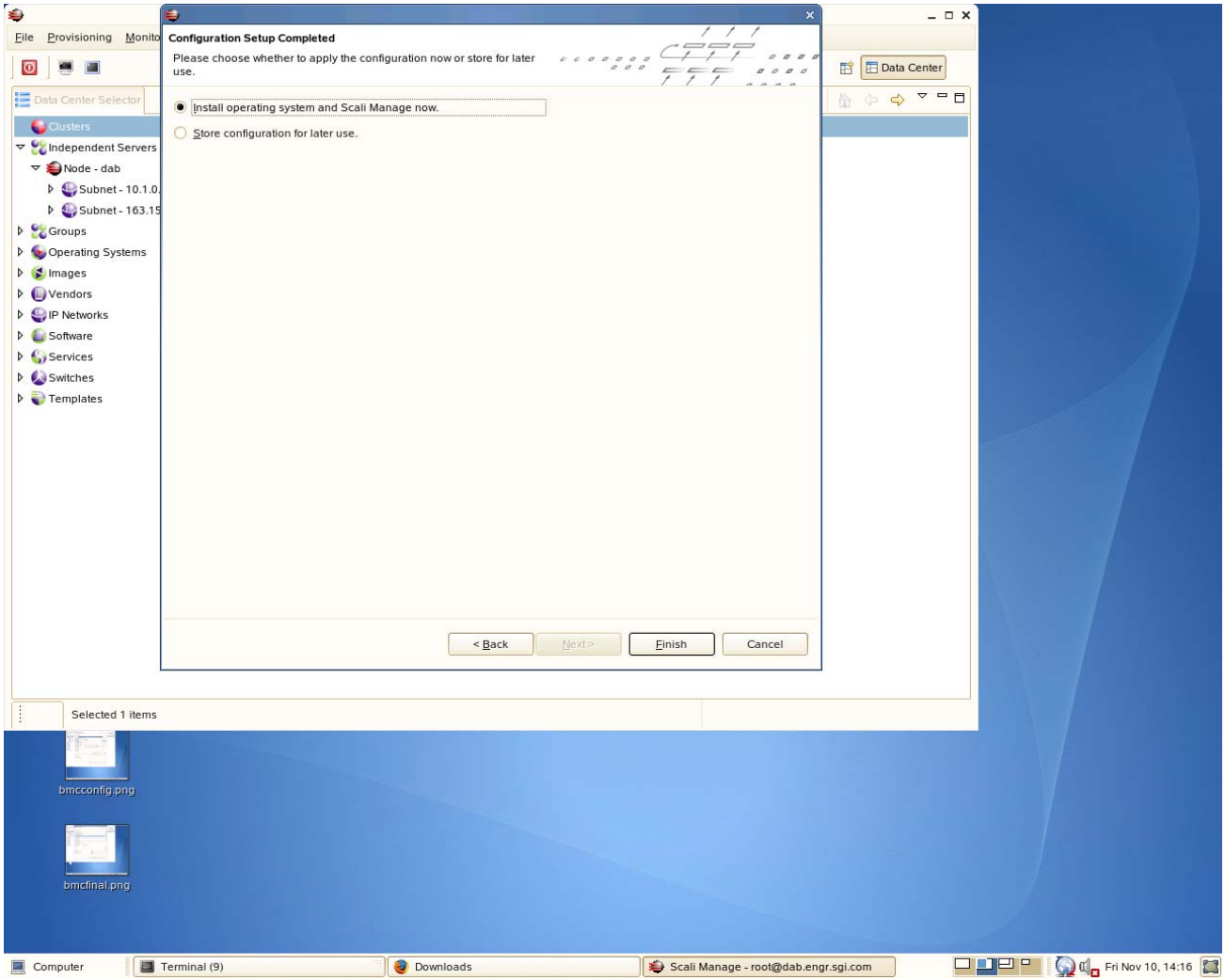

Figure 2-12 Configuration Setup Complete Screen Example

Checking the Log File Entries (Optional)

You can check the log file entries during configuration of the new node to confirm that a log file has been created and to view the entries.

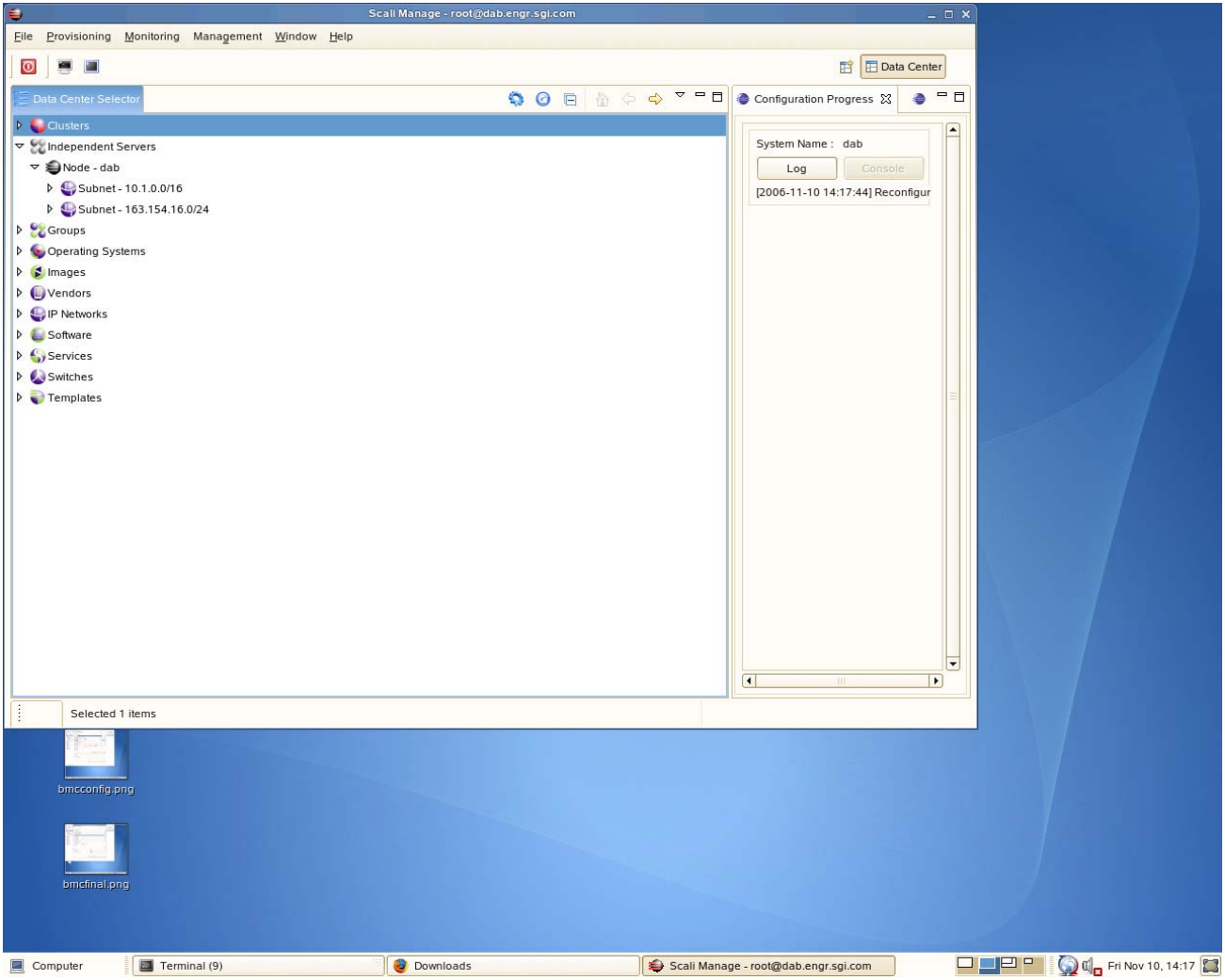

Figure 2-13 Optional Log File Screen Example

Setting a Node Failure Alarm on Scali Manage

This section shows how to create an alarm using a "Node Down" alarm as an example:

1. Start the GUI. See “Start the Scali Manage GUI” on page 25 if needed.
2. Using the mouse, select the "Edit Alarms" submenu from the "Monitoring" menu item.
3. Select a node (or list of nodes) for which you want to define the alarm, see Figure 2-14.

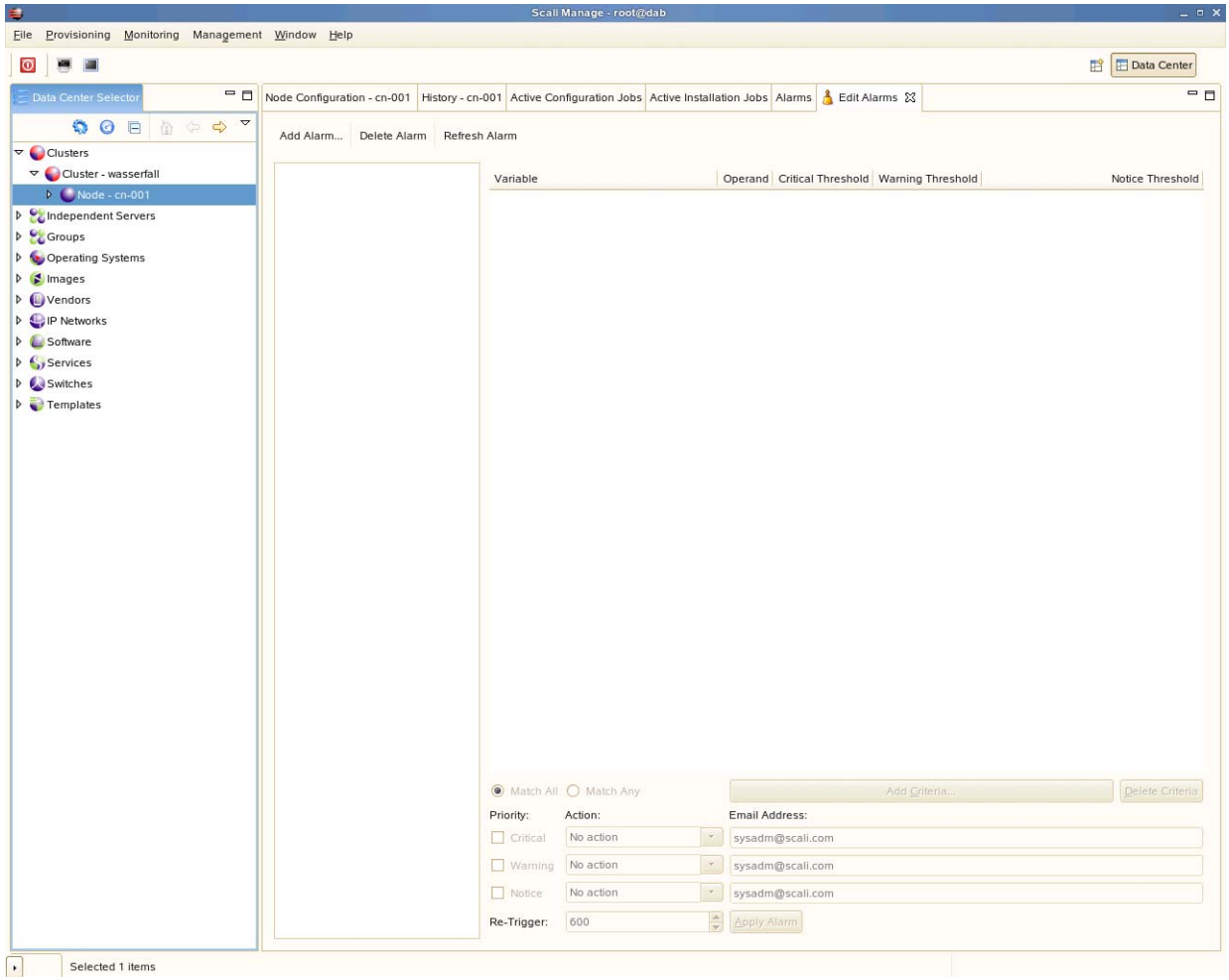

Figure 2-14 Node Selection for Alarm Function Example

4. Then select "Add Alarm" to add the alarm.
5. A popup appears offering input for the alarm name and an optional description, see Figure 2-15.

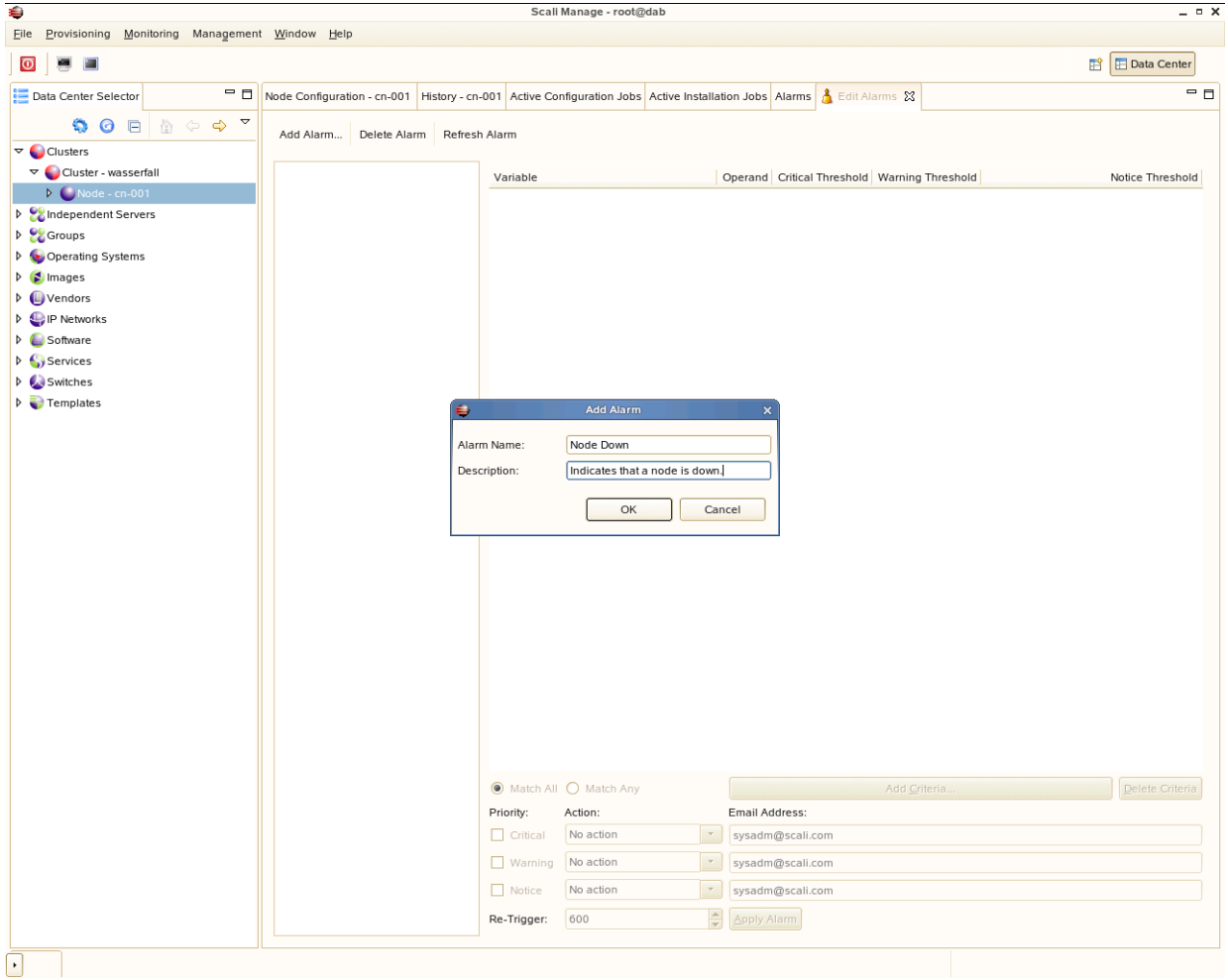

Figure 2-15 Alarm Description Popup Example

6. At this time you must enter the criteria that trigger the alarm. Click on "Add Criteria" (see Figure 2-16 on page 40.)

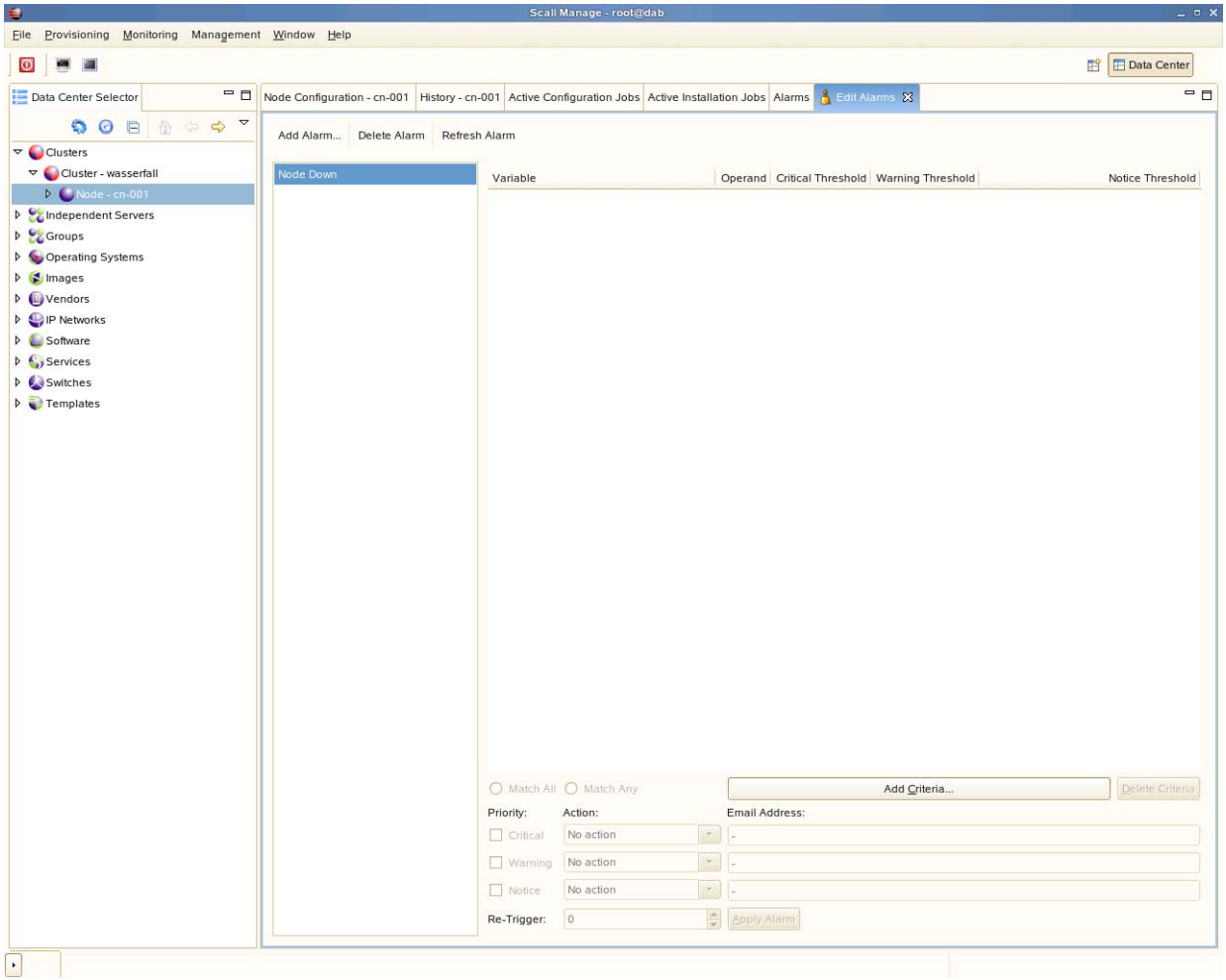

Figure 2-16 Add Criteria Screen Example

7. Another popup presents itself. For this example we picked a “Filter” criteria for the node status. See Figure 2-17 on page 41.

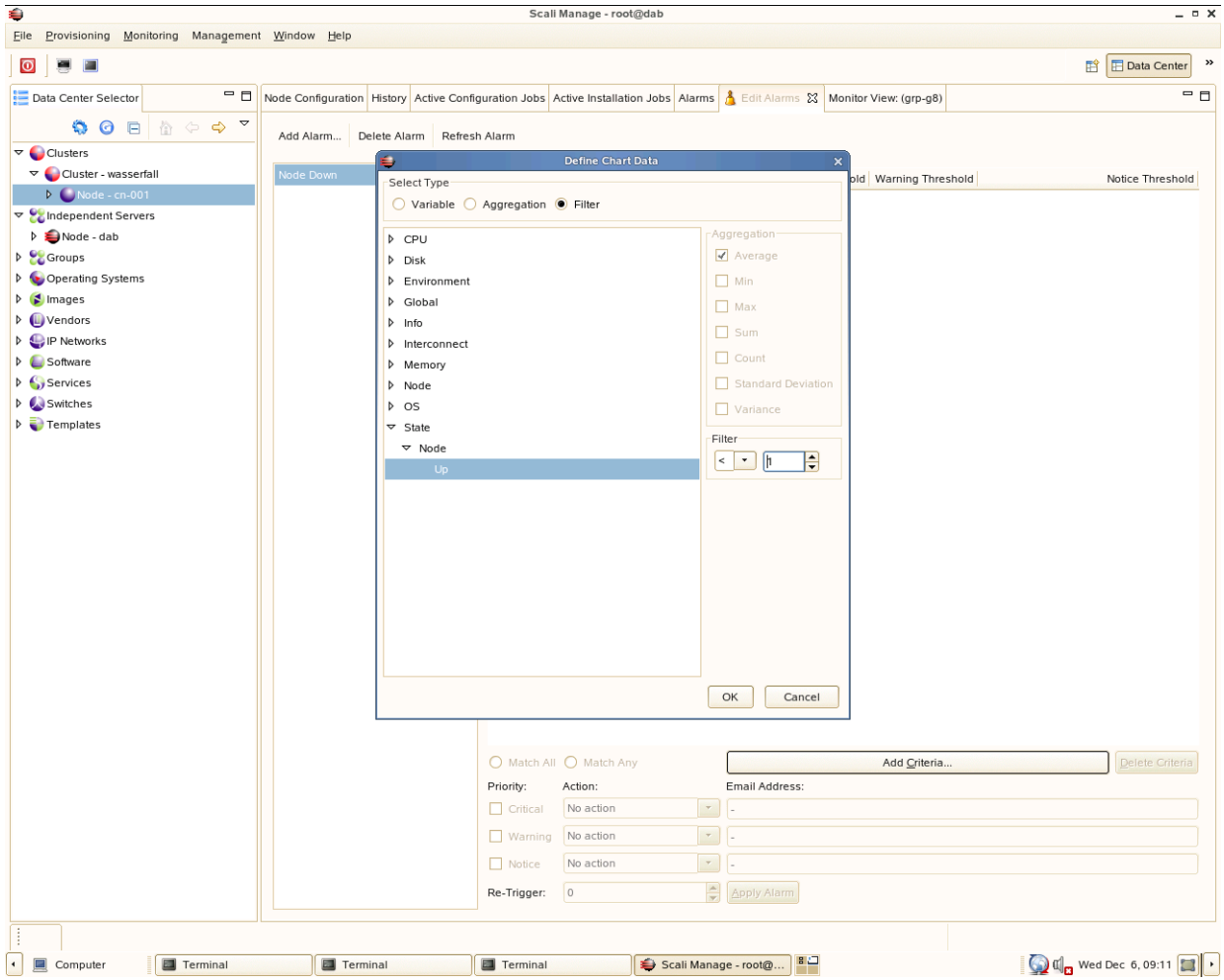

Figure 2-17 Define Chart Data Popup Example (Filter Selected)

Next we need to choose the priority for this alarm. The example assigns a critical priority for the "Node Down" alarm. We want this alarm to be triggered at most once. Therefore we leave the "Re-Trigger" value with 0. To enable this alarm, click on "Apply Alarm", see Figure 2-18 on page 42. An alternative would be to define a re-trigger interval in seconds by providing the amount of seconds for "Re-Trigger". This alarm does not define any action to be taken when the alarm fires. This can be easily done by selecting a predefined action. As an example, Scali can send an

e-mail to a system administrator or email alias. You must pick the appropriate action and supply the e-mail address.

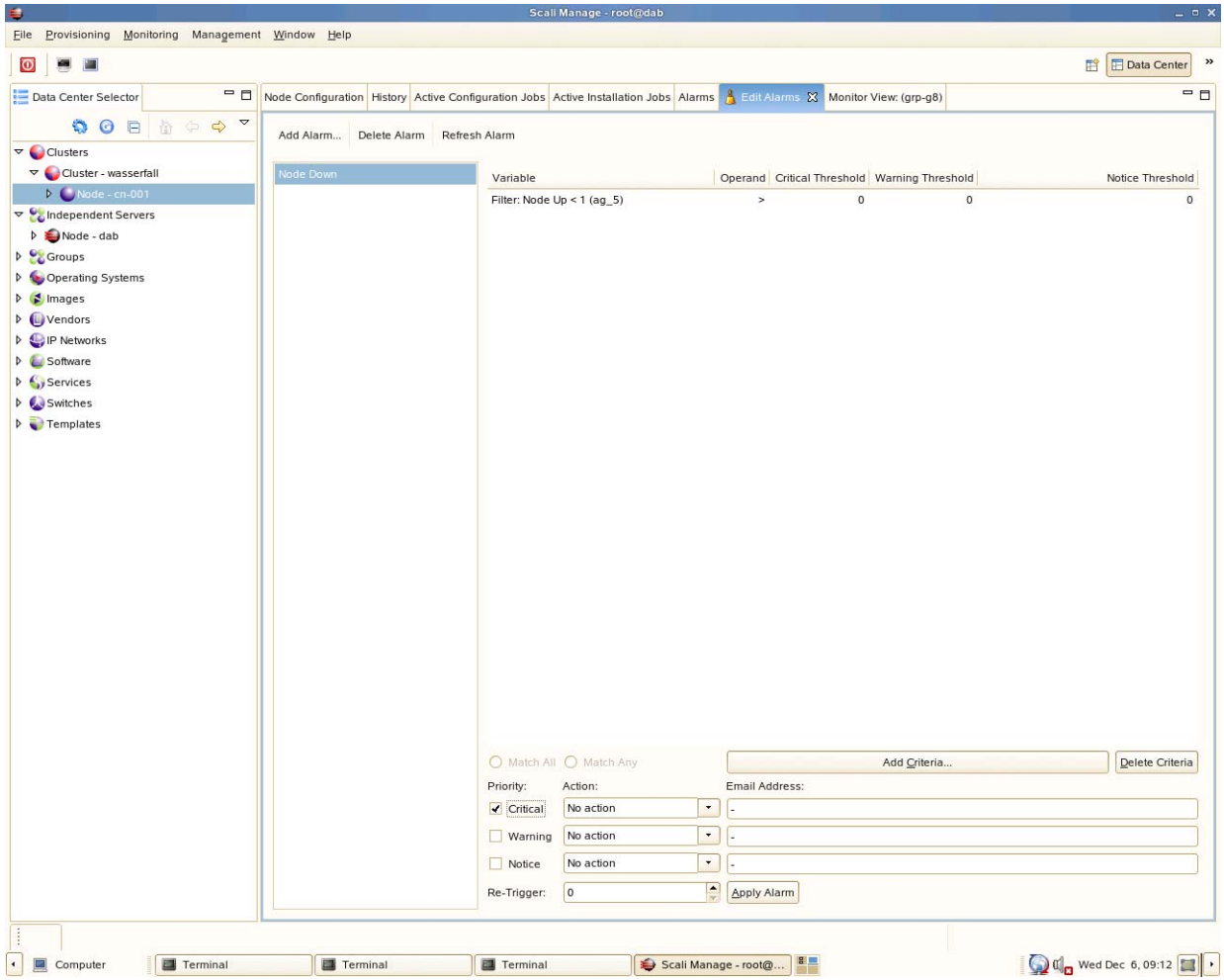

Figure 2-18 Applying the Alarm Example Screen

To illustrate how an alarm makes it's appearance we have purposefully brought down the node. A few seconds thereafter the GUI indicates a node failure by changing the node icon in the cluster tree, see Figure 2-19 on page 43. A few seconds later the alarm gets triggered and shows up in the alarm log, see Figure 2-20 on page 44.

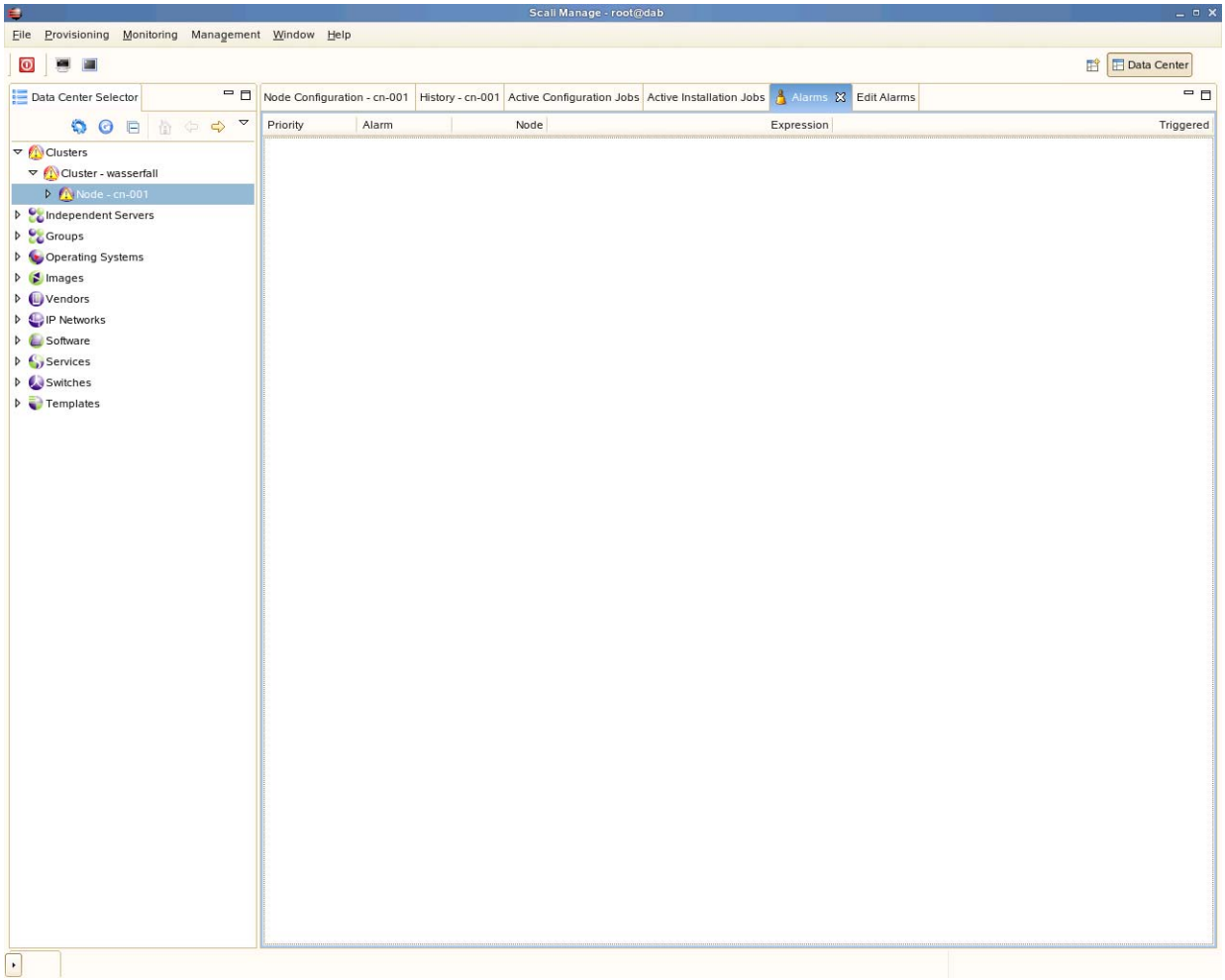

Figure 2-19 Node Failure Icon Example Screen

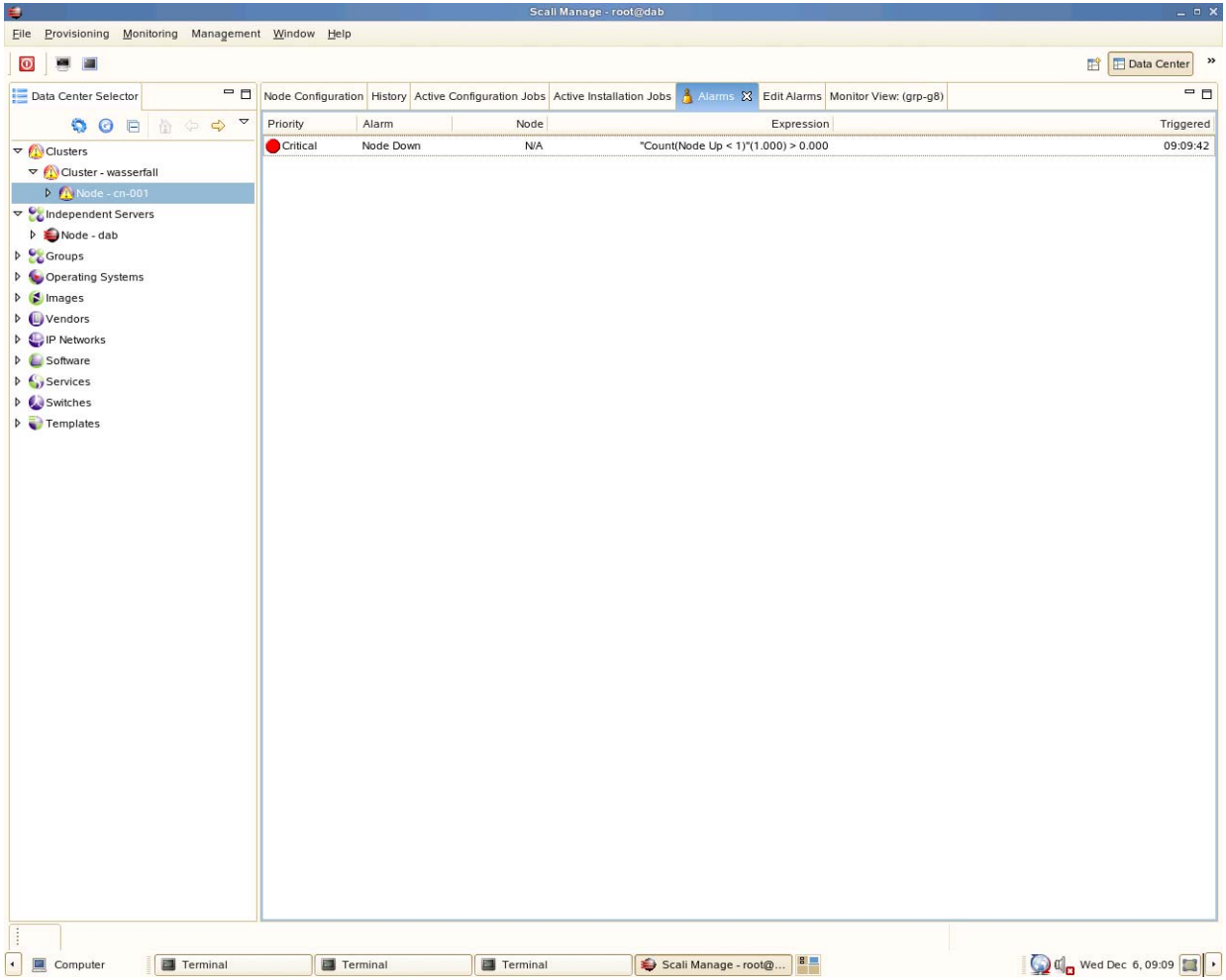

Figure 2-20 Node Down Alarm Screen Example

IPMI Commands Overview

This chapter provides a set of example IPMI commands, and is not meant to be a comprehensive guide in the use of ipmitool. Its purpose is to briefly describe some of the commonly used IPMI commands to help you get started with your cluster administration.

ipmitool

Command-line utility for issuing common IPMI requests allows remote operation usage:

```
ipmitool [-v] [-I intf] [-o eomtype] [-H host] [-k key] [-U user] [-P password] [-E]
                    command...
                    -v : Verbosity, can be specified multiple times -vv
                    -I intf : IPMI interface to use
                    -o oemtype : Select OEM type to support.
```

This usually involves minor hacks in place in the code to work around quirks in various BMCs from various manufacturers. Use -o list to see a list of current supported OEM types.

```
open – OpenIPMI driver (default)
lan – LAN connection (remote connection, requires -H/-U/-P arguments)
lanplus – LANplus connection (IPMI 2.0) Requires H/U/P arguments be
supplied

-H host : Hostname of remote system (-I lan only)

-k key : KG Key (System password) (-I lanplus only)

-U user : Username on remote system (-I lan only)

-P pass : Password for user on remote system (-I lan only)

-E : Read password from IPMI_PASSWORD environment variable
```

If -E and -P are not specified on a remote connection, the utility prompts for a password.

Ipmitool – User administration

BMC Supports multiple users, username/password is required for remote connections. The cluster is shipped with a factory username and password set on user id 2:
Username = **admin**
Password = **admin**

Typical ipmitool command line

```
Ipmitool –I lanplus –o intelplus –H [bmc IP] –U admin –P admin
<command>
```

<opts> references in this document refer to the following command line arguments:

```
–I lanplus –o intelplus –H [bmc ip] –U admin –P admin
```

Adding a user to the BMC

```
ipmitool <opts> user set name <user id>  <username>
Ipmitool <opts> user set password <user id>  <password>
ipmitool <opts> user enable <user id>
```

Ipmitool - Configuring a NIC

Display a current LAN configuration

```
ipmitool <opts> lan print 1
```

Configure a static IP Address

Static IP addresses are already set in the factory on LAN channel 1 of each node. See Table 1-1 on page 3 and Table 1-2 on page 10 for the BMC static IP assignments.

The following commands show how to reconfigure the BMC static IP's. The "1" in the following examples indicate "channel 1" onboard nic1 controller.

```
ipmitool <opts> lan set 1 ipsrc static
ipmitool <opts> lan set 1 ipaddr x.x.x.x
ipmitool <opts> lan set 1 netmask x.x.x.x
impitool <opts> lan set 1 arp respond on
impitool <opts> lan set 1 arp generate on
```

To check your lan settings:

```
impitool <opts>  lan print 1
```

Ipmitool – SOL (serial-over-lan) commands

Serial-Over-Lan comes preconfigured and enabled on each node of your cluster.

Configuring SOL

Here are a few commands that may be useful if reconfiguring is required. The following are SGI recommended:

```
impitool <opts> sol set character-send-threshold 50 1
impitool <opts> sol set character-accumulate-level 004 1
impitool <opts> sol set retry-interval 20 1
impitool <opts> sol set retry-count 6 1
impitool <opts> sol set non-volatile-bit-rate 38.4 1
```

Note: Some systems were set to a 115.2 baud rate. To see your configuration, enter the following:
`impitool <opts> sol info`

Connecting to node console via SOL

```
ipmitool <opts> sol activate
```

Deactivating an SOL connection

In certain cases using the Scali Manage GUI to access a console, you may need to deactivate the SOL connection from the command line to free up the SOL session.

```
ipmitool <opt> sol deactivate
```

Ipmitool – Sensor commands

Displaying all objects in SDR

```
ipmitool <opts> sdr list
Ipmitool <opts> sdr dump <filename> (Dump SDR contents to a file)
```

Displaying all sensors in the system

```
ipmitool <opts> sensor list
```

Displaying an individual sensor

```
   ipmitool <opts> sensor get “Temp”
   Changing sensor threshold
ipmitool <opts> sensor thresh “Temp” ucr 100
```

Thresholds are: unr, ucr, unc, lnc, lcr, lnr.

Ipmitool – Chassis commands

Chassis Identify

```
ipmitool <opts> chassis identify (defaults to 15 seconds)
ipmitool <opts> chassis identify off
```

Controlling System Power

```
ipmitool <opts> chassis power status
ipmitool <opts> chassis power off
ipmitool <opts> chassis power on
ipmitool <opts> chassis power cycle
ipmitool <opts> chassis power soft (Performs safe OS shutdown)
```

Changing System Boot Order

```
ipmitool <opts> chassis bootdev pxe
ipmitool <opts> chassis bootdev harddisk
ipmitool <opts> chassis bootdev cdrom
```

Ipmitool – SEL Commands

Retrieving information about SEL:
`ipmitool <opts> sel info`

Displays date/time of last event, last log clear time, and number of entries.

Displaying SEL:
`ipmitool <opts> sel list`

Clearing SEL:
`ipmitool <opts> sel clear`