


FUJITSU Software PRIMECLUSTER

A horizontal band featuring a red abstract graphic with flowing, curved lines and bright light flares, creating a sense of motion and energy.

Cluster Foundation Configuration and Administration Guide 4.3

Linux

J2UZ-5284-07ENZ0(03)
May 2020

Preface

The Cluster Foundation (CF) provides a comprehensive base of services that user applications and other PRIMECLUSTER services need to administrate and communicate in a cluster.

Target Readers

This manual is intended for all users who use PRIMECLUSTER 4.3 and perform cluster system installation and operation management.

Configuration of This Documentation

This manual is organized as follows:

Chapter	Description
Chapter 1 Cluster Foundation	Describes the administration and configuration of the Cluster Foundation.
Chapter 2 CF Registry and Integrity Monitor	Discusses purpose and physical characteristics of the CF synchronized registry, and it discusses the purpose and implementation of the Cluster Integrity Monitor (CIM).
Chapter 3 Cluster resource management	Discusses the database which is a synchronized clusterwide database holding information specific to several PRIMECLUSTER products.
Chapter 4 GUI administration	Describes the administration features in the CF portion of the Cluster Admin graphical user interface (GUI).
Chapter 5 LEFTCLUSTER state	Discusses the LEFTCLUSTER state, relation between LEFTCLUSTER state and other states. How LEFTCLUSTER state is caused in various ways is also described.
Chapter 6 CF topology table	Discusses the layout and the use of the CF topology table.
Chapter 7 Shutdown Facility (SF)	Describes the components and advantages of PRIMECLUSTER SF and provides administration information.
Chapter 8 Diagnostics and troubleshooting	Provides help for troubleshooting and problem resolution for PRIMECLUSTER Cluster Foundation.
Chapter 9 Manual pages	Lists the manual pages for PRIMECLUSTER.
Appendix A Release information	Describes primary changes in this manual.

Related Documentation

The documents listed below provide details about PRIMECLUSTER products.

- PRIMECLUSTER Concepts Guide
- PRIMECLUSTER Installation and Administration Guide
- PRIMECLUSTER Reliant Monitor Services (RMS) with Wizard Tools Configuration and Administration Guide
- PRIMECLUSTER Web-Based Admin View Operation Guide
- PRIMECLUSTER Global Disk Services Configuration and Administration Guide
- PRIMECLUSTER Global File Services Configuration and Administration Guide
- PRIMECLUSTER Global Link Services Configuration and Administration Guide: Redundant Line Control Function
- PRIMECLUSTER Messages



Note

The PRIMECLUSTER documentation includes the following documentation in addition to those listed above:

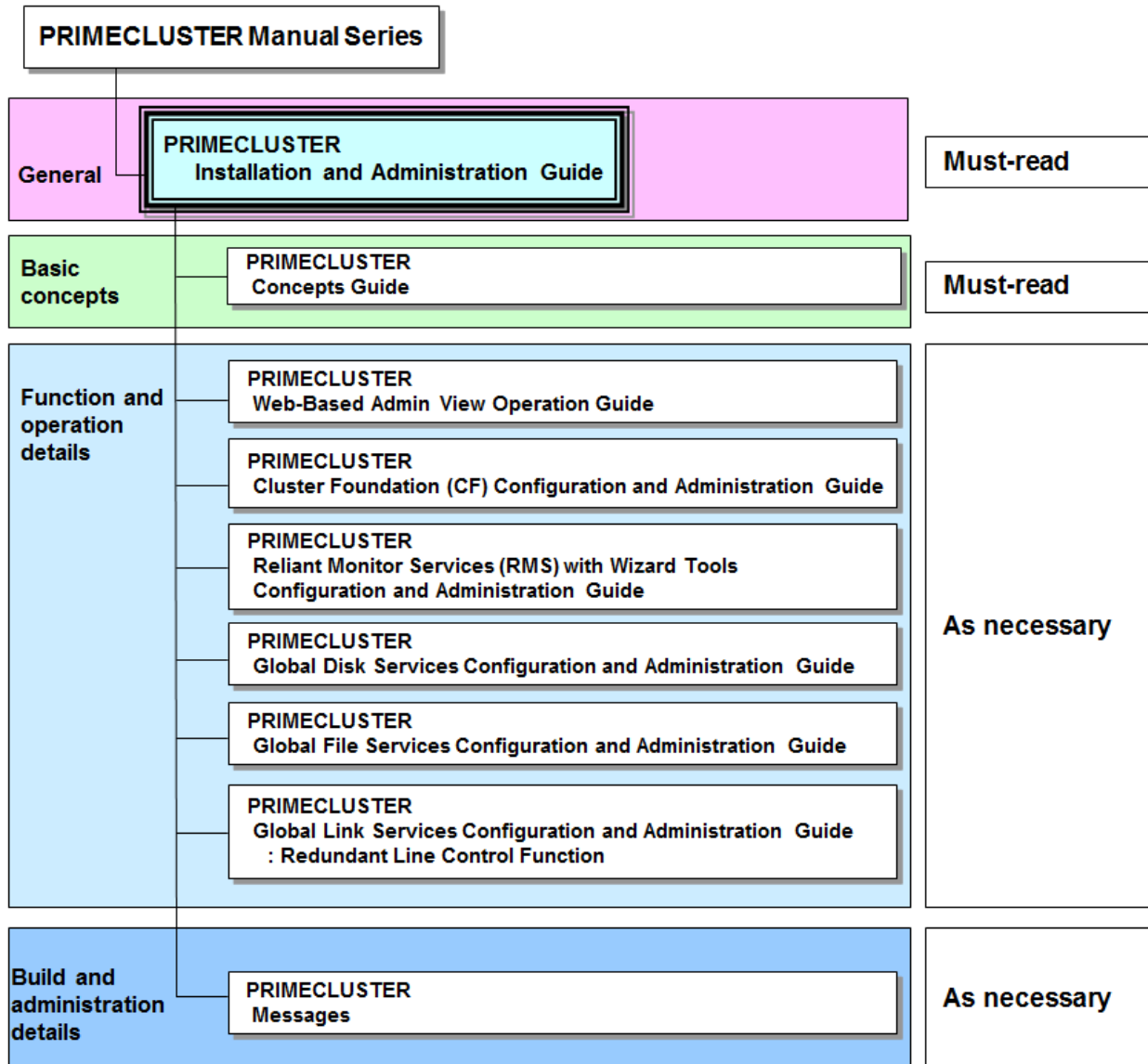
- PRIMECLUSTER Software Release Guide and Installation Guide

This Software Release Guide and Installation Guide are provided with each PRIMECLUSTER product package.

The data is stored on "DVD" of each package. For details on the file names, see the documentation.

.....

Manual Series



Manual Printing

If you want to print a manual, use the PDF file found on the DVD for the PRIMECLUSTER product. The correspondences between the PDF file names and manuals are described in the Software Release Guide for PRIMECLUSTER that comes with the product.

Adobe Reader is required to read and print this PDF file. To get Adobe Reader, see Adobe Systems Incorporated's website.

Online Manuals

To allow users to view the online manuals, use the Cluster management server to register each user name to one of the user groups (wvroot, clroot, cladmin, or clmon).

For information on user group registration procedures and user group definitions, see "4.3.1 Assigning Users to Manage the Cluster" in "PRIMECLUSTER Installation and Administration Guide."

Conventions

Notation

Prompts

Command line examples that require system administrator (or root) privileges to execute are preceded by the system administrator prompt, the hash sign (#). Entries that do not require system administrator rights are preceded by a dollar sign (\$).

In some examples, the notation `node#` indicates a root prompt on the specified node. For example, a command preceded by `fuji3#` would mean that the command was run as user root on the node named fuji3.

The keyboard

Keystrokes that represent nonprintable characters are displayed as key icons such as [Enter] or [F1]. For example, [Enter] means press the key labeled Enter; [Ctrl-b] means hold down the key labeled Ctrl or Control and then press the [B] key.

Typefaces

The following typefaces highlight specific elements in this manual.

Typeface / Symbol	Usage
Constant Width	Computer output and program listings; commands, file names, manual page names and other literal programming elements in the main body of text.
<i>Italic</i>	Variables in a command line that you must replace with an actual value. May be enclosed in angle brackets to emphasize the difference from adjacent text; for example, <code><nodename>RMS</code> ; unless directed otherwise, you should not enter the angle brackets. The name of an item in a character-based or graphical user interface. This may refer to a menu item, a radio button, a checkbox, a text input box, a panel, or a window title.
Bold	Items in a command line that you must type exactly as shown.

Example 1

Several entries from an `/etc/passwd` file are shown below:

```
sysadm:x:0:0:System Admin.:/usr/admin:/usr/sbin/sysadm
setup:x:0:0:System Setup:/usr/admin:/usr/sbin/setup
daemon:x:1:1:0000-Admin(0000)::
bin:x:1:1:bin:/bin:/bin/bash
daemon:x:2:2:daemon:/sbin:/bin/bash
lp:x:4:7:lp daemon:/var/spool/lpd:/bin/bash
```

Example 2

To use the `cat` command to display the contents of a file, enter the following command line:

```
$ cat file
```

Command syntax

The command syntax observes the following conventions.

Symbol	Name	Meaning
[]	Brackets	Enclose an optional item.
{ }	Braces	Enclose two or more items of which only one is used. The items are separated from each other by a vertical bar ().
	Vertical bar	When enclosed in braces, it separates items of which only one is used. When not enclosed in braces, it is a literal element indicating that the output of one program is piped to the input of another.

Symbol	Name	Meaning
()	Parentheses	Enclose items that must be grouped together when repeated.
...	Ellipsis	Signifies an item that may be repeated. If a group of items can be repeated, the group is enclosed in parentheses.

Notation symbols

Material of particular interest is preceded by the following symbols in this manual:



Point

Contains important information about the subject at hand.



Note

Describes an item to be noted.



Example

Describes operation using an example.



Information

Describes reference information.



See

Provides the names of manuals to be referenced.

Abbreviations

Oracle Solaris might be described as Solaris, Solaris Operating System, or Solaris OS.

PRIMEQUEST 2000/1000 Series are abbreviated as PRIMEQUEST.

Export Controls

Exportation/release of this document may require necessary procedures in accordance with the regulations of your resident country and/or US export control laws.

Trademarks

Red Hat is a trademark of Red Hat, Inc. in the United States and other countries.

Linux is a trademark or registered trademark of Mr. Linus Torvalds in the United States and other countries.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

VMware is a registered trademark or trademark of VMware, Inc. in the United States and/or other jurisdictions.

All other hardware and software names used are trademarks of their respective companies.

Requests

- No part of this documentation may be reproduced or copied without permission of FUJITSU LIMITED.
- The contents of this documentation may be revised without prior notice.

Date of publication and edition

September 2015, Seventh edition
August 2016, 7.1 edition
January 2018, 7.2 edition
May 2020, 7.3 edition

Copyright notice

All Rights Reserved, Copyright (C) FUJITSU LIMITED 2008-2020.

Revision History

Revision	Location	Edition
Added the description for the configuration of sharing NIC with administrative LAN and cluster interconnect in the VMware environment.	1.1 CF, CIP, and CIM configuration	7.1
Added the corrective action in case of failure with rcqconfig command.	1.1.7 Example of CF configuration by CLI	
Edited the description for Shutdown Daemon (SD).	7.1 Overview	
Added "Checking SA status."	7.5.2 Checking SA status	
Changed the description of the start CF services status window.	4.6.1 Starting CF	7.2
Changed the description when stopping the CF service.	4.6.2 Stopping CF	
Added the note when displaying logs from Cluster Admin.	4.8 Using PRIMECLUSTER log viewer	7.3

Contents

Chapter 1 Cluster Foundation.....	1
1.1 CF, CIP, and CIM configuration.....	1
1.1.1 Differences between CIP and CF over IP.....	3
1.1.2 cfset.....	5
1.1.3 CF security.....	7
1.1.3.1 cfcf/cfsh.....	7
1.1.3.2 sshconf.....	7
1.1.4 Signed applets.....	8
1.1.5 Example of creating a cluster.....	8
1.1.6 Adding a new node to CF.....	26
1.1.7 Example of CF configuration by CLI.....	27
1.2 CIP configuration file.....	28
Chapter 2 CF Registry and Integrity Monitor.....	30
2.1 CF Registry (CFREG).....	30
2.2 Cluster Integrity Monitor(CIM).....	30
2.2.1 Configuring CIM.....	30
2.2.2 Query of the quorum state.....	31
2.2.3 Reconfiguring quorum.....	31
Chapter 3 Cluster resource management.....	33
3.1 Resource Database configuration.....	33
3.2 Startup synchronization.....	34
3.2.1 Startup synchronization and the new node.....	35
Chapter 4 GUI administration.....	36
4.1 Starting Cluster Admin GUI and logging in.....	36
4.2 Main CF table.....	38
4.3 CF route tracking.....	39
4.4 Node details.....	41
4.5 Displaying the topology table.....	42
4.6 Starting and stopping CF.....	45
4.6.1 Starting CF.....	47
4.6.2 Stopping CF.....	48
4.7 Marking nodes DOWN.....	50
4.8 Using PRIMECLUSTER log viewer.....	50
4.8.1 Search based on time filter.....	51
4.8.2 Search based on keyword.....	51
4.8.3 Search based on severity levels.....	51
4.9 Displaying statistics.....	51
4.10 Heartbeat monitor.....	53
4.11 Adding and removing a node from CIM.....	54
4.12 Unconfigure CF.....	55
4.13 CIM Override.....	56
Chapter 5 LEFTCLUSTER state.....	58
5.1 Description of the LEFTCLUSTER state.....	58
5.2 Recovering from LEFTCLUSTER.....	59
5.2.1 Caused by a panic/hung node.....	60
5.2.2 Caused by staying in the kernel debugger too long.....	60
5.2.3 Caused by a cluster partition.....	60
Chapter 6 CF topology table.....	62
6.1 Basic layout.....	62
6.2 Selecting devices.....	63
6.3 Examples of topology table.....	63

Chapter 7 Shutdown Facility.....	66
7.1 Overview.....	66
7.2 Available SAs.....	67
7.2.1 Blade.....	67
7.2.2 IPMI.....	68
7.2.3 kdump.....	68
7.2.4 MMB.....	68
7.2.5 ICMP.....	68
7.2.6 VMCHKHOST.....	68
7.2.7 libvirt.....	69
7.3 SF split-brain handling.....	69
7.3.1 Administrative LAN.....	69
7.3.2 SF split-brain handling.....	69
7.3.2.1 RMS ShutdownPriority attribute.....	69
7.3.2.2 Shutdown Facility weight assignment.....	69
7.3.3 Runtime processing.....	70
7.3.4 Configuration notes.....	70
7.4 Configuring the Shutdown Facility.....	71
7.4.1 Shutdown Daemon.....	71
7.4.2 Shutdown Agents.....	72
7.5 SF administration.....	76
7.5.1 Starting and stopping SF.....	77
7.5.1.1 Starting and stopping SF manually.....	77
7.5.1.2 Starting and stopping SF automatically (For Red Hat Enterprise Linux 6).....	77
7.5.1.3 Starting and stopping SF automatically (For Red Hat Enterprise Linux 7).....	77
7.5.2 Checking SA status.....	77
7.6 Debugging.....	79
Chapter 8 Diagnostics and troubleshooting.....	80
8.1 Beginning the process.....	80
8.2 Symptoms and solutions.....	82
8.2.1 Join-related problems.....	82
8.2.1.1 Identifying join-related problems.....	82
8.2.1.2 Solving join-related problems.....	83
8.3 Collecting Troubleshooting Information.....	86
Chapter 9 Manual pages.....	87
9.1 CF.....	87
9.2 CIP.....	87
9.3 PAS.....	88
9.4 Cluster Resource Management Facility.....	88
9.5 RMS.....	89
9.6 Shutdown Facility (SF).....	89
9.7 Web-Based Admin View.....	90
9.8 RMS Wizards.....	90
9.9 Monitoring Agent (MA).....	91
Appendix A Release information.....	92
Glossary.....	95
Index.....	107

Chapter 1 Cluster Foundation

This chapter describes the administration and configuration of the Cluster Foundation (CF).

1.1 CF, CIP, and CIM configuration

You must configure CF before any other cluster services, such as Reliant Monitor Services (RMS). CF defines which nodes are in a given cluster. In addition, after you configure CF and CIP, the Shutdown Facility (SF) and RMS can be run on the nodes.

The Shutdown Facility (SF) is responsible for node elimination. This means that even if RMS is not installed or running in the cluster, missing CF heartbeats will cause SF to eliminate nodes.

You can use the Cluster Admin CF Wizard to easily configure CF, CIP, and CIM for all the nodes in the cluster.

A CF configuration consists of the following main attributes:

- Cluster name - This can be any name that you choose as long as it is 31 characters or less per name and each character comes from the set of printable ASCII characters, excluding white space, newline, and tab characters. Cluster names are always mapped to upper case.
- Interconnects - Set of interfaces on each node in the cluster used for CF networking. An Ethernet device on the local node is one example of an interface.
- CF node name - By default, in Cluster Admin, the CF node names are the same as the Web-Based Admin View names; however, you can use the CF Wizard to change them. CF node names are converted to lower case.

The dedicated network connections used by CF are known as interconnects. They typically consist of some form of high speed networking such as 100 MB or Gigabit Ethernet links. These interconnects must meet the following requirements if they are to be used for CF:

- The network links used for interconnects must have low latency and low error rates. This is required by the CF protocol. Private switches and hubs will meet this requirement. Public networks, bridges, and switches shared with other devices may not necessarily meet these requirements, and their use is not recommended.

It is recommended that each CF interface be connected to its own private network with each interconnect on its own switch or hub.

- The interconnects should not be used on any network that might experience network outages of 5 seconds or more. A network outage of 10 seconds will, by default, cause a route to be marked as DOWN. *cfset(1M)* can be used to change the 10 second default. Refer to the Section "1.1.2 cfset" for additional information.

Since CF automatically attempts to bring up downed interconnects, the problem with split clusters only occurs if all interconnects experience a 10-second outage simultaneously. Nevertheless, CF expects highly reliable interconnects.

You should carefully choose the number of interconnects you want in the cluster before you start the configuration process. If you decide to change the number of interconnects after you have configured CF across the cluster, you can either bring down CF on each node to do the reconfiguration or use the *cfrecon* command. To stop CF, it is necessary to stop the higher level services (such as RMS, SF, Global File Services (hereinafter GFS)) on each node. Therefore, the reconfiguration process is complicated to influence other operations. Using the *cfrecon* command will lead to temporary asymmetrical CF configuration.



Note

Your configuration should specify at least two interconnects to avoid a single point of failure in the cluster.

Before you begin the CF configuration process, ensure that all of the nodes are connected to the interconnects you have chosen and that all of the nodes can communicate with each other over those interconnects. For proper CF configuration using Cluster Admin, all of the interconnects should be working during the configuration process.

CIP configuration involves defining virtual CIP interfaces and assigning IP addresses to them. Up to eight CIP interfaces can be defined per node. These virtual interfaces act like normal TCP/IP interfaces except that the IP traffic is carried over the CF interconnects. Because CF is typically configured with multiple interconnects, the CIP traffic will continue to flow even if an interconnect fails. This helps eliminate single points of failure as far as physical networking connections are concerned for intracluster TCP/IP traffic.

Except for their IP configuration, the eight possible CIP interfaces per node are all treated identically. There is no special priority for any interface, and each interface uses all of the CF interconnects equally. For this reason, many system administrators may choose to define only one CIP interface per node.

To ensure that you can communicate between nodes using CIP, the IP address on each node for a specific CIP interface should use the same subnet. Besides, if you use an IPv6 address, use the IPv6 address assigned to the CIP interface for communications. Communications using the link local address are not available.

CIP traffic is really intended only to be routed within the cluster. The CIP addresses should not be used outside of the cluster. Because of this, you should use addresses from the non-routable reserved IP address range.

For the IPv4 address, Address Allocation for Private Internets (RFC 1918) defines the following address ranges that are set aside for private subnets:

Subnets(s)	Class	Subnetmask
10.0.0.0	A	255.0.0.0
172.16.0.0 ... 172.31.0.0	B	255.255.0.0
192.168.0.0 ... 192.168.255.0	C	255.255.255.0

For the IPv6 address, the range where Unique Local IPv6 Unicast Addresses (RFC 4193) defined with the prefix FC00::/7 is used as the address (Unique Local IPv6 Unicast Addresses) which can be allocated freely within the private network.

For CIP nodenames, it is strongly recommended that you use the following convention for RMS:

*cfname*RMS

cfname is the CF name of the node and RMS is a literal suffix. This will be used for one of the CIP interfaces on a node. This naming convention is used in the Cluster Admin GUI to help map between normal node names and CIP names. In general, you only need to configure one CIP interface per node.



Note

In the CIP configuration, CIP names are stored in /etc/hosts. /etc/nsswitch.conf(4) should be set to use files as the first criteria when looking up nodes.

The recommended way to configure CF, CIP and CIM is to use the Cluster Admin GUI. You can use the CF/CIP Wizard in the GUI to configure CF, CIP, and CIM on all the nodes in the cluster in just a few screens. Before running the wizard, however, you must complete the following steps:

1. CF/CIP, Web-Based Admin View, and Cluster Admin should be installed on all the nodes in the cluster.
2. If you are running CF over Ethernet, then all of the interconnects in the cluster should be physically attached to their proper hubs or networking equipment and should be working.
3. Web-Based Admin View configuration must be done. Refer to "2.4.1 Management server configuration" in "PRIMECLUSTER Web-Based Admin View Operation Guide" for details.

In the *cf*tab in Cluster Admin, make sure that the CF driver is loaded on that node. Press the *Load Driver* button if necessary to load the driver. Then press the *Configure* button to start the CF Wizard.

The CF/CIP Wizard is invoked by starting the GUI on a node where CF has not yet been configured. When this is done, the GUI automatically brings up the CF/CIP Wizard in the *cf*tab of the GUI. You can start the GUI by entering the following URL with a browser running the correct version of the Java plug-in:

```
http://management_server:8081/Plugin.cgi
```

management_server is the primary or secondary management server you configured for this cluster. Refer to "4.3.3.1 Initial setup of the operation management server" in "PRIMECLUSTER Installation and Administration Guide" for details on configuring the primary and secondary management servers. Refer to "3.1.2 Prerequisite client environment" in "PRIMECLUSTER Web-Based Admin View Operation Guide" on which browsers and Java plug-ins are required for the Cluster Admin GUI.

In PRIMECLUSTER, it is recommended that you configure the administrative LAN and cluster interconnects on different NICs. However, if you cannot make such a configuration due to restrictions on hardware in KVM environment or VMware environment, the configuration which shares the administrative LAN and cluster interconnects on the NIC is also supported.

KVM environment

In the configuration which shares the administrative LAN and cluster interconnects on the NIC, you must conform all the following conditions for network and Global Link Services (hereinafter GLS):

- Make two NICs redundant by GLS Virtual NIC mode on the Host OS.
- Create the necessary number of the VLAN interfaces for the Host OS, the administrative LAN for the Host OS, public LAN, and cluster interconnects on the virtual interface.
- Create cluster interconnects for the Host OS and guest OS on their VLAN interfaces. They are not made redundant on the cluster interconnect side.
- For the public LAN, create GLS resources on the guest OS and RMS on the guest OS monitors them.

This configuration requires the CF configuration by CLI. For the configuration method, see "[1.1.7 Example of CF configuration by CLI](#)".



In this configuration, there are the following notes:

- Availability in the event of a double failure of network switch
If both network switches where two NICs are connected fail, the administrative LAN, public LAN, and cluster interconnects will enter the fault state. In this state, the Host OS and guest OS cannot be forcibly stopped and no switchover of applications occur. Note that if a double failure occurs on the NIC of a server, switchover of applications occurs because they can be forcibly stopped from the other server.
- Restriction on the timeout value of cluster interconnects
In GLS Virtual NIC mode, it takes 20 seconds to switch a path. On the other hand, the time to detect the failure of cluster interconnects is 10 seconds (default value). Therefore, with the default value, the failure of cluster interconnects will be detected first if one NIC failure occurs.
To solve this problem, change the timeout value (CLUSTER_TIMEOUT) to 40 seconds for the Host OS and 30 seconds for the guest OS.
By this setting change, the time to detect failures of cluster interconnects will be longer (from 10 seconds to 40 seconds).
- Cluster switchover due to overload of the public LAN
If a communication timeout which is more than 30 seconds occurs, PRIMECLUSTER detects a failure of cluster interconnects, forcibly stops the Host OS or guest OS, and a cluster switchover may occur.
- Restriction on the starting and stopping of GLS, and the rebooting for network service of system
When stopping and starting GLS, or rebooting the network service of System, stop CF beforehand. For instructions on stopping CF, refer to the Section "[4.6 Starting and stopping CF](#)".

VMware environment

When sharing NIC with administrative LAN and cluster interconnect in the VMware environment, separate the network allocated to the virtual machine using VMware's function. In this configuration, CF configuration can be conducted from GUI.

1.1.1 Differences between CIP and CF over IP



CF over IP is not available in Linux.

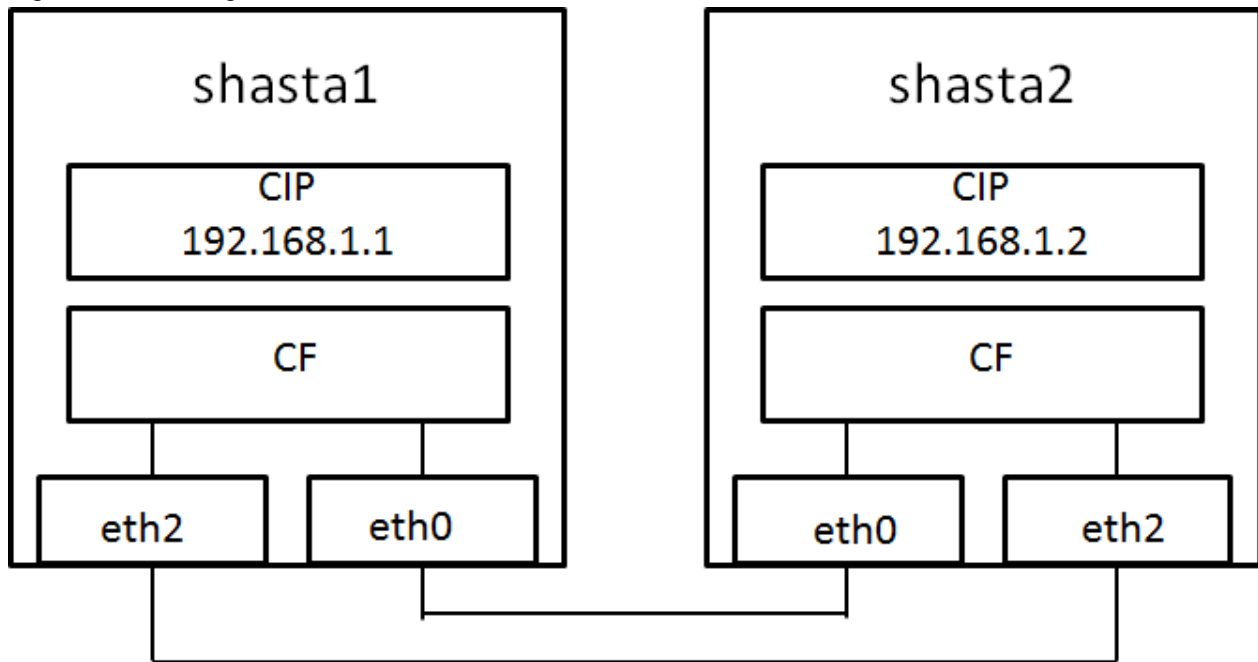
Although the two terms "CF over IP" and "CIP" (also known as "IP over CF") sound similar, they are completely different functions.

In general, the cluster interconnects are separate from the public network and not used by the TCP/IP stack. To allow applications to use TCP/UDP protocols on top of CF, CF uses the CIP driver.

CIP routes the TCP/IP traffic through the cluster interconnects to the other nodes in the cluster. CIP uses the interfaces configured for CF and does failover and load balancing if multiple interfaces are available.

CIP defines a reliable IP interface for applications on top of the Cluster Foundation (CF).

Figure 1.1 CIP diagram



CIP should not be used in a CF over IP configuration. Instead of creating an additional CIP address for applications like RMS, take the IP address assigned to CF directly.

CF over IP is used for configurations where the nodes are separated by long distances and where standard Ethernet wiring cannot be used. This can be due to the maximum cable length or different segments connected through routers. In this case, CF sends its protocol messages through the IPv4 stack instead of using the low-level network interfaces. This is not operated on IPv6. The IP stack needs to be configured up front and must be available before you start the cluster.

The CF node discovery is done by a JOIN message. This message is sent to the configured destination address. This can be the address of the remote node or the broadcast address of this subnet. Broadcast messages do not travel across routers and subnets. If the remote node is on a different subnet the broadcast address cannot be used or the cluster will not join.

It is recommended to separate the CF networks from the public network. Up to four IP devices can be configured for CF but the best way would be to configure only one IP device and let the IP layer deal with multi path or bonding.

To configure the IP interface over the private interconnect, use an IP address designed for the private network, such as in the following example:

192.168.0.x

x is an integer between 1 and 254.

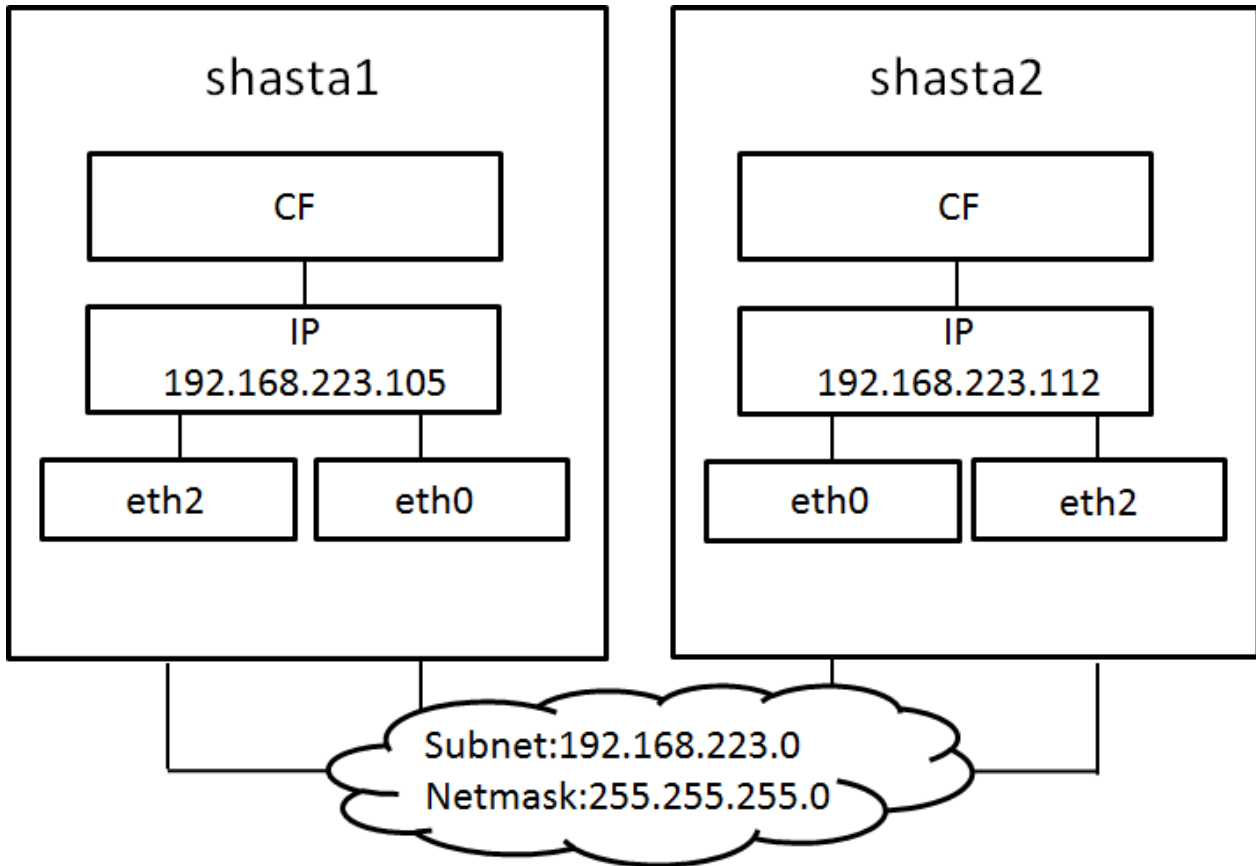
During the cluster joining process, CF sends broadcast messages to other nodes; therefore, all the nodes must be on the same local network. If one of the nodes is on a different network or subnet, the broadcast will not be received by that node. Therefore, the node will fail to join the cluster.



Note

Use CF with the Ethernet link-level connection whenever possible because CF over IP implies additional network/protocol information and usually will not perform as well.

Figure 1.2 CF over IP diagram



1.1.2 cfset

The *cfset(1M)* command can be used to set certain tunable parameters in the CF driver. The values are stored in */etc/default/cluster.config*. The *cfset(1M)* command can be used to retrieve and display the values from the kernel or the file as follows:

- A new file under */etc/default* called *cluster.config* is created.
- The values defined in */etc/default/cluster.config* can be set or changed using the GUI (for *cfcp* and *cfsh* during initial cluster configuration) or by using a text editor.
- The file consists of the following tuple entries, Name and Value:

Name:

- This is the name of a CF configuration parameter. It must be the first token in a line.
- Maximum length for Name is 31 bytes. The name must be unique.
- Duplicate names will be detected and reported as an error when the entries are applied by *cfconfig -I* and by the *cfset(1M)* command (*cfset -r* and *-f* option). This will log invalid and duplicate entries to */var/log/messages*.
- *cfset(1M)* can change the Value for the Name in the kernel if the driver is already loaded and running.

Value:

- This represents the value to be assigned to the CF parameter. It is a string, enclosed in double quotes or single quotes. Maximum length for Value is 4 kilobytes (4K) characters.
- New lines are not allowed inside the quotes.
- A new line or white space marks the close of a token.
- However, if double quotes or single quotes start the beginning of the line, *cfset* treats the line as a continuation value from the previous value.

- The maximum number of Name/Value pair entries is 100.
- The hash sign (#) is used for the comment characters. It must be the first character in the line, and it causes the entries on that line to be ignored.
- Single quotes can be enclosed in double quotes or vice versa.

cfset(1M) options are as follows:

```
cfset [ -r | -f | -a | -o name | -g name | -h ]
```



Note

Refer to the Chapter "[Chapter 9 Manual pages](#)" and to the *cfset*(1M) manual page for more details on options.

The tuneables are as follows:

- CLUSTER_TIMEOUT (refer to "[Example 1 Use cfset\(1M\) to tune timeout as follows:](#)" that follows)
- CFSH (refer to the following Section "[1.1.3 CF security](#)")
- CFCP (refer to the following Section "[1.1.3 CF security](#)")
- CLUSTER_IP_TTL (refer to "[Example 2 To set tuneables to non-default values:](#)" that follows)

This parameter becomes enabled when using CF over IP.

Tunable Description: This is the value of the TTL field in the IP-header for all CF packets.

Default value: 64

Valid values: 1-255

- CLUSTER_IP_CTRL_TOS (refer to "[Example 2 To set tuneables to non-default values:](#)" that follows)

This parameter becomes enabled when using CF over IP.

Tunable Description: This is the value of the TOS 8-bit field in the IP-header for all CF control packets. This includes cluster heartbeat packets. The default value shown below is a best-fit default which sets the 6-bit DSCP field to binary 100010. This is a DSCP forwarding 4F AF class (compatible with older IP precedence) and specifies the lowest AF drop precedence (least likely to be dropped when congestion is encountered).

Default value: 0x88 (136)

Valid values: 0-255

- CLUSTER_IP_DATA_TOS (refer to "[Example 2 To set tuneables to non-default values:](#)" that follows)

This parameter becomes enabled when using CF over IP.

Tunable Description: This is the value of the TOS 8-bit field in the IP-header for all CF data packets (non-control packets). The default value shown below is a best-fit default which sets the 6-bit DSCP field to binary 001010. This is a DSCP forwarding 1F AF class (compatible with older IP precedence) and specifies the lowest AF drop precedence (least likely to be dropped when congestion is encountered).

Default value: 0x28 (40)

Valid values: 0-255

After any change to cluster.config, run the *cfset*(1M) command as follows:

```
# cfset -r
```



Example

Example 1 Use *cfset*(1M) to tune timeout as follows:

```
CLUSTER_TIMEOUT "30"
```

This changes the default 10-second timeout to 30 seconds. The minimum value is 1 second. There is no maximum. It is strongly recommended that you use the same value on all cluster nodes.

CLUSTER_TIMEOUT represents the number of seconds that one cluster node waits for a heartbeat response from another cluster node. Once CLUSTER_TIMEOUT seconds has passed, the non-responding node is declared to be in the LEFTCLUSTER state. The default value for CLUSTER_TIMEOUT is 10, which experience indicates is reasonable for most PRIMECLUSTER installations. We allow this value to be tuned for exceptional situations, such as networks which may experience long switching delays.

Example 2 To set tuneables to non-default values:

1. Edit the /etc/default/cluster.config file and add entries for each tunable:

```
CLUSTER_IP_TTL "64"
CLUSTER_IP_CTRL_TOS "0x88"
CLUSTER_IP_DATA_TOS "0x28"
```

2. Run `cfset -f` to verify settings in file.
3. Run `cfset -r` to load new values to CF.
4. Run `cfset -a` to verify values in kernel.

1.1.3 CF security

PRIMECLUSTER includes the following facilities for cluster communications if you do not want to use .rhosts:

- cfcf/cfsh
- sshconf (not supported by Wizard Tools)

These tools are provided to allow cluster configuration in an environment which does not permit rsh and rcp. They are specialized utilities that do not provide all the functionality of rsh and rcp and are not intended as replacements.

1.1.3.1 cfcf/cfsh

CF includes the ability to allow cluster nodes to execute commands on another node (cfsh) and to allow cluster nodes to copy files from one node to another (cfcf). However, this means that your cluster interconnects must be secure since any node that can join the cluster has access to these facilities. Because of this, these facilities are disabled by default.

PRIMECLUSTER 4.1 and higher offers a chance to configure these facilities. As one of the final steps of the CF Configuration Wizard in the Cluster Adm GUI, there are two checkboxes. Checking one enables remote file copying and checking the other enables remote command execution.

To enable remote access using cfcf/cfsh, set the following parameters in cluster.config:

```
CFCF "cfcf"
CFSH "cfsh"
```

To deactivate, remove the settings from the /etc/default/cluster.config file and run `cfset -r`.

Refer to the Section "[1.1.2 cfset](#)" in this chapter for more information.

1.1.3.2 sshconf

You can use the sshconf tool to set up non-interactive ssh access among a list of nodes. Running sshconf is similar to setting up the .rhosts file for rsh.

sshconf uses the RSA authentication method and protocol version 2. If it exists, sshconf uses the default authentication key \$HOME/.ssh/id_rsa, or it creates the key if it does not already exist.



Example

Examples of the sshconf tool are as follows:

- Enable one way access between nodes:

```
fuji2# sshconf fuji3 fuji4 fuji5
```

Running this command on fuji2 sets up one way ssh access from fuji2 to fuji3, fuji4, and fuji5 respectively.

- Disable one-way access to a node:

```
fuji2# sshconf -d fuji3 fuji4 fuji5
```

Running this command on fuji2 disables ssh access from fuji2 to fuji3, fuji4, and fuji5. This means that fuji2 does not have ssh access to fuji3, fuji4, and fuji5; however, fuji3, fuji4, and fuji5 still have the same ssh access as before running the command.

- Enable two-way access without password:

```
fuji2# sshconf -c fuji3 fuji4 fuji5
```

Running this command on fuji2 sets up ssh access among fuji3, fuji4, and fuji5 without being asked for a password. Note that fuji2 (where the command is run) is not automatically included. fuji2 only has one-way ssh access to fuji3, fuji4, and fuji5.



ssh is not supported by Wizard Tools.

1.1.4 Signed applets

Cluster Admin uses Java applets. The main advantage of trusting signed applets is that Cluster Admin can use the client system's resources. For example, you can copy and paste messages from the Java window into other applications.

When Cluster Admin is first started, a Java security warning dialog allows you to choose the security level for the current and future sessions.

1.1.5 Example of creating a cluster

The following example shows what the Web-Based Admin View and Cluster Admin screens would look like when creating a two-node cluster. The nodes involved are named fuji2 and fuji3, and the cluster name is FUJI.

This example assumes that Web-Based Admin View configuration has already been done. fuji2 is assumed to be configured as the primary management server for Web-Based Admin View, and fuji3 is the secondary management server.

The first step is to start Web-Based Admin View by entering the following URL in a java-enabled browser:

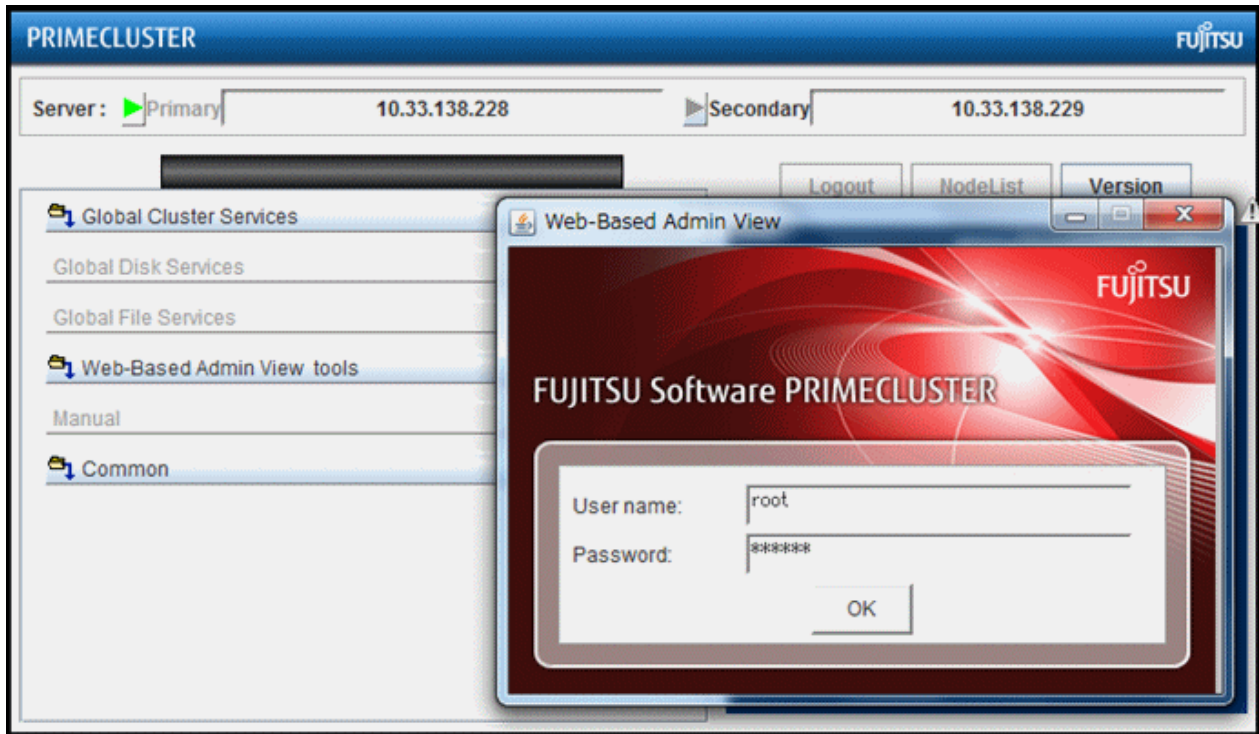
```
http://Management_Server:8081/Plugin.cgi
```

fuji2 is a management server. Enter the following:

```
http://fuji2:8081/Plugin.cgi
```

After a few moments, a login pop-up appears asking for a user name and password.

Figure 1.3 Login pop-up



Since you will be running the Cluster Admin CF Wizard, which does configuration work, you will need a privileged user ID such as root. There are three possible categories of users with sufficient privilege:

- The user root - You can enter root for the user name and root's password on fuji2. The user root is always given the maximum privilege in Web-Based Admin View and Cluster Admin.
- A user in group clroot - You can enter the user name and password for a user on fuji2 who is part of the UNIX group clroot. This user will have maximum privilege in Cluster Admin, but will be restricted in what Web-Based Admin View functions they can perform. This should be fine for CF configuration tasks.
- A user in group wvroot - You can enter the user name and password for a user on fuji2 who is part of the UNIX group wvroot. Users in wvroot have maximum Web-Based Admin View privileges and are also granted maximum Cluster Admin privileges.

For further details on Web-Based Admin View and Cluster Admin privilege levels, refer to "4.3.1 Assigning Users to Manage the Cluster" in "PRIMECLUSTER Installation and Administration Guide."

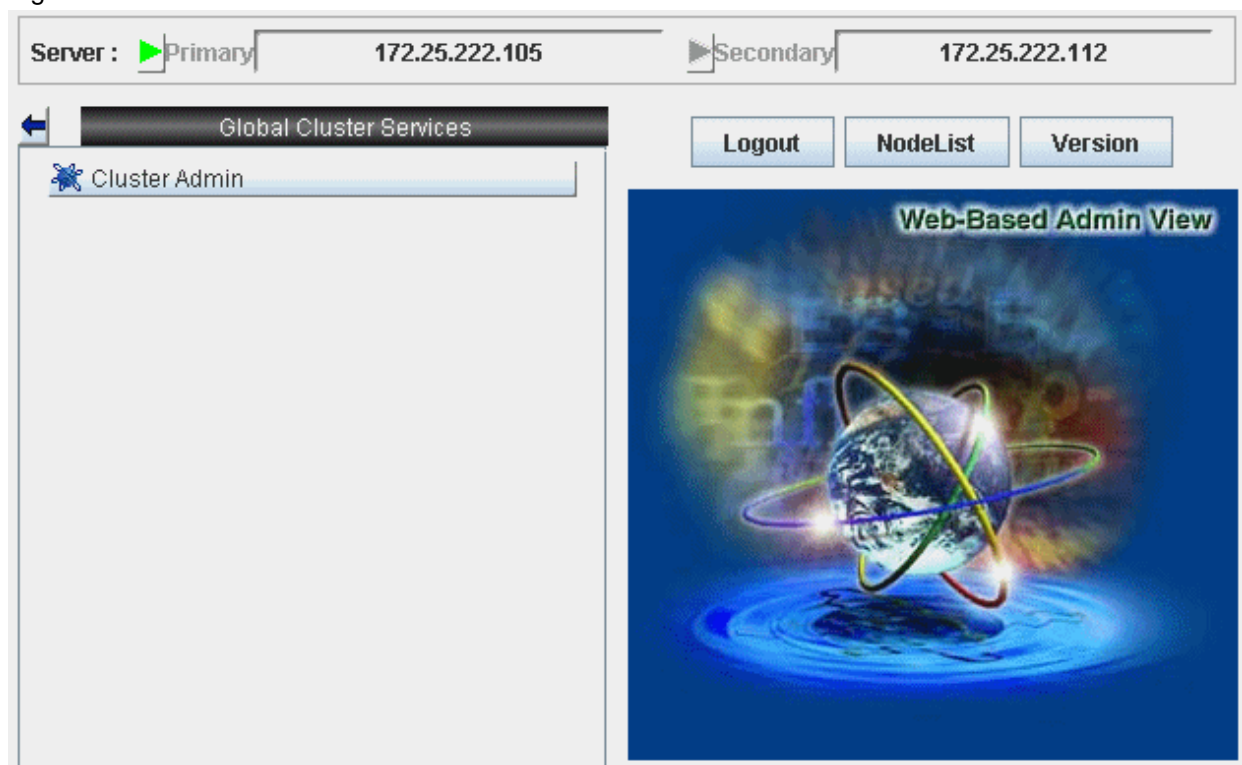
After clicking on the *OK* button, the top menu appears. Click on the button labeled *Global Cluster Services*.

Figure 1.4 Main Web-Based Admin View window after login



The Cluster Admin selection window appears.

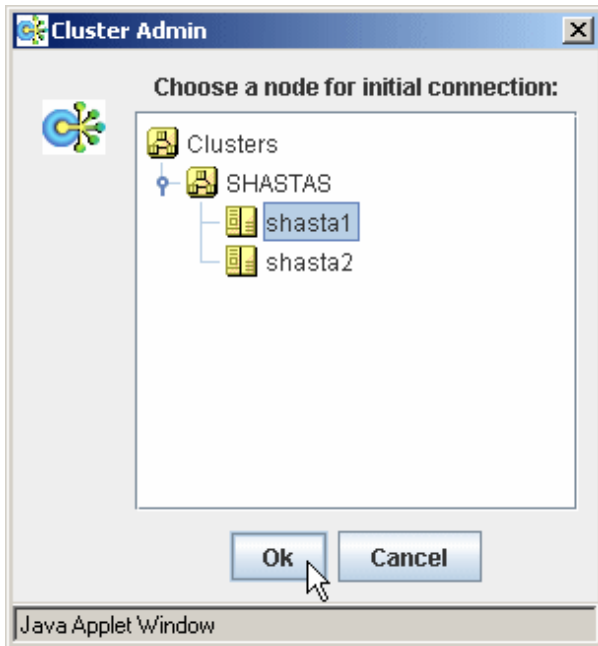
Figure 1.5 Global Cluster Services window in Web-Based Admin View



Click on the button labeled *Cluster Admin* to launch the Cluster Admin GUI.

The *Choose a node for initial connection* window appears.

Figure 1.6 Initial connection pop-up

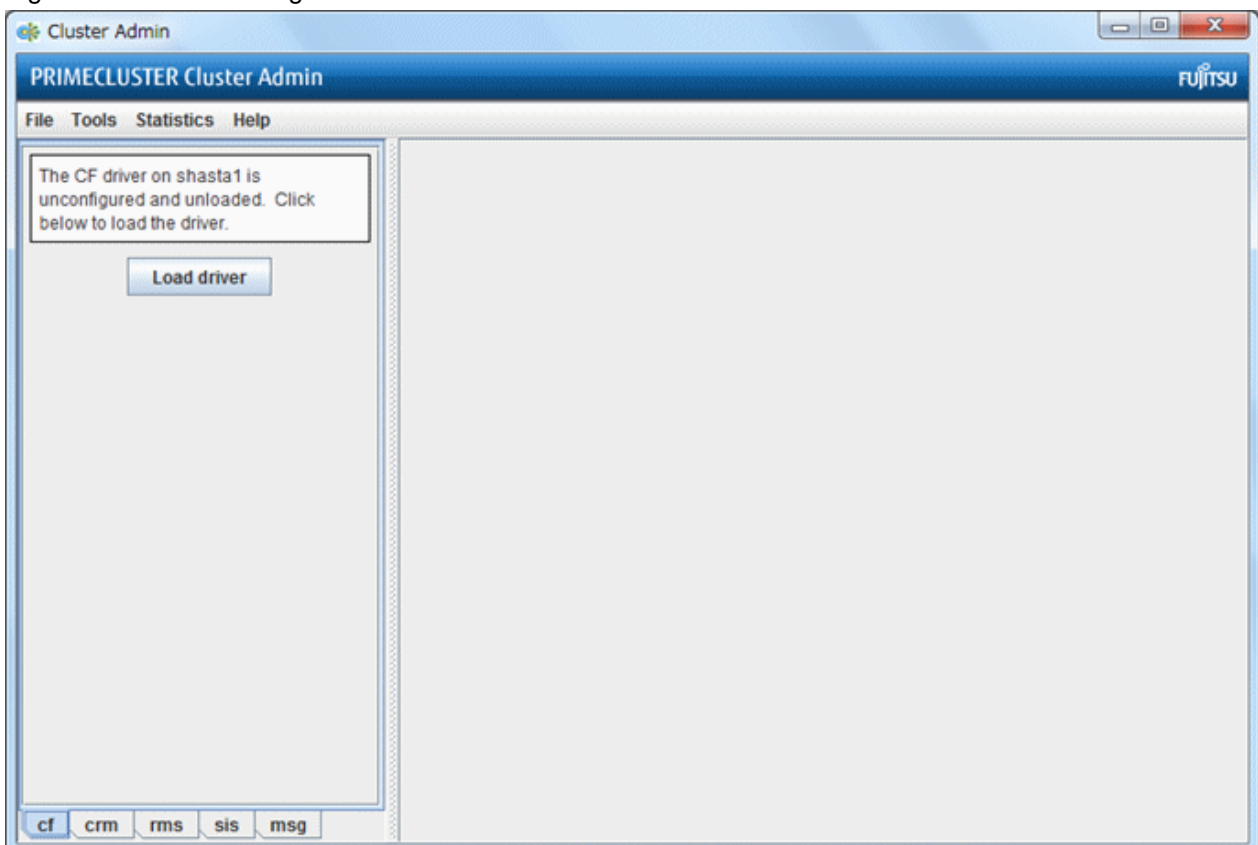


The *Choose a node for initial connection* window lists the nodes that are known to the Web-Based Admin View management station. If you select a node where CF has not yet been configured, then Cluster Admin will let you run the CF Wizard on that node.

In this example, neither fuji2 nor fuji3 have had CF configured, so either would be acceptable as a choice. In Figure 1.6, fuji2 is selected. Clicking on the *OK* button causes the main Cluster Admin GUI to appear. Since CF is not configured on fuji2, a window similar to the following figure appears.

When cancelling the startup of the Cluster Admin GUI, click the <Cancel> button in the initial connection pop-up screen.

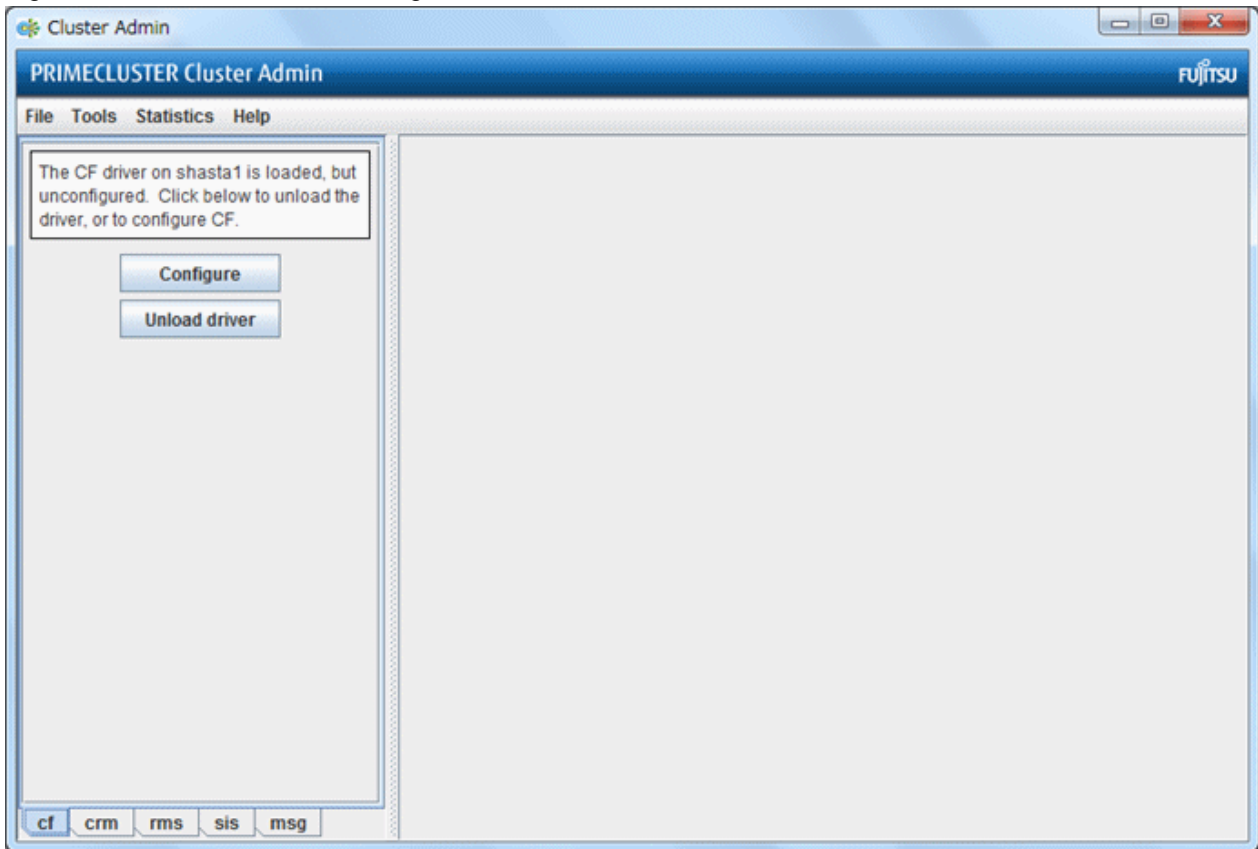
Figure 1.7 CF is unconfigured and unloaded



Click on the *Load driver* button to load the CF driver.

A window indicating that CF is loaded but not configured appears.

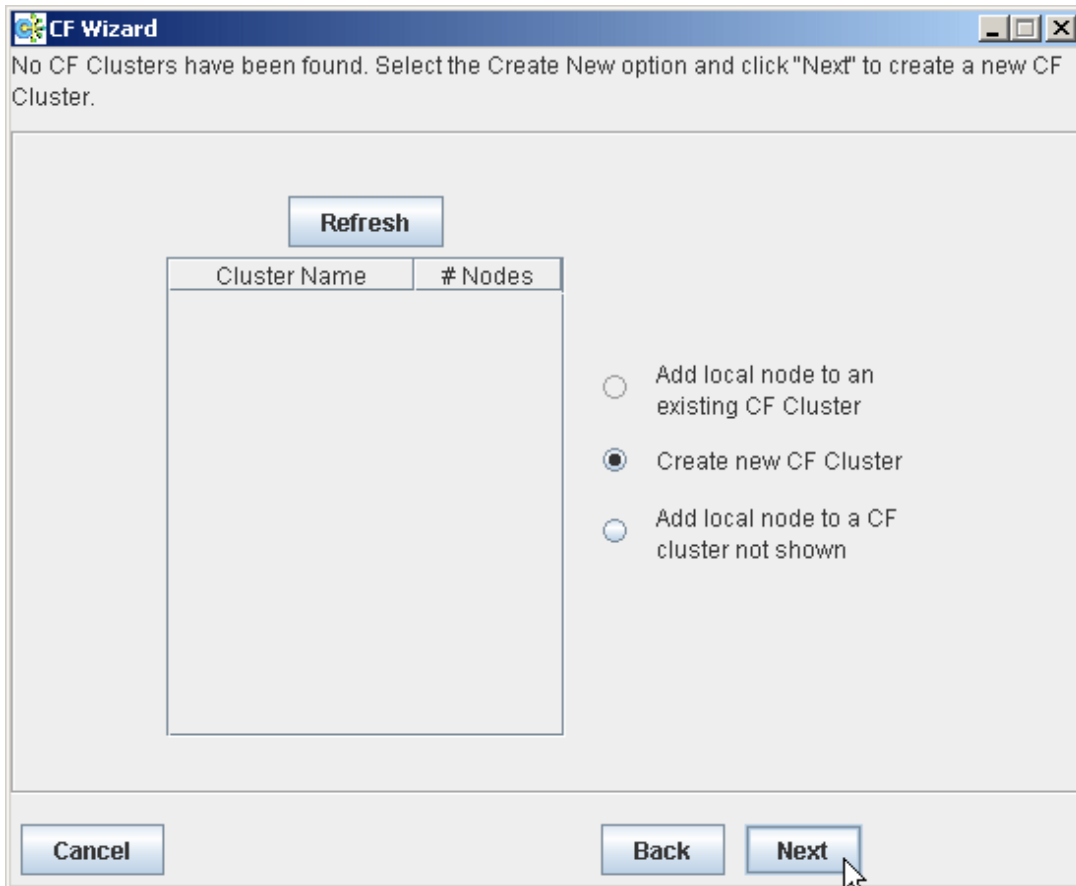
Figure 1.8 CF loaded but not configured



Click on the Configure button to bring up the CF Wizard. The CF Wizard scans for existing clusters.

After the CF Wizard finishes looking for clusters, a window similar to Figure 1.9 appears.

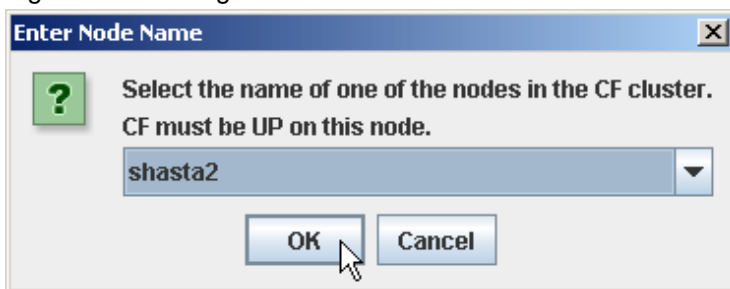
Figure 1.9 Creating or joining a cluster



This window lets you decide if you want to join an existing cluster or create a new one.

A pure CF over IP cluster will not show up in the *Cluster Name* column. To join a CF over IP cluster, select the *Add local node to a CF cluster not shown* radio button and click *Next*.

Figure 1.10 Adding a local node to a CF cluster not shown

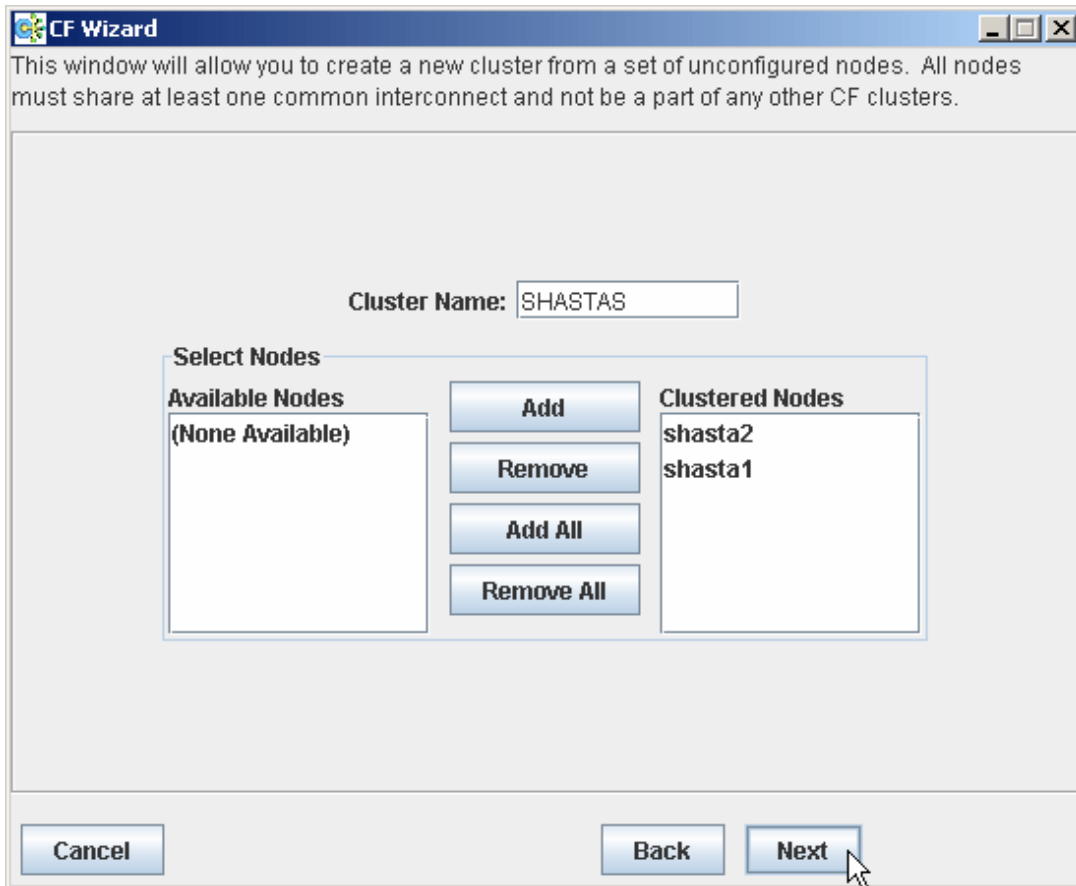


Enter the node name of the CF cluster that you want to join. Click *OK* to proceed. After scanning the node and retrieving the existing cluster's details, the CF wizard takes you to the window for joining an existing cluster.

To create a new cluster, select that the *Create new CF Cluster* radio button as shown in Figure 1.9. Then, click *Next*.

The window for creating a new cluster or for joining an existing cluster appears, depending on your previous selection. Figure 1.11 shows the window for creating a new cluster. The window for joining an existing cluster is very similar, except you cannot change the cluster name.

Figure 1.11 Selecting cluster nodes and the cluster name



This window lets you choose the cluster name and also determine what nodes will be in the cluster. In the example above, we have chosen FUJI for the cluster name.

Below the cluster name are two boxes. The one on the right, under the label *Clustered Nodes*, contains all the nodes that you want to become part of this CF cluster. The box on the left, under the label *Available Nodes*, contains all the other nodes known to the Web-Based Admin View management server. You should select nodes in the left box and move them to the right box using the *Add* or *Add All* button. If you want all of the nodes in the left box to be part of the CF cluster, then just click on the *Add All* button.

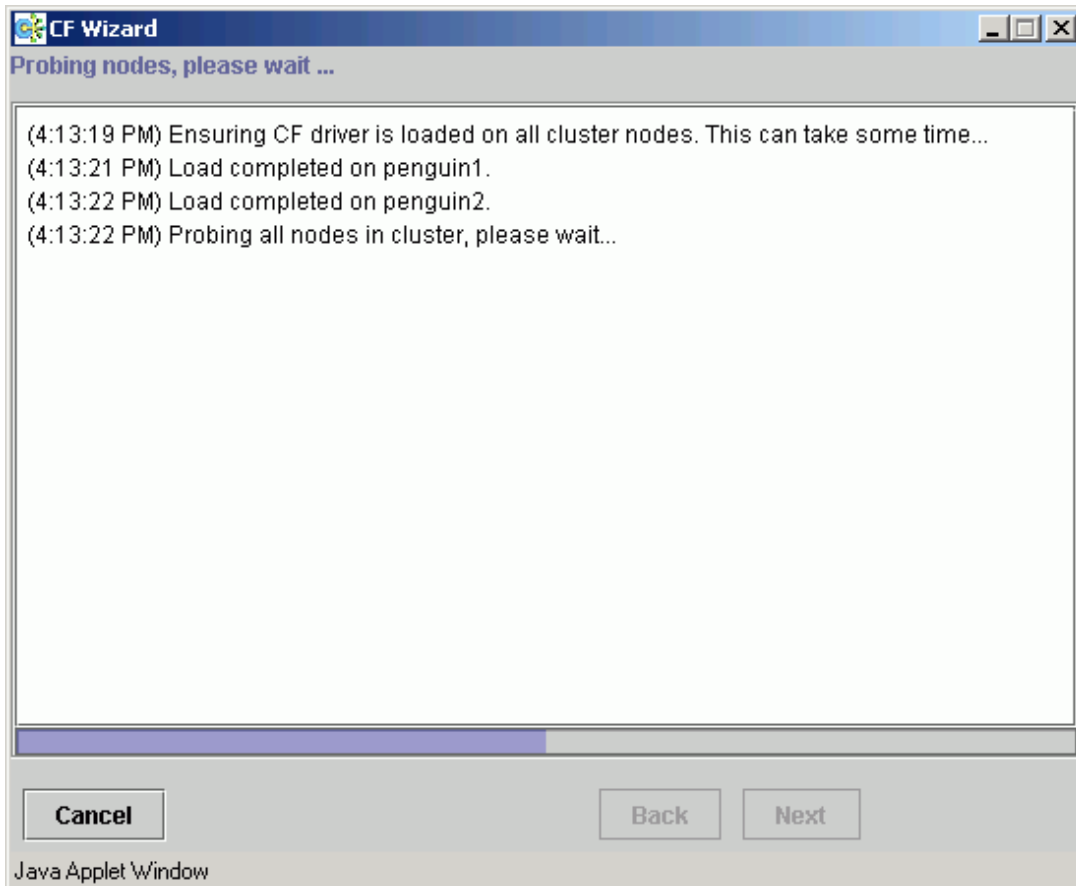
If you get to this window and you do not see all of the nodes that you want to be part of this cluster, then there is a very good chance that you have not configured Web-Based Admin View properly. When Web-Based Admin View is initially installed on the nodes in a potential cluster, it configures each node as if it were a primary management server independent of every other node. If no additional Web-Based Admin View configuration were done, and you started up Cluster Admin on such a node, then Figure 1.11 would show only a single node in the right-hand box and no additional nodes on the left-hand side. If you see this, then it is a clear indication that proper Web-Based Admin View configuration has not been done.

Refer to "4.3 Preparations for Starting the Web-Based Admin View Screen" in "PRIMECLUSTER Installation and Administration Guide."

After you have chosen a cluster name and selected the nodes to be in the CF cluster, click on the *Next* button.

The CF Wizard then loads CF on all the selected nodes and does CF pings to determine the network topology. While this activity is going on, a window similar to Figure 1.12 appears.

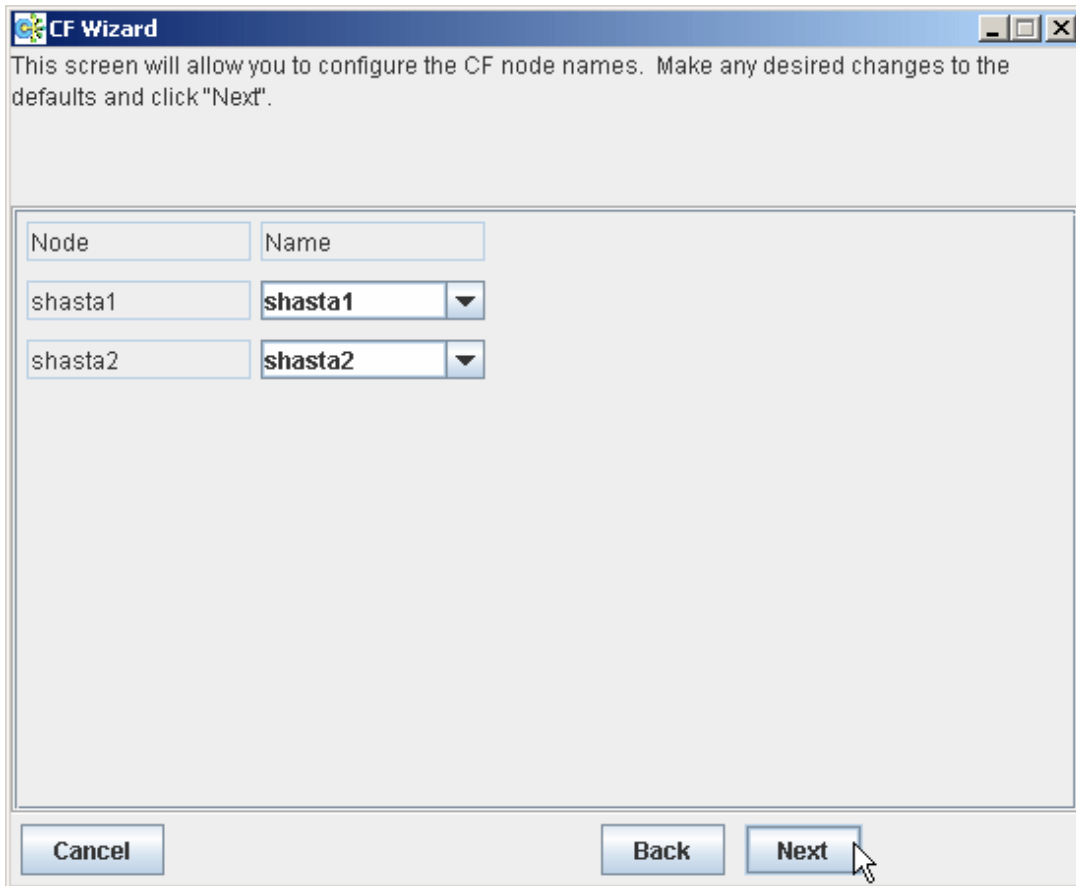
Figure 1.12 CF loads and pings



Usually, loading the CF driver is a relatively quick process. However, on some systems which use large disk arrays, the first CF load can take several minutes.

The window that allows you to edit the CF node names for each node appears. By default, the CF node names, which are shown in the right-hand column, are the same as the Web-Based Admin View names which are shown in the left-hand column.

Figure 1.13 Edit CF node names



The image shows a Windows-style dialog box titled "CF Wizard". The title bar includes standard minimize, maximize, and close buttons. The main text area contains the instruction: "This screen will allow you to configure the CF node names. Make any desired changes to the defaults and click 'Next'." Below this text is a table with two columns: "Node" and "Name". The table has two rows of data. The first row shows "shasta1" in the "Node" column and a dropdown menu in the "Name" column with "shasta1" selected. The second row shows "shasta2" in the "Node" column and a dropdown menu in the "Name" column with "shasta2" selected. At the bottom of the dialog, there are three buttons: "Cancel", "Back", and "Next". A mouse cursor is pointing at the "Next" button.

Node	Name
shasta1	shasta1
shasta2	shasta2

Buttons: Cancel, Back, Next

Make any changes to the CF node name and click *Next*.

After the CF Wizard has finished the loads and the pings, the CF topology and connection table appears.

Figure 1.14 CF topology and connection table

CF Wizard

Select the interconnects to use for CF. Nodes marked with a * will only show interconnects that are configured.

Choose interconnects based on: ☒ Connections ☐ Topology Refresh

SHASTA	<input checked="" type="checkbox"/> Int 1	<input type="checkbox"/> Int 2	<input checked="" type="checkbox"/> Int 3
shasta1 *	eth0	eth1	eth2
shasta2 *	eth0	eth1	eth2

Configuration is OK.

Cancel Back Next

Before using the CF topology and connection table in Figure 1.14, you should understand the following terms:

- Full interconnect - An interconnect where CF communication is possible to all the nodes in the cluster.
- Partial interconnect - An interconnect where CF communication is possible between at least two nodes, but not to all the nodes. If the devices on a partial interconnect are intended for CF communications, then there is a networking or cabling problem somewhere.
- Unconnected devices - These devices are potential candidates for CF configuration, but are not able to communicate with any other nodes in the cluster.

The CF Wizard determines all the full interconnects, partial interconnects, and unconnected devices in the cluster using CF pings. If there are one or more full interconnects, then it will display the connection table shown in the figure above.

Connections table

The connection table lists all full interconnects. Each column with an Int header represents a single interconnect. Each row represents the devices for the node whose name is given in the left-most column. The name of the CF cluster is given in the upper-left corner of the table.

In Figure 1.14, for example, Int 1 has eth0 on shasta1 and shasta2 attached to it. The cluster name is SHASTA.



Note

The connections and topology tables typically show devices that are on the public network. Using devices on a public network is a security risk; therefore, in general, do not use any devices on the public network as a CF interconnect. Instead, use devices on a private network.

Although the CF Wizard may list Int 1, Int 2, and so on, it should be pointed out that this is simply a convention in the GUI. CF itself does not number interconnects. Instead, it keeps track of point-to-point routes to other nodes.

Occasionally, there may be problems setting up the networking for the cluster. Cabling errors may mean that there are no full interconnects. If you click on the button next to *Topology*, the CF Wizard will display all the full interconnects, partial interconnects, and unconnected

devices it has found. If a particular category is not found, it is omitted. For example, in Figure 1.14, only full interconnects are shown because no partial interconnects or unconnected devices were found on fuji2 or fuji3.

To configure CF using the connection table, click on the interconnects that have the devices that you want to use.

When you are satisfied with your choices, click on *Next* to go to the CF over IP configuration window.

Topology table

The topology table gives more flexibility in configuration than the connection table. In the connection table, you could only select an interconnect, and all devices on that interconnect would be configured. In the topology table, you can individually select devices.

While you can configure CF using the topology table, you may wish to take a simpler approach. If no full interconnects are found, then display the topology table to see what your networking configuration looks like to CF. Using this information, correct any cabling or networking problems that prevented the full interconnects from being found. Then go back to the CF Wizard window where the cluster name was entered and click on *Next* to cause the Wizard to reprobe the interfaces. If you are successful, then the connection table will show the full interconnects, and you can select them. Otherwise, you can repeat the process.

The text area at the bottom of the window lists problems or warnings concerning the configuration.

When you are done, click on *Next* to go to the CF over IP configuration window.

Figure 1.15 CF over IP window

CF Wizard

This screen will allow you to configure CF to run over IP. This is optional unless you chose no physical interconnects and is not required for many clusters. If needed, choose a number of IP interconnects, and interfaces for each node on each interconnect. If Auto Subnet Grouping is checked, changing one interface will change all others on the same interconnect to be consistent. You should normally leave this checked.

Enter desired number of IP interconnects: 2

Auto Subnet Grouping ☒

IP Interconnects

Interface	Interconnect 1	Interconnect 2
shasta2	192.168.223.105 [eth2]	172.25.222.105 [eth0]
shasta1	192.168.223.112 [eth2]	172.25.222.112 [eth0]

Cancel Back Next

The screen shown in Figure 1.15 lets you configure CF to run over IP. If you have already configured CF over Ethernet in the topology table or the connection table, you do not have to change any settings on this screen. Leave the number of IP interconnects set to its default of 0, and click *Next*.

Note

CF over IP is not supported in Linux.

Remain the default value of the number of IP interconnects to 0 and click *Next* to skip this screen

CF can use either Ethernet packets or IP for its communication. The topology table and connection table discussed previously allow you to configure CF to use Ethernet packets. This is the preferred CF configuration since CF over Ethernet is significantly faster than CF over IP.

However, CF over Ethernet requires Ethernet link-level connectivity between the nodes in a cluster. In certain disaster recovery scenarios, there may only be IP connectivity between hosts. This is typically the case when the hosts are separated by large geographical distances.

CF over IP uses IP subnetworks in the same way that CF over Ethernet uses physical interconnects. Each IP interconnect must correspond to exactly one subnetwork. For example, suppose that your nodes had the following IP interfaces:

node	subnet1	subnet 2	subnet 3
shasta1	172.25.222.105	192.168.223.105	185.33.48.105
shasta2	172.25.222.112	192.168.223.112	185.33.48.112

Using CF over IP, you might configure one IP interconnect to use the IP addresses 192.168.223.105 and 192.168.223.112. You could configure a second IP interface using the addresses 185.33.48.105 and 185.33.48.112.

But if you need CF over IP, then set the number of IP interconnects to 2 (or more if desired). The Wizard will propose IP interconnects. The IP interconnects are conveniently sorted by subnetwork. If you think that a particular subnetwork is missing a node, then double check that the netmask and broadcast addresses are properly configured for all the nodes on subnetwork.

Select the subnetworks that you want to use as your IP interconnects. You should avoid using addresses on your public network. CF allows promiscuous joins without any limits, so it is best to use private subnetworks for your IP interconnects.

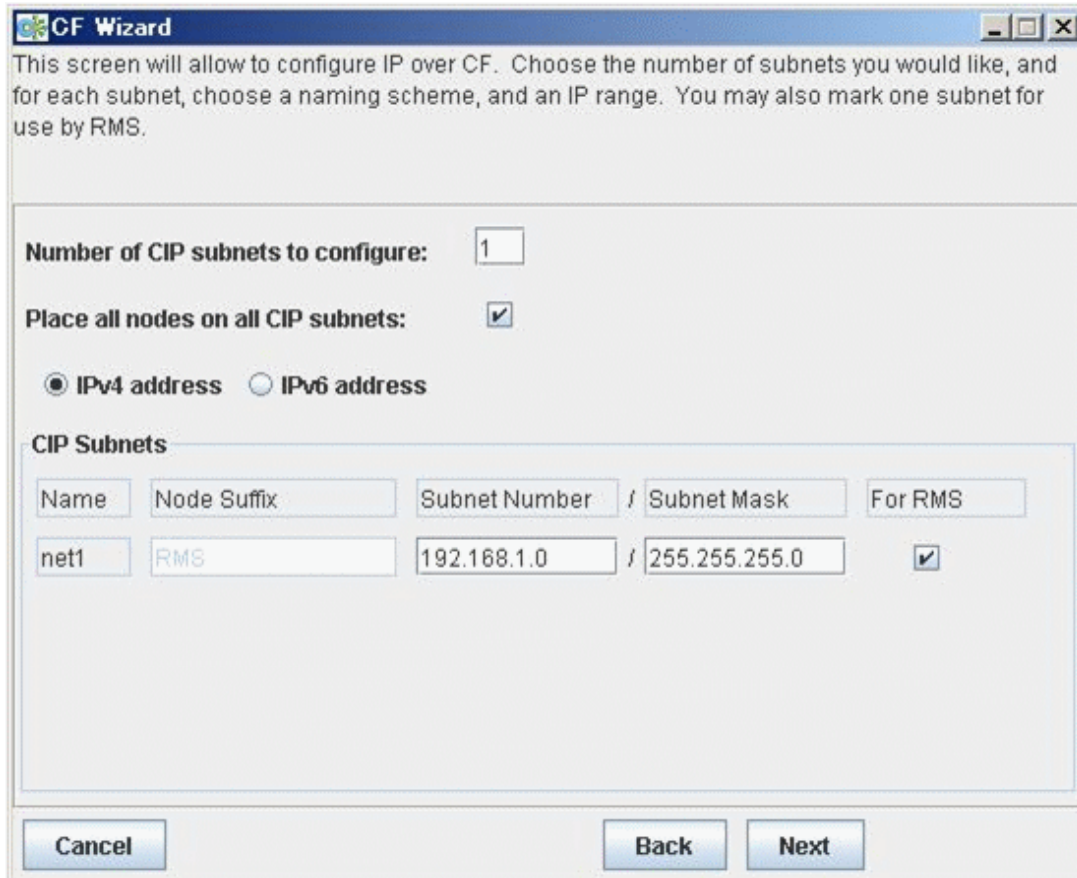
With this setting, CF can be configured to run over the IP interface. After entering the number of required IP interconnects and pressing [Enter], the CF Wizard will display interconnects sorted by available subnetworks, netmasks, and broadcast addresses.

All the IP addresses for all the nodes on a given IP interconnect must be on the same IP subnetwork and should have the same netmask and broadcast address.

Auto *Subnet Grouping* should always be checked in this window. If it is checked and you select one IP address for one node, then all of the other nodes in that column have their IP addresses changed to interfaces on the same subnetwork.

Choose the IP interconnects from the combo boxes on this window, and click on *Next*. The CIP Wizard windows like Figure 1.16 and Figure 1.17 appear.

Figure 1.16 CIP wizard (IPv4) window



The image shows a Windows-style dialog box titled "CF Wizard". It contains instructions at the top and several configuration fields. The "Number of CIP subnets to configure" is set to 1. The "Place all nodes on all CIP subnets" checkbox is checked. The "IPv4 address" radio button is selected. Below is a table for "CIP Subnets" with one row of data. At the bottom are "Cancel", "Back", and "Next" buttons.

CF Wizard

This screen will allow to configure IP over CF. Choose the number of subnets you would like, and for each subnet, choose a naming scheme, and an IP range. You may also mark one subnet for use by RMS.

Number of CIP subnets to configure:

Place all nodes on all CIP subnets: ☒

☒ IPv4 address ☐ IPv6 address

CIP Subnets

Name	Node Suffix	Subnet Number	Subnet Mask	For RMS
net1	RMS	192.168.1.0	255.255.255.0	<input checked="" type="checkbox"/>

Figure 1.17 CIP wizard (IPv6) window

CF Wizard

This screen will allow to configure IP over CF. Choose the number of subnets you would like, and for each subnet, choose a naming scheme, and an IP range. You may also mark one subnet for use by RMS.

Number of CIP subnets to configure:

Place all nodes on all CIP subnets: ☒

☐ IPv4 address ☒ IPv6 address

CIP Subnets

Name	Node Suffix	Network Prefix	Prefix Length	For RMS
net1	RMS	FD00:0:0:1::	64	<input checked="" type="checkbox"/>

Cancel Back Next

This window allows you to configure CIP. You can enter a number in the box after *Number of CIP subnets to configure* to set the number of CIP subnets to configure. The maximum number of CIP subnets is 8.

For each defined subnet, the CIP Wizard configures a CIP interface on each node defined in the CF cluster.

Set either IPv4 or IPv6 as the IP address to set to the CIP interface.

By selecting either of the [IPv4 address] or [IPv6 address] radio button, you can switch "[Figure 1.16 CIP wizard \(IPv4\) window](#)" and "[Figure 1.17 CIP wizard \(IPv6\) window](#)".

When using IPv4 for CIP interface

The following values are assigned for CIP interface:

- The IP address will be a unique IP number on the subnet specified in the *Subnet Number* field. The node portions of the address start at 1 and are incremented by 1 for each additional node.

The CIP Wizard will automatically fill in a default value for the *Subnet Number* field for each CIP subnetwork requested. The default values are taken from the private IP address range specified by RFC 1918. Note that the values entered in the *Subnet Number* field have 0 for their node portion even though the CIP Wizard starts the numbering at 1 when it assigns the actual node IP addresses.

- The IP name of the interface will be of the form *cfnameSuffix* where *cfname* is the name of a node from the CF Wizard, and the *Suffix* is specified in the field *Node Suffix*. If the checkbox *For RMS* is selected, then the *Node Suffix* will be set to RMS and will not be editable. If you are using RMS, one CIP network must be configured for RMS.
- The *Subnet Mask* will be the value specified.

In "[Figure 1.16 CIP wizard \(IPv4\) window](#)", the system administrator has selected 1 CIP network. The *For RMS* checkbox is selected, so the RMS suffix will be used. Default values for the *Subnet Number* and *Subnet Mask* are also selected. The nodes defined in the CF cluster are fuji2 and fuji3. This will result in the following configuration:

- On fuji2, a CIP interface will be configured with the following:

```
CIP nodename: fuji2RMS

IP address: 192.168.1.1

Subnet Mask: 255.255.255.0
```

- On fuji3, a CIP interface will be configured with the following:

```
CIP nodename: fuji3RMS

IP address: 192.168.1.2

Subnet Mask: 255.255.255.0
```

When using IPv6 for CIP interface

The following values are assigned for CIP interface:

- The IP address will be a unique IP number on the network prefix specified in the *Prefix* field. Interface ID of the address start at 1 and are incremented by 1 for each additional node.

The CIP Wizard will automatically fill in a default value for the Prefix field for each CIP subnetwork requested. The default values are taken from the Unique Local Unicast Address range specified by RFC 4193. Note that the values entered in the *Prefix* field have 0 for their interface ID portion even though the CIP Wizard starts the numbering at 1 when it assigns the actual node IP addresses.

- The IP name of the interface is represented in *cfnameSuffix* format.
 - *cfname* is the name of a node given by the CF Wizard
 - *Suffix* is specified in the field *Node Suffix*.

If the checkbox *For RMS* is selected, then the *Node Suffix* will be set to RMS and will not be editable. If you are using RMS, one CIP network must be configured for RMS.

- The *Prefix Length* will be the value specified.

In "Figure 1.17 CIP wizard (IPv6) window", the system administrator has selected 1 CIP network. The *For RMS* checkbox is selected, so the RMS suffix will be used. Default values for the *Prefix* and *Prefix Length* are also selected. The nodes defined in the CF cluster are fuji2 and fuji3. This will result in the following configuration:

- On fuji2, a CIP interface will be configured with the following:

```
CIP nodename: fuji2RMS

IPv6 address: FD00:0:0:1::1

Prefix Length: 64
```

- On fuji3, a CIP interface will be configured with the following:

```
CIP nodename: fuji3RMS

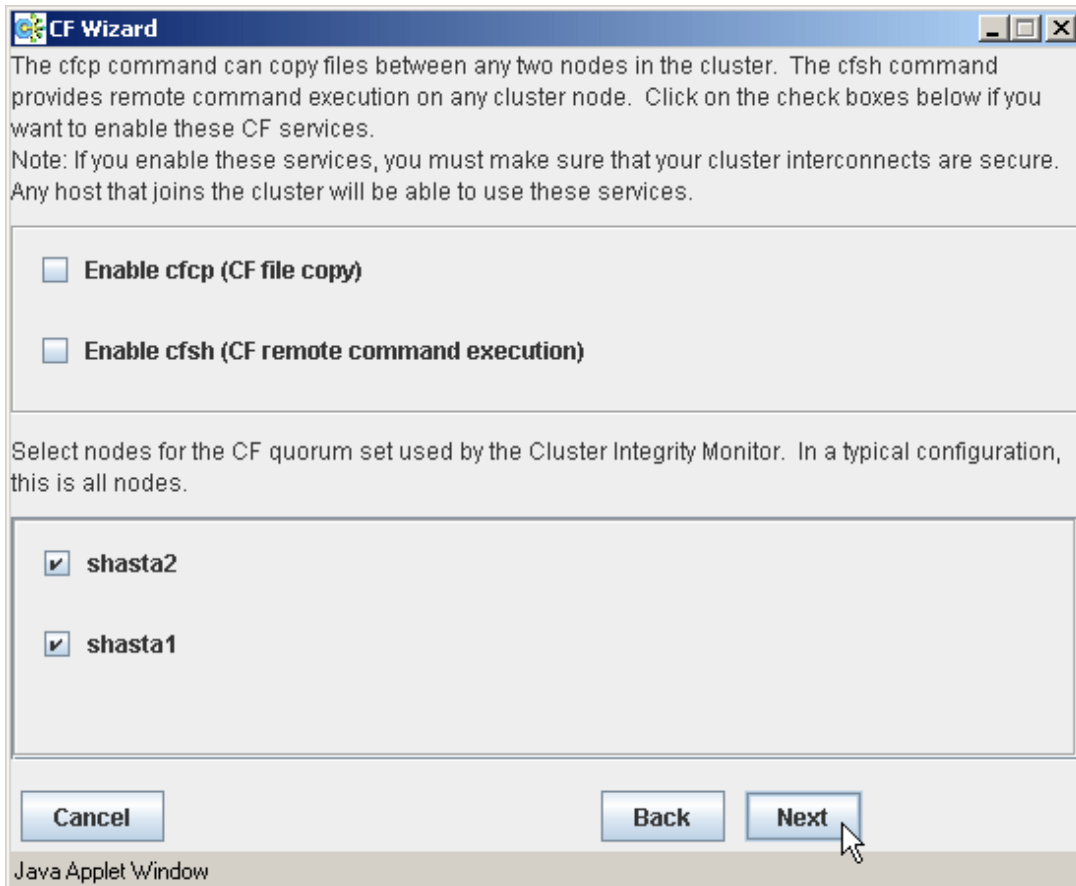
IPv6 address: FD00:0:0:1::2

Prefix Length: 64
```

The CIP Wizard stores the configuration information in the file `/etc/cip.cf` on each node in the cluster. This is the default CIP configuration file. The Wizard will also update `/etc/hosts` on each node in the cluster to add the new IP node names.

When you click on the Next button, CIM configuration window appears.

Figure 1.18 CIM configuration window



The CIM configuration window in Figure 1.18 has the following parts:

- The upper portion allows you to enable *cfcp* and *cfsh*.

cfcp is a CF-based file copy program. It allows files to be copied among the cluster hosts. *cfsh* is a remote command execution program that, similar to *cfcp*, works between nodes in the cluster. The use of these programs is optional. In this example these items are not selected. If you enable these services, however, any node that has access to the cluster interconnects can copy files or execute commands on any node with root privileges.

- The lower portion allows you to determine which nodes should be monitored by CIM.

This window also lets you select which nodes should be part of the CF quorum set. The CF quorum set is used by the CIM to tell higher level services when it is safe to access shared resources.

Note

Do not change the default selection of the nodes

A checkbox next to a node means that node will be monitored by CIM. By default, all the nodes are checked. For almost all configurations, you will want to have all the nodes monitored by CIM.

This window will also allow you to configure CF Remote Services. You can enable either remote command execution, remote file copying, or both.

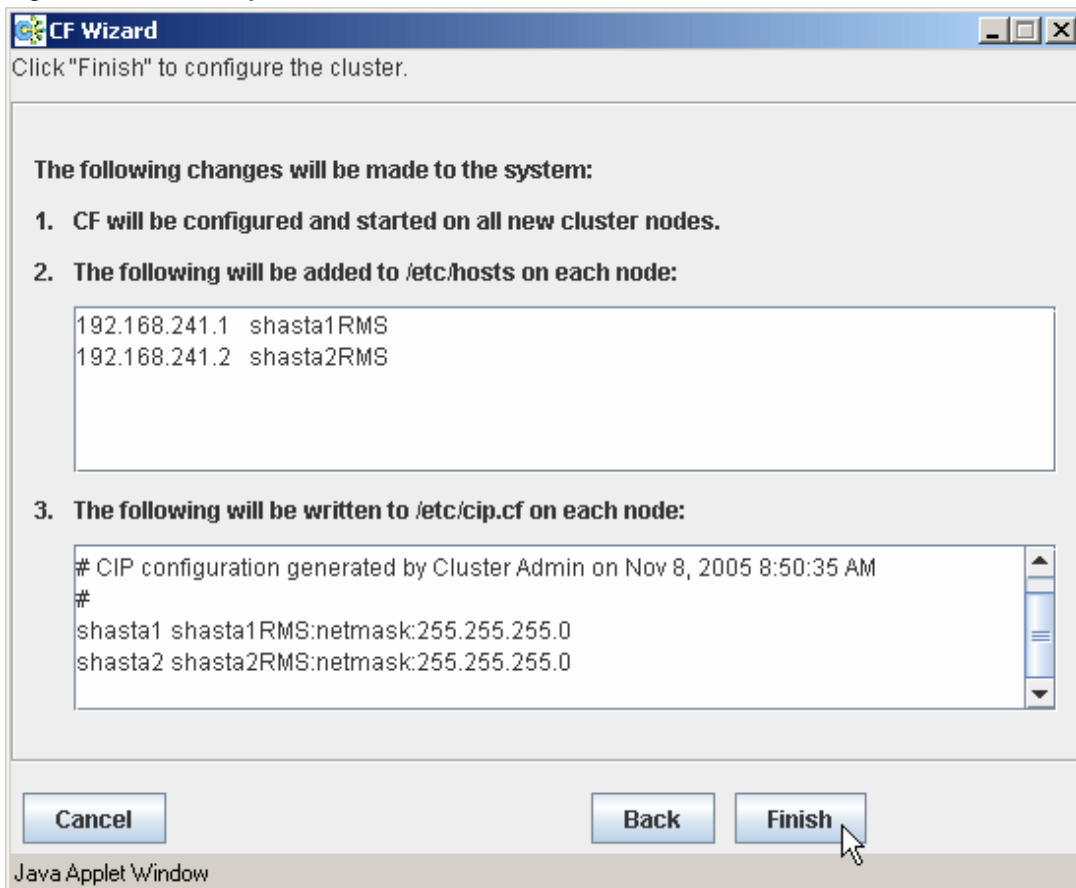
Note

- Enabling either of these means that you must trust all the nodes on the CF interconnects and the CF interconnects must be secure. Otherwise any system able to connect to the CF interconnects will have access to these services.

- To use RMS, make sure to configure *cfc* and *cfs*.

Click on the Next button to go to the summary window.

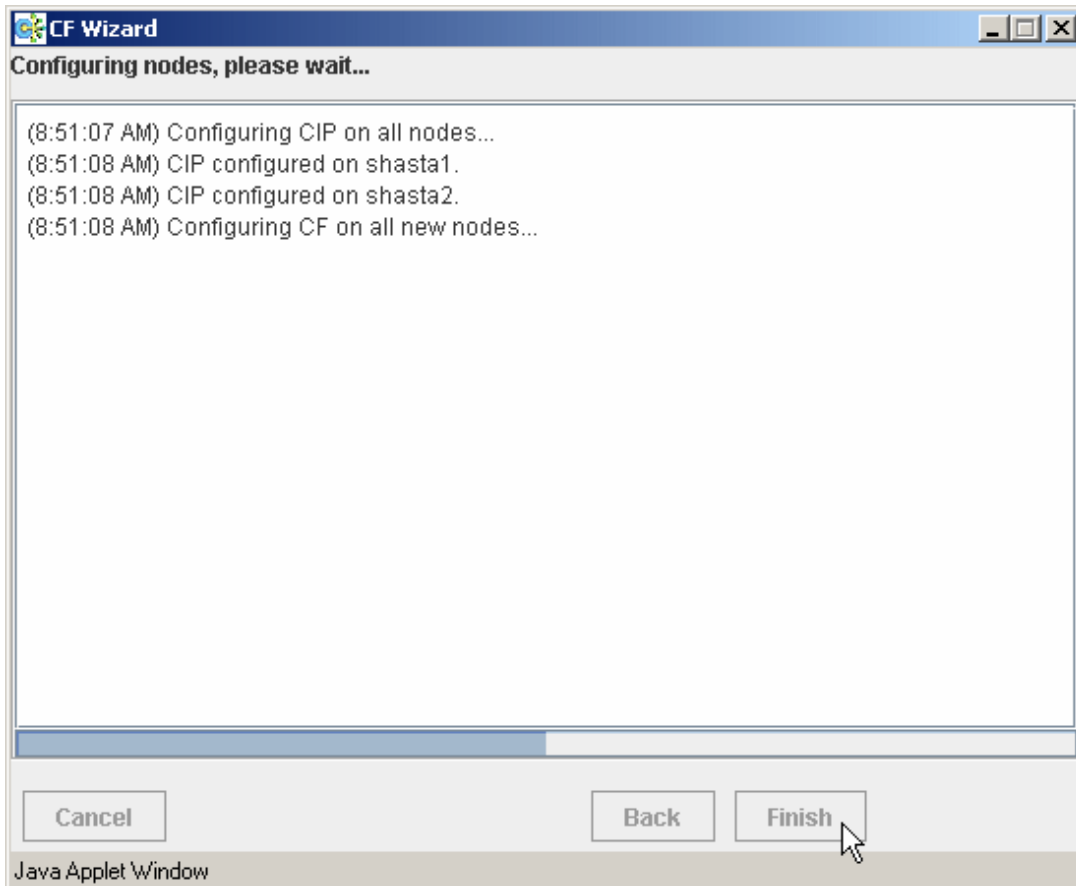
Figure 1.19 Summary window



This window summarizes the major changes that the CF, CIP, and CIM Wizards will perform. When you click on the Finish button, the CF Wizard performs the actual configuration on all the nodes.

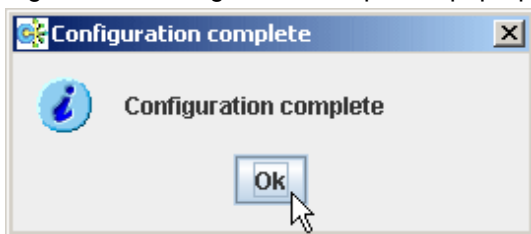
A window similar to the following figure is displayed while the configuration is being done.

Figure 1.20 Configuration processing window



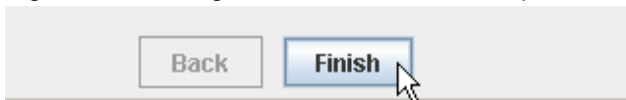
This window is updated after each configuration step. When the configuration successfully completes, a small completion pop-up window appears.

Figure 1.21 Configuration completion pop-up



Click on the OK button to close the pop-up window. The configuration processing window now has a Finish button.

Figure 1.22 Configuration window after completion

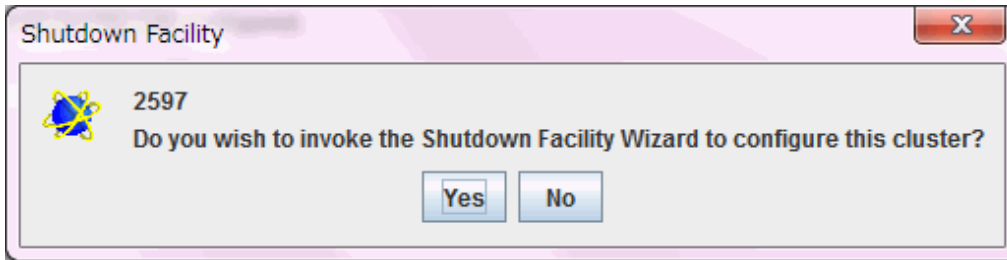


When the CF wizard is executed on the unconfigured node, it requests the CF driver to push CF modules to each Ethernet devices on the system. By this process, CF executes ping of each interface. This will enable the CF wizard to detect the network topology.

However, this unloading process may fail. To solve this problem, unload the driver on the error node, and then load the driver again. The unloading process is simplified by using GUI. See "[4.6 Starting and stopping CF](#)" for details.

When the window is closed by clicking the Finish button, following pop-up window is displayed.

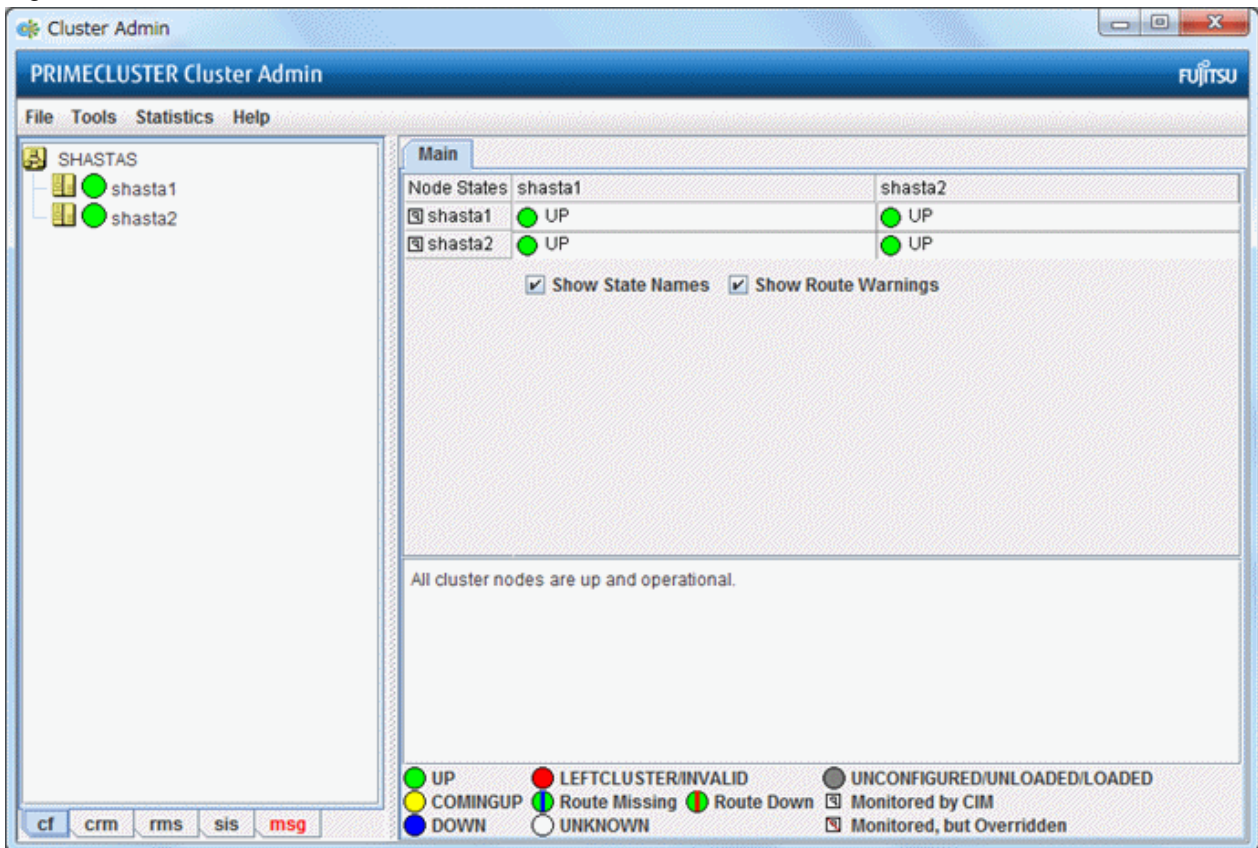
Figure 1.23 SF wizard start notification pop-up



Click on the No button to close the pop-up window.

When the CF Wizard is completed, the main window of Cluster Admin will be displayed as below. After a few seconds, the latest information of the CF's configuration and status will be updated and displayed on it.

Figure 1.24 Main CF window



1.1.6 Adding a new node to CF

This section describes how to add a node to an existing CF cluster.

The first step is to make sure that Web-Based Admin View is properly configured on the new node. Refer to "PRIMECLUSTER Web-Based Admin View Operation Guide" for additional details on Web-Based Admin View configuration options.

After you have properly configured Web-Based Admin on the new node, you should start Cluster Admin. If you are already running the Cluster Admin GUI, exit it and then restart it.

The first window that Cluster Admin displays is the small "Figure 1.6 Initial connection pop-up" window. This window lists all of the nodes which are known to Web-Based Admin View. If the new node is not present in this list, then you should recheck your Web-Based Admin configuration and also verify that the new node is up.

To add the new node, select it in the initial connection pop-up. After making your selection, run the CF Wizard by clicking on the *Configure* button (see "Figure 1.8 CF loaded but not configured") The CF Wizard will appear, and you can use it to join the existing CF cluster.

The CF Wizard will allow you to configure CF, CIM, and CIP on the new node. After it is run, you should configure the Shutdown Facility on the new node.

You will also need to do additional configuration work for other PRIMECLUSTER products you might be using such as the Cluster Resource Manager (CRM), RMS, Global Disk Services (hereinafter GDS),GFS, and so forth.

1.1.7 Example of CF configuration by CLI

When configuring CF by CLI, perform the following steps.

In this section, the cluster system configured with two nodes where the CF node names are "fuji2" and "fuji3", is explained as an example.

1. Create CIP configuration files.

Specify /etc/cip.cf as below on all the nodes which configure the cluster system.

Example:

```
fuji2      fuji2RMS:netmask:255.255.255.0
fuji3      fuji3RMS:netmask:255.255.255.0
```



If you manually create /etc/cip.cf, you cannot reconfigure CF by Cluster Admin. To reconfigure CF by Cluster Admin, delete the /etc/cip.cf file beforehand.

2. Set IP addresses.

Specify /etc/hosts as below on all the nodes which configure the cluster system.

Example:

```
<cip address1>      fuji2RMS
<cip address2>      fuji3RMS
```

3. Enable remote access by using cfcf/cfsh.

Specify /etc/default/cluster.config as below on all the nodes which configure the cluster system.

```
CFCF "cfcf"
CFSH "cfsh"
```

4. Edit /etc/default/cluster on all the nodes.

- a. Edit /etc/default/cluster to create the files of the following contents:

```
nodename <CF node name>
clustername <Cluster name>
device <Cluster interconnect 1>
device <Cluster interconnect 2>
```

Example:

```
nodename fuji2
clustername PRIMECLUSTER1
device eth2
device eth3
```



Make sure that the node name to be defined in nodename is the CF node name, not the node name of the OS.

- b. Set the owner, group, and access permission.

```
# chown root:root /etc/default/cluster
# chmod 600 /etc/default/cluster
```

c. Reboot the nodes.

5. Execute the following command with any node in the cluster system and set the Cluster Integrity Monitor (CIM).

```
# rcqconfig -a <nodename> ...
```

nodename : CF node name

Example:

```
# rcqconfig -a fuji2 fuji3
```

If this command fails, check again that CF node names and cluster names configured in /etc/default/cluster in step 4 are correct.

6. Check that it can be communicated with the RMS node name.

Example: When checking from fuji2

```
# ping fuji3RMS
```

If it cannot be communicated, check again that CF node names, RMS node names, and CIP addresses configured in /etc/cip.cf and /etc/hosts in step 1 and step 2 are correct.

1.2 CIP configuration file

The CIP configuration file is stored in /etc/cip.cf on each node in the cluster. Normally, you can use the GUI to create this file during cluster configuration time. However, there may be times when you wish to manually edit this file.

The format of a CIP configuration file entry is as follows:

```
cfname CIP_Interface_Info [ CIP_Interface_Info... ] [IPv6]
```

- *cfname* tells what node the configuration information is for.
- *CIP_Interface_Info* gives information needed to configure a single CIP interface.

Normally, the configuration information of all the CIP interfaces on all the nodes is contained in the cip.cf configuration file.

- For IPv4, specify *CIP_Interface_Info* with the following format:

```
IPv4-Address[:netmask:<Netmask>]
```

Specify it without any spaces even around colons.

For IPv4-Address, specify as a number in Internet standard dotted-decimal notation or as the Host name. When specifying with the Host name, it needs to be defined in /etc/hosts.

For <Netmask>, specify the netmask value to be set to the IP address as a number in Internet standard dotted-decimal notation.

- For IPv6, specify *CIP_Interface_Info* with the following format:

```
Hostname: " [ IPv6-Address/prefix_length ] "
```

Specify it without any spaces around colons, slashes, and inside of each brackets "[", "]".

For Hostname, describe the Host name to specify the cip address.

For IPv6-Address and *prefix_length*, specify the IPv6 address and the prefix length denoted as a hexadecimal code which is separated by Internet standard colons.

- When using the IPv6 address, specify "IPv6" in the end of the line.

For example, the CIP configuration done in Section "1.1.5 Example of creating a cluster" would produce the following CIP configuration file:

```
fuji2    fuji2RMS:netmask:255.255.255.0
fuji3    fuji3RMS:netmask:255.255.255.0
```

Although not shown in this example, the CIP syntax does allow multiple CIP interfaces for a node to be defined on a single line. The `cip.cf` manual page has more details about the `cip.cf` file.

If you make changes to the `cip.cf` file by hand, you should be sure that the file exists on all the nodes, and all the nodes are specified in the file. Be sure to update all the nodes in the cluster with the new file. Changes to the CIP configuration file will not take effect until CIP is stopped and restarted.

After stopping all applications that use CIP, restart CIP by stopping and starting CF.

For instructions on starting and stopping CF, see Section "[4.6 Starting and stopping CF](#)".

Chapter 2 CF Registry and Integrity Monitor

This chapter discusses the CF registry (CFREG) and the Cluster Integrity Monitor (CIM).

2.1 CF Registry (CFREG)

The CFREG provides a set of CF base product services that allows cluster applications to maintain cluster global data that must be consistent on all of the nodes in the cluster and must live through a clusterwide reboot.

Typical applications include cluster-aware configuration utilities that require the same configuration data to be present and consistent on all of the nodes in a cluster (for example, cluster volume management configuration data).

The data is maintained as named registry entries residing in a data file where each node in the cluster has a copy of the data file. The services will maintain the consistency of the data file throughout the cluster.

A user-level daemon (cfregd), runs on each node in the cluster, and is responsible for keeping the data file on the node where it is running synchronized with the rest of the cluster. The cfregd process will be the only process that ever modifies the data file. Only one synchronization daemon process will be allowed to run at a time on a node. If a daemon is started with an existing daemon running on the node, the started daemon will log messages that state that a daemon is already running and terminate itself. In such a case, all execution arguments for the second daemon will be ignored.

2.2 Cluster Integrity Monitor(CIM)

The purpose of the CIM is to allow applications to determine when it is safe to perform operations on shared resources. It is safe to perform operations on shared resources when a node is a member of a cluster that is in a consistent state.

A consistent state is means that all the nodes of a cluster that are members of the CIM set are in a known and safe state. The nodes that are members of the CIM set are specified in the CIM configuration. Only these nodes are considered when the CIM determines the state of the cluster. When a node first joins or forms a cluster, the CIM indicates that the cluster is consistent only if it can determine the status of the other nodes that make up the CIM set and that those nodes are in a safe state.

As methods for guaranteeing a quorum, PRIMECLUSTER supports the following three CIM methods:

- Node State Management (NSM) method
- MMB method

The NSM method uses the state of the other node as reported by CF and periodically determines whether that node is in a known state. If the state is known, The CF states that are determined as known are UP and DOWN. If the node is being activated or is in LEFTCLUSTER state, that state is determined to be unknown (a quorum is not guaranteed).

The MMB method uses the PRIMEQUEST MMB interface and asynchronously determines the state of a cluster node.

In this way, the CIM provides applications with a set of functions that determine nodes are in a consistent state (whether or not there is a quorum).

PRIMECLUSTER uses CIM to determine whether a user application process that uses resources shared by multiple nodes in a cluster can be processed safely without triggering process contention. In other words, it is safe to perform operations on shared resources if the node executing the process is a member of a cluster system that is in a consistent state (quorum). In the PRIMECLUSTER system, the consistent state is set when all the nodes in the cluster system monitored by the CIM are either in operating (UP) or stopped (DOWN) state and are also in a safe state. The nodes monitored by the CIM to all the nodes that were set when the CIM was configured. Only these nodes are considered when the CIM checks the cluster status. When a node first joins or forms a cluster, the CIM indicates that the cluster is consistent only if it can determine the status of the other nodes that make up the CIM set and that those nodes are in a safe state.

CIM currently supports the Node State Management (NSM) method. The CIM reports on a cluster state that a node state is known and safe (True), or a node state is unknown (False) for the node. True and False are defined as follows:

True - All CIM nodes in the cluster are in a known and safe state.

False - One or more CIM nodes in the cluster are in an unknown or unsafe state.

2.2.1 Configuring CIM

You can perform CIM procedures through the following methods:

- Cluster Admin GUI

This is the preferred method of operation. Refer to the Section "Adding and removing a node from CIM" for the GUI procedures.

- CLI

Refer to the Chapter "Manual pages" for complete details on the CLI options and arguments. The commands can be found in the following directory:

```
/opt/SMAW/SMAWcf/bin
```

CLI

The CIM is configured using the command `rcqconfig(1M)` after CF starts. The `rcqconfig(1M)` command is used to set up or to change the CIM configuration. You only need to run this command if you are not using Cluster Admin to configure CIM.

When `rcqconfig(1M)` is invoked, it checks that the node is part of the cluster. When the `rcqconfig(1M)` command is invoked without any option, it checks if any configuration is present in the CFReg.database after the node joins the cluster. If there is none, it returns as error. When you are using the GUI, these actions are done as part of the configuration process.

`rcqconfig(1M)` configures a quorum set of nodes, among which CF decides the quorum state. `rcqconfig(1M)` is also used to show the current configuration. If `rcqconfig(1M)` is invoked without any configuration changes or with only the `-v` option, `rcqconfig(1M)` will apply any existing configuration to all the nodes in the cluster. It will then start or restart the quorum operation. `rcqconfig(1M)` can be invoked from the command line to configure or to start the quorum.

2.2.2 Query of the quorum state

CIM recalculates the quorum state when it is triggered by some node state change. However you can force the CIM to recalculate it by running `rcquery(1M)` at any time. Refer to the Chapter "Manual pages" for complete details on the CLI options and arguments.

`rcquery(1M)` functions as follows:

- Queries the state of quorum and gives the result using the return code. It also gives you readable results if the verbose option is given.
- Returns True if the states of all the nodes in the quorum set are known. If the state of any node is unknown, then it returns False.
- Exits with a status of zero when a quorum exists, and it exits with a status of 1 when a quorum does not exist. If an error occurs during the operation, then it exits with any other non-zero value other than 1.

2.2.3 Reconfiguring quorum

Refer to the Section "Adding and removing a node from CIM" for the GUI procedures.

CLI

The configuration can be changed at any time and is effective immediately. When a new node is added to the quorum set of nodes, the node being added must be part of the cluster so as to guarantee that the new node also has the same quorum configuration. Removing a node from the quorum set can be done without restriction.

When the configuration information is given to the command `rcqconfig(1M)` as arguments, it performs the transaction to CFREG to update the configuration information. Until CIM is successfully configured and gets the initial state of the quorum, CIM has to respond with the quorum state of False to all queries.



Example

In this example the cluster has formed but no quorum was established.

- Display the states of all the nodes in the cluster as follows:

```
fuji2# cftool -n
Node   Number  State  Os      Cpu
fuji2   1        UP     Linux   Pentium
fuji3   2        UP     Linux   Pentium
```

- Display the current quorum configuration as follows:

```
fuji2# rcqconfig -g
```

Nothing is displayed, since no nodes have been added so far.

- Add new nodes in a quorum set of nodes as follows:

```
fuji2# rcqconfig -a fuji2 fuji3
```

- Display the current quorum configuration parameters as follows:

```
fuji2# rcqconfig -g
QUORUM_NODE_LIST= fuji2 fuji3
```

- Delete nodes from a quorum set of nodes as follows:

```
fuji2# rcqconfig -d fuji2
```

- Display the current quorum configuration parameters after one node is deleted as follows:

```
fuji2# rcqconfig -g
QUORUM_NODE_LIST= fuji3
```

- Add a new node, fuji11 (which is not in the cluster), in a quorum set of nodes as follows:

```
fuji2# rcqconfig -a fuji2 fuji3 fuji11
Cannot add node fuji11 that is not up.
```

- Since CF only configured the cluster to consist of fuji2 and fuji3, fuji11 does not exist. The quorum set remains unchanged.

```
fuji2# rcqconfig -g
QUORUM_NODE_LIST= fuji3
```

.....

Chapter 3 Cluster resource management

This chapter discusses the Resource Database, which is a synchronized clusterwide database.

The cluster Resource Database is intended to be used only by PRIMECLUSTER products. It is not a general purpose database which a customer could use for their own applications.

3.1 Resource Database configuration

This section discusses how to set up the Resource Database for the first time on a new cluster. The following procedure assumes that the Resource Database has not previously been configured on any of the nodes in the cluster.

Before you begin configuring the Resource Database, you must first make sure that CIP is properly configured on all the nodes. The Resource Database uses CIP for communicating between nodes, so it is essential that CIP is working.

The Resource Database also uses the CIP configuration file `/etc/cip.cf` to establish the mapping between the CF node name and the CIP name for a node. If a particular node has multiple CIP interfaces, then only the first one is used. This will correspond to the first CIP entry for a node in `/etc/cip.cf`. It will also correspond to `cip0` on the node itself.

Because the Resource Database uses `/etc/cip.cf` to map between CF and CIP names, it is critical that this file be the same on all the nodes. If you used the Cluster Admin CF Wizard to configure CIP, then this will already be the case. If you created some `/etc/cip.cf` files by hand, then you need to make sure that all the nodes are specified and they are the same across the cluster.

In general, the CIP configuration is fairly simple. You can use the Cluster Admin CF Wizard to configure a CIP subnet after you have configured CF. If you use the Wizard, then you will not need to do any additional CIP configuration. See the Section "[1.1 CF, CIP, and CIM configuration](#)" for more details.

After CIP has been configured, you can configure the Resource Database on a new cluster by using the following procedure. This procedure must be done on all the nodes in the cluster.

1. Log in to the node with system administrator authority.
2. Verify that the node can communicate with other nodes in the cluster over CIP.

To test CIP network connectivity, execute the `ping(1M)` command or the `ping6(8)` command (when using the IPv6 address). The file `/etc/cip.cf` contains the CIP names that you should use in the `ping(1M)` command or the `ping6(8)` command.

If you are using RMS and you have only defined a single CIP subnetwork, then the CIP names will be of the following form:

`cfnameRMS`

For example, if you have two nodes in your cluster named `fuji2` and `fuji3`, then the CIP names for RMS would be `fuji2RMS` and `fuji3RMS`, respectively. You could then run the following commands:

```
fuji2# ping fuji3RMS
fuji3# ping fuji2RMS
```

This tests the CIP connectivity.

3. Execute the `clsetup` command. When used for the first time to set up the Resource Database on a node, it is called without any arguments as follows:

```
# /etc/opt/FJSVcluster/bin/clsetup
```

4. Execute the `clgettree` command to verify that the Resource Database was successfully configured on the node, as shown in the following:

```
# /etc/opt/FJSVcluster/bin/clgettree
```

The command should complete without producing any error messages, and you should see the Resource Database configuration displayed in a tree format.

For example, on a two-node cluster consisting of `fuji2` and `fuji3`, the `clgettree` command might produce output similar to the following:

```
Cluster 1 cluster
  Domain 2 Domain0
    Shared 7 SHD_Domain0
      Node 3 fuji2 UNKNOWN
      Node 5 fuji3 UNKNOWN
```

If you need to change the CIP configuration to fix the problem, you will also need to run the *clinitreset* command and start the information process over.

The format of *clgettree* is more fully described in its manual page. For the purpose of setting up the cluster, you need to check the following:

- Each node in the cluster should be referenced in a line that begins with the word Node.
- The *clgettree* output must be identical on all the nodes.

If either of the above conditions is not met, then it is possible that you may have an error in the CIP configuration. Double-check the CIP configuration using the methods described earlier in this section. The actual steps are as follows:

1. Make sure that CIP is properly configured and running.
2. Run *clinitreset* on all the nodes in the cluster.

```
# /etc/opt/FJSVcluster/bin/clinitreset
```

3. Reboot each node.
4. Rerun the *clsetup* command on each node.

```
# /etc/opt/FJSVcluster/bin/clsetup
```

5. Use the *clgettree* command to verify the configuration.

```
# /etc/opt/FJSVcluster/bin/clgettree
```

3.2 Startup synchronization

A copy of the Resource Database is stored locally on each node in the cluster. When the cluster is up and running, all of the local copies are kept in sync. However, if a node is taken down for maintenance, then its copy of the Resource Database may be out of date by the time it rejoins the cluster. Normally, this is not a problem. When a node joins a running cluster, then its copy of the Resource Database is automatically downloaded from the running cluster. Any stale data that it may have had is thus overwritten.

There is one potential problem. Suppose that the entire cluster is taken down before the node with the stale data had a chance to rejoin the cluster. Then suppose that all the nodes are brought back up again. If the node with the stale data comes up long before any of the other nodes, then its copy of the Resource Database will become the master copy used by all the nodes when they eventually join the cluster.

To avoid this situation, the Resource Database implements a startup synchronization procedure. If the Resource Database is not fully up and running anywhere in the cluster, then starting the Resource Database on a node will cause that node to enter into a synchronization phase. The node will wait up to *StartingWaitTime* seconds for other nodes to try to bring up their own copies of the Resource Database. During this period, the nodes will negotiate among themselves to see which one has the latest copy of the Resource Database. The synchronization phase ends when either all the nodes have been accounted for or *StartingWaitTime* seconds have passed. After the synchronization period ends, the latest copy of the Resource Database that was found during the negotiations will be used as the master copy for the entire cluster.

The default value for *StartingWaitTime* is 60 seconds.

This synchronization method is intended to cover the case where all the nodes in a cluster are down, and then they are all rebooted together. For example, some businesses require high availability during normal business hours, but power their nodes down at night to reduce their electric bill. The nodes are then powered up shortly before the start of the working day. Since the boot time for each node may vary slightly, the synchronization period of up to *StartingWaitTime* ensures that the latest copy of the Resource Database among all of the booting nodes is used.

Another important scenario in which all the nodes may be booted simultaneously involves the temporary loss and then restoration of power to the lab where the nodes are located.

However, for this scheme to work properly, you must verify that all the nodes in the cluster have boot times that differ by less than StartingWaitTime seconds. Furthermore, you might need to modify the value of StartingWaitTime to a value that is appropriate for your cluster.

Modify the value of StartingWaitTime as follows:

1. Start up all of the nodes in your cluster simultaneously. It is recommended that you start the nodes from a cold power on. Existing nodes are not required to reboot when a new node is added to the cluster.
2. After the each node has come up, look in /var/log/messages for message number 2200. This message is output by the Resource Database when it first starts. For example, enter the following command:

```
# grep 2200 /var/log/messages
Feb 23 19:00:41 fuji2 dcmond[407]: [ID 888197 daemon.notice] FJSVcluster: INFO: DCM: 2200:
Cluster configuration management facility initialization started.
```

Compare the timestamps for the messages on each node and calculate the difference between the fastest and the slowest nodes. This will tell you how long the fastest node has to wait for the slowest node.

3. Check the current value of StartingWaitTime by executing the clsetparam command on any of the nodes. For example, enter the following command:

```
# /etc/opt/FJSVcluster/bin/clsetparam -p StartingWaitTime
60
```

The output for our example shows that StartingWaitTime is set to 60 seconds.

4. If there is a difference in startup times found in Step 2, the StartingWaitTime, or if the two values are relatively close together, then you should increase the StartingWaitTime parameter. You can do this by running the clsetparam command on any one node in the cluster. For example, enter the following command:

```
# /etc/opt/FJSVcluster/bin/clsetparam -p StartingWaitTime 300
```

This sets the StartingWaitTime to 300 seconds.

3.2.1 Startup synchronization and the new node

After the Resource Database has successfully been brought up on the new node, then you need to check if the StartingWaitTime used for startup synchronization is still adequate. If the new node boots much faster or slower than the other nodes, then you may need to adjust the StartingWaitTime time.

Chapter 4 GUI administration

This chapter covers the administration of features in the Cluster Foundation (CF) portion of Cluster Admin.

4.1 Starting Cluster Admin GUI and logging in

The first step is to start Web-based Admin View by entering the following URL in a java-enabled browser:

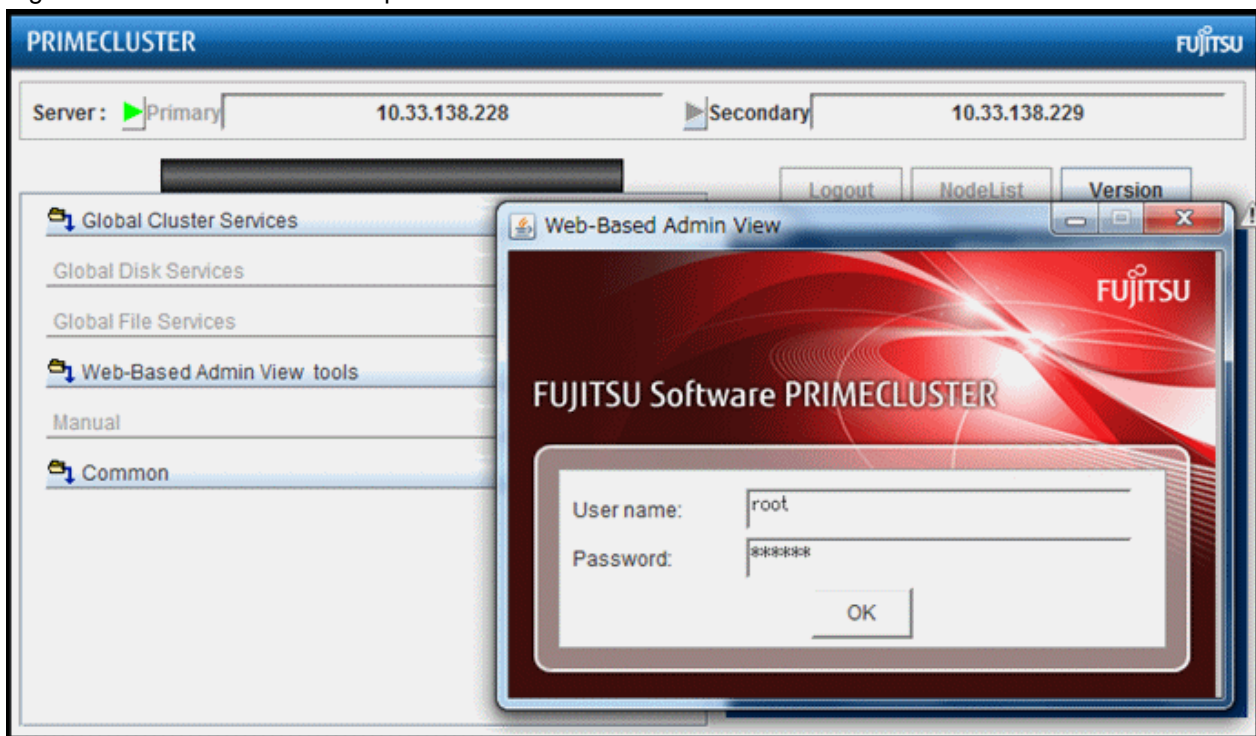
```
http://Management_Server:8081/Plugin.cgi
```

In this example, if fuji2 is a management server, enter the following:

```
http://fuji2:8081/Plugin.cgi
```

This brings up the Web-Based Admin View main window.

Figure 4.1 Cluster Admin start-up window



Enter a user name in the User name field and the password and click on OK.

Use the appropriate privilege level while logging in. There are three privilege levels: root privileges, administrative privileges, and operator privileges.

With the root privileges, you can perform all actions including configuration, administration and viewing tasks. With administrative privileges, you can view as well as execute commands but cannot make configuration changes. With the operator privileges, you can only perform viewing tasks.



Point

In this example we are using root and not creating user groups.

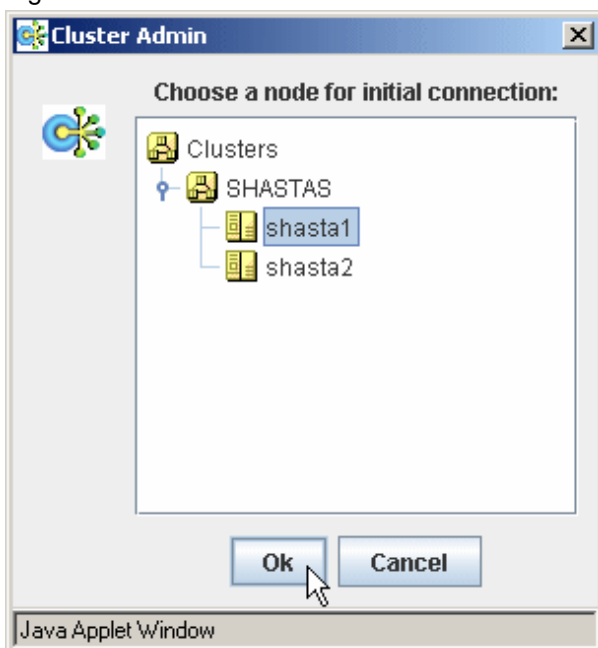
On the "Figure 4.1 Cluster Admin start-up window," click the Global Cluster Services button and the Cluster Admin button appears.

Figure 4.2 Cluster Admin top window



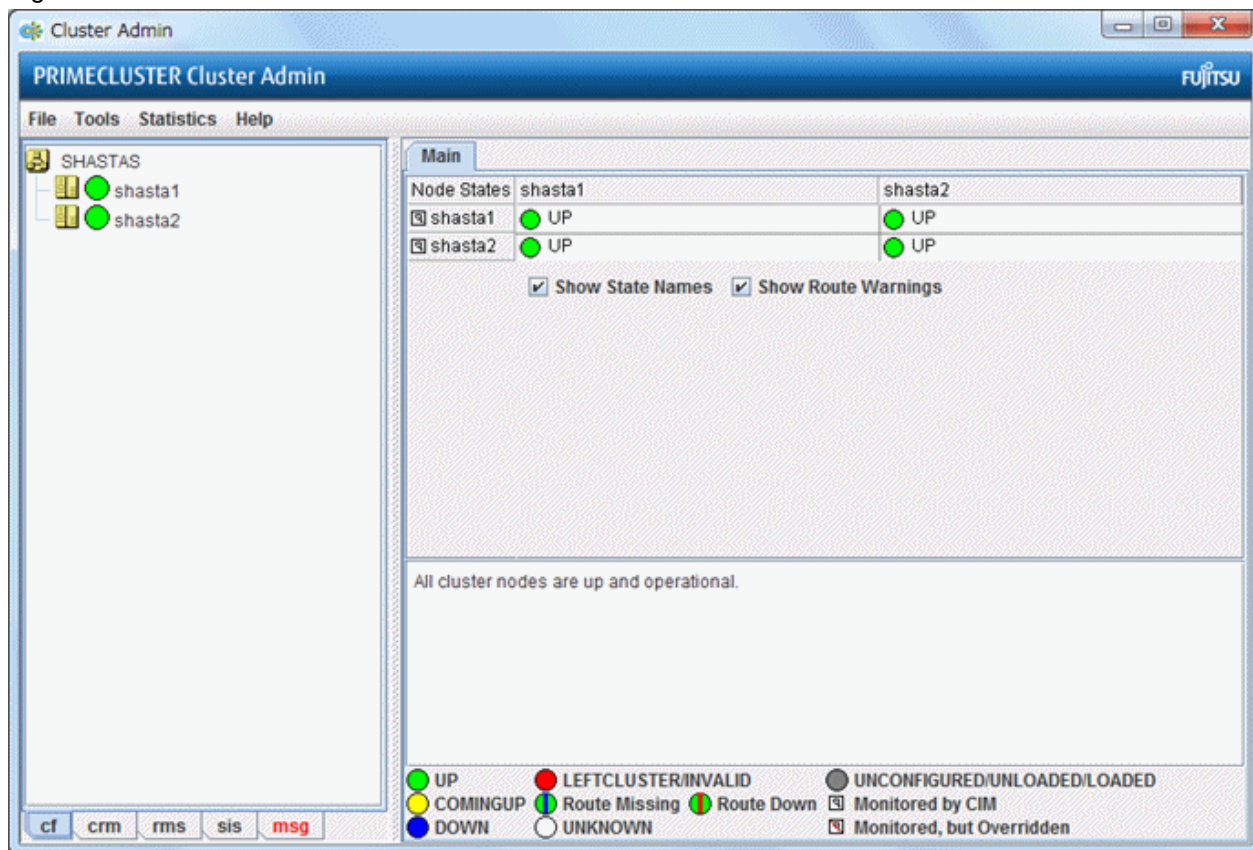
Click the Cluster Admin button. The Choose a node for initial connection window appears.

Figure 4.3 Initial connection choice window



Select a node and click on OK. The Cluster Admin main window appears.

Figure 4.4 Cluster Admin main window



By default, the cf tab is selected and the CF main window is presented. Use the appropriate privilege level while logging in.

4.2 Main CF table

When the GUI is first started, or after the successful completion of the configuration wizard, the main CF table will be displayed in the right panel. A tree showing the cluster nodes will be displayed in the left panel (see "Figure 4.4 Cluster Admin main window.")

The tree displays the local state of each node, but does not give information about how that node considers other nodes. If two or more nodes disagree about the state of a node, one or more colored exclamation marks appear next to the node. Each exclamation mark represents the node state of which another node considers that node to be.

The table in the right panel is called the main CF table. The column on the left of the table lists the CF states of each node of the cluster as seen by the other nodes in the cluster. For instance, the cell in the second row and first column is the state of fuji3 as seen by the node fuji2.




There is an option at the bottom of the table to toggle the display of the state names. This is on by default. If this option is turned off, and there is a large number of nodes in the cluster, the table will display the node names vertically to allow a larger number of nodes to be seen.





There are two types of CF states:

- Local states are the states a node can consider itself in.
- Remote states are the states a node can consider another node to be in.

The following table lists these various states.





Table 4.1 Local states

CF state		Description
UNLOADED		The node does not have a CF driver loaded.
LOADED		The node has a CF driver loaded, but is not running.
COMINGUP		The node is in the process of starting and should be UP soon.

UP		The node is up and running normally.
INVALID		The node has an invalid configuration and must be reconfigured.
UNKNOWN		The GUI has no information from this node. This can be temporary, but if it persists, it probably means the GUI cannot contact that node.
UNCONFIGURED		The CF driver is loaded but the node is not yet configured to run CF.

The following table lists the remote states.

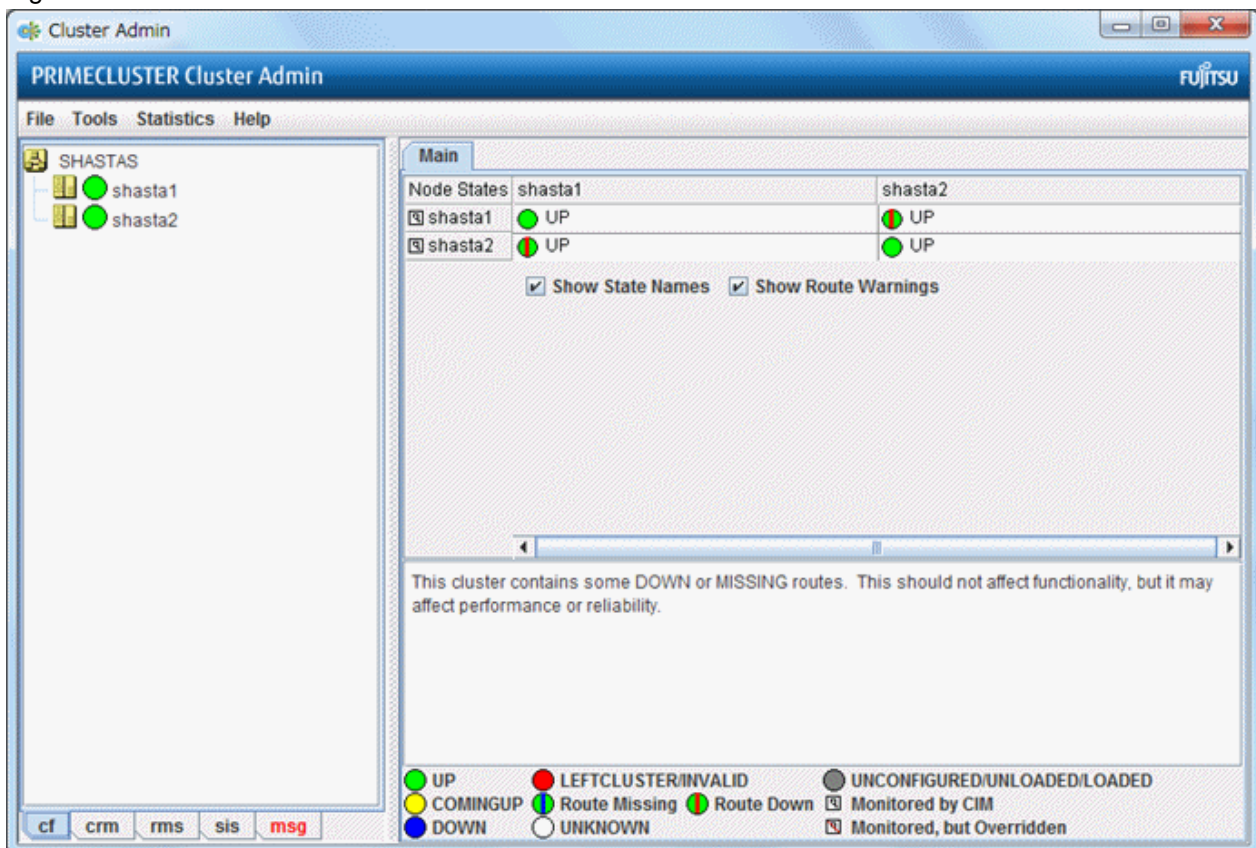
Table 4.2 Remote states

CF state		Description
UP		The node is up and part of this cluster.
DOWN		The node is down and not in the cluster.
UNKNOWN		The reporting node has no opinion on the reported node.
LEFTCLUSTER		The node has left the cluster unexpectedly, probably from a crash. To ensure cluster integrity, it will not be allowed to rejoin until marked DOWN.

4.3 CF route tracking

If a node is UP, but it has one or more DOWN routes, the green circle in the main CF table will have a red line through it (see the following.)

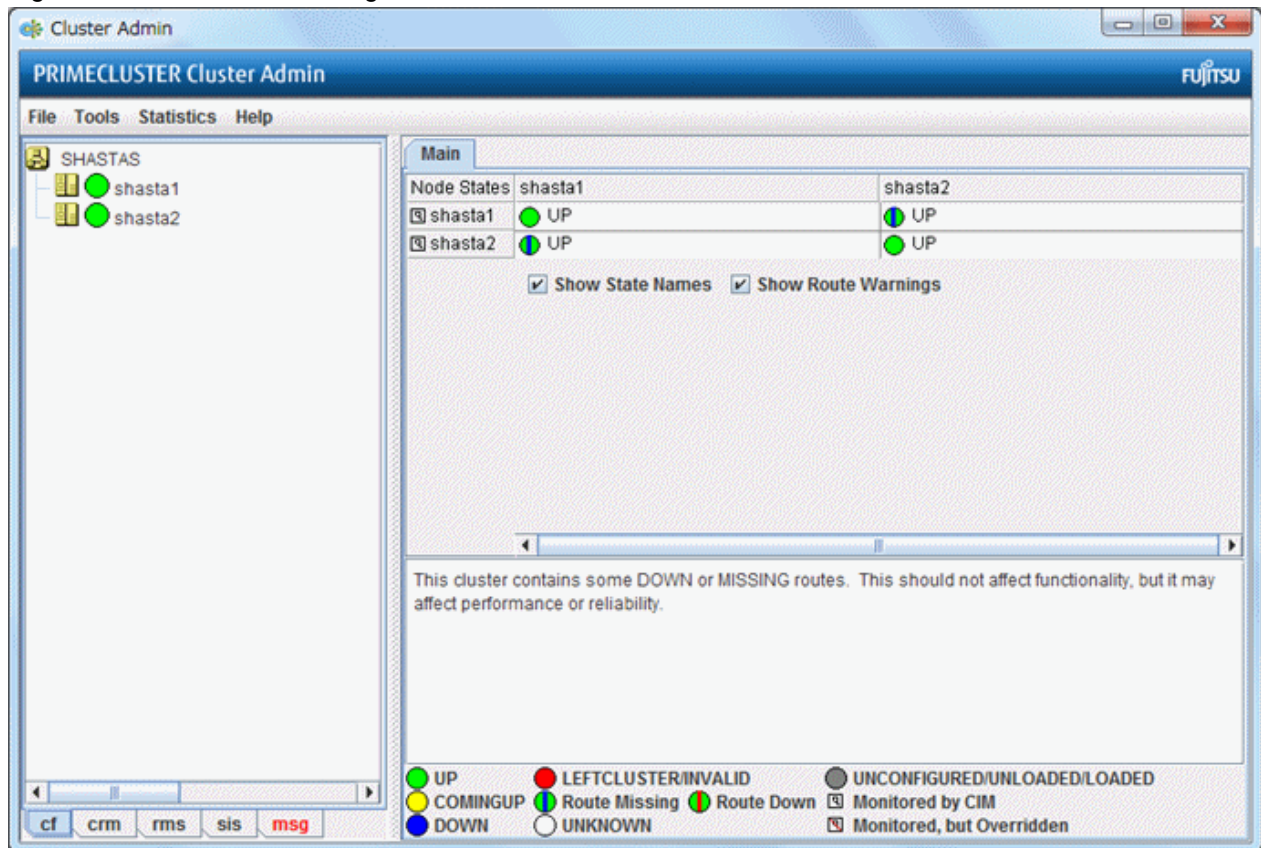
Figure 4.5 CF route DOWN



In this example, one of the network interfaces on fuji2 has been unplugged. Cluster Admin, therefore, shows that a route is DOWN. Since fuji3 cannot contact fuji2 over that interface, it also shows that there is a route down on fuji2. To see which routes are DOWN, click on the node in the left-panel tree and look at the route table.

If CF starts with one or more interfaces missing, then the green circle in the main CF table will have a blue line through it (see the following screen.)

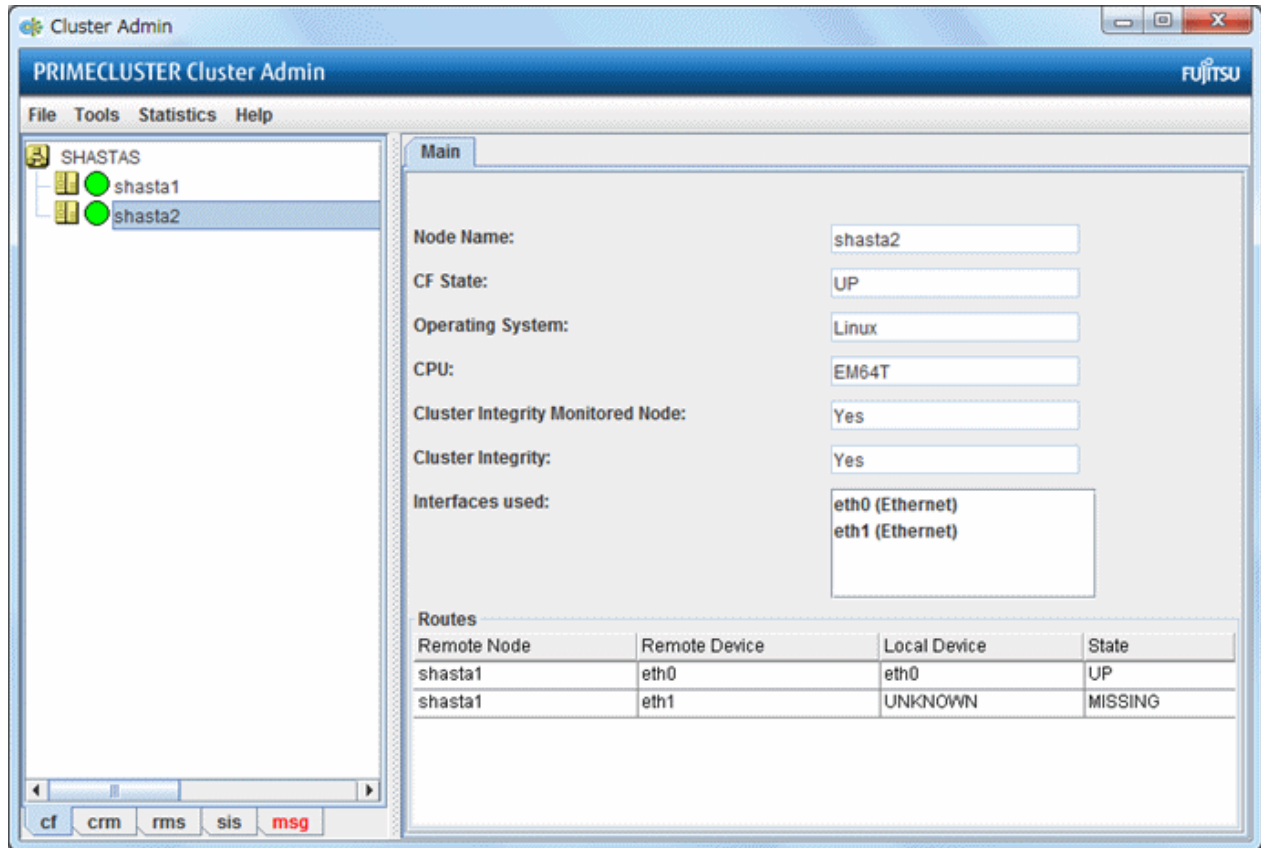
Figure 4.6 CF interface missing



In the above example, fuji3 has a broken connection to fuji2, and Cluster Admin indicates that a route is missing.

In our example, clicking on fuji2 in the left-panel tree shows that there is no route from fuji2 to the eth2 interface on fuji3 (see the following.)

Figure 4.7 CF route table



4.4 Node details

To get detailed information on a cluster node, left-click on the node in the left tree. This replaces the main table with a display of detailed information. (To bring the main table back, left-click on the cluster name in the tree.) This is the example of how the panel is displayed.

Figure 4.8 CF node information

Main

Node Name:

shasta1

CF State:

UP

Operating System:

Linux

CPU:

Pentium

Cluster Integrity Monitored Node:

Yes

Cluster Integrity:

Yes

Interfaces used:

eth0 (Ethernet)
eth2 (Ethernet)

Routes

Remote Node	Remote Device	Local Device	State
shasta2	eth0	eth0	UP
shasta2	eth2	eth2	UP

Shown are the node's name, its CF state(s), operating system, platform, and the interfaces configured for use by CF. The states listed will be all of the states the node is considered to be in. For instance, if the node considers itself UNLOADED and other nodes consider it DOWN, DOWN/UNLOADED will be displayed.

The bottom part of the display is a table of all of the routes being used by CF on this node. It is possible for a node to have routes go down if a network interface or interconnect fails, while the node itself is still accessible.

4.5 Displaying the topology table

To examine and diagnose physical connectivity in the cluster, select

Tools -> Topology. This menu option will produce a display of the physical connections in the cluster. This produces a table with the nodes shown along the left side and the interconnects of the cluster shown along the top. Each cell of the table lists the interfaces on that node connected to the interconnect. There is also a checkbox next to each interface showing if it is being used by CF. This table makes it easy to locate cabling errors or configuration problems at a glance.

This is the example of the topology table.

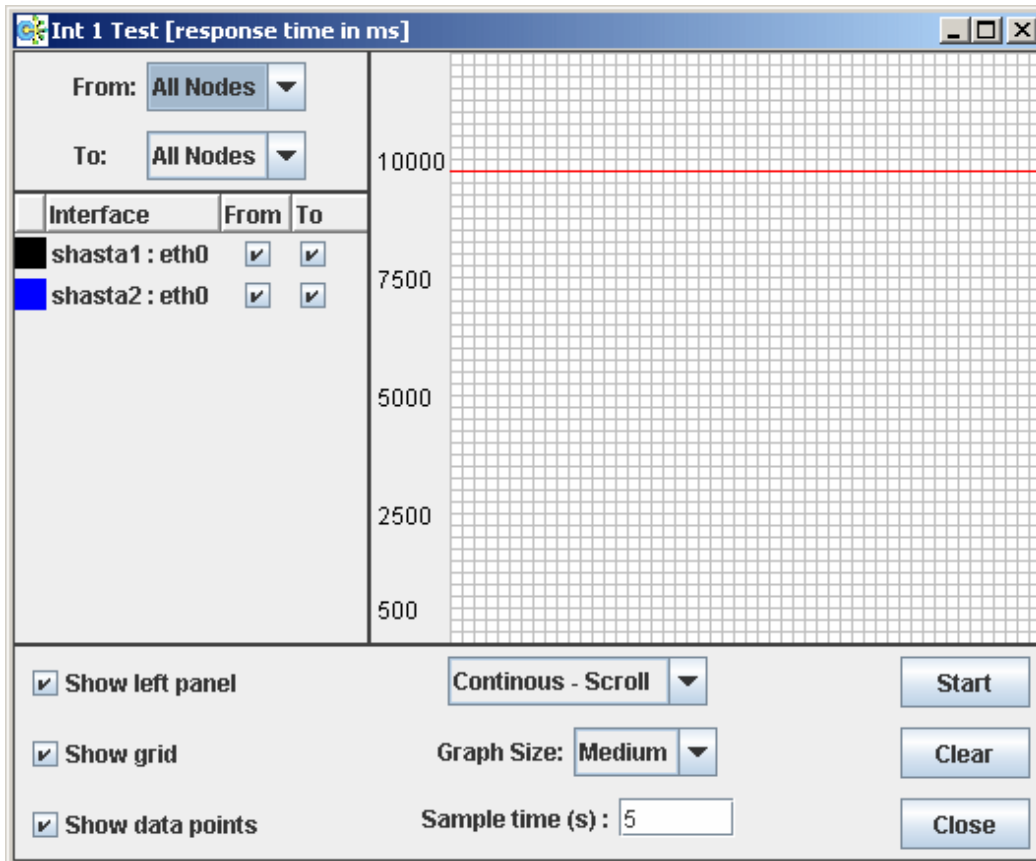
Figure 4.9 CF topology table

SHASTAS: Topology			
SHASTAS	Full Interconnects		
	<input checked="" type="checkbox"/> Int 1 <input type="button" value="Test"/>	<input checked="" type="checkbox"/> Int 2 <input type="button" value="Test"/>	<input checked="" type="checkbox"/> Int 3 <input type="button" value="Test"/>
shasta1	<input type="checkbox"/> eth0	<input checked="" type="checkbox"/> eth1	<input type="checkbox"/> eth2
shasta2	<input type="checkbox"/> eth0	<input checked="" type="checkbox"/> eth1	<input checked="" type="checkbox"/> eth2
<p>This table displays the physical connectivity of the nodes in this cluster. This information is current as of (10:54:58 AM) and will not update. Nodes marked with a * will only show interfaces that are configured.</p>			

Pressing the Test button launches the Response Time monitor.

This tool allows you to see the response time for any combination of two nodes on that interconnect (see the following.).

Figure 4.10 Response Time monitor



The Y axis is the response time for CF pings in milliseconds and the X axis is a configurable period. The red line is the upper limit of the response time before CF will declare nodes to be in the LEFTCLUSTER state.

The controls to the left of the graph determine the nodes for which the graph displays data as follows:

- Set the selection boxes at the top to a specific node name, or to All Nodes.
- Select the check boxes next to the node names to specify specific nodes.

The controls on the left of the bottom panel control how the graphing and information collection is done as follows:

- Check the Show left panel check box to hide the left panel to provide more room for the graph.
- Check the Show grid check box to turn the grid on and off.
- Check the Show data points check box to display a simple line graph.

The controls in the middle of the bottom panel are as follows:

- The top drop-down menu controls how the graph is drawn. The following options are available:
 - Continuous-Scroll - Creates a continuous graph, so that when there are more data points than space, the graph scrolls.
 - Continuous-Clear - Graphs continuously until the graph is full, and then it starts a new graph.
 - Single Graph - Draws a single graph only.
- Graph size - Allows you to control how many data points are drawn.
- Sample time - Controls how often data points are taken.
- The buttons on the lower right control starting and stopping of the graph, clearing it, and closing the graph window.

The buttons on the right of the bottom panel are as follows:

- Start/Stop - Starts or stops the Response Time Monitor.

- Clear - Clears the data and starts a new graph.
- Close - Closes the Response Time Monitor and returns you to the CF Main screen.

Note

The Response Time Monitor is a tool for expert users such as consultants or skilled customers. Its output must be interpreted carefully. The Response Time Monitor uses user-space CF pings to collect its data. If the CF traffic between nodes in a cluster is heavy, then the Response Time Monitor may show slow response times, even if the cluster and the interconnects are working properly. Likewise, if a user does CF pings from the command line while the Response Time Monitor is running, then the data may be skewed.

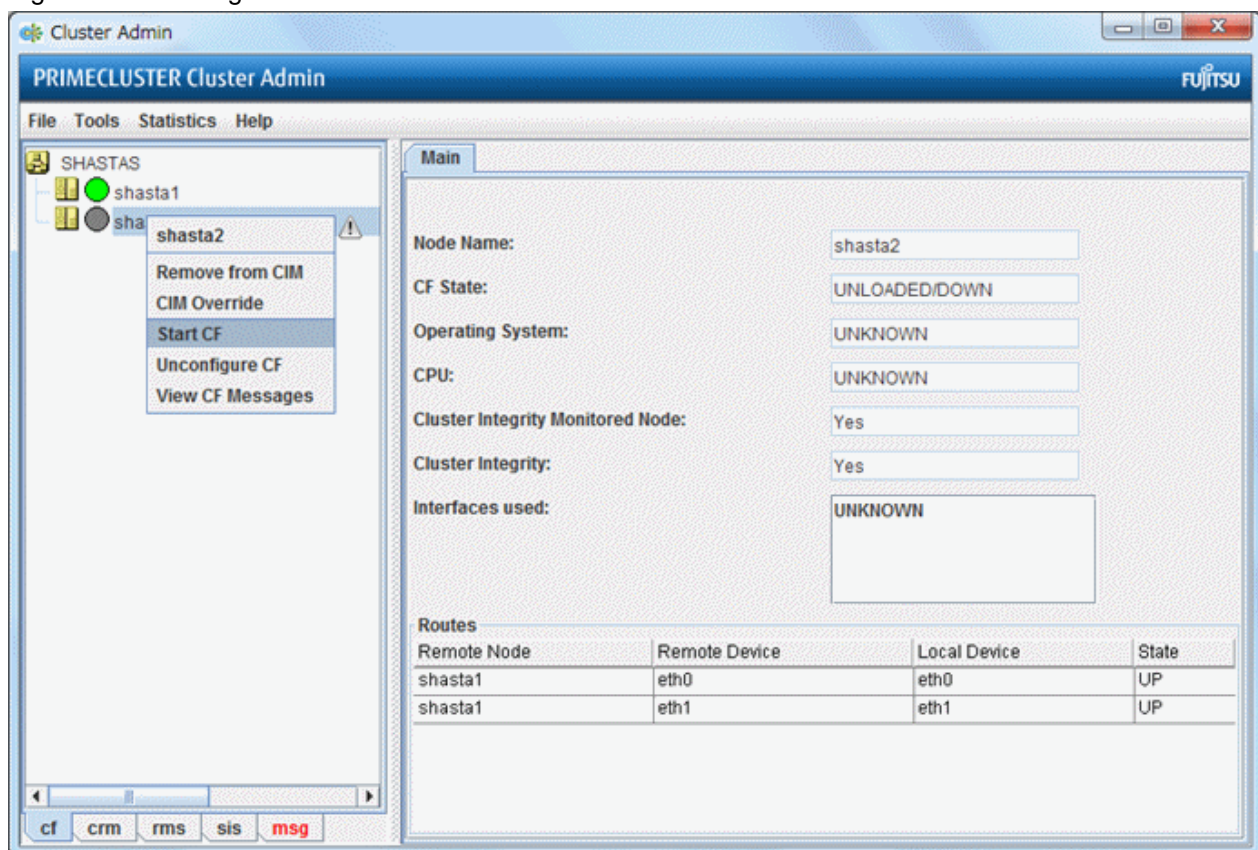
For best results, the Response Time Monitor should be run at times when CF traffic is relatively light, and the CF nodes are only lightly loaded.

4.6 Starting and stopping CF

There are two ways that you can start or stop CF from the GUI. The first is to simply right-click on a particular node in the tree in the left-hand panel. A state sensitive pop-up menu for that node will appear. If CF on the selected node is in a state where it can be started (or stopped), then the menu choice Start CF (or Stop CF) will be offered.

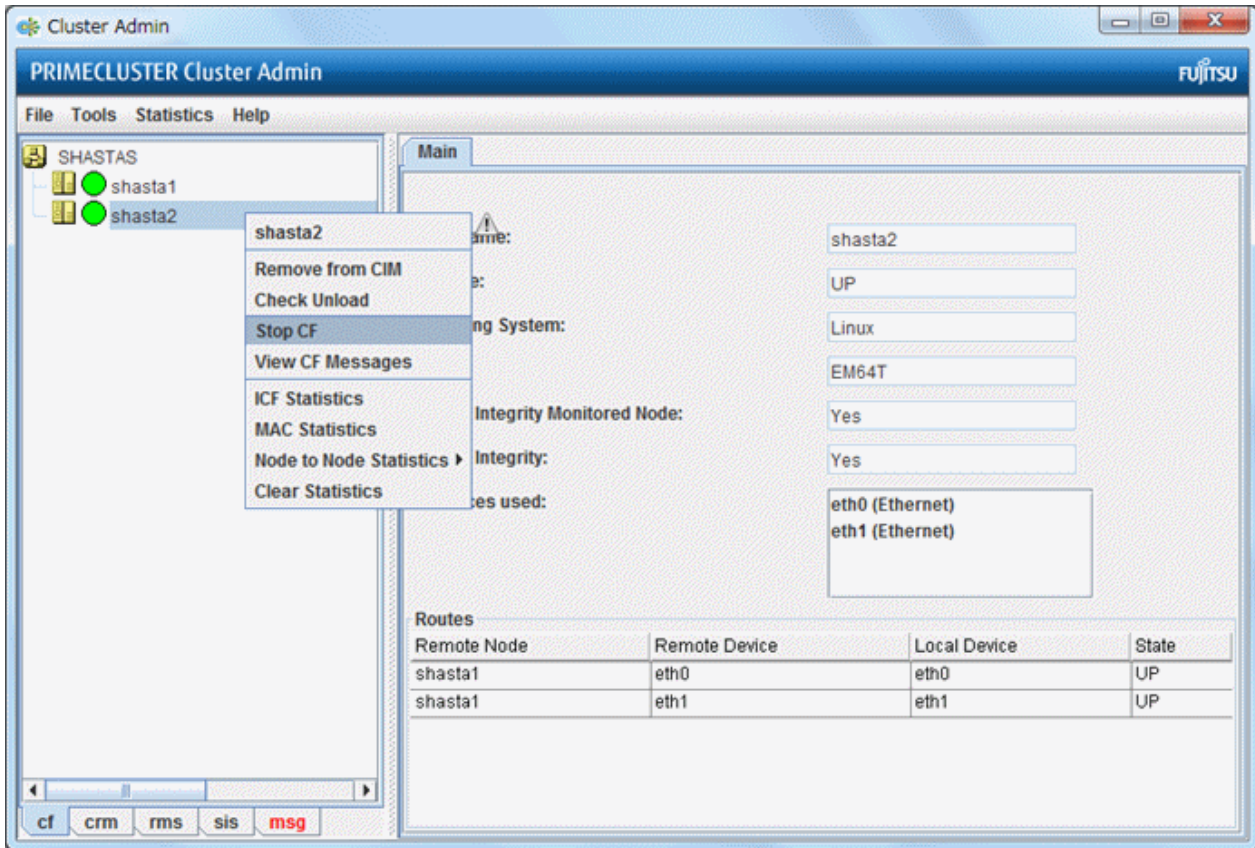
The following screen shows the content-sensitive menu pop-up when you select Start CF.

Figure 4.11 Starting CF



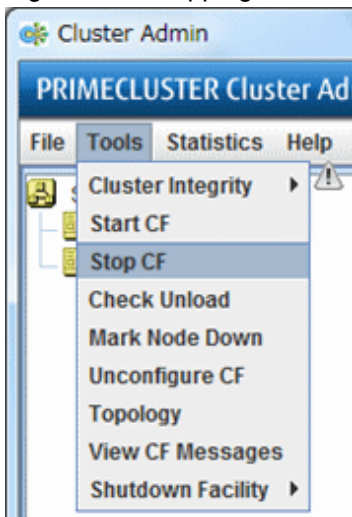
You can also go to the Tools pull-down menu and select either Start CF or Stop CF.

Figure 4.12 Starting and stopping CF from the Tools menu



A pop-up listing all the nodes where CF may be started or stopped will appear (see below.). You can then select the desired node to carry out the appropriate action.

Figure 4.13 Stopping CF and shutting down all the nodes



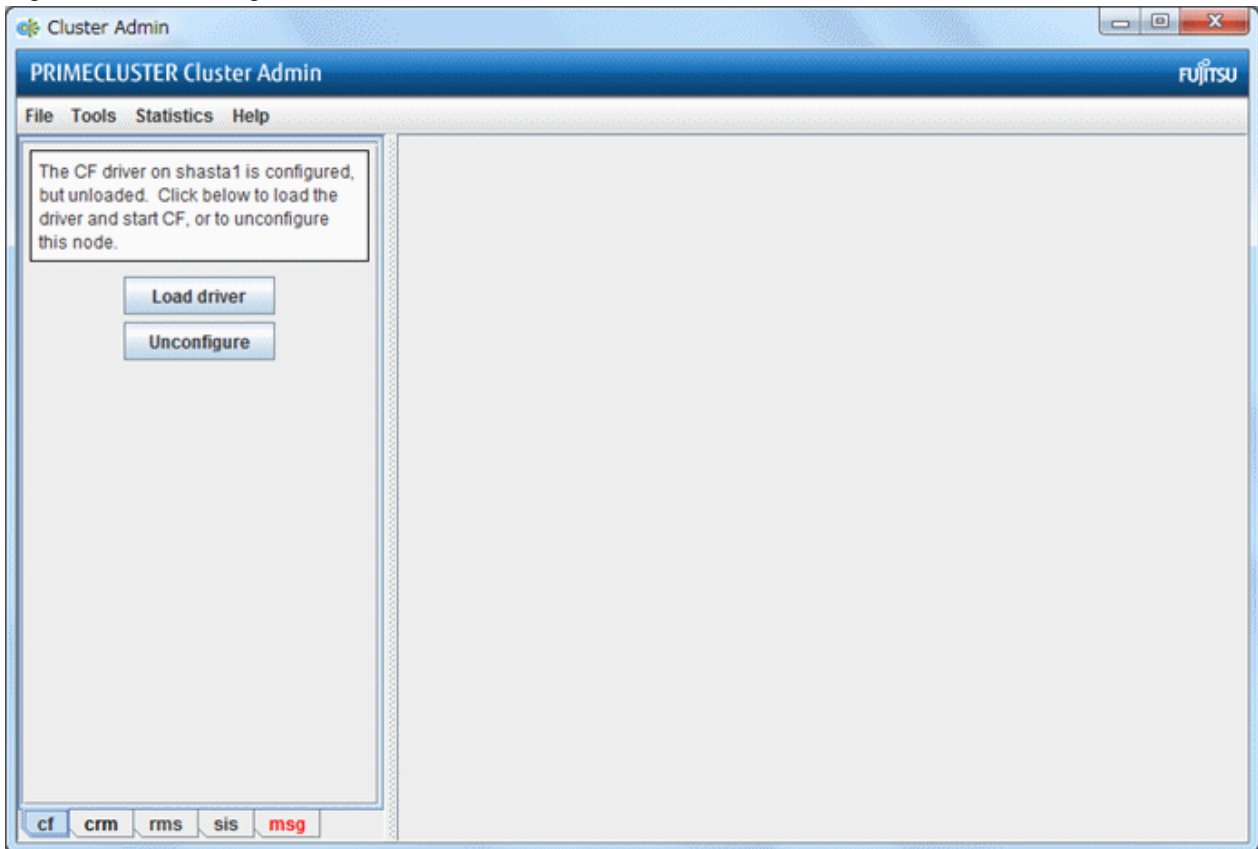
The CF GUI gets its list of CF nodes from the node you selected as the initial connection node that is shown in "[Figure 4.3 Initial connection choice window](#)". If CF is not up and running on the initial connection node, then the CF GUI will not display the list of nodes in the tree in the left panel.

Because of this, if you do not choose the all nodes option, and you want to stop CF on multiple nodes (including the initial node) by means of the GUI, ensure that the initial connection node is the last one on which you stop CF.

4.6.1 Starting CF

If CF is stopped on the initial connection node, the Cluster Admin main window appears with the CF options of Load driver or Unconfigure (see below.) The CF state must be UNLOADED or LOADED to start CF on a node.

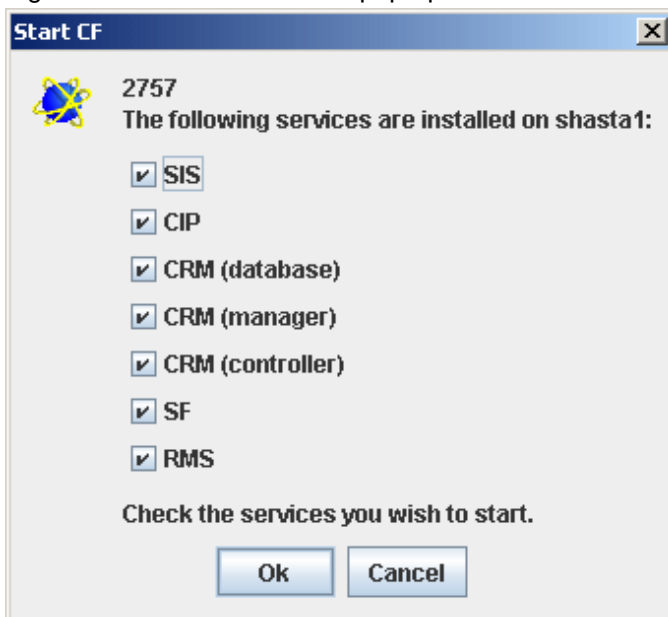
Figure 4.14 CF configured but not loaded



Click on the Load driver button to start the CF driver with the existing configuration.

The Start CF services popup appears (see below.) By default all CF services that have been installed on that node are selected to be started. The contents of this list may vary according to the installed products.

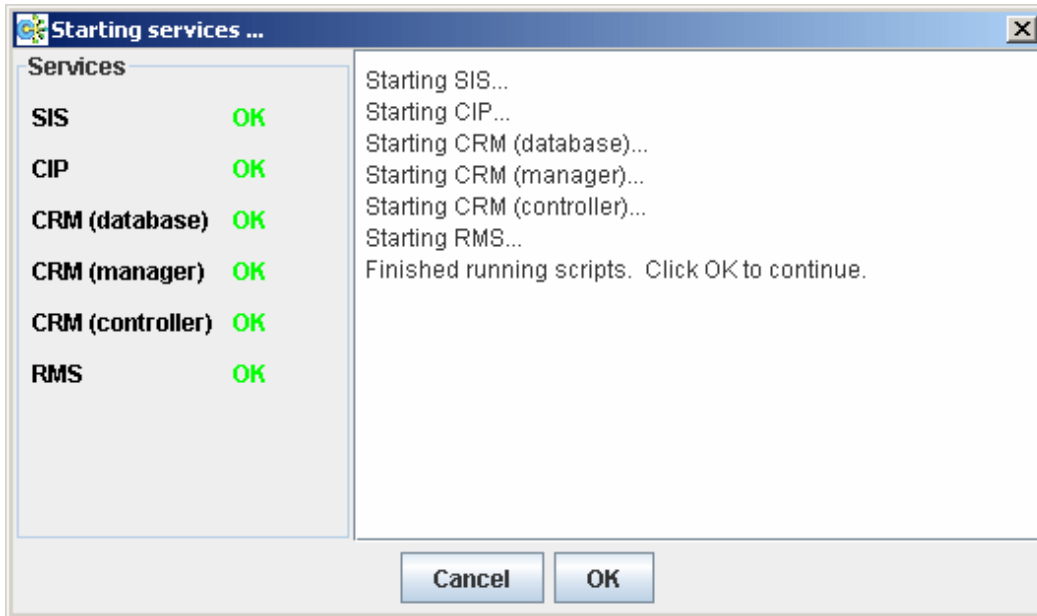
Figure 4.15 Start CF services pop-up



You may exclude CF services from startup by clicking on the selection check box for each service that you do not want to start. This should be done by experts only.

Click on the Ok button and a status popup appears with the results of each service start operation (see below.)

Figure 4.16 Start CF services status window



Click on [OK] to return to the Cluster Admin main window, after all the service startup has finished.

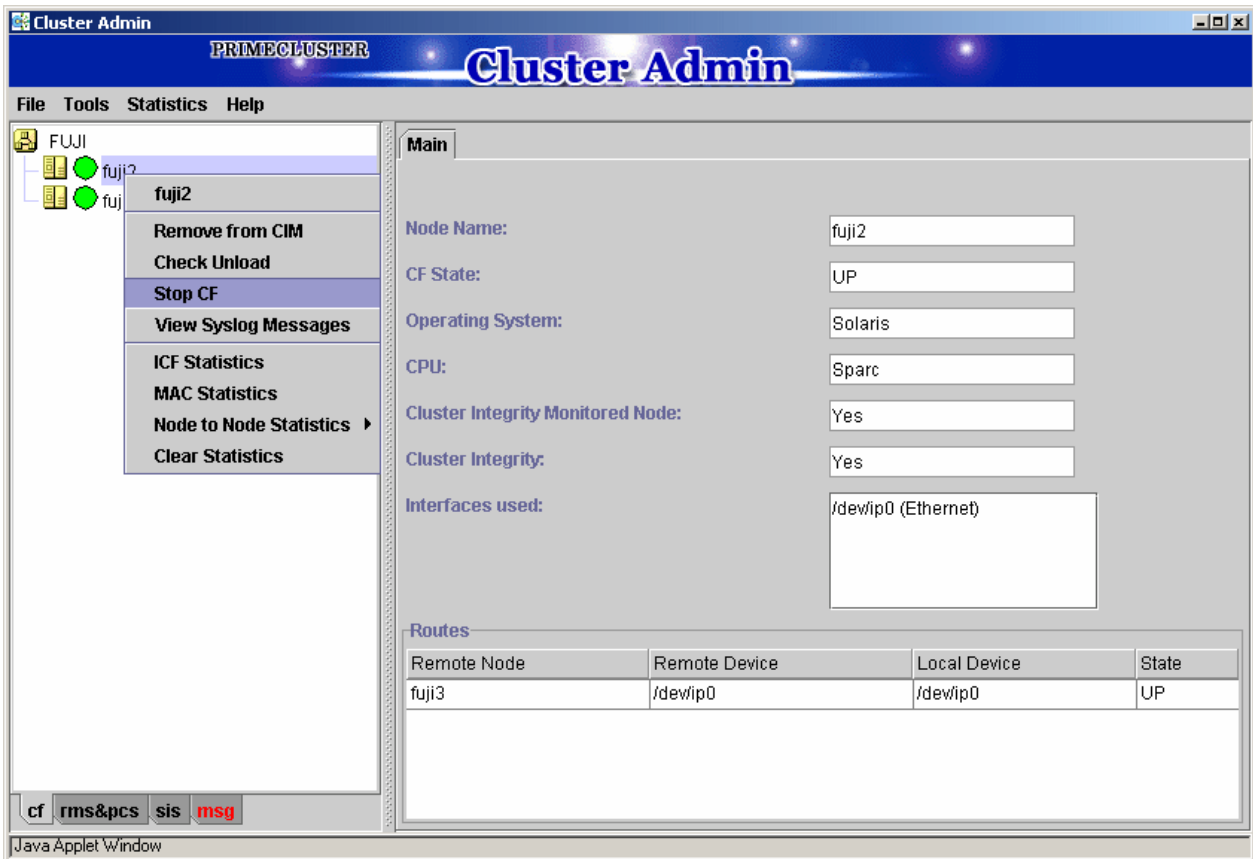
If [Cancel] is clicked, the startup process of the inactivated CF service is canceled.

The already activated CF service will remain in an activated state.

4.6.2 Stopping CF

Right-click on a CF node name and select Stop CF (see below.)

Figure 4.17 Stop CF



A confirmation pop-up appears. Choose Yes to continue.

Figure 4.18 Stopping CF



Before stopping CF, all services that use CF on that node should first be shut down. When you invoke [Stop CF] in the GUI, the GUI checks which services are still running. A list of these services is displayed in the pop-up to ask if you wish to continue the operation.

If you wish to continue the operation, shut down these services.

If the service is canceled, the shutdown processing of CF service that was not shut down will be canceled.

The already shut down CF service will remain shut down.

Note

The dependency scripts currently include only PRIMECLUSTER products. If third-party products, for example Oracle RAC, are using PAS or CF services, then the GUI will not know about them. In such cases, the third-party product should be shut down before you attempt to stop CF.

To stop CF on a node, the node's CF state must be UP, COMINGUP, or INVALID.

4.7 Marking nodes DOWN

If a node is shut down normally, it is considered DOWN by the remaining nodes. If it leaves the cluster unexpectedly, it will be considered LEFTCLUSTER. It is important to mark a node DOWN as SOON as possible to allow normal cluster operation for the remaining nodes. The menu option Tools->Mark Node Down allows nodes to be marked as DOWN.

Note

Marking a node DOWN should be only done if the node is actually down (inoperable or inoperative); otherwise, this could cause data corruption.

To do this, select Tools->Mark Node Down. This displays a dialog of all of the nodes that consider another node to be LEFTCLUSTER. Clicking on one of them displays a list of all the nodes that node considered LEFTCLUSTER. Select one and then click OK. This clears the LEFTCLUSTER status on that node.

Refer to the Chapter "[Chapter 5 LEFTCLUSTER state](#)" for more information on the LEFTCLUSTER state.

4.8 Using PRIMECLUSTER log viewer

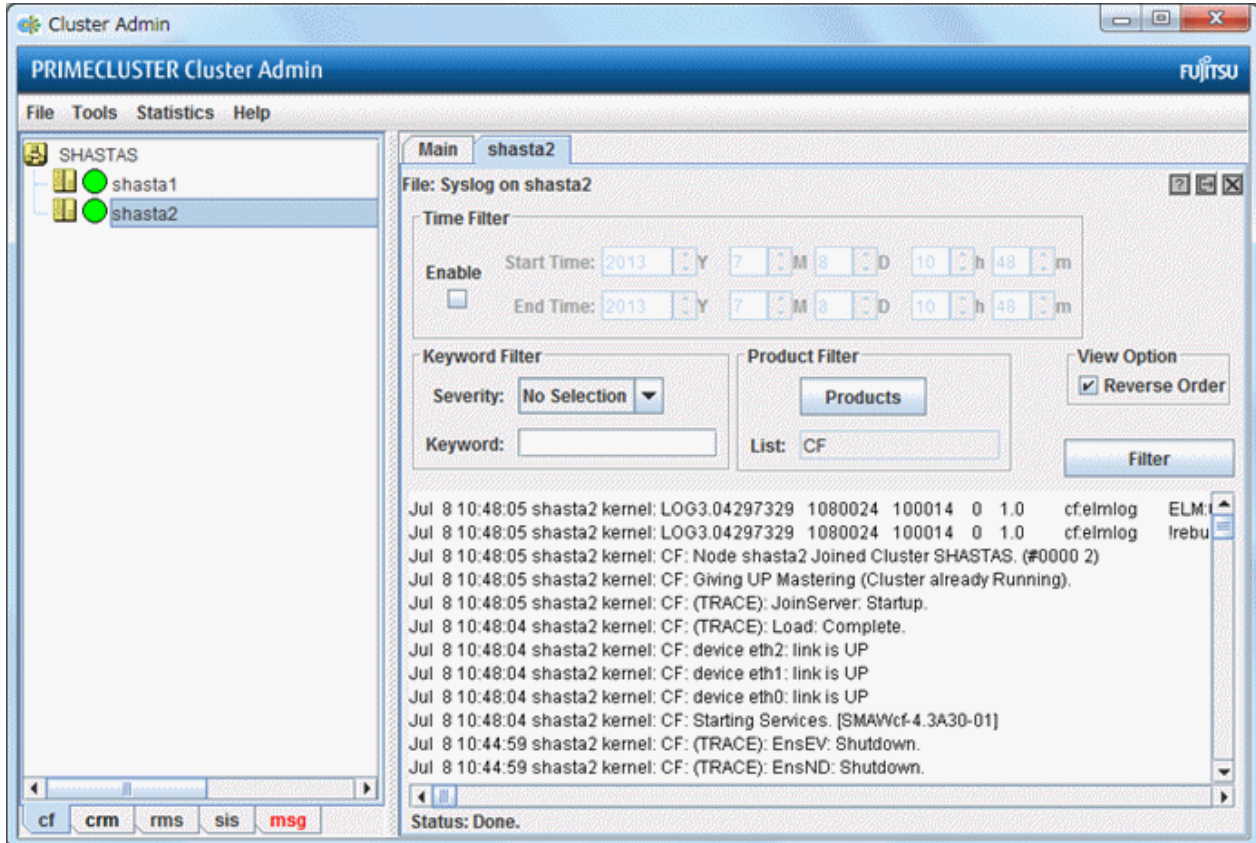
The CF log messages for a given node may be displayed by right-clicking on the node in the tree and selecting View CF Messages.

Alternately, you may go to the Tools menu and select View CF Messages. This brings up a pop-up where you can select the node whose syslog messages you would like to view.

When invoked from within CF, the PRIMECLUSTER log viewer only displays CF syslog messages. To view messages from other products, select the Products button on the Product Filter window pane on the PRIMECLUSTER log viewer screen.

Below is the example of the "[Figure 4.19 PRIMECLUSTER log viewer](#)."

Figure 4.19 PRIMECLUSTER log viewer



The messages appear in the right-hand panel. If you click on the Products button, then only the messages appear for the product that you select. To list the error messages for all of the installed PRIMECLUSTER products, choose All messages. Your choice is then listed in the List field.

The PRIMECLUSTER log viewer has search filters based on date/time/keyword and severity levels.

The Reverse Order checkbox is selected by default. This option reverses the order of the messages. To disable this feature, deselect the checkbox.



Note

If you have built a cluster system in an environment that uses Firewall, these logs cannot be displayed from Cluster Admin.

Display CF messages (/var/log/messages) using a common UNIX text editor such as vi.

4.8.1 Search based on time filter

To perform a search based on a start and end time, click the check box for Enable, specify the start and end times for the search range, and click on the Filter button.

4.8.2 Search based on keyword

To perform a search based on a keyword, enter a keyword and click on the Filter button.

4.8.3 Search based on severity levels

To perform a search based severity levels, click on the Severity pull-down menu. You can choose from the severity levels shown in Table 3 and click on the Filter button.

Table 4.3 Table 3: PRIMECLUSTER log viewer severity levels

Severity level	Severity description
Emergency	Systems cannot be used.
Alert	Immediate action is necessary.
Critical	Error that makes it impossible for the associated PRIMECLUSTER product to continue running.
Error	Error condition that arises unexpectedly, causing the associated PRIMECLUSTER function to terminate abnormally.
Warning	Minor error that does not terminate the offending function.
Notice	Normal but important condition
Info	Provides information on the status of a PRIMECLUSTER operation.
Debug	Verbose message that provides more information on why an error condition occurred.

4.9 Displaying statistics

CF can display various statistics about its operation. There are three types of statistics available:

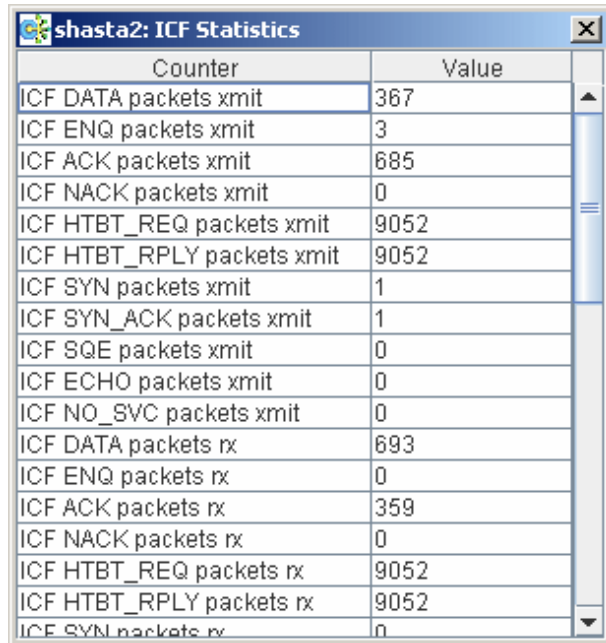
- ICF
- MAC
- Node to Node

To view the statistics for a particular node, right-click on that node in the tree and select the desired type of statistic.

Alternately, you can go to the Statistics menu and select the desired statistic. This will bring up a pop-up where you can select the node whose statistics you would like to view. The list of nodes presented in this pop-up will be all the nodes whose states are UP as viewed from the login node.

The figure below shows the display window for ICF Statistics.

Figure 4.20 ICF statistics

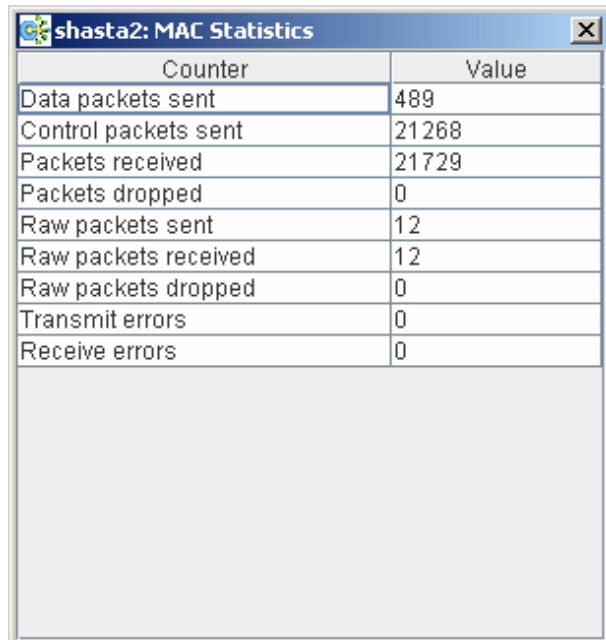


The image shows a window titled "shasta2: ICF Statistics". It contains a table with two columns: "Counter" and "Value". The table lists various ICF packet statistics for transmission (xmit) and reception (rx).

Counter	Value
ICF DATA packets xmit	367
ICF ENQ packets xmit	3
ICF ACK packets xmit	685
ICF NACK packets xmit	0
ICF HTBT_REQ packets xmit	9052
ICF HTBT_RPLY packets xmit	9052
ICF SYN packets xmit	1
ICF SYN_ACK packets xmit	1
ICF SQE packets xmit	0
ICF ECHO packets xmit	0
ICF NO_SVC packets xmit	0
ICF DATA packets rx	693
ICF ENQ packets rx	0
ICF ACK packets rx	359
ICF NACK packets rx	0
ICF HTBT_REQ packets rx	9052
ICF HTBT_RPLY packets rx	9052
ICF SYN packets rx	0

The figure below shows the display window for MAC Statistics.

Figure 4.21 MAC statistics

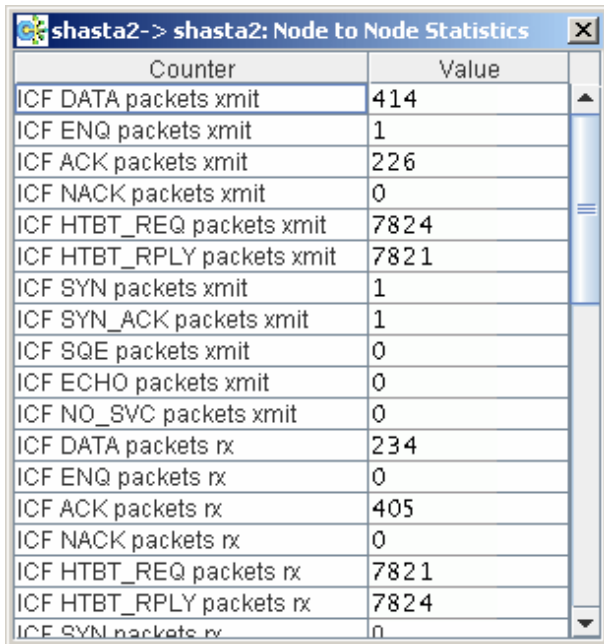


The image shows a window titled "shasta2: MAC Statistics". It contains a table with two columns: "Counter" and "Value". The table lists various MAC layer statistics.

Counter	Value
Data packets sent	489
Control packets sent	21268
Packets received	21729
Packets dropped	0
Raw packets sent	12
Raw packets received	12
Raw packets dropped	0
Transmit errors	0
Receive errors	0

The figure below shows the display window for node to node statistics.

Figure 4.22 Node to Node statistics



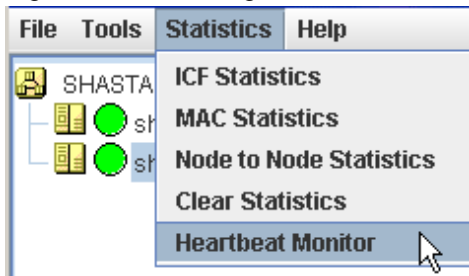
Counter	Value
ICF DATA packets xmit	414
ICF ENQ packets xmit	1
ICF ACK packets xmit	226
ICF NACK packets xmit	0
ICF HTBT_REQ packets xmit	7824
ICF HTBT_RPLY packets xmit	7821
ICF SYN packets xmit	1
ICF SYN_ACK packets xmit	1
ICF SEQ packets xmit	0
ICF ECHO packets xmit	0
ICF NO_SVC packets xmit	0
ICF DATA packets rx	234
ICF ENQ packets rx	0
ICF ACK packets rx	405
ICF NACK packets rx	0
ICF HTBT_REQ packets rx	7821
ICF HTBT_RPLY packets rx	7824
ICF SYN packets rx	0

The statistics counters for a node can be cleared by right-clicking on a node and selecting Clear Statistics from the command pop-up. The Statistics menu also offers the same option.

4.10 Heartbeat monitor

To display the Heartbeat monitor, go to the Statistics menu and select Heartbeat Monitor (see below.)

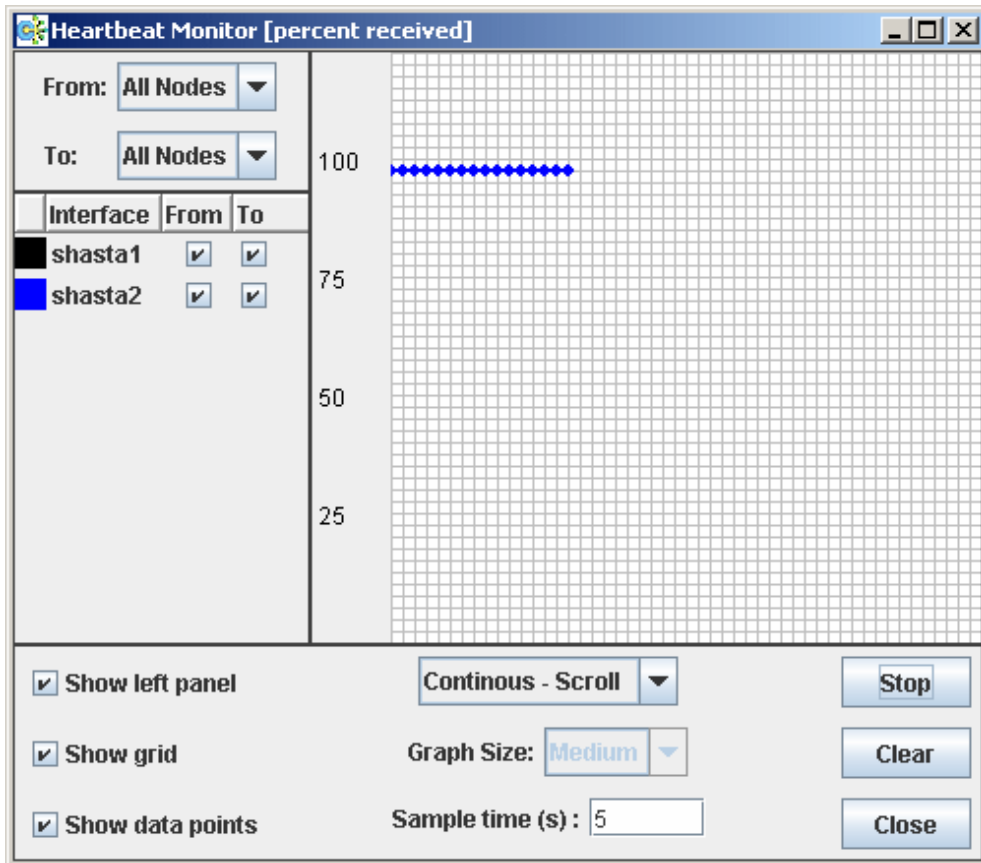
Figure 4.23 Selecting the Heartbeat monitor



The Heartbeat monitor allows you to monitor the percentage of heartbeats that are being received by CF over time. On a healthy cluster, this is normally close to 100 percent.

The Y axis is the percentage of heartbeats that have been successfully received and the X axis is a configurable time interval (see below.)

Figure 4.24 Heartbeat monitor



The controls on the left panel determine which data the graph shows as follows:

- The selection boxes at the top can be set to an individual node, or to All Nodes.
- The check boxes below the selection boxes allow the enabling and disabling of specific nodes.

The controls on the left of the bottom panel control how the graphing and information collection is done as follows:

- The Show left panel check box hides the left panel to provide more room for the graph.
- The Show grid check box turns the grid on and off.
- The Show data points check box can be turned off to display a simple line graph.

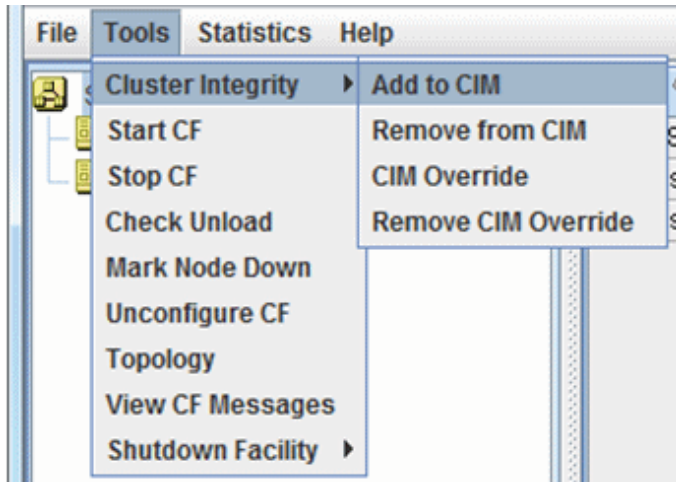
The controls in the bottom panel are as follows:

- The drop-down menu below the graph controls how the graph is drawn. The following options are available:
 - Continuous-Scroll - creates a continuous graph, so that when there are more data points than space, the graph scrolls.
 - Continuous-Clear - graphs continuously, but when the graph is full, clears it and starts a new graph.
 - Single Graph - creates a single graph only.
- Graph size - allows you to control how many data points are drawn.
- Sample time - controls how often data points are taken.
- The buttons on the lower right control starting and stopping of the graph, clearing it, and closing the graph window.

4.11 Adding and removing a node from CIM

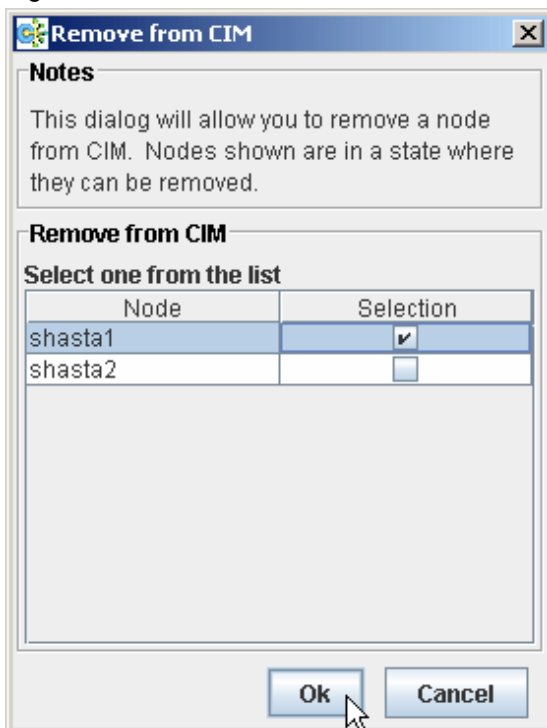
To add a node to CIM, click on the Tools pull-down menu. Select Cluster Integrity and Add to CIM from the expandable pull-down menu (see below.)

Figure 4.25 CIM options



The Add to CIM pop-up display appears. Choose the desired CF node and click on OK.

Figure 4.26 Add to CIM



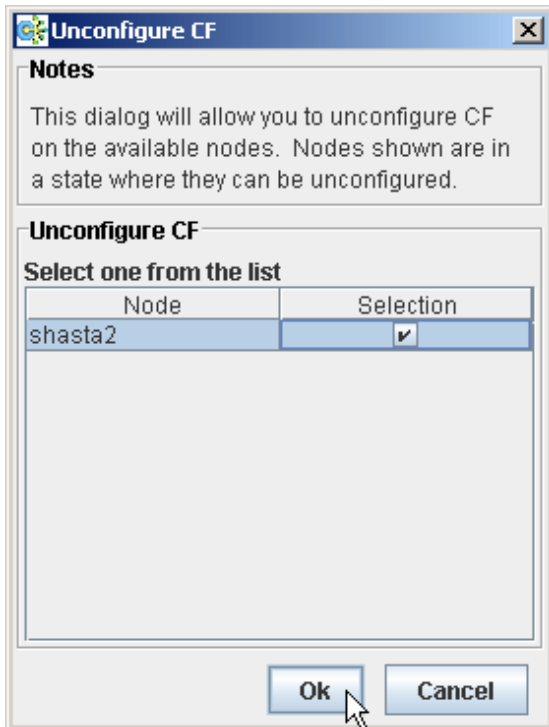
To remove a node from CIM by means of the Tools pull-down menu, select Cluster Integrity and Remove from CIM from the expandable pull-down menu. Choose the CF node to be removed from the pop-up and click on Ok. A node can be removed at any time.

Refer to the Section "[2.2 Cluster Integrity Monitor\(CIM\)](#)" for more details on CIM.

4.12 Unconfigure CF

To unconfigure a CF node, first stop CF on that node. Then, from the Tools pull-down menu, click on Unconfigure CF. The Unconfigure CF pop-up display appears. Select the check box for the CF node to unconfigure, and click Ok (see below.)

Figure 4.27 Unconfigure CF

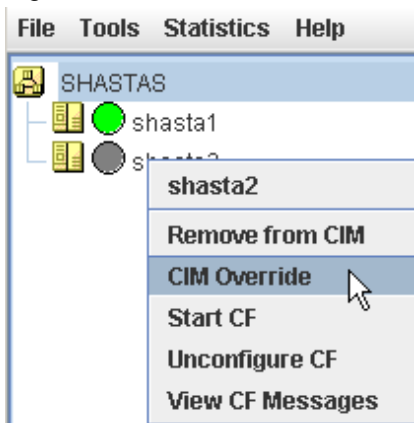


The unconfigured node will no longer be part of the cluster. However, other cluster nodes will still show that node as DOWN until they are rebooted.

4.13 CIM Override

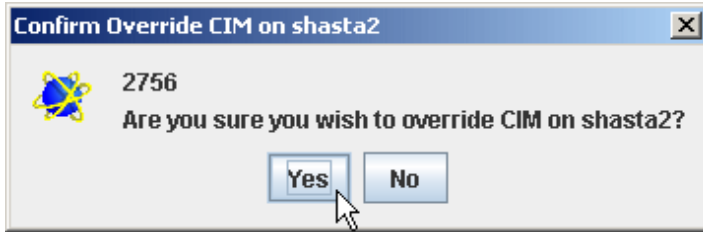
The CIM Override option causes a node to be ignored when determining a quorum. A node cannot be overridden if its CF state is UP. To select a node for CIM Override, right-click on a node and choose CIM Override (see below.)

Figure 4.28 CIM Override



A confirmation pop-up appears (see below.)

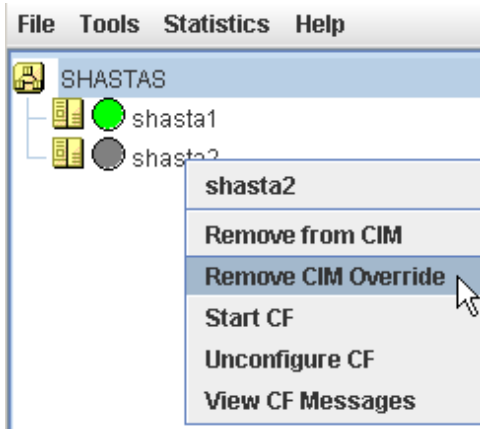
Figure 4.29 CIM Override confirmation



Click Yes to confirm.

Setting CIM override is a temporary action. It may be necessary to remove it manually again. This can be done by right-clicking on a node and selecting Remove CIM Override from the menu (see below.)

Figure 4.30 Remove CIM Override



CIM override is automatically removed when a node rejoins the cluster.

Chapter 5 LEFTCLUSTER state

This chapter describes the LEFTCLUSTER state and how to clear the state.

Occasionally, while CF is running, you may encounter the LEFTCLUSTER state, as shown by running the `cftool -n` command. A message will be printed to the console of the remaining nodes in the cluster. This can occur under the following circumstances:

- Broken interconnects - All cluster interconnects going to another node (or nodes) in the cluster are broken.
- Panicked nodes - A node panics.
- Reboot - Shutting down a node with the reboot command.



Note

Nodes running CF should normally be shut down with the shutdown command or with the init command. If the `reboot -f` command is executed, the node will be in the LEFTCLUSTER state.

The shutdown command or the init command will run the CF's Stop script that will allow CF to be cleanly shut down on that node. However, if you run the `reboot -f`, `halt -f`, or `poweroff -f` command, the CF's Stop script is not run, and the node will go down while CF is running. This will cause the node to be declared to be in the LEFTCLUSTER state by the other nodes.

If the Shutdown Facility (SF) is fully configured and running on all cluster nodes, it will try to resolve the LEFTCLUSTER state automatically. If SF is not configured and running, or the SF fails to clear the state, the state has to be cleared manually. This section explains the LEFTCLUSTER state and how to clear this state manually.

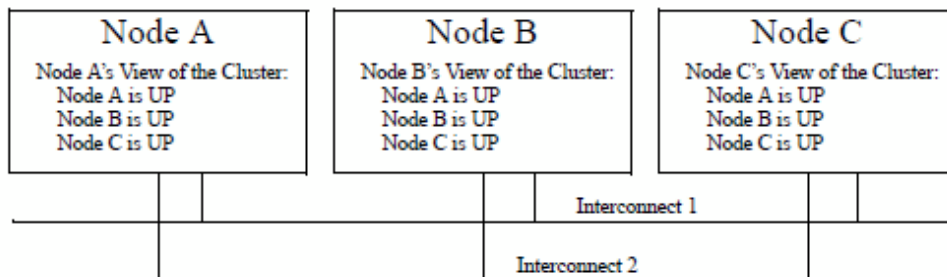
5.1 Description of the LEFTCLUSTER state

Each node in a CF cluster keeps track of the state of the other nodes in the cluster. For example, the other node's state may be UP, DOWN, or LEFTCLUSTER.

LEFTCLUSTER is an intermediate state between UP and DOWN, which means that the node cannot determine the state of another node in the cluster because of a break in communication.

For example, consider the three-node cluster shown in

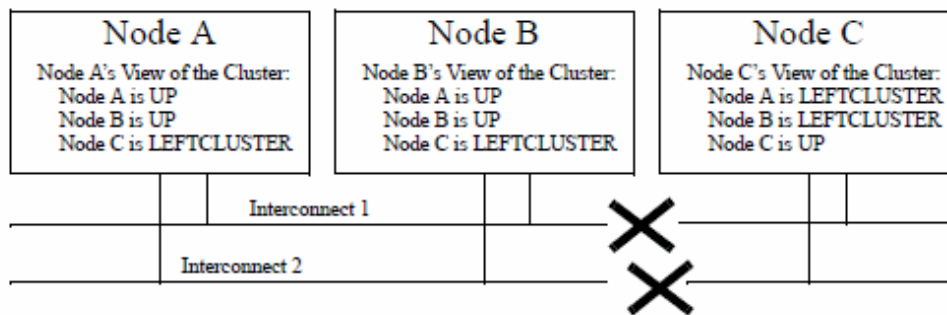
Figure 5.1 Three-node cluster with working connections



Each node maintains a table of what states it believes all the nodes in the cluster are in.

Now suppose that there is a cluster partition in which the connections to Node C are lost. The result is shown in the following figure.

Figure 5.2 Three-node cluster where connection is lost



Because of the break in network communications, Nodes A and B cannot be sure of Node C's true state. They therefore update their state tables to declare that Node C is in the LEFTCLUSTER state. Likewise, Node C cannot be sure of the true states of Nodes A and B, so it marks those nodes as being in the LEFTCLUSTER in its state table.

Note

LEFTCLUSTER is a state that a particular node believes other nodes are in. It is never a state that a node believes that it is in. For example, in, "Figure 5.2 Three-node cluster where connection is lost" each node believes that it is UP.

The purpose of the LEFTCLUSTER state is to warn applications which use CF that contact with another node has been lost and that the state of such a node is uncertain. This is very important for RMS.

For example, suppose that an application on Node C was configured under RMS to fail over to Node B if Node C failed. Suppose further that Nodes C and B had a shared disk to which this application wrote.

RMS needs to make sure that the application is, at any given time, running on either Node C or B but not both, since running it on both would corrupt the data on the shared disk.

Now suppose for the sake of argument that there was no LEFTCLUSTER state, but as soon as network communication was lost, each node marked the node it could not communicate with as DOWN. RMS on Node B would notice that Node C was DOWN. It would then start an instance of the application on Node C as part of its cluster partition processing. Unfortunately, Node C isn't really DOWN. Only communication with it has been lost. The application is still running on Node C. The applications, which assume that they have exclusive access to the shared disk, would then corrupt data as their updates interfered with each other.

The LEFTCLUSTER state avoids the above scenario. It allows RMS and other application using CF to distinguish between lost communications (implying an unknown state of nodes beyond the communications break) and a node that is genuinely down.

When SF notices that a node is in the LEFTCLUSTER state, it contacts the previously configured Shutdown Agent and requests that the node which is in the LEFTCLUSTER state be shut down. With PRIMECLUSTER, a weight calculation determines which node or nodes should survive and which ones should be shut down. SF has the capability to arbitrate among the shutdown requests and shut down a selected set of nodes in the cluster, such that the subcluster with the largest weight is left running and the remaining subclusters are shutdown.

In the example given, Node C would be shut down, leaving Nodes A and B running. After the SF software shuts down Node C, SF on Nodes A and B clear the LEFTCLUSTER state such that Nodes A and B see Node C as DOWN. Refer to the Chapter "Chapter 7 Shutdown Facility" for details on configuring SF and shutdown agents.

Note

Note that a node cannot join an existing cluster when the nodes in that cluster believe that the node is in the LEFTCLUSTER state. Therefore, any nodes in LEFTCLUSTER state have to be recovered before they can join an existing cluster.

5.2 Recovering from LEFTCLUSTER

When a node comes back up after being rebooted and attempts to rejoin the cluster, the join process automatically changes the node's state from LEFTCLUSTER to DOWN so that it can rejoin the cluster. When this occurs, the join server initially sees the node that is attempting to join the cluster as being in the LEFTCLUSTER state. The join server signals the joining node that it is busy because the joining node is

not in the DOWN state. It then notifies all of the remaining nodes in the cluster that the joining node is DOWN and to start the node-down processing, which must be completed before the node is allowed to rejoin the cluster. The joining node continues retrying to join the cluster until the node-down processing is completed on all of the cluster nodes at which time the joining node is allowed to rejoin the cluster.

If SF is not running on all the nodes, or if SF is unable to shut down the node which left the cluster, and the LEFTCLUSTER condition occurs, then the system administrator must manually clear the LEFTCLUSTER state. The procedure for doing this depends on how the LEFTCLUSTER condition occurred.

5.2.1 Caused by a panic/hung node

The LEFTCLUSTER state may occur because a particular node panicked or hung. In this case, the procedure to clear LEFTCLUSTER is as follows:

1. Make sure the node is really down. If the node panicked and came back up, proceed to Step 2. If the node is in the debugger, exit the debugger. The node will reboot if it panicked, otherwise shut down the node, called the offending node in the following discussion.
2. Use the Cluster Admin GUI to log into one of the running nodes in the cluster while the offending node is suspended. Go to the CF main window and select Mark Node Down from the Tools pull-down menu, then mark the offending node as DOWN. This may also be done from the command line by using the following command:

cftool -k

3. Reboot the offending node, it should automatically join the cluster.

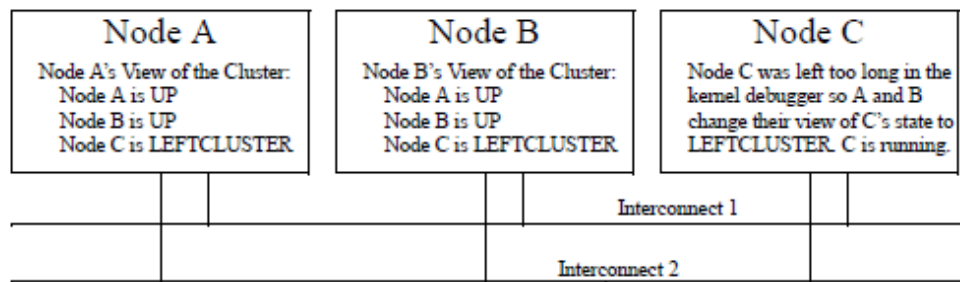


The state of the node is automatically changed from the state of LEFTCLUSTER in the state of DOWN so that the node may enter again when the offending node is reactivated.

5.2.2 Caused by staying in the kernel debugger too long

In the following figure, Node C was placed in the kernel debugger too long so it appears as a hung node. Nodes A and B decided that Node C's state was LEFTCLUSTER.

Figure 5.3 Node C placed in the kernel debugger too long



To recover from this situation, you would need to do the following:

1. Shut down Node C, and bring it back up.
2. If Node C fails to join the cluster and remains in the LEFTCLUSTER state after being shut down and coming back up, start up the Cluster Admin on Node A or B. Use Mark Node Down from the Tools pull-down menu in the CF portion of the GUI to mark Node C DOWN.
3. The node should successfully join the cluster.

5.2.3 Caused by a cluster partition

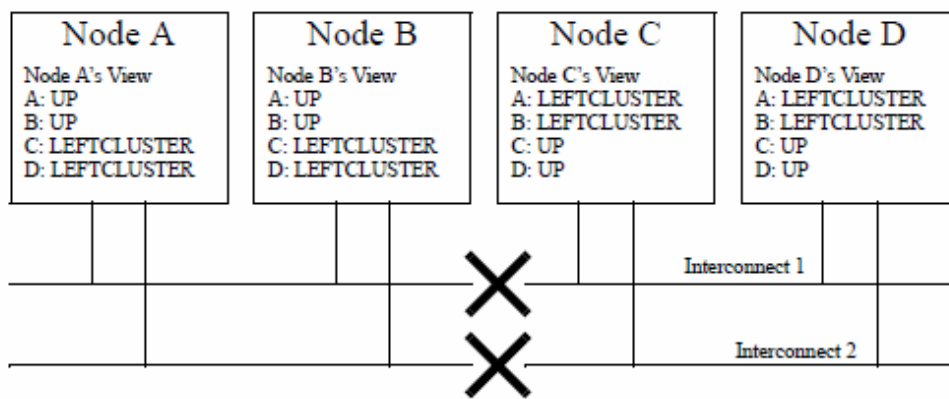
A cluster partition is a communications failure in which all CF communications between sets of nodes in the cluster are lost. In this case, the cluster itself is effectively partitioned into sub-clusters.

To manually recover from a cluster partition, you must do the following:

1. Decide which of the sub-clusters you want to survive. Typically, you will chose the sub-cluster that has the largest number of nodes in it or the one where the most important hardware is connected or the most important application is running.
2. Shut down all of the nodes in the sub-cluster which you don't want to survive.
3. Fix the network break so that connectivity is restored between all the nodes in the cluster.
4. Bring the nodes back up.
5. If the nodes fail to join the cluster and remain in the LEFTCLUSTER state after being shut down and coming back up, use the Cluster Admin GUI to log on to one of the surviving nodes and run the CF portion of the GUI. Select Mark Node Down from the Tools menu to mark all of the shutdown nodes as DOWN.
6. The nodes should successfully join the cluster.

For example, consider the following figure

Figure 5.4 Four-node cluster with cluster partition



In this figure, a four-node cluster has suffered a cluster partition. Both of its CF interconnects (Interconnect 1 and Interconnect 2) have been severed. The cluster is now split into two sub-clusters. Nodes A and B are in one sub-cluster while Nodes C and D are in the other.

To recover from this situation, in instances where SF fails to resolve the problem, you would need to do the following:

1. Decide which sub-cluster you want to survive. In this example, let us arbitrarily decide that Nodes A and B will survive.
2. Shut down all of the nodes in the other sub-cluster, here Nodes C and D.
3. Fix the interconnect break on Interconnect 1 and Interconnect 2 so that both sub-clusters will be able to communicate with each other again.
4. Bring Nodes C and D back up.
5. If the LEFTCLUSTER state persists on Nodes C or D, run the Cluster Admin GUI on either Node A or Node B. Start the CF portion of the GUI and go to Mark Node Down from the Tools pull-down menu. Mark any nodes still in the LEFTCLUSTER state as DOWN.

Chapter 6 CF topology table

This chapter discusses the CF topology table as it relates to the CF portion of the Cluster Admin GUI.

The CF topology table is part of the CF portion of the Cluster Admin GUI. The topology table may be invoked from the *Tools->Topology* menu item in the GUI (refer to the Section ["4.5 Displaying the topology table"](#) in the Chapter "GUI administration"). It is also available during CF configuration in the CF Wizard in the GUI.

The topology table is designed to show the network configuration from perspective of CF. It shows what devices are on the same interconnects and can communicate with each other.

The topology table only considers Ethernet devices. It does not include any IP interconnects that might be used for CF, even if CF over IP is configured.

Displayed devices

The topology table is generated by doing CF pings on all the nodes in the cluster and then analyzing the results.

The rest of this chapter discusses the format of the topology table. The examples implicitly assume that all devices can be seen on each node. Again, this would be the case when first configuring a CF cluster.

6.1 Basic layout

The basic layout of the topology table is shown in the table below.

Table 6.1 Basic layout of the topology table

FUJI	Full interconnects		Partial interconnects		Unconnected devices
	Int 1	Int 2	Int 3	Int 4	
fuji2	eth0 eth2	eth1	eth3	eth5	eth4 eth6
fuji3	eth0	eth2	missing	eth1	
fuji4	eth1	eth2	eth3	missing	eth4

The upper-left-hand corner of the topology table gives the CF cluster name. Below it, the names of all of the nodes in the cluster are listed.

The CF devices are organized into three major categories:

- Full interconnects - Have working CF communications to each of the nodes in the cluster.
- Partial interconnects - Have working CF communications to at least two nodes in the cluster, but not to all of the nodes.
- Unconnected devices - Have no working CF communications to any node in the cluster.

If a particular category is not present, it will be omitted from the topology table. For example, if the cluster in ["Table 6.1 Basic layout of the topology table"](#) had no partial interconnects, then the table headings would list only full interconnects and unconnected devices (as well as the left-most column giving the clustername and node names).

Within the full interconnects and partial interconnects category, the devices are further sorted into separate interconnects. Each column under an Int number heading represents all the devices on an interconnect. (The column header Int is an abbreviation for Interconnect.) For example, in ["Table 6.1 Basic layout of the topology table"](#), there are two full interconnects listed under the column headings of Int 1 and Int 2.

Each row for a node represents possible CF devices for that node.

Thus, in ["Table 6.1 Basic layout of the topology table"](#), Interconnect 1 is a full interconnect. It is attached to eth0 and eth2 on fuji2. On fuji3, it is attached to eth0, and on fuji4, it is attached to eth1.

Since CF runs over Ethernet devices, the ethn devices in ["Table 6.1 Basic layout of the topology table"](#) represent the Ethernet devices found on the various systems. The actual names of these devices will vary depending on the type of Ethernet controllers on the system. For nodes whose CF driver was loaded with -L, only configured devices will be shown.

It should be noted that the numbering used for the interconnects is purely a convention used only in the topology table to make the display easier to read. The underlying CF product does not number its interconnects. CF itself only knows about CF devices and point-to-point routes.

If a node does not have a device on a particular partial interconnect, then the word missing will be printed in that node's cell in the partial interconnects column. For example, in "Table 6.1 Basic layout of the topology table", fuji3 does not have a device for the partial interconnect labeled Int 3.

6.2 Selecting devices

The basic layout of the topology table is shown in the following table. However, when the GUI actually draws the topology table, it puts check boxes next to all of the interconnects and CF devices as shown in the following table.

FUJI	Full interconnects		Partial interconnects		Unconnected devices
	Int 1	Int 2	Int 3	Int 4	
fuji2	eth0 eth2	eth1	eth3	eth5	eth4 eth6
fuji3	eth0	eth2	missing	eth1	
fuji4	eth1	eth2	eth3	missing	eth4

The check boxes show which of the devices were selected for use in the CF configuration. (In the actual topology table, check marks appear instead of x's.)

When the topology table is used outside of the CF Wizard, these check boxes are read-only. They show what devices were previously selected for the configuration. In addition, the unchecked boxes (representing devices which were not configured for CF) will not be seen for nodes where -L was used to load CF.

When the topology table is used within the CF Wizard, then the check boxes may be used to select which devices will be included in the CF configuration. Clicking on the check box in an Int number heading will automatically select all devices attached to that interconnect. However, if a node has multiple devices connected to a single interconnect, then only one of the devices will be selected.

For example, in Table 5, fuji2 has both eth0 and eth2 attached to Interconnect 1. A valid CF configuration allows a given node to have only one CF device configured per interconnect. Thus, in the CF Wizard, the topology table will only allow eth0 or eth2 to be selected for fuji2. In the above example, if eth2 were selected for fuji2, then eth0 would automatically be unchecked.

If the CF Wizard is used to add a new node to an existing cluster, then the devices already configured in the running cluster will be displayed as read-only in the topology table. These existing devices may not be changed without unconfiguring CF on their respective nodes.

For how to delete the CF configuration, see "4.12 Unconfigure CF."

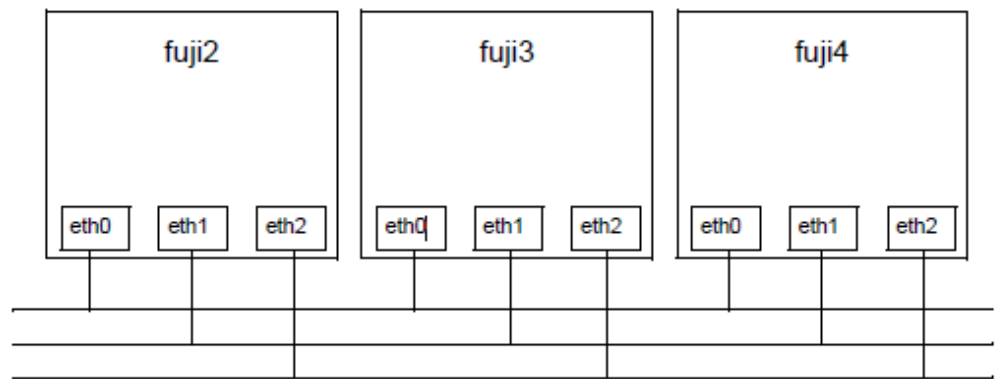
6.3 Examples of topology table

The following examples show various network configurations and what their topology tables would look like when the topology table is displayed in the CF Wizard on a totally unconfigured cluster. For simplicity, the check boxes are omitted.

Example 1

In this example, there is a three-node cluster with three full interconnects.

Figure 6.1 A three-node cluster with three full interconnects



The resulting topology table for the above figure is shown in the following table.

Table 6.2 Topology table for 3 full interconnects

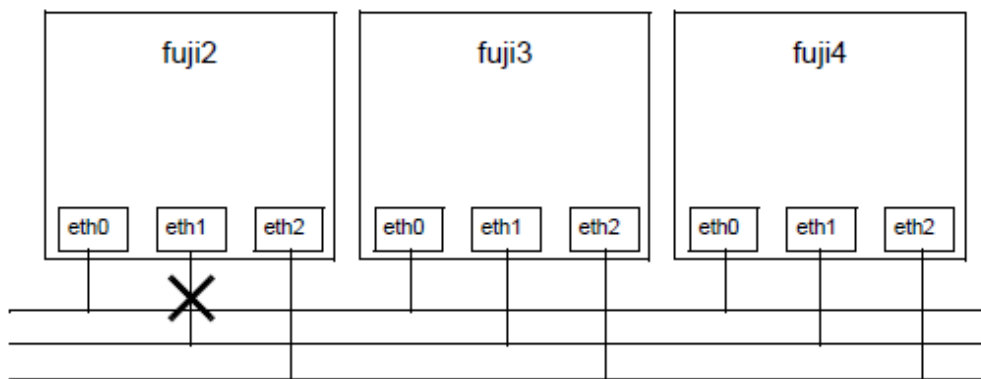
FUJI	Full interconnects		
	Int 1	Int 2	Int 3
fuji2	eth0	eth1	eth2
fuji3	eth0	eth1	eth2
fuji4	eth0	eth1	eth2

Since there are no partial interconnects or unconnected devices, those columns are omitted from the topology table.

Example 2

In this example, fuji2's Ethernet connection for eth1 has been broken.

Figure 6.2 Broken ethernet connection for eth1 on fuji2



The resulting topology table for Figure 6.3 is shown in Table 7.

Table 6.3 Topology table with broken Ethernet connection

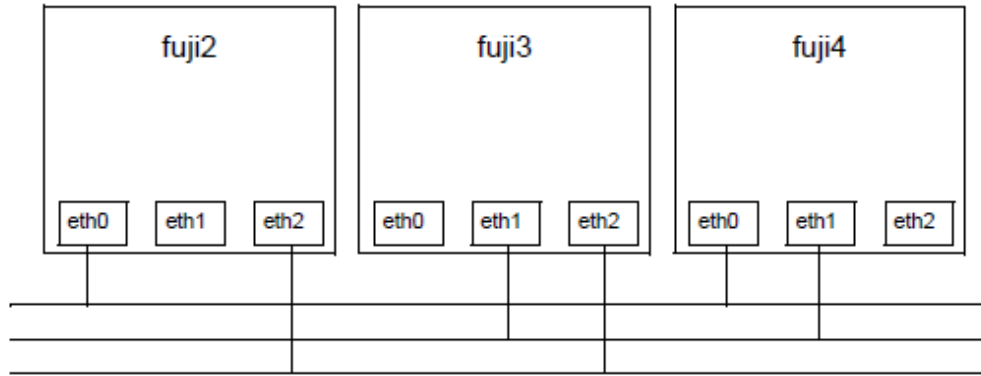
FUJI	Full interconnects		Partial interconnects	Unconnected devices
	Int 1	Int 2	Int 3	
fuji2	eth0	eth2	missing	eth1
fuji3	eth0	eth2	eth1	
fuji4	eth0	eth2	eth1	

In the table "[Table 6.3 Topology table with broken Ethernet connection](#)," eth1 for fuji2 now shows up as an unconnected device. Since one of the interconnects is missing a device for fuji2, the Partial Interconnect column now shows up. Note that the relationship between interconnect numbering and the devices has changed between the table "[Table 6.2 Topology table for 3 full interconnects](#)" and the table "[Table 6.3 Topology table with broken Ethernet connection](#)". In Table, for example, all eth1 devices were on Int 2. In the table "[Table 6.3 Topology table with broken Ethernet connection](#)," the eth1 devices for Nodes B and C are now on the partial interconnect Int 3. This change in numbering illustrates the fact that the numbers have no real significance beyond the topology.

Example 3

This example shows a cluster with severe networking or cabling problems in which no full interconnects are found.

Figure 6.3 Cluster with no full interconnects



The resulting topology table for the above figure is shown in the blow table.

Table 6.4 Topology table with no full interconnects

FUJI	Partial interconnects		Unconnected devices	
	Int 1	Int 2	Int 3	
fuji2	eth0	missing	eth2	eth1
fuji3	missing	eth1	eth2	eth0
fuji4	eth0	eth1	missing	eth2

In this table, the full interconnects column is omitted since there are none. Note that if this configuration were present in the CF Wizard, the wizard would not allow you to do configuration. The wizard requires that at least one full interconnect must be present.

Chapter 7 Shutdown Facility

This chapter describes the components and advantages of PRIMECLUSTER Shutdown Facility (SF) and provides administration information.



Note

Certain product options are region-specific. For information on the availability a specific Shutdown Agent (SA), contact field engineers.

7.1 Overview

The SF provides the interface for managing the shutdown of cluster nodes when error conditions occur. The SF also advises other PRIMECLUSTER products of the successful completion of node shutdown so that recovery operations can begin.

The SF is made up of the following major components:

- The Shutdown Daemon (SD)
- One or more Shutdown Agents (SA)
- *sdtool*(1M) command

Shutdown Daemon (SD)

The SD is started at system boot time and is responsible for the following:

- Monitoring the state of all cluster nodes
- Monitoring the state of all registered SAs
- Reacting to indications of cluster node failure and eliminating the nodes forcibly
- Resolving split-brain conditions
- Notifying other PRIMECLUSTER products that nodes were forcibly eliminated
- Checking the route that forcibly eliminates cluster nodes periodically (in 10-minute intervals)

The SD uses SAs to perform most of its work with regard to cluster node monitoring and forced node elimination. In addition to SA's, the SD interfaces with the Cluster Foundation layer's ENS system to receive node failure indications and to advertise node elimination completion.

The SD starts SA periodically (in 10-minute intervals) to check the route that forcibly eliminates cluster nodes.

The SD reflects the checked route status to the test status of each SA (Test State) displayed by the *sdtool*(1M) command.

Shutdown Agents (SA)

The SA's role is to attempt to shut down a remote cluster node in a manner in which the shutdown can be guaranteed. Some of the SAs are shipped with the SF product, but may differ based on the architecture of the cluster node on which SF is installed. SF allows any PRIMECLUSTER service layer product to shut down a node whether RMS is running or not.

An SA is responsible for shutting down, and verifying the shutdown of a cluster node. Each SA uses a specific method for performing the node shutdown such as:

- SA_blade provides an SA for the Fujitsu Technology Solutions Blade servers.
- SA_ipmi offers the shutdown agent for IPMI-based systems.
- SA_lkcd provides an SA that uses the kernel panic status of other nodes.
- SA_mmb provides an SA that uses the management board (MMB) on PRIMEQUEST nodes.
- SA_icmp provides an SA that checks whether a node to be stopped is in the active or inactive state by using a network route.
- SA_vmchghost provides an SA of the system which uses the KVM virtual machine function.

- SA_libvirtgp and SA_libvirtgr provide an SA of the system which uses the KVM virtual machine function.

See "7.2 Available SAs" for more information on SA.

If more than one SA is used, the first SA in the configuration file is used as the primary SA. SD always uses the primary SA. The other secondary SAs are used as fall back SAs only if the primary SA fails for some reason.

sdtool command

The *sdtool*(1M) command is the command line interface for interacting with the SD. With it the administrator can:

- Start and stop the SD (although this is typically done with an RC script run at boot time)
- View the current state of the SA's
- Force the SD to reconfigure itself based on new contents of its configuration file
- Dump the contents of the current SF configuration
- Enable/disable SD debugging output
- Eliminate a cluster node



Note

Although the *sdtool*(1M) command provides a cluster node elimination capability, the preferred method for controlled shutdown of a cluster node is the */sbin/shutdown* command.

7.2 Available SAs

This section describes the following set of supported SAs:

- Blade
- IPMI Intelligent Platform Management Interface
- kdump
- MMB PRIMEQUEST management board
- ICMP
- VMCHKHOST
- libvirt

7.2.1 Blade

The Blade Shutdown Agent, SA_blade, provides a Shutdown Agent for the Fujitsu Technology Solutions Blade servers. This is used in the SF product to provide a node shutdown facility for these nodes.

Setup and configuration

The Blade server should be configured according to the directions in the manual or manuals shipped with the unit.

For SA_blade to work properly, ensure the following:

- Make sure that ServerView software, containing an SNMP package, is properly installed.
- Ensure that the server Blades can communicate to the management Blade by means of SNMP commands. This includes the setting of proper security groups and communities in the SNMP configuration. Make a note of the SNMP community string that has both read and write permissions. This string is to be mentioned in the SA_blade configuration files. This means that the server Blades can both read and write (or change) the SNMP MIB data of the management blade.

The log file is stored in the following:

```
/var/opt/SMAWsf/log/SA_blade.log
```

7.2.2 IPMI



See

For complete configuration details of BMC or iRMS, refer to the appropriate hardware manual that came with your system.

The IPMI shutdown agent SA_ipmi provides the node shutdown function that uses either the BMC (Baseboard Management Controller) of PRIMERGY server or iRMC (remote management controller).

The log file is stored in the following:

```
/var/opt/SMAWsf/log/SA_ipmi.log
```

7.2.3 kdump

The kdump shutdown agent SA_lkcd is the SA that can be used when kdump is used. After another node panics, this SA executes high-speed switchover while a crash dump is being collected by kdump.

The log file is stored in the following:

```
/var/opt/SMAWsf/log/SA_lkcd.log
```

7.2.4 MMB

The MMB shutdown agent uses the MMB of PRIMEQUEST to provide shutdown mechanisms for nodes.

The MMB shutdown agent provides the following two shutdown mechanisms:

- SA_mmbp Shutdown mechanism that triggers panic in nodes through the MMB.
- SA_mmbp Shutdown mechanism that resets nodes through the MMB.

The log file is stored in the following:

```
/var/opt/SMAWsf/log/SA_mmbp.log  
/var/opt/SMAWsf/log/SA_mmbp.log
```

7.2.5 ICMP

The ICMP shutdown agent, SA_icmp, checks whether a node to be stopped is in the active or inactive state by using a network route.

If there is no response from the node to be stopped in all specified network routes, it determines that the node to be stopped is in the inactive state and terminates normally. If any responses are returned from the node in one or more network routes, it determines that the node to be stopped is in the active state and terminates abnormally.

The log file is stored in the following:

```
/var/opt/SMAWsf/log/SA_icmp.log
```

7.2.6 VMCHKHOST

The VMCHKHOST shutdown agent, SA_vmchghost, is the SA that can be used when the function, which switches the Host OS in the event of an error, is used on the KVM virtual machine function.

The log file is stored in the following:

```
/var/opt/SMAWsf/log/SA_vmchghost.log
```

7.2.7 libvirt

The libvirt shutdown agent provides the shutdown function for nodes (guest OS) in systems that are using the KVM virtual machine function.

There are two types of the libvirt shutdown agent as follows:

- SA_libvirtgp SF that panics nodes (guest OS)
- SA_libvirtgr SF that resets nodes (guest OS)

The log file is stored in the following:

```
/var/opt/SMAWsf/log/SA_libvirtgp.log  
/var/opt/SMAWsf/log/SA_libvirtgr.log
```

7.3 SF split-brain handling

The PRIMECLUSTER product provides the ability to gracefully resolve split-brain state as described in this section.

7.3.1 Administrative LAN

The SF handles a split-brain by using the administrative LAN.

7.3.2 SF split-brain handling

A split-brain condition is one in which one or more cluster nodes have stopped receiving heartbeats from one or more other cluster nodes, yet those nodes have been determined to still be running. Each of these distinct sets of cluster nodes is called a sub-cluster, and when a split-brain condition occurs the Shutdown Facility has a choice to make as to which sub-cluster should remain running.

Only one of the sub-clusters in a split-brain condition can survive. The SF determines which sub-cluster is most important and allows only that sub-cluster to remain. SF determines the importance of each sub-cluster by calculating the total node weight and application weight of each sub-cluster. The sub-cluster with the greatest total weight survives.

Node weights are defined in the SF configuration file `rcsd.cfg`.

Application weights are defined in RMS. Each RMS `userApplication` object can have a `ShutdownPriority` defined for it. The value of the `ShutdownPriority` is that application's weight. RMS calculates the total application weight for a particular node by adding up the weights of all applications that are Online on that node. If an application is switched from one node to another, its weight will be transferred to the new node.

SF combines the values for the RMS `ShutdownPriority` attributes and the SF weight assignments to determine how to handle a split-brain condition.

7.3.2.1 RMS ShutdownPriority attribute

RMS supports the ability to set application importance in the form of a `ShutdownPriority` value for each `userApplication` object defined within the RMS configuration. These values are combined for all `userApplication` objects that are Online on a given cluster node to represent the total application weight of that node. When a `userApplication` object is switched from one node to another, the value of that `userApplication` objects `ShutdownPriority` is transferred to the new node.

The higher the value of the `ShutdownPriority` attribute, the more important the application.

7.3.2.2 Shutdown Facility weight assignment

The Shutdown Facility supports the ability to define node importance in the form of a weight setting in the configuration file. This value represents a node weight for the cluster node.

The higher the node weight value, the more important the node.



Note

Although SF takes into consideration both SF node weights and RMS application weights while performing split-brain handling, it is recommended to use only one of the weights for simplicity and ease of use. When both weights are used, split-brain handling results are much more complex.

It is recommended that you follow the guidelines in "[7.3.4 Configuration notes](#)" for help you with the configuration.

7.3.3 Runtime processing

Split-brain handling may be performed by the following element of the Shutdown Facility:

- The Shutdown Facility internal algorithm

This method uses the node weight calculation to determine which sub-cluster is of greater importance. The total node weight is equal to the value of the defined Shutdown Facility node weight added to the total application weight of the Online applications for this node as calculated within RMS.

SF internal algorithm

When the SF is selected as the split-brain resolution manager, the SF uses the node weight internally.

The SF on each cluster node identifies which cluster nodes are outside its sub-cluster and adds each one of them to an internal shutdown list. This shutdown list, along with the local nodes node weight, is advertised to the SF instances running on all other cluster nodes (both in the local sub-cluster and outside the local sub-cluster) via the admIP network defined in the SF configuration file. After the SFs on each cluster node receive the advertisements, they each calculate the heaviest sub-cluster. The heaviest sub-cluster shuts down all lower weight sub-clusters.

In addition to handling well-coordinated shutdown activities defined by the contents of the advertisements, the SF internal algorithm will also resolve split-brain if the advertisements fail to be received. If the advertisements are not received then the split-brain will still be resolved, but it may take a bit more time as some amount of delay will have to be incurred.

The split-brain resolution done by the SF in situations where advertisements have failed depends on a variable delay based on the inverse of the percentage of the available cluster weight the local sub-cluster contains. The more weight it contains the less it delays. After the delay expires (assuming the sub-cluster has not been shut down by a higher-weight sub-cluster) the SF in the sub-cluster begins shutting down all other nodes in all other sub-clusters.

If a sub-cluster contains greater than 50 percent of the available cluster weight, then the SF in that sub-cluster will immediately start shutting down all other nodes in all other sub-clusters.

7.3.4 Configuration notes

When configuring the Shutdown Facility, RMS, and defining the various weights, the administrator should consider what the eventual goal of a split-brain situation should be.

Typical scenarios that are implemented are as follows:

- Largest Sub-cluster Survival
- Specific Hardware Survival
- Specific Application Survival

The weights applied to both cluster nodes and to defined applications allow considerable flexibility in defining what parts of a cluster configuration should survive a split-brain condition. Using the settings outlined below, administrators can advise the Shutdown Facility about what should be preserved during split-brain resolution.

Largest Sub-cluster Survival

In this scenario, the administrator does not care which physical nodes survive the split, just that the maximum number of nodes survive. If RMS is used to control applications, it will move the applications to the surviving cluster nodes after split-brain resolution has succeeded.

This scenario is achieved as follows:

- By means of Cluster Admin, set the SF node weight values to 1. 1 is the default value for this attribute, so new cluster installations may simply ignore it.
- By means of the RMS Wizard Tools, set the RMS attribute ShutdownPriority of all userApplications to 0. 0 is the default value for this attribute, so if you are creating new applications you may simply ignore this setting.

If no specific action was taken by the system administrator regarding split-brain resolution outcome from the values of both SF weight and RMS ShutdownPriority, the default "Largest Sub-cluster Survival" is selected.

Specific Hardware Survival (SAS)

In this scenario, the administrator has determined that one or more nodes contain hardware that is critical to the successful functioning of the cluster as a whole.

This scenario is achieved as follows:

- Using Cluster Admin, set the SF node weight of the cluster nodes containing the critical hardware to values more than double the combined value of cluster nodes not containing the critical hardware.
- Using the RMS Wizard Tools, set the RMS attribute ShutdownPriority of all userApplications to 0. 0 is the default value for this attribute so if you are creating new applications you may simply ignore this setting.

As an example, in a four-node cluster in which two of the nodes contain critical hardware, set the SF weight of those critical nodes to 10 and set the SF weight of the non-critical nodes to 1. With these settings, the combined weights of both non-critical nodes will never exceed even a single critical node.

Specific Application Survival

In this scenario, the administrator has determined that application survival on the node where the application is currently Online is more important than node survival. This can only be implemented if RMS is used to control the application(s) under discussion. This can get complex if more than one application is deemed to be critical and those applications are running on different cluster nodes. In some split-brain situations, all applications will not survive and will need to be switched over by RMS after the split-brain has been resolved.

This scenario is achieved as follows:

- Using Cluster Admin, set the SF node weight values to 1. 1 is the default value for this attribute, so new cluster installations may simply ignore it.
- Using the RMS Wizard Tools, set the RMS attribute ShutdownPriority of the critical applications to more than double the combined values of all non-critical applications, plus any SF node weight.
- ShutdownPriority is set with a range from 1 to 20.

As an example, in a four-node cluster there are three applications. Set the SF weight of all the nodes to 1, and set the ShutdownPriority of the three applications to 20, 1, 1. This would define that the application with a ShutdownPriority of 20 would survive no matter what, and further that the sub-cluster containing the node on which this application was running would survive the split no matter what. To clarify this example, if the cluster nodes were A, B, C and D all with a weight of 1, and App1, App2 and App3 had ShutdownPriority of 20, 1 and 1 respectively, even in the worst-case split that node D with App1 was split from nodes A, B and C which had applications App2 and App3 the weights of the sub-clusters would be D with 21 and A,B,C with 5. The heaviest sub-cluster (D) would win.

7.4 Configuring the Shutdown Facility

This section describes how to use Command Line Interface (CLI) to configure the Shutdown Facility (SF).

7.4.1 Shutdown Daemon

To configure the Shutdown Daemon (SD), you will need to modify the file

/etc/opt/SMAW/SMAWsf/rcsd.cfg on every node in the cluster.

A file, rcsd.cfg.template, is provided under the /etc/opt/SMAW/SMAWsf directory, which is a sample configuration file for the Shutdown Daemon using fictitious nodes and agents.



Note

It is important that the rcsd.cfg file is identical on all cluster nodes; care should be taken in administration to ensure that this is true.

Here is the example of rcsd.cfg, which is created by editing an rcsd.cfg.template.

```
node1,weight=2,admIP=fuji2:agent=SA_mmbp,timeout=20:agent=SA_mnbr,timeout=20
node2,weight=2,admIP=fuji3:agent=SA_mmbp,timeout=20:agent=SA_mnbr,timeout=20
```

The configuration file must be created in the /etc/opt/SMAW/SMAWsf directory and must use rcsd.cfg as the file name.

The format of the configuration file is as follows:

```
cluster-node1 [,weight=w1][, admIP=admIP1]:agent=SA1, timeout=t1[:agent=SA2, timeout=T2]...
cluster-node2 [,weight=w2][,admIP=admIP2]:agent=SA1, timeout=t1[:agent=SA2, timeout=T2]...
```

```
...
```

- cluster-nodeN is the cfname of a node within the cluster.
- agent and timeout are reserved words.
- SAN is the command name of a SA.
- tN is the timeout duration (seconds) of the SA.

The SA runs for the following cases:

- When a node is forcibly eliminated.
- When checking the connection to the option hardware used when a node is forcibly eliminated.

When a timeout occurs at the forced node elimination, the processing of the SA is stopped, and then the next SA is started. If all SAs fail to perform their processing, the node is left in the LEFTCLUSTER state.

When a timeout occurs when checking the connection to the option hardware used at the forced node elimination, the processing of the SA is stopped, and then the node becomes the TestFailed state.

- *wN* is the node weight.
- *admIPN* is the admin interface on the Administrative LAN on this cluster node. Available IP addresses are IPv4 and IPv6 address. The link local address of IPv6 is not available. When specifying the IPv6 address, enclose it in brackets "[]".
(Example: [1080:2090:30a0:40b0:50c0:60d0:70e0:80f0])

The order of the SAs in the configuration file should be such that the first SA in the list is the preferred SA. If this preferred SA is issued a shutdown request and if its response indicates a failure to shut down, the secondary SA is issued the shutdown request. This request/response is repeated until either an SA responds with a successful shutdown, or all SAs have been tried. If no SA is able to successfully shut down a cluster node, then operator intervention is required and the node is left in the LEFTCLUSTER state.

The location of the log file will be /var/opt/SMAWsf/log/rcsd.log.

7.4.2 Shutdown Agents

This section contains information on how to configure the following SAs with CLI.

- Blade
- IPMI
- ICMP
- VMCHKHOST
- libvirt

Blade

To configure the Blade shutdown agent, you can create or modify the following file:


```
/etc/opt/SMAW/SMAWsf/SA_blade.cfg
```

A sample configuration file can be found at the following location:

```
/etc/opt/SMAW/SMAWsf/SA_blade.cfg.template
```

The format of the SA_blade.cfg file is as follows:

```
community-string      SNMP community string
management-blade-ip    IP address
cfname      slot-no    Action
```

Additionally, you can add the IP address of the management Blade for a second Blade chassis. In this case, you would need to add one or more lines to specify the Blades in that chassis.



Note

management-blade-ip, community-string, cycle, and leave-off are reserved words. They must be described in lowercase letters.

The editable fields are defined as follows:

- SNMP community string
The SNMP community string with read/write permissions for the server Blades. This string is the same value as the SNMP community string in the Management Blade SNMP configuration. By default, this string is usually set to public.
- IP-address
The IP name or address in dot notation of the Management Blade. Available IP addresses are IPv4 and IPv6 address. The link local address of IPv6 is not available.
- cfname
The name of the node in the CF cluster.
- slot-no
The slot number of the Blade server.
- Action
The action can either be cycle or leave-off. If it is cycle, the node will be powered on again after powering off. If it is leave-off, manual action is required to turn the system back on.

The following is an example of the SA_blade configuration file:

```
community-string public
management-blade-ip 123.45.56.78
shasta1 1 cycle
shasta2 3 leave-off
management-blade-ip 123.45.56.79
shasta3 1 cycle
shasta4 2 cycle
```

The log file is stored in the following:

```
/var/opt/SMAWsf/log/SA_blade.log
```

IPMI

To configure the IPMI SA, you need to create or modify the following file:

```
/etc/opt/SMAW/SMAWsf/SA_ipmi.cfg
```

A sample configuration file can be found at the following location:

```
/etc/opt/SMAW/SMAWsf/SA_ipmi.cfg.template
```

The SA_ipmi.cfg configuration file contains lines with four fields (and some subfields) on each line. Each line defines a node in the cluster than can be powered off (leaving it off) or powered off and then on again (power cycle). The fields are as follows:

- cfname
The name of the node in the CF cluster.

Note

cfname must correspond to the IP address of the IPMI compliant onboard LAN interface in these nodes.

- Access-Information
The access information is of the following format:

```
ip-address-of-unit:user:password
```

For ip-address-of-unit, specify the IP address of IPMI (BMC and iRMC). Available IP addresses are IPv4 and IPv6 address. The link local address of IPv6 is not available. When specifying the IPv6 address, enclose it in brackets "[]".
(Example: [1080:2090:30a0:40b0:50c0:60d0:70e0:80f0])

- Action
The action can either be cycle or leave-off. If it is cycle, the node will be powered on again after powering off. If it is leave-off, manual action is required to turn the system back on.

Note

The permissions of the SA_ipmi.cfg file are read/write by root only. This is to protect the password to the BMC/iRMC unit.

The following is an example of the SA_ipmi configuration file:

```
fuji2 192.168.200.1::root:ipmipwd cycle
fuji3 192.168.200.2::root:ipmipwd cycle
fuji4 192.168.200.3::root:ipmipwd leave-off
fuji5 192.168.200.4::root:ipmipwd leave-off
```

The log file is stored in the following:

```
/var/opt/SMAWsf/log/SA_ipmi.log
```

ICMP

To configure the ICMP SA (SA_icmp), you need to create or modify the following file:

```
/etc/opt/SMAW/SMAWsf/SA_icmp.cfg
```

The format of the SA_icmp.cfg file is as follows:

```
TIME_OUT=value
cfname:ip-address-of-node:NIC-name1,NIC-name2
```

The editable fields are defined as follows:

- value
Specify the time in seconds to check for the existence of a node. The recommended value is 5 seconds.
- cfname
Specify a CF node name.
- ip-address-of-node
Specify any of the following IP addresses of cfname. Available IP addresses are IPv4 and IPv6 address. The link local address of IPv6

is not available. When specifying the IPv6 address, enclose it in brackets "[]".
(Example: [1080:2090:30a0:40b0:50c0:60d0:70e0:80f0])

- Administrative LAN
- Public LAN
- Cluster interconnect

You must describe one or more IP addresses for all the nodes in the cluster. Add a new line to specify more than one LAN routes. It is recommended that you specify more than one routes to check the LAN routes.

- NIC-nameX

For ip-address-of-node, specify a network interface used to check the existence of the node. If there are more than one network interfaces, separate them with a comma.



Note

Note the following when you describe network interfaces:

- When duplicating the network by GLS, you need to describe all redundant network interfaces. (For example, eth0,eth1)
- When bonding NICs, you need to describe a bonding device after the IP address. (For example, bond0)
- When describing cluster interconnects, you need to describe all network interfaces used in all paths of the cluster interconnects. (For example, eth2,eth3)

The following is an example of the SA_icmp configuration file:

```
TIME_OUT=5
node1:10.20.30.100:eth0,eth1
node1:10.20.40.200:eth2
node2:10.20.30.101:eth0,eth1
node2:10.20.40.201:eth2
```

The log file is stored in the following:

```
/opt/SMAW/SMAwf/log/SA_icmp.log
```

VMCHKHOST

To configure the Shutdown Agent (SA_vmchkhos) to switch the HOST OS in the event of an error in the virtual machine environment, you must create or modify the following file:

```
/etc/opt/SMAW/SMAwf/SA_vmchkhos.cfg
```

Templates of the configuration file are stored in the following:

```
/etc/opt/SMAW/SMAwf/SA_vmchkhos.cfg.template
```

Each line of the configuration file SA_vmchkhos includes the following 5 fields:

- guest-cfname
CF node name of a guest OS.
- host-cfname
CF node name of the Host OS.
- ip-address
IP address of the Host OS. Available IP addresses are IPv4 and IPv6 address. The link local address of IPv6 is not available.
- user
Account of the Host OS.
For a KVM environment, specify the user name that was created when setting libvirt to the Shutdown Facility.

- password

Login password for the account that was specified with "user."

The encrypted password for the user with general privileges for the Shutdown Facility, which was used when setting libvirt to the Shutdown Facility, is used.

The following is an example of the SA_vmchkhos configuration file:

```
fuji2 hostos2 192.168.200.1 FJSVvmSP 3CA1wxVXKD8a93077BaEkA==
fuji3 hostos3 192.168.200.2 FJSVvmSP 3CA1wxVXKD8a93077BaEkA==
```

The log file is stored in the following:

```
/var/opt/SMAWsf/log/SA_vmchkhos.log
```

libvirt

To configure the Shutdown Agent (SA_libvirtgp) for PANIC or Shutdown Agent (SA_libvirtgr) for RESET of libvirt, you must create or modify the following files respectively.

```
/etc/opt/SMAW/SMAWsf/SA_libvirtgp.cfg
/etc/opt/SMAW/SMAWsf/SA_libvirtgr.cfg
```

Templates of the configuration file are stored in the following:

```
/etc/opt/SMAW/SMAWsf/SA_libvirtgp.cfg.template
/etc/opt/SMAW/SMAWsf/SA_libvirtgr.cfg.template
```

Each line of the configuration files, SA_libvirtgp.cfg and SA_libvirtgr.cfg, includes the following 5 fields:

- cfname
Node name on the CF cluster.
- domain
Domain name of a guest OS.
- ip-address
IP address of the hypervisor. Available IP addresses are IPv4 and IPv6 address. The link local address of IPv6 is not available.
- user
Account of the hypervisor. The user for shutdown facility is specified.
- password
Login password for the account specified in "user" field. A password encrypted by the encryption command, sfcipher(8). For details on sfcipher(8), see "Manual pages."

The following is an example of the SA_libvirtgp configuration file:

```
fuji2 domain2 192.168.200.1 user 3CA1wxVXKD8a93077BaEkA==
fuji3 domain3 192.168.200.2 user 3CA1wxVXKD8a93077BaEkA==
```

The following is an example of the SA_libvirtgr configuration file:

```
fuji2 domain2 192.168.200.1 user 3CA1wxVXKD8a93077BaEkA==
fuji3 domain3 192.168.200.2 user 3CA1wxVXKD8a93077BaEkA==
```

The log file is stored in the following:

```
/var/opt/SMAWsf/log/SA_libvirtgp.log
/var/opt/SMAWsf/log/SA_libvirtgr.log
```

7.5 SF administration

This section provides information on administering SF. SF can be administered with the CLI or Cluster Admin. It is recommended to use Cluster Admin.

7.5.1 Starting and stopping SF

This section describes the following administrative procedures for starting and stopping SF:

- Manually via the CLI
- Automatically via the rc script interface (For Red Hat Enterprise Linux 6)
- Automatically via the systemd service (For Red Hat Enterprise Linux 7)

7.5.1.1 Starting and stopping SF manually

SF may be manually started or stopped by using the `sdtool(1M)` command. The `sdtool(1M)` command has the following options:

```
sdtool [-bsSre] [-k CF-node-name] [-d off|on]
```

-b	Start
-s	State (Human-readable format)
-S	State (Easy-to-analyze format)
-r	Re-configuration
-e	End
-k	Stop
-d	Debug

See the `sdtool` manual pages for details.

7.5.1.2 Starting and stopping SF automatically (For Red Hat Enterprise Linux 6)

SF can be started automatically using the `S13SMAWsf`-script available under the `/etc/rc3.d` directory. The rc start/stop script for SF is installed as `/etc/init.d/SMAWsf`.

7.5.1.3 Starting and stopping SF automatically (For Red Hat Enterprise Linux 7)

SF can be started automatically using the `smawsf.service` of a service of `systemd`. The start/stop script for SF is installed as `/etc/init.d/SMAWsf`.

7.5.2 Checking SA status

This section describes the procedures on how to check the current status of SA.

To check the current status of SA, execute the following command on the node where the status of SA is checked.

```
# sdtool -s
```

Example

```
# sdtool -s
Cluster Host Agent SA State Shut State Test State Init State
-----
node1      SA_xx Idle      Unknown      TestWorked  InitWorked
node1      SA_xx Idle      Unknown      TestWorked  InitWorked
node2      SA_xx Idle      Unknown      TestWorked  InitWorked
node2      SA_xx Idle      Unknown      TestWorked  InitWorked
```

The following items are displayed.

Cluster Host

Node name in the cluster

Agent

Name of the shutdown agent

SA State

Current Status of SA.

The status are as follows.

Status	Description
Idle	No currently running SA
Init-ing	Initializing
InitWorked	Initialization successfully completed
InitFailed	Initialization failed
Testing	Testing (10-minute interval)
TestWorked	Test successfully completed
TestFailed	Test failed
Killing	Force stop node ongoing
KillWorked	Force stop node successfully completed
KillFailed	Force stop node failed
UnInit-ing	Uninitializing
UnInitWorked	Uninitialization successfully completed
UnInitFailed	Uninitialization failed

Shut State

Status of force node stop.

The status are as follows.

Status	Description
Unknown	Force stop node is not executed
Killing	Force stop node ongoing
KillWorked	Force stop node successfully completed
KillFailed	Force stop node failed

Test State

Status of test executed every 10 minutes.

The status are as follows.

Status	Description
Unknown	Test is not executed
Testing	Testing
TestWorked	Test successfully completed
TestFailed	Test failed (force stop cannot be executed)

Init State

Status of initialization/uninitialization of SA.

The status are as follows.

Status	Description
Unknown	Initialization is not executed
Init-ing	Initializing
InitWorked	Initialization successfully completed
InitFailed	Initialization failed
UnInit-ing	Uninitializing
UnInit-Failed	Uninitialization failed



Note

- If Init State is displayed as InitFailed, it indicates that an error occurred in initializing the SA.
- If Test State is displayed as TestFailed, it indicates that an error occurred while the SA is testing if the node displayed as Cluster Host can be stopped, and the node cannot be stopped forcibly.
In this case, an error may be occurring in software, hardware, or network resources that are used by the SA. Solve the error immediately.
- If Test State or Init State is displayed as Unknown, it indicates that SF has not yet stopped the node, tested the route, or initialized SA. Test State and Init State are temporary displayed as Unknown until the actual status is confirmed.
- If TestFailed or InitFailed is displayed, check the SA log file or /var/log/messages. In the log file, you can find the cause of the error why the SA testing or SA initialization failed. Once the cause of the error is solved and SF is restarted, InitWorked or TestWorked will be displayed as the current status.

7.6 Debugging

Whenever there is a recurring problem where the cause cannot be easily detected, turn on the debugger with the following command:

```
# sdttool -d on
```

This will write the debugging information into the log file:

/var/opt/SMAWsf/log/rsd.log, which will provide additional information to find the cause of the problem. You can also use the sdttool -d off command to turn off debugging.

Note that the rsd log file does not contain logging information from any SA. Refer to the SA specific log files for logging information from a specific SA.

Chapter 8 Diagnostics and troubleshooting

This chapter provides help for troubleshooting and problem resolution for PRIMECLUSTER Cluster Foundation. This chapter will help identify the causes of problems and possible solutions. If a problem is in another component of the PRIMECLUSTER suite, the reader will be referred to the appropriate manual. This chapter assumes that the installation and verification of the cluster have been completed.

8.1 Beginning the process

Start the troubleshooting process by gathering information to help identify the causes of problems. You can use the CF log viewer facility from the Cluster Admin GUI, look for messages on the console, or look for messages in the

/var/log/messages file. You can use the *cftool*(1M) command for checking states, configuration information. To use the CF log viewer click on the Tools pull-down menu and select View Syslog messages (refer to the Section "4.8 Using PRIMECLUSTER log viewer" for more details). The log messages are displayed. You can search the logs using a date/time filter or scan for messages based on severity levels. To search based on date/time, use the date/time filter and press the *Filter* button. To search based on severity levels, click on the *Severity* button and select the desired severity level. You can use keyword also to search the log. To detach the CF log viewer window, click on the *Detach* button; click on the *Attach* button to attach it again.

Collect information as follows:

- Look for messages on the console that contain the identifier CF.
- Look for messages in /var/log/messages. You might have to look in multiple files (/var/log/messages.N).
- Use *cftool* as follows:
 - *cftool -l*: Check local node state
 - *cftool -d*: Check device configuration
 - *cftool -n*: Check cluster node states
 - *cftool -r*: Check the route status

Error log messages from CF are always placed in the /var/log/messages file; some messages may be replicated on the console. Other device drivers and system software may only print errors on the console. To have a complete understanding of the errors on a system, both console and error log messages should be examined. The section "4.5 Error Messages" in "PRIMECLUSTER Messages" contains messages that can be found in the /var/log/messages file. This list of messages gives a description of the cause of the error. This information is a good starting point for further diagnosis.

All of the parts of the system put error messages in this file or on the console and it is important to look at all of the messages, not just those from the PRIMECLUSTER suite. The following is an example of a CF error message from the /var/log/messages file:

```
Aug 26 13:31:05 fuji2 kernel: LOG3.0429320 1080024 100014 0 1.0 CF: Giving UP Mastering  
(Cluster already Running)
```

The parts of this message are as follows:

The first 80 bytes are the log3 prefix:

```
Aug 26 13:31:05 fuji2 kernel: LOG3. .0429320 1080024 100014 0 1.0 cf:elmlog
```

This parts of the message is a standard prefix on each CF message in the log file that gives the date and time, the node name, and log3 specific information. Only the date, time, and node name are important in this context. The remainder is the error message from CF as follows:

```
CF: Giving UP Mastering (Cluster already Running).
```

When the node detects a joined server, and it enters an existing cluster instead of making a new cluster, this message is output. Refer to "Chapter 5 CF Messages" in "PRIMECLUSTER Messages" for details of the message.

Several options for the command *cftool*(1M) are available as sources for information. The following is an example:

```
root@fuji2> cftool -l
Node    Number  State  Os      Cpu      Flags
fuji2   2       UP     Linux   Pentium  0
```


This shows that the local node has joined a cluster as node number 2 and is currently UP. This is the normal state when the cluster is operational. Another possible response is as follows:

```
root@fuji2> cftool -l
Node      Number  State      Os      Cpu      Flags
fuji2     --      COMINGUP   --      --
```

This indicates that the CF driver is loaded and that the node is attempting to join a cluster. If the node stays in this state for more than a few minutes, then something is wrong and we need to examine the /var/log/messages file. In this case, we see as follows:

```
root@fuji2> tail /var/log/messages
Aug 28 10:38:25 fuji2 kernel: CF: (TRACE): Load: Complete.
Aug 28 10:38:25 fuji2 kernel: CF: (TRACE): JoinServer: Startup.
Aug 28 10:38:25 fuji2 kernel: CF: Giving UP Mastering (Cluster already Running).
Aug 28 10:38:25 fuji2 kernel: CF: fuji2: busy: local node not DOWN: retrying.
```

We see that this node is in the LEFTCLUSTER state on another node (fuji4). To resolve this condition, see Chapter "5.1 Description of the LEFTCLUSTER state" for a description of the LEFTCLUSTER state and the instructions for resolving the state.

The next option to *cftool*(1M) shows the device states as follows:

```
root@fuji2> cftool -d
Number  Device  Type  Speed  Mtu  State  Configured  Address
1       eth0    4     100    1432  UP     YES         00.03.47.c2.a8.82
2       eth1    4     100    1432  UP     YES         00.02.b3.88.09.f1
3       eth2    4     100    1432  UP     NO          00.02.b3.88.09.ea
```

Here we can see that there are two interconnects configured for the cluster (the lines with YES in the Configured column). This information shows the names of the devices and the device numbers for use in further troubleshooting steps.

The *cftool -n* command displays the states of all the nodes in the cluster. The node must be a member of a cluster and UP in the *cftool -l* output before this command will succeed:

```
root@fuji2> cftool -n
Node      Number  State      Os      Cpu
fuji2     1       UP         Linux   Pentium
fuji3     2       UP         Linux   Pentium
```

This indicates that the cluster consists of two nodes fuji2 and fuji3, both of which are UP. If the node has not joined a cluster, the command will wait until the join succeeds.

cftool -r lists the routes and the current status of the routes as follows:

```
root@fuji2> cftool -r
Node      Number  Srcdev  Dstdev  Type  State  Destaddr
fuji2     1       1       4       4     UP     00.03.47.c2.a8.82
fuji2     1       1       5       5     UP     00.03.47.c2.a8.cc
fuji3     2       2       4       4     UP     00.03.47.d1.af.ec
fuji3     2       2       5       5     UP     00.03.47.d1.af.ef
```

This shows that all of the routes are UP. If a route shows a DOWN state, then the step above where we examined the error log should have found an error message associated with the device. At least the CF error noting the route is down should occur in the error log. If there is not an associated error from the device driver, then the diagnosis steps are covered below.

The last route to a node is never marked DOWN, it stays in the UP state so that the software can continue to try to access the node. If a node has left the cluster or gone down, there will still be an entry for the node in the route table and one of the routes will still show as UP. Only the *cftool -n* output shows the state of the nodes. The following example shows:

```
root@fuji2> cftool -r
Node      Number  Srcdev  Dstdev  Type  State  Destaddr
fuji3     2       3       2       4     UP     00.03.47.d1.af.ec
fuji2     1       3       3       4     UP     00.03.47.c2.a8.82
```

```
root@fuji2> cftool -n
Node      Number  State      Os      Cpu
```

fuji3	1	LEFTCLUSTER	Linux	Pentium
fuji2	2	UP	Linux	Pentium

8.2 Symptoms and solutions

The previous section discussed the collection of data. This section discusses symptoms and gives guidance for troubleshooting and resolving the problems. The problems dealt with in this section are divided into two categories: problems with joining a cluster and problems with routes, either partial or complete loss of routes. The solutions given here are either to correct configuration problems or to correct interconnect problems. Problems outside of these categories or solutions to problems outside of this range of solutions are beyond the scope of this manual and are either covered in another product's manual or require technical support from field engineers. Samples from the error log (/var/log/messages) have the log3 header stripped from them in this section.

8.2.1 Join-related problems

Join problems occur when a node is attempting to become a part of a cluster. The problems covered here are for a node that has previously successfully joined a cluster. If this is the first time that a node is joining a cluster, the *Software Release Guide PRIMECLUSTER* and the Installation Guide for PRIMECLUSTER section on verification covers the issues of initial startup. If this node has previously been a part of the cluster and is now failing to rejoin the cluster, here are some initial steps in identifying the problem.

8.2.1.1 Identifying join-related problems

First, look in the error log and at the console messages for any clue to the problem. Have the Ethernet drivers reported any errors? Any other unusual errors? If there are errors in other parts of the system, the first step is to correct those errors. Once the other errors are corrected, or if there were no errors in other parts of the system, proceed as follows.

Is the CF device driver loaded? The device driver puts a message in the log file when it loads and the *cftool -l* command will indicate the state of the driver. The logfile message looks as follows:

```
CF: (TRACE): JoinServer: Startup.
```

cftool -l prints the state of the node as in the following:

```
root@fuji2> cftool -l
Node      Number  State      Os
fuji2     --      COMINGUP   --
```

This indicates that the driver is loaded and that the node is trying to join a cluster. If the errorlog message above does not appear in the logfile or the *cftool -l* command fails, then the device driver is not loading. If there is no indication in the /var/log/messages file or on the console why the CF device driver is not loading, it could be that the CF kernel binaries or commands are corrupted, and you might need uninstall and reinstall CF. Before any further steps can be taken, the device driver must be loaded.

After the CF device driver is loaded, it attempts to join a cluster as indicated by the following message:

```
CF: (TRACE): JoinServer: Startup
```

The join server will attempt to contact another node on the configured interconnects. If one or more other nodes have already started a cluster, this node will attempt to join that cluster. The following message in the error log indicates that this has occurred:

```
CF: Giving UP Mastering (Cluster already Running).
```

If this message does not appear in the error log, then the node did not see any other node communicating on the configured interconnects and it will start a cluster of its own. The following two messages will indicate that a node has formed its own cluster as follows:

```
CF: Local Node fuji2 Created Cluster FUJI. (#0000 1)
CF: Node fuji2 Joined Cluster FUJI. (#0000 1)
```

At this point, we have verified that the CF device driver is loading and the node is attempting to join a cluster. In the following list, problems are described with corrective actions. Find the problem description that most closely matches the symptoms of the node being investigated and follow the steps outlined there.

8.2.1.2 Solving join-related problems

Problem

The following are typical join problems.

The node does not join an existing cluster; it forms a cluster of its own.

Diagnosis

The error log shows the following messages:

```
CF: (TRACE): JoinServer: Startup.  
CF: Local Node fuji2 Created Cluster FUJI. (#0000 1)  
CF: Node fuji2 Joined Cluster FUJI. (#0000 1)
```

This indicates that the CF devices are all operating normally and suggests that the problem is occurring some place in the interconnect. The first step is to determine if the node can see the other nodes in the cluster over the interconnect. Use *cftool*(1M) to send an echo request to all the nodes of the cluster:

```
root@fuji2> cftool -e  
Localdev Srcdev Address Cluster Node Number Joystate  
3 2 00.03.47.c2.a8.82 FUJI fuji2 2 6  
3 3 00.03.47.d1.af.ec FUJI fuji3 1 6
```

This shows that node fuji3 sees node fuji2 using interconnect device 3 (Localdev) on fuji3 and device 2 (Srcdev) on fuji2. If the *cftool -e* shows only the node itself continue on in this section." If some or all of the expected cluster nodes appear in the list, attempt to rejoin the cluster by unloading the CF driver and then reloading the driver as follows:

```
root@fuji2> cfconfig -u  
root@fuji2> cfconfig -l
```



Note

There is no output from either of these commands, only error messages in the error log.

Problem

The node does not join the cluster and some or all the nodes respond to *cftool -e*.

Diagnosis

At this point, we know that the CF device is loading properly and that this node can communicate with at least one other node in the cluster. We should suspect at this point that the interconnect is missing messages. One way to test this hypothesis is to repeatedly send echo requests and see if the result changes over time, for example:

```
root@fuji2> cftool -e  
Localdev Srcdev Address Cluster Node Number Joystate  
3 2 00.03.47.c2.aa.f9 FUJI fuji2 3 6  
3 2 00.03.47.c2.a8.82 FUJI fuji3 2 6  
3 3 00.03.47.d1.af.ec FUJI fuji4 1 6
```

```
root@fuji2> cftool -e  
Localdev Srcdev Address Cluster Node Number Joystate  
3 2 00.03.47.c2.aa.f9 FUJI fuji2 3 6  
3 2 00.03.47.c2.a8.82 FUJI fuji3 2 6  
3 3 00.03.47.d1.af.ec FUJI fuji4 1 6  
3 3 00.03.47.d1.ae.f9 FUJI fuji5 1 6
```

```
root@fuji2> cftool -e  
Localdev Srcdev Address Cluster Node Number Joystate
```

3	2	00.03.47.c2.aa.f9	FUJI	fuji2	3	6
3	2	00.03.47.c2.a8.82	FUJI	fuji3	2	6
3	3	00.03.47.d1.af.ec	FUJI	fuji4	1	6

```
root@fuji2> cftool -e
```

Localdev	Srcdev	Address	Cluster	Node	Number	Joinstate
3	2	00.03.47.c2.aa.f9	FUJI	fuji2	3	6
3	2	00.03.47.c2.a8.82	FUJI	fuji3	2	6
3	3	00.03.47.d1.af.ec	FUJI	fuji4	1	6
3	3	00.03.47.d1.ae.f9	FUJI	fuji5	1	6

```
root@fuji2> cftool -e
```

Localdev	Srcdev	Address	Cluster	Node	Number	Joinstate
3	2	00.03.47.c2.aa.f9	FUJI	fuji2	3	6
3	2	00.03.47.c2.a8.82	FUJI	fuji3	2	6
3	3	00.03.47.d1.af.ec	FUJI	fuji4	1	6
3	3	00.03.47.d1.ae.f9	FUJI	fuji5	1	6

```
root@fuji2> cftool -e
```

Localdev	Srcdev	Address	Cluster	Node	Number	Joinstate
3	2	00.03.47.c2.aa.f9	FUJI	fuji2	3	6
3	2	00.03.47.c2.a8.82	FUJI	fuji3	2	6
3	3	00.03.47.d1.af.ec	FUJI	fuji4	1	6
3	3	00.03.47.d1.ae.f9	FUJI	fuji5	1	6

Notice that the node fuji5 does not show up in each of the echo requests. This indicates that the connection to the node fuji5 is having errors. Because only this node is exhibiting the symptoms, we focus on that node. First, we need to examine the node to see if the Ethernet commands on that node show any errors. We log on to fuji5 and use the `netstat(8)` or `ip(8)` command to find out the network interface information and errors.

The `netstat(8)` or `ip(8)` command in Linux reports information about the network interfaces.

Further resolution of the problem consists of trying each of the following steps:

- Ensure that the Ethernet cable is securely inserted at each end.
- Try repeated `cftool -e` and look at the `netstat -I` or `ip -s` link. If the results of the `cftool(1M)` are always the same and the input errors are gone or greatly reduced, the problem is solved.
- Replace the Ethernet cable.
- Try a different port in the Ethernet hub or switch or replace the hub or switch, or temporarily use a cross-connect cable.
- Replace the Ethernet adapter in the node.

If none of these steps resolves the problem, then field engineers will have to further diagnose the problem.

Problem

The following console message appears on node fuji3 while node fuji2 is trying to join the cluster with node fuji3:

```
Aug 30 21:31:35 fuji3 kernel: CF: Local node is missing a route from node: fuji2.
Aug 30 21:31:35 fuji3 kernel: CF: missing route on local device: eth1.
Aug 30 21:31:35 fuji3 kernel: CF: Node fuji2 Joined Cluster FUJI. (#0000 3)
```

Diagnosis

Look in `/var/log/messages` on node fuji2.

Same message as on console.

No console messages on node fuji3.

Look in `/var/log/messages` on node fuji3.

```
fuji3:cftool -d
```

Number	Device	Type	Speed	Mtu	State	Configured	Address
--------	--------	------	-------	-----	-------	------------	---------

1	eth0	4	100	1432	UP	YES	00.03.47.c2.a8.82
2	eth1	4	100	1432	UP	YES	00.02.b3.88.09.f1
3	eth2	4	100	1432	UP	NO	00.02.b3.88.09.ea

```
fuji2: cftool -d
```

Number	Device	Type	Speed	Mtu	State	Configured	Address
1	eth0	4	100	1432	UP	YES	00.03.47.c2.a8.3c
2	eth1	4	100	1432	UP	NO	00.02.b3.88.b8.89
3	eth2	4	100	1432	UP	NO	00.02.b3.88.b7.46

Problem

eth1 is not configured are on node fuji2:

Diagnosis

Look in /var/log/messages on node fuji3.

```
Aug 27 16:05:59 fuji3 kernel: e100: eth1 NIC Link is Down
Aug 27 16:06:08 fuji3 kernel: CF: Icf Error: (service err_type route_src route_dst). (#0000 0 2 1 1)
Aug 27 16:06:08 fuji3 kernel: CF: (TRACE): CFSF failure detected: no SFopen: passed to ENS: fuji2.
(#0000 1)
Aug 27 16:06:08 fuji3 kernel: CF: Node fuji2 Left Cluster FUJI. (#00001)
```

Problem

The eth1 device or interconnect temporarily failed. It could be the NIC on either of the cluster nodes or a cable or the following hub problem.

Node in LEFTCLUSTER state

Node fuji2 panicked and has rebooted. The following console message appears on node fuji2:

```
Aug 28 10:38:25 fuji2 kernel: CF: fuji2: busy: local node not DOWN: retrying
```

Diagnosis

Look in /var/log/messages on node fuji2.

```
Aug 28 10:38: fuji2 kernel: CF: (TRACE): JoinServer: Startup.
Aug 28 10:38:25 fuji2 kernel: CF: Giving UP Mastering (Cluster already Running).
Aug 28 10:38:25 fuji2 kernel: CF: fuji3: busy: local node not DOWN: retrying
```

Last message repeats.

No new messages on console or in /var/log/messages on fuji3.

```
fuji3: cftool -n
```

Node	Number	State	Os	Cpu
fuji2	1	LEFTCLUSTER	Linux	Pentium
fuji3	2	UP	Linux	Pentium

Problem

Node fuji2 has left the cluster and has not been declared DOWN.

Fix

```
cftool -k
```

This option will declare a node down. Declaring an operational node down can result in catastrophic consequences, including loss of data in the worst case. If you do not wish to declare a node down, quit this program now.

```
Enter node number: 1
Enter name for node #1: fuji2
```

```
cftool(down): declaring node #1 (fuji2) down  
cftool(down): node fuji2 is down
```

The following console messages then appear on node fuji3:

```
Aug 28 10:47:39 fuji5 kernel: CF: FUJI: fuji2 is Down. (#0000 2)  
Aug 28 10:49:09 fuji5 kernel: CF: Node fuji2 Joined Cluster FUJI. (#0000 2)
```

The following console message appears on node fuji2:

8.3 Collecting Troubleshooting Information

If an error occurs in the PRIMECLUSTER system, collect information required for an investigation from all the nodes that construct the cluster and the cluster management servers. For details on how to collect information, see "Appendix C Troubleshooting" in "PRIMECLUSTER Installation and Administration Guide." Then, contact field engineers.

Chapter 9 Manual pages

This chapter lists the online manual pages for CF, CIP, PAS, Cluster Resource Management Facility, RMS, SF, Web-Based Admin View, RMS Wizards, and Monitoring Agent.

To display a manual page, type the following command:

```
$ man man_page_name
```



Note

To view these manual pages, you must set the MANPATH environment variable so that /etc/opt/FJSVcluster/man is included.

To print a hard copy of a manual page, enter the following command:

```
% man man_page_name | col -b | lpr
```

9.1 CF

System administration

cfconfig

configure or unconfigure a node for a PRIMECLUSTER cluster

cfregd

CF registry synchronization daemon

cfset

apply or modify /etc/default/cluster.config entries into the CF module

cftool

print node communications status for a node or the cluster

rcqconfig

configure or start quorum

rcquery

get quorum state of the cluster

9.2 CIP

System administration

cipconfig

start or stop CIP 2.0

ciptool

retrieve CIP information about local and remote nodes in the cluster

File format

cip.cf

CIP configuration file format

9.3 PAS

System administration

mipcstat

MIPC statistics

clmstat

CLM statistics

9.4 Cluster Resource Management Facility

System administration

clautoconfig

execute of the automatic resource registration

clbackuprdb

save the resource database

clexec

execute the remote command

cldeldevice

delete resource registered by automatic resource registration

clinitreset

reset the resource database

clrestorerdb

restore the resource database

clsetparam

display and change the resource database operational environment

clsetup

set up the resource database

clstartsrc

resource activation

clstopsrc

resource deactivation

clsyncfile

distribute a file between cluster nodes

User command



There is also a "clgettree" command in the Web-Based System Administration tool WSA.

clgettree

display the tree information of the resource database

9.5 RMS

System administration

hvassert

assert (test for) an RMS resource state

hvcmm

start the RMS configuration monitor

hvconfig

display or save the RMS configuration file

hvdisp

display RMS resource information

hvdump

collect debugging information about RMS

hvlogclean

clean RMS log files

hvsetenv

manipulate RMS rc start or AutoStartUp

hvshut

shut down RMS

hvswitch

switch control of an RMS user application resource to another node

hvutil

manipulate availability of an RMS resource

File formats

config.us

RMS configuration file

hvenv.local

RMS local environment variables file

hvgdstartup

RMS generic detector startup file

9.6 Shutdown Facility (SF)

System administration

rcsd

Shutdown Daemon of the Shutdown Facility

sdtool

interface tool for the Shutdown Daemon

sfcipher

password encryption

File formats

rcsd.cfg

configuration file for the Shutdown Daemon

SA_blade.cfg

configuration file for Blade Shutdown Agent

SA_ipmi.cfg

configuration file for an Intelligent Platform Management Interface Shutdown Agent

SA_icmp.cfg

configuration file for ICMP Shutdown Agent

SA_vmchkhosr.cfg

configuration file for shutdown agent of vmchkhosr (Host OS check)

SA_libvirtgp.cfg

configuration file for libvirt Shutdown Agent (for panic)

SA_libvirtgr.cfg

configuration file for libvirt Shutdown Agent (for reset)

9.7 Web-Based Admin View

System administration

fjsvwvbs

start or stop Web-Based Admin View

fjsvwvcnf

start, stop, or restart the web server for Web-Based Admin View

wvCntl

start, stop, or get debugging information for Web-Based Admin View

wvGetparam

display Web-Based Admin View's environment variable

wvSetparam

set Web-Based Admin View environment variable

wvstat

display the operating status of Web-Based Admin View

9.8 RMS Wizards

RMS Wizards and RMS Application Wizards

RMS Wizards are documented as html pages in the SMAWRhv-do package on the DVD. After installing this package, the documentation is available in the following directory:

`/usr/opt/reliant/htdocs./wizards.en` (Solaris)

`/usr/opt/reliant/htdocs.linux/wizards.en` (Linux)

System administration

clrwzconfig

Sets up the linking function between the PRIMECLUSTER resource manager and the middleware products after the RMS configuration definitions are activated.

9.9 Monitoring Agent (MA)

cldevparam

changes and displays the tunable operation environment for asynchronous monitoring

clmmbmonctl

starts, stops, restarts, and displays the operating system of the MMB asynchronous monitoring daemon

clmmbsetup

registers, changes, deletes, or displays MMB information

Appendix A Release information

This appendix describes primary changes in this manual.

No	VL	Edition	Location	Description
1	4.3A10	June 2011	"7.1 Overview" "7.2 Available SAs" "7.4.2 Shutdown Agents"	Added "VMCHKHOST" and "libvirt" to the Shutdown Agent.
2	4.3A10	June 2011	"7.1 Overview" "7.2 Available SAs" "7.4.2 Shutdown Agents" "9.6 Shutdown Facility (SF)"	Deleted descriptions for "RSB" and "RPS".
3	4.3A10	June 2011	"7.1 Overview" "7.2.8 vmSP"	Added descriptions that "SA_vmSPgp" and "SA_vmSPgr" are for the Xen virtual machine function.
4	4.3A10	June 2011	"7.1 Overview" "7.2.6 VMCHKHOST"	Added descriptions that "SA_vmchkhos" is for the Xen/KVM virtual machine function.
5	4.3A10	June 2011	"7.2 Available SAs" "8.3.2 Crash Dump"	Deleted descriptions for "diskdump".
6	4.3A10	June 2011	"7.2.4 MMB"	Changed "PSA" to "PSA/SVmco".
7	4.3A10	June 2011	"7.2.4 MMB"	Added reference manuals.
8	4.3A10	June 2011	"9.6 Shutdown Facility (SF)"	Added "SA_vmchkhos.cfg", "SA_libvirtgp.cfg", and "SA_libvirtgr.cfg" to the File formats.
9	4.3A20	December 2012	"1.1 CF, CIP, and CIM configuration" "1.1.1 Differences between CIP and CF over IP" "1.1.5 Example of creating a cluster " "1.2 CIP configuration file " "3.1 Resource Database configuration" "3.4.3 Configuring the Resource Database on the new node" "7.4.2 Shutdown Agents" 10.1 Overview	Added descriptions of IPv6.
10	4.3A20	December 2012	"1.1 CF, CIP, and CIM configuration"	Changed the important note when searching for nodes in the CIP configuration.
11	4.3A20	December 2012	"1.1 CF, CIP, and CIM configuration"	Described the configuration which shares the administrative LAN and cluster interconnects on the NIC.
12	4.3A20	December 2012	"1.1.3.1 cfcp/cfsh"	Deleted descriptions regarding the security of PRIMECLUSTER.
13	4.3A20	December 2012	"1.1.7 Example of CF configuration by CLI"	Added "Example of CF configuration by CLI."
14	4.3A20	December 2012	"1.2 CIP configuration file "	Added conditions to stop CIP.
15	4.3A20	December 2012	"7.4.2 Shutdown Agents"	Changed descriptions of "user" and "password" in the configuration file for libvirt.
16	4.3A20	December 2012	All	Changed the description of SIS.
17	4.3A30	February 2014	"1.1 CF, CIP, and CIM configuration" CF over IP	Deleted descriptions for CF over IP.

No	VL	Edition	Location	Description
18	4.3A30	February 2014	"1.1 CF, CIP, and CIM configuration"	Added Note "Restriction on the starting and stopping of GLS, and the rebooting for network service of System."
19	4.3A30	February 2014	"1.1.1 Differences between CIP and CF over IP"	Added notes that CF over IP is not supported.
20	4.3A30	February 2014	"1.1.5 Example of creating a cluster "	Changed descriptions when the pop-up window of the configuration window after completion is displayed.
21	4.3A30	February 2014	2.3 Cluster Configuration Backup and Restore (CCBR)	Deleted descriptions for CCBP.
22	4.3A30	February 2014	"7.1 Overview"	Added that the SD checks cluster nodes to be forcibly eliminated to the function of Shutdown Daemon.
23	4.3A30	February 2014	"7.2.4 MMB"	Deleted descriptions for the environment configuration.
24	4.3A30	February 2014	"7.3.1 Administrative LAN"	Changed descriptions for the administrative LAN.
25	4.3A30	February 2014	"8.3.1 Executing the fjsnap or pclsnap Command"	Changed descriptions for the procedure of information collecting.
26	4.3A30	February 2014	"9.5 RMS"	Deleted descriptions for the hvattr command.
27	4.3A30	February 2014	Chapter 11 CF messages and codes	Deleted whole chapter.
28	4.3A40	Sixth edition	"1.1.5 Example of creating a cluster "	Added "Note."
29	4.3A40	Sixth edition	"1.2 CIP configuration file "	Changed the description of the configuration information in the CIP interface for IPv4.
30	4.3A40	Sixth edition	"3.1 Kernel parameters for Resource Database" "3.4 Adding a new node" "7.2.5 vmgp" "7.4.1 Invoking the Configuration Wizard"	Deleted the chapters and sections.
31	4.3A40	Sixth edition	"5.1 Description of the LEFTCLUSTER state" "7.5.1 Starting and stopping SF" "7.5.1.2 Starting and stopping SF automatically (For Red Hat Enterprise Linux 6)" "7.5.1.3 Starting and stopping SF automatically (For Red Hat Enterprise Linux 7)" "8.2.1.2 Solving join-related problems" "9.7 Web-Based Admin View"	Changed the description for RHEL7.
32	4.3A40	Sixth edition	"7.1 Overview" "7.2 Available SAs"	Deleted the descriptions about RHEL5.

No	VL	Edition	Location	Description
			"9.6 Shutdown Facility (SF)"	
33	4.3A40	Sixth edition	"8.3 Collecting Troubleshooting Information"	Changed the description for collecting troubleshooting information.
34	4.3A40	Sixth edition	"9.5 RMS"	Deleted the explanations of the following commands: - hvdist - hvgdmake - hvrclev - hvreset - hvthrottle
35	4.3A40	Seventh edition	7.4.2 Shutdown Agents	Changed the explanations of Blade.

Glossary

Items in this glossary that apply to specific Cluster Foundation components are indicated with the following notation:

- (CF)-Cluster Foundation
- (RMS)-Reliant Monitor Services
- (RCVM)-Volume Manager (not available in all markets)
- (SIS)-Scalable Internet Services

Some of these products may not be installed on your cluster. See field engineers for more information.

AC

See Access Client.

Access Client

GFS kernel module on each node that communicates with the Meta Data Server and provides simultaneous access to a shared file system.

administrative LAN

An optional private local area network (LAN) used for administrative commands to the nodes in the cluster. To provide an extra level of security, normal users do not have access to the administrative LAN. In Cluster Foundation configurations, the System Console and Cluster Console reside on the administrative LAN if one is present.

See also public LAN.

API

See Application Program Interface.

application (RMS)

In the RMS context, an application object is a special resource used to group other resources into a logical collection. Typically, it is used to represent a real-world application or application suite in a high-availability configuration.

Application Program Interface

A shared boundary between a service provider and the application that uses that service.

application template (RMS)

A predefined group of object definition value choices used by the Wizard Tools or the RMS Wizard Kit to create object definitions for a specific type of application.

attribute (RMS)

The part of an object definition that specifies how the base monitor acts and reacts for a particular object type during normal operations.

automatic switchover (RMS)

The procedure by which RMS automatically switches control of a userApplication over to another node after specified conditions are detected.

See also directed switchover (RMS), failover (RMS, SIS), switchover (RMS), symmetrical switchover (RMS).

availability

Availability describes the need of most enterprises to operate applications via the Internet 24 hours a day, 7 days a week. The relationship of the actual to the planned usage time determines the availability of a system.

base cluster foundation (CF)

This Cluster Foundation module resides on top of the basic OS and provides internal interfaces for the CF (Cluster Foundation) functions that the Cluster Foundation services use in the layer above.

See also Cluster Join Services (CF).

base monitor (RMS)

The RMS module that maintains the availability of resources. The base monitor is supported by daemons and detectors. Each node being monitored has its own copy of the base monitor.

Cache Fusion

The improved interprocess communication interface in Oracle 9i that allows logical disk blocks (buffers) to be cached in the local memory of each node. Thus, instead of having to flush a block to disk when an update is required, the block can be copied to another node by passing a message on the interconnect, thereby removing the physical I/O overhead.

CF

See Cluster Join Services (CF).

CF node name (CF)

The CF cluster node name, which is configured when a CF cluster is created.

child (RMS)

A resource defined in the configuration file that has at least one parent. A child can have multiple parents, and can either have children itself (making it also a parent) or no children (making it a leaf object).

See also resource (RMS), object (RMS), parent (RMS).

cluster

A set of computers that work together as a single computing source. Specifically, a cluster performs a distributed form of parallel computing.

See also RMS configuration (RMS).

Cluster Admin

A Java-based, OS-independent management tool for Cluster Foundation products such as CF, SIS, and RMS. Cluster Admin is available from the Web-Based Admin View interface.

See also Cluster Foundation (CF), Scalable Internet Services (SIS), Reliant Monitor Services (RMS), Web-Based Admin View.

Cluster Foundation (CF)

The set of Cluster Foundation modules that provides basic clustering communication services.

See also base cluster foundation (CF).

cluster interconnect

The set of private network connections used exclusively for Cluster Foundation communications.

Cluster Interconnect Protocol

CIP is an interface such as eth0 except the physical layer is built on top of the cluster interconnect.

Cluster Join Services (CF)

This Cluster Foundation module handles the forming of a new cluster and the addition of nodes.

concatenated virtual disk (RCVM)

Concatenated virtual disks consist of two or more pieces on one or more disk drives. They correspond to the sum of their parts. Unlike simple virtual disks where the disk is subdivided into small pieces, the individual disks or partitions are combined to form a single large logical disk.

See also, mirror virtual disk (RCVM), simple virtual disk (RCVM), striped virtual disk (RCVM), virtual disk.

configuration file (RMS)

In the RMS context, the single file that defines the monitored resources and establishes the interdependencies between them. The default name of this file is config.us.

console

See single console.

custom detector (RMS)

See detector (RMS).

custom type (RMS)

See graph (RMS).

daemon

A continuous process that performs a specific function repeatedly.

database node (SIS)

Nodes that maintain the configuration, dynamic data, and statistics in a SIS configuration.

See also gateway node (SIS), service node (SIS), Scalable Internet Services (SIS).

detector (RMS)

A process that monitors the state of a specific object type and reports a change in the resource state to the base monitor.

DHCP

Dynamic Host Control Protocol. A standard method of delivering information to a host at boot time. This is most often used to dynamically assign the host's IP address and netmask, but many other parameters are possible, including domain names, DNS servers, and time servers.

directed switchover (RMS)

The RMS procedure by which an administrator switches control of a userApplication over to another node.

See also automatic switchover (RMS), failover (RMS, SIS), switchover (RMS), symmetrical switchover (RMS).

DOWN (CF)

A node state that indicates that the node is unavailable (marked as down). A LEFTCLUSTER node must be marked as DOWN before it can rejoin a cluster.

See also UP (CF), LEFTCLUSTER (CF), node state (CF).

Enhanced Lock Manager (ELM) (CF)

A light weight, high performance, highly responsive lock manger, specifically designed for providing a high reliability heartbeat messaging mechanism for Cluster Foundation modules.

ENS (CF)

See Event Notification Services (CF).

environment variables

Variables or parameters that are defined globally.

error detection (RMS)

The process of detecting an error. For RMS, this includes initiating a log entry, sending a message to a log file, or making an appropriate recovery response.

Ethernet

LAN standard that is standardized by IEEE 802.3. Currently, except for special uses, nearly all LANs are Ethernets. Originally the expression Ethernet was a LAN standard name for a 10 megabyte per second type LAN, but now it also used as a general term that includes high-speed Ethernets and gigabyte Ethernets.

Event Notification Services (CF)

This Cluster Foundation module provides an atomic-broadcast facility for events.

failover (RMS, SIS)

With SIS, this process switches a failed node to a backup node. With RMS, this process is known as switchover.

See also automatic switchover (RMS), directed switchover (RMS), switchover (RMS), symmetrical switchover (RMS).

gateway node (SIS)

Gateway nodes have an external network interface. All incoming packets are received by this node and forwarded to the selected service node, depending on the scheduling algorithm for the service.

See also service node (SIS), database node (SIS), Scalable Internet Services (SIS).

generic type (RMS)

An object type which has generic properties. A generic type is used to customize RMS for monitoring resources that cannot be assigned to one of the supplied object types.

See also object type (RMS).

GFS Shared File System

A shared file system that allows simultaneous access from multiple Linux(R) systems that are connected to shared disk units, while maintaining data consistency, and allows processing performed by a node to be continued by other nodes even if the first node fails.

The GFS Shared File System can be mounted and used concurrently from multiple nodes.

Global Disk Services

This optional product provides volume management that improves the availability and manageability of information stored on the disk unit of the Storage Area Network (SAN).

Global File Services

This optional product provides direct, simultaneous accessing of the file system on the shared storage unit from two or more nodes within a cluster.

Global Link Services

This optional products provides network high availability solutions by multiplying a network route.

graph (RMS)

See system graph (RMS).

graphical user interface

A computer interface with windows, icons, toolbars, and pull-down menus that is designed to be simpler to use than the command-line interface.

GUI

See graphical user interface.

HBA blockage function

Function that stops the HBA in the switchover source when node switching takes place. This function enables PRIMECLUSTER node switching to be executed at high speed. This function shortens the node switchover time by forcibly stopping all ongoing input/output processes in the HBA of the switchover source.

high availability

A system design philosophy in which redundant resources are employed to avoid single points of failure.

See also Reliant Monitor Services (RMS).

interconnect (CF)

See cluster interconnect.

Internet Protocol address

A numeric address that can be assigned to computers or applications.

See also IP address.

Internode Communications facility

This module is the network transport layer for all Cluster Foundation internode communications. It interfaces by means of OS-dependent code to the network I/O subsystem and guarantees delivery of messages queued for transmission to the destination node in the same sequential order unless the destination node fails.

IP address

See Internet Protocol address.

IP aliasing

This enables several IP addresses (aliases) to be allocated to one physical network interface. With IP aliasing, the user can continue communicating with the same IP address, even though the application is now running on another node.

See also Internet Protocol address.

iRMC (integrated Remote Management Controller)

Abbreviation for integrated Remote Management Controller which is one of the hardware mounted in PRIMEQUEST/PRIMERGY.

JOIN (CF)

See Cluster Join Services (CF).

keyword

A word that has special meaning in a programming language. For example, in an RMS configuration file, the keyword object identifies the kind of definition that follows.

leaf object (RMS)

A bottom object in a system graph. In the configuration file, this object definition is at the beginning of the file. A leaf object does not have children.

LEFTCLUSTER (CF)

A node state that indicates that the node cannot communicate with other nodes in the cluster. That is, the node has left the cluster. The reason for the intermediate LEFTCLUSTER state is to avoid the network partition problem.

See also UP (CF), DOWN (CF), network partition (CF), node state (CF).

link (RMS)

Designates a child or parent relationship between specific resources.

local area network

See Reliant Monitor Services (RMS).

local node

The node from which a command or process is initiated.

See also remote node, mirror virtual disk (RCVM).

log file

The file that contains a record of significant system events or messages. The RMS Wizard Tools, the RMS base monitor, and RMS detectors each maintain their own log files.

Management Board

One of the hardware units installed in PRIMEQUEST.

Management Information Base

A hierarchical database of information about the local network device. The database is maintained by network management software such as an SNMP agent.

See also Simple Network Management Protocol.

MDS

See Meta Data Server.

message

A set of data transmitted from one software process to another process, device, or file.

message queue

A designated memory area which acts as a holding place for messages so they can be processed in the same order they were received.

Meta Data Server

GFS daemon that centrally manages the control information, or meta-data, of a file system.

MIB

See Management Information Base.

MIPC

Mesh Interprocessor Communication

mirror virtual disk (RCVM)

Mirror virtual disks consist of two or more physical devices, and all output operations are performed simultaneously on all of the devices.

See also concatenated virtual disk (RCVM), simple virtual disk (RCVM), striped virtual disk (RCVM), virtual disk.

mirrored disks (RCVM)

A set of disks that contain the same data. If one disk fails, the remaining disks of the set are still available, preventing an interruption in data availability.

See also mirrored pieces (RCVM).

mirrored pieces (RCVM)

Physical pieces that together comprise a mirrored virtual disk. These pieces include mirrored disks and data disks.

See also mirrored disks (RCVM).

MMB

Abbreviation for Management Board, which is one of the hardware units installed in PRIMEQUEST.

mount point

The point in the directory tree where a file system is attached.

multihosting

Multiple controllers simultaneously accessing a set of disk drives.

native operating system

The part of an operating system that is always active and translates system calls into activities.

network partition (CF)

This condition exists when two or more nodes in a cluster cannot communicate over the interconnect; however, with applications still running, the nodes can continue to read and write to a shared device, compromising data integrity.

node

A host that is a member of a cluster.

node state (CF)

Every node in a cluster maintains a local state for every other node in that cluster. The node state of every node in the cluster must be either UP, DOWN, or LEFTCLUSTER.

See also UP (CF), DOWN (CF), LEFTCLUSTER (CF).

object (RMS)

A representation of a physical or virtual resource in the RMS configuration file or in a system graph.

See also leaf object (RMS), object definition (RMS), object type (RMS).

object definition (RMS)

An entry in the configuration file that identifies a resource to be monitored by RMS. Attributes included in the definition specify properties of the corresponding resource.

See also attribute (RMS), object (RMS).

object type (RMS)

A category of similar resources monitored as a group, such as disk drives. Each object type has specific properties, or attributes, which limit or define what monitoring or action can occur. When a resource is associated with a particular object type, attributes associated with that object type are applied to the resource.

See also graph (RMS).

online maintenance

The capability of adding, removing, replacing, or recovering devices without shutting or powering off the node.

operating system dependent (CF)

This module provides an interface between the native operating system and the abstract, OS-independent interface that all Cluster Foundation modules depend upon.

Oracle Real Application Clusters (RAC)

Oracle RAC allows access to all data in a database to users and applications in a clustered or MPP (massively parallel processing) platform. Formerly known as Oracle Parallel Server (OPS).

OSD (CF)

See operating system dependent (CF).

parent (RMS)

An object in the RMS configuration file or system graph that has at least one child.

See also child (RMS), configuration file (RMS), leaf object (RMS), system graph (RMS).

physical IP address

IP address that is assigned directly to the interface (for example, hme0) of a network interface card.

primary node (RMS)

The default node on which a user application comes online when RMS is started. This is always the node name of the first child listed in the userApplication object definition.

PRIMECLUSTER services (CF)

Service modules that provide services and internal interfaces for clustered applications.

private network addresses

Private network addresses are a reserved range of IP addresses specified by the Internet Corporation for Assigned Names and Numbers (ICANN). Modern switches and routers prevent these addresses from being routed to the Internet, allowing two or more organizations to assign the same private addresses for internal use without causing conflicts or security risks.

private resource (RMS)

A resource accessible only by a single node and not accessible to other RMS nodes.

See also resource (RMS), shared resource.

public LAN

The local area network (LAN) by which normal users access a machine.

See also administrative LAN.

queue

See message queue.

redundancy

The capability of one component to assume the resource load of another physically similar component in case the original component fails or is shut down. Common examples include RAID hardware and/or RAID software to replicate data stored on secondary storage devices, multiple network connections to provide alternate data paths, and multiple nodes that can be dynamically reprovisioned to maintain critical services in a cluster.

Reliant Monitor Services (RMS)

The package that maintains high availability of user-specified resources by providing monitoring and switchover capabilities.

remote node

A node that is accessed through a LAN or telecommunications line.

See also local node, node.

reporting message (RMS)

A message that a detector uses to report the state of a particular resource to the base monitor.

resource (RMS)

A hardware or software element (private or shared) that provides a function, such as a mirrored disk, mirrored disk pieces, or a database server. A local resource is monitored only by the local node.

See also private resource (RMS), shared resource.

resource definition (RMS)

See object definition (RMS).

resource label (RMS)

The name of the resource as displayed in a system graph.

resource state (RMS)

Current state of a resource.

RMS

See queue.

RMS commands (RMS)

Commands that enable RMS resources to be administered from the command line.

RMS configuration (RMS)

A configuration made up of two or more nodes connected to shared resources. Each node has its own copy of operating system and RMS software, as well as its own applications.

RMS Wizard Kit (RMS)

RMS configuration products that have been designed for specific applications. Each component of the Wizard Kit includes customized default settings, subapplications, detectors, and scripts.

See also RMS Wizard Tools (RMS), Reliant Monitor Services (RMS).

RMS Wizard Tools (RMS)

A software package composed of various configuration and administration tools used to create and manage applications in an RMS configuration.

See also, RMS Wizard Kit (RMS), Reliant Monitor Services (RMS).

route

In the PRIMECLUSTER Concepts Guide, this term refers to the individual network paths of the redundant cluster interfaces that connect the nodes to each other.

SAN

See Storage Area Network.

scalability

The ability of a computing system to efficiently handle any dynamic change in work load. Scalability is especially important for Internet-based applications where growth caused by Internet usage presents a scalable challenge.

Scalable Internet Services (SIS)

The package that dynamically balances network traffic loads across cluster nodes while maintaining normal client/server sessions for each connection.

script (RMS)

A shell program executed by the base monitor in response to a state transition in a resource. The script may cause the state of a resource to change.

service node (SIS)

Service nodes provide one or more TCP services (such as FTP, Telnet, and HTTP) and receive client requests forwarded by the gateway nodes.

See also database node (SIS), gateway node (SIS), Scalable Internet Services (SIS).

SF

See Shutdown Facility.

shared resource

A resource, such as a disk drive, that is accessible to more than one node.

See also private resource (RMS), resource (RMS).

Shutdown Facility

The Cluster Foundation interface that manages the shutdown and startup of cluster nodes. The SF is automatically invoked during failover operations. It also notifies other Cluster Foundation products of the successful completion of node shutdown so that recovery operations can begin.

Simple Network Management Protocol

A set of protocols that facilitates the exchange of information between managed network devices. The protocols are implemented by software agents residing in the devices. Each agent can read and write data in the local Management Information Base (MIB) in response to SNMP requests from other devices on the network.

See also Management Information Base.

simple virtual disk (RCVM)

Simple virtual disks define either an area within a physical disk partition or an entire partition.

See also concatenated virtual disk (RCVM), mirror virtual disk (RCVM), striped virtual disk (RCVM), virtual disk.

SIS

See Scalable Internet Services (SIS).

SNMP

See Simple Network Management Protocol.

state

See resource state (RMS).

Storage Area Network

The high-speed network that connects multiple, external storage units and storage units with multiple computers. The connections are generally fiber channels.

striped virtual disk (RCVM)

Striped virtual disks consist of two or more pieces. These can be physical partitions or further virtual disks (typically a mirror disk). Sequential I/O operations on the virtual disk can be converted to I/O operations on two or more physical disks. This corresponds to RAID Level 0 (RAID0).

See also concatenated virtual disk (RCVM), mirror virtual disk (RCVM), simple virtual disk (RCVM), virtual disk.

switchover (RMS)

The process by which RMS switches control of a userApplication over from one monitored node to another.

See also automatic switchover (RMS), directed switchover (RMS), failover (RMS, SIS), symmetrical switchover (RMS).

symmetrical switchover (RMS)

This means that every RMS node is able to take on resources from any other RMS node.

See also automatic switchover (RMS), directed switchover (RMS), failover (RMS, SIS), switchover (RMS).

system disk (GDS)

Disk on which the active operating system is installed. System disk refers to the entire disk that contains the slices that are currently operating as one of the following file systems (or the swap area):

/, /usr, /var, /boot, /boot/efi, or swap area

system graph (RMS)

A visual representation (a map) of monitored resources used to develop or interpret the RMS configuration file.

See also configuration file (RMS).

template

See application template (RMS).

type

See object type (RMS).

UP (CF)

A node state that indicates that the node can communicate with other nodes in the cluster.

See also DOWN (CF), LEFTCLUSTER (CF), node state (CF).

virtual disk

A pseudo-device that allows a portion or a combination of physical disks to be treated as a single logical disk. The virtual disk driver is inserted between the highest level of the OS logical input/output (I/O) system and the physical device driver(s), allowing all logical I/O requests to be mapped to the appropriate area on the physical disk(s).

See also concatenated virtual disk (RCVM), mirror virtual disk (RCVM), simple virtual disk (RCVM), striped virtual disk (RCVM).

Web-Based Admin View

A Java-based, OS-independent interface to Cluster Foundation management components.

See also Cluster Admin.

wizard (RMS)

An interactive software tool that creates a specific type of application using pretested object definitions.

Wizard Kit (RMS)

See RMS Wizard Kit (RMS).

Wizard Tools (RMS)

See RMS Wizard Kit (RMS).

Index

[A]		
Adding and removing a node from CIM.....	54	
Add to CIM.....	55	
[C]		
Caused by a cluster partition	60	
Caused by a panic/hung node.....	60	
Caused by staying in the kernel debugger too long.....	60	
CF.....	87	
cfconfig.....	87	
CF Heartbeat monitor.....	53	
CF node information.....	42	
cfregd.....	87	
CF route table.....	41	
CF route tracking.....	39	
cfset.....	87	
cftool.....	87	
CF topology table.....	43	
CIM options.....	55	
CIM Override.....	56	
CIP.....	87	
cip.cf.....	87	
cipconfig.....	87	
CIP name.....	33	
ciptool.....	87	
clautoconfig.....	88	
clbackuprdb.....	88	
cldeldevice.....	88	
cldevparam.....	91	
clxec.....	88	
clgettree.....	88	
clgettree command execution result.....	33	
clgettree command to verify configuration.....	34	
clinitreset.....	34,88	
clmmbmonctl.....	91	
clmmbsetup.....	91	
clmstat.....	88	
clrestorerdb.....	88	
clrwzconfig.....	91	
clsetparam.....	88	
clsetup.....	33,34,88	
clstartsrc.....	88	
clstopsrc.....	88	
clsyncfile.....	88	
Cluster resource management.....	33	
Cluster Resource Management Facility.....	88	
config.us.....	89	
[D]		
Default StartingWaitTime.....	34	
Description of the LEFTCLUSTER state.....	58	
Displaying statistics.....	51	
Displaying the topology table	42	
[F]		
fjsvwvbs.....	90	
fjsvwvcnf.....	90	
[G]		
GUI administration.....	36	
[H]		
hvassert.....	89	
hvcn.....	89	
hvconfig.....	89	
hvdsp.....	89	
hvdump.....	89	
hvenv.local.....	89	
hvgdstartup.....	89	
hvlogclean.....	89	
hvsetenv.....	89	
hvsht.....	89	
hvswhch.....	89	
hvtul.....	89	
[I]		
init command.....	58	
[L]		
LEFTCLUSTER state.....	58	
List of manual pages.....	87	
[M]		
Main CF table.....	38	
Manual pages.....	87	
Marking nodes DOWN.....	50	
mipstat.....	88	
Monitoring Agent (MA).....	91	
[N]		
Node details.....	41	
Node joins a running cluster.....	34	
[P]		
PAS.....	88	
[R]		
rcqconfig.....	87	
rcquery.....	87	
rcsd.....	89	
rcsd.cfg.....	90	
reboot -f command.....	58	
Recovering from LEFTCLUSTER.....	59	
Release information.....	92	
Resource Database configuration.....	33	
Response Time monitor.....	43	
RMS.....	89	
RMS Wizards.....	90	
[S]		
SA_blade.cfg.....	90	
SA_icmp.cfg.....	90	
SA_ipmi.cfg.....	90	
SA_libvirtgp.cfg.....	90	

SA_libvirtgr.cfg.....	90
SA_vmchghost.cfg.....	90
sdtool.....	89
Search based on keyword.....	51
Search based on severity levels.....	51
Search based on time filter.....	51
sfcipher.....	89
Shutdown Agent.....	59
shutdown command.....	58
Shutdown Facility (SF).....	89
Starting and stopping CF	45
Starting CF.....	47
Starting Cluster Admin GUI and logging in.....	36
Startup synchronization.....	34
Stopping CF.....	48
Synchronization phase.....	34

[U]

Unconfigure CF.....	55
Using PRIMECLUSTER log viewer.....	50

[W]

Web-Based Admin View.....	90
wvCntl.....	90
wvGetparam.....	90
wvSetparam.....	90
wvstat.....	90